Circulant-interactive Transformer with Dimension-aware Fusion for Multimodal Sentiment Analysis

Peizhu Gong GONGPEIZHU012@163.COM
Jin Liu (corresponding author) JINLIU@SHMTU.EDU.CN
Xiliang Zhang AZXL1997@163.COM
Xingye Li 1292197258@QQ.COM
Zijun Yu 202130310070@STU.SHMTU.EDU.CN
Shanghai Maritime University

Editors: Emtiyaz Khan and Mehmet Gonen

Abstract

Multimodal sentiment analysis (MSA) is gaining traction as a critical tool for understanding human behavior and enabling a wide range of applications. Since data of different modalities might lie in completely distinct spaces, it is very challenging to perform effective fusion and analysis from asynchronous multimodal streams. Most of the previous works focused on aligned fusion, which is unpractical in real-world scenarios. The recent Multimodal Transformer (MulT) approach attends to model the correlations between elements from different modalities in an unaligned manner. However, it collects temporal information by self-attention transformer which is a sequence model, implying that interactions across distinct time steps are not sufficient. In this paper, we propose the Citculant-interactive Transformer Network with dimension-aware fusion (CITN-DAF), which enables parallel computation of different modalities among different time steps and alleviates inter-modal temporal sensitivity while preserving intra-modal semantic order. By incorporating circulant matrices into the cross-modal attention mechanism, CITN-DAF is aimed to examine all conceivable interactions between vectors of different modalities. In addition, a dimension-aware fusion method is presented, which projects feature representations into different subspaces for an in-depth fusion. We evaluate CITN-DAF on three commonly used sentiment analysis benchmarks including CMU-MOSEI, CMU-MOSI and IEMOCAP. Extensive experimental results reveal that CITN-DAF is superior in cross-modal semantic interactions and outperforms the state-of-the-art multimodal methods.

Keywords: Multimodal sentiment analysis, Transformer, Circulant matrices, Cross-modal attention

1. Introduction

The rapid expansion in the volume of multimedia data has prompted a tremendous need for data management and understanding. Multimodal sentiment analysis (MSA), as a crucial way to determine a human's emotional states towards a certain topic, has gained increasing attention. Early approaches to sentiment analysis are of unimodal nature, with information derived just from one channel, such as images or text (but not both). Recent studies (Koromilas and Giannakopoulos (2021)) have demonstrated that fully exploiting features of different modalities will lead to a further performance gain. Previous researchers

mainly relied on the assumption that multimodal language sequences are already aligned in the resolution of words and considered only short-term multimodal interactions, which is unpractical since the interactions between various modalities are usually more complicated and last for longer than one word. Tsai et al. (2019) proposed the Multimodal Transformer (MulT) approach to fuse crossmodal information from unaligned data sequences. MulT introduces the modality reinforcement unit to reinforce a target modality with information from a source modality by learning the directional pairwise attention between elements across modalities. In their approach, however, the temporal information is collected by self-attention transformer which is a sequence model, implying that fusion among different time steps is not sufficient. By investigating the literature, we find that the performance of MSA models is mainly hindered by three problems:

(i) How to preserve semantic information within each modality. Due to the fact that diverse modalities have different statistical properties and are distributed across distinct feature spaces, extensive studies (Vincent et al. (2010)) design various deep learning methods for monomodal feature representation. For instance, raw images are often processed by spatial hierarchical networks, while raw text is encoded by sequential networks. However, the differences between high-level semantic concepts and low-level values result in semantic gap among intra-modality embeddings. To narrow the gap, self-supervised embedding (SSE) is introduced to represent data of different modalities (Zhai et al. (2019)). Leveraging pre-training on pretext tasks with tremendous amounts of unlabeled data, SSE has supreme generalization capabilities. Nevertheless, inconsistency of representations caused by long-range dependency is always ignored, which is detrimental to semantic correlation preserving.

(ii) How to alleviate temporal interaction sensitivity across diverse modalities. The inputs of MSA are usually composed of multiple sequences that interact in a strictly chronological alignment (Shad Akhtar et al. (2019)). In sequential tasks, recurrent neural networks (RNN) were dominant over the last decades (Liu et al. (2019)). Attention mechanisms (Vaswani et al. (2017)) are always used along with RNN, and Transformer has given a new way of solving sequence problems. However, such a strictly temporally aligned computation is like a glimpse of a video, which leads to poor performance since prior information is overwritten and disturbed by the posterior information. To conquer the problem, memory networks (Li et al. (2017)) construct a repository for rethinking with a complex structure. Cross-modal Transformer (Xi et al. (2020)) achieves relatively satisfactory performance in a more concise manner, but it is still sensitive to temporal order across different modalities.

(iii) How to exploit complementary relations among diverse modalities and eliminate redundancy. It is still common that straightforward fusion methods are employed in MSA, including element-wise product or sum. 1D temporal convolution (Yuan et al. (2021)) with positional embedding is also a popular choice, which extracts the local structure of the input sequences and transforms features of different modalities to the same dimension. A variety of studies (Williams et al. (2018)) have combined the attention mechanism with deep neural networks for feature fusion in recent years. The attention mechanism can not only tell us where to focus, but also improve the modality-specific expression. However, still few studies enable fully exploiting interactions among multimodal features.

