M-BERT: Injecting Multimodal Information in the BERT Structure

Wasifur Rahman
Department of Computer Science
University of Rochester, USA
echowdh2@ur.rochester.edu

Md. Kamrul Hasan
Department of Computer Science
University of Rochester, USA
mhasan8@cs.rochester.edu

Amir Zadeh
Language Technologies Institute
CMU, USA
abagherz@cs.cmu.edu

Louis-Philippe Morency
Language Technologies Institute
CMU, USA
morency@cs.cmu.edu

Mohammed (Ehsan) Hoque
Department of Computer Science, Goergen Institute for Data Science
University of Rochester, USA
mehoque@cs.rochester.edu

Abstract

Multimodal language analysis is an emerging research area in natural language processing that models language in a multimodal manner. It aims to understand language from the modalities of text, visual, and acoustic by modeling both intra-modal and cross-modal interactions. BERT (Bidirectional Encoder Representations from Transformers) provides strong contextual language representations after training on large-scale unlabeled corpora. Fine-tuning the vanilla BERT model has shown promising results in building state-of-the-art models for diverse NLP tasks like question answering and language inference. However, fine-tuning BERT in the presence of information from other modalities remains an open research problem. In this paper, we inject multimodal information within the input space of BERT network for modeling multimodal language. The proposed injection method allows BERT to reach a new state of the art of 84.38% binary accuracy on CMU-MOSI dataset (multimodal sentiment analysis) with a gap of 5.98 percent to the previous state of the art and 1.02 percent to the text-only BERT.

1 Introduction

Human communication flows as a seamless integration of text, acoustic, and vision. In ordinary everyday interactions, we integrate all these modalities to convey our intentions and emotions. The nonverbal context accompanying text during these interaction is crucial for understanding the intentions properly. Integrating this context into NLP models is an active area of research and a step towards achieving more intelligent AI systems.

Previous work in multimodal analysis employs pre-trained word embeddings learned from a large corpus to represent the meaning of a word. These embeddings are non-contextual and hence, they do not necessarily reflect the specific context they are being used. Recently, contextual embeddings have shown remarkable performance in various NLP applications (Peters et al., 2018; Devlin et al., 2018). BERT (Bidirectional Encoder Representations from Transformers) generates contextual word representations using a bi-directional embedding approach (Devlin et al., 2018). Hence, these embeddings are more representative of the textual meaning. Fine-tuning BERT to specific tasks has been identified as a crucial step in achieving higher performance on the task. While such a fine-tuning may seem straightforward for textual NLP data, it is not quite straight forward how fine-tuning can account for nonverbal context. Therefore, modifying BERT structure by *injecting* nonverbal behaviors is of possible high impact for computational modeling of multimodal language.

Using the pre-trained BERT implementation (huggingface, 2019) as baseline model, we present **M-BERT** that injects audio-visual information into the pre-trained BERT model and allows for fine-tuning in presence of nonverbal behaviors. Specifically, this is done by gated-shifting the input embedding of the BERT model using word-level representations of nonverbal behaviors (Wang et al., 2018). Our proposed approach sets a new state of the art of 84.38% binary accuracy on CMU-MOSI dataset of multimodal sentiment analysis; a significant leap from previous state of the art of 78.4% and fine-tuned text-only BERT of 83.36%.

The contributions of this paper are therefore:

- We propose an efficient architecture, named M-BERT, which allows for fine-tuning BERT in presence of multimodal information.
- Our proposed method reaches new state of the art on CMU-MOSI dataset(Zadeh et al., 2016)

2 Related Work

Multimodal Language Analyses: Multimodal language analyses is a recent research trend in natural language processing (Zadeh et al., 2018b; Baltrušaitis et al., 2019) that helps us understand language language from the modalities of text, vision and acoustic by modeling both intermodal and cross-modal interaction. These analyses mostly target the task of sentiment analysis (Poria et al., 2018), emotion recognition (Zadeh et al., 2018d; Busso et al., 2008), and personality traits recognition (Park et al., 2014). Remarkable works in this area have been done on novel multimodal neural architectures (Wang et al., 2018; Pham et al., 2019; Hazarika et al., 2018; Poria et al., 2017; Zadeh et al., 2017) and multimodal fusion approaches (Liang et al., 2018; Tsai et al., 2018; Liu et al., 2018; Barezi et al., 2018).

Pre-trained Language Representations : Learning the word representations has been an active research area in NLP community (Mikolov et al., 2013; Pennington et al., 2014). Recently, pre-trained deep language representation model, trained on vast amount of text with unsupervised objective, have achieved state of the art results on several NLP tasks like question answering, sentiment classification, POS tagging and similarity modeling(Peters et al., 2018; Devlin et al., 2018). In particular, BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018) outperforms its predecessor contextual language model like ELMO (Peters et al., 2018) and GPT (Radford et al., 2018).

