공학석사 학위논문

Grounding Visio-Linguistic Information with Fast and Slow Neural Networks

빠른 신경망과 느린 신경망을 통한 시각-언어 정보 표시

2023년 8월

서울대학교 대학원

협동과정 인지과학전공

설 한 울

Grounding Visio-Linguistic Information with Fast and Slow Neural Networks

빠른 신경망과 느린 신경망을 통한 시각-언어 정보 표시

지도교수 장 병 탁

이 논문을 공학석사 학위논문으로 제출함

2023 년 8 월

서울대학교 대학원

협동과정 인지과학전공

설 한 울

설 한 울의 공학석사 학위논문을 인준함

2023 년 7 월

위 원 장 고 성 룡 (인)

부위원장 장 병 탁 (인)

위 원 박 형 생 (인)

Abstract

Grounding Visio-Linguistic Information with Fast and Slow Neural Networks

Han-wool Sul
Interdisciplinary Program in Cognitive Science
The Graduate School
Seoul National University

The remarkable progress witnessed in the field of Deep Learning has been spurred by the relentless pursuit of emulating the intricate cognitive abilities exhibited by the human brain. This relentless pursuit has led to extraordinary achievements across a myriad of domains, showcasing the exceptional prowess of AI systems. Although significant endeavors have been directed towards refining uni-modal tasks such as natural language processing, computer vision, and speech recognition, it is imperative to acknowledge the fundamental role played by the fusion of multiple modalities in the intricate fabric of human cognition.

The Omnilabel benchmark Schulter et al. (2023) presents a unique and demanding task that necessitates the localization of referenced objects based on textual descriptions. Unlike traditional approaches that rely on predefined and constrained label spaces, the Omnilabel benchmark embraces a vast array of object description variations, spanning from succinct category names to intricate and detailed textual depictions. Furthermore, a distinctive characteristic of this benchmark lies in its allowance for descriptions that can refer to zero,

one, or multiple objects, thereby introducing the intricacy of handling negative pairs wherein the description fails to align with any specific object depicted in the given image. Most studies in the field have primarily focused on handling positive pairs of data, particularly in the context of referring expressions. Consequently, existing models face challenges when confronted with negative pairs in datasets.

Drawing inspiration from the cognitive framework of loss aversion Kahneman and Tversky (1979), which posits that humans tend to weigh losses more heavily than equivalent gains. Loss aversion theory suggests that individuals are inclined to substitute a difficult question with an easier alternative. As a result, the human brain has evolved to operate using two distinct systems Kahneman (2011): Fast Thinking (System 1) and Slow Thinking (System 2).

Taking cues from biological inspiration, we propose the adoption of Fast Neural Network (FNN) as an analog to System 1 and Slow Neural Network (SNN) as an analog to System 2. FNN is trained to determine the positivity or negativity of input data. Following the classification of positive and negative pairs, SNN selectively processes the filtered data obtained from FNN. Our approach demonstrates the efficacy and efficiency of SNN, as it leverages the filtered data without requiring additional training. This strategy proves to be faster and computationally more economical than employing SNN for the inference of the entire dataset.

Keywords: Visual grounding, multi-modal, Referring Expression, Contrastive Learning

Student Number: 2021-29149

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The remarkable progress in Deep Learning has been driven by endeavors to emulate the cognitive capabilities of the human brain. This advancement has empowered AI systems to excel at various tasks, demonstrating impressive performance across domains. While significant attention has been devoted to advancing uni-modal tasks such as natural language processing, computer vision, and speech recognition, it is crucial to acknowledge that human cognition operates through multi-modal processing.

In the realm of human perception, even during activities as seemingly uni-modal as reading a book, the brain dynamically integrates textual information and transforms it into vivid visual and auditory experiences through the power of imagination. Consequently, these mental stimuli manifest themselves within the auditory and visual cortices. Recognizing the intricacies of human cognition, it becomes increasingly evident that addressing multi-modal data is of paramount importance.

With the ongoing advancements in uni-modal deep learning models, there has been a surge of interest in multi-modal models. Researchers have been exploring the potential of building upon the successes of uni-modal models to develop powerful multi-modal architectures. One notable example is the development of GPT-4OpenAI (2023) , which extends the capabilities of models like

ChatGPTOpenAI (2022) to encompass visual information. As a result, these multi-modal models have the capacity to tackle a broader range of tasks, exhibiting enhanced performance compared to their uni-modal counterparts.

Furthermore, significant strides have been made in the field of image generation by leveraging the synergy of textual and visual information. Models such as DALL · E2 Ramesh et al. (2022) and Stable Diffusion Rombach et al. (2022) have demonstrated the ability to generate images based on textual descriptions. What sets these models apart from single diffusion models is their capacity to generate images with a higher level of fidelity and detail, aligning more closely with human intentions and creative expression.

By embracing the challenges and opportunities presented by multi-modal data, AI systems can effectively leverage the rich and interconnected nature of various modalities, unlocking the potential for more comprehensive and nuanced understanding. Advancements in multi-modal deep learning enable models to seamlessly integrate and process information from multiple sources, leading to improved performance across a broad range of tasks that require a holistic comprehension of the world.

Our work focuses on the visual grounding task within the context of multi-modal vision and language processing. Understanding the intricate semantics of the world is a crucial aspect of human visual perception. In our research, we specifically address the challenges posed by the Omnilabel benchmark Schulter et al. (2023), which differs significantly from traditional visual grounding tasks. The primary objective of the Omnilabel benchmark is to thoroughly evaluate models' ability to comprehend complex and unrestricted textual descriptions of objects while accurately localizing the corresponding instances.

Unlike conventional approaches that rely on predefined label spaces, the Omnilabel benchmark embraces a wide range of object description variations,

encompassing both plain category names and intricate textual descriptions. Another distinctive aspect of this benchmark is that a given description can refer to zero to multiple objects, which introduces the challenge of handling negative pairs of descriptions and images where the description does not align with any specific object in the provided image. This nuanced and comprehensive evaluation setup enables us to assess the robustness and versatility of vision and language models in understanding and grounding complex textual descriptions to visual instances.

Omnilabel shares certain characteristics with referring expression tasks, but it also exhibits significant differences. Firstly, unlike traditional referring expression datasetsKazemzadeh et al. (2014)Yu et al. (2016), Omnilabel includes negative pairs, which significantly impact the performance of models. This distinguishes it from referring expression datasets that solely assume positive pairs. The presence of negative samples in Omnilabel introduces additional challenges and complexity.

