EMOTIONET: A Multi-Convolutional Neural Network Hierarchical Approach to Facial and Emotional Classification using TPUs

Rebecca Y. D'Agostino
Pace University
New York, United States
rydagostino@gmail.com

Schmidt, Thomas
Computer Science
Salem State University
Salem, MA
tschmidt@salemstate.edu

Abstract - Machine Learning and the tasks of recognition and detection have become a salient topic in recent years. [1] This study is focused on improving the task of using deep convolutional neural networks for facial emotion recognition in the wild with a two-part hierarchical architecture made up of two deep neural networks that are focused on two different tasks. The first task uses a convolutional neural network to classify face or not face. The second task ascertains using another neural network if a face is displaying the emotion within a selected suite of emotions, depicting joyful and sad faces. Factoring the recognition task in this way has produced an observed improvement on previous published work with an overall accuracy of 90% compared to platforms on a CPU cluster and also using TPUs to increase computational power.

Keywords: Neural Networks; Deep Learning; Facial Recognition; TPU

I. Introduction

Machine learning, a subset of artificial intelligence, is the field that tries to solve problems by training a program through the process of learning, and then using evaluation and testing to assess how the model performed. [2] The study of artificial neural networks, which falls under supervised learning, is a growing field within machine learning that has experienced increased popularity over the years; neural networks are used for many tasks, such as classifying data within areas of images, voice, self-driving applications, and texts. [3]

Facial recognition is defined as the task of being able to detect where a face is present. Facial classification is the task of organizing the faces into pre-defined classes that are created by the user. The authors present a hierarchical approach to improve classification of facial and emotional classification tasks for photos found in the wild and using default databases by using the emotions of happy or sad as example of further selective classification. Two types of training platforms were also used to compare accuracy using different hardware for the same task.

II. Existing Theories & Previous Work

For the task of facial recognition and classification alone, great progress has been made since the 1990s, where the Eigenface approach was considered the standard method to detect and recognize faces. More recently there has been a shift from Eigenface and other methods (like linear subspace, manifold and sparse representation), to different methods that include learning-based local descriptors, which uses local filters that are trained for increased accuracy. Shallow methods, like Eigenface, were seen as the state-of-the-art methods, until the release of AlexNet, which won the ImageNet competition by using deep learning, that is, neural networks with many intermediate layers. Since AlexNet, deep learning has achieved wide acceptance for facial recognition over other methods. [4]

Facial expression classification has also had a similar history to the task of facial recognition. In the late 2000s, Support Vector Machines (SVM) , Adaboost, Non-negative matrix factorization, Sparse learning were used to try to solve and improve results for this problem. [5] However, Convolutional Neural Networks (CNNs), along with deep learning, have become the industry standard [5] ever since due to their ability to achieve better accuracy than other methods previously seen. One study by T. Pham et al. uses transfer learning with a CNN and a SVM as a multi-step to classify in-the-wild emotions similar to the database used in this paper, which yielded a 72.90% accuracy rate.[6] Another approach by Jha et al. uses a CNN with a multi-class SVM loss function, instead of using a cross-entropy that is seen to yield an accuracy of 70.8%. [7] Kim et al. used an ensemble of CNNs that includes the use of discriminative deep convolutional neural networks and alignment-mapping networks to achieve an accuracy of 73.73% on a similar problem of classifying faces and emotions in-the-wild [8]. Sharma et al. applied a combination of a CNN and a recurrent neural network (RNN) to achieve an accuracy on similar emotions found in this paper of 86.8%. [9]

However, while there has been significant progress in this field, functional utility will require improvement on the tasks related to both facial and emotional classification. The purpose of this study was to improve on current benchmarks in deep CNNs on the task of facial classification and emotional classification through a hierarchical model design. This includes multiple convolutional neural networks trained on a specific task.

III. Methods

A. Presented study, Research question & Hypothesis

Identify applicable funding agency here. If none, delete this text box.

The study attempts to improve the classification of facial emotions in-the-wild through the use of a hierarchical approach with CNNs on publicly available datasets, with the comparison of using two hardware configurations. This study also attempts to use a smaller dataset and publicly available datasets only to complete this task.

B. Applied Research Methods

We wrote two models in Python using the framework Tensorflow. Tensorflow is an open-sourced package created by Google for the purpose of creating end-to-end machine learning models. [10] Our first model is focused on the task of classifying if a figure within the image is a face or not. The other model is focused on the task of classifying faces as displaying positive (joyful) emotions, or not. This structure allows both neural networks to be further reduced to binary classes (yes/no), rather than multiple classes that may create greater complication for the neural network. Along with this, the design of two separate neural networks, instead of one larger neural network, allows for both models to focus on one portion of the task, rather than forcing one large neural net to learn both of the tasks at once. We tried other methods with nonbinary classes, but we found those to be worse in trying to learn the images. This may be due to the limited amount of data when the images are divided into many classes, instead of keeping them to two classes only.

