CS588 Midterm Project Presentation

Deepfake Retrieval Systems: Detecting Identity Fraud in Image Databases

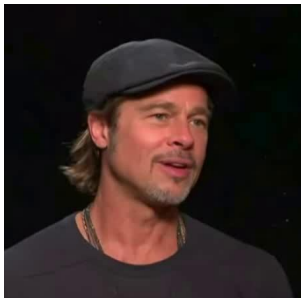
Jumin Lee and Suhyeon Ha (T4)

2024. 05. 08.

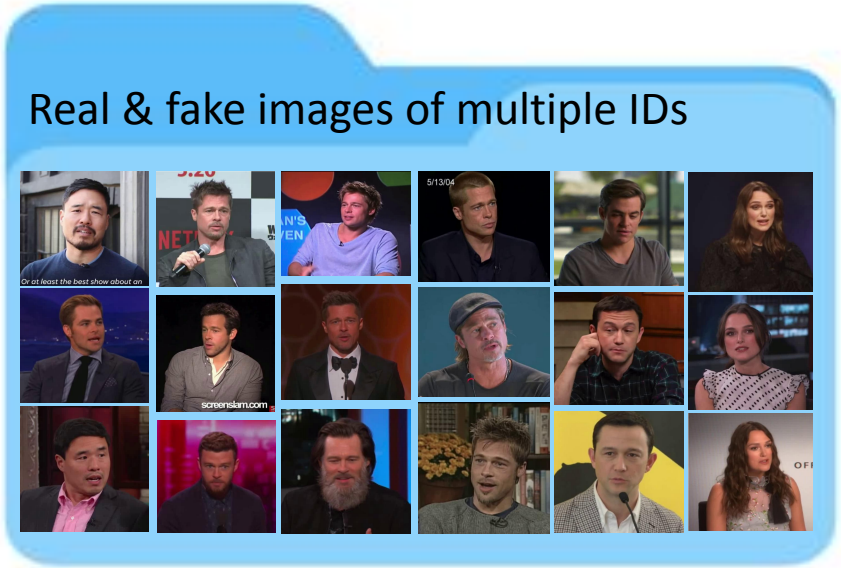
SGVR Lab KAIST

Target Task

- Given an authentic image, our goal is to detect fake images pretending to depict the same person in database.



