Semi-Supervised Learning with Deep Generative Models

G. Gong, F. Giordaniello, J. Swiatkowski

Department of Engineering, University of Cambridge

Introduction

The semi-supervised learning algorithm has important practical impact due to the ever-increasing data size and difficulty of manually labeling the data. One of the goals is to learn the structure of small labeled data and generalize it to larger unlabeled sets. The main contribution of this paper is to develop new semi-supervised classification model using a fusion of deep neural network and probabilistic modeling to parametrize the data density. The experiment results for benchmark data set have quantitatively shown significant improved performance with small number of labeled data compared to previous approaches. The model can also capture the inter-class and intra-class variabilities.

Latent-feature Discriminative Model (M1)

Due to the large amount of unlabeled data, one common idea is to construct a mapping from the original data to its embedding space and then train a separate classifier on this latent space. This will allow the clustering of related data points for accurate classification. In fact, M1 model is equivalent to Variational Auto-Encoder. The model is defined as:

$$p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|\mathbf{0}, \boldsymbol{I});\ \ p_\theta(\boldsymbol{x}|\boldsymbol{z}) = f(\boldsymbol{x}; \boldsymbol{z}, \boldsymbol{\theta}) \qquad (1)$$

where $\boldsymbol{z}$ is the latent representation of data $\boldsymbol{x}$ and $f(\bullet)$ is a Gaussian Distribution over $\boldsymbol{x}$ parameterized by deep neural network, and acting as the decoder (Figure 1). The samples from posterior approximation $q(\cdot)$ can be then used to train separate classifiers. The training minimizes the variational lower bound, which is defined as

$$\log p_\theta(\boldsymbol{x}) \geqslant \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})] \\ - KL[q_\phi(\boldsymbol{z}|\boldsymbol{x})||p_\theta(\boldsymbol{z})] \qquad (2)$$

where $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ is used for feature extraction.

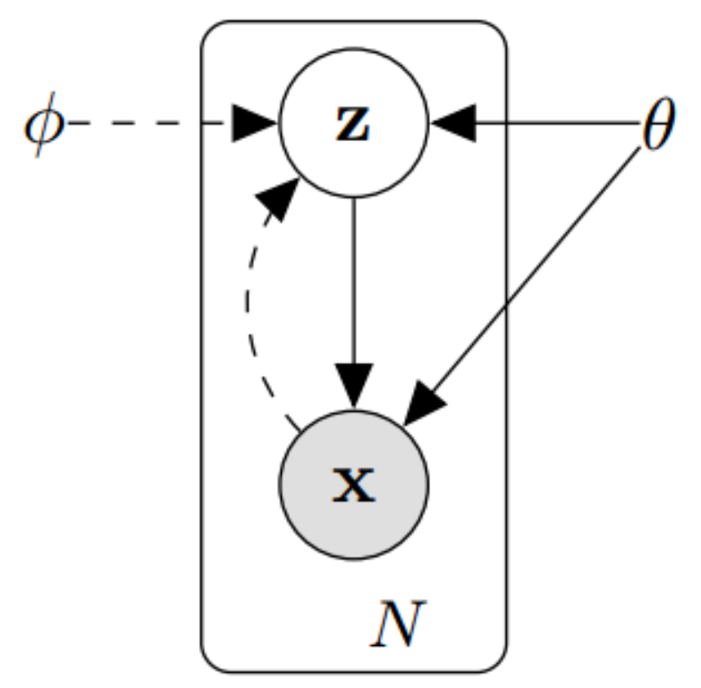

Fig. 1: M1 Graphical Model. The $\boldsymbol{\phi}$ and dashed line represent the variational parameters and encoder distribution $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$.

