INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Deep Hierarchical Fusion with application in Sentiment Analysis

Efthymios Georgiou[1,2], *Charilaos Papaioannou*[1], *Alexandros Potamianos*[1,2]

[1]School of ECE, National Technical University of Athens, Athens, Greece
[2]Behavioral Signal Technologies, Los Angeles, CA, USA

efthymis.g.georgiou@gmail.com, cpapaioan@mail.ntua.gr, potam@central.ntua.gr

Abstract

Recognizing the emotional tone in spoken language is a challenging research problem that requires modeling not only the acoustic and textual modalities separately but also their cross-interactions. In this work, we introduce a hierarchical fusion scheme for sentiment analysis of spoken sentences. Two bidirectional Long-Short-Term-Memory networks (BiLSTM), followed by multiple fully connected layers, are trained in order to extract feature representations for each of the textual and audio modalities. The representations of the unimodal encoders are both fused at each layer and propagated forward, thus achieving fusion at the word, sentence and high/sentiment levels. The proposed approach of deep hierarchical fusion achieves state-of-the-art results for sentiment analysis tasks. Through an ablation study, we show that the proposed fusion method achieves greater performance gains over the unimodal baseline compared to other fusion approaches in the literature.

Index Terms: deep hierarchical fusion, fused representations, multimodal fusion, sentiment analysis

1. Introduction

Human communication is a complex process involving multiple modalities. For example in vocal communication, information about the emotional state and mood of a speaker is evident in (but not limited to) both the audio and text modalities, i.e., what is being said and how things are being said. By the combination of multiple modalities, one can acquire richer representations compared to single modality ones. As expected, approaches in the field of human-machine interaction (HMI) focus on exploiting these modes as much as possible. For a realistic HMI system, attempts should gravitate towards encapsulating both semantic and affective information of a message. To solve this problem, multimodal systems have emerged and it has been shown that they outperform the respective single modal systems [1].

Numerous multimodal fusion strategies have been proposed in the literature and can be separated in two broad categories, the *early* or *feature-level fusion*, which combines the features extracted from various modalities to a unified feature vector [2] and the *late* or *decision-level fusion*, in which classifiers are separately trained in different modalities and their results are fused as a decision vector to produce the final result [3]. Those two approaches can be considered as special cases of the general *hybrid-fusion* [4], which aims to exploit the advantages of both methods [5].

An alternative approach is to learn hierarchical representations between different modalities. Those approaches are inspired by results in neuroscience suggesting that the human cortical networks are hierarchically organized [6]. A recent work [7] has introduced a new approach stated as *dense fusion*, which combines representations in different shared layers and aims to capture the correlations in different levels. The shared layers are also connected between them, providing an efficient way to learn the dependence among hierarchical correlations, taking into account not only the current level, but also the lower ones. This approach not only takes advantage of early and late fusion but also learns multiple hierarchical features, exploiting the notion of re-use [8].

Multimodal Machine Learning is an emerging research field with a large number of major studies being proposed in the last few years [9], [10]. Specifically, for the integration of lexical and acoustic features, early works used Support Vector Machines (SVMs) [11]. In [12], features for each modality were separately extracted before feeding them to the classifier.

Deep learning architectures were also introduced in works for multimodal emotion recognition, due to their ability to obtain higher level multimodal features. Researchers in [13] used bidirectional Long-Short-term-Memory [14] networks (BiLSTM's) to capture long-term dependencies in sequential (video) data. The combination of Convolutional Neural Networks (CNNs) with LSTMs for extracting high quality textual, visual and audio features was also proposed in [15]. In [16], the authors introduce a method for building a highly discriminative multimodal feature space by taking the outter product between all three modalities. In [17], a Generative Adversarial Network (GAN) is utilized, to generate features for missing modalities and project them into a common learned multimodal space.

Despite the extensive research that has been carried out on audio-visual modalities [18], there is limited work on audio-textual sentiment analysis. Moreover, fusion methodologies oftentimes fuse independent modalities at abstract levels, ignoring time-dependent interactions between them, e.g. the simultaneous co-occurrence of a negative sentence with a delayed sigh after the end of the sentence. In [19], a word level alignment between all modalities is proposed. Following the aforementioned idea, authors in [20] use a hierarchical attention architecture. Specifically, they pretrain recurrent networks in order to perform single modal sentiment classification. A fine tuning attention mechanism is then applied and its output vector is given to a CNN to perform the final decision.

