INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

M^3: MultiModal Masking applied to sentiment analysis

Efthymios Georgiou[1,2], *Georgios Paraskevopoulos*[1,2], *Alexandros Potamianos*[1,3,4]

[1]School of ECE, National Technical University of Athens, Athens, Greece
[2]Institute for Language and Speech Processing, Athena Research Center, Athens, Greece
[3]Signal Analysis and Interpretation Laboratory (SAIL), USC, Los Angeles, CA, USA
[4] Behavioral Signal Technologies, Los Angeles, CA, USA

efthygeo@mail.ntua.gr, geopar@central.ntua.gr, potam@central.ntua.gr

Abstract

A common issue when training multimodal architectures is that not all modalities contribute equally to the model's prediction and the network tends to over-rely on the strongest modality. In this work, we present M^3, a training procedure based on modality masking for deep multimodal architectures. During network training, we randomly select one modality and mask its features, forcing the model to make its prediction in the absence of this modality. This structured regularization allows the network to better exploit complementary information in input modalities. We implement M^3 as a generic layer that can be integrated with any multimodal architecture. Our experiments show that M^3 outperforms other masking schemes and improves performance for our strong baseline. We evaluate M^3 for multimodal sentiment analysis on CMU-MOSEI, achieving results comparable to the state-of-the-art.

Index Terms: multimodal, masking, sentiment analysis, dropout, CMU-MOSEI

1. Introduction

Multimodal machine learning approaches process information from different sources and modalities to solve a given task [1]. Human perception itself, also relies on combination of multimodal cues, in order to improve the latent representations of real world entities and concepts. Furthermore, advances in both neuroscience and psychology indicate that multi-sensory inputs are crucial for cognitive functions [2] since infancy [3]. Thus, modeling and understanding multimodal interactions is a potential avenue for building better machine learning models.

Multimodal approaches have shown success in various fields such as Visual Question Answering (VQA) [4], Automatic Speech Recognition (ASR) [5, 6], Sentiment Analysis [7, 8], Machine Translation [9] and Sound Localization [10]. Despite the promising results, it is often observed that multimodal architectures trained from scratch outperform their unimodal counterparts only by small margins [11, 12, 13]. This result indicates that some inputs are able to dominate the learning process in multimodal networks and therefore make the network's prediction over-rely to this modality. One possible cause for this behavior may be attributed to the different generalization rates of each modality [13], as well as the inherent bias in some modalities towards specific tasks, e.g. there are high sentiment cues in the linguistic structure [12]. Current multimodal research is mainly focused on building fusion schemes and learning better representations for the task in hand, but addressing these issues is of vital importance for multimodal learning.

In the emotion recognition and sentiment analysis fields a diverse set of multimodal fusion approaches has been proposed. Zadeh et al. [7] utilize outer products between late representations and construct a high dimensional fusion space from which the prediction is made. Hierarchical fusion schemes have been proposed based on attention mechanisms [14], as well as deep fusion architectures [15]. Moreover, jointly learning multimodal representations has been explored in various ways such as, Deep Canonical Correlation Analysis [16] and cycle-consistency cross modal mappings [17]. Bagher et al. [18] investigate architectures with neural memory-like modules. Different architectural settings such as Transformers [19, 8], Graph Neural Networks [20] and Capsules [21, 22] have also been explored for multimodal fusion, as well as the use of large pre-trained architectures by jointly fine-tuning them for multiple modalities [12] or embedding multimodal information into them [23]. Despite the significant improvements and the challenges tackled in the aforementioned works, none of them directly battles over-reliance on specific modalities.

We propose M^3, which aims at directly tackling over-reliance on specific modalities. Specifically, M^3 takes as input various modalities, e.g. text, audio, visual. It then randomly masks one of them or leaves the total multimodal representation unaffected. By randomly masking the representation of a particular modality, we force the network to solve the task in its absence during training. M^3 is applied at every time step in the multimodal sequence, acting as a form of regularization. Intuitively, our work can be seen as a form of structured dropout [24] but differs from vanilla dropout in the sense that we mask multimodal representations rather than units within an architecture. We integrate M^3 with a strong Long Short Term Memory (LSTM)-based baseline [25] and demonstrate the effectiveness of the proposed approach for sentiment analysis on CMU-MOSEI [18].

