DeepFake Detetion

Nuno Correia (fc58638) e Daniel Neves (fc64504)

Ciências ULisboa — Faculdade de Ciências da Universidade de Lisboa

M.Sc. *Data Science*
Deep Learning class 2024/2025

Objectives

- **The objectives of this project were:**

✓ To develop a model that can correctly identify deep fakes;

✓ To investigate the impact of design choices in deep fake detection.

Problem

- **Problem:** Deep faked videos have become more common and sophisticated.

- **Motivation:** Deep fake detection it is an interesting problem because of the abundance of deep faked videos. Due to this identify fake videos from real ones it is more important than ever.

Methodology

We conducted four experiments to test different design choices:

- **JPEG Augmentation**: To test the importance of compression artifacts, we applied random JPEG compression (quality 10–50, p=0.4) during training;
- **Focal Loss**: We replaced BCE with Label Smoothing with Focal Loss to help detect better subtle fakes;
- **Stronger Backbone**: We used an EfficientNet-B3, training only the classifier head for 3 epochs and then gradually unfreezing network blocks afterwards;
- LSTM Head: To capture temporal cues, we added a Bi-LSTM on top of the XceptionNet outputs, from 10 video frames.

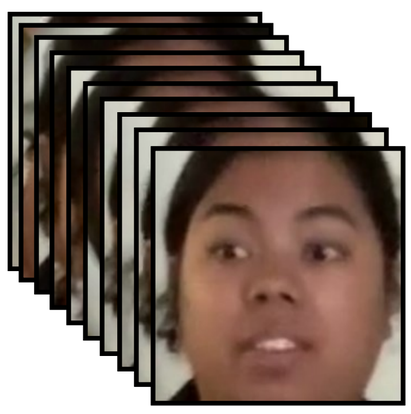
Parameters								
Scheduler	LR	Label Smoothing	Batch	Unfrozen Layers	Decision Threshold	Epoch	TMAX	Betas
AdamW	3,00E-04	0.05	32	18	0.6	10	10	(0.9,0.999)

Augmentation	Parameters
Random Horizontal Flip	p= 0.5
Gaussian Blur	kernel_size= 3, sigma= (0.1,1.5)
Color Jitter	brightness= 0.1, contrast= 0.1, saturation= 0.1,hue= 0.0
Random Resized Crop	224, scale= (0.8,1.0), ratio= (0.9,1.1)
Random Erasing	p= 0.7, scale= (0.02,0.1), ratio= (0.3,3.3)
Normalization	mean = [0.485, 0.456, 0.406], std=[0.229, 0.224, 0.225]

Dataset

- We used 3,136 videos featuring a person. From each video, 10 evenly spaced-out frames were extracted, and faces were cropped using OpenCV.

Real video



Fake video

Results

- Results of the four experiments that were conducted.

Testing

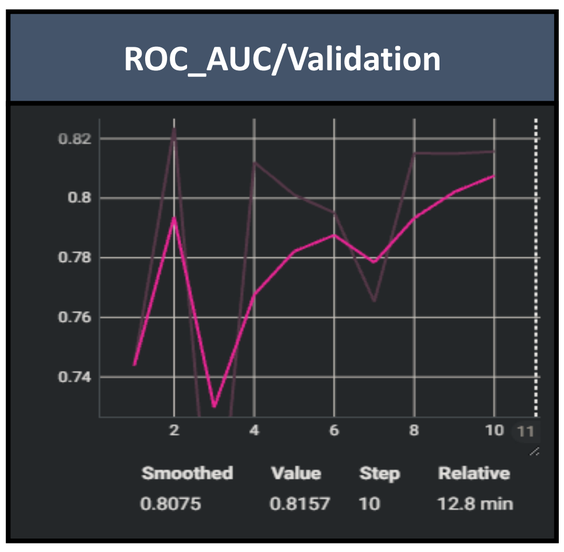
Experiment	AUROC		
	Mean (5 Runs)	CI Mean (**Best Run**)	CI (**Best Run**)
Baseline XceptionNet	0.86	0.93	(0.904, 0.959)
Baseline + JPEG Quality Augmentation	0.81	0.85	(0.803, 0.888)
Baseline + Focal Loss	0.85	0.89	(0.855, 0.923)
EfficientNet B3	0.87	0.90	(0.857, 0.928)
Baseline + LSTM Head	0.92	0.96	(0.931, 0.975)

Experiment	Acc		
	Mean (5 Runs)	CI Mean (**Best Run**)	CI (**Best Run**)
Baseline XceptionNet	0.76	0.88	(0.838, 0.911)
Baseline + JPEG Quality Augmentation	0.71	0.76	(0.707, 0.803)
Baseline + Focal Loss	0.76	0.81	(0.764, 0.850)
EfficientNet B3	0.74	0.82	(0.780, 0.860)
Baseline + LSTM Head	0.83	0.9	(0.866, 0.933)

Experiment	F1		
	Mean (5 Runs)	CI Mean (**Best Run**)	CI (**Best Run**)
Baseline XceptionNet	0.77	0.88	(0.844, 0.917)
Baseline + JPEG Quality Augmentation	0.71	0.74	(0.685, 0.786)
Baseline + Focal Loss	0.71	0.8	(0.743, 0.850)
EfficientNet B3	0.69	0.82	(0.768, 0.857)
Baseline + LSTM Head	0.84	0.91	(0.870, 0.937)

Experiment	EER		GPU Memory	Params (M)
	Mean (5 Runs)	Best Run		
Baseline XceptionNet	0.22	0.13	1686 MB	20.8
Baseline + JPEG Quality Augmentation	0.25	0.25	1701 MB	20.8
Baseline + Focal Loss	0.25	0.20	1686 MB	20.8
EfficientNet B3	0.21	0.17	9526 MB	10.7
Baseline + LSTM Head	0.15	0.09	10866 MB	31.30

- **Best model:** XceptionNet baseline + LSTM Head.



- **AUROC validation learning curve**

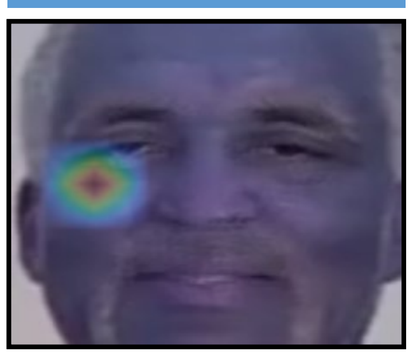
✓ As seen on the graph, validation was still increasing at the 10 epoch where we stopped training.

✓ 5 more epochs might yield +1%/2% AUROC on the training set.

- **Grad-CAM Visualizations of the best model**

Predicted Real

Predicted Fake



- For **"Real"** predictions: barely any activations/none in the face.
- For **"Fake"** predictions: activations are only found on the face.

Conclusions

- **Main takeaway:** Temporal features are more important than image-level detail for this task.
- **Possible future directions:**

✓ Training the best model for 5 more epochs;

✓ Using Focal Loss with the LSTM Head: to help with tricky fakes since our best model only gets 31 videos wrong.

Acknowledgements: This poster session is kindly sponsored by:

NTT DATA