Deep Learning (COSC 2779/2972) – Assignment 1 – 2023

Wing Hang Chan(s3939713)

1 Problem Definition and Analysis

This project involves a supervised learning task utilizing deep Convolutional Neural Networks (CNNs) for the classification of images based on different Facial Action Units (AUs), also known as FACS codes, and high-level emotions using the Extended Cohn-Kanade Dataset (CK+). The CK+ dataset comprises 560 labelled images, with 3 distinct high-level emotions (Figure 1) and 15 AUs. High-level emotion classification is treated as a multi-class problem, while AUs are handled as a multi-label classification task. Notably, each image can have only one high-level emotion label, but it can carry multiple AU labels.

The class distribution within the high-level emotion classes (negative, positive, and surprise) is imbalanced (Figure 2). Similarly, the distribution of AU labels across images is uneven (Figure 3). To address these issues, an over-sampling technique was applied to balance the weight of high-level emotions. Additionally, image pre-processing was performed, involving the conversion of coloured images to grayscale and the cropping and resizing of images to a consistent size. Given the relatively small size of the CK+ dataset, oversampling was chosen over down-sampling, as it was believed to enhance model generalization. Image augmentation techniques, including random zooming and adjustments to brightness, were employed to generate 500 images for each high-level emotion.

The dataset was divided into training, validation, and testing sets, with proportions of 70%, 15%, and 15%, respectively. Specifically, the training set contained 392 instances, while the validation and testing sets each comprised 84 instances. To balance the dataset, the high-level emotion labels were used, as the relationship between images and AUs is many-to-many. Achieving balance solely based on AUs would be a complex and endless endeavor.

The images in the dataset exhibit variations in format, including differences in colour scales (coloured and grayscale) and dimensions (height between 480 to 490 and width from 640 to 720). As a standardization step, the images were resized to 480x480 pixels while ensuring that the centroid of each face remained at the centre of the image. Given that most images were in grayscale and for better reduction, coloured images were converted to monochrome.

In summary, this project involves the use of deep CNNs for classifying images based on high-level emotions and AUs. The CK+ dataset, consisting of 560 images, was pre-processed through oversampling, image augmentation, and resizing for model training, validation, and testing. The class imbalances in both high-level emotions and AUs were addressed through appropriate techniques, and the images were standardized to a consistent format to facilitate effective model training and evaluation.

2 Evaluation Framework

When assessing the performance of models in image classification, the standard practice often involves utilizing top-5 accuracy. However, in the context of the CK+ dataset, which is notably small and contains only 3 high-level emotion categories, the application of top-5 accuracy is not appropriate. Instead, it is more suitable to consider the top-1 accuracy, which has been observed to fall within the range of 70-85% across different models [1]. Establishing a target of 80% for the top-1 accuracy would be a reasonable objective in this scenario.

For the classification of AU labels, a distinct approach is taken. Here, all 15 FACS codes are consolidated into an array for training the model. This consolidation serves to streamline the model's complexity. The model architecture encompasses a solitary output layer featuring 15 tensors, aligning with the 15 AU codes. Wang [2] shows mean average precision (mAP) with a 20-label classification model would be around 75-92%. Thus, a target accuracy of 70% is proposed with comparing Wang's using a larger dataset although that task should be more complex.

The chosen evaluation metric for the model's performance in AU label classification is the binary_accuracy provided by the Keras framework [3]. This metric varies slightly from standard accuracy. While accuracy necessitates an exact match, binary_accuracy operates under the assumption of binary classification, wherein outcomes are typically 1 or 0, signifying yes or no. This metric accommodates situations where predictions are in the vicinity of the true value, such as mapping 0.6 to 1 or 0.33 to 0, employing a threshold of 0.5. This approach aligns well with the FACS code classification, wherein each AU code corresponds to a distinct binary value. Consequently, binary_accuracy emerges as a fitting choice for evaluating the model's proficiency in AU code classification.

To conclude, the evaluation strategy for this project incorporates distinct metrics based on the nature of the tasks. For high-level emotion classification, a top-1 accuracy target of 80% is set, while for AU label classification, a binary_accuracy target of 70% is considered reasonable given the limited dataset size and complexity of the problem.

3 Approach & Justifications

Starting with a simpler network is considered good practice. While various neural networks like VGG, ResNet, and GoogLeNet have earned strong reputations for computer vision tasks, they demand substantial resources. For instance, VGG16, ResNet50, and GoogLeNet (Inception) have parameter counts of 138.4M, 25.6M, and 23.9M respectively. Furthermore, they require 4.2ms, 4.6ms, and 6.9ms per inference step using a Tesla A100 GPU [1]. However, when contrasted with ImageNet, the CK+ dataset is relatively small. In such cases, opting for a lighter model might yield better results in terms of both speed and accuracy. For this purpose, MobileNet presents an appropriate solution.

