

Генеративные модели в машинальном обучении

Лекция 6
Мультиомодальные модели

Михаил Гущин

mhushchyn@hse.ru

НИУ ВШЭ, 2024

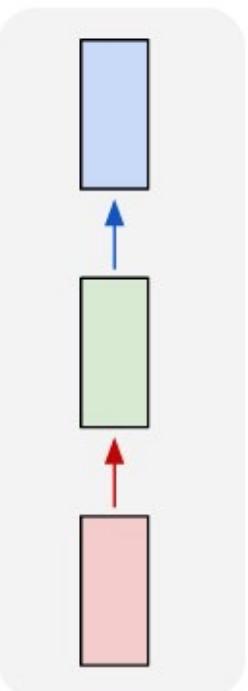


НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ

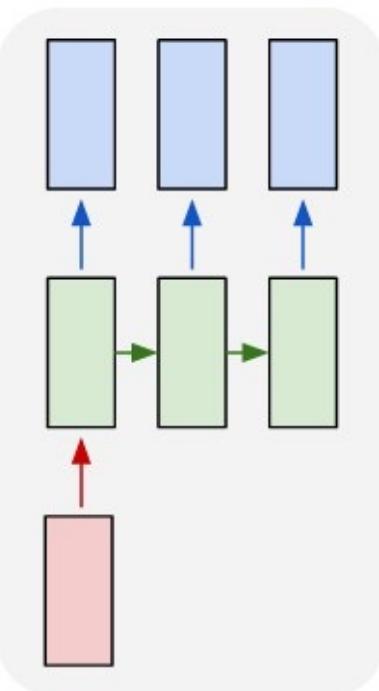
Трансформеры (повтор)

