深度之眼
deepshare.net

本期论文主题:Bert

导师: Yamada

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

预训练的深度双向transformer用于语义理解

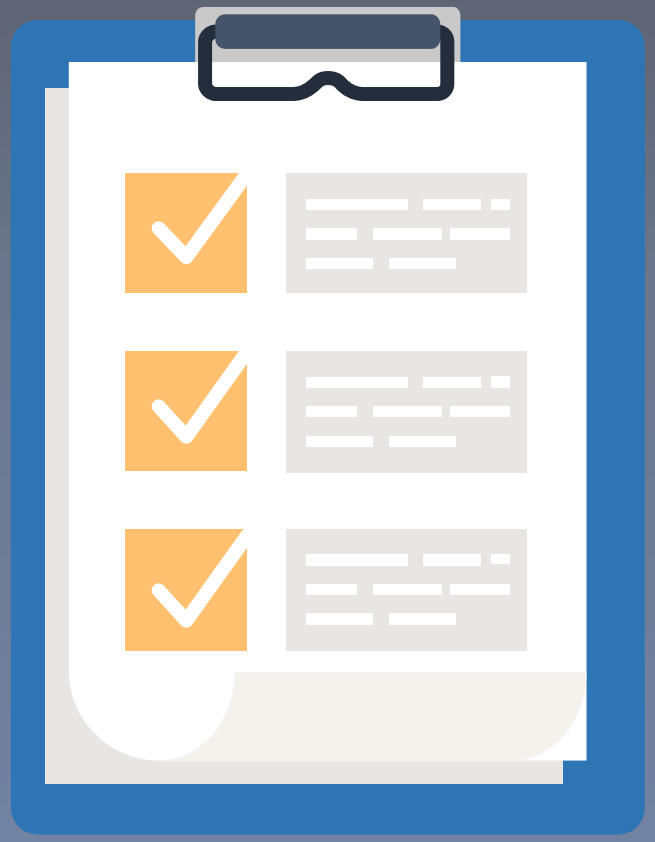
作者：Jacob Devlin

单位：Google

发表会议及时间：2018

上节回顾

Review in the lesson



深度之眼
deepshare.net