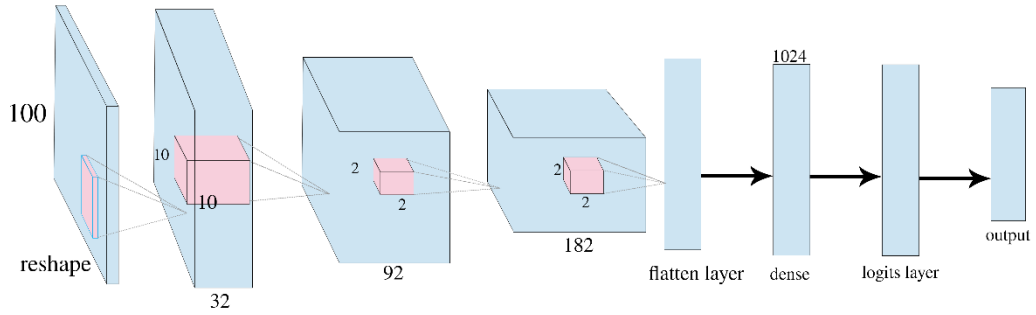

Figure 1. Design of the second model is illustrated above with labels. The first model has one fewer convolutional layers, with 64 features instead of 92.

Data was collected by searching for open source datasets found online; three were selected for specific purposes. Photos were selected at random from each dataset. One is Labeled Faces in the Wild, a database from the University of Massachusetts made available for researchers for studying facial recognition in unconstrained environments, [11] and was chosen to serve as the way for the first neural network to learn what a face looks like and how to classify them. The faces are from the shoulder up, and of random celebrities and other well-known people. Another is called SUN database, which provides a myriad of environmental scenes across many different settings and environments, [12] and was selected to train the other class of the neural network on what is not a face, which includes animals, objects, and other photos surrounded by natural scenes.

For the second neural network, only one database was used to train the neural network on faces that had the emotions happy, sad, or neither, to demonstrate the application of grouping emotions to not only combat the low data size, but to also show the selection process of certain emotions. This database is called AKDEF, otherwise known as The Averaged Karolinska Directed Emotional Faces, [13] also an open source dataset for researchers which has a total of seven emotions for each person, along with five angles to go along with them. This also includes faces that are from the shoulder up, and include various poses as well. Datasets that were used for the “non-face” category were altered to not include photos that had any faces in them, to ensure quality of the training.

C. Analysis Research Model(s) & Instrument(s)

A cluster that has 48 AMD processors and 768 cores, along with 1.5 TB of RAM on a computer using a virtual machine loaded with Ubuntu 18.04 was used in order to run the whole study, in addition to the MATLAB code. The MATLAB code ran separately on a machine that has a i9x processor and 32 GB of RAM on Windows 10 in order to cut the data into 100x100 size, and convert it into black and white format for processing required by the TensorFlow model code.

Prediction was accomplished using two 1080tis Nvidia GPUs in parallel, taking approximately three minutes to perform. Metrics were logged for every 100 steps on both the accuracy and loss for the training and evaluation set split into a 70/30 mix, respectively. For the facial classification task this was done in order to wait for convergence, to see where the optimal values for the model can be observed. After the experiments were run on the cluster, the same methods were applied on a V3-8 Tensor Processing Unit (TPU) using the same Ubuntu OS as the previous method.

Results were observed using TensorBoard to check the results after the model was done training. (TensorBoard is a tool provided by Tensorflow that allows one to visualize the process of training, and also observe the evaluation and other processes; use of this tool allows the researcher to interrupt the often-lengthy training process if underperforming results are expected. Tensorboard also includes features to visualize training data and other features.) After training, the model was applied to the task of prediction with the testing data set. Following the testing of the first model, the second model was evaluated in a similar way, since both of the models used were almost identical besides the weights and bias that were created. The models were saved using Tensorflow's built in functions for saving functions, later to be restored using the same library functions, using the graph feature that is found in the first version of the library.

D. Experiment(s) (setup, sample(s), protocol, briefing of participants)

The first model was trained for 2 hours, 38 minutes, and 22 seconds, and the second model was trained for 14 hours, 17 minutes, and 21 seconds. The datetime package in python was used to time the training to these exact times. The model was also trained many times to see if there would be any improvement from the random weights that are set at the start of the training; the timing results given here are from the final run.

