Learning Hierarchical Features from Generative Models

Shengjia Zhao, Jiaming Song, Stefano Ermon
Stanford University, *ICML 2017*

Presented by Zhe Gan, Duke University

September 22nd, 2017

Main idea

- Background:
  - Deep neural nets have been shown to be very successful at learning feature hierarchies in supervised learning tasks.
  - Generative models, on the other hand, have benefited less from hierarchical models with multiple layers of latent variables.
- Contribution:
  - proved that hierarchical latent variable models do not take advantage of the hierarchical structure when trained with existing variational methods.
  - hence provide limitations on the kind of features existing models can learn.
  - proposed variational ladder autoencoder to learn interpretable and disentangled hierarchical features.

Problem setting: variational autoencoder

- Generative model: $p_\theta(\boldsymbol{x}|\boldsymbol{z})$
- Inference model: $q_\phi(\boldsymbol{z}|\boldsymbol{x})$
- Variational lower bound:

$$\log p(\boldsymbol{x}) \geq \mathcal{L}(\boldsymbol{x}; \theta, \phi) \tag{1}$$

$$= \mathbb{E}_{p_{data}(\boldsymbol{x})}[\mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}[\log p(\boldsymbol{x}|\boldsymbol{z})] - KL(q(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z}))] \tag{2}$$

- Hierarchical VAE:

$$p(\boldsymbol{x}, \boldsymbol{z}) = p(\boldsymbol{x}|\boldsymbol{z}_1) \prod_{\ell=1}^{L-1} p(\boldsymbol{z}_\ell|\boldsymbol{z}_{\ell+1}) p(\boldsymbol{z}_L) \tag{3}$$

Limitations of Hierarchical VAEs

- Representational Efficiency: one layer model is enough for generative modeling.

Proposition

$\mathcal{L}_{ELBO}$ is globally maximized as a function of $q_{z|x}$ and $p(\boldsymbol{x}|\boldsymbol{z})$ when $\mathcal{L}_{ELBO} = -H(p_{data}(\boldsymbol{x}))$. If $\mathcal{L}_{ELBO}$ is globally maximized, the following Gibbs sampling chain converges to $p_{data}(\boldsymbol{x})$ if it is ergodic

$$\boldsymbol{z}_1^{(t)} \sim q(\boldsymbol{z}_1|\boldsymbol{x}^{(t)}) \tag{4}$$

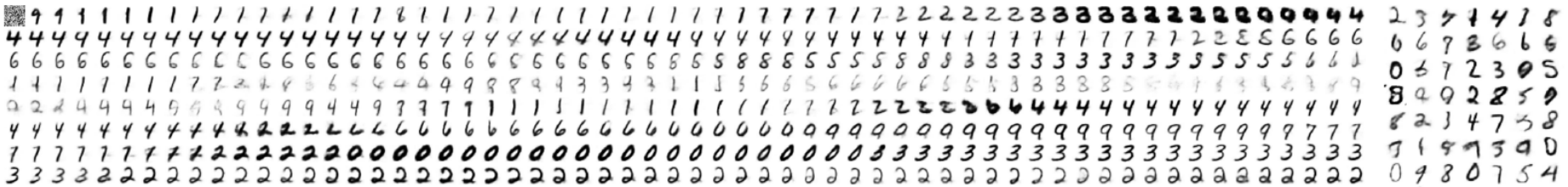
$$\boldsymbol{x}^{(t+1)} \sim p(\boldsymbol{x}|\boldsymbol{z}_1^{(t)}) \tag{5}$$

- Proof: under ideal optimization of $\mathcal{L}_{ELBO}$, $p(\boldsymbol{x}) = \int_{\boldsymbol{z}} p(\boldsymbol{x}, \boldsymbol{z}) d\boldsymbol{z} = p_{data}(\boldsymbol{x})$ and $q(\boldsymbol{z}|\boldsymbol{x}) = p(\boldsymbol{z}|\boldsymbol{x})$, this also implies $q(\boldsymbol{x}|\boldsymbol{z}_1) = p(\boldsymbol{x}|\boldsymbol{z}_1)$. Because the following Gibbs chain converges to $p_{data}(\boldsymbol{x})$ when it is ergodic

$$\boldsymbol{z}_1^{(t)} \sim q(\boldsymbol{z}_1|\boldsymbol{x}^{(t)}), \boldsymbol{x}^{(t+1)} \sim q(\boldsymbol{x}|\boldsymbol{z}_1^{(t)}) \tag{6}$$

