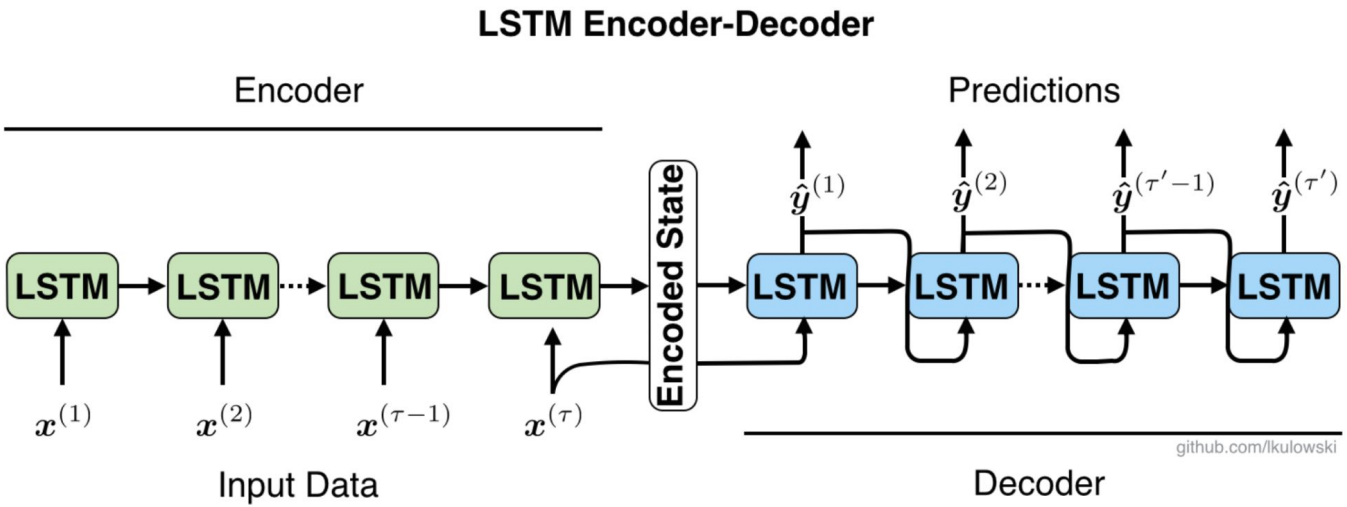
ИТМО

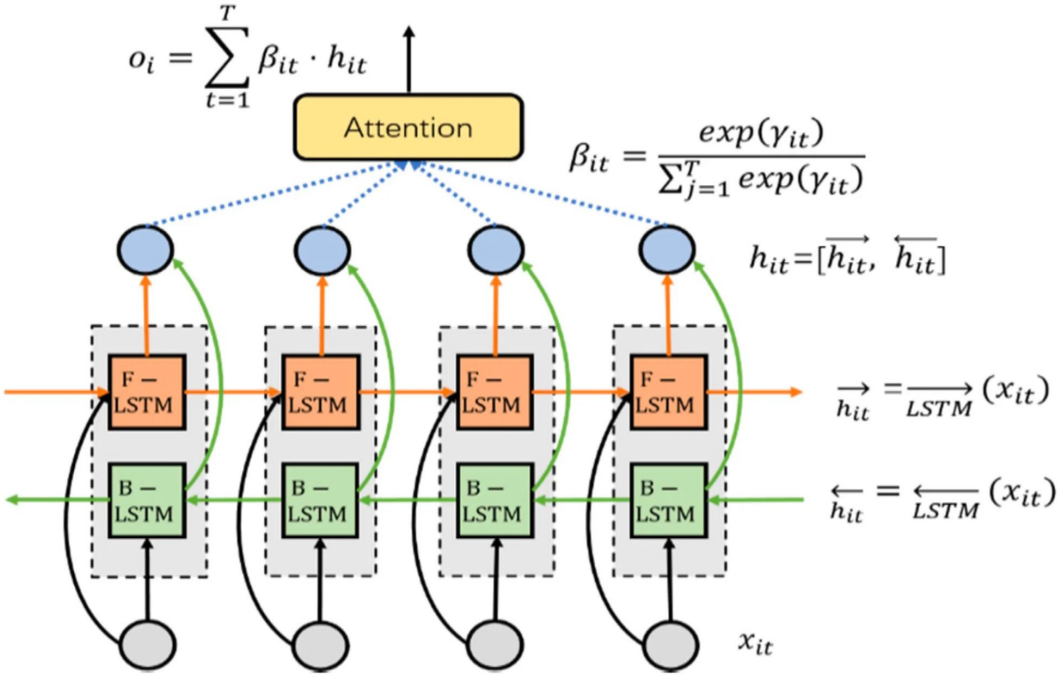
Современные архитектуры нейронных сетей

Трансформеры

LSTM Encoder-Decoder



Attention

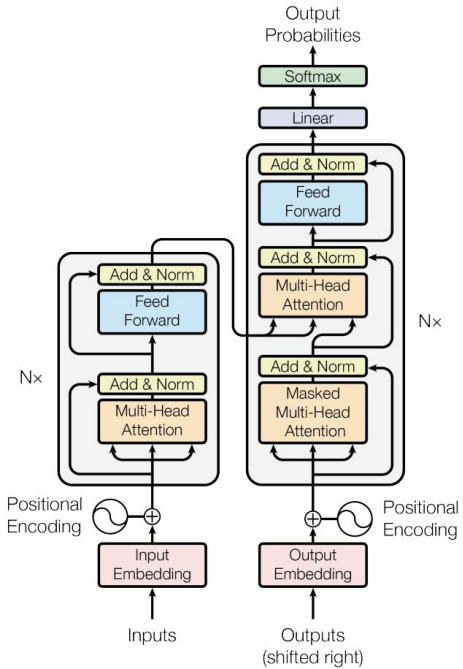

$o_i = \sum_{t=1}^{T} \beta_{it} \cdot h_{it}$
Attention
$\beta_{it} = \frac{exp(\gamma_{it})}{\sum_{j=1}^{T} exp(\gamma_{it})}$
$h_{it} = [\overrightarrow{h_{it}}, \overleftarrow{h_{it}}]$
F – LSTM
B – LSTM
$\overrightarrow{h_{it}} = \overrightarrow{LSTM}(x_{it})$
$\overleftarrow{h_{it}} = \overleftarrow{LSTM}(x_{it})$
x_{it}

Проблемы рекуррентности

- Сложно масштабировать
- Сложно распараллеливать, что значит, что использование видеокарт даст меньший прирост в скорости вычислений

