Exploiting Semantic Embedding and Visual Feature for Facial Action Unit Detection

Huiyuan Yang[1], Lijun Yin[1], Yi Zhou[2], Jiuxiang Gu[3]
[1] Department of Computer Science, Binghamton University, Binghamton, NY, USA
[2] IBM, Singapore, [3] Adobe Research, USA
hyang51@binghamton.edu,lijun@cs.binghamton.edu,joannezhouyi@hotmail.com, jigu@adobe.com

Abstract

*Recent study on detecting facial action units (AU) has utilized auxiliary information (i.e., facial landmarks, relationship among AUs and expressions, web facial images, etc.), in order to improve the AU detection performance. As of now, no semantic information of AUs has yet been explored for such a task. As a matter of fact, AU semantic descriptions provide much more information than the binary AU labels alone, thus we propose to exploit the **S**emantic **E**mbedding and **V**isual feature (**SEV-Net**) for AU detection. More specifically, AU semantic embeddings are obtained through both Intra-AU and Inter-AU attention modules, where the Intra-AU attention module captures the relation among words within each sentence that describes individual AU, and the Inter-AU attention module focuses on the relation among those sentences. The learned AU semantic embeddings are then used as guidance for the generation of attention maps through a cross-modality attention network. The generated cross-modality attention maps are further used as weights for the aggregated feature. Our proposed method is unique in that the semantic features are exploited as the first of this kind. The approach has been evaluated on three public AU-coded facial expression databases, and has achieved a superior performance than the state-of-the-art peer methods.*

1. Introduction

Facial action units (AUs) defined in the Facial Action Coding System (FACS)[5] has been widely used for describing and measuring facial behavior. Automatic action unit detection has been an essential task for facial analysis, with a variety of applications in psychological and behavioral research, mental health assessment and human-robot interaction.

Benefited from the great progress in deep learning research, the performance of AU detection has been improved using the deep-model based methods in recent years [28][30][12][3][19][22][15]. However, the deep-model based methods are starved for labeled data, whereas AU annotation is a highly labor intensive and time consuming process, thus many existing works seek to exploit the auxiliary information for AU detection, which include, for example, *domain knowledge* (*e.g.*, probabilistic dependencies between expressions and AUs as well as dependencies among AUs) [17][18][26]; *facial landmarks and expression labels* [13][15], and *freely web face images* [29]. Although the performance has a certain improvement when utilizing those auxiliary information, the AU semantic descriptions have not yet been explored by any of the previous methods.

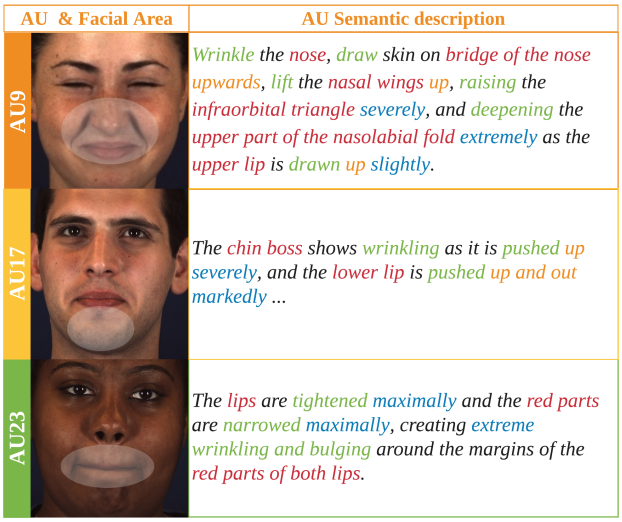

Figure 1. An example of the individual AUs, related facial areas and the corresponding AU semantic descriptions. Red: facial area/position, Green: action, Yellow: motion direction, and Blue: motion intensity. As we can see, the AU related facial areas and their actions are clearly explained in each AU semantic description. The *facial area/position*, *action*, *motion direction and intensity*, and *relation of AUs* will be automatically encoded in the AU semantic embedding, as described in Section 3.2.

FACS provides a complete set of textual descriptions for AU definition, such a set of AU descriptions provide rich semantic information, such as *which facial area is related to the individual AU, what intensity and type of action can be considered as the occurrence of an AU, and what is the relation among AUs, etc.* Such a unique finding motivates us to explore the textual descriptions as an auxiliary information along with the visual information for AU detection. Figure 1 illustrates an example of AU semantic descriptions on three AUs. As we can find that two facial areas (*chin boss* and *lower lip*) and two corresponding actions (*wrinkle chin boss* and *push up the lower lip*) involve in the occurrence of AU17. Besides, we can also obtain the potential relation among AUs. For example, AU23 occurs in the area of lips, which share the *lower lip* with AU17, but they rarely appear together as different actions applied to the lip (*tighten* vs *push up*). A similar example is AU12 (*lip corner puller*) vs AU15 (*lip corner depressor*). Those semantic information (i.e., *facial area/position*, *action*, *motion direction/intensity*, and *relation of AUs*) will be automatically encoded in the AU semantic embedding, which will be described in Section 3.2.

Two recent works [2][24] have been developed to explicitly model the label relationships from the semantic label embeddings using a graph convolutional network (GCN) based method for multi-label image recognition. These two works have demonstrated that explicitly modeling the label relationships from the label embeddings is beneficial for the discovery of meaningful locations and discriminative features. However, both of them rely on the manually defined label relation graph, as used in the GCN module, making them incapable of applications without the ground truth label relation graph.

