Generative Multimodal Learning for Reconstructing Missing Modality

Nishant Mishra(260903177)

Motivation and Introduction

Hypothesis: Training a latent variable based variational inference model on multimodal data, and using it to train similar models with a subset of the possible modalities in order to perform inference with all possible combinations of missing modalities provided as well as get a reconstruction of all modalities.

Solution inspired by the ***Multimodal Variational Autoencoder(MVAE)***[1] model

- We use an ELBO loss which is the combination of the individual reconstruction losses for each modality, and an additional label classification loss.
- The structure of the model follows a tree-structured graph where the different modalities define the observation nodes.
- The model follows a late fusion strategy where fusion is done by taking the product of experts as shown in Fig 1.
- Each modality has its own expert model as an inference network to contribute to z.

$$p(z|x_1,...,x_N) \propto \frac{\prod_{i=1}^{N} p(z|x_i)}{\prod_{i=1}^{N-1} p(z)} \approx \frac{\prod_{i=1}^{N} [\tilde{q}(z|x_i)p(z)]}{\prod_{i=1}^{N-1} p(z)} = p(z)\prod_{i=1}^{N} \tilde{q}(z|x_i).$$

- We use a subsampled setting for training where all (2^N-1) powerset combination is used for joint inference to calculate the ELBO loss with the reconstruction of the modalities.
- This way the model is generalized to perform well in reconstructing given any combination set of the modalities.

Dataset

- We used 3 modalities for our experiment.
- The datasets used for representing the 3 modalities are two MNIST datasets of images in different languages (Farsi and Kannada) and a spoken MNIST dataset consisting of a mixture of 6 speakers.
- Except for the speech data, the other datasets are sampled to form a triplet without replacement.
- The speech data is preprocessed to retrieve 13 dimensional MFCC features.

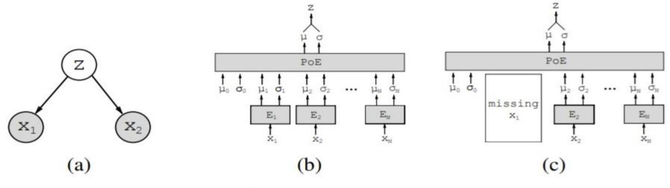


Figure 1 (a) Graphical model of the MVAE. (b) MVAE architecture with N modalities. μi and σi represent the i-th variational parameters; μ0 and σ0 represent the prior parameters. The product-of-experts (PoE) combines all variational parameters (c) If a modality is missing during training, we drop the respective inference network

Results and Method

Combination	Classification (Accuracy)(%)	ELBO	Reconstruction M1 (BCE)	Reconstruction M2 (BCE)	Reconstruction M3 (MSE)
m1	99.6	248.85	89.09	135	0.133
m2	99.7	436.9	348.67	68.96	0.134
m3	99.2	493.03	348.81	135.76	0.0095
m1, m2	99.93	187.64	89.05	69.04	0.133
m2, m3	99.94	427.44	346.19	69.1	0.011
m1, m3	99.88	239.11	89.33	134.8	0.013
m1, m2, m3	99.95	177.62	89.38	69.1	0.014

Table 1: Training performance at different combinations of the modalities and joint inference experts.
(BCE: Binary Cross Entropy; MSE: Mean Squared Error; ELBO: Evidence Lower Bound;
m1: MNIST Language 1 (Farsi); m2: MNIST Language 2 (Kannada); m3: Spoken MNIST (MFCC features))

- We used parallel linear models as the encoders and decoders for each modality respectively
- An additional decoder branch was added for label classification
- The loss function used were Binary Cross Entropy for the images, Mean squared Error for speech and Cross entropy loss for label classification
- These losses when added with KL divergence gave us the ELBO loss.

