

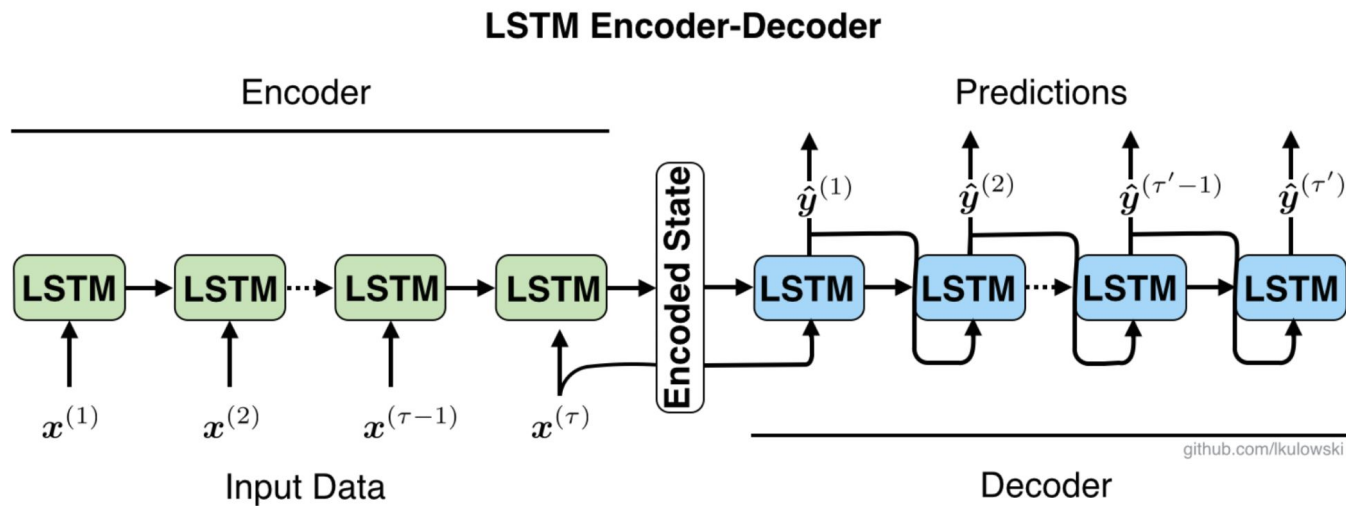


ІТМО

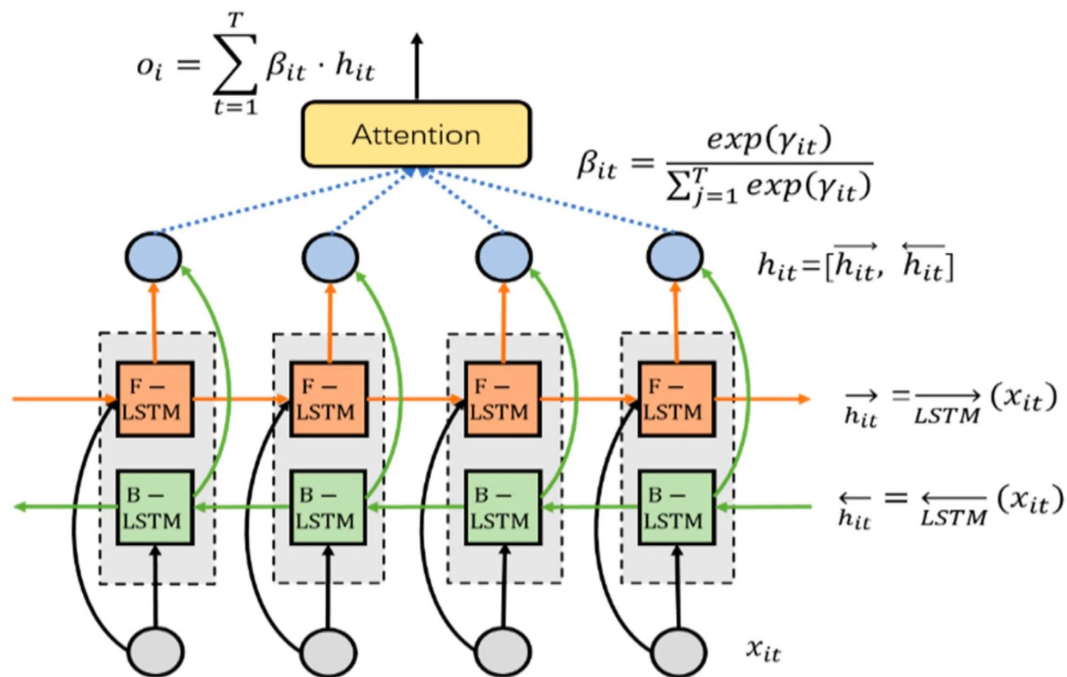
**Современные архитектуры
нейронных сетей**

