北京理工大學
BEIJING INSTITUTE OF TECHNOLOGY

语言预训练模型与应用

常欣煜、徐子懋、赵华耀、禹言江、王子轩、王雪飞

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

目录

CONTENTS

1 引子

2 模型

3 应用

4 Demo

5 前沿进展

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

1

引子

徐子懋

自然语言处理（NLP）是什么

在以前我们做研究，研究的数据都是表格里的数据，比如“关于各个气象数据和天气之间的关系”，我们可以把天气进行编码/处置。然后进行数据分析。

自然语言处理（Natural Language Processing）研究的是对人类所说的语言进行处理研究的问题。让计算机理解自然语言文本的含义，让其理解你想表达的深层意图、思想等。

常见的NLP问题

中文分词

词云画像

词性分析

自动摘要

关系挖掘

情感分析

知识图谱

聊天机器人（小冰等）

机器翻译（谷歌翻译、百度翻译等）

语音识别（小德小德、siri）

……

NLP问题的难点

如何把我们的话让机器听懂？

- 如何分词
- 一词多义

在以前我们做研究，研究的数据都是表格里的数据，比如“关于各个气象数据和天气之间的关系”，我们可以把天气进行编码/处置。然后进行数据分析。

自然语言处理（Natural Language Processing）研究的是对人类所说的语言进行处理研究的问题。让计算机理解自然语言文本的含义，让其理解你想表达的深层意图、思想等。

迁移学习

迁移学习是一种机器学习方法，就是把为一个任务开发的模型作为初始点，重新使用在为另一个任务开发模型的过程中。

在深度学习中常用于计算机视觉任务和自然语言处理任务，可以节省时间资源和算力资源。

语言预训练模型

语言预训练模型是用于自然语言处理（NLP）任务中的模型，我们想让系统“读懂”我们的语言，就要先通过预训练，转化为系统能看的明白的格式。

比如word2vec，就把单词转化为一个个的向量，通过向量在后续任务中继续训练。

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

2

模型

ELMo、ENRIE：赵华耀

BERT、GPTs：禹言江

ELMo算法介绍

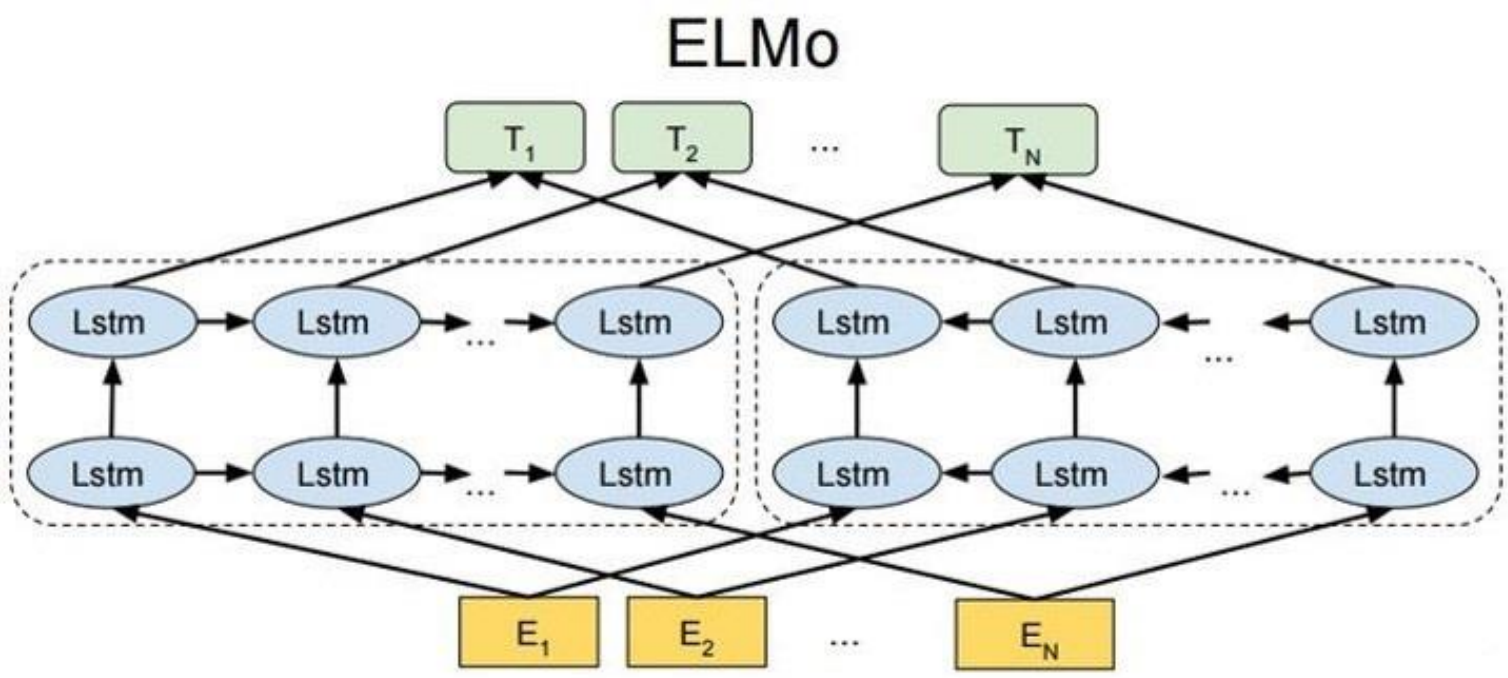
Embeddings from Language Model

2018年NAACL上的Best Paper

对比传统Word2Vec这种形式的词向量，本文提出的模型是一种动态模型。在以往的词向量表示中，词都是一种静态的形式，无论在任何的上下文中都使用同一个向量。这种情况下很难表示一词多义的现象，而ELMo则可以通过上下文动态生成词向量，从理论上会是更好的模型，从实测效果来看在很多任务上也都达到了当时的SOTA成绩。

ELMo是一种是基于特征的语言模型，用预训练好的语言模型，生成更好的特征。

ELMo算法介绍



使用的是一个**双向的LSTM语言模型**，由一个前向和一个后向语言模型构成

目标函数就是取这两个方向语言模型的最大似然。

ELMo算法介绍

优势所在

假设词向量不固定

ELMo的假设前提一个词的词向量不应该是固定的，所以在一词多意方面ELMo的效果一定比word2vec要好。

基于整个语料库学习

ELMo在学习语言模型的时候是从整个语料库去学习的，而后再通过语言模型生成的词向量就相当于基于整个语料库学习的词向量，更加准确代表一个词的意思。

ELMo算法介绍

缺点

特征抽取器

首先，一个非常明显的缺点在特征抽取器选择方面，ELMO 使用了 LSTM 而不是新贵 Transformer，很多研究已经证明了 Transformer 提取特征的能力是要远强于 LSTM 的。

双向拼接的能力

另外一点，ELMO 采取双向拼接这种融合特征的能力可能比 Bert 一体化的融合特征方式弱，但是，这只是一种从道理推断产生的怀疑，目前并没有具体实验说明这一点。

ENRIE-清华算法介绍

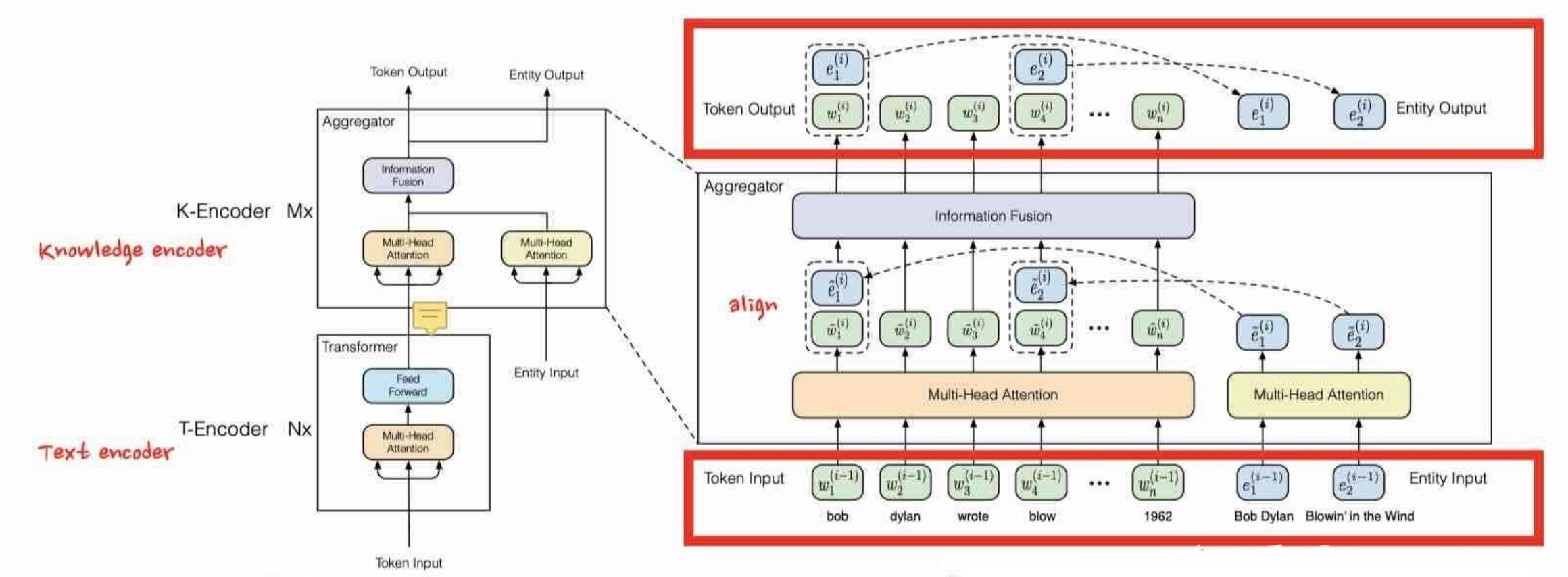
ERNIE:

Enhanced Language Representation with Informative Entities

——清华

提出将知识显性地加入到BERT中

ENRIE-清华算法介绍



模型架构

ERNIE 1.0

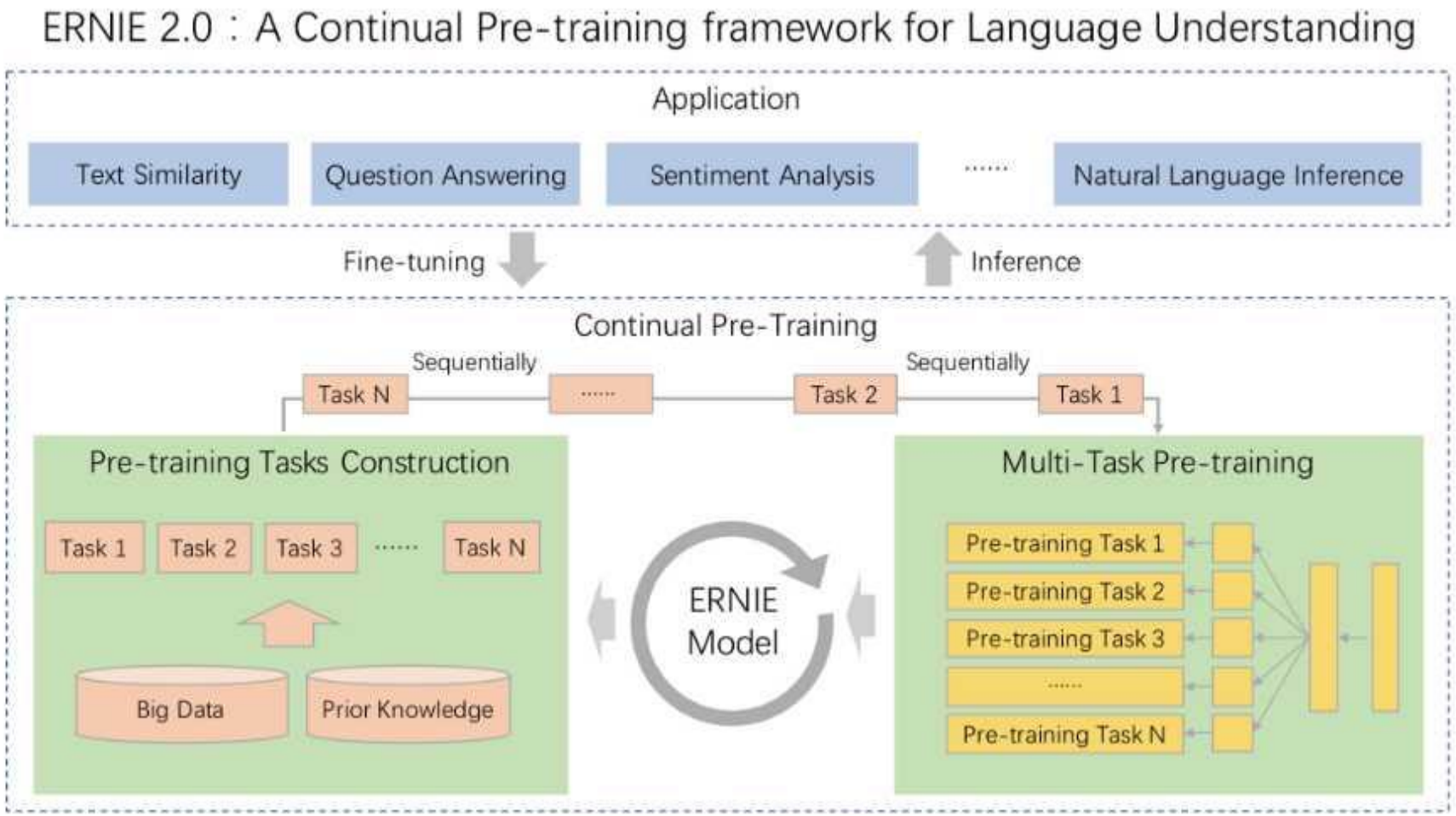
Enhanced Representation through Knowledge Integration

——百度在2019年4月的时候，基于BERT模型，做的进一步优化

缺陷及优点

优点：善于捕获词语之间相互关系，在完型填空等类型的任务中的表现良好。

ENRIE-百度算法介绍



ENRIE-百度算法介绍

百度ERNIE 2.0提出了 **持续预训练** 。通过持续预训练，模型能够持续地学习各类任务，从而使得模型的效果达到了进一步提升。

百度称持续预训练的过程包含两个步骤：首先，需要不断构建无监督的预训练任务，具有大的语料库和/或先验知识。其次，通过多任务学习逐步更新ERNIE模型。

百度提出可持续学习语义理解框架 ERNIE 2.0，支持增量引入词汇（ lexical ）、语法（ syntactic ）、语义（ semantic ）等3个层次的自定义预训练任务，能够全面捕捉训练语料中的词法、语法、语义等潜在信息。

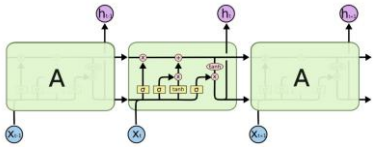
ERNIE2.0总结

ERNIE 2.0 创新地将过去单一的预训练流程拆解为串行的多个预训练任务，无疑是其最大的贡献。如何通过多任务的形式将更多的语法信息有效地融入到模型的自编码中，相信会成为未来新的研究方向。

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

BERT 模型参数

LSTM
(0.08M)



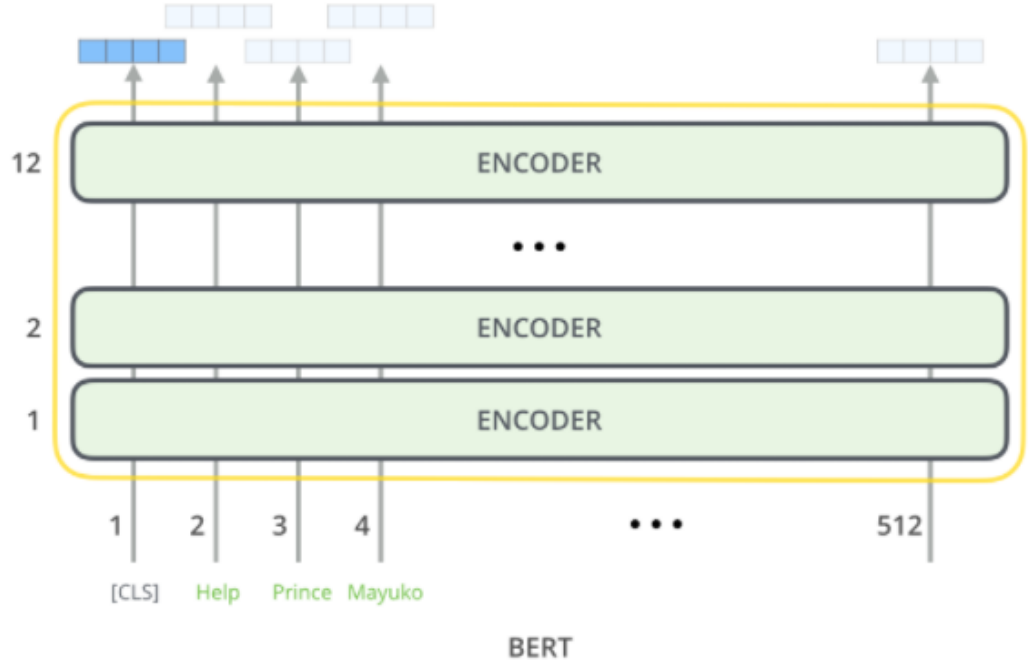
ELMO
(94M)