Вывод: рекуррентность - зло

Attention is all you need

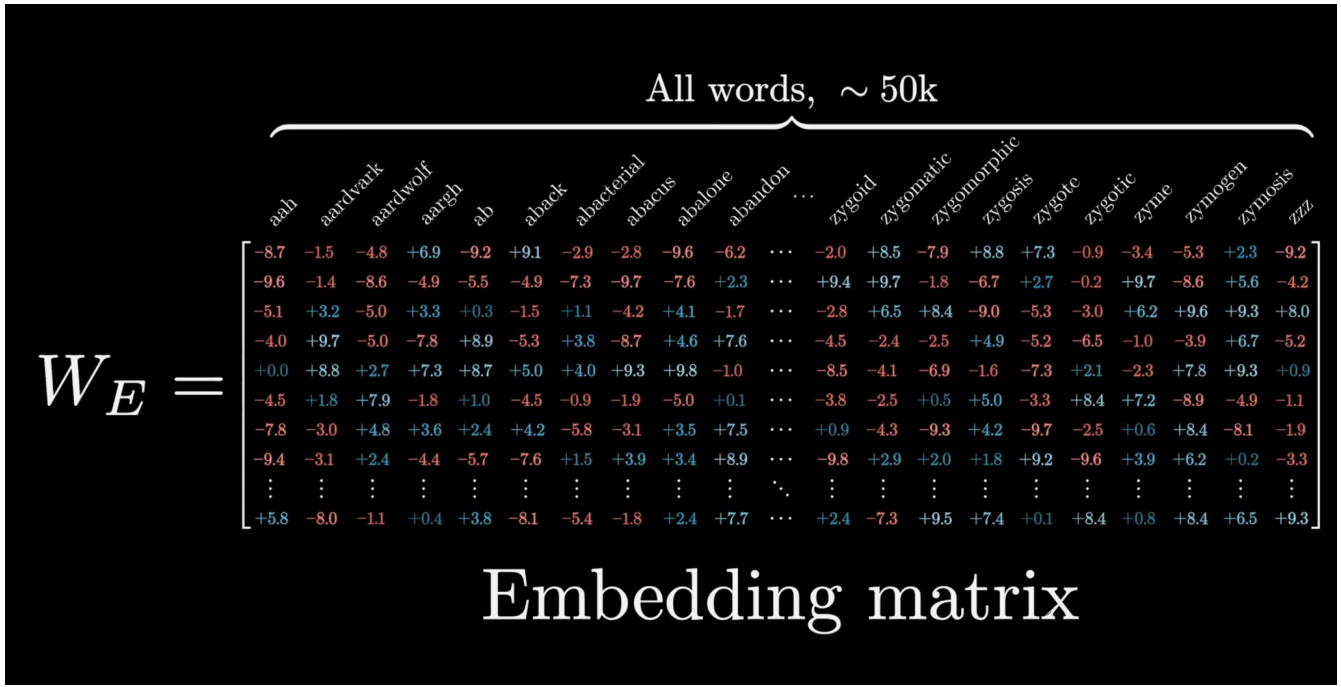


Токенизация по словам

Токенизация по символам

Токенизация по подсловам

Слой Embedding



Позиционное кодирование

ИТМО

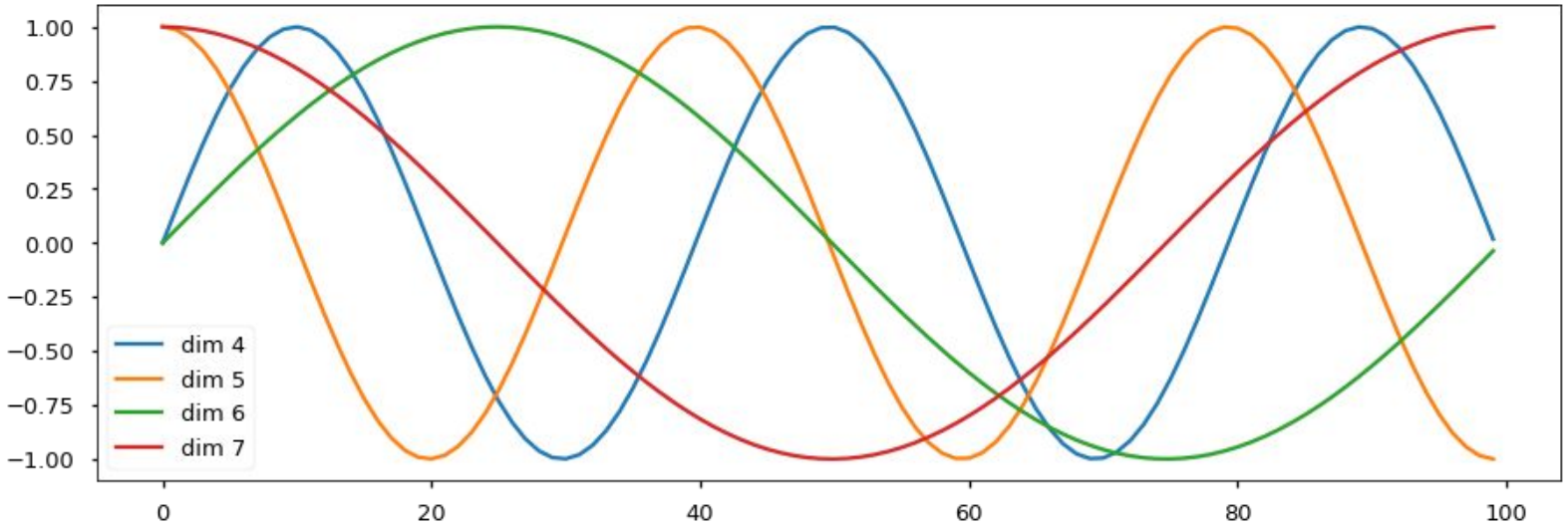
$$PE(pos, 2i) = \sin\left(\frac{pos}{1000^{2i/\text{embedding\_size}}}\right),$$

