# BERT模型深度修炼指南

七月在线 NLP褚博士 zeweichu@gmail.com

七月在线
JULYEDU.COM

# BERT模型深度修炼指南

- 直播课中积极提问的同学可以获得七月在线视频小课奖励（公开课结束时选出）

# BERT模型深度修炼指南

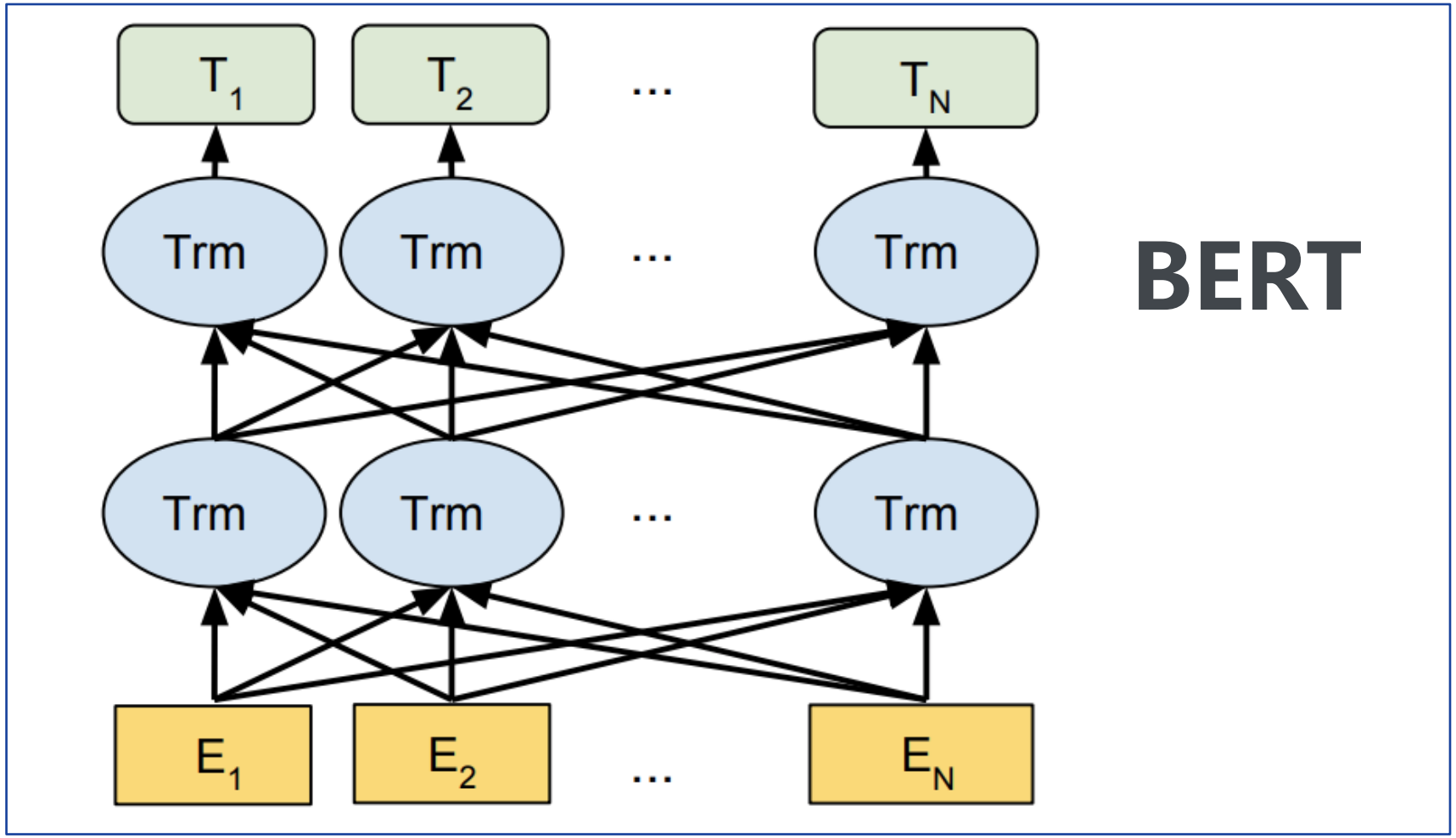
- NLP脱菜：什么是BERT
- 深度揭秘：BERT模型详解
- 一统江湖：BERT模型的应用
- 暴力美学：BERT模型的训练

# NLP脱菜：什么是BERT

BERT是一种词嵌入

(word embedding)

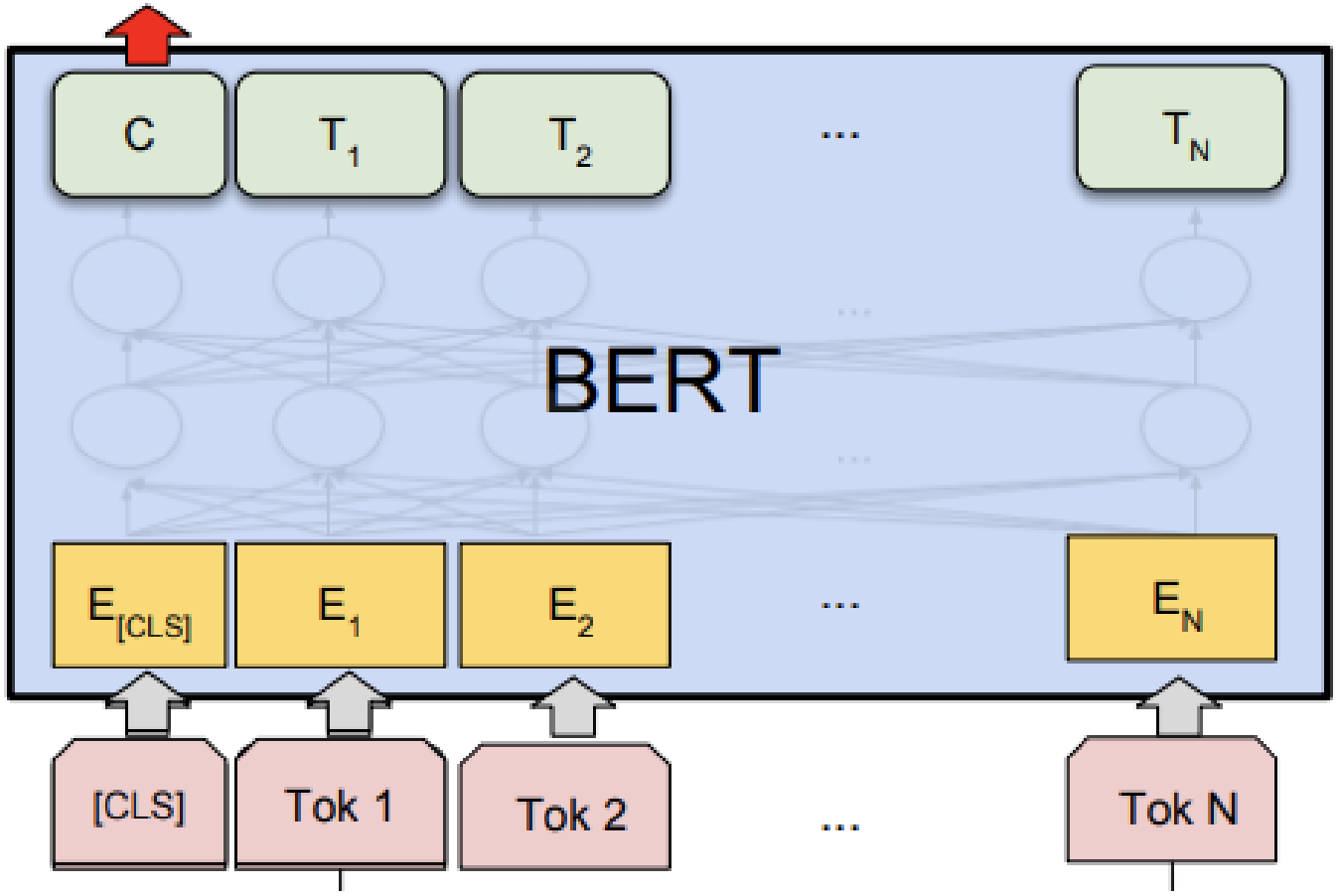
模型




图片来自 BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding https://arxiv.org/pdf/1810.04805.pdf

