ВАРИАЦИОННЫЕ АВТОКОДИРОВЩИКИ

Сергей Николенко

СПбГУ — Санкт-Петербург

6 ноября 2026 г.





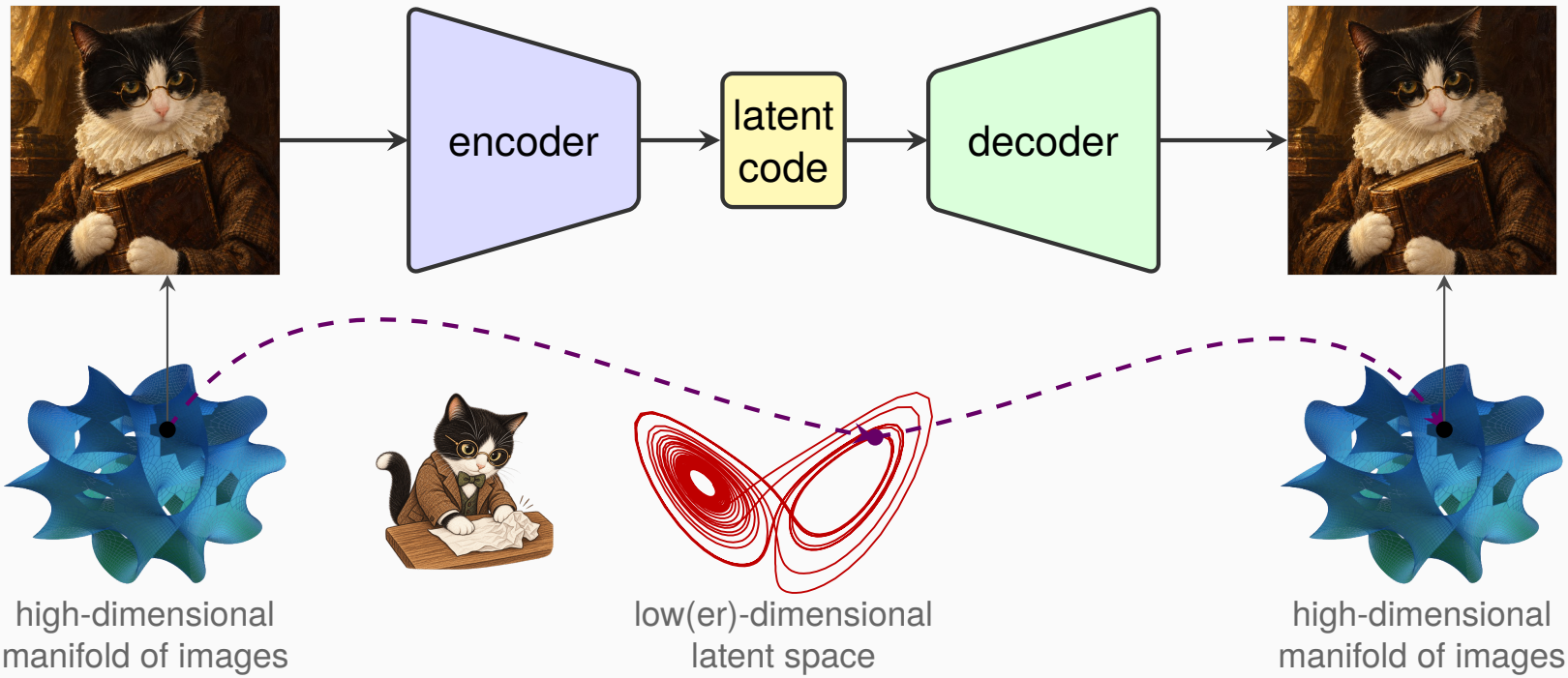
Random facts:

- 6 ноября 1528 г. потерпевший кораблекрушение Кабеса де Вака высадился у берегов Техаса; возвращение к испанцам заняло восемь лет
- 6 ноября 1860 г. Авраам Линкольн выиграл президентские выборы в США
- 6 ноября 1888 г. Бенджамин Гаррисон победил на выборах президента США, хотя Гровер Кливленд получил больше голосов избирателей
- 6 ноября 1935 г. впервые взлетел прототип Hawker Hurricane — одного из основных британских истребителей Битвы за Британию
- 6 ноября 1947 г. на NBC впервые вышла телевизионная программа Meet the Press
- 6 ноября 1986 г. Алекс Фергюсон возглавил «Манчестер Юнайтед»; он оставался главным тренером до 2013 года

