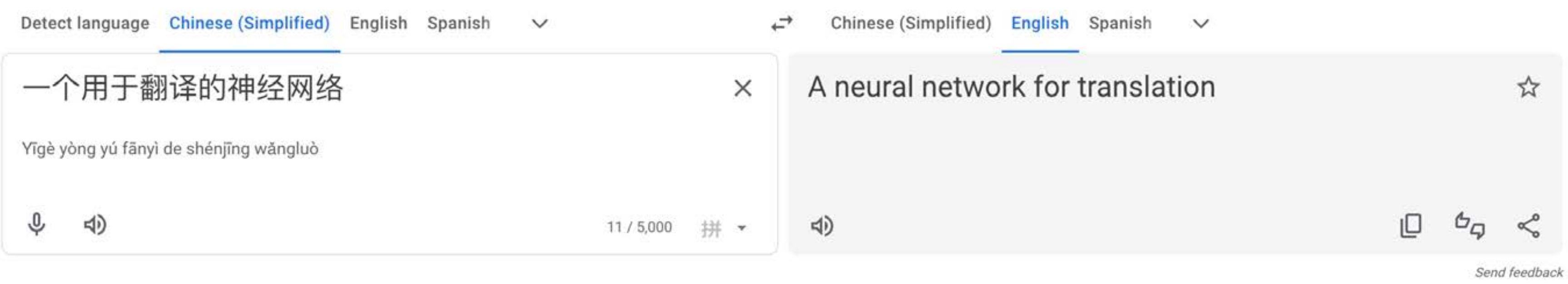
Part 02

变形金刚

- Seq2Seq
- Transformer

2.1 序列到序列

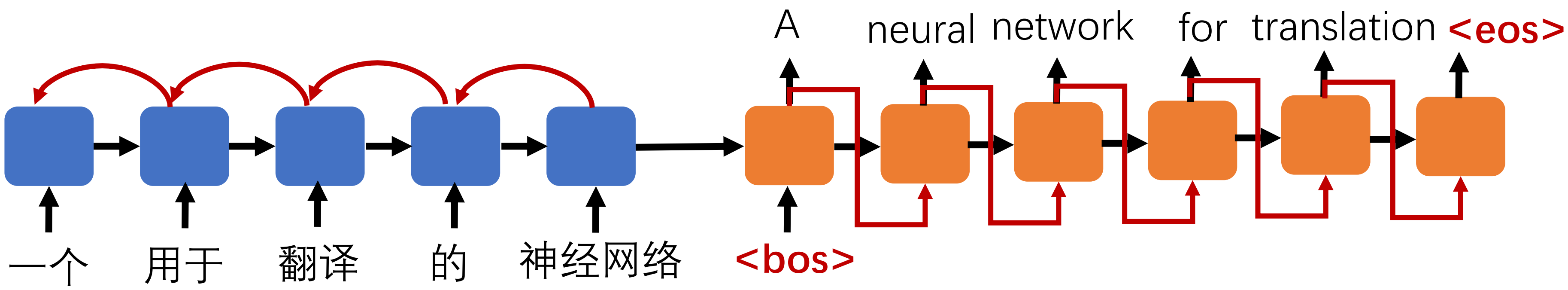


一个用于翻译的神经网络

Yīgè yòng yú fānyì de shénjīng wǎngluò

A neural network for translation

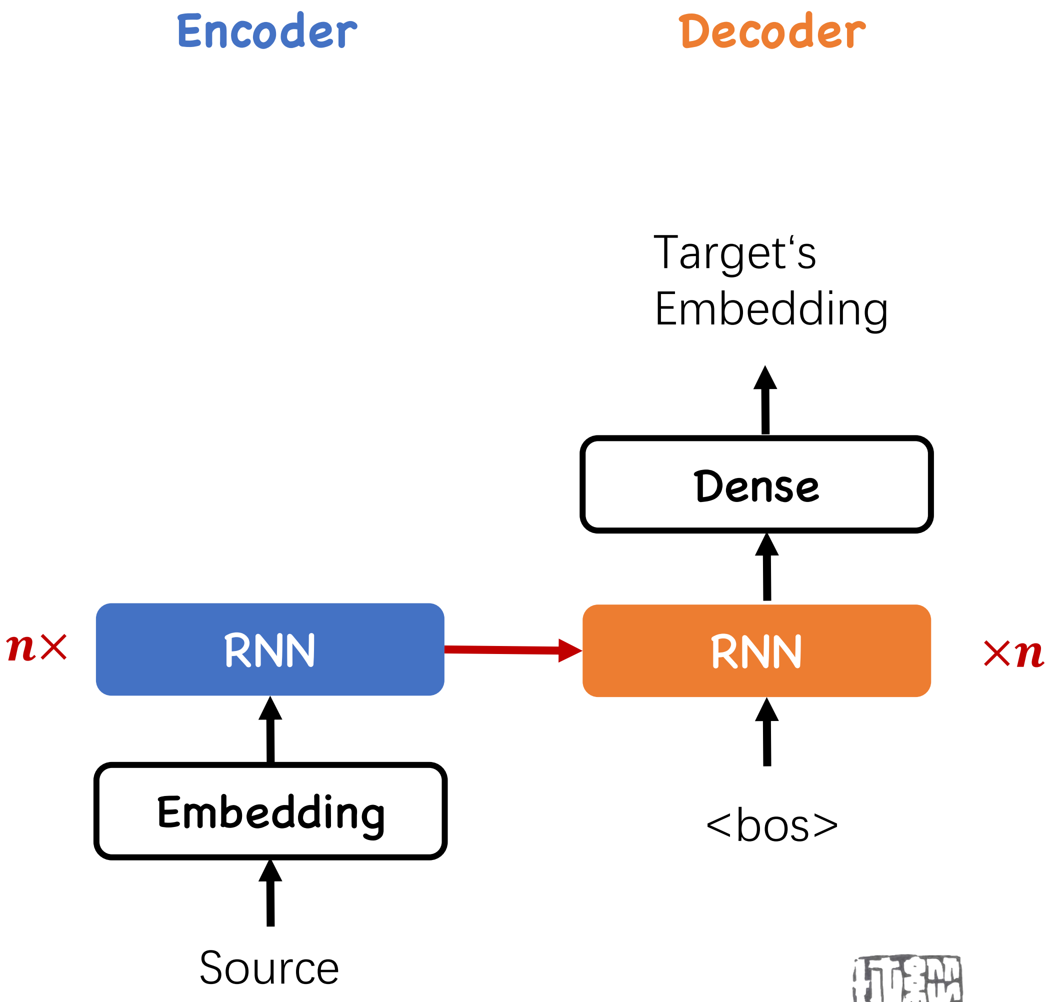
- 序列到序列的任务很常见：翻译、听写、朗读、预测
- 需要掌握输入序列的完整信息，再进行输出：
  - 序列的长度可能不一样，顺序不对应
  - 编码器：尽可能包含序列的更多信息；解码器：将编码复原为目标序列



