# What makes a smile? A Deep Neural Network Point of View

Ivan Čík*, Andrinandrasana David Rasamoelina*, Marián Mach*, Peter Sinčák*
*Department of Cybernetics
and Artificial Intelligence
Technical University of Košice
Slovakia
{ivan.cik, andrijdavid, marian.mach, peter.sincak}@tuke.sk

2021 IEEE 19th World Symposium on Applied Machine Intelligence and Informatics (SAMI) | 978-1-7281-8053-3/21/$31.00 ©2021 IEEE | DOI: 10.1109/SAMI50585.2021.9378634

***Abstract*—Artificial intelligence is the mainstream solution to various problems, and thanks to developments in hardware, it is possible to achieve performance like never before. Continuous data collection provides us with opportunities for the creation of various datasets, which are the basis for various challenges. One of those challenges is recognizing emotions from humans' facial expressions. Multiple deep learning models exist in the wild to solve such a task. They always yield high accuracy on their respective validation and test set. However, the performance of such a model tends to decrease when used on real-world images. This work gives insight into how such a deep learning model can predict facial expression from biases present in training data.**

## I. Introduction

Artificial intelligence algorithms are used today in every field such as healthcare, the military, education, economics, etc. Advance in the field of hardware has enabled the development of more computationally intensive algorithms, the performance of which these algorithms achieved in solving various tasks was unattainable ten years ago. Those algorithms can, nowadays, process various types of data such as text, audio, or video.

Affective computing is one of such domain who got an uplift those last years. Nowadays we have multiple techniques or algorithms to recognize human emotion. It can be from the mean of facial expression, body language [1], voice intonation [2], or other electrical data from invasive sensors (EEG [3, 4], ECG [5, 6], etc. ). Those algorithm or techniques are trained using data collected and annotated by experts (most of the time). But does those data reflect our emotion?

The multiple trained algorithm used in the real world assumes that those annotations are the ground truth. However, it has been shown that human emotion can not be reduced to a set. Furthermore, we see multiple datasets annotated using different conventions or models of emotion. For example in the case of facial expression recognition, we have data annotated using the discrete model (anger, sadness, etc. ) or using the facial action unit or the valance arousal approach. Those data have been collected on a controlled environment setup and therefore the experimenter provided the necessary stimulus to the subject or asked the subject to mimic a certain emotion. After the experimentation, the collected data is annotated by one or multiple domain experts in order to provide the true labels.

Research in facial expression recognition mainly evaluate within a single facial expression database and can achieve promising performances in a controlled laboratory environment. However, lab-controlled databases with posed expressions are too uniform to reflect complex scenarios in real life. In this context, more and more datasets have been collected from the real world for conducting facial expression analysis under unconstrained and challenging conditions (AffectNet [7], EmotioNet [8], Fer2013, Fer+ [9] ), which contain more factors unrelated to facial expressions, such as registration errors and variations in pose, occlusion, illumination, background and subject identity. With this development of diverse facial expression datasets, the dataset bias problem becomes obvious, especially between lab-controlled and real-world data collections. Moreover, in practical application testing, it is hard to satisfy the assumption that the training set and the test sets share the same distribution. In this work, we explore the effect and the usage of those laboratory-controlled data on Deep Neural Networks. We visualize and explore the possible bias incurred by such data to deep learning algorithms.

## II. Convolutional Neural Networks

Regular fully connected Neural Networks use general matrix multiplication to propagate signal from input to output. Convolutional Neural Networks (CNNs) on the other hand are neural networks that use convolution instead of general matrix multiplication in at least one of their layers [10]. CNNs are mostly used to process image data. They are mostly used in computer vision, mainly in face recognition, scene labeling, image classification, action recognition [11], human pose estimation [12], and document analysis [13]. These deep neural networks use different kinds of layers to allow for the computer to discriminate between different features on the subject they deal with. In the field of computer vision they are being more and more widespread as they are understood better, and give the network the ability to learn important filters by itself.

