RESEARCH PROPOSAL
the Fundamental Limitations for Deep Generative Model

Electronic Engineering ***Wenhao Ding***

1 Introduction

Supervised learning has achieved tremendous breakthrough in recent years. Although exceeding human performance in many fields, it still suffers from the problem of requiring enormous data and weak generalization. As a contrast, unsupervised learning has more potential and capability to extract internal and high-level representations. Deep generative model, an important kind of algorithm of unsupervised learning, is considered a promising method to model the data distribution. It is able to represent the data in a latent space and also sample new data from the latent space.

Several successful deep generative models have been widely applied recently. These models include Deep Boltzmann Machine which is based on thermodynamic principles, the directed graph model Deep Belief Nets, and the Autoregressive Model. However, these obsolete technologies either are difficult to train or demand careful structural design, forcing researchers to find alternative models. Fortunately, there are four alternative frameworks having better performance nowadays. The first is Variational Autoencoder (VAE), which is attributed to approximate likelihood method. Next is an implicit likelihood model named Generative Adversarial Networks (GAN). The last two methods are categorized in directly calculation of likelihood: PixelRNN (belongs to Fully Visible Belief Nets) [16] and Glow (belongs to Change of Variables) [8].

Since the likelihood of VAE is intractable, variational inference is implemented, as represented in (2), to get an Evidence Lower Bound (ELBO). This approximate results contains two parts: the first one could be interpreted as a reconstruction error and the second one is a tightness condition to ensue $p_\theta(z|x) = q_\phi(z|x)$ (achieved by Kullback–Leibler divergence).

$$logp_\theta(x) = \int_z q_\phi(z|x) log \frac{p_\theta(x,z)}{q_\phi(z|x)} \geq \mathbb{E}_{q_\phi(z|x)}[logp_\theta(x|z)] - D_{KL}(q_\phi(z|x)||p(z)) \tag{1}$$

GAN, proposed by Ian Goodfellow in 2014, is based on game theory. To avoid the explicit computation of likelihood, the authors build a discriminator to automatically learn it. After reaching a balance point between the generator and the discriminator, the generator will have the ability of describing the distribution of real data. Briefly, the optimization function is shown in (3), in which two terms represent the expectation of the real data and fake data respectively.

$$\min_G \max_D V(G,D) = \mathbb{E}_{x\sim p_{data}(x)}[logD(x)] + \mathbb{E}_{z\sim p_z(z)}[log(1-D(G(z)))] \tag{2}$$

It is more ingenious for methods of directly calculating likelihood. The principle of PixelRNN [16] is quite intuitive. It splits the joint distribution into the product of several conditional distributions with the chain rule. For instance, a whole image could be generated by one pixel after another. Another framework, Glow [8], belongs to the method of *change of variables*. Unlike most

models that regard data X and hidden space Z as a joint distribution, Glow regards X and Z as two reversible differential manifolds. Then with the property of diffeomorphisni, the parameters of the model given a datapoint can be written as (3).

$$\theta^* = \underset{\theta}{argmax}\, logp_x(X;\theta) = \underset{\phi,\psi}{argmax} \sum_{n=1}^{N} logp_z(f(x_n;\phi);\psi) + log\left|\frac{\partial f_\phi}{\partial x_n}\right| \tag{3}$$

Although above mentioned methods have greatly promoted the development of generative models, there are still fundamental limitations remain to be solved. Breaking through these restrictions has enormous effects on improving the performance of most machine learning algorithms. For example, Vapnik's Empirical Risk Minimization (ERM) theory [17, 18] proves that the minimization of errors on training sets gives the model the generalization ability. However, most failure cases fall into the dilemma that remembering all training data promises low training loss. Thus, augmenting the dataset with generative models, to a certain extent, weakens the model's ability of memorizing training samples. Experimental results of [21, 19] also support this conclusion.

Another example is transfer learning. If it is possible to decompose the feature into independent and interpretable feature, we can separate the content and the style. When data with different styles but the same content is provided, we can consider only the content subspace of the manifold to rapidly implement the transfer learning.

Lastly, it is always an important demand to evaluate the problem-solving capability of an invented model, or to evaluate the confidence of the model's result. As discussed in [11], if we have a reasonable description of data space (external resources) and a clear objective of our problem (target function), we can use statistical methods to establish the correlation between them. Then we are able to judge the ability of algorithms.