In this paper we introduce a deep hierarchical fusion method which differs from the aforementioned approaches on the forward propagation of fused representations and the reintegration of them with the unimodal ones in higher representation levels. Specifically, the proposed fusion mechanism for audio and text modalities operates on multiple time-scales (word, sentence) and representation levels of the input. The method is evaluated on sentiment analysis for the MOSI [21] database producing state-of-the art results. The key contributions are: 1) we propose a simple architecture that propagates both the unimodal and fused multimodal representations through our neural network better capturing dependencies between modalities as shown by an ablation study, 2) the multimodal representa-

http://dx.doi.org/10.21437/Interspeech.2019-3243

tion is both fused at each neural layer and propagated forward thus achieving hierarchical deep fusion. The proposed method is general and can be extended to feed-forward architectures of arbitrary depth.

2. Proposed model

As shown in Figure 1, the proposed architecture consists of three parts 1) a text encoder 2) an audio encoder and 3) a Deep Hierarchical Fusion (DHF) network. The two independent modal encoders supply the DHF network with features at each neural layer shown as vertical arrows in Figure 1. The DHF network fuses the information in multiple interconnected levels and finally feeds its output to a classifier that performs sentiment analysis.

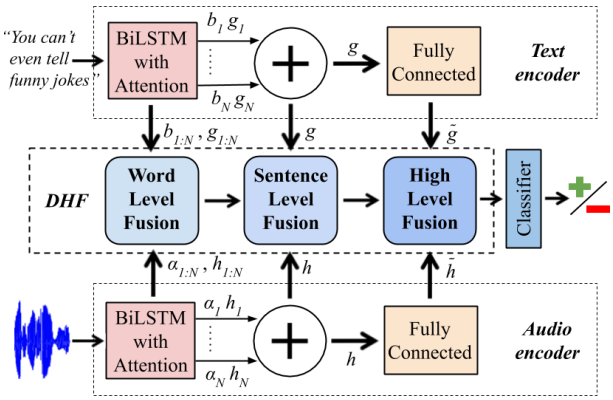

Figure 1: *Overall Architecture*

There are two directions of the flow of the information in the architecture. The first one, illustrated by the vertical arrows, has already been described and depicts the different level representations which are supplied to the DHF. The second one, denoted by the horizontal arrows simulates the forward propagation of the information through the deep network. For the specific task of performing sentiment analysis on spoken sentences, the fusion of textual and acoustic information is performed in three stages. The word-level accepts as inputs two independent modality representations from the encoders. The derived fused representation is then fed-forward to the sentence level which exploits not only the prior fused information, but also re-uses audio and text features, introducing multiple learning paths to the overall architecture. Our DHF network ends up with the high level fusion representation that resides in a (more abstract) multimodal representation space.

2.1. Text encoder

To extract text representations, we use bidirectional LSTM layers [14], which process an input sequentially and are able to capture time-dependencies of language representations. Bidirectional stands for processing an input both forward and backwards. The hidden state g_i of the BiLSTM, at each timestep can be viewed as:

$$g_i = \overrightarrow{g_i}||\overleftarrow{g_i},\ i = 1, \cdots, N \tag{1}$$

where N is the sequence length, $||$ denotes concatenation and $\overrightarrow{g_i}, \overleftarrow{g_i} \in \mathbb{R}^D$ are the forward and backward hidden state representations for the i-th word in the sequence.

