Real and AI Generated Photo Classification Using FCNNs, CNNs, and Transfer Learning

Ryan Baker, Tobin Eberle, Jeff Wheeler, Tom Wilson

ABSTRACT

The purpose of this report was to investigate the classification of real and artificial intelligence (AI) images. Using the *Shutterstock Dataset for AI vs Human Gen. Image* dataset [1], a fully connected neural network (FCNN) trained from scratch, a convolutional neural network (CNN) built and trained from scratch, and a transfer learning-trained convolutional neural network (TL) were examined for accuracy in distinguishing between AI-generated and human-generated images. This resulted in the best model being the CNN (98.47% test accuracy), but the TL model (90.92% test accuracy) and the FCNN (89.04% test accuracy) also performed well. In conclusion, we found that all three models (FCNN, CNN, and TL) are accessible, easy to implement architectures that could be feasible for discerning between AI-generated and human-generated images.

1. INTRODUCTION

AI-generated images are becoming more and more difficult to distinguish from those produced by humans [2]. This blurring of boundaries poses challenges across various sectors, including media verification, digital forensics, and content authentication. Consequently, developing robust, but also available methods to accurately classify images as AI-generated or human-created has become a point of interest in society [3].

The goal of this paper is to assess the feasibility of using commonplace deep learning architectures to distinguish between AI-generated and human-generated images, such as fully connected neural networks (FCNNs), convolutional neural networks trained from scratch (CNNs), convolutional neural networks trained with transfer learning (TLs).

2. RELATED WORK

2.1 Fully Connected Neural Networks (FCNNs)

Some exploration around using fully connected neural networks to detect AI-images has been done [4], however it is generally concluded that while FCNNs can identify complex patterns [5], they are less adept at handling high-dimensional image data due to their dense connectivity and lack of optimization towards spatial patterns [6]. Nevertheless, FCNNs are considered a fundamental, widely available deep learning technique [7]. Additionally, with constant improvements to computing power, it is becoming less and less impractical to use FCNNs [8]. It is therefore useful to examine an FCNN model as part of this analysis.

2.2 Convolutional Neural Networks (CNNs)

Convolutional Neural Networks (CNNs) have become a cornerstone in deep learning image analysis, primarily because of their ability to capture and focus on spatial relationships within data [9][10][11]. CNNs are often more performant than FCNNs at spatial tasks like image analysis, largely because they have *fewer* parameters than FCNNs (making them less prone to overfitting; i.e. only the relevant spatial data is analyzed instead of all of the patterns as a whole, which often includes noise) [12], but also because the convolutional kernel examines only parts of the image at once, allowing it to recognize features like edges (that can ultimately build up to object detection with successive layers) [13]. Given their widespread adoption and design specifically for image classification tasks, it makes sense to include a CNN model in this analysis.

2.3 Transfer Learning-Trained CNNs (TLs)

While transfer learning is often used to solve problems of data scarcity and improve training time [14], they also have been found to enhance the accuracy and stability of model convergence on occasion (even when scarcity is not a concern, as found in [15]), though TLs have usually been found to be similar or slightly less performant than CNNs trained from scratch [16]. It therefore makes sense to add a TL model to this analysis.

3. METHODS

3.1 Data Set Selection

It was imperative to find a data set comprising of a large amount of total images, but also with even distribution of images between each class (AI-generated and human-generated). The *Shutterstock* dataset was the best choice, though another set known as the *SuSy* data set [17], was almost chosen. This data set had more labels (to identify which generative AI model the images came from, along with human-generated images) and fewer human-generated images, so for a basic model that outputs only two classes (AI- and human-generated), this would have resulted in an unbalanced data set.

The *Shutterstock* dataset [1] also had an interesting feature that solidified the choice: the AI-generated images closely matched a real photograph (see Figure 1). This unique pairing of visually similar real and AI-generated images was hypothesized to promote the learning of more fine-grained, discriminative features, potentially improving the model's ability to distinguish subtle cues that separate synthetic and real photographs.

