IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 13, 2020, accepted September 18, 2020, date of publication September 25, 2020, date of current version October 7, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3026823

Multimodal Emotion Recognition With Transformer-Based Self Supervised Feature Fusion

SHAMANE SIRIWARDHANA[1], THARINDU KALUARACHCHI[1], MARK BILLINGHURST[2], AND SURANGA NANAYAKKARA[1]

[1]Augmented Human Laboratory, Auckland Bioengineering Institute, The University of Auckland, Auckland 1010, New Zealand
[2]Empathic Computing Laboratory, Auckland Bioengineering Institute, The University of Auckland, Auckland 1010, New Zealand

Corresponding author: Shamane Siriwardhana (shamane@ahlab.org)

This work was supported by the Assistive Augmentation Research Grant through the Entrepreneurial Universities (EU) initiative of New Zealand.

ABSTRACT Emotion Recognition is a challenging research area given its complex nature, and humans express emotional cues across various modalities such as language, facial expressions, and speech. Representation and fusion of features are the most crucial tasks in multimodal emotion recognition research. Self Supervised Learning (SSL) has become a prominent and influential research direction in representation learning, where researchers have access to pre-trained SSL models that represent different data modalities. For the first time in the literature, we represent three input modalities of text, audio (speech), and vision with features extracted from independently pre-trained SSL models in this paper. Given the high dimensional nature of SSL features, we introduce a novel Transformers and Attention-based fusion mechanism that can combine multimodal SSL features and achieve state-of-the-art results for the task of multimodal emotion recognition. We benchmark and evaluate our work to show that our model is robust and outperforms the state-of-the-art models on four datasets.

INDEX TERMS Multimodal emotion recognition, self-supervised learning, self-attention, transformer, BERT.

I. INTRODUCTION

Multimodal human Emotion Recognition and Sentiment Analysis is an important aspect of many applications, such as customer service, health-care, and education [1]. Advances in Deep Learning [2] (DL) have improved the multimodal emotion recognition by a substantial margin [3], [4]. Two primary research directions in multimodal emotion recognition are (1) how to represent raw data modalities, and (2) how to fuse such modalities before the prediction layer. A good representation of data should capture emotional cues that can generalize over different speakers, background conditions, and semantic contents. A good fusion mechanism should be able to combine input modalities effectively. When it comes to representing data modalities, earlier approaches employed traditional emotion recognition features, such as Mel-frequency cepstral coefficients (MFCC) features [5], facial muscle movement features [6] and glove embeddings [7]. Rather than using low-level features, recent work has also explored the applicability of transfer learning techniques [8]–[10] to extract features from pre-trained DL models. Such work mainly focuses on extracting features related to facial expressions [11] and speech signals [12] from already trained DL networks based on supervised learning methods. Most of the prior work uses both low-level features and deep features (features extracted from pre-trained DL models) [13], [14], rather than representing all modalities with deep features.

In contrast to previous work, we represent all input modalities (audio, video, and text) with deep features extracted from pre-trained Self Supervised Learning (SSL) – a powerful representation learning technique – models [15]–[17]. Although SSL features give powerful representations of the input modalities, it is an extremely challenging task to fuse them before the final prediction due to the following reasons:

1) High dimensionality of SSL embeddings
2) Longer sequence lengths of SSL features
3) Mismatch between sizes and sequence lengths of SSL features across modalities that have extracted from different SSL models

The associate editor coordinating the review of this manuscript and approving it for publication was Kemal Polat.

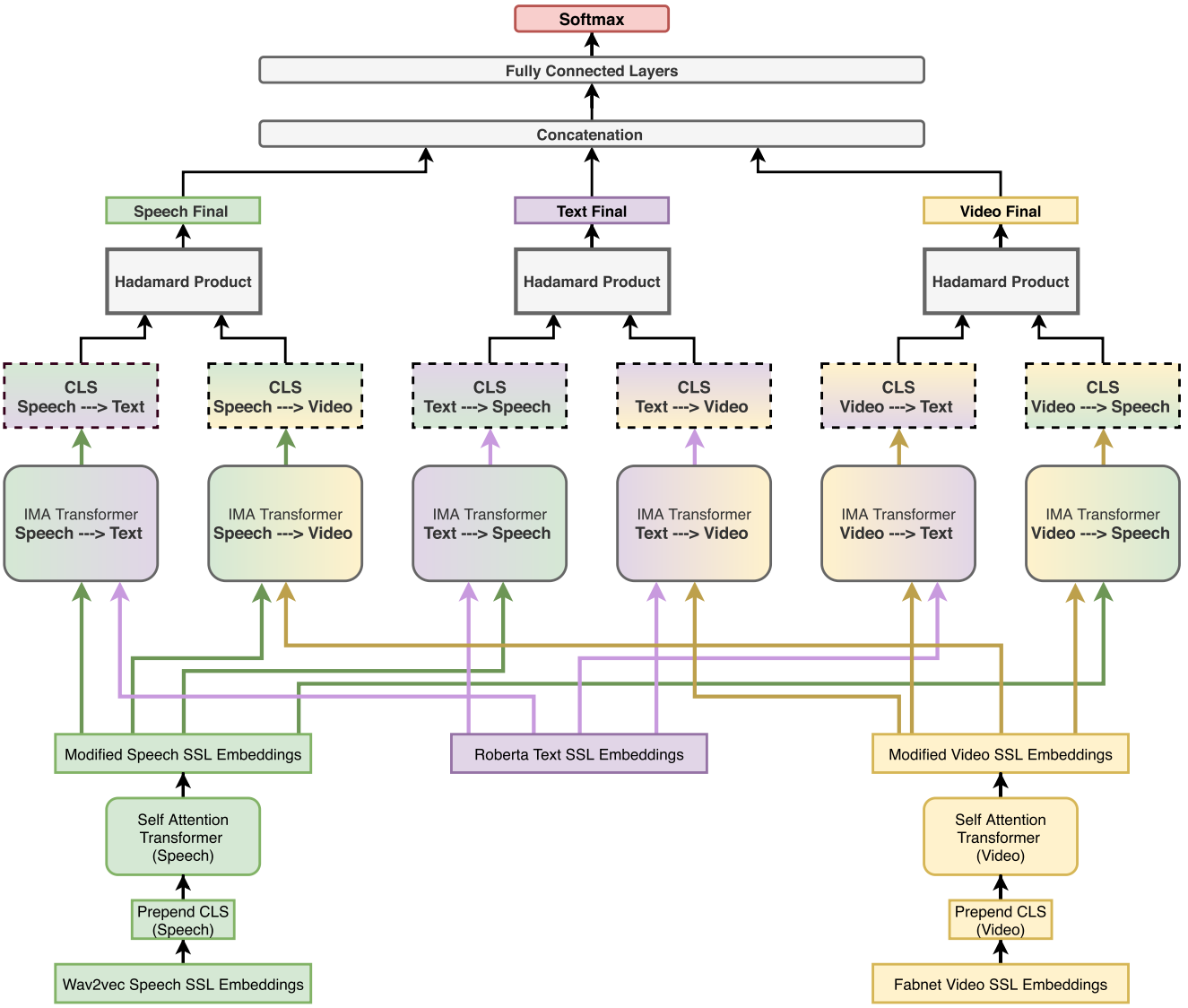

FIGURE 1. **Overview of the Self Supervised Embedding Fusion Transformer (SSE-FT). The proposed fusion mechanism takes SSL embeddings from the three modalities as the inputs. First, speech and video modalities are modified by prepending a unique *CLS* token, followed by a Self-Attention Transformer. These *CLS* tokens for each modality can aggregate the information in the entire sequence (Refer to Section III-B). Then, the information across modalities is extracted with six Inter-Modality-Attention (IMA) based Transformer blocks. Each IMA Transformer takes the *CLS* token from one modality and the entire embedding sequence of other modality as the inputs. For example, the notation *Speech* → *Text* denotes that the *CLS* token comes from the speech modality, while the other embedding sequence comes from the text modality. This can be recognized as attending from speech to text. Each IMA Transformer outputs a *CLS* token enriched with embedding-level information gained from the other modality (Refer to Section III-C). Hadamard product is applied to *CLS* token pairs that belong to the same modality to extract essential features before the concatenation layer (Refer to Section III-D). Finally, the concatenated vectors go through a fully connected layer to produce the output probabilities.**

