CNN with Transfer Learning for Facial Expression Analysis

Manuwai Korber[1]

[1] Research School of Computer Science, Australian National University
u6665970@anu.edu.au

Note to Marker
Please take into consideration I have switched datasets from between assignment 1 and 2. This was with the permission of the course convenor (Ms Plested). For assignment 2 I have used the V2 faces-emotion data set. My previous paper used the V1 thermal-stress data set [1] however the V2 thermal-stress provided was misformatted and the fixed version was too large to work with (~31Gb). Hence, I will be unable to compare results with my assignment 1 and will instead compare to the faces-emotion data set paper. Only a summary of my first paper dataset will be included as it has no relevance to the new dataset or the deep learning techniques implemented.

Abstract. Accurate detection of facial expressions in real world scenarios is a useful tool. It is utilised in human computer interaction (HCI), pain monitoring in patients, and medical conditions such as autism. This paper applies the technique of transfer learning from the pre-trained Convolutional Neural Network (CNN), FaceNet. FaceNet was trained on the VGGFacec2 database with more than 3 million images. The model was retrained to classify a subject's facial expressions as one of seven emotions. Training and testing were conducted with images from the Static Facial Expressions in the Wild database (SFEW). The CNN produced an accuracy of 37%, compared to the complex techniques used in the SFEW paper which predicted with accuracies of 43.71% and 46.28%.

Keywords: Convolutional Neural Network, Transfer Learning, Facial Expression Analysis

0.0 Previous Paper Summary

0.1 Previous Paper Background, Dataset, Technique

Stress is a prevalent problem in society. The analysis of physiological signs can be used to predict whether a person is stressed or not. Being able to recognise stress could produce health and safety benefits in a variety of scenarios such as vehicle safety. Traditional stress recognition systems use self-reporting or invasive sensors to measure physiological signals. More recent techniques have been to use contact-less sensors such as RGB and thermal cameras. Consequently, my first paper looked at predicting whether a subject is stressed from an RGB and thermal camera recording.

The dataset used originates from the paper *"Thermal Super-Pixels for Bimodal Stress Recognition"* [2] and was collected at the Australian National University (ANU) [1]. The principal component analysis data is labelled as "stressed/unstressed" when the label of the film is "stressed/unstressed".

The technique applied comes from the paper "Indicators of Hidden Neuron Functionality: the Weight Matrix versus Neuron Behaviour" by T.D. Gedeon [3]. A common problem with neural networks is that during training models get stuck in local minima. The technique solves this by visualising the angle between activation vectors. Giving an indication whether a network is permanently stuck or whether it is worthwhile continuing to train the network.

0.2 Investigations

A standard neural network was trained and tested on the dataset to give a baseline that allows for the comparison of a neural network enhanced with the angular separation technique. The neural network structure was kept constant for both the standard and technique enhanced network. With 10 input nuerons, 1 hidden layer containing 8 hidden neurons and an output layer of size 2. The benefits of the technique are analysed qualitatively through visual comparisons between different models and datasets. Looking at the wholistic benefits it brings to the process of model development and selection.

0.3 Method

The technique calculates the angles between the activation vectors of the hidden layer neurons. Plotting the angles over the entire training process of a neural network every few epochs. Once training is complete the calculated angles are plotted to easily visualise how the neural network progressed during training. The angles are calculated as below:

0.4 Results

Table 1. Comparison of different models testing accuracy

	Standard Neural Network	Neural Network with Angular Separation Technique	Thermal Stress Paper Neural Network
Test Accuracy	**47%**	**47%**	**89%**

However, the angular separation technique is not designed to increase the direct performance of a neural network either by reducing its loss or increasing testing accuracy. Rather it is to inform a model's engineer of whether the model is permanently stuck within a local minimum. The technique provides the benefit of being able to visualise the training progress of a neural network. Without the technique it would be difficult to confirm whether the network was stuck in a local minimum permanently.

