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 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 sub-network 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.
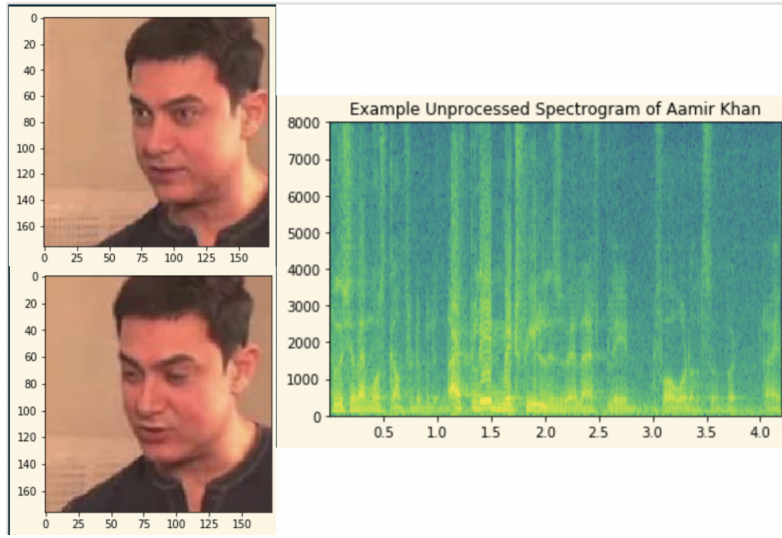

Figure 1: Left: two example frames of a video clip of Aamir Khan. Right: example unprocessed spectrogram used as input in downstream architecture

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 approximate the generative data distribution using a variational 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
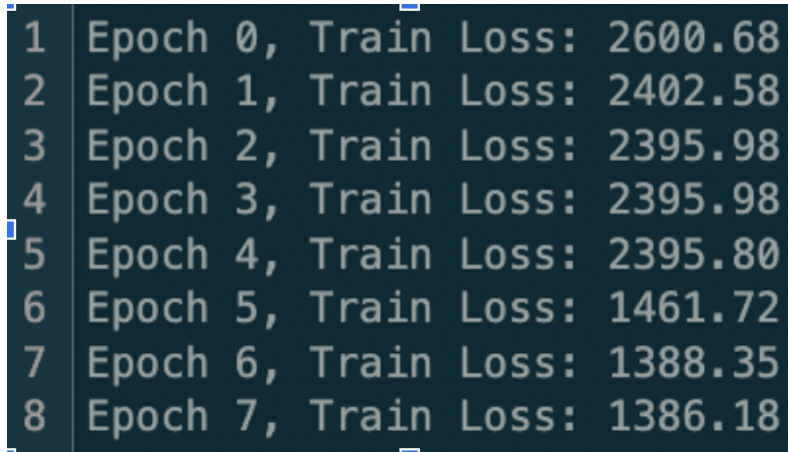
4

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

Our team comprised of two members, Ramana Balla and Harsha Srijay. Both of us spent time on our coming up with the best possible architecture for our model. Harsha implemented the model and the analysis and was able to successfully train the model on the VoxCeleb dataset and perform visualization of high dimensional manifolds in different sub-spaces learnt from the VAE, namely, the image, audio and joint. Ramana attempted to generate visualization from the teacher model implemented in the Cross-Modal Emotions paper. Unfortunately, even after several troubleshooting attempts, he was was not able to have the model up and running.

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 the Adam optimizer to minimize the above described loss between the two modalities. The training log for 8 epochs is shown in figure 2.


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

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

We then evaluate the model on the test set, to further analyze the results and conduct downstream tasks like cross-modal retrieval. Using the trained MS-VAE model, we visualize high dimensional manifolds of the audio, visual, and joint subspaces using t-SNE. The results of this analysis are shown in figure 3. These results were compared with that of figure 4 in the learnable PINS
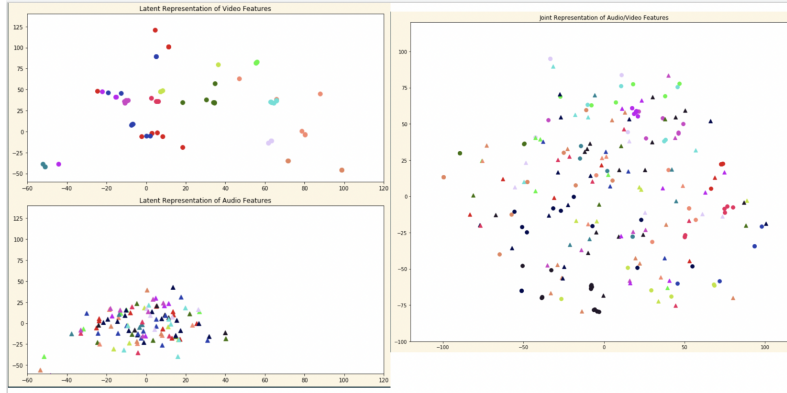
Figure 3: Left: Individual projections of audio and visual manifolds, Right: Joint manifold learned by VAE (circles represent video, triangles represent audio). Each color represents a different celebrity (Johnathon Schaech, Alexandra Roach, etc.) - the same celebrities shown in figure 4 of the learnable PINS paper

