Глубинное обучение

Из NLP в CV

Ирина Сапарина

Vision Transformer

Vision Transformer

Идея: применить архитектуру Transformer для CV (image classification)

- нужна последовательность на вход

Как из картинки сделать последовательность?

Image credit

Vision Transformer

Как из картинки сделать последовательность?

- разделим на 2D патчи

Image credit

Vision Transformer

Как из картинки сделать последовательность?

- разделим на 2D патчи
- вытянем в последовательность 2D картинок

Image credit

Vision Transformer

Как из картинки сделать последовательность?

- разделим на 2D патчи
- вытянем в последовательность 2D картинок
- Linear mapping - в меньшую размерность

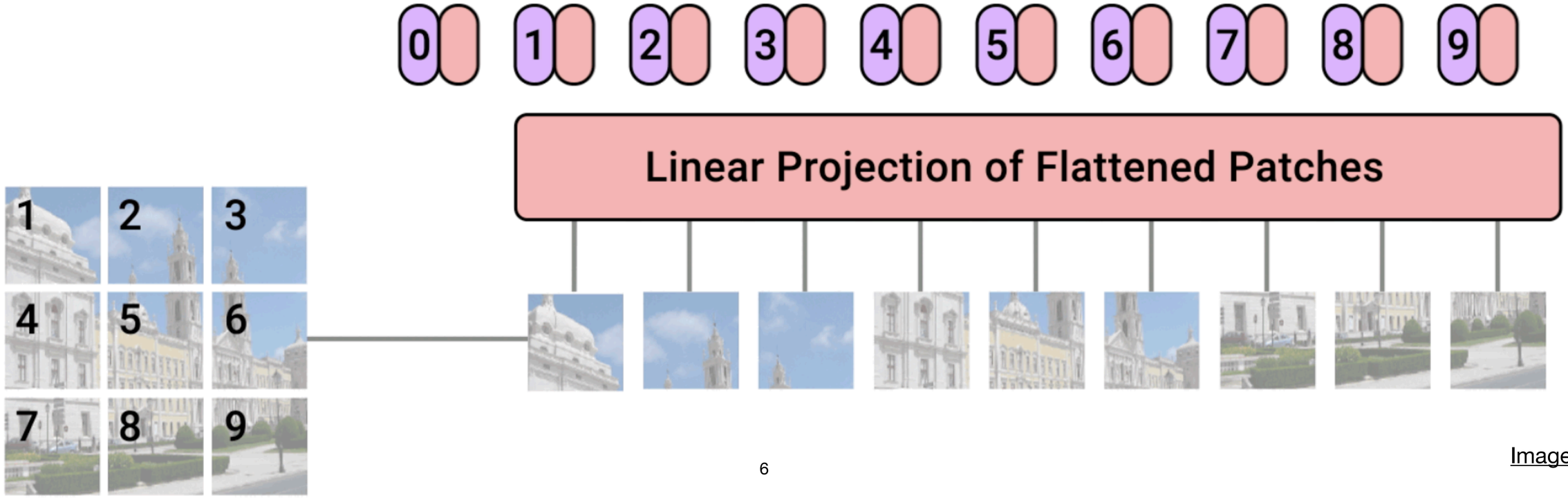


Image credit

Vision Transformer

Как из картинки сделать последовательность?

- разделим на 2D патчи
- вытянем в последовательность 2D картинок
- Linear mapping - в меньшую размерность

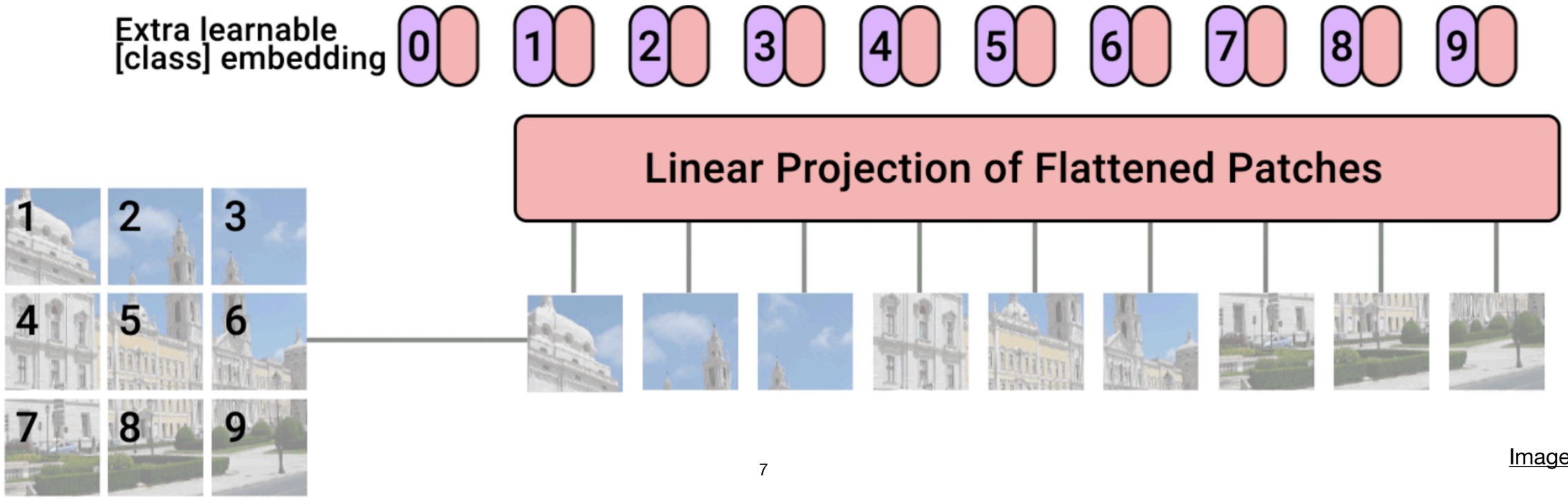


Image credit

Vision Transformer

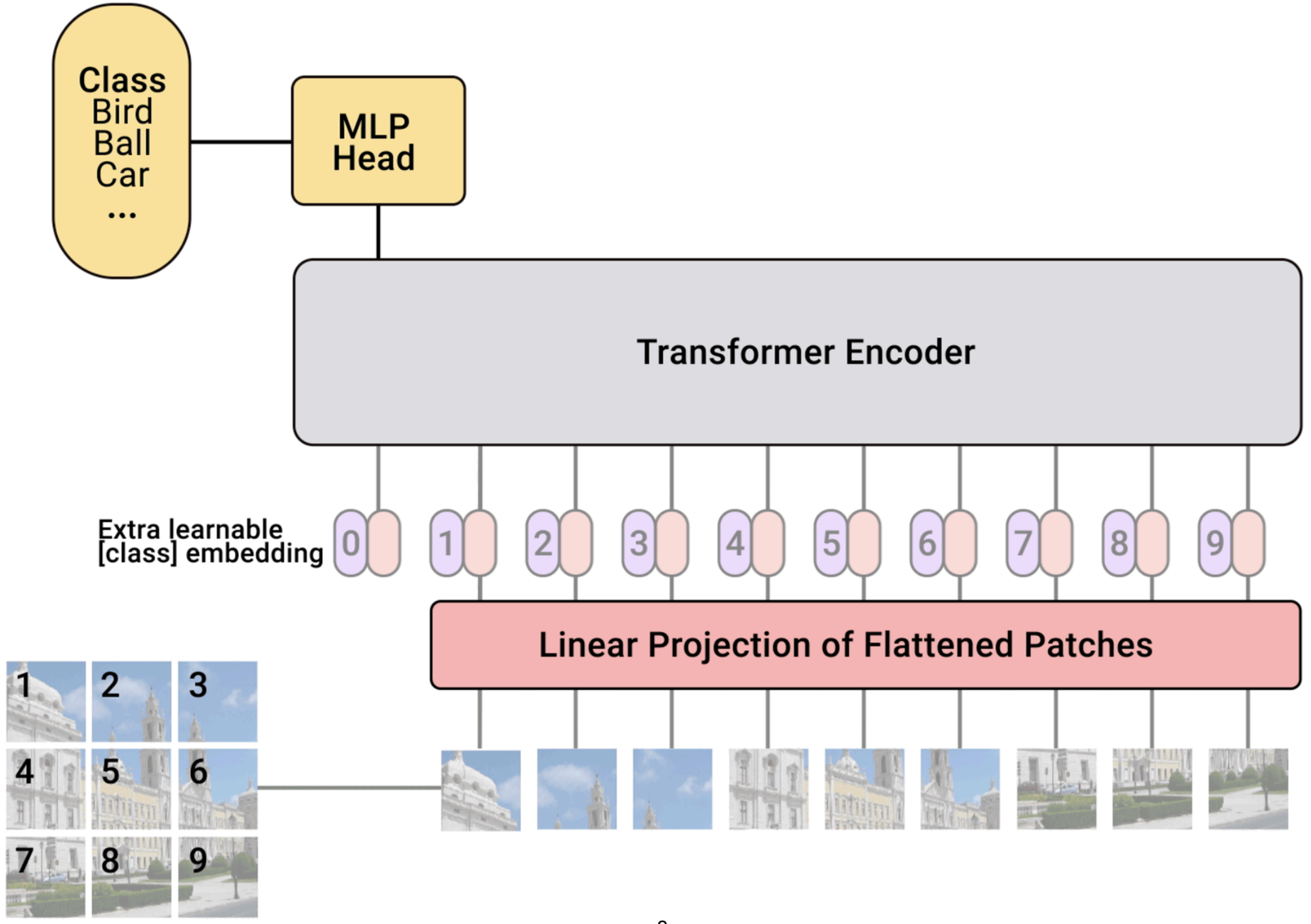


Image credit

Vision Transformer

	ViT-H	Previous SOTA
ImageNet	88.55	88.5
ImageNet-ReaL	90.72	90.55
Cifar-10	99.50	99.37
Cifar-100	94.55	93.51
Pets	97.56	96.62
Flowers	99.68	99.63

Image credit

Swin Transformer

Для других задач CV (сегментация, детекторы) фиксированный патчи работают плохо - это исправляет Swin Transformer

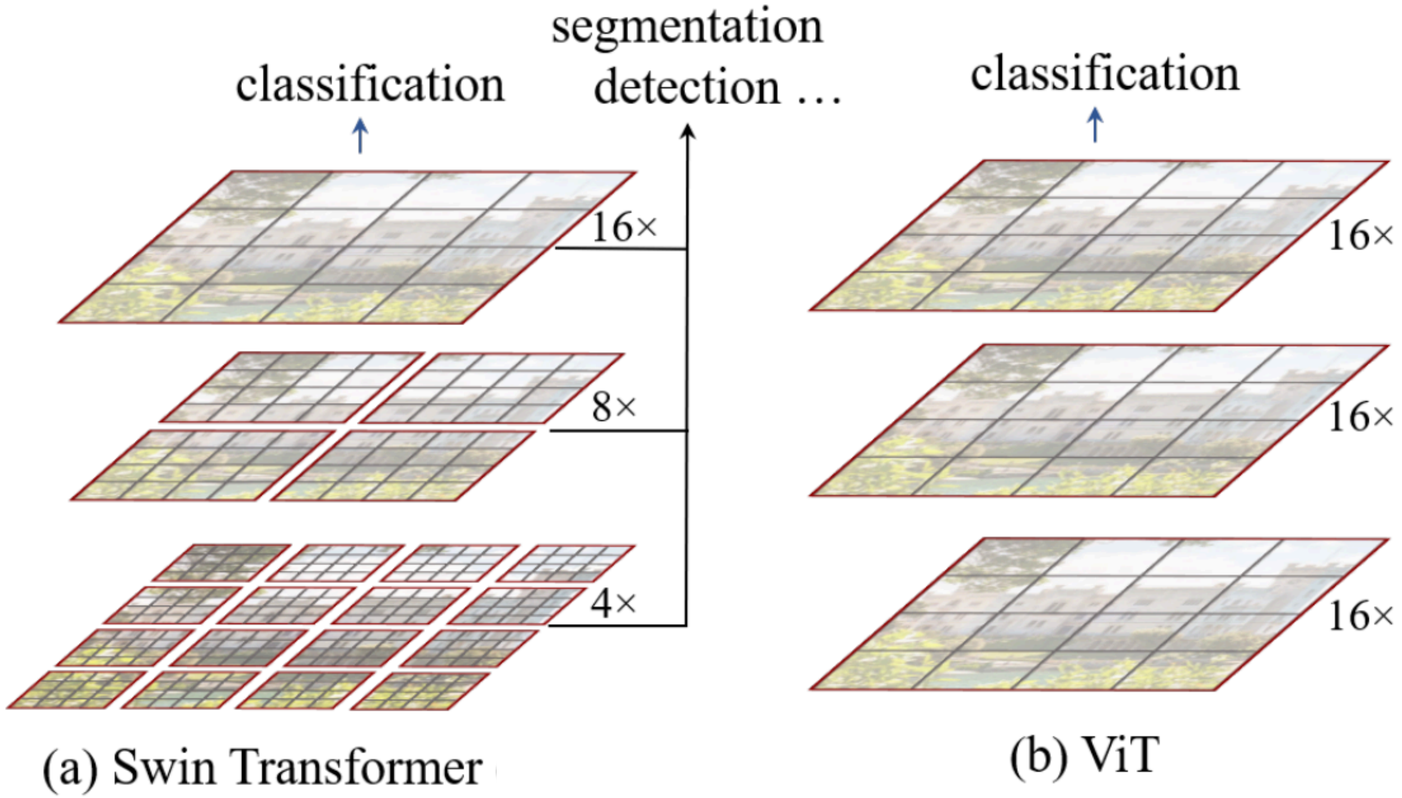


Image credit

Swin Transformer

Shifted Window: self-attention между блоками патчей, сдвигаем окна на следующем слое

