修　士　論　文

Impression Analysis on Presentations using Attention-based LSTM

(Attention-based LSTMを用いたプレゼンテーションの印象解析)

指導教員　山﨑 俊彦　准教授

東京大学大学院
情報理工学系研究科
電子情報学専攻

氏　名　48-186448　易　聖舟

提　出　日　2020 年 1 月 30 日

Abstract

Presentation skills are very important because they can help speakers keep a presentation interesting and motivate the audience to listen. We focus on the impression analysis and the support system of presentation to help the speakers improve their presentation skills. In this thesis, we propose a multimodal neural network to automatically predict impression-related evaluation on presentations. We can predict audiences' impressions for speech videos using text data and audio data. We first used two attention-based LSTMs to extract the linguistic features and the acoustic features, respectively. Especially, the input sequences of audio model are the segmental features that were extracted from audio mel-spectrograms by using CNN. After we got the unimodal features, we used a model-level attention network for feature fusion and final classification. We also applied our language model to another type of presentation: press conference. We only used text data to predict the evaluation of the consultant for press conferences. The word representation consists of token embedding using ELMo and type embedding. We then used the language model to predict consultants' evaluation of Q&A pairs between speakers and journalists.

Contents

1. Introduction **1**

1.1. Oral Presentation . . . 1

1.2. Emotion Recognition . . . 1

1.3. Press Conference . . . 2

1.4. Organization of This Thesis . . . 2

2. Presentation Impression Prediction **3**

2.1. Introduction . . . 3

2.2. Related Work . . . 5

2.2.1. Presentation Analysis . . . 5

2.2.2. Document Classification . . . 6

2.2.3. Audio Classification . . . 7

2.3. Unimodal Feature Learning . . . 7

2.3.1. Linguistic Feature Learning . . . 7

Word Embedding . . . 7

Time-related Linguistic Feature . . . 8

2.3.2. Acoustic Feature Learning . . . 8

Temporal Acoustic Feature . . . 8

Time-related Acoustic Feature . . . 10

2.3.3. Attention Approach . . . 11

Dot-Product Attention . . . 11

Smoothing Normalization . . . 11

2.4. Multimodal Feature Fusion . . . 12

2.5. Experiments . . . 13

2.5.1. Dataset . . . 13

2.5.2. Training . . . 14

2.5.3. Comparative Methods 15
2.6. Results and Analysis 15
2.7. Conclusion and Future Work 18

3. Press Conference Evaluation **20**
3.1. Introduction 20
3.2. Related Work 21
3.2.1. Text Classification 21
3.2.2. Conversation Analysis 21
3.2.3. Word Representations 21
3.3. Automatic Evaluation Method 22
3.3.1. Language Model 22
3.3.2. Word Representation 24
3.4. Experiments 25
3.4.1. Dataset 25
3.4.2. Training 26
3.5. Results and Analysis 27
3.6. Conclusion and Future Work 30

4. Audio Emotion Recognition **31**
4.1. Introduction 31
4.2. CNN Architecture 33
4.3. Experiments 34
4.3.1. Dataset 34
4.3.2. Training 34
4.4. Results and Analysis 35
4.5. Conclusion and Future Work 36

5. Demo System for Presentation Improvement **37**
5.1. Introduction 37
5.2. Presentation Impression Prediction System 37
5.2.1. SVM-MRF 38

5.2.2. Multimodal Neural Network 39

5.3. Slide Analysis System 40

5.4. Experiments 41

5.4.1. Presentation Impression Prediction 41

5.4.2. Slide Analysis 42

5.5. Demo 42

5.6. Conclusion 44

6. Conclusion and Future Work 45

6.1. Conclusion 45

6.2. Future Work 46

References 51

Publications 57

Acknowledgements 60

List of Figures

2.1. The architecture of proposed model . . . 5

2.2. Linguistic feature learning . . . 9

2.3. An example of audio mel-spectrogram . . . 10

2.4. Acoustic feature learning . . . 11

2.5. Feature fusion network . . . 12

2.6. Raw audio data sample . . . 13

2.7. Average attention . . . 17

2.8. Average vote rates . . . 18

2.9. Prediciton accuracy . . . 18

2.10. The prediction result of a presentation by Bill Gates . . . 19

2.11. The prediction result of a presentation by Al Gore . . . 19

3.1. Overview of the method . . . 22

3.2. LSTM with self-attention mechanism . . . 23

3.3. Word representation . . . 23

3.4. Sample construction . . . 25

3.5. Prediction Accuracies . . . 27

4.1. CRNN [60] . . . 32

4.2. Mel-spectrogram . . . 32

4.3. Without downsampling . . . 34

4.4. With downsampling . . . 34

5.1. Continous impression prediction . . . 38

5.2. Negative Class Activation Mapping . . . 39

5.3. Prediction accuracies for different proposals (top and bottom 30%) 40

5.4. Prediction accuracies for different languages using multimodal neural network (top and bottom 50%) 41
5.5. Prediction example 42
5.6. Presentation impression prediction system 42
5.7. Slide analysis system 43

6.1. The architecture of the multi-agent reinforcement learning method [66] 47
6.2. The architecture of the multi-layer residual LSTM network [67] . . . 48
6.3. Word-level aligned multimodal features [1] 49
6.4. Cyclic translation for joint representation [2] 49
6.5. Multimodal Transformer [1] 50

List of Tables

2.1. Official caption samples 13

2.2. Prediction accuracies of linguistic features 16

2.3. Prediction accuracies of acoustic features 16

2.4. Prediction accuracies of multimodal features 16

3.1. Press conferences 25

3.2. Evaluation criterions 26

3.3. Example of prediction result 28

3.4. Task A 29

3.5. Task B 29

4.1. The architecture of 6-layer model 33

4.2. Experiment results 35

4.3. Top-5 and bottom-5 tag-wise AUCs 36

Chapter 1

Introduction

1.1. Oral Presentation

Oral presentation refers to the process of presenting a topic to audiences, and it is thought as the most standard format to express ideas or introduce products in many scences. In our daily lives, we have many opportunities to listen to an oral presentation or make a presentation in front of others. Presentation skills become more and more important because they can help speakers keep a presentation interesting and motivate the audience to listen. For the way of improving our presentation skills, we can practice by ourselves or view others' presentations. Thanks to the information age, we can view colorful presentation videos on many online sites. In our research, we use these presentation videos to develop an automatic evaluation system to support presensters' self practice.

1.2. Emotion Recognition

Emotion recognition is a relatively nascent research area, but there are more and more researchers focusing on it. For example, some researchers worked on multimodal emotion recognition [1, 2, 3]. They used different types of data, including video, text, and audio. There are also many researches on unimodal emotion recognition. For example, Tang et al. [4] and Islam et al. [5] worked on Twitter emotion classification. Some text emotion recognition tasks have become indispensable criterions to evaluate a language model, including Amazon review

[6], Yelp 2015 [1], and IMDB Reviews [2]. There are also many researches on audio emotion recognition, like music emotion recognition [7, 8].

1.3. Press Conference

Press conference is a special type of presentation for handling public relations issues. Different from common oral presentations, they usually take place in the format of Q&A. In the press conferences, multiple presenters need to answer journalists' questions one by one. There are no exsiting press conference dataset, and no authoritative criterions to evaluate speakers' performance. Therefore, in our collaborative research with professional consultants, we create a press conference dataset for speakers' performance evaluation. And we propose a language model for automaticlly predict consultants' evaluation on speakers.

1.4. Organization of This Thesis

This thesis is organized as follows. We first introduce the impression prediction of oral presentations in Chapter 2. We then present the automatic evaluation of press conferences in Chapter 3, and we introduce our fundamental research of audio emotion recognition in Chapter 4. After that, we present a demo of oral presentation support system in Chapter 5. Chapter 6 concludes the thesis and introduces our future works.

[1] https://www.yelp.com/dataset

[2] https://www.imdb.com

Chapter 2

Presentation Impression Prediction

2.1. Introduction

In our daily lives, we have many opportunities to listen to someone's oral presentation or make a presentation in front of a crowd of people. We can also view colorful presentations on online sites, such as Big Think [1], YouTube [2], and TED Talks [3]. Even though oral presentations are everywhere in our daily lives, technical methods for providing audience feedback to the speaker are rare outside of using questionnaire, which is time consuming to create, as well as to collect and analyze the results.

There are some researches about analyzing how audiences think about the presentation that they viewed. For existing methods, Fukushima et al. [9] used presentations in TED Talks to predict audience impressions by using Support Vector Machine (SVM) based on acoustic features and several types of linguistic features. Yamasaki et al. [10] used the same dataset, but they added Markov Random Field (MRF) to consider the correlation between different features instead of simply concatenating all features together. However, the effectiveness of their proposal was limited, because their unimodal classifiers in the SVM-MRF architecture can't be finetuned in the stage of multimodal feature fusion.

In order to provide presentation speakers with more efficient and convincing feedbacks, we propose a multimodal neural network to predict whether their presentation leaves particular impressions on audiences (Figure 2.1). Our proposal

[1] https://bigthink.com

[2] https://www.youtube.com

[3] https://www.ted.com/talks

includes two Long Short-Term Memory (LSTM) [11]-based neural networks to learn linguistic representation for video captioning as well as acoustic representation for raw audio data separated from the video. We then consider the correlation between linguistic representation and acoustic representation by using an attention network, and we use a multilayer perceptron (MLP) as the final classifier. In our experiment, proposed method achieves the average accuracy of 85.0% for independent predictions of all 14 types of impressions. We also evaluate the performance of the language model and the acoustic model, respectively. Their classification accuracies are 81.8% and 70.0%, respectively. Further, we find that impression types have an effect on audiences' attention, given to the presentation contents or to the presenters' manner of speaking.

The contributions of our works are as follows,

- The proposed presentation impression prediction system based on a multimodal neural network achieved significant improvement to accuracy and complexity over existing methods.
- The proposed multimodal neural network overcame the problem of exiting methods by enabling the representation models of unimodal features to be finetuned during multimodal feature fusion.
- We applied an LSTM-based model for oral presentation analysis and adapted the long-duration raw audio data to the capacity of LSTM for modeling long-term dependency by transferring the sequence data from time domain to frequency domain.
- The proposed method clarified the importance relationship between the content of a presentation and the manner of speaking for different audience impressions.

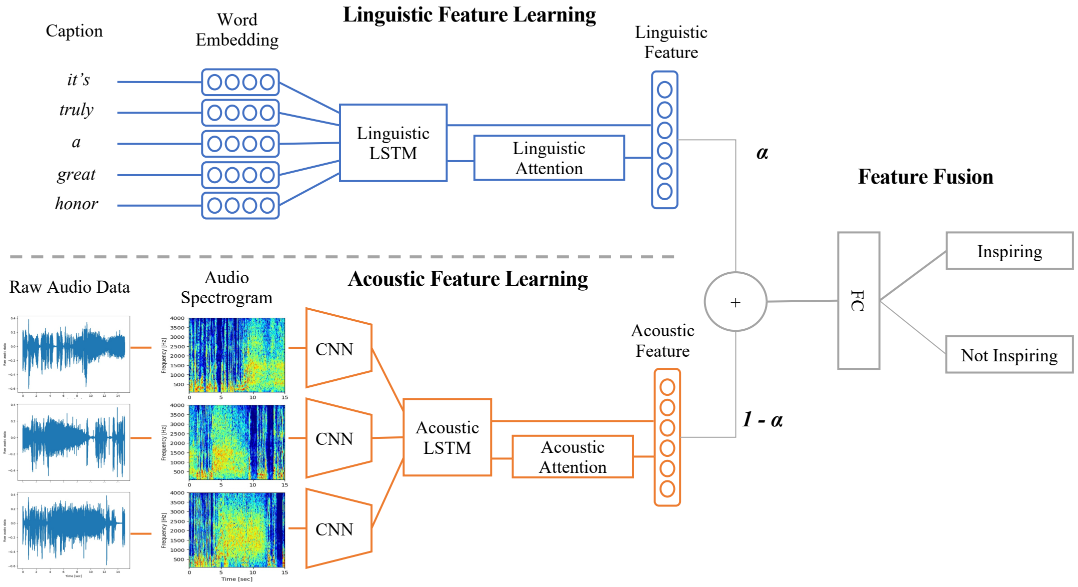


FIGURE 2.1 – The architecture of proposed model

2.2. Related Work

2.2.1. Presentation Analysis

Presentation videos have been used to explore presentation techniques [12]. Some researchers studied the relationship between people's evaluations and their superficial impressions of speakers [13]. Zhang et al. fused the local binary patterns features extracted for facial image representations and typical acoustic features to perform audio-visual emotion recognition of presentation [14]. Xu et al. [15] and Yoon et al. [16] used multimodal neural network for speech emotion recognition The former only applied an attention mechanism on acoustic feature to find the special part of audio that contains a strong emotion information, conditioning on the caption information. Contrarily, the latter applied the attention mechanism on linguistic feature, which conditioned on the acoustic information. Except these late fusion methods, EF-LSTM (Early Fusion LSTM) used a single LSTM on concatenated multimodal inputs [17, 18].

There are some researches about evaluating presentations [19, 20], and some researches about analyzing how audiences consider a presentation they viewed. Fukushima et al. [9] and Yamasaki et al. [10] used presentation data of TED Talks to predict audience impressions. The former used 5 types of representation methods, including Bag-of-Words (BoW) [21], Latent Semantic Indexing (LSI) [22], Latent Dirichlet Allocation (LDA) [23], and word2vec [24], as well as surface-level features to extract linguistic features, while using openSMILE [25] for acoustic features. For feature fusion and classifiers, they concatenated all features together, and fed it to a linear SVM for final prediction. The latter used similar features and input them to SVMs individually. However, they considered the correlation between different features within a single MRF.

The SVM-MRF method used by Yamasaki et al. has a disadvantage that unimodal classifiers are only optimized for unimodal classification tasks, rather than for multimodal task. The parameters of SVMs remain unchanged in the stage of feature fusion. Furthermore, it is relatively inefficient, because it requires to use many word embedding methods and many classifiers. The dimension size of some features they used, such as in BoW (43,408 dim) and openSMILE (6,373 dim) are relatively large, which led to high complexity.

2.2.2. Document Classification

The main challenge we have to face is that real presentation documents usually consist of more than one thousand words. Conventional deep learning models for document classification were designed for much shorter sequences [26, 27, 28]. For example, popular open datasets for the document classification, including Amazon Reviews [6], Yelp 2015 [4], and IMDB Reviews [5], have much smaller lengths than the dataset we used. The average document length of: Amazon, Yelp 2015, and IMDB review is 91.9 words, 151.9 words, 325.6 words, respectively. For comparison, the presentation dataset we used has the average length of 1819.2 words. Most Transformer-based models [29, 30] can only officially support the classification task limited to document length of 512 words or less.

[4] https://www.yelp.com/dataset

[5] https://www.imdb.com

2.2.3. Audio Classification

Niebuhr showed how speech reduction shape the impression that speakers create on audiences [31]. Researches on audio classification are similar to those of document classification in that most works focused on classifying short time audios [32, 33, 34], which last for only several seconds. The datasets that most models concentrated on ESC-50 [35] and UrbanSound8K [36], contain audio of less than 5 seconds and less than 4 seconds, respectively. We find that these models are very difficult to be adapted to our study, because every presentation in our dataset lasts for 825.2 seconds on average. Raw audio data is especially difficult to adapt, because there are over 10 thousand samples in one second of raw audio data. Therefore, it will be very difficult if we simply treat presentation audios as sequence data. As one solution of the long sequence problem, some researchers started to use audio mel-spectrograms as the input [37, 38]. However, most researches using audio mel-spectrograms are in a limited field, acoustic scence recognition.

2.3. Unimodal Feature Learning

The first stage of our proposal is to use two independent networks, one learns linguistic representation of video captions, and the other learns acoustic representation of audio data separated from the video.

