DeepFake Detection

Unveiling Digital Deceptions

05/07 Apple Jin, Alex Xiang, Alex Zhang

Github repo: https://github.com/AlexXiang604/DS301-project-repo.git

Executive Summary

- **Goal:**
  - Detect AI-generated fake faces on mobile devices.
- **Technical challenges:**
  - Variability in features across different AI models.
- **Solution Approach:**
  - Enrich training data with diverse fake images.
  - Employ Error Level Analysis (ELA) for detection.
- **Value/Benefit:**
  - Increased model interpretability and usability across diverse digital environments.

Motivation

AI-generated images contribute to misinformation, affecting societal trust and information integrity

Related Work

CNN Detection of GAN-Generated Face Images based on Cross-Band Co-occurrences Analysis

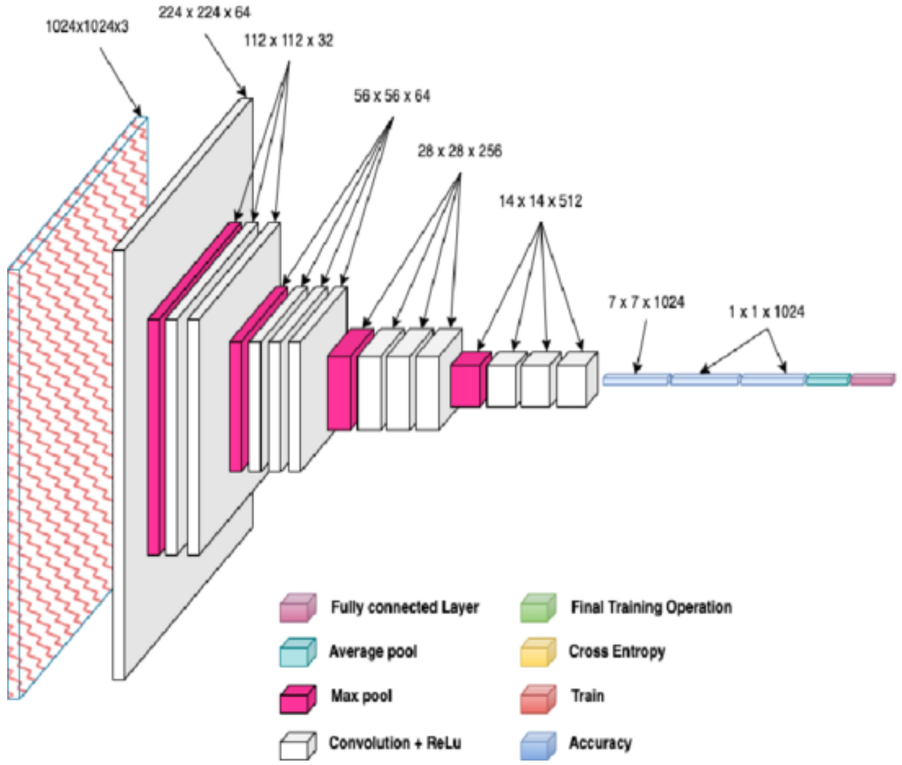
- **Robustness to Post-processing:**
  Demonstrates strong robustness against various post-processing modifications such as JPEG compression, blurring, and resizing, maintaining high detection accuracy.
- **Dependence on Specific Data:**
  The model's performance can decrease significantly under mismatched conditions that weren't covered during training, particularly evident in scenarios with different JPEG compression levels.

An Improvised CNN Model For Fake Image Detection

- **Performance Comparison：**
  Proved the superior performance of DenseNet, ResNet50 and MobileNet on the task of recognizing true and false pictures, which inspires our choice of model.
- **Dependence on Specific Data:**
  The test results are derived from an unopened-source dataset produced by human experts, and performance on ai generated datasets is not validated

