Interpretable and Robust Face Verification

Preetam Prabhu Srikar Dammu*[1], Srinivasa Rao Chalamala*[2], Ajeet Kumar Singh[1] and Yegnanarayana Bayya[2]

[1]*TCS Research, Tata Consultancy Services Ltd., India*
[2]*International Institute of Information Technology, Hyderabad, India*

Abstract

Advances in deep learning have been instrumental in enhancing the performance of face verification systems. Despite their ability to attain high accuracy, most of these systems fail to provide interpretations of their decisions. With the increased demands in making deep learning models more interpretable, numerous post-hoc methods have been proposed to probe the workings of these systems. Yet, the quest for face verification systems that inherently provide interpretations still remains largely unexplored. Additionally, most of the existing face recognition models are highly susceptible to adversarial attacks. In this work, we propose a face verification system which addresses the issue of interpretability by employing modular neural networks. In this, representations for each individual facial parts such as nose, mouth, eyes etc. are learned separately. We also show that our method is significantly more resistant to adversarial attacks, thereby addressing another crucial weakness concerning deep learning models.

Keywords

Face Verification, Interpretability, Adversarial Robustness

1. Introduction

Over the last decade, many deep learning methods for face verification have been proposed, a few of them have even surpassed human performance [1, 2, 3, 4]. These deep learning methods, while enabling exceptional performance does not provide reasoning for their predictions. Blindly relying on the results of these black boxes without interpreting the reasons for their decisions could be detrimental especially in critical applications related to medical, financial, and security domains.

In the context of image recognition, various methods have been proposed to tackle interpretability by attempting to reason why an object has been recognized in a particular way. LRP[5], Grad-CAM[6], LIME[7] have been used widely to highlight regions of the image that the models look at for arriving at the final prediction. Despite the existence of several ways post hoc interpretability methods, it is desirable to have a system that is inherently capable of producing interpretations of its decisions. When the latent features generated by the system represent a logical part of an object, it is convenient to infer the contributions of these features to the final prediction.

Though most of the interpretability method procure heatmaps highlighting the regions that contribute to the decision process of the models, in some applications it is still difficult to understand these heatmaps as they are generated at a pixel-level. If these heatmaps can highlight logical visual concepts in the images then it would be more convenient to interpret. (Please refer Figure. 7 and Section. 5.2).

Another significant drawback of deep learning models is their susceptibility to adversarial attacks. Seemingly insignificant noise which is imperceptible to the human eye can fool deep learning models. Numerous black box and white box adversarial attack methods have been proposed in the literature [8, 9, 10].

The problem of detecting and defending adversarial attacks on deep learning models is still largely unsolved. As these attacks on face verification systems pose a serious security threat, it is imperative to develop trustworthy systems. Our motivation behind this work is to integrate both robustness to attacks as well as interpretability into face verification systems.

Hence, in this work, we propose a face verification system that addresses the aforementioned issues by learning independent latent representations of high-level facial features. The proposed method generates intuitive and easily understood heatmaps on the fly, and is also shown to be much more robust against adversarial examples.

2. Related Work

Face recognition is a non-invasive biometric authentication mechanism and has been in commercial use for several years. It has become one of the preferred choice of authentication for mobile device users as it easy to use and avoids the need of remembering passwords. Though people have some reservations against using face recog-

*Equal Contribution

3rd International Workshop on Privacy, Security, and Trust in Computational Intelligence (PSTCI2021)

d.preetam@tcs.com (P. P. S. Dammu*); srinivas.chalamala@research.iiit.ac.in (S. R. Chalamala*); ajeetk.singh1@tcs.com (A. K. Singh); yegna@iiit.ac.in (Y. Bayya)

nition on large scale systems due to privacy issues, it continues to be one of the widely used technologies for identification.

Deep learning based face recognition has surpassed hand crafted feature-based systems and shallow learning systems in performance. In [2], the authors proposed a deep learning architecture called VGGFace for generating facial feature representations or face embeddings. These face embeddings can be further used for identifying the person using a similarity measure or a classifier. DeepID2[11] uses a Bayesian learning framework for learning metrics for face recognition. In FaceNet[12] authors proposed a compact embedding learned directly from images using triplet-loss for face verification. Different loss functions that maximizes intra-class similarity and improves discriminability for faces have been proposed ArcFace[13], CosFace[14], SphereFace[15], CoCo Loss[16].

