Image Classification: Authentic vs. AI-Gernerated Photos

Anja Stanić *2190471*
Murad Hüseynov *2181584*
Bahareh Najafi *2042940*
Serena Simonetti *1967165*
Piercarlo Risi *1914164*

***Abstract*—This paper investigates the classification of authentic and AI-generated images using subsets of the CIFAKE dataset on two Convolutional Neural Network (CNN) architectures: DenseNet121 and WideResNet. DenseNet121, pretrained on ImageNet, achieved an accuracy of 97% by leveraging its densely connected layers and transfer learning capabilities, even with significant upscaling of images from 32×32 to 224×224 resolution. However, the upscaling introduced blur artifacts, potentially limiting generalizability to non-blurry images. WideResNet, trained directly on the original 32×32 images, avoided these artifacts and demonstrated superior handling of native-resolution data but achieved a lower accuracy of 82.6%.**

I. Introduction

The rapid development of artificial intelligence has enabled the creation of AI-generated images to become more and more sophisticated. Besides having a number of very valuable applications in entertainment, design, and education, these also create significant risks because of the increase of deepfakes and misinformation. Thus, detection of AI-generated content became an urgent problem in machine learning. A variety of methods have been proposed in recent years for performing this task, one of the promising ways being the utilization of CNNs, which are the current standard for image classification tasks due to their capability of learning hierarchical visual features [8].

Given this challenge, most researchers have used CNNs, which have lately given very impressive results in image classification tasks. The CIFAKE dataset is an excellent benchmarking point for the CNN approach on the separation of the two classes into authentic and AI-generated images. However, since its native resolution is 32×32, most pre-trained models, particularly on ImageNet, may expect the usual default size of 224×224.

The two models to be compared for this work are DenseNet121 and WideResNet; both have been proposed to distinguish between real and AI-generated images. DenseNet121 is a densely connected network pre-trained on ImageNet. In this paper, this model is re-targeted for feature extraction for the CIFAKE dataset through transfer learning. However, DenseNet121 expects much bigger input sizes, and the CIFAKE images will be upscaled, which could give rise to blur artifacts and might not be friendly for this model. WideResNet contrasts with this, as this network is specifically designed for Low-resolution datasets such as CIFAKE and thus can have the images processed natively without any need for resolution upscaling.

The main objectives of this study are to:

1) Evaluate the performance of DenseNet121 and WideResNet on a subset of the CIFAKE dataset.
2) Analyze the trade-offs between using a pretrained high-resolution model and a resolution-matched architecture.
3) Provide insights into the implications of upscaling-induced artifacts for AI-generated image detection.

II. Related Work

Classifying authentic and AI-generated images is a very important point of study due to continuous improvement of generative models such as GAN. For most image classification tasks performed by Krizhevsky et al. (2012) [8], CNNs may be applied to learn images in a hierarchical manner. LeCun et al. (2015) [9] indicated that the CNN would show good results in those respective image classification tasks where auto feature extraction is required or involved.

Recently works done such as Wang et al. (2020) [10], Yu et al. (2021) [11], employ CNNs to detect artifacts in generated images, as generative models usually illustrate trivial inconsistencies in the textures, this paper will be utilized in discussing the proposed solutions in detail. However, it can be challenging in some instances when applying these high-resolution pre-trained models such as DenseNet121 to a low-resolution dataset like CIFAKE. Another variant called WideResNet solves these two mentioned disadvantages of DenseNet and other previous models. An upsampled image test conducted by Huang et al. (2017) [1] illustrates good results for DenseNet121. However, upscale presents a chance of introducing blur artifacts. As shown by Zagoruyko and Komodakis (2016) [4], the network width is increased while much stronger performance is achieved without upscaling.

The following study compares DenseNet121 and WideResNet on the CIFAKE dataset, considering some of the trade-offs involving transfer learning with upscaled images and native resolution processing.

III. Proposed Method

Two Convolutional Neural Network (CNN) architectures were used in this work: **DenseNet121** and **WideResNet**. Each of the models was selected with care and optimized based on their respective capabilities to handle the peculiar challenges of the CIFAKE dataset.

