Глубинное обучение

Лазарев Михаил

Pre-trained Transformer

Если есть обученный для перевода Transformer, то:

- **Encoder**: выучил хорошие признаки для слов на языке входа
- **Decoder**: выучил хорошие признаки для слов на языке выхода

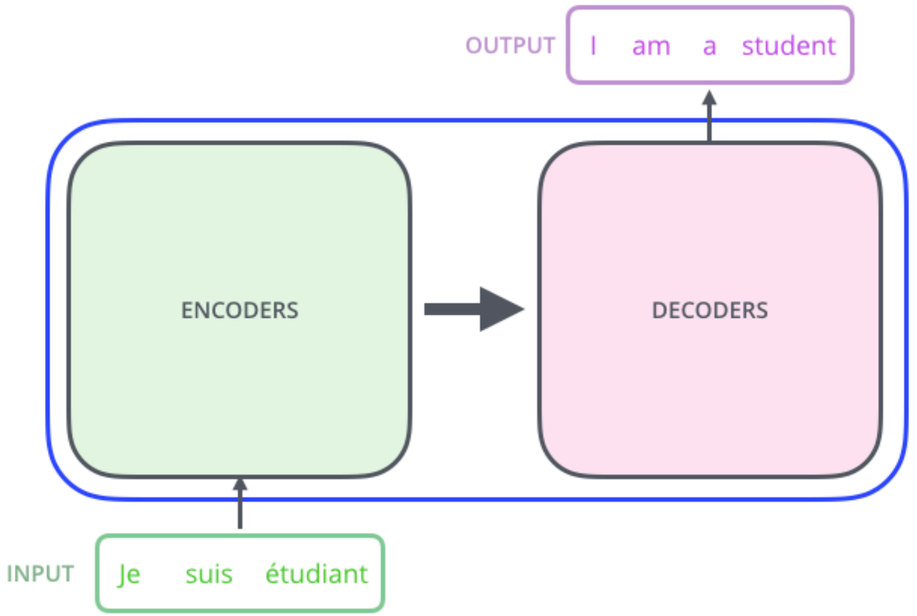


Image credit

Pre-trained Transformer

Если есть обученный для перевода Transformer, то:

- **Encoder**: выучил хорошие признаки для слов на языке входа
- **Decoder**: выучил хорошие признаки для слов на языке выхода

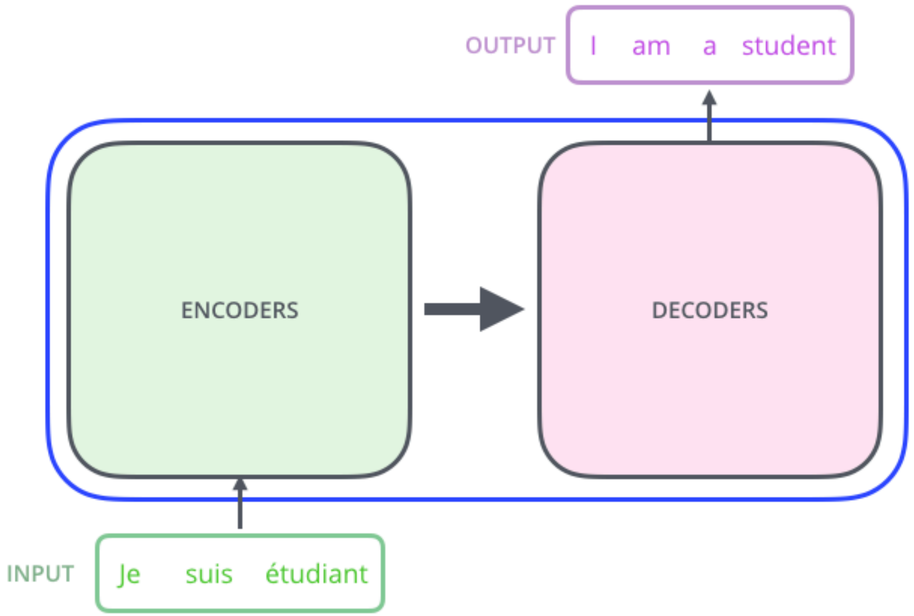


Image credit

Pre-trained Transformer

Для каждого слова Transformer block выдает эмбеддинг, зависящий от всего контекста

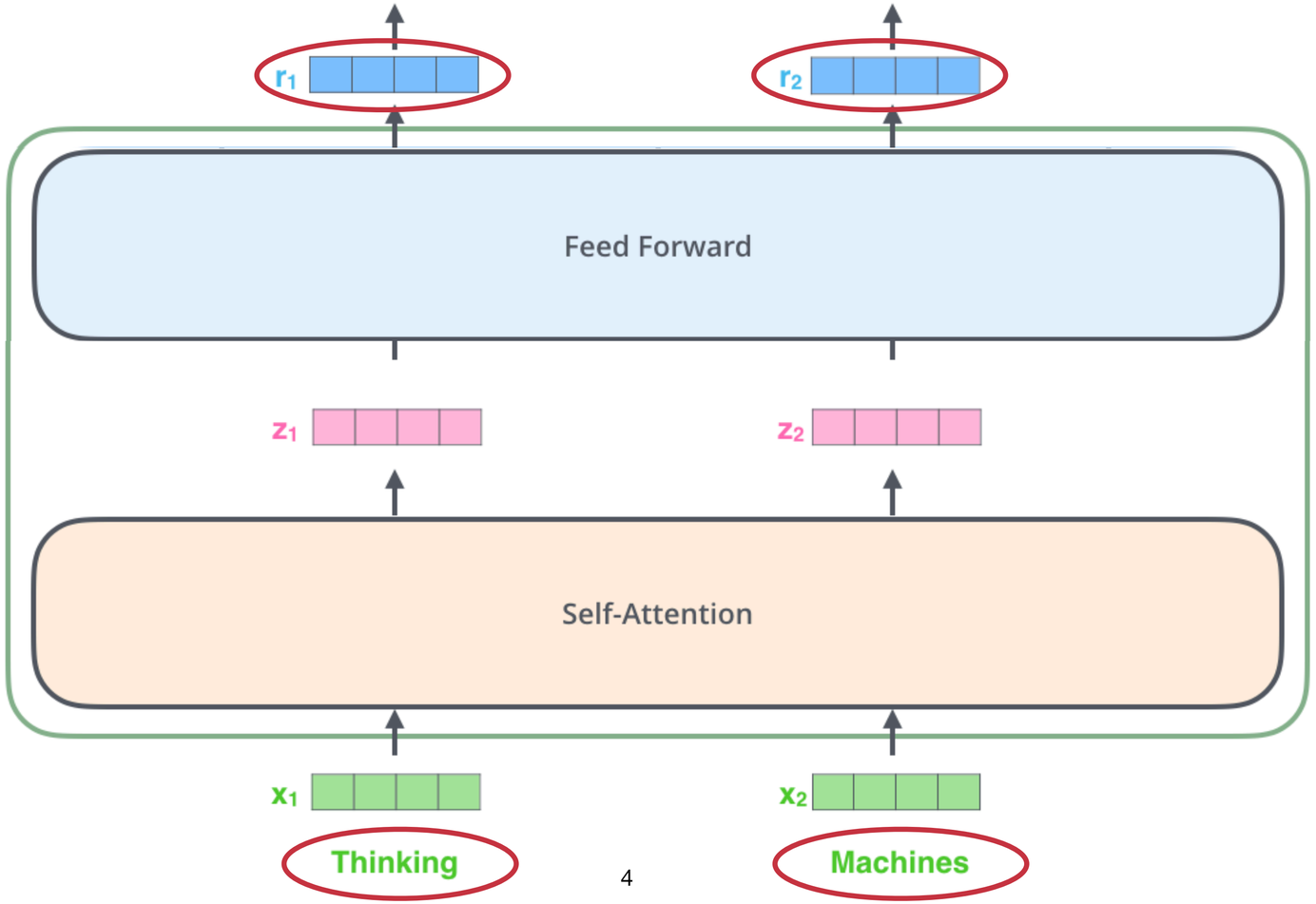


Image credit

Pre-trained Transformer

Для каждого слова Transformer block выдает эмбеддинг, зависящий от всего контекста

- это верно и для train, и для inference

Word2vec: обучаются с учетом контекста, используются независимо

Image credit

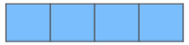
Pre-trained Transformer

Inference:

Pre-trained Transformer

Inference:

*A **plane** crash* → Transformer → *“plane” embedding (in context 1)*

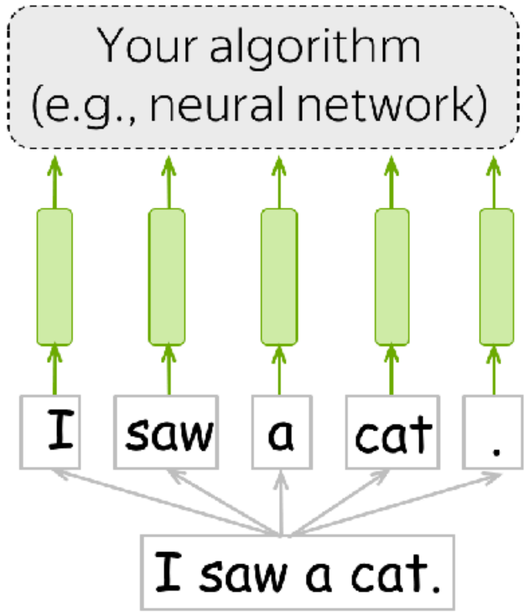


*A **plane** surface* → Transformer → *“plane” embedding (in context 2)*

*A **plane** crash /*
*A **plane** surface* → Word2Vec → *“plane” embedding (equal for both contexts)*

Pre-trained Transformer

Как мы используем эмбеддинги?



Any algorithm for solving a task

Word representation - vector
(input for your model/algorithm)

Sequence of tokens

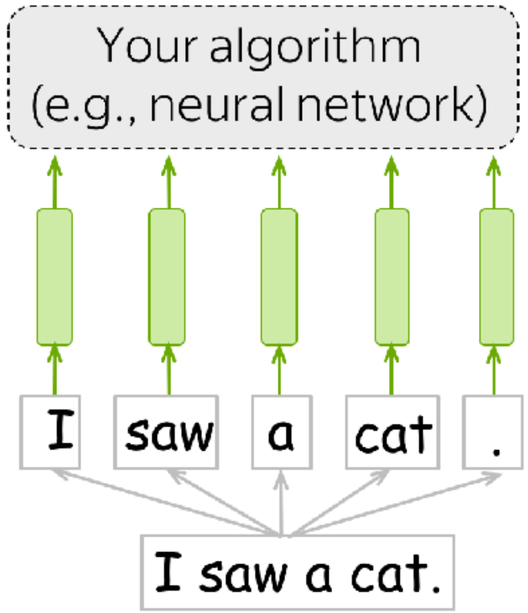
Text (your input)

Image credit

Pre-trained Transformer

Основная идея: взять предобученные эмбеддинги (преобученный Transformer) и дообучить на нужную задачу (возможно, с добавлением “головы”)

Pre-train + Fine-tune (Transfer Learning)



Any algorithm for solving a task

Word representation - vector
(input for your model/algorithm)

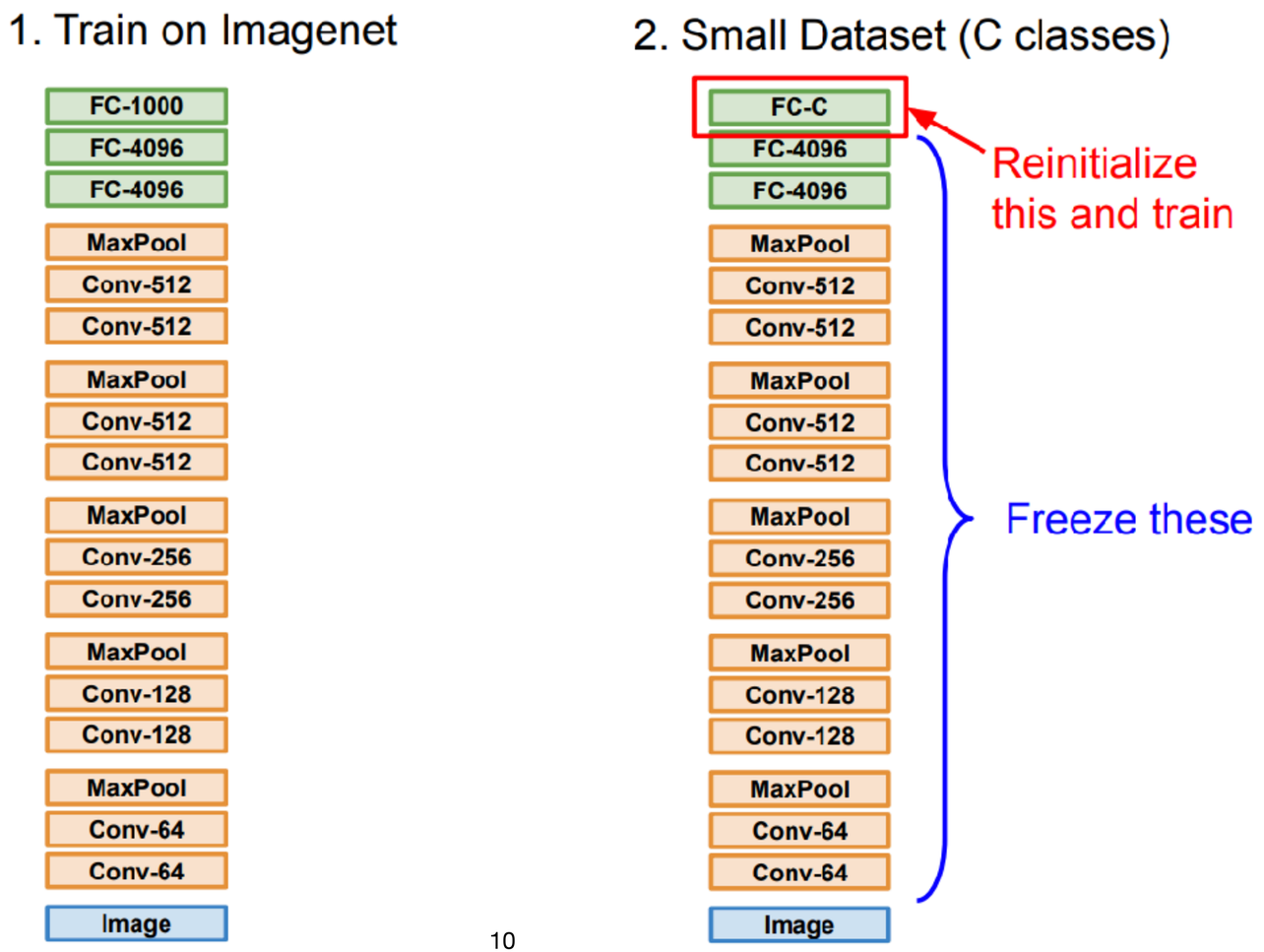
Sequence of tokens

Text (your input)

Image credit

Pre-trained Transformer

Pre-train + Fine-tune (Transfer Learning) в Computer Vision:



GPT

GPT

Generative **P**re-**T**raining

Архитектура: Transformer Decoder

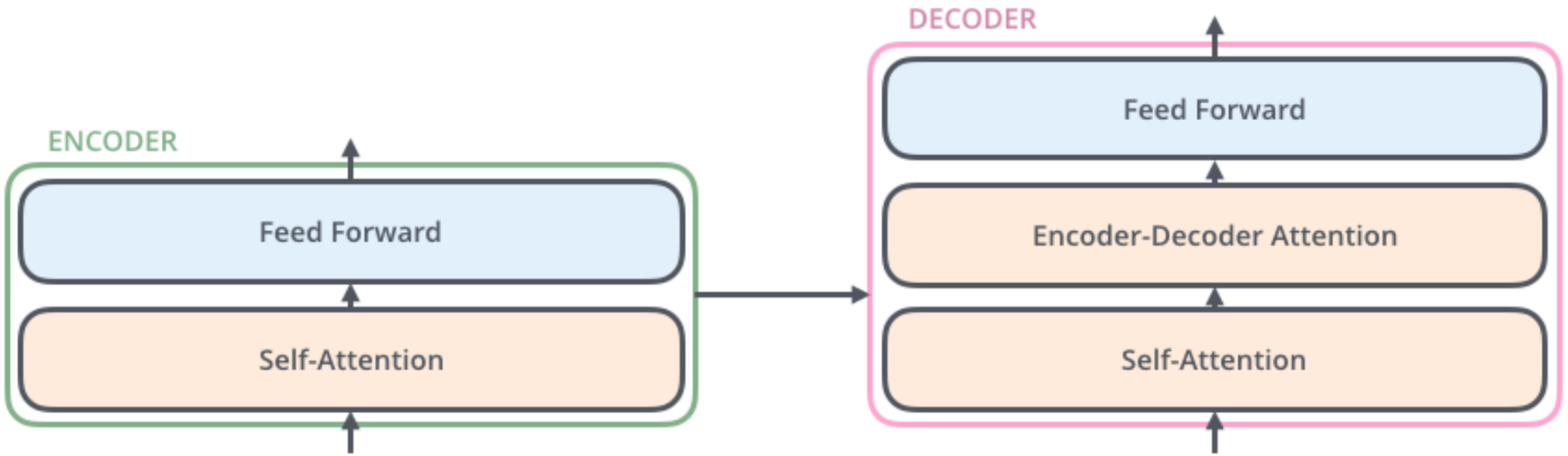


Image credit

GPT

Generative **P**re-**T**raining

Архитектура: Transformer Decoder

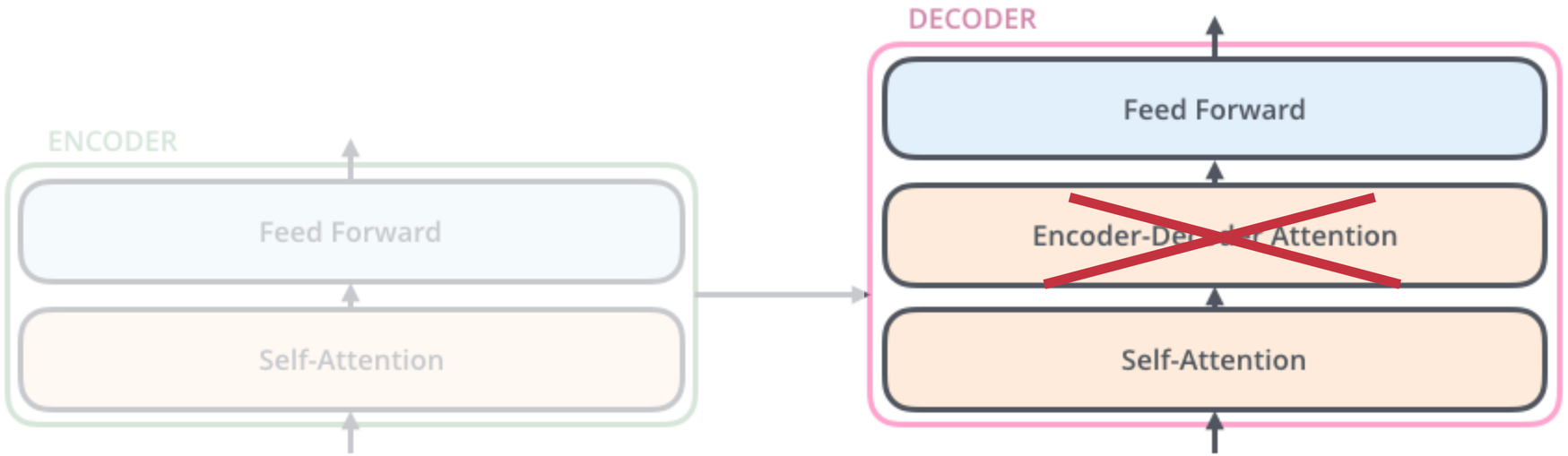
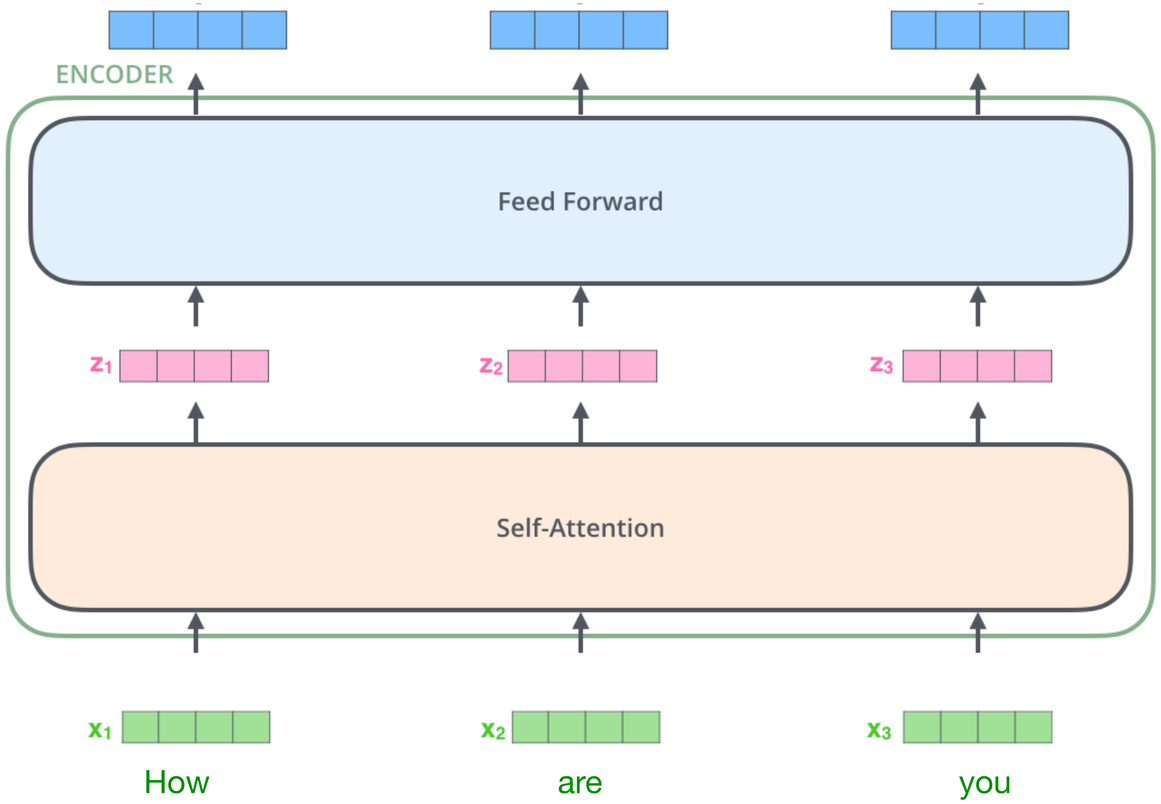


