ECE-685 Final Project S3

Learning Cross-Modal Representations Using Variational Autoencoders

Harshavardhan Srijay, Ramana Balla

November 2022

1 Introduction

Human learning in the real world naturally involves several modes of observation of the same underlying phenomenon. For example, understanding of an environment can be individually learnt through specific modalities, like sight, sound, touch, smell, etc. However, to fully understand and explain the world around us, we often resort to a combination of these individually imperfect senses, but that together, give us a significantly more complete representation of the world. Typically, machine learning models are implemented to retrieve and classify data belonging to a specific modality, like image, audio, text, etc. However, as introduced above, data and observations usually come in multiple modalities, meaning the ability to learn from multiple types of data can be extremely useful to better interpret and identify patterns in the real world.

Because of this, it is important to it is important to use a model which is able to jointly represent the information such that the model can capture the correlation structure between different modalities. For example, detecting a bird can be done through purely vision, purely sound, or more optimally, from both. Furthermore, a thorough understanding of an abstract notion of a bird by a machine must involve not only the individual characteristics of the visual, linguistic, and/or audio features, but the relationships between these features as well. Crucially, this understanding involves the flow of information from observation to representation, and vice versa. This form of generative learning, in which the goal is to learn relationships and correlations between modalities, ideally in an unsupervised manner, is commonly known as ”multi-modal” learning.

In this project, we aim to build multi-modal representations from audio and video data using self-supervised variational autoencoders (VAE), in order to separately learn efficient representations for audio and video data, and then learn a transformation from one modality to another. Specifically, we will be training these VAEs on the VoxCeleb1 dataset to attempt to learn the latent features corresponding to a celebrity's identity that are present in audio and visual modalities. Ideally, this can enable powerful features like cross-modal retrieval, i.e. generating the celebrity's face from their voice, or vice versa.

1.1 Motivation and Importance of Problem

Given that the world is often multi-model, the importance of developing effective machine learning models to learn from multiple data types, and capture non-obvious inter-modal features is obvious. This task however, is quite non-trivial, as multimodal learning typically requires 4 main components: latent factorisation, i.e. factorization of the joint latent space into the shared and private aspects of each modality, coherent joint generation, i.e. generations from different modalities are coherent in the joint latent space (represent the same celebrity), coherent cross generation, i.e. generate coherent data from one modality conditioned on another, and finally, synergy, i.e. generative quality improves with multiple modalities, as opposed to any single modality. Models with the above qualities can enable representation with the goal to learn computer interpretable descriptions of heterogenous data from multiple modalities, translation which represents the process of changing data from one modality to another, alignment where we want to identify relations between elements from two or more different modalities, fusion which represents the process of joining information from two or more modalities to perform a prediction task, and finally co-learning with the goal of transferring knowledge between modalities and their representations. Because of this, multi-modal and cross-modal generative models can prove to be quite useful in downstream tasks such as autonomous driving, healthcare, speech recognition, etc. The recent development of generative multmodal models like DALL-E and CLIP by OpenAI underscore the significant potential of the widespread application of multimodal and cross-modal generative models.

2 Related Works

Previous work has been done in generating join cross-modal and multi-modal embeddings. For example, in a study by Nagrani et al. jointly trained a subnetwork for faces and a subnetwork for voices (both unlabelled) to learn a joint embedding using the correlations between faces and voices of the same identity. This work was based on the psychological study of 'person identity nodes,' or PINS, that are a portion of associative memory holding identity-specific semantic codes that can be accessed via the face, the voice, or other modalities, meaning humans store identity information in a manner that is entirely abstracted from the input modality. This work illustrates the promise of cross-modal and mult-modal genreative models to exploit the redundant information between faces and voices, which forms the groundwork for our work here. Instead however, we train multi-modal VAEs, as opposed to simple networks as in the work of Nagrani et al., due to the ability of VAEs to learn a continuous generative distribution, thereby enabling powerful downstream generative tasks. Biometrics is another active area of research which tries to utilizes multiple modes for their recognition systems. The goal in these recognition systems is to essentially take advantage of the complementary signal components of the different modes (could be facial and speech or any other combination) and through feature fusion from the two modes, achieve a better performance than systems using a single modality for recognition. In our model, we aim to go one step further by taking advantage of the redundancy of the features common in both modalities and thus utilize cross-modal retrieval. Many other recent studies in the audio-visual cross-modal/multi-modal field largely focus on representation learning that incorporates information from both modalities in a discriminative way and/or generative way, which are then used for downstream tasks like sound source separation, sound source localization, cross-modal retrieval, etc. This work differs from this group of related work in that this work tackles the problem via generative models trained through self-supervised learning (label-free). Other VAE-based approaches seek to model the join posterior of multiple modalities as a product-of-experts over marginal posteriors. However, this work differs from other such VAE-based approaches as it is intended specifically to learn joint embeddings between audio and visual modalities with a shared-decoder architecture, which is specifically designed to address the large dimensionality difference between audio and video data.

