Do Deep Generative Models Know What They Don't Know?

Balaji Lakshminarayanan, Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur

DeepMind

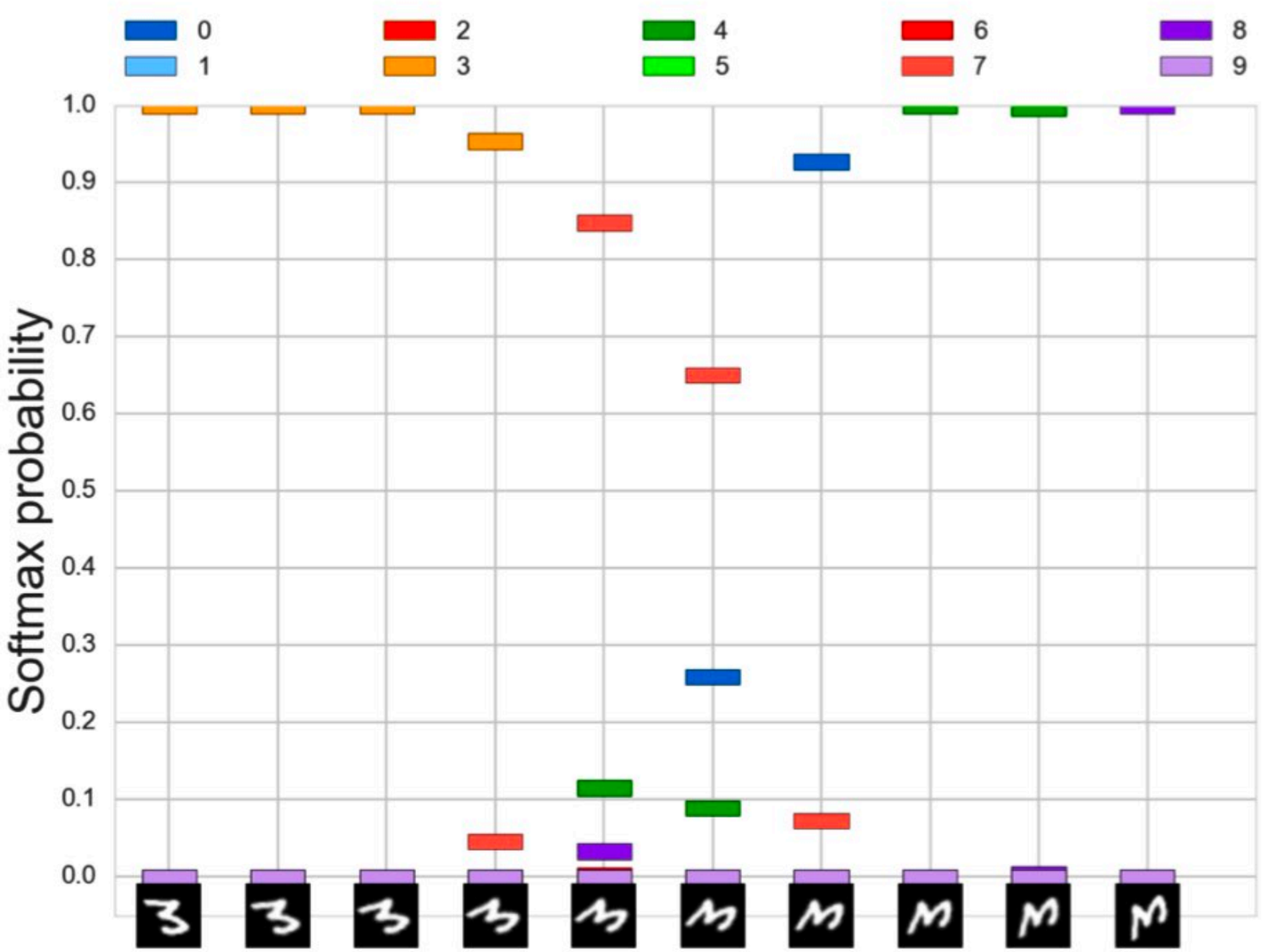
(не) Уверенность модели. Зачем ее оценивать?

- Модели, обученные предсказывать $p(y \mid x)$обычно хорошо работают для $x \sim p_{train}(X)$

- Для x из иного распределения результаты могут быть непредсказуемы

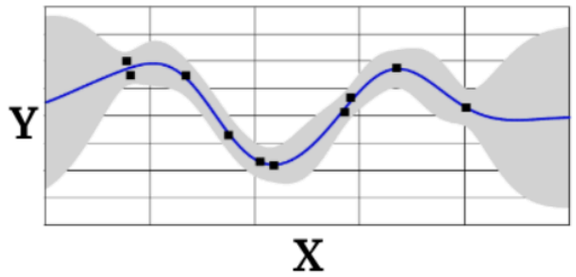
- Зная, что модель не уверенна в предсказаниях мы могли бы не доверять им

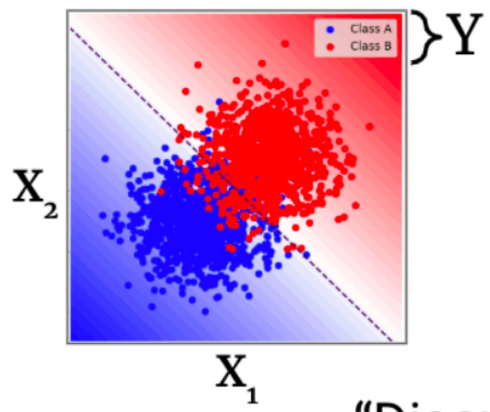
(не) Уверенность модели. Зачем ее оценивать?



Поворот MNIST приводит к тому, что нейронная сеть (LeNet + weight decay) предсказывает неверному классу высокую вероятность

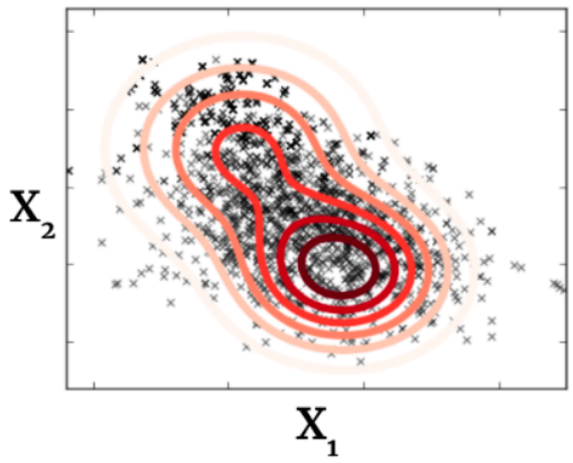
Дискриминативные и Генеративные и модели





