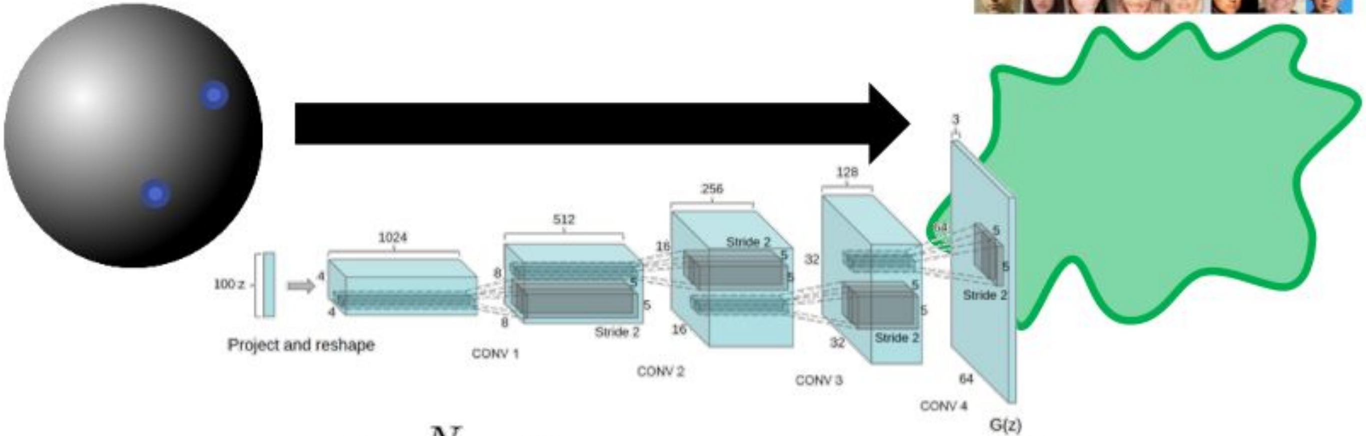
Latent models

Latent models of images

[Bojanowski et al. 2017]: the simplest deep latent model for images:

$$\mathcal{Z} = \mathcal{B}(r, d, p) = \left\{ z \in \mathbb{R}^d : \|z\|_p \leq r \right\}$$



$$\min_{\theta \in \Theta} \frac{1}{N} \sum_{i=1}^{N} \left[\min_{z_i \in \mathcal{Z}} \ell\left(g_\theta(z_i), x_i\right) \right]$$

Latent models of images: reconstructions

d =512

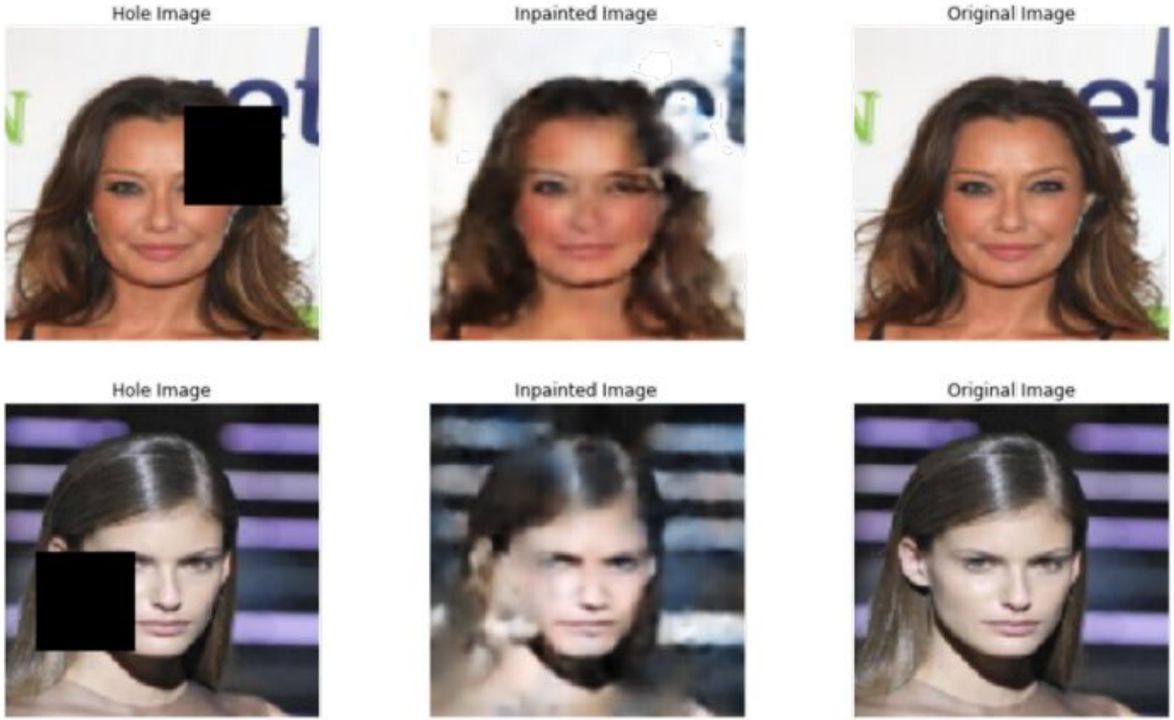
[Bojanowski et al. 2017]

Latent models of images: samples

1. Fit Gaussian in the latent space
2. Sample and generate

[Bojanowski et al. 2017]

Latent models of images: restoration



$$\min_{z \in \mathcal{Z}} \|(g_\theta(z) - x) \odot m\|$$

[thanks to ShahRukh Athar]

Problems with direct optimization

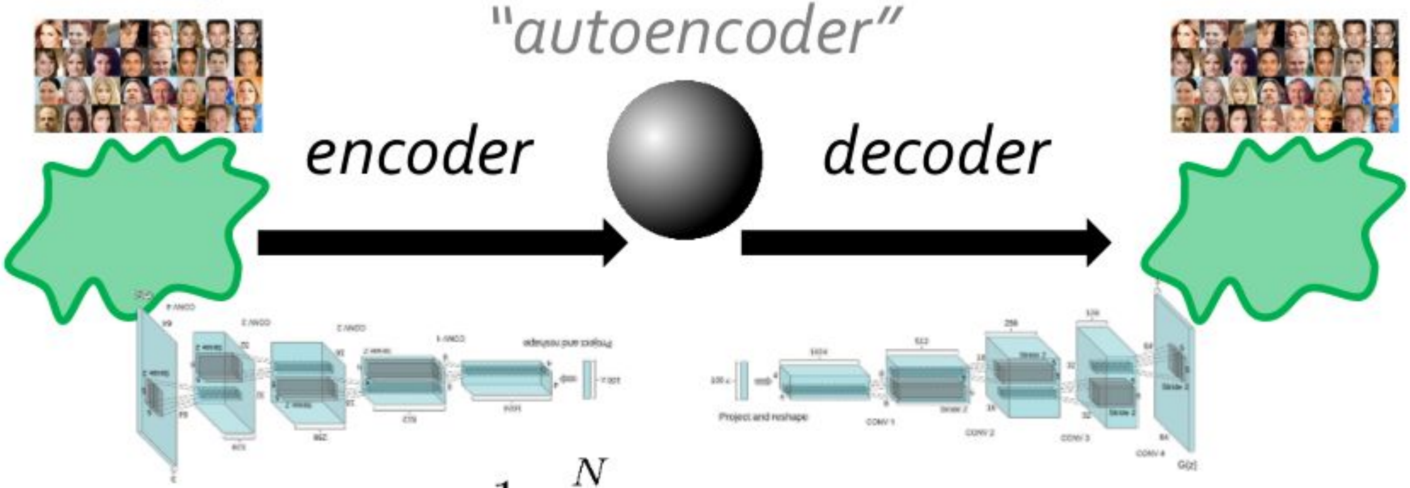
- Previous model requires optimization to fit new images:

$$\min_{\theta \in \Theta} \frac{1}{N} \sum_{i=1}^{N} \left[\min_{z_i \in \mathcal{Z}} \ \ell\left(g_\theta(z_i), x_i\right) \right]$$

