INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Attention Assisted Discovery of Sub-Utterance Structure in Speech Emotion Recognition

Che-Wei Huang, Shrikanth S. (Shri) Narayanan

Signal Analysis and Interpretation Lab (SAIL)
University of Southern California
cheweihu@usc.edu, shri@sipi.usc.edu

Abstract

Recently, attention mechanism based deep learning has gained much popularity in speech recognition and natural language processing due to its flexibility at the decoding phase. Through the attention mechanism, the relevant encoding context vectors contribute a majority portion to the construction of the decoding context, while the effect of the irrelevant ones is minimized. Inspired by this idea, a speech emotion recognition system is proposed in this work for an active selection of sub-utterance representations to better compose a discriminative utterance representation. Compared to the baseline of a model based on the uniform attention, i.e. no attention at all, an attention based model improves the weighted accuracy by an absolute of 1.46% (and relative 57.87% to 59.33%) on the emotion classification task. Moreover, the selection distribution leads to a better understanding of the sub-utterance structure in an emotional utterance.

Index Terms: attention mechanism, speech emotion recognition

1. Introduction

Emotion plays an important role in our daily lives for effective communication. The affective information is often encoded and conveyed through various forms of human behavioral signals [1], including speech, facial expression, (body) language and so on. Emotion recognition aims at decoding the emotional information based on the corresponding multimodal cues. Speech emotion recognition, focused on characterizing the emotional content from the audio signals, has been an active research topic during the last several years [2, 3]. Despite the advances in the field, it is still a challenging task with considerable room for improvement.

Due to the subtlety in human emotion, the type of acoustic features that effectively characterize the affective content in speech is still an open question [4, 5]. With recent advances in machine learning, deep neural network (DNN) has emerged as a powerful tool for pattern classification. In particular, DNN is well-known for its ability to extract high-level representations layer by layer. For example, Han et al. [6] employed a DNN to learn the high-level representations of sub-utterances. A DNN is deep in the model architecture for affording enhanced expressiveness, but it does not consider the temporal information embedded within data.

On the other hand, human emotion is context sensitive with long-range time dependencies. In this context, better modeling of emotion requires an architecture that explicitly takes into account its sequential nature. Metallinou et al. [7] demonstrated that a bi-directional recurrent neural network (RNN) with the long-short term memory gating mechanism (BLSTM) is capable of capturing these long-term dependent contextual affective information in a sequence of consecutive utterances. More recently, Lee et al. [8] proposed an algorithm based on the connectionist temporal classification (CTC) approach to extract frame level representations with regard to its dynamics within an utterance. These studies highlight the potential of the emerging NN architectures in capturing temporal details of emotion expression.

Many speech emotion corpora however provide only utterance level annotations. Speech emotion recognition systems usually have to provide an output to the utterance-level, because of the utterance-based annotation of many speech emotion corpora. Recent, Ghosh et al. [9] proposed to pre-train each frame directly from spectrograms with a deep auto-encoder rather than using the Mel-Frequency Cepstral Coefficients (MFCC). A BLSTM and a multi-layer perceptron are trained on the bottleneck layer activations. This work indicated that the average of hidden vectors at the output of a BLSTM is more discriminative as an utterance representation, compared to the conventional choice of the hidden vector at the last time step.

Although the annotation is given at the utterance level, not all frames within an utterance contain the relevant emotion content, which may be unevenly distributed even among the frames containing emotional information. The recent development in attention mechanism based deep learning [10, 11, 12] adds another functionality to the RNN architecture specifically for addressing problems with a correlated structure between the input and the output sequences. The key idea behind the attention mechanism is to soft/hard align the input-output sequences so that in the decoding phase the major contribution of the context comes from the corresponding encoded information. Speech emotion recognition at the utterance level can be formulated as a many-to-one sequence-to-sequence learning, where the input sequence is the stream of acoustic frames and the output sequence is the emotion label. In this scenario, the average of hidden vectors suggested in [9] becomes a special case of the attention mechanism, i.e. the uniform attention or no attention at all.

