DO DEEP GENERATIVE MODELS KNOW WHAT THEY DON'T KNOW?

Авторы статьи Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, Balaji Lakshminarayanan
Презентация Dayana Savostianova

НИУ ВШЭ

27 02 2020

Мотивация

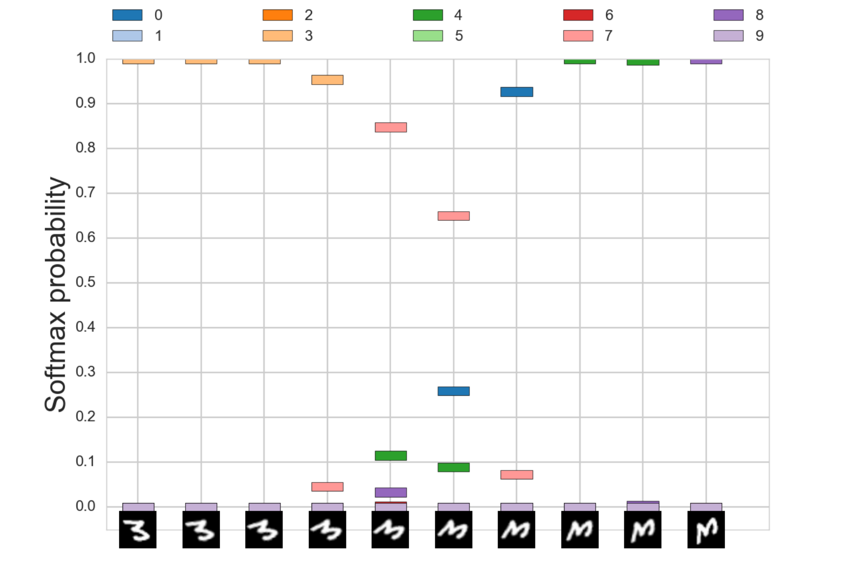


Рис.: Rotated MNIST

Метрика BPD

В процессе обучения мы перестаем смотреть на биты картинок как на дискретное распределение, из-за чего могут получиться неопревданно большие значения $\log q(x)$.
Чтобы это решить добавим шум $u \sim U[0, 1[^D, y = x + u$, $p(x)$ – реальное распределение, $q(x)$ – модельное распределение.

$$\int p(\mathbf{y}) \log q(\mathbf{y})\, d\mathbf{y} = \sum_{\mathbf{x}} P(\mathbf{x}) \int_{[0,1[^D} \log q(\mathbf{x} + \mathbf{u})\, d\mathbf{u}$$

$$\leq \sum_{\mathbf{x}} P(\mathbf{x}) \log \int_{[0,1[^D} q(\mathbf{x} + \mathbf{u})\, d\mathbf{u}$$

$$= \sum_{\mathbf{x}} P(\mathbf{x}) \log Q(\mathbf{x}),$$

Тогда $BPD = -\sum_{\mathbf{x}} P(\mathbf{x}) \log Q(\mathbf{x})$

Используемые модели

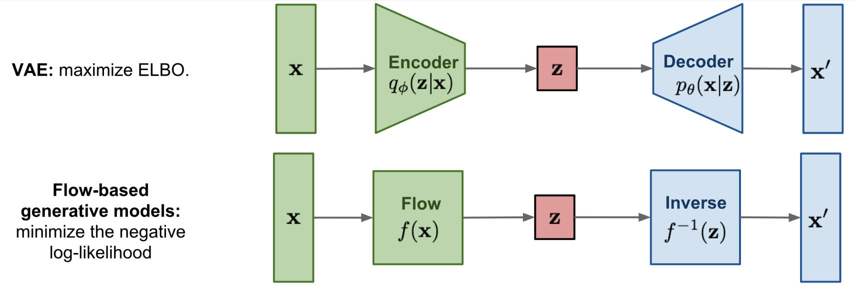


Рис.: Модели

Потоковые (Flow-based) модели в общем случае.