paper, to better visualize the coherence and robustness of the learned cross-modal and multi-modal representations of audio and video. We also used the train model to conduct cross-modal retrieval, in which we assess the strength of the model learnings by detecting, given a set of video frames corresponding to a specific celebrity (in this case, for demonstration purposes, we use Aamir Khan), retrieve the appropriate audio clips (different modality) corresponding to closest audio embeddings within the joint latent space. Ideally, these audio clips should correspond to Aamir Khan, but if not, they should at least be from a celebrity whose voice sounds like that of someone who looks like Aamir Khan. This would indicate a strong ability to learn correlations between modalities, a hallmark of good multi-modal and cross-modal learning. Similarly, given an audio clip of Aamir Khan, a successful multi-modal generative model should be able to retrieve the correct data from a separate modality (in this case visual video frames), that either are of Aamir Khan himself, or are of a celebrity who looks like someone that would have the voice of Aamir Khan. We calculated the MRR (mean reciprocol rank) for the closest audio clip of Aamir Khan in the joint subspace given his image, and vice versa. The results of this analysis are shown in figure 5. Finally, we conducted DBSCAN clustering on the joint representations learnt by the VAE, and found that there were no significnat clusters found in the joint manifold or the audio subspace, but there were in the video only subspace. When compared with the teacher model results on emotions, it is clear that the representations learnt in this implementation are not as discriminated between celebrity identities.

```
Robbie_Kay
Constance_Zimmer
Chris_Lowell
Scott_Porter
Seth_Rogen
5 celebrities most likely to look their voice sounds like Aamir Khan
Angela_Kinsey
Angela_Kinsey
Teri_Hatcher
Angela_Kinsey
Toby_Stephens
```

Figure 4: The first 5 names are the names of the celebrities with the lowest loss given Aamir Khan's visual frame. The next 5 names are the celebrities whose visual modality corresponds to the lowest loss in the joint latent space relative to one of Aamir Khan's audio clip.

5   Concluding Remarks

In general, the results of this project indicate a baseline attempt at using generative methods to learn redundant information between audio and visual modalities, and use this information to enable downstream tasks like cross-modal retrieval. Much of this work was hindered by disk and resource constraints, as we were only able to use 1 GPU for more than 30 GB of data, but the steps described above are likely the most important steps to implement for any multi-modal and cross-modal representation learning generative algorithm. This implementation was also motivated by existing research, indicating that the slightly poor results described here are due to resource limitations, not a lack of implementation. We found a decreasing loss on the training set during training, which was promising, but when visualizing the t-SNE manifolds, and comparing with that of the learnable PINS paper, it is clear the representations are not as robust. It is also clear from this visualization that the transformations conducted on the audio modality were not sufficient, as there was significantly less evidence of useful information in the audio modality, as there are no discernable clusters between celebrity identities, even when compared with that of the video modality. This could be one reason why the joint manifold did not cluster well either. Similarly, in the results for the cross-modal retrieval, we found that given visual information of Aamir Khan, while the closest identities to the video modality based on their audio clips in the joint subspace were not of Aamir Khan, they were mainly of male celebrities, with the exception of Constance Zimmer. This shows promise in the model, as it is still able to pick up that the visual features of Aamir Khan likely indicate lower-frequency audio clips, which are typically found in men. This indicates some useful latent features of identity being learnt by the model. Also, the poor results in the reverse direction of the cross-modal retrieval, i.e. identifying celebrities whose visual modality is closest to Aamir Khan's audio modality in the joint subspace, indicate the lack of information in the audio modality that was aforementioned, which again means future work must improve the representations in the audio modality. To improve the re-

sults, we must also consider improvements like hyperparameter tuning (batch size, learning rate, latent space dimensions, t-SNE/DBSCAN parameters, etc.) through model selection and the incorporation of a validation set, increased dimensionality of audio and video representations (we condensed video representations, for example, to only 23 x 23 x 3, to save space), adding augmentation to the audio modality through datasets like RIRS Noises or Musan, no time-averaging or channel-wise averaging for audio and video respectively.

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.