2.3.1. Linguistic Feature Learning

Word Embedding

We apply the weight of word2vec model, pretrained on part of the Google News dataset [24], as word embedding layer. In order to ensure the training speed, we fix the parameters of the word embedding layer to exclude it from the backward propagation. In addition, we only consider words that exist in the vocabulary of the pretrained word2vec model.

Time-related Linguistic Feature

To handle long sequence problem, we use an LSTM with an attention mechanism to extract time-related linguistic features (Figure 2.2). We input embedded word vectors into the LSTM and get the LSTM’ s output sequence as the time-related linguistic representation based on all input content. Obviously, the importance of all of the LSTM’ s output vectors is not equal. Therefore, we need to use the attention mechanism to find which of them are more relevant to our task. We then assign LSTM’ s output vectors with the adapted weights to produce time-related linguistic representation.

The most common attention functions are additive attention [39] and dot-product attention [40]. They have similar theoretical complexity and performance. Additive attention uses a feed-forward network with a single hidden layer. Dot-product attention uses matrix multiplication, and it is implemented using highly optimized code. It is faster and more space-efficient when compared to additive attention in practice, and it is the reason why we choose dot-product attention. More details are introduced in Section 2.3.3.

As the activation function in the normalization stage of attention weights, we choose a sigmoid function instead of an exponential function, because sigmoid functions are proven to be more suitable when handling very long sequences [41]. A sigmoid function can decrease the possibility that the model focuses on limited feature vectors [41], and this function can slightly improve the performance of our model. We pretrain the LSTM and the attention network on the presentation dataset for the impression prediction task, and we use a matrix with bias as a high-speed classifier.

2.3.2. Acoustic Feature Learning

Temporal Acoustic Feature

For raw audio data, it is common for 10 thousand samples to exist in just one second of data, according to the sampling rate. Even though it is treated as a sequence classification problem for most models, used data is often no more

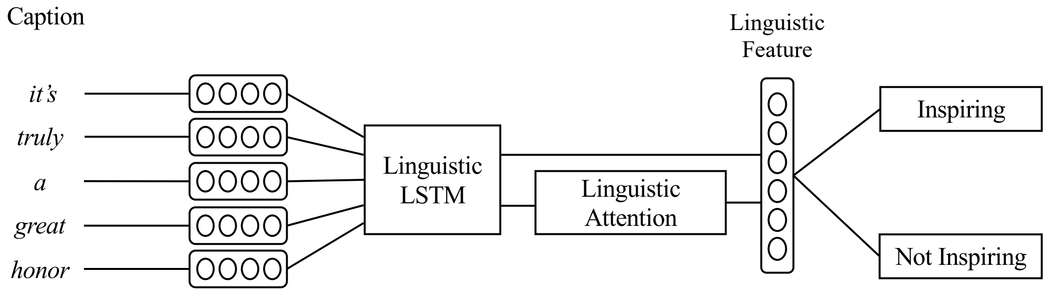


Figure 2.2 – Linguistic feature learning

than several seconds. In contrast, the dataset we use, TED Talks, consists of presentations lasting over 10 minutes. Therefore, it will be very difficult if we take the classification of presentation audios as a sequence classification problem.

In order to handle the long-duration audio data, we need to adapt to the LSTM's capacity of modeling long-term dependency. Therefore, we propose a method to represent audio data in another way. First, audio tracks have different lengths, so we need to split them into segments with the same length. There are several options for the fixed length of audio segments. We evaluate them individually and choose the one that achieves the best performance for later stage of multimodal feature fusion.

After getting the audio segments with the fixed length, we transfer them into frequency domain by Short-Term Fourier Transform (STFT) with the window size of 1024. We only keep the part in the frequency range between 85 Hz and 4,000 Hz, because there is rarely any energy signal out of this range in audio mel-spectrograms. We also take signals with excessively low energy as environment noise and exclude them from the mel-spectrograms. The resulting audio mel-spectrograms are then output as images with a pixel resolution of 224 x 224 (Figure 2.3).

Last, we use a high-performance Deep Convolutional Neural Network (DCNN), ResNet-50 [42], to extract deep visual features. We input mel-spectrogram images and take out the values of nodes immediately prior to the fully connected layer. We take these 2048-dimension vectors as temporal acoustic features, even though they are extracted from mel-spectrogram images.

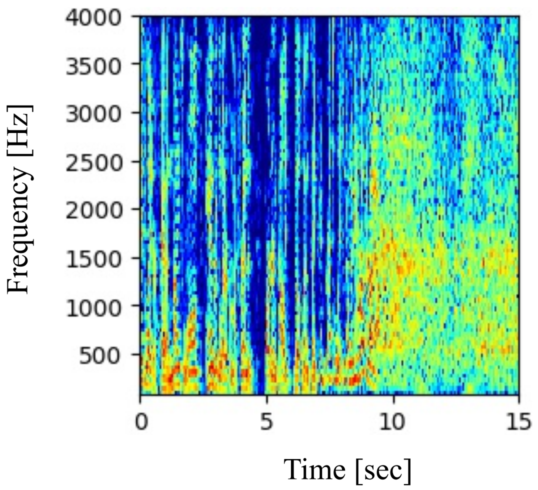


FIGURE 2.3 – An example of audio mel-spectrogram

With the consideration of efficiency, we do not include DCNN in backward propagation. However, due to the high-performance of modern DCNN, acoustic representation can achieve good performance even after several preprocessing that we transfer raw audio data to frequency domain and extract the deep feature.

Time-related Acoustic Feature

After getting temporal acoustic features, we need to embed the feature sequence and model the dependency between them to extract an acoustic feature for the whole presentation. The method of learning time-series audio features is very similar to the method used for extracting time-related linguistic features. We also use an LSTM and assign the LSTM' s output vectors with corresponding weights calculated by dot-production attention (Figure 2.4). The only difference is that we choose the conventional exponential function as our activation function for attention weights normalization, because the length of temporal acoustic feature sequence is relatively short. For example, if we choose 30 seconds as the segment length, there will be only 27.4 audio segments on average for every presentation. Therefore, we do not need to worry about the disproportionate attention which occurs more frequently in a relatively long sequence task.

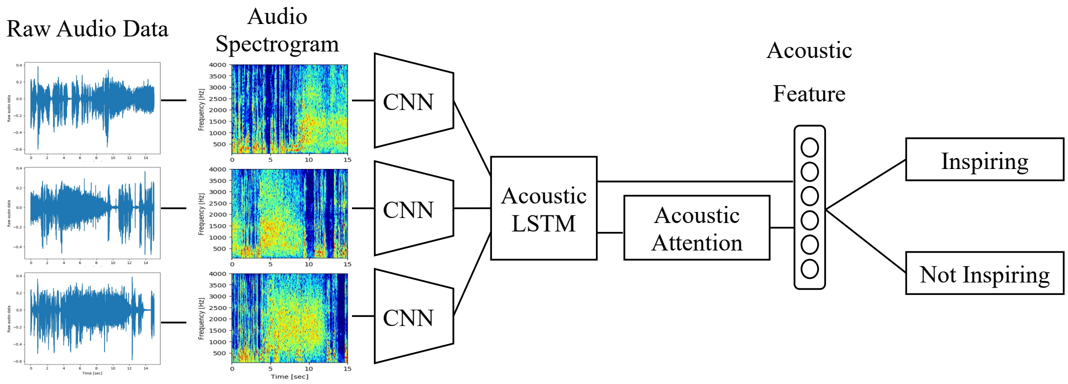


Figure 2.4 – Acoustic feature learning

2.3.3. Attention Approach

Dot-Product Attention

We apply the relatively fast and space-efficient dot-product attention to extracte both linguistic features and acoustic features, used to calculate the weight of each feature vector. There is a context matrix shared by all feature vectors, which acts as a “questioner” to ask feature vectors where the important parts are. The attention network will then get the weight for each feature vector according to their “answers”. We then need a normalization function to make all weights sum up to one. After that, normalized weights will be assigned to corresponding feature vectors, and we will get the final linguistic or acoustic feature representation with time-related information.

Smoothing Normalization

The SoftMax function is commonly used as a normalization function in attention networks:

$$weight_i = \frac{\exp(e_i)}{\sum_{i=1}^{T} \exp(e_i)},$$

where e_i is the i-th feature vector among a set of T feature vectors.

The activation function in SoftMax is an exponential function. However, some researchers note that the exponential function will lead attention work to focus

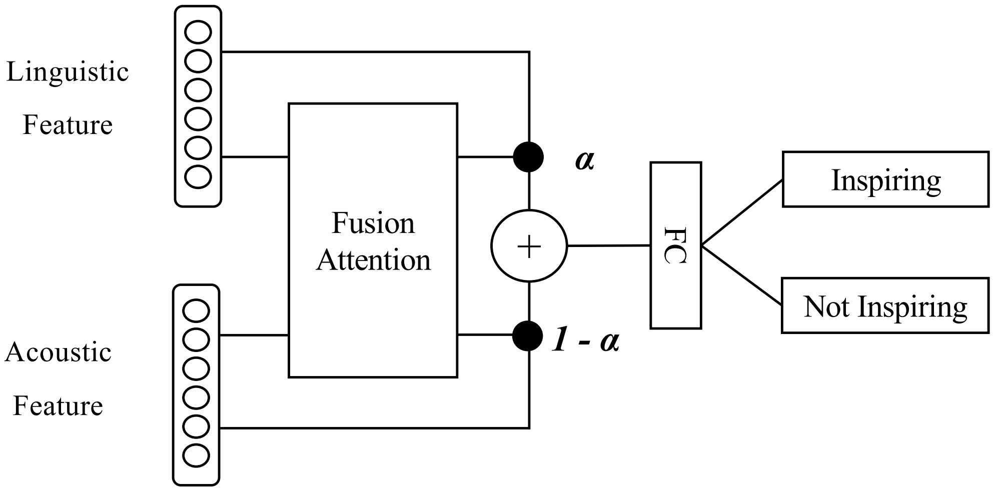


Figure 2.5 – Feature fusion network

on only a few feature vectors, therefore the sigmoid function is more suitable for handling very long sequences.

The smoothing normalization with the logistic sigmoid function σ will then be:

$$weight_i = \frac{\sigma(e_i)}{\sum_{i=1}^{T} \sigma(e_i)}.$$

We try both the exponential function and the sigmoid function as normalization functions in our experiments. We find that using a sigmoid function can slightly improve the performance of our linguistic model.

2.4. Multimodal Feature Fusion

For feature fusion, we use a shared attention network to assign weights for linguistic features and acoustic features. We use the SoftMax function to make these two weights sum up to one (Figure 2.5). By this method, we can clearly understand the importance relationship between the content of the presentation (linguistic) and the presenter's manner of speaking (acoustic) for different audience impressions.

After feature fusion, we input the multimodal feature into a 3-layer MLP for final impression prediction. We can then get the conditional probability of the

Table 2.1 – Official caption samples

No.	Content
1	She took our order, and then went to the couple in the booth next to us, and she lowered her voice so much, I had to really strain to hear what she was saying...
501	There are several scenarios for the future newspaper. Some people say it should be free; it should be tabloid, or even smaller...

targets given by the combination of linguistic features and acoustic features. With the conditional probability of each target, we use negative log likelihood loss to compare them with the ground truth of positive target or negative target. While maximizing the likelihood, we not only adjust feature fusion network but also fine-tune the unimodal networks, because we want to get the most suitable feature representation for multimedia data in both the linguistic and acoustic fields rather than only in one of them.

2.5. Experiments

2.5.1. Dataset

We choose to apply our method on presentation dataset of TED Talks because there are more than 3,000 videos in TED Talks. We eliminate non-oral-presentation types of talks such as playing music, magic shows, and so on. As a result, we collect 2,445 videos from TED Talks as the experiment dataset. All of

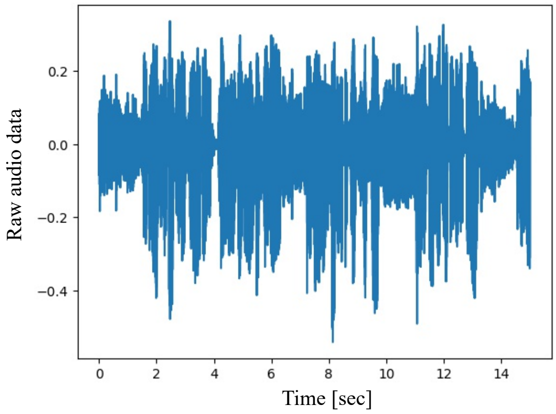


Figure 2.6 – Raw audio data sample

them have official captioning and 14 tags. Each tag tells us the amount of audiences votes for each impression, including *Beautiful, Confusing, Courageous, Fascinating, Funny, Informative, Ingenious, Inspiring, Jaw-dropping, Longwinded, Obnoxious, OK, Persuasive,* and *Unconvincing.*

We rank all videos according to the vote rate of each impression, creating 14 ranking lists in total. Only videos in the top 30% and bottom 30% in each ranking list are used for impression prediction (40% in the middle are not used in our experiments). It becomes a binary classification task with only prediction of positive instances (top 30%) and negative instances (bottom 30%).

The TED Talks dataset includes visual data (videos), acoustic data (raw audio data extracted from videos, Figure 2.6), and linguistic data (captions, Table 2.1). We only use the latter two types of data as input, because the camera view of TED Talks videos often changes. Sometimes it is toward the speakers, sometimes it is toward the slides, and sometimes it is toward the audiences, so handling the video data is very difficult. The sampling rate of the raw audio data is 44.1 kHz. On average, TED Talks presentations last for 825.2 seconds (13 minutes and 45.2 seconds) with 1819.2 words of English captioning.

2.5.2. Training

Each impression uses a different dataset because we only use videos in the top and bottom 30% of each corresponding impression. We split each of them into training, validation, and test datasets with an $8 : 1 : 1$ proportion for 10-fold cross validation. The linguistic LSTM, acoustic LSTM, and all attention layers are 300 dimensions each. We train our network by using the optimizer Adam [43] with a batch size of 64. Different particular learning rates are used during different stages. The learning rate is 5×10^{-5} when pretraining the linguistic network and the acoustic network, but it becomes to be 1×10^{-5} during training of the feature fusion network.

2.5.3. Comparative Methods

The TED Talks dataset is used to predict audience impressions. The average of all word vectors in each document is taken as linguistic feature representation. Several word embedding methods, such as BoW, word2vec, etc., are evaluated on the dataset and many of them demonstrate high performance.

Using openSMILE is a common way to extract acoustic feature. We use openSMILE with the configuration of the INTERSPEECH 2013 Computational Paralinguistic Challenge [25]. The acoustic feature as extracted by openSMILE, is a 6,373-dimensional vector including the information of the pitch, loudness, voice quality, energy, MFCC, etc.

The correlation between different types of features can be considered within a single MRF for impression prediction. We also evaluate a baseline method that simply adding or concatenating features together.

2.6. Results and Analysis

For the impression prediction on the TED Talks dataset, we evaluated our unimodal neural network when using only either linguistic features or acoustic features, as well as a multimodal network after feature fusion. We then compared the performance of our proposal against the comparative methods introduced in Section 2.5.3.

Table 2.2 shows the classification accuracies of our linguistic feature extraction network and Fukushima’ s results of using BoW, LSI, LDA, and word2vec, respectively. The experiment results of our proposal were very similar with BoW, but our linguistic feature has only 300 dimensions, which is much lower than the 43,408 dimensions of BoW. The small size of the feature vector can largely improve the efficiency of training. In addition, for attention normalization functions, we proved that using a sigmoid function can improve the performance of our linguistic model.

