Convolutional Attention Networks for Multimodal Emotion Recognition from Speech and Text Data

Chan Woo Lee[1], Kyu Ye Song[1], Jihoon Jeong[2], Woo Yong Choi[1]*
[1]orbis.ai Inc., Seoul, South Korea
[1]{cwlee, kysong, cchoi}@orbisai.co
[2]Kyung Hee Cyber University, South Korea
[2]jjeong@khcu.ac.kr

Abstract

Emotion recognition has become a popular topic of interest, especially in the field of human computer interaction. Previous works involve unimodal analysis of emotion, while recent efforts focus on multimodal emotion recognition from vision and speech. In this paper, we propose a new method of learning about the hidden representations between just speech and text data using convolutional attention networks. Compared to the shallow model which employs simple concatenation of feature vectors, the proposed attention model performs much better in classifying emotion from speech and text data contained in the CMU-MOSEI dataset.

1 Introduction

Emotion not only is a key driver to people's actions and thoughts, but also is a fundamental part of human communication. As such, emotion recognition technology has become growingly important in improving how humans interact with machines [1]. For instance, emotion recognition has been applied to analyze people's reactions to advertisements, thus creating better neuromarketing campaigns [2]. It has also gained in popularity amongst various other domains such as healthcare [3], customer service, or gaming.

However, effective emotion recognition still remains a challenging task, due to the sheer complexity of generalizing human emotions. For example, individuals express and perceive emotions differently, depending on numerous personal characteristics such as but not limited to age [4], gender [5] and race. Previous efforts have used deep learning based approaches to analyze emotion from single mode of expression, such as facial expression [6] or speech [7]. Since deep learning based approaches have been proven to be effective at learning and generalizing data with high-dimensional feature spaces like images, similar efforts to capture complex feature space of emotional data have also shown promising results with several emotion databases such as EmoDB [8] or IEMOCAP [9]. Unfortunately, human emotion in real-life is often expressed through complex combination of multiple modes of expression, and a lot of information is lost by employing unimodal analysis.

To solve this problem, using deep learning based approaches for multimodal emotion recognition has been researched extensively in recent years. Work of Tzirakis *et al.* uses deep residual networks to extract features from facial expressions, convolutional neural networks to extract features from speech, and concatenates them to input into a LSTM network [10]. Work of Ranganathan *et al.* uses deep believe networks on facial expressions, body expressions, vocal expressions, and physiological signals [11].

Inspired by these approaches, we suggest a new approach to multimodal emotion recognition from just speech and text data. Feature vectors from embedded text sequences and speech spectrograms are extracted using convolutional neural network based architectures. A direct way to learn about the relationship between these two

* Corresponding Author: cchoi@orbisai.co

Proceedings of the First Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML), pages 28–34,
Melbourne, Australia July 20, 2018.

