Deepfakes

Jawwad A Shamsi

What is Deepfake

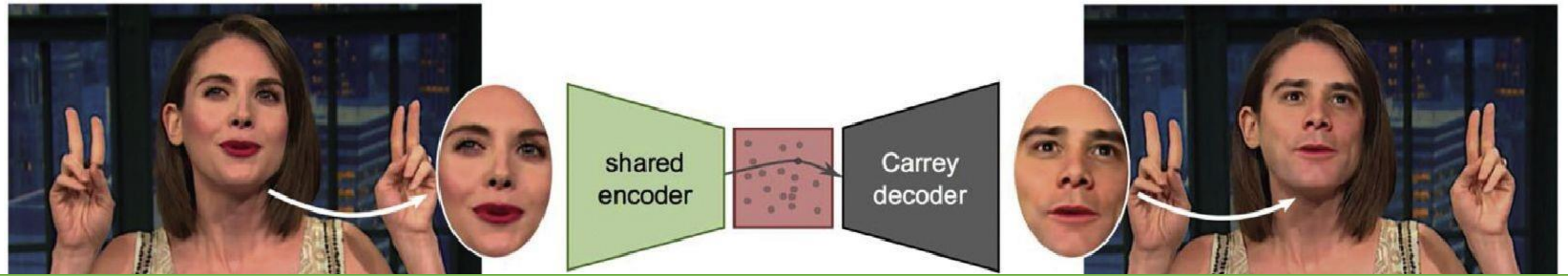
- A synthetic Media
  - In which a person in an existing image or video is replaced with a fake image or a fake video

Alison Brie

Deepfake with Jim Carrey

Deepfakes: Example

- a screenshot
  - of Alison Brie from the original talk show
  - interview on the left,
- and on the right is a frame
  - from the resulting deepfake video featuring Brie's
  - body with Carrey's face.



Step 1: extract Brie face

Step 3: insert fake Carrey face

Step 2: create fake Carrey face

Three Steps

- The region of the image showing Brie's face was extracted from an original movie frame;
- Using DNN, this image was then used to automatically generate a matching image showing Carrey instead (Step 2); and