01 研究背景及成果意义

学习了GLUe以及概念feature-based和fine-tuning、了解了论文的实验结果。

02 论文总览

论文总共包含6个部分，论文主要介绍bert的结构。

03 Bert的衍生模型和Elmo、GPT、Bert的比较

学习了Bert的衍生模型，比较了Elmo、GPT以及Bert。

深度之眼
deepshare.net

第二课：论文精读

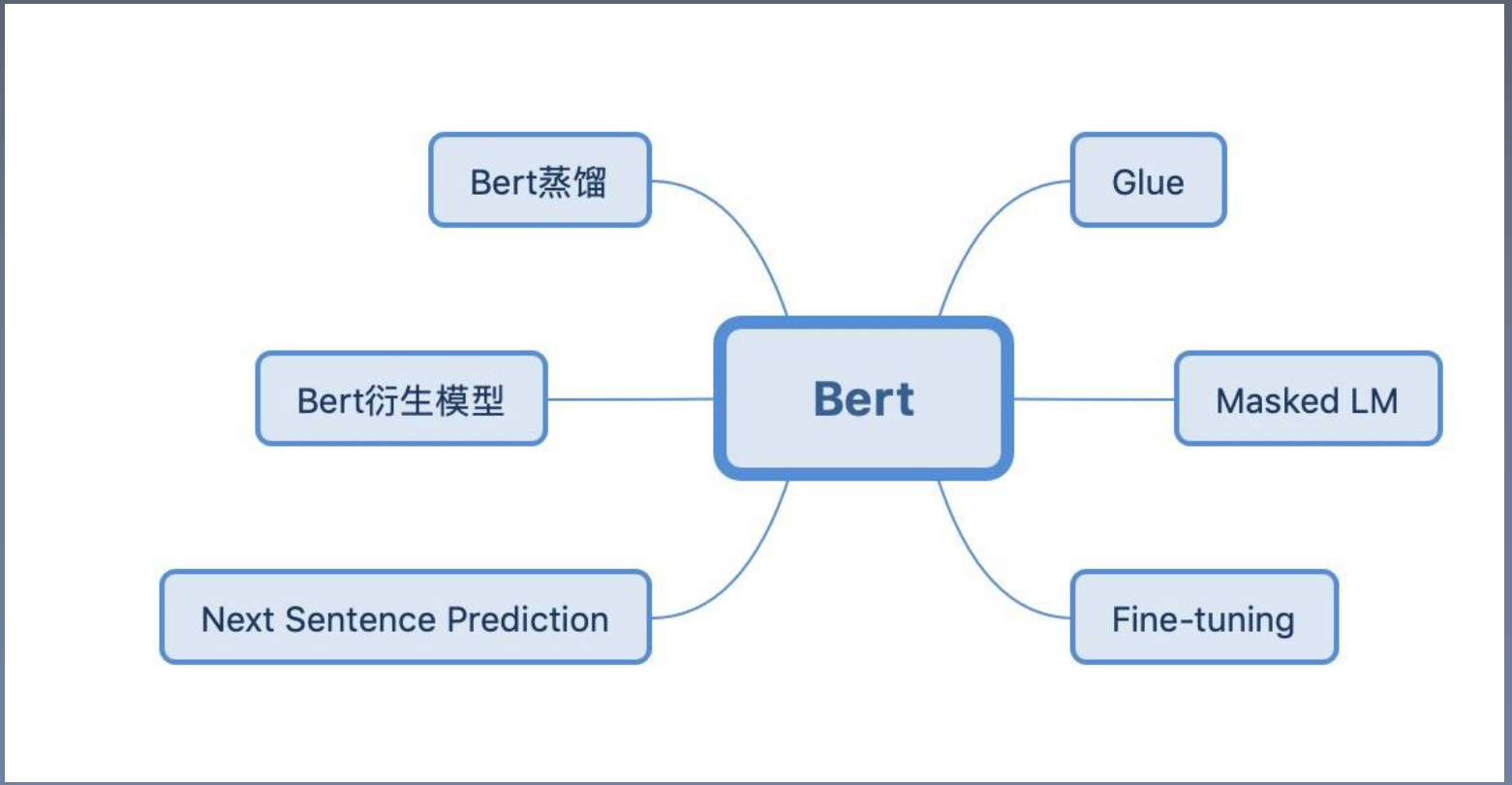
The second lesson: the paper in detail