- Previous model requires storing latent vectors during learning (not scalable)

- **Idea**: predict latent vectors with a new network (*encoder*) from the image

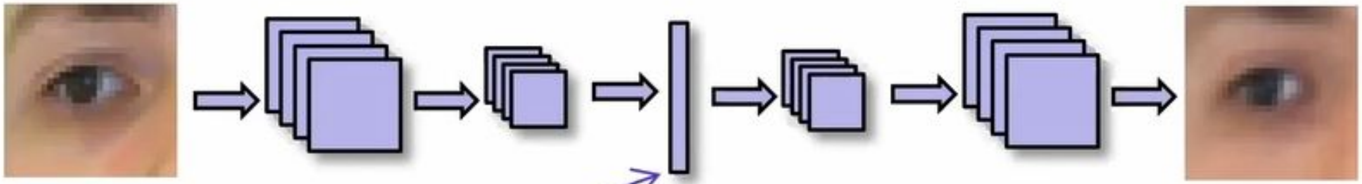
From direct optimization to Autoencoders



$$\min_{\phi,\theta} \frac{1}{N} \sum_{i=1}^{N} l(g_\theta(e_\phi(x_i)), x_i)$$

- Learning still unsupervised
- No scalability issues
- A lot depends on the loss

Smart image editing

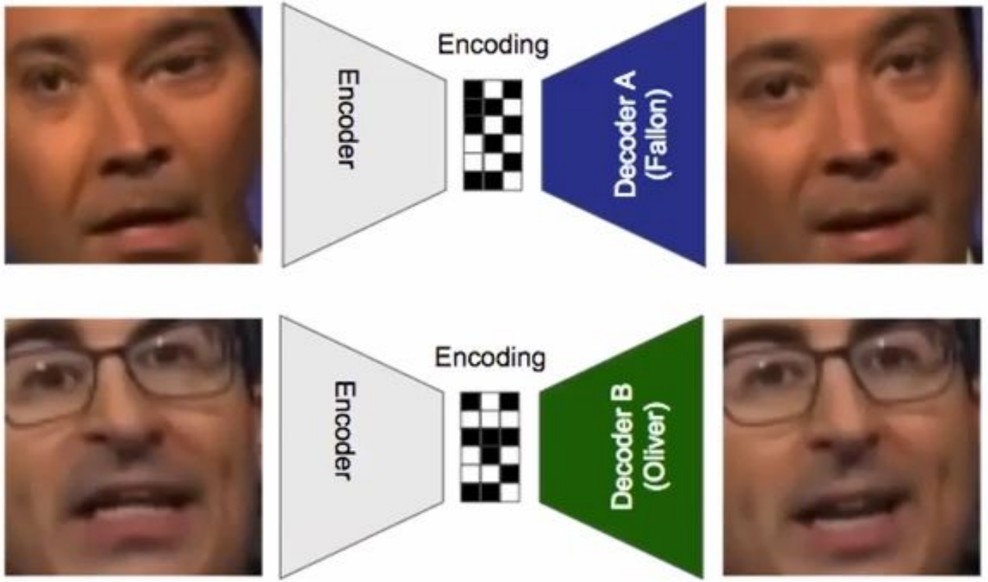


Low-dim space, where we can estimate semantically meaningful directions

Given a bunch of pairs we can estimate a vector for gaze redirection

DeepFake system

Paired "dewarping" autoencoders:



[images source: Gaurav Oberoi blog]

DeepFake system

[images source: Gaurav Oberoi blog]

DeepFake system

DeepFake system

[video source: Gaurav Oberoi blog]

Variational Auto Encoders

Ideally, we want to do maximum likelihood learning:

$$\frac{1}{N}\sum_{i=1}^{N}\log\left(\int_z p_\theta(x_i|z)p(z)dz\right) \to \max_\theta$$

$$p_\theta(x|z)$$

[Kingma & Welling 14]

Variational Auto Encoders

Ideally, we want to do maximum likelihood learning:

$$\frac{1}{N}\sum_{i=1}^{N}\log\left(\int_z p_\theta(x_i|z)p(z)dz\right) \to \max_\theta$$

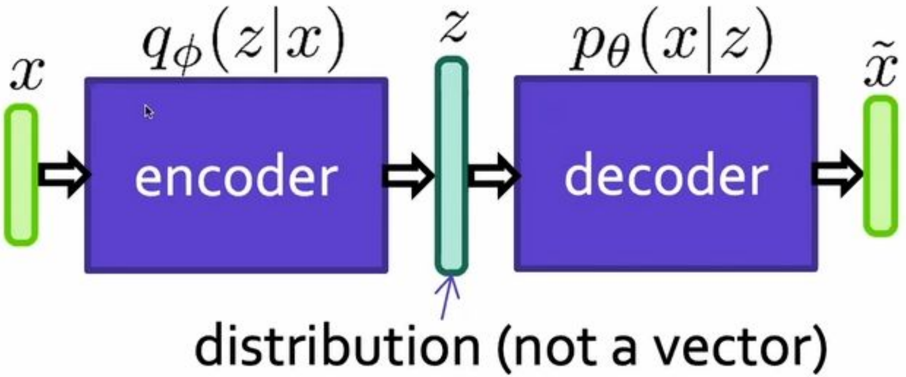
$p_\theta(z|x) =?$

$p_\theta(x|z)$

Idea 1: use $q_\phi(z|x)$ instead of $p_\theta(z|x)$

[Kingma & Welling 14]

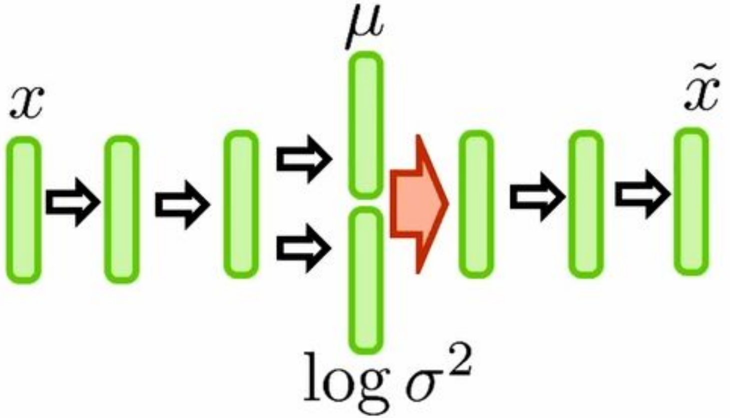
Variational Auto Encoders



[Kingma & Welling 14]