Вариационные автокодировщики

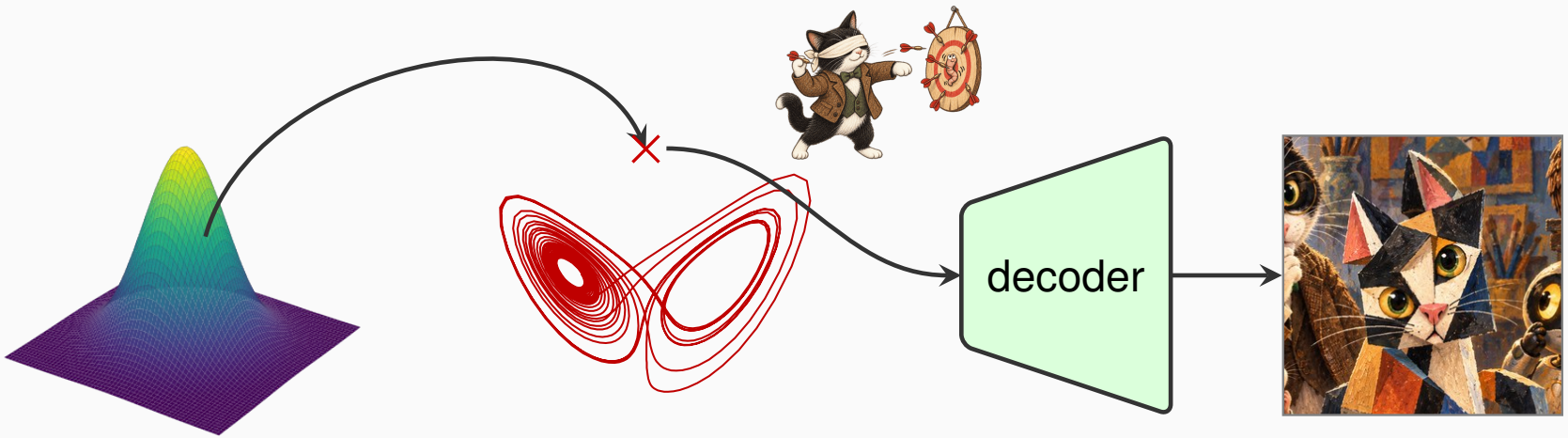
Идея и мотивация

- Главная задача — как всегда: как сделать из обычного автокодировщика порождающую модель?



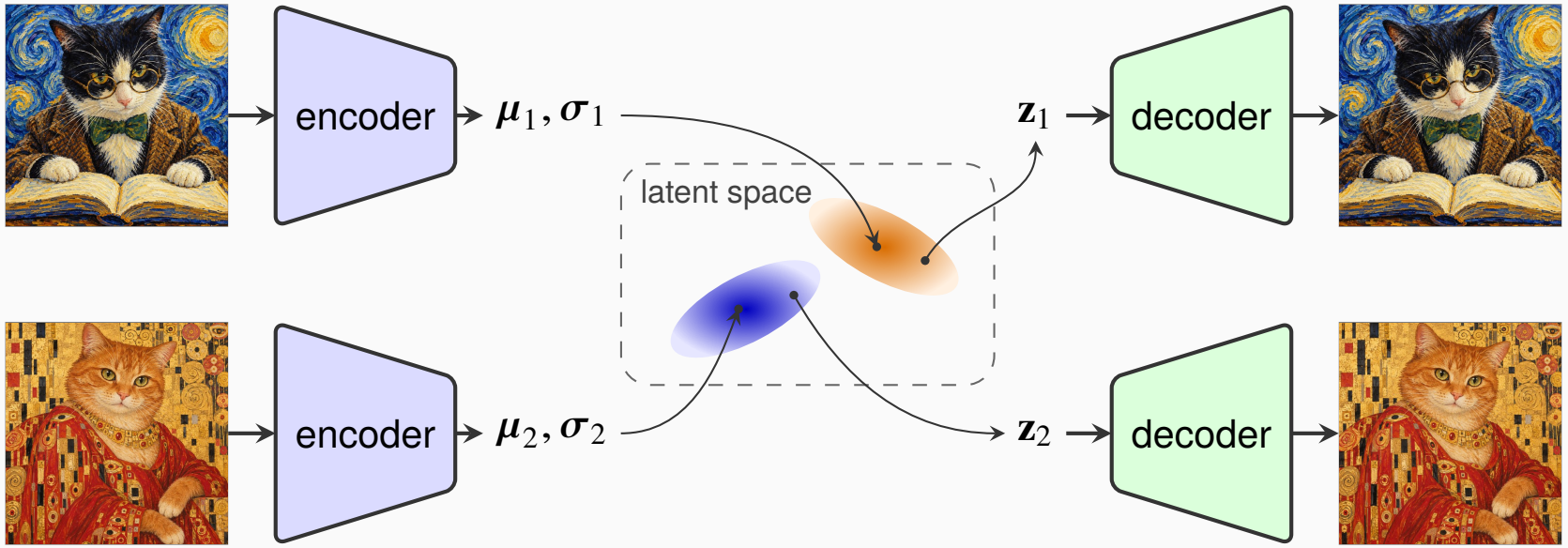
Идея и мотивация

- Даже в пространстве низкой размерности нелегко выбрать хорошую точку:



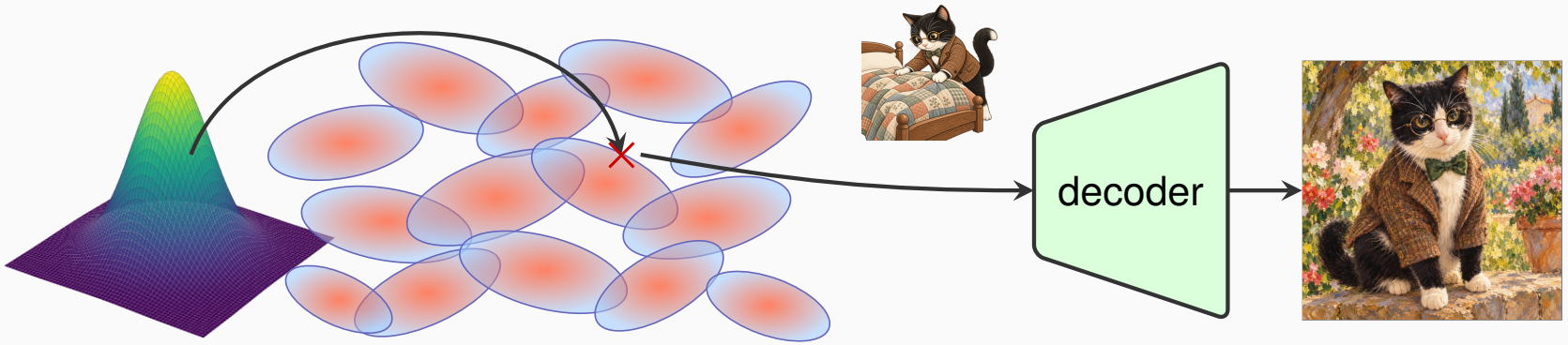
Идея и мотивация

- Главная идея вариационного автокодировщика (Kingma et al., 2013) — пусть из каждой точки $\mathbf{x}$ получается не точка $\mathbf{z}$, а целое распределение $p(\mathbf{z}|\mathbf{x})$:



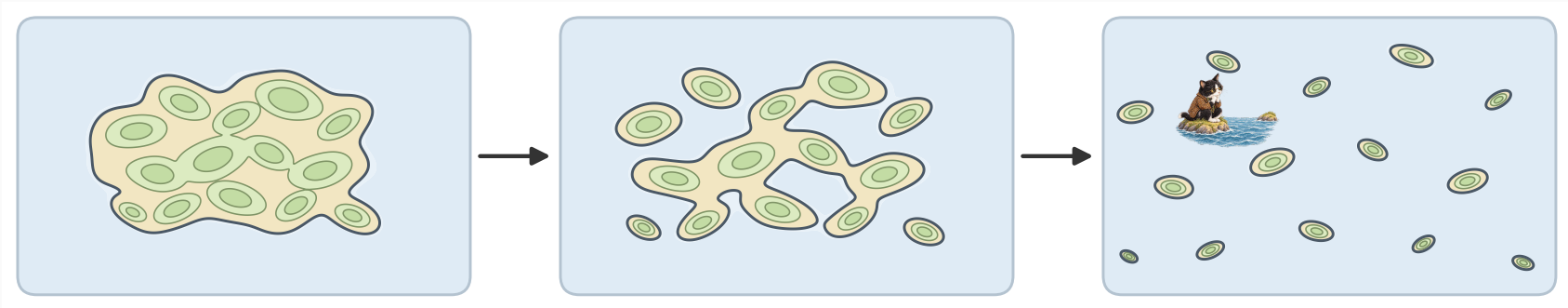
Идея и мотивация

- Интуиция в том, что декодеру приходится быть устойчивым к небольшим изменениям в $\mathbf{z}$; некий более идейный аналог шумоподавляющего автокодировщика
- В идеале мы покроем пространство латентных кодов достаточно «широкими» распределениями:



Идея и мотивация

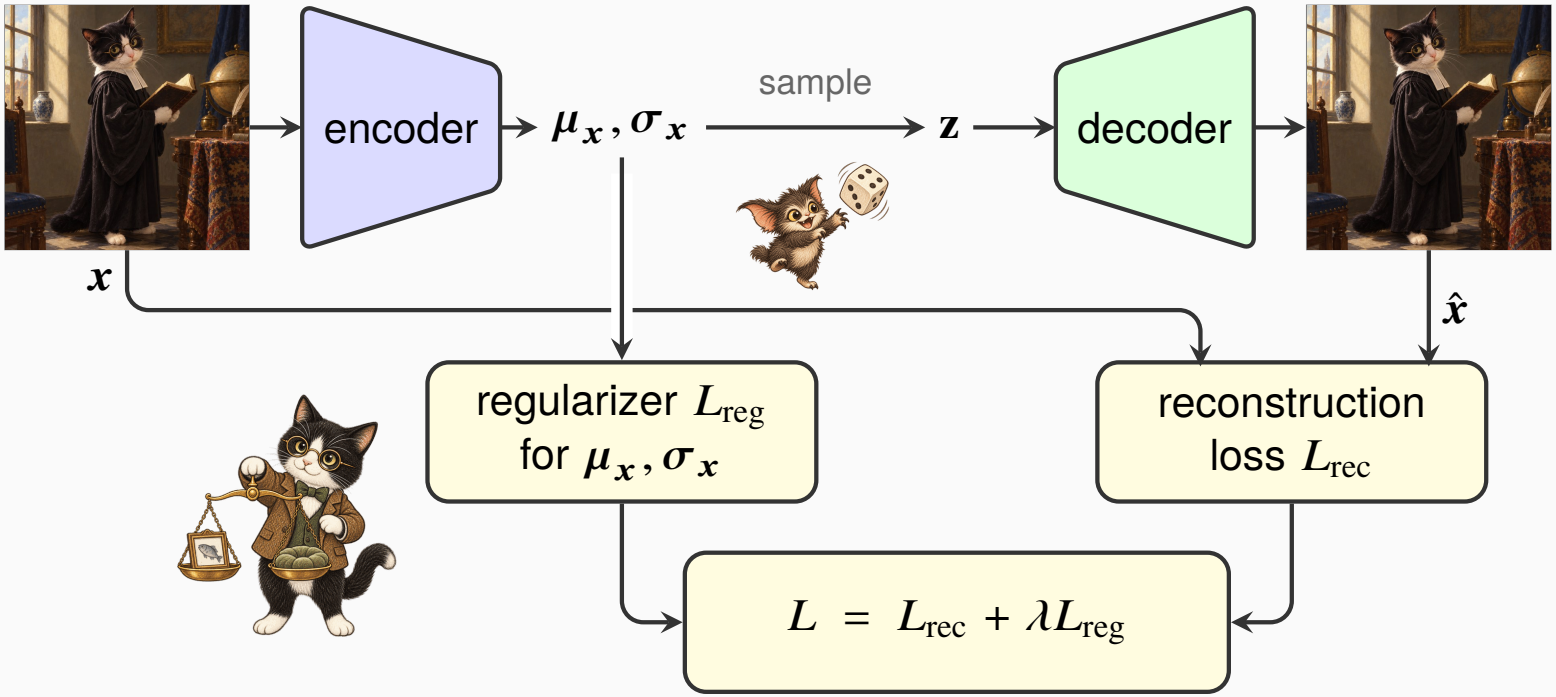
- Однако есть проблема: если просто обучать автокодировщик, ему будет легче декодировать «узкие» распределения, и мы в итоге придём к тем же точкам с нулевой дисперсией



- Как решить эту проблему?