Our contributions are: 1) a generic light-weight layer which can be embedded in multimodal architectures, 2) a comparison of M^3 with other masking schemes which demonstrates the need to mask the whole modality in order to achieve better performance and 3) comparable to the state-of-the-art performance in sentiment analysis task for the CMU-MOSEI dataset. Our code is available as open-source[1].

2. Proposed Approach

In this section we formally describe the proposed approach and we introduce our baseline model and the M^3 layer.

Notation: A Bernoulli distribution is denoted as $\mathcal{B}e(p)$ where p is the probability of sampling the zero value. The same holds for the categorical distribution $\mathcal{C}at(p_1, p_2, p_3)$ where the result is a binary triplet and also $p_1 + p_2 + p_3 = 1$. The modalities

[1]https://github.com/efthymisgeo/multimodal-masking

http://dx.doi.org/10.21437/Interspeech.2021-1739

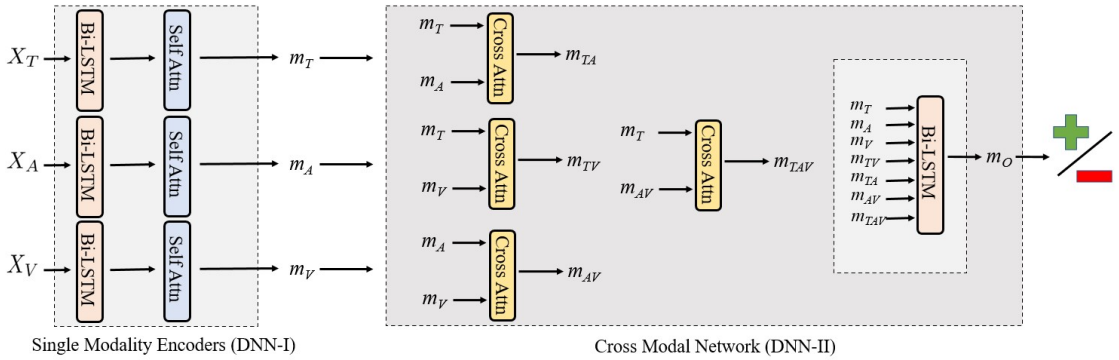

Figure 1: *Baseline Architecture. The architecture consists of two larger parts, namely the Single Modality Encoders (DNN-I) and the Cross Modal Network (DNN-II). The building blocks of the architecture are Bi-LSTMs, Scaled Dot Product Attention (Self Attn) and a symmetric Cross Attention Mechanism (Cross Attn). The M^3 layer is inserted between DNN-I and DNN-II and takes as input the representations m_T, m_A, m_V from DNN-I and feeds its output to DNN-II.*

are denoted as T, A, V for text, audio and video respectively and are compactly represented as an index $l \in \{T, A, V\}$. The representations for each modality, as well as their combinations are denoted as m_l. For example the textual representation is m_T, while the audio-visual is m_{AV}. We also hypothesize that each representation is a sequence of N d-dimensional vectors.

2.1. Baseline Method

Network Modules: The overall architecture consists of three building blocks (Fig. 1). The first is a Bidirectional LSTM (Bi-LSTM), the second is the scaled dot product attention mechanism (Self Attention) [19] and the third is a cross modal attention module (Cross Attention) inspired by the one proposed in Lu et al. [26].

Baseline Architecture: The network consists of two main subnetworks as illustated in Fig. 1, namely the Single Modality Encoders (DNN-I) and the Cross Modal Network (DNN-II). The first network is responsible for processing the unimodal inputs while the cross modal part extracts fused representations which are used for the final prediction.

The *Single Modality Encoder* takes the input features X_A, X_T, X_V for each modality and encodes them using three Bi-LSTMs, one for each modality. The hidden states are reweighted using Self Attention blocks, which in turn produce the unimodal encodings m_T, m_A, m_V.

The *Cross Modal Network* is responsible for capturing cross-modal interactions. It consists of four Cross Attention modules and a Bi-LSTM which performs the final prediction. The Cross Modal Network takes as input the single modality representations and uses symmetric cross attention to capture interactions between T, A, V as illustrated in Fig 1. Specifically it calculates m_{TA}, m_{TV}, m_{AV} which capture all combinations of m_T, m_A, m_V, as well as m_{TAV} which fuses audio-visual with textual information.

