Multitask Multi-database Emotion Recognition

Manh Tu VU
LaBRI
Bordeaux University
33405 Talence, France
manh.vu@labri.fr

Marie BEURTON-AIMAR
LaBRI
Bordeaux University
33405 Talence, France
beurton@labri.fr

Serge MARCHAND
Lucine
Quai Lawton, G4
33300 Bordeaux, France
smarchand@projet-lucine.org

Abstract

This work has been initiated for the 2nd Affective Behavior Analysis in-the-wild (ABAW 2021) competition. We train a unified deep learning model on multi-databases to perform two tasks: seven basic facial expressions prediction and valence-arousal estimation. Since these databases do not contain labels for all the two tasks, we have applied the distillation knowledge technique to train two networks: one teacher and one student model. Both of these models are based on CNN-RNN hybrid architecture. The student model will be trained using both ground truth labels and soft labels derived from the pretrained teacher model. During the training, we have added one more task, which is the combination of the two mentioned tasks, for better exploiting inter-task correlations. We also exploit the sharing videos between the two tasks of the AffWild2 database that is used in the competition for further improving the performance of the network. Experiment results show that with these improvements, our model has reached the performance on par with the state of the art on the test set of the competition. Code and pretrained model are publicly available at https://github.com/glmanhtu/multitask-abaw-2021

1. Introduction

Emotion recognition and analysis are the crucial parts of many applications and human-computer interactive systems, especially in health care and medical fields [24, 1] since it is directly related to the health state of a patient. As results, more and more works have been conducted to try to analyse human emotions and behaviours [23, 22, 26]. In the same sense, the 2nd Affective Behavior Analysis in-the-wild (ABAW 2021) competition by Kollias *et al*. [9, 11, 15, 14, 10, 12, 27, 8] provides a large-scale dataset Aff-Wild2 [13] for analysing human emotion in-the-wild settings. This dataset includes videos with annotations for three tasks including: valence-arousal estimation, action unit (AU) detection, and seven basic facial expression classification. Valence represents how positive or negative an emotional state is, whereas arousal describes how passive or active it is. The seven basic facial expressions include neutral, anger, disgust, fear, happiness, sadness, and surprise. AUs are the basic actions of individuals or groups of muscles for portraying emotions.

In this paper, we focus on two tasks: seven basic facial expressions classification and valence-arousal estimation. Inspired by the multitask training with incomplete label method from Deng *et al*. [3] we propose a method to further exploit the inter-task correlations between these two tasks. Similar to Deng *et al*. [3] we apply the distillation knowledge technique to train two multitask models: a teacher model and a student model. The student model will be trained using both ground truth labels and soft labels derived from the pretrained teacher model. However, instead of treating each task independently when training teacher model as in [3], we add one more task to the training process, which is the combination of the two tasks above to train the network using data coming from AffectNet database [18], in which contains labels for both of the two tasks. Since the data for this task has been annotated for both seven basic expressions and valence-arousal, this task will play the role of guiding the training, i.e. rebalancing the gradient backpropagation of the first two tasks and exploiting the inter-task correlations between the training tasks. Apart from that, taking into account that there are a huge number of videos that are annotated for both seven basic facial expressions and valence-arousal labels in the Affwild2 database, we integrate this information into the student model's training process for better exploiting inter-task correlations. With these improvements, our model has reached the performance on par with the state of the art on the test set of the official dataset Affwild2 of the competition.

2. Related Works

The challenges of human affect analysis have attracted lots of research efforts, especially in in-the-wild settings. In this section, we will briefly introduce some works related to this problem. Pan *et al*. [19] propose a framework to aggregate spatial and temporal convolutional features across the entire extent of a video. Deng *et al*. [3] apply distillation knowledge technique to train their multitask model using data with incomplete labels. Kuhnke *et al*. [17] propose a two stream aural-visual network for multi-task training. Gera *et al*. [4] propose a spatio-channel attention network, which is able to extract local and global attentive features for classifying facial expressions. Kollias *et al*. [11] proposed FaceBehaviorNet for large-scale face analysis, by jointly learning multiple facial affective behaviour tasks and a distribution matching approach. Wei Zhang *et al*. [29] propose a heuristic that the three emotion representations including: categorical emotions, action units and valence-arousal are intrinsically associated with each other. They try to exploit these hierarchical relationships by developing a prior aided streaming network for multitask prediction. Wang *et al*. [25] extend the work of Kuhnke *et al*. [17] by improving the preprocessing method of rendering mask and applying mean teacher model for utilizing the unlabeled data. Su Zhang *et al*. [28] propose an audio-visual spatial-temporal deep neural network with attention mechanism for valence-arousal estimation.