In this work, we apply attention mechanism based BLSTM modeling to speech emotion recognition. Our hypothesis is that such a model would result in a more discriminative utterance representation than the one provided by a model without the attention mechanism. In addition, we will study the structure within the sequence of frames in the composition of an utterance representation.

The outline of this paper is as follows. The next section covers an overview of related work in detail. In the third section we will introduce the proposed algorithm, followed by a section devoted to the experiments. The last section will conclude with our findings.

http://dx.doi.org/10.21437/Interspeech.2016-448

2. Related Work

Deep models, already the state-of-the-art in many areas including speech recognition and computer vision, have been also successfully applied to the emotion recognition task. Han et al. [6] proposed a DNN-ELM algorithm for speech emotion recognition in two steps; the first part extracts the high-level representations of the high-energy segments within an utterance, where a segment is defined to be 265 ms long in order to have enough context. Choosing the high-energy segments is based on the assumption that only these segments carry the relevant emotion information. These segments also share the same label with the utterance they belong to. A DNN trained on these segments predicts a softmax distribution over emotion classes per segment, which is then regarded as a high-level representation of the segment. The second part forms the utterance representation by aggregating the statistical functionals estimated from these sub-utterance representations. An extreme learning machine (ELM) then maps each utterance representation to an emotion state.

However, emotion can be manifested in speech through a long and variable range of temporal dependencies. A simple fixed-length segment may be insufficient for describing the dynamics. Lee et al. [8] employed a BLSTM model with a CTC loss function to encode the temporal information into the frame representations. In essence, a BLSTM transforms the sequence of frames in an utterance into a sequence of high-level representations bearing with the contextual information. The CTC loss function serves the purpose of integrating out all possible alignments between the frames and the sequence of Null and Emo labels. Similarly, certain statistical functionals of these sub-utterance representations form the utterance representation, and the classification task is carried out by a following ELM. Perhaps our work is most similar to Lee et al.'s, but there are several fundamental differences we would like to underscore. First of all, the CTC approach probabilistically minimizes the alignment mismatch via the maximum a posteriori inference while the attention mechanism is deterministic, which could result in a non-monotonic alignment and is therefore open to more future applications [10]. Second, in order to generate all possible alignments, the authors assumed that at least one frame in each voiced region contains the relevant emotion information and a Markovian property for the label transition within a voiced region. The attention mechanism does not rely on such assumptions. Last but not the least, they focused on learning the frame-level contextual representations constrained by the assumption, and constructed the utterance representation via statistical functionals, whereas our goal is to study the feasibility of an end-to-end emotion recognition system and to compare BLSTM models with and without the attention mechanism in the formation of the utterance representations.

A recent work by Ghosh et al. [9] found that the average of all context representations is more indicative of emotion states than the one at the last time step. Either letting the algorithm to passively accumulate the utterance information into the last hidden vector or building the utterance representation in an uninformed way via averaging over all hidden vectors may not be the optimal approach. We hypothesize that an attention mechanism based model could actively help with a structurally meaningful composition of the utterance representation.

3. Proposed Algorithm

Sequence to sequence learning based on the RNN/LSTM architecture has become one of the most popular models in dealing with sequential data. Suppose the input and output sequences

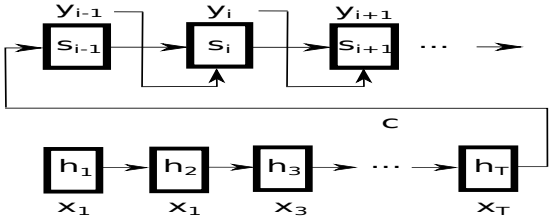


Figure 1: *A diagram of the encoder-decoder framework*

of equal length are $(\mathbf{x}_1, \cdots, \mathbf{x}_T)$ and $(\mathbf{y}_1, \cdots, \mathbf{y}_T)$, respectively. A single-layer RNN can be characterized by these two equations:

$$\mathbf{h}_t = \sigma_h(\mathbf{W}^{\mathbf{hx}}\mathbf{x}_t + \mathbf{W}^{\mathbf{hh}}\mathbf{h}_{t-1}), \quad (1)$$

