Simple and Effective Multimodal Learning Based on Pre-Trained Transformer Models

KAZUKI MIYAZAWA, YUTA KYURAGI, AND TAKAYUKI NAGAI, (Member, IEEE)

1Graduate School of Engineering Science, Osaka University, Osaka 565-0871, Japan
2Artificial Intelligence Exploration Research Center, The University of Electro-Communications, Tokyo 182-8585, Japan

Corresponding author: Takayuki Nagai (nagai@sys.es.osaka-u.ac.jp)

This work was supported in part by the Japan Society for the Promotion of Science (JSPS) KAKENHI under Grant JP19J23364, and in part by the Grant-in-Aid for Scientific Research on Innovative Areas under Grant 20H05565.

ABSTRACT

Transformer-based models have garnered attention because of their success in natural language processing, and in several other fields, such as image and automatic speech recognition. In addition to them being trained on unimodal information, many transformer-based models have been proposed for multimodal information. In multimodal learning, a common problem encountered is the insufficiency of multimodal training data. In this study, to address this problem, a simple and effective method is proposed by using 1) unimodal pre-trained transformer models as encoders for each modal input and 2) a set of transformer layers to fuse their output representations. Further, the proposed method is evaluated by conducting several experiments on two common benchmarks: CMU multimodal opinion sentiment intensity dataset and multimodal internet movie database. The proposed model exhibits state-of-the-art performances on both benchmarks and is robust against the reduction in the amount of training data.

INDEX TERMS Multimodal machine learning, transformers, pre-trained models, emotion recognition.

I. INTRODUCTION

Humans live in a world with multimodal information. They see, hear, smell, touch, and taste many things in their daily lives. In addition, they communicate and understand fellow humans through words, tones, and facial expressions. Thus, learning the relations among different modalities and interpreting their meaning are essential for creating artificial intelligence that can understand the world. To address this problem, multimodal learning is employed for building models that can process and relate information from multiple modalities [1].

Recently, transformer-based [2] models have attracted significant attention in various research fields. Although the original transformer [2] was designed for natural language processing (NLP) tasks, its mechanisms have been applied in various models using other modalities such as image and audio [3]–[6]. Because of the success of these modalities, the interest in applying the transformer mechanism to multimodal learning has increased. However, training these transformer-based models or any other neural-network-based models requires a large amount of training data [1].

Several large datasets for individual modalities are available; however, multimodal datasets are difficult to create and find because the collection of multimodal data is more difficult than the collection of unimodal data. Consequently, in multimodal learning, the problem of insufficient multimodal training data is common. When the amount of training data is insufficient, a multimodal model is more likely to encounter unfamiliar input words or patterns in the test dataset. For example, words that rarely or never appear in the training dataset can be encountered. To address this problem, existing pre-trained models trained on large datasets are used to extract better representations from each modal input. This enables the multimodal model to understand the meaning of such unknown inputs. Thus, in this study, a simple and effective method is proposed for solving the aforementioned problem by using 1) unimodal pre-trained transformer models as encoders for each modal input and 2) a set of transformer layers to fuse their output representations.

In multimodal learning, the fusion of different modality information is considered important. Many types of multimodal fusion exist; however, the major approaches can be classified into two types: early and late fusion [7]. The former fuses low-level features, and the latter fuses prediction-level features. Thus, based on other multimodal transformer
models [8]–[12], the proposed model was structured by using two pre-trained transformer models and a set of transformer layers to fuse the outputs from the two pre-trained transformer models (Fig. 1). Although the outputs of the pre-trained transformer models were concatenated in a late-fusion manner, instead of the [CLS] outputs, which can represent the entire sequence in a single token, being simply concatenated, similar to a shallow-fusion model [10], the output representations were fused using a few transformer layers. Moreover, the self-attention mechanism helped the model to extract the necessary information from both modality outputs, depending on the original inputs.

An advantage of the proposed method is that the pre-trained models can be easily alternated depending on the availability of a better pre-trained model or the input modalities. This can be easily accomplished because fine-tuning in advance with the pre-trained transformer models is not needed, and thus, these pre-trained models can be directly used as encoders for each modality. Another pre-trained transformer model can be easily added depending on the modalities of the training data, because the fusion mechanism is simply a concatenation of the output representations of the pre-trained transformer models. Thus, the proposed method was assumed to be applicable to any combination of modalities.

The effectiveness of the proposed method was verified by evaluating it on two common benchmarks: the CMU multimodal opinion sentiment intensity dataset (CMU-MOSI) [13] and the multimodal internet movie database (MM-IMDb) [14]. The results indicated that the proposed method outperformed the latest models and achieved state-of-the-art performance in terms of most metrics on both the datasets. The proposed method was also evaluated using a reduced amount of training data for each dataset and was found to outperform most of the latest models using only 25% to 50% of the original training dataset. This robustness to the reduction in the training data amount was attributed to the better representations from the pre-trained transformer models and the structure required during their pre-training. Because pre-trained transformer models are trained on a large amount of data within their modality, they provide better representations of the data that never appear in the multimodal training dataset.

II. RELATED WORK
Till date, many transformer-based models have been proposed and used for single-modality and multimodal tasks. Herein, a few of the latest pre-trained transformer models are introduced, and subsequently, models focusing on multimodal fusions are discussed.