We evaluated our acoustic network and compared it with openSMILE, a commonly used tool for acoustic feature extraction. The results show that our model

Table 2.2 – Prediction accuracies of linguistic features

	beau.	conf.	cour.	fasci.	funny	info.	inge.	insp.	jaw-d.	longw.	obnox.	ok	pers.	unconv.	ave.
BoW [9]	**91.1%**	**71.9%**	88.7%	89.1%	83.2%	88.3%	85.9%	87.5%	80.5%	**73.9%**	68.8%	72.8%	91.6%	**73.3%**	**81.9%**
LSI [9]	88.6%	70.8%	88.0%	88.4%	80.0%	87.1%	84.3%	86.2%	78.7%	71.9%	66.7%	66.9%	91.1%	72.7%	80.3%
LDA [9]	84.5%	70.6%	84.4%	86.0%	78.3%	84.4%	82.0%	82.6%	72.9%	69.8%	64.6%	66.6%	88.7%	71.6%	77.6%
Word2vec [9]	90.7%	71.5%	87.3%	88.8%	80.2%	88.4%	87.1%	87.5%	80.1%	73.1%	68.1%	**73.3%**	91.6%	72.3%	81.3%
Exponential function	90.4%	69.6%	87.7%	**91.8%**	**89.0%**	87.7%	86.3%	87.0%	78.8%	66.4%	69.2%	67.8%	**92.5%**	**73.3%**	81.3%
Sigmoid function	90.2%	70.5%	**90.6%**	89.8%	81.8%	**90.1%**	**87.5%**	**87.8%**	**81.9%**	72.5%	**69.3%**	68.2%	91.2%	**73.3%**	**81.8%**

Table 2.3 – Prediction accuracies of acoustic features

	beau.	conf.	cour.	fasci.	funny	info.	inge.	insp.	jaw-d.	longw.	obnox.	ok	pers.	unconv.	ave.
OpenSMILE [9]	**73.5%**	59.4%	69.3%	65.9%	74.7%	**76.4%**	64.2%	66.3%	**66.8%**	73.7%	60.6%	69.9%	70.7%	63.8%	68.2%
15 seconds	67.8%	**63.7%**	**74.0%**	**69.2%**	**78.8%**	69.9%	**68.5%**	**69.2%**	64.4%	74.0%	62.3%	**76.0%**	76.0%	**65.8%**	**70.0%**
30 seconds	65.1%	61.0%	73.3%	62.3%	74.0%	72.6%	63.7%	63.0%	65.8%	**75.3%**	63.0%	75.3%	**76.7%**	61.6%	68.1%
45 seconds	65.8%	56.2%	71.2%	66.4%	71.9%	73.3%	67.1%	61.0%	61.6%	72.6%	**65.8%**	68.5%	76.0%	61.6%	67.1%
60 seconds	61.6%	61.0%	71.2%	61.0%	69.2%	75.3%	67.1%	60.3%	64.4%	72.6%	63.7%	71.9%	74.7%	61.6%	66.8%

Table 2.4 – Prediction accuracies of multimodal features

	beau.	conf.	cour.	fasci.	funny	info.	inge.	insp.	jaw-d.	longw.	obnox.	ok	pers.	unconv.	ave.
Linguistic feature only	90.2%	70.5%	90.6%	89.8%	81.8%	90.1%	87.5%	87.8%	81.9%	72.5%	69.3%	68.2%	91.2%	73.3%	81.8%
Acoustic feature only	67.8%	63.7%	74.0%	69.2%	78.8%	69.9%	68.5%	69.2%	64.4%	74.0%	62.3%	76.0%	76.0%	65.8%	70.0%
Add	90.9%	71.3%	91.6%	90.7%	89.3%	**90.7%**	87.5%	89.1%	86.8%	76.5%	**72.0%**	77.0%	93.0%	73.8%	84.3%
Concatenate	91.1%	70.1%	90.7%	90.9%	88.6%	90.2%	87.7%	89.1%	**87.0%**	77.5%	69.0%	76.8%	93.2%	73.6%	84.0%
SVM-MRF [10]	89.5%	**71.8%**	90.2%	90.8%	82.9%	89.4%	85.8%	88.5%	82.6%	77.7%	71.2%	**77.1%**	92.4%	74.4%	83.1%
Shared attention network	**92.5%**	69.2%	**91.8%**	**91.1%**	**91.1%**	90.4%	**88.4%**	**91.8%**	86.3%	**79.5%**	71.2%	74.0%	**93.8%**	**78.8%**	**85.0%**

has competitive performance when compared to openSMILE. Several options of segment length were compared, including 15 seconds, 30 seconds, 45 seconds and 60 seconds. The best result was achieved when evaluating the with the performance with the lengths of 15 second (Table 2.3). We then used acoustic features with this segment length for feature fusion.

Our unimodal neural network only had a small advantage compared to using linguistic feature alone. However, for some impressions, the performance of our multimodal feature performed much better than only using the linguistic feature, which was further improved by our high-quality acoustic feature and feature fusion network. For example, the prediction accuracy of *Longwinded* was improved from 72.5% to 79.5% after feature fusion.

We got a noticeably better result compared to Yamasaki' s model (Table 2.4), because our unimodal neural network was able to be fine-tuned during feature fusion. In addition, we only utilized the word2vec embedding method and we only needed to build 2 classifiers while preparing the linguistic feature and the acoustic feature. It is a very noticeable improvement to the efficiency of the impression

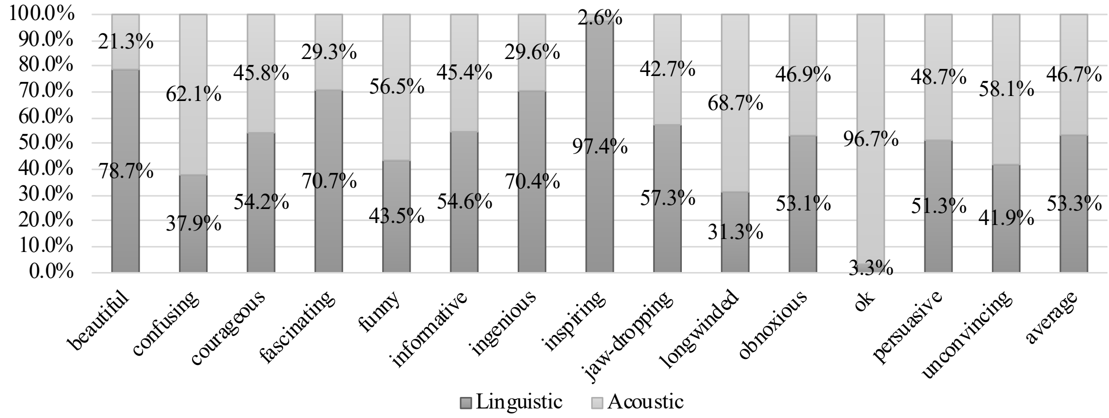


Figure 2.7 – Average attention

prediction system.

For the feature fusion network, we obtained the attention weights of the linguistic feature and the acoustic feature, which stood for presentation content and the presenter's manner of speaking, respectively (Figure 2.7). We could then clearly understand their importance relationship for considering different audience impressions. For example, the content of the presentation is much more important than the presenter's manner of speaking when finding whether audiences consider the presentation to be *Inspiring*. In contrast, if audiences think the presentation is just *OK*, they usually pay attention to how speaker talks. As a whole, audience impressions are more relevant to the content of a presentation than to the presenter's speech mannerisms.

Furthermore, we find that the average vote rates (Figure 2.8) of different impressions are strongly related to the prediction accuracies (Figure 2.9). For the impressions with low average vote rates, including *Confusing*, *Longwinded*, *Obnoxious*, *OK*, and *Unconvincing*, their prediction accuracies are also relatively lower than the impressions with high vote rates.

For the prediction example, Bill Gates has given 6 presentations in TED Talks. Figure 2.10 is the prediction result of one of his presentations. We correctly predicted thirteen impressions. For another example shown in Figure 2.11, for the former American vice-president Al Gore' s presentation, we correctly predicted for twelve impressions.

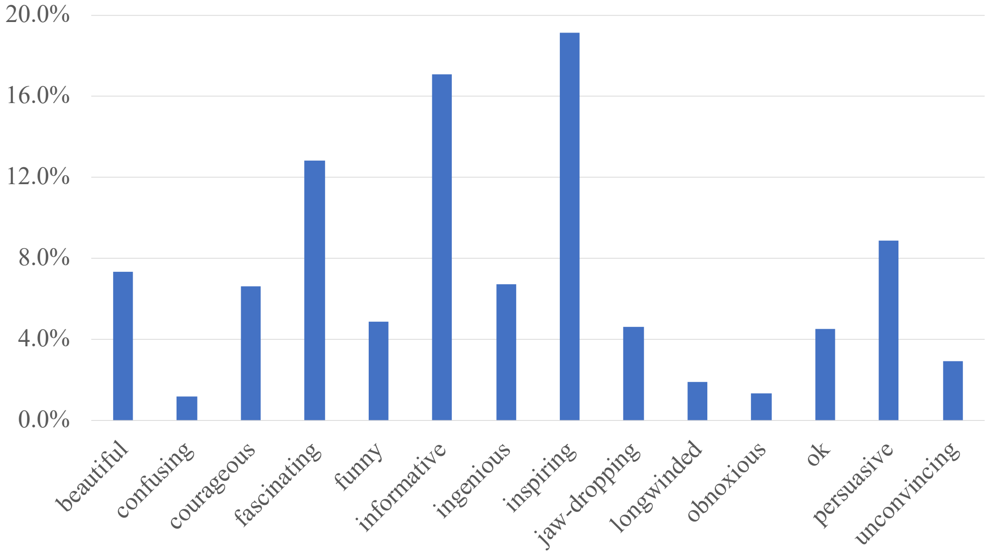


FIGURE 2.8 – Average vote rates

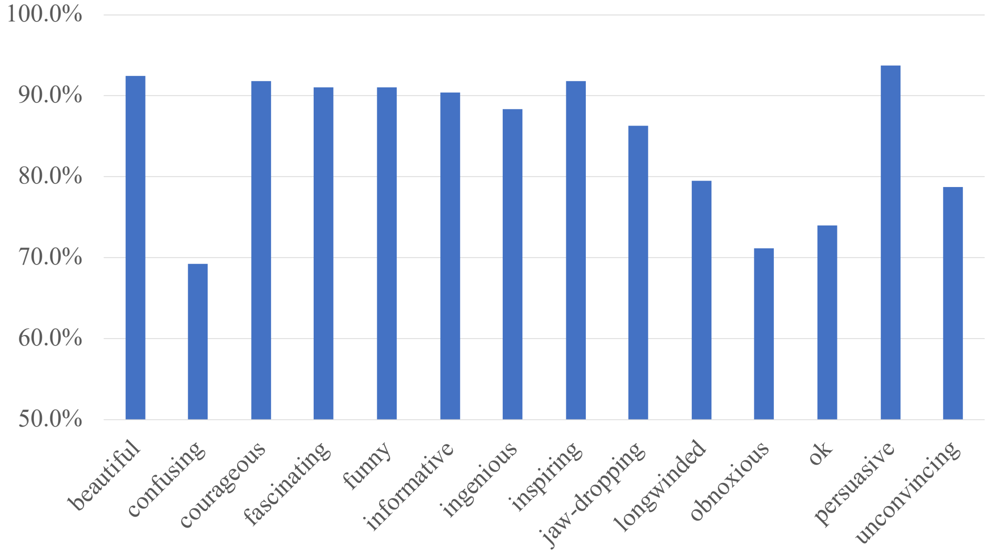


FIGURE 2.9 – Prediciton accuracy

2.7. Conclusion and Future Work

In our study, we successfully apply our proposal on impression prediction to the TED Talks dataset. Our proposal includes two LSTM-based unimodal neural networks for learning linguistic features and acoustic features, respectively. Notably, we successfully transfer the representation of long-duration raw audio data of presentations to adapt them to the capacity of LSTM for modeling long-term dependency. The results show that our proposal achieves high performance with a prediction accuracy of 85.0% for the top and bottom 30% ranked data, and

	Ground Truth	Prediction
beautiful	NO	NO
confusing	YES	YES
courageous	YES	YES
fascinating	NO	NO
funny	NO	NO
informative	YES	YES
ingenious	NO	NO
inspiring	NO	NO
jaw-dropping	YES	NO
longwinded	YES	YES
obnoxious	YES	YES
ok	YES	YES
persuasive	YES	YES
unconvincing	YES	YES

How state budgets are breaking US schools

Correct Total

13/14

FIGURE 2.10 – The prediction result of a presentation by Bill Gates

	Ground Truth	Prediction
beautiful	NO	NO
confusing	YES	YES
courageous	YES	YES
fascinating	NO	NO
funny	NO	NO
informative	YES	YES
ingenious	NO	NO
inspiring	YES	YES
jaw-dropping	YES	NO
longwinded	YES	YES
obnoxious	YES	YES
ok	NO	YES
persuasive	YES	YES
unconvincing	YES	YES

New thinking on the climate crisis

Correct Total

12/14

FIGURE 2.11 – The prediction result of a presentation by Al Gore

a significant improvement to the efficiency of the impression prediction system. Furthermore, we can clarify the importance relationship between the content of a presentation and the presenter's manner of speaking for different audience impressions using an attention network.

In the future, we plan to add the correlation between different impression labels. We also want to search for more videos resources and use the data in different domains for our impression prediction system by using the transfer learning method.

Chapter 3

Press Conference Evaluation

3.1. Introduction

Press conference is a special type of presentation for public relations issues. Press conferences are held in important occasions such as new political actions, new inauguration of Presidents/Governors/CEOs, scandals, and so on. Press conferences have a power to change public opinions. Therefore, training speakers in advance is important and there are some consulting farms for such purposes. In this chapter, we would like to introduce our automatic evaluation system for press conferences that can simulate the professional consultant' evaluations.

In the press conferences, the speakers need to answer journalists' questions. Professional consultants evaluated these answers based on several criterions rather than simply giving them scores. For each criterion, we classified the speech into three classes, including positive, neutral, and negative. Then we built an end-to-end system to automatically predict the evaluation of the consultant using text only.

Our system uses a Long Short-Term Memory (LSTM) with self-attention mechanism. Different from our research of presentation, the samples of press conferences are in the format of Q&A, and the sequences are relatively shorter. We represent each word by token embedding using Embeddings from Language Models (ELMo), assigning tokens representation based on the entire sentences, as well as type embedding, annotating the position of the word.

We applied our proposal on the press conference dataset, containing publicly available press conference videos. We used the speakers' answers only or the speakers' answers along with journalists' questions to predict the evaluation of the consultant. As a result, we achieved the average accuracy of 57.6% for the prediction of 11 evaluation criterions.

3.2. Related Work

3.2.1. Text Classification

The LSTM architecture [11] was motivated by the backpropagated error of Recurrent Neural Network (RNN), blows up or decays exponentially. Now, LSTMs are widely used for text classification tasks [44, 45]. LSTMs can be stacked together to extract features in both low and high levels [46] that can bring performance improvement. LSTMs are also used in text-based speech evaluation tasks [47], and they can achieve better performance than statistical machine learning methods [10]. Different from traditional text classification and speech evaluation tasks, the text sample of the evaluation of press conference includes sentences from two roles, the journalist and the speaker. Their speaking varies in length, style, content, etc.

3.2.2. Conversation Analysis

The Convolutional Neural Network (CNN) architecture is widely used in conversation analysis models [48, 49]. Conversations usually consist of relatively short sentences. CNN can usually achieve higher performance in these tasks. However, as a special example of conversation, the Q&A pairs of press conference consist of relatively longer sentences. In either the question part or the answer part, the sentences include many details that describe the major incident.

