Energy-Based Multimodal VAEs

Anonymous Author(s)
Affiliation
Address
email

Abstract

Multimodal VAEs are a promising class of multimodal generative models that constructs a tractable posterior over the latent space given all modalities. Daunhawer et al. [2022] show that the generative quality of each modality drops as we increase the number of modalities. In this work, we take another direction to address the generative quality of multimodal VAEs by jointly modeling the latent space of unimodal VAEs using energy-based models (EBMs). The role of EBM is to enforce multimodal coherence by learning the correlation among the latent variables. Therefore, our model enjoys the high generative quality of unimodal VAEs while maintaining coherence across different modalities.

1 Introduction

The real-world data often has multiple modalities such as image, text, and audio, which makes learning from multiple modalities an important task. Recently, promising results have been achieved by multimodal generative models [Ramesh et al., 2021, Saharia et al., 2022]. However, these models are often only generative in one modality while conditioning on the rest. On the other hand, mutimodal VAEs are a class of multimodal generative models that are able to generate multiple modalities jointly. To train multimodal VAEs we have to construct a joint posterior over the latent space z: $q(z|\mathbb{X})$, where $\mathbb{X}$ is the set of modalities. To ensure the tractability of the inference network q, prior work has proposed using a product of experts ($q(z|\mathbb{X}) = \prod_i q(z|X_i)$)[Wu and Goodman, 2018], mixture of experts ($q(z|\mathbb{X}) = \sum_i q(z|X_i)$[Shi et al., 2019], or in the generalized form, mixture of the product of experts (MoPoE) [Sutter et al., 2021].

These models rely on modality subsampling during training to have a better performance on inference with missing modality at the test time. Subsampling of the modalities, as pointed out by Daunhawer et al. [2022], results in a generative discrepancy among modalities. We also observe that conditioning on more modalities often reduces the quality of the generated modality, which happens as a result of using the product of experts for combining the modalities. Product of experts constructs a sharper distribution by adding more components. The sharper the distribution is, the more confident it becomes on the agreeing mode (increases coherence). On the other hand, the resulting distribution becomes very picked and loses its generative quality.

To overcome these issues, instead of constructing a joint posterior, we try to explicitly model the joint latent space of individual VAEs: $p_\theta(z_1, z_2, \cdots, z_n)$. The joint latent model learns the correlation among the individual latent space without constructing a joint posterior for all modalities. Therefore, it can ensure prediction coherence while maintaining the generative quality. However, as expected, as we increase the number of modalities, the joint latent model becomes more complicated, which requires an appropriate factorization that is a subject of our future work. Nevertheless, conditioning on more modalities results in more accurate marginal distributions, thus increasing the generative quality.

Under review at the NeurIPS 2022 Workshop on Score-Based Methods. Do not distribute.

2 EB-MVAE

EBMs have been successfully used for modeling text [Deng et al., 2020] and image [Du and Mordatch, 2019, Song and Ermon, 2019] in the original data space. They also have been used to improve the performance of VAEs by modeling the latent space [Aneja et al., 2021, Pang et al., 2020]. In general, deep neural networks are effective in capturing the interaction of the variables, thus making the EBMs a successful model for joint modeling – EBMs parameterize the energy function of a Gibbs distribution over all variables using deep neural networks. We utilize this power to jointly model the latent space of different modalities: $p_\theta(z_1, z_2, \cdots, z_n) \propto \exp(E_\theta(z_1, z_2, \cdots, z_n)$. We cannot directly train the parameters θ using methods such as maximum likelihood, but several alternatives training algorithms have been proposed, including contrastive divergence [Hinton, 2002] and score-matching [Hyvärinen and Dayan, 2005]. In this work, we use score matching as we found it more stable and accurate for our setting. In score matching, we directly train the vector field, $S(\mathbf{z}) = -\nabla_\mathbf{z} E(\mathbf{z})$, by minimizing

$$\mathbb{E}_{p(\mathbf{x})}\mathbb{E}_{q(z_1|x_1)}\mathbb{E}_{q(z_2|x_2)}\cdots\mathbb{E}_{q(z_n|x_n)}\left[\mathrm{tr}(\nabla_\mathbf{z} S_\theta(\mathbf{z})) + \frac{1}{2}||S_\theta(\mathbf{z})||_2^2\right], \tag{1}$$

where $q(z_i|x_i)$ is the unimodal posterior over ith modality and is trained by optimizing the individual ELBO for that modality. We assume all of the modalities are present during training time for optimizing eq. 1 and we leave training with missing modality for future work. On inference time, any of the modalities can be missing.

