

Real vs. AI-Generated Images: A Comparative Study of Neural Network Techniques

Group 2 Yu-Chun (Lila) Su, Chih-Hsin (Olivia) Peng, Pei-Hsin (Bonny) Yang

1. Problem Statement & Motivation

In today's digital age, the internet overwhelms people worldwide with an unprecedented volume of information sourced from various online platforms. Given the human tendency to trust initial impressions, it becomes essential to inspect the reliability of the sources we depend on for information. Regrettably, misinformation and misunderstandings pervade the digital landscape, posing a considerable challenge to the integrity of information dissemination. In light of this, our primary objective is to determine the origin of the images we encounter, discerning whether they have been generated by artificial intelligence or not.

2. Dataset

Dataset: [Kaggle](#)

Our dataset, sourced from Kaggle, contains 120,000 facial images. It comprises 70,000 authentic images and 51,000 AI-generated images, representing diverse demographics. The dataset showcases a wide range of facial images, representing various demographics, including different ages, ethnic backgrounds, and genders. This diversity ensures robustness and generalizability in training our deep-learning models, enhancing performance and reliability.

3. Methodologies

CNN

We use a CNN as our first model. It features three convolutional layers with ReLU activation, followed by max-pooling and dropout to prevent overfitting. The layers have increasing filter counts—32, 64, and 128—with each having a 5x5 kernel, a stride of 2, and padding of 1. After the convolutional layers, the network flattens the feature maps into a 1D vector for the fully connected layers. The model has two intermediate fully connected layers with 500 and 225 features, each with a ReLU activation function. The output layer has two logits, representing raw scores for binary classification. We use cross-entropy as our loss function, which is appropriate for binary classification. Cross-entropy loss applies a sigmoid transformation to the logits, which are the raw outputs from the final linear layer, before calculating the loss. This transformation creates a probabilistic representation of the output, allowing the model to learn to maximize the probability of the correct class. We use SGD as the optimizer, updating model parameters based on mini-batches from the training dataset.

ViT

The pre-trained model we discovered on Hugging Face focuses on determining whether an image was generated by AI or not, aligning closely with our topic of interest. ViT, or Vision Transformer, is often considered superior to CNNs for image training due to its ability to learn attention between individual patches. In the ViT workflow, an input image is initially divided into fixed-size pixels and transformed into a 1D vector. This vector is then projected from a low-dimensional to a high-dimensional space using a linear transformation, enhancing feature capture. For our model, The image is converted to 196 patches, with each capturing 768 features. Subsequently, the high-dimensional vector undergoes processing in the transformer encoder. During this encoding phase, the model crucially learns attention relationships between patches in

a multi-head attention mechanism. Also, the class token is added to the transformer encoder. The class token is initialized as a learnable parameter, allowing it to adapt during training and represent the overall context of the image. This class token's features are then transformed into a set of output scores, generating logits, which are used to predict the final class of the input image. The prediction process involves selecting the highest value, such that if the prediction values for classes 0 and 1 are 2.178 and -3.21, respectively, the prediction will favor class 0.

4. Results

CNN

We chose batch sizes of 32, 64, and 128 and learning rates of 0.01, 0.001, and 0.0001 for hyperparameter tuning. The training accuracy was lower with a batch size of 32, likely because smaller batch sizes increase stochasticity in gradient estimation, potentially requiring a more moderate learning rate to maintain stability. This is especially true when combined with a high learning rate; a learning rate of 0.01 might be too high for smaller batch sizes, leading to training instability or exploding gradients. Additionally, by examining the confusion matrix, we found that the model has a higher recall score on real images, indicating that the CNN model is better at correctly identifying real images than AI-generated ones. This might be because our dataset contains slightly more real images than AI-generated ones, causing the model to learn features more commonly found in real images.

ViT

For this model, the hyperparameter we selected for tuning is the learning rate. We set the batch size to 32 and tuned the learning rate to 0.01, 0.001, and 0.0001. The training accuracies we obtained for each combination were 94.33%, 96.11%, and 80.31%, respectively. The best model was achieved when the batch size was set to 32 and the learning rate to 0.001. Using this model on an unseen test dataset also yielded the highest test accuracy of 95.01%.

In contrast, when setting the batch size to 32 and the learning rate to 0.0001, the model achieved the lowest accuracy of 80.31%, with the testing accuracy dropping to 79.86%. This is probably because our model contains 85,800,194 parameters, requiring a learning rate that facilitates efficient updates to optimize the model. A low learning rate might not provide enough momentum to move toward optimal solutions, resulting in a model that doesn't learn effectively from the data. Furthermore, the model also demonstrates that it can predict real images more accurately than AI-generated images.