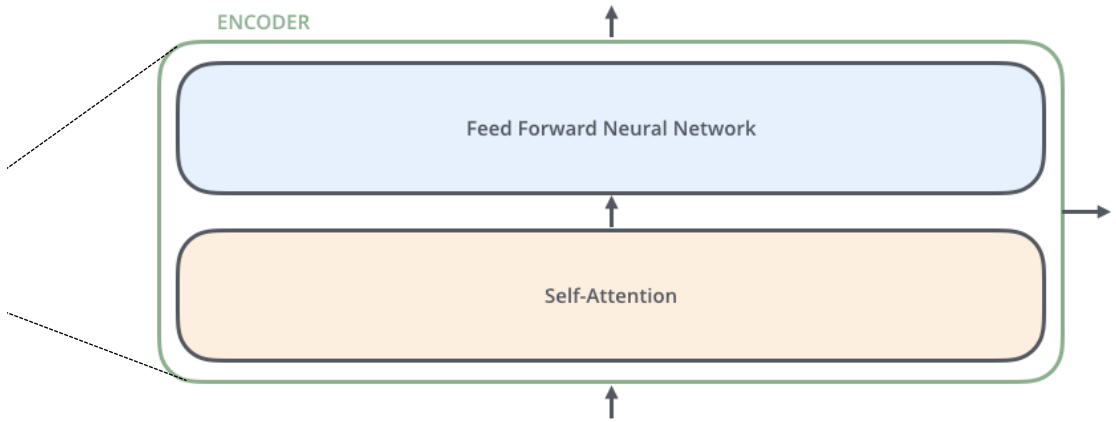
BERT-
base
(101M)

BERT
(340M)

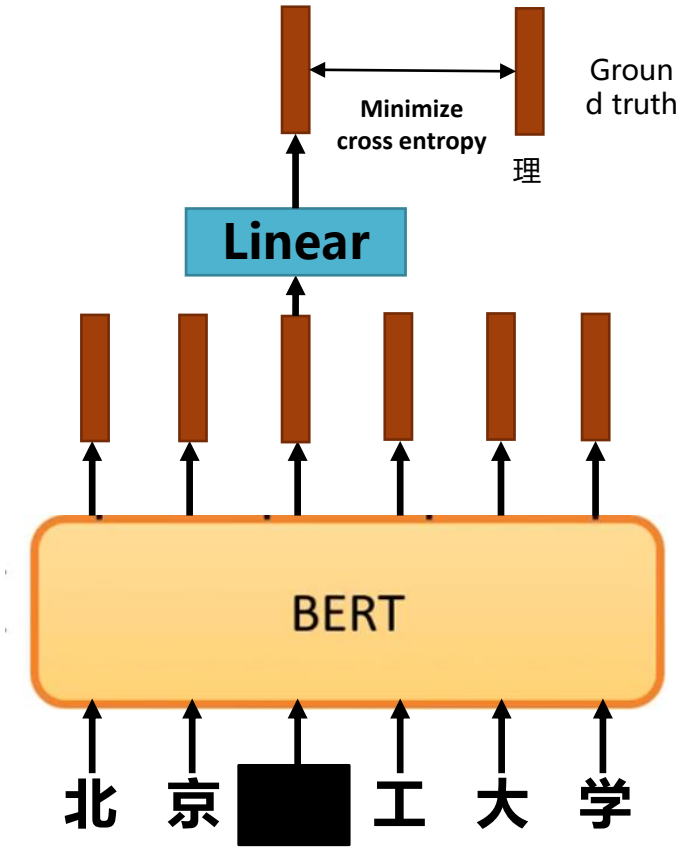
bert: https://arxiv.org/pdf/1810.04805.pdf

BERT Network Architecture





BERT

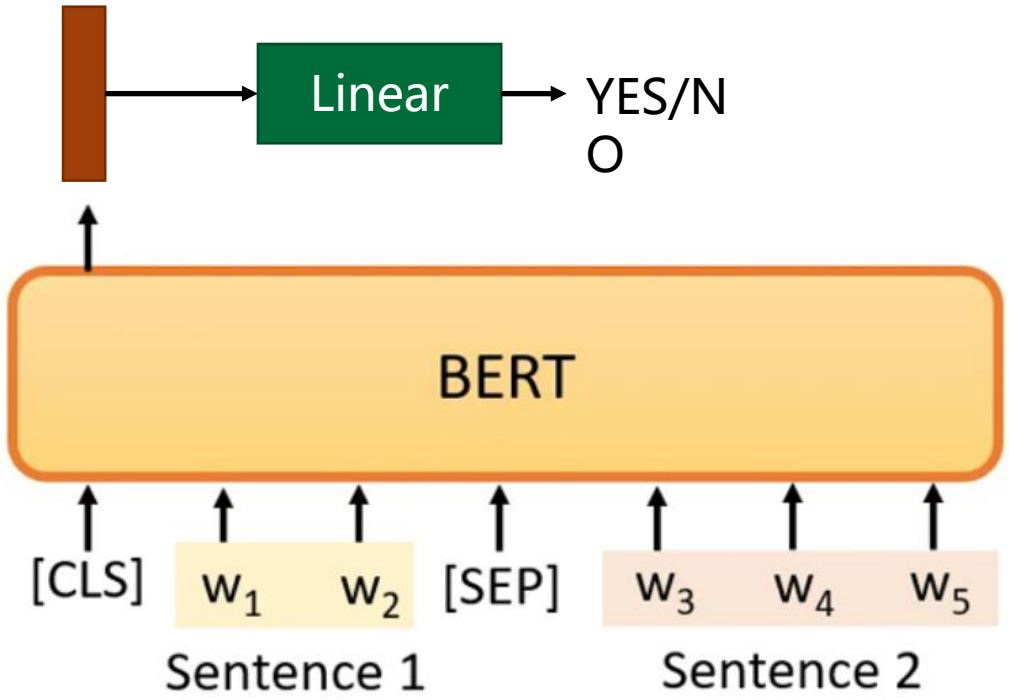


Masked Language Model(MLM)

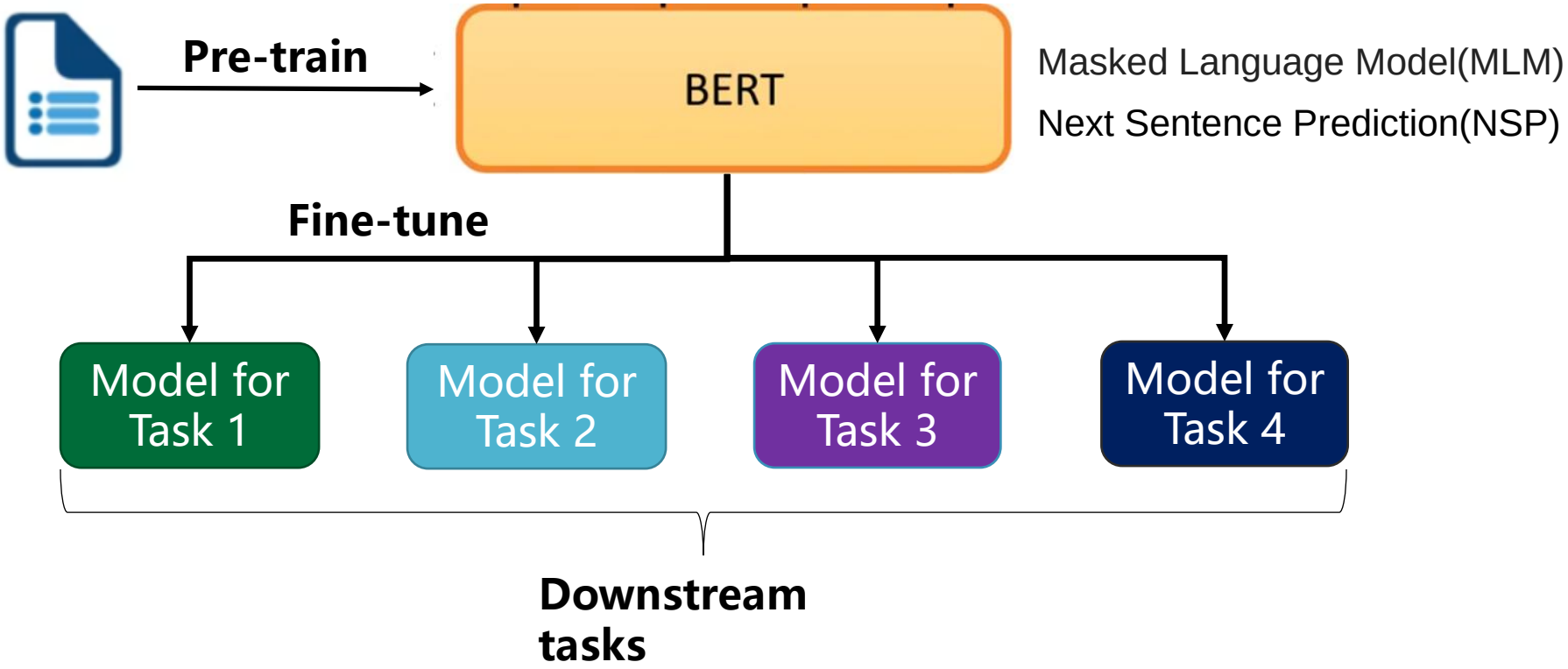
= MASK or Random word

BERT Pre-train

Next Sentence Prediction(NSP)

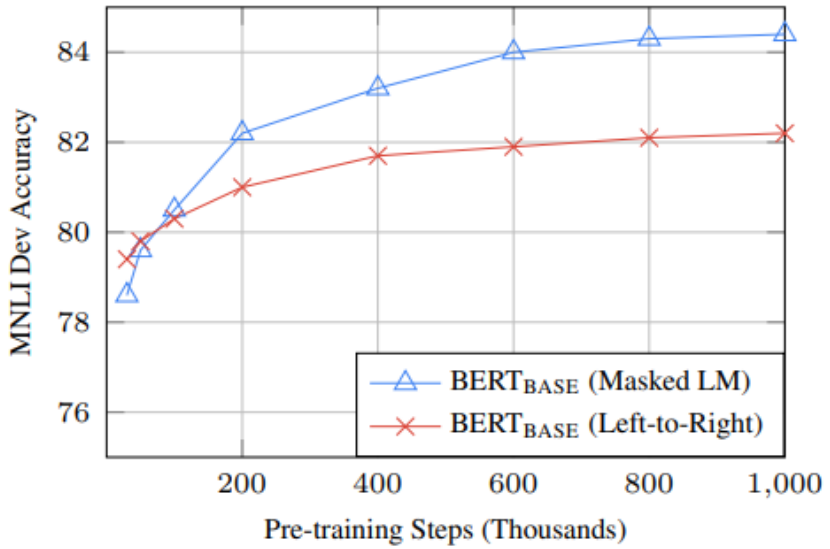


BERT Fine-tune



Pre-training bert is challenging

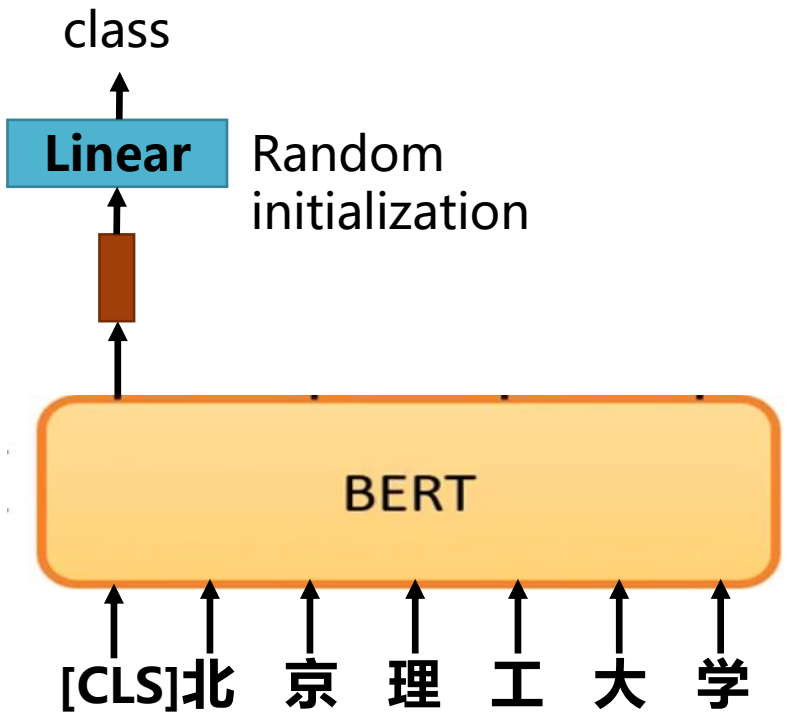
- Pre-training corpus include 3300M words
- 4 days on 16 TPUs（64 TPU chips total)



BERT downstream tasks

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

Sentiment analysis

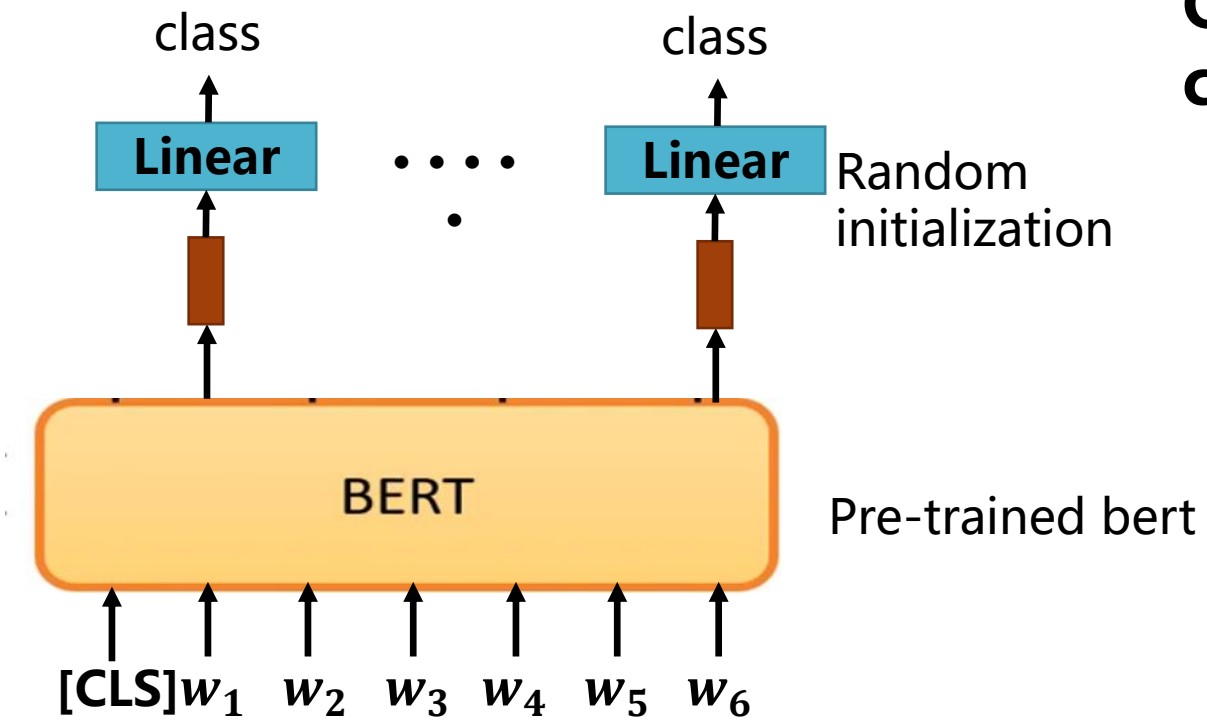


Input: sequence
Output: class

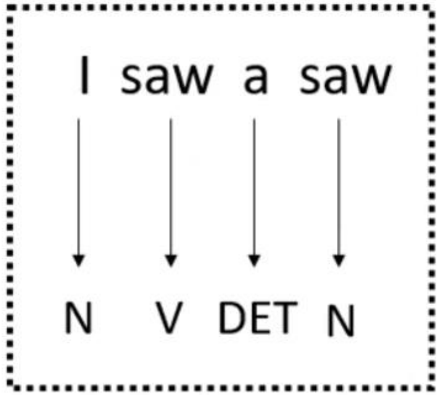
BERT downstream tasks

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

POS tagging



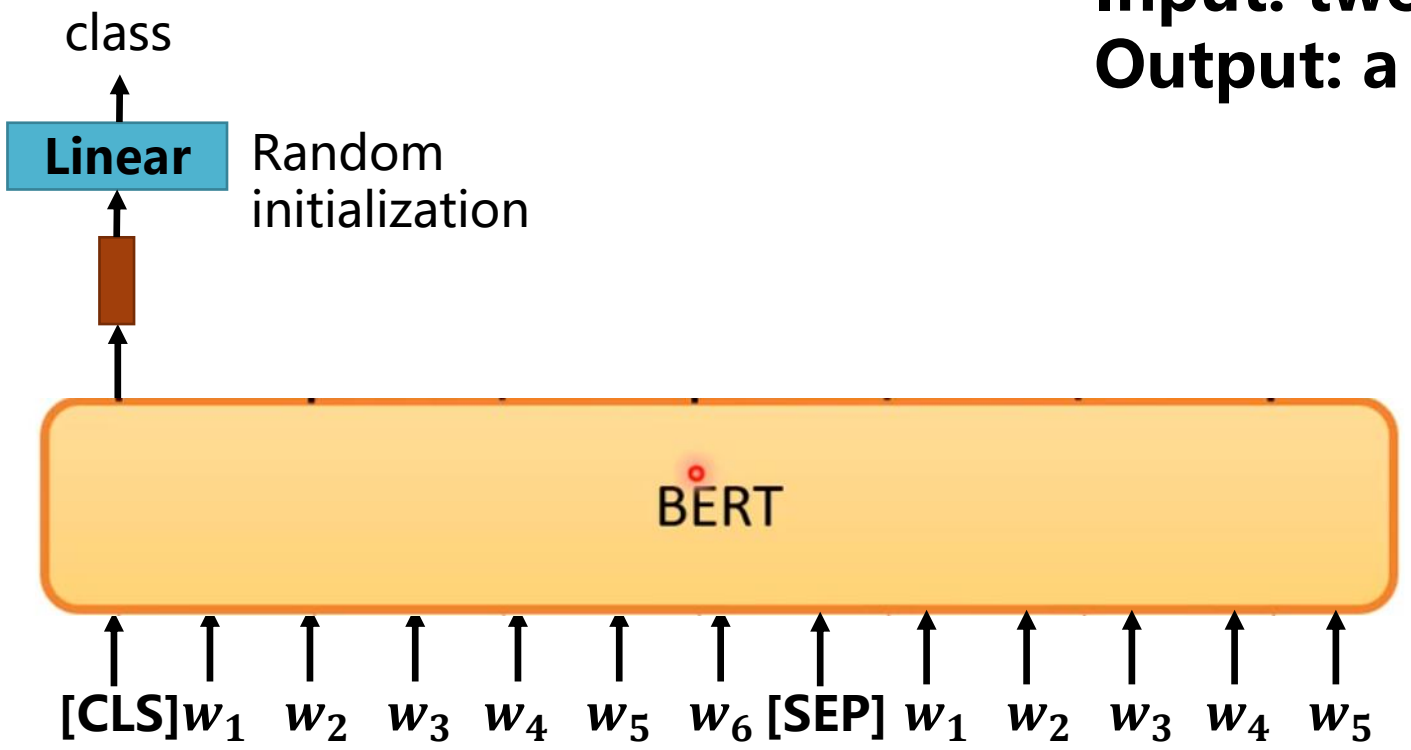
Input: sequence
Output: sequence of class



