EfficientNets for DeepFake Detection: Comparison of Pretrained Models

Conference Paper · January 2021

DOI: 10.1109/ElConRus51938.2021.9396092

Alexey Egorov
National Research Nuclear University MEPhI
15 PUBLICATIONS **139** CITATIONS

EfficientNets for DeepFake Detection: Comparison of Pretrained Models

Artem Pokroy
Department of Computer Systems and Technologies
National Research Nuclear University MEPhI (Moscow Engineering Physics Institute)
Moscow, Russia Federation
pokroy-tema@yandex.ru

Alexey Egorov
Department of Computer Systems and Technologies
National Research Nuclear University MEPhI (Moscow Engineering Physics Institute)
Moscow, Russia Federation
egorovalexeyd@gmail.com

***Abstract*—Rapid advances in media generation techniques have made the creation of AI-generated fake face videos more accessible than ever before. In order to accelerate the development of new ways to expose forged videos, Facebook created Deep Fake Detection Challenge (DFDC), which demonstrated multiple approaches to solve this problem. Analysis of top-performing solutions revealed that all winners used pre-trained EfficientNet networks, which was finetuned on videos containing face manipulations. Because of this observation, we decide to compare the performance of EfficientNets models within the task of detecting fake videos. For comparison, we use models, based on the highest-performing entrant of DFDC, entered by Selim Seferbekov, and the DFDC dataset as training data. Our experiments show that there is no strong correlation between model performance and its size. The best accuracy was achieved by B4 and B5 models.**

Keywords——deepfake videos; deep learning; digital media forensics; detection techniques

I. Introduction

Digital image and video manipulation technologies have been developing rapidly for several decades, and one of these technologies, the transformation of human faces on videos, has achieved tremendous results over the last years. Currently, there are several publicly available solutions [1, 2], that allow any user to perform various facial manipulations on high-quality videos, without requiring a profound knowledge of computer science. In the age of information technologies, when social networks are actively used as a source of information, media data modified in this way can spread rapidly and have a serious social impact [3, 4]. This phenomenon is publicly referred to as a *deepfake*.

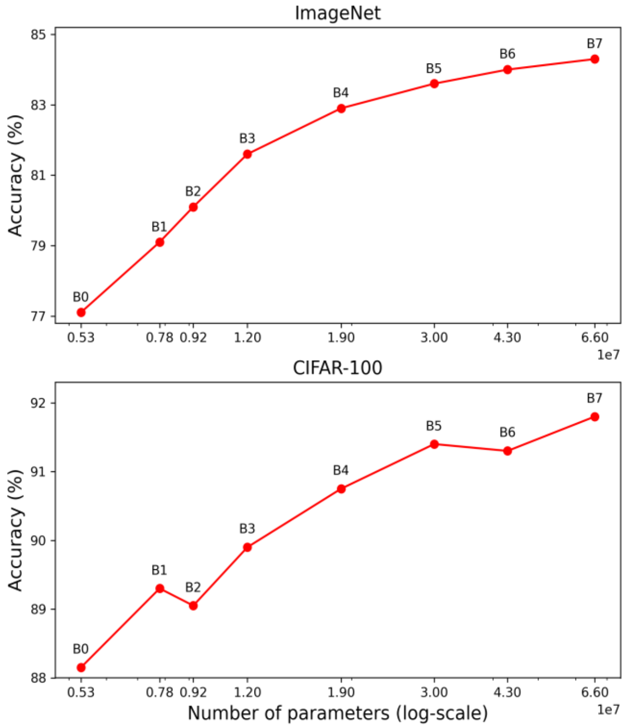
Active development of various deepfake creation methods led to the emergence of research in the field of deepfake detection. These studies gave rise to many deepfake detection methods. Some of these methods search for defects related to human behavior [5, 6], others search for specific artifacts that occur in the process of digital image transformation [7], but most of these methods are data-driven, they are based on sophisticated machine learning models and do not search for specific defects.

