Stronger Baseline for Robust Results in Multimodal Sentiment Analysis

Wei Yang
National Institute of Advanced Industrial Science and Technology (AIST), Japan
wei.yang@aist.go.jp

Jun Ogata
National Institute of Advanced Industrial Science and Technology (AIST), Japan
jun.ogata@aist.go.jp

Abstract

Significant progress in sentiment analysis and emotion recognition has been made recently, and there has also been an increase in the requirements for solving real-world problems. However, human language is complicated and multimodal, making it difficult for computers or artificial intelligence systems to understand. In this study, we adopt several self-supervised learning models to strengthen the learning of representations for multi-modalities (i.e., language, acoustic and visual modalities) to improve the performance of sentiment analysis systems. We also recommend effective fusion and ensemble methods that are easy to apply to obtain a stronger performance. In this study, considering the preliminary experiments, we use a freely available sentiment analysis benchmark, CMU-MOSI. Our recommended techniques and constructed multimodal sentiment analysis systems empirically demonstrate the improvements in the experiment results by approximately 9% and 10.4% for binary classification and multi-classification, respectively, in terms of accuracy, and a 32% reduction in the mean absolute error.

1 Introduction

Sentiment analysis (also known as opinion mining) is a task that aims to systematically identify, extract, quantify, and study emotional states. Sentiment analysis and emotion recognition are both key applications of artificial intelligence (AI). AI has also become indispensable to tasks involving human-computer interaction (HCI).

Some related studies have demonstrated their state-of-the-art (SOTA) result evaluations based on or compared with the sentiment analysis systems constructed using low-level features for single or multiple modalities. Further, it is challenging to determine the amount of improvement that can be achieved using a newly proposed method in comparison to more robust systems built using high-level representations and how effective the method is. This indicates the necessity to know whether the proposed technique or the artificial neural networks (ANNs) can solve new problems or slightly improve the results that have already been addressed or achieved elsewhere.

The main focus of this study was to construct a robust and reliable multimodal sentiment analysis system using several artificial techniques from three perspectives: (1) high-level representations for multi-modalities; (2) robust neural network architecture with an effective fusion strategy; (3) easily applicable and available for solving real-world problems in different fields, such as customer service and clinical medicine (Valstar et al., 2016; Venek et al., 2017).

This study introduces three specific techniques for strengthening the performance of multimodal sentiment analysis and empirically demonstrates how they are applied to improve the experimental results. We also investigate how these techniques improve the performance of sentiment analysis, either singly used or fusion with each other. We also discuss how these empirical experiments and findings can be used to understand the factors affecting the system construction. Our recommended techniques in-

clude:

- Utilize self-supervised learning (SSL) models as feature extractors for every single modality. Considering language modality, we also use a fine-tuned pre-trained SSL model to strengthen the textual (language) representations.
- Introduce a robust crossmodal attention network as the fusion mechanism for overcoming the heterogeneities and long-range dependency problems that may occur across the modalities.
- Predict sentiments by relying on the ensemble of different independent models trained on variant features extracted for the same modality.

The remainder of this paper is organized as follows. Section 2 reviews the related studies and recent progress in deep neural networks (DNNs), sentiment analysis, representations, and fusion techniques. Section 3 introduces a crossmodal attention network for multimodal sentiment analysis. We use it as our basic network architecture for further learning and fusing representations of different modalities. Thereafter, we describe several pre-trained self-supervised learning models applied in our system construction. Section 4 presents the experimental data used, experimental settings, evaluation results, and analysis of the results. We conclude and provide recommendations for future studies in Section 5.

2 Related Studies

Recently, sentiment analysis and emotion recognition have become a popular area for both research and development, meanwhile, a significant progress has been made in the field of machine learning (ML) using deep learning (DL) approaches. The main tasks involved in sentiment analysis include vision/image recognition, natural language processing (NLP), and speech processing.

Some researchers proposed the used of a deep convolutional neural network (CNN) for visual recognition tasks (Simonyan and Zisserman, 2014; Szegedy et al., 2015; He et al., 2016). Several deep learning methods have been applied to different tasks in the field of natural language processing. For instance, the recurrent neural network (RNN) (Sutskever et al., 2014) was used for learning phrase representations for statistical machine translation (SMT) (Cho et al., 2014). Furthermore, by introducing an attention mechanism, RNN's encoder-decoder models improved neural machine translation (NMT) accuracy by jointly learning to align and translate in an end-to-end manner (Bahdanau et al., 2015). On the contrary, the Transformer model (Vaswani et al., 2017) used a new neural network architecture with an attention mechanism that differed from the previous recurrent neural networks or convolutional sequence-to-sequence (ConvS2S) network. The Transformer model computes the representation of a sequence by introducing position-encoding and entirely relying on self-attention mechanisms. The main advantage of the Transformer is that it helps solve the problem of long-range dependencies that are present in the RNNs and ConvS2S. Moreover, it can extract relevant and essential information globally. Another significant advantage is that the Transformer allows the training process to be more parallelizable and faster. In the recent times, the Transformer architecture has also been widely used in speech processing tasks such as automatic speech recognition (ASR) (Wang et al., 2020).