Image credit

Encoder vs Decoder

Inference

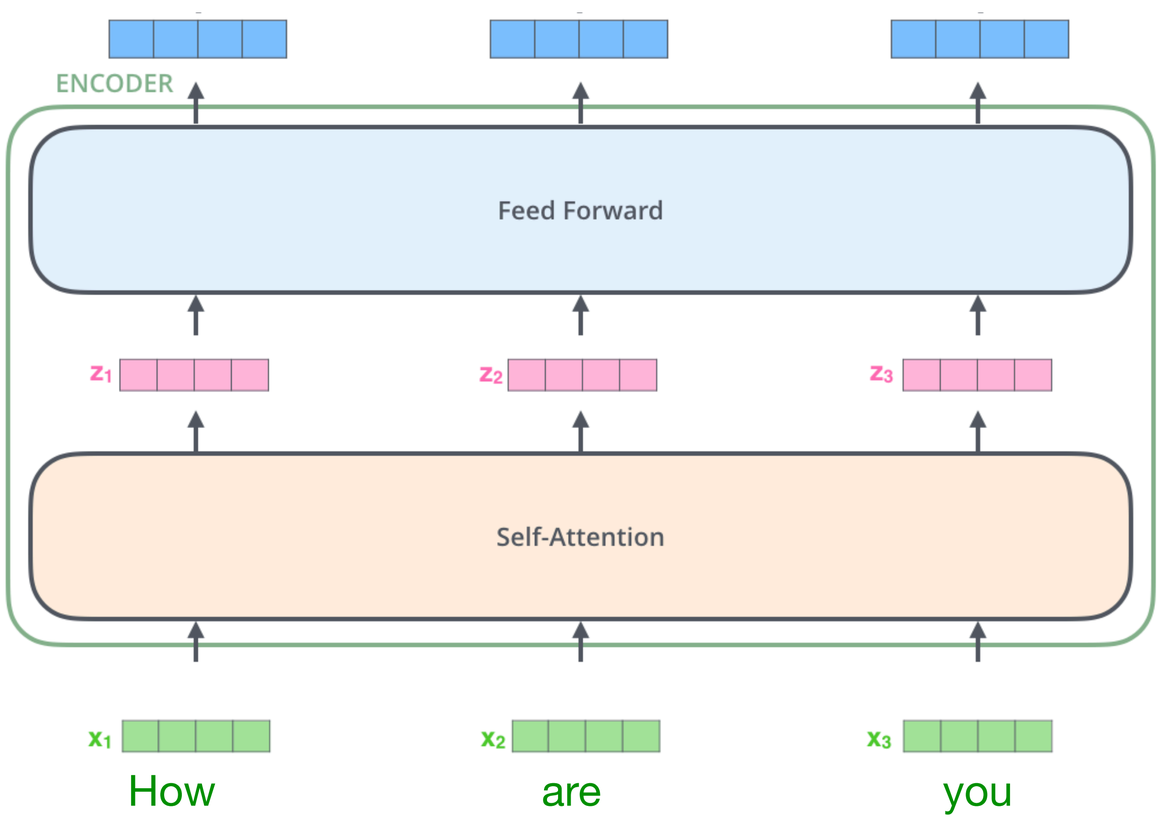


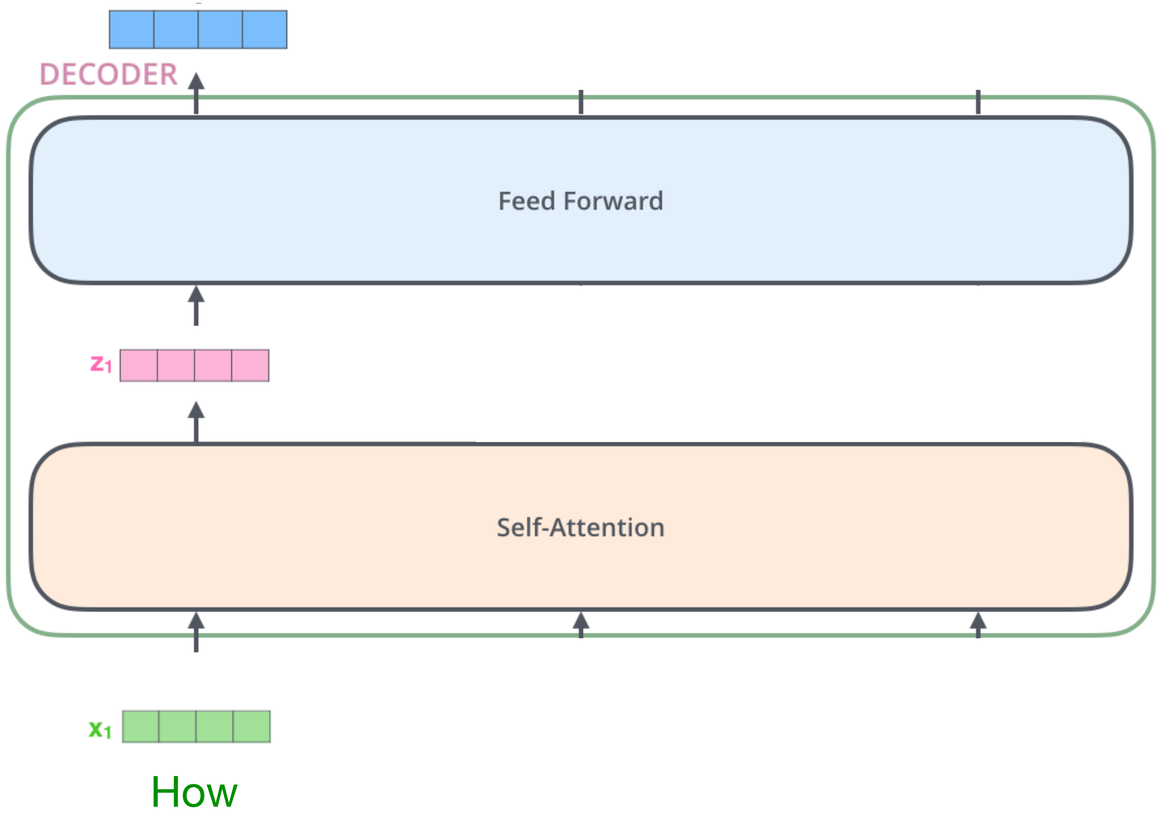
Encoder: одновременно обрабатываем

Encoder vs Decoder

Decoder step 1

Inference





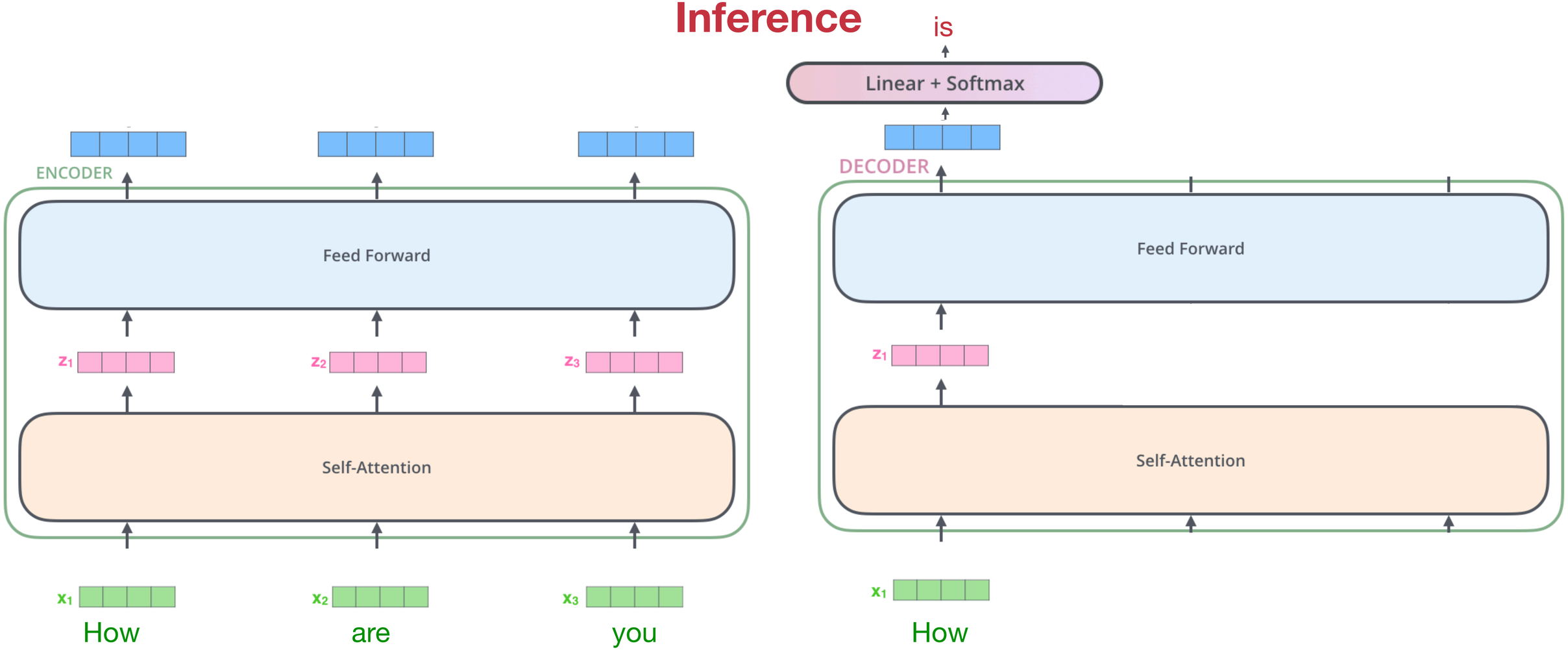
Encoder: одновременно обрабатываем

Decoder: последовательно

Encoder vs Decoder

Decoder step 1

Inference



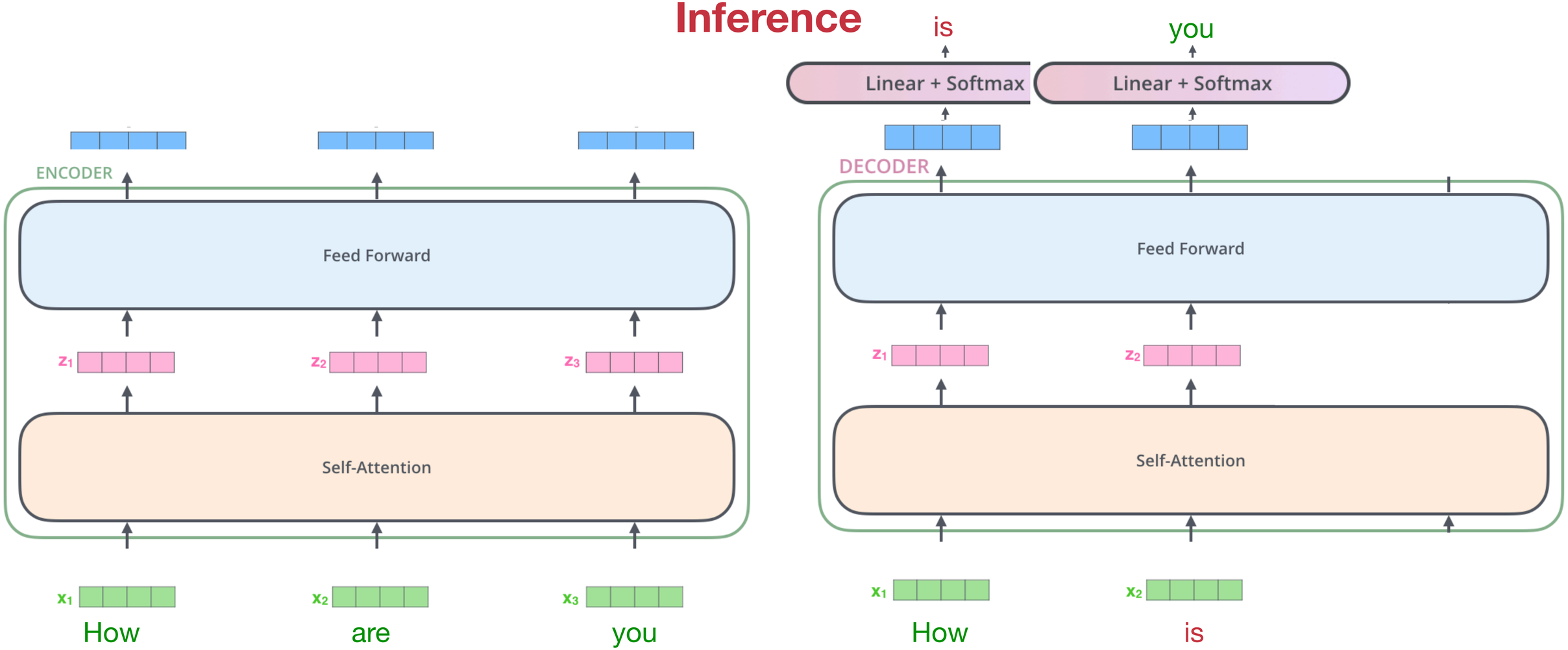
Encoder: одновременно обрабатываем

Decoder: последовательно

Encoder vs Decoder

Decoder step 2

Inference



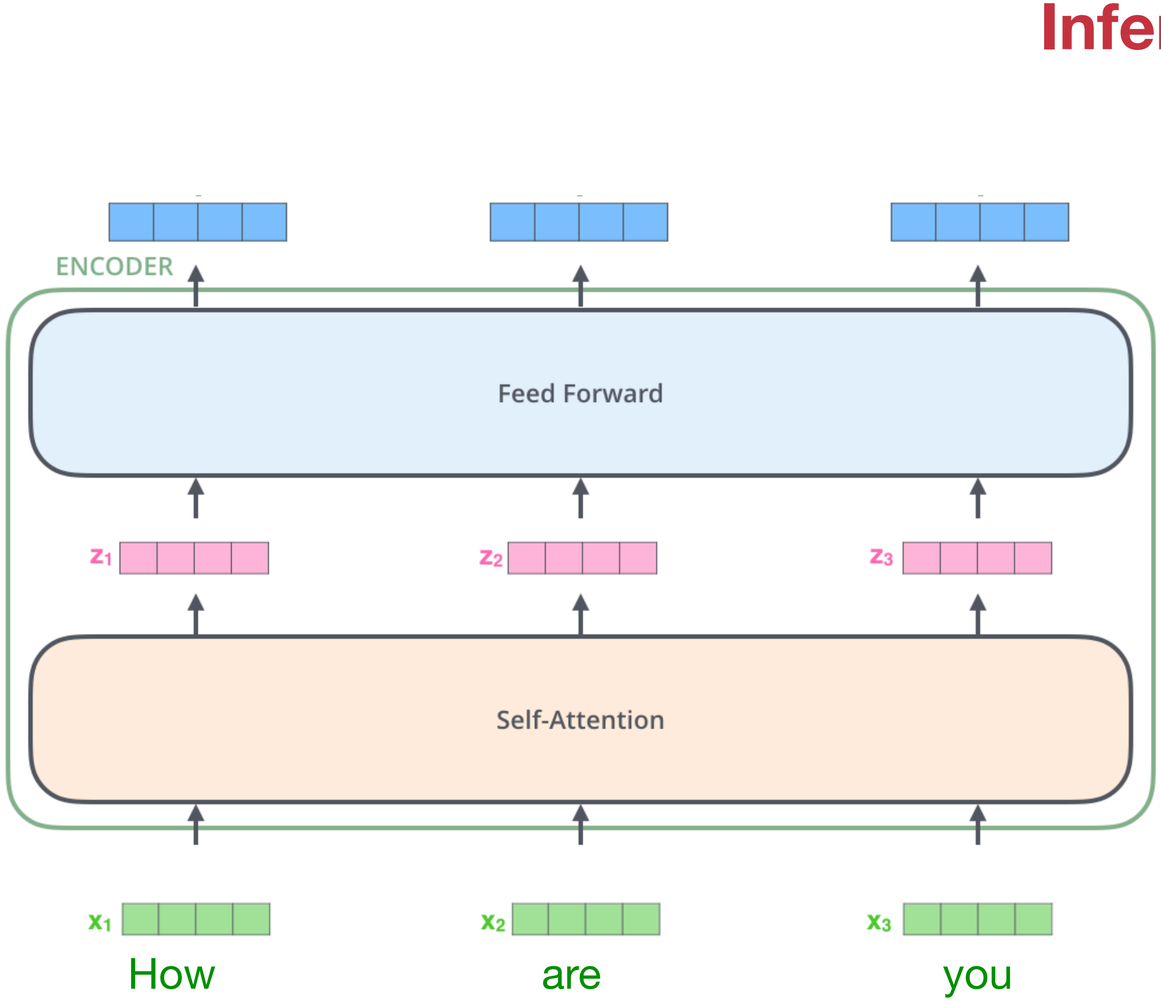
Encoder: одновременно обрабатываем

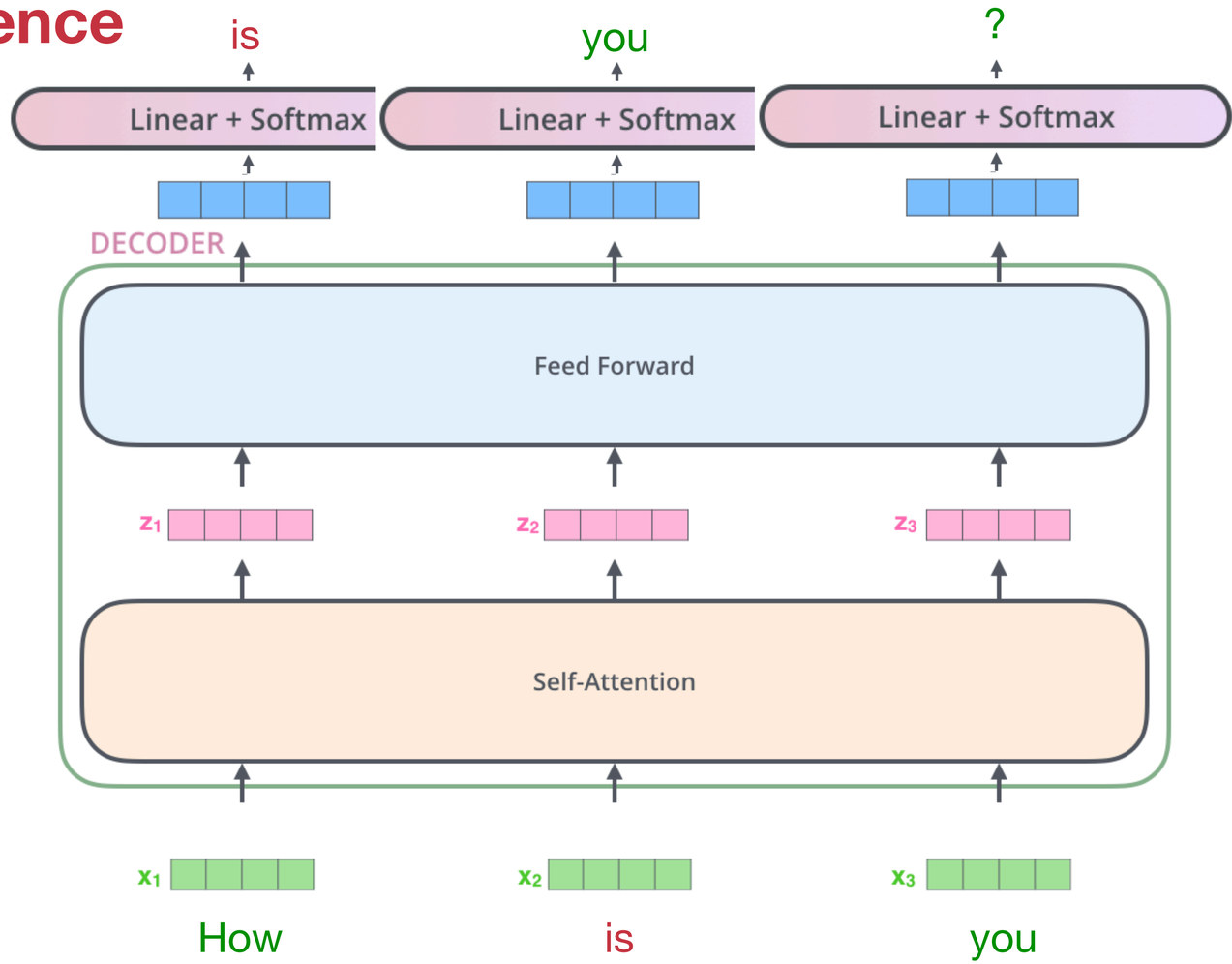
Decoder: последовательно

Encoder vs Decoder

Decoder step 3

Inference



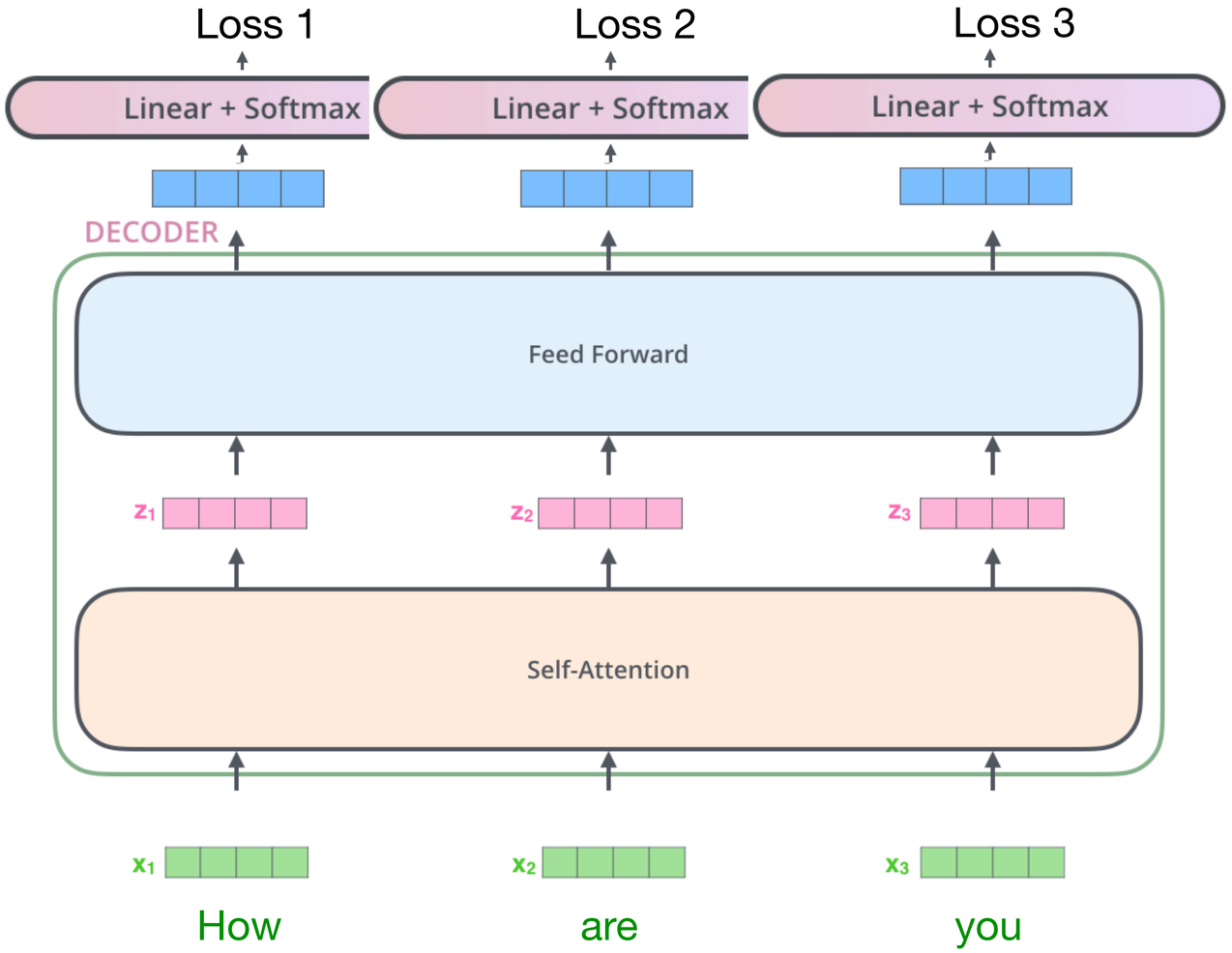


Encoder: одновременно обрабатываем

Decoder: последовательно

Encoder vs Decoder

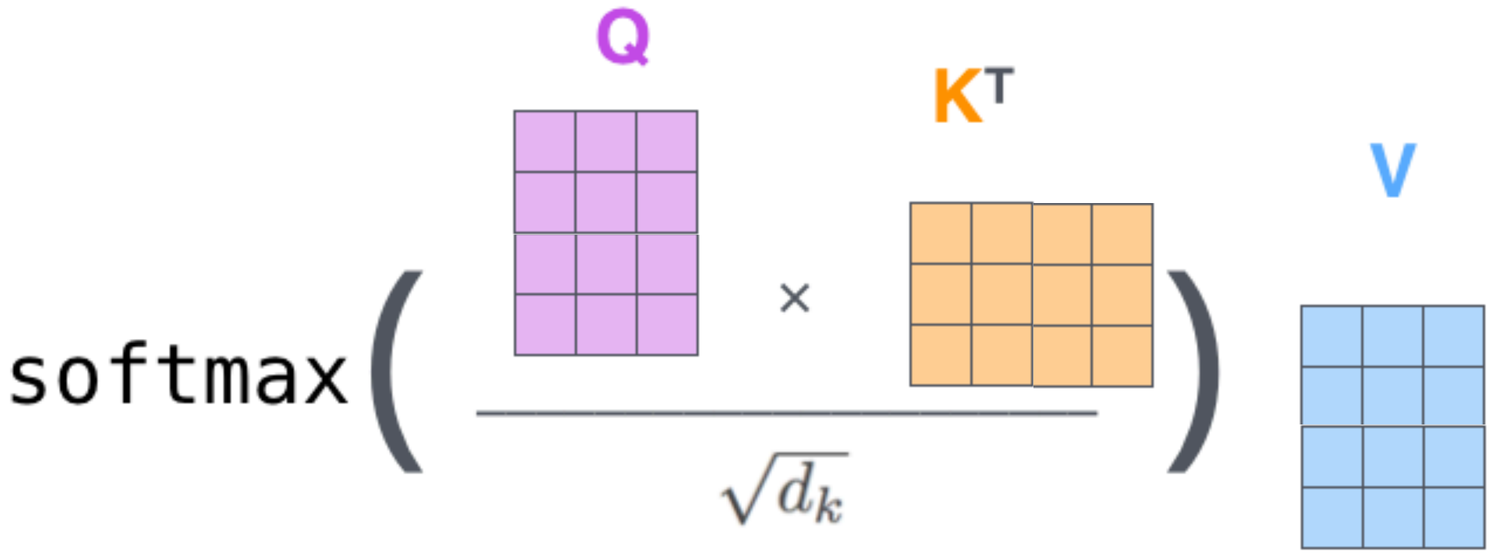
Train

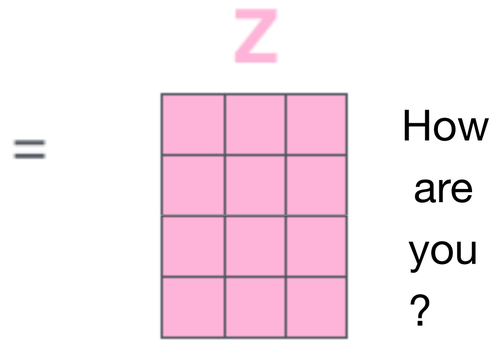


