「GPT前传」Learning to Generate Reviews and Discovering Sentiment

论文地址：https://arxiv.org/abs/1704.01444

这篇文章是 OpenAI 于2017年4月发表，其中 Alec Radford 以及 Ilya Sutskever 也是之后 GPT 论文的作者之一。本篇文章给我感觉其实是 GPT 论文的一个前传，其中的一些方法以及观点都为训练 GPT 做了一个很好的铺垫。论文主要应用于情感分析任务，并具有以下的亮点：

1、不同于原来词粒度的模型编码，作者提出使用了字节「bit-level」粒度进行文本编码

2、作者提出使用无监督学习进行语言模型训练，并发现效果很好

3、作者发现调节情感单元可以生成一些不错的积极和消极的语句

首先作者介绍了当前的不同任务都是基于监督学习进行模型的训练的，但是作者认为无监督学习才是正道，因为相比于监督学习，无监督学习首先在数据收集方面不需要做很多事情其次是它不局限于某些特定领域，说白了就是训练方便至少不用为数据方面发愁。于是作者受到了《Skip-Thought Vectors》的启发，使用了无监督训练文本表征，即 GPT 中预训练方法，输入的 sentence 即为输入文本也为 label，在训练时候时候交叉熵进行 loss 计算，输入的 sentence 取 0:n-1 个字，label 则取1:n 个字，对下一个字进行生成式预训练预测，不同于 GPT 此时用到的模型还是 LSTM，因为 Transformers 还没有被发表。在进行了无监督预训练后使用少量的监督数据在任务上在进行微调，有点后面 SFT 训练的味了。

作者认为数据是一个比较重要的点，总结了之前的方法发现应该在某些针对性领域的数据上进行预训练才能拿到很好的效果，因为作者主要应用于情感分析，因此选取了 Amazon product review dataset 作为训练集，同时选用 mLSTM 结构因为相比 LSTM 训练速度会更加快，如下图所示，具体训练参数可查看原论文

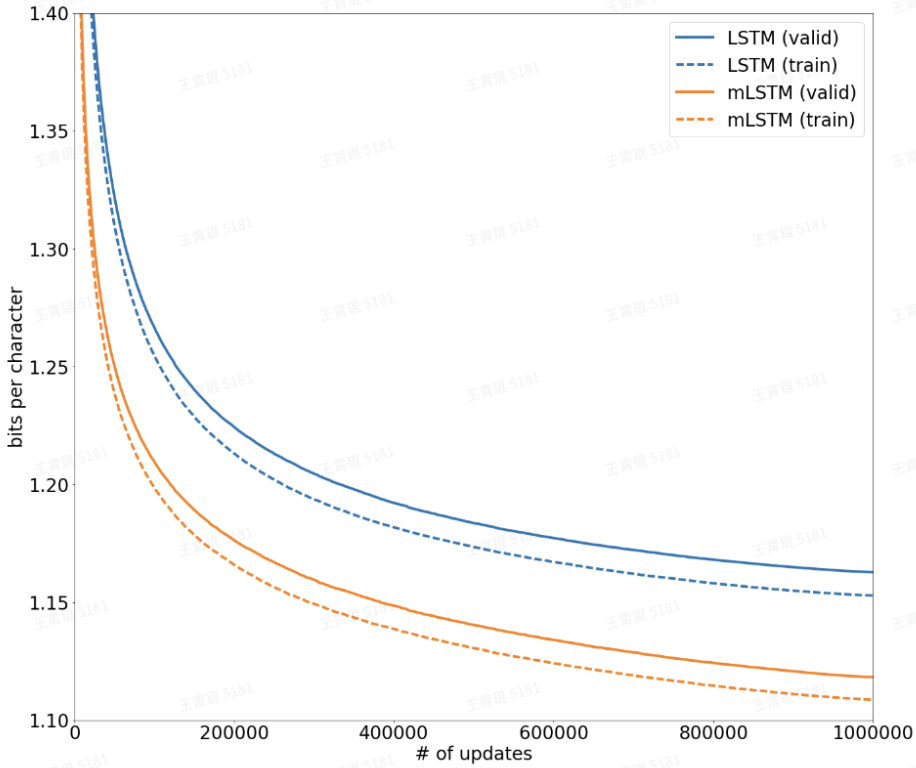

Figure 1. The mLSTM converges faster and achieves a better result within our time budget compared to a standard LSTM with the same hidden state size