To address these challenges, we propose a circulant-interactive Transformer network (CITN-DAF) with dimension-aware fusion to examine all conceivable interactions among diverse modalities while preserving intra-modal semantic representation. Specifically, our proposed model can be divided into three steps: feature extraction, interaction and fusion. Firstly, for problem (i), modality-specific embedding layers with self-attention are introduced to represent feature vectors of image, audio and text sequences. The embedding layers are pre-trained on pretext tasks in advance to provide more valuable feature representations for downstream tasks, while self-attention aids in capturing long-range dependencies. Then, for problem (ii), feature vectors will be passed through proposed circulant-interactive Transformer blocks (CITB) to enrich themselves with useful information from the other modalities. Thanks to circulant matrices and cross-modal attention mechanism (Xu et al. (2020)), CITB can explore all possible interactions between vectors of different modalities. As only regular operations and calculations are involved, CITB avoids increasing parameters or computational costs. Finally, for problem (iii), we introduce a dimension-aware fusion (DAF) module to combine visual, audio and text representations effectively. DAF maps feature representations in three axes: length, width and depth, which allows diverse modalities to be learned in a common space. To evaluate the performance of our proposed model, comprehensive experiments are conducted on three widely used benchmark datasets including IEMOCAP, CMU-MOSEI and CMU-MOSI (Busso et al. (2008); Zadeh et al. (2016); Zadeh and Pu (2018)). The competitive results verify the effectiveness of our approach. The major contributions of our research can be summarized as follows:

- A circulant-interactive Transformer is proposed, which incorporates circulant matrices with cross-modal attention mechanism to explore all possible interactions between vectors of different modalities and reduce temporal interaction sensitivity.
- Modality-specific embedding layers with self-attention are introduced to alleviate the heterogenous discrepancy and preserve semantic correlation.
- A dimension-aware fusion is proposed for in-depth fusion, which enables feature vectors of different modalities can be learned in a comprehensive manner without significant increases in parameters or computational costs.

2. Related Work

2.1. Multimodal Sentiment Analysis

MSA is challenging since different modalities might lie in completely distinct spaces, which is referred as heterogeneity gap. In the early stage, hand-crafted features are designed to bridge this gap in traditional machine learning algorithms, which is time-consuming and labor-intensive. In recent years, deep neural networks (DNN) have emerged as a powerful architecture for capturing the nonlinear distribution of high-dimensional multimedia data in an end-to-end manner. Furthermore, attention mechanisms are widely employed together with DNN to investigate semantic relevance and achieve further performance improvements. Tsai et al. (2019) proposed a Multimodal Transformer (MulT), which attends to interactions between multimodal by repeatedly reinforcing one modality's features with those from the other modalities. Siriwardhana et al. (2020) adopted pre-trained self-supervised

learning models for multimodal feature representation and introduced a transformer-based fusion mechanism to understand inter-modality connections. Hazarika et al. (2020) presented a multimodal affective framework that projects modalities into modality-invariant and modality-specific subspaces and then fuses them to predict emotional states. However, inter-modality correlations across distinct time steps are frequently overlooked. In this paper, we propose CITN-DAF to promote inter-modal semantic interaction and narrow the heterogeneous gap.

2.2. Multimodal Feature Fusion

According to the fusion stage, multimodal sentiment analysis methods can be divided into two categories, feature-level fusion and decision-level fusion. Feature-level fusion (Ben-Ahmed and Huet (2018)) extracts features from various modalities and fuses them by simple accumulation or concatenation. This fusion approach cannot fully explore intra-modality dynamics due to the complexity at the input level. In contrast, decision-level fusion (Gumaei et al. (2022)) refers to combining the results of different classifiers, each trained on separate modalities. However, it fails to model cross-modal interactions for features cannot interact with each other. In addition to the fusion stage, the specific implementation method of the fusion is also a research focus. Common-used multimodal fusion methods including element-wise product or sum may burden the fusion performance. In recent years, a variety of studies have combined the attention mechanism with deep neural networks to ensure the soundness of feature fusion. The attention mechanism can not only tell us where to focus, but also improve the modality-specific expression. Based on the scope of attention, we can divide it into spatial attention (Shi et al. (2021)) and channel attention (Li et al. (2019)). The former is more concerned with the location information embedded in the modalities, while the latter can point to more critical patterns. Although certain research, such as CBAM (Woo et al. (2018)), has combined the advantages of both to obtain good results in computer vision, the integration capability still needs to be improved for multimodal tasks.

3. Method

In this section, we describe our approach for alleviating inter-modal temporal sensitivity while preserving intra-modal semantic order. The CITN-DAF can be segmented into three sub-modules: modality-specific embedding layers (Section 3.2), circulant-interactive Transformer block (Section 3.3) and dimension-aware fusion (Section 3.4). The overall framework is illustrated in Fig. 1.

3.1. Task Setup

The goal of MSA is to judge the emotional state in videos by leveraging multimodal signals. For clarity, we define some notations and describe the MSA task. Three modalities including image, audio, and text are considered, with sequences from each of them denoted $X_V = \{x_p^v\}_{p=1}^{N_v} \in \mathbb{R}^{N_v \times d_v}$, $X_A = \{x_p^a\}_{p=1}^{N_a} \in \mathbb{R}^{N_a \times d_a}$, and $X_T = \{x_p^t\}_{p=1}^{N_t} \in \mathbb{R}^{N_t \times d_t}$, respectively. $N_{(.)}$ and $d_{(.)}$ are used to represent sequence length and feature dimension. Proposed model takes $\{X_i\}_{i \in A,V,T}$ as inputs to predict the affective orientation from either a predefined set of C categories $y \in \mathbb{R}^C$ or as a continuous intensity variable $y \in \mathbb{R}$.