Decoder: можно посчитать одновременно

Encoder vs Decoder

Encoder self-attention: $attention = softmax(\frac{QK^T}{d})V$



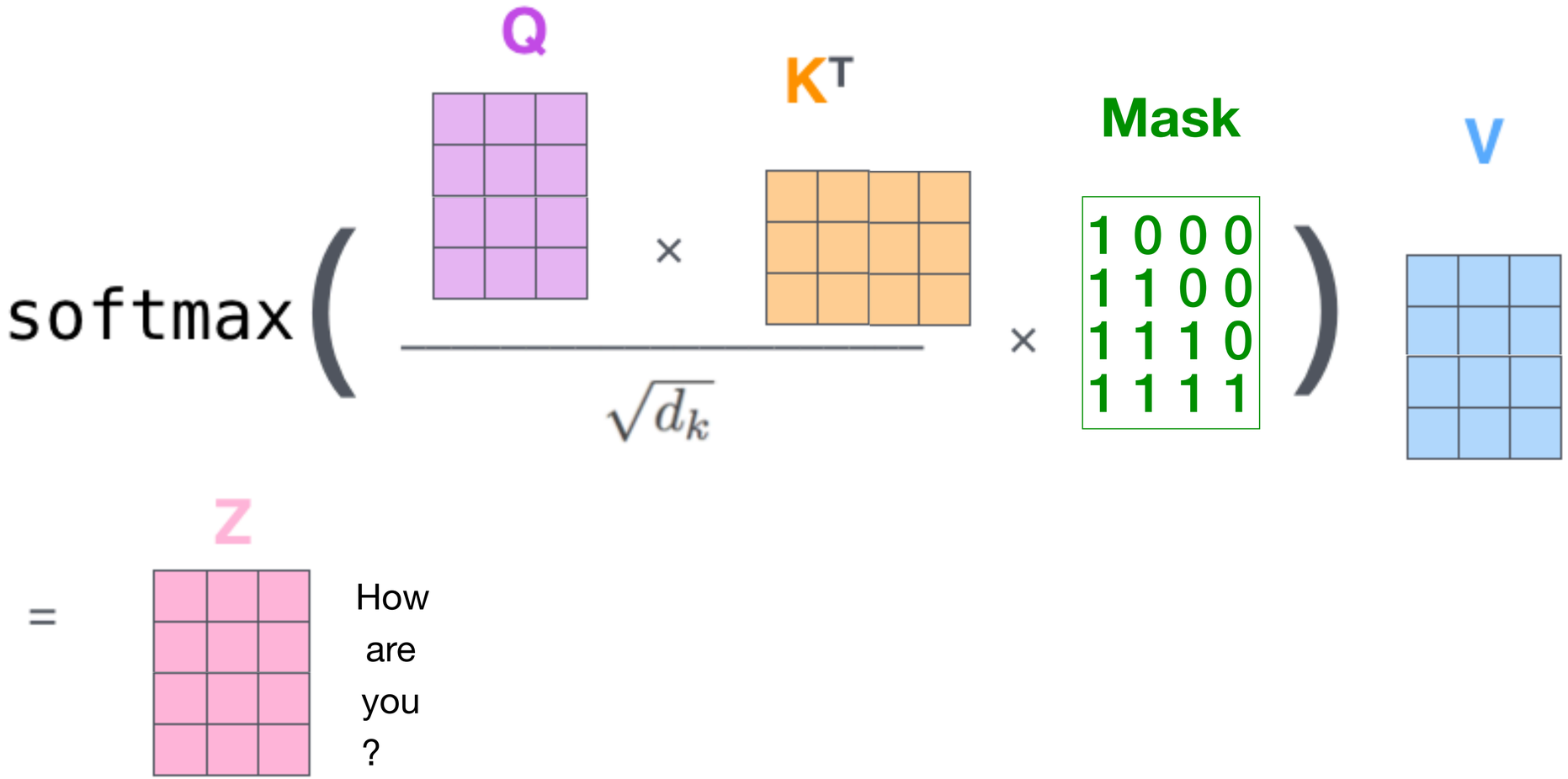


How
are
you
?

Encoder vs Decoder

Decoder (masked) self-attention:

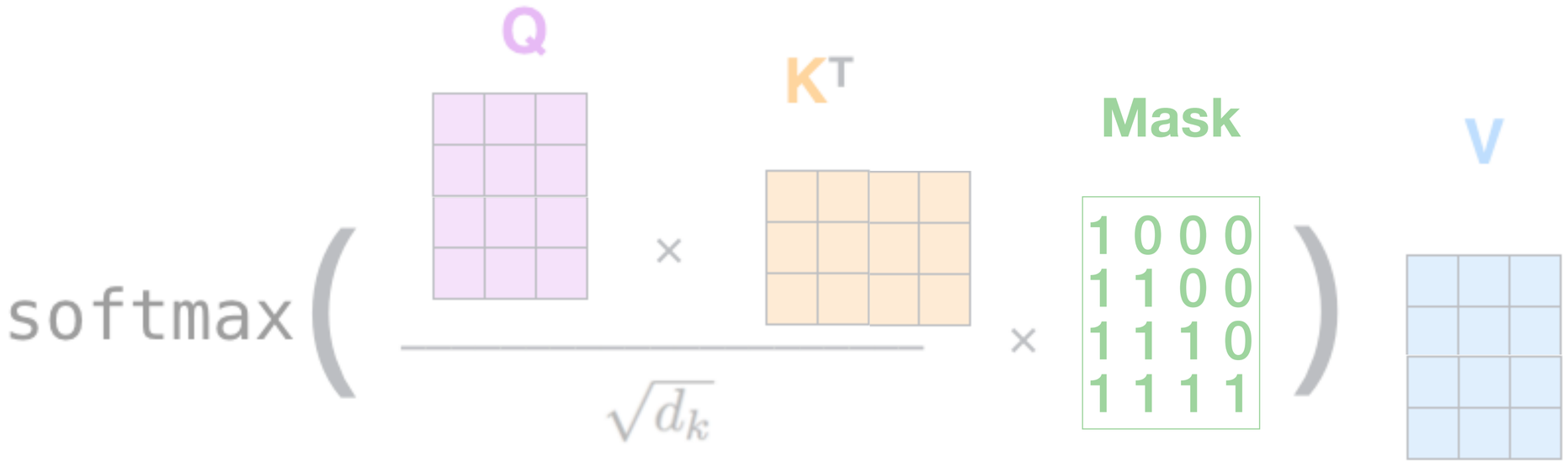
$$attention = softmax(\frac{QK^T + Mask}{d})V$$

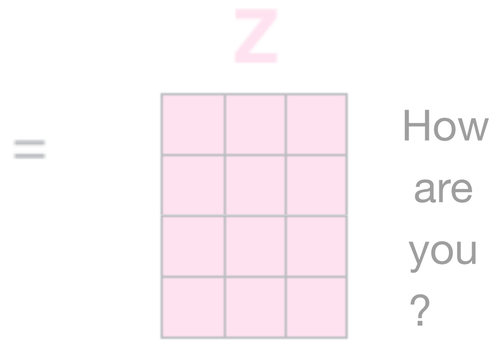


Encoder vs Decoder

Decoder (masked) self-attention:

$$attention = softmax(\frac{QK^T + Mask}{d})V$$



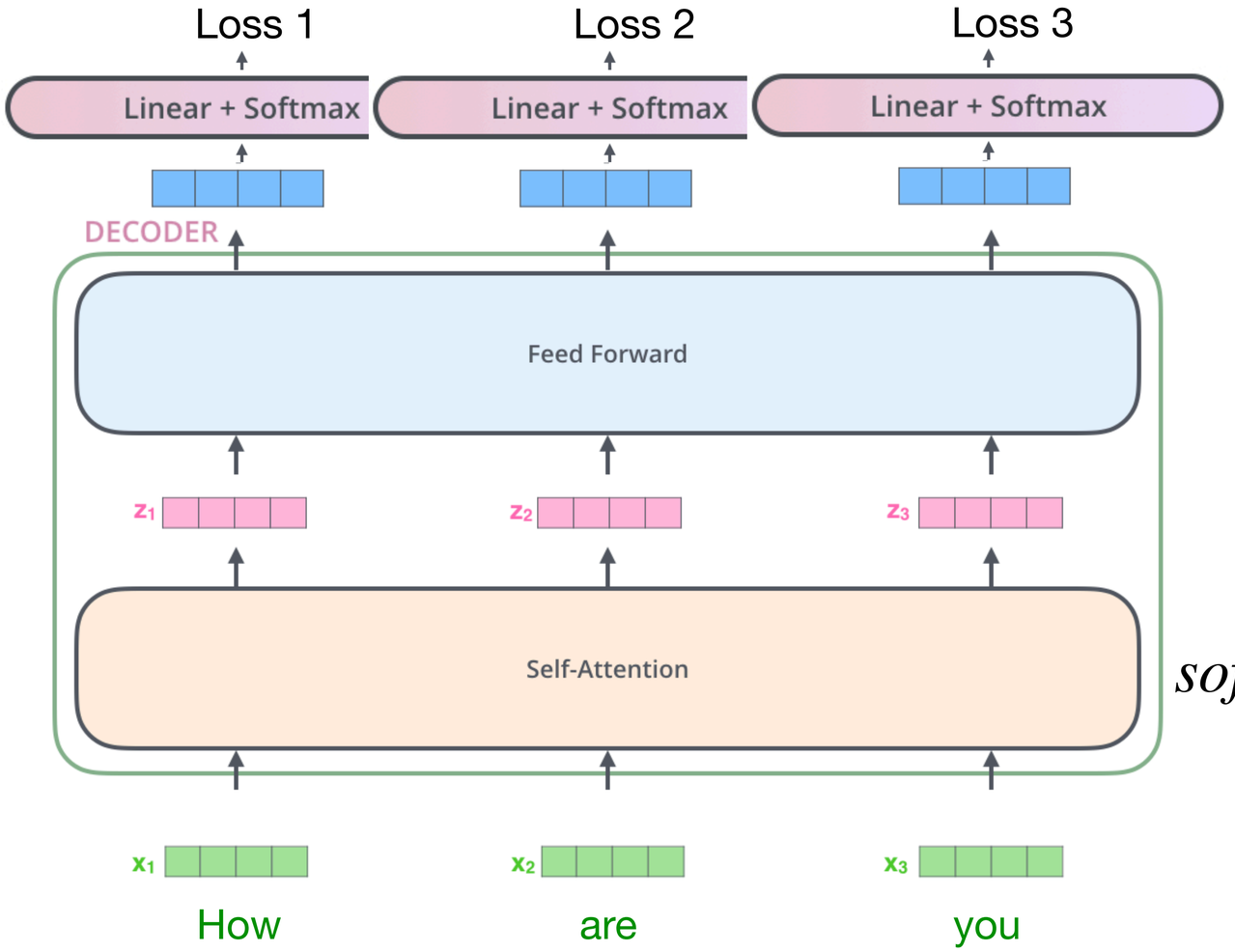


	How	are	you	?
How	1	0	0	0
are	1	1	0	0
you	1	1	1	0
?	1	1	1	1

Masked Attention не будет смотреть вперед

Encoder vs Decoder

Train



$$softmax(\frac{QK^T + Mask}{d})V$$

Decoder: можно посчитать одновременно

GPT

Generative **P**re-**T**raining

Архитектура: Transformer Decoder

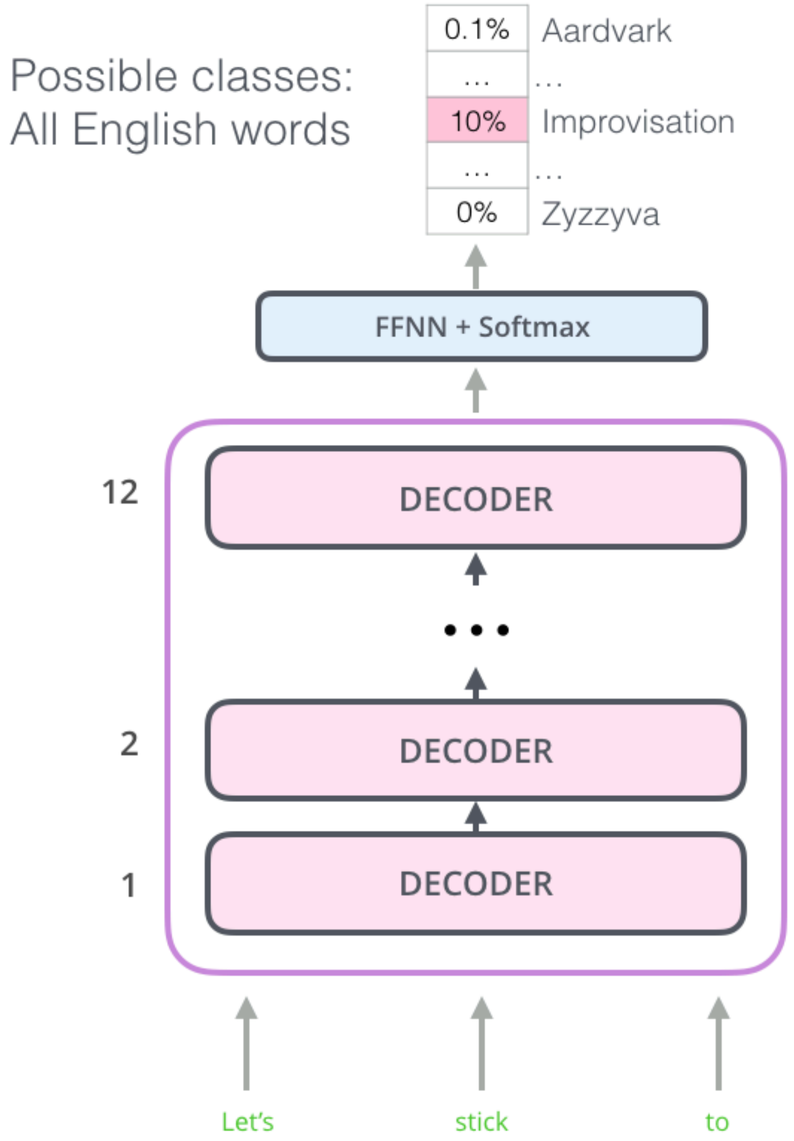


Image credit

GPT

Generative **P**re-**T**raining

Архитектура: Transformer Decoder

Данные: тексты книг (BookCorpus)