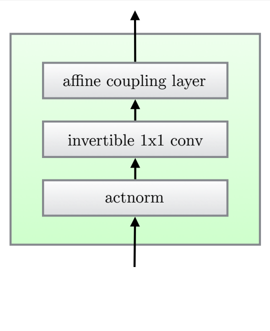
- $x \sim p^*(x)$, где $p^*(\cdot)$ – неизвестное реальное распределение.
- $p_\theta(x)$ – модель с параметрами θ.

Генеративный процесс описывают следующим образом:
$z \sim p_\theta(z), x = g_\theta(z)$, где

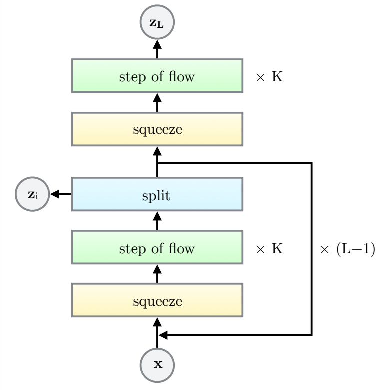
- z скрытая переменная
- $p_\theta(\cdot)$ – какое-нибудь не очень сложное распределение, например сферическая Гауссиана.
- $z = f_\theta(x) = g_\theta^{-1}(x)$ – f, g обратимые

Более того $f = f_1 \circ f_2 \circ \dots \circ f_K$ композиция (и аналогично g), такая что:
$\mathbf{x} \overset{\mathbf{f}_1}{\longleftrightarrow} \mathbf{h}_1 \overset{\mathbf{f}_2}{\longleftrightarrow} \mathbf{h}_2 \cdots \overset{\mathbf{f}_K}{\longleftrightarrow} \mathbf{z}$

Строение модели



(a) One step of our flow.



(b) Multi-scale architecture.

Description	Function	Reverse Function	Log-determinant
Actnorm.	$\forall i,j : \mathbf{y}_{i,j} = \mathbf{s} \odot \mathbf{x}_{i,j} + \mathbf{b}$	$\forall i,j : \mathbf{x}_{i,j} = (\mathbf{y}_{i,j} - \mathbf{b})/\mathbf{s}$	$h \cdot w \cdot \sum(\log\lvert\mathbf{s}\rvert)$
Invertible 1×1 convolution. $\mathbf{W} : [c \times c]$.	$\forall i,j : \mathbf{y}_{i,j} = \mathbf{W}\mathbf{x}_{i,j}$	$\forall i,j : \mathbf{x}_{i,j} = \mathbf{W}^{-1}\mathbf{y}_{i,j}$	$h \cdot w \cdot \log\lvert\det(\mathbf{W})\rvert$ or $h \cdot w \cdot \sum(\log\lvert\mathbf{s}\rvert)$
Affine coupling layer.	$\mathbf{x}_a, \mathbf{x}_b = = \mathrm{split}(\mathbf{x})$ $(\log\mathbf{s}, \mathbf{t}) = = \mathrm{NN}(\mathbf{x}_b)$ $\mathbf{s} = \exp(\log\mathbf{s})$ $\mathbf{y}_a = \mathbf{s} \odot \mathbf{x}_a + \mathbf{t}$ $\mathbf{y}_b = \mathbf{x}_b$ $\mathbf{y} = \mathrm{concat}(\mathbf{y}_a, \mathbf{y}_b)$	$\mathbf{y}_a, \mathbf{y}_b = \mathrm{split}(\mathbf{y})$ $(\log\mathbf{s}, \mathbf{t}) = \mathrm{NN}(\mathbf{y}_b)$ $\mathbf{s} = \exp(\log\mathbf{s})$ $\mathbf{x}_a = (\mathbf{y}_a - \mathbf{t})/\mathbf{s}$ $\mathbf{x}_b = \mathbf{y}_b$ $\mathbf{x} = \mathrm{concat}(\mathbf{x}_a, \mathbf{x}_b)$	$\sum(\log(\lvert\mathbf{s}\rvert))$

Чуть-чуть математики

Хотим уменьшать расстояние между распределениями:

$$\mathrm{KLD}[p^*(\boldsymbol{x})||p(\boldsymbol{x};\theta)] = \int p^*(\boldsymbol{x}) \log \frac{p^*(\boldsymbol{x})}{p(\boldsymbol{x};\theta)}\, d\boldsymbol{x} \approx -\frac{1}{N} \log p(\boldsymbol{X};\theta) - \mathbb{H}[p^*]$$

