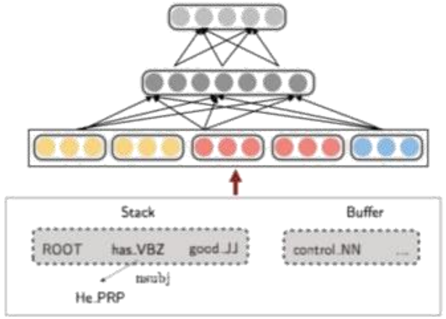
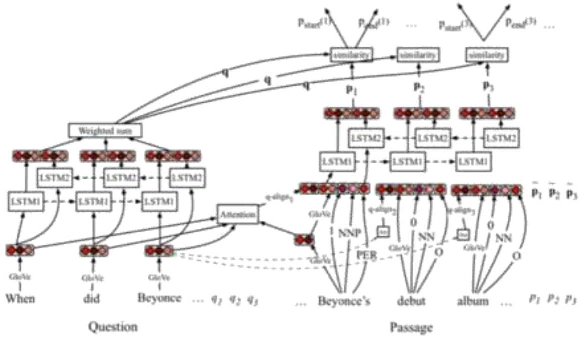
BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding 论文导读
&
Bert 详解

Why Bert ?

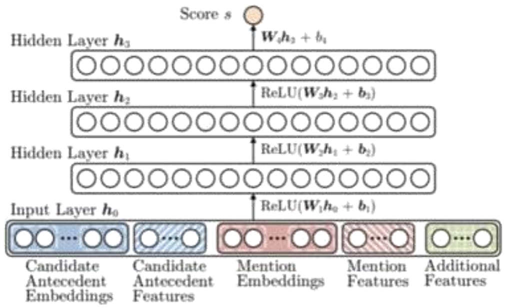
Various Model Architectures for Different NLP Tasks



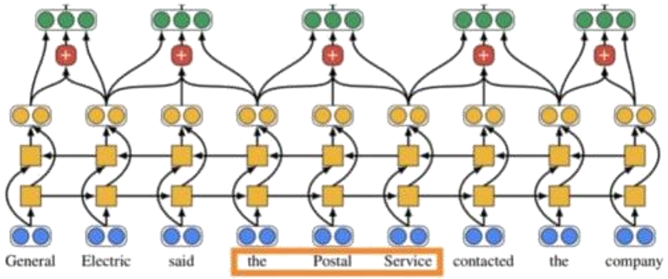
Dependency Parsing



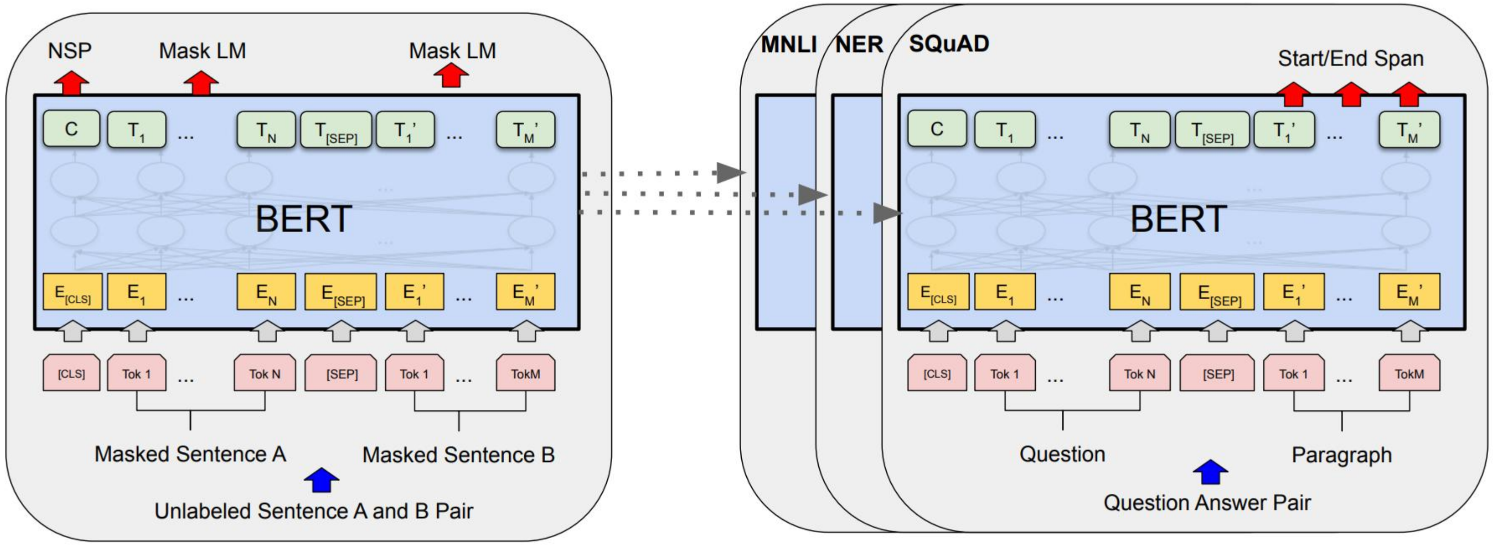
Question Answering



Coreference example 1



Coreference example 2



Pre-training

Fine-Tuning

论文信息

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

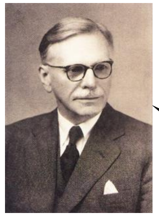
Jacob Devlin **Ming-Wei Chang** **Kenton Lee** **Kristina Toutanova**
Google AI Language
{jacobdevlin,mingweichang,kentonl,kristout}@google.com

Kenton, J. D. M. W. C., & Toutanova, L. K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of NAACL-HLT (pp. 4171-4186).

Language Model

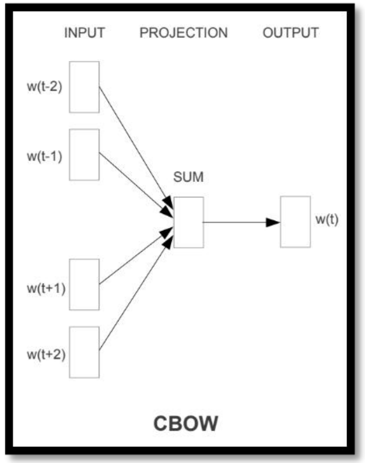
语言模型：

给定词典 V，计算出任意单词序列 w1 , w2 , . . . , wn 是一句话的概率: p(w1 , w2 , . . . , wn), p > = 0.



John Rupert Firth

You shall know a word by the company it keeps.



word embedding

语言模型 — 2003 Bengio

↓

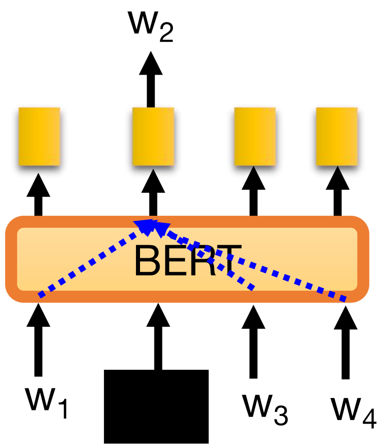
词嵌入阶段（+ 复杂网络模型）:

- word2vec — 2013 Mikolov
- glove — 2014 Jeffrey
- fasttext — 2016 facebook

↓

预训练语言模型阶段（+ 简单 MLP）:

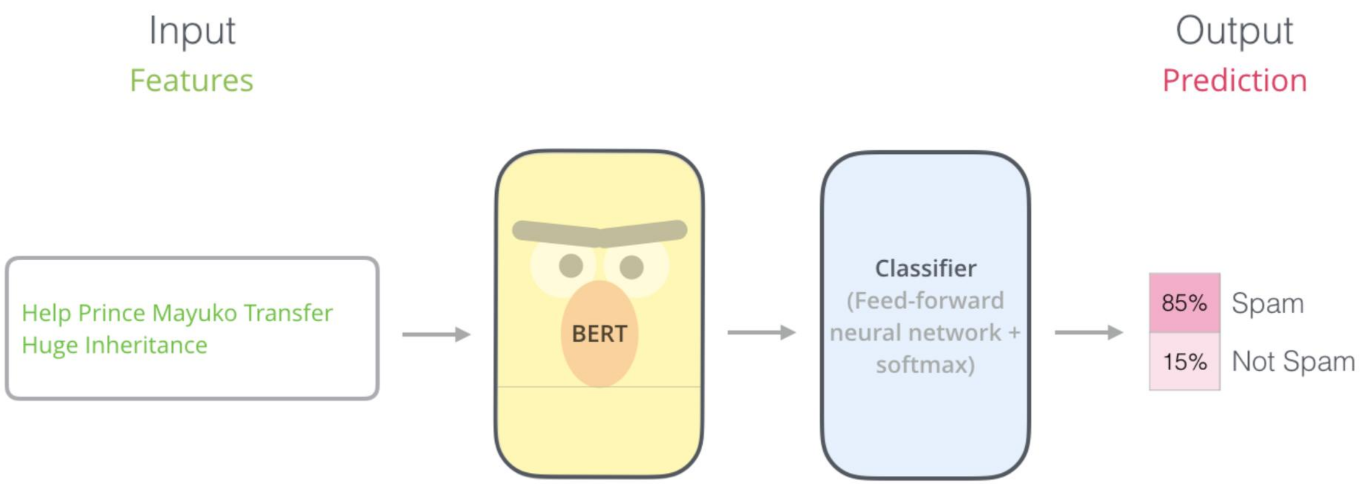
- Elmo — 2018.2 Allen
- GPT — 2018.6 OpenAI
- BERT — 2018.10 google

