Ladder Variational Autoencoders

Casper Kaae Sø nderby, Tapani Raiko, Lars Maalø e, Søren Kaae Sø nderby, and Ole Winther.

Summary

- Variational autoencoders are powerful models for unsupervised learning
- VAE, consisting of hierarchies of conditional stochastic variables, are highly expressive models retaining the computational efficiency of fully factorized models
- LVAE allows the parameterization to interact between the bottom-up and top-down signals
- LVAE improves the generative performance achieving as good or better performance than other
- We may extend its functionality to reduce the dimension of inputs for reinforcement learning trading strategies. Overall It may help us to build a more robust trading system.

Methods – Generative Model

- VAEs and LVAEs simultaneously train a generative model $p_\theta(\mathbf{x}, \mathbf{z}) = p_\theta(\mathbf{x}|\mathbf{z})p_\theta(\mathbf{z})$ (x as data, z as the latent variables), and an inference model $q_\phi(\mathbf{z}|\mathbf{x})$ by optimizing a variational lower bound to the likelihood
- In generative model p_θ, there are L layers z_i, *i = 1…L,* and each stochastic layer is a fully factorized Gaussian distribution conditioned on the layer above: $p_\theta(\mathbf{x}) = \int p_\theta(\mathbf{x}, \mathbf{z})d\mathbf{z}.$

$$p_\theta(\mathbf{z}) = p_\theta(\mathbf{z}_L) \prod_{i=1}^{L-1} p_\theta(\mathbf{z}_i|\mathbf{z}_{i+1})$$

$$p_\theta(\mathbf{z}_i|\mathbf{z}_{i+1}) = \mathcal{N}\left(\mathbf{z}_i|\mu_{p,i}(\mathbf{z}_{i+1}), \sigma^2_{p,i}(\mathbf{z}_{i+1})\right), \quad p_\theta(\mathbf{z}_L) = \mathcal{N}\left(\mathbf{z}_L|\mathbf{0}, \mathbf{I}\right)$$

$$p_\theta(\mathbf{x}|\mathbf{z}_1) = \mathcal{N}\left(\mathbf{x}|\mu_{p,0}(\mathbf{z}_1), \sigma^2_{p,0}(\mathbf{z}_1)\right) \text{ or } P_\theta(\mathbf{x}|\mathbf{z}_1) = \mathcal{B}\left(\mathbf{x}|\mu_{p,0}(\mathbf{z}_1)\right)$$

- Loss function:

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

$$= -KL(q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z})) + E_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right]$$

- Reparameterization trick for stochastic backpropagation $z = \mu + \sigma \odot \varepsilon$

where $\varepsilon \sim$ Normal(0,1)

VAE – Inference Model

- VAE inference models are parameterized as a bottom-up process. Conditioned on the stochastic layer below each stochastic layer is specified as a fully factorized Gaussian distribution:

$$q_\phi(\mathbf{z}|\mathbf{x}) = q_\phi(\mathbf{z}_1|\mathbf{x}) \prod_{i=2}^{L} q_\phi(\mathbf{z}_i|\mathbf{z}_{i-1})$$
$$q_\phi(\mathbf{z}_1|\mathbf{x}) = \mathcal{N}\left(\mathbf{z}_1|\mu_{q,1}(\mathbf{x}), \sigma^2_{q,1}(\mathbf{x})\right)$$
$$q_\phi(\mathbf{z}_i|\mathbf{z}_{i-1}) = \mathcal{N}\left(\mathbf{z}_i|\mu_{q,i}(\mathbf{z}_{i-1}), \sigma^2_{q,i}(\mathbf{z}_{i-1})\right), \; i = 2 \ldots L.$$

- μ() and σ^2() are implemented as:

$$\mathbf{d}(\mathbf{y}) = \texttt{MLP}(\mathbf{y})$$
$$\mu(\mathbf{y}) = \texttt{Linear}(\mathbf{d}(\mathbf{y}))$$
$$\sigma^2(\mathbf{y}) = \texttt{Softplus}(\texttt{Linear}(\mathbf{d}(\mathbf{y})))$$

Where MLP is multilayer perceptron, Linear is a single linear layer and softplus is nonlinearity function to ensure positive variances