Existing face recognition models are extremely vulnerable to adversarial attacks even in black-box setting, which raises security concerns and the requisite for developing more robust face recognition models. Adversarial attacks[17, 18, 19] involve additive small, imperceptible and carefully crafted perturbations to the input with the aim of fooling machine learning models. Adversarial attacks allow an attacker to evade detection or recognition or to impersonate another person.[20] described a method to realize adversarial attacks by introducing a pair of eye glasses. These glasses could be used to evade detection or to impersonate others. Another approach for fooling ArcFace using adversarial patches has been proposed in [21]. In [22], the authors have proposed an approach for detecting adversarial attacks on faces.

Understanding and interpreting the decisions of machine learning systems is of high importance in many applications, as it allows verifying the reasoning of the system and provides information to the human expert or end-user. Early works include direct visualization of the filters [23], deconvolutional networks to reconstruct inputs from different layers [24].

Numerous interpretability methods have been proposed in the literature, some of the widely known ones are Layer-wise Relevance Propagation (LRP) [5], Gradient-weighted Class Activation Mapping (Grad-CAM) [25], Grad-CAM++ [26], SHapley Additive exPlanations (SHAP) values [27] and Local Interpretable Model-Agnostic Explanations (LIME) [7]. Most of these techniques attempt to provide pixel-level explanations to indicate the contribution of each pixel to the classification decision. However, these methods are mostly suitable for tasks such as object recognition where the deep learning models only take a single input image.

Recently, a few methods that attempt to explain the behavior and decisions of face recognition systems have emerged [28, 29, 30, 31, 32]. In [28], the authors rely on controlled degradations using inpainting to generate explanations. In [29], visual psychophysics was used to probe and study the behavior of face recognition systems. In [30], the authors propose a loss function that introduces interpretability to the face verification model through training. In [31], the authors use 3D modeling to visualize and understand how the model represents the information of face images. Fooling techniques [32] have also been used for gaining insights on facial regions that contribute more to the decision.

The recently developed explainability methods for face recognition are considerably different from one another in their approach and form of explanations, unlike saliency methods for object recognition which generate similar form of explanations. Each of these methods have their own pros and cons and are suitable for different purposes. We believe our method has certain characteristics that are well-suited for real world applications: easily interpretable feature level explanations, on-the-fly explanations for every prediction, structurally interpretable model architecture, provides feedback in real time and more importantly robust towards adversarial attacks.

3. Interpretable and Robust Face Verification System

Modular neural networks (MNN) [33] are a class of composite neural networks that were inspired by the biological modularity of the human brain. MNNs are composed of independent neural networks that serve as modules, each of them specializing in a specific task. MNNs are inherently more interpretable than monolithic neural networks due to their architecture and divide-and-conquer methodology. MNNs also intrinsically introduce structural interpretability due to their modular structure. Studies have shown that MNNs are better at handling noise than monolithic networks [33]. Several defense mechanisms against adversarial attacks have been proposed in the literature, some of which have employed deep generative models [34, 35]. One of the main motivations for using generative models is their capability of representing information in a lower-dimensional latent space retaining only the most salient features [36].

3.1. Model

3.1.1. Model Composition Overview

In the proposed MNN architecture, we allocate dedicated modules for eyes, nose, mouth and one for the rest of the features. We employ autoencoders to learn separate and distinct latent representations for different facial features. To achieve this, we mask the input image to retain only the region of interest of that specific module and present

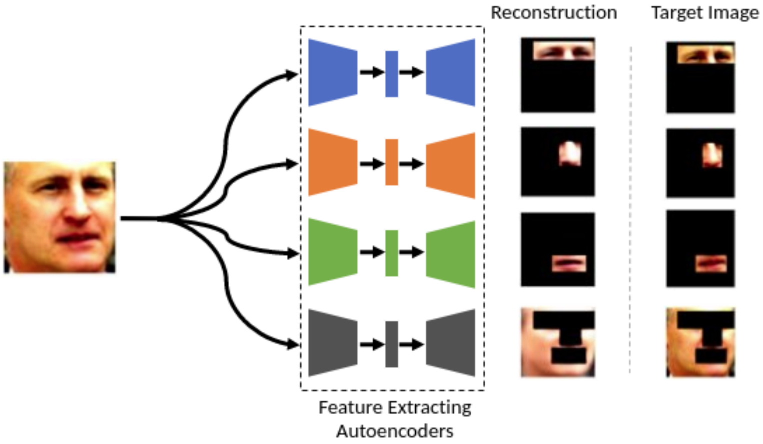

Figure 1: Proposed feature specific latent representations encoding. Images are encoded to feature specific latent representations using feature extracting autoencoders. Reconstructions and corresponding target images are displayed on the right.

it as the target image (See Fig. 1). After the autoencoders have been trained, we retain the encoder and substitute the decoder with Siamese networks in all of the modules, resulting in Modular Siamese Networks (MSN) (See Fig. 2).

