A Novel Multi-purpose Deep Architecture for Facial Attribute and Emotion Understanding

Ankit Sharma(✉), Pooyan Balouchian(✉), and Hassan Foroosh(✉)

Department of Computer Science, University of Central Florida, Orlando, FL, USA
{ankit.sharma285,pooyan}@knights.ucf.edu, foroosh@cs.ucf.edu

Abstract. Facial expression estimation has for years been studied benefiting a wide array of application areas ranging from information retrieval and sentiment analysis to video surveillance and emotion analysis. Methods have been proposed to tackle the problem of facial attribute recognition using deep architectures yielding high accuracies, however less efforts exist to focus on the performance of these architectures. Here in this work, we make use of Squeeze-Net [6] for the first time in the literature to perform facial emotion recognition benchmarked on *Celeb-A* and *AffectNet* datasets. Here we extend Squeeze-Net by introducing a new 5×5 convolution kernel after the last fully-connected layer offered by Squeeze-Net, merging the 1×1 and 3×3 outputs from the last fully-connected layers, to perform a more domain-specific feature extraction. We run extensive experiments using widely-used datasets; i.e. *Celeb-A* and *AffectNet*, using AlexNet and Squeeze-Net in addition to our proposed architecture. Our proposed architecture, an extension to Squeeze-Net, yields results inline with state of the art while offering a simple architecture involving less complexity compared to state of the art, reporting accuracies of 90.47% and 56.38% compared to 90.94% and 52.36%, in *Attribute Prediction* and *Expression Prediction* respectively.

Keywords: Attribute prediction · Emotion recognition · Convolutional neural network

1 Introduction

For the past decade, Facial expression estimation has been a subject of attention in the literature due to large array of application areas it can serve. A wide range of salient information is traceable in a human face, including but not limited to age, gender, race, emotion triggered, etc. Application areas include *social media mining*, *face search and retrieval systems* as well as *video surveillance*, to name a few.

Many conventional methods used for feature extraction in the context of computer vision problems have been replaced by convolutional neural networks [8]. CNNs have proved their effectiveness in attribute classification, hence justifying such replacement. However, CNNs have introduced their own challenges,

R. Vera-Rodriguez et al. (Eds.): CIARP 2018, LNCS 11401, pp. 620–627, 2019.
https://doi.org/10.1007/978-3-030-13469-3_72

such as the constant need for relatively large-scale and reliable labeled training samples as well as the complexity involved in modifying the default functionality of different layers in a given network in an attempt to take advantage of transfer learning.

Most works in the literature focus on improving the classification accuracy by introducing highly complex architectures. To address this problem [6] offers a smaller CNN architecture that achieves AlexNet-level accuracy on ImageNet with 50x fewer parameters. Less works in this area, however, have made use of SqueezeNet's architecture specifically to perform Facial Attribute Estimation. Here in this work, not only we make such use of Squeeze-Net, but we also introduce a completely new convolutional layer at the end of the network with a larger kernel size. Benchmarks run on SqueezeNet compared to our proposed architecture, equipped with our newly introduced layer, suggests a high accuracy inline with state of the art, while keeping the micro-architecture a lot simpler than the traditional CNNs performing feature extraction.

The remainder of this paper is organized as follows. Section 2 provides information on the most recent efforts in the literature on the subject of facial attribute estimation and understanding. Next, we discuss our proposed method in Sect. 3 providing details on *Face Attribute Expression* as well as *Face Expression Recognition*. Moreover, we dig into the experimental setup of our proposed micro-architecture in Sect. 4, providing details on the configuration of the proposed CNN and experiments run. Finally, Sect. 5 provides a thorough analysis of the results of experiments, comparing our work against state of the art. Section 6 concludes the paper and provides potential future work directions.

