Towards Generalization with Deepfake Detection Models

Tyler Vergho

Dartmouth, GOVT 20.12
tyler.k.vergho.23@dartmouth.edu • tvergho.me

Deepfakes: An Introduction

Deepfakes, highly realistic AI-generated images and videos, are increasingly used to spread misinformation and deceive individuals. They pose a threat to the integrity of digital media, necessitating robust and efficient detection methods. However, current detection methods struggle with the rapidly evolving image generation techniques. This research explores the performance of state-of-the-art deepfake detection models and their capacity to generalize to deepfakes produced by unseen generation techniques.

Problem 1: Adversaries Adapt

Deepfake detectors struggle when adversarial manipulators change the models they use for image generation.

When a deepfake detector has been optimized to spot images generated by a specific model, it can become less effective when adversaries switch their image generation methods. For instance, if a detector is trained primarily on images generated by a specific Generative Adversarial Network (GAN), it may struggle to identify deepfakes produced by a different model.

Problem 2: Diffusion Models

Deepfake detectors, often trained on GAN-generated images, fail to keep pace with more realistic outputs from modern diffusion models.

Recent advances in AI technology have seen the emergence of diffusion models like Stable Diffusion, a new generation method rendering anyone with a consumer laptop capable of producing deepfakes that are even more realistic and harder to detect. Unlike GANs, which have been the primary focus of training data for most deepfake detectors, diffusion models leverage a different process to create synthetic images.

Main Findings

There's no "silver bullet" solution for general deepfake detection. However, modern convolutional architectures, ensembles of multiple models, and reconstruction techniques hold promise.

I trained different backbone architectures, ResNet-50 and EfficientNet v2 M (with and without Vision Transformer), using a dataset of images generated by ProGAN.* I found that **EfficientNet** generally matched or outperformed ResNet-50. I tested these **5 models** on **13 datasets**.

*The DIRE-SD model was trained on Stable Diffusion images, to test the hypothesis that training on diffusion outputs better fosters generalization.

	Datasets					
Detectors	ProGAN	StyleGAN	BigGAN	CycleGAN	StarGAN	GauGAN
Wang2020	**100.0/100.0**	73.4/**98.5**	59.0/88.2	80.8/**96.8**	81.0/95.4	79.3/98.1
EFNetV2M	**100.0/100.0**	73.5/98.3	61.9/**88.7**	**83.1**/95.2	76.8/93.9	87.3/98.3
EFNetV2M+ViT	**100.0/100.0**	71.4/96.7	64.5/87.8	82.7/94.5	71.7/92.8	**91.0/98.8**
DIRE	99.8/99.99	**83.9**/93.7	**74.6**/84.6	72.1/80.1	**97.5/99.8**	67.6/73.8
DIRE-SD	56.7/76.5	58.8/78.7	73.6/88.4	59.5/87.0	81.2/94.7	50.4/69.4

	Datasets						
Detectors	CRN	IMLE	Seeing Dark	SAN	DeepFake	StyleGAN2	WhichFaceIsReal
Wang2020	**87.6/99.0**	94.1/99.5	78.3/**92.7**	50.0/63.9	51.1/66.3	68.4/98.0	63.9/88.8
EFNetV2M	79.7/97.8	91.6/99.4	**78.6**/91.3	50.9/64.0	59.0/85.6	70.1/**98.5**	**69.6/91.6**
EFNetV2M+ViT	74.6/98.4	**95.2/99.9**	66.9/83.8	52.1/64.9	51.4/73.5	70.8/97.7	57.4/83.3
DIRE	68.9/87.5	68.6/87.7	60.8/65.0	54.5/60.9	**71.5/91.5**	**83.8**/98.2	65.8/68.4
DIRE-SD	50.3/48.6	49.7/54.1	53.3/77.6	**70.5/87.2**	56.3/59.9	57.0/75.8	53.9/70.5

Table 1: Models' performance on different test sets. Each cell includes accuracy and average precision, separated by slashes.