Conditional generation: We assume at the inference time we have two groups of observed modalities (indexed by $\mathbf{o}$) and unobserved modalities (indexed by $\mathbf{u}$). We define the conditional posterior distribution for unobserved modalities as:

$$q(\mathbf{z_u}|\mathbf{z_o}, \mathbf{x_o}) = \left[\prod_{i\in\mathbf{o}} q(z_i|x_i)\right] p_\theta(\mathbf{z_u}|\mathbf{z_o}) \tag{2}$$

Sampling from $q(\mathbf{z_u}|\mathbf{z_o}, \mathbf{x_o})$ requires samples from unimodal posteriors of given modalities following by sampling from $p_\theta(\mathbf{z_u}|\mathbf{z_o})$. Knowing that $p_\theta(\mathbf{z_u}|\mathbf{z_o}) \propto \exp(-E_\theta(\mathbf{z_u}, \mathbf{z_o}))$, we sample from $p_\theta(\mathbf{z_u}|\mathbf{z_o})$ using Langevin dynamics [Welling and Teh, 2011]:

$$\mathbf{z}^{t+1} = \mathbf{z}^t - \frac{\lambda^2}{2}\nabla_\mathbf{z} E(\mathbf{z}^t, \mathbf{z_o}) + \lambda\mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{3}$$

3 Experiments

We compare EB-MVAE with different multimodal VAEs, including PoE [Wu and Goodman, 2018], MoE [Shi et al., 2019], and MoPoE [Sutter et al., 2021] using PolyMNIST dataset [Sutter et al., 2021]. This dataset consists of five different modalities created by changing the background images of an MNIST dataset. The encoder and decoder architecture of all methods are the same. We train the encoders and decoders using β-VAE [Higgins et al., 2016] with β-scheduling. We construct our energy-based model (EBM) by defining a multi-layer perceptron (MLP) over all five modalities. We assume all modalities are present during training. To train the EBM, we generate the samples from the posterior of each modality and minimize eq. 1.

Both EBM and VAEs are trained using Adam optimizer [Kingma and Ba, 2014] with a constant learning rate of 0.001. The VAEs are trained for 300 epochs with $\beta = 0.1$. We run Langevin dynamics for 40 steps to generate samples from the EBM.

We compare all methods on both prediction coherence and generative quality. We measure the coherence by evaluating the accuracy of the predicted modality based on the digits associated with the observed modalities. We also measure the generative quality of each modality using the FID score.

To evaluate our method, we first generate samples from the unconditional posterior for both EB-MVAE and MoPoE. For EB-MVAE since no modality has been observed, the posterior in eq. 2 becomes equal to the joint distribution over all unimodal latent space ($p_\theta(\mathbf{z_u})$). EB-MVAE has difficulty generating high quality images for modality 1 and modality 5. The main reason is that

(a) Dataset (b) MoPoE (c) EB-MVAE

Figure 1: a) Samples from training data. Each column belongs to one modality (from left to right we name it as modality 1 to 5). b) Unconditional samples from MoPoE (no modality is observed). Each column shows the samples for the corresponding modality. c) Unconditional samples from EB-MVAE.

