공학박사 학위논문

Learning from Imperfect Data: Applicability of A Brain-inspired Algorithm

불완전 데이터 학습 및 뇌 모사 학습의 적용성

2023년 2월

서울대학교 대학원

전기·정보공학부

이 장 호

공학박사 학위논문

Learning from Imperfect Data: Applicability of A Brain-inspired Algorithm

불완전 데이터 학습 및 뇌 모사 학습의 적용성

2023년 2월

서울대학교 대학원

전기·정보공학부

이 장 호

Learning from Imperfect Data: Applicability of A Brain-inspired Algorithm

지도교수 윤 성 로

이 논문을 공학박사 학위논문으로 제출함

2023년 2월

서울대학교 대학원

전기·정보공학부

이 장 호

이장호의 박사 학위논문을 인준함

2023년 2월

위 원 장 조 남 익 (인)

부위원장 윤 성 로 (인)

위 원 곽 노 준 (인)

위 원 정 교 민 (인)

위 원 한 승 주 (인)

Abstract

During the past decade, artificial neural network (ANN) learning through deep learning has achieved significant progress, primarily due to the availability of vast quantities of high-quality data, powerful computer hardware, and effective learning algorithms. High-quality labeled data with accurate labels assigned to the data has enabled artificial neural network models to achieve beyond human capabilities. However, the real-world data can have imperfect supervision that indicates some or all of the data is not labeled or only weak labels are given if the label exists. Another component leading to the success of deep learning is the efficient learning algorithm known as backpropagation. Backpropagation is introduced to learn an ANN which simulates the human activity and cognition performed by the human brain. However, it has been criticized for its biological implausibility in terms of learning algorithms. This dissertation proposes a practical recognition approach for imperfect supervision conditions, as well as findings on how brain-inspired learning algorithms affect imperfect supervision recognition.

The first research of this dissertation is semi-supervised learning, where only a small portion of labeled data exists. In the real world, collected data may exhibit an uneven distribution of classes, and *not all labels* may exist simultaneously. This problem is called class-imbalanced semi-supervised learning. We propose a methodology to address this problem based on an existing semi-supervised learning method. The primary issue with class-imbalanced semi-supervised learning is that the network produces a biased prediction for the majority of classes having relatively large samples. We observe that this problem occurred on the existing semi-supervised learning algorithm and introduce a masking strategy-based objective function that can effectively mitigate the classification bias problem derived from the imbalanced class distribution. Contrary to the previous explicit studies, our approach improves recognition perfor-

mance under the class-imbalanced semi-supervised data protocol.

The second research of this dissertation is weakly supervised learning, where data has *not explicitly labeled*, or approximate labels exist in the data. Human recognition is achieved by consolidating information from multiple sensory organs, commonly known as a multimodal recognition problem in machine learning. We propose a methodology for solving the audio-visual event localization problem, which jointly solves both localization of the temporal boundary of an event and event category recognition. The unconstrained videos have the issue of semantic mismatch between visual and auditory information, particularly at the event transition boundaries. The proposed methodology enhances temporal information within audio and video modalities in feature space and helps to match semantic information between different modalities. The experimental results show that the proposed model effectively aggregates the two modality information to solve the video event identification problem.

The third research of this dissertation is an approach to recognizing unsupervised learning where *labels do not exist* in the data. We consider the deep clustering problem as a representative algorithm of unsupervised learning. To effectively solve this problem, we adopt a contrastive learning approach because it is known to learn discriminative representation without labels. In contrastive learning, two new data are generated through stochastic data augmentation for the same data instance. Next, the distance between the data features created in the same instance is minimized, while the distances between the features created in different instances are maximized. At this time, in contrast to learning-based learning, the class collision problem inevitably occurs. This is a problem that is the same class but is recognized as different classes by the objective function of contrastive learning and learned far away. We effectively improve deep clustering performance by introducing an objective function to suppress this problem and propose a method to use features generated in the middle of the model for contrastive learning.

The final research topic is identifying the effectiveness of a brain-inspired algorithm on imperfect data recognition. Learning by the backpropagation algorithm, which has led to the success of current deep learning, is limited in that it cannot correctly simulate the human brain. We assume that biologically more reasonable learning algorithms, called brain-inspired, can improve the performance of problems that humans perform well. Under these assumptions, we apply the brain-inspired learning algorithms, called predictive coding, to continuous learning, unbalanced data, and adversarial number learning, and perform comparisons with existing error inversion. We analyze the above results based on the neuroplasticity of the human brain, interpret them in terms of the interaction between the hippocampus and the prefrontal cortex, and explore the biologically plausible learning potential.

Through this dissertation, we present a methodology to solve the recognition problem of the imperfect data environment and explore the performance improvement of the recognition problem through brain-inspired learning. The imperfect data recognition problem is one of the most common problems in the real world, and effectively solving it is essential considering data curation's time and cost efficiency. Finally, we discover the potential of brain-inspired learning algorithms to reach the ultimate goal of artificial intelligence.

keywords: Deep Learning, Machine Learning, Imperfect Supervision, Limited Data Recognition, Brain-inspired Learning

student number: 2016-20954

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Modern deep learning research has been studied to achieve artificial general intelligence (AGI) by simulating high-level cognitive activities of the human brain. The primary objective of artificial general intelligence is associated with the cognitive function that the sensory organs operate, such as when we see with our eyes or hear with our ears. It is analogous to the cognitive process in which a person attends to a specific item, observes an object, saves a particular piece of information, acquires new information, and resolves issues. Research to mimic the perception of various sensory organs has been developed from shallow artificial neural networks to modern deep artificial neural networks called deep learning. In particular, it is being intensively researched to replicate human perception in speech recognition related to auditory perception [54, 138], natural language processing related to human language [126, 144], and computer vision representing the human visual systems [125, 188].

In this dissertation, we consider the following question: *Is deep learning an ultimate solution to simulate human perception?* We answer these questions with 'No' for now. Deep learning, also known as artificial neural network learning, has produced achievements in various applications, including image classification [45, 179], object recognition [21, 22], and segmentation [134]. However, previous accomplishments can be achieved when the following three factors are satisfied. The first component for

the successful learning of ANN is a sufficient volume of high-quality labeled data. It assists in learning stabilization and ensures adequate generalization performance. Data-driven features based on these large-scale learning data effectively identify complex and dynamic data relationships. The second factor is computing hardware facilitates the learning of artificial neural networks. The scale of the artificial neural network models developed over the past ten years is progressively growing, and recently announced models like Transformer demand a lot of computational resources [45]. The last element is an efficient algorithm for training artificial neural networks. The backpropagation algorithm proposed by Rumelhart *et al.* [159] is currently the most widely used algorithm for artificial neural network learning. When all three elements mentioned above are fulfilled, we can confirm that artificial intelligence, such as AlphaGo [169], can be developed to outperform human-level performance.

The three success factors discussed above take up a significant portion of deep learning research. Furthermore, satisfying all three elements is challenging in terms of time and cost efficiency. In this dissertation, we identify the limitations in terms of quality data and learning algorithms accessible at the artificial intelligence researcher level and conduct several kinds of research on how to overcome those challenges.

We start by explaining the characteristics of high-quality labeled data as one of the deep learning success factors. The data quality can be evaluated in a variety of ways. We will especially focus on data instances for the recognition problem. Regarding the high-level perspective of the instance, the following factors decide if the data instance is high quality: the target size is appropriate, and the object of recognition is clearly visible inside the image. From the low-level perspective of the instance, the lighting of the target and whether the target is blurred because of hand trembling when taking the target determine the quality of the data. When multiple instances gather, the quality of the data is decided by whether the instances constituting each category are uniformly distributed. Regarding data labels, the degree of label allocation across all data present in the data determines the data quality. The extensively used data for artificial neural

network learning often presupposes that there are an equal number of samples in each category [32, 43, 97]. However, because these are well-balanced data, their properties may differ from those collected in the real world. We collectively refer to data that do not satisfy the enumerated data quality properties as imperfect data. With the recent advances in artificial intelligence research, diverse, imperfect data recognition situations have emerged. Solving problems related to these conditions is essential because these environments are similar to the distribution of data collected in the real world.

Chapter 3 addresses the semi-supervised learning problem, where only a few labeled data are available. At the same time, we consider a class-imbalanced semi-supervised experimental setting where the given data have an unbalanced number for each category. In this environment, the problem called classification bias often occurs, which is known that the prediction for minor classes is biased toward majority classes with a relatively large number of samples. This issue is a type of confirmation bias [5] that occurs in semi-supervised learning, and it is important to prevent bias from occurring in majority classes early stage of learning. This chapter is based on the following paper:

- Jangho Lee, Jaihyun Koh, Seungryong Yoo, Sungroh Yoon, "Rethinking Masked Samples in Class-Imbalanced Semi-Supervised Learning," *International Conference on Pattern Recognition (ICPR)*, August 2022.

In Chapter 4, we address the problem of weakly supervised learning in which human does not explicitly annotate labels. We aim to solve the audio-visual event localization problem, a recognition problem using information from different sensory organs in the video. The semantic label for each time step is set to an event category that occurs equally in the audio and visual modality. At this time, we consider the supervised setting, where the label of the event exists for each time step, and the weakly supervised setting, where the event's label exists at a video level. Since an expert did not take it as an unconstrained video used at this time, visual and auditory information tends to change abruptly. When event information for each modality changes, it

does not produce the same semantic information. For successful audio-visual event localization, it is vital to extract and utilize the characteristics of each modality. Further, it is also crucial to effectively amalgamate information from different modalities. We propose a temporary relationship alignment module to solve the problem of semi-mismatch between the temporary relationship enhancement module and the inter-modality to extract information well within the uni-modality. This chapter is based on the following paper:

- Jangho Lee, Jungbeom Lee, Jaihyun Koh, Sungroh Yoon, "Cross-Modal Temporal Semantic Alignment for Audio-Visual Event Localization," *International Conference on Pattern Recognition (ICPR)*, August 2022.

In Chapter 5, we research unsupervised learning, an algorithm that learns patterns from unlabeled data. Clustering is the representative problem in the absence of labeled data, and deep clustering is the research field that attempts to solve it through deep learning. It is critical to learn discriminative features without label information in deep clustering. We use a contrastive learning-based approach to deep clustering to effectively solve this problem. Contrastive learning, like siamese and triplet networks, reduces distance between instances with similar semantics. In general, it produces two stochastic augmentations on the same image. The distance between the features of the same image is minimized (positive pair), while the distance between the features generated by different images is maximized (negative pair). Features learned through contrastive learning are known to have properties called alignment and uniformity [190]. Alignment denotes a densely mapped distribution within the same category, whereas uniformity denotes that features from different categories fill a given feature space in a spacious manner. Here, we consider the class collision problem, which is unavoidable in the contrastive learning scenario. This problem refers to samples of the same category that are not classified as positive in the mini-batch but are classified as negative. To mitigate these effects, we introduce learning objectives and use network-intermediate features for learning rather than just the encoder's final features for learn-

ing, thereby improving the discriminative property at the low level of the feature. This chapter is based on the following paper:

- Jangho Lee, Seungryong Yoo, Chaehun Shin, Sungroh Yoon, "Fetching Clustering-Favorable Representation via Information Refinement," *International Conference on Pattern Recognition (ICPR)*, August 2022.

Next, this dissertation discusses the learning algorithm aspects among the success factors of deep learning. The algorithm used for modern artificial neural network learning is backpropagation. The backpropagation algorithm performs learning by utilizing the global error signal that occurs in the last layer of the artificial neural network. This error updates the parameters of the neural network by transmitting an error to the front of the artificial neural network through an algorithm such as a gradient descent algorithm. However, the learning behavior of backpropagation has been criticized for being incompatible with the human brain's learning behavior, anatomically and physiologically. Backpropagation is generally considered a biologically inappropriate learning algorithm due to its structure and learning characteristics. Recently, studies have been actively conducted to improve and overcome the biologically incomplete properties of backpropagation.

In Chapter 6, we discuss a deep learning approach using predictive coding, one of the biologically plausible learning algorithms. Predictive coding updates the parameters by local learning rules, where some parameters are updated by the values of parameters located nearby. These properties make predictive coding more biologically plausible. With the development of deep learning, various networks have been developed that exceed human capabilities, but in certain tasks, human performance is still overwhelmingly ahead of artificial intelligence. We collectively refer to these tasks as machine challenge tasks (MCTs), perform learning via predictive coding on three widely known applications in machine learning, and analyze experimental results. This chapter is based on the following paper:

- Jangho Lee, Jeonghee Jo, Byounghwa Lee, Jung-Hoon Lee, Sungroh Yoon,

"Brain-inspired Predictive Coding Improves the Performance of Machine Challenging Tasks," *Frontiers in Computational Neuroscience*, November 2022.

The remainder of this dissertation is structured as follows: The background for this dissertation is presented in Chapter 2, the proposed methods and substantial results are presented in Chapters 3-6, and we conclude this dissertation and suggest future research for imperfect data recognition in Chapter 7.

Chapter 2

Background

This dissertation provides the prerequisite information to understand the following chapters fully. In the previous Chapter, we pointed out data and algorithms as success factors for deep learning. We begin by introducing several imperfect data conditions, such as semi-supervised, weakly supervised, and unsupervised methods. Next, we describe the properties of biological plausibility and the predictive coding algorithm [197], one of the brain-inspired algorithms.

2.1 Imperfect Data Recognition

Imperfect data recognition is important in deep learning because fully supervised learning is sometimes not practical in real-world applications. Since data annotation is a very elaborated and detailed task, the data curator is required substantial competency.

When we curate the classification dataset, the difficulty of data curation increases as the number of categories expands. In the case of segmentation data curation, it requires more time compared to the classification dataset. As it has led to the development of deep learning, it is important to learn a good model using large-scale data, but it is also essential to learn a model that uses limited data to perform fully supervised learning. In this section, we provide three primary imperfect data scenarios depending

on the degree of label annotation and completeness of supervision. Further, we present the current research trends related to each scenario.

2.1.1 Semi-supervised Learning

Semi-supervised learning (SSL) aims to discover an effective way to utilize a vast amount of unlabeled data and enhance the generalization performance of supervised learning [25, 26]. Recent studies on SSL achieved a bit behind but competitive performance in barely supervised experimental settings [177, 211]. Based on several assumptions, the SSL algorithm performed well with a small amount of labeled data. The first is the *manifold assumption* that the decision boundary of the trained network preferably passes through the low-density region [25]. Next is the *cluster assumption* that two samples reside in the same cluster in the input distribution, and they are likely to belong to the same class [26]. Finally, the *smoothness assumption* indicates the correspondence between the input and output spaces [230]. Based on these assumptions, pseudo-labeling [105] and consistency regularization [161] are the most commonly used techniques for SSL. Pseudo-labeling projects unlabeled data points near labeled data points with similar semantics in the feature space. Consistency regularization helps learn the augmentation-invariant representation and makes the network robust for various augmentations [35].

Modern image recognition algorithms [45, 70] are trained using class-wise balanced datasets, such as CIFAR [97] and ImageNet [43]. However, it is laborious to establish balanced datasets with a similar or uniform number of samples for each class [122, 185]. This characteristic is frequently observed in visual inspection [82] and medical image analysis [1]. Previous studies, including re-sampling [27, 68] and re-weighting [36, 131], have been widely used to overcome the class-imbalance problem by adjusting the contribution of each class to class distribution. Furthermore, feature transfer [91, 220] was performed to deliver the acquired information from the majority to the minority classes. Recently, to prevent bias in the classifier, Kang *et al.* [88]

proposed decoupled training that learns the feature extractor and classifier separately.

Class-imbalanced Semi-supervised Learning The class-imbalance problem and SSL have been studied simultaneously. Yang *et al.* [218] began to consider two problems simultaneously and made efforts to learn a balanced feature space through self-supervised learning. Kim *et al.* [90] devised a convex optimization problem to improve the estimated pseudo-labels derived from a biased model. Wei *et al.* [195] introduced a class-rebalancing sampling method that aids in training with less imbalanced data. Nonetheless, the overfitting of minority classifications results from excessive data updates. Previous studies on class-imbalanced SSL can be categorized as *explicit* solutions that can utilize the imbalanced class distribution as prior knowledge to prevent classification bias toward the majority classes. In this dissertation, we propose an *implicit* method that reduces the classification bias using intrinsic properties available in the learning procedure and increases the class-imbalanced classification accuracy without an imbalance prior.

2.1.2 Weakly Supervised Learning

Weakly supervised learning aims to solve the recognition problem when data has not been explicitly labeled or approximate labels exist. Zhou *et al.* [233] categorized the weakly supervised learning into three types: The first scenario is incomplete supervision, in which just a few fully labeled data are provided. The second situation is inexact supervision, which offers merely course labels such as instance category in image segmentation and approximate position in object detection. The final case is inaccurate supervision, which trains on incorrect or mislabeled training data. These data scenarios frequently occur in the real world and are naturally adopted when generation cost is expensive. A representative example is semantic segmentation, which requires pixel-level annotation to perform fully supervised segmentation. Semantic segmentation data construction is labor-intensive and cost-ineffective because it requires as

much complexity as image resolution rather than class label generation. So, weakly supervised semantic segmentation is devised to achieve performance comparable to that of fully supervised semantic segmentation with weak annotations.

Audio-visual Event Localization In this dissertation, we consider the audio-visual event localization (AVE) tasks [182] in a supervised and weakly supervised manner. AVE localization aims to identify the temporal boundary where an event occurs and classify the event category simultaneously. In other words, AVE localization can be regarded as a video temporal segmentation. AVE localization can be approached with two data settings: The first is supervised audio-visual event localization. In this case, both temporal boundary and event category are provided to solve the AVE localization. The second is weakly supervised audio-visual event localization. In this case, the goal is to find an accurate temporal boundary based on only the video-level category.

A dual multimodal residual network (DMRN) [182] introduced the AVE task and its corresponding dataset and established the baseline with audio-guided visual attention and dual long short-term memories (LSTMs) to learn auditory and visual modal inputs, respectively. Multiple studies have implicitly attempted temporal modeling based on the characteristics of the AVE dataset [182]. The audiovisual sequence-to-sequence dual network (AVSDN) [117] created a network with a dual LSTM network to encode global and local information of each modality in a sequence-to-sequence manner. In contrast, the cross-modal attention network (CMAN) [212] expanded the range of attention mechanisms within modal and cross-modal modes, including audio-guided visual attention. Furthermore, the positive sample propagation (PSP) network [231] reinforced the consistency between positive cross-modal connections.

Another approach for solving the AVE task involves learning multiple objectives: Dual attention matching (DAM) [202], wherein a semantic matching mechanism was employed to address the AVE task by jointly solving the cross-modality localization and supervised AVE localization. To learn the interaction between audio and visual

information, the cross-modal relation-aware network (CMRAN) [209] introduced a cross-modal relation-aware module to learn the audiovisual interaction based on self-attention [186]. However, it differs from the original self-attention as the key and value features are the concatenation of the local audio and video features for aggregating the distributed information from the cross-modal inputs. In this dissertation, we attempted to design a module capable of learning temporal semantics by measuring the degree of affinity between temporal features.

2.1.3 Unsupervised Learning

Unsupervised learning is a kind of learning algorithm that learns and identifies patterns from unlabeled data. Because any supervision is provided, it is essential to find meaningful properties of the structure of the dataset. Clustering is a representative unsupervised learning algorithm that groups examples of similar data into identical clusters. In the conventional approach, the K-mean algorithm is the most famous algorithm. It separates the dataset into k discrete clusters with no overlap. At first, all the data instances are randomly assigned into k clusters. K-means algorithm keeps the iteration process until there is no change to the cluster assignments. Unsupervised learning offers the advantage of acquiring effective discriminative features in the absence of labels, but it is more difficult to evaluate the performance of models learned using unsupervised learning than with supervised learning.

Self-supervised learning tries to obtain a flexible and broad representation without human-annotated labels, and it has been intensively researched in the field of machine learning. It helps to learn representation for the unlabeled data by solving user-defined proxy tasks with supervision generated from the data. In the literature, the most dominant approach is contrastive learning, where the information encoded from different views should be similar for a single data instance. In reality, in the picture domain, two randomly enhanced versions of an image referred to as a positive pair should be close to one another in a representation space and far from other images considered negative

pairs. Well-known baselines [28, 69] have shown promising results on diverse downstream tasks [43, 47]. Usually, the quality of a learned representation is evaluated using linear evaluation [28, 69], where a representation that retains discriminative information is regarded as linearly separable. However, Zhao *et al.* [228] recently demonstrated empirically that classification performance alone cannot ensure satisfactory outcomes for subsequent challenges. Several studies utilizing contrastive learning for specific target tasks have been conducted for object detection [206], segmentation [208], and clustering [110].

Despite the success of transfer in downstream tasks, dealing with negative samples that have the same class as the anchor point, called false negatives, is one of the main difficulties in contrastive learning. Pushing against false negatives produces an improper learning signal, so slowing learning and making convergence suboptimal. Chuang *et al.* [30] alleviated the undesirable effects of false negatives by approximating the true negative distribution, and Huynh *et al.* [80] attempted to find candidate false negatives by utilizing an additional augmented view.

Deep Clustering Recently, the previously introduced learning algorithm and technique for unlabeled data are combined into large-scale data training, called deep clustering. Deep clustering is not a well-organized topic because various approaches exist and a data input protocol has not been established clearly. However, deep clustering shares the requirement of discriminative features with image classification. Deep clustering was initially proposed to solve speech-related tasks using deep features [72]. An auto-encoder is utilized to jointly learn better representations and improve performance [62, 207]. DAC [24] proposed a binary pairwise classification framework for image clustering via learnable label features. DCCM [200] not only explored the instance-level information but also analyzed information from different samples to enhance the discriminative power of features. IIC [84] focused on extracting invariant information by applying random perturbation on images and employed an auxiliary

learning objective. PICA [78] employed a partition confidence matrix and explicitly diagonalizes it by minimizing the non-diagonal components. DRC [229] considered and measured the semantics of the row and column space using contrastive learning. CC [110] expanded the work by Zhong *et al.* [229], combining two projection heads composed of two nonlinear layers. Although DRC [229] and CC [110] was based on contrastive learning, they neglect the inherent problems of contrastive learning. We also constructed our method based on the contrastive learning framework, but it overcomed the aforementioned mentioned problems. Previous studies [110, 229] can be regarded as *joint methods* for deep clustering because they solve both the representation learning and cluster assignment simultaneously. Different from joint methods, recently, *sequential methods* [38, 65, 184] achieved better performance rather than joint methods by successively optimizing a representation learning and clustering as a downstream task. In this dissertation, we propose an unsupervised clustering method to consider false negatives and encourage clustering-favorable information.

2.2 Brain-inspired Learning

2.2.1 Biologically Plausible Learning

The backpropagation algorithm [159], which simulates the properties of the human brain, has achieved excellent progress in various machine learning tasks. The algorithm calculates the global error by comparing the predicted outputs and the actual targets at the network's end to achieve an objective. Then, it propagates the error signal to the front of the network to update parameters. Although backpropagation is the most popular learning algorithm for ANNs, it is often regarded as a *biologically implausible* algorithm from a neuroscience perspective. The main reason is that backpropagation does not operate following the local synaptic plasticity [129, 178] as a fundamental property of the nervous system. Synaptic plasticity refers to the ability to reorganize structures or connections by intrinsic or extrinsic stimuli. Another reason is

that the backpropagation requires a copy of the weight matrices to transfer backward error signal [61]. However, retaining synaptic weights on each neuron is impractical in the human brain. So, Lillicrap *et al.* [113] replaced the backward weight matrices with fixed random weights to avoid those problems. According to what Liao *et al.* [112] reported, the signs of backward weight matrices were significant, and when the signs of the forward and backward matrices were concordant, it was possible to obtain the same level of performance or even higher levels of performance. Furthermore, numerous learning algorithms have been developed to improve biological plausibility while preserving classification performance. [2, 106, 119, 150, 197]. Based on the brain's predictive process [155], predictive coding was reported to obtain more biologically plausible properties than the backpropagation algorithm [197]. Furthermore, it achieved comparable performance on arbitrary computational graphs to the backpropagation algorithm.

2.2.2 Predictive Coding

Most architectures in ANNs follow an L-layer structure wherein each layer consists of a set of neurons [159]. The training with the backpropagation algorithm can be explained to minimize a global error generated at the final layer of a network. In the backpropagation algorithm, an activation value of each layer is defined as follows:

$$\hat{v_0} = x \tag{2.1}$$

$$\hat{v_i} = f(\hat{v}_{i-1}; \theta_i) \tag{2.2}$$

where i is the indice of i-th layer, and θ_i is the parameters of i-th layer. The goal of the backpropagation algorithm is to minimize a loss function $\mathcal{L}(\hat{y}, y)$ between the ground-truth target y and the prediction value $\hat{y}$. The final layer output is derived from the forward pass as follows:

$$\hat{y} = f(x; \theta) = \hat{v}_L. \tag{2.3}$$

In the backward pass, the optimization of parameters is performed by the derivative of the loss function. The chain rule and gradient are computed in reverse order as follows:

$$\delta_i = \frac{\partial \mathcal{L}(\hat{y}, y)}{\partial \hat{v}_l} \tag{2.4}$$

and

$$d\theta_i = -\frac{\partial \mathcal{L}(\hat{y}, y)}{\partial \theta_i} \tag{2.5}$$

where δ_i and $d\theta_i$ are the error signal and the gradient from i-th layer, respectively. δ_i and $d\theta_i$ have $\delta_{i+1}\frac{\partial f_{i+1}(\hat{v}_l;\theta_{i+1})}{\partial \hat{v}_l}$ and $-\delta_i\frac{\partial f_i(\hat{v}_{i-1};\theta_i)}{\partial \theta_i}$ from $L-1$ to 1-st layer.

Meanwhile, in the predictive coding algorithm illustrated in Fig. 2.1, an error node e_i is defined in every layer, and the goal of learning is to minimize the collective energy function [14, 16, 51]. The energy function is defined as the sum of prediction errors. A predictive coding network assumes the network as a directed acyclic computational graph $\mathcal{G} = \{\mathcal{E}, \mathcal{V}\}$ to deliver an error from the last layer to the first layer. $\mathcal{E}$ and $\mathcal{V}$ are defined as a set of error nodes $e_i \in \mathcal{E}$ and a set of activation nodes $v_i \in \mathcal{V}$ at every layer.

By analogy to the cortical hierarchy in the human brain, predictive coding can be formulated as a variational inference algorithm [16, 52]. [133] extended predictive coding to an arbitrary computational graph $\mathcal{G}$ considering its hierarchical and generative structure. Given a computational graph $\mathcal{G}$, the feedforward prediction is defined as $p(v_i) = \Pi_i^N p(v_i|\mathcal{P}_i)$ and variational posterior is derived as $Q(\{v_i\}) = \Pi_i^N Q(v_i)$, where $\mathcal{P}(x)$ indicates the set of parent nodes and $\mathcal{C}(x)$ denotes the set of child nodes for the given node x. Each activation node has the prediction $\hat{v}_i = f(\mathcal{P}(v_i); \theta_i) = f(\hat{v}_{i-1}; \theta_i)$ for i-th layer. Based on this, [133] defined an objective function of predictive coding as the variational free energy $\mathcal{F}$ as follows [16, 52]:

$$\mathcal{F} = KL[(Q(\{v_i\})||p(\{v_i\}))] \geq KL[Q(\{v_i\})||p(\{v_{1:N-1}|v_0, v_N\})] \approx \sum_{i=0}^{N} e_i^T e_i \tag{2.6}$$

where a prediction error of each layer e_i.

Recent predictive coding-based studies [133, 158] suppose *fixed prediction assumption*, which indicates the "fixing" the prediction values of the forward pass. We briefly describe the predictive coding mechanism under the fixed prediction assumption. The activation node of the first layer is set to x, and then the following activation nodes are initialized as $v_i = \hat{v}_i$ where $\hat{v}_i$ is calculated by $f_i(\hat{v}_{i-1;\theta})$. Each error node can be calculated as follows:

$$e_i = \hat{v}_i - v_i = f_i(v_{i-1}; \theta_i) - v_i \tag{2.7}$$

and

$$e_L = \frac{\partial \mathcal{L}(\hat{v}_L, y)}{\partial \hat{v}_L}. \tag{2.8}$$

In the backward pass of predictive coding, network parameters θ containing activation nodes $\{v_i\}$ and error nodes $\{e_i\}$ are updated via gradient descent of each layer as follows:

$$v_i \leftarrow v_i + \eta dv_i \tag{2.9}$$

where η is the weight learning rate of predictive coding. dv_i is the gradient of the neuron's activations and is calculated as follows:

$$dv_i = -\frac{\partial \mathcal{F}}{\partial v_i} = e_i - e_{i+1}\frac{\partial f_{i+1}(\hat{v}_i; \theta_{i+1})}{\partial \hat{v}_i}. \tag{2.10}$$

The learning is performed by minimizing the variational free energy $\mathcal{F}$ until converges as follows:

$$\theta_i = \theta_i + \eta d\theta_i \tag{2.11}$$

where η is the weight learning rate. Parameters are updated as follows:

$$d\theta_i = -\frac{\partial \mathcal{F}}{\partial \theta_i} = -e_i\frac{\partial f_i(\hat{v}_{i-1}; \theta_i)}{\partial \theta_i} \tag{2.12}$$

The Eq. 2.11 indicates the local learning rule of the predictive coding where the parameters of i-th layer are only updated based on the e_i and $\hat{v}_{i-1}$.