Variational Auto Encoders

E.g.:

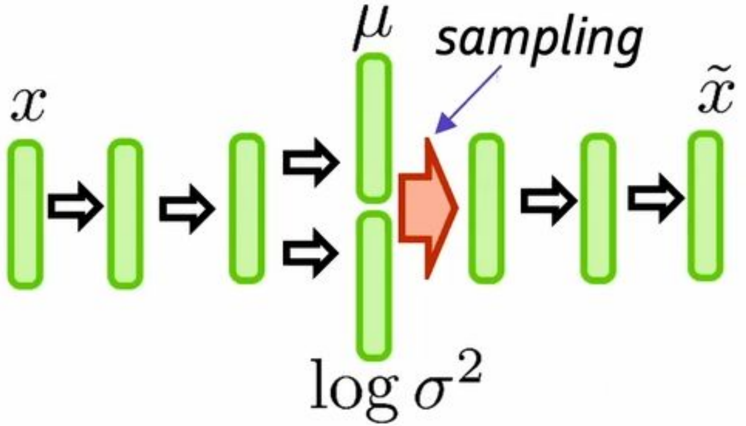


$$[\mu, \sigma] = e_\phi(x) \quad q_\phi(z|x) = \mathcal{N}(\mu, \mathrm{diag}(\sigma^2))$$

[Kingma & Welling 14]

Variational Auto Encoders

E.g.:



$$[\mu, \sigma] = e_\phi(x) \quad q_\phi(z|x) = \mathcal{N}\left(\mu, \mathrm{diag}(\sigma^2)\right)$$

$$\tilde{z} \sim \mathcal{N}\left(\mu, \mathrm{diag}(\sigma^2)\right) \quad \tilde{x} = d_\theta(\tilde{z})$$

[Kingma & Welling 14]

Variational lower bound

$$\log p(x) = \log \int_z p(x, z) dz$$

Variational lower bound

$$\log p(x) = \log \int_z p(x,z)dz$$

$$= \log \int_z p(x,z)\frac{q(z|x)}{q(z|x)}dz$$

Variational lower bound

$$\log p(x) = \log \int_z p(x,z)dz$$

$$= \log \int_z p(x,z)\frac{q(z|x)}{q(z|x)}dz = \log \mathbb{E}_{q(z|x)}\frac{p(x,z)}{q(z|x)}$$

Variational lower bound

$$\log p(x) = \log \int_z p(x, z) dz$$

$$= \log \int_z p(x, z) \frac{q(z|x)}{q(z|x)} dz = \log \mathbb{E}_{q(z|x)} \frac{p(x,z)}{q(z|x)}$$

$$= \log \mathbb{E}_{q(z|x)} \frac{p(x|z) p(z)}{q(z|x)}$$

Variational lower bound

$$\log p(x) = \log \int_z p(x,z)dz$$

$$= \log \int_z p(x,z)\frac{q(z|x)}{q(z|x)}dz = \log \mathbb{E}_{q(z|x)} \frac{p(x,z)}{q(z|x)}$$

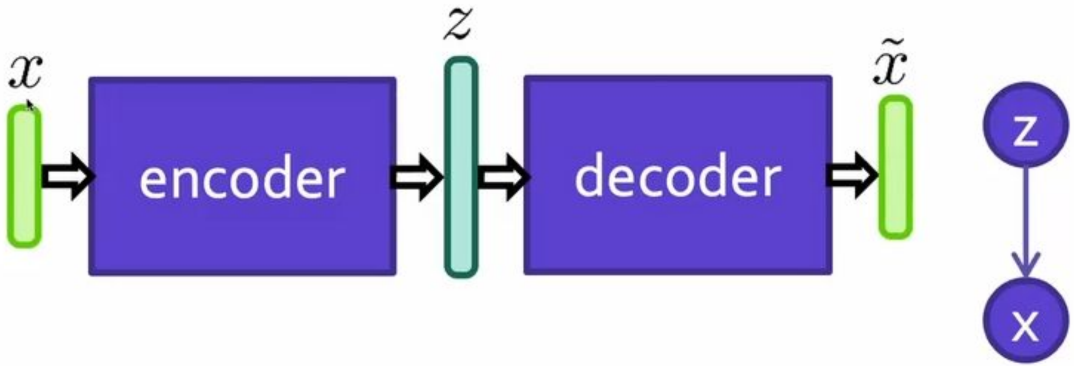
$$= \log \mathbb{E}_{q(z|x)} \frac{p(x|z)\, p(z)}{q(z|x)}$$

$$\geq \mathbb{E}_{q(z|x)} \log p(x|z) + \mathbb{E}_{q(z|x)} \log \frac{p(z)}{q(z|x)}$$

Variational lower bound

$$
\begin{aligned}
\log p(x) &= \log \int_z p(x,z)\,dz \\
&= \log \int_z p(x,z)\frac{q(z|x)}{q(z|x)}\,dz = \log \mathbb{E}_{q(z|x)} \frac{p(x,z)}{q(z|x)} \\
&= \log \mathbb{E}_{q(z|x)} \frac{p(x|z)\,p(z)}{q(z|x)} \\
&\geq \mathbb{E}_{q(z|x)} \log p(x|z) + \mathbb{E}_{q(z|x)} \log \frac{p(z)}{q(z|x)} \\
&= \mathbb{E}_{q(z|x)} \log p(x|z) - \mathrm{KL}\big(q(z|x) \,\|\, p(z)\big)
\end{aligned}
$$

Variational lower bound



$$\log p(x) \geq \underbrace{-\mathrm{KL}\big(q(z|x) \,\|\, p(z)\big)}_{\text{regularization}} + \underbrace{\mathbb{E}_{q(z|x)} \log p(x|z)}_{\text{~denoising auto-encoder}}$$

[Kingma & Welling 14]

Variational lower bound

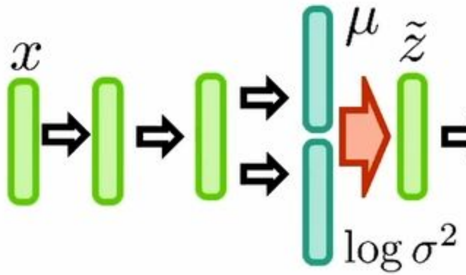
$$\log p(x) \geq \underbrace{-\mathrm{KL}\big(q(z|x) \,\|\, p(z)\big)}_{\text{regularization}} + \mathbb{E}_{q(z|x)} \log p(x|z)$$

$$p(z) = \textstyle\prod_i \mathcal{N}(z_i|0,1)$$

$$\mathrm{KL}\big(q_\phi(z|x) \,\|\, p(z)\big) = \frac{1}{2} \sum_i (\mu_i^2 + \sigma_j^2 - 1 - \log \sigma_j^2)$$

