An Ensemble Approach for Facial Behavior Analysis in-the-wild Video

Hong-Hai Nguyen[1], Van-Thong Huynh[1], Soo-Hyung Kim*
Department of Artificial Intelligence Convergence
Chonnam National University
Gwangju, South Korea
{hhnguyen,vthuynh,shkim}@jnu.ac.kr

Abstract

Human emotions recognization contributes to the development of human-computer interaction. The machines understanding human emotions in the real world will significantly contribute to life in the future. This paper introduces the 3rd Affective Behavior Analysis in-the-wild (ABAW3) 2022 challenge. We focused on solving the problem of the Valence-Arousal (VA) estimation and Action Unit (AU) detection. For valence-arousal estimation, we conducted two stages: creating new features from multimodel and temporal learning to predict valence-arousal. First, we make new features; the Gated Recurrent Unit (GRU) and Transformer are combined using a Regular Networks (RegNet) feature, which is extracted from the image. The next step is the GRU combined with local attention to predict valence-arousal. The Concordance Correlation Coefficient (CCC) was used to evaluate the model. The result achieved 0.450 for valence and 0.445 for arousal on the test set, outperforming the baseline method with a corresponding CCC of 0.180 for valence and 0.170 for arousal. We also performed additional experiments on the action unit task with simple transformer blocks. We achieved a score of 49.04 on the test set in terms of F_1 score, which outperforms the baseline method with a corresponding F1 score of 36.50. Our submission to ABAW3 2022 ranks 3rd for both tasks.

1. Introduction

People's emotions affect their lives and work. Researchers are trying to create machines that can detect and analyze human emotions that contribute to the development of intelligent machines. Therefore, there are a lot of applications in life, such as medicine, health, tracking or driver fatigue [10].

The practical applications have many challenges from an uncontrolled environment. However, data sources are increasingly various from social networks and applications. Besides, the deep learning network improved the analysis and recognition process. Therefore, the ABAW3 2022 [10] was organized for affective behavior analysis in the wild. The challenge includes four tasks: Valence-Arousal (VA) Estimation, Expression (EXPR) Classification, Action Unit (AU) Detection, and Multi-Task-Learning (MTL). This paper only focuses on the VA task. In this task, participants will predict a valence-arousal dimension based on data from the video. In the AU task, participants will predict 12 action units based on data from the video.

In this study, we utilize feature extraction from the deep learning model. The frame-level feature are extracted from the RegNet network [23]. We perform K-fold cross validation with the combination of Gated Recurrent Unit (GRU) [2] and Transformer [6] on RegNet features to evaluate as a baseline model and produce predictions which is used as a new set of features, stage 1. For stage 2, we utilized these features with GRU to fuse them in temporal dimension and improve the results with local attention mechanism.

In this work, we focus on valence-arousal prediction. Our contributions in this paper are summarized as:

- Utilization features from the deep learning model.
- Using multimodal to create new features which increase speed training.
- Combination of local attention with GRU for sentiment analysis.
- Conducting experiments on different models to compare with baseline method.

The next parts of our paper are presented in the following sections: section 2 is related work, section 3 is methodology and the experimental results in section 4 and finally, the conclusion in section 5.

*Corresponding author
[1]Equal contribution

2. Related work

In this section, we shortly summarize some datasets and works related to the problem of affective behavior in the previous challenge.

2.1. Affect Annotation Dataset

In the previous challenge [11–17, 29], the ABAW3 provides a large-scale dataset Aff-Wild2 for affective behavior analysis in-the-wild. The dataset used is the Aff-wild2 which was extended from Aff-wild [29]. The dataset contains annotations for challenges: Valence-Arousal regression, basic emotions, and Action Unit. Aff-wild2 expands the number of videos with 567 videos annotated by valence-arousal, 548 videos annotated by 8 expression categories, 547 videos annotated by 12 AUs, and 172,360 images are used that contain annotations of valence-arousal; 6 basic expressions, plus the neutral state, plus the "other" category; and 12 action units.