In the task of face verification, a pair of images is given as input, which could be either a valid pair or an impostor pair. In the proposed MSN architecture, disentangled embeddings of facial features are generated for both of the input images by the feature extracting encoders present in each feature specific module. These feature embedding pairs are then fed to the Siamese networks present in each module which compute the $L1$ distance vectors for each of the twin feature latent embeddings pairs, similar to the method followed in [37]. The distance vectors from all of the modules are then concatenated and fed to a common decision network which makes the final prediction.

3.1.2. Feature-extracting Autoencoders

In this work, we employ undercomplete autoencoders [36], a type of autoencoder which has a latent dimension lower than the input dimension. Undercomplete autoencoders are trained to reconstruct the original image as accurately as possible while constricting the latent space to a sufficiently small dimension to ensure that only the most salient features are retained in the encoded latent vectors. To achieve our task of extracting feature specific latent vectors, we use a novel technique. In this technique, instead of giving a full image as the target, we mask the input image and retain only a part of the image containing the feature of interest and produce it as the target image. Consequently, the autoencoder learns a latent representation containing important information about the feature and restores only the required part of the image (See Fig. 1, examples in 3.2).

3.1.3. Siamese Networks

Siamese networks have achieved great results in image verification [37, 38]. The two Siamese twin networks share the same weights and parameters. The hypothesis behind this architecture is that if the inputs $x1$ and $x2$ are similar, then the distance between the output vectors $h1$ and $h2$ will be less. The network is trained in such a way that it maximizes the distance between mismatched pairs and minimizes the distance between matched pairs. Loss functions like contrastive loss [39] and triplet loss [40] can be used to achieve this task, few improvised versions of these loss functions have also been proposed in the literature [41, 42].

In our model, we employ Siamese networks for discriminating between feature specific latent vectors of impostors and valid pairs. The latent vectors $x1$ and $x2$ are obtained from the feature-extracting autoencoders described in 3.1.2. L1 distance vectors are computed from the output vectors $h1$ and $h2$ obtained from the Siamese twins for each module. The distance vectors of all of the modules are then concatenated and given as input to the decision network (See Fig. 2).

3.1.4. Decision Network

The decision network is a feed-forward fully connected network that takes the concatenated input from all of the modules. This network enables us to incorporate

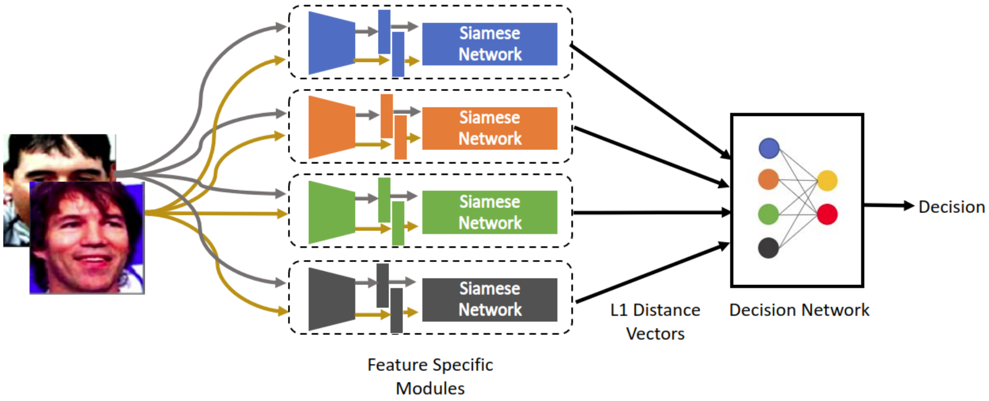

Figure 2: Proposed Modular Siamese Network. Image is initially disentangled by feature-specific encoders to obtain feature-wise embedding pairs, then these embedding pairs are fed to Siamese networks which will compute the distance vectors. All of the distance vectors are then concatenated and fed to the decision network for final verification decision.

information from all of the modules to predict the final decision.