Predictive coding requires n times more computational cost in terms of time complexity because it repeats backward pass n times. In addition, it requires more memory since it has additional parameters, such as error nodes.

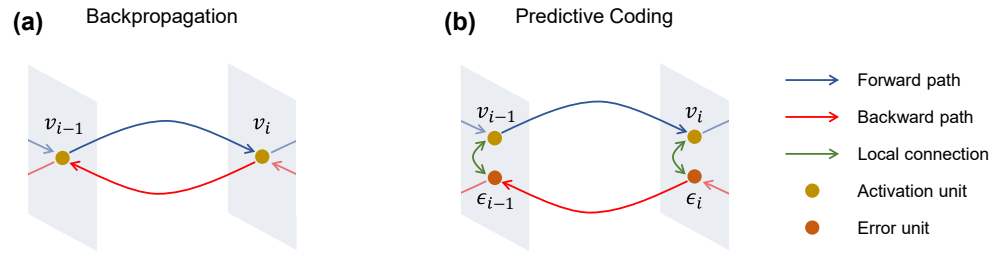

Figure 2.1: Illustration of (a) backpropagation and (b) predictive coding. Different from backpropagation, predictive coding has an error unit ϵ_i for each activation unit v_i and this enables predictive coding to perform local learning.

2.2.3 Machine Challenging Tasks

ANNs have achieved comparable or superior performances to humans by backpropagation in visual recognition [55, 160]. However, ANNs have unsatisfactory performance in certain tasks regarded as simple and easy for human intelligence [20, 57, 171]. As detailed in Section 6, these types of tasks as MCTs (e.g., incremental learning, long-tailed recognition, and few-shot recognition).

Humans ceaselessly take new information from multiple sensory organs and reorganize it in the brain [42, 48]. These processes proceed in a *lifelong manner* because knowledge construction is affected by previous experiences. In addition, humans can refine or transfer knowledge acquired from different types of previous tasks built in an incremental manner [39, 152]. In contrast to human intelligence, ANNs have *catastrophic forgetting* in which the collected information is lost after training of subsequent tasks [57]. Moreover, the human visual system shows robust performances even in limited data recognition, such as long-tailed and few-shot visual recognition. Real-world data commonly follow long-tailed distribution wherein the majority classes occupy the significant part of the dataset and have an open-ended distribution [122]. The primary purpose of long-tailed recognition is to correctly classify the minority class samples to the corresponding targets, reducing the classification bias effect [20]. Further, the classification of tail class samples can be regarded as a few-shot recognition problem

as the degree of imbalance increases [164].

The discrepancy in learning performances between humans and ANNs is closely related to the characteristics of the human brain. First, the human brain operates under two properties: plasticity and stability [178]. Plasticity refers to the brain's change in connectivity and circuitry that enables humans to acquire knowledge, keep memories, and adapt to the external environment [151]. Meanwhile, stability refers to the ability of long-term memory where stable memory is relevant to stable neuron connectivity [174]. A balance between plasticity and stability is achieved with excitatory and inhibitory circuit activity in the visual cortex [178]. Second, the brain engages the hippocampus and neocortex, as explained by the complementary learning system theory that characterizes learning in the brain [152]. The hippocampus focuses on acquiring new knowledge, and knowledge is transferred and generalized to the neocortex via the memory consolidation process. Such mechanisms do not exist in backpropagation. However, they can be indirectly performed in learning predictive coding through the free-energy minimization process of predictive coding. As such, we assume that humans can achieve superior performance in MCTs.

Chapter 3

Learning from Semi-labeled Data

In this dissertation, we propose techniques to handle imperfect supervision scenarios. The degree of imperfect supervision can be categorized into how explicitly the labels are annotated and how many labels are assigned for the whole dataset. The first technique is a method of handling semi-labeled data which implies the scenario under some labeled data and a large amount of unlabeled data.

3.1 Introduction

Semi-supervised learning, which uses unlabeled data to improve the performance of DNNs with minimal labeled data, is an encouraging method for mitigating time-consuming annotation. When labeled data is scarce, such as in medical image analysis [127] and autonomous driving [60], current semi-supervised learning algorithms achieve comparable performance to supervised learning algorithms. Pseudo-labeling [105] and consistency regularization [161] are typically used in semi-supervised learning to handle unlabeled data. *Pseudo-labeling* assigns the highest predicted probability to unlabeled data as its label and trains a classifier with both true and artificial labels. By providing an additional objective function, *consistency regularization* pushes the model to generate a consistent representation or prediction across several views on the unlabeled

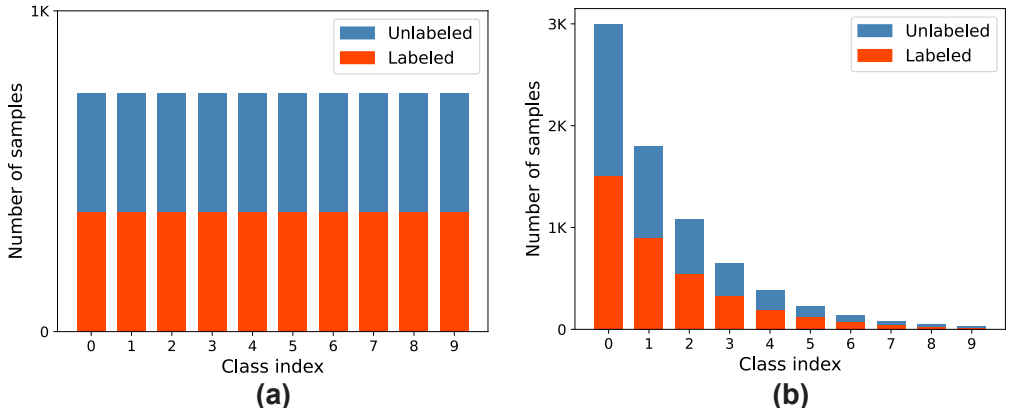

Figure 3.1: The comparison of experimental results on FixMatch with class-imbalanced data (left) and balanced data (right). (a-b) illustrate the data distributions for labeled and unlabeled data. (c-d) represent the class-wise accuracy. (e-f) present the class-wise ratio of samples that exceed the fixed threshold.

data. Modern semi-supervised learning algorithms have significantly improved with multiple labeled data based on these strategies [10, 11, 172].

The most common datasets in machine learning areas are class-wise balanced datasets, where each class is set up to have an equal number of samples [43, 97] or close to them, as seen in Fig. 3.1(a). Nevertheless, it is laborious and time-consuming to generate class-wise balanced datasets, and real-world datasets are frequently substantially imbalanced[109, 122], as depicted in Fig. 3.1(b). The network trained with a class-imbalanced dataset may exhibit a classification bias toward the majority classes because of the data's skewed class distribution [90, 195]. Previous research on resolving the class-imbalance problem in supervised learning settings has primarily focused on reducing classification bias through re-sampling [17, 166], re-weighting [20, 36, 201], and decoupled training [88].

Although class-imbalanced semi-supervised learning is more realistic than class-balanced semi-supervised learning, it has received little attention [90, 195]. Learning with a class-imbalanced dataset reduces the average performance of both super-

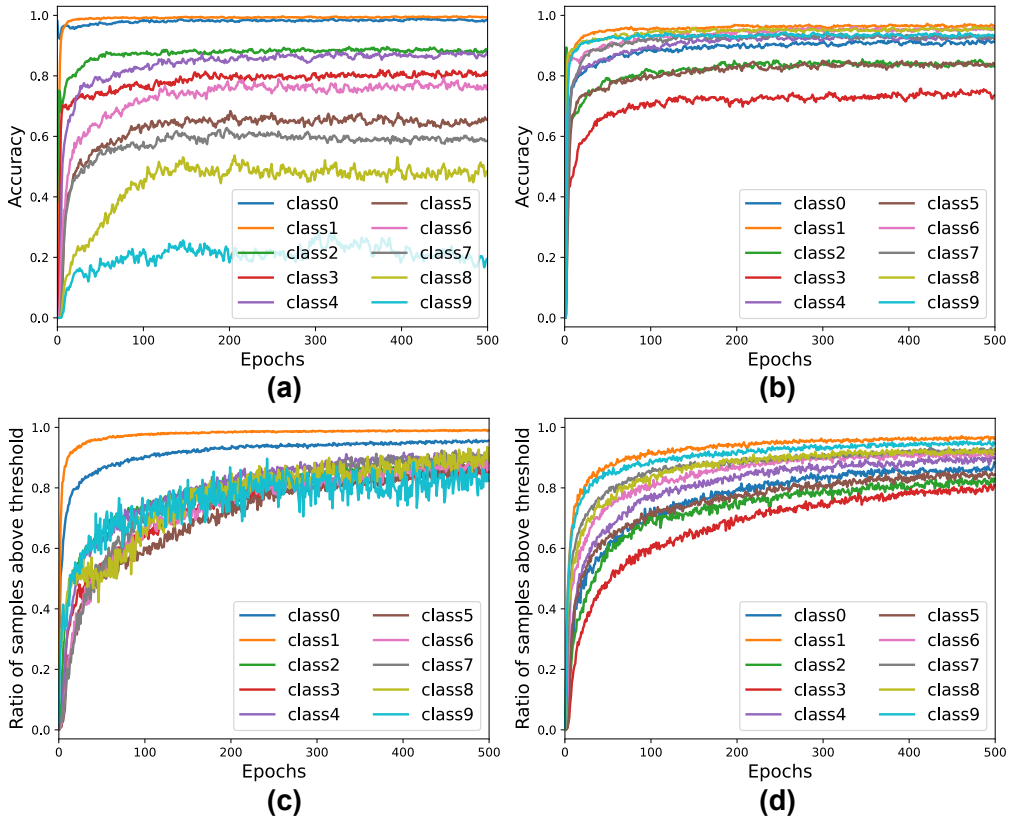

Figure 3.2: The comparison of experimental results on FixMatch with class-imbalanced data (left) and balanced data (right). (a) and (b) represent the class-wise accuracy. (c) and (d) present the class-wise ratio of samples that exceed the fixed threshold.

vised [189] and semi-supervised methods [218]. FixMatch [172] achieved state-of-the-art semi-supervised learning performance on class-balanced data by appropriately incorporating two techniques, pseudo-labeling [105], and consistency regularization [161]. Sohn *et al.* [172] employed a predefined threshold to eliminate samples with low confidence in their prediction and to strengthen the correlation between two images supplemented with different intensities, with the exception of filtered samples. In Fig. 3.2(a)-(b), we observed that FixMatch suffers from a classification bias and confirm that the accuracy of the minority class drastically decreased. We hypothesize that this finding

was influenced by the number of samples from each class. In the early stages of training, a classification bias toward the majority classes increases the number of majority class samples that exceed the predetermined threshold. We visualized the class-wise ratio of the samples that exceeded the fixed threshold in Fig. 3.2(c)-(d). Because of the classification bias, the majority class's ratio rapidly rises compared to the other classes. The classifier is prone to misclassifying true minority class samples as majority class samples.

In this chapter, to mitigate the effects of this skewed data distribution and the resulting bias toward the majority class, we offer a new semi-supervised learning approach that considers the class imbalance. The suggested method presupposes that the class distribution of a mini-batch is skewed. By appropriately excluding the majority class from the minibatch, we could generate a minibatch with less imbalance. Since it is hard to acquire the true label of unlabeled data, we must examine the characteristics of each sample to determine if it belongs to the majority class using intermediate attributes and predictions. First, we focus on the *softmax prediction*. Because the model makes incorrect or ambiguous predictions and produces low confidence in the actual targets, the prediction of minority class samples may make it difficult to pass the threshold. In the latter case, in the semi-supervised learning scenario, minority class samples are near the decision boundary. Second, we focus on *semantic similarity*. Minority class samples in the feature manifold may have lower semantic consistency than majority class samples, particularly when comparing the total number of samples. For this reason, we propose a confidence mask and a semantic mask to build a semi-supervised learning algorithm that is robust to class-imbalanced data.

The contributions of this chapter are summarized as follows:

- We investigated the fixed threshold that causes classification bias and viewed the FixMatch algorithm's learning behavior in semi-supervised learning settings with an uneven number of classes.
- To exclude the majority of class samples from a training minibatch, we propose

a masking method consisting of a confidence mask and a semantic mask.

- We demonstrate that the proposed method can outperform modern semi-supervised learning algorithms without using prior information on three long-tailed image classification datasets.

3.2 Methods

3.2.1 Problem Description

We define semi-supervised image classification for class-imbalanced data as follows: As a L-classification problem, we have a labeled dataset $\mathcal{X} = \{(x_n, y_n) : b \in (1, ..., N)\}$, where x_n and y_n are the labeled data and their corresponding labels. We also have an unlabeled dataset $\mathcal{U} = \{(u_m) : m \in (1, ..., M)\}$, where u_m is the m-th unlabeled data. We denote the number of labeled samples of class l as N_l. The number of samples in l-class are represented in descending order as $N_1 \geq N_2 \geq ... \geq N_L$ and satisfies $\sum_{l=1}^{L} N_l = N$. The imbalance ratio is defined as the proportion of the samples of the highest number of classes to the lowest number of classes as $\gamma = \frac{N_1}{N_L}$. Although it depends on the degree of imbalance, in class-imbalance scenarios, N_1 and N_L satisfy the following relationship, $N_1 \gg N_L$. When an imbalance ratio is determined, the number of samples of each class N_l is parameterized by γ as $N_l = N_1 \cdot \gamma^{-\frac{l-1}{L-1}}$.

Following previous semi-supervised learning algorithms [10, 11, 172], we generate labeled minibatches as $\mathcal{MB}_{\mathcal{X}} = \{(x_b, y_b) : b \in (1, ..., B)\} \subset \mathcal{X}$ and unlabeled minibatches as $\mathcal{MB}_{\mathcal{U}} = \{(u_b) : b \in (1, ..., B)\} \subset \mathcal{U}$ for each iteration, where B is the size of minibatch. We then learn a model that consists of feature extractor f and classifier g, where each subnetwork is parameterized by weights, θ_f and θ_g. The output of classifier indicates the predicted softmax class probability as denoted as $p_m(y|x; \theta)$.

3.2.2 Core Semi-supervised Learning Algorithm

FixMatch [172] is a semi-supervised learning algorithm that makes use of pseudo-labeling and consistency regularization to perform semi-supervised learning. We use it as the backbone semi-supervised learning algorithm because it allows us to guarantee some level of performance on class-imbalanced data. In the supervised loss, FixMatch uses the cross-entropy loss derived from weakly augmented labeled data $\alpha(x)$, as follows:

$$\mathcal{L}_{\mathcal{X}} = \frac{1}{B}\sum_{b=1}^{B}\mathbf{H}(y_b, p_m(y|\alpha(x_b);\theta)) \tag{3.1}$$

where $\mathbf{H}$ represents the cross-entropy loss and y_b are one-hot labels. In the unsupervised loss, FixMatch coerces the coherence between the softmax probability of strongly augmented data $\mathcal{A}(u_b)$ and their pseudo-label q'_b as follows:

$$\mathcal{L}_{\mathcal{U}} = \frac{1}{B}\sum_{b=1}^{B}\sum_{k=1}^{2}\mathbf{I}(\max(q_b) \geq \tau)\mathbf{H}(q'_b, p_m(y|\mathcal{A}(u_b);\theta)) \tag{3.2}$$

where $\mathbf{I}$ is an indicator function, q_b is the softmax probability of u_b, and $\max(q_b)$ is the highest value of the predicted softmax class probability, and τ is the static threshold, and q'_b is the argmax of weakly augmented data $\alpha(u_b)$. Unsupervised loss is selectively applied given the prediction probability of data with the indicator function, where it gives 1 if the provided condition is satisfied, else, 0 is returned.

When it comes to dealing with class-imbalanced data, learning the network only two loss functions does not prevent the classification bias as observed in Fig. 3.2, and this phenomenon is also researched in previous studies through the precision-recall analysis [90, 141, 195]. In particular, the fixed threshold (in general 0.95) is relatively high for tail class samples with a small number of data and thus does not alleviate the class imbalance.

3.2.3 Reuse of Masked Samples

Existing research depends on *explicitly* on the unbalanced distribution to tackle the problem of class imbalanced problem [90, 195]. Wei *et al.* [195] introduced a class-rebalancing sampling approach that combines certain unlabeled images with labeled data by employing the inverted class distribution as the sampling rate. However, it is unsuitable to be aware of and re-sample the data based on the imbalanced distribution of data considering real-world situations. To overcome this issue, we suggest an *implicit* strategy for dealing with the class-imbalanced problem that does neither require a distribution prior nor an additional sampling procedure. Using the intrinsic qualities that occur throughout the training procedure, our method identifies the majority of class samples and rejects them from the learning operation. The proposed strategy used two masks to make the class distribution sensed by the model less uneven in order to find the majority of class samples. The first is a *confidence mask* denoted by $\mathcal{M}_{\mathcal{C}}$, and the second is a *semantic mask* denoted by $\mathcal{M}_{\mathcal{S}}$. Each mask is a binary vector that has the same dimension as a minibatch. With two masks, we propose the following recycling loss:

$$\mathcal{L}_{\mathcal{R}} = \frac{1}{B}\sum_{b=1}^{B}\sum_{k=1}^{2}\mathcal{M}_{\mathcal{C}}(q_b')\mathcal{M}_{\mathcal{S}}(z_b', z_b'')\mathbf{H}(q_b', p_m(y|\mathcal{A}(u_b);\theta)), \tag{3.3}$$

where z' and z'' are the output of feature extractor for $\alpha(u_b)$ and $\mathcal{A}(u_b)$.

3.2.4 Confidence Mask

Our first intuition for efficiently handling the class-imbalanced data is the utilization of masked samples discarded in the learning procedure as described in Fig. 3.2. Due to the high value of fixed threshold, the network is trained with a large number of majority class samples and a small number of minority class samples. Additionally, the minority class samples are less used than the majority class samples when we train the network. These learning properties bring the classification bias toward the majority class. Therefore, we assume that the minority class samples may produce

less peaky softmax class probability compared to those of majority class samples. We try to *implicitly* search minority class samples by using a slightly lower value than the high static threshold of FixMatch and define a confidence mask as follows:

$$\mathcal{M}_{\mathcal{C}}(q') = \mathbf{I}(\max(q') < \tau_c), \tag{3.4}$$

where τ_c denotes the confidence threshold. By applying a confidence mask, we can successfully identify uncertain and ambiguous data that are difficult to assign to a specific class.

3.2.5 Semantic Mask

We focus on the correspondence between distinct augmented perspectives to solve the class-imbalance problem without relying on the prior distribution. In Eq. 3.5, we present a semantic mask that can indirectly filter out samples that belong to the majority class with high-level semantic coherence. In contrast to the confidence mask in Eq. 3.4, we focus on low-level information, such as an intermediate representation, as the softmax probability may contain highly refined information. As a result, to use more low-level information, we create a mask by making use of the features that are output by the backbone. The logic behind the utilization of representation is drawn from the samples that are considered to be members of the minority class. Because of the classification bias, these samples are frequently dispersed across the feature manifold or included within the representations of the samples belonging to the majority class. In class-imbalanced issues, the tendency of majority class samples to exhibit strong recall and minority class samples to produce high precision is well-known [90, 141, 195]. We also made the empirical discovery that the average consistency of weakly and strongly augmented views among minority-class samples was much lower than that of majority-class samples. Based on our data, we constructed the following semantic mask to identify samples with impaired coherence:

$$\mathcal{M}_s(z', z'') = \mathbf{I}(\mathrm{sim}(z', z'') < \tau_s), \tag{3.5}$$

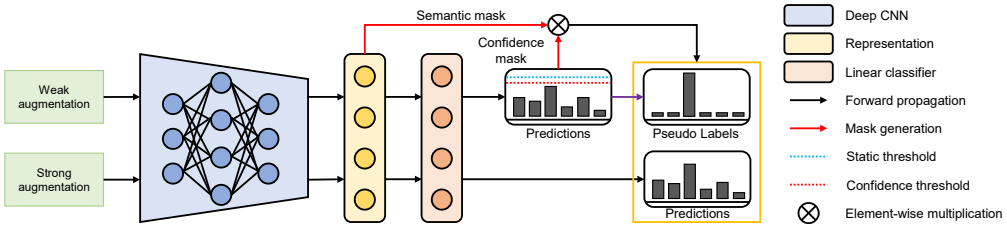

Figure 3.3: Overview of the proposed semi-supervised learning framework. Based on the backbone semi-supervised learning framework, we jointly learn the recycling loss consisting with a confidence mask and a semantic mask generated by the minibatch distribution.

where τ_s is the semantic threshold, and $sim(a, b) = a^\top b/\|a\| \cdot \|b\|$ indicates the dot product between l_2-normalized representation a and b (*i.e.* cosine similarity).

3.2.6 Learning Objectives

In the absence of the class distribution prior, we produce two masks according to Eq. 3.4 and Eq. 3.5. As depicted in Fig. 3.3, we apply them with element-wise multiplication and train the recycling loss collectively with the backbone semi-supervised learning algorithm. The total loss function is expressed as:

$$\mathcal{L} = \mathcal{L}_\mathcal{X} + \mathcal{L}_\mathcal{U} + \mathcal{L}_\mathcal{R}. \tag{3.6}$$

3.3 Experimental Results

3.3.1 Experimental Setup

To demonstrate the efficacy of the proposed method, we considered long-tailed CIFAR-10/100 [97] and STL-10 [32]. Since these datasets were class-wise balanced, we synthetically constructed a class-imbalanced dataset with γ. Additionally, we assumed that unlabeled data share the characteristics of labeled data, as described in [90]. We

used $N_1 = 1500, M_1 = 3000$ for CIFAR-10 and $N_1 = 150, M_1 = 300$ for CIFAR-100, respectively. Since STL-10 does not contain the class information in unlabeled data, we only set N_1 as 450.

3.3.2 Baselines

We compared the proposed method to various baselines, each of which contained supervised algorithms for the class-imbalanced dataset and the current SSL algorithms. First, we look at the most basic kind of supervised learning, which is termed Vanilla and does not include an unlabeled dataset. Then, using the exact same conditions as before, we tested how much of an improvement in performance can be achieved using the re-balancing (RB) methods [20, 83, 88]. Similarly, we conducted experiments on SSL algorithms containing VAT [137], Mean-Teacher [180], MixMatch [11], ReMixMatch [10], and FixMatch [172].

3.3.3 Training Details

All experiments were performed with Wide ResNet-28-2 [222] using a batch size 64. To evaluate classification performance, we measured the *balanced accuracy* (bACC) [77] and *geometric mean scores* (GM) [15]. We report the average performance of the last 20 epochs out of a total of 500 epochs of training. We used random crop and horizontal flip for weak augmentation, while we employed RandomAugment [35] and Cutout [44] for strong augmentation.

3.3.4 Experimental Results on the Same Imbalance Ratio ($\gamma_l = \gamma_u$)

We first experimented when the labeled and unlabeled data shared the same imbalanced distribution in Table 3.1. According to the findings of the experiments, the utilization of recycling loss contributed to a reduction in the severity of the class-imbalance problem under a variety of imbalance settings. In order to conduct a more in-depth evaluation of the usefulness of the suggested approach, the class-wise recall

of CIFAR-10 test and unlabeled data are presented in Table 3.5. We divided CIFAR-10 into three disjoint groups according to the size of the classes: $\{Many, Medium, Few\}$. $Many$ and Few each contain the largest and smallest $\frac{1}{3}$ classes, respectively. For each group, our recycling loss increased the recall by 0.33, 2.25, and 7.80, respectively, from the baseline. From the results, we demonstrated the robustness of our method on class-imbalanced SSL by significantly enhancing the recall of the minority class group.

Table 3.1: Comparison of classification performance (bACC/GM) on CIFAR-10 and CIFAR-100 under class imbalance distribution. We denote the semi-supervised learning approach as SSL and the re-balancing approach as RB. † implies the reproduced results.

			CIFAR-10 ($\gamma = \gamma_l = \gamma_u$)			CIFAR-100 ($\gamma = \gamma_l = \gamma_u$)	
Algorithm	SSL	RB	$\gamma_l = 50$	$\gamma_l = 100$	$\gamma_l = 150$	$\gamma_l = 10$	$\gamma_l = 20$
Vanilla	-	-	65.2 / 61.1	58.8 / 51.0	55.6 / 44.0	55.9 / 50.7	49.5 / 40.3
Re-sampling [83]	-	✓	64.3 / 60.6	55.8 / 45.1	52.2 / 38.2	54.6 / 48.9	48.1 / 38.3
LDAM-DRW [20]	-	✓	68.9 / 76.0	62.8 / 58.9	57.9 / 50.4	55.7 / 51.6	50.4 / 45.4
cRT [88]	-	✓	67.8 / 66.3	63.2 / 59.9	59.3 / 54.6	56.2 / 52.2	50.7 / 43.8
VAT [137]	✓	-	70.6 / 67.8	62.6 / 55.1	57.9 / 46.3	54.6 / 48.6	48.5 / 38.5
Mean-Teacher [180]	✓	-	68.8 / 64.9	60.9 / 52.8	54.5 / 39.8	54.1 / 48.2	48.2 / 37.6
MixMatch† [11]	✓	-	72.9 / 69.4	65.5 / 52.5	62.7 / 42.4	**59.7** / 53.2	**53.2** / 40.1
ReMixMatch† [10]	✓	-	81.4 / 79.8	74.1 / 69.7	69.5 / 61.4	56.7 / 45.0	50.9 / 30.6
FixMatch† [172]	✓	-	80.7 / 79.6	72.7 / 67.6	68.1 / 59.0	56.1 / 49.7	50.6 / 39.5
Ours	✓	-	**82.3 / 81.4**	**75.3 / 72.1**	**70.4 / 65.1**	59.1 / **54.0**	52.2 / **43.5**

3.3.5 Results for Different Imbalance Ratios ($\gamma_l \neq \gamma_u$)

In Table 3.2, we presented additional experimental results under various imbalance ratios, allowing the recycling loss to operate in a more realistic setting. We maintained the imbalance ratio of labeled data at 100 and increased the imbalance ratio of unlabeled data from $1, 50$, and 150. The first condition, when the imbalance ratio is 1, indicates that the unlabeled data follow a uniform distribution. We found that the proposed method consistently achieved better results than the strong baseline known as FixMatch. Moreover, we conducted experiments with reverse-distributed unlabeled data, *i.e.* $M_1 \leq ... \leq M_K$ and $M_k = M_1 \cdot \gamma^{-\frac{k-1}{K-1}}$. Some modern SSL algorithms fail to handle the reversely distributed unlabeled data although they can use unlabeled data compared to re-balancing based supervised learning. However, recycling loss can manage the reversely ordered class distribution because it generates two masks to remove the majority class samples in a minibatch on the fly. In Table 3.3, we provided the experimental results on STL-10. STL-10 naturally shows the different imbalance ratio between labeled and unlabeled data. Under this circumstance, the joint training with recycling loss helped to achieve the superior performance compared to the baselines.

Table 3.2: Comparison of classification performance (bACC/GM) on CIFAR-10 and CIFAR-100 under four different class-imbalance distributions. † implies the reproduced results.

			CIFAR-10 ($\gamma_l = 100$)			
Algorithm	SSL	RB	$\gamma_u = 1$	$\gamma_u = 50$	$\gamma_u = 150$	$\gamma_u = 100$ (reverse)
Vanilla	-	-	58.8 / 51.0	58.8 / 51.0	58.8 / 51.0	58.8 / 51.0
Re-sampling [83]	-	✓	55.8 / 45.1	55.8 / 45.1	55.8 / 45.1	55.8 / 45.1
LDAM-DRW [20]	-	✓	62.8 / 58.9	62.8 / 58.9	62.8 / 58.9	62.8 / 58.9
cRT [88]	-	✓	63.2 / 59.9	63.2 / 59.9	63.2 / 59.9	63.2 / 59.9
VAT [137]	✓	-	65.2 / 59.5	64.0 / 57.3	62.8 / 55.1	59.4 / 50.6
Mean-Teacher [180]	✓	-	73.9 / **71.7**	61.2 / 53.5	59.7 / 50.0	61.0 / 56.4
MixMatch† [11]	✓	-	40.4 / 8.00	64.6 / 50.8	66.3 / 52.9	48.3 / 14.2
ReMixMatch† [10]	✓	-	58.9 / 20.9	75.8 / 71.7	72.2 / 67.4	59.5 / 29.6
FixMatch† [172]	✓	-	68.5 / 38.8	77.1 / 75.3	70.8 / 63.5	61.5 / 30.7
Ours	✓	-	**74.0** / 66.7	**77.9** / **76.2**	**74.1** / **71.4**	**65.1** / **56.9**

3.3.6 Results for Different Imbalance Protocols

There are two main imbalance protocols in class-imbalanced SSL. One is the DARP protocol, where the data distribution is determined by the ratio and proportion of $\frac{M_1}{N_1}$ [90]. The other is the CReST protocol, where the quantity of labeled and unlabeled data is determined by the amount of labeled data for the entire training dataset [195]. Our approach can handle unbalanced data without prior knowledge (PI). In Table 3.4, we compare the performance of classification with that of the latest research requiring data distribution as a precondition. For a fair comparison, we set $\gamma = 100, N_1 = 1000$, and $\beta = 10\%$ for the CIFAR-10. Our proposed recycling loss achieved a competitive classification accuracy for SSL algorithms requiring PI [90, 107, 195].

Table 3.3: Comparison of classification performance (bACC/GM) on STL-10 under class imbalance distribution. † represents the reproduced results.

