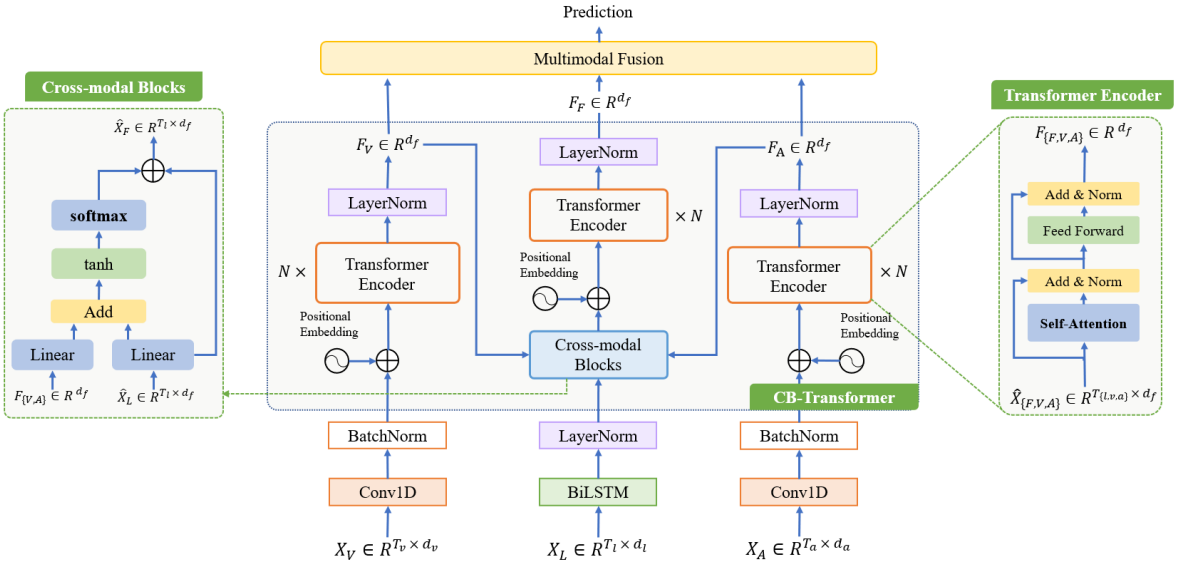
LMR-CBT: Learning Modality-fused Representations with CB-Transformer for Multimodal Emotion Recognition from Unaligned Multimodal Sequences

> Pytorch implementation for Learning Modality-fused Representations with CB-Transformer for Multimodal Emotion Recognition from Unaligned Multimodal Sequences.

Overview

Overall Architecture

In this paper, we propose an efficient neural network to learn modality-fused representations with CB-Transformer (LMR-CBT) for multimodal emotion recognition from unaligned multimodal sequences. Specifically, we first perform feature extraction for the three modalities respectively to obtain the local structure of the sequences. Then, we design a novel transformer with cross-modal blocks (CB-Transformer) that enables complementary learning of different modalities, mainly divided into local temporal learning, cross-modal feature fusion and global self-attention representations. In addition, we splice the fused features with the original features to classify the emotions of the sequences. Finally, we conduct word-aligned and unaligned experiments on three challenging datasets, IEMOCAP, CMU-MOSI, and CMU-MOSEI. The experimental results show the superiority and efficiency of our proposed method in both settings. Compared with the mainstream methods, our approach reaches the state-of-the-art with a minimum number of parameters.



Datasets

Data files (processed MOSI, MOSEI and IEMOCAP datasets) can be downloaded. Due to limitations on the size of the supplemental material and the inability to add links in the double-blind regulations, we were unable to upload the dataset. We will release the data we have processed after the paper is accepted. **You can see the file inside the *result* folder, which contains our word-aligned and unaligned settings for real experiments on the MOSEI dataset.**

Description of Supplementary Files

> Supplementary Material contains four folders, assets, modules, result, and src, in that order.

- The assets folder contains all the overall architecture that appear in the article.
- The modules folder contains the structure of the transformer encoder.
- The result folder contains our word-aligned and unaligned settings for real experiments on the MOSEI dataset.
- The src folder contains the source code of the LMR-CBT.

