深度语义模型以及在淘宝搜索中的应用

阿里云云...
已认证的官方帐号

摘要： 传统的搜索文本相关性模型，如BM25通常计算Query与Doc文本term匹配程度。由于Query与Doc之间的语义gap, 可能存在很多语义相关，但文本并不匹配的情况。为了解决语义匹配问题，出现很多LSA，LDA等语义模型。

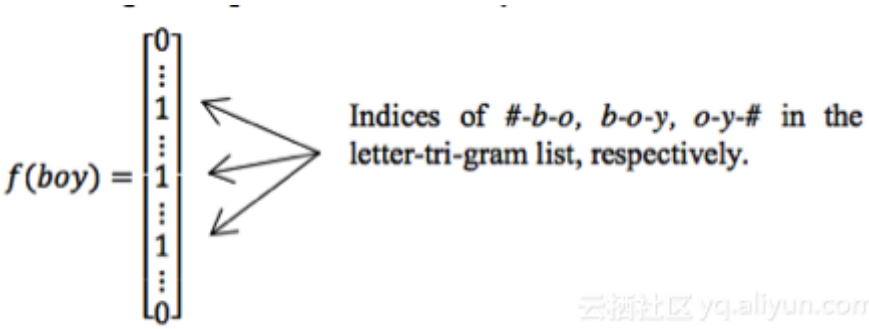
原文：click.aliyun.com/m/4165...

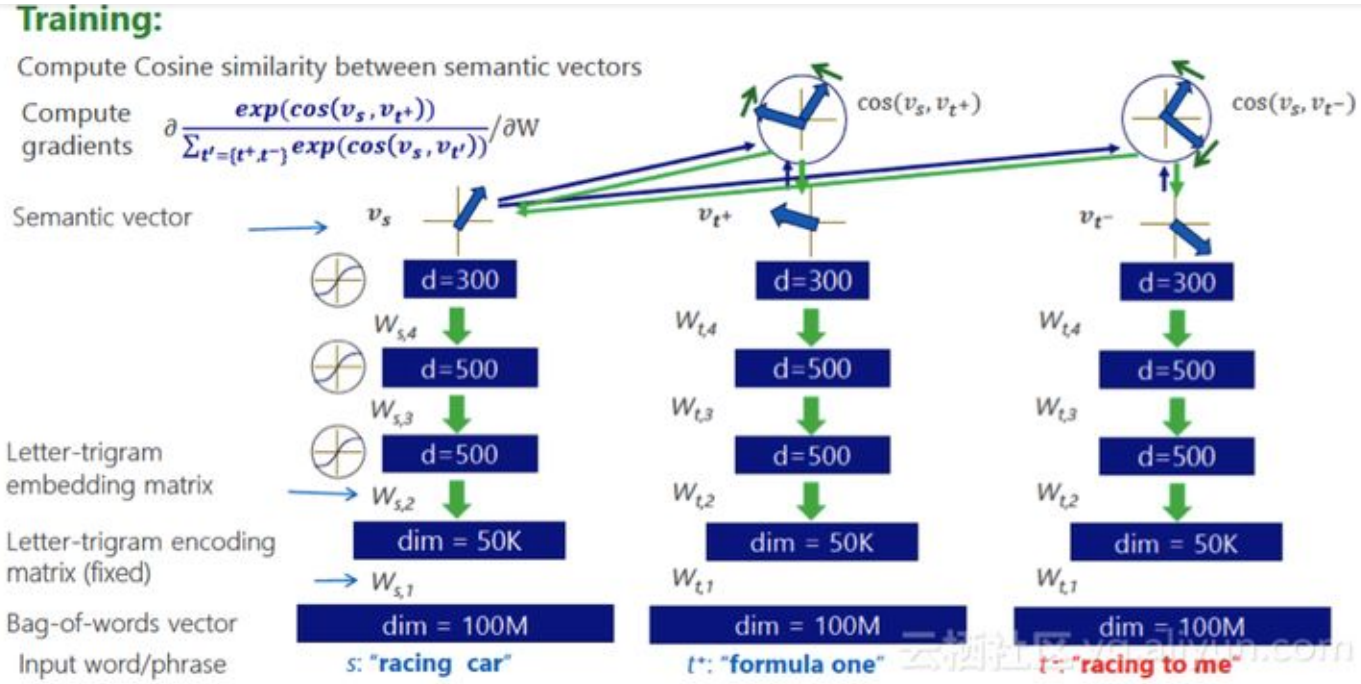
传统的搜索文本相关性模型，如BM25通常计算Query与Doc文本term匹配程度。由于Query与Doc之间的语义gap, 可能存在很多语义相关，但文本并不匹配的情况。为了解决语义匹配问题，出现很多LSA，LDA等语义模型。随着深度学习在NLP的应用，在IR和QA(问答系统)中出现了很多深度模型将query和doc通过神经网络embedding，映射到一个稠密空间的向量表示，然后再计算其是否相关，并取得很好的效果。本文调研了微软，IBM Waston实验室、Google等在这方面的一些工作，并介绍我们在淘宝搜索上做的些工作。