Трансформеры

LSTM Encoder-Decoder



Attention



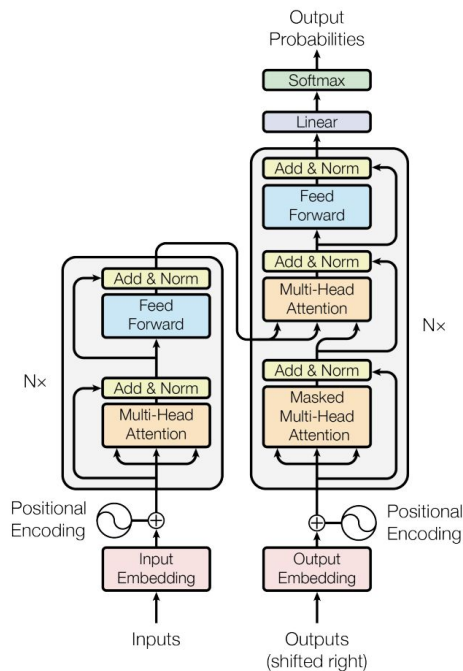
Проблемы рекуррентности



- 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

Всё, что нас не убивает, делает нас сильнее

Токенизация

всё	что	нас	не	убивает	делает	нас	сильнее
-----	-----	-----	----	---------	--------	-----	---------

Токенизация по символам



Всё, что нас не убивает, делает нас сильнее

Токенизация

в	с	ё		ч	т	о		н	а	с		н	е		у	б	и	в	а	е	т		д	е	л	а	е	т		н	а	с		с	и	л	ь	н	е	е
---	---	---	--	---	---	---	--	---	---	---	--	---	---	--	---	---	---	---	---	---	---	--	---	---	---	---	---	---	--	---	---	---	--	---	---	---	---	---	---	---

Токенизация по под словам

Всё, что нас не убивает, делает нас сильнее

Токенизация

всё	что	нас	не	убивает	делает	нас	сильнее
-----	-----	-----	----	---------	--------	-----	---------

Слой Embedding

$W_E =$

All words, ~ 50k

aah	aardvark	aardwolf	aargh	ab	aback	abacterial	abacus	abalone	abandon	...	zygoid	zygomatic	zygomorphic	zygosis	zygote	zygotic	zyme	zymogen	zymosis	zzz
-8.7	-1.5	-4.8	+6.9	-9.2	+9.1	-2.9	-2.8	-9.6	-6.2	...	-2.0	+8.5	-7.9	+8.8	+7.3	-0.9	-3.4	-5.3	+2.3	-9.2
-9.6	-1.4	-8.6	-4.9	-5.5	-4.9	-7.3	-9.7	-7.6	+2.3	...	+9.4	+9.7	-1.8	-6.7	+2.7	-0.2	+9.7	-8.6	+5.6	-4.2
-5.1	+3.2	-5.0	+3.3	+0.3	-1.5	+1.1	-4.2	+4.1	-1.7	...	-2.8	+6.5	+8.4	-9.0	-5.3	-3.0	+6.2	+9.6	+9.3	+8.0
-4.0	+9.7	-5.0	-7.8	+8.9	-5.3	+3.8	-8.7	+4.6	+7.6	...	-4.5	-2.4	-2.5	+4.9	-5.2	-6.5	-1.0	-3.9	+6.7	-5.2
+0.0	+8.8	+2.7	+7.3	+8.7	+5.0	+4.0	+9.3	+9.8	-1.0	...	-8.5	-4.1	-6.9	-1.6	-7.3	+2.1	-2.3	+7.8	+9.3	+0.9
-4.5	+1.8	+7.9	-1.8	+1.0	-4.5	-0.9	-1.9	-5.0	+0.1	...	-3.8	-2.5	+0.5	+5.0	-3.3	+8.4	+7.2	-8.9	-4.9	-1.1
-7.8	-3.0	+4.8	+3.6	+2.4	+4.2	-5.8	-3.1	+3.5	+7.5	...	+0.9	-4.3	-9.3	+4.2	-9.7	-2.5	+0.6	+8.4	-8.1	-1.9
-9.4	-3.1	+2.4	-4.4	-5.7	-7.6	+1.5	+3.9	+3.4	+8.9	...	-9.8	+2.9	+2.0	+1.8	+9.2	-9.6	+3.9	+6.2	+0.2	-3.3
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
+5.8	-8.0	-1.1	+0.4	+3.8	-8.1	-5.4	-1.8	+2.4	+7.7	...	+2.4	-7.3	+9.5	+7.4	+0.1	+8.4	+0.8	+8.4	+6.5	+9.3