Sentiment analysis has already made progress in terms of unimodal performance (Kontopoulos et al., 2013); however, there still exists a challenge with regard to using multimodal representations and different neural networks to make better AI; in the words, to make it possible for AI to "understand" the sentiments/emotions or intentions of humans more precisely. Sentiments/emotions of humans can be expressed verbally (spoken words or natural language), which refers to language/textual modality as well as nonverbally, that is, via nonverbal behaviors (speech and facial attributes), indicating acoustic and visual modalities. Some previous studies on multimodal representations and fusions were based on pre-aligned human multimodal language sequences. For example, Gu et al. (2018) classified emotions and performed sentiment analysis by introducing a hierarchical multimodal architecture with attention mechanism and word-level fusion of textual and speech modalities.

We aimed to construct a robust and reliable multimodal sentiment analysis system based on "unaligned" language sequences. Our cases can be im-

plemented as a regression and classification problem relying on the high-level representations for different modalities. We used language, acoustic, and visual modalities as a trimodal task.

3 Sentiment Analysis System

The robust representations of single modalities are the cornerstone of multimodal intelligence. Based on the success of adopting pre-trained language models to downstream tasks in natural language processing, we propose the use of several pre-trained self-supervised learning (SSL) models trained by leveraging large-scale single modal datasets to improve the fusion of multi-modalities (i.e., the performance of sentiment analysis systems) with high-level representations (see Figure 1).

3.1 Multimodal Transformer

To construct a robust multimodal sentiment analysis system, we decided to use the multimodal Transformer (MulT) (Tsai et al., 2019) as the basic fusion architecture for multi-modalities.

Based on the standard Transformer network (Vaswani et al., 2017), the MulT was built from multiple stacks of pairwise and bidirectional crossmodal attention blocks. The critical procedures of this fusion and learning mechanism for language (L), acoustic (A) and visual (V) modalities can be illustrated using four steps (see Figure 1):

(1) Similar to the idea of position-wise encoding proposed in (Vaswani et al., 2017), absolute positional encoding ($PE(T_{\{L,A,V\}}, d)$, where $T_{\{L,A,V\}}$ is the sequence length), is added to the embeddings ($\hat{X}_{\{L,A,V\}}$) obtained based on one-dimensional (1D) convolutional layer. This process is applied for the initial input features of each modality to obtain the local structure of the sequences, project the features to the same dimension d, and input the resulting features ($Z^{[0]}_{\{L,A,V\}}$) into N layers of crossmodal attention blocks.

(2) The streams from a single modality are potentially transformed to another by repeatedly reinforcing one modality's features (e.g., L as the target modality) with those from the other modalities (e.g., A and V as the source modalities) via the scaled dot-product attention-based multi-head crossmodal attention.

(3) Production of reinforced feature vectors of each target modality is concatenated by a simple concatenation operation. As the example shown in Figure 1, language is the target modality that repeatedly receives information from the acoustic and visual modalities: $Z_L = \left[Z^{[N]}_{A\to L}; Z^{[N]}_{V\to L}\right]$.