3.2.3. Word Representations

Word representations can help learning algorithms to achieve better performance by grouping similar words. However, there are relatively limited number of documents in many datasets that can't support the learning algorithms to get

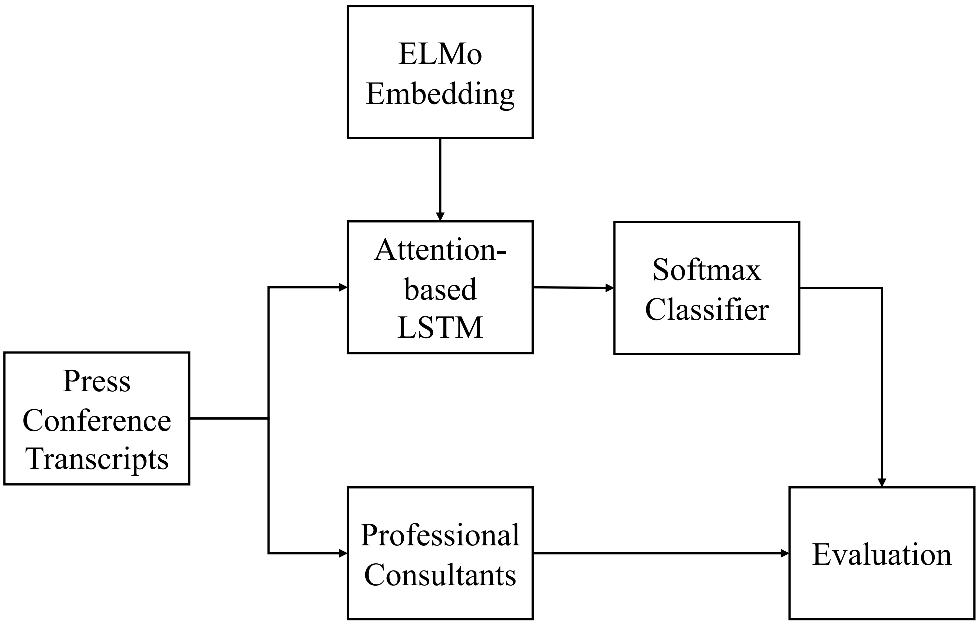


FIGURE 3.1 – Overview of the method

a good word representation. Therefore, pretrained word representations have become more and more necessary.

The training objective of the Skip-gram architecture is to find word representations that are useful and suitable for predicting the surrounding words [24]. The Glove algorithm is based on co-occurrence matrix [50]. The functions of these traditional word representations only consider the co-occurrence of the words and the context words. On the other hand, Embeddings from Language Models (ELMo) assign each token a representation based on a function of the entire sentence [51].

3.3. Automatic Evaluation Method

We show the overview of our method in Figure 3.1. In this section, we will talk about the language model and the word representation of our method.

3.3.1. Language Model

Our language model is an LSTM with self-attention mechanism (Figure 3.2). LSTM uses a memory unit and a gate mechanism to capture and update the information in memory cell. By the memory mechanism, LSTM can capture relatively

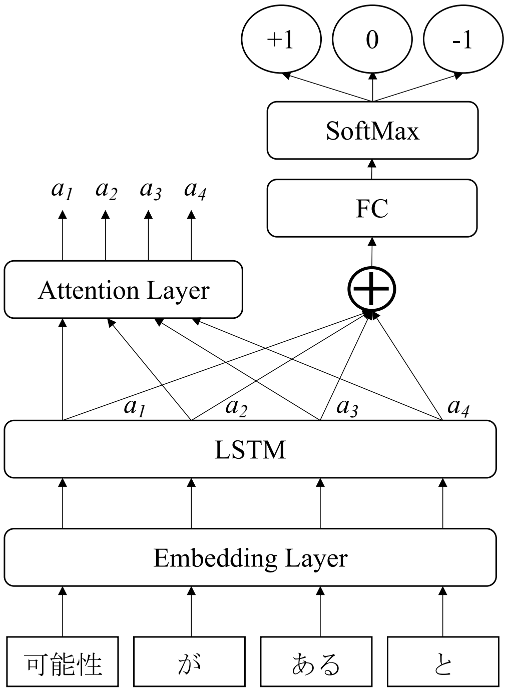


FIGURE 3.2 – LSTM with self-attention mechanism

Input	可能性	が	ある	と	[SEP]	可能性	が	ある
Token Embedding	E_{可能性}	E_が	E_{ある}	E_と	E_[SEP]	E_{可能性}	E_が	E_{ある}
	✚	✚	✚	✚	✚	✚	✚	✚
Type Embedding	E_Q	E_Q	E_Q	E_Q	E_Q	E_A	E_A	E_A

FIGURE 3.3 – Word representation

long-distance information in a sequence. Let us denote $s = [x^{(1)}, x^{(2)}, \ldots, x^{(T)}]$ as the sequence of the input word vectors with the length of T. At the t-th step, LSTM layer updates the network as follows:

$$i^{(t)} = \sigma(W_i x^{(t)} + U_i h^{(t-1)} + b_i),$$

$$f^{(t)} = \sigma(W_f x^{(t)} + U_f h^{(t-1)} + b_f),$$

$$o^{(t)} = \sigma(W_o x^{(t)} + U_o h^{(t-1)} + b_o),$$

$$\tilde{c}^{(t)} = \tanh(W_c x^{(t)} + U_c h^{(t-1)} + b_c),$$

$$c^{(t)} = f^{(t)} \odot c^{(t)} + i^{(t)} \odot \tilde{c}^{(t)},$$

$$h^{(t)} = o^{(t)} \odot \tanh(c^{(t)}),$$

where σ is the logistic sigmoid function. Operator $\odot$ is the element-wise multiplication operation. W, U, and b represent the recurrent weights, input weights, and bias. $i^{(t)}$, $f^{(t)}$, $o^{(t)}$, $\tilde{c}^{(t)}$, $c^{(t)}$, and $h^{(t)}$ respectively represent the input gate, forget gate, output gate, candidate cell state, cell state, and hidden state at step t.

We use an attention mechanism [39] to assign each word with an appropriate weight, because not all words contribute equally to the prediction of the evaluation. The most widely used attention approach is self-attention [27]. We use a three-layer Multilayer Perceptron (MLP) without bias to construct the self-attention layer, and we use the SoftMax function for normalization:

$$y^{(t)} = \tanh(W_1 h^{(t)}),$$

$$a^{(t)} = \frac{\exp(W_2 y^{(t)})}{\sum\limits_{i=1}^{T} \exp(W_2 y^{(i)})},$$

where W denote the fully connection weights. y, h, and a denote the hidden layer of MLP, hidden state of LSTM, and attention score, respectively.

By the attention mechanism, we determine the importance of each word as well as its corresponding LSTM hidden state. Then we represent the sentence by the combination of all LSTM hidden states with considering their importance weights:

$$v_s = \sum_{i=1}^{T} a^{(i)} h^{(t)}.$$

After we get the representation for each sentence, we use a linear layer and the SoftMax function as classifier. We map the sentence representation to the probabilities that the sample is classified into positive (+1), neutral (0), or negative (-1) for each evaluation criterion. The prediction of each evaluation criterion is operated individually.

3.3.2. Word Representation

The word representation of our proposal is the sum of the token embedding and type embedding (Figure 3.3). The token embedding method we used is ELMo. Unlike traditional word embeddings, ELMo is the function of the entire sentence. It uses character convolution and a two-layer Bidirectional LSTM (BLSTM) that is pretrained on large datasets. The type embedding only includes two representations that annotate the position of the word, whether it is in journalist's question or speaker's answer.

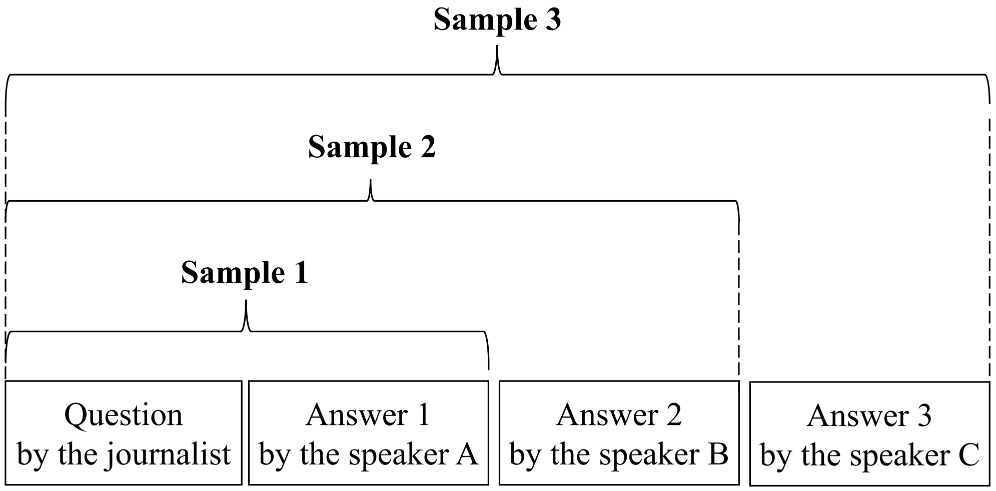

Sample 3
Sample 2
Sample 1
Question by the journalist
Answer 1 by the speaker A
Answer 2 by the speaker B
Answer 3 by the speaker C

Figure 3.4 – Sample construction

Table 3.1 – Press conferences

ID	Conference	Date
1	無資格検査問題でスバルが記者会	2017/12/19
2	不正行為について調査結果を公表 神戸製鋼が会見	2017/11/10
3	神戸製鋼のデータ改ざん問題 川崎社長が辞任表明	2018/3/6
4	スバルがデータ改ざんで新たな不適切行為　吉永社長が会見	2018/6/5
5	バスケ男子代表が不適切な行動で処分 帰国した4選手が謝罪	2018/8/20
6	免震データ改ざん問題 KYBが建物70件を公表	2018/10/19
7	東京医大が記者会見 入試不正で不合格になった受験生の救済は？	2018/11/7

3.4. Experiments

3.4.1. Dataset

We collected seven press conferences videos (Table 3.1) that lasted for 12 hours in total, and then we created the speech transcripts manually. All of these press conferences were held for giving a public apology, such as the apology for the data alteration problem of Kobe Steel, Ltd. The press conferences are usually held in the form of Q&A. After the journalist raised a question, several speakers need to answer the question in succession. Every speaker's answer was strongly connected to what early speakers had said. Therefore, we should not consider an answer independently without the contents before it. For one question, we constructed

Table 3.2 – Evaluation criterions

ID	Criterion
1	Showing the feeling of apology
2	Suitable tone of the voice for an apology
3	Speaking in an appropriate speed
4	Answering immediately
5	Not just using the journalist's words
6	Explaining briefly and easy to understand
7	Explaning the situation based on the facts
8	Speaking honestly if they really don't know
9	Answering assumption question appropriately
10	Not answering emotionally
11	Not expressing personal opinion

several Q&A pair samples according to the number of speakers who answered the same question. Each sample included the question by the journalist, and the answers by the speakers who had given a speech (Figure 3.4). Finally, we collected 1,050 Q&A pairs. We tokenized the corpus by MeCab [1] with IPADIC neologism dictionary [2].

The professional consultant in our team evaluated the Q&A pair samples based on 11 criterions. These criterions are specially designed for press conference that can reflect the speakers' performance roundly (Table 3.2). For each criterion, the sample is classified into three classes, including positive, neutral, and negative.

3.4.2. Training

We designed two prediction tasks. Task A used the speakers' answers only to predict the evaluation of the consultant. Task B used the Q&A pairs that include both journalists' questions and the speakers' answers for prediction. For task B using the speakers' answers only, some samples were considered to be "meaningless" when leaving out the questions like follows (translation):

Yes, it is. / No, it is not.

[1] taku910.github.io/mecab

[2] github.com/neologd/mecab-ipadic-neologd

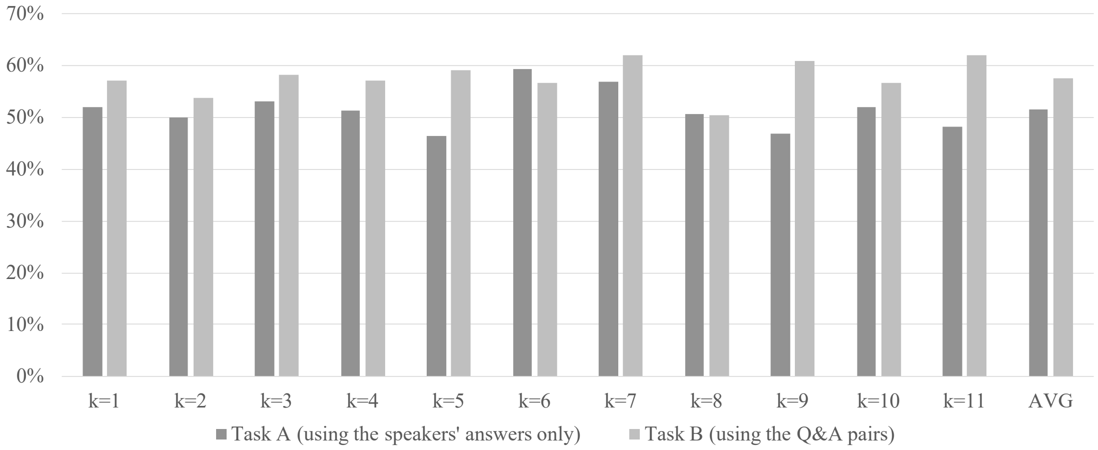


Figure 3.5 – Prediction Accuracies

Sorry, can you repeat the question?

Thank you very much!

I am terribly sorry about that.

Therefore, we deleted these "meaningless" samples, and there remained 810 samples in the dataset of task B.

The splits of the dataset for task A and task B were the same. The proportion of training, validation, and test dataset was 3 : 1 : 1. We loaded ELMo that was pretrained on the Japanese Wikipedia [52] for token embedding and trained our language model on press conference dataset. The dimension of the token embedding and type embedding were 1,024, and the dimension of LSTM hidden states was 512. The optimizer we used was Adam [43] with learning rate of 10^{-5}.