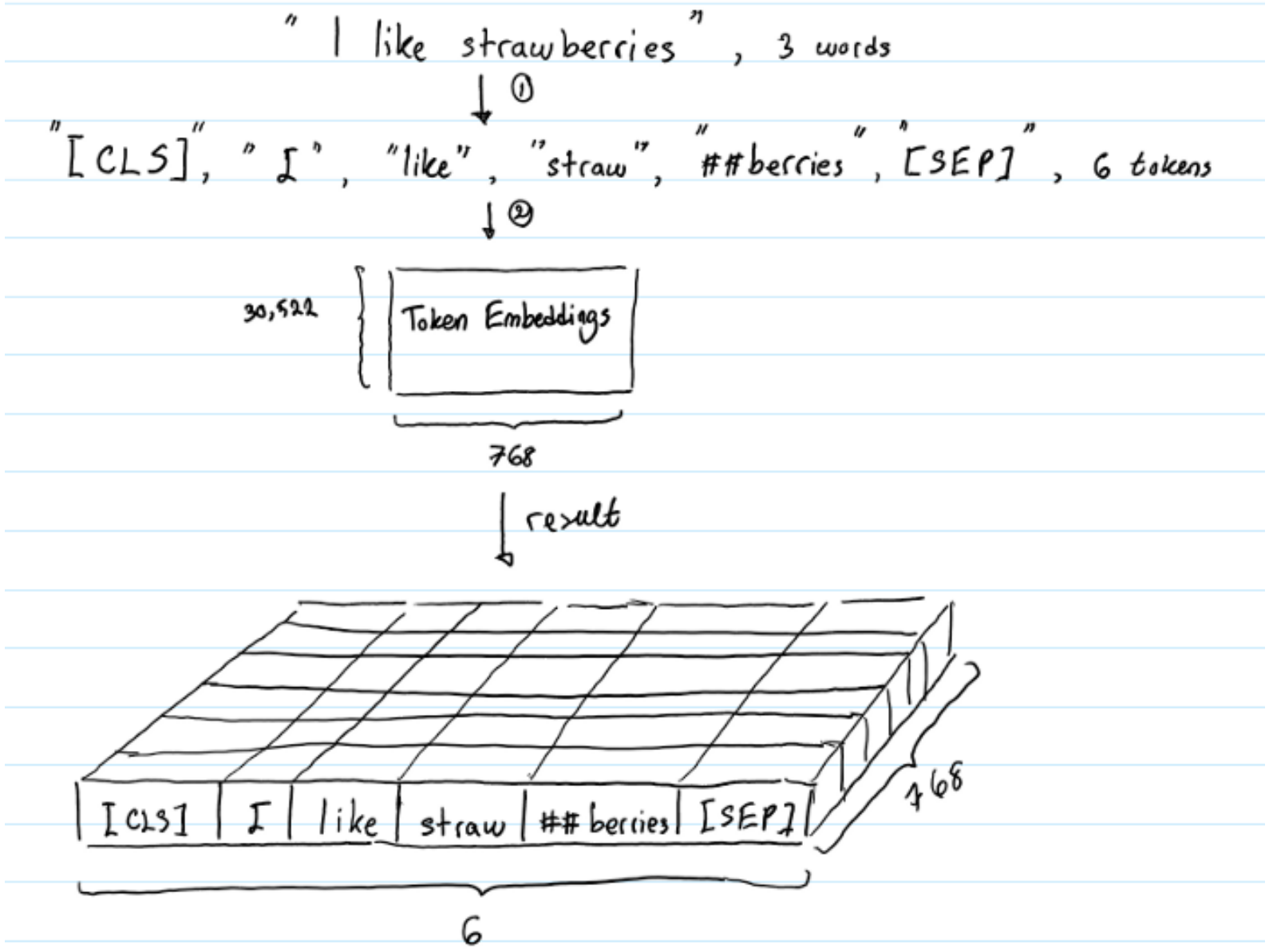
# NLP脱菜：什么是BERT

BERT是一种句子嵌入(sentence embedding)模型



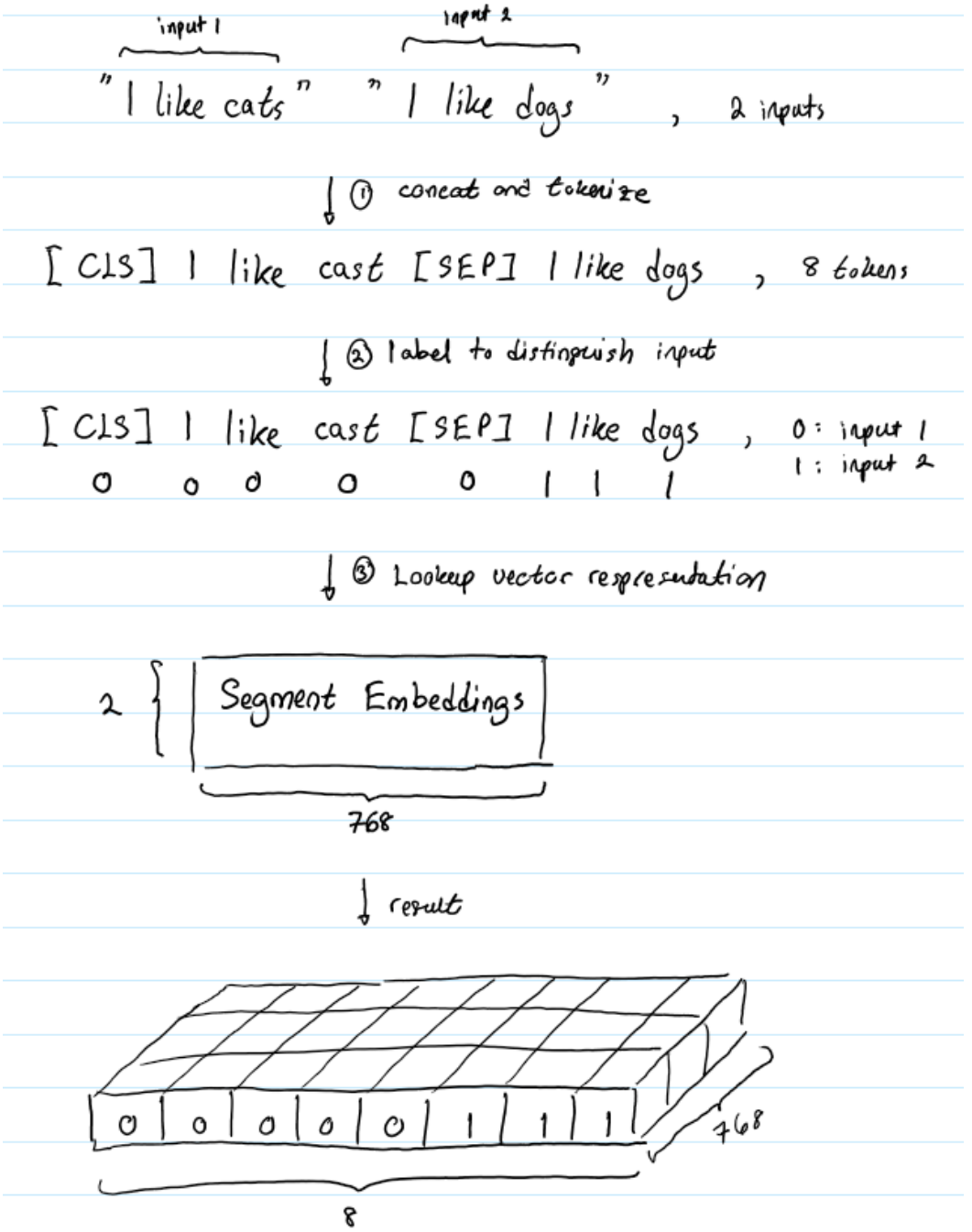

图片来自 BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding https://arxiv.org/pdf/1810.04805.pdf

# NLP脱菜：什么是BERT




图片来自
https://medium.com/@_init_/why-bert-has-3-embedding-layers-and-their-implementation-details-9c261108e28a

# NLP脱菜：什么是BERT




图片来自 https://medium.com/@_init_/why-bert-has-3-embedding-layers-and-their-implementation-details-9c261108e28a

# NLP脱菜：什么是BERT

BERT词向量的优势：

- 包含了语境 (context) 的信息，对比word2vec词向量
- 速度快，并行程度高，对比ELMo模型
- 包含双向 (bidirectional) 语境信息，对比GPT模型
- 在各类NLP任务上效果出众，例如文本分类、问答、词标注（词性标注、实体识别）等等

# 深度揭秘：BERT模型详解

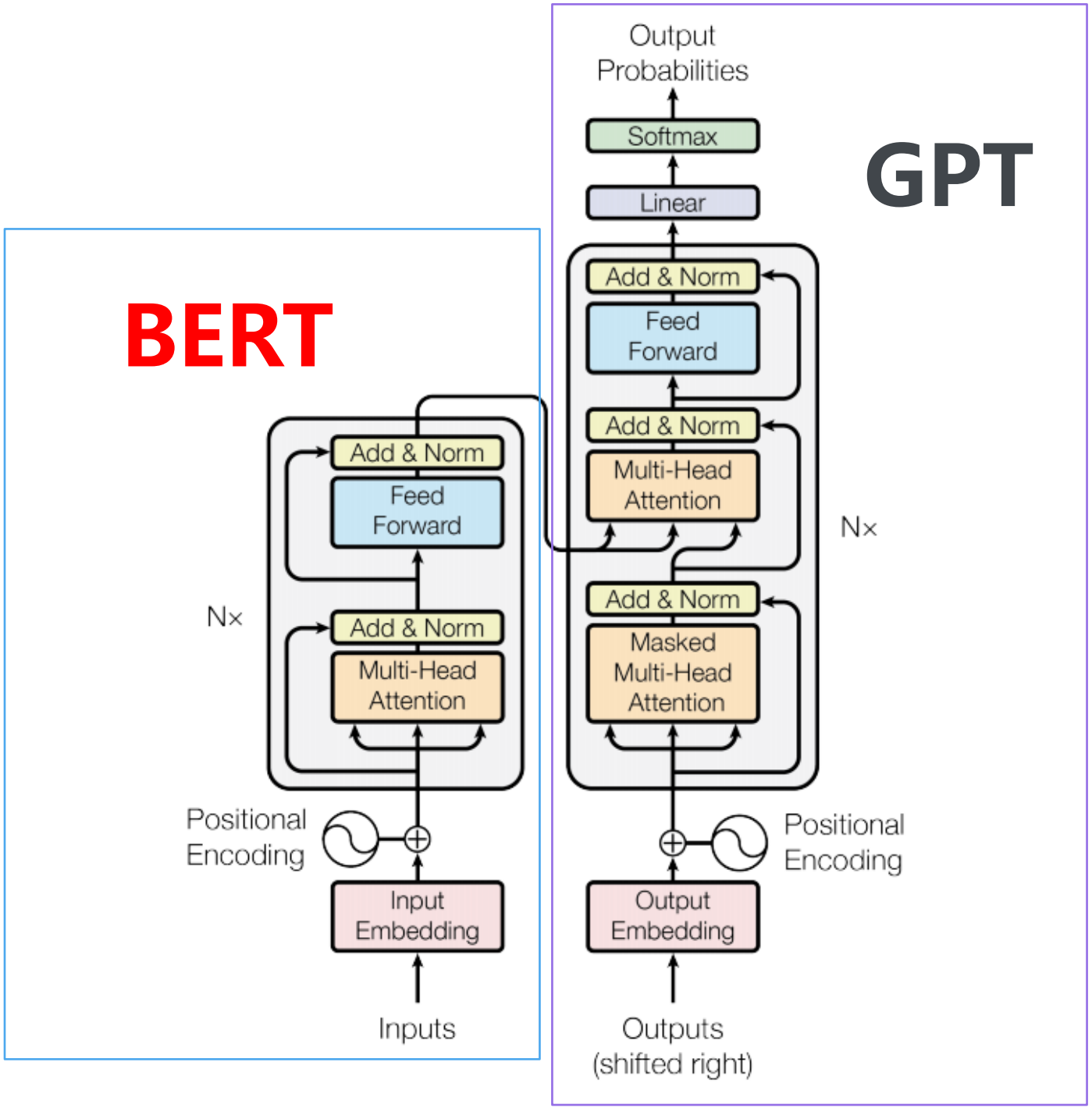
Bidirectional Encoder Representation from Transformers

问题：

- 什么是Transformer?
- 什么是Bidirectional?