$$p(\mathbf{y}|\mathbf{x})$$

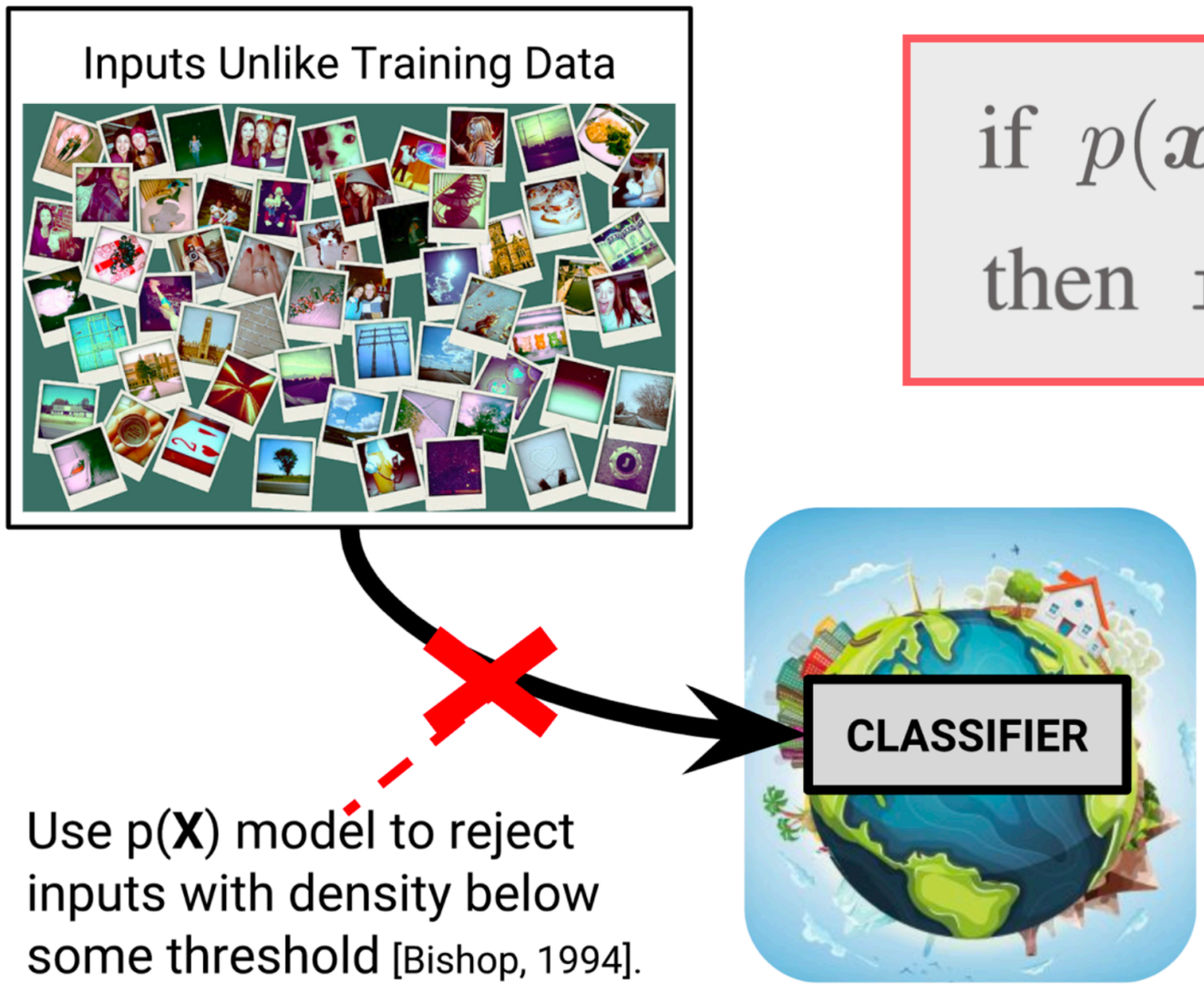
“Discriminative” Model



$$p(\mathbf{x})$$

“Generative” Model

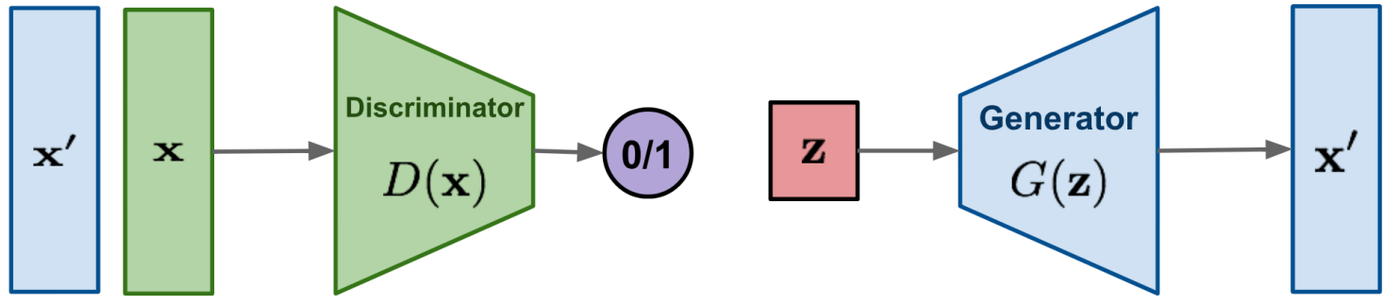
- Знаем: $p(y \mid x)$ ошибается на OOD (out-of-distribution)
- Идея: будем использовать генеративную модель, чтобы отличать x из OOD



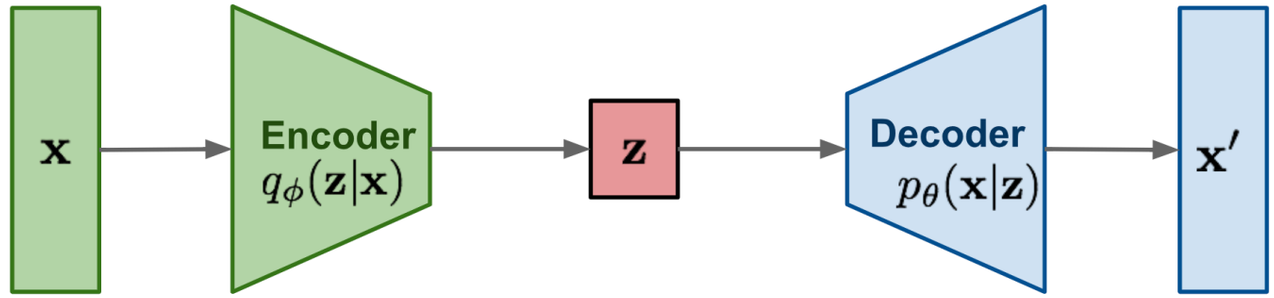
$$\text{if } p(\boldsymbol{x}^*; \boldsymbol{\phi}) < \tau,$$
$$\text{then } \texttt{reject } \boldsymbol{x}^*$$

Типы генеративных моделей

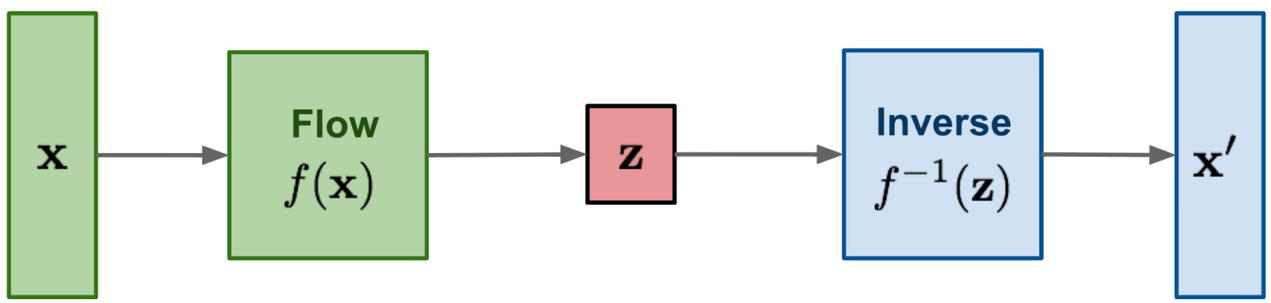
GAN: minimax the classification error loss.



VAE: maximize ELBO.



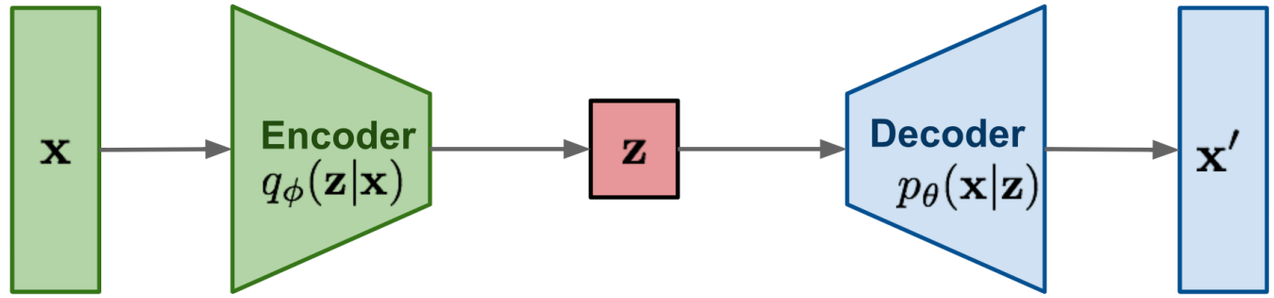
Flow-based generative models: minimize the negative log-likelihood



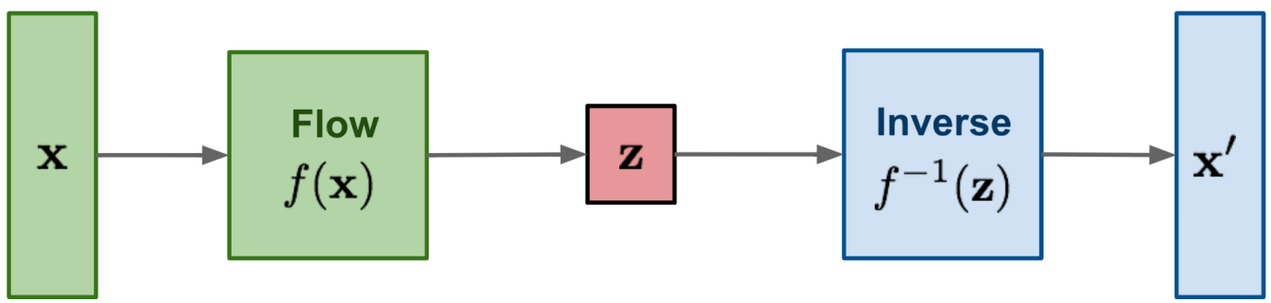
Типы генеративных моделей

GAN: minimax the classification error loss.