So as to accelerate the development of new ways to detect deepfakes, Facebook created Deepfake Detection Challenge (DFDC) [8]. This competition revealed a variety of different approaches to solving this problem. We analyzed the top demonstrated solutions and it appeared that all the winners' solutions used pre-trained models of the EfficientNet family [10], which were finetuned on videos containing facial manipulations.

EfficientNet family of models consists of a single baseline neural network scaled up to various sizes. For all these models there is a dependency between the number of model parameters and its performance. Larger models achieve higher accuracy. But this dependency does not preserve several similar transfer-learning tasks (e.g. CIFAR-10, CIFAR-100, Oxford-IIIT Pets). However, even in these tasks, an overall increase in accuracy was still observed [10]. Figure 1 illustrates the break of dependency with CIFAR-100. In case of deepfake detection, the data model works with is very different from the data it was originally trained on. Therefore, we hypothesize that the tendency to improve accuracy of the model with an increased number of its parameters may not be observed within the deepfake detection problem. If this hypothesis is correct, then while transferring pre-trained EfficientNets to the deepfake detection problem, models with fewer parameters may perform better results.

The analysis of the DFDC winners' solutions and the proposed hypothesis prompted us to compare the performance of pre-trained EfficientNets transferred to deepfake detection task. This comparison will allow us to find out which model achieves the highest performance.

Fig. 1. **The tendency of increasing accuracy.** Generally, as the number of parameters increases, the accuracy increases. But on the CIFAR-100 dataset, models B2 and B6 show worse results than B1 and B5, respectively.



II. Research Materials And Methods

While comparing models, we use a full DFDC dataset consisting of real videos and videos modified by various deepfake creation methods. The task within which we compare EfficientNets is to predict the class of these videos, fake or real. In order to solve this task, we created a baseline classification algorithm inspired by the solution of the winner of the DFDC competition, Selim Seferbekov. The most important step of this algorithm is the use of one model from the EfficientNets family, and since each model has the same output data format, we can simply replace one model with another and compare the algorithm performance without changing other parts of it.

The video classification algorithm we use can be divided into five steps:

- STEP 1: we lower the video frequency by 30 times, then extract an image fragment containing a human face from each frame. After that, we save each video as a sequence of images of people's faces. We use Multi-task Cascaded Convolutional Networks (MTCNN) [9] to detect people's faces.
- STEP 2: for each image in the sequence, we apply a random combination of the following data augmentation methods: rotation by a random angle, flips, blackout random part of the image, Gaussian noise, compression, grayscale, and isotropic resize. During the augmentation process, we do not scale the images and use the highest resolution possible.
- STEP 3: we use the chosen EfficientNet to transform each image in sequence into a feature matrix. In this case, EfficientNets are used as encoders. Each EfficientNet has a strictly defined size of input data, so we pre-scale each image to the required size.
- STEP 4: by using a simple binary classifier, we convert each feature matrix in sequence to a single number. This number represents the probability that the frame originally belongs to a fake video. In fact, we independently classify each frame of the video as fake or real. Figure 3 illustrates the architecture of the classifier we use.
- STEP 5: we average all probabilities in sequence and define the resulting number as the probability that the input video is fake.

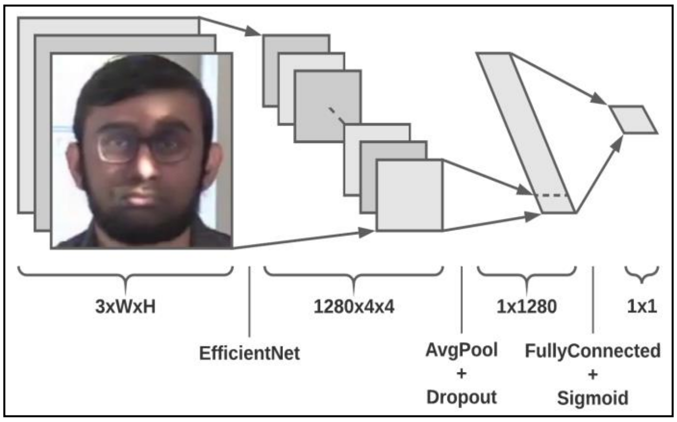


Fig. 3. **Classifier architecture.**

For each pre-trained model from the EfficientNet family we compare (B0-B7), we train our classifier paired with the chosen model. For that, we do 20 epochs of fitting on a train set of videos, computing loss function by each frame separately. Then we test each model by applying our classification algorithm with finetuned EfficientNet and trained classifier to a test set of videos. We use produced predictions to compare EfficientNets using binary classification metrics. Our train and test sets consist of 10000 and 5000 videos from the full DFDC dataset, respectively.

