昇思MindSpore技术公开课
大模型专题

BERT

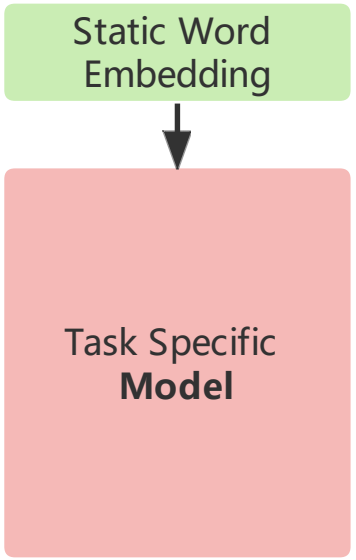
[M]S 昇思 MindSpore

目录

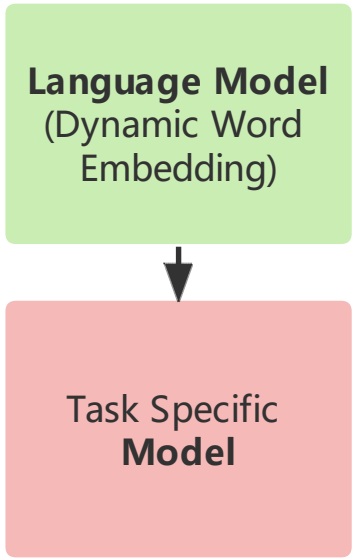
NLP中的预训练模型

语言模型的演变经历了以下几个阶段：

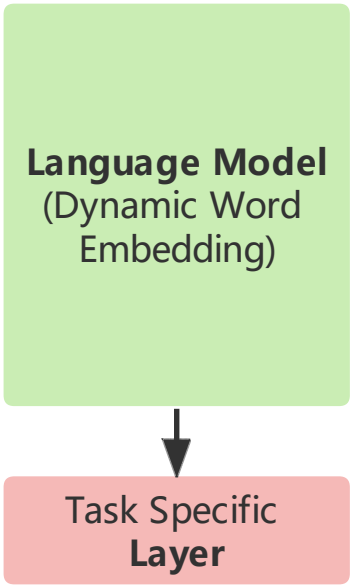
Word2Vec/Glove



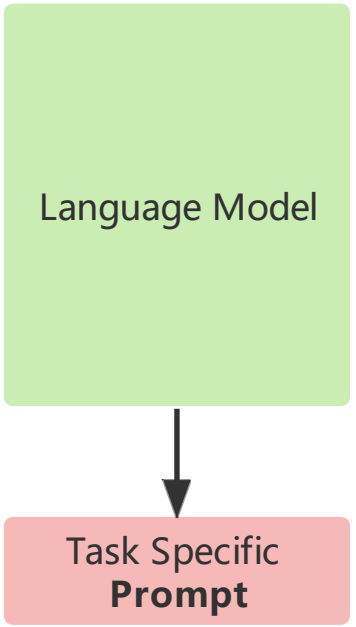
ELMo



BERT (Fine Tuning)

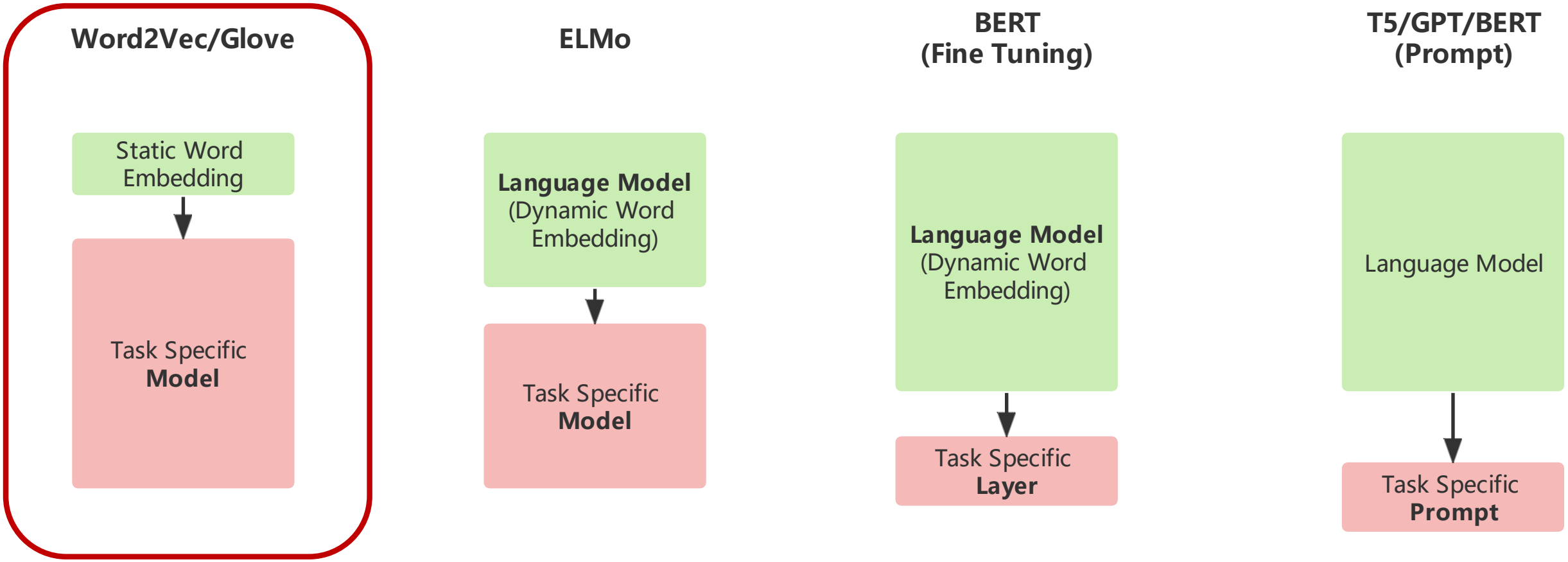


T5/GPT/BERT (Prompt)