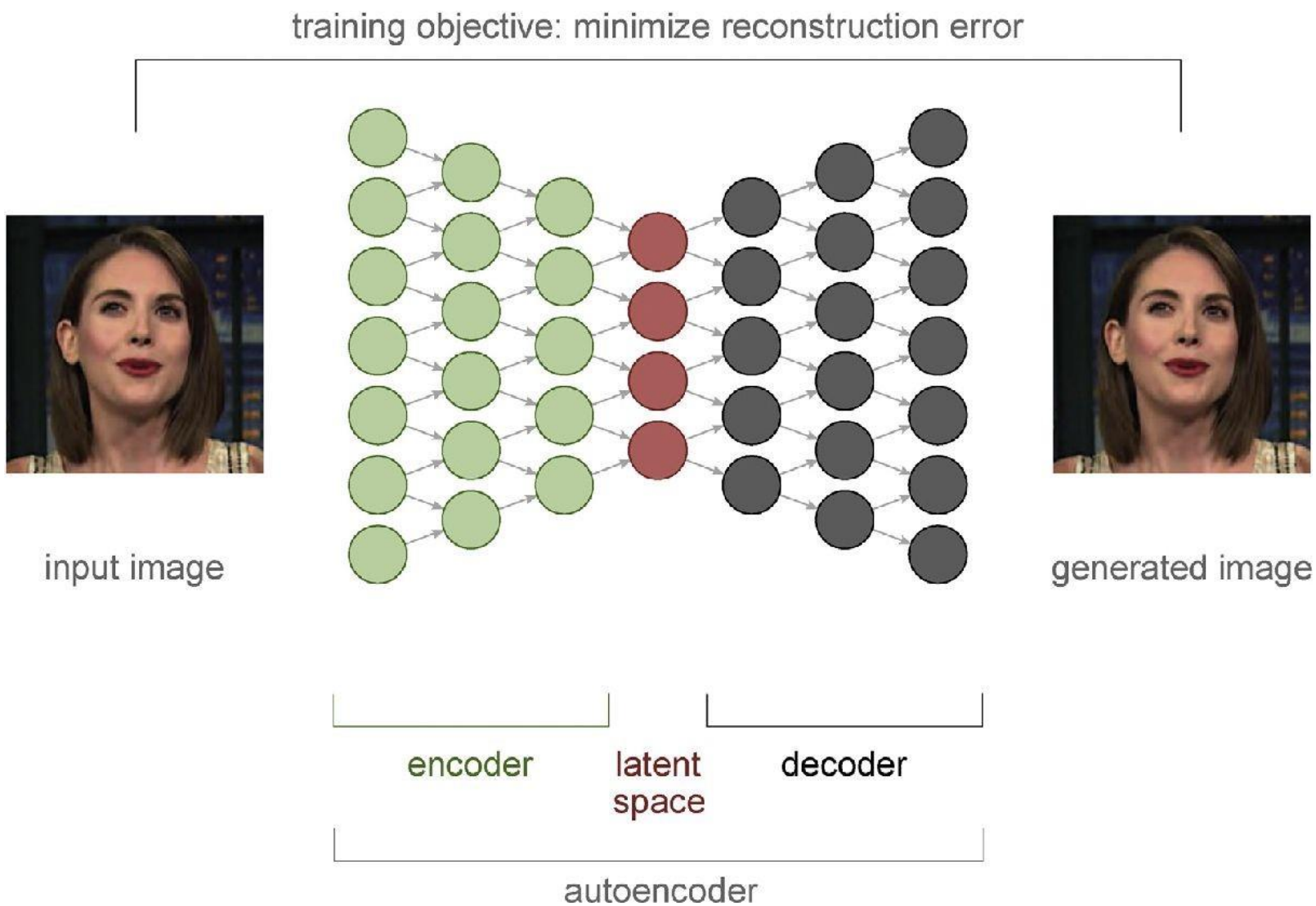
- This generated face was then inserted into the original reference image to create the deepfake.

Auto Encoder

- Based on a given, large set of input images (e.g., all depicting Alison Brie), an autoencoder is trained to recognize key characteristics of a face and subsequently recreate input images as its output.

Three Steps

- Encoder: Recognize a comparably small number of facial characteristics in the input
- Latent Space:
- Decoder: Generate real-looking faces as output


training objective: minimize reconstruction error
input image
generated image
encoder
latent space
decoder
autoencoder

Encoder

- It takes tens of thousands of pixels and compresses them into typically 300 measurements that relate to particular facial characteristics.
  - whether the eyes are open or closed,

- the head pose, the emotional expression, the eye expressions, • ambient light, or skin colore

Latent Space

- Information bottlenecks
- Learn General facial Characteristics

- For the autoencoder, this bottleneck is needed so that the network can learn more general facial characteristics rather than memorizing all input examples of specific people.

Decoder

the purpose of the decoder is to decompress 300 measurements in order to reconstruct an image as truthfully as possible.

Problem

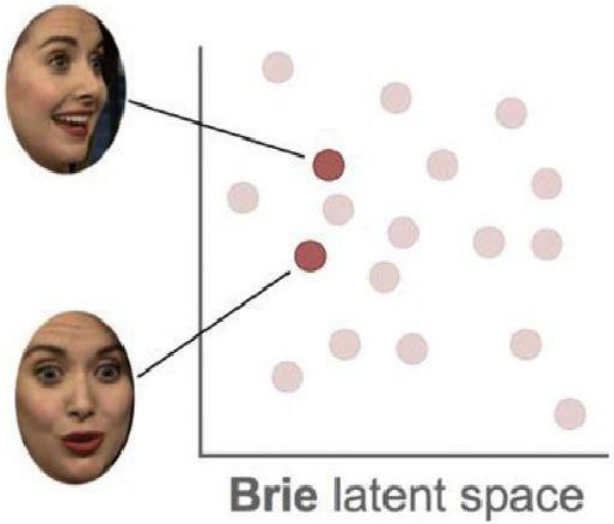
- every point in the latent space corresponds to an image of a given person.
- Which point in this vast space of nearly infinite possibilities will correspond to the image we desire.