https://arxiv.org/pdf/1409.3215.pdf

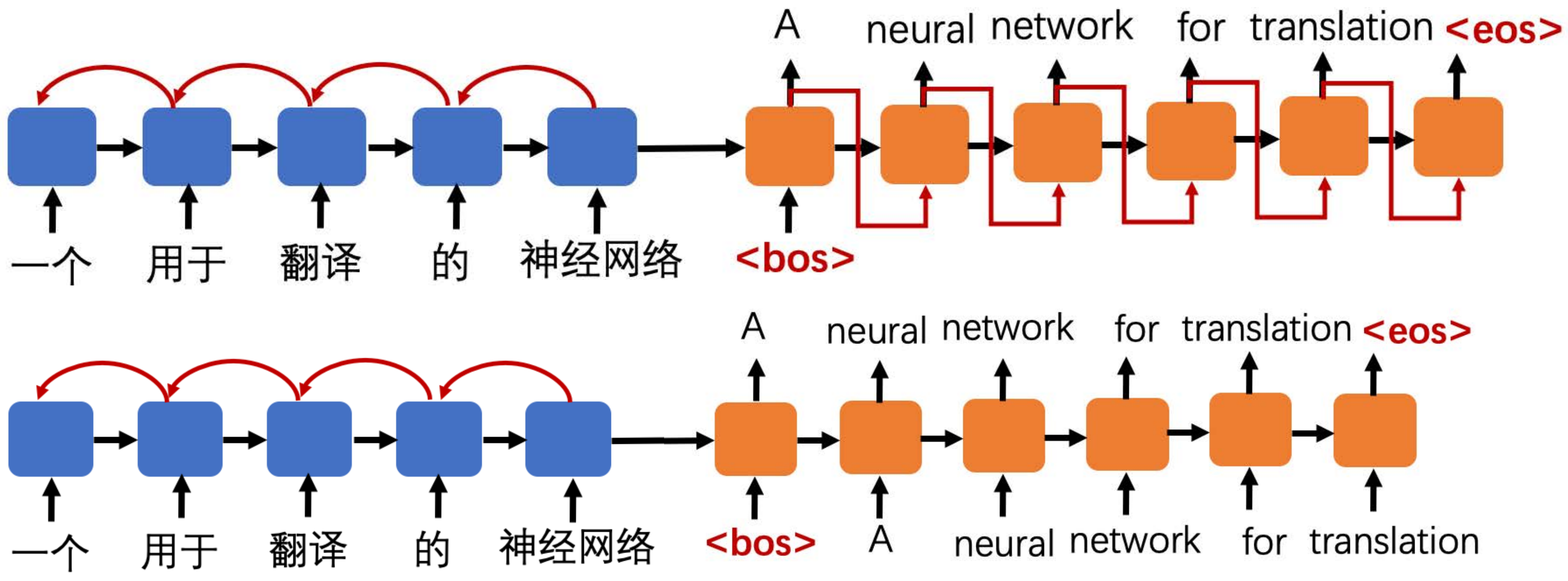
2.1 Seq2Seq （sequence to sequence）模型

- 编码器： 可以是双向RNN；没有输出
- 解码器： 单向RNN
- 编码器的最后一个隐藏层，是解码器的第一个隐藏层
- RNN可以是各种序列模型、可以堆叠很多层