$$PE(pos, 2i+1) = \cos\left(\frac{pos}{1000^{(2i+1)/\text{embedding\_size}}}\right),$$

$$\text{pos\_encoding} = \begin{bmatrix} \sin\left(\frac{pos}{1000^{0/\text{embedding\_size}}}\right) \\ \cos\left(\frac{pos}{1000^{1/\text{embedding\_size}}}\right) \\ \sin\left(\frac{pos}{1000^{2/\text{embedding\_size}}}\right) \\ \dots \\ \cos\left(\frac{pos}{1000^{(\text{embedding\_size - 1})/\text{embedding\_size}}}\right) \end{bmatrix}$$

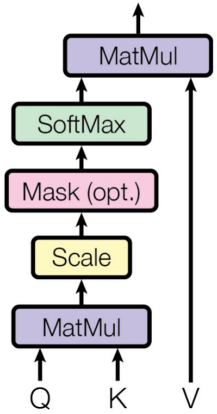
$$\hat{x} = x + \text{pos\_encoding}$$

Позиционное кодирование

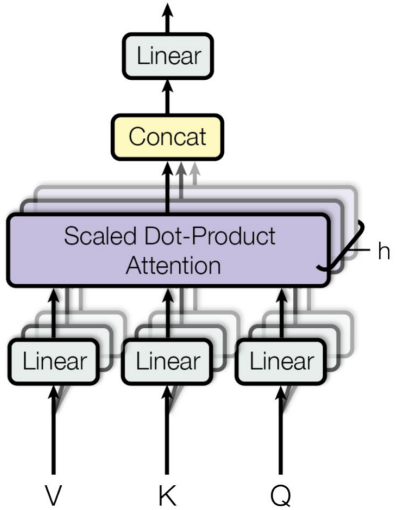


Attention

Scaled Dot-Product Attention

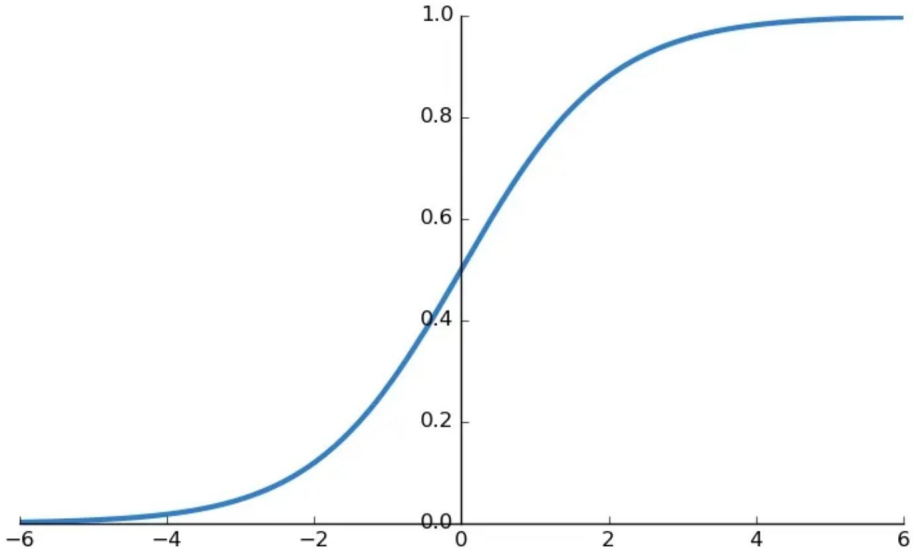


Multi-Head Attention



$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

Attention



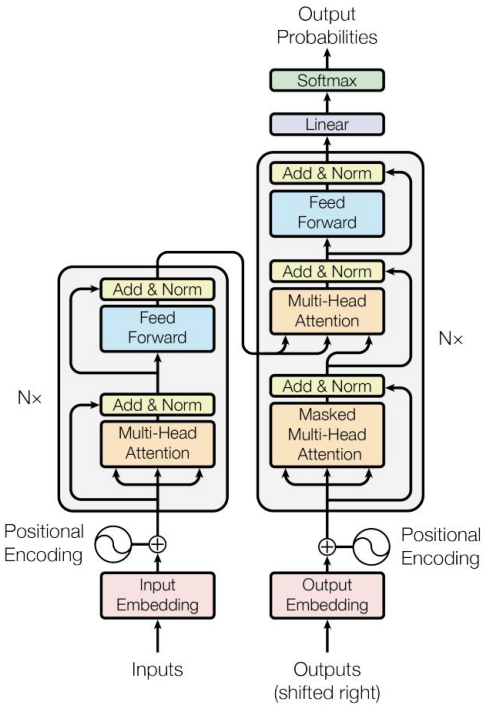
MultiHead Attention

ИТМО

$$\text{MultiHeadAttention} = \text{Concat}(head_1, head_2, ..., head_h)W^O$$

$$head_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