[Kingma & Welling 14]

Variational lower bound



$$\mathbb{E}_{q_\phi(z|x)} \log p_\theta(x|z) \to \max_{\theta,\phi}$$

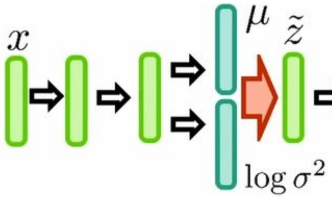
$$[\mu, \log \sigma^2] = e_\phi(x) \quad q_\phi(z|x) = \mathcal{N}\big(z|\mu, \mathrm{diag}(\sigma^2)\big)$$

$$\tilde{z} \sim \mathcal{N}\big(\mu, \mathrm{diag}(\sigma^2)\big) \quad \tilde{x} = d_\theta(\tilde{z})$$

$$p_\theta(x|z) \approx \rho_{\mathcal{N}}\big(x|\tilde{x}, \alpha \mathbf{I}\big)$$

$$\mathbb{E}_{q_\phi(z|x)} \|d_\theta(z) - x\|^2 \to \min_{\theta,\phi}$$

Variational lower bound. ELBO



$$\mathbb{E}_{q_\phi(z|x)} \log p_\theta(x|z) \to \max_{\theta,\phi}$$

$$[\mu, \log\sigma^2] = e_\phi(x) \quad q_\phi(z|x) = \mathcal{N}\big(z|\mu, \mathrm{diag}(\sigma^2)\big)$$

$$\tilde{z} \sim \mathcal{N}\big(\mu, \mathrm{diag}(\sigma^2)\big) \quad \tilde{x} = d_\theta(\tilde{z})$$

$$p_\theta(x|z) \approx \rho_{\mathcal{N}}\big(x|\tilde{x}, \alpha\mathbf{I}\big)$$

$$\mathbb{E}_{q_\phi(z|x)} \|d_\theta(z) - x\|^2 \to \min_{\theta,\phi}$$

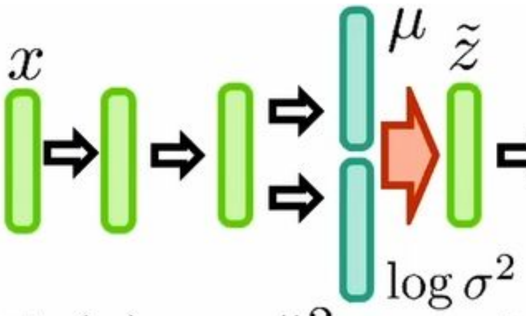
$$\sum_{i=1}^{n} E_{\boldsymbol{z}_i \sim q_\phi(\boldsymbol{z}_i|\boldsymbol{x}_i)} [\log p_\theta(\boldsymbol{x}_i \mid \boldsymbol{z}_i)] - KL(q_\phi(\boldsymbol{z}_i \mid \boldsymbol{x}_i) \,||\, p(\boldsymbol{z}_i))$$

$$\sum_{i=1}^{n} E_{\boldsymbol{z}_i \sim q_\phi(\boldsymbol{z}_i|\boldsymbol{x}_i)} \left[\log \frac{1}{\sqrt{2\pi\sigma^2_{\text{decoder}}}} - \frac{||\boldsymbol{x}_i - f_\theta(\boldsymbol{z}_i)||_2^2}{2\sigma^2_{\text{decoder}}}\right] - KL(q_\phi(\boldsymbol{z}_i \mid \boldsymbol{x}_i) \,||\, p(\boldsymbol{z}_i))$$

$$\mathrm{loss}_{AE} := \frac{1}{n}\sum_{i=1}^{n} ||\boldsymbol{x}_i - f_\theta(h_\phi(\boldsymbol{x}_i))||_2^2$$

link

Reparametrization trick



$$\mathbb{E}_{q_\phi(z|x)} \|d_\theta(z) - x\|^2 \to \min_\theta$$

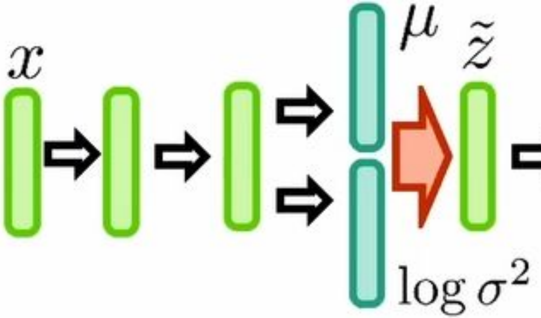
$$\tilde{z} \sim \mathcal{N}\big(\mu, \mathrm{diag}(\sigma^2)\big) \quad \tilde{x} = d_\theta(\tilde{z})$$

Sampling at every iteration

$$\epsilon \sim \mathcal{N}(0, \mathbf{I}), \quad \tilde{z} = \mu + \sigma \odot \epsilon$$

$$L(\phi, \theta) = \sum_i \Big\| d_\theta\big(\mu_\phi(x_i) + \sigma_\phi(x_i) \odot \epsilon_i\big) - x_i \Big\|^2$$

Reparametrization trick

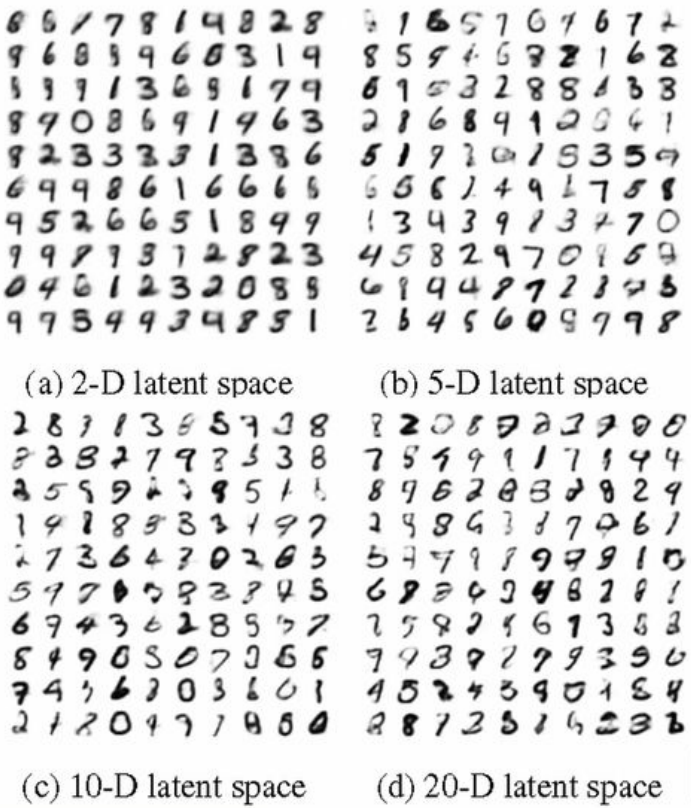


$$\epsilon \sim \mathcal{N}(0, \mathbf{I}), \qquad \tilde{z} = \mu + \sigma \odot \epsilon$$

$$L(\phi, \theta) = \sum_i \left\| d_\theta\big(\mu_\phi(x_i) + \sigma_\phi(x_i) \odot \epsilon_i\big) - x_i \right\|^2$$