3. Methodology

In this section, we introduce our multitask multi-databases training method. Frame images are extracted from video and fed into a Convolution Neural Network (CNN) to train for analysing human's emotion in-the-wild. Then, features extracted from this network will go through a Recurrent Neural Network (RNN) to capture temporal information and finally perform both the seven basic facial expressions classification and valence-arousal estimation. Because in our dataset, we do not always have all labels for all of our tasks, we have applied the multitask training with missing labels method that is described in [3] with some enhancements, which is described in the sections below.

3.1. Data Imbalancing

Similar to [3], we also have used some external datasets to address the data imbalance problem in the Affwild2 dataset, e.g. most of the frames inside the Affwild2 dataset have their valence value in the range of [0-0.4]. The external datasets are including Expression in-the-Wild (ExpW) dataset [31] for expression classification and AFEW-VA dataset [16] for valence-arousal estimation. After merging these datasets, we have applied the same dataset balancing protocol as [3] to improve the balance of the dataset.

Different from [3], as we have mentioned earlier, in this preliminary work we perform only two tasks: seven basic facial expressions prediction and valence-arousal estimation. Apart from that, we also want to include the AffectNet database [18] into the training, since this database is annotated for both seven basic expressions and valence-arousal are available. After this step, for the training process, our dataset is including three parts:

Mixed EXPR The mixing set of the AffWild 2 (expressions part) and ExpW datasets for seven basic expressions. This dataset has no information about valence and arousal.

Mixed VA The mixing set of the AffWild 2 (valence-arousal part) and AFEW-VA datasets for valence and arousal. This dataset has no information about the seven basic expressions.

Affect EXPR_VA The AffectNet dataset, for both seven basic expressions and valence-arousal.

Corresponding to these three dataset's parts are the three training tasks $\mathscr{T} \in \{1, 2, 3\}$, which are including: expression classification (EXPR), valence-arousal estimation (VA) and the mixing of these two tasks (EXPR_VA). One can note that even though we have three training tasks, our model has only two outputs, which are EXPR and VA, since the last training task reuses these two outputs for computing loss.

3.2. Multitask training with missing labels

Here we describe the formulars that are used to train our teacher and student models. Let (X, Y) be the training dataset, where X is a set of input vectors and Y is a set of ground truth training labels. Since our dataset contains three parts including: *Mixed EXPR*, *Mixed VA* and *Affect EXPR_VA*, therefore $(X, Y) = \{(X^{(i)}, Y^{(i)})\}_{i=1}^{3}$. For convenience of notation, we assume each subset i includes an equal number N of instances within a batch, i.e $(X^{(i)}, Y^{(j)}) = \{(x^{(i,n)}, y^{(i,n)})\}_{n=1}^{N}$ where n indexes the instance. Because the data from the last set *Affect EXPR_VA* is including both EXPR and VA annotations, we denote 3_{expr} and 3_{va} as the EXPR annotation and the VA annotation of this set, respectively. For example, instance $x^{(3,1)}$ belongs to *Affect EXPR_VA* dataset and has two annotations: $y^{(3_{expr},1)}$ and $y^{(3_{va},1)}$

The inputs for all instances have the same dimensionality, regardless of task. However, the ground truth labels for different tasks have different dimensionality. The label for the first task (EXPR) is $y^{(1)} \in \{0, 1\}^7$. The label for the second task (VA) is $y^{(2)} \in [-1, 1]^2$. The label for the last task (EXPR_VA) is the mixed of the two tasks above.