A. DenseNet121

First, the DenseNet121 pretrained CNN architecture is considered because it reuses the features continuously via its densely connected convolutional layers [1]. This also reduces redundant computation and guarantees a good gradient flow: for this reason, DenseNet is appropriate for CIFAKE, with its restricted amount of images. The backward process of the feature transmission for DenseNet is made easy by its densely connected structure: every layer provides the input for all previous layers:

$$x_l = H_l([x_0, x_1, \ldots, x_{l-1}]), \tag{1}$$

where x_l represents the output of layer l, H_l is a composite function (batch normalization, ReLU, and convolution), and $[x_0, x_1, \ldots, x_{l-1}]$ is the concatenation of feature maps from all previous layers [2].

To align with the ImageNet dataset on which DenseNet121 was pretrained, all CIFAKE images were resized from their native 32×32 resolution to 224×224 pixels. This pre-processing would be essential, as it gets the input sizes into compatibility with DenseNet121 while still holding the pretrained model's ability to extract generalizable features from larger, high-resolution images [3]. When resizing images from 32×32 to 224×224, each upscaled pixel is calculated as:

$$I_{\text{upscaled}}(u, v) = \sum_{m,n \in \mathcal{N}(u,v)} I_{\text{original}}(m, n) \cdot k(u - m, v - n), \tag{2}$$

where $I_{\text{upscaled}}(u, v)$ is the interpolated pixel intensity, $I_{\text{original}}(m, n)$ are the intensities of nearby pixels, and $k(\cdot)$ is the interpolation kernel (e.g., bicubic or bilinear). Although this upscaling introduced some blur, the model effectively learned consistent patterns in the training data, as both training and validation images were forced to the same preprocessing steps.

B. WideResNet

Given the limitations of upsampling 32×32 images to 224×224, WideResnet architecture was used, because it can natively handle the original 32x32 image resolution. WideResNet is a variation of ResNet that increases the width, or the number of channels, of convolutional layers while reducing the depth [4]. This modification enhances the capacity of feature learning without substantially raising computational complexity.

WideResNet has been shown to achieve state-of-the-art results on CIFAR-10 and similar datasets, which share the same resolution as CIFAKE. Unlike DenseNet121, WideResNet eliminates the need for upscaling, thereby preserving the original image quality and avoiding interpolation artifacts introduced during resizing. The architecture uses residual connections to add inputs directly to the outputs of intermediate layers, enhancing gradient flow:

$$y = F(x, \{W_i\}) + x, \tag{3}$$

where $F(x, \{W_i\})$ represents the residual function parameterized by weights $\{W_i\}$, and x is the input. WideResNet increases the width k (number of filters per layer), enhancing the network's capacity to extract fine-grained features while reducing the depth compared to traditional ResNet architectures.

C. Data Augmentation and Training

To improve data diversity and model robustness, we applied several augmentation techniques, including random horizontal flips, random rotations (up to 15%), and brightness and contrast adjustments (up to 20%) [5]. For both architectures, images were normalized to align with ImageNet statistics using:

$$I_{\text{normalized}} = \frac{I - \mu}{\sigma}, \tag{4}$$

where I is the input image, $\mu = [0.485, 0.456, 0.406]$ is the mean, and $\sigma = [0.229, 0.224, 0.225]$ is the standard deviation.

D. Optimization and Regularization

We optimized the models using the Adam optimizer, with a learning rate initialized at 0.0001 and reduced by half if the validation loss did not improve for three consecutive epochs, implemented via the ReduceLROnPlateau scheduler. Regularization techniques included:

- **L2 Regularization:** Weight decay parameter set at 10^{-5}
- **Dropout:** Applied in DenseNet121 to address overfitting, with rates of 0.3 after the first layer and 0.2 after hidden layers, selected based on earlier experiments. Dropout was not required for WideResNet due to its inherent robustness.

Batch normalization was applied after every fully connected layer in DenseNet121 and after each convolutional layer in WideResNet to stabilize training. Additionally, we implemented early stopping to finish training if validation accuracy did not improve after 10 epochs. Both models were trained for 20 epochs, during which early stopping was not activated for DenseNet121 models, however it was triggered for WideResNet.

E. Evaluation Metrics

We chose loss, accuracy, F1-score, recall, precision, confusion matrix, and the precision-recall curve to provide insights on each model's performance. These metrics provide overall accuracy (accuracy), balance false positives and false negatives (F1-score, recall, precision), and visualize performance trade-offs (precision-recall curve).