BERT downstream tasks

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

Natural language inference



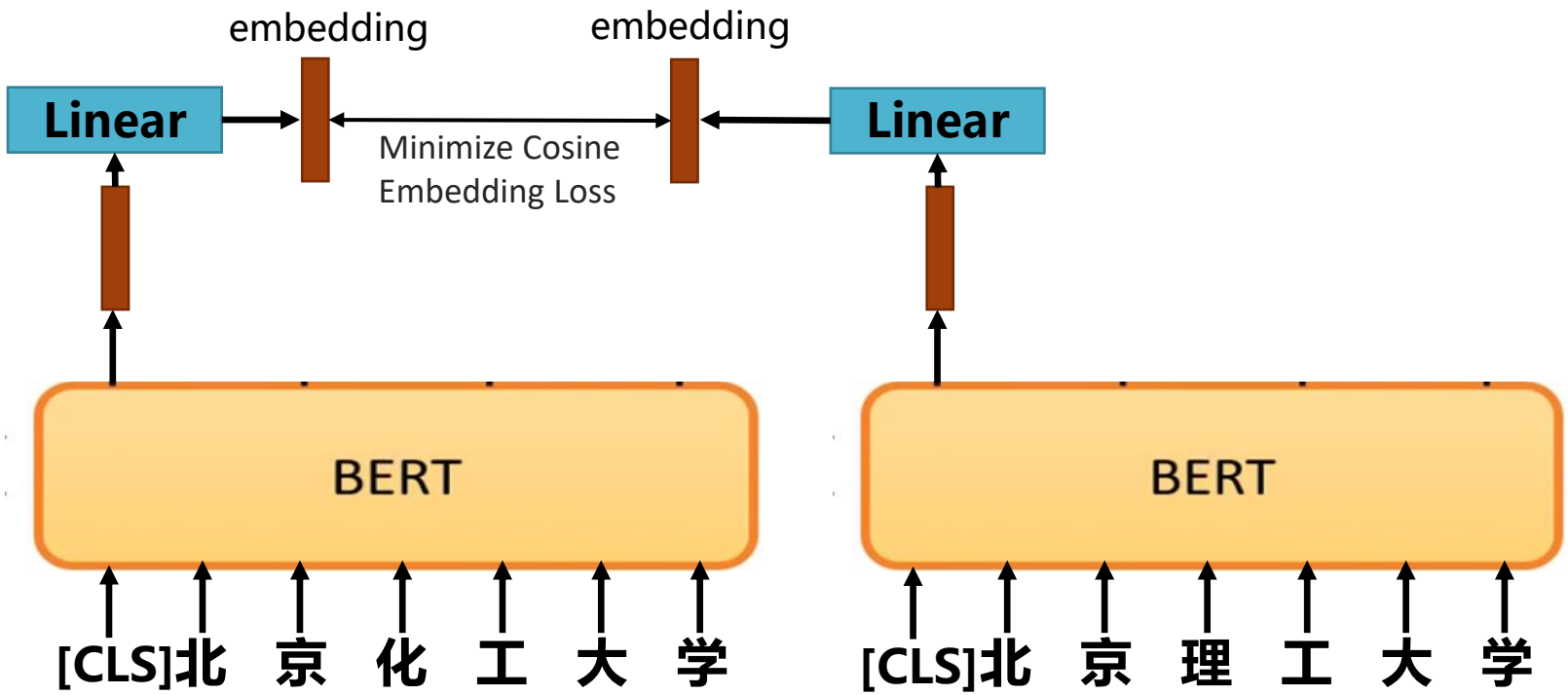
Input: two sequences
Output: a class

Pre-trained bert

BERT down stream tasks

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

Sentence embedding



GPT 模型参数

BERT
(340M)

GPT2.0
(1542M)

Megatron
(8B)

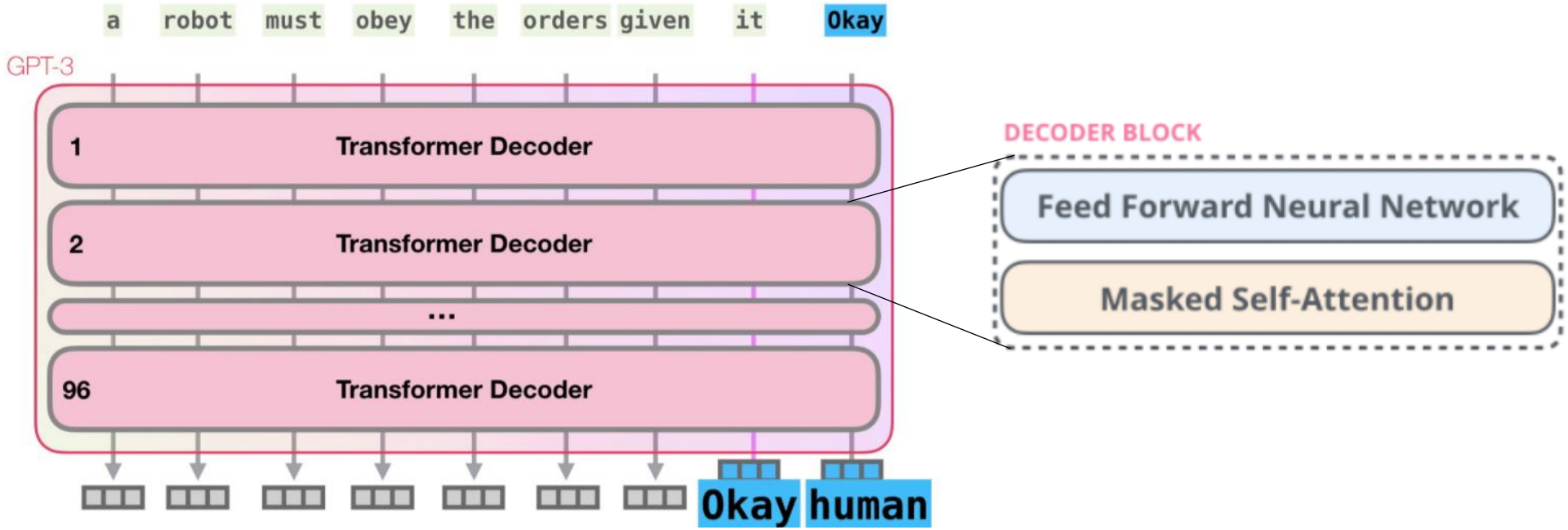
GPT 模型参数

GPT3.0
(175B)

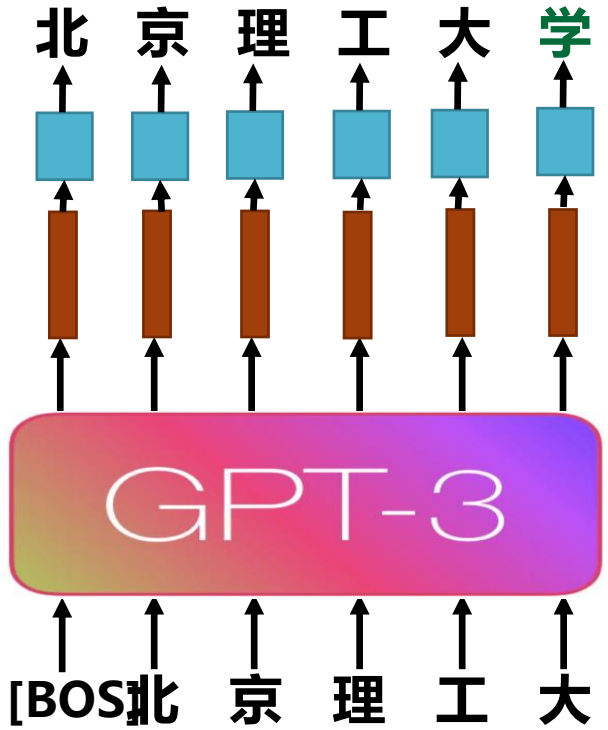
T5
(11B)

Megatron
(8B)

GPT 3.0 Network Architecture



GPT 3.0 Pre-train



Predict Next Token

- Pre-training corpus include 45TB words
- 335 GPU years
- Cost 4.6 million dollars

GPT and "Few-shot" Learning

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

"Few-shot" Learning

```
Translate English to French:        ← task description
sea otter => loutre de mer          ← examples
peppermint => menthe poivrée
plush girafe => girafe peluche
cheese =>                           ← prompt
```

NO gradient descent

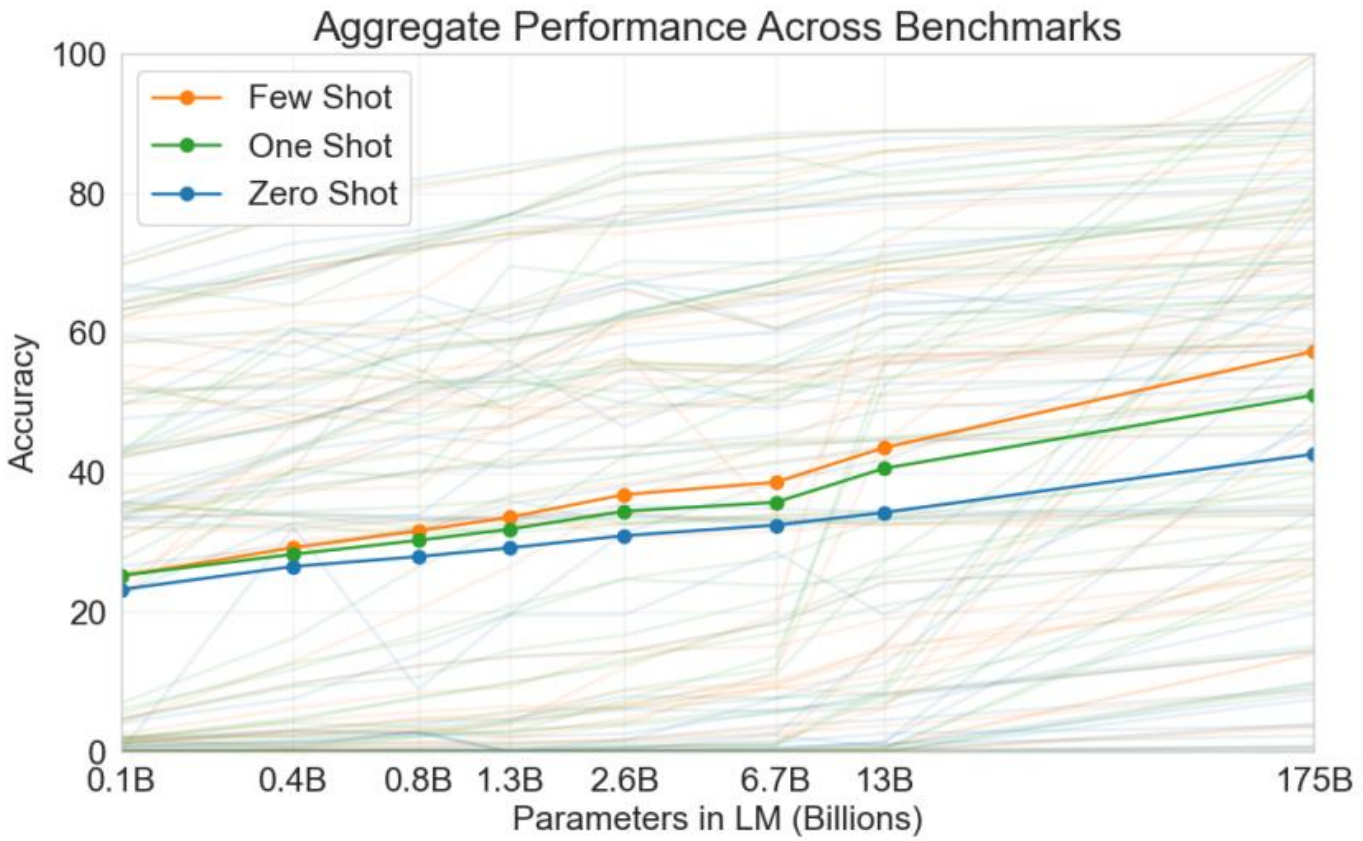
"one-shot" Learning

```
Translate English to French:        ← task description
sea otter => loutre de mer          ← example
cheese =>                           ← prompt
```

"zero-shot" Learning

```
Translate English to French:        ← task description
cheese =>                           ← prompt
```

GPT Performance



Source: https://arxiv.org/pdf/2005.14165.pdf

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

3

应用

王子轩

Application of Pre-trained Model

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

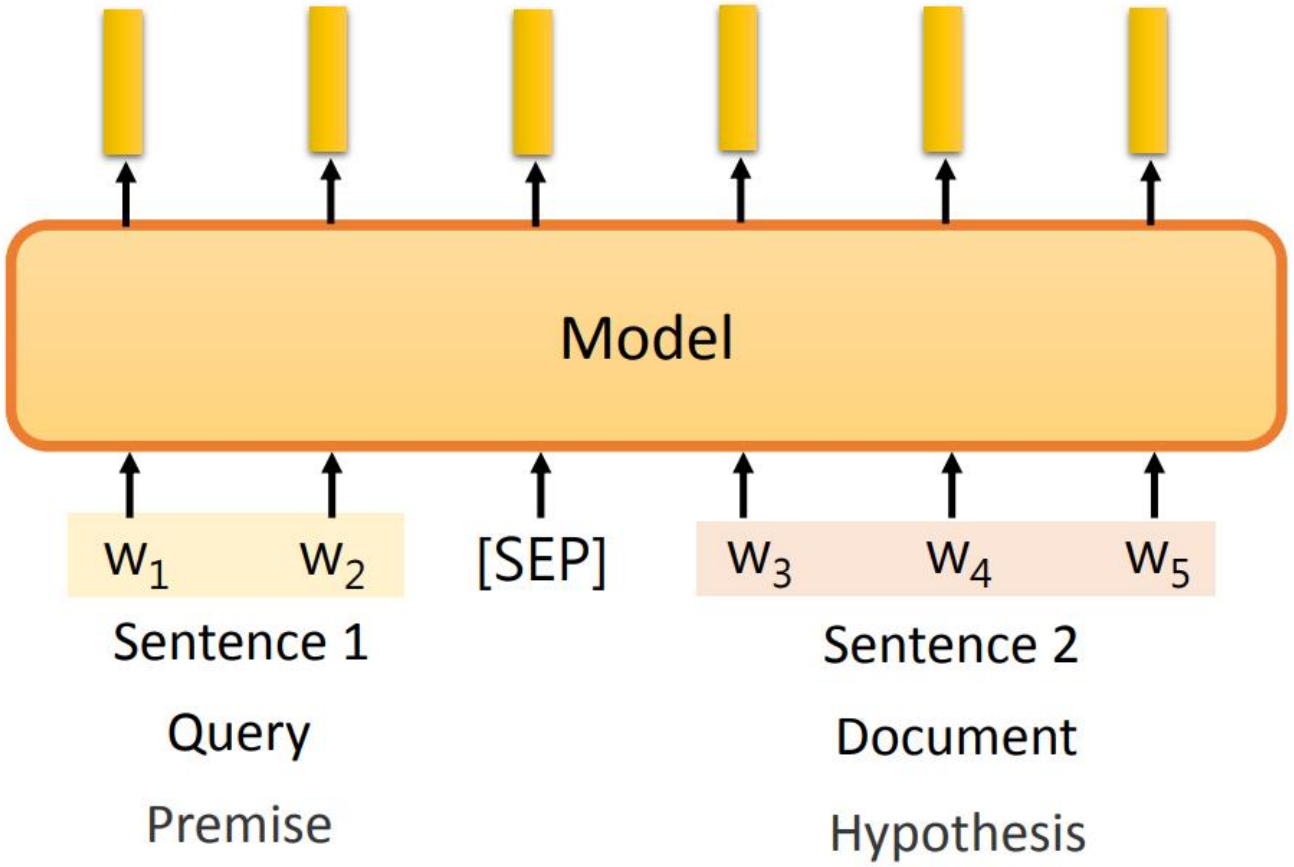
How can the pre-trained model be applied to **NLP tasks**? **Why**?