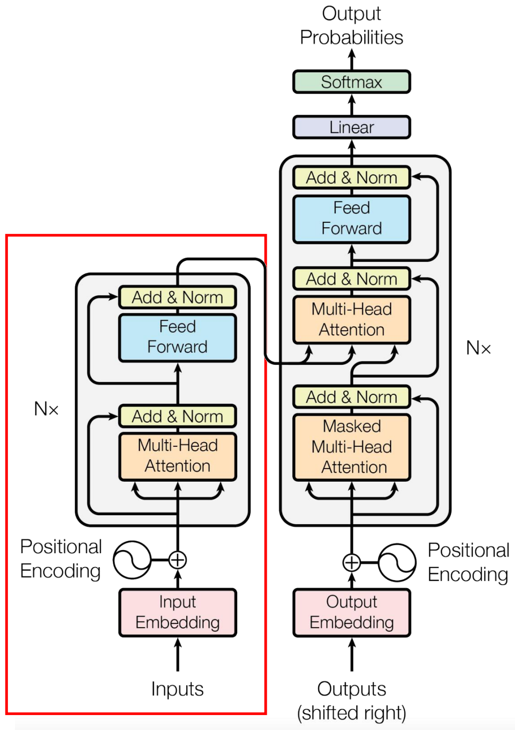


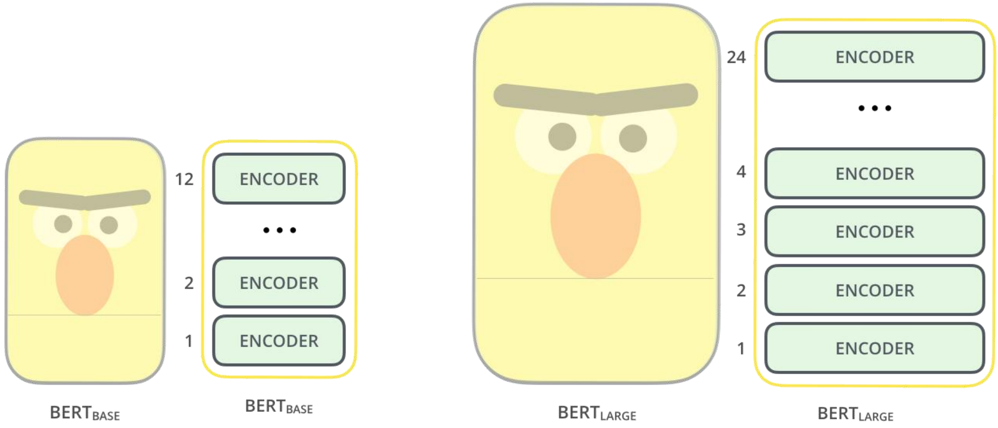
Contextualized word embedding

Bert 架构

Bidirectional Encoder Representation from Transformer







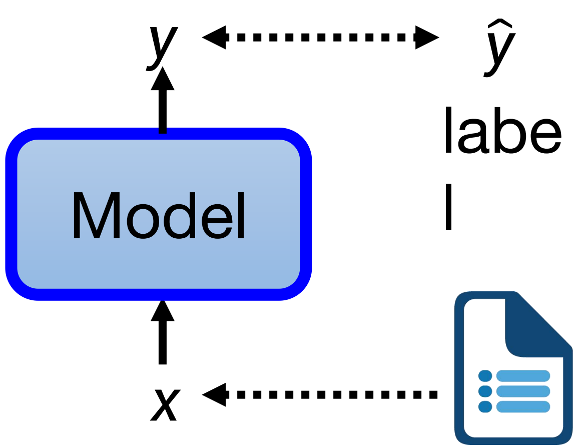
$\text{Bert}_{\text{BASE}}$： L=12， H=768， A=12， 参数总量 110M
$\text{Bert}_{\text{LARGE}}$： L=24， H=1024， A=16， 参数总量 340M

L：block数量
H：隐藏层维度
A：注意力头的数量

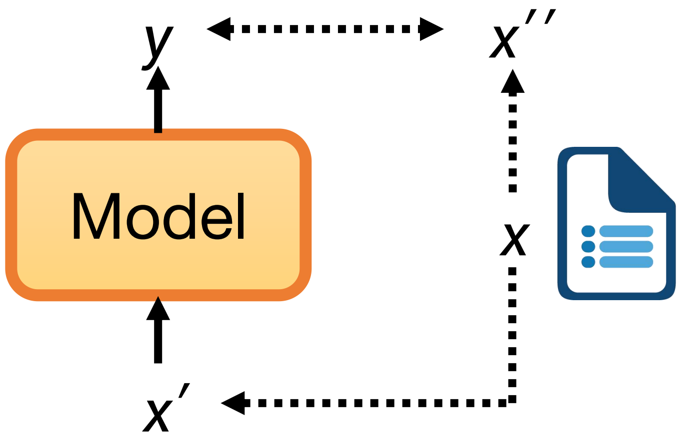
Bert 训练

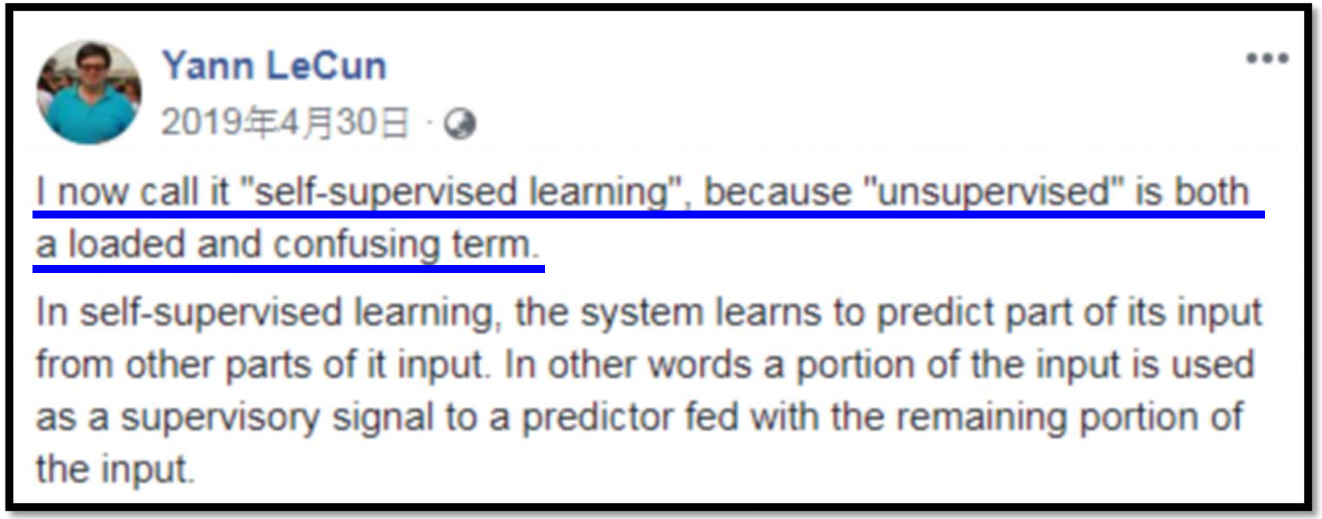
Supervised & Self-Supervised

Supervised



Self-supervised





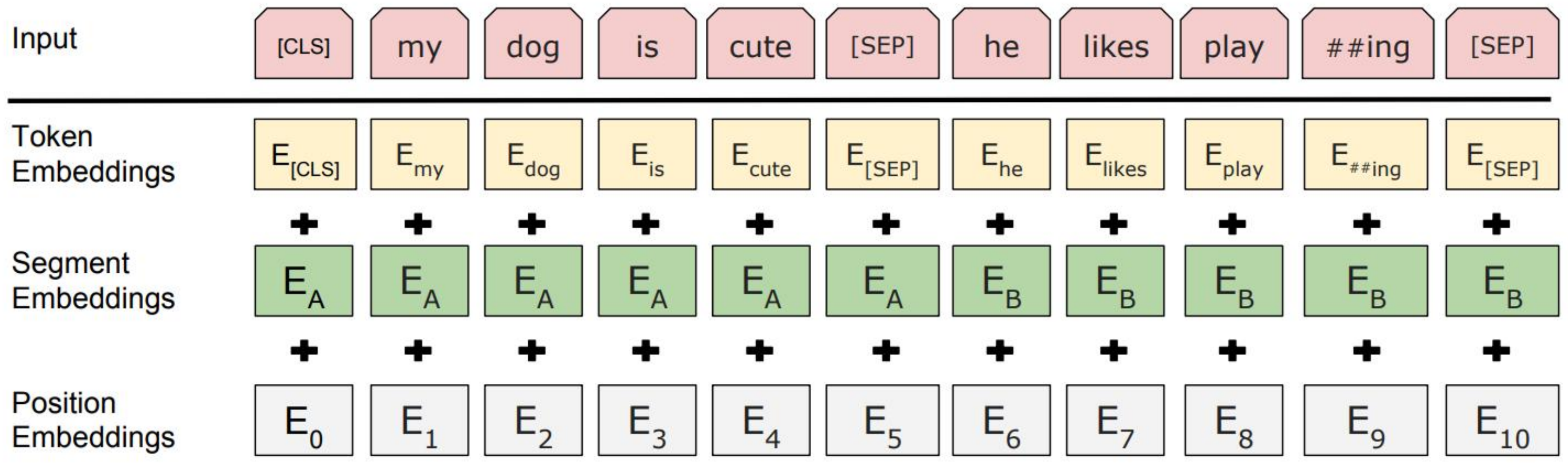
Yann LeCun
2019年4月30日 ·

I now call it "self-supervised learning", because "unsupervised" is both a loaded and confusing term.