VAE: maximize ELBO.

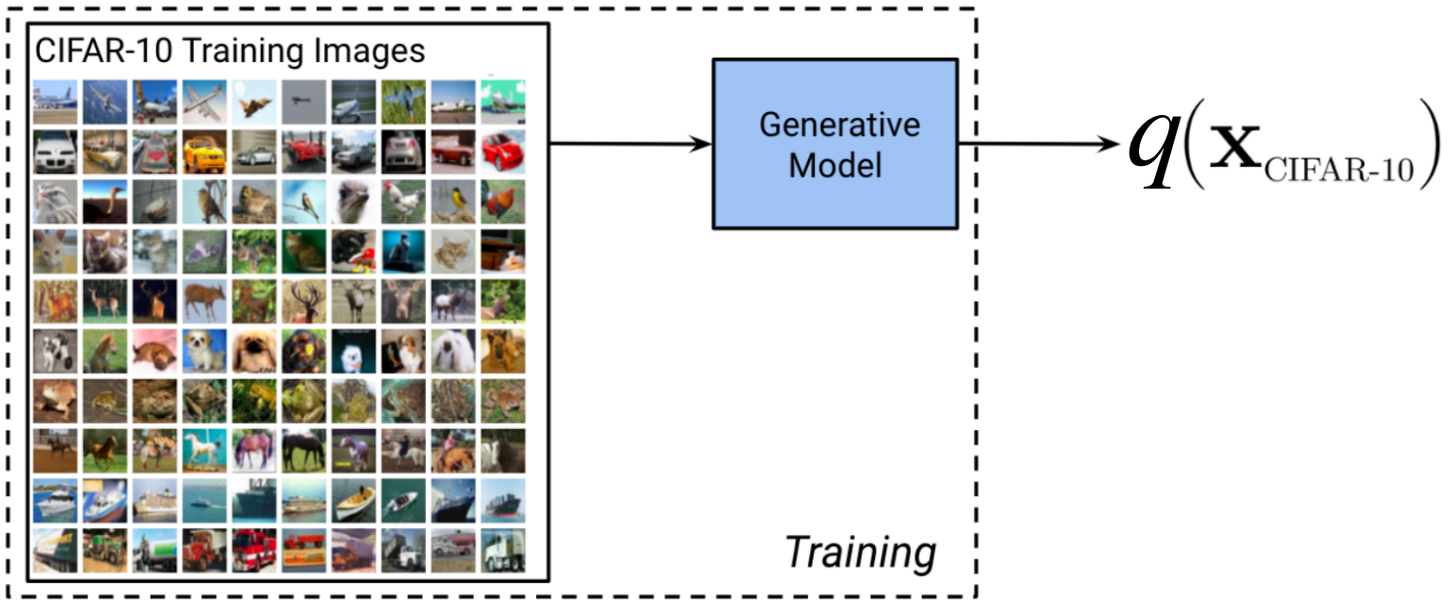


Flow-based generative models: minimize the negative log-likelihood



- Можем вычислить $p(x)$
  - Точно для Flow-based
  - Нижнюю оценку для VAE

Оцениваем лог-правдоподобие данных: bits-per-dimension



- Обычно изображения из дискретного распределения, а модель приближает непрерывное

- Из-за этого $-\log\big(q(x)\big)$ показывает неоправданно высокие значения

- Добавим к входным данным случайны шум

Оцениваем лог-правдоподобие данных: bits-per-dimension

- Пусть $p(x)$ — истинное распределение, $q(x)$ — приближённое
- Добавим к входным данным случайны шум: $y = x + u$

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

Усредненное лог правдоподобие

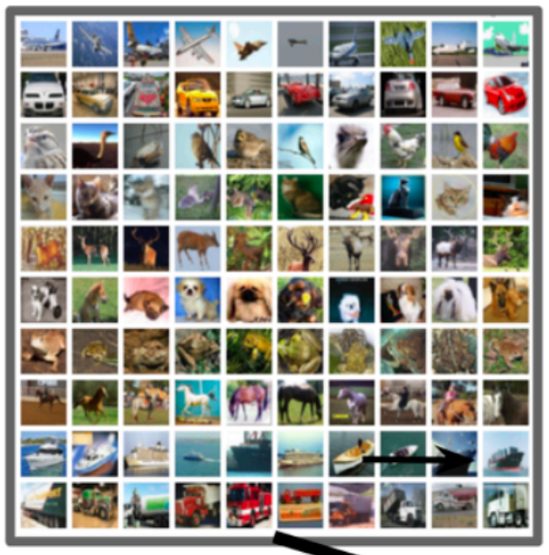
$$\leq \sum_{\mathbf{x}} P(\mathbf{x}) \log \int_{[0,1[^D} q(\mathbf{x}+\mathbf{u})\, d\mathbf{u}$$

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

Отрицание этого выражение — BPD

Обучаем на CIFAR тестируем на SVHN

Training: *CIFAR-10*



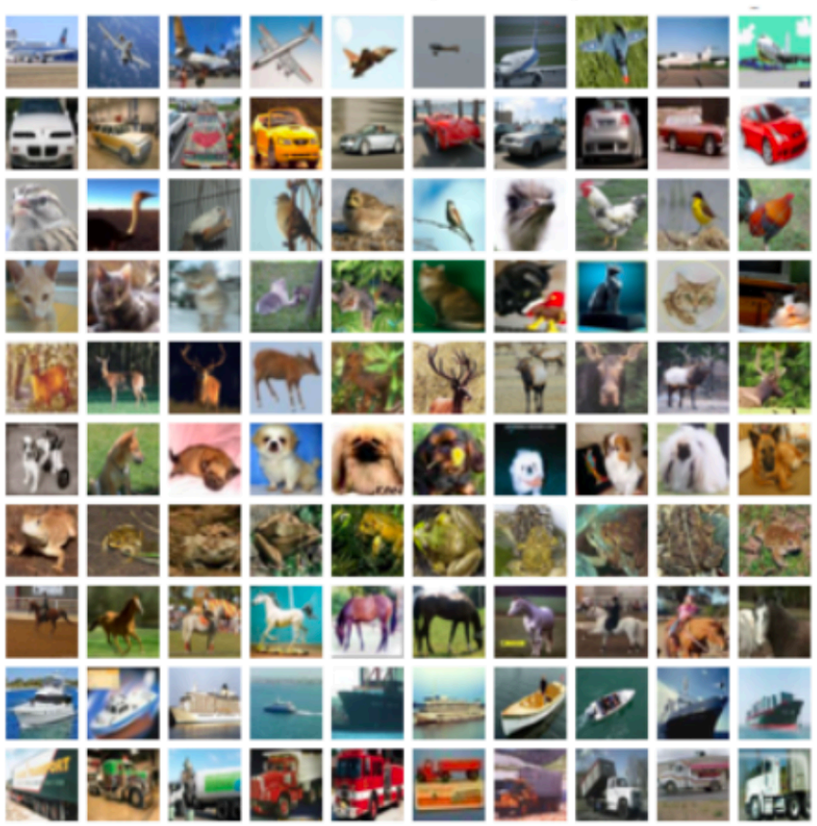
Testing: *SVHN*

GENERATIVE MODEL

$$p(\mathbf{x}_{\text{CIFAR-10}}) \overset{?}{>} p(\mathbf{x}_{\text{SVHN}})$$

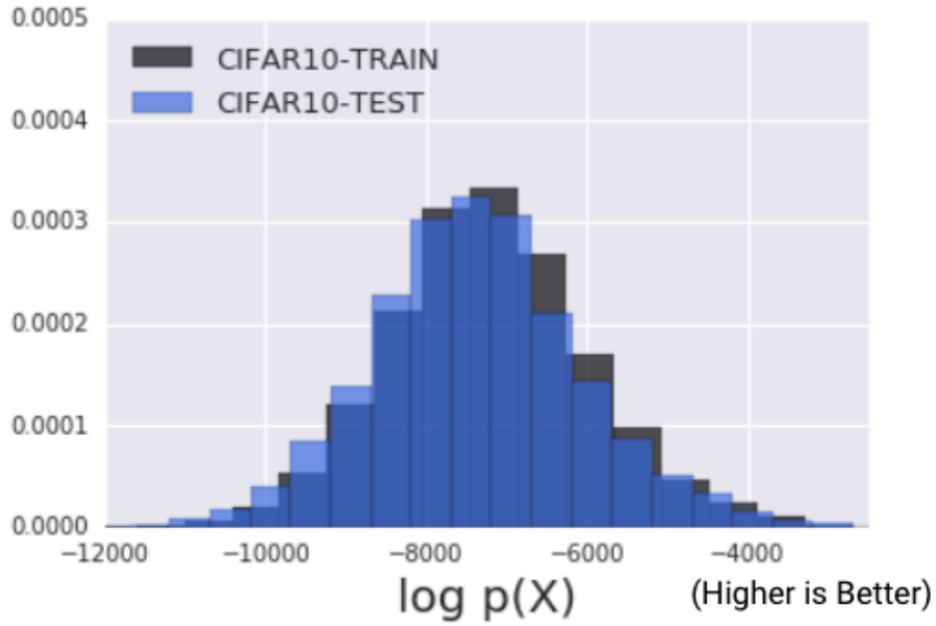
Обучаем на CIFAR тестируем на SVHN

CIFAR-10 Training Images



	Bits Per Dimension (NLL / # dims / log 2)
CIFAR10-Train	3.386
CIFAR10-Test	3.464

(Lower is Better)