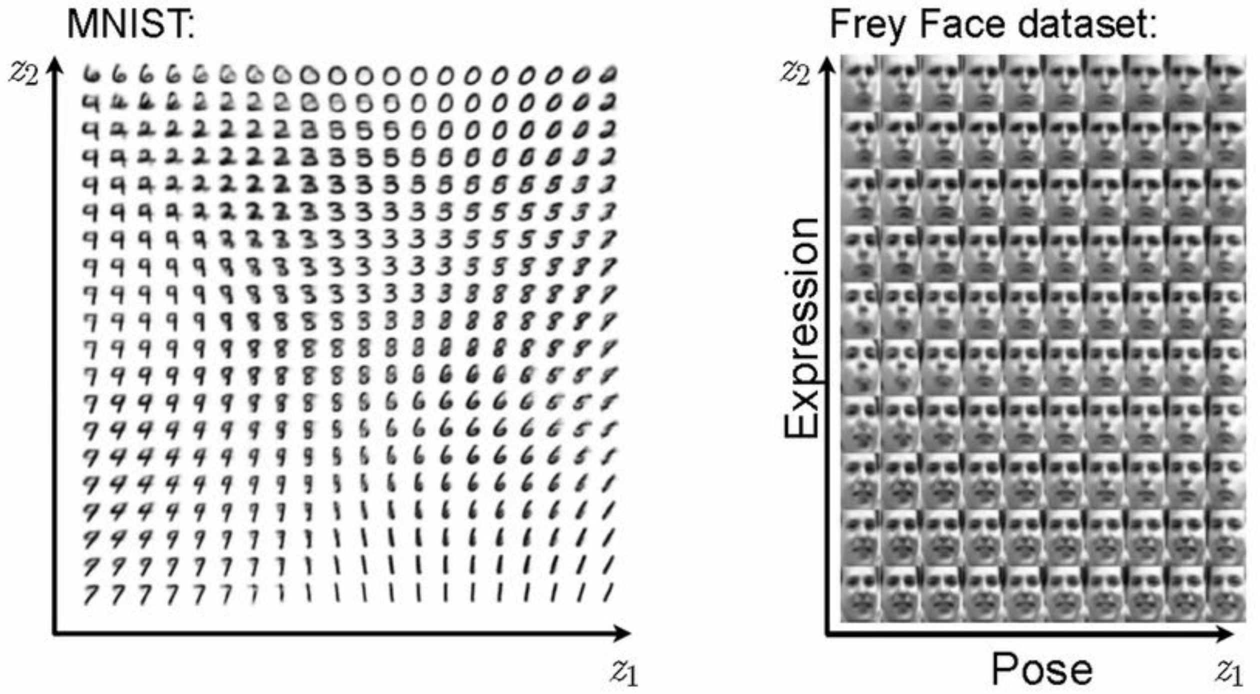
$$\frac{dL}{d\mu} = \frac{dL}{d\tilde{z}} \qquad \frac{dL}{d\sigma} = \frac{dL}{d\tilde{z}} \odot \epsilon$$

VAE learned manifolds



[Kingma & Welling 14]

VAE learned manifolds



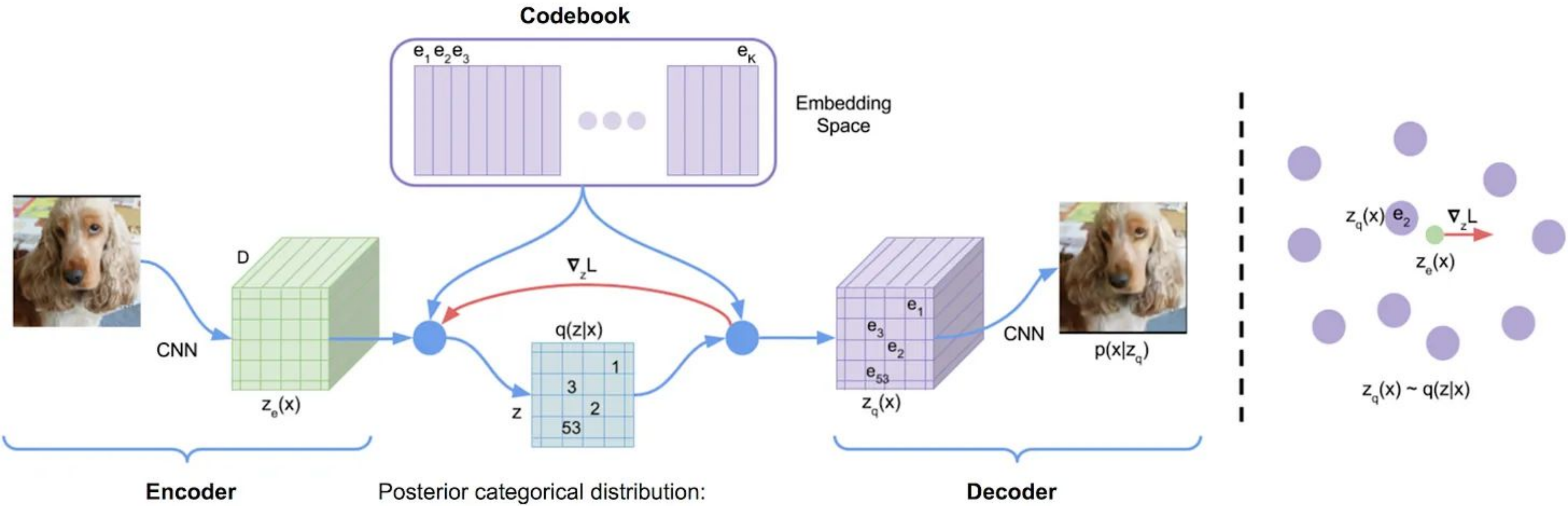
[Kingma & Welling 14]

VQ VAE. Motivation

In general, a lot of the data we encounter in the real world favors a discrete representation. For example, human speech is well represented by discrete phonemes and language. Additionally, images contain discrete objects with some discrete set of qualifiers. You could imagine having one discrete variable for the type of object, one for its color, one for its size, one for its orientation, one for its shape, one for its texture, one for the background color, one for the background texture, etc…