However, both models struggled with generalizing to **unseen deepfake architectures**, even though they achieved near-perfect accuracy on the ProGAN testset—the data they were specifically trained on. The choice of backbone didn't markedly impact the model's ability to generalize to different architectures. I further attempted to improve generalization using the DIRE method, which involves training a classifier to detect **diffusion reconstruction error**.

	Detectors				
Datasets	DIRE-SD	EFNetV2M	Wang2020	EFNetV2M+ViT	DIRE
Stable Diffusion v1.4	**0.886/0.968**	0.504/0.590	0.5/0.534	0.527/0.613	0.502/0.491

Table 2: Models' performance on Stable Diffusion v1.4 images. With the exception of DIRE-SD, models were only trained on ProGAN.

Ensemble Learning

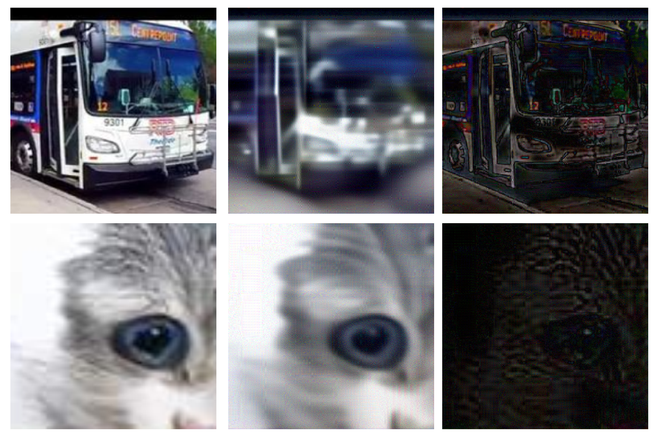
I then compared an EfficientNet v2 S model trained on 5 datasets to an ensemble of "expert" models. The single model displayed broader generalization but didn't reach full accuracy on any dataset. One reason may be the smaller model and dataset, resulting in less accuracy than the ProGAN baseline. Additionally, there may be diminishing returns to increasing dataset diversity. As expected, the ensemble struggled with unseen data but generally performed better on data seen during training. The experiment involved **2 models** tested on **11 datasets**, five seen during training.

	Datasets					
Detectors	**ProGAN**	**BigGAN**	CycleGAN	StarGAN	StyleGAN	**StyleGAN2**
Ensemble	**0.990/0.999**	**0.594/0.793**	**0.701/0.833**	0.530/0.528	0.724/0.815	**0.773**/0.872
EFNetV2S	0.921/0.988	0.570/0.787	0.631/0.810	**0.874/0.956**	**0.728/0.919**	0.673/**0.888**

	Datasets				
Detectors	DaLLe2	Glide	**LDM**	**SD**	Taming
Ensemble	0.544/0.773	0.608/0.852	**0.724/0.913**	0.913/0.983	**0.668/0.855**
EFNetV2S	**0.578/0.839**	**0.635/0.865**	0.659/0.888	**0.926/0.996**	0.583/0.822

Table 3: Comparison on different test sets. Accuracy and average precision are separated by slashes. Bolded models were seen during training.

DIRE



DIRE, or Diffusion Reconstruction Error, compares an original image to its reconstruction using a pre-trained diffusion model. As illustrated above, in theory a real image undergoes significant change during reconstruction, resulting in a high DIRE value (above), while a diffusion-generated image can be reconstructed with greater fidelity, yielding a lower DIRE value (below).

Contrary to the DIRE paper's original claims, my DIRE implementation fell short in universal deepfake detection, despite outperforming other models on StarGAN and showing a 10% accuracy improvement on StyleGAN and StyleGAN2.

Future Study

One promising area that warrants further research is experimenting further with ensemble techniques versus the impact of diverse training datasets in generalizable deepfake detection. Despite not reaching the claimed results, DIRE still holds potential, indicating that more closely replicating the original pipeline could improve deepfake detection capabilities. Future work should also explore other reconstruction error techniques like GAN inversion and VAEs, and experiment with few-shot tuning to simulate real-world scenarios.