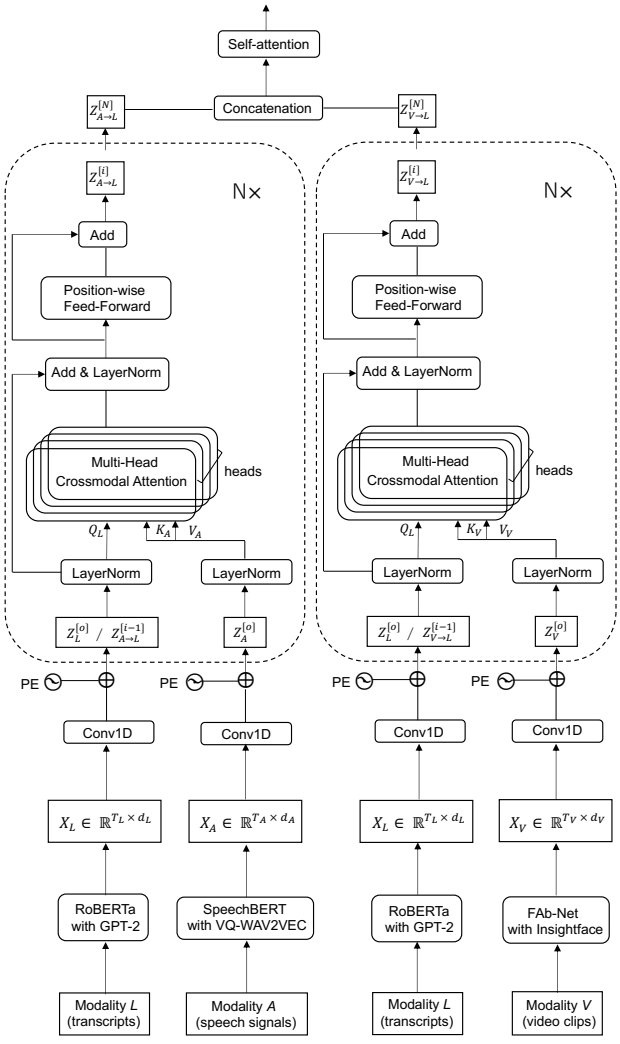

Figure 1: Overview of the sentiment analysis system constructed based on the multimodal Transformer on language (L), acoustic (A), and visual (V) modalities utilizing several pre-trained SSL models. The figure is an illustration of passing an acoustic and visual information to language modality (denoted by $A, V \to L$). This is similar to $A, L \to V$, and $L, V \to A$.

(4) Finally, the concatenated feature vectors can be used with a sequence of the self-attention Transformer blocks to output information and are integrated to pass through fully-connected layers for the final prediction.

3.2 RoBERTa with GPT-2

In our study, instead of using a pre-trained "language understanding" model, BERT (Devlin et al., 2019) for language modality, we used the RoBERTa (Liu et al., 2019). RoBERTa is a robustly optimized BERT-like pre-trained model that can be trained on longer sequences with bigger batches and using more training data by following the basic pre-training procedure of BERT. However, compared to BERT, during RoBERTa's pre-training, the subsequent sentence prediction objective is removed, and the masking pattern is dynamically changed. RoBERTa has demonstrated better performances on several downstream tasks (Rajpurkar et al., 2016; Rajpurkar et al., 2018; Wang et al., 2018; Wang et al., 2019).

Regarding text encoding, we used a byte-level language model known as GPT-2 (Radford et al., 2019), which introduces mapping tables between bytes and their corresponding Unicode strings for subword tokenization, regardless of the word segmentation.

3.3 SpeechBERT with VQ-WAV2VEC

The VQ-WAV2VEC (Baevski et al., 2020), as a new discretization approach, was proposed to learn discrete representations for speech by encoding raw consecutive audio signals into fixed-length discretized sequences. It facilitates the direct use of discretized speech sequences in several NLP tasks and applications, including speech recognition and spoken question answering (SQA) and can also be used to train a "speech understanding" model. In our study, we propose the use of a BERT-like pre-trained model trained with the RoBERTa training task on discrete representations of audio segments obtained by VQ-WAV2VEC, to obtain the initial representations for the acoustic modality.

As the acoustic feature extractor, we used a pre-trained SpeechBERT model trained on a discretized 960 h Librispeech dataset (Panayotov et al., 2015).

3.4 FAb-Net with Insightface

For pre-processing, we used "Insightface" (Deng et al., 2019) to perform face recognition from raw video clips. This pre-trained face recognition model trained using an Additive Angular Margin Loss function (ArcFace) capable of highly discriminating features is expected to further improve the accuracy of the face recognition task. This method and its corresponding pre-trained model have been proven to be more discriminatively efficient, compared to the recent state-of-the-art face recognition methods (Liu et al., 2017) applied on different large-scale, image and video datasets (Huang et al., 2007; Wolf et al., 2011).

We also used a pre-trained model trained on a self-supervised framework Facial Attributes-Net (FAb-Net) (Wiles et al., 2018) to encode each facial frame recognized by "Insightface". The FAb-Net is known for its ability to learn meaningful face embeddings that encode facial attributes such as head pose, expression, and facial landmarks. A pre-trained *affectnet* model enables the acquisition of emotion-related representations for the visual modality.