2 Related Works

There exist considerable amount of research on Attribute Estimation taking advantage of different CNN architectures and multi-label classification. The most popular open source packages available for training and testing of deep CNNs include, but are not limited to Caffe [7], TensorFlow [1] and Keras [3]. Deep-Face applied both siamese deep CNN and a classification CNN in order to maximize the distance between impostors and minimize the distance between true matches. Efforts in the field of Face Recognition mainly focus on developing deeper and more complex architectures, yielding relatively higher accuracies at the cost of higher complexity introduced to the architectures.

[5] takes advantage of the discriminative power of CNNs to learn semantic attribute classifiers as a mid-level representation for subsequent use in recognition and verification systems. In a close work to ours, [4] presents a Deep Multi-Task Learning approach to jointly estimate multiple heterogeneous attributes from a single face image. They tackle attribute correlation and heterogeneity with convolutional neural networks (CNNs) consisting of shared feature learning for all the attributes, and category-specific feature learning for heterogeneous attributes. [4] reports an average accuracy of 86.1% for smile and gender classification.

Approaches using hand-crafted and deep learning features can be grouped into two categories: (i) single-task learning of per attribute classifier; and (ii) multi-task learning of a joint attribute classifier. A known caveat with single-task learning is lack of attention to the correlation between the tasks, hence estimating each task separately. Here in this work, however, we propose a multi-task approach where multiple models are learned for multi-attribute estimation using a shared representation. This approach can also be observed in [2] tackling human attribute prediction problem.

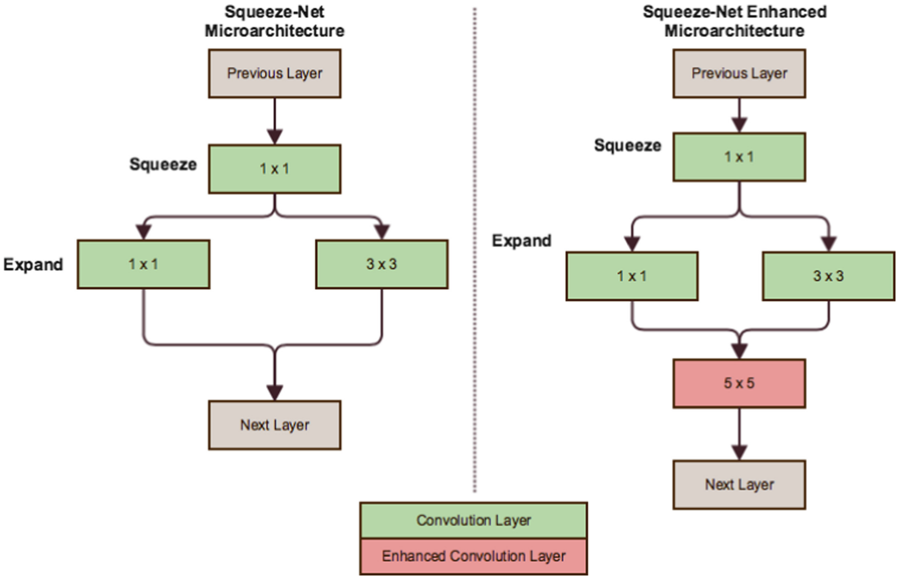

Fig. 1. Microarchitectural view: organization of convolution filters in the **Fire module**. The microarchitecture depicted on the left shows default organization of Squeeze-Net. The microarchitecture depicted on the right shows Squeeze-Net enhanced with the proposed 5 × 5 convolution layer.