1.DSSM、CDSSM，LSTM-DSSM及相关系列工作

微软的DSSM及相关系列模型是深度语义模型中比较有影响力的。集团内PS上有DSSM分布式实现，而且也有多业务应用.

DSSM首先将query和doc表示成一个高维且稀疏的BOW向量，向量的维度即词典的大小，每一维表示该term在query或doc中出现的频次；如果向量每一位直接用单词，会出现维度非常高，而且对一些未登录词无法处理。作者做了一个非常有用的trick word-hash: 将每个单词表示成一个letter-tri-gram的序列， 例如：boy切分成#-b-o, b-o-y, o-y-#， 然后再表示成letter-tri-gram向量。把每个单词向量累加起来即表示整段文本的向量。





由于DSSM对文本embedding时没有考虑term的顺序信息，又陆续提出了采用Convolution和LSTM对文本embedding，可以保留词序信息。其中，Convolution是实现方式通过对query或doc用固定大小滑动窗口取片段，对每个片段内文本用word-hash+dnn压缩， 然后取max-pooling表示整个query或doc向量。

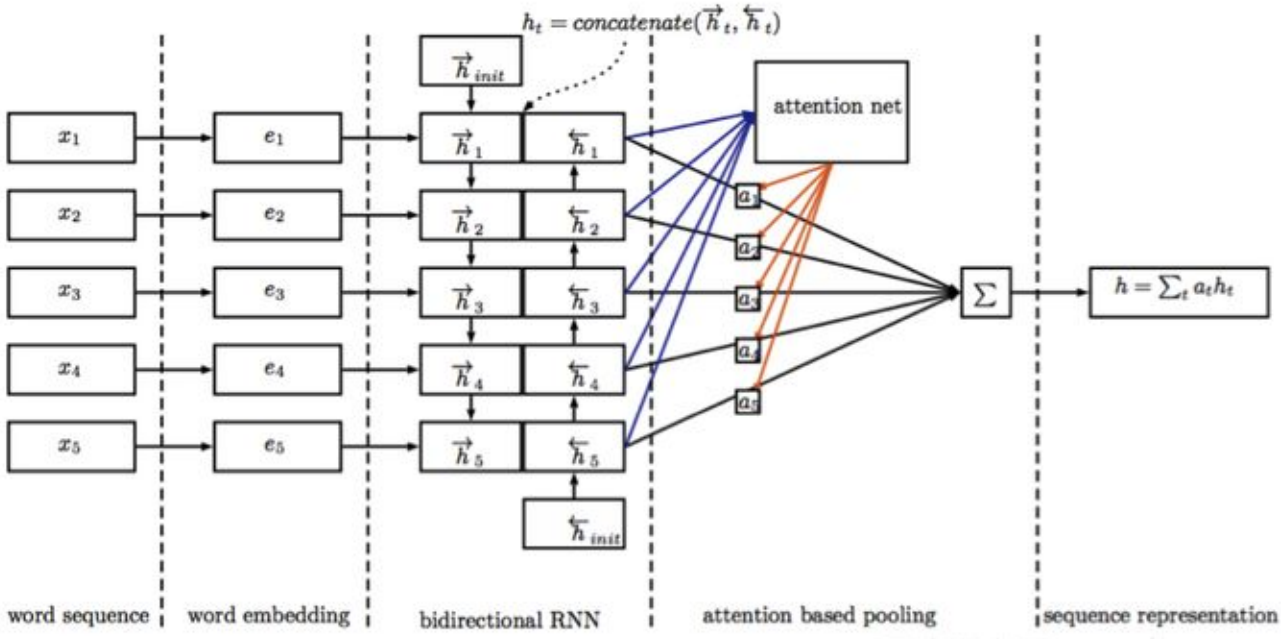
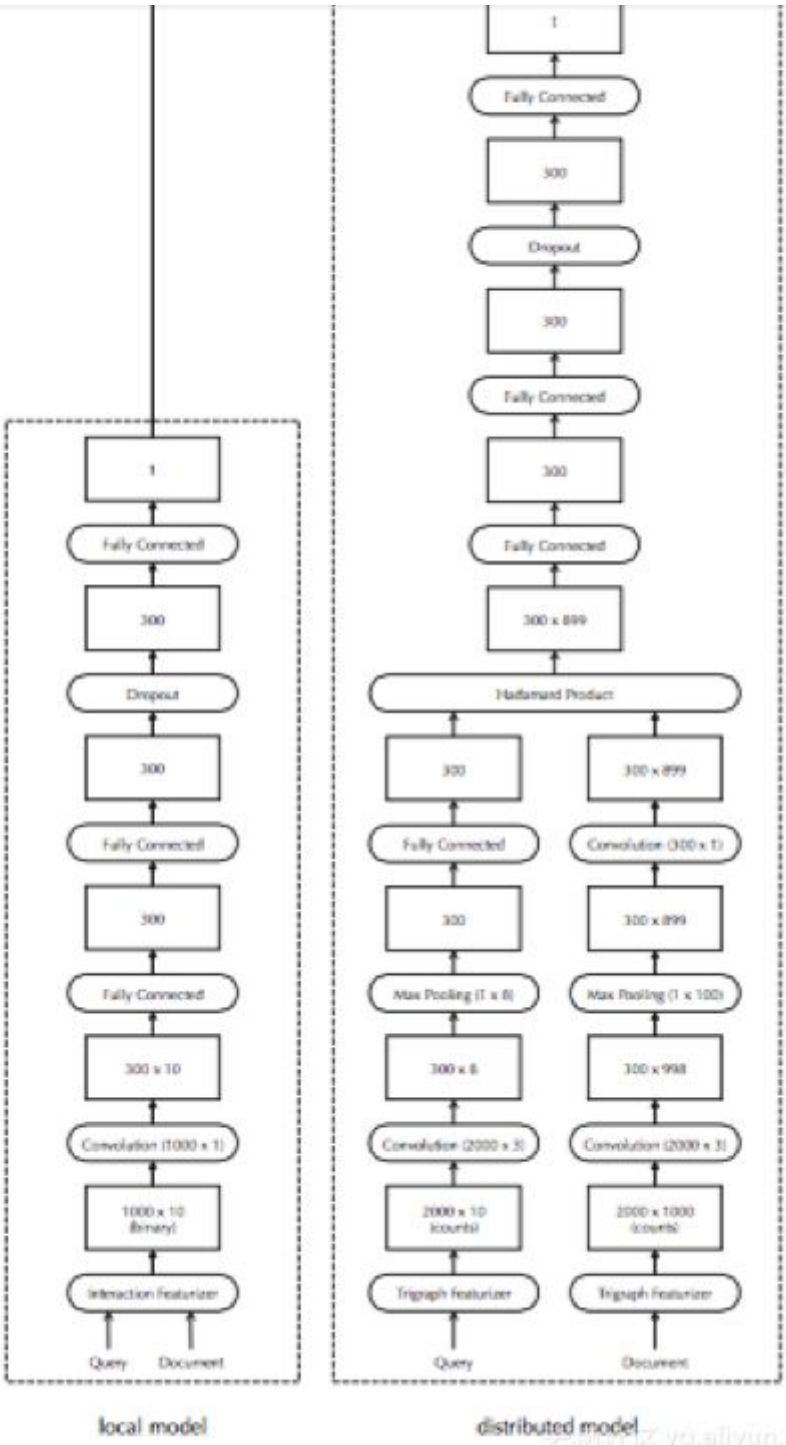


Figure 2: The encoding architecture of attention based pooling, demonstrated with a bidirectional RNN.