NLP中的预训练模型

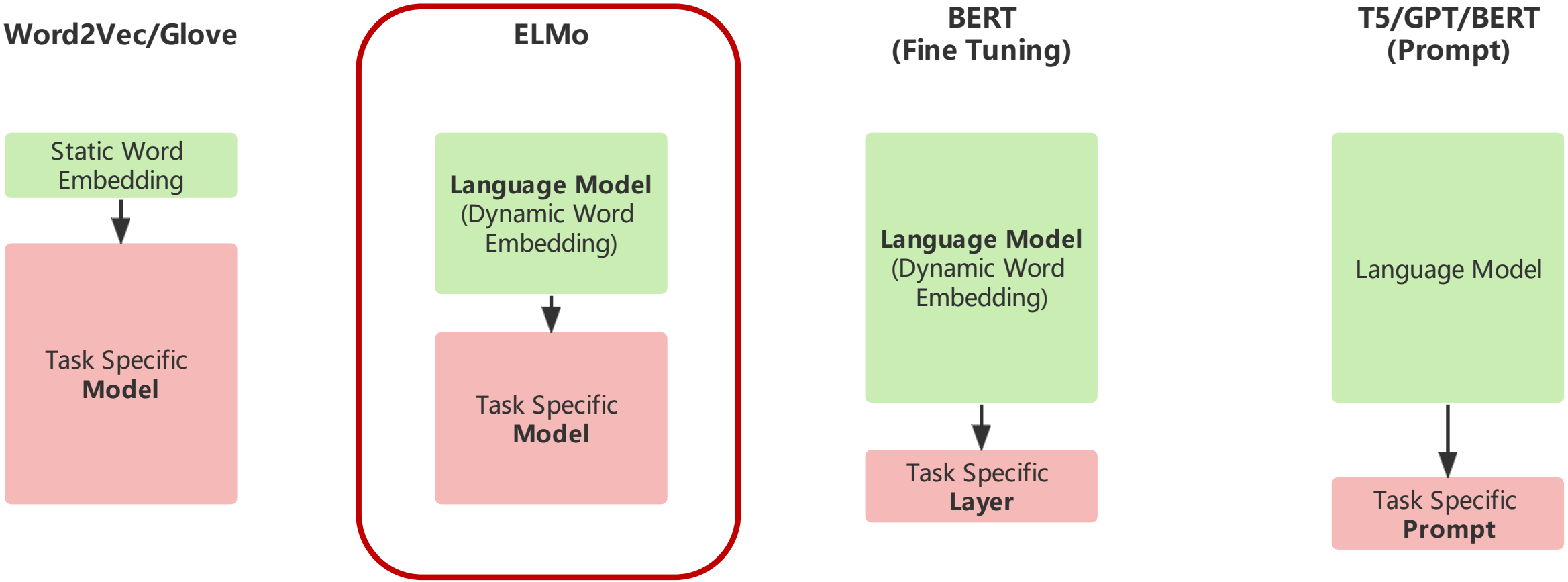
语言模型的演变经历了以下几个阶段：



1. word2vec/Glove将离散的文本数据转换为固定长度的**静态词向量**，后根据下游任务训练不同的**语言模型**

NLP中的预训练模型

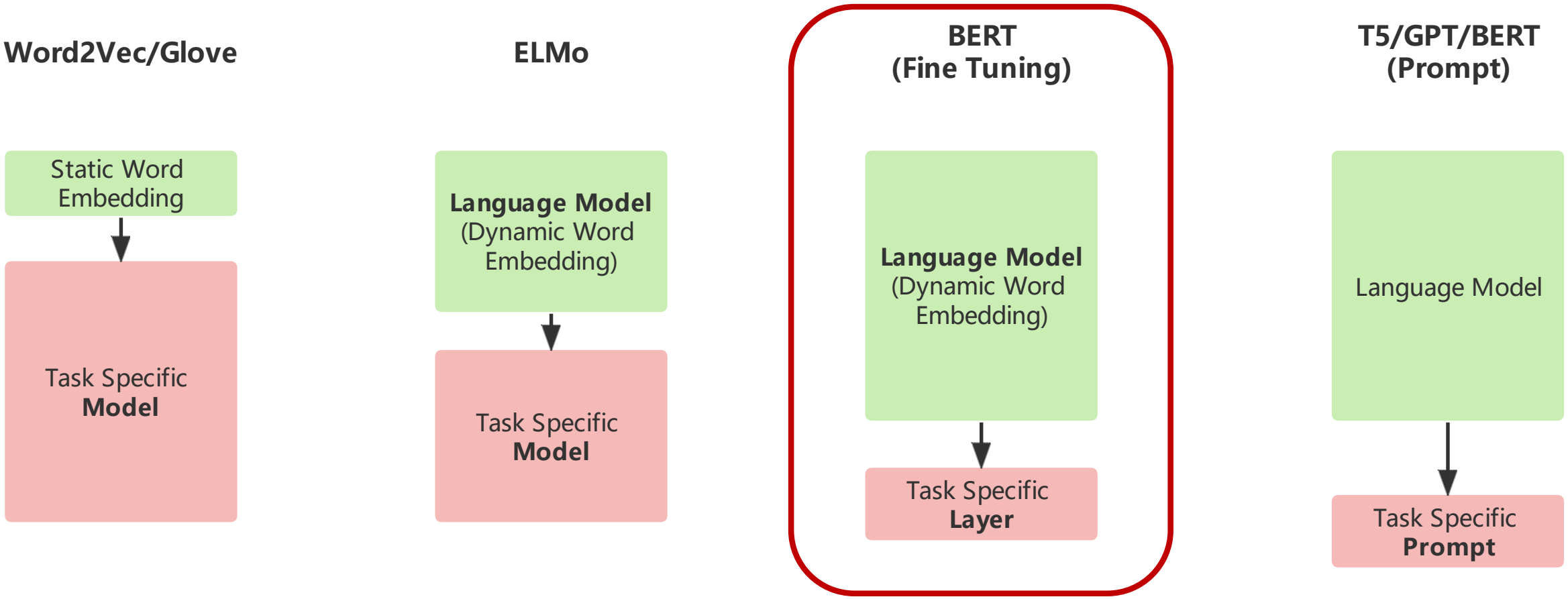
语言模型的演变经历了以下几个阶段：



2. ELMo预训练模型将文本数据**结合上下文信息**，转换为**动态词向量**，后根据下游任务训练不同的**语言模型**

NLP中的预训练模型

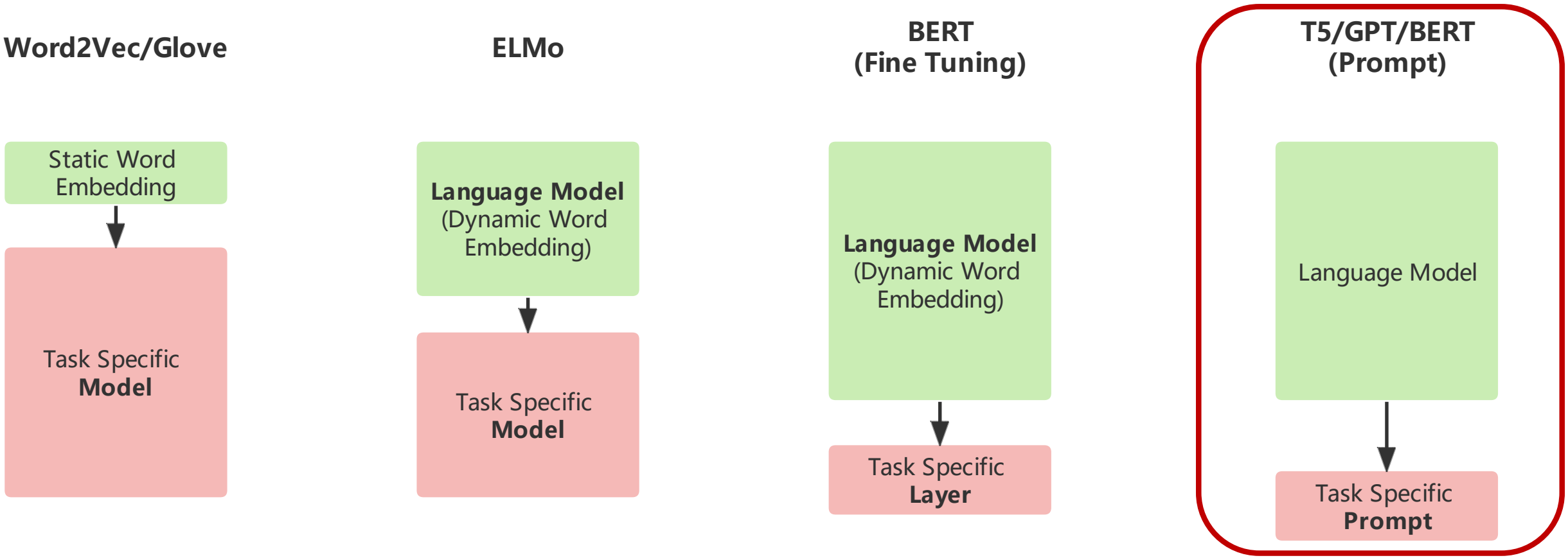
语言模型的演变经历了以下几个阶段：



3.BERT同样将文本数据转换为**动态词向量**，能够更好地捕捉**句子级别的信息与语境信息**，后续只需对BERT参数进行微调，仅重新训练最后的**输出层**即可适配下游任务

NLP中的预训练模型

语言模型的演变经历了以下几个阶段：

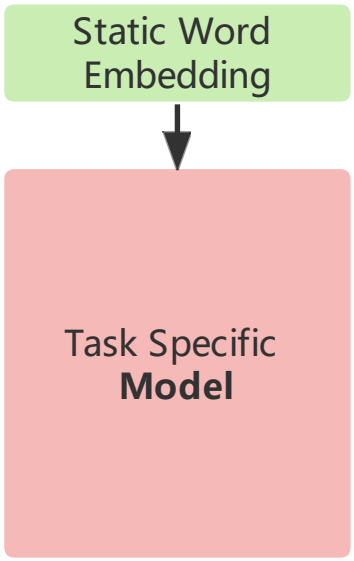


4.GPT等预训练语言模型主要用于**文本生成类任务**，需要通过**prompt方法**来应用于下游任务，指导模型生成特定的输出。

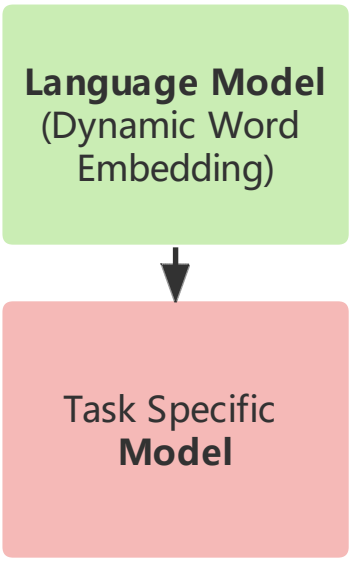
NLP中的预训练模型

语言模型的演变经历了以下几个阶段：

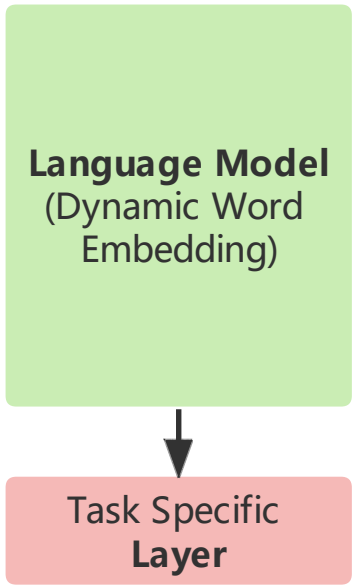
Word2Vec/Glove



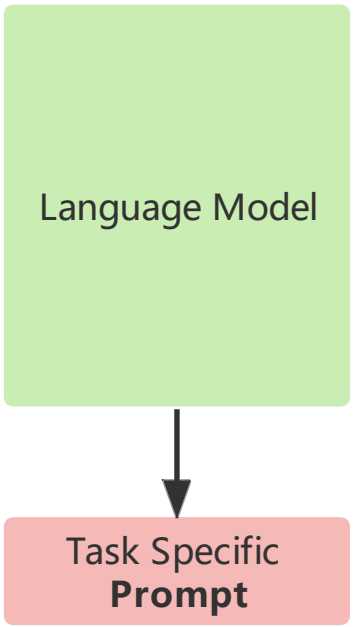
ELMo



BERT (Fine Tuning)



T5/GPT/BERT (Prompt)



BERT模型本质上是结合了ELMo模型与GPT模型的优势。

- 相比于ELMo，BERT仅需改动最后的输出层，而非模型架构，便可以在下游任务中达到很好的效果；
- 相比于GPT，BERT在处理词元表示时考虑到了双向上下文的信息；

目录

BERT Introduction

1 - Semi-supervised training on large amounts of text (books, wikipedia..etc).

The model is trained on a certain task that enables it to grasp patterns in language. By the end of the training process, BERT has language-processing abilities capable of empowering many models we later need to build and train in a supervised way.

Semi-supervised Learning Step

Model:

Dataset:

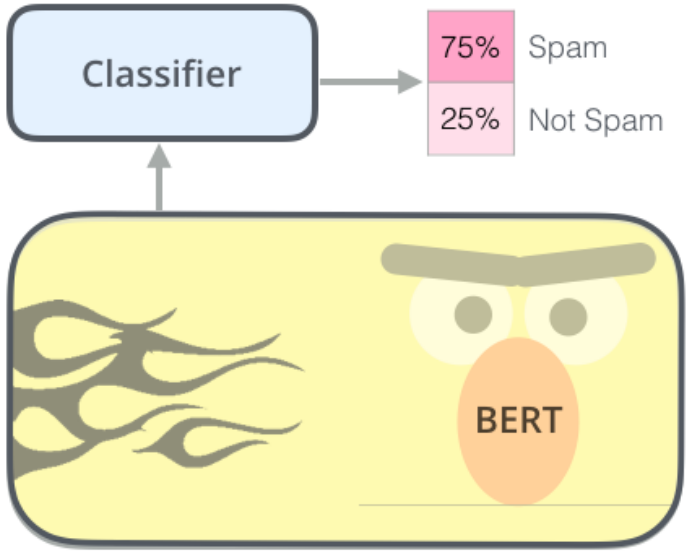
Objective:

Predict the masked word (langauge modeling)

2 - Supervised training on a specific task with a labeled dataset.

Supervised Learning Step

Model: (pre-trained in step #1)

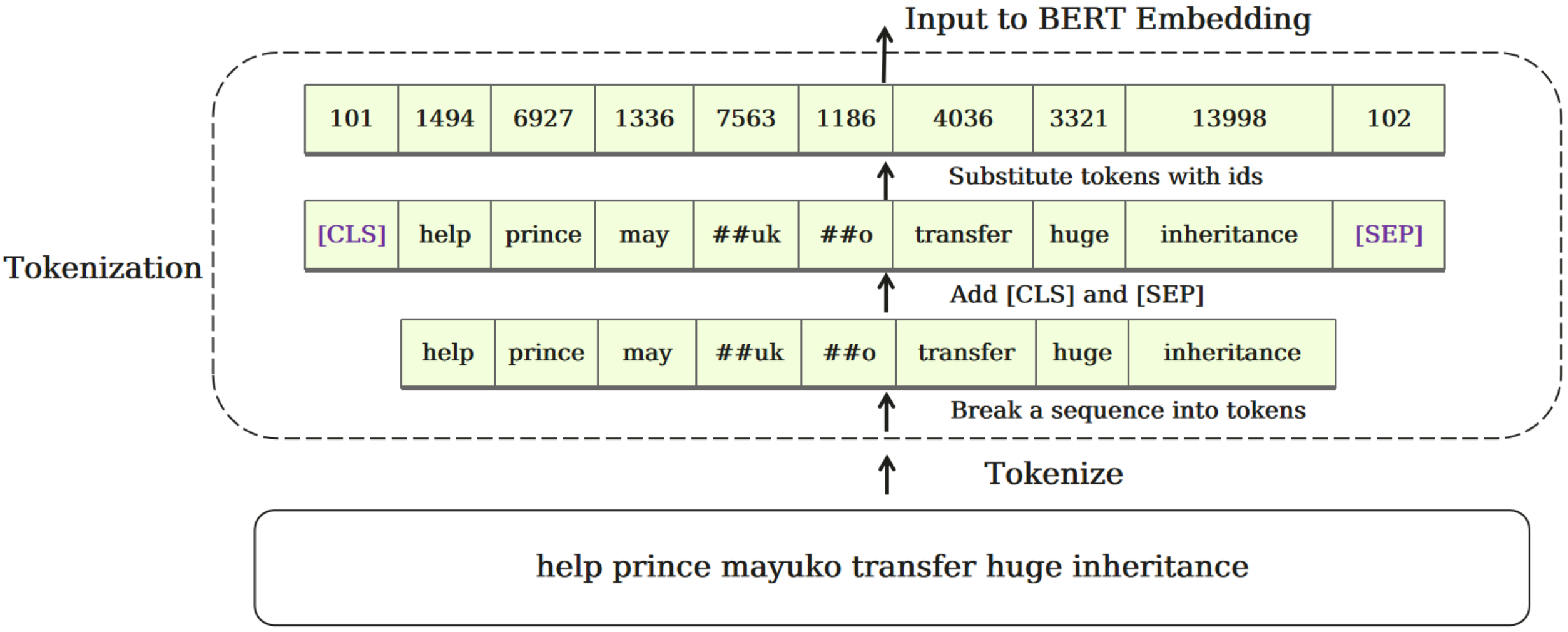


Dataset:

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

BERT Input

第一步：Tokenization， 输入的句子经过分词后，首尾添加[CLS]与[SEP]特殊字符，后转换为数字id



BERT Input

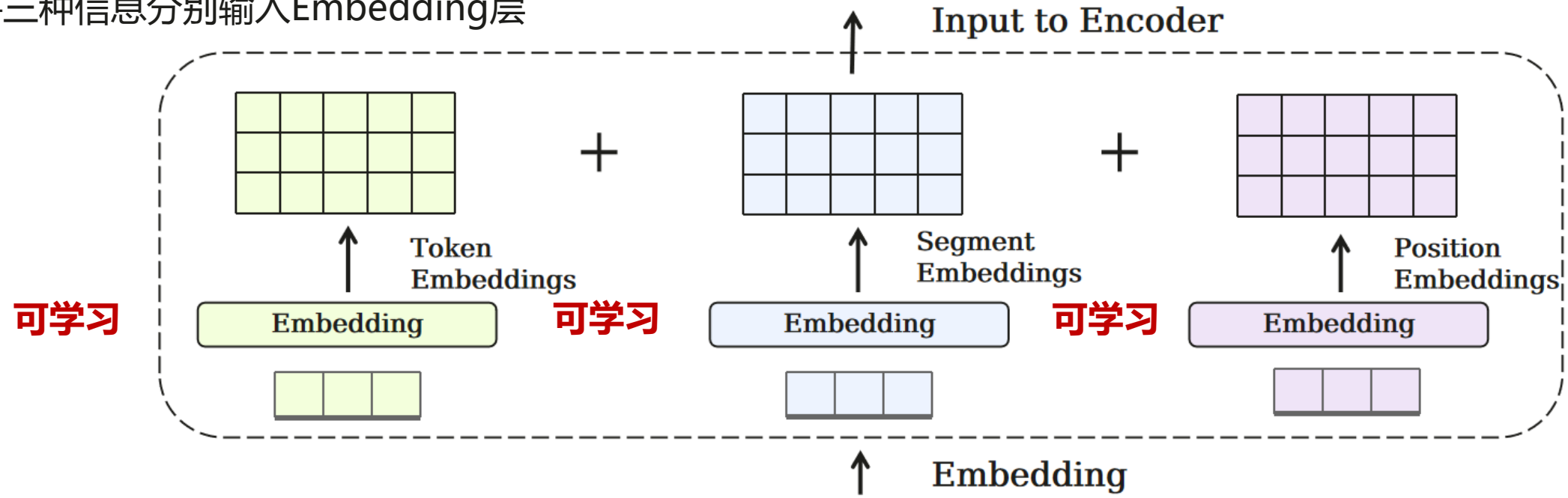
第二步：Embedding， 输入到BERT模型的信息由三部分内容组成：

- 表示内容的token ids
- 表示位置的position ids
- 用于区分不同句子的token type ids

token type ids	0	0	0	0	0	0	0	0	0	0
position ids	0	1	2	3	4	5	6	7	8	9
token ids	101	1494	6927	1336	7563	1186	4036	3321	13998	102
model input	[CLS]	help	prince	may	##uk	##o	transfer	huge	inheritance	[SEP]