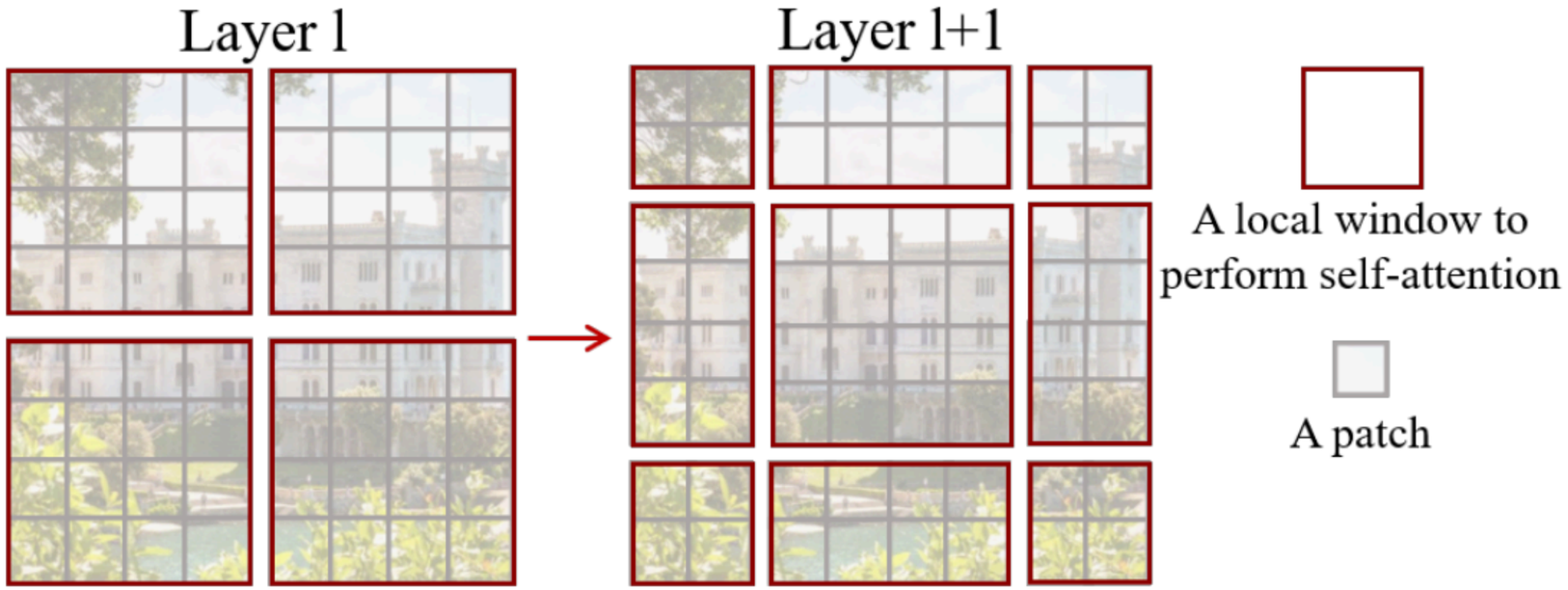


Image credit

CLIP: Contrastive Language–Image Pre-training

CLIP (Contrastive Language–Image Pre-training)

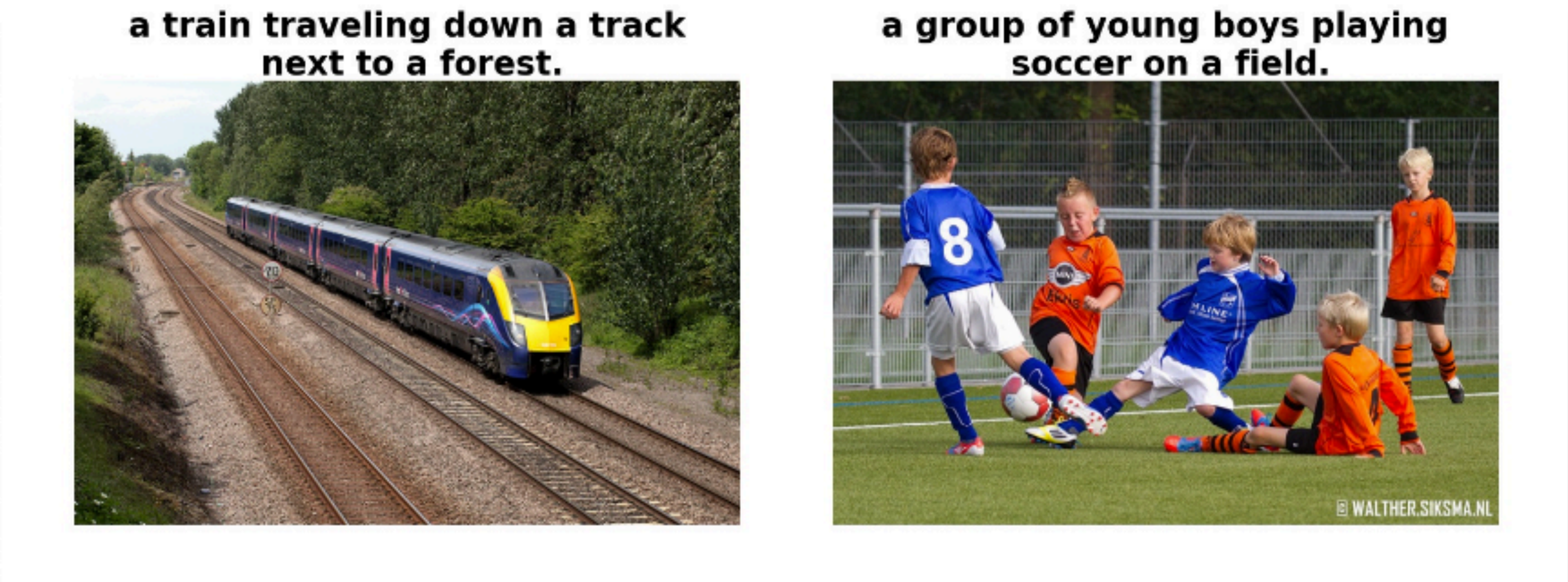
BERT-like модели предобучаются **на неразмеченных данных** (тексты)

Image classifiers учат **на размеченных людьми данных** (ImageNet, etc.)

Идея: предобучить image classifier на больших данных без использования аннотации (краудсорсинга)

CLIP (Contrastive Language–Image Pre-training)

- Собрали 400 миллионов пар (картинка, подпись) - из интернета



CLIP (Contrastive Language–Image Pre-training)

- Предобучение: энкодим N картинок (ViT) и N подписей к ним (Transformer)

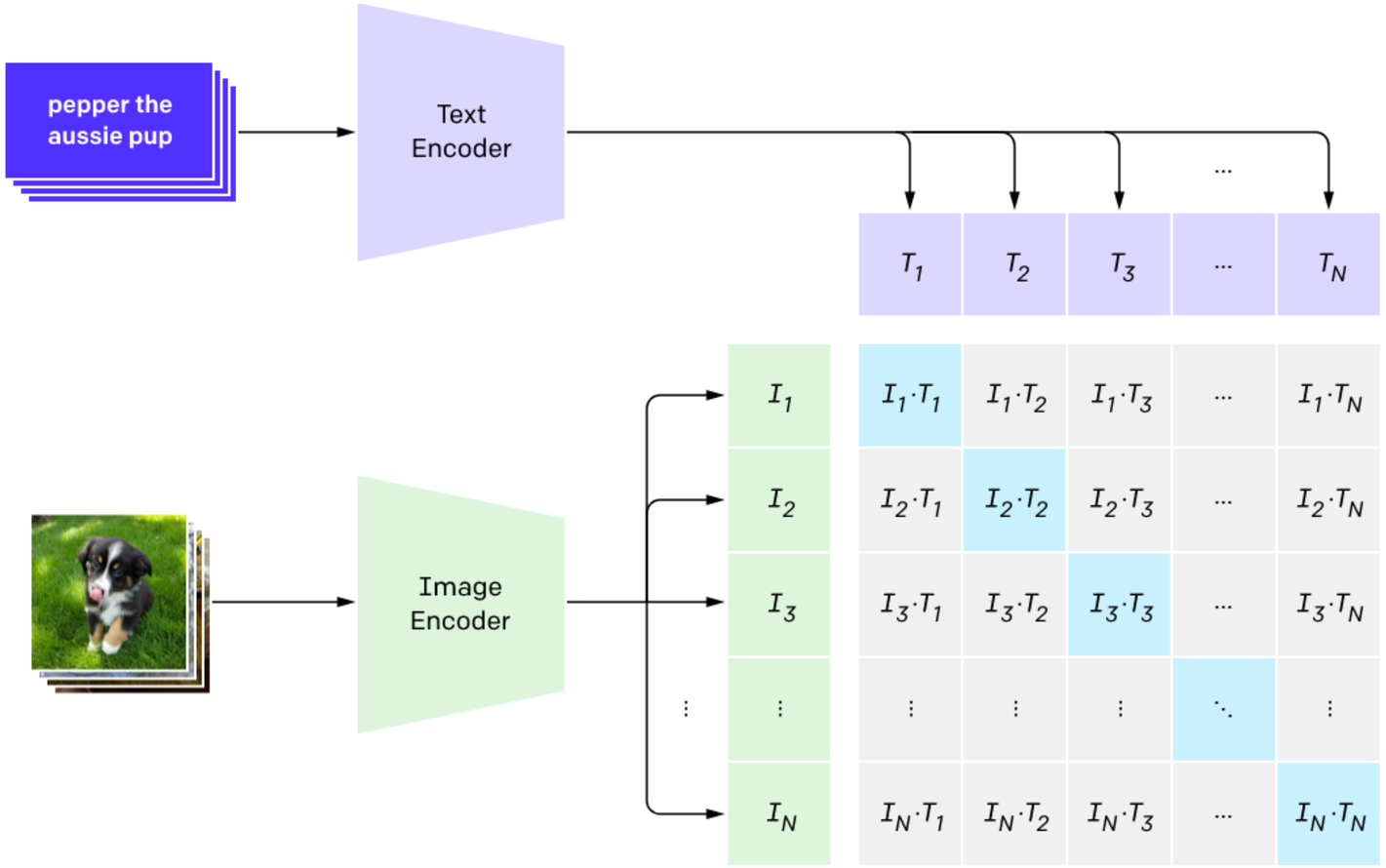
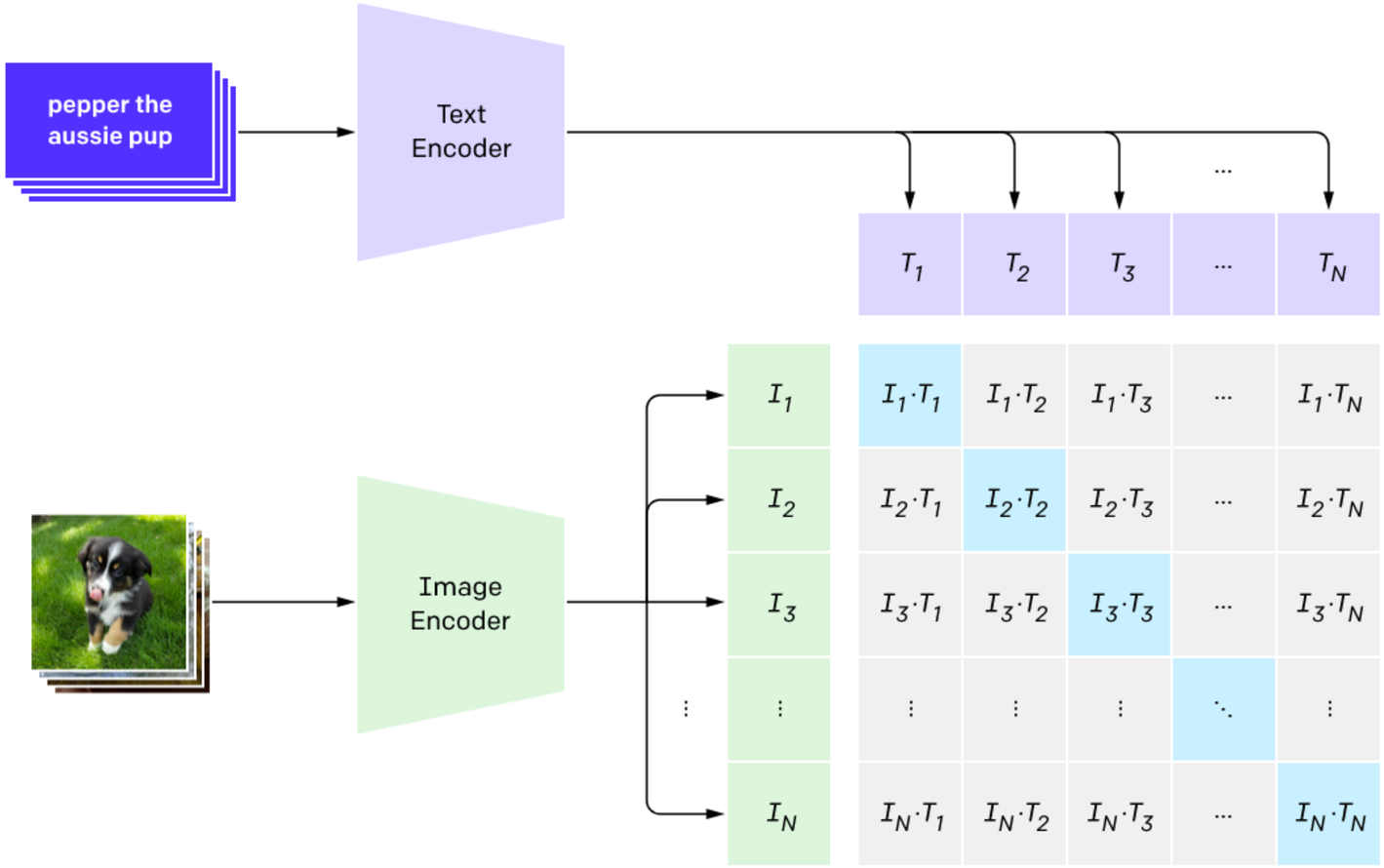