Secondly, the descriptions in Omnilabel datasets have the capability to refer to multiple objects simultaneously. While datasets such as PhraseCut Wu et al. (2020) also involve multiple objects associated with a single description, it is important to note that the descriptions in PhraseCut primarily consist of simple templated phrases. In contrast, Omnilabel incorporates more diverse and complex textual descriptions. Other datasets like RefCOCO/+/g Mao et al. (2016) Yu et al. (2016) typically involve descriptions that refer to a single object per expression.

These differences in the negative pair inclusion and the ability of descriptions to refer to multiple objects make Omnilabel a unique benchmark that requires models to handle both the challenges of negative samples and the complexities of multi-object references in textual descriptions.

Research conducted in the fields of human brain science, psychology, and economics has shed light on the differential responses individuals exhibit towards gains and losses. Notably, findings have revealed a heightened sensitivity to losses compared to equivalent gains, a phenomenon commonly referred to as loss aversion Kahneman and Tversky (1979). Building upon this theory, scholars have explored its implications for human cognitive processes and decision-making Kahneman (2011). Their analysis posits that individuals tend to gravitate towards substituting complex questions with simpler alternatives, a cognitive strategy aimed at alleviating cognitive load. To expound further, they delineate the human thinking system into two distinct modes: fast thinking and slow thinking. Fast thinking is primarily observed in rapid, automatic, frequent, and stereotypic cognitive events, whereas slow thinking encompasses effortful, infrequent, logical, calculating, and conscious cognitive processes. This dichotomy in thinking patterns is attributed to the evolutionary adaptations of human cognition, which prioritize energy conservation and temporal efficiency through the utilization of fast thinking when facing uncertainty.

Drawing inspiration from this biological processes, we have proposed a novel approach for tackling the Omnilabel task by introducing the Fast Neural Network (FNN) and Slow Neural Network (SNN). Our methodology leverages the inherent strengths of these two distinct networks in a two-stage manner. Initially, the FNN swiftly processes the data pairs and classifies them as either positive or negative instances. Subsequently, the positive data pairs are forwarded to the SNN, which operates on a more deliberate timescale. The SNN effectively integrates the image and description components of the positive pairs and generates corresponding bounding boxes. By adopting this two-stage inference framework, we anticipate significant enhancements in both model performance and inference time. Consequently, our approach exhibits superior results

compared to baseline approaches and other competitive models, while achieving remarkable efficiency gains.

Chapter 2

Related Works

2.1 Datasets for visual grounding

A multitude of datasets exists pertaining to visual grounding tasks, sharing a commonality in their provision of both images and corresponding captions for object identification, alongside ground truth bounding boxes.

Referring Expression Referring Expression datasets, similar to other visual grounding datasets, exhibit the inclusion of image-text pairs accompanied by ground truth bounding boxes. However, what sets them apart is the emphasis on captions referring to multiple objects within an image. Among these datasets, the prominent ones include RefCOCO/+/g Mao et al. (2016) Yu et al. (2016) and PhraseCut Wu et al. (2020).

The RefCOCO/+ datasets were curated through the employment of the ReferItGame Kazemzadeh et al. (2014) methodology, which employs a game-like approach to crowd-source natural language referring expressions. In this game, two players assume distinct roles. Player 1 observes an object within a given image and precisely generates a referring expression for the identified object. Subsequently, Player 2 is tasked with localizing the exact object referred to by the given expression. On the other hand, the RefCOCOg dataset follows a similar data collection methodology as ReferItGame, but in a non-interactive

setting. A notable distinction between RefCOCO and RefCOCO+ lies in the usage of "taboo" words. RefCOCO allows unrestricted generation of referring expressions, while RefCOCO+ imposes restrictions on the expression-generating player, prohibiting the usage of localization words. This added constraint renders the expressions more intricate and poses a greater challenge for models to effectively address the task. In our study, we leverage the RefCOCO/+/g datasets to fine-tune our network in the image-text matching task, enabling the classification of negative and positive pairs within the Omnilabel datasets, thus facilitating the subsequent processing of only positive pairs.

Another notable dataset for referring expression is PhraseCut, which is constructed based on the Visual Genome Krishna et al. (2017) dataset. The primary objective of models in this task is to generate segmentation masks for objects referenced by provided phrases. The phrases in PhraseCut are formulated according to specific criteria, often comprising object categories along with their attributes or object categories in relation to other objects. One key distinction from RefCOCO/+/g is that a single phrase in PhraseCut can refer to multiple objects simultaneously. This feature enables a more complex and diverse range of expressions, thereby posing a greater challenge for models in accurately identifying and segmenting the relevant objects.

Phrase Grounding In the context of referring expression tasks, the primary objective for models is to accurately localize the referenced object within an image. However, in contrast, other visual grounding datasets like Flickr30K Plummer et al. (2015) or Visual Genomes Krishna et al. (2017) have a broader focus, aiming to localize multiple objects mentioned in the given textual captions. These datasets serve as valuable resources for fine-grained pre-training of vision language models, providing a comprehensive understanding of object

localization in various contexts and facilitating the development of more robust and versatile vision language models.

2.2 Transformers

The Transformers architecture Vaswani et al. (2017) stands as one of the most pervasive and versatile architectures in modern deep learning. Its widespread adoption spans across diverse domains, including natural language processing, computer vision, speech recognition, reinforcement learning, and various multimodal applications. At its core, the Transformers architecture leverages self-attention mechanisms, which were introduced to extract richer and more informative representations by assigning higher weights to target representations based on their relative importance. This attention-based approach has proven to be highly effective in capturing complex dependencies and relationships within data, enabling Transformers to achieve remarkable performance across a wide range of tasks and domains.

Subsequent to the introduction of the Transformers architecture, researchers further refined the architecture by dividing it into distinct encoder and decoder components. The encoder architecture, inspiring BERT Devlin et al. (2018), underwent training using masked language modeling and next sentence prediction techniques. These Transformers encoder-based architectures have exhibited remarkable performance in various classification-oriented natural language processing (NLP) tasks. These tasks include entity recognition, textual entailment, coreference resolution, and more. The encoder-focused Transformers models have demonstrated their efficacy in capturing and representing contextual information within text, leading to state-of-the-art results in a wide array of NLP applications.