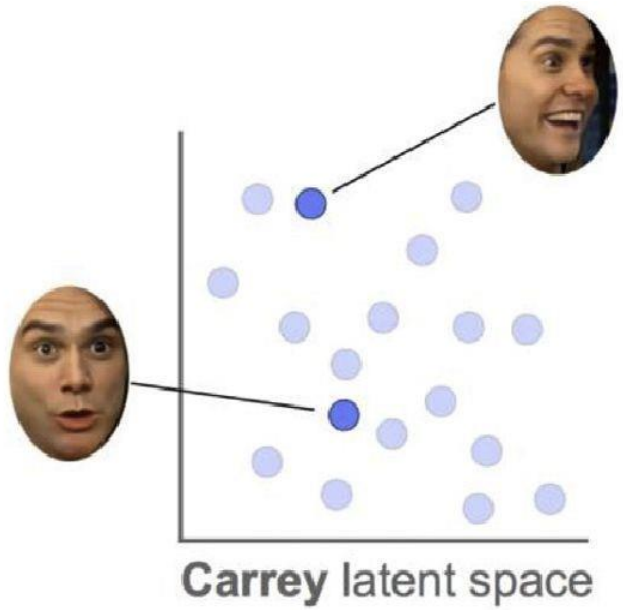
The deepfake trick

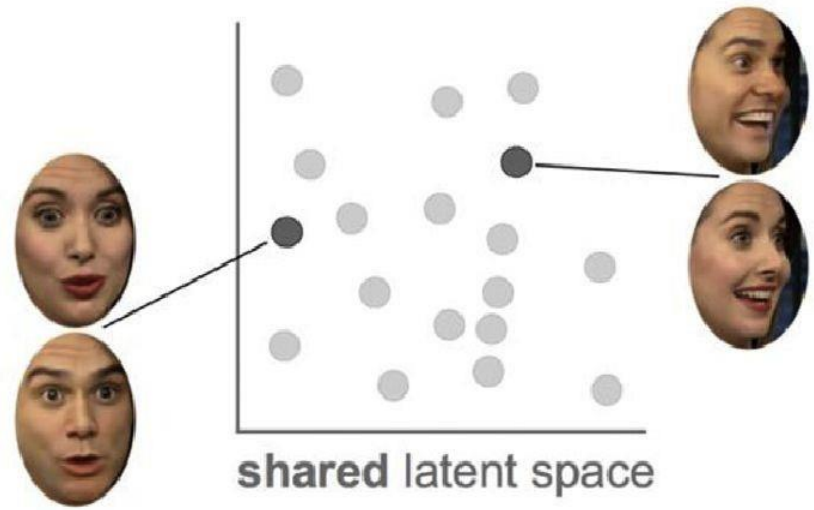
- The trick for creating deepfakes is to set up the structure of the autoencoder in a way that an image of another person can act as a guide to help find the specific combination of measurements that yields the desired image

The deepfake trick (2)

- The trick to making this possible is using the same shared encoder for both people. In the encoding process, the DNN selects 300 measurements it deems meaningful based on the training images for each person. If images of two people are compressed on separate encoders different features would be seen as meaningful and we could not combine them in a valuable way


Brie latent space


Carrey latent space


shared latent space

The deepfake trick(3)

- The autoencoder trick is to train two autoencoders, each with a person-specific decoder, using the exact same encoder. This encoder will learn to use general features that the faces of both people have in common.

