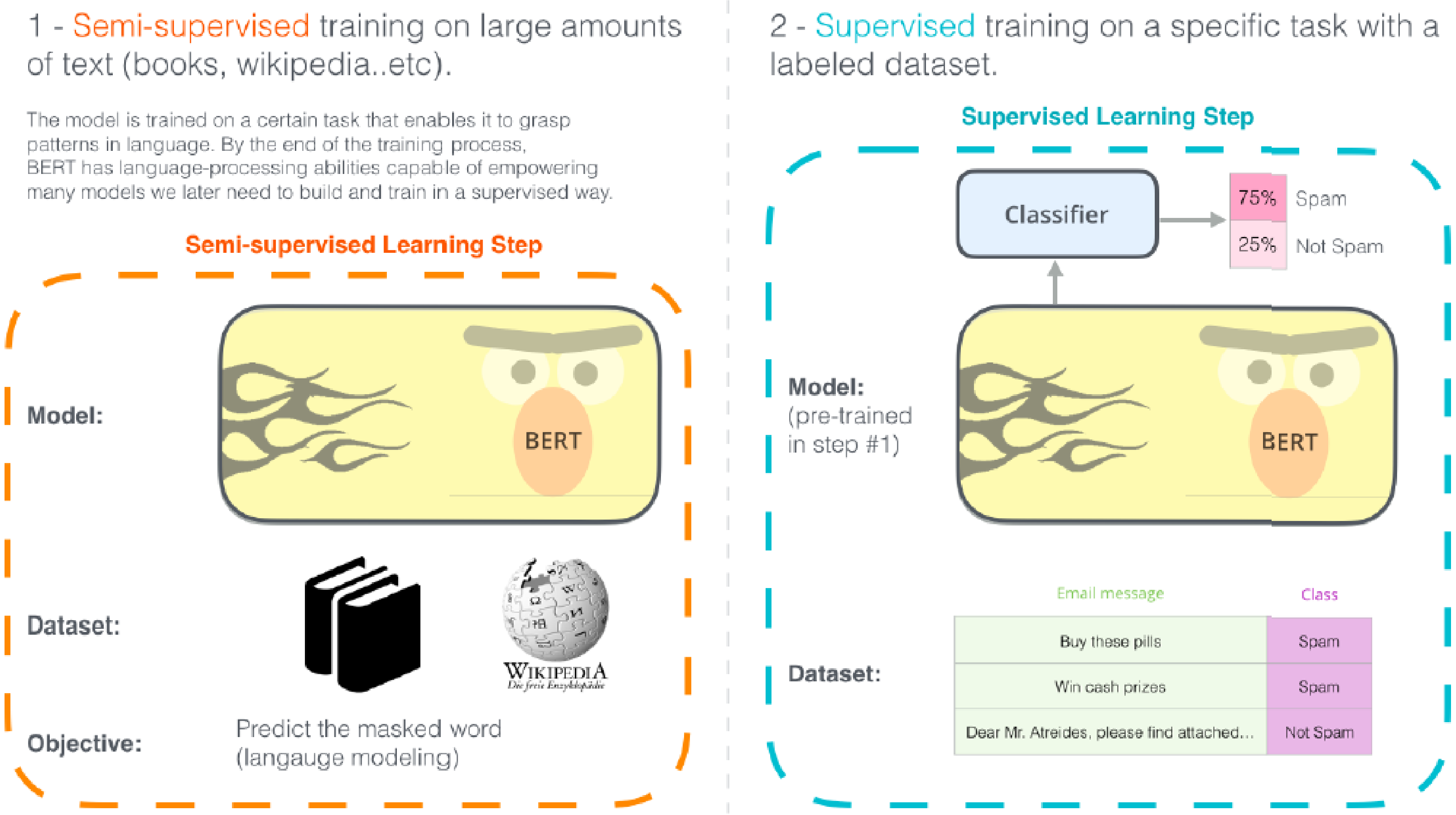
Bert 简介

Bidirectional Encoder Representations from Transformers

蔡云麒

Bert——NLP预训练模型：

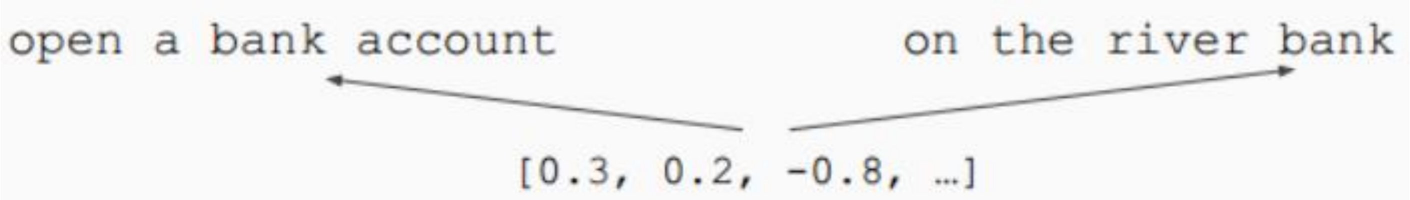
- ◆1.训练数据少，不足以训练复杂的网络
- ◆2.加快训练速度
- ◆3.参数初始化，先找到好的初始点，有利于优化。