Similar to [3], we also apply the two steps training for capturing inter-task correlations. We train a single teacher

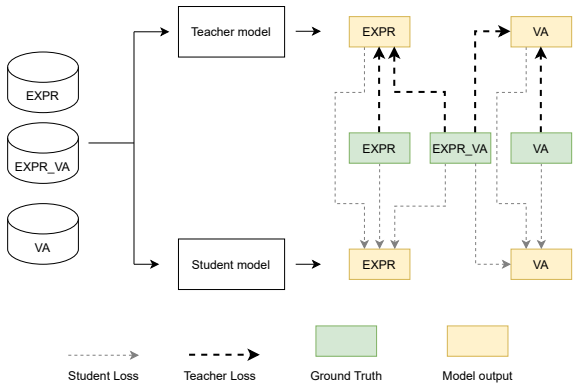

Figure 1. The overview of our multitask training with missing labels.

model using only the ground truth labels in the first step. In the second step, we replace the missing labels with soft labels derived from the outputs of the teacher model. We then use the ground truth and soft labels to train a single student model. Different from [3], we do not train multi student models for model ensemble because this approach is too costly in term of computation and the gain in performance is not significant. The overview of our network can be seen in Fig 1 and the architecture of our model is in Fig 2.

To be in the same line with [3] in the sense of notation, we also denote the output of our multitask network by $f_\theta^{(i)}(\cdot)$ where θ contains the model parameters of either teacher model or student model, and $i \in \{1, 2\}$ indicates the current task. For example, $f_\theta^{(1)}(x^{(3)})$ indicates the output of the network for task 1 (EXPR) for an instance in the *Affect EXPR_VA* set. To avoid clutter, we will often refer to the output of the teacher network on task i by $t^{(i)}$ irrespective of what the input label is, i.e. $t^{(i)} = f_\theta^{(i)}(x^{(j)})$ for some $j \in \{1, 2\}$ and similarly to the output of the student network on task i by $s^{(i)}$.

Regarding the objective loss functions, similar to [3], we also treat the problem of expression classification as a multiclass classification problem, and the problem of valence-arousal estimation as a combination of multiclass classification and regression problem. We will use the same Softmax Function SF, the Cross Entropy function CE and the Concordance Correlation Coefficient function CCC, which have already been defined in [3].

3.2.1 Supervision loss functions

Here we denote the loss functions that are used for optimizing our models parameters with the supervision of the ground truth labels for each of our training tasks.

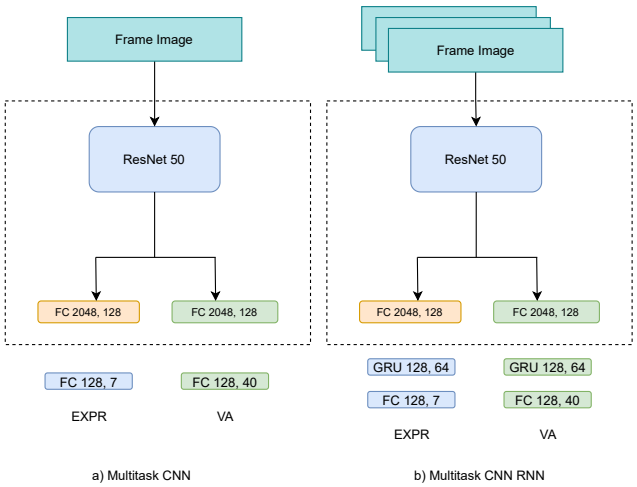

Figure 2. The multitask CNN (a) and CNN-RNN (b) architectures, The two architectures share the same ResNet spatial feature extractor shown in the dashed box.