			STL-10 ($\gamma = \gamma_l \neq \gamma_u$)	
Algorithm	SSL	RB	$\gamma_l = 10$	$\gamma_l = 20$
Vanilla	-	-	56.4 / 51.8	48.1 / 38.2
Re-sampling [83]	-	✓	57.8 / 53.6	47.4 / 35.8
LDAM-DRW [20]	-	✓	58.0 / 54.4	50.2 / 42.4
cRT [88]	-	✓	59.2 / 55.7	49.2 / 42.3
VAT [137]	✓	-	64.2 / 61.1	56.2 / 50.5
Mean-Teacher [180]	✓	-	57.7 / 54.8	48.0 / 35.3
MixMatch† [11]	✓	-	55.4 / 45.3	43.1 / 20.6
ReMixMatch† [10]	✓	-	71.6 / 67.7	61.5 / 45.6
FixMatch† [172]	✓	-	<u>74.4</u> / <u>72.2</u>	<u>67.4</u> / <u>61.3</u>
Ours	✓	-	**76.7 / 75.2**	**70.8 / 67.0**

Table 3.4: Comparison of classification accuracy on CIFAR-10 under CReST protocol [195]. We compare the results reported in [107].

		CIFAR-10
Algorithm	PI	$\gamma = 100, \beta = 10\%$
FixMatch [172]	-	70.0
DARP [90]	✓	74.6
CReST [195]	✓	73.9
ABC [107]	✓	**77.2**
Ours	-	<u>76.9</u>

Table 3.5: Per-class classification recall (%) on CIFAR-10 under class imbalance

		Class Index										
Algorithm	Split	0	1	2	3	4	5	6	7	8	9	Avg.
FixMatch [172]	test	98.4	**99.4**	88.5	79.6	**89.8**	66.8	72.4	61.7	47.3	23.2	72.7
Ours	test	**98.6**	**99.4**	**89.3**	**80.7**	89.4	**67.2**	**80.3**	**64.4**	**52.7**	**38.5**	**76.1**
		+0.2	+0.0	+0.8	+1.1	-0.4	+0.4	+7.9	+2.7	+5.4	+15.3	+3.4
FixMatch [172]	unlabeled	**98.1**	**98.9**	**88.0**	82.2	**88.9**	70.7	75.5	63.9	58.0	16.7	74.1
Ours	unlabeled	97.7	98.7	87.4	**83.3**	87.9	**72.8**	**80.6**	**66.3**	**60.0**	**46.7**	**78.1**
		-0.4	-0.2	-0.6	+1.1	-1.0	+2.1	+5.1	+2.4	+2.0	+30.0	4.0

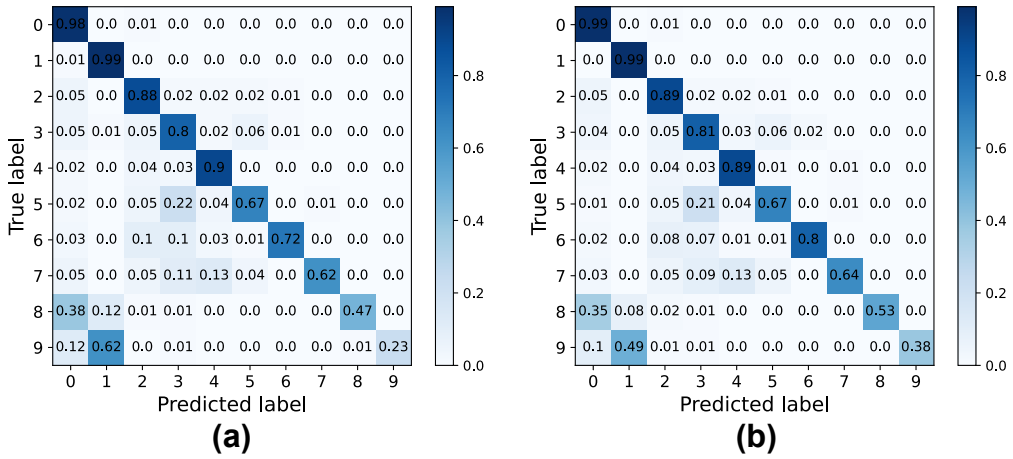

Figure 3.4: Confusion matrices of (a) FixMatch and (b) the proposed algorithm. Our method effectively reduces false-negative predictions of the minority class.

3.4 Discussion

3.4.1 Qualitative Analysis

We compared confusion matrices as to how much the recycling loss reduces the false positives, as shown in Fig. 3.4. Tian *et al.* [181] reported that a strong correlation exists between several classes in the CIFAR-10, such as class 0 (airplane)-class 8 (ship) and class 1 (automobile)-class 9 (truck). We confirmed that such a correlation causes classification bias in conjunction with class-imbalanced characteristics and verified that our method alleviates bias toward majority classes.

To evaluate the representation quality learned using the proposed method, we presented the t-SNE [183] of the learned representations in Fig. 3.5 and visually observe the results in Fig. 3.4. Considering the assumptions of semi-supervised learning mentioned in Sec 3.2.1, FixMatch failed to learn separable representations (*e.g.*, classes 5 and 7), while joint training with recycling loss reduced the inter-class overlap and guided the decision boundary to penetrate low-density regions.

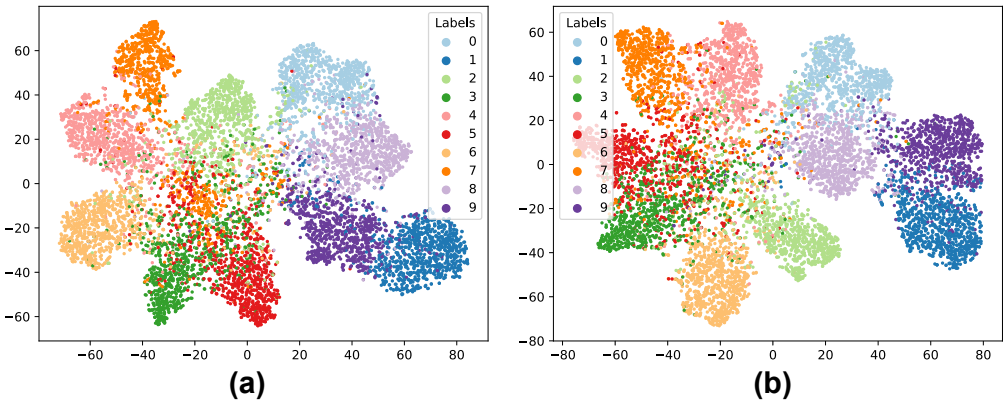

Figure 3.5: t-SNE of (a) FixMatch and (b) the proposed algorithm

Table 3.6: Effect of each mask on CIFAR-10 under class imbalance

	CIFAR-10 ($\gamma = \gamma_l = \gamma_u$)		
Method	$\gamma_l = 50$	$\gamma_l = 100$	$\gamma_l = 150$
w/o $\mathcal{L}_{\mathcal{R}}$	80.7 / 79.6	72.7 / 67.6	68.1 / 59.0
w/ $\mathcal{L}_{\mathcal{R}}$ (only $\mathcal{M}_{\mathcal{C}}$)	81.6 / 80.6	75.0 / 71.8	69.7 / 64.0
w/ $\mathcal{L}_{\mathcal{R}}$ (only $\mathcal{M}_{\mathcal{S}}$)	81.4 / 80.4	74.5 / 71.4	69.4 / 62.4
w/ $\mathcal{L}_{\mathcal{R}}$ ($\mathcal{M}_{\mathcal{C}}$ and $\mathcal{M}_{\mathcal{S}}$)	**82.3 / 81.4**	**75.3 / 72.1**	**70.4 / 65.1**

3.4.2 Ablation Studies

To demonstrate the cooperation between the two masks in Eq. 3.3, we evaluated both the bACC and GM by sequentially inserting the suggested masks in Table 3.6. When analyzing the effects of each mask, we observed that applying the confidence mask results in more improvements than using the semantic mask. We obtained the best performance when the proposed masks are applied together rather than individually.

3.5 Summary

In this chapter, we presented a novel but straightforward method for dealing with class-imbalanced semi-supervised learning. We investigated the behavior of a cutting-edge SSL algorithm in a class-imbalanced scenario and discovered a classification bias toward the majority classes. We propose recycling loss to engage abandoned samples in a learning procedure with two masks inspired by the observation of the state-of-the-art SSL algorithm. The samples with the highest softmax prediction are filtered out using a confidence mask, while inconsistent samples are identified using a semantic mask. A confidence mask filters out majority-class samples based on the assumption that minority-class samples may produce relatively low confidence. In contrast, a semantic mask utilizes intermediate features to filter out data with low coherence between different views. Our experiments show that combining the two proposed masks improves the state-of-the-art SSL algorithm on various long-tail datasets in a class-imbalanced scenario. In particular, our methodology delivers competitive performance on state-of-the-art class-imbalanced methods that explicitly re-balance the classifier depending on the imbalance distribution of labeled data. To further demonstrate the effectiveness of each mask, we also provide a qualitative analysis and ablation research.

Chapter 4

Learning from Weakly Labeled Data

In Chapter 2, we performed the image classification task, which requires explicit image-level annotations among the various types of imperfect supervision. On the other hand, the other tasks, such as image segmentation and video-related tasks, demand pixel-level annotations or labels for each time step. Therefore, it tasks a significant amount of effort to construct the datasets for these tasks. In this situation, the significance of conducting research with a relatively weak label rather than an explicit label is now becoming spotlighted.

4.1 Introduction

Humans unconsciously perform *multisensory integration* of different sensory modalities exhibiting neural mechanisms and causal inferences based on perceptions [95, 167]. In fact, the human perceptual system can discover higher-order knowledge by associating *heterogeneous stimuli* in cognitive science. Concerning the audio-visual perception task, the efficiency of the visual system was observed to have improved with the aid of auditory stimuli. There exist evidence supporting the assumption that multisensory cue combinations result in high-level cognition. Consequently, it has been proposed that the harmonization of auditory and visual stimuli enhances the sig-

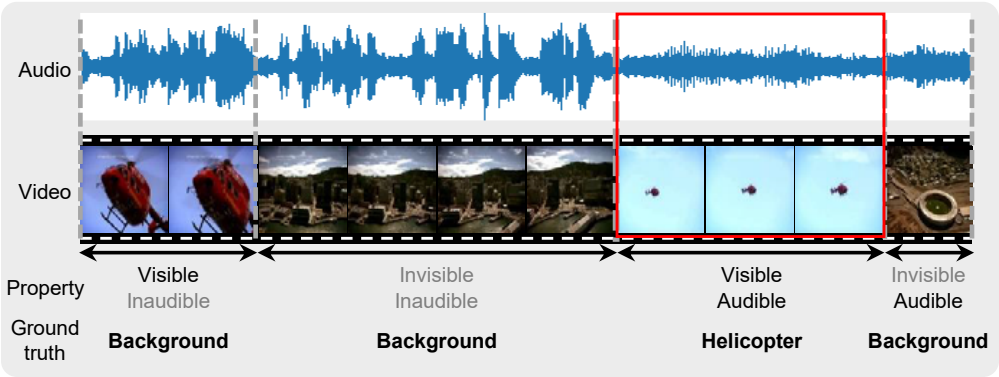

Figure 4.1: Illustration of the audio-visual event localization. The event boundary, as denoted in red box, is labeled when both audio and visual events are jointly observed.

nal detection efficiency owing to the reduction in the disparity between two stimuli sources [49, 95]. Various studies related to machine learning have been conducted on modeling human multisensory integration for cross-modal representation [4], action recognition [89], source localization [76], and conditional generation [234]. Accordingly, audio-visual information has been interpreted as closely related to human perception. Moreover, this information has been proven to act as cues that aid in the development of perceptual inference on counterpart modalities [145, 146, 147, 182]. Thus, in this chapter, how audio-visual event (AVE) localization task can be effectively addressed via understanding the relationship between *heterogeneous stimuli* was examined.

AVE localization aims to temporally localize where specific events have occurred and correctly classify events that occur in an unconstrained video as described in Fig. 4.1. Therefore, the task can be interpreted as jointly performing an event segmentation on the temporal dimension and video recognition problem. Recent studies [117, 182, 202, 209, 212] have proposed various solutions to the AVE localization through attention mechanisms and recorded noteworthy performance. However, the task of AVE localization remains a challenge owing to the existence of *semantic inconsistency* in the unconstrained videos that result from the semantic mismatch

between auditory and visual information. Semantic inconsistency implies that the semantic information included in the visual modality may not necessarily correspond to the semantics of the audio modality as described in Fig. 4.1. Under the circumstances, in unconstrained videos, there exist frequent abrupt cross-modal event transitions, such as video scene shifts and sound source changes. Consequently, these properties may interfere with the correct inference of the model. In this chapter, we propose a novel AVE localization that demonstrates the efficacy of solving the aforementioned problems in cross-modal data by facilitating temporal modeling, as illustrated in Fig 4.2. A *temporal relation enhancement module (TREM)* that guides the model to learn temporal information via simulating the 3D convolutional neural networks (CNNs) as well as learning the temporal properties from discrete features was proposed. In addition, the semantic inconsistency was resolved to a certain extent owing to the expansion of the temporal field of view with the shifting mechanism of the auditory and visual features. Next, a *temporal relation alignment module (TRAM)* reinforces temporal consistency by learning the global scope relation combined with self-attention [186] as proposed. Studies based on the attention mechanism [202, 209] have focused only on local information corresponding to the time step; however, a temporal relation-aware attention mechanism that considers the global relation between cross-modal information was proposed. Moreover, the semantics from different modalities were empirically demonstrated to be well aligned by employing a relation-aware process.

The contributions of this chapter can be summarized as follows:

- We propose a novel architecture composed of two modules to address the AVE localization problem under semantic inconsistency.
- We focus on sequential modeling to enhance the temporal characteristics between the audio and visual modalities. Especially we reinforce the network to be aware of the temporal relation and align the semantic information between cross-modal representations.

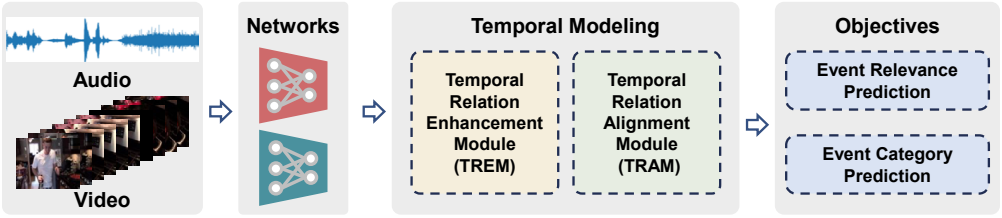

Figure 4.2: Overview architecture of our proposed model for the audio-visual event localization. Audio and visual features are extracted from the pre-trained backbone networks and pass through the two temporal modeling modules. The proposed module is jointly trained with two objective functions in a supervised setting.

- We propose two evaluation metrics for measuring the degree of global and local semantic correspondences.
- The experimental results indicate that the proposed modules exhibit new state-of-the-art performance for AVE localization under supervised and weakly supervised settings.

4.2 Methods

4.2.1 Problem Statement

Adhering to the problem definition of [182], an audio-visual event can be defined as the simultaneous occurrence of audio and visual events as illustrated in Fig. 4.1. Specifically, each video $\{\boldsymbol{V}, \boldsymbol{A}\}$ is divided into T non-overlapping segments, where $\boldsymbol{V} = [\boldsymbol{v}^1, ..., \boldsymbol{v}^T]$, and $\boldsymbol{A} = [\boldsymbol{a}^1, ..., \boldsymbol{a}^T]$ denote visual and audio features that align with the video content, respectively. Each data possesses a second-level label indicating the presence of an event as $\boldsymbol{y}^t \in \{0, 1\}$, and the event-relevance region is defined as $\boldsymbol{y} = \{\boldsymbol{y}^{t,c} | \boldsymbol{y}^{t,c} \in \{0, 1\}, \sum_c^C \boldsymbol{y}^{t,c} = 1\}$, where C is the total number of event categories including the background. In the case of supervised event localization, the model is trained using second-level event annotations y. In contrast, in a weakly su-

pervised event localization, only the video-level labels can be used by averaging the second-level event labels $y = \frac{1}{T}\sum_T y^{t,c}$.

4.2.2 Temporal Relation Enhancement

Because the semantic inconsistency problem is inevitable in unconstrained videos, a model that can explore semantic causality between adjacent data must be designed. Thus, to discover temporal properties embedded in inputs, *TREM*, which encourages the network to simulate 3D CNNs via broadening of the temporal field of view, was introduced. For example, TREM splits the visual feature v into a specific number of chunks in visual features. The 1^{st} chunk was moved to the positive direction of the temporal axis denoted as $\boldsymbol{v}^{+1}$, while the 2^{nd} chunk was shifted in the opposite direction of the temporal axis denoted as $\boldsymbol{v}^{-1}$. The remaining chunks were not subjected to a shift operation denoted as $\boldsymbol{v}^{0}$. Subsequently, the visual features were reconfigured using Eq. 4.1. In addition, the same operation was applied to audio features. Considering the forward propagation of the visual and auditory features on the TREM, a temporal relation of enhanced features can be obtained by merging the features as follows:

$$\begin{aligned} \boldsymbol{v}_e &= w_1\boldsymbol{v}^{-1} + w_2\boldsymbol{v}^{0} + w_3\boldsymbol{v}^{+1}, \\ \boldsymbol{a}_e &= w_1\boldsymbol{a}^{-1} + w_2\boldsymbol{a}^{0} + w_3\boldsymbol{a}^{+1}, \end{aligned} \tag{4.1}$$

where w_1, w_2, and w_3 represent the weight of each feature combination. Following the feature shifting along the temporal dimension, two empty regions on the opposite side of the two shifted features were found to exist, which were then filled with zero values. The TREM acted as a guide for learning the cross-modal relationship via the utilization of the in-sync features between auditory and visual modalities $\{\boldsymbol{v}^t, \boldsymbol{a}^t\}$ coupled with the cooperating off-sync features containing $\{\boldsymbol{v}^t, \boldsymbol{a}^{t-1}\}$ and $\{\boldsymbol{v}^{t+1}, \boldsymbol{a}^t\}$. Consequently, TREM assisted in discovering the relationships between the neighboring features by increasing the temporal field of view. In addition, the shifting mechanism also strengthened the temporal property without additional computation cost

and brings a dramatic performance improvement in the AVE localization without additional memory usage. The proposed TREM differs from TSM [115], wherein the temporal convolution is inserted between specific layers of the network. However, in the TREM, the shifting mechanism is directly applied to the discrete spatial features extracted from cross-modal sources with uniform intervals.

4.2.3 Temporal Relation Alignment

For the model to perform a reliable AVE localization, the auditory and visual features should contain similar semantic information for each time step. However, not all data have the aforementioned characteristics. Thus, *TRAM*, which learns the global scope relation on the cross-modality features, was proposed as illustrated in Fig. 4.3. Motivated by the global attention mechanism [227], the TRAM grasps cross-modal alignments by calculating the pairwise affinity of the t-th temporal feature component with all the temporal relation-aware features as follows:

$$\begin{aligned} \boldsymbol{v}_r &= [\boldsymbol{v}_e^{t,1}, \boldsymbol{v}_e^{t,2}, ..., \boldsymbol{v}_e^{t,T}], \\ \boldsymbol{a}_r &= [\boldsymbol{a}_e^{t,1}, \boldsymbol{a}_e^{t,2}, ..., \boldsymbol{a}_e^{t,T}], \end{aligned} \tag{4.2}$$

where $\boldsymbol{v}_e^{t_1,t_2}$ and $\boldsymbol{a}_e^{t_1,t_2}$ represent the d-dimensional affinity features. For example, $\boldsymbol{v}_e^{t_1,t_2}$ is obtained via the inner product of the two features, $\boldsymbol{v}_e^{t_1}$ and $\boldsymbol{v}_e^{t_2}$. Further, layer normalization [6] was applied to standardize the degree of semantics contained in the multimodal features. Moreover, to determine the optimal temporal agreement between the auditory and visual features, it was applied to the concatenation of different distributions as follows:

$$[\boldsymbol{v}_r; \boldsymbol{a}_r] = \mathrm{LayerNorm}([\boldsymbol{m}_v; \boldsymbol{m}_a]), \tag{4.3}$$

where $\boldsymbol{m}_v = [\boldsymbol{v}_e; \boldsymbol{v}_r]$, and $\boldsymbol{m}_a = [\boldsymbol{a}_e; \boldsymbol{a}_r]$, respectively.

Next, the cross-modal attention was applied to the audiovisual temporal relation-aware features. Similar to the explanation by Xu *et al.* [209], the visual feature $\boldsymbol{v}_r$ was

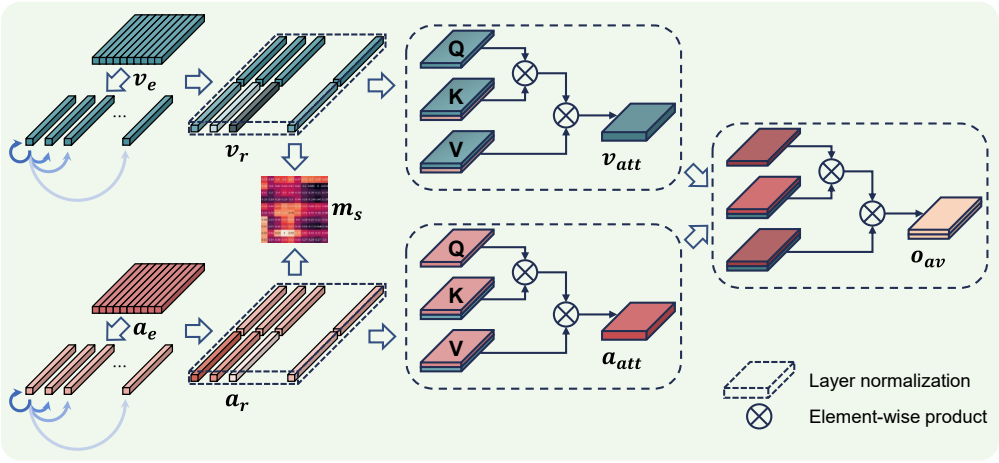

Figure 4.3: Illustration of the proposed TRAM. To find optimal harmonization between cross-modal information, TRAM calculates the temporal relation-aware feature derived by measuring the temporal affinity and normalizing the concatenation of two features.

employed as a query feature, denoted as $\boldsymbol{Q}_v \in \mathbb{R}^{T\times d}$, with the projection parameter $\boldsymbol{W}_v^Q$. In addition, as a key and value feature, the concatenation of visual and auditory features $\boldsymbol{m}_{v,a} \in \mathbb{R}^{2T\times d}$ was employed as well. By projecting temporal relation-aware features with $\boldsymbol{W}_v^K$ and $\boldsymbol{W}_v^V$, the key and value features, $\boldsymbol{K}_{v,a}$ and $\boldsymbol{V}_{v,a}$ dimensions of $2T \times d$ were derived. Further, the cross-modal attentive features were calculated as follows:

$$\boldsymbol{v}_{att} = \sigma(\frac{\boldsymbol{Q}_v \boldsymbol{K}_{v,a}^T}{\sqrt{d}})\boldsymbol{V}_{v,a}, \quad \boldsymbol{a}_{att} = \sigma(\frac{\boldsymbol{Q}_a \boldsymbol{K}_{a,v}^T}{\sqrt{d}})\boldsymbol{V}_{a,v}. \tag{4.4}$$

where $\boldsymbol{Q}, \boldsymbol{K}$, and $\boldsymbol{V}$ are the query, key, and value of attention for each modality, respectively, and are determined as follows:

$$\begin{aligned} \boldsymbol{Q}_v &= \boldsymbol{v_r}\boldsymbol{W}_v^Q, \quad \boldsymbol{K}_{v,a} = \boldsymbol{m}_{v,a}\boldsymbol{W}_v^K, \quad \boldsymbol{V}_{v,a} = \boldsymbol{m}_{v,a}\boldsymbol{W}_v^V \\ \boldsymbol{Q}_a &= \boldsymbol{a_r}\boldsymbol{W}_a^Q, \quad \boldsymbol{K}_{a,v} = \boldsymbol{m}_{a,v}\boldsymbol{W}_a^K, \quad \boldsymbol{V}_{a,v} = \boldsymbol{m}_{a,v}\boldsymbol{W}_a^Q \\ \boldsymbol{m}_{v,a} &= [\boldsymbol{v}_r, \boldsymbol{a}_r], \quad \boldsymbol{m}_{a,v} = [\boldsymbol{a}_r; \boldsymbol{v}_r] \end{aligned} \tag{4.5}$$

where $\boldsymbol{W}_v^Q$, $\boldsymbol{W}_v^K$, and $\boldsymbol{W}_v^V$ denote the learnable parameters for audio-guided visual

self-attention. In a similar manner, $\boldsymbol{W}_a^Q$, $\boldsymbol{W}_a^K$, and $\boldsymbol{W}_a^V$ denote learnable parameters for video-guided audio attention; and σ represents the softmax function. Finally, a similar self-attention was applied to the multiplication of two features, $\boldsymbol{v}_{att}$ and $\boldsymbol{a}_{att}$, to reinforce the cross-modal congruency, and thus obtain $\boldsymbol{o}_{av}$ as a dual-modality feature.

4.2.4 Audio-visual Event Localization

After finishing the audiovisual interaction with the two proposed modules, $\boldsymbol{o}_{av}$ can be obtained with dimensions of $T \times d$-dimensional features. As shown in Fig 4.2, the event relevance prediction and event category prediction were performed with two linear layers, $\boldsymbol{W}_r$ and $\boldsymbol{W}_c$. Consequently, two prediction scores, an event-relevance score $\hat{s}_r$ and an event category score $\hat{s}_c$, were obtained. In the case of the supervised setting, the network was jointly trained using two objective functions for event-relevance prediction and event category prediction, where the former was optimized by a *binary cross-entropy* loss and the latter by a *cross-entropy loss*. In contrast, in the weakly supervised setting, the multi-instance learning (MIL) formulation was followed [199].

However, because only video-level labels can be accessed, as mentioned in 4.2.1, the video-level label was inferred by aggregating individual predictions following MIL pooling. The inference procedure followed is similar to that for the supervised task.

4.3 Experimental Results

4.3.1 Experimental Setup

To demonstrate the effectiveness of the proposed method on AVE localization, it was evaluated by applying it to the AVE dataset [182]. The dataset contains $3339, 402$ and 402 videos for training, validation, and testing, respectively, with each video sampled from AudioSet [56]. The dataset is categorized into 28 event categories, including classes closely related to daily life. Further details regarding the AVE dataset can be found in the original paper [182].

The proposed method used PyTorch 1.2.0 [148] for joint training of two proposed modules on Ubuntu 16.04 LTS. All experiments are performed with 2 $\times$ Intel(R) Xeon(R) CPU E5-2630 v4 @ 2.20GHz, 256GB RAM, 4 $\times$ NVIDIA TESLA V100 GPU.

4.3.2 Implementation detail

The VGG-19 [170] was employed for extracting visual features pre-trained with ImageNet [160]. To create 1-second visual features, global average pooling was applied on the 16 consecutive visual features, based on which $512 \times 7 \times 7$ feature map was generated. To produce the auditory feature, this study employed VGGish [73] pre-trained with AudioSet [56], which produces a 128-dimensional feature per second. Consequently, for training, a batch size of 32, Adam optimizer [92], and a learning rate of 5×10^{-4} were used to learn the proposed model. The learning rate was progressively decayed by multiplying by 0.5 for every 10 epoch until learning reached 30 epochs to avoid overfitting.

To implement the proposed module, we divided the 512-dimensional features into the duration of an input video (10 seconds), and the shift operation was performed on the first two chunks. To calculate the temporal relation-aware features, we calculated the global affinity when i is 1 for the convenience of the computation.

4.3.3 Comparison to the State-of-the-Art: Supervised Training

The effectiveness of the proposed method in the supervised AVE localization setting was demonstrated through comparisons with the recently introduced AVE localization methods in terms of the AVE localization accuracy in Table 4.1. The experimental results indicated that the joint training with the two proposed modules outperformed the strongest competitor, PSP [231] by 0.1%. Although the feature extracted from VGG-19 rather than ResNet-151 was experimented with, the proposed method achieved promising results by outperforming the result obtained via CMAN [212] by 0.8%.

Table 4.1: Comparison to state-of-the-art approaches in supervised and weakly supervised classification accuracy (%) on the AVE dataset.

Method	Supervised	Weakly supervised
ED-TCN [102]	46.9	-
Audio [73]	59.5	-
Visual [170]	55.3	-
AVSDN [117]	72.6	66.8
DMRN [182]	72.7	66.7
DAM [202]	74.5	-
AVIN [154]	75.2	69.4
AV-transformer [118]	76.8	70.2
CMAN [212]	77.1	-
CMRAN [209]	77.4	72.9
PSP [231]	77.8	73.5
Ours	**77.9**	**73.9**

4.3.4 Comparison to the State-of-the-Art: Weakly supervised Training

Table 4.1 reports the performance comparison of the proposed temporal modeling against state-of-the-art methods in a weakly supervised setting. The proposed method exhibited leading performance because it surpassed state-of-the-art performance [231] by 0.4%. Consequently, the comparison indicates that the proposed method overcomes the audiovisual inconsistency, despite the existence of video-level labels only.

