Improving Semi-Supervised Learning with Auxiliary Deep Generative Models

Lars Maaløe
larsma@dtu.dk

Casper Kaae Sønderby
casper.sonderby@bio.ku.dk

Søren Kaae Sønderby
soren.sonderby@bio.ku.dk

Ole Winther
olwi@dtu.dk

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

Abstract

Deep generative models based upon continuous variational distributions parameterized by deep networks give state-of-the-art performance. In this paper we propose a framework for extending the latent representation with extra auxiliary variables in order to make the variational distribution more expressive for semi-supervised learning. By utilizing the stochasticity of the auxiliary variable we demonstrate how to train discriminative classifiers resulting in state-of-the-art performance within semi-supervised learning exemplified by an 0.96% error on MNIST using 100 labeled data points. Furthermore we observe empirically that using auxiliary variables increases convergence speed suggesting that less expressive variational distributions, not only lead to looser bounds but also slower model training.

1 Introduction

Deep neural networks have recently showed impressive performance on a wide range of tasks. These improvements have been achieved by better algorithms, faster computers and increased availability of large labeled datasets in many areas, e.g. image recognition (Deng et al., 2009). In many practical situations it is relatively inexpensive to acquire data but expensive to label it. This makes semi-supervised learning attractive. There exist many approaches to performing semi-supervised learning, e.g. transductive SVM (TSVM) (Joachims, 1999), EM methods (Nigam et al., 2000), graph-based methods (Blum and Chawla, 2001; Zhu et al., 2003; Pitelis et al., 2014) and deep auto-encoders (Rifai et al., 2011; Ranzato and Szummer, 2008; Weston et al., 2008).

Recently several different (deep) models have significantly improved the performance on semi-supervised classification tasks. Kingma et al. (2014) introduced a deep generative model for semi-supervised learning (DGM) by augmenting the auto-encoding variational Bayes (AEVB) model (Kingma, 2013; Rezende et al., 2014) algorithm with labeled units. Miyato et al. (2015) introduced an improved semi-supervised learner by applying adversarial training to deep networks. Finally Rasmus et al. (2015) introduced a generalization of the ladder network (Valpola, 2014) that has the ability to learn a latent classification variable.

In this article we introduce the auxiliary deep generative model (ADGM) and apply it to semi-supervised learning. In the ADGM, the variational encoder model has an extra set of stochastic variables compared to the generative decoder model. These extra so-called auxiliary variables makes the variational model more flexible and thus able to improve the variational lower bound (Agakov and Barber, 2004). The auxiliary variable and the input data is fed into a variational auto-encoder and a discriminative classifier. Empirically we show that the ADGM, (i) obtain state-of-the-art results on semi-supervised classification, (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. The Hierarchical variational models approach by Ranganath et al. (2015) is a closely related recent independently proposed framework that treats the parameters...
of the variational distribution as auxiliary variables as a means to get a more expressive variational approximation.

2 Methods

[Kingma et al., 2014] introduced a probabilistic approach to semi-supervised learning by stacking a generative feature extractor (called M1) and a generative semi-supervised model (M2) into a stacked generative semi-supervised model (M1+M2). M1 is a variational auto-encoder, where the generative model is defined as \( p_\theta(z)p_\theta(x|z) \) (decoder with parameters \( \theta \)), with the variational approximation being \( q_\phi(z|x) \) (encoder with parameters \( \phi \)), as replacement for the intractable posterior \( p_\theta(z|x) \) (Kingma, 2013). M2 includes labels \( y \) in the generative model: \( p_\theta(x|y,z)p_\theta(z)p_\theta(y) \), decoder \( q_\phi(z|y,x) \) and a discriminative classifier, \( q_\phi(y|x) \). The generative model \( p \) combining M1 and M2 called M1+M2 is (cf. fig. 1a):

\[
\begin{align*}
M1 &: p_\theta(x|z_1) = f(x; z_1, \theta), \\
M2 &: p(y) = \text{Cat}(y|x); \quad p(z_2) = \mathcal{N}(z_2|0, I); \quad p_\theta(z_1|y,z_2) = f(z_1; y, z_2, \theta),
\end{align*}
\]

where \( f(\cdot) \) are the decoders and \( \text{Cat}(\cdot) \) is the multinomial distribution. The corresponding inference model \( Q \) is (cf. fig. 1b):

\[
\begin{align*}
M1 &: q_\phi(z_1|x) = \mathcal{N}(z_1|\mu_\phi(x), \text{diag}(\sigma^2_\phi(x))), \\
M2 &: q_\phi(z_2|y, z_1) = \mathcal{N}(z_2|\mu_\phi(y, z_1), \text{diag}(\sigma^2_\phi(y, z_1))); \quad q_\phi(y|z_1) = \text{Cat}(y|\pi_\phi(z_1)).
\end{align*}
\]

Since latent variables \( z_2 \) and \( y \) in M2 are marginally independent, the class specific information can be modeled through \( y \) and remaining information through \( z_2 \).