A. PRE-TRAINED TRANSFORMER MODELS
Similar to BERT [15], many transformer-based models are pre-trained to learn general representations before they are fine-tuned to other downstream tasks. This pre-training phase is often performed with a large amount of unsupervised data for an extended period of time, and subsequently, the models are expected to be fine-tuned to a certain smaller downstream dataset for a short period of time. Recently, many pre-trained transformer models have been proposed. In the NLP field, many improved models of the BERT [15] and transformer [2] have been proposed, such as RoBERTa [16], XL-Net [17], Longformer [18], and GPT-2 [19]. Although the transformer was originally designed for NLP tasks, the model has also been successful in other fields that involve the use of different modality information. The vision transformer (ViT) [4] has been proposed for image recognition and classification. The ViT is often referred to as a state-of-the-art image classification model, and it has been successful as an image encoder in other models, such as CLIP [20] and video transformer networks [21]. The video transformer network was designed for video recognition, where a ViT was used to extract the features for each frame. Further, models such as Wav2Vec [3], VQ-Wav2Vec [6], Wav2Vec 2.0 [5], and Conformer [22] have been demonstrated to be successful in audio processing and speech recognition tasks. These pre-trained transformer models are composed of similar architectures using transformers [2]; however, they are trained on a large dataset of their target modality.

B. MULTIMODAL FUSION
In multimodal learning, a model must combine the information of different modalities to improve the performance of multimodal tasks. Variational autoencoder (VAE)-based models, such as multimodal VAEs [23] and joint multimodal VAEs [24], have been proposed to learn a shared representation across modalities. Since the proposal of the transformer [2], its performance and ability to visualize an entire input sequence have attracted attention, and consequently, many models have applied its attention mechanism to multimodal learning using the models presented in the previous section. The multimodal bitransformer [8], LXMERT [11], ViLBERT [9], VisualBERT [12], and PixelBERT [25] extend the BERT [15] architecture and change the inputs to text and images. These models use text-pre-trained transformer models as text encoders and other mechanisms for image encoders such as ResNet-152 [26] or trained their own transformer as an image encoder. In contrast, CLIP [20] uses the ViT [4] as its image encoder and text information as weakly supervising information. These models either use a pre-trained transformer model as an encoder for one of their modalities or simply use the structure of the transformer. The shallow fusion of self-supervised learning (SSL) models [10] uses two single-modality-pre-trained transformer models, where the [CLS] outputs from the two pre-trained models are concatenated in a late-fusion manner, and subsequently, the model is fine-tuned for multimodal emotion analysis tasks. Similar to the proposed model, this model relies on pre-trained transformer encoder models. However, the difference lies in the fact that the proposed model uses the self-attention mechanism to
The overall architecture of the proposed model is shown and the pre-trained models used are explained. In this section, the overall structure of the proposed model attention mechanism. the necessary information from the outputs using the self-
make predictions from the [CLS] outputs of the pre-trained in shallow fusion, a multi-layer perceptron (MLP) is used to combine the whole output of the pre-trained models, whereas in shallow fusion, a multi-layer perceptron (MLP) is used to make predictions from the [CLS] outputs of the pre-trained models. Thus, because of the application of the transformer layers over the entire output, the proposed model is assumed to be capable of determining a method to effectively extract the necessary information from the outputs using the self-attention mechanism.

III. PROPOSED METHOD

In this section, the overall structure of the proposed model and the pre-trained models used are explained.

A. PROPOSED MODEL

The overall architecture of the proposed model is shown in Fig. 1. The proposed model comprises three parts: 1) a pre-trained transformer model for modality A, 2) a pre-trained transformer model for modality B, and 3) the top transformer encoder layers for multimodal fusion. For the pre-trained models, RoBERTa [16], Wav2Vec 2.0 [5], and ViT [4] can be used depending on the modality of the input. Further, RoBERTa and Wav2Vec 2.0 were employed as the two pre-trained models for the later experiment using CMU-MOSI, whereas RoBERTa and ViT were employed as the two pre-trained models for the subsequent experiment using MM-IMDb. Although in principle, the proposed method can work with models other than these pre-trained models, it is validated in this study by using these pre-trained models. For both experiments, transformer encoder layers were used for multimodal fusion. Moreover, because positional embeddings were already added to the pre-trained models, they were not considered in the top transformer layers.

The output hidden states from each pre-trained encoder model, \( o_A \) and \( o_B \), with sizes \( o_A \in \mathbb{R}^{s \times d_{model}} \) and \( o_B \in \mathbb{R}^{s \times d_{model}} \), respectively, were concatenated for use as the input for the top transformer encoder layers as follows:

\[
h_0 = \text{concat}(o_A, o_B),
\]

where \( \text{concat}(a, b) \) represents the concatenation of \( a \) and \( b \), \( s \) is the sequence length, and \( d_{model} \) is the dimension size of the model. In this study, we used a pre-trained model with the same \( d_{model} \). If the \( d_{model} \) is different for each model, linear embedding can be used to make the \( d_{model} \) the same for each modal. Using the concatenated output, \( h_0 \), as the input for the top transformer layers, we calculated the hidden states in each layer, \( h_\ell (\ell \in \{1, 2, 3\}) \), using a multi-head self-attention mechanism, which can be expressed as

\[
\text{MultiHead} (h_\ell) = \text{concat}(m_1, m_2, \ldots, m_h)W^Q_\ell,
\]

where

\[
Q_\ell = K_\ell = V_\ell = h_\ell,
\]

\[
m_i = \text{attention}(Q_\ell W^Q_{\ell,i}, K_\ell W^K_{\ell,i}, V_\ell W^V_{\ell,i}),
\]

\[
\text{attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V.
\]

