Multimodal Embeddings From Language Models for Emotion Recognition in the Wild

Shao-Yen Tseng, *Member, IEEE*, Shrikanth Narayanan, *Fellow, IEEE*, and Panayiotis Georgiou, *Senior Member, IEEE*

***Abstract*—Word embeddings such as ELMo and BERT have been shown to model word usage in language with greater efficacy through contextualized learning on large-scale language corpora, resulting in significant performance improvement across many natural language processing tasks. In this work we integrate acoustic information into contextualized lexical embeddings through the addition of a parallel stream to the bidirectional language model. This multimodal language model is trained on spoken language data that includes both text and audio modalities. We show that embeddings extracted from this model integrate paralinguistic cues into word meanings and can provide vital affective information by applying these multimodal embeddings to the task of speaker emotion recognition.**

***Index Terms*—Machine learning, unsupervised learning, natural language processing, speech processing, emotion recognition.**

I. Introduction

ACOUSTIC and visual elements in human communication, such as vocal intonation and facial expressions, incorporate semantic information and paralinguistic cues conveying intent and affect [1]. For this reason many multimodal systems have been proposed which integrate information from multiple modalities to improve natural language understanding. This effort has many applications such as in video summarization [2], [3], dialogue systems [4], [5], and emotion and sentiment analysis [6]–[8].

The study of multimodal fusion in affective systems is an especially prevalent and important topic. This follows from the fact that human behavioral expression is fundamentally a multifaceted phenomenon that manifests over multiple modalities [9] and can be more accurately identified through multimodal models [10]. Another factor is the importance of affective information as an ingredient in a variety of downstream tasks such as in language modeling [11], dialogue system design [12], [13], and video summarization [14].

Many multimodal systems for recognition of sentiment, emotion, and behaviors have been proposed in prior work, including recent neural network based approaches. In feature-level fusion, Tzirakis *et al.* [15] combined auditory and visual modalities by extracting features using convolutional neural networks (CNN) on each modality which then were concatenated as input to an LSTM network. Hazarika *et al.* [16] proposed the use of a self-attention mechanism to assign scores for weighted combination of modalities. Other works have applied multimodal integration using late fusion methods [17], [18].

For deeper integration between modalities many have proposed the use of multimodal neural architectures. Lee *et al.* [19] have proposed the use of an attention matrix calculated from speech and text features to selectively focus on specific regions of the audio feature space. The memory fusion network was introduced by Zadeh *et al.* [20] which accounted for intra- and inter-modal dependencies across time. Akhtar *et al.* [21] have proposed a contextual inter-modal attention network that leverages sentiment and emotion labels in a multi-task learning framework.

The strength of deep models arises from the ability to learn meaningful representations of, and association between, features from multiple modalities. This is learned implicitly by the model in the course of training [22]. In this work we propose a model to explicitly learn informative joint representations of speech and text. This is achieved by modeling the dynamic relations between lexical content and acoustic paralinguistics through a language modeling task on spoken language. We augment a bidirectional language model (biLM) with word-aligned acoustic features and optimize the model first using large-scale text corpora, and then followed by speech recordings. In this work we hypothesize that (1) representations from this model capture multimodal information that can be used to improve speaker emotion recognition (SER), and (2) the benefits of such methods increase with more diverse data. We evaluate the effectiveness of the representations in SER on three datasets of increasing diversity with respect to naturalness and recording conditions, namely IEMOCAP, MSP-IMPROV and CMU-MOSEI.

II. Related Work

Lexical representations such as ELMo [23] and BERT [24] have recently been shown to model word semantics and syntax with greater efficacy. This is achieved through contextualized learning on large-scale language corpora which allows internal states of the model to capture both the complex characteristics of word use as well as polysemy due to different contexts. The integration of these word embeddings into downstream models has improved the state of the art in many NLP tasks through their rich representation of language use.

Manuscript received December 13, 2020; accepted March 5, 2021. Date of publication March 11, 2021; date of current version April 7, 2021. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Sandro Cumani. *(Corresponding author: Shao-Yen Tseng.)*

Shao-Yen Tseng and Shrikanth Narayanan are with the Department of Electrical and Computer Engineering, University of Southern California, Los Angeles, CA 90089-0001 USA (e-mail: shaoyen.t@gmail.com; shri@ee.usc.edu).