3 Details of Project

In this project, we used the VoxCeleb1 dataset (https://mm.kaist.ac.kr/datasets/voxceleb/) to train a VAE-based architecture for multi-modal learning and cross-modal retrieval between video and audio modalities. The architecture is implemented with inspiration from https://arxiv.org/pdf/2102.03424.pdf. We chose the VoxCeleb1 dataset as it had less data (we faced memory/disk constraints throughout this project), and because it had a publically available list for train/test splits. From our preliminary research, there does not seem to be significant differences between VoxCeleb1 and VoxCeleb2 in terms of data quality for the purposes of this project. We downloaded the audio files, and the images corresponding to the video frames from this link. We split the data into a train and test set in accordance with the suggested split (class 1/2 is in the training set, class 3 is in the test set). For each celebrity in the dataset, there are several video clips, and for each video clip, there are several audio clips (.wav files) and cropped frames of the celebrity during the utterance of the audio clip. There can be multiple audio clips per video clip, but each audio clip corresponds to a set of visual frames. One challenge when handling this dataset is that each audio clip corresponds to a varying number of video frames, meaning each sample in the training and test set consisted of an audio clip, a tensor of each of the variable amount of video frames corresponding to the audio clip, and the label indicating the celebrity's identity. For each sample, we sampled a 15 ms from the audio clip and then segmented the clip into 2 to save disk space, removed the DC component of the signal, added a small dither, and then transformed the audio clip into a mean-var normalized spectrogram of size [297, 512]. We standardized and scaled the images to [3,23,23] in order to reduce disk space and computation time. This resulted in poorer resolution of the images, but from visual inspection, we felt there were still noticeable characteristics that can be used with the audio modality, even at this lower resolution. Then, in order to further reduce the dimensionality and improve disk space and runtime, we averaged the spectrogram across time, under the hypothesis that frequency differences relevant to person identification should generally be time-invariant, and averaged the video frames across the 3 RGB values, again hypothesizing that a grayscale image should be sufficient to learn the latent features of a person's identity. An example of a single training/testing sample is shown in Fig 1.

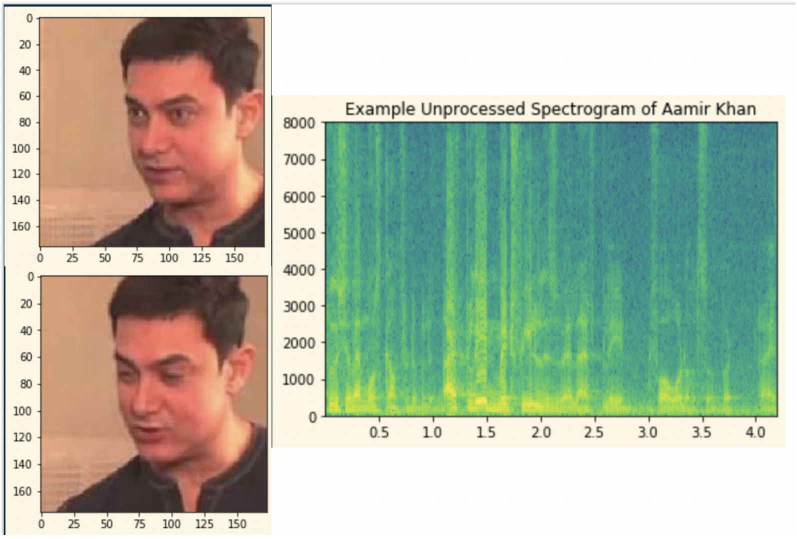


Figure 1: Left: two example frames of a video clip of Aamir Khan. Right: example unprocessed spectrogram used as input in downstream architecture