BERT Input

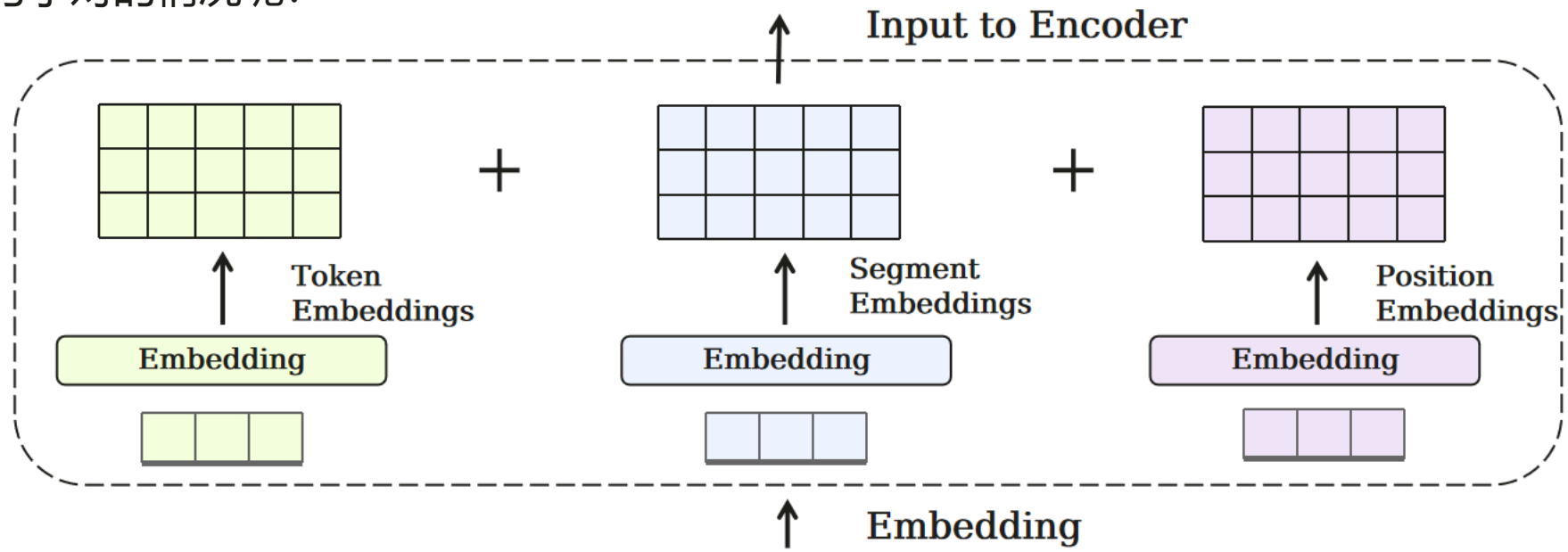
将三种信息分别输入Embedding层



token type ids	0	0	0	0	0	0	0	0	0	0
position ids	0	1	2	3	4	5	6	7	8	9
token ids	101	1494	6927	1336	7563	1186	4036	3321	13998	102
model input	[CLS]	help	prince	may	##uk	##o	transfer	huge	inheritance	[SEP]

BERT Input

如果出现输入是句子对的情况呢?

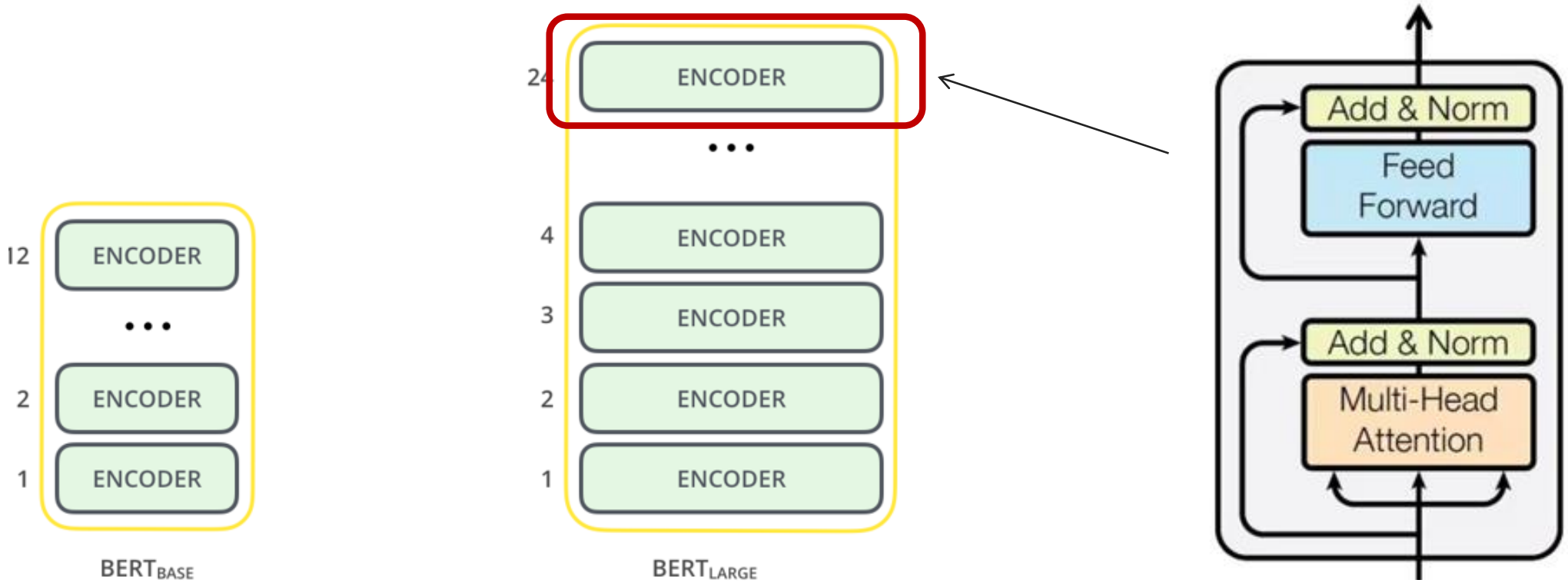


token type ids	0	0	0	0	0	0	1	1	1	1	1
position ids	0	1	2	3	4	5	6	7	8	9	10
token ids	102	1139	3676	1110	10509	102	2002	7777	2377	1158	102
model input	[CLS]	my	dog	is	cute	[SEP]	he	likes	play	##ing	[SEP]

BERT Architecture

BERT由Encoder Layer堆叠而成，Encoder Layer的组成与Transformer的Encoder Layer一致：

- 自注意力层 + 前馈神经网络，中间通过residual connection和LayerNorm连接



BERT Architecture

BERT（Bidirectional Encoder Representation from Transformers）是由Transformer的Encoder层堆叠而成，

BERT的模型大小有如下两种：

- BERT BASE：与Transformer参数量齐平，用于比较模型效果（110M parameters）
- BERT LARGE：在BERT BASE基础上扩大参数量，达到了当时各任务最好的结果（340M parameters）

model	blocks	hidden size	attention head
Transformer	6	512	8
BERT BASE	12	768	12
BERT LARGE	24	1024	16

Reference: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

BERT Architecture

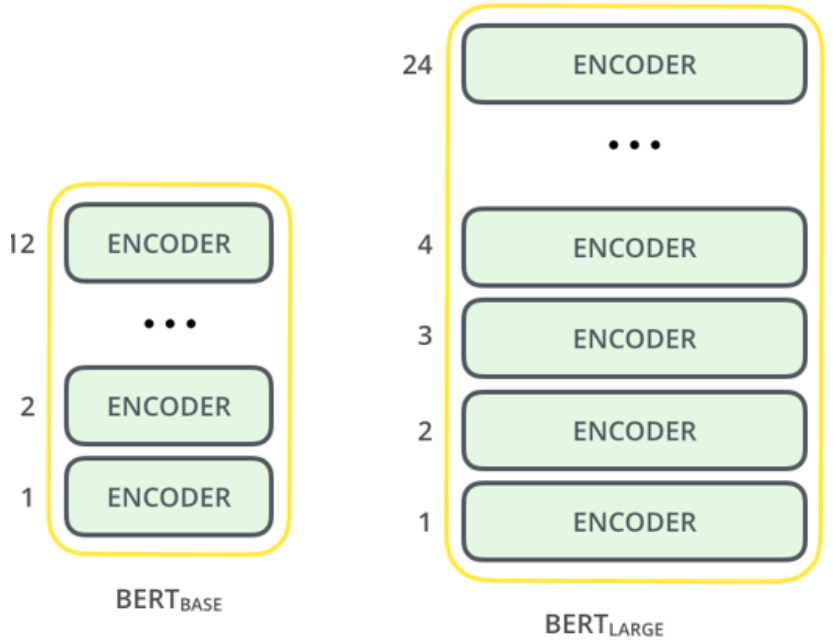
BERT（Bidirectional Encoder Representation from Transformers）是由Transformer的Encoder层堆叠而成，

BERT的模型大小有如下两种：

- BERT BASE：与Transformer参数量齐平，用于比较模型效果（110M parameters）
- BERT LARGE：在BERT BASE基础上扩大参数量，达到了当时各任务最好的结果（340M parameters）

model	blocks	hidden size	attention head
Transformer	6	512	8
BERT BASE	12	768	12
BERT LARGE	24	1024	16

堆叠的encoder层数



Reference: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

BERT Architecture

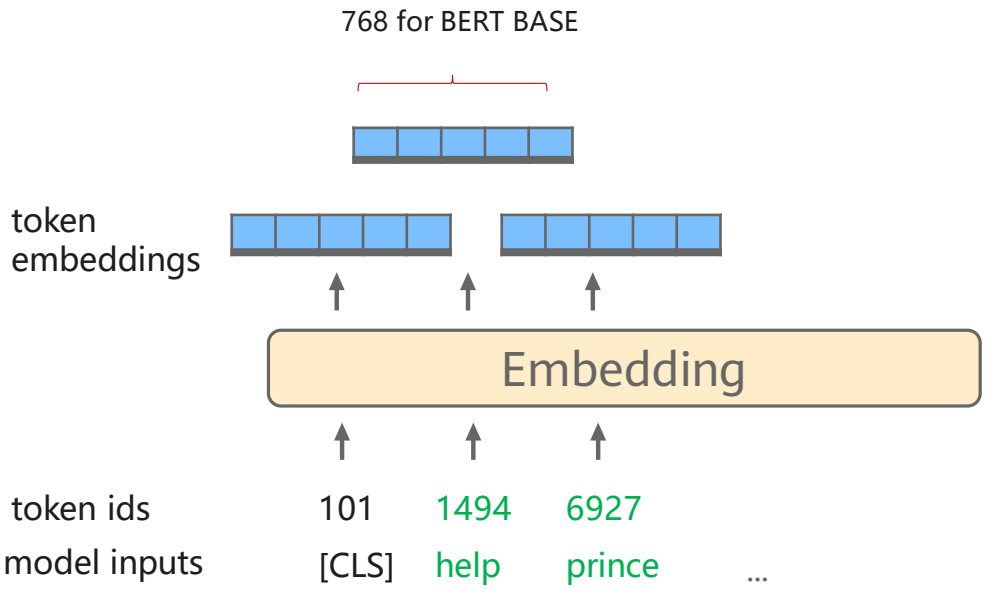
BERT（Bidirectional Encoder Representation from Transformers）是由Transformer的Encoder层堆叠而成，