Image credit

CLIP (Contrastive Language–Image Pre-training)

- Предобучение: энкодим N картинок (ViT) и N подписей к ним (Transformer)



Лосс:

```
# symmetric loss function
labels = np.arange(n)
loss_i = cross_entropy_loss(logits, labels, axis=0)
loss_t = cross_entropy_loss(logits, labels, axis=1)
loss   = (loss_i + loss_t)/2
```

Image credit

CLIP (Contrastive Language–Image Pre-training)

- Применение для Image Classification: меняем метку на текст “the photo of …”

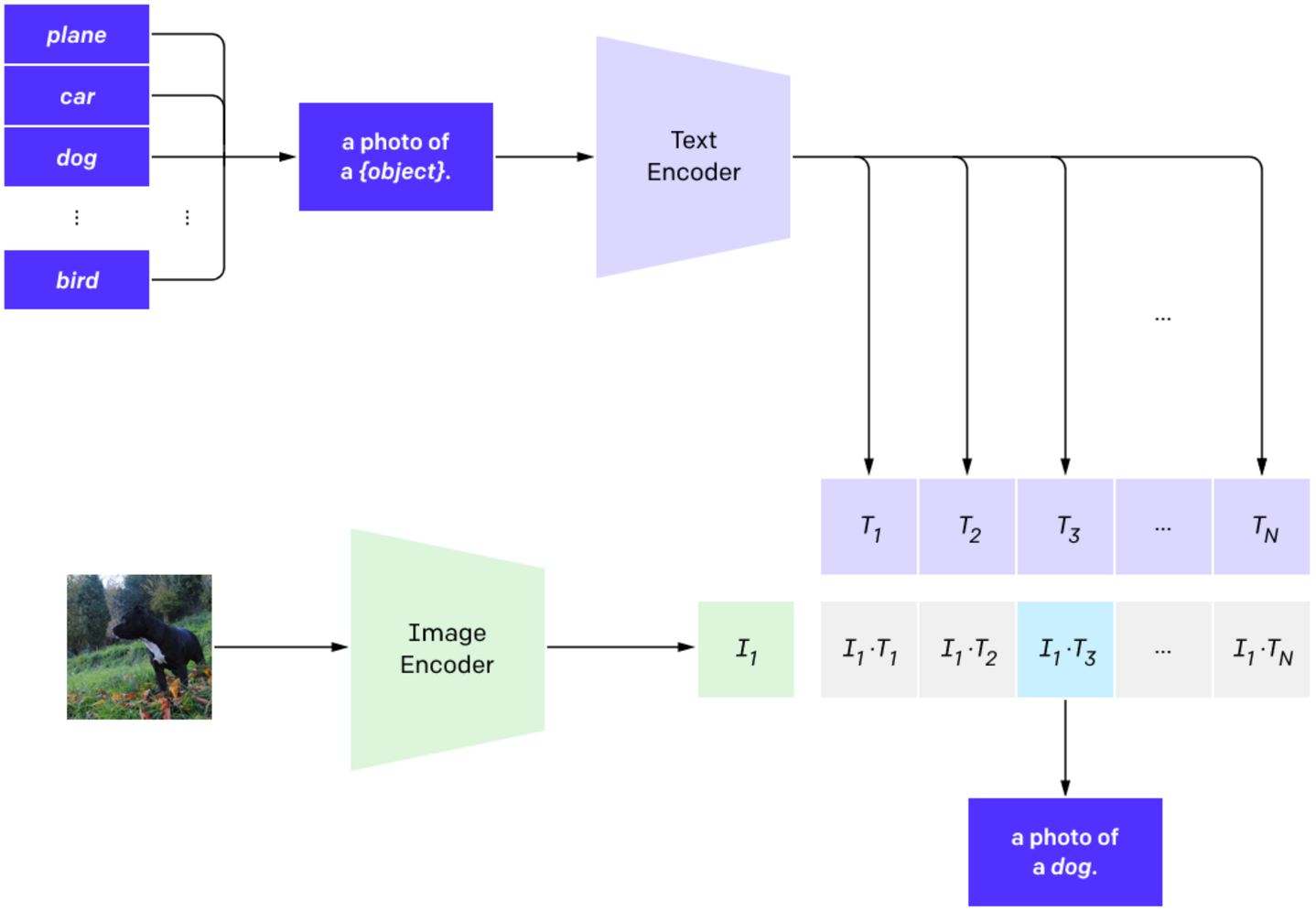


Image credit

CLIP (Contrastive Language–Image Pre-training)

FOOD101

guacamole (90.1%) Ranked 1 out of 101 labels

✓ a photo of **guacamole**, a type of food.

× a photo of **ceviche**, a type of food.

× a photo of **edamame**, a type of food.

× a photo of **tuna tartare**, a type of food.

× a photo of **hummus**, a type of food.

SUN397

television studio (90.2%) Ranked 1 out of 397

✓ a photo of a **television studio**.

× a photo of a **podium indoor**.

× a photo of a **conference room**.

× a photo of a **lecture room**.

× a photo of a **control room**.

YOUTUBE-BB

airplane, person (89.0%) Ranked 1 out of 23

✓ a photo of a **airplane**.

× a photo of a **bird**.

× a photo of a **bear**.

× a photo of a **giraffe**.

× a photo of a **car**.

EUROSAT

annual crop land (12.9%) Ranked 4 out of 10

× a centered satellite photo of **permanent crop land**.

× a centered satellite photo of **pasture land**.

× a centered satellite photo of **highway or road**.

✓ a centered satellite photo of **annual crop land**.

× a centered satellite photo of **brushland or shrubland**.

Image credit

CLIP (Contrastive Language–Image Pre-training)

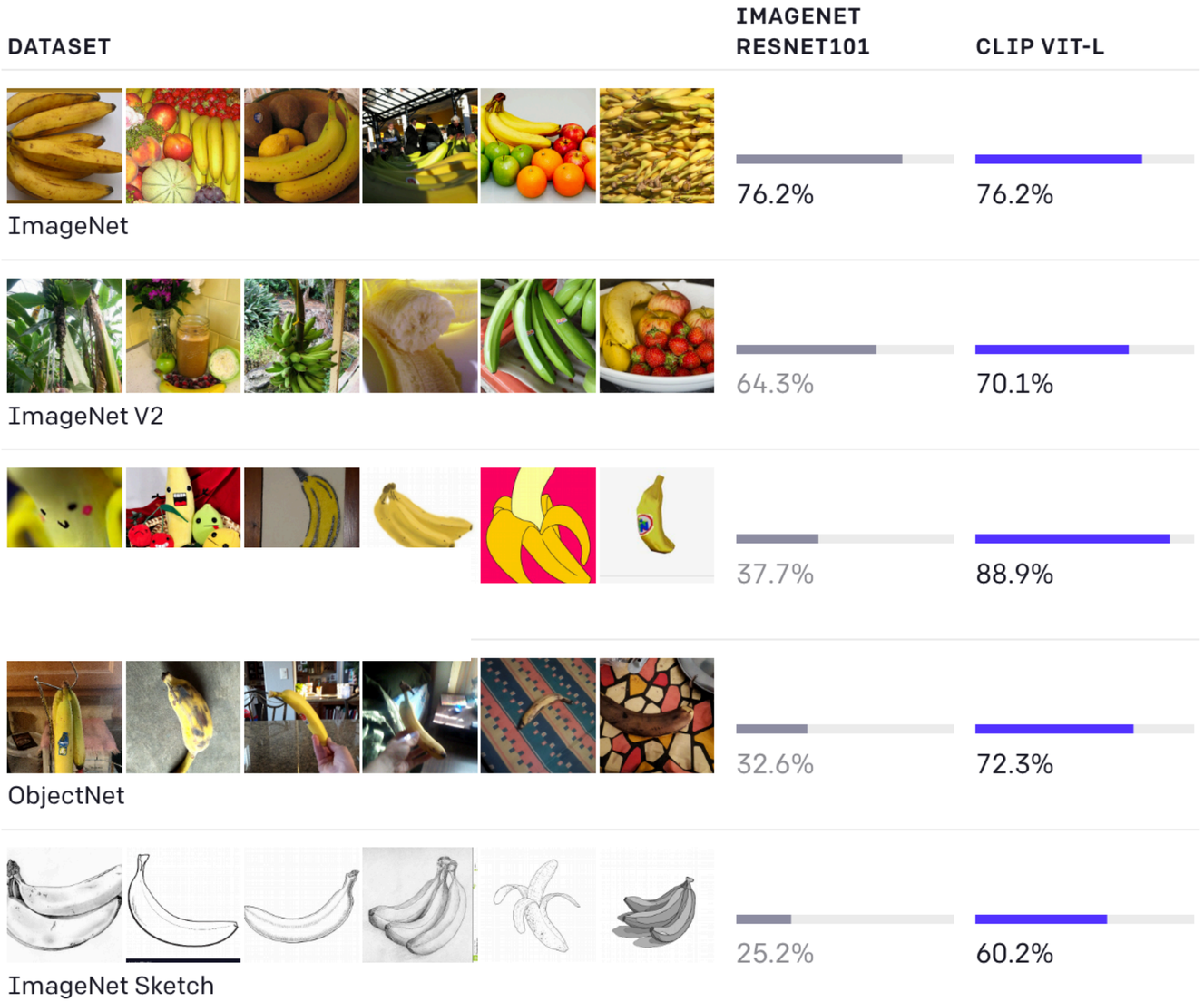


Image credit

CLIP (Contrastive Language–Image Pre-training)

- Чувствителен к описанию

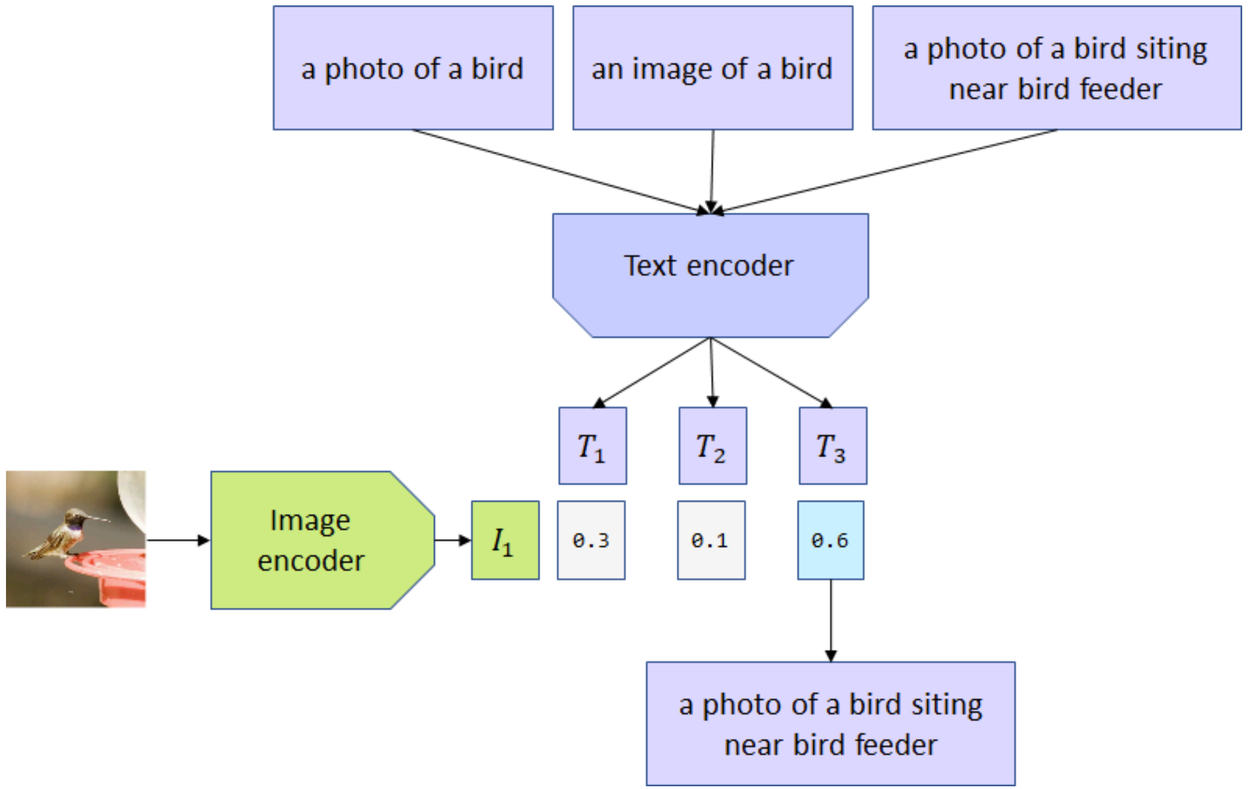


Image credit

CLIP + StyleGAN

- Используют преодобученные StyleGAN и CLIP
- Цель - выучить трансформации над исходной картинкой, чтобы результат генерации StyleGAN был близок к текстовому описанию (эмбеддингам CLIP)

