大语言模型

Large Language Model (LLM)

背景介绍

随着近年来人工智能技术的飞速发展，大模型的发展趋势也可谓是“激流勇进”，但在迅猛发展的背后，技术的局限性也日益凸显。研究者们针对这些局限性也在积极探索并提出了不计其数的新方法和研究方向。

章节概述

CHAPTER OVERVIEW

章节概述

- 超大规模预训练网络的基本原理及典型网络代表
- 主流模型压缩方法
  - 模型剪枝
  - 模型量化
  - 模型蒸馏
- 其他自然语言处理热门研究点
  - 问答系统
  - 机器阅读理解
- 多模态任务

小节介绍

SECTION INTRODUCTION

1

超大规模预训练网络

预训练技术，BERT，GPT系列

2

模型压缩方法

模型剪枝，模型量化，模型蒸馏

3

其它热门的研究点

问答系统，机器阅读理解

4

多模态任务的举例与现状

多模态学习，图像-文本多模态任务

超大规模预训练网络

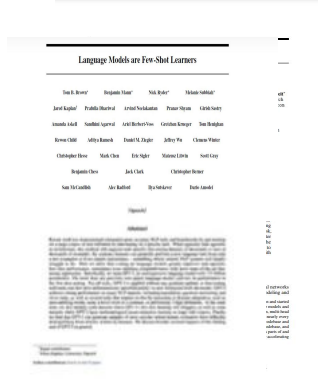
预训练是通过设计好一个网络结构来做语言模型任务，然后把大量甚至是无穷尽的无标注自然语言文本利用起来，预训练任务把大量语言学知识抽取出来编码到网络结构中，在当前任务带有标注信息的数据有限时，这些先验的语言学特征会对当前任务有极大的特征补充作用。

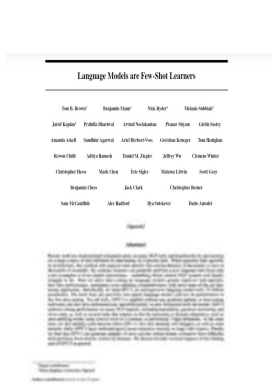
大语言模型(Large Language Model, LLM)发展史

OpenAI


Improving Language Understanding by Generative Pre-Training


Language Models are Unsupervised Multitask Learners


Language Models are Few-Shot Learners


Language Models are Few-Shot Learners

GPT 06/2018

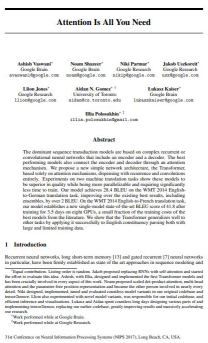
GPT-2 02/2019

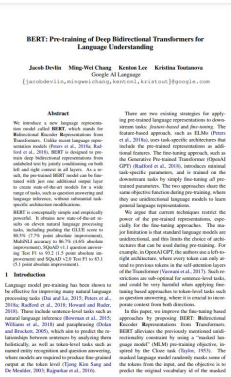
GPT-3 05/2020

GPT-4 05/2023

Transformer 06/2017

BERT 10/2018

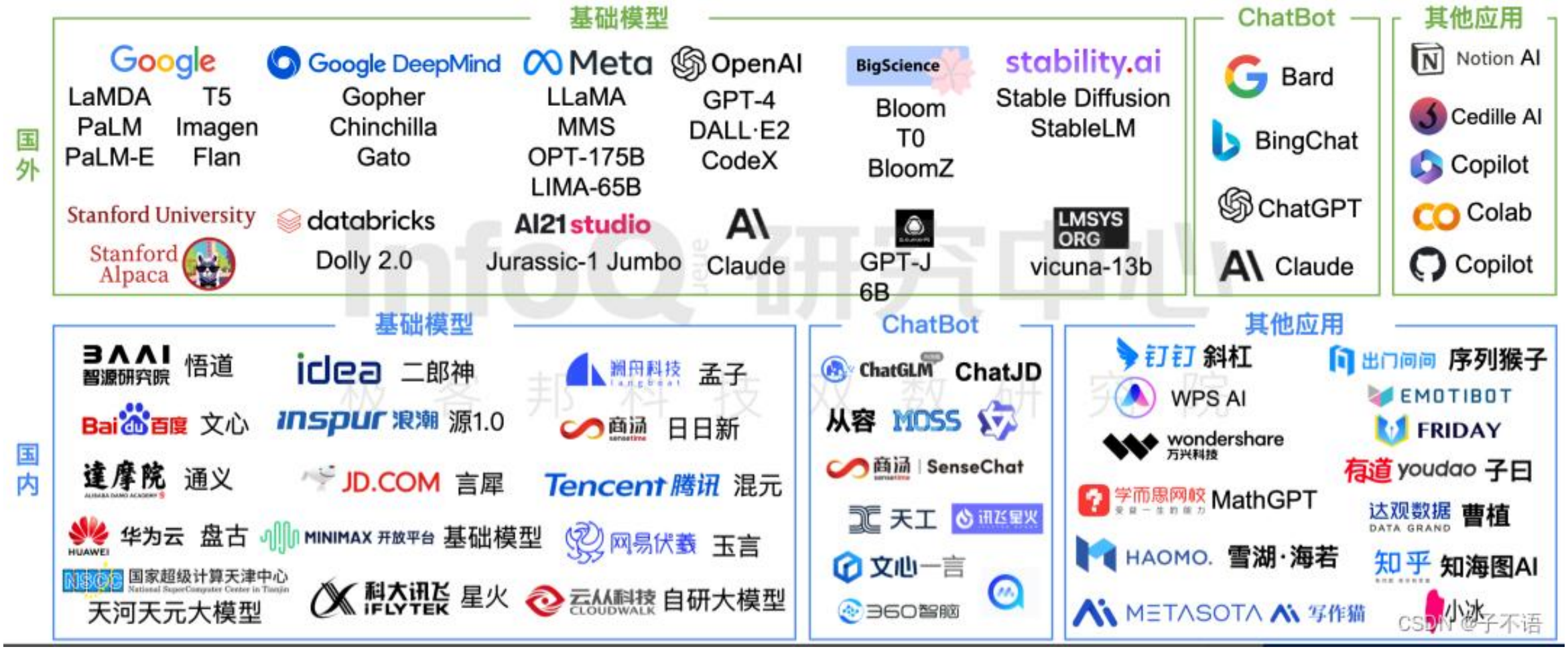

Attention Is All You Need


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

- 大语言模型的“三大”特点：
  - 大数据
  - 大模型
  - 大算力

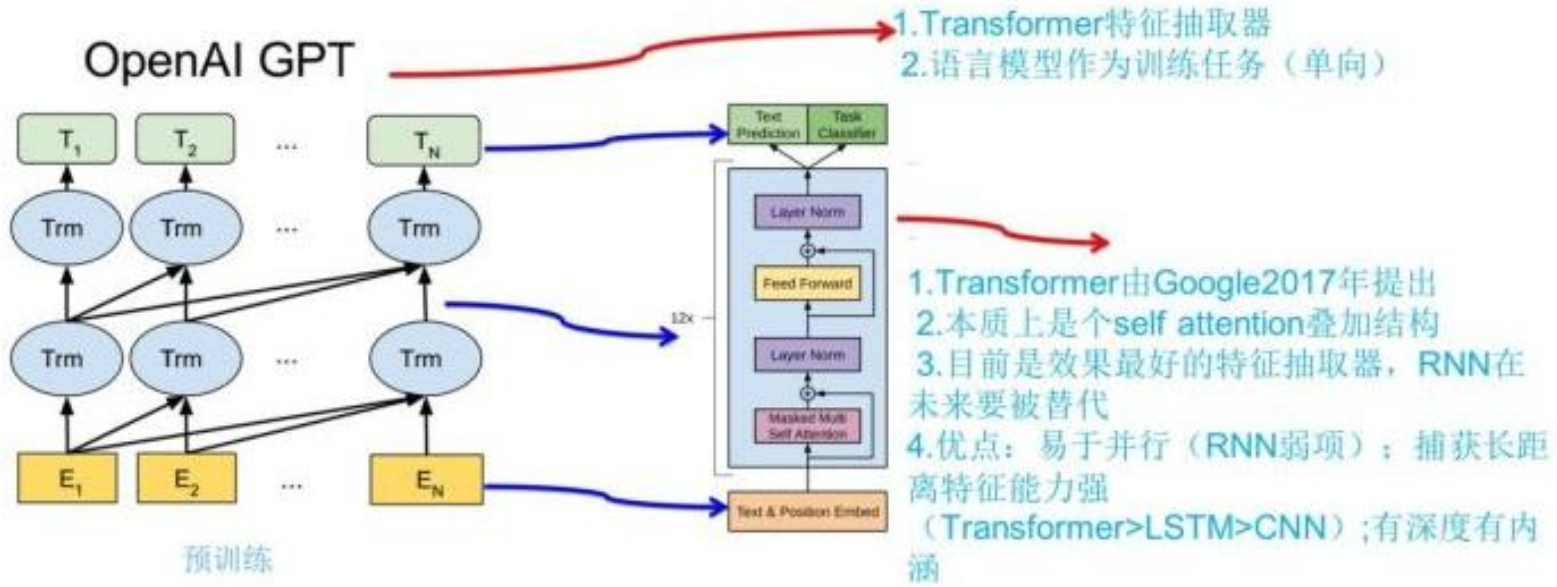
大语言模型(Large Language Model, LLM)

国内外大模型百花齐放



Generative Pre-training Transformer（GPT）

从WE到GPT：Pretrain+Finetune两阶段过程



“基于微调的模式” 论文：Improving Language Understanding by Generative Pre-Training

GPT

- 第一阶段 → 利用语言模型进行预训练
- 第二阶段 → 通过微调的模式解决下游任务

Generative Pre-Training（GPT）预训练方式

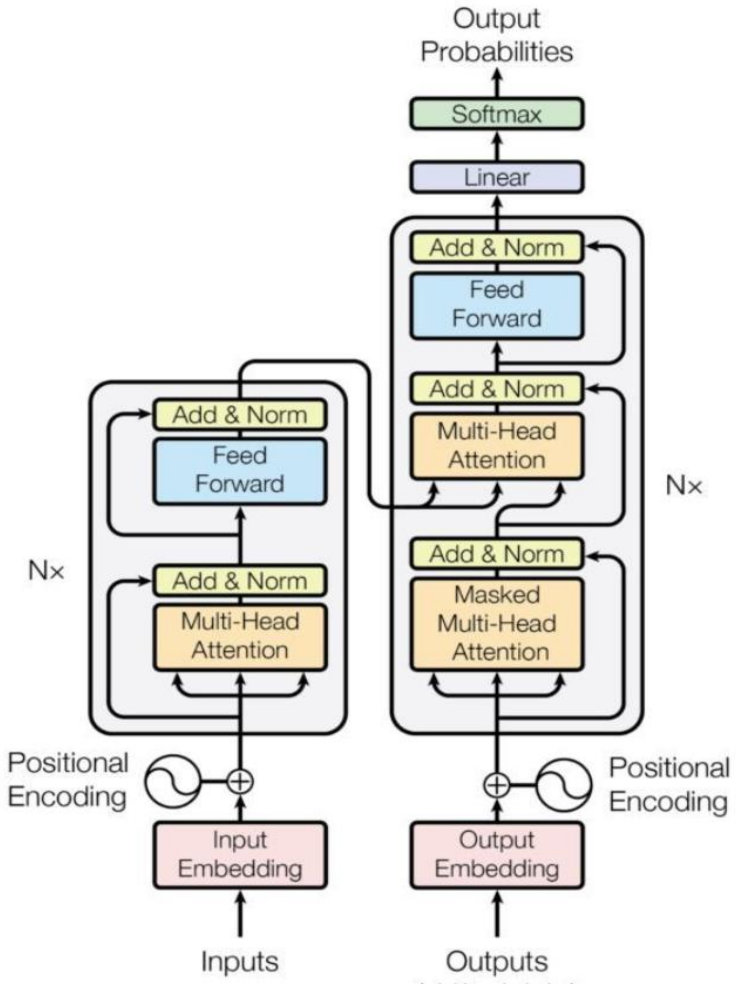
Transformer在做机器翻译任务时，需要平行数据集。但是我们身边存在大量没有标注的数据，例如文本、图片、代码等等。标注这些数据需要花费大量的人力和时间，标注的速度远远不及数据产生的速度，所以带有标签的数据往往只占有总数据集很小的一部分。随着算力的不断提高，计算机能够处理的数据量逐渐增大。如果不能很好利用这些无标签的数据就显得很浪费。

半监督学习和预训练+微调的二阶段模式整变得越来越受欢迎。最常见的二阶段方法就是Word2Vec，使用大量无标记的文本训练出带有一定语义信息的词向量，然后将这些词向量作为下游机器学习任务的输入，就能够大大提高下游模型的泛化能力。但是Word2Vec有一个问题，就是单个单词只能有一个Embedding。这样一来，一词多义就不能很好地进行表示。

