DEEP DIRECTED GENERATIVE MODELS WITH ENERGY-BASED PROBABILITY ESTIMATION

Taesup Kim, Yoshua Bengio*
Institut des Algorithmes d'Apprentissage de Montréal
Université de Montréal
Montréal, QC, H3T 1J4, Canada
{taesup.kim,yoshua.bengio}@umontreal.ca

ABSTRACT

Energy-based probabilistic models have been confronted with intractable computations during the learning that requires to have appropriate samples drawn from the estimated probability distribution. It can be approximately achieved by a Monte Carlo Markov Chain sampling process, but still has mixing problems especially with deep models that slow the learning. We introduce an auxiliary deep model that deterministically generates samples based on the estimated distribution, and this makes the learning easier without any high cost sampling process. As a result, we propose a new framework to train the energy-based probabilistic models with two separate deep feed-forward models. The one is only to estimate the energy function, and the another is to deterministically generate samples based on it. Consequently, we can estimate the probability distribution and its corresponding deterministic generator with deep models.

1 INTRODUCTION

Energy-based models have been used to capture dependencies over variables by defining an energy function. The energy function associates each configuration of the variables with a scalar energy value. Lower energy values should be assigned to more likely or plausible configurations and conversely higher values to others. This has been used for example to estimate the probability distribution based on a Boltzmann distribution defined by an energy function and appropriate normalization factor. In this case, the energy function is defined to assign a probability value that is not normalized. The normalization factor plays an important factor that constrains the energy function to properly estimate the probability distribution. However, it introduces difficulties during the learning procedure that requires a number of samples appropriately drawn from the estimated probability distribution. This is typically computationally intractable and makes the learning progress slow and noisy or requires certain model structures to get samples. Especially, a bipartite undirected graph structure with stochastic hidden variables, such as restricted Boltzmann machines(Hinton, 2012), is possible to approximately draw samples through Monte Carlo Markov chain (MCMC) methods, although as the model becomes sharper (with more energy differences between nearby configurations), mixing between modes becomes difficult for MCMC methods. We explore an approach that circumvents this problem by introducing an auxiliary generative model, which is a deep directed generative deep model to that trained with generative adversarial networks (Goodfellow et al., 2014), to efficiently draw samples from the estimated probability distribution without any Markov chain.

2 THE PROPOSED MODEL

The Boltzmann distribution is a probability distribution $P_\Theta(x) = \frac{1}{Z_\Theta} e^{-E_\Theta(x)}$ based on the energy function $E_\Theta(x)$ with trainable parameter set Θ, and it can be used to define energy-based probabilistic models. Moreover, it simply can be extended to the product of experts model (PoE), which formulates the energy function as a sum of experts $E_\Theta(x) = \sum_i \tilde{E}_{\theta_i}(x)$. In this paper, we only

*CIFAR Senior Fellow

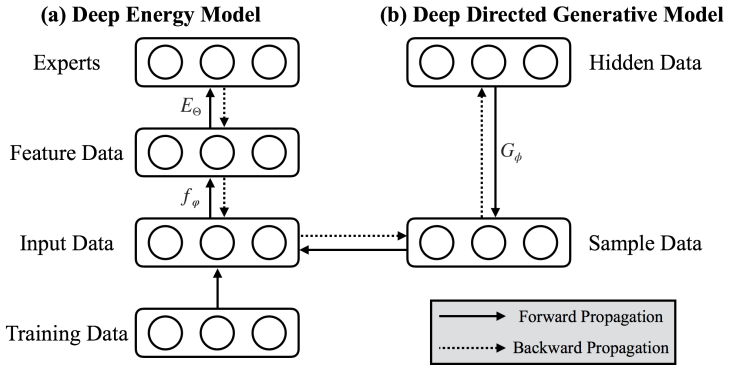

Figure 1: The proposed model has two separate deep models, (a) deep energy model is defined to estimate the probability distribution by learning the energy function based on the feature data, and (b) deep directed generative model is a deterministic generator that generates samples based the deep energy model. Like in GAN, the input into the energy model either comes from the training data or from the deep directed generative model.

consider each expert based on a logistic regression model to detect or penalize certain patterns, and it is exactly derived from the form found in restricted Boltzmann machines, except that the experts depend on non-linear features extracted from the raw data, following the idea presented by (Ngiam et al., 2011).

The energy-based probabilistic model is trained to estimate the data distribution $P_{\mathcal{D}}(x)$ by fitting to it the model distribution $P_\Theta(x)$, which is defined by the energy function. This is usually done by minimizing the Kullback-Leibler divergence between two distributions $\mathrm{D}_{KL}\big(P_{\mathcal{D}}(x)||P_\Theta(x)\big)$, and it is exactly same as to minimize the expected negative log-likelihood under the data distribution. This can done by using the gradient descent method, and the gradient with respect to the model distribution parameters Θ is given by,

$$\nabla_\Theta \mathcal{L} = \underbrace{\mathrm{E}_{P_{\mathcal{D}}(x)}\left[\frac{\partial E_\Theta(x)}{\partial \Theta}\right]}_{\text{Positive Phase}} - \underbrace{\mathrm{E}_{P_\Theta(\tilde{x})}\left[\frac{\partial E_\Theta(\tilde{x})}{\partial \Theta}\right]}_{\text{Negative Phase}} \tag{1}$$

This shows an interesting learning rule with two different terms, which are referred to as positive and negative phase respectively. However, the negative phase is typically intractable because it is difficult to get samples from the model distribution . It typically requires sampling methods such as Monte Carlo Markov chains, making the learning slow and noisy, and possibly limiting its ability to learn sharp distributions, as argued by Bengio et al. (2013).