The decoder architecture, inspiring GPT (Generative Pre-trained Transformer) Radford et al. (2018), is specifically designed for language generation tasks. It leverages masked self-attention mechanisms to enable auto-regressive generation, where each token is generated conditioned on the previously generated tokens. GPT models have gained significant recognition for their impressive generation capabilities and their remarkable parameter size. The advancements in the GPT series include the introduction of GPT2 Radford et al. (2019) and GPT3 Brown et al. (2020). GPT3, in particular, underwent a transformative process called reinforcement learning with human feedback (RLHF), resulting in the development of InstructGPT Ouyang et al. (2022), colloquially referred to as GPT3.5. Subsequently, the dialog version of GPT3.5 emerged as ChatGPT OpenAI (2022), which further extended the capabilities of the model to engage in interactive conversations.

The success of the encoder and decoder Transformer architecture in natural language processing tasks has motivated researchers to explore its application in other domains. In the field of computer vision, researchers have been exploring the application of Transformers by transforming continuous data into a discrete format. This approach involves breaking down continuous visual data, into smaller, manageable components. One notable adaptation is the Vision Transformer (ViT) Dosovitskiy et al. (2021), which partitions the input image into small patches, typically 16x16, and feeds them into the Transformer architecture. This approach has demonstrated remarkable performance in image recognition tasks, showcasing the effectiveness of Transformers in the visual domain. Another advancement in this direction is the Swin Transformer Liu et al. (2021), which introduces a shifted window mechanism for dividing images into patches. By incorporating this innovative strategy, the Swin Transformer enhances the efficiency and effectiveness of vision Transformers even further.

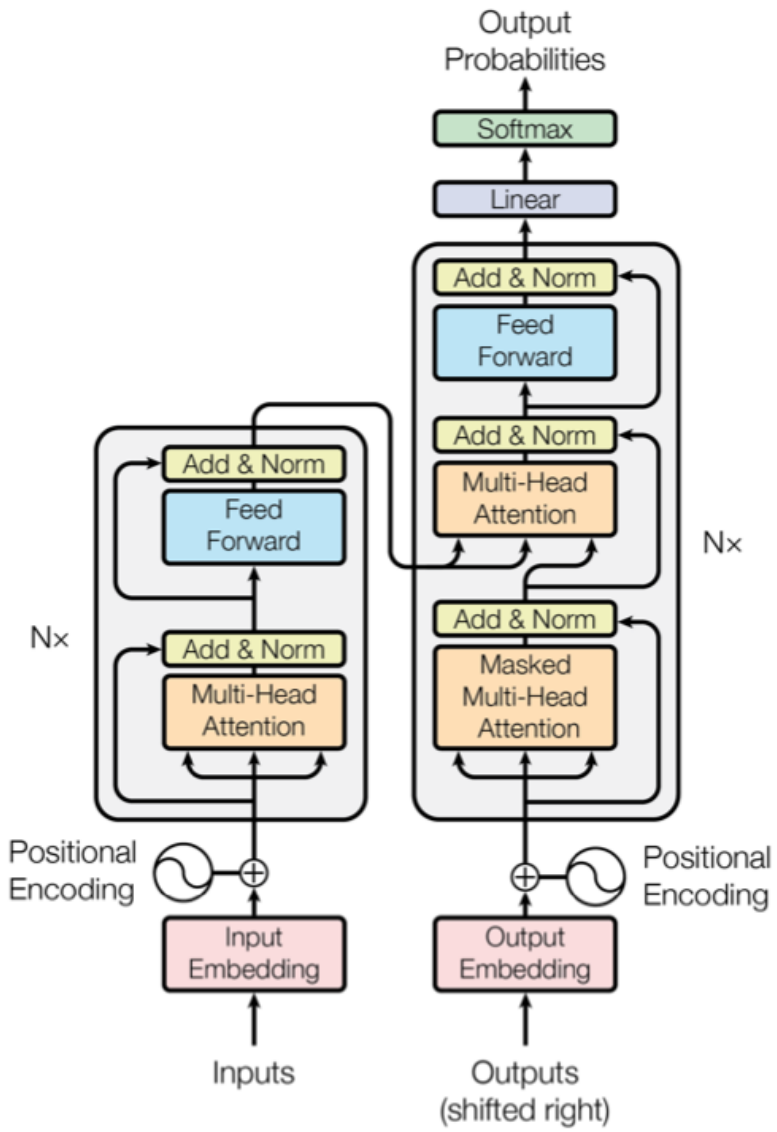

Figure 2.1: The Transformer Architecture. left encoders and right decoders, either in isolation or in combination, has been developed as a prominent architectural design within the field of deep neural networks, stemming from extensive research and serving as a fundamental framework. Figure is from (Vaswani et al., 2017).

2.3 Vision-Language Pretraining

Since the introduction of the Transformer architecture, the development of Vision-Language Pretraining (VLP) models has witnessed remarkable progress. These models can be classified into two categories based on their methodologies: single stream models and dual stream models.

Single stream models, such as VisualBERTLi et al. (2019) and OFAWang et al. (2022), adopt the BERT architecture and input both language and image tokens combined into the model. This approach allows the model to process and integrate information from both modalities jointly.

In contrast, dual stream models like ViLBERTLu et al. (2019), GLIPZhang et al. (2022), and FIBERDou et al. (2022) employ separate encoders for language and image tokens. This separation enables independent processing of language and image inputs. To facilitate cross-modal information fusion, these models either fuse the information in one of the encoders or employ specialized architectures for cross-modal integration.

Dual stream models leverage modified cross-attention mechanisms in the Transformer decoder to capture cross-modal dependencies effectively. Additionally, contrastive learning techniques are often employed to facilitate learning meaningful representations by aligning related language and image inputs.

By combining the power of Transformers with these different VLP model architectures, researchers have made significant strides in addressing vision-language tasks and achieving impressive performance on a range of benchmarks. These models have greatly advanced our ability to understand and interpret visual and textual information in a unified framework.

In the current landscape of vision-language models, dual stream methods have gained significant popularity and are often preferred in state-of-the-art

architectures. The main advantage of dual stream methods lies in their ability to leverage the pretrained parameters of single modal models such as BERT Devlin et al. (2018), RoBERTa Liu et al. (2019), and Swin Transformers Liu et al. (2021). By incorporating these pretrained models, dual stream methods can effectively capture both language and visual information.

A key aspect contributing to the success of dual stream methods is the use of contrastive learning techniques. Contrastive learning, employed in notable models like CLIP Radford et al. (2021), represents a fundamental approach in vision-language pretraining. This technique aims to bring paired data points closer together in the data representation space while maintaining a relative distance between non-paired data points. By doing so, the model learns to differentiate between positive and negative pairs, facilitating the learning of meaningful visual and textual representations.