Embedding matrix

Позиционное кодирование



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

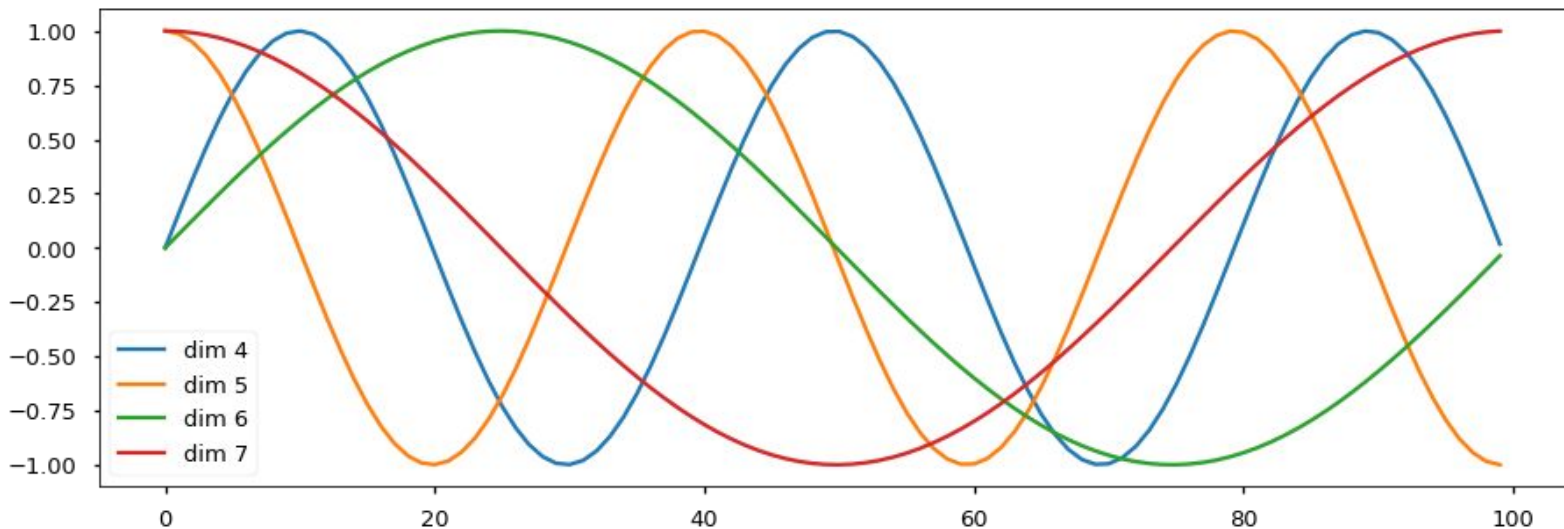
$$PE(pos, 2i + 1) = \cos\left(\frac{pos}{1000^{(2i+1)/\text{embedding_size}}}\right),$$