3 Our Proposed Method

The goal in this work is to demonstrate that our proposed architecture, for feature extraction, depicted in Fig. 2, outperforms Alex-Net as well as Squeeze-Net architectures specifically when dealing with Face Attribute Estimation. Towards this aim, here we extend the microarchitecture offered by Squeeze-Net as depicted in Fig. 1. Squeeze-Net begins with a standalone convolution layer (conv1), followed by 8 fire modules (fire2–9), ending with a final convolution layer (conv10). The number of filters per fire module increases gradually from the beginning to the end of the network. Squeeze-Net performs max-pooling with a stride of 2 after layers conv1, fire4, fire8, and conv10. A Fire module is comprised of a squeeze convolution layer (only 1×1 filters), feeding into an expand

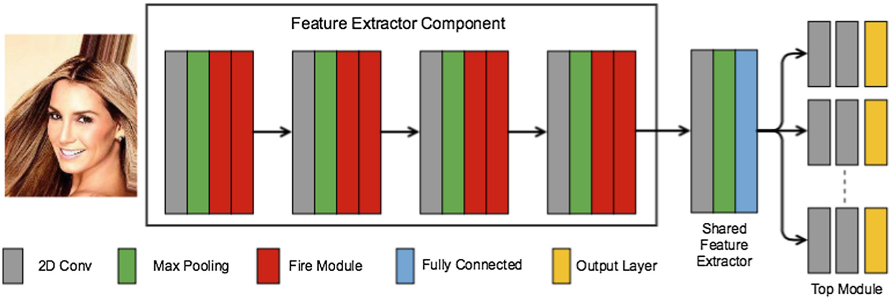

Fig. 2. System architecture depicting 4 blocks, each containing a 2D convolutional layer, two fire modules and max pooling layer, a shared feature extractor, followed by the top module

layer that has a mix of 1 × 1 and 3 × 3 convolution filters. The output of the fire-module is the concatenated outputs of the *expand* layer.

In our proposed micro-architecture, we (i) add a batch normalization layer after the squeeze convolution layer, and (ii) introduce dropout layers after the *expand* layers as depicted in Fig. 1. The result of our experiments further discussed in Sect. 5 demonstrate that our proposed microarchitecture yields a higher generalization ability compared to state of the art.

It is worth mentioning that in fine-grained tasks, such as Face Attribute Prediction and Face Expression Recognition, the objective is to find features that are capable of capturing the subtle highly localized intra-class variations. Therefore, here, we inject our newly developed convolution layer of 5 × 5 filters into our microarchitecture after every few fire modules. We borrow our intuition from the fact that a convolution layer with a larger kernel size provides us with better discriminative features after a sequence of convolution layers with 1 × 1 and 3 × 3 filters.

As depicted in Fig. 1, our microarchitecture begins with a block of two fire modules (squeeze filters, expand filters), followed by a convolution layer of 5 × 5 of expand filters, further followed by a max-pooling layer of size 3 with stride 2. Our experiments, further explained in Sect. 4 are run on the proposed architecture depicted in Fig. 2.

The proposed feature extractor, as part of our proposed architecture, benefits from the following setup:

- A convolution layer of 7 × 7 kernel of 96 filters
- A micro-architecture with (squeeze filters = 16, expand filters = 64)
- A micro-architecture with (squeeze filters = 32, expand filters = 128)
- A micro-architecture with (squeeze filters = 48, expand filters = 192)
- A micro-architecture with (squeeze filters = 64, expand filters = 256)
- A fully-connected layer of size 1024

This feature extractor is attached to a top-network that is domain-specific:

- To perform Face Attribute Prediction, as a multi-label problem, our top network consists of multiple independent sub-networks of fully connected layers.
- To perform Face Expression Recognition, as a multi-classification problem, our top network consists of up of 2 fully connected layers with dropouts.

3.1 Face Attribute Prediction

Face Attribute Prediction is a multi-label classification problem that aims at determining if a given face matches attributes among a set of binary attributes. The *CelebA* [9] dataset benefits from 40 attributes, such as eyeglasses, wearing hat etc. Excerpts from *CelebA* dataset are shown in Fig. 3. In this work, the architecture designed for Face Attribute Prediction makes use of the proposed feature extractor with the top network consisting of 40 independent sub-networks; i.e. the same number as the number of attributes supported by *CelebA* dataset. Each sub-network consists of a fully-connected layer of size 512, followed by a fully-connected layer of size 256 and a final output layer with sigmoid activation.