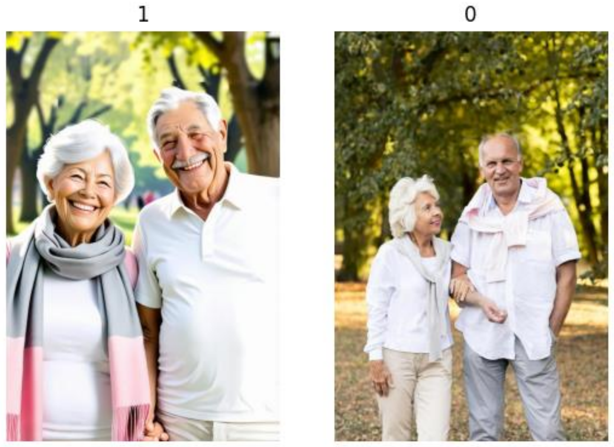

Figure 1: Example pair of an AI-Generated Image (Left) With Its Complementary Real Photograph (Right) From the Shutterstock Dataset.

3.2 Fully Connected Neural Network (FCNN)

The approach taken to create and train the FCNN was to build a basic model with 4 fully connected layers to serve as a starting point to begin training and hyperparameter tuning. With this basic model in place, the initial results were quite low. To improve the initial accuracy, the hyperparameters were adjusted and results were tracked in output files from the completed training runs. The focus was on the number of training epochs, learning rate, and weight decay. Making this hyper parameter adjustment improved test accuracy.

The next step in improving the FCNN model was adding batch normalization to the input layers. Batch normalization can improve gradient flow and prevent vanishing or exploding gradients. This had an immediate impact on test accuracy and results improved further.

To obtain a final FCNN model, we continued to fine tune the hyperparameters and kept the batch normalization. Adding dropout layers had a negative impact on the test accuracy so it was left out of the final model. The last addition to the FCNN model was the inclusion of a learning rate scheduler, which made adjustments to the learning rate between training epochs.

3.3 Transfer Learned Convolutional Neural Network (TL)

Transfer learning is the process of using an existing neural network model to improve the generalization of a model on a new dataset. By transferring the weights and biases from the model that is pre-trained on a larger dataset, we are effectively increasing the size of our target dataset, leading to improved results. The caveat here is that the datasets need to be closely related, and the data of the target dataset needs to be transformed to match the data used in pre-training.

The TL model in this project utilizes the Resnet18 model outlined in "Deep Residual Learning for Image Recognition" provided by Torchvision and Cornell University. Resnet18 is trained on the ImageNet dataset: a large image database containing over 14 million images [18]. The weights are initially loaded into our TL model in a frozen state and the final fully connected layer is adjusted to have the same number of nodes as number of classes (in this case, 2). The TL model is then trained for a few iterations with the weights frozen to tune the fully connected output layer. The weights are then unfrozen so that the TL model can adapt to the new dataset.

Similar to the FCNN, we focus on tuning number of epochs, learning rate, and weight decay. Test results are used to guide the tuning of these parameters until we reach a TL model that has generalized well to our dataset.

3.4 Convolutional Neural Network (CNN)

An additional effort was undertaken to train the Resnet18 model by initializing its weights with random numbers and allowing the training process to optimize more extensively than when Resnet18 was used in the transfer learning model. This allowed the existing Resnet18 network topology to be applied to the domain of artificial image detection, without relying on existing weights trained on a different domain. To get the model to train with maximum flexibility, the layers which were frozen during transfer learning were unfrozen, and gradient computation was enabled for all layers. After increasing the number of epochs during training, this approach yielded very near-optimal test accuracy.