Encoder и Decoder

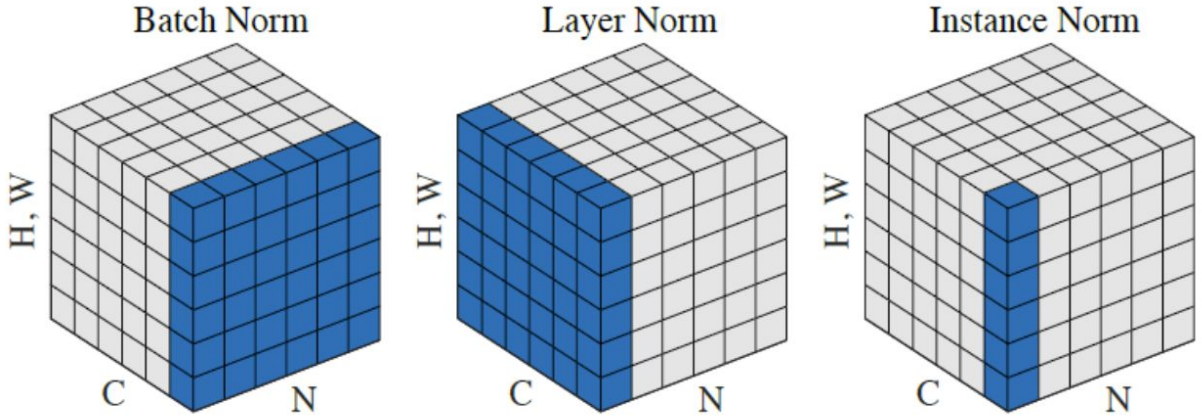


LayerNorm

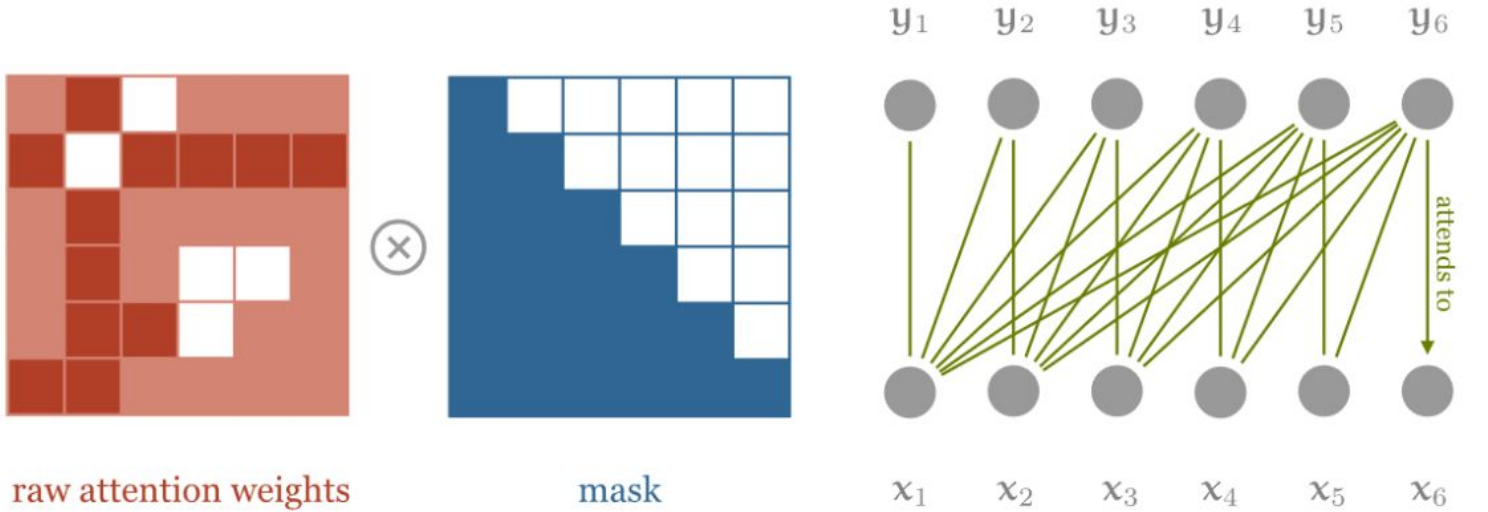
ИТМО

Формула общая с пакетной нормализацией

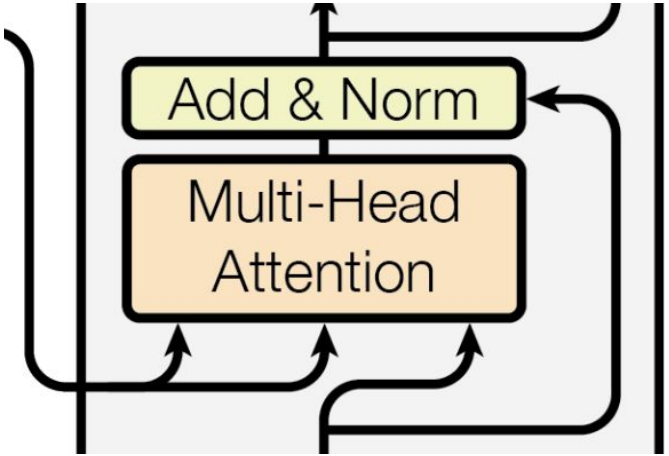
$$y = \frac{x - \mathrm{E}[x]}{\sqrt{\mathrm{Var}[x] + \epsilon}} * \gamma + \beta$$

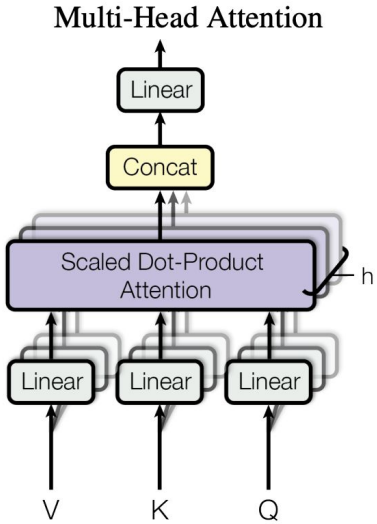


Masked MultiHead Attention



Cross attention





Температура трансформера

Softmax без температуры (обычный)