Although a simple concatenation seems like a viable option, additional trainable parameters required to fully connect the high dimensional SSL embeddings make the network prone to overfitting. Considering these problems, we propose a reliable and effective SSL feature fusion mechanism based on the core concepts of Self-Attention [18] and Transformers [18]–[20]. We used three publicly available pre-trained SSL models of RoBERTa [19] to represent text, Wav2Vec [17] to represent speech and FAb-net [16] to represent facial expressions.

We introduce our novel fusion mechanism as *Self Supervised Embedding Fusion Transformer (SSE-FT)*. As depicted in Figure 1, our framework mainly consists of two Self Attention-based Transformers and six Inter-modality Attention (IMA) based Transformers blocks (refer to Section III). First, two Self Attention-based Transformers modify the speech and video SSL embeddings. This modification step adds a special token named *CLS* to both speech and video sequences that can aggregate the information embedded in the entire sequence. We did not modify the text SSL embedding sequence since they get extracted from the Transformer based

model where the embedding sequence already contains a *CLS* token. Then, all three SSL embedding sequences pass through six IMA based Transformers that enrich each modality's sequence representation with useful information from other modality. In this step, we specifically use *CLS* tokens related to each modality. Finally, we introduce a Hadamard product based computation to compute the most important feature in each modality. In summary, our main contributions are as follows[1]:

- Use of tri-modal SSL features extracted from three independently pre-trained SSL architectures in multimodal emotion recognition.
- Introduction of a novel Transformers based fusion mechanism that fuses SSL features having arbitrary embeddings, sizes, and sequence lengths.
- Evaluation and comparison of robustness and generalizability of our model on four publicly available multimodal datasets.
- Conduct a series of ablation studies to understand the effect of each major component in the architecture.

II. BACKGROUND AND RELATED WORK

In this section, we cover background and closely related work to our research. First, we give a brief introduction to feature extraction mechanisms used in multimodal emotion recognition. Then, we summarize the theory of SSL and explain the three pre-trained SSL models used in this research. Finally, we highlight closely related work on multimodal fusion.

A. FEATURE EXTRACTION MECHANISMS

In multimodal emotion recognition settings, most of the prior work [11] use a mix of low-level and deep features. This section gives an overview of different feature extraction mechanisms used in prior work.

1) LOW-LEVEL FEATURE EXTRACTION MECHANISMS

Usually, multimodal emotion recognition algorithms consist of a feature extraction mechanism and fusion methods [21]. Previous work have discussed several feature extraction mechanisms for commonly used data modalities, such as audio, video, and text. MFCC [22] and COVAREP [5] can be identified as typical speech feature extraction mechanisms. Word-to-Vector methods such as Skip-gram and Glove [7] are usual examples for text features. There are direct tools like FACET [23] for facial feature extraction to understand emotions.

2) DEEP FEATURE EXTRACTION MECHANISMS

Features extracted from pre-trained DL models are known as deep features. Usually, such DL models first get trained with one or more large supervised datasets. The previous work [9] have used pre-trained facial recognition networks to extract facial features. Previous work has also [24], [25] used the pre-trained speech to text models to extract speech features for the sentiment analysis task. Such work highlights that the deep features extracted from input modalities performed better compared to low-level features.

B. MULTIMODAL FEATURES EXTRACTED FROM PRE-TRAINED-SSL ALGORITHMS

Features extracted from pre-trained SSL models are known as Self Supervised Embeddings (SSE). SSL has become a prominent representation learning paradigm in both Natural Language Processing (NLP) and Computer Vision (CV) communities [15], [20], [26]. SSL algorithms have two stages. The first stage is known as pre-training, while the second stage uses pre-trained SSL models to extract features for downstream tasks. The pre-training stage utilizes a given set of pretext tasks and a large number of unlabelled data. Such pretext tasks use regularities and connections in existing data to design a supervisory signal. Tasks like determining image rotation [27], [28], finding missing words in a sentence [19], [20] can be stated as examples for pretext tasks used in NLP and CV domains. The features generated from pre-trained SSL algorithms have problem agnostic qualities since they have not been trained with problem specific manual labels [29].

Recent literature describes how bigger SSL models consist of billions of parameters like GPT-2, GPT-3 [30]–[32] that outperform baseline models in different NLP tasks. However, training such models from scratch is a very computationally expensive task. Therefore, we highlight the importance of exploiting the pre-trained versions of publicly available SSL models as feature extractors for multimodal raw data streams. In our research, we use three publicly available pre-trained SSL models to extract features. Previous work [13], [14] have used SSL features extracted from BERT [20] to represent text modality in multimodal emotion recognition. To the best of our knowledge, this is the first time two or more pre-trained SSL models are used to extract features in multimodal emotion recognition.

C. SUMMARY OF SSL MODELS USED IN MULTIMODAL FEATURE EXTRACTION

We use three pre-trained SSL models in this research. All the model checkpoints were taken from publicly available repositories. We did not fine-tune any SSL model with multimodal emotion recognition datasets. Features for each data modality were extracted from frozen SSL models.

1) RoBERTa

RoBERTa [19] is an extension of the BERT [20] model, which has shown competitive results in GLUE language modelling tasks [33]. The main difference between RoBERTa and BERT is the training mechanism. RoBERTa does not use the next sentence prediction task. We use the pre-trained RoBERTa from the open-sourced fairseq toolkit [34]. The model consists of 355M parameters and has been pre-trained on large English-language text datasets [35]. The network architecture

[1]Pytorch implementation available at https://github.com/shamanez/Self-Supervised-Embedding-Fusion-Transformer

is similar to BERT, which contains a 24-layer transformer encoder. We input tokenized raw text to the model and use the output from the final layer as the feature representation. RoBERTa can handle large tokenized sentences with a maximum length of 512 words where each sentence is mapped to an embedding of 1024 floating points.

2) Wav2Vec

The architecture of the Wav2vec [17] has been developed on layers of temporal convolution, and the pretext task used for Self Supervised Training leverages the concept of Contrastive Predictive Coding [36]. As authors in wav2Vec suggested, the context representation C can be used as an embedding to represent the raw audio waveform. The authors set the size of the embedding to 512 and the maximum audio waveform length to 9.5 seconds. The network consisted of 35M parameters and was pre-trained on 960 hours of audio taken from the Librispeech dataset [37]. The pre-trained model checkpoint was downloaded from the Fairseq repository [34].