Since elements of the input sequence do not contribute equally to the expression of the sentiment in a message, we use an *attention mechanism* that aggregates all hidden states g_i, using their relative importance b_i by putting emphasis on the impactful components of the sequence [22]. This structure is described as follows:

$$e_i = \tanh\left(W_g g_i + b_g\right),\ e_i \in [-1, 1] \tag{2}$$

$$b_i = \frac{\exp\left(e_i\right)}{\sum_{i=1}^{N} \exp\left(e_t\right)},\ \sum_{i=1}^{N} b_i = 1 \tag{3}$$

$$g = \sum_{i=1}^{N} b_i g_i,\ g \in \mathbb{R}^{2D} \tag{4}$$

where the attention weights W_g, b_g adapt during training. Formally, the attention mechanism feeds every hidden state g_i to a nonlinear network that assigns an *energy value* e_i to every element (2). These values are then normalized via (3), to form a probability distribution and a weight b_i is attached to each hidden representation. We compute the representation g of the whole message as the sum (4) of the weighted representations. Since the sequential information is modeled, a fully connected network is applied to perform the classification task. The high-level representation $\widetilde{g} \in \mathbb{R}^{2D}$ extracted by the fully connected layers can be described as:

$$\widetilde{g} = W_t g + b_t \tag{5}$$

where W_t and b_t are the trainable parameters. After the training procedure we strip the output layer off and we use the text subnetwork as the text encoder, as it can be seen in Figure 1. This encoder provides the DHF network with three different high-level representations, namely word-level features $b_{1:N}, g_{1:N}$, sentence-level representations g and high-level features $\widetilde{g}$.

2.2. Audio encoder

A similar approach is followed regarding the acoustic module, since speech features are aligned in word-level and then averaged, resulting in an audio representation for each word. We use a BiLSTM (6):

$$h_i = \overrightarrow{h_i}||\overleftarrow{h_i},\ i = 1, \cdots, N \tag{6}$$

where $h_i \in \mathbb{R}^H$ describes the hidden unit of the i-th timestep. An attention mechanism (2), (3), (7) is also applied:

$$h = \sum_{i=1}^{N} a_i h_i,\ h \in \mathbb{R}^{2H} \tag{7}$$

with the respective attention layer parameters denoted as W_h and b_h. Similarly to the text encoder 2.1, a high-level audio representation $\widetilde{h} \in \mathbb{R}^{2H}$ is learned, via a fully connected network with trainable weight parameters W_a, b_a. This representation is, in turn, given to an output softmax layer which performs the classification. After the learning process, the softmax layer of the speech classifier is no longer considered as part of the network. The remaining sub-modules form the audio encoder of Figure 1 and the word-level $a_{1:N}, h_{1:N}$, sentence-level h and high-representation-level $\widetilde{h}$ features are fed to the DHF.

2.3. DHF

As shown in Figure 1, the DHF network is made up of three hierarchical levels, which are described in the following subsections.

2.3.1. Word-level fusion module

The word-level is the first fusion stage and aims to capture the time-dependent cross-modal correlations. This subnetwork accepts as inputs the word-level features $a_{1:N}, h_{1:N}$ and $b_{1:N}, g_{1:N}$ from audio and text encoder respectively. At every i-th timestep, we apply the following fusion-rule:

$$c_i = a_i h_i || b_i g_i || h_i \odot g_i, \tag{8}$$

where $\odot$ denotes the Hadamard product and $c_i \in \mathbb{R}^{2(2H+D)}$ is the fused time-step representation. These representations form a sequence of length N and are passed to a BiLSTM network with an attention mechanism (2), (3), (9), which outputs the word-level fused representations:

$$f_W = \sum_{i=1}^{N} k_i f_i, \ f \in \mathbb{R}^{2W} \tag{9}$$

where k_i is the fused attention weight at i-th timestep and f_i is the concatenation of hidden states $\overrightarrow{f_i}, \overleftarrow{f_i}$, which belong to a W-dimensional space. We consider W_f and b_f as the respective attention trainable parameters.

2.3.2. Sentence-level fusion module

This is the second level in the fusion hierarchy and as stated by its name, it fuses sentence-level representations. This module accepts as inputs three information flows 1) sentence-level representation g from the text encoder, 2) sentence-level representation h from the audio encoder and 3) the previous-level fused representation f_W. The architecture of the network consists of three fully connected layers. Instead of directly fusing g with h, we apply two fully connected networks which learn some intermediate representations which are then fused with f_W through a third network and produce a new fused representation $f_U \in \mathbb{R}^{2W}$.