4 Experiments

This section presents the results of a sequence of experiments that were performed for constructing robust multimodal sentiment analysis systems. Beyond reporting the significant improvements in different evaluation metrics compared to the baseline system and the previous studies, we also conducted an in-depth analysis to show the underlying reason for the effectiveness and improvements that we achieved and to describe the inherent weaknesses (of the baseline system and the previous studies) that we addressed. We also performed a sequence of ablation studies.

4.1 Data Preparation and Evaluation Metrics

CMU-MOSI (Zadeh et al., 2016) is one of the most frequently used human multimodal sentiment analysis datasets in English, which comprises 2,199 short monologue video clips (train: 1,284; validation: 229; test: 686) with their corresponding transcripts, and audio information. Human annotators label each example (lasting the duration of a sen-

tence) in CMU-MOSI with a sentiment score from -3 (strongly negative) to 3 (strongly positive).

CMU-MOSI is also a speaker-independent and fine-grained sentiment analysis dataset that does not focus on only the [negative and positive] sentiments. Thus, we also evaluate the performance of our constructed models using various metrics, similar to those applied in previous studies, including the seven-class (i.e., Acc_7: sentiment score classification in [-3, 3]) and binary (that is, Acc_2: positive/negative sentiments) accuracies, weighted *F1* score (based on the binary classification), *MAE* of the scores (mean absolute error, based on the seven-class sentiment score prediction), and correlation (*Corr*) of the model's prediction with humans.

4.2 Baseline System

The experiments for CMU-MOSI reported in (Tsai et al., 2019) has two versions: (i) "word-aligned" and (ii) "unaligned" versions. The former aligns the timesteps of acoustic and visual modalities based on the words in the language modality. Considering the latter, the length of the textual sequences is the same as the length obtained in the former; nevertheless, it keeps the original length of the acoustic and visual features without any word-segmented alignment. The feature dimension mentioned in (Tsai et al., 2019) is 35 for Facet (iMotions, 2017) and 74 for COVAREP (Degottex et al., 2014).

The initial features used as our baseline for the "unaligned" version is different from the feature dimensions used in (Tsai et al., 2019). Here, we prefer our "unaligned" baseline features to be the feature sets used in (Chen et al., 2017) for CMU-MOSI provided on the Web[1]. These features are 20 best selected features from Facet feature and five features from COVAREP feature by using univariate linear regression tests. Furthermore, the selected Facet and COVAREP features are linearly normalized by the maximum absolute value in the training set. In this study, we use these three groups of evaluation results in Table 1.

4.3 Experimental Setting

To compare the behavior of our systems to the baseline system and the previous studies, we conduct experiments using the same hyperparameter settings described in (Tsai et al., 2019) for the CMU-MOSI dataset. The main differences between our experimental setting and performance provided in (Tsai et al., 2019) are as follows.

(1) We use subword-based tokenization instead of word-based tokenization to prepare for the initial textual representations. In (Tsai et al., 2019), textual data are segmented per word and are expressed as discrete word embeddings using pre-trained Glove word embeddings (*glove.840B.300d*) (Pennington et al., 2014). The embedding for each word is a 300-dimensional vector.

(2) There is maximum utilization of subword-based pre-trained SSL models. Alternatively, the RoBERTa language model prefers *roberta.large* or *roberta.large.mnli*, which is a fine-tuned *roberta.large* on a sentence pair classification task (model the textual interaction between a pair of sentences, e.g., contradiction or entailment). Because the last layer of the pre-trained model is too close to the target functions during the pre-training, we average the second-to-last hidden layer of each token ([CLS] + subword tokens + [SEP]) to produce a single 1024-dimension embedding as the entire sentence representation.

(3) Another difference refers to the use of COVAREP. We use a pre-trained model SpeechBERT (*bert_kmeans*) with *vq-wav2vec_kmeans* for audio tokenization and speech feature extraction as described in Section 3.3. This is a similar process for acoustic representation by averaging the second-to-last hidden layer of each speech token to obtain a single 768-dimension embedding as an entire audio representation.

(4) Regarding visual modality, instead of using Facet as described in (Chen et al., 2017; Tsai et al., 2019), we extracted emotion-related facial attributes using the FAb-Net-based pre-trained model for frames that were recognized based on "Insightface" facial frame recognition model introduced in Section 3.4. The maximum time-depth is set to 300 for frame recognition, and the feature dimension for each frame is 256.

(5) The final technique that we proposed refers to combine using multiple sentiment analysis models for the final prediction, indicating the ensemble of the two best models by averaging their predic-

[1] http://immortal.multicomp.cs.cmu.edu/raw_datasets/processed

tions (regression results). The idea of ensembling multiple models is powerful and is widely used in machine learning and NLP fields.

