Problem Statement

From education [1] to art [2] to software development [3], the recent increase in public access to powerful Artificial Intelligence (AI) tools has affected a wide variety of fields. Not all of these changes have been negative, but many have expressed concern about the authenticity of AI-generated content. With this concern comes the question: how does one differentiate AI work from human work? The unfortunate answer is that it can be very difficult for humans to tell the difference [4]. This project uses machine learning to address the issue, focusing specifically on AI-generated images.

Solution

The solution to this problem is an image classification algorithm that can take a given image and accurately identify whether it was created by AI or by a human being. Preliminary research on general image classification suggested that a Convolutional Neural Network (CNN) would make an effective model [5], so that was the algorithm implemented.

Demo

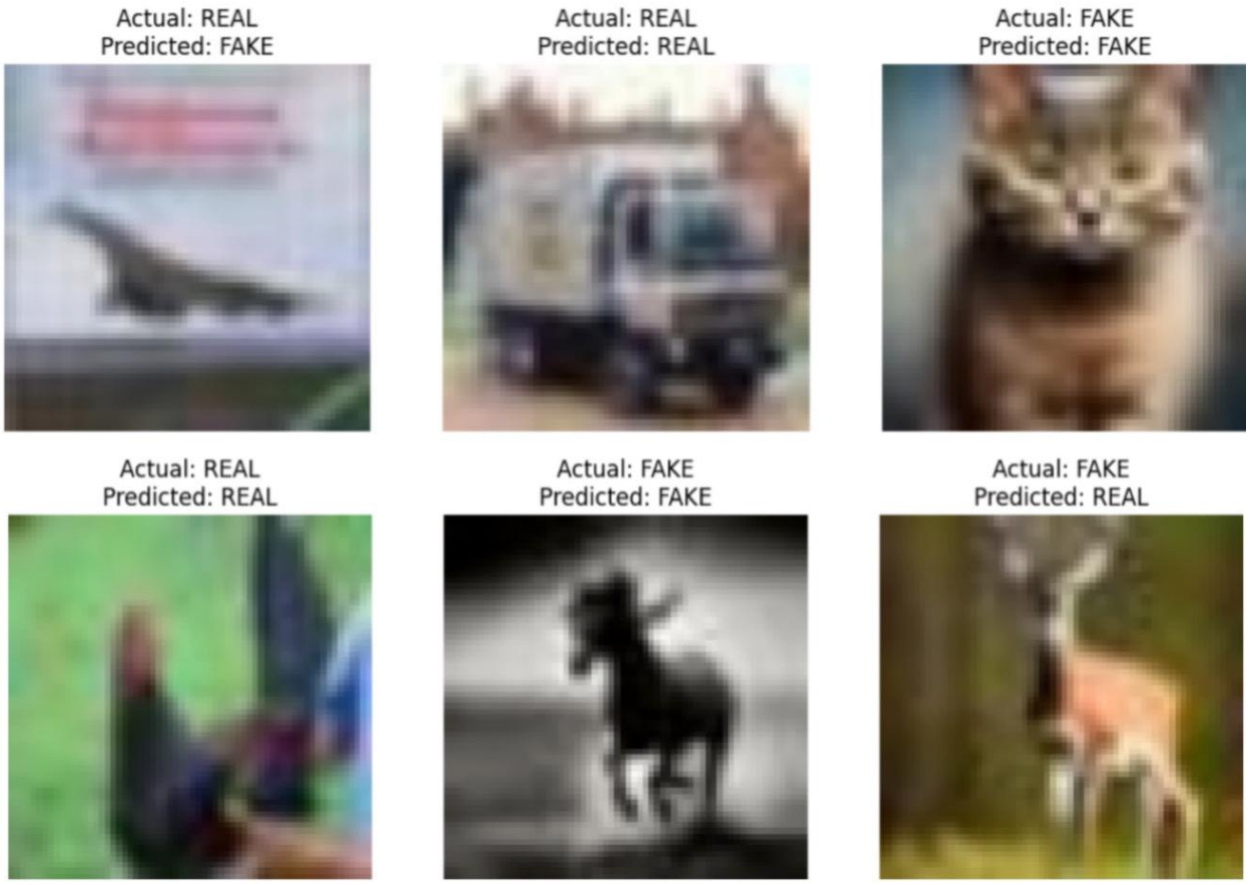

Figure 1. *Model Results/Demonstration*

Figure 1 shows the prediction results of the final trained CNN model on 6 different images (3 real and 3 fake). Images can be used as an input to the model, and are correctly classified most of the

time, although the accuracy is not 100%. In a real-world application, of course, we would not necessarily know the true label of the images– the true labels were only included here for context and model evaluation purposes.