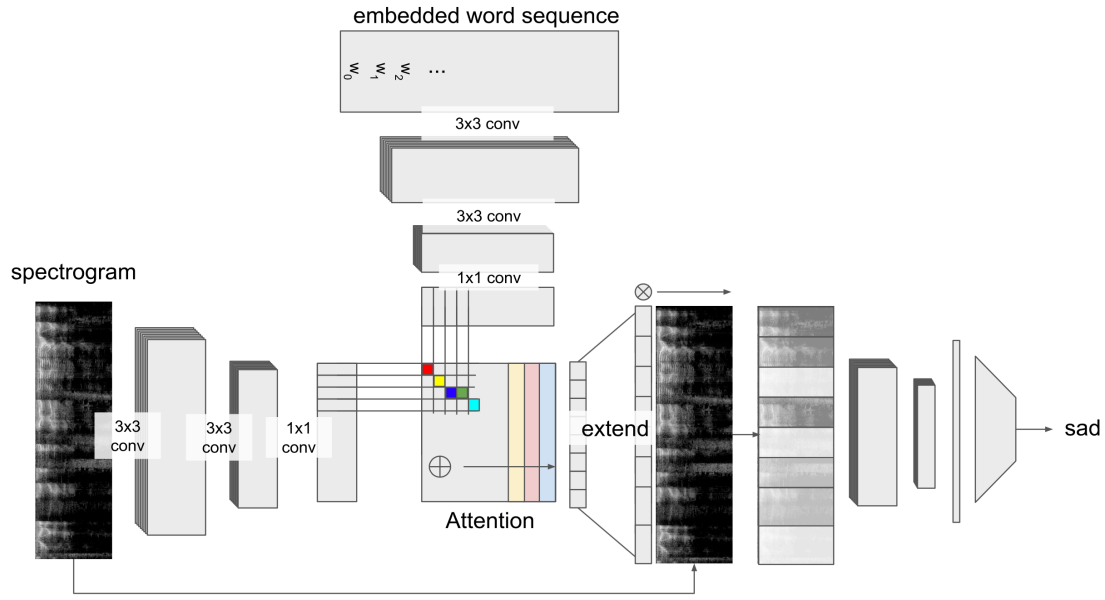

Figure 1 Attention Networks for multimodal representation learning between speech and text data for emotion classification. Separate CNNs are used to extract features from speech spectrograms and embedded word sequences. An attention matrix of m x n dimension is calculated by simply taking a softmax of the dot products of the feature vectors. This attention matrix is then multiplied to the spectrogram input, and goes through a third CNN for emotion classification.

feature vectors would be to utilize a *shallow model*, which is a simple concatenation of two feature vectors. However, since the correlations between feature vectors from speech and text is highly non-linear, it is difficult for a shallow model to properly learn multimodal representations. Therefore, we utilize trainable attention mechanisms to learn nonlinear correlations between these feature vectors. Attention mechanisms also help retain information in the time-domain by forming temporal embedding between two feature vectors. Since speech features and context shares the same time domain, using attention mechanism may help to discover new information for emotion classification. Attention models have previously been successfully applied to tasks such as image caption generation [12], machine translation [13], and speech recognition [14].

To demonstrate the benefits of this new approach, we use it to classify emotions from speech and text data provided in the CMU-MOSEI dataset into six classes: happy, angry, sad, surprised, disgusted, and fear [15]. We also compare this approach to the shallow model approach to show how the attention mechanism can improve capturing of multimodal correlations between text and speech.

2 Model

The attention network shown in figure 1 is comprised of three separate convolutional neural networks: one each for feature extraction from speech spectrogram and word embedding sequence, and one for emotion classifier. Outputs from each of the CNNs from word embedding and spectrogram are used to compute an attention matrix for representing word embedding's correlation to the spectrogram with respect to the emotion labelling. This attention matrix combined with the input spectrogram to be inputted into the CNN based classifier for emotion.

Input embedded word sequences have a size of $R^{e \times L}$ (e: embedding size, L: max sequence length), while input spectrograms have a size of $R^{f \times t}$ (f: frequency range, t: time domain after FT). Word embedding size is fixed at 300, and raw text sentence length was capped at 40 words. Thereby, total word embedding sequence dimension results to 300 by 40. Input spectrograms are derived from transforming raw audio signals with a sample rate of 8000 Hz in the frequency ranges of 0~4kHz, with a fixed size of 200 x 400.

To find the attention matrix between the two feature vectors, 1 by 1 convolution is conducted before calculating the dot product. The resulting

attention matrix has a size of m x n, determined by the last feature vector after 1 by 1 convolution. The column of the attention matrix is the attention of word sequence with respect to the spatial distribution of the input spectrogram. At the extend stage, feature dimensions that are lost due to max pooling in the convolutional layers is recovered. By broadcasting attention values by 2^P, where P is the number of max pooling layers applied, attention values applied to the entire width of the spectrogram.

