BERT 论文阅读

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

由 @快刀切草莓君 阅读编写。

1 引言

两种为下游任务应用预训练模型表示的现存策略

- 基于特征 e.g. ELMo：使用包括预训练表示作为额外特征的特定任务架构
- 精调 e.g. GPT Generative Pre-trained Transformer 引入最少的特定任务参数
- 这两种策略都使用了**单一方向语言模型** 限制了与训练表示的能力

benchmark：GLUE, MultiNLI, SQuAD

2 相关工作

简要回顾使用最广的**预训练泛化语言表示**（pre-training general language representations) 方式。

2.1 基于特征的无监督方式

预训练的**词嵌入**(word embedding)，使用了从左至右的语言模型目标以及从左到右找不正确的词 discriminate correct from incorrect word。这些方法泛化后，得到句子和段落的嵌入。

训练句子的表示：给下个句子打分；根据给定的句子，从左至右生成下一个句子的词。

ELMo，从双向语言模型中提取语境相关特征（contextual representation）

2.2 无监督精调方式 Fine-tuning

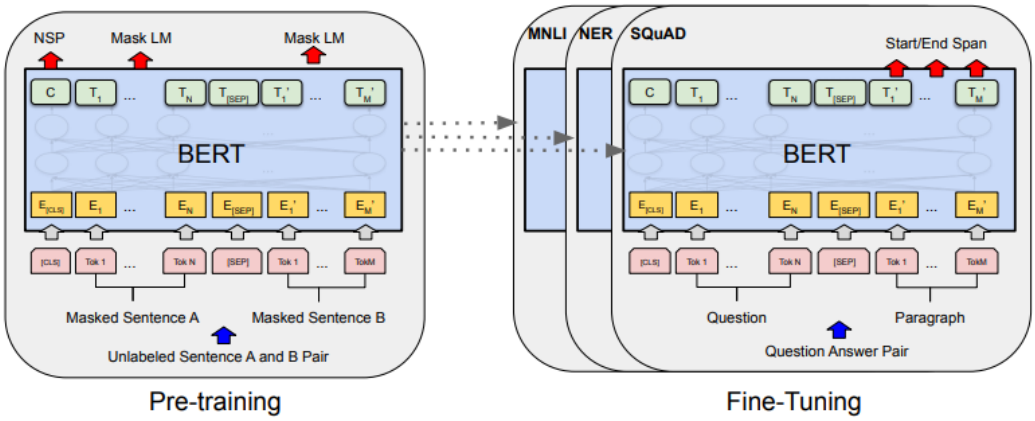
GPT，左到右的语言建模和自动编码目标。

2.3 有监督数据的迁移学习

大数据集监督学习任务：机器翻译、自然语言推理；ImageNet：迁移学习的重要性

3 BERT

两个阶段：预训练（使用无标签数据在不同的预训练任务）、精调（用与训练参数初始化后，用有标签数据精调）



[CLS]：在每个输入例子前；[SEP]：特殊分隔符

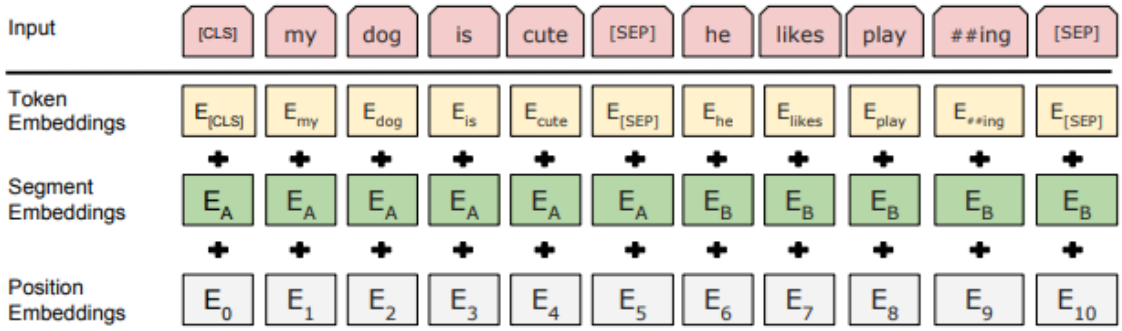
不同的下游任务都有独立的精调模型，但是这些精调模型和预训练模型差异很小。

模型架构：多层双向Trnsformer编码器

- BERT-BASE (L=12, H=768, A=12, Total Parameters=110M)； GPT的规模
- BERT-LARGE (L=24, H=1024, A=16, Total Parameters=340M).
- L：Transformer块，H：隐藏层size，A：自注意力机制头数

输入输出表示 Input/Output Representations

- 输入表示可以在一个token sentence中明确表示 句子和一些句子 ，BERT是根据输入标记来定义句子。
- WordPiece embedding，30000个token的词表。