By leveraging the combination of dual stream architectures and contrastive learning methods, state-of-the-art vision-language models have achieved remarkable performance on a wide range of tasks. These models have demonstrated their effectiveness in capturing intricate relationships between images and text, enabling more sophisticated and comprehensive understanding of multimodal data.

Chapter 3

Visio-Linguistic Grounding, Fast and Slow

In this particular section, we expound upon the intricate Omnilabell Schulter et al. (2023) task and our algorithm, employing two-stage Fast and Slow neural networks, designed to tackle this complex problem.

Omnilabel The Omnilabel task, in essence, entails language-based object detection, encompassing the realms of Referring Expression and open-vocabulary detection. Within this task, The model, denoted as M, adeptly processes both the RGB image I_i and the corresponding description D_i as input, generating predicted bounding boxes B_i based on the provided description D_i. The input image I_i comprises carefully curated images meticulously selected from the Valid and Test sets of prominent datasets such as COCO Lin et al. (2014), Objects-365 Shao et al. (2019), and OpenImages-V5 Kuznetsova et al. (2020). the description D_i and image I_i are painstakingly assembled through the Amazon Mechanical Turk (AMT) contributors, who leverage pre-existing images and ground-truth bounding boxes from the original data annotations.

In Omnilabel benchmark, there are crucial difference with exisiting datasets: firstly, it encompasses far more intricate and nuanced descriptions when compared to conventional object detection and visual grounding datasets. Unlike

conventional object detection datasets, Omnilabel liberates itself from restrictive text assumptions, embracing the vast realm of free-form expressions. Secondly, the descriptions within the Omnilabel dataset span a wide spectrum, ranging from rudimentary categorical labels to remarkably specific depictions, setting it apart from referring expression datasets. Lastly, each description within the Omnilabel dataset possesses the capacity to refer to zero or even multiple objects, presenting a distinctive challenge. In other words, a description within this dataset may not solely pertain to a single object within the paired image; rather, it can encompass a multitude of objects. This poses a formidable obstacle for standard models accustomed to the referring expression task, as they are primarily geared towards scenarios where each description solely corresponds to a single object in the image, thus rendering them unsuitable for the Omnilabel benchmark.

Fast and Slow neural networks In the context of this task framework, we endeavored to address its complexities through the utilization of two neural networks, named the Fast Neural Network (FNN) and the Slow Neural Network (SNN). The FNN serves as the initial processing stage, deftly analyzing the complete pairs of images and descriptions. Furthermore, the SNN is employed as the subsequent stage, where it efficiently processes the positive pairs of the dataset and generates accurate bounding boxes in accordance with the corresponding descriptions.

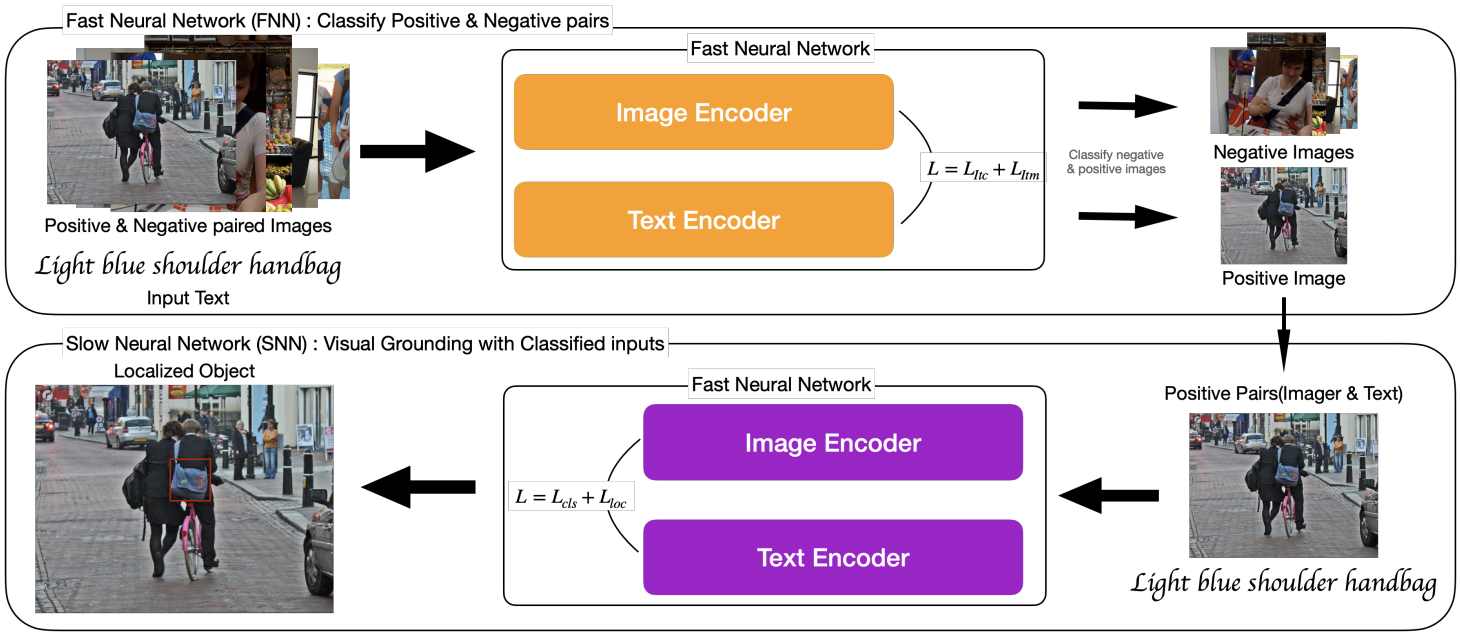

Figure 3.1: The proposed framework comprises two components: the Fast Neural Network (FNN) and the Slow Neural Network (SNN). In the first stage, the FNN efficiently processes a batch of input images paired with their corresponding text descriptions. It performs a binary classification task, determining whether the given image-text pairs are positively matched or not. The FNN acts as a swift decision-maker, quickly filtering out negative pairs and retaining only the positive ones for further analysis. In the second stage, the selected positive pairs are passed to the SNN. The role of the SNN is to perform more detailed and sophisticated processing on the positive pairs, specifically focusing on localizing the objects referred to in the text descriptions. By leveraging the filtered positive pairs, the SNN can concentrate its computational resources on precisely localizing the objects of interest, thereby achieving higher accuracy and more efficient inference.