Evaluation is conducted by averaging the evaluation scores obtained over multiple runs. We shuffled the training, validation, and test data for each experimental condition, performed ten runs, and reported averaged metric scores. All the test results (evaluation scores) are presented in Table 1, 2, and 3.

4.4 Analysis of the Results and Discussion

As can be seen from Table 1, the independent application or a combination of these techniques led to a consistent improvement in a large margin across all the scenarios compared to their corresponding baselines.

Our systems outperform the baseline system by a margin of 5.1∼9% and 3∼10.4% on the binary accuracy and seven-class metric, respectively. The proposed methods also attribute to *MAE* and *Corr* because they decrease or increase with a large margin of 32.2% and 16.7%, respectively. Our experiments show that when the techniques are combined using several pre-trained SSL models with the crossmodal Transformer as the fusion mechanism, it facilitates the achievement of the SOTA results compared to the "unaligned" scenario and "word-aligned" setting with low-level features.

Another observation is that the advantages provided by a well pre-trained SSL model comprising rich semantic and contextual information benefits the training process of sentiment analysis systems. In our study, instead of using *roberta.large* ($L1$), we used *roberta.large.mnli* ($L2$), which was fine-tuned on the sentence pairs classification task, for the representation extraction for language modality. Using a fine-tuned pre-trained SSL model overcomes the underlying problems of initial low-level features of baseline systems. It also alleviates the "non-finetuned" problem in which we did not perform any "fine-tuning" on the sentiment analysis task itself.

In addition, in our experiments, we maintained the original features (A and V) and only replaced L with $L1/L2$ or replaced both L and V with $L1/L2$ and $V1$, respectively, enabled us to obtain a better Acc_7, *MAE*, and *Corr* at the cost of a decreased Acc_2 sometimes ($L1+A+V$ and $L2+A+V$), compared to replacing A with $A1$ in any case. We hypothesized that such a performance drop occurred because $A1$ was the only pre-trained SSL model unrelated to emotions, and it lacked fine-tuning. Nonetheless, they still outperformed the baseline system.

	Acc_7	Acc_2	*F1*	*MAE*	*Corr*
CMU-MOSI Sentiment in (Tsai et al., 2019)					
(i) Word-aligned	40.0	83.0	82.8	0.871	0.698
(ii) Unaligned	39.1	81.1	81.0	0.889	0.686
CMU-MOSI Sentiment (Multi-modal, unaligned)					
Baseline ($L+A+V$)	34.5	78.3	78.2	1.013	0.648
$L1+A1+V$	37.5 (+3.0)	84.3 (+6.0)	84.3 (+6.1)	0.841 (-0.172)	0.740 (+0.092)
$L1+A1+V1$	40.3 (+5.8)	83.9 (+5.6)	83.9 (+5.7)	0.799 (-0.214)	0.758 (+0.110)
$L1+A+V1$	41.5 (+7.0)	84.3 (+6.0)	84.3 (+6.1)	0.755 (-0.258)	0.787 (+0.139)
$L1+A+V$	41.6 (+7.1)	83.4 (+5.1)	83.5 (+5.3)	0.757 (-0.256)	0.777 (+0.129)
$L2+A1+V$	39.7 (+5.2)	85.9 (+7.6)	85.9 (+7.7)	0.792 (-0.221)	0.783 (+0.135)
$L2+A1+V1$	40.7 (+6.2)	86.0 (+7.7)	86.0 (+7.8)	0.775 (-0.238)	0.786 (+0.138)
$L2+A+V1$	41.9 (+7.4)	85.9 (+7.6)	85.9 (+7.7)	0.733 (-0.280)	0.810 (+0.162)
$L2+A+V$	43.3 (+8.8)	85.2 (+6.9)	85.1 (+6.9)	0.740 (-0.273)	0.795 (+0.147)
Ensemble of two $L2+A+V$	**44.9 (+10.4)**	86.6 (+8.3)	86.6 (+8.4)	0.703 (-0.310)	0.808 (+0.160)
Ensemble of $L2+A1+V$ & $L2+A1+V1$	43.1 (+8.6)	**87.3 (+9.0)**	**87.3 (+9.1)**	0.758 (-0.255)	0.796 (+0.148)
Ensemble of $L2+A+V1$ & $L2+A+V$	44.5 (+10.0)	87.0 (+8.7)	87.0 (+8.8)	**0.691 (-0.322)**	**0.815 (+0.167)**

Table 1: Evaluation results of the construction of stronger multimodal sentiment analysis systems on CMU-MOSI benchmark using our proposed techniques. The evaluation metrics (excluding *MAE*) indicate a desirable performance when their values are high. $\{L, A, V\}$ represents the {Glove, COVAREP, Facet}-based features for each modality. $\{L1, L2, A1, V1\}$ denotes {roberta.large, roberta.large.mnli, SpeechBERT, FAb-Net}-based new representations. "+" refers to multimodal Transformer functions for each modality and fusion making.

