Master's Thesis of Electrical and Computer Engineering

An Attention-Based Speaker Naming Method for Online Adaptation in Non-Fixed Scenarios

고정되지 않은 시나리오에서의 온라인 적응을 위한 어텐션 기반 화자 명명 기법

February 2020

Graduate School of Engineering
Seoul National University
Electrical and Computer Engineering Major

Jungwoo Pyo

An Attention-Based Speaker Naming Method for Online Adaptation in Non-Fixed Scenarios

Professor. Sang Kyun Cha

Submitting a master's thesis of Public Administration

November 2019

Graduate School of Engineering
Seoul National University
Electrical and Computer Engineering Major
Jungwoo Pyo

Confirming the master's thesis written by
Jungwoo Pyo

December 2019

Chair	Bo Hyung Han	(Seal)
Vice Chair	Sang Kyun Cha	(Seal)
Examiner	Young Min Kim	(Seal)

Abstract

A speaker naming task, which finds and identifies the active speaker in a certain movie or drama scene, is crucial for dealing with high-level video analysis applications such as automatic subtitle labeling and video summarization. Modern approaches have usually exploited biometric features with a gradient-based method instead of rule-based algorithms. In a certain situation, however, a naive gradient-based method does not work efficiently. For example, when new characters are added to the target identification list, the neural network needs to be frequently retrained to identify new people and it causes delays in model preparation. In this paper, we present an attention-based method which reduces the model setup time by updating the newly added data via online adaptation without a gradient update process. We comparatively analyzed with three evaluation metrics(accuracy, memory usage, setup time) of the attention-based method and existing gradient-based methods under various controlled settings of speaker naming. Also, we applied existing speaker naming models and the attention-based model to real video to prove that our approach shows comparable accuracy to the existing state-of-the-art models and even higher accuracy in some cases.

Keyword : speaker naming, attention, online adaptation

Student Number : 2017-27184

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Biometric recognition plays an important role in advanced authentication systems. It identifies individuals based on physical or behavioral characteristics. The speaker naming task, which is to identify visible speaking characters in multimedia videos, consists of multiple types of biometric recognition. Most of the speaker naming methods distinguish an active speaker based on biometric features like face image or voice. The necessity of this task is proven to be essential for high-level video analysis problems such as summarization [1], interaction analysis [2], and semantic indexing [3]. In particular, the identification of active speaking characters with automatic subtitling can help deaf audiences to enjoy the videos more without any difficulties in understanding the context.

Most of existing speaker naming models mainly focus on boosting up the accuracy for finding active speaker among fixed origin character list. Gradient-based methods are considered one of the right solutions to get higher accuracy, and these methods have been proposed in various ways using multiple modalities for speaker naming. [4] proposes a deep multimodal model based on a CNN architecture to extract the facial and acoustic features from videos,

then combine them through a fusion function. Correspondingly, a multimodal Long Short-Term Memory(LSTM) architecture [5] merges visual and auditory modalities from the beginning of each input sequence. [6] improves the performance of talking-face detection by capturing the lip motion.

However, it is not always possible to occur expected or fixed situation in a video. In the real world, there are several uncertain situations which cause the difficulties of identifying active speaker such as appearing new characters or misinterpretation due to lack of labeled training data. Most of the gradient-based identification approaches cannot immediately adapt to a change in predicted character list or update the newly added data to model. Traditionally, these methods have to predefine a set of targeted characters before the training period. In particular, an existing model has to be retrained with a new set of targeted characters, which consist of both origin classes and the new ones, from scratch. It causes much time-consuming to rebuild the model. Transfer learning [7] and domain adaptation [8] are proved to be efficient for faster adaptation of neural network through initialization based on origin data. Nonetheless, these methods still require considerable time in the training phase adapting the newly added data to original model, which takes considerable time.

Besides, it is hard to contain a sufficient amount of labeled data in real-world datasets since labeling task is costly. The availability

of labeled data poses a major practical issue for many gradient-based models.

To overcome these problems, we apply an attention module with few-shot learning [9] for making our identification model flexible to accommodate changes at run-time. The attention module, which is based on scaled dot-product attention structure [10], represents the similarities between prior knowledge embeddings and extracted features from the target video. The prior knowledge embeddings are the given data, which consist of facial and vocal embeddings of the predicted classes from the training dataset.

Next, few-shot learning is used for dealing with the scarcity of labeled data and imbalanced class distribution. Attention mechanism and few-shot learning are effectively combined in our model since they are both linear and straightforward. The essential component of the few-shot learning method derives feature embeddings based on a distance function. The attention mechanism consists of a linear combination with scaling and a softmax operation among these feature embeddings as shown in Figure 1.1. This combination makes the model consider every single embedding of prior knowledge carefully. Therefore, our method works well under the conditions even with a small amount of data or a highly imbalanced class distribution. More importantly, our model only utilizes the pretrained neural networks to extract embeddings. It means our model does not comprise a backpropagation process unlike other gradient-based

models. Namely, the setup time is significantly decreased by updating the new information on the attention module in run-time.

However, our proposed method is not always an optimal solution for all situations. In situations where character changes are not frequent or there are many IDs to identify, a deep-learning approach that guarantees robust performance may be more suitable even if the model setup takes a long time. Consequently, we compared attention-based method with gradient-based methods under various conditions of speaker naming by adjusting two variables: the number of target IDs to be identified, and the number of shots per each character. Furthermore, we compared our proposed model with existing speaker naming models on real video.