Attention values are calculated using the following equations:

$$a_{it} = \frac{exp(e_i \cdot f_t)}{\sum_{t=1}^{T} \exp(e_t \cdot f_t)} \quad (1)$$

$$c_t = \sum_{t=1}^{m} a_{it} f_t \quad (2)$$

e_i stands for the word embedded latent vector, while f_t stands for the spectrogram latent vector. By taking a dot product of e_i and f_t and taking a softmax of it, we are able to calculate a_{it}. Since taking a dot product of e_i and f_t essentially equates to calculating the similarity between to vectors, a_{it} is the similarity distribution with respect to time domain. Next, by multiplying a_{it} and f_t element-wise, c_t can be obtained, which essentially is the input spectrogram with attention information added. As shown in Figure 1, the attention matrix can be constructed with m x n dimensions, and when visualized looks like Figure 2.

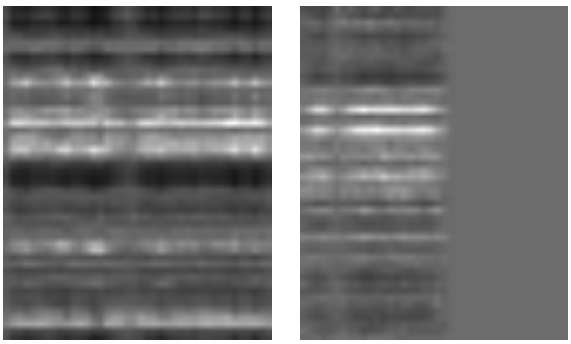

Figure 2 Visualization of the attention matrix. Row means time domain matching the input spectrogram, column means word sequence

After the model learns the representation of each features for attention, the last CNN layer computes the weighted sum of all the information extracted from the attention input. The output vector is then fed into a fully connected softmax layer for classification.

3 Data and preprocessing

3.1 Dataset

We use audio and text data from CMU-Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) dataset for all experiments [15]. The videos, totaling 23,141 files, are chosen from YouTube speakers including various topics and monologue, and are gender balanced.

Annotations consist of six emotion indexes: sadness (2843), angry (6794), happy (10028), disgust (1845), surprise (349), fear (817) with value ranges of [0,4.6], and sentiment label with a value range of [-3,3]. The dataset is organized by video IDs and corresponding segments with six emotion and sentiment labels. Video IDs are then further split into segments. The training set consists 3303 video ID and 23453 segments, while the validation set consists of non-overlapping 300 video IDs and 1834 segments.

Text embedding was prepared using GloVe word2vec method. Each word embedding is fixed at a length of 300. The duration of each word utterance is also provided by the P2FA forced alignment [15].

3.2 Data preprocessing

Speech raw signals are converted to spectrograms before being input into the attention network using Short Time Fourier Transform (STFT) after resampling with a reduced sample rate from 44100 Hz to 8000Hz, as seen in Figure 3. Hamming window is used during STFT, and the length of each segment is 800. The transformed spectrogram is then converted to log-scale to make the vertical axis units of dB, with a frame size of 200x400.

4 Experimental results

In this section, we describe the experiment methodologies and report the recognition performance proposed attention network architecture on the CMU-MOSEI dataset [15].

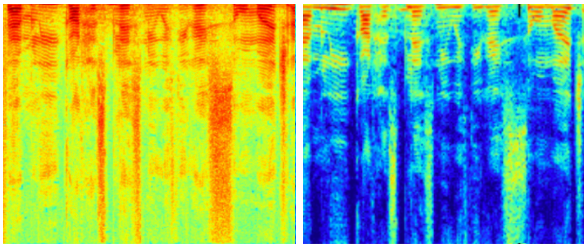

Figure 3 Speech spectrogram after STFT (Left: after STFT, Right: log scale)

4.1 Methods

All models are trained with the training dataset provided by the ACL 2018 Multimodal Challenge. This training dataset is a subset of the entire CMU-MOSEI dataset. The models are validated using the provided validation dataset, again as part of the Challenge. Two sets of experiments are conducted: First, the shallow model architecture (Figure 3) is trained with the training set. The proposed attention network architecture is trained end-to-end, and validated for performance. We then train a shallow model as outlined in Figure 3 to use as a baseline to track how much improvement the attention network provides in learning the correlation between word embedding and corresponding spectrogram features.

