Стохастическое обратное распространение ошибки и приближенный вариационный вывод в глубинных генеративных моделях

2014

Stochastic Backpropogation

Задача оптимизации в моделях с латентными переменными:

$$\max_{\theta} \mathbb{E}_{q(\xi|\theta)}[f(\xi)]$$

Stochastic Backpropogation

Задача оптимизации в моделях с латентными переменными:

$$\max_{\theta} \mathbb{E}_{q(\xi|\theta)}[f(\xi)]$$

Пример:

$$\begin{aligned}\log p(v, \theta^r, \theta^g) &= \log \sum_h q(h|\theta^r) \log p(v|h, \theta^g) \\ &= \mathbb{E}_{q(h|\theta^r)}[\log p(v|h, \theta^g)]\end{aligned}$$

Stochastic Backpropogation

Задача оптимизации в моделях с латентными переменными:

$$\max_{\theta} \mathbb{E}_{q(\xi|\theta)}[f(\xi)]$$

Пример:

$$\begin{aligned}\log p(v, \theta^r, \theta^g) &= \log \sum_h q(h|\theta^r) \log p(v|h, \theta^g) \\ &= \mathbb{E}_{q(h|\theta^r)}[\log p(v|h, \theta^g)]\end{aligned}$$

Проблема:

$$\nabla_\theta \mathbb{E}_{q(\xi|\theta)}[f(\xi)] - ?$$

Likelihood Maximization

$$
\begin{aligned}
\log p(\boldsymbol{v}) &= \sum_{\boldsymbol{h}} \log p(\boldsymbol{v}) q(\boldsymbol{h}) \\
&= \sum_{\boldsymbol{h}} q(\boldsymbol{h}) \log p(\boldsymbol{v}, \boldsymbol{h}) - \sum_{\boldsymbol{h}} q(\boldsymbol{h}) \log p(\boldsymbol{h}|\boldsymbol{v}) \\
&= \sum_{\boldsymbol{h}} q(\boldsymbol{h}) \log \frac{p(\boldsymbol{v}, \boldsymbol{h})}{q(\boldsymbol{h})} - \sum_{\boldsymbol{h}} q(\boldsymbol{h}) \log \frac{p(\boldsymbol{h}|\boldsymbol{v})}{q(\boldsymbol{h})} \\
&= \mathcal{L}(q) + \mathrm{KL}\left(q||p\right) \\
&= \sum_{\boldsymbol{h}} \log p(\boldsymbol{v}, \boldsymbol{h}) q(\boldsymbol{h}) + \mathcal{H}(q) + \mathrm{KL}\left(q||p\right)
\end{aligned}
$$

Gaussian Backpropogation

$$\nabla_\theta \, \mathbb{E}_{q(\xi|\theta)}[f(\xi)] - ?$$

Gaussian Backpropogation

$$\nabla_\theta \mathbb{E}_{q(\xi|\theta)}[f(\xi)] - ?$$

Предположим, что $q - \mathcal{N}(\xi|\boldsymbol{\mu}, \boldsymbol{C})$, тогда

$$\nabla_{\mu_i} \mathbb{E}_{\mathcal{N}(\boldsymbol{\xi}|\boldsymbol{\mu},\boldsymbol{C})}[f(\boldsymbol{\xi})] = \mathbb{E}_{\mathcal{N}(\boldsymbol{\xi}|\boldsymbol{\mu},\boldsymbol{C})}[\nabla_{\xi_i} f(\boldsymbol{\xi})]$$

$$\nabla_{C_{i,j}} \mathbb{E}_{\mathcal{N}(\boldsymbol{\xi}|\boldsymbol{\mu},\boldsymbol{C})}[f(\boldsymbol{\xi})] = \frac{1}{2} \mathbb{E}_{\mathcal{N}(\boldsymbol{\xi}|\boldsymbol{\mu},\boldsymbol{C})}[\nabla^2_{\xi_i,\xi_j} f(\boldsymbol{\xi})]$$