Run the Code

1. Create (empty) folders for data and pre-trained models:

```
mkdir data pre_trained_models
```

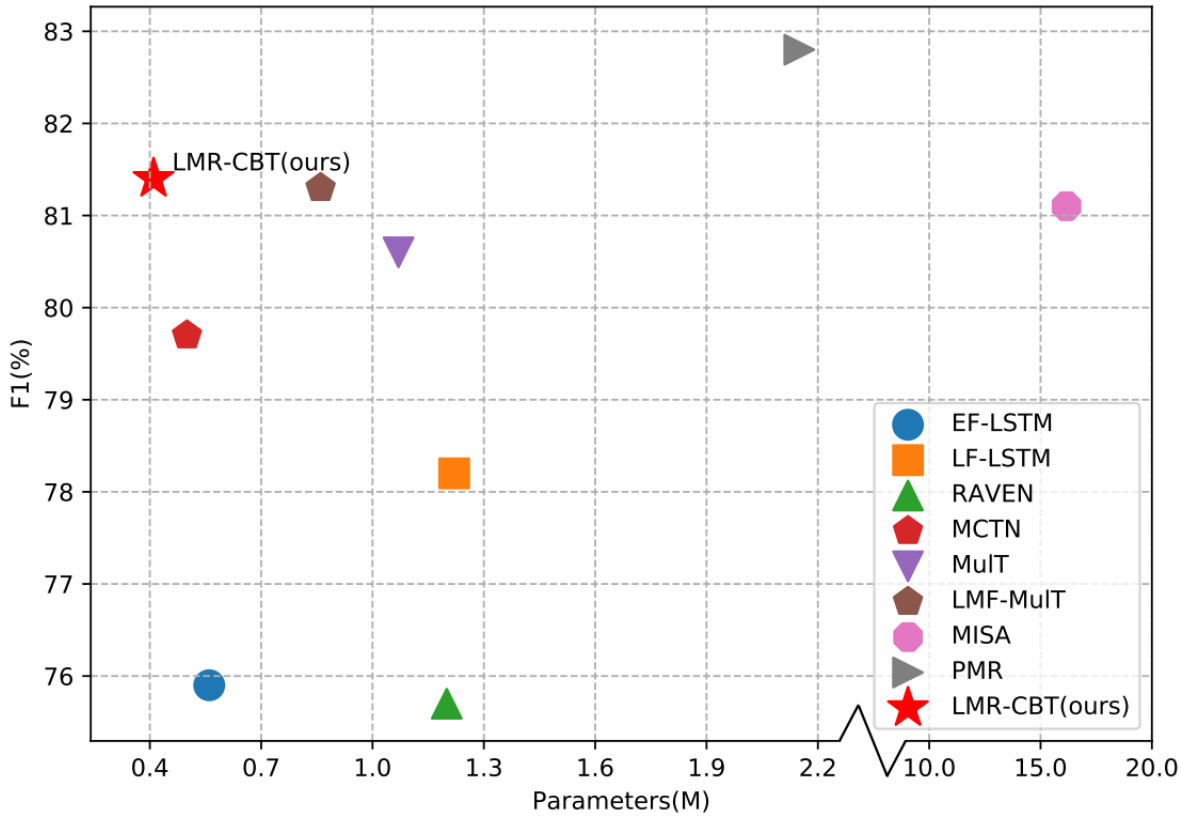
and put the processed data in 'data/'.

2. Command as follows

```
python main.py [--FLAGS]
```

Result

Figure demonstrates that our proposed model reaches the state-of-the-art with a minimum number of parameters (**only 0.41M**) on the CMU-MOSEI dataset. Compared with other approaches, our proposed light-weight network is more applicable to real scenarios.



Model	CMU-MOSEI F1 score	the number of parameters
LMR-CBT(ours)	81.5	0.41M
PMR	82.8	2.15M
MISA	81.1	15.9M
LMF-MulT	81.3	0.86M
MulT	80.6	1.07M
MCTN	79.7	0.50M
RAVEN	75.7	1.20M
LF-LSTM	78.2	1.22M
EF-LSTM	75.9	0.56M