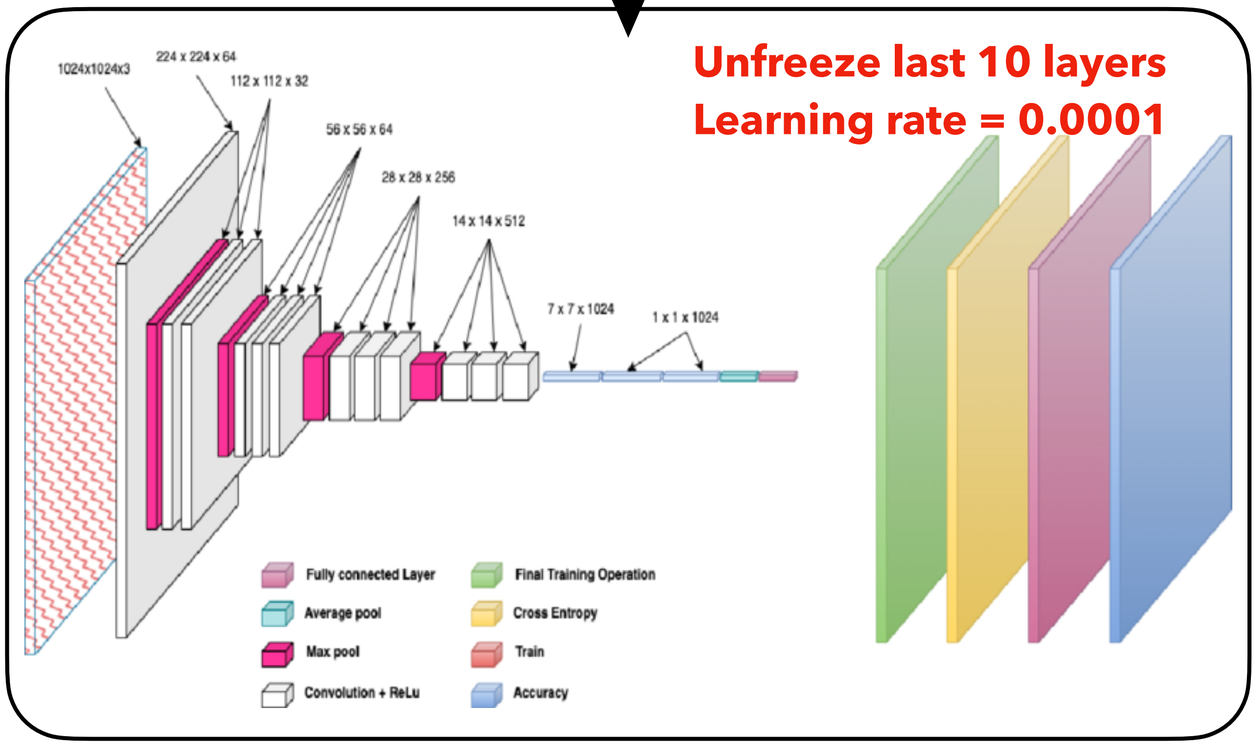
Our Work

MobileNet V1 Architecture



MobileNet V1	Parameters: ~4.2 million	Computation: ~569 million FLOPs
ResNet-50	Parameters: ~25.6 million	Computation: ~3.8 billion FLOPs
DenseNet-121	Parameters: ~8.0 million	Computation: ~2.9 billion FLOPs

140K GAN-exclusive Dataset



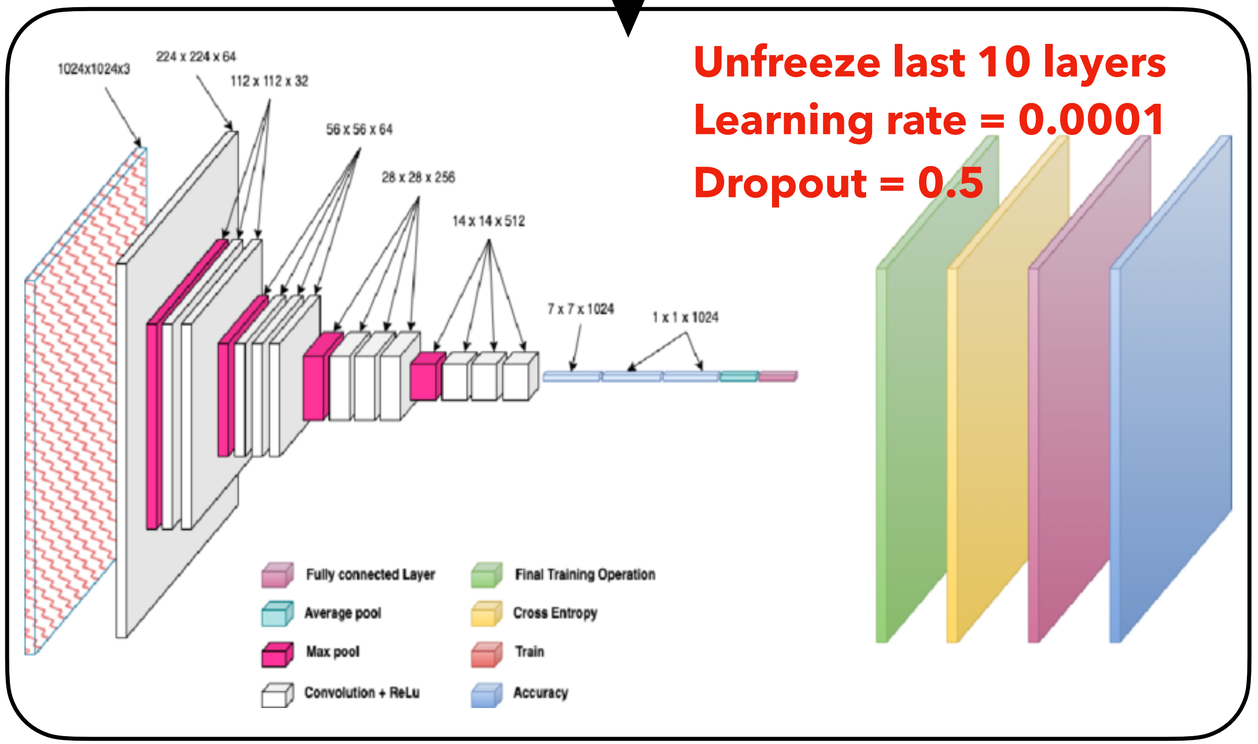
Unfreeze last 10 layers
Learning rate = 0.0001