\( Q, K, \) and \( V \) are the query, key, and value, respectively, and \( d_k \) is the scaling factor that satisfies \( d_k = d_v = d_{model}/h \), where \( d_{model} \) and \( h \) are the model dimensions and number of heads, respectively. \( W^Q_{\ell,i}, W^K_{\ell,i}, \) and \( W^V_{\ell,i} \) are the projection parameters, whose dimensions are the following:

\[
W^Q_{\ell,i} \in \mathbb{R}^{d_{model} \times d_k}, \quad W^K_{\ell,i} \in \mathbb{R}^{d_{model} \times d_k}, \quad W^V_{\ell,i} \in \mathbb{R}^{d_{model} \times d_v},
\]

and \( W^O_{\ell,i} \in \mathbb{R}^{d_{model} \times d_{model}} \). \( \text{softmax}(\cdot) \) denotes the softmax function. Each transformer encoder layer has two sublayers: a multi-head self-attention layer and a fully connected feed-forward network layer \( \text{FeedForward}(\cdot) \). Further, for each sublayer, a residual connection and layer-wise normalization \( \text{LayerNorm}(\cdot) \) were employed. Therefore, the output of each layer can be calculated as

\[
h_{\ell-1}' = \text{LayerNorm} (h_{\ell-1} + \text{MultiHead} (h_{\ell-1})),
\]

\[
h_{\ell} = \text{LayerNorm} (h_{\ell-1}' + \text{FeedForward} (h_{\ell-1}')).
\]

For the purpose of classification, the first token embedding in the last layer \( h_3 \) was used, which corresponded to
the [CLS] token in the output. The hyperparameters used in the experiments are presented in Table 3.

**B. ROBERTA**
RoBERTa [16] is an extension model of BERT that removes the next sentence prediction objective in BERT and pre-trains the model longer with a larger amount of data. The architecture of RoBERTa is similar to that of BERT, except for some minor changes in its key parameters. With only a few changes, RoBERTa has been shown to outperform BERT and XLNet in all nine tasks in GLUE [27], and many other studies have further confirmed its usefulness in other tasks. Thus, RoBERTa was used as the encoder for the text information.

For the experiments using CMU-MOSI, roberta-base and roberta-large were used as the pre-trained model weights for the text encoder. In addition, for the experiments using MM-IMDb, roberta-large was used as the pre-trained model weight for the text encoder.

**C. WAV2VEC 2.0**
Wav2Vec 2.0 [5] is a model pre-trained in a self-supervised manner similar to masked language modeling in BERT. It randomly masks a certain proportion of time steps in the latent feature encoder space and solves a contrastive task [28]. During this pre-training, wav2vec learns speech representation by contrastive task, where the true latent is to be distinguished from its distractors. Following the pre-training phase, the model was fine-tuned on the Librispeech-960 dataset [29] for speech recognition tasks using connectionist temporal classification (CTC) loss [30].

In our experiment using CMU-MOSI, we use facebook/wav2vec2-base-960h and facebook/wav2vec2-large-960h as the pretrained weights for our audio encoder.

**D. VISION TRANSFORMER**
The ViT [4] is a model that facilitates image classification using transformer encoder models with a slight change in its encoder architecture. For a particular image input, the model splits the image into a sequence of fixed-size non-overlapping patches, which are then linearly embedded and used as the input for the transformer layers. The ViT primarily relies on transformer architectures and outperforms models using convolutional architectures. Further, it is successful as an image encoder in other models, such as CLIP [20], a weakly supervised image representation learning model that succeeded in few-shot learning tasks.

Thus, in the experiment using MM-IMDb, google/vit-large-patch16-224 were used as the pre-trained model weights for the image encoder.

---

1. https://huggingface.co/roberta-base
2. https://huggingface.co/roberta-large
3. https://huggingface.co/facebook/wav2vec2-base-960h
4. https://huggingface.co/facebook/wav2vec2-large-960h
5. https://huggingface.co/google/vit-large-patch16-224

**E. TRAINING**
The proposed model can be trained in two ways:

1) **Full Training**: The entire model, including the pre-trained models and the top transformer layers, was optimized simultaneously.

2) **Frozen Pre-trained Weights**: The top transformer layers were optimized using multimodal training data, while the pre-trained model weights were fixed.

Intuitively, it is expected that “full training results in a higher performance.” Further, in the experiments described later, these two learning methods were compared in terms of performance and the amount of training data. In addition, “Full Training w/o Pre-trained Weights,” was compared as well, wherein the learning begins from scratch without using pre-trained weights while maintaining the model structure.

**F. HOW DOES OUR PROPOSED MODEL WORK?**
In this section, the effectiveness of the proposed model for multimodal learning has been discussed considering two perspectives.

The first is the over-fitting perspective. In general, a large model with many parameters, such as a transformer, should be trained on a large amount of data, because if the amount of data is insufficient, a high possibility of overfitting exists. One solution to this problem involves setting good initial values. As the proposed model is a late-fusion-type multimodal learning model, wherein the layers close to the input signal are independent embeddings for each modality and these independent embeddings are achieved using pre-trained models which provide good initial values for training the model.