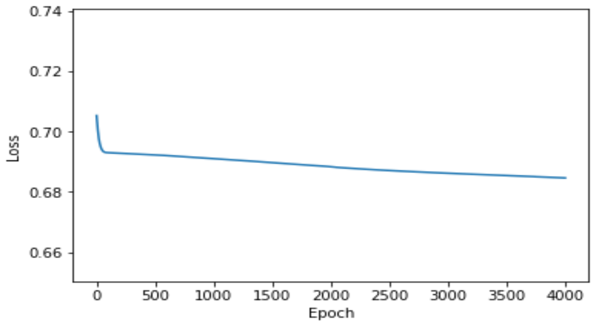

Figure 1. Training loss.

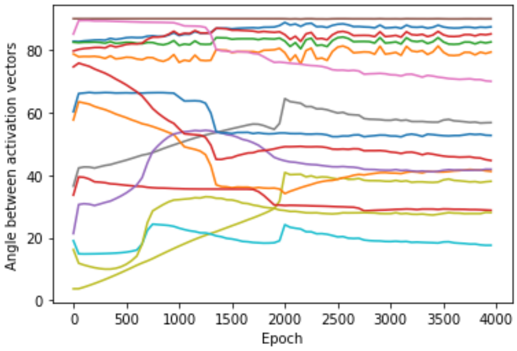

Figure 2. Angle between activation vectors.

The plots from Figure 2 and 3 result from the neural network enhanced with the angular separation technique. In Figure 2 the loss is still decreasing but it difficult to tell whether the network is stuck in a local minimum. However, from Figure 3 after epoch 2000 it is clear there is little change to the angles between activation vectors. We can then determine that this model is indeed stuck in a local minimum and it is not worth training further. Hence, the technique has provided benefit by saving time and effort during model selection.

0.5 Conclusion & Discussion

The angular separation technique works well for any dataset including the one chosen. This property results from the techniques dependence on the network architecture and not the dataset. For the thermal-stress dataset the technique works as it gives the insight that the model is permanently stuck in a local minimum.

However, there are a few problems with the technique which will require future work. For instance, with neural networks that have large hidden layer sizes it becomes difficult to visualise the angles between the activation vectors. Increases in the number of hidden layers and their sizes increases the computational cost of calculating all the angles, slowing down the training of the network itself. A live plotting of the network's training progression could be provided instead of having to wait until the network has completed training.

1 Introduction

1.1 Background

Accurate detection of facial expressions in real world scenarios is a useful tool. It is utilised in human computer interaction (HCI), affective computing, human behaviour analysis, pain monitoring in patients, stress, anxiety and depression analysis, lie detection and medical conditions such as autism. Facial expression analysis has previously used images of facial expressions taken in controlled 'lab like' conditions. Hence, the Static Facial Expressions in the Wild database [4] was created to produce a data set that resembles facial expressions in real world environments. Consequently, this report looks at predicting the emotion of a subject from an image in a real-world environment.

1.2 Dataset

The Static Facial Expressions in the Wild (SFEW) database originates from the paper *"Static Facial Expression Analysis in Tough Conditions: Data, Evaluation Protocol and Benchmark"* [4]. SFEW is a collection of 675 images from various movies labelled with six basic expressions angry, disgust, fear, happy, sad, surprise and neutral. It features images with unconstrained facial expressions, varied head poses, large age range, different face resolutions, occlusions, varied focus and close to real world illumination.

1.3 Angular Separation Technique

The technique utilised on the original thermal-stress dataset comes from the paper *"Indicators of Hidden Neuron Functionality: the Weight Matrix versus Neuron Behaviou*r" by T.D. Gedeon [2]. A common problem with neural networks is that during training models get stuck in local minima. A common training indicator used is the loss value plotted over epochs. However, from the loss function alone it is difficult to determine whether the model is a) stuck in a local minimum and will eventually leave or b) will never leave the local minima.