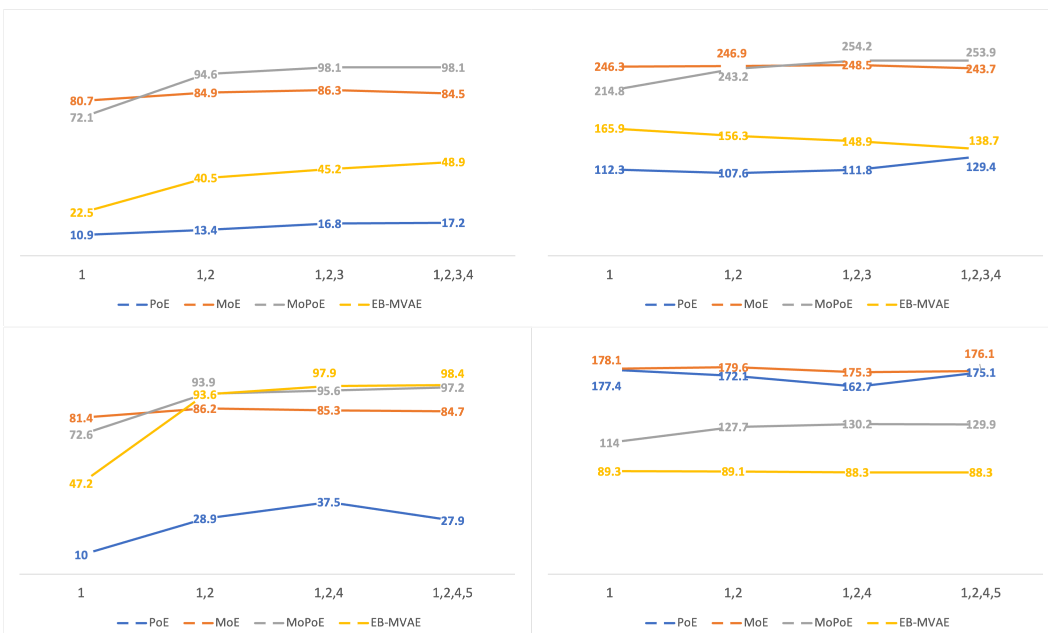
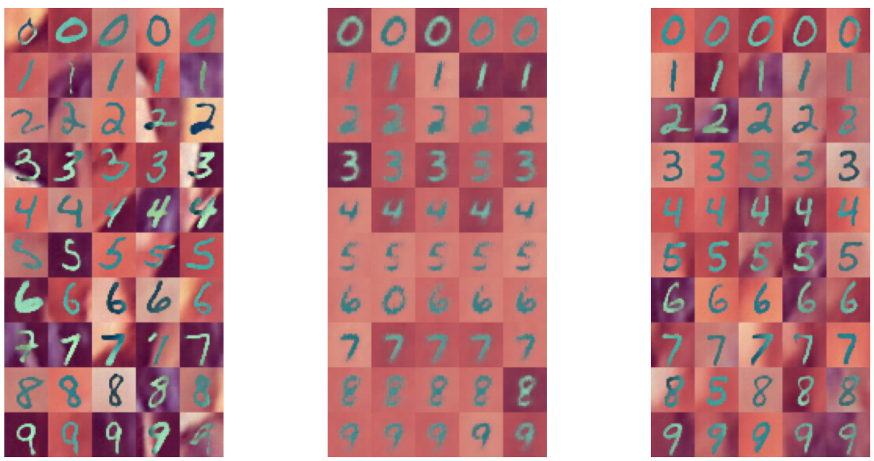


Figure 2: Left: Conditional coherence measured using prediction accuracy. Right: Conditional generative quality measured using FID score. The target modality in the first row is modality 5 and in the second row is modality 3.

fitting a joint model to data becomes more difficult as we increase the number of variables (modalities) and also the digits in these two modalities are more obscured by the background. The unimodal VAE tries to learn the background pattern as well as the digits and that propagates to the joint EBM model. MoPoE, on the other hand, tries to learn a common latent space for all modalities, thus emphasizes more on the common digit rather than the background information.

