Comparative Analysis of CNNs & ViTs for classification of AI generated images

Pushparaj K S
pes1202102119@pesu.pes.edu

Shauryadeepsinh G Raolji
pes1202100807@pesu.pes.edu

K Siddharth Rao
pes1202100577@pesu.pes.edu

K Ganesh Vaidyanathan
pes1202100629@pesu.pes.edu

Dr. Shylaja S S
shylaja.sharath@pes.edu

Monday 6th May, 2024

Abstract

The rise of artificial intelligence has led to the creation of hyper-realistic images, blurring the lines between reality and simulation. This paper investigates the challenge of distinguishing between real and AI-generated images, exploring the effectiveness of Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) in image analysis. Through a comparative study, we address key challenges including classification difficulties, ViTs' viability, and optimal model selection. Experiments conducted on the CIFAKE dataset reveal that ViTs generally outperform CNNs in accuracy and recall for classifying AI-generated images, with varying computational complexities. Our findings contribute to understanding the nuances of navigating AI-generated visual content and offer insights into selecting appropriate models for image classification tasks.

Keywords: *Generative Adversarial Networks, Convolutional Neural Networks, Vision Transformers, Image classification, Synthetic images, CIFAKE dataset, Model architectures*

1 Introduction

In the era of artificial intelligence, the capacity to generate hyper-realistic images indistinguishable from authentic photographs has surged. This phenomenon, fueled by advancements in generative models like GANs and VAEs, challenges our ability to discern between reality and simulation. Consequently, it has sparked a pressing need to develop techniques capable of differentiating between real and AI-generated images across various domains, from verifying the authenticity of media content to ensuring the integrity of digital forensics.

This paper delves into this critical task, exploring the intricate features and perceptual nuances that distinguish real photographs from their AI-generated counterparts. Moreover, it evaluates the effectiveness of Convolutional Neural Networks (CNNs) and Vision Transformers, two prominent architectures in image analysis, in accurately identifying the origin of images. By examining both the visual characteristics that betray the artificial nature of generated images and the performance of state-of-the-art neural networks, this study seeks to provide a comprehensive understanding of the challenges and opportunities inherent in navigating the realm of AI-generated visual content

2 Related Work

After the advent of synthetic image generators, there has been a lot of scrutiny regarding lack of capable means to distinguish AI generated images from real ones. Various works across the globe, have tried to solve this issue. One notable paper is the "Analysis of Artificial Intelligence based Image Classification Techniques" paper, which performs a comparative study among the various traditional Machine learning methodologies, such as Support vector machines, Random Forests, Discriminant

Analysis to name a few. The authors found that the KNN classifier outperforms the others yielding the highest accuracy.

The paper, " Online Detection of AI-Generated Images", discusses the pixel level approach for image classification. This paper also talks about the results produced by simple CNN models such as ResNet. Another interesting paper is "Raising the Bar of AI-generated Image Detection with CLIP", this paper indicates that the CLIP-based approach for detecting AI-generated images is highly effective, even with limited training data, and robust against various image impairments.

In our paper we present the following:

1. Comparison study of CNNs and ViTs for classification of images as AI generated or real images.
2. Achieved the highest CNN based Accuracy on CIFAKE dataset(as far we have researched!)

3 Problem Statement Formulation

The advent of Artificial Intelligence (AI) has revolutionized various domains, including image processing and classification tasks. With recent advancements in AI, particularly in the field of Generative Adversarial Networks (GANs), there has been a surge in the generation of synthetic images that closely mimic real-world images. These AI-generated images present unique challenges for traditional classification algorithms due to their complex and identical nature.