In self-supervised learning, the system learns to predict part of its input from other parts of it input. In other words a portion of the input is used as a supervisory signal to a predictor fed with the remaining portion of the input.

Bert 训练

输入向量



将单词转换为固定维度的向量

区分句子对的上下句

标记句中每个词的位置

Segment Embeddings 示例：

[CLS] 我 的 狗 很 可 爱 [SEP] 企 鹅 不 擅 长 飞 行 [SEP]
0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1

Bert 训练

任务一：MLM（Masked Language Modeling）

MASK 策略示例：

对于语句“my dog is hairy”，随机把句中15%的token替换为以下内容：

80%的几率被替换成[MASK]：

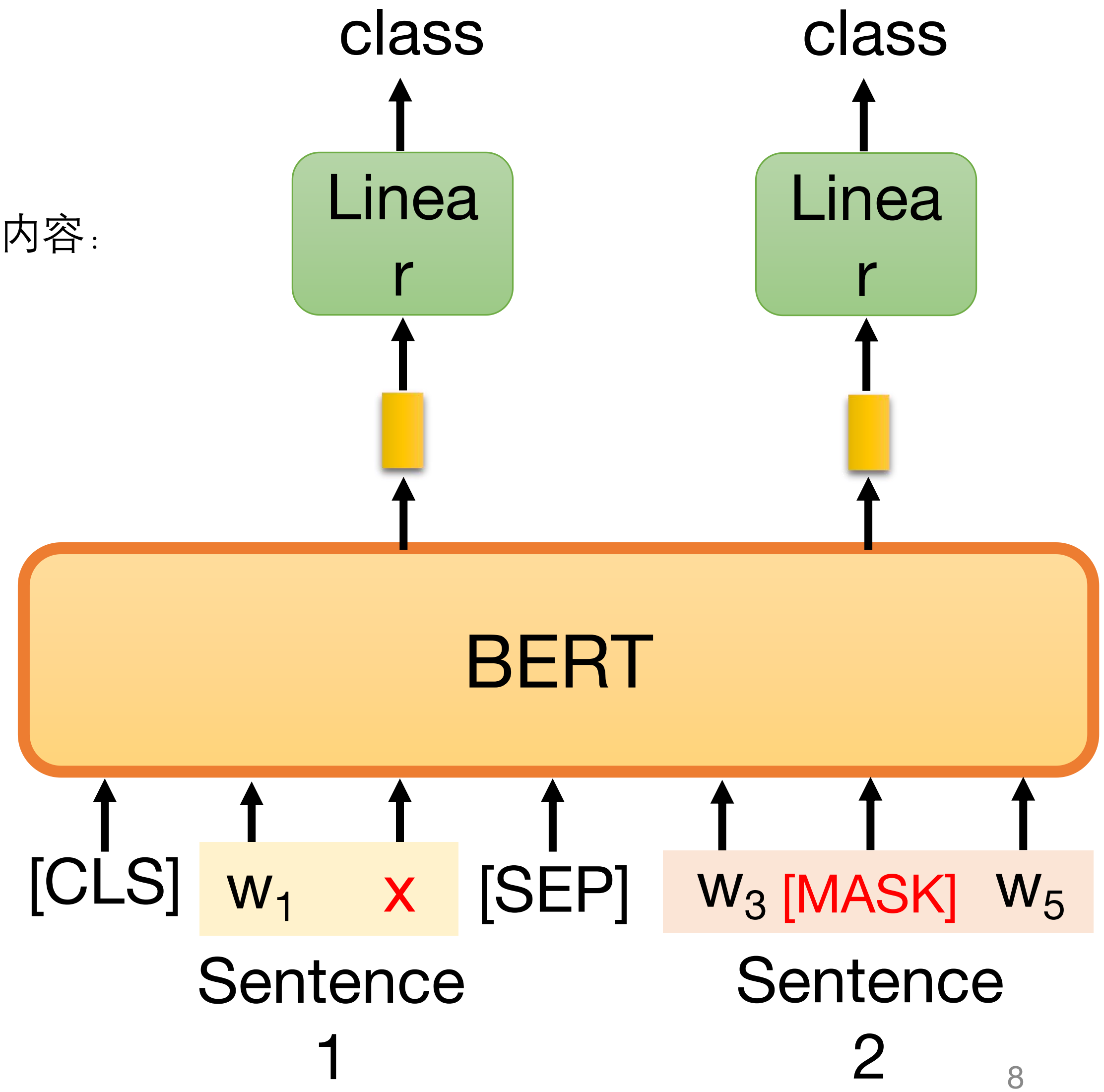
“my dog is hairy” → “my dog is [MASK]”

10%的几率被替换成其他token：

“my dog is hairy” → “my dog is apple”

10%的几率原封不动：

“my dog is hairy” → “my dog is hairy”



Bert 训练

任务二：NSP（Next Sentence Prediction）

正负句子对样本：

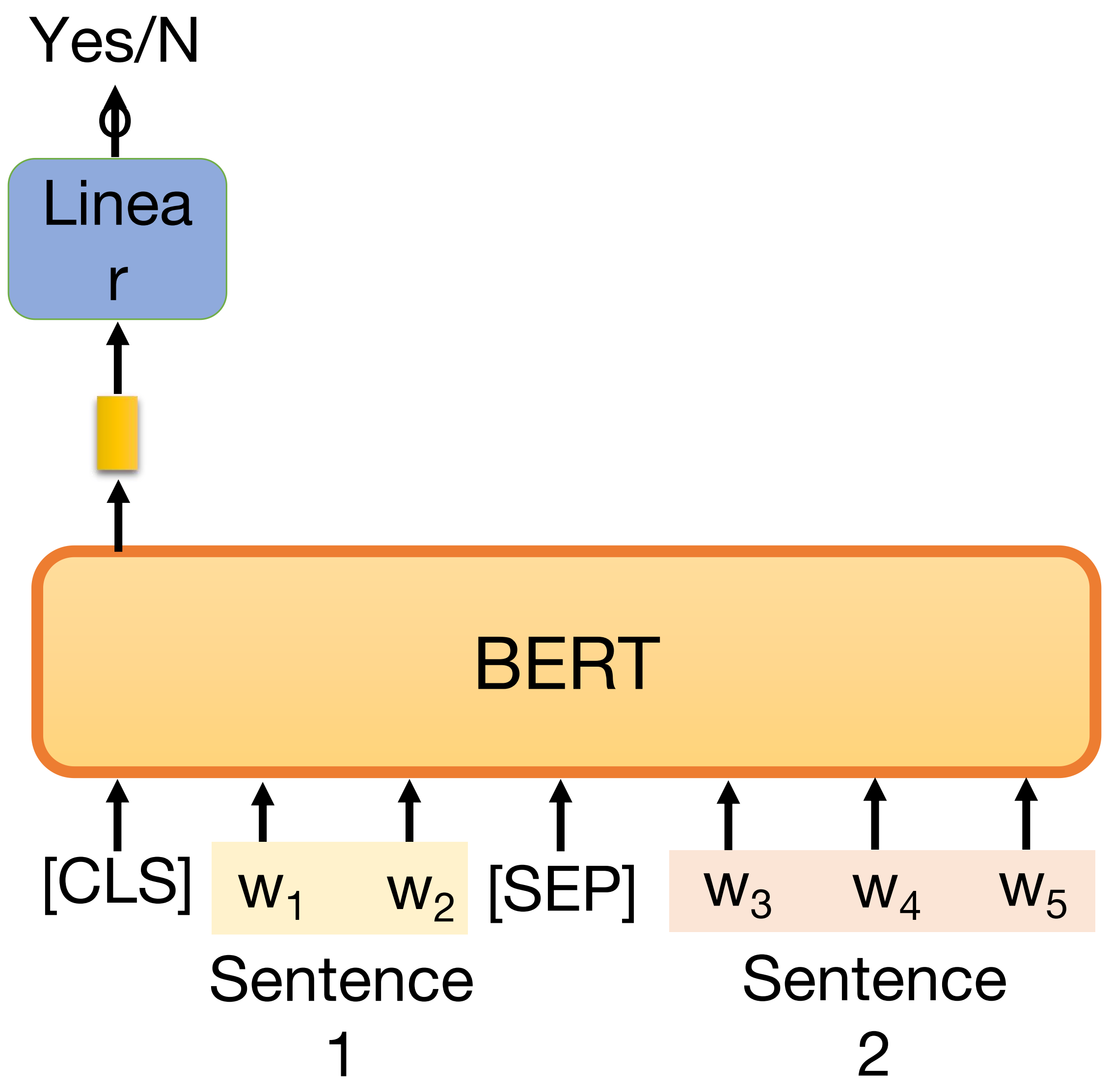
50% 的正样本：训练语料库中的两个连续段落

50% 的负样本：来自不同文档的两个随机段落

示例：

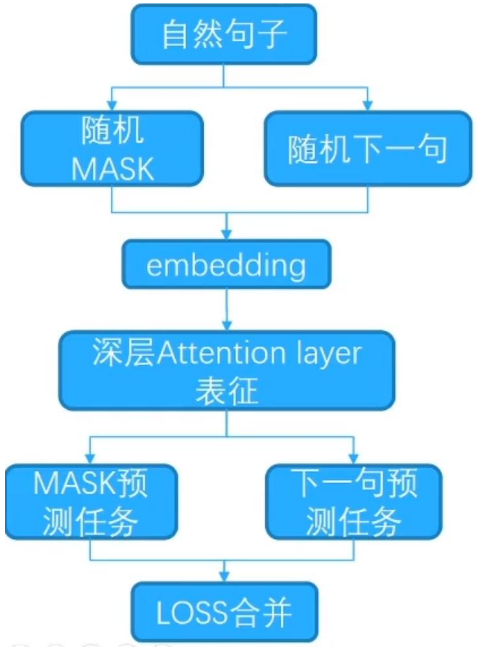
Input：[CLS] 博学而笃志 [SEP] 切问而近思 [SEP]
Target：Yes