- 每个句子的第一个token是 [CLS]，对于分类任务，最后一层和此token相关的隐藏状态可以被用于聚合句子的表示。
- 句子对 被打包成一个句子，用两种方式区分：1.[SEP]； 2.加入一个学习到的嵌入至每个token中指示其归属。
- 每个token的输入由相应的token、段落和位置编码。



3.1 预训练 BERT

使用两个无监督任务来预训练BERT：Masked LM, Next Sentence Prediction(NSP)

任务1：Masked LM

- 标准的条件语言模型只能用单向的顺序来训练，因为双向会让每个字不能正确地看自己，模型会琐碎地预测目标此在一个多层的上下文中？
- 随机应该一定比重的输入token，然后预测这些token，这个过程称MLM，完形填空；与mask token对应的最后的隐藏向量送去词表上做softmax
- 尽管可以获得双向预训练模型，但是由于[MASK]token不会在精调的时候出现，预训练和精调出现了不匹配；为了减轻这一情况，不都使用[MASK]token；80% mask, 10% 随机token；10%原先的不变的词；最后的输出还是预测原先的token

任务2：预测下一个句子 NSP

- 预训练一个二值化的预测下一个句子任务，50%是正确的；在这种任务上预训练对QA和NLI（自然语言推理）非常有帮助。
- NSP任务和表示学习(representation-learning)目标有关系，但是在早期的工作中，只有句子嵌入被传给下游任务；然而BERT将所有的参数用来给下游任务模型做初始化
- 预训练数据：BooksCorpus，英文维基百科。

3.2 精调 BERT

- 对于涉及**文本对**的应用，一般的方法是在应用双向交叉注意力机制前独立地对他们进行编码；而BERT使用**自注意力机制**将这两个阶段联合起来，因为使用自注意力机制连接文本对有效地包括了两个句子之间地双向交叉注意力。 bidirectional cross attention
- 对于每个任务，简单地加入特定的输入输出到BERT中，对所有参数进行端到端精调。
    - 在输入部分，预训练的句子AB就类似 1.转述中的句子对；2.蕴含(entailment)中的假设前提对；3.问答中的问题文章对；4.文本分类或序列标注中的退化的text-none对
    - 在输出部分，token表示被送到 token-level 的任务中去，像序列标注和问答；[CLS]表示可以**被喂给用于分类的输出层**，例如蕴含和情感分析(sentiment analysis)
- 和预训练相比，精调是非常便宜的。

4 实验

BERT在11个NLP任务上精调后的结果。

4.1 一般语言阅读理解评估 （GLUE）

GLUE：不同的自然语言理解任务的集合：MNLI、QQP、QUNLI、SST-2、CoLA ....