Using the TPU, the same data were cut into a larger image size of 256x256. The first model was trained within 5.3 minutes, and the second model was trained within 13.3 minutes, showing the performance improvement from using Google's cloud-based TensorFlow infrastructure. Both models were trained using 15 steps for each epoch, with the first model trained for 30 epochs and the second model trained on 40 epochs. Prediction was made using the TPU as well, yielding to an improvement on inference time as well.

IV. FINDINGS

A. Collected data

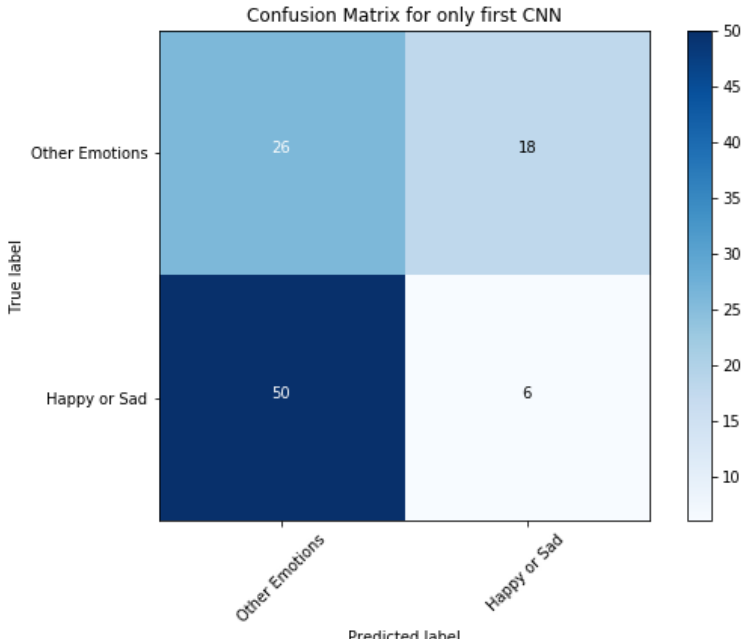

Figure 2. Confusion matrix of results of model two alone without the filter design with the validation set.

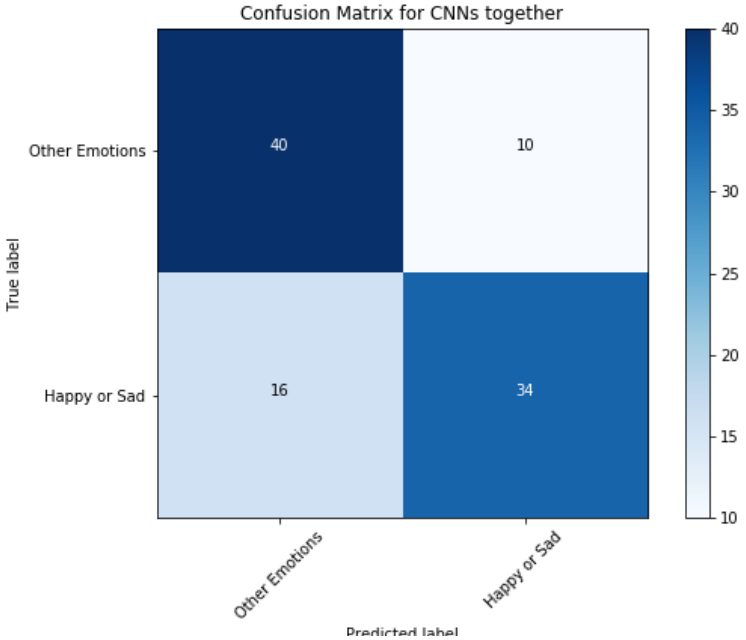

Figure 3. Confusion matrix of results of model two with the filter design with the validation set.