Inspired by the self-attention mechanism from transformer [20] and Inter/Intra attention modules in [7], we propose a novel framework to exploit the **S**emantic **E**mbedding and **V**isual feature (**SEV-Net**) for AU detection, which will automatically learn the AU relations from the AU semantic descriptions. *First of all*, in order to capture the semantic relations among AUs, we introduce two new attention modules, which are so-called Intra-AU and Inter-AU attention module, where the Intra-AU attention module targets at the word-level attention among the AU semantic descriptions (i.e., $<$*lip corner*$>$ $-<$*raised*$>$), while the Inter-AU attention module focuses on the relation among sentences (i.e., *both AU12 and AU15 occur at the lip corner, but they cannot happen concurrently because opposite actions are associated with the corresponding AUs (puller vs depressor)*). *Second*, the learned AU semantic embeddings are further combined with the visual features to generate the attention map through a cross-modality attention module. Unlike the traditional self-attention methods, the cross-modality attention module benefits from the rich semantic information (i.e.,*facial area/position*, *action*, *motion direction/intensity*, and *relation of AUs*), hence being able to learn more useful and discriminative features from more meaningful facial areas. The attention maps are further utilized as weights for the aggregated feature for AU classification.

In summary, the contributions of this work are two-fold:

1. We proposed a unified framework that applying the attention into three different levels to capture different AU semantic relations: Intra-AU attention (*Words level: location, action type/intensity, etc*), Inter-AU attention (*Sentence level: AU relations, can two AUs happen concurrently?*) and cross-modality attention (*Modality level: connecting the AU semantic embedding to visual features*). As a result, the model is able to learn more discriminative features from more meaningful areas.

2. To the best of our knowledge, this is the first work to introduce AU semantic description as an auxiliary information for AU detection, achieving significant improvement for AU detection than SOTA in three widely used datasets.

2. Related works

AU detection with auxiliary information Current works on facial action (AU) recognition typically rely on fully AU-annotated training data. However, as compared to the other computer vision tasks, the publicly available AU-labeled datasets are quite small due to the labor-intensive work on AU annotation. Therefore, the research community started to utilize the auxiliary information for robust AU detection. Zhao et al.[29] proposed a weakly spectral clustering approach to use freely downloaded web images for learning action units. An embedding space is learned by exploiting web images with inaccurate annotations, and then a rank-order clustering method is applied to re-annotate these images for training AU classifiers. Peng and Wang[17] utilized the domain knowledge for AU detection. Here, the domain knowledge refers to the probabilistic dependencies between expressions and AUs as well as dependencies among AUs. To train a model from partially AU-labeled and fully expression labeled facial images, Peng and Wang[18] used the dual learning method to model the dependencies between AUs and expressions for AU detection. By leveraging prior expression-independent and expression-dependent probabilities on AUs, Zhang et al.[26] proposed a knowledge-driven method for jointly learning multiple AU classifiers without AU annotations, and achieved a good performance on five benchmark databases. Li et al.[11] designed an attention map and facial area cropping network based on facial landmarks. Niu et al.[15] proposed to use the facial landmarks as person-specific shape regularization for

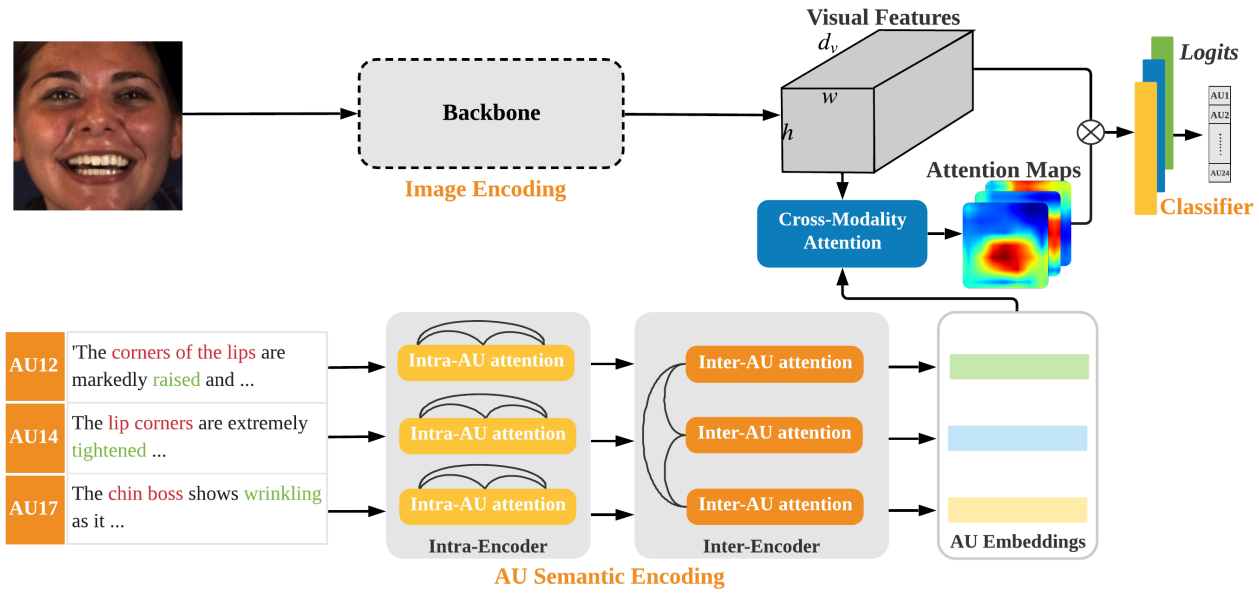

Figure 2. The overall framework of the proposed method. The visual features are first extracted by a backbone network. The AU embeddings are obtained through feeding the AU description sentence to an Intra-AU attention module to capture the relation among words in each sentence, followed by an Inter-AU attention module to capture the relation among AU sentences. The learned AU embeddings and visual features are combined together to generate the attention map through a cross-modality attention module, and the attention maps will be further utilized as weights for the aggregated feature. Finally, the classifier is applied for AU detection.