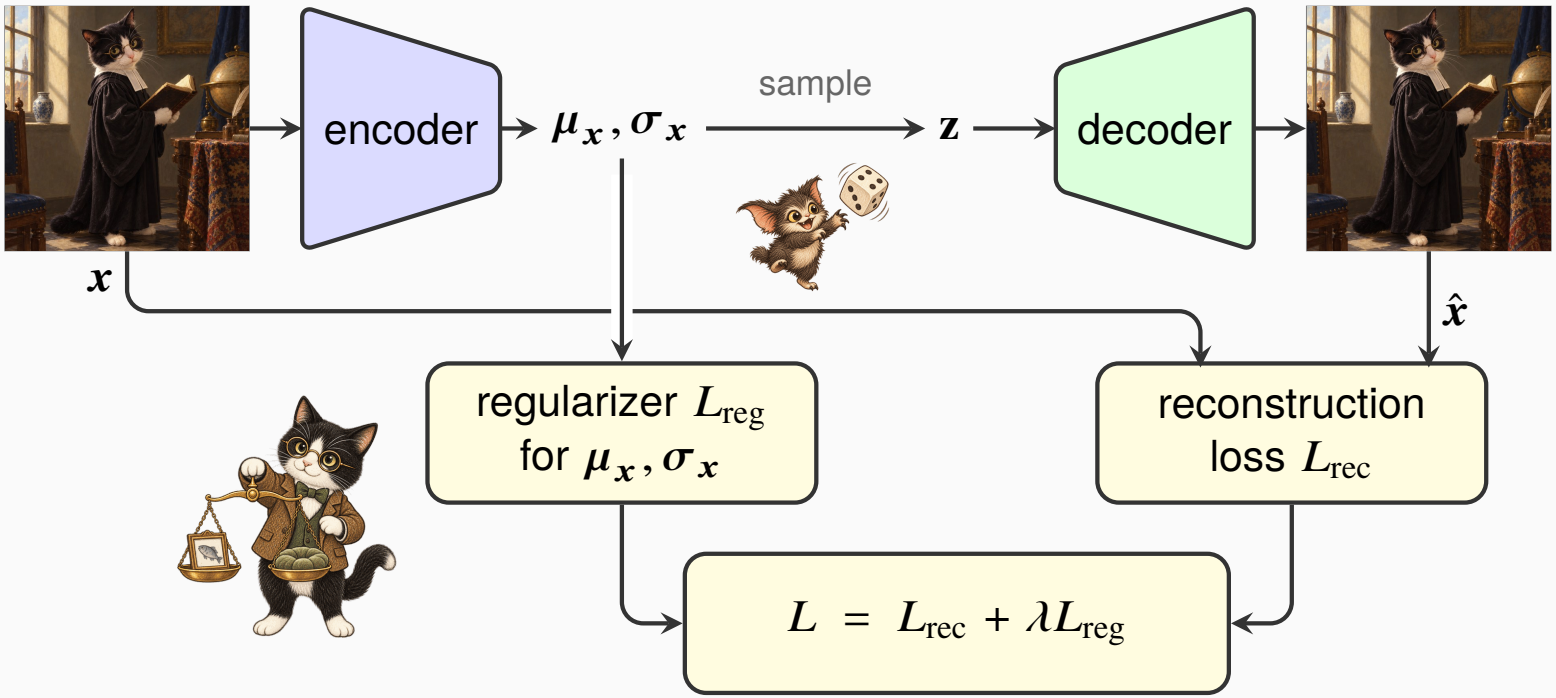
Идея и мотивация

- Нужно добавить некое ограничение на $p(\mathbf{z}|\mathbf{x})$, то есть регуляризовать его:



Идея и мотивация

- Теперь остались два вопроса:
  - какие выбрать функцию ошибки и регуляризатор?
  - как протаскивать градиенты через сэмплирование?
- Давайте на оба этих вопроса ответим...



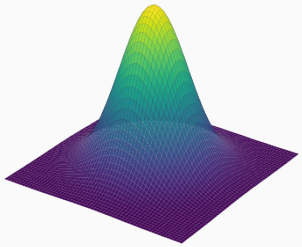
Вариационные приближения в VAE

- Давайте начнём с начала — мы хотим обучить кодировщик делать латентные коды $\mathbf{z}$ из картинок $\mathbf{x}$ и декодировщик, который $\mathbf{z}$ превращает обратно в $\mathbf{x}$.
- Совместное распределение можно разложить по-разному:

$$p(\mathbf{z})\,p(\mathbf{x}|\mathbf{z}) = p(\mathbf{x},\mathbf{z}) = p(\mathbf{x})\,p(\mathbf{z}|\mathbf{x})$$

$$p(\mathbf{z})\;\; p(\boldsymbol{x}|\mathbf{z}) \;\; = \;\; p(\boldsymbol{x},\mathbf{z}) \;\; = \;\; p(\boldsymbol{x})\;\; p(\mathbf{z}|\boldsymbol{x})$$

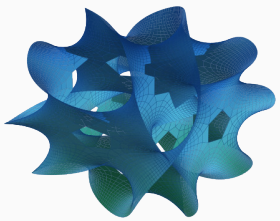
distribution we want to sample $\mathbf{z}$ from



decoder distribution



high-dimensional manifold of real images



encoder distribution

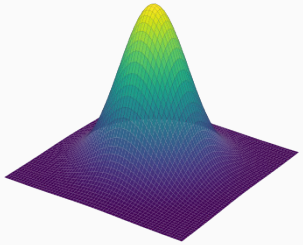


Вариационные приближения в VAE

- Здесь:
  - $p(\mathbf{x})$ — это распределение изображений,
  - $p(\mathbf{z})$ — это распределение латентных кодов, т.е. простое распределение, из которого можно сэмплировать;
  - а остальные два нам нужно найти — это распределение кодировщика $p(\mathbf{z}|\mathbf{x})$ и декодировщика $p(\mathbf{x}|\mathbf{z})$.

$$p(\mathbf{z})\ p(\boldsymbol{x}|\mathbf{z}) = p(\boldsymbol{x}, \mathbf{z}) = p(\boldsymbol{x})\ p(\mathbf{z}|\boldsymbol{x})$$

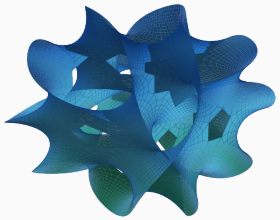
distribution we want to sample $\mathbf{z}$ from



decoder distribution



high-dimensional manifold of real images



encoder distribution



Вариационные приближения в VAE

- Для порождающей модели надо обучить и $p(\mathbf{x}|\mathbf{z})$, и $p(\mathbf{z}|\mathbf{x})$.
- Но у нас $p(\mathbf{z})$ — это простое распределение, а $p(\mathbf{x})$ — очень сложное!
- Значит, придётся выбрать одно из двух:
  - или предположим, что $p(\mathbf{x}|\mathbf{z})$ простое, а затем будем искать сложное $p(\mathbf{z}|\mathbf{x})$;
  - или наоборот, предположим, что $p(\mathbf{z}|\mathbf{x})$ простое, а затем будем искать сложное $p(\mathbf{x}|\mathbf{z})$.