Query



Database

Results

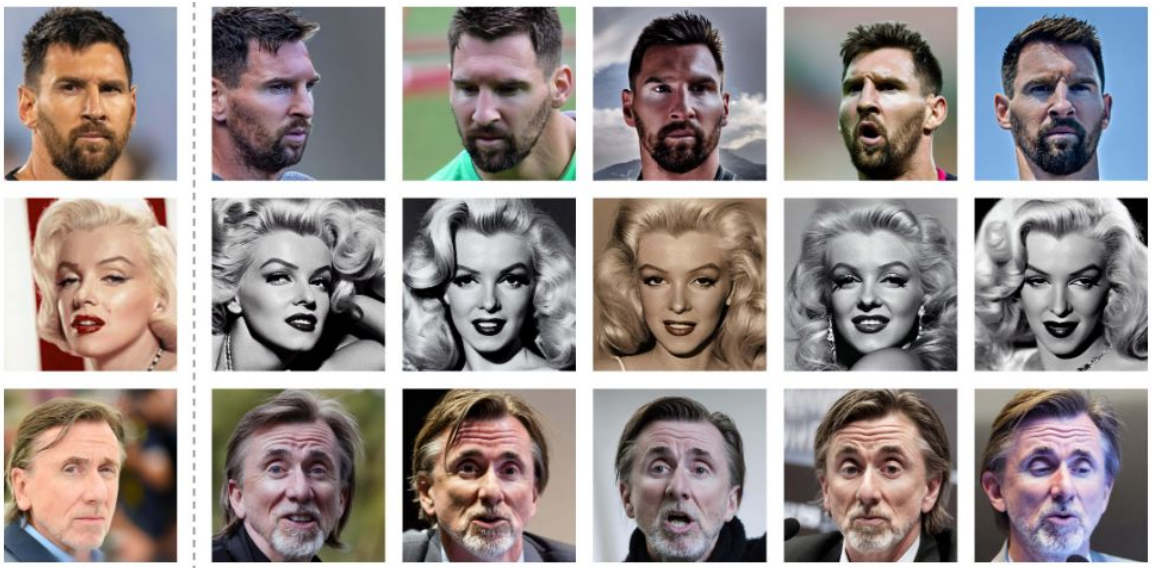
Introduction

- Why it is important?
  : The rise of deepfake presents significant risks in misinformation, identity theft, and privacy invasion.

- Detecting deepfakes that impersonate a specific identity is critical for safeguarding individuals and society.

Introduction

- Why it is important?
  - Arc2Face: A Foundation Model of Human Faces, arxiv 2024



ID Generated images with consistent ID

Introduction

- Why it is important?

- Arc2Face: A Foundation Model of Human Faces, arxiv 2024

  : Introduce a large dataset of high-resolution facial images with consistent ID and intra-class variability, and an ID-conditioned face model trained on it, which:

✓ generates high-quality images only its ArcFace embedding

✓ offers superior ID similarity compared to existing models

✓ can be extended to different input modalities, e.g. pose/expression

Related Work

- No existing work to retrieve deepfakes of the query image.

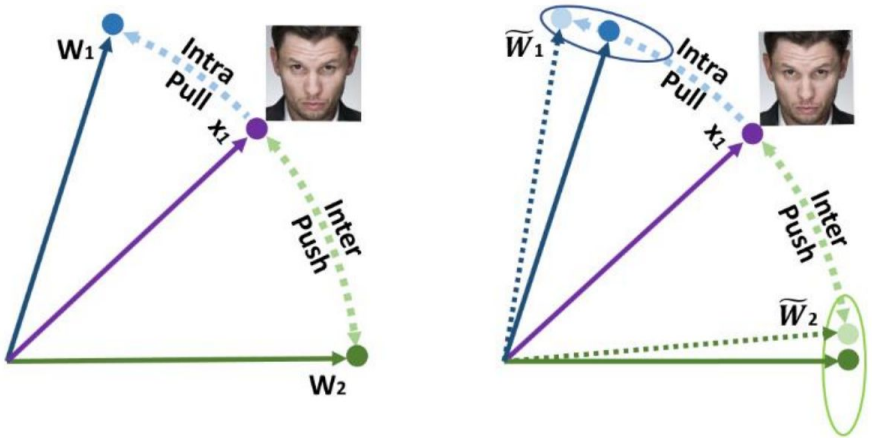
- A combination of face recognition and deepfake detection can be utilized.
  - Stage 1. Face retrieval
    : Identify images that match the given identity.
    Being unrecognized as someone's identity suggests its quality is doubtful.
    ex) Variational Prototype Learning for Deep Face Recognition, CVPR 21

  - Stage 2. Deepfake detection
    : Determine whether the identified **face images** have been manipulated.

Related Work

- Variational Prototype Learning for Deep Face Recognition, CVPR 2021

  : Propose a novel **Variational Prototype Learning** method which represents each class as a <u>distribution</u> instead of a <u>point</u> by using the margin-based softmax loss.



(a) Prototype Learning (b) Variational Prototype Learning

Related Work

- No existing work to retrieve deepfakes of the query image.

- A combination of face recognition and deepfake detection can be utilized.
  - Stage 1. Face retrieval
    : Identify images that match the given identity.
    Being unrecognized as someone's identity suggests its quality is doubtful.
    ex) Variational Prototype Learning for Deep Face Recognition, CVPR 21

  - Stage 2. Deepfake detection
    : Determine whether the identified **face images** have been manipulated.