While WideResNet would ultimately only reach its highest at a validation/test accuracy of 82.6%/82.4%, which is far lower than the performance achieved with DenseNet121, the latter performed over images with native or required CIFAKE resolution from 32x32, which is a configuration where the blurring impact arising due to upscale wouldn't have a place. The performance compared with DenseNet121 is relatively low probably because it could not exploit the features from pretraining on ImageNet, which were quite relevant for this task, in view of the domain shift from low-resolution CIFAKE images to higher-resolution ImageNet features.

IV. Dataset and Benchmark

"CIFAKE: Real and AI-Generated Synthetic Images" is a dataset containing 60,000 real images and 60,000 AI-generated images, making it suitable for training models to distinguish between the two categories [6]. While the REAL images are sourced from CIFAR-10 dataset [7], the FAKE images were generated using AI. For this project, we used a balanced subset that consists of 30,000 authentic images and 30,000 AI-generated images, totaling 60,000 images to keep equal distribution between two classes, as well as, not to exceed computational limitations and sources, as we used Google Colab free version with its daily usage limits, and it was not feasible to train on the whole dataset.

V. Experimental Results

A. DenseNet121 results

Based on the training of several models over 20 epochs, the table below presents their evaluation using the previously mentioned metrics:

TABLE I
METRICS FOR INITIAL, TRANSITIONAL AND FINAL MODEL

	initial model	transitional model	final model
train loss	0.4298	0.2210	0.0073
validation loss	0.4388	0.2006	0.0981
train accuracy	0.8764	0.9079	0.9976
validation accuracy	0.8688	0.9155	0.9745
validation precision	0.8689	0.9158	0.9745
validation recall	0.8688	0.9155	0.9745
validation f1-score	0.8688	0.9155	0.9745
test precision	0.8685	0.9125	0.9735
test recall	0.8685	0.9122	0.9735
test f1-score	0.8685	0.9122	0.9735

B. WideResNet results

Based on the training of several models, the table below presents their evaluation using the previously mentioned metrics:

TABLE II
METRICS FOR INITIAL, TRANSITIONAL AND FINAL MODEL

	model 1	model 2	model 3
train loss	0.2112	0.0206	0.2603
validation loss	0.6412	0.0430	0.6399
train accuracy	0.9102	0.8587	0.8870
validation accuracy	0.8261	0.8071	0.7974
validation precision	0.8423	0.8450	0.8384
validation recall	0.8261	0.8071	0.7974
validation f1-score	0.8240	0.8017	0.7911
test precision	0.8396	0.8440	0.8359
test recall	0.8241	0.8093	0.7956
test f1-score	0.8221	0.8044	0.7893

Fig. 1. DenseNet121

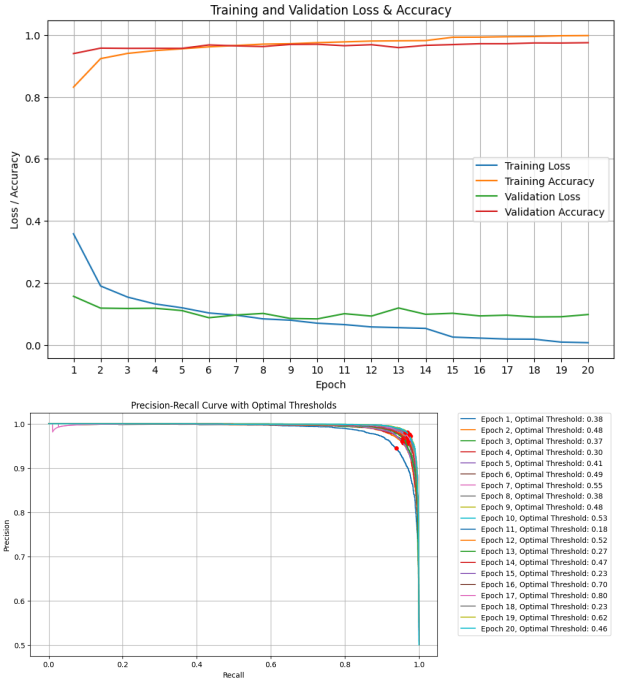
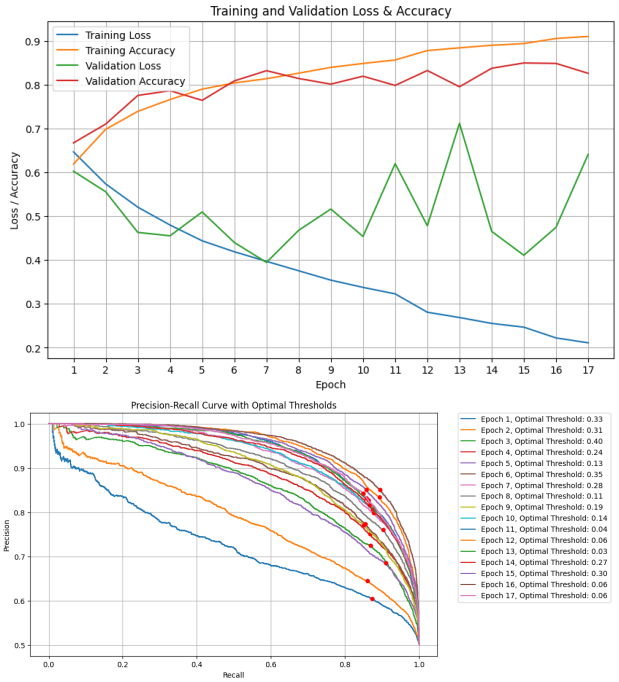