Our project aims to address the following key problem statements:

1. **Challenges in Classification:** The primary challenge lies in effectively classifying AI-generated images using conventional Convolutional Neural Networks (CNNs), which are primarily designed for processing natural images. These images often contain intricate patterns, textures, and structures that may not conform to traditional image classification methodologies.
2. **Viability of Vision Transformers (ViTs):** Vision Transformers (ViTs) have emerged as a promising alternative to CNNs for image classification tasks. However, their performance and suitability for classifying AI-generated images remain largely unexplored. It is essential to investigate the efficacy of ViTs in handling the unique characteristics of AI-generated images and compare their performance against conventional CNNs.
3. **Optimal Model Selection:** Given the diverse nature of AI-generated images and the evolving landscape of AI models, selecting the most suitable model architecture for image classification tasks becomes crucial. By conducting a comparative analysis between CNNs and ViTs, we aim to provide insights into the optimal model selection for classifying AI-generated images based on various performance metrics.

In summary, our project seeks to explore and compare the effectiveness of CNNs and ViTs for the classification of AI-generated images. By addressing the aforementioned challenges and objectives, we aim to contribute to the advancement of AI technologies and their applications in image processing and classification tasks.

4 Experiments and Results

In this section, we describe the experiments conducted to compare the performance of Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) for the classification of AI-generated images. We present the experimental setup, including dataset preparation, model architectures, training procedures, and evaluation metrics. Subsequently, we provide a detailed analysis of the obtained results.

4.1 Experimental Setup

4.1.1 Dataset

The CIFAKE dataset was used for the experiments, containing both real and AI-generated synthetic images. It consists of a diverse range of images suitable for image classification tasks, with a total of [number] images divided into [number] classes.

4.1.2 Model Architectures

For CNNs, we experimented with several popular architectures, including EfficientNetB4, EfficientNetB5, EfficientNetB6, InceptionV3, MobileNetV2, MobileNetV3, ResNet101V2, and VGG16. These architectures are widely used for image classification tasks and were fine-tuned for the specific task of classifying AI-generated images.

For ViTs, we explored a variety of models, including amunchet-rorshark-vit-base, base-patch16-224, base-patch32-384, google-vit-base-patch16-224-in21k, tiny-patch16-224, and dima806. Due to the computational complexity of training ViTs, we limited the training epochs to 2.

4.1.3 Training Procedure

Both CNNs and ViTs were trained using similar procedures. The models were initialized with pre-trained weights from ImageNet and fine-tuned on the CIFAKE dataset. We used the Adamax optimizer with a learning rate of 0.001 for CNNs and 3e-6 for ViTs. The training was conducted for 100 epochs for CNNs and 2 epochs for ViTs, with early stopping criteria based on validation loss.

All our CNNs and ViTs were trained on V100 GPU compute resources available on Paperspace. V100 GPU has 32 GB GPU RAM and took around 30- 45 mins to run one code of our CNN and about 90 mins for the ViT models.

4.1.4 Evaluation Metrics

We evaluated the performance of CNNs and ViTs using standard classification metrics, including accuracy, precision, and recall. Additionally, we analyzed the computational complexity of each model in terms of the number of trainable parameters (in millions).

4.2 Results

Table (1) presents the results of our experiments, showing the accuracy, precision, and recall for each model architecture. We observe that ViTs generally outperform CNNs in terms of accuracy and recall, although some ViT models exhibit higher precision as well. The number of trainable parameters varies across models, with ViTs generally having a larger parameter count compared to CNNs.

Also, as seen in the figure (1), CNNs are able to perform very high classification accuracy in general with almost one-fourth the number of model parameters.

ViTs, in general with less training epochs, do not match the CNNs, but using a completely fine tuned model of ViT on the CIFAKE dataset, gives us a whopping 98.37 classification accuracy!

4.3 Discussion

The experimental results demonstrate that ViTs generally outperform CNNs in the classification of AI-generated images. However, the choice of model architecture also depends on factors such as computational complexity and training time. Further analysis is required to understand the trade-offs between model performance and resource requirements for real-world deployment.