NLP Tasks

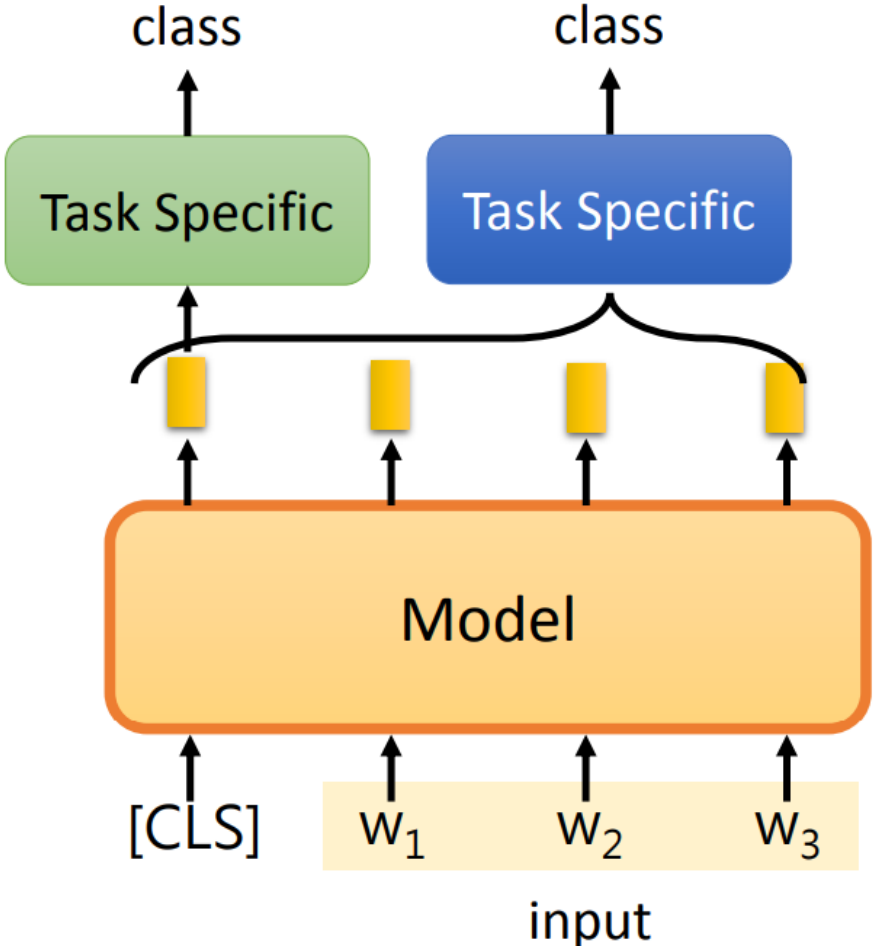
Input
- one sentence
- multiple sentences

Output
- one class
- class for each token
- copy from input
- general sequence

Input


Model
w_1
w_2
[SEP]
w_3
w_4
w_5
Sentence 1
Query
Premise
Sentence 2
Document
Hypothesis

Output-One Class

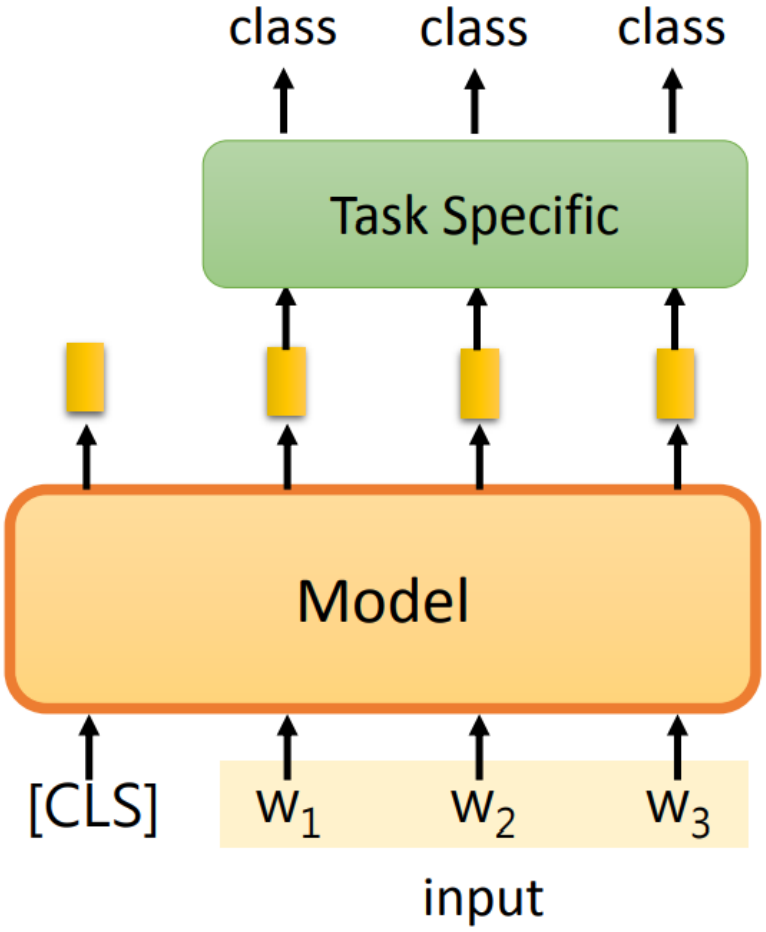


Output-One Class

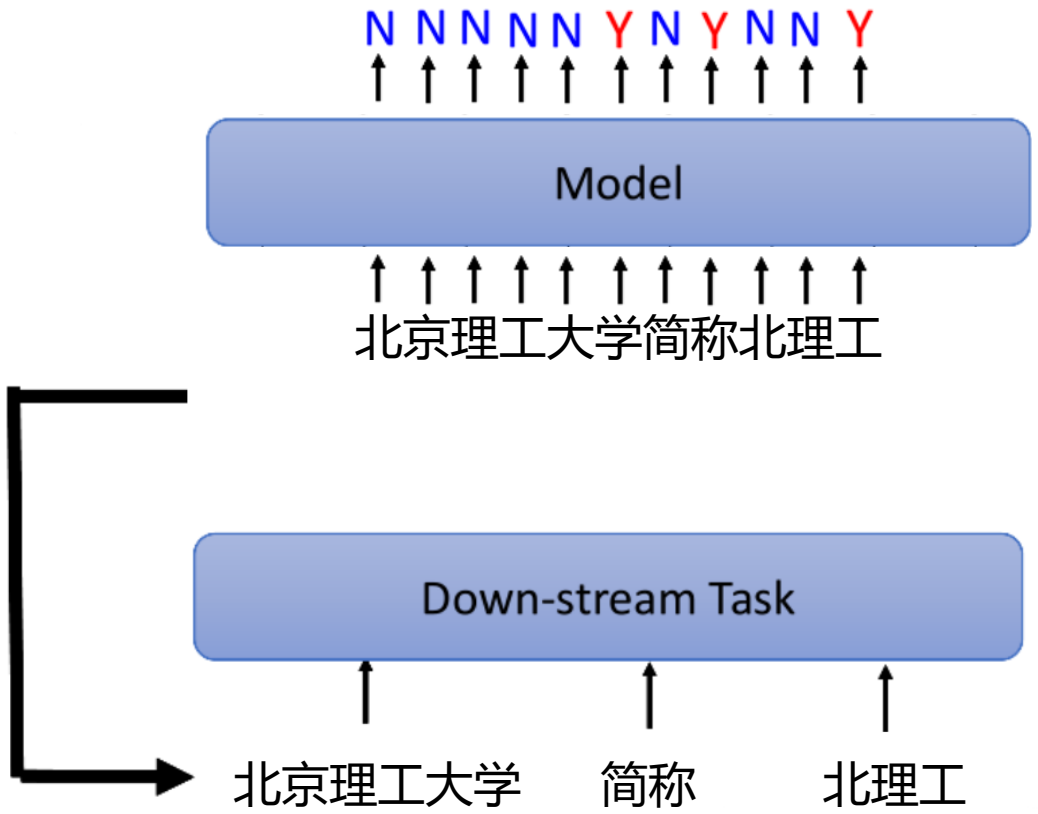
柯南劇場版《紺青之拳》還蠻有趣的

柯南劇場版《紺青之拳》槽點很多

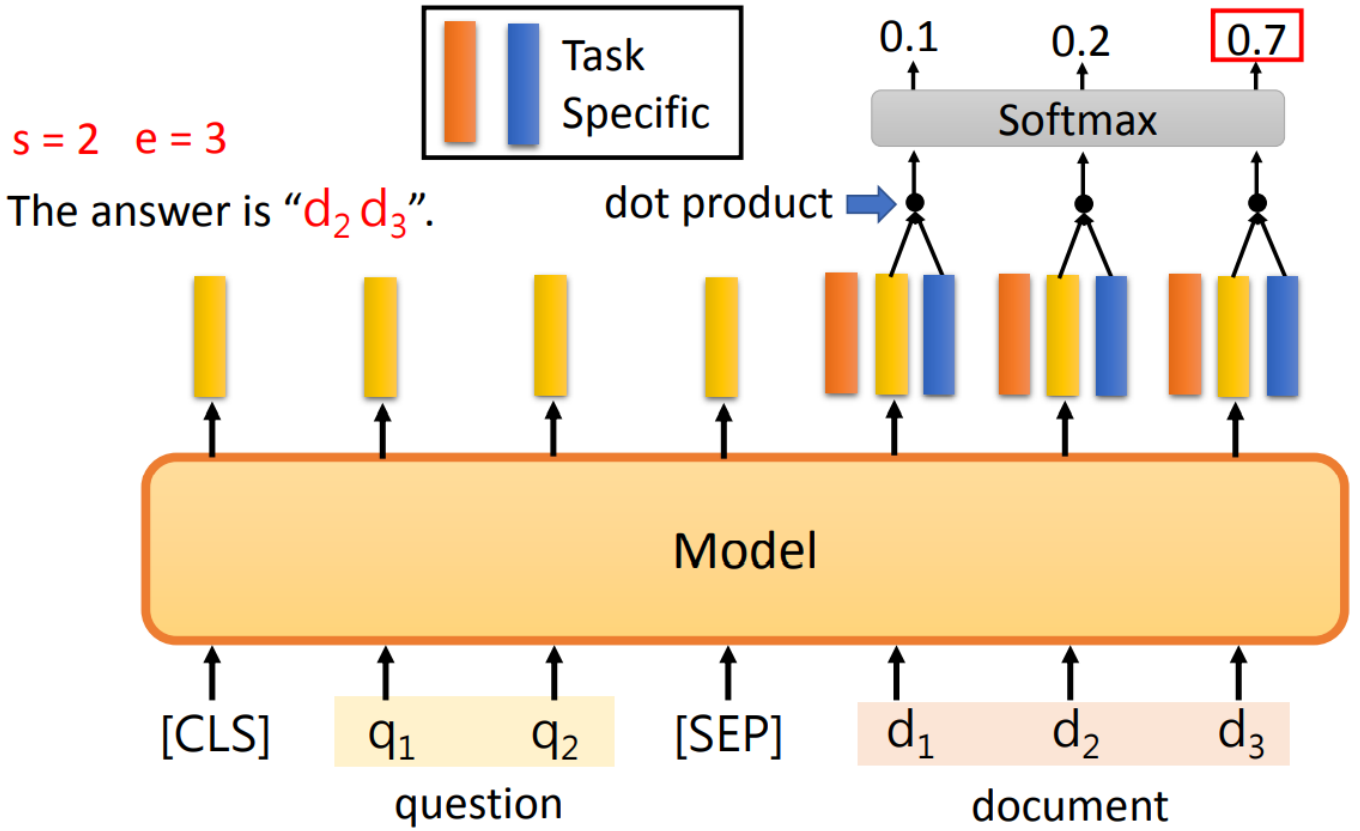
Output-Class for Each Token



Output-Class for Each Token



Output-Copy from Input



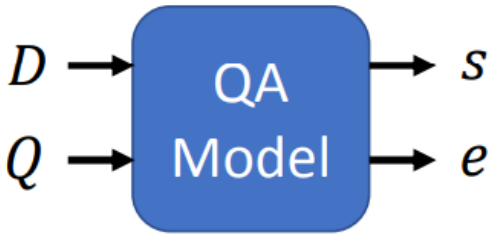
Output-Copy from Input

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

Extraction-based QA

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... Preci 77 on 79 as smaller droplets coalesce via ion other rain drops or ice crystals within a cloud. Short, intense periods of rain in scattered locations are called "showers".

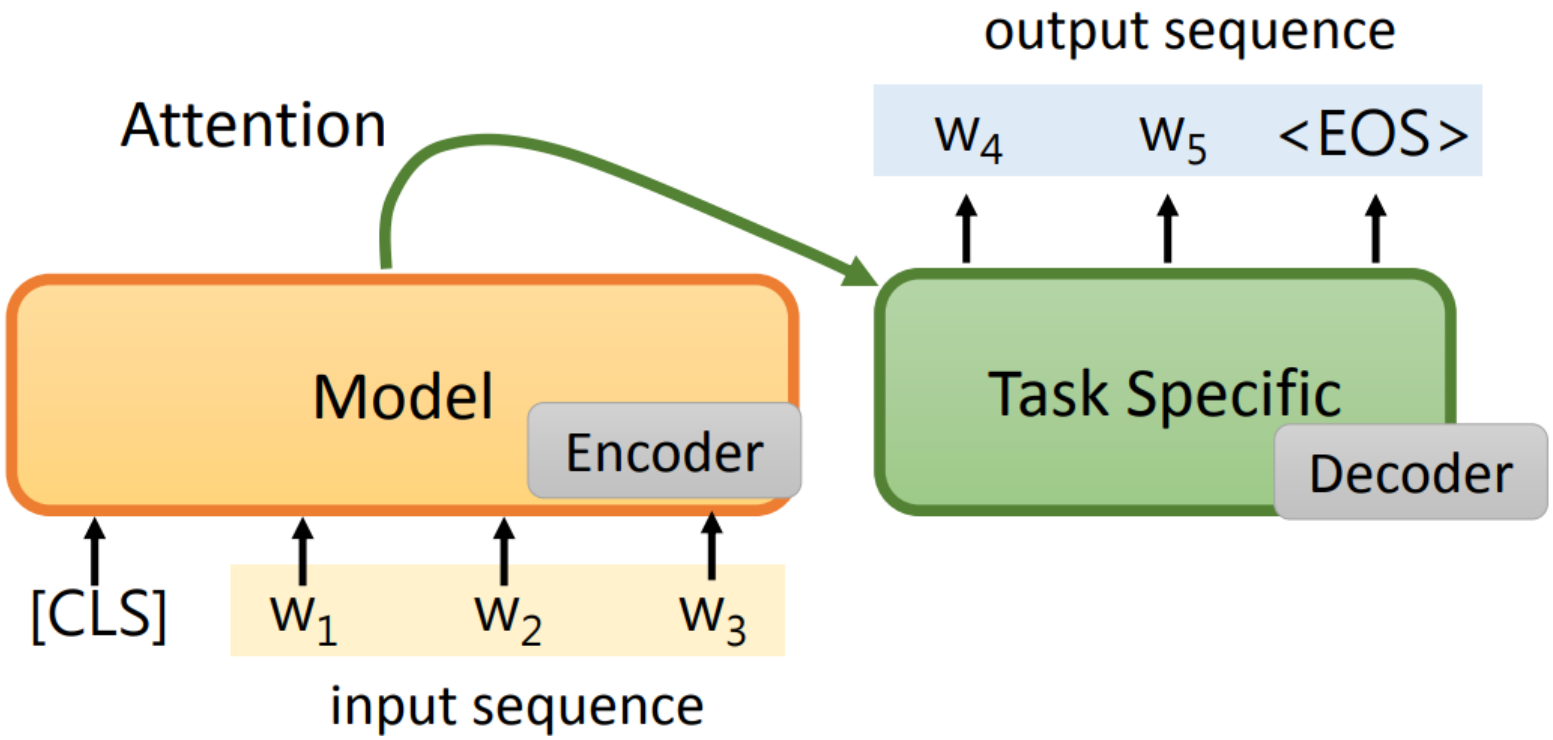
What causes precipitation to fall?
gravity

Where do water droplets collide with ice crystals to form precipitation?
within a cloud $s = 77, e = 79$