The other is the perspective of association between different modalities. As mentioned earlier, the proposed method is a multimodal learning method of late fusion type; that is, this model implies that each modality is embedded independently in the first half, with the top transformer learning the associations between them. Therefore, if the pre-trained model has already learned the embedding, the top-level transformer with relatively fewer parameters only needs to learn the correspondence between modalities, implying that it can learn with less training data. Furthermore, because the pre-trained model is considered to represent a certain category in its embedding space, the amount of data required to learn the associations between them is expected to be less than that required to learn the correspondence at the sample level. Finally, fine-tuning optimizes the entire system to be more adaptive to the task.

These properties are very convenient for multimodal learning because collecting data on a single modality is relatively simple with availability of many published pre-trained models. However, collecting synchronized multimodal data or preparing a large amount of training data is both expensive and challenging.
TABLE 1. Dataset details.

| Dataset       | Type       | Train | Validation | Test  | Labels           |
|---------------|------------|-------|------------|-------|------------------|
| CMU-MOSI [13] | Multiclass | 1284  | 229        | 686   | [-3, +3]         |
| MM-IMDb [14]  | Multilabel | 15552 | 2608       | 7799  | 23 classes       |

IV. EXPERIMENTS

The effectiveness of the proposed method was verified by evaluating it on two common benchmarks for the classification tasks. First, the dataset details shown in Table 1 and evaluation settings are explained, and then, the experimental results are presented.

A. DATASET

1) CMU-MOSI

The CMU-MOSI dataset [13] is a human multimodal sentiment analysis dataset comprising 2,199 monologue opinion video clips from YouTube movie reviews. Each clip contains three modalities: acoustic, facial, and text.

The experiments in this study were conducted by employing only acoustic and text information because no single-modality-pre-trained transformer model for facial information was found. In CMU-MOSI, each sample is labeled by human annotators with a sentiment score in the range of [-3, +3]. The objective was to predict the sentiment score of each segment. There were 1,284, 229, and 686 segments in the training, validation, and test sets, respectively. Following [10], [32], 7-class accuracy, 2-class accuracy, mean average error (MAE), and correlation were adopted as the evaluation metrics for the experiments.

2) MM-IMDb

The MM-IMDb dataset [14] consists of 25,959 movie plot outlines and images of movie posters. The objective was to classify the genres of each movie. This was a multi-label classification problem, implying that one movie may belong to multiple genres, as in the example in Table 2. There were 15,552, 2,608, and 7,799 movies in the training, validation, and test sets, respectively. To perform the evaluation, the process outlined in [14], [32], [33] was followed, and the F1 macro, F1 micro, F1 samples, and F1 weighted were adopted as the evaluation metrics.

B. BASELINE METHODS

The effectiveness of the proposed method was verified by comparing it with the latest models on the aforementioned datasets.

- CMU-MOSI
  - MulT
    Multimodal transformer (MulT) [34] is a model designed for unaligned multimodal language sequences, wherein a directional pairwise cross-modal transformers are adopted to merge the multimodal information. The MulT architecture pairs all modalities with cross-modal transformers, followed by a transformer that performs predictions using the fused features.

- ICCN
  Interaction canonical correlation network (ICCN) [35] extracts the interaction features of a CNN in a deep canonical correlation analysis (DCCA)-based network, with the core idea of the model being to learn the hidden correlations between the features extracted from the outer product of text and audio and text and videos.

- MAG-BERT/MAG-XLNet
  MAG-BERT and MAG-XLNet are models with multimodal adaptation gate (MAG) [36] applied to a certain layer of BERT [15] and XLNet [17]. MAG allows BERT and XLNet to have multimodal nonverbal data during fine-tuning. In contrast to the proposed model, wherein each modality is treated equally, MAG-BERT and MAG-XLNet use non-text information as complementary information to text information.

- Shallow-Fusion
  Similar to the proposed model, the shallow fusion of SSL models [10] employs RoBERTa [16] as the language encoder and the BERT-like transformer architecture trained on discretized speech tokens based on VQ-Wav2Vec [6]. However, in contrast to the proposed model, this model only concatenates the [CLS] output from each pre-trained model, instead of considering all hidden states.

- MM-IMDb

- BM-NAS
  Bilevel multimodal neural architecture search (BM-NAS) [33] is a model designed for more generalized and flexible DNNs for multimodal learning, wherein a bilevel searching scheme that learns the unimodal feature selection strategy at the upper level and the multimodal feature fusion strategy at the lower level is adopted.

- SMIL
  Multimodal learning with severely missing modalities (SMIL) [32] is a Bayesian meta-learning-based model proposed to solve the problem of missing modalities in multimodal datasets. This model aims to employ a feature reconstruction network, which generates an approximation of the missing modality, thereby enabling the model to obtain complete data in the latent feature space.

- GMU
  Gated multimodal units (GMUs) [14] are recurrent units designed for multimodal fusion

VOLUME 10, 2022 29825
TABLE 2. Example data for each dataset.

| Dataset   | Image/Audio | Text                                                                 | Label   |
|-----------|-------------|----------------------------------------------------------------------|---------|
| MM-IMDb   |             | Harry Graham leads a double life and is caught out when he and his  | Drama   |
|           |             | wife, Eve, apply to adopt a baby. They live in San Francisco but    | Film-Noir |
| CMU-MOSI  | Sampling    | I like to tease some mystery and great plot what could we been but  | -1.800  |
|           | Rate 16kHz  | than all is explained and dispppare are ten minutes later its just a   |         |
|           |             | clear remake of predator from back in 1997 with the premise being  |         |
|           |             | almost identical                                                    |         |