# Transormer模型

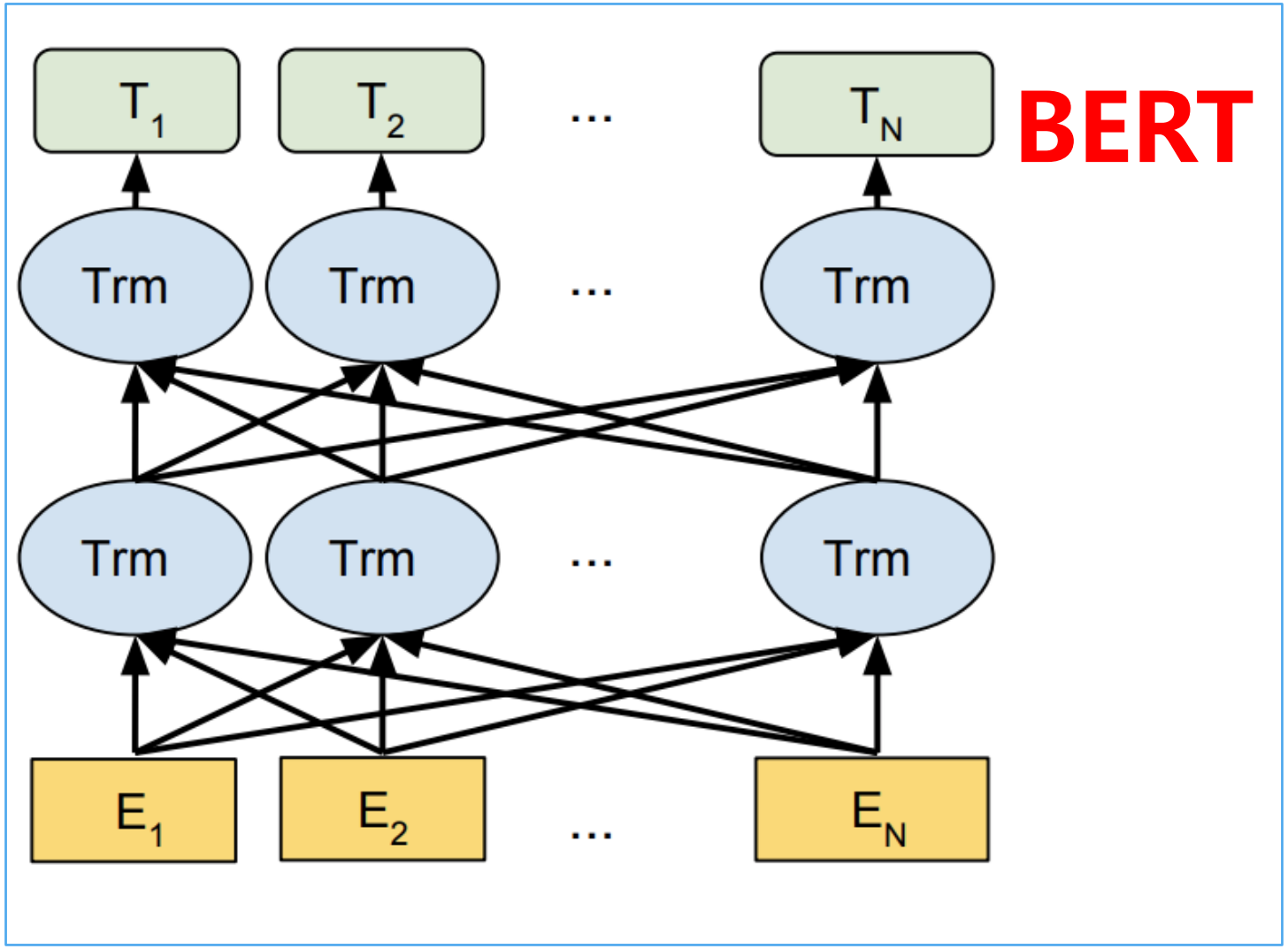
图片来自Vaswan et al., Attention Is All You Need https://arxiv.org/pdf/1706.03762.pdf



# Bidirectional

每一个词向量的产生都同时依赖该单词的

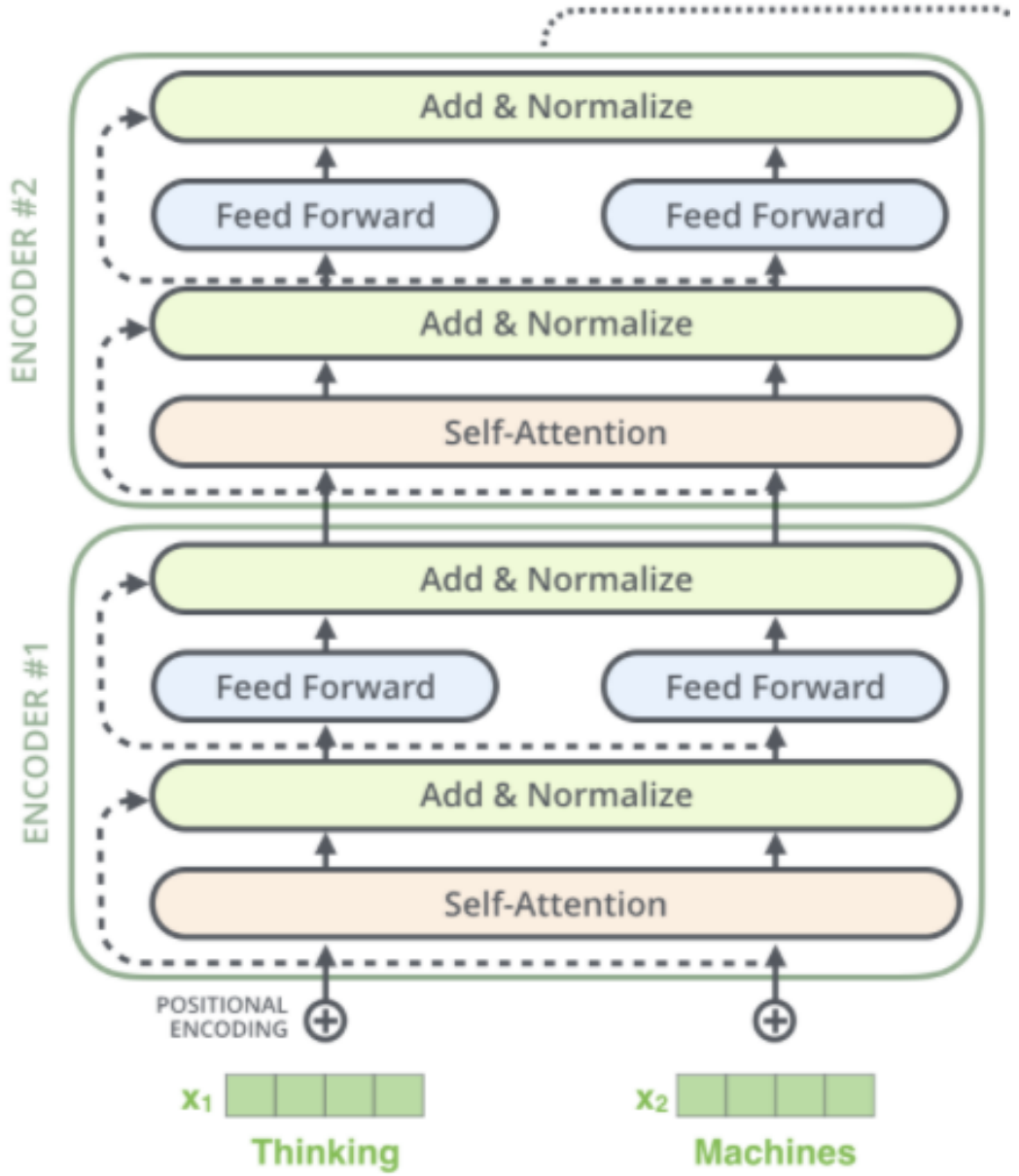
左侧与右侧的语境信息




图片来自 BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding https://arxiv.org/pdf/1810.04805.pdf