link

Vector quantized VAE

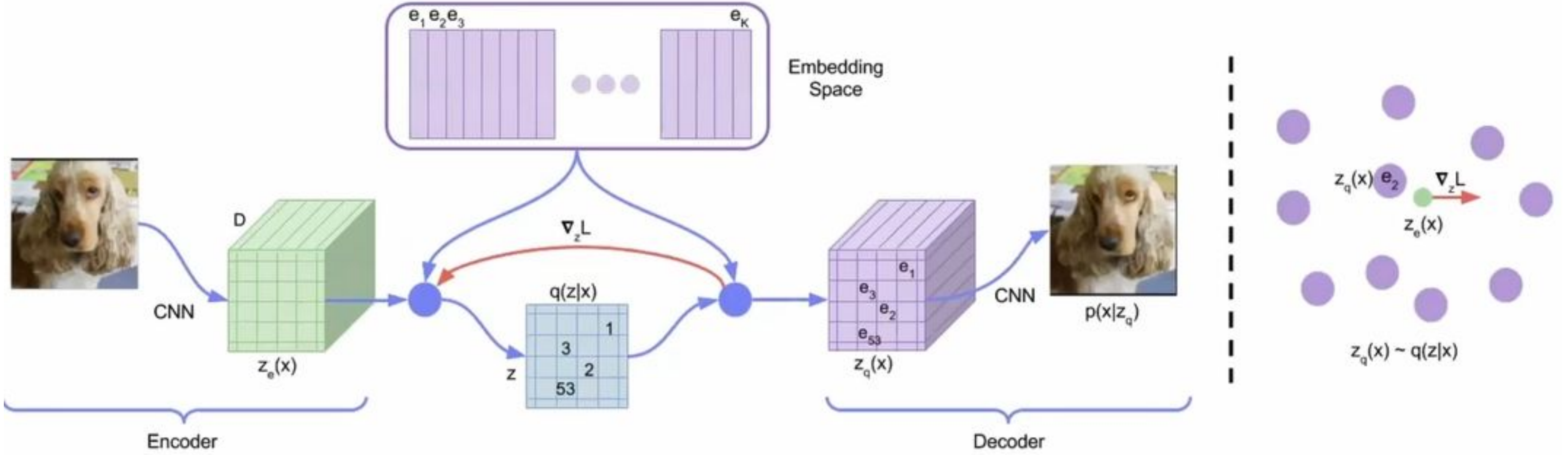


Posterior categorical distribution:

$$q(\mathbf{z} = \mathbf{e}_k|\mathbf{x}) = \begin{cases} 1 & \text{if } k = \arg\min_i \|\mathbf{z}_e(\mathbf{x}) - \mathbf{e}_i\|_2 \\ 0 & \text{otherwise.} \end{cases}$$

Vector quantized VAE

Latent space is a learned lattice:

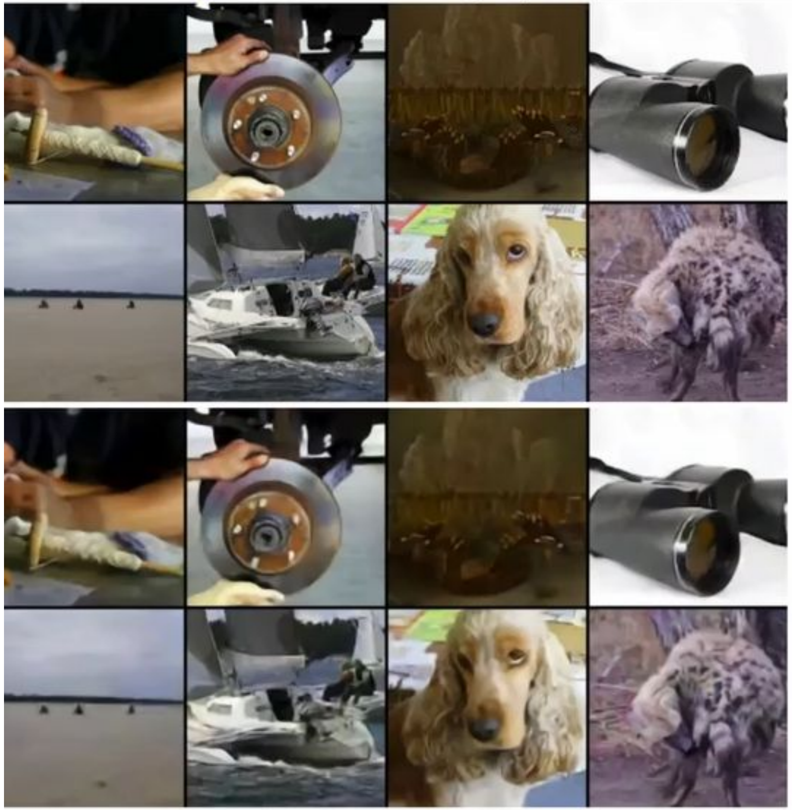


$$L = \log p(x|z_q(x)) + \|\text{sg}[z_e(x)] - e\|_2^2 + \beta\|z_e(x) - \text{sg}[e]\|_2^2$$

- "Straight-through" gradient estimation used to backprop through quantization (gradient over z_q is copied to the gradient over z_e)

[van den Oord et al. NeurIPS17]

Vector quantized VAE



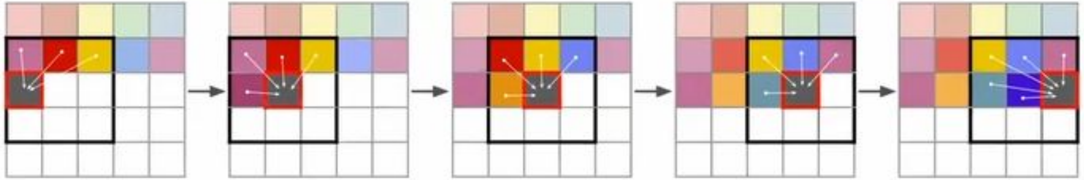
128x128 24bit RGB images

32x32 9 bit codewords (K=512)

[van den Oord et al. NeurIPS17]

Vector quantized VAE: samples

Samples from autoregressive model in the latent space (ImageNet):



"PixelCNN" prior

Image from [Esser et al.CVPR21]

[van den Oord et al. NIPS16, NeurIPS17]

Backprop through sampling from discrete distribution

- Need a layer that samples from a discrete categorical distribution $(\pi_1, \pi_2, \ldots \pi_N)$

$$\rho(x) = \exp(-x + \exp(-x))$$

- Implementation of the forward pass (reparameterization trick again!):

Gumbel (0,1):

$$Z = \mathtt{onehot}(\mathtt{argmax}\{G_i + \log \pi_i\})$$

[Jang et al. ICLR17, Maddison et al. ICLR17]

Backprop through sampling from discrete distribution

- Need a layer that samples from a discrete categorical distribution $(\pi_1, \pi_2, \dots \pi_N)$

- Implementation of the forward pass (reparameterization trick again!):

$\rho(x) = \exp(-x + \exp(-x))$

Gumbel (0,1):