2.2. Affective Behavior Analysis in the wild

The affective behavior analysis in the wild challenge has attracted a lot of researchers. Deng et al. [4] applied deep ensemble models learned by a multi-generational self-distillation algorithm to improve emotion uncertainty estimation. Regarding architectures, the author used features extractors from the efficient CNN model and applied GRU as temporal models. Zhang et al. [30] introduced multi-task recognition, which is a streaming network by exploiting the hierarchical relationship of emotional expression in the second ABAW challenge. In the paper, Vu et al. [27] used a multi-task deep learning model for continuous emotion recognition and facial expressions prediction. The authors conducted the knowledge distillation architecture to two networks training: teacher and student model. Kuhnke et al. [18] introduced a two-stream network for multi-task training. The model used the multimodel information extracted from audio and vision. The authors [3] solved two challenges of the competition. First, the problem is highly imbalanced in the dataset. Second, the datasets do not include labels for all three tasks. The author applied balancing techniques and proposed a teacher-student structure to learn from the imbalance labels to tackle the challenges.

3. Methodology

For this section, we introduce the proposed method for continuous emotion estimation and action unit detection. For the VA task, our approach contains two stages: create new features to increase training speed in Figure 1 and use GRU to learn temporal information, illustrated in Figure 2. Besides, local attention was applied to improve the model. In the AU task, Transformers blocks were combined, Figure 3.

3.1. Valence and arousal estimation

Visual feature extraction: Our visual feature is based on RegNet [23] architecture, a lightweight and efficient network. RegNet consists four stages to operate progressively reduced resolution with sequence of identical blocks. The pretrained weight from ImageNet [5] was used as initial training, and the last three stages are unfreezing to learn new representation from facial data.

Gated Recurrent Unit: Introduced by Cho et al. [2], GRU solved the problem of vanishing gradient, which comes with a standard recurrent neural network. There are primarily two gates in a GRU. The first gate is the reset gate (gr as Equation 1), and the other is the update gate (gu as Equation 3). The reset gate determines how much past information to forge and keep short-term dependencies in sequences as Equation 2.

$$gr_t = \sigma(W_x^{(gr)} x_t + W_h^{(gr)} h_{t-1}) \tag{1}$$

where x_t is the input at the time step t, corresponding weight is $W_x^{(gr)}$; h_{t-1} is hidden state of the previous $t-1$, corresponding weight is $W_h^{(gr)}$; σ is the sigmoid activation.

$$r = tanh(gr_t \odot (W^{(h)} h_{t-1}) + W^{(x)} x_t) \tag{2}$$

where $tanh$ is tanh function and $\odot$ is Hadamard product operation.

The update gate help to decide what information to throw and what new information to add as Equation 4, which helps pick up long-term information.

$$gu_t = \sigma(W_x^{(gu)} x_t + W_h^{(gu)} h_{t-1}) \tag{3}$$

where x_t is the input at the time step t, corresponding weight is $W_x^{(gu)}$; h_{t-1} is hidden state of the previous $t-1$, corresponding weight is $W_h^{(gu)}$; σ is the sigmoid activation.

$$u = gu_t \odot h_{t-1} \tag{4}$$

Finally, the network calculates the h_t vector (Equation 5), storing information for the current unit and input for the next step.

$$h_t = r \odot (1 - gu_t) + u \tag{5}$$

Transformer: Transformer architecture is firstly introduced in 2017 [26] as one of the most potent models created to date. The Transformer consists of encode-decoder stacks. The encoder represents the input to hold the information to e learned. The decoder uses the information presented by the encoder to predict the output. A Transformer model has learned relationships in sequential data. The mathematical technique used in the Transformer model is called the attention mechanism, which enables the transformers to capture long-term memory. Input of attention mechanism consists

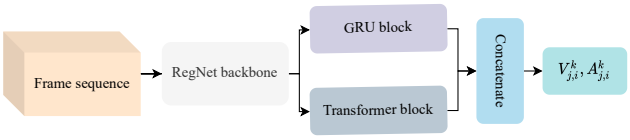

Figure 1. Feature extraction module: Valence-Arousal Predictor (VAP).

of Query (Q), Value (V) with dimension d_v, and Key (K) with dimension d_k. The dot product of Query and Key after that are divided each by $\sqrt{(d_k)}$. The softmax function will be applied to obtain the weights on the values. The Equation 6 is an attention formula.