EXPR task The supervision loss for the samples from the *Mixed EXPR* set is denoted as:

$$\mathscr{L}^{(1)}(y^{(1)}, t^{(1)}) = CE\left(y^{(1)}, SF(t^{(1)}, 1)\right) \quad (1)$$

VA task The supervision loss for the samples from the *Mixed VA* set is denoted as:

$$\mathscr{L}^{(2)}(y^{(2)}, t^{(2)}) = \sum_{i=1}^{2}\left\{CE\left(onehot(y_i^{(2)}), SF(t_i^{(2)}, 1)\right) + \frac{1}{B}\left(1 - CCC(y_i^{(2)}, t_i^{(2)})\right)\right\} \quad (2)$$

EXPR_VA task For the samples from Affect EXPR_VA set, since the samples of this set are annotated for both VA and EXPR, the supervision loss for this task is denoted as:

$$\mathscr{L}^{(3)}(y^{(3)}, t^{(3)}) = CE\left(y^{(3_{expr})}, SF(f_{\theta_t}^{(1)}(x^{(3)}), 1)\right) + \sum_{i=1}^{2}\left\{CE\left(onehot(y_i^{(3_{va})}), SF(f_{\theta_t i}^{(2)}(x^{(3)}), 1)\right) + \frac{1}{B}\left(1 - CCC(y_i^{(3_{va})}, f_{\theta_t i}^{(2)}(x^{(3)}))\right)\right\} \quad (3)$$

From this equation, we can see that for each sample of the dataset, we calculate the loss for both EXPR and VA tasks. Therefore, the gradient backpropagation derived from this task's loss is the most accurate one compared to the other two tasks. Because we can see that the loss of the EXPR task can be used to adjust the model's parameters for better EXPR prediction, but it has absolutely no idea of

whether the VA estimation is correct or not, and the same goes for the loss of the VA task. Therefore, the EXPR_VA task plays the role of guiding the training process, i.e. rebalance the gradient backpropagation for the whole training process. In the same time, since this task compute the loss for both EXPR and VA tasks, it can exploit the inter-task correlations, which typically can help the network for better prediction.

3.2.2 Distillation loss functions

Here we denote the loss functions that are used to optimise our student model parameters with the supervision of both the ground truth labels (hard targets) and the pre-trained teacher model's outputs (soft targets) for each of our training tasks. Similar to [3], we use the KL divergence to measure the difference between two probability distributions (output of teacher model and student model). The KL divergence of two vectors p and q is denoted as: $KL(p,q) = \sum_i p_i log\left(\frac{p_i}{q_i}\right)$.

EXPR task Distillation loss for the samples from the *Mixed EXPR* set:

$$\mathscr{H}^{(1)}(t^{(1)}, s^{(1)}) = KL\left(SF(t^{(1)}, T), SF(s^{(1)}, T)\right) \quad (4)$$

VA task Distillation loss for the samples from the *Mixed VA* set:

$$\mathscr{H}^{(2)}(t^{(2)}, s^{(2)}) = \sum_{i=1}^{2} KL\left(SF(t_i^{(2)}, T), SF(s_i^{(2)}, T)\right) \quad (5)$$

EXPR_VA task Distillation loss for the samples from the *Affect EXPR_VA* set is the combination of the EXPR and VA distillation losses, which is denoted as:

$$\begin{aligned}\mathscr{H}^{(3)}(t^{(3)}, s^{(3)}) = \\ KL\left(SF\left(f_{\theta_t}^{(1)}(x^{(3)}), T\right), SF\left(f_{\theta_s}^{(1)}(x^{(3)}), T\right)\right) \\ + \sum_{i=1}^{2} KL\left(SF\left(f_{\theta_t i}^{(2)}(x^{(3)}), T\right), SF\left(f_{\theta_s i}^{(2)}(x^{(3)}), T\right)\right)\end{aligned} \quad (6)$$

3.2.3 Batch-wise loss functions

Given a batch of data $(X, Y) = \{\{(x^{(i,n)}, y^{(i,n)})\}_{n=1}^{N}\}_{i=1}^{3}$, the parameters of teacher network and student networks are denoted as θ_t and θ_s, respectively. Since our last dataset *Affect EXPR_VA* contains annotation for both EXPR and VA, therefore, when $i = 3$ then $y^{(3,n)}$ contains both $y^{(3_{expr},n)}$ and $y^{(3_{va},n)}$.