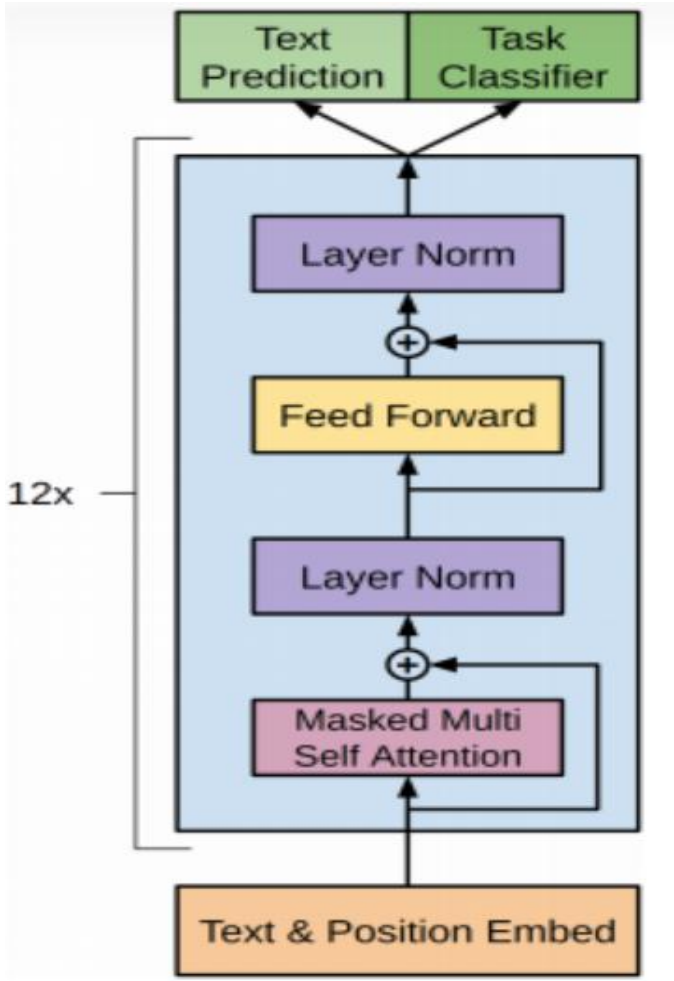
Generative Pre-Training（GPT）预训练方式

单向Transformer结构

在Transformer的文章中，提到了Encoder与Decoder使用的Transformer Block是不同的。在Decoder Block中，使用了Masked Self-Attention，即句子中的每个词，都只能对包括自己在内的前面所有词进行Attention，这就是单向Transformer。GPT使用的Transformer结构就是将Encoder中的Self-Attention替换成了Masked Self-Attention。



Transformer



GPT

Generative Pre-Training（GPT）训练方式

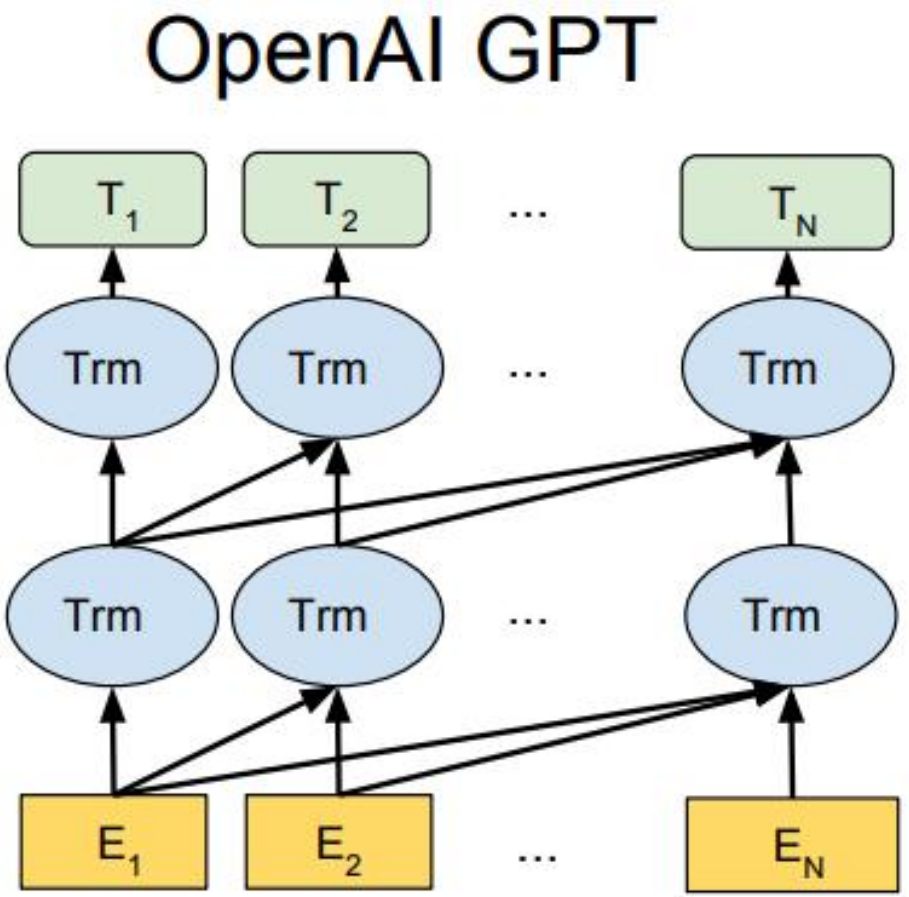
GPT预训练过程

由于采用的是单向的Transformer，只能看到上文的词，所以语言模型为：

$$L_1(\mathcal{U}) = \sum_i \log P(u_i | u_{i-k}, \ldots, u_{i-1}; \Theta)$$

训练过程就是将句子n个词的词向量(第一个为<BOS>)加上Positional Encoding后输入到前面提到的Transfromer中，n个输出分别预测该位置的下一个词(<BOS>预测句子中的第一个词，最后一个词的预测结果不用于语言模型的训练)。

由于使用了Masked Self-Attention，所以每个位置的词都不会“看见”后面的词，也就是预测的时候是看不见“答案”的，保证了模型的合理性，这也是为什么OpenAI采用了单向Transformer的原因。



Generative Pre-Training（GPT）微调过程

接下来就进入模型训练的第二步，运用少量的带标签数据对模型参数进行微调。

上一步中最后一个词的输出我们没有用到，在这一步中就要使用这一个输出来作为下游监督学习的输入。

$$L_2(\mathcal{C}) = \sum_{(x,y)} \log P(y|x^1, \ldots, x^m).$$

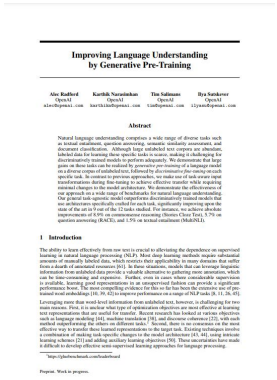
$$P(y|x^1, \ldots, x^m) = \texttt{softmax}(h_l^m W_y).$$

为避免Fine-Tuning使得模型陷入过拟合，文中还提到了辅助训练目标的方法，类似于一个多任务模型或者半监督学习。具体方法就是在使用最后一个词的预测结果进行监督学习的同时，前面的词继续上一步的无监督训练，使得最终的损失函数成为：

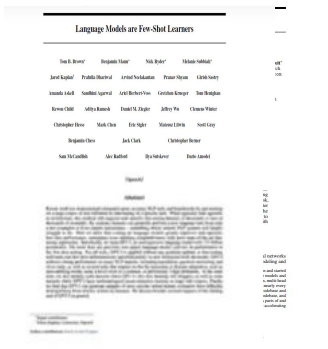
$$L_3(\mathcal{C}) = L_2(\mathcal{C}) + \lambda * L_1(\mathcal{C})$$

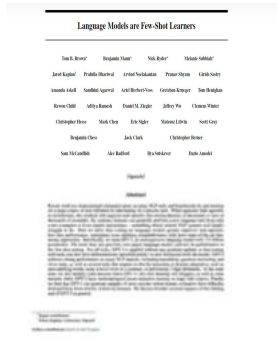
大语言模型(Large Language Model, LLM)发展史

OpenAI


Improving Language Understanding by Generative Pre-Training


Language Models are Unsupervised Multitask Learners


Language Models are Few-Shot Learners


Language Models are Few-Shot Learners

GPT
06/2018

GPT-2
02/2019

GPT-3
05/2020

GPT-4
05/2023

Transformer
06/2017


Attention Is All You Need

BERT
10/2018


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

- 大语言模型的“三大”特点：
  - 大数据
  - 大模型
  - 大算力

BERT（Bidirectional Encoder Representation from Transformers）

BERT与GPT的相似之处：

BERT(Bidirectional Encoder Representation from Transformer)，是Google Brain在2018年提出的基于Transformer的自然语言表示框架。是一提出就大火的明星模型。BERT与GPT一样，采取了Pre-training + Fine-tuning的训练方式，在分类、标注等任务下都获得了更好的效果。

BERT与GPT非常的相似，都是基于Transformer的二阶段训练模型，都分为Pre-Training与Fine-Tuning两个阶段，都在Pre-Training阶段无监督地训练出一个可通用的Transformer模型，然后在Fine-Tuning阶段对这个模型中的参数进行微调，使之能够适应不同的下游任务。

BERT与GPT的不同之处：

训练目标和模型结构和使用上不同：

GPT采用的是单向的Transformer，而BERT采用的是双向的Transformer，也就是不用进行Mask操作；

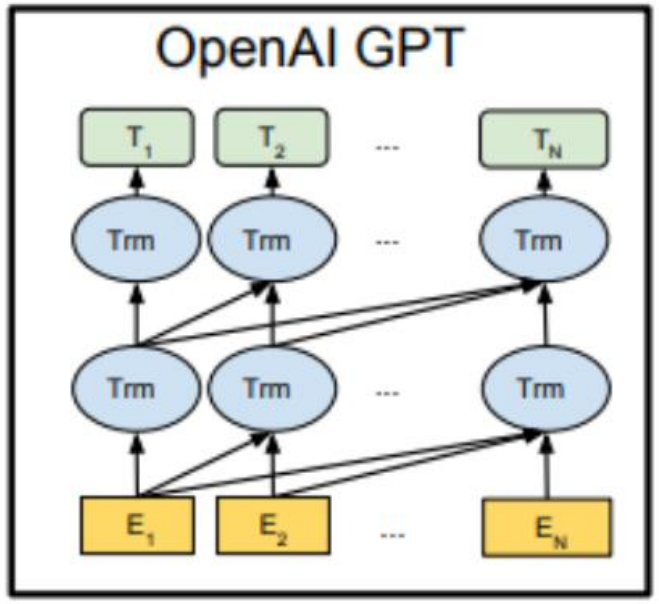
使用的结构的不同，直接导致了它们在Pre-Training阶段训练目标的不同；

超大规模预训练网络介绍-BERT

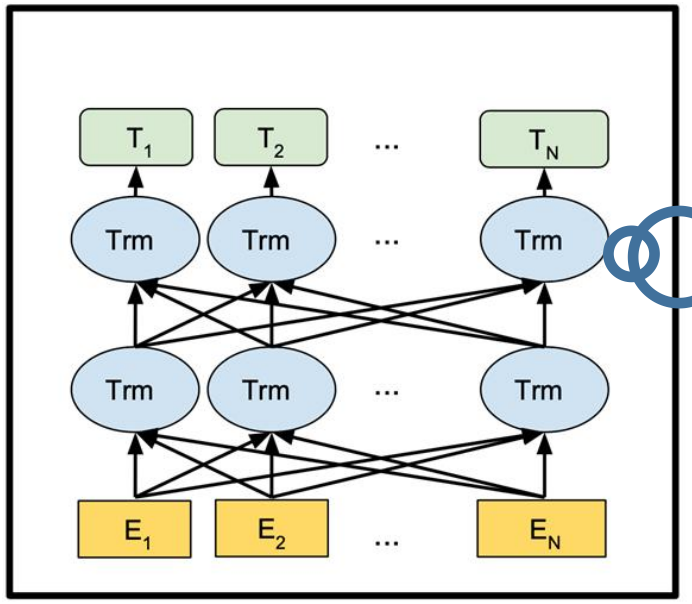
BERT（Bidirectional Encoder Representation from Transformers）