2.3.3. High-level fusion module

The last fusion hierarchy level combines the high-level representations of the textual and acoustic modalities, $\widetilde{g}$ and $\widetilde{h}$, with the sentence-level fused representation f_U. This high-dimensional representation is passed through a Deep Neural Network (DNN), which outputs the sentiment level representation $f_S \in \mathbb{R}^M$. The goal of this module is to project this concatenated representation to a common multimodal space.

2.4. Output Layer

After the multimodal information is propagated through the DHF network, we get a high-level representation f_S for every spoken sentence. The role of the linear output layer is to transform these representation to a sentiment prediction. Consequently, this module varies according to the task. For binary classification, we use a single sigmoid function with binary cross entropy loss, whereas a softmax function with a cross entropy loss is applied in the multi-class case.

3. Experimental Methodology

Our experiments were carried out in the *CMU-MOSI* [21] database, a collection of online videos in which a speaker is expressing an opinion towards a movie. Every video consists of multiple clips, where each clip contains a single opinion which is expressed in one or more spoken sentences. MOSI database contains 2199 opinion segments with a unique continuous sentiment label in the interval $[-3, +3]$. We make use of binary, five-scale and seven-scale labels.

3.1. Data Preprocessing

We preprocess our data with *CMU-Multimodal SDK* (mmsdk) [23] tool, which provides us with an easy way for downloading, preprocessing, aligning and extracting acoustic and textual features. For the text input we use *GloVe* embeddings [24]. Specifically, each spoken sentence is represented as a sequence of 300-dimensional vectors. As for the acoustic input, useful features such as MFCCs, pitch tracking and voiced/unvoiced segmenting [16] are used. All acoustic features (72-dimensional vectors) are provided by mmsdk-tool, which uses *COVAREP* [25] framework. Word-alignment is also performed with mmsdk tool through *P2FA* [26] to get the exact time-stamp for every word. The alignment is completed by obtaining the average acoustic vector over every spoken word.

3.2. Baseline Models

We briefly describe the baseline models which our proposed approach is compared to.
C-MKL [15]: uses a CNN structure to capture high-level features and feeds them to a multiple kernel learning classifier.
TFN [16]: uses Kronecker products to capture unimodal, bimodal and trimodal feature interactions. Authors use the same feature set with the one described in subsection 3.1.
FAF [20]: uses hierarchical attention with bidirectional gated recurrent units at word level and a fine tuning attention mechanism at each extracted representation. The extracted feature vector is passed to a CNN which performs the final decision.
MFM [17]: is a GAN, which defines a joint distribution over multimodal data. It takes into account both the generative and the discriminative aspect and aims to generate missing modality values, while projecting them into a common learned space. The feature set in this study is the same with the one we describe in 3.1.

3.3. Experimental Setup

The hidden state hyperparameters H, D, W are chosen as $128, 32, 256$, respectively. A 0.25 dropout rate is picked for all attention layers. Furthermore, fully connected layers in both encoders use Rectified Linear Units (ReLU) and dropout with 0.5 value is applied to the audio encoder. The DHF hyperparameter M is chosen as 64 and all its fully connected layers use ReLU activation functions and a 0.15 dropout probability. Moreover, a gradient clipping value of 5 is applied, as a safety measure against exploding gradients [27]. Our architecture's trainable parameters are optimized using Adam [28] with $1e-3$ learning rate and $1e-5$ as weight decay regularization value. For all models, the same $80-20$ training-testing split is used and we further separate 20% of the training dataset for validation. A 5-fold cross validation is used. All models are implemented using PyTorch [29] framework.

4. Results

As shown in Table 1, the proposed method consistently outperforms other well-known approaches. Specifically in binary classification task, which is the most well-studied, the proposed architecture outperforms by a small 0.5% margin all other models. As for the five and seven class task we outperform other

approaches by 5.87% and 2.14% respectively, which imply the efficacy of the DHF model. Missing values indicate non reported performance measure in the corresponding papers.