Multimodal Variation of Language Representations: Words can have different meaning when they appears with different non-verbal intents (Zadeh et al., 2017; Wang et al., 2018). Dynamic shifting of word representations based on their accompanying non-verbal context has shown promising performance on the task of multimodal sentiment and emotion recognition (Wang et al., 2018).

3 Multimodal BERT

Fine-tuning BERT in presence of multimodal information is a fundamental NLP resarch question in modeling multimodal language. In this section, we introduce Multimodal BERT (M-BERT) that injects non-verbal information in BERT structure by shifting the text *input embedding* through a **Multimodal Shifting Gate**. Figure.1 shows the overview of M-BERT model and its component.

We represent an input as three N length sequences $\{L, A, V\}$ – representing language(text), audio and video respectively. We represent language as a sequence of word-piece tokens $L = [L_1, L_2, \dots L_N]$. Similarly, audio is represented as $A = [A_1, A_2, \dots, A_N]$ and video as $V = [V_1, V_2, \dots, V_N]$. Here, A_i and V_i are the average acoustic and visual features corresponding to word-piece tokens L_i (following the alignment technique of (Chen et al., 2017)).

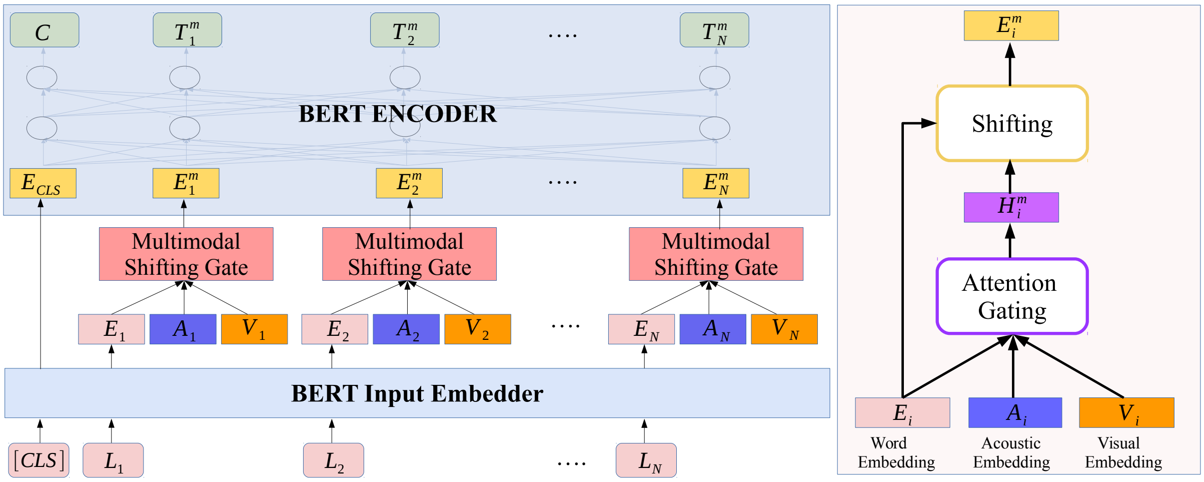
3.1 BERT

We use the variant of BERT used for *Single Sentence Classification Tasks* (Devlin et al., 2018). First, *input embeddings* are generated from a sequence of word-piece tokens by adding token embeddings, segment embeddings and position embeddings (we define this component as **BERT Input Embedder**). Then **BERT Encoder** applies multiple self-attention layers on top of these input embeddings to attend to inputs with respect to each other. A special [CLS] token is appended in front of the input token sequence. So, for a N length input sequence, we get N+1 vectors from the **BERT Encoder** – the first of them representing the class of this input sequence. This class vector is used to predict the label of the input.

3.2 M-BERT

Our proposed M-BERT model (Fig. 1.a) infuses audio-visual information with *input embeddings* using the **Multimodal Shifting Gate** (discussed in Sec 3.3). Subsequently, it inputs the modified embeddings to **BERT Encoder**.

Given the language sequence $L = [L_1, L_2, \dots L_N]$, a [CLS] token is appended to L and inputted to the **BERT Input Embedder** which outputs $E = [E_{CLS}, E_1, E_2, \dots E_N]$. We prepare a sequence of triplets $[(E_i, A_i, V_i) : \forall i \in$



(a) Multimodal BERT

(b) Multimodal Shifting Gate

Figure 1: Overview of Multimodal BERT Network

$[1, N]]$ by pairing E_i with the (A_i, V_i) corresponding to L_i. Each of these triplets are passed through the **Multimodal Shifting Gate** which transforms the ith triplet into E_i^m – a unified multimodal representation of the corresponding word-piece token. We concatenate all of these output to form a sequence $[E_{CLS}, E_1^m, E_2^m, \dots, E_N^m]$ which is used as input to the **BERT Encoder**. The **BERT Encoder** transforms the input into $[C, T_1^m, T_2^m, \dots, T_N^m]$: each element in the sequence attended by every other elements. We pass C – the element corresponding to class label of this sequence – through an affine transformation to produce the final output.

