Transfer Learning for Deepfake Detection in Static Facial Images

Milica Papić
Blink.ing D.O.O.
Belgrade, Serbia *milica.papic@blinking.id*

Pavle Milošević
University of Belgrade – Faculty of Organizational Sciences
Belgrade, Serbia *pavle.milosevic@fon.bg.ac.rs*

Ivan Milenković
University of Belgrade – Faculty of Organizational Sciences
Belgrade, Serbia *ivan.milenkovic@fon.bg.ac.rs*

Miloš Milovanović
University of Belgrade – Faculty of Organizational Sciences
Belgrade, Serbia *milos.milovanovic@fon.bg.ac.rs*

Miroslav Minović
University of Belgrade – Faculty of Organizational Sciences
Belgrade, Serbia *miroslav.minovic@fon.bg.ac.rs*

Abstract

Digital authentication systems that rely on biometric recognition are especially vulnerable to deepfake attacks, which can be used to impersonate legitimate users and bypass security protocols. As deepfake attacks become increasingly sophisticated, detection methods must evolve rapidly. In this paper, we propose the usage of transfer learning instead of standard deep learning to provide a fast response to novel threats. We evaluate 12 approaches, combining three deep neural networks as feature extractors with four traditional machine learning algorithms as classifiers. Finally, the best-performing model, i.e. ConvNeXt with a support vector classifier, is fine-tuned and evaluated on a real-world dataset, demonstrating strong performance.

Keywords: deepfake, deep neural network, transfer learning, classification.

1. Introduction

The rapid advancements in artificial intelligence (AI) have made it possible to manipulate multimedia content swiftly and easily. AI-generated face images, videos, and speech pose a major challenge to the protection of personal digital identity in today's world [3]. One of the most prominent threats to personal digital identity is considered to be deepfakes. Deepfake refers to generated content, including images, videos, and audio recordings, created using AI methods, primarily through deep neural networks (DNNs) [9]. As the creation of deepfakes through applications has become accessible to a broader population, more of such content is being shared on both conventional and social media. Celebrities are becoming frequent targets [9], politicians' statements are falsified to create division in society [3], etc. Therefore, the detection of deepfakes emerges as one of the most important topics for both practitioners and

researchers in this field. In this paper, we will focus on deepfake images/videos, as a primary threat to digital authentication systems.

Techniques for deepfake detection vary from traditional statistical methods and machine learning (ML) models to applications of deep learning [5], including convolutional neural networks (CNNs), Transformers, and Generative AI approaches. Most commonly, the detection process relies on knowledge of how deepfakes are potentially created, where detection algorithms focus on artefacts left by the creation method [8]. Other approaches offer greater explainability in deepfake detection and try to find exact image areas that have been manipulated [4]. According to most studies, DNNs can relatively easily learn to distinguish real images from deepfakes if they are trained on a sufficiently large dataset generated in a similar manner. However, with advances in technology, such as more sophisticated generative adversarial network (GAN) architectures or other advanced image generators, deepfake content has become a significant threat in this area. The issue arises because the classifier has not learned the discriminative features of these new images. Additionally, the new training procedures are time-consuming and highly dependent on the quantity of images, which is one of the most common limitations.

On the other hand, transfer learning is a well-known ML technique that involves using a model trained on one task as the starting point for a model applied to a second, related task. Starting models, typically DNNs, are trained on large datasets, while smaller datasets are sufficient to adapt the model to a new task [10]. This technique involves changing several layers of the starting model, or using it as a feature extraction model. Transfer learning offers a simple and fast answer to a specific problem, bearing in mind that it saves resources.

In this paper, we aim to explore the potential of using transfer learning to detect visual deepfake attacks. Specifically, we use a pre-trained feature extractor and train only a simple traditional ML classifier when faced with more sophisticated attacks or new technologies. This approach would provide a robust ML framework for detecting deepfake face images, while also allowing for a swift response to novel, more advanced threats that are based on limited data. Furthermore, the best-performing classification model is analyzed and fine-tuned in a real-world environment to demonstrate the practical value of the proposed approach.