基于Transformer的双向编码器表示

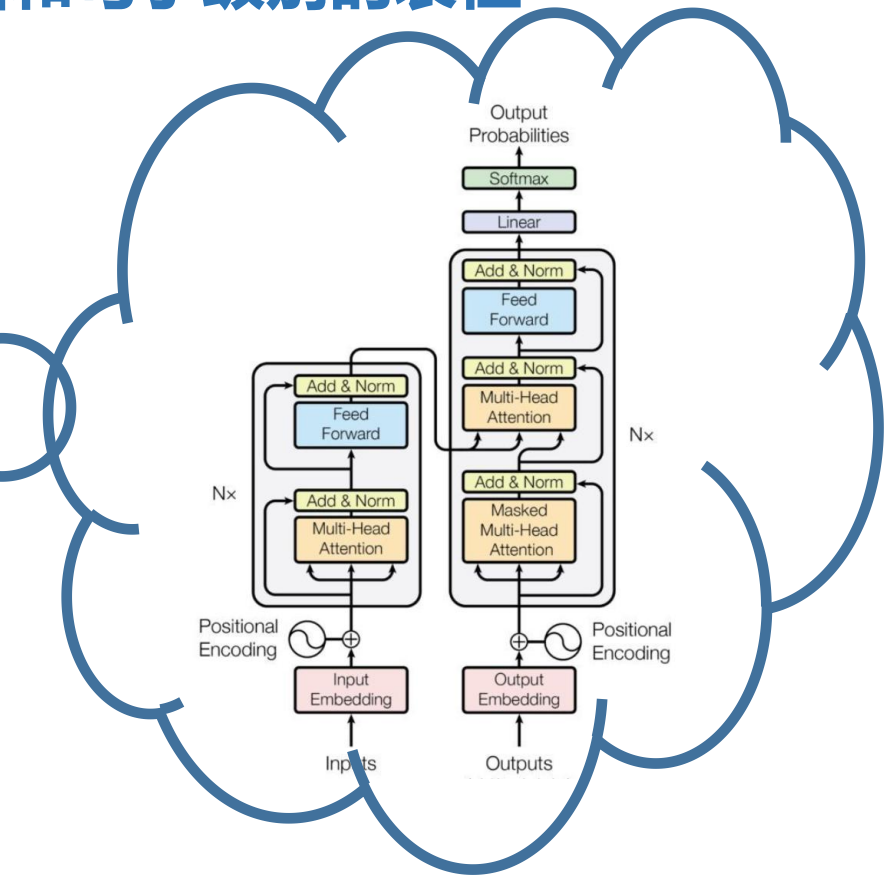
预训练方法的创新 → **用了Masked LM和Next Sentence Prediction两种方法分别捕捉词语和句子级别的表征**



GPT模型结构



BERT模型结构



BERT（Bidirectional Encoder Representation from Transformers）

Masked Language Model（MLM）

在Transformer中，我们即想要知道上文的信息，又想要知道下文的信息，但同时要保证整个模型不知道要预测词的信息，那么就干脆不要告诉模型这个词的信息就可以了。也就是说，BERT在输入的句子中，挖掉一些需要预测的词，然后通过上下文来分析句子，最终使用其相应位置的输出来预测被挖掉的词。这其实就像是在做完形填空 (Cloze)一样。

但是，直接将大量的词替换为<MASK>标签可能会造成一些问题，模型可能会认为只需要预测<MASK>相应的输出就行，其他位置的输出就无所谓。同时Fine-Tuning阶段的输入数据中并没有<MASK>标签，也有数据分布不同的问题。为了减轻这样训练带来的影响，BERT采用了如下的方式：

1. 输入数据中随机选择15%的词用于预测，这15%的词中，
2. 80%的词向量输入时被替换为<MASK>
3. 10%的词的词向量在输入时被替换为其他词的词向量
4. 另外10%保持不动

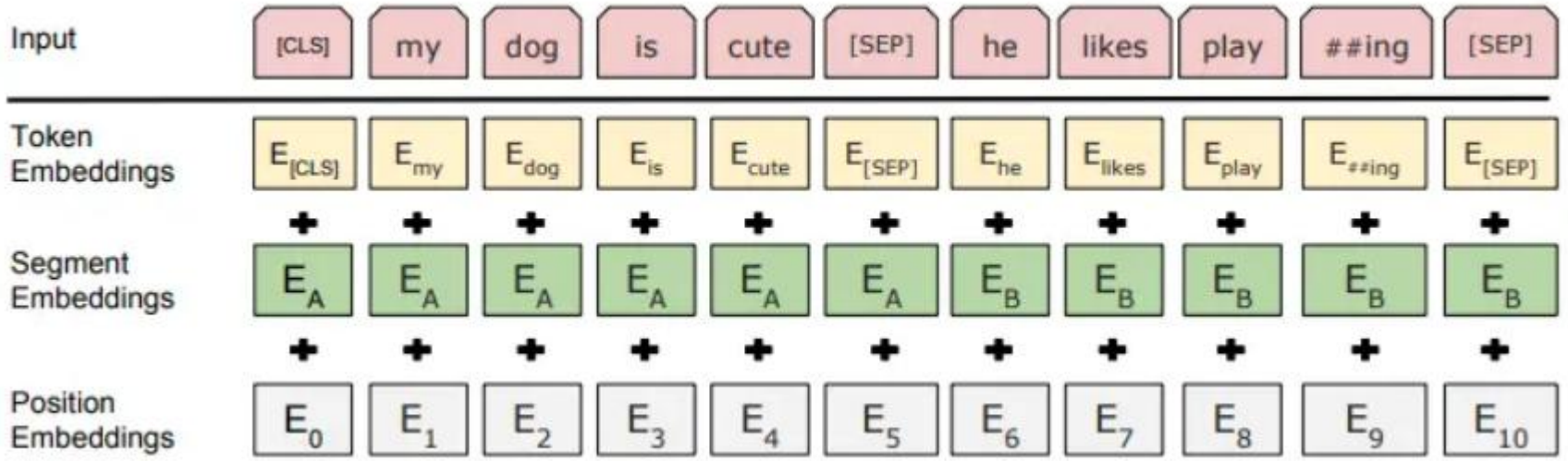
这样一来就相当于告诉模型，我可能给你答案，也可能不给你答案，也可能给你错误的答案，有<MASK>的地方我会检查你的答案，没<MASK>的地方我也可能检查你的答案，所以<MASK>标签对你来说没有什么特殊意义，所以无论如何，你都要好好预测所有位置的输出。

BERT（Bidirectional Encoder Representation from Transformers）

Next Sentence Prediction (NSP)

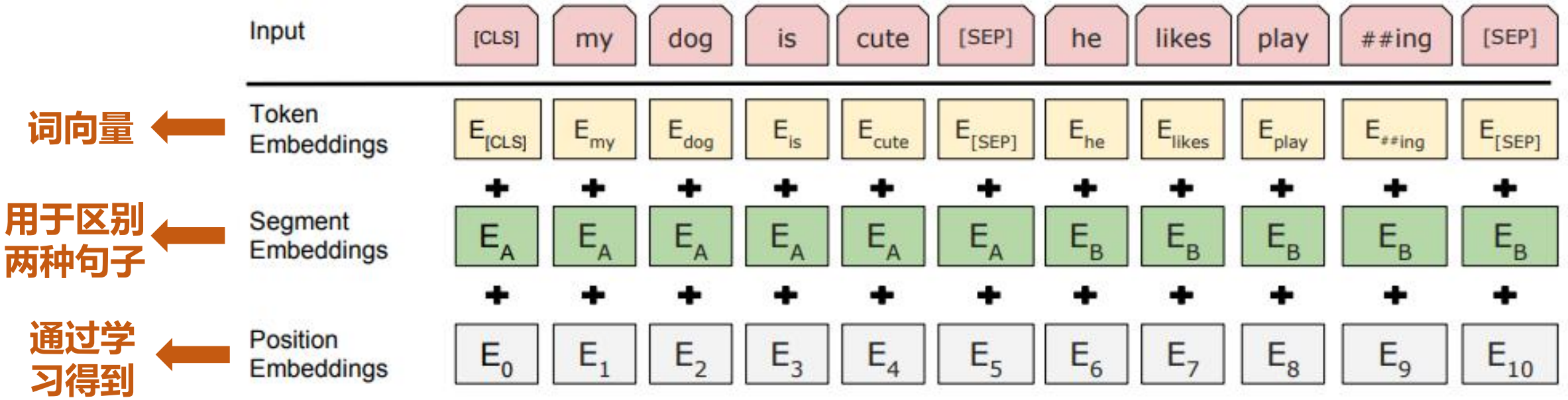
BERT还提出了另外一种预训练方式NSP，与MLM同时进行，组成多任务预训练。这种预训练的方式就是往Transformer中输入连续的两个句子，左边的句子前面加上一个<CLS>标签，它的输出被用来判断两个句子之间是否是连续上下文关系。采用负采样的方法，正负样本各占50%。

为了区分两个句子的前后关系，BERT除了加入了Positional Encoding之外，还两外加入了一个在预训练时需要学习的Segment Embedding来区分两个句子。这样一来，BERT的输入就由词向量、位置向量、段向量三个部分相加组成。此外，两个句子之间使用<SEP>标签予以区分。



超大规模预训练网络介绍-BERT

BERT的Embedding组成



第一步预训练 ➡ 做语言模型 ➡ 深度双向的模型

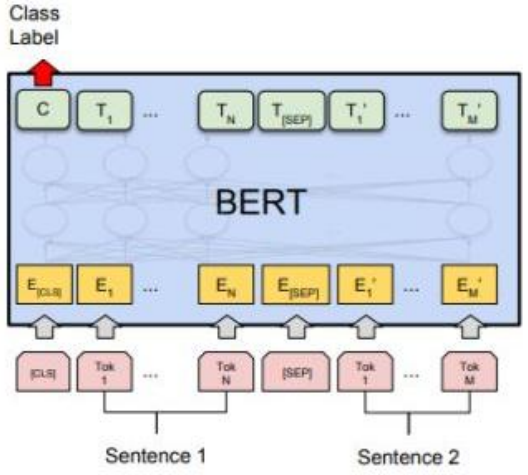
加mask的技巧 ➡ 训练过程中随机mask 15%的token ➡ 最终的损失函数只计算被mask掉的那个token

第二步预训练 ➡ 让模型理解两个句子之间的联系

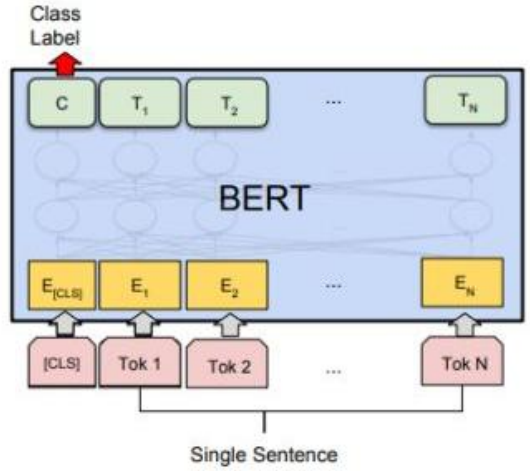
超大规模预训练网络介绍-BERT

其他任务的BERT参数调整

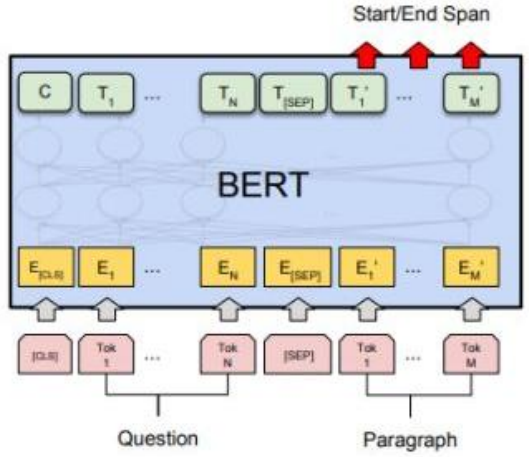
Fine-tunning阶段



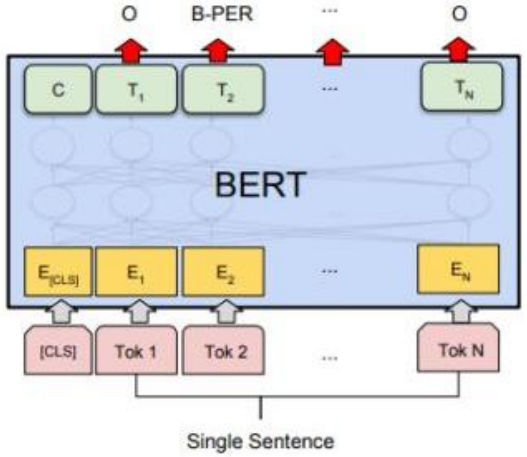
(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG