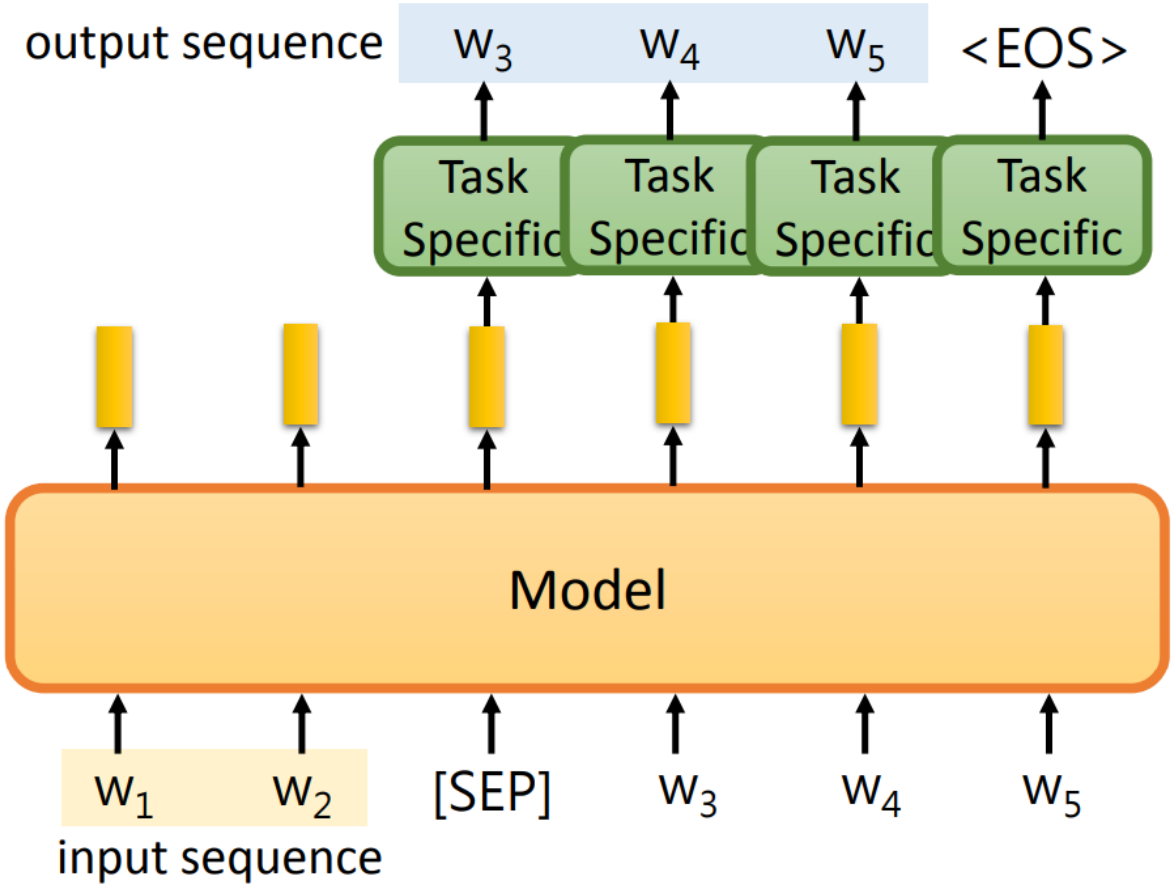
Output-General Sequence

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

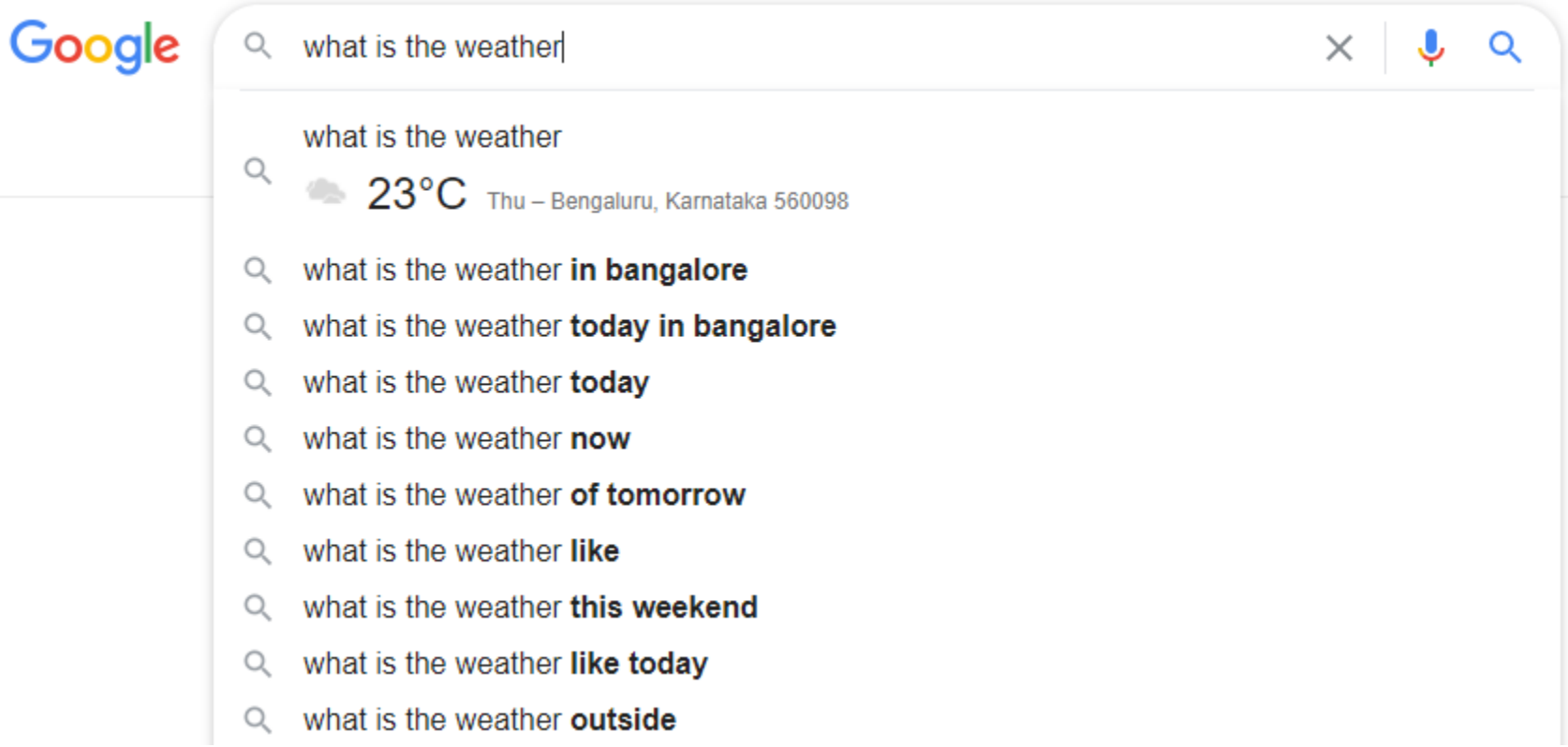
Seq2seq Model



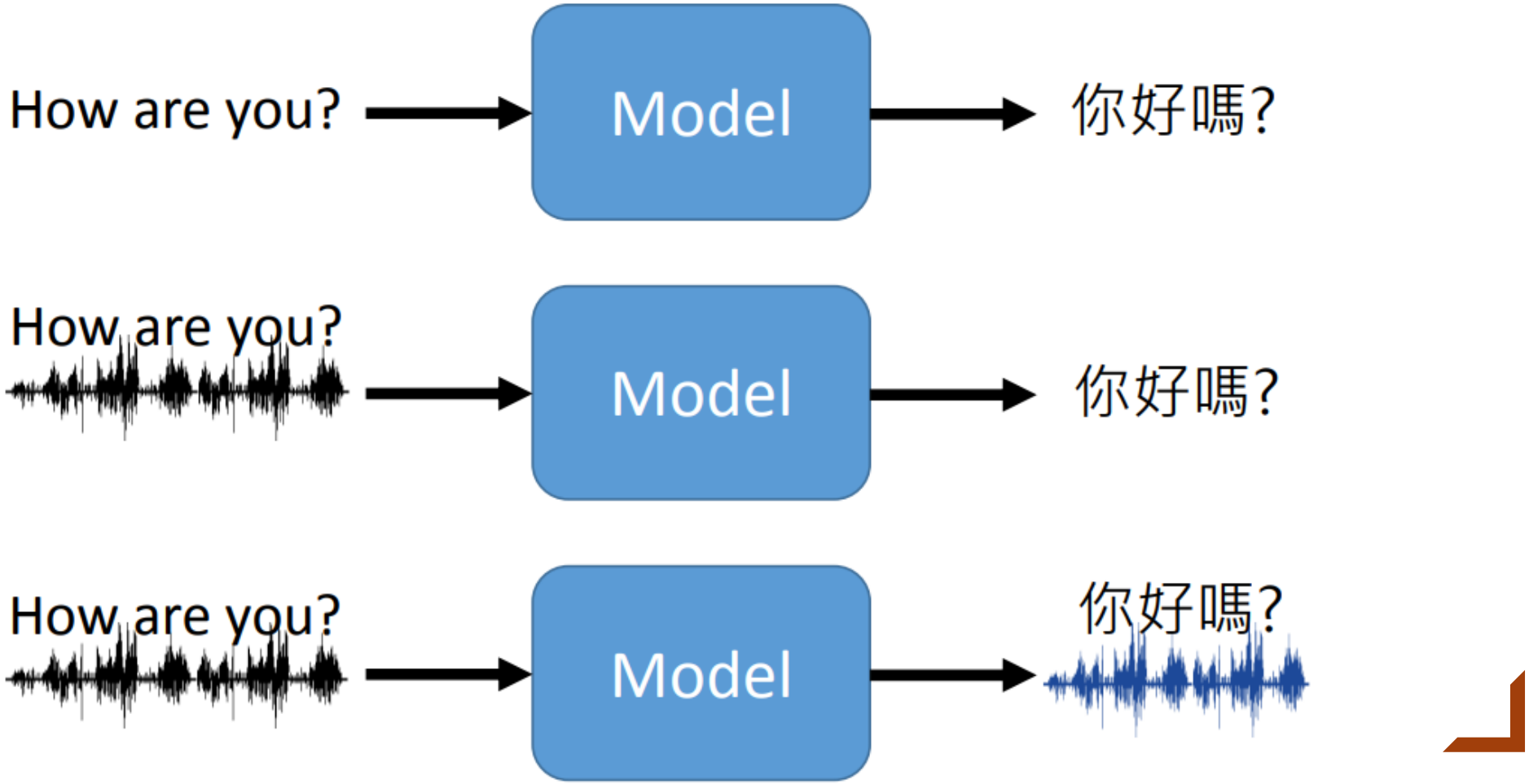
Output-General Sequence



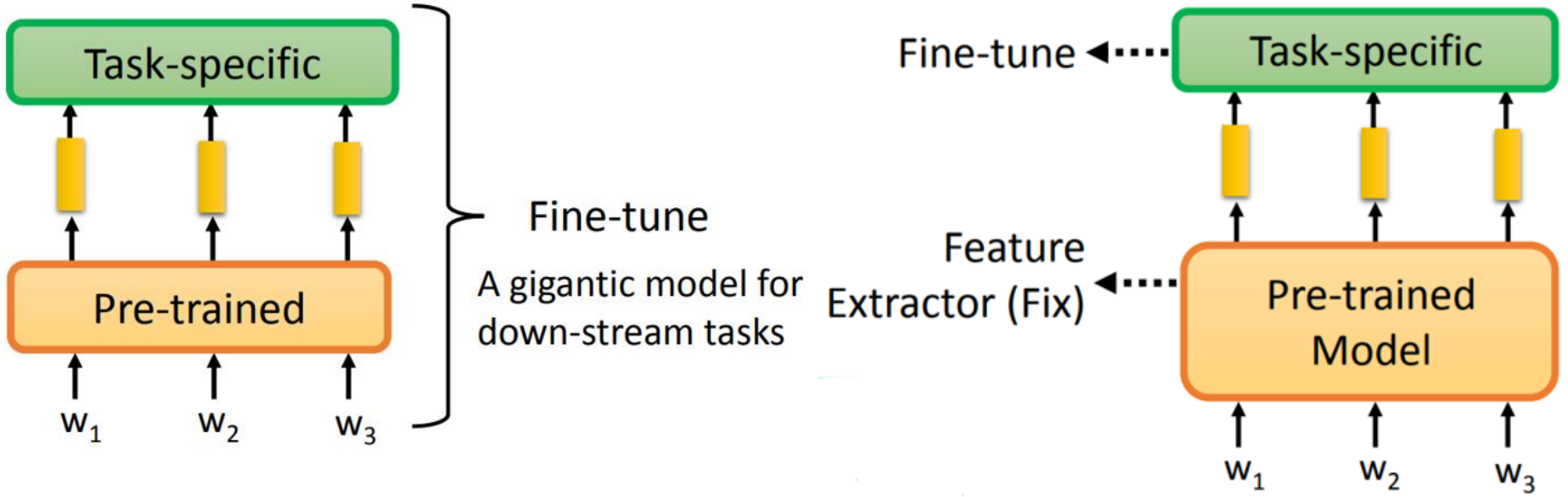
Output-General Sequence



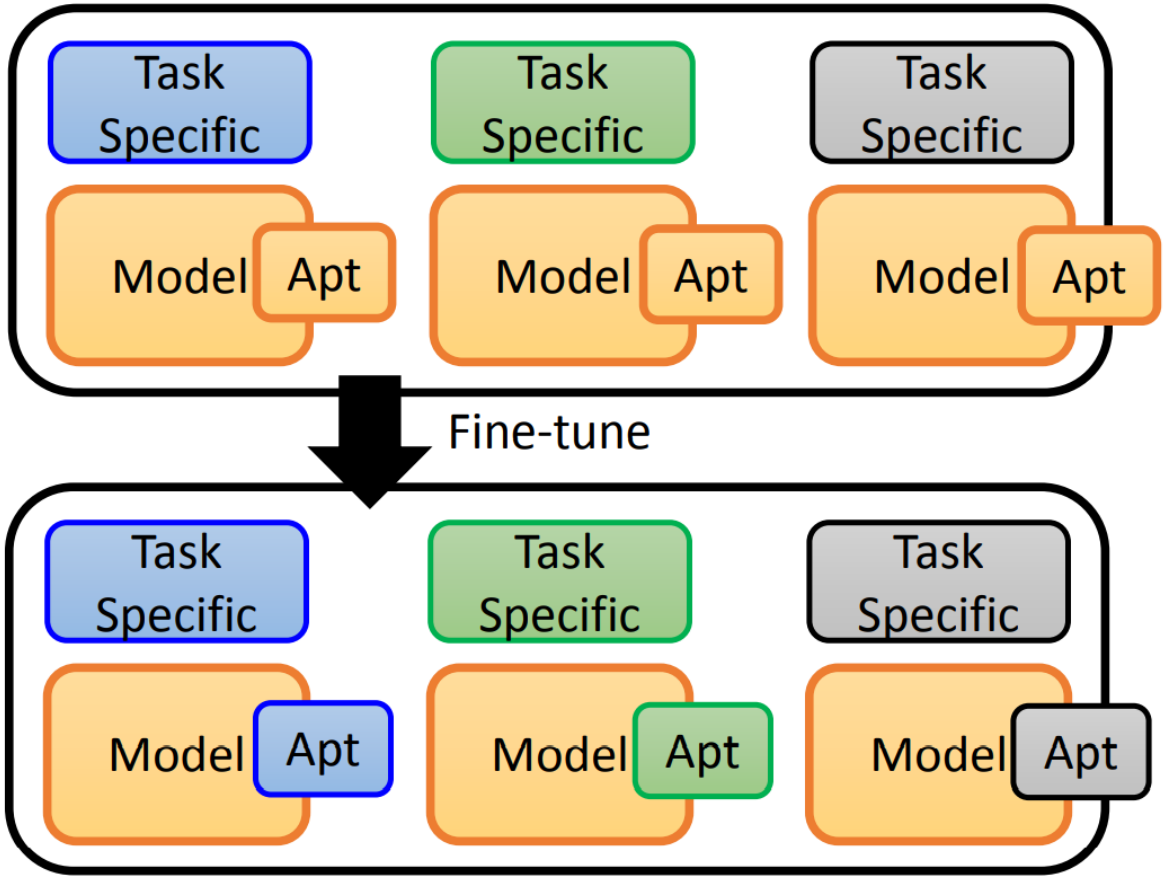
Output-General Sequence



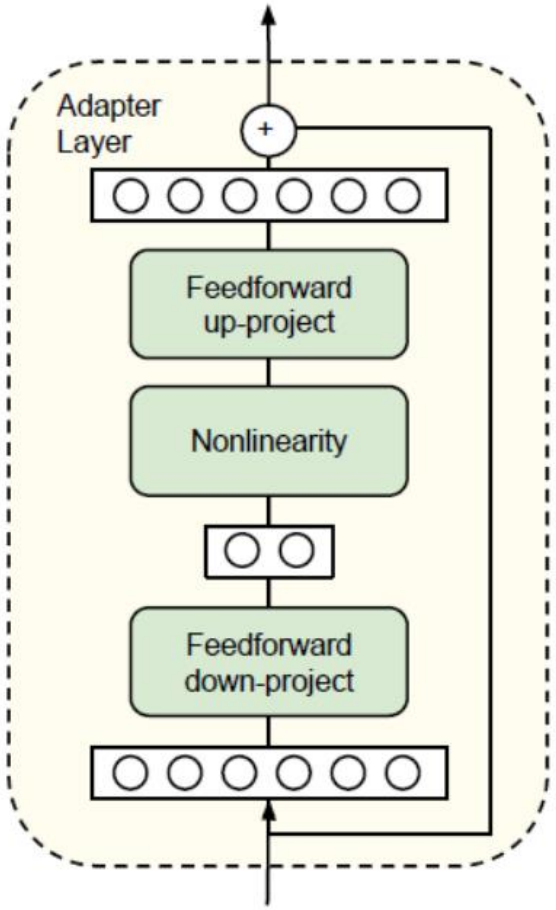
Fine-tune

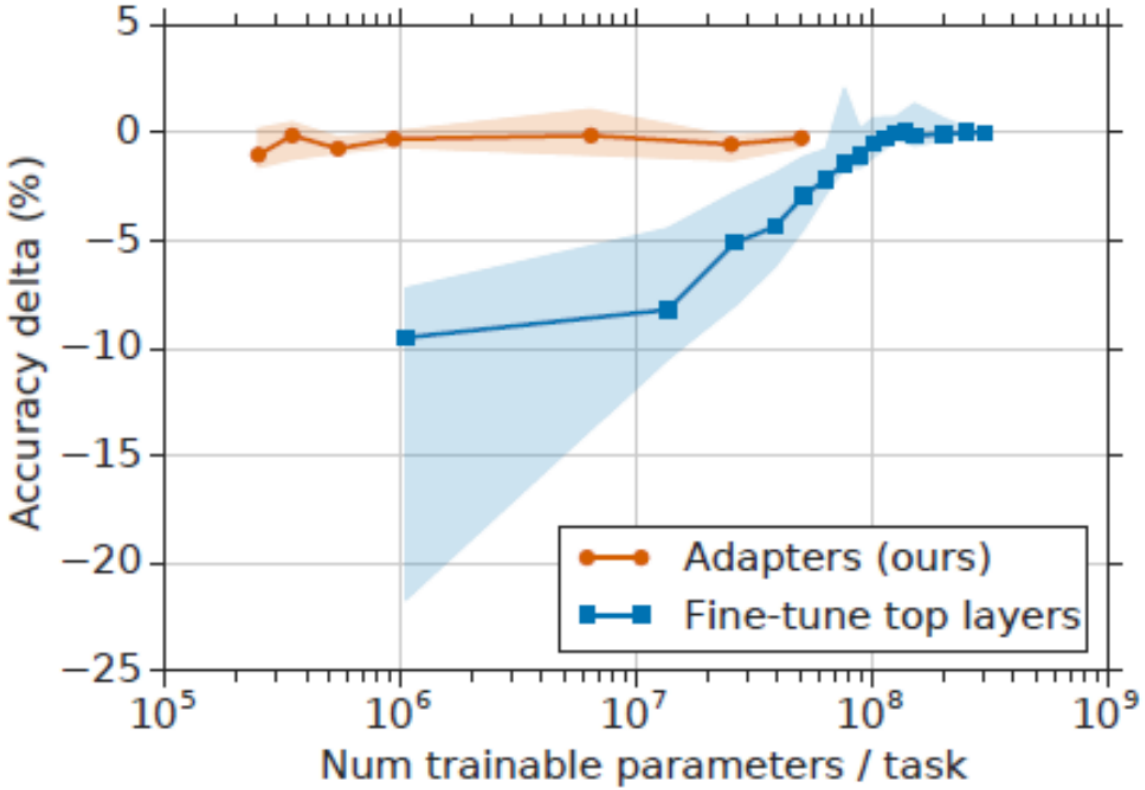

Task-specific
Pre-trained
w_1
w_2
w_3
Fine-tune
A gigantic model for down-stream tasks
Fine-tune
Task-specific
Feature Extractor (Fix)
Pre-trained Model
w_1
w_2
w_3

Fine-tune


Task Specific
Task Specific
Task Specific
Model
Apt
Model
Apt
Model
Apt
Fine-tune
Task Specific
Task Specific
Task Specific
Model
Apt
Model
Apt
Model
Apt

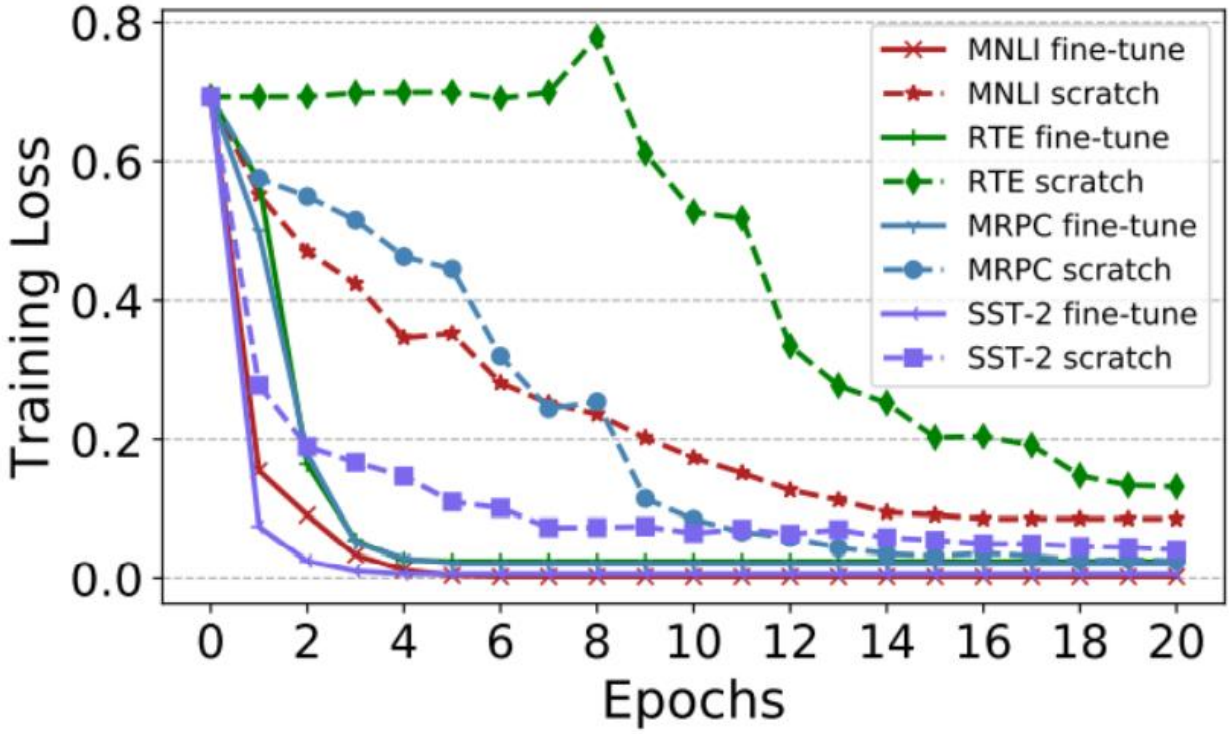
Fine-tune





Source of image: https://arxiv.org/abs/1902.00751

Why Fine-tune?



[Hao, et al., EMNLP'19] Source of image: https://arxiv.org/abs/1908.05620

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

4

Demo

王雪飞

CPM-LM

模型论文

CPM: A Large-scale Generative Chinese Pre-trained Language Model

清源 CPM 介绍

清源 CPM(Chinese Pretrained Models)是北京智源人工智能研究院和清华大学研究团队合作开展的，以中文为核心的大规模预训练模型。

特点

1、GPT-3含有175B参数使用了570GB的数据进行训练。但大多数语料是基于英文（93%），并且GPT-3的参数没有分布，所以提出了CPM（Chinese Pretrained language Model）：包含2.6B参数，使用100GB中文训练数据。

2、CPM可以对接下游任务：对话、文章生成、完形填空、语言理解。

Demo

3、随着参数规模的增加，CPM在一些数据集上表现更好，表示大模型在语言生成和理解上面更有效。

4、在分词语料库的基础上构建了一种新的子词词汇来适应汉语语料库，并将批处理量增加到3072个，以获得更稳定的模型训练，且在小样本甚至0样本中得到较好的性能

可以采用官方的推理工具BMInf，也可以自己下载模型，定制使用方式

Demo

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

实现方法

CPM的模型架构与GPT的架构类似（从左到右的Transformer encoder）

·BERT-Chinese(2019)中的字符级词汇结构不适用于中文，使用新的子词词汇表，它包括字符和常见词汇

·由于中文比英文分布更离散，训练时采用比GPT-3一百万的tokens更大的batch size（三百万tokens）

	n_{param}	n_{layers}	d_{model}	n_{heads}	d_{head}
CPM-Small	109M	12	768	12	64
CPM-Medium	334M	24	1,024	16	64
CPM-Large	2.6B	32	2,560	32	80

Table 1: Model sizes. n_{param} is the number of parameters. n_{layers} is the number of layers. d_{model} is the dimension of hidden states, which is consistent in each layer. n_{heads} is the number of attention heads in each layer. d_{head} is the dimension of each attention head.

Data Source	Encyclopedia	Webpage	Story	News	Dialog
Size	~ 40GB	~ 39GB	~ 10GB	~ 10GB	~ 1GB

Table 2: Details of training data.

Demo

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

官方案例：

Relation:	首都 (Capital)
Prompt:	美国 首都 华盛顿 America Capital Washington 中国 首都 北京 China Capital Beijing 日本 首都 Japan Capital
CPM:	东京 Tokyo
Relation:	主要工艺 (Main Process)
Prompt:	酱焖辣椒 主要工艺 焖 (Sauce Braised Chili) (Main Process) Stew 当归鸭肉煲 主要工艺 煲 (Duck with Angelica) (Main Process) Boil 韭菜煎蛋饼 主要工艺 (Leek Omelette) (Main Process)
CPM:	煎 Fried
Relation:	学科 (Subject)
Prompt:	恒星级黑洞 学科 宇宙论 (Stellar Black Hole) Subject Cosmology 品类需求强度 学科 品牌经济学 (Category Demand Intensity) Subject Economics 大地构造学 学科 (Tectonic Geology) Subject
CPM:	地质学 Geology

Post:	现在高中数学竞赛这么凶残了！ How crazy the math competition of high school is!
CDial-GPT: (Few-shot)	你还真信? Do you really believe it?
CDial-GPT: (Supervised)	数学让我们学会了很多，好好学，明天回家 Math lets us learn a lot. Study hard and go home, tomorrow.
CPM-Large: (Few-shot)	现在都是微积分,大学的东西比高中简单多了,还有什么呢? Now it's all calculus. The things in college are much simpler than those of high school. What else?
CPM-Large: (Supervised)	这是要逼死我吗?每道题都不会做啊 Is this going to kill me? I can't answer any question.
Post:	每天嘻嘻哈哈的笑着，只有自己才知道活的有多累。 I laugh every day, but only I know how tired I am.
CDial-GPT: (Few-shot)	我也是 Me, too.
CDial-GPT: (Supervised)	只有自己才知道活的有多累。 Only I know how tired I am.
CPM-Large: (Few-shot)	当你想要哭泣的时候,就去那些熟悉的地方吧。 When you want to cry, go to those familiar places.
CPM-Large: (Supervised)	真的不知道,生活怎么会这么累 I really don't know how my life could be so tiring.

