Word-Aware Modality Stimulation for Multimodal Fusion

Shuhei Tateishi, NTT Docomo, Inc.

Yasuhito Ohsugi, NTT Docomo, Inc.

Makoto Nakatsuji, NTT Human Informatics Laboratories

Contents

1. Introduction
2. Methodology
3. Evaluation Results
4. Conclusion, and Challenges for the future

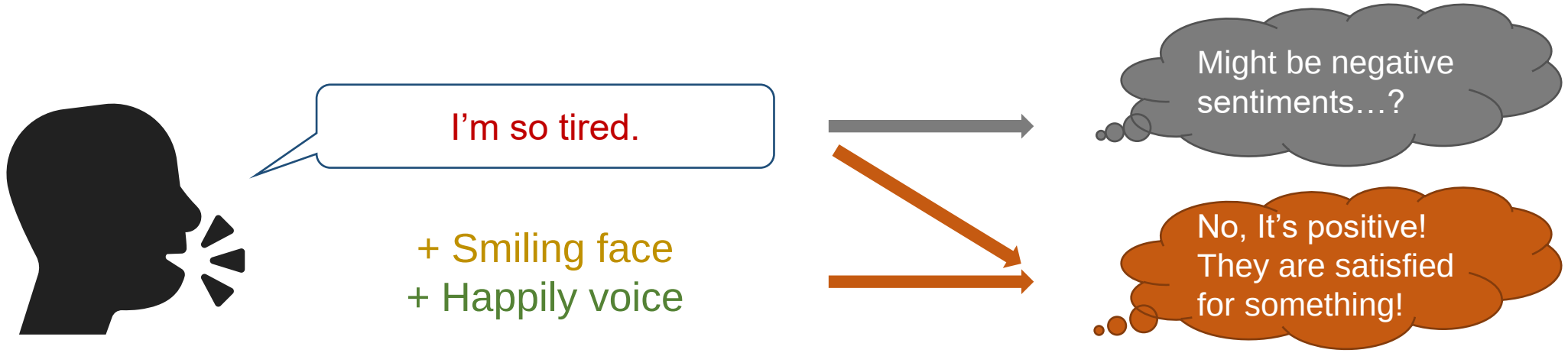
Contents

1. Introduction
2. Methodology
3. Evaluation Results
4. Conclusion, and Challenges for the future

Introduction – Multimodal learning

Multimodal – a multiple data source input is expected to improve the accuracy of machine learning tasks.

The typical task by multimodal is: **“Sentiment Analysis”**



Introduction – Former studies

Thus, several multimodal fusion methods have been proposed:

- Bidirectional LSTMs have been employed to capture long-range dependencies from low-level acoustic descriptors and visual features (Eyben et al. (2010), Wöllmer et al. (2010))
- CNNs have been used to extract both textual and visual features (Poria et al., 2015)
- Tensor multiplication among the language, audio, and visual modality feature vectors (Zadeh et al., 2017)
- Transformer layers to fuse multimodal embedding streams (Tsai et al., 2019)
- Encodes the language modality by using BERT and the speech modality by using COVAREP, then fuse them by source-target attention (Yang et al., 2020)
- The speech modality as a dynamic prefix alongside the textual modality, in contrast to conventional language models like RoBERTa (Arjmand et al., 2021)

… And so on.

Introduction – Problem

However, we found a huge challenge against such former methods in respect of accuracy:

“Mono-modal learning by BERT is so powerful that it outperforms almost all existing multimodal methods”

Introduction – “BERT, the Great”

Previous multimodal studies claimed their method has a superior performance to BERT in sentiment analysis task, but they only conducted a few (e.g., up to five) iterations for BERT learning.