2.1

Seq2Seq训练与评估



- 训练：
  - Encoder是原始序列
  - Decoder提供真实目标序列
  - 只预测one step

2.1 Transformer

- Encoder Decoder
- 所有的元件都以Attention为核心
- 位置编码
  - 在原始数据的Embedding上
  - 加一个表示位置的值
- Decoder的Attention被改装
- Transformer块
- 特殊的通信机制

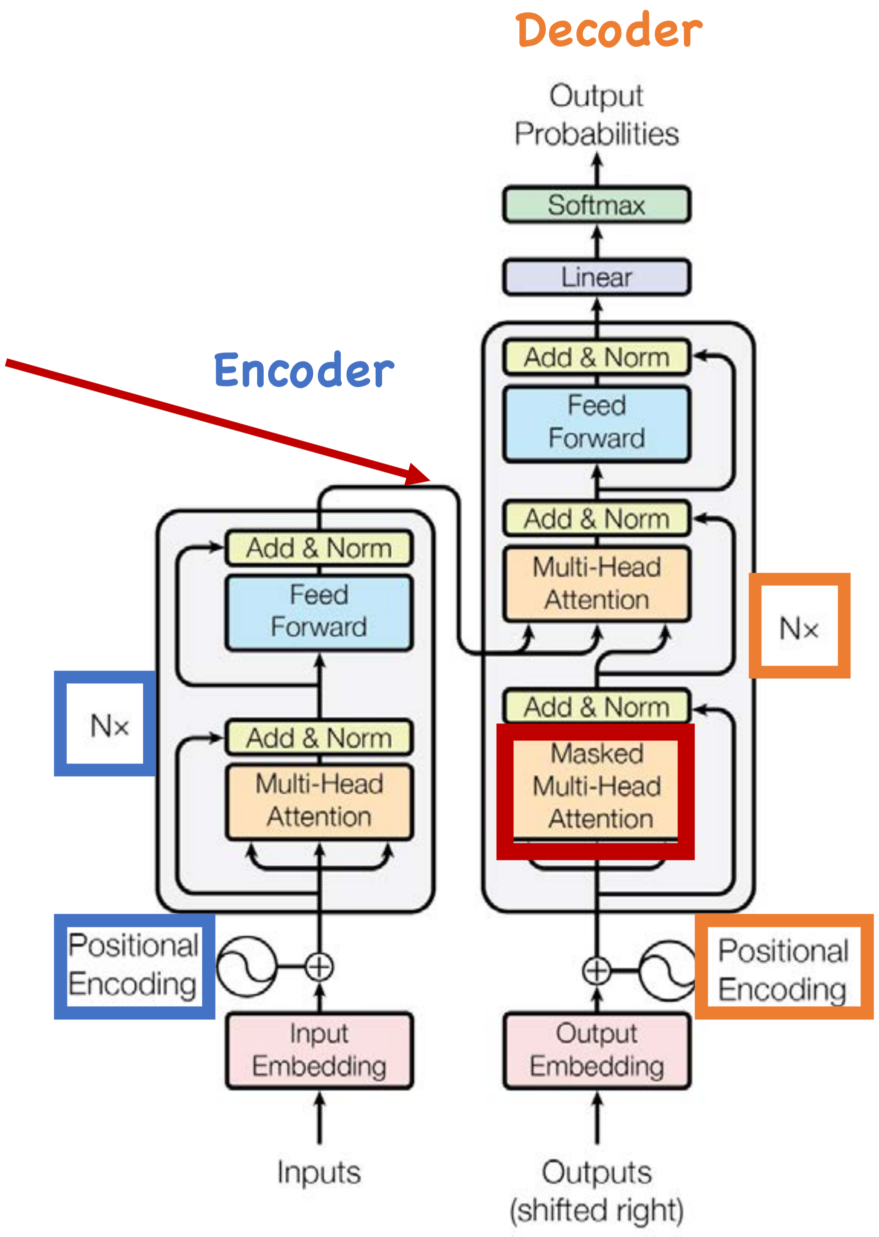
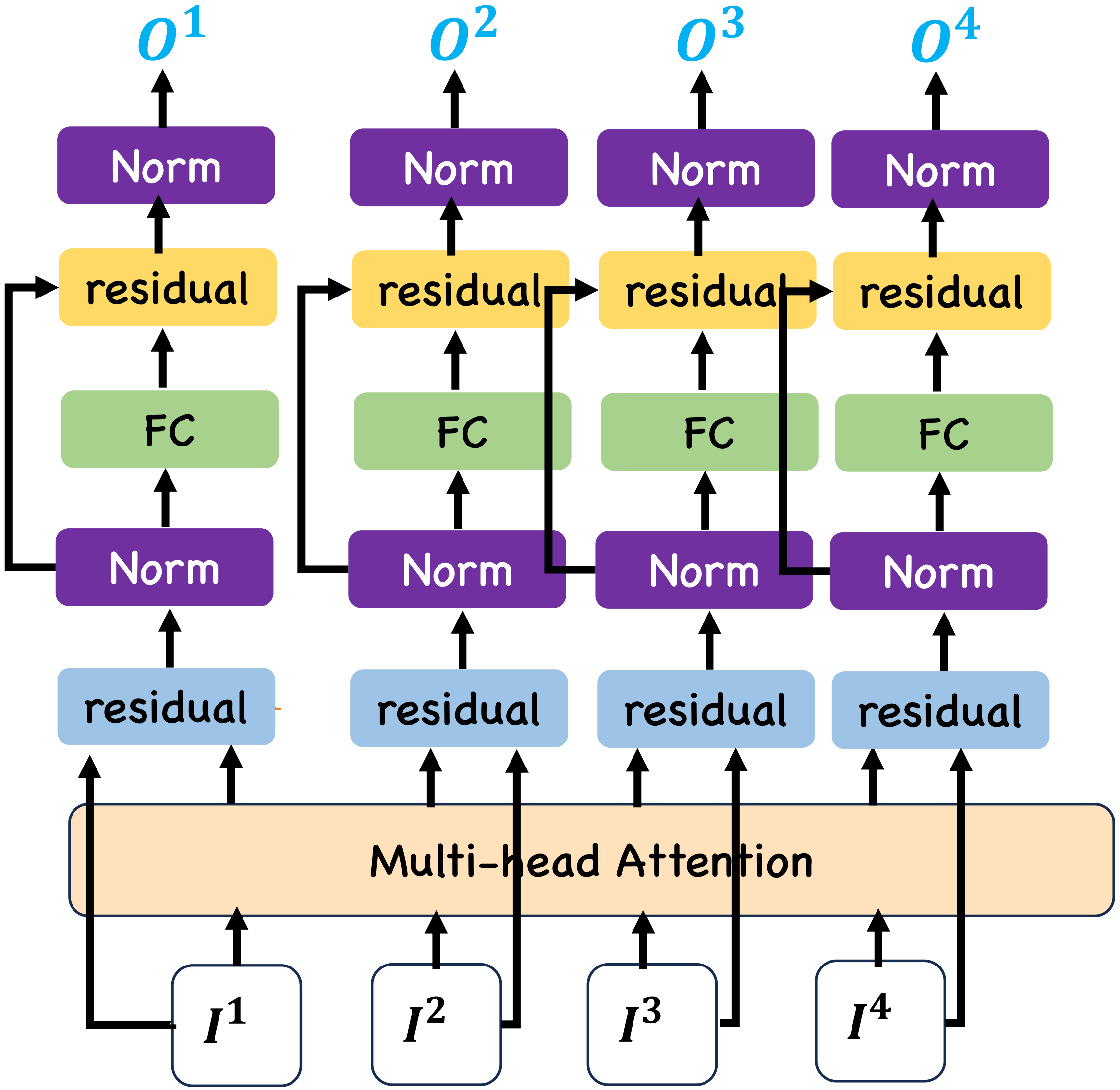
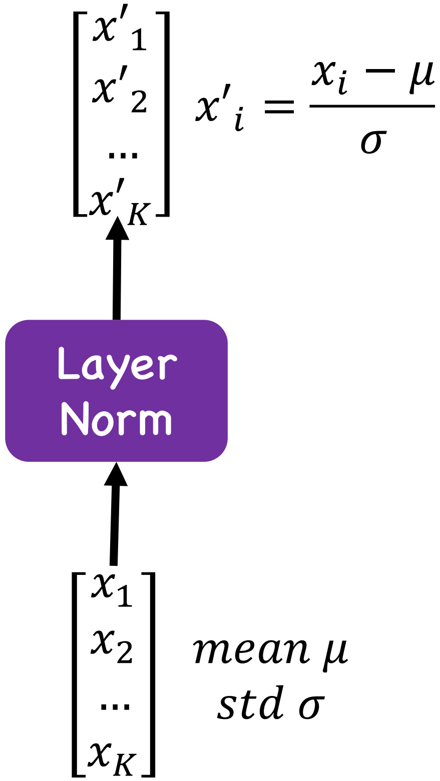


