Emotion Recognition From Facial Expressions: A Preliminary Report

Tanja Jaschkowitz* Leah Kawka* Mahdi Mohammadi* Jiawen Wang*

{Tanja.Jaschkowitz,Leah.Kawka,Mahdi.Mohammadi,Jiawen.Wang}@campus.lmu.de

1. Introduction

Facial emotion recognition (FER) [3] is not only an interesting in our daily life, but also important in the realm of artificial intelligence and computer vision. In this short proposal, we aim to leverage several deep neural networks to analyze and interpret different human facial emotions.

The structure of this report is arranged as follows. In Section 2, we provide the datasets we used, the model architecture we implemented. The preliminary evaluation results of our models are given inSection 3. In Section 4 describes the optimization strategies we have already used and plan to investigate in the coming weeks. Finally, an overview of our time schedule for the entire final project is given in Figure 1. Our code and supplementary material are available at https://github.com/werywjw/SEP-CVDL.

2. Approach

2.1. Dataset Acquisition and Processing

Firstly, for all the image data from the training dataset RAF-DB[1] [4, 5], we filter out neutral instances from the original dataset, the emotion labels are denoted as 1 (Surprised), 2 (Fearful), 3 (Disgusted), 4 (Happy), 5 (Sad), and 6 (Angry) for simplicity (Our first dataset is downloaded from https://www.kaggle.com/datasets/shuvoalok/raf-db-dataset/code with the addition CSV file to their labels). The test result in Figure 2 is also aggregated from this specific dataset. To transform and resize the images to `(64,64)`, we convert the images to greyscale with three channels as our original convolutional neural network (CNN) is designed to work with three-channel inputs. Also, we randomly flip the images horizontally with a default 50% chance. This kind of augmentation helps in making the model more robust to orientation changes and thus improves the generalization ability. Our training dataset is aggregated from FER+ [1], CK+ [6].

2.2. Model Architecture

We implemented from scratch an emotion classification model with four convolution layers at the very beginning. Following each convolutional layer, batch normalization is applied. This stabilizes learning by normalizing the input to each layer. Then three linear layers are applied to extract features to the final output. We also add a `dropout` layer to prevent overfitting. The activation function used after each layer is Rectified Linear Unit (ReLU), since it introduces the non-linearity into the model, allowing it to learn more complex patterns. In order to find the best hyperparameter configuration (see Table 2 for details) of the model, we utilize the parameter grid from sklearn [2].

[1] http://www.whdeng.cn/raf/model1.html

Models	Accuracy (Train)	Accuracy (Test)	Accuracy (Vali)
CNN (Baseline)	66.3	75.2	52.6
CNN (SE)	74.3	79.9	59.6
CNN (SE+Residual)	71.5	78.9	56.4
ResNet18 [2]	76.8	79.8	60.3

Table 1. Accuracy (%) for different models in our experiments

Hyperparameter	Configuration
Learning rate	{0.1, 0.01, 0.001, 0.0001}
Batch size	{8, 16, 32, 64}
Dropout rate	{0.5}
Epoch	{10, 20, 30}
Early stopping	{`True`, `False`}
Patience	{5}

Table 2. Explored hyperparameter space for our models

3. Preliminary Results

For evaluation, we use the metric accuracy. As seen in Table 1, we report all the training, testing, and validation accuracy in % to improve the performance of our models. The loss function employed for all models is cross-entropy, which is typically for multi-class classification. The best results of the performance of the model with respect to loss and accuracy are depicted in Figure 2.