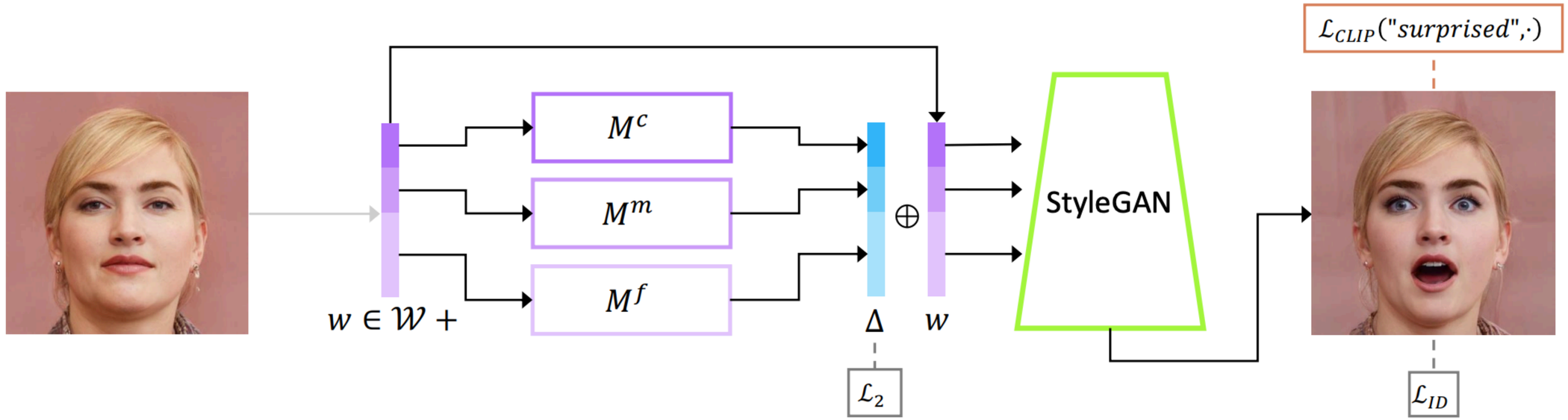
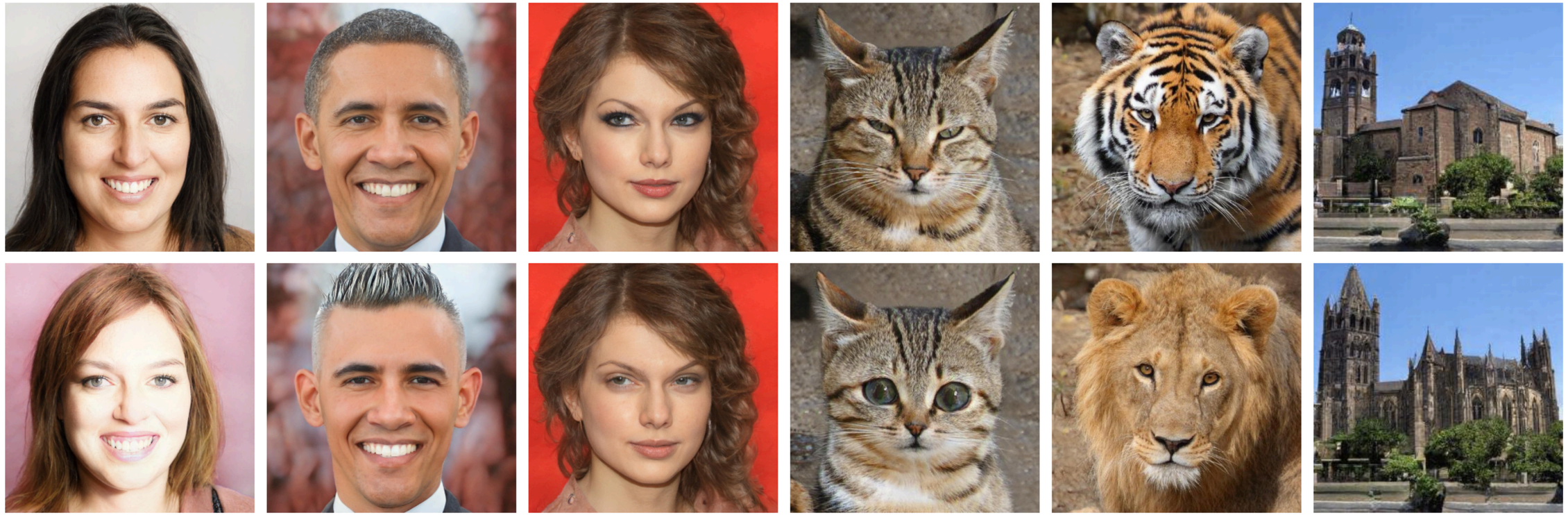


Image credit

CLIP + StyleGAN



“Emma Stone” “Mohawk hairstyle” “Without makeup” “Cute cat” “Lion” “Gothic church”

Image credit

DALL-E

DALL-E

Задача: по текстовому описанию сгенерировать картинку

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

AI-GENERATED IMAGES

Image credit

DALL-E

Задача: по текстовому описанию сгенерировать картинку

Авторегрессионная модель:

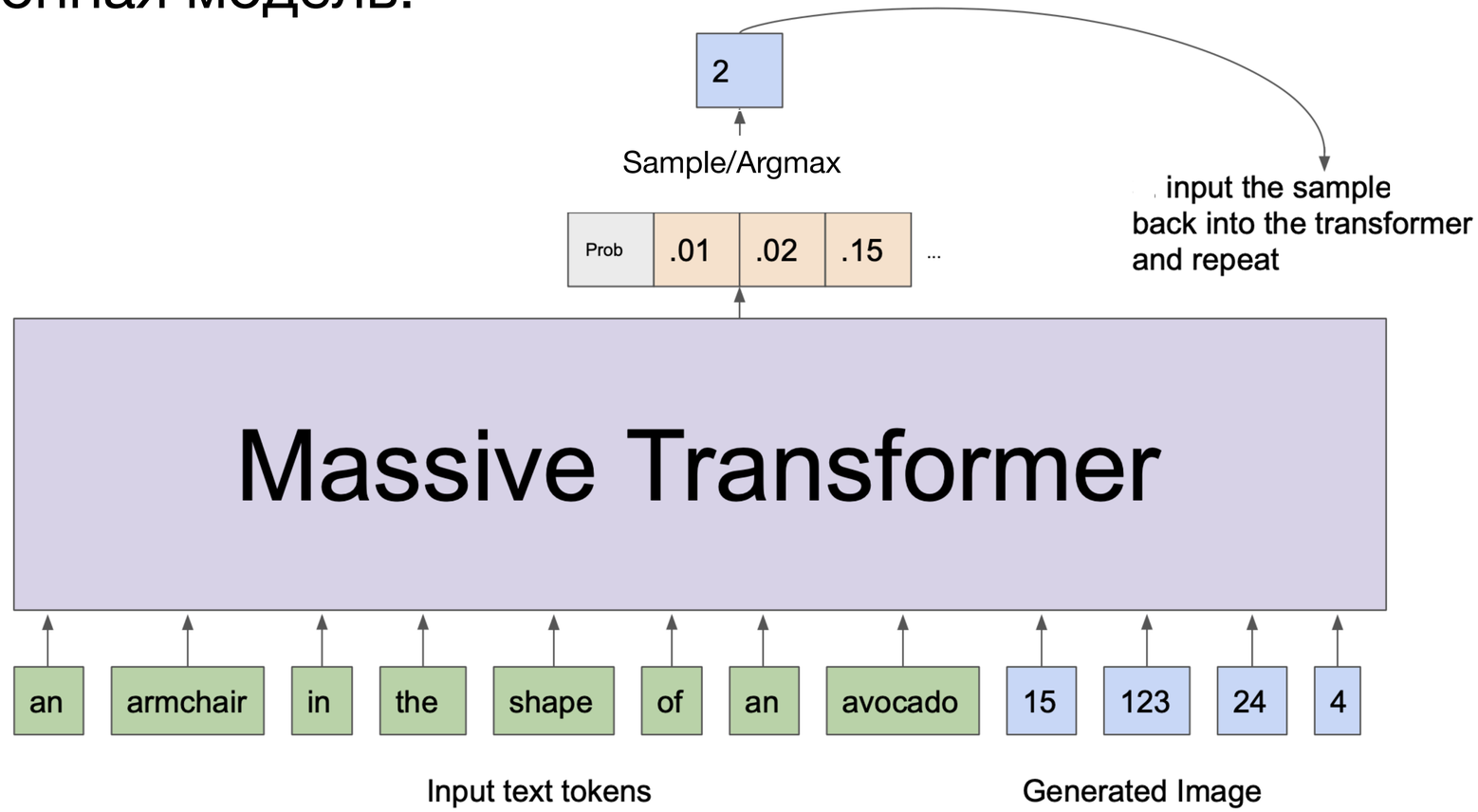
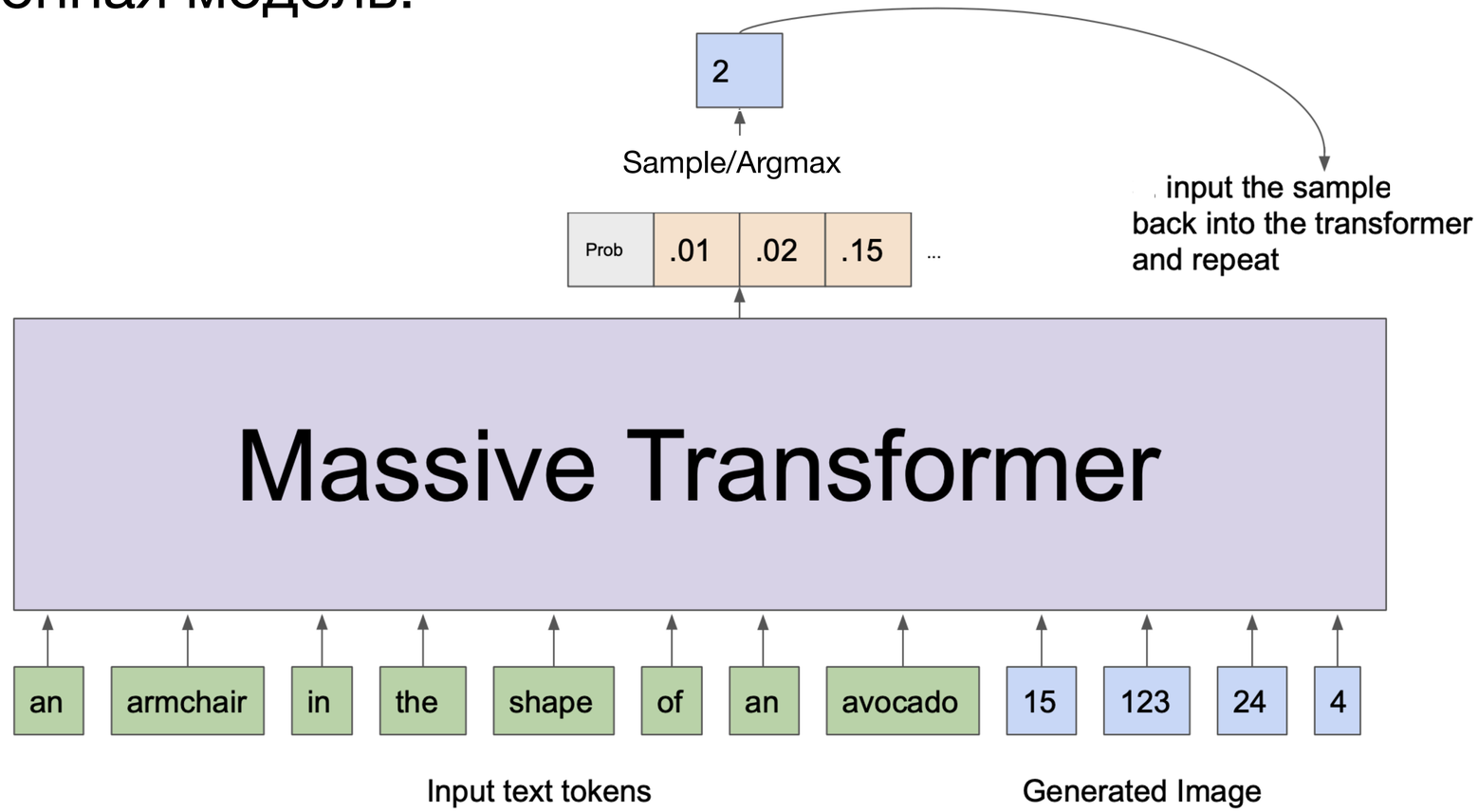


Image credit

DALL-E

Задача: по текстовому описанию сгенерировать картинку

Авторегрессионная модель:



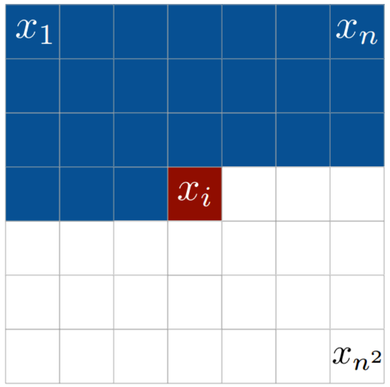


Image credit

DALL-E

Генерировать по пикселям долго - переведем картинку в пространство меньшей размерности