Типы задач

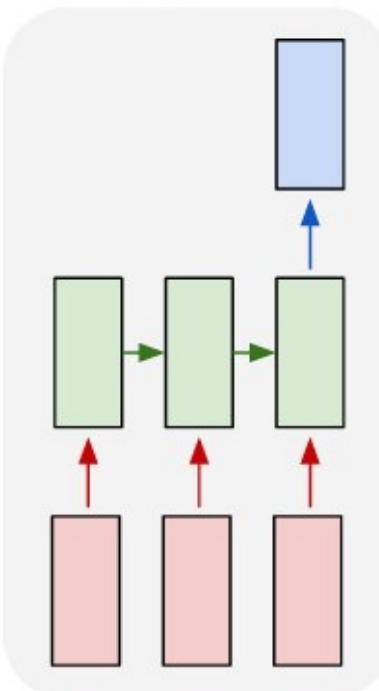
one to one



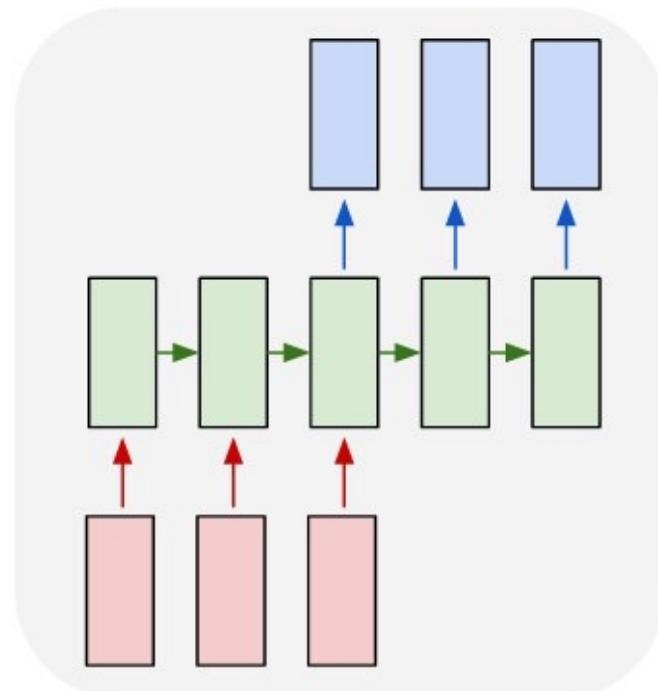
one to many



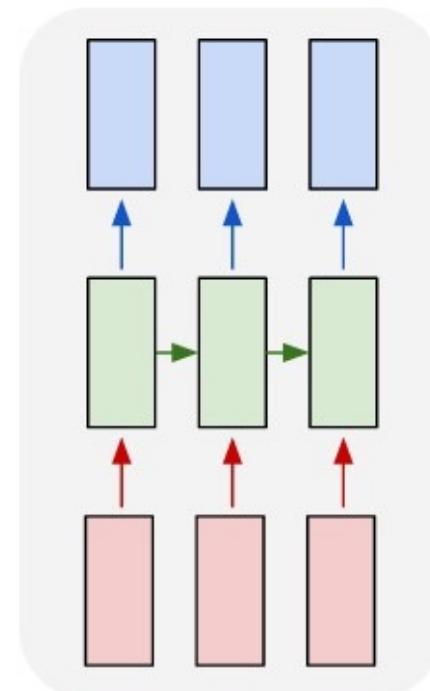
many to one



many to many

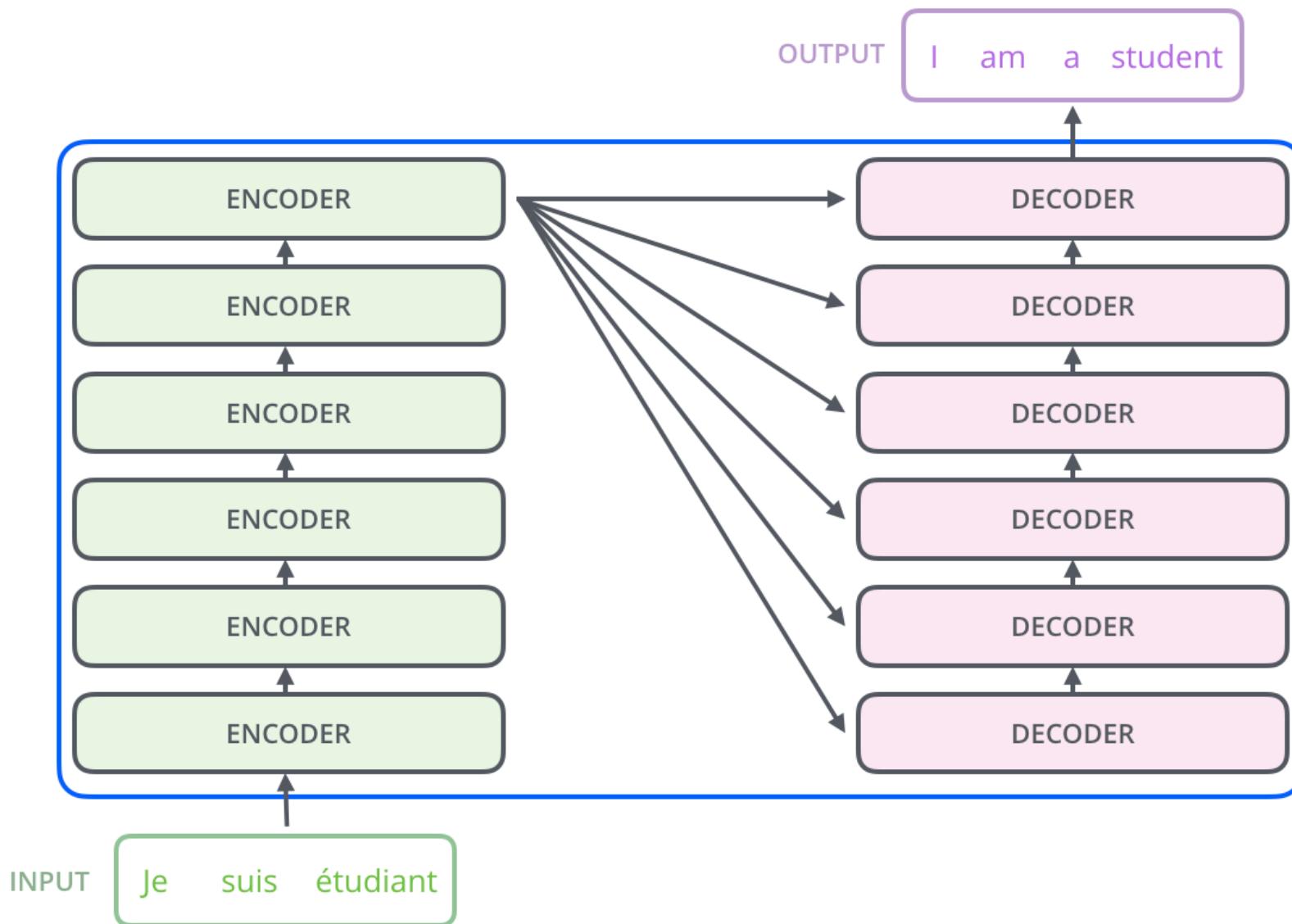


many to many

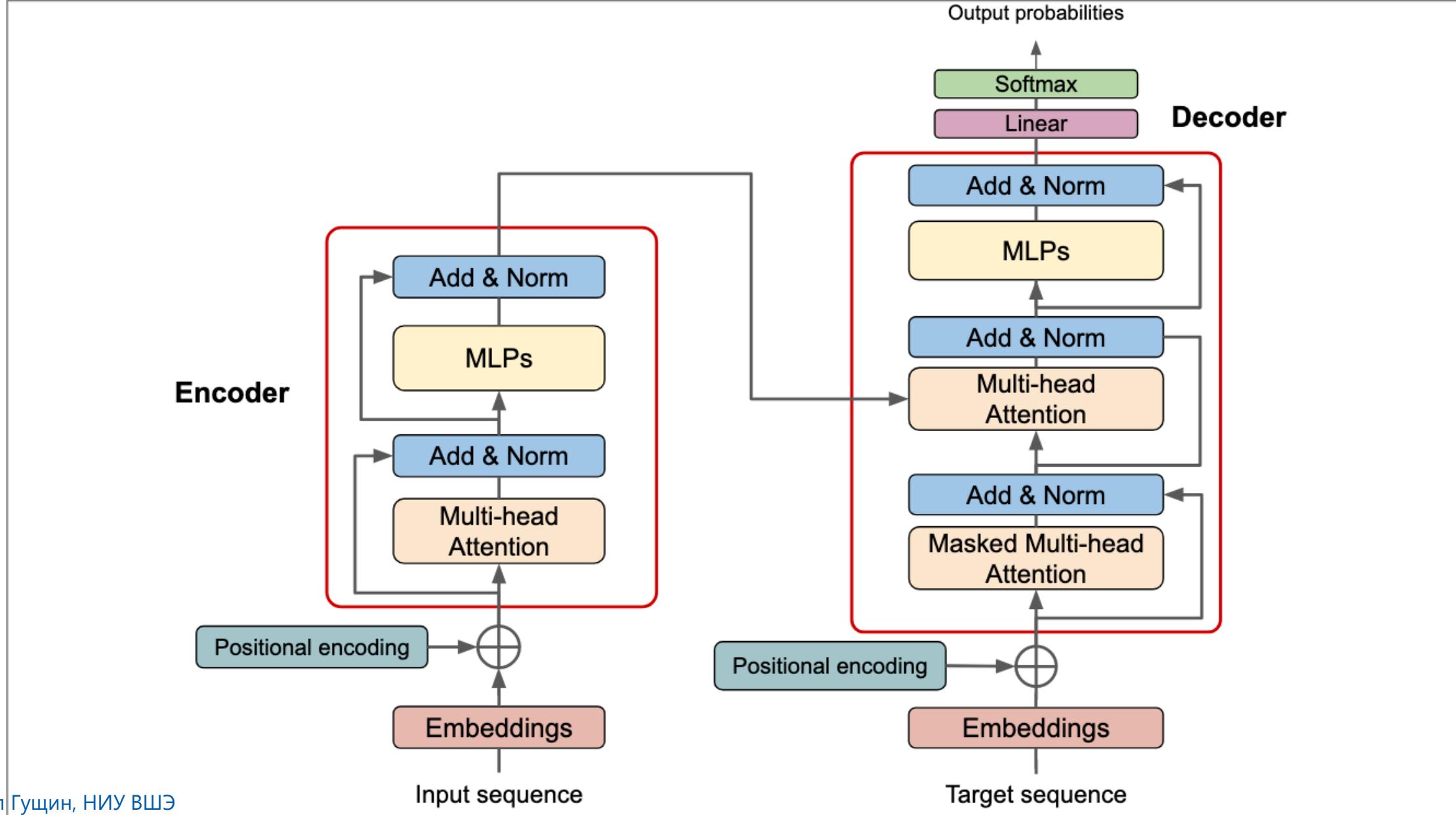


Источник: <http://karpathy.github.io/2015/05/21/rnn-effectiveness/>

Общая схема трансформера

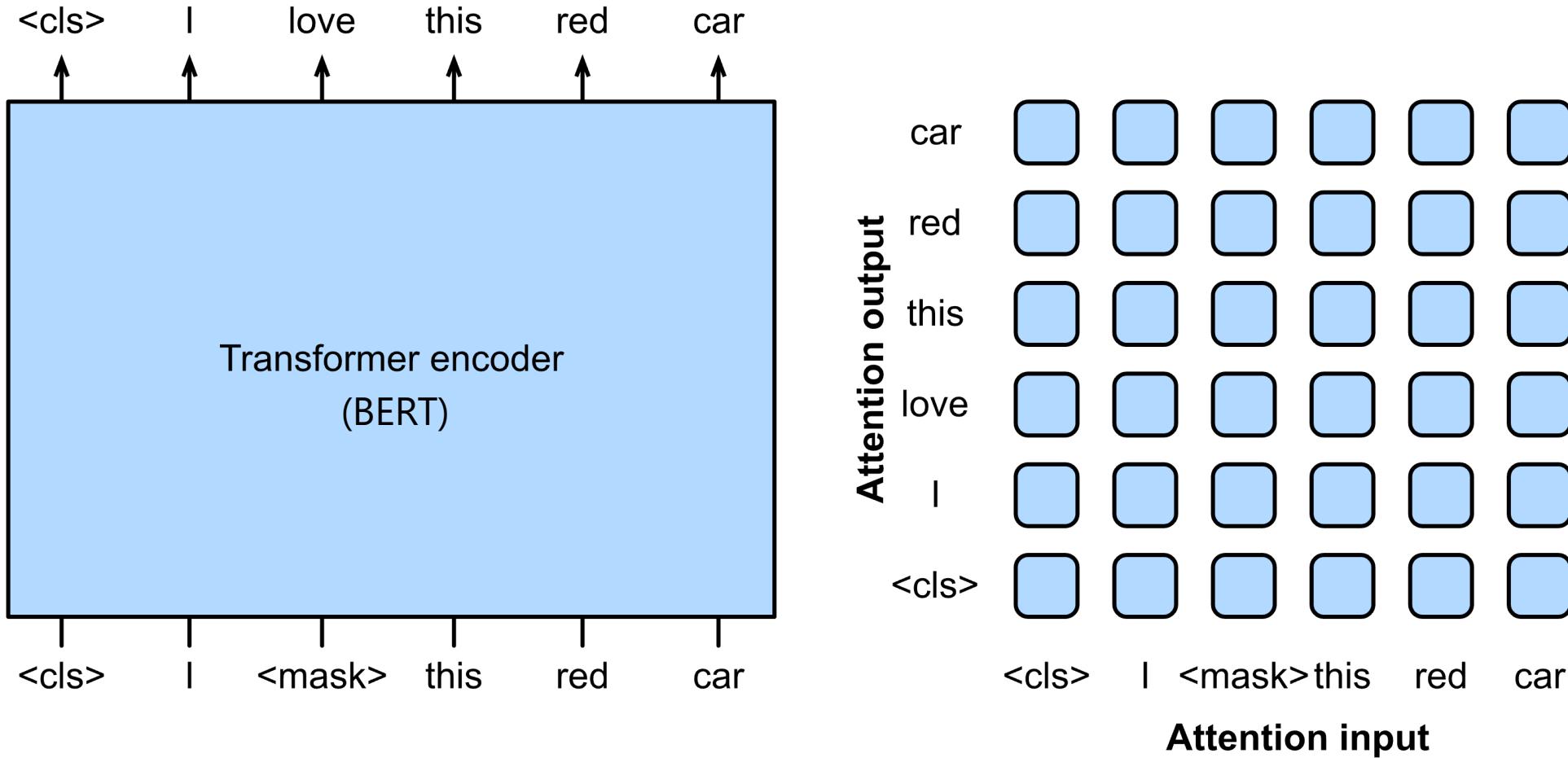


Общая схема трансформера



BERT
(Bidirectional Encoder Representations
from Transformers)
(повтор)

Semi-supervised: Masked Language Model



Semi-supervised: Masked Language Model

- ▶ Выкидываем часть слов (обычно 15%), модель должна их восстановить:
 - 80% меняем на токен <MASK>
 - 10% меняем на случайные
 - 10% не меняем
- ▶ Мы предсказываем пропуски, считаем Loss по всем 15%. Случайные и неизменные токены позволяют не переобучиться под токен <MASK>.

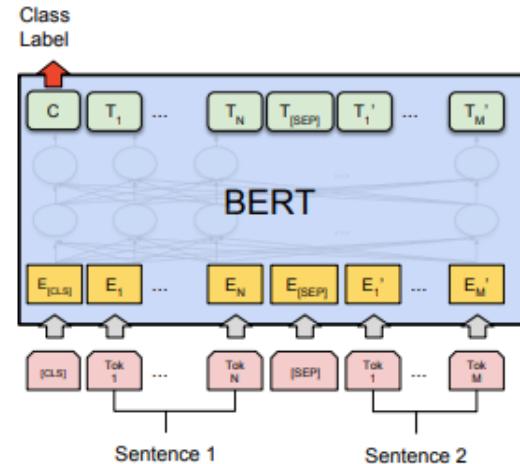
Специальные токены

- ▶ Пример токенезированного предложения:

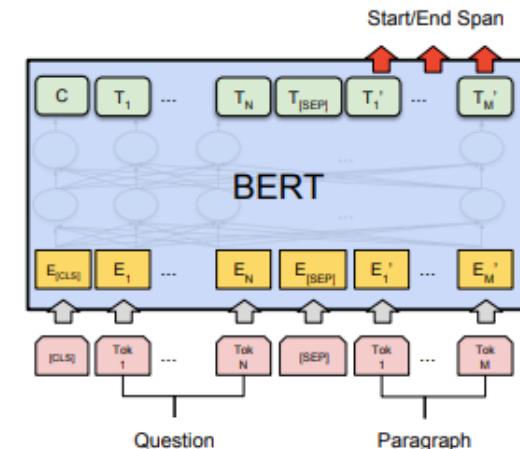
[CLS] my dog is cute [SEP] he likes play ##ing [SEP]

- ▶ [CLS] — специальный токен, который мы добавляем к началу любого предложения, он несет информацию о всей входной последовательности
- ▶ [SEP] — специальный токен, который говорит, что слева от него первое предложение, а справа — второе
- ▶ ## — специальный токен, обозначающий, что это кусок слова

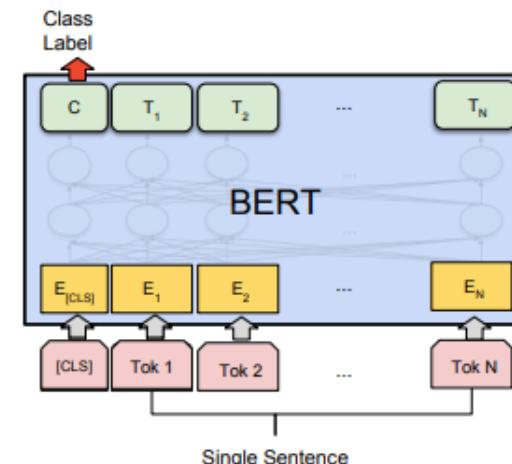
Task specific BERT



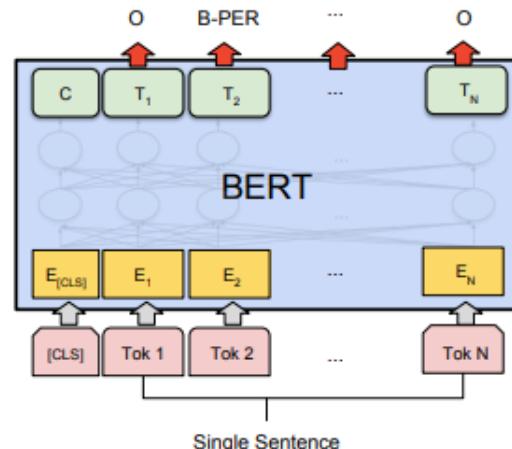
(a) Sentence Pair Classification Tasks:
MNLI, QQP, QNLI, STS-B, MRPC,
RTE, SWAG