TABLE 3. Hyperparameters of our model in each experiment. Epoch on the left shows those used in “Full Training” model, epoch in the middle shows the epochs used in “Frozen Pre-trained Weights”, and epoch on the right shows those used in “w/o Pre-trained Weights” models. “Lr,” “Decay,” and “Warmup” represent the learning rate, weight decay, and the warm up steps used in training. “#layers” is the number of layers used in our top transformer layers.

| Dataset (model size) | Metric  | Lr  | Decay | Warmup | #layers | Epochs  |
|----------------------|---------|-----|-------|--------|---------|---------|
| CMU-MOSI (base)      | 7-class | 1e-5 | 0.05  | 100    | 3       | 20/200/200 |
|                      | 2-class | 1e-5 | 0.01  | 100    | 3       | 20/200/200 |
|                      | MAE/Corr| 1e-5 | 0.01  | 100    | 2       | 150/200/200 |
| CMU-MOSI (large)     | 7-class | 1e-5 | 0.01  | 100    | 3       | 20/50/150  |
|                      | 2-class | 1e-5 | 0.01  | 100    | 3       | 20/50/150  |
|                      | MAE/Corr| 1e-5 | 0.01  | 100    | 3       | 50/150/150 |
| MM-IMDb (large)      | Multilabel | 5e-5 | 0.01  | 100    | 3       | 20/50/150  |

that have one modality gate over the other. This model resulted in the proposal of the MM-IMDb dataset [14].

**MMBT**

Multimodal bitransformers (MMBT) [8] use BERT [15] as their primary architecture and ResNet-152 [26] to extract image features. Further, the image features are concatenated with sentence tokens to form an input to the BERT structure. Moreover, this functions on the idea of applying transformer architecture for multimodal fusion, similar to the proposed model.

**C. EXPERIMENTAL SETTINGS**

For both datasets, the architecture shown in Fig. 1 was used. For CMU-MOSI, RoBERTa and Wav2Vec 2.0 were employed as the two pre-trained models, whereas for MM-IMDb, RoBERTa and ViT were used. Hyperparameters used in both datasets are shown in Table 3. The transformer encoder were used as the top transformer layers to fuse the output representations from the two pre-trained models. Further, for both datasets, the model was applied under three different conditions with different amounts of training data (10%, 25%, 50%, 75%, 90%, and 100%) as follows:

1) Full Training:
   - Fine-tuning the whole model including the pre-trained models
2) Frozen Pre-trained Weights:
   - Fine-tuning the top transformer layers with the pre-trained model weights frozen
3) w/o Pre-trained Weights:
   - With same model size but directly fine-tuning the whole model without loading the pre-trained weights

In the experiment using the CMU-MOSI dataset, the model was evaluated using base and large models for the pre-trained transformer models, wherein $d_{model} = 768$ and $1024$ were used for the base and large models, respectively. However, for the experiment using the MM-IMDb dataset, the performance of the proposed model was compared with two other conditions: text-only and image-only. For each model, transformer encoder were added on top of each single-modality-pre-trained model (RoBERTa and ViT) and subsequently the model was fine-tuned to the classification task. However, similar experiments with the CMU-MOSI dataset were not conducted because Wav2Vec 2.0, does not have a [CLS] token, as in RoBERTa and ViT. We only used the large models for the MM-IMDb dataset, because we found that the large models achieved better performance than base models in our first experiment on the CMU-MOSI dataset. We used Hugging Face’s Transformers library [37] to build our models and load the pre-trained weights. We used eight NVIDIA A100 (40GB) GPUs for each experiment.
We compare our models on three different conditions: 1) fine-tuning the whole model including the pretrained models (Full Training), 2) fine-tuning the top transformer layers with the pretrained model weights frozen (Frozen Pretrained Weights), and 3) same model size but directly fine-tuning the whole model without loading the pretrained weights (w/o Pretrained Weights). The purple lines show the previous state-of-the-art model result trained on full training data. For our results in all metrics, please see Tables 6, 7, and 8 in the Appendix.

Table 4. Results for multimodal sentiment analysis on the CMU-MOSI dataset [13]. Performances of the other models are obtained from [32], [34], [35], [38], [10], [36].

| Algorithm       | Acc-7 | Acc-2 | MAE  | Corr |
|-----------------|-------|-------|------|------|
| SMIL [32]       | -     | 71.0  | -    | -    |
| MFM [38]        | 36.2  | 78.1  | 0.951| 0.662|
| MultiF [34]     | 39.1  | 81.1  | 0.889| 0.686|
| ICCN [35]       | 39.01 | 83.07 | 0.862| 0.714|
| MAG-BERT [36]   | -     | 86.1  | 0.712| 0.796|
| MAG-XLNet [36]  | -     | 87.9  | 0.675| 0.821|
| Shallow-Fusion [10] | 48.927| 88.275| 0.577| -    |
| Ours (base)     | 54.956| 87.043| 0.624| 0.836|
| Ours (large)    | **56.268** | **86.890** | **0.580** | **0.868** |

Table 5. Results for multimodal classification task on the MM-IMDb dataset [14]. Performances of the other models are taken from [8], [14], [32], [33].

| Algorithm       | Macro | Micro | Samples | Weighted |
|-----------------|-------|-------|---------|----------|
| SMIL [32]       | -     | 62.0  | 61.7    | -        |
| GMU [14]        | 54.1  | 63.0  | 63.0    | 61.7     |
| BM-NAS [33]     | -     | -     | -       | 62.92    |
| MMBT [8]        | **61.6** | 66.8  | -       | -        |
| Ours(Text-only) | 59.723| 66.829| 67.235  | 66.108   |
| Ours(Image-only)| 34.719| 50.131| 49.157  | 47.058   |
| Ours            | **61.537** | **68.211** | **69.250** | **67.527** |

V. RESULTS AND DISCUSSION
This section presents the results and summarizes the observations. The main results for the CMU-MOSI and MM-IMDb experiments are listed in Tables 4 and 5, respectively.