接下来作者做了多个实验尝试，首先是针对情感分析任务，在四个数据集上进行了 acc 测试，如下图所示，模型「BYTE mLSTM」展示出了很不错的效果。

Table 1. Small dataset classification accuracies

METHOD	MR	CR	SUBJ	MPQA
NBSVM [49]	79.4	81.8	93.2	86.3
SKIPTHOUGHT [23]	77.3	81.8	92.6	87.9
SKIPTHOUGHT(LN)	79.5	83.1	93.7	89.3
SDAE [12]	74.6	78.0	90.8	86.9
CNN [21]	81.5	85.0	93.4	89.6
ADASENT [56]	83.1	86.3	**95.5**	**93.3**
BYTE MLSTM	**86.9**	**91.4**	94.6	88.5

同时为了验证模型的效果，作者在一个更宽泛更难的情感分析数据集 The Stanford Sentiment Treebank (SST)上与基线模型进行了对比，作者发现基于无监督预训练+有监督微调竟然打败了当前最优的模型效果。其实感觉之后的 SFT 模型训练是有一部分这里思想的启发的。

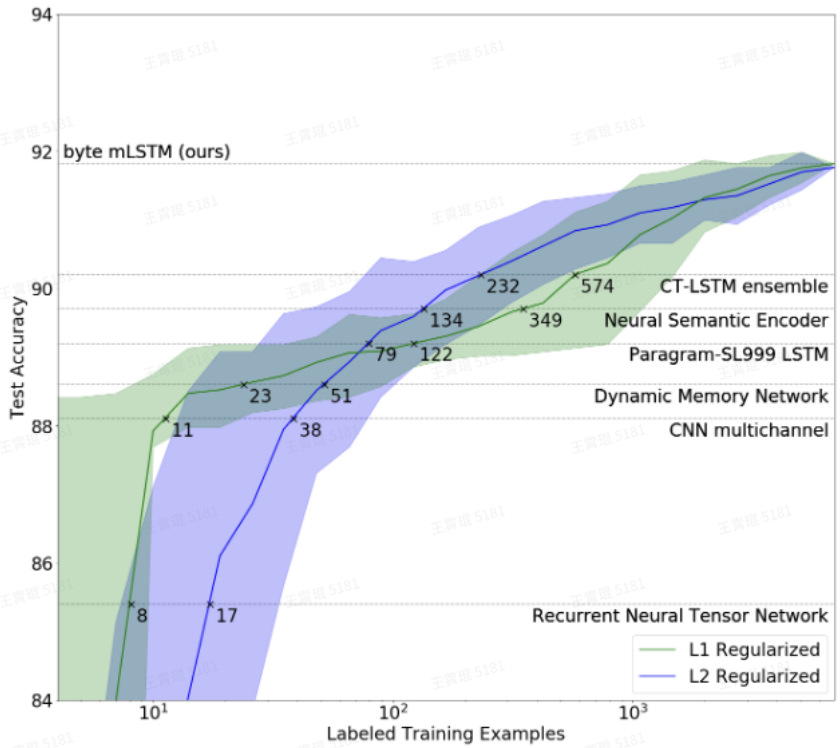

Figure 2. Performance on the binary version of SST as a function of labeled training examples. The solid lines indicate the average of 100 runs while the sharded regions indicate the 10th and 90th percentiles. Previous results on the dataset are plotted as dashed lines with the numbers indicating the amount of examples required for logistic regression on the byte mLSTM representation to match their performance. RNTN (Socher et al., 2013) CNN (Kim, 2014) DMN (Kumar et al., 2015) LSTM (Wieting et al., 2015) NSE (Munkhdalai & Yu, 2016) CT-LSTM (Looks et al., 2017)

第二个实验叫 Sentiment Unit，大概就是他们通过实验发现 mLSTM 中的隐藏层向量可以很好的表达情感，原话如下：we discovered a single unit within the mLSTM that directly corresponds to sentiment. 这里的实验我没有看的太懂，就先跳过了。

第三个实验叫做 Capacity Ceiling，作者认为既然通过这种方法这个模型的表现能够这么好，他想看看在更大更多领域的数据集上效果怎么样，于是用 Yelp Dataset Challenge 数据集，进行了不同量级的监督训练，然而结果不尽如人意，并没有跑赢 state-of-art。他分析原因可能有两点：一个是无监督训练只是在情感分析领域的数据集上进行，并没有在其他领域进行训练，可能模型的表示泛化性有一定的缺乏；另一个就是认为模型直接从句子级训练跳跃到文档级训练有点吃不消，说白了就是对长序列的注意力不够，找不到关键信息。我认为是当时 LSTM 模型本身的原因，长序列的记忆缺失通病。作者的原因分析的非常的准确，因此在后面才能用更广泛的数据集以及更好的 transformers 结构有了出名的 GPT 模型。