Gaussian Backpropogation

$$\nabla_\theta \mathbb{E}_{q(\xi|\theta)}[f(\xi)] - ?$$

Предположим, что q — $\mathcal{N}(\xi|\boldsymbol{\mu}, \boldsymbol{C})$, тогда

$$\nabla_{\mu_i} \mathbb{E}_{\mathcal{N}(\boldsymbol{\xi}|\boldsymbol{\mu},\boldsymbol{C})}[f(\boldsymbol{\xi})] = \mathbb{E}_{\mathcal{N}(\boldsymbol{\xi}|\boldsymbol{\mu},\boldsymbol{C})}[\nabla_{\xi_i} f(\boldsymbol{\xi})]$$

$$\nabla_{C_{i,j}} \mathbb{E}_{\mathcal{N}(\boldsymbol{\xi}|\boldsymbol{\mu},\boldsymbol{C})}[f(\boldsymbol{\xi})] = \frac{1}{2} \mathbb{E}_{\mathcal{N}(\boldsymbol{\xi}|\boldsymbol{\mu},\boldsymbol{C})}[\nabla^2_{\xi_i,\xi_j} f(\boldsymbol{\xi})]$$

Если $\boldsymbol{\mu}$ и $\boldsymbol{C}$ зависят от параметра θ, то

$$\nabla_{\boldsymbol{\theta}} \mathbb{E}_{\mathcal{N}(\boldsymbol{\xi}|\boldsymbol{\mu},\boldsymbol{C})}[f(\boldsymbol{\xi})] = \mathbb{E}_{\mathcal{N}(\boldsymbol{\xi}|\boldsymbol{\mu},\boldsymbol{C})} \left[\boldsymbol{g}^T \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\theta}} + \frac{1}{2} Tr \left(\boldsymbol{H} \frac{\partial \boldsymbol{C}}{\partial \boldsymbol{\theta}} \right) \right],$$

где $\boldsymbol{g}$ — градиент $f(\boldsymbol{\xi})$, H — гессиан $f(\boldsymbol{\xi})$.

Gaussian Backpropogation

$$
\begin{aligned}
\nabla_{\mu_i} \mathbb{E}_{\mathcal{N}(\boldsymbol{\xi}|\boldsymbol{\mu},\boldsymbol{C})}[f(\boldsymbol{\xi})] &= \int \nabla_{\mu_i} \mathcal{N}(\boldsymbol{\xi}|\boldsymbol{\mu},\boldsymbol{C}) f(\boldsymbol{\xi}) d\boldsymbol{\xi} \\
&= -\int \nabla_{\xi_i} \mathcal{N}(\boldsymbol{\xi}|\boldsymbol{\mu},\boldsymbol{C}) f(\boldsymbol{\xi}) d\boldsymbol{\xi} \\
&= -\left[\int \mathcal{N}(\boldsymbol{\xi}|\boldsymbol{\mu},\boldsymbol{C}) f(\boldsymbol{\xi}) d\boldsymbol{\xi}_{\neg i} \right]_{\xi_i=-\infty}^{\xi_i=+\infty} \\
&+ \int \mathcal{N}(\boldsymbol{\xi}|\boldsymbol{\mu},\boldsymbol{C}) \nabla_{\xi_i} f(\boldsymbol{xi}) d\boldsymbol{\xi} \\
&= \mathbb{E}_{\mathcal{N}(\boldsymbol{\xi}|\boldsymbol{\mu},\boldsymbol{C})}[\nabla_{\xi_i} f(\boldsymbol{\xi})]
\end{aligned}
$$

$$
\nabla_{\mu_i} \mathcal{N}(\boldsymbol{\xi}|\boldsymbol{\mu},\boldsymbol{C}) = -\nabla_{\xi_i} \mathcal{N}(\boldsymbol{\xi}|\boldsymbol{\mu},\boldsymbol{C})
$$

$$
\nabla_{C_{i,j}} \mathcal{N}(\boldsymbol{\xi}|\boldsymbol{\mu},\boldsymbol{C}) = \frac{1}{2} \nabla_{\xi_i,\xi_j} \mathcal{N}(\boldsymbol{\xi}|\boldsymbol{\mu},\boldsymbol{C})
$$