The technique aims to solve this problem by giving a view into a neural network's learning progress via the angle between activation vectors. Which shall be referred to as the 'angular separation technique' henceforth. The angular separation technique does not strive to improve the outright performance of a network but to give insight into the training progression of a network. Achieved by giving an indication whether a network is permanently stuck or whether it is worthwhile continuing to train the network. The technique will be applied to the neural network built for the faces emotion dataset to determine whether we can gain additional insight into their training progression.

1.4 Deep Learning Techniques

Two main deep learning techniques were combined in producing the final model, that each address a different issue. These are the Convolutional Neural Network (CNN) and transfer learning.

A CNN is used to transform the images into a more workable format from a large size of 576x720 pixels with three colour channels. Otherwise for the input alone there would be 1,244,160 input neurons and as a result many more parameters to adjust within the neural network. Making it extremely time consuming and resource intensive to work with image inputs. A CNN solves this problem by performing convolutions on the inputs. Basically, condensing the information within the input to a lower dimension and therefore a more manageable size.

Transfer learning helps avoid the limitations of a small dataset. Deep learning often requires large datasets to work properly however the SFEW dataset is relatively small. It only contains a total of 675 samples with 100 samples for each class bar one, the disgust category which contains 75 samples. In comparison, the dataset ImageNet [5] contains more than 1 million images and 1000 classes. Transfer learning of models trained on large datasets such as ImageNet allows for the transfer of knowledge from a similar problem, saving time and resources. The lower layer parameters are frozen and only the deeper layer parameters are fine-tuned with stochastic gradient descent. The lower layers extract features such as lines and edges which are common to all images. Whereas the deeper layers extract shapes such as eyes and mouths.

2 Method

2.1 CNN Implementation

Initially, tried to implement a CNN from scratch and train on the dataset. The structure was two convolutional layers with ReLU activations functions and pooling layers in between, then two fully connect layers. However, there was little success with such a shallow network in combination with having little data to train with.

2.2 Transfer Learning

Transfer learning offers the ability to transfer knowledge from models trained on large datasets to other datasets in a similar domain. A few pre-trained models were trialled including SqueezeNet [6], ResNet18 [7] and FaceNet [8]. The research tested transfer learning with Resnet18 [RESNET] which was trained on ImageNet with a Top-1 error of 30.24 and Top-5 error of 10.92. However, ImageNet contains only a small number of samples containing human faces and hence FaceNet was selected.

2.2.1 FaceNet

FaceNet was chosen as a pre-trained model to fine tune. It is a model trained on VGGFace2 [9], a large-scale face recognition dataset containing nine thousand identities with between 80 and 800 images for each identity, and more than 3 million images in total. Images are downloaded from Google Image Search and have large variations in pose, age, illumination, ethnicity and profession.

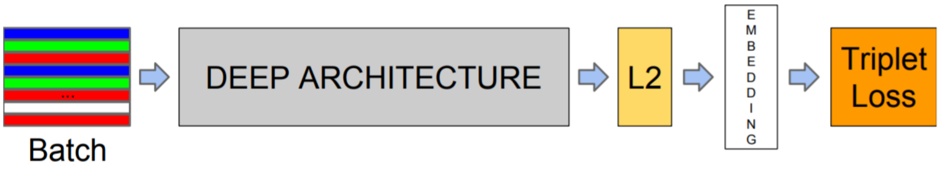

Figure 3. FaceNet Structure [8]

The FaceNet structure, involves a batch size of 32 images. The “Deep Architecture” referenced in Figure 1 is the Inception-ResNetV1 (Figure 2). It is a hybrid model of the Inception and Residual Networks.