Limitations of Hierarchical VAEs

- Under the assumptions of the above Proposition, we can sample from $p_{data}(\boldsymbol{x})$ without using the latent code $(\boldsymbol{z}_2, \ldots, \boldsymbol{z}_L)$ at all.
- The Gibbs chain generates samples (left figure) with similiar visual quality as ancestral sampling with the entire model (right figure), even though the Gibbs chain only used the bottom layer of the model

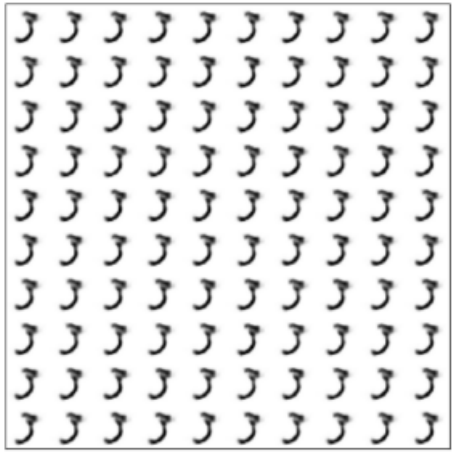
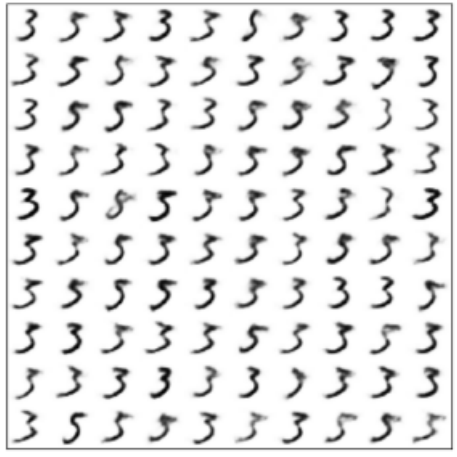


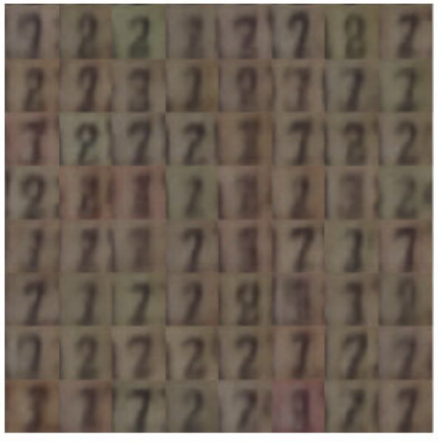
Limitations of Hierarchical VAEs

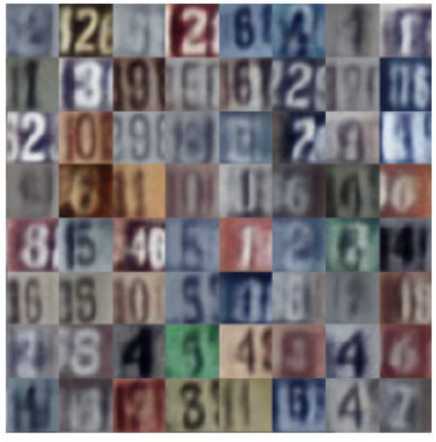
- Feature learning: no hierarchical and disentangled features are learned in a typical HVAE.
- We can consider $q(\boldsymbol{z}|\boldsymbol{x})$ as a (probabilistic) feature detector. It is natural to think that q might learn hierarchical features similarly to a feed-forward network $\boldsymbol{x} \to \boldsymbol{z}_1 \to \cdots \to \boldsymbol{z}_L$
- $q(\boldsymbol{z}_{>\ell}|\boldsymbol{z}_\ell)$ maps low-level features to high-level features, then $q(\boldsymbol{z}_\ell|\boldsymbol{z}_{>\ell})$ do the reverse.
- Under ideal optimization of $\mathcal{L}_{ELBO}$, $q(\boldsymbol{z}_\ell|\boldsymbol{z}_{>\ell}) = p(\boldsymbol{z}_\ell|\boldsymbol{z}_{>\ell})$
- $p(\boldsymbol{z}_\ell|\boldsymbol{z}_{>\ell})$ is typically a simple distribution family such as Gaussians, therefore, under this case, the only type of feature hierarchy we can hope to learn is one under which $q(\boldsymbol{z}_\ell|\boldsymbol{z}_{>\ell})$ is Gaussian.