$$\mathbf{y}_t = \sigma_y(\mathbf{W}^{\mathbf{yh}}\mathbf{h}_t), \quad (2)$$

where σ_h and σ_y are activation functions. The hidden vector, $\mathbf{h}_t$, is the cell state of a RNN at time t, exemplifying the memory stored in the model. The matrices $\mathbf{W}^{\mathbf{ba}}$ indicate the transformations from a vector of type $\mathbf{a}$ to a vector of type $\mathbf{b}$. Here, the superscripts $\mathbf{x}$, $\mathbf{h}$ and $\mathbf{y}$ stand for the types of input, hidden and output vectors, respectively. To allow the input and the output sequences to have different lengths, Cho et al. [13] and Sutskever et al. [14] came up with the so-called encoder-decoder framework. This framework first maps the input sequence to a fixed dimensional vector as a representation for the entire sequence, called the context vector $\mathbf{c}$. The mapping from an input sequence to its context vector is referred to as an encoder RNN/LSTM. From there on, a decoder RNN/LSTM takes the context vector as its input and generates the output sequence conditioned on the context vector. In the equations above, $\mathbf{h}_T$ is regarded as the context vector $\mathbf{c}$ for it being a non-linear function of the entire input sequence. Fig. 1 describes the encoder-decoder framework, where $\mathbf{s}_i$ is the cell state of the decoder.

Speech emotion recognition can be formulated as a many-to-one sequence to sequence learning problem, where the input sequence is the stream of frames in an utterance, and the output sequence is the predictive distribution of the emotion states and of length one. One immediate advantage of this formulation is to encapsulate the modeling process from the sub-utterance level to the utterance level representations in a systematic way, rather resorting to statistical functionals. This opportunity toward an end-to-end system is actually one attractive point the attention mechanism offers [10]. Another potential advantage may be the future accommodation for more diverse tasks, where speech emotion recognition could serve as a building component of a larger system.

In spite of the merit of a minimal assumption on the sequence structure, Sutskever et al. [14] found that reversing the the input order improves the performance for free. This phenomenon stems from the learning mechanism in the LSTM architecture. Notice the term $\mathbf{W}^{\mathbf{hh}}\mathbf{h}_{t-1}$ in Eq. (1). At every step into the time, the memory stored in the cell state is scaled by a matrix $\mathbf{W}^{\mathbf{hh}}$. Therefore, $\mathbf{h}_T$ is more representative of $\mathbf{x}_T$ and less of previous inputs. Since the decoding is in the order of time, this choice of a context vector would demand a very long term memory if the length of the input sequence is long. By reversing the input order, $\mathbf{h}_T$ explicitly exploits the sequence structure to be a more suitable candidate as a context vector. But, $\mathbf{h}_T$ is not the only way to make a context vector.

3.1. Baseline: Uniform Attention based BLSTM model

One variant of the LSTM architecture is to equip it with bi-directional information, i.e. memory from both the past and the future, where the hidden vector at time step t is the concatenation of those vectors from both directions. A BLSTM generally outperforms a LSTM when the amount of training data is sufficient, and hence the BLSTM architecture is gradually becoming the new standard. With a BLSTM it does not matter which order the input sequence is.

We can view these previous options for constructing the context vector as instances of a *passive* approach, where the algorithm accumulates the information in the cell state until the end of input sequence. On the other hand, an *active* alternative should be able to effectively select the semantically relevant hidden vectors $\mathbf{h}_t$ from *all* of the time steps. In this regard, perhaps to take the average of these hidden vectors $\mathbf{h}_t$ is the most simplistic option. Despite its simplicity, Ghosh et al. [9] found it superior than the passive approach based on a BLSTM.

We will take this model as our baseline in this work. Since averaging amounts to an application of the uniform distribution, that is, an uninformed prior, we call this model the uniform attention based model, implying no attention at all.