BERT的模型大小有如下两种：

- BERT BASE：与Transformer参数量齐平，用于比较模型效果（110M parameters）
- BERT LARGE：在BERT BASE基础上扩大参数量，达到了当时各任务最好的结果（340M parameters）

model	blocks	hidden size	attention head
Transformer	6	512	8
BERT BASE	12	768	12
BERT LARGE	24	1024	16

token embedding的词向量长度



Reference: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

BERT Architecture

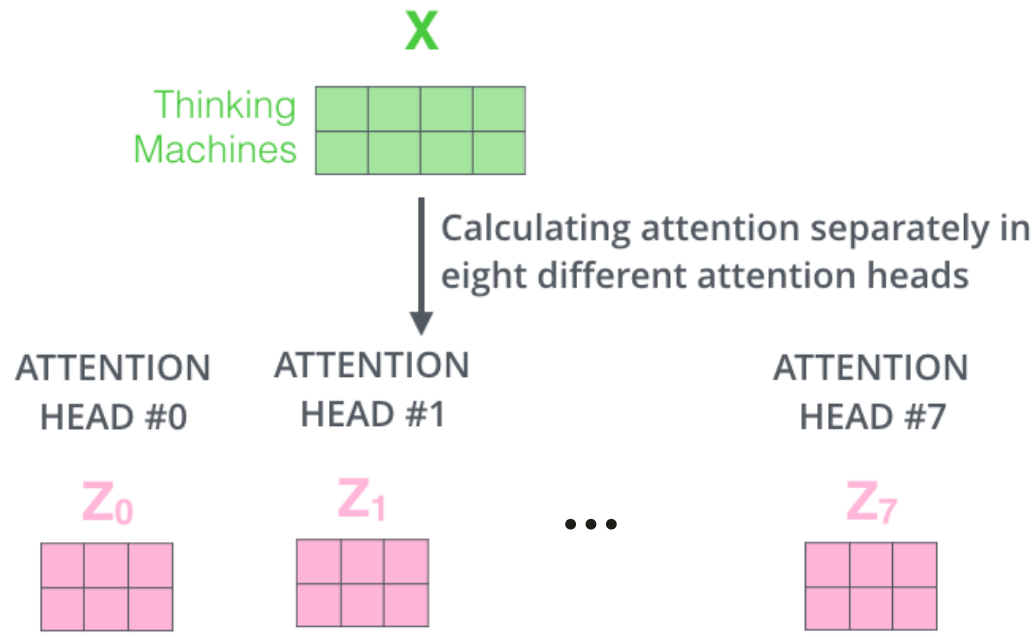
BERT（Bidirectional Encoder Representation from Transformers）是由Transformer的Encoder层堆叠而成，

BERT的模型大小有如下两种：

- BERT BASE：与Transformer参数量齐平，用于比较模型效果（110M parameters）
- BERT LARGE：在BERT BASE基础上扩大参数量，达到了当时各任务最好的结果（340M parameters）

model	blocks	hidden size	attention head
Transformer	6	512	8
BERT BASE	12	768	12
BERT LARGE	24	1024	16

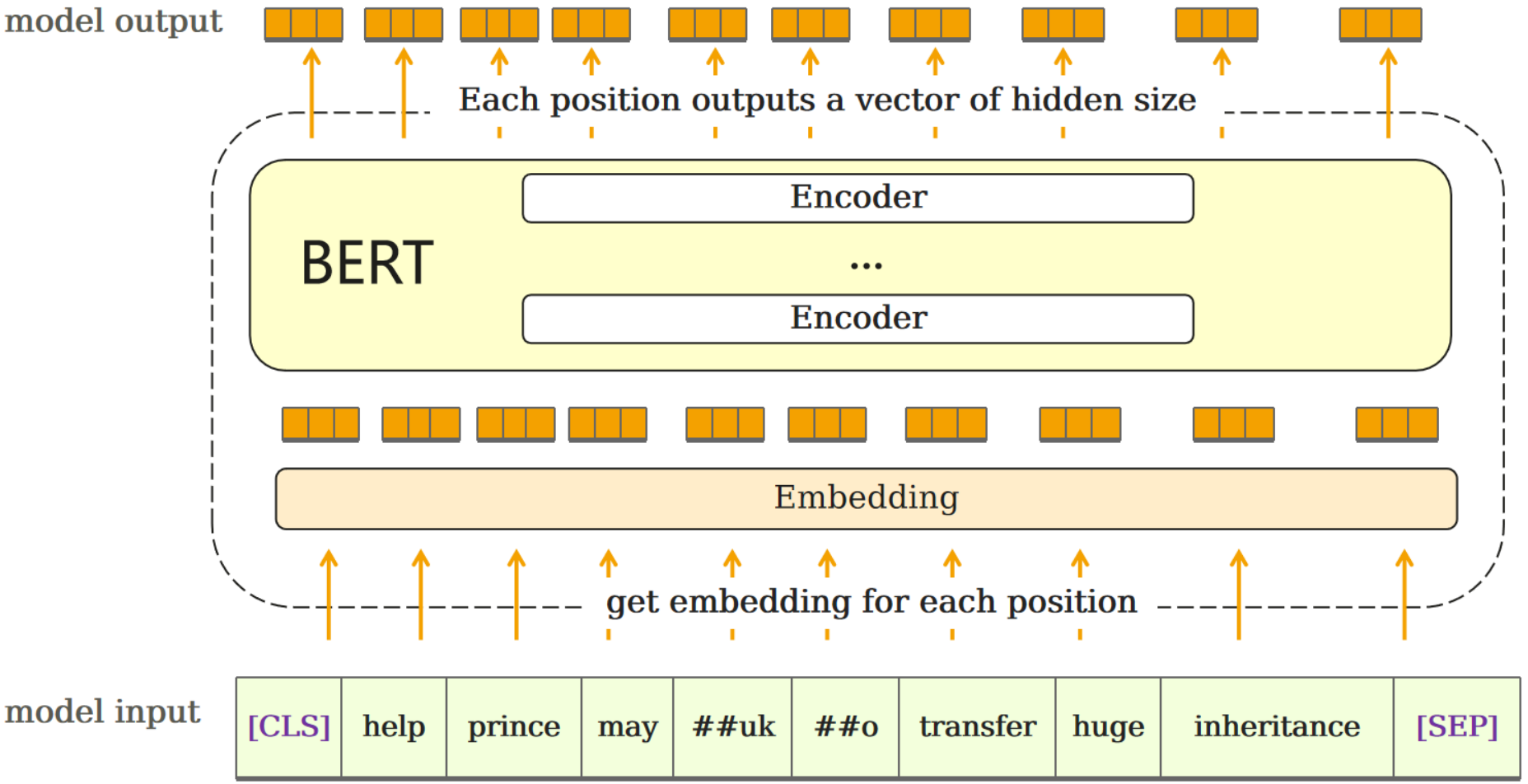
多头注意力头的数量



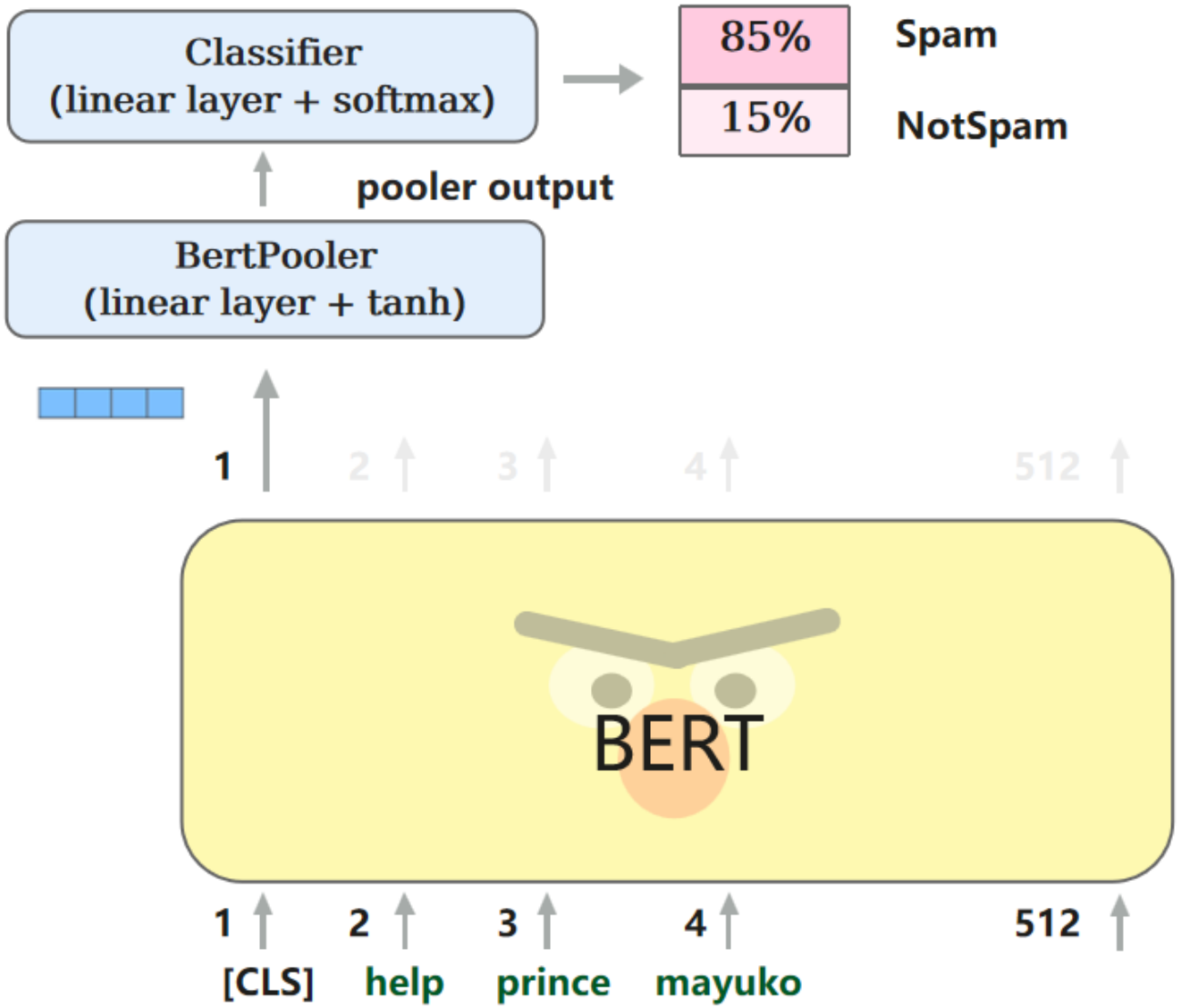
Reference: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

BERT Output

BERT会针对每一个位置输出大小为hidden size的向量，在下游任务中，会根据任务内容的不同，选取不同的向量放入输出层



BERT Output



例如，在诈骗邮件分类任务中，我们会将表示句子级别信息的`[CLS]` token所对应的向量，经过Bert Pooler放入classfier中，得到对spam/not spam分类的预测。

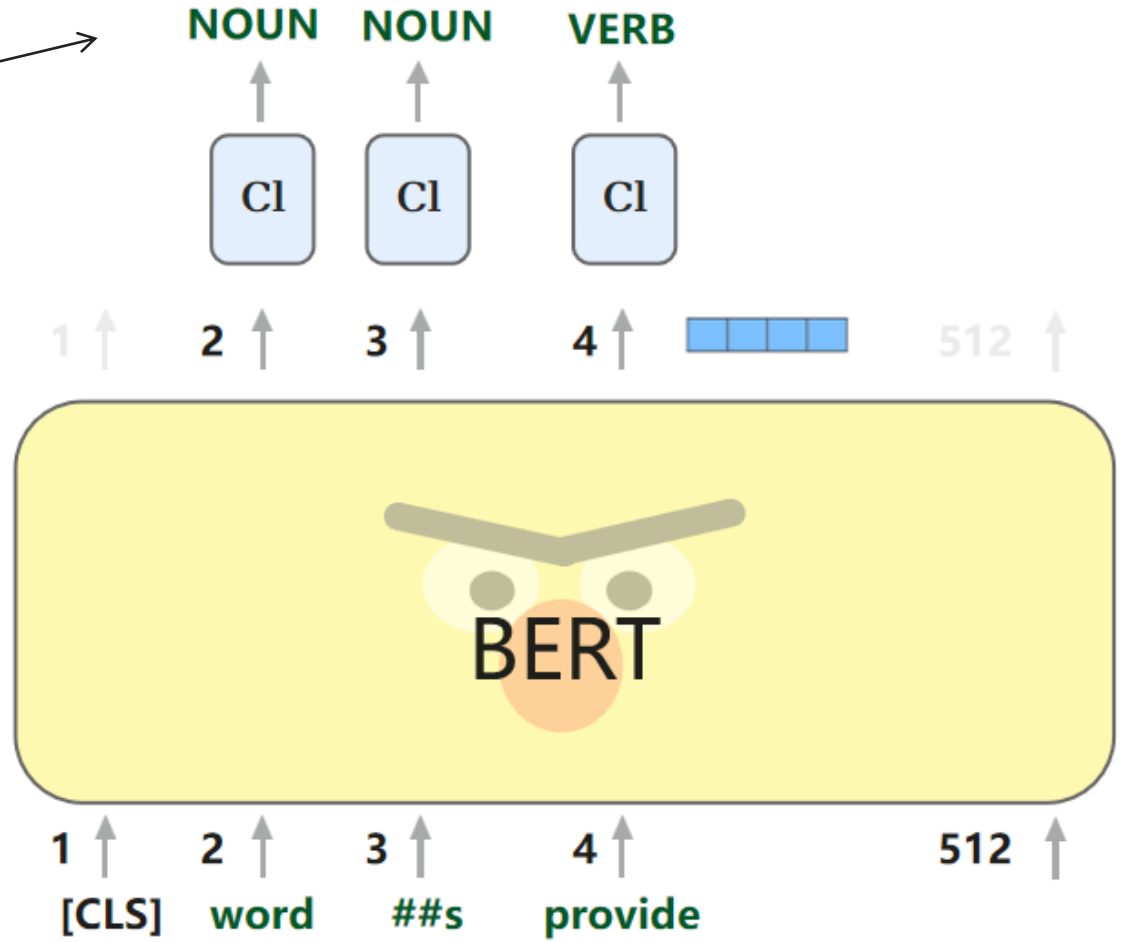
我们一般称[CLS]经过线性层+激活函数tanh的输出为**pooler output**，用于句子级别的分类/回归任务