Our contributions are summarized as follows:

- We proposed a non-gradient-based method using attention module with few-shot learning, which can efficiently deal with the scarcity of labeled data as well as imbalanced class distribution.
- Our model significantly reduces the setup time of the model by removing the gradient descent process and updating the new data to model online.
- Under various environments adjusting both the number of target IDs to be identified and the number of shots per each character, we conducted comparative analyses with real-world dataset between our proposed method and

existing gradient-based methods through three metrics: accuracy, memory usage, and setup time.

- Our model shows comparable accuracy to the state-of-the-art speaker naming models on real video.

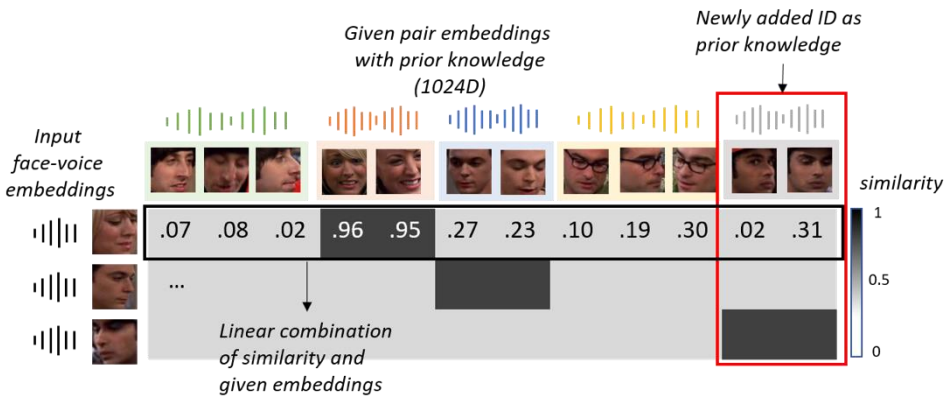

Figure 1.1: Visualization of predicting ID of face-voice pair with few-shot learning based attention module. The predicted ID of the target can be inferred to linear combination of cosine similarity between every prior knowledge embedding and target embedding. It also shows the simplicity of dealing with data of newly added ID inserting into the attention module.

Chapter 2

Related Work

2.1. Speaker Naming

Speaker naming is a task to identify the speaker in a video source. Recent studies about automatic speaker naming used deep neural networks to get each speaker's name from the multimodal face-voice sources. In [4], they proposed a convolutional neural network (CNN) based multimodal framework that automatically learns from the combined face and acoustic features. They trained the SVM classifier to reject all non-matching face-voice pairs and get identification results. Likewise, [5] improved the accuracy by changing the CNN based model to Long Short-Term Memory (LSTM) based model. This change gave robust identification results for face distortion. [12] used attention architecture to accommodate the face variations.

2.2. Feature Extractors for Face and Audio Cues

The primary purpose of feature extractors is to express a particular type of data to distilled numerical embeddings, which has lower dimension than original data. Several feature extractors have been studied in each field according to various types of data. Most

feature extractor operates by setting appropriate loss function and distance metric, then optimizing them.

2.2.1. Face

Various types of loss functions have been tried to use as facial feature extraction. [13], [14] used cross-entropy loss to minimize euclidean distance. FaceNet [15] introduced triplet loss based on Euclidean distance to train the face feature extractor. Also, FaceNet utilized MTCNN [16] to extract the aligned cropped face images from raw image dataset. SphereFace [17], CosFace [18], and ArcFace [19] used angular loss to minimize the cosine similarity.

2.2.2. Audio

The area of feature extraction of audio data has also been studied in various directions. There have been many useful methods, such as using MFCC [20] and using CNN [21]. Recently, [22] suggested a new model using "thinResNet" trunk architecture and dictionary-based NetVLAD layer. This method successfully performed the speaker identification task on audio data with varying lengths and mixing of unrelated signals.

2.3. Attention Mechanism

Attention mechanism is first proposed in [23] for neural machine translation(NMT) field. The attention mechanism looks up all of the

input elements (e.g., sequential input such as frames in video or words in a sentence) at every decoding time, calculates the attention map which is a matrix that reflects the relevance of present input and previous input elements.

Attention map is a probability matrix that each target word is aligned to, or translated from source word. Each element of the attention map is computed as the softmax value, which means the similarity of source word and each target word.

Some papers have brought attention mechanism to speaker naming task. In [12], they proposed attention guided deep audio-face fusion approach to detect active speaker. They also used individual network model to convert from face and voice sources to each embedding as ours. Before fusing the face and voice embeddings, they applied the attention module only for face embeddings to consider the relationship between other face embeddings. However, our work applied attention mechanism to fusion of face-voice pair embeddings and focused on the relevance of target embeddings and prior knowledge embeddings.

Chapter 3

Methodology

Speaker naming contains all processes from detecting faces, recognizing voice, and matching these embeddings to identifying the current speaker. As shown in Figure 3.1, we regard the speaker naming problem in two cases. The first case is to find out the pair embeddings that both face and voice embedding are identified as same ID(so-called "matched-pair"). The second one is to pick out the pair embeddings where ID of face and voice do not match(so-called "non-matched-pair"). We propose a non-gradient based method using attention networks with few-shot learning to solve this problem. In this section, we formulate our problem precisely and elaborate on our proposed model.

Figure 3.1: Speaker naming task contains two situations: i) finding matched face with corresponding voice if speaker appears in the scene, ii) picking out all distractors if speaker is out of the scene.

