Affective Dynamic based Technique for Facial Emotion Recognition (FER) to Support Intelligent Tutors in Education

Affective Dynamic based Technique for Facial Emotion Recognition (FER) to Support Intelligent Tutors in Education

Xingran Ruan[1 [0000−0001−9462−8816]], Charaka Palansuriya[1,2 [0000−0002−7130−5659]], and Aurora Constantin[1 [0000−0001−5352−5300]]

[1] The University of Edinburgh, Edinburgh, UK
{xruan, aurora.constantin}@ed.ac.uk
[2] Edinburgh Parallel Computing Centre, Edinburgh, UK
charaka@epcc.ed.ac.uk

Abstract. Facial expressions of learners are relevant to their learning outcomes. The recognition of their emotional status influences the benefits of instruction or feedback provided by the intelligent tutor in education. However, learners' emotions expressed during interactions with the intelligent tutor are mostly detected by self-reports of learners or judges who observe them in manually. The automated Facial Emotion Recognition (FER) task has been a challenging problem for intelligent tutors. The state-of-art automated FER methods target six basic emotions instead of learning-related emotions (e.g., neutral, confused, frustrated, and bored). Thus our research contributes to training a machine learning (ML) model to recognise learning-related emotions for intelligent tutors automatically, based on an Affective Dynamics (AD) model. We implement the AD model into our loss function (AD-loss) to fine-tune the ML model. In the test scenario, the AD-loss method improves the performance of state-of-art FER algorithms.

Keywords: Facial emotion recognition · Intelligent tutors · Epistemic emotion · Affective dynamics model.

1 Introduction

Self-regulated learning (SRL) is a complex educational construct that describes how students as regulated learners manage cognitive functions to achieve educational goals [15]. Under the framework of SRL, a considerable number of variables that influence learning outcomes are studied. In terms of emotions, research has shown that emotions can have positive and negative influences on learning [7, 9]. Significantly, the analysis of D'Mello et al. [7] points out that engagement, boredom, confusion, frustration, and curiosity are the most frequent learning-related emotions that predict learning. Researchers have proposed a number of intelligent tutors that are adaptive to learners' emotions to provide support during

their learning [3, 5]. Thus, emotion is a critical component of learning activity and FER of learners has become an essential task in education [9].

However, the state-of-art FER techniques in intelligent tutors are recognising emotions manually which limits to support learners adaptively [3, 8]. In [8], subjects are invited to learn about computer literacy with AutoTutor and researchers record subjects' facial expressions during their learning. The expressed affective statuses of subjects in videos are recognised by themselves in a post-learning session. In the study of Craig et al. [3], learners' emotions (e.g. boredom, flow, and confusion) during learning activities are recognised by five observers who were given a training session. On the other hand, the state-of-art automatic FER techniques mostly focus on six basic emotions (sadness, happiness, fear, anger, surprise and disgust) [12], but affective studies have shown that basic emotions occurred considerably rare when compared to epistemic emotions (boredom, confusion, frustration, flow/engagement) during learning. For example, D'Mello et al. [5] analysed studies of affective status of learning activity and claimed that these epistemic emotions were the most frequent. Thus our research is motivated to propose the FER technique to support intelligent tutors in recognising the epistemic emotions (neutral, confusion, frustration, and boredom) of learners during their learning process automatically.

In this paper, Section 2 introduces the affective dynamic model and explains using AD-Loss to fine-tune the ML model. Experimental results are presented in Section 3. Finally, conclusions and further works are provided in Section 4.

2 Proposed Method

In this section, we describe an AD-Loss framework which helps to improve and solve FER to support intelligent tutors to interpret learners' epistemic emotions. Our design is inspired by a model of affective dynamics during learning [6] which proposed that the experience of learners' emotions during SRL depends on their cognitive disequilibrium status. To imitate this affective dynamics model in our FER approach, we simultaneously take a pair of joint images as input to a pre-trained network then sequentially apply Multilayer perception (MLP), and an affective dynamic transition operation. The whole framework is shown in Fig.1.