4. RESULTS AND DISCUSSION

4.1 Best Models and Hyperparameters

The best model overall was found to be the CNN model with 98.47% test accuracy and 2791.6 seconds training time (NUM_EPOCHS=6, BATCH_SIZE=24, WEIGHT_DECAY=1e-04, LEARNING_RATE=0.001), followed by the TL model with 90.92% test accuracy and 2558.0 seconds training time (NUM_EPOCHS=3, BATCH_SIZE=24, WEIGHT_DECAY=0.0001, LEARNING_RATE=0.001), and finally the FCNN model with 89.04% test accuracy and 4157.2 seconds training time (NUM_EPOCHS=10, BATCH_SIZE = 24, WEIGHT_DECAY = 1e-05, LEARNING_RATE = 0.0001).

4.2 Confusion Matrices

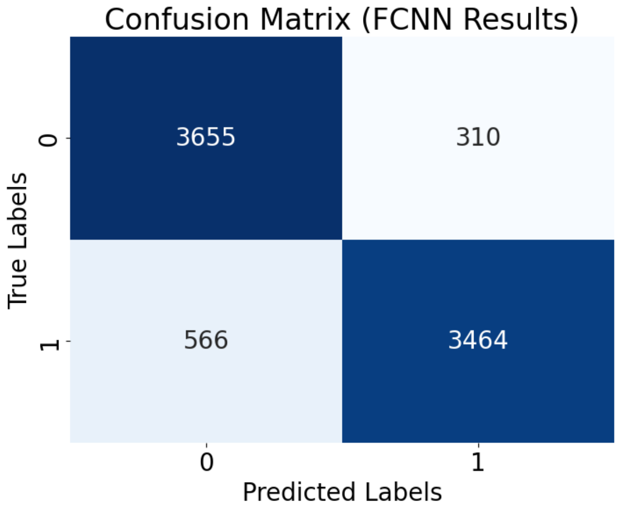

Figure 2: Confusion Matrix for Best Performing FCNN Model

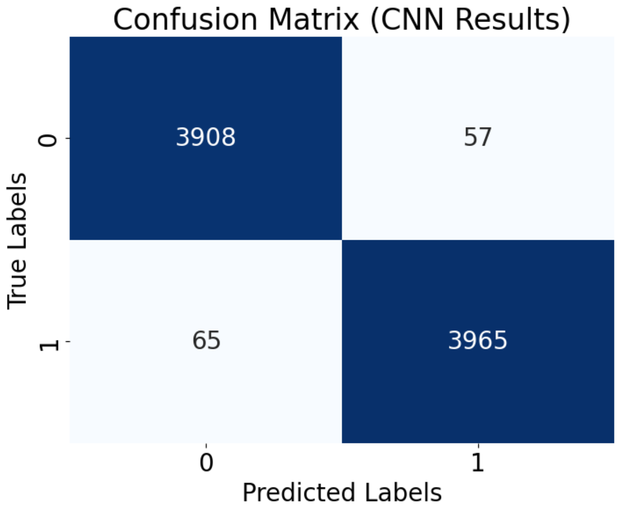

Figure 3: Confusion Matrix for Best Performing CNN Model

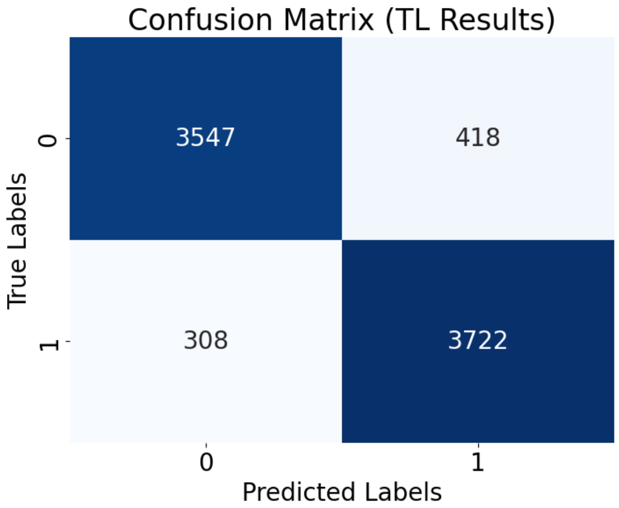

Figure 4: Confusion Matrix for Best Performing TL Model

4.3 Sample Misclassified Images

Below is one example misclassified image from each model, along with their human-generated or AI-generated complement. The actual image that was misclassified has the text ‘Mislabeled Image’ above it, and the complementary image is provided for visual reference. More samples can be found in the jupyter notebook that accompanies this paper.