$$Softmax = \frac{e^{x_k}}{\sum_{i=0}^{n-1} e^{x_i}}$$

Softmax с температурой

$$Softmax_T = \frac{e^{x_k/T}}{\sum_{i=0}^{n-1} e^{x_i/T}}$$

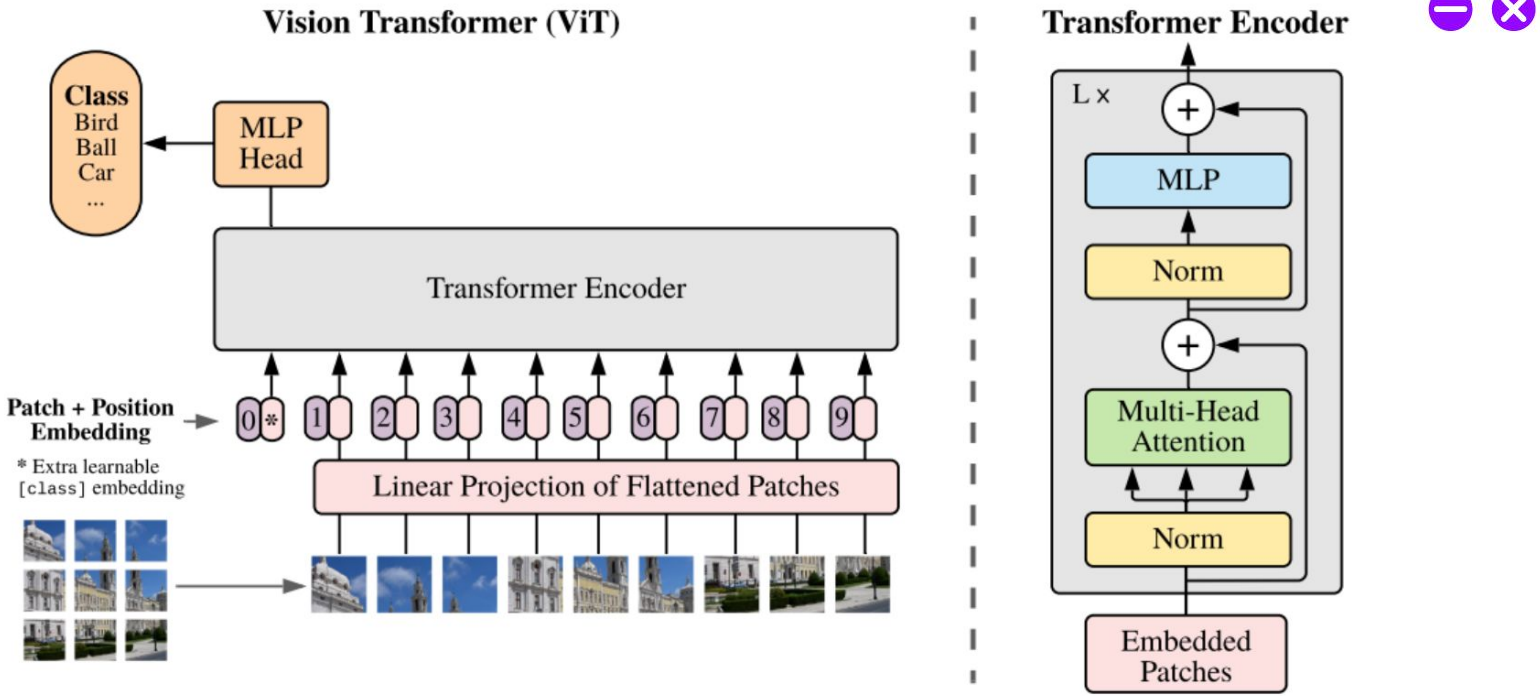
ViT. Проблемы сверточных нейронных сетей

- Отсутствие восприятия глобального контекста
- Фиксированная структура

С глобальным контекстом может помочь механизм внимания, а проблему фиксированной структуры хорошо решает архитектура трансформера

Архитектура ViT

ИТМО



Обучаемое позиционное кодирование