CNN-generated images detection

Deep Learning Final Project

Michele Sanna, Cecilia Zagni

Università degli Studi di Trieste
Data Science and Scientific Computing

January 10, 2024

Table of contents

1. Introduction
2. Paper description
3. Introduction to ResNet18 and EfficientNet
4. Project overview
5. Our results

Introduction

The objective of this project is to replicate the work proposed by Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, Alexei A. Efros in the paper **CNN-generated images are surprisingly easy to spot...for now** [4].

The goal of the paper is to build a **universal detector** for CNN generated images. Universal means that the "real-or-fake" classifier should work not only on the networks it was directly trained on, but also generalizes on future unseen networks.

We tried to reproduce the results of the paper using two smaller nets, ResNet18 and EfficientNet, both pretrained on ImageNet.

Paper description

Since the goal of the experiments is to train a detector able to work on future networks, one can think of training it on images generated by all the available CCN architectures. But we don't have the future model rigth now, so the detector was trained on images generated by only one network, and tested on other CNN-synthesized images, to see how well it generalizes.

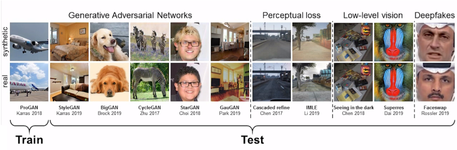
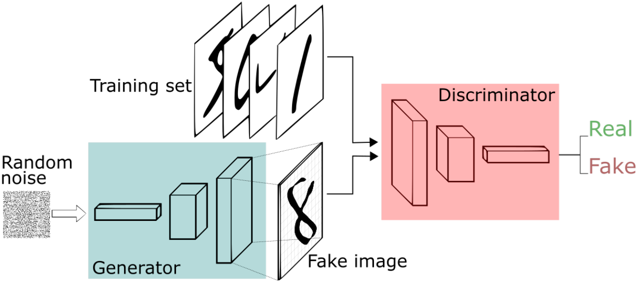


Figure: Image source [4]

For the train images ProGAN was chosen, because it generates high quality images and has a simple convolutional network structure.
For the classifier, the authors chose ResNet-50 pre-trained with ImageNet, and trained it in a binary classification setting.

Paper description - GAN generated images

The GANs (Generative Adversarial Nets) [1] are a type of neural network intended to generate synthetic data (in our case images), trained through a process of competition between a generative model and a discriminative model.



This simple game produces very convincing images, often indistinguishable from real ones by a human eye

Paper description

The dataset was trained on 720K images and validate on 4K.

To evaluate the model's performance, the average precision (AP) for each test set is computed separately and reported.

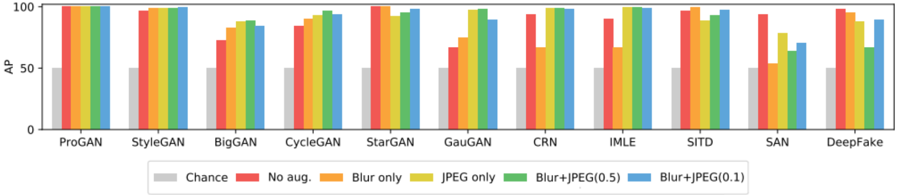


Figure: Image source [4]

The result of the model trained with no augmentation applied were above chance, but the performances were inconsistent. Hence several data augmentation were applied to avoid overfitting. In general this improved the classification performances (exceptions are super-resolution and DeepFake).

Paper description

The dataset contains images belonging to 20 different LSUN categories. To check whether image diversity improves performance, the authors performed several experiments varying the number of classes in the dataset used to train the real-or-fake classifier.