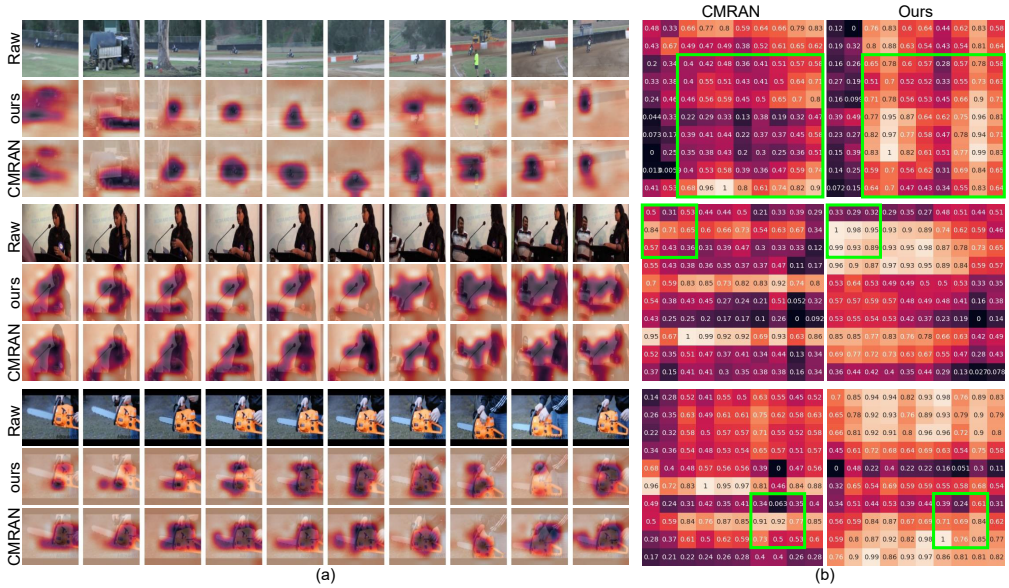

Figure 4.4: Qualitative visualization of (a) audio-guided visual attention and (b) temporal semantic consistency map. The area of green box denotes the ground-truth event boundary where both audio and visual event happens simultaneously.

4.4 Discussion

4.4.1 Effectiveness of TREM

The TREM was introduced to boost temporal information based on the shifting mechanism [115] and thereafter expand it from the visual to auditory features. To verify whether the TREM can aid in addressing the AVE localization task, ablation studies were conducted via the addition or deletion of the TREM in the networks. As reported in Table 4.2 and 4.3, the application of the TREM on both modalities enhanced the event localization performance in supervised and weakly supervised experimental settings. In the supervised setting, the joint training with the TREM achieved 77.99%, a remarkable improvement of 0.55% compared to the previous best method. In the weakly supervised setting, the joint training with the TRAM exhibited a performance improvement pattern similar to that of the supervised setting by increasing the baseline

performance by 0.67%. These promising results can be attributed to the enlargement of the temporal receptive field. Owing to the increase in the temporal field of view for effective cross-modal interaction, various combinations of views were created, and consequently, the network achieved the best fusion. Thus, expanding the temporal field of view can aid in determining at *sweet spot* between neighboring segments and enhance the temporal modeling power of the network. Further, in the supervised setting, it is considered that the performance with the TREM is higher than that of the joint training with the two modules for the following reasons: Considering the data statistics, 66.4% of AVE spans over the duration of the video, boosting the temporal property may have aided n matching the correct event labels in the supervised setting, in contrast to the weakly supervised settings that only use video-level labels.

4.4.2 Effectiveness of TRAM

The TRAM was proposed to determine the relationships and mitigate the temporal semantic inconsistency between cross-modal features. Further, to validate its benefits, the performance of the training with the TRAM was compared while maintaining the cross-modal self-attention as reported in Table 4.2 and 4.3. In the supervised and weakly supervised settings, the TRAM was observed to enhance the baseline performance by 0.20 and 0.19%, respectively. However, because of the different distributions of cross-modality, effectively training cross-modal data is a challenge [191]. Hence, layer normalization was deployed before cross-modal self-attention to harmonize cross-modality representations along the channel axis. Thereafter, the assumption was experimentally demonstrated via the application of layer normalization to the concatenation of the cross-modal representation rather than passing each representation on each normalization layer. Certain examples wherein the proposed modules were added are shown in Fig. 4.5.

4.4.3 Resolve the Temporal Semantic Inconsistency

The effectiveness of the proposed method in alleviating audiovisual temporal semantic inconsistency was demonstrated. Considering the property of unconstrained videos, each video contained candidate objects capable of producing various sounds, including abrupt scene changes, such as camera movement, and included semantic inconsistency. Hence, correctly localizing the temporal event boundary by semantically aligning auditory and visual information remains a challenging problem.

4.4.4 Quantitative Analysis

Two evaluation metrics, that is, the semantic consistency score (SCS), to quantify the extent to which the semantic inconsistency has been resolved and the semantic alignment score (SAS) to measure the degree of semantic alignment under the condition of an unstrained environment were proposed in this study. SCS indicates the *global* correspondence between two modalities, while SAS implies the number of cross-modality features temporally and semantically aligned in the *local* temporal interval. To calculate SCS and SAS, the intermediate features $\boldsymbol{v}_r$ and $\boldsymbol{a}_r$ were considered before entering the cross-modal self-attention in the TRAM, which yielded the temporal semantic consistency map $m_s \in \mathbb{R}^{T\times T}$. Each component of the map, $m_s(t_1, t_2)$, which signifies the semantic affinity between $\boldsymbol{v}_r^{t_1}$ and $\boldsymbol{a}_r^{t_2}$, is measured by $m_s(t_1, t_2) = \frac{\boldsymbol{v}_r^{t_1}\cdot\boldsymbol{a}_r^{t_2}}{\left\|\boldsymbol{v}_r^{t_1}\right\|\cdot\left\|\boldsymbol{a}_r^{t_2}\right\|}$, where $\boldsymbol{v}_r^{t_1}$ and $\boldsymbol{a}_r^{t_2}$ indicate the visual and auditory temporal relation-aware features corresponding to the time steps t_1 and t_2, respectively. Thereafter, the map was normalized to $[0, 1]$ to adjust the degree of semantic alignment. Consequently, using the temporal semantic consistency map m_s, the SCS and SAS were evaluate as follows:

$$\begin{aligned} SCS &= \mathbb{E}[\frac{1}{T}\sum_{t_1,t_2\in[0,T]} m_s(t_1, t_2)] \quad 1 \le t_1, t_2 \le T, \\ SAS &= \mathbb{E}[\sum_{t=t_s}^{t_e} \mathrm{tr}(m_s)] \quad t_s \le t \le t_e, \end{aligned} \tag{4.6}$$

where l_s and l_e indicate the event boundary. When the semantics of audio and visual features match in each time step, the lower bound of SCS is 0, and vice versa SCS has the upper bound of 1. SAS evaluates the degree of alignment for the temporal boundary zone, as opposed to SCS, which analyzes semantic consistency for a given video. SAS has the average temporal length of the ground truth event boundary as the upper bound. The results of SCS and SAS for all the testing datasets are presented in Table 4.4. The results show that the proposed method outperformed the baseline in terms of the proposed semantic alignment metrics.

4.4.5 Qualitative Analysis

In Fig. 4.4, the qualitative results of the audio-guided visual attention and temporal semantic consistency map are shown. The CMRAN [209] was set as the baseline, based on which an AVE localization model was built employing the two learning objectives. Comparing the attention map of CMRAN [209], it was confirmed that the proposed modules accurately focused on the region where the audio-visual correspondence occurred. In particular, the proposed attention maps were found to reduce the false-positive attention region by separating the sounding objects. For example, in the middle sample of Fig. 4.4(a), the results of this study correctly localized the part where the woman speaks into the microphone and the part where another person appears and speaks. Next, how the proposed method reinforced the temporal relationship between neighboring features has been described in Fig. 4.4(b). The intensity of each map indicates the degree of normalized semantic affinity. The bright region represents the semantically well-aligned temporal region, whereas the dark area indicates the semantic misalignment between the visual and auditory inputs. The observation verified the proposed coercing of the temporal semantic consistency compared with the baseline.

In Fig. 4.6 and Fig. 4.7, we visualized additional examples to support the proposed method. The attention maps of the proposed method obtained in a supervised and weakly supervised manner are more precise compared to the CMRAN [209]. The

proposed method correctly localizes and explicitly distinguishes the sound source.

4.5 Summary

This chapter proposed two modules for addressing audio-visual semantic inconsistency in unconstrained videos. TREM was introduced to extract high-quality temporal representation by broadening the temporal field of view on multimodal features. Thereafter, TRAM was proposed for exploring the global scope semantic relation with cross-modal self-attention on the cross-modal features. By jointly training two modules, state-of-the-art AVE localization performance was realized in supervised and weakly supervised experimental settings on the AVE dataset. Moreover, extensive ablation studies verified that the proposed method resolved temporal inconsistency and improved temporal semantic alignment both quantitatively and qualitatively.

Table 4.2: Ablation studies on the supervised setting. † is the reported results of the CMRAN [209]. $\boldsymbol{v}$ and $\boldsymbol{a}$ indicate the usage of each module on the visual and auditory modalities.

TREM ($\boldsymbol{v}$)	TREM ($\boldsymbol{a}$)	TRAM ($\boldsymbol{v}$)	TRAM ($\boldsymbol{a}$)	Acc. (%)
-	-	-	-	77.44†
✓	-	-	-	76.14
-	✓	-	-	77.94
-	-	✓	-	75.97
-	-	-	✓	78.09
✓	✓	-	-	**77.99**
✓	-	✓	-	76.12
✓	-	-	✓	77.44
-	✓	✓	-	76.27
-	✓	-	✓	76.72
-	-	✓	✓	**77.63**
✓	✓	✓	-	76.84
✓	✓	-	✓	76.79
✓	-	✓	✓	77.16
-	✓	✓	✓	77.07
✓	✓	✓	✓	**77.86**

Table 4.3: Ablation studies on the weakly supervised setting. † is the reported results of the CMRAN [209]. $\boldsymbol{v}$ and $\boldsymbol{a}$ indicate the usage of each module on the visual and auditory modalities.

TREM ($\boldsymbol{v}$)	TREM ($\boldsymbol{a}$)	TRAM ($\boldsymbol{v}$)	TRAM ($\boldsymbol{a}$)	Acc. (%)
-	-	-	-	72.94†
✓	-	-	-	73.38
-	✓	-	-	73.06
-	-	✓	-	72.61
-	-	-	✓	72.83
✓	✓	-	-	**73.61**
✓	-	✓	-	72.49
✓	-	-	✓	73.49
-	✓	✓	-	72.39
-	✓	-	✓	72.99
-	-	✓	✓	**73.13**
✓	✓	✓	-	71.34
✓	✓	-	✓	73.06
✓	-	✓	✓	73.86
-	✓	✓	✓	73.86
✓	✓	✓	✓	**73.86**

Table 4.4: Evaluation of the effectiveness of proposed modules on the temporal semantic consistency. †We set the CMRAN as a baseline and compare two metrics.

Method	SCS (↑)	SAS (↑)
Proposed	**0.493**	**3.64**
Baseline†	0.482	3.61

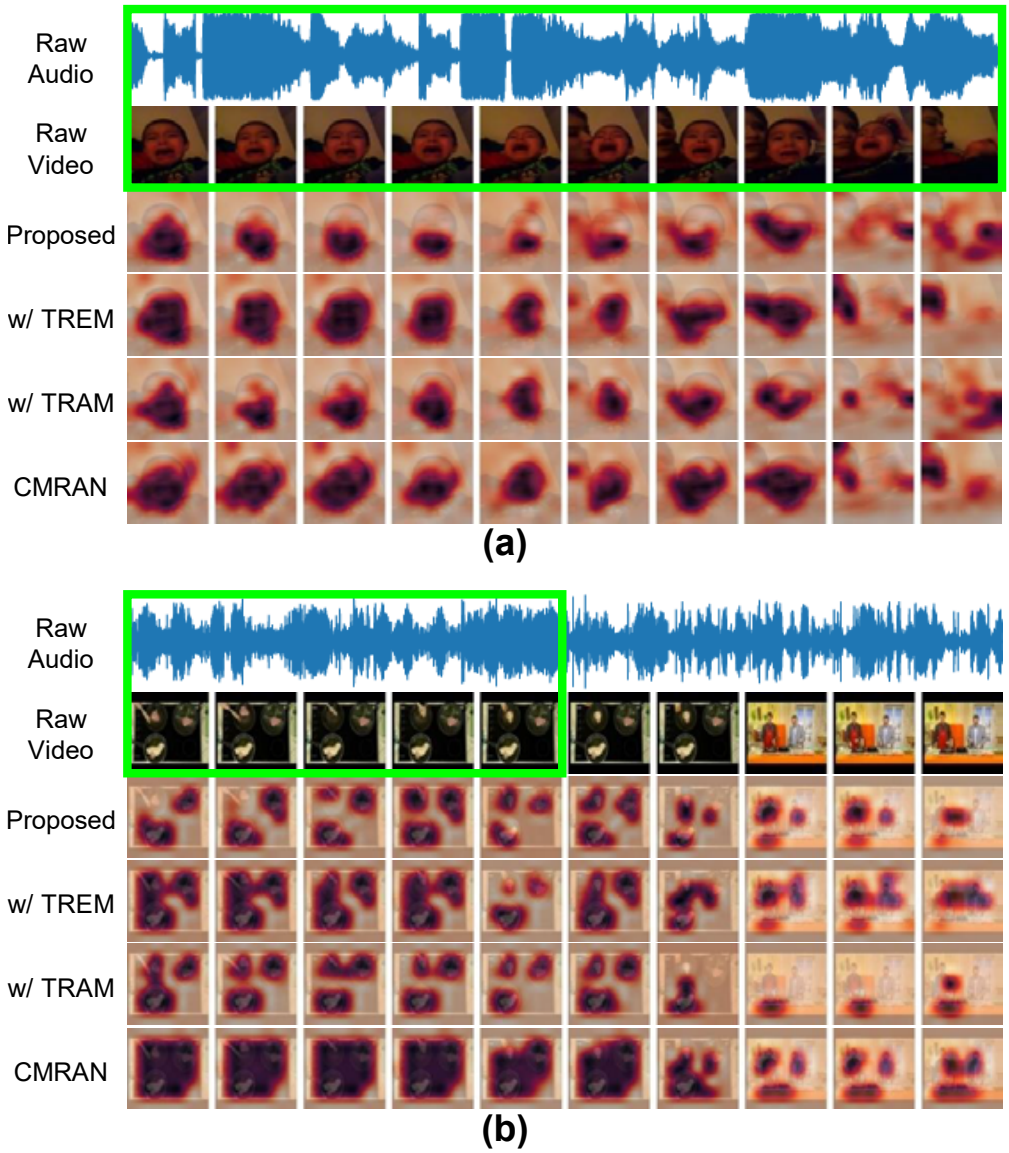

Figure 4.5: Illustration for the ablation study. The area of green box denotes the ground-truth event boundary. (a) represents the example of ”Baby cry, infant cry”, and (b) is the example of ”Frying (food)”. The more temporal modeling is applied, the more accurate localization performance is.

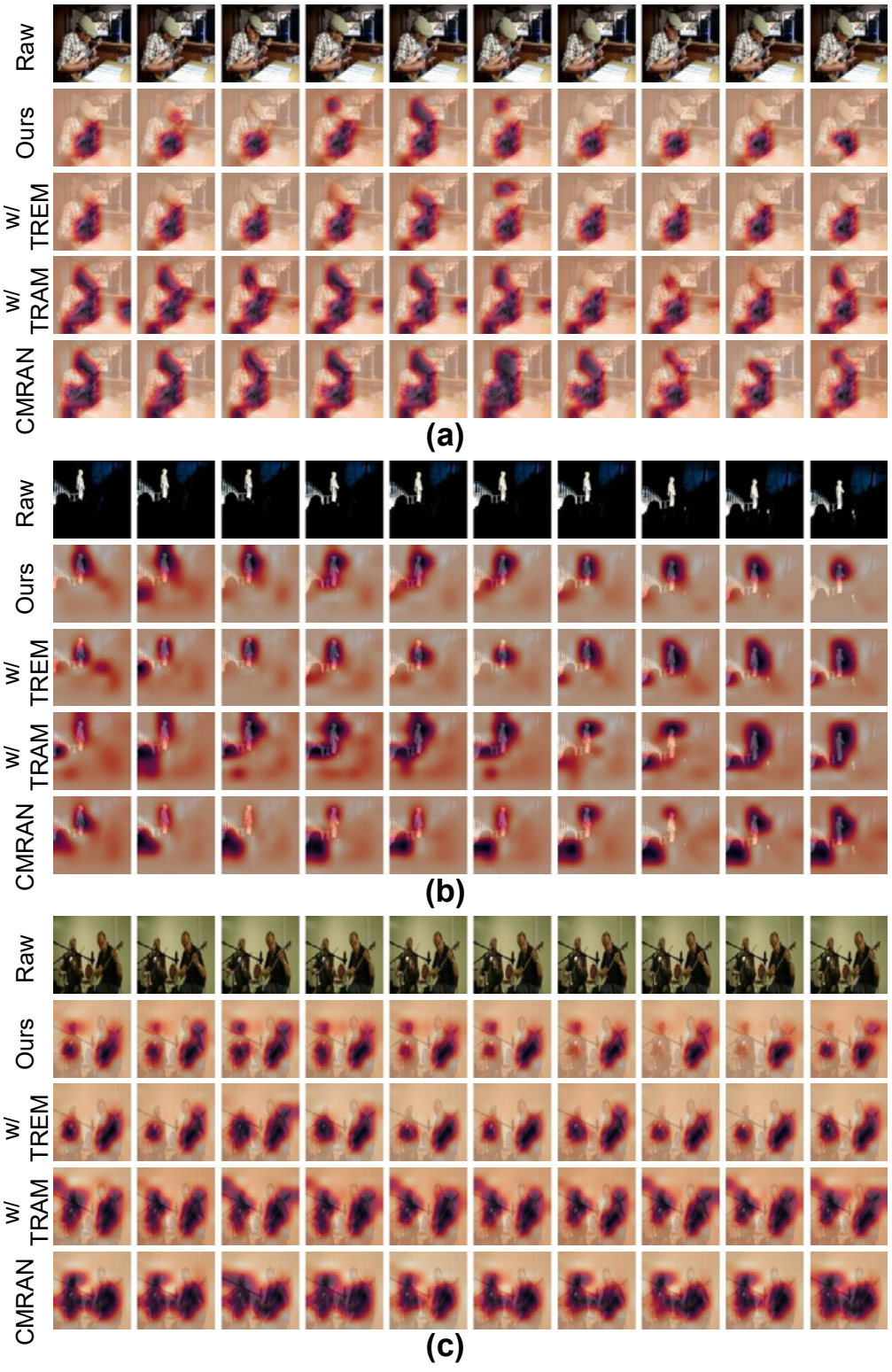

Figure 4.6: Additional visualization examples of the supervised experiments

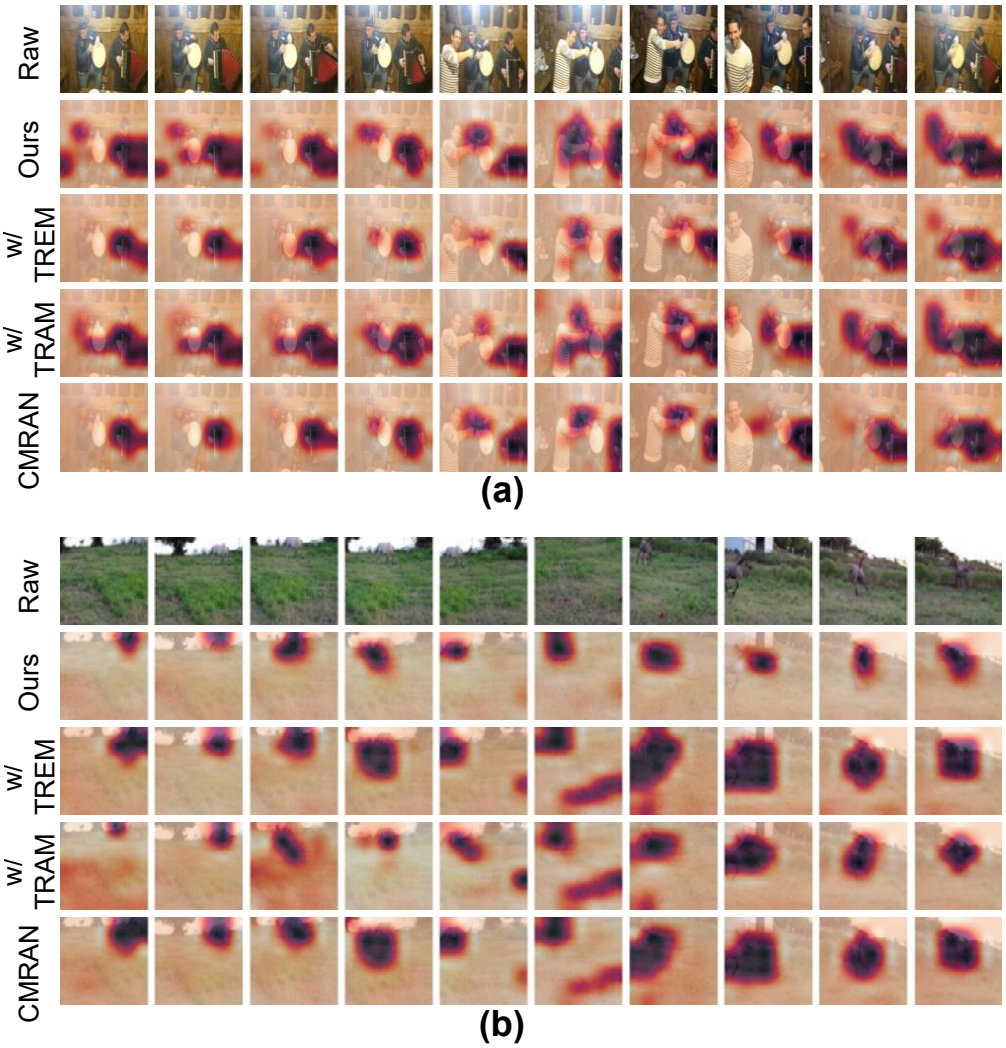

Figure 4.7: Additional visualization examples of the weakly supervised experiments

Chapter 5

Learning from Unlabeled Data

In this chapter, we address the problem with no supervision. Since no supervision is accessible in such situations, it is common to utilize proxy supervision that can be generated from the training data itself. The most representative example of proxy supervision is the generated images through data augmentation. In this case, two images generated from the same data instance are regarded as a positive pair, while augmented images from two different images are considered negative pairs. By utilizing the proxy annotation from the mini-batch, the network is guided to learn discriminative representation.

5.1 Introduction

Clustering is the process of grouping similar data points into the same clusters by minimizing the intra-cluster variance while maximizing the inter-cluster variance without annotations. Traditional approaches [19, 66, 75, 86, 104, 124] have been proposed to discover an optimal partitioning with local descriptors [37, 123, 143]. However, finding optimal clusters with hand-crafted features has two limitations: 1) *capacity of features*: the discriminative features of an entire image cannot be properly extracted. 2) *algorithm susceptibility*: the clustering performance heavily depends on the choice

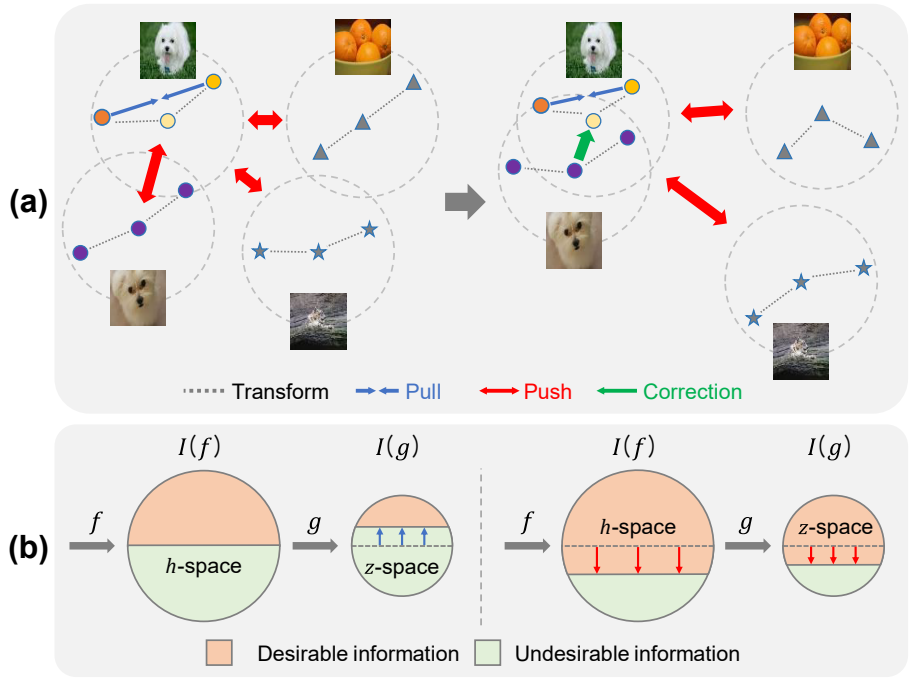

Figure 5.1: Illustration of our motivations and proposed solutions. (a) compares the clustering procedure using naïve contrastive learning and our approach. Grey dashed circles indicate the boundary of random perturbation. We guide that the false positive is located near the anchor in the feature space. (b) visualizes the amount of information before and after the projection head. To make a clustering-favorable space, we apply the contrast on the h-space to provide more information to the projection head.

of similarity metric [168]. Therefore, it is important to learn high-quality features and build a robust algorithm that is less affected by visual similarity.

Deep clustering aims to perform representation learning and clustering without annotations jointly. Early studies [24, 72] directly estimated cluster membership. Subsequently, some studies [62, 63, 78, 140, 200, 207] associated multiple objectives to extract various properties of the input. The latest trend in deep clustering [78, 84, 110, 200, 229] is to not only predict the cluster assignments but solve the contrastive prediction task [28, 69]. In this chapter, we employed the contrastive learning framework [28]

to establish our deep clustering method.

Recently, contrastive learning has become the main idea of self-supervised learning [28, 69, 136, 204]. As illustrated in Fig. 5.1(a), it first generates two perturbed images using stochastic augmentations and learns discriminative representation by encouraging two representations to be pulled closer but pushes apart all the others. Although contrastive learning has proven its effectiveness in deep clustering [110, 229], an inherent problem exists; because no labels are provided, *false negatives* inevitably arise from the learning procedure. Some data that belong to the positive category can be regarded as negative and generate improper learning signals. Another problem is *information compression* which occurs at a nonlinear projection head [28] and may restrict the information that is beneficial for downstream tasks as illustrated in Fig. 5.1(b). To solve the above issues, we developed a novel deep clustering method that leverages a learning objective that can purify a learning signal contaminated by false negatives. In addition, we enforced the latent features having clustering-relevant information by contrasting positive features against negative features.

The contributions of this chapter are summarized as follows:

- We propose a novel end-to-end deep clustering method that creates a clustering-favorable representation to overcome the two abovementioned drawbacks of contrastive learning.
- We propose a feature refinement strategy to correct undesirable learning signals derived from false negatives and exploit informative negatives to improve the volume of desirable information in downstream tasks.
- We apply contrastive learning to the latent space so that the projection head can receive a desirable learning signal.
- We achieve state-of-the-art performance on five challenging datasets.

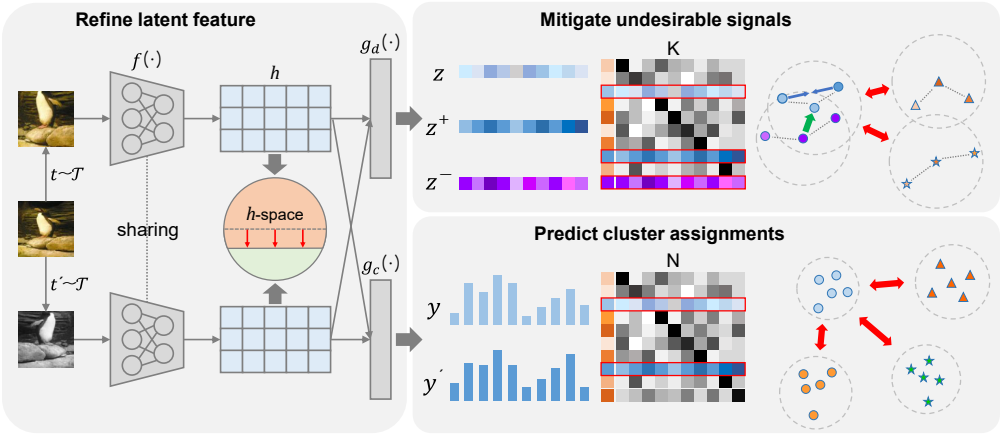

Figure 5.2: The proposed method consists of three parts, base encoder $f(\cdot)$, projection head $g_d(\cdot)$, and clustering head $g_c(\cdot)$. On the multiple heads, we perform instance-level discrimination and cluster-level discrimination. Further, to refine the hidden representation, which is the output of the base encoder, we applied the contrastive loss on the output of the encoder h.

5.2 Methods

In this section, we provide a method to learn a clustering-favorable feature space and focus on the following questions: (1) *How can the undesirable learning signal be alleviated from false negatives?* and (2) *How can the latent feature space be refined to help the network maintain the desirable learning signal?* We start with an overview of the deep clustering approach and describe the components of our method as illustrated in Fig. 5.1.