Stochastic Backpropogation

Какие рапсределния q можно использовать?
Два требования:

- $\nabla_\theta p(x|\theta) = \nabla_x p(x|\theta) B(x, \theta)$;
- $p(x, \theta) = 0$ на границах области определения x.

Экспоненциальное семейство:

$$p(x|\theta) = h(x) \exp(\eta(\theta)^T \phi(x) - A(\theta))$$

$$B(x, \theta) = \frac{\nabla_\theta \eta(\theta) \phi(x) - \nabla_{theta} A(\theta)}{\nabla_x \log h(x) + \eta(\theta)^T \nabla_x \phi(x)}$$

Преобразование координат

$$\boldsymbol{\xi} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{C})$$

$$\boldsymbol{\xi} = \boldsymbol{\mu} + \boldsymbol{R}\boldsymbol{\epsilon}, \text{ где } \boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{I}) \text{ и } \boldsymbol{C} = \boldsymbol{R}\boldsymbol{R}^T$$

$$\begin{aligned}\nabla_{\boldsymbol{R}} \mathbb{E}_{\mathcal{N}(\boldsymbol{\xi}|\boldsymbol{\mu},\boldsymbol{C})}[f(\xi)] &= \nabla_{\boldsymbol{R}} \mathbb{E}_{\mathcal{N}(\epsilon|0,\boldsymbol{I})}[f(\boldsymbol{\mu} + \boldsymbol{R}\boldsymbol{\epsilon})] \\ &= \mathbb{E}_{\mathcal{N}(\epsilon|0,\boldsymbol{I})}[\epsilon \boldsymbol{g}^T],\end{aligned}$$

где $\boldsymbol{g}$ — градиент f в точке $\boldsymbol{\mu} + \boldsymbol{R}\boldsymbol{\epsilon}$.

Deep Latent Gaussian Model

$$\boldsymbol{\xi}_l \sim \mathcal{N}(\boldsymbol{\xi}|0, \boldsymbol{I}), \qquad l = 1, \ldots, L;$$
$$\boldsymbol{h}_L = \boldsymbol{G}_L \boldsymbol{\xi}_L$$
$$\boldsymbol{h}_l = T_l(\boldsymbol{h}_{l+1}) + \boldsymbol{G}_l \boldsymbol{\xi}_l, \qquad l = 1, \ldots, L;$$
$$v \sim \pi(v|T_0(\boldsymbol{h}_l))$$

T_l — многослойный персептрон;

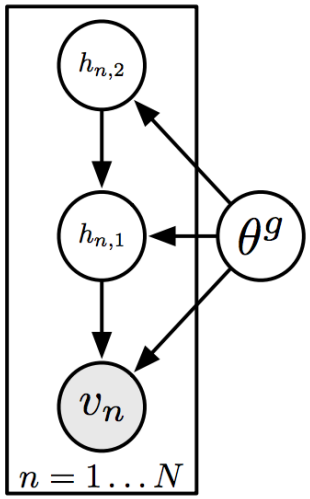
$\boldsymbol{G}_l$ — матрица.