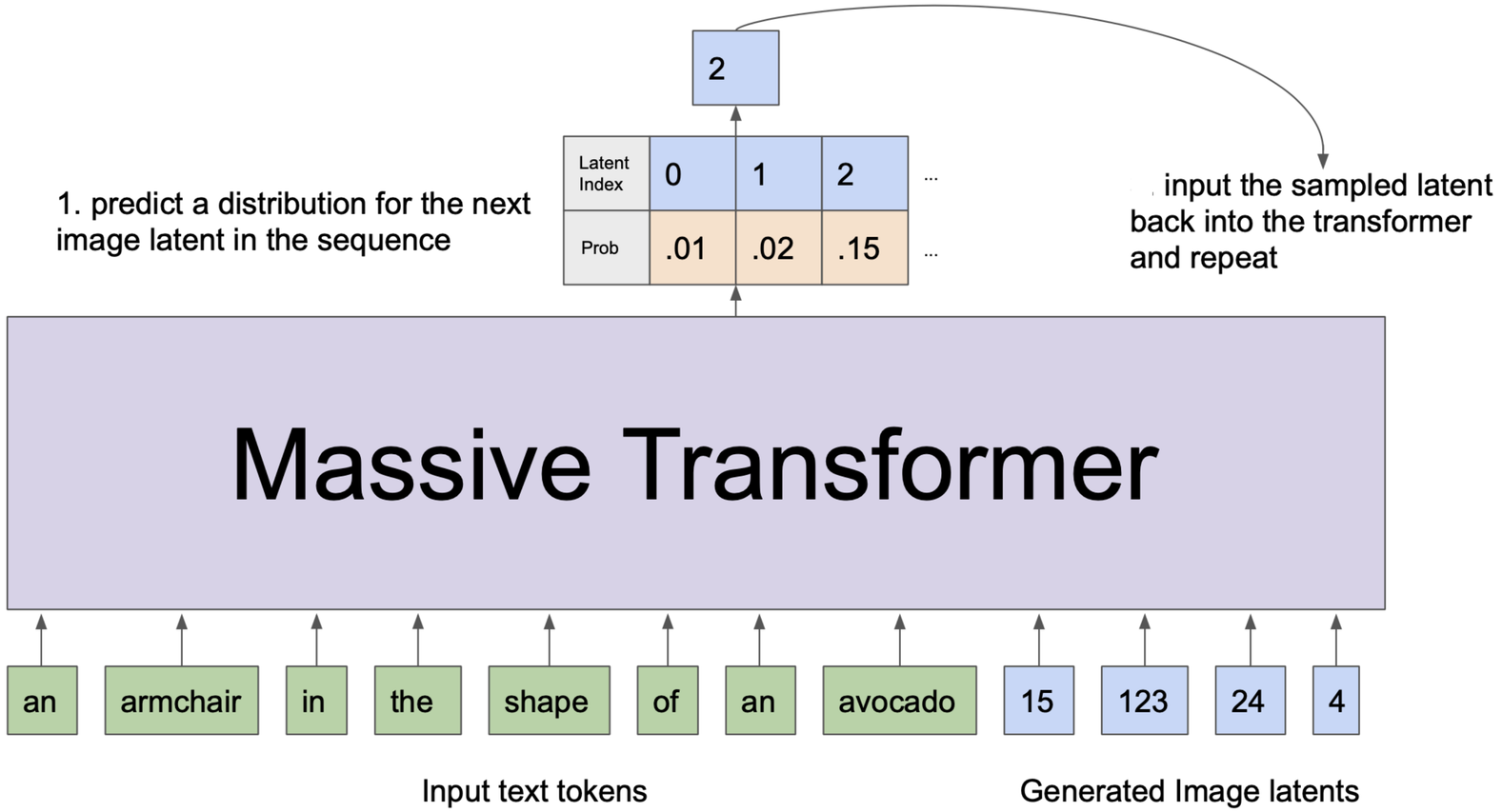


Image credit

VQ-VAE

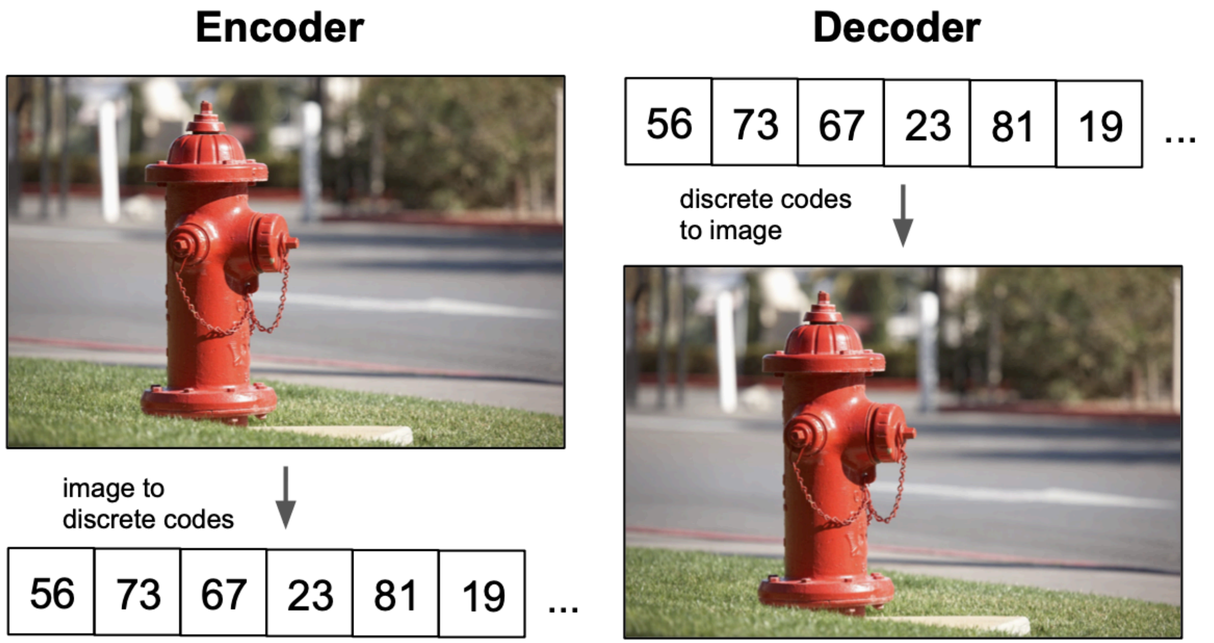


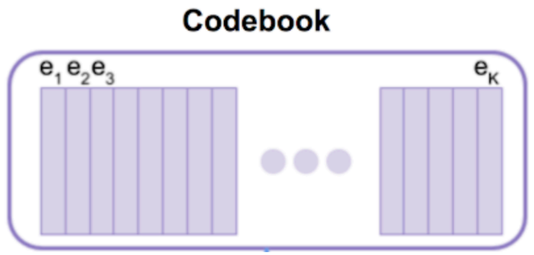
Image credit

VQ-VAE

Reconstructed Input: $\hat{x}$

↑ Decoder

z_q



Embedding Space

↑ **Vector Quantization Layer**

$$z_q(x) = argmin_i \|e_i - z_e(x)\|_2^2$$

z_e

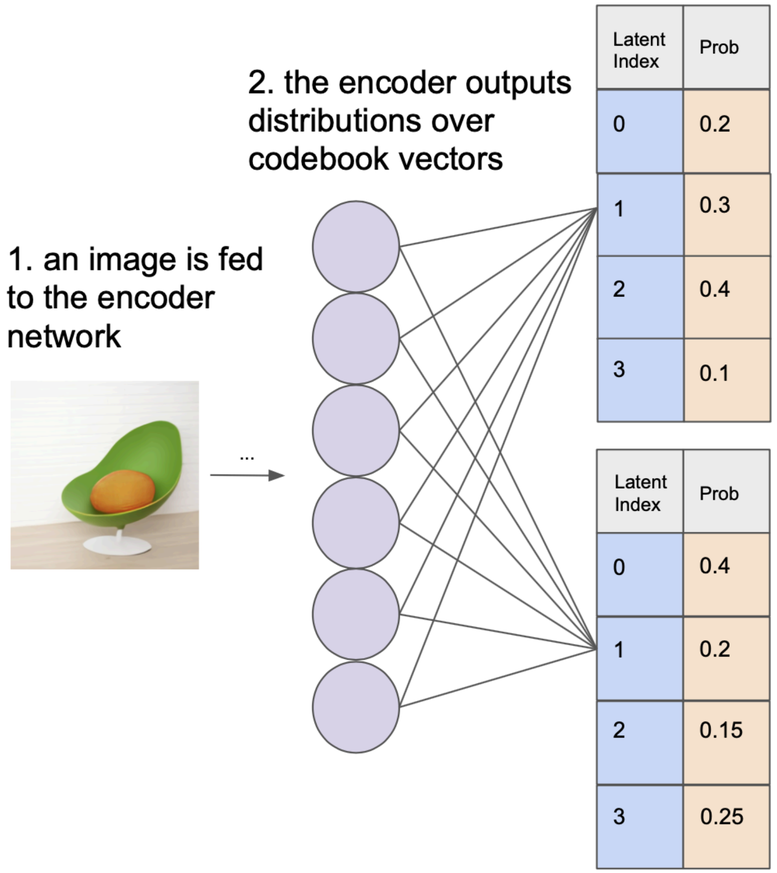
↑ Encoder

Input: x

dVAE

Отличия от VQ-VAE:

- вместо выбора одного индекса для codebook предсказываем распределение



Latent Index	Codebook Vector
0	[0.01, -2.3, 5.6, 0.04, -0.1, 8.92, 3.24, ...]
1	[5.4, 0.65, 0.2, 4.6, 8.9, -2.43, 0.07, ...]
2	[9.78, 0.67, -3.4, 0.2, -1.0, 7.2, 13.8, ...]
3	[2.45, -8.9, 0.3, 2.04, -0.89, 19.1, 0.3, ...]

Image credit

dVAE

Отличия от VQ-VAE:

- чтобы текли градиенты - используем дифференцируемое близкое распределение (Gumbel Softmax Relaxation)