深度之眼
deepshare.net

目 录

学习目标

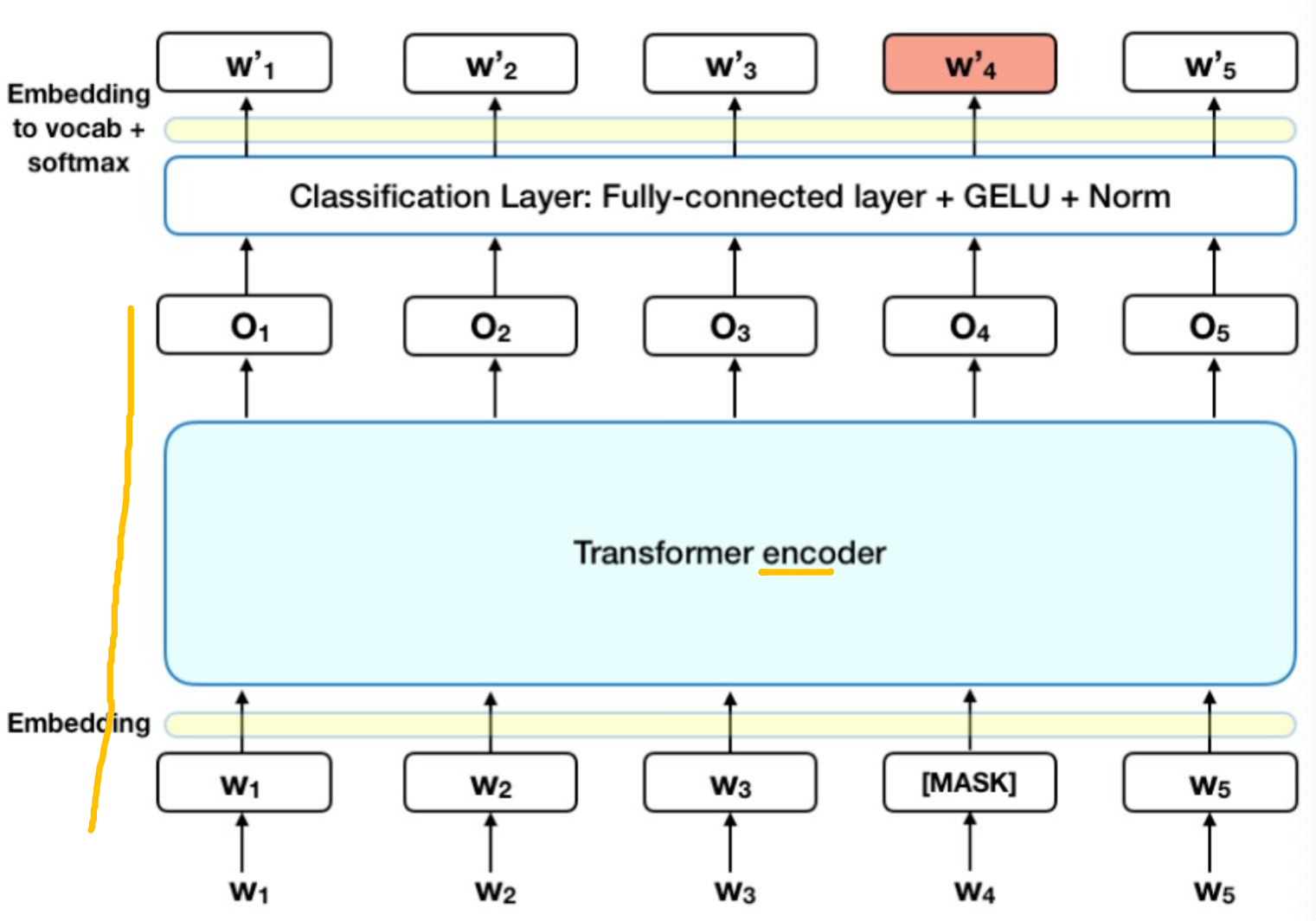
Learning objectives



深度之眼
deepshare.net

Model Architecture

Model Architecture

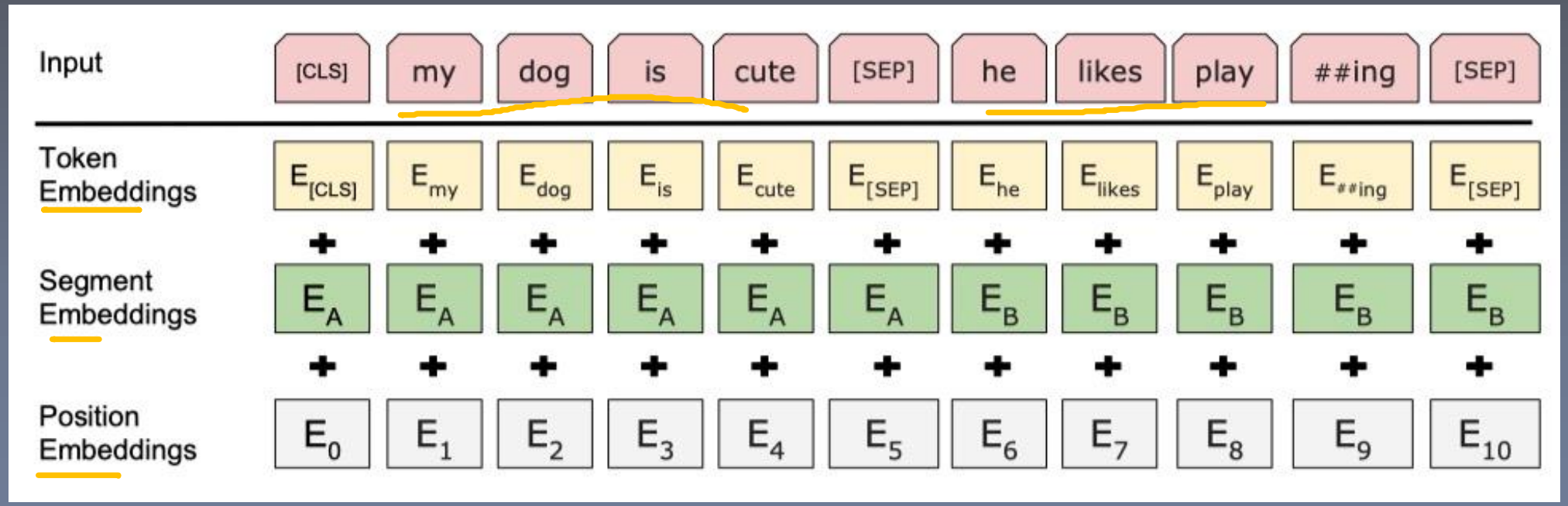


$$BERT_{BASE}(L = 12, H = 768, A = 12, TotalParameters = 110M)$$