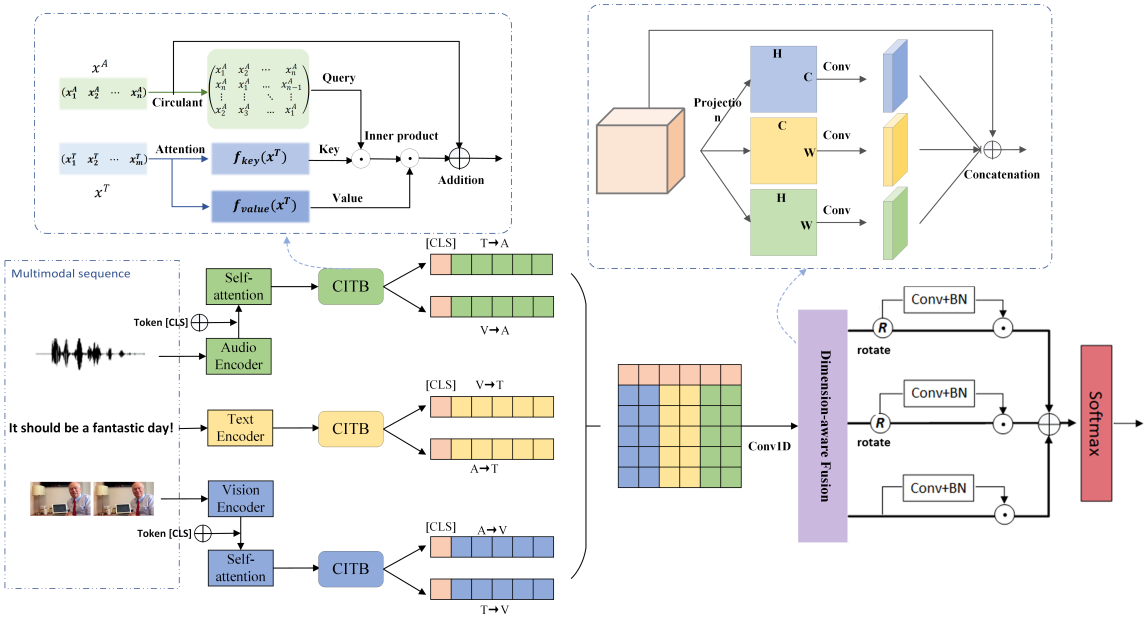

Figure 1: The overall framework of CITN-DAF. CITN-DAF consists of three modules: modality-specific embedding layers with self-attention, circulant-interactive Transformer block (upper left) and dimension-aware fusion (upper right).

3.2. Modality-specific Embedding Layer

Different modalities are characterized by different statistical properties. For example, an image is a three-channel RGB array while text is often symbolic. Therefore, we introduce modality-specific embedding layers to represent feature vectors of different modalities. These embedding layers are trained on pretext tasks in advance to provide more valuable feature representations for downstream tasks. Implementation details are presented in Section 4.4.

For semantic correlation preserving, we introduce a self-attention-based embedding modification module to unify feature representation of different modalities. Inspired by BERT, a special token *CLS* is added to the first element of the sequence to encode the global semantics and obtain bidirectional information. Self-attention aids in capturing long-range correlations within each modality. The details are presented in Algorithm 1.

3.3. Circulant-interactive Transformer

Circulant-interactive Transformer block (CITB) is designed to explore all possible interactions between vectors of different modalities by incorporating circulant matrices with cross-modal attention mechanism. The detailed procedures of CITB are illustrated in upper left of Fig. 1.

Without losing generality, we discuss the CITB for bimodal data. Given two feature vectors in different modalities, e.g., the text vector $O_T \in \mathbb{R}^{(N_t+1)\times d_t}$ and the audio vector

Algorithm 1: Embedding Modification Algorithm

Function Embedding Modification (X_i) *return* O^{seq}, *the modified feature sequence*

Input:

$\{X_i\}_{i\in A,V}$, sequences of audio or video sliced by sentence

Output:

O^{seq}, the modified feature sequence, initially empty

Begin:

Initialize special token $CLS \leftarrow []$

Slice the video and audio into a data sequence $X_i \leftarrow \left[x_1^i, x_2^i, \ldots, x_m^i\right]$

Concatenate the special token CLS with the original data sequence:

$$concate([CLS],\ X_i)$$

Encode sequence position information $Pos = [p_0, p_1, p_2 \ldots, p_n]$

Incorporate position information into data sequence

$$I_{\text{seq}} = Pos \bigoplus concate([CLS],\ X_i)$$

for m iterations do

Set query vector Q, key vector K, value vector V

Calculate the similarity between Q and K

$$Sim_i\left(Q, K_i\right) = \frac{Q \cdot K_i}{\|Q\| \cdot \|K_i\|}$$

Get the final feature vector

$$O^{seq} = \sum_{i=1}^{L_x} softmax \cdot \left(Sim_i\left(Q, K_i\right)\right) V_i = \sum_{i=1}^{L_x} \frac{e^{Sim_i}}{\sum_{j=1}^{L_x} e^{Sim_j}} V_i$$

return O^{seq}

End

$O_A \in \mathbb{R}^{(N_a+1)\times d_a}$.Take text as the target modal, we define the query vectors as $Q_T = O_T W_{Q_T}$, keys as $K_A = O_A W_{K_A}$, and values as $V_A = O_A W_{V_A}$, where $W_{Q_T} \in \mathbb{R}^{d_t \times d}, W_{K_A} \in \mathbb{R}^{d_a \times d}, W_{V_A} \in \mathbb{R}^{d_a \times d}$ are weights. In particular, we use the query vectors to construct circulant matrix. Then the mapping process from A to T can be expressed as $Y_{A\to T} = CITB_{A\to T}\left(O_T, O_A\right)$, the specific process can be seen in the following formula:

$$\begin{aligned} Y_{A\to T} &= CITB_{A\to T}\left(O_T, O_A\right) \\ &= \text{softmax}\left(\text{mat}\, T\left(Q_T\right) K_A\right) V_A \\ &= \text{softmax}\left(\frac{\sum_{i=1}^{N_t+1} Q_{T_i} \odot K_A}{d}\right) V_A \end{aligned} \tag{1}$$

where $Q_{T_i} \in \mathbb{R}^d$ is row vector of circulant matrix $mat\_T\left(Q_T\right)$, $\odot$ denotes dot product. After cross-modal interaction, the mixed feature vectors will go through a multi-layer perceptron with residual connections. In all, the CITB algorithm can be expressed as:

$$\begin{aligned} Y_{A\to T}^{[0]} &= Y_T^{[0]} \\ \tilde{Y}_{A\to T}^{[i]} &= CITB_{V\to T}^{[i],multi}\left(Y_{A\to T}^{[i-1]}, Y_V^{[0]}\right) + Y_{A\to T}^{[i-1]} \\ Y_{A\to T}^{[i]} &= MLP\left(\tilde{Y}_{A\to T}^{[i]}\right) + \tilde{Y}_{A\to T}^{[i]} \end{aligned} \tag{2}$$

where $MLP(.)$ is multi-layer perceptron, and $CITB_{A\to T}^{[i],\ \text{multi}}$ represents CITB with multi-head attention mechanism. In the above process, the circulant matrix shifts elements of each row once. The newly-defined interaction operations enables that all possible correlation across different modalities can be mined.