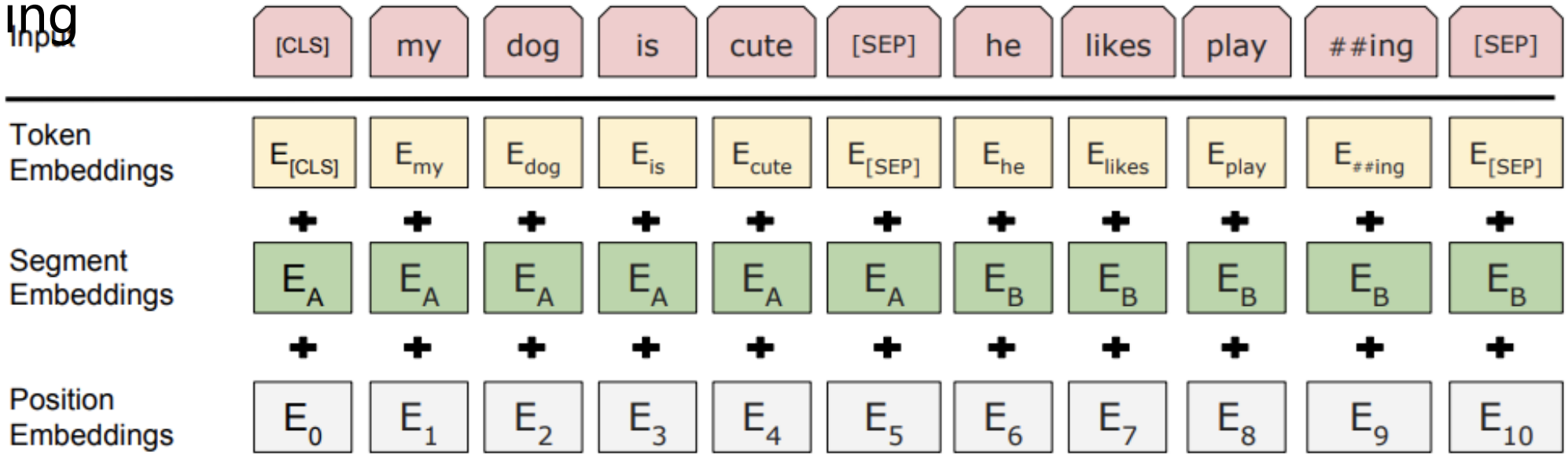
# 深度揭秘：BERT模型详解

图片来自
https://jalammar.github.io/illustrated-transformer/



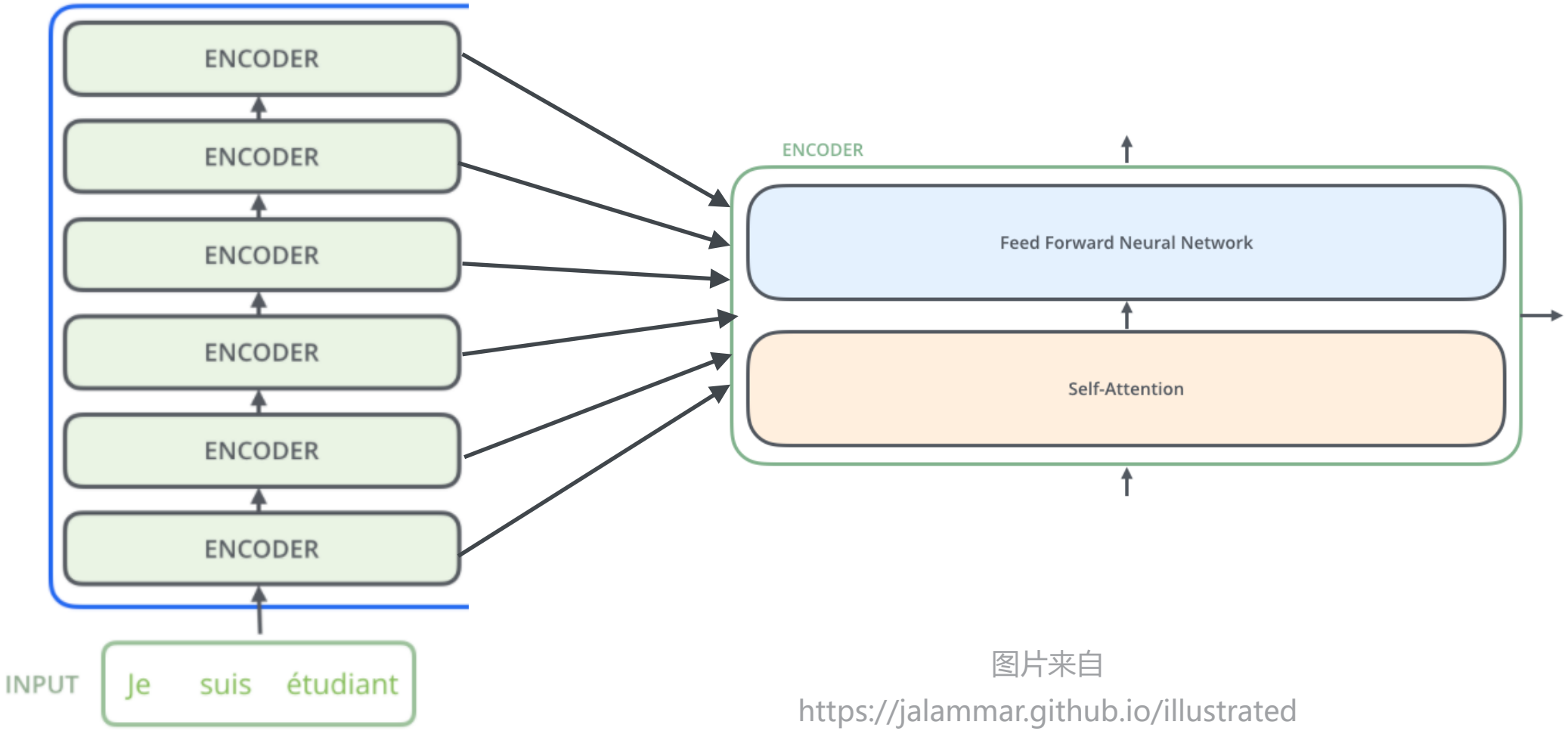
# 深度揭秘：BERT模型详解

BERT模型的输入：token embedding + segment embedding + position embedding



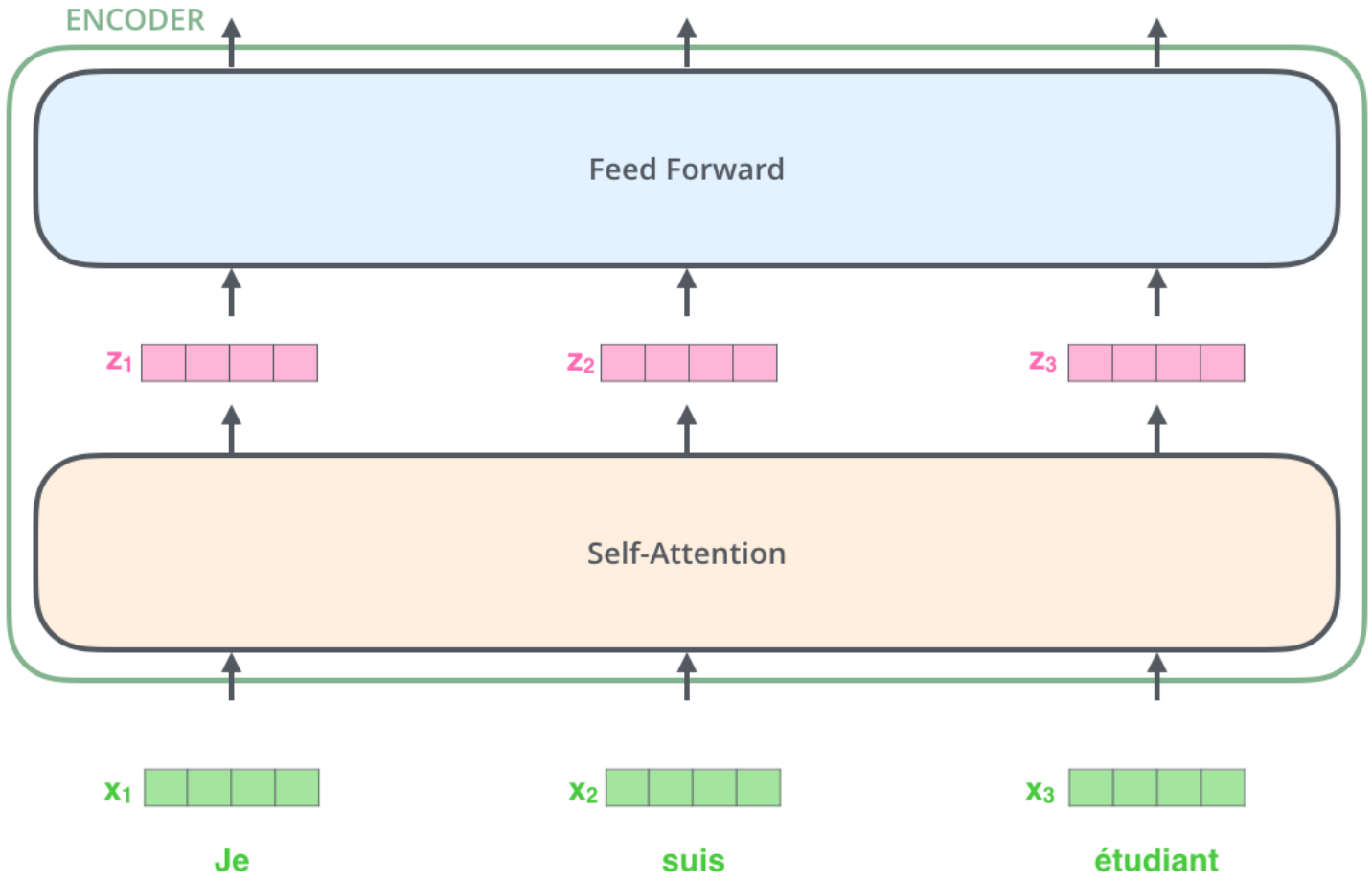

图片来自 BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding https://arxiv.org/pdf/1810.04805.pdf