5.2.1 Deep Clustering

Deep clustering separates N images $\mathcal{I} = \{I_i\}_{i=1}^{N}$ into k clusters $\mathcal{Y} = \{Y_j\}_{j=1}^{k}$ following semantic characteristics in an unsupervised manner. Two major components are the (1) base encoder $f(\cdot)$, which produces a latent feature $h = f(I)$, where $h \in \mathbb{R}^d$ for a given input I, and (2) classifier $g_c(\cdot)$, which maps each latent feature h into class

assignments $y = g_c(h)$, where $y \in \mathbb{R}^k$. After training, each cluster is assigned a maximum likelihood as $y^* = \arg\max_y(y_j)$, where $j \in \{1, 2, ..., k\}$.

5.2.2 Contrastive Learning

The goal of contrastive learning is to reduce the distance between positive samples while enlarging the distances between negative samples by solving pretext tasks [28, 69]. We suspend a projection head $g_d(\cdot)$ after the base encoder to acquire a view-invariant representation. The projection head $g_d(\cdot)$ projects an output of the base encoder, called a base feature h, onto $z \in \mathbb{R}^l$. For each image, two stochastic transformations t and t' sampled from the augmentation family $\mathcal{T}$ produce a pair of correlated images (x, x'), and we perform this action for N images. Among $2N$ augmented images, there exists a semantically similar pair called positive (x, x^+) and dissimilar pairs called negative (x, x^-). Afterward, the network is trained to maximize the similarities between positive pairs and minimize similarities between an anchor and negative pairs as follows:

$$\mathcal{L}_{cont} = -\log s(z, z^+) + \log[s(z, z^+) + \sum_{i=1}^{N} s(z, z_i^-)], \tag{5.1}$$

where $s(a, b) = e^{a^\top b}/\tau$, and $\|a\|_2 = \|b\|_2 = 1$.

5.2.3 Mitigate Undesirable Learning Signal

Due to the sampling bias problem noted by Chuang *et al.* [30], solving naïve contrastive prediction tasks in Eq. 5.1 inevitably encounter the class collision problem because it performs instance-level discrimination without considering class information. A class collision problem is a phenomenon in which instances regarded as negative samples contain the same or similar semantics, thereby degrading the quality of the presentation. Here, the harmful learning signal arising from false negatives may interfere with building a clustering-favorable feature space. Since it is impossible to exclude the false negatives, we propose a solution to handle false negatives to improve

the clustering performance effectively. Therefore, we employ a learning objective to purify the corrupted learning signal from the aforementioned problem based on debiased contrastive learning [30]. We expect that applying a debiased contrastive prediction task will help to move nearer to the positives in the feature space as follows:

$$\mathcal{L}_{deb} = -\log \frac{s(z, z^+)}{s(z, z^+) + \frac{Q}{\tau^-}\Big[\mathbb{E}_{p^-}\big[s(z, z^-)\big] - \tau^+ \mathbb{E}_{p^+}\big[s(z, z_v)\big]\Big]}, \tag{5.2}$$

where Q is a weighting parameter, and τ^+ is the class probability.

Negative samples are sampled from the data distribution p rather than the true negative distribution p^- because the ground truth class information is unavailable. Assuming that the latent class c follows a uniform distribution $p(c) = \tau^+$, in the denominator of Eq. 5.2, the sum of true negative pair logits is approximated by subtracting the expectation of auxiliary positive pair logits from the expectation of negative pair logits is rather than naive summation over all negatives, as in Eq. 5.1. However, Eq. 5.2 is satisfied as the number of negative samples N tends to infinity. Thus, in practice, the expectation terms in Eq. 5.2 are replaced by the empirical estimation of expectation within the mini-batch. Additionally, to consider the hard negatives located near the positives, we modify Eq. 5.2 to $\mathcal{L}_{hard}$ based on the hard negative sampling strategy [157] and reorganize the learning objective as:

$$\mathcal{L}_{hard} = -\log \frac{s(z, z^+)}{s(z, z^+) + \frac{Q}{\tau^-}\Big[\mathbb{E}_{z^-}\big[s(z, z^-)\big] - \tau^+ \mathbb{E}_{z_v}\big[s(z, z_v)\big]\Big]}. \tag{5.3}$$

Here, we can approximate two expectation terms, $\mathbb{E}_{z^-}$ and $\mathbb{E}_{z_v}$, in the denominator, using the Monte-Carlo importance sampling as follows:

$$\begin{aligned} \mathbb{E}_{z^- \sim q_\beta^-}\big[s(z, z^-)\big] &\approx \mathbb{E}_{x^- \sim p}\big[e^{(\beta+1)s(z,z^-)}/\hat{Z}_\beta\big] \\ \mathbb{E}_{z_v \sim q_\beta^+}\big[s(z, z_v)\big] &\approx \mathbb{E}_{z_v \sim p^+}\big[e^{(\beta+1)s(z,z_v)}/\hat{Z}_\beta^+\big] \end{aligned} \tag{5.4}$$

where $\hat{Z}_\beta$ and $\hat{Z}_\beta^+$ are the partition functions of q_β and q_β^+, respectively. For more detailed information, please refer to the original papers [30, 157].

5.2.4 Refining Latent Features

As mentioned by Chen *et al.* [28], the projection head leaves only a part of the information on z. By optimizing the contrastive loss with restricted information, z becomes view-invariant; however, it could lose the task-favorable information as a result. Furthermore, Minderer *et al.* [135] noted z can be easily biased towards containing undesirable information, called a *shortcut problem*, which largely deteriorates the representation quality important for downstream tasks.

We propose a feature contrast regularizer to include clustering-relevant information in z while preventing the unwanted situation described above. Our intuition is that organizing learning signals for the base encoder not only from a restricted feature space but also from a feature space with more information can avoid z convergence to a trivial solution in which the trained network yields exactly the same z as different input images. Straightforwardly, we choose the feature space, mapped by the layer just before the projection head, in which the information remains intact. Therefore, we minimize the contrastive loss [28] on the intact features, h as follows:

$$\mathcal{R} = -\log s(h, h^+) + \log[s(h, h^+) + \sum_{i=1}^{N} s(h, h_i^-)]. \tag{5.5}$$

In this manner, the final layers contain more task-favorable information by jointly optimizing the spaces of z and h.

5.2.5 How to Estimate Cluster Assignments?

We attach another projection head $g_c(\cdot)$ to directly estimates the cluster assignments distribution similar to [110] in addition to the projection head $g_d(\cdot)$, which learns view-invariant information from data instances. The projection head $g_c(\cdot)$ maps a latent feature h to $y \in \mathbb{R}^k$ with a Softmax function. Similar to in Eq. 5.1, we define contrastive loss [28] to distinguish the estimated cluster assignments as follows:

$$\mathcal{L}_c = -\log s(y, y^+) + \log[s(y, y^+) + \sum_{i=1}^{N} s(y, y_i^-)]. \tag{5.6}$$

We also utilize the negative entropy to make cluster assignments closer to a uniform distribution, and it finally prevents the trivial solution that most data points are allocated to the same cluster. The overall loss for the cluster assignments prediction can be written as $\mathcal{L}_{clus} = \mathcal{L}_c - \mathcal{H}(Y)$. Here, $\mathcal{H}(Y)$ is the entropy of cluster assignments probabilities which defined by $-\sum_{i=1}^{2K} p(y_k) \log p(y_k)$.

5.2.6 Objective Function

We jointly optimize the three objectives from two projection heads and one objective from the base encoder in an end-to-end manner. The overall learning objective function is formulated as:

$$\mathcal{L} = \mathbb{E}_{\substack{(x,x^+)\sim p^+ \\ x_i^- \sim p^-}} \big[\mathcal{L}_{hard} + \mathcal{L}_{cont} + \mathcal{L}_{clus} + \lambda \mathcal{R}\big]. \tag{5.7}$$

5.3 Experimental Results

5.3.1 Experimental Setup

We conducted experiments on five benchmark datasets. (1) CIFAR-10/100 [97]: A natural image with $50,000/10,000$ samples from 10/100 categories for training and testing, respectively. In CIFAR-100, we regarded the 20 super-classes as the target classes to reduce semantic granularity. (2) ImageNet-10 and ImageNet-Dogs: As a subset of ImageNet [98], the former with 10 randomly sampled subjects and the latter with 15 dog breeds. (3) Tiny-ImageNet: A subset of ImageNet [98] with 200 categories. It consists of $100,000/10,000$ images for training/testing, respectively.

We used three standard evaluation metrics: (a) normalized mutual information (NMI), (b) accuracy (ACC), and (c) adjusted rand index (ARI). All metrics ranged from 0 to 1, and the higher value indicates better clustering performance.

5.3.2 Implementation Details

For a fair comparison, we followed the same data protocol as those in the previous studies [78, 84, 110]. All experiments were performed using ResNet-34 [70] to yield a hidden representation of an image, and we set the dimension of the projection head $g_d(\cdot)$ as 128. We resized all images to 224×224 to fit the input resolution of the backbone network and randomly applied Gaussian blur, horizontal flip, color jittering, and conversion to grayscale. We excluded the Gaussian blur on CIFAR-10/100 since it may prevent the network from accessing the correct semantic information. We used the Adam optimizer [92] with an initial learning rate of $3e-4$ to optimize the backbone network and two projection heads. No weight decay or scheduler was not used in the experiments, and all models were trained for $1,000$ epochs.

We utilize the deep learning library PyTorch 1.7.0 [148] to implement the proposed model. We conduct all experiments with 2 $\times$ Intel(R) Xeon(R) Gold 6258R @ 2.70GHz, 512GB RAM, 8 $\times$ NVIDIA Quadro RTX 8000 GPU with the batch size 256.

5.3.3 Main Results

Comparison to the State-of-the-Art We first compared the clustering performance of our method with state-of-the-art algorithms in Table 5.1 and observed the following characteristics. (1) Our method outperformed the previous state-of-the-art methods on five datasets. In particular, we surpassed the best competitor [110] by effectively considering false negatives and utilizing hard negative mining for deep clustering. (2) In terms of accuracy, we observed the proposed model showed greater performance improvements when the number of classes we wanted to cluster was large. We achieved the accuracy gain on CIFAR-100 of 1.9% and ImageNet-Dogs of 8.1%. We achieved a 2.5% performance improvement on the Tiny-ImageNet, which is the most challenging dataset among the five. These results demonstrate that our approach to creating a clustering-favorable feature space appropriately applies to representation learning.

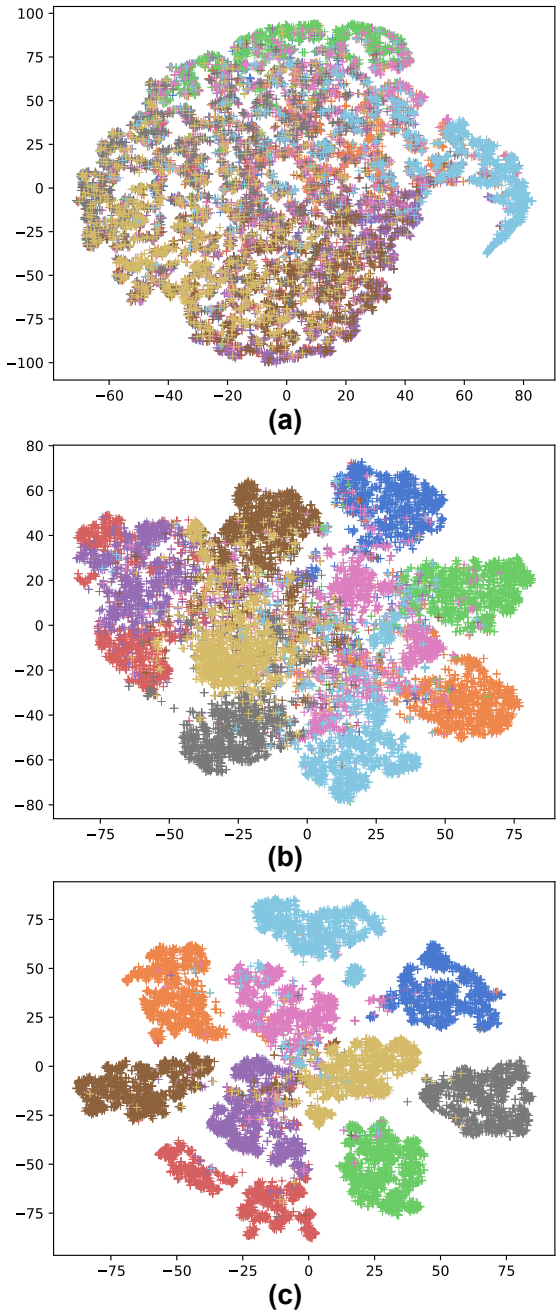

Figure 5.3: Cluster evolution on ImageNet-10. (a)-(c) represent the visualization of feature space at the early, middle, and final stages of learning.

Table 5.1: Comparison with existing methods. We evaluate our method on five challenging benchmark datasets.

Dataset	CIFAR-10			CIFAR-100			ImageNet-10			ImageNet-Dogs			Tiny-ImageNet		
Metrics	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI
K-means	0.087	0.229	0.049	0.084	0.130	0.028	0.119	0.241	0.057	0.055	0.105	0.020	0.065	0.025	0.005
SC [225]	0.103	0.247	0.085	0.090	0.136	0.022	0.151	0.274	0.076	0.038	0.111	0.013	0.063	0.022	0.004
AC [59]	0.105	0.228	0.065	0.098	0.138	0.034	0.138	0.242	0.067	0.037	0.139	0.021	0.069	0.027	0.005
NMF [18]	0.081	0.190	0.034	0.079	0.118	0.026	0.132	0.230	0.065	0.044	0.118	0.016	0.072	0.029	0.005
AE [9]	0.239	0.314	0.169	0.100	0.165	0.048	0.210	0.317	0.152	0.104	0.185	0.073	0.131	0.041	0.007
DAE [187]	0.251	0.297	0.163	0.111	0.151	0.046	0.206	0.304	0.138	0.104	0.190	0.078	0.127	0.039	0.007
DCGAN [153]	0.265	0.315	0.176	0.120	0.151	0.045	0.225	0.346	0.157	0.121	0.174	0.078	0.135	0.041	0.007
DeCNN [224]	0.240	0.282	0.174	0.092	0.133	0.038	0.186	0.313	0.142	0.098	0.175	0.073	0.111	0.035	0.006
VAE [93]	0.245	0.291	0.167	0.108	0.152	0.040	0.193	0.334	0.168	0.107	0.179	0.079	0.113	0.036	0.006
JULE [214]	0.192	0.272	0.138	0.103	0.137	0.033	0.175	0.300	0.138	0.054	0.138	0.028	0.102	0.033	0.006
DEC [207]	0.257	0.301	0.161	0.136	0.185	0.050	0.282	0.381	0.203	0.122	0.195	0.079	0.115	0.037	0.007
DAC [24]	0.396	0.522	0.306	0.185	0.238	0.088	0.394	0.527	0.302	0.219	0.275	0.111	0.190	0.066	0.017
ADC [63]	–	0.325	–	–	0.160	–	–	–	–	–	–	–	–	–	–
DDC [23]	0.424	0.524	0.329	–	–	–	0.433	0.577	0.345	–	–	–	–	–	–
DCCM [200]	0.496	0.623	0.408	0.285	0.327	0.173	0.608	0.710	0.555	0.321	0.383	0.182	0.224	0.108	0.038
IIC [84]	–	0.617	–	–	0.257	–	–	–	–	–	–	–	–	–	–
GATC [140]	0.475	0.610	0.402	0.215	0.281	0.116	0.609	0.762	0.572	0.322	0.333	0.200	-	-	-
PICA [78]	0.591	0.696	0.512	0.310	0.337	0.171	0.802	0.870	0.761	0.352	0.352	0.201	0.277	0.098	0.040
DRC [229]	0.612	0.716	0.547	0.356	0.367	0.208	0.830	0.884	0.798	0.384	0.389	0.233	0.321	0.139	0.056
CC [110]	0.705	0.790	0.637	0.431	0.429	0.266	**0.859**	0.893	0.822	0.445	0.429	0.274	0.340	0.140	0.071
Ours	**0.715**	**0.807**	**0.658**	**0.444**	**0.448**	**0.288**	0.857	**0.899**	**0.824**	**0.497**	**0.510**	**0.356**	**0.363**	**0.165**	**0.086**

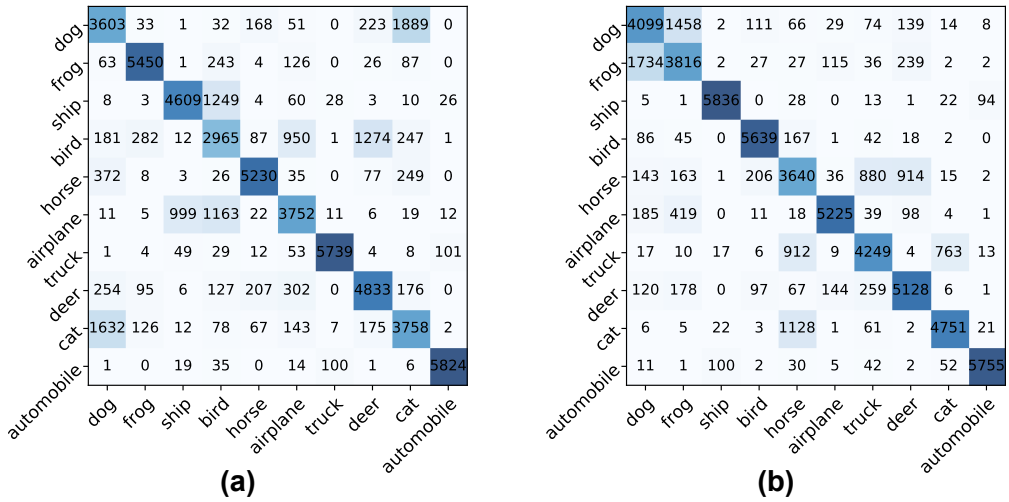

Figure 5.4: Confusion matrices of (a) baseline and (b) our results. Each row of the matrix is the predicted cluster, while each column is the ground-truth.

Cluster Evolution We designed our learning objectives under the motivation of not only reducing undesirable signals in z but also refining the information contained in the latent feature h. In Fig. 5.3, we visualized trained latent features with t-SNE [183] to observe how data instances are progressively grouped. All samples were distributed sporadically during the early stages of training and formed a clustering-favorable distribution, except for some overlapped inter-class distribution.

Confusion Matrix Analysis As shown in Fig. 5.4, we compared two confusion matrices of the baseline [110] and proposed method. The top-right area of the confusion matrix indicates the false negatives that interfere with providing desirable learning signals to the projection heads. From this result, our deep clustering approach helps create a clustering-favorable feature space by filtering out a large number of false negatives. Specifically, we effectively reduced the number of false negatives from 7524 to 5526.

Backbone Reliance As shown in Table 5.2, we evaluated the clustering performance while changing the backbone network that constructs the feature space. We discovered

Table 5.2: Reliance of backbone networks. We evaluate the performance on various base encoders.

Dataset	Backbone	NMI	ACC	ARI
CIFAR-10	ResNet-18	0.686	0.777	0.618
	ResNet-34	**0.718**	**0.802**	**0.653**
	ResNet-50	0.687	0.778	0.620
ImageNet-10	ResNet-18	0.852	0.895	0.819
	ResNet-34	0.858	**0.898**	0.826
	ResNet-50	**0.859**	**0.898**	**0.827**

Table 5.3: Transferability of learned features

Train \ Test	ImageNet-10	CIFAR-10	STL-10
ImageNet-10	-	0.340	0.478
CIFAR-10	0.296	-	0.464
STL-10	0.618	0.488	-

the following tendencies: (1) When we experimented on a relatively small dataset, such as ImageNet-10, we acquired a similar performance in terms of accuracy. However, we observed that NMI and ARI increased as the network became deeper. These results indicate that we can build a more robust feature space as the network depth increases. Based on the rich expressiveness of the deep network, we formed a more robust feature space. This means that deep networks generate more expressive features in terms of the security of cluster results. (2) Meanwhile, we confirmed that deep networks do not guarantee high performance for relatively large datasets.

Figure 5.5: Cases studies on ImageNet-10. Each row represents ”soccer ball”, ”trailer truck”, and ”orange”, respectively. (a) green, (b) red, and (c) blue boxes indicate the True Positive, False Positive, and False Positive examples, respectively.

Transferability As shown in Table 5.3, to evaluate the generality of our method, we assessed the cross-data transferability among ImageNet-10, CIFAR-10, and STL-10 with accuracy scores. Because there was no additional training on the target datasets, we demonstrate that Table 5.3 shows meaningful transfer performance between datasets. We observed that the transfer results on STL-10 are outstanding, and we conjecture that it would be more advantageous for a model to capture general semantics due to the relatively large dataset size.

Case Studies We examined the success and failure examples to deliver insights into our method, as shown in Fig. 5.5. We consider three types of examples: true positives, false negatives, and false positives. False negative samples were frequently generated when images with different labels contained two different semantics, such as “dog” and “ball”. While false positive samples arose for images with similar color and structure priors, such as “ball” and “orange”, in false positive samples, we speculate that color and structure information is a major factor in mis-clustering. For example, an orange-colored soccer ball with a “soccer ball” class was mis-clustered in the “orange” class. Based on the above-mentioned problems, we believe unsupervised clustering must

Table 5.4: Ablation results on CIFAR-10. † indicates the reproduced results [110].

Method	MNI	ACC	ARI
baseline	0.705	0.790	0.637
†baseline	0.680	0.765	0.607
†baseline + $\mathcal{R}$	0.672	0.763	0.600
$\mathcal{L}_{clus}$	0.553	0.584	0.411
$\mathcal{L}_{clus}$ + $\mathcal{L}_{deb}$	0.695	0.784	0.628
$\mathcal{L}_{clus}$ + $\mathcal{L}_{hard}$	0.668	0.726	0.573
$\mathcal{L}_{clus}$ + $\mathcal{R}$	0.609	0.675	0.508
$\mathcal{L}_{clus}$ + $\mathcal{L}_{deb}$ + $\mathcal{R}$	0.693	0.784	0.625
$\mathcal{L}_{clus}$ + $\mathcal{L}_{hard}$ + $\mathcal{R}$	0.701	0.788	0.634
$\mathcal{L}_{clus}$ + $\mathcal{L}_{hard}$ + $\mathcal{L}_{cont}$	**0.718**	0.802	0.653
$\mathcal{L}_{clus}$ + $\mathcal{L}_{hard}$ + $\mathcal{L}_{cont}$ + $\mathcal{R}$	0.715	**0.807**	**0.658**

learn low-level features while exploiting high-level features.

5.3.4 Ablation Study

In Table 5.4, we report the ablation results on CIFAR-10 to show the efficacy of each component of the proposed method. We assessed the three evaluation metrics by adding each component in sequence. Here, we observed three appealing results. First, when replacing $\mathcal{L}_{cont}$ with $\mathcal{L}_{deb}$ or $\mathcal{L}_{hard}$, the clustering performance was slightly reduced. This pattern shows different characteristics compared to previous studies [30, 157], which performed linear classification as a downstream task. We consider that this result is due to some loss of downstream task-relevant information through the projection head. Second, when we complemented $\mathcal{R}$ as a learning objective, we achieved improved clustering performance compared to the reproduced baseline. We explain this phenomenon as helping to boost the clustering-favorable information in the h-

space suppressing the improper learning signal for downstream tasks. Further, joint training of the $\mathcal{L}_{clus}$ and $\mathcal{R}$ enhanced the clustering accuracy from 0.584 to 0.675. Finally, we achieved the best performance when we trained the network with four learning objectives. We interpret this result as an undesirable learning signal triggered by false negatives damaging the representation quality, even though the regularizer aims to construct a better h-space. From the ablation results, we confirmed that our method to correct learned information in the base feature h jointly and that in the projected feature z is effective for learning a clustering-favorable representation.

5.4 Discussion

5.4.1 Representation Quality

We perform a series of experiments to evaluate the representation quality trained with our method besides deep clustering. Table 5.5 shows the Top-1 accuracy under linear evaluation, the most widely used evaluation protocol for self-supervised learning. To measure the Top-1 accuracy, we follow to employ three stages: 1) learn the representation of $f(\cdot)$ on the training dataset under multiple conditions (SimCLR [28], w/o FCR, and w/ FCR). 2) train a linear classifier $g(\cdot)$ on top of $f(\cdot)$ on the training dataset by freezing the weights of the encoder. 3) evaluate the accuracy of the classifier on the testing dataset. Here, we trained all the models with 200 epochs with 32×32 resized images. In four different datasets, we demonstrated that the proposed method is effective to enhance the discriminative properties.

In Table 5.6, we employed the non-parametric classifier based on k-Nearest Neighbor (k-NN) on the representation trained in a self-supervised manner. We applied the feature contrast regularizer to features in other layers before the projection head to observe which intermediate feature helps to propagate the task-relevant properties. We selected five-layer features from conv1 to layer4 in ResNet [70] and appended a linear layer to extract the same dimensional features with the output of the base encoder. Each

Table 5.5: Linear evaluation for a model trained with different methods. We report Top-1 classification accuracy (%).

Method	CIFAR-10	CIFAR-100	ImageNet-10	ImageNet-Dogs
SimCLR [28]	79.94	52.90	79.80	53.07
SimCLR [28] w/o $\mathcal{R}$	82.39	54.47	82.60	53.73
SimCLR [28] **w/** $\mathcal{R}$	**82.78**	**54.84**	**84.60**	**53.87**

Table 5.6: k-NN classification accuracy. We applied a feature contrast regularizer on the features from conv1 to layer4 in ResNet [70]. The baseline indicates the learning without feature contrast regularizer.

Layer name	conv1	layer1	layer2	layer3	layer4	baseline	proposed
Top-1	37.71	59.83	67.66	71.64	72.35	68.4	**73.49**
Top-5	42.95	65.63	72.45	75.53	76.58	73.77	**77.41**

linear layer is jointly trained with the proposed method following the first stage mentioned in the previous paragraph. From the experimental results, using features before the projection head shows the best performance in the k-NN classifier. It is well known that the convolutional neural network learns the hierarchical nature of features [223]. Based on this attribute, it can be seen that applying the feature contrast regularizer to features, including high-level semantics, helps to learn discriminative features. From two additional experiments, we have shown that our method helps to propagate task-relevant information to the end of the network. Though it is challenging the direct estimation of information [8], our method shares the same spirit as some studies that extract task-relevant information [176, 193] for the downstream task.

5.4.2 Behavior of Representation

Due to the difficulty of ensuring a clustering-favorable space, we indirectly proved the degree of clustering favorability with two properties: alignment and uniformity [190]. Alignment indicates the average distance between intra-class pairs, while uniformity represents the degree to which all samples are distributed, preserving maximal information. We visualized two properties of the model trained with ImageNet-10, as shown in Fig. 5.6. From the qualitative results, our method is effective for both representation learning and deep clustering. It can be seen that the representation learned by our method is distributed widely on the unit hypersphere and is well-located in a local area for each class. To provide more information on our results, we visualize uniformity for all classes of CIFAR-10 in Fig. 5.7.

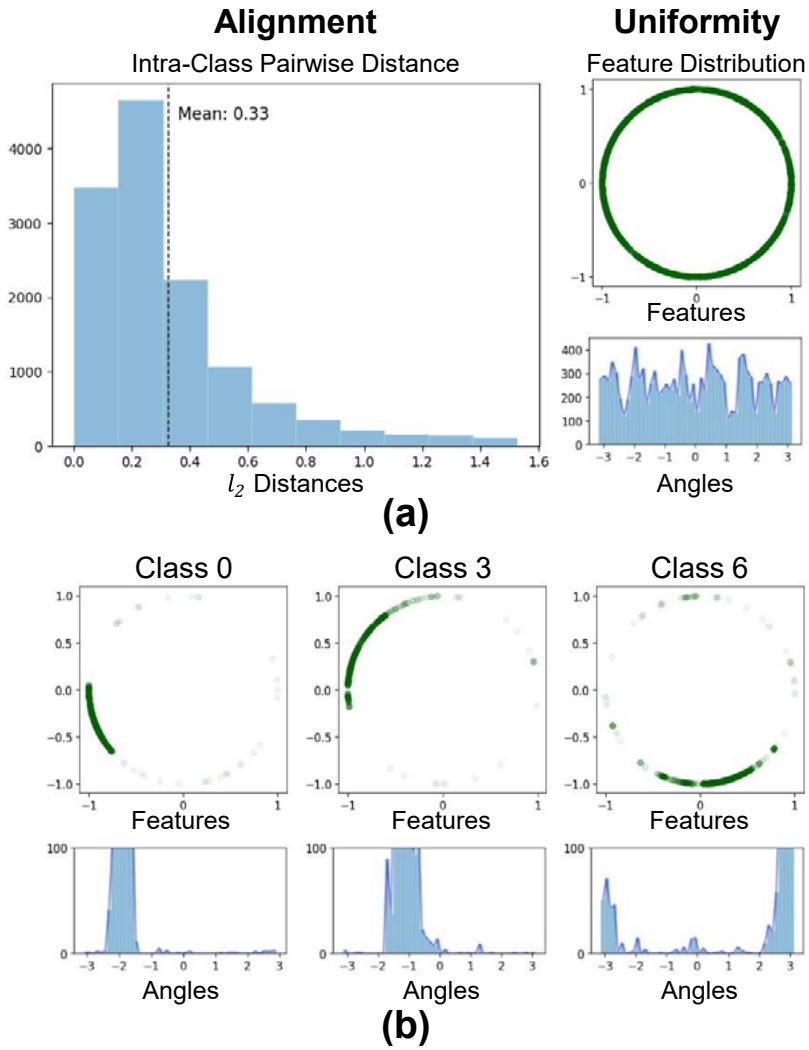

Figure 5.6: Alignment and uniformity analysis. We plot feature distributions with the dimensionality reduction with t-SNE.

