Incorporating visual features into word embeddings: A bimodal autoencoder-based approach

一. 写作背景及目的

1. 背景

多模态语义表示是自然语言处理和计算机视觉研究中的一个不断发展的领域。而且最近将感知信息(如视觉特征)和语言特征融合在一起的研究,正在变得越来越热门.

2. 目的

作者提出了一个新颖的用于多模态表征学习的双模自动编码器模型,其中自编码器可以结合对应的视觉特征来增强对应的语言特征向量.同时作者进一步研究了增强型词嵌入在区分正反义词与模糊的相关词时所具有的潜在的效果

二. 核心问题

目前多模态语义表示方面主要存在以下几点问题：

1. 没有对单词进行零样本表征学习
2. 没有充分利用有用的资源(如使用图像特征进行增强型词嵌入)

三. 现有状况及相关研究

目前的多模态表示学习方法主要可以通过信息的融合或者整合方式来进行分类.作者在论文中主要列举了两类方法.

1. 一类是对单词-特征矩阵进行奇异值分解,其中的单词-特征矩阵主要由语言向量和对应的视觉特征向量串联组成.其中语言向量通过Strudel方法生成,然后视觉特征向量使用依赖于BOVW的传统卷积特征提取方式来获得.
2. 第二类是简单的进行向量拼接.其中的视觉特征通过卷积神经网络提取,语言特征通过Word2Vec方式提取.但是这些方法都无法处理零样本学习.
   为了解决多模态表示背景下的零样本学习问题.Lazaridou等人提出了可以合并视觉特征的skip-gram模型,该模型通过学习共同预测语言和视觉特征来建立单词向量.但是由于该模型是一个联合学习的过程,所以无法独立的使用现在已经存在的独立的语言资源或者相关的视觉特征资源.

四. 作者所做的工作

1. 作者提出的ViEW模型解决了多模态表示背景下的零样本表示学习问题,同时该模型没有采用联合学习方法,所以不会产生上述的问题.该模型的核心是一个双模的自编码器,通过设置特定的损失函数来集成双模输入.
2. 该模型在运行时,首先把字嵌入向量送入网络中进行正向计算,然后h2层进行视觉特征与正向计算的词嵌入向量进行合并,并采用H3层向量作为多模态语义表示.

五. 实验验证

1. 任务与评估方式

由于语义相关性涵盖了词语之间比语义相似性更广泛的词汇或语义关系,所以作者采用了标准语义相关性任务来评估多模态表示模型的性能.

2. 数据集

作者采用了三个数据集来评估多模态语义模型:

1. MEM数据集.此数据集专门用于评估多模态语义模型.总共包含了由751个不同单词构成的3000个单词对.同时数据集中的每个单词都包含了一个词性标签.
2. SimLex-999数据集.此数据集由999个单词对构成,主要用来评估模型捕获语义相似性的能力.
3. SemSim/VisSim数据集.此数据集总共包含了7576个单词对,并且同时提供了语义相似度和视觉相似度的标注.

3. 语言特征

作者通过采用skip-gram模型提取了300维的词嵌入.采用的语料库是enwiki9.

4. 视觉特征

作者采用了GoogLeNet来提取图像的视觉特征,具体方式是每个数据从图像数据集中选择50-100个相应的图像,然后求得所有对应图像通过GoogLeNet提取的隐藏层向量的平均向量,作为视觉特征,获取到的视觉特征为1024维.图像数据集主要采用了ImageNet数据集与ESP-Game数据集.
因视觉特征维数必须与语言特征相同,所以必须对视觉特征进行降维操作,作者采用了PCA(主成分分析)和AE(自动编码器)两种降维方式进行对比分析.

六. 结论

1. 在非零样本学习条件的设置下(即测试数据集中所有单词都参与过双模自编码器训练),ViEW模型采用的多模态表示方式性能优于单模态模型.同时相比于其他模型,ViEW模型达到了最先进的性能.
2. 经过对比分析,发现PCA是减少视觉特征维数的更好办法.
3. 在零样本学习条件的设置下(即测试数据集中包含了没有相应训练图像的单词),ViEW模型在MEN数据集上达到了最好的性能,但是在其他两个数据集上MMSG模型达到了最好的效果,这表明,ViEW模

型还可以再继续提升优化.

4. 多模态语义表示具有过滤语义相关但是上下文不相关的单词的潜能.
5. 同时通过对从ImageNet和ESP-Game获得的视觉特征进行t-SNE可视化后发现,ImageNet中的图像在描绘目标概念时是比较规则的,而ESP-Game图像通常是嘈杂的。