ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0.
Available from http://www.i6doc.com/en/.

Real-time Convolutional Neural Networks for emotion and gender classification

Octavio Arriaga[1] and Matias Valdenegro-Toro[2] and Paul G. Plöger[3]

[1] AG Robotik, University of Bremen, Bremen, Germany.
[2] German Research Center for Artificial Intelligence, Bremen, Germany.
[3] Hochschule Bonn-Rhein-Sieg, Sankt Augustin Germany.

Abstract.
Emotion and gender recognition from facial features are important properties of human empathy. Robots should also have these capabilities. For this purpose we have designed special convolutional modules that allow a model to recognize emotions and gender with a considerable lower number of parameters, enabling real-time evaluation on a constrained platform. We report accuracies of 96% in the IMDB gender dataset and 66% in the FER-2013 emotion dataset, while requiring a computation time of less than 0.008 seconds on a Core i7 CPU. All our code, demos and pre-trained architectures have been released under an open-source license in our repository at https://github.com/oarriaga/face_classification.

1 Introduction

Most of the human communication is done trough speech, hand gestures and facial expressions. Interpreting correctly any of these elements using machine learning (ML) techniques has proven to be complicated due the high variability of the samples within each task [7]. This leads to models with millions of parameters trained on thousands of samples [5]. Furthermore, the human accuracy for classifying an image of a face in one of 7 different emotions is 65% $\pm$ 5% [7].

In spite of these difficulties, robot platforms (like AILA [10]) oriented to attend and solve household tasks (as described in [3]) require facial expression2s systems that are robust and computationally efficient. These tasks require CNN architectures with millions of parameters; therefore, their deployment in robot platforms and real-time systems becomes unfeasible. In this paper we propose an implement a general CNN building framework for designing real-time CNNs. The implementations have been validated in a real-time facial expression system that provides face-detection, gender classification and that achieves human-level performance when classifying emotions.

2 Related Work

Commonly used CNNs for feature extraction include a set of fully connected layers at the end. Specifically, VGG16 [11] contains approximately 90% of all its parameters in their last fully connected layers. Recent architectures such as Inception V3 [12], reduced the amount of parameters in their last layers by including a Global Average Pooling operation. Global Average Pooling reduces

ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0. Available from http://www.i6doc.com/en/.

each feature map into a scalar value by taking the average over all elements in the feature map, which forces the network to extract global features from the input image. Modern CNN architectures such as Xception [2] leverage from the combination of two of the most successful experimental results in CNNs: the use of residual modules [8] and depth-wise separable convolutions [4].

Furthermore, the state-of-the-art model for the FER-2013 dataset is based on CNN trained with square hinged loss [13]. This model achieved an accuracy of 71% [7] using approximately 24 million parameters. In this architecture 98% of all parameters are located in the last fully connected layers.

The second-best methods presented in [7] achieved an accuracy of 66% using an ensemble of CNNs.

3 Low-Parameter Model with Separable Depthwise Convolutions

We are motivated to build a model with the highest accuracy and lowest number of parameters. Reducing the number of parameters help to overcome two important problems. First, the number of parameter defines computational performance, which we aim to improve. Second, a model with less parameters provides a better generalization under an Occam's razor assumption. Our first model relies on the idea of eliminating completely the fully connected layers. The second architecture combines the removal of the fully connected layer and the inclusion of the combined depth-wise separable convolutions and residual modules.

Following the previous architecture schemas, our initial architecture used Global Average Pooling to completely remove any fully connected layers. This was achieved by having in the last convolutional layer the same number of feature maps as number of classes, and applying a softmax activation function to each reduced feature map. Our initial proposed architecture is a standard fully-convolutional neural network composed of 9 convolution layers, ReLUs, Batch Normalization [9] and Global Average Pooling. This model contains approximately 600,000 parameters. We will refer to this model as *Fully Convolutional VGG-Like*.

Our second model is inspired by the Xception [2] architecture. This architecture combines the use of residual modules [8] and depth-wise separable convolutions [4].