The training teacher loss is denoted as:

$$\mathscr{F}_t(X, Y, \theta_t) = \sum_{i=1}^{3}\sum_{n=1}^{N} \mathscr{L}^{(i)}\left(y^{(i,n)}, f_{\theta_t}^{(i)}(x^{(i,n)})\right) \quad (7)$$

The student loss of a sample x with ground truth y from dataset i with $i \in \{1, 2, 3\}$ is denoted as:

$$\begin{aligned}\mathscr{G}_i(x, y, \theta_t, \theta_s) = \lambda \times \mathscr{L}^{(i)}\left(y, f_{\theta_s}^{(i)}(x)\right) \\ + \ (1-\lambda) \times \mathscr{H}^{(i)}\left(f_{\theta_t}^{(i)}(x), f_{\theta_s}^{(i)}(x)\right)\end{aligned} \quad (8)$$

Similar to [3], we also use the parameter λ to weight the supervision loss versus the distillation loss. The λ parameter is set to 0.6 to weight the ground truth slightly more than the soft labels.

The student loss is denoted as:

$$\begin{aligned}\mathscr{F}_t(X, Y, \theta_t, \theta_s) = \sum_{n=1}^{N} \mathscr{G}_3\left(x^{(3,n)}, y^{(3,n)}, \theta_t, \theta_s\right) \\ + \sum_{i=1}^{2}\sum_{n=1}^{N}\Bigg\{\mathscr{G}_i\left(x^{(i,n)}, y^{(i,n)}, \theta_t, \theta_s\right) \\ + \sum_{j\neq i} \mathscr{H}^{(j)}\left(f_{\theta_t}^{(j)}(x^{(j,n)}), f_{\theta_s}^{(j)}(x^{(j,n)})\right)\Bigg\}\end{aligned} \quad (9)$$

As we have mentioned earlier, there are 164 videos that are annotated for both EXPR and VA in the Affwild2 database. Instead of treating all of these videos as if they are annotated with only one label like [3], we check if the given video frame has been annotated with one or both EXPR and VA labels. Then, we compute the objective loss of the secondary task using the distillation loss alone or supervision loss plus distillation loss, respectively. Particularly, the student loss for taking into account this characteristic is denoted as:

$$\begin{aligned}\mathscr{F}_t(X, Y, \theta_t, \theta_s) = \sum_{n=1}^{N} \mathscr{G}_3\left(x^{(3,n)}, y^{(3,n)}, \theta_t, \theta_s\right) \\ + \sum_{i=1}^{2}\sum_{n=1}^{N}\left\{\mathscr{G}_i\left(x^{(i,n)}, y^{(i,n)}, \theta_t, \theta_s\right)\right. \\ \left. + \sum_{j\neq i}\begin{cases}\mathscr{H}^{(j)}\left(f_{\theta_t}^{(j)}(x^{(j,n)}), f_{\theta_s}^{(j)}(x^{(j,n)})\right), & \text{if } y^{j,n} \text{ is NA} \\ \mathscr{G}_j\left(x^{(j,n)}, y^{(j,n)}, \theta_t, \theta_s\right), & \text{otherwise}\end{cases}\right\}\end{aligned} \quad (10)$$

3.3. Frame images analysis

For the video's frame images, face images with the size of 112×112 pixels are aligned and extracted from each frame. Then, we use these images to train a CNN model using the method mentioned in Section 3.2. For this CNN model, we have selected the ResNet 50 [5] architecture as

base network and added two head layers corresponding to the two outputs of the model: EXPR and VA (see Figure 2). During training, we have applied some image-wise augmentation process with some filters to improve the performance of the model. These filters are including: random image translation [21] and random image horizontal flip.

3.4. Temporal information exploitation

Once the CNN student model has been trained, we use this model to extract features from each video frame. Then, we group these features together to form a new dataset ds of feature's sequences with the sequence length of 32 frames per sequence. Finally, we fed data from this new dataset ds into a bidirectional RNN network for exploiting temporal information, as well as predicting EXPR and VA. For this RNN network, we have selected the Gated Recurrent Units (GRU) architecture [2] as it has been proven to be efficient in remembering long-term dependencies. Regarding this GRU model's parameters, we also use the training method in Section 3.2 to train them. During the training, we have used the same augmentation process with filters that are mentioned in Section 3.3 but in sequence level.

