MIL多示例学习知识

参考文献：

From Captions to Visual Concepts and Back

这篇文献的第一部分就是用MIL的方法从图片中检测出词及相应的框，这些词来自一个general的词库；

下面解释一下MIL方法用于detection：

我们从一张图中框出多个patch，这些patch构成一个bag，一张图对应一个bag；

我们已经有一个标签词库（vocabulary size 是k），

这张图（一个bag）的标签就是k维向量，其中，在这张图的描述句子中出现的词，标为1，其余词标为0；

接下来，

我们用一个经典的CNN分类网络（Alex或者VGG等）来从每一个patch中学习特征，并把最后一层输出换成k维，输出每一个patch中包含不同词的概率，

然后用一个bag中所有patch的概率计算得到一个bag中包含不同词的概率（k维）；

最后，

对bag的预测概率和标签用交叉熵损失函数训练网络（实际上相当于对k维词库中的每个词独立训练了一个detector），

每次用训练好的网络来更新bag的标签，

再用更新的标签重新训练网络，

一定次数后停止，得到训练好的检测模型；

得到的这个模型可以用来对测试图片的每个patch进行预测，来做detection任务，得到检测的词和相应的框，

在image caption的attribute中只关注一张图片中包含的词，而不关注对应的框，那么可以继续用patches的概率得到bag的概率， 即为最可能词的预测结果。