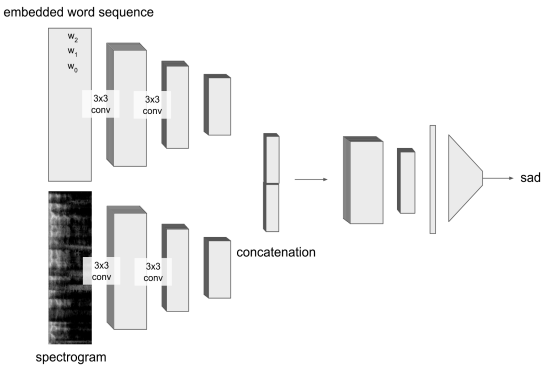

Figure 4 Shallow model diagram

4.2 Hyperparameters

Stochastic gradient descent with a set learning rate is employed during training. For regularization, dropout is applied to the last hidden layer. The system's hyperparameters are: 32 kernels with 3 kernel size; a batch size of 32; a dropout rate of 0.1; learning rate of 1e-3; a pool size of 2 and stride of 2; the dense layer units after final CNN are 1024, 512, and 128 for all configurations.

4.3 Evaluation

For each experiment, we report an overall accuracy (each sentence across the dataset has an equal weight; weighted accuracy) and a class accuracy (first evaluated for each emotion and then averaged; unweighted accuracy). All the classification results are listed in Tables 1-2, including precision, recall, and f-1 score. Confusion matrices are also provided to show how well the model correctly classifies each emotion, using the top-1 class prediction as a metric.

4.4 Experiment 1: shallow model

In this section, we report the results of training the shallow model with the CMU-MOSEI dataset. Since the shallow model is a common and the simplest method of multimodal emotion classification, we use it as a baseline model for comparison.

The overall validation accuracy (weighted) is 83.11% and class validation accuracy (unweighted) is 77.23% as shown in Table 1. The multi-class confusion matrix is shown in Figure 5, showing the highest accuracies for anger and happy emotions, and lowest accuracies for fear and surprise emotions.

Emotion	Precision	Recall	f-1 score
sadness	0.82	0.65	0.73
happy	0.93	0.88	0.91
anger	0.75	0.90	0.82
disgust	0.75	0.75	0.75
surprise	0.98	0.55	0.70
fear	0.83	0.63	0.72
average	0.85	0.84	0.84
class accuracy	77.23%	**Overall accuracy**	83.11%

Table 1 The results of shallow model

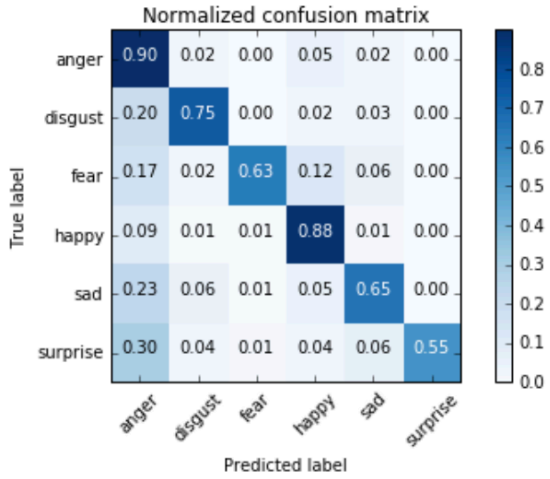

Figure 5 Confusion matrix of shallow model

4.5 Experiment 2: attention model

In this section, we report the results of attention model to compare to the baseline results.

The overall accuracy (weighted) is 88.89% and class accuracy (unweighted) is 84.08 % as shown in Table 2 for the attention model, a significant improvement from the same metrics of shallow model. According to the confusion matrix shown in Figure 6, validation accuracies have increased throughout all emotion classes compared to the baseline.