Related Work

- Prompt-guided inpainting can modify images while preserving their identities.
- (Stage 2. Deepfake detection) can't handle this issue.

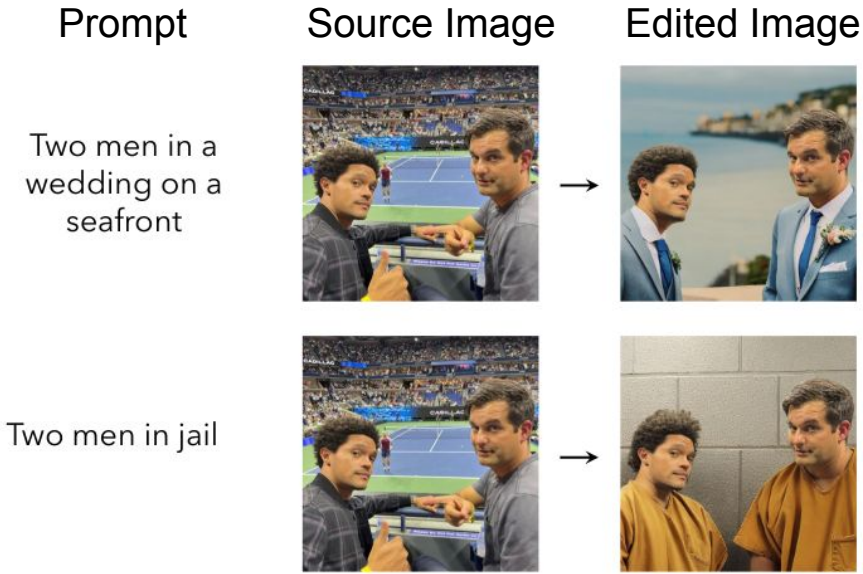


Image Credit: Raising the Cost of Malicious AI-Powered Image Editing

Related Work

- No existing work to retrieve deepfakes of the query image.

- A combination of face recognition and forgery detection can be utilized.
  - Stage 1. Face retrieval

    : Identify images that match the given identity.

    Being unrecognized as someone's identity suggests its quality is doubtful.

    ex) Variational Prototype Learning for Deep Face Recognition, CVPR 21

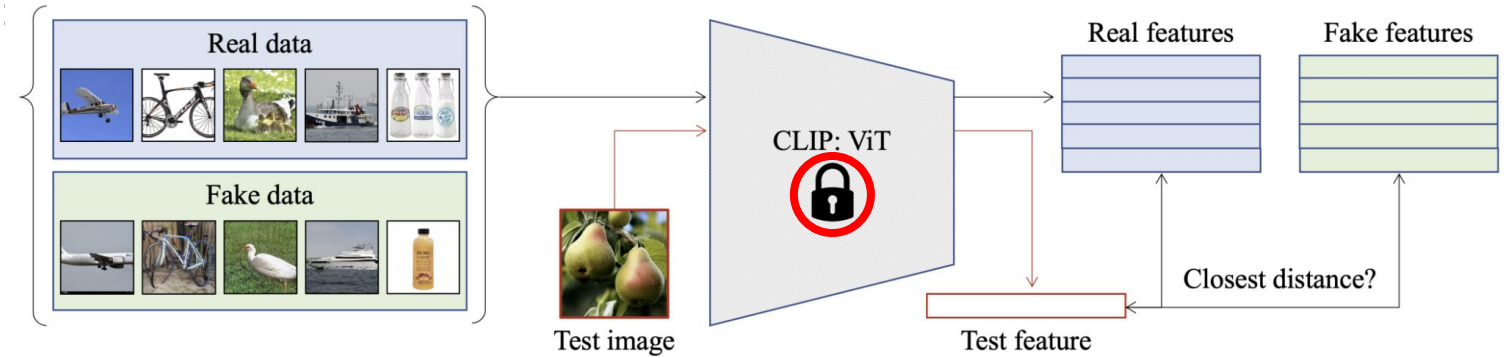
  - Stage 2. Forgery detection

    : Determine whether the identified **arbitrary images** have been manipulated.

    ex) Towards Universal Fake Image Detectors that Generalize Across Generative Models, CVPR 23

Related Work

- Towards Universal Fake Image Detectors that Generalize Across Generative Models, CVPR 23



- No training of real vs. fake classifiers

  : The classification process should happen in a **feature space** which has not been trained to separate images from the two classes.