3.4. Dimension-aware Fusion

Dimension-aware fusion can be divided into three branches, as shown in upper right of Fig. 1. Given a feature $\mu \in R^{C\times H\times W}$ as input, each branch is responsible for capturing the cross-modal interaction between spatial dimension H or W and the channel dimension C. In the first branch, we construct the interaction between the height dimension H and the channel dimension C. Specifically, we perform matrix transposition on , acting on the first and second dimensions, and the new feature tensor after rotation is expressed as $\mu_1 \in R^{W\times H\times C}$. Then μ_1 will go through a special comprehensive pooling layer to squeeze the first dimension of the input tensor. This pooling operation performs average pooling and maximum pooling on the input tensor respectively. Finally, we enhance the fusion of the global feature by a multi-perspective aggregation:

$$Co\text{-}Pool(\mu) = Concate\left[\mu_{\max}; \mu_{\text{avg}}\ ; \mu_{\max} - \mu_{\text{avg}}\ ; \mu_{\max} * \mu_{\text{avg}}\ \right] \tag{3}$$

where

$$\mu_{max} = MaxPool_{1d}(\mu), \mu_{avg} = AvgPool_{1d}(\mu) \tag{4}$$

where $Concate[.;.;.;.]$refers to the concatenation operation. For example, the dimension size of the result obtained by *Co-Pool* (μ_1) is $4\times H\times C$, denoted as $\widetilde{\mu_1}$.Then $\widetilde{\mu_1}$ will go through a convolutional layer with batch normalization to obtain an intermediate result $\widetilde{\mu}_1^*$ whose dimension size is corrected to $1\times H\times C$. After $\widetilde{\mu}_1^*$ passes through the sigmoid activation layer, an attention weight matrix is generated. Multiplying the matrix with μ_1, the interaction information between H and C can be obtained. Finally, to keep the shape consistent with the original input, we transpose the first and second dimensions of the output again. The calculation process of the second branch is similar to that of the first branch. The difference is that we rotate the second and third dimensions of to obtain a new feature tensor $\mu_2 \in R^{H\times C\times W}$. And for the last branch, there is no need to perform a rotation operation. Finally, we perform element-wise sum and average on feature maps from three branches, which can be expressed as:

$$\begin{aligned}
\zeta &= \sum_{i=1}^{n} \left(\mu\sigma\left(\psi_3\left(\widetilde{\mu_3}^*\right)\right) + rotate\left(\widetilde{\mu_1}\sigma\left(\psi_1\left(\widetilde{\mu_1}^*\right)\right)\right)\right. \\
&\quad \left. + rotate\left(\widetilde{\mu_2}\sigma\left(\psi_2\left(\widetilde{\mu_2}^*\right)\right)\right)\right) \\
\cdots &= \sum_{i=1}^{n} \left(rotate\left(\widetilde{\mu_1}\omega_1\right) + rotate\left(\widetilde{\mu_2}\omega_2\right) + \mu\omega_3\right) \\
\cdots &= \sum_{i=1}^{n} \left(\zeta_1 + \zeta_2 + \zeta_3\right)
\end{aligned} \tag{5}$$

where *rotate(.)* represents the rotation operation, σ represents the sigmoid activation function, n is the number of branches and ψ_1, ψ_2 and ψ_3 respectively represent the *2D* convolutional layers of three different convolution kernels.

4. Experimental Setup

4.1. DataSet

IEMOCAP. The IEMOCAP dataset (Busso et al. (2008)) contains five sets of dialogues with ten male and female actors, where each set of dialogues is performed by two regular actors. IEMOCAP is annotated into categorical labels. Since the dataset is unevenly distributed among each category, following other previous studies, we select four of the most common labels, namely Happy, Sad, Anger, and Excitement. We divide the entire dataset into three subsets, using the first four dialogues as training and validation and the last one for testing. Thus, the two actors in the test set are not present in the training set and validation set, which excludes speaker-related interference and is helpful for real scenario applications.

CMU-MOSEI & CMU-MOSI. CMU-MOSEI (Zadeh and Pu (2018)) is a typical multimodal affective computing dataset that contains 22,000 exemplars, each with associated audio, video and text input streams. Unlike the way other discrete datasets are annotated, each example in CMU-MOSEI is assigned an emotion rating (between -3 and +3), with -3 corresponding to extreme negative emotions and +3 representing extreme positive emotions. We directly use the segmentation methods provided in the CMU-SDK to launch our experiments. CMU-MOSI (Zadeh et al. (2016))is similar to CMU-MOSEI in all respects, except for the number of samples.

4.2. Baselines

In order to reflect the effectiveness of CITN-DAF, it is compared with the existing sentiment analysis methods including Early Fusion LSTM (EF-LSTM) (Williams et al. (2018)), Recurrent Attended Variation Embedding Network (RAVEN) (Wang et al. (2019)), Multimodal Cyclic Translation Network (MCTN) (Pham et al. (2019)), Multimodal Transformer (MulT) (Tsai et al. (2019)), Modality-Invariant and -Specific Representations (MISA) (Hazarika et al. (2020)), Self-supervised Multi-Task Learning (Self-MM) (Yu et al. (2021)), Multimodal Adaptation Gate for Bert (MAG-BERT) (Rahman et al. (2020)), Bimodal Information-augmented Multi-head Attention (BIMHA) (Wu et al. (2022)), Transformer-based Feature Reconstruction Network (TFR-Net) (Yuan et al. (2021)), Learning Modality-fused Representations with CB-Transformer (LMR-CBT) (Fu et al. (2021)), Multi-Scale Representation with Shared Vectors of Locally Aggregated Descriptors (ScaleVLAD) (Luo et al. (2021)) and Progressive Modality Reinforcement (PMR) (Lv et al. (2021)). Among them, since EF-LSTM, RAVEN and MCTN rely on the assumption that multimodal language sequences are already aligned, we introduce the connectionist temporal classification (CTC) module (Graves et al. (2006)) to make them applicable to unaligned settings.

