NVAE: A Deep Hierarchical Variational Autoencoder

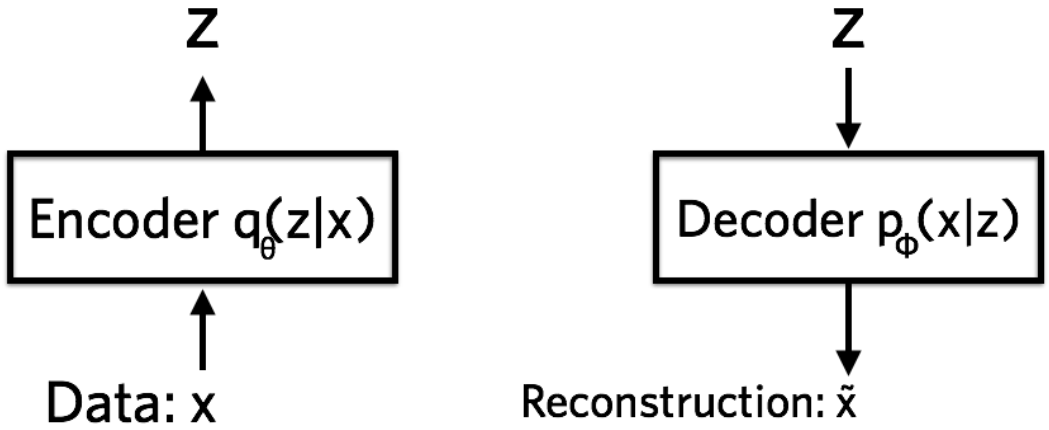
Ахметов Артемий

VAE

VAE - variational autoencoders. Энкодер переводит объект в скрытое пространство z, после чего декодер пытается восстановить исходный объект.

Скрытое пространство - bottleneck, поэтому энкодеру нужно сохранить максимальное количество информации об объекте в z

Отличие VAE от обычных автоэнкодеров - p(z)~Normal(0,1)



$$l_i(\theta, \phi) = -\mathbb{E}_{z \sim q_\theta(z|x_i)}[\log p_\phi(x_i \mid z)] + \mathbb{KL}(q_\theta(z \mid x_i) \,||\, p(z))$$

В чем проблема?

- VAE пытаются сохранить как можно больше информации, однако строятся в основном на архитектурах для классификации, отбрасывающих много информации.
- VAE часто по-разному реагируют на большое количество параметров.
- VAE моделируют долговременные корелляции в данных из за чего в сети должно быть большие поля восприятия(receptive field)
- Из за KL дивергенции обучение глубоких VAE нестабильно

NVAE

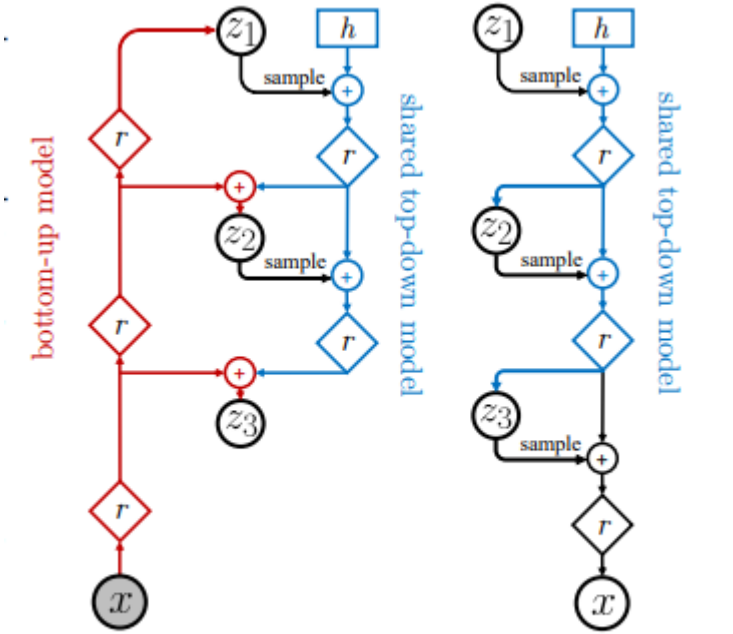
Nouveau VAE - глубокий иерархический VAE

- В NVAE используются глубокие свертки
- Residual блоки
- Спектральная регуляризация для стабилизации обучения
- Уменьшение затрат по памяти в сравнении с VAE
- NVAE первое успешное применение VAE на изображения разрешения больше 256x256

Архитектура

В глубоких иерархических VAE скрытые переменные z распределяются на L групп

Чтобы уменьшить затратность обучения, используется bidirectional encoder, часть которое позже становится генеративной моделью.