We describe in detail the symmetric cross-modal attention mechanism we use. Consider $m_k, m_l \in \mathbb{R}^{B \times N \times d}$, where $k \neq l$ are modality indicators. Given the input modality representations, we can construct keys $K_l = W_l^K m_l$, queries $Q_k = W_k^Q m_k$ and values $V_l = W_l^V m_l$, which are learnable due to their corresponding projection matrices $W_{\{k,l\}}^{\{K,Q,V\}}$. We can now define the cross-modal attention layer as:

$$a_{kl} = s\left(\frac{K_l^T Q_k}{\sqrt{d}}\right) V_l + m_k, \tag{1}$$

where $s(\cdot)$ denotes the softmax function and B, N, d are the batch size, sequence length and hidden size respectively. The symmetric attention is defined by summing two cross-modal attentions as:

$$m_{kl} = a_{kl} + a_{lk} \tag{2}$$

These crossmodal representations are concatenated ($\|$), along with the unimodal representations m_A, m_T, m_V to produce the fused feature vector $\tilde{m} \in \mathbb{R}^{B \times N \times 7d}$ in Eq. (3).

$$\tilde{m} = m_T \parallel m_A \parallel m_V \parallel m_{TAV} \parallel m_{AV} \parallel m_{TV} \parallel m_{TA} \tag{3}$$

We feed $\tilde{m}$ into a Bi-LSTM and we then take the last hidden state from both directions and concatenate them into a single representation which is denoted as m_O. This final representation is fed to a linear layer which performs regression.

2.2. MultiModal Masking ($\mathbf{M^3}$)

The proposed method is illustrated in Fig. 2. It takes as input representations from three modalities, e.g. text, audio and visual, which are extracted from a neural architecture. Formally described M^3 takes as input m_T, m_A, m_V which lie in $\mathbb{R}^{B \times N \times d}$, where B is the batch size, N the sequence length and d the dimension of the processed multimodal input sequence[2]. It is then applied at every sample of the batch and at every time step i in the multimodal sequence of length N.

In particular it decides whether to mask one of the given modalities or leave them unaffected, based on a *masking probability*, denoted as p_M. We introduce this hyperparameter because it allows us to control the rate of masking. Formally this is described as sampling from a Bernoulli distribution, i.e. $\mu_M \sim \mathcal{B}e(p_M)$, where μ_M is the probability which decides whether to mask or not. In the case of not masking ($\mu_M = 1$), M^3 simply leaves all the modality representations unaffected and feeds them to the next module.

[2]One may use different sets and number of aligned input modalities with varying feature dimensions. We describe the case of aligned text, audio and visual modalities with common dimension d.

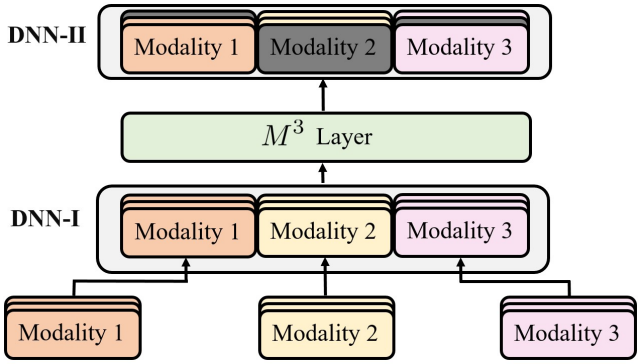

Figure 2: *Each modality is illustrated with a different color. The modality features are fed to a neural architecture (DNN-I) which creates high level unimodal embeddings. The output sequence of the first DNN is in turn fed to the M^3 Layer. The M^3 layer processes these sequential representations by randomly choosing whether to mask or not at each time step. In the case of masking a modality is chosen based on the modality probabilities. The masked output is fed to the rest pipeline (DNN-II) which is responsible for making the final prediction. M^3 can be used as a drop-in module in any multimodal architecture and can be extended to an arbitrary number of modalities.*