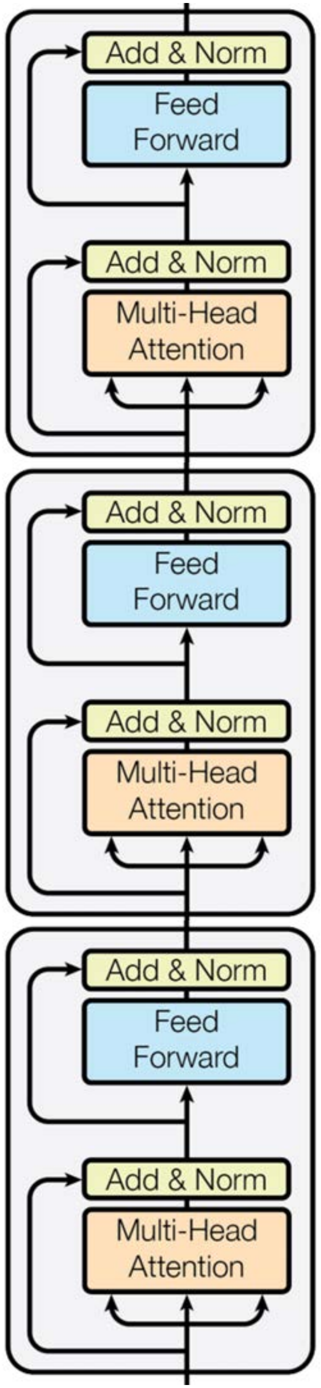
Figure 1: The Transformer - model architecture.

