Custom attribution loss for improving generalization and interpretability of deepfake detection

Pavel Korshunov, Anubhav Jain, and Sébastien Marcel
Idiap Research Institute, Martigny, Switzerland

Motivation

Can we improve generalization and interpretability of deepfake detection?

- Typical deepfake detector is a binary classifier.
- Most detection approaches employ a *blackbox* strategy.
- Can we train models for attribution?
- Can training for attribution help in forensics analysis?

Datasets and approaches

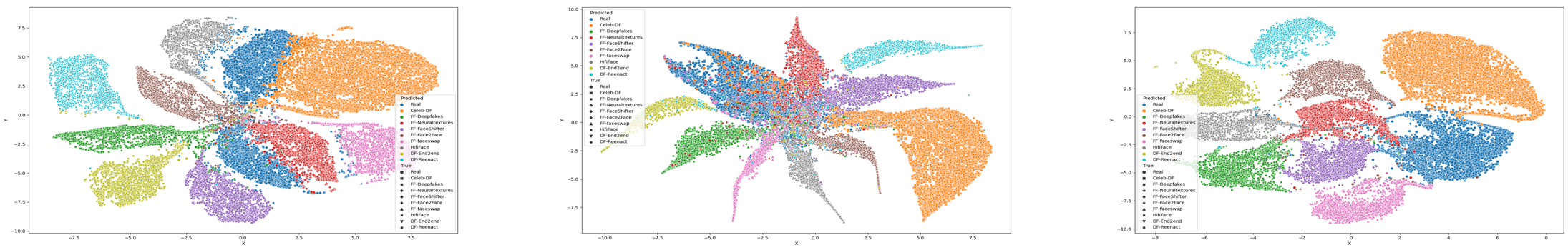
- Datasets
  - Train on FaceForensics++, HifiFace, DeeperForensics, and Celeb-DF
  - Test on the same databases
  - Test on the whole Google and Jigsaw, DF-Mobio, and DeepfakeTIMIT databases
- Approaches (on top of Xception model)
  - Binary
    - Fake and real labels
  - Attribution
    - Classifier with 9 fake classes and one real
  - Triplet-loss based
    - 64-size embedding with semi-hard triplets
    - Convert embedding space into 10 classes using SVM, nearest neighbor (NN), and logistic regression (LR)
  - Our variant of ArcFace-loss (ArcFaceMod)
    - Increase margin for real class and decrease for deepfakes
    - Convert 64-sized embeddings to 10 classes with SVM, NN, and LR

(a) Original (b) Neural Text. (c) Deepfakes (d) Face Shifter (e) Face2Face

(f) Face Swap (g) HifiFace (h) DF End2end (i) DF Reenact

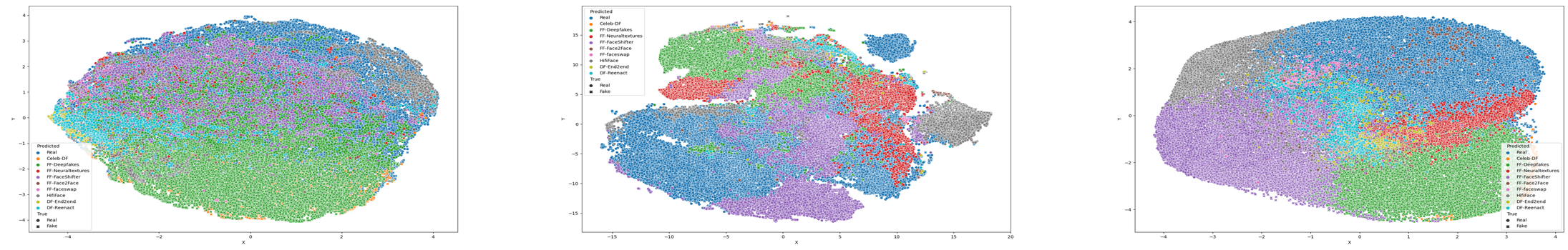
Figure: Cropped faces from FaceForensics++, HifiFace, and DeeperForensics (DF) databases.

Modeling space of deepfakes



(a) Attribution (b) ArcFaceMod-LR (c) TripletLoss-NN

Figure: Tested on the combined Celeb-DF, FF++, HifiFace, and DF. Real videos in blue; marker style – true labels.



(a) Attribution (b) ArcFaceMod-LR (c) TripletLoss-NN

Figure: Tested on DF-Mobio database. Color shows predicted labels; marker style – true labels.

Same- and cross-database evaluation results

Table: Same-database evaluation of deepfake detection.

Approach	Test DB	AUC	FNR (%)	FPR (%)	HTER (%)
Binary	Celeb-DF	98.56	1.12	26.76	13.94
	FaceForensics++	48.03	95.00	4.20	49.60
	HifiFace	25.29	95.00	8.57	51.79
Attribution	Celeb-DF	**100.00**	1.69	0.00	**0.84**
	FaceForensics++	**99.14**	0.71	10.36	5.54
	HifiFace	96.57	0.71	35.00	17.86
TripletLoss-NN	Celeb-DF	99.76	0.56	10.88	5.72
	FaceForensics++	98.84	1.43	9.20	**5.31**
	HifiFace	98.63	1.43	15.00	**8.21**
ArcFaceMod-LR	Celeb-DF	99.87	0.00	5.59	2.79
	FaceForensics++	99.06	1.43	11.34	6.38
	HifiFace	**98.81**	1.43	16.43	8.93

Table: Cross-database evaluation of deepfake detection.

Approach	Test DB	AUC	FNR (%)	FPR (%)	HTER (%)
Binary	DF-Mobio	36.90	95.43	3.70	49.56
	Google	54.01	54.27	34.58	44.43
	DeepfakeTIMIT	70.54	38.60	34.38	36.49
Attribution	DF-Mobio	75.52	12.36	59.66	36.01
	Google	87.89	2.20	56.39	**29.30**
	DeepfakeTIMIT	**84.97**	4.88	46.56	**25.72**
TripletLoss-NN	DF-Mobio	**83.15**	22.60	26.03	**24.32**
	Google	84.15	6.06	54.11	30.08
	DeepfakeTIMIT	70.08	52.33	21.25	36.79
ArcFaceMod-LR	DF-Mobio	79.98	9.73	58.65	34.19
	Google	**88.79**	0.55	69.46	35.00
	DeepfakeTIMIT	63.55	27.21	62.19	44.70

Contact: Pavel.Korshunov@idiap.ch
Source code: https://gitlab.idiap.ch/bob/bob.paper.deepfake_attribution/-/tree/icassp2022