Параметры генеративной части $\boldsymbol{\theta}^g$ — $(\{\boldsymbol{G}_l\}_{l=1}^K, \{T_l\}_{l=1}^L)$
Априорное распределение на параметры:
$p(\boldsymbol{\theta}^g) = \mathcal{N}(\boldsymbol{\theta}|0, \kappa \boldsymbol{I})$

Deep latent Gaussian Model

Совместное распределние можно записать двумя эквивалентными способами:

$$p(\boldsymbol{v}, \boldsymbol{h}) = p(\boldsymbol{v}|\boldsymbol{h}_1, \boldsymbol{\theta}^g) p(\boldsymbol{h}_L|\boldsymbol{\theta}^g) p(\boldsymbol{\theta}^g) \prod_{l=1}^{L-1} p(\boldsymbol{h}_l|\boldsymbol{h}_{l+1}, \boldsymbol{\theta}^g),$$



где $p(\boldsymbol{h}_l|\boldsymbol{h}_{l+1}, \boldsymbol{\theta}^g) = \mathcal{N}(\boldsymbol{h}_l|T_l(\boldsymbol{h}_{l+1}), \boldsymbol{G}_l \boldsymbol{G}_l^T)$

$$p(\boldsymbol{v}, \boldsymbol{\xi}) = p(\boldsymbol{v}|\boldsymbol{h}_1(\xi_{1,\dots,L}), \boldsymbol{\theta}^g) p(\boldsymbol{\theta}^g) \prod_{l=1}^{L} \mathcal{N}(\boldsymbol{\xi}_l|0, \boldsymbol{I})$$

Lower bound on the marginal likelihood

$\boldsymbol{V}$ — датасет.
$\boldsymbol{V} = \{\boldsymbol{v}_n\}_{n=1}^{N}$, где $\boldsymbol{v}_n = (v_{n,1}, \ldots, v_{n,D})$

$$
\begin{aligned}
-\mathcal{L}(\boldsymbol{V}) &= -\log \int p(\boldsymbol{V}|\boldsymbol{\xi}, \boldsymbol{\theta}^g) p(\boldsymbol{\xi}, \boldsymbol{\theta}^g) \\
&= -log \int \frac{q(\boldsymbol{\xi})}{q(\boldsymbol{\xi})} p(\boldsymbol{V}|\boldsymbol{\xi}, \boldsymbol{\theta}^g) p(\boldsymbol{\xi}, \boldsymbol{\theta}^g) \\
\leq \mathcal{F}(\boldsymbol{V}) &= KL[q(\boldsymbol{\xi}) \| p(\boldsymbol{\xi})] - \mathbb{E}_{q(\boldsymbol{\xi})}[\log p(\boldsymbol{V}|\boldsymbol{\xi}, \boldsymbol{\theta}^g) p(\boldsymbol{\theta}^g)]
\end{aligned}
$$

Approximate posterior

$q(\boldsymbol{\xi}|v)$ — распознающая часть модели.

$$q(\boldsymbol{\xi}|\boldsymbol{V}, \boldsymbol{\theta}^r) = \prod_{n=1}^{N} \prod_{l=1}^{L} \mathcal{N}(\xi_{n,l}|\boldsymbol{\mu}_l(v_n), \boldsymbol{C}_l(v_n)),$$

где $\boldsymbol{\mu}_l(\cdot)$ и $\boldsymbol{C}_l(\cdot)$ — функции, представленные глубинными нейросетями. Будем обозначать параметры распределния q — $\boldsymbol{\theta}^r$.

Lower bound of marginal likelihood

$$KL[\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{C}) \| \mathcal{N}(0, \boldsymbol{I})] = \frac{1}{2}\left(\mathrm{Tr}(\boldsymbol{C}) + \|\boldsymbol{\mu}\|^2 - k - \log|\boldsymbol{C}|\right)$$

$$\begin{aligned}
\mathcal{F}(\boldsymbol{V}) &= KL[q(\boldsymbol{\xi}) \| p(\boldsymbol{\xi})] - \mathbb{E}_{q(\boldsymbol{\xi})}[\log p(\boldsymbol{V}|\boldsymbol{\xi}, \boldsymbol{\theta}^g) p(\boldsymbol{\theta}^g)] \\
&= \frac{1}{2} \sum_{n,l} \left[\mathrm{Tr}(\boldsymbol{C}_{n,l}) + \|\boldsymbol{\mu}_{n,l}\|^2 - k_l - \log|\boldsymbol{C}_{n,l}|\right] \\
&\quad - \sum_n \mathbb{E}_{q(\boldsymbol{\xi}|\boldsymbol{v}_n)} \left[\log p(\boldsymbol{v}_n|\boldsymbol{h}(\boldsymbol{\xi}))\right] + \frac{1}{2\kappa}\|\boldsymbol{\theta}^g\|^2
\end{aligned}$$