此外， 无论是Convolution还是LSTM对文本embedding, 都涉及到通过词或局部片段的向量生成整个句子的向量，比较简单粗暴的方法是直接取sum、avg或者max等。微软的学者们进一步做了改进，提出利用Attention机制来学习各个词组合成句子向量的权重。以LSTM-DSSM为例，

知乎 首发于 我是程序员

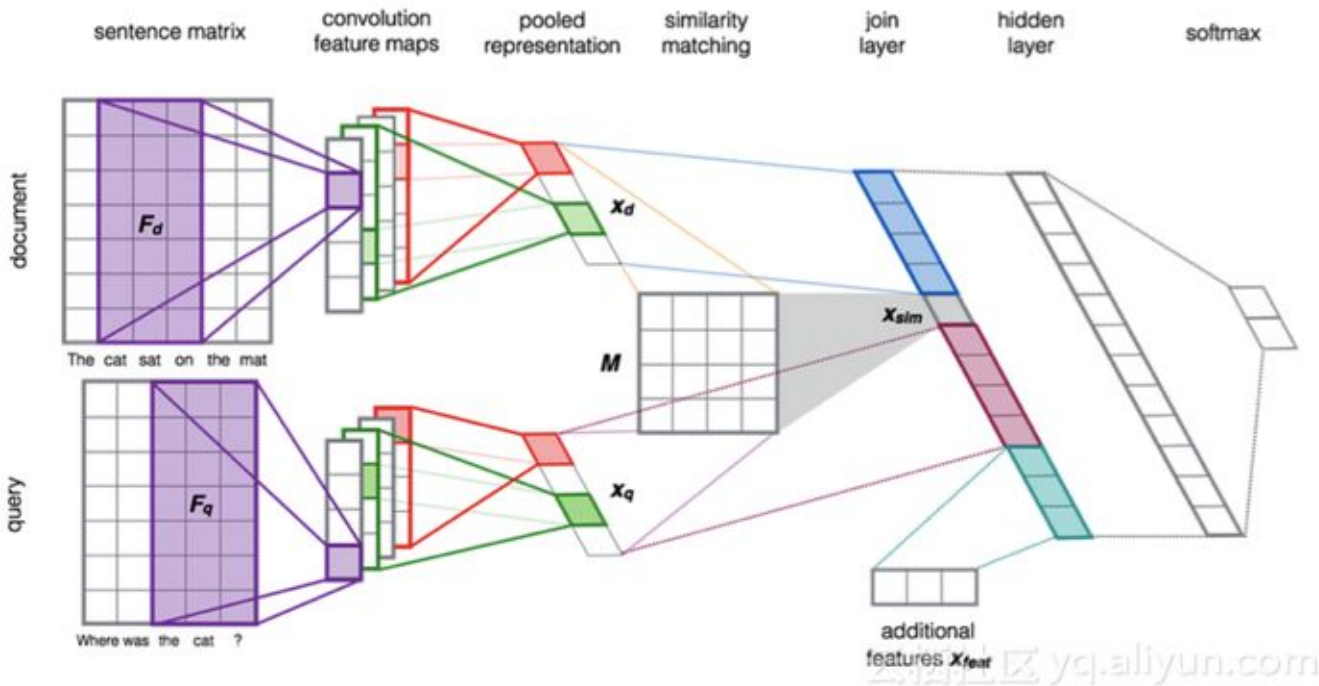
最近，微软的学者们又提出了一个观点：query与doc的相关程度是由query里的term与doc文本精准的匹配，以及query语义与doc语义匹配程度共同决定。而且，term匹配与term在doc中的位置和紧密度有较大关系。因此，他们用一个local model来表达term匹配程度，distribute model表达语义匹配程度，把这两个子模型放在同一个模型来训练。distribute model类似与DSSM来学习语义匹配关系。Local model的输入是一个nq∗ndnq∗nd的矩阵mm，nqnq是query中term个数，ndnd是doc中term个数，位置m(i,j)=0or1m(i,j)=0or1表示query里的第i个词是否与doc里的第j个词匹配，对这个输入矩阵通过convolution抽取特征并向量化。据其实验结果，这种结合term匹配信息的模型效果要优于DSSM等语义模型。