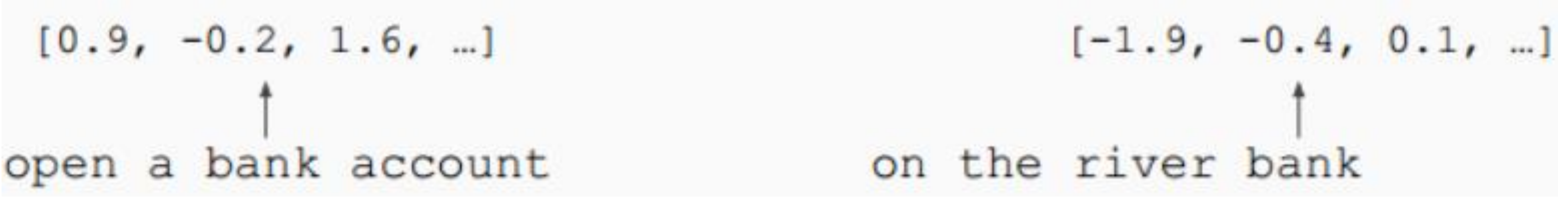
背景：从Word Embedding说起

说到利用深度学习来进行自然语言处理，必然绕不开的一个问题就是“Word Embedding”也就是将词转换为计算机能够处理的向量。

随之而来的人们也碰到到了一个根本性的问题，我们通常会面临这样的一个问题，同一个单词在不同语境中的**一词多义**问题

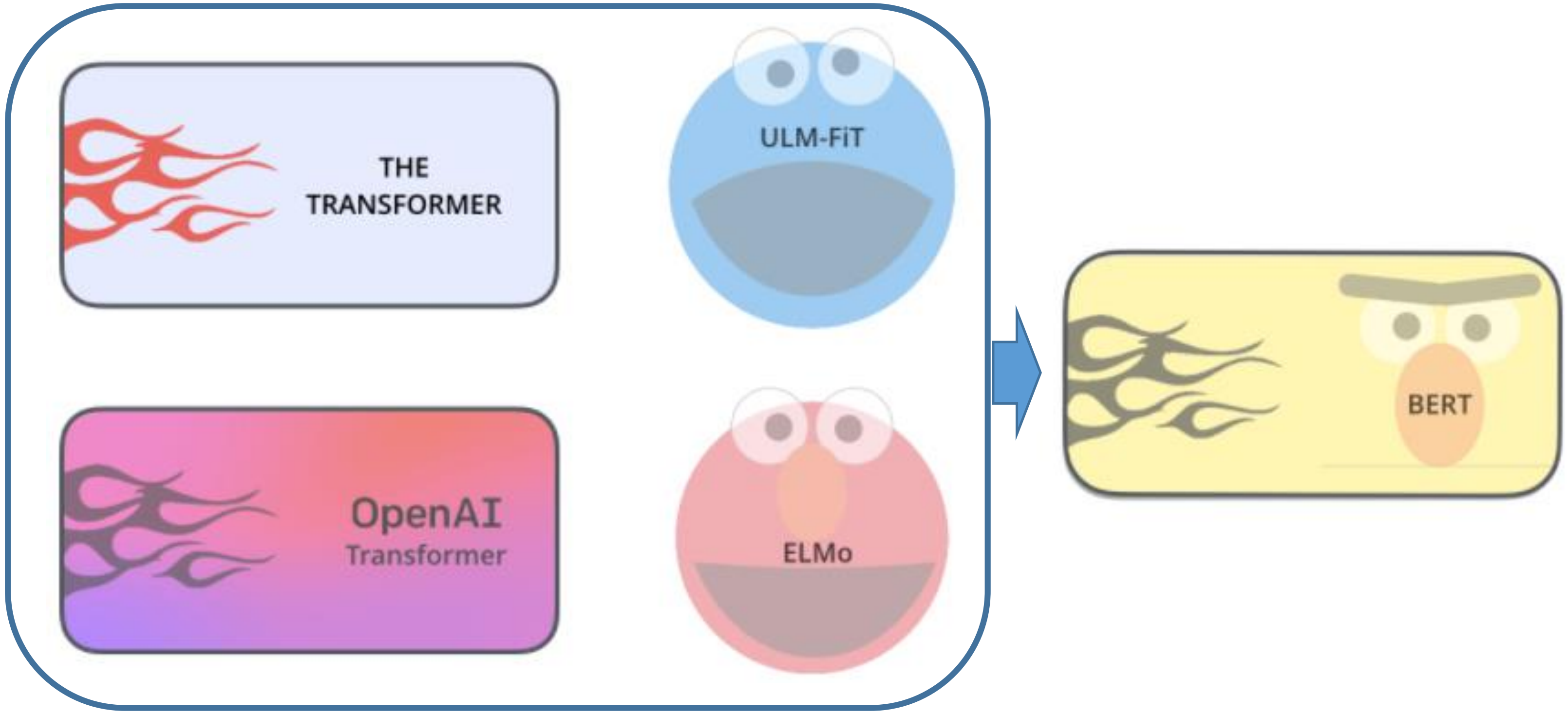


研究人员对此也想到了对应的解决方案，例如在大语料上训练**语境表示**，从而得到不同的上下文情况的不同向量表示。



Bert有什么黑科技?

- Bert在模型层面上并没有新的突破，准确来说它更像是NLP领域近期优秀模型的**集大成者** 

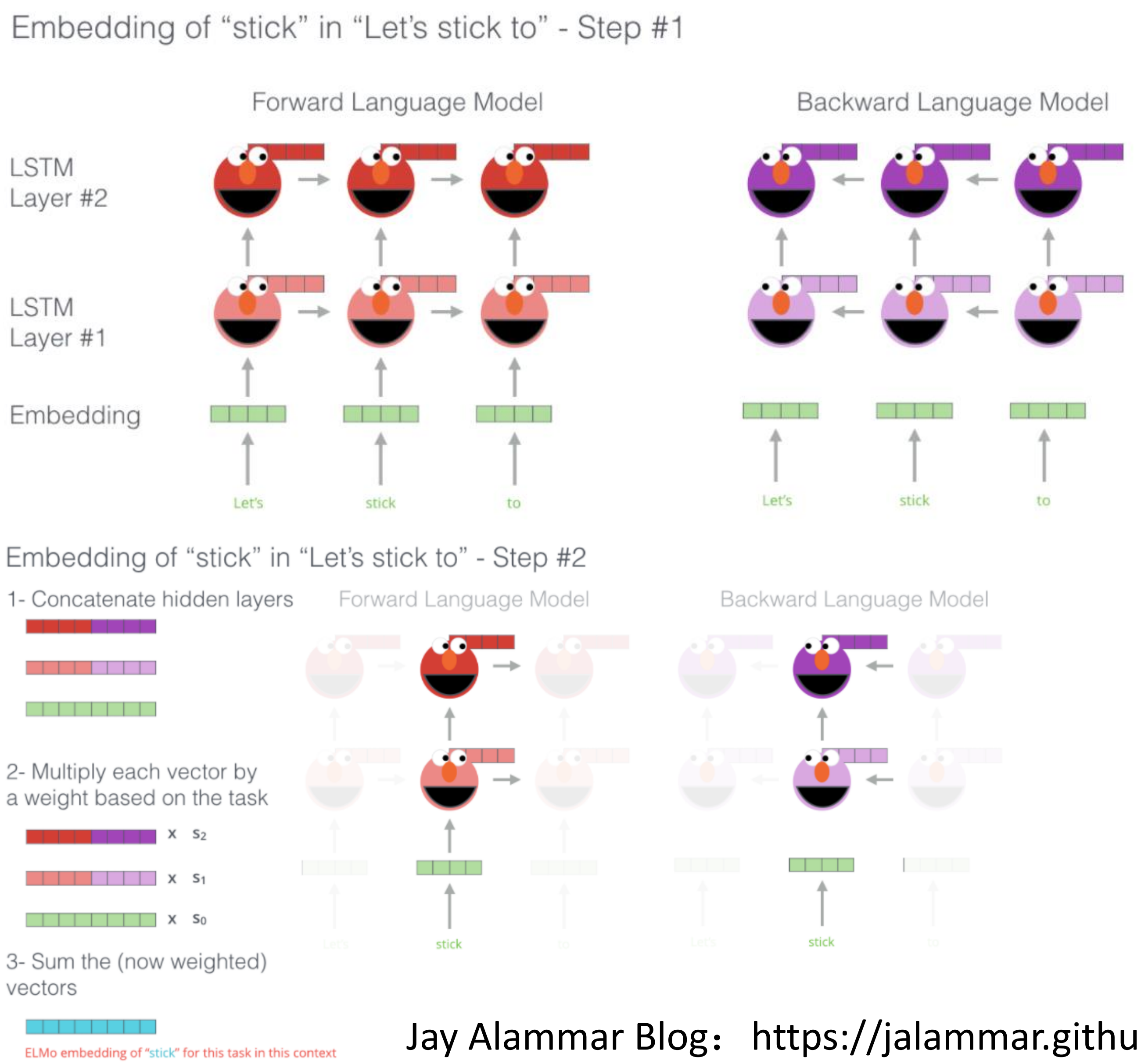


*Semi-supervised Sequence Learning (by Andrew Dai 和 Quoc Le)；ELMo (by Matthew Peters和来自 AI2 and UW CSE的研究人员)；ULMFiT (by fast.ai 创始人 Jeremy Howard 和 Sebastian Ruder)； OpenAI transformer (by OpenAI 研究员Radford, Narasimhan, Salimans, and Sutskever)；Transformer (Vaswani et al)。*Jay Alammar Blog：https://jalammar.github.io/

ELMo的语境学习

*ELMo*实际上训练了一个**双向的LSTM**——这样它的语言模型不仅能预测下一个词，还有预测上一个词.

