**华东师范大学数据科学与工程学院上机实践报告**

|  |  |  |
| --- | --- | --- |
| **课程名称**：人工智能 | **年级 大三** | **上机实践成绩**： |
| **指导教师**：李翔 | **姓名**：俞致远 |  |
| **上机实践名称**：多模态情感分析 | **学号**：10195501403 | **上机实践日期**：6.10-6.20 |

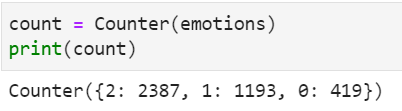
1. **实验目的**

给定配对的文本和图像，预测对应的情感标签。

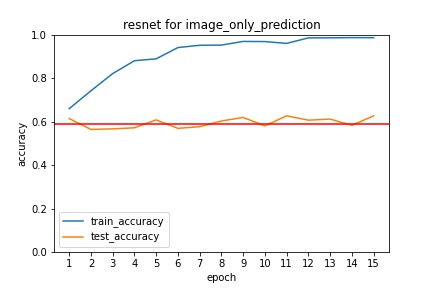
1. **实验过程**

**（1）仅使用图片，采用不同的模型和训练方式，查看验证集上结果。我将有标签数据的20%作为验证集，其余作为训练集。**

值得注意的是，虽然该问题为三分类问题，但各个类别的数据是极不均匀的，如下图所示，2代表正面情感，1代表负面情感，0代表中立情感，可以看到一半还多的样本都是正面情感，这样一来，如果我将所有结果均预测为positive，那么我最后的准确率仍能达到60%左右，因此如果模型在验证集上不能达到这个精度，那么还不如都预测为正面情感来的准。



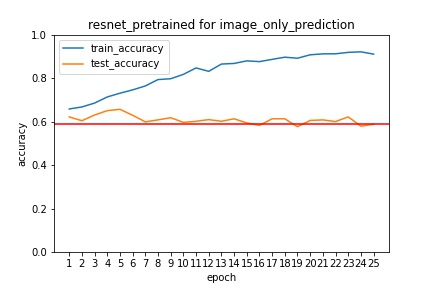
**1、仅给定图片数据，使用ResNet网络从头训练。**最后训练以上精确率几乎达到100%，但是验证集上只有60%左右，红色的线是正面情感的样本在所有样本中的占比，验证机准确率几乎是贴着红线在走，说明模型是不合格的。



我觉得可能的原因有两点：第一是数据量可能太小，训练数据一共三千多条，而ResNet是针对ImageNet设计的，数据量显然不对称，如果从头训练必然会导致过拟合；第二是有可能好多图片本身没有表达出情感信息，比如下面这张图片，需要文字来进一步阐释。

****

**2、使用ImageNet上预训练好的ResNet提取特征，在后面添加一层全连接做分类，训练时仅调整最后的全连接参数**，实验结果如下：



可以看到结果依然不尽如人意，我想到原因仍然有两点：一是ImageNet是针对“图片主体是什么”这个问题做图像分类，而没有涉及情感信息，和本次任务之间存在语义鸿沟；第二点仍然是因为仅从图片本身无法看出情感信息，就算叫人类来识别，也很难通过图片识别出情感信息。

1. **使用CLIP模型做预测。**既然之前效果不好的原因有可能是图像的情感分类数据量不足，一个自然的想法就是使用更大规模的数据集，那么数据集从哪来呢？目前比较有名CV数据集中似乎没有和本任务相关的。但是我突然想到之前看到的OpenAI的CLIP工作，该工作将大量的图片-文本对做对比学习，使用数据为网上爬取得到，数据量极大，泛化能力极强，能衡量任意图片和文本的相似度，比如要区分一张图片是猫还是狗，只要比较该图像和“this is a dog”与“this is a cat”两句话之间的各自的相似度即可。那么本实验是否也可以将三分类的情感问题转换为图像与三句话之间的相似度比较问题呢？这三句话又应该如何设定呢？我尝试了不同的话语来表达情感并观察模型效果，实验结果以表格形式给出：

|  |  |
| --- | --- |
| 文本端输入 | Precision |
| "That is positive", "That is negative", "That is neutral" | 0.422 |
| "it's nice", "it's sad", "it's ok" | 0.574 |
| "it's good", "it's bad", "that's the case" | 0.621 |