Limitations of Hierarchical VAEs

Only very minor variations correspond to lower layers (left and center), and almost all the variation is represented by the top layer (right).

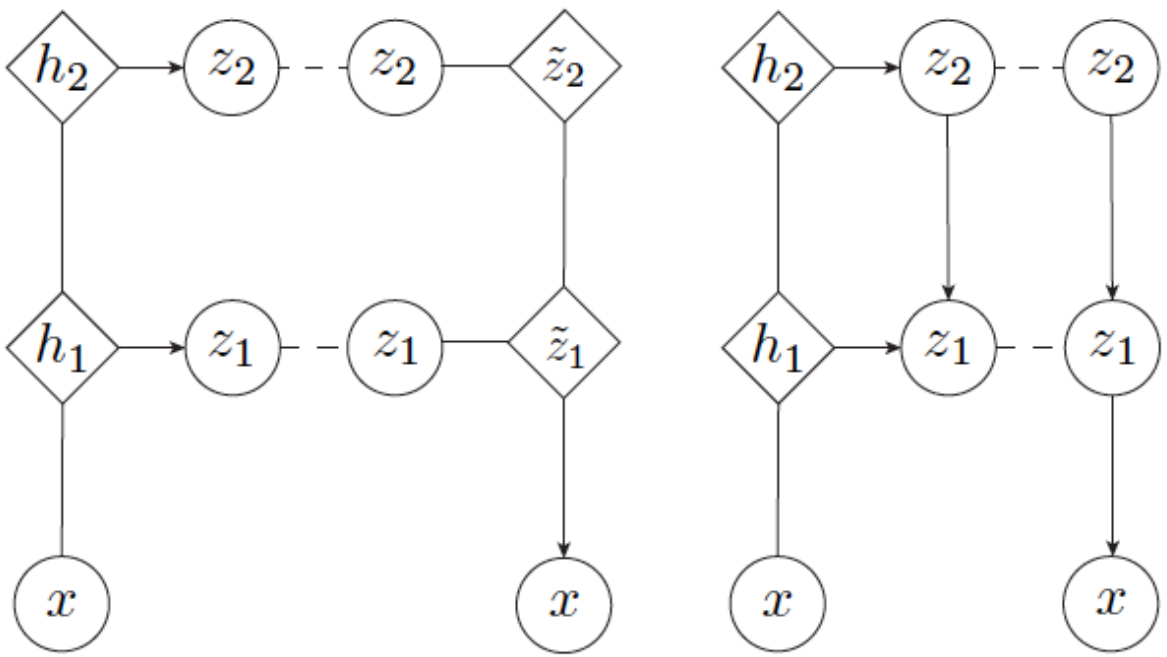







Variational Ladder Autoencoders

- Proposed: one-layer model that learns feature hierarchies
- Intuition: If $\boldsymbol{z}_i$ is more abstract than $\boldsymbol{z}_j$, then the inference mapping $q(\boldsymbol{z}_i|\boldsymbol{x})$ and generative mapping when other layers are fixed $p(\boldsymbol{x}|\boldsymbol{z}_i, \boldsymbol{z}_{-i} = \text{fixed})$ requires a more expressive network to capture.
- (Left) VLAE; (Right) LVAE.