Gradient of lower bound

$$\mathcal{F}(\boldsymbol{V}) = \frac{1}{2}\sum_{n,l}\left[\mathrm{Tr}(\boldsymbol{C}_{n,l}) + \|\boldsymbol{\mu}_{n,l}\|^2 - k_l - \log|\boldsymbol{C}_{n,l}|\right]$$
$$- \sum_{n}\mathbb{E}_{q(\boldsymbol{\xi}|\boldsymbol{v}_n)}\left[\log p(\boldsymbol{v}_n|\boldsymbol{h}(\boldsymbol{\xi}))\right] + \frac{1}{2\kappa}\|\boldsymbol{\theta}^g\|^2$$

$$\nabla_{\theta_j^g}\mathcal{F}(\boldsymbol{v}) = -\mathbb{E}_q\left[\nabla_{\theta_j^g}\log p(\boldsymbol{v}|\boldsymbol{h})\right] + \frac{1}{\kappa}\theta_j^g$$

Gradient of lower bound

$$\mathcal{F}(\boldsymbol{V}) = \frac{1}{2}\sum_{n,l}\left[\mathrm{Tr}(\boldsymbol{C}_{n,l}) + \|\boldsymbol{\mu}_{n,l}\|^2 - k_l - \log|\boldsymbol{C}_{n,l}|\right]$$
$$- \sum_{n}\mathbb{E}_{q(\boldsymbol{\xi}|\boldsymbol{v}_n)}\left[\log p(\boldsymbol{v}_n|\boldsymbol{h}(\boldsymbol{\xi}))\right] + \frac{1}{2\kappa}\|\boldsymbol{\theta}^g\|^2$$

$$\nabla_{\boldsymbol{\mu}_l}\mathcal{F}(\boldsymbol{v}) = -\mathbb{E}_q\left[\nabla_{\boldsymbol{\xi}_l}\log p(\boldsymbol{v}|\boldsymbol{h}(\boldsymbol{\xi}))\right] + \boldsymbol{\mu}_l$$

Gradient of lower bound

$$\mathcal{F}(\boldsymbol{V}) = \frac{1}{2}\sum_{n,l}\left[\mathrm{Tr}(\boldsymbol{C}_{n,l}) + \|\boldsymbol{\mu}_{n,l}\|^2 - k_l - \log|\boldsymbol{C}_{n,l}|\right]$$

$$- \sum_{n}\mathbb{E}_{q(\boldsymbol{\xi}|\boldsymbol{v}_n)}\left[\log p(\boldsymbol{v}_n|\boldsymbol{h}(\boldsymbol{\xi}))\right] + \frac{1}{2\kappa}\|\boldsymbol{\theta}^g\|^2$$

$$\boldsymbol{C}_l = \boldsymbol{R}_l\boldsymbol{R}_l^T$$

$$\nabla_{R_{l,i,j}}\mathcal{F}(\boldsymbol{v}) = -\mathbb{E}_q\left[\epsilon_{l,j}\nabla_{\xi_{l,i}}\log p(\boldsymbol{v}|\boldsymbol{h}(\boldsymbol{\xi}))\right]$$

$$+ \frac{1}{2}\nabla_{R_{l,i,j}}\left[\mathrm{Tr}\boldsymbol{C}_l - \log|\boldsymbol{C}_l|\right]$$