2.1 Affective Dynamics Model

D'Mello and Graesser's Model of Affective Dynamics [6] aims to highlight the critical role of cognitive disequilibrium in the learning activity and explain the trajectory of cognitive-affective states that are spawned by cognitive disequilibrium. The model emphasizes that cognitive disequilibrium occurs when a learner is confronted with obstacles whilst trying to achieve learning goals (e.g., interruptions, contradictions, unexpected feedback etc..) and claimed that cognitive disequilibrium is the key to understanding the affective states that underlie complex learning. In the form of the transition network, a learner's affective state of confusion is one key signature of cognitive disequilibrium. Learners must engage

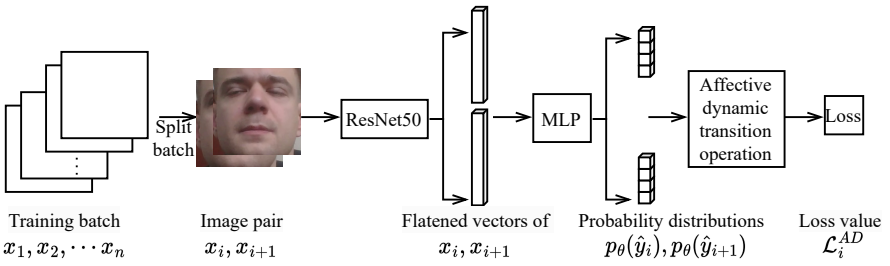

Fig. 1. AD-Loss framework. We split each incoming training batch of size n into $n-1$ pairs and feed them into pre-trained ResNet50 (pre-trained on ImageNet) and MLP. Each pair of frames will be converted to two probability distributions and be fed into affective dynamic transition operation to compute loss value.

in effortful problem-solving activities in order to resolve the impasse and resolve equilibrium. A hopeless confusion occurs when the impasse can not be resolved, the learner will get lost and experience frustration. Furthermore, persistent frustration may transition into boredom.

2.2 Affective Dynamics based FER Technique

Using a pre-trained deep convolutional neural network (DCNN) model through appropriate transfer learning is the main baseline for solving the FER task. Since training a DCNN model from scratch (with all weights randomly initialized) is a computing intensity workload, a DCNN which already been pre-trained on another large dataset (e.g. ImageNet [11]) can be fine-tuned on the target dataset for emotion classification. The following subsection describes implementing the proposed AD-Loss to fine-tune a pre-trained network for emotion classification.

We employ a pre-trained ResNet50 network [10] as the foundation of our work, incorporating flatten layers, dropout layers, and dense layers based on empirical evidence [1]. The first convolutional layer of ResNet50 takes the input of size $248 \times 248 \times 3$ for input RGB colour images with size 248×248; after successive convolutional blocks and pooling average layers, the output size is $9 \times 9 \times 2048$ (2048 volumes with the size of 9×9). Following the ResNet50 network, we add a flatten layer which converts the output into a linear vector of size $165,888 (= 9 \times 9 \times 2048)$. After that, we add three dense layers with lengths of 1000, 128, and 4 respectively behind the flatten layer and one dropout layer with the rate of 0.3 between each of the two dense layers to increase the robustness. The last dense layer with size 4 represents 4 different emotion statuses (neutral, confused, frustrated, and bored).

During fine-tuning on learners' facial expressions during learning $\mathcal{X} = \{x_1, x_2, \ldots x_n\}$ [3], we freeze all parameters in the pre-trained ResNet50 and update the parameters in three dense layers. The ResNet50-based network converts each

[3] $\mathcal{X}$ is a sequence of frames of facial emotion expressions of learners.

frame pair (x_i, x_{i+1}) to two conditional probability distributions $p_\theta(\hat{y}_i \mid x_i)$ and $p_\theta(\hat{y}_{i+1} \mid x_{i+1})$ [4], where $\hat{y}_i$ is the prediction of classifier given x_i. Each of these distributions is a four-dimensional vector, with each element (e.g. the element j) denoting the probability of the classifier in class j given input x_i. The θ are the parameters in the network and the final AD-Loss of x_i and x_{i+1} is $\mathcal{L}_i^{AD}$.

