

LeC8 Pre-trained LMs

可用于多任务学习，训练不可直接

用于NLP任务

Encoder → ELMo, BERT

Decoder → GPT

En-De → BART, T5

△ BERT

ELMo: 上下文语义分割 (上、下句LSTM)

(多生成词表)

BERT: 使用 Masked-LM → 级-上-下-3

For LM: 上级单词，通过从左到右逐字从后向前

→ ELMo不使用双向LSTM

→ 随机选择 k% 词再预测

* Encoder

Contextualized Word Embeddings (词嵌入通过语义化)

(如NN, Transformer: 都可将上下文嵌入)

改变上下文嵌入 → 改变词嵌入

机制:

抽 15% 词进行预测

其中 80% → 换为 [MASK] → 常规匹配

其中 10% → 换为任意词 → 增强, Robustness

其中 10% → 另一替换

build strong representations of non-masked words

Because: No masking during finetuning

缺点:

MLM: 抽 15%

↑ 和 PR 存在冗余 (完整上下文)

LM: 100%

优势: 抽取词表 → 更少的参数

△ ELMo

1. 一个2层LSTM, 一个2层双LSTM (二部分离)

2. 字符级CNN计算二词表示并输入

3. 残差连接

(底层输入) (顶层输出)

4. 2个LSTM共享词表 / Softmax参数

(见 LeC8 P15图)

输入词表示: (第k词) → 第k词 (应该为2)

$$ELMo_k = \sum_{j=0}^L s_j \cdot h_{k,j} \quad \rightarrow \text{第 } j \text{ 层对 } k \rightarrow \text{第 } k \text{ 词}$$

j=0: 输入表示 → 第0层 (可学) end-task mode

1. 运行准备: 由于 ELMo 通过表示 → 下游任务模型

对于 end-task model: 直接使用 ELMo (不再训练)

学习 ELMo) 或 保持相对较低的速率。→ Finetune

→ 逐层冻结 + 微调 *

大模型直接用
冻结层

结合下游任务
模型 + 特例训练

ELMo TO BERT:

LSTM 换为 Transformer

使用子词分割

Tips: 影响训练 Embeddings 指向:

① 直接施加词频惩罚来施加词频 → 词频
入世会发生成效.

② 整个 Transformer 网络 \rightarrow 随机 + 任务
双面俱进

不再直接经 LM (单向生成)

而是填完式生成

并派生指代于语言生成任务

预测值:

1. 句内掩码

2. Next Sentence Prediction

(预测句与上一句与 | 下一句)

\rightarrow (其实 Not necessary)

\rightarrow 将 [CLS] 产生的量通过 FNN

Others:

BERT \rightarrow 将 token 嵌入 其他方法 \rightarrow / 另

文本修改 / 句子对表

\rightarrow 表示 [CLS]

下游任务使用 / 微调

① 微调: 连续下游任务, 设置低学习率

热启动.

② 指示学习: 在输入 -> 指示语 (一)
(Prompting)

（减少中高频预测词率）

③ parameter-efficient fine-tuning \rightarrow 轻量级

微调: $\begin{cases} \text{模型头部} \\ \text{参数} \end{cases}$ 可学参数 \rightarrow

当输入及输出固定时模型

\rightarrow 只需调整模型中部分参数

（模型输出层. 指示参数）

Decoder PLMs

(因 Lee b: LM)

GPT: $\begin{cases} \text{单向} \\ \text{损失函数} \end{cases}$

$\begin{cases} \text{反向} \\ \text{损失函数} \end{cases}$

副作用:

① 微调

GPT: 最后一个 token 包含完整输入.

\Rightarrow 最后 token 连接下流任务微调

② 指示学习

将任务描述输入模型 \rightarrow 完成任务

Other prompting: 单试学习 (one-shot),

少试学习 (few-shot) \rightarrow 任务描述 + 指示语

插入例子

Other prompting: Chain-of-thought prompting

提供样例及求解思路

⇒ Encoder & Decoder

BART: Pre-trained by arbitrary noisy function

Learning a model to reconstruct the original text

T5: supervised & unsupervised tasks → text to

text format to train.

GLM: Decoder only, but used like end-to model.

只输入掩码，返回预测