(a) Bidirectional Encoder (b) Generative Model

$$\mathcal{L}_{\text{VAE}}(\boldsymbol{x}) := \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}\left[\log p(\boldsymbol{x}|\boldsymbol{z})\right] - \text{KL}(q(\boldsymbol{z}_1|\boldsymbol{x})||p(\boldsymbol{z}_1)) - \sum_{l=2}^{L} \mathbb{E}_{q(\boldsymbol{z}_{<l}|\boldsymbol{x})}\left[\text{KL}(q(\boldsymbol{z}_l|\boldsymbol{x}, \boldsymbol{z}_{<l})||p(\boldsymbol{z}_l|\boldsymbol{z}_{<l}))\right]$$

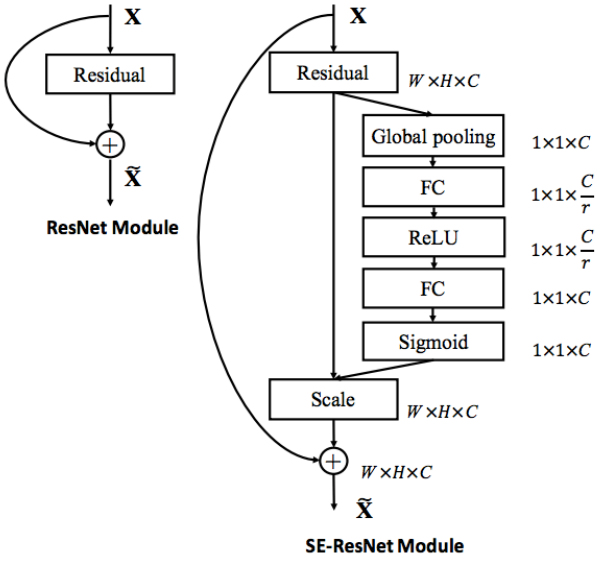
Residual блоки

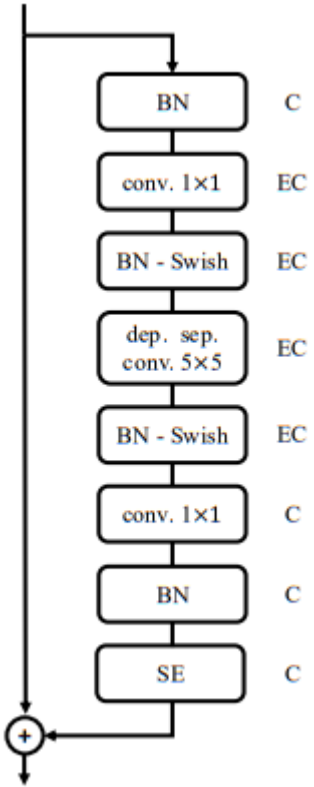
Residual блоки используются для сохранения долговременных корреляций при их прохождении сквозь сеть.

Блоки генеративной модели

- Используется свертка с большим ядром
- Используется особенный BN
- Swish активация показывает обнадеживающие результаты
- Squeeze and Excitation (SE) - gating слой

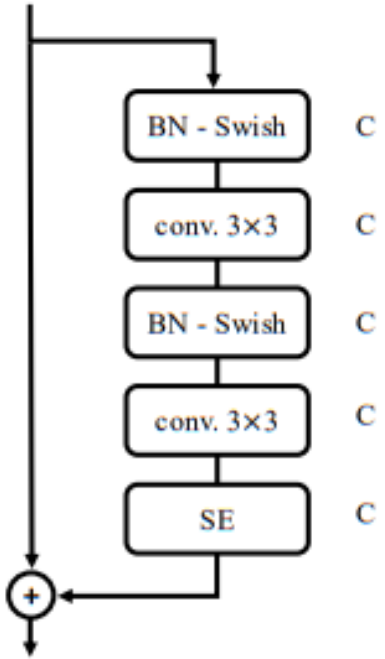
$$f(u) = \frac{u}{1+e^{-u}}$$





Блоки в энкодере

Сделан вывод что BN-Activation-Conv лучше чем Conv-BN-Activation



Проблемы с памятью

Из за использования глубинных сверток сильно возрастают требования к памяти. Что делать?

1. Использовать NVIDEA APEX, для переведения float в half-precision float. Уменьшает затраты по памяти на 40%!
2. Копия карты признаков хранится при каждом backward проходе. Используется gradient check-pointing и BN пересчитывается при каждом проходе назад. Не сильно замедляет обучение но памяти требуется еще на 18% меньше.

