Training and Evaluating Multimodal Word Embeddings with Large-scale Web Annotated Images

一. 写作背景及目的

1. 背景

目前最先进的词嵌入模型仅仅在纯文本上进行训练,缺乏对于视觉语义信息的有效利用.同时一些包含视觉语义信息的数据集规模却非常小,难以达到与纯文本数据集同等的规模,或者一些数据集中仅仅包含某些有限的且预先已经定义好的视觉概念.

2. 目的

提出一种更有效的词嵌入方法，即同时利用文本信息与图像信息进行词向量嵌入，同时整理并开放了一个大规模的包含句子描述的图像的数据集

二. 核心问题

目前词嵌入主要存在以下几点问题：

1. 缺少大规模的包含图像与文本描述的数据集。
2. 如何直接评估不同任务的模型的质量(比如寻找相关短语或者单词的任务),即缺少评估词嵌入有效性的数据集.

三. 现有状况及相关研究

1. 包含句子描述的图像数据集规模大多比较小,比如ms coco数据集近包含100万个句子.同时这些数据集中描述图像所用的语言比较简单.
2. 最近发布的YFCC100M数据集,虽然提供了丰富的图像相关信息,但是大部分与摄影技术相关的信息,并未直接描述图像所包含的内容.
3. 单词相似性评估基准数据集.目前有WordSim-353/WS-Sim,MEN],SimLex-999等数据集,但是大部分都是由几百个相关的单词对组成的,规模不是很大.
4. 目前RNN-CNN模型大量被用于作为图像描述生成训练的基本模型,作者受此启发,采用了RNN的变体(GRU)
5. 目前对于利用视觉信息进行词嵌入的训练主要有两种方式.其中一种是首先分别提取文本和图像特征,然后使用奇异值分解,再通过堆叠的自编码器或者简单的串联过后,然后在skip-gram模型中通过融合视觉与感知信息,来共同学习文本与图像的特征.但是因为缺乏大规模的多模态数据集,这种方式只能将视觉内容与预先定义好的有限的名词集进行关联,达到的词嵌入效果有限.

四. 作者所做的工作

1. 构建了一个用于训练多模态信息嵌入的训练数据集.作者首先在Pinterest上抓取了超过4000万张图片,然后每张图片平均与12个句子相关联,然后去除了重复的句子或者少于四个单词的短句.判断某个句子是通过计算单词的unigram重叠比率来实现的.最后构建出来的这个数据集包含4000万张图片,以及3亿个相关句子描述,远远大于之前的图像描述数据集.同时因为这些描述信息主要来自对于该图片比较感兴趣的用户,所以这个数据集中的描述信息相比别的数据集更加的自然与丰富.
2. 通过整理Pinterest图像搜索系统的用户点击数据,作者又构建了一个评估数据集.这个数据集由一些包含短语A,短语B与短语C的三元组构成,其中短语A与短语B在语义上比短语A与短语C更加相近点.这种相对比较的方法通常用于评估和比较不同的词嵌入模型.同时作者又将这个数据集分成了两部分:

- 基于用户点击数据的原始评估数据集.首先,给出一个用户的搜索信息,这个搜索系统会返回一系列图片以及描述这些图片的标志信息,然后我们选择点击率最高的标注信息添加到正样本集合中.同时为了增加难度,我们需要从正样本列表中删除包含与用户输入的信息含有公共单词的样本.然后再从10亿个短语中随机的抽取一个短语作为负样本.评估时,通过分别计算正负样本短语的与原始短语的相似性来区分正负样本短语,以其分类正确率来作为模型评价标准.作者采用此种方式采集了接近1千万个三元组样本.
- 黄金标准版评估数据集.因为原始数据集主要基于用户点击信息构建的,所以里面可能存在不符合规范的数据.比如正短语与原始短语不相关或者负短语与原始词强相关.所以作者使用了CrowdFlower众包平台对原始数据集进行清理.当标注时,正短语与负短语分别随机的放置到A或者B选项上,标注者需要选择哪个短语与原始短语更相关,或者都完全不相关.每个词组被分到三到五个标注者进行标注,只有当超过百分之五十的标注者的结果相同时,此样本才会被添加到最终数据集上.作者采用这种方式最终收集了超过一万个三元组样本.

3. 作者提出了三种基于RNN-CNN的模型来学习多模态词嵌入.所有的模型都包含两部分:一部分是用于提取视觉特征的卷积神经网络,另一部分是用于对句子进行建模的RNN网络.在CNN中,采用224x224的图像输入大小,再采用16层的VGGNet进行视觉特征提取.然后将softMax层之前的层的数据进行二值化作为图像特征,然后在模型A与B中映射入RNN中,在模型三中图像特征被嵌入到与句子相同的空间中.在RNN部分中,作者使用了包含512个GRU神经元的网络结构.

五. 实验验证

1. 首先将Printerest 40M数据集中的句子中的单词转换为小写.然后分别在句子的开头与结尾部分添加开始与结束符号.
2. 训练时采用随机梯度下降法,采用256的批次大小,初始学习率为1.0,然后开始最模型进行训练,知道损失不会继续下降为止.
3. 评估时使用训练的嵌入模型来提取短语中的单词的嵌入表示,然后比较原始短语与正短语和负短语之间的余弦距离来判断嵌入表示是否正确.

六. 结论

1. 根据最终的测试结果表明,当模型成功的将视觉信息与文本信息融合在一起时,视觉信息可以有效的帮助提高单词嵌入的训练效果.模型A分别在Gold RP10K(黄金标准评估数据集)和RP10M(原始评估数据集)数据集上优于Word2Vec模型9.5%和9.2%。模型C也优于纯文本RNN模型的表现.
2. 权重共享策略可以增强模型将视觉信息融合到词嵌入训练的能力.例如,模型A采用权重共享策略时,其正确率在Gold RP10K(黄金标准评估数据集)和RP10M(原始评估数据集)数据集上分别相比没有采用时高了2.5%与4.8%.
3. 模型A在所有三种模型中表现最佳。它表明,权重分享策略所施加的软监督比直接监督更有效。
4. 在Pinterest40M数据集上训练的所有模型都比在更大纯文本数据集上训练的skip-gram模型表现的更好。