Now we will explain how to adapt the affective dynamics model in AD-Loss. Because the transition of a learner's emotional status mostly follows a partition that can transfer from equilibrium (neutral) to confusion when encountering impasses; transfer from confusion to frustration if the confronted impasses can not be resolved; furthermore, a persistent frustration can transition into boredom, we create a transition matrix $\mathcal{T}$ that describes the emotion status trajectory based on the affective dynamics model, see below. $\mathcal{T}(i, j)$ describes current emotional status $i \in [1, 4]$ transfers to the next (predicted) emotional status $j \in [1, 4]$ ($\mathcal{T}(i, j) = 0$ or 1). When $\mathcal{T}(i, j) = 0$, a penalty λ_{ad} will be aggregated on $\mathcal{L}_i^{AD}$, otherwise the penalty is zero. In order to implement the transition matrix into AD-Loss, we create the loss function $\mathcal{L}^{AD}$ as described in Eq.1 and set a parameter $\lambda_{ad} \in \mathcal{R}$ to control the penalty from affective dynamics transition operation.

$$\mathcal{T} = \begin{array}{c} \begin{array}{cccc} \text{Neutral} & \text{Confused} & \text{Frustrated} & \text{Boredom} \end{array} \\ \left[\begin{array}{cccc} 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 \end{array}\right] \end{array} \begin{array}{l} \text{Neutral} \\ \text{Confused} \\ \text{Frustrated} \\ \text{Boredom} \end{array}$$

$$\mathcal{L}_i^{AD} = \sum_{k=i}^{i+1} \mathcal{L}_k^{CE} + \lambda_{ad}(1 - \mathcal{T}(\lfloor argmax(p_\theta(\hat{y}_i)) \rfloor, \lfloor argmax(p_\theta(\hat{y}_{i+1})) \rfloor)) \quad (1)$$

$$\mathcal{L}_i^{CE} = p_\theta(\hat{y}_i)^T log(y_i) \quad (2)$$

We proceed to learn parameters θ^* with the following learning objective function for a sequence of emotional frames $\mathcal{X} = \{x_1, x_2, \ldots x_n\}$:

$$\theta^* = argmin \sum_{i=1}^{n-1} \mathcal{L}_i^{AD} \quad (3)$$

This objective function can be explained as: for ith frame in the training sequence, we select its successive frame (the $i+1$th), then search for an optimal θ^* to minimize the cross-entropy loss $\mathcal{L}^{CE}$ and the penalty of AD-Loss $\mathcal{L}^{AD}$.

3 Experimental Evaluations

To evaluate our method on recognising the epistemic emotions of learners during their learning activity, we tested our AD-Loss on PUZZLED [13] [5] and compared with other state-of-art FER methods that support intelligent tutors in education.

[4] We note $p_\theta(\hat{y}_i)$ as the abbreviation of $p_\theta(\hat{y}_i \mid x_i)$ in the rest of paper.

[5] The PUZZLED consists of 10 videos of students when they are watching educational videos (contains neutral, confused, frustrated, and boredom).

3.1 Results

In the pre-processing phase, we first recognised and cropped face areas from video frames. The transformed face was considered as the face region and we resized the face region to 248×248 pixels. In the test scenario, PUZZLED was split into training, and test sets by subject ID. We randomly chose 10% of subjects as the test set and the rest of the data as the training set in each test round. To increase the robustness of the network, we augmented the origin training dataset by performing colour shift (maximum value of 20), rotation, smoothing (maximum window size 5), mirror, and multi-crop [2] before the training.

