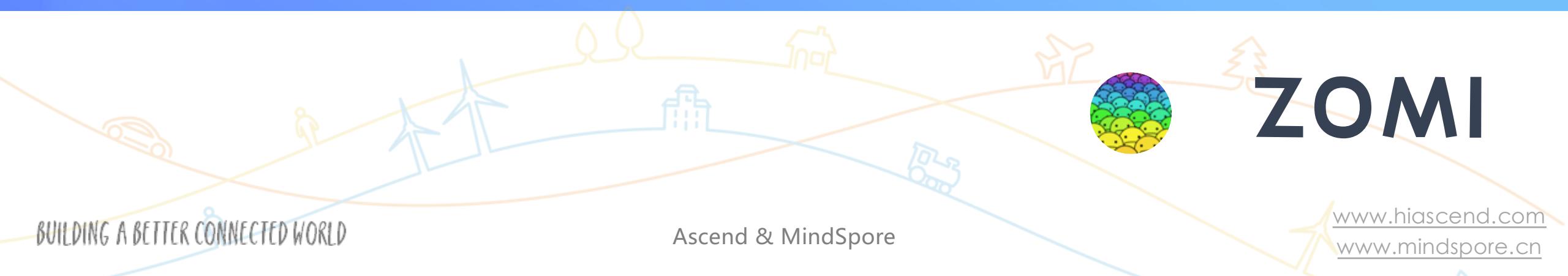


分布式训练系列

大模型算法结构



关于本内容

I. 内容背景

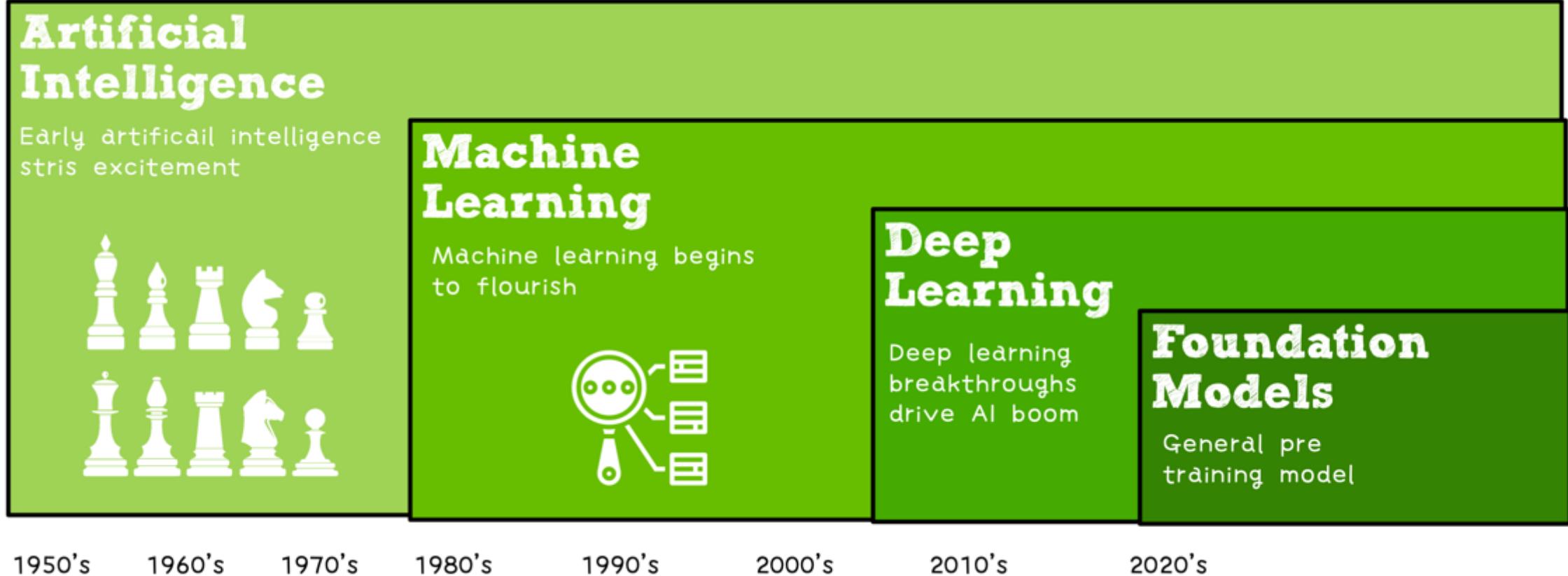
- AI集群+大模型+分布式训练系统

2. 具体内容

- 分布式+AI集群：服务器架构 – 集群软硬件通信 - 通信原语 - AI框架分布式功能
- 大模型与训练挑战：什么是大模型 – 大模型训练的四个挑战
- 大模型算法结构：大模型算法发展 – Transformer结构 – MOE结构
- SOTA大模型算法：BERT – GPT3 – Switch Transformer

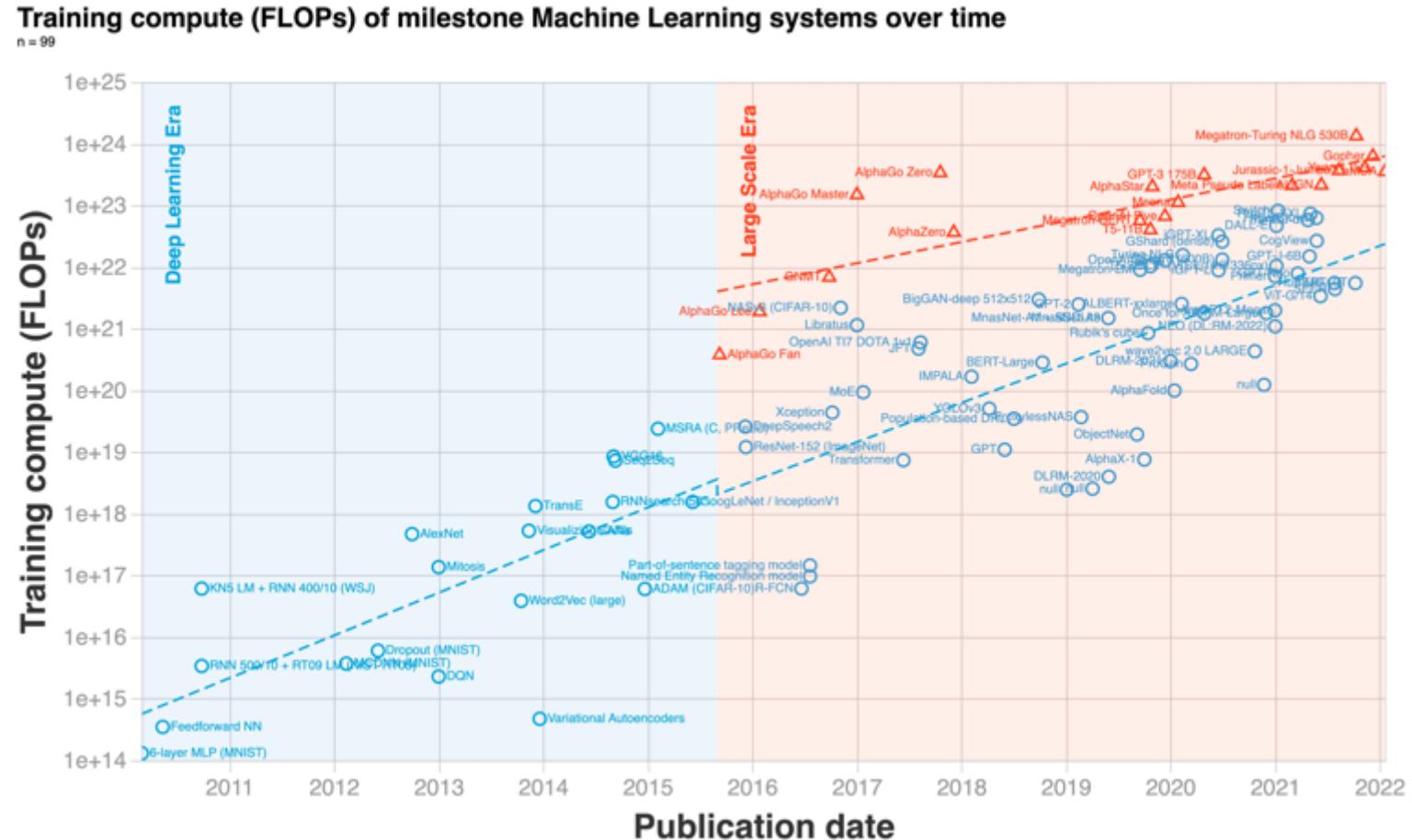
◦ 分布式并行：数据并行 – 张量并行 – 自动并行 – 多维混合并行

人工智能发展与大规模分布式训练关系



深度学习迎来大模型 (Foundation Models)

1. 自监督学习方法，可以减少数据标注，降低训练研发成本
2. 模型参数规模越大，有望进一步现有模型结构的精度局限突破
3. 解决模型碎片化，提供预训练方案
 - e.g. 语言模型 GPT-3
 - 8 张 V100，训练时长 36 年
 - 512 张 V100，训练近 7 个月