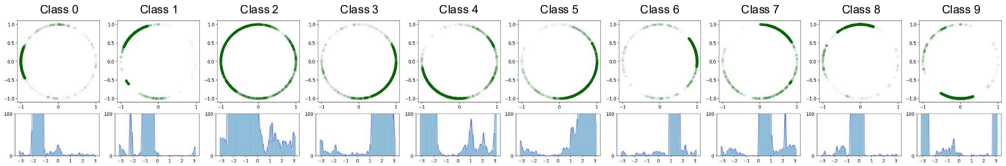

Figure 5.7: Uniformity analysis on CIFAR-10. We plot feature distributions with the dimensionality reduction with t-SNE.

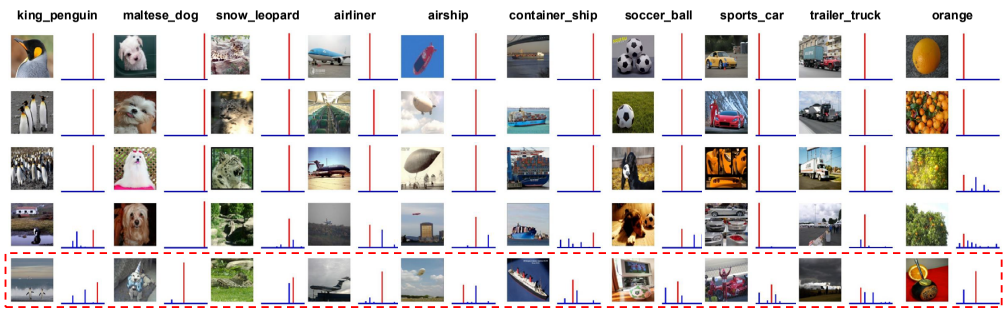

Figure 5.8: Examples of cluster assignment confidence. For ImageNet-10, we represent the correct labels on the upward side and draw the confidence distribution on the right side of images. The bottom line shows some failure cases.

5.4.3 Clustering Results

We report the example of clustering results with the cluster assignment confidence in Fig. 5.8. We observe the two trends from the clustering results. 1) When a small object appears on an image, the network tends to be less confident about the target cluster. 2) When the other objects exist on an image, the network will likely assign the image to other clusters.

5.4.4 Implicit Feature Decorrelation

Drawing the correlation matrix for feature z extracted from the projection head, we discovered that the proposed method implicitly performs feature decorrelation, as illustrated in Fig. 5.9. It is known that unexpected correlation between features may degenerate the performance of clustering [192] and some methods [72, 101] adopt the explicit idea of spectral clustering [225] as their learning objectives to decorrelate their features. However, the proposed method implicitly obtains the effect of spectral clustering by directly contrasting the feature space. This feature decorrelation helps to build a clustering-favorable space by removing the dependency among features.

5.5 Summary

In this chapter, we present a novel deep clustering method to correct the bias arising from the unsupervised experimental setting and propose the feature contrast regularizer which learns the clustering-favorable representation. We have conducted extensive experiments on the five challenging datasets and outperformed state-of-the-art joint deep clustering approaches. With extensive ablation studies, we analyzed the effect of each learning objective. We implicitly demonstrated that our method generates a more clustering-favorable representation feature extractor having more discriminative properties by providing the results on the linear evaluation and k-NN classifier performance.

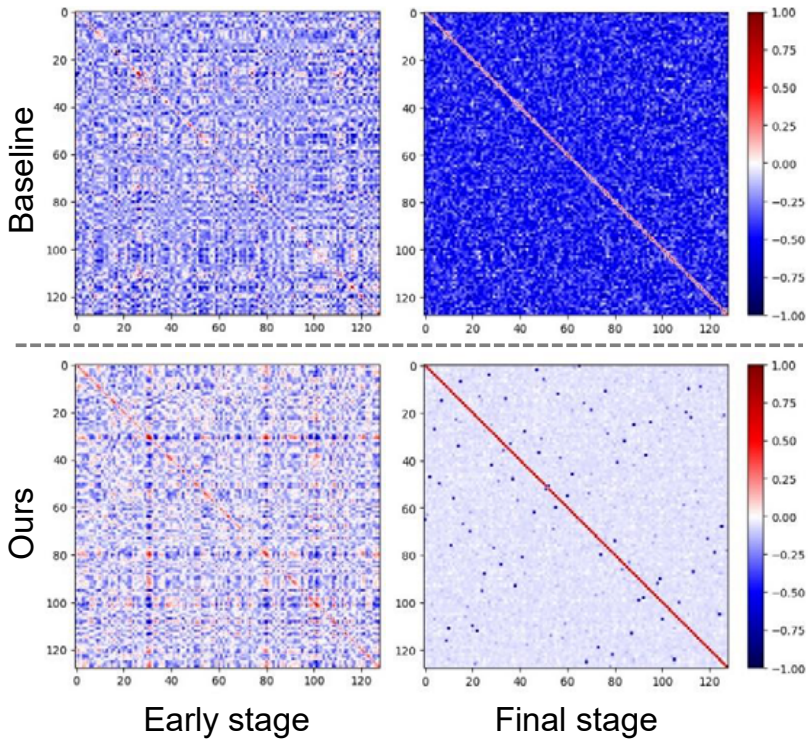

Figure 5.9: Feature correlation matrix on CIFAR-10

Chapter 6

Learning from Brain-inspired Approach

In this chapter, we explore the potential applicability of a brain-inspired algorithm on imperfect data recognition. Though ANN-based learning has achieved successful performance in various areas of machine learning, the underlying learning algorithm is criticized for its biological implausibility. We experimentally demonstrate that a brain-inspired algorithm is effective for imperfect data recognition based on previous studies, showing noise-resistant learning under imperfect supervision.

6.1 Introduction

The human brain has an intricate and heterogeneous structure that consists of a high recurrent and nonlinear neural network [12, 48, 53]. It is commonly understood that the learning system of the human brain operates on the synaptic plasticity mechanism [71], wherein the modulation in synaptic weights varies according to the intrinsic or extrinsic stimuli [151]. Specifically, neural plasticity regulates the process of synaptic transmission as a fundamental property of neurons [31, 129]. Based on this property, the neuronal responses to sensory stimuli enable the robust recognition [42, 55, 142, 194] and noise-resistance learning [149, 175] in human perception.

Based on the human brain architecture, artificial neural networks were suggested

to simulate the pattern of the human decision-making process for recognition tasks. Rumelhart *et al.* [159] introduced the backpropagation algorithm that adjusts the network parameters to achieve reliable performance. Backpropagation iteratively updates the network parameters relying on the error signal generated at the network's end between the produced output and the desired output. In the last decade, with the benefits of backpropagation [159], ANNs have exceeded human-level performance on various machine learning applications such as classification, segmentation, and detection [45, 70]. However, learning ANNs with backpropagation have been criticized for their biological implausibility, wherein its behavior conflicts with the activity of real neurons in the human brain [3, 81]. First, the human brain operates according to *neural plasticity*, which indicates the capability for modifying neural circuit connectivity or degree of interaction [139]. Second, global error-guided learning requires the forward weight matrices to propagate the error signal flow to the lower layer, that is *weight transport problem* [61]. Multiple learning algorithms have been proposed to alleviate the previously mentioned obstacles based on strong constraints of backpropagation and enhance the biological properties [41, 112, 113, 197, 198]. This chapter explored the predictive coding network [197] among the various biologically plausible learning and its characteristics.

A predictive coding network [197] was introduced to resolve the biological limitations of backpropagation depending on the hierarchically organized visual cortex of the human brain [53, 155]. With respect to biological plausibility, a predictive coding network concentrates on local and Hebbian plasticity by minimizing the prediction errors between expected and actual inputs [133, 155]. The learning mechanism of the predictive coding network is different from that of backpropagation, which updates the network parameters using only one error derived from the last layer [159]. Predictive coding is regarded as a local learning algorithm because its learning is performed with local error nodes and global error nodes. A learning network with predictive coding approximates the learning dynamics of backpropagation [197] and can also be ex-

panded to arbitrary computational graphs [133]. Multiple works [29, 64, 196] inspired by the property of prediction itself have been proposed, and some studies [29, 162] demonstrated that the potential of the predictive manner related to human perception.

However, despite the remarkable accomplishment of ANN architectures and their learning algorithms, there remains a performance gap between machine and human intelligence in some applications. We collectively refer to these tasks as *machine-challenging tasks* (MCTs); MCTs are difficult for machine intelligence while easy for human intelligence. This study considers the representative MCTs as incremental learning, long-tailed recognition, and few-shot learning inspired by [67]. Humans progressively and ceaselessly acquire new knowledge and preserve it by virtue of the hippocampus [152]. The primary function of the hippocampus is that it enables long-term memory of the spatial and sequential order from the human experience [13, 39]. This property makes the human intelligence exhibits robust and performs better than machine intelligence [58, 120, 232]. Meanwhile, ANNs trained with backpropagation tend to forget what they learned when it learns new information, that is *catastrophic forgetting* [50, 57, 130]. As another example, machine intelligence shows unsatisfactory performance under limited or imperfect training data recognition [40, 121]. When training ANNs for classification tasks in a long-tail scenario, the classifier can be easily biased toward the majority classes that contain the most data and show poor performance in minority classes [85]. These phenomena result from the fundamental differences in visual processing between the brain and ANNs [210]. Inspired by Hassabis *et al.* [67], we hypothesized that the closer the learning algorithm is to the human brain, the more effective it is for the MCTs.

Similar to our assumption on the MCTs, the learning algorithms inspired by the brain are consistently studied to reduce the performance gap between machine intelligence and human intelligence based on human's various attributes. In terms of human learning mechanisms, a spiking neural network (SNN) is considered a promising solution to replicate the neural processing process of the brain. Yang *et al.* [217] pro-

posed an SNN-based continual meta-learning framework and demonstrated that the suggested model improves the accuracy and robustness of the continual meta-learning tasks. Yang *et al.* [216] also established the ensemble framework with multiple spike-driven few-shot online learning and confirmed the effectiveness of the brain-inspired paradigm. On the other hand, recent studies reported that the neural network trained biologically plausible manner embodies specific memory functions in the human memory system. Salvatori *et al.* [162] discovered that the network trained with predictive coding can naturally implement the associative memory function, such as reconstructing incomplete regions. Yang *et al.* [215] verified that the multicompartmental spiking neural network incorporates the working memory satisfying four essential components of brain-inspired mechanisms. Therefore, based on previous studies, we speculated that predictive coding has other latent properties. This study aimed to discover hidden properties and extend the scope of predictive coding to MCTs. Contrary to the conventional solutions for the MCTs, this chapter focused on the predictive coding algorithm itself employed for the optimization of the network parameters. In incremental learning, it is confirmed that predictive coding better reveals the plasticity-stability property and enables faster adaptation to new tasks than backpropagation. In long-tailed recognition, it reduces the classification bias problem of minority classes.

The contributions of this chapter can be summarized as follows:

- The study characterized the MCTs, which are easy for human intelligence and difficult for machine intelligence, in machine learning fields and proposed a hypothesis that the brain-inspired learning algorithm improves the performance of MCTs.

- Predictive coding, a biologically plausible learning algorithm, was adopted for MCTs, such as incremental learning and limited data recognition. In addition, extensive experiments were performed by reimplementing the learning with backpropagation with predictive coding.

Table 6.1: Classification accuracy (%) on the Moon dataset. We denoted the learning with backpropagation as BP and learning with the predictive coding framework as PC. σ indicates the added Gaussian noise to the data.

σ	10 labels		20 labels		30 labels	
	BP	PC	BP	PC	BP	PC
0.1	86.60	87.26 (+0.66)	89.04	90.94 (+1.90)	94.28	94.64 (+0.36)
0.2	86.40	86.56 (+0.16)	87.68	89.84 (+2.16)	93.10	93.22 (+0.12)
0.3	84.12	86.76 (+2.64)	86.06	87.66 (+1.60)	89.32	89.60 (+0.28)

- The effect of learning algorithms close to brain learning on MCTs in terms of neuroscience was presented. Mainly, the experimental results were analyzed with respect to the plasticity-stability dilemma and interplay between the hippocampus and prefrontal cortex.

6.2 Exploration Study

We performed the SSL training covered in Chapter 3 with predictive coding to explore our assumptions on imperfect data recognition. We applied the predictive coding-based SSL for the Moon dataset, a toy dataset for machine learning, such as classification and clustering, as visualized in Fig. 6.1. We generated the $10,000$ training samples with 0.1 Gaussian noise to construct data for the experiments. We sampled $1,000$ data for the test dataset following the same training set distribution. We designed the multilayer perceptron with two hidden layers, each with 100 hidden neurons. We measured the performance of pseudo-labeling [105] by increasing the number of labels data. The classification results are presented in Table 6.2. We performed experiments on five different seeds and reported the average classification accuracy. From the experimental results, we observed that predictive coding-based learning achieves better classification accuracy than that backpropagation. The efficiency of predictive coding

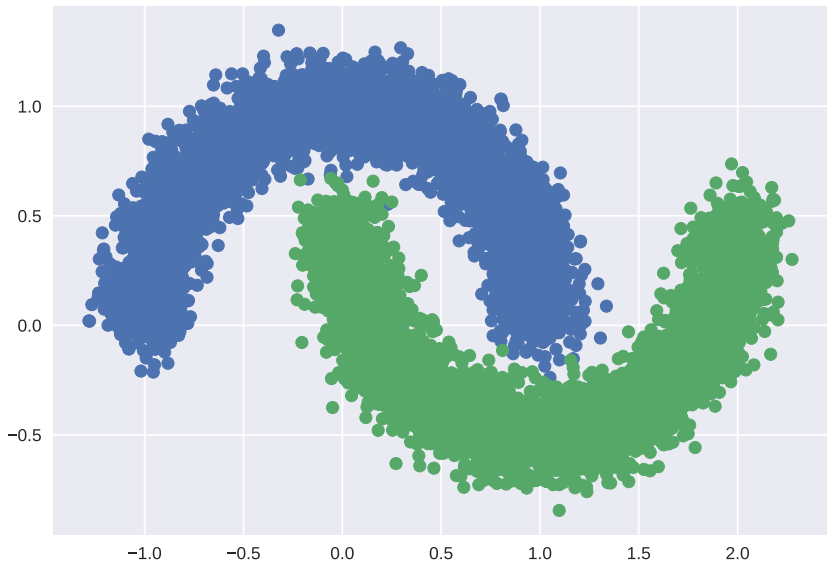

Figure 6.1: Moon data visualization

is especially noticeable as the number of labeled data reduces from 30 to 10. These experimental results correspond to what we have assumed, and there is potential room to improve the performance of the other MCTs.

6.3 Incremental Learning with Predictive Coding

Based on previous studies [67, 149], our fundamental assumption is that the more biologically plausible the learning algorithm, closely replicating the learning mechanism of the brain, the more effective it will be for MCTs. Previous studies focused on confirming that the predictive coding network itself inherits the physiological characteristics of the brain. Salvatori *et al.* [162] recently explored that predictive coding networks naturally implement associative memory, which plays a vital role in human intelligence [33]. Motivated by the previous study, the current research assumed that

Table 6.2: Classification accuracy (%) on the Moon dataset. We denoted the learning with backpropagation as BP and learning with the predictive coding framework as PC. σ indicates the added Gaussian noise to the data.

σ	10 labels		20 labels		30 labels	
	BP	PC	BP	PC	BP	PC
0.1	86.60	87.26 (+0.66)	89.04	90.94 (+1.90)	94.28	94.64 (+0.36)
0.2	86.40	86.56 (+0.16)	87.68	89.84 (+2.16)	93.10	93.22 (+0.12)
0.3	84.12	86.76 (+2.64)	86.06	87.66 (+1.60)	89.32	89.60 (+0.28)

predictive coding networks have a latent ability to consolidate the sequentially acquired knowledge in the human memory system. Therefore, we propose a predictive coding framework for incremental learning and verify the efficacy of MCTs. The task of incremental learning can be mainly categorized into two categories [128]: class-incremental learning and task-incremental learning. The current study focused on the former. In class-incremental learning, the knowledge from previously seen classes is no longer available when a network learns the knowledge of unseen classes, and the learned network aims to achieve favorable classification accuracy for all tasks without forgetting. Multiple tasks were sequentially learned based on the pre-defined order to validate our assumption, and each task with its validation set finishing the training of the given task was evaluated. The algorithms are detailed in Algorithm 1.

Algorithm 1 Predictive Coding for Incremental Learning

Input: Dataset $\mathcal{D}_{t=1}^{T}$, Computational Graph $\mathcal{G} = \{\mathcal{E}, \mathcal{V}\}$, inference learning rate η_v, weight learning rate η_θ

for all dataset for each task $\mathcal{D}_t \in \mathcal{D}$ **do** ▷ For each minibatch in the sequential tasks

  $\hat{v}_0 \leftarrow x_t$ ▷ Initialize the graph with inputs

  for all $\hat{v}_i \in \mathcal{V}$ **do** ▷ Forward phase: calculate predictions

    $\hat{v}_i \leftarrow f(\mathcal{P}(\hat{v}_i); \theta)$

  end for

  $\epsilon_L \leftarrow L - \hat{v}_L$ ▷ Compute output error

  while *not converged* **do** ▷ Backward phase: backward iteration

    for all $(v_i, \epsilon_i) \in \mathcal{G}$ **do**

    $\epsilon_i \leftarrow v_i - \hat{v}_i$ ▷ Compute prediction errors

    $v_i \leftarrow v_i + \eta_v \frac{d\mathcal{F}}{dv_i}$ ▷ Update the vertex values

    end for

  end while

end for

for all $\theta_i^t \in \mathcal{E}$ **do** ▷ Update weights at equilibrium

  $\theta_i^t \leftarrow \theta_i^{t+1} + \eta_\theta \frac{d\mathcal{F}}{d\theta_i}$

end for

  return θ^t

Table 6.3: Details of the tasks in the disjoint-MNIST and disjoint-FMNIST benchmarks

Task id	MNIST classes	FMNIST classes	Training	Testing
1	[0, 1, 2, 3, 4]	[T-shirt/top, Trouser, Pullover, Dress, Coat]	25k	5k
2	[5, 6, 7, 8, 9]	[Sandal, Shirt, Sneaker, Bag, Ankle boot]	25k	5k

6.3.1 Experimental Settings

A 3-layer predictive coding network with ReLU non-linearity, where the number of the hidden nodes was 800 for the simple dataset such as MNIST [103] and FMNIST [205], was employed. Similar to the study by [165], a simplified Alexnet architecture [98] consisting of three convolutional layers was used for the complex dataset such as CIFAR-10 [96]. The three convolutional layers comprised 64, 128, and 256 channels.

We refined the data to formulate sequential incremental tasks. The data were divided into multiple portions following the previous studies [108, 173]. Then, we constructed four datasets: 1) disjoint-MNIST, 2) disjoint-FMNIST, 3) split-MNIST, and 4) split-CIFAR-10. Disjoint-MNIST and disjoint-FMNIST were organized by separating MNIST and FMNIST into two tasks. In addition, a more complex dataset, called split-MNIST and split-CIFAR-10, was also established, where all classes were separated into five tasks, and each task contained two categories. The details of the tasks on the multiple datasets are described in Tables 6.3 and 6.4. Finally, we evaluated incremental learning performance. We trained a network with sequential order and measured that the acquired knowledge was maintained after each task's training, same as [165].

A learning rate of 0.05 was used, and the learning rate was divided by $1/3$ to perform incremental learning if there was no advancement in the validation loss for five consecutive epochs. In predictive coding, the weight learning rate was set as 0.1 while keeping the other hyperparameters. The minimum learning rate was set as $1e^{-4}$ and batch size as 64. All experiments were conducted using data split according to five different seeds. We provide the code to reproduce the results in the manuscript at `https:`

Table 6.4: Details of the tasks in the split-CIFAR-10 benchmark

Task id	CIFAR-10 classes	Category	Training	Testing
1	[airplane, car]	vehicle	10k	2k
2	[bird, cat]	animal	10k	2k
3	[deer, dog]	animal	10k	2k
4	[frog, horse]	animal	10k	2k
5	[ship, truck]	vehicle	10k	2k

`//github.com/jangho2001us/PredictiveCoding_IncrementalLearning`.

6.3.2 Experiments on Incremental Learning

Incremental learning was performed on disjoint-MNIST and disjoint-FMNIST using the predictive coding framework to validate our hypothesis. To implement the incremental learning task in a predictive coding manner, we integrated the code of [165] and [158] by replacing the network learning from the backpropagation with the predictive coding networks. The performance of each task was evaluated after completing the learning of each task in Tables 6.5 and 6.6. The performance in all tasks learned was evaluated using the best model of the last task. In this case, the best model refers to the model with the highest performance in the given task. Moreover, the other backpropagation-based incremental approaches containing SGD [57], SGD-F [57], EWC [94], IMM [108], LFL [87], and LWF [111] were evaluated to observe whether the predictive coding framework itself is effectual for preventing catastrophic forgetting. For all datasets, the average performance of the network trained with SGD based on the predictive coding manner outperformed the performance of the network trained with SGD based on backpropagation. Furthermore, learning with predictive coding exceeds strong competitor EWC [94] on disjoint-MNIST and split-MNIST.

To make the challenging experimental settings, we combined two classes into one task and created five tasks using MNIST and CIFAR-10, similar to the study by [173].

Table 6.5: Comparison of incremental learning performance (%) on disjoint-MNIST. We denoted the learning with backpropagation as BP and learning with the predictive coding framework as PC. We used the five random seeds in the experiments and reported the average performance between task 1 and task 2.

Algorithm	Method	Task1	Task2	Average
BP	SGD [57]	88.19	98.99	93.59
	SGD-F [57]	99.61	84.56	92.09
	EWC [94]	92.29	98.99	95.64
	IMM-MEAN [108]	98.22	97.10	97.66
	IMM-MODE [108]	85.51	98.47	91.99
	LFL [87]	93.20	65.78	79.49
	LWF [111]	99.43	98.84	99.13
PC	SGD [57]	92.80	98.91	95.85

Incremental learning performance of backpropagation and predictive coding on split-MNIST and split-CIFAR-10 is shown in Tables 6.7 and 6.8. The performance of incremental learning based on predictive coding was also compared with that of conventional approaches [57, 87, 94, 108, 111]. To observe its ability to retain previously obtained knowledge, we visualized the average accuracy of trained tasks in Fig. 6.2. Fig. 6.2 and Table 6.7 are the experimental results from the same protocol (split-MNIST). After finishing every epoch, we evaluated the performance of all the tasks and drew Fig. 6.2. While Table 5 shows the results of the average evaluation five times using the best model derived from each task. It was confirmed that catastrophic forgetting occurred in both learning algorithms, but the degree of forgetting was certainly more severe in the experimental results of backpropagation. Learning with predictive coding showed stable performance even when the learning task changed, in contrast to the pattern of backpropagation. In the backpropagation experiment, when the network acquired the knowledge of task 3, the knowledge of task 2 was forgotten. Further,

Table 6.6: Comparison of incremental learning performance (%) on disjoint-FMNIST. We denoted the learning with backpropagation as BP and learning with the predictive coding framework as PC. We used the five random seeds in the experiments and reported the average performance between task 1 and task 2.

Algorithm	Method	Task1	Task2	Average
BP	SGD [57]	67.37	97.47	82.42
	SGD-F [57]	91.87	82.06	86.96
	EWC [94]	88.79	96.66	92.72
	IMM-MEAN [108]	85.70	95.46	87.78
	IMM-MODE [108]	64.15	96.33	80.24
	LFL [87]	79.00	83.01	81.00
	LWF [111]	91.24	97.35	94.30
PC	SGD [57]	75.68	97.11	86.40

when the network learned knowledge of task 5, it was confirmed that the discriminative information of tasks 1 and 2 was removed from the memories. These experimental results confirm that a biologically plausible learning algorithm reduces catastrophic forgetting in incremental learning and enhances the performance of incremental learning as one of MCTs.

We carried out additional experiments to demonstrate the advantages of learning with the brain-inspired algorithm. We implemented the predictive coding version of EWC [94], IMM-MEAN [108], and IMM-MODE [108] algorithms and evaluated their performance on disjoint-MNIST. In the EWC algorithm, learning with predictive coding improves the average performance from 95.64% to 97.52%. In addition, learning with predictive coding enhances the average performance 0.21% and 5.42% in IMM-MEAN and IMM-MODE, respectively.

Table 6.7: Comparison of incremental learning performance (%) on split-MNIST. We denoted the learning with backpropagation as BP and learning with the predictive coding framework as PC. We used the five random seeds in the experiments and reported the average performance from task 1 to task 5.

Algorithm	Method	Task1	Task2	Task3	Task4	Task5	Average
BP	SGD [57]	98.52	74.06	93.74	96.43	99.61	92.47
	SGD-F [57]	99.95	90.52	95.43	98.06	87.38	94.27
	EWC [94]	99.41	75.24	94.21	96.34	99.60	92.96
	IMM-MEAN [108]	99.94	98.67	94.38	96.55	88.33	95.57
	IMM-MODE [108]	99.88	74.20	95.27	97.47	99.42	93.25
	LFL [87]	94.34	52.62	54.34	70.63	89.36	72.26
	LWF [111]	99.95	99.10	99.77	99.83	99.76	99.68
PC	SGD [57]	99.89	97.09	99.28	99.39	98.37	98.80

6.4 Limited Data Recognition with Predictive Coding

The potential of predictive coding networks for limited data recognition was then investigated. Specifically, the efficacy of predictive coding networks in long-tailed recognition and few-shot recognition type of MCTs was analyzed. First, real-world datasets are often highly imbalanced following long-tail distribution in which data category accounts for a significant portion of the overall data [85, 122]. Owing to the skewed class distribution of the dataset, the network trained with a class-imbalanced dataset may show a classification bias problem in which the samples of tail classes are predicted as head classes [20]. In addition, managing few-shot samples in an open-world setting is crucial because it is similar to the situation in which the human recognition system can be encountered. Second, to achieve more human-like recognition performance, effectively managing few-shot examples in an open-world setting is crucial. Two experimental scenarios are significant because it is realistic situations that human

Table 6.8: Comparison of incremental learning performance (%) on split-CIFAR-10. We denoted the learning with backpropagation as BP and learning with the predictive coding framework as PC. We used the five random seeds in the experiments and reported the average performance from task 1 to task 5.

Algorithm	Method	Task1	Task2	Task3	Task4	Task5	Average
BP	SGD [57]	72.17	66.08	71.44	84.17	93.71	77.51
	SGD-F [57]	95.72	67.96	60.03	69.97	77.38	74.15
	EWC [94]	72.76	64.90	67.53	73.99	72.15	70.26
	IMM-MEAN [108]	89.71	78.35	78.51	74.73	78.91	80.04
	IMM-MODE [108]	76.14	67.07	73.63	84.79	93.87	79.10
	LFL [87]	71.50	59.30	71.71	84.47	84.85	74.37
	LWF [111]	76.95	70.58	78.46	94.34	93.99	82.86
PC	SGD [57]	70.42	74.27	80.70	87.21	90.96	80.71

recognition can encounter.

The cortical neuron in the human brain can learn with only a few repetitions owing to the local synaptic plasticity [219], and it is widely known that such plasticity contributes to the interactions between limited data [203]. It has been demonstrated that the changes in synaptic connections assist in learning new information and long-term memory formation [213]. Given the characteristics of synaptic plasticity, experiments with a predictive coding framework were performed on the class-imbalanced data, and the biologically plausible learning algorithm that helped limited data recognition was identified.