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



(c) Question Answering Tasks: SQuAD v1.1

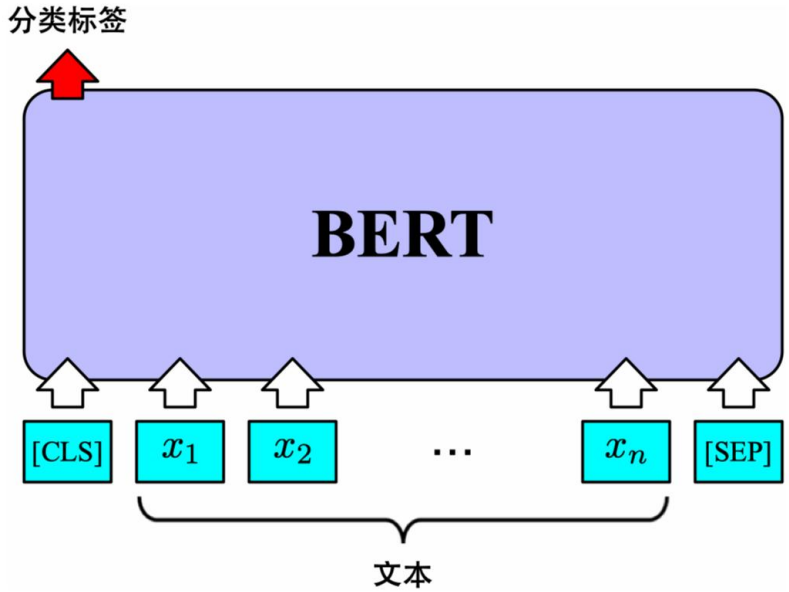


(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

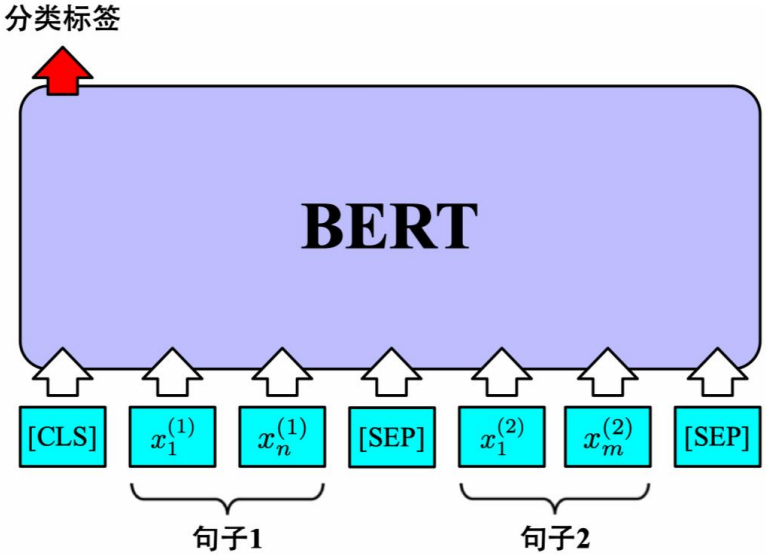
超大规模预训练网络介绍-BERT

下游任务应用

- 预训练语言模型应用
  - 单句文本分类
    - 最常见的NLP任务，需要将输入文本分成不同类别
    - 例如：情感分类等



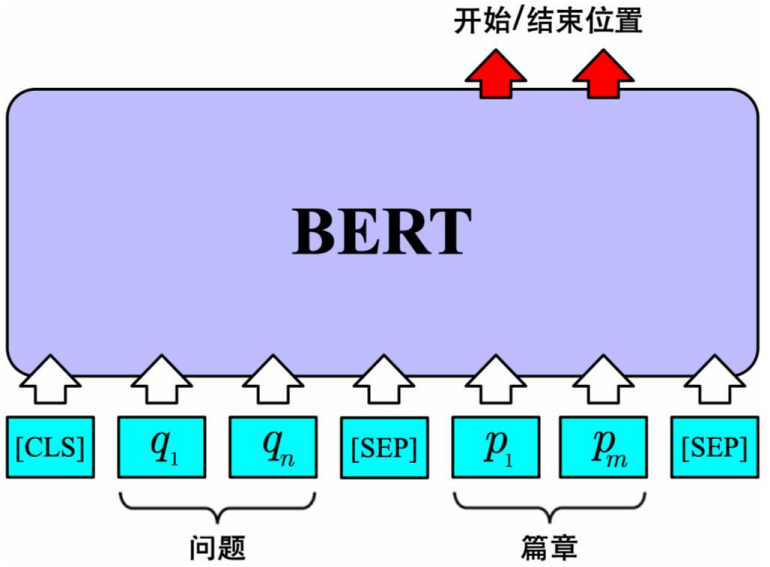
  - 句对文本分类
    - 与单句分类任务类似，需要将一对文本分成不同类别
    - 例如：文本蕴涵任务中，将句对分成“蕴涵”或“冲突”类别



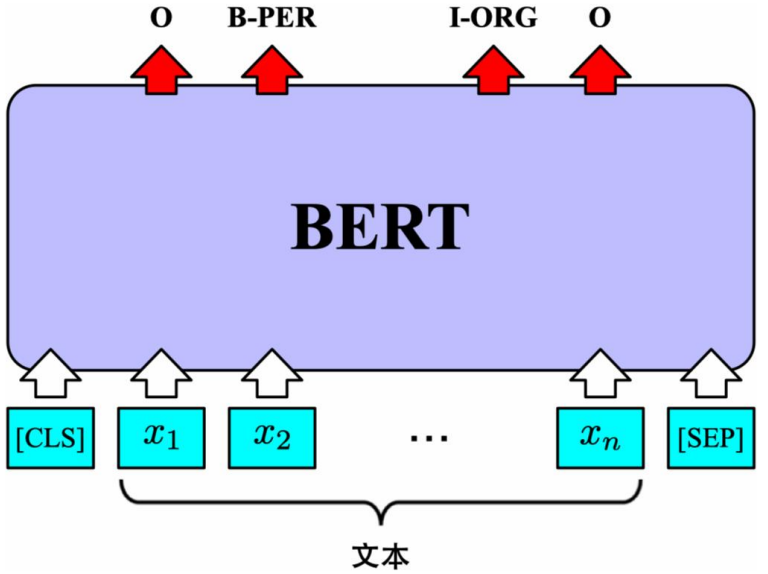
超大规模预训练网络介绍-BERT

下游任务应用

- 预训练语言模型应用
  - 阅读理解
    - 以抽取式阅读理解为例，要求机器在阅读篇章和问题后给出相应的答案，而答案要求是从篇章中抽取出一个文本片段（Span）



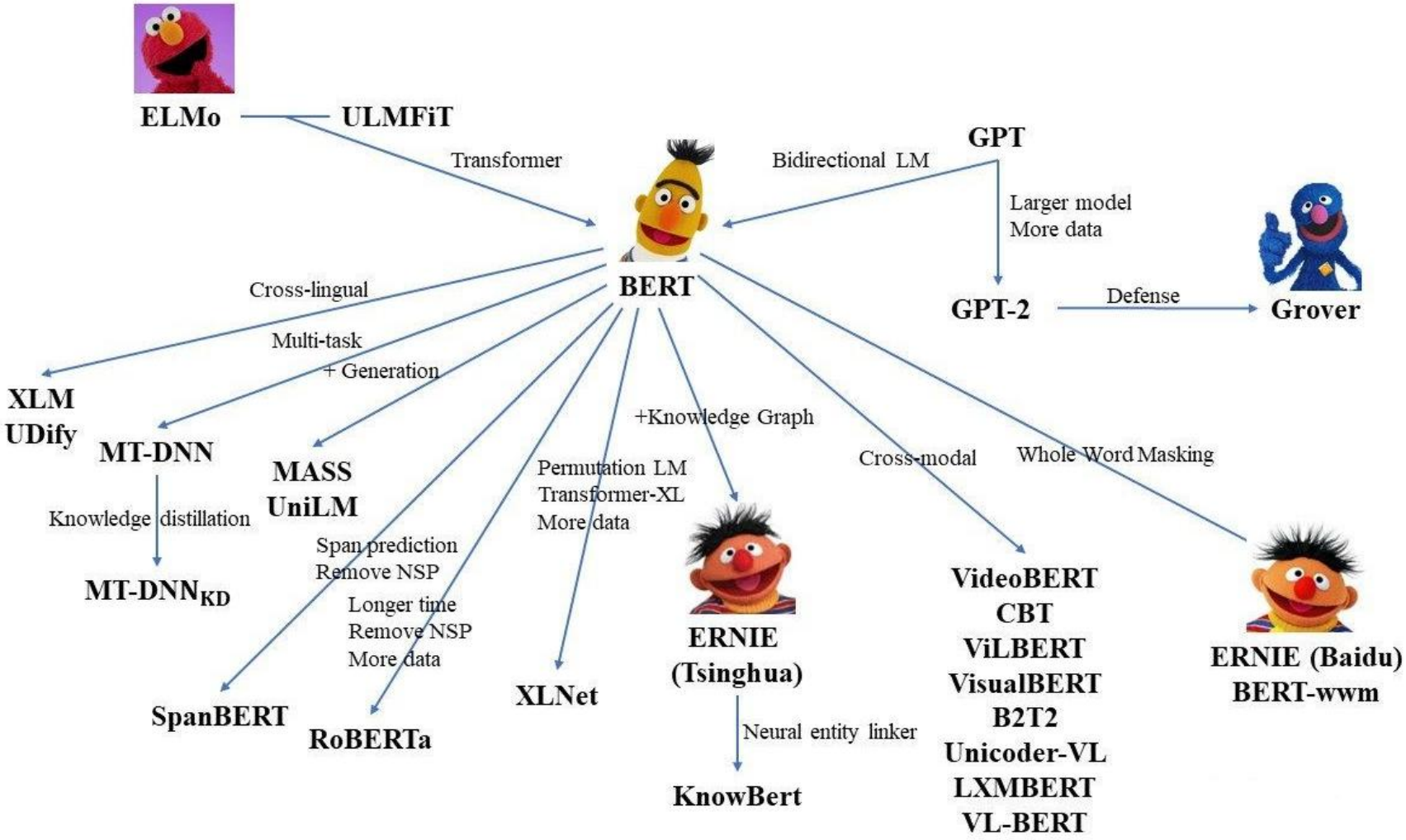
  - 序列标注
    - 以命名实体识别任务为例，对给定输入文本的每一个词输出一个标签，以此指定某个命名实体的边界信息



原始标签	B-PER	I-PER	O	O	O	O	B-LOC	
原始输入	John	Smith	has	never	been	to	Harbin	
处理后的标签	B-PER	I-PER	O	O	O	O	B-LOC	B-LOC
处理后的输入	John	Smith	has	never	been	to	Ha	##rbin

超大规模预训练网络介绍-BERT

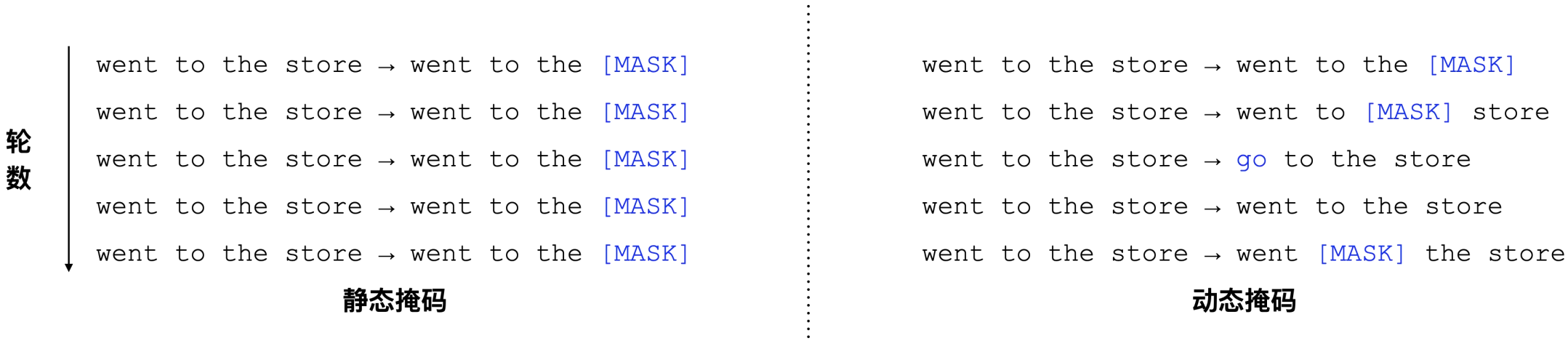
BERT和它的朋友们



超大规模预训练网络介绍-BERT

RoBERTa

- 在BERT的基础上引入了**动态掩码**技术，同时**舍弃了NSP任务**
- 采用了**更大规模的预训练数据**，并以**更大的批次**和BPE词表训练了**更多的步数**
- 动态掩码
  - 即决定掩码位置和方法是在模型的训练阶段实时计算的，这样能保证无论训练多少轮，都能够最大限度地保证同一段文本能够**在不同轮数下产生不同的掩码模型**



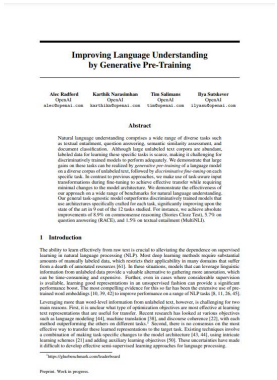
RoBERTa

- 舍弃NSP任务
  - 实验结果表明不使用NSP任务能够带来下游任务的性能提升
  - 最后采用了“跨文档整句输入”并舍弃了NSP任务的方案

实验设置	SQuAD 1.1	SQuAD 2.0	MNLI-m	SST-2	RACE
文本对输入 + NSP	90.4	78.7	84.0	92.9	64.2
句子对输入 + NSP	88.7	76.2	82.9	92.1	63.0
跨文档整句输入	90.4	79.1	84.7	92.5	64.8
文档内整句输入	90.6	79.7	84.7	92.7	65.6

大语言模型(Large Language Model, LLM)发展史

OpenAI


Improving Language Understanding by Generative Pre-Training

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever

Abstract

Natural language understanding comprises a wide range of diverse tasks such as textual entailment, question answering, semantic similarity assessment, and document classification. Although large unlabeled text corpora are abundant, labeled data for learning these specific tasks is scarce, making it challenging for discriminatively trained models to perform adequately. We demonstrate that large gains on these tasks can be realized by *generative pre-training* of a language model on a diverse corpus of unlabeled text, followed by *discriminative fine-tuning* on each specific task. In contrast to previous approaches, we make use of task-aware input transformations during fine-tuning to achieve effective transfer while requiring minimal changes to the model architecture. We demonstrate the effectiveness of our approach on a wide range of benchmarks for natural language understanding. Our general task-agnostic model outperforms discriminatively trained models that use architectures specifically crafted for each task, significantly improving upon the state of the art in 9 out of the 12 tasks studied. For instance, we achieve absolute improvements of 8.9% on commonsense reasoning (Stories Cloze Test), 5.7% on question answering (RACE), and 1.5% on textual entailment (MultiNLI).