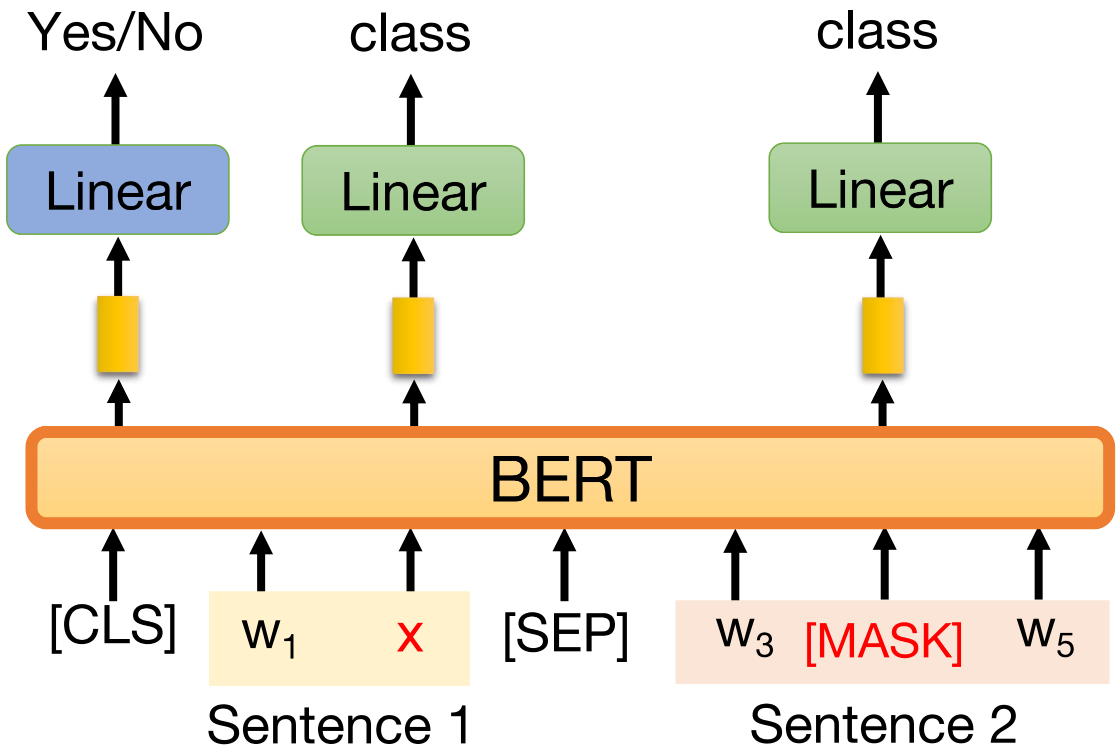
Input：[CLS] 博学而笃志 [SEP] 今天风好大 [SEP]
Target：No



Bert 训练

Multi-Task Learning

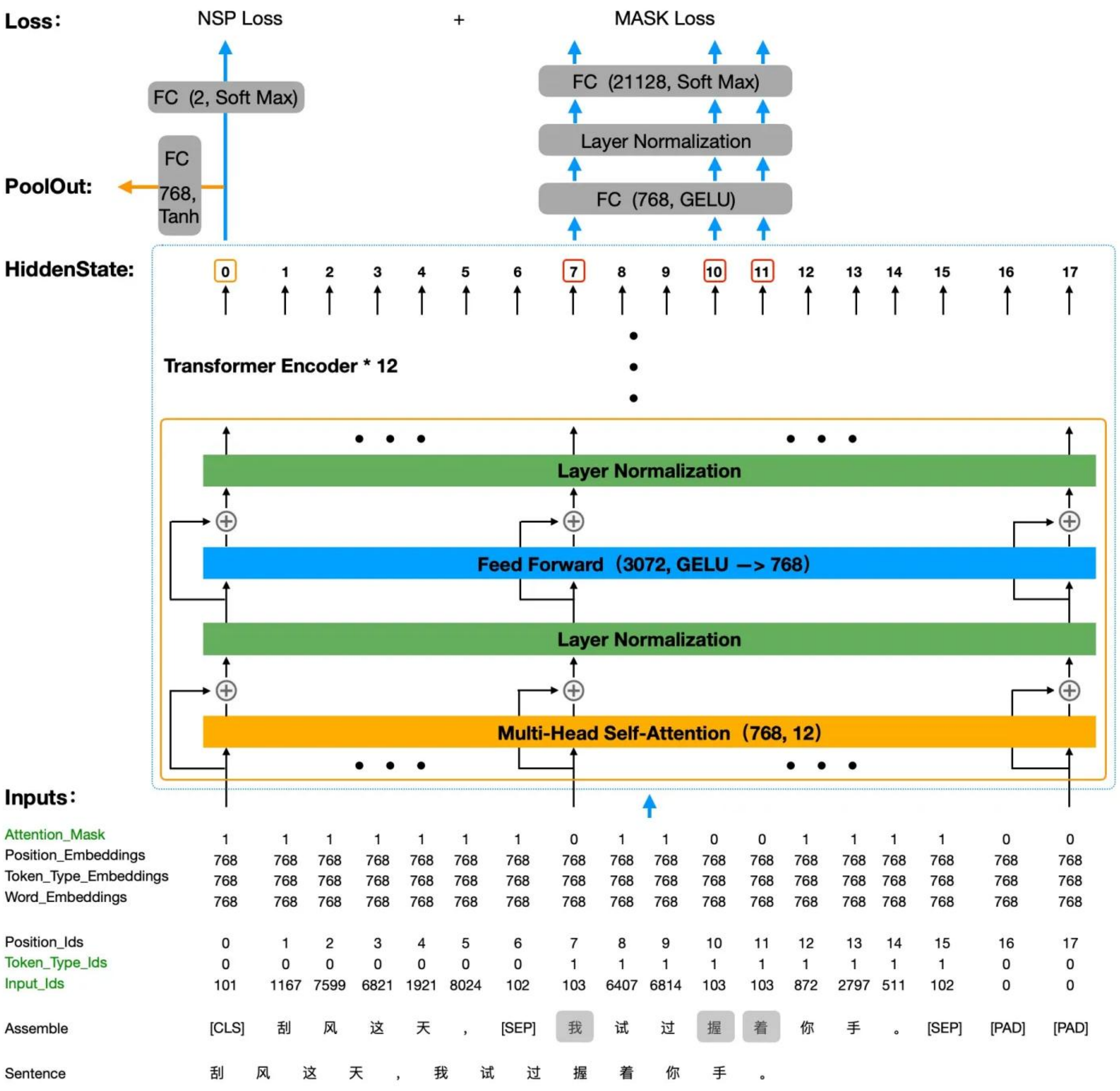




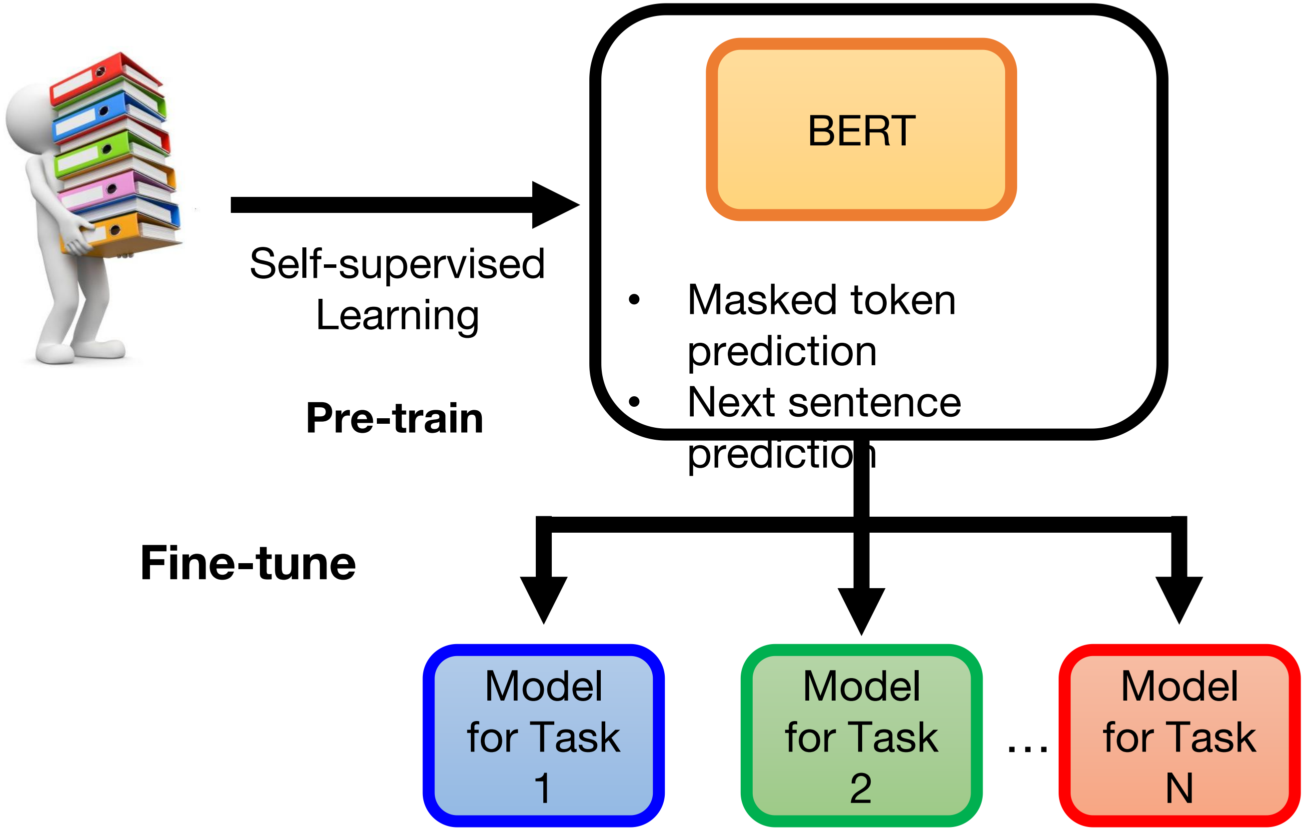
- Input: [CLS] calculus is a branch of math [SEP] panda is native to [MASK] central china [SEP]
- Targets: false, south
- ------------------------------------
- Input: [CLS] calculus is a [MASK] of math [SEP] it [MASK] developed by newton and leibniz [SEP]
- Targets: true, branch, was