Gradient of lower bound

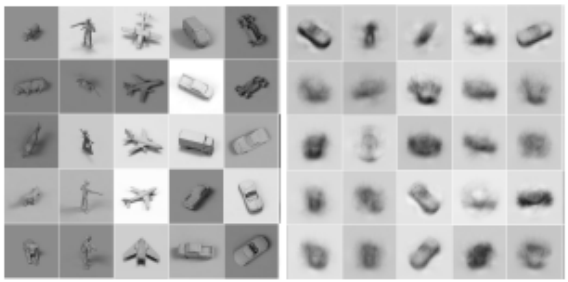
$$\mathcal{F}(\boldsymbol{V}) = \frac{1}{2}\sum_{n,l}\left[\mathrm{Tr}(\boldsymbol{C}_{n,l}) + \|\boldsymbol{\mu}_{n,l}\|^2 - k_l - \log|\boldsymbol{C}_{n,l}|\right]$$
$$- \sum_{n}\mathbb{E}_{q(\boldsymbol{\xi}|\boldsymbol{v}_n)}\left[\log p(\boldsymbol{v}_n|\boldsymbol{h}(\boldsymbol{\xi}))\right] + \frac{1}{2\kappa}\|\boldsymbol{\theta}^g\|^2$$

$$\nabla_{\boldsymbol{\theta}^r}\mathcal{F}(\boldsymbol{v}) = \nabla_{\boldsymbol{\mu}}\mathcal{F}(\boldsymbol{v})^T\frac{\partial\boldsymbol{\mu}}{\partial\boldsymbol{\theta}^r} + \mathrm{Tr}\left(\nabla_{\boldsymbol{R}}\mathcal{F}(\boldsymbol{v})\frac{\partial\boldsymbol{R}}{\partial\boldsymbol{\theta}^r}\right)$$

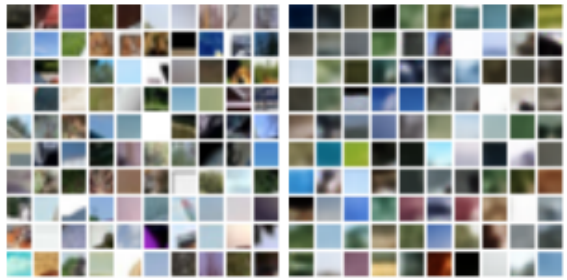
Алгоритм обучения

1: **while** hasNotConverged() **do**
2: $\quad \boldsymbol{V} \leftarrow$ getMiniBatch()
3: $\quad \boldsymbol{\xi}_n \sim q(\boldsymbol{\xi}_n | \boldsymbol{v}_n)$
4: $\quad \boldsymbol{h} \leftarrow \boldsymbol{h}(\boldsymbol{\xi})$
5: $\quad \triangle \boldsymbol{\theta}^{g,r} \leftarrow$ calcGradients()
6: $\quad \boldsymbol{\theta}^{g,r} \leftarrow \boldsymbol{\theta}^{g,r} + \triangle \boldsymbol{\theta}^{g,r}$
7: **end while**