3.1. Problem Formulation

We formulate our problem as follows. Let t be the index of time window, $I = \{i_1, i_2, \dots, i_N\}$ denotes the ID of characters. J_t is the number of faces which are captured in t. f_j^t is the j−th number of face embeddings cropped in time window t. Likewise, v_t represents the voice embedding in time window t. Then the maximum probability of facial embedding whose ID is i_k in time window t is as follows.

$$F_{prob}(i_k, t) = \max_{1 \le j \le J} p(i_k | f_j^t) \tag{1}$$

By multiplying $F_{prob}(i_k, t)$ and the probability that predicted ID of voice embedding in t is i_k, we can infer the ID of speaker in t as below.

$$Spk_{ID}(t) = \underset{i_k \in I}{\operatorname{argmax}}(F_{prob}(i_k, t) \cdot p(i_k | v^t)) \tag{2}$$

Based on Equation (2), we calculate the accuracy of the speaker naming model if it correctly estimates the ID of matched-pair, or picks out the non-matched-pair in the time window t. After all, we aggregate $Spk_{ID}(t)$ over all time windows to get the total accuracy of the target video.

3.2. Attention-Based Method for Speaker Naming

The speaker naming problem consists of two parts: finding out matched face-voice pairs to predict current speaker, and picking out the non-matched-pairs. Our approach to solve the problem is as follows. First, we capture the face images and voice chunks by every fixed size of the time window. Then, we convert face images and voice chunks to extracted embeddings with pre-trained face and audio feature extractor. We concatenate both face and voice embeddings to make candidates of pair embeddings by each frame. Then, we calculate the attention map with this concatenated embedding. Attention map applies scaling and a softmax function to cosine similarity matrix among all of the characters' prior knowledge embeddings and extracted target embeddings. We predict the IDs of target embeddings based on the attention mechanism. Then, the proposed method aggregates the prediction result by each time window, and it determines the active speaker in the scene. Finally, we measure the prediction accuracy of the model by aggregating

every result of all time windows. We describe our method's overall architecture and flow in Figure 3.2.

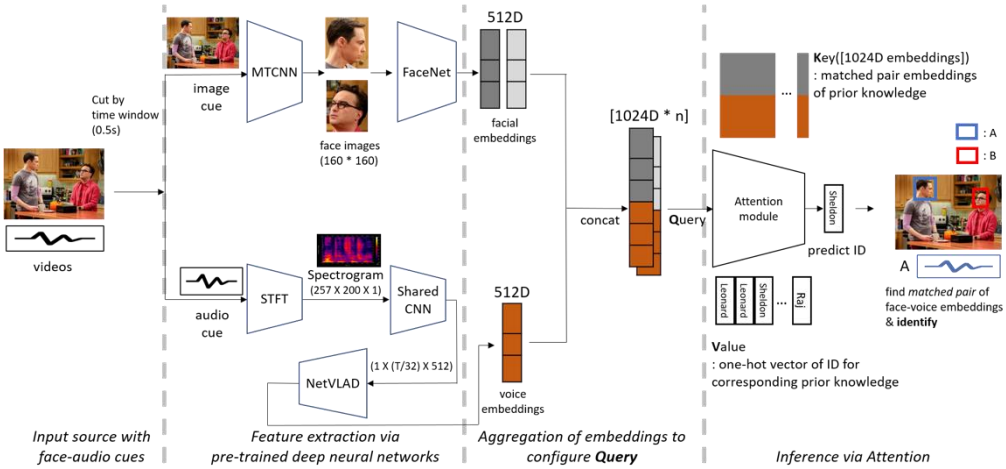

Figure 3.2: Overall architecture of attention-based model for speaker naming

3.2.1. Feature Extraction

For generating embeddings which contain the features of facial appearance and voice, we use pre-trained feature extractors, which convert raw input sources to numerical vectors with reduced dimensions.

Our network uses FaceNet as facial feature extractor, NetVLAD as voice feature extractor. The weights of these extractors are fixed while updating the attention module or end-to-end inferencing phase.

3.2.2. Attention Module with Few-Shot Learning

Our attention module with few-shot learning consists of multiple components. Let Q denotes the query matrix, which is the extracted face-voice pair embeddings from target video. Q contains several matched-pairs and non-matched-pairs, which we will predict within a certain time window. Thus, Q is a variable for time window t. K,V belong to prior knowledge for our network. K denotes the matrix of multiple face-voice pair embeddings extracted from the training data. V is a one-hot vector matrix of IDs corresponding to K. These KV set work as proofs for our decision whether the pair embeddings are matched-pair or not, and classifying the pair's ID.

Detailed process of attention mechanism is shown in Figure 3.3. The intuitive role of attention module is to consider the correlations between every pair of Q and K. In our case, computing attention map and context vectors in attention module correspond to computing similarity and matrix of predicted IDs, respectively. As a distance metric, we use cosine distance because embeddings in Q and K are unit vectors, we can get cosine similarity using inner product. Before performing matrix multiplication of Q and K, we use transposed matrix Q^T to match the dimension.

Our method performs a few additional operations after doing matrix multiplication with Q^T and K. First, multiply scale factor sf to $Q^T K$. Then, apply the softmax function to all elements. Our network is set to the value of sf as $\sqrt{d_K}$, where d_K is the dimension

of K, set to 1024. The reason for multiplying sf is that after performing multiplication between unit vectors to calculate cosine similarity, the value becomes so small that it interferes with subsequent softmax operation. If the input parameter's scale of softmax function is too small or large, it cannot express the appropriate probability distribution. Our setting can arrange the scale of value at the proper level to perform softmax.