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (6)$$

Local Attention: Local attention mechanisms [1, 22, 24, 25] consider a small subset of the source. The mechanism focuses on a small window that is distinguishable. Therefore, the local attention has the advantage of decreasing computation and being easy to train.

Our temporal learning for unimodal consists of a combination of Gated recurrent units (GRU) block [2] - a standard recurrent network, and Transformer block [26] - attention based for sequential learning, as shown in Figure 1. The representations from GRU and Transformer are concatenated to form a new feature vector and fed to a fully connected (FC) layer for producing valence and arousal scores. These emotion scores are used to calculate the loss function for the respective blocks. The mean of two-loss functions will be calculated, and the final loss representative for optimizing the whole system.

We conducted K-fold cross validation to obtain different models for ensemble learning with $K = 5$. The scores from each fold are combined together and to form a single vector for each frame in video, F_j, which can be formulated as

$$F_j = \{V^1_{j,i}, V^2_{j,i}, V^3_{j,i}, V^4_{j,i}, V^5_{j,i}, A^1_{j,i}, A^2_{j,i}, A^3_{j,i}, A^4_{j,i}, A^5_{j,i}\} \quad (7)$$

where $V^k_{j,i}, A^k_{j,i}$ are valence and score for fold k^{th} of i^{th} frame in j^{th} video with $k = \overline{1, K}$, $i = \overline{1, N_j}$, and N_j is the length of j^{th} video. To ensemble results from K-fold models, we deployed GRU-based architecture for modelling temporal relationship, followed by two local attention layers to adjust the contribution of features, as in Figure 2.

3.2. Action unit detection

In this task, we also deploy RegNet [23] as visual feature extraction as described in subsection 3.1 to obtain X, a sequence of L feature vectors with size D from a sequence of L frames, in which each frame has dimensions of $112 \times 112 \times 3$. Then two branches are deployed with

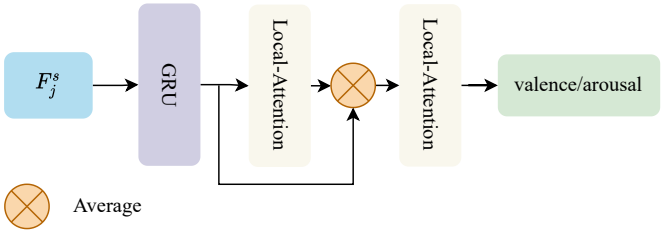

Figure 2. Overview of prediction model: Gated Recurrent Unit combined with local attention. Where s is sequence and F_j is vector scrores of j^{th} video.

Transformer [26] blocks, T_1 and T_2, for temporal learning and fully connected layers for prediction of 12 action units, as in Figure 3. These blocks has the same number of parameters but different in term of input (source) and output (target). T_1 directly use the output from RegNet, in which the weight is initialized from a pretrained model and fine-tune on Affwild2 dataset, as the source and target which contain solid knowledge. In case of T_2, we attach a *feature expansion and compression* block, $\mathcal{F}_{ec}$, in which the feature from RegNet are expanded to higher dimension with a factor of k, and then reduce to original dimension, which can be described as

$$\mathcal{F}_{ec} = \sigma(W_2\sigma(W_1X)), \quad (8)$$

where W_1, W_2 are learned weights with size $L \times kD$ and $L \times D$, respectively, used for the linear transformation of the input, and σ is a non-linear function, as we used ReLU in our approach. W_1 and W_2 are initialized randomly with He initialization [7], which contain noise and weak knowledge. To tackle of that, we attached a full connected layer for directly optimize it on Affwild2 and fused together with predictions from $T_1, T2$ block to obtain final prediction for 12 action units.