2.1 Transformer 的 Encoder



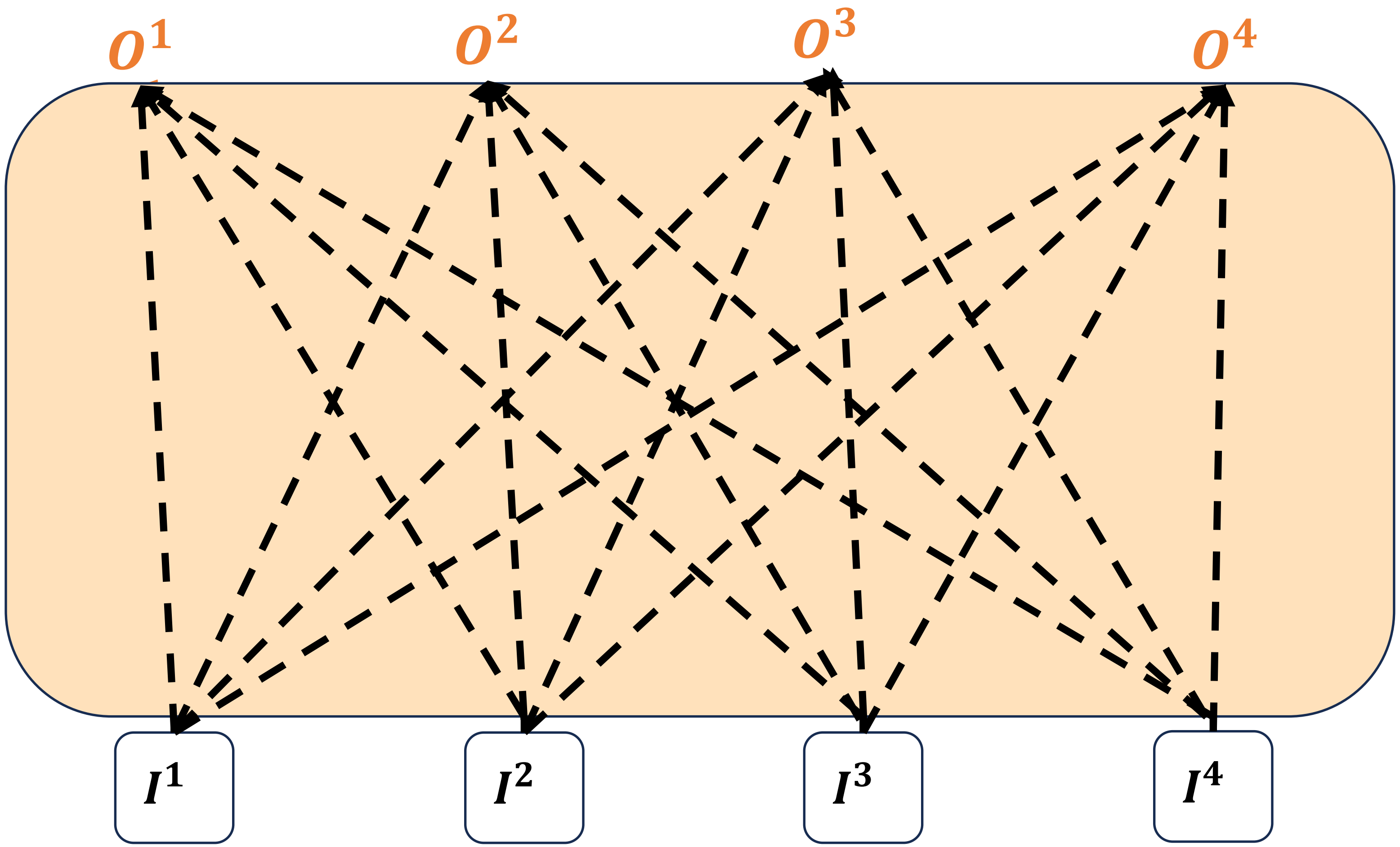

$\begin{bmatrix} x'_1 \\ x'_2 \\ \cdots \\ x'_K \end{bmatrix}$
Layer Norm
$\begin{bmatrix} x_1 \\ x_2 \\ \cdots \\ x_K \end{bmatrix}$
mean μ
std σ