MobileNet [6] makes use of depthwise separable convolutions which is first presented by Xception network to build an efficient and light model. Depthwise separable convolutions is a combination of depthwise and pointwise convolutions. Depthwise convolution is more efficient than standard Conv but it does not come up with new features. Therefore, pointwise (1 x 1) convolution is added for creating the features. Choosing MobileNet over MobileNetV2 is due to GPU resource limitations. Despite MobileNetV2's deeper and more accurate architecture, it's chosen due to constrained GPU resources. The model encompasses two 1024-unit dense layers atop MobileNet, incorporating a global average pooling layer for tensor flattening. Two output layers employing softmax activation are connected to a 1024-unit dense layer. One output layer, with 15 units, signifies individual AU codes. The second output layer, with 3 units, represents high-level emotions. While inputs are normalized by scaling the grayscale range, the ADAM optimizer is favored due to its adaptive learning rate, enhancing prediction precision and optimization likelihood.

The base model employs two regularization techniques: data augmentation and early stopping. Data augmentation is imperative since oversampling can yield duplicate inputs. Applying random zoom and brightness ensures uniqueness. Random zoom rectifies excess margin and unhelpful information in 480 x 480 images, even after centroid cropping. Moreover, random brightness counteracts over-lit images, preserving essential features like lips, eyes, and nose.

The second regularization approach is early stopping. For single-output models, patience can be easily determined by monitoring a selected metric. However, this model has multiple outputs, necessitating monitoring of multiple metrics. To ensure both outputs excel, two metrics are monitored, guaranteeing optimal performance for both outputs. The early stopping mechanism halts training when the model reaches the pre-set target—70% accuracy for AU labels and 80% for high-level emotions. If the target isn't met, the model persists beyond the patience threshold. This approach accounts for CK+ dataset size limitations and ensures the model converges towards the objectives.

Though this early stopping approach has limitations—for instance, if one output excels while the other lags—it ensures the chosen model aligns with target accuracy requirements, making it an appropriate strategy.

4 Experiments & Tuning

Several hyperparameters can be tuned in the model, though the code presented only focuses on the primary aspects. The first parameter is the sample size for each class. Since data is oversampled, the class sizes in the training set can be adjusted. In the base model, this size is set as the maximum count of the mode of the high-level emotion, which is typically negative. This number might range between 250 and 330, depending on the split in the training data. Thus, the training data size could fall between 750 and 990. Despite the training data performing well across epochs, a larger number of epochs is required to achieve the target result. Usually, it takes 10 to 15 epochs to reach the target. However, a larger dataset can reach the target sooner, generally taking 5 to 8 epochs. This larger dataset also demands more time per epoch. However, the dataset size shouldn't be overly large due to potential similarities among images within the same training epoch. These images are augmented, not new. The batch size is set at 16 due to the dataset's small size, while the training error rate often recommends a batch size of 32 for optimal error rate results [4].

Another parameter influencing training performance is the input dimensions of the images. Although pre-processed as 480 x 480, these images may contain extraneous information like timestamps at the bottom or hair. Downsizing to 300 x 300 and focusing on the faces is feasible, given that faces are commonly centred in the images. Cropping from the centre streamlines the process, potentially enhancing precision and speed.
The larger sample model introduces two additional augmentation methods, potentially leading to improved generalization. Given the need for more augmentation to maintain freshness in the training set, random contrast and rotation are chosen. Both factors are deliberately kept low, considering the standardized nature of the images. Heightened factors could affect validation and testing outcomes.

The optimizer learning rate is set at the default rate, as altering it to 0.01 or 0.0001 significantly impacts model accuracy. This may be due to the small dataset size, and multi-label classification might lack sufficient data for training, even within the oversampled model.

A hint of overfitting is observable in the base model. While binary accuracy and categorical accuracy for training continue to rise, validation remains stagnant. This is a common occurrence with minuscule datasets and complex models. Consequently, more intricate models such as ResNet or GoogLeNet were not employed, as they tend to exacerbate overfitting. Additionally, L1 regularization, along with increased data augmentation and early stopping—both previously mentioned—is employed to address the overfitting concern. L1 regularization is preferred over dropout due to the latter's random feature exclusion, potentially compounded by multiple dropout layers. In contrast, L1 regularization often rests atop the fully connected layer, allowing the model to decide which features fade out.