1 Introduction

The ability to learn effectively from raw text is crucial to alleviating the dependence on supervised learning in natural language processing (NLP). Most deep learning methods require substantial amounts of manually labeled data, which restricts their applicability in many domains that suffer from a dearth of annotated resources [61]. In these situations, models that can leverage linguistic information from unlabeled data provide a valuable alternative to gathering more annotation, which can be time-consuming and expensive. Further, even in cases where considerable supervision is available, learning good representations in an unsupervised fashion can provide a significant performance boost. The most compelling evidence for this so far has been the extensive use of pre-trained word embeddings [10, 39, 42] to improve performance on a range of NLP tasks [8, 11, 26, 45].

Leveraging more than word-level information from unlabeled text, however, is challenging for two main reasons. First, it is unclear what type of optimization objectives are most effective at learning text representations that are useful for transfer. Recent research has looked at various objectives such as language modeling [44], machine translation [38], and discourse coherence [22], with each method outperforming the others on different tasks. Second, there is no consensus on the most effective way to transfer these learned representations to the target task. Existing techniques involve a combination of making task-specific changes to the model architecture [43, 44], using intricate learning schemes [21] and adding auxiliary learning objectives [50]. These uncertainties have made it difficult to develop effective semi-supervised learning approaches for language processing.

[1] http://gluebenchmark.com/leaderboard

Preprint. Work in progress.

GPT
06/2018


Language Models are Unsupervised Multitask Learners

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever

Abstract

Natural language processing tasks, such as question answering, machine translation, reading comprehension, and summarization, are typically approached with supervised learning on task-specific datasets. We demonstrate that language models begin to learn these tasks without any explicit supervision when trained on a new dataset of millions of webpages called WebText. When conditioned on a document plus questions, the answers generated by the language model reach 55 F1 on the CoQA dataset - matching or exceeding the performance of 3 out of 4 baseline systems without using the 127,000+ training examples. The capacity of the language model is essential to the success of zero-shot task transfer and increasing it improves performance in a log-linear fashion across tasks. Our largest model, GPT-2, is a 1.5B parameter Transformer that achieves state of the art results on 7 out of 8 tested language modeling datasets in a zero-shot setting but still underfits WebText. Samples from the model reflect these improvements and contain coherent paragraphs of text. These findings suggest a promising path towards building language processing systems which learn to perform tasks from their naturally occurring demonstrations.

1. Introduction

Machine learning systems now excel (in expectation) at tasks they are trained for by using a combination of large datasets, high-capacity models, and supervised learning (Krizhevsky et al., 2012) (Sutskever et al., 2014) (Amodei et al., 2016) but these systems are brittle and sensitive to slight changes to the data distribution (Recht et al., 2018) and task specification (Kirkpatrick et al., 2017). Current systems are better characterized as narrow experts rather than competent generalists. We would like to move towards more general systems which can perform many tasks – eventually without the need to manually create and label a training dataset for each one.

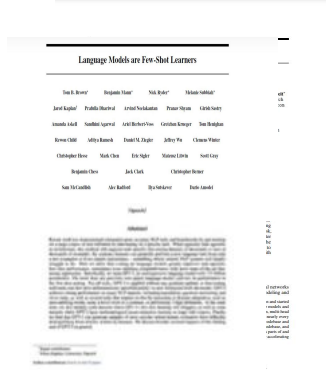
The dominant approach to creating ML systems is to collect a dataset of training examples demonstrating correct behavior for a desired task, train a system to imitate these behaviors, and then test its performance on independent and identically distributed (IID) held-out examples. This has served well to make progress on narrow experts. But the often erratic behavior of captioning models (Lake et al., 2017), reading comprehension systems (Jia & Liang, 2017), and image classifiers (Alcorn et al., 2018) on the diversity and variety of possible inputs highlights some of the shortcomings of this approach.

Our suspicion is that the prevalence of single task training on single domain datasets is a major contributor to the lack of generalization observed in current systems. Progress towards robust systems with current architectures is likely to require training and measuring performance on a wide range of domains and tasks. Recently, several benchmarks have been proposed such as GLUE (Wang et al., 2018) and decaNLP (McCann et al., 2018) to begin studying this.

Multitask learning (Caruana, 1997) is a promising framework for improving general performance. However, multitask training in NLP is still nascent. Recent work reports modest performance improvements (Yogatama et al., 2019) and the two most ambitious efforts to date have trained on a total of 10 and 17 (task, dataset) pairs respectively (McCann et al., 2018) (Bowman et al., 2018). From a meta-learning perspective, each (dataset, objective) pair is a single training example sampled from the distribution of datasets and objectives. Current ML systems need hundreds to thousands of examples to induce functions which generalize well. This suggests that multitask training many need just as many effective training pairs to realize its promise with current approaches. It will be very difficult to continue to scale the creation of datasets and the design of objectives to the degree that may be required to brute force our way there with current techniques. This motivates exploring additional setups for performing multitask learning.

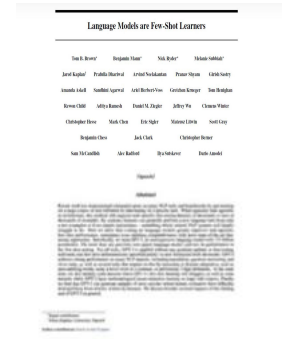
The current best performing systems on language tasks

GPT-2
02/2019


Language Models are Few-Shot Learners

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, Dario Amodei

GPT-3
05/2020


Language Models are Few-Shot Learners

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, Dario Amodei

GPT-4
05/2023

Transformer
06/2017


Attention Is All You Need

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin

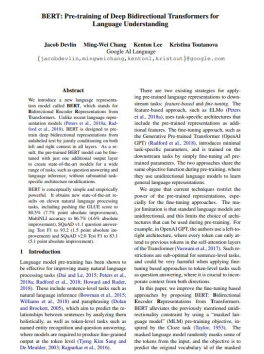
Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

1 Introduction

Recurrent neural networks, long short-term memory [13] and gated recurrent [7] neural networks in particular, have been firmly established as state of the art approaches in sequence modeling and

BERT
10/2018


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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova

Abstract

We introduce a new language representation model called **BERT**, which stands for Bidirectional Encoder Representations from Transformers. Unlike recent language representation models (Peters et al., 2018a; Radford et al., 2018), BERT is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task-specific architecture modifications.

BERT is conceptually simple and empirically powerful. It obtains new state-of-the-art results on eleven natural language processing tasks, including pushing the GLUE score to 80.5% (7.7% point absolute improvement), MultiNLI accuracy to 86.7% (4.6% absolute improvement), SQuAD v1.1 question answering Test F1 to 93.2 (1.5 point absolute improvement) and SQuAD v2.0 Test F1 to 83.1 (5.1 point absolute improvement).

1 Introduction

Language model pre-training has been shown to be effective for improving many natural language processing tasks (Dai and Le, 2015; Peters et al., 2018a; Radford et al., 2018; Howard and Ruder, 2018). These include sentence-level tasks such as natural language inference (Bowman et al., 2015; Williams et al., 2018) and paraphrasing (Dolan and Brockett, 2005), which aim to predict the relationships between sentences by analyzing them holistically, as well as token-level tasks such as named entity recognition and question answering, where models are required to produce fine-grained output at the token level (Tjong Kim Sang and De Meulder, 2003; Rajpurkar et al., 2016).

There are two existing strategies for applying pre-trained language representations to downstream tasks: *feature-based* and *fine-tuning*. The feature-based approach, such as ELMo (Peters et al., 2018a), uses task-specific architectures that include the pre-trained representations as additional features. The fine-tuning approach, such as the Generative Pre-trained Transformer (OpenAI GPT) (Radford et al., 2018), introduces minimal task-specific parameters, and is trained on the downstream tasks by simply fine-tuning all pre-trained parameters. The two approaches share the same objective function during pre-training, where they use unidirectional language models to learn general language representations.

We argue that current techniques restrict the power of the pre-trained representations, especially for the fine-tuning approaches. The major limitation is that standard language models are unidirectional, and this limits the choice of architectures that can be used during pre-training. For example, in OpenAI GPT, the authors use a left-to-right architecture, where every token can only attend to previous tokens in the self-attention layers of the Transformer (Vaswani et al., 2017). Such restrictions are sub-optimal for sentence-level tasks, and could be very harmful when applying fine-tuning based approaches to token-level tasks such as question answering, where it is crucial to incorporate context from both directions.

In this paper, we improve the fine-tuning based approaches by proposing BERT: Bidirectional Encoder Representations from Transformers. BERT alleviates the previously mentioned unidirectionality constraint by using a "masked language model" (MLM) pre-training objective, inspired by the Cloze task (Taylor, 1953). The masked language model randomly masks some of the tokens from the input, and the objective is to predict the original vocabulary id of the masked

- 大语言模型的“三大”特点：
  - 大数据
  - 大模型
  - 大算力

GPT-2

GPT-2继续沿用了原来在GPT种使用的单向Transformer模型，目的就是尽可能利用单向Transformer的优势，做一些BERT使用的双向Transformer所做不到的事。那就是通过上文生成下文文本。

GPT-2的想法就是**完全舍弃Fine-Tuning过程**，转而使用一个容量更大、无监督训练、更加通用的语言模型来完成各种各样的任务。我们完全不需要去定义这个模型应该做什么任务，因为很多标签所蕴含的信息，就存在于语料当中。就像一个人如果博览群书，自然可以根据看过的内容轻松的做到自动摘要、问答、续写文章这些事。

严格来说GPT-2可能不算是一个多任务模型，但是它确实使用相同的模型、相同的参数完成了不同的任务。**那么GPT-2是怎么使用语言模型完成多种任务的呢？**

主流超大规模预训练网络介绍-GPT-2

GPT-2 多任务学习

通常我们针对特定任务训练的专用模型，给定输入，就可以返回这个任务相应的输出，也就是

$$p(output \mid input)$$

那么如果我们希望设计一个通用的模型，这个模型在给定输入的同时还需要给定任务类型，然后根据给定输入与任务来做出相应的输出，那么模型就可以表示成下面这个样子

$$p(output \mid input, task)$$

就好像原来我需要翻译一个句子，需要专门设计一个翻译模型，想要问答系统需要专门设计一个问答模型。但是如果一个模型足够聪明，并且能够根据你的上文生成下文，那我们就可以通过在输入中加入一些标识符就可以区分各种问题。比如可以直接问他：（‘自然语言处理', 中文翻译)来得到我们需要的结果Nature Language Processing。在我的理解中GPT-2更像是一个无所不知的问答系统，通过告知一个给定任务的标识符，就可以对多种领域的问答、多种任务做出合适的回答。GPT-2满足零样本设置 (zero-shot setting)， 在训练的过程中不需要告诉他应该完成什么样的任务，预测是也能给出较为合理的回答。