In the scenario where modalities with stronger relative performance are present, one may wish to increase the masking rate of the dominant modalities. This can lead to increased contribution of the weaker modalities and stronger fusion results. To this end we introduce $p_L = \{\pi_T, \pi_A, \pi_V\}$, which is set of hyperparameters that control the individual modalities masking rate and are called *modal probabilities*. A categorical distribution is then sampled as of $\mu_\Pi \sim Cat(\pi_T, \pi_A, \pi_V)$ to decide which of the three modalities will be masked. The total sampling process of the M^3 layer can be described as is represented as:

$$\mu_{M^3} = \mu_M + (1 - \mu_M)\mu_\Pi = (\mu_T, \mu_A, \mu_V) \quad (4)$$

and is a binary value for each of the involved modalities. The output of M^3 for a given sample i of the multimodal sequence can be expressed as

$$M^3(m_T^i, m_A^i, m_V^i; p_M, p_L) = [\mu_T m_T^i, \mu_A m_A^i, \mu_V m_V^i] \quad (5)$$

where m_l^i denotes the representation for each modality $l \in \{T, A, V\}$ and lies in $\mathbb{R}^d$ and the index i depicts the time step of the sequence.

$\mathbf{M^3}$ Architecture: The augmented baseline with M^3 layer is the following. The Single Modality Encoders (DNN-I) (Fig. 1) output is fed to the M^3 layer and its output is given as input to the Cross Modal Network (DNN-II).

3. Experimental setup

We evaluate M^3 on CMU-MOSEI for sentiment analysis. The dataset contains $23,454$ YouTube video clips of movie reviews annotated at the video level by humans for sentiment and emotion scores. Sentiment scores range from -3 (strongly negative) to 3 (strongly positive) and emotion annotations. Audio sequences are represented with 74 COVAREP features [27] and visual sequences using Facet features[3]. Video transcriptions are segmented in words and represented using GloVe [28]. All sequences are aligned to the transcribed words. Standard train, validation and test splits are provided. For all our experiments we use Bi-LSTMs with hidden size 100. All projection sizes for the attention modules are set to 100. We use dropout with drop rate 0.2. We use Adam [29] with learning rate $5e-4$ and halve the learning rate if the validation loss does not decrease for 2 epochs. All models are trained using batch size $B = 32$. We use early stopping on the validation loss (patience 10 epochs). Models are trained for regression on sentiment values using Mean Absolute Error (MAE) loss. We use standard evaluation metrics: 7-class (i.e. classification in $\mathbb{Z} \cap [-3, 3]$), binary accuracy and F1-score (negative in $[-3, 0)$, positive in $(0, 3]$), MAE and Pearson correlation coefficient between model and human predictions.

[3]iMotions: https://imotions.com/guides/facial-expression-analysis/

4. Experiments

In this section we compare M^3 with the baseline and the state-of-the-art fusion methods for sentiment classification for CMU-MOSEI. We evaluate different masking schemes and analyze the performance of M^3 for different sets of hyperparameters.

4.1. Masking Schemes Ablation

In Table 1 we compare M^3 with different masking schemes. The baseline model is presented in Section 2.1 and its performance is shown in the first row. Random Mask does not take into account structural modality information. Similar to dropout we sample random feature indices from a Bernoulli distribution across modalities and set those features to zero. For Soft M^3 we choose a random modality, as described in Eq. (4) and sample a mask only for the chosen modality. In other words, we mask only a part of the chosen modality. M^3 takes a more radical approach, i.e. we randomly choose a modality for each sample and time step in the batch and set all features to zero for the chosen modality. In Table 1 we can see that M^3 outperforms the baseline, as well as all other masking schemes for all metrics by a significant margin. Interestingly Soft M^3 does not always outperform Random Mask, so limiting dropout-like masking within a specific modality is not enough. One has to go all the way during training and force the networks predictions to be made in the absence of masked modalities to observe a performance gain. For all the examined schemes the masking rates are tuned independently for fair comparison. Also note that we did not tune the modal probabilities for the M^3 approaches, i.e. $\pi_l = 1/3$.

Table 1: *CMU-MOSEI Sentiment Analysis performance. We randomly drop a modality for each method. The modality masking probability is tuned separately for each method for fair comparison. The modalities pose equal masking probability. Averages over 5 runs.*

	Acc^2	$F1$	Acc^7	Corr	MAE
Baseline	82.07 ± 0.6	82.23 ± 0.5	51.49 ± 0.6	0.695 ± 0.005	0.590 ± 0.004
Random Mask	81.97 ± 0.7	82.29 ± 0.5	51.66 ± 0.8	0.700 ± 0.009	0.588 ± 0.008
Soft M^3	82.31 ± 0.6	82.59 ± 0.5	51.13 ± 0.6	0.691 ± 0.007	0.592 ± 0.006
M^3	$\mathbf{82.46} \pm 0.5$	$\mathbf{82.64} \pm 0.5$	$\mathbf{51.78} \pm 0.6$	$\mathbf{0.702} \pm 0.004$	$\mathbf{0.586} \pm 0.004$