## III. Related Work and Background

Emotions perform many functions in humans, including communication functions when interacting with others. Emotions are the subject of research in several scientific disciplines

such as psychology, ethology, neuroscience, etc. Attempts to objectively measure emotions in humans have their roots in the beginnings of experimental psychology of the nineteenth century [14].

The term "emotion" is a commonly used term, but it is difficult to define precisely. Intuitively, we think that emotion is a physical and physiological phenomenon. But what are the characteristics of emotion? Anger is undoubtedly an emotion. How about love, pressure, tension or hostility?

Computer scientists started to looked at emotion in the nineties. In 1995, Picard [15], a computer scientist at the Massachusetts Institute of Technology (MIT), published an article that laid the foundation of Affective computing [15]. To quote Tan and Picard [14]:

> If we want computers to be genuinely intelligent and to interact naturally with us, we must give computers the ability to recognize, understand, even to have and express emotions.

Emotions have a universal and innate character. They have an adaptive function. The most easily observed performance of them is facial expressions. These are all recognized and undoubtedly play an important role in social interaction. Human faces are an important source of information. In addition to the synchronization between voice and facial animation, facial details also play an important role in conveying emotions [16].

Emotions have a huge influence on human behavior and can have a significant influence on many areas such as: learning, memory, problem solving, humor, perception and health [17–20]. For many years, facial emotion recognition has been an active field in both research and business where lots of challenges were made. Research in this field have made a big success those last years. For e.g. Pramerdorfer and Kampel [21] used modern deep CNN, the test accuracy of 75.2% was achieved without additional training data. Georgescu et al. [22] combined depth features and manual features, an accuracy of 75.42% was obtained on FER 2013. A hybrid Convolutional Neural Network with dense-scale invariant feature transform aggregator was proposed by Connie et al. [23] for facial expression recognition, which achieves 73.4% accuracy on FER-2013. Lee et al. [24] propose an efficient deep convolutional neural network for real-time facial expression inference on embedable device while preserving state-of-the-art accuracy and providing highly efficient and low latency model.

In the following section we explain the method used to interpret and explain the results of DNNs on facial expression images.

## IV. Grad-CAM

Deep neural models are often considered as "black box", meaning that these models are almost impossible to interpret. This phenomenon is caused by the ever-increasing complexity non linearity within those models. Developed algorithms are becoming too complicated, and the process how they map inputs to outputs is often opaque [25]. Explainable Artificial Intelligence is a research area that aims to explain and interpret such algorithms to developers as well as end users. Deep convolutional networks have demonstrated very high predictive performance [26, 27] on very complex datasets like ImageNet [28]. Some of the first successes in interpreting deep neural networks have occurred in the context of image classification. Interpretation methods allow for the first time to open these "black boxes" and gain insights about what the model actually learns and how they arrive at predictions.

In contrast, the gradient-based interpretability method uses the reverse transmission of information in the neural network to understand the influence of neurons and the correlation between input $x$ and output $y$. Gradient-weighted Class Activation Mapping(Grad-CAM) method [29] is a class-discriminative attribution technique used to distinguish the neuronal activity of CNN network neurons. GradCAM, uses the gradients of any class-specific image that flows into the final convolutional layer to generate a map, highlighting the important areas in the image which contributed to particular model output.

With Grad-CAM it is possible to get importance map without re-training or changing the base model. To obtain the class-discriminative localization map Grad-CAM $L^C_{Grad-CAM} \in R^{u \times v}$ of width $u$ and height $v$ for any class $c$, we first compute the gradient of $y_c$ (score for any class $c$) with respect to feature maps $A^k$ of a convolutional layer, $i.e. \frac{\partial y^c}{\partial A^k}$. To obtain the neuron importance weights $\alpha^c_k$:

$$\alpha^c_k = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A^k_{ij}}$$

the gradients that are flowing back are global-average-pooled over the width and height dimensions (indexed by $i$ and $j$). Grad-CAM heat-map is a weighted combination of feature maps, but followed by a ReLU

$$L^C_{Grad-CAM} = ReLU(\sum_k \alpha^c_k A^k)$$

which will remove the pixels with values less than 0.