Fig. 2. WideResnet



C. Overall results and model comparison

DenseNet121 achieved 97% accuracy, meaning it generalized very well to the validation set. Still, the model was a bit overfitting: the training loss continued to decrease, while the validation loss had plateaued. By only knowing the DenseNet architecture in application to such huge upscaled CIFAKE images from the resolution of 32x32 to 224x224,

this good accuracy was achieved with DenseNet due to its high propensity for densely connected ConvNets. The low-level information got nicely propagated in the earlier layers to subsequent layers within this model without the model being badly blurry in the upper layers, hence making it able to get adapted against the increased artifacts with consistent preprocess and normalization over the upscaling of images during training.

VI. Conclusion and Future work

This paper explored authentic and AI-generated image classification using DenseNet121 and WideResNet architectures on subsets of the CIFAKE dataset. DenseNet121 reached 97% accuracy by leveraging transfer learning but at the cost of upsampling, which caused blur artifacts. WideResNet did not perform upsampling, operating on the images at their native resolution of 32×32, and achieved a lower accuracy of 82.6%. These results expose the trade-offs of different approaches between pretraining on high-resolution models and low-resolution models, providing insights for fine-tuning detection strategies concerning AI-generated images.

In the future, we would like to further work on:

1) **Advanced Preprocessing:** Investigate image enhancement methods to address blur from upscaling.
2) **Hybrid Architectures:** Explore combining pretrained models with low-resolution architectures for improved performance.
3) **Broader Dataset Usage:** Evaluate on the full CIFAKE dataset and real-world data with diverse resolutions.

References

[1] Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely Connected Convolutional Networks. In CVPR. https://arxiv.org/abs/1608.06993.

[2] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A Large-Scale Hierarchical Image Database. In CVPR. https://ieeexplore.ieee.org/document/5206848.

[3] Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How Transferable Are Features in Deep Neural Networks? In NeurIPS. https://arxiv.org/abs/1411.1792.

[4] Zagoruyko, S., & Komodakis, N. (2016). Wide Residual Networks. arXiv. https://arxiv.org/abs/1605.07146.

[5] Shorten, C., & Khoshgoftaar, T. M. (2019). A Survey on Image Data Augmentation for Deep Learning. Journal of Big Data. https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0197-0.

[6] CIFAKE: Real and AI-Generated Synthetic Images from Kaggle. https://www.kaggle.com/datasets/birdy654/cifake-real-and-ai-generated-synthetic-images.

[7] CIFAR-10 dataset. https://www.cs.toronto.edu/ kriz/cifar.html

[8] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. https://dl.acm.org/doi/pdf/10.1145/3065386

[9] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep Learning. https://www.nature.com/articles/nature14539

[10] Wang, S., Yu, L., Jiang, Z., & others (2020). Detecting AI-Generated Images with Residual Artifacts. https://arxiv.org/abs/2007.12349

[11] Yu, F., Bao, J., Zhou, Y., & others (2021). Multi-Scale Forensic Detection of AI-Generated Images. https://arxiv.org/abs/2104.11227

GitHub Repository:
https://github.com/Anjaas85/CIFAR10_vs_CIFAKE

VII. Group Member Roles

- Murad Huseynov - training the models, report, data analythics
- Anja Stanic - training the models, repository, data analythics
- Bahareh Najafi - data analythics, visualisation, presentation
- Piercarlo Risi - data analythics, initial report
- Serena Simonetti - preprocessing, presentation