Two strategies apply the ensemble for constructing our more robust sentiment analysis systems: (1) ensembling two best models trained on the same feature sets with different runs; (2) ensembling two best models trained on different feature sets, indicating that they are fully independent and more diverse models.

We empirically found that a more significant improvement was obtained when the prediction moved from a single model to model ensembles. Furthermore, an ensemble of two best fully independent

models trained on different feature sets (ensemble of $L2{+}A1{+}V$ & $L2{+}A1{+}V1$ and $L2{+}A{+}V$ & $L2{+}A{+}V1$) achieved better results than ensembling two best models trained on the same feature sets with different runs (ensemble of two $L2+A+V$) on Acc_2, *F1*, *MAE*, and *Corr*. Similar to the trends of the predictions performed based on the single model, two models that include the COVAREP-based features for acoustic modality always yielded better seven-class accuracy (Acc_7). In contrast, replacing the COVAREP-based features (A) with SpeechBERT-based representations ($A1$) led to a better binary accuracy (Acc_2).

Considering these experiments, we empirically concluded that the ensemble of $L2 + A + V$ & $L2 + A + V1$ is a better option for achieving more balanced and robust results. Our experimental results demonstrated both the effectiveness of ensemble and the importance of model diversity for obtaining more robust sentiment analysis systems.

4.5 Ablation Studies

To further understand the origin of these improvements and the influence of individual representation functions for each modality in our sentiment analysis system construction, we performed ablation analysis, similar to the performance by (Tsai et al., 2019) for CMU-MOSEI (Bagher Zadeh et al., 2018).

We separated the performance for different cases: (1) using only a single modality with self-attention Transformer (uni-modal); (2) repeatedly reinforcing only one modality's features with those from other modalities (e.g., Only A, $V \rightarrow L2$).

Based on the results presented in Table 2, the performance of uni-modal is extremely good under both "non-finetuned" and "fine-tuned" conditions for the language modality represented as entire sentence embeddings using the pre-trained RoBERTa model. Using only "non-finetuned" RoBERTa resulted in a better performance on Acc_7, *MAE*, and *Corr*, compared to $L1 + A1 + V$ and $L1 + A1 + V1$ (see Table 1). However, considering other cases, the crossmodal Transformer performed better. Shifting the "non-finetuned" to "fine-tuned" RoBERTa led to approximately the best results on Acc_2, compared to the multimodal case presented in Table 1. Language modality may be negatively affected by acoustic and visual modalities during crossmodal fusion. Nonetheless, we empirically observed that better results cannot be obtained by ensembling two language-only "*roberta.large.mnli*"-based single models. Considering the experimental results, we also found that the lower the level of the features for the single modalities, the more necessary the multi-modality fusion becomes. The specific improvement in a single modality (which is language) can be particularly helpful for handling low-level feature problems in baseline systems.

Replacing the original features used in the baseline system with new ones for acoustic or visual modalities in uni-modal performance did not show large improvements. This may be because they are either not sentiment/emotion-related or not fine-tuned on any related task.

Regarding the previous MulT experiments, among the three targeted modality-dependent cross-modal Transformers, language (as the target modality) worked the best, compared to the other two cases on CMU-MOSEI. Considering our experiments, we empirically observed that language and acoustic be-