$$BERT_{BASE}(L = 24, H = 1024, A = 16, TotalParameters = 340M)$$

L 代表Transformer的层数
H 代表hidden size
A 代表head的数量

Model Architecture



整理输入:(1)在第一个句子的开头开入[CLS]标价，在末尾加入[SEP]标记。

(2)将表示句子A或句子B的embedding添加到每个token上。

(3)给每个token准备对应的embedding和postion embedding

(4)句子和句子之间用[SEP]隔开

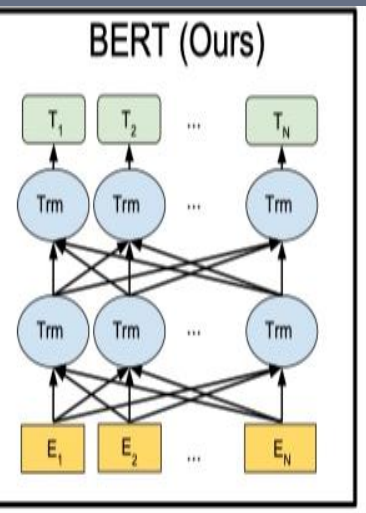
深度之眼
deepshare.net

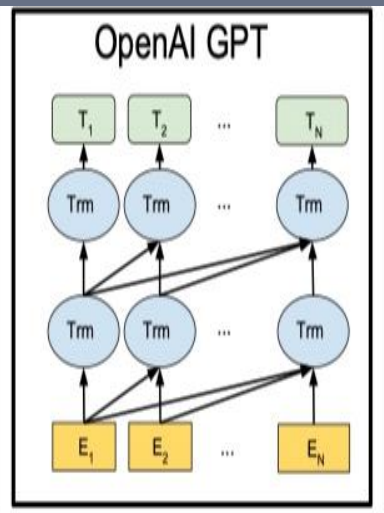
Pre-training BERT

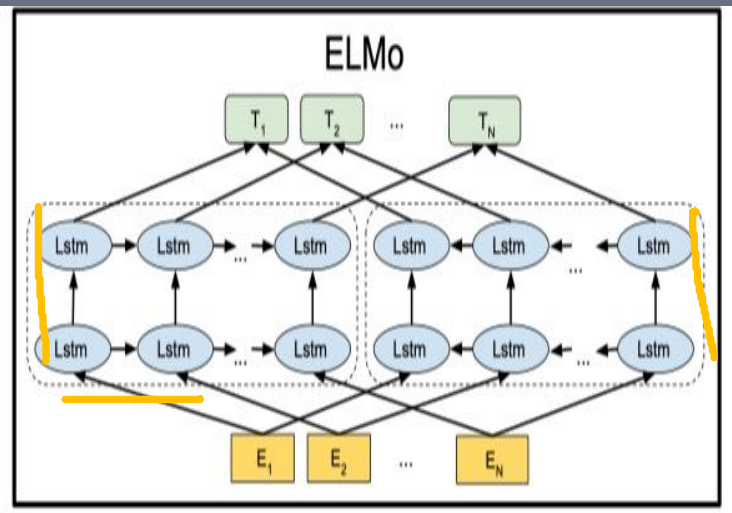
Pre-training fine-tuning

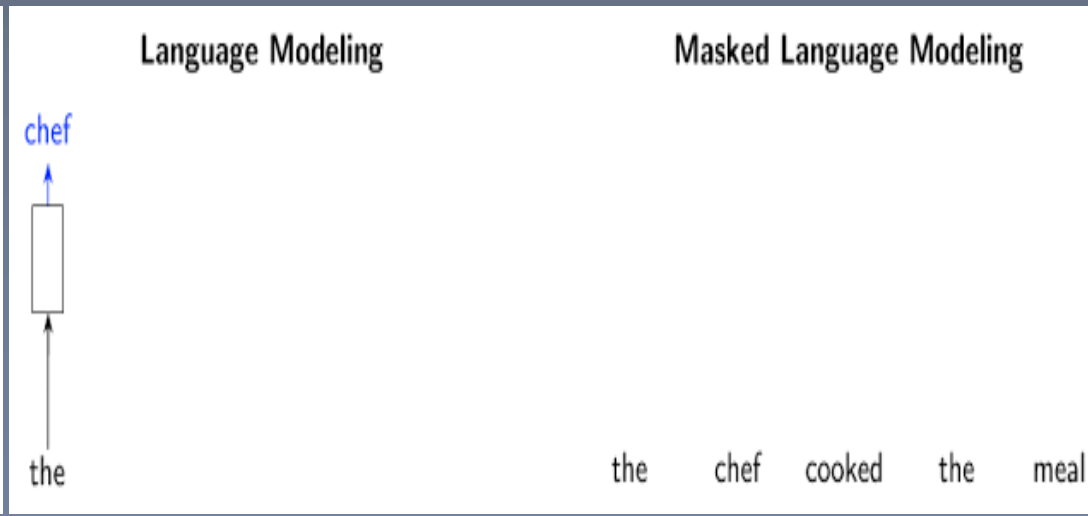
MLM