分布式训练与模型算法关系

- 深度学习训练耗时：

$$\text{训练耗时} = \text{训练数据规模} \times \text{单步计算量} / \text{计算速率}$$

模型相关

可变因素

- 计算速率：

$$\text{计算速率} = \text{单设备计算速率} \times \text{设备数} \times \text{多设备并行效率 (加速比)}$$

混合精度
算子融合
梯度累加

服务器架构
通信拓扑优化

数据并行
模型并行
流水并行

深度学习迎来大模型（ Foundation Models ）

谷歌Flan-T5诞生！1800种语言任务超大规模微调

夕小瑶的卖萌屋 2022-10-25 12:05 发表于四川

「多语言图像描述」最强评估基准XM3600来了！涵
盖36种语言

新智元 新智元 2022-10-24 13:13 发表于北京

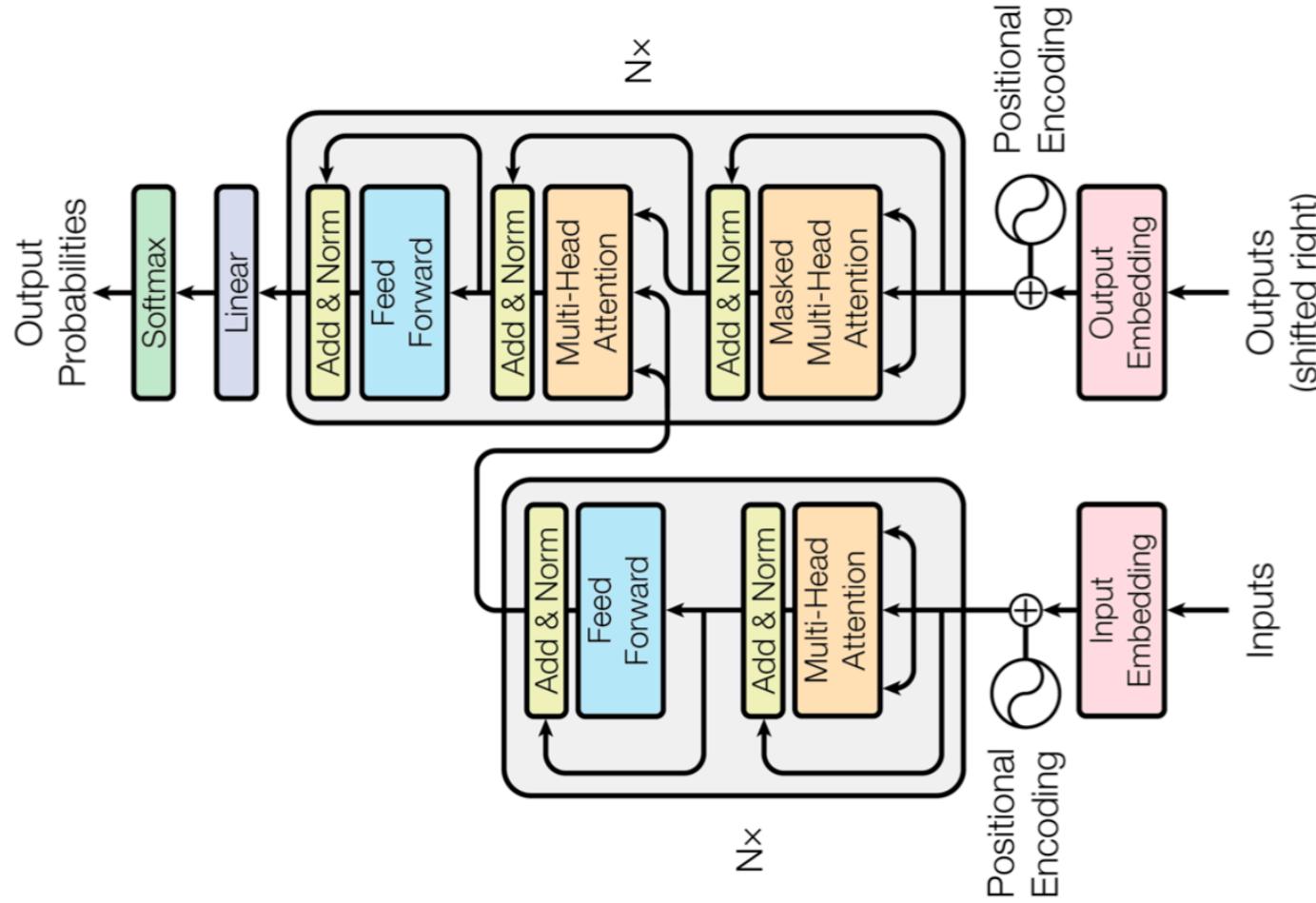
小扎亲自演示首个「闽南语」翻译系统！主攻3000种
无文字的语言

新智元 新智元 2022-10-24 13:13 发表于北京

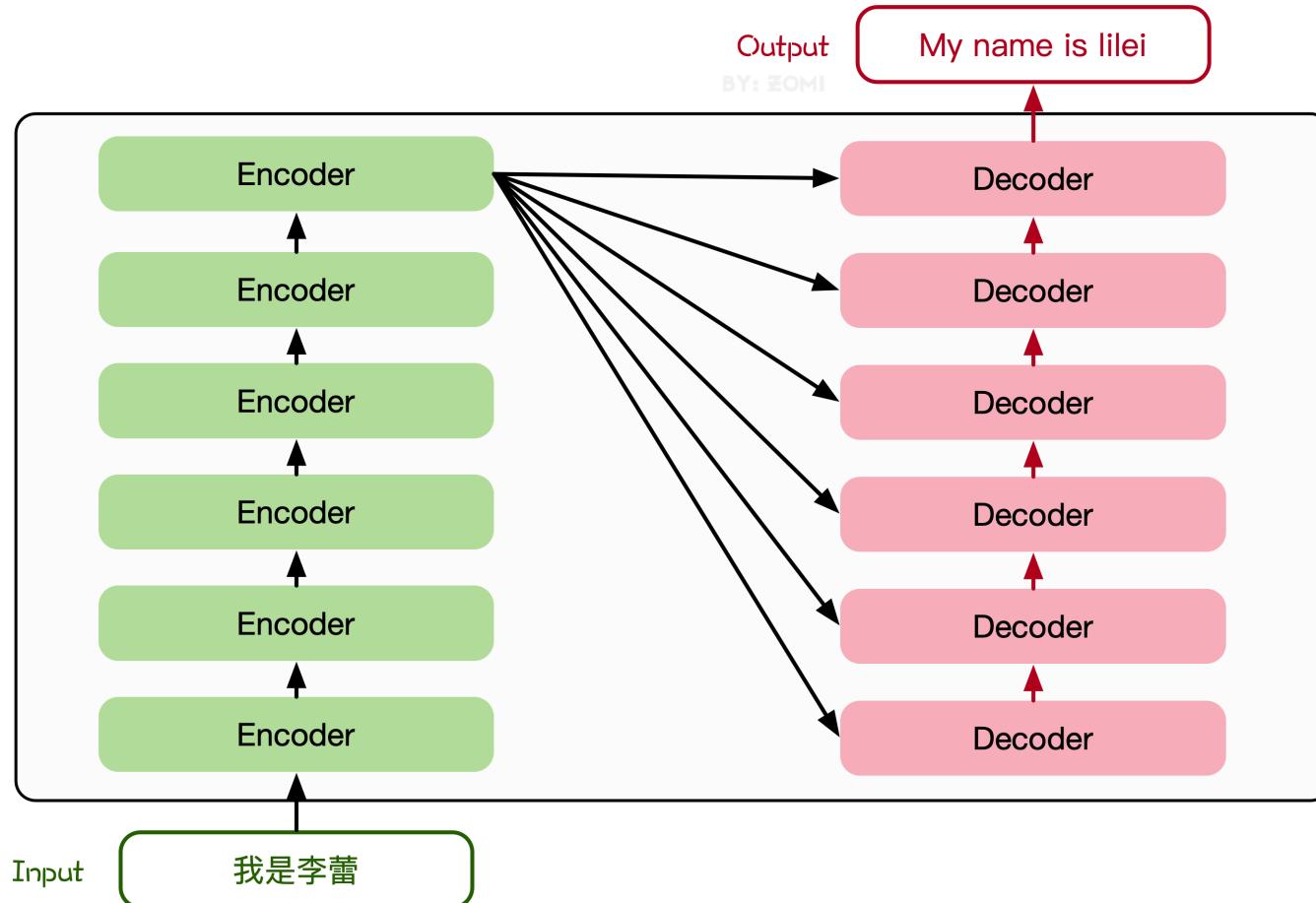
大模型结构演进

1. **Transformer** 取代RNN、CNN进入大模型时代
2. **MoE** 稀疏混合专家结构模型参数量进一步突破
3. **Bert** 突破十亿规模NLP大模型
4. **GPT3** 全新语言模型 1750 亿参数大模型
5. **Switch Transformer** 首个突破万亿大模型
6. **GLaM** 1.2万亿参数通用稀疏语言模型