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

$$\hat{x} = x + \text{pos_encoding}$$

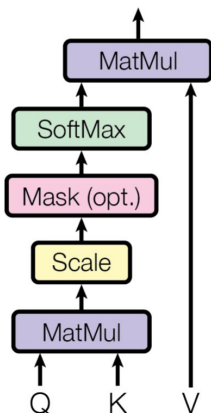
Позиционное кодирование

ИТМО

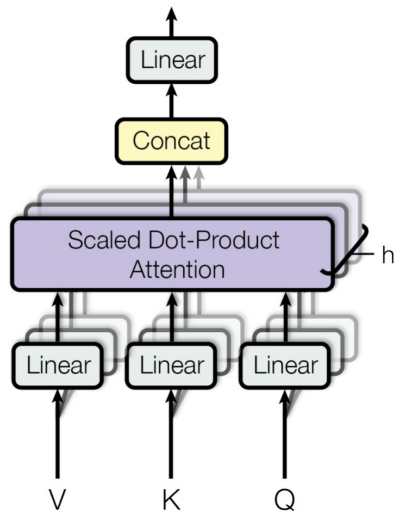


Attention

Scaled Dot-Product Attention



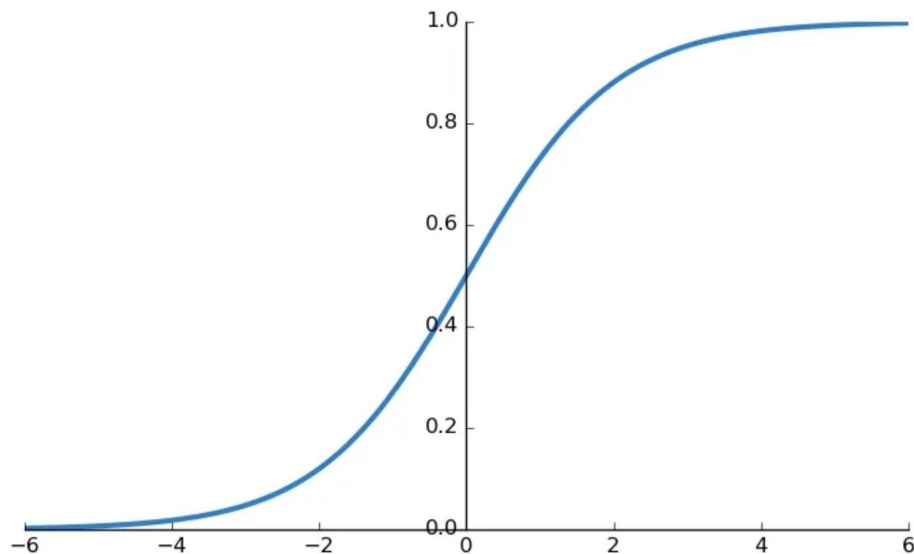
Multi-Head Attention



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

Attention

VITMO



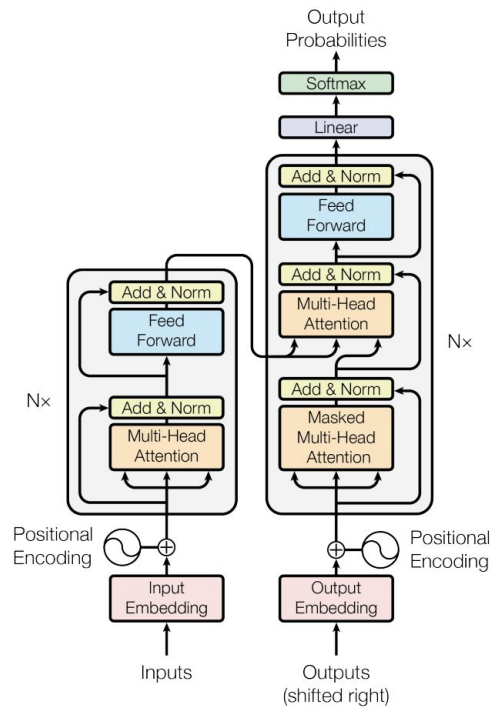
MultiHead Attention



$$\text{MultiHeadAttention} = \text{Concat}(\text{head}_1, \text{head}_2, \dots, \text{head}_h)W^O$$