$$x'_i = \frac{x_i - \mu}{\sigma}$$



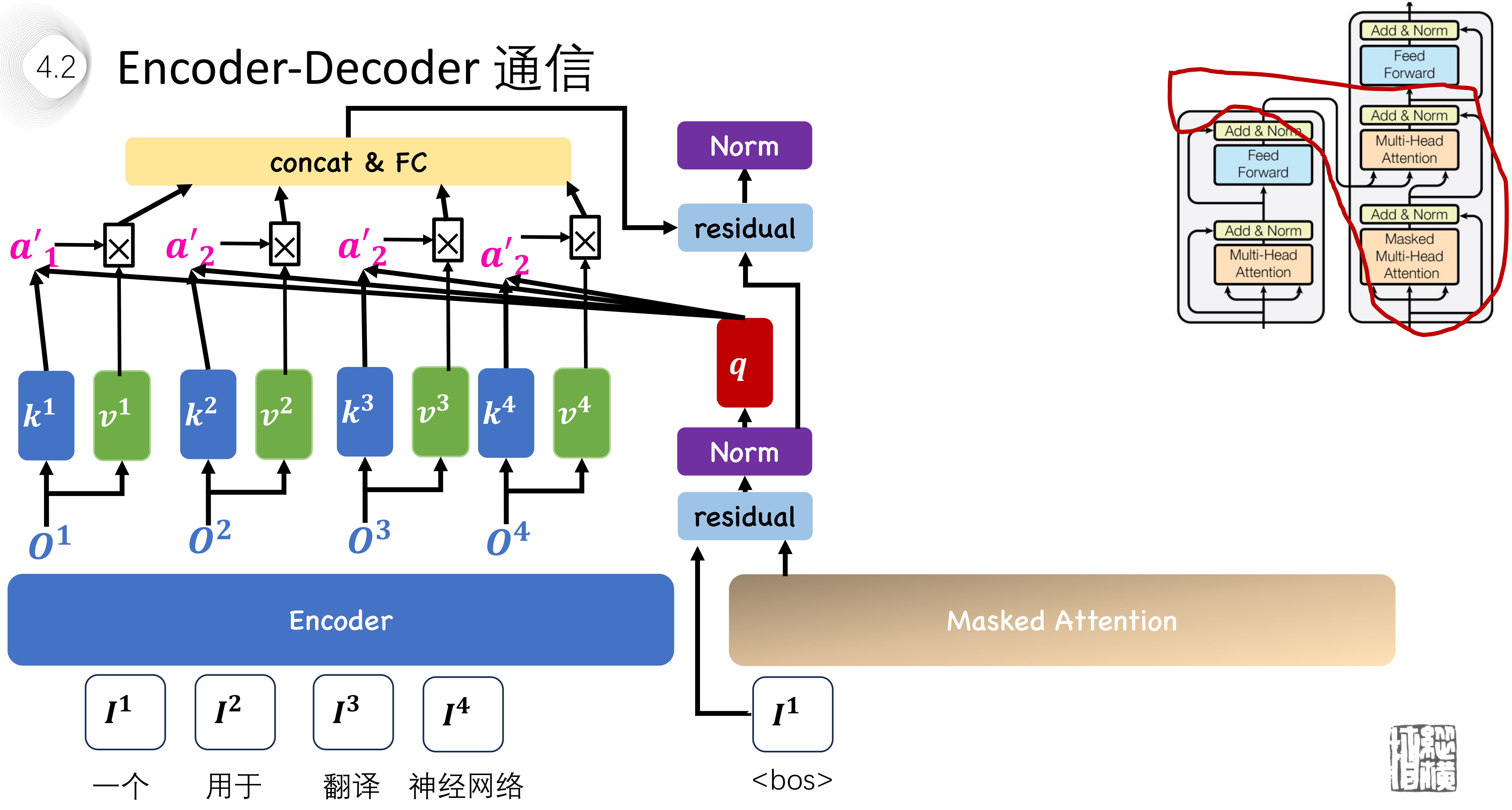
4.2

Masked Attention



- 我们稍微改动Attention，让它不能偷看后面的数据
- 生成时还是顺序的表现好

4.2 Encoder-Decoder 通信



2.3 Transformer的应用

Cong, Lin William, et al. "AlphaPortfolio: Direct construction through deep reinforcement learning and interpretable AI." *Available at SSRN 3554486* (2021).