Transformer 取代RNN、CNN进入大模型时代



Transformer 取代RNN、CNN进入大模型时代



Transformer 取代RNN、CNN进入大模型时代

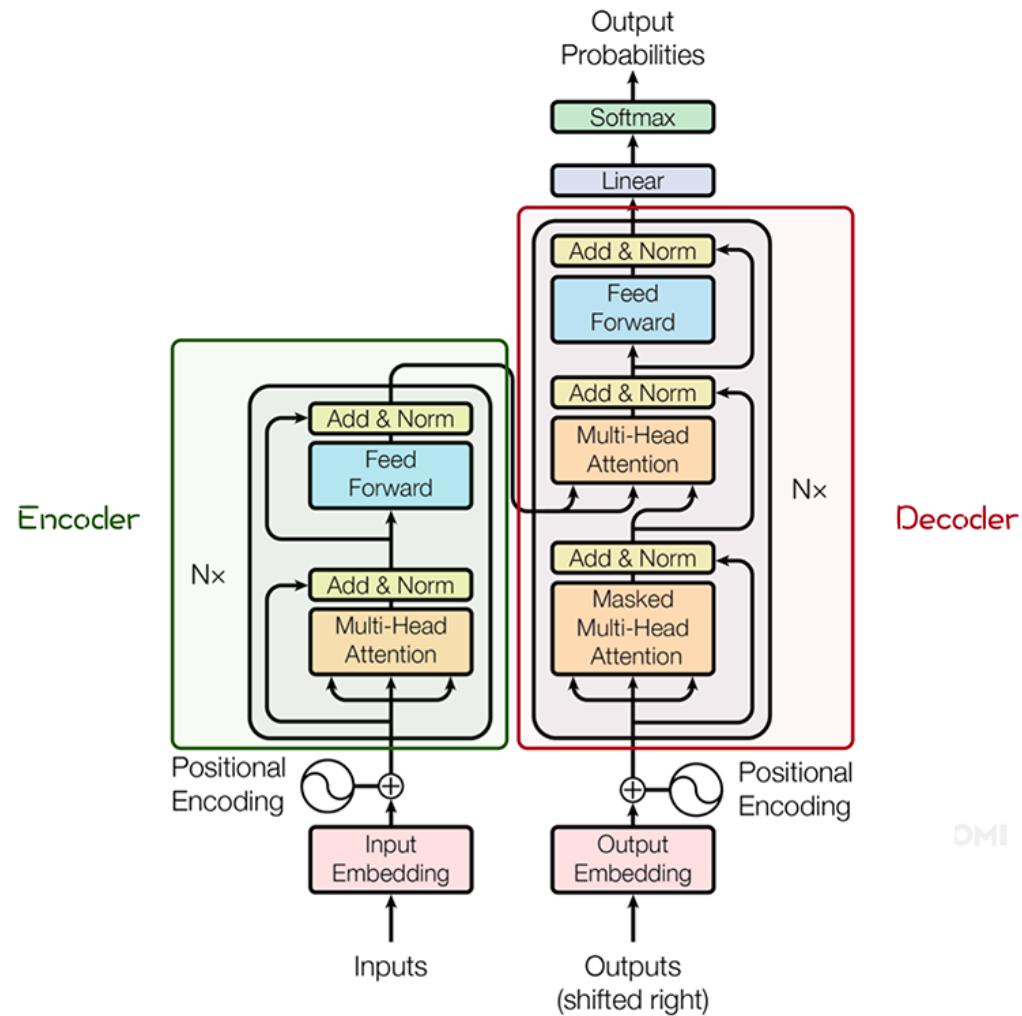
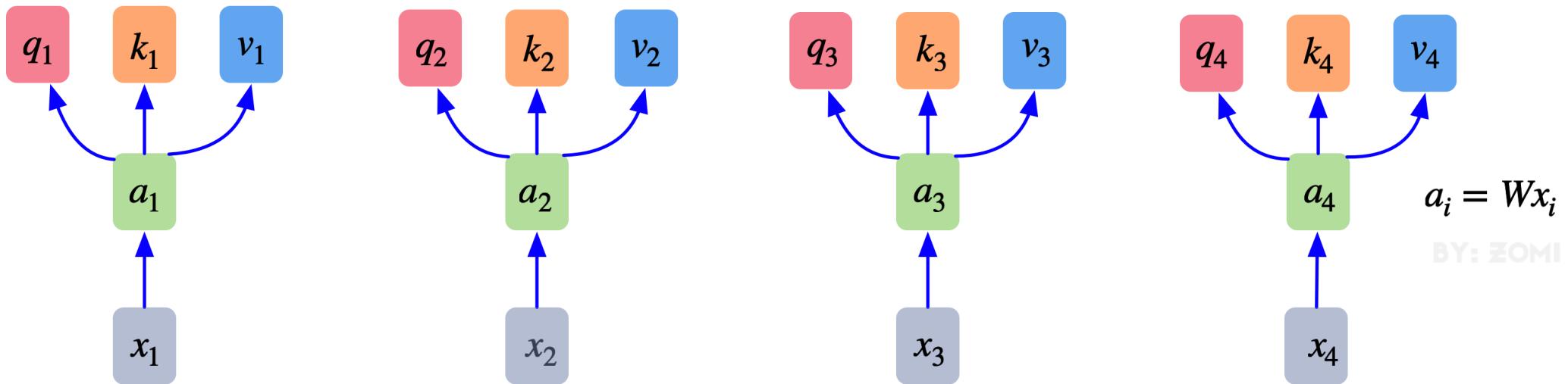


Figure 1: The Transformer - model architecture.

Transformer 取代RNN、CNN进入大模型时代

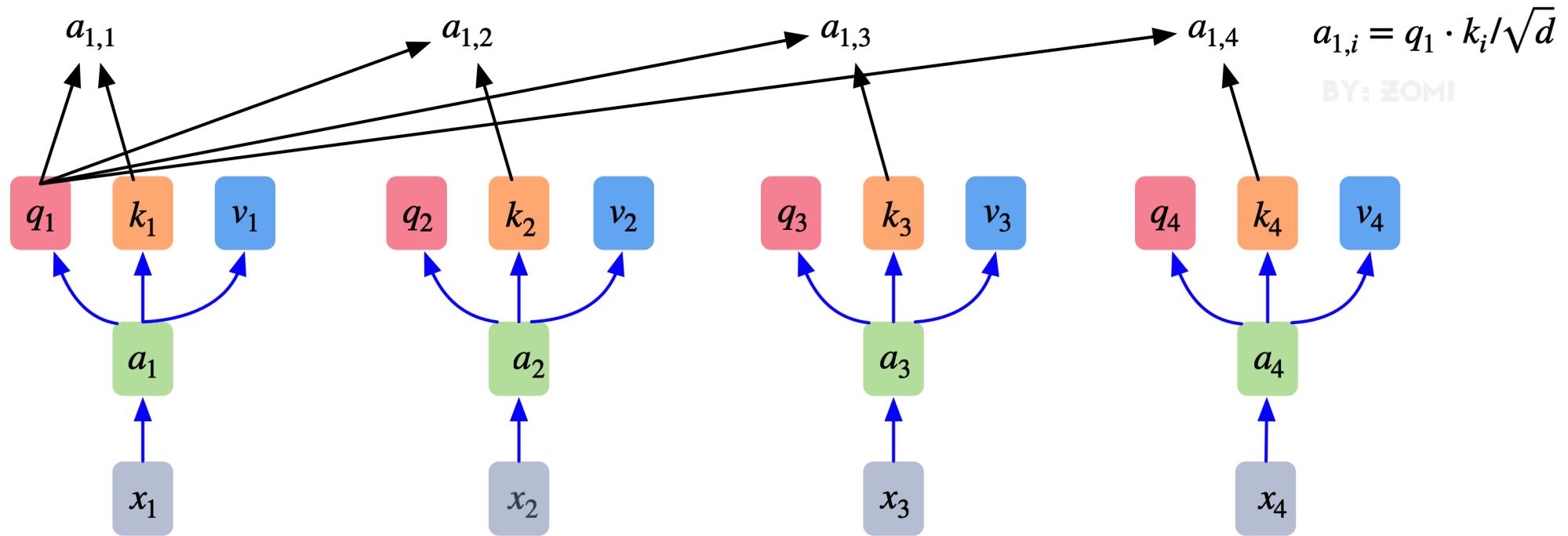
Attention模块：核心内容是为输入向量的每个单词学习一个权重。通过给定一个任务相关的查询向量Query向量，计算Query和各个Key的相似性或者相关性得到注意力分布，即得到每个Key对应Value的权重系数，然后对Value进行加权求和得到最终的Attention数值。



