Improving Semi-Supervised Learning with Auxiliary Deep Generative Models

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

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

UNIVERSITY OF COPENHAGEN

Auxiliary Deep Generative Models

We introduce the *auxiliary deep generative model* (ADGM) and apply it to semi-supervised learning. Contrary to previous deep generative models for semi-supervised learning[1] the ADGM is trainable end-to-end and achieve state-of-the-art on semi-supervised classification of MNIST (cf. **Fig. 1, 2**). The generative model is defined as

$$p_\theta(x|z,y) = f(x;z,y,\theta); \quad p(z) = \mathcal{N}(z|0,\mathrm{I});$$
$$p(y) = \mathrm{Cat}(y|\pi); \quad p(a) = \mathcal{N}(a|0,\mathrm{I}).$$

And the corresponding inference model is

$$q_\phi(a|x) = \mathcal{N}(a|\mu_\phi(x), \mathrm{diag}(\sigma^2_\phi(x))),$$
$$q_\phi(z|y,x) = \mathcal{N}(z|\mu_\phi(y,x), \mathrm{diag}(\sigma^2_\phi(y,x))),$$
$$q_\phi(y|a,x) = \mathrm{Cat}(y|\pi_\phi(a,x)).$$

The key point of the ADGM is that the auxiliary unit introduces a class specific latent distribution between input and output of the classifier allowing a more expressive discriminative distribution. Further the stochasticity of the auxiliary unit maps each input into a latent distribution used for the discriminative classifier, which is richer than a deterministic dependency. We show that the ADGM,

(i) have state-of-the-art results (0.96% error) on semi-supervised classification on MNIST with 100 labels,
(ii) is trainable end-to-end without the need for any pre-training,
(iii) have good convergence properties and
(iv) that its stochastic auxiliary variable is essential for good discriminative classification.

Variational Lower Bound

We optimize the model by maximizing the lower bound on the likelihood. The variational lower bound on the marginal likelihood for a single *labeled data point* is

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

For *unlabeled data* we further introduce:

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

where $\mathcal{H}(.)$ is the entropy. Since the classification loss is not part of the labeled data lower bound we introduce:

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

where α is a weight between generative and discriminative learning. The variational lower bound for labeled and unlabeled data is

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

Experiments

The ADGM achieves state-of-the-art results (0.96% error) on semi-supervised classification on MNIST with 100 labels (cf. **Table 1**). The information contribution from the auxiliary units and the latent units are seen in **Fig. 3**. The number of active units in the latent space is around 20. The number of active auxiliary units, on the other hand, is much larger. We speculate that this is due to the up-weighting of the discriminative classification in the lower bound. **Fig. 4** shows how the ADGM outperforms both a similarly optimized M2 model and an ADGM where the auxiliary unit is deterministic and further that the convergence rate of the ADGM is the fastest. In **Fig. 5** we visualize 10 Gaussian distributed random samples conditioned on each class *y*.

Conclusion

We have shown that making the discriminative distribution more flexible by introducing extra auxiliary variables gives state-of-the-art performance on the 100 labeled examples MNIST benchmark. We are in the progress of extending this to other semi-supervised scenarios. It is also of interest to extend this approach to both fully unsupervised and supervised generative settings. Currently we are combing the proposed framework with the new tighter bound by Burda et. al. [5].

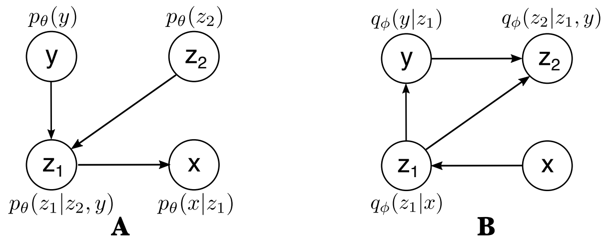

Figure 1: Graphical representation of Kingma et al. (M1+M2) model [1]. **A)** The generative model *P*. **B)** The inference model *Q*. The M1 model acts as a feature extractor, and these features are fed into the semi-supervised classifier M2. Although both M2 and M1+M2 should be powerful generative models for semi-supervised learning, direct application of these models *failed to deliver good results*. Instead Kingma et al. [1] trained M1 and M2 independently.