Panayiotis Georgiou is with the Siri Understanding, Apple Inc., Culver City, CA 90016 USA (e-mail: panayiotis.georgiou@gmail.com).

Digital Object Identifier 10.1109/LSP.2021.3065598

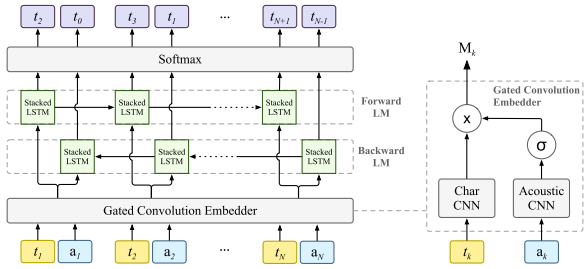

Fig. 1. Architecture of the multimodal bidirectional language model.

To learn representations from multimodal data Hsu *et al.* [25] proposed the use of variational autoencoders to encode inter- and intra-modal factors into separate latent variables. Later, Tsai *et al.* [26] factorized representations into multimodal discriminative and modality-specific generative factors using inference and generative networks. Recent work by Rahman *et al.* [27] has concurrently proposed the infusion of multimodal information into the BERT model. There the authors have combined the generative capabilities of the BERT model with a sentiment prediction task to allow the model to implicitly learn rich multimodal representations through a joint generative-discriminative objective.

In this work we propose to explicitly learn multimodal representations of spoken words by augmenting the biLM model in ELMo with acoustic information. This is motivated from how humans integrate acoustic characteristics in speech to interpret the meaning of lexical content from a speaker. Our work differs from prior work in that we do not include or target any discriminative objectives and instead rely on generative tasks and contextual learning to learn meaningful multimodal representations. As a first step in this direction we adopt the ELMo architecture for its use of a language modeling task. One motivation behind this is to allow better multimodal integration using token-level tasks, as opposed to sequence prediction used in BERT. We show how this unsupervised model can be easily trained with large-scale unlabeled data and also demonstrate the usefulness of the resulting multimodal embeddings on emotion recognition in the wild.

III. Multimodal Embeddings

We extract multimodal embeddings through the use of a bidirectional language model (biLM) infused with acoustic information. The biLM comprises stacked layers of bidirectional LSTMs which operate over lexical and audio embeddings. The lexical and audio embeddings are calculated from respective convolutional layers and combined using a sigmoid-gating function. Multimodal embeddings are then computed using a linear function over the internal states of the recurrent layers. The architecture of the mutlimodal biLM is shown in Fig. 1.

A. Bidirectional Language Model

A language model (LM) computes the probability distribution of a sequence of words by approximating it as the product of conditional probabilities of each word given previous words. This has been implemented using neural networks in many prior work yielding state of the art results [28]. In this work we applied the biLM model used in ELMo, which is based on the character-level RNN-LM [29].

The biLM is composed of a forward and backward LM each implemented by a multi-layer LSTM. The forward LM predicts the probability distribution of the next token given past context while the backward LM predicts the probability distribution of the previous token given future context. Each LM operates on the same input, which is a token embedding of the current token calculated through a character-level convolutional neural network (CharCNN) [30]. A softmax layer is used to estimate token probabilities from the output of the two-layer LSTM in the LMs. The parameters of the softmax layer are shared between the LMs in both directions.

Different from ELMo, our input to the biLM includes acoustic features in addition to word tokens. Now the forward LM aims to model, at each time step, the conditional probability of the next token t_{k+1} given the current token t_k, acoustic features $\mathbf{a}_k$, and previous internal states of the stacked LSTM $\vec{\mathbf{s}}_{k-1}$:

$$P(t_{k+1} \mid t_k, \mathbf{a}_k, \vec{\mathbf{s}}_{k-1}) \tag{1}$$

The backward LM operates similarly but predicts the previous token t_{k-1} given the current token t_k, acoustic features $\mathbf{a}_k$, and internal states resulting from future context $\overleftarrow{\mathbf{s}}_{k+1}$. Details of the acoustic features are given in Section IV-C.