		Acc_7	Acc_2	*F1*	*MAE*	*Corr*
CMU-MOSI Sentiment (Uni-modal)						
Language-only	Glove	35.2	77.4	77.3	0.996	0.640
	roberta.large	41.1	83.1	83.1	0.781	0.764
	roberta.large.mnli	**42.6**	**86.7**	**86.6**	**0.722**	**0.807**
Audio-only	COVAREP	16.8	49.5	54.8	1.443	0.183
	SpeechBERT	16.2	**51.8**	54.7	**1.419**	0.116
Vision-only	Facet	19.2	56.1	55.8	1.389	0.174
	FAb-Net	18.1	**57.6**	**59.7**	1.392	0.101
CMU-MOSI Sentiment (Multi-modal, target modality-dependent)						
Only $A, V \rightarrow L$ (Baseline)		35.9	76.4	76.3	1.047	0.592
Only $A1, V \rightarrow L2$		39.2	85.3	85.3	0.810	0.775
Only $A1, V1 \rightarrow L2$		40.4	85.4	85.4	0.794	0.773
Only $A, V1 \rightarrow L2$		42.5	86.2	86.2	0.735	0.809
Only $A, V \rightarrow L2$		**43.3**	85.6	85.6	0.727	0.793
Only $L, V \rightarrow A$ (Baseline)		33.7	77.1	77.0	0.999	0.645
Only $L2, V \rightarrow A1$		41.1	86.6	86.6	0.764	0.792
Only $L2, V1 \rightarrow A1$		40.0	86.6	86.6	0.769	0.798
Only $L2, V1 \rightarrow A$		40.9	85.8	85.8	0.731	0.810
Only $L2, V \rightarrow A$		43.0	86.4	86.4	**0.716**	0.808
Only $L, A \rightarrow V$ (Baseline)		34.7	79.0	79.0	0.971	0.652
Only $L2, A1 \rightarrow V$		39.3	**86.7**	**86.7**	0.808	0.786
Only $L2, A1 \rightarrow V1$		38.4	85.6	85.6	0.830	0.772
Only $L2, A \rightarrow V1$		40.6	86.1	86.0	0.755	**0.815**
Only $L2, A \rightarrow V$		41.8	86.3	86.2	0.732	0.801

Table 2: Experimental results of the ablation studies on the benefit of using several pre-trained self-supervised models for constructing more robust sentiment analysis systems on unaligned CMU-MOSI.

	Acc_7	Acc_2	*F1*	*MAE*	*Corr*
CMU-MOSI Sentiment (Bi-modal, unaligned)					
$L + A$ (Baseline)	35.3	79.5	79.4	0.965	0.675
Only $L \rightarrow A$	34.6	76.8	76.7	1.033	0.620
Only $A \rightarrow L$	34.9	78.5	78.4	1.005	0.649
$L2 + A1$	40.3	85.5	85.5	0.802	0.774
Only $L2 \rightarrow A1$	41.1	86.4	86.3	0.764	0.798
Only $A1 \rightarrow L2$	39.3	84.9	84.8	0.817	0.764
$L2 + A$	42.2	86.5	86.5	0.733	0.809
Only $L2 \rightarrow A$	41.9	86.8	86.8	0.709	0.814
Only $A \rightarrow L2$	42.6	86.3	86.3	0.719	0.809
Ensemble of two $A \rightarrow L2$	43.3	**87.8**	**87.8**	**0.690**	**0.824**
Ensemble of $L2 + A$ & $A \rightarrow L2$	**44.6**	87.3	87.4	**0.690**	**0.821**

Table 3: Evaluation results for the bi-modal sentiment analysis system for unaligned CMU-MOSI.

ing the target modalities were of approximately the similar strength (only $A, V \rightarrow L2$, and only $L2, V \rightarrow A$) as a full crossmodal Transformer ($L2 + A + V$). Moreover, the target modality-dependent Transformer performed slightly better on most of the evaluation metrics.

In summary, the designed and finished experiments confirmed that the strength of the language representation plays a crucial role in multimodal sentiment analysis (i.e., the tri-modal case). On the contrary, this study also proved that it should work for bi-modal sentiment analysis if one of the non-verbal modalities' information may be missing (e.g., only the language and acoustic modalities). Table 3 shows our confirmed results. Without the negative effect during the crossmodal fusion, only using a fine-tuned RoBERTa ($L2$) modal and COVAREP features (A) for language and acoustic modalities enabled us to obtain even better results for approximately all the evaluation metrics. Furthermore, although only one modality with strong representations performs well (i.e., language), the ensemble mechanism, using more diverse models and leveraging the complementarity of multi-modalities, can represent richer information to make the predictions more robust. Our analysis also indicated that using the target modality-dependent crossmodal Transformer with strong representations instead of the full crossmodal Transformer is beneficial for faster training and saves time without any loss of accuracy. This will be especially helpful for training models on large datasets.

5 Conclusion and Future Studies

This study empirically recommends and introduces different pre-trained SSL models that can be relatively easily applied on different modalities to extract robust representations, resulting in stronger sentiment analysis systems.

Considering the predictive power of deep learning, we empirically demonstrate the effectiveness of high-level representations, effective fusion mechanism, and independent model ensembling for constructing a stronger sentiment analysis system for robust results. Our proposed system can serve as a strong baseline capable of capturing long-range contingencies, regardless of the alignment assumption, and solving lower accuracy suffers from the low-level feature problem.