Generative semi-supervised model (M2)

Compared to M1 model, M2 assumes the unlabeled data $\boldsymbol{x}$ is generated by the latent class variable y in addition to a continuous latent variable $\boldsymbol{z}$. To be precise, the generative model is defined as:

$$p(y) = \mathrm{Cat}(y|\boldsymbol{\pi});\ p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|\mathbf{0}, \boldsymbol{I}); \\ p_\theta(\boldsymbol{x}|y, \boldsymbol{z}) = f(\boldsymbol{x}; y, \boldsymbol{z}, \boldsymbol{\theta}) \qquad (3)$$

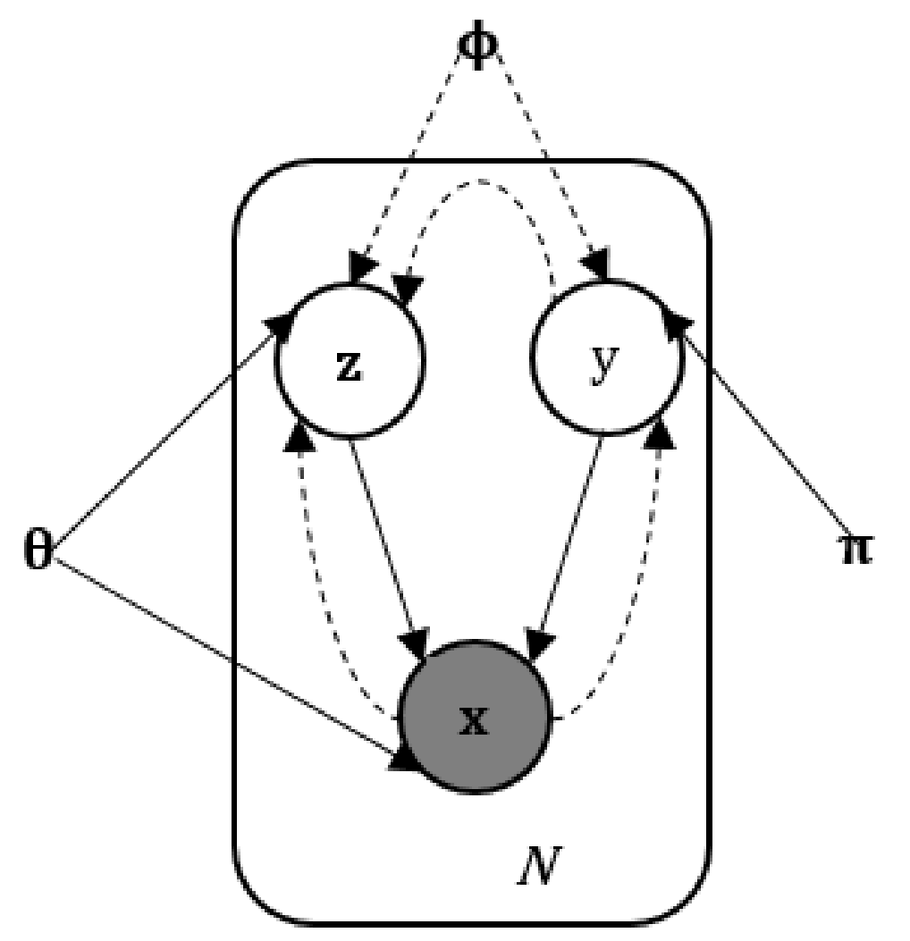

Fig. 2: The M2 graphical model. $f(\cdot)$ is a Gaussian distribution parametrized by MLPs.

The system can be trained by optimizing the variational lower bound. For data with observed labels:

$$\log p_\theta(\boldsymbol{x}, y) \geqslant \mathcal{L}(\boldsymbol{x}, y) = \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x}, yy)}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z}, y) \\ + \log p_\theta(y) + \log p(\boldsymbol{z}) - \log q_\phi(\boldsymbol{z}|y, \boldsymbol{x})] \qquad (4)$$

For unlabeled data:

$$\log p_\theta(\boldsymbol{x}) \geqslant -\mathcal{U}(\boldsymbol{x}) = \mathbb{E}_{q_\phi(\boldsymbol{z}, y|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z}, y) \\ + \log p_\theta(y) + \log p(\boldsymbol{z}) - \log q_\phi(y, \boldsymbol{z}|\boldsymbol{x})] \qquad (5)$$