Bert、GPT、Elmo比较图



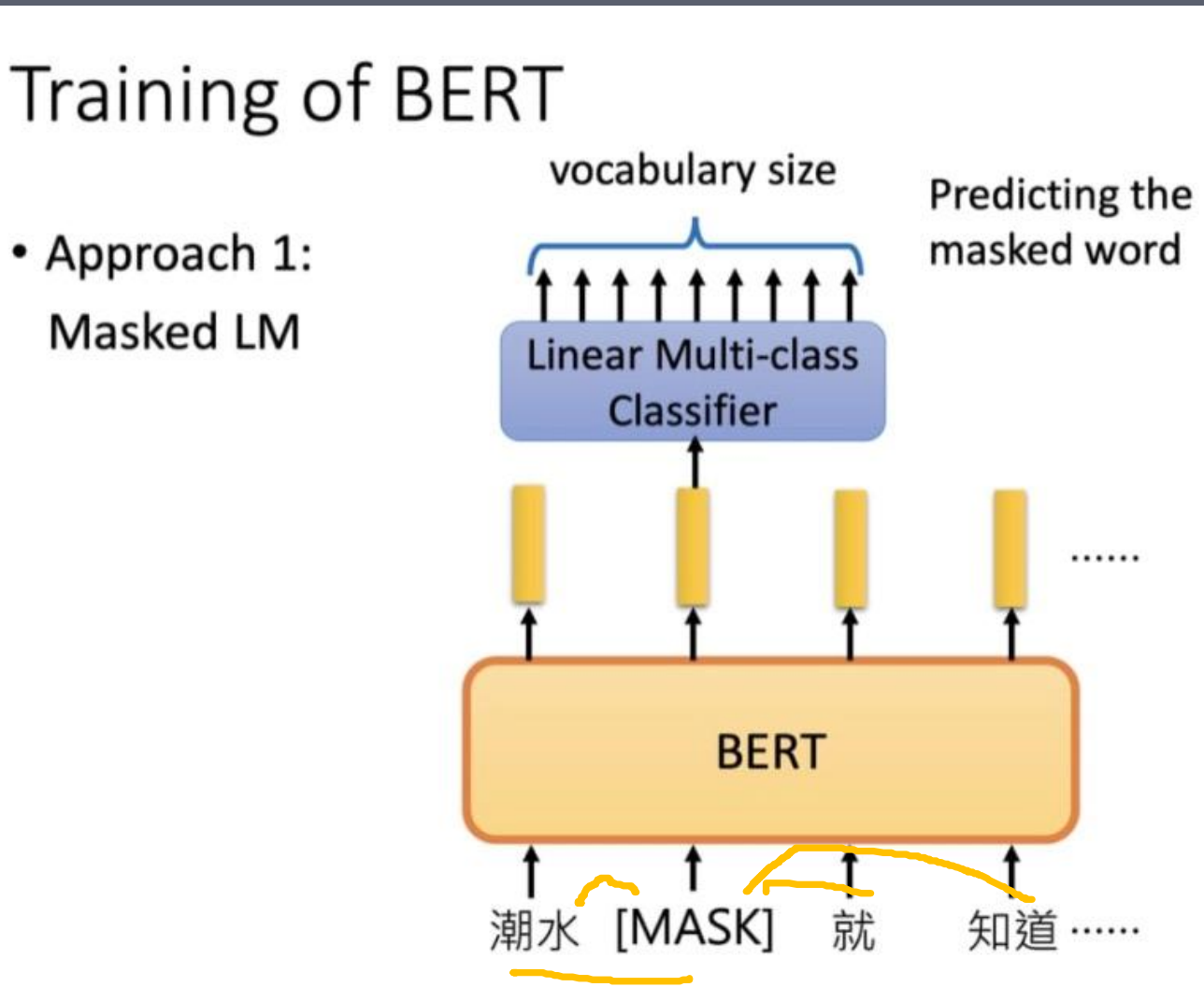






Pre-training fine-tuning

MLM



为了使模型学习到句子left-to-right和right-to-left的全局的信息，采取两种策略：

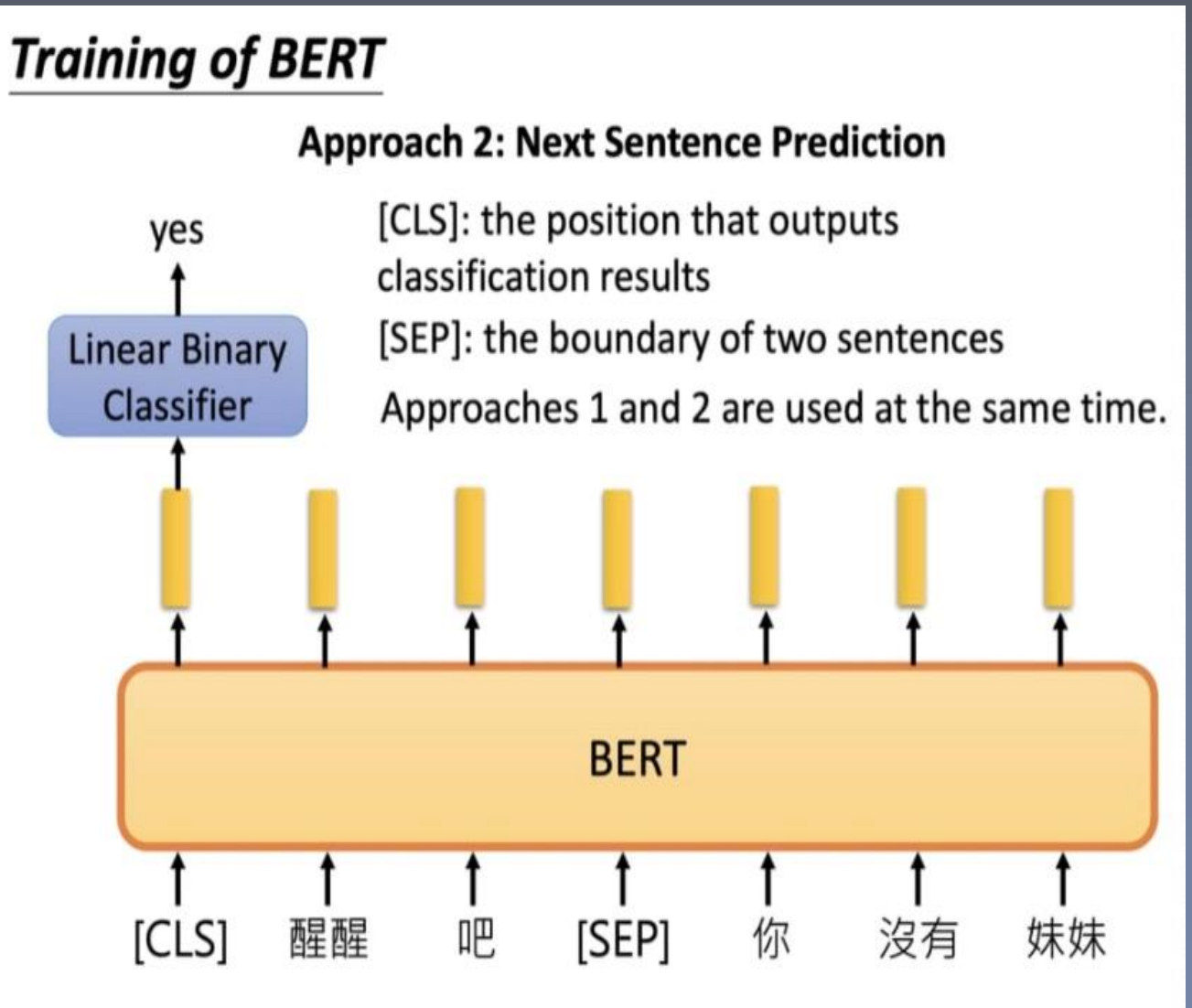
策略1:Masked LM，随机的mask一个句子中15%的词，用其上下文做预测
my dog is hairy--> my dog is [mask]