# 深度揭秘：BERT模型详解



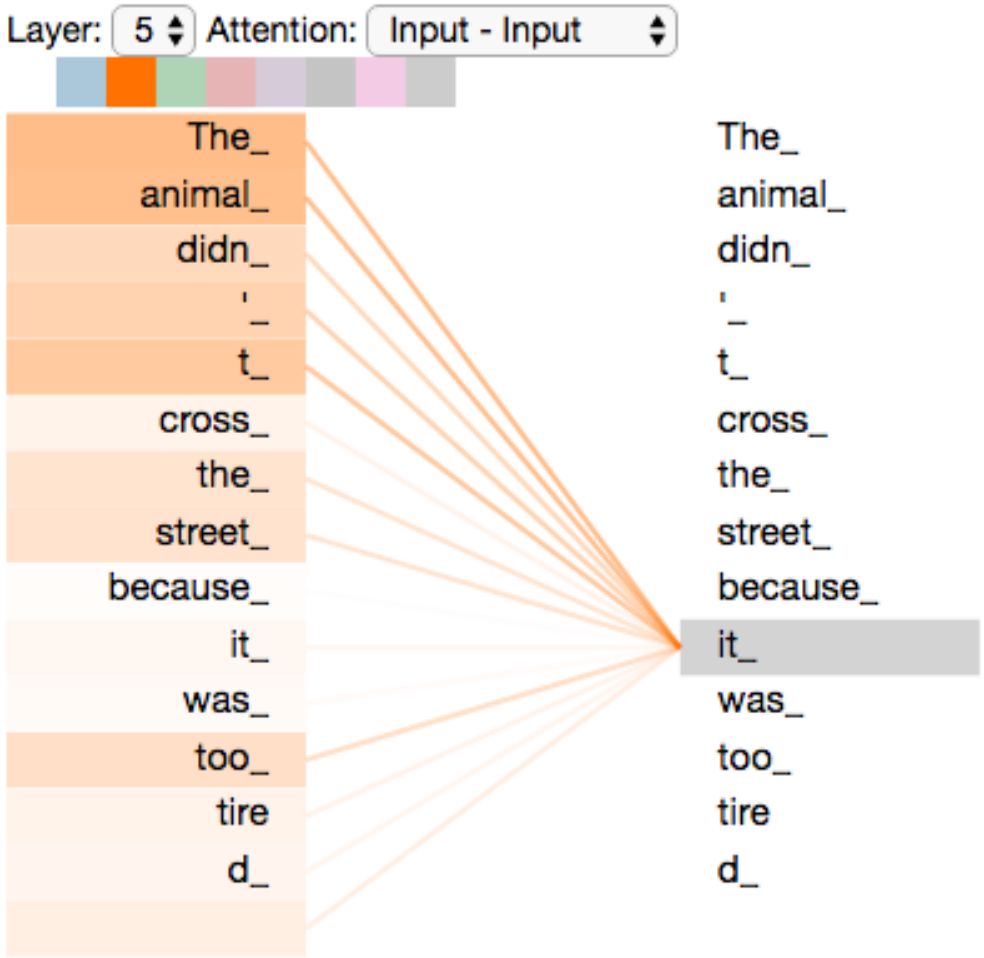

图片来自
https://jalammar.github.io/illustrated-transformer/

# 深度揭秘：BERT模型详解



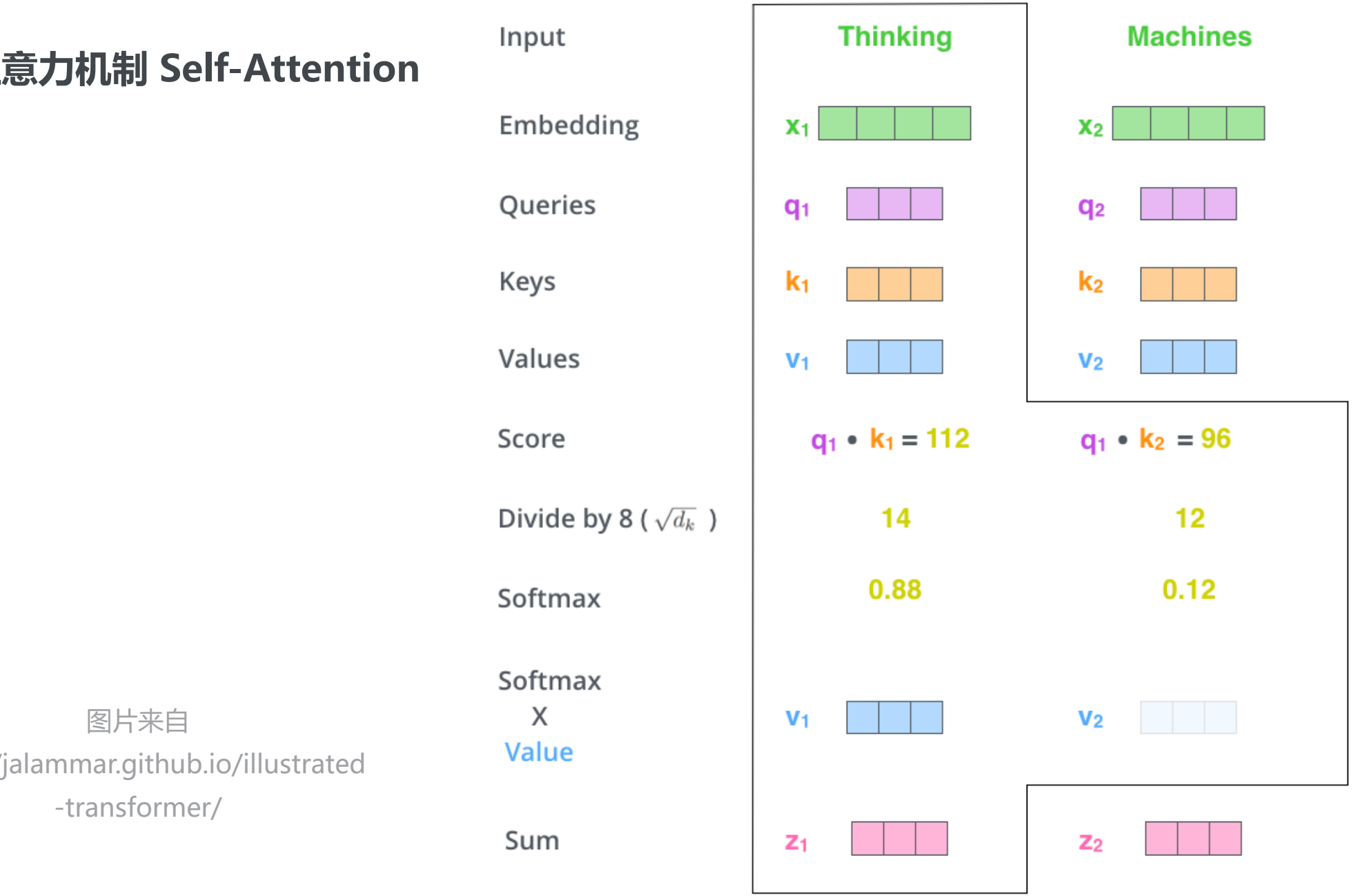

图片来自
https://jalammar.github.io/illustrated-transformer/

# 自注意力机制 Self-Attention



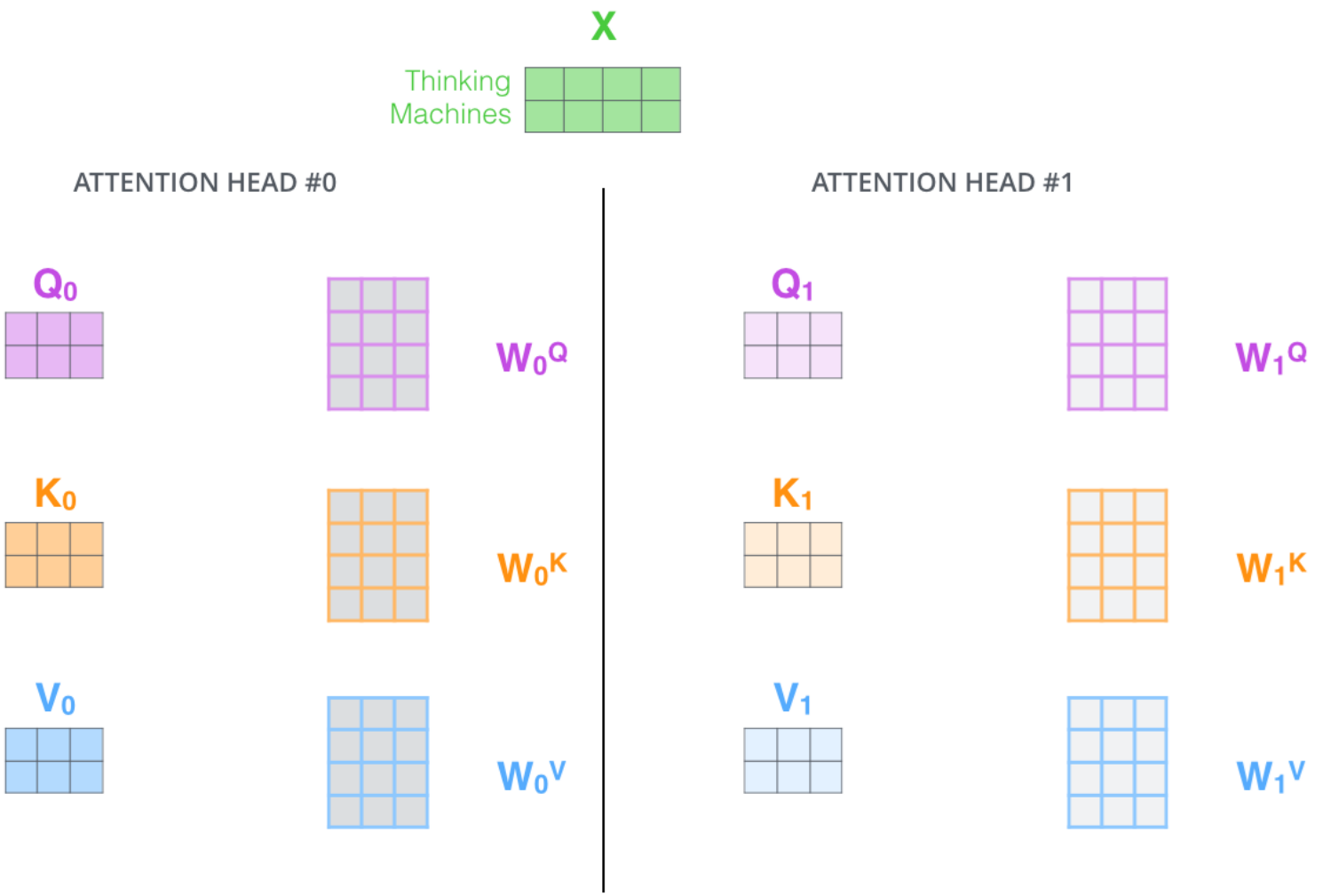

图片来自
https://jalammar.github.io/illustrated-transformer/

# 自注意力机制 Self-Attention



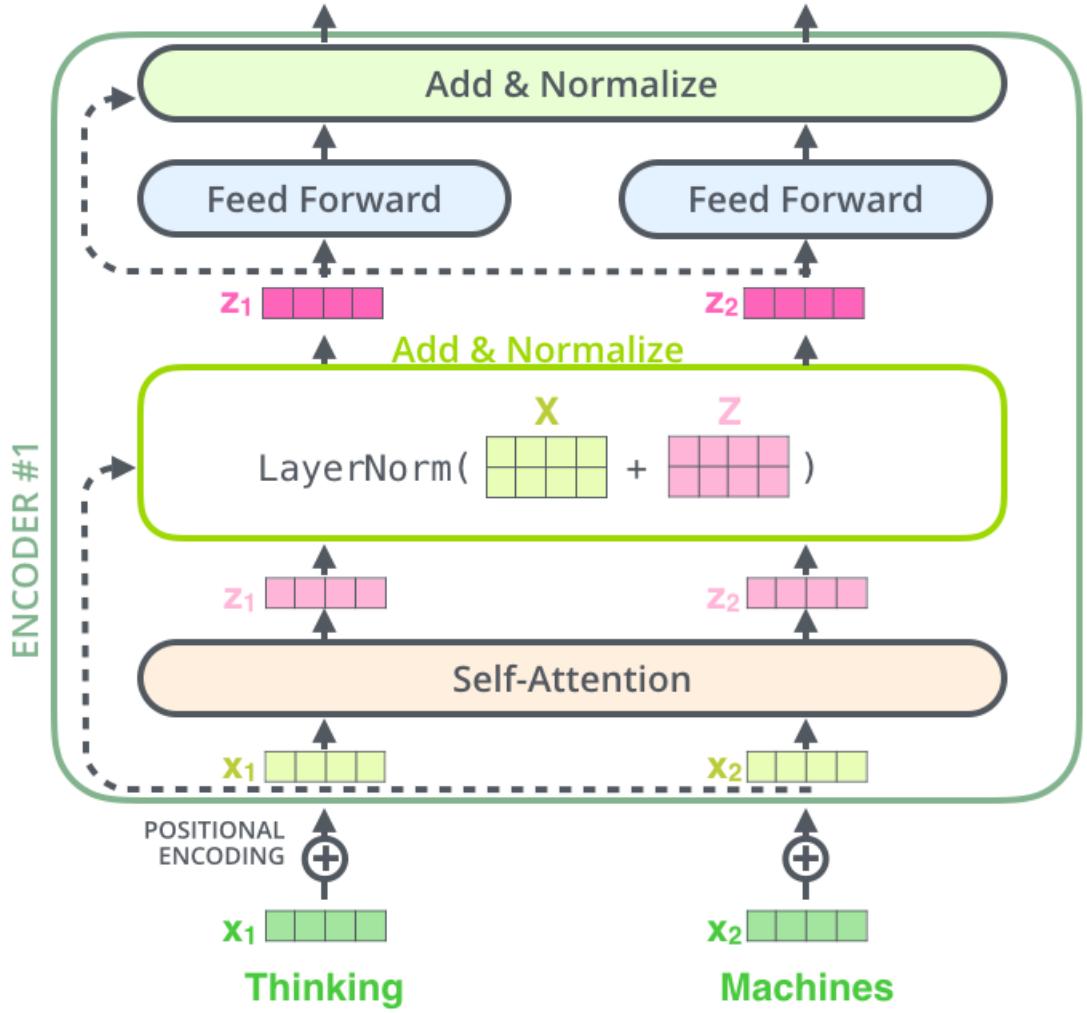

图片来自
https://jalammar.github.io/illustrated-transformer/

# 多头注意力 Multi-Headed Attention




图片来自
https://jalammar.github.io/illustrated-transformer/

# Residual, LayerNorm, Feed Forward

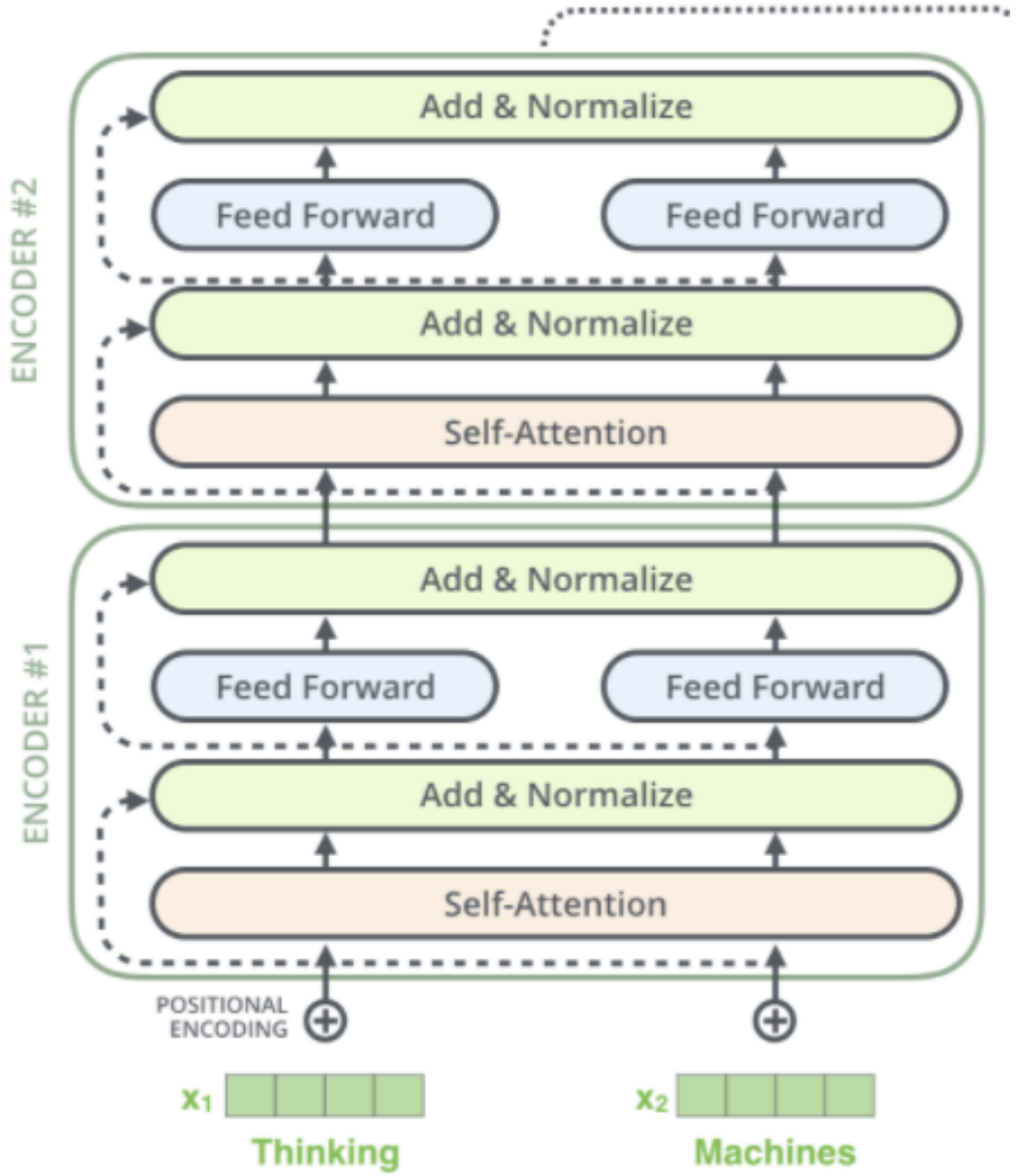



图片来自
https://jalammar.github.io/illustrated-transformer/

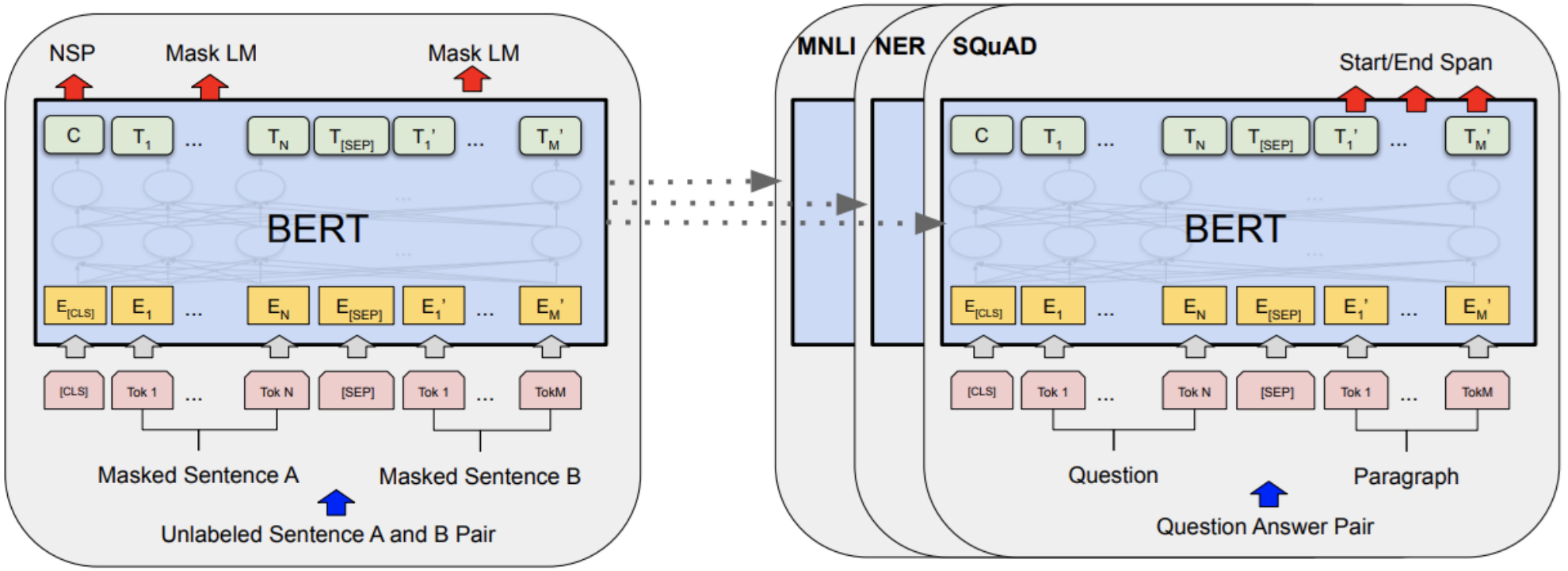
# 深度揭秘：BERT模型详解

$$W_1 \mathrm{ReLU}(W_2 x + b_2) + b_2$$

图片来自
https://jalammar.github.io/illustrated-transformer/



# 一统江湖：BERT模型的应用




预训练　　微调