4. Experiments and Results

4.1. Implementation details

The whole network system is implemented using PyTorch framework [20]. During the training phase, Adam optimizer [7] was employed with the initial learning rate is set to $1e^{-4}$. The maximum number of epochs is 40 and the training process will stop when there is no improvement after five consecutive epochs. The number of batch size for the CNN part of the network is set to 64. For RNN network, the batch size is 16. The training and validating processes were performed on an Intel Workstation machine with a NVIDIA Gerforce RTX 2080 Ti 11G GPU.

4.2. Results

Here we report the results of different experiments to demonstrate the effectiveness of each of our changes comparing to the original method [3]. For the evaluation metrics, we use the same criterion as outlined in [9]. Valence and Arousal estimation is based on the mean Concordance Correlation Coefficient (CCC). The seven basic expressions classification is measured by $0.67 \times F_1$ score $+ 0.33 \times$ total accuracy. For each of our experiments, we run it 10 times and report the mean of the evaluation results on the Validation set of the AffWild2 dataset.

Table 1 shows the performance of the teacher network when training using Equation 7 with only the first two tasks ($\mathscr{T} \in \{1, 2\}$) and with all the three tasks ($\mathscr{T} \in \{1, 2, 3\}$). From this table, we can see that when training with only two tasks, our model has already outperformed the baseline results of the competition. When we add the third task EXPR_VA into the training process

($\mathscr{T} \in \{1, 2, 3\}$), we can see that the performance of both EXPR and Valence have increased quite a lot, especially the later with 17% of improvement. Despite of having a slightly decreasing in term of Arousal (about 2%), the performance of the network has been improved in overall by a large margin, compared to the model trained without the EXPR_VA task.

Table 1. Performance results of the teacher CNN models on the validation set of the Affwild2 database. The baseline results are provided by the ABAW 2021 competition organiser.

Method	EXPR	Valence	Arousal
Baseline	0.366	0.23	0.21
Multitask $\mathscr{T} \in \{1, 2\}$	0.498	0.374	**0.407**
Multitask $\mathscr{T} \in \{1, 2, 3\}$	**0.513**	**0.438**	0.398

After training the teacher model, we train student models with the supervision of both ground truth and the pretrained teacher model using Equation 9 for the case of not using the shared annotations (No sharing), and using Equation 10 for the case of using the shared annotations (With sharing). The results are shown in Table 2. From this table, it can be seen that the performance of the model trained using the shared annotations (With sharing) is better than the one trained without using it (No sharing). This results indicate the importance of exploiting the sharing annotations in the database. Once the student model is trained, we use this

Table 2. Performance results of the student CNN models on the validation set of the Affwild2 database. The student models are trained using all three tasks $\mathscr{T} \in \{1, 2, 3\}$.

Method	EXPR	Valence	Arousal
No sharing	0.513	**0.472**	0.412
With sharing	**0.525**	0.471	**0.421**

Table 3. Performance results of the CNN + GRU model. Both teacher and student models are trained using all three tasks $\mathscr{T} \in \{1, 2, 3\}$.

Method	EXPR	Valence	Arousal
Teacher model	0.555	0.523	0.543
Student model	**0.555**	**0.526**	**0.551**

CNN model to extract features to train GRU network for exploiting temporal information. We train a teacher model using Equation 7 and a student model using Equation 10. Table 3 shows the results of these models. From this table,