However, we found many (up to 50) iterations for BERT learning makes great improvement on the accuracy score of BERT.

e.g.) Sentiment-analysis, CMU-MOSI, Acc^7 (7-class accuracy), BERT-large:
5 iterations: **41.5 %** → 50 iterations (w/early-stopping): **50.51 %**

→ **50.51 % Acc^7 outperforms results of almost all previous methods**

Introduction – Our method

The basic strategies for our study are:

1. **“Beat the BERT”**: Inventing how to surpass BERT with the multimodal fusion method in the sentiment analysis task.
2. **“Use the BERT”**: Seeking multimodal fusion methods that do not hinder the superior expressive power of BERT.

That is our method:

Word-aware Modality Stimulation Fusion (WA-MSF).

Contents

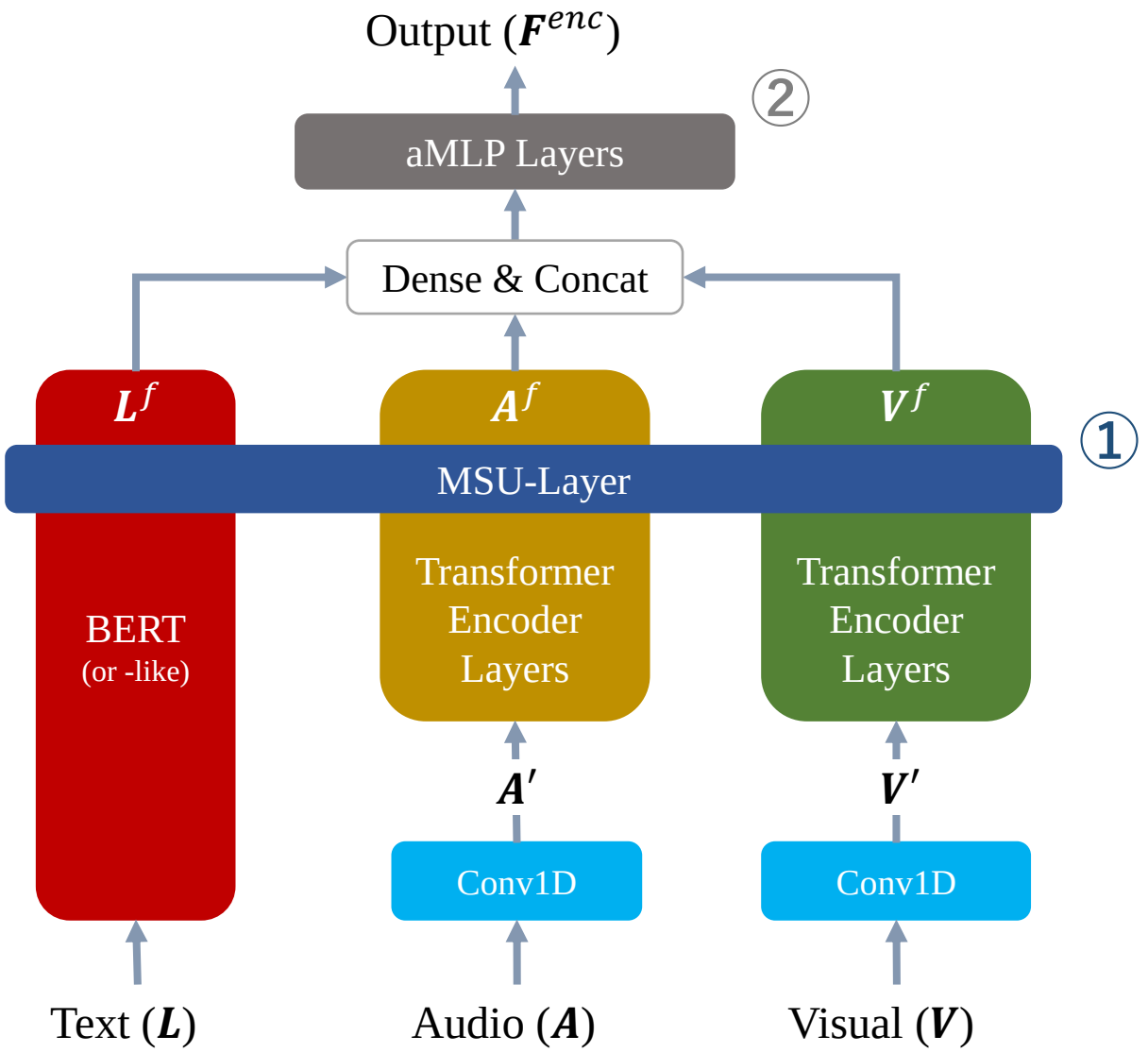
1. Introduction
2. Methodology
3. Evaluation Results
4. Conclusion, and Challenges for the future