3) FABNET

We used the pre-trained Fabnet [16] model to obtain embeddings for each frame in the video that contained the speaker's face. The pretext task of Fabnet is specially designed to encourage the network to learn the facial attributes that encodes the landmarks, pose, and emotions. Given only the embeddings correspond to the source and target frames, the network is asked to map the source frame to the target frame by predicting a flow field between them, thus forcing the network to understand the offset that should occur in the source image pixels to obtain the target image. This proxy task forces the network to distil the information required to compute the flow field (e.g. the head pose and expression) into the source and target embeddings. The source and target frames are taken from the same face-track of a person with the same identity but with different expressions/poses. The network is pre-trained on two large datasets from voxceleb datasets [38]. The embedding dimension is 256. We use this network to obtain the representation for each video frame.

D. MULTIMODAL FEATURE FUSION MECHANISMS

A wide range of previous work uses Convolutional Neural Networks (CNN), and LSTM based DL models as fusion mechanisms [3], [39]. Recent work has explored the effectiveness of novel DL architectures like Transformers [40] and Graph Convolution Nets [41] as fusion methods. When comparing sequential deep learning architectures similar to LSTM and RNN, recent work highlights both computational efficiency and effectiveness of Transformer [18] based methods. In contrast to our work, all these methods work with low-level features. There is prior work using BERT-based [20] SSL features for text, while other modalities are represented with low-level features. These works discuss fusion mechanisms based on RNN and Self Attention mechanisms [13]. To the best of our knowledge, this is the first time that a fusion mechanism is proposed when representing all three modalities with SSL features. Since SSL features have high dimensional embeddings, larger sequence sizes, and different sequence lengths and embedding dimensions in between modalities, we designed a Transformer based fusion mechanism that is both efficient and more accurate than the previous state-of-the-art.

III. METHODOLOGY

In this section, we present each component of our novel fusion mechanism which is **S**elf **S**upervised **E**mbedding **F**usion **T**ransformer (SSE-FT). First, we describe the feature extraction process that uses pretrained SSL models. Next, we explain the core concept behind the modification of speech and video SSL embeddings. After that, we introduce our cross model fusion method that builds upon the idea of Inter-Modality-Attention (IMA). Finally, we explain the Hadarmard computation based feature selection.

A. SELF SUPERVISED EMBEDDING EXTRACTION

As the first step, we extracted features from raw data modalities using three pre-trained SSL models described in Section II-C. As depicted in the Table 1, dimensions and maximum training sequence lengths of SSL features vary across each modality. Both pre-trained models of RoBERTa [19] and Wav2Vec [17] were accessed from Fairseq code-base [34] and used to extract text and speech SSL features. To download the pre-trained Fabnet model and extract features for video modality, we referred to their publication [16]. To extract features from videos, we cropped faces from each video frame using Retina-Face [42] facial recognition model. Then, we sent each frame that consist of a face through the pre-trained Fabnet model to obtain features of the video modality.

TABLE 1. Statistics of the embedding extracted from pre-trained SSL models.

Model Name	Embedding Size	Max Sequence Length
Text - RoBERTa Embeddings	1024	512
Audio - Wave2Vec Embeddings	512	935
Video - Fabnet Embeddings	256	300

B. MODIFICATION OF SSL EMBEDDING

Figure 2 consists of two Transformer blocks that illustrates the process of speech and video embedding sequence modification. Features extracted from SSL models have large embedding size and long sequence length. We wanted to develop a mechanism where a single embedding can represent a long embedding sequence related to a modality. To achieve this, we modify both Wav2Vec embeddings (A) and Fabnet embeddings (V) by prepending a trainable vector named *CLS* and apply Self Attention to each embedding sequence as depicted in Equation 1. The Self Attention mechanism in Equation 2 works similar to the original

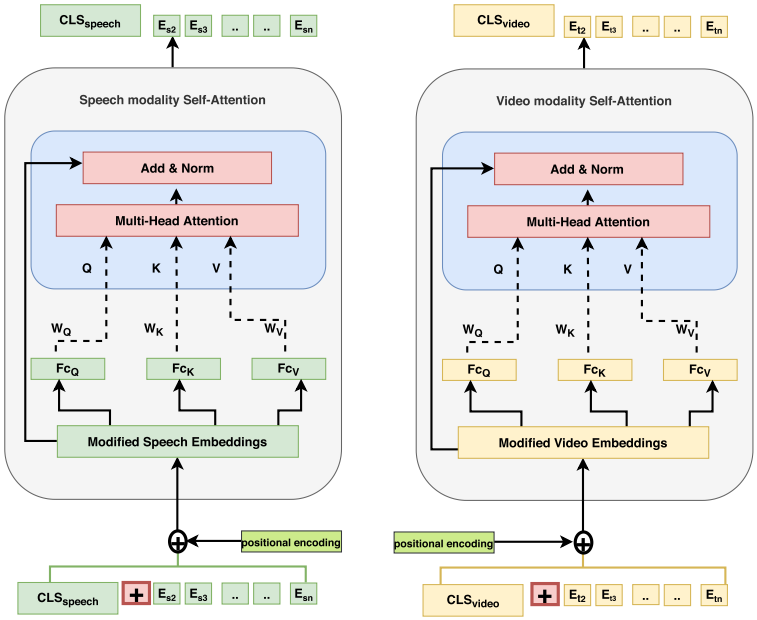

FIGURE 2. Modification of Wav2Vec and Fabnet SSL embeddings with Transformer blocks. First, each embedding sequence for speech and text modalities are modified by prepending unique tokens of CLS_{speech} and CLS_{video}. Then, each of the modified embedding sequences goes though two separate Self Attention-based Transformers blocks.

Transformer mechanism [18]. In Equation 2, symbols Q, K, V, and d_Q refers to Query, Key, Value, and dimensionality of Query vector, respectively.

$$\begin{aligned} &V_{embeddings} \\ &\quad = \text{Self Attention}[prepend\left([CLS]_V, Fabnet_{seq}\right)] \\ &A_{embeddings} \\ &\quad = \text{Self Attention}[prepend\left([CLS]_A, Wav2Vec_{seq}\right)] \quad (1) \\ &\text{Self Attention} \\ &\quad = softmax\left(\frac{QK^T}{\sqrt{d_Q}}\right)V \quad (2) \end{aligned}$$

In our embedding sequence modification phase, we drew inspiration from how BERT [20] or RoBERTa [19] models represent an entire sequence using the first unique token called *CLS* (stands for classification). Since the Self Attention mechanism in BERT-based models is bidirectional (past and future), the *CLS* token, which is the first token of the sequence, is encoded with all information to its right, which is the future sequence. Therefore, the *CLS* token can be used as a compressed representation to solve classification problems like sentiment analysis. In our model, we only prepended *CLS* tokens to Wav2Vec and FabNet embeddings sequences because they do not follow a similar architecture to BERT. Since RoBERTa is a BERT based model, we used the text embedding sequence as it is. Having access to three *CLS* tokens representing three modalities helped us efficiently compute IMA and design a straight forward late fusion mechanism.

