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 performance on semi-supervised classification of MNIST (cf. Fig. 1, 2). The generative model is defined as

\[ p_{\theta}(x|y, \alpha) = \mathcal{N}(z|\mu_{\phi}(x, y)), \]

and the corresponding inference model

\[ q_{\phi}(z|x) = \mathcal{N}(a|\mu_{\phi}(x), \sigma_{\phi}^2(x)), \]

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

(i) have state-of-the-art results (9.06% 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 performance.

**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

\[ \log p_{\theta}(x, y) \geq E_{q_{\phi}(z|x, y)} \left[ \log \frac{p_{\theta}(y)p_{\theta}(z|x, y)}{q_{\phi}(z|x, y)} \right] = -\mathcal{L}(x, y) \]

For unlabeled data we further introduce:

\[ \log p_{\theta}(x) \geq E_{q_{\phi}(z|x, y)} \left[ \log \frac{p_{\theta}(y)p_{\theta}(z|x, y)}{q_{\phi}(z|x, y)} \right] = -U(x) \]

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

\[ -\mathcal{L}(x, y) = -\mathcal{L}(x, y) - \alpha \cdot E_{q_{\phi}(z|x, y)} \left[ -\log q_{\phi}(y|a, x) \right] \]

where \( \alpha \) is a weight between generative and discriminative learning. The variational lower bound for labeled and unlabeled data is

\[ J = \sum (x, y) \mathcal{L}(x, y) + \sum U(x_n) \]

**Experiments**

The ADGM achieves state-of-the-art results (9.06% 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.

**Conclusion**

We have shown that making the discriminative distribution more flexible by introducing 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 combining the proposed framework with the new tighter bound by Burda et. al. [5].