## V. Gradient SHAP

The Shapley value [30] were proposed by Lloyd Shapley in 1953. It is a classic method to distribute the total gains of a collaborative game to a coalition of cooperating players. Whenever we have a complex model (gradient boosting, neural network or anything that takes certain features as input and produces some predictions as output) and we want to understand the decisions made by the model, we can use the SHAP value. In this method, a data feature can be individual categories in tabular data or superpixel groups in images. The goal of SHAP is to explain the prediction of an instance x by calculating the contribution to the prediction of each feature. One of the advantages of the SHAP method is that the Shapley value explanation is represented as an additive feature attribution method.

Gradient SHAP [31] applies Gaussian noise several times to each input sample, selects a random point along the path

Fig. 1. Samples from Chicago face database

between baseline and input, and calculates the output gradient with respect to the random points selected. The final SHAP values represent the **expected value of gradients * (inputs - baselines)**.

## VI. Chicago Face Dataset

A group of scientists in the field of psychology and the study of emotions has compiled the Chicago Face Database [32]. It is a freely available database consisting of high-resolution, standardized photos of 158 subjects of various ethnicity aged 17 to 65 years, expressing different emotions. Each face in the database had one expression among five : **neutral, angry, fearful, happy with mouth closed, happy with mouth opened**. Each image has a resolution of 2444 (wide) by 1718 (high) pixels, making really small details recognizable. Images were equated for color temperature and placed onto a plain white background. Processing one of these images in the network would take some time, and knowing that our goal was to train on more than a thousand images, that would make the process exhaustive. In addition to these performance issues, it carries the risk of containing too many data points for the network to work with, which would drastically slow down the training process.

## VII. Experiments

Experiments were written in Python using PyTorch which is a open source machine learning library [33] that was built to allow for quick development of machine learning solutions, providing many abstract methods and objects to specialize into the desired parts for our deep convolutional neural network. The high resolution photographs in the Chicago Face Dataset would cause our network to train too long and be sensitive to small details, so to avoid these problems, all the images were resized to 128 by 128 pixels and normalized using the mean and standard deviation of the dataset. That way the data will be processed faster and the training will be easier, while still having enough salient features on the images to be relevant.

### A. Architecture

The neural network that was used in experiments in itself was made of 5 layers. The first 3 layers are intertwined convolution layers and batch normalization layers. The convolution layers let the network pick out features from the picture, and each batch normalization keep the values in ranges that let the network reach useful conclusions followed by a concat pooling layers: one is a maximum pooling, the second an average pooling. Finally, the last 2 layers are fully dense layer used for the classification of features learned by the convolutions.

## VIII. Results

For the training of the network, a learning rate of $0.001$ was chosen after some trial and error. The loss was calculated using the negative log likelihood function of PyTorch. Adam [34] was used as optimizer. The final accuracy acquired on validation dataset were of around 65% across 5 differents run.

### A. GradCam

To add a measure of understanding to the artificial neural network, we decided to add a visualization of the areas that hold a high importance in the decision process of the neural network. To accomplish that, we relied on the Gradient-weighted Class Activation Mapping(Grad-CAM).

Fig. 2. Output of Grad-CAM on facial expression with Neutral Emotion. Note that the pixels that were crucial in recognizing the emotion are mostly allocated to the hair, and we would rather expect it to be on the face.

The GRAD-CAM method did not add much insight into the process of the artificial neural network either. Method that would display this information as heat maps over the images. The result of Grad-CAM (see Fig. 2) of that endeavor is something that we were not expecting: the heat-map displays the hair as being the highest importance for the network, with the facial features coming after in intensity. As an output, we would expect the class-discriminative localization map to be located mainly in the facial area, as we trained our network to recognize emotions using facial expression.

### B. Gradient SHAP

GradientShap makes an assumption that the input features are independent and that the explanation model is linear, meaning that the explanations are modeled through the additive composition of feature effects. Under those assumptions, SHAP value can be approximated as the expectation of gradients that are computed for randomly generated $n$ input samples after adding gaussian noise $n$ times to each input for different baselines/references.