B. Acoustic Convolution Layers

To integrate paralinguistic information into the language model, time-aligned acoustic features of each word are provided in adjunct to word tokens. We add additional convolutional layers at the input of the biLM to compute acoustic embeddings from the acoustic features. The convolutional layers provide a feature transformation of the acoustic features which are then combined with token embeddings using a gating function.

Due to the varying duration of spoken words, acoustic features are zero-padded to a fixed frame size before being passed to the CNN. This is similar to the use of a maximum number of characters per word in the CharCNN. The acoustic CNN is implemented by series of 1-D convolution layers each followed by a max-pooling layer. The final feature map is then projected to the same dimension size as token embeddings to allow for element-wise combination.

C. Multimodal ELMo

We combine token and acoustic embeddings using a sigmoid gating function:

$$\mathbf{M}_k = \mathbf{U}(t_k) \odot \sigma(\mathbf{V}(\mathbf{a}_k)) \tag{2}$$

where $\mathbf{U}(t_k)$ and $\mathbf{V}(\mathbf{a}_k)$ are the embeddings calculated from the word token and acoustic feature, respectively, σ is the sigmoid function, and $\odot$ represents element-wise multiplication. The sigmoid gate is a useful mechanism in language modeling [31] as it allows the network to select relevant features in the token embedding. In our case it serves to modify semantic meaning of words through scaling of the token embedding based on acoustic information. The embeddings after the gated sigmoid function are considered to be multimodal and are used as input to both the forward and backward LM.

Word embeddings are extracted for use in downstream models in a similar fashion to ELMo. That is, we define each word vector as a task-specific weighted sum of all LSTM outputs as well as the input token embedding $\mathbf{M}_k$. To form sentence embeddings for use in downstream models we additionally average over all word vectors in a sentence This approach of obtaining sentence embeddings from a weighted average of word vectors has been shown to be an effective method of information aggregation in many NLP tasks [32].

The final multimodal ELMo (M-ELMo) sentence embedding for a sentence consisting of N words is given as

$$\textbf{M-ELMo} = \gamma \frac{1}{N} \sum_{k=1}^{N} \sum_{j=0}^{L} c_j \mathbf{h}_{k,j} \tag{3}$$

where $\mathbf{h}_{k,j}$ are the concatenated outputs of LSTMs in both directions at the $j\text{th}$ layer for the $k\text{th}$ token and $j = 0$ corresponds to $\mathbf{M}_k$. Values $\{c_j\}$ are softmax-normalized weights for each layer and γ is a scalar value, all of which are task-specific and tunable parameters in the downstream model.

IV. Experiments

A. Multimodal biLM Architecture

The final model architecture proposed in [23] was used for the lexical and recurrent components of the multimodal biLM. This model comprises a character CNN with 2048 character n-gram convolutional filters followed by a two-layer biLSTM ($L = 2$) with 4096 units and a projection size of 512 dimensions.

The architecture of the acoustic CNN was based on keyword spotting CNNs proposed in [33], however we applied 1-D convolution since our acoustic features include non-spatial categories. We also used a smaller kernel size in the time dimension to model acoustic variations at a finer scale. The acoustic CNN comprises three 1-D convolutional layers using kernels of size 3 and a stride of 1. Each layer is followed by a max-pooling function over three frames.

B. Pre-Training the Multimodal biLM

The multimodal biLM was pre-trained in two stages. In the first stage the lexical components of the biLM were optimized prior to the inclusion of acoustic features. This was achieved by training on a text corpus and fixing the acoustic input as zero. The 1 Billion Word Language Model Benchmark [34] was used to train the lexical components of the biLM for 10 epochs. After training, the model achieved perplexities of around 35 which is similar to values reported for pretrained models in [23].

In the second stage of pre-training we optimized the biLM using the multimodal dataset CMU-MOSEI (described in Section IV-D). In our experiments we used text and audio which were not in the testing split of the dataset to train the biLM. In terms of word count CMU-MOSEI contains around 447 K words which is much smaller than the 1-billion word LM benchmark. Therefore, to prevent over-fitting we reduced the learning rate used in the previous stage by a factor of 10 and fine-tuned the biLM for an additional 5 epochs.