To wrap up, the careful selection of hyperparameters, including sample size, batch size, input dimensions, and regularization strategies, is crucial for effectively training models, especially with small datasets. Employing augmentation techniques and striking a balance between complexity and regularization helps mitigate challenges like overfitting. These considerations collectively contribute to improved model performance, generalization, and convergence.

5 Ultimate Judgment, Analysis & Limitations

In examining the challenges faced by the base model, several factors could contribute to the consistent validation accuracy during the initial tenth epoch. One potential factor relates to the binary and categorical accuracy metrics, as previously discussed. These metrics are governed by a threshold, and the validation accuracy might appear stagnant because the model fails to generate results beyond the threshold. Alternatively, the size of the training set might be insufficient to produce the diverse features the model requires. Moreover, the small validation set might not adequately represent the breadth of features captured by the training model. The expectation was that a more substantial number of augmented images would lead to improved validation results after the initial epochs. To address this, an enhanced model employs a larger training sample size.

The accuracy graph for the base model (Figure 4) indicates a conspicuous overfitting issue concerning high-level emotions. While training accuracy continues to rise, validation accuracy plateaus. However, caution must be exercised before categorizing this as overfitting. The discrepancy could be attributed to the limited size of the validation set. Notably, the test results of the base model reveal a precision, recall, and F1-score of 0 for AU1 (inner brow raiser) with a sample size of 14 in the test set (Figure 10). Although other AUs exhibit 0 accuracy, their data sizes are smaller compared to AU1. Concerning high-level emotions, the model achieves a weighted F1-score of 0.8099 (Figure 5), driven by 69 true positives out of a total sample of 84. Nevertheless, the model retains a tendency to predict negative labels, reflecting the prevalence of the majority class in the balanced training set.

The implementation of L1 regularization in the large sample model yields significant improvements in combating overfitting in high-level emotions. The gap between training and validation accuracy narrows appreciably (Figure 6). Despite the training accuracy converging toward 1, suggesting possible exhaustion of new features in the training data, the validation accuracy makes commendable progress. Regrettably, AU1 accuracy remains unaltered in the test set, maintaining a value of 0 (Figure 11). In contrast, AU17 accuracy exhibits substantial enhancement, contributing to a higher weighted accuracy. Moreover, the larger sample model registers more true positive counts (Figure 7) while minimizing predicted negative labels, culminating in a weighted F1-score of 0.8915—a noteworthy 8% improvement. This underscores the significance of dataset size in image classification tasks.

Transfer learning emerges as a viable approach to address the challenge of limited dataset size. Notably, Tapotosh [5] demonstrates that augmenting CK+ with other datasets yields improved performance. Subsequently, a transfer learning model utilizing the MobileNet weights from ImageNet [7] is introduced for comparison with the base and large sample models. The transfer learning model uses the training set of the base model, leveraging the wealth of data from over a million images. The model's parameters are locked to avert catastrophic forgetting and expedite training, reducing trainable parameters from 5.3M to 2.1M. Additionally, the model benefits from features derived from ImageNet. Validation outcomes are promising, with the model achieving the target within three epochs—compared to the base model's 11 epochs. However, a challenge arises due to the RGB nature of ImageNet, as opposed to the grayscale images of CK+. This necessitates adapting the input to accommodate colored images. Despite similarities in the AU label test results (Figure 12), the persistent issue with AU1 persists. True positive counts fall between those of the large sample and base models, suggesting improved performance if data sizes were equal.

It is imperative to note that certain AU labels are absent from the test set, preventing their validation. Limited data availability for certain AUs presents a challenge in this classification task. The model's inability to predict some AUs is attributed to its tendency to treat all AUs as an average, potentially benefitting from a weight-based approach.

6 Discussion on Ethical issues and biases

Ethical issues

1. Informed Consent - Adequate consent should be obtained from participants and data users, including students.
2. Privacy Concerns - Emotion classification often involves analysing facial expressions, raising privacy concerns due to potential identity disclosure.

Biases

1. Gender - Emotion expression can differ based on gender. If the dataset is skewed in gender distribution, the model might not perform well for certain groups.
2. Skin Color - Given the grayscale nature of the dataset, biases might arise concerning individuals with darker skin tones.
3. Bias in Data Collection - Data collected from specific environments might not effectively generalize to real-world scenarios.