Table 4. Comparison with other works on the test set of the Affwild2 database

Method	Expression			CCC		
	F_1	Acc	Criterion	Valence	Arousal	Mean
Top entries to ABAW 2020:						
ICT-VIPL [30]	0.287	0.652	0.408	0.361	0.408	0.385
NISL2020 [3]	0.270	0.680	0.405	0.440	**0.454**	**0.447**
TNT [17]	**0.398**	**0.734**	**0.509**	**0.448**	0.417	0.433
Top entries to ABAW 2021:						
FlyingPigs [28]	–	–	–	0.463	0.492	0.478
STAR [25]	0.476	0.732	0.560	0.478	**0.498**	0.488
Netease Fuxi Virtual Human [29]	**0.763**	**0.807**	**0.778**	0.486	0.495	0.491
CPIC-DIR2021 [6]	0.683	0.771	0.712	–	–	–
NISL2021 (no publication)	0.431	0.654	0.505	**0.533**	0.454	**0.494**
Our model	0.351	0.668	0.456	0.505	0.475	0.490

we can see that: the performance of the student model is equivalent to the performance of the teacher model in the task of EXPR prediction and better than the teacher model in all the other cases. When we compare the CNN + GRU model with the CNN model alone (in Table 2), the former model outperformed the latter by a large margin.

4.3. Comparison with State of the art

Here we compare the performance of our model with the state of the art on the test set of Affwild2 dataset. In this 2nd challenge, the database has been updated by adding more videos and labels for the AU detection task, but since the data for EXPR recognition task and VA estimation task are almost unchanged, we are still able to compare the performance of our model with the works on the previous ABAW 2020 challenge [9].

Table 4 shows the comparison results between the works on Affwild2 database. One can note that these results are the results of the test set of the database and have been computed by the organiser of the competition for fair comparison. For the prior works on this dataset (ABAW 2020) we have: ICT-VIPL team [30] with their M^3T model, NISL2020 team [3] with their multitask model trained on multiple datasets with incomplete labels and TNT team [17] with their two streams aural-visual network. For the top entries to the challenge (ABAW 2021), we have: FlyingPigs team [28] with their Audio-visual Attentive Fusion model, STAR team [25] with their multitask aural-visual model, Netease Fuxi Virtual Human team [29] with their Prior Aided Streaming network, CPIC-DIR2021 team [6] with their multitask multimodal method for detecting AUs and classifying facial expressions. Apart from that, we also have the NISL2021 team, but without publicly available article. From this table, we can see that our model is significantly outperformed the original model (which we have adapted from) of the NISL2020 team [3] in both of the two tasks and outperformed all other prior works in term of VA estimation. This results are clearly showing that our changes and improvements in the approach have improved the overall performance of the model significantly.

For the top entries to ABAW 2021 competition, our model has reached the third place in the VA estimation track leaderboard, over 40 teams that have participated in this track of the challenge. Taking into account that the difference between the mean VA of our model and the best model (NISL2021) is only 0.82%, we can say that our model has reached the same performance level compared to the state of the art in term of VA estimation on the official Affwild2 database of the competition. In term of EXPR recognition, we are in the 8th place in the EXPR track leaderboard, over 55 teams that participated in this track of the challenge. The reason for this results could be because we are not using AUs annotations as the other competitors. e.g. the top two teams in this track: Netease Fuxi Virtual Human and CPIC-DIR2021, they are both trying to detect AUs beside recognise EXPR and have achieved a good performance compared to the others. There could be a strong link between action units and facial expressions that need to be identified in the future works.

5. Conclusion

In this paper, we have presented a method to optimise the multitask training with incomplete labels approach. On top of the original method based on teacher-student architecture, we have added a new task to train the deep neural network on a dataset that contains both seven basic expressions and valence-arousal values for better exploiting the inter-task correlations between the two tasks. In the same time, we have exploited the shared annotations inside the Affwild2 database during the training process of the student model. With these improvements, we have obtained a model that is on par with state of the art in term of valence and arousal estimation on the test set of the Affwild2

database. In future work, we will investigate about the link between action units and facial expressions, which could be the key to further improve the performance of both the facial expressions classification and valence-arousal estimation tasks.