4.3. Evaluation Metrics

We evaluate our experimental results in two forms: classification and regression. For classification, we report weighted F1 score (F1-Score) and binary classification accuracy (Acc-2). Following prior works (Hazarika et al. (2020); Tsai et al. (2019)), the Acc-2 and F1-score on MOSEI & MOSI datasets are calculated in two distinct approaches. The first is a negative/non-negative classification (non-exclude zero) and the other is a more accurate

formulation of negative/positive classes (exclude zero). We report results on both these metrics using the segmentation marker -/-, where the left-side score is for neg./non-neg. while the right-side score is for neg./pos. classification. For regression, we report Mean Absolute Error (MAE) and Pearson correlation (Corr). Except for MAE, higher values denote better performance for all metrics.

Table 1: The implementation details of CITN-DAF on CMU-MOSEI, CMU-MOSI and IEMOCAP.

Dataset	**CMU-MOSEI**	**CMU-MOSI**	**IEMOCAP**
Batch size	**32**	**16**	**32**
Learning rate	3×10^{-4}	3×10^{-4}	3×10^{-4}
Optimization	**Adam**	**Adam**	**Adam**
Self-attention blocks	**2**	**1**	**2**
Self-attention heads	**4**	**1**	**4**
CITB	**2**	**1**	**2**
CITB heads	**4**	**1**	**4**
Dropout Rate	**0.1**	**0.5**	**0.1**
Epochs	**20**	**20**	**20**

4.4. Implentation Details

For visual modality, we adopt an ensemble model (Mao et al. (2021)) with a convolution neural network, a CNN-RNN and a CNN-Transformer to incorporate both spatial and temporal information for facial expression recognition. In particular, CNN is a ResNet50, and the depth of GRU is 2. The multi-head attention of Transformer utilizes multiple queries, keys, and values to focus on the most related information at each query. The ensemble is done through a weighted summation among each model.

The wav2vec 2.0 (Baevski et al. (2020)) is used to extract features from original audio clips. The model consists of three sub-modules, feature encoder, transformer, and quantization module. Feature encoder is a multi-layer CNN that represents the input signal as low-level feature vectors. The transformer module is further applied to incorporate the contextual content. The quantization module discretizes the low-level features to a finite speech representation. To train the model, part of the low-level features is masked from the transformer module, and the objective is to identify the quantized version of the masked features based on its context. We take the wav2vec 2.0 model pre-trained on LibriSpeech.

We used ALBERT (Lan et al. (2019)) as our language model for extracting the feature from the text modality. ALBERT is an extension of the BERT, which has established new state-of-the-art (SOTA) results on the GLUE, RACE and SQuAD benchmarks. ALBERT improves parameter efficiency of BERT by incorporating two parameter reduction techniques. The first is decomposing the large vocabulary embedding matrix into small matrices, and the other is sharing cross-layer parameters. In addition, a self-supervised loss for sentence-order prediction is introduced to focus on inter-sentence coherence. We pre-trained ALBERT on large English text datasets including BOOKCOPUS and Wikipedia

for 125k steps. The model uses a 12-layer transformer as encoder with a maximum input length of 512 and an output embedding of 128.

The hyperparameters of CITN-DAF are shown in Table 1. We develop different training schemes according to the amount of data and the number of categories.

Table 2: The comparison experiments on CMU-MOSEI & CMU-MOSI containing word-aligned and unaligned versions. For Acc-2 and F1, the number on the left of / denotes "negative/non-negative" and the right is "negative/positive".KEY - CM: circulant matrix; DAF: dimension-aware fusion.