Based on the above explanation, the attention map is mathematically written as:

$$A = \mathrm{softmax}(sf(Q^T K)) \tag{3}$$

The context vectors which represents the prediction of IDs to Q is written as:

$$C = VA^T \tag{4}$$

C represents the probability of which face−voice pair in Q is regarded as a particular ID. The probability of face and voice are separated in C. Our method uses confidence score vector c_p as the criteria for the decision to distinguish whether the p−th embeddings in Q is the active speaker or not. We apply Hadamard product [25], which multiplies the face part and voice part of C element−wise, to consider both features of face and voice. From this operation, we get $1 \times N$ vector of confidence score where N is the number of IDs. The maximum value and its index in c_p will be regarded as a confidence score and $Spk_{ID}(t)$, respectively.

We elaborate on the procedure of overall flow in Algorithm 1.

Algorithm 1. End−to−End Speaker Naming Prediction

Let Q: Query, K: Key, V: Value, A: Attention map
$I \leftarrow \{i_1, i_2, \cdots, i_N\}$: a set of characters' IDs
Cut video by 0.5s interval of time window
for each time window $t \leftarrow 1\ to\ T$ **do**
    Rep_t: representative frame in t
    $\{f_1, \cdots, f_{J_t}\}$: cropped J_t faces from Rep_t
    $\{q_{f_1}, \cdots, q_{f_{J_t}}\}$: facial embeddings from $\{f_1, \cdots, f_{J_t}\}$
    q_v: voice embedding extracted from audio in t
    $Q \leftarrow \begin{pmatrix} q_{f_1} & q_{f_2} & \cdots & q_{f_{J_t}} \\ q_v & q_v & \cdots & q_v \end{pmatrix}$
    $A \leftarrow \mathrm{softmax}(sf(Q^T K))$
    $C \leftarrow VA^T$
    $max\_conf \leftarrow 0$
    for $p \leftarrow 1$ to J **do**
        $c_p \leftarrow c_{f_p} \odot c_{v_p}$ *// Hadamard product*
        $max\_conf \leftarrow \max(max\_conf, \max(c_p))$
        if $max\_conf == \max(c_p)$ **then**
            $Spk_{ID}(t) \leftarrow argmax_{i \in I}(c_p)$
        end if
    end for
end for

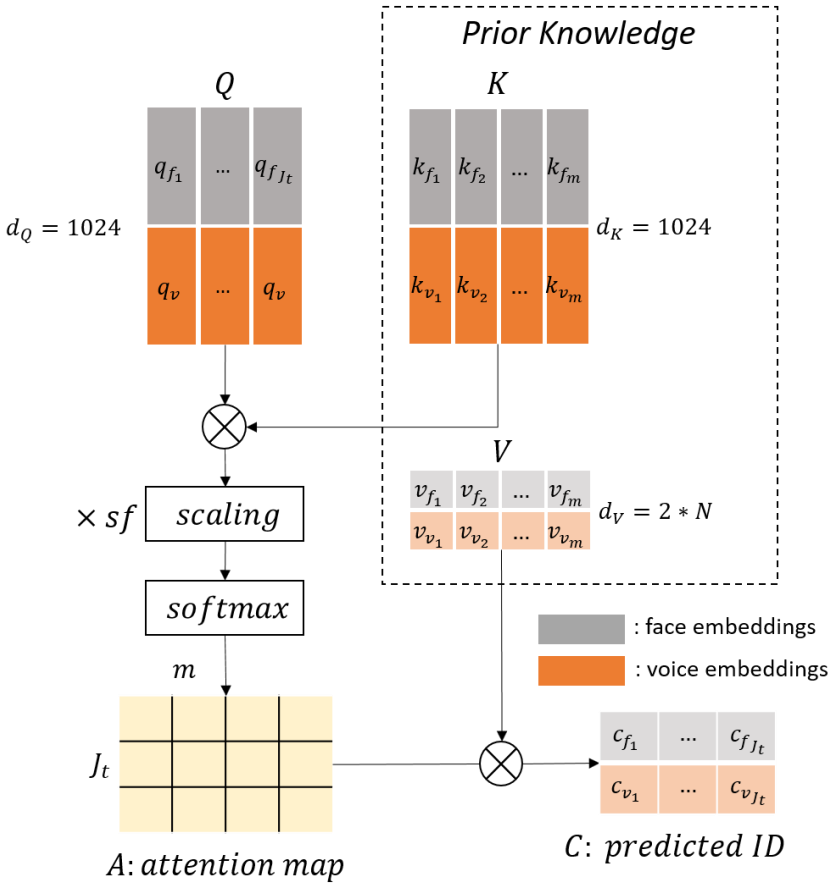

Figure 3.3: Mechanism of attention module with few−shot learning

Chapter 4

Experiments

4.1. Dataset Overview

In experiments, we used two public datasets: the utterance videos of celebrities(*VoxCeleb2* [11]) and a TV show(*The Big Bang Theory(BBT)*). For the experiment in incremental settings, we randomly chose 500 people from *VoxCeleb2*, which contain more than 10 videos. Then, we split the train set and valid set by 5:2 per each ID. For *BBT*, we selected 5 episodes (*S01E02, S01E03, S01E04, S01E05, S01E06*). Each episode consists of the whole video, face images with various poses and illumination, and aggregated voice file without silence.

4.2. Data Preprocessing

We used FaceNet and NetVLAD, which are the same extractors used in our model, for extracting train and test embeddings from the raw datasets. First of all, *BBT* dataset consists of multiple cropped face images and merged voice files per ID, by each episode. Cropped face images were resized to 160×160 for fitting the image size with the input of FaceNet model. After that, we used pre-trained FaceNet

to convert resized face images into 512-dimension embeddings. Similarly, we converted our audio file into 512-dimension voice embeddings. The window size of each audio chunk is 2s, cut with 0.1s stride.