2. Problem setup and Data

In this study, we aim to explore the possibility of detecting deepfake attacks based on a single image using transfer learning. The detection is performed on a single still image to align with the requirements of most commercial identity authentication platforms. On one hand, this presents a much more challenging task compared to detection based on a video feed, as there is no interaction with the user and only limited data is available from a single image. However, this approach offers greater convenience to the user, making the authentication process more user-friendly. Transfer learning is used instead of deep learning to reduce model training time and allow for a fast response to novel threats.

The dataset used in the research was obtained from the Kaggle platform [1] and includes a combination of images taken from three large datasets: *FaceForensics++*, *Celeb-DF(v1)*, and *Celeb-DF(v2)*. Since this Kaggle dataset mainly consists of video data that vary in length, frame extraction is performed, followed with face detection in order to reduce background noise and have a consistency in data. Finally, all face images are scaled to 224x224 resolution, since it is image size suitable for all chosen DNNs. For the sake of this experiment, we have randomly selected a subset of images, containing 16,433 face images of all available identities due to limited computer power and limitations of transfer learning approach, i.e. some of chosen ML classification algorithms are not suitable for big data problems.

Additionally, a real-world dataset was collected for testing purposes. In fact, we utilized 1,000 images of real users acquired using the Blink.ing identity authentication platform, along with a certain number of deepfake attacks. To balance our dataset, we supplemented it with images from a Kaggle dataset that were not included in the training dataset. As for the training dataset, face detection and image scaling were performed. Due to GDPR compliance, this dataset cannot be made publicly available and is used only for the purposes of this study.

3. Experimental setup and Results

In this experiment we evaluate the transfer learning approach for deepfake detection, i.e. we utilize DNNs for feature extraction, followed by standard ML algorithms for classification. The experiment has two phases. First, we aim to choose the best combination of deep learning feature extractor and ML classifier on the Kaggle dataset, and then to test the selected model on the real-world dataset and perform threshold optimization.

We have employed three DNNs with different characteristics: Inception [6], a representative of CNNs; DeiT [7], a representative of transformer networks; and ConvNeXt [2], a representative of CNNs that incorporate attention principles. All chosen feature extractors are pretrained on ImageNet-1K dataset and not additionally fine-tuned. The feature vectors are taken from the last fully connected layer and used without any additional preprocessing. As classifies we have applied logistic regression, support vector classifier (SVC), random forest and XGBoost. That resulted in 12 models to be trained and evaluated.

To obtain reliable results, we performed a 5-fold cross-validation procedure combined with a parameter grid search based on the F1 score. During data partitioning, we ensured that there was no identity overlap, i.e. images of a given identity were included exclusively in either the training or the test partition, but not both. In the case of logistic regression and SVC, we have optimized the regularization parameter *C*. For random forest, we have taken into account number of trees, maximal tree depth, and minimal sample leaves. Finally, for the XGBoost algorithm we were focused on the number of trees, maximal tree depth, learning rate, and fraction of features that are randomly selected for building each individual tree. Evaluation of the performance is conducted based on AUC, recall, F1, accuracy and precision metrics. For the final evaluation, we have conducted trial-and-error procedure for determining the optimal threshold value for our particular case.

Table 1 presents a summary of the results for the classifiers based on all three types of feature vectors. The best results for traditional ML classifiers are highlighted in bold, while the best results for DNN extractors are marked with gray shading. The results of the experiment indicate that the choice of deep learning architecture for feature vector extraction significantly impacts the success of deepfake image detection. Models trained on ConvNeXt and DeiT vectors demonstrated notably better performance compared to those trained on Inception vectors. The logistic regression and SVC models trained on vectors extracted from the ConvNeXt and DeiT architectures showed very similar results, with a slight advantage for ConvNeXt. In contrast, all random forest and XGBoost models exhibited substantially worse results, primarily due to a high rate of Type II errors. Finally, the SVC model trained on ConvNeXt feature vectors, with the optimal hyperparameter *C* set to 0.1, was selected as the best performing model.

