总结

黄飞虎

2021.1.6

学习情况

使用图像分类数据集完成了图像分类

在完成作业3

论文情况

Deep contextualized word representations

深层语境词表征

本文的工作

推出了一种新的基于深度学习框架的词向量表征模型：ELMo（Embeddings from Language Models）模型，从缩写就可以看出模型本质是从语言模型而来的。

比较了ELMo与其他词向量的训练结果，突出了ELMO的优势

常见的词向量的比较

- Word2vec和glove都属于静态的词向量，无法解决一词多义的问题。而ELMo、GPT、bert词向量，它们都是基于语言模型的动态词向量。

- 从特征提取器方面说，ELMo采用LSTM，GPT和BERT用Transformer。从单、双向语言模型方面说，BERT和ELMo采用双向语言模型，GPT采用单向语言模型。但是ELMo实际上是两个单向语言模型（方向相反）的拼接，这种融合特征的能力比bert一体化融合特征方式弱。

ELMo原理

与其他广泛使用的单词嵌入不同，ELMo单词表示是整个输入句子的功能，它是在具有字符卷积的两层biLM上计算的是内部网络状态的线性函数。 此设置使它能够半监督学习，其中对biLM进行了大规模的预培训，并且可以轻松地将其纳入各种现有的神经NLP体系结构中。

双向LSTM语言模型有两个特点，一是用了多层LSTM，二是结合了forward和backward LM。elmo使用的双向lstm语言模型，由一个forward LM和一个backward LM构成。所要优化的目标：最大化对数前向和后向的似然概率。

前向语言模型：

$$p(t_1, t_2, \ldots, t_N) = \prod_{k=1}^{N} p(t_k \mid t_1, t_2, \ldots, t_{k-1})$$

后向语言模型：

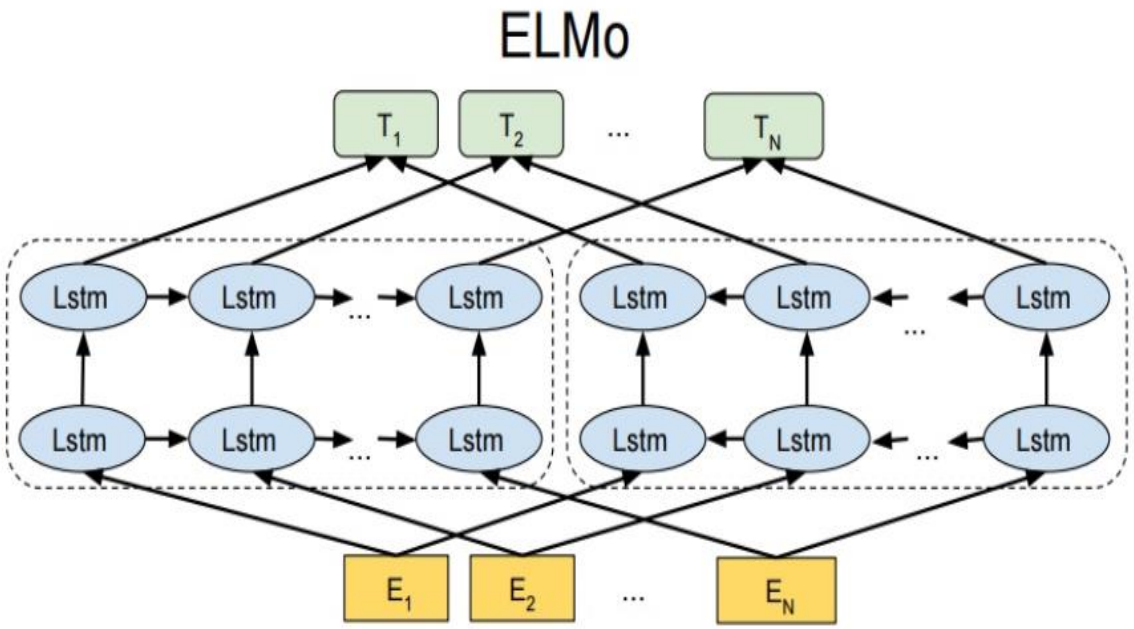
$$p(t_1, t_2, \ldots, t_N) = \prod_{k=1}^{N} p(t_k \mid t_{k+1}, t_{k+2}, \ldots, t_N)$$