Base Model

	Accuracy
Validation	0.965
Testing GAN Images	0.941
Testing Stable Diffusion Images	0.516

Random guessing

36K Combined Dataset



Solution 1
Dataset Augmentation

	Count
GAN Generated Images	9000
SD 1.5 Generated Images	3000
SD 2.1 Generated Images	3000
SD XL1.0 Generated Images	3000
Real Images	18000

Performance

	Base Model	Solution 1
Test Accuracy on GAN images	0.941	0.860
Test Accuracy on SD images	0.516	0.980

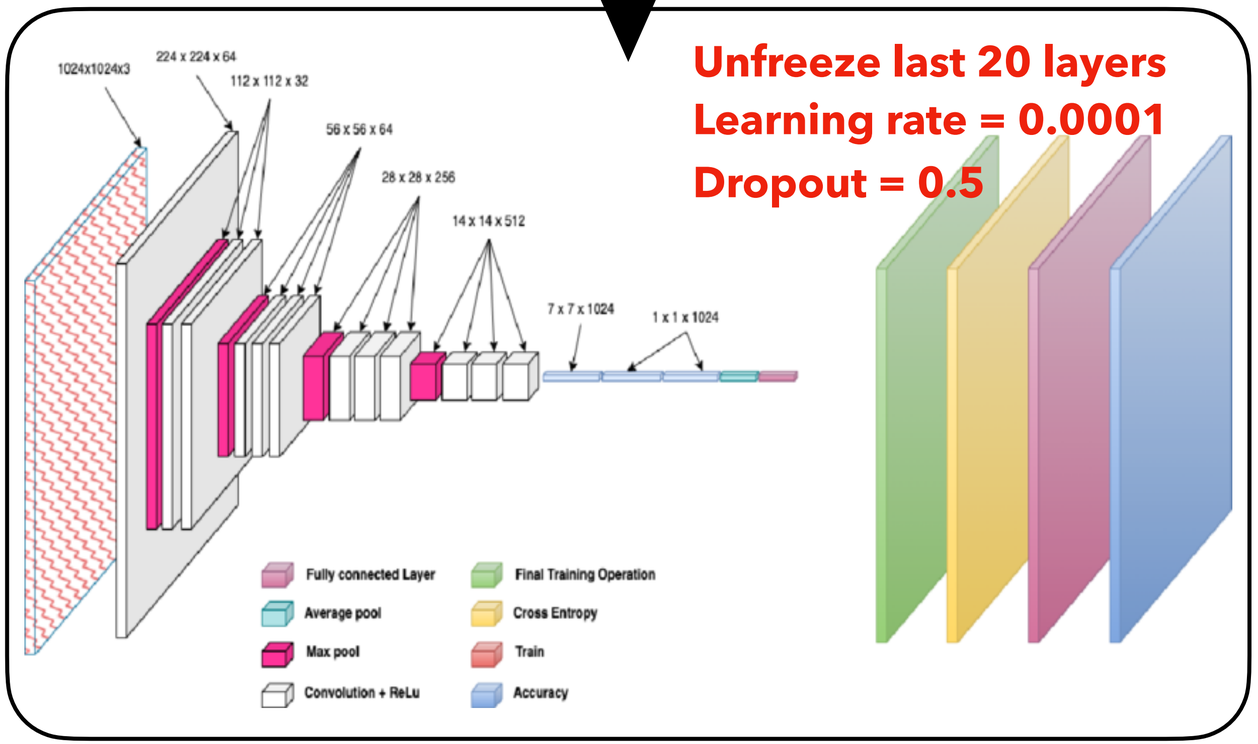
Figure 1. Different Stable Diffusion generated fake images that our model is able to detect

