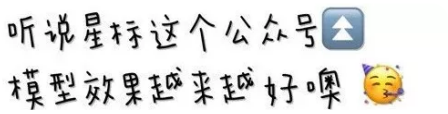
ACL2020 | 什么时候值得用BERT上下文嵌入

Viktor Karlsson NewBeeNLP 2020-08-07



作者 | Victor Karlsson
原文 | 见页面左下角『阅读原文』
编译 | NewBeeNLP

不知道大家在平时使用时有没有发现，BERT的上下文嵌入非常『昂贵』，并且可能无法在所有情况下带来价值。分享一篇ACL2020的论文，介绍了一些思路。

- 论文：Contextual Embeddings: When Are They Worth It?
- 代码：https://github.com/HazyResearch/random_embedding

写在前面

诸如BERT或其改进后代之类的SOTA模型，使用起来十分"昂贵"。仅仅是预训练的『BERT-base』模型（用今天的标准几乎可以认为是很小的模型），也需要在16个TPU芯片上花费了超过4天的时间，而这需要花费数千美元。这甚至都没有考虑对模型进行进一步的微调或最终使用，这两者都只会增加最终的总成本。

与其尝试找出创建更小的Transformer模型的方法（如何修剪BERT达到加速目的？理论与实现），不如退后一步去问：**「基于Transformer模型的上下文嵌入何时真正值得使用？」** 在什么情况下，使用GloVe或甚至是随机嵌入等计算成本较低的非上下文嵌入（non-contextual embeddings），可能达到类似的性能？

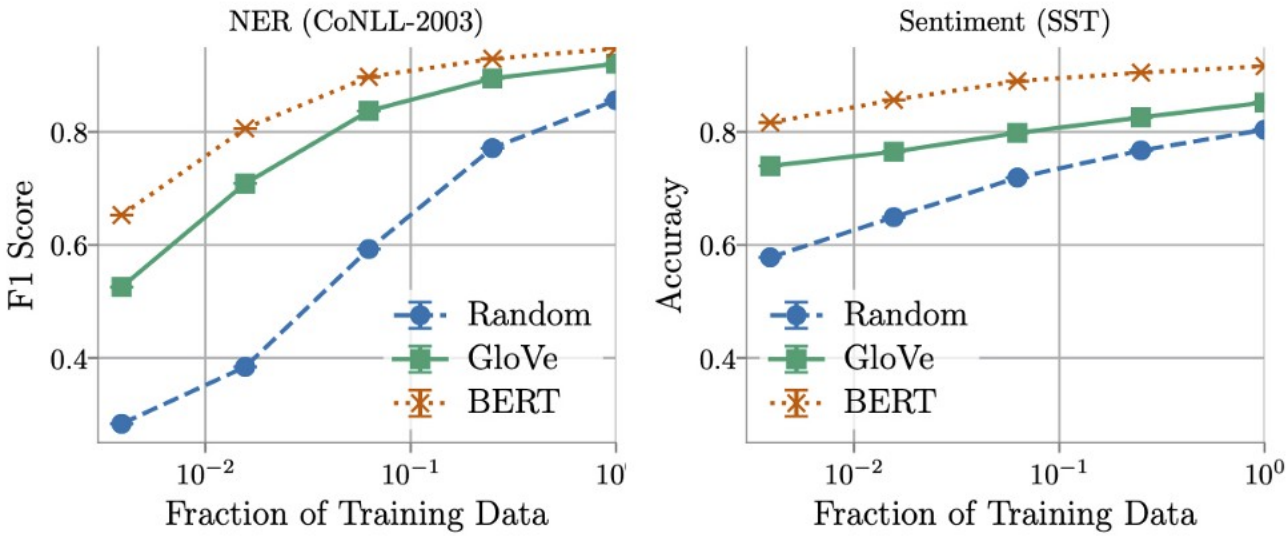
这些是Arora等人提出的一些问题，而答案就在他们的研究中：Contextual Embeddings: When Are They Worth It[1]?本文将概述他们的研究并重点介绍他们的主要发现。

研究内容

该研究分为两个部分，首先检查训练数据量的影响，然后检查这些数据集的语言特性。

训练数据大小

作者发现，在决定BERT-embedding和Glove-embedding的效果性能方面，训练数据量起着关键作用。通过使用更多的训练数据，非上下文嵌入很快得到了改善，并且在使用所有可用数据时，通常能够在BERT模型用时的5-10%之内完成。



另一方面，作者发现在某些情况下，可以用少于16倍的数据来训练上下文化嵌入，同时仍然与非上下文化嵌入所获得的最佳性能相当。这就需要在推理（计算和内存）和标记数据的成本之间进行了权衡，或者如Arora等人所说：

> ML practitioners may find that for certain real-world tasks the large gains in efficiency [when using non-contextual embeddings] are well worth the cost of labelling more data. ——— Arora et al