Table 1: *CMU-MOSI Sentiment Analysis performance.*

Task	**Binary**		**5 class**	**7 class**
	Acc(%)	F1	Acc(%)	Acc(%)
CMK-L	73.6	75.2	-	-
TFN	75.2	76.0	39.6	-
FAF	76.4	76.8	-	-
MFM	76.4	76.3	-	35.0
DHF	**76.9**	**76.9**	**45.47**	**37.14**

Table 2 illustrates a comparison between the text, the audio and the fusion classifier within the proposed model. Every column describes a unique approach. The most interesting part of our experiments is that the proposed method achieves larger performance gains, $\Delta Fusion$, than the other proposed approaches, as it can be seen in Table 2. Even though the unimodal classifiers for the binary task are not as accurate as in other approaches (FAF, TFN), the DHF boosts the performance enough to outperform them in the multimodal classification. Specifically, the results indicate that our method improves the performance by 3.1%, whereas the state-of-the-art approach FAF shows a relative improvement of 1.4%.

Table 2: *CMU-MOSI Delta Fusion in Binary task.*

Model	**FAF**	**TFN**	**DHF**
	Acc(%)	Acc(%)	Acc(%)
Text	75.0	74.8	73.8
Audio	60.2	65.1	63.3
Fusion	76.4	75.2	76.9
$\Delta Fusion$	↑1.4	↑0.4	↑**3.1**

Table 3 shows the results of an ablation study regarding the contribution of different DHF modules. Three experiments are carried out and in each one, a level of hierarchy is being subtracted. Specifically the first row corresponds to a DHF architecture without the High-Level Fusion module (see Figure 1). The Sentence-Level representation is fed to a softmax classifier in this case. The next two rows describe the DHF without the Sentence-Level and Word-Level Fusion modules respectively. We notice that higher hierarchy levels are more important for the model performance. This demonstrates that the impact of the earlier levels of hierarchy is being decreased as new representations are extracted in the following levels, denoting that the model deepens its learning on feature representations.

Finally, we tested the robustness of the proposed model, by adding Gaussian noise upon the input data. The first two columns of table 4 detail the noise deviation T_{std}, A_{std} on the text and audio data respectively. The next three columns describe each classifier's accuracy. We notice that a 4.7% performance decay in the text classifier, yields a 4.1% decay in the fusion method. This is expected while the input noise affects both the text and multimodal classifier. Additionally, the third row shows a 4% and 8.3% reduction in text and audio performance

Table 3: *CMU-MOSI Ablation study in Binary task.*

Model	**Accuracy(%)**	**F1**
DHF *No High-Level*	75.0	74.8
DHF *No Sent-Level*	75.5	75.4
DHF *No Word-Level*	75.7	75.6
DHF	76.9	76.9

respectively, while fusion model only shows a 6.5% decay. It can be observed that for reasonable amounts of input data noise, the DHF outperforms the textual classifier. Finally, the last row makes it clear that a significant noise injection to the system ends in ruining our approach. The potential reason that may influence our model is that DHF learns how to combine features from both modalities and thus noise affects the propagation of fusion information significantly. This result indicates that DHF learns cross-modal representations, a result that lies in the core of DHF's initial purpose.

Table 4: *Noise Effect on Binary Task*

Noise		**Accuracies (%)**		
T_{std}	A_{std}	Text	Audio	DHF
0.0	**0.0**	**73.81**	**63.33**	**76.91**
0.3	0.0	69.05	63.33	72.86
0.3	0.01	69.32	55	70.48
0.25	0.05	68.09	52.62	65.95

5. Conclusions and Future Work

We proposed a deep hierarchical architecture for modality fusion and applied it to the problem of sentiment analysis from the audio and text modalities. Our implementation uses three layers roughly corresponding to the word, sentence and high/sentiment levels. The proposed method achieves state-of-the-art sentiment analysis scores on the MOSI database. Interestingly, although the baseline audio-only and text-only sentiment analysis systems are not state-of-the-art, the multimodal sentiment analyzer that employs hierarchical fusion gives enough of a performance boost to beat the competition.

In future work, we aim at developing a fusion scheme where the single modal encoders are not frozen but instead allow for weight adaptation, potentially by using different optimizers for each modality. Furthermore, we will experiment with different neural architectures for the single-modality encoders, such as pretrained CNNs that are able to extract high-level audio features. We will also investigate how to better synchronize the different layers of the single-modality encoders, as well as experiment with deeper architectures.