+ Разнообразие

+ Большой объем

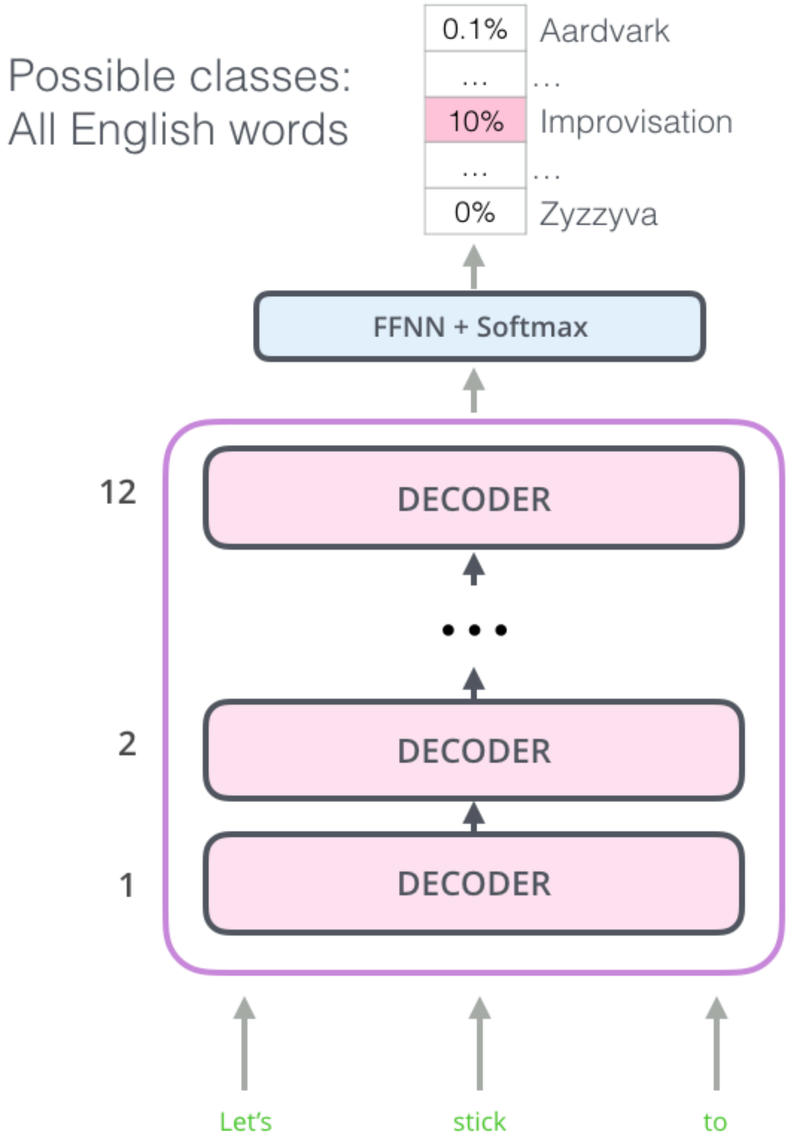


Image credit

GPT

Generative **P**re-**T**raining

Архитектура: Transformer Decoder

Данные: тексты книг (BookCorpus)

Задача для обучения: Language Modeling

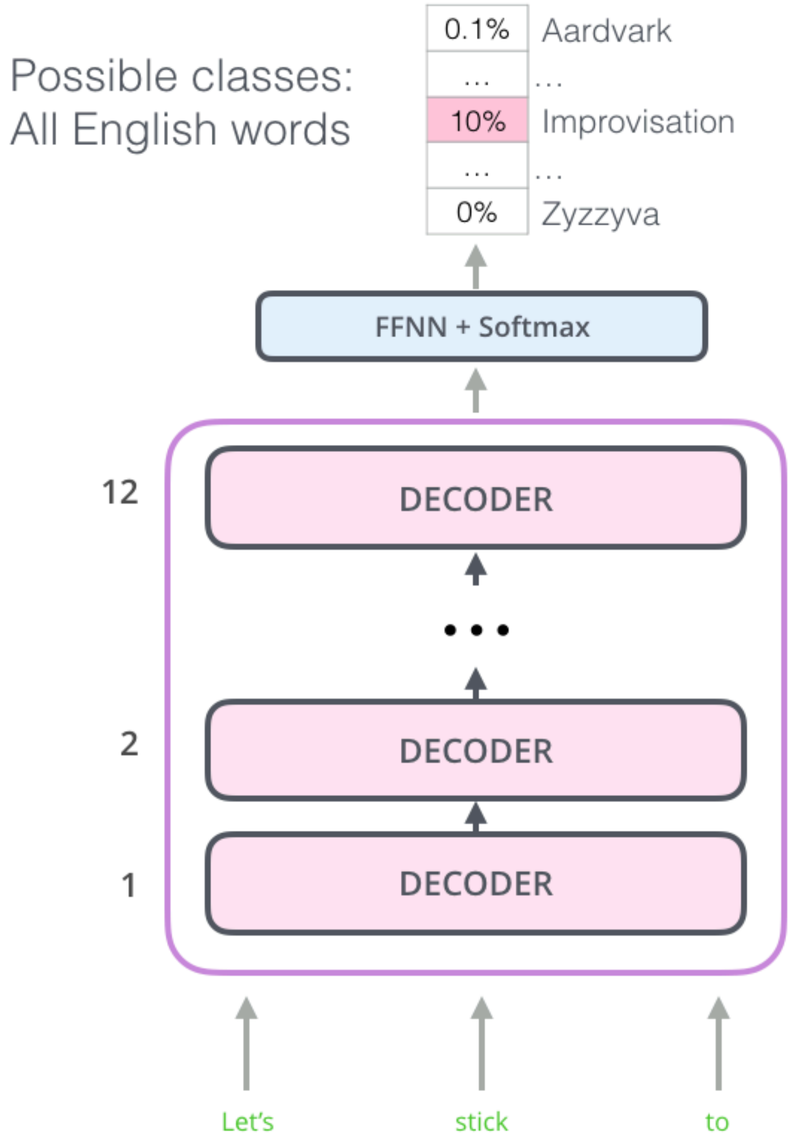


Image credit

GPT

Как использовать?

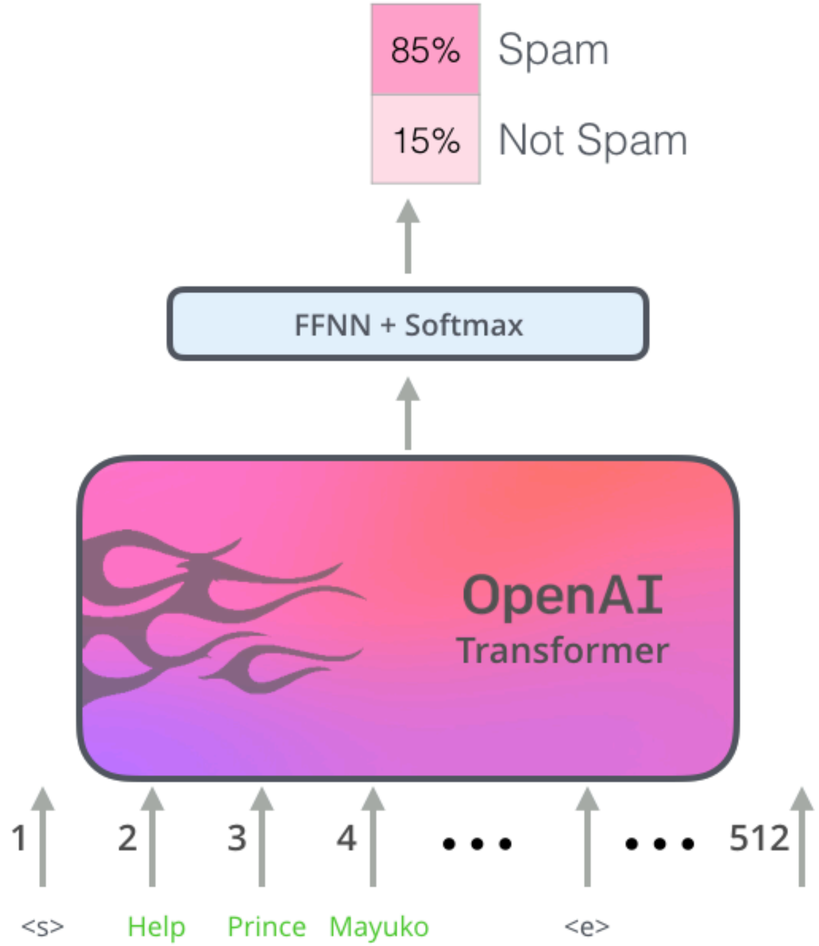
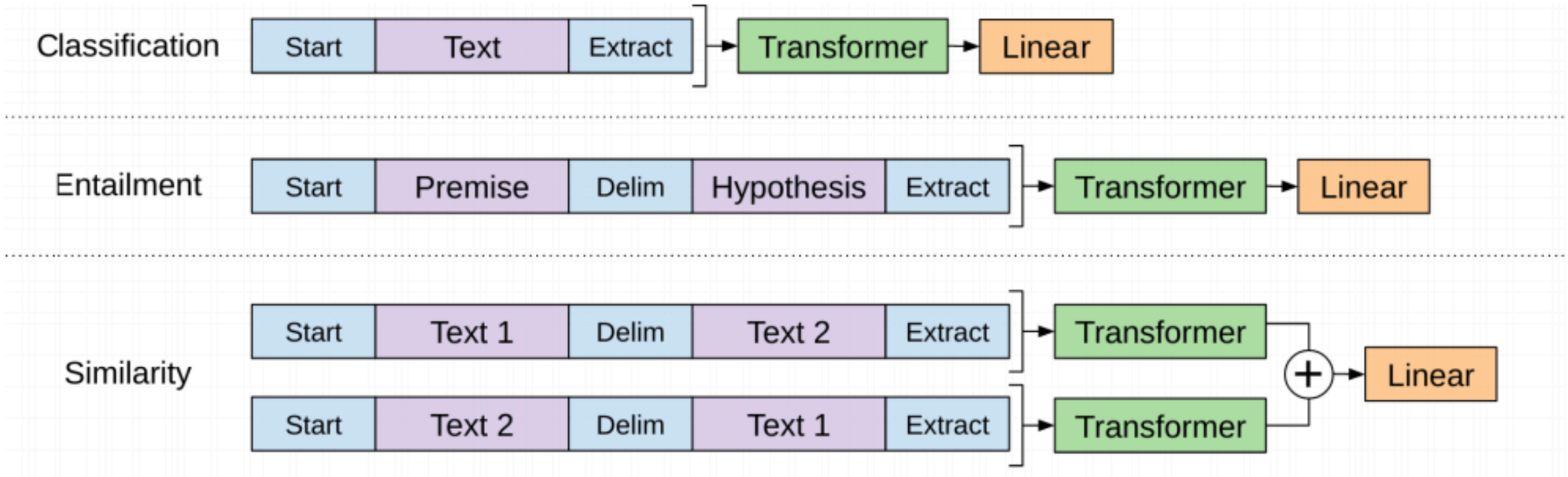


Image credit

GPT

Как использовать?

Для разных задач - разный формат входа



GPT-2

x10 параметров

x10 данных

GPT-2

Как использовать? **zero-shot task transfer**

Нет дообучения на новую задачу (fine-tuning)

Форматируем input, чтобы была понятна задача

GPT-2

Форматируем input, чтобы была понятна задача

Пример: задача суммаризации

Article: Amina Ali Qassim is sitting with her youngest grandchild on her lap, wiping away tears with her headscarf. Only a few months old, this is the baby girl whose ears she desperately tried to cover the night the aerial bombardment started. She lay awake, she says, in a village mosque on the Yemeni island of Birim, counting explosions as the baby cried.
It could have been worse though. They could have still been in their house when the first missile landed.
"Our neighbor shouted to my husband 'you have to leave, they're coming.' And we just ran. As soon as we left the house, the first missile fell right by it and then a second on it. It burned everything to the ground," Qassim tells us
...

TL;DR:

Input

GPT-2

Форматируем input, чтобы была понятна задача

Пример: задача суммаризации

Article: Amina Ali Qassim is sitting with her youngest grandchild on her lap, wiping away tears with her headscarf. Only a few months old, this is the baby girl whose ears she desperately tried to cover the night the aerial bombardment started. She lay awake, she says, in a village mosque on the Yemeni island of Birim, counting explosions as the baby cried.
It could have been worse though. They could have still been in their house when the first missile landed.
"Our neighbor shouted to my husband 'you have to leave, they're coming.' And we just ran. As soon as we left the house, the first missile fell right by it and then a second on it. It burned everything to the ground," Qassim tells us
...

Input

TL;DR: Yemen is in the middle of a civil war. Saudi Arabia is leading the coalition bombing campaign. It's been bombing Yemen for more than two months now.

GPT-2 prediction

GPT-2

Language Modeling (генерация текста)

Context (human-written): In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.
GPT-2: The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science. Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved. Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

GPT-3

x100 параметров (по сравнению с GPT-2)

x5 данных (по сравнению с GPT-2)

* нет в свободном доступе

GPT-3

x100 параметров (по сравнению с GPT-2)

x5 данных (по сравнению с GPT-2)

Как использовать? **zero-shot/one-shot/few-shot settings**

- показываем 0/1/несколько примеров (promts)

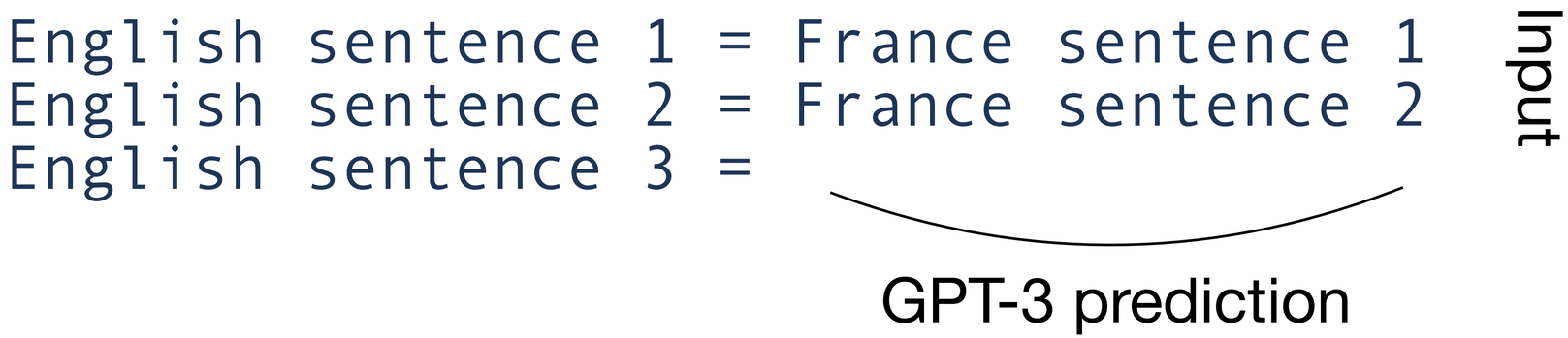
GPT-3

x100 параметров (по сравнению с GPT-2)

x5 данных (по сравнению с GPT-2)

Как использовать? **zero-shot/one-shot/few-shot settings**

- показываем 0/1/несколько примеров (promts)



GPT-3

Prompt:

```
The following are ideas for startups, apps, marketplaces, platforms:
###
A startup that helps content creators, such as photographers and
videographers, monetize their work.
###
An online forum that helps people find roommates
###
```

GPT-3:

```
A community hub for entrepreneurs to share ideas and ask questions of
experienced entrepreneurs
###
A website that helps people discover new interests and hobbies.
###
A company that helps brands advertise their products to potential buyers on
Instagram.
###
```

Image credit

BERT

BERT

Bidirectional **E**ncoder **R**epresentations from **T**ransformers

Архитектура: Transformer Encoder

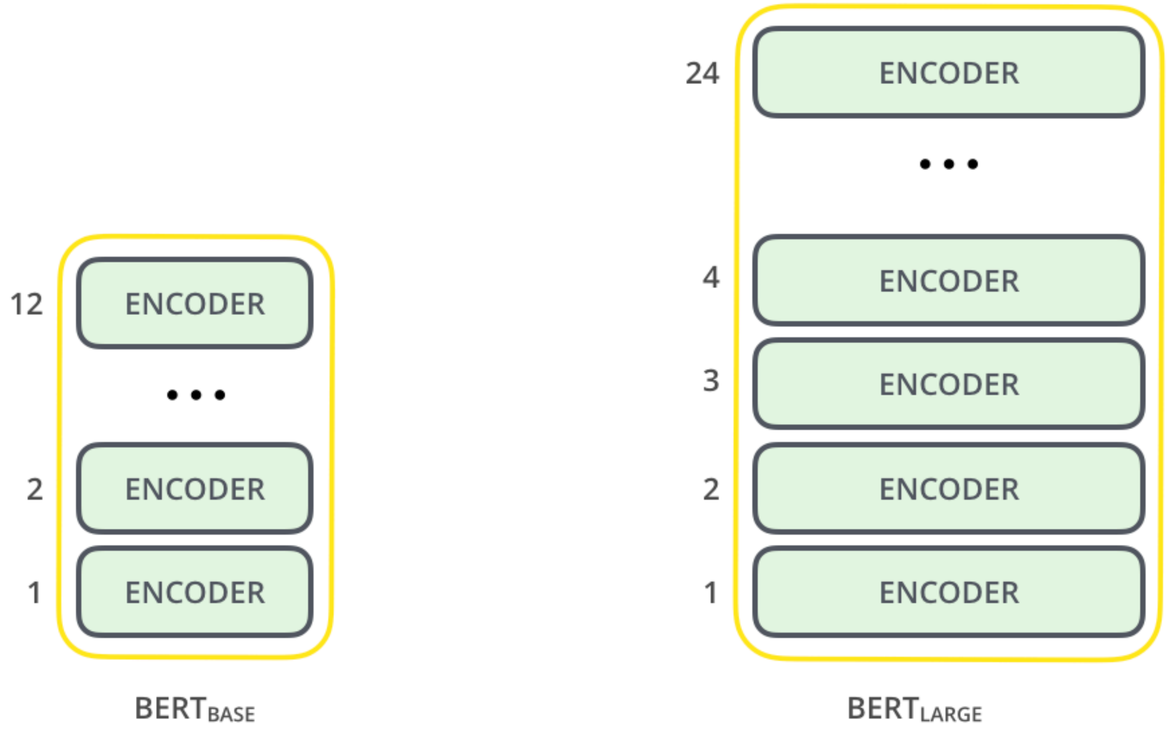


Image credit

BERT

Bidirectional **E**ncoder **R**epresentations from **T**ransformers

Архитектура: Transformer Encoder

Данные: Wikipedia и тексты книг (BookCorpus)

Image credit

BERT

Bidirectional **E**ncoder **R**epresentations from **T**ransformers

Архитектура: Transformer Encoder

Данные: Wikipedia и тексты книг (BookCorpus)

Задачи для обучения: Masked LM и Next Sentence Prediction

Image credit

BERT

Masked Language Model

Случайно выбираем 15% позиций и заменяем на <MASK>

Задача: предсказать исходный токен

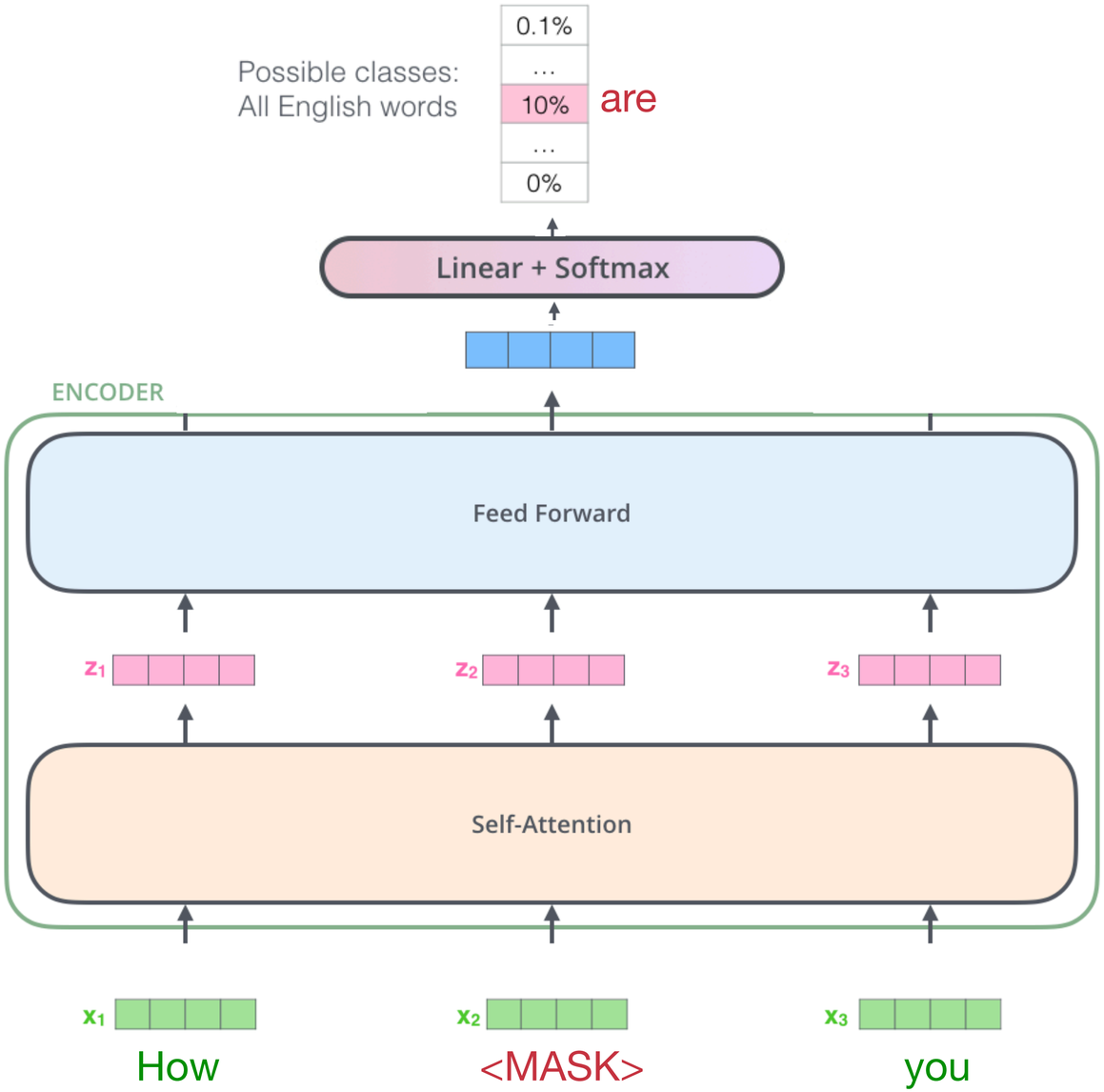


Image credit

BERT

Masked Language Model

Случайно выбираем 15% позиций:

- 80% заменяем на <MASK>
- 10% заменяем на случайный токен
- 10% оставляем

Задача: предсказать исходный токен

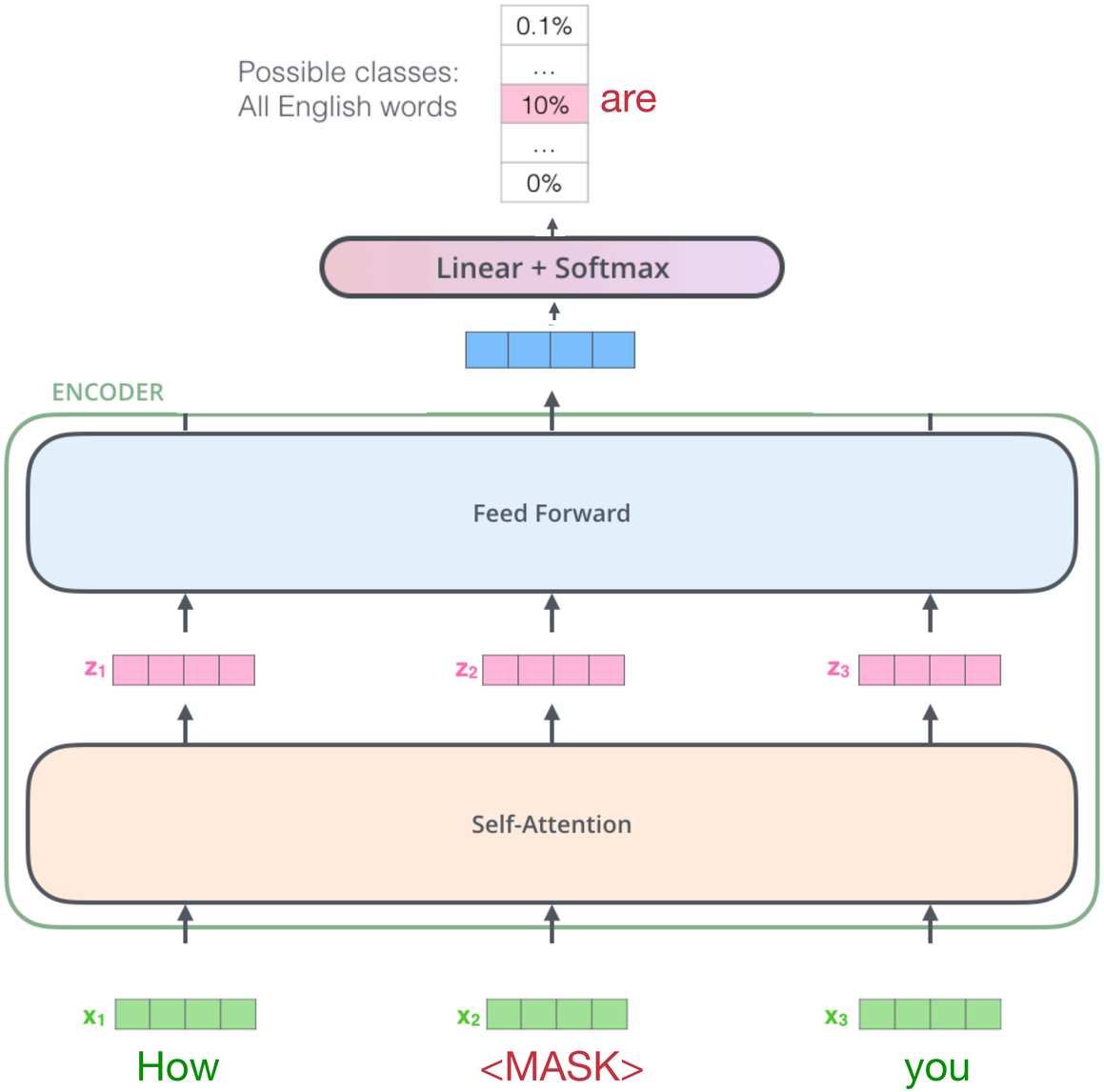


Image credit

BERT

Masked Language Model

Случайно выбираем 15% позиций:

- 80% заменяем на <MASK>
- 10% заменяем на случайный токен
- 10% оставляем

Задача: предсказать исходный токен

Обученные эмбеддинги учитывают контекст слева и справа

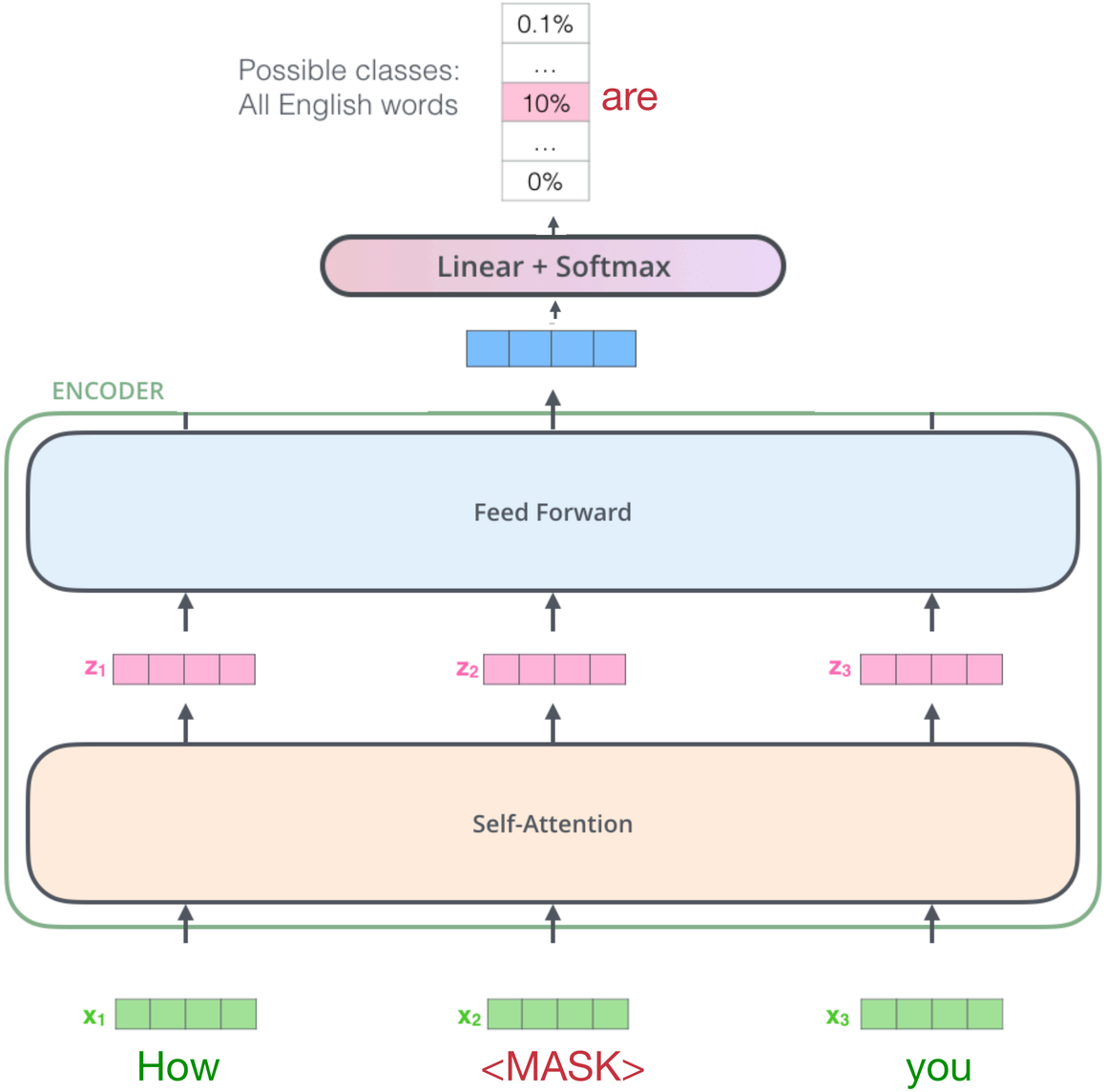


Image credit

BERT

Next Sentence Prediction

Для некоторых задач нужно понимать взаимоотношения между двумя предложениями:

- Similarity
- Entailment
- Question Answering
- …

BERT

Next Sentence Prediction

Вход: 2 предложения (A и B)

50% - B следует за A в тексте
50% - B выбрано случайно

Формат входа:

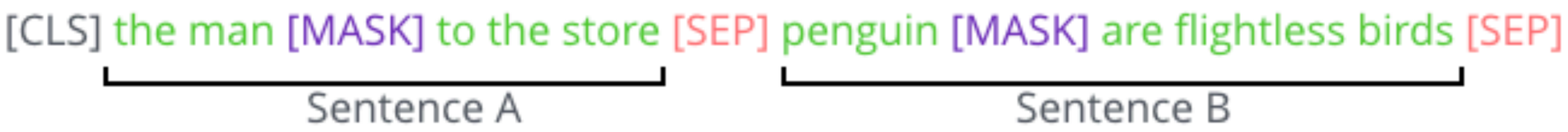


Image credit

BERT

Next Sentence Prediction

Вход: 2 предложения (A и B)

50% - B следует за A в тексте
50% - B выбрано случайно

Задача: предсказать, следует ли B за A

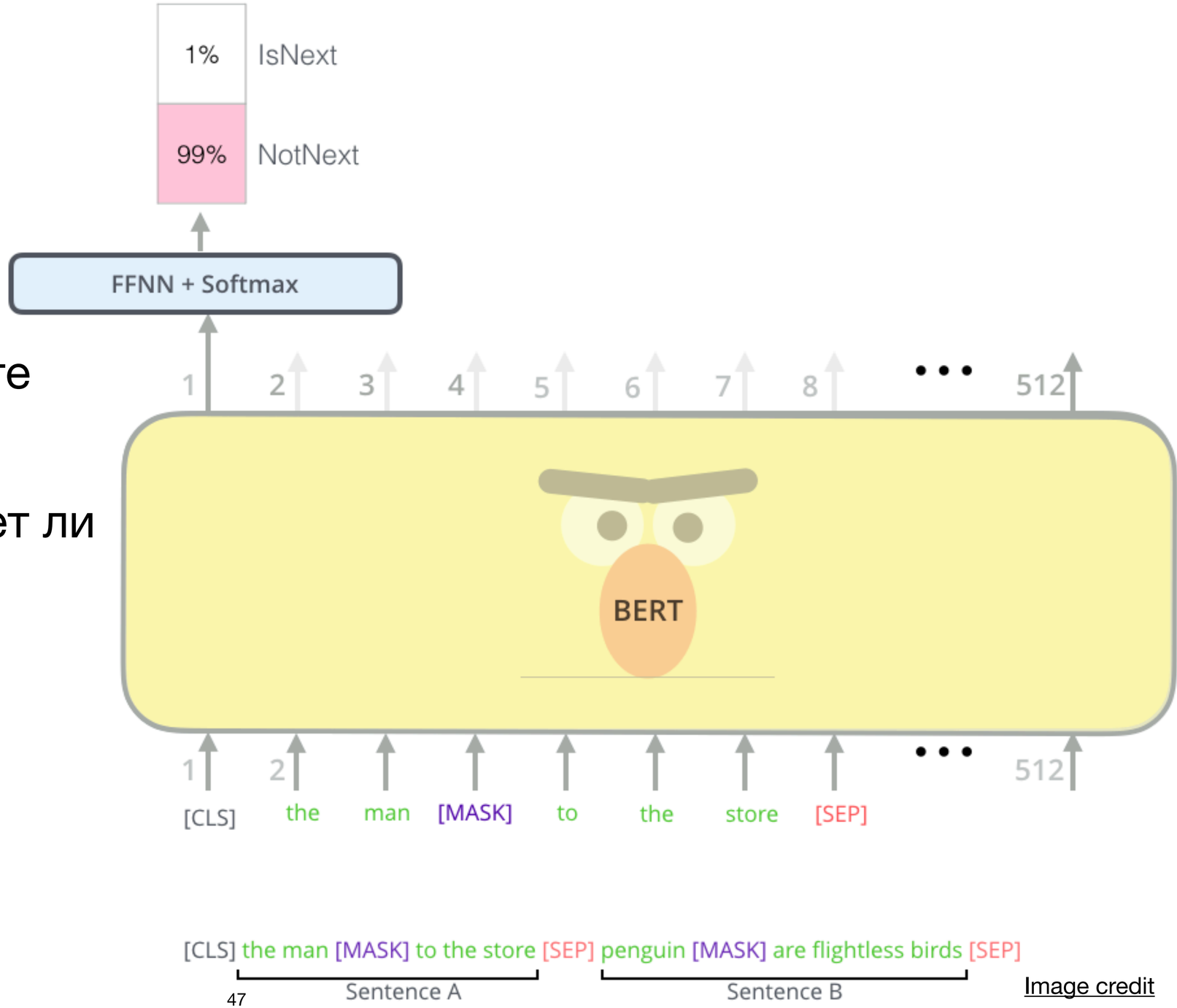


Image credit

BERT

Next Sentence Prediction

Вход: 2 предложения (A и B)

50% - B следует за A в тексте
50% - B выбрано случайно

Задача: предсказать, следует ли B за A

<CLS> - выучивает агрегированную информацию

<SEP> - разделитель

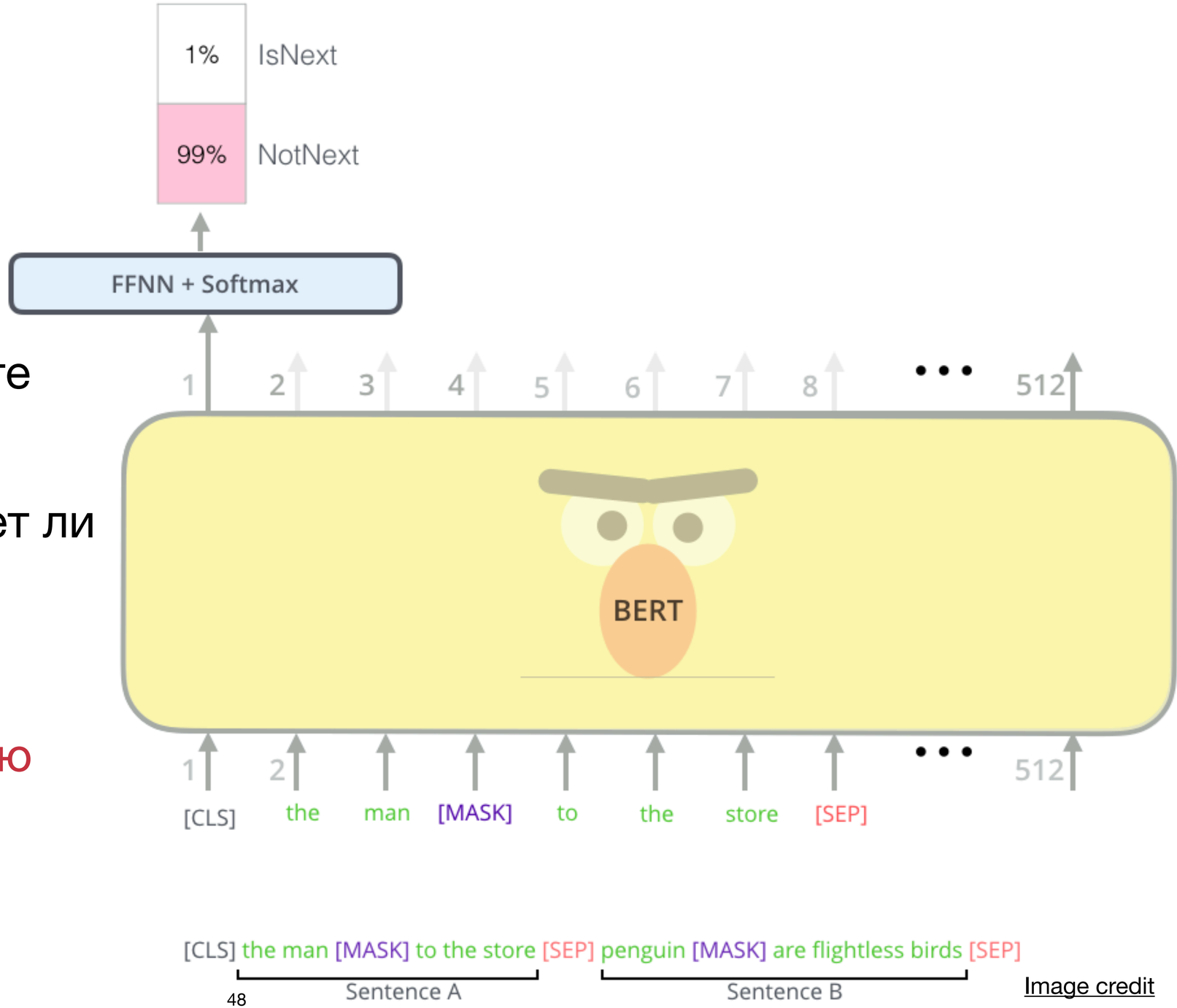


Image credit

BERT

Как использовать?

- Linear+Softmax поверх <CLS> - для задач классификации предложения (или двух)
- Linear+Softmax поверх всех выходов - для задач классификации токенов
- Выходы BERT - как вход в другие модели (task-specific)
- …

RoBERTa

Robustly **O**ptimized **BERT** Pretraining **A**pproach

Улучшенная версия BERT:

- x10 данных
- x10 вычислительных ресурсов (дольше обучение, больше batch size)
- Не использовали задачу Next Sentence Prediction

Улучшение в 2-20% (в зависимости от задачи)