主流超大规模预训练网络介绍-GPT-2

性能稳定、优异。

有着超大规模的在海量数据集上训练的基于Transformer的巨大模型

模型与只带有解码器的Transformer模型类似

GPT-2 → GPT-2将Transformer堆叠的层数增加到48层，隐层的维度为1600，参数量达到了15亿。

GPT-2将词汇表提升到50257。

处理最长单词序列数由GPT的512扩展到1024。

此外还对Transformer做出了小调整。

GPT-2中Transformer模块的处理方式

每个Transformer模块的处理方式都是一样的，但每个模块都会维护自己的自注意力层和神经网络层中的权重

通过自注意力层处理

↓

将其传递给神经网络层

↓

模块处理完单词后，将结果向量传入堆栈中的下一个Transformer模块，继续进行计算

↓

⋮

↓

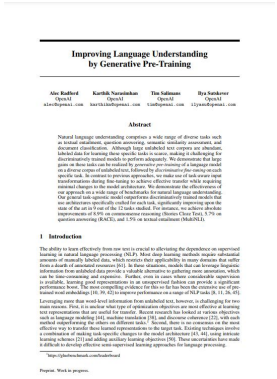
最后一个Transformer模块产生输出后，模型会将输出的向量乘上嵌入矩阵得到词汇表中每个单词对应的注意力得分

GPT-2主要贡献

GPT-2的最大贡献是验证了通过海量数据和大量参数训练出来的词向量模型有迁移到其它类别任务中而不需要额外的训练。但是很多实验也表明，GPT-2的无监督学习的能力还有很大的提升空间，甚至在有些任务上的表现不比随机的好。尽管在有些zero-shot的任务上的表现不错，但是我们仍不清楚GPT-2的这种策略究竟能做成什么样子。GPT-2表明随着模型容量和数据量的增大，其潜能还有进一步开发的空间，基于这个思想，诞生了我们下面要介绍的GPT-3。

大语言模型(Large Language Model, LLM)发展史

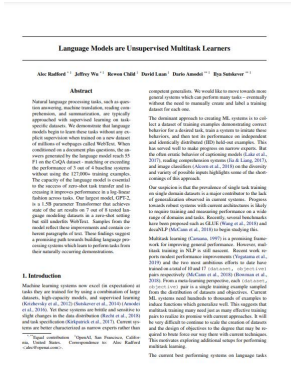
OpenAI


Improving Language Understanding by Generative Pre-Training

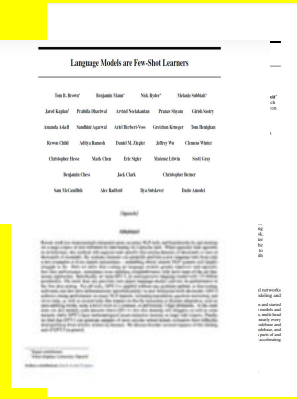
Transformer 06/2017

GPT 06/2018


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


Language Models are Unsupervised Multitask Learners

GPT-2 02/2019


Language Models are Few-Shot Learners

GPT-3 05/2020


Language Models are Few-Shot Learners

GPT-4 05/2023

- 大语言模型的“三大”特点：
  - 大数据
  - 大模型
  - 大算力

主流超大规模预训练网络介绍-GPT-3

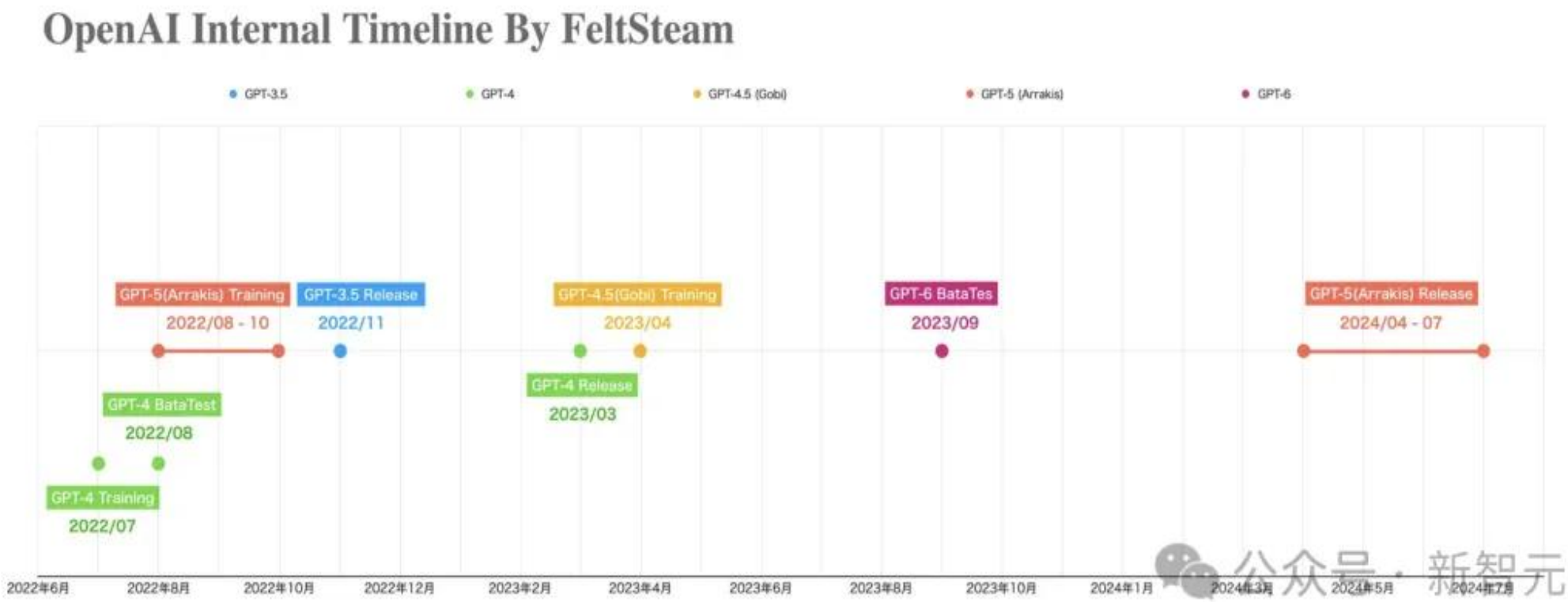
GPT-3

模型	发布时间	参数量	预训练数据量
GPT	2018 年 6 月	1.17 亿	约 5GB
GPT-2	2019 年 2 月	15 亿	40GB
GPT-3	2020 年 5 月	1,750 亿	45TB

GPT-3沿用GPT-2的网络结构。为了使模型在少样本的任务中尽快收敛，提出In-context learning的概念，核心思想在于通过少量的数据寻找一个合适的初始化范围，使得模型能够在有限的数据集上快速拟合，并获得不错的效果。

主流超大规模预训练网络介绍

GPT-4, 4o, o1 ...



主流超大规模预训练网络介绍

通往AGI

GPT-6也被电力卡脖子了——部署十万个H100时，整个电网发生了崩溃！

就在刚刚，微软工程师爆料，10万个H100基建正在紧锣密鼓地建设中，目的就是训练GPT-6。

微软工程师吐槽说，团队在部署跨区域GPU间的infiniband级别链接时遇到了困难。

Corbitt：为何不考虑直接将所有设备部署在同一个地区呢？

微软工程师：这确实是我们最初的方案。但问题是，一旦我们在同一个州部署超过100,000个H100 GPU，电网就会因无法负荷而崩溃。

1

超大规模预训练网络

预训练技术，BERT，GPT-2

2

模型压缩方法

模型剪枝，模型量化，模型蒸馏

3

其它热门的研究点

问答系统，机器阅读理解

4

多模态任务的举例与现状

多模态学习，图像-文本多模态任务

模型压缩方法

复杂的模型同时带来了高额的存储空间及计算资源消耗，使其较难落实到各个硬件平台。为了解决这些问题，需对模型进行模型压缩以最大限度地减小模型对计算空间和时间的消耗。

模型剪枝

基于的假设（共识） → **DNN的过参数化（Over-parameterization）**

DNN的过参数化：指训练阶段网络需要大量的参数来捕捉数据中的微小信息，而当训练完成并进入预测阶段后，网络通常并不需要这么多的参数。

基本思想 → **剪裁最不重要的部分**

思路
- **贪心法**
- **考虑参数裁剪对损失的影响**
- **考虑对特征输出的可重建性的影响**

模型剪枝

贪心法（saliency-based方法）→ 按重要性进行排序，之后将不重要的部分去除

magnitude-based weight pruning方法 → 按参数（或特征输出）绝对值大小来评估重要性，然后用贪心法对重要性较低的部分进行剪枝

结构化剪枝 → 用Group LASSO算法来得到结构化的稀疏权重

缺点！ → 忽略了参数间的相互关系而只能找到局部最优解

考虑参数裁剪对损失的影响

考虑对特征输出的可重建性的影响 → 最小化裁剪后网络对于特征输出的重建误差

基于的认知 → 如果对当前层进行裁剪后对后面的输出没有较大影响，则说明裁掉的是不太重要的信息

模型量化

量化模型（Quantized Model）→ 在硬件上移植非常方便

→ 模型加速（Model Acceleration）方法中其中一类方法的总称

→ 包括二值化网络（Binary Network）、三值化网络（Ternary Network），深度压缩（Deep Compression）等

→ 通往高速神经网络最佳的方法，但仍面临实现难度大、准确性不稳定，使用门槛较高的多方面问题

量化 → 一种权值共享的策略

量化后的权值张量 → 一个高度稀疏的有较多共享权值的矩阵

对于非零参数 → 定点压缩 → 更高的压缩率

e.g. Deepcompression

模型蒸馏

采用的方法 → **迁移学习**

主要思想 → **通过预先训练好的复杂模型（Teacher model）的最后输出结果来作为先验知识，结合One-Hot label数据，共同指导一个简单的网络（Student model）学习**

目标 → **让student学习到teacher的泛化能力**

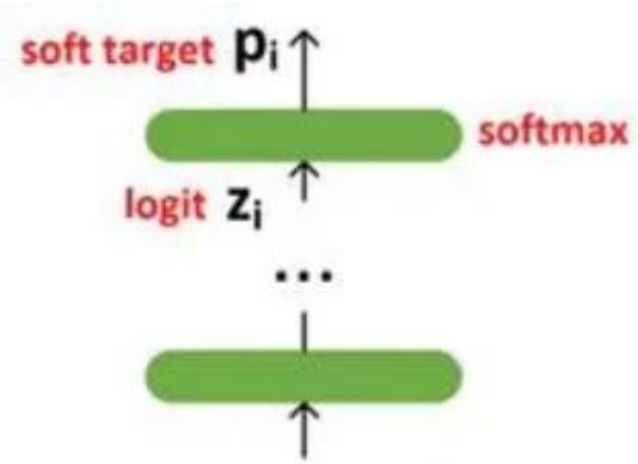
常用名词解释

teacher	原始模型或模型ensemble
student	新模型
transfer set	用来迁移teacher知识、训练student的数据集合
soft target	teacher输出的预测结果（一般是Softmax之后的概率）
hard target	样本原本的标签
temperature	蒸馏目标函数中的超参数
born-again network	蒸馏的一种，指student和teacher的结构和尺寸完全一样
teacher annealing	防止student的表现被teacher限制，在蒸馏时逐渐减少soft targets的权重

模型蒸馏

原始模型训练阶段 → **根据目标问题，设计一个大模型或者多个模型集合（N1，N2,...,Nt）即teacher，然后并行训练集合中的网络**

精简模型训练阶段



$$q_i = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)}$$

z_i：Logit　　例如：SoftMax层的输入

q_i：SoftMax层计算的分类概率

T：温度，通常被设为1

精简模型训练阶段

设计一个简单网络N0即student并收集简单模型训练数据

↓

将收集到的样本输入原始模型（N1,N2,...,Nt），修改原始模型teacher的Softmax层中温度参数T为一个较大值

↓

选取每一个样本在每个原始模型得到其最终的分类概率向量中概率值最大的即为该模型对于当前样本的判定结果

↓

对所有概率向量求取均值作为当前样本最后的概率输出向量并保存

↓

标签融合前面收集到的数据定义为样本原本的标签，即hard_target

↓

根据公式Target=a*hard_target+b*soft_target（a+b=1），Target最终作为训练数据的标签去训练精简模型

1 超大规模预训练网络

预训练技术，BERT，GPT-2

2 模型压缩方法

模型剪枝，模型量化，模型蒸馏

3 其它热门的研究点

问答系统，机器阅读理解

4 多模态任务的举例与现状

多模态学习，图像-文本多模态任务

其他热门的研究点

自然语言处理的具体表现形式包括机器翻译、文本摘要、文本分类、文本校对、信息抽取、语音合成、语音识别等。自然语言处理就是要计算机理解自然语言，自然语言处理机制涉及两个流程，包括自然语言理解和自然语言生成。

热门研究点介绍-问答系统

问答系统 → **信息检索系统的一种高级形式，通过Web搜索或链接知识库等方式，检索到用户问题的答案，并用准确、简洁的自然语言回答用户**

分类-根据其问题答案的数据来源和回答的方式

- **基于Web信息检索的问答系统（Web Question Answering, WebQA）**
- **基于知识库的问答系统（Knowledge Based Question Answering, KBQA）**
- **社区问答系统（Community Question Answering, CQA）**