当然期间还有一些别的尝试：



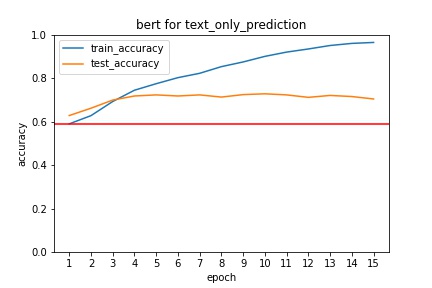
多次尝试后，似乎0.621就是极限了，而且由于CLIP模型太大，微调根本跑不动，我花了好长时间才跑了一个epoch，验证集上精确率反而降到了0.61，我就想到，仅使用图片预测结果应该是不合理的，0.6的预测准确率无法突破了，需要文字来辅助。

**（2）仅使用文字，采用不同的模型和训练方式，查看验证集上结果。**

从直观的角度看，文字对情感判别的贡献应该比图片更大一些，比如下面这张图片及其配对的文字，如果仅看图片，图片里只有一台老式的唱片机，但是配图的文字是***Just finished off installing this simple yet superb Hi-Fi system。***可以看到作者的心情应当是非常喜悦的。



1. **仅给定文本数据，使用Bert模型做微调。**微调结果如下，看到使用文本后预测准确率可以比红线（全预测为正面情感的准确率）高了将近十个百分点，说明模型是合格的。

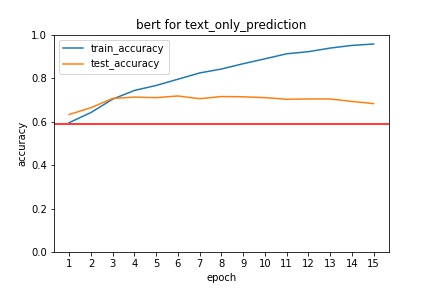
****

能否进一步提高呢？我想到由于bert预训练的句子都是比较完整流畅的，而数据中的句子有大量标签，网站等混杂信息，破坏了句子完整性，是否可以做些预处理呢？

**2、对文本做预处理后再用bert微调。**由于这个应当是社交媒体上的数据，这些文字大多凌乱不堪，好多不是一句规整的句子，我观察了一下，基本是以下三种形式的文字导致的：

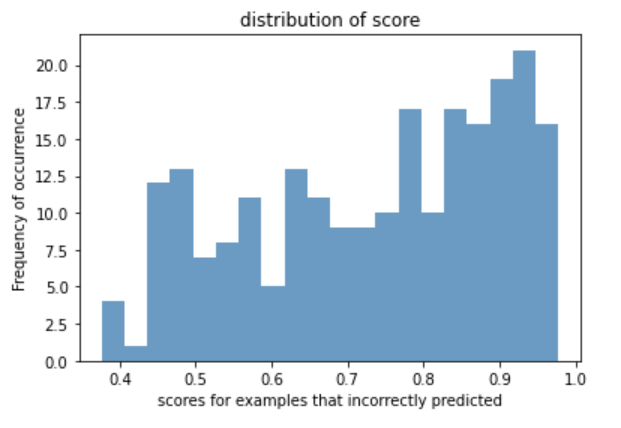
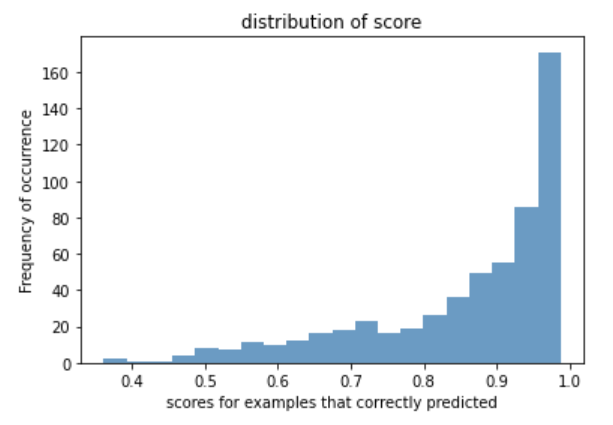
|  |  |  |
| --- | --- | --- |
| 形式 | 描述 | 是否对分类有作用 |
| @…… | @某人的文字，通常在开头或结尾 | 大部分情况下没有用 |
| http:…… | 网站的链接，通常在开头或结尾 | 大部分情况下没有用 |
| #…… | 标签的信息，开头结尾和正文均有存在 | 非常有用 |