(c) Question Answering Tasks:
SQuAD v1.1

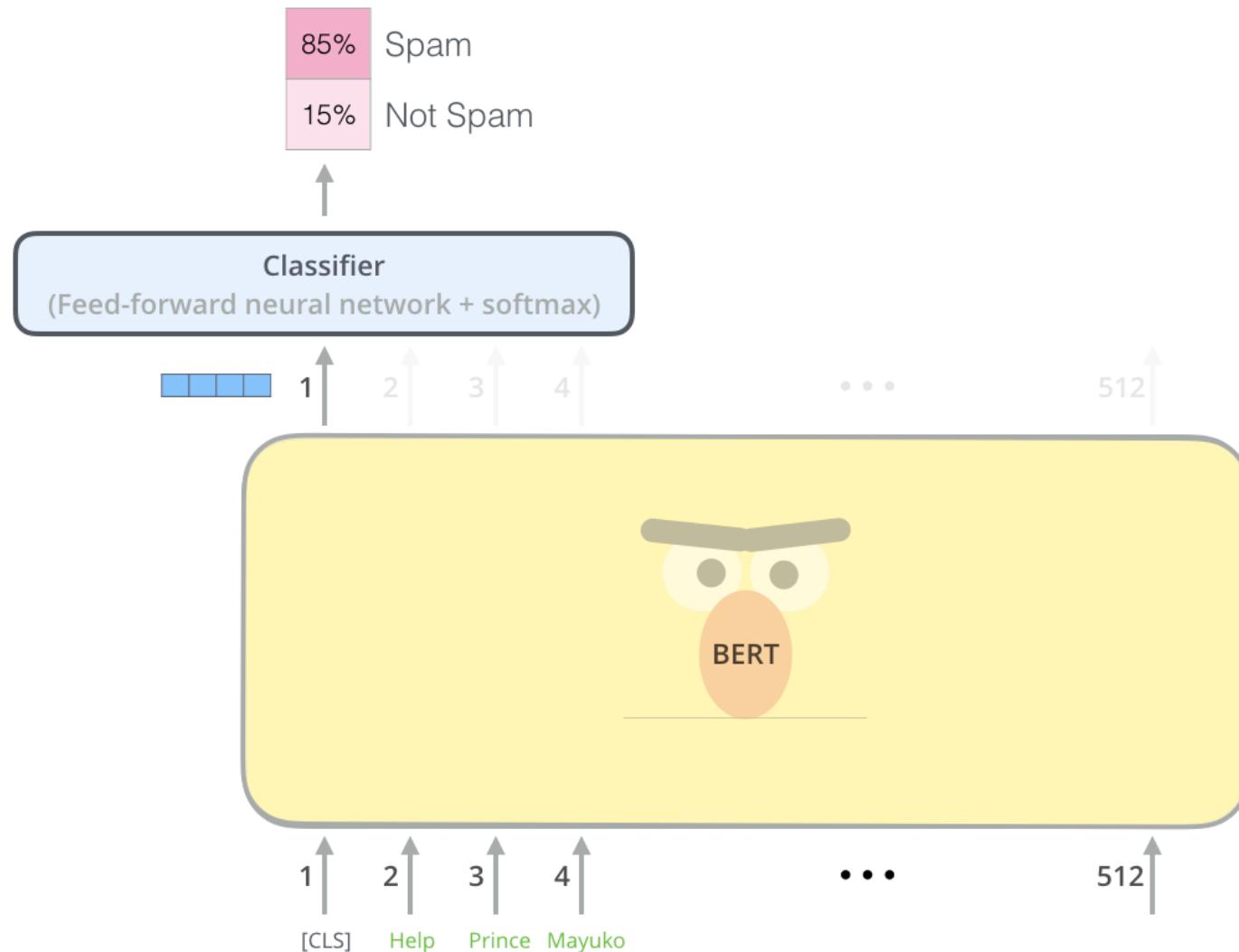


(b) Single Sentence Classification Tasks:
SST-2, CoLA

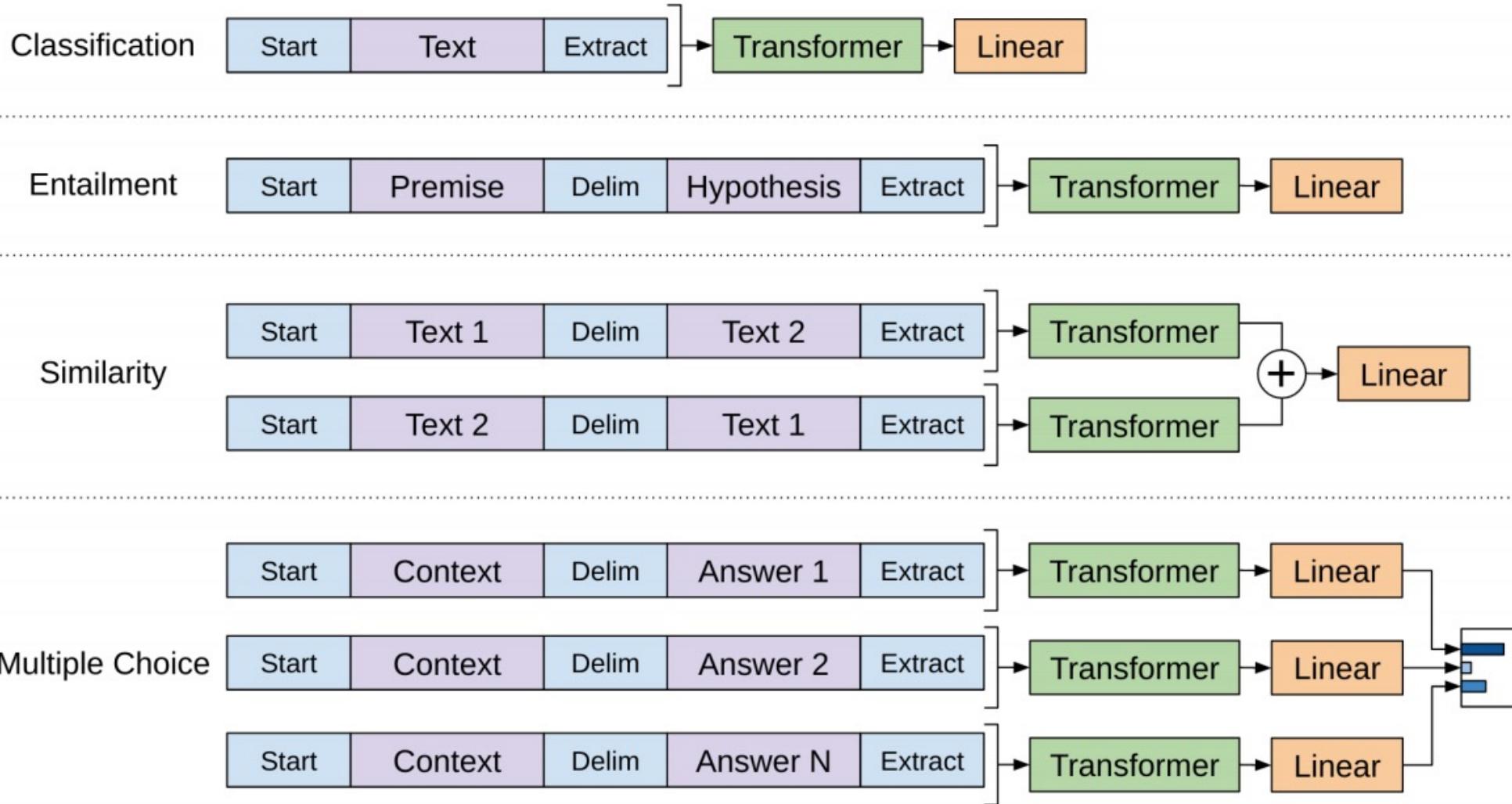
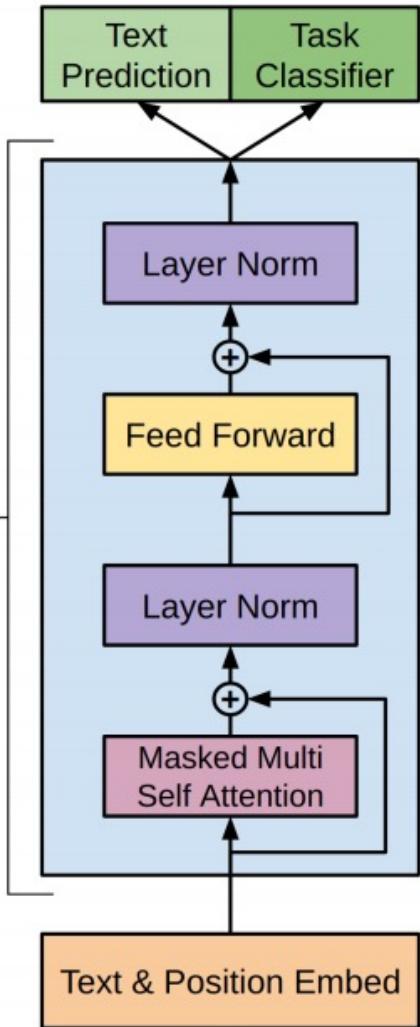