Reference: The Illustrated BERT, ELMo, and co. (How NLP Cracked Transfer Learning)

BERT Output

例如，在词性标注任务（POS Tagging）中，我们需要获得每一个token所对应的类别，因此需要将[CLS]和[SEP]中有实际意义的token输出，分别输入对应的classifier中。

NOUN	60%
VERB	20%
ADJ	10%
…	…

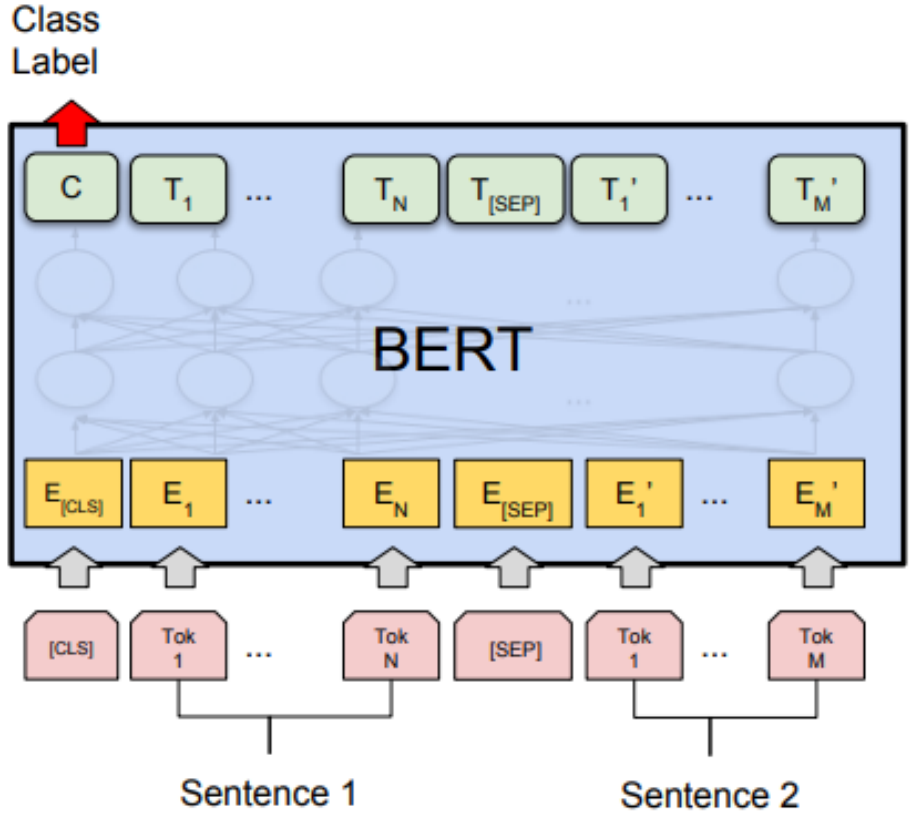


我们一般称BERT输出的每个位置对应的vector为 **sequence output**

Reference: The Illustrated BERT, ELMo, and co. (How NLP Cracked Transfer Learning)

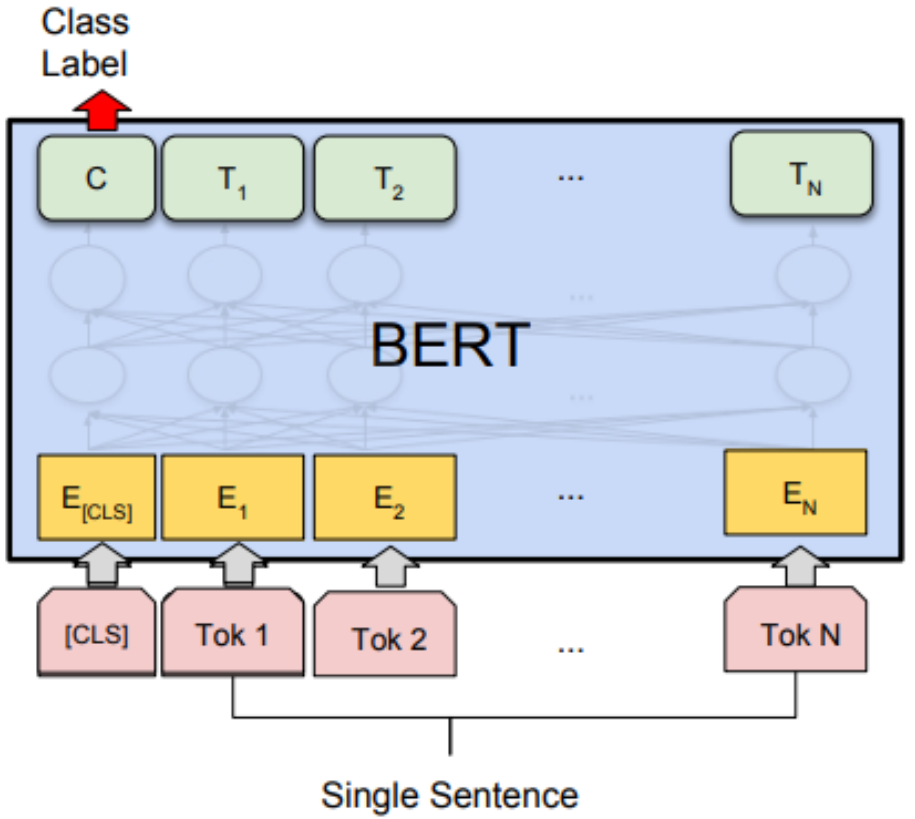
BERT的不同下游任务

句子对分类：[CLS]



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

单句子分类：[CLS]



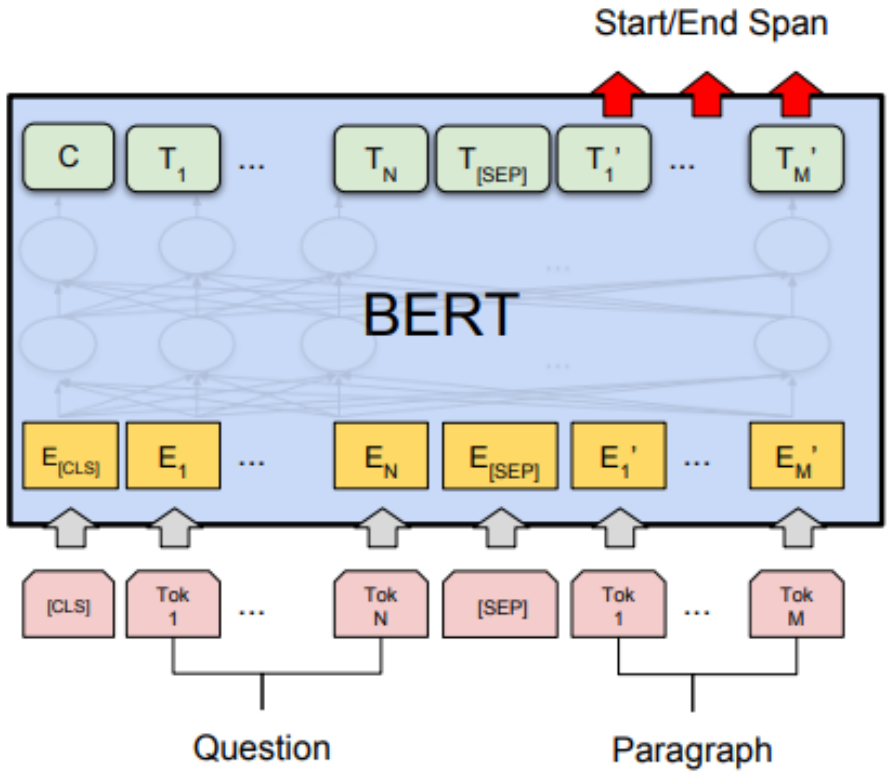
(b) Single Sentence Classification Tasks: SST-2, CoLA

Reference: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

BERT的不同下游任务

Q&A：第二个序列

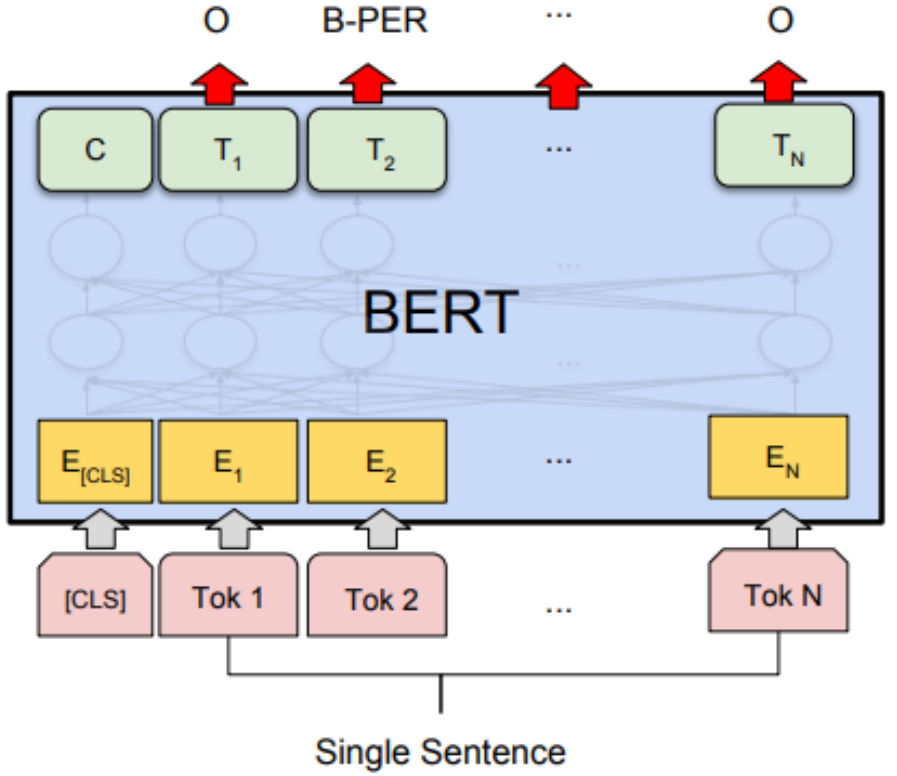
找出答案的起始与结束token



(c) Question Answering Tasks:
SQuAD v1.1

单句子Tagging：

有实际意义的token



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

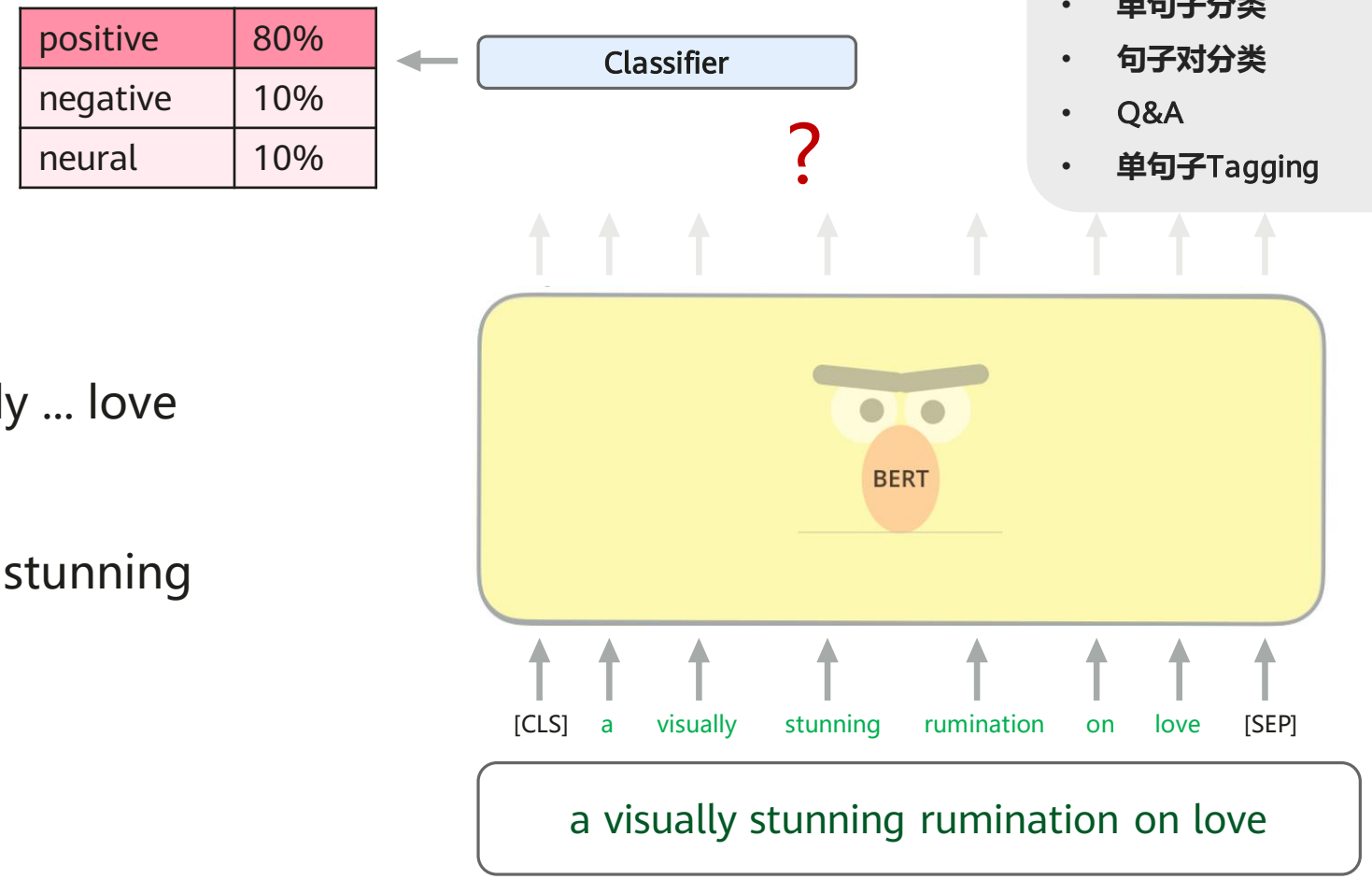
Reference: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Quiz