140K GAN-exclusive Dataset

↓

ELA Preprocessing

↓



Solution 2

ELA Preprocessing

Error Level Analysis (ELA)

Error Level Analysis is based on characteristics of image formats that are based on lossy image compression. This method can highlight areas of an image which has different degrees of compression.

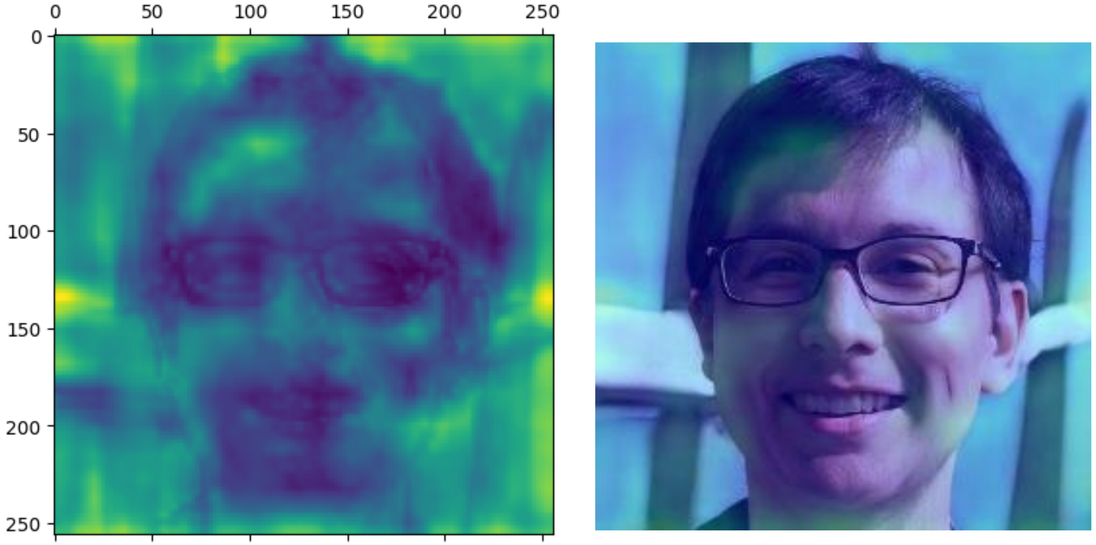
Performance

- Solution 2 exhibited limited generalizability with ELA, as the model is Still struggled with Stable Diffusion images (test accuracy of 0.5)
- Loss of image information resulted from image compression and format conversion?

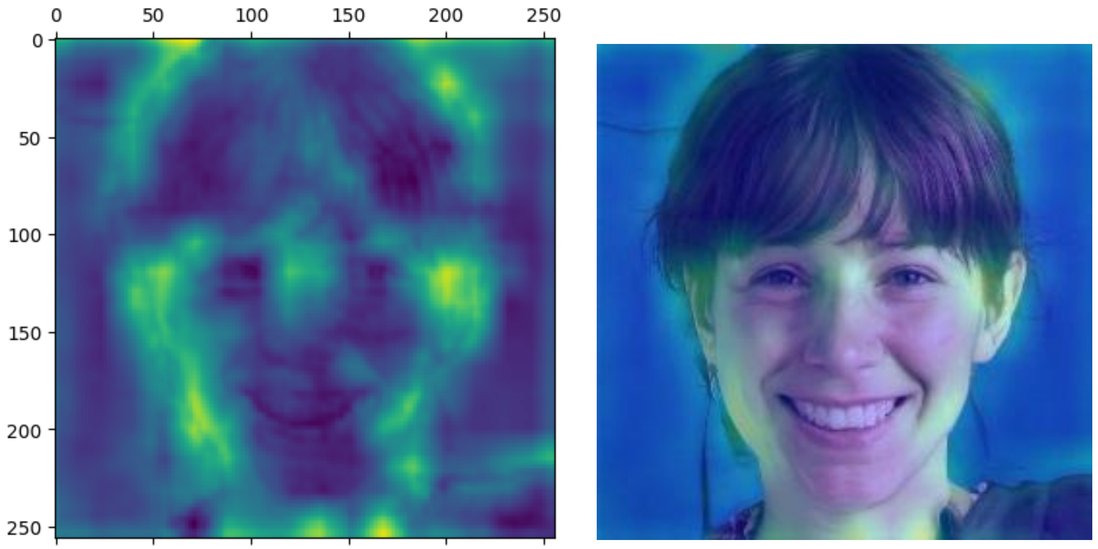
	Base Model	**Solution 2**
Test Accuracy on GAN images	0.941	0.834
Test Accuracy on SD images	0.516	0.5