3.5. Results and Analysis

We automatically evaluated the press conference based on 11 evaluation criterions. For each criterion, we classified samples into three classes (positive, neutral, or negative), and we compared the automatically operated evaluations with the evaluation of the consultant. The prediction accuracy of each criterion, whose IDs are annotated with k, for task A and task B are shown in Figure 3.5. As a result, we achieved the average accuracy of 51.5% and 57.6% for task A and task B respectively. In comparison, we can only get the prediction accuracy of 33.7% and

TABLE 3.3 – Example of prediction result

Question (translation)	Answer (translation)		Criterion ID										
			1	2	3	4	5	6	7	8	9	10	11
Is there a date for the seismic isolation testing?	Yes, there is an inspection report for the seismic isolation testing.	GT	1	1	-1	-1	1	1	-1	0	0	-1	-1
		Task A	1	-1	1	-1	1	-1	1	-1	-1	-1	1
		Task B	1	-1	**-1**	-1	1	-1	1	**0**	**0**	-1	1
Did President Yoshinaga agree on the judgment to continue the shipment?	It ’ s a little difficult to answer.	GT	-1	-1	1	1	1	0	0	1	1	1	1
		Task A	-1	-1	-1	-1	0	0	1	-1	-1	-1	-1
		Task B	-1	-1	**1**	-1	1	0	**0**	-1	**1**	-1	0
Some products were out of specification, is it right?	It’s a fact that we took products out of specification.	GT	-1	-1	1	1	1	1	1	1	1	1	1
		Task A	-1	-1	1	-1	-1	1	1	-1	-1	-1	-1
		Task B	-1	-1	1	-1	**1**	-1	1	-1	-1	-1	**1**

Table 3.4 – Task A

k = 1		Prediction		
GT		-1	0	1
	-1	**48**	8	18
	0	10	**4**	5
	+1	35	2	**32**
k = 2		Prediction		
GT		-1	0	1
	-1	**53**	7	21
	0	8	**4**	5
	+1	39	1	**24**
k = 3		Prediction		
GT		-1	0	1
	-1	**43**	10	25
	0	4	**4**	2
	+1	34	1	**39**
k = 4		Prediction		
GT		-1	0	1
	-1	**55**	8	21
	0	0	**4**	2
	+1	38	1	**24**
k = 5		Prediction		
GT		-1	0	1
	-1	**31**	8	14
	0	13	**13**	7
	+1	36	9	**31**
k = 6		Prediction		
GT		-1	0	1
	-1	**52**	9	17
	0	4	**9**	4
	+1	29	3	**35**
k = 7		Prediction		
GT		-1	0	1
	-1	**25**	4	25
	0	6	**10**	6
	+1	22	7	**57**
k = 8		Prediction		
GT		-1	0	1
	-1	**43**	12	8
	0	20	**14**	6
	+1	27	7	**25**
k = 9		Prediction		
GT		-1	0	1
	-1	**42**	15	5
	0	30	**25**	6
	+1	17	13	**9**

Table 3.5 – Task B

k = 1		Prediction		
GT		-1	0	1
	-1	**68**	9	14
	0	18	**32**	9
	+1	29	10	**20**
k = 2		Prediction		
GT		-1	0	1
	-1	**67**	12	19
	0	14	**32**	9
	+1	33	10	**14**
k = 3		Prediction		
GT		-1	0	1
	-1	**44**	16	23
	0	10	**35**	7
	+1	26	6	**43**
k = 4		Prediction		
GT		-1	0	1
	-1	**60**	14	15
	0	17	**32**	2
	+1	36	6	**28**
k = 5		Prediction		
GT		-1	0	1
	-1	**23**	5	21
	0	17	**57**	10
	+1	27	5	**45**
k = 6		Prediction		
GT		-1	0	1
	-1	**44**	7	22
	0	13	**47**	5
	+1	36	8	**28**
k = 7		Prediction		
GT		-1	0	1
	-1	**26**	4	24
	0	12	**52**	6
	+1	27	7	**52**
k = 8		Prediction		
GT		-1	0	1
	-1	**32**	16	12
	0	22	**60**	6
	+1	34	14	**14**
k = 9		Prediction		
GT		-1	0	1
	-1	**42**	18	4
	0	26	**72**	5
	+1	19	10	**14**

k = 10		Prediction		
GT		-1	0	1
	-1	**66**	13	8
	0	18	**9**	1
	+1	32	6	**9**

k = 11		Prediction		
GT		-1	0	1
	-1	**32**	19	5
	0	20	**29**	9
	+1	18	13	**1**

k = 10		Prediction		
GT		-1	0	1
	-1	**63**	13	7
	0	29	**51**	1
	+1	36	5	**5**

k = 11		Prediction		
GT		-1	0	1
	-1	**39**	19	13
	0	21	**70**	3
	+1	18	8	**19**

33.4% by randomly guessing the labels. The results indicate that the information of the journalists' questions can improve the prediction performance.

Let us show some examples of the prediction results for our test data in Table 3.3. Moreover, we respectively show the confusion matrixes of Task A and Task B in Table 3.4 and Table 3.5. We find that adding the question information can correct some wrong prediction results caused by only considering the answer, especially for the contents with neutral evaluations.

3.6. Conclusion and Future Work

We constructed an automatic evaluation system for press conference by using LSTM and self-attention mechanism. Our proposal achieved good performance even though the tasks are based on relatively complex criterions. The experiment results indicate that the Q&A pairs include more information than only the speakers' answers and they can improve the performance of our model.

For future work, we plan to extend the press conference dataset. For criterion IDs of 2, 3, and 4, these evaluation criterions are supposed to be considered based on audios, but we only used the text data until now. Therefore, we are going to include the using of the audio features in our automatic evaluation system.

Chapter 4

Audio Emotion Recognition

4.1. Introduction

We propose a method for the fundamental research of audio emotion recognition. We want to predict moods and themes conveyed by a music track. Moods are often defined as feelings conveyed by the music (e.g. happy, sad, dark, melancholy, etc.), and themes are associated with events or contexts where the music is suited to be played (e.g. epic, melodic, christmas, love, film, space, etc.).

Image classification performance has improved greatly with the advent of large datasets such as ImageNet [53] using CNN architectures such as VGG [54], Inception [55], and ResNet [42]. There are also many researches on music emotion recognition or music classification using CNN architectures [56, 57]. Even though statistical machine learning (e.g. Support Vector Machine [58] and Random Forest [59]) can still achieve good performance in some tasks, deep learning, especially CNN based method, is more popular and can achieve better performance in most tasks. For large-scale datasets, deep learning is much more practicable than statistical machine learning.

We tried to do emotion and theme prediction using a dataset in three types of audio representations including traditional handcrafted audio features, mel-spectrograms, and raw audio inputs. We only used the mel-spectrograms in our experiments (Figure 4.2).

For audio recognition, the most typical model is CRNN [60] (Figure 4.1) that includes a Convolutional Neural Network (CNN) to extract temporal features and a Recurrent Neural Network (RNN) for aggregation. For this research, we

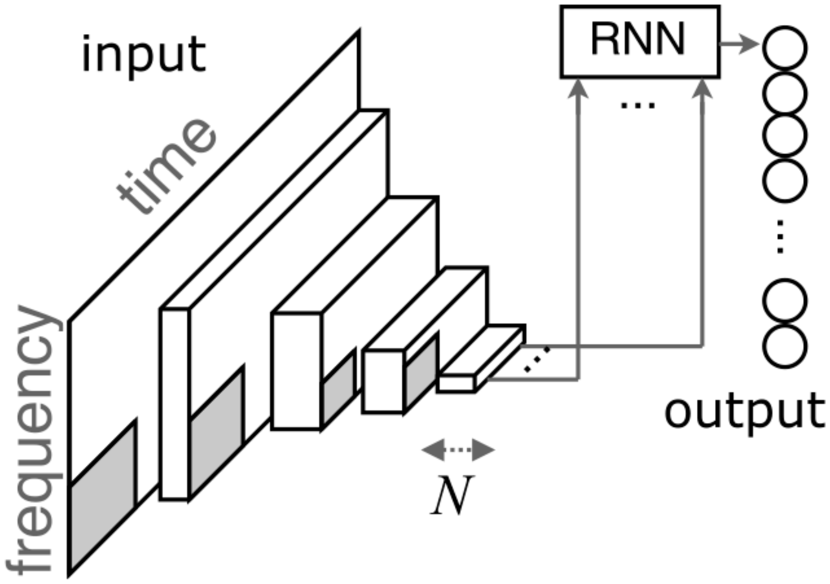


FIGURE 4.1 – CRNN [60]

focus on using CNN to find a suitable model to extract temporal feature for the emotion and theme prediction. A simple but effective model we tried consists of five convolutional layers. We also tried other models with more layers, but they didn't always achieve better results. As a result, the model that achieved the best performance in our experiments is a shallow neural network with six convolutional layers.

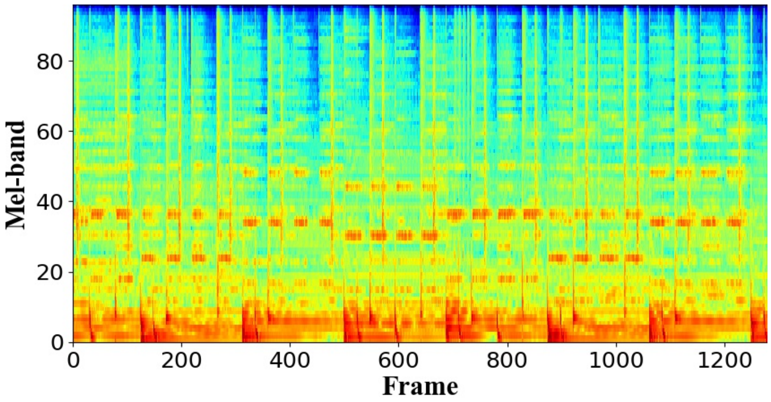


FIGURE 4.2 – Mel-spectrogram

Table 4.1 – The architecture of 6-layer model

Mel-spectrogram	Input: 96x1280x1
Conv 3x3x32	
MP (2, 2)	Output: 48x640x32
Conv 3x3x64	
MP (2, 4)	Output: 24x160x64
Conv 3x3x128	
MP (2, 2)	Output: 12x80x128
Conv 3x3x256	
MP (2, 4)	Output: 6x20x256
Conv 3x3x512	
MP (3, 5)	Output: 2x4x512
Conv 3x3x256	
MP (2, 4)	Output: 1x1x256
Dense	
Sigmoid	Output: 56x1

4.2. CNN Architecture

We concentrated on finding the most suitable CNN architecture for the task. The baseline is a simple but effective model consisting of five convolutional layers and a final dense layer. We also tried other models with deeper architecture. We tried models with 6, 16, 18 or 25 convolutional layers. In particular, the most shallow model we considered is a fully convolution neural network with ELU activations, six 3x3 convolutional layers, and 32, 64, 128, 256, 512, 256 units for each layer respectively (Table 4.1).

We also tried some models with the residual architecture [42]. The convolutional block consists of 1x1, 3x3 and 1x1 convolutional layer sequentially (Figure 4.3). This is the architecture for inputs and outputs with the same size and unit number. For the block that maps inputs to outputs with smaller size and more units, the stride of 3x3 convolutional layer is two and the shortcut is a 1x1 convolutional layer for downsampling (Figure 4.4).

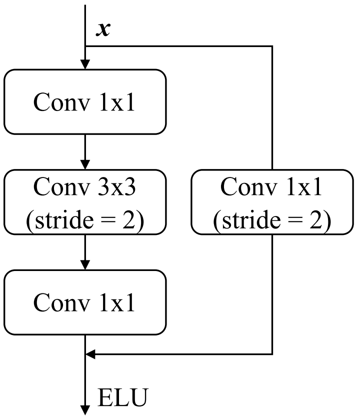


FIGURE 4.3 – Without downsampling

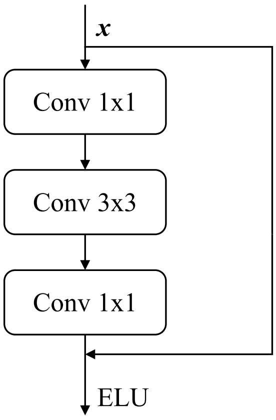


FIGURE 4.4 – With downsampling

4.3. Experiments

4.3.1. Dataset

The dataset is provided by one task of MediaEval 2019: Emotion and Theme Recognition in Music Using Jamendo [61]. It includes 17,982 music tracks with mood and theme annotations. The split for training, validation and test is about $2 : 1 : 1$. In total, there are 56 tags, and tracks can possibly have more than one tag. There are three types of audio representations, including traditional handcrafted audio features, mel-spectrograms, and raw audio inputs. The traditional handcrafted audio features are from Essentia [62] using the feature extractor for AcousticBrainz. These features were used in the MediaEval genre recognition tasks. The number of mel-bands of the mel-spectrograms is 96. The raw audio inputs are in MP3 format with 44.1 kHz sampling rate.

4.3.2. Training

We used the pre-computed mel-spectrograms (Figure 4.2) as inputs, and we used different data augumentation methods in training, validation and test dataset. Let T be the length of input section [frame]. For training dataset, we randomly cropped a T-frame section from each audio track in every epoch. For validation and test dataset, we respectively cropped 10 and 20 T-frame sections from each audio track at regular intervals. We averaged the predictions over all sections of

Table 4.2 – Experiment results

Conv Layers	Residual	PR-AUC-macro	ROC-AUC-macro
5 (baseline)	No	0.1161	0.7475
6	**No**	**0.1256**	**0.7532**
16	Yes	0.1125	0.7393
18	Yes	0.1135	0.7460
25	No	0.1009	0.7319

each audio track. The length of input section T is 1,280 frames. We trained our networking using Adam with the batch size of 64 and the learning rate of 0.001.

4.4. Results and Analysis

We compared the performance of the models that have different architectures or mechanisms in Table 4.2. Suprisingly, the model that achieved the best performance in our experiments was a relatively shallow model that only consists of six convolutional layers, the architecture of which is detailed introduced in Section 4.2. Moreover, the top-5 and bottom-5 tag-wise AUCs of the 6-layer model are showned in Table 4.3. The performance achieved by the best 6-layer model is in the fifth place among all 29 submissions.

The network with 25 convolutional layers consists of one 7x7, twenty-four 3x3 convolutional layers and five max pooling layers for downsampling. It's commonly believed that deep models can achieve a better performance in image classification task. However, the model with deep architecture didn't always achieve a better performance in this task. We also tried residual architecture that commonly used for improving the performance of neural networks. However, the models with residual architecture didn't have an advantage in performance.

The number of samples (18K) in the dataset is relatively smaller than some image datasets (e.g. CIFAR-10: 60K, MS-COCO: 200K, ImageNet: 517K) and the length of audio data (>30s) is relatively longer than some sound datasets (e.g. UrbanSound8K: <4s, ESC-50: 5s, AudioSet: 10s). According to our experience, the generalization ability of models is especially important in this task.

Table 4.3 – Top-5 and bottom-5 tag-wise AUCs

Tag	Rank	PR-AUC-macro	ROC-AUC-macro
summer	1	0.4698	0.9033
deep	2	0.4435	0.9137
corporate	3	0.4017	0.8849
epic	4	0.3886	0.8384
film	5	0.3606	0.7709
etro	52	0.0213	0.7943
holiday	53	0.0186	0.6856
cool	54	0.0185	0.6763
sexy	55	0.0145	0.7327
travel	56	0.0117	0.5990

Therefore, it is reasonable that relatively shallow VGG-based network with strong generalization ability can achive better performance. Furthermore, we think that the number of music tracks in the dataset is relatively not enought, due to the difficulty of this task.

4.5. Conclusion and Future Work

In our experiments, we applied several convolutional neural networks to recogonize the emotion and theme of music. A shallow VGG-based network that consists of six convolutional layers achieved the best performance with PR-AUC-macro of 0.1256 and ROC-AUC-macro of 0.7532. We think that the generalization ability of the models is very important in this task.

In the future, we plan to use all types of the audio representations because we think it is interesting that we treat audio recognition as a multimodal task. Traditional handcrafted audio features and the raw audio inputs may bring great improvement in the performance of our model.

Chapter 5

Demo System for Presentation Improvement

5.1. Introduction

Oral presentation is the most standard format to express ideas or introduce products in many scences. However, there are few efficient tools that can automatically evaluate the presentation. We developed a demo system, which includes a presentation impression prediction system and a presentation slide analysis system, to evaluate presenters' performance and provide impression-related feedback.

We respectively used statistical machine learning method [10] and deep learning method [63] to predict the impressions, a presentation could give to the audiences. In [64], we proposed a method to evaluate slides and to provide a visual feedback to tell presenters which areas of their slides are better to be modified in order to make a better impression.

In our demo, we present the presentation impression prediction system that predicts the audiences' impressions from multiple aspects. We also present the slide analysis system that gives each slide a score and tells presenters the area that needs to be modified.