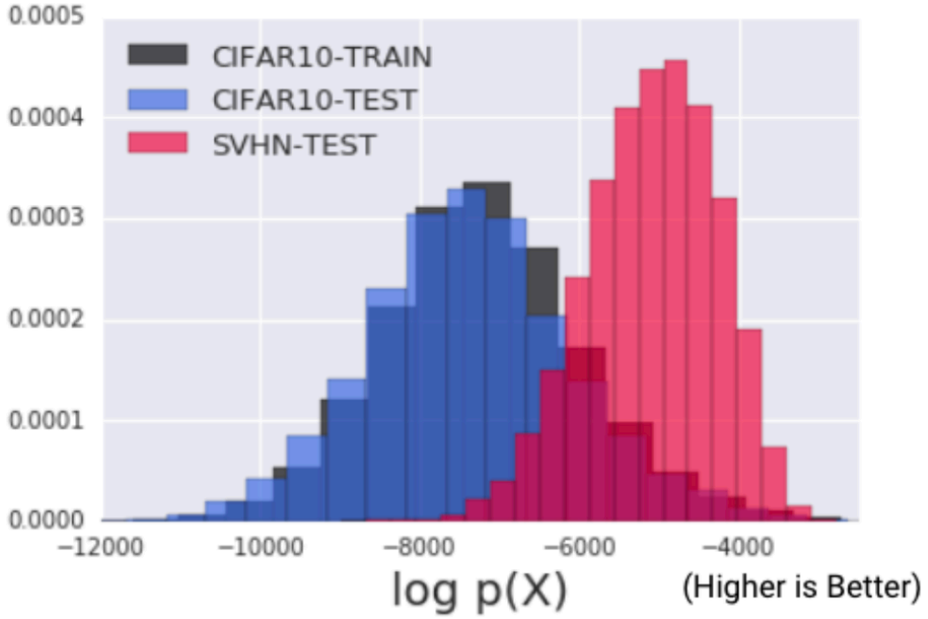
(Higher is Better)

Обучаем на CIFAR тестируем на SVHN

SVHN Test Images

	Bits Per Dimension (NLL / # dims / log 2)
CIFAR10-Train	3.386
CIFAR10-Test	3.464
SVHN-Test	**2.389**

(Lower is Better)

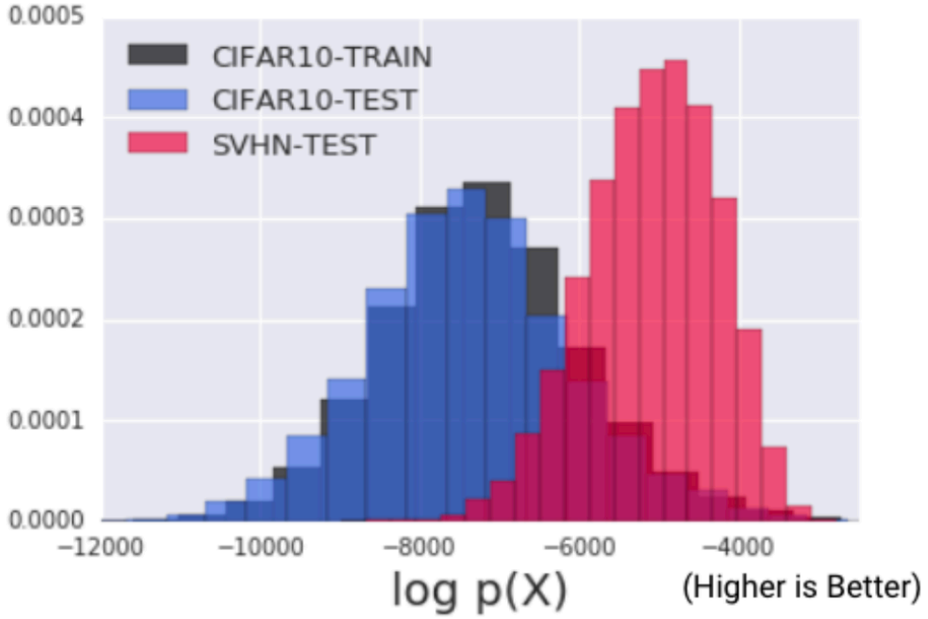


Обучаем на CIFAR тестируем на SVHN

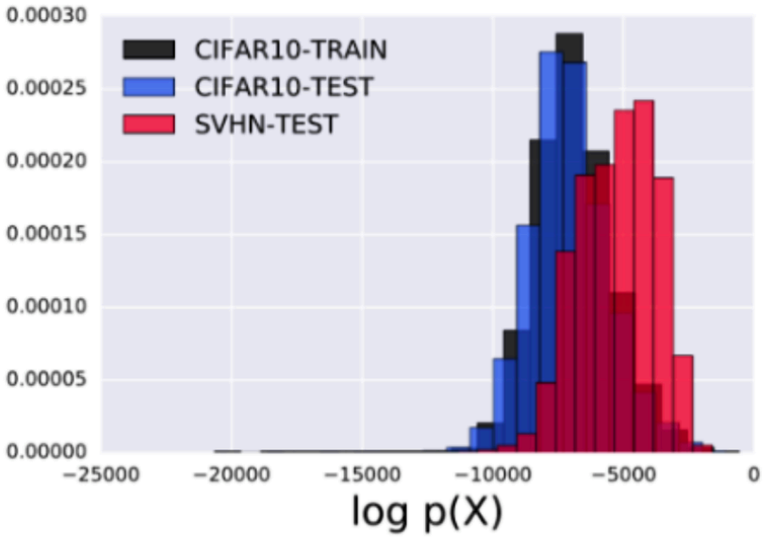
SVHN Test Images

	Bits Per Dimension (NLL / # dims / log 2)
CIFAR10-Train	3.386
CIFAR10-Test	3.464
SVHN-Test	**2.389**

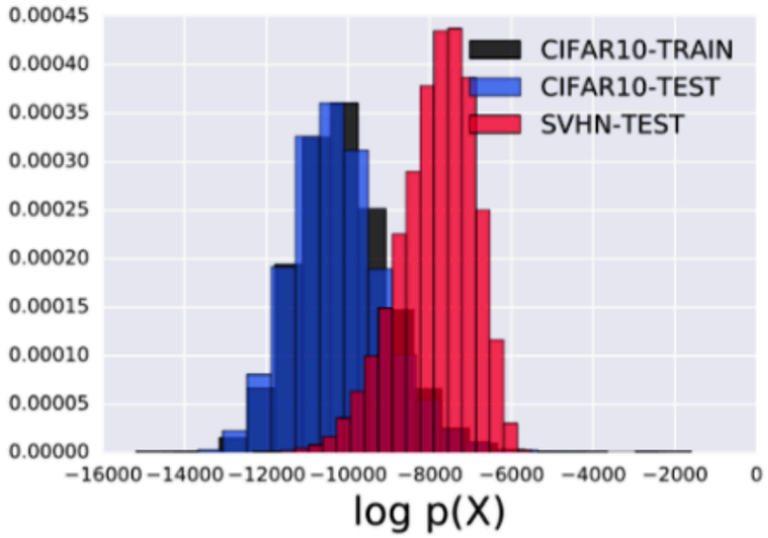
(Lower is Better)