Methodology

The overview figure is the right.

Our method contains **two core concepts:**

1. **Modality Stimulation Unit Layer (MSU-Layer)**
2. **aMLP multimodal fusion**



Methodology – MSU-Layer

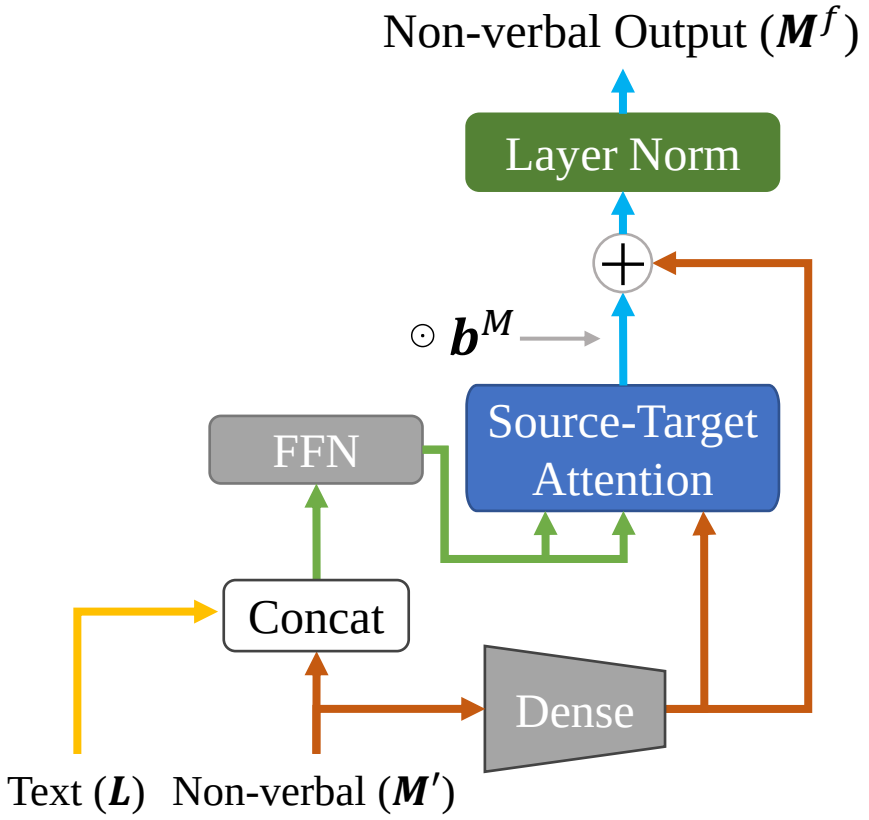
Modality Stimulation Unit Layer (MSU-Layer) is designed for non-verbal (audio, and visual) modalities to infuse linguistic, semantic information from the text modality and import into each non-verbal modality.