BY: ZOMI

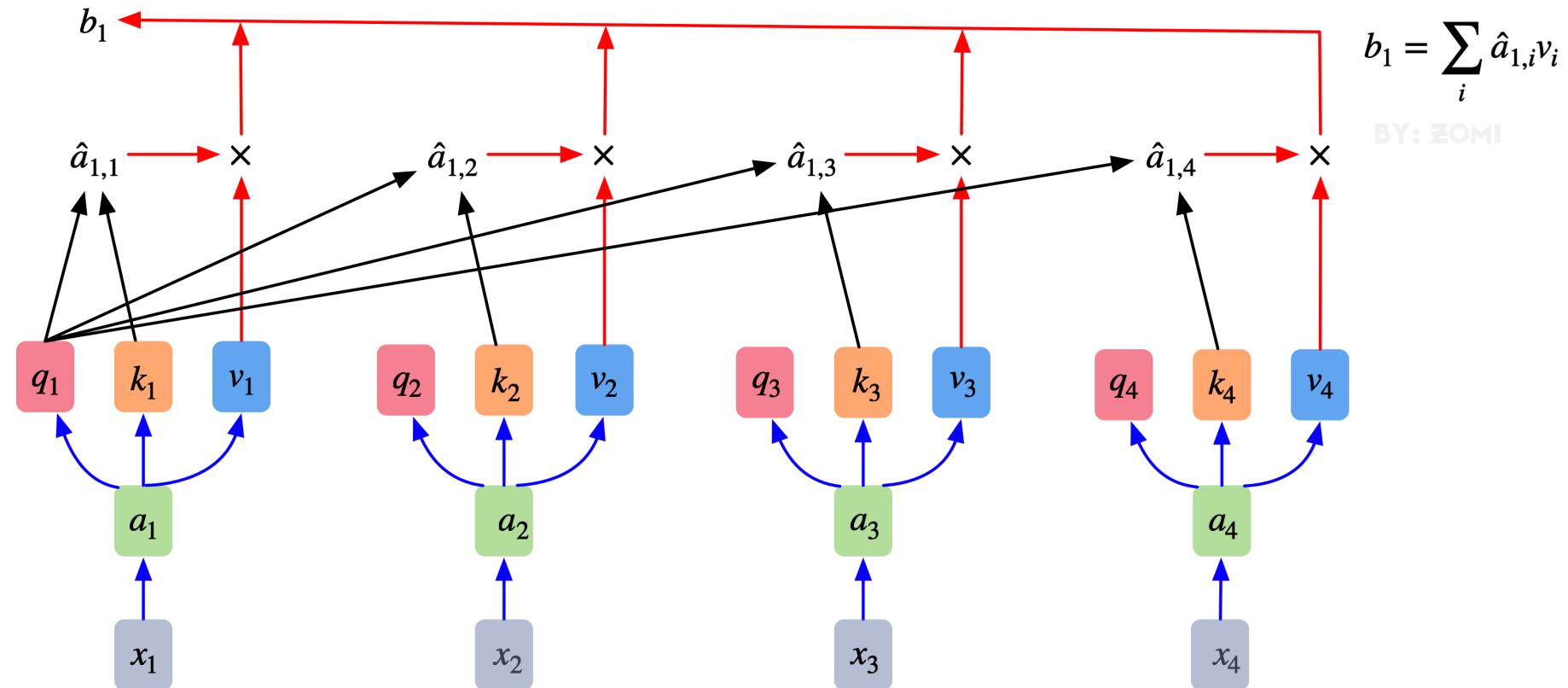
Transformer 取代RNN、CNN进入大模型时代

需要对Q和K进行点乘并除以维度的平方根，对所有向量的结果进行Softmax处理，通过公式(2)的操作，我们获得了向量之间的关系权重。



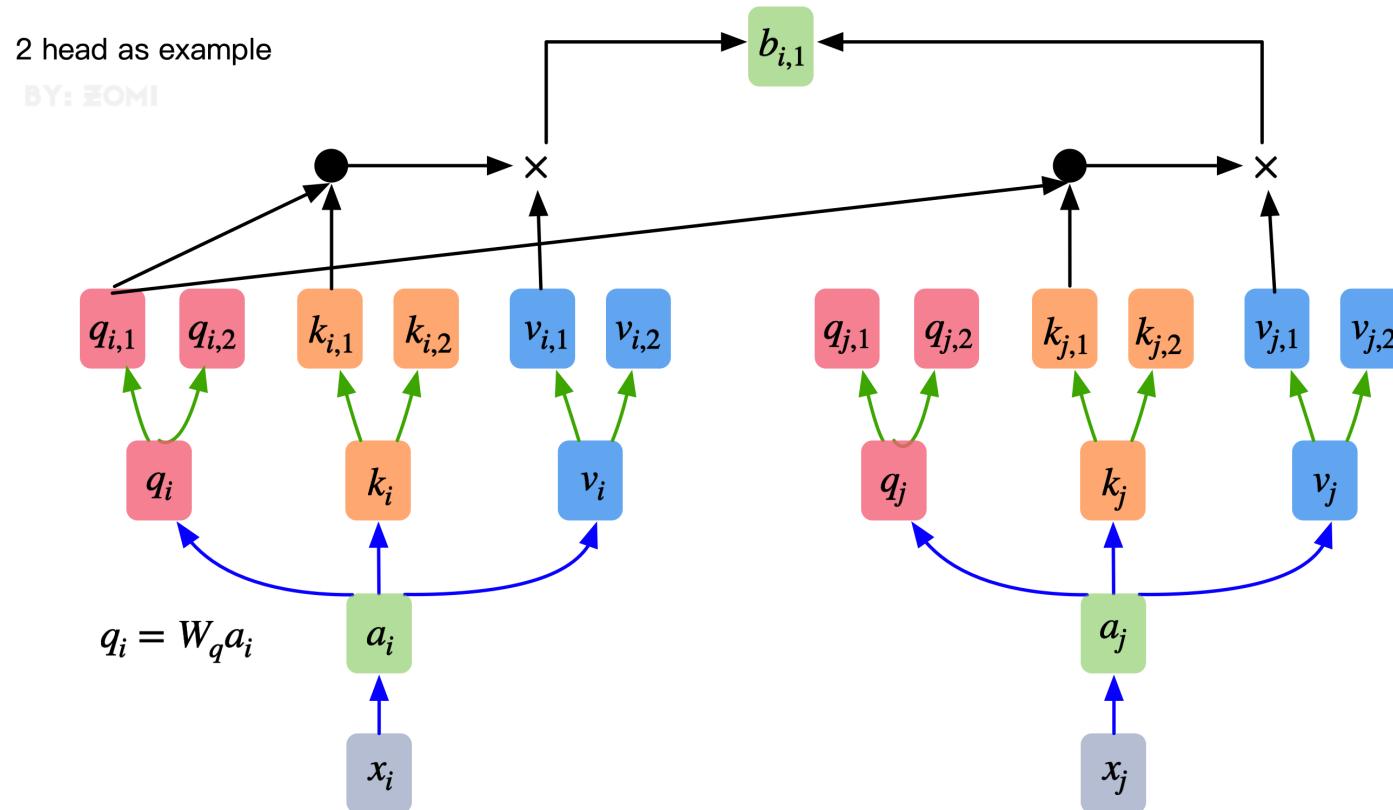
Transformer 取代RNN、CNN进入大模型时代

每一组QKV最后都有一个V输出，这是Self-Attention得到的最终结果，是当前向量在结合了它与其他向量关联权重后得到的结果



Transformer 取代RNN、CNN进入大模型时代

Multi-Head Attention



Transformer 取代RNN、CNN进入大模型时代

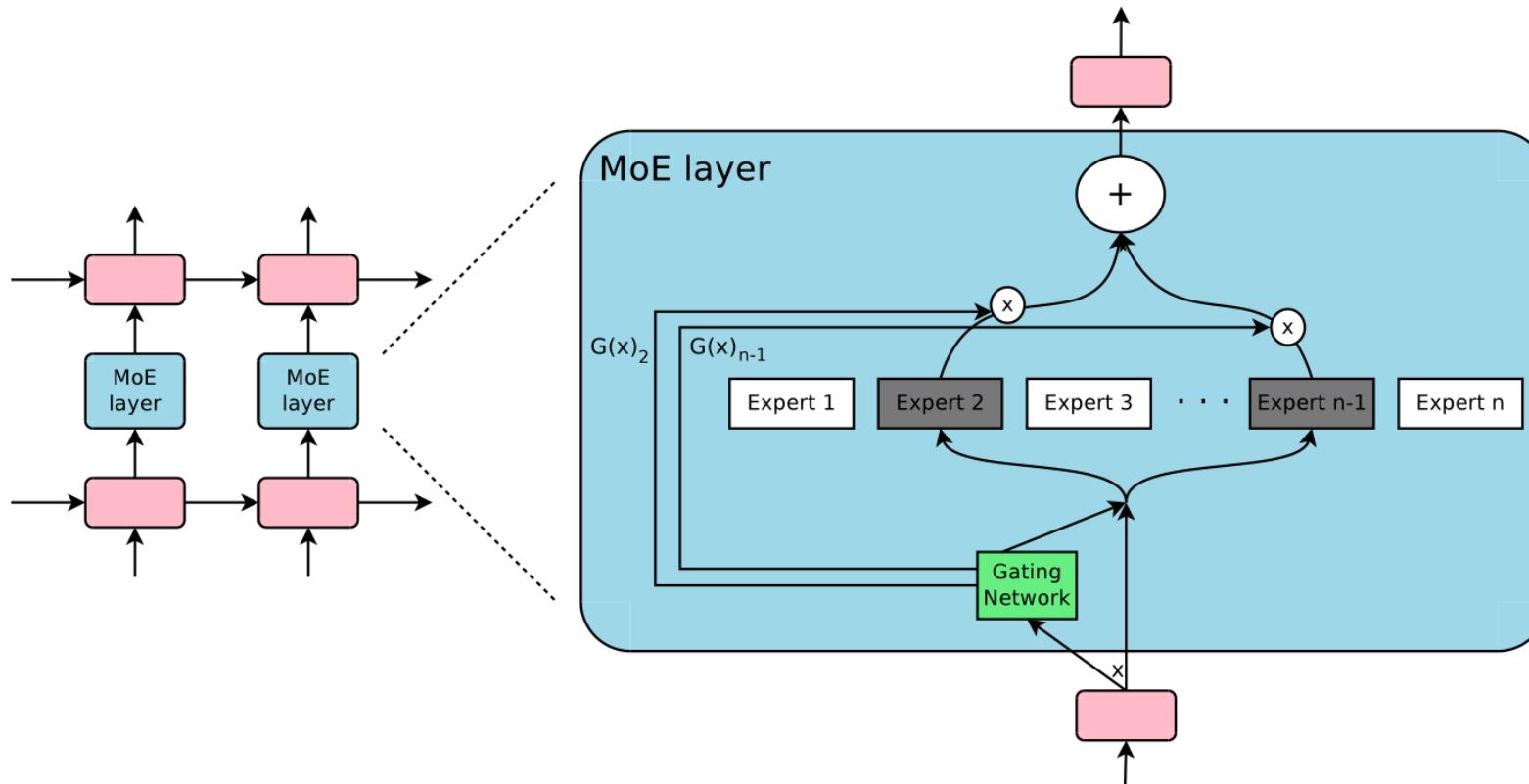
Transformer 解决 Seq2Seq 问题的 Transformer 模型，用全 attention 的结构代替了 LSTM。

- 每层计算复杂度更优；
- 可直接计算点乘结果；
- 模型更具有可解释性；
- 一步计算解决长时依赖问题；
- 模型参数量急剧膨胀；

Layer Type	Complexity per Layer	Sequential Operations	Maximum Path Length
Self-Attention	$O(n^2 \cdot d)$	$O(1)$	$O(1)$
Recurrent	$O(n \cdot d^2)$	$O(n)$	$O(n)$
Convolutional	$O(k \cdot n \cdot d^2)$	$O(1)$	$O(\log_k(n))$
Self-Attention (restricted)	$O(r \cdot n \cdot d)$	$O(1)$	$O(n/r)$

MoE 稀疏混合专家结构模型参数量进一步突破

稀疏门控专家混合模型（ Sparsely-Gated MoE ）：旨在实现条件计算，即神经网络的某些部分以每个样本为基础进行激活，作为一种显著增加模型容量和能力而不必成比例增加计算量的方法。



MoE 稀疏混合专家结构模型参数量进一步突破

为了保证稀疏性和均衡性，对softmax做了如下处理：

- 引入KeepTopK，这是个离散函数，将top-k之外的值强制设为负无穷大，从而softmax后的值为0。
- 加noise，这个的目的是为了做均衡，这里引入了一个Wnoise的参数，后面还会在损失函数层面进行改动。

$$y = \sum_{i=1}^n G(x)_i E_i(x) \quad (1)$$

$$G_\sigma(x) = \text{Softmax}(x \cdot W_g) \quad (2)$$

$$G(x) = \text{Softmax}(\text{KeepTopK}(H(x), k)) \quad (3)$$

$$H(x)_i = (x \cdot W_g)_i + \text{StandardNormal()} \cdot \text{Softplus}((x \cdot W_{\text{noise}})_i) \quad (4)$$

$$\text{KeepTopK}(v, k)_i = \begin{cases} v_i & \text{if } v_i \text{ is in the top } k \text{ elements of } v. \\ -\infty & \text{otherwise.} \end{cases} \quad (5)$$