图片来自 BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

https://arxiv.org/pdf/1810.04805.pdf

# 一统江湖：BERT模型的应用

BERT模型及BERT模型的变种刷新了各大NLP评测任务的榜首

- 文本分类
- 阅读理解
- 命名实体识别
- 自然语言推断
- 。。。

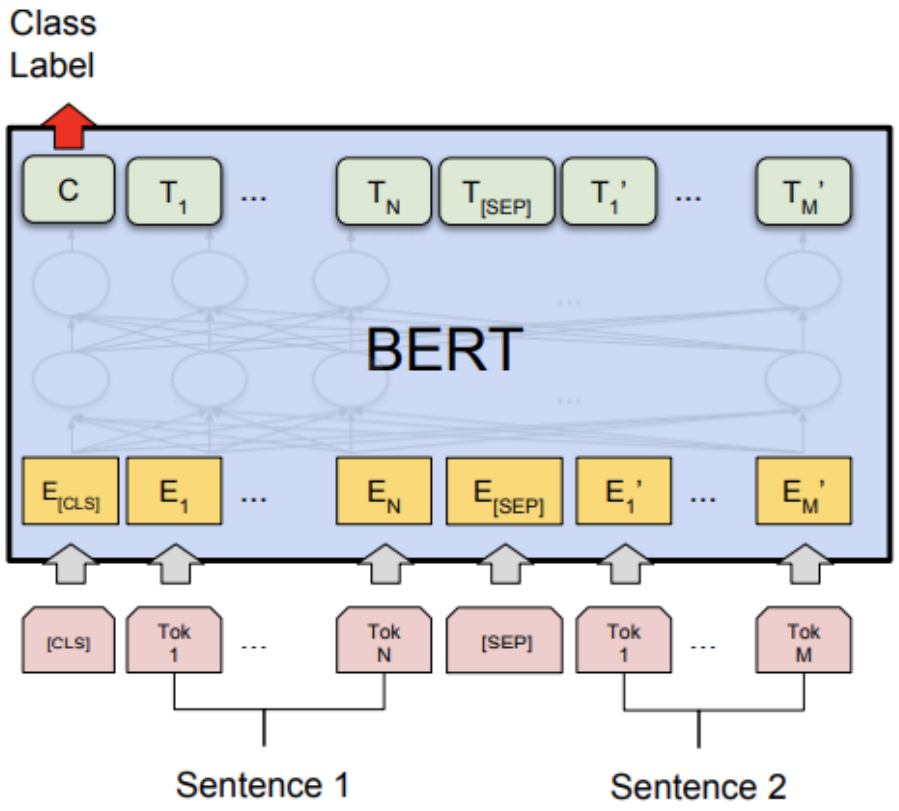
# 文本分类任务

| System | MNLI-(m/mm) 392k | QQP 363k | QNLI 108k | SST-2 67k | CoLA 8.5k | STS-B 5.7k | MRPC 3.5k | RTE 2.5k | **Average** - |
|---|---|---|---|---|---|---|---|---|---|
| Pre-OpenAI SOTA | 80.6/80.1 | 66.1 | 82.3 | 93.2 | 35.0 | 81.0 | 86.0 | 61.7 | 74.0 |
| BiLSTM+ELMo+Attn | 76.4/76.1 | 64.8 | 79.8 | 90.4 | 36.0 | 73.3 | 84.9 | 56.8 | 71.0 |
| OpenAI GPT | 82.1/81.4 | 70.3 | 87.4 | 91.3 | 45.4 | 80.0 | 82.3 | 56.0 | 75.1 |
| $\text{BERT}_{\text{BASE}}$ | 84.6/83.4 | 71.2 | 90.5 | 93.5 | 52.1 | 85.8 | 88.9 | 66.4 | 79.6 |
| $\text{BERT}_{\text{LARGE}}$ | **86.7/85.9** | **72.1** | **92.7** | **94.9** | **60.5** | **86.5** | **89.3** | **70.1** | **82.1** |