因此，我选择删去“@某人”的文字和网页链接的文字，保留标签的文字，对文本预处理后，尽可能保证文本句子的连贯性和完整性。将预处理好的句子送进bert做微调，最终实验结果如下图所示，令我比较失望的是，结果和没有预处理的文本信息几乎一致：

****

我想可能是因为@某人和网页的文本信息本身就因为OOV的缘故没有被送进bert，就算送进去了也没有太多干扰信息，因此预处理可能是不需要的。

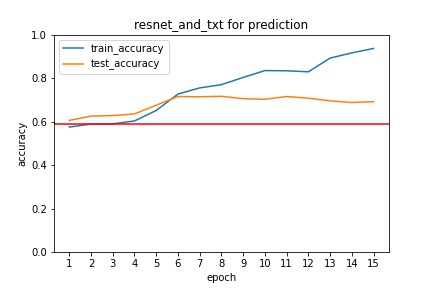
**3、观察究竟是哪些样本分错了。**观察分类结果的“得分”的分布，这里的得分定义为三分类最终输出的三维向量进行softmax后最大值的大小，比如某个样本最终输出结果为[0.1, 0.1, 0.8]，那么他的“得分”即为0.8。显然，得分越高说明模型越确定该样本属于某个类别，得分分布如下图所示：



由图可知，对于正确分类的样本，模型大多“确信”它属于这个类别，而对于错误分类的样本，模型似乎不能确定他们究竟该分为哪一类，以至于得分分布看上去像是一个均匀分布。观察分错的语句，可以看到大部分语句中几乎没有情感信息（但是分类却不是中性），要么就是语言并非英语。

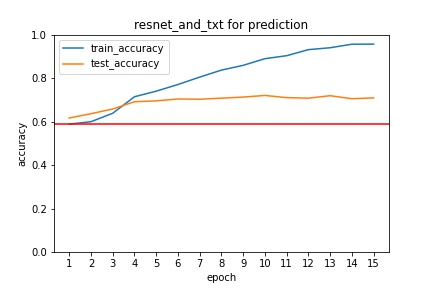
**（3）使用图片结合文字的形式，采用不同的模型和训练方式，查看验证集上结果。**

**1、concat方式结合。**结合图片和文字，最简单的方式将两者的特征经过线性变换后做一个拼接，文字的表征通过bert得到，而图片的表征通过resnet得到，两者都映射到128维后做一个拼接。最后映射到三维向量上做分类。最后结果如下图所示。