BART and T5

BART

Bidirectional and **A**uto-**R**egressive **T**ransformers

Идея: соединить преимущества BERT и GPT

Архитектура: Transformer Encoder-Decoder

Данные: как в RoBERTa

BART

Bidirectional and **A**uto-**R**egressive **T**ransformers

Идея: соединить преимущества BERT и GPT

Архитектура: Transformer Encoder-Decoder

Данные: как в RoBERTa

Задача для обучения: восстановить последовательность

Последовательность: ABCDE

- Маскируем C, D
- Шум: лишняя маска перед B

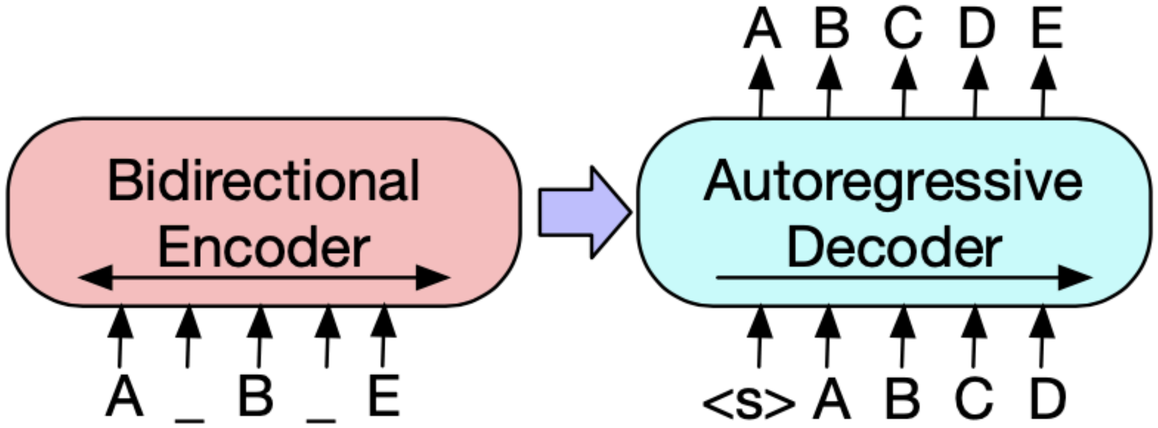


Image credit

T5

Text-**t**o-**T**ext **T**ransfer **T**ransformer

Идея: соединить преимущества BERT и GPT

Архитектура: Transformer Encoder-Decoder

Данные: x5 от данных RoBERTa

Задача для обучения: восстановить последовательность

T5

Text-**t**o-**T**ext **T**ransfer **T**ransformer

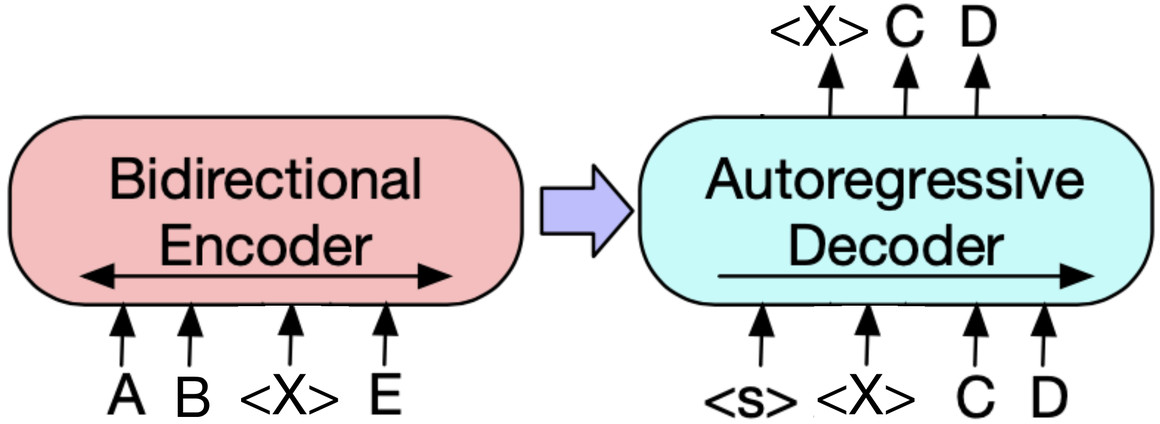
Идея: соединить преимущества BERT и GPT

Архитектура: Transformer Encoder-Decoder

Данные: x5 от данных RoBERTa

Задача для обучения: восстановить последовательность

Вместо маски используем специальные токены



Сравнение

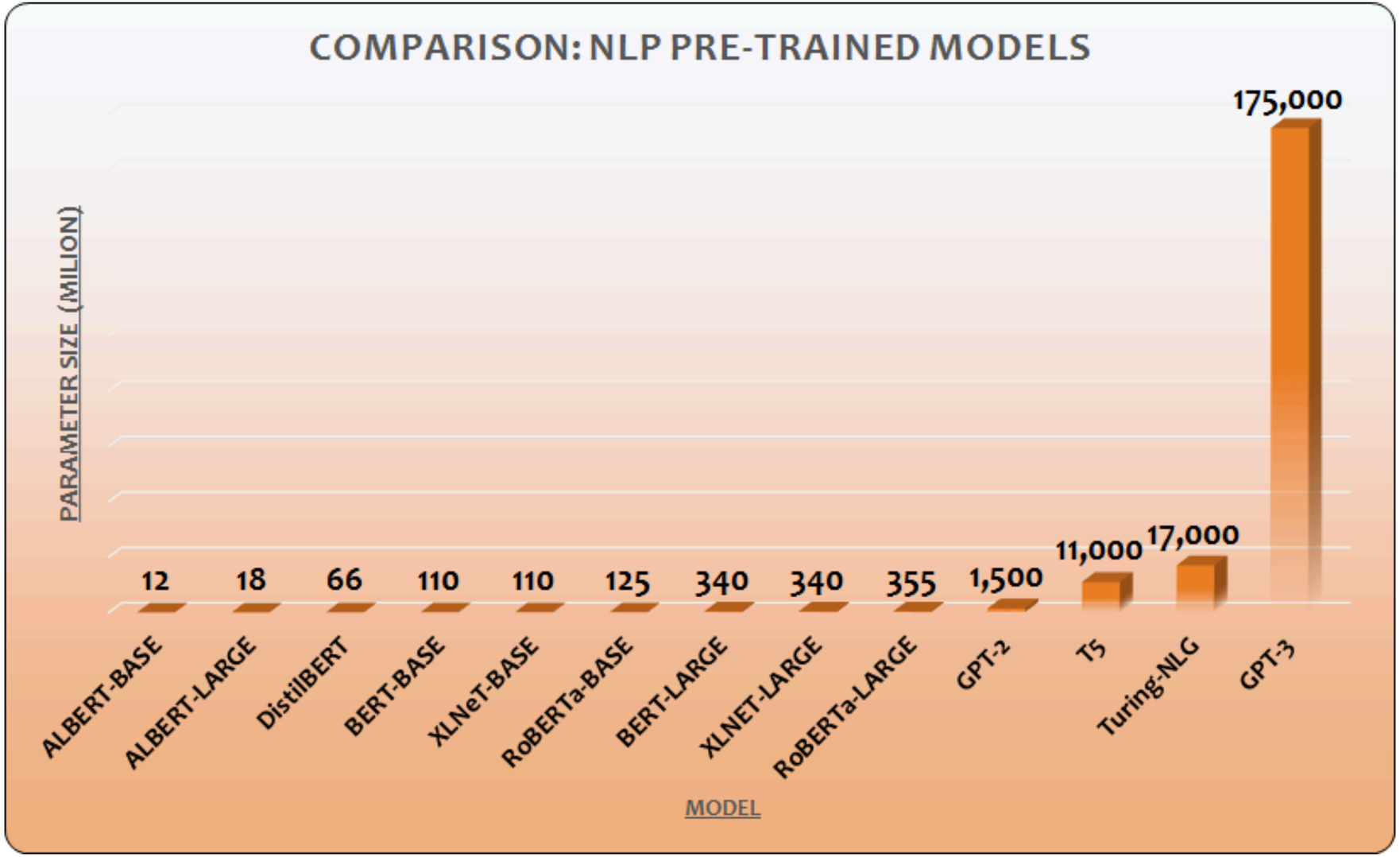


Image credit

Сравнение

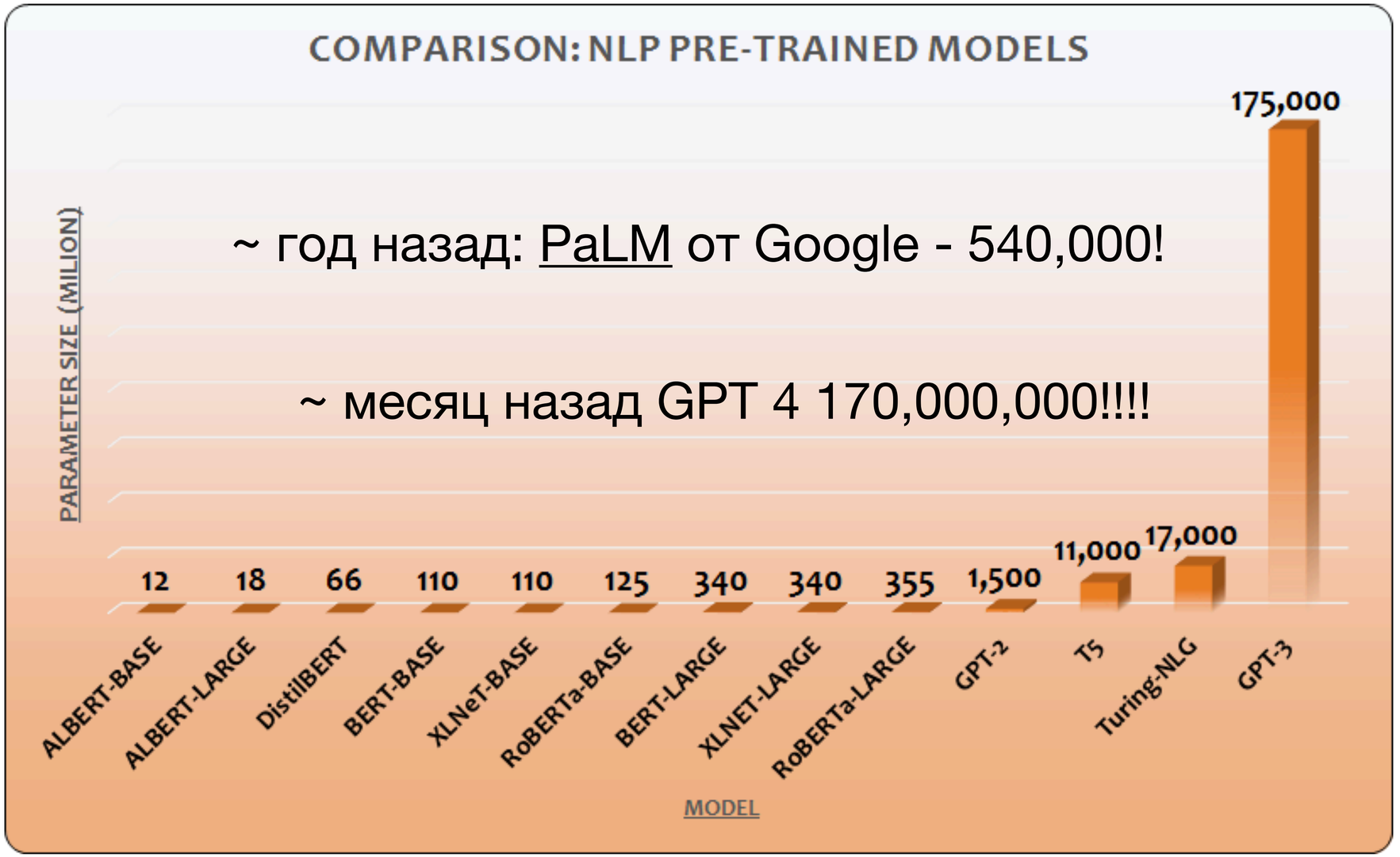


Image credit

Как использовать?

Большой список доступных моделей и удобный интерфейс - библиотека HuggingFace Transformers 🤗

Как использовать?

Большой список доступных моделей и удобный интерфейс - библиотека HuggingFace Transformers 🤗

- GPT-3.5 - хорошая генерация текста
- BERT, RoBERTa - классификация, использование в качестве эмбеддингов
- T5, BART - seq2seq задачи

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

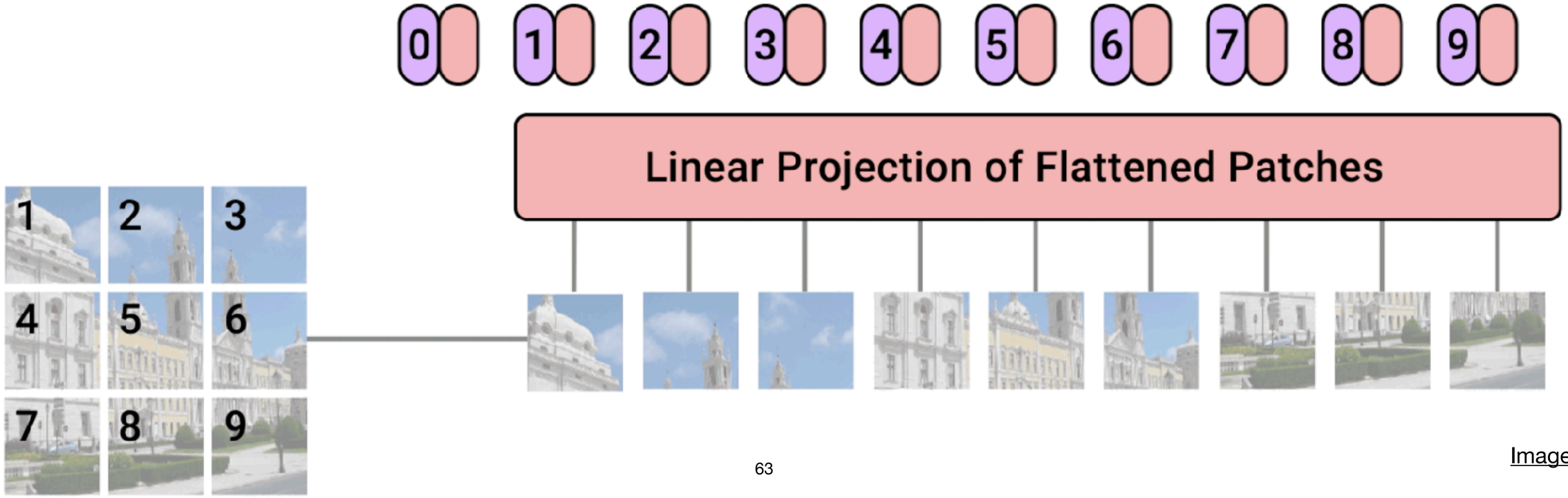


Image credit

Vision Transformer

Как из картинки сделать последовательность?

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

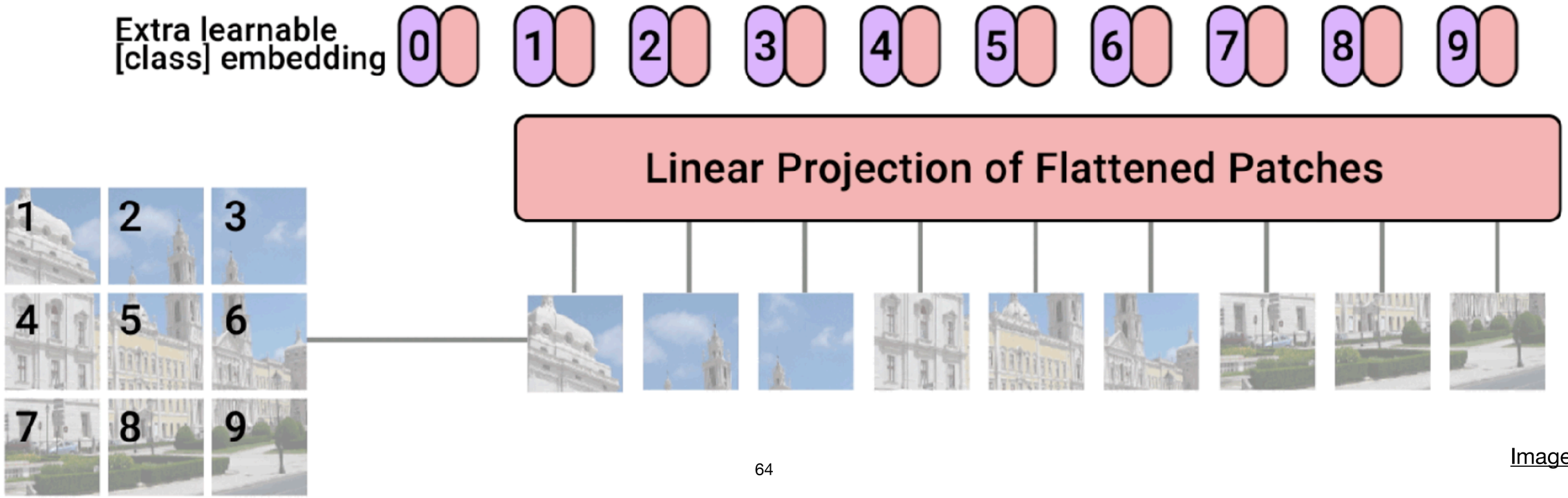


Image credit

Vision Transformer

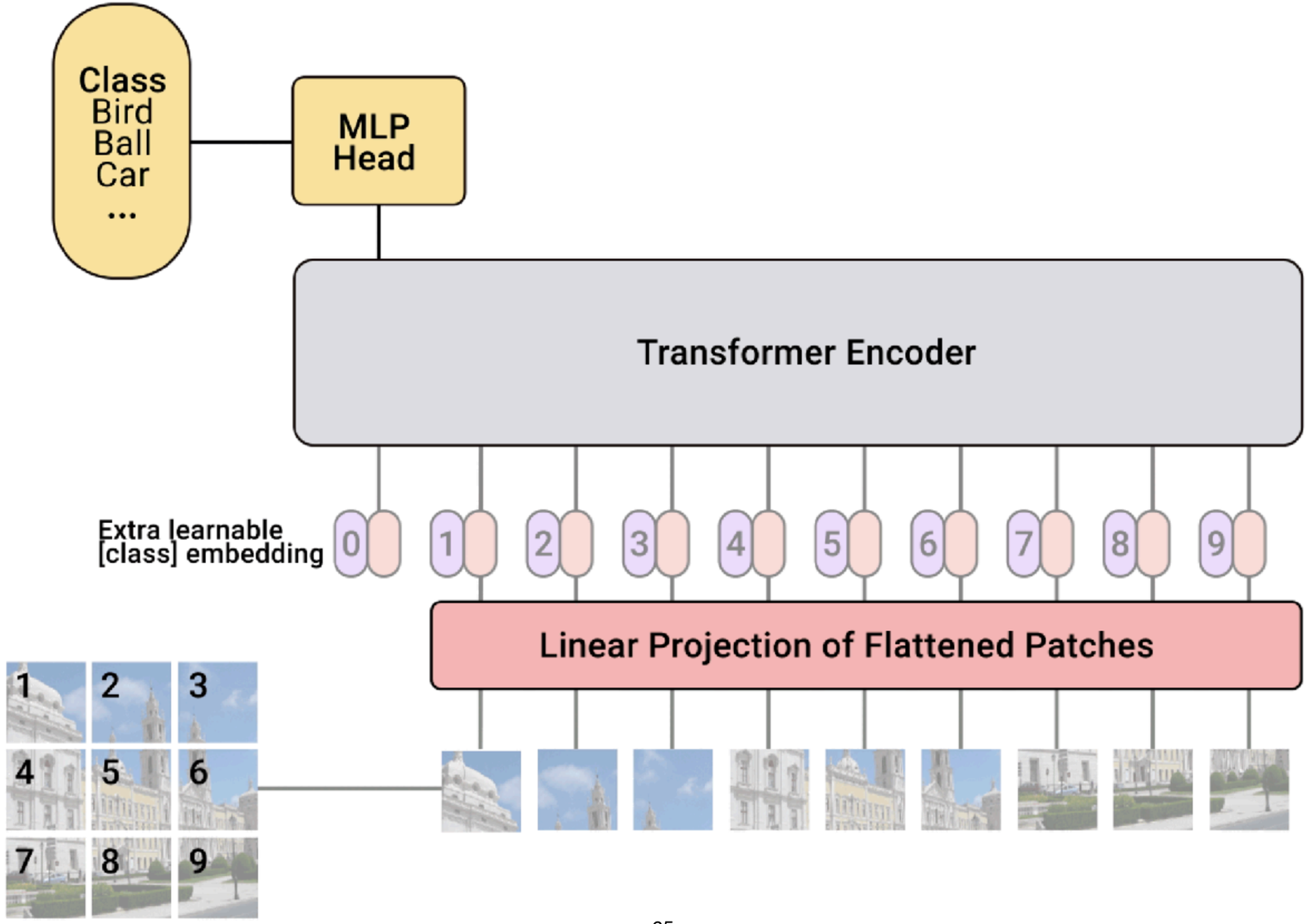


Image credit

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

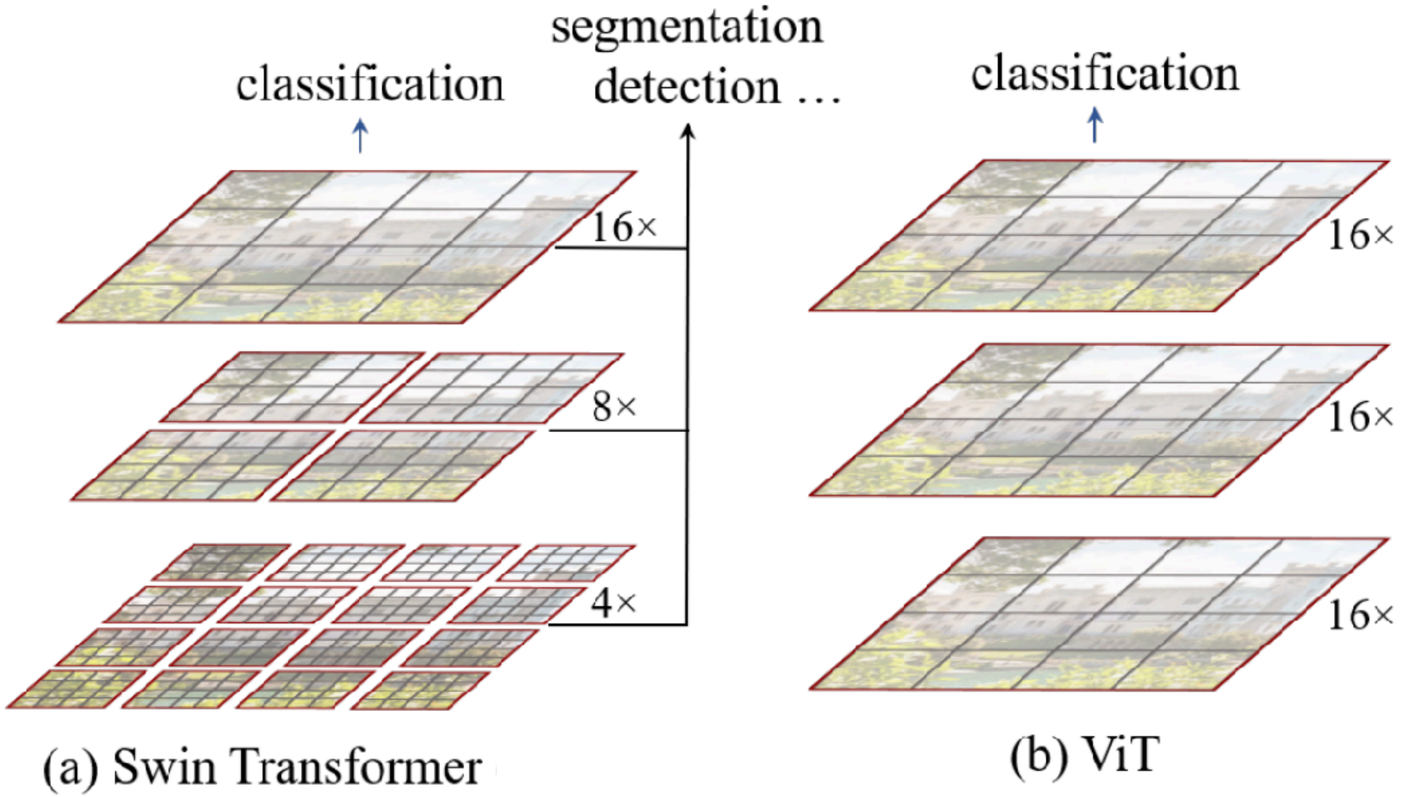


Image credit

Swin Transformer

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

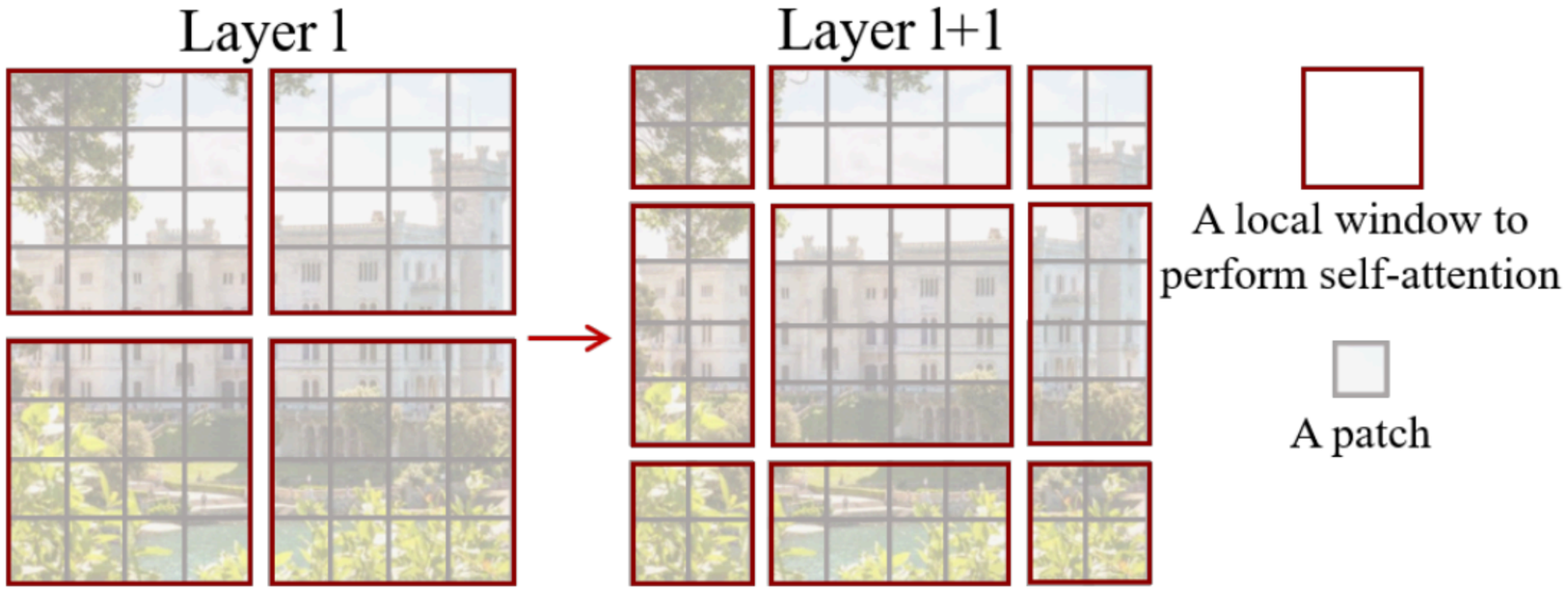


Image credit

CLIP: Contrastive Language–Image Pre-training

CLIP (Contrastive Language–Image Pre-training)