热门研究点介绍-问答系统

分类-根据其问题答案的数据来源和回答的方式

基于Web信息检索的问答系统（Web Question Answering, WebQA）

- 以搜索引擎为支撑，理解分析用户的问题意图后并在全网范围内搜索相关答案反馈给用户

  e.g. 早期的 Ask Jeeves 和 AnswerBus 问答系统

基于知识库的问答系统（Knowledge Based Question Answering, KBQA）

- 通过结合一些已有的知识库或数据库资源，以及利用非结构化文本的信息，使用信息抽取的方法提取有价值的信息，并构建知识图谱作为问答系统的后台支撑，再结合知识推理等方法为用户提供更深层次语义理解的答案

社区问答系统（Community Question Answering, CQA）

- 基于社交媒体的问答系统，大多数问题的答案由网友提供，问答系统会检索社交媒体中与用户提问语义相似的问题，并将答案返回给用户

热门研究点介绍-问答系统

KBQA → 目前应用最广泛的问答系统之一

知识库（Knowledge Base, KB）→ 用于知识管理的一种特殊的数据库，用于相关领域知识的采集、整理及提取

- **表示形式是一个对象模型（object model），通常称为本体，包含一些类、子类和实体**
- **常见的知识库有Freebase、DBPedia等**
- **一般采用RDF格式对其中的知识进行表示，知识的查询主要采用RDF标准查询语言SPARQL**

热门研究点介绍-问答系统

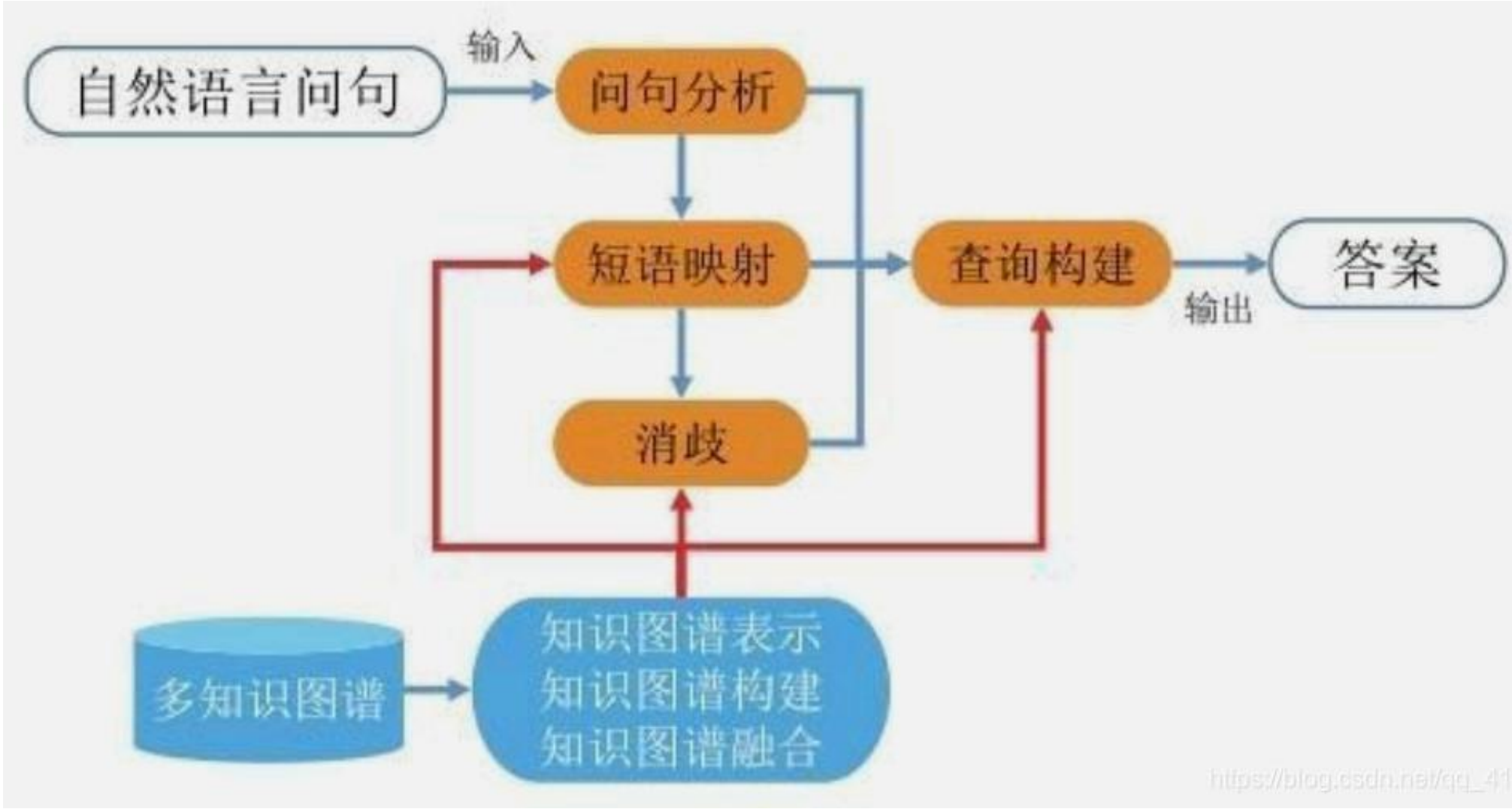
KBQA → 目前应用最广泛的问答系统之一

知识库（Knowledge Base, KB）→ 用于知识管理的一种特殊的数据库，用于相关领域知识的采集、整理及提取

- **表示形式是一个对象模型（object model），通常称为本体，包含一些类、子类和实体**
- **常见的知识库有Freebase、DBPedia等**
- **一般采用RDF格式对其中的知识进行表示，知识的查询主要采用RDF标准查询语言SPARQL**

热门研究点介绍-问答系统

KBQA基本架构 ➡ 包含问句理解、答案信息抽取、答案排序和生成等核心模块



热门研究点介绍-机器阅读理解

机器阅读理解（Machine Reading Comprehension, MRC）

→ 给定一篇文章以及基于文章的一个问题，让机器在阅读文章后对问题进行作答

常见任务主要

- **完形填空** → 将文章中的某些单词隐去，让模型根据上下文判断被隐去的单词最可能是哪个
- **多项选择** → 给定一篇文章和一个问题，让模型从多个备选答案中选择一个或多个最有可能是正确答案的选项
- **片段抽取** → 给定一篇文章和一个问题，让模型从文章中抽取连续的单词序列，并使得该序列尽可能的作为该问题的答案
- **自由作答** → 给定一篇文章和一个问题，让模型生成一个单词序列，并使得该序列尽可能的作为该问题的答案

经典机器阅读理解的基本框架

- 嵌入编码（Embedding）
- 特征抽取（Feature Extraction、Encode）
- 文章-问题交互（Context-Question Interaction）
- 答案预测（Answer Prediction）

热门研究点介绍-机器阅读理解

研究趋势

- 基于外部知识的机器阅读理解 → 相关外部知识的检索及外部知识的融合
- 带有不能回答的问题的机器阅读理解 → 不能回答的问题的判别、干扰答案的识别等
- 多条文档机器阅读理解 → 相关文档的检索、噪声文档的干扰、检索得到的文档中没有答案、可能存在多个答案、需要对多条线索进行聚合等
- 对话式阅读理解 → 对话历史信息的利用、指代消解等

1 超大规模预训练网络

预训练技术，BERT，GPT-2

2 模型压缩方法

模型剪枝，模型量化，模型蒸馏

3 其它热门的研究点

问答系统，机器阅读理解

4 多模态任务的举例与现状

多模态学习，图像-文本多模态任务

多模态任务的举例与现状

每一种信息的来源或者形式，都可以称为一种模态。例如，人有触觉，听觉，视觉等；信息的媒介，有语音、视频、文字等。以上的每一种都可以称为一种模态。多模态机器学习旨在通过机器学习的方法实现处理和理解多源模态信息的能力。目前比较热门的研究方向是图像、视频、音频、语义之间的多模态学习。

研究方向

- 多模态表示学习
- 模态转化
- 对齐
- 多模态融合
- 协同学习

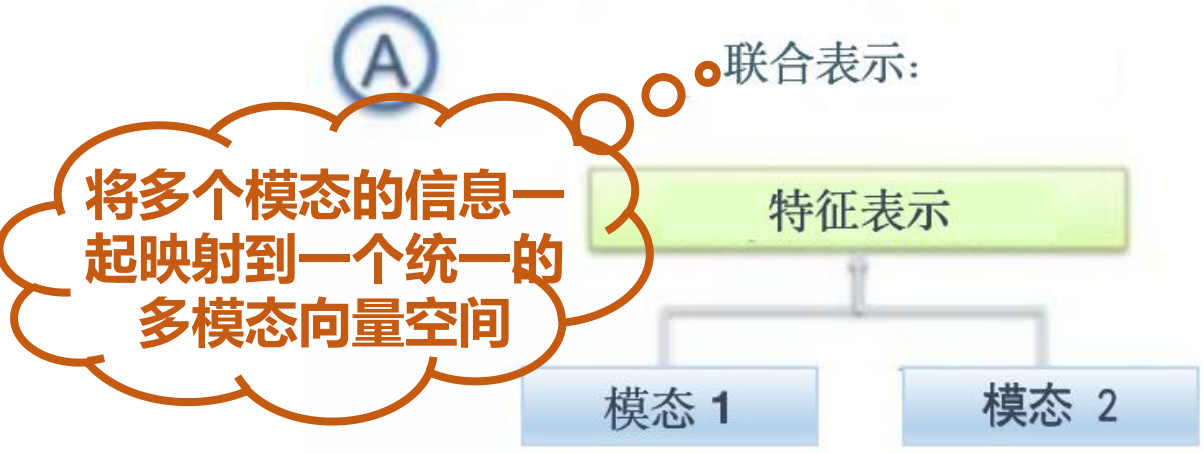
多模态学习的概念

多模态表示学习 → **利用多模态之间的互补性，剔除模态间的冗余性，从而学习到更好的特征表示**

两大研究方向

- **联合表示（Joint Representations）**
- **协同表示（Coordinated Representations）**

将多模态中的每个模态分别映射到各自的表示空间，但映射后的向量之间满足一定的相关性约束



将多个模态的信息一起映射到一个统一的多模态向量空间

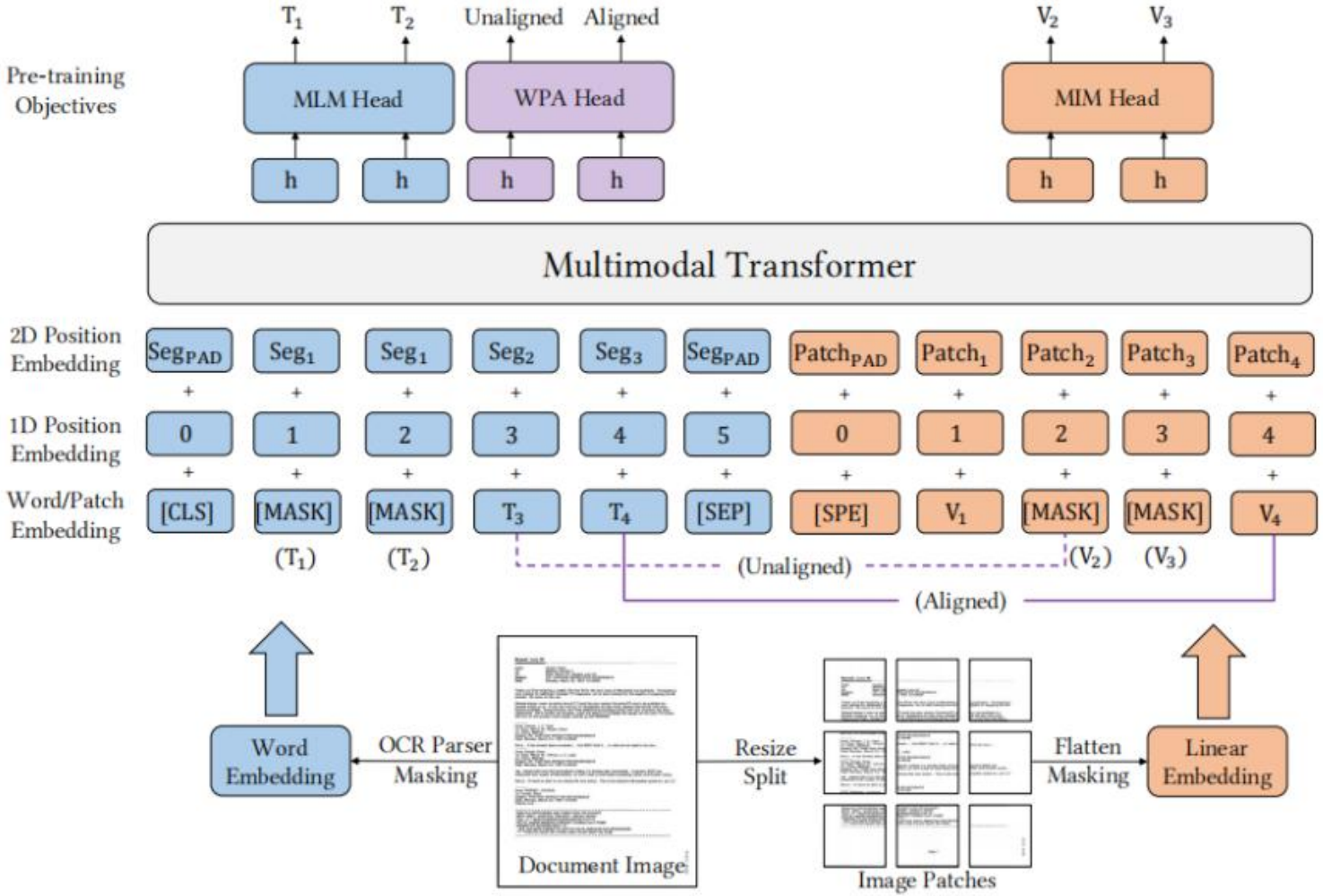
多模态学习的概念

Document AI

海量纸质数据需要电子化存档，通用的做法是先将纸质文件进行扫描，得到扫描文本图像，将文本图像进行OCR识别（计算机视觉，CV），得到文本信息，再对文本信息进行信息抽取和结构化存储（自然语言处理，NLP）。该任务在教育、金融、医疗等领域均有广泛的应用价值，因而得到了广泛的关注和研究，并形成了大数据、人工智能领域一个热点研究方向：文档智能（Document AI）。微软亚洲研究院在该领域提出了文档多模态预训练模型LayoutLM系列，得到了广泛的关注。利用多模态预训练模型LayoutLM可对多种文档领域下游任务带来提升，比如文档分类，版面分析，信息提取，文档VQA（视觉问答）等。