Demo

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

使用该模型进行推理任务：

```python
import paddle
import argparse
import numpy as np
from GPT2 import GPT2Model, GPT2Tokenizer

# 参数设置
parser = argparse.ArgumentParser()
parser.add_argument("--pretrained_model", type=str, required=True, help="the detection model dir.")
args = parser.parse_args()

# 初始化GPT-2模型
model = GPT2Model(
    vocab_size=30000,
    layer_size=32,
    block_size=1024,
    embedding_dropout=0.0,
    embedding_size=2560,
    num_attention_heads=32,
    attention_dropout=0.0,
    residual_dropout=0.0)

print('正在加载模型，耗时需要几分钟，请稍后...')

# 读取CPM-LM模型参数(FP16)
state_dict = paddle.load(args.pretrained_model)

# FP16 -> FP32
for param in state_dict:
    state_dict[param] = state_dict[param].astype('float32')

# 加载CPM-LM模型参数
model.set_dict(state_dict)

# 将模型设置为评估状态
model.eval()
```

```python
# 加载编码器
tokenizer = GPT2Tokenizer(
    'GPT2/bpe/vocab.json',
    'GPT2/bpe/chinese_vocab.model',
    max_len=512)

# 初始化编码器
_ = tokenizer.encode('_')

print('模型加载完成.')
```

Demo

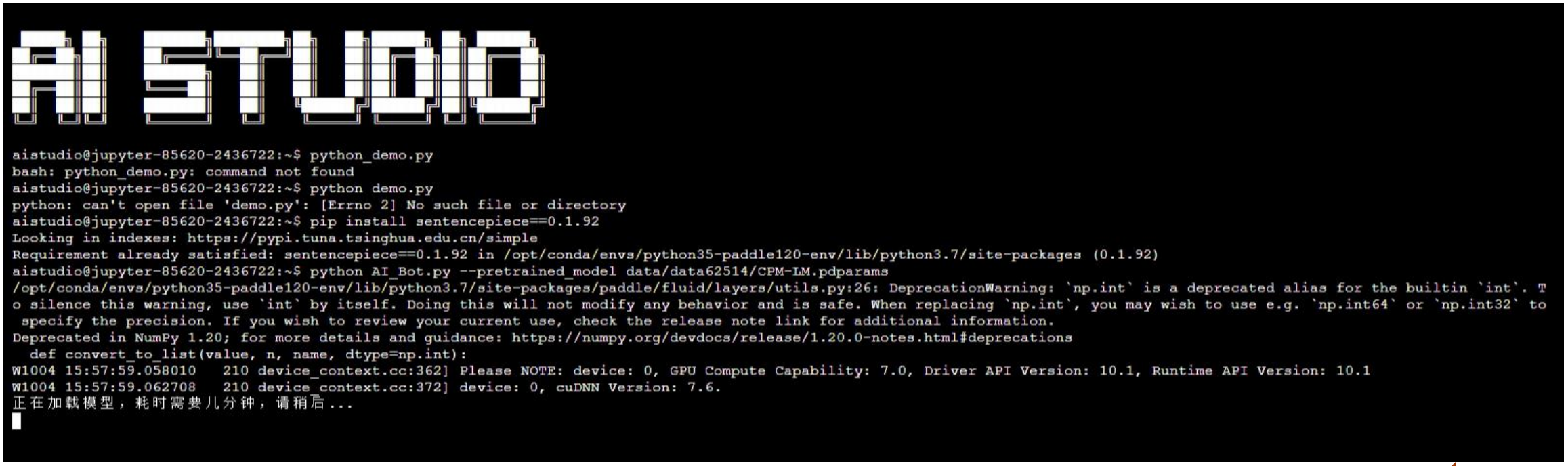

```python
# 主程序
mode = 'q'
funs = ask_question
print('输入“切换”更换问答和古诗默写模式，输入“exit”退出')
while True:
    if mode == 'q':
        inputs = input("当前为问答模式，请输入问题：")
    else:
        inputs = input("当前为古诗默写模式，请输入古诗的上半句：")
    if inputs=='切换':
        if mode == 'q':
            mode = 'd'
            funs = dictation_poetry
        else:
            mode = 'q'
            funs = ask_question
    elif inputs=='exit':
        break
    else:
        funs(inputs)
```

```python
# 基础预测函数
def predict(text, max_len=10):
    ids = tokenizer.encode(text)
    input_id = paddle.to_tensor(np.array(ids).reshape(1, -1).astype('int64'))
    output, cached_kvs = model(input_id, use_cache=True)
    nid = int(np.argmax(output[0, -1].numpy()))
    ids += [nid]
    out = [nid]
    for i in range(max_len):
        input_id = paddle.to_tensor(np.array([nid]).reshape(1, -1).astype('int64'))
        output, cached_kvs = model(input_id, cached_kvs, use_cache=True)
        nid = int(np.argmax(output[0, -1].numpy()))
        ids += [nid]
        # 若遇到'\n'则结束预测
        if nid==3:
            break
        out.append(nid)
    print(tokenizer.decode(out))

# 问答
def ask_question(question, max_len=10):
    predict('''问题：中国的首都是哪里？
    答案：北京。
    问题：李白在哪个朝代？
    答案：唐朝。
    问题：%s
    答案：''' % question, max_len)

# 古诗默写
def dictation_poetry(front, max_len=10):
    predict('''默写古诗：
    白日依山尽，黄河入海流。
    %s，''' % front, max_len)
```

Demo

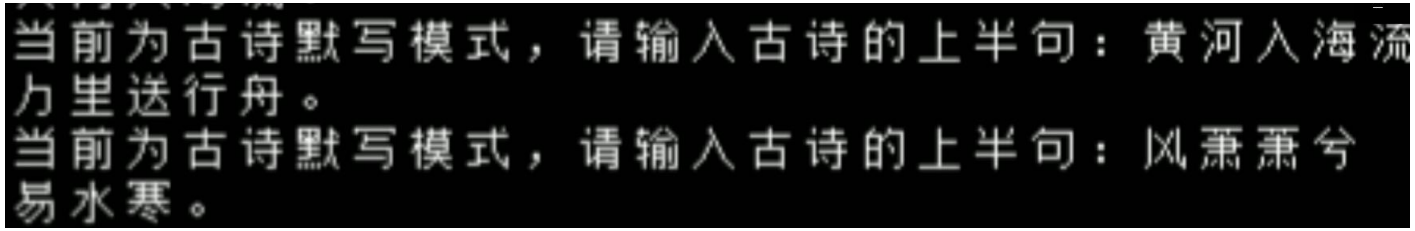
北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

启动



Demo

古诗模式



该模式下不管输入什么都会对成对子

Demo

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

问答模式

```
当前为问答模式，请输入问题：北京理工大？
北京理工大学。
当前为问答模式，请输入问题：今天的日期？
今天是农历的四月初八。
当前为问答模式，请输入问题：Hello World?
你好，世界。
当前为问答模式，请输入问题：pardon?
“pardon”。
当前为问答模式，请输入问题：文学家？
鲁迅。
当前为问答模式，请输入问题：大文学家？
杜甫。
当前为问答模式，请输入问题：林黛玉？
林黛玉。
当前为问答模式，请输入问题：林黛玉葬花日期？
“黛玉”。
当前为问答模式，请输入问题：新中国成立日期？
1949年10月1日。
当前为问答模式，请输入问题：核聚变？
铀235。
当前为问答模式，请输入问题：感冒了吃什么药？
板蓝根。
当前为问答模式，请输入问题：你叫什么名字？
我叫“??”。
当前为问答模式，请输入问题：gpt-3是什么？
“歼-5”。
当前为问答模式，请输入问题：中文预训练模型是什么？
中文预训练模型是“中文”
当前为问答模式，请输入问题：你好？
你好。
当前为问答模式，请输入问题：你会算微积分吗？
不会。
当前为问答模式，请输入问题：你是什么？
我是一个“??”。
当前为问答模式，请输入问题：歼-20发明时间？
1969年。
当前为问答模式，请输入问题：人的化学构成有？
碳、氢、氧、
```

Demo

此外，GPT系列模型还有续写功能，在尝试了清源续写和彩云小梦等模型后，可以发现GPT有较强的模仿推理能力，但并未真正对其中的关系有所了解。即使如此，GPT结构的模型在短问答中的准确性也比较优秀了。

黛玉刚进贾府，正和贾母等谈论着自己的体弱多病和吃药等事，“一语未了，只听后院中有人笑声，说：“我来迟了，未曾迎接远客！”黛玉纳罕道：“这些人个个皆敛声屏气，恭肃严整如此，这来者系谁，这样放诞无礼？”“这个人打扮与众姑娘不同，彩绣辉煌，恍若神妃仙子：头上戴着金丝八宝攒珠髻，绾着朝阳五凤挂珠钗；项上带着赤金盘螭璎珞圈；裙边系着豆绿宫绦，双衡比目玫瑰佩；身上穿着缕金百蝶穿花大红洋缎窄褙袄，外罩五彩刻丝石青银鼠褂；下着翡翠撒花洋绉裙。一双丹凤三角眼，两弯柳叶吊梢眉，身量苗条，体格风骚，粉面含春威不露，丹唇未启笑先闻。”黛玉听了,只觉这人眼熟,一时想不起来,便道:“既是不认得,就请回罢,我这里不留人。”

单击其他卡片试试快速切换

起，毕竟妹妹是男孩子。

"哦~~~我知道了。"

"好了，快睡吧！"...