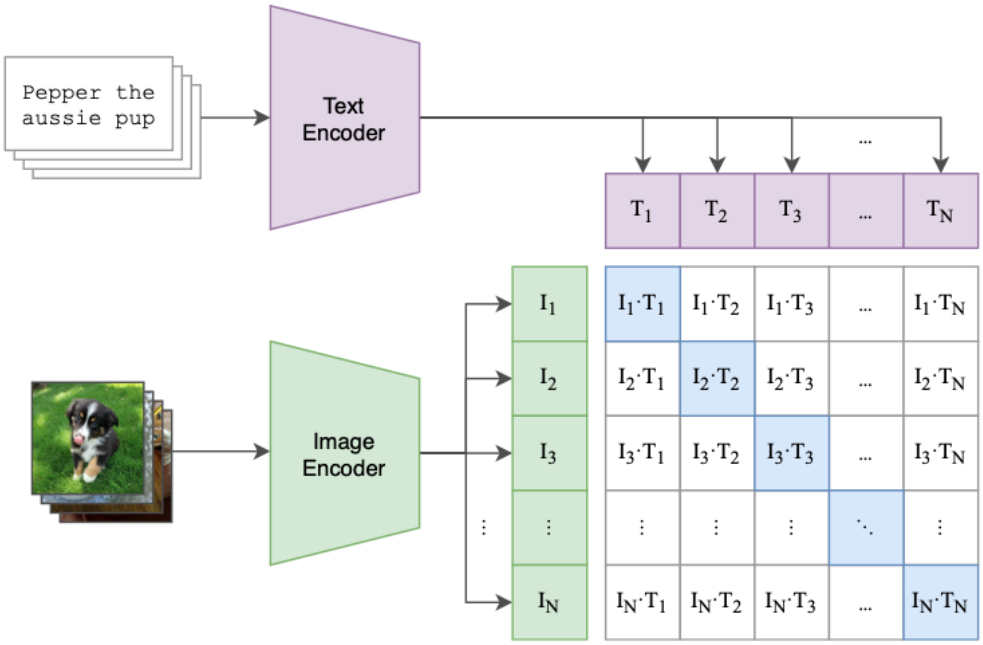

Figure 3.2: Contrastive Learning. This methodology employs the joint embedding derived from the vision encoder and the text encoder, enabling the prediction of positive pairs within a batch of training examples. This figure is from (Radford et al., 2021)

3.1 Fast Neural Networks

Within the framework of the Fast Neural Network (FNN), the BLIP-2 Li et al. (2023) model is employed as the underlying architecture. In BLIP-2, a pre-trained vision transformer model is adapted as an image encoder, with its parameters frozen to ensure stability and efficiency. Two distinct transformer models are employed within BLIP-2: an image transformer and a language transformer. The image transformer interacts with the pre-trained and frozen

image encoder, while the text transformer serves the dual purpose of encoding and decoding textual information.

During the training process, two key components, named as Image-Text Contrastive Learning (ITC) and Image-Text Matching (ITM), are utilized within this architecture:

$$L = L_{Itc} + L_{Itm} \tag{3.1}$$

ITC aims to establish a mutual representation between paired images by enhancing the image-text similarity for positive pairs and diminishing it for negative pairs:

$$L_{Itc} = -log\frac{exp(I_i D_i/\tau)}{\sum_{j=0}^{K} exp(I_i D_j/\tau)} \tag{3.2}$$

I_i and D_i represent the output embeddings acquired from the vision transformer and the language transformer, respectively.Additionally, the hyper-parameter τ governs the influence of the dot product between these embeddings. Furthermore, the variable K signifies the number of pairs calculated during the computation of the contrastive loss.

On the other hand, ITM is employed to align the representations of image and text through binary classification, allowing the model to determine whether a given image-text pair is a match or not:

$$O_{Itm} = Linear(I_i, D_i), \quad L_{Itm} = SoftMax(O_{Itm}) \tag{3.3}$$

The linear layer takes embedding from vision transformer I_i and language transformer D_i as input and produces the output O_{Itm}. By subjecting O_{Itm} to the softmax function, the probability distribution P_{Itm} is obtained. Note that the decision of whether a certain probability corresponds to a match or

an unmatched case is determined by applying a threshold. The BLIP-2 model undergoes a pre-training phase utilizing an extensive corpus of 129 million images, encompassing datasets such as COCO Lin et al. (2014), Visual Genome Krishna et al. (2017), CC3M Sharma et al. (2018), CC12M Changpinyo et al. (2021), SBU Ordonez et al. (2011) and LAION400M Schuhmann et al. (2021). Fine-tuning takes place employing the Refcoco-itm datasets, which consist of 60 million image-text pairs derived from the Refcoco datasets (including Refcoco, Refcoco+, and Refcocog). The Refcoco-itm datasets are intentionally constructed to enable models to perform image-text matching specifically with object descriptions and image pairs, rather than with image captions and image pairs.

3.2 Slow Neural Networks

The Slow Neural Network (SNN) serves as the actual model responsible for generating the intended outcomes. In our implementation, we leverage the GLIPv2 model Zhang et al. (2022), which is one among various fine-grained vision language pretrained models. The GLIPv2 model adopts pretrained transformer models for both the vision encoder and the language encoder. However, in contrast to the BLIP2 model, the transformer models in GLIPv2 are initialized with existing models and subsequently fine-tuned using input data and various loss functions.

The authors of GLIPv2 propose a unified formulation that merges the tasks of visual grounding and object detection. Instead of treating object detection solely as the task of identifying bounding boxes and classifying them based on given labels, GLIPv2 considers the matching of given label-categories as textual inputs. Through this reformulation, the distinct tasks of object detection and

visual grounding are harmoniously unified. To achieve this, GLIPv2 introduces a loss function comprising of both the classification loss L_{cls} and the localization loss L_{loc}. These components collectively facilitate the seamless integration of object detection and visual grounding tasks.

$$L = L_{cls} + L_{loc} \tag{3.4}$$

Following the contrastive alignment loss discussed in Equation 3.2, the two losses, namely the box classification loss L_{cls} and the localization loss L_{loc}, are computed. The box classification loss L_{cls} is calculated using a straightforward linear layer. Mathematically, the expression for L_{cls} can be written as:

$$Emb_I = Enc_I(I), \quad S_{cls} = Emb_I W^T, \quad L_{cls} = loss(s_{cls}; T) \tag{3.5}$$

The feature embeddings, denoted as $Emb \in \mathbf{R}^{N \times d}$, are obtained from the input image through the encoder. The weight matrix W facilitates the box classifier, while T serves as a mechanism for binary matching between regions and classes. Notably, in the context of object detection as phrase grounding, $Emb_T = Enc_T(T)$ assumes a similar role as that of the weight matrix W.