Thus, putting them together and adding an extra term accounting for the effect of labeled data to discriminative distribution $q_\phi(y|\boldsymbol{x})$:

$$\mathcal{J}^\alpha = \Sigma_{(\boldsymbol{x}, y)\sim\tilde{p}_l}\mathcal{L}(\boldsymbol{x}, y) + \Sigma_{\boldsymbol{x}\sim\tilde{p}_u}\mathcal{U}(\boldsymbol{x}) \\ + \alpha \mathop{\mathbb{E}}_{p_l(\boldsymbol{x}, y)}[-\log q_\phi(y|\boldsymbol{x})] \qquad (6)$$

Stacked Model (M1+M2)

The two models can be stacked together for better performance. M1 will first map raw data to its embedding space and then use generative M2 to perform classification on this latent space. The model can be described as:

$$p_\theta(\boldsymbol{x}, y, \boldsymbol{z}_1, \boldsymbol{z}_2) = p(y)p(\boldsymbol{z}_2)p_\theta(\boldsymbol{z}_1|y, \boldsymbol{z}_2)p_\theta(\boldsymbol{x}|\boldsymbol{z}_1) \quad (7)$$

where raw data $\boldsymbol{x}$ is replaced with its latent representation $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ acts as $\boldsymbol{z}$ in M2 model. $p_\theta(\boldsymbol{z}_1|y, \boldsymbol{Z}_2)$ and $p_\theta(\boldsymbol{x}|\boldsymbol{z}_1)$ are parametrized as MLPs.

Scalable Variational Inference

Due to the non-linearity and non-conjugated dependencies in the model, exact posterior cannot be obtained, instead, recent advances in variational inference can be used to approximate the marginal and posterior distribution by optimizing the variational lower bound. Equations 2 and 6 show the lower bound for M1 and M2 model. The reparameterization trick and Monte Carlo approximation can be used to estimate the objective function. This approach is referred as *stochastic gradient variational Bayes* (SGVB). For example, for M1 we can write:

$$\mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})] = \mathbb{E}_{\mathcal{N}(\boldsymbol{\epsilon}|\mathbf{0}, \boldsymbol{I})}[p_\theta(\boldsymbol{x}|\boldsymbol{\mu}_\phi(\boldsymbol{x}) + \boldsymbol{\sigma}_\phi(\boldsymbol{x}) \odot \boldsymbol{\epsilon})] \qquad (8)$$

Thus, the generative parameter $\boldsymbol{\theta}$ and variational parameter $\boldsymbol{\phi}$ can be easily computed by gradient based optimization method.