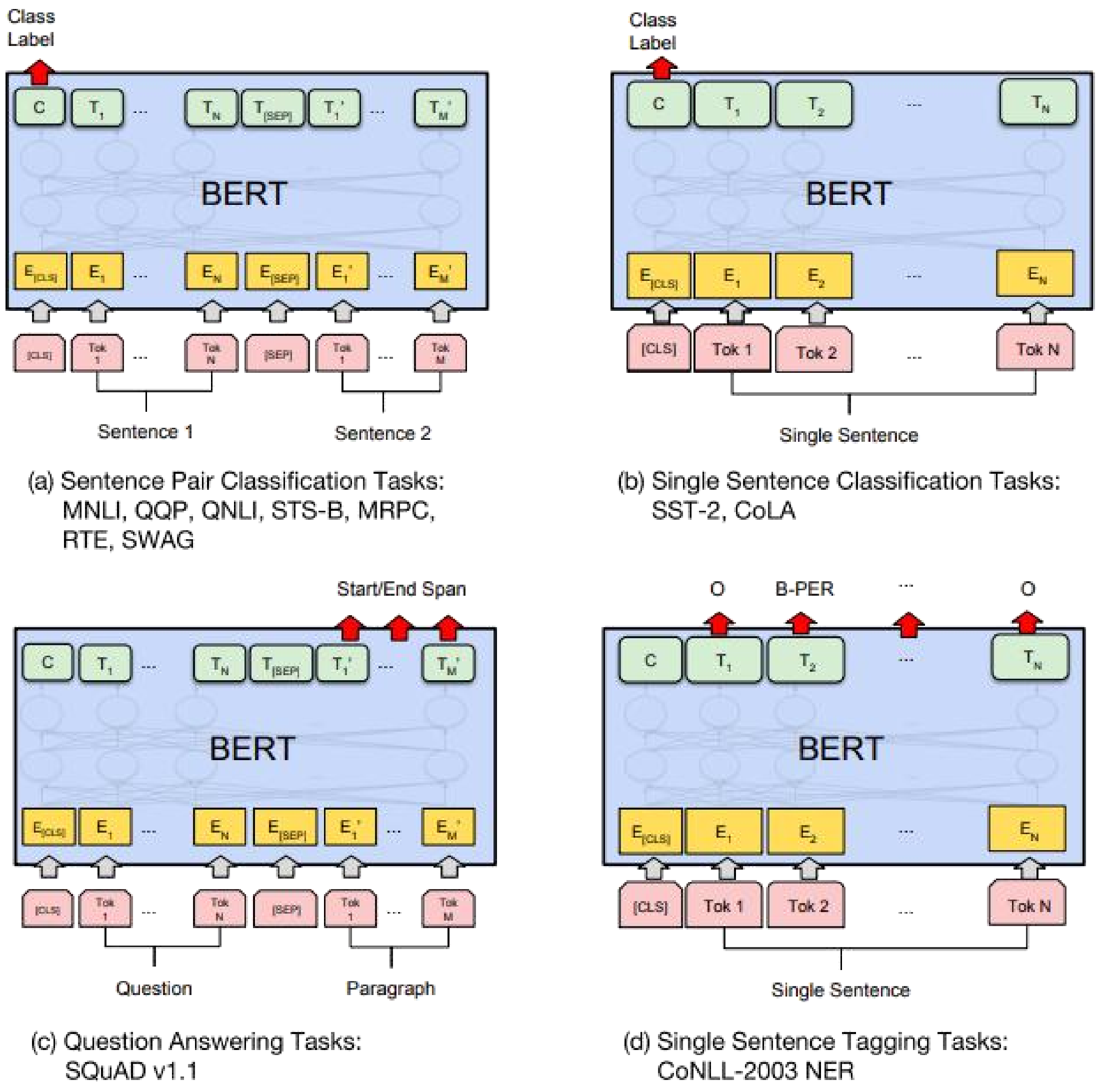
Bert 训练

Multi-Task Learning



Bert 使用



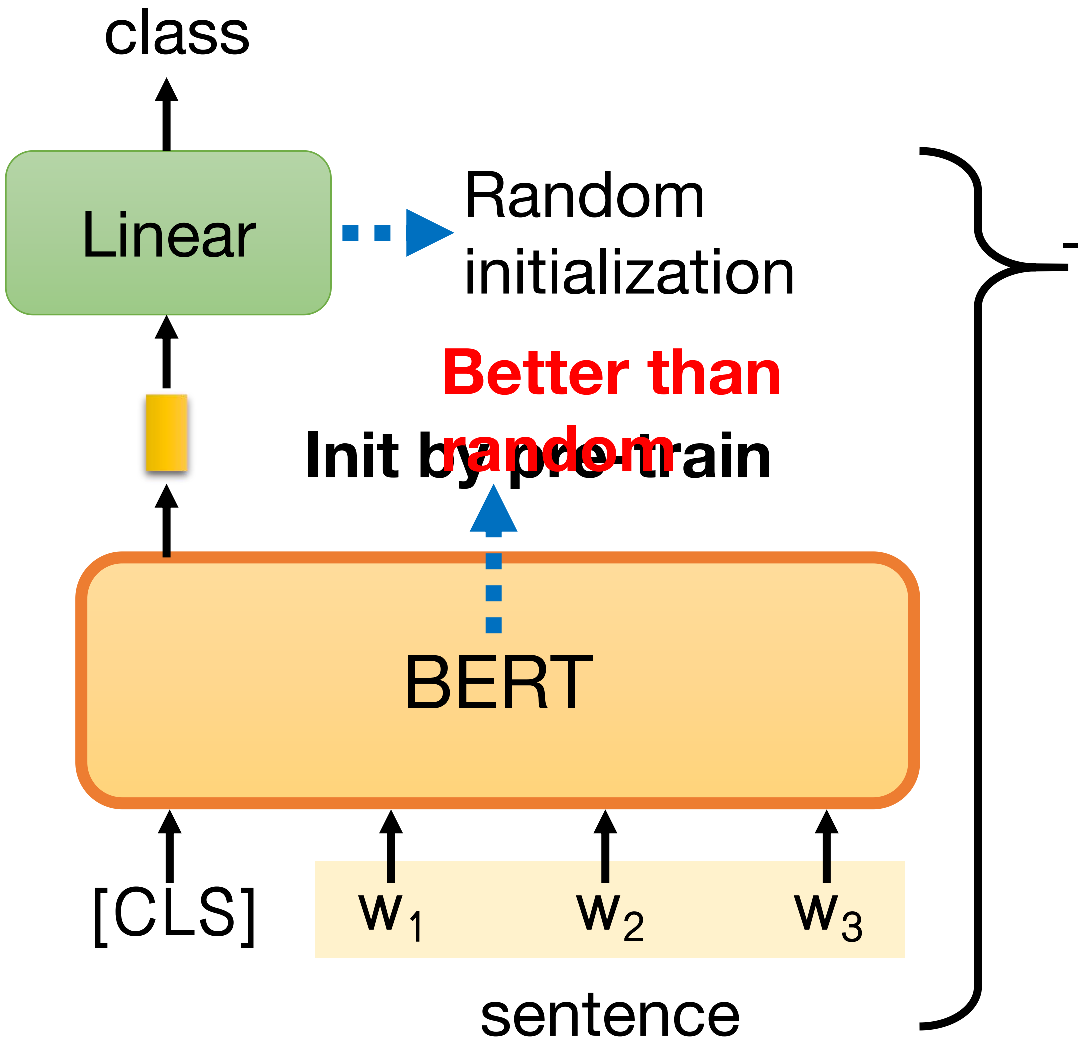


Downstream Tasks

- The tasks we care
- We have a little bit labeled data.

Bert 使用

应用场景 1 文本分类



This is the model to be learned.