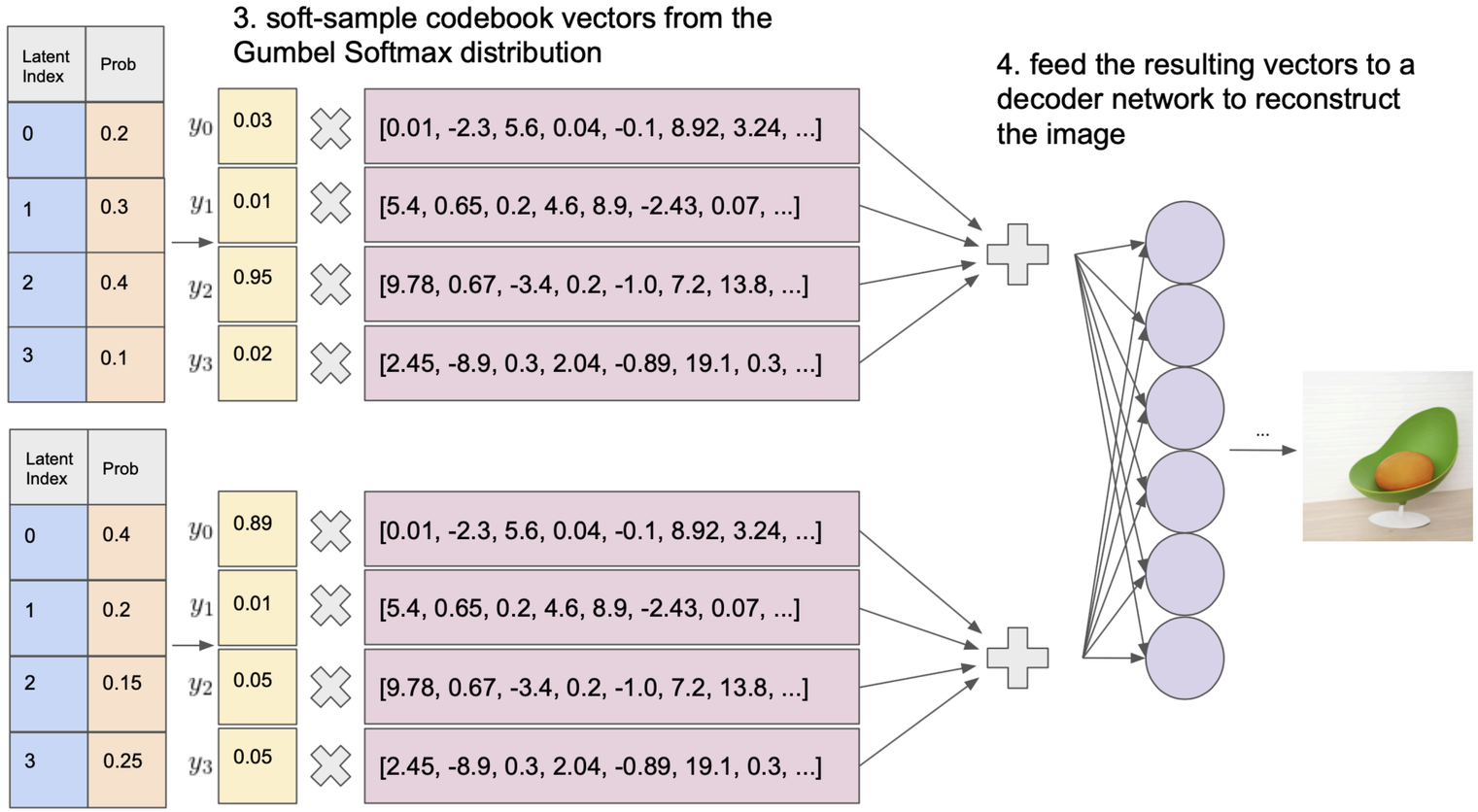


Image credit

dVAE

Loss: $-\mathrm{E}_{z\sim q(z|x)} \log(p(x\,|\,z)) + KL(q(z\,|\,x)\,||\,p(z))$

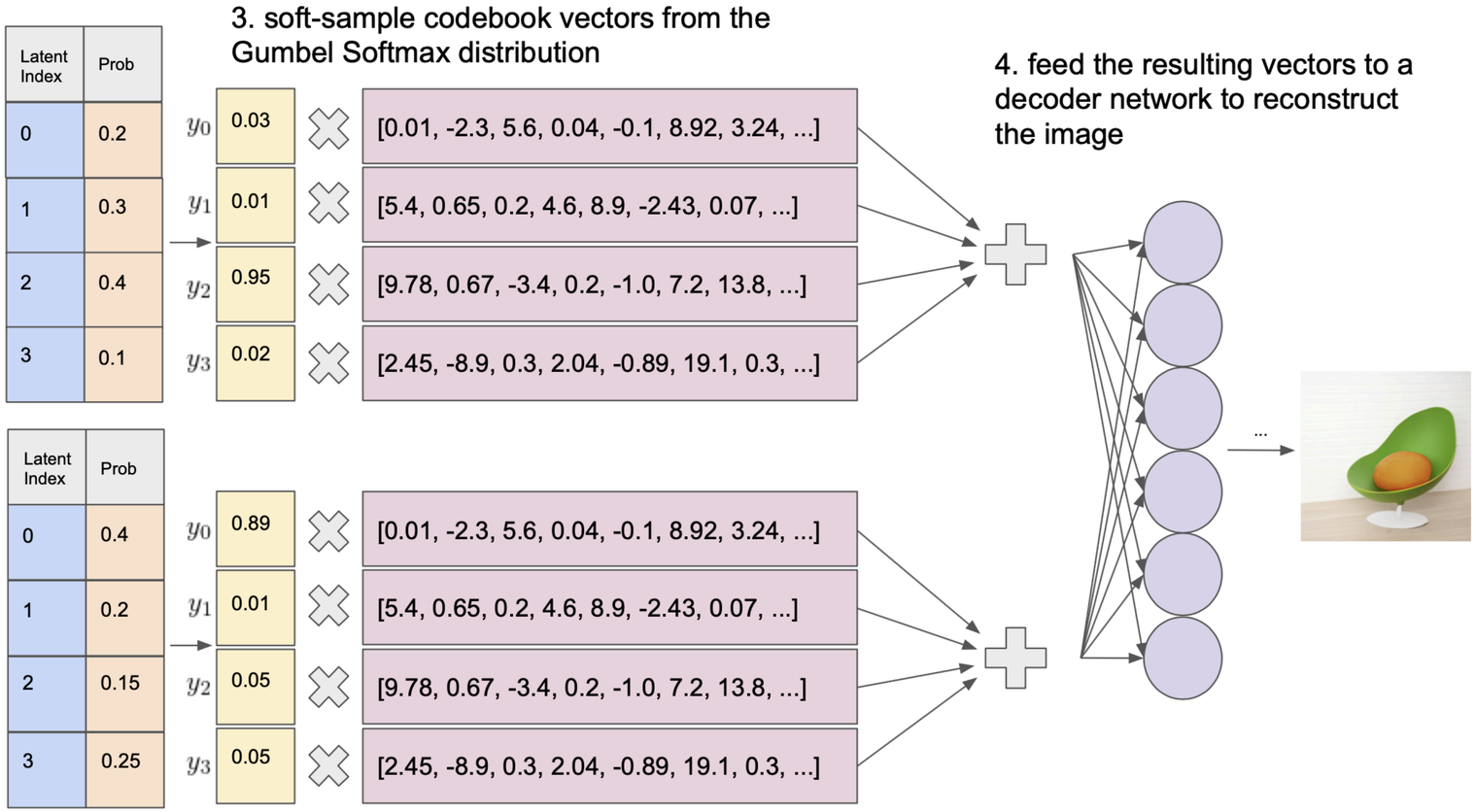


Image credit

dVAE

relaxation distribution (from encoder)

uniform distribution

Loss: $-\mathrm{E}_{z\sim q(z|x)}\log(p(x\,|\,z)) + KL(q(z\,|\,x)\,||\,p(z))$

KL - “расстояние” между распределениями

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x) \log\left(\frac{P(x)}{Q(x)}\right).$$

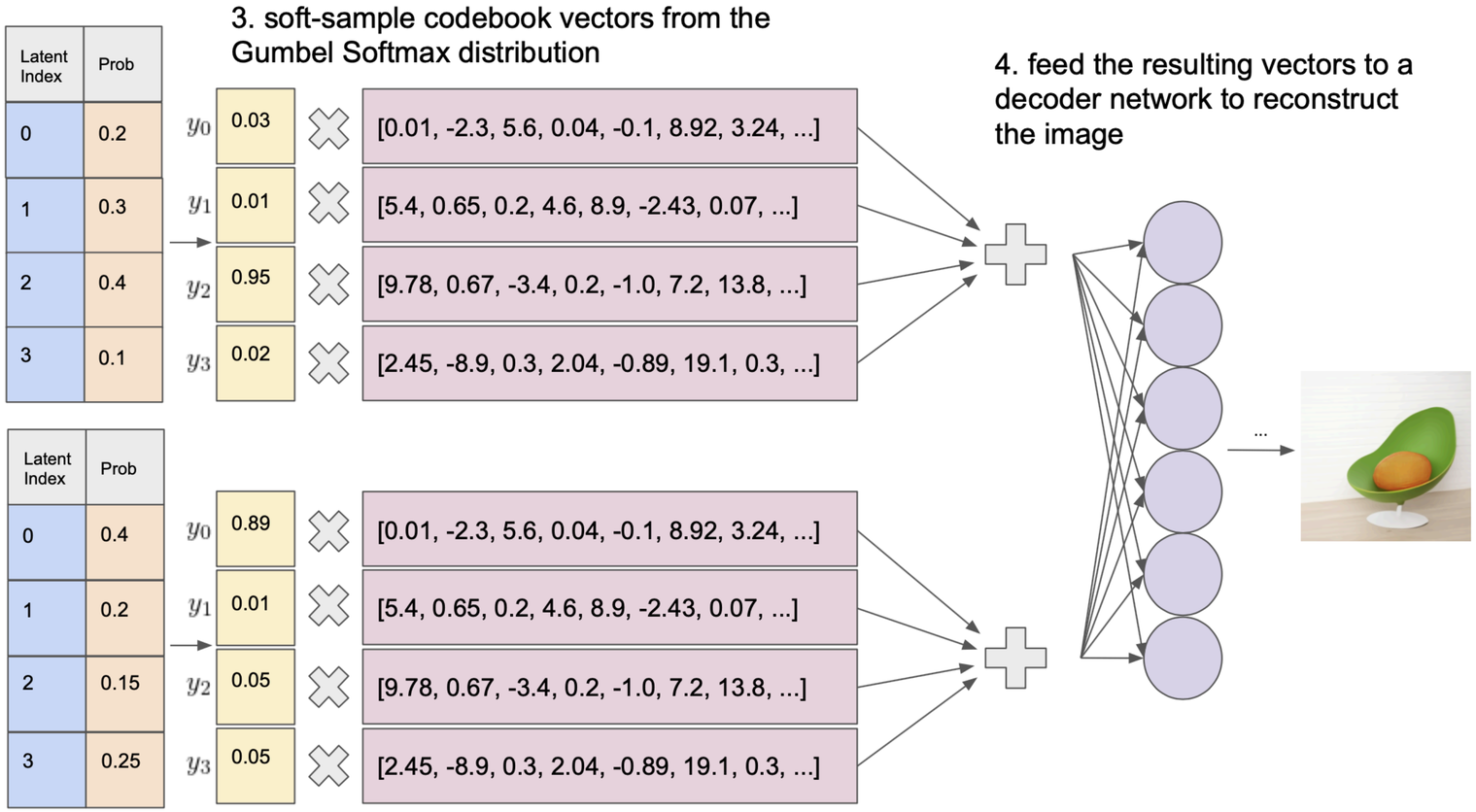


Image credit

dVAE

Decoder output

Loss: $-\mathrm{E}_{z\sim q(z|x)} \log(p(x|z)) + KL(q(z|x)||p(z))$

Полученная картинка должна быть похожа на исходную

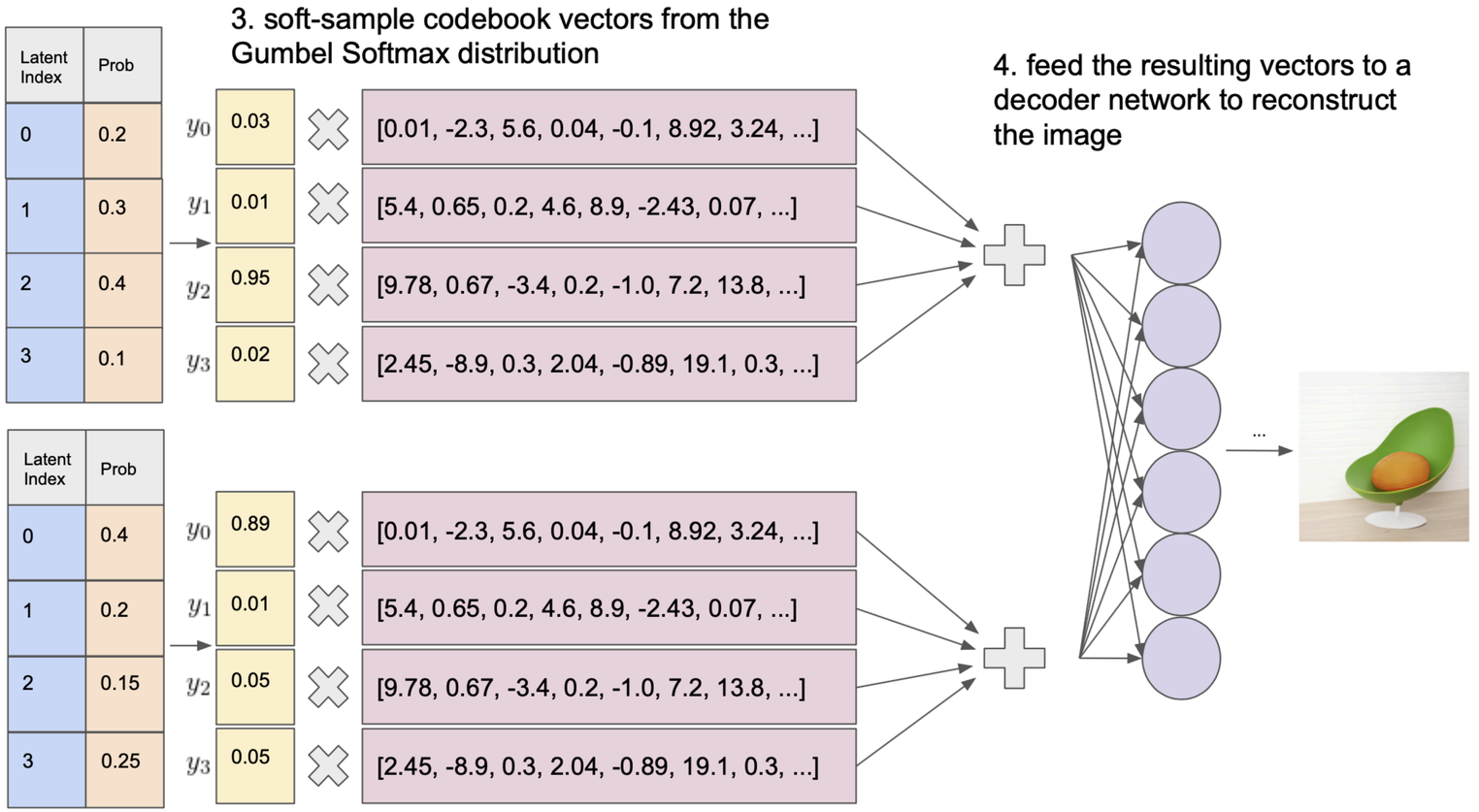


Image credit

dVAE

Если есть распределение над индексами codebook - можно генерировать картинку (decoder dVAE)

