Universal fake detection

Florete Fabian-Andrei, Nica Maria-Catalina
University of Bucharest, Romania

fabianflorete@yahoo.com, mariacatalina27.nica@gmail.com

1. Introduction

The main objective of our project is expanding upon a model that can accurately detect artificially generated images using either GAN or Stable diffusion methods.

2. Dataset and Preprocessing

Dataset Description:
These are the datasets used for training, validation and testing. Each have different generation methods: stargan, cyclegan, progan are generated using GAN algorithms, while dalle, deepfake, glide and ldm are generated using stable-diffusion algorithms.

Dataset total samples	
TRAIN	72.000 images
VALIDATION	1.000 images
TEST	total images
CYCLE-GAN	2642 images
DEEPFAKE	5405 images
PROGRAN	8000 images
STARGAN	3998 images
DALLE	2000 images
GLIDE_100_10	2000 images
LDM_200_CFG	2000 images

Data preprocessing: Preprocessing these images is done by feeding the image through the encoding models.

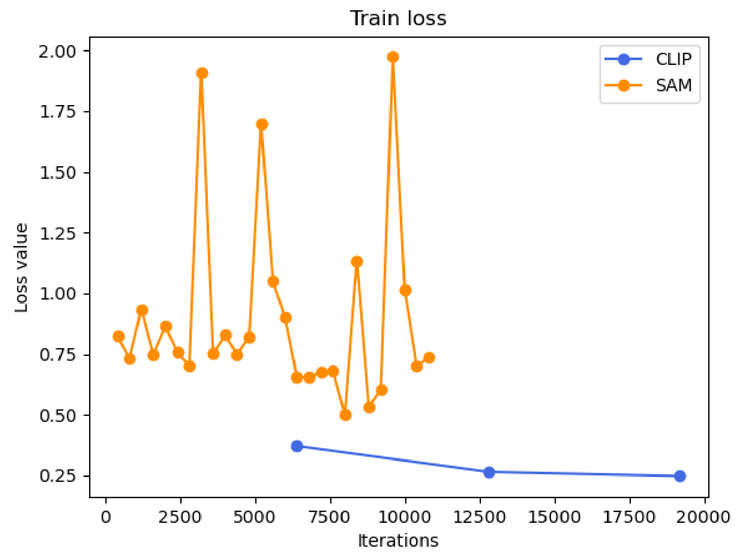
Image samples (real and fake):

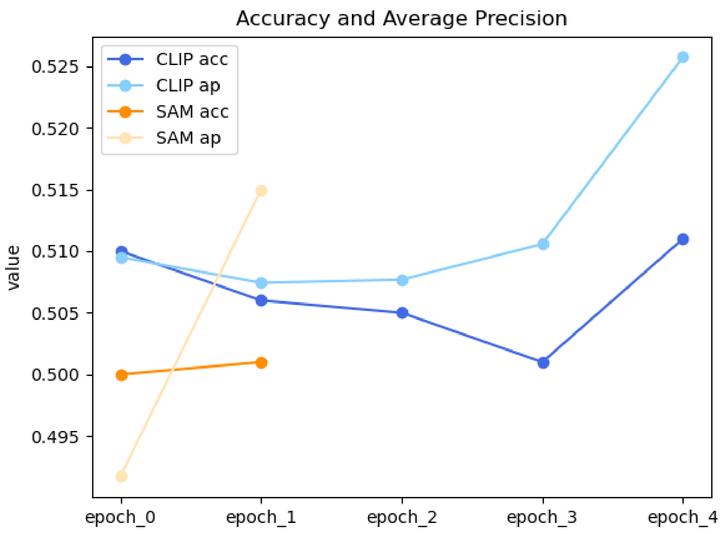
3. Models

The approach to this problem consists in the usage of two types of image encoders over which the authors have added another linear layer. The encoding models are CLIP and SAM(Segment-Anything).
CLIP: The CLIP model was already implemented by the authors. We retrained it on our training dataset and noted the results.
SAM: For the SAM model we recalculated the linear layer size and integrated it accordingly into our project. We then trained it and noted the results from the test set.

4. Results

Test on	**Pre-trained** (CLIP feats)	**Trained on train_data** (CLIP feats)	**Trained on train_data** (SAM feats)
Deepfakes	82.04	87.87	46.36
CycleGAN	99.81	99.76	54.73
ProGAN	100.00	100.00	55.81
StarGAN	99.35	99.69	54.85
LDM	93.22	95.00	54.03
Glide	95.48	95.45	52.57
DALL-E	97.73	98.62	56.49
D3	60.52	56.18	58.86





Train loss and validation loss for both SAM and CLIP encoders.

Both models have varied linear layer neuron sizes
CLIP
CLIP's linear layer has a total of 768 trainable parameters.
SAM
SAM's linear layer has a total of 1.048.576 trainable parameters.

From the learning curves and accuracies we can see that both models were slowly improving both the AP and the Accuracy as well as the loss.

Training time: Each training epoch took 30 minutes for the **CLIP** encoder and over 4 hours for the **SAM** encoder.

5. Conclusion

Experiments concluded that the CLIP-based approach significantly outperformed the SAM approach both in training and inference time and in accuracy. One hypothesis for this is due to SAM being trained to outline objects in an image while CLIP was trained to recognize the objects and thus learn visual concepts. This may help the model with embedding the images differently when they are artificial rather than just outline them in the case of SAM who we suggest does not take into account the actual 'real texture' of an image.