3.1.5. Model Architectural Details

The model architecture and training setting described in [43] were used for training the feature extracting autoencoders. The Siamese networks consist of four fully connected layers with ELU activation functions. The final decision network that takes the concatenated distance vectors from the modules has two fully connected layers with ReLU activation functions.

3.2. Training details

The training of the proposed MSN is carried out in 3 training phases. In the first phase, the feature extracting autoencoders are trained with perceptual loss [43]. In the next phase, the decoder parts in each of the modules are replaced with the Siamese network and trained using the triplet loss, freezing the layers trained in the previous phase. Finally, the decision network is trained using Binary Cross-Entropy (BCE). The Adam optimization technique [44] was used for training the network in all of the three training phases.

From Fig. 3, 4, 5 and 6, we observe that the feature extracting autoencoders are able to generate high quality reconstructions of the intended facial feature. Once training is complete, the autoencoders take unmasked full images as input and reconstruct only the required facial region by incorporating relevant information of that facial feature into the latent feature vector.

The subnetworks can be trained in parallel as they are independent of each other. Once the training is complete, we obtain a complete end-to-end face verification system.

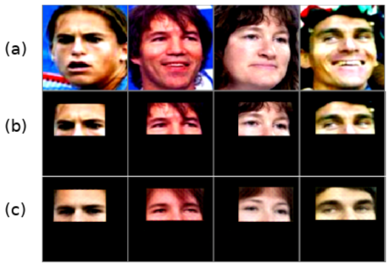

Figure 3: Reconstruction of eyes. (a) input image, (b) masked target image, (c) reconstructed image

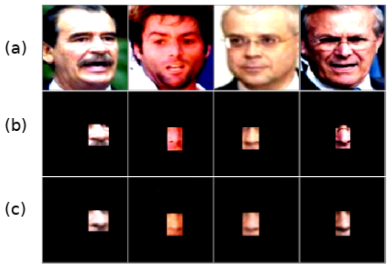

Figure 4: Reconstruction of nose. (a) input image, (b) masked target image, (c) reconstructed image

Facial landmarks used for masking were generated by using MTCNN [45].

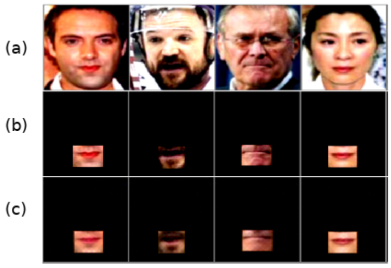

Figure 5: Reconstruction of mouth. (a) input image, (b) masked target image, (c) reconstructed image

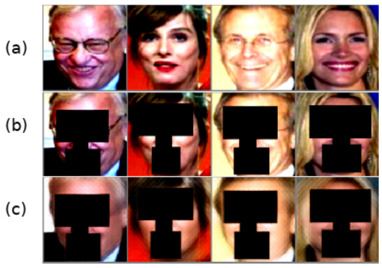

Figure 6: Reconstruction of remaining facial region. (a) is the input image, (b) is the masked target image, (c) is the reconstructed image

4. Interpretability in Modular Siamese Networks

The proposed system generates inherently feature-level heatmaps that are intuitive and easily interpreted, as humans naturally observe the similarity of high-level visual concepts instead of pixels. Each subnetwork of the MSN generates a distance measure that reflects the visual similarity of the features. This is achieved by computing the euclidean distance between the twin output vectors produced by the Siamese networks for each module representing a certain feature. Using these distance measures, a pairwise heatmap incorporating the similarity or dissimilarity of the features is generated and overlayed on both of the images. As can be seen in Fig. 7, the proposed system is able to effectively localize the similarities and dissimilarities of features in a pair of images. These heatmaps could be used as a tool for understanding the decisions taken by the verification system (Refer section 5.2).

5. Experimental Results

The face verification system was trained on the VGGFace2 dataset [46] and evaluated on Labeled Faces in the Wild (LFW) dataset [47]. For reporting performance, we use 10-fold cross validation using the splits defined by LFW protocol which serves as a benchmark for comparison [47].

5.1. Verification

The accuracies of the individual modules and the proposed MSN model have been presented in Table 1. The accuracies for individual modules have been calculated by finding the optimum distance threshold that maximizes accuracy.