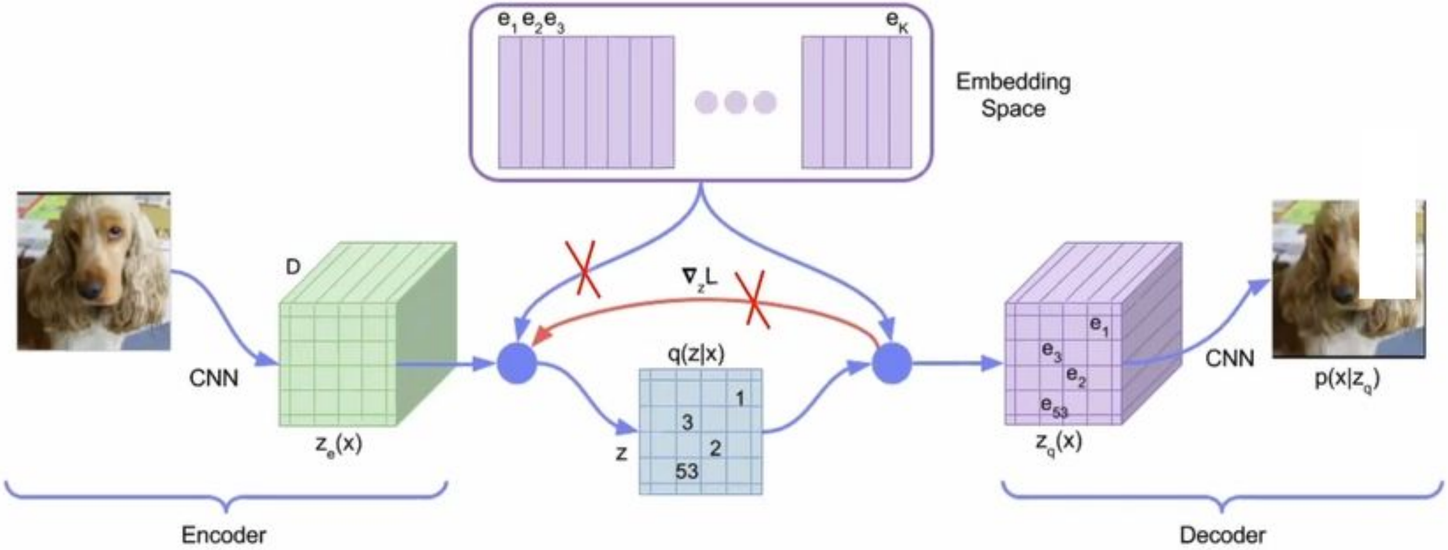
$$Z = \texttt{onehot}(\texttt{argmax}\{G_i + \log \pi_i\})$$

- Differentiable approximation via softmax:

$$Z = \texttt{softmax}\{G_i + \log \pi_i\}$$

[Jang et al. ICLR17, Maddison et al. ICLR17]

From VQ-VAE to dVAE (DALL-E, part I)



- Instead, predict a K=8192-bin cathegorical distribution at each location
- Each distribution is transformed with Gumbel-softmax and used to weight codebook vectors

$$\{y_i\} = \texttt{GumbelSoftmax}\{q(e_i|x)\} \qquad z = \sum_{j=1}^{K} y_j e_j$$

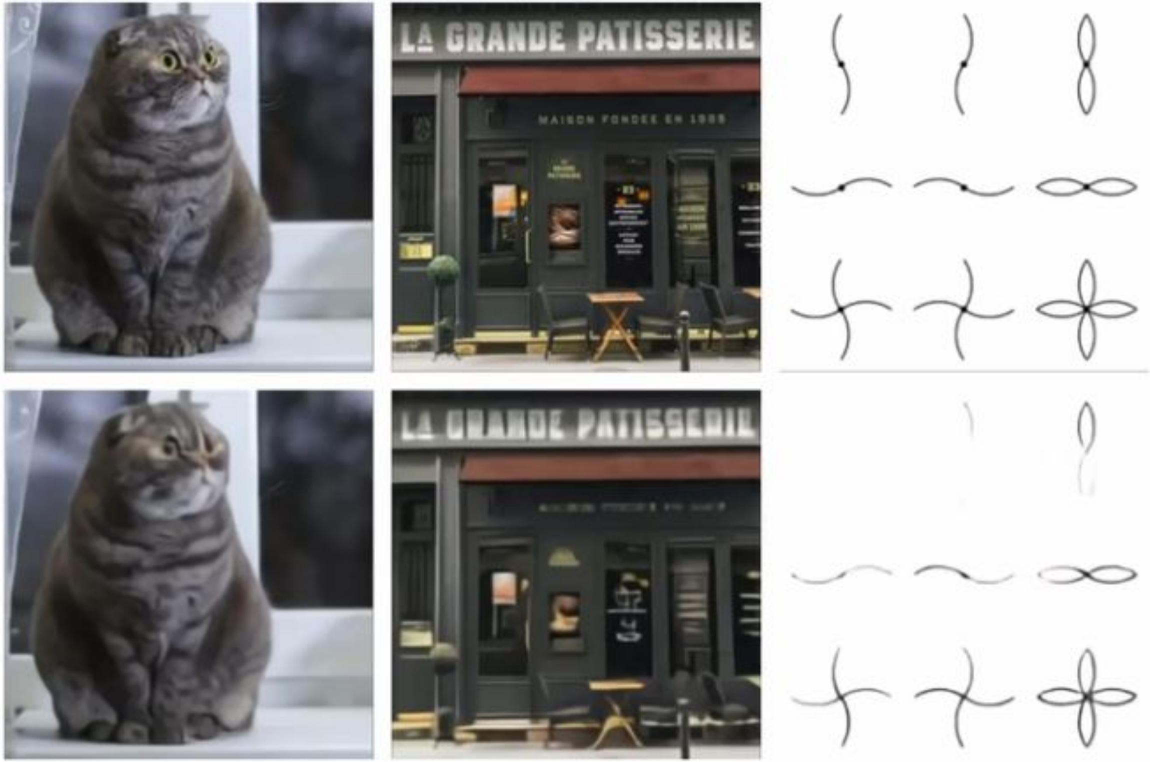
[Ramesh et al. ICML21]

dVAE decoding

Training dVAE gives us:

1) a mapping from an image to 1024 tokens from a 8192-lexicon (encoder)

2) a mapping back (decoder)



[Ramesh et al. ICML21]

DALL-E: part 2

- A 250M dataset of text-image pairs is used for training
- (caption, image) is mapped to 256+1024 token sequence. BPE encoding is used for text
- A large-scale sequence model (SparseTransformer) with 12 B params is trained to model the token symbols
- Inside the attention layers, each image token can attend to all caption tokens as well as to nearby image tokens that are to the left or above