We did additional preprocessing in order to get cropped face images and voice chunks when preprocessing *VoxCeleb2* because it only consists of video files. First of all, we cut the videos every 30 frames per second. Then, MTCNN [16] cropped face images for all frames. The captured images which are not actual face image were removed. About voice file, we applied the same setting of what *BBT* was preprocessed.

4.3. Comparative Analysis among Speaker Naming Methods under Various Settings

Previous studies [4], [5], [12] have evaluated the accuracy of their methods in a refined setting, which has purified voice and appears a small, fixed number of characters(5-6 IDs) in the scene. Also, they use sufficient pair embeddings per each character for training model. In this experiment, we compare our speaker naming model with existing gradient-based methods in detail under more various environments unlike previous work. By considering the advent of new characters in the story, we can precisely evaluate the performance of speaker naming methods in a more realistic situation with *VoxCeleb2* dataset.

4.3.1. Evaluation Metric

Speaker naming is to find the matched-pairs of face and voice embeddings and predict its identity. To compare how well the speaker naming method can identify the ID of matched-pair, we defined matching pair accuracy(mpA) as follows:

$$mpA = \frac{N_{id_{pred}==id_{gt}}}{N_{total}} \times 100\%$$

The second metric is the number of parameters of speaker naming model loaded in memory. If the model consists of a neural network, the weights and biases belonged to these parameters. In the case of attention-based model, pair embeddings were counted as parameters. We convert these parameters into kilobytes(KB) and compare them.

The third metric is the setup time of model. In the case of neural network, we calculated setup time by adding data loading in memory, calculating the gradient, and updating it to the weights. In the case of an attention-based model, we measure the setup time by adding the loading time of prior knowledge embeddings and the calculation time of attention module to derive prediction results.

4.3.2. Experimental Setup

We conducted the experiment adjusting two main variables in situations: the number of target IDs for prediction, and the number of

shots(pair embeddings of face−voice) for prior knowledge per each target ID.

About the number of target IDs, we separated the situation into two parts: the number of target IDs is small or large. In each case, the number of target IDs started to be set from 5 to 50 with five increments, and from 50 to 500 with fifty increments, respectively. We also adjust the number of shots per each character set to 5(small), 50(large) shots in both situations to consider the effects of the number of labeled training data to performances of speaker naming methods.

As the baseline methods, we selected two representative gradient−based methods to compare with our $Attention-based(Att-based)$ method. The first method is $Training\ from\ Scratch(TfS)$ which trains the neural network with both original and new data. Most of the deep neural networks normally use TfS in training phase. The second one is $Learning\ without\ Forgetting(LwF)$, which generates the new branch on top of the network and trains with only new data.

We followed the same neural network structure with one of the previous work [4] on both methods for fair comparison. The maximum training epoch is 500, which is sufficient to converge the loss function. If the network reached the optimal cost before the maximum epoch while training, we took the accuracy and the setup time at the moment the optimal cost was derived. Transfer learning

was applied in every stage of all gradient−based methods when the number of IDs was increased.

4.3.3. Results

As shown in Figure 4.1 and Table 4.1, we conducted both quantitative and qualitative analysis based on the experimental results. Most notably, our method($Att-based$) significantly reduced the setup time of model compared to other gradient−based methods about tens to hundreds of times regardless of conditions.

The mpA was high in order of TfS, $Att-based$, LwF. However, when the number of target IDs was 450 with large shots, the LwF gradually surpassed $Att-based$ as shown in the "Large IDs−mpA" graph in Figure 4.1. Generally, gradient−based methods showed a big difference in mpA depending on the number of shots. In contrast, $Att-based$ worked well in both situations and had less effect in terms of the number of shots.

$Att-based$ utilized small number of parameters when the number of target IDs or its shots are small. However, as the number of target IDs is increased with large(50) shots, $Att-based$ showed memory inefficiency because the number of parameters increased quadratically with (the number of IDs × the number of shots per ID). In contrast, TfS occupies constant number of parameters; it is only related to the structure of the neural network. LwF is proportional to the number of times that ID is added. Because LwF has a multi−

branch structure, the new branch is generated when the new target character comes in.

To sum up, $\boldsymbol{Att-based}$ is the most appropriate method when new people appear frequently, and the shots per each character is not sufficient. Also, $\boldsymbol{Att-based}$ works effectively where immediate update for hard-to-recognize data such as various facial poses is needed. Overall, $\boldsymbol{TfS}$ is the best suited for situations where the new people are not frequently updated and high accuracy is required. $\boldsymbol{LwF}$ locates in the middle of other two methods, because it shows faster setup of model than $\boldsymbol{TfS}$, but compromising its $\boldsymbol{mpA}$ and memory usage.