Models	**MOSEI (aligned)**				**MOSI (aligned)**			
	Acc2(↑)	MAE(↓)	Corr(↑)	F1(↑)	Acc2(↑)	MAE(↓)	Corr(↑)	F1(↑)
EF-LSTM	77.84/80.79	0.601	0.683	78.34/80.67	77.38/78.48	0.949	0.669	77.35/78.51
MCTN	79.8/-	0.609	0.67	80.6/-	79.3/-	0.909	0.677	79.1/-
RAVEN	79.1/-	0.614	0.662	79.5/-	78/-	0.915	0.691	76.6/-
MulT	80.2/-	0.657	0.661	79.8/-	78.7/-	0.964	0.662	78.4/-
MISA	83.6/85.5	0.555	0.756	83.8/85.3	81.8/83.4	0.783	0.761	81.7/83.6
MAG-BERT	83.79/85.23	0.539	0.753	83.74/85.08	**82.54/84.3**	**0.731**	**0.789**	**82.59/84.3**
BIMHA	83.19/83.93	0.562	0.729	83.21/83.64	78.57/80.18	0.929	0.663	78.55/80.23
CITN-DAF	**84.5/86.72**	**0.56**	**0.763**	**84.72/87.45**	**82.73/84.13**	0.791	0.703	82.53/84.3
- w/o CM	81.2/82.53	0.641	0.677	81.36/82.7	79.75/80.53	0.875	0.665	79.35/80.2
- w/o DAF	83.1/84.32	0.583	0.792	83.7/85.15	81.73/82.53	0.805	0.687	81.7/82.64
Models	**MOSEI (unaligned)**				**MOSI (unaligned)**			
	Acc2(↑)	MAE(↓)	Corr(↑)	F1(↑)	Acc2(↑)	MAE(↓)	Corr(↑)	F1(↑)
EF-LSTM+CTC	76.1	0.68	0.585	75.9/-	73.6	1.078	0.542	74.5/-
MCTN+CTC	79.3/-	0.631	0.645	79.7/-	75.9/-	0.991	0.613	76.4/-
RAVEN+CTC	75.4/-	0.664	0.599	75.7/-	72.7/-	1.076	0.544	73.1/-
MulT	81.15/84.63	0.559	0.733	81.56/84.52	79.71/80.98	0.88	0.702	79.63/80.95
MISA	80.67/84.67	0.558	0.752	81.12/84.66	81.84/83.54	0.777	0.778	83.36/85.43
Self-MM	83.76/85.15	0.531	0.765	83.82/84.9	83.44/85.46	0.708	0.796	83.36/85.43
BIMHA	84.07/83.96	0.559	0.731	83.35/83.5	78.57/80.3	0.925	0.671	78.50/80.03
TFR-Net	83.75/-	0.598	0.749	83.58/-	81.73/-	0.754	0.783	81.50/-
ScaleVLAD	84.2/85.73	0.603	0.771	**87.3/89.3**	81.5/83.43	0.827	**0.781**	81.7/83.43
CITN-DAF	**85.91/86.73**	**0.519**	**0.773**	85.53/86.31	**83.73/84.87**	**0.733**	0.778	**83.35/85.28**
- w/o CITB	82.17/84.83	0.557	0.735	83.56/84.36	80.71/81.57	0.873	0.729	81.17/81.95
- w/o DAF	83.39/85.1	0.554	0.754	83.71/84.48	80.97/82.5	0.785	0.766	81.76/83.1

Table 3: The comparison experiments on IEMOCAP dataset (unaligned).

Models	**Happy**		**Sad**		**Angry**		**Neural**	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1
EF-LSTM+CTC	76.2	75.7	70.2	70.5	72.7	67.1	58.1	57.4
RAVEN+CTC	77	76.8	67.6	65.6	65	64.1	62	59.5
MCTN+CTC	80.5	77.5	72	71.7	64.9	65.6	49.4	49.3
MulT	84.8	81.9	77.7	74.1	73.9	70.2	62.5	59.7
PRM	86.4	83.3	78.5	75.3	75	71.3	63.7	60.9
BIMHA	86.3	85.3	**83.6**	82.9	74.2	**74.4**	**70.1**	**71.2**
LMR-CBT	85.7	79.5	79.4	72.6	**76**	70.7	63.6	60.5
CITN-DAF	**86.7**	**85.4**	83.2	**83.6**	**74.4**	73.7	65.7	61.2
- w/o CM	85.2	82.3	78.7	75.3	72.2	71.7	63.5	60.7
- w/o DAF	86.3	84.2	81.32	77.6	73.1	72.9	63.9	61.1

Table 4: The ablation study in modalities on CMU-MOSEI.

	Acc2(↑)	**F1(↑)**	**MAE(↓)**	**Corr(↑)**
Text only	78.15/80.2	77.4/79.62	0.653	0.631
Audio only	63.1/64.75	65.63/67.17	0.764	0.310
Vision only	65.31/66.16	66.47/69.36	0.759	0.343
Text & Vision	80.17/81.63	81.9/82.57	0.543	0.752
Text & Audio	82.15/83.71	82.74/83.63	0.534	0.776
Vision & Audio	67.72/69.33	68.2/69.45	0.702	0.381
V, A→T	80.79/81.57	80.17/82.13	0.605	0.670
T, A→V	79.45/80.37	79.73/80.53	0.611	0.651
T, V→A	79.55/80.73	79.24/80.17	0.620	0.648
CITN-DAF	85.91/86.73	85.53/86.31	0.519	0.773

5. Results and Disscussion

5.1. Quantitative Results

The comparative results for MSA on three widely used benchmarks are presented in Table 2 (MOSEI & MOSI) and Table 3 (IEMOCAP). For a fair comparison, the evaluation on MOSEI & MOSI consists of a word-aligned and an unaligned version. Generally, CITN-DAF achieves better performance than the other baselines in most evaluation metrics, especially on MOSEI dataset. However, the improvement on the MOSI dataset is not so obvious. We infer that this might be caused by the capacity of data, since MOSI is a similar task to MOSEI. We also find that the performance improvement of CITN-DAF in the unaligned version is more significant than that in the word-aligned version, which indicates that our approach is superior in cross-modal semantic interactions across distinct time steps. In addition, to demonstrate the effectiveness of each module in the proposed method, we provide two ablated models for comparison. The experimental results reveal the role of the circular matrix and the DAF, respectively.

The results on the discrete IEMOCAP dataset are reported in Table 3. Compared with the SOTA unaligned methods like BIMHA and LMR-CBT, we can see that CITN-DAF achieves the best performance on both Acc-2 and F1-score of Happy emotion, competitive performance on Angry and Sad emotion, and the worst on Neural emotion. We believe that the unbalanced distribution of data in the dataset is one of the major causes of this occurrence. Additionally, the ablation of CITB and DAF proves their effectiveness in all metrics and datasets.

