SUPPLEMENTARY MATERIAL

1 Visualization of Latent Representations

Since the main basis of building multimodal models is that they learn better representations compared to unimodal networks, therefore for above models it becomes imperative to check as if how similar the embeddings are when we have all modalities and when we have just one modality. Ideally, we will want the unimodal encoders to produce the same embeddings which the encoder which sees all the modalities does. Therefore, in this section, we show the tSNE plots of latent embeddings generated by all the models on our trimodal dataset, these embeddings basically are mean of the predicted posterior distribution. Note that, intuitively, we can assume that the more separated the embeddings are, the higher will be the accuracy in cross-modal generation.

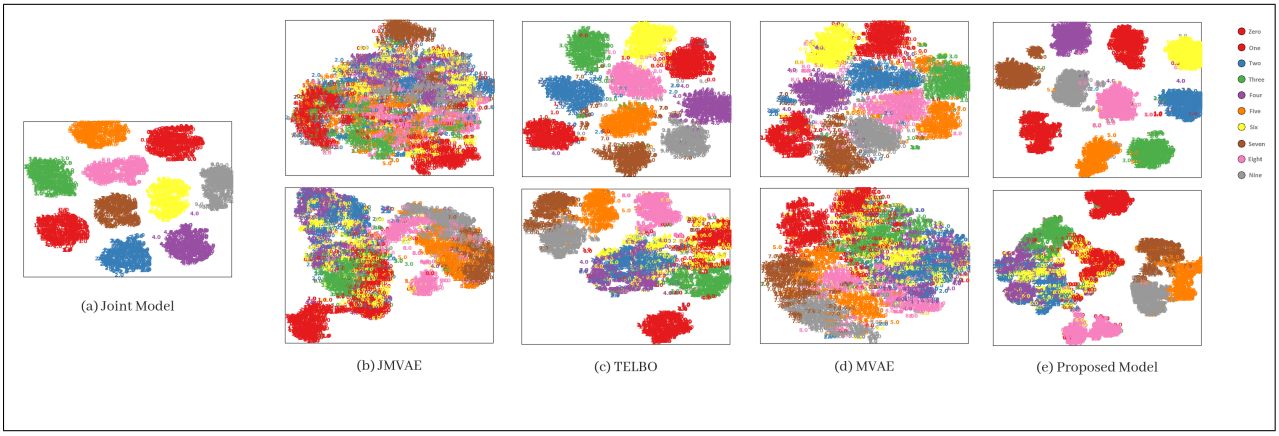

Figure 1: tSNE plots of mean of posterior distributions (a) Joint Multimodal VAE ($q_\phi(z|x, y, w)$) and, (b,c,d,e) shows $q_\phi(z|x)$ (upper figure) and $q_\phi(z|y)$ (lower figure) plots obtained using various models. (x and y denotes MNIST and Fashion-MNIST images respectively)

As expected, these means forms the clearly separated clusters (fig 1(a)) when all the three modalities are present because the model now have access to all the information that it needs. The distinction starts to appear when some modalities are missing in input.

For remaining models, this distribution is computed for two cases when only MNIST modality is available, and second when only images from Fashion-MNIST are available, in both the cases the other two modalities are missing from the input and not being shown to the model. We can see that for JMVAE model the clusters starts to merge together, thus making is decoders to err as we have seen in fig 6 and 7 (section 4.4 of the paper). The TELBO model performs better than the JMVAE, while the MVAE model performs worst among all primarily because the sub-sampling training approach deteriorates learning as we increase number of modalities in the input. Our proposed model, On the other hand, preserves the distinction of clusters much better than any of these models. This shows that our model achieves better transfer learning. Because, unlike retrofit models, our proposed model uses the pretrained encoder networks of joint model only, and does not train completely new set of encoders.

2 Adversarial training and Hyperparameter tuning

There are two possible ways we can reduce blurriness for a VAE: a) through adversarial training by using a discriminator network at the end of the network. As the focus was on a parameter efficient multimodal VAE and hence we had not considered this earlier. b) through hyperparameter tuning by varying the parameter β which controls how much weigh we want to give to KL-divergence between posterior and prior, in comparison to the reconstruction term.

$$\mathcal{L} = \quad \mathbb{E}_{q_\phi(z|x,y)}[\log p_{\theta_x}(x|z) + \log p_{\theta_y}(y|z)] - \beta\mathcal{D}(q_\phi(z|x,y)||p_\theta(z)) \tag{1}$$

. Results using both these approaches are shown below. For both approaches the given attribute was {Female, HeavyMakeup}, and network has to generate corresponding images.

- **Adversarial training**: Adversarial training though significantly improves the images quality but training becomes unstable, and this also makes model more susceptible to mode collapse.

Figure 2: CelebA results with adversarial training

- **Hyperparameter tuning**: We trained our model with various different values of β, which are 0.5, 1.0, 2.0 and 3.0. As we increase the β value both the diversity and image quality starts to improve. But as we increase this value even further the model starts generating incorrect images that does not match the query attribute. It shows that it is necessary to choose a appropriate value for β if our objective is cross-modal generation.

Figure 3: Training with $\beta = 0.5$

Figure 4: Training with $\beta = 1.0$

Figure 5: Training with β = 1.5

Figure 6: Training with β = 2.0