Our work was performed using NRNU MEPhI high-performance computing center, even so, we had to use only a part of the full DFDC dataset and low frame rate of videos, due to the high computational complexity of EfficientNets B6 and B7. However, we used the full DFDS dataset instead of the preview DFDS dataset, since the full release contains videos that have been modified by a larger number of different methods.

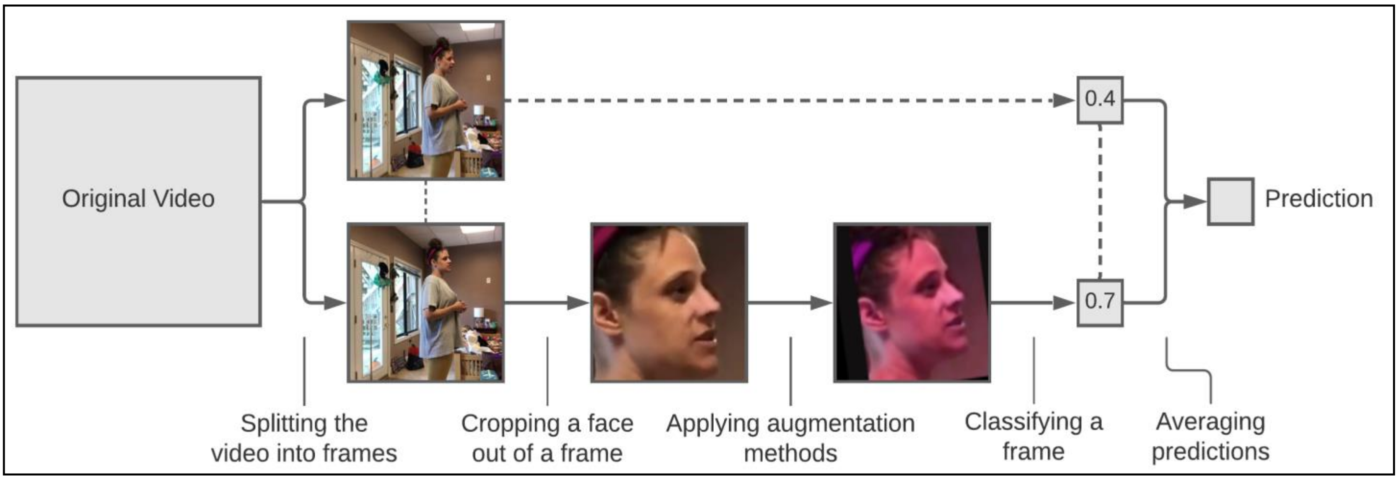

Original Video
0.4
0.7
Prediction
Splitting the video into frames
Cropping a face out of a frame
Applying augmentation methods
Classifying a frame
Averaging predictions

Fig. 2. **Video classification algorithm.**

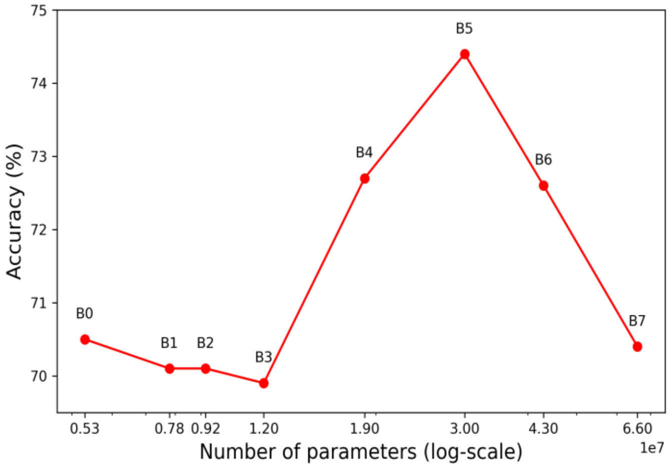
III. Results

Using the computed predictions, we compare the models by two metrics: accuracy and AUC-ROC. The results are presented in Table 1.