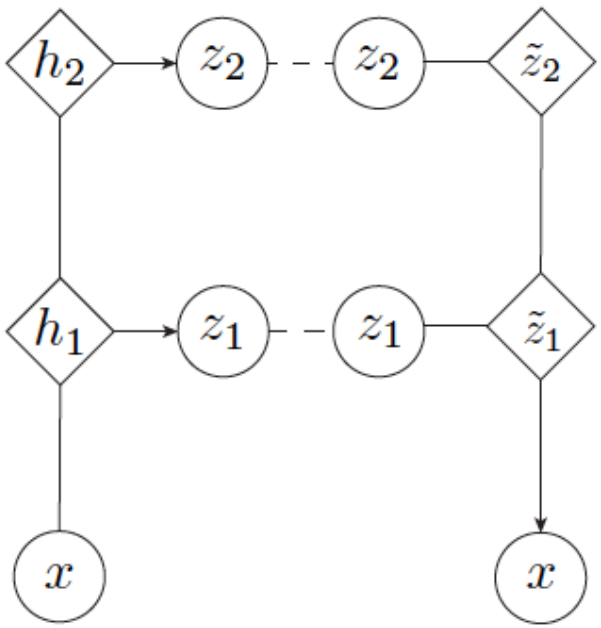
Variational Ladder Autoencoders

- Generative model:
  $p(\boldsymbol{z}) = p(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_L) = \mathcal{N}(0, \mathbf{I})$.
  $p(\boldsymbol{x}|\boldsymbol{z}_1, \ldots, \boldsymbol{z}_L)$ is defined as

$$\tilde{\boldsymbol{z}}_L = f_L(\boldsymbol{z}_L) \tag{7}$$
$$\tilde{\boldsymbol{z}}_\ell = f_\ell(\tilde{\boldsymbol{z}}_{\ell+1}, \boldsymbol{z}_\ell) \tag{8}$$
$$\boldsymbol{x} \sim r(\boldsymbol{x}; f_0(\tilde{\boldsymbol{z}}_1)) \tag{9}$$

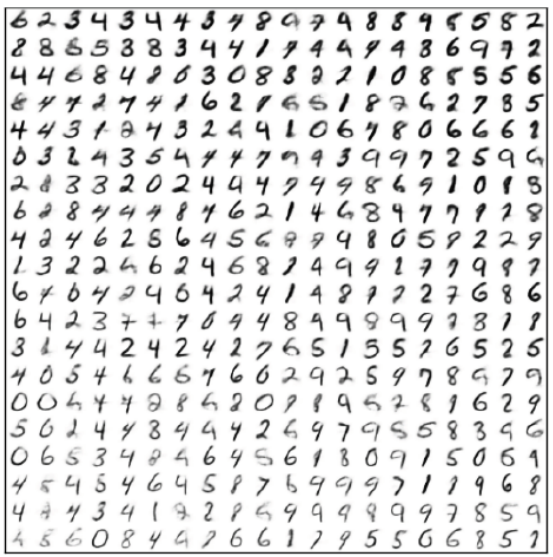
- Inference network:

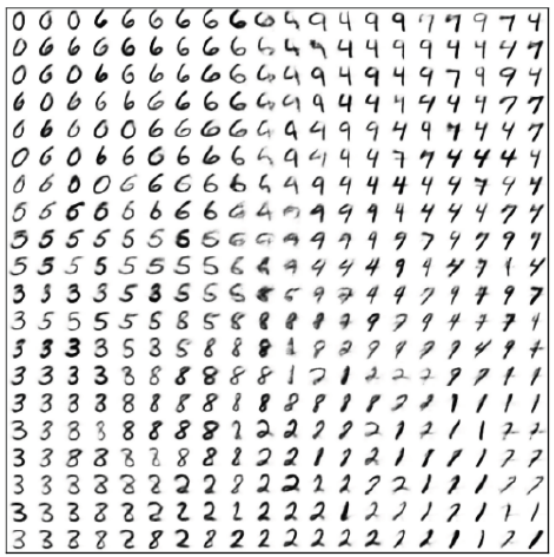
$$\boldsymbol{h}_\ell = g_\ell(\boldsymbol{h}_{\ell-1}) \tag{10}$$
$$\boldsymbol{z}_\ell \sim \mathcal{N}(\boldsymbol{\mu}_\ell(\boldsymbol{h}_\ell), \boldsymbol{\sigma}_\ell(\boldsymbol{h}_\ell)) \tag{11}$$



Experiments

1st layer encodes stroke width; 2nd layer encodes digit width and tilt; 3rd layer encodes digit identity.





Experiments

1st: color schemes; 2nd: shape variations; 3rd: digit identity; 4th: the general structure of the image.