Related Work

- Towards Universal Fake Image Detectors that Generalize Across Generative Models, CVPR 23
- Results

< Classification accuracy (averaged over real and fake images) >

Detection method	Variant	Generative Adversarial Networks						Deep fakes	Guided	LDM			Glide			DALL-E	Total
		Pro-GAN	Cycle-GAN	Big-GAN	Style-GAN	Gau-GAN	Star-GAN			200 steps	200 w/ CFG	100 steps	100 27	50 27	100 10		Avg. acc
Ours	NN, $k = 1$	99.58	94.70	86.95	80.24	96.67	98.84	80.9	68.76	89.56	68.99	89.51	86.44	88.02	87.27	77.52	82.30
	NN, $k = 3$	99.58	95.04	87.63	80.55	96.94	98.77	83.05	70.02	90.37	70.17	90.57	87.84	89.34	88.78	79.29	83.28
	NN, $k = 5$	99.60	94.32	88.23	80.60	97.00	98.90	83.85	70.55	90.89	70.97	91.01	88.42	90.07	89.60	80.19	83.72
	NN, $k = 9$	99.54	93.49	88.63	80.75	97.11	98.97	**84.5**	**71.06**	91.29	72.02	91.29	**89.05**	**90.67**	**90.08**	81.47	**84.25**
	LC	**100.0**	98.50	**94.50**	82.00	**99.50**	97.00	66.60	70.03	**94.19**	73.76	**94.36**	79.07	79.85	78.14	**86.78**	81.38

  - Reproduce official code using a linear classifier

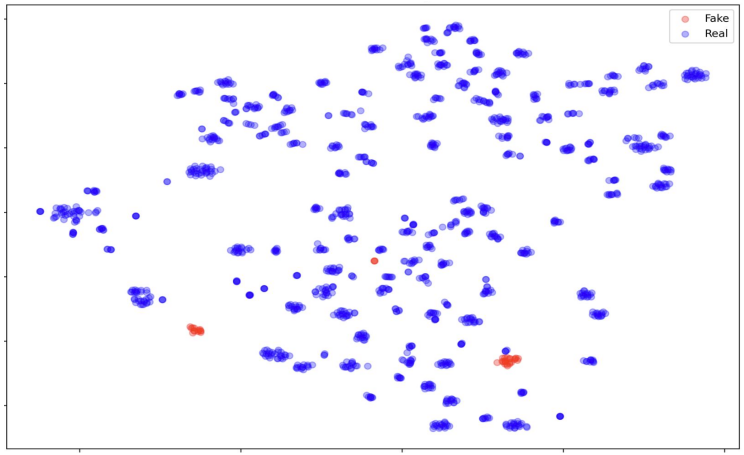
    : Real Accuracy = 94%, **Fake Accuracy = 42%**, Average Accuracy = 68%

Related Work Analysis

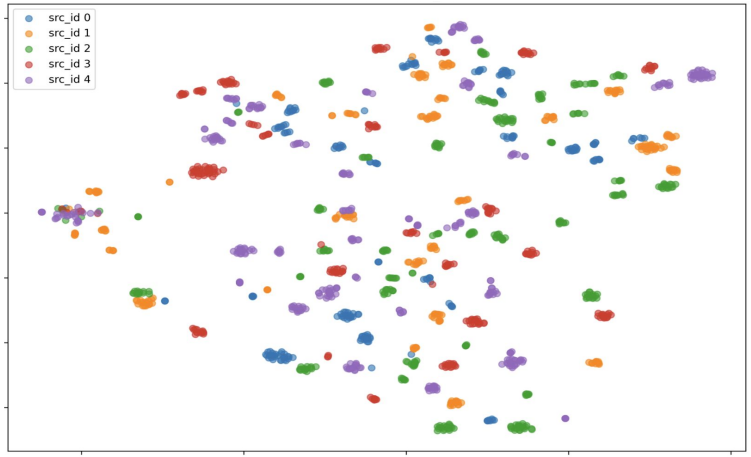
- Checked the performance of clip embedding for Face retrieval and Forgery detection.
- Results