Assumptions, Constraints & Implications

This model assumes that there is a mathematically perceptible difference between AI-generated and human-generated images. Should AI-generated images continue to increase in quality to the point of complete mathematical indistinguishability, this model would no longer be useful or accurate. The model also assumes that images are either entirely AI-generated or entirely human-generated. Some images can be a mix of both, but images such as these were not included in training and it is uncertain how the model would behave under such conditions. Overall, the implications of the model are that AI-generated images can still be differentiated from human-generated ones via machine learning, even if not by the human eye.

How the Solution was Built

The data set that this project will use is called “CIFAKE: Real and AI-Generated Synthetic Images” and is available for download through Kaggle [6]. This data consists of 60,000 AI-generated images and 60,000 human-generated images. Features include the image, the group label (AI-generated or human-generated), and any other potentially useful features that can be derived from the images (number of pixels, colors of pixels, etc). The final solution was a CNN image classifier that can accurately predict whether an image is AI-generated (“Fake”) or human-generated (“Real”). A baseline CNN model was generated, then modifications were made in an attempt to improve the model accuracy.

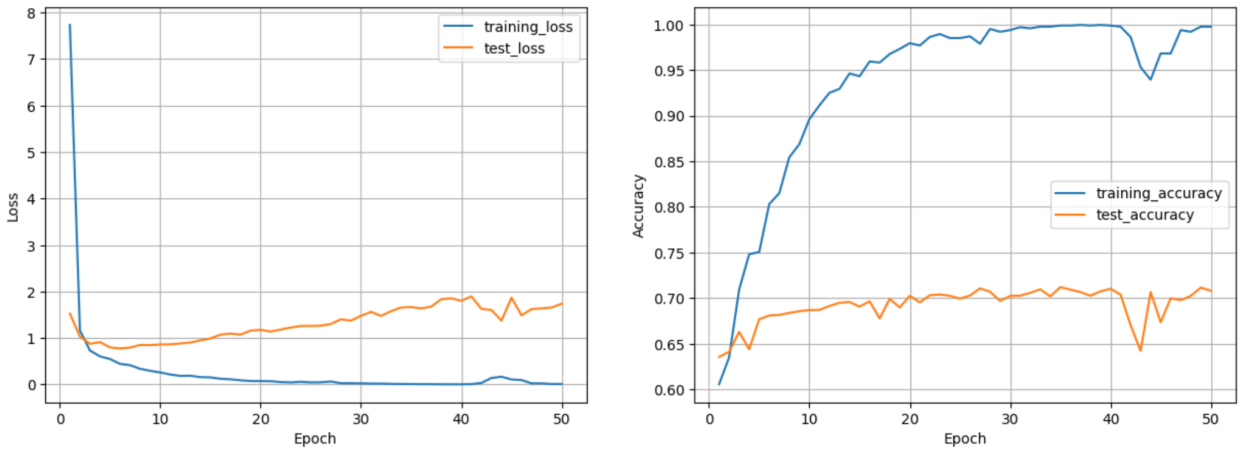

Figure 2. *Base Model from CSCI470 resource*

Figure 2 shows the results of the first model attempted, which was the original baseline model from the CSCI470 resource (which can be found in project/code/references). This model had high training accuracy, but low testing accuracy and relatively large testing loss– an indicator of an overfit and poor model.