6. Acknowledgements

We would like to thank our anonymous reviewers for their feedback. Special thanks to Nikos Athanasiou and Georgios Paraskevopoulos for their suggestions. This work has been supported by computational time granted from the Greek Research & Technology Network in the National HPC facility-ARIS.

7. References

[1] S. K. D'mello and J. Kory, "A review and meta-analysis of multimodal affect detection systems," *ACM Computing Surveys (CSUR)*, vol. 47, no. 3, p. 43, 2015.

[2] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, "Multimodal fusion for multimedia analysis: a survey," *Multimedia systems*, vol. 16, no. 6, pp. 345–379, 2010.

[3] C. G. Snoek, M. Worring, and A. W. Smeulders, "Early versus late fusion in semantic video analysis," in *Proceedings of the 13th annual ACM international conference on Multimedia*. ACM, 2005, pp. 399–402.

[4] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2018.

[5] M. Wöllmer, M. Al-Hames, F. Eyben, B. Schuller, and G. Rigoll, "A multidimensional dynamic time warping algorithm for efficient multimodal fusion of asynchronous data streams," *Neurocomputing*, vol. 73, no. 1-3, pp. 366–380, 2009.

[6] P. Taylor, J. Hobbs, J. Burroni, and H. Siegelmann, "The global landscape of cognition: hierarchical aggregation as an organizational principle of human cortical networks and functions," *Scientific reports*, vol. 5, p. 18112, 2015.

[7] D. Hu, C. Wang, F. Nie, and X. Li, "Dense multimodal fusion for hierarchically joint representation," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3941–3945.

[8] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[9] N. Srivastava and R. R. Salakhutdinov, "Multimodal learning with deep boltzmann machines," in *Advances in neural information processing systems*, 2012, pp. 2222–2230.

[10] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proceedings of the 28th international conference on machine learning (ICML-11)*, 2011, pp. 689–696.

[11] D. Seppi, A. Batliner, B. Schuller, S. Steidl, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, N. Amir, and V. Aharonson, "Patterns, prototypes, performance: classifying emotional user states," in *Proc. 9th Interspeech 2008 incorp. 12th Australasian Int. Conf. on Speech Science and Technology SST 2008, Brisbane, Australia*, 2008, pp. 601–604.

[12] V. Rozgić, S. Ananthakrishnan, S. Saleem, R. Kumar, and R. Prasad, "Ensemble of svm trees for multimodal emotion recognition," in *Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*. IEEE, 2012, pp. 1–4.

[13] F. Eyben, M. Wöllmer, A. Graves, B. Schuller, E. Douglas-Cowie, and R. Cowie, "On-line emotion recognition in a 3-d activation-valence-time continuum using acoustic and linguistic cues," *Journal on Multimodal User Interfaces*, vol. 3, no. 1-2, pp. 7–19, 2010.

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

[15] S. Poria, E. Cambria, and A. Gelbukh, "Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis," in *Proceedings of the 2015 conference on empirical methods in natural language processing*, 2015, pp. 2539–2544.

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

[17] Y.-H. H. Tsai, P. P. Liang, A. Zadeh, L.-P. Morency, and R. Salakhutdinov, "Learning factorized multimodal representations," *arXiv preprint arXiv:1806.06176*, 2018.

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

[19] M. Chen, S. Wang, P. P. Liang, T. Baltrušaitis, A. Zadeh, and L.-P. Morency, "Multimodal sentiment analysis with word-level fusion and reinforcement learning," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. ACM, 2017, pp. 163–171.

[20] Y. Gu, K. Yang, S. Fu, S. Chen, X. Li, and I. Marsic, "Multimodal affective analysis using hierarchical attention strategy with word-level alignment." in *Proceedings of the conference. Association for Computational Linguistics. Meeting*, vol. 2018, 2018, p. 2225.

[21] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos," *arXiv preprint arXiv:1606.06259*, 2016.

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

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

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

[25] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "Covarepa collaborative voice analysis repository for speech technologies," in *2014 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 2014, pp. 960–964.

[26] J. Yuan and M. Liberman, "Speaker identification on the scotus corpus," *Journal of the Acoustical Society of America*, vol. 123, no. 5, p. 3878, 2008.

[27] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *International conference on machine learning*, 2013, pp. 1310–1318.

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

[29] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.