4.2. Comparison with the state-of-the-art

Table 2: *Comparison with the state-of-the-art in CMU-MOSEI for Sentiment Analysis task.*

	Acc^2	$F1$	Acc^7	Corr	MAE
RAVEN [30]	79.1	79.5	50.0	0.662	0.614
MCTN [17]	79.8	80.6	49.6	0.670	0.609
M.Rout [22]	81.7	81.8	51.6	-	-
MulT [8]	**82.5**	82.3	51.8	**0.703**	**0.580**
M^3 (ours)	**82.5**	**82.92**	**51.89**	0.700	0.586

In Table 2 we compare M^3 with state-of-the-art approaches for MOSEI sentiment classification. M^3 achieves state-of-the-art performance for binary F1 score with 0.62% absolute improvement and seven class accuracy with 0.09% absolute improvement, while achieving comparable performance for binary accuracy and correlation. Note that our baseline model consists of LTSMs with attention modules instead of Transformers as in [8]. Finally, observe that just the introduction of M^3 in our baseline model yields a performance boost that leads to comparable to the state-of-the-art results.

4.3. Modality Masking Hyperparameter Analysis

We also perform an analysis of the performance of M^3 with respect to the hyperparameters p_M and π_T. Specifically p_M is the probability that controls if masking is performed, or if modality features are left unaltered. If masking is performed, we mask text features with probability π_T, while audio and visual features with equal probabilities $\frac{1-\pi_T}{2}$. We choose to investigate the text masking probability, because text is the dominant modality in CMU-MOSEI for sentiment classification. Intuitively we want to mask the dominant modality more often, so that we force the network to use information encoded in the weaker audio and visual modalities. In Table 3 we see the performance of M^3 with respect to these hyperparameters. We experiment with $\pi_T = 0.33$ and $\pi_T = 0.6$ for various values of p_M. All results are averaged over 5 runs (stds are in the same range as in Table 1). We observe that generally $p_M \in \{0.2, 0.25\}$ yields the best performance. As expected the larger value for π_T yields better models in general that achieve higher accuracy scores, though by a small margin. This is an indication that randomly forcing the absence of the dominant modality during training can lead to better solutions.

Table 3: *M^3 hyperparameter analysis. The parameter p_M denotes the probability of masking one out of the involved modalitites, while π_T is the probability of dropping the text modality.*

π_T	p_M	Acc^2	$F1$	Acc^7	Corr	MAE
0.33	0.10	82.15	82.45	51.80	0.699	0.588
0.33	0.20	82.46	82.64	51.78	**0.702**	**0.586**
0.33	0.25	82.21	82.60	51.10	0.695	0.589
0.33	0.40	81.84	82.19	51.50	0.695	0.589
0.60	0.10	82.23	82.52	51.88	0.699	0.588
0.60	0.20	82.31	82.53	51.84	0.699	0.587
0.60	0.25	**82.50**	**82.92**	**51.89**	0.700	**0.586**
0.60	0.40	82.12	82.25	51.25	0.699	0.590

5. Conclusions

In this work we present M^3, a light-weight masking layer for multimodal training. M^3 can be integrated into any multimodal architecture to improve performance for multimodal tasks and is also extendable to an arbitrary number of modalities. The core motivation of M^3 is to randomly force the network predictions to be made without use of the dominant modalities during training, in order to enhance the contribution of the weaker modalities. Therefore M^3 randomly selects one of the modalities for a percentage of the training samples and sets its features to zero. We experiment with different masking strategies and M^3 yields the best results. In addition, M^3 consistently improves performance for our strong LSTM with attention baseline, especially when we select to mask the dominant modality (text) with higher probability. The performance improvement obtained by integrating M^3 into the baseline model leads to comparable with the state-of-the-art results. In the future we plan to integrate M^3 with more architectures (Transformers) and experiment with more diverse multimodal tasks, e.g. Multimodal ASR and VQA, which would involve more than three modalities. Finally we plan to experiment with scheduling the mask probability of each modality during training in order to better control the learning rate of each modality.