Emotion	Precision	Recall	f-1 score
sadness	0.88	0.86	0.87
happy	0.92	0.92	0.92
anger	0.85	0.92	0.88
disgust	0.88	0.81	0.84
surprise	0.98	0.62	0.76
fear	0.94	0.65	0.77
average	0.89	0.89	0.89
class accuracy	84.08%	**Overall accuracy**	88.89%

Table 2 The results of attention model

5 Discussion

Comparing the attention model to the shallow model, shallow model utilizes a superficial feature concatenation, while attention model calculates the similarity between two feature vectors that can be trained with learnable weights. In the context of the feature space, concatenating two feature vectors in the shallow model essentially is a simple increase in dimensionality. On the other hand, the feature space in the attention model is fixed to the audio feature space. However, since the features now depend on a new variable called attention, the model can selectively utilize different features in the audio feature space to different extents for better classification. In other words, text data now plays an important role in determining whether a speech feature is important or not in classifying certain emotions, an especially important benefit for training datasets with limited size or data balance.

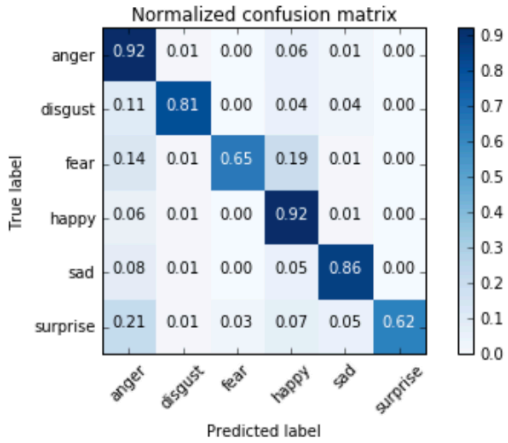

Figure 6 Confusion matrix of attention model

In addition, correlation information between text and speech with respect to the time domain can be easily lost when shallow concatenation is utilized. Meanwhile, calculation of the attention matrix requires matrix multiplication between embedded word and spectrogram feature for a given time. Hence, time series information is retained in the calculated attention matrix through temporal embedding, and to the resulting attention applied spectrogram. Since context and its vocal style of delivery plays an important role in communicating emotion, retaining the time information provides huge benefits in classifying emotions from just speech and text.

Furthermore, while the shallow model is merely an analysis of a union of text and speech infor-

mation, the proposed attention model aims to discover new meaningful methods of how two feature vectors intersect. In other words, shallow model is highly single feature dependent, while attention model is not. This means that if each of the feature vectors contain inadequate information to begin with, shallow model will perform much worse than attention model.

Since the attention model provides newly discovered correlation between the two feature vectors, this new information can be used in ensemble with the original text and speech feature vectors.

Of course, attention models aren't silver bullets in choosing the desired features and discarding the rest. Without careful training of the model, distribution of the attention values can flatten out. For instance, if the input data contains too much padding, and the network has a big bias causing little optimization, the feature vector used to calculate the attention values will approximate to 0, and subsequently attention values will also approximate to 0. One possible solution is the utilize loss masking on the padding of the input data so that a more dynamic softmax distribution in the attention matrix can be obtained.

It is worth noting that for both experiments, f-1 scores of select classes, namely happy and anger are much higher than those of other classes. This is mainly due to a considerable class imbalance of the training set, in which ~44% of the data is happy, and ~30% of the data is angry.

6 Conclusion

The attention model proposed for multimodal emotion recognition from speech and text data provides an effective method of learning about the correlation between the two output feature vectors from separate yet jointly trained CNNs. This method is especially effective for correlation information between speech and text, because the context and the way it is delivered plays a crucial role in affective communication, and the attention model retains temporal information well throughout its model. For future work, syncing the input text and speech data in the temporal dimension may help the attention network focus on learning the relationship between one speech segment and one word, instead of the relationship between whole speech segment and whole text segment.