C. INTER-MODALITY-ATTENTION (IMA) BASED FUSION LAYER

Figure 3 which consists of six Transformer blocks, illustrates the functionality of the IMA based fusion layer. The primary purpose of the IMA fusion layer is to share relevant information across modalities. The IMA fusion layer was designed to embed the representation of one modality with information gained from representations of other modalities. The IMA layer works similarly to the Self Attention in Equation 2, except it creates the Query (Q) vector from the *CLS* token of one modality and Key (K) -Value (V) vectors from the embedding sequence of the other modality.

The IMA fusion layer's inputs consist of three embedding sequences where the first token of each embeddings sequence is the *CLS* token. Since the *CLS* token of each modality aggregates the sequence's information, the IMA attention is computed between the *CLS* token of one modality and the entire embedding sequence of the other modality. This way, there are six IMA Transformer blocks, where each Transformer block's Q vector is calculated from one modality's *CLS* token, and K-V vectors are calculated from another modality's entire embedding sequence.

D. APPLICATION OF HADARMARD PRODUCT PRIOR TO THE PREDICTION LAYER

As the next step, we explored the possible ways of combining them prior to the prediction layer. The concatenation of six tokens seems to be the obvious way to combine information. However, in our work, we further simplify the *CLS* tokens

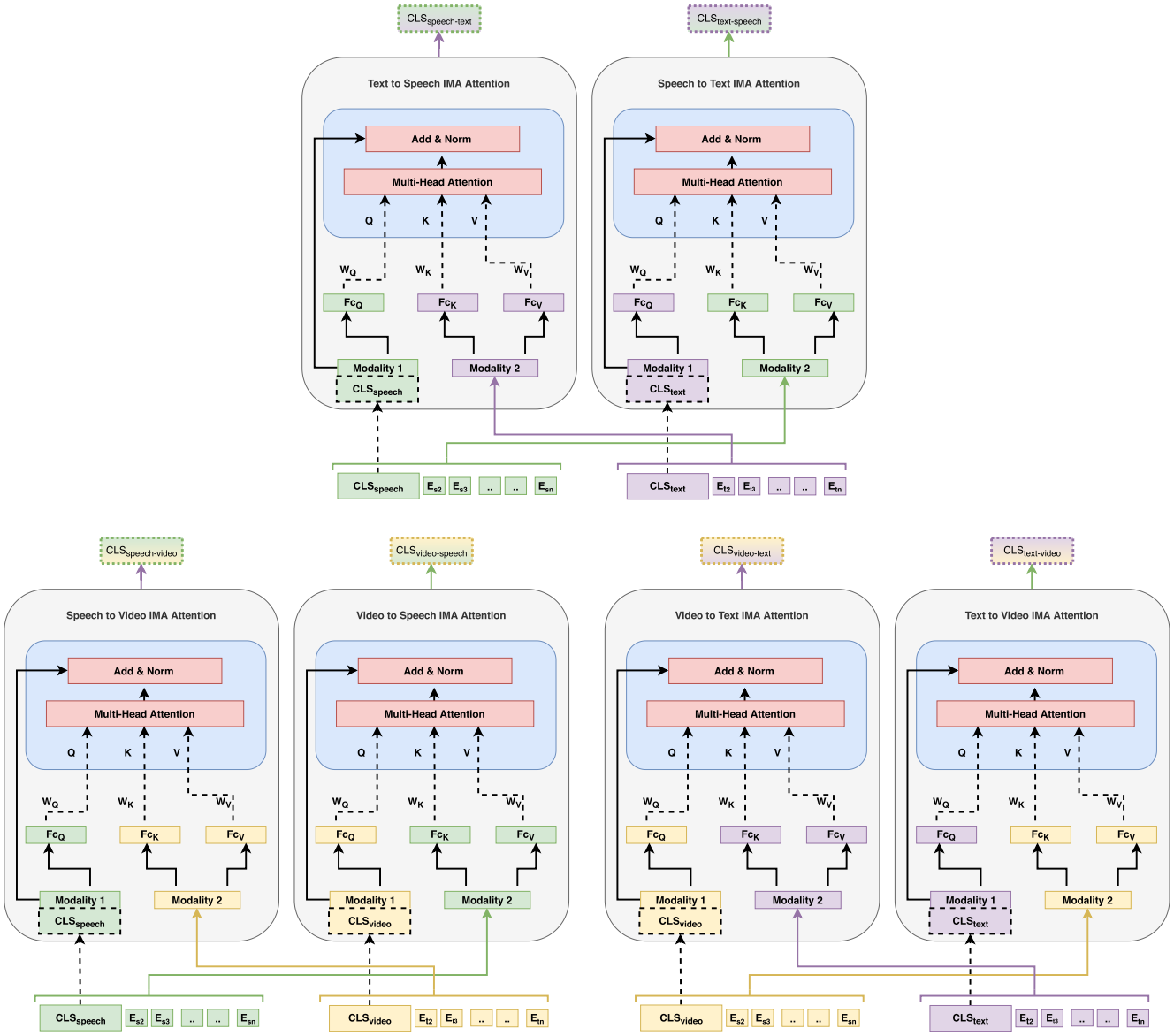

FIGURE 3. Extended view of IMA fusion layer. The fusion mechanism consists of six Inter-Modality-Attention (IMA) based Transformer blocks. Each of the Transformer blocks enables a *CLS* token from one modality to attend to the entire embedding sequence of other modality and gather vital cross-modal information. The fusion process finally outputs six *CLS* tokens that are enriched with inter-modality information.

before concatenation. As illustrated in Figure 1, the six *CLS* embeddings can be grouped into three pairs considering their core-modality (modality of the Q vector in the IMA computation). Finally, to extract the essential information that stems from one modality, we take the Hadamard product ($\odot$) between *CLS* tokens pairs of the same Core-Modality. Equation 3 illustrates the computation of the Hadamard product between six *CLS* tokens that are computed by the IMA layer (Figure 3). Symbols v, a, and t are used to represent video, speech (audio), and text modalities, respectively. v_{final}, a_{final}, and t_{final} are the three resultant vectors after computing the Hadamard product between IMA fusion layer outputs that belong to the same core-modality. Intuitively, the Hadamard product is used to extract the mutual information among the two *CLS* representations. Prior work [43] also has highlighted the effectiveness of using the Hadamard computation to enrich the information in BERT embeddings. Finally, after the Hadamard computation, we concatenate the final three representations and send them through a prediction layer. We compare the use of Hadamard computation based fusion with 6-vector concatenation in our ablation study (see Section VI-E) and empirically show that the proposed method works better.

$$\begin{aligned} v_{final} &= [CLS]_{video \rightarrow speech} \odot [CLS]_{video \rightarrow text} \\ a_{final} &= [CLS]_{speech \rightarrow video} \odot [CLS]_{speech \rightarrow text} \\ t_{final} &= [CLS]_{text \rightarrow video} \odot [CLS]_{text \rightarrow speech} \\ Final_{fusion} &= concatenation\left(v_{final}, a_{final}, t_{final}\right) \end{aligned} \tag{3}$$

TABLE 2. Hyperparameter tuning. Hyperparameters of SSE-FT that we use for the various tasks. The "# of Attention Blocks." and "# IMA Heads" are for each transformer. We performed a basic grid search for hyperparameters, such as the dropout rate, the number of transformer blocks, and attention heads.