A. IMPROVEMENT IN PERFORMANCE
Despite its simplicity, the proposed model achieved the state-of-the-art performance in terms of most metrics on both datasets, as shown in Tables 4 and 5. A few other models also performed as well as the proposed model, such as the shallow fusion of SSL models [10] and MMBTs [8], which share similar ideas. For instance, the shallow fusion of SSL
models has two unimodal pre-trained transformer models and concatenates their [CLS] output to make predictions, similar to the proposed model. However, MMBTs use transformer encoder layers to combine concatenated multimodal information, and this is an aspect of the proposed model. The results obtained by using these models [8], [10] reveal that the use of pre-trained transformer models as encoder components for the entire model and the fusion mechanism with the transformer encoder layers are effective in multimodal learning. We consider that the transformer encoder layers were able to successfully aggregate the information from the two pre-trained models by applying the self-attention mechanism. However, we have not conducted comparative experiments on the structure of this part and will investigate this in the future.

B. ROBUSTNESS TO LESS TRAINING DATA

In Fig. 2, the results obtained using the proposed model for three different conditions using different amounts of training data are compared. The purple lines in Fig. 2 illustrate the previous state-of-the-art performances, trained on 100% of the original training data, presented in Tables 4 and 5. The results indicate that the “Full Training” model outperformed those proposed in certain previous works [14], [32]–[34], [36], [38] despite using only 25% to 50% of the original training data in both datasets. This robustness to the reduction of the training data amount is attributed to the use of pre-trained transformer model weights. Because the unimodal pre-trained models are trained on a large dataset for their modality, they provide better embeddings for many words or patterns, including those that never appear in the multimodal training dataset. Moreover, they are expected to have structured knowledge of the modality in the pre-trained model weights. To examine whether the improvement in the performance was not because of the number of parameters or the transformer structure but rather owing to the pre-trained model weights and structured knowledge of the modality, the proposed model was tested with pre-trained weights being frozen (“Frozen Pre-trained Weights”) and without using the pre-trained model weights (“w/o Pre-trained Weights”). The results obtained using “w/o Pre-trained Weights” were inferior to those with the “Full Training” model, regardless of the amount of data, implying that pre-trained model weights were the key to the improvements in the performance. Although the “Frozen Pre-trained Weights” model performed better than the “w/o Pre-trained Weights” model, its performance dropped more significantly when compared with ‘Full Training” and “w/o Pre-trained Weights.” Thus, the entire model, including the pre-trained models, should be trained to improve the performance and render it more robust to the reduction in the training data amount.

C. COMPARISON WITH UNIMODAL MODELS

Table 5 presents the results of our method using only unimodal information. The results indicate that the text-only model performs almost as well as the “Full Training” model, which was trained on information from both modalities. This
FIGURE 6. Top 10 Image Dominant Movies. Examples with F1 Score > 0.8 are selected. Missing labels are shown in red, and the mispredicted labels are shown with a strike-out line.

FIGURE 7. Attention in each head in the most text dominant example in MM-IMDb test dataset. Attention for the [CLS] token in the first layer of the top transformer layers is shown. Purple represents the text tokens and the orange represents the image tokens.

may suggest that the full training model attracts attentions primarily from the text encoder output; however, the results shown in Fig. 3 suggest that the top layer attracts more attention from the image encoder output. The attention is captured from the first layer of the top transformer encoder layers, and for text data $\alpha_{\text{text}}$, it is calculated as

$$\alpha_{\text{text}} = \frac{\sigma_{\text{text attention}}}{\sigma_{\text{text attention}} + \sigma_{\text{image attention}}},$$

where $\sigma_{\text{text attention}}$ and $\sigma_{\text{image attention}}$ represent the summation of attention over each modality and $N_{\text{tokens}}$ and $N_{\text{img tokens}}$ are the number of tokens in each modality. This indicates that the proposed model uses the outputs from both the text and image encoder models to improve the performance.

D. ATTENTION RATIO FOR EACH MODALITY

Fig. 4 shows the average attention ratio for each label in MM-IMDb dataset, which indicates that the modality to be focused on is decided by the model depending on the data.
A clear correlation exists between the class and modality in which the model focuses. Fig. 4 (a) depicts that “Sport,” “Documentary,” and “Biography” movies are classified based on text information rather than image information. In contrast, “Horror,” “Western,” and “Film-Noir” movies are classified based on image information rather than text information. Fig. 4 (b) shows the number of examples that belong to each class in the top 100 text dominant examples (top 100 data with the highest attention ratio for text), whereas Fig. 4 (c) shows the same in the top 100 image dominant examples. Thus, the model is more focused on text data for classes such as “Comedy” and “Romance,” and on image data for classes such as “Thriller,” “Horror,” and “Crime.”