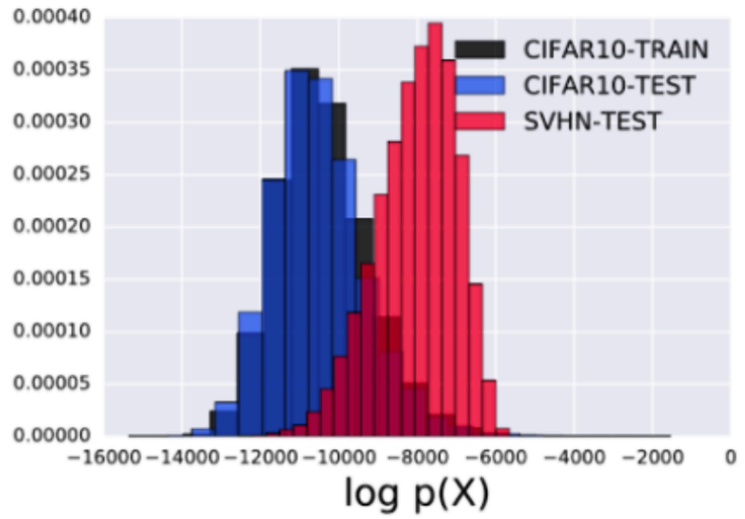
Воспроизводимость на других моделях



(a) PixelCNN

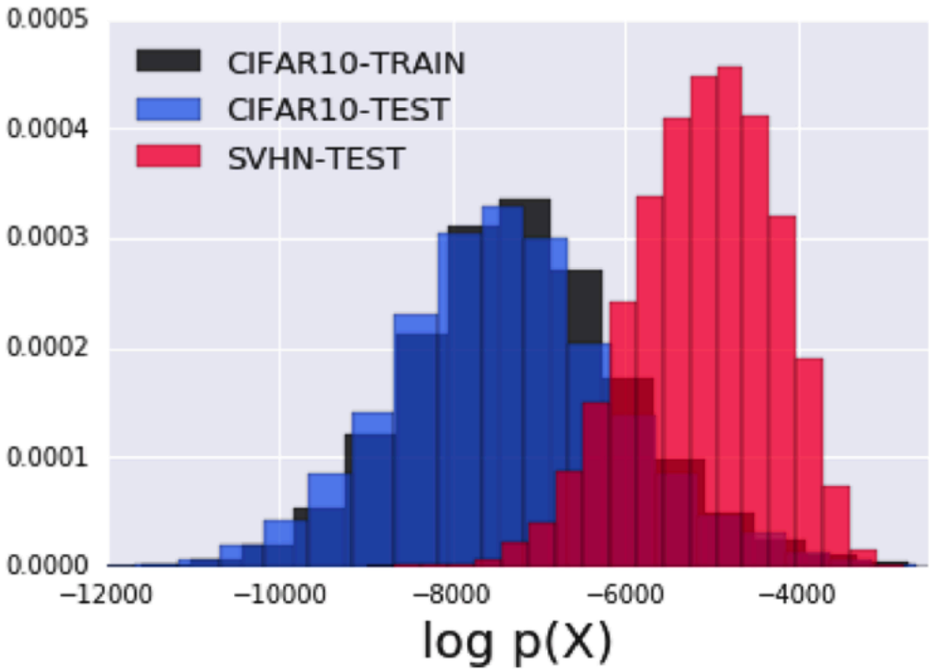


(b) VAE with RNVP as encoder

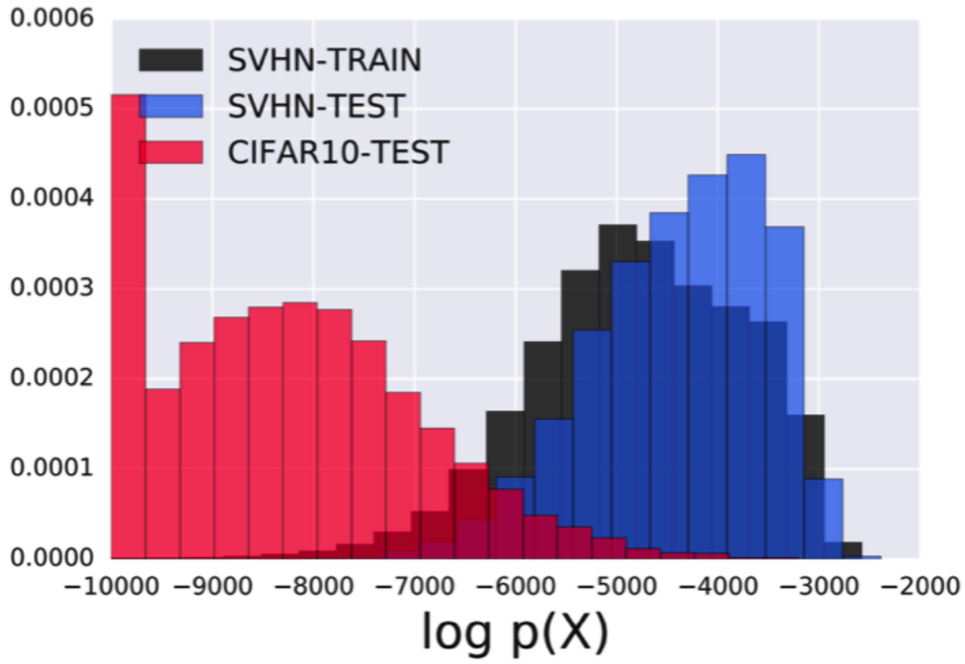


(c) VAE conv-categorical likelihood

Асимметричность по данным в обучении

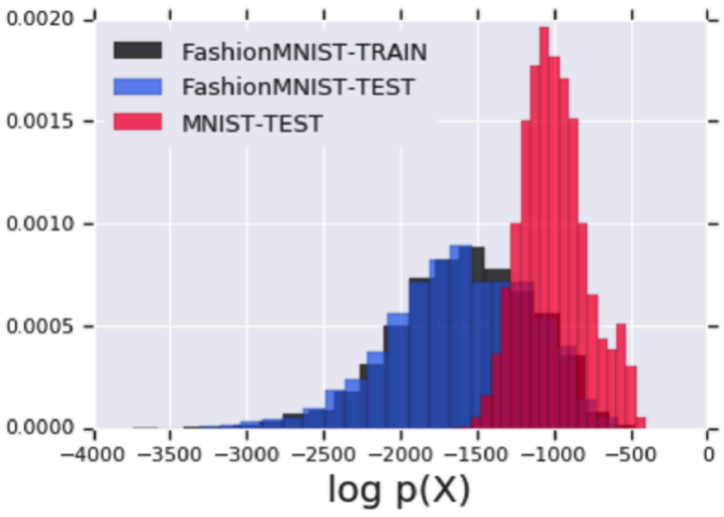


CIFAR-10 vs SVHN

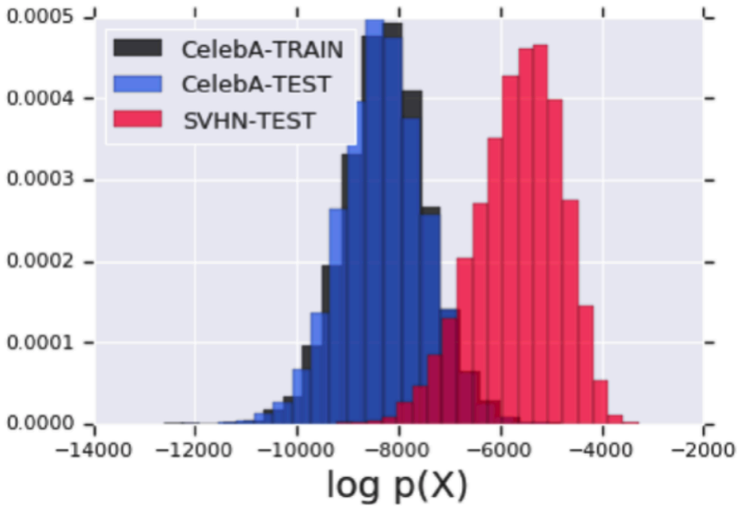


SVHN vs CIFAR-10

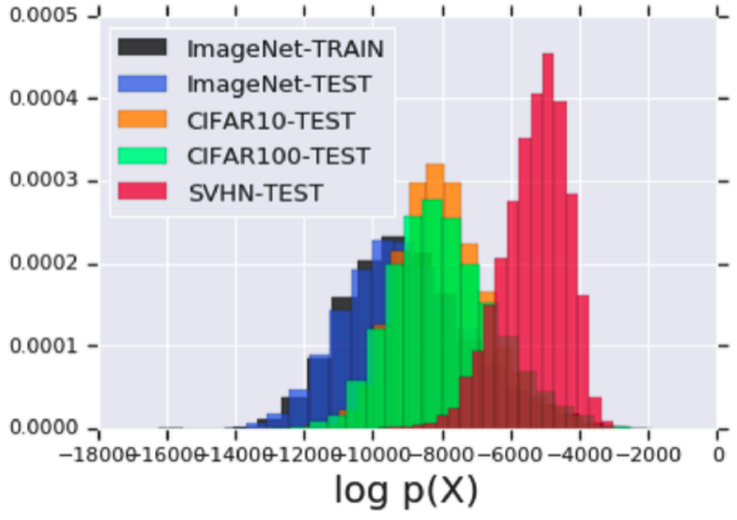
Другие пары наборов данных



FashionMNIST vs MNIST



CelebA vs SVHN

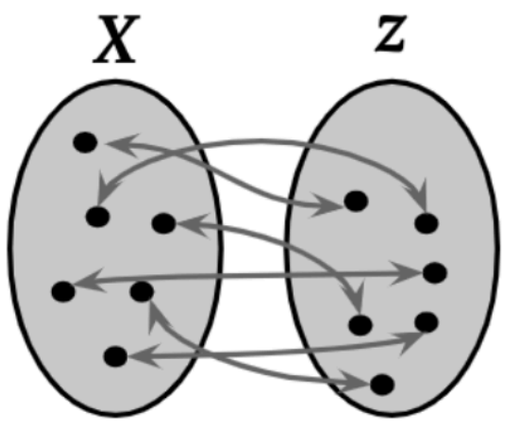


ImageNet vs CIFAR-10 vs SVHN

Коротко про потоки

Define *Z* by a transformation of another variable *X*: $Z = f(X)$

f(**x**) is a *bijection*
(invertible 1:1 mapping)