Table 1**. EfficientNet Performance Results on DFDC**

Model	Accuracy	AUC-ROC	Params
EfficientNet-B0	70.5	0.785	5.3M
EfficientNet-B1	70.1	0.779	7.8M
EfficientNet-B2	70.1	0.769	9.2M
EfficientNet-B3	69.9	0.785	12M
EfficientNet-B4	72.7	0.828	19M
EfficientNet-B5	**74.4**	**0.829**	30M
EfficientNet-B6	72.6	0.807	43M
EfficientNet-B7	70.4	0.789	66M

Fig. 4. **Model parameters vs. accuracy.**



These results confirm the hypothesis we propose. A tendency to increase the accuracy of the model with an increase in the number of its parameters is not observed in our results. The highest accuracy is achieved by using the B5 model. Figure 4 demonstrates the dependency between the accuracy of the model and the number of its parameters.

IV. Discussion And Conclusions

At present, the problem of deepfake detection is still relevant, but there are already many solutions that allow us to find deepfakes with different accuracy degree. Among these solutions, high results are achieved by methods that use models of the EfficientNet family for feature extraction.

According to our results, the use of pre-trained EfficientNets with a larger number of parameters does not always lead to increase in accuracy. The accuracy of our solution retains similar results when using B0-B3 models. The use of B4 and B5 models it increases and reaches the peak value, but the use of B6 and B7 models leads to a significant decrease in accuracy, despite the great advantage in the number of parameters.

A decrease in accuracy of our solution with the use of B6 and B7 models may be related to the fact that convolutional neural networks of this size begin to work with more complex patterns that are much more difficult to transfer to a different task. In this case, models with a larger number of parameters can potentially achieve better results, but this will require much longer training.

References

[1] I. Perov *et al.*, "DeepFaceLab: A simple, flexible and extensible face swapping framework," *arXiv*, May 2020, Accessed: Nov. 30, 2020. [Online]. Available: http://arxiv.org/abs/2005.05535.

[2] "deepfakes/faceswap: Deepfakes Software For All." https://github.com/deepfakes/faceswap (accessed Nov. 30, 2020).

[3] "Tech - Disinfo and 2020 Election — NYU Stern Center for Business and Human Rights." https://bhr.stern.nyu.edu/tech-disinfo-and-2020-election (accessed Nov. 30, 2020).

[4] R. Chesney and D. K. Citron, "Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security," *SSRN Electron. J.*, Aug. 2018, doi: 10.2139/ssrn.3213954.

[5] Y. Li, M. C. Chang, and S. Lyu, "In Ictu Oculi: Exposing AI created fake videos by detecting eye blinking," Jan. 2019, doi: 10.1109/WIFS.2018.8630787.

[6] X. Yang, Y. Li, and S. Lyu, "Exposing Deep Fakes Using Inconsistent Head Poses," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing -*

Proceedings, May 2019, vol. 2019-May, pp. 8261–8265, doi: 10.1109/ICASSP.2019.8683164.

[7] F. Matern, C. Riess, and M. Stamminger, "Exploiting visual artifacts to expose deepfakes and face manipulations," in *Proceedings - 2019 IEEE Winter Conference on Applications of Computer Vision Workshops, WACVW 2019*, Feb. 2019, pp. 83–92, doi: 10.1109/WACVW.2019.00020.

[8] "Deepfake Detection Challenge | Kaggle." https://www.kaggle.com/c/deepfake-detection-challenge (accessed Nov. 30, 2020).

[9] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016, doi: 10.1109/LSP.2016.2603342.

[10] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," *36th Int. Conf. Mach. Learn. ICML 2019*, vol. 2019-June, pp. 10691–10700, May 2019, Accessed: Nov. 30, 2020. [Online]. Available: http://arxiv.org/abs/1905.11946.