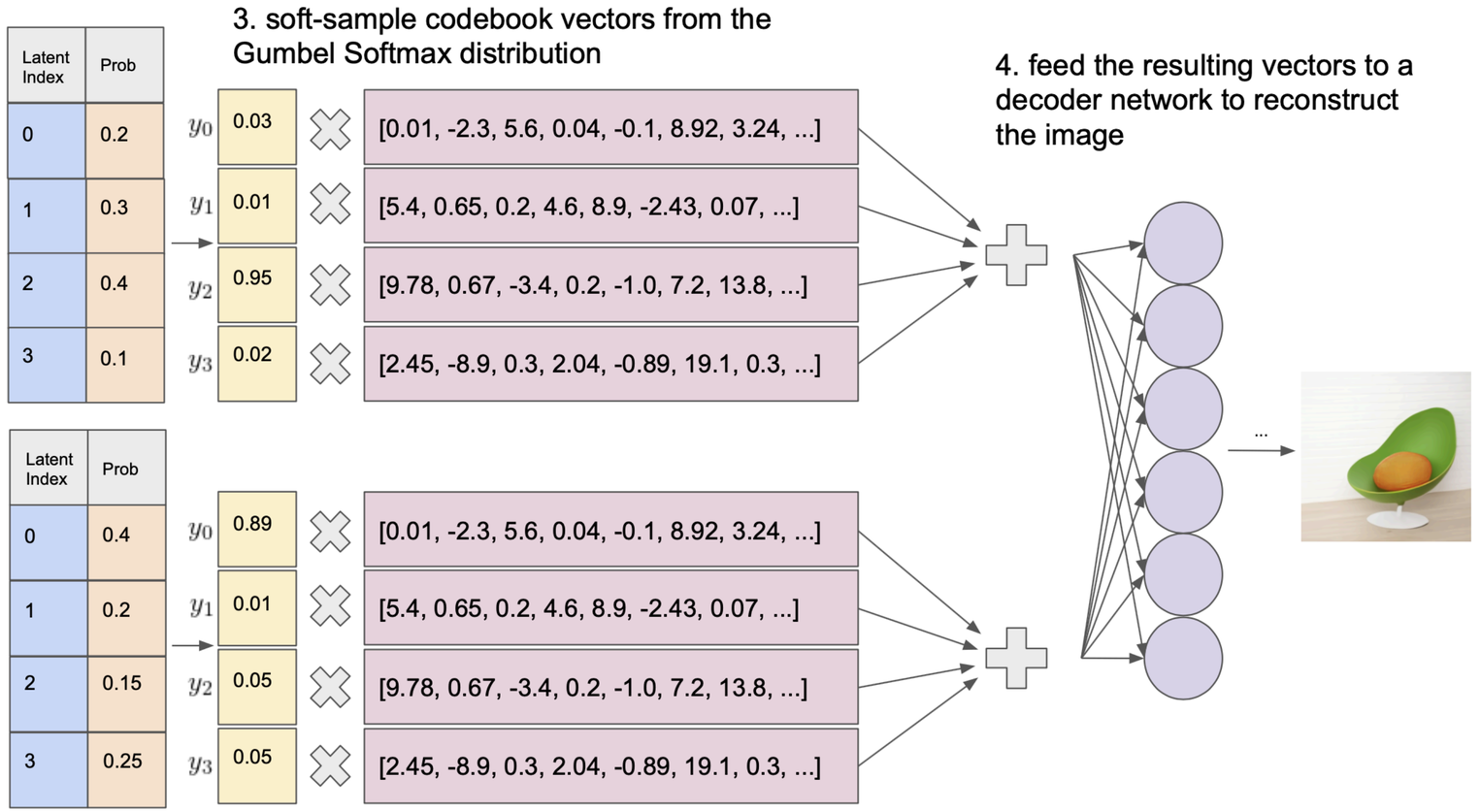


Image credit

DALL-E

1. Transformer по входному тексту авторегрессионно генерирует индексы

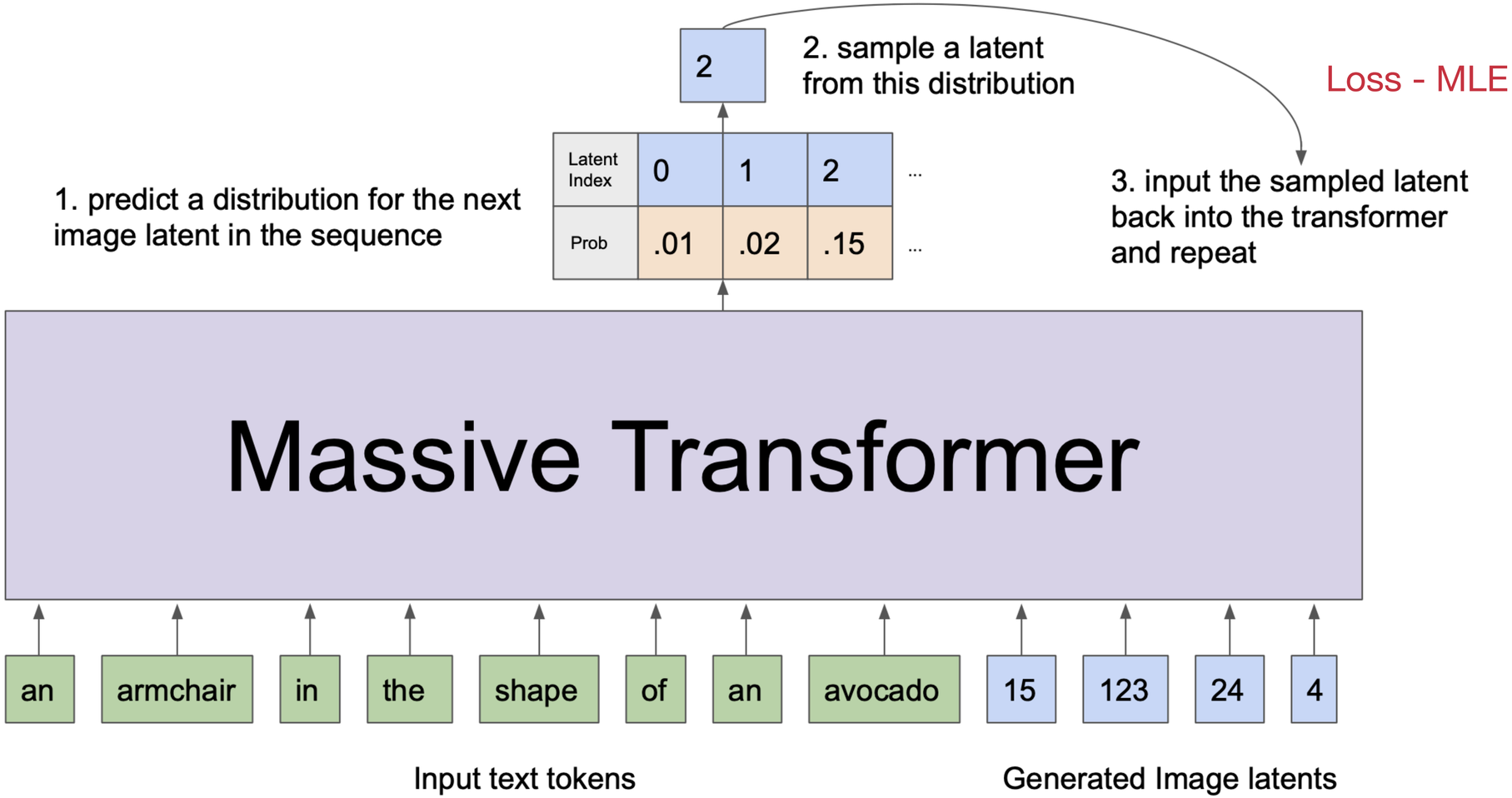


Image credit

DALL-E

2. dVAE по сгенерированным индексам генерирует картинку

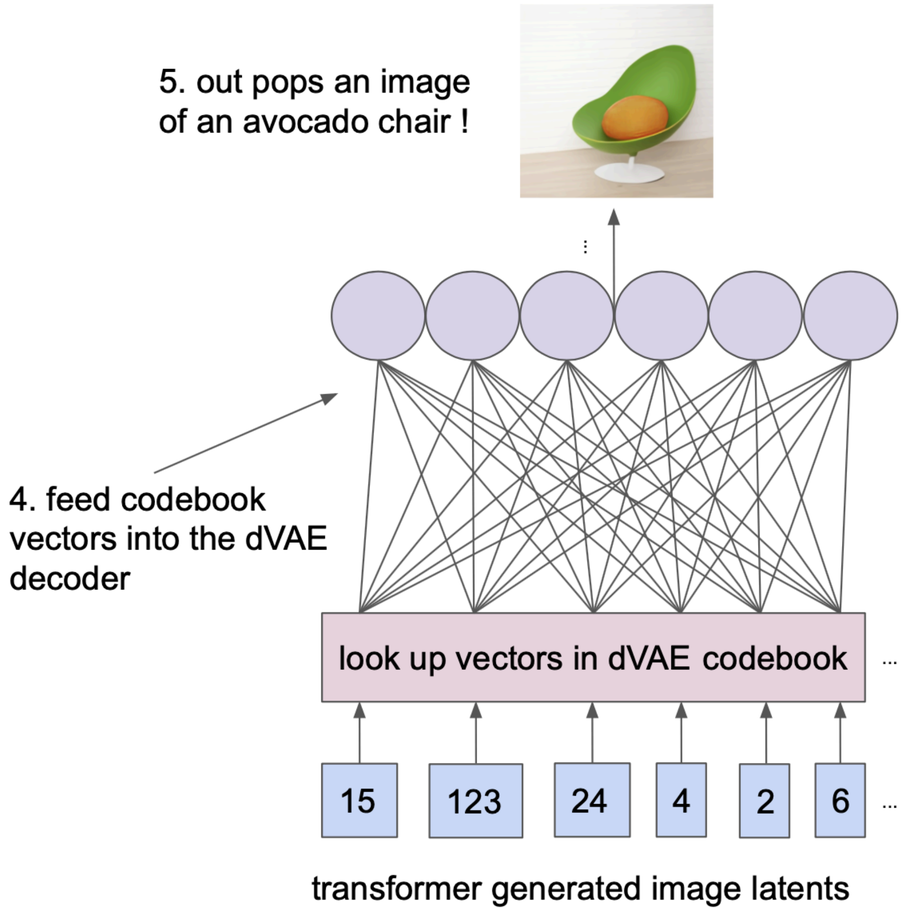


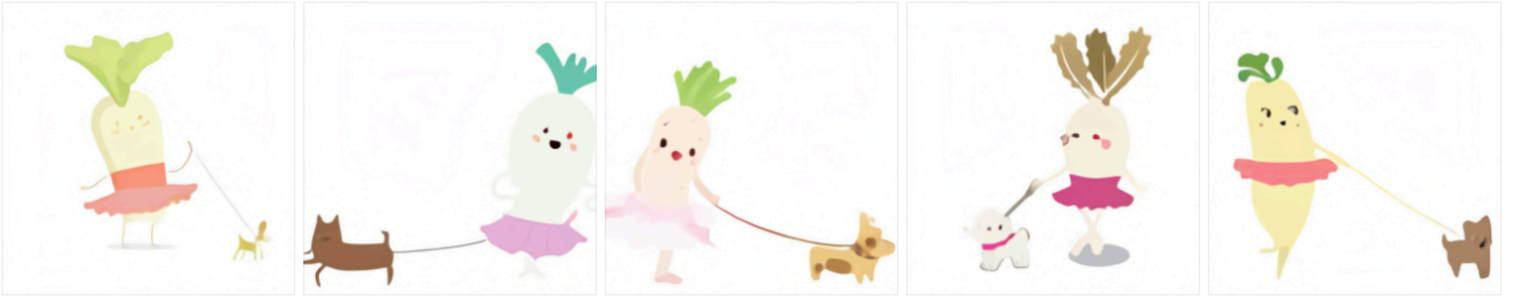
Image credit

DALL-E

TEXT PROMPT

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

AI-GENERATED IMAGES



TEXT & IMAGE PROMPT

the exact same cat on the top as a sketch on the bottom

AI-GENERATED IMAGES

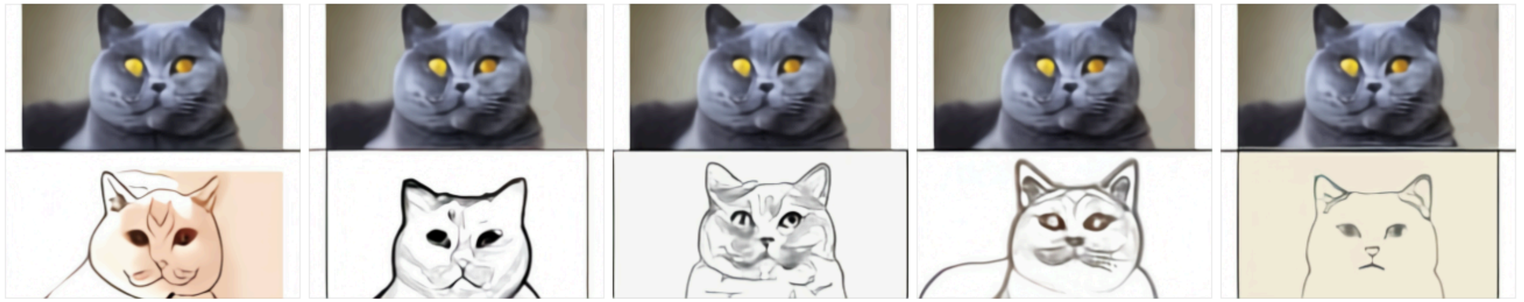


Image credit

DALL-E 2

Модель: CLIP + GLIDE (Denoising Diffusion Probabilistic Model)

a dolphin in an astronaut suit on saturn, artstation

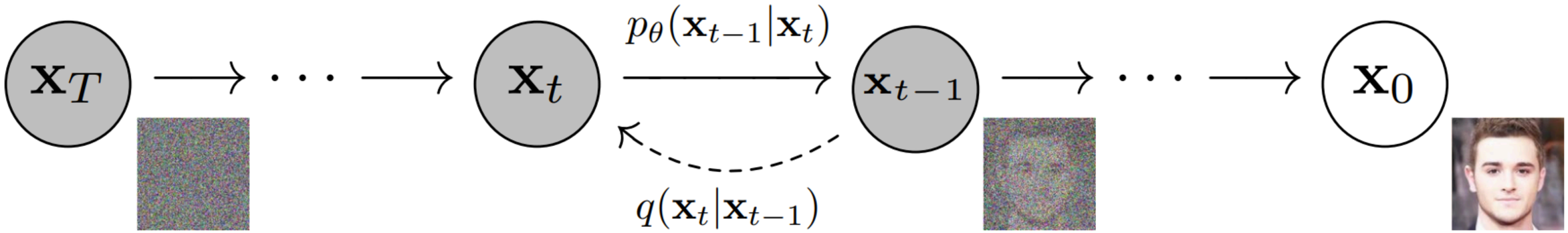
a propaganda poster depicting a cat dressed as french emperor napoleon holding a piece of cheese

a teddy bear on a skateboard in times square

Image credit

DDPM: overview

DDPM:

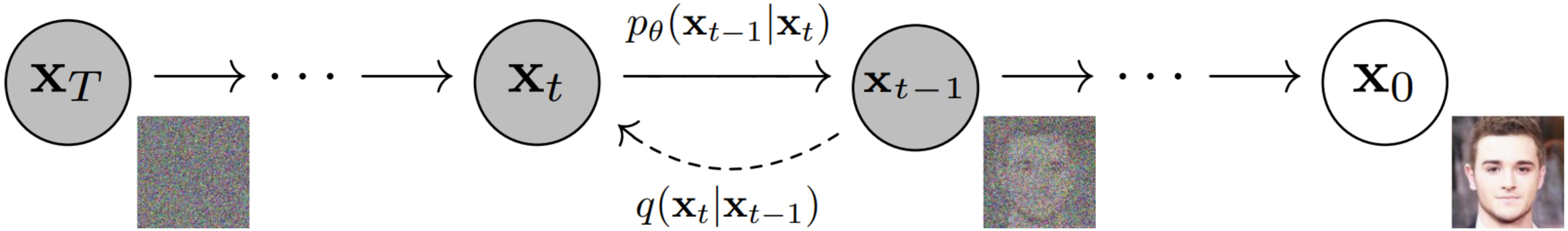


Добавляем случайный шум на каждом шаге

Image credit

DDPM: overview

DDPM:



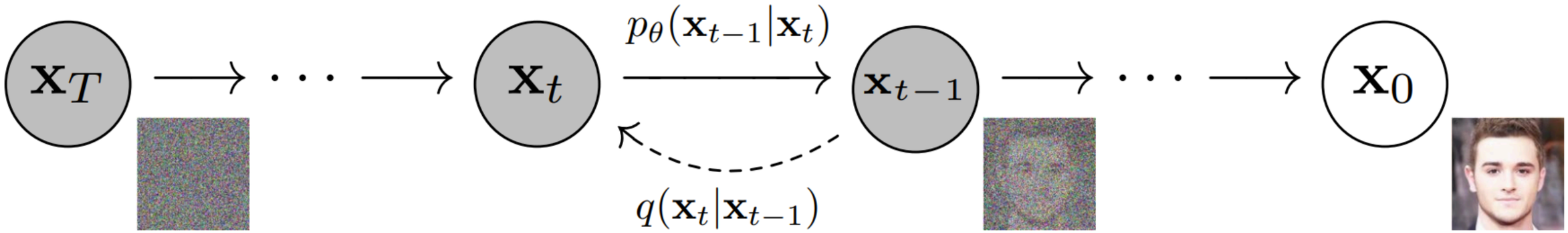
Восстанавливаем (убираем добавленный шум)

Image credit

DDPM: overview

DDPM:

Markov chain



Восстанавливаем (убираем добавленный шум)

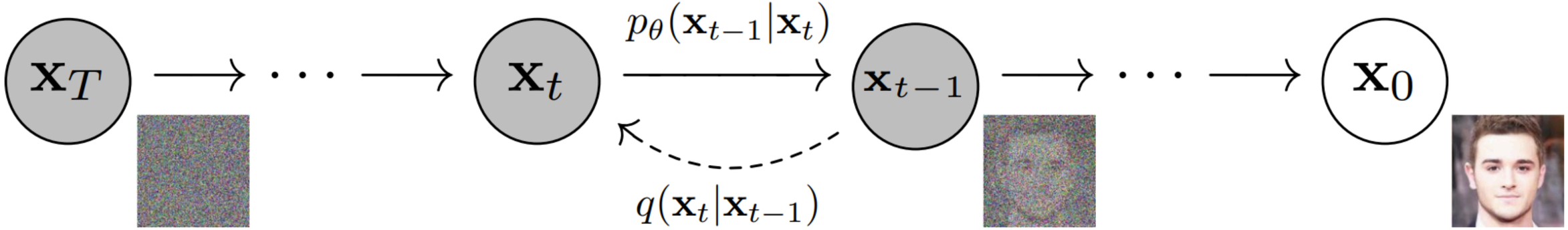
Каждый раз применяем модель типа UNet

Image credit

DDPM: overview

DDPM:

Markov chain

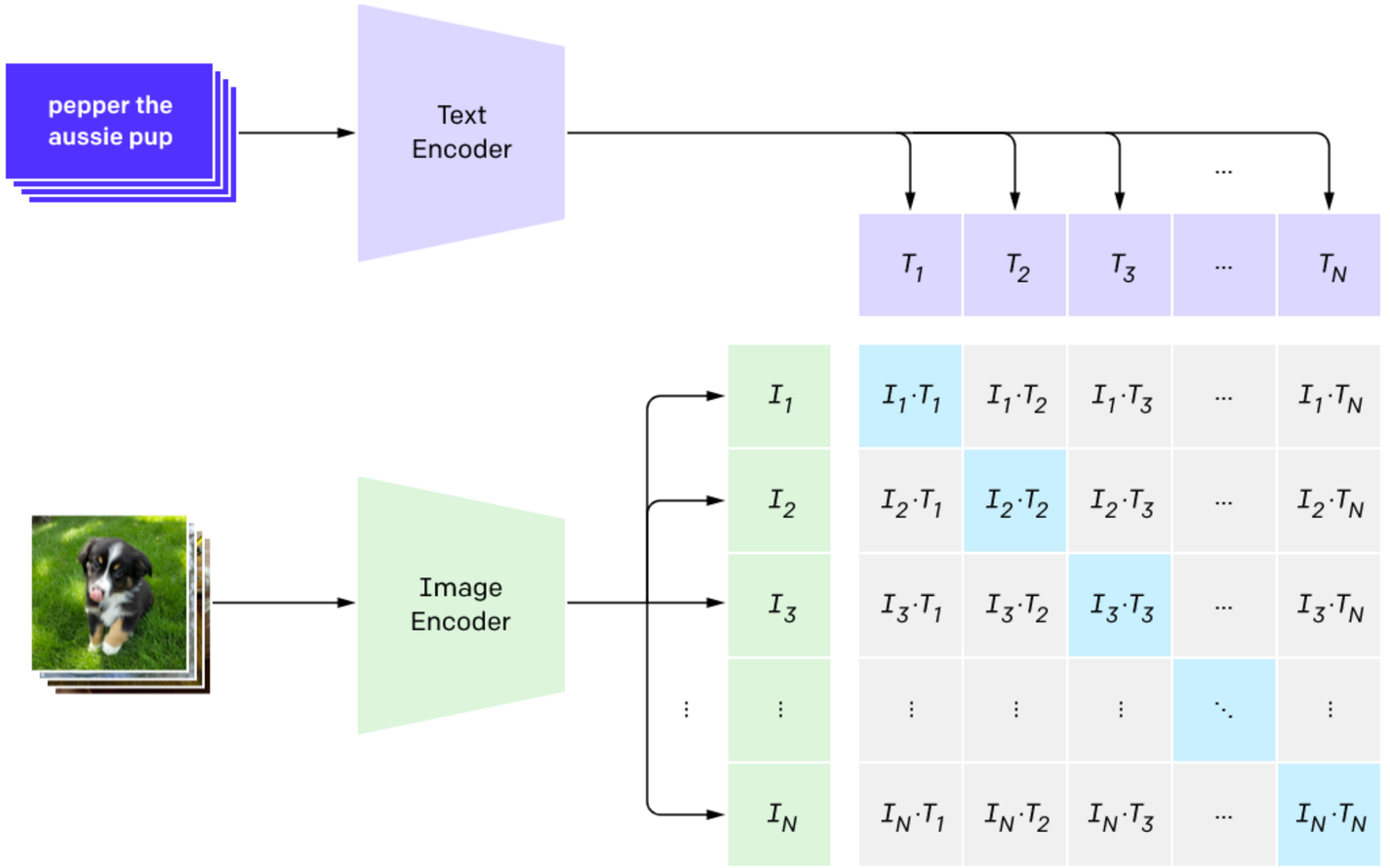


$$L_{\text{simple}} := E_{t\sim[1,T],x_0\sim q(x_0),\epsilon\sim\mathcal{N}(0,\mathbf{I})}[||\epsilon - \epsilon_\theta(x_t, t)||^2]$$

Image credit

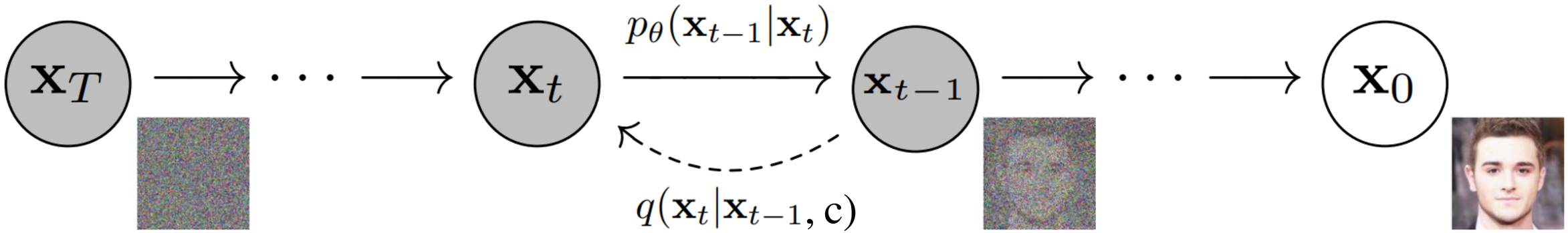
DALL-E 2

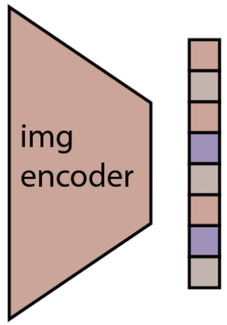
1. Обучаем CLIP



DALL-E 2

1. Обучаем CLIP
2. Обучаем DDPM обусловленную на CLIP image encoding (GLIDE)





c - CLIP image encoding

Image credit

DALL-E 2

1. Обучаем CLIP
2. Обучаем DDPM обусловленную на CLIP image encoding (GLIDE)
3. Обучаем другую DDPM предсказывать CLIP image encoding (по text encoding)

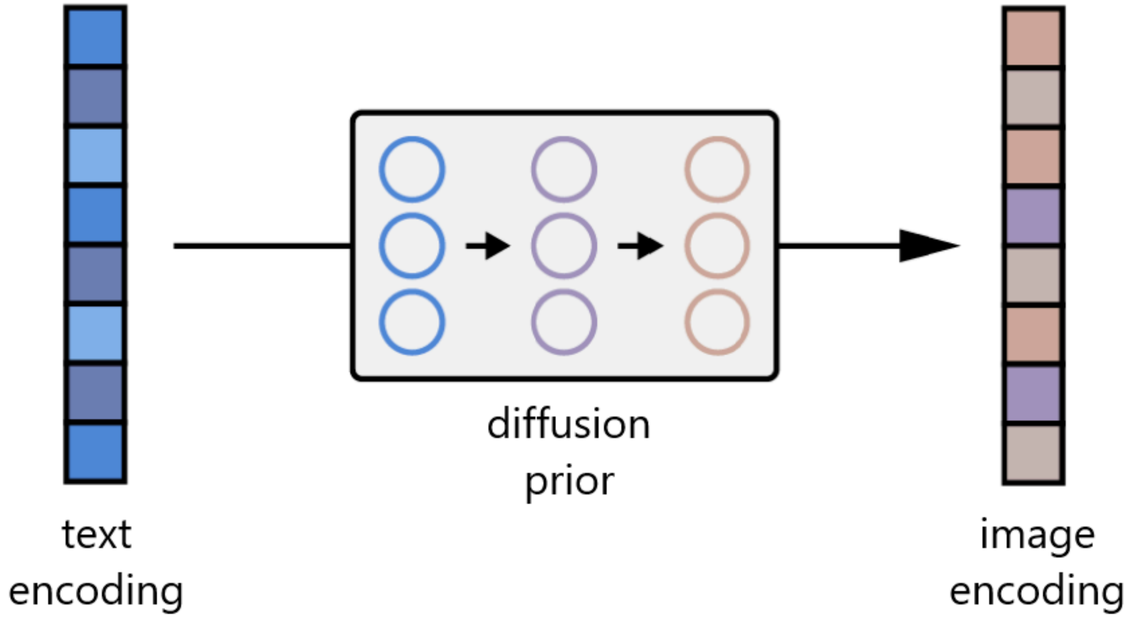


Image credit

DALL-E 2

Генерация:

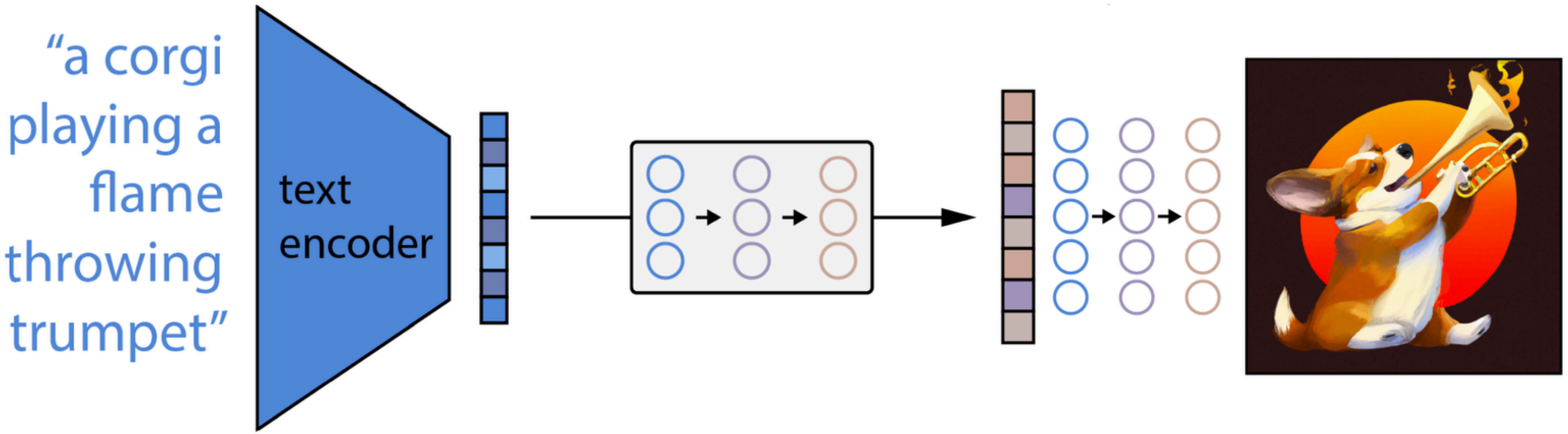


Image credit

DALL-E 2

an espresso machine that makes coffee from human souls, artstation

panda mad scientist mixing sparkling chemicals, artstation

a corgi’s head depicted as an explosion of a nebula