3. 如果我希望基于BERT完成情感分类任务，即输入一句话，输出语句的情感色彩（positive、negative or neural），我应该选择哪一部分的输出放入classifier中?

Hint: 该任务属于什么类型的任务?

- 单句子分类
- 句子对分类
- Q&A
- 单句子Tagging



A. [CLS]

B. a visually ... love

C. [SEP]

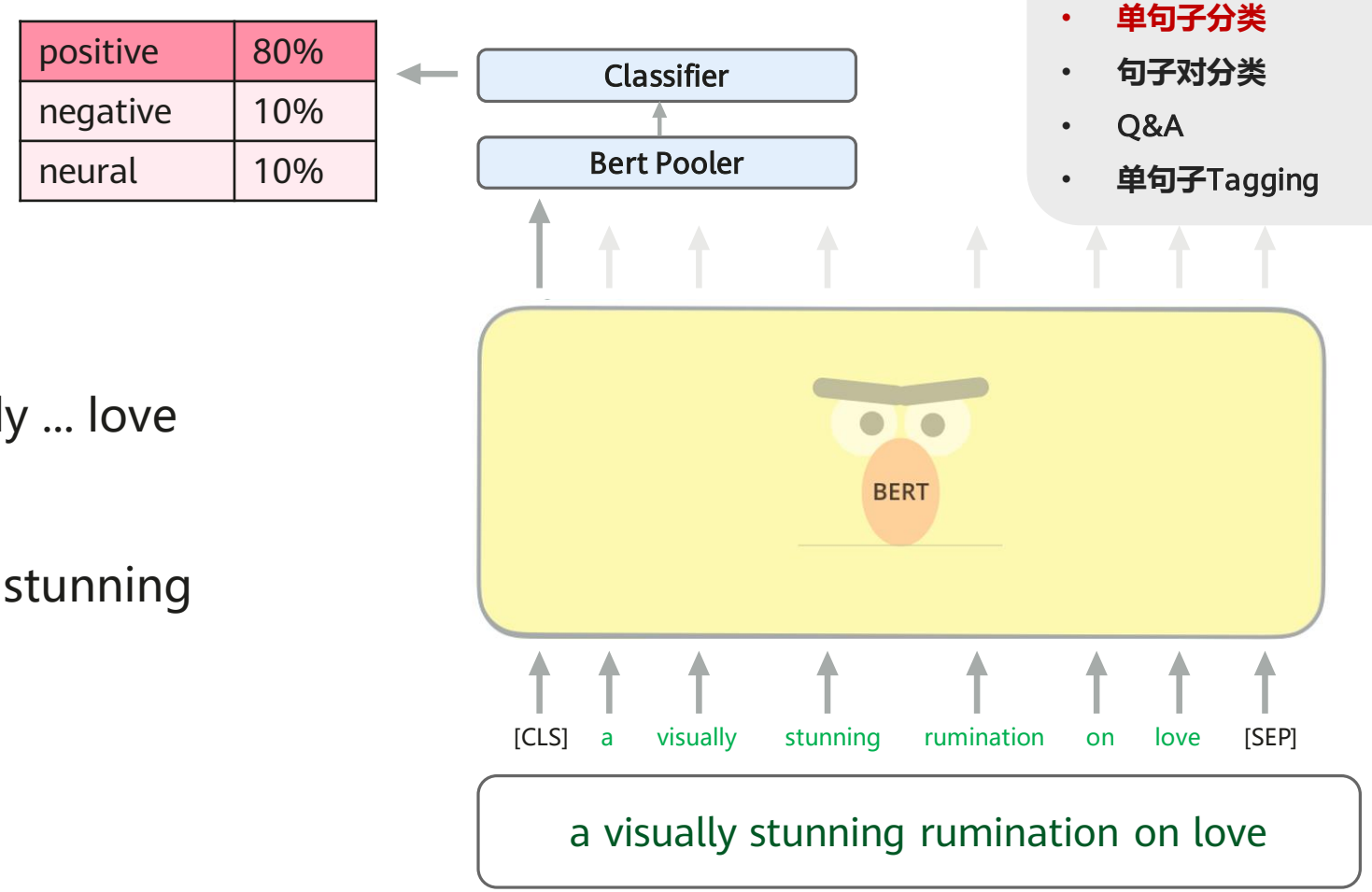
D. visually stunning

Quiz

3. 如果我希望基于BERT完成情感分类任务，即输入一句话，输出语句的情感色彩（positive、negative or neural），我应该选择哪一部分的输出放入classifier中?

Hint: 该任务属于什么类型的任务?

- **单句子分类**
- **句子对分类**
- Q&A
- **单句子Tagging**



A. [CLS]

B. a visually ... love

C. [SEP]

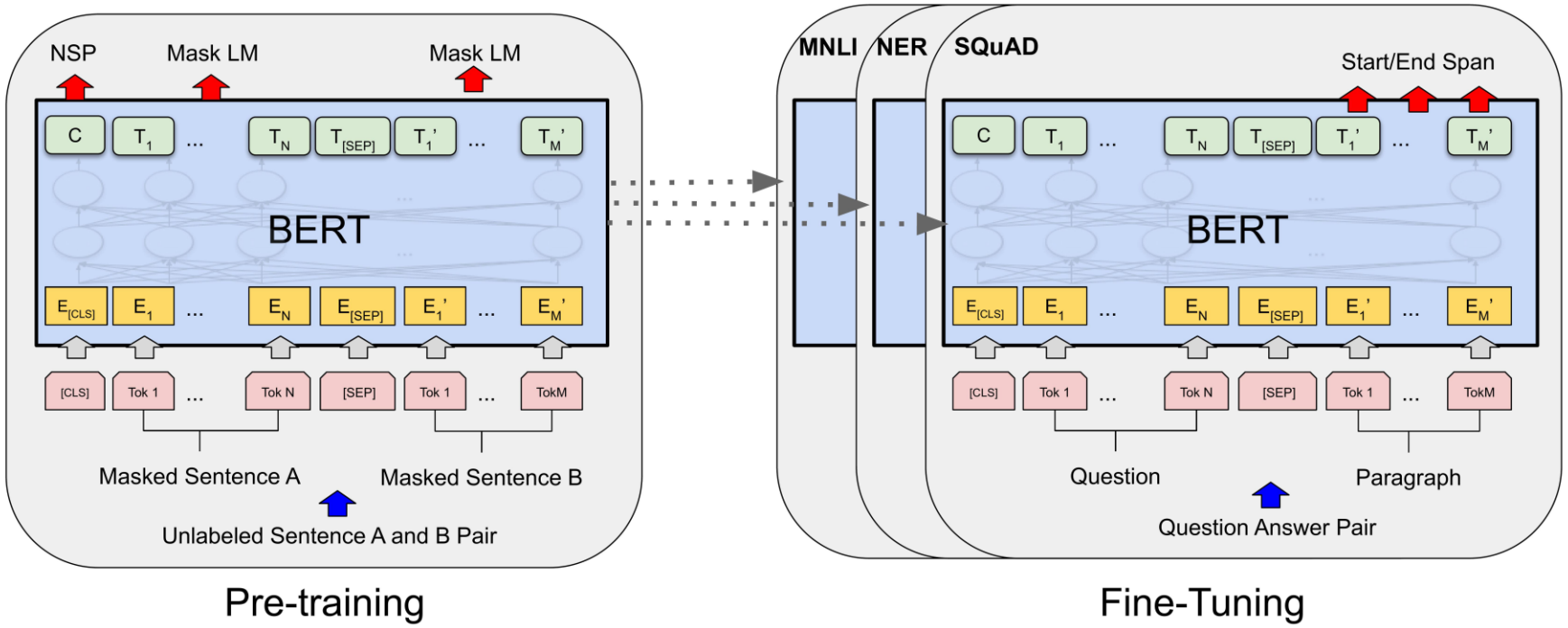
D. visually stunning

目录

BERT预训练

BERT预训练任务有两种：Masked Language Modelling（MLM） 和 Next Sentence Prediction（NSP）。

- MLM：随机遮盖输入句子中的一些词语，并预测被遮盖的词语是什么（完形填空）
- NSP：预测两个句子是不是上下文的关系



Reference: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Masked Language Model（MLM）