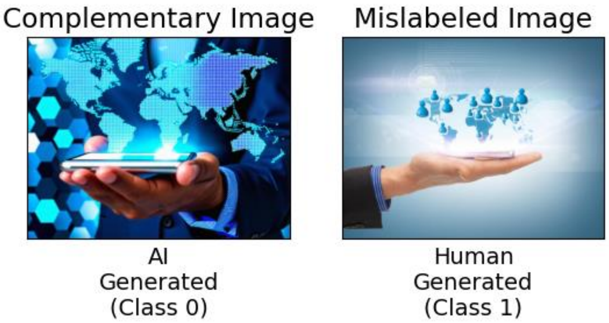

Figure 5: Sample Mislabeled Image from the Best FCNN Model

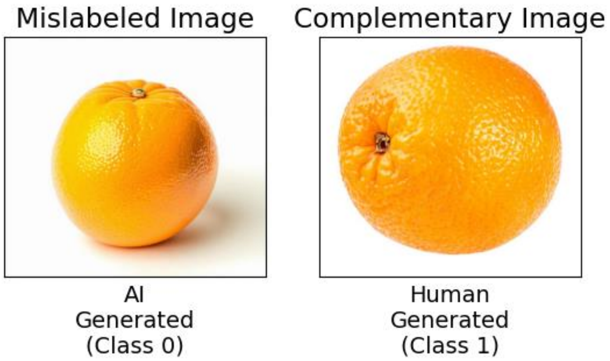

Figure 6: Sample Mislabeled Image from the Best CNN Model

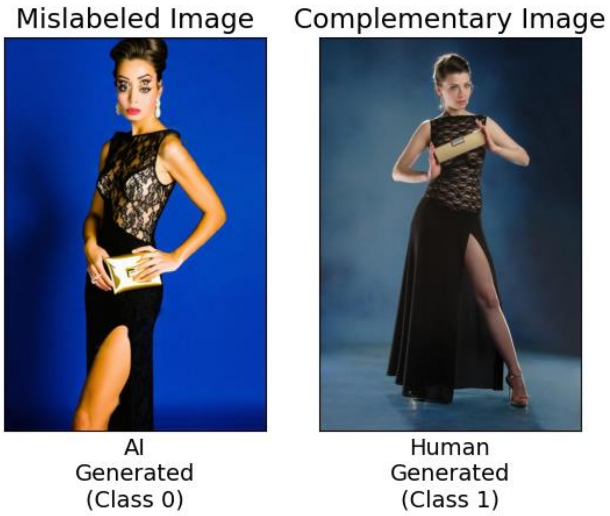

Figure 7: Sample Mislabeled Image from the Best TL Model

4.4 Discussion

Overall, the results obtained from all three models were surprisingly performant (CNN: 98.47%, TL: 90.92%, FCNN: 89.04%). This could most likely be attributed to the very large dataset used for training (0.7 * 79,950 images / 2 classes = 27,982 training images for each class). It was unsurprising that the FCNN model was least performant, however for the purposes of general AI-detection, it could be accurate enough for some applications.

Based on the previous work reviewed, there was an expectation that the TL model would train faster than the CNN model, and this was confirmed, however, only marginally faster (2558.0 seconds vs 2791.6 seconds; a difference of less than 4 minutes). It was also unsurprising to see that the CNN model outperformed the TL model, even though the TL model still performed very well (a difference of 7.55% in test accuracy). The FCNN model was less performant than the CNN and TL models (as expected; a difference of 9.43% and 1.45% respectively), and also took longer to train (expected; a difference of 1365.6 seconds and 1599.2 seconds respectively).