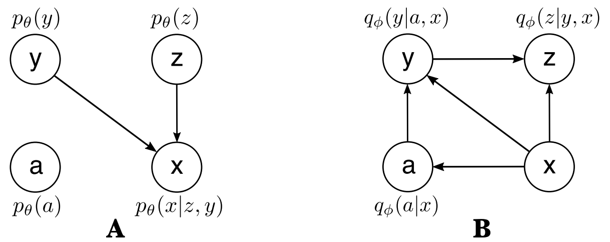

Figure 2: Graphical representation of *the auxiliary deep generative model* (ADGM). **A)** The generative model *P*. **B)** The inference model *Q*. By adding the auxiliary unit to the inference model it acts as a feature extraction as well as a data augmentation step. This way the discriminative classifier receives a much more expressive input.

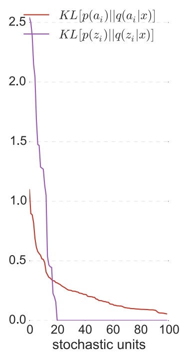

Figure 3: Calculating the KL divergence between prior and approximate posterior to determine active stochastic units in the auxiliary and latent space.

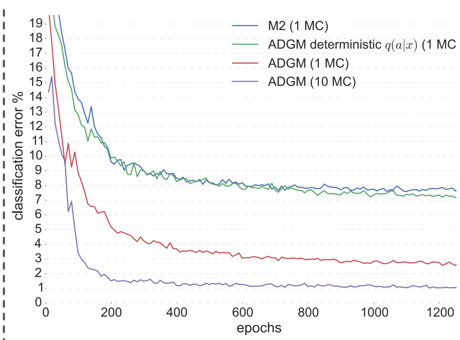

Figure 4: Comparison of the auxiliary deep generative model to Kingma et al.'s M2 model [1] and a model where the auxiliary variable is deterministic. All models are trained using same hyperparameters. The stochastic auxiliary units are important for both classification and convergence rate.

Model	*Err. (%)*
AtlasRBF [2]	8.10% (±0.95)
Deep Generative Model (M1+M2) [1]	3.33% (±0.14)
Virtual Adversarial [3]	2.12%
Ladder [4]	1.06% (±0.37)
Auxiliary Deep Generative Model (1 Monte Carlo samples)	2.25% (±0.08)
Auxiliary Deep Generative Model (10 Monte Carlo samples)	**0.96% (±0.02)**

Table 1: Classification errors with standard deviation for the best performing semi-supervised models on the MNIST dataset with 100 labels.

Figure 5: 100 Gaussian distributed random samples drawn from the 100 dimensional latent distribution in the auxiliary deep generative model with a fixed class *y*.

References

[1] Kingma, D. P., Rezende D.J., Mohamed, S., and Welling, M. (2014). Semi-Supervised Learning with Deep Generative Models. arXiv preprint arXiv: 1406.5298.
[2] Pitelis, N., Russell, C., and Agapito, L. (2014). Semi-supervised Learning Using an Unsupervised Atlas. In Calders, T., Esposito, F., Hllermeier, E., and Meo, R., editors, *Machine Learning and Knowledge Discovery in Databases*, volume 8725 of *Lecture Notes in Computer Science*, pages 565–580. Springer Berlin Heidelberg.
[3] Miyato, T., Maeda, S.-i., Koyama, M., Nakae, K., and Ishii, S. (2015). Distributional Smoothing with Virtual Adversarial Training. *arXiv preprint arXiv:1507.00677*.
[4] Rasmus, A., Valpola, H., Honkala, M., Berglund, M., and Raiko, T. (2015). Semi-Supervised Learning with Ladder Network. *arXiv preprint arXiv:1507.02672*.
[5] Burda, Y., Grosse, R., & Salakhutdinov, R. (2015). Importance Weighted Autoencoders. arXiv preprint arXiv:1509.00519