Input:
sequence
output: class

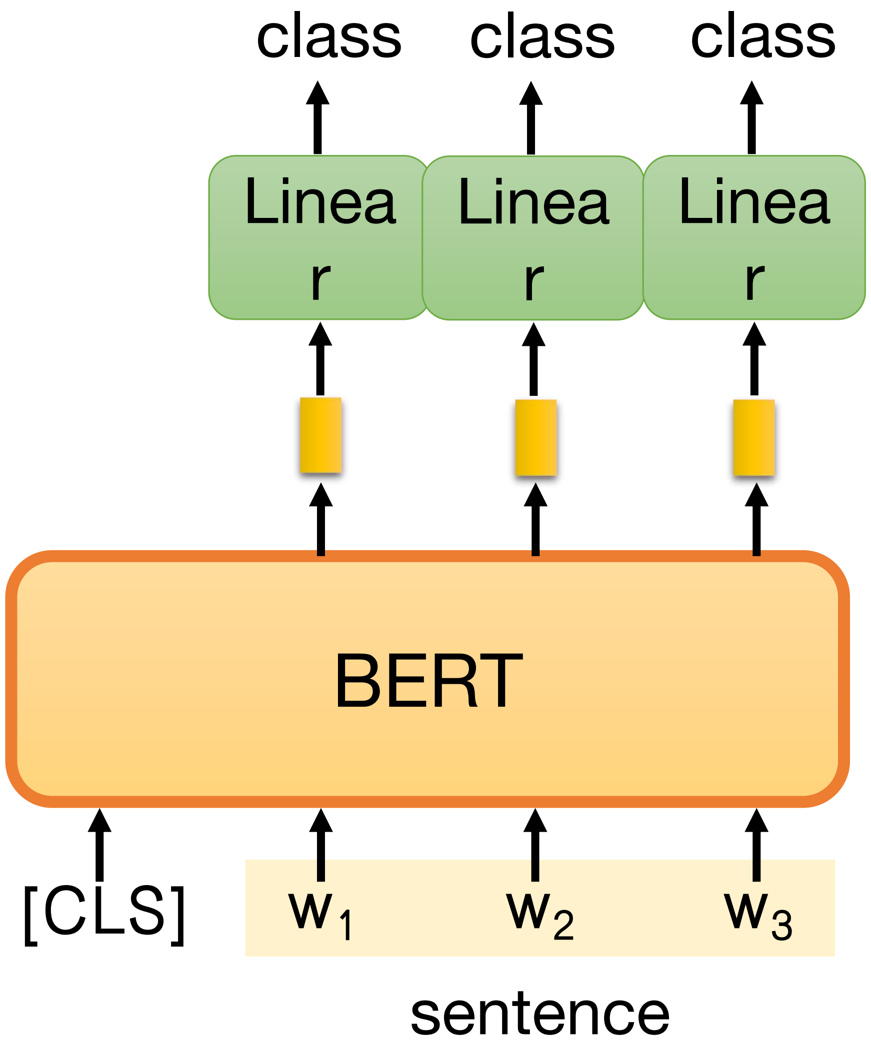
Example: Sentiment analysis

this is good

positive

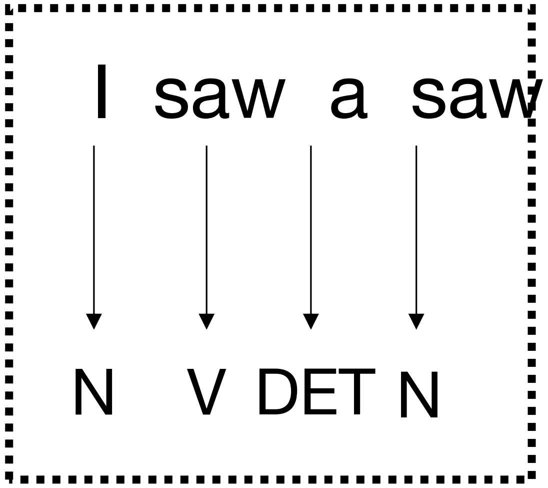
Bert 使用

应用场景 2 序列标注



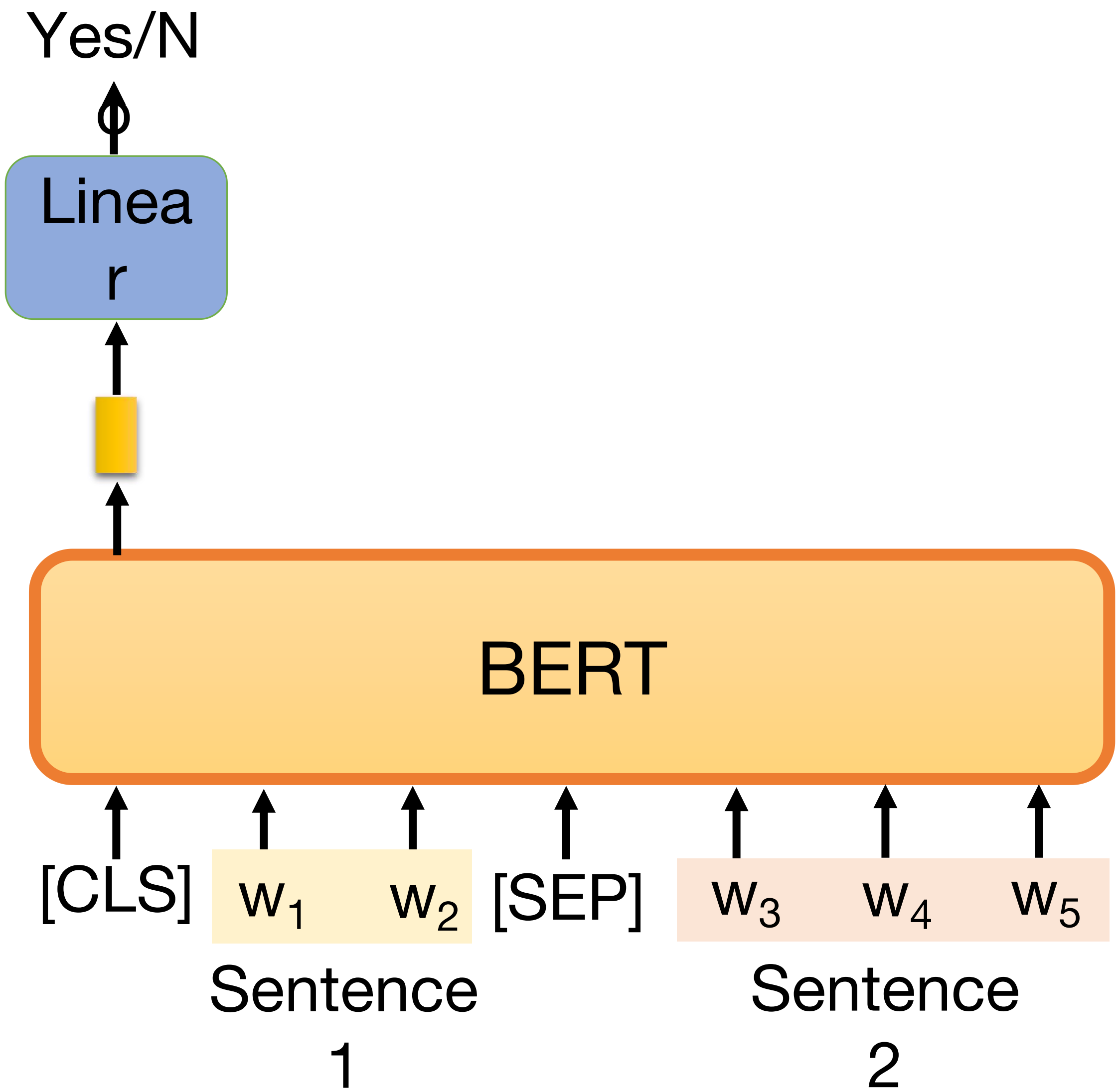
Input: sequence
output: same as input

Example: POS tagging



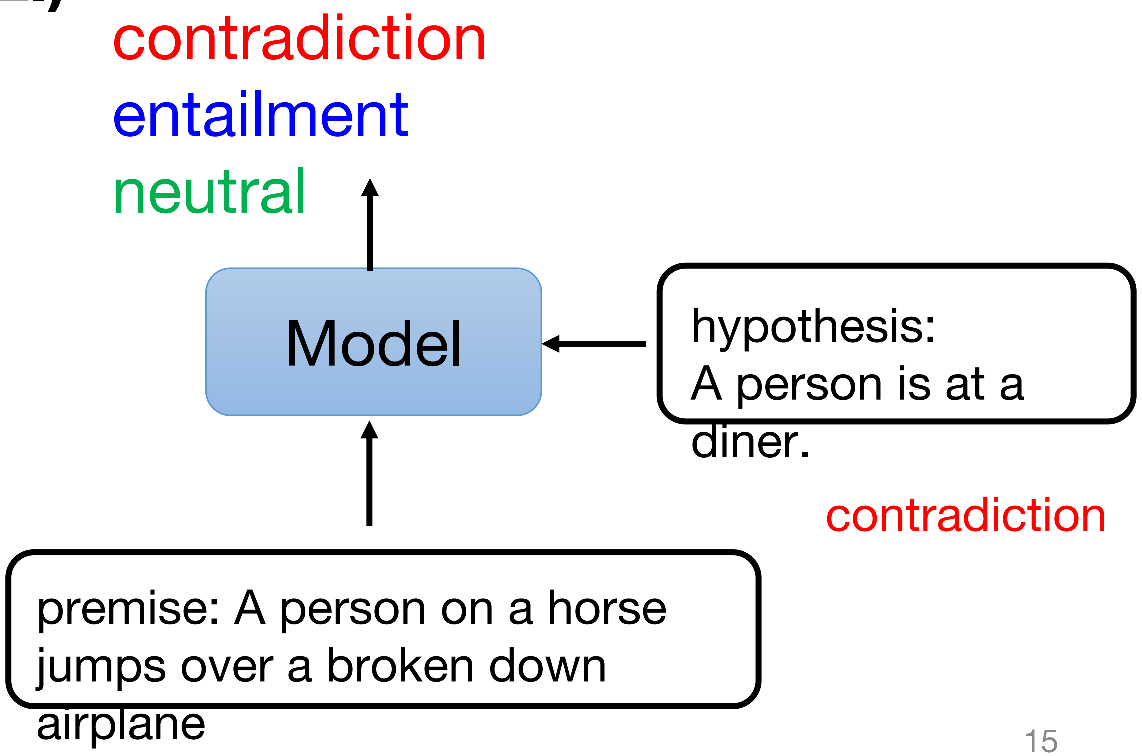
Bert 使用

应用场景 3 自然语言推理



Input: two
sequences
Output: a class

Example: Natural Language Inferencee (NLI)



Bert 使用

应用场景 4 抽取式问答

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

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

output: two integers (s, e)

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

In meteorology, precipitation is any product of the condensation of 17 spheric 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 77 atte 79 cations are called "showers".