(d) Single Sentence Tagging Tasks:
CoNLL-2003 NER

Пример: классификация текста

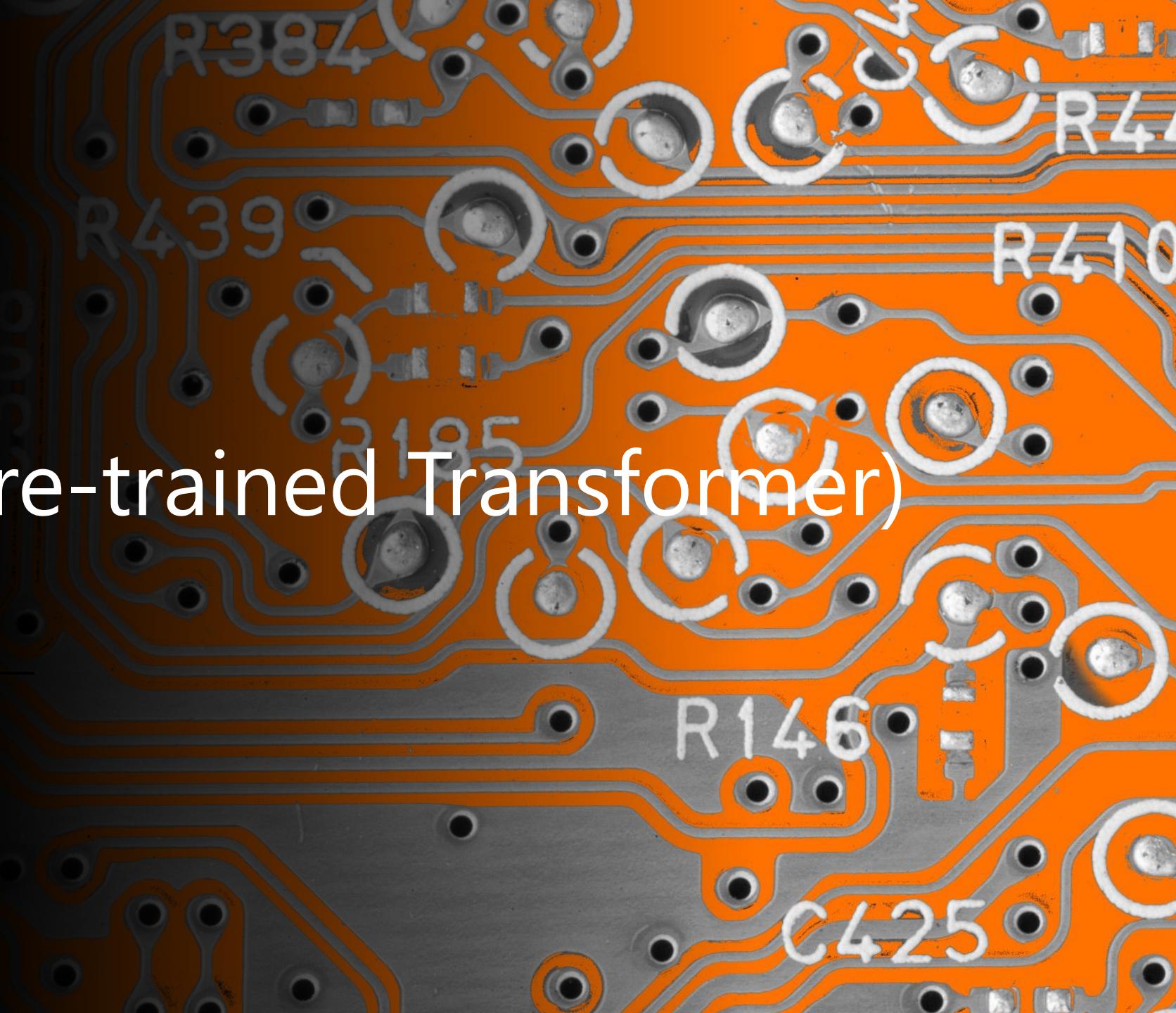


Примеры

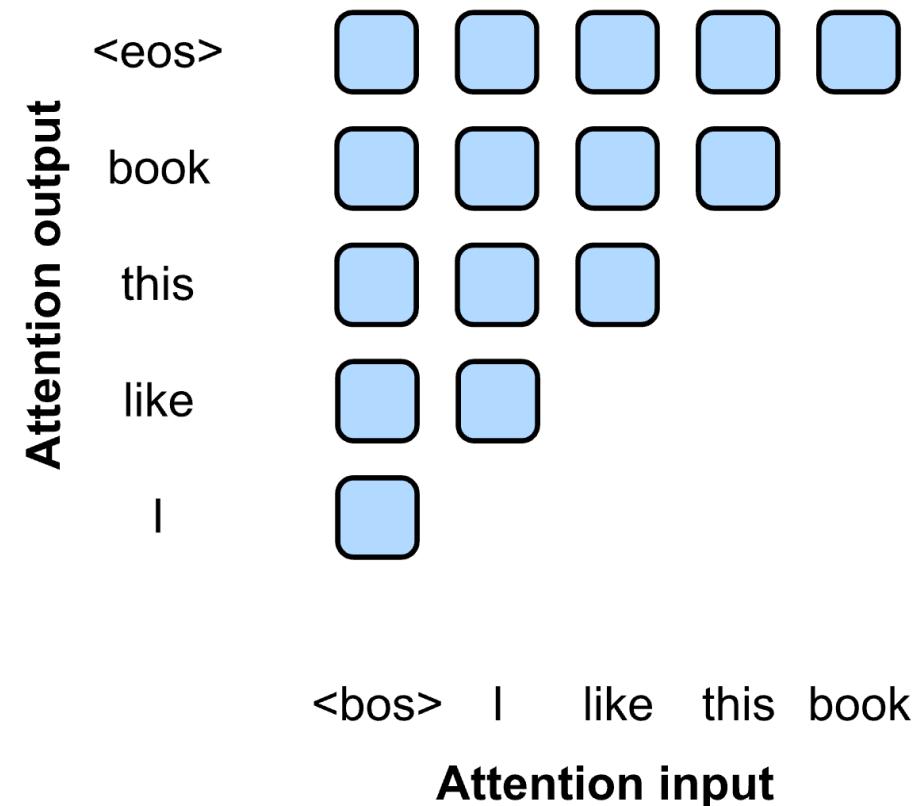
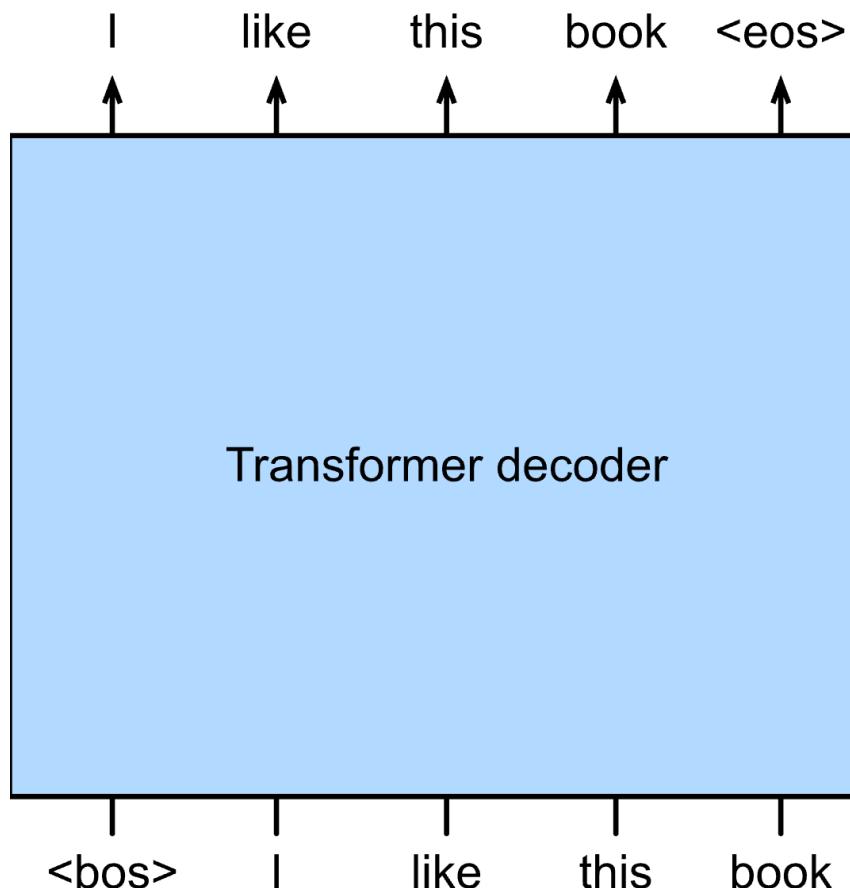


Источник: <https://paperswithcode.com/method/gpt>

GPT
(Generative Pre-trained Transformer)
(повтор)



GPT



Источник: https://d2l.ai/chapter_attention-mechanisms-and-transformers/large-pretraining-transformers.html

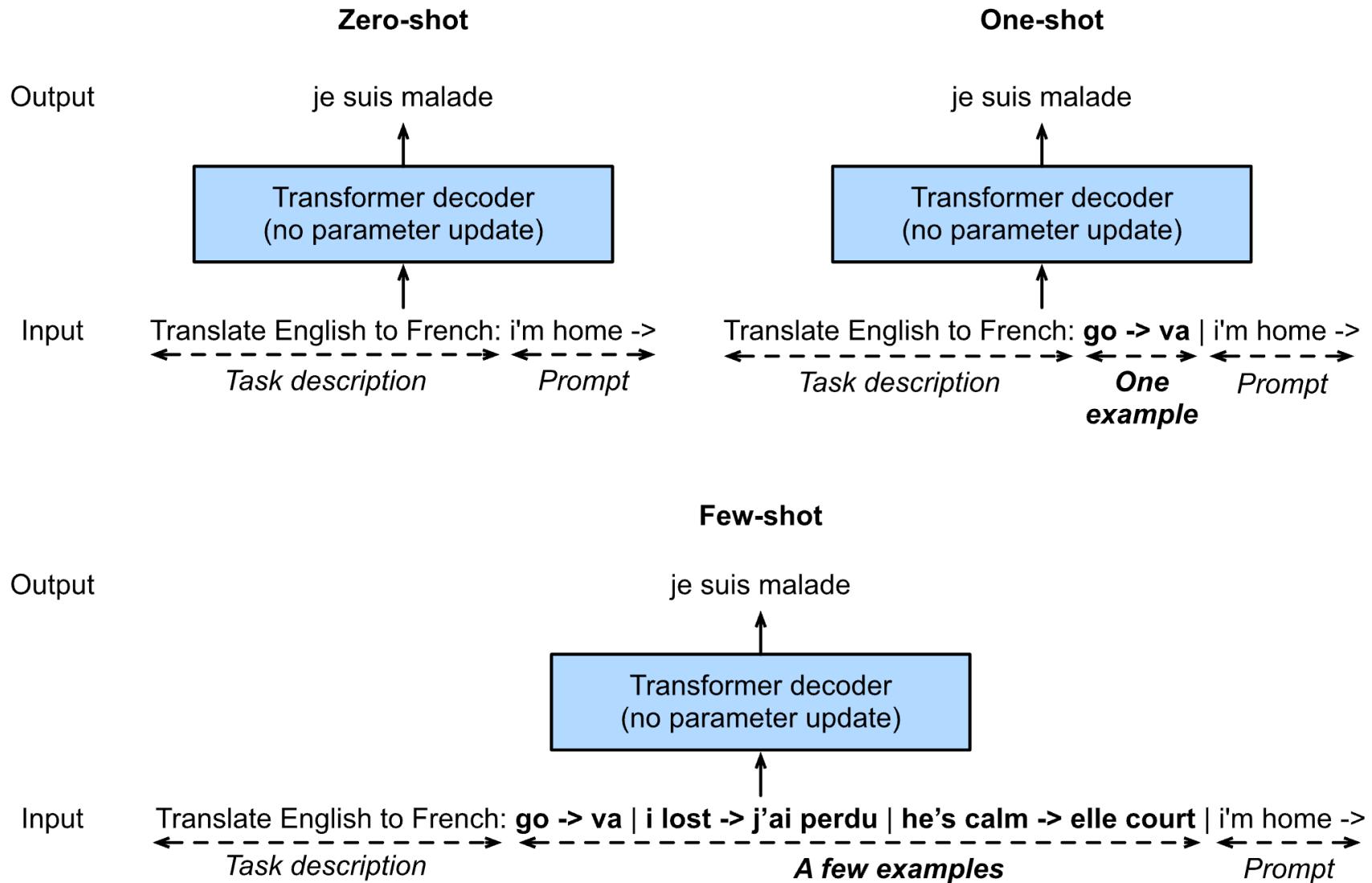
Обучение GPT

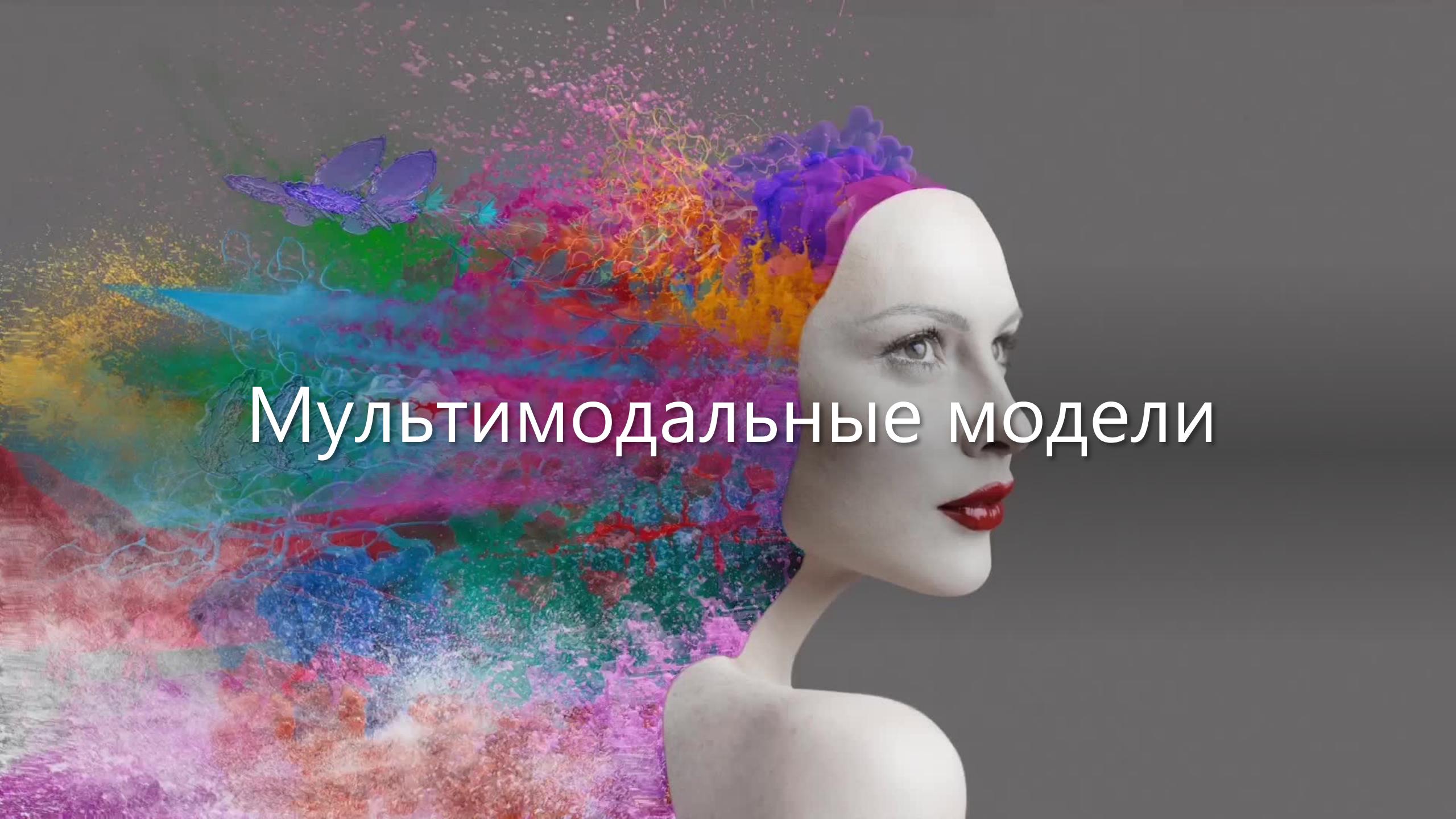
- ▶ Обучаем модель предсказывать на один токен вперед
- ▶ Выход – это вход, смещенный на 1 токен
- ▶ «<eos>» - специальный токен начала предложения
- ▶ «<bos>» - специальный токен конца предложения
- ▶ При подсчете attention смотрим только на предыдущие токены (не знаем будущее)

