**Generative Multimodal Learning for Reconstructing Missing Modality**

**Ankur Agarwal, Nishant Mishra (GROUP-7)**

**Introduction**

We are working on generative learning with variational inference to solve the multimodal learning problem. The goal is to be able to train and analyze a latent variable based inference model that is able to perform inference with all possible combinations of missing modalities provided. As such we train in a subsampled setting Multimodal Variational AutoEncoder ([Multimodal Generative Models for Scalable Weakly-Supervised Learning](https://arxiv.org/abs/1802.05335)) for our multimodal dataset. We use an ELBO loss which is the combination of the individual reconstruction losses for each modality, and an additional label classification loss. The results of our experiments are on the following page.

**Datasets and Model Architecture**

The structure of the model follows a tree-structured graph where the different modalities define the observation nodes. The datasets used for representing the 3 modalities are two MNIST datasets of different languages (Farsi and Kannada) and a spoken MNIST dataset consisting of a mixture of 6 speakers. The three modalities are sampled with the instance of a given digit. Except for the spoken data, the other datasets are sampled to form a triplet without replacement. The speech data is processed to retrieve 13 dimension MFCC features.

**Model Training Strategy**

The model follows a late fusion strategy where fusion is done by taking the product of experts. Each modality has its own expert model is an inference network to contribute to z. A (2N-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. A digit classification task is added. The model is trained at 128 batch size with a learning rate of 1e-3 for a latent dimension of 300. An 80:20 training-validation split is used on a sampling set of 60,000 unique triplets.

**Questions to the instructor:**

Which all analysis/experiments mentioned in the *“Next Steps”* section seem to be the most interesting/important analysis.

**Next Steps**

* We will further evaluate the individual performances and ELBO loss values on the validation and test set. We shall obtain and analyze the reconstruction of missing modality in inference mode when not using all the modalities. Additionally, we will perform ablation studies with hyperparameters like the latent size, coefficients of individual reconstruction losses. We will also experiment with other loss functions for our continuous-valued speech features apart from MSE loss which might correlate better with BCE loss(e.g bucketed softmax+BCE).
* We will analyze the individual modality reconstructions. We want to see if the model has learned a disentangled representation in the latent space and if so how does it affect the reconstruction of individual modalities in presence as combination and absence as individual inputs.
* An ambitious goal is to study how an input modality to the model improves the classification performance and how does the performance vary when the model is to explicitly learn about the missing modality via reconstruction from the latent created by the other modalities

**Results**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Combination** | **Classification (Accuracy)** | **ELBO** | **Reconstruction M1 (BCE)** | **Reconstruction M2 (BCE)** | **Reconstruction M3 (MSE)** |
| **m1** | 0.996 | 248.85 | 89.09 | 135 | 0.133 |
| **m2** | 0.997 | 436.9 | 348.67 | 68.96 | 0.134 |
| **m3** | 0.992 | 493.03 | 348.81 | 135.76 | 0.0095 |
| **m1, m2** | 0.9993 | 187.64 | 89.05 | 69.04 | 0.133 |
| **m2, m3** | 0.9994 | 427.44 | 346.19 | 69.1 | 0.011 |
| **m1, m3** | 0.9988 | 239.11 | 89.33 | 134.8 | 0.013 |
| **m1, m2, m3** | 0.9995 | 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))