References

[1] Arkin, R. C.; Fujita, M.; Takagi, T.; and Hasegawa, R. 2003. An ethological and emotional basis for human–robot interaction. Robotics and Autonomous Systems 42(3):191–201.

[2] F.Burkhardt, J.Ajmera, R.Englert, J.Stegmann, and W.Burleson, "Detecting anger in automated voice portal dialogs," in Proc. Annu. Conf. Int. Speech Commun. Assoc., 2006, pp. 1053–1056.

[3] Q.Ji, Z.Zhu, and P.Lan, "Real-time non intrusive monitoring and prediction of driver fatigue," IEEE Trans. Veh. Technol., vol. 53, no. 4, pp. 1052– 1068, Jul. 2004.

[4] A.Mill, J.Alliketal.,"Age-related differences in emotion recognition ability: a cross-sectional study." Emotion, vol. 9, no. 5, p.619, 2009.

[5] T.Vogtand, E.Andre´, "Improving automatic emotion recognition from speech via gender differentiation," in Proc. Language Re-sources and Evaluation Conference (LREC 2006), Genoa, 2006.

[6] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).

[7] C.N.Anagnostopoulos, T.Iliou, andI. Giannoukos, "Features and classifiers for emotion recognition from speech : A survey from 2000 to 2011," Artif. Intell. Rev., vol. 43, no. 2, pp. 155–177, 2015.

[8] Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., and Weiss, B. A database of german emotional speech. In Proc. INTERSPEECH 2005, Lissabon, Portugal (2005), pp. 1517–1520.

[9] Busso, Carlos, et al. "IEMOCAP: Interactive emotional dyadic motion capture database." Language resources and evaluation 42.4 (2008): 335

[10] Tzirakis, P., Trigeorgis, G., Nicolaou, M. A., Schuller, B. W., & Zafeiriou, S. (2017). End-to-end multimodal emotion recognition using deep neural networks. IEEE Journal of Selected Topics in Signal Processing, 11(8), 1301-1309.

[11] Ranganathan, H., Chakraborty, S., & Panchanathan, S. (2016, March). Multimodal emotion

recognition using deep learning architectures. In Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on (pp. 1-9). IEEE.

[12] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov,R. S. Zemel, and Y. Bengio, “Show, attend and tell: Neural image caption generation with visual attention.” in ICML, vol. 14,2015,pp.77–81.

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

[14] J.K.Chorowski, D.Bahdanau, D.Serdyuk, K.Cho, and Y.Ben-gio, “Attention-based models for speech recognition,” in Advances in Neural Information Processing Systems, 2015, pp. 577–585.

[15] Zadeh, Amir, et al. "Human Multimodal Language in the Wild: A Novel Dataset and Interpretable Dynamic Fusion Model" Association for Computational Linguistics (2018)

[16] Poria, Soujanya, et al. "Context-dependent sentiment analysis in user-generated videos” Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) pp. 873-883 (2017)

[17] Zadeh, Amir, et al. "Multi-attention recurrent network for human communication comprehension" arXiv preprint arXiv:1802.00923 (2018)

[18] Poria, S., Cambria, E., Hazarika, D., Majumder, N., Zadeh, A. and Morency, L.P., 2017. Context-dependent sentiment analysis in user-generated videos. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (Vol. 1, pp. 873-883).

[19] Poria, S., Cambria, E., Hazarika, D., Mazumder, N., Zadeh, A. and Morency, L.P., 2017, November. Multi-level Multiple Attentions for Contextual Multimodal Sentiment Analysis. In 2017 IEEE International Conference on Data Mining (ICDM) (pp. 1033-1038). IEEE.

[20] Zadeh, A., Liang, P., Vanbriesen, J., Poria, S., Cambria, E., Chen, M., Morency, L., 2018. Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph. Association for Computational Linguistics.