Masked Language Modelling（MLM） 捕捉词语级别的信息

1. 在输入中随机遮盖15%的token

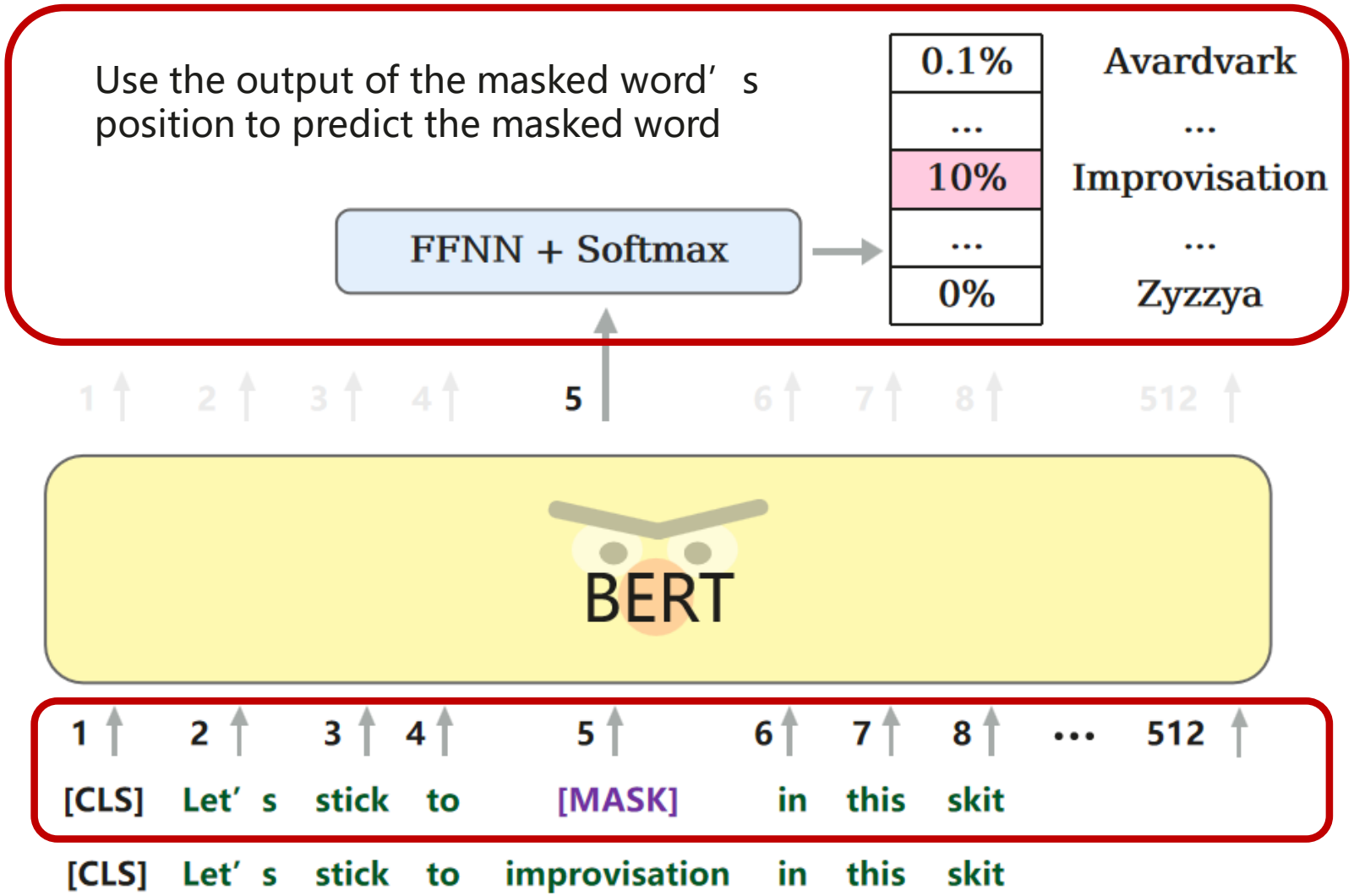
   （即将token替换为[MASK]）

2. 将[MASK]位置对应的BERT输出

   放入输出层中，预测被遮盖的

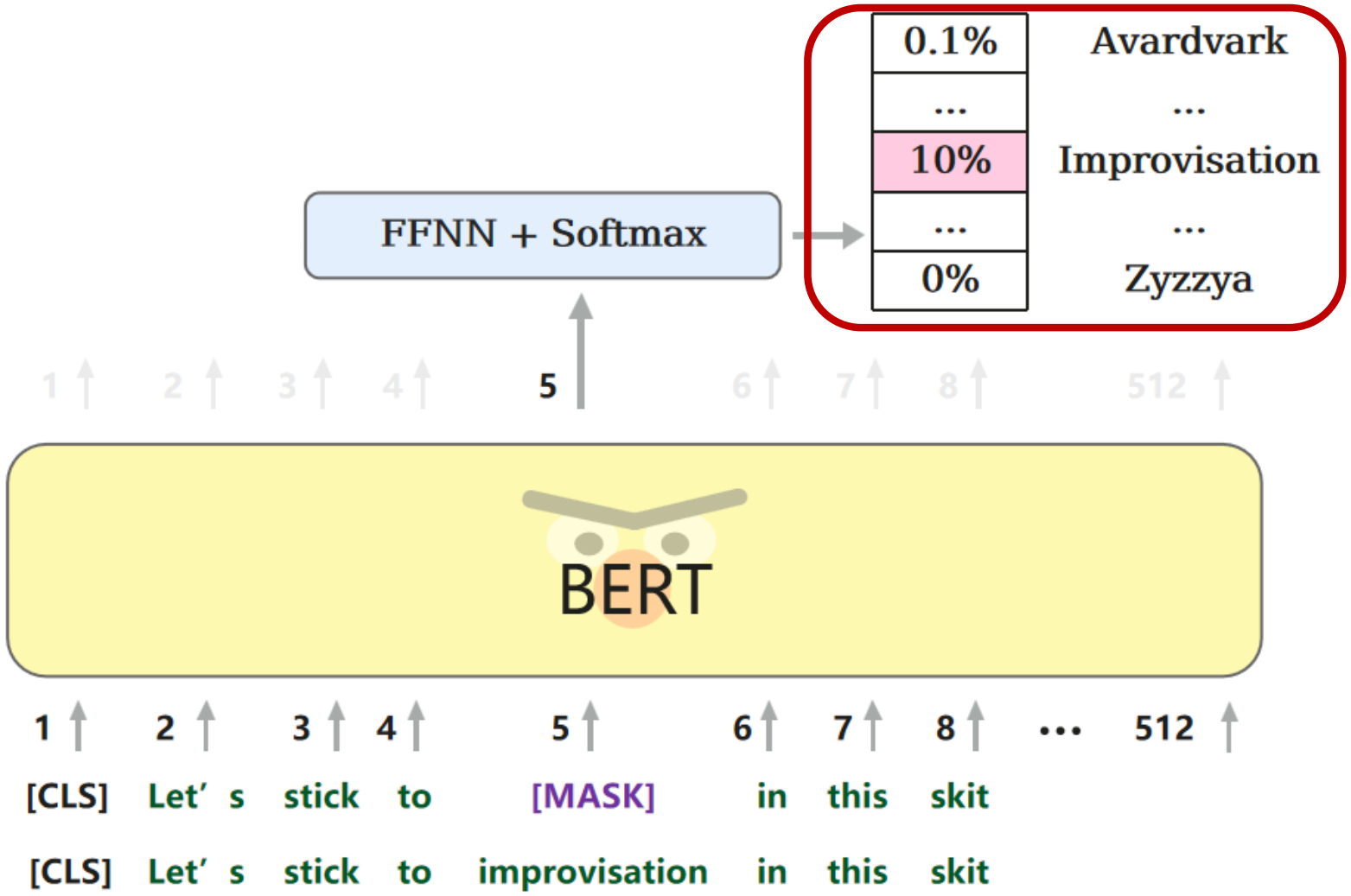
   token

问题一：如何进行预测?



Masked Language Model（MLM）

在将[MASK]位置所对应的BERT输出放入输出层后，本质上是在进行一个多分类任务



1. 我们事先会建立一个词表，覆盖所有英语单词

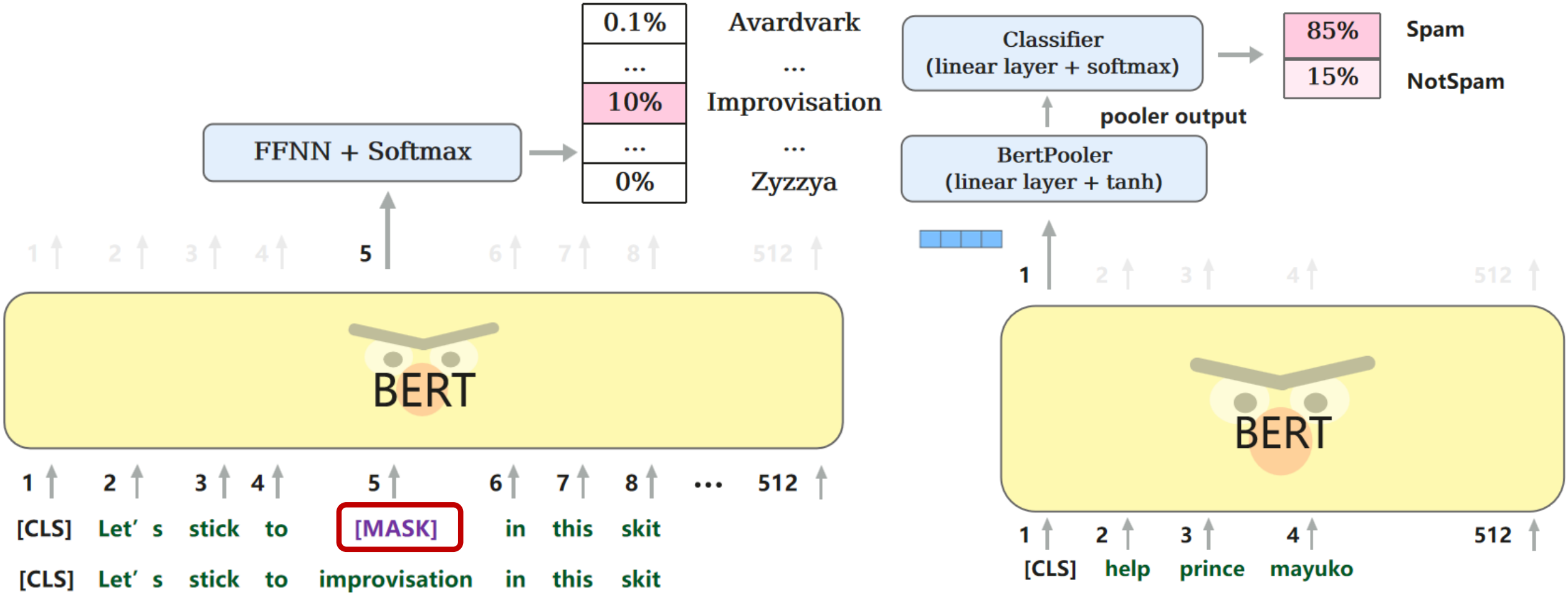
aardvark
aarhus
...
improvisation
...
zyzzyva

2. classifier计算每个词对应被遮盖的token的概率，并选取概率最高的词语作为预测结果

aardvark	0.1%
aarhus	0.1%
...	...
improvisation	10%
...	...
zyzzyva	0%

Masked Language Model（MLM）

在将[MASK]位置所对应的BERT输出放入输出层后，本质上是在进行一个多分类任务

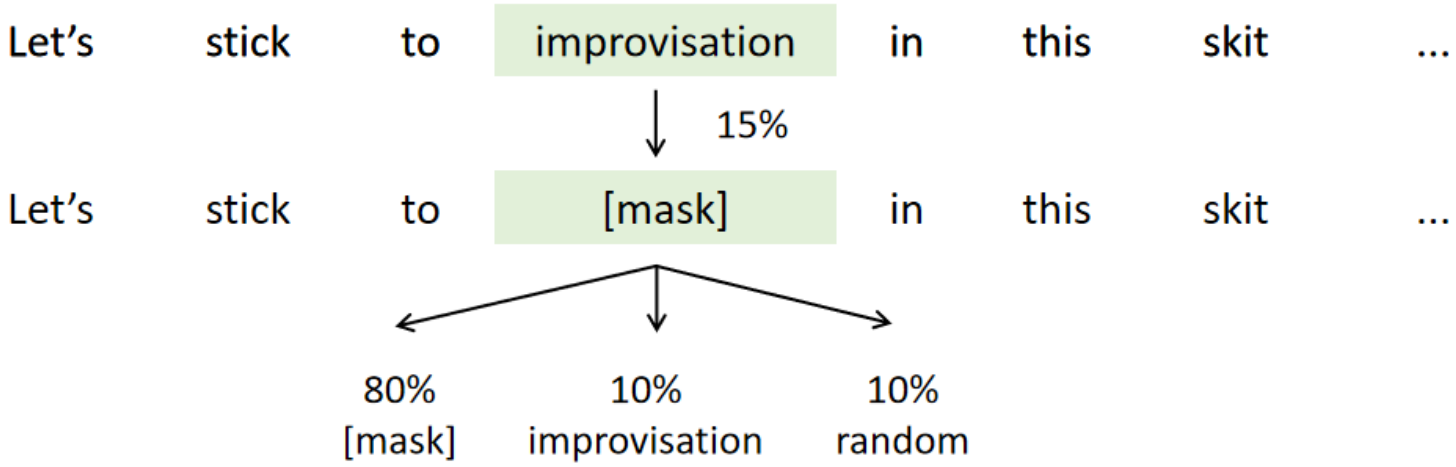


问题二：BERT进行推理时，输入并没有[MASK]

Masked Language Model（MLM）

为了使得预训练任务和推理任务尽可能接近，BERT在随机遮盖的15%的tokens中又进行了进一步的处理：

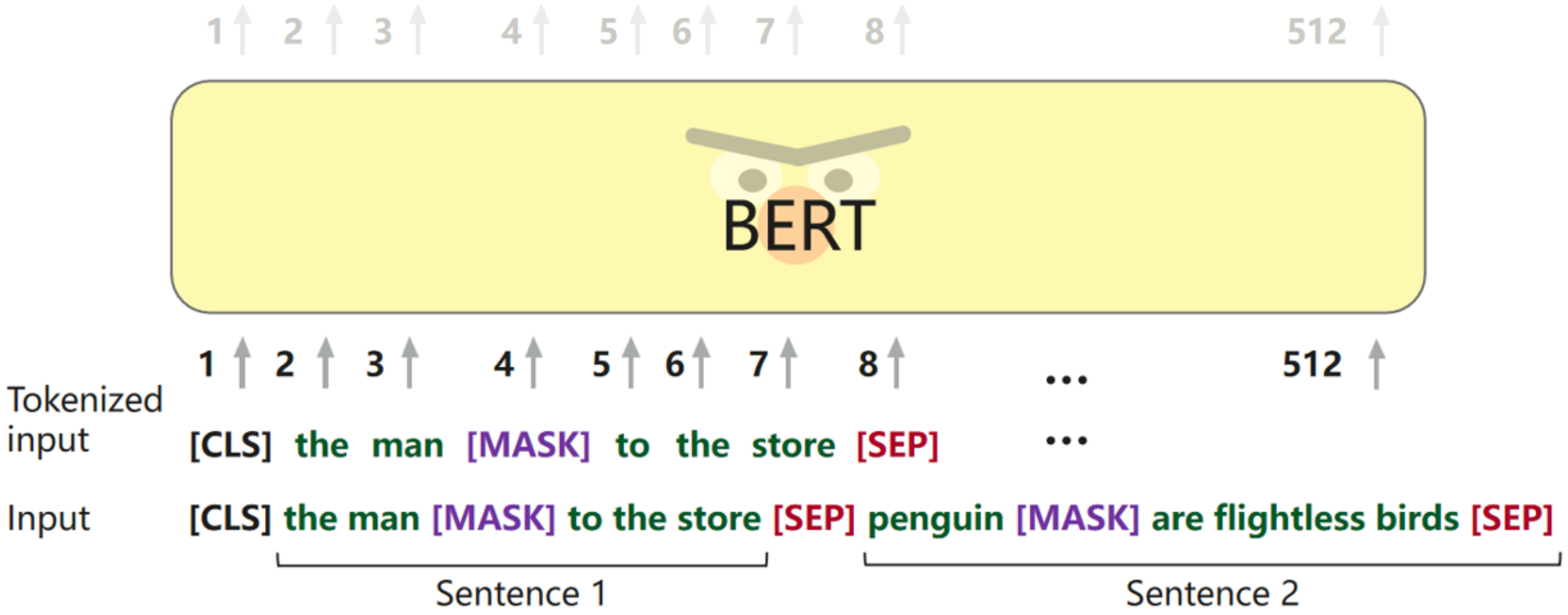
- 80%的概率替换为[MASK]
- 10%的概率替换为文本中的随机词
- 10%的概率不进行替换，保持原有的词元