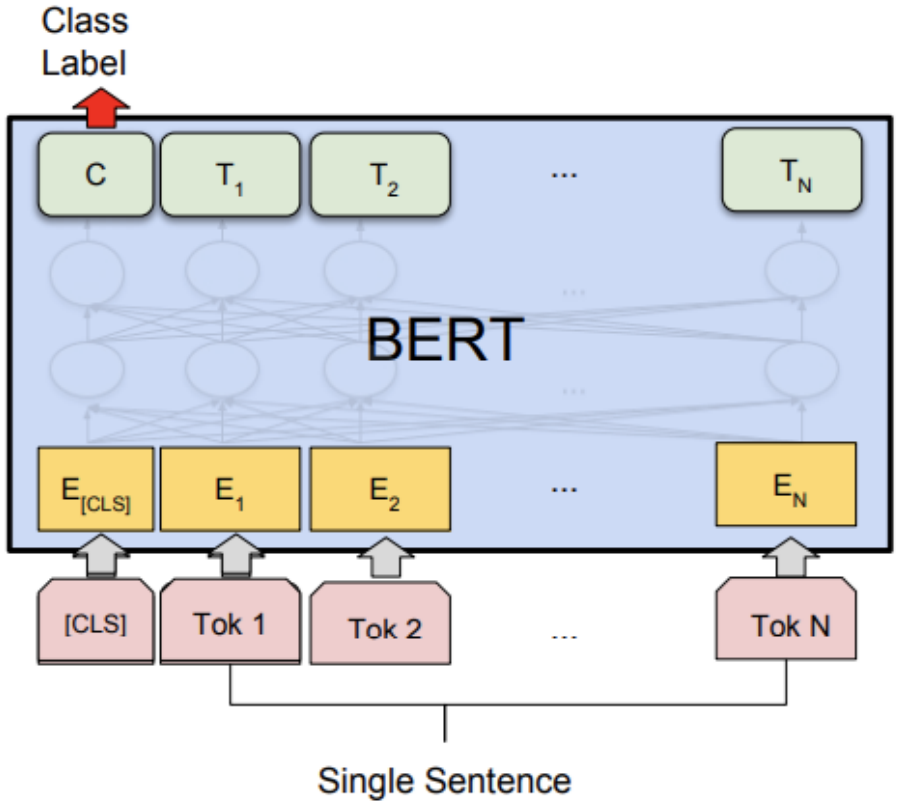
图片来自 BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding https://arxiv.org/pdf/1810.04805.pdf

https://gluebenchmark.com/leaderboard

# 文本分类任务




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




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

图片来自 BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding https://arxiv.org/pdf/1810.04805.pdf

# 问答系统

| System | Dev EM | Dev F1 | Test EM | Test F1 |
|---|---|---|---|---|
| Top Leaderboard Systems (Dec 10th, 2018) | | | | |
| Human | - | - | 82.3 | 91.2 |
| #1 Ensemble - nlnet | - | - | 86.0 | 91.7 |
| #2 Ensemble - QANet | - | - | 84.5 | 90.5 |
| Published | | | | |
| BiDAF+ELMo (Single) | - | 85.6 | - | 85.8 |
| R.M. Reader (Ensemble) | 81.2 | 87.9 | 82.3 | 88.5 |
| Ours | | | | |
| $\text{BERT}_{\text{BASE}}$ (Single) | 80.8 | 88.5 | - | - |
| $\text{BERT}_{\text{LARGE}}$ (Single) | 84.1 | 90.9 | - | - |
| $\text{BERT}_{\text{LARGE}}$ (Ensemble) | 85.8 | 91.8 | - | - |
| $\text{BERT}_{\text{LARGE}}$ (Sgl.+TriviaQA) | **84.2** | **91.1** | **85.1** | **91.8** |
| $\text{BERT}_{\text{LARGE}}$ (Ens.+TriviaQA) | **86.2** | **92.2** | **87.4** | **93.2** |

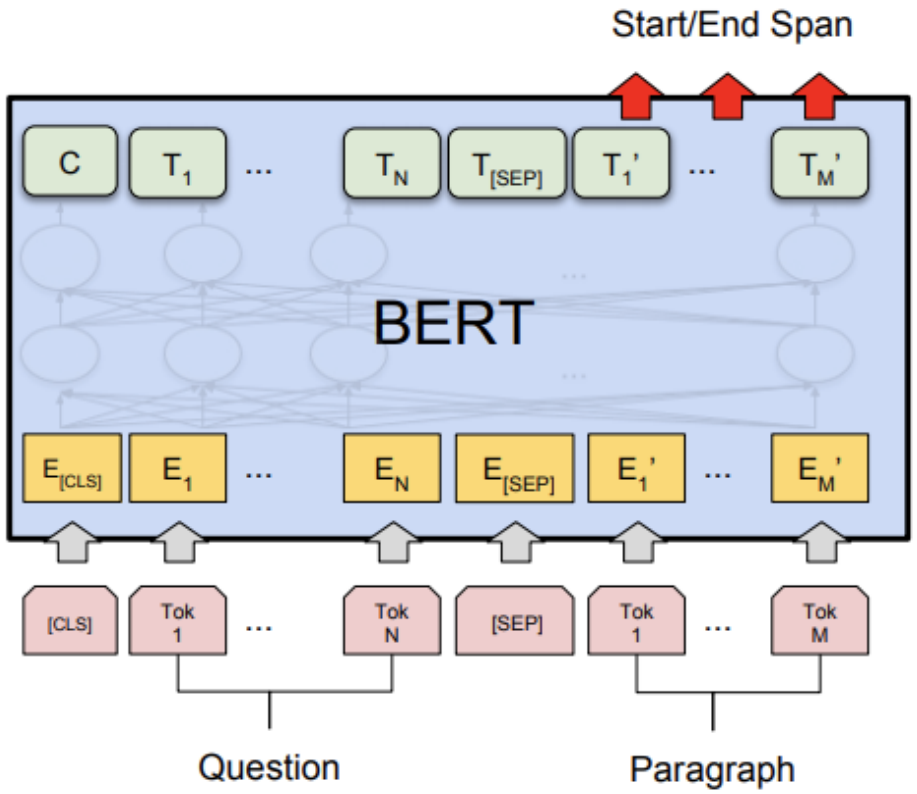
Table 2: SQuAD 1.1 results. The BERT ensemble is 7x systems which use different pre-training checkpoints and fine-tuning seeds.

| System | Dev EM | Dev F1 | Test EM | Test F1 |
|---|---|---|---|---|
| Top Leaderboard Systems (Dec 10th, 2018) | | | | |
| Human | 86.3 | 89.0 | 86.9 | 89.5 |
| #1 Single - MIR-MRC (F-Net) | - | - | 74.8 | 78.0 |
| #2 Single - nlnet | - | - | 74.2 | 77.1 |
| Published | | | | |
| unet (Ensemble) | - | - | 71.4 | 74.9 |
| SLQA+ (Single) | - | | 71.4 | 74.4 |
| Ours | | | | |
| $\text{BERT}_{\text{LARGE}}$ (Single) | 78.7 | 81.9 | 80.0 | 83.1 |

Table 3: SQuAD 2.0 results. We exclude entries that use BERT as one of their components.

图片来自 BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
https://arxiv.org/pdf/1810.04805.pdf

# 问答系统




(c) Question Answering Tasks:
SQuAD v1.1

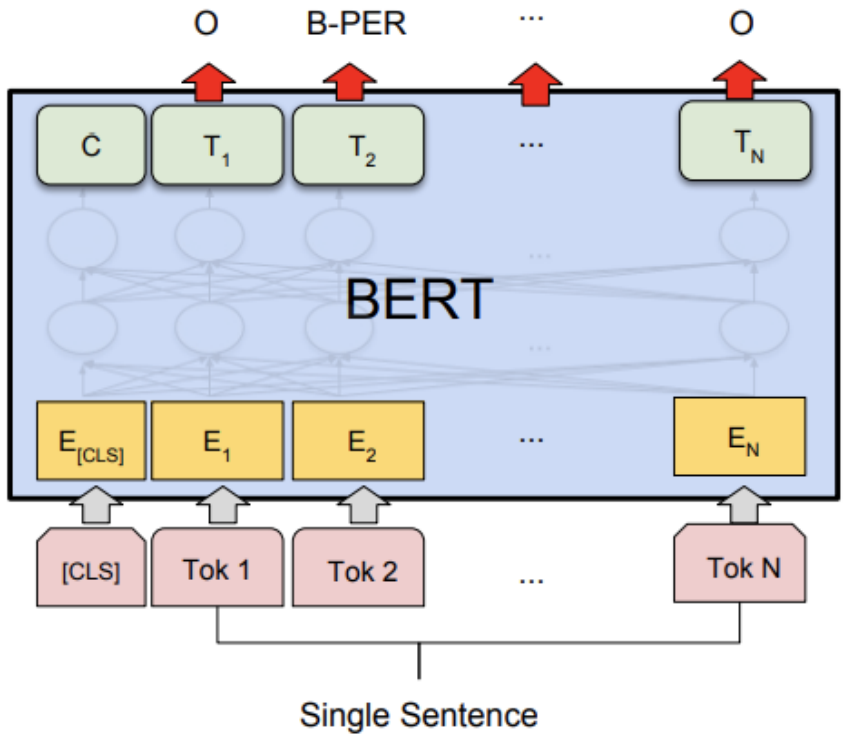
预测答案在文中出现的“开始”和“结束”位置

图片来自 BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
https://arxiv.org/pdf/1810.04805.pdf

# 命名实体识别

| System | Dev F1 | Test F1 |
|---|---|---|
| ELMo (Peters et al., 2018a) | 95.7 | 92.2 |
| CVT (Clark et al., 2018) | - | 92.6 |
| CSE (Akbik et al., 2018) | - | **93.1** |
| Fine-tuning approach | | |
| $\text{BERT}_{\text{LARGE}}$ | 96.6 | 92.8 |
| $\text{BERT}_{\text{BASE}}$ | 96.4 | 92.4 |
| Feature-based approach ($\text{BERT}_{\text{BASE}}$) | | |
| Embeddings | 91.0 | - |
| Second-to-Last Hidden | 95.6 | - |
| Last Hidden | 94.9 | - |
| Weighted Sum Last Four Hidden | 95.9 | - |
| Concat Last Four Hidden | 96.1 | - |
| Weighted Sum All 12 Layers | 95.5 | - |

