Tradeoffs in Neural Variational Inference

Ashwin D'Cruz
Department of Engineering
University of Cambridge
add39@cam.ac.uk

Sebastian Nowozin
Microsoft Research
Cambridge
Sebastian.Nowozin@microsoft.com

Bill Byrne
Department of Engineering
University of Cambridge
wjb31@cam.ac.uk

1 Introduction

Representation learning is useful for a wide range of applications (Bengio et al. (2013)). Learned good representations can be used as inputs for supervised machine learning systems (Salakhutdinov (2009)) such as speech recognition (Seide et al. (2011)), object recognition (Krizhevsky et al. (2012)), and natural language processing(Bengio (2008)). A recent approach to representation learning has been the Variational Auto-Encoder (VAE)(Kingma and Welling (2013); Rezende et al. (2014)). This is an efficient probabilistic deep learning method that has gained a lot of popularity recently. It is widely applied due to its ease of use and promising results (Doersch (2016)).

The VAE is comprised of a generative process $p_\theta(\mathbf{x}|\mathbf{z})$ and an approximate recognition or inference process $q_\phi(\mathbf{z}|\mathbf{x})$. Both these processes can be parametrized with neural networks allowing joint optimization using common techniques such as stochastic gradient descent. The quality of the inference and generative process are dependent on the accuracy of the inference network. There have been several proposals on how to improve the inference network of the VAE. While these improvements draw comparisons to the original VAE work, a thorough comparison between the different improvements is lacking in the community.

We present a quantitative evaluation between the following models: Importance Weighted Auto-Encoder (IWAE) (Burda et al. (2015)), Auxiliary Deep Generative Model (ADGM)(Maaløe et al. (2016)), Skip Deep Generative Model (SDGM) (Maaløe et al. (2016)), Householder Flow Model (Tomczak and Welling (2016)), and Inverse Autoregressive Flow (IAF) (Kingma et al. (2016)) model.

The common metric used in comparing deep generative models is the variational lower bound (ELBO) as shown in Equation 1 below:

$$\mathbf{E}_{q_\phi}[logp(x,z) - logq_\phi(z|x)] \tag{1}$$

We report on this metric but additionally investigate the modeling time taken by these various approaches as this is an important practical consideration. For diversity, we examine the performance of these models on three datasets of varying complexity in our original work(D'Cruz et al. (2017)): MSRC-12 (Pose) (Fothergill et al. (2012)), MNIST (LeCun (1998)), and celebA (Liu et al. (2015)). In this work, we examine some of the results on the Pose dataset.

Table 1: Pose data: average ELBO over the test set (175,638 samples). A single Monte Carlo sample was used for the expectation unless otherwise stated.

Model	ELBO Mean
IAF (8 transformations)	146.21
SDGM	142.96
IAF (4 transformation)	136.24
IAF (3 transformation)	135.82
Householder (10 transformations)	133.03
IAF (2 transformation)	132.08
IWAE (2 MC samples)	130.81
ADGM	127.08
VAE (1 MC sample)	126.03
IWAE (1 MC sample)	125.80
IAF (1 transformation)	125.78
IWAE (5 MC samples)	122.91
VAE (2 MC samples)	122.34
Householder (1 transformation)	119.85
VAE (5 MC samples)	119.34

Table 2: Pose data: average times (seconds) over the training set (351,275 samples). We also report the standard deviation in brackets.

Model	Encoding	Decoding	Update
IAF (8 transformations)	0.0041 ($\pm$ 0.00140)	0.0277 ($\pm$ 0.00203)	0.1397 ($\pm$ 0.00293)
IAF (4 transformations)	0.0041 ($\pm$ 0.00183)	0.0169 ($\pm$ 0.00154)	0.1326 ($\pm$ 0.00275)
IAF (3 transformations)	0.0028 ($\pm$ 0.00913)	0.0128 ($\pm$ 0.01430)	0.0117 ($\pm$ 0.01518)
IAF (2 transformations)	0.0044 ($\pm$ 0.00654)	0.0107 ($\pm$ 0.00801)	0.0123 ($\pm$ 0.01152)
IAF (1 transformation)	0.0040 ($\pm$ 0.00171)	0.0077 ($\pm$ 0.00068)	0.0102 ($\pm$ 0.00152)
Householder (10 transformations)	0.0046 ($\pm$ 0.00498)	0.0229 ($\pm$ 0.01109)	0.0220 ($\pm$ 0.00825)
Householder (1 transformation)	0.0041 ($\pm$ 0.00126)	0.0073 ($\pm$ 0.01774)	0.0110 ($\pm$ 0.00132)
IWAE (5 MC samples)	0.0037 ($\pm$ 0.00138)	0.0033 ($\pm$ 0.00026)	0.0325 ($\pm$ 0.00278)
IWAE (2 MC samples)	0.0037 ($\pm$ 0.00142)	0.0033 ($\pm$ 0.00028)	0.0154 ($\pm$ 0.00150)
IWAE (1 MC sample)	0.0037 ($\pm$ 0.00129)	0.0034 ($\pm$ 0.00018)	0.0087 ($\pm$ 0.00074)
VAE (5 MC samples)	0.0036 ($\pm$ 0.00188)	0.0034 ($\pm$ 0.00047)	0.0245 ($\pm$ 0.00350)
VAE (2 MC samples)	0.0038 ($\pm$ 0.00134)	0.0033 ($\pm$ 0.00187)	0.0126 ($\pm$ 0.00454)
VAE (1 MC sample)	0.0076 ($\pm$ 0.03913)	0.0036 ($\pm$ 0.00101)	0.0095 ($\pm$ 0.01733)
SDGM	0.0068 ($\pm$ 0.00198)	0.0064 ($\pm$ 0.00452)	0.2467 ($\pm$ 0.00330)
ADGM	0.0086 ($\pm$ 0.01838)	0.0065 ($\pm$ 0.00203)	0.2474 ($\pm$ 0.00306)