Moving on to the localization loss L_{loc}, it is computed utilizing the concept of smooth L1 loss, employing the coordinates of both the ground truth bounding boxes and the predicted bounding boxes. This calculation ensures a smooth and robust estimation of the localization loss, contributing to the overall optimization process.

$$L_{loc} = \begin{cases} 0.5 \left(Cor_p - Cor_{gt}\right)^2 / \beta, & \text{if } |Cor_p - Cor_{gt}| < \beta \\ |Cor_p - Cor_{gt}| - 0.5 * \beta, & \text{otherwise} \end{cases} \tag{3.6}$$

Cor_p and Cor_{gt}, which represent the predicted and ground truth coordinates, respectively. Additionally, we introduce the parameter β, which serves as a threshold to determine the type of loss to be employed. When the difference between Cor_p and Cor_{gt} falls below the threshold β, the loss function behaves akin to L2 loss. Conversely, if the difference exceeds the threshold, the loss function transitions to L1 loss, which emphasizes robustness against outliers and promotes more localized predictions.

Chapter 4

Experiments

This section provides a concise overview of our experiments on the omnilabel datasets. We introduce the datasets, evaluation metrics, and implementation details in Sections 4.1, 4.2, and 4.3, respectively. Additionally, we present the experimental results, highlighting key findings and outcomes.

4.1 Datasets

We employ a rigorous evaluation benchmark known as Omnilabel omn to assess the performance of our approach. The Omnilabel benchmark, solely designed for evaluation purposes, lacks training data.

The curated image collection within the Omnilabel dataset comprises the validation and test sets of prominent datasets such as MS COCO Lin et al. (2014), Object-365 Shao et al. (2019), and Openimagesv5 Kuznetsova et al. (2020). Within these datasets, images were carefully selected based on specific criteria: (1) Ensuring the presence of at least two instances pertaining to a (super) category, (2) Encompassing a diverse range of sub-categories within each super-category, and (3) Incorporating descriptions that focus on the object's appearance, relations, and actions. However, certain data points were excluded from the dataset: (1) Instances exceeding a count of ten, (2) Boxes with a size smaller than 80 pixels, and (3) Boxes exhibiting an overlap greater than 50%

with other boxes or being flagged as "iscrowd". These rigorous criteria ensure the integrity and reliability of the dataset, enabling precise evaluation of the models.

Regarding the validation data, the fundamental components of the input data points encompass essential elements such as "description" and "image-ids". The description represents the textual input that references the objects within the images, ranging from zero to multiple instances. The image-ids comprise a list of unique identifiers corresponding to the images, which can be either positive or negative in nature. In the case of negative image pairs, they must share the same object (super-) category as the positive pairs.

Turning to the ground-truth data, it encompasses crucial elements such as "image-id", "bbox" (bounding box), and "description-ids". Each ground-truth entry pertains to a unique bounding box within an image. Notably, within a single image, the model's output can potentially refer to the same object using different descriptions. For instance, a specific bounding box within an image could be referred to as a "car" based on the category description, while simultaneously being pointed to using a sentence-based description like "vehicles on the road". In the case of test data, only input data is available, devoid of corresponding ground-truth annotations.

4.2 Evaluation Metrics

The evaluation of model outputs within the Omnilabel benchmark entails the utilization of two essential metrics: Intersection over Union (IoU) and Average Precision (AP). These metrics serve as pivotal indicators for assessing the quality and effectiveness of the models' predictions.

The evaluation process within Omnilabel is stratified into two distinct groups:

object-categories and object-descriptions. Within the object-description group, further stratification is carried out based on the length of the input description. Specifically, the descriptions are categorized into three groups: "short" for descriptions containing less than four words, "middle" for descriptions ranging from four to eight words, and "long" for descriptions exceeding eight words in length. This categorization facilitates a nuanced analysis of the model's performance based on the complexity and length of the provided descriptions. By employing this meticulous stratification, Omnilabel enables a comprehensive evaluation of the model's ability to handle descriptions of varying lengths and intricacies.

$$Intersection\ Over\ Union\ (IoU) = \frac{Area\ of\ Intersection}{Area\ of\ union} \tag{4.1}$$

In the context of object detection tasks, the evaluation metric of Average Precision (AP) assumes a significant role. AP is calculated independently for each category and subsequently averaged to obtain the mean Average Precision (mAP) across all categories. In traditional object detection benchmarks, the predicted bounding boxes are matched with the corresponding ground-truth boxes based on their Intersection over Union (IoU) exceeding a certain threshold.

However, in the case of Omnilabel, some distinctions arise. The initial categorization based solely on object categories is no longer applicable, given the inclusion of object descriptions within the dataset. As a result, the predicted bounding boxes are matched with the respective ground-truth boxes, and the AP is computed accordingly. This modified evaluation process in Omnilabel accounts for the unique characteristics of the dataset, ensuring accurate assessment of the model's performance in both object detection and visual grounding tasks.

4.3 Implementation Details

Our framework is implemented using the PyTorch deep learning library `http://pytorch.org`. Specifically, for the Fast Neural Network (FNN) component, we employ the BLIP2 model `https://github.com/salesforce/LAVIS`. As for the Slow Neural Network (SNN), we utilize the GLIPv2 model `https://github.com/microsoft/GLIP`.

For the FNN component, we experiment with three distinct BLIP2 models: a pretrained model, a model fine-tuned on the COCO dataset, and a model fine-tuned on the Refcoco-itm dataset. Additionally, an ensemble approach is employed, where the predictions of the three models are combined using a threshold of 0.3 on the total sum of probability. This ensemble strategy aims to leverage the strengths of multiple models to improve the overall performance of the system.

$$\begin{cases} \text{False}, & \text{if } (P_r + C_r + R_r) < th \\ \text{True}, & \text{otherwise} \end{cases} \tag{4.2}$$

where P_r, C_r and R_r are Pretrained result, COCO finetuned result, Refcoco finetuned result for each data point, respectively.