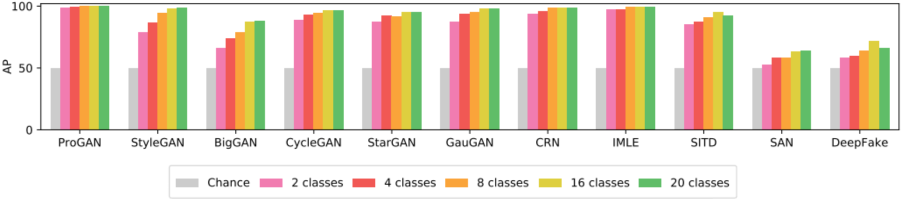


Figure: Image source [4]

Increasing the training set diversity improves performance only up to a point. When the number of classes used increases from 2 to 16, AP consistently improves, minimal improvement is observed when increasing from 16 to 20 classes.

Paper description - conclusions

These experiments suggest that:

1. Today's CNN-generated images retain detectable fingerprints that distinguish them from real photos,
2. These common artifacts can be detected by a classifier. The trik to make it work well are:
   - Large-scale dataset
   - Data augmentation
   - Long training time
3. The situation may not persist because of the advancement in technology.

ResNet18

We trained the first models with ResNet18, pretrained on ImageNet.

Here the ResNet18 architecture, each dotted module in the picture is a residual block.

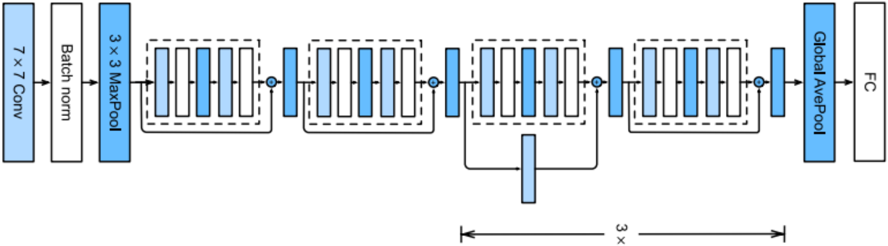


Figure: Image source [2]

Residual Block

A detail of a residual block. Left: standard version. Right: version with 1x1 convolution, to allow a change in the number of channels between the input to the block and the output.

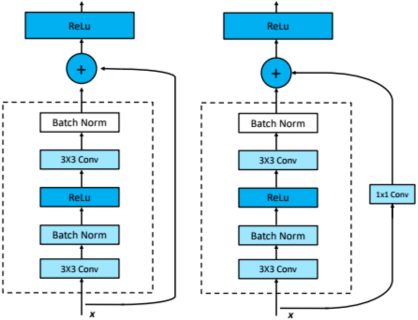
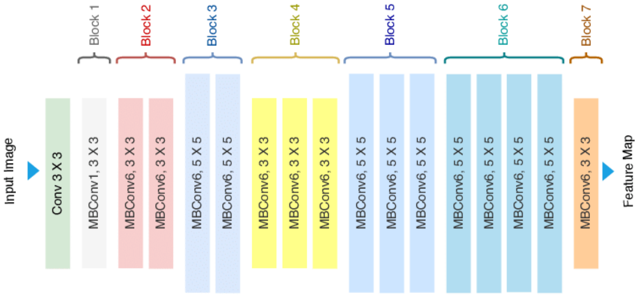


Figure: Image source [2]

EfficientNet

As the second model, we're using an EfficientNet, more precisely the EfficientNet B0, the smaller of the models presented in the EfficientNet paper [3].



In the related paper [3] it's shown that this model has an accuracy on ImageNet close to the ResNet-50's, so we expect similiar performance with respect to [4].

Project overview

Due to the long time necessary to set up the dataset and training the model, we couldn't reproduce all the experiments of the paper.

We chose to train two different nets (ResNet18 and EfficientNet) and test them on all the test sets, in order to confront the results with the paper in terms of AP.