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

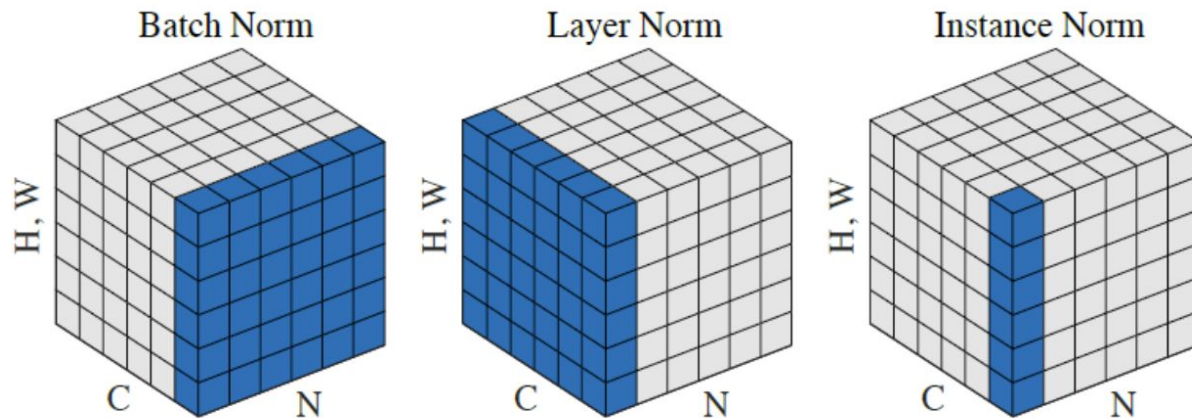
Encoder и Decoder



LayerNorm

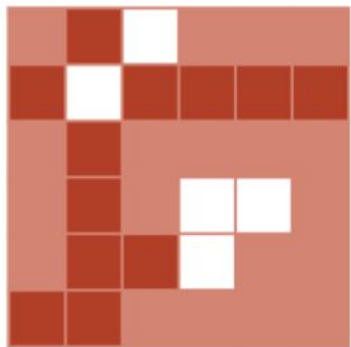
Формула общая с пакетной нормализацией

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



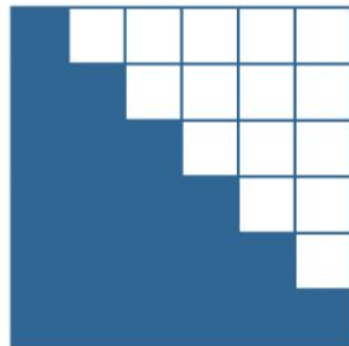
Masked MultiHead Attention

VITMO

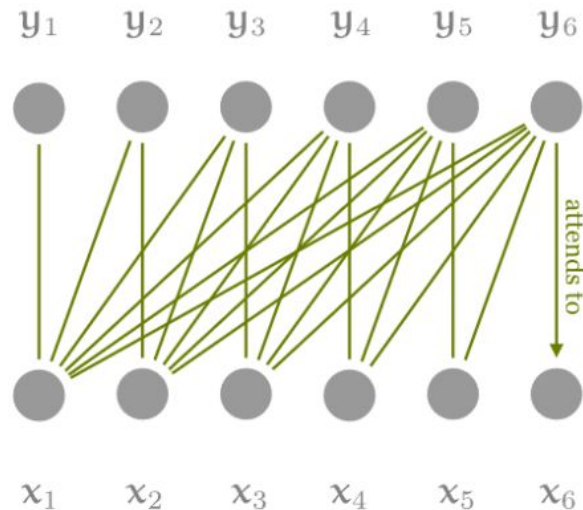


raw attention weights

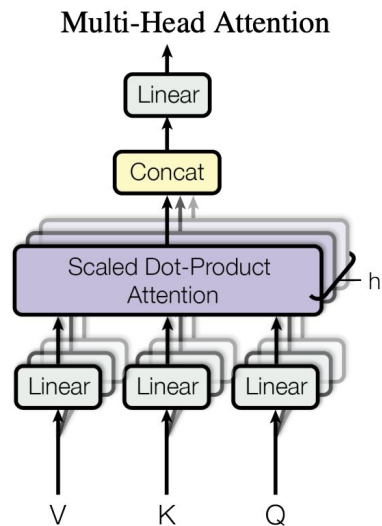
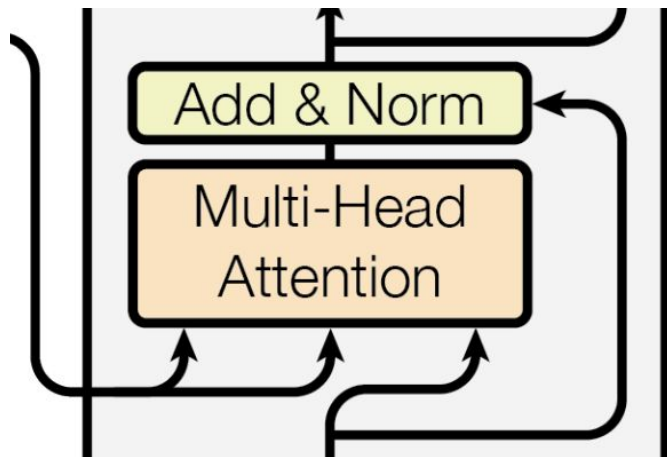
\otimes