No.	Model	Accuracy
1.	Module 1 - Eyes	80.8%
2.	Module 2 - Nose	73.2%
3.	Module 3 - Mouth	74.5%
4.	Module 4 - Rest	78.3%
5.	Modular Siamese Network	98.5%

Table 1
Accuracies of modular siamese network and sub-modules.

We observe that the eyes module outperforms other modules, indicating that it could be the most discriminating feature. The accuracy of MSN is 98.5% which is comparable to the SOTA accuracies that have been reported in the literature which are greater than 99%.

5.2. Feature-level Heatmaps

Feature-level heatmaps are intuitive and easily interpretable as humans, unlike computers, look at features as whole and not at pixels individually. The pairwise heatmaps that are inherently generated by the proposed method incorporate relative information taking both of the input images into consideration. The feature-wise euclidean distances computed by individual modules in MSN are used to generate the heatmaps. As can be seen in Figure. 7, features that look visually similar are colored blue and colored red when dissimilar in all of the images. For true positives, the heatmaps are indicating high similarity for features that are visually close, as expected. The system shows high dissimilarity between the nose regions of the first impostor pair in 5.b, which is in line with human perception as their shapes are significantly different. Studying when the system fails could be helpful, since these visual cues may help rectify the workings of the system. In the first pair of 5.c, we observe that both of the persons wearing eye glasses caused the eyes module to assign low distance score and when accompanied another similar looking feature resulted in misclassification. The heatmap of the second pair of 5.c demonstrates how spectacles and similar looking facial hair fooled the system. The heatmaps in 5.d illustrate

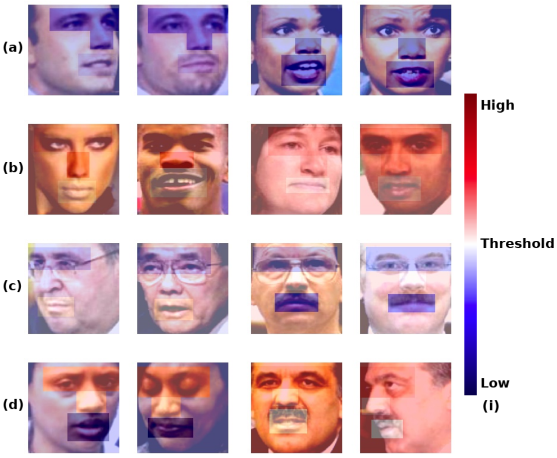

Figure 7: Demonstration of facial feature explanations: Each facial factor and its relevance to face verification. Green indicates similarity while red indicates dissimilarity. (a) True Positives (b) True Negatives (c) False Positives (d) False Negatives (e) Color map indicating dissimilarity. Best viewed in color. (Refer Section. 5.2)

how closing eyes and significant difference in pose can affect the verification. In the first pair, the same person closing eyes in one of the images made the eyes module to compute a high distance score. In the second, significantly different pose which resulted in partial visibility of facial features in one of the images led the system to predict high dissimilarity score.

Since these computations at feature level are carried out in live, the system could instantly generate meaningful messages that can help the user to correct any issues in case of a failure, like removing eye glasses or changing pose for better lighting.

5.3. Performance under adversarial attacks

We tested the robustness and resistance of the proposed method against the widely known adversarial attacks such as the Fast Gradient Sign Method (FGSM) [8], DeepFool [48] and FGSM in fast adversarial training (FFGSM)[49].

Assuming the first image in the two image pairs to be the test image, and the other one to be the anchor image, we attack only test image similar to the experiments conducted in the studies [50, 51]. For comparison, we have considered the well-known FaceNet model which has report SOTA performance earlier. The results have been plotted in Figures. 8, 9 and 10.

The proposed method has shown significantly higher robustness than FaceNet against all three adversarial attacks.

