Word-Aware Modality Stimulation for Multimodal Fusion

Shuhei Tateishi, Yasuhito Ohsugi (NTT Docomo, Inc.), Makoto Nakatsuji (NTT Human Informatics Laboratories)

Background

We expect multimodal machine learnings provide superior accuracy to unimodal learning, especially for multimodal sentiment analysis task.

However, we found the monomodal learning by BERT with many training iterations (up to 50 times) is so powerful that it outperforms almost all existing multimodal methods.

In this regard, we determined that there is room for improvement in existing methods, leading us to undertake this study.

Aim

We aimed to elucidate why most other multimodal approaches, including those utilizing BERT for text processing, fall short in accuracy compared to standalone BERT, and to propose a superior model for multimodal sentiment analysis surpassing BERT.

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.

Methodology

Our method, **Word-aware Modality Stimulation Fusion (WA-MSF),** contains **two core concepts:**

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

MSU-Layer is designed for non-verbal (audio and visual) modalities to infuse linguistic 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.

A) For non-verbal modalities

MSU-layer for non-verbal modalities will be injected immediately after a specific layer of the transformer encoder for each modality.

Before this encoder, a Conv1D layer will be applied to match the sequence length of each modality with the length of words of the textual modality.

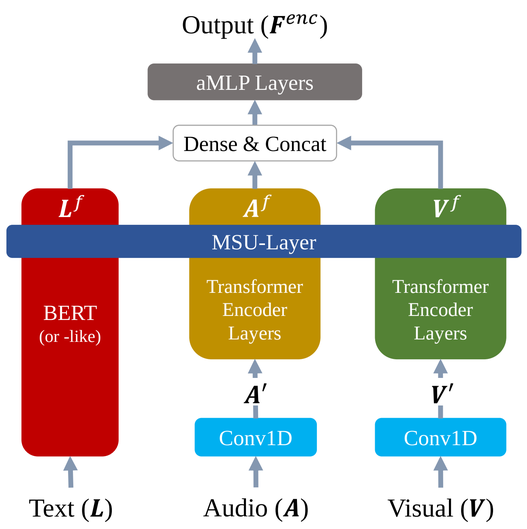

Fig. 1. Overview of WA-MSF

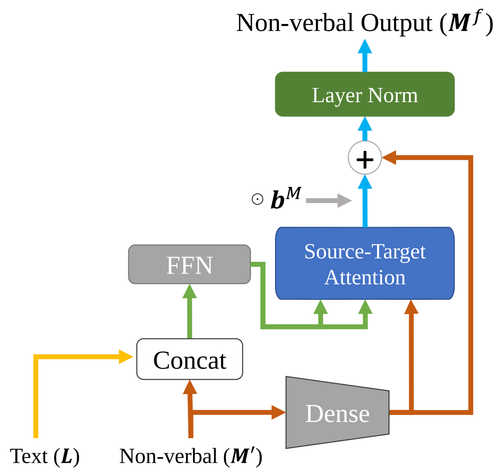

Fig. 2. MSU-layer for non-verbal modalities

B) For textual modality

MSU-layer for the textual modality also will be injected immediately after a specific layer of the BERT (will be detailed later).

However, an effect from nonverbal modalities must be limited for the layers **after the MSU-layer.**

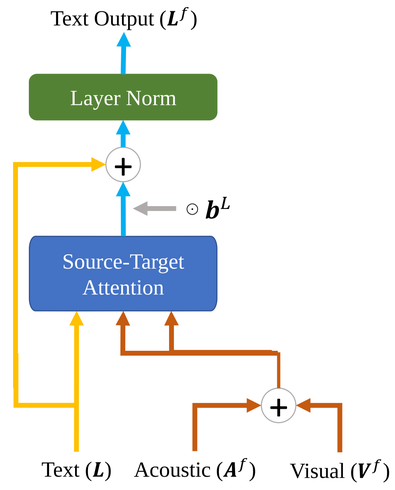

Fig. 3. MSU-layer for textual modality

This is for reducing negative impacts from non-verbal modalities.

Thus, for layers before MSU-layer is inserted, only the textual part of the dataset is pre-trained, then its weights are reused and these layers are frozen while the multimodal training.

After that, for multimodal fusion step, we employ **aMLP** (gMLP with tiny attention) as a multimodal fusion encoder. This is because aMLP has the potential to reconcile the temporal-spatial aspects of non-verbal modalities with textual semantic understanding.

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.

Table 1: Evaluation Result (CMU-MOSI).
$\mathrm{XX_h}$: "higher is better", $\mathrm{XX_l}$: "lower is better".

Method	$\mathrm{F1_h}$	$\mathrm{Acc_h^2}$	$\mathrm{Acc_h^7}$	$\mathrm{MAE_l}$	$\mathrm{Corr_h}$
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	86.42	86.9	48.68	0.691	0.809
BERT-large	86.04	85.98	50.51	0.636	0.838
Ours (Max)	**86.97**	86.86	**51.82**	**0.623**	**0.842**
Ours (Avg)	85.99	85.96	49.99	**0.629**	0.838

Table 2: Evaluation Result (CMU-MOSEI)

Method	$\mathrm{F1_h}$	$\mathrm{Acc_h^2}$	$\mathrm{Acc_h^7}$	$\mathrm{MAE_l}$	$\mathrm{Corr_h}$
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**	54.39	0.523	0.773
BERT-large	N/A	N/A	53.38	0.531	0.775
Ours (Max)	86.09	86.26	**54.63**	**0.515**	**0.785**
Ours (Avg)	85.67	85.80	53.71	**0.520**	**0.782**

Bolded is the best score, and Underlined is the second-best. "Max" score of our method is the best score from 100 attempts of our method evaluation, and "Avg" score is the mean score from them.

The above tables are the results. According to them, our method marks best performance for the **regression task** ($\mathrm{Acc^7}$, MAE, Corr).

The results of ablation studies are as below table:

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

Method	$\mathrm{F1_h}$	$\mathrm{Acc_h^2}$	$\mathrm{Acc_h^7}$	$\mathrm{MAE_l}$	$\mathrm{Corr_h}$
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**

For ease of viewing, We prepared violin charts to compare transformer fusion with aMLP fusion.

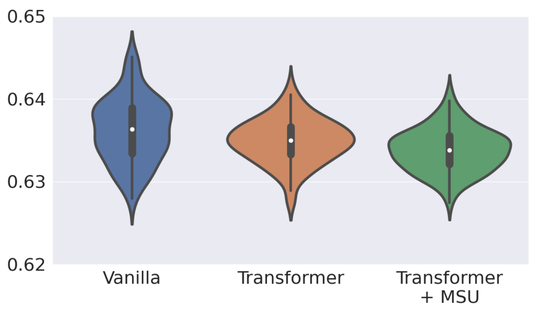

Fig. 4. Mean absolute error, transformer fusion. Lower is better

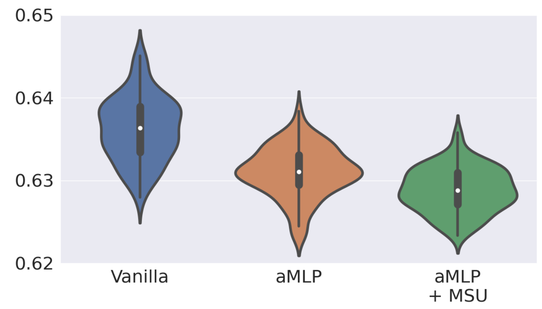

Fig. 5. Mean absolute error, aMLP fusion. Lower is better

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.

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.

Future prospects

- 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?