[2] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.ParameterGrid.html

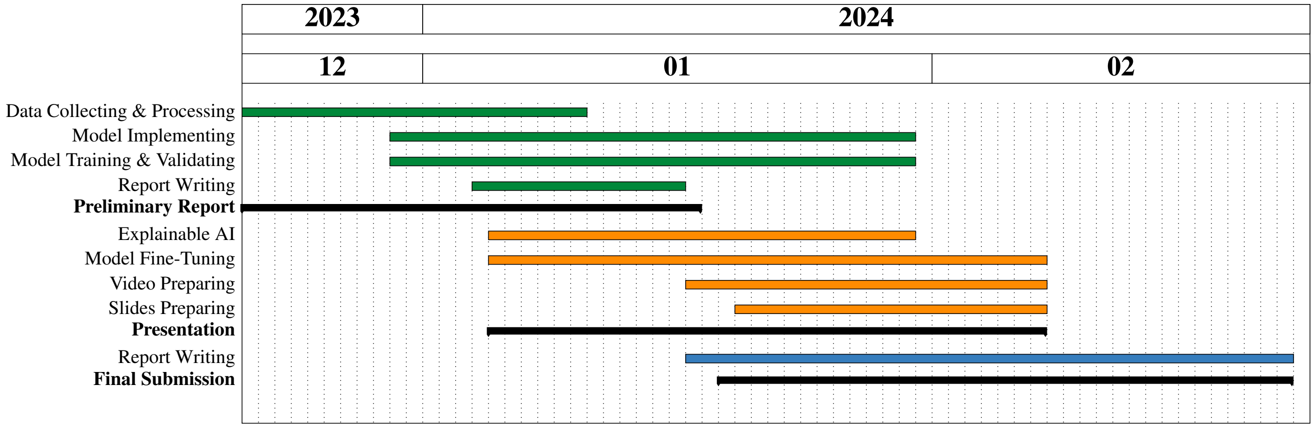

Figure 1. Overview of the schedule for the final project

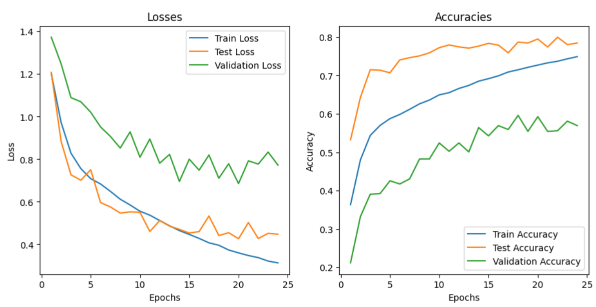

Figure 2. Empirical results in terms of the loss and accuracy on different training epochs

4. Optimization Strategies

We increase the depth of the network by adding some convolutional layers to learn more complex features. We also add the residual connections to help the training of deeper networks more efficiently, as they allow gradients to flow through the network more easily, improving the training for deep architectures. Moreover, we add squeeze and excitation (SE) blocks to apply channel-wise attention. In the coming weeks, we

Author Contributions

Equal contributions are listed by alphabetical order of surnames. Every author did the literature research and contributed to the writing of the paper.

- **Tanja Jaschkowitz** implemented the model architecture, training and testing infrastructure,
- **Leah Kawka** collected the training data, prepared data processing, implemented augmentation, ran the results, Explainable AI & Video-green square
- **Mahdi Mohammadi** implemented the , augmentation,
- **Jiawen Wang** implemented the model architecture, training and testing infrastructure, and optimization strategies. In the specific writing part, she also checked and aggeregated this report from other team members.

Acknowledgements

We are deeply grateful to our advisors **Johannes Fischer** and **Ming Gui** for their helpful and valuable support during the entire semester. We also thank **Prof. Dr. Björn Ommer** for providing this interesting practical course.

References

[1] Emad Barsoum, Cha Zhang, Cristian Canton-Ferrer, and Zhengyou Zhang. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction, ICMI 2016, Tokyo, Japan, November 12-16, 2016*, pages 279–283. ACM, 2016. 1

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016. 1

[3] ByoungChul Ko. A brief review of facial emotion recognition based on visual information. *Sensors*, 18(2):401, 2018. 1

[4] Shan Li and Weihong Deng. Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. *IEEE Transactions on Image Processing*, 28 (1):356–370, 2019. 1

[5] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2584–2593. IEEE, 2017. 1

[6] Patrick Lucey, Jeffrey F. Cohn, Takeo Kanade, Jason M. Saragih, Zara Ambadar, and Iain A. Matthews. The extended cohn-kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR Work-*

shops 2010, San Francisco, CA, USA, 13-18 June, 2010, pages 94–101. IEEE Computer Society, 2010. 1