Обучение

Из за присутствия KL в функции потерь, при большой разнице в распределениях энкодера и декодера обучение происходит нестабильно. Есть два решения:

1. Переопределение KL
2. Спектральная регуляризация. Добавим к L_vae регуляризацию на константу Липшица

$$p(z_l^i|\boldsymbol{z}_{<l}) := \mathcal{N}\left(\mu_i(\boldsymbol{z}_{<l}), \sigma_i(\boldsymbol{z}_{<l})\right)$$
$$q(z_l^i|\boldsymbol{z}_{<l}, \boldsymbol{x}) := \mathcal{N}\left(\mu_i(\boldsymbol{z}_{<l}) + \Delta\mu_i(\boldsymbol{z}_{<l}, \boldsymbol{x}), \sigma_i(\boldsymbol{z}_{<l}) \cdot \Delta\sigma_i(\boldsymbol{z}_{<l}, \boldsymbol{x})\right)$$

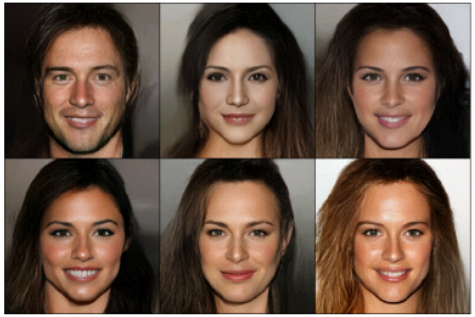
$$\mathrm{KL}\big(q(z^i|\boldsymbol{x})||p(z^i)\big) = \frac{1}{2}\left(\frac{\Delta\mu_i^2}{\sigma_i^2} + \Delta\sigma_i^2 - \log\Delta\sigma_i^2 - 1\right)$$

$$\mathcal{L}_{SR} = \lambda \sum_i s^{(i)}$$

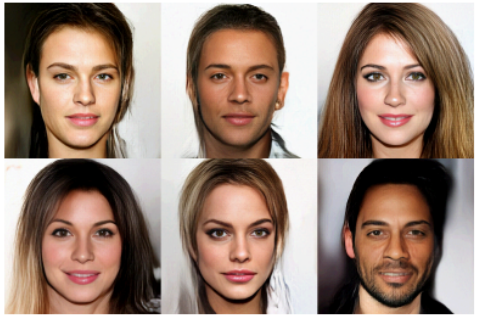
Эксперименты

(d) CelebA HQ ($t = 0.6$)

(e) FFHQ ($t = 0.5$)



(f) MaCow [67] trained on CelebA HQ ($t = 0.7$)



(g) Glow [62] trained on CelebA HQ ($t = 0.7$)

Method	MNIST 28×28	CIFAR-10 32×32	ImageNet 32×32	CelebA 64×64	CelebA HQ 256×256	FFHQ 256×256
NVAE w/o flow	**78.01**	2.93	-	2.04	-	0.71
NVAE w/ flow	78.19	**2.91**	3.92	**2.03**	**0.70**	**0.69**
VAE Models with an Unconditional Decoder						
BIVA [36]	78.41	3.08	3.96	2.48	-	-
IAF-VAE [4]	79.10	3.11	-	-	-	-
DVAE++ [20]	78.49	3.38	-	-	-	-
Conv Draw [42]	-	3.58	4.40	-	-	-
Flow Models without any Autoregressive Components in the Generative Model						
VFlow [59]	-	2.98	-	-	-	-
ANF [60]	-	3.05	3.92	-	0.72	-
Flow++ [61]	-	3.08	**3.86**	-	-	-
Residual flow [50]	-	3.28	4.01	-	0.99	-
GLOW [62]	-	3.35	4.09	-	1.03	-
Real NVP [63]	-	3.49	4.28	3.02	-	-
VAE and Flow Models with Autoregressive Components in the Generative Model						
δ-VAE [25]	-	2.83	3.77	-	-	-
PixelVAE++ [35]	78.00	2.90	-	-	-	-
VampPrior [64]	78.45	-	-	-	-	-
MAE [65]	77.98	2.95	-	-	-	-
Lossy VAE [66]	78.53	2.95	-	-	-	-
MaCow [67]	-	3.16	-	-	0.67	-
Autoregressive Models						
SPN [68]	-	-	3.85	-	0.61	-
PixelSNAIL [34]	-	2.85	3.80	-	-	-
Image Transformer [69]	-	2.90	3.77	-	-	-
PixelCNN++ [70]	-	2.92	-	-	-	-
PixelRNN [41]	-	3.00	3.86	-	-	-
Gated PixelCNN [71]	-	3.03	3.83	-	-	-

Источники

1. VAE - https://jaan.io/what-is-variational-autoencoder-vae-tutorial/
2. NVAE - https://arxiv.org/abs/2007.03898