Auxiliary Deep Generative Models

Lars Maaløe[1], Casper Kaae Sønderby[2], Søren Kaae Sønderby[2] and Ole Winther[1,2]

[1]Cognitive Systems, Department of Applied Mathematics and Computer Science, Technical University of Denmark
[2]Bioinformatics Centre, Department of Biology, University of Copenhagen

Introduction

We train deep generative models with multiple stochastic layers. The Auxiliary Deep Generative Models (ADGM) utilize an extra set of auxiliary latent variables to increase the flexibility of the variational distribution. We also introduce a slight change to the ADGM, a 2-layered stochastic model with skip connections, the Skip Deep Generative Model (SDGM). Both models are trainable end-to-end and offer state-of-the-art performance when compared to other semi-supervised methods. We demonstrate:

(i) The auxiliary variable models can fit complex latent distributions.

(ii) The ADGM utilizes the data manifold for semi-supervised classification.

(iii) State-of-the-art results on end-to-end semi-supervised classification.

Auxiliary method

We extend the variational distribution of deep generative models with auxiliary variables *q(z,a) (cf.* ***Fig. 3, 4****)*. We demonstrate hat the marginal distribution of the variational approximation *q(z)* can fit more complicated posteriors (Agakov and Barber, 2004). In order to have an unchanged generative model, it is required that the joint mode *p(x,z,a)* gives back the original *p(x,z)* under marginalization over *'a'*, thus *p(x,z,a) = p(a|x,z)p(x,z) (cf.* ***Fig. 3****)*.

The lower bound of the auxiliary variational auto-encoder is defined as

$$\log p(x) = \log \int_a \int_z p(x,a,z) da dz$$

$$\geq \mathbb{E}_{q_\phi(a,z|x)} \left[\log \frac{p_\theta(a|z,x) p_\theta(x|z) p(z)}{q_\phi(a|x) q_\phi(z|a,x)} \right] \equiv -\mathcal{U}_{\mathrm{AVAE}}(x) \ .$$

The stochastic variables are defined as in a normal variational auto-encoder (Kingma et al., 2013; Rezende and Mohamed, 2014).

For semi-supervised learning we represent the class label with a multinomial latent variable y. y can be explicitly marginalized when unobserved *(cf.* ***Fig. 4****)*. The lower bound for the labeled data is

$$\log p(x,y) = \log \int_a \int_z p(x,y,a,z) dz da$$

$$\geq \mathbb{E}_{q_\phi(a,z|x,y)} \left[\log \frac{p_\theta(x,y,a,z)}{q_\phi(a,z|x,y)} \right] \equiv -\mathcal{L}(x,y) \ .$$

The performance is improved by introducing an explicit classification loss

$$\mathcal{L}_l(x_l,y_l) = \mathcal{L}(x_l,y_l) + \alpha \cdot \mathbb{E}_{q_\phi(a|x_l)}[-\log q_\phi(y_l|a,x_l)] \ .$$

The lower bound for the unlabelled data points is

$$\log p(x) = \log \int_a \int_y \int_z p(x,y,a,z) dz dy da$$

$$\geq \mathbb{E}_{q_\phi(a,y,z|x)} \left[\log \frac{p_\theta(x,y,a,z)}{q_\phi(a,y,z|x)} \right] \equiv -\mathcal{U}(x) \ .$$

The objective function for the semi-supervised model is

$$\mathcal{J} = \sum_{(x_l,y_l)} \mathcal{L}_l(x_l,y_l) + \sum_{(x_u)} \mathcal{U}(x_u) \ .$$

The SDGM has a slightly changed generative model compared to the ADGM, so that the auxiliary variable is now in *p(x|a,z,y)*.