3.2 Face Expression Recognition

Face Expression Recognition, a multi-classification problem, is a well-studied problem in computer vision. We use a subset of *AffectNet* [10] as our dataset to run experiments for Face Expression Recognition. Here we sub-sampled the dataset in an attempt to avoid the class imbalance problem posed by the original *AffectNet*. Our sub-sampled dataset offers 8 categories with each category containing a maximum of 5,000 images. For the Face Expression Recognition, our proposed architecture makes use of the feature extractor with a top network of fully connected layers of sizes [512,512] and a final layer of size 8 with softmax activation.

Table 1. Experiments table showing benchmarked datasets, number of epochs, number of classes as well as the batch size used

Dataset	No. of epoch	No. of classes	Batch size
CelebA	20	-	64
Affect-Net	50	8	64

4 Experimental Setup

All experiments pointed out in this work are run on a AWS p2.x Large instance with a memory of 61 GiB. The implementation is done using the Keras Python Deep Learning Library. Table 1 shows the number of epochs, batch size and number of classes for each task.

I2 regularization value of 0.0001 is used for the convolution layers for both implementations. Bath size is set to 64 and the learning rate is configured to 0.001. Adam Optimizer was used to perform optimization.

Fig. 3. Excerpts from *AffectNet* and *CelebA* datasets.

Table 2. Attribute estimation accuracies across multiple methodologies benchmarked on *AffectNet* and *CelebA*. AffectNet* refers to the sub-sampled balanced dataset to avoid class imbalance problem.

Problem	Dataset	Method	Accuracy
Attribute prediction	CelebA	PANDA [12]	85%
Attribute prediction	CelebA	Zhong [13]	89.8%
Attribute prediction	CelebA	MOON [11]	90.94%
Attribute prediction	CelebA	SqueezeNet [6]	82.14%
Attribute prediction	CelebA	SqueezeNet-Enhanced	**90.47%**
Attribute prediction	CelebA	Hand [5]	91.26%
Attribute prediction	CelebA	Han [4]	93%
Expression prediction	AffectNet*	AlexNet [8]	52.36%
Expression prediction	AffectNet*	SqueezeNet [6]	48.16%
Expression prediction	AffectNet*	SqueezeNet-Enhanced	**56.38%**

5 Results and Analysis

In this section, we analyze the results reported by state of the art as well as our proposed architecture, depicted in Fig. 2.

5.1 Attribute Prediction Results

As shown in Table 2, our proposed architecture outperforms state of the art in Expression Prediction and yields almost the same accuracy compared to state of the art when tackling Attribute Prediction, while avoiding the complexities introduced in [12,13] and [11]. [4] proposes a method for inferring human attributes, such as gender, hair style, etc., from images of people under large variation of viewpoint pose, appearance, articulation and occlusion, hence offering a part-based model. This method, while yielding reasonable accuracy, requires more training as well as labeled data when compared to our method proposed here. [13], on the other hand, considers mid-level CNN features as an alternative to the high-level ones for attribute prediction. Their intuition is based on the observation that the mid-level deep representations outperform the prediction accuracy achieved by the fine-tuned high level abstractions. This work requires transfer learning as opposed to our proposed methodology, where all features are learned from the beginning of the network from scratch, eliminating the need to perform transfer learning, while not sacrificing the achieved accuracy. In [11], the focus is addressing the multi-label imbalance problem by introducing a novel mixed objective optimization network (MOON) with a loss function that mixes multiple task objectives with domain adaptive re-weighting of propagated loss. This work yields an accuracy of 90.94%, which is closest to the accuracy reported here in this work; i.e. 90.47%. The marginal difference in the accuracy reported by MOON compared to the accuracy reported by our method can be attributed to the fact that MOON's loss function implementation is more complex than our loss function; i.e. standard cross entropy.

