階層的Tensor Fusionを用いた交渉対話における嘘検出

Hierarchical tensor fusion for deception detection of negotiation dialog

グエン テトウン[1*] 吉野幸一郎[1] サクリアニ サクティ[1] 中村哲[1]
Nguyen The Tung[1] Koichiro Yoshino[1] Sakriani Sakti[1] Satoshi Nakamura[1]

[1] 奈良先端科学技術大学院大学
[1] Nara Institute of Science and Technology

Abstract: In negotiation dialogs, it is common for negotiator to use lies for getting a more beneficial outcome. Therefore, to negotiate efficiently, it is important to know whether the other party is telling the truth or not. This deception information can be determined from various visual and acoustic clues. In this paper, we proposed a new method called hierarchical tensor fusion network for combining visual and acoustic modalities for the task of deception detection in negotiation dialog. The hierarchical tensor fusion model has the best performance in our experiments, outperforming the existing approaches used in previous studies (hierarchical and tensor fusion). We inspected the performance of negotiation dialog system when using output labels from multiple deception detection methods and saw that the dialog system achieves highest performance when using labels from the hierarchical TFN model.

1 Introduction

In a negotiation, interlocutors can use lies to gain advantages and reach outcomes that are more beneficial for them. In order to successfully negotiate with human users, dialog systems need an accurate deception detection module [7].

According to existing studies, acoustic factors (pitch, intensity, and speaking rate) and visual clues (facial expressions) are good indicators to tell if someone is being deceptive [1, 2]. From these literatures, it is expected that incorporating visual features to acoustic features has the potential to achieve high deception detection accuracy.

There are many different methods to combine features from multiple modalities. A majority of multimodal processing works still use traditional methods such as early or late fusion. On the other hand, more advanced combination methods like hierarchical fusion [3] or tensor fusion network [4] have been proposed and was shown to outperform the early and late fusion in tasks such as emotion recognition or sentiment analysis.

In this study, we propose the hierarchical tensor fusion network (Hierarchical TFN), a new type of method for fusing multi-modal features efficiently using neural network. Our proposal can be seen as an integration of the hierarchical and the tensor fusion network. The proposed method balances the feature abstraction level by a hierarchical structure and explicitly combines different types of features by outer product, thus eliminating the weaknesses of previous methods.

We conducted experiment on deception detection task and observed that the model based on the proposed hierarchical tensor fusion method achieved the best score, outperformed the existing methods. In addition, we used the predicted labels from deception detection models based on different fusion methods as inputs of an reinforcement-learning-based negotiation dialog manager [5]. We observed that when using the predicted deception labels from the hierarchical tensor fusion based classifiers, the dialog manager achieved a high score.

*連絡先： 奈良先端科学技術大学院大学
奈良県生駒市高山町８９１６番地の５
E-mail: nguyen.tung.np5@is.naist.jp

2 Related works

There are a lot of studies about how to combine multi-modal features efficiently. Tian et.al., [3] proposed hierarchical fusion for emotion recognition. This network has a hidden layer that connects with both acoustic and facial input feature vectors. The architecture was constructed under the assumption that acoustic and facial features have different levels of abstractions. Tensor fusion network (TFN) [4] composes a tensor of the acoustic features and facial features. This fusion method achieved high performance in the tasks of face recognition and sentiment analysis. Such performance can be explained by the fact that TFN can consider the outer product of different type features explicitly. Despite of those achievements, hierarchical and tensor fusion still contain some weaknesses. For the tensor fusion network, it assumes that all modalities have the same level of abstraction, thus making the network structure large and complex. This property makes it difficult to train a network from small amount of data. The hierarchical fusion uses concatenation when fusing different modalities, which makes learning of multi-modal fusion very complicated. From these observations, we proposed the Hierarchical TFN and expected that it would able to model the important features contributing to the improvement of deception detection accuracy.

3 Method

The structure of our proposed network is shown in Figure 1. As can be seen from this figure, the hierarchical TFN structure is similar to a hierarchical network, but multi-modality fusion is performed by using the outer product (same as TFN) instead of concatenating. Our proposed method's advantages over a hierarchical network is similar to that of TFN over early fusion thanks to the use of the outer product for fusion. The hierarchical TFN also resembles a TFN structure; however, raw acoustic features are used as inputs of tensor fusion instead of an acoustic embedding vector as we can see in TFN.