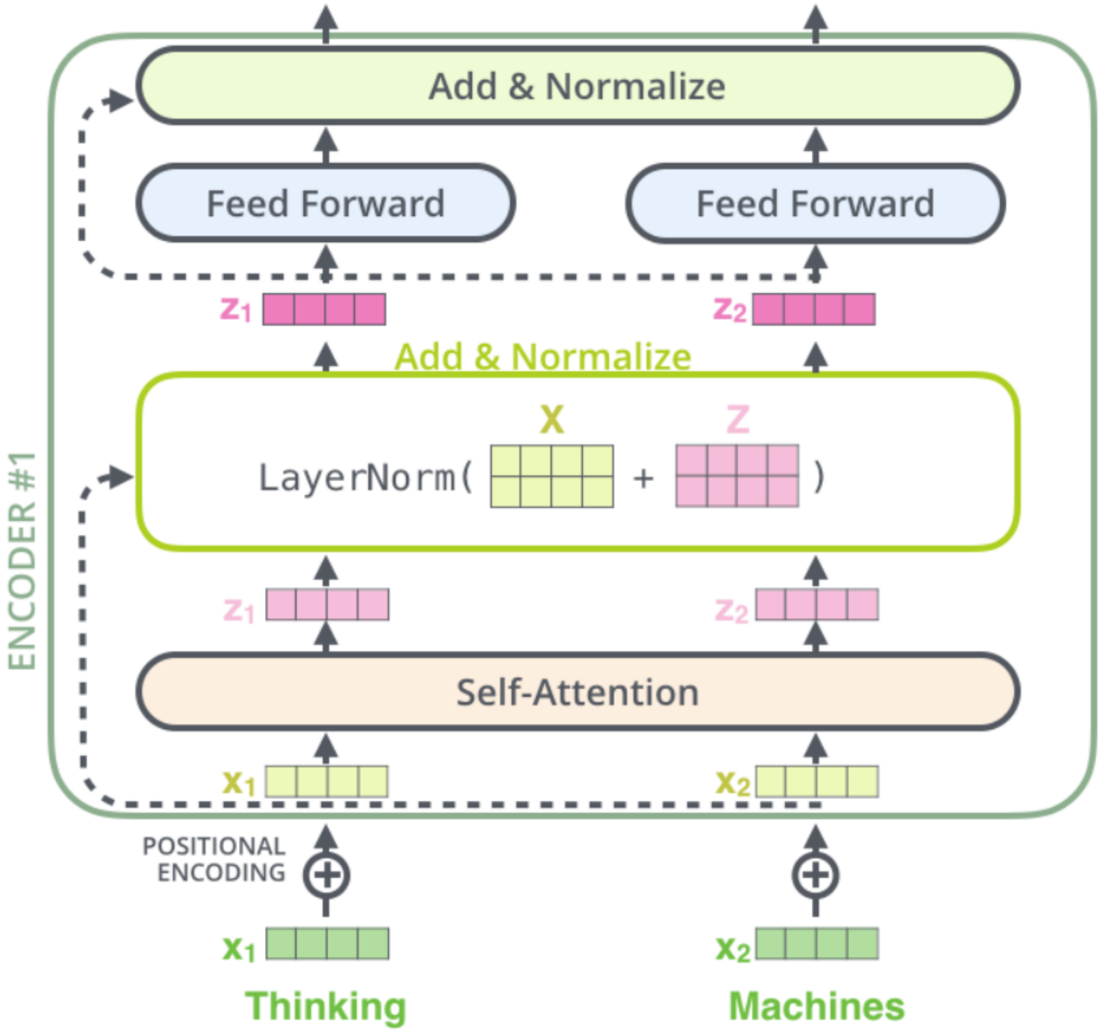
通过将隐藏状态(和初始嵌入)以某种方式(拼接之后加权求和)组合在一起, ELMo提出了**语境化的词嵌入**。



Jay Alammar Blog： https://jalammar.github.io/

GPT的Transformer

Transformer是一个非常强大的特征提取器，相比于LSTM它不管词间距离的长短，更能处理**长期依赖关系**。并且具有**可并行**的优势。



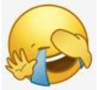
Jay Alammar Blog：https://jalammar.github.io/

Bert的兼收并蓄

特征提取能力 

语境表达能力 

特征提取能力 

语境表达能力 

特征提取能力 

语境表达能力 

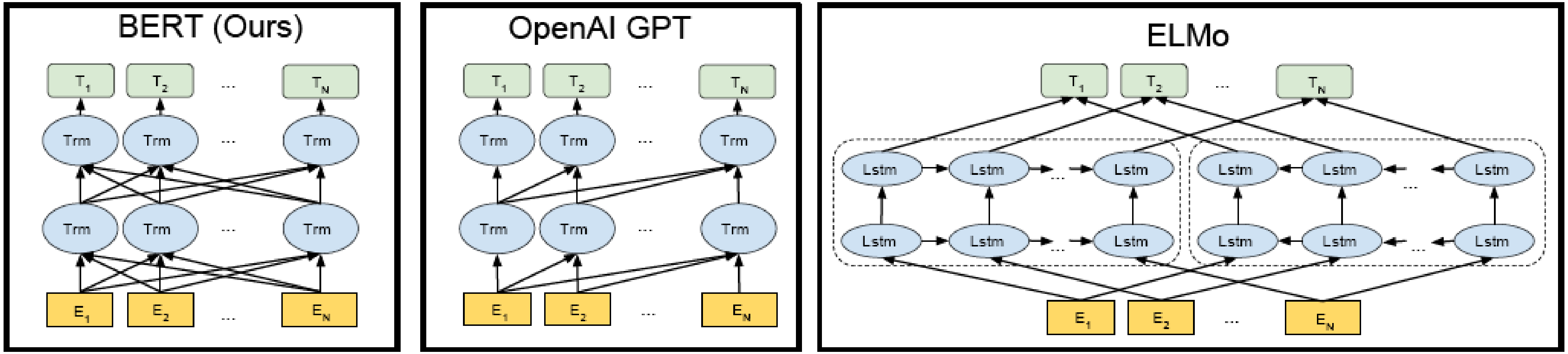


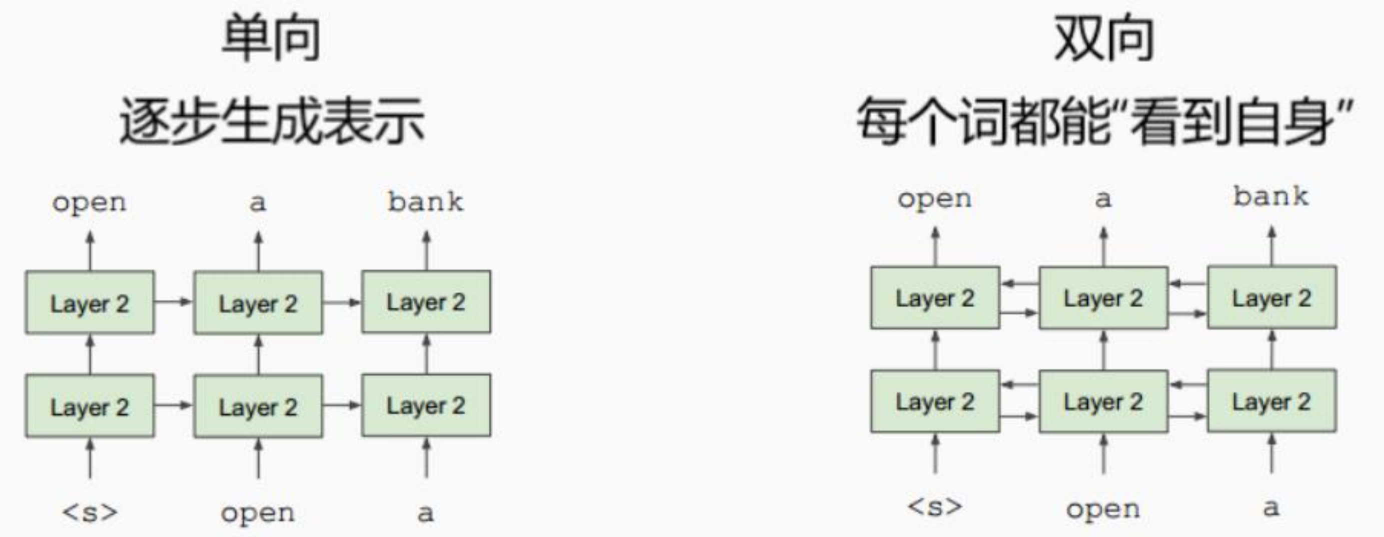
Figure 1: Differences in pre-training model architectures. BERT uses a bidirectional Transformer. OpenAI GPT uses a left-to-right Transformer. ELMo uses the concatenation of independently trained left-to-right and right-to-left LSTM to generate features for downstream tasks. Among three, only BERT representations are jointly conditioned on both left and right context in all layers.