2. Google相关工作

Google的学者在用convolution对文本向量化是相比CDSSM做了些改进。Convolution的方法参考了Nal Kalchbrenner等对文本用卷积来做分类的方法。

首先，对句子中的每个词做embedding, 然后将词的embedding concat起来组合成一个矩阵，有点类似图像的表达。然后，在这个矩阵上通过不同feature map抽取特征，然后pooling生成一个维度的向量来表达句子。对Query和Doc的语义向量，再通过一个bilinear的模型计算其语义相似

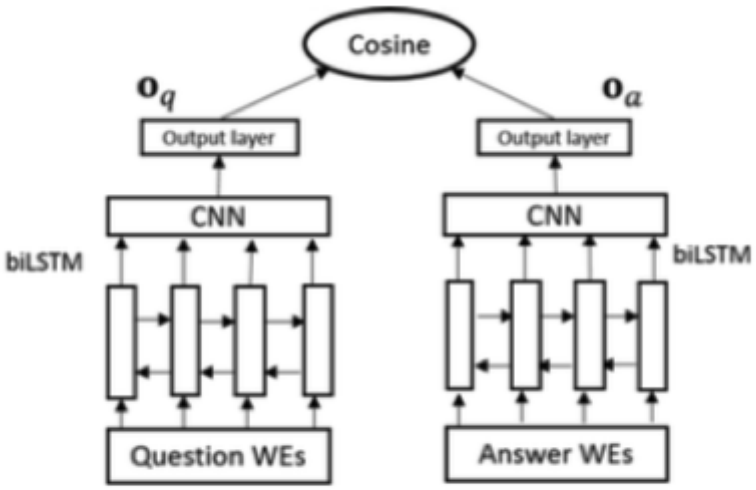


3. IBM Waston实验室相关工作

问答系统有很多种类型，其中给定一个Question和候选Answer，从候选Answer中挑选最合适的答案，这个过程与信息检索中的相关性模型非常相似。Waston实验室在InsuranceQA数据集实验了上述类似的模型，并综合CNN和LSTM的优势，提出了几种有意思的混合模型：

(1) Convolutional-pooling LSTM

用一个Bi-LSTM作为word embedding的方法，然后word embedding concat成矩阵表达句子，用卷积来抽取组合特征作为question和anwser的向量表达，再计算cosin loss.



先对原始文本用卷积捕捉局部的N-gram信息， 然后在这个基础上用Bi-LSTM来学习更大范围的上下文依赖关系。

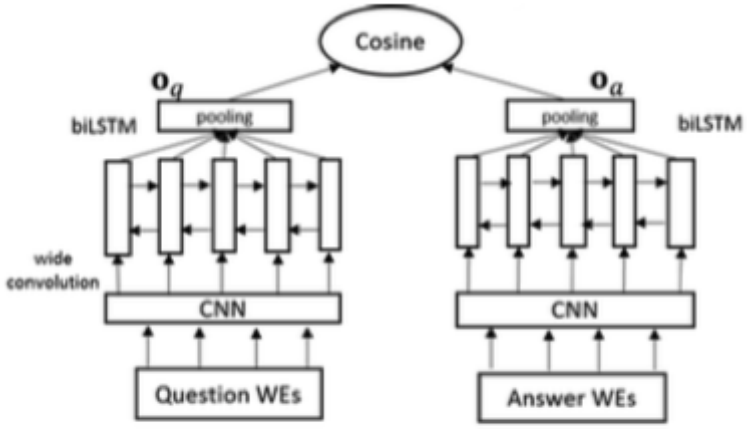


Figure 3: Convolution-based LSTM

(3) Attentive-LSTM

相比LSTM-DSSM, 在Attention机制上做了些改进，与NMT的Attention机制接近，即：通过Answer中的词向量加权平均生成整个Answer的向量时，每个词的权重是由Question向量和词向量来决定的。Question的表达仍由其所有词向量的avg或sum，max来表示。