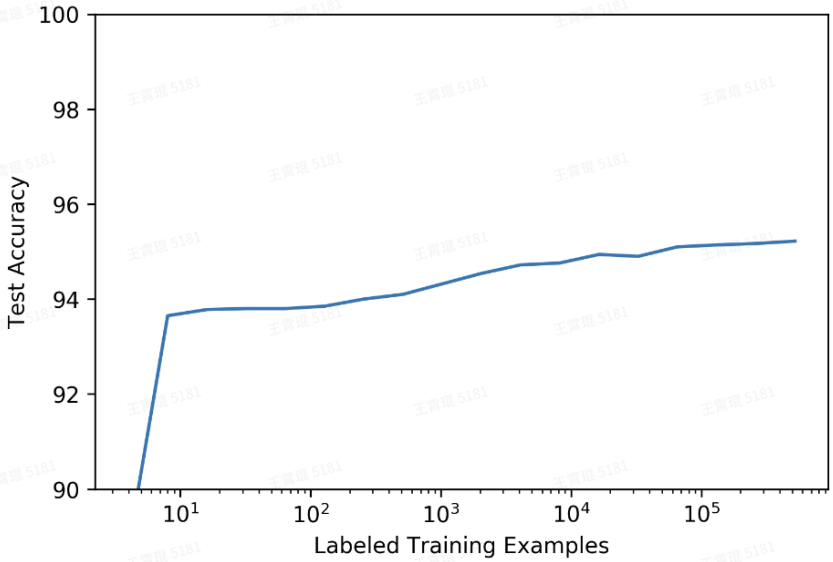

Figure 5. Performance on the binary version of the Yelp reviews dataset as a function of labeled training examples. The model's performance plateaus after about ten labeled examples and only slow improves with additional data.

最后总结一下作者的生成实验，作者虽然是应用于情感分析任务的，但是他在做预训练的时候是采用的 GPT 那套生成式训练，作者这里同样感兴趣它的生成效果怎么样，结果发现还真不错，主要是将情绪单元简单的设置1和-1来产出积极和消极的生成语句，原话如下：In Table 5 we show that by simply setting the sentiment unit to be positive or negative, the model generates corresponding positive or negative reviews. 说实话我这里仍然没弄懂情绪单元是怎么进行设置的，所以就不多解释了，最后作者挑选了一些不错的生成 case 供我们欣赏。

Sentiment fixed to positive	**Sentiment fixed to negative**
Just what I was looking for. Nice fitted pants, exactly matched seam to color contrast with other pants I own. Highly recommended and also very happy!	The package received was blank and has no barcode. A waste of time and money.
This product does what it is supposed to. I always keep three of these in my kitchen just in case ever I need a replacement cord.	Great little item. Hard to put on the crib without some kind of embellishment. My guess is just like the screw kind of attachment I had.
Best hammock ever! Stays in place and holds it's shape. Comfy (I love the deep neon pictures on it), and looks so cute.	They didn't fit either. Straight high sticks at the end. On par with other buds I have. Lesson learned to avoid.
Dixie is getting her Doolittle newsletter we'll see another new one coming out next year. Great stuff. And, here's the contents - information that we hardly know about or forget.	great product but no seller. couldn't ascertain a cause. Broken product. I am a prolific consumer of this company all the time.
I love this weapons look . Like I said beautiful !!! I recommend it to all. Would suggest this to many roleplayers , And I stronge to get them for every one I know. A must watch for any man who love Chess!	Like the cover, Fits good. . However, an annoying rear piece like garbage should be out of this one. I bought this hoping it would help with a huge pull down my back & the black just doesn't stay. Scrap off everytime I use it.... Very disappointed.

Table 5. Random samples from the model generated when the value of sentiment hidden state is fixed to either -1 or 1 for all steps. The sentiment unit has a strong influence on the model's generative process.

最后，作者总结通过分层等方法可以提升长文本的表示，以及使用更多更广泛的数据集去训练会得到更好的效果，同时作者认为当前这一套方法可以针对语言模型继续研究下去，不需要在做任何特

殊处理下就可以拿到高质量的文本表示。可以说之后 Google 发表的 Transformer 成就了之后的 GPT。

1、Since our model operates at the byte-level, hierarchical/multi-timescale extensions could improve the quality of representations for longer documents.

2、The sensitivity of learned representations to their training domain could be addressed by training on a wider mix of datasets with better coverage of target tasks

3、Finally, our work encourages further research into language modelling as it demonstrates that the standard language modelling objective with no modifications is sufficient to learn high-quality representations.