*Jacob Devlin Ming-Wei Chang Kenton Lee Kristina Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, **arXiv:1810.04805v1***

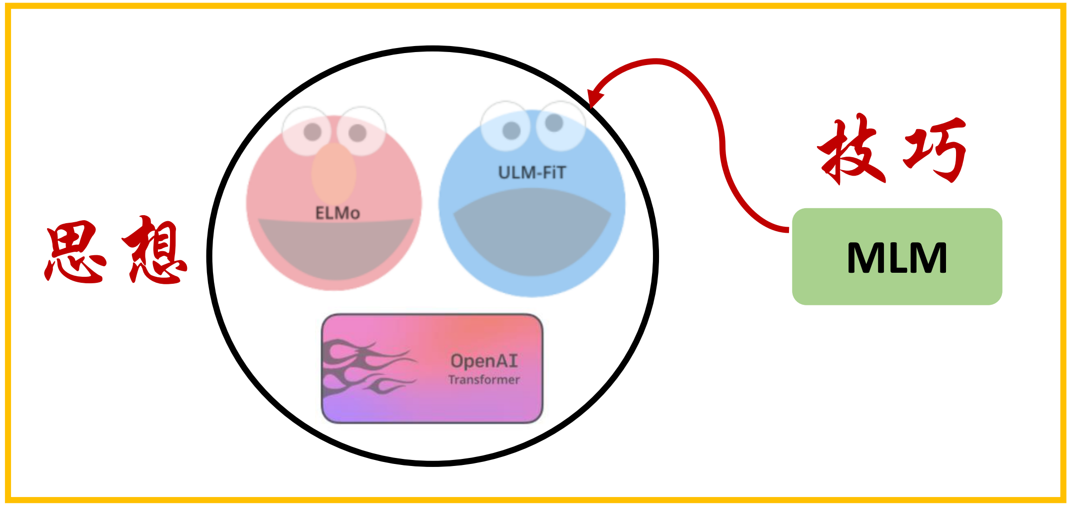
Bert需要克服的最大障碍

GPT中存在的**最大问题**：

Transformer虽然强大，但是如果使用双向的编码器，会产生一些循环，在这些循环中，单词会间接地“**窥见**”自己



Bert采用的解决办法： “masked language model” (MLM) ； *Cloze task (Taylor,1953)*



Bert采用了一个非常巧妙的小技巧将各家的思想恰到好处的融合起来了

Bert的效果

刷新了目前几乎所有的NLP记录

System	MNLI-(m/mm)	QQP	QNLI	SST-2	CoLA	STS-B	MRPC	RTE	Average
	392k	363k	108k	67k	8.5k	5.7k	3.5k	2.5k	-
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	79.9	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	88.1	91.3	45.4	80.0	82.3	56.0	75.2
$\text{BERT}_{\text{BASE}}$	84.6/83.4	71.2	90.1	93.5	52.1	85.8	88.9	66.4	79.6
$\text{BERT}_{\text{LARGE}}$	**86.7/85.9**	**72.1**	**91.1**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**81.9**

MultiNLI

Premise: Hills and mountains are especially sanctified in Jainism.
Hypothesis: Jainism hates nature.
Label: Contradiction

CoLa

Sentence: The wagon rumbled down the road.
Label: Acceptable

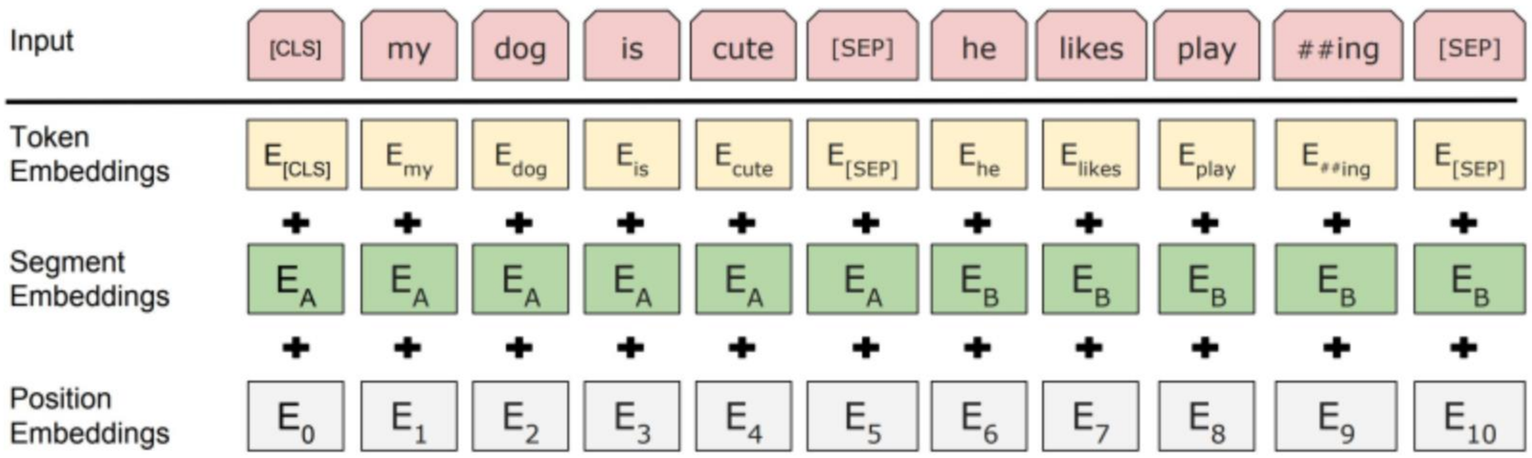
Sentence: The car honked down the road.
Label: Unacceptable

来自Bert论文中的一句话：**the bidirectional nature of our model is the single most important new contribution.**

Bert的运转过程

Input_data准备

第一步：数据预处理




```
*** Example ***
tokens: [CLS] for more than a [MASK] up [MASK] great main street , [MASK] [MASK] [MASK] centre of the city , at right angles ,
input_ids: 101 2005 2062 2084 1037 103 2039 103 2307 2364 2395 1010 103 103 103 2803 1997 1996 2103 1010 2012 2157 12113 1010
input_mask: 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
segment_ids: 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
masked_lm_positions: 5 7 12 13 14 24 32 55 59 60 71 94 103 104 105 109 112 114 117 0
masked_lm_ids: 3542 1996 4625 1999 1996 2011 1997 5472 1996 5532 1997 2008 1037 16465 10427 1998 2037 2000 2098 0
masked_lm_weights: 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 0.0
next_sentence_labels: 1
```

Bert的运转过程

模型训练

第二步:Bert训练的核心思想

① **Masked LM**

输入序列中每个单词有K%的几率被替换

K=15%

```
                       store              gallon
                         ↑                  ↑
the man went to the [MASK] to buy a [MASK] of milk
```

每一次以80%的概率用[MASK]替换

went to the store → went to the [MASK]

每一次以10%的概率随机替换

went to the store → went to the running

每一次以10%的概率不进行替换

went to the store → went to the store

② **预测下一句**

Sentence A = The man went to the store.
Sentence B = He bought a gallon of milk.
Label = IsNextSentence

Sentence A = The man went to the store.
Sentence B = Penguins are flightless.
Label = NotNextSentence