GPT-3 и далее

- ▶ Обучаем одну модель сразу решать несколько задач
- ▶ Для этого на вход подаем описание самой задачи:
 - Перевод с русского на английский
 - Саммаризация
 - Генерация текста на тему
 - И другие

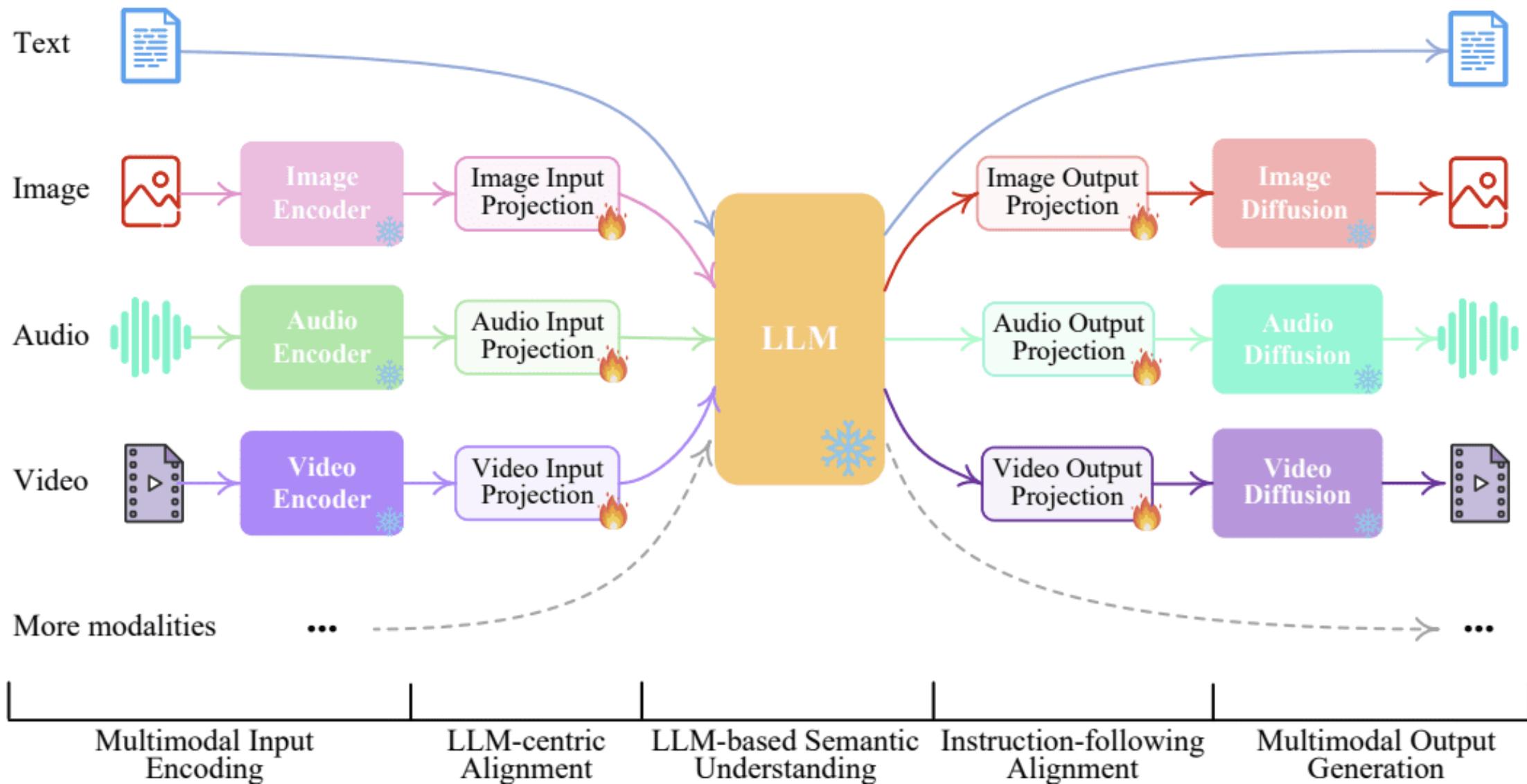
GPT-3 и далее



A woman's head is shown in profile, facing right. She has white skin, dark hair, and is wearing red lipstick. The background behind her head is a vibrant, abstract collage of various colors including purple, blue, green, yellow, and red, with some foliage and butterflies visible.

Мультимодальные модели

Мультимодальность



Пример

- Example-1: Text → Text+Image+Audio

I've recently started practicing mindfulness and meditation as part of my daily routine, but sometimes it's challenging. Can you help me with some techniques?

Of course! Mindfulness and meditation can be a wonderful way to cultivate a sense of calm and focus. One technique you might find helpful is called "mindful breathing." It involves taking slow, deep breaths and focusing your attention on the sensation of the breath entering and leaving your body. Another technique is "body scan meditation," where you consciously scan your entire body from head to toe, paying attention to any sensations or discomfort. You can also try guided meditations that provide a structured practice. Would you like me to find a video that demonstrates mindful breathing or body scan meditation?

I wish to listen to a music for the meditation first.

I guess you may want some musics that sound soft.

▶ 0:00 / 0:09

Base on the audio content, visualize it in an image.



Can you infer what time is in the picture?

Based on the image, it is likely that the time is either early morning or late afternoon. The sunlight is either low or setting, creating a serene atmosphere.

Пример

How many cars are in this picture?



There are two cars shown in this picture.

Tell me what kind of car are they?

The two cars in the picture are SUVs.

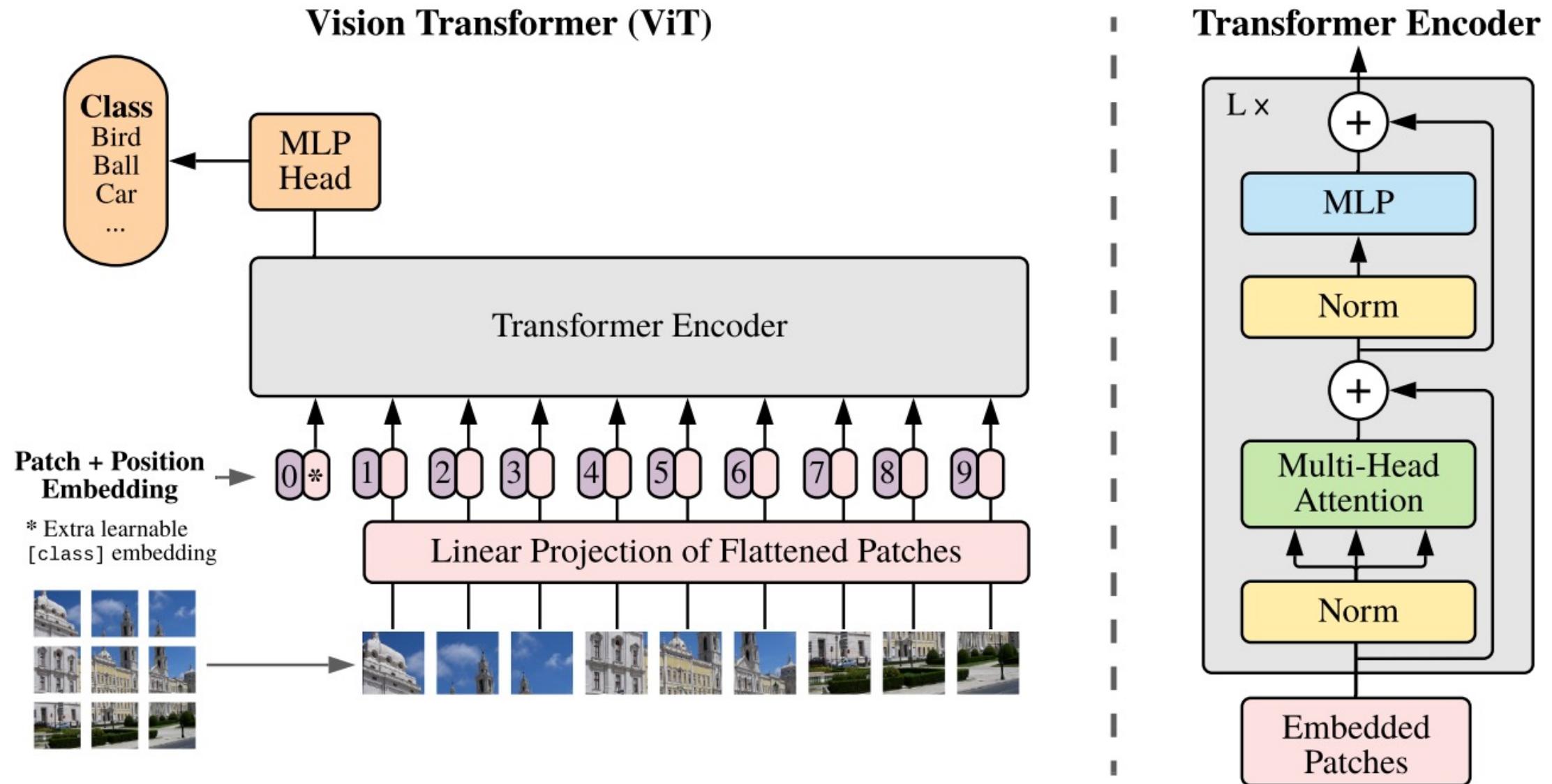
Can you show me the sound of starting the engines of them.