Поскольку энтропия реального распределения константа, оптимизируем только левую часть:

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \log p(\boldsymbol{X};\theta) = \arg\max_{\boldsymbol{\theta}} \sum_{n=1}^{N} \log p(\boldsymbol{x}_n;\theta).$$

Хотим извлекать вероятности из плотностей, для этого надо учиться интегрировать $P(\boldsymbol{\Omega}) = \int_{\boldsymbol{\Omega}} p(\boldsymbol{x};\theta) d\boldsymbol{x}$
Можно имитировать интегрирование возмущением:

$$\log \int p(\boldsymbol{x}_n + \delta;\theta) p(\delta)\, d\delta \geq \mathbb{E}_{\boldsymbol{\delta}} \left[\log p(\boldsymbol{x}_n + \delta;\theta)\right] \approx \log p(\boldsymbol{x}_n + \tilde{\delta};\theta)$$

где $\tilde{\boldsymbol{\delta}}$ сэмпл из $p(\boldsymbol{\delta})$.

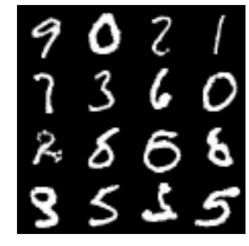
ВАЖНЫЙ ТРЮК про замену переменных:

$$p_z(f_\theta(X)) \mid \frac{df_\theta(X)}{dX} \mid = p(X)$$

$\mid \frac{df_\theta(X)}{dX} \mid$ – компенсация объема диффеоморфизма.
Тогда для Glow

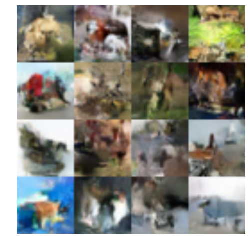
$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \log p_X(\boldsymbol{X}; \boldsymbol{\theta}) = \arg\max_{\phi,\psi} \sum_{n=1}^{N} \log p_z(f(\mathbf{x}_n; \phi); \psi) + \log \left| \frac{\partial \boldsymbol{f}_\phi}{\partial \mathbf{x}_n} \right| .$$

Картинки



(a) MNIST

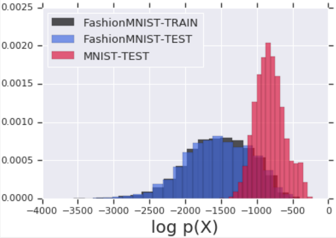
(b) FashionMNIST



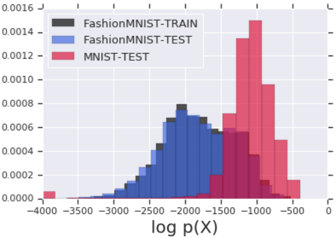
(c) CIFAR-10

(d) SVHN

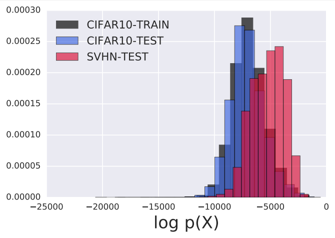
PixelCNN и VAE



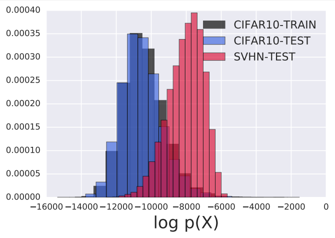
(a) PixelCNN: FashionMNIST vs MNIST



(b) VAE: FashionMNIST vs MNIST

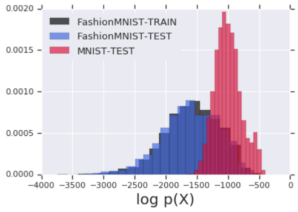


(c) PixelCNN: CIFAR-10 vs SVHN

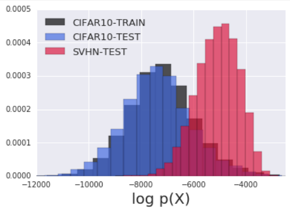


(d) VAE: CIFAR-10 vs SVHN

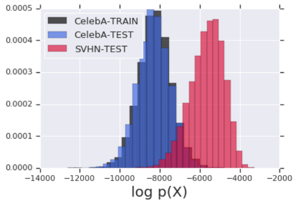
Glow



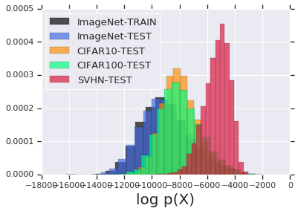
(a) Train on FashionMNIST, Test on MNIST