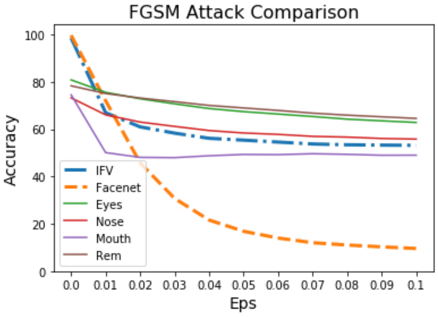

Figure 8: Robustness of proposed approach against FGSM Attack. (IFV: Interpretable and Robust Face Verification system (proposed method))

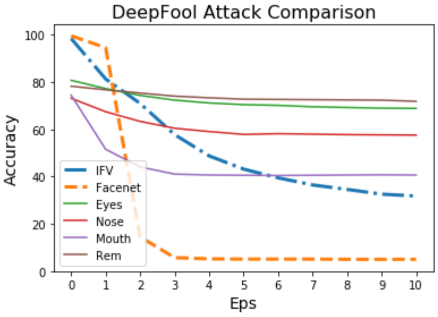

Figure 9: Robustness of proposed approach against DeepFool Attack. (IFV: Interpretable and Robust Face Verification system (proposed method))

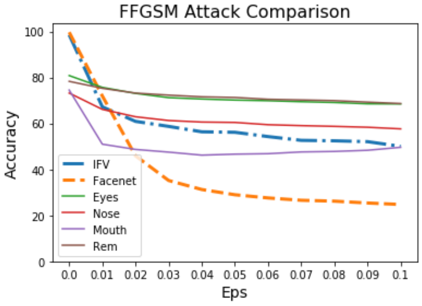

Figure 10: Robustness of proposed approach against FFGSM Attack. (IFV: Interpretable and Robust Face Verification system (proposed method))

For FGSM, the accuracy of FaceNet falls below 20% when ϵ is 0.05 while MSN is still close to 60% accurate (See Figure. 8). In the case of DeepFool attack, we notice a sharp drop of accuracy to below 10% on step 2 in facenet, while MSN shows a lot more resilience by being more than 70% accurate. Similarly for FFGSM, accuracy of FaceNet drops to just above 30% while MSN has an accuracy still above 60% at ϵ equals 0.03. In all of these attacks, we notice that individual modules are noticably more resistant. Since MSN makes the final prediction based on these functionally independent modules, it consequently inherits its robustness from them.

The enhanced robustness could be attributed to the fault tolerant nature of MNN [52, 33]. Additionally, the encoders used for extracting feature specific latent representations are trained to retain only the most salient features because of the bottleneck latent layer and as a result, they may be able to provide some immunity against noise or perturbations.

6. Conclusion and Future Work

Numerous face verification methods have been proposed in the literature, most of which focus solely on improving the performance. Consequently, super-human accuracy has already been achieved in face verification. The real need for improvement in this domain is in the areas of robustness, explainability and fairness. The most important attribute of the proposed method is that it is both robust to adversarial attacks and inherently interpretable. To the best of our knowledge, there is no other published method for face verification that provides both of these qualities at the same time. We believe that pursuing this direction is essential for developing more trustworthy systems.

Having the interpretations of predictions or decisions while they are being taken by deep learning models could prove to be paramount in many applications. While post-hoc interpretations might help in understanding the behavior of the model, they may not be of much help in generating real-time explanations. Incorporating interpretability to the system itself could allow us to handle human errors by enabling communication with the user, informing them of what went wrong and suggesting rectifications.

In this paper, we have presented a new technique to learn latent representations of high-level facial features. We proposed a modular face verification system that inherently generates interpretations of its decisions with the help of the learned feature-specific latent representations. The need and importance of having such a readily interpretable systems were discussed. Further, we have demonstrated that the proposed system a has higher resistance to adversarial examples.

In summary, we have introduced and validated a face verification system that: provides on-the-fly and easily interpretable feature level explanations, has structurally interpretable model architecture, is able to provide feedback in real time, and has increased robustness towards adversarial attacks.

References

[1] N. Kumar, A. C. Berg, P. N. Belhumeur, S. K. Nayar, Attribute and simile classifiers for face verification, in: 2009 IEEE 12th international conference on computer vision, IEEE, 2009, pp. 365–372.

[2] O. M. Parkhi, A. Vedaldi, A. Zisserman, Deep Face Recognition, in: Procedings of the British Machine Vision Conference 2015, British Machine Vision Association, 2015, pp. 41.1–41.12. URL: http://www.bmva.org/bmvc/2015/papers/paper041/index.html. doi:10.5244/C.29.41.

[3] F. Schroff, D. Kalenichenko, J. Philbin, Facenet: A unified embedding for face recognition and clustering, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 815–823.

[4] F. Wang, J. Cheng, W. Liu, H. Liu, Additive margin softmax for face verification, IEEE Signal Processing Letters 25 (2018) 926–930.