6. Acknowledgements

We would like to thank the anonymous reviewers for their feedback. This research has been co-financed by the European Regional Development Fund of the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH – CREATE – INNOVATE (project safety4all with code:T1EDK04248).

7. References

[1] 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, pp. 423–443, 2018.

[2] "Current perspectives and methods in studying neural mechanisms of multisensory interactions," *Neuroscience and Biobehavioral Reviews*, vol. 36, pp. 111 – 133, 2012.

[3] P. A. Neil, C. Chee-Ruiter, C. Scheier, D. J. Lewkowicz, and S. Shimojo, "Development of multisensory spatial integration and perception in humans," *Developmental science*, vol. 9, pp. 454–464, 2006.

[4] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "VQA: visual question answering," in *2015 IEEE International Conference on Computer Vision, ICCV*, 2015, pp. 2425–2433.

[5] G. Paraskevopoulos, S. Parthasarathy, A. Khare, and S. Sundaram, "Multimodal and multiresolution speech recognition with transformers," in *Proc. of the 58th Annual Meeting of the Association for Computational Linguistics, ACL*, 2020, pp. 2381–2387.

[6] T. Afouras, J. S. Chung, and A. Zisserman, "ASR is all you need: Cross-modal distillation for lip reading," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2020, pp. 2143–2147.

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

[8] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proc. of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

[9] O. Caglayan, P. Madhyastha, L. Specia, and L. Barrault, "Probing the need for visual context in multimodal machine translation," in *Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, 2019, pp. 4159–4170.

[10] A. Senocak, T. Oh, J. Kim, M. Yang, and I. S. Kweon, "Learning to localize sound source in visual scenes," in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2018, pp. 4358–4366.

[11] Z. Sun, P. K. Sarma, W. A. Sethares, and Y. Liang, "Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis," in *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, EAAI*, 2020, pp. 8992–8999.

[12] S. Siriwardhana, A. Reis, R. Weerasekera, and S. Nanayakkara, "Jointly Fine-Tuning "BERT-Like" Self Supervised Models to Improve Multimodal Speech Emotion Recognition," in *Proc. Interspeech 2020*, pp. 3755–3759.

[13] W. Wang, D. Tran, and M. Feiszli, "What makes training multimodal classification networks hard?" in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, 2020, pp. 12 692–12 702.

[14] 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 *Proc. of the conference. Association for Computational Linguistics. Meeting*, 2018, p. 2225.

[15] E. Georgiou, C. Papaioannou, and A. Potamianos, "Deep hierarchical fusion with application in sentiment analysis," *Proc. Interspeech 2019*, pp. 1646–1650.

[16] Z. Sun, P. K. Sarma, W. Sethares, and E. P. Bucy, "Multi-modal sentiment analysis using deep canonical correlation analysis," *Proc. Interspeech 2019*, pp. 1323–1327.

[17] H. Pham, P. P. Liang, T. Manzini, L.-P. Morency, and B. Póczos, "Found in translation: Learning robust joint representations by cyclic translations between modalities," in *Proc. of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 6892–6899.

[18] A. Bagher 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. of the 56th Annual Meeting of the Association for Computational Linguistics*, 2018.

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 6000–6010.

[20] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, and A. F. Gelbukh, "Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation," in *Proc. of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP*, 2019, pp. 154–164.

[21] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems*, 2017, pp. 3856–3866.

[22] Y. H. Tsai, M. Ma, M. Yang, R. Salakhutdinov, and L. Morency, "Multimodal routing: Improving local and global interpretability of multimodal language analysis," in *Proc. of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP*, 2020, pp. 1823–1833.

[23] W. Rahman, M. K. Hasan, S. Lee, A. Bagher Zadeh, C. Mao, L.-P. Morency, and E. Hoque, "Integrating multimodal information in large pretrained transformers," in *Proc. of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

[24] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *CoRR*, vol. abs/1207.0580, 2012.

[25] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, pp. 1735–1780, 1997.

[26] J. Lu, D. Batra, D. Parikh, and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *NeurIPS 2019*, 2019, pp. 13–23.

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

[28] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in *Proc. of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543.

[29] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR*, 2015.

[30] Y. Wang, Y. Shen, Z. Liu, P. P. Liang, A. Zadeh, and L.-P. Morency, "Words can shift: Dynamically adjusting word representations using nonverbal behaviors," in *Proc. of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 7216–7223.