AU detection, where the features extracted from the facial landmarks guide the extraction of visual features through an orthogonal regularization, thus the model is subject-independent, as well it is generalizable to unseen subjects.

Note that AU textual descriptions provide rich AU semantic information about *facial area/position*, *action*, *motion direction/intensity*, and *relation of AUs*, but there is no reported work that utilizes such an auxiliary information for AU detection.

Learning label relation from label semantic embedding Several previous methods are proposed to explicitly model the label relationship from the the label semantic representation for multi-label classification. Chen et al.[2] proposed to explicitly model the label dependencies through a GCN from prior label representations for multi-label image recognition. As a result, the proposed method can effectively alleviate over-fitting and over-smoothing issues. You et al.[24] proposed a GCN based method to learn semantic label embeddings for multi-label classification. The semantic label embeddings explicitly model the label relationships, and further used as a guidance for learning cross-modality attention maps. As to AU detection, Li et al.[10] proposed to incorporate a GCN based AU relationship model to the visual features for the representation learning. To the best of our knowledge, there is no reported method that exploits the AU semantic description for AU detection so far.

[2] and [24] are the most related works, even though they target on the task of image classification. Our method differs significantly in following facts: (1) both [2] and [24] rely on a manually constructed label relation graph, while our method does not. Instead, our method automatically learn the AU relations from the AU semantic descriptions. (2) In contrast to [2][24] that utilize a graph neural network to model the label dependencies, our method utilizes the attention mechanism at three different levels: Intra-AU(*Words level*), Inter-AU(*Sentence level*) and Cross-modality(*Modality level*), thus capturing the rich semantic information for AU detection. The ablation studies in section 4.4 has demonstrated that our method is move effective than [2][24] in capturing the AU relations.

3. Proposed method

Fig 2 gives an overview of our proposed framework. It consists of three parts: image feature encoder, Intra-AU and Inter-AU attention-based encoders, and cross-modality attention network. As shown in the top part of Fig 2, a CNN-based backbone is used to encode the g iven image into visual features $\mathbf{V}$. The bottom part shows the textual feature encoding process. The Inter-AU attention and Intra-AU attention encoders are Transformer-based encoders, which

capture the intra-AU relations and inter-AU relations respectively. The cross-modality attention network, followed by a classifier, captures the cross-modality relations between visual features and textual features. In the following, we describe the individual modules of our framework as well as the loss functions.

3.1. Image Encoding

We first encode the given image I to the spatial visual features $\boldsymbol{V} = \{\mathrm{V}_1, \ldots, \mathrm{V}_{w\times h}\}, \mathrm{V}_i \in \mathbb{R}^{d_v}$ with a backbone model, where $w \times h$ is the size of feature, and d_v is the dimension of each feature channel in V_i.

3.2. AU Semantic Encoding

The overall AU semantic encoding contains both an Intra-AU encoder and an Inter-AU encoder, where the Intra-AU encoder is shared by words among AU description sentence, and the Inter-AU encoder is shared by AU sentence embeddings.

Input Embeddings The purpose of Intra-AU encoder is to model the intra-relations among the words in AU semantic description. The input to the Intra-AU encoder includes AU descriptions $S = \{S_1, \ldots, S_{N_S}\}$, where N_S is the number of AU descriptions. For each AU description S_i, we use the WordPiece tokenizer [21] to split it into tokens $\{s_{i,1}, \ldots, s_{i,N_i}\}$, where N_i is the number of tokens for each AU description. Apart from the token embeddings, we also assign positional encoding $\mathrm{p}_{s_{i,j}}$ to each word $s_{i,j}$. In particular, for token $s_{i,j}$, its input representations $\mathrm{w}_{i,j}$ is the sum of its trainable word embedding, segment embedding, and positional embedding:

$$\mathrm{w}_{i,j} = f_{\mathrm{LN}}\big(\mathrm{WordEmb}(s_{i,j}) + \mathrm{SegEmb}(j) + \mathrm{p}_{s_{i,j}}\big) \quad (1)$$

where $f_{\mathrm{LN}}(\cdot)$ stands for layer normalization [1].

Intra-AU Encoder Following the embedding layers, we apply the multi-layer transformer encoder to encode each AU description S_i. Like BERT [4], our Inter-AU encoder is used to encode contextual information for tokens within each sentence. Each layer of the Intra-AU encoder is the same as the vanilla transformer encoder layer [20]. Let $\boldsymbol{W}^l = (\boldsymbol{w}_1, \cdots, \boldsymbol{w}_{N_i})$ be the encoded features at the l-th transformer layer, $\boldsymbol{W}^0$ being the input layer. The features at the $(l+1)$-th layer are obtained by applying a transformer block defined as:

$$\boldsymbol{H}^{l+1} = f^l_{\mathrm{LN}}\big(\boldsymbol{W}^l + f^l_{\text{Self-Att}}(\boldsymbol{W}^l)\big) \quad (2)$$

$$\boldsymbol{W}^{l+1} = f^l_{\mathrm{LN}}\big(\boldsymbol{H}^{l+1} + f^l_{\mathrm{FF}}(\boldsymbol{H}^{l+1})\big) \quad (3)$$

where $f_{\text{Self-Att}}(\cdot)$ is the multi-headed self-attention module introduced in [1], which makes each token attend the other tokens with attention weights. The feed-forward (FF) sub-layer $f_{\mathrm{FF}}(\cdot)$ in Eq. 3 is further composed of two fully-connected (FC) sub-layers: $f^l_{\mathrm{FC}_2}(f_{\mathrm{GELU}}(f^l_{\mathrm{FC}_1}(\cdot)))$, where f_{GELU} represents the GeLU activation [9].