2 Results and Discussion

Table 1 displays the ELBO achieved by the various models on a held out test set of the Pose dataset, ranked in order of decreasing performance. We also report on the timings of different sections of the pipeline for these models in Table 2.

2.1 Discussion

While both the Householder and IAF models utilize normalizing flows, we note that the reduced complexity of the Householder flow leads to less powerful transformations of the latent space compared to IAF. Fewer IAF transformations are required to achieve better performances than the Householder models. The SDGM achieves strong results despite not using normalizing flows. We can consider normalizing flows as a way of adding depth in the latent space by repeatedly transforming the latent space. With the SDGM, depth in the latent space is explicitly addressed in the modeling assumptions. With the VAE and IWAE, we note that increasing the number of Monte Carlo (MC) samples can sometimes lead to over-fitting on the training set and poor fits on the held out test set.

The encoding time captures the time taken to move a data point through the encoder network to the first stage of the latent space. Further transformations of the latent space and the eventual decoding

back into a data point are considered part of the decoding time. The update time is the how long the network took to update its parameters via back propagation. For all models, we recorded the timings across 100 epochs and display the average in Table 2. Due to the our particular definition of encoding, we note that most models take the same amount of time to encode a data point. The exceptions are the ADGM and SDGM model which take longer due to the additional auxiliary variable. Due to the batch processing, we observed that while the IWAE takes longer to train, during inference it isn't significantly slower than the standard VAE. Comparing the normalizing flow methods, we see that adding another Householder flow is far less costly than adding another IAF transformation for the training stage.

2.2 Practical Considerations

Across the datasets and models, we observe that training times have larger variations compared to inference timings. If training times are not an issue, the IAF and SDGM model perform quite well during testing. The IWAE is also able to perform quite well though the number of MC samples is an extra parameter to tune as too many leave the model liable to over-fitting. This is in contrast to the normalizing flow methods which seem to improve as more transformations are added. Compute resources permitting, more transformations are recommended.

3 Conclusions and Future Work

A more complete set of results and evaluations can be found at D'Cruz (2017). We offer a thorough comparison of several modifications to the VAE, examining both ELBO and modeling time. By considering both these aspects, we are able to offer practical guidelines on choosing an appropriate model for the task at hand. Through examining trends on different datasets, we are able to state our conclusions more decisively. In addition, we have also contributed an extensive Chainer code repository (D'Cruz (2017)) to allow others access to these new probabilistic deep generative models for the purpose of replicating our experiments and also applying these models to new tasks.

Moving forward, we will consider more models and carry out ablative studies. We also propose using convolutional layers instead of standard multi-layer perceptrons as the former have shown promise for image data (Krizhevsky et al. (2012); Radford et al. (2015)). Depth in the stochastic layer will also be investigated.

References

Y. Bengio. Neural net language models. *Scholarpedia*, 3(1):3881, 2008.

Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

Y. Burda, R. Grosse, and R. Salakhutdinov. Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*, 2015.

A. D'Cruz. Deep generative models. `https://github.com/ashwindcruz/dgm`, 2017.

A. D'Cruz, S. Nowozin, and B. Byrne. Tradeoffs in neural variational inference, 2017. URL `http://www.mlsalt.eng.cam.ac.uk/foswiki/pub/Main/ClassOf2017/Cruz_Dissertation.pdf`.

C. Doersch. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*, 2016.

S. Fothergill, H. Mentis, P. Kohli, and S. Nowozin. Instructing people for training gestural interactive systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1737–1746. ACM, 2012.

D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

D. P. Kingma, T. Salimans, and M. Welling. Improving variational inference with inverse autoregressive flow. *arXiv preprint arXiv:1606.04934*, 2016.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

Y. LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015.

L. Maaløe, C. K. Sønderby, S. K. Sønderby, and O. Winther. Auxiliary deep generative models. *arXiv preprint arXiv:1602.05473*, 2016.

A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014.

R. Salakhutdinov. *Learning deep generative models*. University of Toronto, 2009.

F. Seide, G. Li, and D. Yu. Conversational speech transcription using context-dependent deep neural networks. In *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

J. M. Tomczak and M. Welling. Improving variational auto-encoders using householder flow. *arXiv preprint arXiv:1611.09630*, 2016.