Since our initial proposed architecture removed the last fully connected layer, we reduced further the amount of parameters by eliminating them now from the convolutional layers. This was done trough the use of depth-wise separable convolutions. The main purpose of these layers is to separate the spatial cross-correlations from the channel cross-correlations [2]. They do this by first applying a $D \times D$ filter on every M input channels and then applying N $1 \times 1 \times M$ convolution filters to combine the M input channels into N output channels. Applying $1 \times 1 \times M$ convolutions combines each value in the feature map without considering their spatial relation within the channel.

ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0.
Available from http://www.i6doc.com/en/.

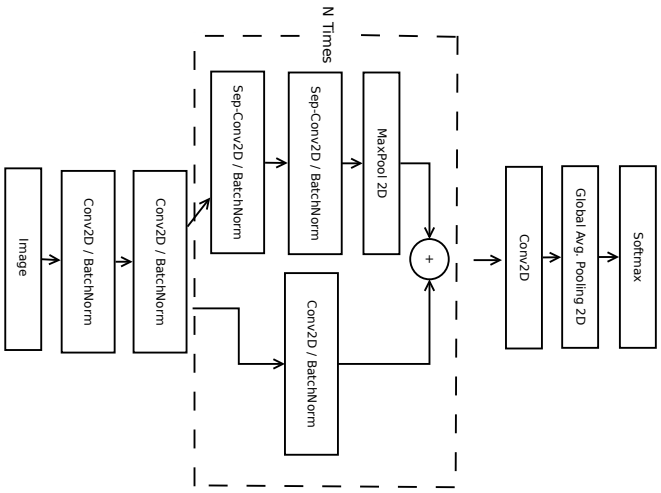


Fig. 1: Our mini-Xception model for real-time classification.

Depth-wise separable convolutions reduces the computation with respect to the standard convolutions by a factor of $\frac{1}{N} + \frac{1}{D^2}$ [4].

Our final architecture is a fully-convolutional neural network that contains 4 residual depth-wise separable convolutions where each convolution is followed by a batch normalization operation and a ReLU activation function. The last layer applies a global average pooling and a soft-max activation function to produce a class prediction. This architecture has approximately $60,000$ parameters; which corresponds to a reduction of $10\times$ when compared to our initial naive implementation, and $80\times$ when compared to the original CNN in [7]. Figure 1 displays our complete final architecture which we refer to as mini-Xception.

4 Experimental Evaluation

We evaluate our models in three datasets: FER 2013 [7] and FER+ [1] datasets for emotion classification. FER 2013 contains 35,887 grayscale images where each image belongs to one of the following classes {*angry*, *disgust*, *fear*, *happiness*, *sadness*, *surprise*, *neutral*}. FER+ is a crowd-sourced relabeling of the data in FER 2013, as it contains many incorrect labels. For gender classification, we evaluate on the IMDB gender dataset, which contains 460,723 RGB images of class *woman* or *man*. We standarized all images to 64×64 pixels.

Table 1 contains our principal result comparison, including emotion classification results from Georgescu et al. [6]. Our proposed models obtain a small loss in accuracy, from $3 - 10\%$, but work well with a considerable reduction in parameters. Particularly the VGG ensemble uses VGG-face, VGG-f, VGG-13, and some engineered features to obtain the best result in the FER and FER+ datasets, but this comes at the expense of over 400 million trainable parameters across different networks in the ensemble. For real-time applications these en-

ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0. Available from http://www.i6doc.com/en/.

Model	# of Params	FER Acc	FER+ Acc	IMDB Acc
FC VGG-Like	600K	66%	78%	96%
mini-Xception	60K	66%	81%	95%
VGG-f [6]	24M	71%	84%	-
VGG ensemble [6]	425M	75%	88%	-

Table 1: Gender and Emotion Classification Results on the FER, FER+, and IMDB datasets

semble models are unusable, and our methods have a higher parameter efficiency with respect to accuracy.

For computational time comparison, we evaluated the computation time on a Core i7-6700HQ CPU using the `timeit` python module. One evaluation of the VGG network takes 0.2 ± 0.01 seconds, a 5 element ensemble takes at least 1 second per image. Our FC VGG-Like takes 0.007 ± 0.002 seconds , while the mini-Xception takes 0.008 ± 0.0001 seconds, which shows the advantage of having a model with a low parameter count. These networks can be used for real-time applications, such as gender and emotion recognition in a service robot.