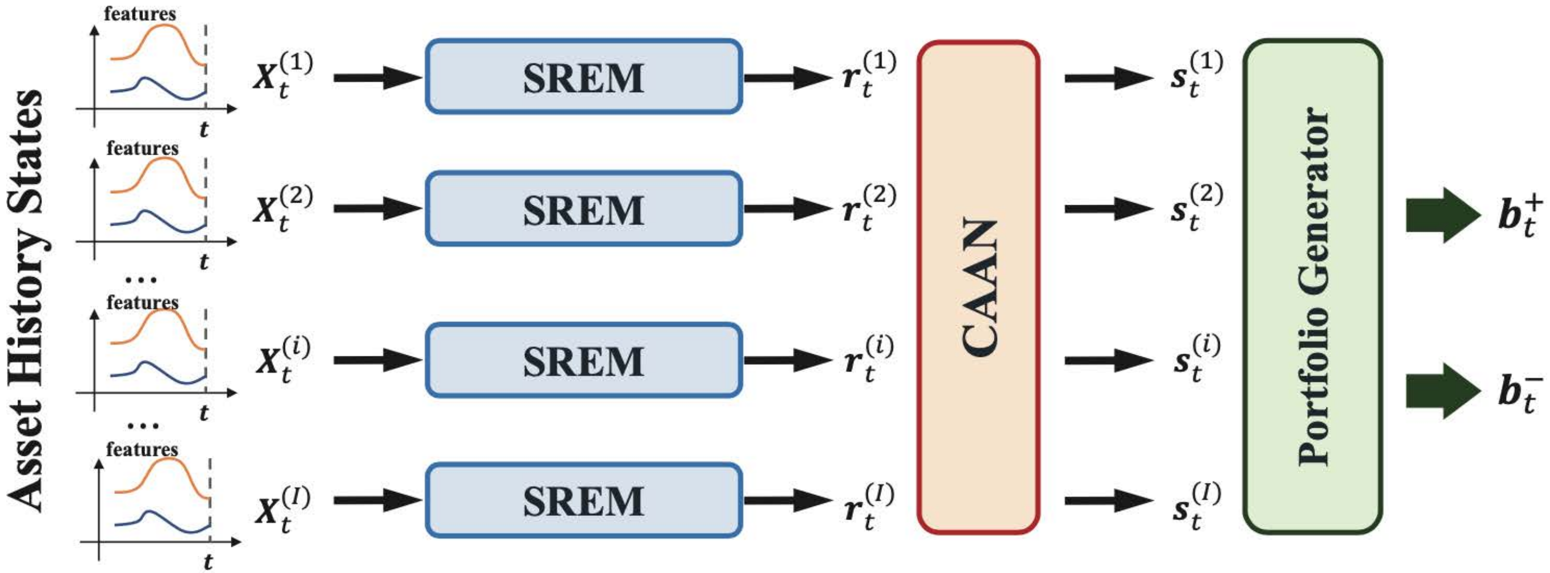
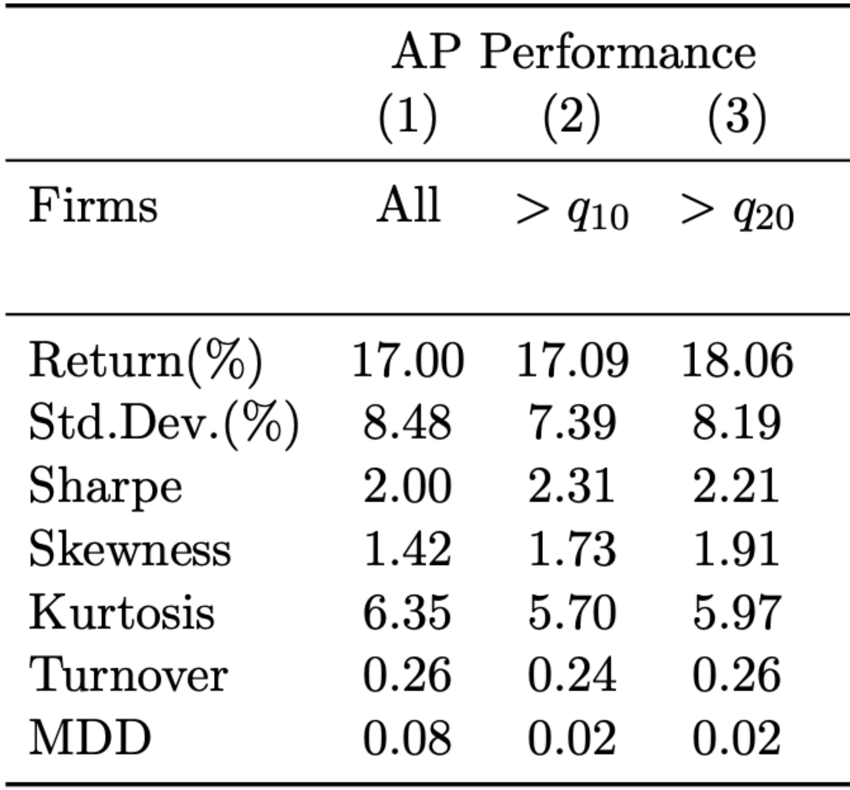


Figure 1: Overall Architecture of AP.



