CCA-Transformer: Cascaded Cross-Attention Based Transformer for Facial Analysis in Multi-modal Data

Jun-Hwa Kim , Namho Kim , Minsoo Hong , Cheesun Won *

Posted Date: 27 March 2024

doi: 10.20944/preprints202403.1629.v1

Keywords: face analysis; expression; valence-arousal; action unit

Article

CCA-Transformer: Cascaded Cross-Attention Based Transformer for Facial Analysis in Multi-Modal Data

Junhwa Kim [1,†], Namho Kim [2,†], Minsoo Hong [2,†] and Cheesun Won [3,*]

[1] Department of Artifical Intelligence, Konyang University, Daejeon, South Korea
[2] Korea Broadcasting System(KBS), Seoul, South Korea
[3] Department of Electronics and Electrical Engineering, Dongguk University, Seoul, South Korea
* Correspondence: cswon@dongguk.edu
† These authors contributed equally to this work.

Abstract: One of the most crucial elements in deeply understanding humans on a psychological level is manifested through facial expressions. The analysis of a human behavior can be informed by their facial expressions, making it essential to employ indicators such as expression (Expr), valence-arousal (VA), and action units (AU). In this paper, we introduce the method proposed in the Challenge of the 6th Workshop and Competition on Affective Behavior Analysis in-the-wild (ABAW) at CVPR 2024. Our proposed method utilizes the multi-modal Aff-wild2 dataset, which is splitted into spatial and audio modalities. For the spatial data, we extract features using a SimMiM model that was pre-trained on a diverse set of facial expression data. For the audio data, we extract features using a WAV2VEC model. To fusion the extracted spatial and audio features, we employed the cascaded cross-attention mechanism of a transformer.

Keywords: face analysis; expression; valence-arousal; action unit

1. Introduction

With the advancement of deep learning technologies, the interaction between humans and machines [1–4] has increased significantly. Deep learning has enabled machines to understand and interpret human behaviors [5], emotions [6], and speech [7] more accurately, leading to more intuitive and efficient human-machine interfaces. This progress has facilitated the development of sophisticated applications such as virtual assistants, chatbots, autonomous vehicles, and personalized recommendation systems, enhancing the way humans interact with technology in various domains. For such interactions, understanding a human's intentions is crucial, but grasping their mental and psychological state is equally important. In this context, human behavior analysis becomes vital as it offers a window into the emotional and cognitive processes of individuals, enabling machines to interpret and respond to human needs and states more effectively. For such facial behavior analysis, identifying expression, valence-arousal, and action units is crucial as tools. Expression refers to the observable manifestations of emotions and feelings through facial movements. Valence-arousal represents the spectrum of emotional states, where valence indicates the positivity or negativity, and arousal denotes the level of excitement or calmness. Low arousal corresponds to states like boredom or relaxation, while high arousal is linked to excitement or fear. Valence differentiates emotions based on their positive or negative nature, with fear being highly negative and happiness being positive. Action units are the fundamental actions of individual muscles or groups of muscles that compose the facial expressions, providing a detailed map of facial behavior.

The ABAW6 [8] - the sixth competition on Affective Behavior Analysis in-the-wild is organized for human facial behavior analysis, focusing on VA estimation, Expr Recognition, and AU Detection challenges. The ABAW competition [9–20] began in 2017 and has been annually since 2020. The Aff-wild and Aff-wild2 datasets are specialized for large-scale human facial behavior analysis and are composed of multi-modal data, making them suitable for experimenting with the Transformer-based algorithms and conducive to technological advancement.

In this paper, we introduce the methods that we used in the VA, AU, and Expr competitions of ABAW6 [8]. For the aff-wild2 videos that include audio, we utilized two types of multi-modal

 doi:10.20944/preprints202403.1629.v1

data:image frames and audio. Our approach is divided into two stages. Firstly, we extract the features from each type of multi-modal data. Secondly, we fus the different extracted features. We used SimMIM [21] model to extract features from image frame and WAV2VEC [22] model for audio feature extraction. For the fusion method, we employed a cascading structure based on cross-attention of Transformer.

2. Methodology

In this section, we introduce the method proposed for Expression, Valence-arousal estimation, and AU (Action Unit) detection at the 6th ABAW competition [8]. The overall model structure is depicted in the Figure 1. The training approach is divided into two main steps. The first step involves extracting spatial and audio features from the Aff-Wild2 dataset. The second step involves training these extracted features using the cascaded cross-attention mechanism of transformer. Each step is conducted independently.