LVAE – Inference Model

- A new inference model that recursively corrects the generative distribution with a data dependent approximate likelihood term:

$$\begin{aligned}
\mathbf{d}_n &= \texttt{MLP}(\mathbf{d}_{n-1}) \\
\hat{\mu}_{q,i} &= \texttt{Linear}(\mathbf{d}_i), i = 1 \dots L \\
\hat{\sigma}^2_{q,i} &= \texttt{Softplus}(\texttt{Linear}(\mathbf{d}_i)), i = 1 \dots L
\end{aligned}$$

Where $d_0 = x$

- This is followed by a stochastic downward pass recursively computing both the approximate posterior and generative distribution:
  - Together these form the approximate posterior distribution

$$q_\phi(\mathbf{z}|\mathbf{x}) = q_\phi(\mathbf{z}_L|\mathbf{x}) \prod_{i=1}^{L-1} q_\phi(\mathbf{z}_i|\mathbf{z}_{i+1}, \mathbf{x})$$

$$\sigma_{q,i} = \frac{1}{\hat{\sigma}^{-2}_{q,i} + \sigma^{-2}_{p,i}}$$

$$\mu_{q,i} = \frac{\hat{\mu}_{q,i}\hat{\sigma}^{-2}_{q,i} + \mu_{p,i}\sigma^{-2}_{p,i}}{\hat{\sigma}^{-2}_{q,i} + \sigma^{-2}_{p,i}}$$

$$q_\phi(\mathbf{z}_i|\cdot) = \mathcal{N}\left(\mathbf{z}_i|\mu_{q,i}, \sigma^2_{q,i}\right),$$

Warm-up trick

- The variational regularization term causes some of the latent units to become uninformative during training because the approximate posterior for unit k, is regularized towards its own prior p. From observation, it is presumably trapped in a local minima or saddle point at KL(q|p) =0.

$$\mathcal{L}(\theta, \phi; \mathbf{x})_{WU} = -\beta KL(q_\phi(z|x)||p_\theta(\mathbf{z})) + E_{q_\phi(z|x)}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right]$$

- One way to avoid this problem is to add a beta variable and increased linearly from 0 to 1 during the first N_t epochs of training.

Experiments

- Data set: MNIST, OMNIGLOT and NORB
- Largest models trained used a hierarchy of five layers of stochastic latent variables of sizes 64, 32,16,8 and 4, going from bottom to top.
- MLP's with two layers of deterministic hidden units

Result – log-likelihood performance

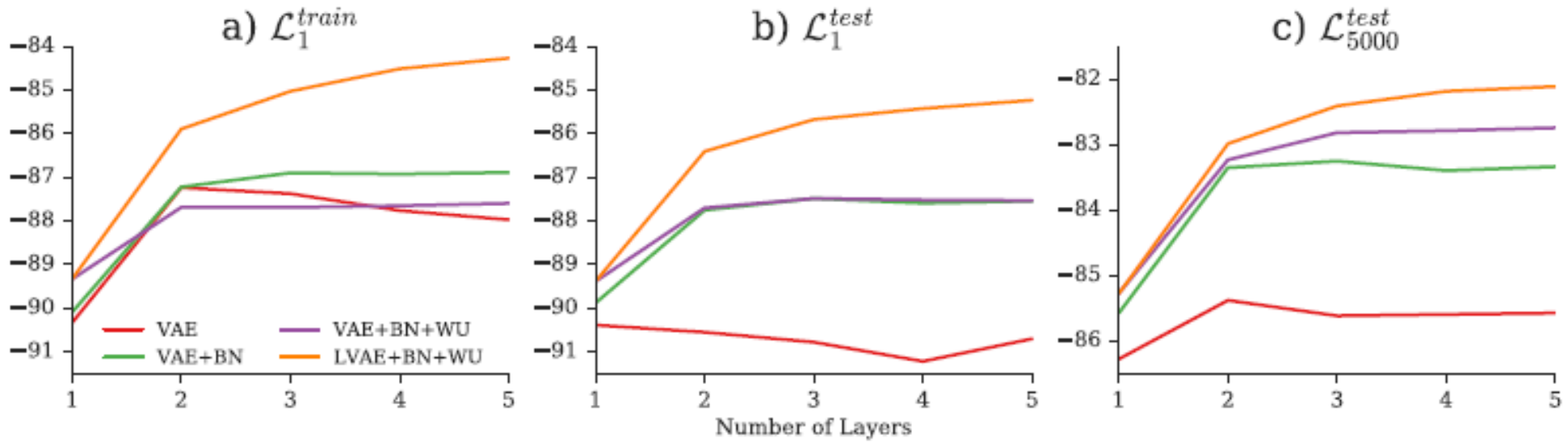

Figure 3: MNIST log-likelihood values for VAEs and the LVAE model with different number of latent layers, Batch-normalization (*BN*) and Warm-up (*WU*). a) Train log-likelihood, b) test log-likelihood and c) test log-likelihood with 5000 importance samples.