mask



Cross attention





Softmax без температуры
(обычный)

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

Softmax с
температурой

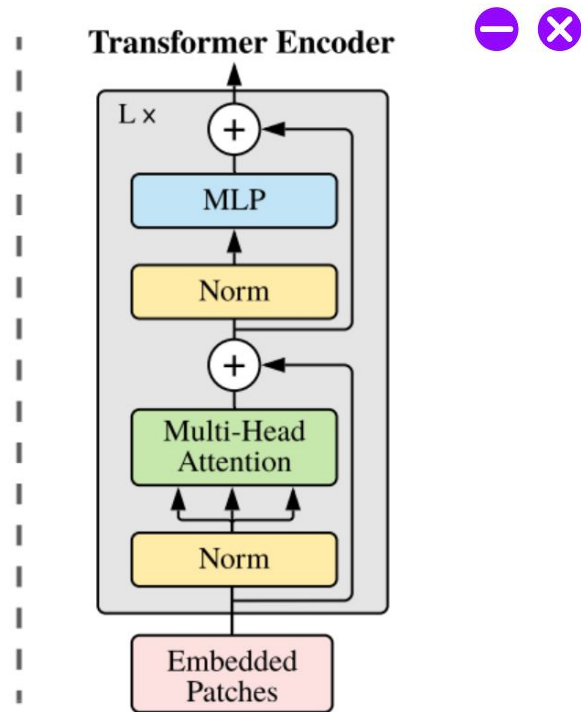
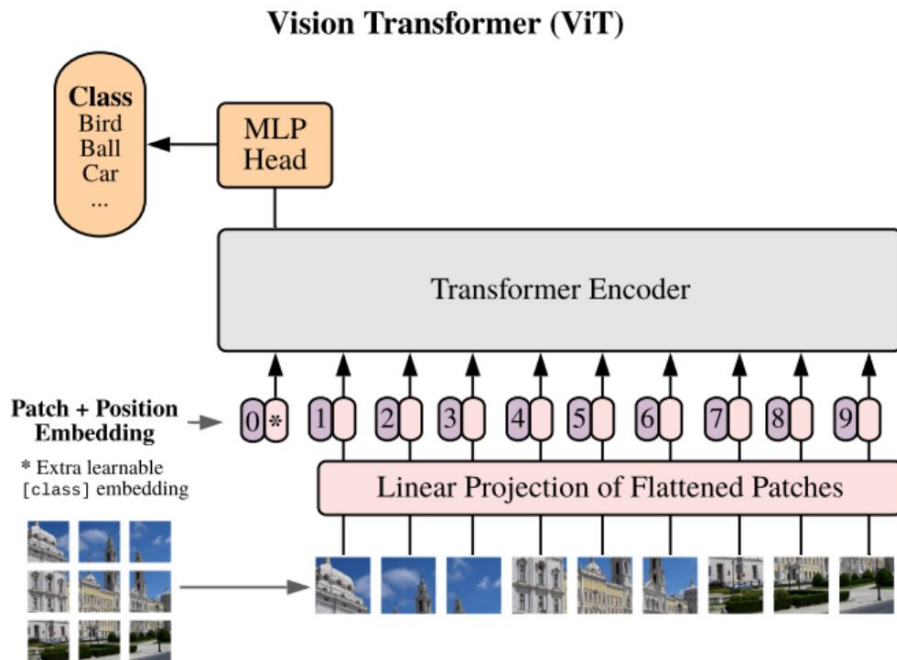
$$\textit{Softmax}_T = \frac{e^{x_k/T}}{\sum_{i=0}^{n-1} e^{x_i/T}}$$

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



Обучаемое позиционное кодирование