What causes precipitation to fall?

gravity $s = 17, e = 17$

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

graupel

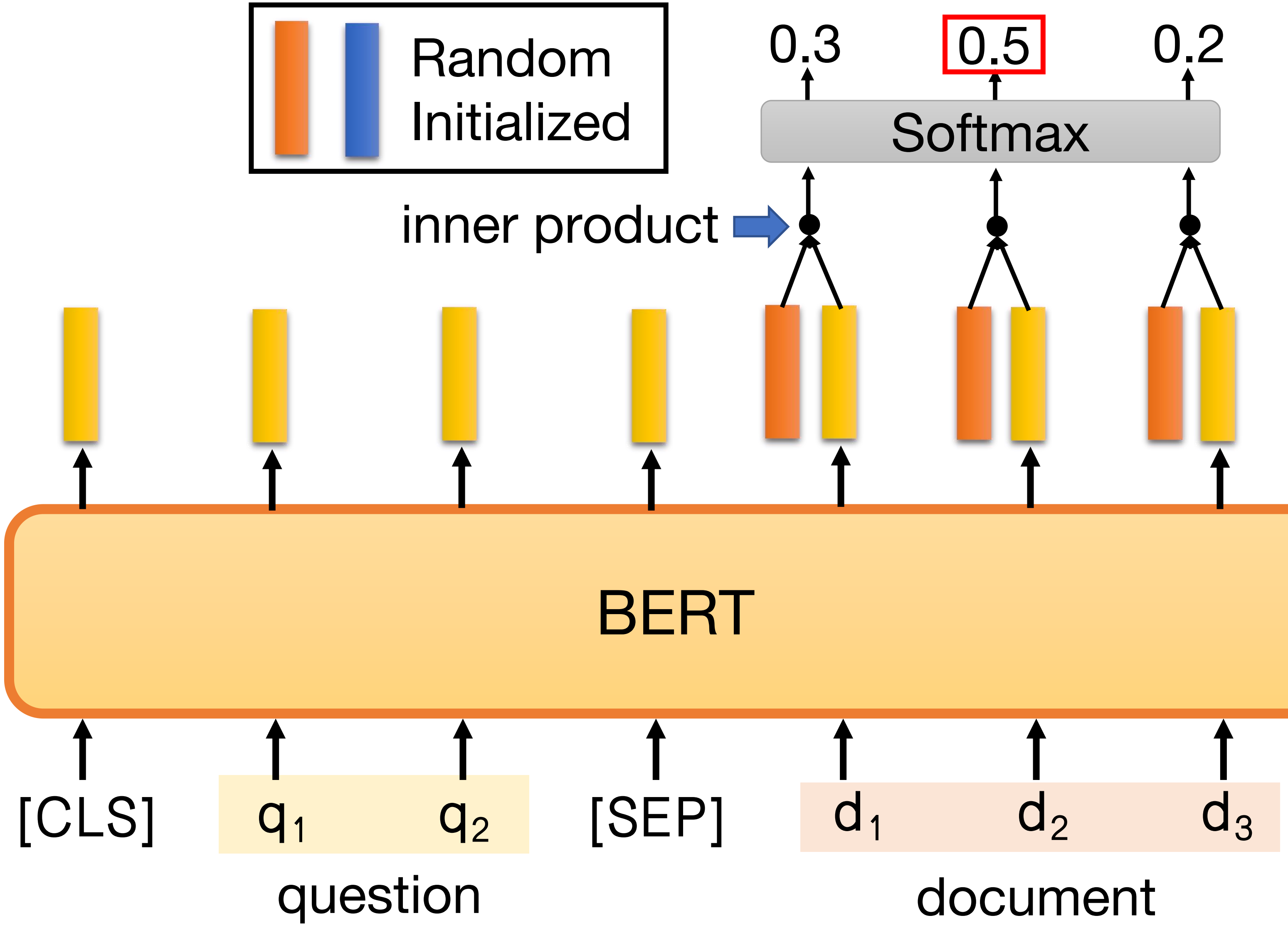
Where do water droplets collide with ice crystals to form precipitation?