Dataset	Batch Size	Initial Learning Rate	Learning Rate Shedular	# Self Attention Blocks	# IMA Blocks	# Self Attention Heads	# IMA Heads	Dropout Rate	Epochs
CMU-MOSEI	32	3.00E-04	Polynomial decay	2	2	4	4	0.1	20
CMU-MOSI	16	3.00E-04	Polynomial decay	1	1	1	1	0.5	20
IEMOCAP	32	3.00E-04	Polynomial decay	2	1	4	4	0.1	20
MELD	32	3.00E-04	Polynomial decay	1	1	2	2	0.1	20

E. SUMMARY OF THE FUSION METHOD

In summary, our fusion method consists of three major parts. First, we modify speech and video embedding sequences by adding two trainable *CLS* embeddings and send them through two different Self Attention based Transformers. After the embedding modification step, we send the embeddings through six Transformer blocks that consist of IMA to capture cross-modality information. Finally, we computed the Hadamard product between *CLS* tokens related to the same modality to enrich the token with the most relevant information. We use only three modalities in our experiments, but this method can be easily extended when using more than three modalities of SSL features.

IV. IMPLEMENTATION DETAILS

This section presents the details of the model implementation and the experimental setup. We implemented our model with *Fairseq*[2] [34] a sequential data processing framework which is built on *Pytorch*[3] DL framework. The training was conducted in distributed GPU settings using two *1080 NVIDIA Titan* GPUs. Details of the final hyperparameters are illustrated in the Table 2.

V. EVALUATION DATASETS

To proceed with our experiments, we used four publicly available datasets. All these datasets consist of speech, text, and video modalities. We compare our proposed method with the state-of-the-art results published for each dataset, as described in the Results Section VI. It is important to note the variations present in the evaluation metrics used by prior work for different datasets. Thus, to have a fair evaluation of our model's performance, we followed the same evaluation procedure many prior works used [3], [4], [4], [40], which showed the state-of-the-art results for each dataset. We used Accuracy, F1-score, Mean Average Error, and Correlation Coefficient as the main evaluation metrics. A summary of the statistics of all datasets used can be found in Table 3. Both IEMOCAP [44] and MELD [3] datasets were annotated with common categorical emotion classes such as Happy, Sad, Angry, Neutral, and Excitement. Both CMU-MOSI [45] and CMU-MOSEI [4] are annotated with sentiment scores that varies between -3 to $+3$. In Section VI, we further explain the datasets and evaluation metrics used in the evaluation process.

[2]https://github.com/pytorch/fairseq
[3]https://pytorch.org

TABLE 3. Summary of the datasets statistics. The number of training, validation, and testing examples in different datasets that we used for conduct experiments.

Dataset	Training	Validation	Testing
CMU-MOSEI - seven class	16326	1871	4659
CMU-MOSI - seven classes	1283	229	686
IEMOCAP - four classes	2717	789	938
MELD - seven classes	9989	1107	2610

VI. RESULTS

In this section, we first explain the experiments conducted to evaluate our model's performance on four datasets and then present the ablation studies conducted to understand the functionality of our proposed model. This work aims to design an effective fusion mechanism when representing all input feature modalities with SSL features. Mainly, we focused on designing a fusion mechanism that can be easily extended with SSL features of several modalities. We also wanted to highlight the effectiveness of Self Supervised features in the task of multimodal emotion recognition. For comparison and evaluation purposes, we mainly used MuLT [40] as the closest multimodal fusion mechanism to our proposed method. Although MuLT [40] uses Transformer based fusion mechanism, their work has not focused on using SSL features, which allows us to highlight the effectiveness of SSL features. Our Transformer based fusion mechanism consists of unique components such as embeddings modification with *CLS* token, modality-specific *CLS* token-based IMA, and Hadamard computation to extract information. These components were explicitly introduced to work with SSL features. At the time of writing, we did not find similar work that specifically focuses on using high dimensional SSL features to represent all modalities.

A. IEMOCAP EXPERIMENTS

The IEMOCAP [44] dataset contains conversation data of 10 male and female actors collected in 5 sessions, with each session consisting of 2 unique actors. The data is segmented by utterances, where each utterance is transcribed and

TABLE 4. Results of multimodal emotion analysis on IEMOCAP with non-aligned multimodal sequences. We report Binary Accuracy (BA) ad F1 score for each emotion. Performances of other models are taken from the MULT [40].

Task	Happy		Sad		Angry		Neutral	
Algorithm Metric	Acc(h)	F1(h)	Acc(h)	F1(h)	Acc(h)	F1(h)	Acc(h)	F1(h)
CTC + RAVEN (Wang et al., 2019)	77	76.8	67.6	65.6	65	64.1	62	59.5
CTC + MCTN (Pham et al., 2019)	80.5	77.5	72	71.4	64.9	65.6	49.4	49.3
MulT (Tsai et al., 2019)	84.8	81.9	77.7	74.1	73.9	70.2	62.5	59.7
SSE-FT (Ours)	**86.5**	**85.7**	**86.7**	**86.2**	**89.4**	**89**	**76**	**75.9**

TABLE 5. Results of multimodal emotion analysis on CMU-MOSEI with non-aligned multimodal sequences. We report Seven Class accuracy, BA (binary accuracy) and F1 score (*for all the scores up to here, the higher the better), MAE (Mean-absolute Error, the lower the better), and Corr (Pearson Correlation Coefficient, the higher the better). Performance's of other models are taken from the MULT [40].

Algorithm Metric	Acc (7 classes - h)	Acc (2 classes - h)	f1 score (h)	MAE (l)	Corr (h)
CTC + RAVEN ((Wang et al., 2019)	45.5	75.4	75.7	0.664	0.599
CTC + MCTN (Pham et al., 2019)	48.2	79.3	79.7	0.631	0.645
MulT (Tsai et al., 2019)	50.7	81.6	81.6	0.591	0.694
SSE-FT (Ours)	**55.7**	**87.3**	**87**	**0.529**	**0.792**

annotated. Labels are chosen from emotion classes of Anger, Happy, Sad, Neutral, Excitement, Frustration, Fear, Surprise, and others. Since the dataset is not uniformly distributed among the classes, we followed prior work [4], [40], and only used the four most frequent labels, which are Happy, Sad, Anger, and Excitement.

To provide a fair evaluation, we followed prior work [4], [40] when designing the final step of the output layer of our model (to compute the binary accuracy for each emotion). For each utterance, the algorithm predicts the availability of each emotion. We split the dataset into training and testing by taking examples from the first four sessions for training and the last session for testing. Therefore, the training and testing datasets have data from 8 and 2 different actors respectively. This way of splitting also allows us to evaluate the algorithm in speaker-independent settings, which is essential for real-world scenarios. Table 4 shows the superior accuracy and the f1 score for each emotion compared to previous work.

B. CMU-MOSEI EXPERIMENTS

For multimodal language analysis that consists of 22000 examples with each having relevant audio (speech), video, and text input streams. The dataset is mainly used to analyze sentiment, and the dataset is created by extracting videos from YouTube, with three people annotating each example to reduce the bias. Unlike annotations in other datasets, which consist of discrete emotions, such as happy and sad, this data set is annotated by assigning a sentiment score to each example that varies from -3 to $+3$, where -3 corresponds to extremely negative sentiment and $+3$ means extremely positive. To evaluate our model with the CMU-MOSEI dataset, we followed the latest prior work [40], which uses seven class accuracy and binary accuracy to evaluate their models. Unlike a usual classification task, here, the model is trained on predicting the sentiment score by minimizing the mean absolute loss (L1 loss). Once the algorithms are trained, the predicted score is rounded off to the nearest integer in the integer set -3 to $+3$, which classifies the data into seven classes. The binary accuracy is then calculated using zero as the threshold for the sentiment score (following the prior work [4], [40], label zero was removed when evaluating the binary accuracy). Similar to previous work [4], [40], we used labels and dataset splits provided in the CMU-SDK [46]. Table 5 shows the performance of our model compared to the state-of-the-art models. As the results suggest, our model outperforms state-of-the art models on each evaluation metric by a fair margin.