4. Experiments and results

4.1. Dataset

The VA task includes 567 videos with 455 subjects (277 males and 178 females) was annotated by four experts for valence and arousal score in range of $[-1, 1]$. The AU task contains 547 videos that include annotations in terms of 12 AUs, namely AU1, AU2, AU4, AU6, AU7, AU10, AU12, AU15, AU23, AU24, AU25, and AU26 of around 2.7M frames, with 431 subjects (265 males and 166 females), have been annotated in a semi-automatic method.

4.2. Valence and arousal estimation

The networks were conducted with the PyTorch Deep Learning toolkit. For stage 1, the GRU was set with 256-dimensional hidden states and two layers. The transformer

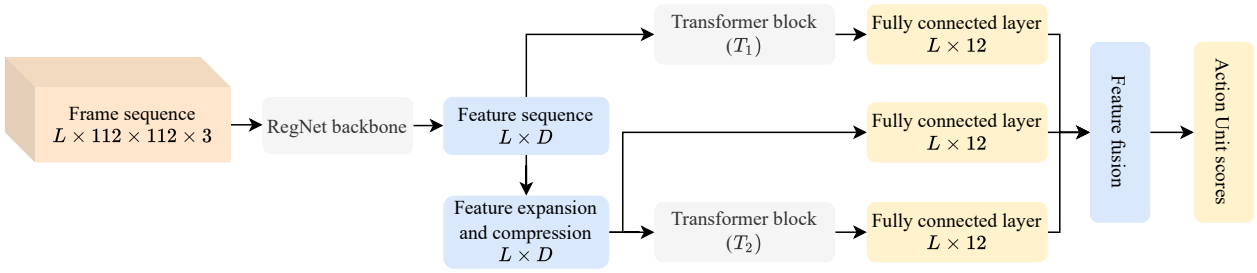

Figure 3. An overview of our action unit detection system.

was set with eight heads of multi-head attention, three for the sub-layer in the decoder and encoder, 1024-dimensional for the feedforward network, and 0.3 for the dropout value. For stage 2, the GRU was set with 256-dimensional hidden states, four-layered. The cosine annealing warm restarts [21] were used for training both stages. The networks are trained in 25 epochs with initial learning rate of 0.001 and Adam optimizer [9].

The mean between valence and arousal Concordance Correlation Coefficient (CCC) [19], $\mathcal{P}$, is used to evaluate the performance of the model as

$$\mathcal{P}_{\text{VA}} = \frac{\mathcal{P}_{\text{V}} + \mathcal{P}_{\text{A}}}{2}, \tag{9}$$

where $\mathcal{P}_{\text{V}}$ and $\mathcal{P}_{\text{A}}$ are the CCC of valence and arousal, respectively, which is defined as

$$\mathcal{P} = \frac{2\rho\sigma_{\hat{Y}}\sigma_{Y}}{\sigma_{\hat{Y}}^2 + \sigma_{\hat{Y}}^2 + (\mu_{\hat{Y}} - \mu_{Y})^2} \tag{10}$$

where μ_Y was the mean of the label Y, $\mu_{\hat{Y}}$ was the mean of the prediction $\hat{Y}$, $\sigma_{\hat{Y}}$ and σ_Y were the corresponding standard deviations, ρ was the Pearson correlation coefficient between $\hat{Y}$and Y.

The CCC loss function was used all continuous-time based on the VA task of the challenge. The loss function was defined as follows:

$$\mathcal{L} = 1 - \mathcal{P}_{\mathcal{VA}} \tag{11}$$

We describe the results of our K-fold cross validation experiments with training set of Affwild2, and evaluate on Affwild2 validation set separately, as shows in Table 1. All folds give better results than baseline with combined valence and arousal. In addition, the best result, 0.465, is obtained by averaging 5 folds.

Table 2 present the results of our final models with prior works. We conducted experiments with GRU separated and GRU combined with local attention by k-fold features. Moreover, we also experimented on the combination of

Table 1. Evaluation of VAP with K-fold for valence and arousal estimation on the original Affwild2 validation set. K-fold splits are done on Affwild2 training set only.