5.2. Presentation Impression Prediction System

We proposed two methods to predict the audiences' impressions based on linguistic feature as well as acoustic feature. One method used Support Vector

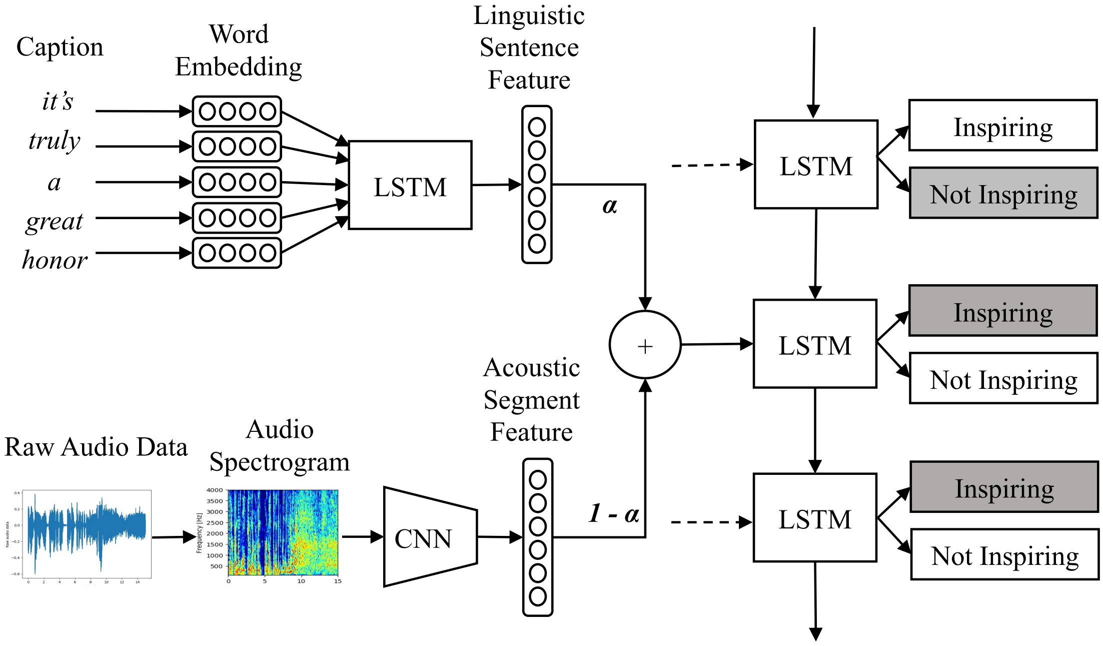


FIGURE 5.1 – Continous impression prediction

Machine (SVM) and Markov Random Field (MRF), and the other method used a multimodal neural network and an attention mechanism.

5.2.1. SVM-MRF

We extracted linguistic feature and acoustic feature from the captions and the audio data, respectively. For the linguistic feature, we used multiple word embedding methods, including Bag-of-words, Latent Semantic Indexing, Latent Dirichlet Allocation, skipgram, and surface-level features. We averaged the vector of all words as document embedding. We extracted the acoustic feature by using openSMILE.

We used SVM to predict the impression labels with only one type of document embedding or only acoustic feature at a time. We then used MRF to consider the correlations between different features and even different impressions to relabel the results of SVMs. However, this method can only predict the impressions for complete presentation videos.

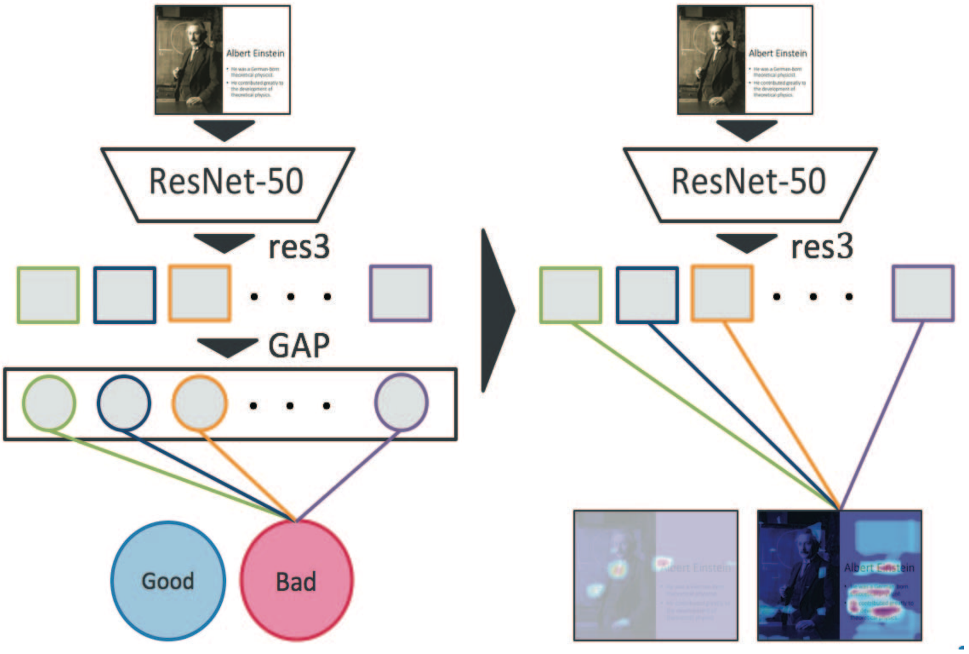


Figure 5.2 – Negative Class Activation Mapping

5.2.2. Multimodal Neural Network

We used a multimodal neural network to provide presenters with continuous feedback (Figure 5.1). We used skip-gram for word embedding, and input the word vectors in each sentence to Long Short-Term Memory (LSTM) in order to get sentence-level linguistic features. We extracted the acoustic features of audio segments, corresponding to each sentence. We took out audio segments and transfered it into frequency domain by Short-Term Fourier Transform. We then used Convolutional Neural Network to extract the deep feature from audio spectrograms as segmental acoustic features.

After we got the sentence-level features, we used an attention mechanism for feature fusion. We then used LSTM to consider the correlations between sentences and predicted audiences' impressions on each sentence. According to the prediction results, the presenters can understand which sentences may leave bad impressions and need to be modified.

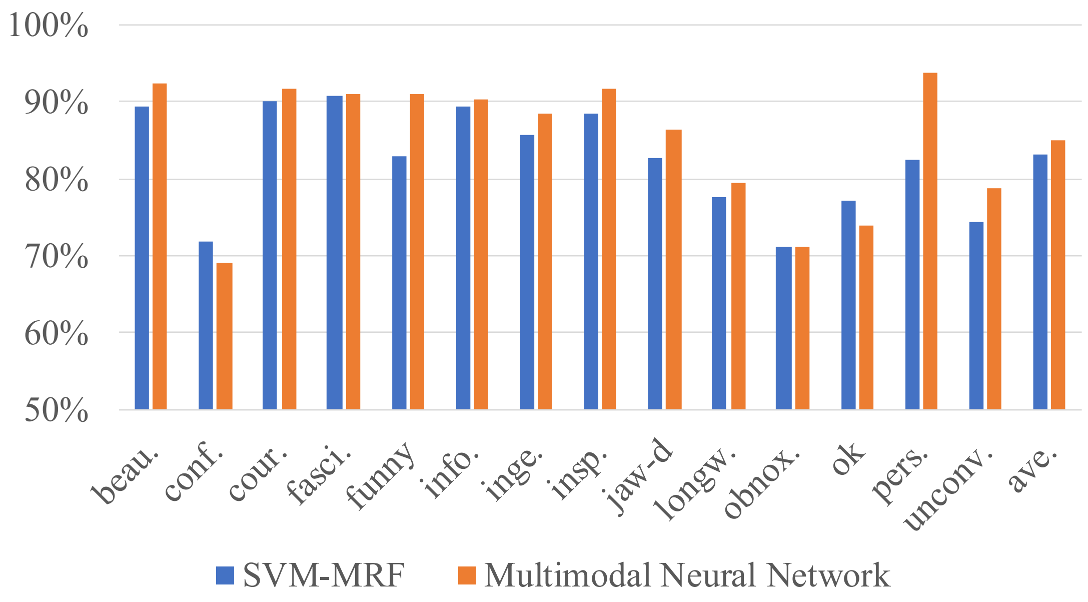


FIGURE 5.3 – Prediction accuracies for different proposals (top and bottom 30%)

5.3. Slide Analysis System

Apart from the presenters's performance, the design of slides is also a key element to a successful presentation. We extracted and concatenated image features, structural features, and content features of slides. We used SVM to predict whether the designs are good or not. If users only get a score, they may not be convinced and don't know how to modify their slides. Therefore, we also proposed a visualization method by using Class Activation Mapping [65] to tell presenters which areas of their slides may give negative impressions (Figure 5.2). We put a Global Average Pooling (GAP) after the Res3 layer of ResNet-50 to learn the weights of class "Positive" and class "Negative". We then gave the weights of "Negative" to the feature maps and got the heatmaps that can show the "Negative" areas.

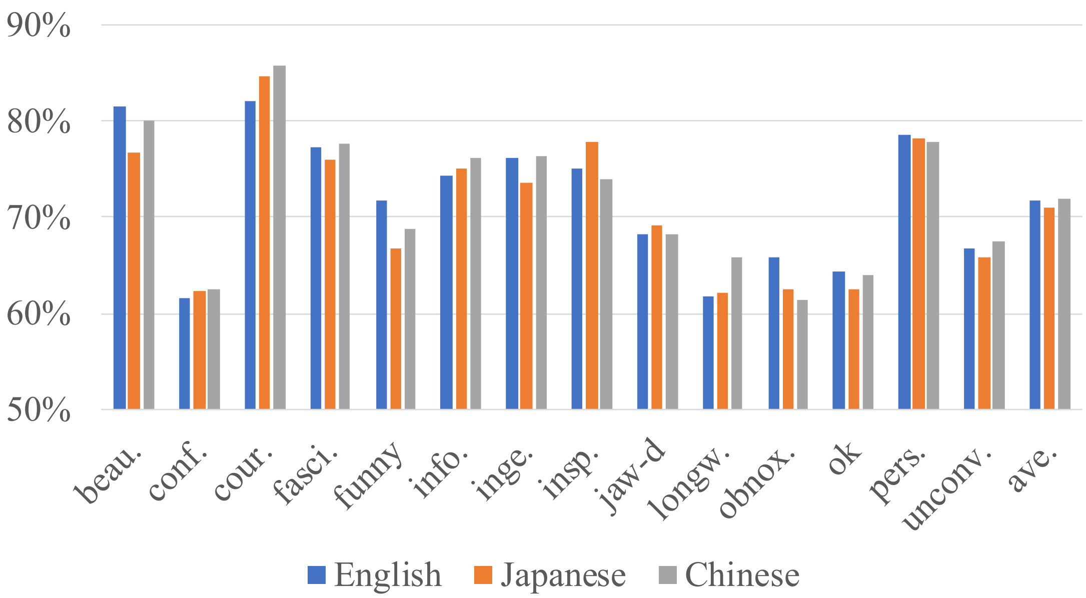


Figure 5.4 – Prediction accuracies for different languages using multimodal neural network (top and bottom 50%)

5.4. Experiments

5.4.1. Presentation Impression Prediction

We used the captions and the audio data of 2,445 presentations. Each presentation has 14 impression tags, including *Beautiful*, *Confusing*, *Courageous*, *Fascinating*, etc. These tags are based on all audiences' votes and tell us whether the audiences have corresponding impressions.

We evaluated our proposals on top and bottom 30% and 50% of the presentation samples, respectively. The prediction results of complete presentations impressions of two proposals (top and bottom 30%) are shown in Figure 5.3. SVM-MRF achieved higher efficiency but can't predict impressions continuously as multimodal neural network does. Therefore, we predict the impressions of the complete presentation and each sentence by these two methods, respectively.

Besides the original English captions, we also used the official translation in Japanese and Chinese. We applied our proposals with Japanese and Chinese captions as well as English audio, because only English audio is available, Figure 5.4 shows that mismatched data can also achieve a high performance.

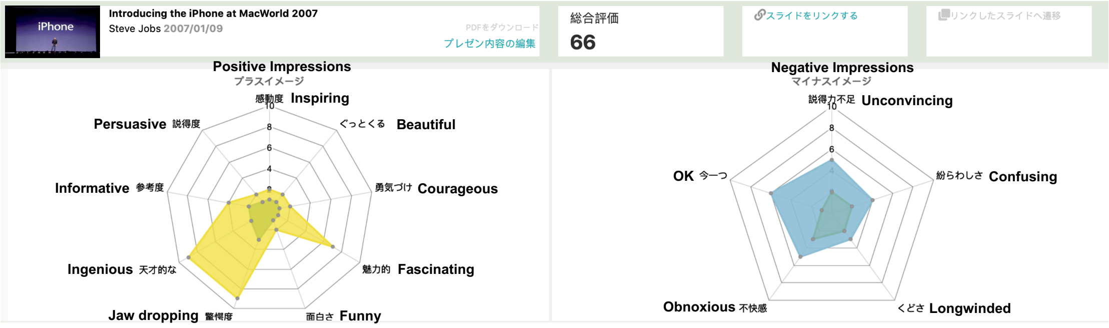


FIGURE 5.5 – Prediction example

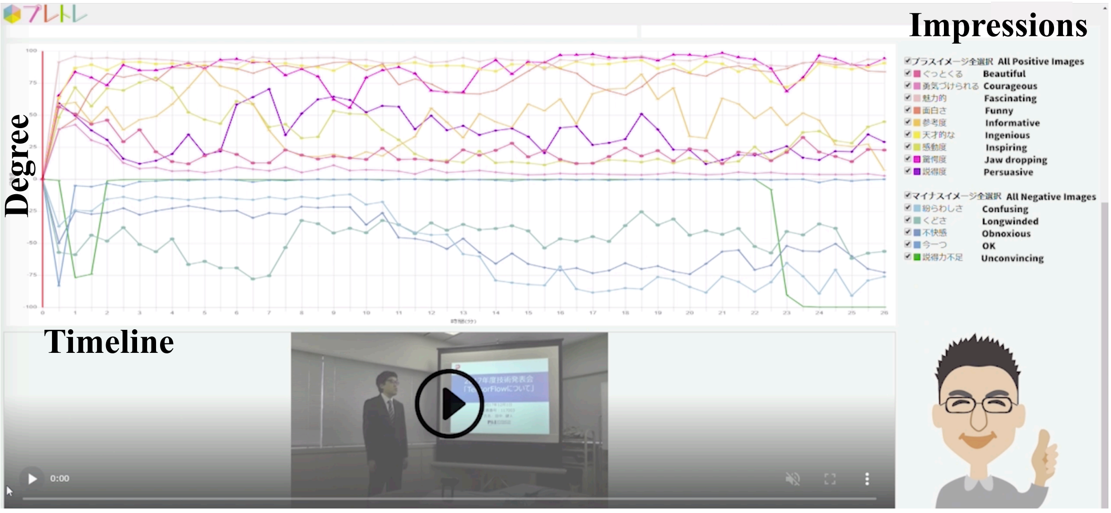


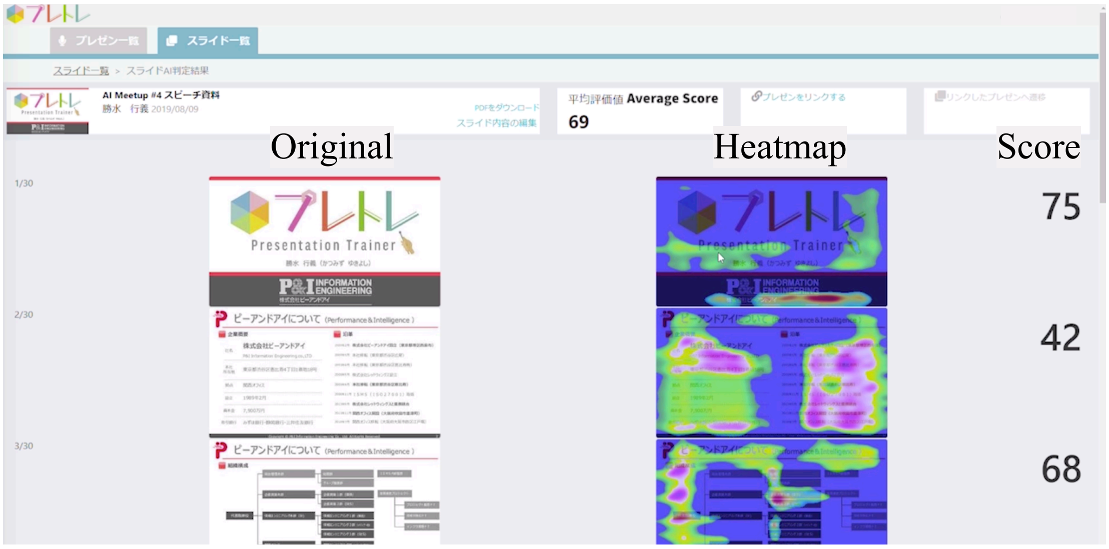
FIGURE 5.6 – Presentation impression prediction system

5.4.2. Slide Analysis

We hired 100 workers to create 1,000 PowerPoint slides in 10 topics, and they gave each slide with a visual clarity rank (1 to 100). We treated top and bottom 30% of the slides as “Positive” and “Negative” samples, respectively. We achieved the classification rate of 90.3%. We then further proposed a feedback system that can provide visual clarity scores and point out areas that should be modified.