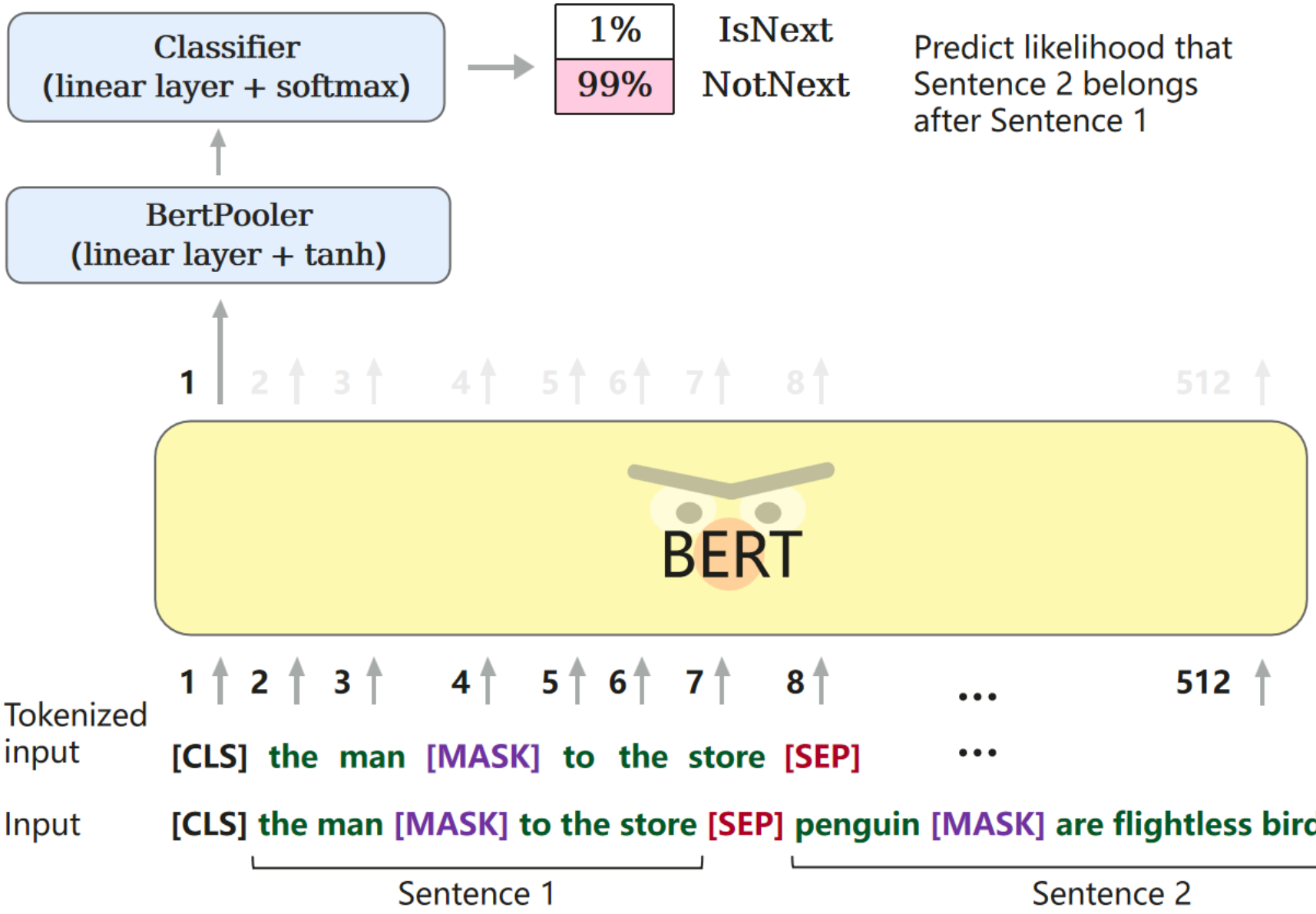
Next Sentence Prediction（NSP）

Next Sentence Prediction (NSP) 捕捉句子级别信息，简单来说是一个针对句子对的分类问题，判断一组句子中，句子B是否为句子A的下一句（IsNext or NotNext）

句子级别的分类任务，应该选取哪个位置的输出放入classifier?



Next Sentence Prediction（NSP）



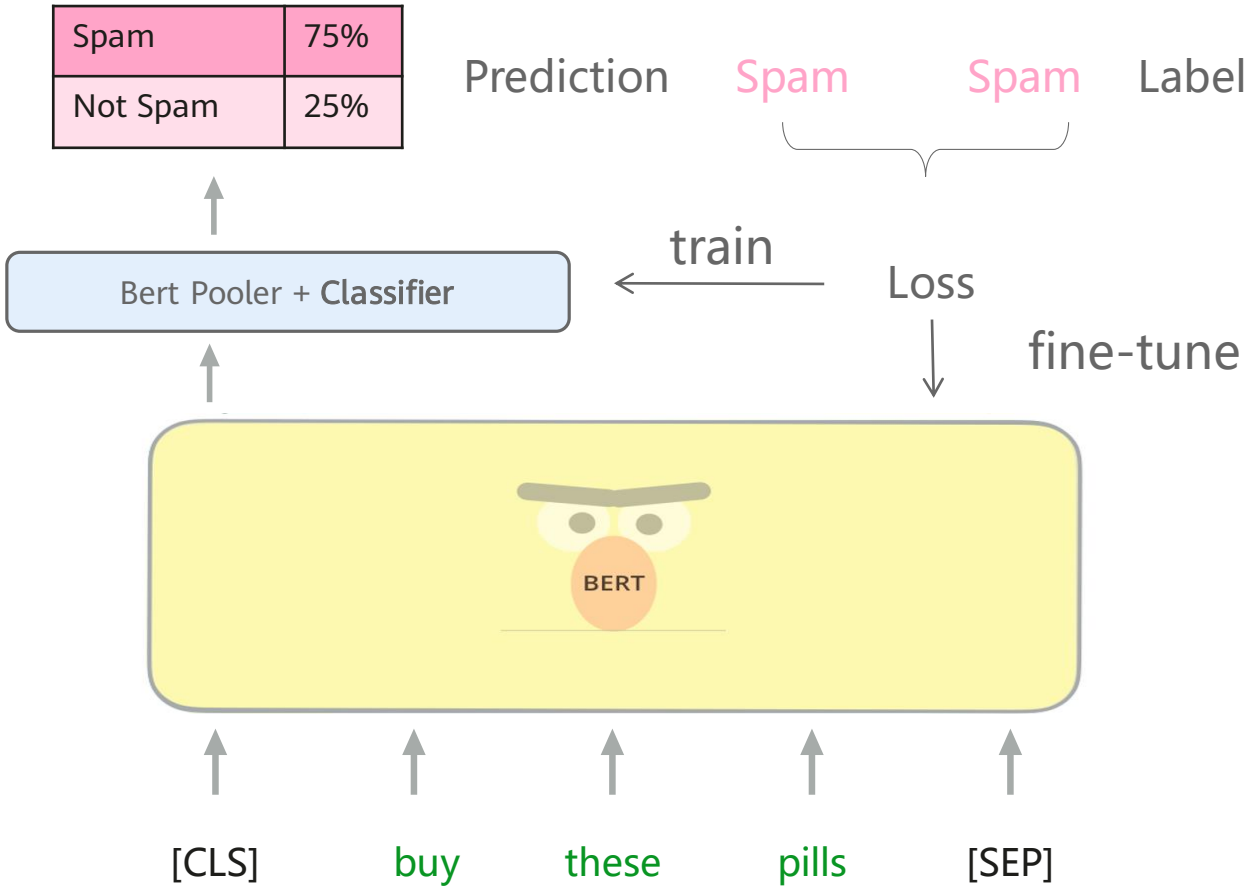
Next Sentence Prediction (NSP) 捕捉句子级别信息

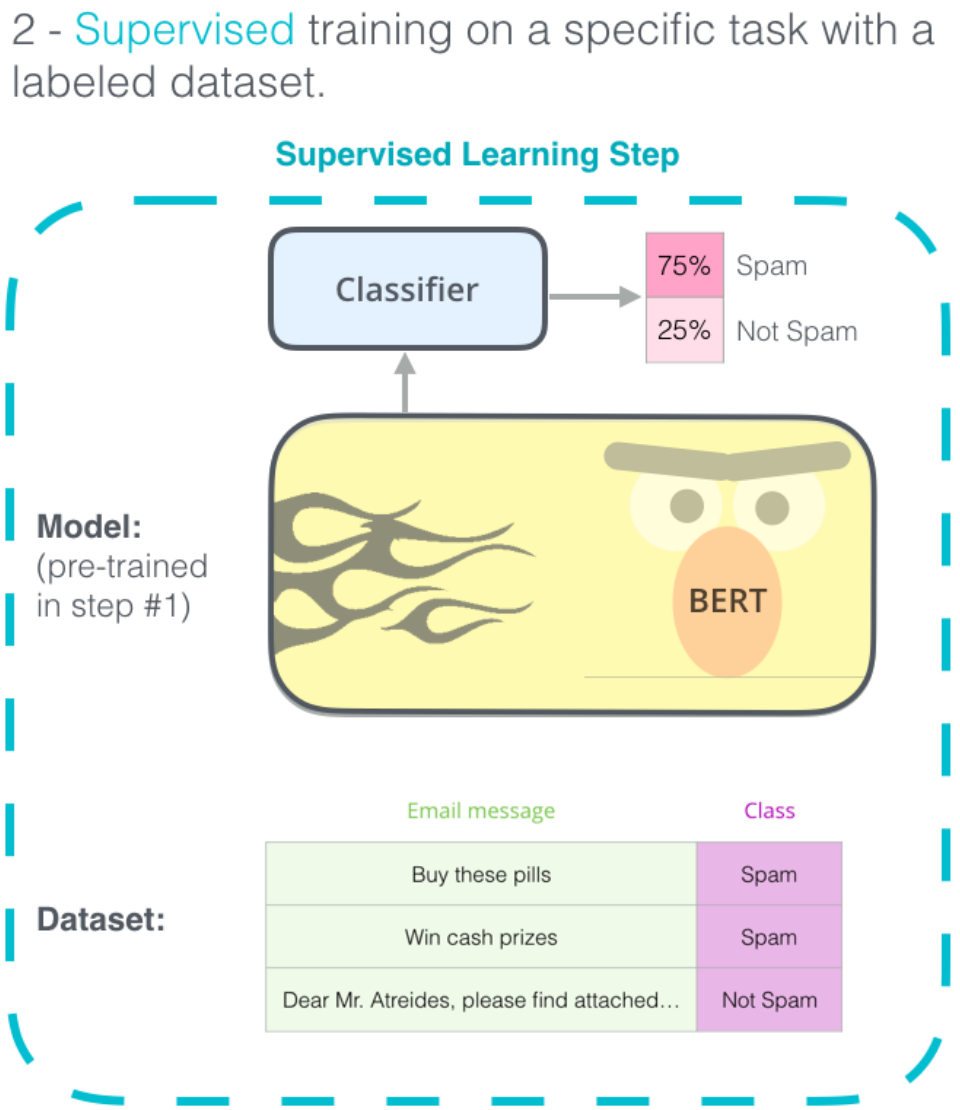
- 针对句子对的二分类问题
- 判断一组句子中，句子B是否为句子A的下一句（IsNext or NotNext）

目录

BERT Fine-tuning

在下游任务中，我们使用少量的标注数据（labelled data）对预训练Transformer编码器的所有参数进行微调，额外的输出层将从头开始训练。





Reference: The Illustrated BERT, ELMo, and co. (How NLP Cracked Transfer Learning)

Thank you.

把数字世界带入每个人、每个家庭、
每个组织，构建万物互联的智能世界。

Bring digital to every person, home, and organization for a fully connected, intelligent world.

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.

HUAWEI