[5] S. Bach, A. Binder, G. Montavon, F. Klauschen, K. R. Müller, W. Samek, On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation, in: PLoS ONE, volume 10, Public Library of Science, 2015, p. e0130140. doi:10.1371/journal.pone.0130140.

[6] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization, in: ICCV, 2017, pp. 618–626. URL: http://gradcam.cloudcv.org.

[7] M. T. Ribeiro, S. Singh, C. Guestrin, " why should i trust you?" explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp. 1135–1144.

[8] I. J. Goodfellow, J. Shlens, C. Szegedy, Explaining and harnessing adversarial examples, arXiv preprint arXiv:1412.6572 (2014).

[9] J. Su, D. V. Vargas, K. Sakurai, One pixel attack for fooling deep neural networks, IEEE Transactions on Evolutionary Computation 23 (2019) 828–841.

[10] N. Carlini, D. Wagner, Towards evaluating the robustness of neural networks, in: 2017 ieee symposium on security and privacy (sp), IEEE, 2017, pp. 39–57.

[11] Y. Sun, X. Wang, X. Tang, Deep Learning Face Representation by Joint Identification-Verification, undefined (2014). arXiv:1406.4773v1.

[12] F. Schroff, D. Kalenichenko, J. Philbin, FaceNet: A unified embedding for face recognition and clustering, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, volume 07-12-June, IEEE Computer Society, 2015, pp. 815–823. doi:10.1109/CVPR.2015.

7298682. arXiv:1503.03832.

[13] J. Deng, J. Guo, N. Xue, S. Zafeiriou, Arcface: Additive angular margin loss for deep face recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 4690–4699.

[14] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, W. Liu, Cosface: Large margin cosine loss for deep face recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5265–5274.

[15] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, L. Song, Sphereface: Deep hypersphere embedding for face recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 212–220.

[16] Y. Liu, H. Li, X. Wang, Rethinking Feature Discrimination and Polymerization for Large-scale Recognition, in: undefined, 2017. URL: http://arxiv.org/abs/1710.00870. arXiv:1710.00870.

[17] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, P. Frossard, Universal adversarial perturbations (2016). URL: http://arxiv.org/abs/1610.08401. arXiv:1610.08401.

[18] I. J. Goodfellow, J. Shlens, C. Szegedy, Explaining and harnessing adversarial examples, in: 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings, International Conference on Learning Representations, ICLR, 2015. arXiv:1412.6572.

[19] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, R. Fergus, Intriguing properties of neural networks, in: 2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings, International Conference on Learning Representations, ICLR, 2014. arXiv:1312.6199.

[20] M. Sharif, S. Bhagavatula, L. Bauer, M. K. Reiter, Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition, in: Proceedings of the ACM Conference on Computer and Communications Security, volume 24-28-October-2016, Association for Computing Machinery, New York, New York, USA, 2016, pp. 1528–1540. URL: http://dl.acm.org/citation.cfm?doid=2976749.2978392. doi:10.1145/2976749.2978392.

[21] M. Pautov, G. Melnikov, E. Kaziakhmedov, K. Kireev, A. Petiushko, On adversarial patches: real-world attack on arcface-100 face recognition system, in: 2019 International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON), IEEE, 2019, pp. 0391–0396.

[22] A. J. Bose, P. Aarabi, Adversarial attacks on face detectors using neural net based constrained optimization, in: 2018 IEEE 20th International Workshop on Multimedia Signal Processing (MMSP), IEEE, 2018, pp. 1–6.

[23] M. D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), volume 8689 LNCS, Springer Verlag, 2014, pp. 818–833. doi:10.1007/978-3-319-10590-1_53. arXiv:1311.2901.

[24] M. D. Zeiler, G. W. Taylor, R. Fergus, Adaptive deconvolutional networks for mid and high level feature learning, in: Proceedings of the IEEE International Conference on Computer Vision, IEEE, 2011, pp. 2018–2025. URL: http://ieeexplore.ieee.org/document/6126474/. doi:10.1109/ICCV.2011.6126474.

[25] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, 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.

[26] A. Chattopadhay, A. Sarkar, P. Howlader, V. N. Balasubramanian, Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks, in: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, 2018, pp. 839–847.

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

[28] J. R. Williford, B. B. May, J. Byrne, Explainable face recognition, in: A. Vedaldi, H. Bischof, T. Brox, J.-M. Frahm (Eds.), Computer Vision – ECCV 2020, Springer International Publishing, Cham, 2020, pp. 248–263.