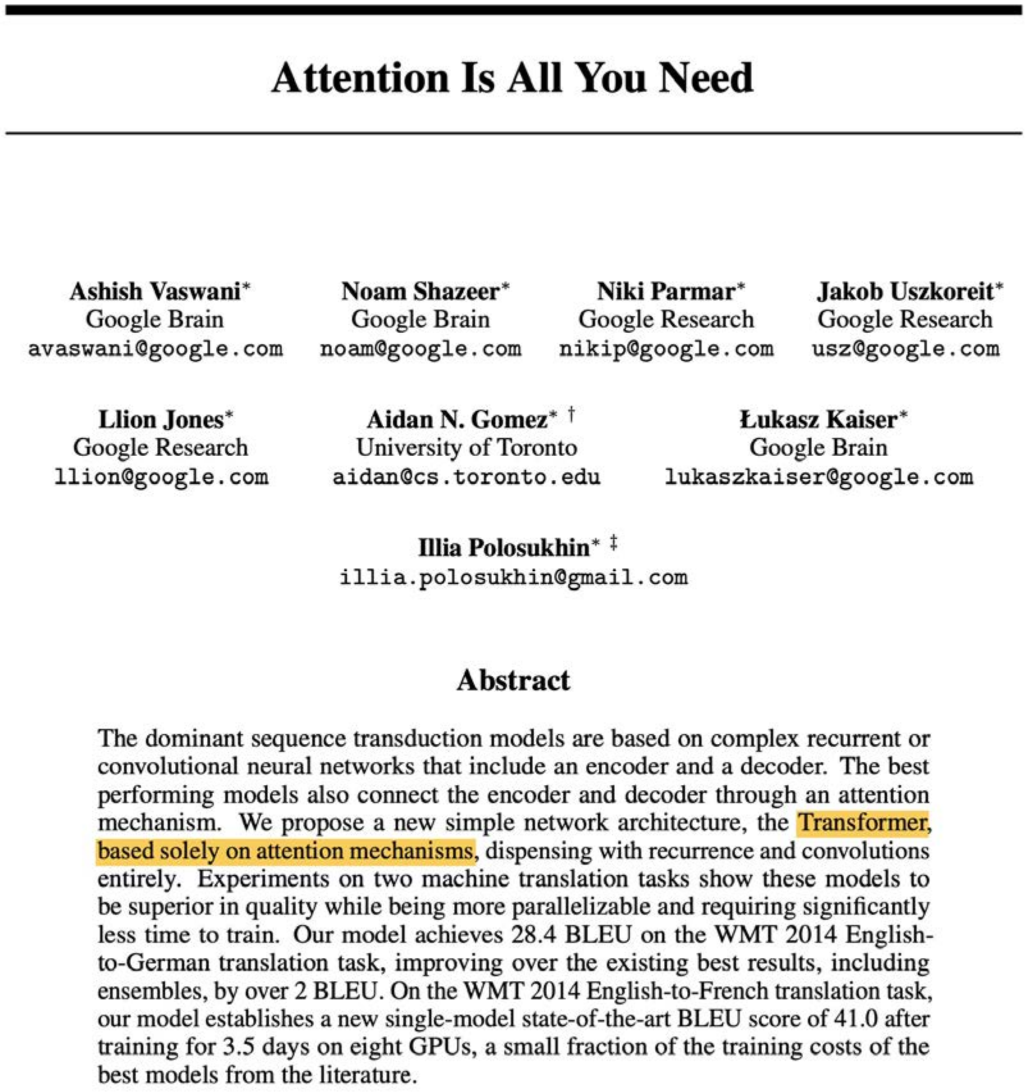
	AP Performance		
	(1)	(2)	(3)
Firms	All	$> q_{10}$	$> q_{20}$
Return(%)	17.00	17.09	18.06
Std.Dev.(%)	8.48	7.39	8.19
Sharpe	2.00	2.31	2.21
Skewness	1.42	1.73	1.91
Kurtosis	6.35	5.70	5.97
Turnover	0.26	0.24	0.26
MDD	0.08	0.02	0.02

SREM (Sequence Representation Extraction Module)
Transformer Encoder Based

CAAN Cross-Asset Attention Network

1.3 Attention is all you need （2017）

https://arxiv.org/abs/1706.03762



Attention Is All You Need

Ashish Vaswani*
Google Brain
avaswani@google.com

Noam Shazeer*
Google Brain
noam@google.com

Niki Parmar*
Google Research
nikip@google.com

Jakob Uszkoreit*
Google Research
usz@google.com

Llion Jones*
Google Research
llion@google.com

Aidan N. Gomez* †
University of Toronto
aidan@cs.toronto.edu

Łukasz Kaiser*
Google Brain
lukaszkaiser@google.com

Illia Polosukhin* ‡
illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.0 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature.

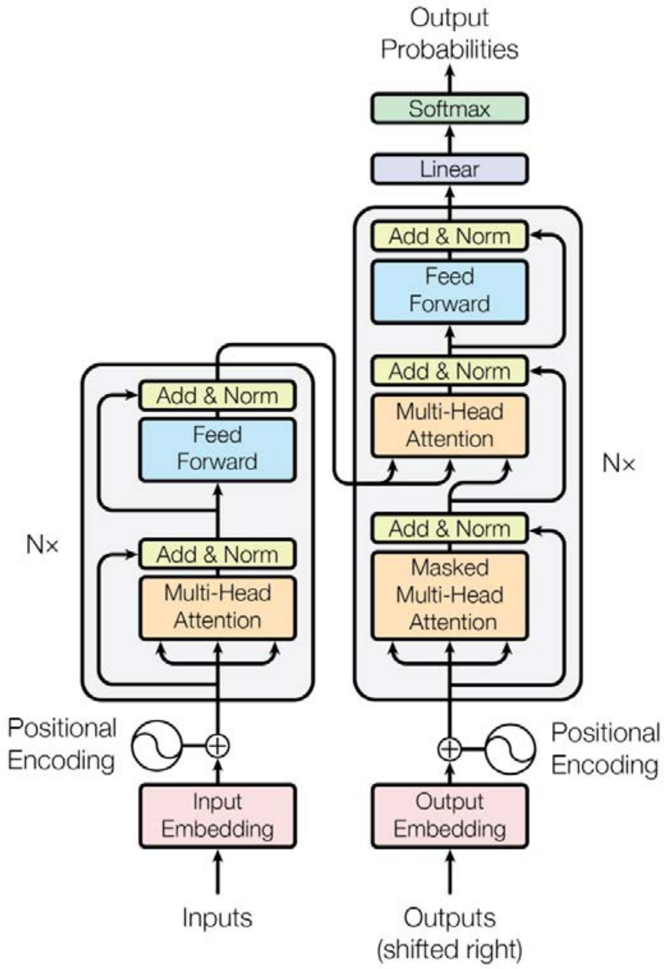


Figure 1: The Transformer - model architecture.