多模态学习的概念

Document AI



多模态模型LayoutLM，处理图像、文本和版式信息。

多模态学习的概念

模态转化（映射） → **将一个模态的信息转换为另一个模态的信息**

模态间转换的难点
- **未知结束位**
- **主观评判性** → **很多模态转换问题的效果没有一个比较客观的评判标准，目标函数的确定是非常主观的**

对齐 → **对来自同一个实例的不同模态信息的子分支/元素寻找对应关系**

多模态学习的概念

多模态融合 → **联合多个模态的信息，进行目标预测（分类或者回归）**

按照融合的层次分类

- **pixel level**
- **feature level**
  - **early**
  - **late**
- **decision level**

代表了融合发生在特征抽取的早期和晚期

分别对应对原始数据进行融合、对抽象的特征进行融合和对决策结果进行融合

协同学习 → **使用一个资源丰富的模态信息来辅助另一个资源相对贫瘠的模态进行学习**

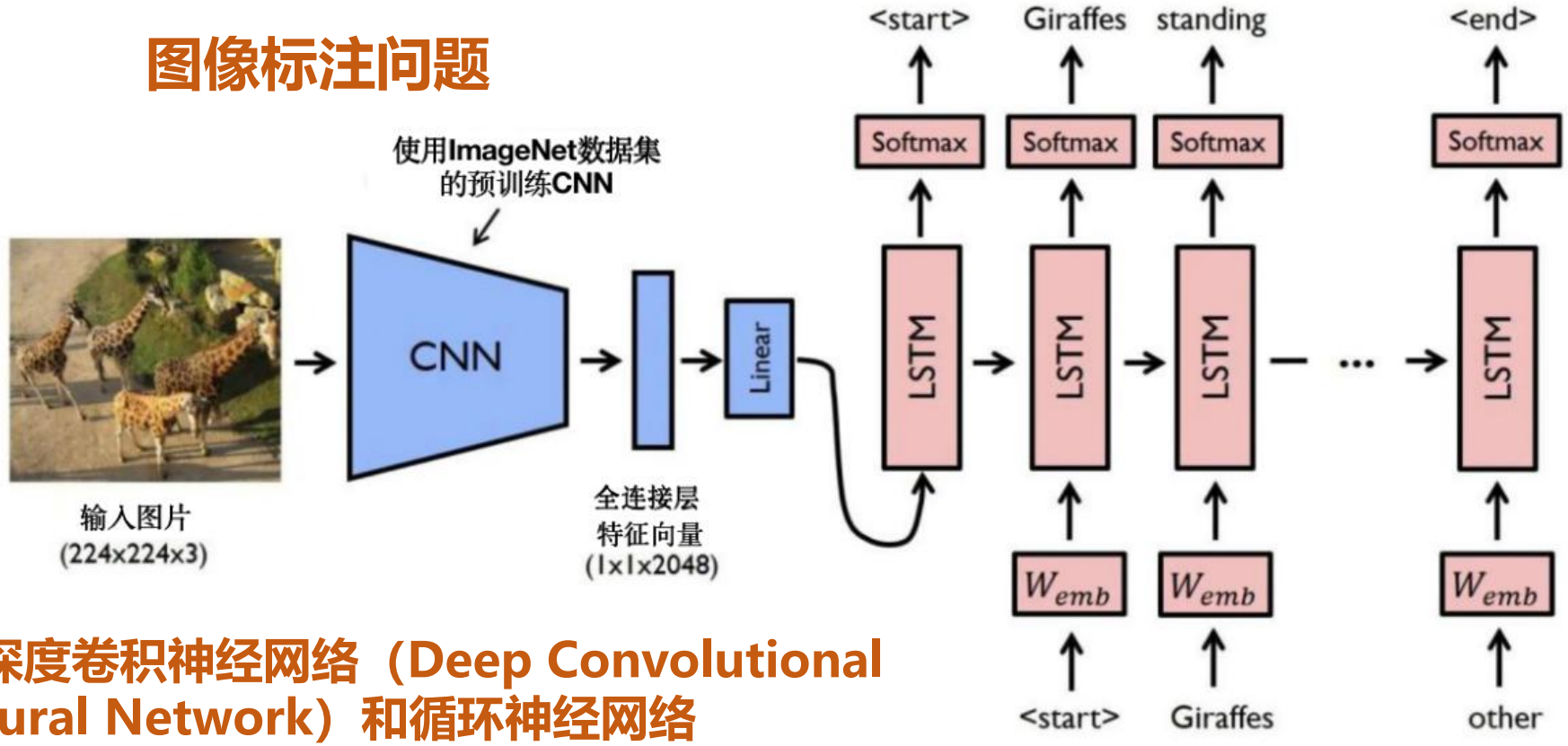
e.g. 迁移学习

图像-文本多模态任务举例及研究现状

图像描述（Image Caption）⟶ 融合计算机视觉、自然语言处理和机器学习的综合问题

图像描述主体网络结构

图像标注问题



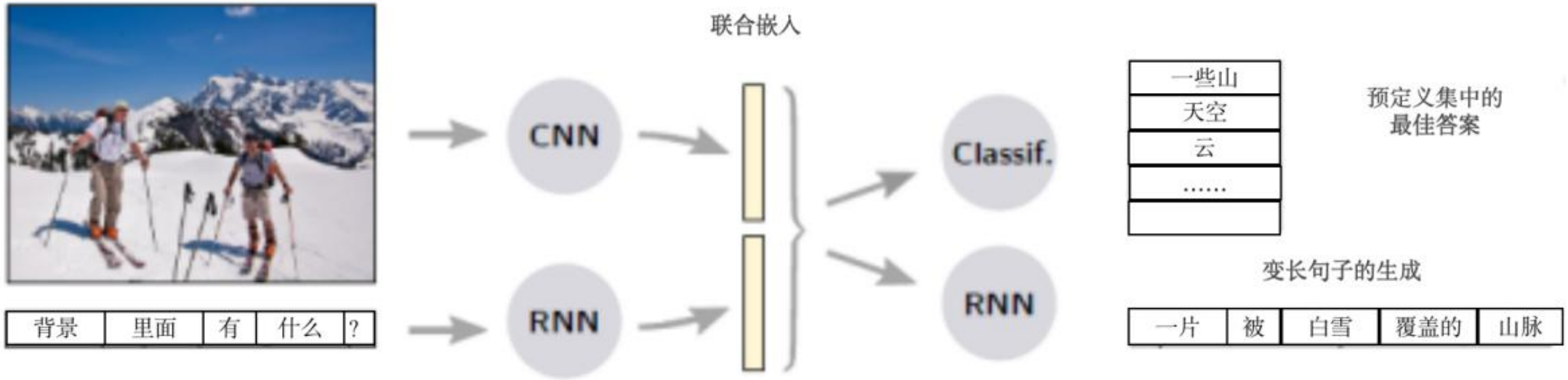
将深度卷积神经网络（Deep Convolutional Neural Network）和循环神经网络（Recurrent Neural Network）结合

图像-文本多模态任务举例及研究现状

视觉问答（VQA）⟶ 给定一张图片和一个与该图片相关的自然语言问题，计算机能产生一个正确的回答

融合CV与NLP的技术 ⟶ 计算机需要同时学会理解图像和文字

融合嵌入方法 ⟶ 对图像和问题进行联合编码



图像-文本多模态任务举例及研究现状

注意力机制 → 源于机器翻译问题

→ 让模型动态地调整对输入项各部分的关注度，从而提升模型的“专注力”

复合模型 → 设计一种模块化的模型，可根据问题的类型动态组装模块来产生答案

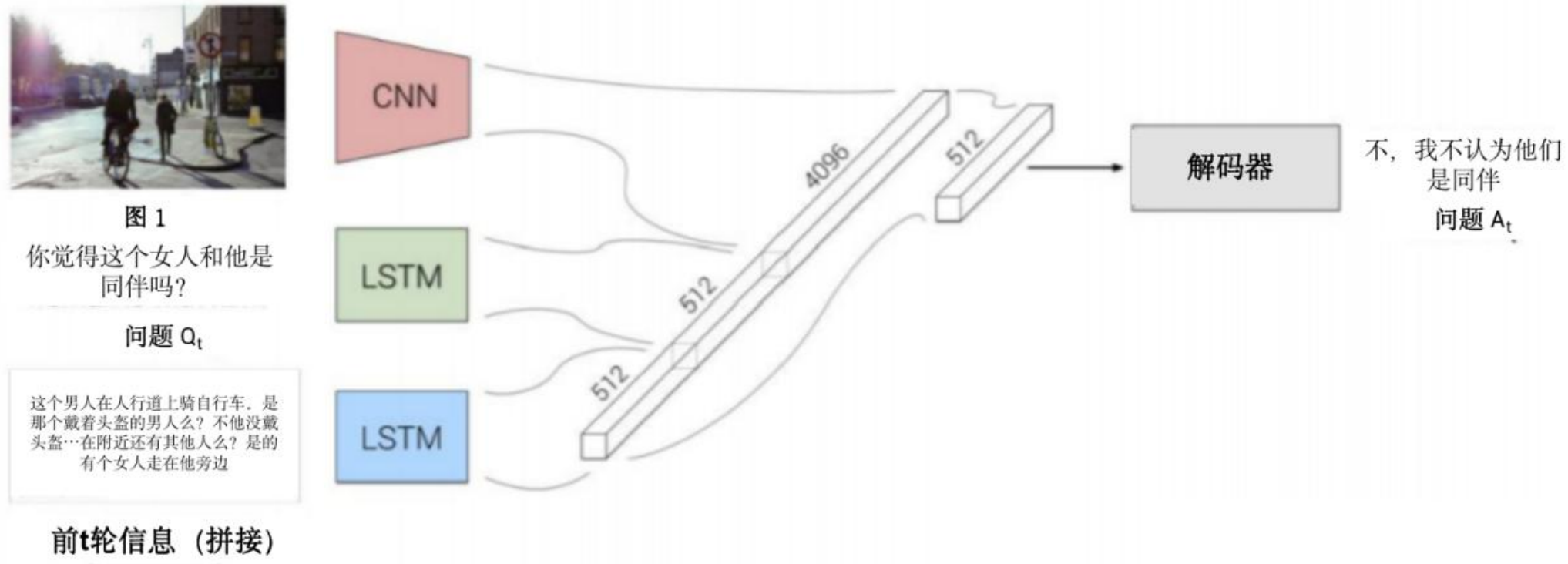
使用先验知识库方法

图像-文本多模态任务举例及研究现状

视觉对话 ⟶ AI代理与人类以自然的会话语言对视觉内容进行有意义的对话

具有访问和理解的多轮对话历史 ⟶ 需要一个可以组合多个信息源的编码器

主体网络结构



本章总结

CHAPTER SUMMARY

本章总结

本章首先介绍了超大规模预训练网络的基本原理及典型网络代表，之后从模型剪枝、模型量化及模型蒸馏三个方面讲解了主流模型压缩方法，然后从理解与生成两个方面，汇总介绍了其它热门研究点并介绍了问答系统与机器阅读理解等任务，最后在多模态学习概念的基础之上，介绍了图像-文本多模态任务的任务定义及网络架构。