数据集的语言特性

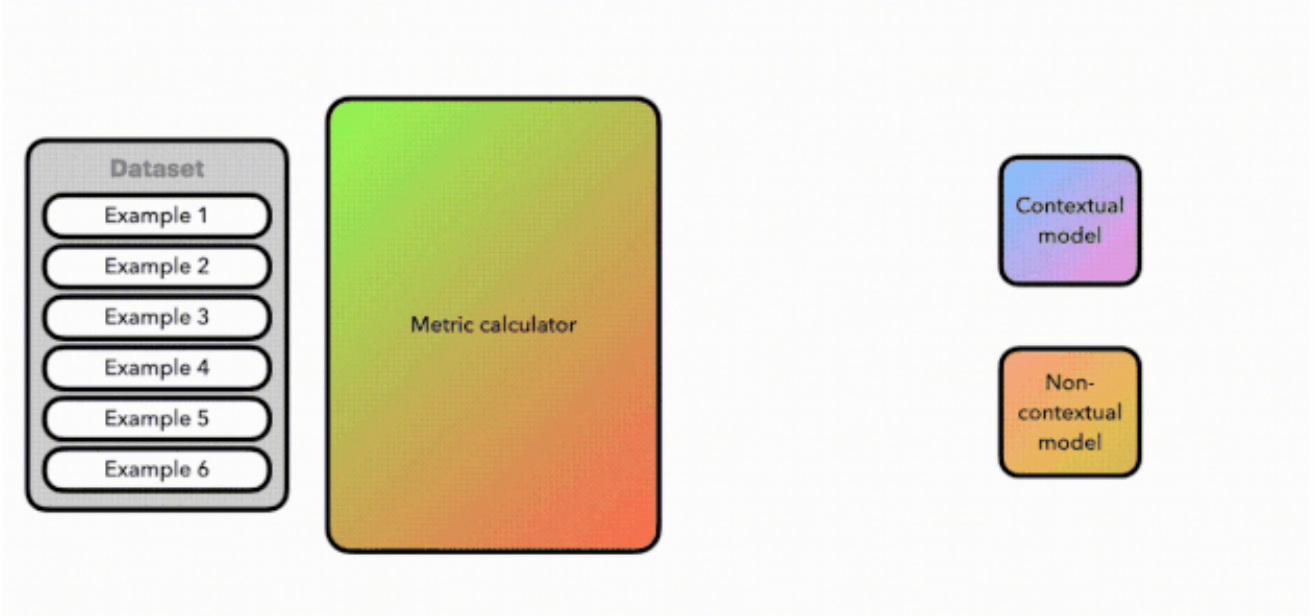
对训练数据量的研究表明，在某些任务中，上下文嵌入比非上下文嵌入的表现要好得多，而在其他情况下，这些差异要小得多。这些结果激发了作者们的思考，是否有可能找到并量化语言特性，以表明这种情况何时发生。

为此，他们定义了三个度量标准，用于量化每个数据集的特征。根据设计，这些度量没有给出一个单一的定义，而是用来编码哪些特征影响模型性能的直觉。这使得我们可以对它们进行解释，然后对它们进行严格的定义，以用于我们研究的任务。因此，下面以命名实体识别数据集举例作者提出的指标：

- **文本结构的复杂性**：表示一个句子中词与词之间的依赖性。在NER中表现为每个实体跨越的token数量，如“George Washington”横跨两个token。

- **词义模糊**：每个token在训练数据集中分配的不同标签的数量，如“Washington”可以指定人员、地点和组织，这需要考虑到它的背景。
- **未出现词的流行度**：表示在训练过程出现从未见过词的概率。在NER中定义为token出现次数的倒数。

这些指标被用来给数据集中的每一项打分，以便我们将它们分成“困难”和“容易”。这使得我们能够比较来自同一数据集的这两个分区的嵌入性能。



如果这些指标是非信息性的，那么这两个分区的性能差异将是相等的。幸运的是，作者们发现并非如此。相反，他们观察到，在42个案例中，有30个案例，上下文嵌入和非上下文嵌入之间的差异在困难分区上高于简单分区。

这意味着，这些指标可以作为一个代理，来自BERT之类模型的上下文嵌入将优于非上下文嵌入！然而，从另一个角度来看，它可能更有用——用于指示来自glove的非上下文嵌入何时足以达到最先进的性能。

结论

在研究Contextual Embeddings: When Are They Worth It? 中，Arora等人强调了数据集的关键特征，这些特征指示上下文嵌入何时值得使用。首先，训练数据集大小决定了非上下文化嵌入的潜在有用性，即越多越好。其次，数据集的特征也起着重要作用。作者定义了三个指标，即文本结构的复杂性，词汇使用的模糊性，以及未出现词的流行度，这有助于我们理解使用上下文嵌入可能带来的潜在好处。

本文参考资料