C. Features

Since a CharCNN was used as the lexical embedder, input words to the biLM were first transformed into a character mapping then padded to a fixed length. The character-level representation of each word is given as a $c \times l_c$ matrix, where c is the dimension size of the character embedding and l_c is the maximum number of characters in a word.

We used acoustic features extracted using COVAREP (v1.4.2) [35] similar to [21], [36]. There are 74 features in total and include, among others, pitch, voiced/unvoiced segment features, mel-frequency cepstral coefficients, glottal flow parameters, peak slope parameters, and harmonic model parameters.

The acoustic features were aligned with word timings to provide acoustic information for each word. Since the time duration varies between words we padded the number of acoustic frames per token to a fixed length. Thus, word-aligned acoustic features were given as a $d \times l_a$ matrix, where d is the number of acoustic features and l_a is the maximum number of acoustic frames in a word. In our experiments we used a maximum frame length of 2 seconds per word. This corresponds to more than 99.9% of all words in the dataset. We assume any truncated words to be unrepresentative of conventional articulation during conversations (*e.g.*, purposely drawn-out words) which may require specific modeling outside the scope of this study.

D. Datasets

After pre-training we extract sentence embeddings from the multimodal biLM for use in emotion recognition. In our experiments we evaluated on emotion datasets IEMOCAP [37], MSP-IMPROV [38], and CMU-MOSEI.

IEMOCAP is a multimodal emotion dataset consisting of 10 actors displaying emotion in scripted and improvised hypothetical scenarios. This dataset is considered as the least diverse in our experiments having been collected in clean recording conditions in addition to consisting of acted scenarios with few speakers. Although this is not an ideal test set we include it for completeness. Similar to [39] we evaluate on the emotion categories *angry* (1103), *excite* (1041), *happy* (595), *sad* (1084), *frustrated* (1849), *surprise* (107), and *neutral* (1708) for a total of 7487 utterances. We report results from 10-fold cross validation where in each fold one speaker is used as the test set while their conversation partners and remaining speakers are used as the development and training set, respectively.

MSP-IMPROV is another collection of emotional audiovisual recordings performed by actors. However in this dataset the authors strove to increase naturalness by using target sentences within improvised conversational scenarios which provides a slightly improved setup for evaluation over IEMOCAP. Additional segments of natural interaction of actors during breaks were also included to make this dataset slightly more diverse. We evaluate on the emotion categories *angry* (789), *happy* (2603), *sad* (882), *neutral* (3340), for a total of 7714 utterances. The dataset consists of 6 sessions between male and female pairs which leads to 12-fold cross validation.

CMU-MOSEI contains 23 453 single-speaker video segments from YouTube which have been manually transcribed and annotated for sentiment and emotion. Emotions are annotated on a [0, 3] Likert scale and include multiclass categories such

TABLE I
EMOTION RECOGNITION RESULTS ON IEMOCAP AND MSP-IMPROV

Model	UAR (%)	
	IEMOCAP	MSP-IMPROV
Chance	14.3	25
MDRE [6]	53.6	-
MHA [40]	**55.5**	-
CNN + 40 MFBs [41]	-	52.6
CNN-LSTM-DNN [42]	-	52.4
ELMo + NN	49.2 $\pm$ 0.2	49.5 $\pm$ 0.3
M-ELMo + NN	51.7 $\pm$ 0.3	**53.2 $\pm$0.2**

TABLE II
EMOTION RECOGNITION RESULTS ON CMU-MOSEI

Model	WA(%)						
	Ang	Dis	Fear	Hap	Sad	Sur	Average
(Single modality)							
(A) Graph-MFN [36]	56.4	60.9	62.7	61.5	62.0	54.3	59.6
(L) Graph-MFN [36]	56.6	64.0	58.8	54.0	54.0	54.3	57.0
(L) ELMo+NN	64.4	73.6	61.8	65.4	60.1	62.5	64.6$\pm$0.3
(A + L + V)							
Graph-MFN [36]	62.6	69.1	62.0	66.3	60.4	53.7	62.3
CIM-Att-STL [21]	64.5	72.2	51.5	61.6	65.4	53.0	61.3
(A + L)							
CIM-Att-STL [21]	-	-	-	-	-	-	59.6
M-ELMo+NN	66.0	73.8	63.2	67.0	63.0	64.2	**66.2$\pm$0.2**