Our experiments only used CMU-MOSI in the sequence of in-depth studies that we conducted several times to obtain stable and reliable results. In the future, we plan to use CMU-MOSEI (approximately ten times the size of CMU-MOSI) and IEMOCAP (Busso et al., 2008) as the experimental data for emotion recognition. We also plan to apply a similar idea to other sentiments or emotion benchmarks in different languages and modalities. Fine-tuning for the pre-trained SSL models used in this study on the sentiment analysis task itself is one of our future work. We intend to perform another comparison in our subsequent studies.

We believe that the current work will facilitate future studies in the field of multimodal sentiment analysis. We also believe that our proposed systems can be used for different applications to solve real-world problems because our methods can be easily extended and work effectively, regardless of the size of the training dataset. Therefore, we recommend the evaluation of new model extensions and algorithms for use in sentiment analysis systems that are as strong as the ones we have proposed, in order to obtain reliable results.

Acknowledgments

This work was supported by the Council for Science, Technology, and Innovation (CSTI), "Cross-ministerial Strategic Innovation Promotion Program (SIP), Big-data and AI-enabled Cyberspace Technologies" (funding agency: NEDO).

References

Alexei Baevski, Steffen Schneider, and Micheal Auli. 2020. VQ-WAV2VEC: Self-supervised learning of discrete speech representations. In *Proceedings of Eighth International Conference on Learning Representations (ICLR 2020)*, pages 1–12.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. 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*, pages 2236–2246.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

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):335–359.

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.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734.

G. Degottex, John Kane, Thomas Drugman, T. Raitio, and Stefan Scherer. 2014. COVAREP–A collaborative voice analysis repository for speech technologies. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2014)*, pages 960–964.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. ArcFace: Additive angular margin loss for deep face recognition. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2019)*, pages 4690–4699.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Yue Gu, Kangning Yang, Shiyu Fu, Shuhong Chen, Xinyu Li, and Ivan Marsic. 2018. Multimodal affective analysis using hierarchical attention strategy with word-level alignment. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 2225–2235.

Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. 2007. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst.

iMotions. 2017. Facial expression analysis.

Efstratios Kontopoulos, Christos Berberidis, Theologos Dergiades, and Nick Bassiliades. 2013. Ontology-based sentiment analysis of twitter posts. *Expert Systems with Applications*, 40(10):4065–4074.

Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. 2017. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6738–6746.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, M. Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized bert pretraining approach. *arXiv preprint*, arXiv: 1907.11692.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. In *Proceesings of IEEE international Conference on Acoustics, Speech and Signal Processing (ICASSP2015)*, pages 5206–5210.

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.

A. Radford, Jeffrey Wu, R. Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1:9.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable Questions

for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 784–789.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *arXiv:1409.3215v3*.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9.

Yao-Hung Hubert Tsai, Shaojie Bai, and Paul Pu Liang *et al.* 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6558–6569.

Michel Valstar, Jonathan Gratch, Björn Schuller, Fabien Ringeval, Denis Lalanne, Mercedes Torres Torres, Stefan Scherer, Giota Stratou, Roddy Cowie, and Maja Pantic. 2016. AVEC 2016: Depression, mood, and emotion recognition workshop and challenge. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, pages 3–10.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*, pages 6000–6010.

Verena Venek, Stefan Scherer, Louis-Philippe Morency, Albert Skip Rizzo, and John Pestian. 2017. Adolescent suicidal risk assessment in clinician-patient interaction. *IEEE Transactions on Affective Computing*, 8(2):204–215.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In *arXiv:1905.00537*.

Yongqiang Wang, Abdelrahman Mohamed, Duc Le, Chunxi Liu, Alex Xiao, Jay Mahadeokar, Hongzhao Huang, Andros Tjandra, Xiaohui Zhang, Frank Zhang, Christian Fuegen, G. Zweig, and M. Seltzer. 2020. Transformer-based acoustic modeling for hybrid speech recognition. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6874–6878.

O. Wiles, A.S. Koepke, and A. Zisserman. 2018. Self-supervised learning of a facial attribute embedding from video. In *arXiv:1808.06882*.

Lior Wolf, Tal Hassner, and Itay Maoz. 2011. Face recognition in unconstrained videos with matched background similarity. *CVPR 2011*, pages 529–534.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. MOSI: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *IEEE Intelligent Systems 31.6 (2016): 82-88*.