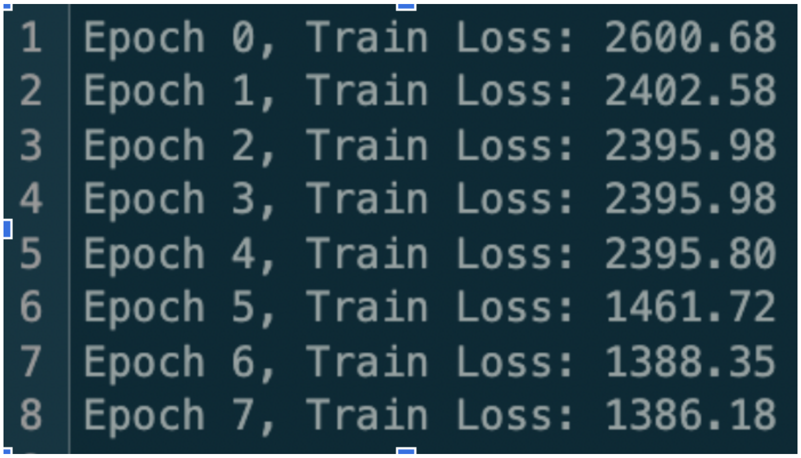
The model architecture consists of separate encoders that individually encode the audio and video modalities into separate subspaces, followed by a shared decoder that transforms one modality to the other. The shared decoder helps to enforce mutuality between the two modalities, in order to better capture the redundant information present in both that have a synergistic effect in their ability to learn human identity and enable downstream tasks like cross-modal retrieval.

We follow the traditional VAE formulation of attempting to derive the latent posterior, here using unlabeled, multi-modal input. We use a latent space dimension of 100. We derive an objective function using ELBO with a Wasserstein distance between the latent representations of audio and video, in order to ensure proper alignment between the two distinct subspaces of audio and video. Otherwise, the transformation from one modality to another is less robust. The loss is the sum of the standard reconstruction loss (MSE), to ensure strong generative reconstructions by the VAE, the Wasserstein latent loss to ensure proper alignment and transformation from one modality's subspace to another, and a KL divergence term, similar to c-VAE, to enforce normality between the prior and the variational distribution.

3.1 Contributions of each Team Member

4 Experimental Results

We train our model for only 8 epochs, due to size and complexity constraints, on the aforementioned training set of VoxCeleb, and use Adam optimization to minimize the above described loss between the two modalities. The training log for 8 epochs is shown in figure 2.



```
Epoch 0, Train Loss: 2600.68
Epoch 1, Train Loss: 2402.58
Epoch 2, Train Loss: 2395.98
Epoch 3, Train Loss: 2395.98
Epoch 4, Train Loss: 2395.80
Epoch 5, Train Loss: 1461.72
Epoch 6, Train Loss: 1388.35
Epoch 7, Train Loss: 1386.18
```

Figure 2: Training log of MS-VAE architecture for 8 epochs

5 Concluding Remarks

hyperparameter tuning (batch, lr, etc.), less signal reduction of audio and video, add augmentation to audio (rirs/musan), no channel avg pooling, time-averaging

6 References

1. https://arxiv.org/pdf/1911.03393.pdf
2. https://arxiv.org/pdf/2102.03424.pdf
3. https://mm.kaist.ac.kr/datasets/voxceleb/
4. https://www.robots.ox.ac.uk/ vgg/research/LearnablePins/
5. https://github.com/v-iashin/VoxCeleb

6. https://github.com/clovaai/voxcelebtrainer

7. Bruce, V., Young, A.: Understanding face recognition. British journal of psychology 77(3), 305–327 (1986)

8. Barnard, K., Duygulu, P., Forsyth, D., Freitas, N.d., Blei, D.M., Jordan, M.I.: Matching words and pictures. Journal of machine learning research 3(Feb), 1107–1135 (2003)

9. Nicolas Bonneel, Julien Rabin, Gabriel Peyre, and ´ Hanspeter Pfister, “Sliced and radon wasserstein barycenters of measures,” Journal of Mathematical Imaging and Vision, vol. 51, no. 1, 2015.

10. Diederik P Kingma and Max Welling, “Auto-encoding variational bayes,” in ICLR, 2014. 5. A. Nagrani*, S. Albanie*, A. Zisserman, Learnable PINs: Cross-Modal Embeddings for Person Identity, European Conference on Computer Vision, 2018

11. Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schoelkopf, “Wasserstein auto-encoders,” in ICLR, 2017.