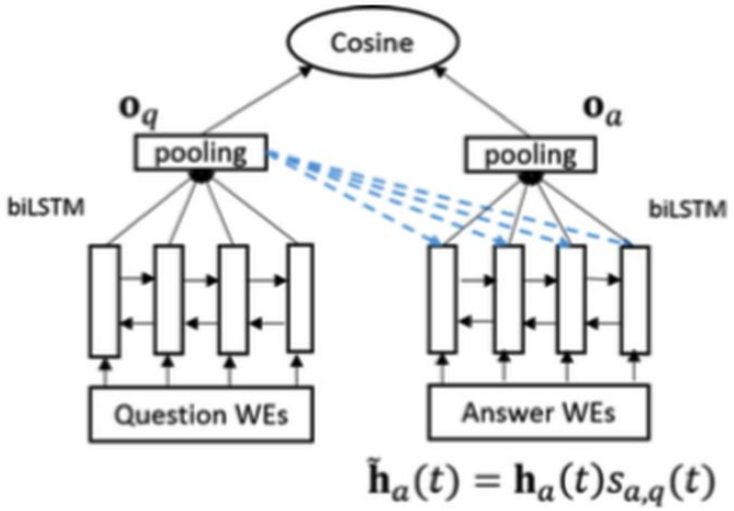
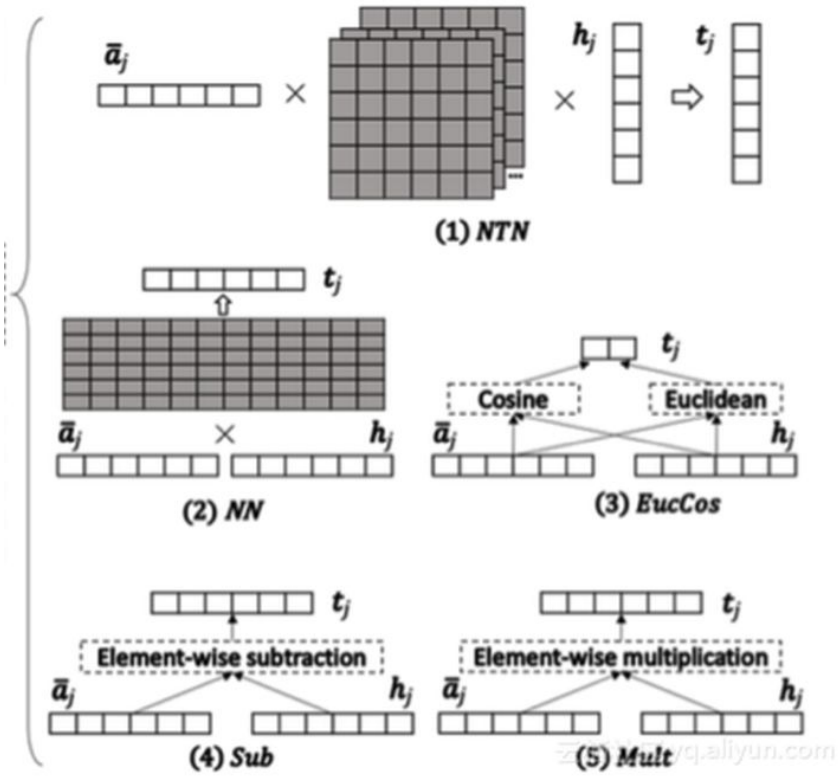