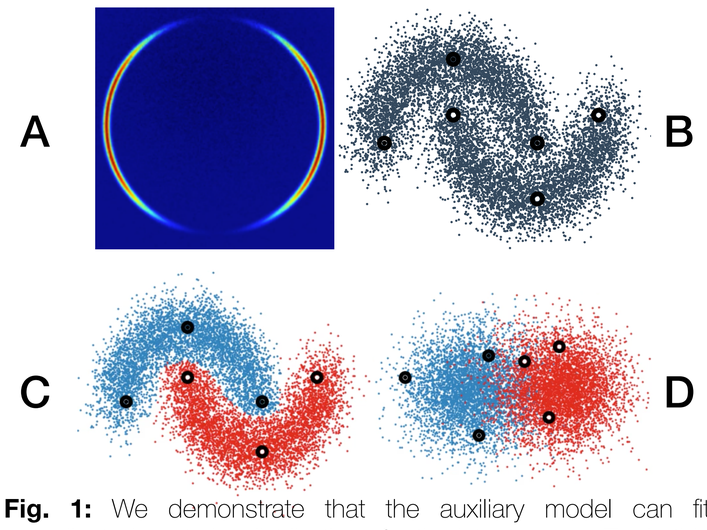

Fig. 1: We demonstrate that the auxiliary model can fit complicated posterior distributions for the latent space. To do this we consider one of the 2D potential models from Rezende and Mohamed, 2015. We also show how the auxiliary model captures the data manifold by applying the auxiliary variables, and thereby manages to distinguish between the half-moons. **A)** The approximation of the energy with two modes. The most frequent solution found in optimization is not the one shown, but one where the latent distribution fits only one of the two equivalent modes. **B)** A semi-supervised problem consisting of two classes/ half-moons and 6 labeled data points. **C)** The result of the ADGM fitting the two half-moons. **D)** The auxiliary latent space of a fully trained ADGM.

	MNIST 100 labels	SVHN 1000 labels	NORB 1000 labels
M1+TSVM	11.82% (±0.25)	55.33% (±0.11)	18.79% (±0.05)
M1+M2	3.33% (±0.14)	36.02% (±0.10)	-
VAT	2.12 %	24.63 %	9.88 %
Ladder Network	1.06% (±0.37)	-	-
ADGM	**0.96% (±0.02)**	22.86 %	10.06% (±0.05)
SDGM	1.32% (±0.07)	**16.61% (±0.24)**	**9.40% (±0.04)**

Table 1: Semi-supervised test error % benchmarks on MNIST, SVHN and NORB for randomly labeled and evenly distributed data points from Kingma et al., 2014, Miyato et al., 2015 and Rasmus et al., 2015. The lower section demonstrates the benchmarks of the contribution of this article.

	ELBO
VAE+NF, L=1 (Rezende and Mohamed, 2015)	-85.10
IWAE, L=1, IW=1 (Burda et al., 2015)	-86.76
IWAE, L=1, IW=50 (Burda et al., 2015)	-84.78
IWAE, L=2, IW=1 (Burda et al., 2015)	-85.33
IWAE, L=2, IW=50 (Burda et al., 2015)	-82.90
VAE+VGP, L=2 (Tran et al., 2015)	-81.90
LVAE, L=5, IW=1 (Sønderby et al., 2016)	-82.12
LVAE, L=5, IW=10. FT (Sønderby et al., 2016)	-81.74
Auxiliary VAE, L=1, IW=1	-84.59
Auxiliary VAE, L=2, IW=1	-82.97

Table 2: Unsupervised test log-likelihood on permutation invariant MNIST for the normalizing flows VAE (VAE+NF), importance weighted auto-encoder (IWAE), variational Gaussian process VAE (VAE+VGP) and Ladder VAE (LVAE) with FT denoting the finetuning procedure from Sønderby et al. (2016), IW the importance weighted samples during training, and L the number of stochastic latent layers.

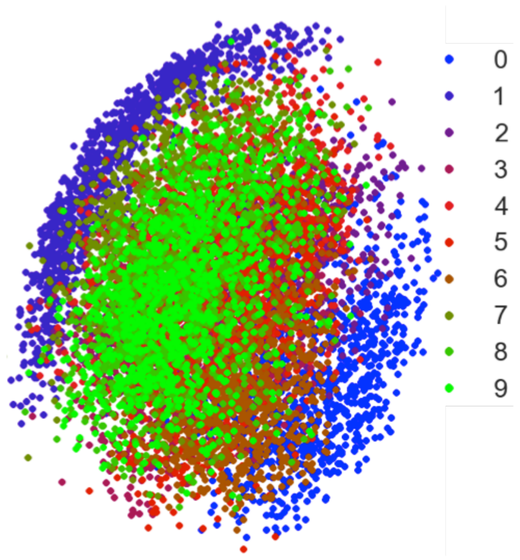

Fig. 2: SDGM trained on 100 labeled MNIST. PCA on the 1st and 2nd principal component of the auxiliary latent space.