C. CMU-MOSI EXPERIMENTS

CMU-MOSI [45] multimodal sentiment analysis dataset is similar to the CMU-MOSEI [4] dataset in all aspects, except for the number of examples. It consists of 2200 examples of YouTube movie reviews. Similar to MOSEI, we used the labels and dataset splits provided in CMU-SDK. Table 6 shows the comparison of the performance of our model with recently published work. Although the CMU-MOSI dataset consists of fewer training examples than other datasets, our model can still outperform prior work by a fair margin.

D. MELD EXPERIMENTS

MELD [3] dataset has more than 12000 utterances from *Friends* television series. In contrast to other datasets, MELD is a conversational dataset that has some examples with several actors in a single utterance. Each utterance is annotated, selecting from one of the seven emotion classes: Anger, Disgust, Sadness, Joy, Surprise, Fear, and Neutral. To give a fair comparison, we provide the seven class accuracy of our model classified using a Softmax layer. Even though we used MuLT [40] as our closest benchmark for all other datasets, for MELD [3] dataset, we could not find the performance

TABLE 6. Results of multimodal emotion analysis on CMU-MOSI with non-aligned multimodal sequences. We report Seven Class accuracy, Binary Accuracy (BA), F1 score, MAE (Mean-absolute Error), and Corr (Pearson Correlation Coefficient). Performances of other models are taken from the MULT [40].

Algorithm Metric	Acc (7 classes - h)	Acc (2 classes - h)	f1 score (h)	MAE (l)	Corr (h)
CTC + RAVEN ((Wang et al., 2019)	45.5	75.4	75.7	0.664	0.599
CTC + MCTN (Pham et al., 2019)	48.2	79.3	79.7	0.631	0.645
MulT (Tsai et al., 2019)	50.7	81.6	81.6	0.591	0.694
SSE-FT (Ours)	**55.7**	**87.3**	**87**	**0.529**	**0.792**

TABLE 7. Results of multimodal emotion analysis on MELD with non-aligned multimodal sequences. We report Seven Class weighted accuracy and the F1 score. Performances of other models are taken from the QIN [47].

Metric	Acc (7 classes)	F1 score
Hierarchical biLSTM (Zhang et al 2019)	60.8	56.3
QIN (Zhang et al 2019)	61.9	57.8
SSE-FT (Ours)	**64.3**	**63.9**

of MuLT [40] to compare due to the unavailability of the low-level features. Table 7 shows the comparison and superiority of our model with other recent evaluations on the MELD dataset for seven class emotion recognition accuracy.

E. ABLATION STUDIES

As depicted in Table 8, we conducted a series of ablation studies using the CMU-MOSEI [4] dataset to understand the influence of different components in the proposed fusion mechanism. We selected CMU-MOSEI because it has the highest number of training examples, compared to other datasets. Mainly there are three types of ablation studies described as follows:

- Ablation study on speech, text, and video input modalities.
- Ablation study on the use of IMA layers (Pre-IMA layer)
- Ablation study on the use of Hadamard product (Post-IMA layer)

1) UNIMODAL INPUT

As illustrated in the first part of the Table 8 (Unimodal Transformers), we checked the impact of each input modality on the final accuracy. In the unimodal experiments, the *CLS* token extracted after the Self Attention Transformer was taken as the final representation (only for speech and video modalities). The unimodal results highlight the significance of text features. Text modality gives 80.2% of binary sentiment accuracy and 47.7% for 7-class sentiment accuracy. The model with only speech modality shows 67.5% and 43.8% for binary and seven-class accuracy. Finally, the model with only video modality shows 66.3% and 43.6% for binary and seven-class accuracies.

It is essential to highlight that the model with only text modality performs significantly better than other modalities. A possible reason could be the power of the RoBERTa embeddings compared to other SSL feature modalities. We compare the results of the unimodal ablation study with recorded MuLT [40] ablation study results that show a similar trend. Since CMU-MOSEI dataset is collected from real-world YouTube review videos examples, most of the sentiment contents can be understood with the text modality.

2) DUAL-MODAL INPUTS

Next, we conducted experiments to understand the model's performances with dual-modal inputs. In this experiment, we used the two *CLS* tokens after the IMA fusion layer as the final representation. As illustrated in Table 8 (dual-modal section), a model that takes text and speech as inputs gives the best results of 54.1% of seven class sentiment accuracy and 86% binary sentiment accuracy. The model that takes speech and video gives the lowest results of 44.18% for seven class sentiment accuracy and 68.2% for binary sentiment accuracy. Dual modality results also highlight the high informative nature of text modality concerning CMU-MOSI dataset.

3) PRE-IMA LAYER

In this study, we compared the model's performance without the IMA block under the tri-modal setting. We mainly wanted to explore the cause of the improvement by the six IMA fusion Transformer blocks. We used three *CLS* tokens extracted from three embedding sequences prior to the IMA layer. The *CLS* tokens for speech and video modality extracted after Self Attention blocks. The *CLS* token for text sequence was directly extracted from RoBERTa embeddings. Finally, we concatenated three vectors before sending them through the prediction layer. This setting achieves an accuracy of 47.5% for seven class sentiment and 81.9% for binary accuracy. When compared with the best performing model, which gives 55.5% accuracy for seven class sentiments classification and 87.3% for binary sentiment classification, thus highlighting the effectiveness of the proposed IMA fusion Transformers.

4) POST-IMA LAYER

In this ablation study, we explored the effectiveness of our proposed Hadarmad product based fusion mechanism that comes after the IMA fusion. In our final model, we extract *CLS* tokens from all six IMA Transformers. Then, we computed the Hadamard product between *CLS* tokens that belong to the same modality. In this experiment, we did not apply

TABLE 8. Evaluation results of the ablation studies. SSE-FT ablation studies on CMU-MOSI dataset in terms of the significance of individual modalities (Unimodal), combinations of two modalities (Dual-modal) pre-IMA fusion mechanism and psot-IMA fusion mechanism. L, A, V denote language (text), audio (speech), and visual. The notation of (h) means higher the better and the notation of (l) means lower the better.