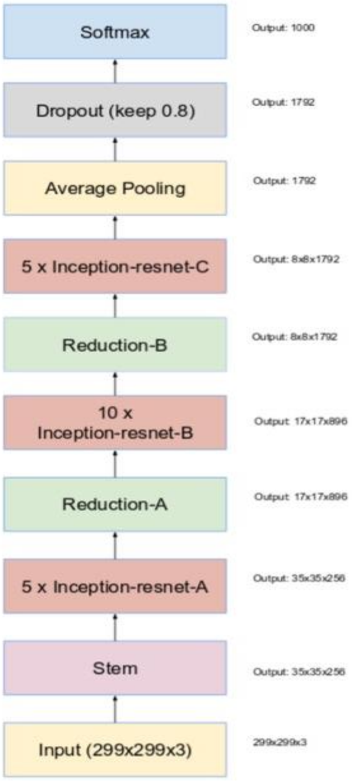

Figure 4. Inception-ResNetV1 Model Structure [10]

Inception Networks utilise “Inception Modules” to perform multiple convolutions with different kernel sizes as well as implement pooling within one layer. The modules produce great results as information with varying sizes can be identified at the same layer. In addition, it removes the need for having to fine tune the kernel size hyperparameter. Residual Networks (ResNet) solve the problem of vanishing gradients that occur in deep neural networks. Achieved through introducing skip connections that pass outputs of a layer to a layer deeper within the network.

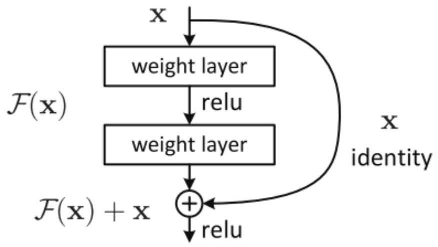
x
weight layer
$\mathcal{F}(\mathbf{x})$
relu
weight layer
x
identity
$\mathcal{F}(\mathbf{x}) + \mathbf{x}$
relu

Figure 5. Residual Learning [11]

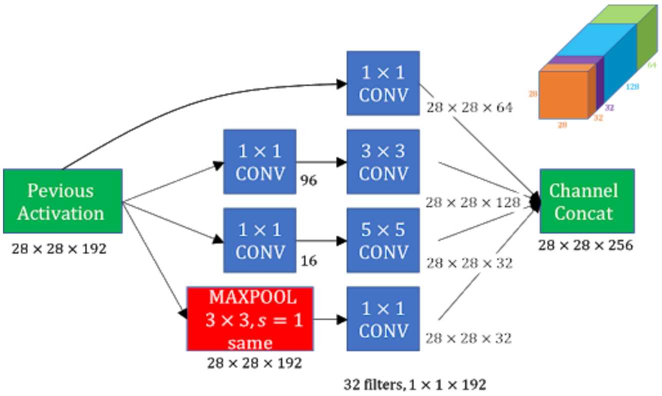

Figure 6. Inception Module [12]

2.4 Investigations/ Experiment details

To investigate how well the deep learning techniques implemented perform we utilise the same testing framework as the SFEW dataset paper. The dataset was split into 50% train and 50% test sets. To test the performance of the transfer learning, the number of layers unfrozen and retrained was experimented with.

3 Results

The transfer learning CNN does not perform as well as the original dataset paper methods, Local Phase Quantisation (LPQ) and pyramid of histogram of oriented gradients (PHOG). However, these are complex methods which do not appear to be as easily transferred between problems compared to transfer learning.

Table 2. Comparison with Dataset Paper

Transfer Learning Model	Dataset Paper LQ	Dataset Paper PHOG
37%	**43.71%**	**46.28%**

Table 3. Transfer Learning with FaceNet Results

Layers Unfrozen	1	2	4	6	9	9
Test Accuracy	**23%**	**24%**	**29%**	**33%**	**37%**	**40% (80/20 split)**
Number of Trainable Parameters	**3591**	**4615**	**922,119**	**2,522,887**	**20,773,383**	**20,773,383**

As more layers are unfrozen and retrained the accuracy of the model increases. However, the increased accuracy comes at the cost of having to train exponentially more parameters. This requires more time and higher computation and energy consumption.

Experiments were constrained to the training of a maximum of 9 layers. This is because the GPU used cannot handle retraining more than nine layers as the number of trainable parameters becomes too large. The entire FaceNet model contains 23,486,215 parameters. Therefore ~88% of the network parameters were retrained when 9 layers are unfrozen.

The need to transfer such a large number of layers could stem from the lack of pre-processing applied to input images. For FaceNet, images are usually cropped to include only the face removing lots of noise that is contained within the background of an image.