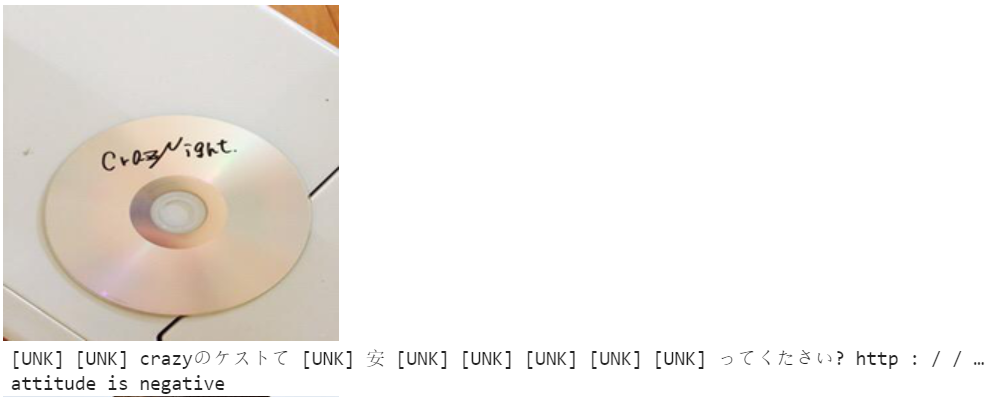
可以看到结果并没有比单纯的文字好多少，可能的原因是该模式相当于静态的给图片和文字加权，对某些样本可能文字蕴含的情感信息更多，而对另一些可能图片更多。因此可能动态加权要好一些。

1. **利用attention机制对图片和文字特征进行动态加权后相加。**我设计的模型是，增加一个query向量作为参数，与图片文本特征分别做内积后求得权重，并以此权重对特征做加权求和。最终结果如下：



模型最高到了74%，但还是没有太大提升。

1. **观察究竟是哪些出错，以及能否避免。**我打印出了部分分类错误的图片和文字以及对应的情感，代码中有展示，这里选几张展示以及分析预测错误的原因。



此图片对应的文字为日语，bert-base-uncased虽然是多语言的，但是显然有大量的日语没能识别出来，这个数据可能是twitter上爬来的，因此有多种语言，日语韩语法语都有，这些语言一方面不能被完全识别，另一方面数据量少，因此分错是可以理解的。



此条推特我也是没法分辨的，我无法体会其中的负面情感。



这个可能模型也是懵的，为啥有slave这样的词汇依然是正面情感（少儿不宜）。

因此可以说，相当一部分图片无法正确被分类是由于语言问题，数据脏，数据量不足导致的。

1. **实验总结**
2. **遇到的困难和bug以及解决方案**

本次实验没有遇到什么bug，非要说有的话就是维度不匹配，还有模型和数据device不匹配这样的小问题，都是粗心导致的，并且报错以后很快能找到问题所在。

本次实验首先自己搭建了一个resnet网络，但是自己搭建的网络需要输入图片大小为3\*224\*224，因此要对原图像进行调整，不能直接cut，因为部分图片大小没有224\*224这么大，因此使用PIL模块自带的resize方法。随后使用torchvision提供的预训练好的resnet网络，这个就没有图像大小的要求了。

另外，最后上传到github上时遇到两个问题，一个是密钥没有更新，另一个是data文件太大传不上去，前者很容易解决，后者我尝试使用网上推荐的LFS，但是只能上传压缩文件，因此我最终选择了分批上传。

1. **实验的总结以及有待完善的工作**

本次实验是多模态情感分类，仅通过图片判断准确率在60%左右，仅通过文字判断准确率在70%左右，通过图片和文字共同判断准确率也仅仅上升到74%，不尽如人意。导致这个现象的根本原因我认为是数据量不足，只有4000个训练样本，相比较之下ImageNet有一千万的训练样本。此外，文本和图片彼此相对独立也是模型效果不好的原因，虽然我用到的CLIP模型是对文本和图片做了语义的匹配，但这个模型也仅仅只是做了“匹配”的工作，而不能根据两者之间的关联关系互相补充信息，生成一个更具泛化能力的表征。另外多种语言混杂，文本格式混乱，部分图片不含有充分语义信息也是次要原因。