To obtain a fixed-length sentence representation for each AU, we get the AU representation $\mathbf{w}_i^{L_{\text{Intra}}}$ by computing the mean of all outputs among each sentence: $\mathbf{w}_i^{L_{\text{Intra}}} = \frac{1}{N_i}\sum_j^{N_i} \boldsymbol{w}_j^{\text{Intra}}$, where L_{Intra} is the final layer of Intra-AU encoder, N_i is the number of tokens in each AU sentence description. After the Intra-AU encoding, for AU descriptions $\{S_1, \ldots, S_{N_S}\}$, we have a set of embeddings: $\{\mathbf{w}_1^{L_{\text{Intra}}}, \ldots, \mathbf{w}_{N_S}^{L_{\text{Intra}}}\}$.

Inter-AU Encoder The Inter-AU encoder is designed to exchange information across multiple AU embeddings. Like Intra-AU encoder, we also apply the multi-layer transformer network to encode the input embeddings. Note that, the input to the Inter-AU encoder is the set of embeddings from the Intra-AU encoder, not tokens. The final output of Inter-AU encoder represents as $\{\mathbf{w}_1^{L_{\text{Inter}}}, \ldots, \mathbf{w}_{N_S}^{L_{\text{Inter}}}\}$, where L_{Inter} is the final layer of Inter-AU attention encoder, N_S is the number of AUs.

3.3. Cross-Modality Attention

From the AU semantic learning, we can obtain the AU embeddings, which encodes information of both Intra-/Inter-AU relation and area of interest. To fully utilize the rich information encoded in AU semantic embedding, we let the AU embeddings guide the generation of attention maps through the cross-modality attention module, defined as:

$$\mathrm{z}_k^i = \mathrm{ReLU}\Big(\frac{\mathrm{V}_i^T \bar{\mathbf{w}}_k^{L_{\text{Inter}}}}{||\mathrm{V}_i|| \cdot ||\bar{\mathbf{w}}_k^{L_{\text{Inter}}}||}\Big) \quad (4)$$

where $k \in \{1, 2, ..., N_s\}$,$i \in \{1, 2, ..., w \times h\}$, $||\cdot||$ represents the norm function, $\bar{\mathbf{w}}_k^{L_{\text{Inter}}}$is the linear projection of $\mathbf{w}_k^{L_{\text{Inter}}}$ from $\mathbb{R}^{d_w}$ to $\mathbb{R}^{d_v}$, and $\mathrm{V}_i \in \mathbb{R}^{d_v}$. We can obtain the category-specific cross-modality attention map z_k^i for AU_k at location i, which is further normalized to:

$$\alpha_k^i = \frac{\mathrm{z}_k^i}{\sum_{i=1}^{w\times h} \mathrm{z}_k^i} \quad (5)$$

The normalized cross-modality attention map can be further utilized as weight for the aggregated feature, because the high value in a specific location i of AU_k can be interpreted as the location i is more important than other locations for recognizing AU_k, thus the model needs to pay more attention to that location when detect AU_k.

$$\mathrm{x}_k = \sum_{i=1}^{w\times h} \alpha_k^i \mathrm{V}_i \quad (6)$$

where, $x_k \in \mathbb{R}^{d_v}$ is the final feature vector for AU_k. From this step, we can obtain N_s features for each input image.

3.4. AU detection

A classifier $f_C : \mathbb{R}^{d_v} \rightarrow \mathbb{R}^1$ is then shared by the N_s image features for estimating probability of AUs. A binary cross-entropy (BCE) loss function is used as the final loss function for AU recognition:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{N_s}\Big(y_k^i \times \log(\widehat{y_k^i}) + (1-y_k^i) \times \log(1-\widehat{y_k^i})\Big) \quad (7)$$

where N is the total number of training images, N_s is the number of AU, y_k^i and $\widehat{y_k^i}$ represent the ground truth label and prediction for AU_k in image i respectively.

4. Experiments

To evaluate our proposed method, we perform experiments on three public benchmark datasets: BP4D[25], DISFA[14] and BP4D+[27]. By comparing with the GCN based methods[2][24], we validate the effectiveness of the proposed Intra-AU and Inter-AU attention modules for automatically learning of the AU relations from AU semantic description. We also demonstrate that AU semantic embedding is beneficial for the discovery of more meaningful facial areas through visualization of the cross-modality attention maps.

4.1. Data

BP4D[25] is a widely used dataset for evaluating AU detection performance. The dataset contains 328 2D and 3D videos collected from 41 subjects (23 females and 18 males) under eight different tasks. As mentioned in the dataset, the most expressive 500 frames (around 20 seconds) are manually selected and labeled for AU occurrence from each one-minute long sequence, resulting in a dataset of around 140,000 AU-coded frames. For a fair comparison with the state-of-the-art methods, a three-fold subject-exclusive cross validation is performed on 12 AUs.

DISFA[14] is another benchmark dataset for AU detection, which contains videos from left view and right view of 27 subjects (12 females, 15 males). 12 AUs are labeled with AU intensity from 0 to 5, resulting in around 130,000 AU-coded images. Following the experimental setting in [10], 8 of 12 AUs with intensity greater than 1 from the left camera are used. F1-score is reported based on subject-exclusive 3-fold cross-validation.