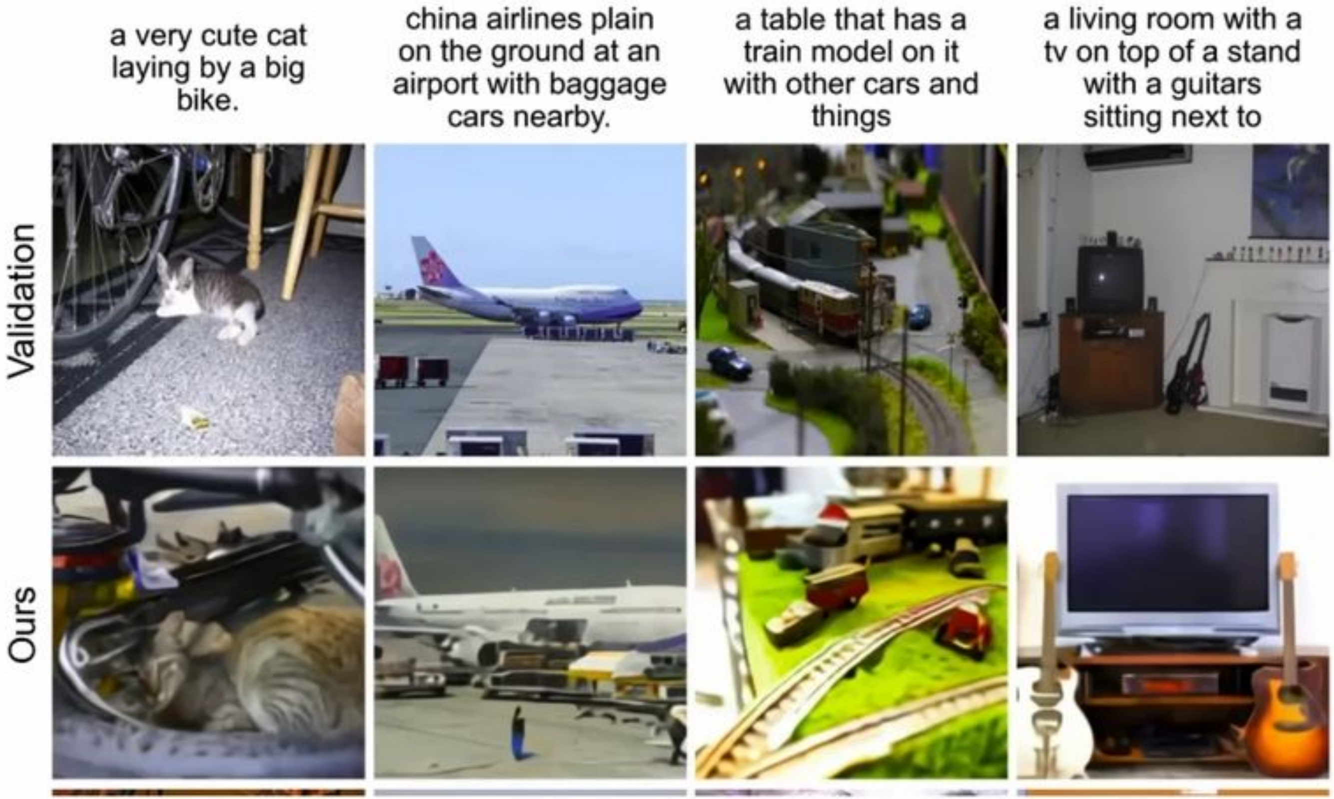
[Ramesh et al. ICML21]

DALL-E: part 2

- A 250M dataset of text-image pairs is used for training
- (caption, image) is mapped to 256+1024 token sequence. BPE encoding is used for text
- A large-scale sequence model (SparseTransformer) with 12 B params is trained to model the token symbols
- Inside the attention layers, each image token can attend to all caption tokens as well as to nearby image tokens that are to the left or above
- After training, image tokens can be sampled sequentially conditioned on text and previously sampled tokens
- Sampled images can be reranked using CLIP

[Ramesh et al. ICML21]

DALL-E results



[Ramesh et al. ICML21]

DALL-E results

TEXT PROMPT an armchair in the shape of an avocado. . . .

AI-GENERATED IMAGES

TEXT PROMPT an illustration of a baby daikon radish in a tutu walking a dog

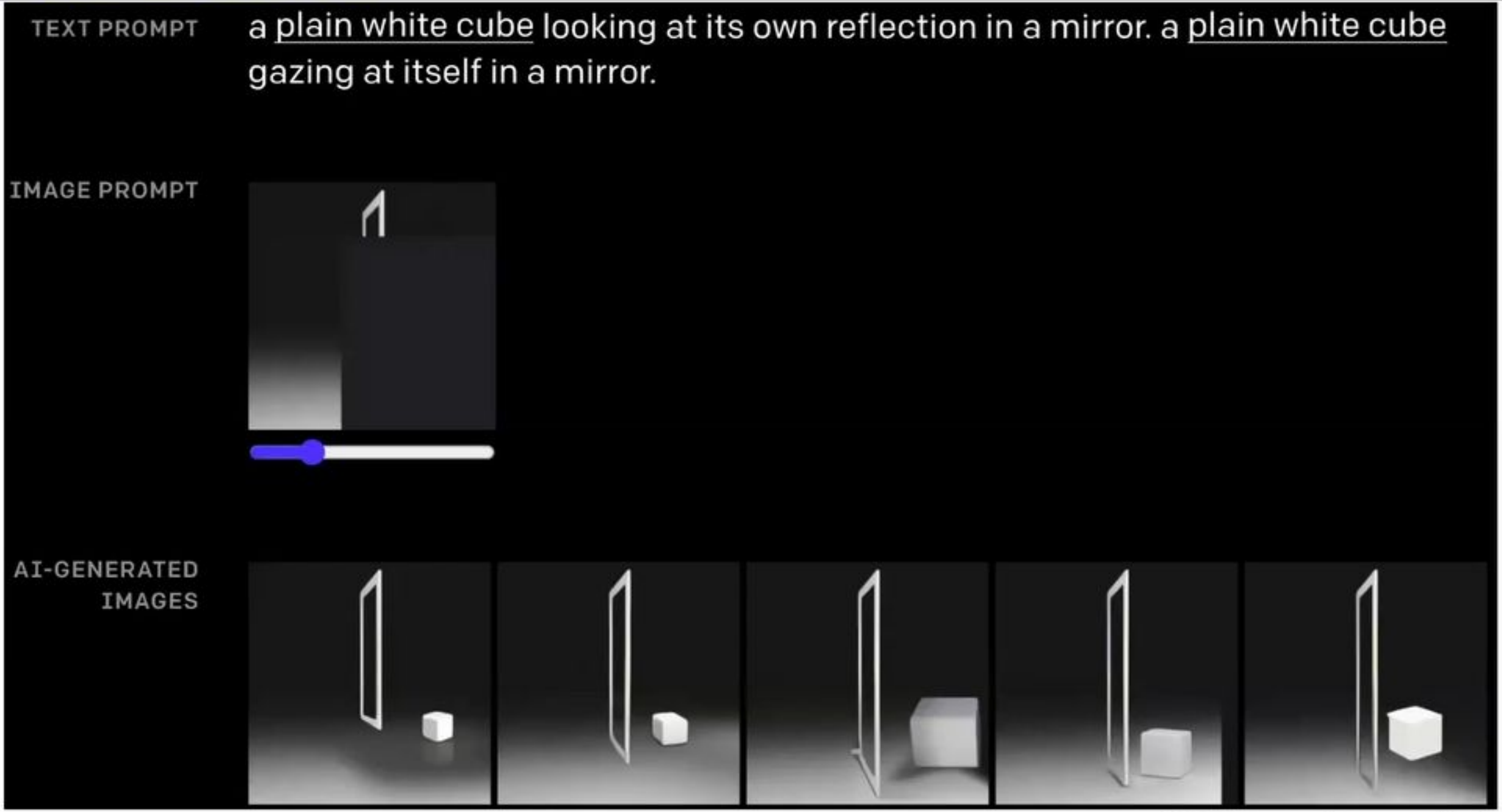
AI-GENERATED IMAGES

[Ramesh et al. ICML21]

DALL-E results

[Ramesh et al. ICML21]

DALL-E results



[Ramesh et al. ICML21]