Global Vectors

YWJ

2017年7月17日

1 Introduction

步Skip-Gram的后尘，这篇文章 [1]设计了一种结构上大同小异而单纯扩大统计范围的词向量模型。它不直白地使用NCE采样而直接对整个文本做统计，得到关联词的准确频率作为优化目标。这么做的正面意义，以及作者声称Glov相对Word2vec的一系列优势在我看来是存疑的，在最后一节会详细讨论这些。另外值得一提的是这篇文章对词向量模型启发式建模过程的描述是有概括意义而值得参考的。

2 目标函数的生成过程

2.1 考虑在全文范围内做统计

Global Vector与Skip-Gram的主要区别是Glov使用整个文本的统计信息做为训练数据，而SG只以一个小窗口在文本上滑动(类似阅读的过程)而不全局的考虑整个文本。

Glov受到的启发来自古老的HAL（Hyperspace Analogue to Language）模型，HAL使用一个词到词的二维矩阵X来描述整个文本的信息，矩阵元素X_{ij}的意义是词j出现在词i上下文的频数，确定了训练文本和上下文窗口后，矩阵X就被确定了。

Glov希望由词i,j对应的向量通过一些对称的运算得到X_{ij}，以此作为词向量的优化目标。

2.2 考虑使用概率的比值而不直接使用概率（这其实就是NCE）

如果对矩阵X的每一行做归一化，可以得到词$i$fi$j$互为上下文的统计概率，显然统计文本足够多时矩阵元素趋于真实分布。记归一化后的矩阵为P，元素p_{ij}表示词$i$fi$j$互为上下文的概率，直观的想法是优化下面的函数：

$$F(w_i, \widetilde{w_j}) = p_{ij}$$

w_i是词i作为中心词的词向量，$\widetilde{w_j}$是j作为上下文词的词向量，由于上下文是对称的，在理想的学习器中，一个词作为中心词或上下文词的向量表示是相等的，但由于实际中学习器对两者的更新频率以及顺序不同，这两者是有区别的(在section 6的推导以及最后的更新算法中能很直观的看到这一点)，但这并不妨碍在完成训练后将中心词或上下文的向量表示混用或单独取某一方作为最终的词向量。

但作者的想法是简单的使用单个关联词对的概率不如比较两个关联词对的概率靠谱，或者说在确定了中心词k的前提下，用词i和j出现在k上下文的概率之比作为优化函数的右端。目标函数变成了：

$$F(w_i, w_j, \widetilde{w_k}) = \frac{p_{ki}}{p_{kj}}$$

如果熟悉NCE损失的形式就不难发现，把上面的式子右边取自然对数就成了在单个样本在单个负采样下的NCE损失函数（作者后面确实就这么干了）。但作者似乎在刻意回避或根本没有意识到这一点，只是经验主义地说明了这样做的优势。

2.3 考虑词间的差异而忽略共性

作者的想法是：如果词i与k相近而词j与k不怎么相关，比率(或者叫NCE似然)$r = \frac{p_{ki}}{p_{kj}}$应该是个远大于1的数，反之$r$是个小于1而接近0的数，如果两个词都不与$k$相关或者都与$k$相同程度地相关，$r$应该接近1。而两个词如果本身互相靠近，它们到$k$的距离应该是相似的，而两个词互相远离，那么它们到$k$的距离可能差异较大，于是目标函数的左部关于$i, j$的向量形式应该取它们的差值：

$$F(w_i - w_j, \widetilde{w_k}) = \frac{p_{ki}}{p_{kj}}$$

但这样做显然会把一些有三角关系的词判为近义词。比如‘ice’, ‘cream’, ‘icecream’,显然‘ice’和‘cream’的语义是隔得很远的，但由于有‘icecream’这样一个兼具两者性质的词存在，三个词仿佛组成了以‘icecream’为顶点的等腰三角形，而上面的模型显然会因为两个脚上的词到顶点距离相似而认为它们是近义词。增加负采样的数量从而破坏这种对称的三角结构应该能解决这个问题。当然如果那样做了，这个模型就更接近Word2vec了。

注意到式子左边是两个同型向量经由F运算，而右边是个标量，于是左部只能表示为这种形式：

$$F((w_i - w_j)^T \widetilde{w_k}) = \frac{p_{ik}}{p_{jk}}$$

2.4 考虑保持上下文的对称性

以词A为中心考察词B做上下文的结果应该与以词B做中心考察词A做上下文的结果相同，即AB对称地互为上下文。

$$F((w_i - w_j)^T \widetilde{w_k}) = \frac{p_{ik}}{p_{jk}}$$
$$F((\widetilde{w_i} - \widetilde{w_j})^T w_k) = \frac{p_{ik}}{p_{jk}}$$

或者：

$$\widetilde{w_i}^T w_k = w_i^T \widetilde{w_k}$$
$$\widetilde{w_j}^T w_k = w_j^T \widetilde{w_k}$$