Fold	Valence	Arousal	Combined
1	0.290	0.491	0.391
2	0.348	0.435	0.391
3	0.339	0.500	0.419
4	0.294	0.491	0.392
5	0.362	0.492	0.427
Average	**0.390**	**0.540**	**0.465**
Baseline [10]	0.31	0.17	0.24

GRU and Transformer based on RegNet feature. All methods give better results than baseline by approximate twice time. The local attention made a small improvement of the ensemble model, which proving the potential of applying attention to the system. Furthermore, our method is higher than previous works [4, 30], respectively 0.494 and 0.495.

In total, 33 Teams took part in this sub-challenge in which 16 Teams submitted and 7 Teams better than the baseline. Our method got 0.450, 0.445, 0.448 corresponding to valence, arousal, and combined, respectively. Moreover, we ranked 3rd in this competition's task.

4.3. Action unit detection

Our network architectures is trained by using SGD with learning rate of 0.9 combine and Cosine annealing warm restarts scheduler [21]. We optimized the network in 20 epochs with focal loss function [20] and evaluate with the average of F_1 score across 12 categories. The results of our experiment are shown in Table 3. The result with single Transformer (T_1) are slightly better than the whole system (including both T_1 and T_2) which need more additional experiments on the fusion mechanism to verify the effectiveness of T_2.

In total, 38 Teams participate in this sub-challenge in which 19 Teams submitted and 8 Teams better than the baseline. Our method got 49.04 of the average F1 Score. We ranked 3rd in this competition's task.

Table 2. The comparison with previous works on valence arousal estimation with Affwild2 validation set.

Method	Feature	$\mathcal{P}_{\text{V}}$	$\mathcal{P}_{\text{A}}$	$\mathcal{P}_{\text{VA}}$
Baseline [10]	ResNet	0.31	0.17	0.24
Deng et al. [4]	MobilefaceNet + MarbleNet	0.442	0.546	0.494
Zhang et al. [30]	Expression Embedding [31]	**0.488**	0.502	0.495
GRU + Transformer	RegNet	0.391	0.565	0.478
GRU	k-fold	0.432	0.575	0.504
GRU + Attention	k-fold	0.437	**0.576**	**0.507**

Table 3. The comprasion with prior methods on Affwild2 for action unit detection. A, V, T are indicate for audio, visual, and text, respectively.

Method	Feature	F_1-val	F_1-test
Baseline [10]	V	0.39	0.365
NetFuxiVirHuman [32]	A+V+T	-	**0.499**
SituTech [8]	V	0.731	0.498
STAR-2022 [28]	A+V	0.518	0.488
Our method	RegNet	0.533	0.479
Our method - weighted loss	V	0.539	0.483
Our method - only T_1	V	**0.544**	0.483
Use T_1 as target for T_2	V	0.541	0.488
Voting fusion	-	-	**0.490**

5. Conclusions

This aspect of the research suggested that utilizing features from deep learning representations to the Valence-Arousal Estimation sub-challenge of ABAW3 2022. To extract information over time, GRU is used for sentiment analysis. To enhance the advantages of GRU, we have connected the local attention mechanism into our model. The CCC function was used to predict arousal/valence. Experimental results show that our proposed model outperforms the baseline method. We demonstrated that our results were better when combined GRU and local attention. Furthermore, we introduced a new feature extracted from multimodal by combining folds. We showed that the new features improve not only speed but also accuracy. Moreover, we used the simple Transformer block for the Action Unit Detection task. Two branches of the feature are conducted. Specifically, the second branch experimented with expanding to higher dimensions and then compressing to the original dimensions with the aim of improving the robustness of the model.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NRF-2020R1A4A1019191) and Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2021R1I1A3A04036408).