MoE 稀疏混合专家结构模型参数量进一步突破

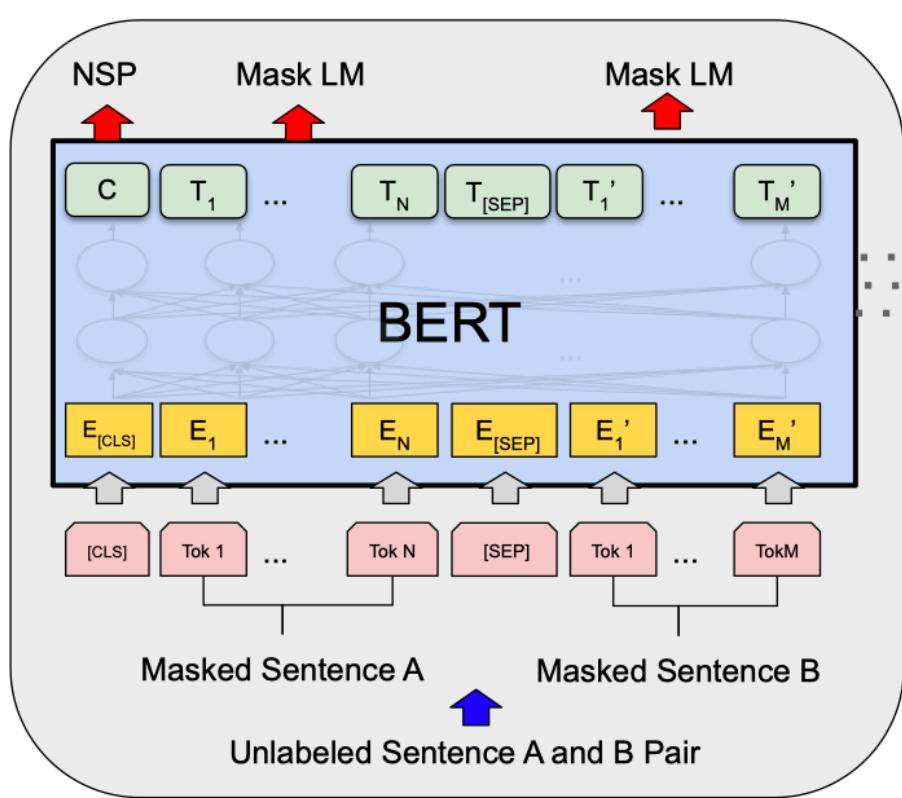
将大模型拆分成多个小模型，对于一个样本来说，无需经过所有的小模型去计算，而只是激活一部分小模型进行计算，这样就节省了计算资源。稀疏门控 MoE，实现了模型容量超过1000倍的改进，并且在现代 GPU 集群的计算效率损失很小。

Table 1: Summary of high-capacity MoE-augmented models with varying computational budgets, vs. best previously published results (Jozefowicz et al., 2016). Details in Appendix C.

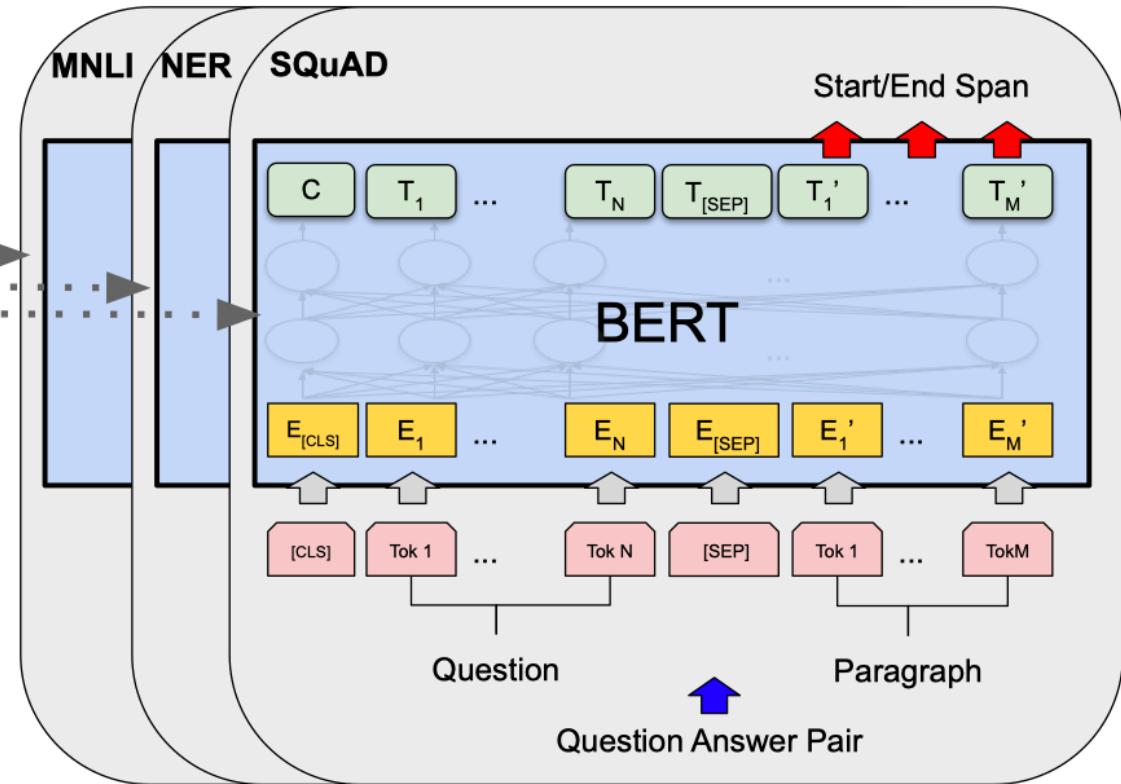
	Test Perplexity 10 epochs	Test Perplexity 100 epochs	#Parameters excluding embedding and softmax layers	ops/timestep	Training Time 10 epochs	TFLOPS /GPU
Best Published Results	34.7	30.6	151 million	151 million	59 hours, 32 k40s	1.09
Low-Budget MoE Model	34.1		4303 million	8.9 million	15 hours, 16 k40s	0.74
Medium-Budget MoE Model	31.3		4313 million	33.8 million	17 hours, 32 k40s	1.22
High-Budget MoE Model	28.0		4371 million	142.7 million	47 hours, 32 k40s	1.56