(b) Train on CIFAR-10, Test on SVHN

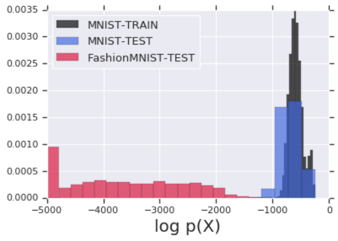


(c) Train on CelebA, Test on SVHN

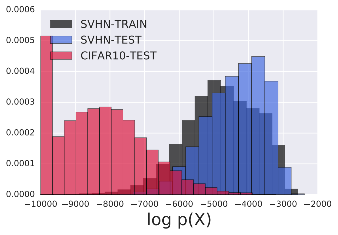


(d) Train on ImageNet, Test on CIFAR-10 / CIFAR-100 / SVHN

Несимметричность



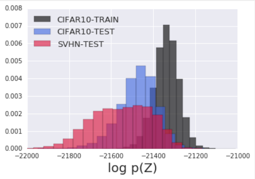
(a) Train on MNIST, Test on FashionMNIST



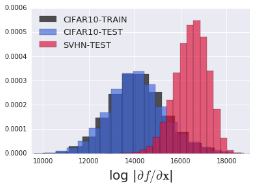
(b) Train on SVHN, Test on CIFAR-10

CV vs. NVP

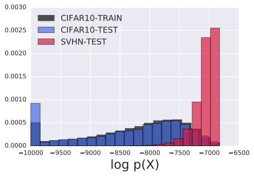
- NVP – Non Volume Preserving (классический вариант)
- CV – Constant Volume (отличается отсутствием скейлинга в ACL)



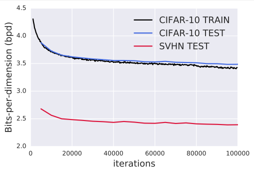
(a) CIFAR-10: $\log p(\boldsymbol{z})$



(b) CIFAR-10: Volume



(c) CV-Glow Likelihoods



(d) LL vs Iter.

Анализ CV-GLOW

Хотим анализировать, но сложно:

$$\mathbb{E}_q[\log p(\boldsymbol{x};\theta)] - \mathbb{E}_{p^*}[\log p(\boldsymbol{x};\theta)] > 0$$

Берем Тейлора

$$\log p(\boldsymbol{x};\theta) \approx \log p(\boldsymbol{x}_0;\theta) +$$
$$+\nabla_{\boldsymbol{x}_0} \log p(\boldsymbol{x}_0;\theta)^T(\boldsymbol{x}-\boldsymbol{x}_0) + \frac{1}{2}\operatorname{Tr}\{\nabla^2_{\boldsymbol{x}_0} \log p(\boldsymbol{x}_0;\theta)(\boldsymbol{x}-\boldsymbol{x}_0)(\boldsymbol{x}-\boldsymbol{x}_0)^T\}$$

Вот сомнительная часть:
Утверждается что раз Тейлор работает только около $\boldsymbol{x}_0$, тогда
$\mathbb{E}_q[\boldsymbol{x}] = \mathbb{E}_{p^*}[\boldsymbol{x}] = \boldsymbol{x}_0$
Тогда получаем

$$0 < \mathbb{E}_q[\log p(\boldsymbol{x};\theta)] - \mathbb{E}_{p^*}[\log p(\boldsymbol{x};\theta)] \approx \frac{1}{2}\operatorname{Tr}\{\nabla^2_{\boldsymbol{x}_0} \log p(\boldsymbol{x}_0;\theta)(\boldsymbol{\Sigma}_q - \boldsymbol{\Sigma}_{p^*})\}$$
$$= \frac{1}{2}\operatorname{Tr}\left\{\left[\nabla^2_{\boldsymbol{x}_0} \log p_z(f(\boldsymbol{x}_0;\phi)) + \nabla^2_{\boldsymbol{x}_0} \log\left|\frac{\partial \boldsymbol{f}_\phi}{\partial \boldsymbol{x}_0}\right|\right]\left(\boldsymbol{\Sigma}_q - \boldsymbol{\Sigma}_{p^*}\right)\right\},$$