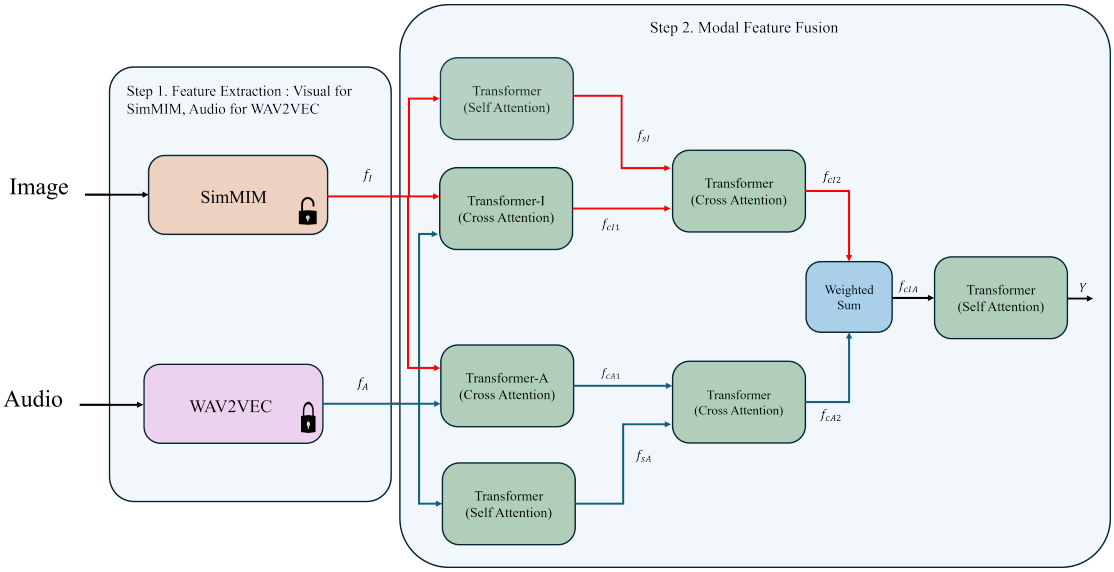

Figure 1. Overall structure of the proposed network.

2.1. Step 1: Feature Extraction

The SimMIM [21] model, based on masked image modeling, is particularly suited for face-related tasks due to its ability to capture fine-grained facial features through its innovative masking strategy. Masked image modeling allows the model to focus on reconstructing missing parts of the input, which makes the model to develop a deep understanding of facial structures and expressions. This attribute makes SimMIM well-suited for tasks involving facial behavior analysis, as it can learn nuanced differences in face. In our approach, the backbone of the SimMIM model employs the Swin Transformer [23], leveraging a pre-training step similar to the method used in previous winner [24] of ABAW 5. Specifically, we combined various datasets(AffectNet [25], CASIA-WebFace [26], CelebA [27], and IMDB-WIKI [28]), for an initial training phase to equip the model with a broad understanding of facial features across different contexts. Subsequently, we fine-tuned this pre-trained model using the ground truth expressions from the Aff-wild2 dataset, enabling it to adapt more closely to the specific characteristics of this dataset. Finally, by removing the fully connected layer at the end of the model, we were able to extract features related to Valence-Arousal, Action Units, and expressions.

We adapted the WAV2VEC [22] model to extract features from audio. The WAV2VEC model is highly effective due to its advanced self-supervised learning mechanism, which enables it to learn robust representations from raw audio by predicting the masked acoustic units. Using the WAV2VEC, we did not fine-tune the WAV2VEC model, instead, we used the pre-trained model weights to extract audio features directly.

 doi:10.20944/preprints202403.1629.v1

2.2. Step 2: Modal Feature Fusion

In Step 1, the obtained image feature f_I and audio feature f_A forward to cross attention transformer centered on each modality to yield f_{cI_1} and f_{cA_1}. When performing cross attention with the image, the audio feature is forward as the query, while the image feature is used for the key and value. Conversely, when cross attention with the audio, the image feature is forward as the query, with the audio feature is used for the key and value. The equations for the cross attention output vectors f_{cI_1} and f_{cA_1} are as follows:

$$f_{cI_1} = \text{Cross Attention}(f_A, f_I, f_I) = \text{Softmax}\left(\frac{f_A f_I^T}{\sqrt{d_k}}\right) f_I \tag{1}$$

$$f_{cA_1} = \text{Cross Attention}(f_I, f_A, f_A) = \text{Softmax}\left(\frac{f_I f_A^T}{\sqrt{d_k}}\right) f_A \tag{2}$$

where d_k is the dimensionality of the Key vector and T denotes the transpose operation. Additionally, self attention is applied to f_I and f_A, resulting in f_{sI} and f_{sA}. The equations for the self attention output vectors f_I and f_A are as follows:

$$f_{sI} = \text{Self Attention}(f_I, f_I, f_I) = \text{Softmax}\left(\frac{f_I f_I^T}{\sqrt{d_k}}\right) f_I \tag{3}$$

$$f_{sA} = \text{Self Attention}(f_A, f_A, f_A) = \text{Softmax}\left(\frac{f_A f_A^T}{\sqrt{d_k}}\right) f_A \tag{4}$$