The parameters used for both the trainings were:

- Loss function: Cross Entropy Loss
- Oprimizer: Adam ($\beta_1 = 0.9, \beta_2 = 0.999$)
- Number of classes: 20
- Batch size: 64
- N° of epochs: 20

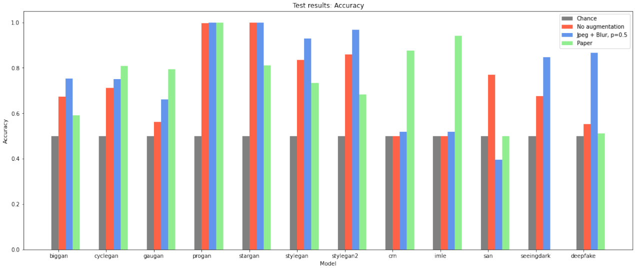
We implemented also the possibility of dropping the LR by 10× if after 5 epochs the validation accuracy does not increase by 0.1%, and the stopping at LR 10^{-6}, but we didn't use this criterion due to time limitations.

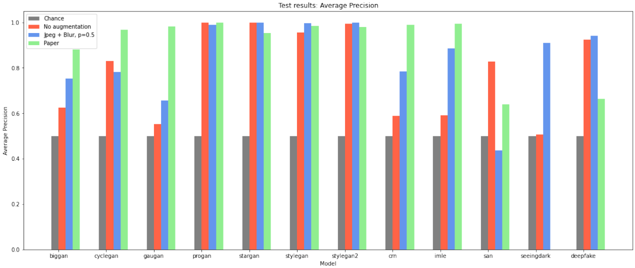
Project overview

We performed 2 trainings for each net: one with no augmentations (only 224x224 cropping), and the other with the following transormations:

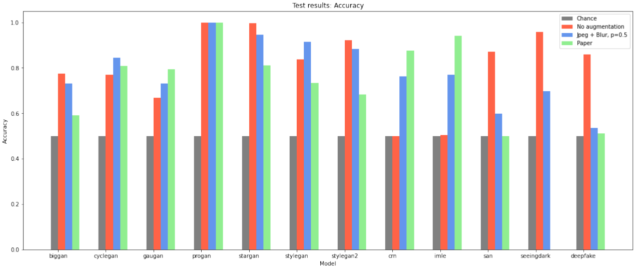
- 224x224 cropping
- JPEG with 50% probability
- Random horizontal flip
- Gaussian blur ($\sigma \sim \mathit{Uniform}[0, 3]$) with 50% probability

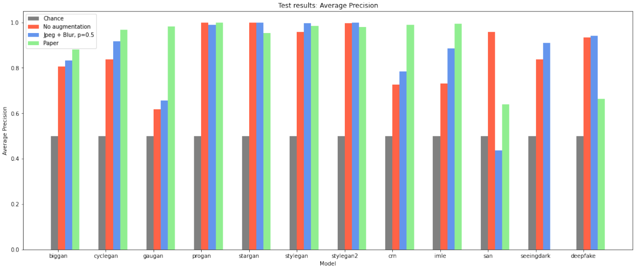
Our results - ResNet18



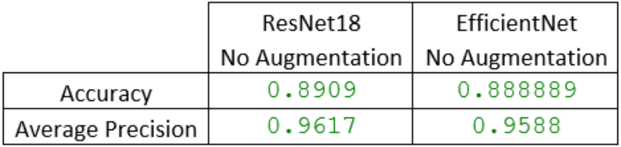


Our results - EfficientNet





A test with a more recent (and sophisticated) GAN: StyleGan3



	ResNet18 No Augmentation	EfficientNet No Augmentation
Accuracy	0.8909	0.888889
Average Precision	0.9617	0.9588

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio.
Generative adversarial networks, 2014.

Kevin P. Murphy.
Probabilistic Machine Learning: An introduction.
MIT Press, 2022.

Mingxing Tan and Quoc V. Le.
Efficientnet: Rethinking model scaling for convolutional neural networks.
CoRR, abs/1905.11946, 2019.

Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros.
Cnn-generated images are surprisingly easy to spot... for now.
CoRR, abs/1912.11035, 2019.

The End