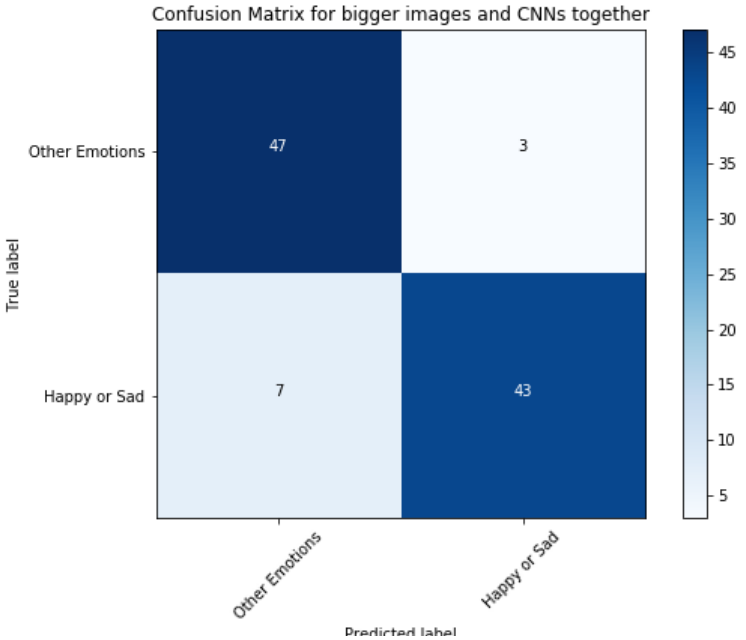

Figure 4. Confusion matrix of results of model two alone without the filter design with the validation set.

The first model achieved an accuracy on the training dataset of 94%, an accuracy of 87% on the evaluation data set, a loss of .18 on the training data set, and a loss of .58 on the evaluation data set. The second model was able to achieve an accuracy on the training dataset of 98%, an accuracy of 85% on the evaluation data set, a loss of 2.2*e^16 on the training data set, and a loss of .13 on the evaluation data set (Figure 2,3).

Using the TPU, the first model had an observed training accuracy of 99%, and a testing accuracy of 94%. The second model had an observed training accuracy of 97%, and a testing accuracy of 96% (Figure 4). The first model has a loss of 0.0028, and the second model has a loss of 0.0933.

B. Analysis of collected data

Figure 5. Comparison between training times between TPUs and CPUs for the same models.

When comparing the model results on the CPU vs. TPU, there was an observed improvement on accuracy of emotion of 90%, but not for the accuracy on the facial classification. This improvement on the accuracies may be attributed to the increase in image size due to the increased computing resources as seen in the training times (Figure 5) and the ability to use larger image size from 100x100 to 256x256.

C. Discussion

TABLE I. SUMMARY OF RESULTS AS SEEN FROM THE THREE TYPES OF MODELS.

Summary of Results	Correctly Classified	Incorrectly Classified	Accuracy
Only One Neural Network	32	68	32%
Both	74	27	74%
Both trained on TPUs	90	10	90%

TABLE II. SUMMARY OF RESULTS AS SEEN FROM THE TWO MODELS AND COMPARED TO OTHER METHODS.

Approach	Error Rate
AlexNet	34.33%
Imagenet	34.33%
Proposed Approach	26%
Proposed Approach with TPUs	10%

Without the first neural network acting as a filter, the results on the validation set on the first method were observed to be 50% (Figure 1). With the first neural network acting as a filter, the results on the validation set were observed to be 74% (Figure 2). Most of the images are successfully classified into face or not a face class, which allowed the second neural network to be able to classify the faces correctly. This is compared to the results all together, as seen in (Table 1). When the second neural network tried to classify the non-face images into the emotional classes, most of the classifications were borderline classes (Figure 1). This confusion could be attributed to the fact that it is trained on the task of emotional classification, and it could possibly be looking for human features that signify emotions, rather than in objects that do not contain those. The confusion matrix of the second neural network without the first neural network as a filter is shown to have a high amount of false positive and false negatives (Figure 1). The confusion matrix of both the first and second neural networks working together show a low amount of false positives and false negatives, with improvement that could be contributing to the first neural network that is acting as a filter. (Figure 2) Most of the outlying cases such as objects being allowed through the filter have been due to the objects being either round or having certain shapes within the image that may be similar to human features.

V. CONCLUSION

A. Limitations

Compared to the datasets mentioned previously, the authors have only used publicly-available datasets, leading to less training data that would occur if the authors collected their own private dataset. This could lead to issues where the model that has been trained to incorrectly assign a label to a photo it has not seen in the public dataset.

Along with this, the model is found to be much slower when using hardware other than the TPU. This shortcoming may lead to the research being out of reach for duplication by some entities.

B. Concluding Remarks

Considering the neural networks as a combined system, the first neural network is able to filter out most of the images that are classified as faces or not to an accuracy of 92%, and then filter the faces into the two groups, happy, sad or neither of those with an accuracy of 74%. Without this filter provided by the first neural network, the second neural network is not able to filter out any of the faces or not faces, and tries to classify them as facial emotions, and thus gives a low accuracy of 32%. This is also compared to the results seen in other research studies [14], where there has been an improvement observed (Table 2). Previous work using a similar idea, but with AlexNet and Imagenet has shown an error of 34.33%, compared to our work that has 26% error rate. [15] It has also been observed that the results in this work have improved on results seen using *Yu et al.* 9-layer convolutional network structure on similar faces in the wild that achieved 61.29% accuracy. [16]

Using the TPU, there has been an observed improvement on the accuracy on the emotions with an accuracy of 90%, but not for the accuracy on the facial classification. The first model had an observed training accuracy of 99%, and a testing accuracy of 94%. The second model had an observed training accuracy of 97%, and a testing accuracy of 96% (Figure 3). The first model has a loss of 0.0028, and the second model has a loss of 0.0933. This improvement on the accuracies may be attributed to the increase in image size due to the increased computing resources (Figure 4) and the ability to use larger image size from 100x100 to 256x256.