2 Limitations

(1) Lack of suitable metric to measure the distance. How to describe the distance between the model and the real data distribution remains unsolved. Directly optimizing likelihood may result in weak correlation between the likelihood and the quality of sample [3][15]. Specifically, the condition of ELBO in VAE is often difficult to satisfy, and the asymmetry of KL divergence and the prior assumption of Gaussian also lead to poor quality of generated samples. In addition, models that straightly calculates likelihood, such as Glow and PixelRNN, suffer from the problem that high likelihood is not equal to being similar distribution. Models may provide larger likelihood for out-of-training distribution with small variance [12]. On the other hand, GAN, the representative of indirect likelihood-based methods, optimizes discriminative metric and generator simultaneously. This will lead to an extremely unstable training process (problem like mode collapse frequently occurs).

(2) Lack of interpretability for latent codes. In order to facilitate the generative model with a better understanding of the data distribution, we hope not only to reconstruct the real samples from the low-dimensional manifold, but also to give interpretability to each dimension of it. GANs only reconstruct samples from random noise [4], and outputs random results. The output samples of conditional GAN [10] are controlled by external supervised label. InfoGAN [2] is essentially an external condition model. Although label is not required, its latent codes are often entangled. Though VAE obtains latent codes that can be continuously sampled, these codes, as same as infoGAN, are often related to entangled features [7]. Additionally, the lack of interpretability also

comes from the mismatching between dimension of manifold and dimension of latent codes. For example, the noise's dimension in GAN, latent code's dimension of infoGAN and VAE are all arbitrary selected. This dimension mismatching makes it difficult to train the model. An intuitive example is the impossibility of fitting a curve to a plane.

(3) Lack of suitable evaluation to compare models. Because of the lack of evaluation or the unreasonable evaluation, it is impossible to compare the performance of different models from multiple perspectives and to judge what progress the new model has made. At present, the standard metric for GAN is to compare the image's quality, using FID [5] or IS [14]. However, these methods only consider the fidelity of images, ignoring the fact that great samples and poor likelihood may occur simultaneously [15]. As for VAE, there have been attempts to analyze the latent space in [6] and [7], but no comprehensive evaluation method is proposed.

3 Technical Description and Tasks

The limitations described in previous section do not exist independently. The solution to one problem might provide insights to another. Therefore, I intend to push forward every limitation and finally establish a complete and self-consistent theoretical framework. Applications could be improved by this framework as well. To that end, I propose three kinds of solutions corresponding to the aforementioned limitations.

(1) Finding stable and precise metrics. Firstly, it is worthwhile to find other approximative approaches for intractable likelihood. Since probability density is intractable in most cases, better approximation methods, such as tightness conditions that are easier to satisfy than VAE, will be widely utilized. On the other hand, inspired by current non-likelihood-based methods, we can also convert the likelihood into implicit expression, or boldly jump out of the traditional MLE framework of Bayesian. For example, GAN, a successful attempt beyond the traditional frameworks, achieves the best generative results at present [1]. Works like Wasserstein distance [13] and Maximum Mean Discrepancy (MMD) [20] also provide some guidance in this direction.

(2) Finding strategies of decomposing feature. In order to have interpretable latent codes, a key point is to force them to be independent. In other words, the model should have the ability to extract independent features. An intuitive implementation is to convert the latent space with orthogonal basis. The latent codes could be separated into independent components. Essentially, these methods are highly correlated with compressed sensing, sparse coding and dictionary learning, thus some mature schemes could be used for reference. In addition, we often have a priori knowledge in most tasks. These priori knowledge can not only be used in supervised learning, but also play a guiding role in decoupling entangled feature. Another important point is the separation of discrete and continuous features. Due to the limitation of KL divergence, current VAE framework assumes the prior distribution of latent codes is Gaussian. However, lots of content information is defined by human and is discrete. GAN is difficult to sample in a discrete space as well.