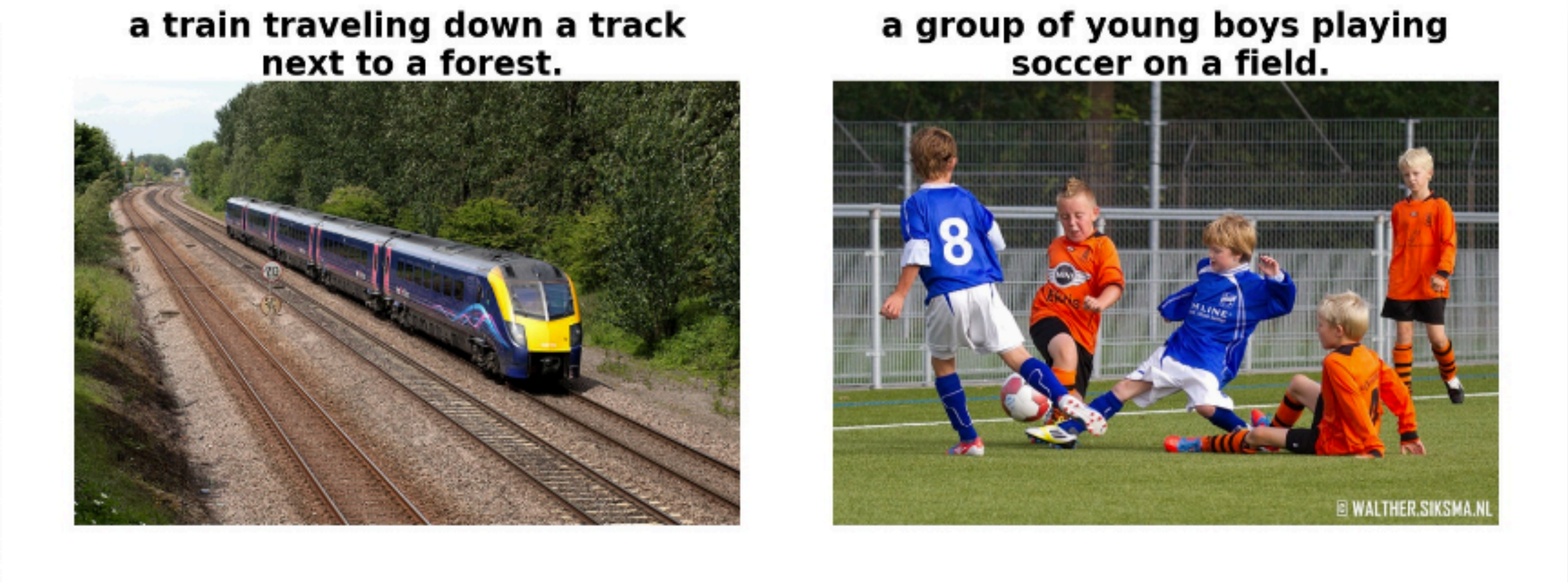
BERT-like модели предобучаются **на неразмеченных данных** (тексты)

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)

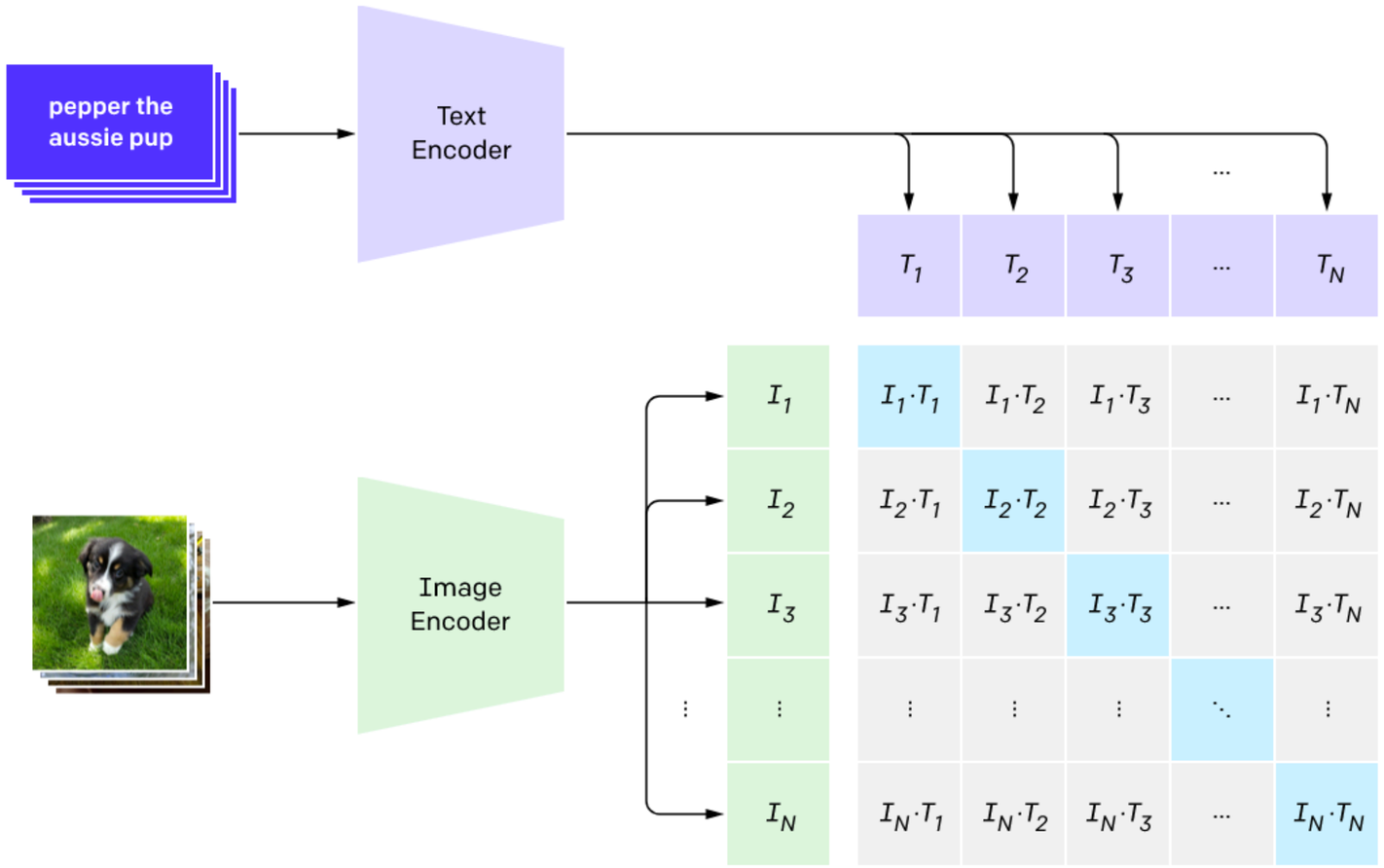
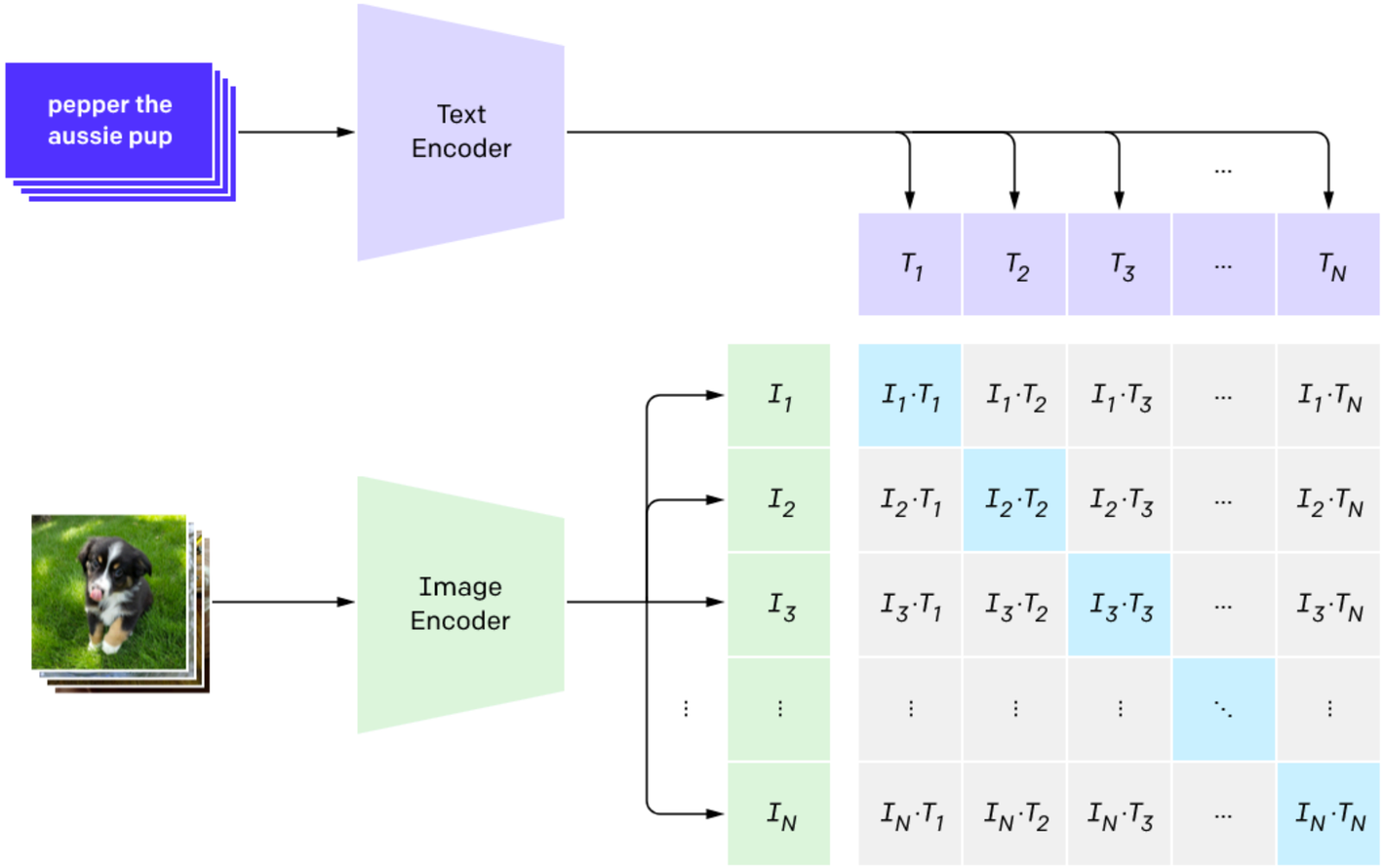


Image credit

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 …”

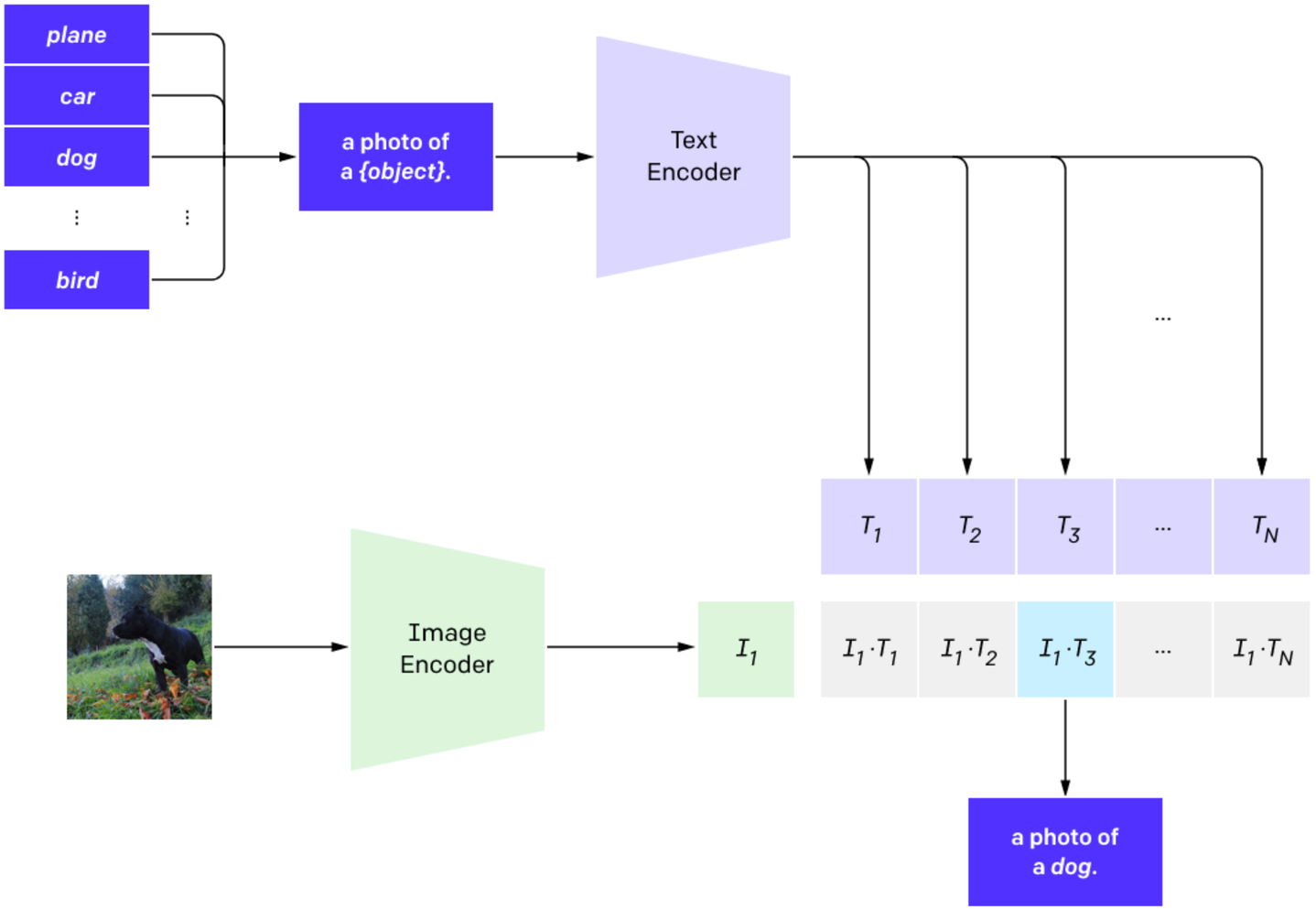


Image credit

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)

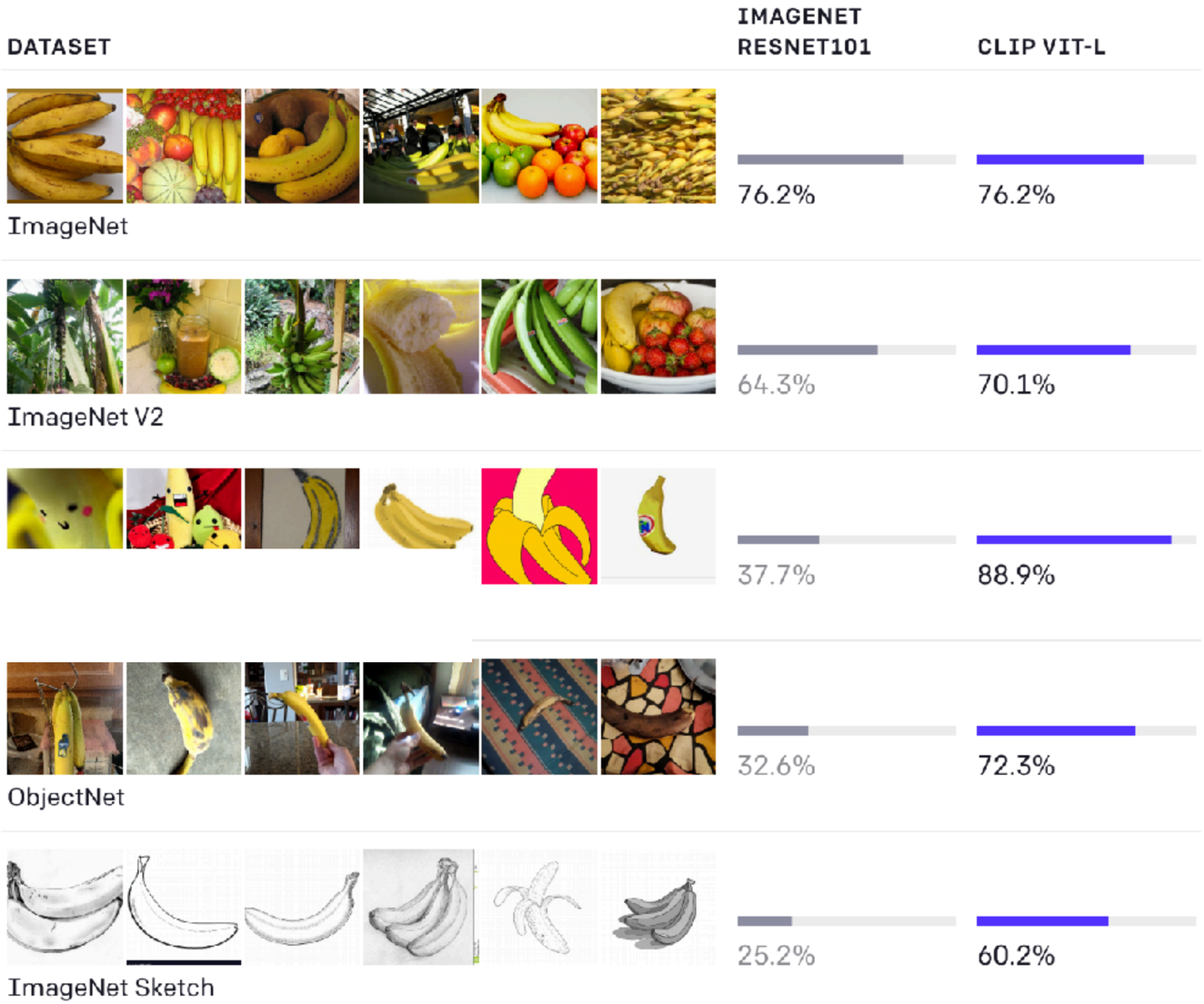


Image credit

CLIP (Contrastive Language–Image Pre-training)