使用单个句子或句子对输入，使用与[CLS]对应的token作为聚合表示；仅引入了一个新的分类层。

4.2 斯坦福问题回答数据集（SQuAD） v1.1

SQuAD：100k 源于大众的问题/答案对；给定一个问题 和 wikipedia 上包含答案的一篇文章，预测文章中答案文本的标签。

使用单一的语句包来表示输入的问题和文章，问题使用A embedding，文章使用 B嵌入。

精调过程中引入了开始向量S和结束向量E，通过 Ti 和 S 的点积来计算ith词是答案范围(answer span)的开始之概率，最后对所有的词进行softmax。 Tj 和 E 计算结束。i -> j 备选范围的分数 $S \cdot Ti + E \cdot Tj$ ，`j>i`的分数最大的范围被当作输出。

训练目标：正确的开始和结束位置的log概率之和最大。

4.3 SQuAD v2.0

和1.1相比，考虑到提供的段落内不存在短的答案，使得问题更真实。

在v1.1的BERT基础上，对不含答案的问题，将start和end 定在[CLS]token上；start和end的可能空间包括了[CLS] token。预测的时候比较不含答案范围的分数与非空范围：$Snull = S \cdot C + E \cdot C$，$Sij = max(S \cdot Ti + E \cdot Tj)$，当$Sij > Snull + tao$的时候预测非空回答，tao：门槛。

4.4 常识推理的对抗性数据集 SWAG

SWAG: A Large-Scale Adversarial Dataset for Grounded Commonsense Inference；包含113k 句子对 竞争样本，用于评估常识推理。给定一个句子，任务是从四个选项中选择最可信的延续。

精调：每个输入序列 包含给定句子的连接（A）和 一个可能的延续（B）；唯一一个 task-specific 参数，它点积 [CLS] token 得到C，用于表示每个选择的分数，后接softmax。

5 Ablation Studies 消融实验

观察各个部分的影响

5.1 预训练任务的影响

展示深度双向的重要性

- No NSP: 双向，有MLM，但是没有NSP
- LTR(left to right) & No NSP：单向语言模型
- BiLSTM & No NSP

Tasks	Dev Set MNLI-m (Acc)	QNLI (Acc)	MRPC (Acc)	SST-2 (Acc)	SQuAD (F1)
$\text{BERT}_{\text{BASE}}$	84.4	88.4	86.7	92.7	88.5
No NSP	83.9	84.9	86.5	92.6	87.9
LTR & No NSP	82.1	84.3	77.5	92.1	77.8
+ BiLSTM	82.1	84.1	75.7	91.6	84.9

5.2 模型大小的影响

不同数量 layer, hidden units, attention heads；BERTbase BERTlarge

5.3 在基于特征的方法运用BERT

基于特征的方法，从预训练模型中提取合适的特征，优点:

- 不是所有任务可以轻易地用Transformer编码器结构表示，需要为特定任务准备的模型架构。
- 预先计算一次训练数据的昂贵表示，然后在该表示的基础上用更便宜的模型运行许多实验，这有很大的计算好处。

NER任务 CoNLL-2003，输入使用case-preserving WordPiece model；将其当作一个不使用CRF层的标签任务；从一个甚至更多的BERT层中提取激活值，在分类之前使用两个双向LSTM。

System	Dev F1	Test F1
ELMo (Peters et al., 2018a)	95.7	92.2
CVT (Clark et al., 2018)	-	92.6
CSE (Akbik et al., 2018)	-	**93.1**
Fine-tuning approach		
$\text{BERT}_{\text{LARGE}}$	96.6	92.8
$\text{BERT}_{\text{BASE}}$	96.4	92.4
Feature-based approach ($\text{BERT}_{\text{BASE}}$)		
Embeddings	91.0	-
Second-to-Last Hidden	95.6	-
Last Hidden	94.9	-
Weighted Sum Last Four Hidden	95.9	-
Concat Last Four Hidden	96.1	-
Weighted Sum All 12 Layers	95.5	-

6 结论

最近用语言模型迁移学习而来的经验带来的提升，已经说明了昂贵的、无监督学习的预训练是很多语言理解系统整体的一部分。这些结果使得一些资源较少的任务从中受益。

参考

Bert论文

NLP常见任务