$\mathbf{x} = f^{-1}(\mathbf{z})$ $\mathbf{z} = f(\mathbf{x})$

Change of Variables Formula (X → Z):

$$p_z(f(X)) \left| \frac{df(X)}{dX} \right| = p(X)$$

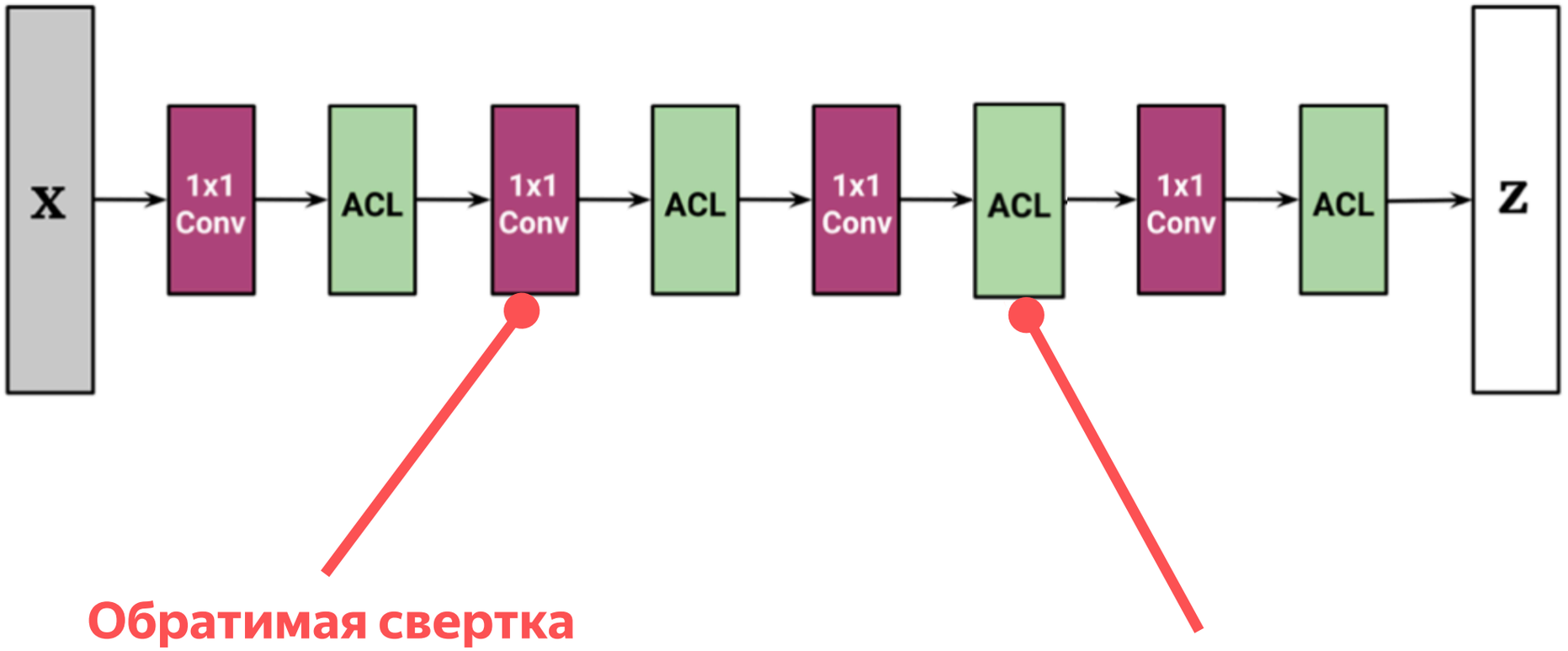
Use simple p_z distribution (e.g. standard normal)

(Distribution term)

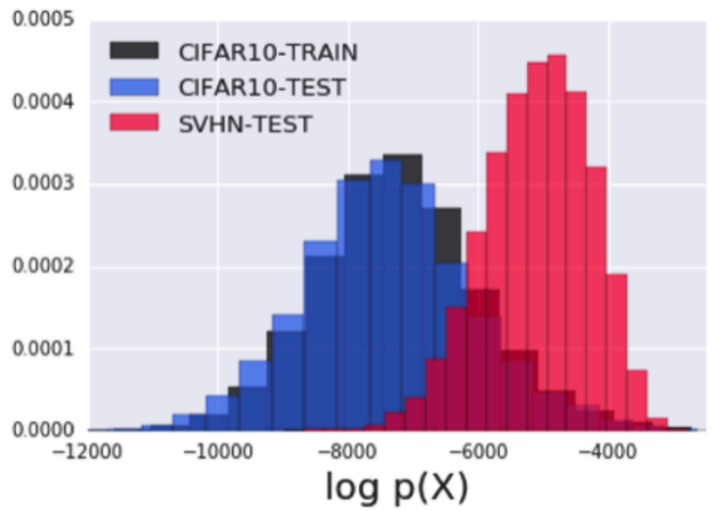
Use f such that the Jacobian df/dx is easy to compute

(Volume term)

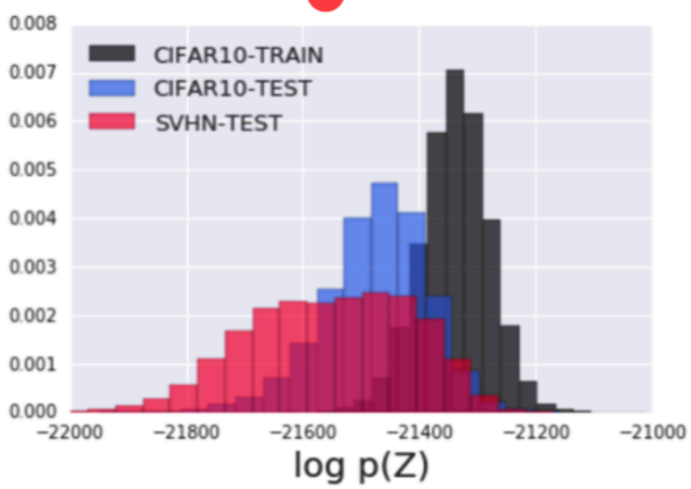
Коротко про потоки



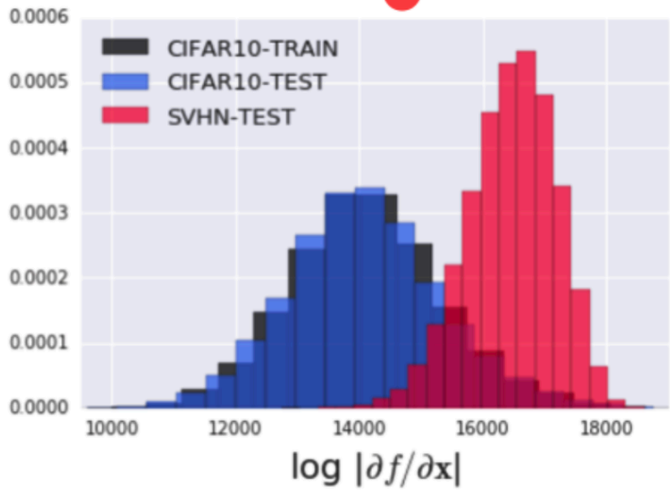
Разложение правдоподобия для flow-based моделей