- Чувствителен к описанию

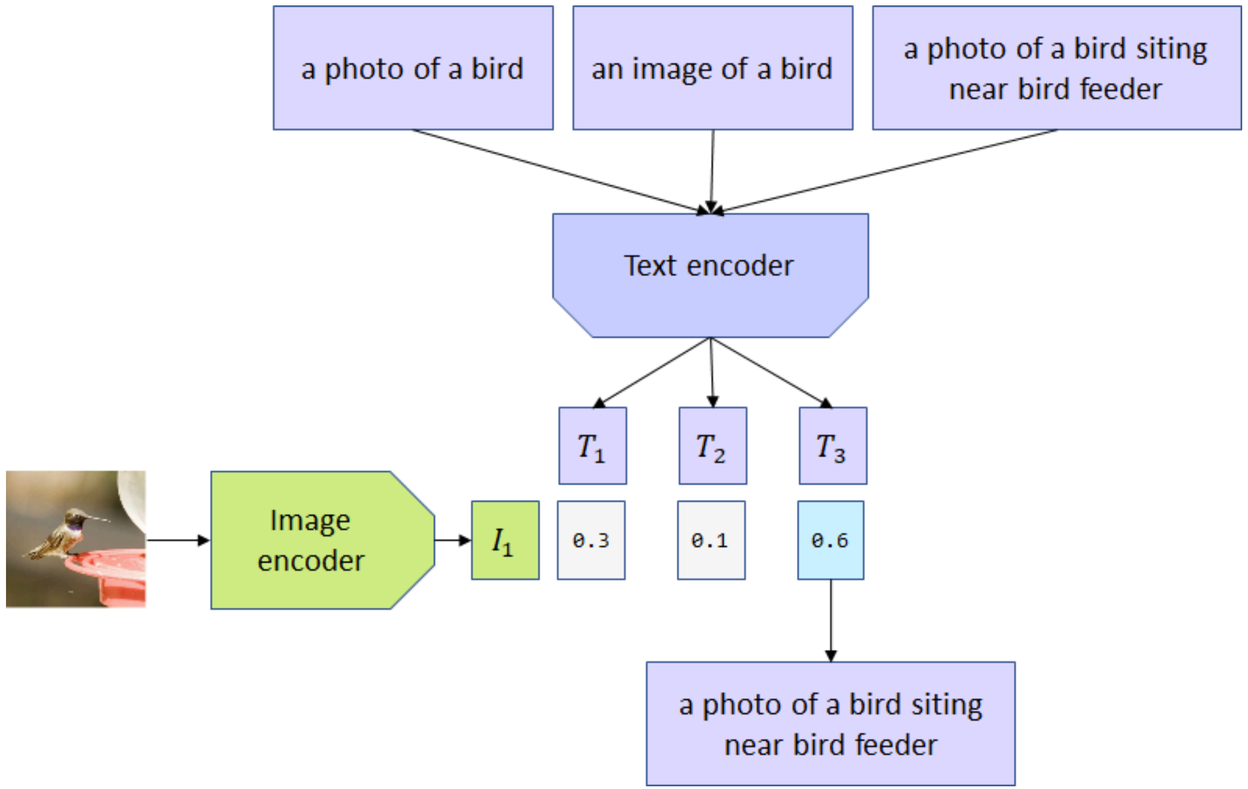


Image credit

CLIP + StyleGAN

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

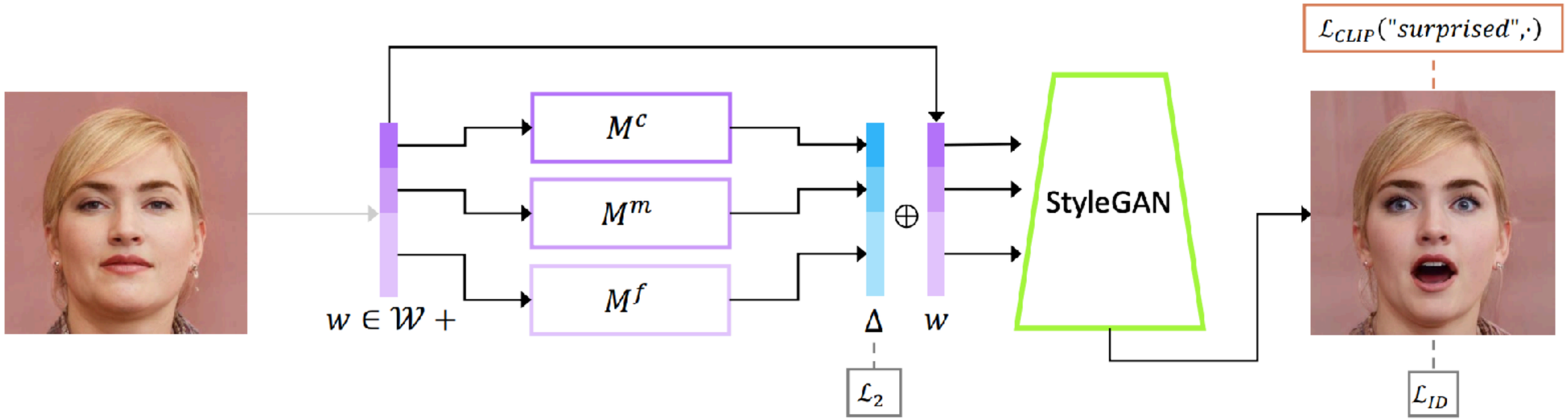


Image credit

CLIP + StyleGAN

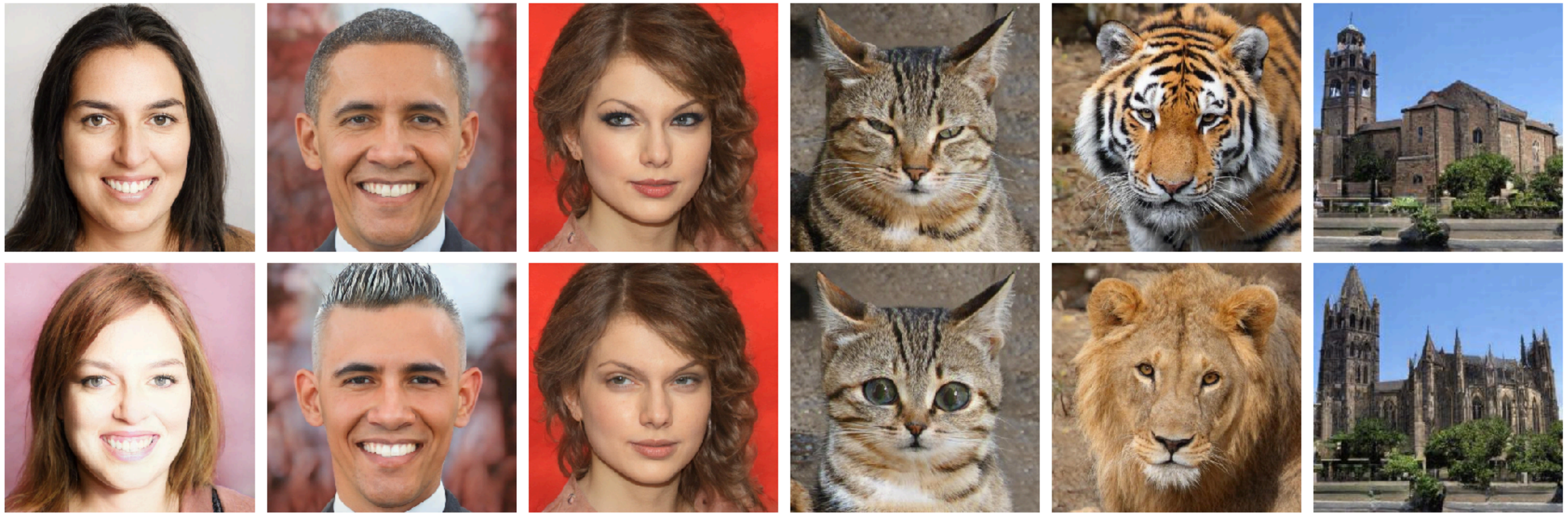


Image credit

DALL-E 2

Модель: CLIP + GLIDE (Denoising Diffusion Probabilistic Model)

a dolphin in an astronaut suit on saturn, artstation

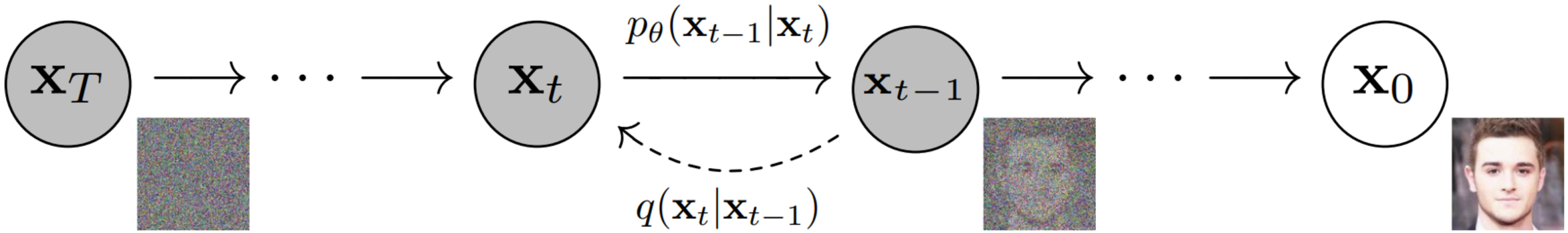
a propaganda poster depicting a cat dressed as french emperor napoleon holding a piece of cheese

a teddy bear on a skateboard in times square

Image credit

DDPM: overview

DDPM:

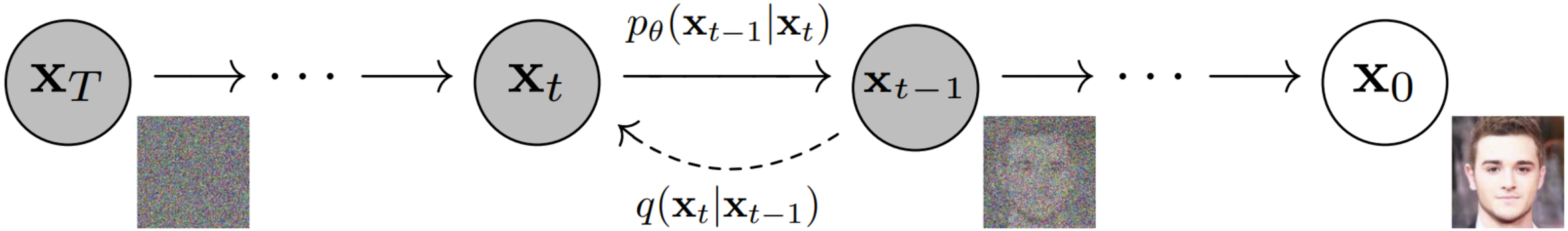


Добавляем случайный шум на каждом шаге

Image credit

DDPM: overview

DDPM:



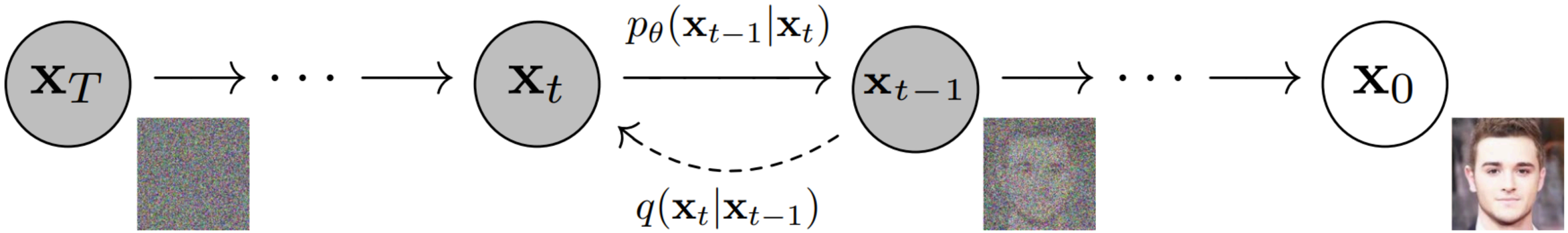
Восстанавливаем (убираем добавленный шум)

Image credit

DDPM: overview

DDPM:

Markov chain



Восстанавливаем (убираем добавленный шум)

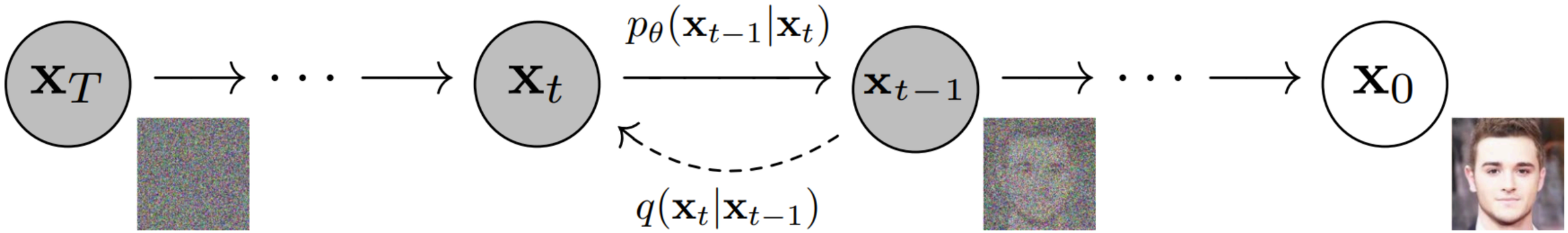
Каждый раз применяем модель типа UNet

Image credit

DDPM: overview

DDPM:

Markov chain

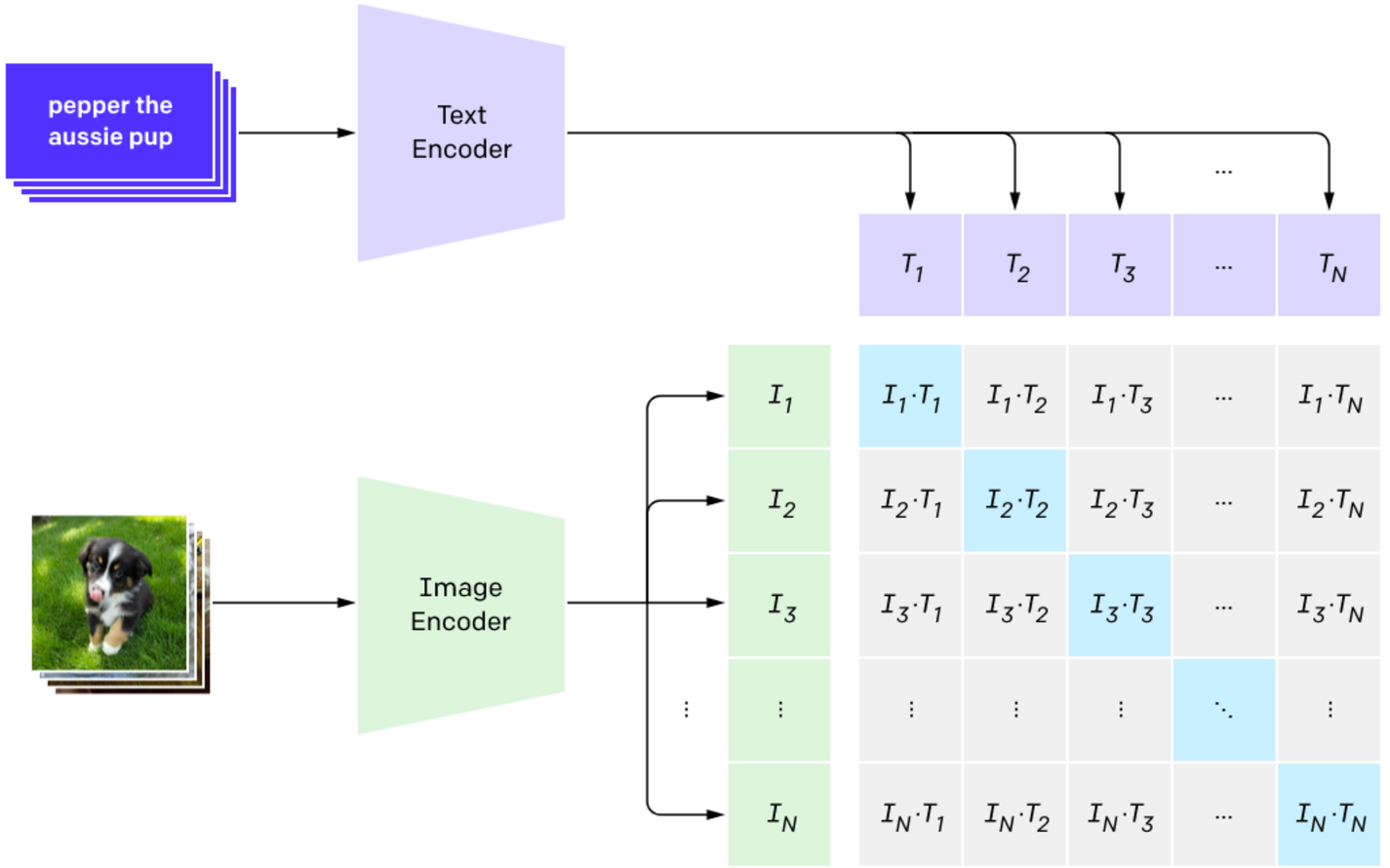


$$L_{\text{simple}} := E_{t\sim[1,T],x_0\sim q(x_0),\epsilon\sim\mathcal{N}(0,\mathbf{I})}[||\epsilon - \epsilon_\theta(x_t, t)||^2]$$

Image credit

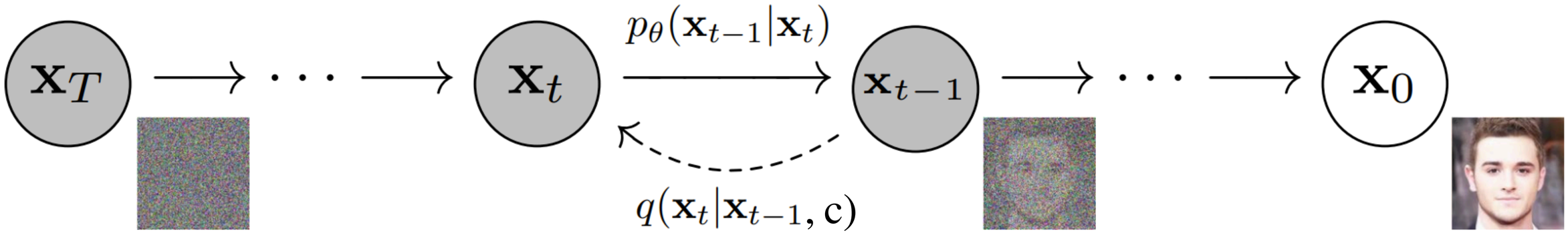
DALL-E 2

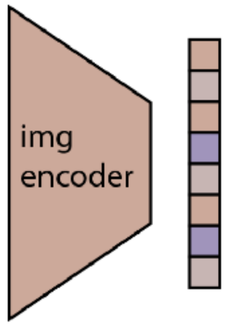
1. Обучаем CLIP



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)

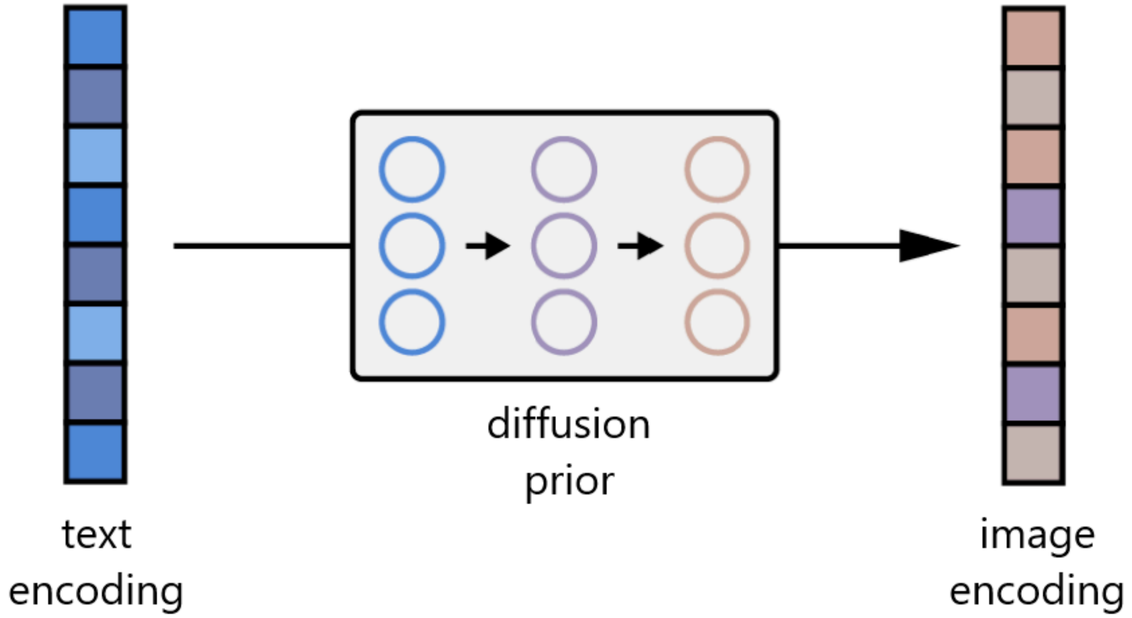


Image credit

DALL-E 2

Генерация:

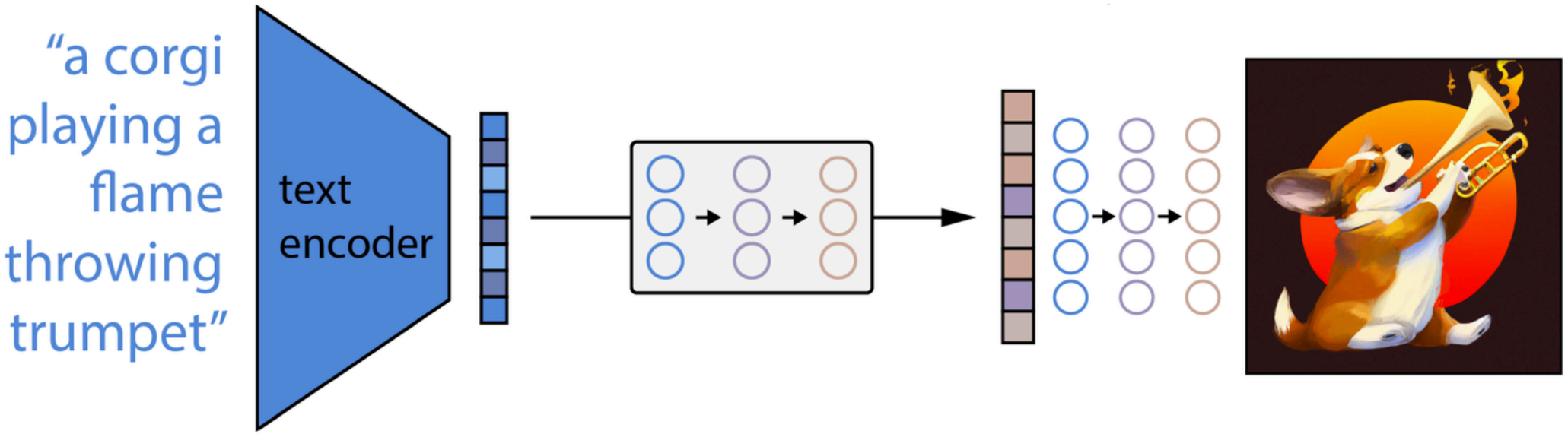


Image credit

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