Modalities: acoustic (A), lexical (L), visual (V). Note that chance is 50%.

as *happiness*, *sadness*, *anger*, *fear*, *disgust*, and *surprise*. We binarize these annotations to arrive at class labels by predicting the presence of emotions, *i.e*. any emotion with a rating greater than one. Since this dataset includes natural interactions from YouTube there is a greater degree of variance in recording conditions, expression, and speaker differences compared to previous datasets. Therefore we expect this dataset to be a better indicator of effectiveness in emotion recognition in the wild.

E. Emotion Recognition Using DNN

As our goal is to evaluate the efficacy of the multimodal sentence embeddings we used a simple feedforward neural network for emotion recognition. The multimodal sentence embeddings were used as inputs to a neural network (NN) to predict emotions. The tunable parameters described in Section III-C are also included in this network. The network is trained using the training split of each dataset and the validation split is used in choosing hyper-parameters of the network. As a baseline we also compared to a neural network model using sentence embeddings from a text-only ELMo model. Since the biLM is not updated in this stage only the feedforward NN is considered in-domain to emotion recognition.

F. Evaluation Methods

Different metrics were used to evaluate each dataset to be comparable with other work. To evaluate IEMOCAP and MSP-IMPROV the unweighted average recall (UAR) was calculated for each fold and the average value over all folds is reported. As CMU-MOSEI is a multi-class task we evaluated this dataset using weighted accuracy (WA) [36] on each emotion. The metric can also be viewed as the macro-average recall value for each emotion label. We also averaged the resulting weighted accuracy metric across all emotions to obtain an average value.

Due to the lack of work on MSP-IMPROV and CMU-MOSEI which apply SER specifically to text and audio we compared with prior work using different modalities. We also compared with studies that match our train and test splits in IEMOCAP and MSP-IMPROV. For CMU-MOSEI we compare with models that additionally consider the visual modality.

V. RESULTS & DISCUSSION

The results of emotion recognition for the different datasets are shown in Table I (IEMOCAP and MSP-IMPROV) and Table II (CMU-MOSEI). We report the mean and standard deviation for each model over 10 runs. The use of mutliomdal sentence embeddings in a simple model did not improve over prior work on the IEMOCAP dataset. However on MSP-IMPROV the M-ELMo embeddings showed improvement over prior work at UAR 53.2%. On CMU-MOSEI our feedforward neural network using multimodal embeddings improves over prior work in terms of average WA over all emotions at 66.2%.

In all cases we observe that the M-ELMo embeddings show improvement over standard ELMo embeddings in emotion recognition. The improvement of M-ELMo over ELMo embeddings is significant with p-value < 0.05 under McNemar's test. Surprisingly, a neural network using ELMo embeddings led to higher performance than other advanced models using multiple modalities on CMU-MOSEI. This result might demonstrate the effectiveness of unsupervised contextualized embeddings over other methods such as GloVe [43], which were used in [21] and [36].

While the multimodal embeddings did not show improvement on IEMOCAP, we observe a gradual improvement as the evaluation moved to MSP-IMPROV, which was carefully designed to promote naturalness, and CMU-MOSEI, which is real-world data collected from YouTube videos. This might indicate benefits of large-scale pretraining and the suitability of the multimodal embeddings for in the wild SER with more diverse conditions.

VI. CONCLUSION

In this paper we proposed a method for learning and extracting multimodal embeddings from pretrained language models. Our model used convolutional layers to calculate acoustic embeddings from audio features which were then combined with text embeddings in a biLM using a sigmoid gating function. This model was trained in an audio-based language modeling task using audio and text from YouTube videos. We then showed the effectiveness of sentence embeddings extracted from this multimodal biLM in the task of speaker emotion recognition. The results are promising especially given that our downstream model using a simple neural network outperforms state of the art architectures on MSP-IMPROV and CMU-MOSEI. This demonstrates the benefits of using unsupervised pretraining on multimodal models to obtain representations that are effective in capturing inter- and intra-modal dynamics in spoken natural language. Future work includes exploration of different multimodal fusion methods and models as well as unsupervised pretraining on additional tasks at a greater scale.