Confusion matrices were not overly skewed; there were relatively even amounts of False Positives (FP) and False Negatives (FN) across all three models (FCNN: [FP: 310, FN: 566], CNN: [FP: 57, FN: 65], [FP: 418, FN: 308]).

Some non-photographic images were found after using the data set extensively, such as the human-generated image in Figure 5igure 5; while this image may technically be ‘human generated’ with the assistance of photo editing software, it may be desirable to remove these types of images to identify between photographs from a camera and synthetic images (regardless if they were AI-generated or photo edited). Other mislabeled images appear like they could be real or AI-generated (even to the human eye, such as Figure 6). This may suggest that generative AI algorithms may be better at making some types of images look very close to their real counterparts. It was also interesting to find images like Figure 7igure 7, where the mislabeled image is clearly very easy to spot as AI-generated upon visual inspection, but the classifier model got it wrong. This likely indicates that the models don’t use the same or similar criteria or patterns that humans use to determine whether an image is likely AI-generated or human-generated.

5. CONCLUSIONS

This study evaluated the performance of three deep learning architectures on the task of classifying AI-generated and human-generated images using the Shutterstock dataset: fully connected neural networks (FCNNs), convolutional neural networks (CNNs), and transfer learning-based CNNs (TLs). While all models achieved relatively high classification accuracies, the CNN trained from scratch achieved the best overall performance at 98.47%, followed by the TL model at 90.92%, and the FCNN at 89.04%.

When factoring the time taken to train each model, there can be clear tradeoffs seen between the CNN and TL models (the TL model was faster to train but was less performant), though most scenarios would likely favor a 7.55% improvement in accuracy over a 4 minute longer time taken to train (therefore CNN trained from scratch may be preferable in most scenarios). The FCNN was also surprisingly performant, with only a 2% weaker accuracy than the TL model (though it took over 60% longer to train).

These results support several established conclusions in the literature. The CNN's superior performance highlights its natural suitability for image-related tasks. Transfer learning, although slightly less performant in this case, remains an effective and practical approach, especially when training time and computational resources are constrained. Even the FCNN, typically not preferred for image data, achieved a surprisingly competitive result, likely bolstered by the large training dataset.

The confusion matrices and misclassified image analysis offered insight into the limitations and blind spots of each architecture and the data set. Some misclassified images were visually ambiguous even to human observers, while others revealed a disconnect between the model's learned features and the visual abilities of most humans. This finding suggests opportunities for future work in interpretability and adversarial robustness.

In summary, all three models proved accessible and reasonably effective for real vs. AI-generated image classification, with CNNs offering the best performance, TL models offering a compelling trade-off between performance and efficiency, and FCNNs serving as a viable baseline. These findings support the feasibility of using common deep learning architectures for digital media verification and similar tasks where distinguishing between real and generated content is of increasing societal importance.

REFERENCES

[1] "ShutterStock Dataset for AI vs Human-Gen. Image." Accessed: Apr. 03, 2025. [Online]. Available: https://www.kaggle.com/datasets/shreyasraghav/shutterstock-dataset-for-ai-vs-human-gen-image

[2] S. S. Baraheem and T. V. Nguyen, "AI vs. AI: Can AI Detect AI-Generated Images?," *J. Imaging*, vol. 9, no. 10, Art. no. 10, Oct. 2023, doi: 10.3390/jimaging9100199.

[3] Y. Wang, Y. Hao, and A. X. Cong, "Harnessing Machine Learning for Discerning AI-Generated Synthetic Images," May 23, 2024, *arXiv*: arXiv:2401.07358. doi: 10.48550/arXiv.2401.07358.

[4] H. Agrawal, R. Parada, and C. Sullivan, "On the Detection of GAN-Generated Facial Imagery".