Metric	Acc (7 classes - h)	Acc (2 classes - h)	f1 score (h)	MAE (l)	Corr (h)
Uni-modal Transformers					
Text only	47.7	80.2	80.4	0.636	0.677
Video only	43.6	66.3	65.9	0.729	0.345
Speech only	43.8	67.5	67.8	0.709	0.374
Dual-modal Transformers					
Text and Video	53.9	85.9	85.7	0.543	0.752
Text and Speech	54.1	86	85.8	0.534	0.776
Video and Speech	44.18	68.2	67.8	0.702	0.381
Pre-IMA Fusion Mechanisms					
A+V+T **Three CLS** token concatenation	47.5	81.9	81.8	0.618	0.685
Post--IMA Fusion Mechanisms					
A+V+T **Six CLS** token concatenation	53.3	84.6	84.1	0.567	0.737
Our Final Model (SSE-FT)					
Our Final Model (SSE-FT)	**55.5**	**87.3**	**87**	**0.529**	**0.792**

the Hadarmad product to the *CLS* tokens but concatenated all six *CLS* tokens and sent it to the final prediction layer. For this experiment, the model achieves 53.3% for seven class sentiment accuracy and 84.6% for binary sentiment accuracy. The comparison of results with our final model highlights the effectiveness of the Hadamard product-based information extraction. As the final result suggests, the use of Hadamard computation improves both binary and seven-class sentiment accuracy by nearly 3% while reducing the number of trainable parameters since pure concatenation of all six vectors adds three times more parameters to the final prediction layer.

VII. DISCUSSION

To the best of our knowledge, there is no prior work that uses SSL features to represent all three input modalities. Recent work have only used SSL features to represent text modality with pre-trained BERT features [48]. These work mainly use common DL architectures like CNN and LSTM. For the first time in the literature, we comprehensively explore the representation of all modalities with SSL features, overcoming the challenge of the high dimensional nature of SSL features.

SSL has become a popular research area and already shown improvements in NLP and CV. Since the SSL paradigm can use widely available unlabelled data, an increasing number of pre-trained SSL models for different data streams are being open-sourced to the research community. Usually, these models have different architectures, and they are trained independently with different pretext tasks. Therefore, in this research, we highlight the importance of introducing effective and reliable fusion mechanisms that can be used to fuse multimodal SSL features.

Our proposed fusion mechanism mainly uses two Self-Attention based Transformers and six IMA based Transformers. The fusion mechanism was mainly designed to work with discrete sequences of SSL embeddings while carefully considering differences in SSL embeddings generated from different pre-trained architectures. Due to these reasons, we can easily extend this proposed mechanism with more or new SSL features extracted from different pre-trained models. The experiments conducted with four publicly available datasets highlight the ability of SSL features to provide better results for the task of multimodal emotion recognition. Then, the results for ablation studies show the state-of-the-art performance of our proposed fusion mechanism.

VIII. CONCLUSION AND FUTURE WORK

In this work, we focused on using pre-trained SSL models as feature extractors to improve the task of emotion recognition. To achieve our goal, we designed a transformer-based multimodal fusion mechanism that has the ability to perform well by understanding inter-modality connections. We first evaluated our model with strong baselines from four well-established multimodal affective datasets and demonstrated that our method can outperform previous state-of-the-art methods. Next, we conducted strong ablation studies to understand the important components in our fusion mechanism. It is important to have a stable and well-investigated fusion mechanism when using SSL features as inputs the features generated from SSL techniques are generally high dimensional and can be considered as high-level features. The results suggest that we can effectively use SSL features from different pretrained models to solve the task of multimodal emotion recognition. The use of SSL algorithms allows us to leverage the potential within largely available unsupervised data to tasks like emotion recognition. This method also enables us to use already available pre-trained SSL models that are usually very expensive to train and takes considerable amount of training time, without re-training or training from scratch.

Although we only focused on speech, video, and text in this work, we would like to explore ways to fuse SSL features from other modalities like electroencephalogram (EEG) data [49] in future work. With pre-trained SSL models, we used independently trained models for each modality; however, recent literature shows that certain SSL algorithms can learn joint information between video and text for tasks like video question answering [48]. Therefore, we aim to explore such models to extract features and ways to design SSL models that can learn joint representations between audio (speech), video, and text in future research.

REFERENCES

[1] R. W. Picard, *Affective Computing*. Cambridge, MA, USA: MIT Press, 2000.
[2] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.
[3] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "MELD: A multimodal multi-party dataset for emotion recognition in conversations," 2018, *arXiv:1810.02508*. [Online]. Available: http://arxiv.org/abs/1810.02508
[4] A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, 2018, pp. 2236–2246.
[5] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "COVAREP—A collaborative voice analysis repository for speech technologies," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 960–964.
[6] R. Ekman, *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*. New York, NY, USA: Oxford Univ. Press, 1997.
[7] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.
[8] H.-W. Ng, V. D. Nguyen, V. Vonikakis, and S. Winkler, "Deep learning for emotion recognition on small datasets using transfer learning," in *Proc. ACM Int. Conf. Multimodal Interact. (ICMI)*, 2015, pp. 443–449.
[9] Z. Han, H. Zhao, and R. Wang, "Transfer learning for speech emotion recognition," in *Proc. IEEE IEEE 5th Int. Conf. Big Data Secur. Cloud (BigDataSecurity) Int. Conf. High Perform. Smart Comput., (HPSC) IEEE Int. Conf. Intell. Data Secur. (IDS)*, May 2019, pp. 96–99.
[10] M. Ezzeldin A. ElShaer, S. Wisdom, and T. Mishra, "Transfer learning from sound representations for anger detection in speech," 2019, *arXiv:1902.02120*. [Online]. Available: http://arxiv.org/abs/1902.02120
[11] K. Feng and T. Chaspari, "A review of generalizable transfer learning in automatic emotion recognition," *Frontiers Comput. Sci.*, vol. 2, p. 9, Feb. 2020.
[12] B. Nagarajan and V. R. M. Oruganti, "Deep net features for complex emotion recognition," 2018, *arXiv:1811.00003*. [Online]. Available: http://arxiv.org/abs/1811.00003
[13] N.-H. Ho, H.-J. Yang, S.-H. Kim, and G. Lee, "Multimodal approach of speech emotion recognition using multi-level multi-head fusion attention-based recurrent neural network," *IEEE Access*, vol. 8, pp. 61672–61686, 2020.
[14] Z. Sun, P. Sarma, W. Sethares, and Y. Liang, "Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis," 2019, *arXiv:1911.05544*. [Online]. Available: http://arxiv.org/abs/1911.05544
[15] L. Jing and Y. Tian, "Self-supervised visual feature learning with deep neural networks: A survey," 2019, *arXiv:1902.06162*. [Online]. Available: http://arxiv.org/abs/1902.06162
[16] O. Wiles, A. S. Koepke, and A. Zisserman, "Self-supervised learning of a facial attribute embedding from video," 2018, *arXiv:1808.06882*. [Online]. Available: http://arxiv.org/abs/1808.06882
[17] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "Wav2vec: Unsupervised pre-training for speech recognition," 2019, *arXiv:1904.05862*. [Online]. Available: http://arxiv.org/abs/1904.05862
[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
[19] 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*. [Online]. Available: http://arxiv.org/abs/1907.11692
[20] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*. [Online]. Available: http://arxiv.org/abs/1810.04805
[21] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognit.*, vol. 44, no. 3, pp. 572–587, Mar. 2011.
[22] M. Swain, A. Routray, and P. Kabisatpathy, "Databases, features and classifiers for speech emotion recognition: A review," *Int. J. Speech Technol.*, vol. 21, no. 1, pp. 93–120, Mar. 2018.
[23] S. Stöckli, M. Schulte-Mecklenbeck, S. Borer, and A. C. Samson, "Facial expression analysis with affdex and facet: Avalidation study," *Behav. Res. Methods*, vol. 50, no. 4, pp. 1446–1460, 2018.
[24] Z. Lu, L. Cao, Y. Zhang, C.-C. Chiu, and J. Fan, "Speech sentiment analysis via pre-trained features from end-to-end ASR models," in *Proc. ICASSP-IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 7149–7153.
[25] N. Tits, K. El Haddad, and T. Dutoit, "ASR-based features for emotion recognition: A transfer learning approach," 2018, *arXiv:1805.09197*. [Online]. Available: http://arxiv.org/abs/1805.09197
[26] P. Sermanet, C. Lynch, Y. Chebotar, J. Hsu, E. Jang, S. Schaal, S. Levine, and G. Brain, "Time-contrastive networks: Self-supervised learning from video," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2018, pp. 1134–1141.
[27] C. Doersch, A. Gupta, and A. A. Efros, "Unsupervised visual representation learning by context prediction," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1422–1430.
[28] T. N. Mundhenk, D. Ho, and B. Y. Chen, "Improvements to context based self-supervised learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9339–9348.
[29] A. Kolesnikov, X. Zhai, and L. Beyer, "Revisiting self-supervised visual representation learning," 2019, *arXiv:1901.09005*. [Online]. Available: http://arxiv.org/abs/1901.09005
[30] I. Solaiman, M. Brundage, J. Clark, A. Askell, A. Herbert-Voss, J. Wu, A. Radford, G. Krueger, J. Wook Kim, S. Kreps, M. McCain, A. Newhouse, J. Blazakis, K. McGuffie, and J. Wang, "Release strategies and the social impacts of language models," 2019, *arXiv:1908.09203*. [Online]. Available: http://arxiv.org/abs/1908.09203
[31] L. R. Varshney, N. S. Keskar, and R. Socher, "Pretrained AI models: Performativity, mobility, and change," 2019, *arXiv:1909.03290*. [Online]. Available: http://arxiv.org/abs/1909.03290
[32] T. B. Brown *et al.*, "Language models are few-shot learners," 2020, *arXiv:2005.14165*. [Online]. Available: http://arxiv.org/abs/2005.14165
[33] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "GLUE: A multi-task benchmark and analysis platform for natural language understanding," 2018, *arXiv:1804.07461*. [Online]. Available: http://arxiv.org/abs/1804.07461
[34] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli, "Fairseq: A fast, extensible toolkit for sequence modeling," 2019, *arXiv:1904.01038*. [Online]. Available: http://arxiv.org/abs/1904.01038
[35] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 19–27.
[36] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2018, *arXiv:1807.03748*. [Online]. Available: http://arxiv.org/abs/1807.03748
[37] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 5206–5210.
[38] J. Son Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," 2018, *arXiv:1806.05622*. [Online]. Available: http://arxiv.org/abs/1806.05622
[39] J. K. P. Seng and K. L.-M. Ang, "Multimodal emotion and sentiment modeling from unstructured big data: Challenges, architecture, & techniques," *IEEE Access*, vol. 7, pp. 90982–90998, 2019.