Comparison

- Comparing the results with current state of the art results on permutation invariant MNIST
- VAE+NF(normalizing flow VAE), IWAE(importance weighted VAE), VAE+VGP(Variational Gaussian Process)

	$\leq \log p((x))$
VAE 1-layer + NF [18]	-85.10
IWAE, 2-layer + IW=1 [3]	-85.33
IWAE, 2-layer + IW=50 [3]	-82.90
VAE, 2-layer + VGP [21]	-81.90
LVAE, 5-layer	-82.12
LVAE, 5-layer + finetuning	-81.84
LVAE, 5-layer + finetuning + IW=10	-81.74

Examination of Warm-up affect

- For models with warm-up we initially see many active units which are then gradually pruned away as the variational regularization term is introduced
- At the end of training warm-up results in more active units indicating a more distributed representation and further that the LVAE model produces both the deepest and most distributed latent representation

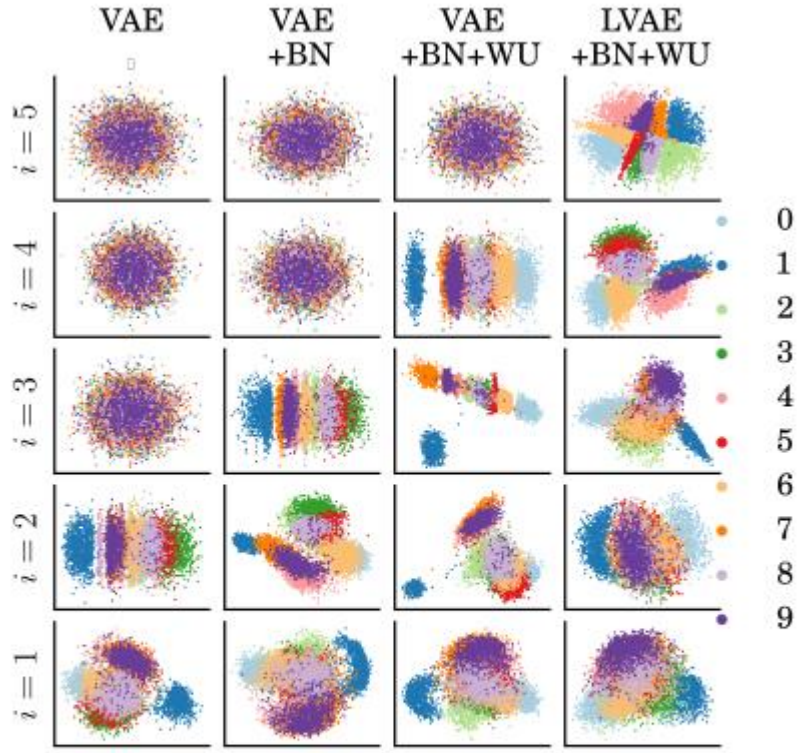

Figure 6: PCA-plots of samples from $q(z_i|z_{i-1})$ for 5-layer VAE and LVAE models trained on MNIST. Color-coded according to true class label

Conclusion

- This paper introduced a new inference model for VAEs combining a bottom-up data dependent approximate likelihood term with prior information the generative distribution
- It increases the approximated log-likelihood compared to VAEs; it provides a tighter bound on the log-likelihood; it learns a deeper and qualitatively different latent representation of the data
- Warm-up and batch-normalization are important for optimizing deep VAEs and LVAEs.
- Further work: combining with other models (Normalizing flow, Variational Gaussian Process, Auxiliary Deep Generative models); semi-supervised learning

Reference

- Casper Kaae Sø nderby, Tapani Raiko, Lars Maalø e, Søren Kaae Sø nderby, and Ole Winther. Ladder variational autoencoders. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 3738–3746. Curran Associates, Inc., 2016.