[5] J. Janke, M. Castelli, and A. Popovič, "Analysis of the proficiency of fully connected neural networks in the process of classifying digital images. Benchmark of different classification algorithms on high-level image features from convolutional layers," *Expert Syst. Appl.*, vol. 135, pp. 12–38, Nov. 2019, doi: 10.1016/j.eswa.2019.05.058.

[6] S. H. S. Basha, S. R. Dubey, V. Pulabaigari, and S. Mukherjee, "Impact of fully connected layers on performance of convolutional neural networks for image classification," *Neurocomputing*, vol. 378, pp. 112–119, Feb. 2020, doi: 10.1016/j.neucom.2019.10.008.

[7] I. H. Sarker, "AI-Based Modeling: Techniques, Applications and Research Issues Towards Automation, Intelligent and Smart Systems," *SN Comput. Sci.*, vol. 3, no. 2, p. 158, Feb. 2022, doi: 10.1007/s42979-022-01043-x.

[8] L. F. S. Scabini and O. M. Bruno, "Structure and Performance of Fully Connected Neural Networks: Emerging Complex Network Properties," Jul. 29, 2021, *arXiv*: arXiv:2107.14062. doi: 10.48550/arXiv.2107.14062.

[9] N. Sharma, V. Jain, and A. Mishra, "An Analysis Of Convolutional Neural Networks For Image Classification," *Procedia Comput. Sci.*, vol. 132, pp. 377–384, Jan. 2018, doi: 10.1016/j.procs.2018.05.198.

[10] L. Chen, S. Li, Q. Bai, J. Yang, S. Jiang, and Y. Miao, "Review of Image Classification Algorithms Based on Convolutional Neural Networks," *Remote Sens.*, vol. 13, no. 22, Art. no. 22, Jan. 2021, doi: 10.3390/rs13224712.

[11] J. Gu *et al.*, "Recent advances in convolutional neural networks," *Pattern Recognit.*, vol. 77, pp. 354–377, May 2018, doi: 10.1016/j.patcog.2017.10.013.

[12] M. M. Bejani and M. Ghatee, "A systematic review on overfitting control in shallow and deep neural networks," *Artif. Intell. Rev.*, vol. 54, no. 8, pp. 6391–6438, Dec. 2021, doi: 10.1007/s10462-021-09975-1.

[13] N. Sharma, V. Jain, and A. Mishra, "An Analysis Of Convolutional Neural Networks For Image Classification," *Procedia Comput. Sci.*, vol. 132, pp. 377–384, Jan. 2018, doi: 10.1016/j.procs.2018.05.198.

[14] M. Gholizade, H. Soltanizadeh, M. Rahmanimanesh, and S. S. Sana, "A review of recent advances and strategies in transfer learning," *Int. J. Syst. Assur. Eng. Manag.*, vol. 16, no. 3, pp. 1123–1162, Mar. 2025, doi: 10.1007/s13198-024-02684-2.

[15] Y. Yu, E. Hu, and Q. Bi, "Efficient prediction of machine tool position-dependent dynamics based on transfer learning and adaptive sampling," *CIRP J. Manuf. Sci. Technol.*, vol. 58, pp. 62–79, Jun. 2025, doi: 10.1016/j.cirpj.2025.01.009.

[16] T. T. Cai, D. Kim, and H. Pu, "Transfer Learning for Functional Mean Estimation: Phase Transition and Adaptive Algorithms," Mar. 27, 2024, *arXiv*: arXiv:2401.12331. doi: 10.48550/arXiv.2401.12331.

[17] "Papers with Code - SuSy Dataset Dataset." Accessed: Apr. 04, 2025. [Online]. Available: https://paperswithcode.com/dataset/susy-dataset

[18] "Deep Residual Learning for Image Recognition" Accessed: Apr 04, 2025. [Online]. Available: https://arxiv.org/abs/1512.03385