为了保证预训练和微调时的一致性，采取以下措施：

（1）80%的机会是采用mask
my dog is hairy -->my dog is [mask]
（2）10%的机会是随机一个词替代mask
my dog is hairy -->my dog is apple
（3）10%的机会是保持不变的
my dog is hairy -->my dog is hairy

Pre-training fine-tuning

NSP



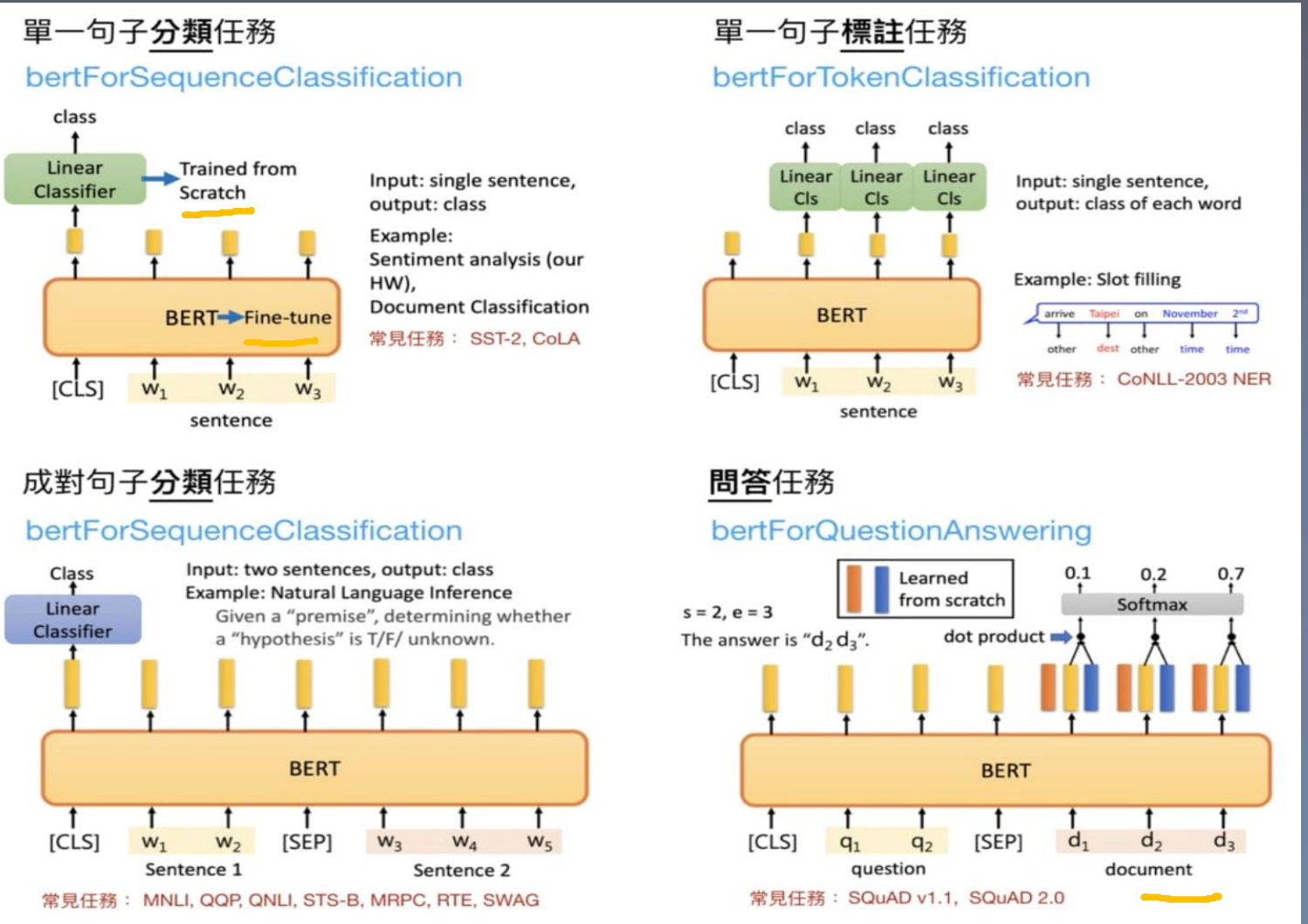
为了使模型学习到句子left-to-right和right-to-left的全局的信息，采取两种策略:

策略2:Next Sentence Prediction
选择一些句子对A和B，其中50%的数据B是A的下一条句子，50%的数据是从语料库中选取的，这样做是针对句子间的任务，例如SNLI。

深度之眼
deepshare.net

Fine-tuning BERT

Fine-tuning BERT



在句子分类任务中，就是在[CLS]对应的token处接相应的线性层，BERT结构的参数是fine-tunied，但是线性层的参数是从头到尾训练的。

ner任务中每个token后面都会有对应的输出。

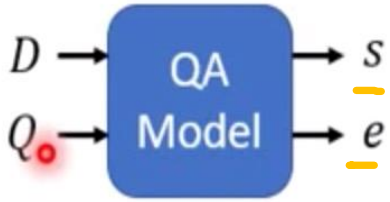
Fine-tuning BERT

How to use BERT – Case 4

- Extraction-based Question Answering (QA) (E.g. SQuAD)

Document: $D = \{d_1, d_2, \cdots, d_N\}$

Query: $Q = \{q_1, q_2, \cdots, q_M\}$



output: two integers (s, e)

Answer: $A = \{d_s, \cdots, d_e\}$

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity. The main forms of precipitation include drizzle, rain, sleet, snow, graupel and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals within a cloud. Short, intense periods of rain in scattered locations are called "showers".

What causes precipitation to fall?
gravity

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupel

Where do water droplets collide with ice crystals to form precipitation?
within a cloud

在句子分类任务中，就是在[CLS]对应的token处接相应的线性层，BERT结构的参数是fine-tunied，但是线性层的参数是从头到尾训练的。

ner任务中每个token后面都会有对应的输出。

Fine-tuning BERT

深度之眼
deepshare.net

模型蒸馏

1、首先需要训练一个大的模型，这个大模型也称为 teacher 模型。

2、利用 teacher 模型输出的概率分布训练小模型，小模型称为 student 模型。

3、训练 student 模型时，包含两种 label，soft label 对应了 teacher 模型输出的概率分布，而 hard label 是原来的 one-hot label。

4、模型蒸馏训练的小模型会学习到大模型的表现以及泛化能力。

$$KL(p||q) = E_p(log(p/q)) = \sum_i p_i log(p_i) - \sum_i p_i log(q_i)$$

i代表当前的token，p代表teacher模型的分布，q代表student模型的分布，采用KL散度作为损失函数

深度之眼
deepshare.net

实验设置和结果分析

Experiment results

实验结果及分析

Results and Discussion

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**

Bert在下游任务中的表现完全远超之前的模型。

论文总结

论文总结

Summary of the paper

深度之眼
deepshare.net

A

关键点

- MLM思想
- NSP
- fine-tuning BERT

B

小细节

- BETRT缺点。
- 模型蒸馏。

论文总结

Summary of the paper

C

启发点

- 写模型可以采用bert+的形式
- 写模型可以采用蒸馏的形式

深度之眼
deepshare.net

本课回顾及下节预告

Review in the lesson and Preview of next lesson

本节回顾

Preview of next lesson

01 Pre-training Bert

学习Bert的pre-training部分

02 Fine-tuning Bert

学习Bert的fine-tuning部分

03 实验设置及结果分析

比较了模型在几个数据集上的表现情况。

04 论文总结

总结论文中创新点、关键点及启发点

下节课前准备

Preview of next lesson

- 再次阅读BERT论文
- 熟悉BERT模型结构及数据预处理方式
- 配置PyTorch开发环境
- 下载BERT代码
- https://github.com/huggingface/transformers

结 语

循循而进，欲速则不达也。

深度之眼
deepshare.net

联系我们：
电话：18001992849
邮箱：service@deepshare.net
ＱＱ：2677693114

公众号

客服微信