References

[1] Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer, 2020. 3

[2] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, 2014. 1, 2, 3

[3] Didan Deng, Zhaokang Chen, and Bertram E Shi. Multitask emotion recognition with incomplete labels. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 592–599. IEEE, 2020. 2

[4] Didan Deng, Liang Wu, and Bertram E Shi. Iterative distillation for better uncertainty estimates in multitask emotion recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3557–3566, 2021. 2, 4, 5

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1

[7] 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. 3

[8] Wenqiang Jiang, Yannan Wu, Fengsheng Qiao, Liyu Meng, Yuanyuan Deng, and Chuanhe Liu. Facial action unit recognition with multi-models ensembling. *arXiv preprint arXiv:2203.13046*, 2022. 5

[9] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4

[10] Dimitrios Kollias. Abaw: Valence-arousal estimation, expression recognition, action unit detection & multi-task learning challenges. *arXiv preprint arXiv:2202.10659*, 2022. 1, 4, 5

[11] D Kollias, A Schulc, E Hajiyev, and S Zafeiriou. Analysing affective behavior in the first abaw 2020 competition. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)(FG)*, pages 794–800. 2

[12] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Face behavior a la carte: Expressions, affect and action units in a single network. *arXiv preprint arXiv:1910.11111*, 2019. 2

[13] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Distribution matching for heterogeneous multitask learning: a large-scale face study. *arXiv preprint arXiv:2105.03790*, 2021. 2

[14] Dimitrios Kollias, Panagiotis Tzirakis, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, Björn Schuller, Irene Kotsia, and Stefanos Zafeiriou. Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. *International Journal of Computer Vision*, pages 1–23, 2019. 2

[15] Dimitrios Kollias and Stefanos Zafeiriou. Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface. *arXiv preprint arXiv:1910.04855*, 2019. 2

[16] Dimitrios Kollias and Stefanos Zafeiriou. Affect analysis in-the-wild: Valence-arousal, expressions, action units and a unified framework. *arXiv preprint arXiv:2103.15792*, 2021. 2

[17] Dimitrios Kollias and Stefanos Zafeiriou. Analysing affective behavior in the second abaw2 competition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3652–3660, 2021. 2

[18] Felix Kuhnke, Lars Rumberg, and Jörn Ostermann. Two-stream aural-visual affect analysis in the wild. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 600–605. IEEE, 2020. 2

[19] I Lawrence and Kuei Lin. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, pages 255–268, 1989. 4

[20] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 4

[21] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 4

[22] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*, 2015. 3

[23] Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing network design spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10428–10436, 2020. 1, 2, 3

[24] Jack Rae and Ali Razavi. Do transformers need deep long-range memory? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, July 2020. Association for Computational Linguistics. 3

[25] Aurko Roy*, Mohammad Taghi Saffar*, David Grangier, and Ashish Vaswani. Efficient content-based sparse attention with routing transformers, 2020. 3

[26] 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. 2, 3

[27] Manh Tu Vu, Marie Beurton-Aimar, and Serge Marchand. Multitask multi-database emotion recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3637–3644, 2021. 2

[28] Lingfeng Wang, Shisen Wang, and Jin Qi. Multi-modal multi-label facial action unit detection with transformer. *arXiv preprint arXiv:2203.13301*, 2022. 5

[29] Stefanos Zafeiriou, Dimitrios Kollias, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, and Irene Kotsia. Aff-wild: Valence and arousal 'in-the-wild'challenge. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*, pages 1980–1987. IEEE, 2017. 2

[30] Wei Zhang, Zunhu Guo, Keyu Chen, Lincheng Li, Zhimeng Zhang, and Yu Ding. Prior aided streaming network for multi-task affective recognitionat the 2nd abaw2 competition. *arXiv preprint arXiv:2107.03708*, 2021. 2, 4, 5

[31] Wei Zhang, Xianpeng Ji, Keyu Chen, Yu Ding, and Changjie Fan. Learning a facial expression embedding disentangled from identity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6759–6768, 2021. 5

[32] Wei Zhang, Zhimeng Zhang, Feng Qiu, Suzhen Wang, Bowen Ma, Hao Zeng, Rudong An, and Yu Ding. Transformer-based multimodal information fusion for facial expression analysis. *arXiv preprint arXiv:2203.12367*, 2022. 5