However, 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 in benchmarks. [Kingma et al., 2014] reported 11.97\% (±1.71) classification error for MNIST with 100 labeled examples and no results were reported for M1+M2. Instead [Kingma et al., 2014] trained M1 to get latent features \( z_1 \) that was then used as input for training M2.

In this contribution we propose an alternative formulation with two sets of stochastic variables that converges end-to-end and achieves state-of-the-art performance. The ADGM has the generative model \( p \) defined as \( p_\theta(a)p_\theta(y)p_\theta(z)p_\theta(x|y, z) \) (cf. fig. 1c), where \( a, y, z \) are the auxiliary variable, class label, and latent features, respectively. Learning the posterior distribution is intractable, thus we define the approximation as \( q_\phi(a|x)q_\phi(z|y, x) \) and a classifier \( q_\phi(y|a, x) \) (cf. fig. 1d). The distributions of the generative model \( p \) are

\[
\begin{align*}
p_\theta(x|z, y) &= f(x; z, y, \theta); & p(z) &= \mathcal{N}(z|0, I); \\
p(y) &= \text{Cat}(y|x); & p(a) &= \mathcal{N}(a|0, I).
\end{align*}
\]

And for the corresponding inference model \( q \)

\[
\begin{align*}
q_\phi(a|x) &= \mathcal{N}(a|\mu_\phi(x), \text{diag}(\sigma^2_\phi(x))), \\
q_\phi(z|y, x) &= \mathcal{N}(z|\mu_\phi(y, x), \text{diag}(\sigma^2_\phi(y, x))), \\
q_\phi(y|a, x) &= \text{Cat}(y|\pi_\phi(a, x)).
\end{align*}
\]

The key point of the ADGM is that the auxiliary unit \( a \) introduces a class specific latent distribution between \( x \) and \( y \) allowing a more expressive discriminative distribution \( q_\phi(y|a, x) \). Further the stochasticity of \( a \) maps each input into a distribution \( q_\phi(a|x) \) used for the discriminative classifier, which is richer than a deterministic dependency between \( x \) and \( y \). Note that it is possible to let \( a \)
be conditioned on $x, y, z$ in the generative model, but we found that this did not improve the performance. The ADGM use multi-layered perceptrons (MLP) to model $q_\phi(a|x)$, $q_\phi(z|y, x)$, $q_\phi(y|a, x)$ and $p_\theta(x|z, y)$.

For Gaussian distributions we apply the reparameterization trick, introduced in (Kingma [2013]) to backpropagate the error signal through the latent variables. We approximate the expectations by drawing unbiased Monte Carlo (MC) estimates $\tilde{a} \sim q_\phi(a|x)$ and $\tilde{z} \sim q_\phi(z|y, x)$.

**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

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

with $q_\phi(a, z|x, y) = q_\phi(a|x)q_\phi(z|y, x)$. For unlabeled data we further introduce the variational distribution for $y, q_\phi(y|a, x)$:

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

with $q_\phi(a, y, z|x) = q_\phi(z|y, x)q_\phi(y|a, x)q_\phi(a|x)$.

The discriminative classifier for the auxiliary variable model is defined as $\mathbb{E}_{q_\phi(a|x)}[q_\phi(y|a, x)]$. As noted by Kingma et al. (2014) the discriminative term (6) appears in the objective $-\mathcal{U}(x)$, but not in $-\mathcal{L}(x, y)$. Kingma et al. (2014) observed that for this model reads:

$$
\mathcal{L}_t(x_t, y_t) = \mathcal{L}(x_t, y_t) - \alpha \cdot \log \mathbb{E}_{q_\phi(a|x_t)}[q_\phi(y_t|a, x_t)],
$$

where $\alpha$ is a weight between generative and discriminative learning. The variational lower bound for labeled $x_t, y_t$ and unlabeled data $x_u$ is

$$
\mathcal{J} = \sum_{(x_t, y_t)} \mathcal{L}_t(x_t, y_t) + \sum_{(x_u)} \mathcal{U}(x_u).
$$

### 3 Results

Table 1 shows that the ADGM outperforms all previously proposed models on the MNIST dataset. The model’s convergence to around 1.5% is fast, and from that point the convergence speed declines (cf. Fig. 2a). In Fig. 2b we visualize 10 Gaussian distributed random samples conditioned on each class $y$.