REFERENCES

[1] E. Rodero, "Intonation and emotion: Influence of pitch levels and contour type on creating emotions," *J. Voice*, vol. 25, no. 1, pp. 25–34, 2011.

[2] F. Nihei, Y. I. Nakano, and Y. Takase, "Fusing verbal and nonverbal information for extractive meeting summarization," in *Proc. ACM, Group Interact. Front. Technol.*, 2018, pp. 1–9.

[3] S. Palaskar, J. Libovický, S. Gella, and F. Metze, "Multimodal abstractive summarization for How2 videos," in *ACL*, 2019, pp. 6587–6596.

[4] L. Liao, Y. Ma, X. He, R. Hong, and Tat-seng Chua, "Knowledge-aware multimodal dialogue systems," in *Proc. 26th ACM Int. Conf. Multimedia*, 2018, pp. 801–809.

[5] C. Hori *et al.*, "End-to-end audio visual scene-aware dialog using multimodal attention-based video features," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 2352–2356.

[6] S. Yoon, S. Byun, and K. Jung, "Multimodal speech emotion recognition using audio and text," in *Proc. IEEE Spoken Lang. Technol. Workshop*, 2018, pp. 112–118.

[7] P. P. Liang, Z. A. A. Liu, B. Zadeh, and L.-P. Morency, "Multimodal language analysis with recurrent multistage fusion," in *Proc. Empirical Methods Natural Lang. Process.*, 2018, pp. 150–161.

[8] C. Busso *et al.*, "Analysis of emotion recognition using facial expressions, speech and multimodal information," in *Proc. Int. Conf. Multimodal Interfaces*, Oct. 2004, pp. 205–211.

[9] S. Narayanan and P. G Georgiou, "Behavioral signal processing: Deriving human behavioral informatics from speech and language," *Proc. IEEE*, vol. 101, pp. 1203–1233, 2013.

[10] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Inf. Fusion*, vol. 37, pp. 98–125, 2017.

[11] P. G. S. Kumar, S.-Y. Panayiotis, T. Georgiou, and S. Narayanan, "Behavior gated language models," 2019, *arXiv:1909.00107*.

[12] J. Pittermann, A. Pittermann, and W. Minker, "Emotion recognition and adaptation in spoken dialogue systems," *Int. J. Speech Technol.*, vol. 13, no. 1, pp. 49–60, 2010.

[13] D. Bertero, F. B. Siddique, C.-S. Y. Wu, R. Wan, H. Y. Chan, and P. Fung, "Real-time speech emotion and sentiment recognition for interactive dialogue systems," in *Proc. Empirical Methods Natural Lang. Process.*, 2016, pp. 1042–1047.

[14] A. Singhal, P. Kumar, R. Saini, P. P. Roy, D. P. Dogra, and B.-G. Kim, "Summarization of videos by analyzing affective state of the user through crowdsource," *Cogn. Syst. Res.*, vol. 52, pp. 917–930, 2018.

[15] P. Tzirakis *et al.*, "End-to-end multimodal emotion recognition using deep neural networks," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 8, pp. 1301–1309, Dec. 2017.

[16] D. Hazarika, S.Gorantla, S. Poria, and R. Zimmermann, "Self-attentive feature-level fusion for multimodal emotion detection," in *Proc. IEEE Conf. Multimedia Inform. Process. and Retrieval*, 2018, pp. 196–201.

[17] S.-Y. H. Tseng, B. L.Baucom, and P. Georgiou, "'Honey, I learned to talk': Multimodal fusion for behavior analysis," in *Proc. 20th ACM Int. Conf. Multimodal Interact.*, 2018, pp. 239-243.

[18] N. Blanchard, D. Moreira, A. Bharati, and W. J Scheirer, "Getting the subtext without the text: Scalable multimodal sentiment classification from visual and acoustic modalities," in *Proc. Grand Challenge Workshop Hum. Multimodal Lang. (Challenge-HML)*, pp. 1–10.