Bert的运转过程

模型训练

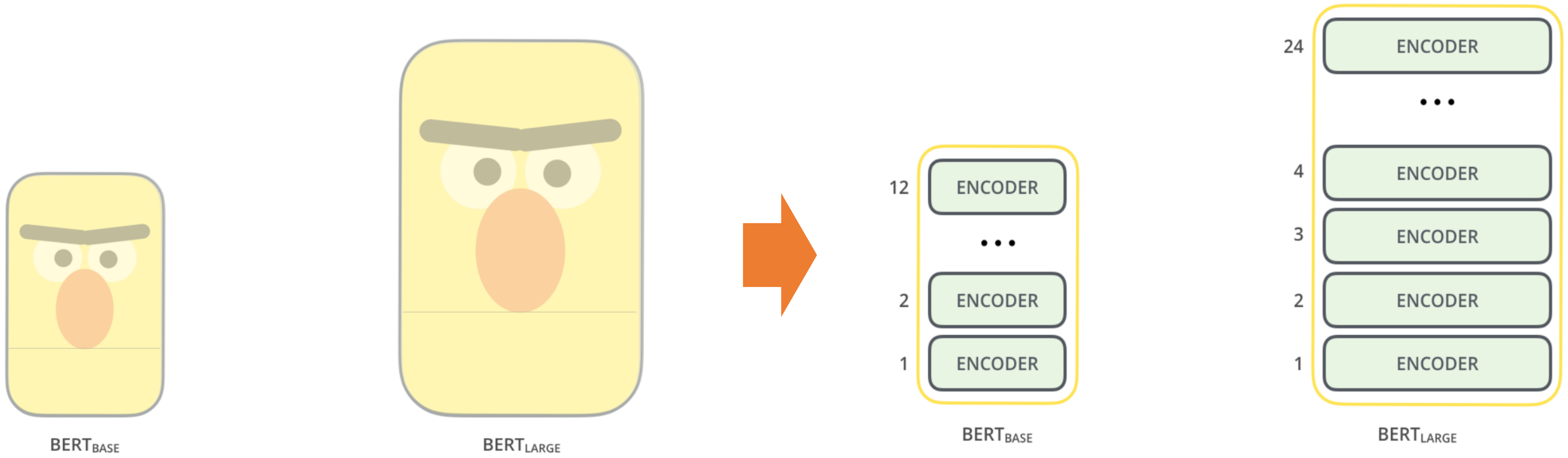
遮罩策略的影响：

Masking Rates			Dev Set Results		
MASK	SAME	RND	MNLI Fine-tune	NER Fine-tune	NER Feature-based
80%	10%	10%	84.2	95.4	94.9
100%	0%	0%	84.3	94.9	94.0
80%	0%	20%	84.1	95.2	94.6
80%	20%	0%	84.4	95.2	94.7
0%	20%	80%	83.7	94.8	94.6
0%	0%	100%	83.6	94.9	94.6

Jacob Devlin Ming-Wei Chang Kenton Lee Kristina Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, ***arXiv:1810.04805v1***

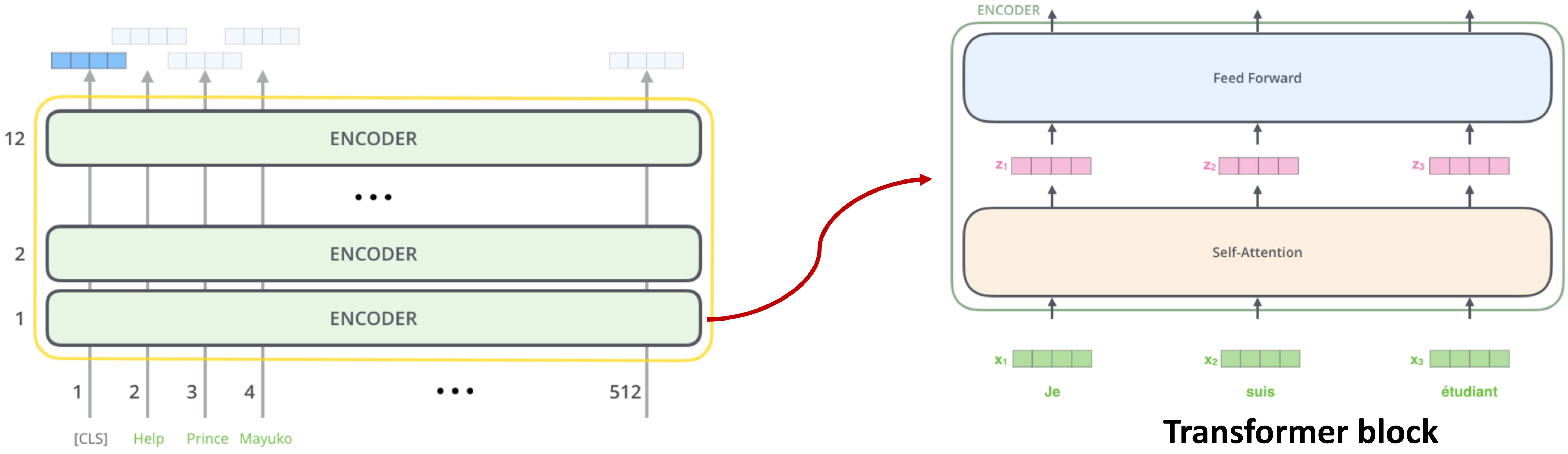
- 一次遮住100%的词让基于特征的方法在性能上有所下降
- 而100%使用随机替换的方式让基于特征的方法在性能上下降幅度较小

Bert模型架构



Jay Alammar Blog：https://jalammar.github.io/

Bert模型架构



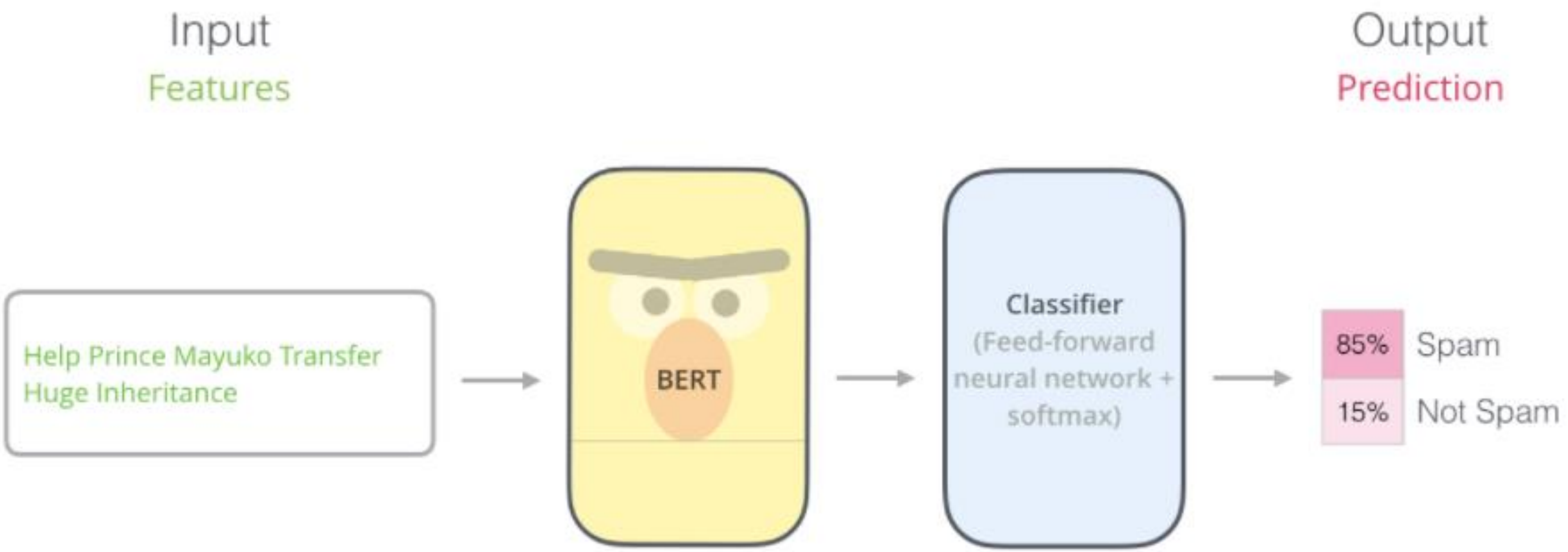
BERT

Transformer block

号称：*NLP最强特征提取器*

Bert有哪些用途?