Sure, here's the sound of the engine of the SUV:

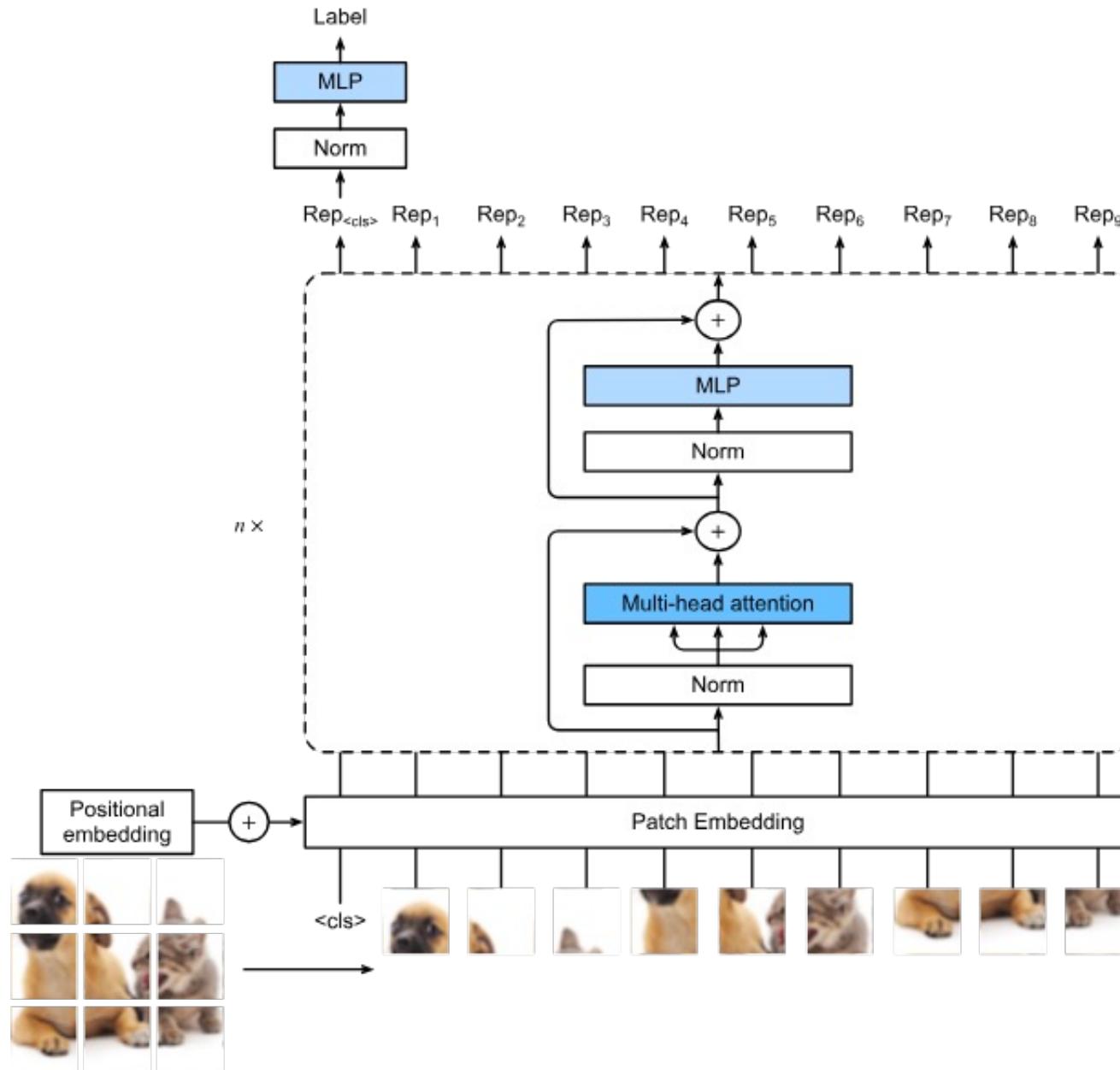
▶ 0:00 / 0:09

Vision Transformer (ViT)





- ▶ Делим изображение на фрагменты фиксированного размера.
Например, на 9 одинаковых фрагментов
- ▶ Вытягиваем фрагменты в вектор и пропускаем через полно связные нейронные слои. Например, один линейный слой без функции активации
- ▶ Добавляем *positional encoding*
- ▶ Последовательность эмбеддингов подаем на вход **Transformer Encoder (как BERT)**



- ▶ На вход подаем специальный токен «<cls>» и девять фрагментов изображения
- ▶ Они преобразуются в десять представлений соответственно
- ▶ Представление «<cls>» далее преобразуется в выходную метку

Пример

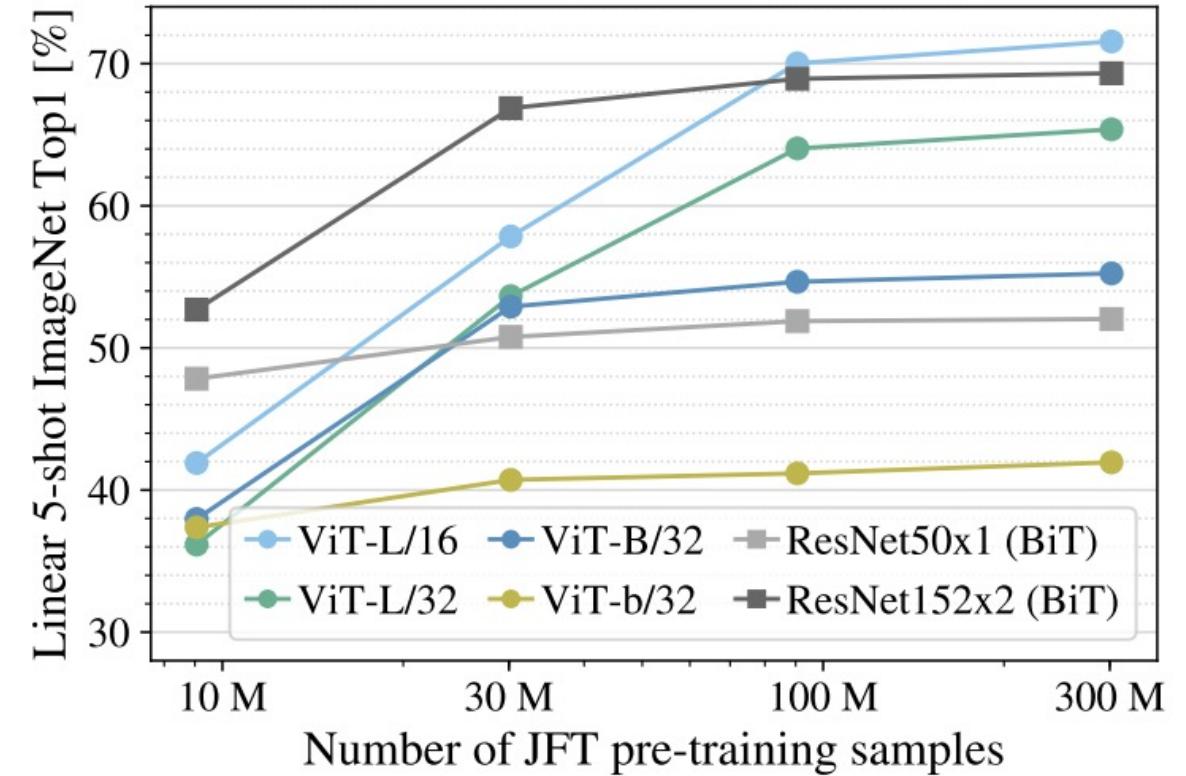
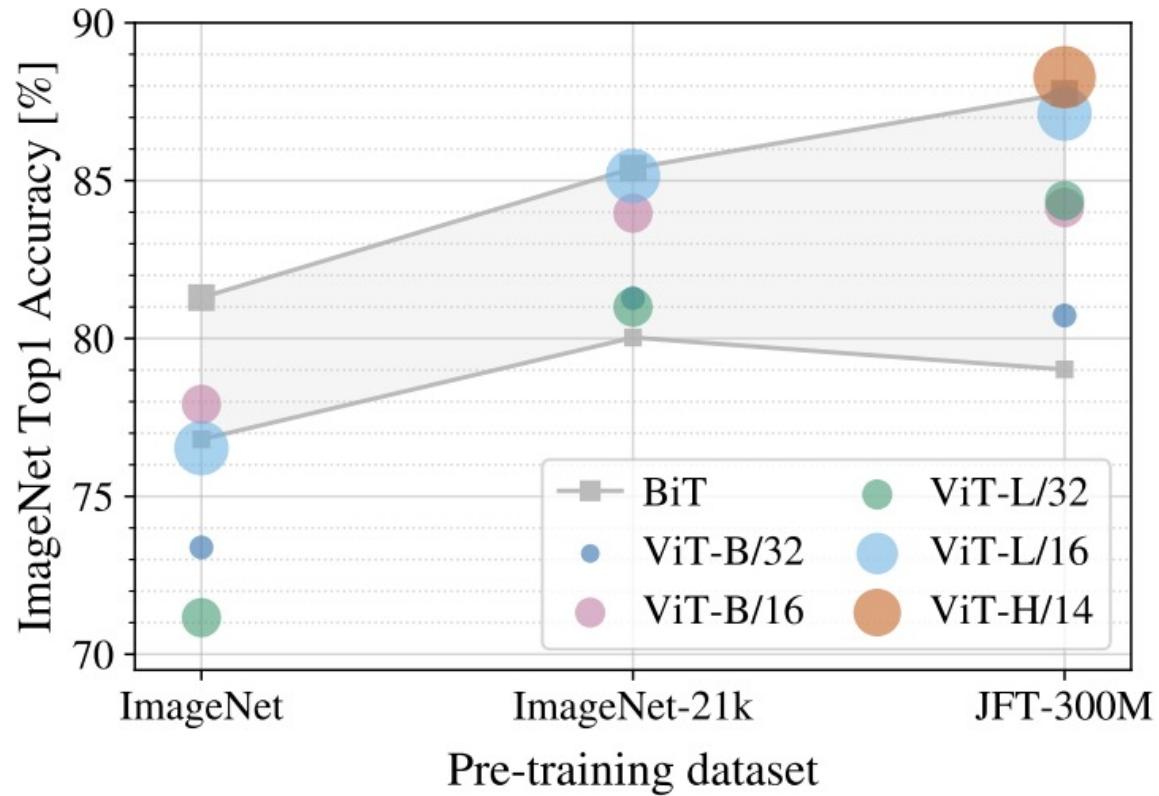
Model	Layers	Hidden size D	MLP size	Heads	Params
ViT-Base	12	768	3072	12	86M
ViT-Large	24	1024	4096	16	307M
ViT-Huge	32	1280	5120	16	632M

Источник: <https://arxiv.org/pdf/2010.11929>

Пример

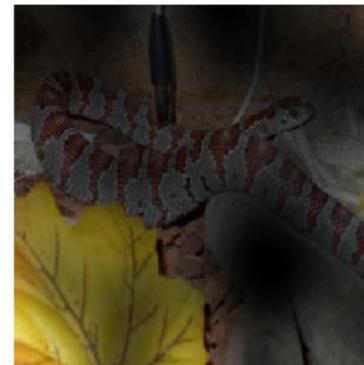
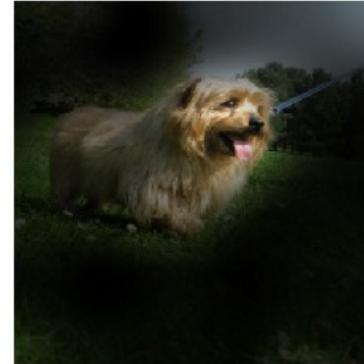
	Ours-JFT (ViT-H/14)	Ours-JFT (ViT-L/16)	Ours-I21k (ViT-L/16)	BiT-L (ResNet152x4)	Noisy Student (EfficientNet-L2)
ImageNet	88.55 ± 0.04	87.76 ± 0.03	85.30 ± 0.02	87.54 ± 0.02	88.4/88.5*
ImageNet ReaL	90.72 ± 0.05	90.54 ± 0.03	88.62 ± 0.05	90.54	90.55
CIFAR-10	99.50 ± 0.06	99.42 ± 0.03	99.15 ± 0.03	99.37 ± 0.06	—
CIFAR-100	94.55 ± 0.04	93.90 ± 0.05	93.25 ± 0.05	93.51 ± 0.08	—
Oxford-IIIT Pets	97.56 ± 0.03	97.32 ± 0.11	94.67 ± 0.15	96.62 ± 0.23	—
Oxford Flowers-102	99.68 ± 0.02	99.74 ± 0.00	99.61 ± 0.02	99.63 ± 0.03	—
VTAB (19 tasks)	77.63 ± 0.23	76.28 ± 0.46	72.72 ± 0.21	76.29 ± 1.70	—
TPUv3-core-days	2.5k	0.68k	0.23k	9.9k	12.3k