合起来就是双向语言模型：

$$\sum_{k=1}^{N} \Big(\log p(t_k \mid t_1, \ldots, t_{k-1}; \Theta_x, \overrightarrow{\Theta}_{LSTM}, \Theta_s) + \log p(t_k \mid t_{k+1}, \ldots, t_N; \Theta_x, \overleftarrow{\Theta}_{LSTM}, \Theta_s) \Big)$$

Θx和Θs表示两个网络共享的参数。其中Θx表示映射层的共享，将单词映射为word embedding的共享，就是说同一个单词，映射为同一个word embedding。Θs表示上下文矩阵的参数，这个参数在前向和后向lstm中是相同的。

ELMo: Embeddings from Language Models



ELMO 模型不同于之前的其他模型只用最后一层的输出值来作为word embedding的值，而是用所有层的输出值的线性组合来表示word embedding的值。

对于每个token，一个L层的biLM要计算出 2L+12L+1 个表征：

$$
\begin{aligned}
R_k &= \{\mathbf{x}_k^{LM}, \overrightarrow{\mathbf{h}}_{k,j}^{LM}, \overleftarrow{\mathbf{h}}_{k,j}^{LM} \mid j = 1, \ldots, L\} \\
&= \{\mathbf{h}_{k,j}^{LM} \mid j = 0, \ldots, L\},
\end{aligned}
$$

评估

下表是ELMo 增强神经模型和当前最优（SOTA）单个模型基线在六个 NLP 基准任务上的测试集性能对比。

TASK	PREVIOUS SOTA		OUR BASELINE	ELMO + BASELINE	INCREASE (ABSOLUTE/ RELATIVE)
SQuAD	Liu et al. (2017)	84.4	81.1	85.8	4.7 / 24.9%
SNLI	Chen et al. (2017)	88.6	88.0	88.7 ± 0.17	0.7 / 5.8%
SRL	He et al. (2017)	81.7	81.4	84.6	3.2 / 17.2%
Coref	Lee et al. (2017)	67.2	67.2	70.4	3.2 / 9.8%
NER	Peters et al. (2017)	91.93 ± 0.19	90.15	92.22 ± 0.10	2.06 / 21%
SST-5	McCann et al. (2017)	53.7	51.4	54.7 ± 0.5	3.3 / 6.8%

回答问题
文本蕴含
语音角色标签
共指解析
命名实体提取
情绪分析

分析

	Source	Nearest Neighbors
GloVe	play	playing, game, games, played, players, plays, player, Play, football, multiplayer
biLM	Chico Ruiz made a spectacular play on Alusik 's grounder {...}	Kieffer , the only junior in the group , was commended for his ability to hit in the clutch , as well as his all-round excellent play .
	Olivia De Havilland signed to do a Broadway play for Garson {...}	{...} they were actors who had been handed fat roles in a successful play , and had talent enough to fill the roles competently , with nice understatement .

ELMo消除多义词的效果展示

Model	F_1
WordNet 1st Sense Baseline	65.9
Raganato et al. (2017a)	69.9
Iacobacci et al. (2016)	**70.1**
CoVe, First Layer	59.4
CoVe, Second Layer	*64.7*
biLM, First layer	67.4
biLM, Second layer	*69.0*

biLM表示训练的模型。第一层，第二层分别使用biLM结果显示，越高层，对语义理解越好，表示对词义消歧做的越好。这表明，越高层，越能捕获词意信息。

POS tagging（词性标注）

Model	Acc.
Collobert et al. (2011)	97.3
Ma and Hovy (2016)	97.6
Ling et al. (2015)	**97.8**
CoVe, First Layer	*93.3*
CoVe, Second Layer	92.8
biLM, First Layer	*97.3*
biLM, Second Layer	96.8

这是另一个任务的实验了，第一层效果好于第二层。表明，低层的更能学到词的句法信息和词性信息。

ELMo的优点与不足

优点：

1. ELMo能够学习到词汇用法的复杂性，比如语法、语义。
2. ELMo能够学习不同上下文情况下的词汇多义性。
3. 能轻松添加到现有模型

不足：

1. ELMo使用了 LSTM 而不是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。
2. 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。

谢谢观看