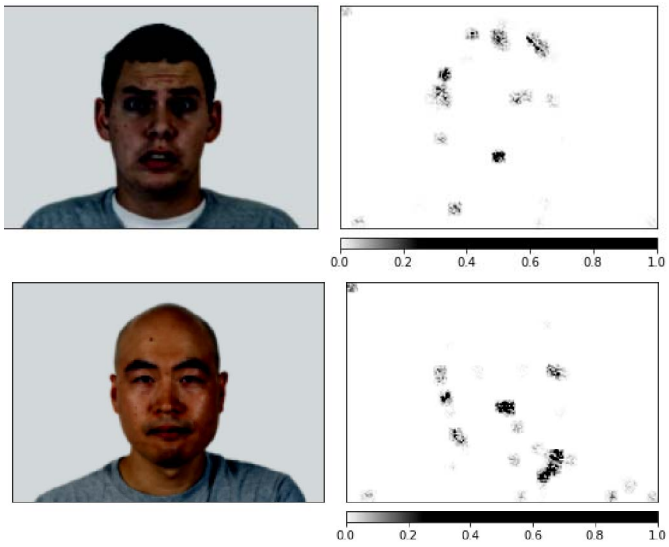


Fig. 3. Output of GradientShap on facial expressions with Neutral Emotion. Note that the pixels that were crucial in recognizing the emotion are on the face, but also on the hair and around the head, which means that our model is not decided mainly on the basis of facial expressions, but also on non-essential features.

The result of GradientShap (see Fig. 3) indicates that black pixels increase the model's output. The input image is shown on the left. Note that the pixels that were crucial in recognizing the emotion are located on the face, but also in the hair area and outside the face.

## IX. Conclusion

Despite the fact that our trained Deep Neural Networks correctly classified those images in fig. 3 and fig. 2 with very high accuracy. Our analysis showed that it did so by focusing on parts of the images that are not so important. This mainly shows that our networks tries to identify patterns that are not important. Another hypothesis and more probable is that the network identified and learned a bias present in the data that the annotator nor we humans did not identify. Such bias generally harms the performance of Deep Neural Networks when used in a real-world application. Therefore choosing the right dataset and transformation is useful for DNN networks to properly learn the foreground pattern. In future works, we will investigate models trained on wild datasets such as AffectNet.

## Acknowledgment

This research work was supported by APVV project 015-0730 "Cloud Based Human Robot Interaction"

## References

[1] L. Abramson, R. Petranker, I. Marom, and H. Aviezer, "Social interaction context shapes emotion recognition through body language, not facial expressions." *Emotion*, 2020.

[2] S. Yacoub, S. Simske, X. Lin, and J. Burns, "Recognition of emotions in interactive voice response systems," in *Eighth European conference on speech communication and technology*, 2003.

[3] Y.-P. Lin, C.-H. Wang, T.-P. Jung, T.-L. Wu, S.-K. Jeng, J.-R. Duann, and J.-H. Chen, "Eeg-based emotion recognition in music listening," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 7, pp. 1798–1806, 2010.

[4] R. Jenke, A. Peer, and M. Buss, "Feature extraction and selection for emotion recognition from eeg," *IEEE Transactions on Affective computing*, vol. 5, no. 3, pp. 327–339, 2014.

[5] J. Cai, G. Liu, and M. Hao, "The research on emotion recognition from ecg signal," in *2009 International Conference on Information Technology and Computer Science*, vol. 1. IEEE, 2009, pp. 497–500.

[6] S. Katsigiannis and N. Ramzan, "Dreamer: A database for emotion recognition through eeg and ecg signals from wireless low-cost off-the-shelf devices," *IEEE journal of biomedical and health informatics*, vol. 22, no. 1, pp. 98–107, 2017.

[7] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18–31, 2017.