Further, compared with the text dominant examples shown in Fig. 5, the image dominant examples shown in Fig. 6 tend to have more terrifying-looking movie posters. Moreover, the difference in the attention ratio among the labels suggests that the model can decide which modality should be focused on according to the inputs. In shallow fusion models [10], the [CLS] output from each pre-trained model are used, and an MLP layer is directly used for classification. Figures 7 and 8 show the attention in each head for the first [CLS] token in the first layer of the top transformer layers. The attention of the first token for each modality is not the highest, and each head appears to gather different types of information from each modality. This indicates that the model efficiently extracts the necessary information from the entire output for classification tasks.

### E. PRE-TRAINED MODELS FOR OTHER MODALITIES

The proposed model fused multimodal information by simply concatenating the outputs from the pre-trained transformer models; therefore, the number of modalities it can handle depends on the number of existing pre-trained models. For example, pre-trained transformer models exist for human poses [39], biological signals [40], videos [21], [41], [42], and robot dynamics [43]. Consequently, the model can be easily fine-tuned to a multimodal task by combining these pre-trained models and using the transformer layers on top. However, in cases where a pre-trained model for the modality in a multimodal dataset does not exist, a new transformer model can be pre-trained using a large unsupervised dataset of the modality, or a method reported in a recent work [44] can be used, which showed that the pre-trained...
**TABLE 6.** Results for all metrics in our training data reduction experiment on the CMU-MOSI dataset. We compare our base models on three different conditions: 1) fine-tuning the whole model including the pre-trained models (Full Training), 2) fine-tuning the top transformer layers with the pre-trained model weights frozen (Frozen Pre-trained Weights), and 3) using same model size but directly fine-tuning the whole model without loading the pre-trained weights (w/o Pre-trained Weights).

| Condition                     | Data Amount | Acc-7 | Acc-2 | MAE  | Corr |
|-------------------------------|-------------|-------|-------|------|------|
| Full Training                 | 100%        | 54.956| 87.043| 0.624| 0.336|
| 90%                           | 52.770      | 85.823| 0.628 | 0.835|
| 75%                           | 53.499      | 85.976| 0.639 | 0.832|
| 50%                           | 50.729      | 85.518| 0.695 | 0.804|
| 25%                           | 45.335      | 86.433| 0.699 | 0.409|
| 10%                           | 41.983      | 82.774| 0.847 | 0.766|
| Frozen Pre-trained Weights    | 100%        | 44.752| 80.183| 0.890| 0.691|
| 90%                           | 44.752      | 79.573| 0.942 | 0.659|
| 75%                           | 45.918      | 78.384| 0.940 | 0.676|
| 50%                           | 41.983      | 76.372| 0.987 | 0.622|
| 25%                           | 36.443      | 73.171| 1.127 | 0.542|
| 10%                           | 31.195      | 64.787| 1.273 | 0.324|
| w/o Pre-trained Weights       | 100%        | 37.318| 68.598| 1.150| 0.470|
| 90%                           | 36.297      | 68.140| 1.191 | 0.414|
| 75%                           | 35.277      | 65.750| 1.152 | 0.443|
| 50%                           | 34.544      | 65.701| 1.216 | 0.379|
| 25%                           | 32.070      | 64.024| 1.254 | 0.358|
| 10%                           | 31.050      | 58.537| 1.372 | 0.068|

**F. LIMITATIONS OF THE PROPOSED METHOD**

In theory, the proposed model can handle more modalities than those used in our evaluations. However, in practice, it is dependent on the available computational resources. Because the model has multiple pre-trained transformer models, which can result in large size even by using only one of them, the proposed method has a limitation with respect to the model size. In addition to the pre-trained transformer models, a few transformer encoder layers were used for multimodal fusion. In these layers, the self-attention mechanism [2] was used instead of the cross-attention mechanism mentioned in [11] because the application of the former to the concatenated outputs is expected to facilitate the model in extracting the necessary information from both modality outputs. However, the self-attention mechanism requires results at greater computational costs than the cross-attention mechanism [11]. Thus, using the cross-attention or other mechanisms in the latest efficient transformer models can reduce the model size and increase the training speed.

**VI. CONCLUSION**

In this study, an architecture that is heavily reliant on existing pre-trained transformer models was proposed. Unimodal pre-trained transformer models, such as RoBERTa and Vision Transformer, were employed as encoders for each modality, and self-attention transformer layers were used to fuse their output representations. The proposed model aimed to reduce the amount of training data required to achieve state-of-the-art performance in multimodal learning. Further, the model was evaluated on two common benchmarks for multimodal classification. State-of-the-art performance was achieved on both benchmarks, and the robustness of the model to the reduction in the training data amount was demonstrated as the model outperformed the previous state-of-the-art models with only 25% to 50% of the original training dataset. The results revealed that the proposed method is simple and effective for multimodal learning in both text-image and text-audio combinations, thereby indicating that it can be effective for any other combination of modalities.

In the future, the multimodal learning method will be extended to generation tasks using multiple modalities. In addition, as the proposed method has a limitation in terms of the computational costs involved, in future, the use of more efficient transformer models will be considered to reduce the model size and increase the training speed.
APPENDIX
A. RESULT DETAILS
Tables 6, 7, and 8 present the results for all metrics in the training-data reduction experiment conducted. Most of the metrics exhibit patterns similar to the results shown in Fig. 2. The training time for the “Full Training” model with 100% data amount is approximately 40, 120, and 300 minutes for CMU-MOSI (base acc-7), CMU-MOSI (large acc-7), and MM-IMDb, respectively.