MSU-Layer has two kinds of process:

1. For non-verbal modality, and
2. For textual modality.

Methodology – MSU-Layer (nonverbal)

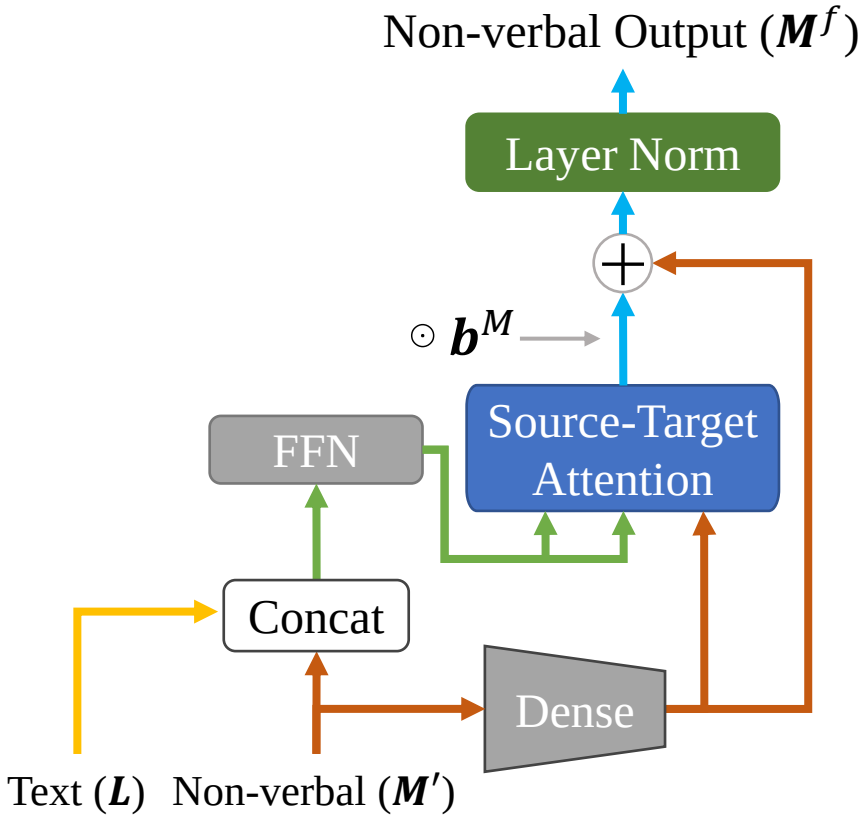
- MSU-layer for non-verbal modalities will be injected immediately after a specific layer of the transformer encoder for each modality.
- Before this transformer encoder, a Conv1D layer will be applied to match the sequence length of each modality with the length of words of the textual modality (see overview figure).



Methodology – MSU-Layer (nonverbal)

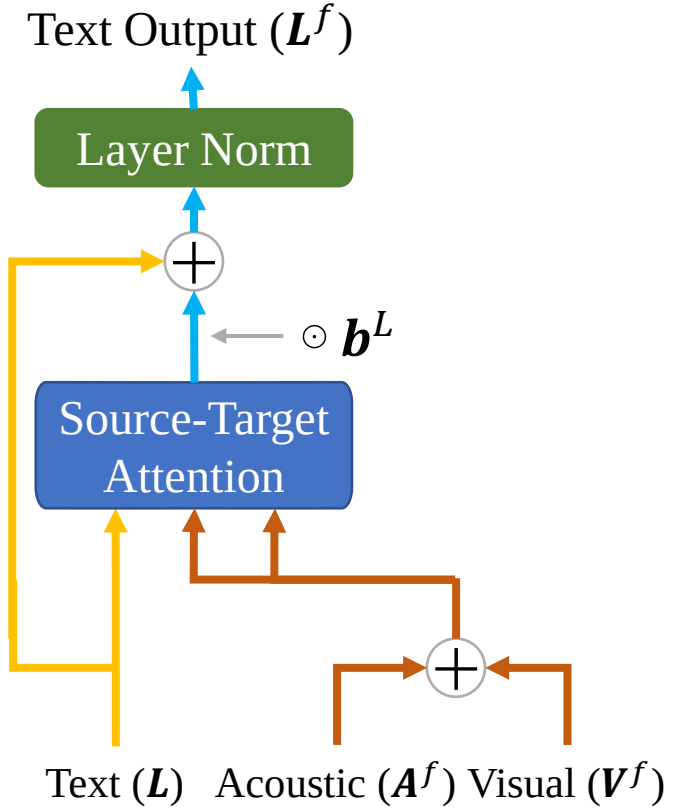
The main aims of this layer are:

1. importing the semantic information of the textual modality to align the attention target
2. Optimize the weights of the Conv1D layer for word alignment by back propagation – enabling consideration of the semantics of the textual modality



Methodology – MSU-Layer (textual)

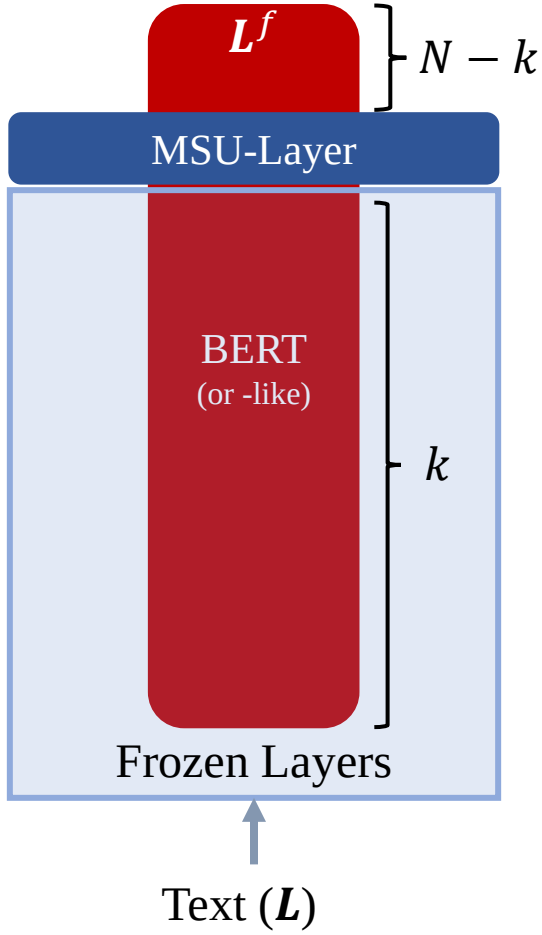
- MSU-layer for the textual modality also will be injected immediately after a specific layer of the BERT (will be detailed later).
  → To enable the textual encoder to also reference information from each nonverbal modality after the MSU-layer.
- However, an effect from nonverbal modalities must be limited for the layers **after the MSU-layer is inserted.**



Methodology – Partial freeze for BERT

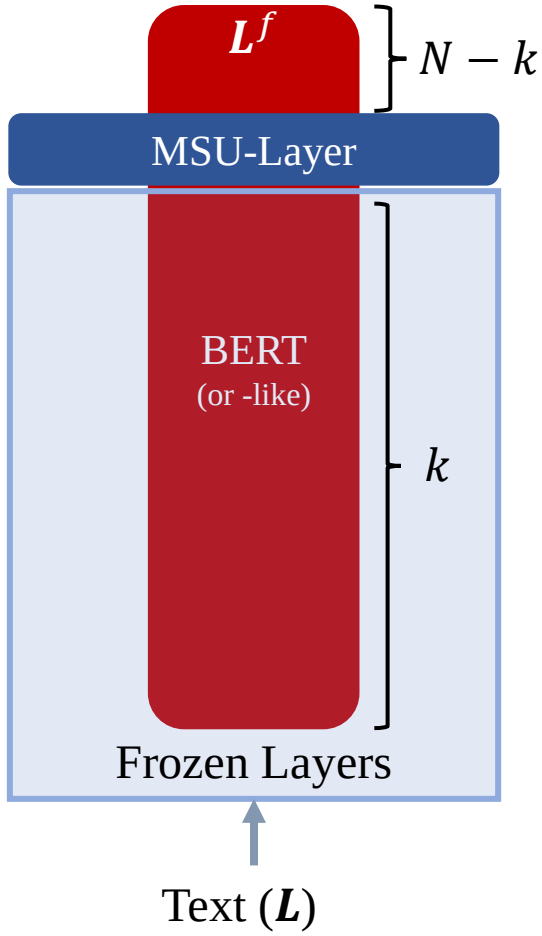
We hypothesized:

- Non-verbal modalities extending its influence to the word understanding part of BERT's layers results in lower accuracy for former multimodal methods compared to using a single BERT.
- By influencing only the layers of BERT responsible for contextual understanding, non-verbal modalities hold the potential for improving accuracy.