3.2. Attention mechanism based BLSTM model

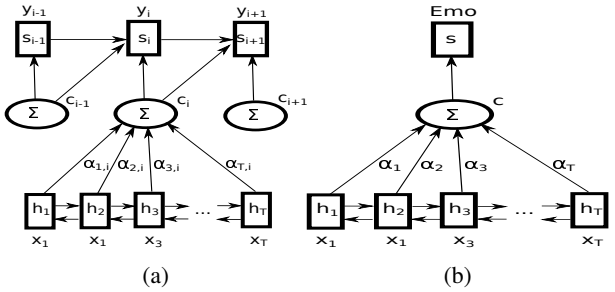


Figure 2: *The subplot (a) depicts the working process of the attention mechanism at the moment of decoding the output* $\mathbf{y}_i$*. The subplot (b) describes the application of the attention mechanism to speech emotion recognition.*

Attention mechanism based recurrent neural network has found its application in a variety of sequence to sequence learning tasks, including speech recognition [10], machine traslantion [11], part-of-speech tagging [12], to name just a few. The basic idea is to select relevant encoded hidden vectors through an informative sequence of weights, called the attention weights, in the decoding phase. Such an architecture provides an opportunity toward building an end-to-end system, which is already a popular research topic in speech recognition.

In the setting of the attention mechanism, the context vector $\mathbf{c}$ is no longer static through the whole decoding iterations. Therefore, for decoding each output $\mathbf{y}_i$, the corresponding context vector is denoted as $\mathbf{c}_i$ with a subscript index i to highlight the variability at each time step. Fig. 2a illustrates the working process of the attention mechanism at the moment of decoding the output $\mathbf{y}_i$, where the sequence of $\mathbf{h}_t$ is the hidden vectors of a BLSTM at each time step, the sequence of $\alpha_{t,i}$ the attention weights for composing the context vector $\mathbf{c}_i$, and $\mathbf{s}_i$ the cell state of the decoder LSTM. The Σ notation stands for the summation as usual.

The decoding of the i-th output $\mathbf{y}_i$ relies on the relevant context $\mathbf{c}_i$. Borrowing a similar set of notations from [10], we can summarize them symbolically:

$$\alpha_i = Attend(\mathbf{s}_{i-1}, \alpha_{i-1}, \mathbf{h}), \quad (3)$$

$$\mathbf{c}_i = \sum_{t=1}^{T} \alpha_{t,i}\mathbf{h}_t, \quad (4)$$

$$\mathbf{y}_i = Generate(\mathbf{s}_{i-1}, \mathbf{c}_i). \quad (5)$$

In the formulation of speech emotion recognition, the output sequence has a length of one, and thus Eq. (3, 5) can be simplified into

$$\alpha = Attend(\mathbf{h}), \quad (6)$$

$$\mathbf{y} = Generate(\mathbf{c}), \quad (7)$$

while Eq. (4) remains unchanged. The *Attend()* function learns the attention weights based on the input sequence of $\mathbf{h}_t$ and the *Generate()* function symbolizes the decoder. Moreover, since the output consists of no temporally dependent structure in itself, the decoder need not to be a recurrent neural network. Instead, a DNN should suffice. To be noted, the attention mechanism is still in its burgeoning stage and there has not been a conventionally converged view on the implementation of the attention mechanism. Chorowski et al. [10] proposed to distinguish three different implementations of the attention mechanism: the location-based, content-based and hybrid attention mechanisms. Limited by the simplicity of the output sequence, the content-based approach is the only viable one in this study as described by Eq. (6).

The content-based attention weights in our work follow the implementation in the literature [10, 12]

$$\alpha_t = softmax((\mathbf{w}^a)^T \sigma_a(\mathbf{W}^{ah}\mathbf{h}_t)). \quad (8)$$

We further simplified the formula into:

$$\alpha_t = softmax((\mathbf{w}^a)^T \mathbf{h}_t) \quad (9)$$

by removing the intermediate hidden layer. The need for a further simplification is simply to prevent over-fitting. The extra component for computing the attention weights has the number of parameters proportional to the length of the input sequence. Therefore it is rational to simplify its architecture as long as the spirit of the attention mechanism remains. On the output side, we employed a decoder DNN for generating the output emotion state. Fig. 2b gives a diagram of the attention based speech emotion recognition system.