# of target IDs	# of shots per each ID	Metric	$Training\ from\ Scratch(TfS)$	$Learning\ without\ Forgetting\ (LwF)$	$Att-based\ (Ours)$
Small IDs (5-50 people)	Small shots (5 shots)	mpA(%)	**86.05**	61.02	83.38
		# of params(KB)	6248.4 (constant)	15473.4 (linear)	**564.39 (quadratic)**
		Setup time(s)	60.02	7.06	**0.29**
	Large shots (50 shots)	mpA(%)	**94.80**	80.29	88.06
		# of params(KB)	6248.4 (constant)	15473.4 (linear)	**5644.03 (quadratic)**
		Setup time(s)	381.21	86.28	**2.64**
Large IDs (50-500 people)	Small shots (5 shots)	mpA(%)	**78.67**	23.06	67.68
		# of params(KB)	6248.4 (constant)	33923.4 (linear)	**5752.3 (quadratic)**
		Setup time(s)	437.20	85.14	**2.50**
	Large shots (50 shots)	mpA(%)	**84.20**	66.94	70.39
		# of params(KB)	**6248.4 (constant)**	33923.4 (linear)	57523.15 (quadratic)
		Setup time(s)	1973.08	909.22	**25.05**

Table 4.1: Summary for comparison of performances between gradient-based methods($\boldsymbol{TfS,\ LwF}$) and attention-based($\boldsymbol{Att-based}$) method under various settings. The numbers in the table are the average of the measurements in the range.

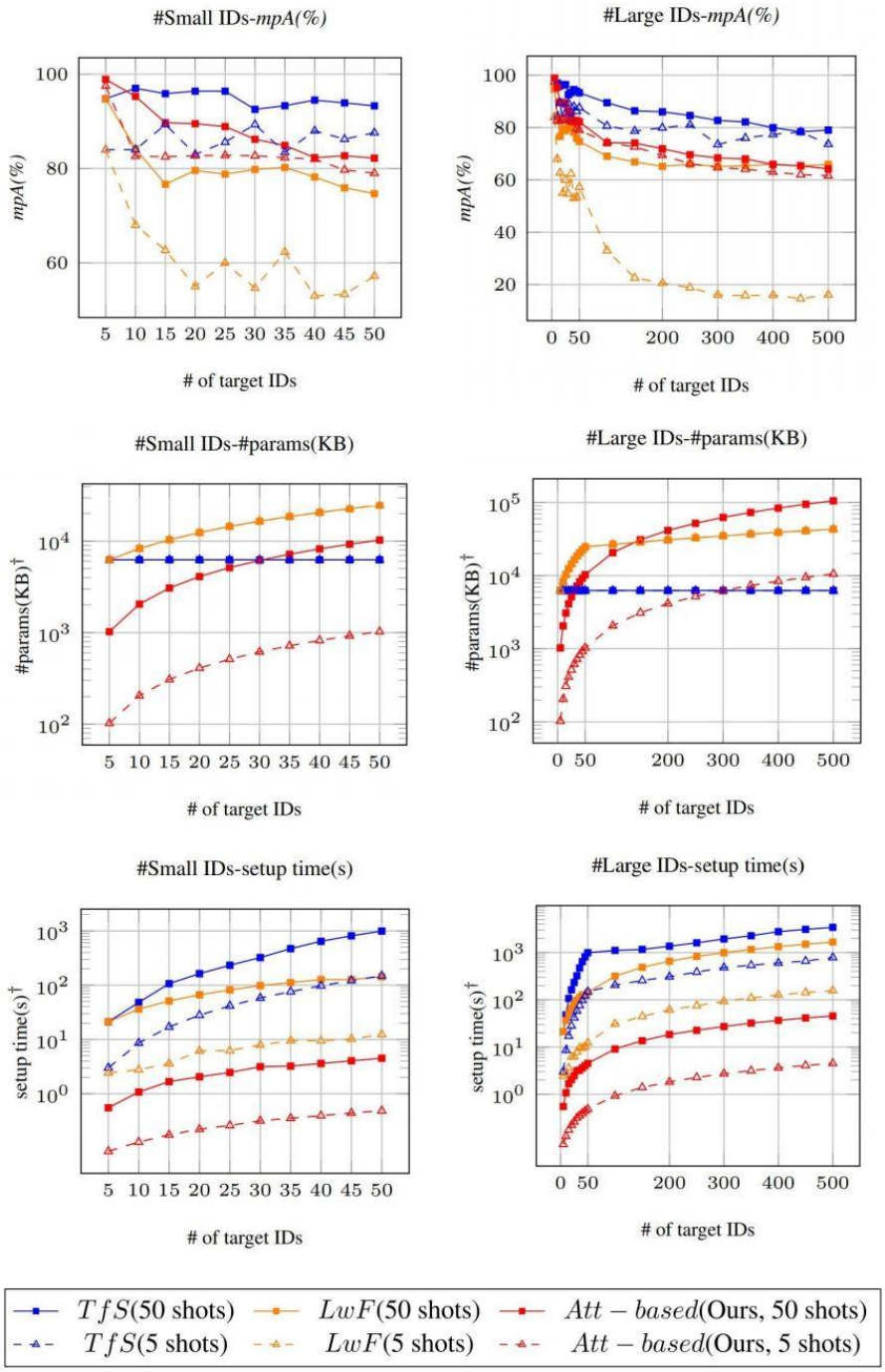

Figure 4.1: Comparative Analyses between speaker naming methods under various settings. Three metrics(***mpA***, the number of parameters, setup time) are measured for each situation where the number of target IDs and the number of shots per character are changed. †The y−axis of the graph is in logarithmic scale.

4.4. Speaker Naming Accuracy for Real Video

In this experiment, we applied our model in real video to compare the accuracy with previous gradient-based speaker naming models.

4.4.1. Evaluation Metric

Speaker naming accuracy(snA) had used broadly in multiple speaker naming related papers before and was formulated in [12], which is from one of our baseline. We used this metric to measure the performance of real video inference for comparing our model with a well-known speaker naming baselines. They define snA as follows:

$$snA = \frac{N_{p^{sn}==s^{tr}}}{N_{s^{tr}}} \times 100\%$$

where p^{sn} and s^{tr} denote the labels of predicted samples and ground truth, respectively. $N_{p^{sn}==s^{tr}}$ is the number of correctly predicted time windows and $N_{s^{tr}}$ is the total number of time windows.