We also provide a comparison of the learned features between several emotions and both of our proposed models, which can be observed in Figure 2. We can observe that our model learned to get activated by considering features such as the frown, the teeth, the eyebrows and the widening of one's eyes, and that each feature remains consistent within the same class. These results reassure that our models learned to interpret understandable human-like features, that provide generalizable elements. These interpretable results have helped us understand several common misclassification such as persons with glasses being classified as *angry*. This happens since the label *angry* is highly activated when it believes a person is frowning and frowning features get confused with darker glass frames. Moreover, we can also observe that the features learned in our mini-Xception model are more interpretable than the ones learned from our FC VGG-Like. Consequently the use of more parameters in our naive implementations leads to less robust features.

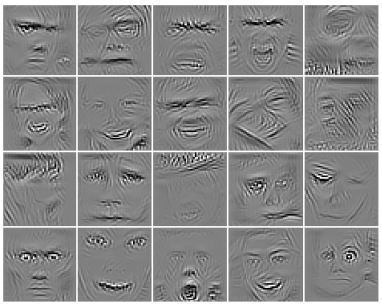
5 Future work

In our specific application we have empirically found that our trained CNNs for gender classification are biased towards western facial features and facial accessories. We hypothesize that this misclassfications occurs since our training dataset consist of mostly western: actors, writers and cinematographers.

Furthermore, as discussed previously, the use of glasses might affect the emotion classification by interfering with the features learned. However, the use of glasses can also interfere with the gender classification. This might be a result from the training data having most of the images of persons wearing glasses assigned with the label *man*. We believe that uncovering such behaviours is

ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0.
Available from http://www.i6doc.com/en/.

(a) (b)



(c)

Fig. 2: Every row starting from the top corresponds to each emotion (a) Samples from the FER-2013 dataset (b) Guided back-propagation visualization of our mini-Xception model (c) Guided back-propagation visualization of our FC VGG-Like model

of extreme importance when creating robust classifiers, and that the use of the visualization techniques such as guided back-propagation will become invaluable when uncovering model biases.

6 Conclusions and Future Work

We have developed a vision system that performs face detection, gender classification and emotion classification in a single integrated system. Our proposed architectures have been systematically built in order to reduce the amount of parameters. Specifically, we achieve 66% accuracy on emotion classification on the FER-2013 dataset, and 95% on gender classification on the IMDB dataset, all while reducing the number of parameters 80 times. Finally, we presented a visualization of the learned features in the CNN using the guided back-propagation visualization, which shows the high-level features learned by our models and discuss their interpretability.

ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0. Available from http://www.i6doc.com/en/.

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*. ACM, 2016.

[2] François Chollet. Xception: Deep learning with depthwise separable convolutions. *CoRR*, abs/1610.02357, 2016.

[3] Markus Eich, Malgorzata Dabrowska, and Frank Kirchner. Semantic labeling: Classification of 3D entities based on spatial feature descriptors. In *IEEE ICRA*, 2010.

[4] Andrew G. Howard et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *CoRR*, abs/1704.04861, 2017.

[5] Dario Amodei et al. Deep speech 2: End-to-end speech recognition in english and mandarin. *CoRR*, abs/1512.02595, 2015.

[6] Mariana-Iuliana Georgescu, Radu Tudor Ionescu, and Marius Popescu. Local learning with deep and handcrafted features for facial expression recognition. *arXiv preprint arXiv:1804.10892*, 2018.

[7] Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al. Challenges in representation learning: A report on three machine learning contests. In *International Conference on Neural Information Processing*. Springer, 2013.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of IEEE CVPR*, 2016.

[9] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, 2015.

[10] Johannes Lemburg, José de Gea Fernández, Markus Eich, Dennis Mronga, Peter Kampmann, Andreas Vogt, Achint Aggarwal, Yuping Shi, and Frank Kirchner. Aila-design of an autonomous mobile dual-arm robot. In *IEEE ICRA*. IEEE, 2011.

[11] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[12] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of IEEE CVPR*, 2016.

[13] Yichuan Tang. Deep learning using linear support vector machines. *arXiv preprint arXiv:1306.0239*, 2013.