Вариационные приближения в VAE

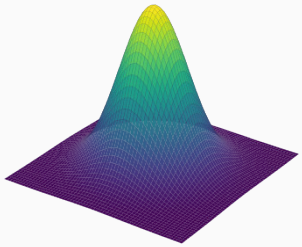
- В VAE удобнее использовать первый вариант: предположим, что

$$p(\mathbf{x}|\mathbf{z}) = N(\mathbf{x}|f(\mathbf{z}), c\mathbf{I}), \quad \text{где} \quad f(\mathbf{z}) = \text{Decoder}(\mathbf{z})$$

- Слева получается $p(\mathbf{x}|\mathbf{z})$ умножить на $p(\mathbf{z}) = N(\mathbf{z}|0, \mathbf{I})$, то есть совместное распределение $p(\mathbf{x}, \mathbf{z})$.

$$p(\mathbf{z})\ p(\boldsymbol{x}|\mathbf{z}) = p(\boldsymbol{x}, \mathbf{z}) = p(\boldsymbol{x})\ p(\mathbf{z}|\boldsymbol{x})$$

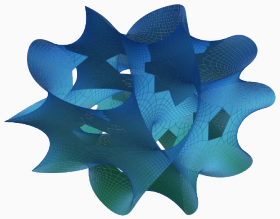
distribution we want to sample $\mathbf{z}$ from



decoder distribution



high-dimensional manifold of real images



encoder distribution



Вариационные приближения в VAE

- А справа будем делать приближение $q(\mathbf{z}) \approx p(\mathbf{z}|\mathbf{x})$:

$$p(\mathbf{z})\ p(\boldsymbol{x}|\mathbf{z}) = p(\boldsymbol{x}, \mathbf{z}) = p(\boldsymbol{x})\ p(\mathbf{z}|\boldsymbol{x}) \approx q(\mathbf{z})$$



distribution we want to sample $\mathbf{z}$ from

decoder distribution $\mathcal{N}(\boldsymbol{x}|f(\mathbf{z}), c\mathbf{I})$

high-dimensional manifold of real images

encoder distribution $\mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_{\boldsymbol{x}}, \boldsymbol{\sigma}^2_{\boldsymbol{x}})$



- А именно вариационное приближение...

Вариационные приближения в VAE

- Вспомним идею вариационных приближений:

$$p(\mathbf{x}, \mathbf{z}) = p(\mathbf{x})\, p(\mathbf{z}|\mathbf{x}),$$

$$\log p(\mathbf{x}) = \log p(\mathbf{x}, \mathbf{z}) - \log p(\mathbf{z}|\mathbf{x}),$$

возьмём ожидание по $q(\mathbf{z})$:

$$\mathbb{E}_{q(\mathbf{z})}[\log p(\mathbf{x})] = \mathbb{E}_{q(\mathbf{z})}[\log p(\mathbf{x}, \mathbf{z})] - \mathbb{E}_{q(\mathbf{z})}[\log p(\mathbf{z}|\mathbf{x})],$$

а затем сделаем стандартные преобразования:

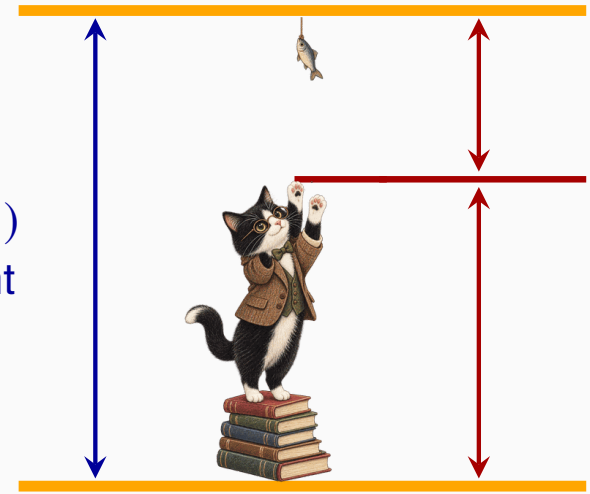
$$\begin{aligned}\log p(\mathbf{x}) =& \mathbb{E}_{q(\mathbf{z})}[\log p(\mathbf{x}, \mathbf{z})] - \mathbb{E}_{q(\mathbf{z})}[\log q(\mathbf{z})] + \\ &+ \mathbb{E}_{q(\mathbf{z})}[\log q(\mathbf{z})] - \mathbb{E}_{q(\mathbf{z})}[\log p(\mathbf{z}|\mathbf{x})], \\ \log p(\mathbf{x}) =& \int q(\mathbf{z}) \log \frac{p(\mathbf{x}, \mathbf{z})}{q(\mathbf{z})} \mathrm{d}\mathbf{z} + \int q(\mathbf{z}) \log \frac{q(\mathbf{z})}{p(\mathbf{z}|\mathbf{x})} \mathrm{d}\mathbf{z}.\end{aligned}$$