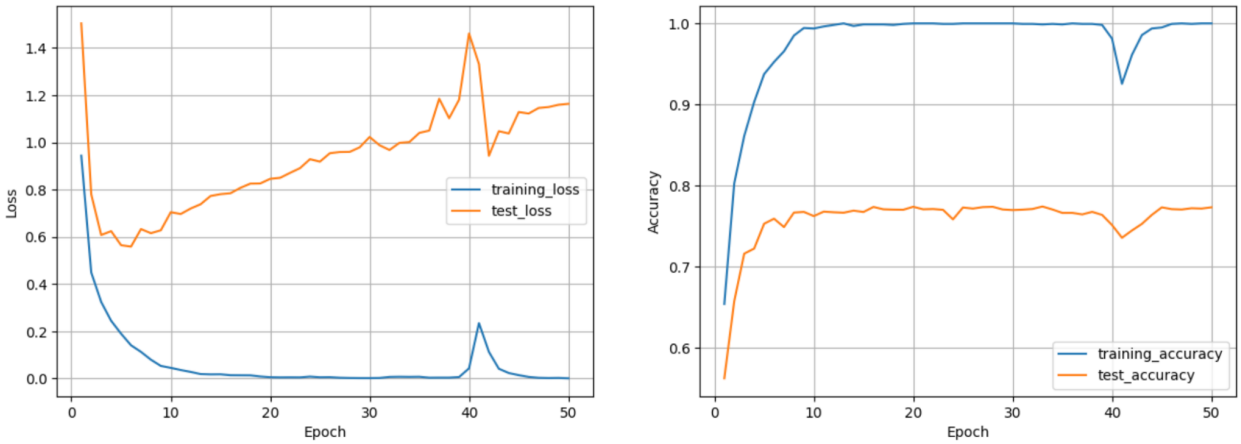

Figure 3. *Modification 1: Adding Batch Normalization layers*

Figure 3 shows the results of adding Batch Normalization layers to the model (the location of these added layers can be seen in the project source code). The graphs demonstrate that the model was more accurate on test data than the baseline model (~75% accuracy instead of ~70%), but there are still signs of overfitting and the testing loss is still quite high.

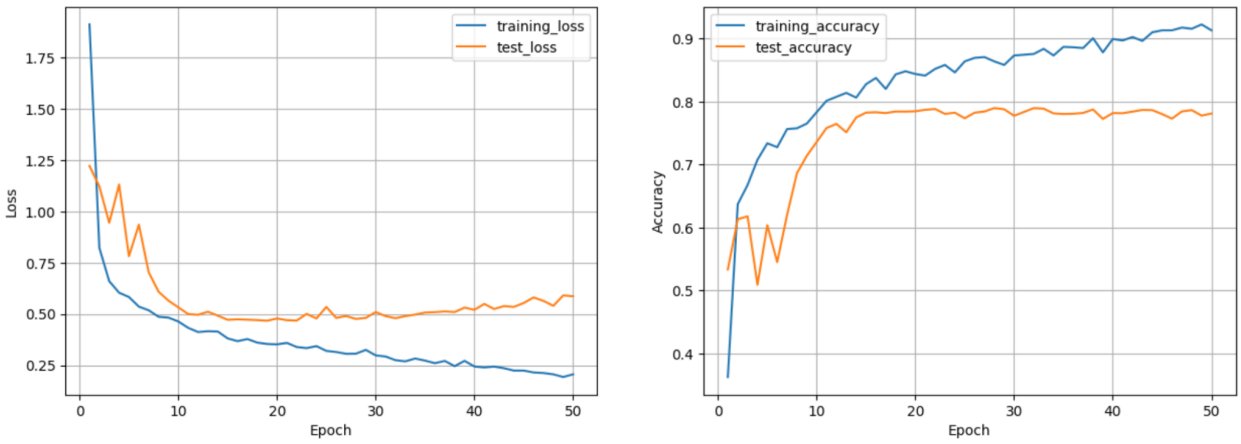

Figure 4. *Modification 2: Adding more convolutional, batch normalization, and max pooling layers*

Figure 4 shows the results of adding more of the same layers from the original model (the exact layer details can again be found in the project source code). This shows a slight improvement over the previous model, with training accuracy reaching almost 80% on average. The testing loss is also much closer to the training loss– an indicator of a less overfit model. While this model is certainly less than perfect, this was the most accurate model achieved within the limited time frame of this project.

Note that many other modifications– including more layers, dropout layers, alternate activation functions, etc.– to the model were attempted, but changes that did not yield a more accurate model were not included in this report.

Summary

This model does a decent job of predicting whether an image is human- or AI-generated, thus proving that machine learning is an effective solution to this problem. Given more time and expertise, this model likely could have been improved significantly, although it was an interesting learning experience nonetheless. Machine learning will likely continue to be applicable to this and other AI-related problems as technology continues to develop.

Links

[1]
https://www.insidehighered.com/opinion/letters/2023/06/05/classroom-implications-when-ai-plagiarizes-and-fabricates-letter

[2]
https://beautifulbizarre.net/2023/03/11/ai-art-ethical-concerns-of-artists/#:~:text=AI%20Art%20takes%20jobs%20from,human%20artists%20work%20hard%20for.

[3]
https://www.guardrails.io/blog/ai-assisted-coding-a-double-edged-sword/

[4]
https://www.tidio.com/blog/ai-test/

[5]
https://medium.com/analytics-vidhya/image-classification-techniques-83fd87011cac#:~:text=Supervised%20classification%20uses%20classification%20algorithms,%2C%20and%20k%2Dnearest%20neighbor.

[6]
https://www.kaggle.com/datasets/birdy654/cifake-real-and-ai-generated-synthetic-images