5.5. Demo

In the demo, we presented our web application of presentation support system. Users only need to upload their presentation videos or presentation slides. Our system will automatically analyze them behind the scenes. We first presented our presentation impression prediction system. This system predicted audiences’



AI Meetup #4 スピーチ資料
勝水 行義 2019/08/09

平均評価値 Average Score
69

FIGURE 5.7 – Slide analysis system

impressions for each complete presentation from 14 aspects, including 9 positive impressions and 5 negative impressions. The prediction results of positive and negative impressions were presented in the left and the right, respectively. Figure 5.5 is an example of analysis results of a presentation from Steve Jobs, in which he introduced the original iPhone. This presentation is predicted to be ingenious, jaw-dropping, fascinating, obnoxious and OK. Our system can also show the temporal change of these impressions continuously during the presentation (Figure 5.6). There were 14 lines in the figure, and each line represents how the prediction of corresponding impression changed every minute. Furthermore, the demo system can tell the presenter what are the top-10 similar presentations in TED Talks that create similar impressions on audiences. Now, our demo system can analyze presentation videos in English or Japanese, and we plan to make Chinese available.

We then presented the slide analysis system (Figure 5.7). The system gave each single-page slide with a score and told us which areas gave audiences negative impressions. In our experience, these bad areas often include too many letters, blanks, or inappropriate font size.

5.6. Conclusion

In our study, we successfully applied our proposal of impression prediction on the TED Talks dataset. The results showed that our proposal achieved high performance with a prediction accuracy of 85.0% for the top and bottom 30% ranked data. And our proposal of slide evaluation can achieve the prediction accuracy of 90.3% and give a useful visual feedback to point out areas that need to be modified. Our presentation analysis system is now available as a WEB service, which were also demonstrated in this section.

Chapter 6

Conclusion and Future Work

6.1. Conclusion

In this thesis, we presented our study of impression analysis on presentations using attention-based LSTM. We began with the background introduction of presentations, emotion recognition, and the press conference. Then we proposed the multimodal neural network for the impression prediction of oral presentations. Our proposal included two LSTM-based unimodal neural networks with the attention mechanism for extracting linguistic features and acoustic features, as well as an attention network for feature fusion. We successfully applied our proposal for impression prediction on the TED Talks dataset. The results showed that our proposal achieved high performance with a prediction accuracy of 85.0% for the top and bottom 30% ranked data, and a significant improvement to the efficiency of the impression prediction system. Furthermore, we can clarify the importance relationship between the content of a presentation and the presenter's manner of speaking for different audience impressions using an attention network. A strong relationship between the average vote rates and the prediction accuracies were found that the prediction accuracies of impressions with a low average vote rate were relatively lower than the impressions with a high vote rate.

After that, we introduced our proposal for the press conferences, a special type of oral presentation. In the press conferences, the speakers need to answer journalists' questions. Professional consultants evaluated these answers based on 11 criterions, and classified the samples into positive, neutral, or negative. We constructed an automatic evaluation system to predict the consultants' evaluation by using LSTM and self-attention mechanism. The word representation of our proposal was the sum of the token embedding and type embedding. From experiment results, our proposal achieved relatively good performance of the average accuracy of 57.6% for the three classification task, even though it was based on relatively complex criterions. Furthermore, the experiment results of using the speakers' answers only, or using the Q&A pairs that include both journalists' questions and the speakers' answers, indicated that the Q&A pairs can include more information than only the speakers' answers, and they can help to improve the performance of our model.

We also presented our fundamental research of audio emotion recognition. We applied several CNNs to recogonize the emotion and theme of music, using the audio mel-spectrograms as input. These CNNs mainly differ in the depth of the architecture and whether have the residual architecture. A shallow VGG-based network, consisting of six convolutional layers and without residual architecture, achieved the best performance with PR-AUC-macro of 0.1256 and ROC-AUC-macro of 0.7532. We pointed out that the generalization ability of the models is very important for this task, and complex models may bring overfitting problems. Furthermore, we also pointed out that the number of music tracks in the dataset is relatively not enought, due to the difficulty of this task.

At last, we presented the demo of oral presentation support system, including the presentation impression prediction system and the slide analysis system. The presentation impression prediciton system can predict the audiences' impressions created by the presentation from multiple aspects. And it can tell the presenter what are the top-10 similar presentations in TED Talks that create similar impressions on audiences. In the demo, we also presented the slide analysis system that gave each slide a score and told presenters the area that needs to be modified.

6.2. Future Work

In our research, we have found that most recent langue models and acoustic models are designed for relatively short length data, and there is a performance

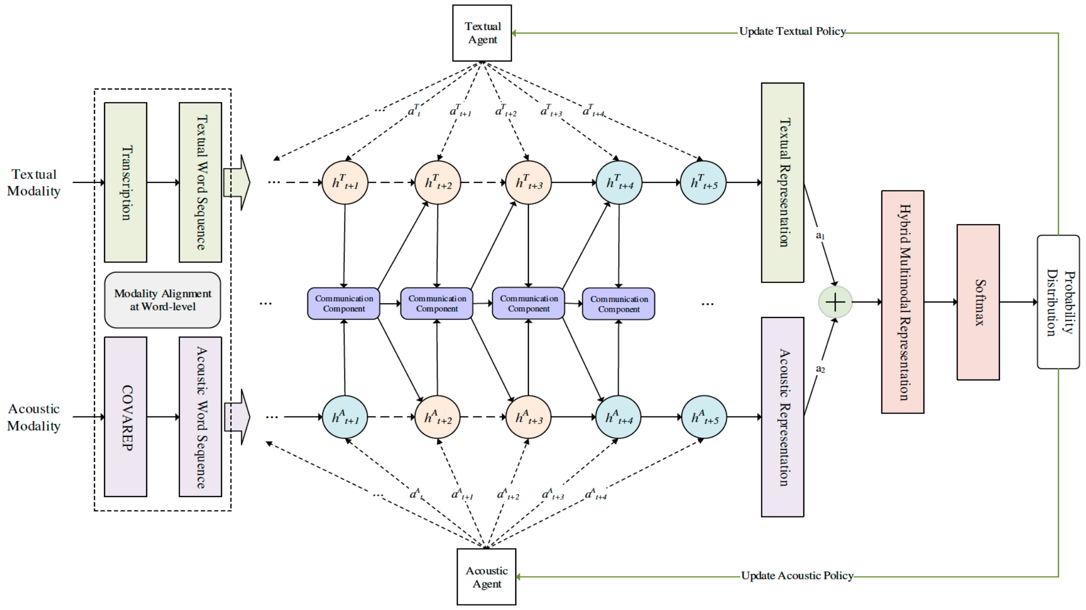


FIGURE 6.1 – The architecture of the multi-agent reinforcement learning method [66]

bottleneck of learning long-term dependency. However, the ability to model long-term dependency is extremely important for presentation analysis, because presentations often last for a relatively long time and the length of their linguistic information and acoustic information is also very long. Even though the average length of documents for presentation analysis can be more than 1800 words/doc, recent models, like LSTM based model, can only learn dependency of no more than several hundred words long. There are already some researchers focusing on the long sequence problem of multimodal recognition. Zhang et al. used a multi-agent reinforcement learning method (Figure 6.1) to select effective sentiment-relevant words for multi-modal sentiment analysis with focus on both the textual and acoustic modalities [66]. Ma et al. used a multi-layer residual LSTM network (Figure 6.2) for better obtaining long-term dependency [67]. These methods tried to solve the long dependency learning problem from the way of denoising or adding a shortcut, respectively. In the future, we will continue our research on presentation analysis, and focus on handling the super-long sequence of multimedia data. For linguistic data, we will dedicate to solving the problem of how to learn long-term dependency in language model.

We have been focusing on multimodal emotion recognition of unaligned wild

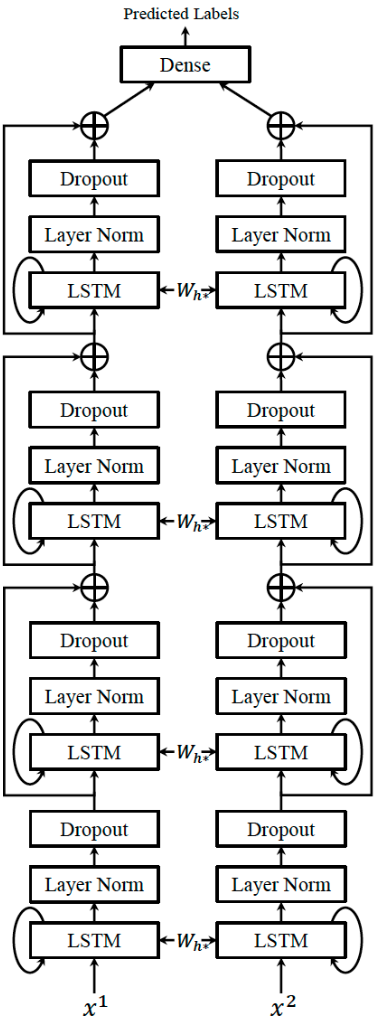


Figure 6.2 – The architecture of the multi-layer residual LSTM network [67]

data, and the sequence length of modalities have a big difference. In the future, we are going to experiment on the word-level aligned datasets, including CMU-MOSI [68], CMU-MOSEI [69], IEMOCAP [70]. In these word-level aligned datasets, the sequence of the segmental features of video and audio data are strictly corresponding to each word (Figure 6.3). The preprocessing of word-level alignment can bring a great improvement in the performance of models [1]. Because of the relatively large length of used data, we only applied LSTM-based methods until now. If we start to handle the word-level aligned datasets above, most sequence data will be much shorter. Therefore, we will use Transformer-based methods that can achieve better performance than LSTM-based methods.

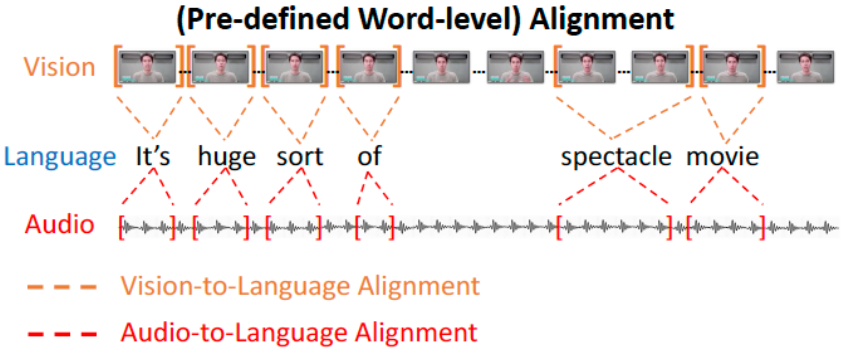

(Pre-defined Word-level) Alignment
Vision
Language
It's
huge
sort
of
spectacle
movie
Audio
Vision-to-Language Alignment
Audio-to-Language Alignment

FIGURE 6.3 – Word-level aligned multimodal features [1]

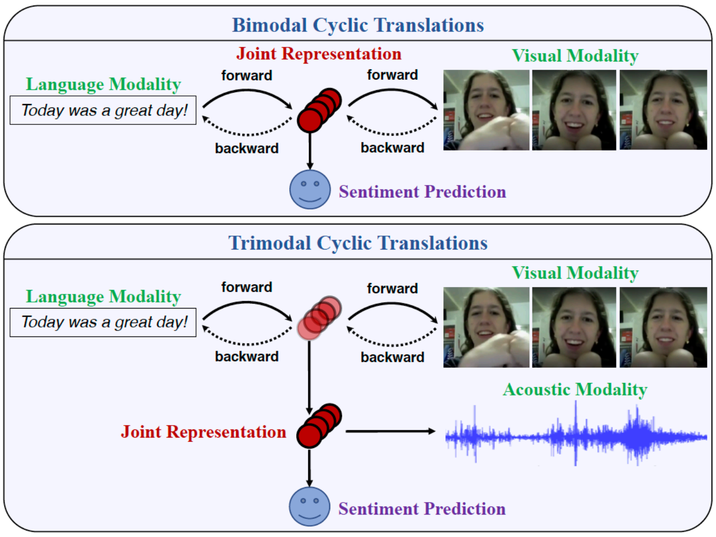


FIGURE 6.4 – Cyclic translation for joint representation [2]

For the feature fusion of multimodal features, we used an attention network as the late fusion method. However, before the feature fusion, we didn't consider any correlation between low-level unimodal features. Therefore, we will try to use some crossmodal methods. Pham et al. worked on the cyclic translation between different modalities (Figure 6.4) to learn joint representations from additional information present in multi modalities [2]. Tsai et al. used crossmodal Transformers (Figure 6.5) to repeatedly reinforce a target modality with the low-level features from another source modality by learning the attention across the two modalities' features [1]. In the future, we will combine the cyclic translation with the crossmodal Transformers. The corssmodal Transformer architecture is expected to achieve better performance than LSTM-based methods. The cyclic translation

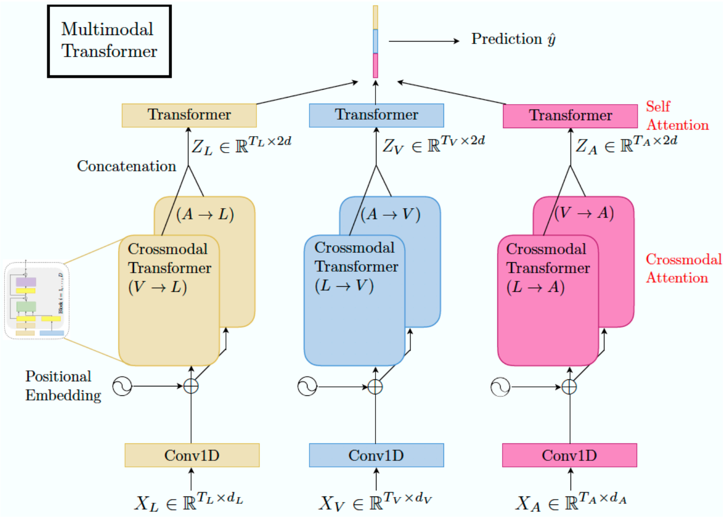

Multimodal Transformer
Prediction $\hat{y}$
Transformer
Transformer
Transformer
Self Attention
$Z_L \in \mathbb{R}^{T_L \times 2d}$
$Z_V \in \mathbb{R}^{T_V \times 2d}$
$Z_A \in \mathbb{R}^{T_A \times 2d}$
Concatenation
$(A \rightarrow L)$
$(A \rightarrow V)$
$(V \rightarrow A)$
Crossmodal Transformer $(V \rightarrow L)$
Crossmodal Transformer $(L \rightarrow V)$
Crossmodal Transformer $(L \rightarrow A)$
Crossmodal Attention
Positional Embedding
Conv1D
Conv1D
Conv1D
$X_L \in \mathbb{R}^{T_L \times d_L}$
$X_V \in \mathbb{R}^{T_V \times d_V}$
$X_A \in \mathbb{R}^{T_A \times d_A}$

FIGURE 6.5 – Multimodal Transformer [1]

with using shared architecture can decrease the complexity of Transformer-based methods and bring the performance improvement by adding pretraining stage.