Subsequently, f_{cI_1} and f_{sI} are input to cross-attention to obtain f_{cI_2}. Similarly, f_{cA_1} and f_{sA} are input to cross-attention to obtain f_{cA_2}. The equations for the cross attention output vectors f_{cI_5} and f_{cA_5} are as follows:

$$f_{cI_2} = \text{Cross Attention}(f_{sA}, f_{cI_1}, f_{cI_1}) \tag{5}$$

$$f_{cA_2} = \text{Cross Attention}(f_{sI}, f_{cA_1}, f_{cA_1}) \tag{6}$$

f_{cI_2} and f_{cA_2} are combined through a weighted sum where we assign a weight of 0.7 to f_{cI_2} and 0.3 to f_{cA_2}, to obtain f_{cIA}. Finally, self attention is applied to f_{cIA} to generate the output Y. The equations for the self attention output vectors Y is as follows:

$$Y = \text{self Attention}(f_{cIA}, f_{cIA}, f_{cIA}) \tag{7}$$

Finally, for the VA task, the module uses a 1D convolutional layer to transform the input feature dimension from 768 to 256. Following this, it defines two separate sequential models (vhead and ahead) for predicting valence and arousal, respectively. Each model consists of a series of linear layers, batch normalization, GELU activations, and dropout layers, culminating in a single output passed through a Tanh activation function. For the AU detection task, the module defines a simpler sequential model (head) that consists of a single linear layer mapping the input features to the number of classes, followed by a Sigmoid activation function to predict the presence of each action unit. In the Expr recognition task, the model simplifies even further, utilizing only a single linear layer (head) that maps the input features directly to the number of expression classes, enabling the direct categorization of facial expressions.

 doi:10.20944/preprints202403.1629.v1

3. Experiments

3.1. Implementation details

Our experiments were conducted in the Pytorch environment with the following specifications: Ubuntu 18.04.6 LTS, 128 GB RAM, and 2 NVIDIA RTX A6000 GPUs. We used the Aff-wild2 dataset for the experiments. To extract image features, we utilized the SimMIM model [21], which was not directly trained on the Aff-Wild2 dataset. Instead, it was initially trained using datasets AffectNet [25], CASIA-WebFace [26], CelebA [27],and IMDB-WIKI [28], and then fine-tuned using the ground truth for Expression, Valence-arousal, and Action Unit from the Aff-Wild2 dataset. To obtain audio features, we employed the Wav2Vec model [22], utilizing its pre-trained version without any additional fine-tuning.

The aff-wild2 dataset is in-the-wild and provides extensive annotations across three key tasks: valence-arousal (VA), expression recognition (Expr), and action unit detection (AU). For the VA task, the dataset includes 594 videos with around 3 million frames from 584 subjects, annotated for valence and arousal. In the Expr task, there are 548 videos with approximately 2.7 million frames annotated for the six basic expressions (anger, disgust, fear, happiness, sadness, surprise), the neutral state, and an 'other' category for additional expressions or affective states. Finally, for the AU task, the dataset comprises 547 videos with around 2.7 million frames annotated for 12 action units (AU1, AU2, AU4, AU6, AU7, AU10, AU12, AU15, AU23, AU24, AU25, AU26), offering a detailed framework for action unit analysis.

3.2. Evaluation Metrics

For the Affective Behavior Analysis in-the-wild (ABAW) competition, each of the three tasks, Expression (EXPR), Valence-Arousal (VA), and Action Unit (AU), has a specific performance measure.

Expression (EXPR) Task: The F1 score is used as an evaluation metric to assess the performance. The F1 score combines precision and recall into a single measure, which is particularly useful in imbalanced datasets. The F_1 score is defined as:

$$F_1 = \frac{1}{n}\sum_{i}^{n}\frac{2 \times precision_i \times recall_i}{precision_i + recall_i}, \tag{8}$$

where n is the number of emotion classes, $precision_i$ is the precision of the i-th class, and $recall_i$ is the recall of the i-th class.