Figure 4: Attentive LSTM

4. 其它相关工作

上述工作主要集中在如何更好生成Query和Doc向量表达，如何设计两个向量comparision



另外在机器阅读理解也有很多类似工作，本文就不展开描述了。下面介绍下我们的相关工作。

5. 我们的工作

我们对淘宝搜索做了大量的语义改写后，matching不仅局限于term的匹配了，下面分别从数据和模型介绍下我们的工作。

5.1 深度模型通常大量的训练数据，而对商品搜索相关性这个问题，获取大量高质量训练数据并不容易。网页搜索通常直接采用点击数据作为是否相关的label，在商品搜索上不是很有效：用户点击行为与价格、图片、个性化偏好等很多因素相关，仅依赖点击数据对相关性样本有太多噪声；而采用人工标注数据，准确率相对较高，但受时效性、成本等因素限制较大。最近学术界也逐渐意识到这个问题，提出BM25等无监督模型生成大量样本。我们获取训练数据的方式有：

(1) 对行为数据采样，并用一些类似图像Data Augmentation的手段获取大量(亿级别)准确率相对较低的训练数据，先用这些数据training一个较好的模型；这些方法包括：

a. query下取CTR正常的商品作为正样本， CTR低于平均值较多的商品作为负样本

b. query能召回的类目下随机采样商品作为负样本

c. 对query中的term做一些变换，用变换后的query下点击商品作为原始query的负样本，例如“红色长袖连衣裙”变换成“蓝色短袖连衣裙”， 而“蓝色短袖连衣裙”下点击商品可以作为“红色长袖连衣裙”下的负样本；

