Privacy preserving representation learning for multimodal behavior understanding

Aabha Ranade (7579610559)
Ravi Vivek Agrawal (3320524348)
Shivangi Kochrekar (3320265465)
Vaidehi Vatsaraj (2671157553)

Problem definition

Features commonly used for behavior understanding often contain person specific (biometric) information. The goal of this project is to learn representations that are useful for the downstream task (emotion recognition/speech recognition) that cannot be used to re-identify the speaker so that their voice patterns or facial features, are not exploited for unauthorized identification or surveillance purposes.

Background

Paper Title	Methodology	Results
Privacy Enhanced Multimodal Neural Representations for Emotion Recognition Mimansa Jaiswal, Emily Mower Provost (University of Michigan)	The main network consists of three components: 1) Embedding sub-network: multimodal late fusion. 2) Emotion classifier 3) Gender classifier The main network is trained to unlearn gender Attacker network (used 12 variants)	Privacy preserving networks trained for emotion recognition can be used to protect against gender and membership identification Future work: explore how privacy can be enhanced for multiple primary tasks
3d Face Reconstruction With Dense Landmarks Wood, Erroll and Baltru (Microsoft)	Predict probabilistic dense landmarks L, each with position μ and certainty σ using CNN. Then, fit a 3D face model to L, minimizing an energy E by optimizing model parameters.	Dense landmarks are ideal for integrating face shape information across frames by demonstrating accurate and expressive facial performance capture
Facial Expression Translation using Landmark Guided GANs Hao Tang and Nicu Sebe	LandmarkGAN: 1) category -guided landmark generation 2) Landmark-guided expression-to-expression translation	Experimental results on 4 challenging datasets Capable of generating higher quality faces with correct expression than the state-of-the-art approaches and is task agnostic
LandmarkGAN: Synthesizing faces from landmarks Pu Sun, Yuezun Li, Honggang Qi, Siwei Lyu	Input: Facial landmarks → Converted facial landmarks → Target face with source expressions The landmark convertor has an autoencoder structure which is then fed to a TL2F generator	Train and test the method on CelebV dataset. Face Synthesis method is flexible to larger changes to the landmarks but has a limit. Used 3 metrics for quantitative evaluation: LMK, SSIM, ID.

Background

Paper Title	Methodology	Results
Privacy-preserving Representation Learning for Speech Understanding Tran, M. and Soleymani, M.	Privacy Transformer: trained on Voice cloning dataset Pretrained encoder : HuBERT :Raw waveform → low-level features → high level representations. Privacy Transformer : ● Embedding layers (2 layers) - Transformer encoder (does not use positional encoding) - Fully connected layer	Speaker identification: achieves similar accuracy while outperforming baselines on paralinguistic tasks Future work: boost with Distill-HuBERT
Privacy and Utility Preserving Data Transformation for Speech Emotion Recognition Tiantian Feng, and Shrikanth Narayana	Uses a Gradient Reversal Layer to unlearn the gender information, minimize reconstruction error 1) Risk Model 2) Replacement AutoEncoder 3) Gradient Reversal Layer	Varied accuracy in predicting different emotions. Transformed data using the RAE decreases the accuracy in predicting gender
Representation Learning For Audio Privacy Preservation Using Source Separation And Robust Adversarial Learning Diep Luong, Minh Tran, Shayan Gharib, Konstantinos Drossos, Tuomas Virtanen	Proposed method: RDAL-M: 1) Learns the latent representations of audio recordings 2) Prevents the differentiation between speech and non-speech recordings. 3) Uses source separation: removes privacy-sensitive signals 4) Adversarial learning setup between a feature extractor and a speech classifier to preserve audio privacy.	Obscures speech presence within audio recordings while significantly preserving the performance of the utility task

Background

Paper Title	Methodology	Results
Deep autoencoders for attribute preserving face de-identification Paraskevi Nousi, Sotirios Papadopoulos, Anastasios Tefas, Ioannis Pitas	Face de-identification using Deep Autoencoders, by fine tuning the encoder to perform face de-identification. Various methods to finetune the encoder in both a supervised and unsupervised fashion to preserve facial attributes, while generating new faces which are both visually and quantitatively different from the original ones. Dataset - LFW dataset, CelebA	Lightweight de-identification pipeline for embedded systems. Tested on NVIDIA Jetson TX2, suitable for UAVs. Quick processing with CUDA GPU: 0.65 ms for 64×64 face images. High de-identification rates with realistic results. Capable of real-time operation with an efficient face detector.
Voice privacy using CycleGAN and time-scale modification Gauri P. Prajapati, Dipesh K. Singh, Preet P. Amin, Hemant A. Pat	Cycle Consistent Generative Adversarial Network (CycleGAN) to modify (transform) the speaker's gender as well as the other prosodic aspects using their Mel cepstral coefficients (MCEPs) and fundamental frequency	Results were tested for 101 speakers
X-vector anonymization using autoencoders and adversarial training for preserving speech privacy Juan M. Perero-Codosero a b 1, Fernando	Extracted an x-vector, which is an utterance-level embedding and then transformed by an autoencoder where speaker, gender, and accent information are suppressed through adversarial training. Anonymized speech generated through a neural speech synthesizer	Proposed system outperforms the baseline in terms of privacy and utility for the majority of the evaluated attack scenarios