5 Conclusion

In this study, we explored the effectiveness of Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) in distinguishing between authentic and AI-generated images. Our experiments on the CIFAKE dataset showed that ViTs generally outperformed CNNs in accuracy and recall for classifying AI-generated images, despite their higher computational complexity. While some ViT models also exhibited higher precision, considerations such as computational resources and training time are important in selecting the optimal model architecture. These results suggest the potential of ViTs as a promising alternative to CNNs for image classification tasks involving AI-generated content. Further research is needed to understand the trade-offs between model performance and resource requirements in real-world scenarios. Our study contributes to advancing AI technologies in image processing and classification tasks.

All the codes can be found on our github : `https://github.com/kganeshv12/DeepLearningProject`

Table 1: Performance Comparison between CNNs and ViTs

Model	Accuracy	Precision	Recall
EfficientNetB4	97.35	96.08	98.72
EfficientNetB5	97.54	97.41	97.69
EfficientNetB6	96.29	98.89	93.64
InceptionV3	96.31	95.95	96.7
MobileNetV2	93.36	92.54	94.32
MobileNetV3	96.47	97.49	95.4
ResNet101V2	95.15	95.64	94.61
VGG16	96.57	97.24	95.86
amunchet-rorshark-vit-base	66.95	67.9	66.95
base-patch16-224	72.35	72.29	72.35
base-patch32-384	73.65	74.61	73.65
google-vit-base-patch16-224-in21k	55.5	69.2	55.5
tiny-patch16-224	83.66	84.44	83.66
dima806	98.37	98.37	98.37

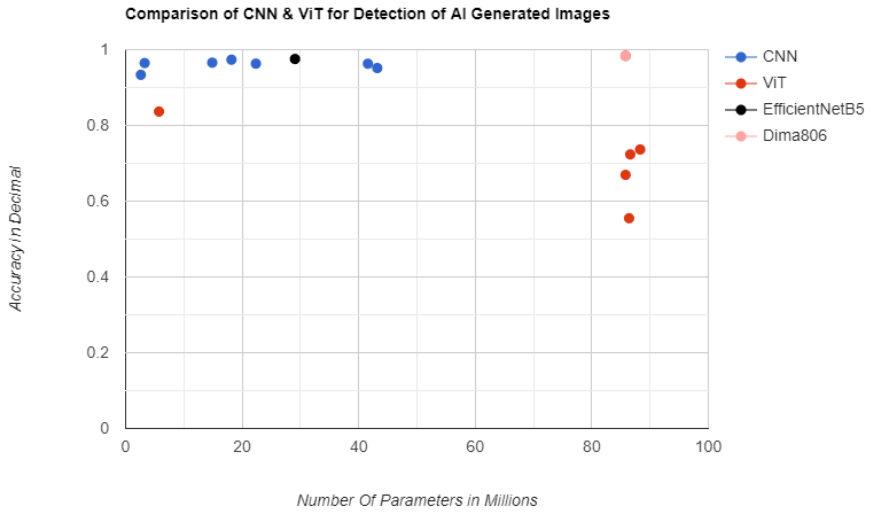

Figure 1: Results:ViTs vs CNNs

6 Bibliography

- **Bird, J. J. and Lotfi, A.** (2023). CIFAKE: Image Classification and Explainable Identification of AI-Generated Synthetic Images. *arXiv:2303.14126* [cs.CV].
- **Tan, M. and Le, Q. V.** (2020). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *arXiv:1905.11946* [cs.LG].
- **Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.** (2014). Going Deeper with Convolutions. *arXiv:1409.4842* [cs.CV].
- **Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.** (2017). MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv:1704.04861* [cs.CV].
- **He, K., Zhang, X., Ren, S., Sun, J.** (2015). Deep Residual Learning for Image Recognition. *arXiv:1512.03385* [cs.CV].

- **Simonyan, K., Zisserman, A.** (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv:1409.1556* [cs.CV].
- **Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.** (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv:2010.11929* [cs.CV].