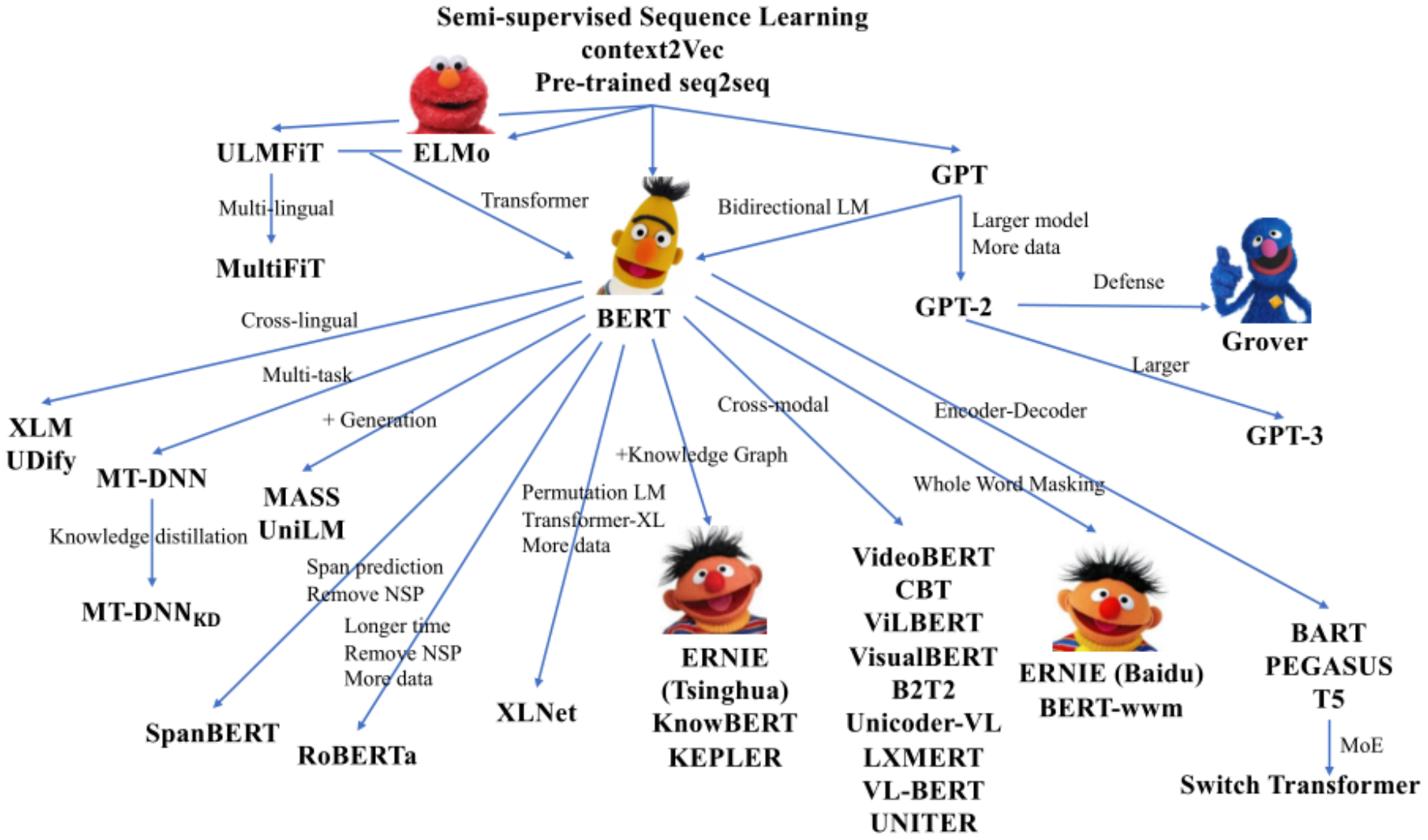
北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

5

前沿进展

常欣煜

模型特点


Semi-supervised Sequence Learning
context2Vec
Pre-trained seq2seq
ELMo
ULMFiT
Multi-lingual
MultiFiT
Transformer
BERT
Bidirectional LM
GPT
Larger model
More data
GPT-2
Defense
Grover
Larger
GPT-3
Cross-lingual
XLM
UDify
Multi-task
MT-DNN
Knowledge distillation
MT-DNN$_{KD}$
+ Generation
MASS
UniLM
Span prediction
Remove NSP
SpanBERT
Longer time
Remove NSP
More data
RoBERTa
Permutation LM
Transformer-XL
More data
XLNet
Cross-modal
+Knowledge Graph
ERNIE
(Tsinghua)
KnowBERT
KEPLER
VideoBERT
CBT
ViLBERT
VisualBERT
B2T2
Unicoder-VL
LXMERT
VL-BERT
UNITER
Encoder-Decoder
Whole Word Masking
ERNIE (Baidu)
BERT-wwm
BART
PEGASUS
T5
MoE
Switch Transformer

模型特点

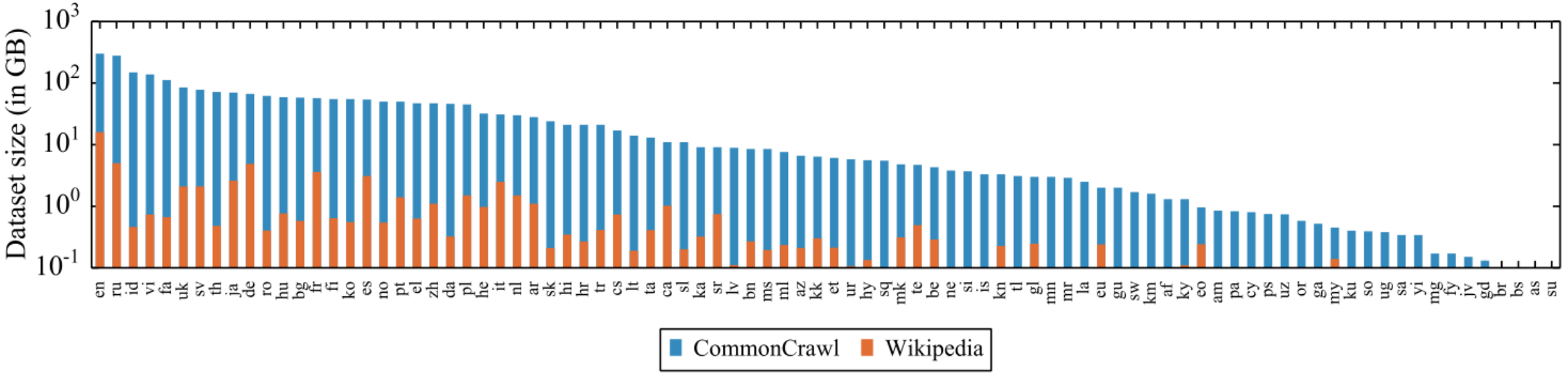
北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

PTMs	Architecture[a)]	Input	Pre-training task	Corpus	Params	GLUE[b)]	FT?[c)]
ELMo [14]	LSTM	Text	BiLM	WikiText-103			No
GPT [15]	Transformer Dec.	Text	LM	BookCorpus	117M	72.8	Yes
GPT-2 [60]	Transformer Dec.	Text	LM	WebText	117M–1542M		No
BERT [16]	Transformer Enc.	Text	MLM & NSP	WikiEn+BookCorpus	110M–340M	81.9[c)]	Yes
InfoWord [54]	Transformer Enc.	Text	DIM+MLM	WikiEn+BookCorpus	=BERT	81.1[c)]	Yes
RoBERTa [42]	Transformer Enc.	Text	MLM	BookCorpus+CC-News+OpenWebText+ STORIES	355M	88.5	Yes
XLNet [48]	Two-Stream Transformer Enc.	Text	PLM	WikiEn+ BookCorpus+Giga5 +ClueWeb+Common Crawl	≈BERT	90.5[d)]	Yes
ELECTRA [55]	Transformer Enc.	Text	RTD+MLM	same to XLNet	335M	88.6	Yes
UniLM [43]	Transformer Enc.	Text	MLM[f)]+ NSP	WikiEn+BookCorpus	340M	80.8	Yes
MASS [40]	Transformer	Text	Seq2Seq MLM	*Task-dependent			Yes
BART [49]	Transformer	Text	DAE	same to RoBERTa	110% of BERT	88.4[c)]	Yes
T5 [41]	Transformer	Text	Seq2Seq MLM	Colossal Clean Crawled Corpus (C4)	220M–11B	89.7[c)]	Yes
ERNIE(THU) [61]	Transformer Enc.	Text+Entities	MLM+NSP+dEA	WikiEn + Wikidata	114M	79.6	Yes
KnowBERT [62]	Transformer Enc.	Text	MLM+NSP+EL	WikiEn + WordNet/Wiki	253M–523M		Yes
K-BERT [63]	Transformer Enc.	Text+Triples	MLM+NSP	WikiZh + WebtextZh + CN-DBpedia + HowNet + MedicalKG	=BERT		Yes
KEPLER [59]	Transformer Enc.	Text	MLM+KE	WikiEn + Wikidata/WordNet			Yes
WKLM [56]	Transformer Enc.	Text	MLM+ERD	WikiEn + Wikidata	=BERT		Yes

更丰富的数据

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

- 多语种
  - 尽管来自世界各地的人们使用不同的语言，他们可以表达相同的意思
  - 用一个模型表征多种语言模型



XLM使用CC-100数据库，包含100种语言

Unsupervised Cross-lingual Representation Learning at Scale

更丰富的数据

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

MODEL	en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur	Avg
Evaluation of cross-lingual sentence encoders (Cross-lingual transfer)																
BiLSTM	73.7	67.7	68.7	67.7	68.9	67.9	65.4	64.2	64.8	66.4	64.1	65.8	64.1	55.7	58.4	65.6
mBERT	81.4	-	74.3	70.5	-	-	-	-	62.1	-	-	63.8	-	-	58.3	-
XLM	85.0	78.7	78.9	77.8	76.6	77.4	75.3	72.5	73.1	76.1	73.2	76.5	69.6	68.4	67.3	75.1
Unicoder	85.1	79.0	79.4	77.8	77.2	77.2	76.3	72.8	73.5	76.4	73.6	76.2	69.4	69.7	66.7	75.4
XLM-R_{Base}	85.8	79.7	80.7	78.7	77.5	79.6	78.1	74.2	73.8	76.5	74.6	76.7	72.4	66.5	68.3	76.2
$HICTL_{Base}$	**86.3**	**80.5**	**81.3**	**79.5**	**78.9**	**80.6**	**79.0**	**75.4**	**74.8**	**77.4**	**75.7**	**77.6**	**73.1**	**69.9**	**69.7**	**77.3**
Machine translate at training (Translate-train)																
BiLSTM	73.7	68.3	68.8	66.5	66.4	67.4	66.5	64.5	65.8	66.0	62.8	67.0	62.1	58.2	56.6	65.4
mBERT	81.9	-	77.8	75.9	-	-	-	-	70.7	-	-	76.6	-	-	61.6	-
XLM	85.0	80.2	80.8	80.3	78.1	79.3	78.1	74.7	76.5	76.6	75.5	78.6	72.3	70.9	63.2	76.7
Unicoder	85.1	80.0	81.1	79.9	77.7	80.2	77.9	75.3	76.7	76.4	75.2	79.4	71.8	71.8	64.5	76.9
$HICTL_{Base}$	**85.7**	**81.3**	**82.1**	**80.2**	**81.4**	**81.0**	**80.5**	**79.7**	**77.4**	**78.2**	**77.5**	**80.2**	**75.4**	**73.5**	**72.9**	**79.1**

ON LEARNING UNIVERSAL REPRESENTATIONS ACROSS LANGUAGES

更丰富的数据

- 多模态：V（Vision）&L（Language）
- 难点：将非文本信息融合进BERT等预训练模型
- 训练任务：
  - 掩蔽文本预测
  - 掩蔽动作预测
  - 掩蔽物体推测
  - 视频-文本对齐

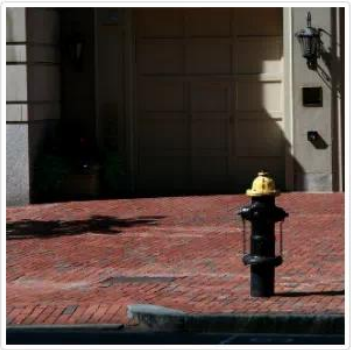
更丰富的数据

- 多模态任务：视觉问答

Question : What color is the hydrant?

Original Image | **red**

Complementary Image | **black and yellow**



Question : What color is the ground?

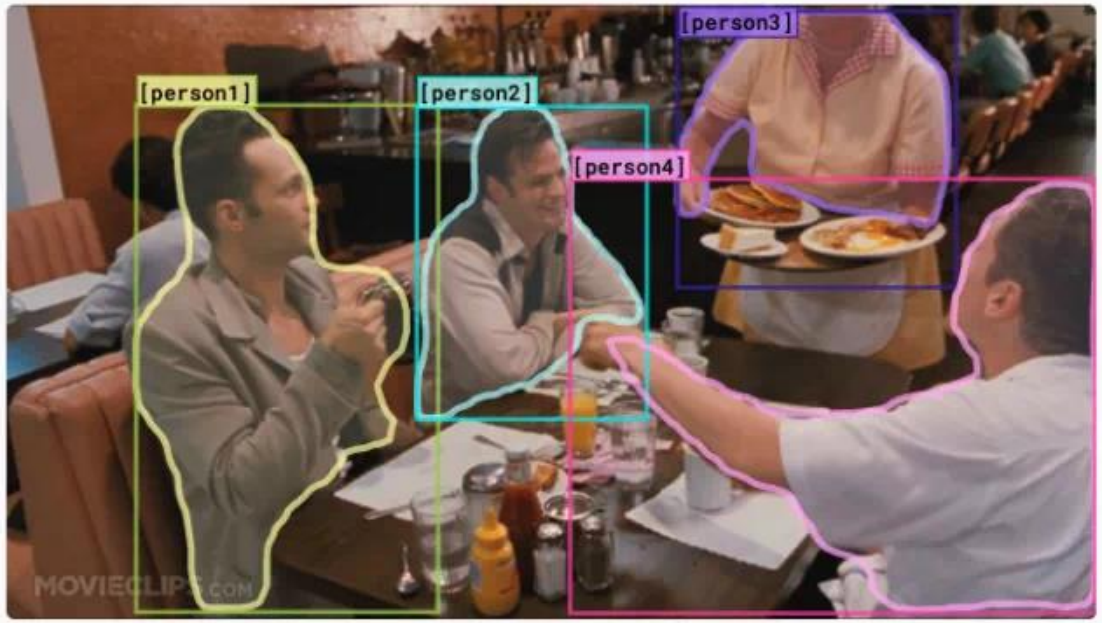
Original Image | **brown**

Complementary Image | **brown and green**

更丰富的数据

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

- 多模态任务：视觉常识推理



Why is [person4] pointing at [person1]?

a) He is telling [person3] that [person1] ordered the pancakes.

b) He just told a joke.

c) He is feeling accusatory towards [person1].

d) He is giving [person1] directions.

Rationale: I think so because...

a) [person1] has the pancakes in front of him.

b) [person4] is taking everyone's order and asked for clarification.

c) [person3] is looking at the pancakes both she and [person2] are smiling slightly.

d) [person3] is delivering food to the table, and she might not know whose order is whose.