4 Discussion

Inspecting the confusion matrix, the model misclassifies labels which even a human may misclassify. For instance, many of the images classified as happy have true labels of surprised.

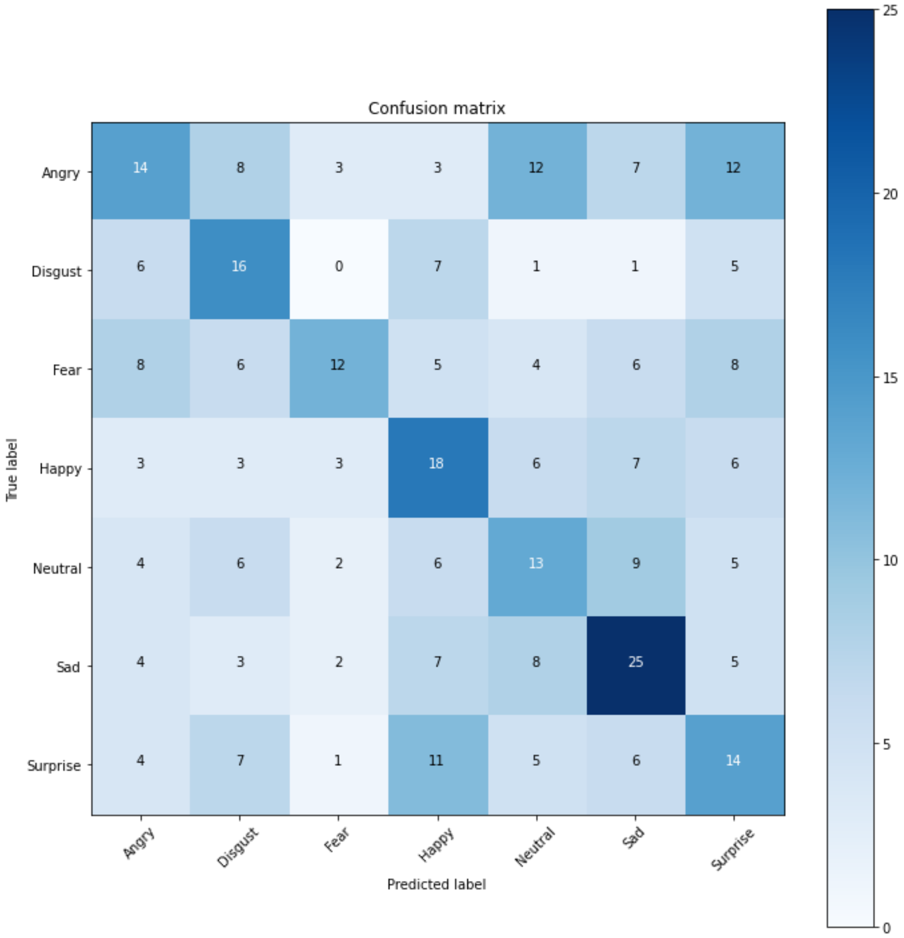

Figure 7. Confusion Matrix of Predictions and True Labels

Here is a test for the reader. What emotion do you perceive in Figure 8? Happy or Surprised? You may be surprised to know that the image is actually labelled as happy. What emotion do you perceive in Figure 9? Surprised or Angry? In fact, the image is labelled as Angry. This shows that the dataset itself does not always contain distinct samples. Although this is not to the fault of the dataset but is due to human emotions being a spectrum rather than purely discrete and up to individual interpretation.