References

[1] Rasha Al-Eidan, Hend Al-Khalifa, and AbdulMalik Al-Salman. Deep-learning-based models for pain recognition: A systematic review. *Applied Sciences*, 10:5984, 08 2020. 1

[2] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation, 2014. 5

[3] Didan Deng, Zhaokang Chen, and Bertram E. Shi. Multitask emotion recognition with incomplete labels, 2020. 1, 2, 3, 4, 5, 6

[4] Darshan Gera and S Balasubramanian. Affect expression behaviour analysis in the wild using spatio-channel attention and complementary context information, 2020. 2

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. 4

[6] Yue Jin, Tianqing Zheng, Chao Gao, and Guoqiang Xu. A multi-modal and multi-task learning method for action unit and expression recognition, 2021. 6

[7] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. 5

[8] Dimitrios Kollias, Irene Kotsia, Elnar Hajiyev, and Stefanos Zafeiriou. Analysing affective behavior in the second abaw2 competition. *arXiv preprint arXiv:2106.15318*, 2021. 1

[9] 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. 1, 5, 6

[10] 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. 1

[11] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Distribution matching for heterogeneous multitask learning: a large-scale face study. *arXiv preprint arXiv:2105.03790*, 2021. 1, 2

[12] 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. 1

[13] Dimitrios Kollias and Stefanos Zafeiriou. Aff-wild2: Extending the aff-wild database for affect recognition, 2019. 1

[14] Dimitrios Kollias and Stefanos Zafeiriou. Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface. *arXiv preprint arXiv:1910.04855*, 2019. 1

[15] 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. 1

[16] Jean Kossaifi, Georgios Tzimiropoulos, Sinisa Todorovic, and Maja Pantic. Afew-va database for valence and arousal estimation in-the-wild. *Image and Vision Computing*, 65:23–36, 2017. 2

[17] Felix Kuhnke, Lars Rumberg, and Jorn Ostermann. Two-stream aural-visual affect analysis in the wild. *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, Nov 2020. 2, 6

[18] Ali Mollahosseini, Behzad Hasani, and Mohammad H. Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, Jan 2019. 1, 2

[19] Xianzhang Pan, Guoliang Ying, Guodong Chen, Hongming Li, and Wenshu Li. A deep spatial and temporal aggregation framework for video-based facial expression recognition. *IEEE Access*, 7:48807–48815, 2019. 2

[20] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. 5

[21] Simone Porcu, Alessandro Floris, and Luigi Atzori. Evaluation of data augmentation techniques for facial expression recognition systems. *Electronics*, 9(11), 2020. 5

[22] Anvita Saxena, Ashish Khanna, and Deepak Gupta. Emotion recognition and detection methods: A comprehensive survey. *Journal of Artificial Intelligence and Systems*, 2(1):53–79, 2020. 1

[23] Nicu Sebe, Ira Cohen, Theo Gevers, and Thomas S Huang. Multimodal approaches for emotion recognition: a survey. In *Internet Imaging VI*, volume 5670, pages 56–67. International Society for Optics and Photonics, 2005. 1

[24] Jérôme Thevenot, Miguel Bordallo Lopez, and Abdenour Hadid. A survey on computer vision for assistive medical diagnosis from faces. *IEEE Journal of Biomedical and Health Informatics*, PP:1–1, 10 2017. 1

[25] Lingfeng Wang and Shisen Wang. A multi-task mean teacher for semi-supervised facial affective behavior analysis, 2021. 2, 6

[26] Philipp Werner, Daniel Lopez-Martinez, Steffen Walter, Ayoub Al-Hamadi, Sascha Gruss, and Rosalind Picard. Automatic recognition methods supporting pain assessment: A survey. *IEEE Transactions on Affective Computing*, page 1–1, 2019. 1

[27] 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. 1

[28] Su Zhang, Yi Ding, Ziquan Wei, and Cuntai Guan. Audio-visual attentive fusion for continuous emotion recognition, 2021. 2, 6

[29] 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, 2021. 2, 6

[30] Yuan-Hang Zhang, Rulin Huang, Jiabei Zeng, Shiguang Shan, and Xilin Chen. m^3t: Multi-modal continuous valence-arousal estimation in the wild, 2020. 6

[31] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. From facial expression recognition to interpersonal relation prediction. *International Journal of Computer Vision*, 126(5):550–569, 2018. 2