CIFAR-10 vs SVHN



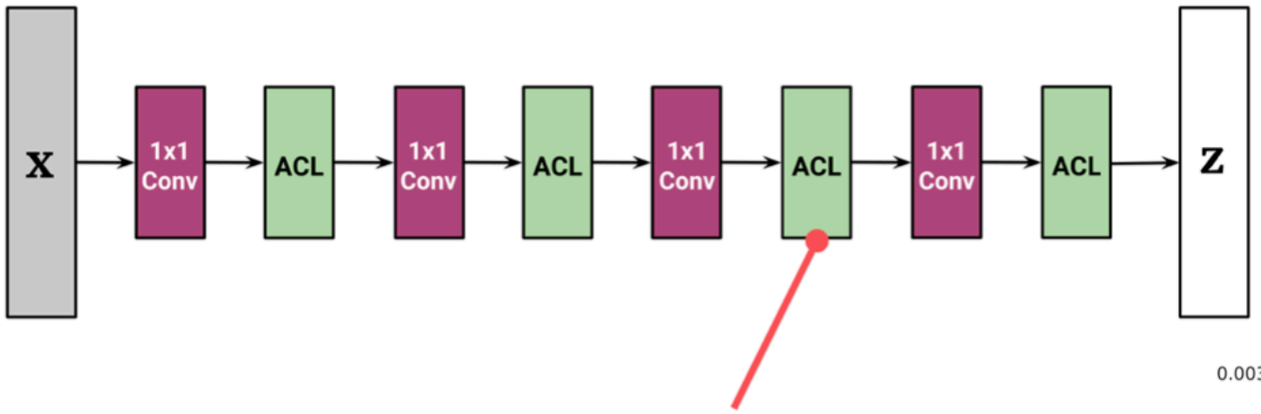
Distribution Term



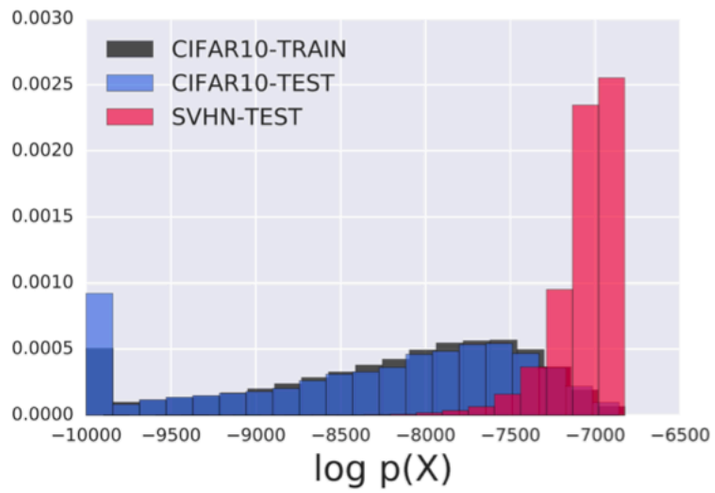
Volume Term

Constant-Volume GLOW

We define a sub-class we term *constant-volume* (w.r.t. input) flows.



CIFAR-10 vs SVHN



Анализ CV-GLOW

Mathematical characterization:

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

Non-Training Distribution

Training Distribution

Анализ CV-GLOW

Mathematical characterization:

$$0 < \underbrace{\mathbb{E}_q[\log p(\boldsymbol{x};\boldsymbol{\theta})]}_{\text{Non-Training Distribution}} - \underbrace{\mathbb{E}_{p^*}[\log p(\boldsymbol{x};\boldsymbol{\theta})]}_{\text{Training Distribution}}$$

$$\log p(x;\theta) \approx \log p(x_0;\theta) + \nabla x_0 \log_p(x_0;\theta)^T(x - x_0) + \frac{1}{2} Tr\left\{ \nabla^2 x_0 \log p(x_0;\theta)(x - x_0)(x - x_0)^T \right\}$$

Анализ CV-GLOW

Mathematical characterization:

$$0 < \underbrace{\mathbb{E}_q[\log p(\boldsymbol{x};\boldsymbol{\theta})]}_{\text{Non-Training Distribution}} - \underbrace{\mathbb{E}_{p^*}[\log p(\boldsymbol{x};\boldsymbol{\theta})]}_{\text{Training Distribution}}$$

$$\approx \nabla x_0 \log p(x_0;\theta)^T(\mathbb{E}q[x] - \mathbb{E}p_*[x]) \quad \text{Zero for equal means}$$

$$+\frac{1}{2}Tr\left\{\nabla^2 x_0 \log p(x_0;\theta)(\sum q - \sum p_*)\right\}$$

Анализ CV-GLOW

Mathematical characterization:

$$0 < \underbrace{\mathbb{E}_q}_{\text{Non-Training Distribution}}[\log p(\boldsymbol{x};\boldsymbol{\theta})] - \underbrace{\mathbb{E}_{p^*}}_{\text{Training Distribution}}[\log p(\boldsymbol{x};\boldsymbol{\theta})]$$

$$\approx \frac{1}{2}\operatorname{Tr}\left\{\left[\underbrace{\nabla^2_{\boldsymbol{x}_0}\log p_z(f(\boldsymbol{x}_0;\boldsymbol{\phi})) + \nabla^2_{\boldsymbol{x}_0}\log\left|\frac{\partial \boldsymbol{f}_{\boldsymbol{\phi}}}{\partial \boldsymbol{x}_0}\right|}_{\text{Change-of-Variable Terms}}\right](\underbrace{\boldsymbol{\Sigma}_q}_{\text{Second Moment of Non-Training Distribution}} - \overline{\boldsymbol{\Sigma}_{p^*}})\right\}$$

Second Moment of Training Distribution

Анализ CV-GLOW

Mathematical characterization:

$$0 < \underbrace{\mathbb{E}_q[\log p(\boldsymbol{x};\boldsymbol{\theta})]}_{\text{Non-Training Distribution}} - \underbrace{\mathbb{E}_{p^*}[\log p(\boldsymbol{x};\boldsymbol{\theta})]}_{\text{Training Distribution}}$$

$$\approx \frac{1}{2}\operatorname{Tr}\left\{\left[\nabla^2_{\boldsymbol{x}_0}\log p_z(f(\boldsymbol{x}_0;\boldsymbol{\phi})) + \nabla^2_{\boldsymbol{x}_0}\log\left|\frac{\partial f_{\boldsymbol{\phi}}}{\partial \boldsymbol{x}_0}\right|\right](\underbrace{\boldsymbol{\Sigma}_q}_{\text{Second Moment of Non-Training Distribution}} - \overline{\boldsymbol{\Sigma}_{p^*}})\right\}$$

Change-of-Variable Terms

Second Moment of Training Distribution

Анализ CV-GLOW

Plugging in the CV-Glow transform:

$$\mathrm{Tr}\left\{\left[\nabla^2_{\boldsymbol{x}_0}\log p(\boldsymbol{x}_0;\boldsymbol{\theta})\right](\boldsymbol{\Sigma}_q-\boldsymbol{\Sigma}_{p^*})\right\}$$

$$=\frac{\partial^2}{\partial z^2}\log p(\boldsymbol{z};\boldsymbol{\psi})\sum_{c=1}^{C}\left(\prod_{k=1}^{K}\sum_{j=1}^{C}u_{k,c,j}\right)^2\sum_{h,w}(\overline{\sigma^2_{q,h,w,c}}-\overline{\sigma^2_{p^*,h,w,c}})$$

Second Moment of Non-Training Distribution

Second Moment of Training Distribution

1x1 Conv. Params

Анализ CV-GLOW

Plugging in the CV-Glow transform:

$$\mathrm{Tr}\left\{\left[\nabla^2_{\boldsymbol{x}_0}\log p(\boldsymbol{x}_0;\boldsymbol{\theta})\right](\boldsymbol{\Sigma}_q-\boldsymbol{\Sigma}_{p^*})\right\}$$

$$=\frac{\partial^2}{\partial z^2}\log p(\boldsymbol{z};\boldsymbol{\psi})\sum_{c=1}^{C}\left(\prod_{k=1}^{K}\sum_{j=1}^{C}u_{k,c,j}\right)^2\sum_{h,w}(\overline{\sigma^2_{q,h,w,c}}-\overline{\sigma^2_{p^*,h,w,c}})$$

Second Moment of Non-Training Distribution

Second Moment of Training Distribution

1x1 Conv. Params

Анализ CV-GLOW

Plugging in the CV-Glow transform:

$$\mathrm{Tr}\left\{\left[\nabla^2_{\boldsymbol{x}_0}\log p(\boldsymbol{x}_0;\boldsymbol{\theta})\right](\boldsymbol{\Sigma}_q-\boldsymbol{\Sigma}_{p^*})\right\}$$

$$=\frac{\partial^2}{\partial z^2}\log p(\boldsymbol{z};\boldsymbol{\psi})\sum_{c=1}^{C}\left(\prod_{k=1}^{K}\sum_{j=1}^{C}u_{k,c,j}\right)^2\sum_{h,w}(\overline{\sigma^2_{q,h,w,c}}-\overline{\sigma^2_{p^*,h,w,c}})$$