Valence-Arousal (VA) Task: The performance measure is the mean Concordance Correlation Coefficient (CCC) of valence and arousal. The CCC measures the agreement between observed and predicted scores. The performance P is given by:

$$P = 0.5 \times (CCC_{arousal} + CCC_{valence}), \tag{9}$$

where $CCC_{arousal}$ is the CCC for arousal and $CCC_{valence}$ is the CCC for valence.

Action Unit (AU) Task: The performance measure is the average F1 Score across all 12 categories. The performance P is computed as:

$$P = \frac{\sum(F1_i)}{12}, \tag{10}$$

where $F1_i$ is the F1 score for the i-th AU category.

3.3. Results

Expression (EXPR) Task: For the EXPR task, our model achieved varying levels of performance across different emotion classes. The highest F1-score was observed in 'Anger', 'Disgust', and 'Fear' categories, each achieving a perfect score of 1.0, indicating that our model could identify these expressions with high precision and recall. The 'Neutral' expression received an F1-score of 0.6063,

 doi:10.20944/preprints202403.1629.v1

which suggests a reasonably good recognition rate. However, the model struggled with 'Surprise' and 'Sadness' expressions, evidenced by F1-scores of 0.0028 and 0.0178, respectively. This indicates a need for further model refinement to improve its sensitivity to these expressions. The 'Happiness' expression, often easily recognizable, had a lower-than-expected F1-score of 0.2857, which could be attributed to the complexity of the dataset. The 'Other' category, encompassing various non-standard expressions, achieved a moderate F1-score of 0.4238. Overall, the average F1-score for the EXPR task was 0.5420, providing a baseline for future improvements. The detailed results for each emotion class can be found in Table 1.

Table 1. Experimental results for the EXPR Task on the Aff-Wild2 validation set.

EXPR Task	F1-Score
Neutral	0.6063
Anger	1.0
Disgust	1.0
Fear	1.0
Happiness	0.2857
Sadness	0.0178
Surprise	0.0028
Other	0.4238
Average	0.5420

Action Unit (AU) Task: The AU task results were promising, with an average F1-score of 0.7077 across all 12 categories. The model was particularly effective in detecting 'AU15' and 'AU23', where it reached the maximum F1-score of 1.0, indicating a perfect match between predictions and the ground truth. Other AUs like 'AU6', 'AU7', 'AU10', 'AU12', and 'AU25' also showed high F1-scores, all above 0.69, demonstrating the model's strong capability in recognizing these facial muscle movements. 'AU24' received the lowest score of 0.4886, suggesting areas where the model may require additional training data or feature engineering. The detailed F1-scores for each AU category are presented in Table 2.

Table 2. Experimental results for the AU Task on the Aff-Wild2 validation set.

AU Task	F1-Score
AU1	0.5909
AU2	0.6365
AU4	0.6285
AU6	0.7126
AU7	0.6841
AU10	0.6930
AU12	0.7017
AU15	1.0
AU23	1.0
AU24	0.4886
AU25	0.8148
AU26	0.5423
Average	0.7077

Valence-Arousal (VA) Task: For the VA task, the model's performance was quantified using the Concordance Correlation Coefficient (CCC), with 'Arousal' obtaining a CCC of 0.5906 and 'Valence' a CCC of 0.4328. The average CCC for the VA task was 0.5117, indicating a moderate agreement with the ground truth. These results highlight the challenges in accurately predicting the subtle variations in emotional intensity represented by valence and arousal dimensions. Detailed performance metrics for valence and arousal can be seen in Table 3.

 doi:10.20944/preprints202403.1629.v1

Table 3. Experimental results for the VA Task on the Aff-Wild2 validation set.

AU Task	CCC
Valence	0.4328
Arousal	0.5906
Average	0.5117

4. Conclusions

In this paper, we have presented a method that extracts features from visual information using SimMIM and from auditory information using WAV2VEC. Subsequently, in the fusion process of the two modalities, we introduced a cascaded approach based on cross-attention to integrate visual and auditory information for the 6th Affective Behavior Analysis in-the-wild (ABAW) Competition at CVPR 2024.