[40] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," 2019, *arXiv:1906.00295*. [Online]. Available: http://arxiv.org/abs/1906.00295

[41] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, and A. Gelbukh, "DialogueGCN: A graph convolutional neural network for emotion recognition in conversation," 2019, *arXiv:1908.11540*. [Online]. Available: http://arxiv.org/abs/1908.11540

[42] J. Deng, J. Guo, Y. Zhou, J. Yu, I. Kotsia, and S. Zafeiriou, "Retinaface: Single-stage dense face localisation in the wild," 2019, *arXiv:1905.00641*. [Online]. Available: https://arxiv.org/abs/1905.00641

[43] O. Sido, "SQuAD: Integrating PCE and Non-PCE approaches," Tech. Rep., 2019. [Online]. Available: https://www.semanticscholar.org/paper/SQuAD%3A-Integrating-PCE-and-Non-PCE-approaches-Sido/52992010675e411808c5eea61826609521904be7?p2df

[44] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, no. 4, p. 335, Dec. 2008.

[45] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "MOSI: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos," 2016, *arXiv:1606.06259*. [Online]. Available: http://arxiv.org/abs/1606.06259

[46] A. Zadeh, P. P. Liang, S. Poria, P. Vij, E. Cambria, and L.-P. Morency, "Multi-attention recurrent network for human communication comprehension," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, p. 5642.

[47] Y. Zhang, Q. Li, D. Song, P. Zhang, and P. Wang, "Quantum-inspired interactive networks for conversational sentiment analysis," Tech. Rep., 2019, pp. 5436–5442. [Online]. Available: https://doi.org/10.24963/ijcai.2019/755, doi: 10.24963/ijcai.2019/755.

[48] J. Lu, D. Batra, D. Parikh, and S. Lee, "ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," 2019, *arXiv:1908.02265*. [Online]. Available: http://arxiv.org/abs/1908.02265

[49] P. Sarkar and A. Etemad, "Self-supervised learning for ECG-based emotion recognition," 2019, *arXiv:1910.07497*. [Online]. Available: http://arxiv.org/abs/1910.07497

[50] B. Knyazev, R. Shvetsov, N. Efremova, and A. Kuharenko, "Convolutional neural networks pretrained on large face recognition datasets for emotion classification from video," 2017, *arXiv:1711.04598*. [Online]. Available: http://arxiv.org/abs/1711.04598

SHAMANE SIRIWARDHANA received the bachelor's degree in electrical and electronic engineering from the University of Peradeniya, Sri Lanka, in 2016, and the M.Eng. degree from The University of Auckland, New Zealand, in 2019, where he is currently pursuing the Ph.D. degree from the Auckland Bioengineering Institute, The University of Auckland. During his masters, he worked in the field of transfer reinforcement learning. His research interests are related to multimodal deep learning, emotion recognition, unsupervised learning, natural language understanding, and human–computer interaction.

THARINDU KALUARACHCHI received the bachelor's degree in electronic and telecommunication engineering from the University of Moratuwa, Sri Lanka, in 2016. He is currently pursuing the Ph.D. degree with the Auckland Bioengineering Institute, The University of Auckland. His research interests are related to human-centered machine learning, emotion recognition, unsupervised learning, and human–computer interaction.

MARK BILLINGHURST received the Ph.D. degree in electrical engineering from the University of Washington under the supervision of Prof. T. Furness III and Prof. L. Shapiro. He is currently working as a Professor and leading the Empathetic Computing Laboratory, Auckland Bioengineering Institute, The University of Auckland. He has coauthored several books and publications in peer-reviewed books, journals, and conference proceedings leading to more than 24,000 citations. He was awarded a Discover Magazine Award in 2001 for creating the Magic Book technology. In 2019, he was awarded with the VGTC Virtual Reality Career Award.

SURANGA NANAYAKKARA received the B.Eng. and Ph.D. degrees from the National University of Singapore, in 2005 and 2010, respectively. Later, he was a Postdoctoral Researcher with the Pattie Maes's Fluid Interfaces Group, MIT Media Lab. In 2011, he founded the Augmented Human Lab to explore ways of creating novel human–computer interfaces as natural extensions of our body, mind, and behaviour. He is currently working as an Associate Professor and leading the Augmented Human Laboratory, Auckland Bioengineering Institute, The University of Auckland. For the totality and breadth of achievements, he has won many awards including young inventor under 35 (TR35 award) in the Asia Pacific region by MIT TechReview, Outstanding Young Persons of Sri Lanka (TOYP), and INK Fellowship 2016.