6.4.1 Experimental Settings

The same architecture used in the previous section consisting of three-layer MLP was used in long-tailed recognition. The number of hidden neurons was set as 800 with ReLU non-linearity and dropout. We used MNIST [103] for our experiment and syn-

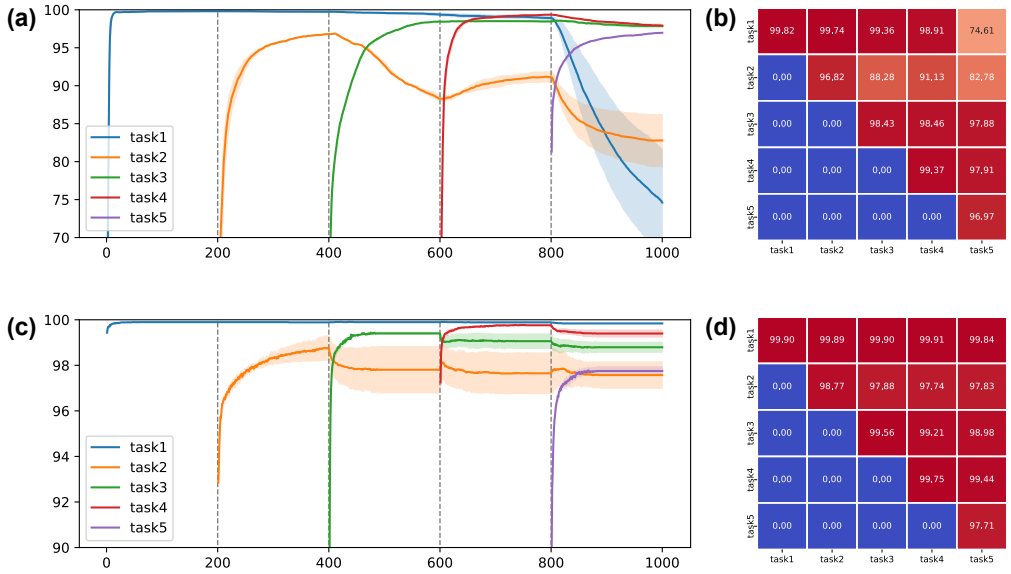

Figure 6.2: Qualitative and quantitative performance comparison on two learning schemes for (A-B) backpropagation and (C-D) predictive coding on split-MNIST. In (A) and (C), the solid line indicates the average accuracy for each task and the transparent region represents the standard deviation on five random seeds. The vertical dashed line refers to the point at which the task to be learned changes. In (B) and (D), each value indicates the performance of each task measured by the final model.

thesized the long-tailed data with an imbalance ratio γ. The imbalance ratio was defined as the proportion of the samples of the highest number of classes to the lowest number of classes as $\frac{N_{max}}{N_{min}}$. Although it differed depending on the imbalance ratio, in general, N_{max} and N_{min} usually followed the relationship, $N_{max} \gg N_{min}$. Exponential distribution and the number of samples N_l in l-th class was defined as $N_l = N_{max} \cdot \gamma^{-\frac{l-1}{L-1}}$. The four types of imbalanced data distribution were then synthesized as previously described [90]. To train a network, we set a batch size of 128 and optimized a model until 100 epochs. When backpropagation was used for learning, the learning rate was increased from 0.0001 to 0.5 by growing five times, and the best performance results among them were determined. When predictive coding was

used for the optimization, a learning rate of 0.002 with a weight decay of $2e^{-4}$ was used. Additionally, the weight learning rate η was set as 0.1 and the number of iterations as 20 as hyperparameters for predictive coding networks. All the experiments with predictive coding were performed under the fixed prediction assumption. We provide the code to reproduce the results in the manuscript at `https://github.com/jangho2001us/PredictiveCoding_LongTailedRecognition`.

In few-shot recognition, the same experimental settings with those of [171], which comprised four convolutional blocks with Batch normalization, ReLU, and MaxPool were used. Experiments on few-shot recognition were conducted with Omniglot [100] dataset containing 1623 categories of handwritten characters. The performance of few-shot recognition is commonly measured by N-way k-shot classification, where N implies the number of given classes and k indicates the number of samples in each category. The current study extended the experimental protocol of the original paper to 30-way k-shot experiment settings because those evaluation protocols are more difficult because the number of classes for the candidate group increases. The learning rate was set to $1e^{-3}$ and then reduced by $1/10$ every 20 epoch to train a network. The same learning rate, weight decay, weight learning rate, and iterations were used for learning networks with a predictive coding framework. For more information, please refer to the original paper [171]. We provide the code to reproduce the results in the manuscript at `https://github.com/jangho2001us/PredictiveCoding_FewShotRecognition`.

6.4.2 Experiments on Long-tailed Recognition

In Table 6.9, we compared the long-tailed recognition performance with Cross-Entropy (CE) loss, Mixup approach [226], Focal loss [116], Class-Balanced Focal (CB Focal) loss [36], Label-Distribution-Aware-Margin (LDAM) loss [20], and Balanced Meta-Softmax (BALMS) loss [156]. The experimental results showed the benefit of learning with predictive coding networks. First, the long-tailed recognition performance was

higher by 4.45% in learning the network with a predictive coding framework than in learning with CE loss under severe class imbalance of data distribution. Similar results in the following experiments were observed when the network was trained with other learning objectives such as Focal [116] and BALMS [156]. In this experiment, the performance improvement is evaluated using the predictive coding framework rather than comparing performance between different learning objectives. The results shown in Table 6.9 indicate that the learning algorithm close to the human brain brings a positive effect on MCTs, confirming our assumption.

6.4.3 Experiments on Few-shot Recognition

The few-shot recognition performance trained with backpropagation and predictive coding framework is shown in Table 6.10. Learning with predictive coding enabled robust recognition under the various few-shot experimental protocols. Additionally, predictive coding networks showed their potential ability under challenging inference settings such as 20-way 1-shot and 30-way 1-shot rather than 20-way 5-shots and 30-way 5-shots. The experimental results confirmed our assumptions and supported that the brain-like learning algorithm was effective for MCTs.

6.5 Discussion

6.5.1 Analysis of Plasticity-stability Aspects

The plasticity-stability dilemma is a well-known problem widely studied in biology [129]. This phenomenon is related to the power of consolidating new information without forgetting previously acquired information [132]. Further, it is an essential issue in incremental learning with ANNs [114]. The human brain is well-controlled to learn new information and to prevent the learned information from being overridden by the new information [178]. However, ANNs naturally induce catastrophic forgetting and expose the trade-off between plasticity and stability [94].

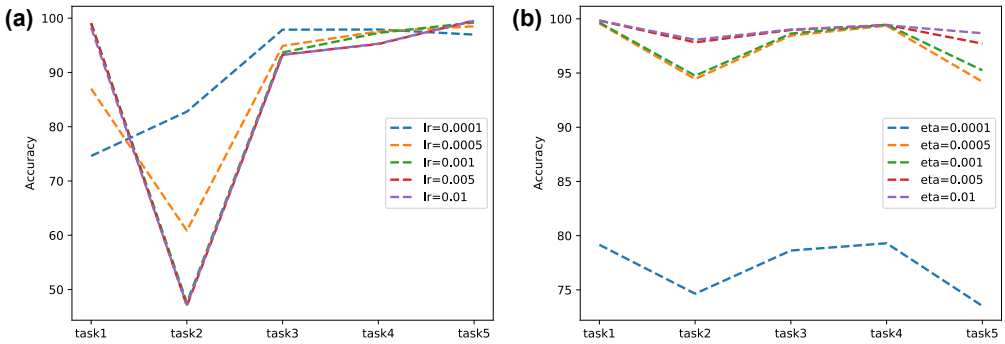

Figure 6.3: Comparison of learning with (A) backpropagation and (B) predictive coding on split-MNIST in two learning schemes. To adjust network stability, the learning rate of backpropagation and the weight learning rate of predictive coding are varied.

To confirm that predictive coding achieves a better plasticity-stability trade-off than backpropagation, we experimented with split-MNIST by controlling the stability of two learning mechanisms. Adjusting the learning rate is not directly related to stability, but it was used because it was considered a factor that could adjust stability in our experiments. In Fig. 6.3, we report the experimental results and compare the learning schemes by adjusting the learning rate of backpropagation and the weight learning rate of predictive coding. In backpropagation experiments, the learning is reduced from 0.01 to 0.0001 to decrease forgetting of acquired knowledge. When the learning rate was 0.0001, the network forgot less information to perform task 2. However, it still showed limited performance in tasks 1 and 2. Thus, maintaining stability by reducing the learning rate may not be acceptable because it deteriorates the overall performance. Meanwhile, performance was consistently high for each task in predictive coding experiments. These results implied predictive coding had better plasticity properties than backpropagation while maintaining stability.

6.5.2 Interplay of Hippocampus and Prefrontal Cortex

The hippocampus plays an essential role in episodic memory at the top of the cortical processing hierarchy [48]. In incremental learning, the ability to regulate learned information and retrieve context-appropriate memories is essential. We can understand the effectiveness of predictive coding in incremental learning as the interaction between the hippocampus and the prefrontal cortex in the human brain [7, 46]. It is well known that the hippocampus can quickly encode new information, stabilize memory traces, and organize memory networks [152]. In addition, this mechanism has been physiologically proven through functional magnetic resonance imaging studies [74].

We have shown that the learning process of predictive coding networks is analogous to the interaction between the hippocampus and the prefrontal cortex in the human brain [46]. As described in Algorithm 1, the learning process based on predictive coding networks can be divided into two phases: forward and backward pass. In the forward phase, the predictive coding network computes its predictions for every layer. In the backward phase, the predictive coding network minimizes the free-energy summation as a learning objective. The two-phase learning of predictive coding networks corresponds to acquiring and consolidating information in the hippocampus and prefrontal cortex. The predictive coding framework promotes the two processes and enables accurate inference when data containing information corresponding to the previously learned task are received.

6.5.3 Rationale for Selecting Predictive Coding

The reason why we selected predictive coding as a brain-inspired algorithm is as follows. Predictive coding is potentially more biologically plausible because *local* learning rules perform parameter updates. This property is distinct from the update of backpropagation executed from the global error signal. It will be ideal if the parameter update is performed asynchronously in a different layer, such as the neural plasticity of the human brain. However, the parameter update of predictive coding occurs under

the *fixed prediction assumption* [133]. The fixed prediction assumption implies that the parameters are updated based on the *fixed* predictions of the forward phase. Whittington *et al.* [197] demonstrated that a predictive coding network with a fixed prediction assumption performs the same parameter updates as backpropagation. Another limitation of predictive coding is the degree of convergence of variational free energy used as a learning objective. The convergence of the backward phase is achieved by setting a specific number of iterations [158]. In general, predictive coding requires n times computational cost than that of backpropagation. We set the number of iterations for all examples as 20. It means that the training will be 20 times slower than the backpropagation training. Depending on the number of backward iterations, learning with predictive coding may converge or diverge. Millidge *et al.* [133] performed 100 iterations, but we successfully trained the networks with 20 iterations. Although these two issues introduced earlier remain open questions, we conducted our experiments using predictive coding because we thought its advantages outweighed its disadvantages.

6.6 Summary

This study empirically demonstrated the potential effectiveness of predictive coding in MCTs. However, despite this, the predictive coding algorithm still has some limitations. First, predictive coding requires a longer training time than backpropagation because it executes backward iteration until the error nodes and activation nodes converge. Although we expanded our experiments for large networks such as VGGNet and ResNet [70, 99], we could not perform the experiments on MCTs because of the excessive training time. Second, to conduct learning with predictive coding, the network should be an architecture composed of sequential layers. For example, if shortcut connections exist, it is challenging to implement them into a predictive coding layer. In this case, we set the block unit, which is the boundary of the shortcut, as the predictive coding layer. If predictive coding combines learning speed and scalability, there

will be infinite opportunities for development as a learning algorithm that can replace backpropagation.

In summary, we extensively analyze the benefits of learning ANNs with predictive coding frameworks for MCTs. MCTs can be described as tasks that are easy for human intelligence while difficult for machine intelligence. Based on our hypothesis, we empirically demonstrate that brain-inspired predictive coding has advantages in incremental learning on MNIST and CIFAR, long-tailed recognition on MNIST, and few-shot recognition on Omniglot. In neuroscience, especially the intrinsic properties of the human brain, we discuss why training ANNs with a predictive coding framework improves the performance of MCTs. The study concludes that predictive coding learning is similar to the plasticity-stability property of the human brain and mainly mimics the interaction between the hippocampus and prefrontal cortex. Finally, it is an interesting avenue for future work to reduce the training time under the fixed prediction assumption and relax the constraint of predictive coding while maintaining the performance.

Table 6.9: Comparison of classification performance (%) on MNIST under four different imbalance distributions. Experiments are performed with five random seeds, and the average performance is reported. Relative variance is provided in the bracket. Increments are presented as red and decrements as blue.

Algorithm	Objective Function	Imbalance Ratio (γ)			
		200	100	50	10
BP	CE	68.78	78.06	89.63	97.17
	Mixup [226]	67.60	76.69	86.97	96.15
	Focal [116]	70.92	79.42	90.89	97.31
	CB Focal [36]	69.93	79.72	91.26	97.09
	LDAM [20]	65.17	75.58	84.91	97.14
	BALMS [156]	72.25	81.34	92.50	97.23
PC	CE	73.23 (+4.45)	79.26 (+1.20)	90.10 (+0.47)	97.37 (+0.20)
	Mixup [226]	67.77 (+0.17)	77.60 (+0.91)	88.26 (+1.29)	96.27 (+0.12)
	Focal [116]	71.99 (+1.07)	79.57 (+0.15)	91.18 (+0.29)	97.03 (-0.28)
	CB Focal [36]	70.19 (+0.26)	80.28 (+0.56)	91.40 (+0.14)	97.24 (+0.14)
	LDAM [20]	65.54 (+0.37)	76.05 (+0.47)	85.08 (+0.17)	97.20 (+0.06)
	BALMS [156]	74.22 (+1.97)	82.28 (+0.94)	93.50 (+1.00)	97.45 (+0.22)

Table 6.10: Experimental results on the low-shot recognition on the Omniglot dataset. Five random seeds are used in the experiment, and the average performance is reported. Relative variance is shown in the bracket. Increments are presented as red.

Algorithm	Method	5-way Acc.		10-way Acc.		20-way Acc.		30-way Acc.	
		1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
BP	ProtoNet [171]	98.41	99.56	96.87	99.18	94.64	98.54	92.97	97.98
PC	ProtoNet [171]	98.46 (+0.05)	99.59 (+0.03)	96.98 (+0.11)	99.19 (+0.01)	94.88 (+0.24)	98.59 (+0.05)	93.14 (+0.17)	98.05 (+0.07)

Chapter 7

Conclusion

Modern deep learning relies substantially on high-quality labeled data, computing hardware, and an effective learning algorithm to be successful in sophisticated applications. In this dissertation, we consider high-quality labeled data and learning algorithms necessary for deep learning to produce human-level artificial intelligence. In reality, constructing high-quality labeled data demands time and cost-inefficient operations, which is also a research direction that explores effective learning methods using imperfect data. Although the ANN was introduced to simulate human brain behavior, there is a problem that the widely used learning algorithm [159] itself does not match that of a human brain in terms of physiology and anatomy. These properties contrast with the perception of humans, showing noise-resistant and excellent generalization performance under imperfect training data. In this chapter, we summarize our contributions in terms of data to effectively handle several imperfect data conditions and the algorithm to simulate imperfect data recognition with a human-like learning mechanism.

7.1 Dissertation Summary

In Chapter 3, we introduced a novel but simple method for solving class-imbalanced semi-supervised learning. We analyzed the behavior of the state-of-the-art SSL algorithm under a class-imbalanced scenario and confirmed that there exists a classification bias toward the majority classes. Motivated by the observation of a state-of-the-art SSL algorithm, we proposed the recycling loss to engage abandoned samples in a learning procedure with two masks. We generated a confidence mask to filter out samples producing high softmax prediction and created a semantic mask to detect samples with low consistency on different views. Our experiments demonstrated that the ensemble of the two proposed masks improves the state-of-the-art SSL algorithm on various long-tailed datasets under a class-imbalanced scenario. Our method especially achieved competitive performance on modern class-imbalanced approaches that explicitly re-balance the classifier based on the imbalance distribution of labeled data. We also presented a qualitative analysis and ablation study to prove the efficacy of each mask.

In Chapter 4, we proposed two modules, TREM and TRAM, for addressing audio-visual semantic inconsistency in unconstrained videos. TREM was designed to extract high-quality temporal representation by broadening the temporal field of view on multimodal features. Thereafter, TRAM was proposed for exploring the global scope semantic relation with cross-modal self-attention on the cross-modal features. By jointly training two modules, state-of-the-art AVE localization performance was achieved in supervised and weakly supervised experimental settings on the AVE dataset. Moreover, we conducted extensive ablation studies to verify that the proposed method resolved temporal inconsistency in unconstrained videos and improved temporal semantic alignment both quantitatively and qualitatively.

In Chapter 5, we presented a novel deep clustering method to correct the bias arising from the unsupervised experimental setting. We proposed the feature contrast regularizer which learns the clustering-favorable representation. We have conducted

extensive experiments on the five challenging datasets and outperformed state-of-the-art joint deep clustering approaches. With extensive ablation studies, we analyzed the effect of each learning objective. We implicitly demonstrated that our method generates a more clustering-favorable representation feature extractor having more discriminative properties by providing the results on the linear evaluation and k-NN classifier performance.

In Chapter 6, we extensively analyzed the benefits of learning ANNs with predictive coding frameworks for MCTs. MCTs can be described as tasks that are easy for human intelligence while difficult for machine intelligence. Based on our hypothesis, we empirically demonstrated that brain-inspired predictive coding has advantages in incremental learning on MNIST and CIFAR, long-tailed recognition on MNIST, and few-shot recognition on Omniglot. In neuroscience, especially the intrinsic properties of the human brain, we discussed why training ANNs with a predictive coding framework improves the performance of MCTs. The study concluded that predictive coding learning is similar to the plasticity-stability property of the human brain and mainly mimics the interaction between the hippocampus and prefrontal cortex.

7.2 Suggestion for Future Research

7.2.1 Overcoming Limitations of Predictive Coding

In Chapter 6, we presented empirical evidence proving the potential benefits of predictive coding in MCTs. However, despite these advantages, the predictive coding algorithm still has some limitations. First, predictive coding requires a longer training time than backpropagation because it executes backward iteration until the error nodes and activation nodes converge. Although we expanded our experiments for large networks such as VGGNet and ResNet [70, 99], we could not perform the experiments on MCTs because of the excessive training time. To effectively improve this, parallelization of predictive coding learning is essential. In backpropagation, decoupled parallel

backpropagation algorithm [79] guaranteed the convergence of parallelized backpropagation and significantly reduced the training time. However, in predictive coding, the study focused on attaining competitive performance with backpropagation. A recent study called incremental predictive coding [163] is the first approach to improve the learning speed of predicated coding. If predictive coding combines learning speed and scalability, there will be infinite opportunities for development as a learning algorithm that can replace backpropagation. The second limitation is the fixed prediction assumption, which is a strong constraint for successful learning. The experiments in our dissertation are all based on this assumption. According to Rosenbaum *et al.* [158], it was proved that backpropagation learning could not be completely approximated in the absence of fixed prediction assumption. Therefore, it is crucial to simulate the learning aspect of backpropagation by alleviating the fixed prediction assumption. Finally, for predictive coding-based learning, the network's architecture should consist of consecutive layers. For instance, replacing a shortcut connection in ResNet with a predictive coding layer is difficult. In this case, a new local learning rule should be proposed to transform this structure into the predictive coding layer. It is an interesting avenue for future work to reduce the training time under the fixed prediction assumption and relax the constraint of predictive coding while maintaining the performance.

7.2.2 Exploration of the Human Memory Properties

As a future work, it would be interesting to explore the properties of the visual and human memory systems related to predictive coding. First, regarding the visual system, Rao and Ballad [155] demonstrated that predictive coding brings extra-classical receptive-field effects, and this discovery expanded to the temporal properties of the human visual system. Second, exploration between predictive coding and the human memory system. Recently, some studies [162, 221] demonstrated that the model trained with predictive coding naturally implements associative memory in the brain. The authors experimented the associative properties such as image retrieval and recov-

ery experiments. Inspired by this study, we speculated that learning with predictive coding can potentially implement other memory functions in the human brain. The memory function we are paying attention to is long-term memory. Humans have an exceptional capacity for long-term memory storage and recall. At this time, a person expands short-term memory to long-term memory by introducing rehearsal [34]. By introducing a human-performed rehearsal process into predictive coding to verify the long-term memory function, we can make the basis for our assumptions more reliable.

Bibliography

[1] Sitara Afzal, Muazzam Maqsood, Faria Nazir, Umair Khan, Farhan Aadil, Khalid M Awan, Irfan Mehmood, and Oh-Young Song. A data augmentation-based framework to handle class imbalance problem for alzheimer's stage detection. *IEEE Access*, 2019.

[2] Nasir Ahmad, Marcel A van Gerven, and Luca Ambrogioni. Gait-prop: A biologically plausible learning rule derived from backpropagation of error. *Advances in Neural Information Processing Systems*, 33:10913–10923, 2020.

[3] Mohamed Akrout, Collin Wilson, Peter Humphreys, Timothy Lillicrap, and Douglas B Tweed. Deep learning without weight transport. *Advances in Neural Information Processing Systems*, 32, 2019.

[4] H. Alwassel, D. Mahajan, B. Korbar, L. Torresani, B. Ghanem, and D. Tran. Self-supervised learning by cross-modal audio-video clustering. *Advances in Neural Information Processing Systems*, 33, 2020.

[5] Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *International Joint Conference on Neural Networks*, pages 1–8. IEEE, 2020.

[6] J. L. Ba, J. R. Kiros, and G. E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[7] Helen C Barron, Ryszard Auksztulewicz, and Karl Friston. Prediction and memory: A predictive coding account. *Progress in neurobiology*, 192:101821, 2020.

[8] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm. Mine: mutual information neural estimation. *arXiv preprint arXiv:1801.04062*, 2018.

[9] Yoshua Bengio, Pascal Lamblin, Dan Popovici, Hugo Larochelle, et al. Greedy layer-wise training of deep networks. *Advances in Neural Information Processing Systems*, 2007.

[10] David Berthelot, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. In *International Conference on Learning Representations*, 2019.

[11] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *Advances in Neural Information Processing Systems*, 2019.

[12] Maxwell A Bertolero, BT Thomas Yeo, and Mark D'Esposito. The modular and integrative functional architecture of the human brain. *Proceedings of the National Academy of Sciences*, 112(49):E6798–E6807, 2015.

[13] Chris M Bird and Neil Burgess. The hippocampus and memory: insights from spatial processing. *Nature Reviews Neuroscience*, 9(3):182–194, 2008.

[14] Rafal Bogacz. A tutorial on the free-energy framework for modeling perception and learning. *Journal of mathematical psychology*, 76:198–211, 2017.

[15] Paula Branco, Luís Torgo, and Rita P Ribeiro. A survey of predictive modeling on imbalanced domains. *ACM Computing Surveys*, 2016.

[16] Christopher L Buckley, Chang Sub Kim, Simon McGregor, and Anil K Seth. The free energy principle for action and perception: A mathematical review. *Journal of Mathematical Psychology*, 81:55–79, 2017.

[17] Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 2018.

[18] Deng Cai, Xiaofei He, Xuanhui Wang, Hujun Bao, and Jiawei Han. Locality preserving nonnegative matrix factorization. In *International Joint Conference on Artificial Intelligence*, 2009.

[19] Xiao Cai, Feiping Nie, Heng Huang, and Farhad Kamangar. Heterogeneous image feature integration via multi-modal spectral clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2011.

[20] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems*, 2019.

[21] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[22] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021.

[23] Jianlong Chang, Yiwen Guo, Lingfeng Wang, Gaofeng Meng, Shiming Xiang, and Chunhong Pan. Deep discriminative clustering analysis. *arXiv preprint arXiv:1905.01681*, 2019.

[24] Jianlong Chang, Lingfeng Wang, Gaofeng Meng, Shiming Xiang, and Chunhong Pan. Deep adaptive image clustering. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[25] Olivier Chapelle, Bernhard Scholkopf, and Alexander Zien. Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. *IEEE Transactions on Neural Networks*, 20(3):542–542, 2009.

[26] Olivier Chapelle and Alexander Zien. Semi-supervised classification by low density separation. In *International workshop on artificial intelligence and statistics*, 2005.

[27] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 2002.

[28] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, 2020.

[29] Bhavin Choksi, Milad Mozafari, Callum Biggs O'May, Benjamin Ador, Andrea Alamia, and Rufin VanRullen. Predify: Augmenting deep neural networks with brain-inspired predictive coding dynamics. *Advances in Neural Information Processing Systems*, 34, 2021.

[30] Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning. *Advances in Neural Information Processing Systems*, 2020.

[31] Ami Citri and Robert C Malenka. Synaptic plasticity: multiple forms, functions, and mechanisms. *Neuropsychopharmacology*, 33(1):18–41, 2008.

[32] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Society for Artificial Intelligence and Statistics*, 2011.

[33] Roberto Colom, Sherif Karama, Rex E Jung, and Richard J Haier. Human intelligence and brain networks. *Dialogues in clinical neuroscience*, 2022.

[34] Fergus IM Craik and Michael J Watkins. The role of rehearsal in short-term memory. *Journal of verbal learning and verbal behavior*, 12(6):599–607, 1973.

[35] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020.

[36] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[37] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2005.

[38] Zhiyuan Dang, Cheng Deng, Xu Yang, Kun Wei, and Heng Huang. Nearest neighbor matching for deep clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[39] Lila Davachi and Sarah DuBrow. How the hippocampus preserves order: the role of prediction and context. *Trends in cognitive sciences*, 19(2):92–99, 2015.

[40] Ruben De Man, Grace J Gang, Xin Li, and Ge Wang. Comparison of deep learning and human observer performance for detection and characterization of simulated lesions. *Journal of Medical Imaging*, 6(2):025503, 2019.

[41] Giorgia Dellaferrera and Gabriel Kreiman. Error-driven input modulation: Solving the credit assignment problem without a backward pass. In *International Conference on Machine Learning*, pages 4937–4955, 2022.

[42] Sophie Denève, Alireza Alemi, and Ralph Bourdoukan. The brain as an efficient and robust adaptive learner. *Neuron*, 94(5):969–977, 2017.

[43] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2009.

[44] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

[45] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.

[46] Howard Eichenbaum. Prefrontal–hippocampal interactions in episodic memory. *Nature Reviews Neuroscience*, 18(9):547–558, 2017.

[47] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 2010.

[48] Daniel J Felleman and David C Van Essen. Distributed hierarchical processing in the primate cerebral cortex. *Cerebral cortex (New York, NY: 1991)*, 1(1):1–47, 1991.

[49] F. Frassinetti, N. Bolognini, and E. Làdavas. Enhancement of visual perception by crossmodal visuo-auditory interaction. *Experimental brain research*, 2002.

[50] Robert M French. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3(4):128–135, 1999.

[51] Karl Friston. Learning and inference in the brain. *Neural Networks*, 16(9):1325–1352, 2003.

[52] Karl Friston. A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1456):815–836, 2005.

[53] Karl Friston. Hierarchical models in the brain. *PLoS computational biology*, 4(11):e1000211, 2008.

[54] Santosh K Gaikwad, Bharti W Gawali, and Pravin Yannawar. A review on speech recognition technique. *International Journal of Computer Applications*, 10(3):16–24, 2010.

[55] Robert Geirhos, David HJ Janssen, Heiko H Schütt, Jonas Rauber, Matthias Bethge, and Felix A Wichmann. Comparing deep neural networks against humans: object recognition when the signal gets weaker. *arXiv preprint arXiv:1706.06969*, 2017.

[56] J. F. Gemmeke, D. PW. Ellis, D. Freedman, et al. Audio set: An ontology and human-labeled dataset for audio events. In *IEEE International Conference on Acoustics Speech and Signal Processing*, 2017.

[57] Ian J Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. An empirical investigation of catastrophic forgetting in gradient-based neural networks. *arXiv preprint arXiv:1312.6211*, 2013.

[58] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[59] K Chidananda Gowda and G Krishna. Agglomerative clustering using the concept of mutual nearest neighbourhood. *Pattern Recognition*, 1978.

[60] Sorin Grigorescu, Bogdan Trasnea, Tiberiu Cocias, and Gigel Macesanu. A survey of deep learning techniques for autonomous driving. *Journal of Field Robotics*, 2020.

[61] Stephen Grossberg. Competitive learning: From interactive activation to adaptive resonance. *Cognitive Science*, 11(1):23–63, 1987.

[62] Xifeng Guo, Long Gao, Xinwang Liu, and Jianping Yin. Improved deep embedded clustering with local structure preservation. In *International Joint Conference on Artificial Intelligence*, 2017.

[63] Philip Haeusser, Johannes Plapp, Vladimir Golkov, Elie Aljalbout, and Daniel Cremers. Associative deep clustering: Training a classification network with no labels. In *German Conference on Pattern Recognition*, 2018.

[64] Kuan Han, Haiguang Wen, Yizhen Zhang, Di Fu, Eugenio Culurciello, and Zhongming Liu. Deep predictive coding network with local recurrent processing for object recognition. *Advances in Neural Information Processing Systems*, 31, 2018.

[65] Sungwon Han, Sungwon Park, Sungkyu Park, Sundong Kim, and Meeyoung Cha. Mitigating embedding and class assignment mismatch in unsupervised image classification. In *Proceedings of the European Conference on Computer Vision*, 2020.

[66] John A Hartigan and Manchek A Wong. A k-means clustering algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 1979.

[67] Demis Hassabis, Dharshan Kumaran, Christopher Summerfield, and Matthew Botvinick. Neuroscience-inspired artificial intelligence. *Neuron*, 95(2):245–258, 2017.

[68] Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 2009.

[69] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[70] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016.

[71] Donald Olding Hebb. *The organization of behavior: A neuropsychological theory*. Psychology Press, 2005.

[72] John R Hershey, Zhuo Chen, Jonathan Le Roux, and Shinji Watanabe. Deep clustering: Discriminative embeddings for segmentation and separation. In *IEEE International Conference on Acoustics Speech and Signal Processing*, 2016.

[73] S. Hershey, S. Chaudhuri, D. PW. Ellis, et al. Cnn architectures for large-scale audio classification. In *IEEE International Conference on Acoustics Speech and Signal Processing*, 2017.

[74] Nicholas C Hindy, Emily W Avery, and Nicholas B Turk-Browne. Hippocampal-neocortical interactions sharpen over time for predictive actions. *Nature Communications*, 10(1):1–13, 2019.