within a cloud $s = 77, e = 79$

Bert 使用

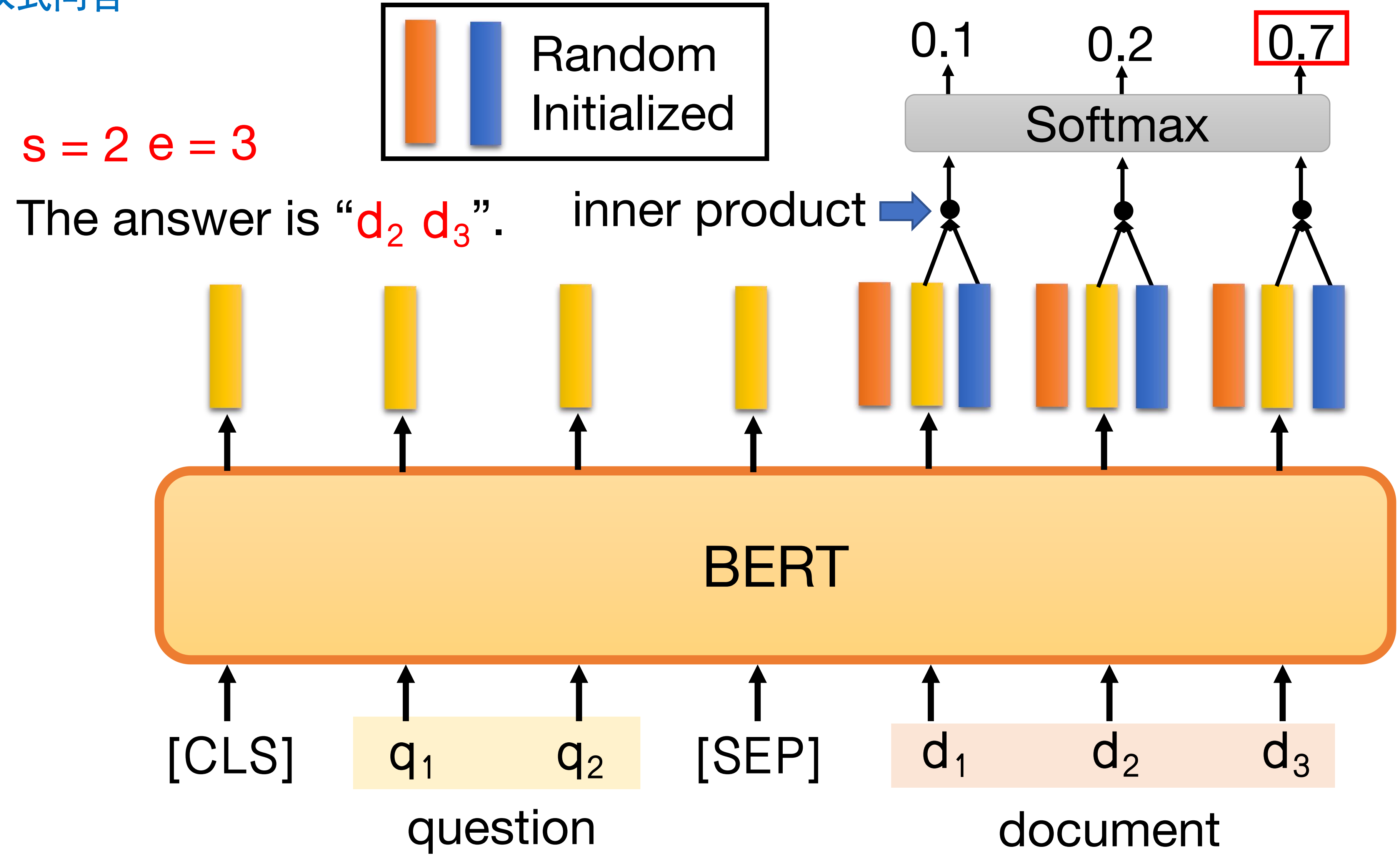
应用场景 4 抽取式问答

s = 2



Bert 使用

应用场景 4 抽取式问答



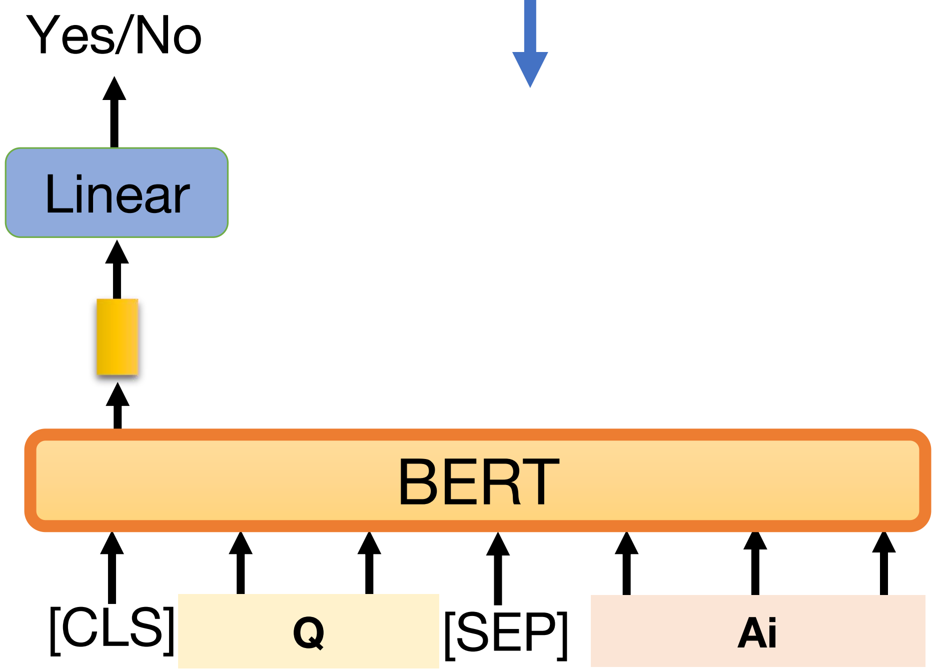
Bert 使用

应用场景 4 抽取式问答

1. 【Q】 【A1】 【label】
2. 【Q】 【A2】 【label】
3. 【Q】 【A3】 【label】
4. 【Q】 【A4】 【label】
5. 【Q】 【A5】 【label】

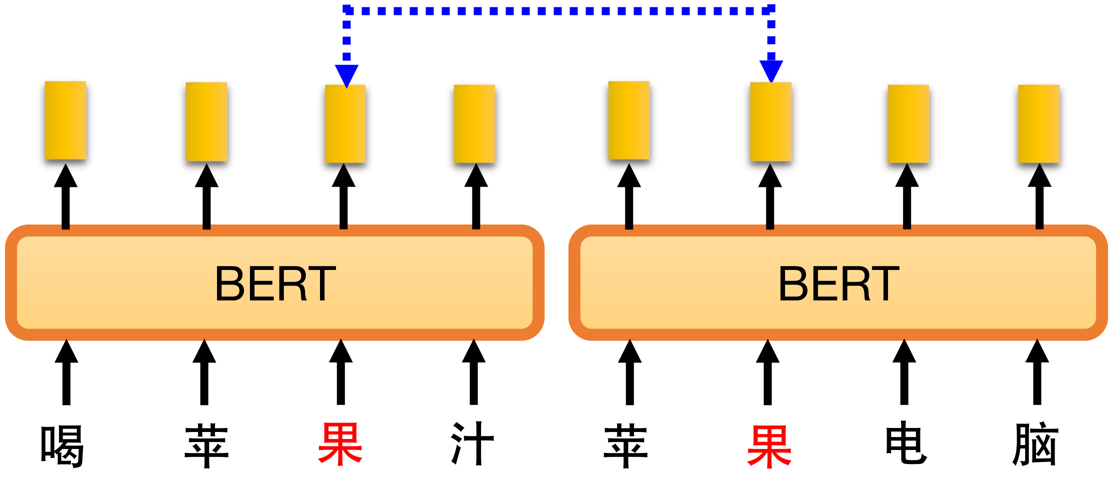
北京奥运会是哪年举办?

第29届夏季奥林匹克运动会（Games of the xxix olympiad），又称2008年北京奥运会，2008年8月8日晚上8时整在中华人民共和国首都北京举办。【1】
2008年北京奥运会主办城市是北京，上海、天津、沈阳、秦皇岛、青岛为协办城市。【2】
香港承办马术项目。【3】
2008年北京奥运会共有参赛国家及地区204个，参赛运动员11438人，设302项（28种）运动，共有60000多名运动员、教练员和官员参。加【4】
2008年北京奥运会共创造43项新世界纪录及132项新奥运纪录，共有87个国家和地区在赛事中取得奖牌，中国以51枚金牌居金牌榜首名，是奥运历史上首个登上金牌榜首的亚洲国家。【5】

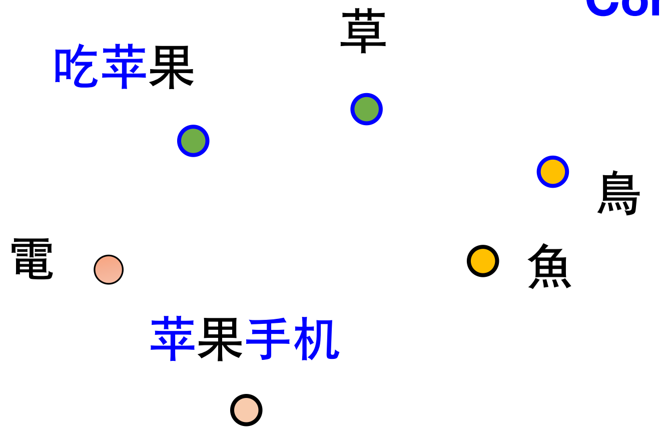


BERT 表征可视化

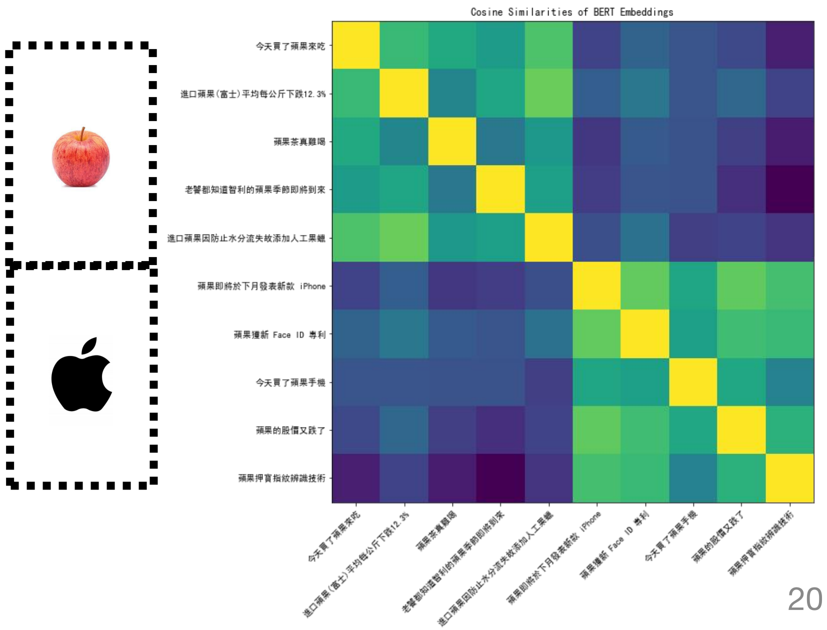
compute cosine similarity



The tokens with similar meaning have similar embedding.



Context is considered.



总结

1. **预训练的有效性**： BERT 改变了游戏规则，是因为相比设计复杂巧妙的网络结构，在海量无监督数据上预训练得到的BERT语言表示+少量训练数据微调的简单网络模型的实验结果取得了很大的优势。

2. **网络深度**：基于 传统语言模型 (NNLM，CBOW等) 获取词向量的表示已经在 NLP领域获得很大成功，而 BERT 预训练网络基于 Transformer 的 Encoder，可以做得很深。

3. **双向语言模型**：在 BERT 之前，ELMo 和 GPT 的主要局限在于标准语言模型是单向的，GPT 使用 Transformer 的 Decoder 结构，只考虑了上文的信息。ELMo 从左往右的语言模型和从右往左的语言模型其实是分开训练的，共享 embedding，将两个方向的 LSTM 拼接并不能真正表示上下文，其本质仍是单向的，且多层 LSTM难训练。

4. **目标函数**：对比以往语言模型任务只做预测下一个位置的单词，想要训练包含更多信息的语言模型，就需要让语言模型完成更复杂的任务，BERT 主要完成完形填空和句对预测的任务，即两个 loss：一个是 Masked Language Model，另一个是 Next Sentence Prediction。

参考资料

- BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (ailab-ua.github.io)
- The Illustrated BERT, ELMo, and co. (How NLP Cracked Transfer Learning) – Jay Alammar – Visualizing machine learning one concept at a time. (jalammar.github.io)
- Hung-yi Lee (ntu.edu.tw)
- WordEmbedding发展史（语言模型演变史） - 知乎 (zhihu.com)
- BERT详解（附带ELMo、GPT介绍）_bert算法 数学家是我的理想_数学家是我理想的博客-CSDN博客
- BERT模型详解 - 李理的博客 (fancyerii.github.io)
- BERT论文的解读 PPT_bert介绍ppt_SimonChenHere的博客-CSDN博客
- 一张图看懂BERT - 知乎 (zhihu.com)
- NLP——Bert核心内容 - 知乎 (zhihu.com)
- 关于Cbow，Transformer，Elmo，GPT，Bert - 知乎 (zhihu.com)
- BERT详解：概念、原理与应用__StarryNight_的博客-CSDN博客
- LeeMeng - 進擊的 BERT：NLP 界的巨人之力與遷移學習