In our test experiment, we 'plugged-in' AD-Loss (with $\lambda_{ad} = 1$) on three different state-of-art FER methods and compared the accuracy and F1 score with their origin. The results were summarized in Table 1. It can be seen from Table 1 that on average our method improved Cross-Entropy (CE) loss [1] 5% on accuracy and 9% on F1 score, respectively; improved Additive Margin Softmax (AM) loss [14] 6% on accuracy and 5% on F1 score, respectively; improved ArcFace (AF) loss [4] 2% on accuracy and 1% on F1 score, respectively. [6]

Table 1. Recognition Rate and F1 score of testing on PUZZLED. **CE**: Cross Entropy; **AD**: AD-Loss (we proposed); **AM**: Additive Margin Softmax; **AF**: ArcFace.

Base Model	Loss Function	Weighted F1 score (Ave)	Accuracy (Ave)
ResNet50	CE	0.4866	0.5299
ResNet50	**CE + AD**	**0.5296**	**0.5541**
ResNet50	AM	0.5194	0.6376
ResNet50	**AM + AD**	**0.5216**	**0.6419**
ResNet50	AF	0.5055	0.6027
ResNet50	**AF + AD**	**0.5320**	**0.6414**

4 Conclusion

In this paper, we proposed an ML-based FER technique and showed the affective dynamics model was able to support intelligent tutors to recognise learners' epistemic emotions during their learning. We implemented the affective dynamics model into the AD-Loss to fine-tune the pre-trained ResNet50 on the learners' emotion dataset. We tested our proposed technique on the PUZZLED dataset and compared it with other state-of-art FER techniques. It shows that AD-Loss improves the recognition rate of baseline FER methods. Our research introduced psychological aspects into the ML framework to recognise learners' emotional expressions during learning activities more accurately. This will aid in providing

[6] To assess the performance, we utilized CE to measure the loss of the output from both AM and AF.

effective real-time support for learners during learning. In future work, our research will focus on implementing the proposed FER technique for intelligent tutors to interpret learners' (especially neural typical children's) emotions and provide real-time dynamic feedback to promote them reaching educational goals.

References

1. Akhand, M., Roy, S., Siddique, N., Kamal, M.A.S., Shimamura, T.: Facial emotion recognition using transfer learning in the deep cnn. Electronics **10**(9), 1036 (2021)
2. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. arXiv preprint arXiv:2006.09882 (2020)
3. Craig, S., Graesser, A., Sullins, J., Gholson, B.: Affect and learning: an exploratory look into the role of affect in learning with autotutor. Journal of educational media **29**(3), 241–250 (2004)
4. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4690–4699 (2019)
5. D'mello, S., Graesser, A.: Autotutor and affective autotutor: Learning by talking with cognitively and emotionally intelligent computers that talk back. ACM Transactions on Interactive Intelligent Systems (TiiS) **2**(4), 1–39 (2013)
6. D'Mello, S., Graesser, A.: Dynamics of affective states during complex learning. Learning and Instruction **22**(2), 145–157 (2012)
7. D'Mello, S., Kappas, A., Gratch, J.: The affective computing approach to affect measurement. Emotion Review **10**(2), 174–183 (2018)
8. Graesser, A., Chipman, P., King, B., McDaniel, B., D'Mello, S.: Emotions and learning with auto tutor. Frontiers in Artificial Intelligence and Applications **158**, 569 (2007)
9. Graesser, A.C.: Emotions are the experiential glue of learning environments in the 21st century. Learning and Instruction **70**, 101212 (2020)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
11. ImageNet: ImageNet. https://www.image-net.org/ (2021), [Online; accessed 28-December-2022]
12. Li, S., Deng, W.: Deep facial expression recognition: A survey. IEEE transactions on affective computing (2020)
13. Linson, A., Xu, Y., English, A.R., Fisher, R.B.: Identifying student struggle by analyzing facial movement during asynchronous video lecture viewing: Towards an automated tool to support instructors. In: International Conference on Artificial Intelligence in Education. pp. 53–65. Springer (2022)
14. Wang, F., Liu, W., Liu, H., Cheng, J.: Additive margin softmax for face verification. arXiv preprint arXiv:1801.05599 (2018)
15. Winne, P.H.: A cognitive and metacognitive analysis of self-regulated learning. Handbook of self-regulation of learning and performance pp. 15–32 (2011)