References

[1] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," *arXiv preprint arXiv:1906.00295*, 2019.

[2] H. Pham, P. P. Liang, T. Manzini, L.-P. Morency, and B. Póczos, "Found in translation: Learning robust joint representations by cyclic translations between modalities," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 6892–6899.

[3] A. Zadeh, P. P. Liang, S. Poria, P. Vij, E. Cambria, and L.-P. Morency, "Multi-attention recurrent network for human communication comprehension," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[4] D. Tang, F. Wei, N. Yang, M. Zhou, T. Liu, and B. Qin, "Learning sentiment-specific word embedding for twitter sentiment classification," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Baltimore, Maryland: Association for Computational Linguistics, Jun. 2014, pp. 1555–1565. [Online]. Available: https://www.aclweb.org/anthology/P14-1146

[5] M. R. Islam and A. Elchouemi, "Feature selection approach for twitter sentiment analysis and text classification based on chi-square and naïve bayes," in *International Conference on Applications and Techniques in Cyber Security and Intelligence ATCI 2018: Applications and Techniques in Cyber Security and Intelligence*, vol. 842. Springer, 2019, p. 281.

[6] J. McAuley and J. Leskovec, "Hidden factors and hidden topics: understanding rating dimensions with review text," in *Proceedings of the 7th ACM conference on Recommender systems*, 2013, pp. 165–172.

[7] M. Soleymani, M. N. Caro, E. M. Schmidt, C.-Y. Sha, and Y.-H. Yang, "1000 songs for emotional analysis of music," in *Proceedings of the 2nd ACM international workshop on Crowdsourcing for multimedia*. ACM, 2013, pp. 1–6.

[8] D. Bogdanov, A. Porter, P. Tovstogan, and M. Won, "Mediaeval 2019: Emotion and theme recognition in music using jamendo," *MediaEval Benchmarking Initiative for Multimedia Evaluation. Sophia Antipolis, France*, 2019.

[9] Y. Fukushima, T. Yamasaki, and K. Aizawa, "Presentation video assessment based on text and acoustic analysis," *IEICE Transactions on Information and Systems*, vol. J99-D, no. 8, pp. 699–708, 2016 (in Japanese).

[10] T. Yamasaki, Y. Fukushima, R. Furuta, L. Sun, K. Aizawa, and D. Bollegala, "Prediction of user ratings of oral presentations using label relations," in *Proceedings of the 1st ACM International Workshop on Affect & Sentiment in Multimedia*, 2015, pp. 33–38.

[11] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[12] A. Wu and H. Qu, "Multimodal analysis of video collections: Visual exploration of presentation techniques in ted talks," *IEEE transactions on visualization and computer graphics*, 2018.

[13] A. I. Gheorghiu, M. J. Callan, and W. J. Skylark, "A thin slice of science communication: Are people' s evaluations of ted talks predicted by superficial impressions of the speakers?" *Social Psychological and Personality Science*, p. 1948550618810896, 2019.

[14] S. Zhang, L. Li, and Z. Zhao, "Audio-visual emotion recognition based on facial expression and affective speech," in *International Conference on Multimedia and Signal Processing.* Springer, 2012, pp. 46–52.

[15] H. Xu, H. Zhang, K. Han, Y. Wang, Y. Peng, and X. Li, "Learning alignment for multimodal emotion recognition from speech," *arXiv preprint arXiv:1909.05645*, 2019.

[16] S. Yoon, S. Byun, and K. Jung, "Multimodal speech emotion recognition using audio and text," in *2018 IEEE Spoken Language Technology Workshop (SLT).* IEEE, 2018, pp. 112–118.

[17] L.-P. Morency, R. Mihalcea, and P. Doshi, "Towards multimodal sentiment analysis: Harvesting opinions from the web," in *Proceedings of the 13th international conference on multimodal interfaces*, 2011, pp. 169–176.

[18] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *2013 IEEE international conference on acoustics, speech and signal processing*, 2013, pp. 6645–6649.

[19] F. Weninger, J. Krajewski, A. Batliner, and B. Schuller, "The voice of leadership: Models and performances of automatic analysis in online speeches," *IEEE Transactions on Affective Computing*, vol. 3, no. 4, pp. 496–508, 2012.

[20] G. Luzardo, B. Guamán, K. Chiluiza, J. Castells, and X. Ochoa, "Estimation of presentations skills based on slides and audio features," in *Proceedings of the ACM workshop on Multimodal Learning Analytics Workshop and Grand Challenge*, 2014, pp. 37–44.

[21] Z. S. Harris, "Distributional structure," *Word*, vol. 10, no. 2-3, pp. 146–162, 1954.

[22] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American society for information science*, vol. 41, no. 6, pp. 391–407, 1990.

[23] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[24] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[25] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia*, 2013, pp. 835–838.

[26] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 1480–1489.

[27] 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*, 2017, pp. 5998–6008.

[28] S. Lee, X. Jin, and W. Kim, "Sentiment classification for unlabeled dataset using doc2vec with jst," in *Proceedings of the 18th Annual International Conference on Electronic Commerce: e-Commerce in Smart connected World.* ACM, 2016, p. 28.

[29] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[30] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," in *Advances in neural information processing systems*, 2019, pp. 5754–5764.

[31] O. Niebuhr, "Clear speech-mere speech? how segmental and prosodic speech reduction shape the impression that speakers create on listeners." in *INTERSPEECH*, 2017, pp. 894–898.

[32] Y. Tokozume, Y. Ushiku, and T. Harada, "Learning from between-class examples for deep sound recognition," *arXiv preprint arXiv:1711.10282*, 2017.

[33] H. B. Sailor, D. M. Agrawal, and H. A. Patil, "Unsupervised filterbank learning using convolutional restricted boltzmann machine for environmental sound classification." in *INTERSPEECH*, 2017, pp. 3107–3111.

[34] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Processing Letters*, vol. 24, no. 3, pp. 279–283, 2017.

[35] K. J. Piczak, "Esc: Dataset for environmental sound classification," in *Proceedings of the 23rd ACM international conference on Multimedia*, 2015, pp. 1015–1018.

[36] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 1041–1044.

[37] K. J. Piczak, "Environmental sound classification with convolutional neural networks," in *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2015, pp. 1–6.

[38] K. Koutini, H. Eghbal-zadeh, and G. Widmer, "Receptive-field-regularized cnn variants for acoustic scene classification," in *Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, 2019, p. 124.

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

[40] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," *arXiv preprint arXiv:1508.04025*, 2015.

[41] S.-s. Shen and H.-y. Lee, "Neural attention models for sequence classification: Analysis and application to key term extraction and dialogue act detection," *arXiv preprint arXiv:1604.00077*, 2016.

[42] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

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

[44] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional lstm and other neural network architectures," *Neural networks*, vol. 18, no. 5-6, pp. 602–610, 2005.

[45] J. Nowak, A. Taspinar, and R. Scherer, "Lstm recurrent neural networks for short text and sentiment classification," in *International Conference on Artificial Intelligence and Soft Computing.* Springer, 2017, pp. 553–562.

[46] J. Zhou, Y. Lu, H.-N. Dai, H. Wang, and H. Xiao, "Sentiment analysis of chinese microblog based on stacked bidirectional lstm," *IEEE Access*, vol. 7, pp. 38 856–38 866, 2019.

[47] S. Yi, X. Wang, and T. Yamasaki, "Impression prediction of oral presentation using lstm and dot-product attention mechanism," in *IEEE Fifth International Conference on Multimedia Big Data (BigMM)*, 2019, pp. 242–246.

[48] D. Hazarika, S. Poria, R. Mihalcea, E. Cambria, and R. Zimmermann, "Icon: Interactive conversational memory network for multimodal emotion detection," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 2594–2604.

[49] D. Hazarika, S. Poria, A. Zadeh, E. Cambria, L.-P. Morency, and R. Zimmermann, "Conversational memory network for emotion recognition in dyadic dialogue videos," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, 2018, pp. 2122–2132.

[50] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[51] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," *arXiv preprint arXiv:1802.05365*, 2018.

[52] E. Bataa and J. Wu, "An investigation of transfer learning-based sentiment analysis in japanese," *arXiv preprint arXiv:1905.09642*, 2019.

[53] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, 2009, pp. 248–255.

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

[55] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[56] K. Choi, G. Fazekas, and M. Sandler, "Automatic tagging using deep convolutional neural networks," *arXiv preprint arXiv:1606.00298*, 2016.

[57] 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 *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017, pp. 131–135.

[58] R. Panda, R. Malheiro, and R. P. Paiva, "Musical texture and expressivity features for music emotion recognition," in *Proceedings of the International Society for Music Information Retrieval*, 2018, pp. 383–391.

[59] M. A. F. Ballester, "A novel approach to string instrument recognition," in *Proceedings of Image and Signal Processing: 8th International Conference*, vol. 10884, 2018, pp. 165–175.

[60] B. Shi, X. Bai, and C. Yao, "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 11, pp. 2298–2304, 2016.

[61] D. Bogdanov, A. Porter, P. Tovstogan, and M. Won, *MediaEval 2019: Emotion and Theme Recognition in Music Using Jamendo*, 2019, in MediaEval Benchmark Workshop.

[62] D. Bogdanov, N. Wack, E. Gómez Gutiérrez, S. Gulati, H. Boyer *et al.*, “Essentia: An audio analysis library for music information retrieval,” in *Proceedings of the International Society for Music Information Retrieval*, 2013, pp. 493–498.

[63] S. Yi, X. Wang, and T. Yamasaki, “Impression prediction of oral presentation using lstm and dot-product attention mechanism,” in *IEEE Fifth International Conference on Multimedia Big Data (BigMM)*, 2019, pp. 242–246.

[64] S. Oyama and T. Yamasaki, “Visual clarity analysis and improvement support for presentation slides,” in *IEEE Fifth International Conference on Multimedia Big Data (BigMM)*, 2019, pp. 421–428.

[65] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Learning deep features for discriminative localization,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.

[66] D. Zhang, S. Li, Q. Zhu, and G. Zhou, “Effective sentiment-relevant word selection for multi-modal sentiment analysis in spoken language,” in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 148–156.

[67] J. Ma, H. Tang, W.-L. Zheng, and B.-L. Lu, “Emotion recognition using multimodal residual lstm network,” in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 176–183.

[68] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, “Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages,” *IEEE Intelligent Systems*, vol. 31, no. 6, pp. 82–88, 2016.

[69] A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, “Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph,” in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 2236–2246.

[70] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, “Iemocap: Interactive emotional dyadic motion capture database,” *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

Publications

International Conferences and Workshops

[1] Shengzhou Yi, Xueting Wang, and Toshihiko Yamasaki. Impression Prediction of Oral Presentation using LSTM and Dot-product Attention Mechanism. *In IEEE Fifth International Conference on Multimedia Big Data (BigMM 2019), pp. 242- 246, Sep. 11-13, 2019, Singapore.*

[2] Shengzhou Yi, Xueting Wang, and Toshihiko Yamasaki. Emotion and Theme Recognition of Music Using Convolutional Neural Networks. *In MediaEval Benchmark Workshop (MediaEval 2019), Oct. 27-29, 2019, Sophia Antipolis, France.*

[3] Shengzhou Yi, Hiroshi Yumoto, Xueting Wang, Toshihiko Yamasaki. PresentationTrainer: Oral Presentation Support System for Impression-related Feedback. *In Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI 2020), Demo. Feb. 7-12, 2020, New York, USA. (Accepted)*

[4] Shengzhou Yi, Koshiro Mochitomi, Isao Suzuki, Xueting Wang, Toshihiko Yamasaki. Attention-based LSTM for Automatic Evaluation of Press Conferences. *In IEEE Third International Conference on Multimedia Information Processing and Retrieval (MIPR 2020). Apr. 9-11, 2020, Shenzhen, China. (Accepted)*

International Conferences and Workshops without Review

[5] Shengzhou Yi, Xueting Wang, and Toshihiko Yamasaki. Impression Prediction of Oral Presentation using LSTM and Dot-product Attention Mechanism. *Third International Workshop on Symbolic-Neural Learning (SNL-2019), P-16, July 11-12, 2019, Miraikan hall, Odaiba Miraikan 7F (Tokyo, Japan).*

Domestic Conferences and Symposia

[6] Shengzhou Yi, Toshihiko Yamasaki, Izumi Masumura, Yoshinori Yasui, Takako Misaki, Nobuhiko Okabe. Prediction of the National Epidemiological Surveillance of Infectious Diseases Using LSTM. *Image Media Processing Symposium (IMPS), P-1-11, Nov. 19-21, 2018, Gotemba.*

[7] Shengzhou Yi, Wang Xueting, and Yamasaki Toshihiko. Impression Prediction of Oral Presentation Using LSTM with Dot-product Attention Mechanism. *Media Experience Virtual Environment (MVE), IEICE Technical Report, vol. 119, no. 75, pp. 1-6, Jun. 10-11, 2019, Tokyo.*

[8] Shengzhou Yi, Koshiro Mochitomi, Isao Suzuki, Xueting Wang, Toshihiko Yamasaki. Automatic Evaluation of Press Conferences Using LSTM with Self-Attention Mechanism. *Human Communication Group (HCG) Symposium, Dec. 11-13, 2019, Hiroshima.*

[9] Shengzhou Yi, Xueting Wang, and Toshihiko Yamasaki. CNN-based Music Emotion and Theme Recognition Featuring Shallow Architecture. *Media Experience Virtual Environment (MVE), IEICE Technical Report, Jan. 23-24, 2020, Nara. (Submitted)*

[10] Shengzhou Yi, Takuya Yamamoto, Osamu Yamamoto, Yukiyoshi Katsumizu, Hiroshi Yumoto, Xueting Wang, Toshihiko Yamasaki. Make Your Presentation Better: Oral Presentation Support System using Linguistic and Acoustic Features. Image Engineering Technical Group (IE), Feb. 27-28, 2020, Hokkaido. (Submitted)

Acknowledgements

It is a wonderful experience to study at The University of Tokyo for last two years. In these two years, there are many persons who used the best effort to help me, but I have no chance to express my thanks to them. Therefore, please let me say "Thank you!" to everyone who helped me here.

Please first let me express my deepest gratitude to my supervisor, Prof. Toshihiko Yamasaki. Without his guidance and monitoring, I can't smoothly study about the field of multimodal analysis that I am really interested in. He provided me with the best support on both my research and daily life. Under his help, I learned how to present my works as well as communicating with other researchers or cooperators. I am very thankful to Prof. Kiyoharu Aizawa. He can always find the weak points of my research and give me with his professional advice. I am also very grateful to Mrs. Wang, providing great advice to help me edit my papers, and sharing her academic knowledge.

Thanks to all lab members, making me feel that I am living in a big familly. Thanks to Mrs. Egawa, providing me with the best supports for my school affairs and daily life. Thanks to Mr. Inoue, and Mr. Nakamura, giving their helpful advice and opening my window into researching. Thanks to Ms. Yiwei, and Mr. Sourav, telling me what they have experienced to help me towards the right direction. Thanks to Mr. Kosugi, Mr. Kato, Mr. Zhang, Mr. Lin, and Mr. Chen, who entered the lab at the same time with me, talking with me everyday and never let me feel longly. Thanks to Mr. Tsubota, Mr. Yu, Mr. Kaneko, Ms. Kawanarada, and Mr. Tanaka, making me feel very happy in our monthly party. Thanks to Mr. Ikeda, who often raise some questions to me, making me feel that my knowledge is meaningful to others. May all lab members good health, a beautiful future and can achieve their dreams.

Shengzhou Yi
January 30, 2020