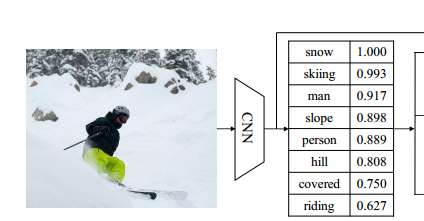
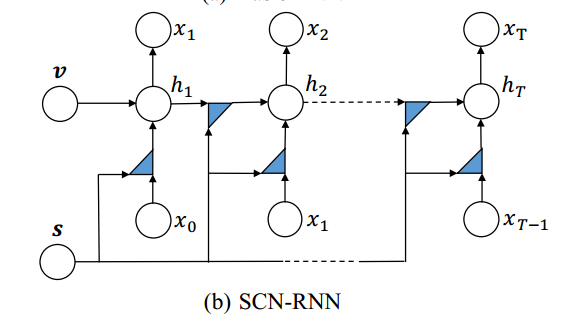
**一、主要思想**

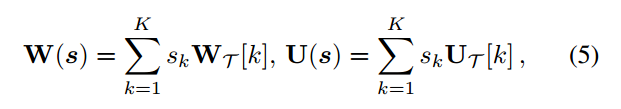
1、引入“semantic”概念，类似于attribute，即把caption库中的词汇作为标签，用CNN学习图片的语义特征，此语义特征作为后续RNN的输入；

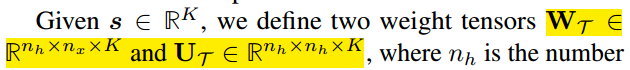


2、本文在图片语义信息输入RNN的方式上有所创新，把一个Embedding矩阵改为多个semantic dependent的矩阵，再对这些矩阵进行ensemble；



C:\Users\YanJack\AppData\Local\Temp\enhtmlclip\Image(2).png

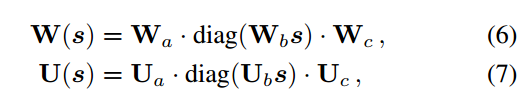


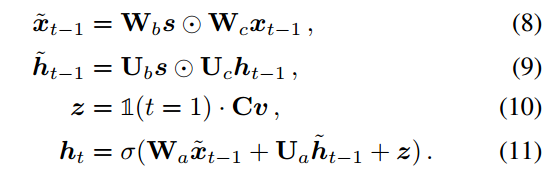


3、通过同时使用2D和3D（加入时序）的CNN提取视频特征，可以把本文generate caption的方法应用到video中；

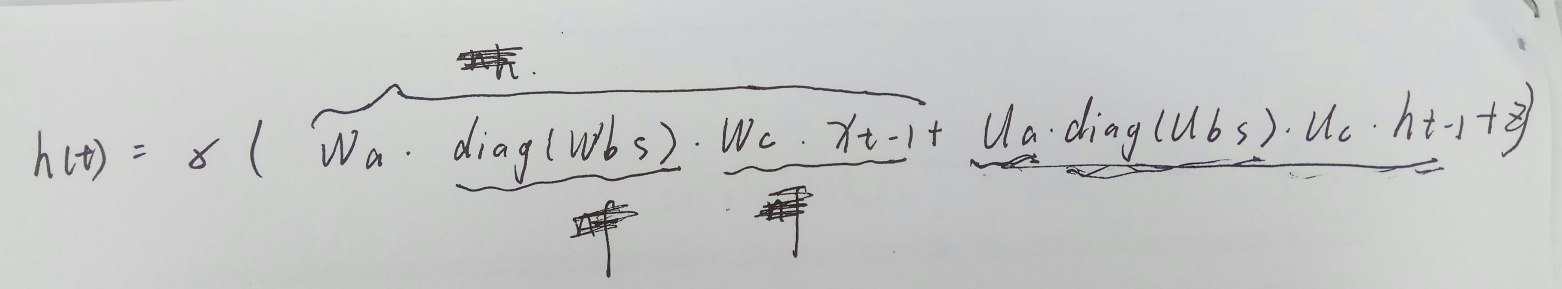
**二、细节**

1、由于semantic特征的维度很大，故提出一种矩阵分解的方法把semantic特征映射到nf维（nf<K）；公式如下，其中Wb是nf\*K维，s是K维，Wb\*s是nf维，diag表示把Wb\*s的元素作为对角元素展成nf\*nf维矩阵；





2、本文的参数量是nf\*（nx+2K+3nh）= nf\*nx + nf\*2K + nf\*3nh，其中nf\*nx对应Wc的参数，nf\*2K对应Wb和Ub的参数，nf\*3nh对应Uc，Wa，Ua的参数；



**三、idea**

本文提出的“semantic”其实就是“attribute”，

区别在于“attribute”是以guide的形式作为LSTM的输入，“semantic”是以权重的形式对Embedding矩阵进行ensemble，

由此想到可以在输入方式上做文章；