Table 1. Summary of the classification model results.

Traditional ML classifier	Inception		ConvNeXt		DeiT	
	AUC	Recall	AUC	Recall	AUC	Recall
Logistic regression	**83.12%**	**64.75%**	91.62%	72.40%	90.68%	**76.67%**
SVC	82.42%	63.67%	**91.69%**	**74.42%**	**90.86%**	76.42%
Random forest	74.55%	61.33%	76.51%	53.42%	73.39%	44.83%
XGBoost	75.94%	61.17%	79.51%	54.58%	73.42%	52.25%

Additional test was conducted using the best performing model on real-world Blink.ing dataset in order to determine the generalization ability on data from different sources. Results for chosen different threshold are given in Table 2. Due to the nature of the problem, i.e. higher security of authentication system, lower threshold values are more suitable for real-world application.

Table 2. Overview of SVC model results for different decision thresholds.

Threshold	Accuracy	Precision	Recall	F1
0	82.40%	**99.54%**	65.10%	78.72%
-3.58	**94.35%**	94.04%	**94.70%**	**94.37%**

4. Conclusion

In this paper, we focused on the transfer learning approach to this problem, as it is a fast and easy way to fine-tune the model to tackle novel threats in this turbulent environment. Three feature extraction networks and four standard ML methods have been included in our experiments. First, the transfer learning models are trained and evaluated on the well-known dataset used in literature and for ML competitions. It is shown that the recent neural networks incorporating the attention mechanism, i.e., ConvNeXt and DeiT, outperformed purely convolutional architectures. Also, it is shown that there is no need for sophisticated ML classification algorithms, since logistic regression and SVC proved to be a solid choice. Finally, ConvNeXt in combination with SVC provided the best performance. Afterwards, the best model, ConvNeXt with SVC, is evaluated based on a real-world dataset obtained from a digital identity authentication company, achieving promising results for application.

Acknowledgement

This study was supported by the Innovation Fund of the Republic of Serbia through the project BlinkGUARD (ID: 53154).

References

1. FaceForensics++, CelebV1 & V2-DF Combined Dataset, https://www.kaggle.com/datasets/jimmy98/faceforensics-celebv1-and-v2-df-combined-dataset. Accessed April 17, 2024
2. Liu, Z., Mao, H., Wu, C. Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11976-11986. IEEE (2022)
3. Masood, M., Nawaz, M., Malik, K. M., Javed, A., Irtaza, A., Malik, H.: Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward. Appl. Intell. 53 (4), 3974-4026 (2023)
4. Nguyen, H.H., Fang, F., Yamagishi, J., Echizen, I.: Multi-task learning for detecting and segmenting manipulated facial images and videos. In: 2019 IEEE 10th international conference on biometrics theory, applications and systems (BTAS), pp. 1-8. IEEE (2019)
5. Rana, M.S., Nobi, M.N., Murali, B., Sung, A.H.: Deepfake detection: A systematic literature review. IEEE Access 10, 25494-25513 (2022)
6. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: Proceedings of the AAAI conference on artificial intelligence (AAAI-17) 31 (1). AAAI (2017)
7. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: Proceedings of the 38th International Conference on Machine Learning, pp. 10347-10357. PMLR (2021)
8. Waseem, S., Bakar, S.A.R.S.A., Ahmed, B.A., Omar, Z., Eisa, T.A.E.: DeepFake on face and expression swap: A review. IEEE Access 11, 117865-117906 (2023)
9. Westerlund, M.: The emergence of deepfake technology: A review. Technol. Innov. Manag. Rev. 9 (11), 39-52 (2019)
10. Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., & He, Q.: A comprehensive survey on transfer learning. In: Proceedings of the IEEE 109, pp. 43-76. IEEE (2021)