[19] C. W. Lee, Y. Kyu, J. S. Jeong, and W. Y. Choi, "Convolutional attention networks for multimodal emotion recognition from speech and text data," *ACL*, 2018.

[20] A. Zadeh, P. P. Liang, N. Mazumder, S. Poria, E. Cambria, and L.-P. Morency, "Memory fusion network for multi-view sequential learning," in *Proc. AAAI Conf. Artif. Intell.*, 2018.

[21] Md. S. Akhtar, D. Chauhan, D. Ghosal, S. Poria, A. Ekbal, and P. Bhattacharyya, "Multi-task learning for multi-modal emotion recognition and sentiment analysis," in *Proc. Conf. North Amer. Chap. Assoc. Comput. Linguistics: Hum. Lang. Techno.*, Volume 1 (Long and Short Papers), pp. 370–379, 2019.

[22] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013.

[23] E. P. Matthew *et al.*, "Deep contextualized word representations," in *Proc. NAACL-HLT*, 2018, pp. 2227–2237.

[24] J. Devlin, M-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171–4186.

[25] W.-Ni. Hsu and J. Glass, "Disentangling by partitioning: A representation learning framework for multimodal sensory data," 2018, *arXiv:1805.11264*.

[26] Y.-H. H. Tsai, P. P. Liang, A. Zadeh, L.-P. Morency, and R. Salakhutdinov, "Learning factorized multimodal representations," in *Proc. Int. Conf. Learn. Representations*, 2019.

[27] W. Rahman, Md. K. Hasan, A. Zadeh, L.-P. Morency, and M. E. Hoque, ": M-BERT Injecting multimodal information in the BERT structure," 2019, *arXiv:1908.05787*.

[28] C. Gong, D. He, X. Tan, T. Qin, L. Wang, and T.-Y. Liu, "FRAGE: Frequency-agnostic word representation," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2018.

[29] R. Józefowicz, O. Vinyals, M. Schuster, N. Shazeer, and Y. Wu, "Exploring the limits of language modeling," 2016, *arXiv:1602.02410*.

[30] Y. Kim, Y. Jernite, D. Sontag, and A. M Rush, "Character-aware neural language models," in *Proc. 13th AAAI Conf. Artif. Intell.*, 2016.

[31] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 933–941.

[32] S. Arora, Y. Liang, and T. Ma, "A simple but tough-to-beat baseline for sentence embeddings," in *Proc. 5th Int. Conf. Learn. Representations*, 2017, pp. 24–26.

[33] T. N. Sainath and C. Parada, "Convolutional neural networks for small-footprint keyword spotting," in *Proc. 16th Annu. Conf. Int. Speech Commun. Assoc.*, 2015, pp. 1478–1482.

[34] C. Chelba *et al.*, "One billion word benchmark for measuring progress in statistical language modeling," in *Proc. Interspeech*, 2014, pp. 2635–2639.

[35] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "COVAREP – A collaborative voice analysis repository for speech technologies," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2014, pp. 960–964.

[36] A. 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 *Proc. Assoc. Comput. Linguistics*, 2018, pp. 2236–2246.

[37] C. Busso *et al.*, "IEMOCAP: Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, no. 4, pp. 335–359, 2008.

[38] C. Busso, S.Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. M. Provost, "MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception," *IEEE Trans. Affective Comput.*, vol. 8, no. 1, pp. 67–80, May 2016.

[39] S. Yoon, S. Dey, H. Lee, and K. Jung, "Attentive modality hopping mechanism for speech emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 3362–3366.

[40] S. Yoon, S. Byun, S. Dey, and K. Jung, "Speech emotion recognition using multi-hop attention mechanism," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 2822–2826.

[41] Z. Aldeneh and E. M. Provost, "Using regional saliency for speech emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2017, pp. 2741–2745.

[42] S. Latif, R. Rana, S. Khalifa, R. Jurdak, and J. Epps, "Direct modelling of speech emotion from raw speech," *in Proc. Interspeech*, 2019, pp. 3920–3924.

[43] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for word representation," in *Proc. Empirical Methods Natural Lang. Process.*, 2014, pp. 1532–1543.