< t-SNE Visualization >

< Real/Fake >



< Selected IDs >



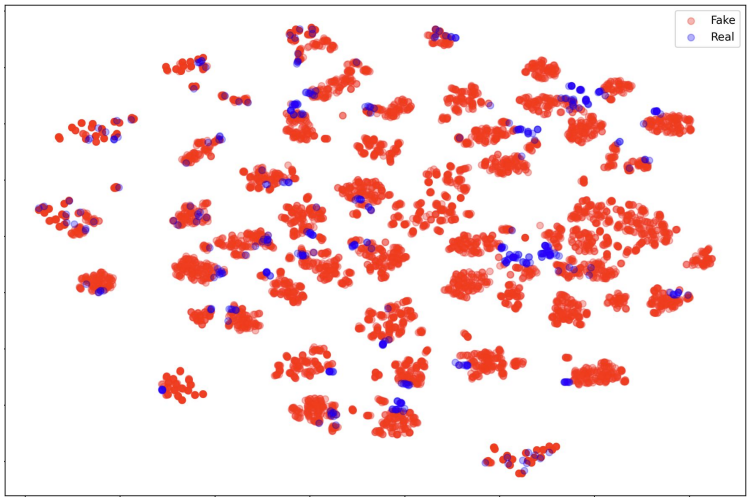
  - Reproduce official code using a linear classifier

    : Real Accuracy = 94%, **Fake Accuracy = 42%**, Average Accuracy = 68%

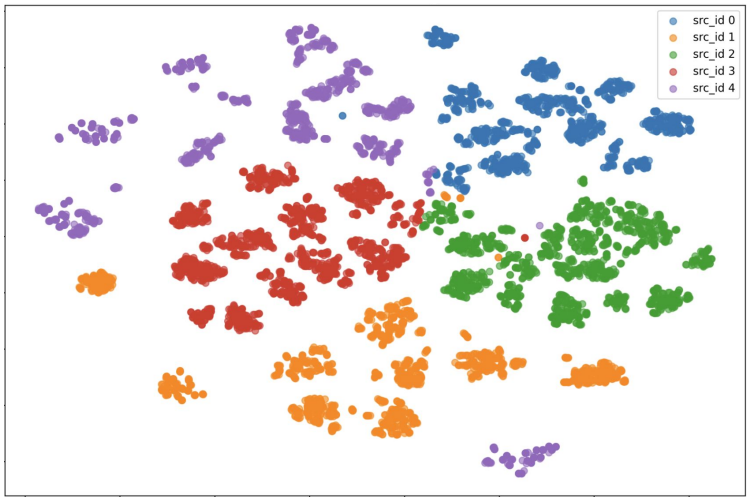
Related Work Analysis

- Checked the performance of clip embedding for Face retrieval and Forgery detection.
- We tried applying the model to another dataset (Celeb-DF).

