Transfer Learning for Facial Expression Classification

Pierre Bréchet[1], Zhenyu Chen[2], Nicolas Jakob[1], and Sebastian Wagner[2]

[1]Technical University of Munich [2]Ludwig Maximilian University of Munich

Overview

We tackle the task of Facial Expression Classification that is challenging because of the lacking availability of large data sets. To cope with this problem, we combine three commonly used data sets. Most approaches to facial expression classification have relied on training a network from scratch. We train three different architectures, two trained from scratch and one using transfer learning. We show that features from a pre-trained network, trained for pose and age estimation, are useful for fine-tuning and improve the accuracy over models trained from scratch by a large margin. We achieve a state-of-the-art performance [6] of **76.5 % on FER2013** with an architecture that is **far simpler than previous work**.

Data sets

We combined several available data sets to cope with the small size of available data. The different data sets are

1. The **FER 2013** (Kaggle Competition) data set [1]: 30 000 48x48 pixels images of human faces, in unconstrained environments and poses
2. The extended Cohn-Kanade (**CK+**) data set [3]: 167 faces in constrained illumination settings but in posed and non-posed (spontaneous) expressions
3. The Japanese Female Facial Expression (**JAFFE**) data set [4]: 213 images from 10 Japanese models

All images are in gray scale values. The data set merging implied cropping and re-scaling the images to the same size (64x64 pixels).
Even though the relative sizes of the CK+ and JAFFE data sets are small, our hypothesis is that combining them improves classification accuracy.
The resulting labels are **Angry**, **Fear**, **Happy**, **Sad**, **Surprise**, and **Neutral**.

Fig. 1: Excerpt from the data set. From left to right, the emotions are Angry, Fear, Happy, Sad, Surprise, and Neutral

In order to train FaceNet (our most performant model), the images were mirrored, randomly cropped, and randomly rotated with a low angle.

Network Architectures

We apply three different networks for processing the data set. EmoNet was used for verifying the consistency of the combined data set, while EasyNet and FaceNet were used to classify the images – from scratch and using transfer learning respectively.

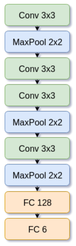

Fig. 2: EmoNet Architecture. It checks if the combined data set is consistent and can be processed later on. Based on [2]

Fig. 3: EasyNet Architecture. EasyNet is an attempt to improve the architecture over the baseline and explore the boundaries of models trained from scratch

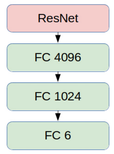

Fig. 4: Module of FaceNet Architecture. FaceNet is a proof-of-concept of how transfer learning works for facial expression classification. Based on VGGFace2 [5]

Training FaceNet

FaceNet takes advantage of transfer learning to outperform EasyNet for our purpose. VGGFace2 is trained on 3.31 million faces, and its features proved to be more relevant for us than networks trained on e.g ImageNet.
For fine-tuning the model, the first 100 layers are fixed and only the 101st layer and upwards are trained with moderate learning rate. In addition, the top layer of the Resnet-50 is removed and replaced by two fully connected layers as well as a final Softmax layer that provides a probability for each of the 6 emotions.
Due to the comparatively large amount of layers being trained on little amounts of data, there is a high sensitivity of the network especially regarding the learning rate and the upper fully connected layers. We therefore recommend validating these hyperparameters when working with this architecture as this can improve results dramatically (more than 13 % in our case).
Also, the whole architecture has been trained using SGD with Nesterov's Momentum. The optimization process was better than when using Adam optimizer.

Classification Results

The best performance on the data set is achieved using FaceNet. While EasyNet gives a 68.9 % test accuracy score on FER2013, FaceNet reaches **76.5% test accuracy on the FER2013** data set. The classification accuracy for FaceNet on FER2013 as compared with previous work can be seen in Fig. 6.

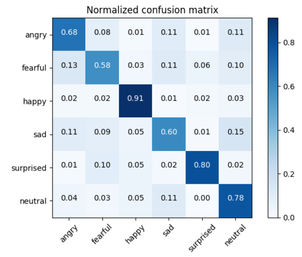

Fig. 5: Confusion matrix for the emotion classification

	Test accuracy
FaceNet	**76.5 %**
State-of-the-art [6]	75.1 %
Kaggle Winner	71.2 %
HappyNet [2]	57.1 %

Fig. 6: The FaceNet accuracy results on FER2013 as compared with previous work

Looking at the confusion matrix in Fig. 5, we can see a disparity between the scores, depending on the detected emotions they refer to. One explanation we have for this result is that we found it even as humans hard to distinguish the emotions without doubt.
Because of our limited amount of time for hyper-parameter tuning, we expect the network to perform even better when carefully tuned and when fed with more data such as the AffectNet dataset.

Conclusion and Further Work

We experiment with emotion detection in facial images based on a wider dataset than commonly used in the literature. We thus propose a combination of three common datasets. Furthermore, we propose an architecture for facial expression classification that is far simpler than previous methods and is almost on-par with the state-of-the-art.

One of the interesting observations we made is that deep networks help a lot in comparison to shallow networks for this task when applied with care.

An interesting direction of research is to add more data sets such as the AffectNet data set to achieve higher generalization. Another direction to explore is how temporal change in videos can be incorporated into the method.

References

[1] url: https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge/data.
[2] Chris English Dean Ducan Gautam Shine. *Facial Emotion Recognition in Real Time*. 2017.
[3] Patrick Lucey et al. *The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression*.
[4] *The Japanese Female Facial Expression (JAFFE) Database*. url: http://www.kasrl.org/jaffe.html.
[5] *VGG Face Descriptor*. 2015. url: http://www.robots.ox.ac.uk/~vgg/software/vgg_face/.
[6] Zhanpeng Zhang et al. *Learning Social Relation Traits from Face Images*. 2015. arXiv: 1509.03936. url: http://arxiv.org/abs/1509.03936.