Bert 突破十亿规模NLP大模型

Transformer的双向编码表示来改进基于架构微调的方法



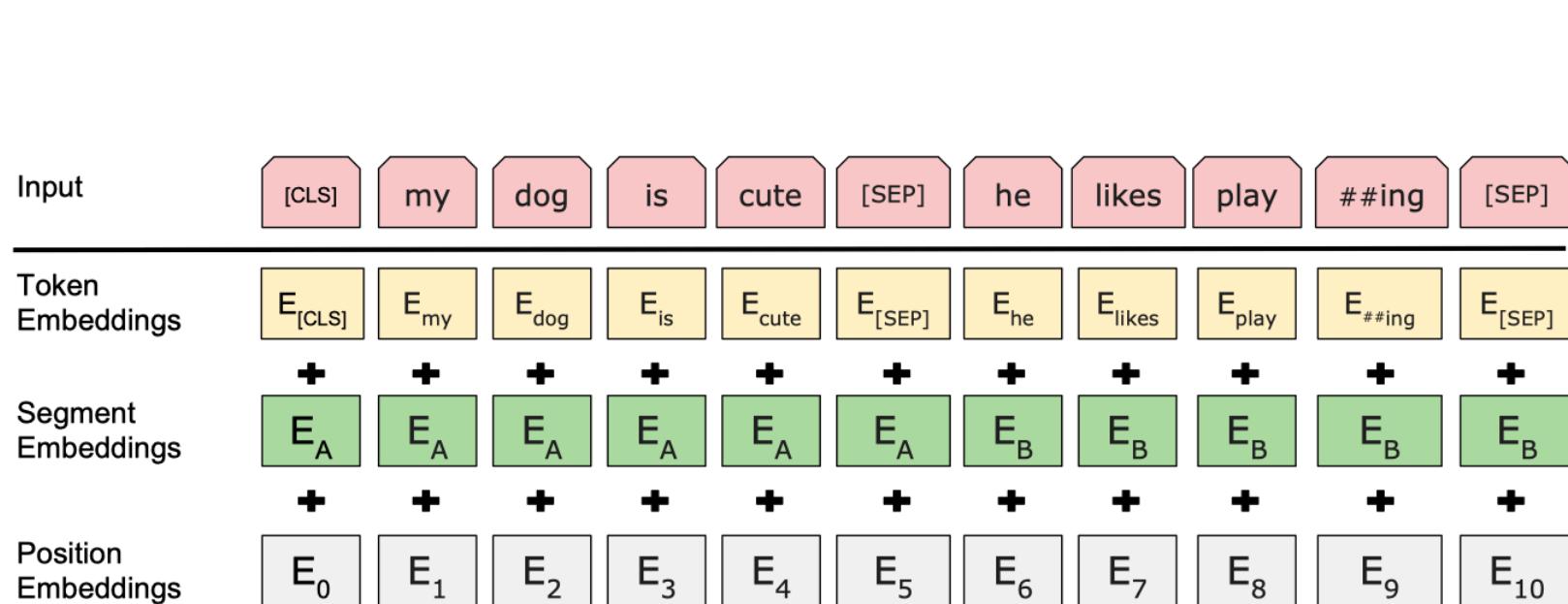
Pre-training



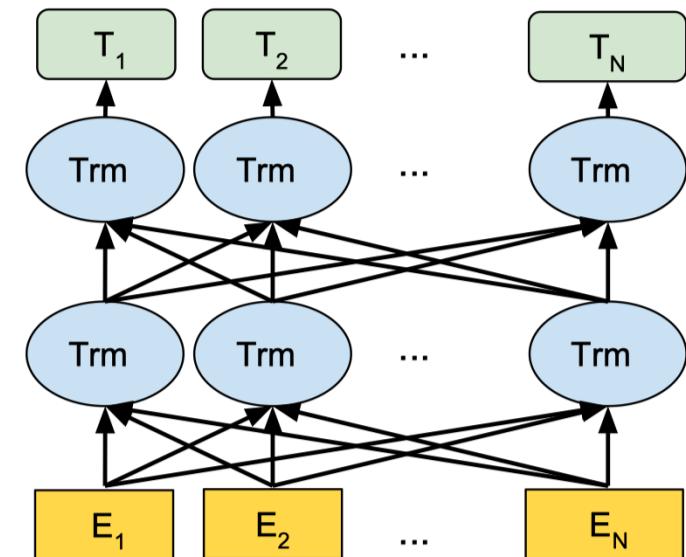
Fine-Tuning

Bert 突破十亿规模NLP大模型

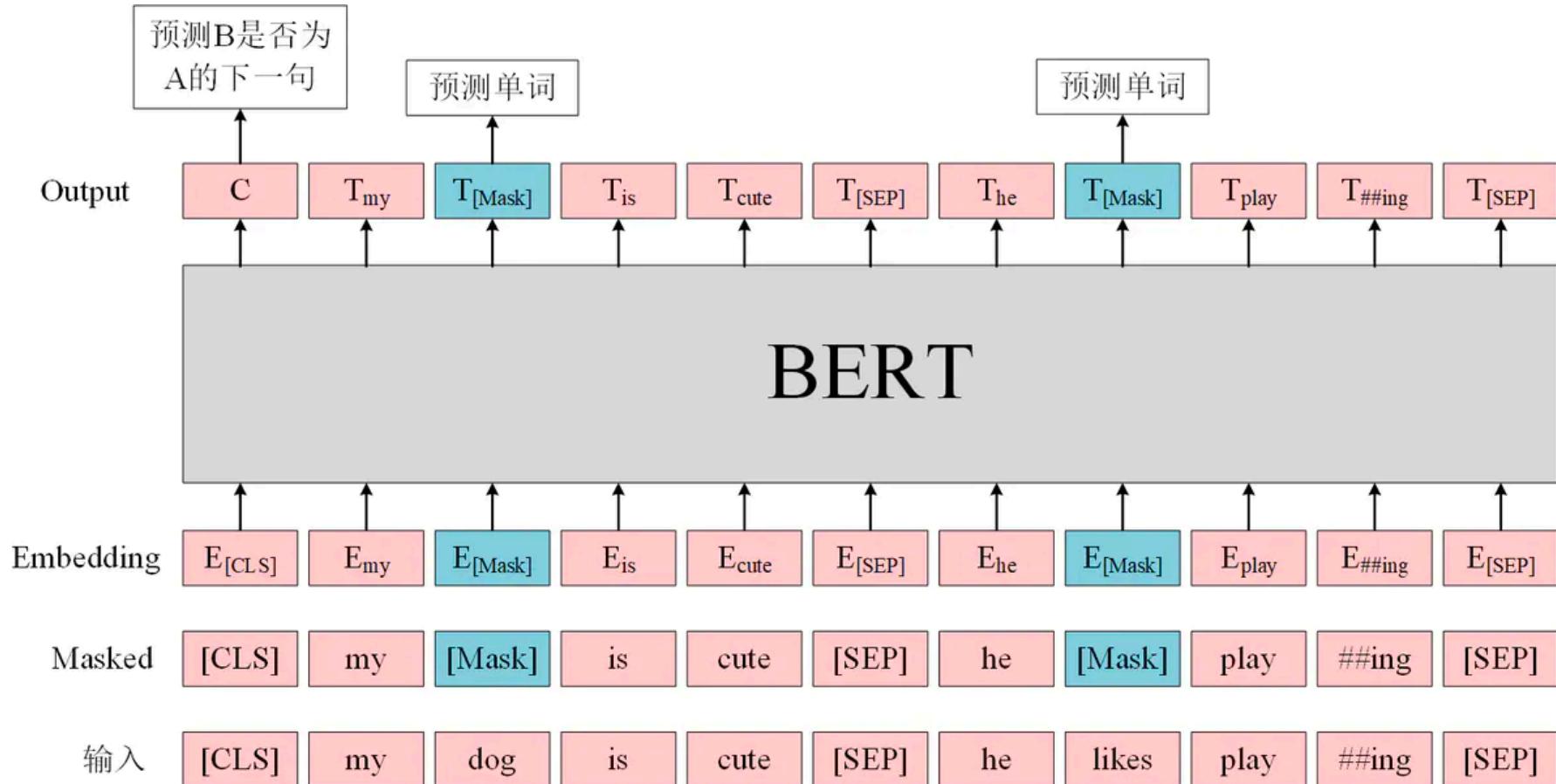
- [CLS] 标志放在第一个句子的首位，经过 BERT 得到的表征向量 C 可以用于后续的分类任务。
- [SEP] 标志用于分开两个输入句子，例如输入句子 A 和 B，要在句子 A, B 后面增加 [SEP] 标志。
- [MASK] 标志用于遮盖句子中单词，将单词用 [MASK] 遮盖之后，再利用 BERT 输出 [MASK] 向量预测单词。



BERT (Ours)

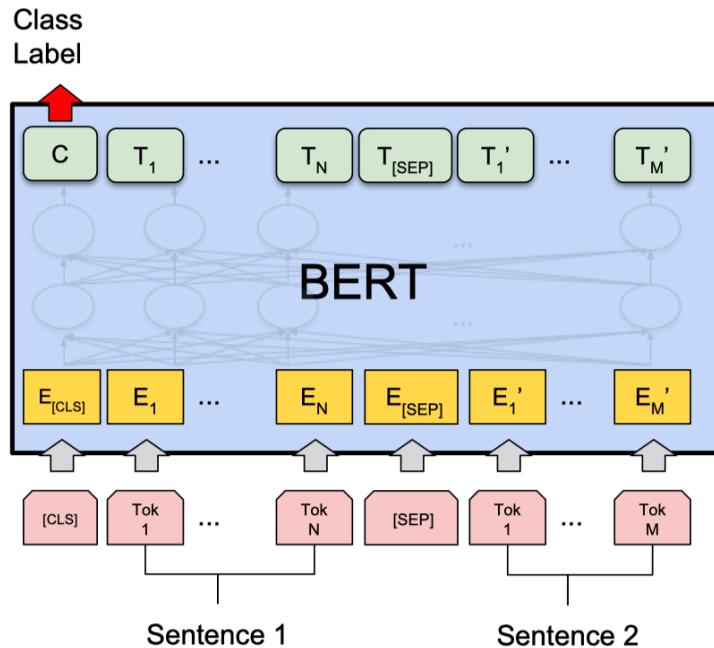


Bert 突破十亿规模NLP大模型

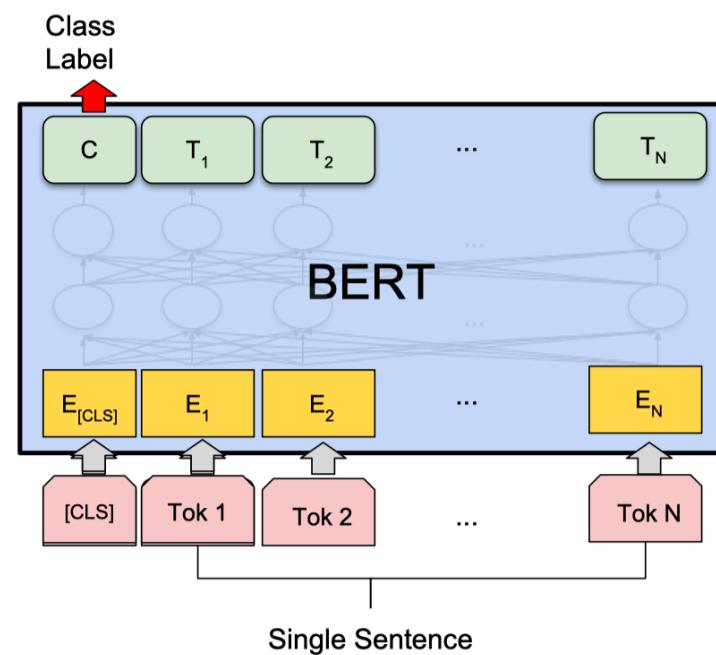


Bert 突破十亿规模NLP大模型

Token级任务：BERT 的第二个预训练任务是 Next Sentence Prediction (NSP) , 下一句预测任务，这个任务主要是让模型能够更好地理解句子间的关系。



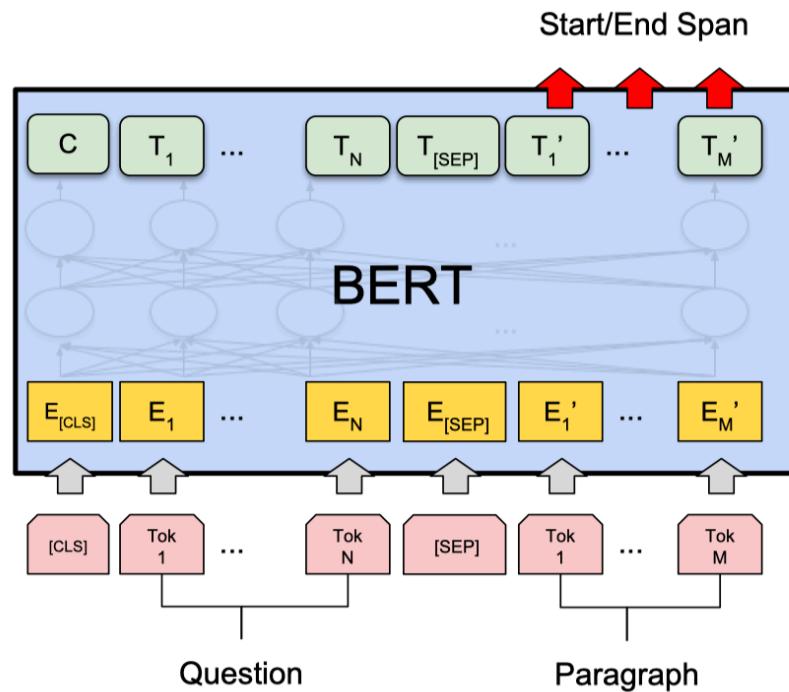
(a) Sentence Pair Classification Tasks:
MNLI, QQP, QNLI, STS-B, MRPC,
RTE, SWAG



