[CS5242 - Group 7] Evaluating the effectiveness of image classification models in detecting AI generated images

Group 7:
David Aaron Chan (A0274917H
Lee Young Ho (A0113602Y)
Tan Longbin (A0155353H)
Yap Wei Xuan (A0183226H)

Objective

Given the increasing prevalence and quality of generative AI, it is becoming harder to distinguish between real and AI-generated images. This poses issues in the fields of fraud detection / evidence verification.

Project goal is to evaluate the **effectiveness of different image classification models in detecting AI generated images**. Further analysis is done to assess the generalizability of the different models and cost.

Methodology

Overview of methodology

We consider 3 image classification models: Residual Neural Network (ResNet) , Densely Connected Convolutional Neural Network (DenseNet) and Visual Geometry Group (VGG) as well as the 3 image generation models: Big Generative Adversarial Network (BigGAN), Stable Diffusion Version 4(SD) and MidJourney (MJ).

We use images from the GenImage dataset, created by Huawei's Noah's Ark Lab (Zhu et al., 2023), which comprises of more than 1000 image classes, giving a good representation of potential real-world scenarios.

Evaluation task design

To evaluate the performance of ResNet, DenseNet and VGG16 with the chosen parameters, each model is tested in 2 tasks.

Task A: Differentiating Real vs Fake/AI-Generated images (Each model fine-tuned with each generator dataset and is tested against all the generators' datasets) (33000 Training Dataset images / 10 Epochs / 12000 Test Dataset images).

Task B: Predicting the correct Generator for each image (Each model fine-tuned with the combined AI images from all the generators and tested against another batch of AI images) (49500 Training Dataset images / 10 Epochs/ 18000 Test Dataset images). Fig 1 below illustrates these tasks.

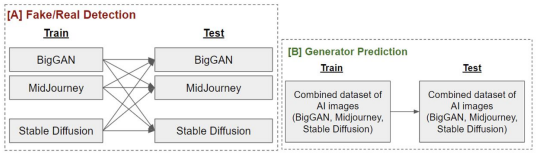

Fig 1: Evaluation Tasks A & B

Optimisation

We further optimise the learning rate hyper parameter by testing multiple variants of ResNet and DenseNet against learning rates of 0.01, 0.001 and 0.0001.

2-model approach

Beyond Tasks A and B, we also explore a 2-model approach which first classified an image under the most probable generator and subsequently use a more specific model to do fake image detection.

Results

Task A: Fake image detection results

Table 1: Accuracy comparisons for each model across the 3 training data subsets and tested against each test dataset.

Train Data	BG			MJ			SD		
Test Data	BG	MJ	SD	BG	MJ	SD	BG	MJ	SD
ResNet50	99.5%	**50.1%**	49.8%	48.9%	**94.8%**	**73.2%**	49.2%	**66.1%**	**95.8%**
DenseNet121	**99.7%**	**50.1%**	49.8%	**50.2%**	93.5%	72.7%	**51.3%**	63.7%	95.0%
VGG16	99.5%	49.9%	**49.9%**	46.9%	92.2%	70.0%	47.4%	59.8%	93.2%

Table 2: Average accuracy comparison with same generator performance and generalizable performance across all 3-image classification models

Train Data	Average accuracy (%)	
Test Data	Same generator performance[2] (eg., train BG, detect BG)	Generalisable performance[3] (e.g., train BG, detect MJ)
ResNet	**96.7%**	56.2%
DenseNet	96.0%	**56.3%**
VGG	95.9%	54.2%

Per Table 1, we found that ResNet outperformed both DenseNet and VGG in same model train and test performance. This is likely due to the larger amount of parameters as compared to DenseNet. Surprisingly, VGG performed worse despite having more parameters, suggesting issues stemming from exploding gradients. Also notable is that MJ and SD runs have similar accuracies, likely a result of their similar latent diffusion based generation methods.

Per Table 2, we found that all models generalised equally poorly.

Task B: Fake image detection results

Table 3: Accuracy comparison between ResNet and DenseNet when trained and tested from data across all the 3 image classification models

Model	Combined test dataset comprising of images from BigGAN, Mid Journey and Stable Diffusion (Accuracy %)
ResNet50	**96.5%**
DenseNet121	96.0%

We omit VGG given its poor performance in Task A. We found that ResNet performed marginally better. This is aligned with results from Task A where ResNet was better at identifying images from MJ and SD. Given two-thirds of Task B data set comprises of MJ and SD, it is natural that performance is similar.

Optimisation

To test the effect learning rate optimisation, an initial learning rate of .001 was used. We that a learning rate of 0.01 was ideal, given that it produced greater lower average loss as compared to alternatives, while maintaining similar train / test times.

2-model approach

Table 4: Accuracy of 2-model approach with ResNet50

Model	Accuracy (%)
ResNet50	**94.4%**

We attempted the 2-model approach with ResNet50, utilising a combined validation set from Task A, producing an accuracy of 94.4%, justifying this as a valid approach in producing high fake image detection accuracy.

Conclusion

Our project demonstrated that the relatively lighter ResNet model outperforms models such as VGG. We also demonstrate the utility of a novel 2-model approach; it allows for more efficient detection of images from new generators, by only requiring generator specific models to be trained, while still being able to manage mixed generator test sets.

References

Idowu, Jamiu, and Ahmed Almasoud. Uncertainty in AI: Evaluating Deep Neural Networks on Out-of-Distribution Images. arXiv:2309.01850, arXiv, 4 Sept. 2023. arXiv.org, https://doi.org/10.48550/arXiv.2309.01850.
Wang, Yuyang, et al. Harnessing Machine Learning for Discerning AI-Generated Synthetic Images. arXiv:2401.07358, arXiv, 14 Jan. 2024. arXiv.org, https://doi.org/10.48550/arXiv.2401.07358.
Zhu, Mingjian, et al. GenImage: A Million-Scale Benchmark for Detecting AI-Generated Image. arXiv:2306.08571, arXiv, 24 June 2023. arXiv.org, https://doi.org/10.48550/arXiv.2306.08571.