Datasets

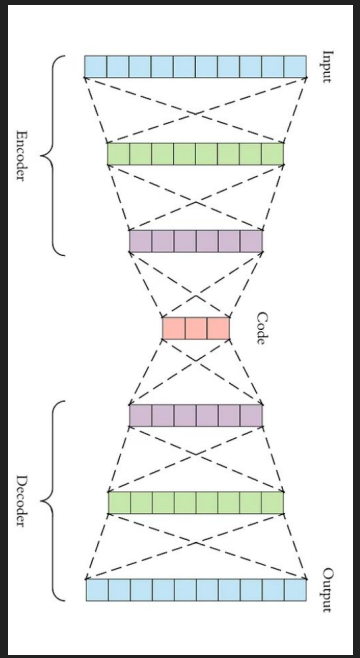
CREMA- D

- 7,442 original clips from 91 actors.
- 48 male and 43 female actors
- Ages between 20 - 74 years
- Variety of races and ethnicities (African American, Asian, Caucasian, Hispanic)
- Actors spoke from a selection of 12 sentences
- Average length of 2.63 ± 0.53 seconds
- Six different emotions (Anger, Disgust, Fear, Happy, Neutral and Sad)
- Four different emotion levels (Low, Medium, High and Unspecified)

Flickr-Faces-HQ (FFHQ)

- 70,000 high-quality PNG images
- 1024×1024 resolution
- Considerable variation in terms of age, ethnicity and image background
- Accessories such as eyeglasses, sunglasses, hats, etc.

Method



Video Dataset

Visual Component

Video to Frames

Image Augmentation

Encoding - Decoding

Combining reconstructed frames

Audio Component

Extract mfccs, introduced noise

Vector quantization using K means clustering

Reconstruct audio using centroids

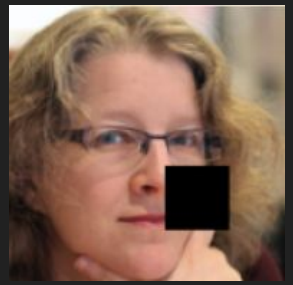
- Speaker Identification
- Emotion Recognition

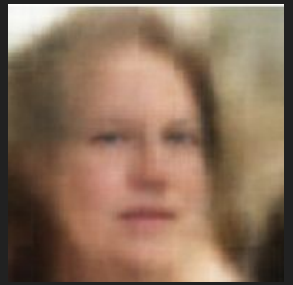
Method - Video

1. Image preprocessing - resizing, adding noise
2. Custom generator function to load images in the batches.
3. Train the autoencoder on 54k images

   (learning rate = [0.01, 0.001], optimizers = [Adam, SGD], Loss = mse, Batch size= [32, 64], Epochs = [50, 100, 150] Latent space size = 300)

4. Save the model with training accuracy 84.65%
5. Convert video to frames (Crema-D) and preprocess the frames
6. Get reconstructed image of each frame using the saved model
7. Combine reconstructed frames to make video
8. Test video for emotion recognition using a pretrained emotion recognition model (FER)
9. Test for speaker identification using pretrained vgg16 model.





Autoencoder Model Architecture

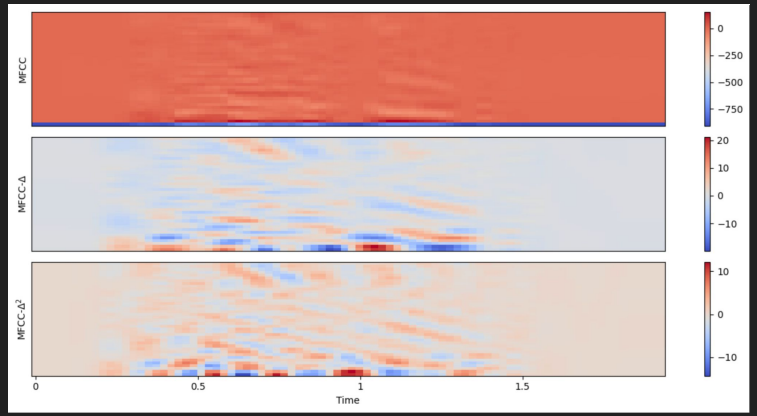
Layer (type)	Output Shape	Param #
input_layer (InputLayer)	(None, 128, 128, 3)	0
conv2d (Conv2D)	(None, 64, 64, 128)	9,728
max_pooling2d (MaxPooling2D)	(None, 32, 32, 128)	0
leaky_re_lu (LeakyReLU)	(None, 32, 32, 128)	0
batch_normalization (BatchNormalization)	(None, 32, 32, 128)	512
conv2d_1 (Conv2D)	(None, 16, 16, 64)	73,792
max_pooling2d_1 (MaxPooling2D)	(None, 8, 8, 64)	0
leaky_re_lu_1 (LeakyReLU)	(None, 8, 8, 64)	0
batch_normalization_1 (BatchNormalization)	(None, 8, 8, 64)	256
flatten (Flatten)	(None, 4096)	0
dense (Dense)	(None, 4096)	16,781,312