[29] B. RichardWebster, S. Y. Kwon, C. Clarizio, S. E. Anthony, W. J. Scheirer, Visual psychophysics for making face recognition algorithms more explainable, in: V. Ferrari, M. Hebert, C. Sminchisescu, Y. Weiss (Eds.), Computer Vision – ECCV 2018, Springer International Publishing, Cham, 2018, pp. 263–281.

[30] B. Yin, L. Tran, H. Li, X. Shen, X. Liu, Towards interpretable face recognition, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 9348–9357.

[31] T. Xu, J. Zhan, O. G. Garrod, P. H. Torr, S.-C. Zhu, R. A. Ince, P. G. Schyns, Deeper interpretability of deep networks, arXiv preprint arXiv:1811.07807 (2018).

[32] T. Zee, G. Gali, I. Nwogu, Enhancing human face recognition with an interpretable neural network, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, 2019.

[33] A. Schmidt, Z. Bandar, Modularity - a concept for new neural network architectures, Proc. IASTED International Conference on Computer Systems and Applications, Irbid, Jordan, 1998 (1998).

[34] U. Hwang, J. Park, H. Jang, S. Yoon, N. I. Cho, Puvae: A variational autoencoder to purify adversarial examples, IEEE Access 7 (2019) 126582–126593.

[35] P. Samangouei, M. Kabkab, R. Chellappa, Defense-gan: Protecting classifiers against adversarial attacks using generative models, arXiv preprint arXiv:1805.06605 (2018).

[36] I. Goodfellow, Y. Bengio, A. Courville, Deep learning, MIT press, 2016.

[37] G. Koch, R. Zemel, R. Salakhutdinov, Siamese neural networks for one-shot image recognition, in: ICML deep learning workshop, volume 2, Lille, 2015.

[38] O. M. Parkhi, A. Vedaldi, A. Zisserman, Deep face recognition (2015).

[39] R. Hadsell, S. Chopra, Y. LeCun, Dimensionality reduction by learning an invariant mapping, in: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), volume 2, IEEE, 2006, pp. 1735–1742.

[40] G. Chechik, V. Sharma, U. Shalit, S. Bengio, Large scale online learning of image similarity through ranking, Journal of Machine Learning Research 11 (2010) 1109–1135.

[41] W. Chen, X. Chen, J. Zhang, K. Huang, Beyond triplet loss: a deep quadruplet network for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 403–412.

[42] D. Cheng, Y. Gong, S. Zhou, J. Wang, N. Zheng, Person re-identification by multi-channel parts-based cnn with improved triplet loss function, in: Proceedings of the iEEE conference on computer vision and pattern recognition, 2016, pp. 1335–1344.

[43] X. Hou, L. Shen, K. Sun, G. Qiu, Deep feature consistent variational autoencoder, in: 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), 2017, pp. 1133–1141. doi:10.1109/WACV.2017.131.

[44] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).

[45] K. Zhang, Z. Zhang, Z. Li, Y. Qiao, Joint face detection and alignment using multitask cascaded convolutional networks, IEEE Signal Processing Letters 23 (2016) 1499–1503. doi:10.1109/LSP.2016.2603342.

[46] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, A. Zisserman, Vggface2: A dataset for recognising faces across pose and age, in: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), IEEE, 2018, pp. 67–74.

[47] G. B. Huang, M. Ramesh, T. Berg, E. Learned-Miller, Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments, Technical Report 07-49, University of Massachusetts, Amherst, 2007.

[48] S. Moosavi-Dezfooli, A. Fawzi, P. Frossard, Deepfool: A simple and accurate method to fool deep neural networks, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2574–2582. doi:10.1109/CVPR.2016.282.

[49] E. Wong, L. Rice, J. Z. Kolter, Fast is better than free: Revisiting adversarial training, arXiv preprint arXiv:2001.03994 (2020).

[50] F. Zuo, B. Yang, X. Li, Q. Zeng, Exploiting the inherent limitation of l0 adversarial examples, in: 22nd International Symposium on Research in Attacks, Intrusions and Defenses ({RAID} 2019), 2019, pp. 293–307.

[51] M. Kulkarni, A. Abubakar, Siamese networks for generating adversarial examples, arXiv preprint arXiv:1805.01431 (2018).

[52] G. Auda, M. Kamel, Modular neural networks: a survey, International Journal of Neural Systems 9 (1999) 129–151.