3.3 Multimodal Shifting Gate

Using (Wang et al., 2018) as motivation, we shift the *input embeddings* by a shift vector learned from the acoustic and visual features (Fig 1.b). We deploy an **Attention Gating** mechanism to create a shift vector by controlling the influence of acoustic and visual features with respect to the *input embedding*. Given the ith triplet (E_i, A_i, V_i), we create two different vectors $[E_i; A_i]$ and $[E_i; V_i]$ by concatenating *input embedding* with acoustic and visual information respectively and use them to produce two gating vectors g_i^v and g_i^a:

$$g_i^v = R(W_{gv}[E_i; V_i] + b_v) \tag{1}$$

$$g_i^a = R(W_{ga}[E_i; A_i] + b_a) \tag{2}$$

where W_{gv}, W_{ga} are weight matrices for visual and acoustic modality and b_v and b_a are scalar biases. $R(x)$ is the ReLU activation function.

We then create a non-verbal shift vector H_i^m by fusing together A_i and V_i multiplied by their respective gating vectors:

$$H_i^m = g_i^a \cdot (W_a A_i) + g_i^v \cdot (W_v A_v) + b_H \tag{3}$$

where W_a and W_v are weight matrices for acoustic and visual information respectively and b_H is the bias vector.
Then, we use the following **Shifting** mechanism to add H_i^m with the Input Embedding E_i to create a **Multimodal Input Embedding** E_i^m:

$$E_i^m = E_i + \alpha H_i^m \tag{4}$$

$$\alpha = min(\frac{\|E_i\|_2}{\|H_i^m\|_2}\beta, 1) \tag{5}$$

where β is a hyper-parameter selected through cross-validation. We use the scaling factor α so that the effect of non-verbal shift H_i^m remains within a desirable range.

4 Experiments

4.1 Datasets

We evaluate our model on the CMU-MOSI (Zadeh et al., 2016) dataset of multimodal sentiment analysis. This dataset – containing 2199 video segments taken from 93 Youtube movie review videos – has real-valued sentiment intensity annotations in the [-3,+3] range from human annotators. The train, validation and test folds of the CMU-MOSI contains 1248, 229 and 686 segments respectively (Chen et al., 2017).

4.2 Features

For each modality, the extracted features are as follows:

Language: We use P2FA forced alignment model (Yuan and Liberman, 2008) to align the text and audio on word level. From the forced alignment, we extract the timing annotations on word and sentence level. We also interpolate the acoustic and visual cues on word level (Chen et al., 2017).

Acoustic: COVAREP tool (Degottex et al. 2014) is used to extract following relevant features: low level acoustic features including 12 Mel-frequency cepstral coefficients (MFCCs), pitch tracking and voiced/unvoiced segmenting features, glottal source parameters, peak slope parameters and maxima dispersion quotients.

Visual: We extract visual features like facial landmarks, facial action units, gaze tracking and head pose using the Facet(iMotions, 2017) library.

4.3 Experimental Design

The goal of our experiment is to show that audio-visual information infusion can improve performance metrics on an established multimodal dataset like CMU-MOSI. We use the following variants of BERT to study multimodal sentiment analysis task:

BERT: In this variant, we fine-tune the standard BERT network described in 3.1 using text information only.

M-BERT: In this variant, we inject audio-visual information with text in BERT structure through the **Multimodal Shifting Gate** (Sec 3.2).

We can also determine the effect of non-verbal information injection in BERT structure by comparing the performance of M-BERT with BERT.

4.4 Baseline Models

We compare the performance of M-BERT with the following models on the multimodal sentiment analysis task: **RMFN** (SOTA1)[1] fuses multimodal information in multiple stages by focusing on a subset of signals in each stage (Liang et al., 2018). **MFN** (SOTA2) synchronizes states of three separate LSTMs with a multi-view gated memory (Zadeh et al., 2018a). **MARN** (SOTA3) models view-specific interactions using hybrid LSTM memories and cross-modal interactions using a Multi-Attention Block(MAB) (Zadeh et al., 2018c).