Methodology – Partial freeze for BERT

- From past research, it is known that each layer of BERT serves separate roles such as word and contextual comprehension. (Ethayarajh, 2019)
- From several experiments, we identified the appropriate number of layers, denoted as k, to insert the MSU-layer to infuse non-verbal modalities' information.
- For layers before k, only the textual part of the dataset is pre-trained, then its weights are reused and these layers are frozen while the multimodal training.



Methodology – aMLP fusion

MSU-layer is a **pre-fusion** layer, so our method prepares separated fusion mechanism to maximize the benefit of MSU-layer and non-verbal modality.

That is: **aMLP** (gMLP with tiny attention)

Methodology – aMLP fusion

- **gMLP** (Liu et al., 2021) is a new multi-layer encoder model without attention mechanism, It possesses strengths in capturing temporal-spatial features.
- **aMLP** is a variant of gMLP, a tiny attention mechanism is added to gMLP to reinforce the ability of text processing.

→ aMLP has the potential to reconcile the temporal-spatial aspects of non-verbal modalities with textual semantic understanding.

Contents

1. Introduction
2. Methodology
3. Evaluation Results
4. Conclusion, and Challenges for the future

Evaluation Result

We were conducted the evaluation for our method:

1. Evaluated by two datasets: **CMU-MOSI** and **CMU-MOSEI**
2. For the textual modality, "**BERT-large**" (24 layers, 1024 dimension embeddings) model is employed
3. Inserting a MSU-layer after the **20th layer** of BERT
4. Focusing on two aspects as an ablation study:
    1. Combination of modalities (textual, acoustic, visual)
    2. Fusion method (aMLP vs Transformer)

For other parameter configurations, please refer to the paper.

Evaluation Result

- **Bolded** is the best score
- <u>Underlined</u> is the second-best
- "Max" score of our method is the best score from 100 attempts of our method evaluation
- "Avg" score is the mean score from 100 attempts

→ Our method marks best performance for the **regression task** (Acc^7, MAE, Corr)

Table 1: Evaluation Result (CMU-MOSI).
XX_{h}: "higher is better", XX_{l}: "lower is better".

Method	$\mathrm{F1}_{\mathrm{h}}$	$\mathrm{Acc}^2_{\mathrm{h}}$	$\mathrm{Acc}^7_{\mathrm{h}}$	$\mathrm{MAE}_{\mathrm{l}}$	$\mathrm{Corr}_{\mathrm{h}}$
CM-BERT	84.5	84.5	44.9	0.729	0.791
MAG-BERT	82.5	82.37	43.62	0.727	0.781
MAG-XLNet	85.7	85.6	N/A	0.675	0.821
TEASEL	85	**87.5**	47.52	0.64	0.836
CHFN	86.2	86.4	48.6	0.689	0.809
UniMSE	<u>86.42</u>	<u>86.9</u>	48.68	0.691	0.809
BERT-large	86.04	85.98	<u>50.51</u>	<u>0.636</u>	<u>0.838</u>
Ours (Max)	**86.97**	86.86	**51.82**	**0.623**	**0.842**
Ours (Avg)	85.99	85.96	49.99	**0.629**	<u>0.838</u>

Table 2: Evaluation Result (CMU-MOSEI)

