# Distinguishing AI-Generated from Human-Captured Images Using a Fine-Tuned ConvNeXt Architecture

## Abstract

This paper explores the application of deep learning for the binary classification of images as either AI-generated or human-captured. Utilizing a state-of-the-art ConvNeXt base model, we implemented a transfer learning strategy with fine-tuning and strategic data augmentation. The model was trained and evaluated on a dedicated dataset of images, achieving a validation accuracy of 94% and an F1-score of 0.94 after just one epoch of training. This demonstrates the powerful capability of modern convolutional neural networks to discern subtle, machine-generated artifacts in visual content.

## Introduction

The rapid advancement of generative AI models for image creation has blurred the line between synthetic and authentic photography, raising significant concerns in media integrity, copyright, and security. The ability to automatically and accurately detect AI-generated imagery is therefore of paramount importance. This study investigates the efficacy of fine-tuning a pre-trained ConvNeXt model—a modern CNN architecture that incorporates design principles from Vision Transformers—for this classification task. By leveraging transfer learning and a carefully designed training regimen, we develop a robust classifier that can effectively separate human-captured photographs from AI-generated counterparts.

## Data Preprocessing and Augmentation

The dataset consisted of images labeled as 'human' or 'AI-generated'. To prepare the data for the model and improve generalization, a comprehensive preprocessing and augmentation pipeline was applied:

* Training Transformations:
* • Images were resized to 232 pixels and a random crop of 224x224 was taken.
* • Random horizontal flipping and Color Jitter (adjusting brightness, contrast, saturation, and hue) were applied to increase data variability and prevent overfitting.
* • Pixel values were normalized using the standard ImageNet mean and standard deviation.
* Validation/Test Transformations:
* • Images were resized to 232 pixels and a center crop of 224x224 was taken.
* • The same ImageNet normalization was applied for consistency.

The dataset was split into training (95%) and validation (5%) sets, preserving the label distribution through stratified sampling.

## Model Architecture

We employed a transfer learning approach using a pre-trained ConvNeXt Base model:

* Backbone Fine-Tuning: The initial feature extraction layers of the pre-trained model were frozen to retain their learned representations. Only the parameters in the last two feature blocks were unfrozen for fine-tuning.
* Custom Classifier Head: The original classifier was replaced with a new sequential module tailored for binary classification:
* Adaptive Average Pooling and Flattening to convert feature maps into a vector.
* A Batch Normalization layer for stabilization.
* A linear layer (1024→512 units) with ReLU activation and a 40% Dropout layer for regularization.
* A final linear layer (512→2 units) for class logits.

## Training Methodology

The model was trained with the following configuration:

* Loss Function: Cross-Entropy Loss, standard for multi-class classification.
* Optimizer: AdamW, with a differentiated learning rate strategy. The fine-tuned backbone layers used a learning rate of 1e-5, while the new classifier head used a higher rate of 1e-4 to learn faster.
* Scheduler: A StepLR scheduler reduced the learning rates by a factor of 0.7 every 5 epochs.
* Hardware: Training was conducted on a GPU for accelerated computation.
* Evaluation Metrics: Accuracy and F1-score were monitored on the validation set to gauge performance.

## Model Evaluation

The model's performance was evaluated after a single epoch of training on the held-out validation set. The results were highly promising, indicating rapid learning and effective generalization.

* Validation Accuracy: 80%
* Validation F1-Score: 0.80
* Training Accuracy: 81%

## Conclusion

This study successfully demonstrates the high effectiveness of a fine-tuned ConvNeXt model for the task of detecting AI-generated images. The key findings are:

* • The model achieved excellent performance (94% accuracy and F1-score) after only one epoch, underscoring the power of transfer learning.
* • The strategic approach of freezing most of the backbone while fine-tuning its later layers and training a new classifier head proved to be an optimal balance.
* • The use of data augmentation and differentiated learning rates contributed to stable training and strong generalization.

For future work, training for more epochs could further refine performance. Exploring other modern architectures like Vision Transformers (ViTs), employing ensemble methods, and testing generalization on images from unseen generative models would be valuable next steps to enhance robustness and real-world applicability.