例子：句子分类



训练数据:

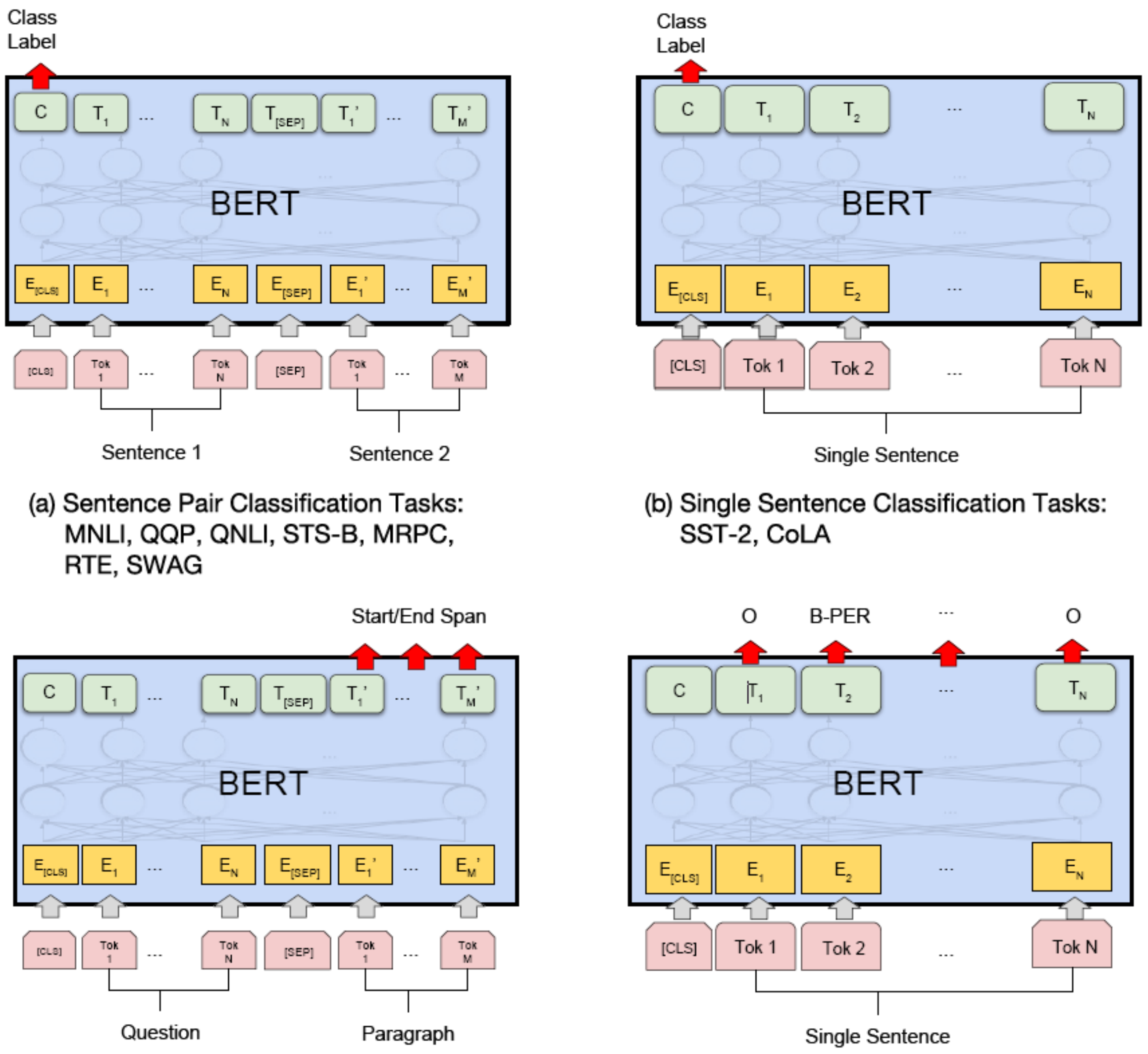
Email message	Class
Buy these pills	Spam
Win cash prizes	Spam
Dear Mr. Atreides, please find attached…	Not Spam

这种方式源于:

- **Semi-supervised Sequence Learning**
- **ULMFiT**

Bert用来解决特定任务

- **情感分析**
  - 输入: 一条影评/商品评价。
  - 输出: 正面评价还是负面评价?
  - 数据集如: [SST]()
- **事实核查**
  - 输入:一个句子。
  - 输出: 是不是一个断言
- **阅读理解任务**
- **文本摘要任务**
- **序列标注任务**
- **等等**



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

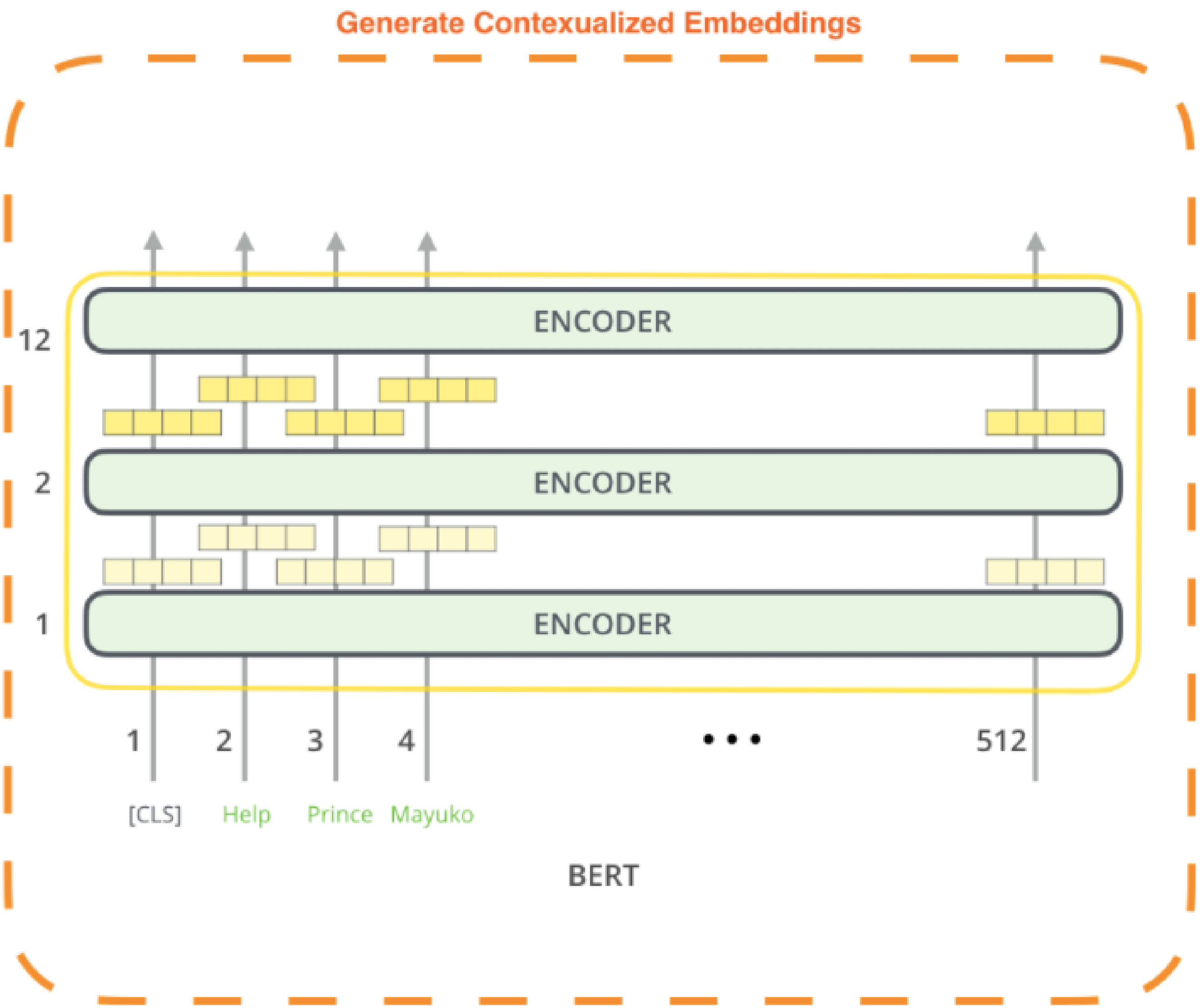
(b) Single Sentence Classification Tasks: SST-2, CoLA