4.4.2. Experimental Setup

For evaluation, we followed the same settings of previous works [4], [5], [27], [28] on speaker naming experiment. The four-minute-long *BBT S01E03* video clip was used for evaluation dataset.

In real situation, there occurs many non-matched-pairs of face-voice embeddings per period. Unlike previous controlled settings, we put 30 shots of matched and non-matched-pairs at a ratio of 1 to 4

in prior knowledge embeddings because end-to-end inference detects not only an active speaker but also distractors.

We tested our model to video with the time window of multiple periods of 0.5s. Time windows more than 0.5s were also tested for comparing existing methods to clarify the result. If the time window is more than 0.5s, the prediction of the model is determined by the majority vote of multiple 0.5s-sized windows as previous work [4] did.

4.4.3. Results

As shown in Table 2, $Att-based$ showed comparable snA as other gradient-based speaker naming models in most cases. In certain circumstances, such as the size of the time window is 1s, 2s, and 3s, our model even outperformed the other state-of-the-art models.

Time window(s)	[27]	[28]	[4]	[5]	[2]	$Att-based$ $(Ours)$
0.5s	–	–	74.93	86.59	**87.73**	84.34
1s	–	–	77.24	89.00	–	**92.50**
1.5s	–	–	79.35	**90.45**	–	88.89
2s	–	–	82.12	90.84	–	**92.45**
2.5s	77.81	80.80	82.81	**91.17**	–	89.89
3s	–	–	83.42	91.38	–	**93.12**

Table 4.2: Speaker naming accuracy(snA $(\%)$) comparison between attention-based model and existing speaker naming models on real video of *BBT S01E03*.

4.5. Ablation Studies

We conducted this experiment to study our non-gradient based speaker naming model's performance in various type of cues and different dimension of embeddings. We tested using matched pairs of multimodal face-voice cues and uni-modal cue(face-only, voice-only) when the dimension of feature is 128D and 512D respectively. Despite the voice-only model was impractical in our defined speaker naming task due to lack of localizing active speaker, we prepared voice-only model for systematic comparing with other cases. The number of test target embeddings for *BBT* was set to 40000. About prior knowledge embeddings, we set 50 shots per each ID. Thus, the total number of prior knowledge embedding set of *BBT* is 250. The tested models were face-voice, face, voice model based on our method.

Dimension of *BBT* embeddings	Face-Voice	Face-Only	Voice-Only
128D	**96.23**	93.73	82.65
512D	97.26	89.62	**98.36**

Table 4.3: *mpA*(%) of face-voice fusion model, uni-modal(face-only, voice-only) model on *BBT* dataset with 128D and 512D. We add voice model only for comparing performance, though it is impractical one in speaker naming task.

Table 4.3 showed the *mpA* results of ablation studies about our tested models. Comparing with face-voice fusion model and face-only model, the fusion models(96.23%, 97.26%) were superior to the

face-only models. About voice models, face-voice fusion model outperformed voice model with a huge difference on 128D model. However, on 512D model, the voice model showed slightly higher mpA than face-voice model. We see that using matched pairs of face-voice fusion cues is more effective than using single face cue, or voice cue in some cases.

Chapter 5

Conclusion and Future Work

5.1. Conclusion

In this paper, we presented an attention-based speaker naming method for online adaptation in non-fixed scenarios. The key idea is to predict the ID of the matched-pair based on attention mechanism, which considers the correlation between all pairs of prior knowledge embeddings and extracted target embeddings.

Our proposed approach significantly reduced the model setup time by keeping comparable accuracy to existing state-of-the-art models, as demonstrated in our experiments. Also, the model can be updated online by only changing information on the attention module.

5.2. Future Work

Our further research aims to solve the current limitations and improve the method well-applied to more generalized situations. Now, our current method was using only two modalities and showed low accuracy when the number of target IDs for identification is large. Also, it can occur memory inefficiency if the number of IDs and the number of shots per ID are increased. If we properly combine the advantages of the gradient-based methods with our method, the

integrated method will be one of the solutions to cover more various situations adequately in the future.

Bibliography

[1] K. Takenaka, T. Bando, S. Nagasaka, and T. Taniguchi, “Drive video summarization based on double articulation structure of driving behavior,” in *Proceedings of the 20th ACM international conference on Multimedia,* pp. 1169-1172, ACM, 2012.

[2] C. Liu, S. Jiang, and Q. Huang, “Naming faces in broadcast news video by image google,” in *Proceedings of the 16th ACM international conference on Multimedia,* pp. 717-720, ACM, 2008.

[3] H. Zhang, Z.-J. Zha, Y. Yang, S. Yan, Y. Gao, and T.-S. Chua, “Attribute-augmented semantic hierarchy: towards bridging semantic gap and intention gap in image retrieval,” in *Proceedings of the 21st ACM international conference on Multimedia,* pp. 33-42, ACM, 2013.

[4] Y. Hu, J. S. Ren, J. Dai, C. Yuan, L. Xu, and W. Wang, “Deep Multimodal Speaker Naming,” in *Proceedings of the 23rd Annual ACM International Conference on Multimedia,* pp. 1107-1110, ACM, 2015.