C. Future Work

To see improvements on the first CNN results, future studies can change the design of the model, such as changing the hyper parameters or make the model deeper within the hidden layers. This increases the ability of the CNN to pick out features within the images.

Possible applications of this research includes environments within the public, or anywhere outside of a controlled setting. This model can be used for fields such as robotics, video processing, and many more that require the use of faces that are not in a controlled setting. The research can also be used for other applications of in-the-wild faces where there is the need for combining a controlled and non-controlled dataset.

ACKNOWLEDGMENT

I would like to thank my advisor Dr. Benjamin for helping me focus, guiding me along the way, and allowing me to explore my interests in machine learning. Thank you Anqi Zhang for creating the model in one of my figures.

Research supported with Cloud TPUs from Google's TensorFlow Research Cloud (TFRC).

REFERENCES

[1] T. Haunert, (July, 2016) The Rise of Machine Learning. [Online] Available: https://blogs.oracle.com/oraclemagazine/the-rise-of-machine-learning

[2] M. I. Jordan, T. M. Mitchell, (July, 2015) Machine learning: Trends, perspectives, and prospects. [Online] Available: http://www.cs.cmu.edu/~tom/pubs/Science-ML-2015.pdf

[3] W. Liu, Z. Wang, X. Liu, N. Zeng, Y. Liu and F. Alsaadi, "A survey of deep neural network architectures and their applications", Neurocomputing, vol. 234, pp. 11-26, 2017. Available: 10.1016/j.neucom.2016.12.038.

[4] I. Masi, Y. Wu, T. Hassner, P. Natarajan, “Deep Face Recognition: a Survey”, Conference on Graphics, Patterns and Images (SIBGRAPI), October 2018

[5] S. Li and W. Deng, “Deep Facial Expression Recognition: A Survey,” International Conference on Intelligent and Interactive Systems and Applications.

[6] T. Pham, C. Won, “Facial Action Units for Training Convolutional Neural Networks” , IEEE Access, p. 77816 – 77824, 2019.

[7] V. Jha, P. D. Shenoy and V. K. R, "Development of Facial Expression Classifier using Neural Networks," 2019 IEEE International WIE Conference on Electrical and Computer Engineering (WIECON-ECE), Bangalore, India, 2019, pp. 1-4, doi: 10.1109/WIECON-ECE48653.2019.9019937.

[8] B.-K. Kim, S.-Y. Dong, J. Roh, G. Kim and S.-Y. Lee, "Fusing aligned and non-aligned face information for automatic affect recognition in the wild: A deep learning approach", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 48-57, 2016.

[9] N. Sharma and C. Jain, "Characterization of Facial Expression using Deep Neural Networks," 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS), Coimbatore, India, 2019, pp. 492-495, doi: 10.1109/ICACCS.2019.8728386.

[10] M.Abadi, et al. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[11] E. Learned-Miller, G. B. Huang, A. Roychowdhury, H. Li, and G. Hua, “Labeled Faces in the Wild: A Survey,” Advances in Face Detection and Facial Image Analysis, pp. 189–248, 2016.

[12] A. Barriuso and A. Torralba, “Notes on image annotation,” Computer Science and Artificial Intelligence Laboratory (CSAIL), Oct. 2012.

[13] D. Lundqvist, A. Flykt, and A. Öhman, “Karolinska Directed Emotional Faces,” PsycTESTS Dataset, 1998.

[14] A. Bilal, A. Jourabloo, M. Ye, X. Liu, and L. Ren, “Do Convolutional Neural Networks Learn Class Hierarchy?,” IEEE Transactions on Visualization and Computer Graphics, vol. 24, no. 1, pp. 152–162, 2018.

[15] A. Savoiu and J. Wong, “Recognizing Facial Expressions Using Deep Learning,” Recognizing Facial Expressions Using Deep Learning, 2017.

[16] Z Yu and C. Zhang, "Image Based Dtatic Facial Expression Recognition with Multiple Deep Network Learning[C]", Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, pp. 435-442, 2015.