(c) Question Answering Tasks: SQuAD v1.1

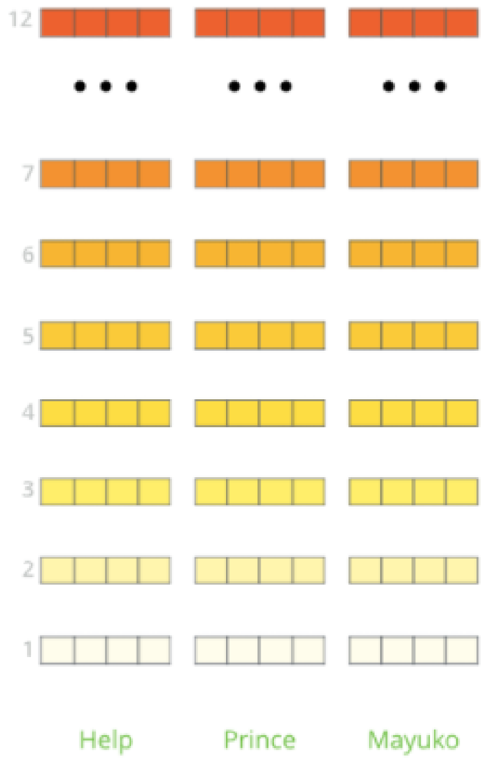
(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

arXiv:1810.04805v1

Bert用于特征提取

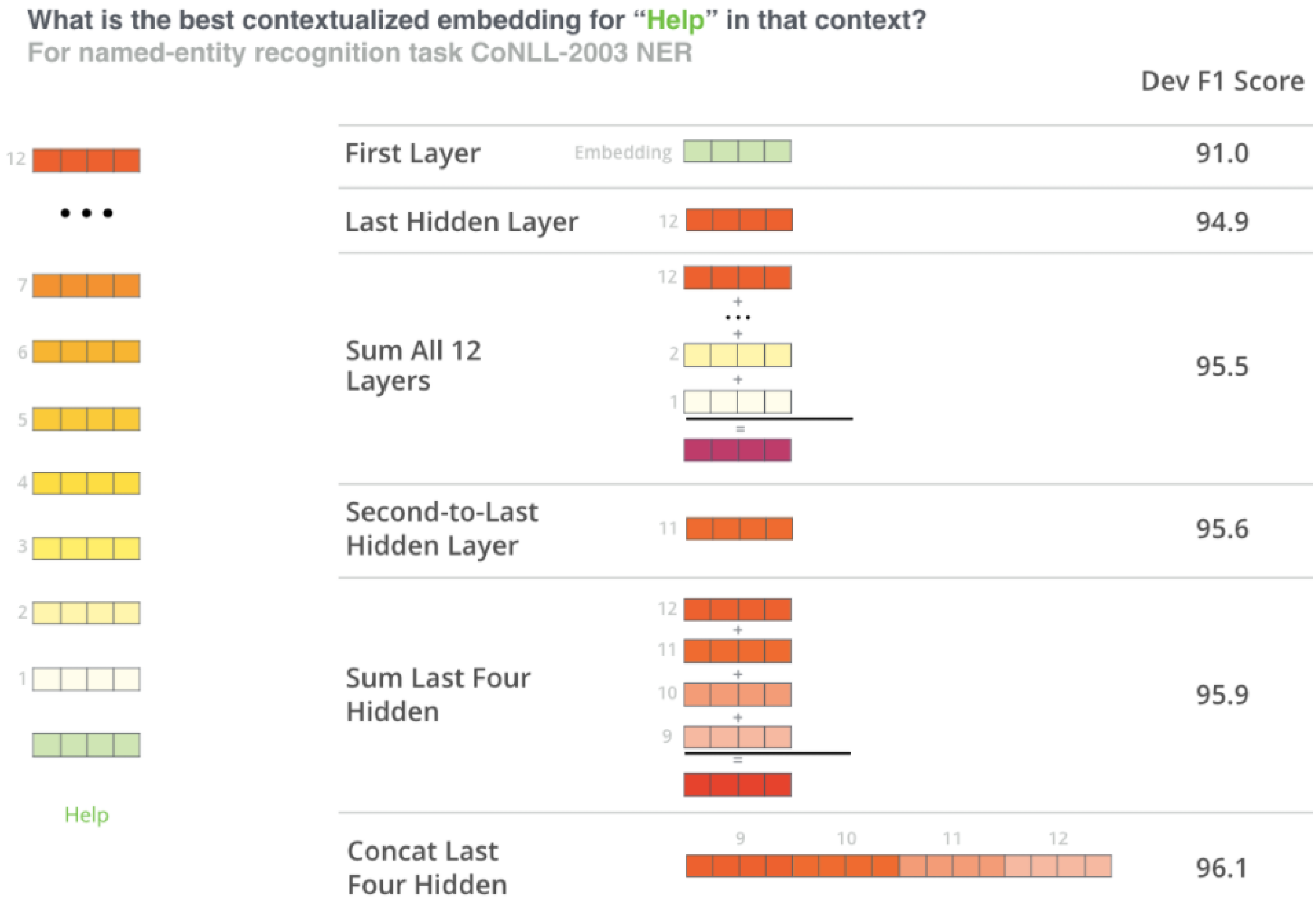


The output of each encoder layer along each token's path can be used as a feature representing that token.



But which one should we use?

Bert用于特征提取



Layers	Dev F1
Finetune All	96.4
First Layer (Embeddings)	91.0
Second-to-Last Hidden	95.6
Last Hidden	94.9
Sum Last Four Hidden	95.9
Concat Last Four Hidden	96.1
Sum All 12 Layers	95.5

Table 7: Ablation using BERT with a feature-based approach on CoNLL-2003 NER. The activations from the specified layers are combined and fed into a two-layer BiLSTM, without backpropagation to BERT.