< Real/Fake >



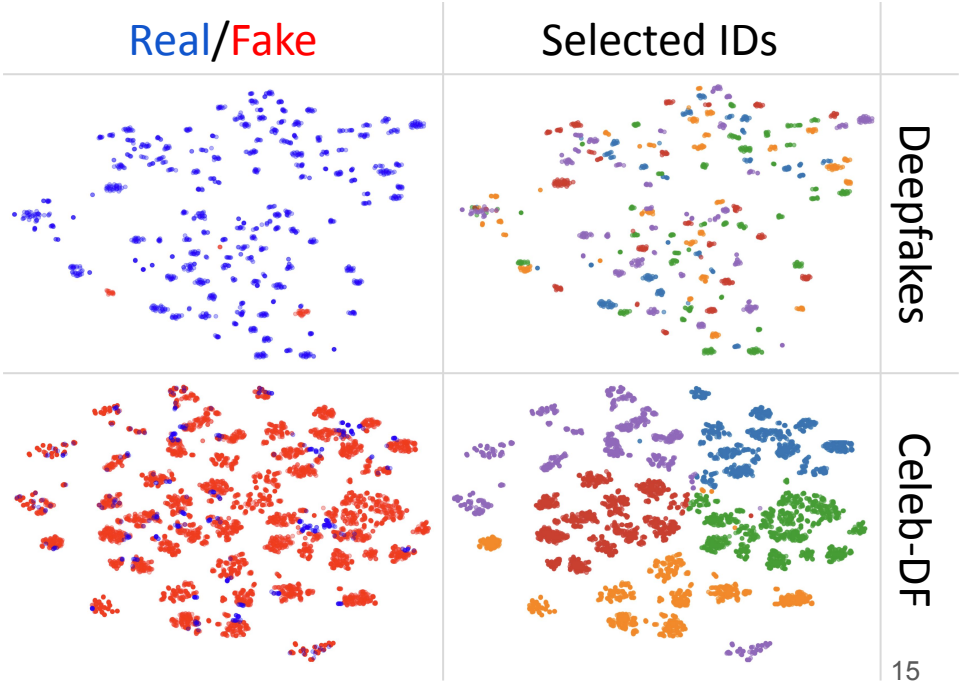
< Selected IDs >



- Real Accuracy = 99%, Fake Accuracy = 1%, Average Accuracy = 11%

Related Work Analysis

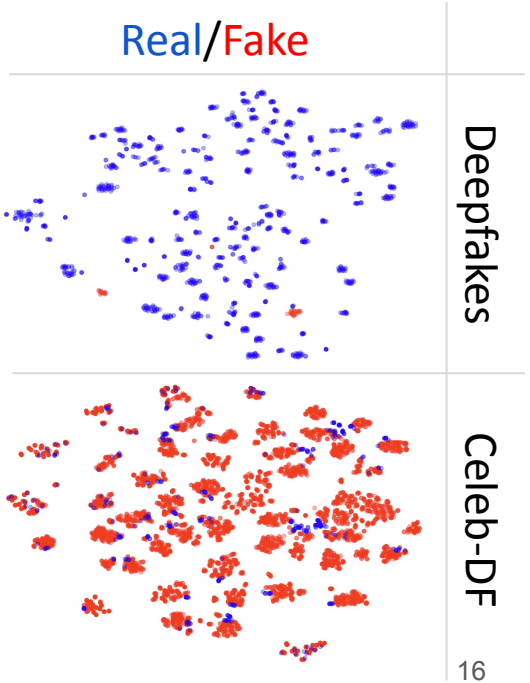
- Show different pattern between 2 datasets.
- Deepfakes Dataset
  - Face retrieval 😭
  - Forgery detection 😐
- Celeb-DF Dataset
  - Face retrieval 😊
  - Forgery detection 😭



Challenges

- Performance of forgery detector highly depends on facial datasets.
- There is still room for improvement in forgery detection for both datasets.

→ Our goal is to improve facial forgery detection of UniDet for deepfake retrieval systems



Roles

- Jumin
  - Pre-process datasets and visualize embedding spaces (done)
  - Analyze the features of the cropped face and the whole image (~ May 18th)
- Suhyeon
  - Reproduce UniDet (done)
  - Analyze the given query image features (~ May 18th)
- Integrate cropped face and whole image features with query image features (~ May 30th)

Q&A

Thank you