Method	$\mathrm{F1}_{\mathrm{h}}$	$\mathrm{Acc}^2_{\mathrm{h}}$	$\mathrm{Acc}^7_{\mathrm{h}}$	$\mathrm{MAE}_{\mathrm{l}}$	$\mathrm{Corr}_{\mathrm{h}}$
MMIM	85.94	85.97	54.24	0.526	0.772
MAG-BERT	84.5	84.7	N/A	N/A	N/A
UniMSE	**87.46**	**87.50**	<u>54.39</u>	<u>0.523</u>	0.773
BERT-large	N/A	N/A	53.38	0.531	<u>0.775</u>
Ours (Max)	<u>86.09</u>	<u>86.26</u>	**54.63**	**0.515**	**0.785**
Ours (Avg)	85.67	85.80	53.71	**0.520**	**0.782**

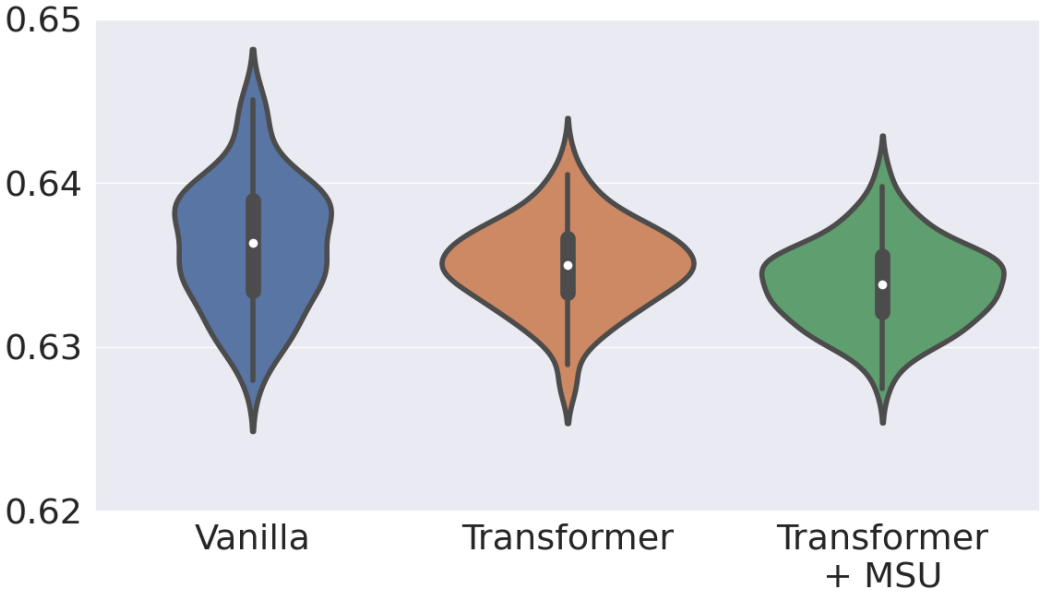
Evaluation Result

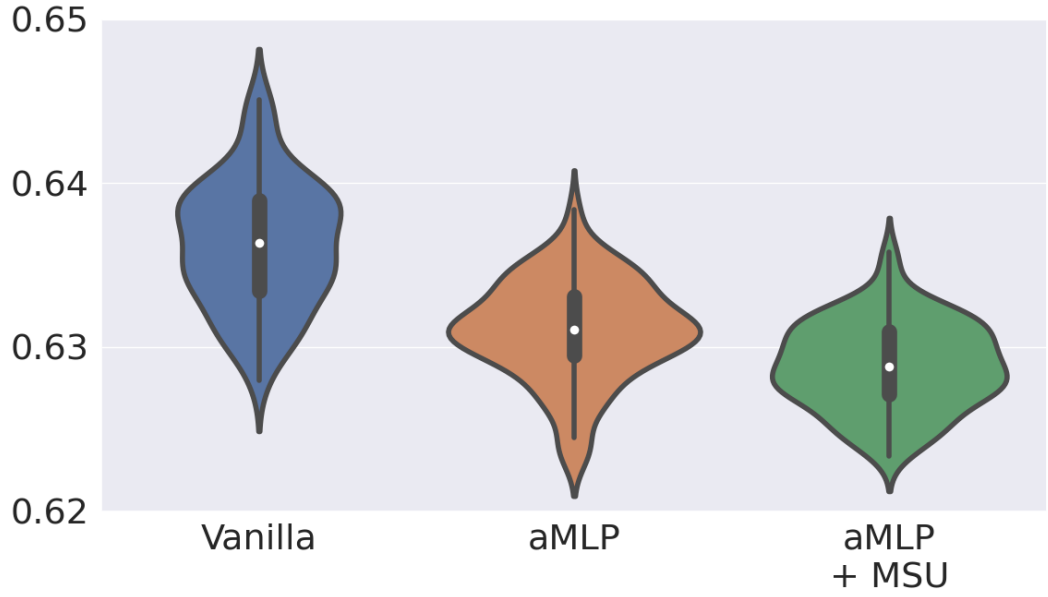
Table 3: Ablation Study Results - mean and standard deviation (CMU-MOSI)