arXiv:1810.04805v1

Bert模型版本及细节

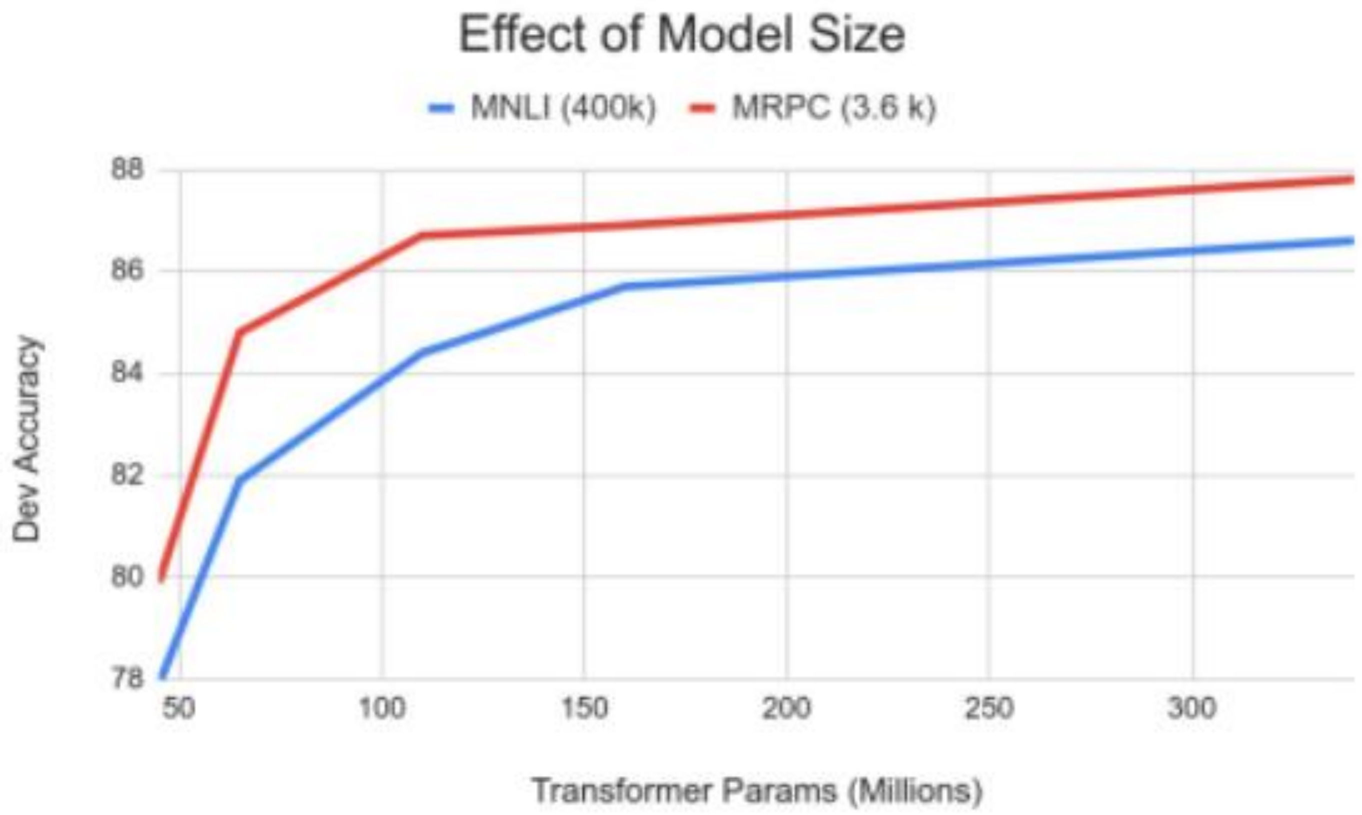
模型版本：

- `BERT-Base, Uncased`: 12-layer, 768-hidden, 12-heads, 110M parameters
- `BERT-Large, Uncased`: 24-layer, 1024-hidden, 16-heads, 340M parameters
- `BERT-Base, Cased`: 12-layer, 768-hidden, 12-heads , 110M parameters
- `BERT-Large, Cased`: 24-layer, 1024-hidden, 16-heads, 340M parameters
- `BERT-Base, Multilingual Cased (New, recommended)`: 104 languages, 12-layer, 768-hidden, 12-heads, 110M parameters
- `BERT-Base, Multilingual Uncased (Orig, not recommended)` **(Not recommended, use `Multilingual Cased` instead)**: 102 languages, 12-layer, 768-hidden, 12-heads, 110M parameters
- `BERT-Base, Chinese`: Chinese Simplified and Traditional, 12-layer, 768-hidden, 12-heads, 110M parameters

模型细节：

语料：Wikipedia (2.5B words) + BookCorpus (800M words)
Batch Size: 131,072 个词 (1024 个序列* 每个序列128个词或者256个序列 *每个序列512个词)
训练步数：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
在4x4 或 8x8 的TPU slice上训练了4天

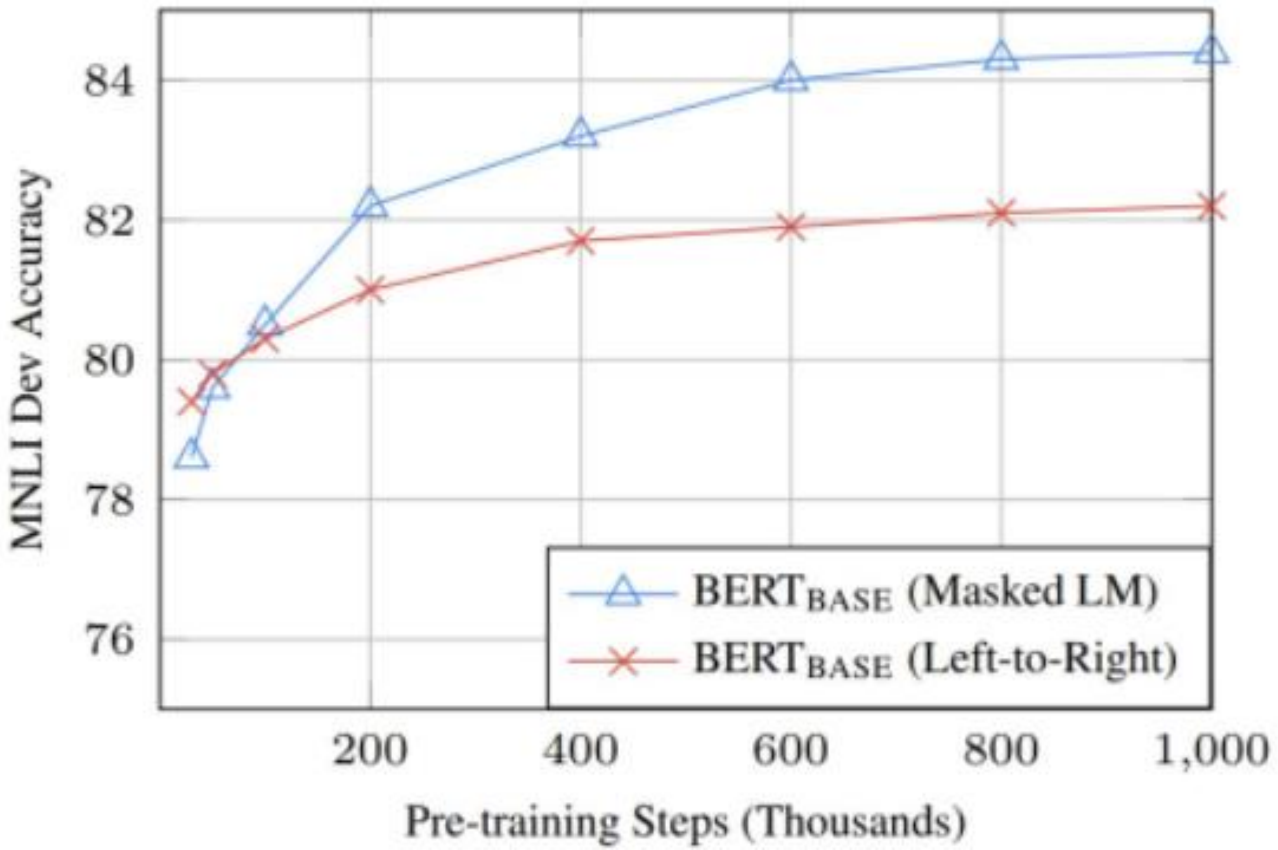
Bert模型版本及细节



arXiv:1810.04805v1

- 大模型更有效
- 即使在仅有3,600个标注样本的数据集上，把参数从110M调整到340M也带来了性能提升
- 性能提升并不是渐进的

Bert模型版本及细节



arXiv:1810.04805v1

- 由于仅仅预测15%的词，因此Masked LM收敛速度仅仅慢了一点点
- Masked LM的绝对结果显然更好