4 Experiments

The dataset we used for deception detection includes two types of data. The first one is created by splitting videos from [6] into utterances. The second

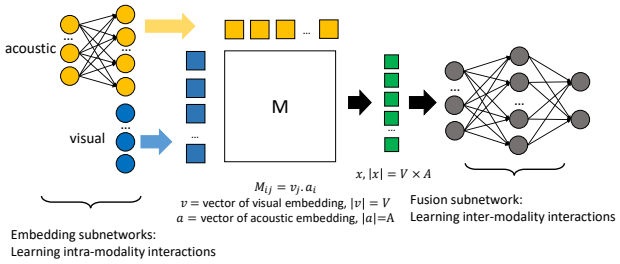

図 1: *Hierarchical tensor fusion network.*

data consists of recorded videos of health consultation dialogs [5]. In total, our deception detection dataset used in the experiment includes 1265 utterances.

表 1: Results of deception detection.

Model	Accuracy	Precision	Recall	F1-score
single acoustic	53.78%	0.4747	0.5000	0.4870
single visual	49.28%	0.4095	0.3525	0.3879
multi early	53.42%	0.4603	0.3566	0.4018
multi late	54.68%	0.4794	0.3811	0.4247
multi hierarchical	53.78%	0.4733	0.4713	0.4723
multi TFN	50.36%	0.4216	0.3525	0.3839
multi hierarchical TFN	**58.63%**	**0.5304**	**0.5000**	**0.5148**

Table 1 summarizes performance of different models in the deception detection task. We can see that the hierarchical TFN outperforms all the other methods.

表 2: Accuracy of dialog acts selection when using different deception labels result.

Deception labels used for dialog management	DA accuracy
chance rate deception	65.69%
gold-label deception	80.31%
single visual prediction	70.15%
single acoustic prediction	66.22%
multi early prediction	66.48%
multi late prediction	68.58%
multi hierarchical prediction	69.10%
multi TFN prediction	69.66%
multi hierarchical TFN prediction	71.20%

In the next experiment, we used predicted labels from deception detection models for a dialog manager that decides output dialog acts (DA) on the basis of the user's deception information (whether the user is

lying or not). The dialog act labels are decided by POMDP-based dialogue manager [5], which can incorporate the deceptive labels with the user's dialog act labels for decide a better dialogue act of the system. It is clear that the dialog manager achieved high score when using deception labels from hierarchical TFN model.

5 Conclusion

In this research, we proposed a new method called hierarchical tensor fusion network for combining features from different modalities. We applied the proposed method to build a classification model for the task of deception detection. Results from our experiment showed that the hierarchical tensor fusion model has the best performance, outperforming the existing methods. We also conducted experiment about the DA selection accuracy of a negotiation dialog system when using output labels from multiple deception detection models and observed that the dialog manager achieves high performance when using labels from the hierarchical TFN fusion.

Acknowledgment

Part of this work was supported by JSPS KAKENHI Grant Number JP17H06101.

参考文献

[1] Hirschberg, J., Benus, S., Brenier, J.M., Enos, F., Friedman, S., Gilman, S., Girand, C., Graciarena, M., Kathol, A., Michaelis, L.,et.al: Distinguishing deceptive from non-deceptive speech, *Interspeech*, pp. 1833–1836 (2005)

[2] Ekman, P. and Friesen, W.V.: Detecting deception from the body or face, *Journal of personality and Social Psychology,*, Vol.29, No.3, pp. 288 (1974)

[3] Tian, L., Moore, J., and Lai, C.,: Recognizing emotions in spoken dialogue with hierarchically fused acoustic and lexical features, *Spoken Language Technology Workshop (SLT)*, pp. 565–572 (2016)

[4] Zadeh, A., Chen, M., Poria, S., Cambria, E., and Morency, L.P.: Tensor fusion network for multimodal sentiment analysis, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 1103–1114 (2017)

[5] The Tung, N., Yoshino, K., Sakti, S., and Nakamura, S.: Impact of deception information on negotiation dialog management, *IWSDS,*, (2018)

[6] Pérez-Rosas, V., Abouelenien, M., Mihalcea, R., and Xiao, Y.: Verbal and nonverbal clues for real-life deception detection. *EMNLP*, pp.2336–2346 (2015)

[7] The Tung, N., Yoshino, K., Sakti, S., and Nakamura, S.: Dialog Management of Healthcare Consulting System by Utilizing Deceptive Information *JSAI*, (2019)