However, we still can expect that we get better conditional performance as we observe more modalities. To confirm this, we increase the number of observed modalities from 1 to 4 and report the accuracy and FID score for modality 3 and 5 in Figure 2. For a multimodal generative model, as we condition on more modalities, we expect improvement in both prediction accuracy (coherent cross generation) and generative quality (synergy) [Shi et al., 2019]. Among PoE, MoE, and MoPoE, only MoE loosely follow the expected patterns, while MoPoE and PoE only respect coherent cross generation pattern and violates expected synergy pattern. EB-MVAE, on the other hand, shows better accuracy as we conditioned on more modality and at the same time its generative quality improves. This behavior is describable via its joint latent model. As we condition on more modality the marginal distribution gets closer to the target unimodal distribution. It is worth noting that the PoE, MoE, and MoPoE either have high quality generative capability (PoE) or high coherence (MoE and MoPoE), while EB-MVAE has no fundamental limitation (because of its joint modeling of the latent space of individual modalities) to have both properties. For predicting modality 3 given the rest of modalities, EB-MVAE has the best accuracy and generative quality among the methods.

We also qualitatively compare the conditional posterior of modality 3 given the rest of the modalities for EB-MVAE and MoPoE. In Figure 3 we draw one generated output using one sample from $q(z_3|z_1, z_2, z_4, z_5, x_1, x_2, x_4, x_5)$ for five different assignments to x_1, x_2, x_4, x_5 (that has the same digits) at each row. We also show the generated samples using unimodal posterior $q(z|x_i)$ for different data points with the same digits (each row). EB-MVAE samples have more variety than MoPoE samples and better capture the background, and the generative quality of samples is closer to those of unimodal VAE. This is evidence that the common latent space is more restricted than the joint model of unimodal latent spaces, which results in lower generative quality.



(a) Unimodal-VAE (b) MoPoE (c) EB-MVAE

Figure 3: Samples from predicted modality 3. a) Sample from the unimodal VAE using posterior distribution. b) Samples from EB-MVAE conditioned on modalities 1,2,4,5. c) Samples from MoPoE conditioned on modalities 1,2,4,5.

4 Conclusion

Multimodal VAEs are an important tool for modeling multimodal data. In this paper, we provide a different multimodal posterior using energy-based models. Our proposed method learns the correlation of latent spaces of unimodal VAEs using a joint model in contrast to the traditional multimodal VAE construction that learns a common latent space for all modalities. We show that our method (EB-MVAE) can generate high quality and coherent samples.

References

Jyoti Aneja, Alex Schwing, Jan Kautz, and Arash Vahdat. {NCP}-{vae}: Variational autoencoders with noise contrastive priors, 2021. URL `https://openreview.net/forum?id=c1xAGI3nYST`.

Imant Daunhawer, Thomas M. Sutter, Kieran Chin-Cheong, Emanuele Palumbo, and Julia E Vogt. On the limitations of multimodal VAEs. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=w-CPUXXrAj`.

Yuntian Deng, Anton Bakhtin, Myle Ott, Arthur Szlam, and Marc'Aurelio Ranzato. Residual energy-based models for text generation. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=B1l4SgHKDH.

Yilun Du and Igor Mordatch. Implicit generation and modeling with energy based models. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/378a063b8fdb1db941e34f4bde584c7d-Paper.pdf.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. 2016.

Geoffrey E Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, 2002.

Aapo Hyvärinen and Peter Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014. URL https://arxiv.org/abs/1412.6980.

Bo Pang, Tian Han, Erik Nijkamp, Song-Chun Zhu, and Ying Nian Wu. Learning latent space energy-based prior model, 2020. URL https://arxiv.org/abs/2006.08205.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation, 2021. URL https://arxiv.org/abs/2102.12092.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding, 2022. URL https://arxiv.org/abs/2205.11487.

Yuge Shi, Brooks Paige, Philip Torr, et al. Variational mixture-of-experts autoencoders for multi-modal deep generative models. *Advances in Neural Information Processing Systems*, 32, 2019.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution, 2019.

Thomas M. Sutter, Imant Daunhawer, and Julia E. Vogt. Generalized multimodal elbo, 2021. URL https://openreview.net/pdf?id=5Y21V0RDBV.

Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688. Citeseer, 2011.

Mike Wu and Noah Goodman. Multimodal generative models for scalable weakly-supervised learning. *Advances in Neural Information Processing Systems*, 31, 2018.