Task Metric	BA↑	F1↑	MAE↓	Corr↑
SOTA3	77.1	77.0	0.968	0.625
SOTA2	77.4	77.3	0.965	0.632
SOTA1	78.4	78.0	0.922	0.681
BERT	83.36	85.53	0.736	0.777
M-BERT	**84.38**	**86.34**	**0.732**	**0.790**
Δ_{SOTA}	↑ **5.98**	↑ **8.34**	↓ **0.19**	↑**0.11**

Table 1: Sentiment prediction results on CMU-MOSI. SOTA1, SOTA2 and SOTA3 refer to the previous best, second best and third best state of the art models respectively. Best results are highlighted in bold and Δ_{SOTA} represents the change in performance of M-BERT model over SOTA1. Our model significantly outperforms the current SOTA across all evaluation metrics.

4.5 Evaluation Metrics

We perform two different evaluation tasks on CMU-MOSI datset: i) Binary Classification, and ii) Regression. We formulate it as a regression problem and report Mean-absolute Error (MAE) and the correlation of model predictions with true labels. Besides, we convert the regression outputs into categorical values to obtain binary classification accuracy (BA) and F1 score. Higher value means better performance for all the metrics except MAE.

5 Results and Discussion

The performances of **M-BERT** and **BERT** are described in Table 1. **M-BERT** model outperforms all the baseline models (described in Sec.4.4) on every evaluation metrics with large margin. It sets new state-of-the-art performance for this task and achieves 84.38% accuracy, a 5.98% increase with respect to the SOTA1 and 1.02% increase with respect to **BERT** (text-only).

Even **BERT** (text-only) model achieves 83.36% accuracy, an increase of 4.96% from the SOTA1 78.4%, using text information only. It achieves higher performance in all evaluation metrics compare to SOTA1; reinforcing the expressiveness and utility of BERT contextual representation.

We can clearly see that **M-BERT** model achieves better performance than **BERT** (text-only) model across all metrics. It is a testament to the usefulness of injecting audio-visual information into the BERT structure and the flexibility of the network in incorporating that information fruitfully.

[1] SOTA=State of The Art

6 Conclusion

In this paper, we designed **M-BERT**, an intuitive extension of the BERT network capable of injecting non-verbal information into the BERT structure for fine-tuning. We validated our model by advancing the state of the art on an established multimodal sentiment analysis dataset. Our model demonstrates the capability of the pre-trained BERT model to incorporate non-verbal information through fine-tuning.

References

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2019. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443.

Elham J Barezi, Peyman Momeni, Pascale Fung, et al. 2018. Modality-based factorization for multimodal fusion. *arXiv preprint arXiv:1811.12624*.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335.

Minghai Chen, Sen Wang, Paul Pu Liang, Tadas Baltrušaitis, Amir Zadeh, and Louis-Philippe Morency. 2017. Multimodal sentiment analysis with word-level fusion and reinforcement learning. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pages 163–171. ACM.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann. 2018. Conversational memory network for emotion recognition in dyadic dialogue videos. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, pages 2122–2132.

huggingface. 2019. pytorch-pretrained-bert. https://github.com/huggingface/pytorch-pretrained-BERT.

iMotions. 2017. Facial expression analysis.

Paul Pu Liang, Ziyin Liu, Amir Zadeh, and Louis-Philippe Morency. 2018. Multimodal language analysis with recurrent multistage fusion. *arXiv preprint arXiv:1808.03920*.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2018. Efficient low-rank multimodal fusion with modality-specific factors. *arXiv preprint arXiv:1806.00064*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Sunghyun Park, Han Suk Shim, Moitreya Chatterjee, Kenji Sagae, and Louis-Philippe Morency. 2014. Computational analysis of persuasiveness in social multimedia: A novel dataset and multimodal prediction approach. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 50–57. ACM.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabas Poczos. 2019. Found in translation: Learning robust joint representations by cyclic translations between modalities. *arXiv preprint arXiv:1812.07809*.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Mazumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Multi-level multiple attentions for contextual multimodal sentiment analysis. In *2017 IEEE International Conference on Data Mining (ICDM)*, pages 1033–1038. IEEE.

Soujanya Poria, Amir Hussain, and Erik Cambria. 2018. *Multimodal Sentiment Analysis*, volume 8. Springer.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/research-covers/languageunsupervised/language understanding paper. pdf*.

Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2018. Learning factorized multimodal representations. *arXiv preprint arXiv:1806.06176*.

Yansen Wang, Ying Shen, Zhun Liu, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2018. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. *arXiv preprint arXiv:1811.09362*.

Jiahong Yuan and Mark Liberman. 2008. Speaker identification on the scotus corpus. *Journal of the Acoustical Society of America*, 123(5):3878.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*.

Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018a. Memory fusion network for multi-view sequential learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Amir Zadeh, Paul Pu Liang, Louis-Philippe Morency, Soujanya Poria, Erik Cambria, and Stefan Scherer. 2018b. Proceedings of grand challenge and workshop on human multimodal language (challenge-hml). In *Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML)*.

Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. 2018c. Multi-attention recurrent network for human communication comprehension. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6):82–88.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018d. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 2236–2246.