BP4D+[27] is a multimodal spontaneous emotion dataset, where high-resolution 3D dynamic model, high-resolution 2D video,thermal (infrared) image and physiological data were acquired from 140 subjects. There are 58 males and 82 females, with ages ranging from 18 to 66 years old. Each subject experienced 10 tasks corresponding to 10 different emotion categories, and the most facially-expressive 20 seconds from four tasks were AU-coded from all 140 subjects, resulting in a database contains around 192,000 AU-coded frames. Following a similar setting in BP4D dataset, 12 AUs are selected and performance of 3-fold cross-validation is reported.

4.2. Implementation details

All the face images are aligned and cropped to the size of 256×256 using affine transformation based on the provided facial landmarks, randomly cropped to 224×224 for training, and center-cropping for testing. Random horizontal flip is also applied during training. To analyze the impact of our proposed method, we use the ResNet-18[8] architecture as the backbone and baseline.

Based on the FACS manual, we have summarized 15 AU semantic descriptions (i.e., *AU1, AU2, AU4, AU6, AU7, AU9, AU10, AU12, AU14, AU15, AU17, AU23, AU24, AU25, AU26*). The details are listed in the supplemental material.

The Intra-AU encoder has the same configuration as $\text{BERT}_{\text{Large}}$. More specifically, we set the number of layers L_{Intra} to 24, the hidden size to 1,024, and the number of heads to 16. The parameter of the Intra-AU encoder is initialized with pre-trained parameter[1], and freezed during training. For the Inter-AU encoder, we build upon the encoder block as used in transformer[20]. Note that the input here is the sentence embedding, rather than tokens. We set the number of layers L_{Inter} to 2, the hidden size to 1024, and the number of heads to 6. The size of final AU semantic embedding is 768, which is then projected to 512 through a linear function, and the visual features from ResNet-18 we used is $7 \times 7 \times 512$.

All the modules, except Intra-AU encoders, are randomly initialized. We use an Adam optimizer with initial learning rate of 0.001, and the learning rate is decayed after each epoch with momentum 0.85. The batch size is 100, and we train the model for 50 epochs with early stopping. We implement our method with the Pytorch [16] framework and perform training and testing on the NVIDIA GeForce 2080Ti GPU.

To evaluate the performance, we use the F1-score for comparison study with the state of the arts. F1-score is defined as the harmonic mean of the precision and recall. As the distribution of AU labels are unbalanced, F1-score is a preferable metric for performance evaluation.

4.3. Comparison with related methods

We compare our method to alternative methods, including Liner SVM (LSVM) [6], Joint Patch and Multi-label

[1] https://github.com/UKPLab/sentence-transformers

Table 1. F1 scores in terms of 12 AUs are reported for the proposed method and the state-of-the-art methods on the BP4D database. Bold numbers indicate the best performance; bracketed numbers indicate the second best.

Method	AU1	AU2	AU4	AU6	AU7	AU10	AU12	AU14	AU15	AU17	AU23	AU24	***Avg***
LSVM [6]	23.2	22.8	23.1	27.2	47.1	77.2	63.7	[64.3]	18.4	33.0	19.4	20.7	35.3
JPML[28]	32.6	25.6	37.4	42.3.	50.5	72.2	74.1	**65.7**	38.1	40.0	30.4	42.3	45.9
DRML[30]	36.4	41.8	43.0	55.0	67.0	66.3	65.8	54.1	33.2	48.0	31.7	30.0	48.3
EAC-net[11]	39.0	35.2	48.6	76.1	72.9	81.9	86.2	58.8	37.5	59.1	35.9	35.8	55.9
DSIN [3]	[51.7]	40.4	56.0	76.1	73.5	79.9	85.4	62.7	37.3	[62.9]	38.8	41.6	58.9
JAA [19]	47.2	44.0	54.9	[77.5]	74.6.	[84.0]	86.9	61.9	43.6	60.3	42.7	41.9	60.0
OF-Net [22]	50.8	[45.3]	[56.6]	75.9	75.9	80.9	**88.4**	63.4	41.6	60.6	39.1	37.8	59.7
LP-Net [15]	43.4	38.0	54.2	77.1	[76.7]	83.8	87.2	63.3	45.3	60.5	**48.1**	**54.2**	61.0
SRERL [10]	46.9	[45.3]	55.6	77.1	**78.4**	83.5	[87.6]	63.9	[52.2]	**63.9**	[47.1]	[53.3]	[62.9]
Ours (SEV-Net)	**58.2**	**50.4**	**58.3**	**81.9**	73.9	**87.8**	87.5	61.6	**52.6**	62.2	44.6	47.6	**63.9**

Table 2. F1 scores in terms of 8 AUs are reported for the proposed method and the state-of-the-art methods on DISFA dataset. Bold numbers indicate the best performance; bracketed numbers indicate the second best.

Method	AU1	AU2	AU4	AU6	AU9	AU12	AU25	AU26	***Avg***
LSVM[6]	10.8	10.0	21.8	15.7	11.5	70.4	12.0	22.1	21.8
DRML[30]	17.3	17.7	37.4	29.0	10.7	37.7	38.5	20.1	26.7
EAC-net[11]	41.5	26.4	66.4	[50.7]	**80.5**	**89.3**	88.9	15.6	48.5
DSIN[3]	42.4	39.0	[68.4]	28.6	46.8	70.8	90.4	42.2	53.6
JAA[19]	43.7	46.2	56.0	41.4	44.7	69.6	88.3	58.4	56.0
OF-Net[22]	30.9	34.7	63.9	44.5	31.9	[78.3]	84.7	[60.5]	53.7
LP-Net[15]	29.9	24.7	**72.7**	46.8	[49.6]	72.9	[93.8]	**65.0**	[56.9]
SRERL [10]	[45.7]	[47.8]	59.6	47.1	45.6	73.5	84.3	43.6	55.9
Ours (SEV-Net)	**55.3**	**53.1**	61.5	**53.6**	38.2	71.6	**95.7**	41.5	**58.8**

(JPML) [28], Deep Region and Multi-label (DRML) [30], Enhancing and Cropping Network (EAC-net)[11], Deep Structure Inference Network (DSIN) [3], Joint AU Detection and Face Alignment (JAA) [19],Optical Flow network (OF-Net) [22], Local relationship learning with Person-specific shape regularization (LP-Net) [15], and Semantic Relationships Embedded Representation Learning (SRERL) [10].