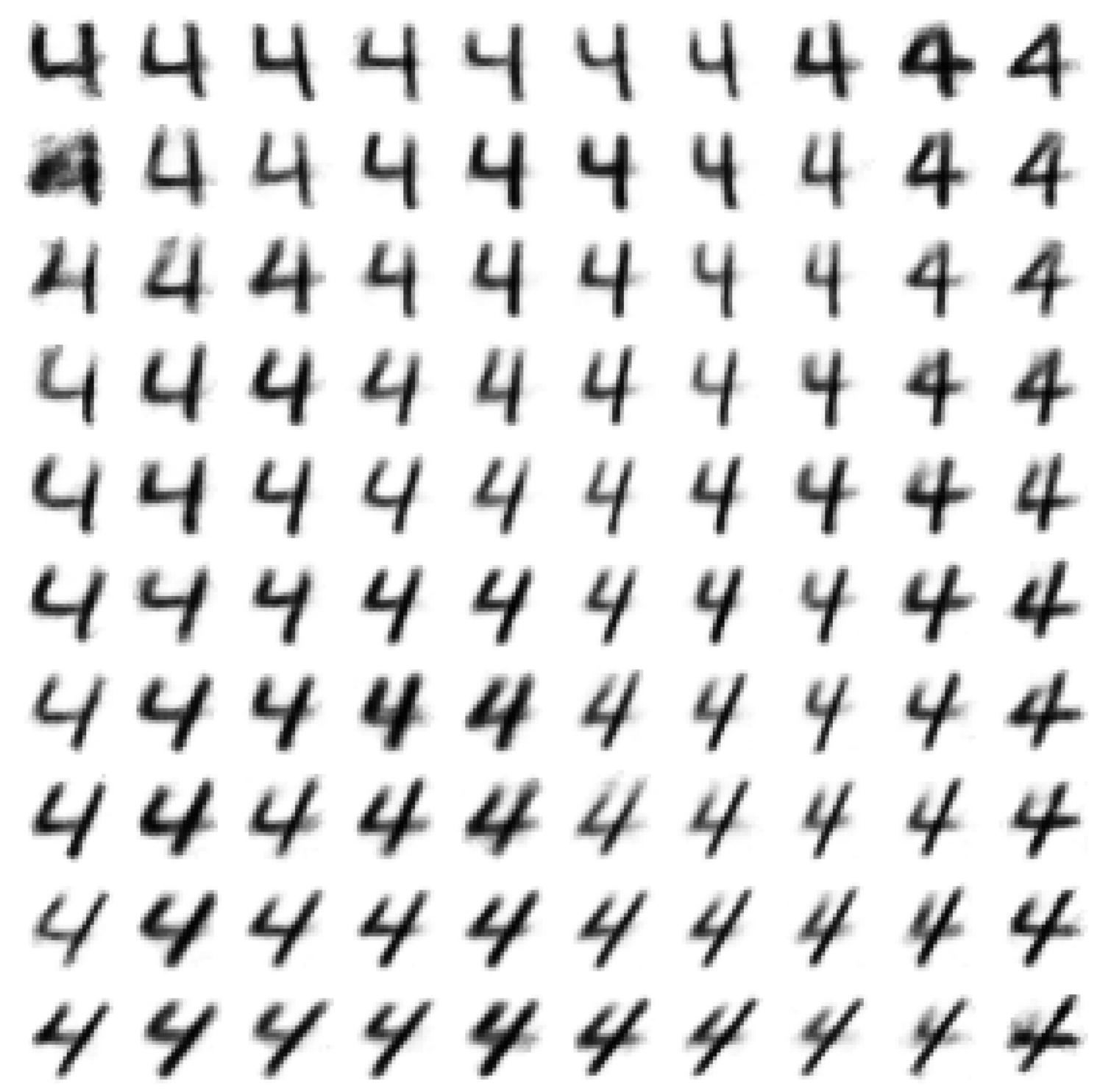

Fig. 3: Visualisation of samples from **x** generated by fixing the class label y and varying the 2D latent variable **z** on MNIST.

Results

N	M1+TSVM	M1+M2
100	20.82 (± 2.49)	4.68 (± 0.22)
300	4.98 (± 0.57)	2.49 (± 0.07)
1000	3.68 (± 0.28)	2.57 (± 0.23)
3000	2.36 (± 0.09)	2.27 (± 0.03)

Table 1: Semi-supervised classification on MNIST. M1+M2 results after 1000 epochs averaged over 3 runs.

Experiments

Performance of the stacked M1+M2 model is evaluated using the MNIST dataset split into labeled (N) and unlabeled data points. Both M1 and M2 models use 50-dimensional latent space, 500 hidden units per layer and softplus activation function. M1 and M2 models use respectively two and one hidden layers. Transductive SVM (TSVM) and the M2 model are trained on the reduced-dimension **z** from M1. Table 1 shows significant improvement of the stacked M1+M2 model with respect to TSVM on small number of labeled points N. In addition, Fig. 3 illustrates that points close in latent space in fact have similar writing style. Finally, Fig. 4 visualizes SVHN images generated using inferred style **z** from a sample image and varying class y. This demonstrates separation of latent style and class labels.