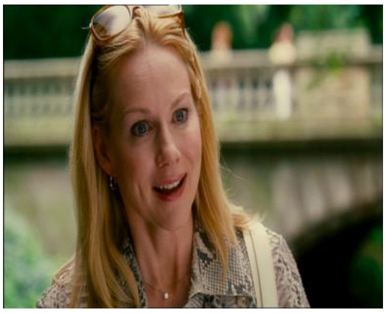

Figure 8. Sample from data set

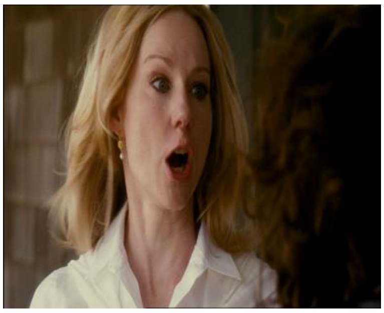

Figure 9. Sample from data set

4.1 Angular Separation Technique

The angular separation technique was attempted on the CNN with transfer learning from FaceNet. The technique however failed to work for two main reasons. Firstly, the CNN contains too many layers to use the technique. However, even if one hidden layer was selected for the technique to be applied to it would still fail. This is due to the application being image processing, the CNN contains many hidden neurons in each layer. Compared to my first paper that implemented a simple neural network containing one hidden layer with only 8 neurons. This CNN is order of magnitudes larger. Secondly, producing the activation vectors each forward and backward pass through the network is more time and resource intensive than passing through PCA components

5 Conclusion and Future Work

The deep learning technique of transferring a pre-trained neural network from one task to another is an extremely powerful tool. In this specific instance the transferred CNN is almost able to match sophisticated techniques such as Local Phase Quantisation (LPQ) and pyramid of histogram of oriented gradients (PHOG). With future work and adjustment, it is expected that the CNN will be able to predict with an accuracy above those techniques.

Pre-processing of the images before being fed into the CNN would result in improved performance. If the images had face alignment and centring the background noise in the images would be removed. Letting the model focus on analysing the facial expressions, speeding up training and improving accuracy.

Data augmentation could be another technique used to improve model performance. Since the SFEW dataset is relatively small, data augmentation can help by producing more training samples. This is done by creating new images each epoch through augmentations such as flips, tilts and resizing. Making the model more robust to variance in poses, lighting, and location.

References

[1] N. Sharma, A. Dhall, T. Gedeon, and R. Goecke, “Thermal spatiotemporal data for stress recognition,” EURASIP Journal on Image and Video Processing, vol. 2014, no. 1, 2014. [Online]. Available: http://dx.doi.org/10.1186/1687-5281-2014-28

[2] R. Irani, K. Nasrollahi, A. Dhall, T. B. Moeslund and T. Gedeon, "Thermal super-pixels for bimodal stress recognition," 2016 Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA), Oulu, 2016, pp. 1-6.

[3] T. D. Gedeon, "Indicators of hidden neuron functionality: the weight matrix versus neuron behaviour," *Proceedings 1995 Second New Zealand International Two-Stream Conference on Artificial Neural Networks and Expert Systems*, Dunedin, New Zealand, 1995, pp. 26-29, doi: 10.1109/ANNES.1995.499431.

[4] Dhall, A., Goecke, R., Lucey, S., & Gedeon, T. (2011, November). Static facial expressions in tough conditions: Data, evaluation protocol and benchmark. In 1st IEEE International Workshop on Benchmarking Facial Image Analysis Technologies BeFIT, ICCV2011.

[5] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition (pp. 248–255).

[6] SqueezeNet. 2016 arXiv. Available at: https://arxiv.org/abs/1602.07360

[7] He, Kaiming, 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, pp. 770-778. 2016.

[8] Schroff, F., Kalenichenko, D., & Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 815–823).

[9] Cao, Qiong & Shen, Li & Xie, Weidi & Parkhi, Omkar & Zisserman, Andrew. (2018). VGGFace2: A Dataset for Recognising Faces across Pose and Age. 67-74. 10.1109/FG.2018.00020.

[10] Image. Bharath Raj. 2018. Available at: https://towardsdatascience.com/a-simple-guide-to-the-versions-of-the-inception-network-7fc52b863202
[11] Image. Vincent Fung. 2017. Available at: https://towardsdatascience.com/an-overview-of-resnet-and-its-variants-5281e2f56035

[12] Image. Datahacker.rs. 2018. Available at: http://datahacker.rs/building-inception-network/