Пример



Пример

Input Attention



Другие задачи

- ▶ ViT можно использовать не только для классификации изображений, но и для задач детектирования объектов и сегментации
- ▶ Примеры моделей:
 - Segmentation Transformer (SETR)
 - Swin Transformer
 - Segmenter
 - ViT-YOLO
 - и много других

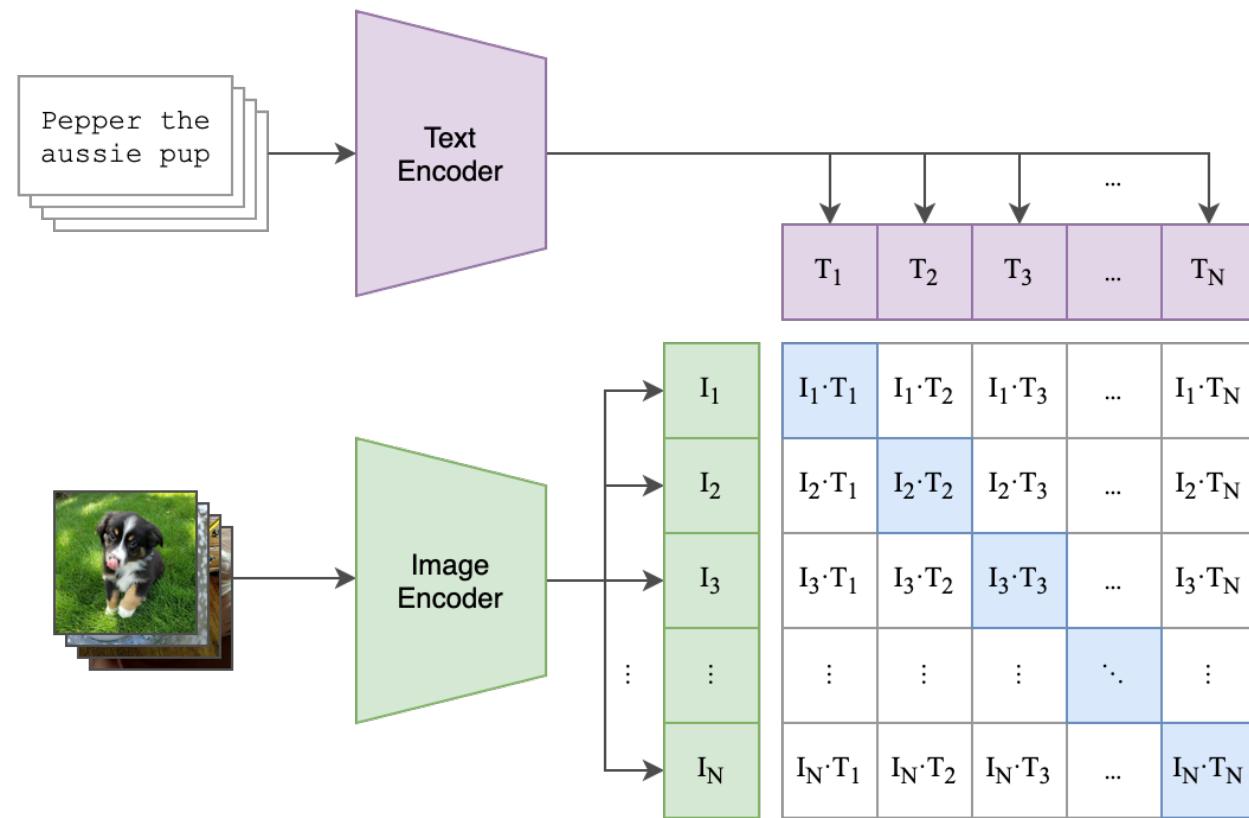
Contrastive Language- Image Pre-training (CLIP)

CLIP

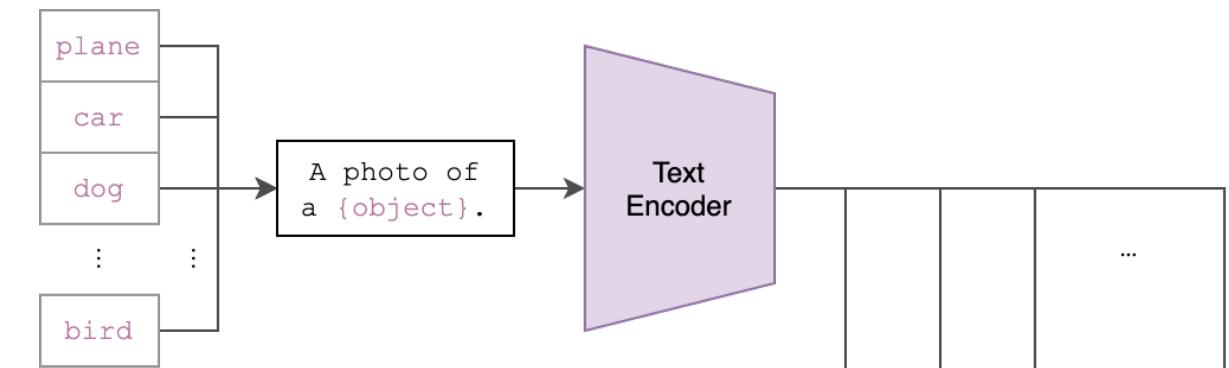
- ▶ CLIP (Contrastive Language-Image Pre-Training) – это нейронная сеть, обученная на парах (картинка, текст)
- ▶ Она может предсказать наиболее подходящее текстовое описание для заданной картинки без обучения на конкретной задаче (zero-shot learning)

CLIP

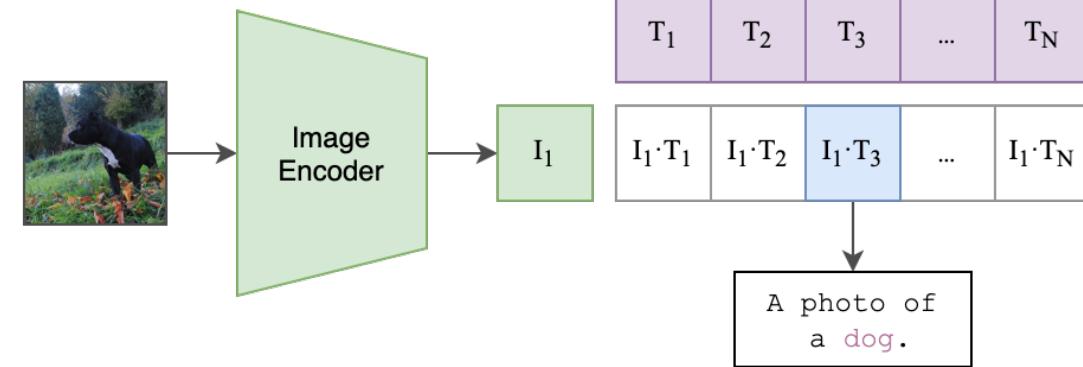
(1) Contrastive pre-training



(2) Create dataset classifier from label text



(3) Use for zero-shot prediction



Источник: <https://github.com/OpenAI/CLIP>

Архитектура CLIP

- ▶ Image Encoder:
 - Предобученные ResNet или ViT
- ▶ Text Encoder:
 - Предобученный GPT
- ▶ Линейные слои для согласования размерностей эмбеддингов изображений и текстов
- ▶ Матрица парных косинусных расстояний между эмбеддингами

Обучение CLIP

- ▶ Для обучения используем cross entropy loss
- ▶ Максимизируем косинусное расстояние НА диагонали матрицы (текст соответствует картинке)
- ▶ Минимизируем косинусное расстояние ВНЕ диагонали матрицы (текст не соответствует картинке)
- ▶ Модель обучалась на 400M парах (картинка, текст)

Пример матрицы

Cosine similarity between text and image features



Обучение CLIP

```
# image_encoder - ResNet or Vision Transformer
# text_encoder - CBOW or Text Transformer
# I[n, h, w, c] - minibatch of aligned images
# T[n, l] - minibatch of aligned texts
# W_i[d_i, d_e] - learned proj of image to embed
# W_t[d_t, d_e] - learned proj of text to embed
# t - learned temperature parameter

# extract feature representations of each modality
I_f = image_encoder(I) #[n, d_i]
T_f = text_encoder(T) #[n, d_t]

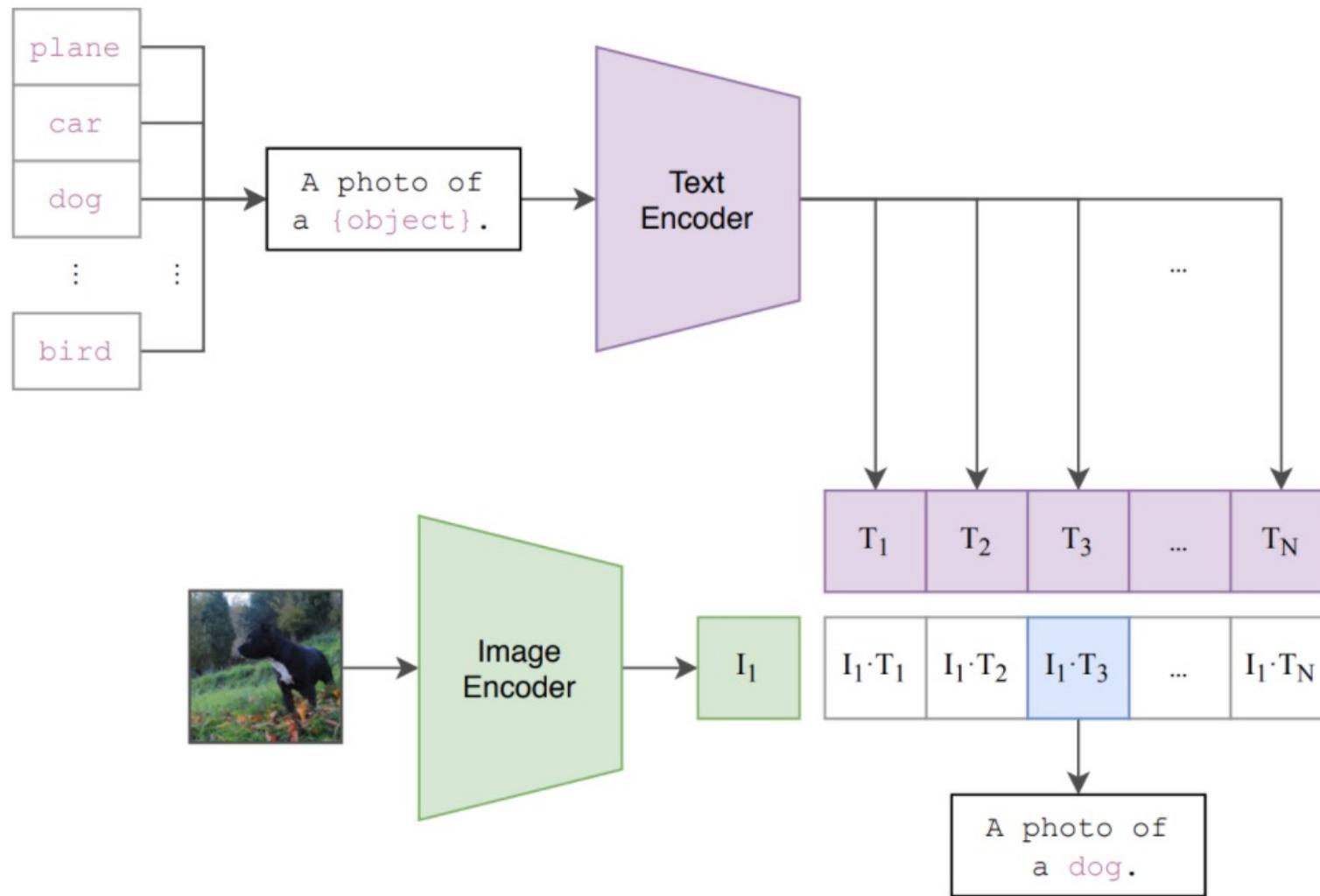
# joint multimodal embedding [n, d_e]
I_e = l2_normalize(np.dot(I_f, W_i), axis=1)
T_e = l2_normalize(np.dot(T_f, W_t), axis=1)

# scaled pairwise cosine similarities [n, n]
logits = np.dot(I_e, T_e.T) * np.exp(t)

# 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
```

Figure 3. Numpy-like pseudocode for the core of an implementation of CLIP.