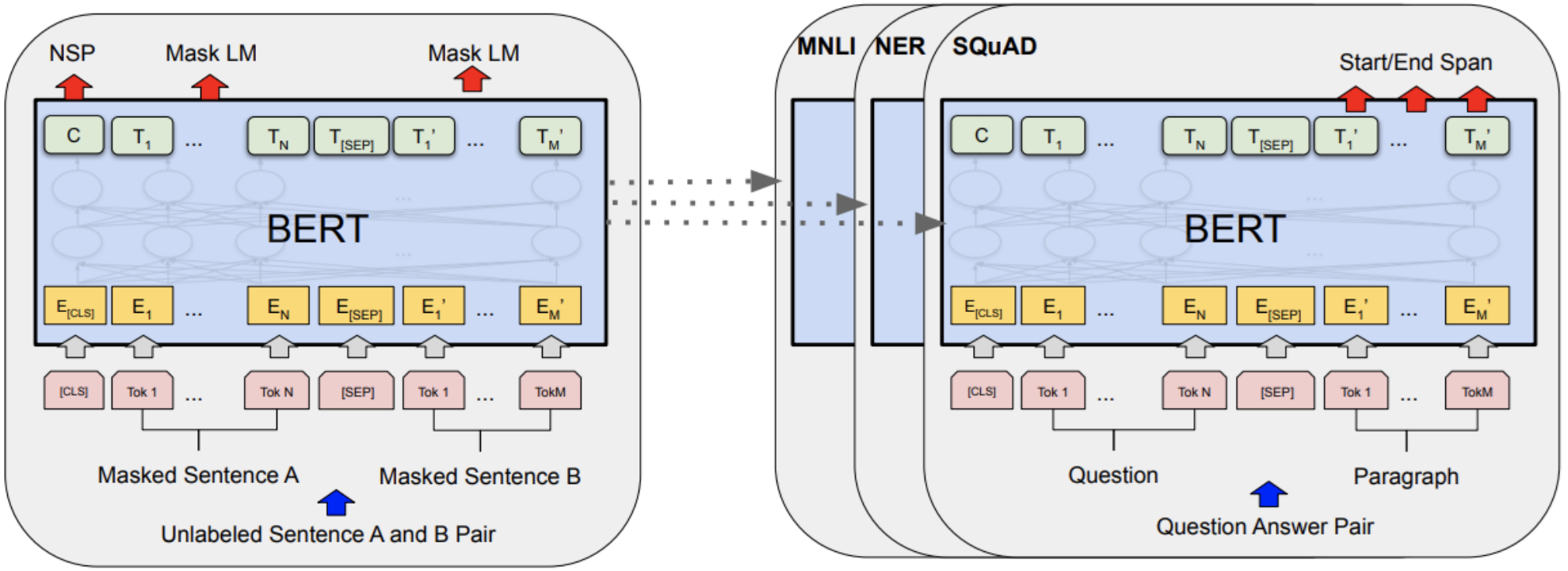
Table 7: CoNLL-2003 Named Entity Recognition results. Hyperparameters were selected using the Dev set. The reported Dev and Test scores are averaged over 5 random restarts using those hyperparameters.




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

图片来自 BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding https://arxiv.org/pdf/1810.04805.pdf

# 暴力美学：BERT模型的训练

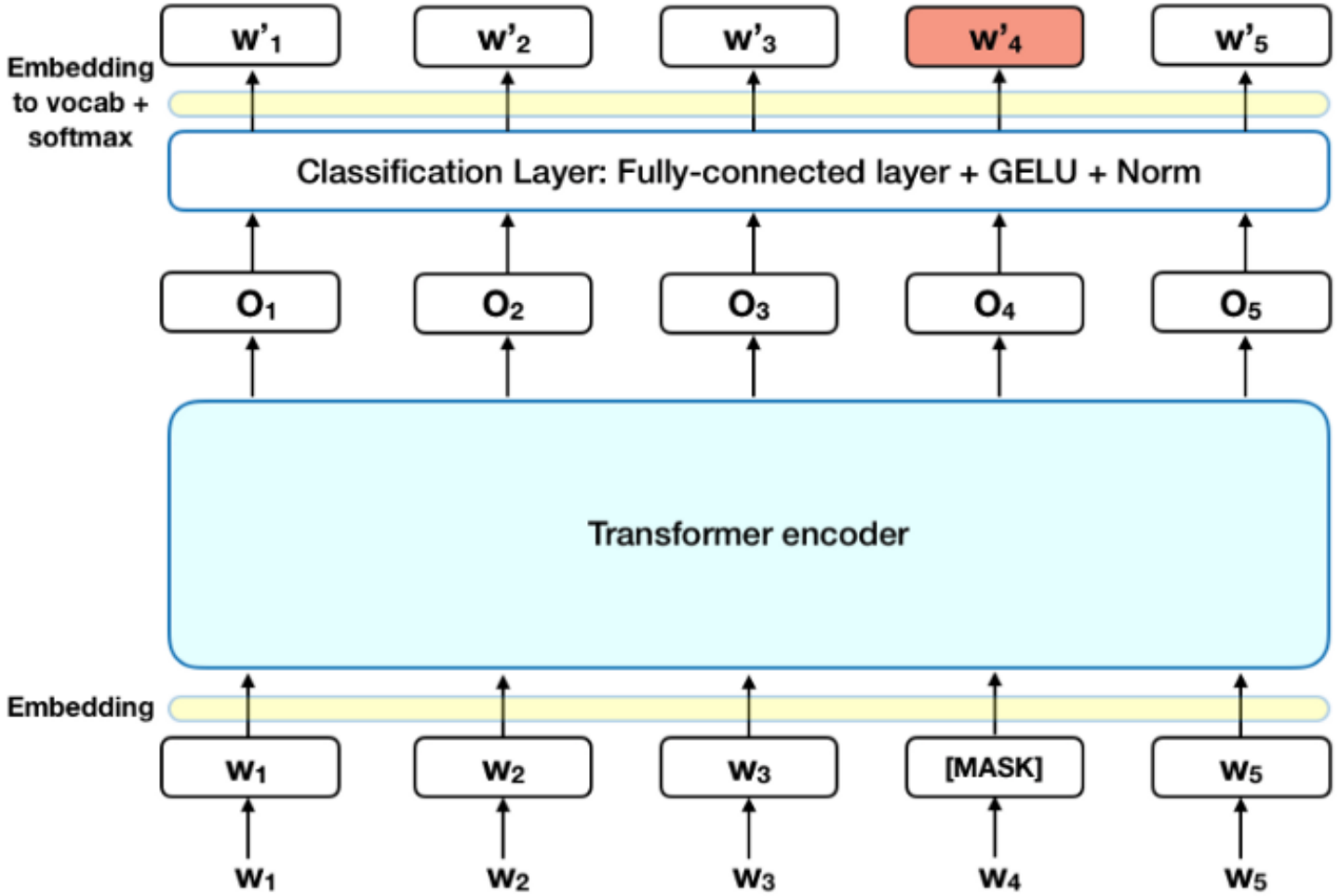



预训练　　微调

图片来自 BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

https://arxiv.org/pdf/1810.04805.pdf

# 暴力美学：BERT模型的训练




预测 [MASK]

图片来自 https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270

# 暴力美学：BERT模型的训练

MASKING策略：

- 15%的token会被替换成[MASK]，但是并不是这15%的token都会被替换，而是:
- 80%的时候，替换成[MASK]
- 10%的时候，换成另一个随机单词
- 10%的时候，不作任何替换

NSP训练：

- 判断两个句子A和B是在文中连续（相邻）的两个句子，还是不相邻的句子

# 暴力美学：BERT模型的训练

- 训练数据: Wikipedia (2.5B words) + BookCorpus (800M words)
- Batch Size: 131,072 words (1024 sequences * 128 length or 256 sequences * 512 length)
- 训练时间: 1M steps (~40 epochs)
- Optimizer: AdamW, 1e-4 learning rate, linear decay
- BERT-Base: 12-layer, 768-hidden, 12-head
- BERT-Large: 24-layer, 1024-hidden, 16-head
- Trained on 4x4 or 8x8 TPU slice for 4 days

# 比BERT更强的BERT

RoBERTa:

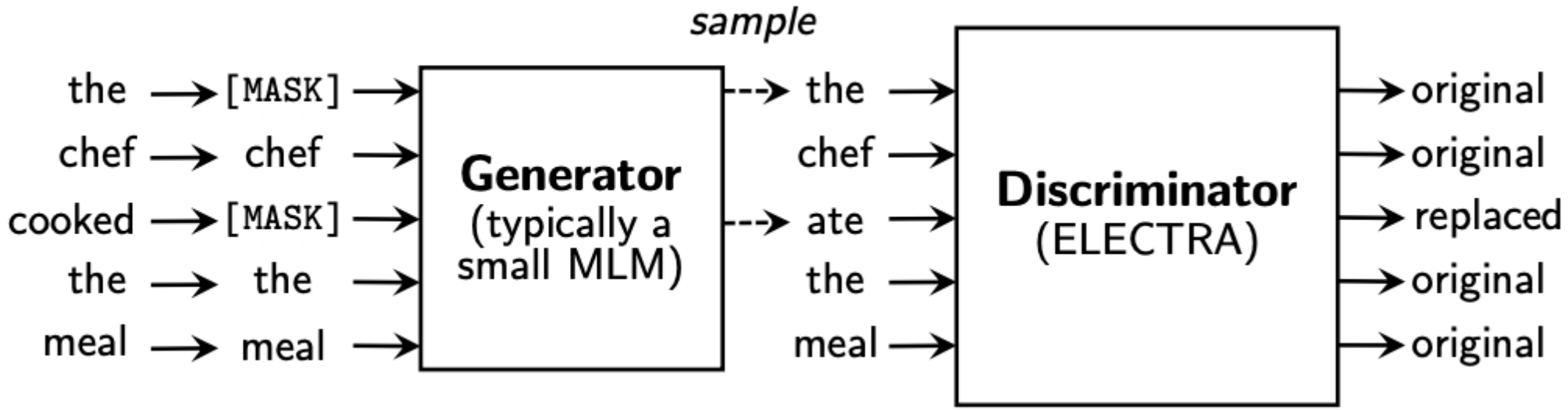
- 训练时间更长，epoch更多
- 采用更大量的训练数据

ALBERT：

- embedding层改为factorized embedding parameterization
- 层间参数共享
- 把NSP换成sentence ordering objective

# 比BERT更强的BERT

ELECTRA：




图片来自 ELECTRA: PRE-TRAINING TEXT ENCODERS AS DISCRIMINATORS RATHER THAN GENERATORS
https://arxiv.org/pdf/2003.10555.pdf

# 更多BERT

更小更快的BERT：

- DistilBERT
- TinyBERT
- MobileBERT
- FastBERT

融入更多知识的BERT：

- K-BERT

# 广告时间

NLP就业班第五期，报上“BERT模型深度修炼指南：NLP褚博士”的名号，可以获得优惠，详情请点击课程页面的咨询按钮 http://www.julyedu.com/weekend/nlpjiuye5

在直播课中提问的几位同学请联系以下几位老师之一（请不要重复添加）：

- 詹老师：julyedukefu_02
- 杨老师：julyedukefu05
- 杜老师：julyedukefu09

# 参考资料


- BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding https://arxiv.org/pdf/1810.04805.pdf
- Attention Is All You Need https://arxiv.org/pdf/1706.03762.pdf
- https://jalammar.github.io/illustrated-transformer/
- https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270
- http://web.stanford.edu/class/cs224n/slides/Jacob_Devlin_BERT.pdf
- ELECTRA: PRE-TRAINING TEXT ENCODERS AS DISCRIMINATORS RATHER THAN GENERATORS https://arxiv.org/pdf/2003.10555.pdf
- https://github.com/huggingface/transformers/blob/master/src/transformers/modeling_bert.py

# 谢谢大家

# 中文预训练模型

- bert-base-chinese
- ernie
- 哈工大训练的模型 https://github.com/ymcui/Chinese-BERT-wwm 有很多

知识图谱， triple

subject-predicate-object

杭州-省会-浙江

北京-首都-中国

entity-relation-entity

实体—>embedding

关于实体的描述 —> encoder —> embedding

实体在知识图谱当中的位置 —> encoder —> embedding

GNN, GCN graph convolutional network, kipf 2015

GPT-3