(3) 采用数量相对少(100w)、准确率高的人工标注数据fine-tuning用上述两种方法pre_training

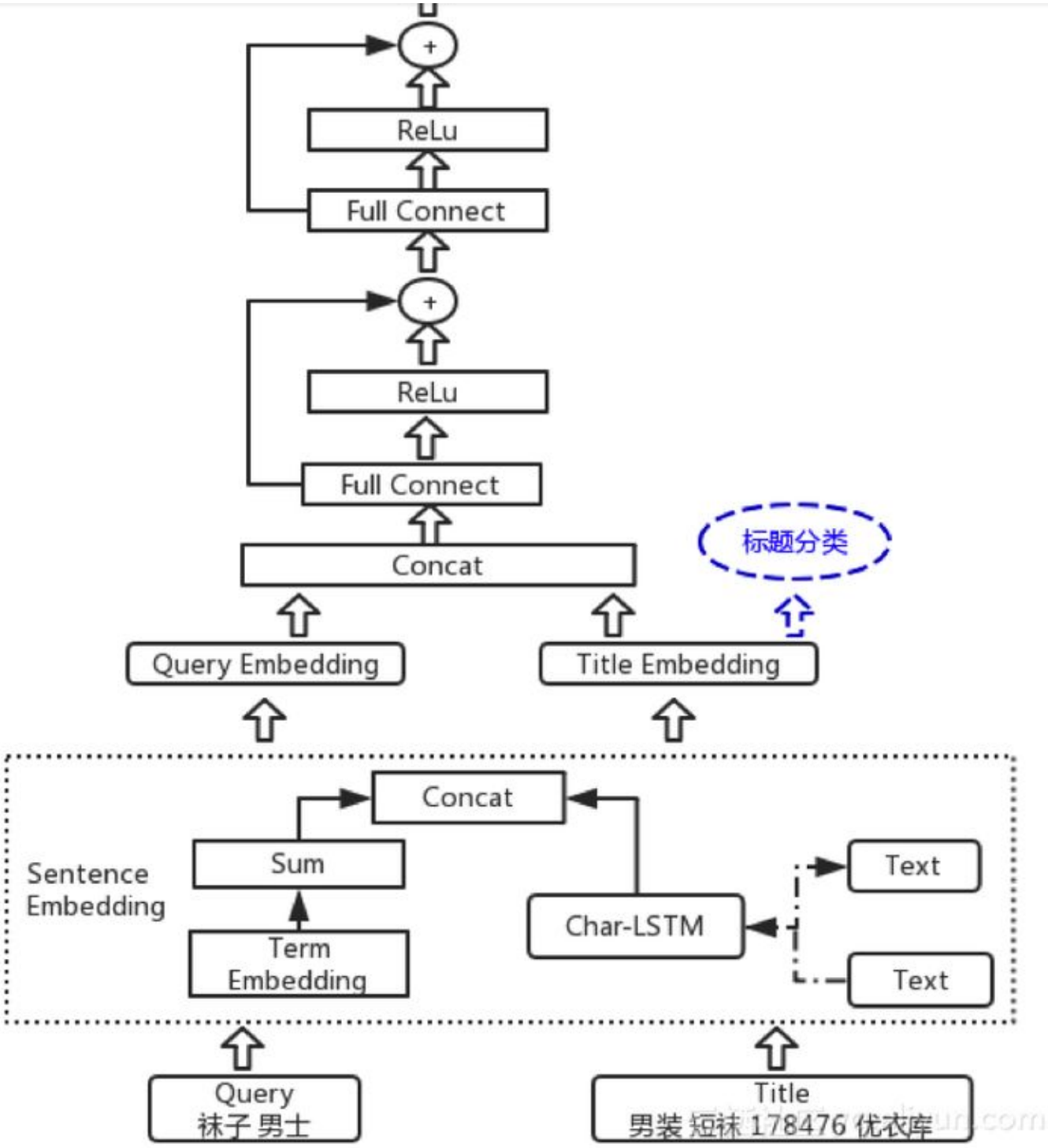
5.2 模型设计主要考虑的几个因素：

(1) 淘宝上Query和商品标题存在大量长尾词，尤其大量数字和英文组合的货号、型号、容量等，分词无法穷尽。仅通过词来对query和标题embedding会损失很多信息，需要考虑字符维度。

(2) 商品除了标题外了，还有图片、类目、属性等信息可以利用。

(3) 工程实现线上计算要轻量，两个向量的compare function要控制计算复杂度。

我们现在采用的模型如下：



(1) 对Query和标题向量我们采用DNN + Char-LSTM组合的方式：DNN能高效地学到TOP词的embedding, Char-LSTM能捕获到较长尾的字符组合。引入Char-LSTM后模型比较难训练，我们使用query和标题文本语料pretraining LSTM-AutoEncoder, 获得比较好的初始参数；同时TOP词的embedding采用word2vec初始化，模型能更快收敛。

(2) 在商品标题的embedding上增加了一个类目预测的辅助task, 使得不同类目的商品在向量空间内有更好的区分度，对模型效果和收敛速度都有比较好的提升。

好，但计算量增加会很大；借鉴ResNet全连层设置窄一些，并加深模型，可以保证效果同时较大减少计算量。

我们抽样部分query抓取线上排序结果， 与该模型排序后TOP30人工评测GOOD比例提升1.31%。

5.3 后续计划

商品除了标题和类目，图片也是很重要的信息来源，后续加入图片信息，同时也在尝试用query和商品向量做召回，实现multi-modal检索。

另外，Attention机制也是一个被证明重要的提升点。受限于线上ranking latency的要求，不可能对每个商品标题根据query来计算其"关注"的部分，但可以引入一些self-attention的方法来生成更好的标题向量。

参考文献：

[1] Shen, Y., He, X., Gao, J., Deng, L., & Mesnil, G. (2014). A Latent Semantic Model with Convolutional-Pooling Structure for Information Retrieval (pp. 101–110). Presented at the the 23rd ACM International Conference, New York, New York, USA: ACM Press. doi.org/10.1145/2661829...

[2] Services, E. U. C. (2014). Learning Deep Structured Semantic Models for Web Search using Clickthrough Data, 1–8.

[3] Deep Sentence Embedding Using Long Short-Term Memory Networks: Analysis and Application to Information Retrieval. (2016). Deep Sentence Embedding Using Long Short-Term Memory Networks: Analysis and Application to Information Retrieval, 1–25.

[4] Zhai, S., Chang, K.-H., Zhang, R., & Zhang, Z. M. (2016). DeepIntent: Learning Attentions for Online Advertising with Recurrent Neural Networks

(pp. 1295–1304). Presented at the the 22nd ACM SIGKDD International Conference, New York, New York, USA: ACM Press. doi.org/10.1145/2939672...

[5] Mitra, B., Diaz, F., & Craswell, N. (2016). Learning to Match Using Local and Distributed Representations of Text for Web Search, 1–9.

[6] Improved Representation Learning for Question Answer Matching. (2016). Improved