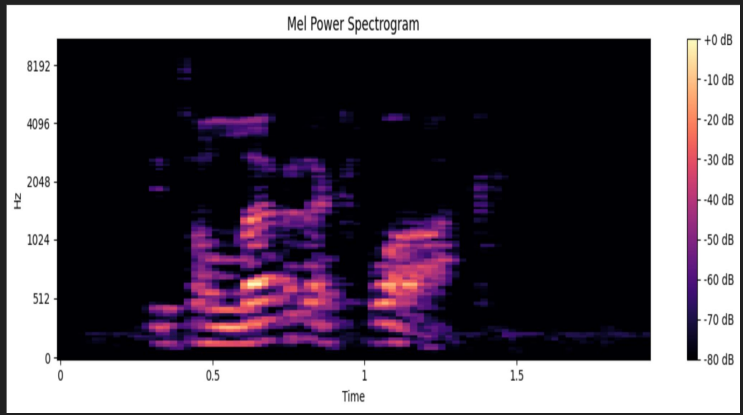
Layer (type)	Output Shape	Param #
input_layer_1 (InputLayer)	(None, 4096)	0
dense_1 (Dense)	(None, 4096)	16,781,312
reshape (Reshape)	(None, 8, 8, 64)	0
conv2d_transpose (Conv2DTranspose)	(None, 16, 16, 64)	36,928
up_sampling2d (UpSampling2D)	(None, 32, 32, 64)	0
leaky_re_lu_2 (LeakyReLU)	(None, 32, 32, 64)	0
batch_normalization_2 (BatchNormalization)	(None, 32, 32, 64)	256
conv2d_transpose_1 (Conv2DTranspose)	(None, 64, 64, 128)	204,928
up_sampling2d_1 (UpSampling2D)	(None, 128, 128, 128)	0
leaky_re_lu_3 (LeakyReLU)	(None, 128, 128, 128)	0
batch_normalization_3 (BatchNormalization)	(None, 128, 128, 128)	512
conv2d_transpose_2 (Conv2DTranspose)	(None, 128, 128, 3)	3,459
activation (Activation)	(None, 128, 128, 3)	0

Layer (type)	Output Shape	Param #
input_layer (InputLayer)	(None, 128, 128, 3)	0
functional_1 (Functional)	(None, 4096)	16,865,600
functional_3 (Functional)	(None, 128, 128, 3)	17,027,395

Method - Audio

1) Extracting audio
2) Extracting MFCCs
3) Creating noise by setting sensitivity and epsilon [Sensitivity: 0.5 and Epsilon: 0.8]
4) Clustering the MFCCs using K-means algorithm [Experimenting with different no of clusters]
5) Reconstructing the audio by using centroids of the MFCCs
6) Training a Random Forest for emotion recognition
7) Speaker identification:
    a) Trained our own CNN model for speaker identification
    b) Applying a pre-trained model





Results

1) Original audio: Happy

2) Quantized audio

3) Original audio: Sad

4) Quantized audio

Results

Audio Emotion Recognition

Class	Precision	Recall	F1-Score
Angry	0.76	0.89	0.82
Happy	0.73	0.68	0.70
Sad	0.94	0.84	0.88

Video Emotion Recognition

Class	Precision	Recall	F1-Score
Angry	0.15	0.20	0.17
Happy	0.55	0.61	0.58
Sad	0.49	0.53	0.51

Late Fusion Combined Emotion Recognition Results (Using 60-40 Weighted Voting System)

Class	Precision	Recall	F1-Score
Angry	0.52	0.61	0.56
Happy	0.66	0.64	0.65
Sad	0.75	0.70	0.73

Results

- The Emotion Recognition Task tested with late fusion gave us an accuracy of 64.3% with the audio and video module giving 78% and 39.7% accuracy respectively
- Our video privacy model successfully hides the identity of the user with 95% accuracy
- Our audio privacy model successfully hides the identity of the user with a considerable accuracy

Takeaways

- Video privacy while maintaining emotional features is a more convoluted task than we initially estimated.
- Audio privacy seems to be a relatively less intricate task as evident from the results.
- 3d mesh generation is a highly computationally heavy task, which doesn’t justify the trade off for it be used in the pipeline.
- Privacy is a major concern with the increasing use of AI models in the present scenario and this domain requires further research

Contribution

Aabha Ranade - Conducted literature review for face reconstruction techniques like autoencoders and 3d mesh construction, Code for parsing videos to frames and corrupting the frames, autoencoder architecture, video emotion recognition

Ravi Vivek Agrawal - Conducted literature review for multimodal privacy preservation techniques, emotion recognition for videos model, combining the results for Emotion Recognition Task

Shivangi Kochrekar - Conducted literature review for video anonymization techniques, extracted the facial expressions, autoencoder architecture, making the video privacy,emotion and re-identification pipeline

Vaidehi Vatsaraj - Conducted literature review for audio and video speaker anonymization techniques, worked on anonymization methods, training an RF model for audio emotion recognition, testing for emotion and speaker recognition