式子左部时差值，而右部是商的形式，于是自然想到运算F是对右部取对数：

$$F(x) = e^x$$

$$F(w_i^T \widetilde{w_k}) = p_{ik}$$

实际上这样的解是不对称的：

$$w_i^T \widetilde{w_k} = \ln p_{ik} = \ln X_{ik} - \ln X_i$$

$$\widetilde{w_i}^T w_k = \ln p_{ki} = \ln X_{ki} - \ln X_k$$

由于$X_{ik} = X_{ki}$,不对称的根源在于$X_i \neq X_k$,由于X_i与词k无关，考虑将X_k吸收进w_i的偏置b_i,再考虑对称性，也给$\widetilde{w_k}$添加偏置项$\widetilde{b_k}$,此时有：

$$w_i^T \widetilde{w_k} + b_i + \widetilde{b_k} = \ln X_{ik}$$

$$\widetilde{w_i}^T w_k + \widetilde{b_i} + b_k = \ln X_{ki}$$

这两个式子是互相对称的，任意一个都可以作为Glov的优化目标。

2.5 调整关联词对的权重

通过引入偏置得到关于上下文对称的F函数后，仍然存在的问题是所有被当作互为上下文的词对对目标函数的影响是等重的。但从经验上来说，样本的数量越少，样本提供的统计结果越不可靠。或者说，之所以能根据统计结果估计样本背后的概率分布是因为有强弱大数定理做支撑，但样本很少的时候大数定理的结论很松，因此少量样本的统计信息价值很低。援引吴军的《数学之美》中的例子说明这种情况：

如果你在一天的某时刻观察到窗外有200个男生和800个女生，你很有理由相信明天这个时候窗外的男女比例接近1：4，但如果你只观察到了1个男生和4和女生，上面这个信念就显得不可靠了。

同样，在文本统计中，对出现频率很低的中心词直接用2.2中的方法估计上下文概率是不可靠的。直接的想法是根据样本的基数给予不同的权重。作者给出了这样一个损失和函数：

$$J = \sum_i \sum_k f(X_{ik})(\widetilde{w_i}^T w_k + \widetilde{b_i} + b_k - \ln X_{ki})$$

$$f(x) = \begin{cases} (\frac{x}{x_{\max}})^\alpha & x < x_{\max} \\ 1 & x \geq x_{\max} \end{cases}$$

至于x_{max}和α两个超参，作者给出的建议值是100和0.75。

3 一些看法

下面逐条分析作者声称的Glov对Word2vec的优势：

1. Glov使用整个文本的统计信息而Word2vec每次只对局部文本做优化而不考虑整体文本。

Glov所谓的利用全局的统计信息指的是：直接统计词频作为上下文词的概率，并将之作为两向量积的优化目标，也就是说所有的关联词对不管在文本中出现的频率高低参与优化的次数是相等的（随机梯度下降在矩阵中随机地取元素，每个词被取到的概率是相等的），但不同的是出现次数更多的关联词对优化速度更快，或者在学习率一定的情况下，每次优化的变化值更大。

Word2vec对特定关联词对的优化速度则是固定的，由于负采样的存在这个速率是随机的而且是一个与关联词对在全文出现频次无关的值，但Word2vec中每个关联词对的优化次数是与其出现的频次有关的，

Word2vec在窗口遍历文本的过程中取样，那么出现频率高的关联词对被取到的概率越大，被优化的次数越多。

可以看见两者在对文本全局统计信息的利用上是对称的，一个固定优化次数而根据全局词频决定优化速率，一个固定优化速率根据全局词频决定优化次数。说Word2vec只在局部取样优化而不考虑全局统计信息显然是想当然的说法。

反而是Glov这种先统计再训练的办法导致了这个模型无法做增量训练。Word2vec可以直接在原来的训练文本上拼接新增的训练内容，而不影响正在进行的训练过程，因为Word2vec在训练前面的文本时不会预设后面还有没有或者有什么样的文本。Glov则不得不对增加内容后的训练集从新做统计，得到一个与之前完全不同的训练矩阵再从头开始整个训练过程。

1. Glov使用权重削弱低频统计信息。

实际上我认为这是Glov对Word2vec的主要优势所在，Word2vec由于没有预先统计全文，即使窗口采样到不可信的低频词，也与高频词一视同仁地做优化，而Glov使用了经验函数来削弱低频词带来的噪声，这可能是Glov在一些测试上跑分领先Word2vec的主要原因。

3. Glov使用比率而不直接使用单个关联词对的概率。

前文已经提过，所谓比率就是NCE似然,这样看来Word2vec不仅也使用了比率，而且由于负采样数量不只一个，正负样本数量不均衡，不会出现上面提到过的等腰三角误判的问题。

4 一些后话

词向量模型的不足根源在于一个这些模型所遵循的一个共同的基础假设：有相似上下文的词有相似的语义，它们在向量空间中的表示应当相似。

这个假设造成的后果是一些语气，时态不同的词可以用在完全相同的上下文中，但它们表达的语义可能是截然不同的。比如所有的反义词对，你会发现把句子中的某个词替换为它的反义词仍然是通顺的，这说明这些反义词的上下文几乎是一样的，那么它们在向量空间中的地位是相似的，但包含它们的句子可能表达了完全相反的意思。如果仅仅使用很小的窗口，学习器不论统计多少文本也不会分清‘cold’和‘warm’的区别，增大上下文窗口可以解决这个问题，或者说这类词在当前句子中可以任意替换而不显得违和，只有联系更远的上下文才能将之区分开，但增大上下文窗口的代价是显而易见的，Skip-Gram和Glov都没能解决这个问题。

参考文献

[1] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543, 2014.