5. Image Comparison

Based on the results obtained from CNN and ViT, we selected our top-performing model for further evaluation, which is ViT with a hyperparameter combination of a batch size of 32 and a learning rate of 0.001. Upon assessing the detection capability of the model, we observed certain distinguishing characteristics between images predicted as real and those generated by AI. Specifically, AI-generated images tend to exhibit smiles that appear more artificial and less genuine compared to real images. Additionally, AI-generated images often feature fewer wrinkles than their real counterparts.

6. Insights

Based on the outcomes from both CNN and ViT experiments, three key insights emerge:

(1) **Purpose recommendation:** Due to ViT's ability to learn attention across image patches, it excels in capturing global context through self-attention, making it well-suited for tasks requiring an understanding of the entire image at once. In contrast, CNNs are adept at capturing local features, relying on fixed-size inputs and kernel-based feature extraction. Therefore, CNNs are effective for tasks that prioritize localized information, while ViT is better suited for tasks that necessitate a holistic understanding of the image.

(2) **Model recommendation:** CNNs exhibit impressive accuracy despite requiring fewer hyperparameters, making them an excellent choice for obtaining quick initial results with enhanced computational efficiency. However, ViT stands out as a sophisticated model ideal for comprehensive data exploration and in-depth analysis. Leveraging advanced pre-trained models like ViT is recommended for tasks requiring nuanced insights and understanding of complex data. Therefore, if the dataset is relatively simple, CNNs are more suitable to avoid potential overfitting that could occur when using more advanced models like ViT.

(3) **Business recommendation:** Our objective is to leverage trained models capable of discerning the origin of encountered images, distinguishing between those generated by artificial intelligence and real-world sources. By harnessing these models, we not only augment our ability to filter and authenticate information but also gain invaluable insights into the evolving landscape of AI-generated content. As technology advances, maintaining proficiency in discerning between authentic and AI-generated imagery is vital for upholding integrity and cultivating trust in the digital realm.

7. Challenges















During our model training, we faced challenges due to ViT's complexity. It includes high-dimensional embeddings and self-attention mechanisms, which demand substantial memory, limiting our batch sizes on Google Colab. We addressed this by optimizing ViT with a batch size of 32. Additionally, our large dataset led to RAM crashes, resolved by leveraging SCC and Colab PRO. Despite time constraints, we focused on CNN and ViT models but aim to explore advanced models like VGG and Inception in the future for improved image classification accuracy.

8. Conclusion

In conclusion, our study aimed to differentiate between real and AI-generated images to uphold online information integrity. Evaluating CNNs and ViTs, we found CNNs suitable for local feature capture, while ViTs excel in understanding holistic image context. We provided model recommendations based on dataset complexity and task needs. Despite challenges, our study underscores the significance of advanced models in addressing evolving digital image classification and information verification needs.

9. Contribution Table

https://docs.google.com/spreadsheets/d/1coPbsQOv0C2J85SruElp4PF3GA_y2nPSv6a6K6movDA/edit?usp=sharing

1	🕒 Problem Statement & Motivation	 bonniyang, lilasu086, and Olivia-Peng	▼	Done	▼
2	🕒 Dataset	 bonniyang, lilasu086, and Olivia-Peng	▼	Done	▼
3	🕒 Data Loading & Package Import	 bonniyang, lilasu086, and Olivia-Peng	▼	Done	▼
4	🕒 Data Augmentation	 bonniyang, lilasu086, and Olivia-Peng	▼	Done	▼
5	🕒 Define Classes	 bonniyang, lilasu086, and Olivia-Peng	▼	Done	▼
6	🕒 Model Training - CNN	 bonniyang, lilasu086, and Olivia-Peng	▼	Done	▼
7	🕒 Model Training - ViT	 bonniyang, lilasu086, and Olivia-Peng	▼	Done	▼
8	🕒 Images Present	 bonniyang, lilasu086, and Olivia-Peng	▼	Done	▼
9	🕒 Insights	 bonniyang, lilasu086, and Olivia-Peng	▼	Done	▼
10	🕒 Challenges	 bonniyang, lilasu086, and Olivia-Peng	▼	Done	▼
11	🕒 Conclusions	 bonniyang, lilasu086, and Olivia-Peng	▼	Done	▼
12	🕒 Document	 bonniyang, lilasu086, and Olivia-Peng	▼	Done	▼
13	🕒 WandB Reports	 bonniyang, lilasu086, and Olivia-Peng	▼	Done	▼
14	🕒 Slide	 bonniyang, lilasu086, and Olivia-Peng	▼	Done	▼

10. Use of AI

- (1) After we come up with our original ideas, we write them down and use chatGPT to help us refine the wordings and sentences.
- (2) We use ChatGPT to assist us in debugging when we encounter issues in our code.

11. Appendix

CNN Wandb Report: <https://api.wandb.ai/links/bostonuniversity-olivia/1d0szl0l>

ViT Wandb Report: <https://api.wandb.ai/links/bostonuniversity-olivia/tr3whds4>

GitHub Link: <https://github.com/lilasu086/ba865.git>