Zero-shot



Zero-shot

- ▶ Модель обучалась на большом наборе разных текстов и изображений
- ▶ Из-за чего она хорошо работает на парах из других датасетов, которых не было в обучающей выборке
- ▶ Нужно помнить, что такой *transfer learning* хорошо работает для похожих задач
- ▶ Не стоит ожидать, что CLIP будет хороша для медицинских, сейсмических или астрономических снимков

Примеры

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

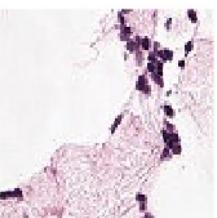
annual crop land (12.9%) Ranked 4 out of 10



- ✗ 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**.

PATCHCAMELYON (PCAM)

healthy lymph node tissue (22.8%) Ranked 2 out of 2



- ✗ this is a photo of **lymph node tumor tissue**
- ✓ this is a photo of **healthy lymph node tissue**

IMAGENET-A (ADVERSARIAL)

lynx (4.2%) Ranked 5 out of 200



- ✗ a photo of a **fox squirrel**.
- ✗ a photo of a **mongoose**.
- ✗ a photo of a **skunk**.
- ✗ a photo of a **red fox**.
- ✓ a photo of a **lynx**.

Примеры

RESISC45

correct label: roundabout



correct rank: 1/45 correct probability: 96.39%

satellite imagery of roundabout.

satellite imagery of intersection.

satellite imagery of church.

satellite imagery of medium residential.

satellite imagery of chaparral.

Flowers-102

correct label: great masterwort



correct rank: 1/102 correct probability: 74.25%

a photo of a great masterwort, a type of flower.

a photo of a bishop of llandaff, a type of flower.

a photo of a pincushion flower, a type of flower.

a photo of a globe flower, a type of flower.

a photo of a prince of wales feathers, a type of flower.

Stanford Cars

correct label: 2012 Honda Accord Coupe correct rank: 1/196 correct probability: 63.30%



a photo of a 2012 honda accord coupe.

a photo of a 2012 honda accord sedan.

a photo of a 2012 acura tl sedan.

a photo of a 2012 acura tsx sedan.

a photo of a 2008 acura tl type-s.

ImageNet

correct label: King Charles Spaniel



correct rank: 1/1000 correct probability: 91.61%

a photo of a king charles spaniel.

a photo of a brittany dog.

a photo of a cocker spaniel.

a photo of a papillon.

a photo of a sussex spaniel.

Примеры

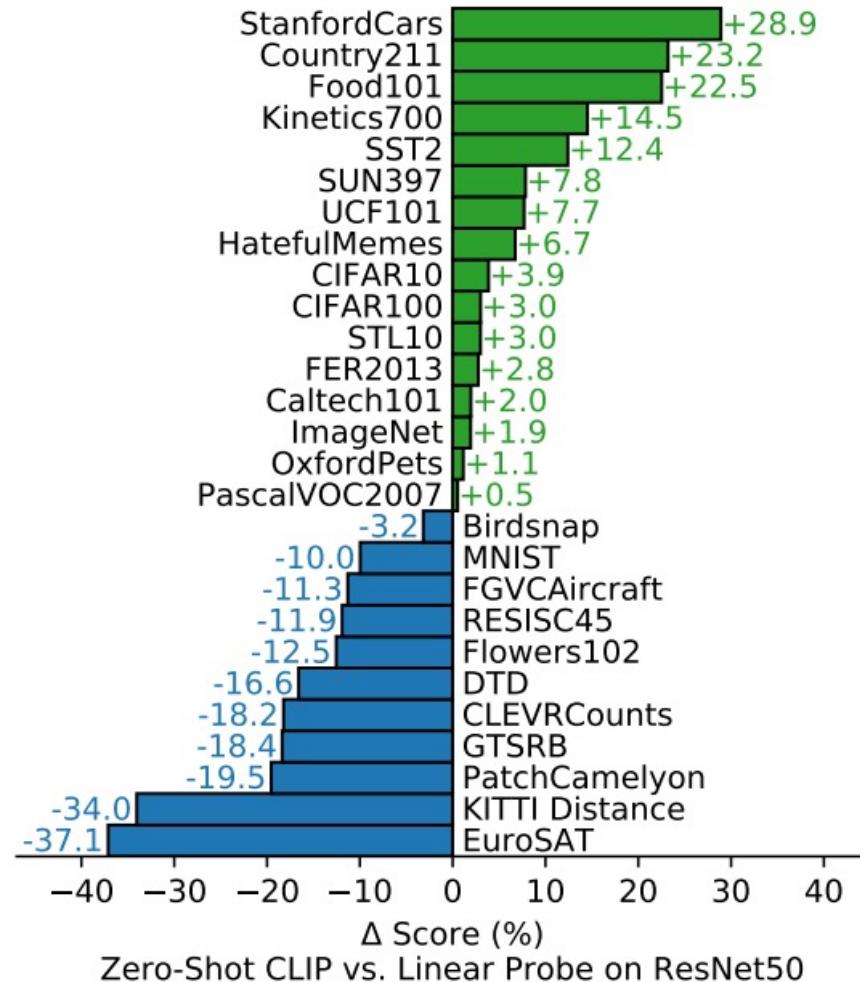


Figure 5. Zero-shot CLIP is competitive with a fully supervised baseline. Across a 27 dataset eval suite, a zero-shot CLIP classifier outperforms a fully supervised linear classifier fitted on ResNet-50 features on 16 datasets, including ImageNet.

Примеры

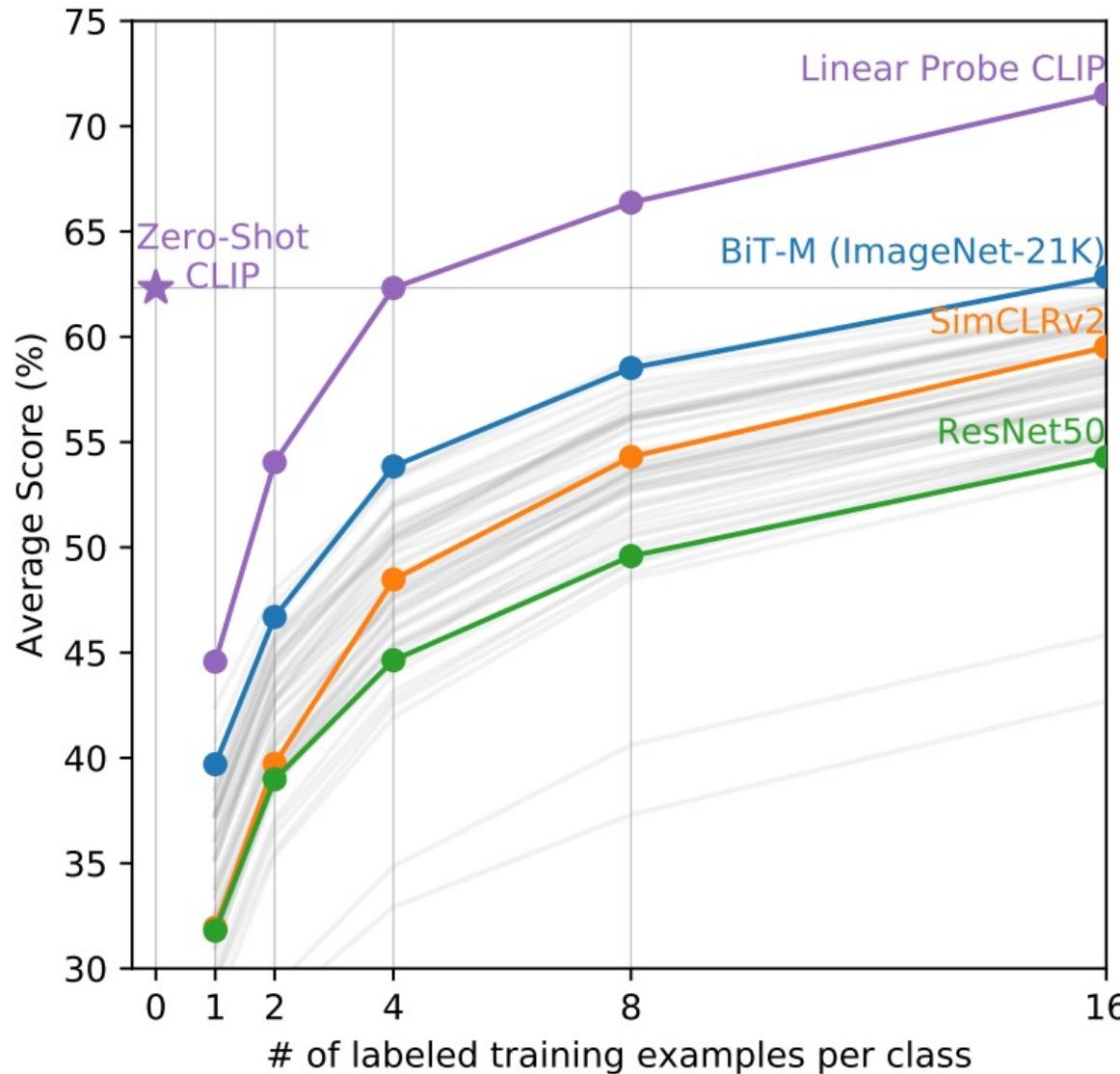


Figure 6. Zero-shot CLIP outperforms few-shot linear probes. Zero-shot CLIP matches the average performance of a 4-shot linear classifier trained on the same feature space and nearly matches the best results of a 16-shot linear classifier across publicly available models. For both BiT-M and SimCLRv2, the best performing model is highlighted. Light gray lines are other models in the eval suite. The 20 datasets with at least 16 examples per class were used in this analysis.

N-shot learning – это когда на каждый класс есть только N размеченных объектов в обучающей выборке