5.2. Ablation Study

5.2.1. Role of Modality

In order to investigate the influence of different modalities, ablation experiments are conducted on the MOSEI dataset. Our experiments involve the following scenarios: (1) unimodal, where only text, audio or image is considered as input for affective computing; (2) bimodal, where two of the three modalities are selected as input, i.e., text and video, text

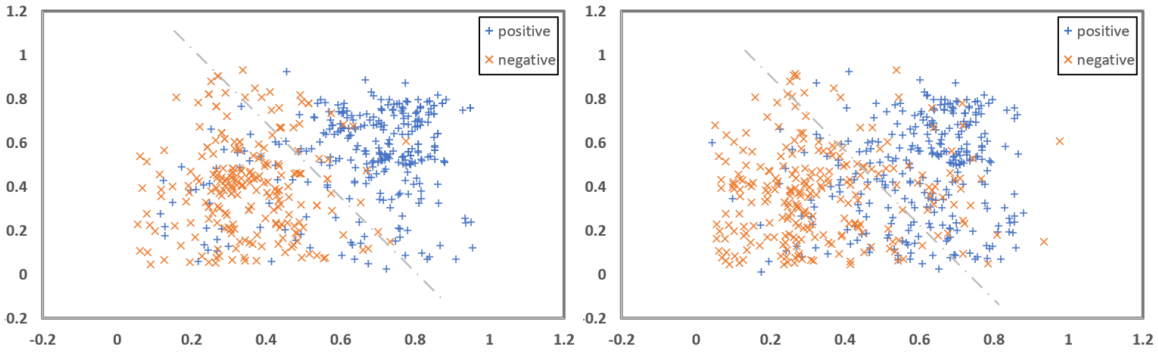

Figure 2: Visualizing joint embedding on CMU-MOSEI. The left is the SSE-based method, and the right is based on default feature extractor.

and audio or video and audio; (3) multimodal, we take different branches for ablation, considering only one modality as the target. The experimental results are shown in Table 4. According to the results, it is believed that the performance of model dramatically drops without text modality, while similar drops do not occur when the other two modalities are removed. We speculate that there are two reasons for this phenomenon. The first is that text contains more semantic information and emotional tendencies. The other is that ALBERT has a stronger representation capability than the other two embedding models.

5.2.2. Visualizing Joint Embedding

We compare SSE-based CITN-DAF(with modality-specific embedding) with that using the default feature extractor of CMU-MOSEI and provide a visualization of joint embedding sentiment distributions. The t-SNE algorithm is used to transform the integrated vectors into a two-dimensional space before to the fusion stage. Fig. 2 shows that the points within respective classes of positive and negative samples are more intensive and the inter-class interval is clearer in the SSE-based CITN-DAF, demonstrating that modality-specific embedding plays an important role in feature extraction.

5.2.3. Visualizing Attention Mechanism

This visualization is conducted to present the effects of attention mechanism over time series. According to the Fig. 3, we can find that cross-modal attention has learned to attend to meaningful signals across various modalities. For example, stronger attention is given to the intersection of words that tend to suggest emotional tendencies (e.g., “dark”, “starring”) and drastic facial expression changes in the video. This observation demonstrates one of the aforementioned advantages of CITN-DAF that cross-modal attention allows CITN-DAF to directly capture potentially long-range signals.

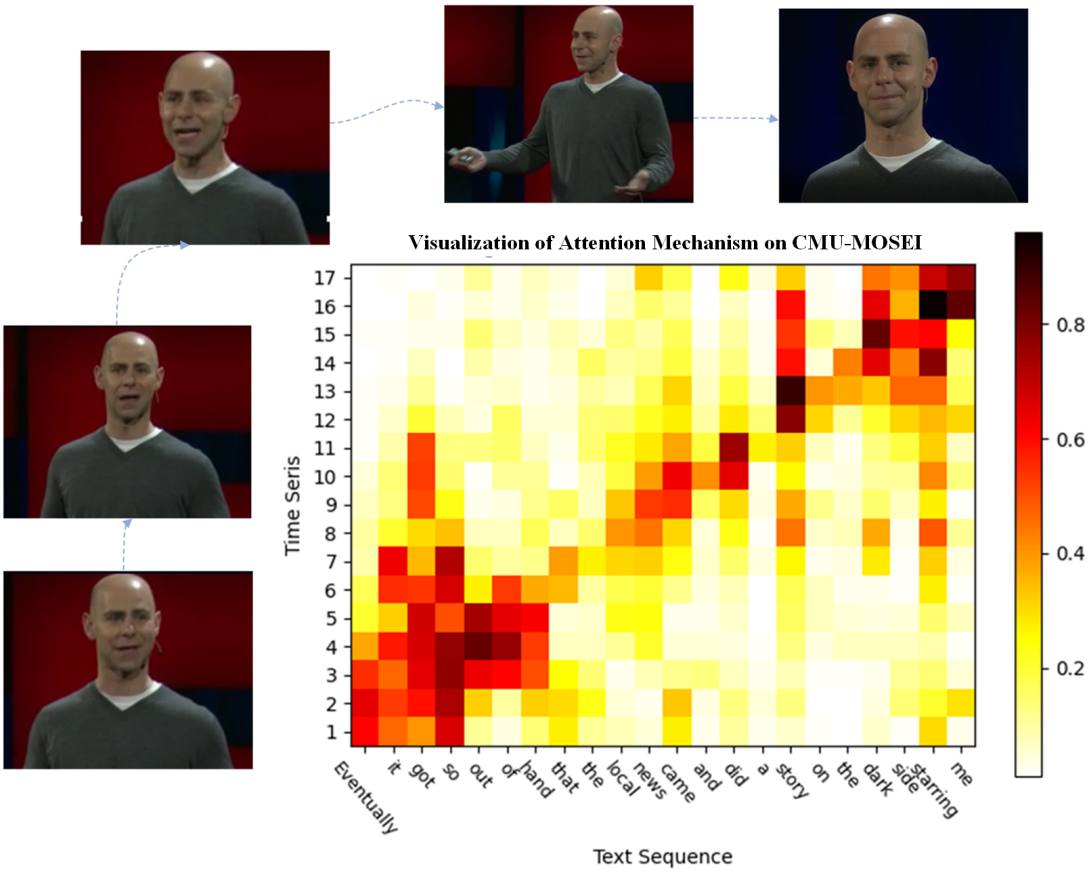

Figure 3: Visualization of attention mechanism over time series. Utterances and keyframes involving changes in expression are presented.