4.4 Results

We conducted a comprehensive quantitative evaluation of our proposed method on the Omnilabel benchmark. The evaluation results are presented in Table 4.1, offering valuable insights into the performance of various baseline models. Notably, a substantial performance disparity emerges when comparing the inclusion of negative descriptions in the label space (AP-descr) versus their exclusion (AP-descr-pos). This disparity is particularly pronounced in the case of COCO images, as highlighted by the Omnilabel dataset, as this subset of the dataset exhibits a higher prevalence of negative descriptions relative to the number of images. Moreover, we computed the Average Precision for descriptions of different lengths: short (AP-descr-S), medium (AP-descr-M), and long (AP-descr-L). These metrics provide a comprehensive assessment of our model's efficacy in handling descriptions consisting of less than four words, between four and eight words, and exceeding eight words, respectively.

Our proposed methods have demonstrated remarkable performance in the evaluation metrics, achieving the highest scores in AP-descr for All (+2.2), COCO (+5.0), and OpenImagesv5 (+7.6). Notably, our model outperforms the baseline models, such as GLIP-L, by even larger margins, with improvements of 3.3, 5.4, and 7.6, respectively. Unlike other models, our approach exhibits a smaller performance gap between AP-descr and AP-descr-pos. This indicates that while our fast neural network classification may have slightly missed some positive pairs, it effectively filters out negative pairs. In comparison, competitive models like GLIP-L and FIBER-B show larger gaps in performance, with disparities of 12.25, 23.7, 11.75, and 11.0 in average for All, COCO, Object-365, and OpenImagesv5, respectively. In contrast, our methods demonstrate gaps of only 6.4, 13.1, 6.4, and 5.7, respectively. This showcases the efficacy of our meth-

ods in enabling models to work effectively, efficiently, and as a model-agnostic solution, thus highlighting their value and significance.

Moreover, it is worth noting that our model achieves significantly higher scores in AP-descr-L compared to other models, with substantial improvements observed in ALL (6.1), COCO (9.4), and OpenImagesv5 (10.7) evaluations, respectively. This observation suggests that solely relying on the SNN model for the identification of positive data pairs can be more challenging. However, when combining the capabilities of both the FNN and SNN models, our approach demonstrates superior performance across different lengths of data. This highlights the synergistic effect of utilizing both fast and slow neural networks, leading to higher scores compared to other models, particularly for longer descriptions.

Table 4.1: The evaluation results presented herein pertain to the Omnilabel validation 0.3.1 version. The AP-descr denotes the overall outcomes encompassing both positive and negative pairs, while the AP-descr-pos signifies the evaluation solely focused on positive pairs. Furthermore, the AP for short (AP-descr-S), medium (AP-descr-M), and long (AP-descr-L) length sentences was computed.

	Method	AP-descr	AP-descr-pos	AP-descr-S	AP-descr-M	AP-descr-L
All	RegionCLIP	2.6	3.2	3.6	2.7	2.3
	Detic	5.4	8.0	5.7	5.4	6.2
	MDETR	4.7	9.1	6.4	4.6	4.0
	GLIP-T	16.4	25.8	29.4	14.8	8.2
	GLIP-L	21.2	33.2	37.7	18.9	10.8
	FIBER-B	22.3	**34.8**	38.6	19.5	12.4
	Ours	**24.5**	30.9	**39.8**	**20.1**	**18.5**
COCO	RegionCLIP	3.5	5.1	6.1	3.3	4.1
	Detic	4.6	9.9	10.2	3.5	7.2
	MDETR	13.2	31.6	15.4	13.5	12.4
	GLIP-T	11.7	31.2	27.0	10.9	10.2
	GLIP-L	13.9	36.8	28.9	12.9	11.5
	FIBER-B	14.3	**38.8**	**31.3**	12.7	14.2
	Ours	**19.3**	32.4	28.7	**17.5**	**23.6**
Object-365	RegionCLIP	3.6	4.1	5.0	3.5	3.0
	Detic	5.7	8.4	6.6	5.9	6.9
	MDETR	3.2	5.9	3.0	3.2	2.7
	GLIP-T	18.1	26.9	34.2	16.0	9.1
	GLIP-L	24.0	35.2	44.5	20.5	11.8
	FIBER-B	**25.9**	**38.2**	**44.7**	**22.5**	14.1
	Ours	22.1	28.5	37.7	17.3	**14.8**
OpenImagesv5	RegionCLIP	2.7	2.9	3.4	2.7	2.0
	Detic	5.4	6.9	5.4	5.6	5.8
	MDETR	6.1	10.6	9.6	5.7	4.1
	GLIP-T	15.7	24.4	25.8	14.9	7.5
	GLIP-L	20.1	31.2	33.3	18.7	10.3
	FIBER-B	20.1	30.9	34.1	18.5	10.5
	Ours	**27.7**	**33.4**	**43.2**	**23.3**	**21.2**

Chapter 5

Conclusion

OmniLabel introduces a pioneering benchmark designed to evaluate language-based object detectors. Traditional language-based object detection models encounter challenges in discerning whether a given pair of language description and image is positively paired, as they typically assume the presence of only positive pairs within the dataset. To address this issue, our framework proposes a two-way strategy, leveraging the Fast Neural Network (FNN) and Slow Neural Network (SNN) approaches, which prove to be effective and efficient in resolving this problem.

Given the substantial scale of the Omnilabel benchmark, consisting of approximately 250 million pairs in total, with only 0.16 million pairs representing positive description-image pairs, it becomes highly inefficient for models to infer predictions for all data points. In our framework, we optimize the time and computational resources by employing a selective inference strategy. This approach significantly enhances the efficiency and effectiveness of the models by focusing on the most relevant data pairs rather than processing the entire dataset indiscriminately.

5.1 Limitation

The primary focus of this study is the resolution of negative and positive pairs within the Omnilabel benchmark. However, it is essential to recognize that positive and negative pairs represent just one aspect that distinguishes Omnilabel from other existing datasets. As discussed in Chapter 3, the Omnilabel benchmark encompasses a broader range of complexities, including intricate descriptions and varying quantities of bounding boxes for object-descriptions. Moreover, when considering object-categories, Omnilabel embraces a vast realm of free-form expressions, further highlighting its unique characteristics compared to other datasets in the field.

5.2 Future Work

While this work primarily focuses on addressing positive and negative pairs within the Omnilabel benchmarks, there remain other distinctive features within the Omnilabel dataset that warrant attention. Developing new datasets for this task is a resource-intensive endeavor; hence, alternative approaches such as pseudo labeling and semi-supervised learning methods can be explored to tackle these remaining challenges.