<table>
<thead>
<tr>
<th>Model</th>
<th>Semi-Supervised (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AtlasRBF (Pitelis et al. [2014])</td>
<td>8.10% (±0.95)</td>
</tr>
<tr>
<td>Deep Generative Model (M1+M2) (Kingma et al. [2014])</td>
<td>3.33% (±0.14)</td>
</tr>
<tr>
<td>Virtual Adversarial (Miyato et al. [2015])</td>
<td>2.12%</td>
</tr>
<tr>
<td>Ladder (Rasmus et al. [2015])</td>
<td>1.06% (±0.37)</td>
</tr>
<tr>
<td>Auxiliary Deep Generative Model (1 MC)</td>
<td>2.23% (±0.08)</td>
</tr>
<tr>
<td>Auxiliary Deep Generative Model (10 MC)</td>
<td>0.96% (±0.02)</td>
</tr>
</tbody>
</table>

Table 1: Semi-supervised benchmarks on MNIST for 100 randomly labeled data points. The ADGM was trained by performing 1 and 10 samples.

Fig. 3 shows the information contribution from the auxiliary units $a$ and the latent units $z$. Like in Burda et al. [2015], the number of contributing/activated units in $z$ is quite low $\sim 20$. The number of contributing auxiliary units $a$, on the other hand, is much larger. We speculate that this is due to the upweighting of the discriminative classification in the lower bound. Fig. 2a shows how the ADGM outperforms a similarly optimized M2 model and an ADGM where the auxiliary unit is deterministic. We found that convergence of the M2 model was highly unstable. The result in Fig. 2a is the best achieved.

**Implementation details**

The ADGM is implemented in Python using Theano (Bastien et al. [2012]) and Lasagne (Dieleman et al. [2015]) libraries. For training, we have used the Adam (Kingma and Ba [2014]) optimization framework with a learning rate of 3e-4 and exponential decay rate for the 1st and 2nd moment.
at 0.9 and 0.999, respectively. The learning rate was annealed by .75 every 200 epochs. $\alpha$ was defined as $0.01 \cdot N$, where $N$ is the number of data points. We parameterized the MLPs with two layers of 500 hidden units using rectified linear units. All stochastic layers used 100 linear hidden units to parameterize $\mu$ and $\log(\sigma^2)$. The weights and biases are initialized using the Glorot and Bengio (2010) scheme. The averaging over the $a$ and $z$ variables were performed by sampling using the reparameterization trick (Kingma, 2013; Rezende et al., 2014) and the average over $y$ by exact enumeration so

$$E_{q(a, y, z|x)} [f(a, x, y, z)] \approx \frac{1}{N_{\text{samp}}} \sum_{i=1}^{N_{\text{samp}}} \sum_{y} q(y|a_i, x) f(a_i, x, y, z_{yi})$$  \quad (11)$$

with $a_i \sim q(a|x)$ and $z_{yi} \sim q(z|x, y)$. 1 or 10 samples was used for training and 100 for predictions.

We used MNIST as benchmark. In order to make a fair comparison with the ladder network, we have combined the training set of 50000 examples with the validation set of 10000 examples. The test set of 10000 remained as is. We used a batch size of 200 with the first half of the batch always being the 100 labeled samples. The labeled data are chosen randomly, but distributed evenly across classes. Before each epoch the normalized MNIST images were binarized by sampling Bernoulli distributions.

All experiments were carried out on GeForce GTX TITAN X GPUs. With 10 samples the runtime was $\sim$64 s/epoch. The training converged to around 1.5% classification error in 200 epochs corresponding to around 3 hours.

## 4 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. (2015), where a tight bound on $p(y, x)$ may be used directly for classification through $p(y|x) \propto p(y, x)$.
Acknowledgements
We thank Durk P. Kingma and Shakir Mohamed for helpful discussions. This research was supported by the Novo Nordisk Foundation and NVIDIA Corporation with the donation of TITAN X and Tesla K40 GPUs.

References