To overcome this problem, we introduce a new framework that has two separate deep models and is motivated from the generative adversarial networks(Goodfellow et al., 2014), with feed-forward neural networks as depicted in Figure 1. The first deep model is only used to define the energy-function, as a sum of terms each of which is associated with an expert $\tilde{E}_{\theta_I}$ and deep feature extractor f_φ,

$$E_\Theta(x) = \sum_i \tilde{E}_{\theta_I}(f_\varphi(x)) = -\sum_i \textit{softplus}\big(w_i^T f_\varphi(x) + b_i\big) \tag{2}$$

and we call this neural network the *deep energy model*, which defines the model distribution P_Θ. Next, a second deep neural network deterministically generates samples from given latent variables z that are randomly sampled from the uniform distribution U.

$$\tilde{x} = G_\phi(z) \qquad \text{where } \forall i, z_i \sim U\big[-c, c\big] \tag{3}$$

and we call this model G_ϕ the *deep directed generative model*. We also assume this model has an underlying distribution P_ϕ that we would like to be as aligned as possible to the model distribution P_Θ. In this framework, we expect the energy and the generative model to learn from each other to be aligned, approximately forming two sides of the same coin.

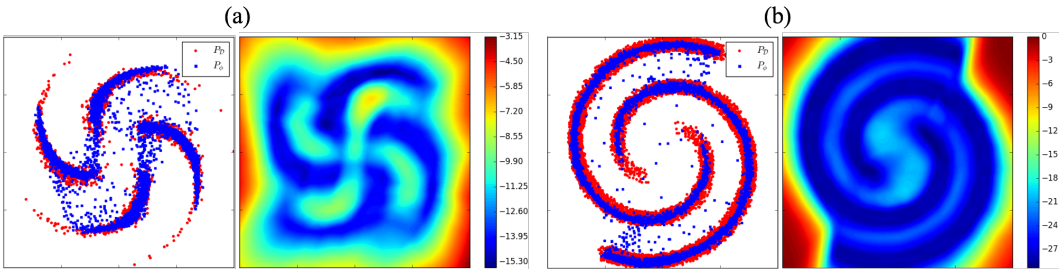

Figure 2: Results on (a) four-spin dataset and (b) two-spiral dataset. Left : Samples from the training dataset (red) and the generative model (blue). Right : The estimated energy function, with blue indicating low energy and red high energy.

We now reformulate the original problem by modifying the negative phase in Equation 1 by introducing two deep models E_Θ and G_ϕ. First, we assume that the model and generator distributions are approximately aligned $P_\theta(x) \approx P_\phi(x)$, and use samples generated from the generative model for the negative phase.

$$\mathrm{E}_{P_\Theta(\tilde{x})}\left[\frac{\partial E_\Theta(\tilde{x})}{\partial \Theta}\right] \approx \mathrm{E}_{P_\phi(\tilde{x})}\left[\frac{\partial E_\Theta(\tilde{x})}{\partial \Theta}\right] = \mathrm{E}_{U(z)}\left[\frac{\partial E_\Theta(G_\phi(z))}{\partial \Theta}\right]. \tag{4}$$

Then, the modified learning rule can be viewed as training a classifier based on E_Θ that discriminates between data from the training dataset $\mathcal{D}'$ and samples from the generative model G_ϕ(Welling et al., 2002; Bengio, 2009). Next, we keep aligning the energy and generative model by minimizing the expected negative log-likelihood over samples generated from the generative model, and the corresponding gradient is as,

$$\nabla_\phi \mathcal{L}' = \mathrm{E}_{U(z)}\left[\frac{\partial E_\Theta(G_\phi(z))}{\partial \phi}\right] \tag{5}$$

With our proposed framework, we can obtain at the end a deterministic generator to efficiently draw samples based on the estimated data distribution that is represented by our trained energy function.

3 Experiments

We experimented our proposed model with 2D-synthetic dataset to show that the deep energy and generative models are properly learned and aligned to each other. We generated two types of datasets, four-spin and two-spiral, respectively. Each dataset has 10,000 points randomly generated from a fixed distribution. For simplicity, we set the number of experts as the dimension of hidden data, and used the same structure for both models but in reverse order. We used *AdaGrad* (Duchi et al., 2011) optimizer with learning rate 0.01. In Figure 2, we visualized our results, which are generated samples from the generative model and the energy function surface, and it can be observed that the generator properly draws samples as the data distribution based on the energy function.

4 Conclusion

The energy-based probabilistic models have been broadly used to define generative processes with estimating the probability distribution. In this paper, we showed that the intractability can be avoided by using two separate deep models. In future work, we expect to extend this work with more complex and high-dimensional data to efficiently generate samples, and also approximately visualize the energy function in a low-dimensional space.

REFERENCES

Yoshua Bengio. Learning deep architectures for ai. *Found. Trends Mach. Learn.*, 2(1):1–127, January 2009. ISSN 1935-8237.

Yoshua Bengio, Grégoire Mesnil, Yann Dauphin, and Salah Rifai. Better mixing via deep representations. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, pp. 552–560, 2013.

John C. Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159, 2011.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pp. 2672–2680, 2014.

Geoffrey E. Hinton. A practical guide to training restricted boltzmann machines. In *Neural Networks: Tricks of the Trade - Second Edition*, pp. 599–619. 2012.

Jiquan Ngiam, Zhenghao Chen, Pang Wei Koh, and Andrew Y. Ng. Learning deep energy models. In *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*, pp. 1105–1112, 2011.

Max Welling, Richard S. Zemel, and Geoffrey E. Hinton. Self supervised boosting. In *Advances in Neural Information Processing Systems 15: Neural Information Processing Systems, NIPS 2002, December 9-14, 2002, Vancouver, British Columbia, Canada*, pp. 665–672, 2002.