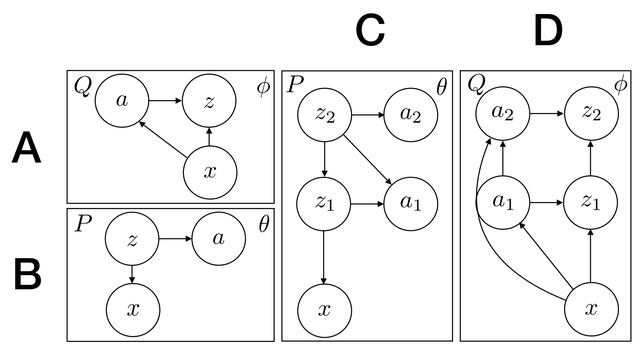

Fig. 3: The unsupervised auxiliary variational auto-encoder. The incoming connections to each variable are deep neural networks with parameters θ and ϕ. **A)** 1-layered inference model and **B)** corresponding generative model. **C)** 2-layered inference model and **D)** generative model.

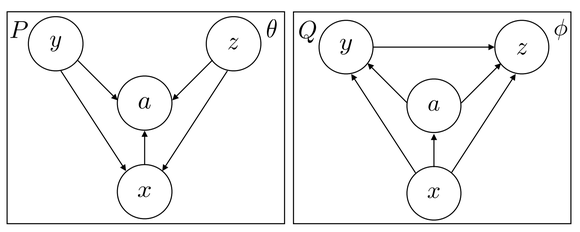

Fig. 4: The ADGM for semi-supervised learning. The incoming connections to each variable are deep neural networks with parameters θ and ϕ. **A)** Generative model and **B)** inference model.

Results

We evaluate the generative performance of the unsupervised auxiliary model, AVAE, using the MNIST dataset. The inference and generative models are visualised in **Fig. 3**. We report the lower bound for 5000 importance weighted samples and use the same training and parameter settings as in Sønderby et al. (2016) with warm-up, batch normalization and 1 Monte Carlo and IW sample for training. **Table 2** shows the results and even though they are not directly comparable it is evident that the auxiliary variational auto-encoder is outperforming a regular variational auto-encoder.

In order to compare the ADGM and SDGM to other methods we evaluated the semi-supervised performance on the MNIST, SVHN and NORB datasets (cf. **Table 1**). The SDGM were generally much more stable in convergence and speed especially on Gaussian distributed observed data points.

Conclusion

We have introduced a method for making the variational distributions used in deep generative models more expressive. We have demonstrated that the method gives state-of-the-art performance in a number of semi-supervised benchmarks and is trainable end-to-end.

References

Agakov, F., Barber, D., (2004), An Auxiliary Variational Method, In Neural Information Processing, volume 3316, pages 561-566, Springer Berlin Heidelberg.

Burda, Y., Grosse, R., Salakhutdinov, R., (2015), Importance Weighted Autoencoders, arXiv preprint: 509.00519.

Kingma, D.P., Rezende, D.J., Mohamed, S., Welling, M., (2014), Semi-Supervised Learning with Deep Generative Models, ICML, pages 3581–3589.

Kingma, D.P., Welling, M., (2013), Auto-Encoding Variational Bayes, arXiv preprint: 1312.6114.

Miyato, T., Maeda, S.-i., Koyama, M., Nakae, K., and Ishii, S. (2015), Distributional Smoothing with Virtual Adversarial Training, arXiv preprint arXiv: 1507.00677.

Rezende, D.J., Mohamed, S., (2015), Variational Inference with Normalising Flows, ICML, pages 1530-1538.

Rezende, D.J., Mohamed, S., Wierstra, D., (2014), Stochastic Backpropagation and Approximate Inference in Deep Generative Models, arXiv preprint: 1401.4082.

Rasmus, A., Berglund, M., Honkala, M., Valpola, H., and Raiko, T. (2015), Semi-supervised learning with ladder networks, NIPS, pages 3532–3540.

Sønderby, C.K.., Raiko, T., Maaløe, L., Sønderby, S.K., Winther, O., (2016), Ladder Variational Autoencoders, arXiv preprint: 1602.02282.

Tran, D., Raganath, R., Blei, D.M., (2015), Variational Gaussian process, arXiv preprint: 1511.06499.