7 References:

[1] Keras - “Keras API reference - Keras Applications”, [online] Available: https://keras.io/api/applications/
[2] Zhouxia Wang - “Multi-label Image Recognition by Recurrently Discovering Attentional Regions” in Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 464-472
[3] Keras - “Keras API reference / Metrics / Accuracy metrics”, [online] Available: https://keras.io/api/metrics/accuracy_metrics/
[4] Ayush Thakur (Jul 9, 2023) – “What's the Optimal Batch Size to Train a Neural Network?” , [online] Available: https://wandb.ai/ayush-thakur/dl-question-bank/reports/What-s-the-Optimal-Batch-Size-to-Train-a-Neural-Network---VmlldzoyMDkyNDU
[5] T. Ghosh, "Impact of Facial Expressions on the Accuracy of a CNN Performing Periocular Recognition," in 2019 8th Brazilian Conference on Intelligent Systems (BRACIS), 2019.
[6] A. G. Howard, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," arXiv preprint arXiv:1704.04861, 2017.
[7] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, 2012: 1097–1105

8 Appendix:

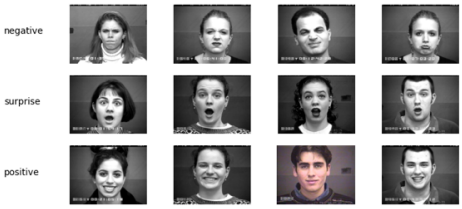

Figure 1 Image Sample of High Level Emotions

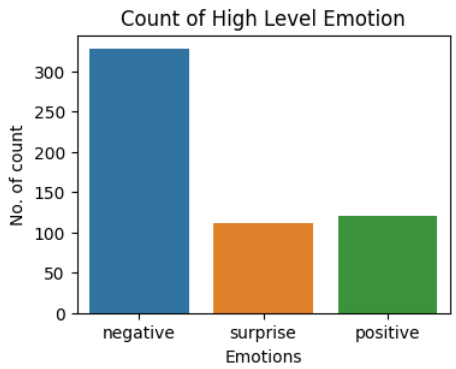

Figure 2 Bar Chart Count of High Level Emotion

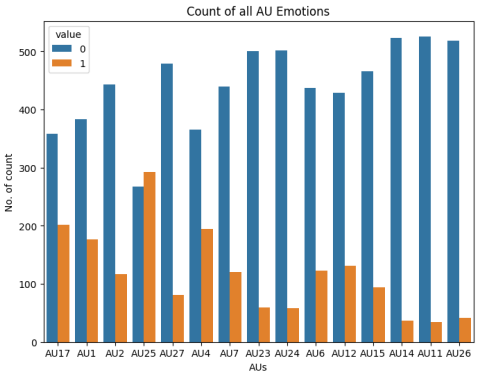

Figure 3 Bar Chart Count of AU Labels

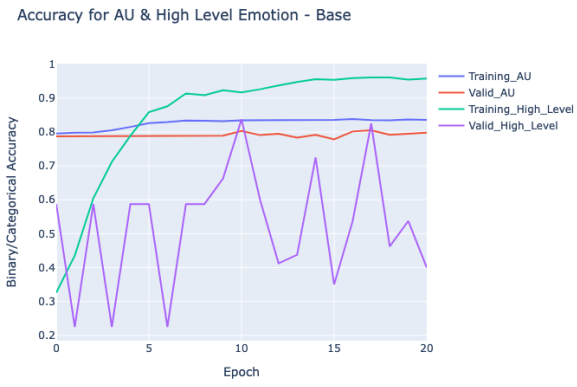

Figure 4 Training & Validation Accuracy - Base Model

