Real-time Convolutional Neural Networks for Emotion and Gender Classification

Octavio Arriaga and Paul G. Plöger

Hochschule Bonn-Rhein-Sieg
Department of Computer Science
Grantham-Allee 20, 53757 Sankt Augustin, Germany

octavio.arriaga@smail.inf.h-brs.de paul.ploeger@h-brs.de
Web: www.b-it-bots.de

Abstract. This paper presents the design and implementation of a real-time vision system which accomplishes the tasks of face detection, gender classification and emotion classification simultaneously all in one blended step using convolutional neural networks (CNNs). After presenting the details of the setup and training we proceed to evaluate on standard benchmark sets. We report accuracies of 96% in the IMDB gender dataset and 66% in the FER-2013 emotion dataset. Along with this we also introduced the very recent real-time enabled guided back-propagation (GBP) visualization technique. GBP uncovers the dynamics of weight changes and evaluates the learned features. We argue that the careful implementation of modern CNN architectures, the use of the current regularization methods and the visualization of previously hidden features are necessary in order to reduce the gap between slow performances and real-time architectures. Our system has been validated by its deployment on a Care-O-bot 3 robot used during RoboCup@Home competitions. All our code, demos and pre-trained architectures have been released under a public license in our public repository.

1 Introduction

The success of service robotics decisively depends on a smooth robot to user interaction. Thus a robot should be able to extract information just from the face of its user, e.g. identify the emotional state or deduce gender. 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 [3]. Specifically in the facial expression task, the human accuracy for classifying the image of a face in one of 7 different emotions is $65\% \pm 5\%$. Moreover, the state-of-the-art methods in image-related tasks such as image classification [1] and object detection are all based on Convolutional Neural Networks (CNNs). These tasks require CNN architectures with millions of parameters; therefore, their application in robot platforms and real-time systems becomes unfeasible. Consequently, in this paper we focus on the design and implementation of an open-source, real-time facial expression system that provides face-detection, gender classification

and that achieves human-level performance in emotion classification. This system was developed for general robot platforms and the RoboCup@Home competition challenges. Furthermore, CNNs are used as black-boxes and often their learned features remain hidden, making them complicated to establish a balance between their classification accuracy and unnecessary parameters. Therefore, we implemented a real-time guided-gradient visualization proposed by Springenberg [5] in order to validate the features learned by the CNN.

2 Method

Our initial model used a standard fully-convolutional neural network architecture. This architecture is composed of 9 convolution layers, with rectified linear units (ReLUs) as activations and average pooling. This model contains approximately 600,000 parameters and it achieved an accuracy of 96% in the IMDB gender dataset and 66% in the FER-2013 dataset. The IMDB dataset contains 460,723 RGB images where each image belongs to the class "woman" or "man", while the FER-2013 dataset images consists of 35,887 grayscale images where each image belongs to one of the following classes {"angry", "disgust", "fear", "happy", "sad", "surprise", "neutral"}. The state-of-the-art model for the FER2-2013 dataset is undisclosed and it achieved an accuracy of 71% [3]. Other methods achieved the same accuracy as our model using an ensemble of CNNs which can prove slow for real-time systems [3].

Our second model is inspired by the Xception [1] architecture. This architecture combines two of the most successful experimental assumptions in CNNs: the use of residual modules [4] and depth-wise separable convolutions [2]. Residual modules modify the desired mapping between two subsequent layers, so that the learned features become the difference of the original feature map and the desired features. Depth-wise separable convolutions are composed of two different layers: depth-wise convolutions and point-wise convolutions. The main purpose of these layers is to separate the spatial cross-correlations from the channel cross-correlations [1]. 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. The use of depth-wise separable convolutions reduces the computation with respect to the standard convolutions by a factor of $\frac{1}{N} + \frac{1}{D^2}$ [2]. 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. This architecture has approximately $60,000$ parameters; which corresponds to a reduction of $10\times$ when compared to our initial naive implementation, while achieving accuracies of 95% for gender classification and again 66% for the emotion classification task. Our complete pipeline including the openCV face detection module, the gender classification and the emotion classification takes 0.048 ± 0.015 seconds on a low-end GTX 860M GPU and $.051 \pm 0.022$ seconds on a i5-4210M CPU. This implementation has been successfully implemented as a real-time system in the Care-O-bot 3 robot. We also added to our

implementation a real-time guided back-propagation visualization to observe which pixels in the image activate an element of a higher-level feature map. Given a CNN with only ReLUs as activation functions for the intermediate layers, guided-back propagation takes the derivative of every element (x, y) of the input image I with respect to an element (i, j) of the feature map f^L in layer L. The reconstructed image R filters all the negative gradients; consequently, the remaining gradients are chosen so that they only increase the value of the chosen element of the feature map.

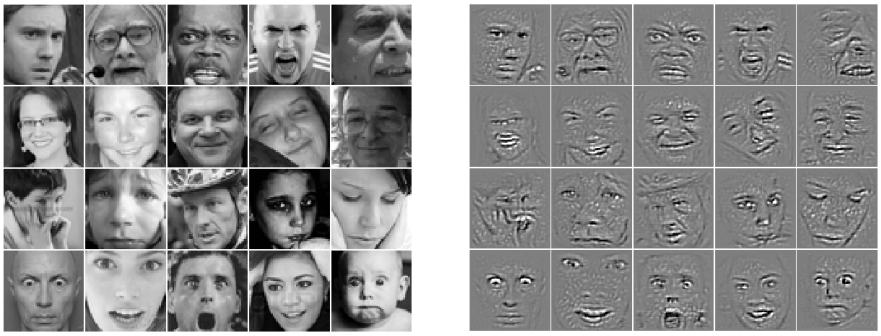
3 Results

Some results from the demos provided in our repository are shown Figure 1. We also provide the scripts to perform both static and real-time inference, training of new architectures and the requirements and the setup installation in our public repository `https://github.com/oarriaga/face_classification`. Our project has been positively received in the open-software community by acquiring over 1.3 thousand stars and over 190 forks.

(a) (b)

Fig. 1: (a) Results of the provided real-time emotion classification demo (b) Results of the provided combined gender and emotion inferences. The color blue represents the assigned class "woman" and red the class "man".

The GBP example can be observed in figure 2. To the authors knowledge this is the first work that visualizes the learned features inside a CNN for face classification. The white areas in figure 2b correspond to a higher activation of a neuron. We can observe that the CNN learned to get activated by considering features such as the frown, the teeth, the eyebrows and the widening of one's eyes. This results reassures that the CNN learned to interpret understandable human-like features, that provide generalizable elements.



(a) (b)

Fig. 2: Both figures contain the same images in the same order. Every row starting from the top corresponds respectively to the emotions {"angry", "happy", "sad", "surprise"} (a) Samples from the FER-2013 dataset (b) GBP visualization.

4 Conclusions

We have created a robot framework-independent vision system that performs face detection, gender classification and emotion classification in a single integrated module. We have achieved human-level performance in our classification tasks using a single CNN that leverages modern architecture constructs. Our complete pipeline performs as a real-time system and has been successfully integrated in a Care-O-bot 3 robot. Finally we presented a visualization of the learned features in the CNN using the guided back-propagation visualization.

References

1. François Chollet. Xception: Deep learning with depthwise separable convolutions. *CoRR*, abs/1610.02357, 2016.
2. Andrew G. Howard et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *CoRR*, abs/1704.04861, 2017.
3. Ian Goodfellow et al. Challenges in Representation Learning: A report on three machine learning contests, 2013.
4. Kaiming He, 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*, pages 770–778, 2016.
5. Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.