6. Conclusion

In this paper, we propose a circulant-interactive Transformer network for multimodal sentiment analysis. At the heart of the model is to alleviate temporal sensitivity across diverse modalities while preserving semantic information within each modality. By incorporating circulant matrices and cross-modal attention mechanism, our model can explore all possible interactions between vectors of different modalities. Moreover, we introduce a dimension-aware fusion module to project integrated representations into different subspaces for a holistic view. Comprehensive experiments on four widely used benchmarks indicate that our model is superior in cross-modal interactions and achieved comparable or better results compared to the existing state-of-the-art methods.

We also find that current model performance is limited when faced with multiple classifications or insufficient datasets. In future work, we plan to explore one-shot learning methods and factorized representations.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (No. 2021YFC2801000), the National Natural Science Foundation of China (No. 61872231), and the Major Research plan of the National Social Science Foundation of China (No. 2000ZD130).

References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460, 2020.

Olfa Ben-Ahmed and Benoit Huet. Deep multimodal features for movie genre and interestingness prediction. In *2018 international conference on content-based multimedia indexing (CBMI)*, pages 1–6. IEEE, 2018.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4): 335–359, 2008.

Ziwang Fu, Feng Liu, Hanyang Wang, Siyuan Shen, Jiahao Zhang, Jiayin Qi, Xiangling Fu, and Aimin Zhou. Lmr-cbt: Learning modality-fused representations with cb-transformer for multimodal emotion recognition from unaligned multimodal sequences. *arXiv preprint arXiv:2112.01697*, 2021.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376, 2006.

Abdu Gumaei, Walaa N Ismail, Md Rafiul Hassan, Mohammad Mehedi Hassan, Ebtsam Mohamed, Abdullah Alelaiwi, and Giancarlo Fortino. A decision-level fusion method for covid-19 patient health prediction. *Big Data Research*, 27:100287, 2022.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM international conference on multimedia*, pages 1122–1131, 2020.

Panagiotis Koromilas and Theodoros Giannakopoulos. Deep multimodal emotion recognition on human speech: A review. *Applied Sciences*, 11(17):7962, 2021.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.

Xiang Li, Wenhai Wang, Xiaolin Hu, and Jian Yang. Selective kernel networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 510–519, 2019.

Zheng Li, Yun Zhang, Ying Wei, Yuxiang Wu, and Qiang Yang. End-to-end adversarial memory network for cross-domain sentiment classification. In *IJCAI*, pages 2237–2243, 2017.

Jin Liu, Yihe Yang, Shiqi Lv, Jin Wang, and Hui Chen. Attention-based bigru-cnn for chinese question classification. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–12, 2019.

Huaishao Luo, Lei Ji, Yanyong Huang, Bin Wang, Shenggong Ji, and Tianrui Li. Scalevlad: Improving multimodal sentiment analysis via multi-scale fusion of locally descriptors. *arXiv preprint arXiv:2112.01368*, 2021.

Fengmao Lv, Xiang Chen, Yanyong Huang, Lixin Duan, and Guosheng Lin. Progressive modality reinforcement for human multimodal emotion recognition from unaligned multimodal sequences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2554–2562, 2021.

Shuyi Mao, Xinqi Fan, and Xiaojiang Peng. Spatial and temporal networks for facial expression recognition in the wild videos. *arXiv preprint arXiv:2107.05160*, 2021.

Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabás Póczos. Found in translation: Learning robust joint representations by cyclic translations between modalities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6892–6899, 2019.

Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, Amir Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. Integrating multimodal information in large pretrained transformers. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2020, page 2359. NIH Public Access, 2020.

Md Shad Akhtar, Dushyant Singh Chauhan, Deepanway Ghosal, Soujanya Poria, Asif Ekbal, and Pushpak Bhattacharyya. Multi-task learning for multi-modal emotion recognition and sentiment analysis. *arXiv e-prints*, pages arXiv–1905, 2019.

Wenling Shi, Huiqian Du, Wenbo Mei, and Zhifeng Ma. (sarn) spatial-wise attention residual network for image super-resolution. *The Visual Computer*, 37(6):1569–1580, 2021.

Shamane Siriwardhana, Tharindu Kaluarachchi, Mark Billinghurst, and Suranga Nanayakkara. Multimodal emotion recognition with transformer-based self supervised feature fusion. *IEEE Access*, 8:176274–176285, 2020.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2019, page 6558. NIH Public Access, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, and Léon Bottou. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11 (12), 2010.

Yansen Wang, Ying Shen, Zhun Liu, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7216–7223, 2019.

Jennifer Williams, Steven Kleinegesse, Ramona Comanescu, and Oana Radu. Recognizing emotions in video using multimodal dnn feature fusion. In *Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML)*, pages 11–19, 2018.

Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

Ting Wu, Junjie Peng, Wenqiang Zhang, Huiran Zhang, Shuhua Tan, Fen Yi, Chuanshuai Ma, and Yansong Huang. Video sentiment analysis with bimodal information-augmented multi-head attention. *Knowledge-Based Systems*, 235:107676, 2022.

Chen Xi, Guanming Lu, and Jingjie Yan. Multimodal sentiment analysis based on multi-head attention mechanism. In *Proceedings of the 4th international conference on machine learning and soft computing*, pages 34–39, 2020.

Xing Xu, Tan Wang, Yang Yang, Lin Zuo, Fumin Shen, and Heng Tao Shen. Cross-modal attention with semantic consistence for image–text matching. *IEEE transactions on neural networks and learning systems*, 31(12):5412–5425, 2020.

Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 10790–10797, 2021.

Ziqi Yuan, Wei Li, Hua Xu, and Wenmeng Yu. Transformer-based feature reconstruction network for robust multimodal sentiment analysis. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4400–4407, 2021.

Amir Zadeh and Paul Pu. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th annual meeting of the association for computational linguistics (Long Papers)*, 2018.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*, 2016.

Xiaohua Zhai, Avital Oliver, Alexander Kolesnikov, and Lucas Beyer. S4l: Self-supervised semi-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1476–1485, 2019.