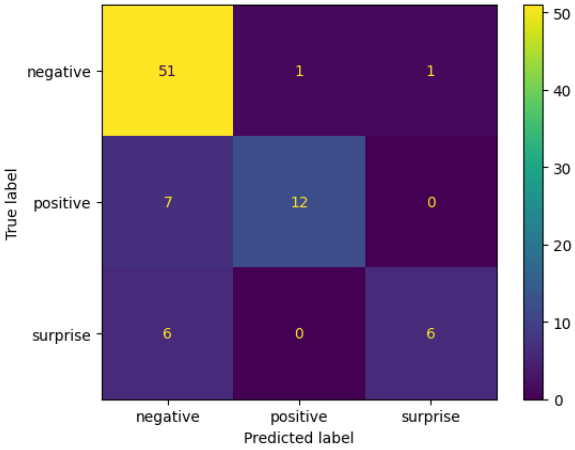

Figure 5 Confusion Matrix - Base Model

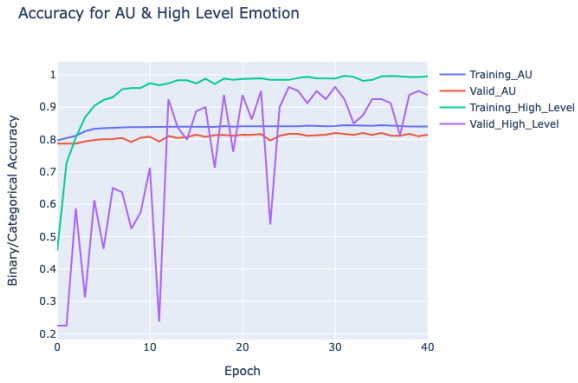

Figure 6 Training & Validation Accuracy - Large Sample Model

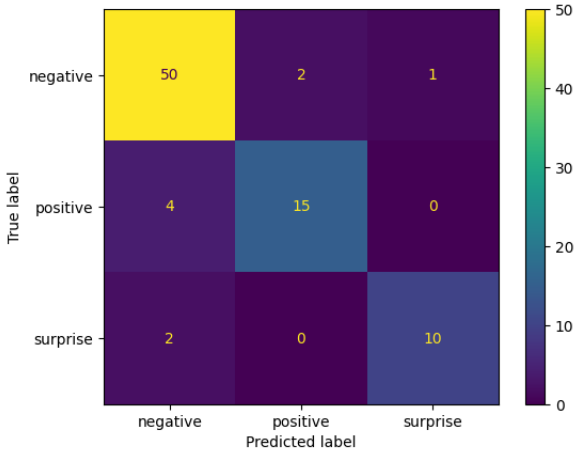

Figure 7 Confusion Matrix - Large Sample Model

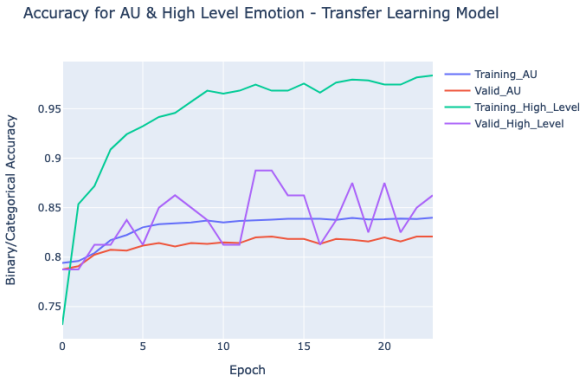

Figure 8 Training & Validation Accuracy - Transfer Learning Model

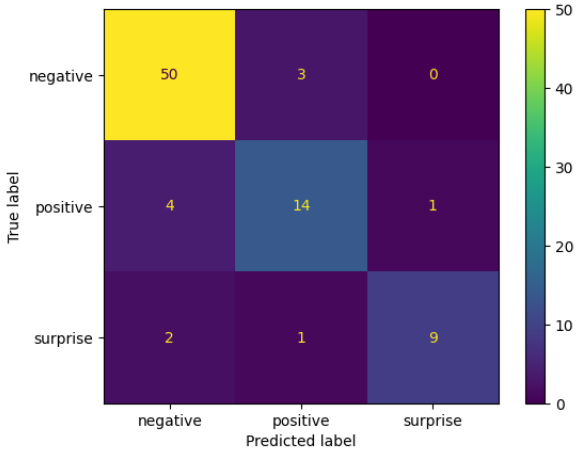

Figure 9 Confusion Matrix - Transfer Learning Model

	precision	recall	f1-score	support
0	0.56	1.00	0.72	38
1	0.00	0.00	0.00	14
3	0.20	0.04	0.06	27
5	0.00	0.00	0.00	2
6	0.00	0.00	0.00	2
8	0.00	0.00	0.00	1
10	0.00	0.00	0.00	0
accuracy			0.46	84
macro avg	0.11	0.15	0.11	84
weighted avg	0.32	0.46	0.34	84

Figure 10 Classification Report - Base Model

	precision	recall	f1-score	support
0	0.65	0.92	0.76	38
1	0.00	0.00	0.00	14
3	0.36	0.19	0.24	27
5	0.00	0.00	0.00	2
6	0.00	0.00	0.00	2
8	0.00	0.00	0.00	1
10	0.00	0.00	0.00	0
accuracy			0.48	84
macro avg	0.14	0.16	0.14	84
weighted avg	0.41	0.48	0.42	84

Figure 11 Classification Report - Large Sample Model

	precision	recall	f1-score	support
0	0.65	0.97	0.78	38
1	0.00	0.00	0.00	14
3	0.50	0.26	0.34	27
5	0.00	0.00	0.00	2
6	0.00	0.00	0.00	2
8	0.00	0.00	0.00	1
10	0.00	0.00	0.00	0
accuracy			0.52	84
macro avg	0.16	0.18	0.16	84
weighted avg	0.45	0.52	0.46	84

Figure 12 Classification Report - Transfer Learning Model