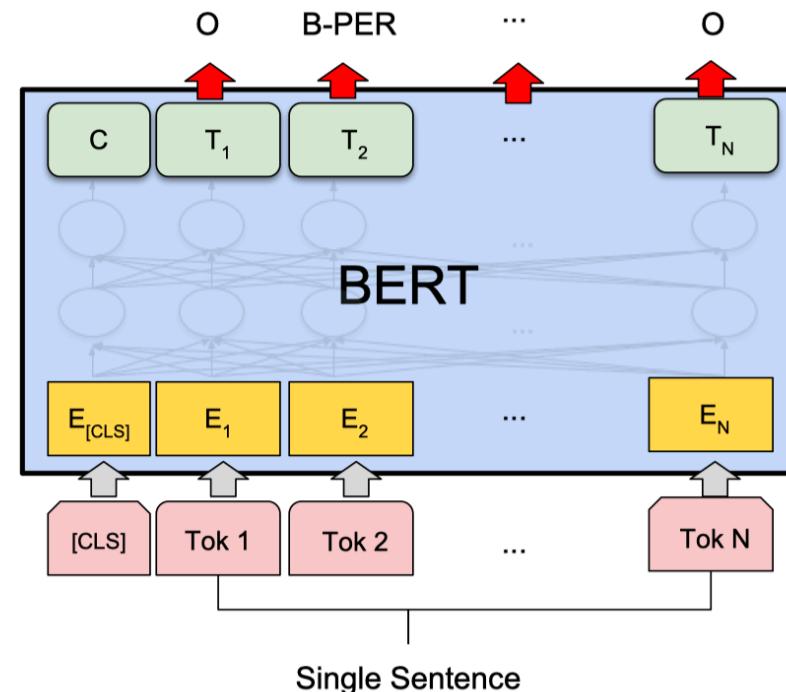
(b) Single Sentence Classification Tasks:
SST-2, CoLA

Bert 突破十亿规模NLP大模型

序列级任务：BERT 的第一个预训练任务是 Masked LM，在句子中随机遮盖一部分单词，然后同时利用上下文的信息预测遮盖的单词，这样可以更好地根据全文理解单词的意思。



(c) Question Answering Tasks:
SQuAD v1.1



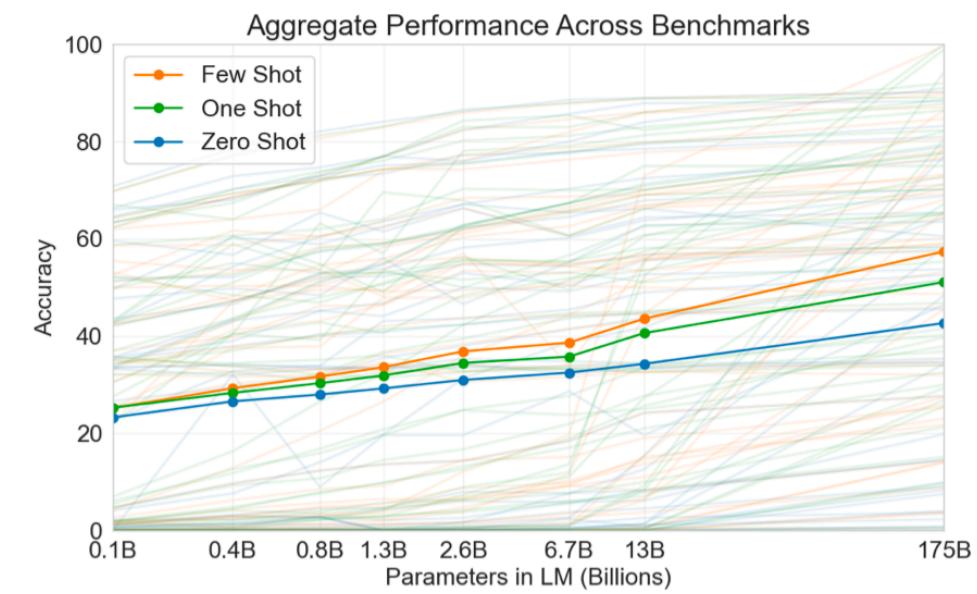
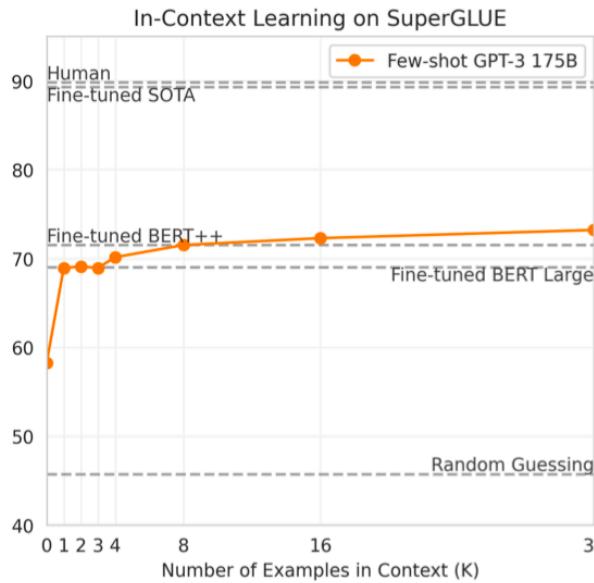
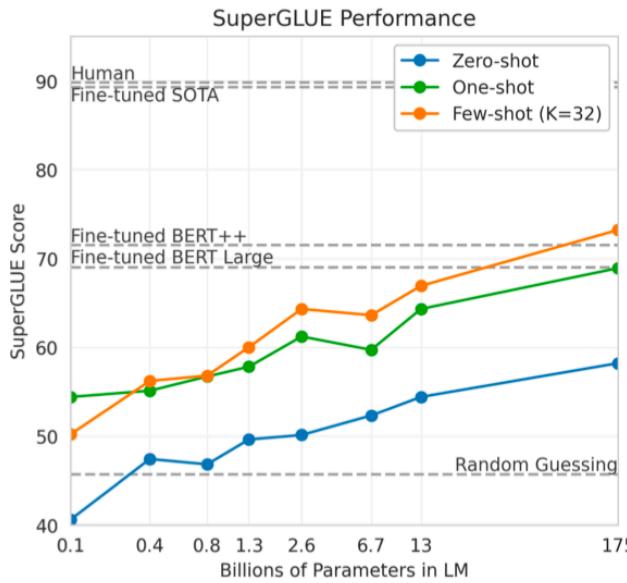
(d) Single Sentence Tagging Tasks:
CoNLL-2003 NER

Bert 突破十亿规模NLP大模型

- BERT-base 模型包含 12 个 Encoder block , BERT-large 包含 24 个 Encoder block , 通过超大数据、巨大模型、和极大的计算开销训练而成，在11个自然语言处理的任务中取得了最优 (state-of-the-art, SOTA) 。
- 用了超大的数据集 (BooksCorpus 800M + English Wikipedia 2.5G单词) 和超大的算力 (对应于超大模型) 在相关的任务上做预训练，实现了在目标任务上表现的单调增长；
- 训练主要分为两个阶段：预训练阶段和 Fine-tuning 阶段。 Fine-tuning 阶段是后续用于一些下游任务的时候进行微调，例如文本分类，词性标注，问答系统等，BERT 无需调整结构就可以在不同的任务上进行微调。

GPT3 全新语言模型 1750 亿参数大模型

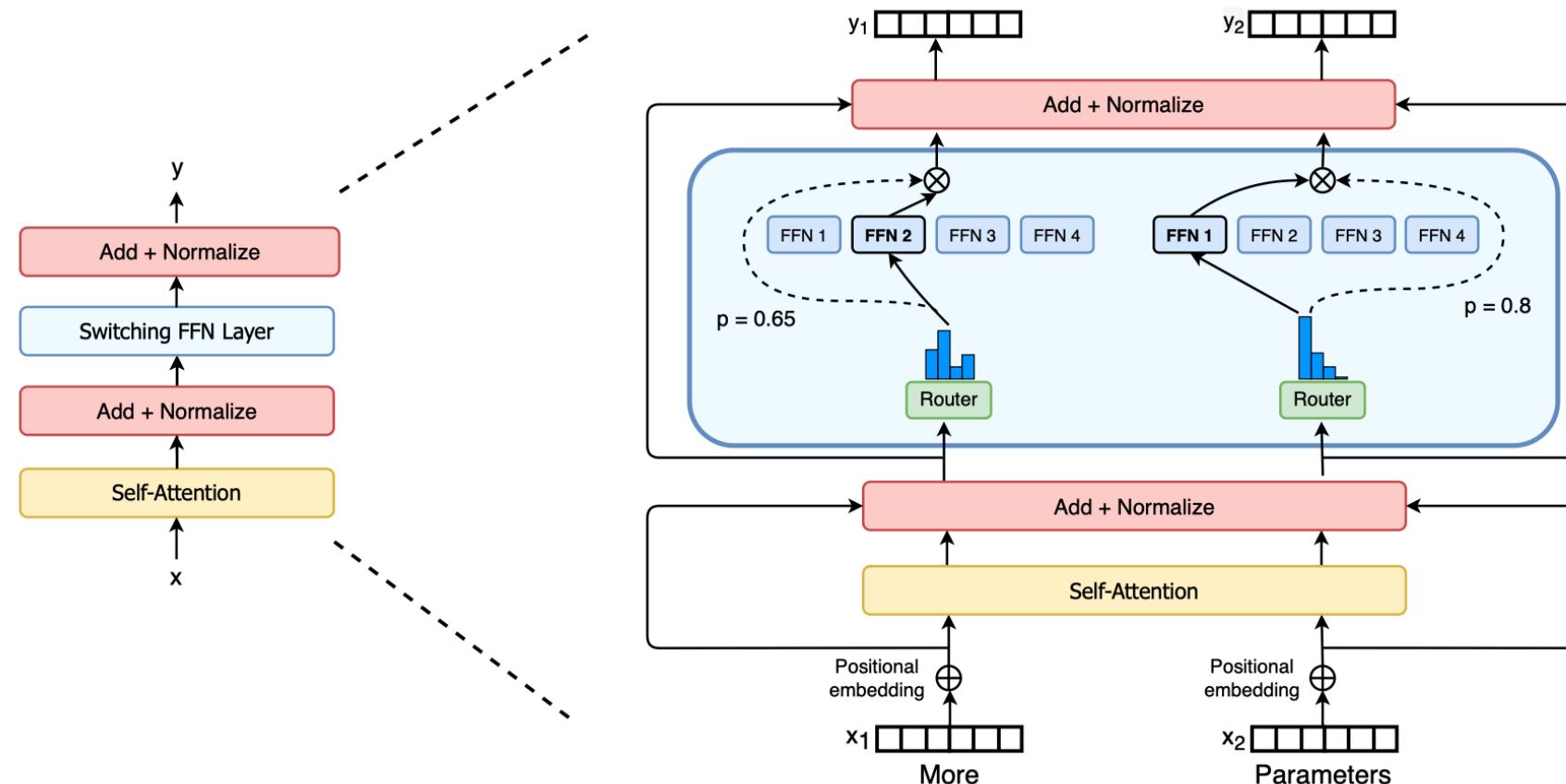
GPT-3毫无疑问是NLP领域乃至AI领域取得重大突破的一项工作，1750亿参数的超大规模，使得语言模型具备了生成难辨真假的新闻文章的能力。