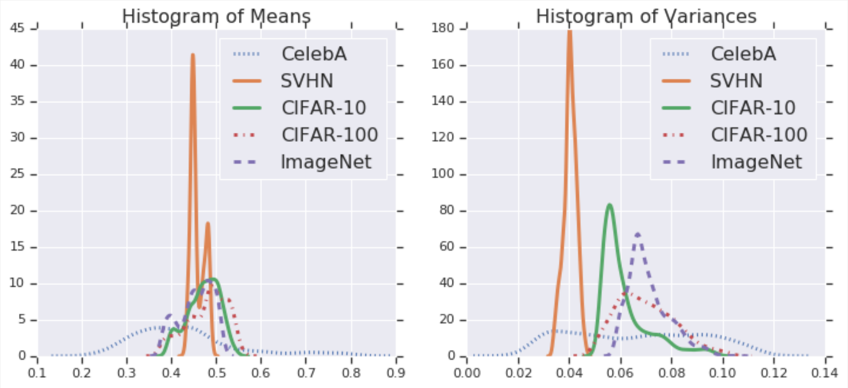


Рис.: Histogram of per-dimension means and variances (empirical).

- Пусть U_k ядро $C \times C$, где C – количество каналов, а k – номер потока.
- $\partial f_{h,w,c} / \partial x_{h,w,c} = \prod_k \sum_{j=1}^{C} u_{k,c,j}$
- $\partial^2 f_{h,w,c} / \partial x_{h,w,c}^2 = 0$

$$
\begin{aligned}
&\operatorname{Tr}\left\{\left[\nabla^2_{\mathbf{x}_0} \log p(\mathbf{x}_0; \theta)\right]\left(\mathbf{\Sigma}_q - \mathbf{\Sigma}_{p^*}\right)\right\} \\
&\quad = \frac{\partial^2}{\partial z^2} \log p(\mathbf{z}; \psi) \sum_{c=1}^{C} \left(\prod_{k=1}^{K} \sum_{j=1}^{C} u_{k,c,j} \right)^2 \sum_{h,w} (\sigma^2_{q,h,w,c} - \sigma^2_{p^*,h,w,c})
\end{aligned}
$$

$$
\begin{aligned}
&\mathbb{E}_{\text{SVHN}}[\log p(\boldsymbol{x};\theta)] - \mathbb{E}_{\text{CIFAR-10}}[\log p(\boldsymbol{x};\theta)] \\
&\qquad \approx \frac{-1}{2\sigma_{\psi}^2}\left[\alpha_1^2(49.6-61.9)+\alpha_2^2(52.7-59.2)+\alpha_3^2(53.6-68.1)\right] \\
&\qquad = \frac{1}{2\sigma_{\psi}^2}\left[\alpha_1^2\cdot 12.3+\alpha_2^2\cdot 6.5+\alpha_3^2\cdot 14.5\right] \geq 0 \\
&\qquad \text{где} \quad \alpha_c = \prod_{k=1}^{K}\sum_{j=1}^{C} u_{k,c,j}
\end{aligned}
$$

Выводы

- Существует проблема с OOD
- ВОЗМОЖНО для FLow объясняется разницей в дисперсии распределений
- Пока эта проблема существует, нужно думать прежде чем искать аномалии такими моделями.

Вопросы

- Из каких элементов состоит один слой Flow. Выписать, что происходит в каждом из них.
- Формулировка проблемы поставленной статьей. Перечислить возможные причины ее возникновения.
- Формула замены переменных для Glow. На аккие части благодаря этому разбивается поиск оптимального параметра.