[8] C. F. Benitez-Quiroz, R. Srinivasan, Q. Feng, Y. Wang, and A. M. Martinez, "Emotionet challenge: Recognition of facial expressions of emotion in the wild," *arXiv preprint arXiv:1703.01210*, 2017.

[9] E. Barsoum, C. Zhang, C. Canton Ferrer, and Z. Zhang, "Training deep networks for facial expression recognition with crowd-sourced label distribution," in *ACM International Conference on Multimodal Interaction (ICMI)*, 2016.

[10] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*. MIT press Cambridge, 2016, vol. 1.

[11] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 3551–3558.

[12] J. Martinez, R. Hossain, J. Romero, and J. J. Little, "A simple yet effective baseline for 3d human pose estimation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2640–2649.

[13] D. Augusto Borges Oliveira and M. Palhares Viana, "Fast cnn-based document layout analysis," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*, Oct 2017.

[14] J. T. T. Tan and R. W. Picard, "Affective computing and

intelligent interaction," in *Second International Conference, ACII*. Springer, 2007.

[15] R. W. Picard, *Affective computing*. MIT press, 2000.

[16] N. Tan, C. Clavel, M. Courgeon, and J. Martin, "Postural expression of action tendency," in *Proceedings of the ACM Multitimedia 2010's Workshop on Social Signal Processing*, 2010.

[17] H. A. Simon, "Models of Man, Social and Rational: Mathematical Essays on Rational Human Behavior in a Social Setting," *Journal of Philosophy*, vol. 59, no. 7, pp. 177–182, 1962.

[18] A. D. Rasamoelina, F. Adjailia, and P. Sinčàk, "Deep convolutional neural network for robust facial emotion recognition," in *2019 IEEE International Symposium on INnovations in Intelligent SysTems and Applications (INISTA)*. IEEE, 2019, pp. 1–6.

[19] D. Gelernter, *The Muse in the Machine: Computerizing the Poetry of Human Thought*. Simon and Schuster, Jun. 2010, google-Books-ID: 4_UTzq4sZDcC.

[20] D. R. Hofstadter and D. C. Dennett, *The Mind's I: Fantasies And Reflections On Self & Soul*. Basic Books, Jan. 2001, google-Books-ID: SSBJPwAACAAJ.

[21] C. Pramerdorfer and M. Kampel, "Facial Expression Recognition using Convolutional Neural Networks: State of the Art," Dec. 2016. [Online]. Available: https://arxiv.org/abs/1612.02903v1

[22] M.-I. Georgescu, R. T. Ionescu, and M. Popescu, "Local Learning with Deep and Handcrafted Features for Facial Expression Recognition," Apr. 2018. [Online]. Available: https://arxiv.org/abs/1804.10892v6

[23] T. Connie, M. Al-Shabi, W. P. Cheah, and M. Goh, "Facial Expression Recognition Using a Hybrid CNN–SIFT Aggregator," in *Multi-disciplinary Trends in Artificial Intelligence*, ser. Lecture Notes in Computer Science, S. Phon-Amnuaisuk, S.-P. Ang, and S.-Y. Lee, Eds. Springer International Publishing, 2017, pp. 139–149.

[24] J. R. H. Lee, L. Wang, and A. Wong, "Emotionnet nano: An efficient deep convolutional neural network design for real-time facial expression recognition," *arXiv preprint arXiv:2006.15759*, 2020.

[25] W. Samek, T. Wiegand, and K.-R. Müller, "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models," *arXiv preprint arXiv:1708.08296*, 2017.

[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[28] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.

[29] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[30] L. S. Shapley, "A value for n-person games," *Contributions to the Theory of Games*, vol. 2, no. 28, pp. 307–317, 1953.

[31] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in neural information processing systems*, 2017, pp. 4765–4774.

[32] D. S. Ma, J. Correll, and B. Wittenbrink, "The chicago face database: A free stimulus set of faces and norming data," *Behavior research methods*, vol. 47, no. 4, pp. 1122–1135, 2015.

[33] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in neural information processing systems*, 2019, pp. 8026–8037.

[34] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.