4. Experiments

To evaluate the effectiveness of the proposed algorithm, we performed our experiments on the Interactive Emotional Dyadic Motion Capture (IEMOCAP) database [15]. IEMOCAP consists of rich information about speech, facial expressions and hand gestures of ten actors in dyadic sessions. The actors were asked to perform selected emotional scripts and pre-defined improvised scenarios. There are five sessions in the corpus with two actors, one from each gender, in each session. The total amount of data in this modest sized corpus amounts to roughly 12 hours. For speech emotion recognition, we only considered the audio tracks labelled as one of the four categorical emotion types, including *Angry*, *Happy*, *Sad* and *Neutral* as they are the majority of the categorical emotion types, where the numbers of utterances in each category are 1103, 595, 1084 and 1708, respectively, with a sum of 4490. In the experiments, we followed

a leave-one-speaker-out approach for cross validation. Specifically, we took four sessions as the training data, while in the remaining one session, one speaker is used for validation, model selection and parameter tuning and the other for testing.

The low-level descriptors comprised of the 13-dimensional MFCC including the zero-th order coefficient, the pitch and their first order derivatives. Therefore, each frame had a dimensionality of 28. The baseline in our work was a uniform attention mechanism based BLSTM. It had two LSTM layers (128 forward and 128 backward). The size of the hidden layer was chosen by cross validation. The context vector $\mathbf{c}$ then passed through a softmax layer to give the emotion state distribution.

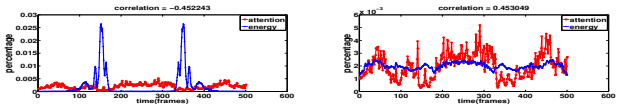
Next, we substituted the attention mechanism defined in Eq. (8, 9) for the averaging part in the baseline. In our proposed model, the BLSTM and the attention mechanism were the largest two components but they did not need to be coupled together. To mitigate over-fitting, we took a greedy approach by training these two components separately. A greedy layer-wise training technique could avoid co-adaptation, a phenomenon which tends to cause over-fitting during the training phase. This approach is similar to the method proposed by Hinton et al. in [16] for training deep models except for the fact that our model was a discriminative one. In addition to the greediness, we also applied dropout to the scaled hidden vectors ($\alpha_t \mathbf{h}_t$ in Fig. 2b) before the summation. Dropout [17] is known to be an effective tool to regularize training and consequently to reduce the chance of over-fitting. A final model further added a hidden layer to the decoder. All model architectures and parameters were selected based on optimizing the un-weighted accuracy of the validation set. To begin with, we extracted the hidden vectors $\mathbf{h}_t$ from the baseline model, and then introduced the modifications step by step. The final model consisted of an encoder BLSTM with a cell size of 256, a dropout with a probability of 0.5 and a hidden layer of size 128 in the decoder.

Table 1: *The performance of the proposed algorithms in comparison to the baseline model. In each row, the top number is based on Eq. (8), and the bottom one is based on Eq. (9). UA stans for the un-weighted accuracy and WA for the weighted accuracy.*

	UA (%)	WA (%)
Baseline	48.54	57.87
Baseline + Att.	48.13 48.70	55.88 56.38
Greedy + Att.	48.71 49.30	56.41 57.33
Greedy + Dropout + Att.	48.82 49.21	56.18 57.36
Greedy + Dropout + Att. + MLP	49.58 49.96	57.26 59.33

We observed that an implementation of the BLSTM coupled with the attention mechanism resulted in an over-fitted model. In *Baseline+Att.*, only the model based on Eq. (9) slightly improved upon *Baseline* in terms of the un-weighted accuracy (48.70%). With additional techniques to prevent over-fitting one at a time, *Greedy+Att.* and *Greedy+Dropout+Att.* gradually outperformed *Baseline* in terms of UA. At last, the final model gained an improvement by a definite margin in both measures. The results are summarized in Table 1.

Not only did the attention mechanism help in the task of emotion recognition, it also provided useful insights into the sub-utterance structure. Fig. 3 gives an example plot of the attention weights and the frame energy curves based on $\mathbf{x}_t$ and $\mathbf{h}_t$, respectively. Note first that the attention mechanism followed the energy curve of the hidden vectors $\mathbf{h}_t$ instead of the raw PCM signals. Also, from the right panel in Fig. 4, the average of the correlations over all utterances is rather small.