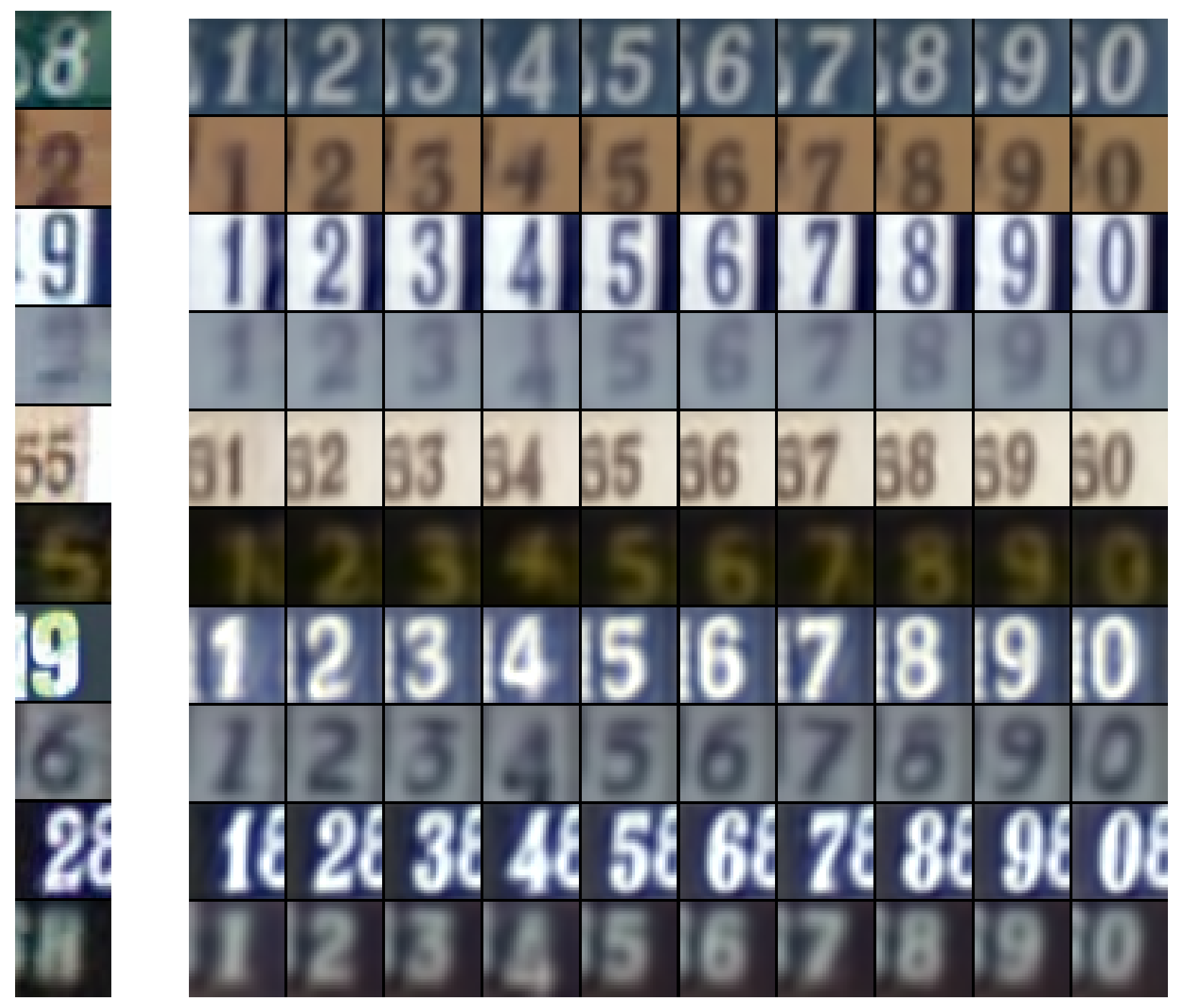

Fig. 4: SVHN analogies (right) generated from samples (left).

References

[1] D. Kingma, D. Rezende, S. Mohamed, M. Welling. *Semi-supervised Learning with Deep Generative Models.* arXiv:1406.5298, 2014.