Table 1 shows the results of different methods on the BP4D database. We can see that our method achieves the best accuracy in recognizing *AU1*, *AU2*, *AU4*, *AU6*, *AU10*, and *AU15*, and outperforms all of the SOTA methods. Compared with the patch or region-based methods: JPML and DRML, our method achieves 18.0% and 15.6% higher performance on BP4D database. Compared with JAA and LP-Net, which used facial landmarks as a joint task or regularization for AU detection, our method still shows 3.9% and 2.9% improvement in terms of F1-score on the BP4D database. SRERL is the previous state-of-the-art method, and related to our method in terms of the use of AU relationships. But our method is significantly different with the SRERL: *First*, SRERL only use the visual modality, while our method not only use both the visual and textual modalities, but also consider the correlation through the cross-modality attention network; *Second*, a manually constructed AU relation graph from label distribution is needed for the GCN module in SRERL; while our model does not rely on the pre-defined AU relation graph, instead, it will automatically learn the AU relations from the AU semantic description through the Intra-AU and Inter-AU attention modules. With regard to the performance, our method is 1.0% higher in terms of F1-score than the SRERL.

The performance comparison in terms of 8 AUs on the DISFA database are reported in Table 2. As we can see, our method achieves 58.8% F1-score, outperforming all of the related works. Our method shows 32.1%, 2.8% and 1.9% improvement than DRML, JAA and LP-Net respectively. As compared to the related work of SRERL, our method shows 2.9% improvement.

Our method is also evaluated on the BP4D+ database, which contains more AU-labeled frames from more subjects, and the results are shown in Table 3. Except using the reported results from [23], we also report the results of ML-GCN[2] and MS-CAM[24] based on our own implementation. Note that both ML-GCN and MS-CAM are not originally designed for AU detection, we extend

Table 3. F1 scores in terms of 12 AUs are reported for the proposed method and the state-of-the-art methods on the BP4D+ database. Bold numbers indicate the best performance; bracketed numbers indicate the second best; * indicate the result from our own implementation.

Method	AU1	AU2	AU4	AU6	AU7	AU10	AU12	AU14	AU15	AU17	AU23	AU24	***Avg***
FACS2D-Net[23]	34.6	32.6	[44.1]	82.1	85.3	87.6	87.2	65.9	44.0	44.3	44.8	[29.6]	56.8
FACS3D-Net[23]	[43.0]	[38.1]	**49.9**	82.3	85.1	87.2	87.5	66.0	48.4	[47.4]	50.0	**31.9**	59.7
ML-GCN[2]*	40.2	36.9	32.5	84.8	[88.9]	89.6	**89.3**	81.2	[53.3]	43.1	55.9	28.3	60.3
MS-CAM[24]*	38.3	37.6	25.0	[85.0]	**90.9**	**90.9**	[89.0]	[81.5]	**60.9**	40.6	[58.2]	28.0	[60.5]
ResNet18	34.6	34.6	33.1	84.9	87.0	[90.0]	88.9	80.4	[53.3]	38.7	54.7	13.4	57.8
Ours (SEV-Net)	**47.9**	**40.8**	31.2	**86.9**	87.5	89.7	88.9	**82.6**	39.9	**55.6**	**59.4**	27.1	**61.5**

and re-implement the two methods during our experiments, more details will be described in Section 4.4. Our method achieves 61.5%, the highest F1-score when compared with related methods. It is worth noting that although FACS3D-Net[23] detects the AUs from the image sequence, our method still shows 1.8% improvement.

4.4. Ablation study

Self-attention based *vs* GCN based semantic encoding: ML-GCN[2] and MS-CMA[24] are the two recently proposed methods that leverage the label semantic embedding for multi-label classification. In ML-GCN[2], the authors proposed to explicitly model the label dependencies through a GCN from prior label representations for multi-label image recognition. MS-CAM[24] is an extension of the ML-GCN, which not only explicitly models the label relationships, but also adds a cross-modality attention module between the visual features and label embeddings. It is worth noting that a manually constructed graph (*label relation*) is necessary for both ML-GCN and MS-CAM. Although they are not originally designed for AU detection, the framework can be extended for AU detection. To demonstrate the effectiveness of the proposed Intra-AU and Inter-AU attention encoding module of our proposed method, we extend and re-implement the ML-GCN and MS-CAM methods for comparison. As a pre-defined graph is needed for the GCN module in both the ML-GCN and MS-CAM, we manually construct such a graph that starts from computing the Pearson correlation coefficient (PCC) between each pair of the AUs in the dataset, and then converts the AU relationship into positive and negative connections based on two thresholds (we use 0.2 and -0.03 for positive and negative thresholds respectively in our experiment, the same setting can be found in [10] as well).

The result is reported in Fig.3. As we can see, ML-GCN clearly shows improved performance than the ResNet-18 baseline in three datasets, demonstrating the effectiveness of adding the GCN modeled label relations for detection. Compared to the ML-GCN, MS-CAM achieves even better performance. Except the extra cross-modality attention module, MS-CAM is similar to ML-GCN in terms of the GCN based label relation encoding, so the improved performance can be used to demonstrate that the cross-modality attention mechanism is beneficial for the combination of semantic embedding and visual features.