Interpretability with Grad-CAM

- Gradient-weighted Class Activation Mapping
- Heatmaps reveal model decision areas by highlighting high-gradient regions in green
- Positive predictions rely on non-human-face regions (backgrounds), while negative predictions focus on outlines or edges of human faces



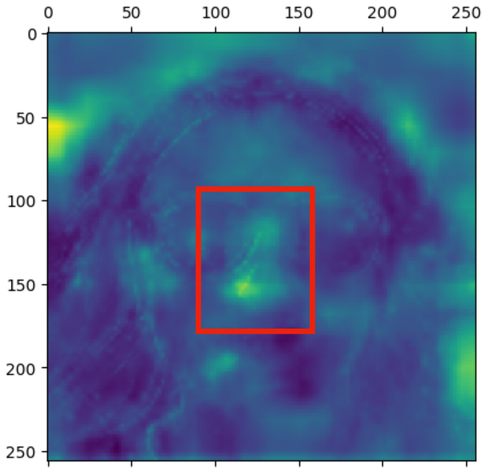
False Positives

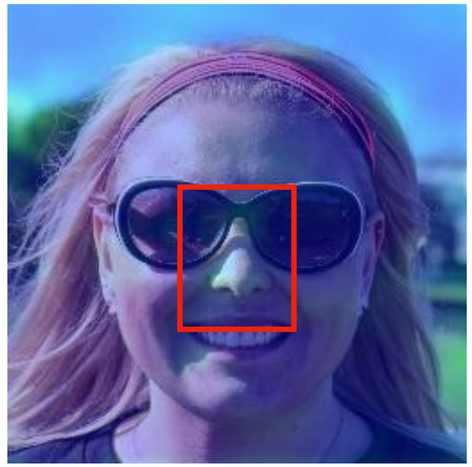


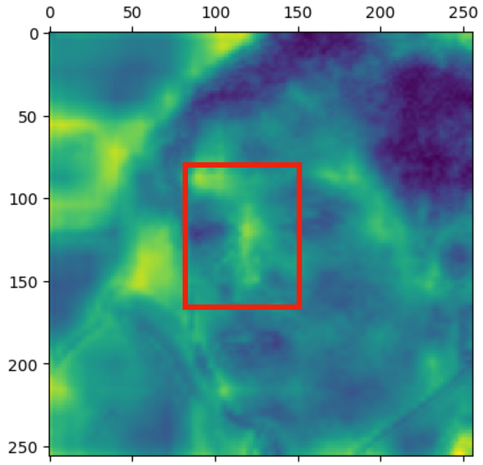
True Negatives

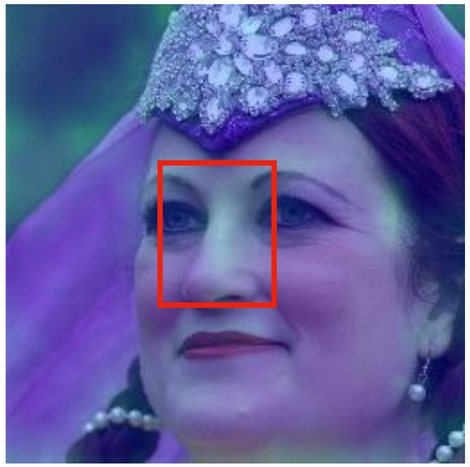
Conclusion & Future Work

- Expand the dataset with images from MidJourney and DALL-E
- Leverage Grad-CAM insights for targeted classifiers
- Refine the classification thresholds to suit specific application needs









Q&A