更丰富的数据

- 多模态任务：多模态检索

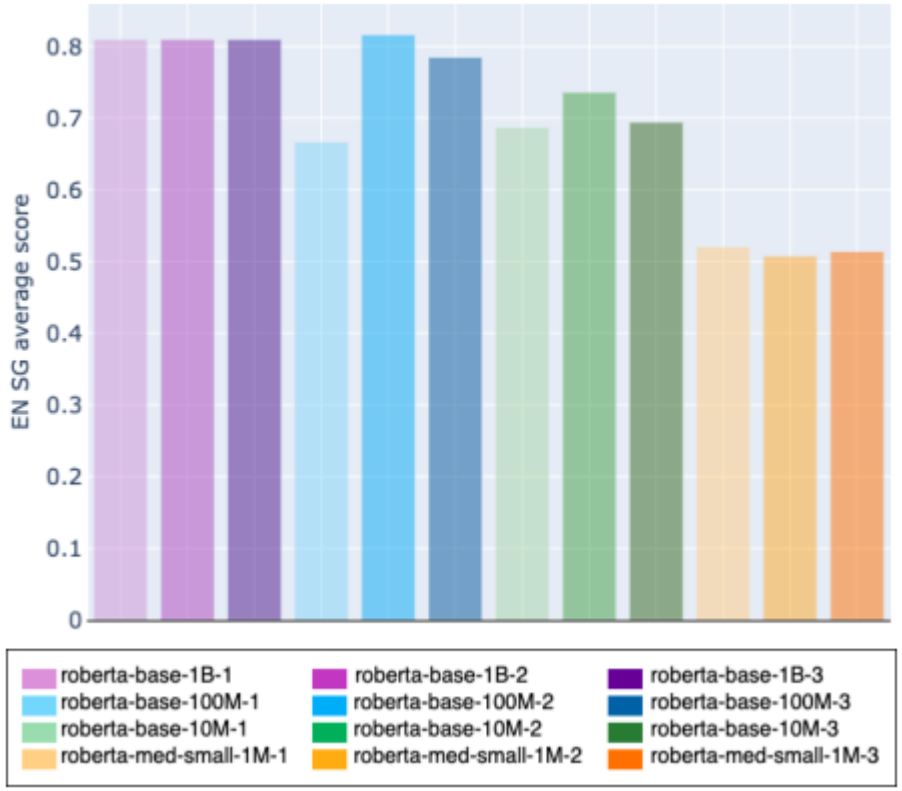
Query: leopard-print women's shoes

Relevant product

Irrelevant product

更大规模的模型

- 仅仅扩大预训练模型规模的情况下，模型就可以取得更好的效果
- 将预训练模型的参数规模从百万级进一步推进到十亿级，甚至万亿级



基于RoBERTa原始文本数据，训练数据大小对模型句法能力的影响

How much pretraining data do language models need to learn syntax?

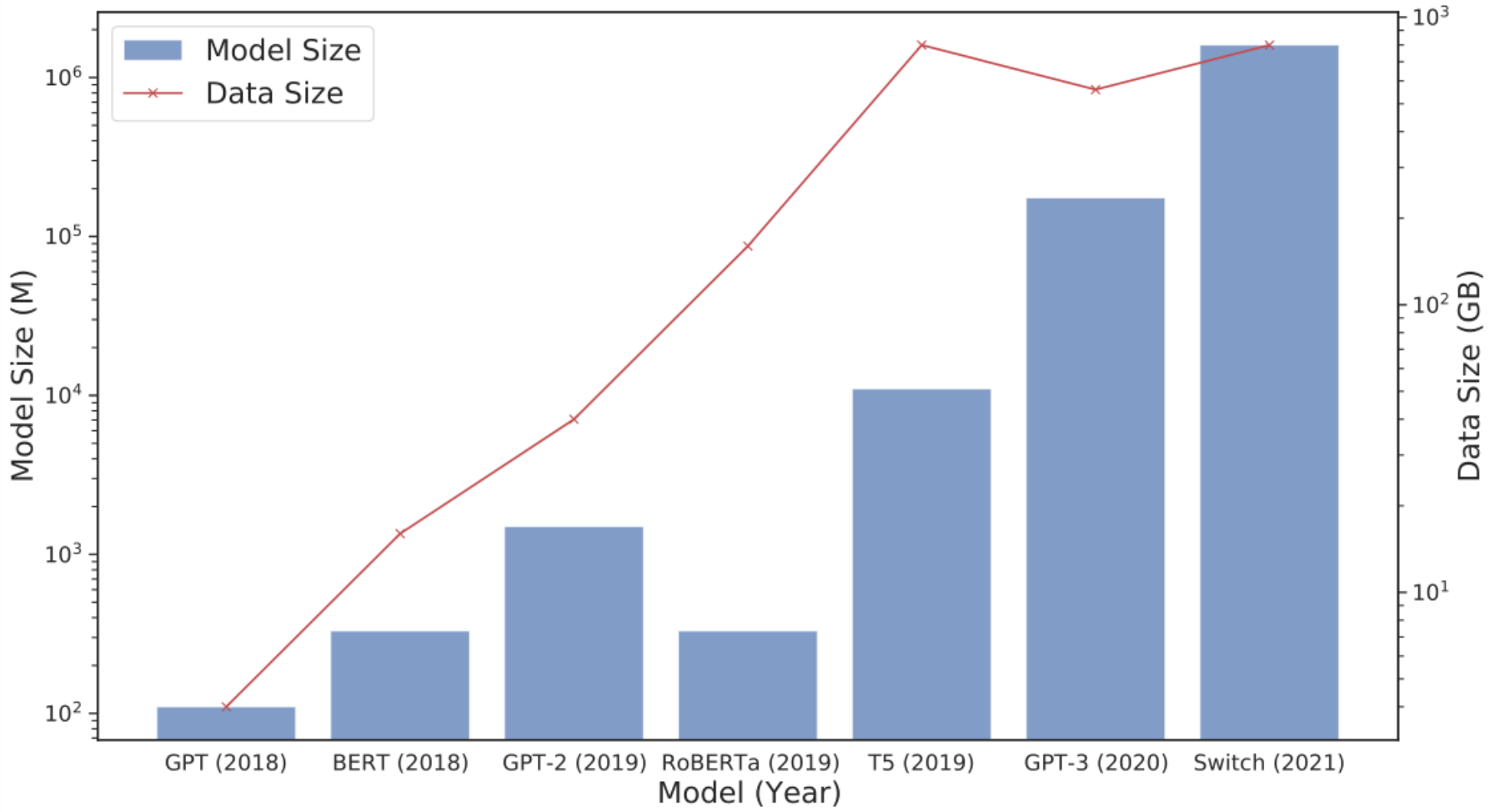
更大规模的模型

- MiniBERTas的云计算成本和二氧化碳排放以及它们在词性标注（PoS）、依存句法分析（Dep. parsing）和释义识别（Paraphrase id.）方面的平均性能

Model family	Cost	CO_2e (lbs)	PoS	Dep. parsing	Paraphrase id.
roberta-base-1B	$20320	2330	96.03 (+0.5%)	85.73 (+1.76%)	89.59 (+2.02%)
roberta-base-100M	$5075	582.5	95.53 (+1.11%)	83.97 (+4.04%)	87.57 (+2.79%)
roberta-base-10M	$500	58.25	94.42 (+2.73%)	79.93 (+14.48%)	84.78 (+5.34%)
rob-med-small-1M	$50	5.825	91.69 (base)	65.45 (base)	79.44 (base)

How much pretraining data do language models need to learn syntax?

更大规模的模型



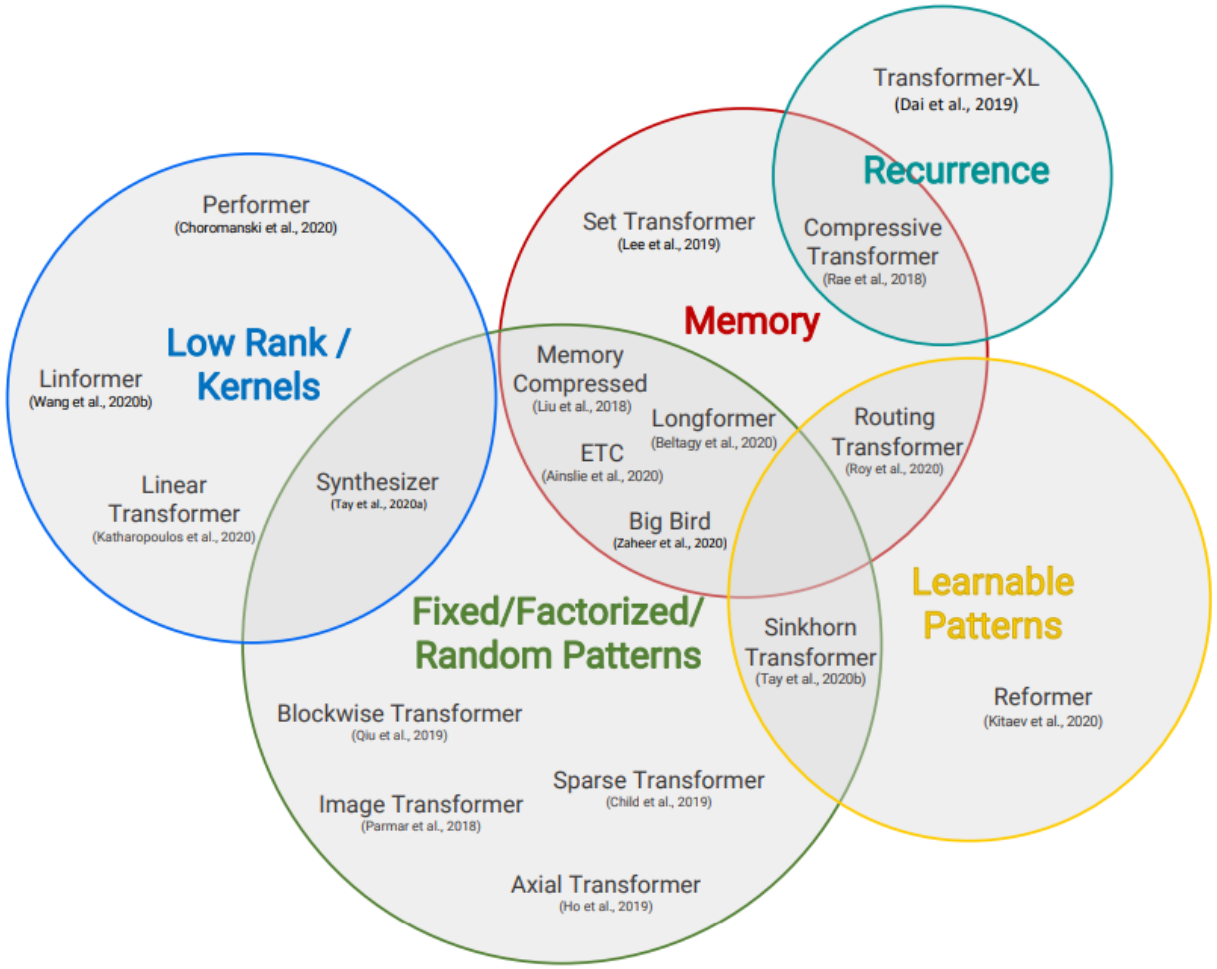
近年来语言预训练模型大小和数据大小

提高训练效率

- 训练Transformer已经被证明是一个强大的模型，但是效率低下，需要消耗大量计算资源
- 解决方案：
  - 并行训练（模型并行、数据并行）
  - 模型压缩
    - 参数共享（ALBERT）
    - 知识蒸馏（将一个小的学生模型训练成一个大的教师模型的行为，例如：DistillBERT、TinyBERT、MiniLM）
    - 模型量化（Q8-BERT）、模型剪枝（CompressingBERT）……

提高训练效率

- 新的架构

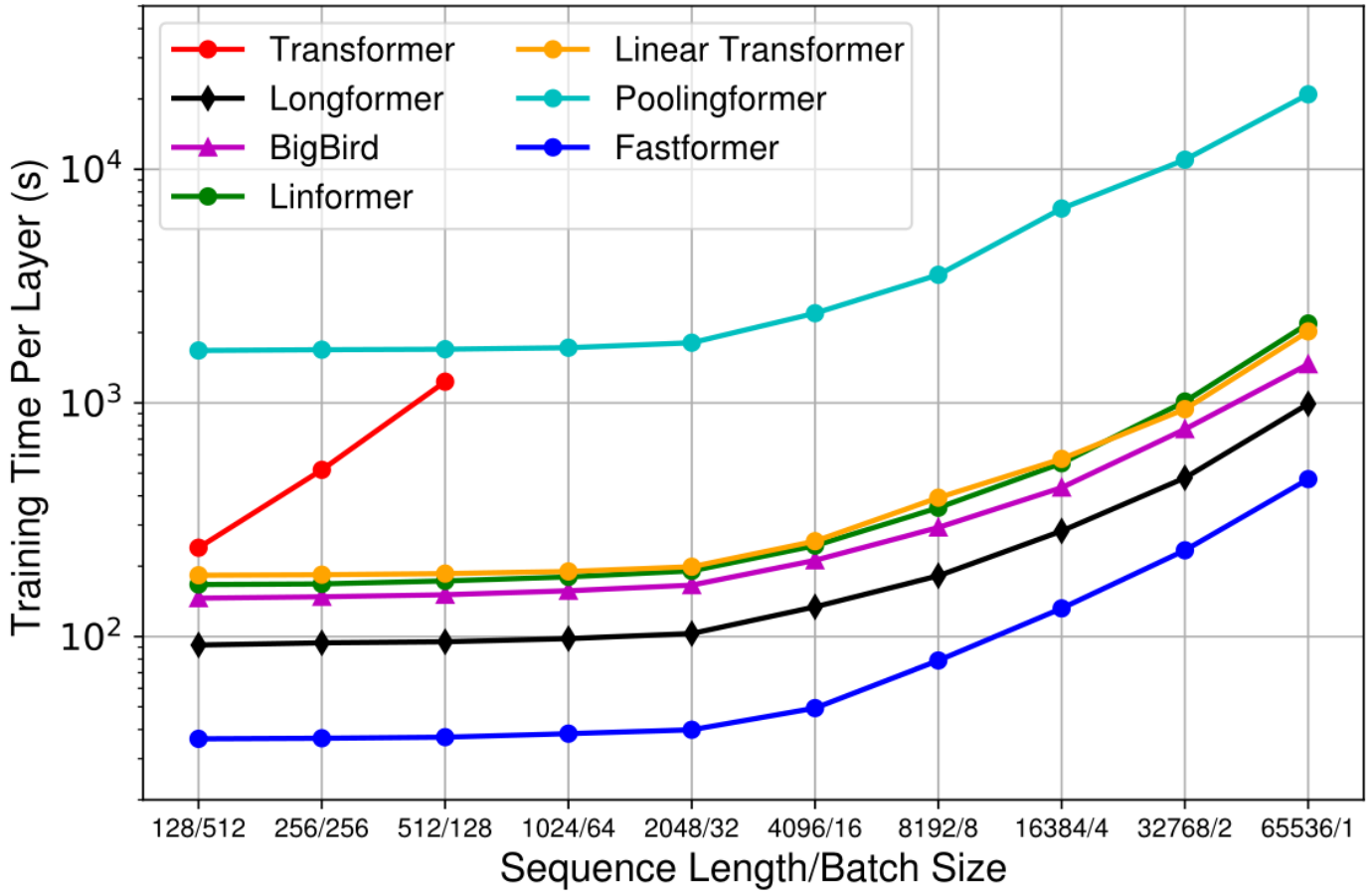


Efficient Transformers: A Survey

提高训练效率

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

- 新的架构



Fastformer: Additive Attention Can Be All Y ou Need

中文预训练模型：盘古-α

- 今年4月，鹏城实验室发布了GPT-3的中文版本，命名为盘古-α，基于“鹏城云脑Ⅱ”和MindSpore框架的自动混合并行模式
- 模型750 GB，收集了 1.1 TB 的中文语料库，包括：
  - 语言电子书
  - 百科全书
  - 消息
  - 社交媒体
  - 网页

中文预训练模型：盘古-α

- 该模型包含超过 2000 亿个参数，比 GPT 3 多出 2500 万个参数。它以更高效的方式完成了跨文本摘要、问答和对话生成等语言任务。

Task	Dataset	Input&Prompt
Cloze and completion	WPLC	/
	CHID	/
	PD&CFT	/
	CMRC2017	/
	CMRC2019	/
	CMRC2018	/
Reading comprehension	DRCD	阅读文章：$Document\n问：$Question\n答：(Read document: $Document\nQuestion：$Question\nAnswer:)
	DuReader	阅读文章：$Document\n问：$Question\n答：(Read document: $Document\nQuestion：$Question\nAnswer:)
Closed book QA	WebQA	阅读文章：$Document\n问：$Question\n答：(Read document: $Document\nQuestion：$Question\nAnswer:
Winograd-Style	CLUEWSC2020	问：$Question\n答：(Question：$Question\nAnswer:)
Common sense reasoning	C^3	/
NLI	CMNLI	问: $Question\n答:$Choice\n该答案来自对话: $Passage (Question: $Question\nAnswer:$Choice\nAnswer from dialogue: $Passage)
	OCNLI	$S1?对/或许/错，$S2 ($S1?Yes/Maybe/No, $S2)
Text classification	TNEWS	$S1?对/或许/错，$S2 ($S1?Yes/Maybe/No, $S2)
	IFLYTEK	这是关于$label的文章：$passage (This passage is about$label: $passage)
	AFQMC	这是关于$label的应用程序：$passage (This application is about: $passage)
	CSL	下面两个句子语义相同/不同：$S1。$S2 (The following two sentences have the same/different semantics: $S1. $S2)
		摘要：$passage，关键词：$keyword是/不是真实关键词 (Abstract: $passage, keywords: $keyword True/False keywords)

谢谢老师和各位同学
敬请批评指正

Thanks for your listening

北京理工大學
BEIJING INSTITUTE OF TECHNOLOGY

常欣煜、徐子懋、王雪飞、王子轩、禹言江、赵华耀