The deepfake trick(4)

- This allows for similar pictures of two different people to be positioned in a similar location of the latent space. For example, pictures showing either a smiling Carrey or Brie will lead to very similar measurements, or unit activations, in the latent space

Table 1. Types and examples of deepfakes

Type	Description	Current example	Business application
Photo deepfakes	**Face and body-swapping** Making changes to a face, replacing or blending the face (or body) with someone else's face (or body)	FaceApp's aging filter alters your photo to show how you might look decades from now (Kaushal, 2019)	Consumers can try on cosmetics, eyeglasses, hairstyles, or clothes virtually
Audio deepfakes	**Voice-swapping** Changing a voice or imitating someone else's voice	Fraudsters used AI to mimic a CEO's voice and then tricked a manager into transferring $243,000 (Suwajanakorn et al., 2017)	The voice of an audiobook narration can sound younger, older, male, or female and with different dialects or accents to take on different characters
	Text-to-Speech Changing audio in a recording by typing in new text	Users made controversial Jordan B. Peterson, a famous professor of psychology and author, say anything they wanted until his threat of legal action shut the site NotJordanPeterson down (Cole, 2019)	Misspoken words or a script change in a voiceover can be replaced without making a new recording
Video deepfakes	**Face-swapping** Replacing the face of someone in a video with the face of someone else	Jim Carrey's face replaces Alison Brie's in a *Late Night with Seth Meyers* interview.	Face-swapped video can be used to put the leading actor's face onto the body of a stunt double for more realistic-looking action shots in movies.
	Face-morphing A face changes into another face through a seamless transition	Former *Saturday Night Live* star Bill Hader imperceptibly morphs in and out of Arnold Schwarzenegger on the talk show *Conan*	Video game players can insert their faces onto their favorite characters
	Full-body puppetry Transposing the movement from one person's body to that of another	"Everybody Dance Now" shows how anyone can look like a professional dancer	Business leaders and athletes can hide physical ailments during a video presentation.
Audio & video deepfakes	**Lip-syncing** Changing the mouth movements and words spoken in a talking head video	In "You Won't Believe What Obama Says In This Video!" Jordan Peele edits Obama to use profanity in a public service announcement	Ads and instructional videos can be 'translated' into other languages using the same voice used in the original recording

Positives: In movies

- voice dubbing
- fix misspoken lines or make script changes
  - without rerecording footage
  - create seamless dubs of actors speaking different languages.
- Stunt doubles can be created

▪Actors can look older or younger with the use of deepfakes ▪

instead of time-consuming make-up.

Negatives

•the entire dubbing and re-voicing industry, which has long translated movies so that the new words match the original lip movement of the actor, is endangered and at risk of becoming extinct now that languages and lips can be changed.

Negatives (2)

- anyone with deepfake technology can make a powerful politician an artificial intelligence (AI) puppet and make them say things, which they may have never said in real life.

Malicious usage

- President Obama video for President Donald Trump
- deepfake audio was used toscam a CEO of a UK-based energy firm and robbedC220,000
- Real video of US speaker, believed to be appear as fake.

How to detect

- Pay attention to the face. High-end DeepFake manipulations are almost always facial transformations.

- · Pay attention to the cheeks and forehead. Does the skin appear too smooth or too wrinkly? Is the agedness of the skin similar to the agedness of the hair and eyes? DeepFakes are often incongruent on some dimensions.
- · Pay attention to the eyes and eyebrows. Do shadows appear in places that you would expect? DeepFakes often fail to fully represent the natural physics of a scene.
-
- · Pay attention to the glasses. Is there any glare? Is there too much glare? Does the angle of the glare change when the person moves? Once again, DeepFakes often fail to fully represent the natural physics of lighting.

How to detect(2)

- Pay attention to the facial hair or lack thereof. Does this facial hair look real? DeepFakes might add or remove a mustache, sideburns, or beard. But, DeepFakes often fail to make facial hair transformations fully natural.
- Pay attention to facial moles. Does the mole look real?
- Pay attention to blinking. Does the person blink enough or too much?
- Pay attention to the size and color of the lips. Does the size and color match the rest of the person's face?

Conclusion

- Emerging topic
  - Lots of Challenges and Opportunities

References

- Deepfakes : Trick or Treat
- Youtube video https://www.youtube.com/watch?v=l82PxsKHxYc
- The Emergence of Deepfake Technology: A Review. MIT Technology Review
- Deepfake Detection using Capsule Networks withLong Short-Term Memory Networks
- Detect Deepfakes: Howto counteract. MIT Media

https://www.media.mit.edu/projects/detect-fakes/overview/