Sampling from model



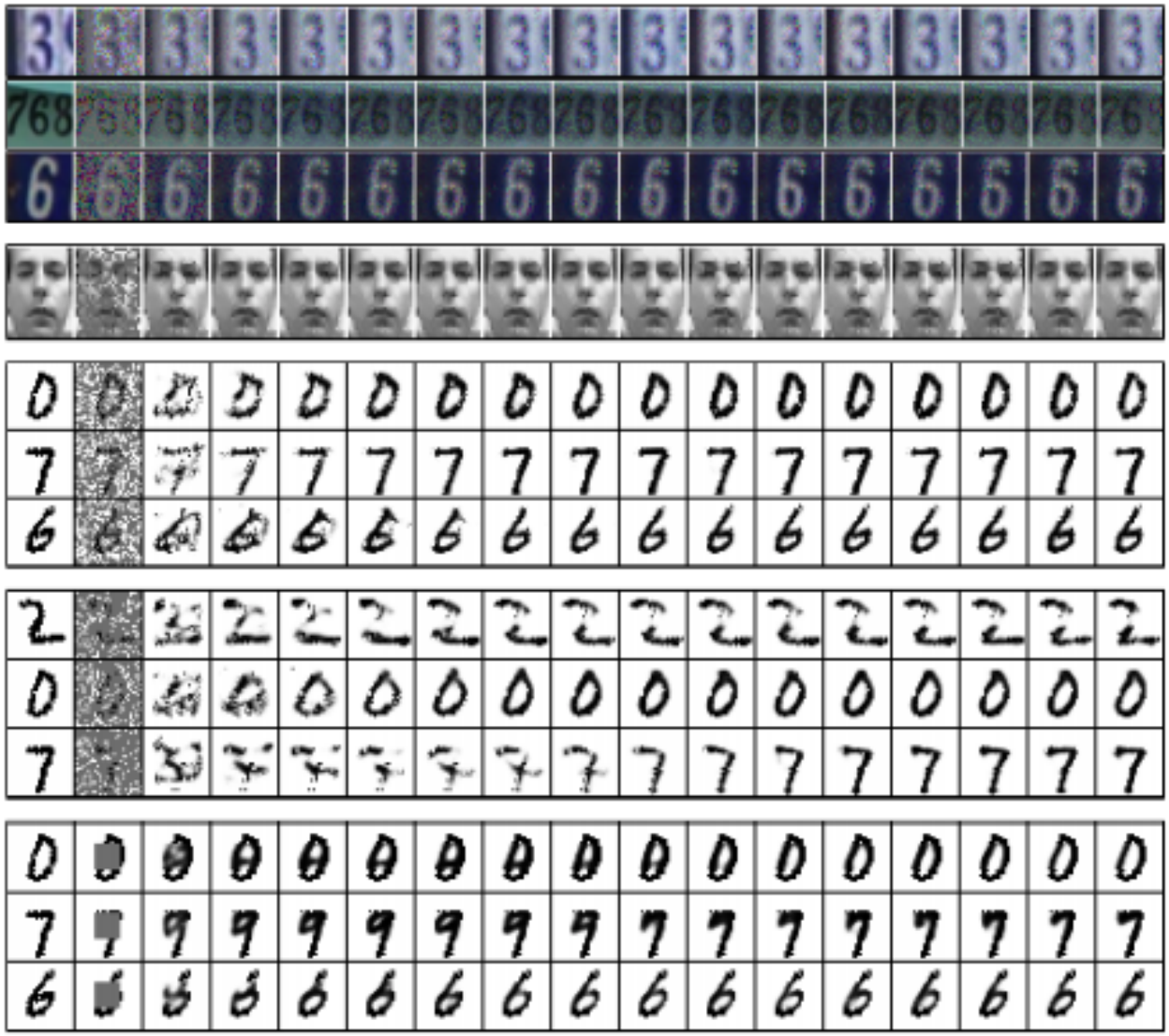
(a) NORB



(b) CIFAR

(c) Frey

Imputation problem



Probability of MNIST test set

Table 1. Comparison of negative log-probabilities on the test set for the binarised MNIST data.

Model	$-\ln p(\mathbf{v})$
Factor Analysis	106.00
NLGBN (Frey & Hinton, 1999)	95.80
Wake-Sleep (Dayan, 2000)	91.3
DLGM diagonal covariance	87.30
DLGM rank-one covariance	86.60
Results below from Uria et al. (2014)	
MoBernoullis K=10	168.95
MoBernoullis K=500	137.64
RBM (500 h, 25 CD steps) approx.	86.34
DBN 2hl approx.	84.55
NADE 1hl (fixed order)	88.86
NADE 1hl (fixed order, RLU, minibatch)	88.33
EoNADE 1hl (2 orderings)	90.69
EoNADE 1hl (128 orderings)	87.71
EoNADE 2hl (2 orderings)	87.96
EoNADE 2hl (128 orderings)	85.10