(a) (b)

Figure 3: *The attention weights and the frame energy curve of an utterance. In the left panel, the frame energy is based on the raw PCM signals, while in the right panel it is based on the hidden vectors* $\mathbf{h}_t$.

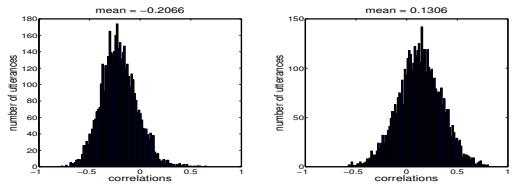


Figure 4: *The histograms of correlations between the attention weights and the frame energy curve. The left panel is based on the raw PCM signals, while in the right panel it is based on the hidden vectors* $\mathbf{h}_t$.

5. Discussion

Based on the improved performance, it is clear that the attention mechanism indeed provides an informative selection of frames. The result also shows that the selection distribution need not to be correlated to the frame energy curve. However, the improvement margin is not very significant. On this aspect, we conjecture there are two factors. On the one hand, the greedy approach might lead to a sub-optimal region that the attention mechanism has little to contribute. On the other hand, even though the attention mechanism offers an active selection distribution, there is a room for further improvement. Currently, the combination of frames is through a weighted sum, which is linear. A non-linear combination of the attention weights and the frame hidden vectors would be an interesting direction for future study.

6. Conclusion

In this work, we investigated the application of attention mechanism based BLSTM model to the task of speech emotion recognition. Despite the limited amount of data to fit such a big model, we adopted an greedy approach to minimize the effect of over-fitting. The preliminary experimental results show that the attention mechanism based system outperforms an system without the attention mechanism. Further, we also found out that the attention selection distribution is not just correlated to the frame energy curve, underscoring more complex speech property evolution related to emotion.

7. Acknowledgements

This research is supported by NSF, NIH, DARPA and Google Inc.

8. References

[1] S. S. Narayanan and P. Georgiou, "Behavioral signal processing: Deriving human behavioral informatics from speech and language," *Proceedings of IEEE*, vol. 101, no. 5, pp. 1203 – 1233, May 2013.

[2] C. M. Lee and S. S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 2, pp. 293–303 [IEEE Signal Processing Society Best Paper Award 2009], Mar. 2005.

[3] M. Grimm, K. Kroschel, E. Mower, and S. S. Narayanan, "Primitives-based evaluation and estimations of emotions in speech," *Speech Communication*, vol. 49, no. 10-11, pp. 787–800, Nov. 2007.

[4] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recogn.*, vol. 44, no. 3, pp. 572–587, Mar. 2011.

[5] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Commun.*, vol. 53, no. 9-10, pp. 1062–1087, Nov. 2011.

[6] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association, Singapore, September 14-18, 2014*, 2014, pp. 223–227.

[7] A. Metallinou, M. Wollmer, A. Katsamanis, F. Eyben, B. Schuller, and S. Narayanan, "Context-sensitive learning for enhanced audiovisual emotion classification," *IEEE Trans. Affect. Comput.*, vol. 3, no. 2, pp. 184–198, Apr. 2012.

[8] J. Lee and I. Tashev, "High-level feature representation using recurrent neural network for speech emotion recognition," in *Interspeech 2015*. ISCA - International Speech Communication Association, September 2015.

[9] S. Ghosh, E. Laksana, L.-P. Morency, and S. Scherer, "Learning representations of affect from speech," in *arXiv:1511.04747*.

[10] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Advances in Neural Information Processing Systems 28*, 2015, pp. 577–585.

[11] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *CoRR*, vol. abs/1409.0473, 2014.

[12] O. Vinyals, L. u. Kaiser, T. Koo, S. Petrov, I. Sutskever, and G. Hinton, "Grammar as a foreign language," in *Advances in Neural Information Processing Systems 28*, 2015.

[13] K. Cho, B. Van Merriënboer, Ç. Gülçehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder–decoder for statistical machine translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Oct. 2014, pp. 1724–1734.

[14] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in Neural Information Processing Systems 27*, 2014, pp. 3104–3112.

[15] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Journal of Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, Dec. 2008.

[16] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, Jul. 2006.

[17] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014.