The utilization of models tailored for Omnilabel in the realm of robotics holds tremendous potential, offering manifold advantages across a myriad of robotics tasks. Omnilabel encompasses a diverse array of descriptions, exhibiting a rich spectrum of linguistic expressions and capturing a wide range of semantic nuances. Making it highly applicable to various vision-language tasks. However, it is important to acknowledge that working solely with vision-language datasets imposes inherent limitations. Leveraging the well-performing models developed for vision-language tasks holds the potential to extend their applica-

bility to various other tasks for robotics.

Furthermore, understanding human natural language is not always straightforward for artificial agents. Ensuring clear communication between agents and humans is crucial, and incorporating dialogue systems may facilitate better comprehension of tasks and enable effective task execution by the agents. This highlights the potential benefits of incorporating dialogue systems into the framework to enhance task understanding and performance.

Bibliography

OmniLabel benchmark. https://www.github.com. Accessed: 2023-01-24.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568, 2021.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Zi-Yi Dou, Aishwarya Kamath, Zhe Gan, Pengchuan Zhang, Jianfeng Wang, Linjie Li, Zicheng Liu, Ce Liu, Yann LeCun, Nanyun Peng, Jianfeng Gao,

and Lijuan Wang. Coarse-to-fine vision-language pre-training with fusion in the backbone, 2022.

Daniel Kahneman. *Thinking, Fast and Slow*. Penguin Books, 2011.

Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–291, 1979. ISSN 00129682, 14680262. URL `http://www.jstor.org/stable/1914185`.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7):1956–1981, 2020.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language, 2019.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, 2019.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016.

OpenAI. OpenAI: Introducing ChatGPT. `https://openai.com/blog/chatgpt`, 2022.

OpenAI. Gpt-4 technical report, 2023.

Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24, 2011.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen.

Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

Samuel Schulter, Vijay Kumar B G, Yumin Suh, Konstantinos M. Dafnis, Zhixing Zhang, Shiyu Zhao, and Dimitris Metaxas. Omnilabel: A challenging benchmark for language-based object detection, 2023.

Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8430–8439, 2019.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. *arXiv preprint arXiv:2202.03052*, 2022.

Chenyun Wu, Zhe Lin, Scott Cohen, Trung Bui, and Subhransu Maji. Phrasecut: Language-based image segmentation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10216–10225, 2020.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 69–85. Springer, 2016.

Haotian Zhang, Pengchuan Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. Glipv2: Unifying localization and vision-language understanding. *Advances in Neural Information Processing Systems*, 35:36067–36080, 2022.

국문초록

딥 러닝 분야의 놀라운 발전은 인간의 뇌가 보여주는 복잡한 인지 능력을 모방하려는 노력에 의해 촉진되었다. 그 결과로, AI 시스템의 탁월한 성능을 보여주면서 수많은 영역에서 놀라운 성과를 거두었다. 특히 자연어 처리, 컴퓨터 비전 및 음성 인식과 같은 단일 모달 분야에서 많은 발전이 이루어졌으나, 인간 인식은 복잡한 구조에서 다중 양태의 융합에 의해 수행되고 있다는 근본적인 차이점이 있다. 그럼에도 단일 모달에서의 발전에 힘입어 여러 모달을 다루는 연구도 많은 관심을 받아 발전을 거듭하고 있다.

멀티-모달을 다루는 문제 중에서, Omnilabel 벤치마크 Schulter et al. (2023)는 텍스트 설명을 기반으로 지칭된 물체의 위치를 찾는 문제를 제시한다. 사전에 정의되고 제한된 레이블의 공간에 의존하는 기존의 접근 방식과 달리, Omnilabel 벤치마크는 물체의 이름과 같은 간결한 단어 형태부터 복잡하고 상세한 자연어 설명에 이르기 까지 광범위한 객체 설명을 포함하고 있다. 또한 이 벤치마크의 특징은 하나의 물체에 대한 설명문이 0개 에서부터 여러 개의 객체를 지칭할 수 있다는 것이다. 따라서 설명문은 주어진 이미지에 존재하는 물체를 지칭하지 않고 있을 수도 있다. 인공지능 모델은 이러한 불일치 쌍에는 물체가 없음을 인지하여야 하며 일치하는 쌍에는 문장이 지칭하는 물체의 위치를 표시하여야 한다. 이 분야의 연구는 주로 기존 연구인 지칭표현에서 수행되었고, 일치 데이터 쌍을 입력으로 한다는 전제하에 수행되어왔다. 그러므로 기존 모델은 데이터셋에서 불일치 쌍을 입력받을 때 그동안 학습하지 않았던 문제에 직면하게 된다.

본 연구는 인간이 동등한 이득보다 손실을 더 무겁게 따지는 경향이 있다고 가정하는 손실 혐오 Kahneman and Tversky (1979)의 인지 과정에 기반하였다. 손실 혐오 이론은 개인이 어려운 질문을 더 쉬운 질문으로 대체하려하는 경향이 있다는 것을 나타낸다. 결과적으로, 인간의 뇌는 입력받은 데이터를 두 가지 시스

템을 사용하여 작동하도록 진화하였다 Kahneman (2011). 두 가지 시스템은 각각 빠른 생각(시스템 1)과 느린 생각(시스템 2)으로 불리며 빠른 생각은 직관적이고 자주 등장하는 문제를 처리하고, 느린 생각은 논리적이며 깊은 사고를 필요하는 문제를 처리할 때 사용된다.

우리는 이러한 생물학적인 기전에서 영감을 받아 시스템 1에 대응하는 Fast neural network(FNN)과 시스템 2에 해당하는 느린 신경망(SNN)을 제안하였다. FNN은 입력 데이터의 일치 혹은 불일치 여부에 따라 데이터를 분류하도록 학습된다. SNN은 FNN에서 일치 데이터 쌍으로 분류된 데이터만을 입력으로 받아 해당 쌍의 이미지에서 텍스트가 지칭하는 물체의 위치를 출력한다. 이러한 방식은 계산적으로 복잡한 SNN이 추가적인 학습 없이도 필터링된 데이터를 활용할수 있도록 하기 때문에 효과적이며 효율적인 결과를 보여준다. 이러한 방식을 통해 SNN만을 활용하여 데이터 셋 전체를 추론하는 기존의 방식보다 더 빠르고 좋은 성능을 낼 수 있음을 보였다.

주요어: 시각 표현, 다중 양태, 지칭표현, 대조학습
학번: 2021-29149