Вариационные приближения в VAE

- Итого получается, что

$$\log p\left(\mathbf{x}\right) = L(q) + \mathrm{KL}\left(q\left(z\right) \| p\left(\mathbf{z}|\mathbf{x}\right)\right), \quad \text{где}$$

$$L(q) = \int q\left(\mathbf{z}\right) \log \frac{p\left(\mathbf{x}, \mathbf{z}\right)}{q\left(\mathbf{z}\right)} \mathrm{d}\mathbf{z}.$$



Kullback–Leibler divergence

$$\mathrm{KL}\left(q\left(\mathbf{z}\right) \| p\left(\mathbf{z}|\boldsymbol{x}\right)\right) = \int q\left(\mathbf{z}\right) \log \frac{q\left(\mathbf{z}\right)}{p\left(\mathbf{z}|\boldsymbol{x}\right)} \mathrm{d}\mathbf{z}$$

evidence lower bound (ELBO)

$$\mathcal{L}(q) = \int q\left(\mathbf{z}\right) \log \frac{p\left(\boldsymbol{x}, \mathbf{z}\right)}{q\left(\mathbf{z}\right)} \mathrm{d}\mathbf{z}$$

Вариационные приближения в VAE

- Значит, можно аппроксимировать $p(\mathbf{z}|\mathbf{x})$ через $q(\mathbf{z})$, максимизируя $L(q)$. В случае VAE его можно найти:

$$
\begin{aligned}
L(q) &= \int q(\mathbf{z}) \log \frac{p(\mathbf{x}, \mathbf{z})}{q(\mathbf{z})} \mathrm{d}\mathbf{z} = \int q(\mathbf{z}) \log \frac{p(\mathbf{z})\, p(\mathbf{x}|\mathbf{z})}{q(\mathbf{z})} \mathrm{d}\mathbf{z} \\
&= \int q(\mathbf{z}) \log p(\mathbf{x}|\mathbf{z}) \mathrm{d}\mathbf{z} + \int q(\mathbf{z}) \log \frac{p(\mathbf{z})}{q(\mathbf{z})} \mathrm{d}\mathbf{z} \\
&= \int q(\mathbf{z}) \log N(\mathbf{x}|f(\mathbf{z}), c\mathbf{I}) \mathrm{d}\mathbf{z} - \mathrm{KL}(q(\mathbf{z}) \| p(\mathbf{z})) \\
&= -\frac{1}{2c} \mathbb{E}_{q(\mathbf{z})} \left[\|\mathbf{x} - f(\mathbf{z})\|^2 \right] - \mathrm{KL}(q(\mathbf{z}) \| p(\mathbf{z})) .
\end{aligned}
$$

Вариационные приближения в VAE

- И вот получилось как раз то, чего мы ожидали интуитивно:

$$-\mathcal{L}(q) = \frac{1}{2c}\mathbb{E}_{q(\mathbf{z})}\left[\|\boldsymbol{x} - f(\mathbf{z})\|^2\right] + \mathrm{KL}\left(q\left(\mathbf{z}\right)\|p\left(\mathbf{z}\right)\right)$$



reconstruction loss L_{rec}

regularizer L_{reg} for $\boldsymbol{\mu}_{\boldsymbol{x}}, \boldsymbol{\sigma}_{\boldsymbol{x}}$

$$L = L_{\text{rec}} + \lambda L_{\text{reg}}$$

Вариационные приближения в VAE

- Только теперь мы знаем, как выглядит L_{rec}, да и KL-дивергенцию между двумя гауссианами можно подсчитать:

$$\mathrm{KL}\left(q\left(\mathbf{z}\right)\|p\left(\mathbf{z}\right)\right) = \frac{1}{2}\sum_{j=1}^{d}\left(\sigma_{\mathbf{x},j}^2 + \mu_{\mathbf{x},j}^2 - \log\sigma_{\mathbf{x},j}^2 - 1\right).$$