Second Moment of Non-Training Distribution

Second Moment of Training Distribution

1x1 Conv. Params

Sums over channel dimensions

Product over steps in flow

Sum over spatial dimensions

Анализ CV-GLOW

Plugging in the CV-Glow transform:

$$\mathrm{Tr}\left\{\left[\nabla^2_{\boldsymbol{x}_0}\log p(\boldsymbol{x}_0;\boldsymbol{\theta})\right](\boldsymbol{\Sigma}_q-\boldsymbol{\Sigma}_{p^*})\right\}$$

$$=\underbrace{\frac{\partial^2}{\partial z^2}\log p(\boldsymbol{z};\boldsymbol{\psi})}_{\text{< 0 for all log-concave densities (e.g. Gaussian)}}\sum_{c=1}^{C}\underbrace{\left(\prod_{k=1}^{K}\sum_{j=1}^{C}u_{k,c,j}\right)^2}_{\text{Non-negative due to square}}\sum_{h,w}(\overline{\sigma^2_{q,h,w,c}}-\overline{\sigma^2_{p^*,h,w,c}})$$

Second Moment of Non-Training Distribution

Second Moment of Training Distribution

Анализ CV-GLOW

Plugging in the CV-Glow transform:

$$\mathrm{Tr}\left\{\left[\nabla^2_{\boldsymbol{x}_0}\log p(\boldsymbol{x}_0;\boldsymbol{\theta})\right](\boldsymbol{\Sigma}_q-\boldsymbol{\Sigma}_{p^*})\right\}$$

$$=\underbrace{\frac{\partial^2}{\partial z^2}\log p(\boldsymbol{z};\boldsymbol{\psi})}_{\text{< 0 for all log-concave densities (e.g. Gaussian)}}\sum_{c=1}^{C}\underbrace{\left(\prod_{k=1}^{K}\sum_{j=1}^{C}u_{k,c,j}\right)^2}_{\text{Non-negative due to square}}\sum_{h,w}(\overline{\sigma^2_{q,h,w,c}}-\overline{\sigma^2_{p^*,h,w,c}})$$

Second Moment of Non-Training Distribution

Second Moment of Training Distribution

Sign boils down to difference in moments. Speaks to asymmetric behavior.

Анализ CV-GLOW

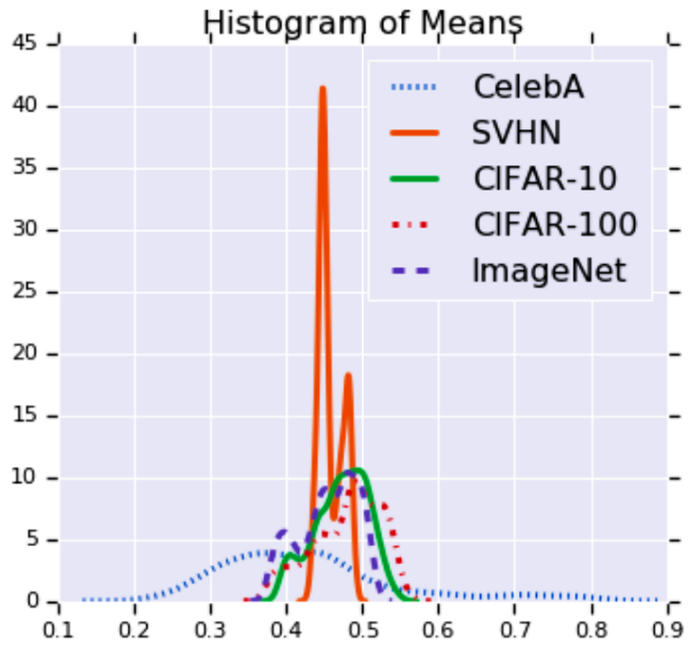
Plugging in the CIFAR-10 and SVHN statistics:

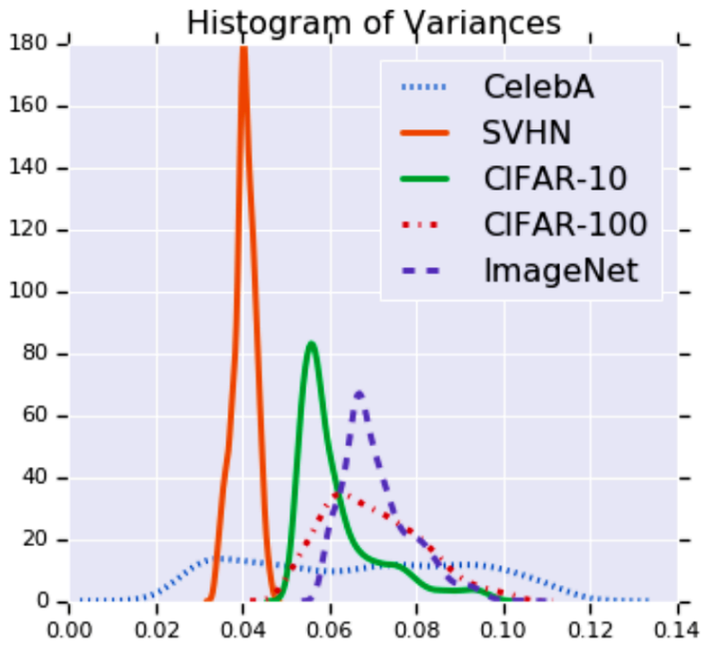
$$\mathbb{E}_{\text{SVHN}}[\log p(\boldsymbol{x};\boldsymbol{\theta})] - \mathbb{E}_{\text{CIFAR10}}[\log p(\boldsymbol{x};\boldsymbol{\theta})]$$

$$\approx \frac{1}{2\sigma_{\psi}^2}\left[\alpha_1^2 \cdot 12.3 + \alpha_2^2 \cdot 6.5 + \alpha_3^2 \cdot 14.5\right] \geq 0 \quad \text{where} \;\; \alpha_c = \prod_{k=1}^{K}\sum_{j=1}^{C} u_{k,c,j}$$

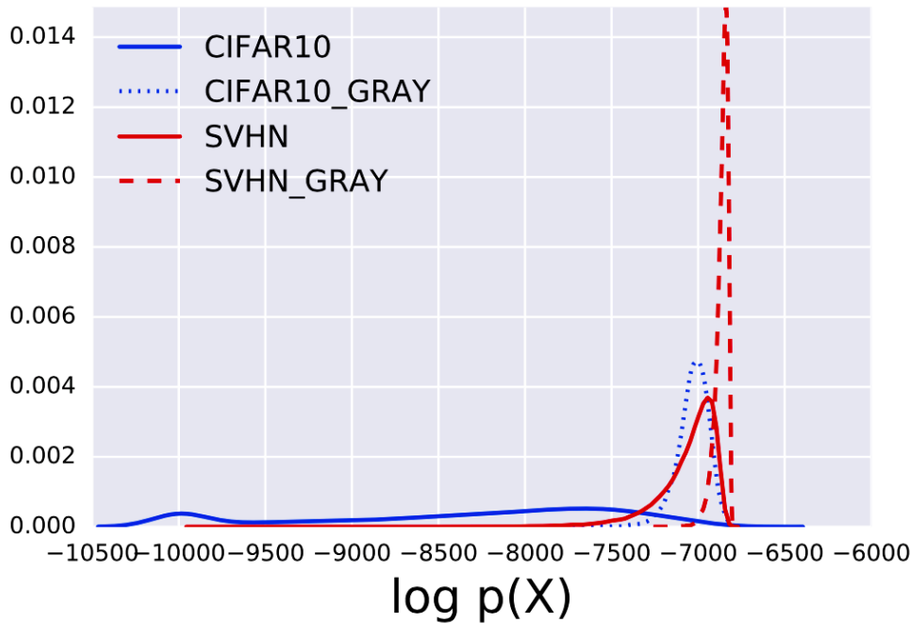
Differences in variances in the three spatial dimensions

Анализ CV-GLOW





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



(b) Graying images increases likelihood.

Выводы

- У существующих генеративных моделей есть проблемы с выявлением OOD
- Для flow-based моделей это объясняется разницей в дисперсии распределений
- Нужно быть осторожным при использовании генеративных моделей для выявления аномалий

Вопросы

- Почему нужно прибавлять случайный шум к данным при обучении генеративных моделей? Что такое BPD?

- Какой эффект обнаружили авторы при сравнении правдоподобия данных из обучения и из другого домена? Ожидаемый ли он? Почему?

- Выписать неравенство, которым объясняется феномен для Constant-Volume GLOW. Что значит каждый член этого неравенства?