[75] Seunghoon Hong, Jonghyun Choi, Jan Feyereisl, Bohyung Han, and Larry S Davis. Joint image clustering and labeling by matrix factorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015.

[76] D. Hu, R. Qian, M. Jiang, et al. Discriminative sounding objects localization

via self-supervised audiovisual matching. *Advances in Neural Information Processing Systems*, 33, 2020.

[77] Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. Learning deep representation for imbalanced classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016.

[78] Jiabo Huang, Shaogang Gong, and Xiatian Zhu. Deep semantic clustering by partition confidence maximisation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[79] Zhouyuan Huo, Bin Gu, Heng Huang, et al. Decoupled parallel backpropagation with convergence guarantee. In *International Conference on Machine Learning*, pages 2098–2106. Proceedings of Machine Learning Research, 2018.

[80] Tri Huynh, Simon Kornblith, Matthew R Walter, Michael Maire, and Maryam Khademi. Boosting contrastive self-supervised learning with false negative cancellation. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022.

[81] Bernd Illing, Jean Ventura, Guillaume Bellec, and Wulfram Gerstner. Local plasticity rules can learn deep representations using self-supervised contrastive predictions. *Advances in Neural Information Processing Systems*, 34, 2021.

[82] Jiyong Jang and Sungroh Yoon. Feature concentration for supervised and semisupervised learning with unbalanced datasets in visual inspection. *IEEE Transactions on Industrial Electronics*, 2020.

[83] Nathalie Japkowicz. The class imbalance problem: Significance and strategies. In *International Conference on Artificial Intelligence*, 2000.

[84] Xu Ji, João F Henriques, and Andrea Vedaldi. Invariant information clustering

for unsupervised image classification and segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019.

[85] Justin M Johnson and Taghi M Khoshgoftaar. Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1):1–54, 2019.

[86] Armand Joulin, Francis Bach, and Jean Ponce. Discriminative clustering for image co-segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2010.

[87] Heechul Jung, Jeongwoo Ju, Minju Jung, and Junmo Kim. Less-forgetting learning in deep neural networks. *arXiv preprint arXiv:1607.00122*, 2016.

[88] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *International Conference on Learning Representations*, 2020.

[89] E. Kazakos, A. Nagrani, A. Zisserman, and D. Damen. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019.

[90] Jaehyung Kim, Youngbum Hur, Sejun Park, Eunho Yang, Sung Ju Hwang, and Jinwoo Shin. Distribution aligning refinery of pseudo-label for imbalanced semi-supervised learning. In *Advances in Neural Information Processing Systems*, 2020.

[91] Jaehyung Kim, Jongheon Jeong, and Jinwoo Shin. M2m: Imbalanced classification via major-to-minor translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[92] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[93] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[94] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

[95] Konrad P Körding, Ulrik Beierholm, Wei Ji Ma, Steven Quartz, Joshua B Tenenbaum, and Ladan Shams. Causal inference in multisensory perception. *PLoS one*, 2007.

[96] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical Report 0, University of Toronto, Toronto, Ontario, 2009.

[97] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[98] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25, 2012.

[99] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[100] Brenden Lake, Ruslan Salakhutdinov, Jason Gross, and Joshua Tenenbaum. One shot learning of simple visual concepts. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 33, 2011.

[101] Marc T Law, Raquel Urtasun, and Richard S Zemel. Deep spectral clustering learning. In *International Conference on Machine Learning*, 2017.

[102] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, and G. D. Hager. Temporal convolutional networks for action segmentation and detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017.

[103] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[104] Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 1999.

[105] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, In International Conference on Machine Learning*, 2013.

[106] Dong-Hyun Lee, Saizheng Zhang, Asja Fischer, and Yoshua Bengio. Difference target propagation. In *Joint european conference on machine learning and knowledge discovery in databases*, pages 498–515. Springer, 2015.

[107] Hyuck Lee, Seungjae Shin, and Heeyoung Kim. Abc: Auxiliary balanced classifier for class-imbalanced semi-supervised learning. *Advances in Neural Information Processing Systems*, 2021.

[108] Sang-Woo Lee, Jin-Hwa Kim, Jaehyun Jun, Jung-Woo Ha, and Byoung-Tak Zhang. Overcoming catastrophic forgetting by incremental moment matching. *Advances in Neural information processing systems*, 30, 2017.

[109] Yu Li, Tao Wang, Bingyi Kang, Sheng Tang, Chunfeng Wang, Jintao Li, and Jiashi Feng. Overcoming classifier imbalance for long-tail object detection with

balanced group softmax. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[110] Yunfan Li, Peng Hu, Zitao Liu, Dezhong Peng, Joey Tianyi Zhou, and Xi Peng. Contrastive clustering. *arXiv preprint arXiv:2009.09687*, 2020.

[111] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947, 2017.

[112] Qianli Liao, Joel Leibo, and Tomaso Poggio. How important is weight symmetry in backpropagation? In *Proceedings of the AAAI Conference on Artificial Intelligence Conference on Artificial Intelligence*, volume 30, 2016.

[113] Timothy P Lillicrap, Daniel Cownden, Douglas B Tweed, and Colin J Akerman. Random synaptic feedback weights support error backpropagation for deep learning. *Nature Communications*, 7(1):1–10, 2016.

[114] Guoliang Lin, Hanlu Chu, and Hanjiang Lai. Towards better plasticity-stability trade-off in incremental learning: A simple linear connector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 89–98, 2022.

[115] J. Lin, C. Gan, and S. Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[116] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[117] Y Lin, Y. Li, and Y. F. Wang. Dual-modality seq2seq network for audio-visual event localization. In *IEEE International Conference on Acoustics Speech and Signal Processing*, 2019.

[118] Y. Lin and Y. F. Wang. Audiovisual transformer with instance attention for audio-visual event localization. In *Asian Conference on Computer Vision*, 2020.

[119] Jack Lindsey and Ashok Litwin-Kumar. Learning to learn with feedback and local plasticity. *Advances in Neural Information Processing Systems*, 33:21213–21223, 2020.

[120] Kaiyuan Liu, Xingyu Li, Yi Zhou, Jisong Guan, Yurui Lai, Ge Zhang, Hang Su, Jiachen Wang, and Chunxu Guo. Denoised internal models: a brain-inspired autoencoder against adversarial attacks. *arXiv preprint arXiv:2111.10844*, 2021.

[121] Xiaoxuan Liu, Livia Faes, Aditya U Kale, Siegfried K Wagner, Dun Jack Fu, Alice Bruynseels, Thushika Mahendiran, Gabriella Moraes, Mohith Shamdas, Christoph Kern, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *The Lancet Digital Health*, 1(6):e271–e297, 2019.

[122] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[123] David G Lowe. Object recognition from local scale-invariant features. In *Proceedings of the IEEE International Conference on Computer Vision*, 1999.

[124] David G Lowe. Local feature view clustering for 3d object recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2001.

[125] Dengsheng Lu and Qihao Weng. A survey of image classification methods and techniques for improving classification performance. *International Journal of Remote Sensing*, 28(5):823–870, 2007.

[126] Christopher Manning and Hinrich Schutze. *Foundations of Statistical Natural Language Processing*. MIT press, 1999.

[127] Yassine Marrakchi, Osama Makansi, and Thomas Brox. Fighting class imbalance with contrastive learning. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, 2021.

[128] Marc Masana, Xialei Liu, Bartlomiej Twardowski, Mikel Menta, Andrew D Bagdanov, and Joost van de Weijer. Class-incremental learning: survey and performance evaluation on image classification. *arXiv preprint arXiv:2010.15277*, 2020.

[129] Pedro Mateos-Aparicio and Antonio Rodríguez-Moreno. The impact of studying brain plasticity. *Frontiers in Cellular Neuroscience*, 13:66, 2019.

[130] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of Learning and Motivation*, volume 24, pages 109–165. Elsevier, 1989.

[131] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *International Conference on Learning Representations*, 2020.

[132] Martial Mermillod, Aurélia Bugaiska, and Patrick Bonin. The stability-plasticity dilemma: Investigating the continuum from catastrophic forgetting to age-limited learning effects. *Frontiers in Psychology*, 4:504, 2013.

[133] Beren Millidge, Alexander Tschantz, and Christopher L Buckley. Predictive coding approximates backprop along arbitrary computation graphs. *arXiv preprint arXiv:2006.04182*, 2020.

[134] Shervin Minaee, Yuri Y Boykov, Fatih Porikli, Antonio J Plaza, Nasser Kehtar-

navaz, and Demetri Terzopoulos. Image segmentation using deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[135] Matthias Minderer, Olivier Bachem, Neil Houlsby, and Michael Tschannen. Automatic shortcut removal for self-supervised representation learning. In *International Conference on Machine Learning*, 2020.

[136] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[137] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[138] Ali Bou Nassif, Ismail Shahin, Imtinan Attili, Mohammad Azzeh, and Khaled Shaalan. Speech recognition using deep neural networks: A systematic review. *IEEE Access*, 7:19143–19165, 2019.

[139] Guilherme Neves, Sam F Cooke, and Tim VP Bliss. Synaptic plasticity, memory and the hippocampus: a neural network approach to causality. *Nature Reviews Neuroscience*, 9(1):65–75, 2008.

[140] Chuang Niu, Jun Zhang, Ge Wang, and Jimin Liang. Gatcluster: Self-supervised gaussian-attention network for image clustering. In *Proceedings of the European Conference on Computer Vision*, 2020.

[141] Youngtaek Oh, Dong-Jin Kim, and In So Kweon. Daso: Distribution-aware semantics-oriented pseudo-label for imbalanced semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9786–9796, 2022.

[142] Shay Ohayon, Winrich A Freiwald, and Doris Y Tsao. What makes a cell face selective? the importance of contrast. *Neuron*, 74(3):567–581, 2012.

[143] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 2001.

[144] Daniel W Otter, Julian R Medina, and Jugal K Kalita. A survey of the usages of deep learning for natural language processing. *IEEE Transactions on Neural Networks and Learning Systems*, 32(2):604–624, 2020.

[145] A. Owens and A. A. Efros. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European Conference on Computer Vision*, 2018.

[146] A. Owens, P. Isola, J. McDermott, et al. Visually indicated sounds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016.

[147] A. Owens, J. Wu, J. H. McDermott, et al. Ambient sound provides supervision for visual learning. In *Proceedings of the European Conference on Computer Vision*, 2016.

[148] A. Paszke, S. Gross, S. Chintala, et al. Automatic differentiation in pytorch. 2017.

[149] Nicolas Perez-Nieves, Vincent CH Leung, Pier Luigi Dragotti, and Dan FM Goodman. Neural heterogeneity promotes robust learning. *Nature Communications*, 12(1):1–9, 2021.

[150] Roman Pogodin and Peter Latham. Kernelized information bottleneck leads to biologically plausible 3-factor hebbian learning in deep networks. *Advances in Neural Information Processing Systems*, 33:7296–7307, 2020.

[151] Jonathan D Power and Bradley L Schlaggar. Neural plasticity across the lifespan. *Wiley Interdisciplinary Reviews: Developmental Biology*, 6(1):e216, 2017.

[152] Alison R Preston and Howard Eichenbaum. Interplay of hippocampus and prefrontal cortex in memory. *Current biology*, 23(17):R764–R773, 2013.

[153] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[154] J. Ramaswamy. What makes the sound?: A dual-modality interacting network for audio-visual event localization. In *IEEE International Conference on Acoustics Speech and Signal Processing*, 2020.

[155] Rajesh PN Rao and Dana H Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1):79–87, 1999.

[156] Jiawei Ren, Cunjun Yu, Xiao Ma, Haiyu Zhao, Shuai Yi, et al. Balanced meta-softmax for long-tailed visual recognition. *Advances in Neural Information Processing Systems*, 33:4175–4186, 2020.

[157] Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. *arXiv preprint arXiv:2010.04592*, 2020.

[158] Robert Rosenbaum. On the relationship between predictive coding and backpropagation. *arXiv preprint arXiv:2106.13082*, 2021.

[159] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986.

[160] O. Russakovsky, J. Deng, H. Su, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 2015.

[161] Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In *Advances in Neural Information Processing Systems*, 2016.

[162] Tommaso Salvatori, Yuhang Song, Yujian Hong, Lei Sha, Simon Frieder, Zhenghua Xu, Rafal Bogacz, and Thomas Lukasiewicz. Associative memories via predictive coding. *Advances in Neural Information Processing Systems*, 34, 2021.

[163] Tommaso Salvatori, Yuhang Song, Beren Millidge, Zhenghua Xu, Lei Sha, Cornelius Emde, Rafal Bogacz, and Thomas Lukasiewicz. Incremental predictive coding: A parallel and fully automatic learning algorithm. *arXiv preprint arXiv:2212.00720*, 2022.

[164] Dvir Samuel, Yuval Atzmon, and Gal Chechik. From generalized zero-shot learning to long-tail with class descriptors. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 286–295, 2021.

[165] Joan Serra, Didac Suris, Marius Miron, and Alexandros Karatzoglou. Overcoming catastrophic forgetting with hard attention to the task. In *International Conference on Machine Learning*, pages 4548–4557. Proceedings of Machine Learning Research, 2018.

[166] Li Shen, Zhouchen Lin, and Qingming Huang. Relay backpropagation for effective learning of deep convolutional neural networks. In *Proceedings of the European Conference on Computer Vision*, 2016.

[167] S. Shimojo and L. Shams. Sensory modalities are not separate modalities: plasticity and interactions. *Current opinion in neurobiology*, 2001.

[168] Ali Seyed Shirkhorshidi, Saeed Aghabozorgi, and Teh Ying Wah. A comparison study on similarity and dissimilarity measures in clustering continuous data. *PloS one*, 2015.

[169] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

[170] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[171] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in Neural Information Processing Systems*, 30, 2017.

[172] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in Neural information processing systems*, 33:596–608, 2020.

[173] Ghada Sokar, Decebal Constantin Mocanu, and Mykola Pechenizkiy. Addressing the stability-plasticity dilemma via knowledge-aware continual learning. *arXiv preprint arXiv:2110.05329*, 2021.

[174] Lee Susman, Naama Brenner, and Omri Barak. Stable memory with unstable synapses. *Nature Communications*, 10(1):1–9, 2019.

[175] Yoshinori Suzuki, Hideaki Ikeda, Takuya Miyamoto, Hiroyoshi Miyakawa, Yoichi Seki, Toru Aonishi, and Takako Morimoto. Noise-robust recognition of wide-field motion direction and the underlying neural mechanisms in drosophila melanogaster. *Scientific Reports*, 5(1):1–12, 2015.

[176] Saeid A Taghanaki, Kristy Choi, Amir Hosein Khasahmadi, and Anirudh Goyal. Robust representation learning via perceptual similarity metrics. In *International Conference on Machine Learning*, 2021.

[177] Kai Sheng Tai, Peter Bailis, and Gregory Valiant. Sinkhorn label allocation: Semi-supervised classification via annealed self-training. In *International Conference on Machine Learning*, 2021.

[178] Anne E Takesian and Takao K Hensch. Balancing plasticity/stability across brain development. *Progress in Brain Research*, 207:3–34, 2013.

[179] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. Proceedings of Machine Learning Research, 2019.

[180] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems*, 2017.

[181] Qi Tian, Kun Kuang, Kelu Jiang, Fei Wu, and Yisen Wang. Analysis and applications of class-wise robustness in adversarial training. *arXiv preprint arXiv:2105.14240*, 2021.

[182] Y. Tian, J. Shi, B. Li, et al. Audio-visual event localization in unconstrained videos. In *Proceedings of the European Conference on Computer Vision*, 2018.

[183] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 2008.

[184] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. Scan: Learning to classify images without labels. In *Proceedings of the European Conference on Computer Vision*, 2020.

[185] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[186] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

[187] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, and Léon Bottou. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 2010.

[188] Athanasios Voulodimos, Nikolaos Doulamis, Anastasios Doulamis, and Eftychios Protopapadakis. Deep learning for computer vision: A brief review. *Computational Intelligence and Neuroscience*, 2018, 2018.

[189] Peng Wang, Kai Han, Xiu-Shen Wei, Lei Zhang, and Lei Wang. Contrastive learning based hybrid networks for long-tailed image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[190] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, 2020.

[191] W. Wang, D. Tran, and M. Feiszli. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[192] Xiao Wang, Shaohua Fan, Kun Kuang, Chuan Shi, Jiawei Liu, and Bai Wang. Decorrelated clustering with data selection bias. In *International Joint Conference on Artificial Intelligence*, 2020.

[193] Yulin Wang, Zanlin Ni, Shiji Song, Le Yang, and Gao Huang. Revisiting locally supervised learning: an alternative to end-to-end training. *arXiv preprint arXiv:2101.10832*, 2021.

[194] Susan G Wardle, Jessica Taubert, Lina Teichmann, and Chris I Baker. Rapid and dynamic processing of face pareidolia in the human brain. *Nature Communications*, 11(1):1–14, 2020.

[195] Chen Wei, Kihyuk Sohn, Clayton Mellina, Alan Yuille, and Fan Yang. Crest: A class-rebalancing self-training framework for imbalanced semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[196] Haiguang Wen, Kuan Han, Junxing Shi, Yizhen Zhang, Eugenio Culurciello, and Zhongming Liu. Deep predictive coding network for object recognition. In *International Conference on Machine Learning*, pages 5266–5275. Proceedings of Machine Learning Research, 2018.

[197] James CR Whittington and Rafal Bogacz. An approximation of the error backpropagation algorithm in a predictive coding network with local hebbian synaptic plasticity. *Neural Computation*, 29(5):1229–1262, 2017.

[198] Sunghyeon Woo, Jeongwoo Park, Jiwoo Hong, and Dongsuk Jeon. Activation sharing with asymmetric paths solves weight transport problem without bidirectional connection. *Advances in Neural Information Processing Systems*, 34, 2021.

[199] J. Wu, Y. Yu, C. Huang, and K. Yu. Deep multiple instance learning for image classification and auto-annotation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2015.

[200] Jianlong Wu, Keyu Long, Fei Wang, Chen Qian, Cheng Li, Zhouchen Lin, and Hongbin Zha. Deep comprehensive correlation mining for image clustering. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019.

[201] Tong Wu, Qingqiu Huang, Ziwei Liu, Yu Wang, and Dahua Lin. Distribution-

balanced loss for multi-label classification in long-tailed datasets. In *Proceedings of the European Conference on Computer Vision*, 2020.

[202] Y. Wu, L. Zhu, Y. Yan, and Y. Yang. Dual attention matching for audio-visual event localization. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019.

[203] Yujie Wu, Rong Zhao, Jun Zhu, Feng Chen, Mingkun Xu, Guoqi Li, Sen Song, Lei Deng, Guanrui Wang, Hao Zheng, et al. Brain-inspired global-local learning incorporated with neuromorphic computing. *Nature Communications*, 13(1):1–14, 2022.

[204] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[205] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

[206] Enze Xie, Jian Ding, Wenhai Wang, Xiaohang Zhan, Hang Xu, Peize Sun, Zhenguo Li, and Ping Luo. Detco: Unsupervised contrastive learning for object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, 2021.

[207] Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *International Conference on Machine Learning*, 2016.

[208] Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[209] H. Xu, R. Zeng, Q. Wu, et al. Cross-modal relation-aware networks for audio-visual event localization. In *ACM International Conference on Multimedia*, 2020.

[210] Yaoda Xu and Maryam Vaziri-Pashkam. Limits to visual representational correspondence between convolutional neural networks and the human brain. *Nature Communications*, 12(1):1–16, 2021.

[211] Yi Xu, Lei Shang, Jinxing Ye, Qi Qian, Yu-Feng Li, Baigui Sun, Hao Li, and Rong Jin. Dash: Semi-supervised learning with dynamic thresholding. In *International Conference on Machine Learning*, 2021.

[212] H. Xuan, Z. Zhang, S. Chen, et al. Cross-modal attention network for temporal inconsistent audio-visual event localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

[213] Guang Yang, Feng Pan, and Wen-Biao Gan. Stably maintained dendritic spines are associated with lifelong memories. *Nature*, 462(7275):920–924, 2009.

[214] Jianwei Yang, Devi Parikh, and Dhruv Batra. Joint unsupervised learning of deep representations and image clusters. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016.

[215] Shuangming Yang, Tian Gao, Jiang Wang, Bin Deng, Mostafa Rahimi Azghadi, Tao Lei, and Bernabe Linares-Barranco. Sam: A unified self-adaptive multi-compartmental spiking neuron model for learning with working memory. *Frontiers in Neuroscience*, 16, 2022.

[216] Shuangming Yang, Bernabe Linares-Barranco, and Badong Chen. Heterogeneous ensemble-based spike-driven few-shot online learning. *Frontiers in Neuroscience*, 16, 2022.

[217] Shuangming Yang, Jiangtong Tan, and Badong Chen. Robust spike-based continual meta-learning improved by restricted minimum error entropy criterion. *Entropy*, 24(4):455, 2022.

[218] Yuzhe Yang and Zhi Xu. Rethinking the value of labels for improving class-imbalanced learning. In *Advances in Neural Information Processing Systems*, 2020.

[219] Pierre Yger, Marcel Stimberg, and Romain Brette. Fast learning with weak synaptic plasticity. *Journal of Neuroscience*, 35(39):13351–13362, 2015.

[220] Xi Yin, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, and Manmohan Chandraker. Feature transfer learning for face recognition with under-represented data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[221] Jason Yoo and Frank Wood. Bayespcn: A continually learnable predictive coding associative memory. *arXiv preprint arXiv:2205.09930*, 2022.

[222] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *British Machine Vision Conference*, 2016.

[223] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Proceedings of the European Conference on Computer Vision*, 2014.

[224] Matthew D Zeiler, Dilip Krishnan, Graham W Taylor, and Rob Fergus. Deconvolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2010.

[225] Lihi Zelnik-manor and Pietro Perona. Self-tuning spectral clustering. In *Advances in Neural Information Processing Systems*, 2005.

[226] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[227] Z. Zhang, C. Lan, W. Zeng, et al. Relation-aware global attention for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[228] Nanxuan Zhao, Zhirong Wu, Rynson WH Lau, and Stephen Lin. What makes instance discrimination good for transfer learning? *arXiv preprint arXiv:2006.06606*, 2020.

[229] Huasong Zhong, Chong Chen, Zhongming Jin, and Xian-Sheng Hua. Deep robust clustering by contrastive learning. *arXiv preprint arXiv:2008.03030*, 2020.

[230] Dengyong Zhou, Olivier Bousquet, Thomas Lal, Jason Weston, and Bernhard Schölkopf. Learning with local and global consistency. *Advances in Neural Information Processing Systems*, 2003.

[231] Jinxing Zhou, Liang Zheng, Yiran Zhong, Shijie Hao, and Meng Wang. Positive sample propagation along the audio-visual event line. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[232] Zhenglong Zhou and Chaz Firestone. Humans can decipher adversarial images. *Nature Communications*, 10(1):1–9, 2019.

[233] Zhi-Hua Zhou. A brief introduction to weakly supervised learning. *National Science Review*, 5(1):44–53, 2018.

[234] H. Zhu, H. Huang, Y. Li, A. Zheng, and R.S He. Arbitrary talking face generation via attentional audio-visual coherence learning. In *International Joint Conference on Artificial Intelligence*, volume 4, 2020.

초 록

딥러닝을 통한 인공신경망 학습은 지난 10년간 눈부신 발전을 이루어 왔으며, 이는 크게 대용량 데이터, 컴퓨팅 하드웨어, 및 학습 알고리즘에 의해 이루어졌다. 데이터에 정확한 라벨이 할당된 양질의 대용량 데이터는 사람의 역량을 뛰어넘는 인공신경망 모델 학습을 가능하게 했다. 하지만 현실 세계에서 수집된 데이터는 데이터의 일부 또는 전체에 라벨이 존재하지 않거나, 라벨이 존재하더라도 대략적인 라벨이 주어지는 등 불완전한 데이터 상황이 빈번하게 발생한다. 또한 딥러닝의 성공을 가져온 요소는 오류역전파 알고리즘을 통한 효과적인 학습이다. 오류역전파를 통해 뇌가 수행하는 인식 활동 자체를 잘 모사할 수 있으나, 학습 알고리즘 측면에서 이는 생물학적으로 타당하지 못한 한계점을 갖는다. 본 학위논문에서는 다양한 불완전한 데이터 조건에서 효과적인 인식 접근 방법을 제안하고, 생물학적으로 타당한 학습 알고리즘이 불완전한 데이터 학습에 어떤 영향을 미치는지에 대한 연구 결과물을 포함한다.

첫 번째 연구는 데이터 일부의 라벨만 존재하는 준지도학습의 인식 문제에 대한 접근이다. 현실 세계에서 수집된 데이터는 클래스별 데이터 분포가 고르지 않을 수 있으며, 수집된 데이터의 라벨이 존재하지 않는 속성을 갖는다. 이 문제를 해결하고자 우리는 준지도학습 방법에 기반한 클래스 불균형 문제를 해결하기 위한 방법론을 제안한다. 준지도학습 상황의 클래스 불균형 문제는 상대적으로 수가 많은 클래스로의 편향을 발생시키는 경향을 보인다. 우리는 기존의 준지도학습 알고리즘이 이러한 경향을 발생시킴을 관찰하고, 불균형한 클래스 분포를 상대적으로 완화할 수 있는 마스킹 전략기반 목적함수를 도입한다. 이를 기존의 준지도학습 알고리즘의 목적함수와 함께 학습함으로써 클래스 불균형 문제를 암시적으로 해결하였다.

두 번째 연구는 데이터에 대략적인 라벨이 존재하는 약지도학습의 인식 문제에 대한 접근이다. 사람의 인식은 다양한 감각기관의 정보를 종합하여 이루어지며, 이는 기계학습에서 멀티모달 인식 문제로 통용된다. 우리는 시각 및 청각 정보가 결합한 비디오의 이벤트 식별 문제에서 초 단위의 정확한 라벨 대신 비디오 단위의 약한 라벨이 주어진 상황의 인식 문제를 해결하기 위한 방법론을 제안한다. 전문적으로 촬영되지 않은 비디오는 시각 정보와 청각 정보 사이의 시맨틱이 불일치하는 문제를 가지며 이는 특히 이벤트가 전환되는 경계에서 빈번하게 일어난다. 우리는 특징 공간상에서 오디오 및 비디오 모달리티 각각의 시간적 속성 정보를 강화하는 방법과 서로 다른 모달리티 사이의 시맨틱 정보를 일치하는 방법론을 제안한다. 이는 두 모달리티 정보를 효율적으로 융합하여 비디오 이벤트 식별 문제를 효과적으로 해결한다.

세 번째 연구는 데이터에 라벨이 존재하지 않는 비지도학습의 인식 문제에 대한 접근이다. 우리는 대표적인 비지도학습 문제인 딥 클러스터링을 대조학습을 활용하여 해결한다. 일반적으로 대조학습을 위해서는 동일한 데이터 인스턴스에 대해 서로 다른 데이터 증강을 통해 두 개의 새로운 데이터를 생성한다. 이 때 동일한 인스턴스에서 만들어진 데이터들 사이의 거리를 가깝게하고, 서로 다른 인스턴스들에서 만들어진 데이터들 사이의 거리를 멀게 학습한다. 이러한 학습은 라벨 없이도 학습된 특징들의 차별성을 강화하는 속성을 갖는다. 앞선 상황에서 대조학습 기반 학습에서는 클래스 충돌 문제가 반드시 발생한다. 이는 동일한 클래스이지만 대조학습의 목적함수에 의해 서로 다른 클래스로 인식되어 거리가 멀게 학습되는 문제이다. 우리는 이러한 현생을 억제하기 위한 목적함수를 도입하고, 모델의 중간에서 생성되는 특징 또한 함께 대조학습에 사용하는 방법을 제안함으로써 딥 클러스터링 성능을 효과적으로 개선하였다.

마지막 방법은 사람의 뇌에 기반한 학습 알고리즘을 통한 접근이다. 현재의 딥러닝의 성공을 가져온 오류역전파에 의한 학습은 정작 뇌를 제대로 모사하지 못하는 한계점을 갖고 있다. 우리는 생물학적으로 더 타당한 학습 알고리즘이라면, 사람이 잘 수행하는 문제들의 성능을 개선할 수 있다는 가정에 기반한다. 우리는 이러한 가정 하에 생물학적으로 타당한 학습 알고리즘을 연속학습, 불균형 데이터 및 적

의 수의 학습에 적용하고, 기존의 오류역전파를 통한 학습과의 비교를 수행한다. 우리는 위의 결과를 사람의 뇌가 갖고 있는 신경가소성에 기반한 분석과, 해마와 전두엽의 상호작용 관점에서 해석하며 생물학적으로 타당한 학습의 잠재력에 대한 탐구를 수행한다.

본 학위논문을 통해 불완전한 데이터 환경의 인식 문제를 해결하는 방법론을 제시하고, 뇌 모사 학습을 통한 인식 문제 성능 개선에 대해 탐색하였다. 불완전한 데이터 인식 문제는 현실 세계에서 접하기 쉬운 문제 중 하나이며, 이를 효과적으로 해결하는 것은 데이터 생성의 시간, 비용 효율적인 측면에서 중요하다. 또한 사람의 뇌에서 수행되는 학습과 유사한 학습을 수행하는 것은 궁극적인 인공지능에 도달하기 위한 방법으로써 많은 가능성을 품고 있다.

주요어: 딥러닝, 기계학습, 불완전 데이터, 제한된 데이터 인식, 뇌 모사 학습
학번: 2016-20954