<https://jalammar.github.io/how-gpt3-works-visualizations-animations/>

Switch Transformer 首个突破万亿大模型

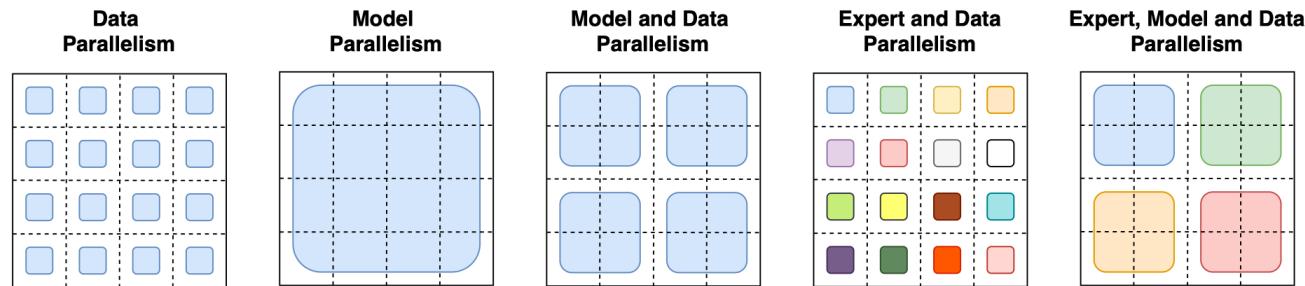
将开关层添加到Transformer自我注意层中，生成查询、键和值的可训练权重矩阵替换为交换层；



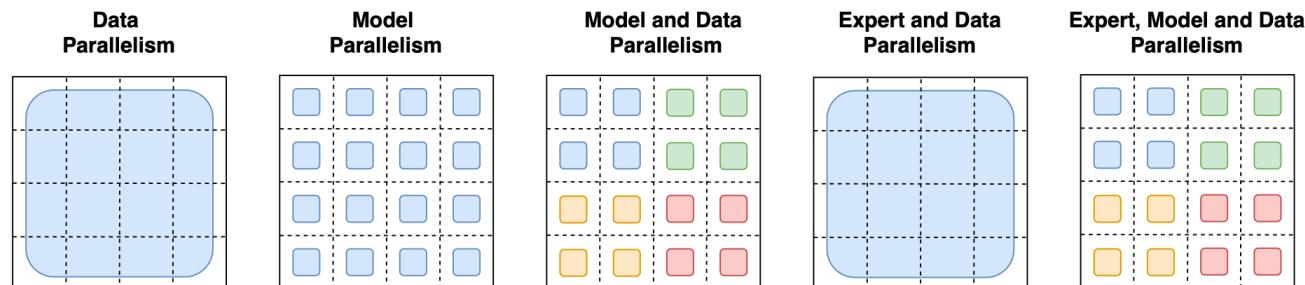
Switch Transformer 首个突破万亿大模型

- 基于Transformer MoE网络结构，简化了MoE的routing机制，降低了计算量；
- 进一步通过数据并行、模型并行、Expert并行的方式降低了训练通信量，提升训练性能；

How the *model weights* are split over cores

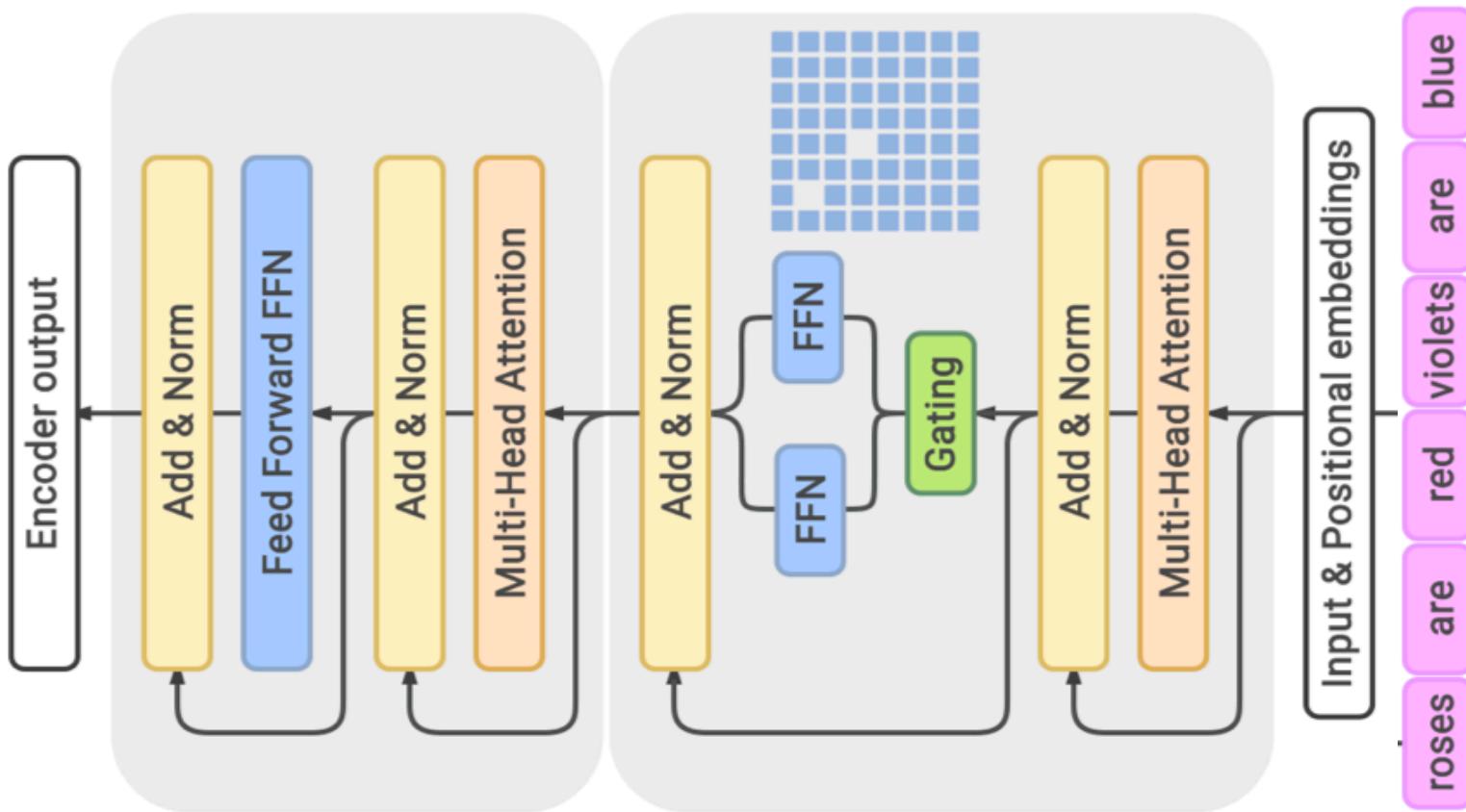


How the *data* is split over cores



GLaM 1.2万亿参数通用稀疏语言模型

每个 MoE 层的 64 个专家中有 1.2T 的总参数，总共有 32 个 MoE 层



Summary

1. 一切都只需要注意力的Transformer引发AI迈进大模型时代
2. 稀疏门控专家混合MoE，计算比最先进的密集 LSTM 模型少10倍
3. Transformer的双向无监督编码器表示的Bert椒麻鸡
4. 突破千亿规模的自回归万能语言模型 GPT-3
5. Switch Transformers 将 MoE 风格的架构与 Transformer 模型架构相结合
6. 谷歌1.2万亿通用稀疏语言模型GLaM，小样本学习打败GPT-3



BUILDING A BETTER CONNECTED WORLD

THANK YOU

Copyright©2014 Huawei Technologies Co., Ltd. All Rights Reserved.

The information in this document may contain predictive statements including, without limitation, statements regarding the future financial and operating results, future product portfolio, new technology, etc. There are a number of factors that could cause actual results and developments to differ materially from those expressed or implied in the predictive statements. Therefore, such information is provided for reference purpose only and constitutes neither an offer nor an acceptance. Huawei may change the information at any time without notice.

引用

- I. Transformer. Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).
- II. MoE. Shazeer, Noam, et al. "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer." *arXiv preprint arXiv:1701.06538* (2017).
- III. BERT Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).
- IV. GPT3. Brown, Tom, et al. "Language models are few-shot learners." *Advances in neural information processing systems* 33 (2020): 1877-1901.
- V. Switch Transformer. Fedus, William, Barret Zoph, and Noam Shazeer. "Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity." (2021).
- VI. GLAM. Du, Nan, et al. "Glam: Efficient scaling of language models with mixture-of-experts." *International Conference on Machine Learning.* PMLR, (2022).