5.2 Expression Prediction Results

In order to demonstrate the effectiveness of our proposed method in other application areas, here we run experiments for Expression Prediction on a subsampled balanced version of *AffectNet*. As reported in Table 2, our proposed method offers an accuracy of 56.38%, beating accuracies yielded by AlexNet and SqueezeNet, reported as 52.36% and 48.16% respectively. Here the observation is that the standard AlexNet and SqueezeNet implementations are more challenged to extract localized information compared to our proposed architecture, equipped with our 5×5 convolution filter, as part of its Fire Module.

6 Conclusion

In this work, we propose a novel CNN architecture, an enhanced version of Squeeze-Net, which extends Squeeze-Net's fire module by adding a 5×5 convolution kernel to perform a more accurate feature extraction. To demonstrate the effectiveness of our proposed architecture, we ran experiments on two wildly-used datasets; i.e. *CelebA* and *AffectNet*, across two separate problem domains; i.e. Face Attribute Prediction and Face Expression Recognition. Our results provide proof that while inline with accuracies reported by state of the art; i.e. beating state of the art in Expression Prediction and reporting a very close accuracy in Attribute Prediction, less complexity is involved in the proposed architecture. In the Attribute Prediction and Expression Prediction domains, our system yields accuracies of 90.47% and 56.38% respectively, compared to best accuracies reported by the state-of-the-art methods; i.e. 90.94% and 52.36% on the mentioned domains. Work is currently in progress to run similar experiments with a slightly different architecture; i.e. adding a 7×7 convolution kernel instead of the proposed 5×5 kernel currently in use and analyze the architecture's effectiveness accordingly.

References

1. Abadi, M., et al.: Tensorflow: a system for large-scale machine learning. In: OSDI, vol. 16, pp. 265–283 (2016)
2. Abdulnabi, A.H., Wang, G., Lu, J., Jia, K.: Multi-task CNN model for attribute prediction. IEEE Trans. Multimedia **17**(11), 1949–1959 (2015)
3. Chollet, F., et al.: Keras: deep learning library for theano and tensorflow, vol. 7(8) (2015). https://keras.io/k
4. Han, H., Jain, A.K., Shan, S., Chen, X.: Heterogeneous face attribute estimation: a deep multi-task learning approach. IEEE Trans. Pattern Anal. Mach. Intell. **40**(11), 2597–2609 (2017)
5. Hand, E.M., Chellappa, R.: Attributes for improved attributes: a multi-task network utilizing implicit and explicit relationships for facial attribute classification. In: AAAI, pp. 4068–4074 (2017)
6. Iandola, F.N., Han, S., Moskewicz, M.W., Ashraf, K., Dally, W.J., Keutzer, K.: Squeezenet: alexnet-level accuracy with 50x fewer parameters and < 0.5 mb model size. arXiv preprint arXiv:1602.07360 (2016)
7. Jia, Y., et al.: Caffe: convolutional architecture for fast feature embedding. In: Proceedings of the 22nd ACM International Conference on Multimedia, pp. 675–678. ACM (2014)
8. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
9. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3730–3738 (2015)
10. Mollahosseini, A., Hasani, B., Mahoor, M.H.: Affectnet: a database for facial expression, valence, and arousal computing in the wild. arXiv preprint arXiv:1708.03985 (2017)
11. Rudd, E.M., Günther, M., Boult, T.E.: MOON: a mixed objective optimization network for the recognition of facial attributes. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9909, pp. 19–35. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46454-1_2
12. Zhang, N., Paluri, M., Ranzato, M., Darrell, T., Bourdev, L.: Panda: pose aligned networks for deep attribute modeling. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1637–1644 (2014)
13. Zhong, Y., Sullivan, J., Li, H.: Leveraging mid-level deep representations for predicting face attributes in the wild. In: 2016 IEEE International Conference on Image Processing (ICIP), pp. 3239–3243. IEEE (2016)