(3) Finding evaluation based on statistical method. Such comprehensive evaluations should be considered in the following aspects: **(a) STABILITY:** Stability has always been an important indicator of a system. In terms of generative model, we can analyze the stability from the variance of its output. **(b) DEPENDENCE:** This point corresponds to the evaluation of the decomposition ability. Methods calculating the correlation between different latent codes are highly demanded to judge whether the model obtains the independence or not. **(c) RELIABILITY:** The neural network, which is not different from other modules, should be regarded as only a part of a complete

system. In practice, we often encounter the phenomenon that neural network outputs wrong results with high confidence [9]. This is harmful for the whole system. Therefore, we need to determine how much we can trust the a neural network.

4 Timeline

The whole timeline is divided into two stages in general. In the first stage, I plan to solve the problem of interpretability for latent codes, and try to establish some evaluation metrics. The plan of second stage is to improve the distance metric between distributions, in which more statistical knowledge is required.

Stage 1 ($12 \sim 18$ months): Since the goal of decomposition and interpretability of features is very clear, it is possible to achieve some results in the short term. At the same time, in order to compare the new invented models with the old ones, new evaluations will be proposed while designing the new algorithm.

Stage 2 ($24 \sim 36$ months): First of all, I need to learn more about classical technologies and related mathematical tools. Then, on the basis of this, I can explore better methods of similarity measurement.

5 Social Impact

The establishment of a complete framework for generative models is not only of academic significance, but also of technological innovation in many areas of society and industry.

(1) Semi-supervised learning and one-short learning. In industry, the amount of data is limited or the labeling work is time-consuming. If the target data space could be well described, the performance of the algorithm could be improved by using data augmentation and feature separation.

(2) Super resolution of audio and image, text to speech and speech to text It is quite intuitive to use conditional generative models for the task of image restoration, audio de-noising, and image de-blurring. Different resolutions belong to different data domains, and generative model plays the role as a bridge.

(3) Rare data generation. Merging two extreme data distributions to an intermediate dataset can be applied to scenarios where the intermediate data is rare. For example, the dangerous vehicle intersection data in real traffic.

(4) Abnormal sample detection. First, a generative model for normal data should be constructed. Then, with the normal data distribution , anomalous sample will have different output.

(5) Other disciplines. Disciplines like Material Science, Medical Science and Chemistry are all able to benefit from the deep generative model. For instance, new materials or chemical compounds can be generated with the knowledge of existing substances and basic rules.

References

[1] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

[2] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in neural information processing systems*, pages 2172–2180, 2016.

[3] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. The MIT Press, 2016.

[4] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[5] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, Günter Klambauer, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a nash equilibrium. *CoRR*, abs/1706.08500, 2017.

[6] 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. In *International Conference on Learning Representations*, 2017.

[7] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. *arXiv preprint arXiv:1802.05983*, 2018.

[8] Diederik P. Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *arXiv preprint arXiv:1807.03039*, 2018.

[9] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016.

[10] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[11] George Montañez. Why machine learning works. 2018.

[12] Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. Do deep generative models know what they don't know? *arXiv preprint arXiv:1810.09136*, 2018.

[13] Paul Rubenstein, Bernhard Schoelkopf, and Ilya Tolstikhin. On the latent space of wasserstein auto-encoders. *arXiv preprint arXiv:1802.03761*, 2018.

[14] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *arXiv preprint arXiv:1606.03498*, 2016.

[15] Lucas Theis, Aäron van den Oord, and Matthias Bethge. A note on the evaluation of generative models. *arXiv preprint arXiv:1511.01844*, 2015.

[16] Aaron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. In *International Conference on Machine Learning*, 2016.

[17] V Vapnik. Principles of risk minimization for learning theory. *Advances in Neural Information Processing Systems*, 4:831–838, 1992.

[18] V. N. Vapnik and A. Ya. Chervonenkis. *On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities*. Springer International Publishing, 2015.

[19] Vikas Verma, Alex Lamb, Christopher Beckham, Aaron Courville, Ioannis Mitliagkis, and Yoshua Bengio. Manifold mixup: Encouraging meaningful on-manifold interpolation as a regularizer. *arXiv preprint arXiv:1806.05236*, 2018.

[20] Hongliang Yan, Yukang Ding, Peihua Li, Qilong Wang, Yong Xu, and Wangmeng Zuo. Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation. In *Computer Vision and Pattern Recognition*, 2017.

[21] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018.