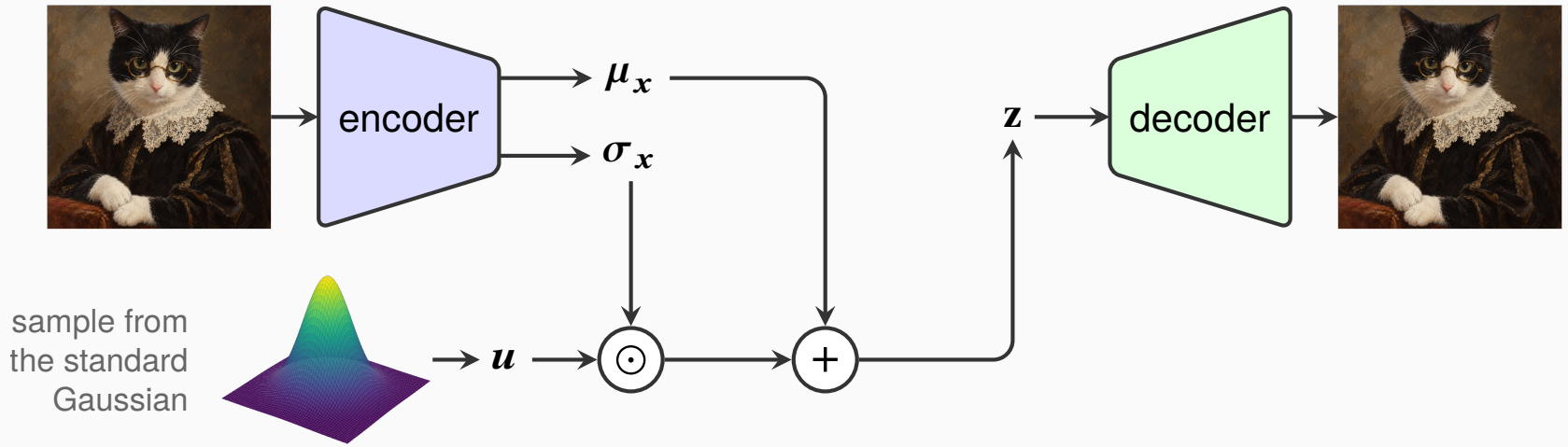
- Получается окончательная функция ошибки для вариационного автокодировщика:

$$L(q) = -\frac{1}{2c}\mathbb{E}_{q(\mathbf{z})}\left[\|\mathbf{x} - f(\mathbf{z})\|^2\right] - \frac{1}{2}\sum_{j=1}^{d}\left(\sigma_{\mathbf{x},j}^2 + \mu_{\mathbf{x},j}^2 - \log\sigma_{\mathbf{x},j}^2 - 1\right).$$

- Осталась только проблема с сэмплированием...

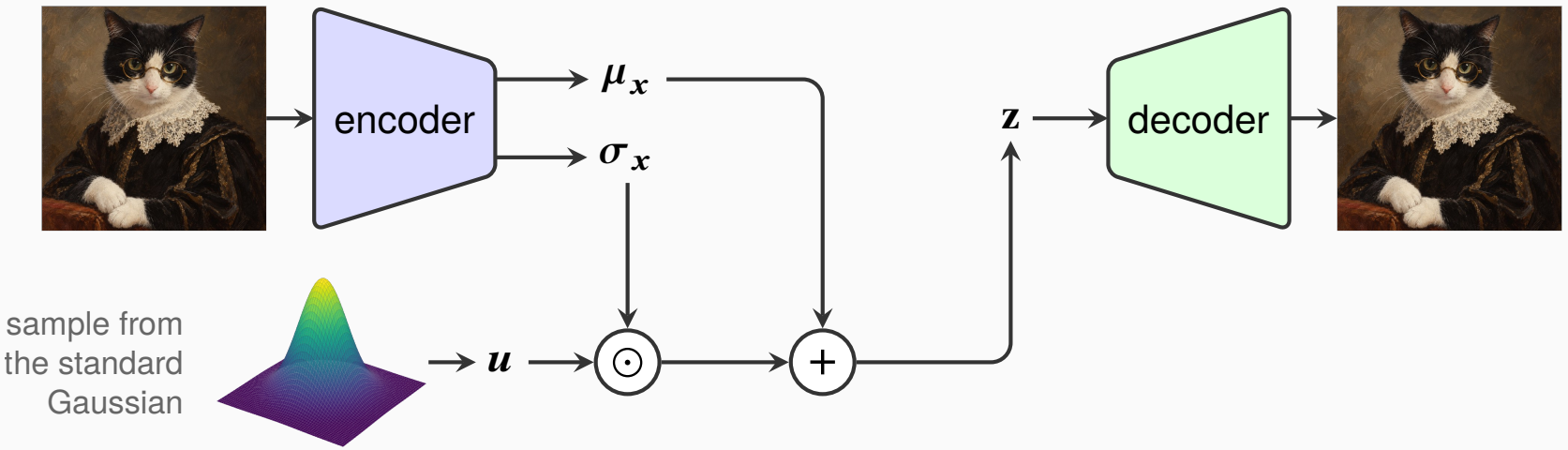
Reparametrization trick

- Чтобы обучить модель, нужно пропустить градиент через $\mathbf{z}$ обратно к кодировщику, но там сэмплирование
- Reparametrization trick: давайте сначала сэмплировать, а потом преобразовывать результат!



Reparametrization trick

- И теперь уже точно все компоненты на месте: сэмплируем мини-батч $\mathbf{u}$ для мини-батча $\mathbf{x}$ и используем их для обучения



Спасибо за внимание!