Our method also contains a cross-modality attention module, however, it is significantly different with MS-CAM in how to encoder the semantic information: our method does not rely on the manually constructed AU relation graph, instead, it automatically learns the AU relation from the AU semantic descriptions through our Intra-AU and Inter-AU attention modules. The highest performance in three datasets demonstrates the effectiveness of the self-attention based AU semantic encoding.

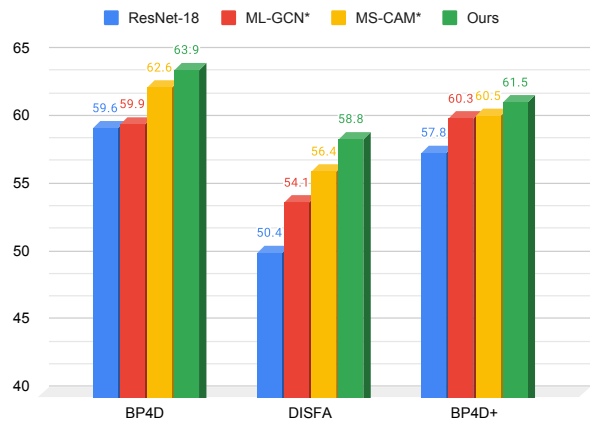

Figure 3. The comparison between Resnet-18, ML-GCN*[2], MS-CMA*[24], and our method on three datasets. * indicates the result from our own implementation.

Visualization of the cross-modality attention maps We visualize the learned cross-modality attention maps for several AUs to illustrate the ability of capturing meaningful regions for AU detection. We also compare with the attention maps learned in MS-CAM [24] in Fig.4. We can observe that our proposed method concentrate more on semantic regions, thus it is capable of exploiting more discriminative and meaningful information. For example, both

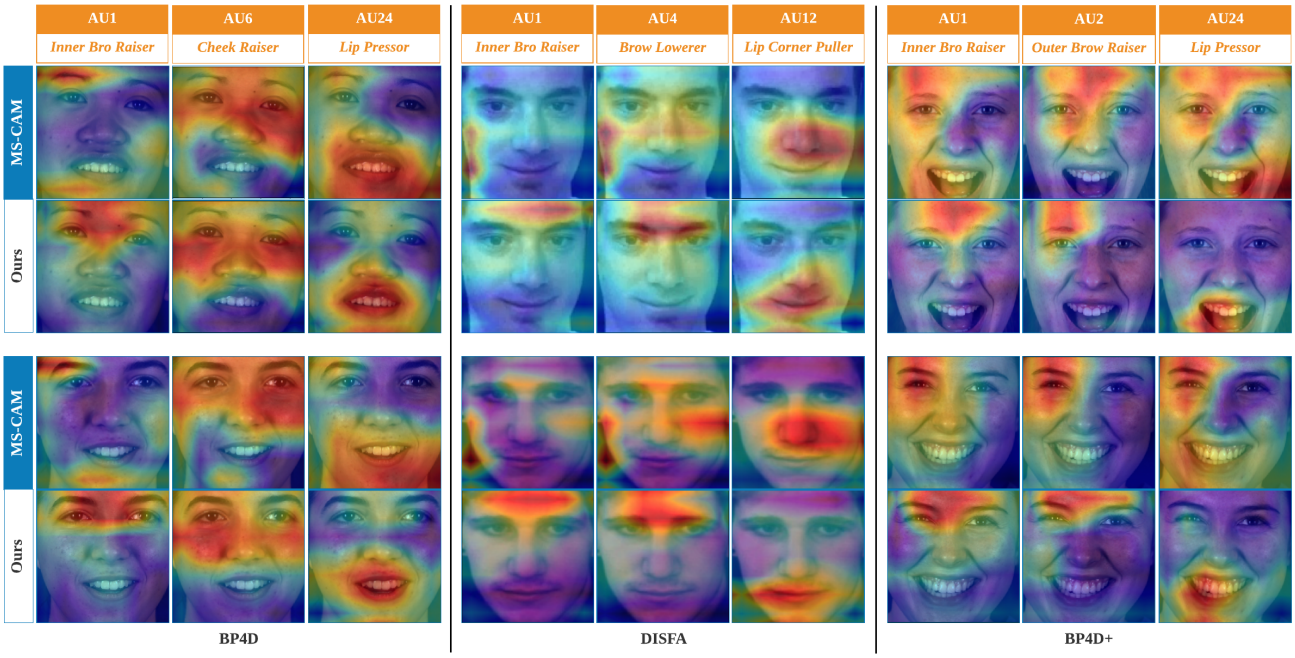

Figure 4. Visualization of the learned cross-modality attention maps for several AUs (as examples) of different subjects on BP4D, DISFA, and BP4D+ datasets, respectively. Each row shows the results of the same method; the first and third rows show the attention maps obtained through MS-CAM [24], and the second and the last rows represent the attention maps obtained by our method. Attention maps are visualized using heat-map and projected on the original images as well.

AU24: Lip Pressor (third column) and AU12:Lip Corner Puller (sixth column) occur around the lip area, our model shows a much better focus on the specific area than the corresponding attention maps in MS-CAM. As to the reason for the difference, we suspect that MS-CAM relies on the manually constructed AU relation graph, which is usually constructed using a statistic model from the AU label distribution; however the distribution is likely to be biased due to only part of the videos being selected and AU labeled (*for example only 30% frames are AU labeled in the BP4D dataset*). On the other hand, the AU relation is clearly illustrated in the AU semantic description. Our model does not rely on the statistic metric based AU relation graph, instead it will automatically learn the AU relations through Intra-AU and Inter-AU attention modules, thus resulting in a better result.