The ELBO losses for all 7 combinations of input modalities were added to get the training loss which was optimized..

$$\text{ELBO}(x_1,...,x_N) + \sum_{i=1}^{N}\text{ELBO}(x_i) + \sum_{j=1}^{k}\text{ELBO}(X_j)$$

The results are summarized in Table 1 and reconstruction outputs shown in Fig 2. Fig 4 shows the loss and accuracy curve

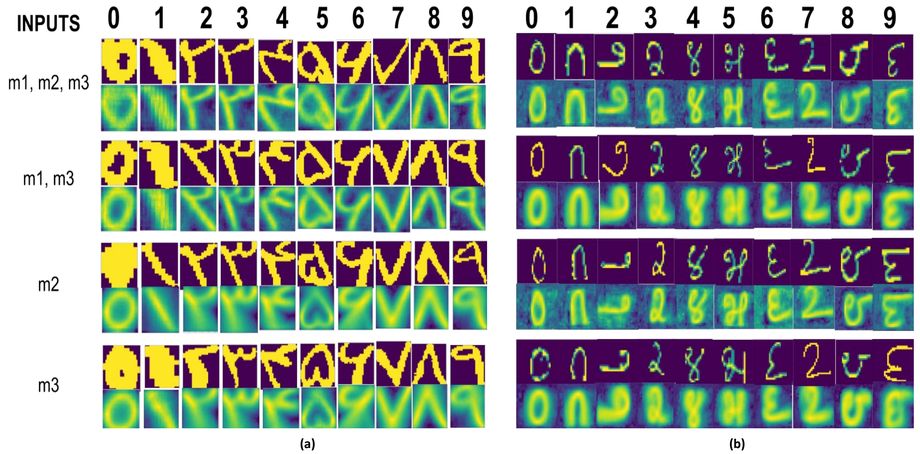


Figure 2: Original Image and Image reconstruction outputs in various combination of modalities. (a) Farsi MNIST reconstruction. (b) Kannada image Reconstruction.

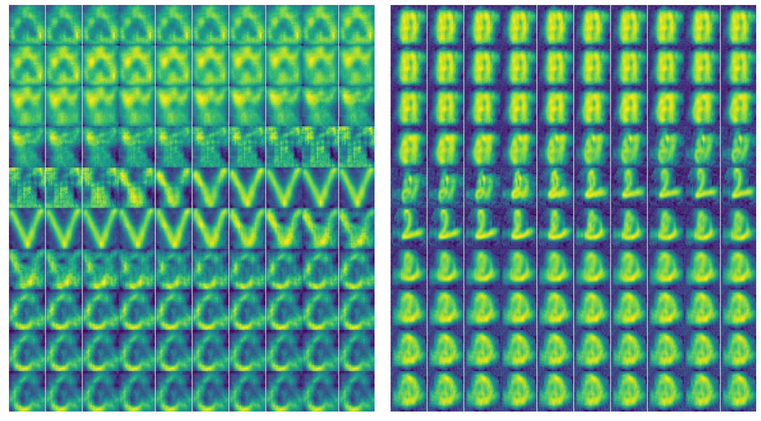
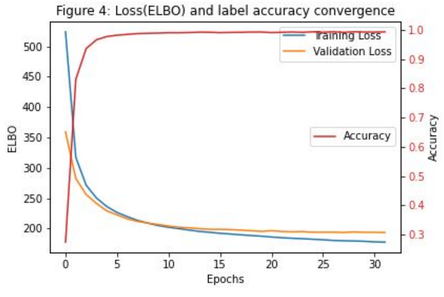


Figure 3: Output Reconstructions obtained by perturbing index 201 of latent variable by amounts -1500 to 1500 in intervals of 30 shown in 10x10 grid. Output reconstruction transitions as [5,3,7,0], original input label 7. (left: Farsi), (right :Kannada),

Conclusion

Disentanglement of Representation

We also studied the disentanglement property of the latent space representation. We perturbed the latent space at particular indices with varying noises, we observed a consistent pattern of variations in reconstruction output of 2 different reconstructions. We observe that the model learns disentangled representations some of which are modality agnostic as shown in Fig 3. From the results, we can conclude that the model learns a robust representation that helps in dealing with missing modality scenarios



Implementation Details

The code for the mentioned paper [1] was legacy. We took inspiration from their model definition and central idea, but rest of the implementation, including preprocessing, data loader, loss functions, training regime,experiments, and metric calculation was our own.

References

[1] Wu, Mike, and Noah Goodman. "**Multimodal generative models for scalable weakly-supervised learning.**" *Advances in Neural Information Processing Systems*. 2018.

[2] Zhi-Xuan, Tan, Harold Soh, and Desmond C. Ong. "**Factorized inference in Deep Markov Models for incomplete multimodal time series.**" AAAI, 2020.