Method	$F1_h$	Acc_h^2	Acc_h^7	MAE_l	$Corr_h$
Modality combination					
Text (BERT-large)	85.70 ± 0.66	85.67 ± 0.64	47.73 ± 1.52	0.6591 ± 0.0149	0.8270 ± 0.0082
Text + Video	**86.05 ± 0.43**	**86.01 ± 0.41**	49.87 ± 0.78	0.6319 ± 0.0027	0.8363 ± 0.0024
Text + Audio	85.87 ± 0.46	85.85 ± 0.43	49.94 ± 0.69	0.6298 ± 0.0027	0.8368 ± 0.0021
Full	85.99 ± 0.47	85.96 ± 0.44	**49.99 ± 0.74**	**0.6288 ± 0.0027**	**0.8376 ± 0.0019**
Fusion method					
Vanilla	85.54 ± 0.45	85.54 ± 0.43	49.65 ± 0.87	0.6362 ± 0.0039	0.8357 ± 0.0022
+ Transformer	85.89 ± 0.43	85.86 ± 0.41	49.64 ± 0.77	0.6349 ± 0.0027	0.8361 ± 0.0019
+ MSU-Lyr	85.85 ± 0.37	85.82 ± 0.39	49.78 ± 0.77	0.6338 ± 0.0026	0.8358 ± 0.0019
+ aMLP	85.95 ± 0.43	85.92 ± 0.40	49.85 ± 0.78	0.6310 ± 0.0030	0.8370 ± 0.0022
+ MSU-Lyr	**85.99 ± 0.47**	**85.96 ± 0.44**	**49.99 ± 0.74**	**0.6288 ± 0.0027**	**0.8376 ± 0.0019**

- According to the ablation study, Full modalities with aMLP fusion provides the best performance, especially for the regression task.
- aMLP fusion also synergizes with our MSU-layer. It is a remarkable result.

Evaluation Result





- According to the ablation study, Full modalities with aMLP fusion provides the best performance, especially for the regression task.
- aMLP fusion also synergizes with our MSU-layer. It is a remarkable result.

Contents

1. Introduction
2. Methodology
3. Evaluation Results
4. Conclusion, and Challenges for the future

Conclusion

- We proposed a new method of multimodal-fused sentiment analysis, called **Word-Aware Modality Stimulation Fusion** (WA-MSF).
- Introducing the new concept, **Modality Stimulation Unit layer** (MSU-layer) designed to activate linguistic information within non-verbal modalities by referencing the textual modality sequence prior to the fusion process.
- We also discovered that **aMLP** is the most applicable multimodal fusion method because it has the potential to reconcile the temporal-spatial aspects of non-verbal modalities with textual semantic understanding.

Challenges for the future

- For linguistic: In the field of language processing, LLMs such as GPT-4 are flourishing, but is it possible to incorporate our multimodal fusion method into models (e.g., by Q-former)?
- For non-verbal: Is it possible to leverage particularly Transformer-based embeddings that do not rely on feature extraction methods like OpenFace or COVAREP?