5. Conclusion

In this paper, we have proposed a novel framework by combining the visual features and AU semantic embeddings for the task of AU detection. There exist a number of works that have applied a variety of auxiliary information (such as *facial landmarks*, *relation among AUs and expressions*, *web facial images*, etc.) for AU detection. However, there is no AU semantic information from the textual domain that has ever been explored. Our new framework exploits the AU semantic description, which is believed to have much more rich information than the traditional binary AU labels, thus becomes the first of this kind for improving the performance of AU detection. In order to make full use of AU semantic information, we propose two new modules (so-called Intra-AU and Inter-AU attention modules) to capture the AU semantic embedding, which is further combined with the visual features for computing the cross-modality attention maps. Our proposed method is evaluated on three widely used facial expression databases, and has achieved superior performance over the peer SOTA methods.

6. Acknowledgement

The material is based on the work supported in part by the NSF under grant CNS-1629898 and the Center of Imaging, Acoustics, and Perception Science (CIAPS) of the Research Foundation of Binghamton University.

References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[2] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. Multi-label image recognition with graph convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5177–5186, 2019.

[3] Ciprian Corneanu, Meysam Madadi, and Sergio Escalera. Deep structure inference network for facial action unit recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 298–313, 2018.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019.

[5] Paul Ekman. Facial action coding system (facs). *A human face*, 2002.

[6] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. Liblinear: A library for large linear classification. *Journal of machine learning research*, 9(Aug):1871–1874, 2008.

[7] Jiuxiang Gu, Jason Kuen, Shafiq Joty, Jianfei Cai, Vlad Morariu, Handong Zhao, and Tong Sun. Self-supervised relationship probing. *Advances in Neural Information Processing Systems*, 33, 2020.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[9] Dan Hendrycks and Kevin Gimpel. Bridging nonlinearities and stochastic regularizers with gaussian error linear units. 2016.

[10] Guanbin Li, Xin Zhu, Yirui Zeng, Qing Wang, and Liang Lin. Semantic relationships guided representation learning for facial action unit recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8594–8601, 2019.

[11] Wei Li, Farnaz Abtahi, Zhigang Zhu, and Lijun Yin. Eac-net: A region-based deep enhancing and cropping approach for facial action unit detection. *arXiv preprint arXiv:1702.02925*, 2017.

[12] Wei Li, Farnaz Abtahi, Zhigang Zhu, and Lijun Yin. Eac-net: Deep nets with enhancing and cropping for facial action unit detection. *IEEE transactions on pattern analysis and machine intelligence*, 40(11):2583–2596, 2018.

[13] Yong Li, Jiabei Zeng, Shiguang Shan, and Xilin Chen. Self-supervised representation learning from videos for facial action unit detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10924–10933, 2019.

[14] S Mohammad Mavadati, Mohammad H Mahoor, Kevin Bartlett, Philip Trinh, and Jeffrey F Cohn. Disfa: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 4(2):151–160, 2013.

[15] Xuesong Niu, Hu Han, Songfan Yang, Yan Huang, and Shiguang Shan. Local relationship learning with person-specific shape regularization for facial action unit detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11917–11926, 2019.

[16] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8024–8035, 2019.

[17] Guozhu Peng and Shangfei Wang. Weakly supervised facial action unit recognition through adversarial training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2188–2196, 2018.

[18] Guozhu Peng and Shangfei Wang. Dual semi-supervised learning for facial action unit recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8827–8834, 2019.

[19] Zhiwen Shao, Zhilei Liu, Jianfei Cai, and Lizhuang Ma. Deep adaptive attention for joint facial action unit detection and face alignment. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 705–720, 2018.

[20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[21] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

[22] Huiyuan Yang and Lijun Yin. Learning temporal information from a single image for au detection. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–8. IEEE, 2019.

[23] Le Yang, Itir Onal Ertugrul, Jeffrey F Cohn, Zakia Hammal, Dongmei Jiang, and Hichem Sahli. Facs3d-net: 3d convolution based spatiotemporal representation for action unit detection. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 538–544. IEEE, 2019.

[24] Renchun You, Zhiyao Guo, Lei Cui, Xiang Long, Yingze Bao, and Shilei Wen. Cross-modality attention with semantic graph embedding for multi-label classification. In *AAAI*, pages 12709–12716, 2020.

[25] Xing Zhang, Lijun Yin, Jeffrey F Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, Peng Liu, and Jeffrey M Girard. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32(10):692–706, 2014.

[26] Yong Zhang, Weiming Dong, Bao-Gang Hu, and Qiang Ji. Classifier learning with prior probabilities for facial action unit recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5108–5116, 2018.

[27] Zheng Zhang, Jeff M. Girard, Yue Wu, Xing Zhang, Peng Liu, Umur Ciftci, Shaun Canavan, Michael Reale, Andy Horowitz, Huiyuan Yang, Jeffrey F. Cohn, Qiang Ji, and Lijun Yin. Multimodal spontaneous emotion corpus for human behavior analysis. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[28] Kaili Zhao, Wen-Sheng Chu, Fernando De la Torre, Jeffrey F Cohn, and Honggang Zhang. Joint patch and multi-label learning for facial action unit detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2207–2216, 2015.

[29] Kaili Zhao, Wen-Sheng Chu, and Aleix M Martinez. Learning facial action units from web images with scalable weakly

supervised clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2090–2099, 2018.

[30] Kaili Zhao, Wen-Sheng Chu, and Honggang Zhang. Deep region and multi-label learning for facial action unit detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3391–3399, 2016.