REFERENCES
[1] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, “Multimodal machine learning: A survey and taxonomy,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 2, pp. 423–443, Feb. 2019.
[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. U. Kaiser, and I. Polosukhin, “Attention is all you need,” in Proc. Adv. Neural Inf. Process. Syst., 2017, pp. 5998–6008.
[3] S. Schneider, A. Baevski, R. Collobert, and M. Auli, “Wav2vec: Unsupervised pre-training for speech recognition,” in Proc. Interspeech, Sep. 2019, pp. 3465–3469.
[4] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An image is worth 16 × 16 words: Transformers for image recognition at scale,” in Proc. Int. Conf. Learn. Represent. (ICLR), 2021.
[5] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “Wav2vec 2.0: A framework for self-supervised learning of speech representations,” in Proc. Adv. Neural Inf. Process. Syst., 2020, pp. 12449–12460.
[6] A. Baevski, S. Schneider, and M. Auli, “Vq-wav2vec: Self-supervised learning of discrete speech representations,” in Proc. Int. Conf. Learn. Represent. (ICLR), 2020.
[7] K. P. Arety, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, “Multimodal fusion for multimedia applications: A survey,” Multimedia Syst., vol. 16, no. 6, pp. 345–379, Nov. 2010.
[8] D. Kiela, S. Bhooshan, H. Firooz, and D. Testuggine, “Supervised multitask learning of discrete speech representations,” in Proc. Adv. Neural Inf. Process. Syst., 2019, pp. 13–23.
[9] S. Sirivardhana, A. Reis, R. Weerasekera, and S. Nanayakkara, “Jointly fine-tuning, BERT-like self supervised models to improve multimodal speech emotion recognition,” in Proc. Interspeech, 2020, pp. 3755–3759.
[10] H. Tan and M. Bansal, “XLERT: Learning cross-modality encoder representations from transformers,” in Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP), 2019, pp. 5100–5111.
[11] L. Harold Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, “VisualBERT: A simple and performant baseline for vision and language,” 2019, arXiv:1908.03557.
[12] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, “Multimodal sentiment intensity analysis in video: Facial gestures and verbal messages,” IEEE Intell. Syst., vol. 36, no. 6, pp. 82–88, Nov. 2016.
[13] J. E. A. Ovalle, T. Solorio, M. Montes-y-Gómez, and F. A. A. Gonz lez, “Gated multimodal units for information fusion,” in Proc. Int. Conf. Learn. Represent. (ICLR), 2017.
[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in Proc. Hum. Lang. Technol. Conf. North Am. Chapter Assoc. Comput. Linguistics (HLT-NAACL), 2019, pp. 4171–4186.
[15] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “RoBERTa: A robustly optimized BERT pretraining approach,” 2019, arXiv:1907.11692.
[16] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, “XLNet: Generalized autoregressive pretraining for language understanding,” in Proc. Adv. Neural Inf. Process. Syst., 2019, pp. 5753–5765.
[17] I. Beltagy, M. E. Peters, and A. Cohan, “Longformer: The long-document transformer,” 2020, arXiv:2004.05150.
K. Miyazawa et al.: Simple and Effective Multimodal Learning Based on Pre-Trained Transformer Models

[41] H. Akbari, L. Yuan, R. Qian, W.-H. Chuang, S.-F. Chang, Y. Cui, and B. Gong, “VATT: Transformers for multimodal self-supervised learning from raw video, audio and text,” in Proc. Adv. Neural Inf. Process. Syst., 2021.

[42] H. Tan, J. Lei, T. Wolf, and M. Bansal, “VIMPAC: Video pre-training via masked token prediction and contrastive learning,” 2021, arXiv:2106.11250.

[43] D. Xing, J. Li, Y. Yang, and B. Xu, “General robot dynamics learning and Gen2Real,” 2021, arXiv:2104.02402.

[44] K. Lu, A. Grover, P. Abbeel, and I. Mordatch, “Pretrained transformers as universal computation engines,” 2021, arXiv:2103.05247.

KAZUKI MIYAZAWA received the M.E. degree in engineering from The University of Electro-Communications, Tokyo, Japan, in 2019. He is currently pursuing the Ph.D. degree with the Department of System Innovation, Graduate School of Engineering Science, Osaka University, Osaka, Japan. Since 2019, he has been a JSPS Research Fellowship for Young Scientists. His current research interests include multimodal data integration, robot learning, concept formation, natural language processing, and reinforcement learning.

YUTA KYURAGI was born in Kariya, Aichi, Japan, in 1996. He received the B.S. degree in engineering science from Osaka University, Osaka, Japan, in 2020, where he is currently pursuing the M.S. degree in engineering science.

TAKAYUKI NAGAI (Member, IEEE) received the B.E., M.E., and Ph.D. degrees from the Department of Electrical Engineering, Keio University, in 1993, 1995, and 1997, respectively. Since 1998, he has been with The University of Electro-Communications. From 2002 to 2003, he was a Visiting Scholar at the Department of Electrical Computer Engineering, University of California at San Diego, San Diego, CA, USA. Since 2018, he has been a Professor with the Graduate School of Engineering Science, Osaka University. He currently works as a Specially-Appointed Professor at UEC AIX, a Visiting Researcher at the Tamagawa University Brain Science Institute, and a Visiting Researcher at the AIST AIRC. His research interests include intelligent robotics, cognitive developmental robotics, and robot learning. He aims at realizing flexible and general intelligence like human by combining AI and robot technologies. He has received the IROS Best Paper Award Finalist, the Advanced Robotics Best Paper Award, and the JSAI Best Paper Award.

***