[5] J. S. Ren, Y. Hu, Y.-W. Tai, C. Wang, L. Xu, W. Sun, and Q. Yan, “Look, Listen and Learn – A Multimodal LSTM for Speaker Identification,” in *Proceedings of the 30th AAAI Conference on Artificial Intelligence,* pp. 3581-3587, 2016.

[6] H. Bredin and G. Gelly, “Improving speaker diarization of tv series using talking-face detection and clustering,” in *Proceedings of the 24th ACM International Conference on Multimedia,* MM’ 16, (New York, NY, USA), pp. 157-161, ACM, 2016.

[7] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, “How transferable are features in deep neural networks?,” in *Advances in neural information processing systems,* pp. 3320-3328, 2014.

[8] Ganin, Y., and Lempitsky, V., “Unsupervised domain adaptation by backpropagation.” in *Proceedings of the 32nd International Conference on International Conference on Machine Learning,* Volume 37, ICML’15, 1180–1189.

[9] L. Fei-Fei, R. Fergus, and P. Perona, “One-shot learning of object categories,” *IEEE transactions on pattern analysis and machine intelligence,* vol. 28, no. 4, pp. 594-611, 2006.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in neural information processing systems,* pp. 5998-6008, 2017.

[11] J. S. Chung, A. Nagrani, and A. Zisserman, “Voxceleb2: Deep speaker recognition,” *arXiv preprint arXiv:1806.05622*, 2018.

[12] X. Liu, J. Geng, H. Ling, and Y. ming Cheung, “Attention guided deep audio−face fusion for efficient speaker naming,” *Pattern Recognition,* vol. 88, pp. 557 - 568, 2019.

[13] Y. Sun, Y. Chen, X. Wang, and X. Tang, “Deep learning face representation by joint identification−verification,” in *Advances in neural information processing systems,* pp. 1988-1996, 2014.

[14] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, “A discriminative feature learning approach for deep face recognition,” in *European conference on computer vision,* pp. 499-515, Springer, 2016.

[15] F. Schroff, D. Kalenichenko, and J. Philbin, “Facenet: A unified embedding for face recognition and clustering,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815-823, 2015.

[16] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, “Joint face detection and alignment using multitask cascaded convolutional networks,” *IEEE Signal Processing Letters,* vol. 23, no. 10, pp. 1499-1503, 2016.

[17] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, “Sphereface: Deep hypersphere embedding for face recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition,* pp. 212- 220, 2017.

[18] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, “Cosface: Large margin cosine loss for deep face recognition,” in

Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5265–5274, 2018.

[19] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* pp. 4690– 4699, 2019.

[20] L. Muda, M. Begam, and I. Elamvazuthi, "Voice recognition algorithms using mel frequency cepstral coefficient (mfcc) and dynamic time warping (dtw) techniques," *arXiv preprint arXiv:1003.4083,* 2010.

[21] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, et al., "Cnn architectures for large-scale audio classification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* pp. 131–135, IEEE, 2017.

[22] W. Xie, A. Nagrani, J. S. Chung, and A. Zisserman, "Utterance-level aggregation for speaker recognition in the wild," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* pp. 5791–5795, IEEE, 2019.

[23] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473,* 2014.

[24] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018),* pp. 67–74, IEEE, 2018.

[25] R. A. Horn, "The hadamard product," in *Proc. Symp. Appl. Math,* vol. 40, pp. 87–169, 1990.

[26] Z. Li and D. Hoiem, "Learning without forgetting,*" IEEE transactions on pattern analysis and machine intelligence,* vol. 40, no. 12, pp. 2935–2947, 2017.

[27] M. Bauml, M. Tapaswi, and R. Stiefelhagen, "Semi-supervised learning with constraints for person identification in multimedia data," in *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition,* CVPR' 13, (Washington, DC, USA), pp. 3602–3609, IEEE Computer Society, 2013.

[28] M. Tapaswi, M. Buml, and R. Stiefelhagen, "knock! knock! who is it? probabilistic person identification in tv-series," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2658–2665, June 2012.

초 록

특정 영화 또는 드라마 장면에서 활성 화자를 찾고 식별하는 화자 명명 작업은 자동 자막 라벨링 및 비디오 요약과 같은 고급 비디오 분석 응용 프로그램을 처리하는 데 중요하다. 현대의 접근 방식은 일반적으로 규칙 기반 알고리즘 대신 기울기 기반 방법으로 생체 인식 기능을 활용했다. 그러나 특정 상황에서는 단순한 기울기 기반 방법이 효율적으로 작동하지 않는다. 예를 들어, 새로운 인물이 목표 식별 리스트에 추가 될 때, 뉴럴 네트워크는 새로운 사람들을 식별하기 위해 자주 재훈련 되어야 하고 이것이 모델 준비를 지연시킨다. 이 논문에서는 기울기 업데이트 프로세스 없이 온라인 적응 기법을 통해 새로 추가된 데이터를 업데이트하여 모델 준비 시간을 줄이는 어텐션 기반 방법을 제시한다. 우리는 3 가지 평가 지표 (정확도, 메모리 사용량, 모델 준비 시간)를 통해 어텐션 기반 방법과 기존의 기울기 기반 방법들을 화자 명명 작업의 다양한 제어된 설정 하에서 비교 분석했다. 또한 기존의 화자 명명 모델과 어텐션 기반 모델을 실제 비디오에 적용한 결과 우리의 접근 방식이 기존의 최첨단 모델과 비슷한 수준의 정확도를 보여주거나, 경우에 따라 더 높은 정확도를 보여주었다.

Keyword : 화자 명명, 어텐션, 온라인 적응 기법

Student Number : 2017-27184