Deepfake Detection for Faces

Petre Eftimie, Radu Buzaș, Teodora Diaconescu and Elisabeta Oneață*

University of Bucharest, Romania

*Bitdefender, Romania

petre-laurentiu.eftimie@s.unibuc.ro, radu-gabriel.buzas@s.unibuc.ro, teodora-cosmina.diaconescu@s.unibuc.ro, eoneata@bitdefender.com

1. Introduction

1. We want to evaluate the generalization capabilities of deepfake detection methods.
2. For the first task, we will train on images coming from one generator and test on images coming from other generators.
3. For the second task, we will train on all images and try to tell if the image is real or specify which generator produced it if it's not.

2. Dataset and Preprocessing

- **Dataset Description:**

  The dataset contains real images from the CelebAHQ dataset and locally manipulated images produced by four generators: LDM, Pluralistic, LAMA, Repaint. Each class has 9000 images for training, 900 for validation and 900 for testing.

- **Data preprocessing:** Preprocessing done only when using CLIP embeddings as described in their paper.

3. Models

- We trained each model for 10 epochs, using AdamW optimizer with 0.01 learning rate and 0.01 weight decay.
- Applied linear classifier (Fully-Connected output layer) over each model backbone.

1. **Linear classifier over CLIP:** Trained a fully-connected layer over CLIP embeddings.
2. **ResNet18 backbone trained from scratch**
3. **ResNet18 backbone pretrained with ImageNet**: We kept the backbone weights frozen for the first 5 epochs (fine-tuning) and then unfroze them for the last 5 epochs, but with a lower learning rate (10^{-6}).

6. Conclusion

- The linear classifier over CLIP performs the best on both tasks, proving the effectiveness of pretrained embeddings.
- Training using the ResNet18 backbone from scratch is not only slow but also does not provide significantly better results than the pretrained version.
- Faces from the Repaint dataset seem to be the hardest to classify correctly in both tasks.

4. Cross-generator deepfake detection

Test on	**Train on**			
	LAMA	LDM	Pluralistic	Repaint
LAMA	0.999	0.583	0.752	0.506
LDM	0.574	0.999	0.937	0.653
Pluralistic	0.765	0.967	0.993	0.644
Repaint	0.504	0.885	0.869	0.744

Table 1: AP for linear classifier over CLIP

Test on	**Train on**			
	LAMA	LDM	Pluralistic	Repaint
LAMA	1	0.503	0.579	0.509
LDM	0.458	0.591	0.580	0.515
Pluralistic	0.682	0.517	0.995	0.534
Repaint	0.524	0.493	0.530	0.498

Table 2: AP for ResNet18 backbone trained from scratch

Test on	**Train on**			
	LAMA	LDM	Pluralistic	Repaint
LAMA	0.999	0.412	0.631	0.494
LDM	0.343	0.966	0.550	0.621
Pluralistic	0.873	0.600	0.943	0.543
Repaint	0.533	0.810	0.632	0.775

Table 3: AP for ResNet18 backbone pretrained with ImageNet

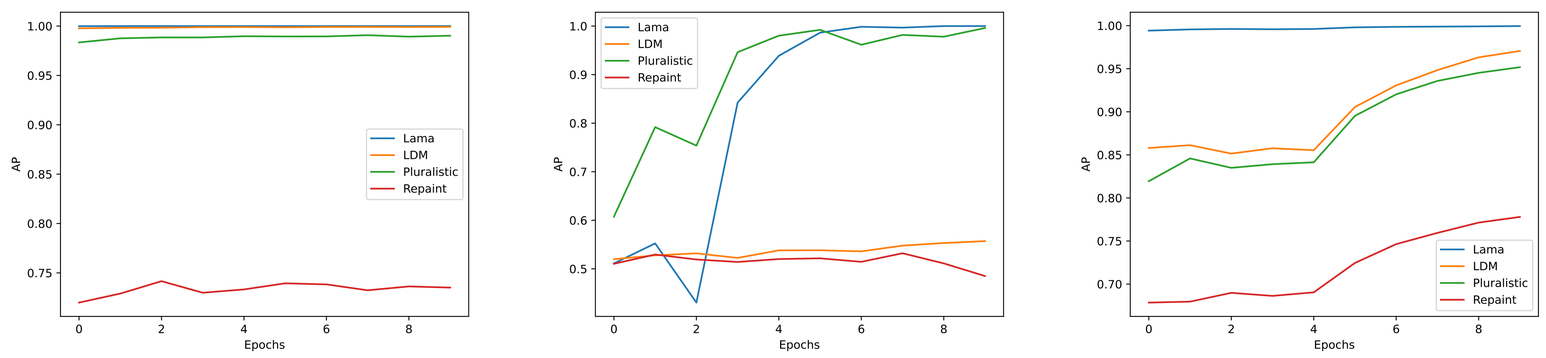

Figure 1: Validation AP plots for CLIP, ResNet18 Scratch and Pretrained

5. Model attribution

Model	Overall Accuracy
Linear classifier over CLIP	0.804
ResNet18 backbone trained from scratch	0.726
ResNet18 backbone pretrained with ImageNet	0.744

Table 4: Overall Accuracy for model attribution

Model	Real	LAMA	LDM	Pluralistic	Repaint
CLIP	0.746	0.995	0.971	0.881	0.43
ResNet18 Scratch	0.237	0.964	0.961	0.994	0.474
ResNet18 Pretrained	0.592	0.994	0.835	0.844	0.455

Table 5: Per class Accuracy for model attribution

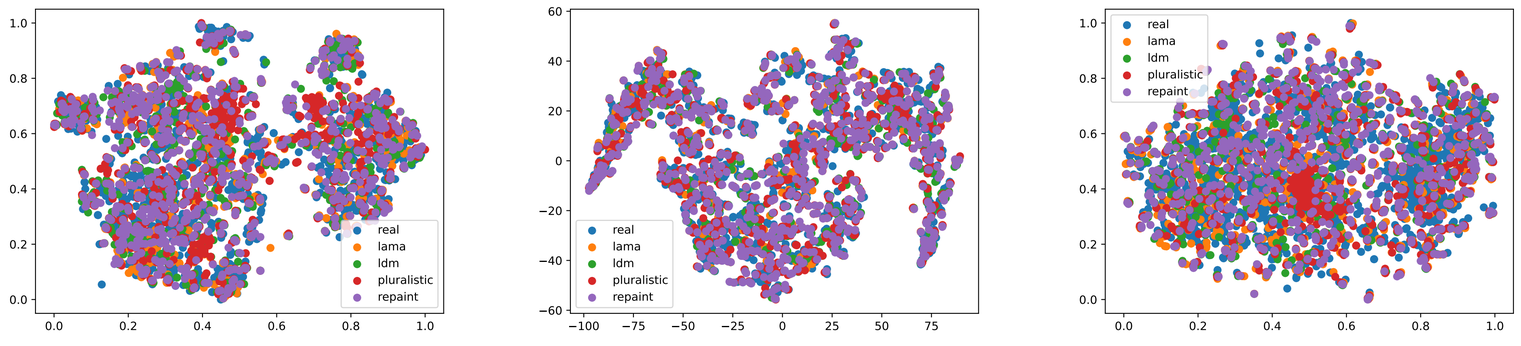

Figure 2: TSNE plots for CLIP, ResNet18 Scratch and Pretrained