References

1. Muhammad, G.; Alsulaiman, M.; Amin, S.U.; Ghoneim, A.; Alhamid, M.F. A facial-expression monitoring system for improved healthcare in smart cities. *IEEE Access* **2017**, *5*, 10871–10881.
2. Davoudi, A.; Malhotra, K.R.; Shickel, B.; Siegel, S.; Williams, S.; Ruppert, M.; Bihorac, E.; Ozrazgat-Baslanti, T.; Tighe, P.J.; Bihorac, A.; others. Intelligent ICU for autonomous patient monitoring using pervasive sensing and deep learning. *Scientific reports* **2019**, *9*, 8020.
3. Rouast, P.V.; Adam, M.T.; Chiong, R. Deep learning for human affect recognition: Insights and new developments. *IEEE Transactions on Affective Computing* **2019**, *12*, 524–543.
4. Suma, V. Computer vision for human-machine interaction-review. *Journal of trends in Computer Science and Smart technology (TCSST)* **2019**, *1*, 131–139.
5. Hu, K.; Jin, J.; Zheng, F.; Weng, L.; Ding, Y. Overview of behavior recognition based on deep learning. *Artificial intelligence review* **2023**, *56*, 1833–1865.
6. Chowdary, M.K.; Nguyen, T.N.; Hemanth, D.J. Deep learning-based facial emotion recognition for human–computer interaction applications. *Neural Computing and Applications* **2023**, *35*, 23311–23328.
7. Hema, C.; Marquez, F.P.G. Emotional speech recognition using cnn and deep learning techniques. *Applied Acoustics* **2023**, *211*, 109492.
8. Kollias, D.; Tzirakis, P.; Cowen, A.; Zafeiriou, S.; Shao, C.; Hu, G. The 6th Affective Behavior Analysis in-the-wild (ABAW) Competition. *arXiv preprint arXiv:2402.19344* **2024**.
9. Zafeiriou, S.; Kollias, D.; Nicolaou, M.A.; Papaioannou, A.; Zhao, G.; Kotsia, I. Aff-wild: Valence and arousal 'in-the-wild'challenge. Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on. IEEE, 2017, pp. 1980–1987.
10. Kollias, D.; Sharmanska, V.; Zafeiriou, S. Face Behavior a la carte: Expressions, Affect and Action Units in a Single Network. *arXiv preprint arXiv:1910.11111* **2019**.
11. Kollias, D.; Tzirakis, P.; Nicolaou, M.A.; Papaioannou, A.; Zhao, G.; Schuller, B.; Kotsia, I.; Zafeiriou, S. Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. *International Journal of Computer Vision* **2019**, pp. 1–23.
12. Kollias, D.; Zafeiriou, S. Expression, Affect, Action Unit Recognition: Aff-Wild2, Multi-Task Learning and ArcFace. *arXiv preprint arXiv:1910.04855* **2019**.
13. Kollias, D.; Schulc, A.; Hajiyev, E.; Zafeiriou, S. Analysing Affective Behavior in the First ABAW 2020 Competition. 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)(FG), pp. 794–800.
14. Kollias, D.; Sharmanska, V.; Zafeiriou, S. Distribution Matching for Heterogeneous Multi-Task Learning: a Large-scale Face Study. *arXiv preprint arXiv:2105.03790* **2021**.
15. Kollias, D.; Zafeiriou, S. Affect Analysis in-the-wild: Valence-Arousal, Expressions, Action Units and a Unified Framework. *arXiv preprint arXiv:2103.15792* **2021**.
16. Kollias, D.; Zafeiriou, S. Analysing affective behavior in the second abaw2 competition. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 3652–3660.

17. Kollias, D. Abaw: Valence-arousal estimation, expression recognition, action unit detection & multi-task learning challenges. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 2328–2336.
18. Kollias, D. ABAW: learning from synthetic data & multi-task learning challenges. European Conference on Computer Vision. Springer, 2023, pp. 157–172.
19. Kollias, D. Multi-Label Compound Expression Recognition: C-EXPR Database & Network. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 5589–5598.
20. Kollias, D.; Tzirakis, P.; Baird, A.; Cowen, A.; Zafeiriou, S. Abaw: Valence-arousal estimation, expression recognition, action unit detection & emotional reaction intensity estimation challenges. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 5888–5897.
21. Xie, Z.; Zhang, Z.; Cao, Y.; Lin, Y.; Bao, J.; Yao, Z.; Dai, Q.; Hu, H. Simmim: A simple framework for masked image modeling. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 9653–9663.
22. Schneider, S.; Baevski, A.; Collobert, R.; Auli, M. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862* **2019**.
23. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 10012–10022.
24. Zhang, W.; Ma, B.; Qiu, F.; Ding, Y. Multi-modal facial affective analysis based on masked autoencoder. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 5792–5801.
25. Mollahosseini, A.; Hasani, B.; Mahoor, M.H. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing* **2017**, *10*, 18–31.
26. Yi, D.; Lei, Z.; Liao, S.; Li, S.Z. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923* **2014**.
27. Liu, Z.; Luo, P.; Wang, X.; Tang, X. Deep learning face attributes in the wild. Proceedings of the IEEE international conference on computer vision, 2015, pp. 3730–3738.
28. Rothe, R.; Timofte, R.; Van Gool, L. Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision* **2018**, *126*, 144–157.