论文笔记《Towards generating long and coherent text》

论文来源： 2019 ACL

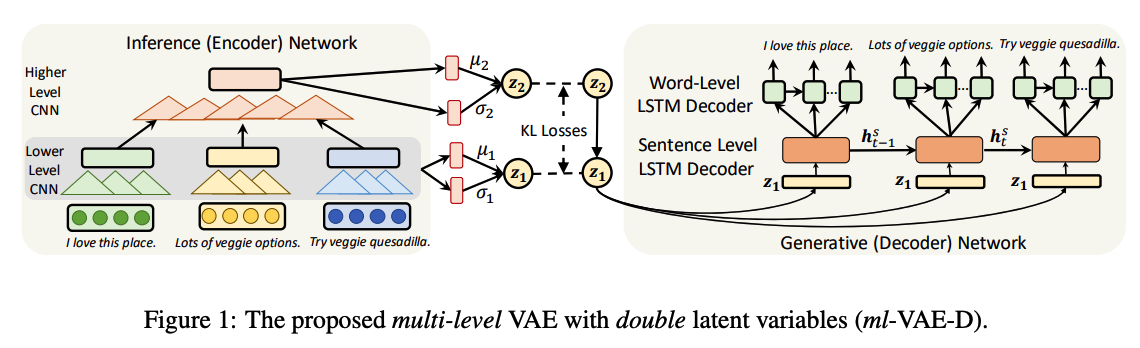
论文主要内容：

本文提出了various multi-level network structures for the VAE model(ml-VAE)，将VAE结构适用于生成全局语义连贯的长文本序列，不像以前使用VAE大多是在一个句子（少于20词）上的。

ml-VAE主要包含两部分，一个是higher-level提取出的文本语义，论点等特征，一个是lower-level提取的是文本的一些词特征，如特殊词的选择。

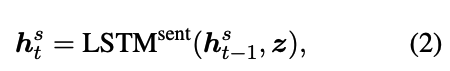
文中定义了一个hierarchical RNN decoder，分别获取sentence和word-level的表示。文中经过实验发现，与其将隐层变量直接作为RNN decoder的初始状态，不如先将隐层变量经过一个higher-level(sentence) RNN decoder产生一个输出，然后再把这个输出输入到一个lower-level(word) RNN decoder来生成词，这样的实验效果更好。这样的话lower-level decoder网络就会对隐层变量有个强依赖，也就避免了autoregression(自回归，也就是2016年第一篇用VAE做文本生成提出的问题，RNN decoder不依靠隐变量，完全靠自己的强大能力就可以生层目标结果)。

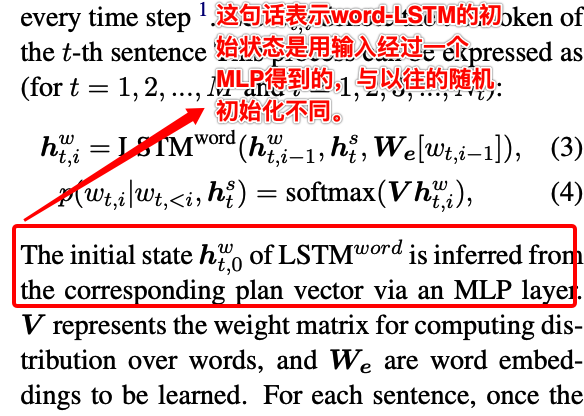
文中提出的结构：



**Single Latent Variable(ml-VAE-S)**

这里主要对应上图中的decoder部分，作者提出了一个plan vector的向量，这个向量其实就是上图中sentence level LSTM decoder的输出结果，注意这里是将隐变量作为LSTM的每个时刻的输入的（文中解释，这里的可以看做是获取了整个段落的语义表示），然后sentence level LSTM decoder的输出再作为word level LSTM decoder的输入来决定输出的词。





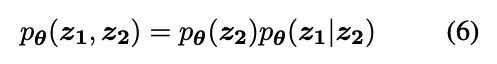
**Double Latent Variables(ml-VAE-D)**

这一部分主要是对应生成多个隐变量，对应图中的左边encoder部分，文中的图是有两层的隐变量和，encoder采用CNN逐步往上提取特征（这里就对应着句子不同维度的特征，比如char -> word -> sentence -> paragraph -> document），文中的在低层的CNN的参数是共享的，并计算均值方差生成各个层的隐变量分布【这里文中没有具体介绍，如何生成均值和方差，我觉得应该是和以前一样，经过全连接层生成两个向量】。但是图中decoder还是直接从进行生成的，那它是怎么利用到的信息的呢，文中说在计算Loss时用到了（目前这一块还没弄得很是清楚，并且图中的从指向的实线也没看到对应的转换）。

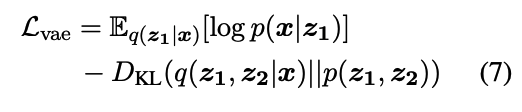
多层的multi-layer隐变量的另一个好处是可以解决VAE的后验爆炸问题(posterior collapse)或者说是KL散度消失。文中有一个假设就是和在encoder部分是相互独立的，因此多层次的隐变量的联合后验分布可以表示为（此时的参数为）：



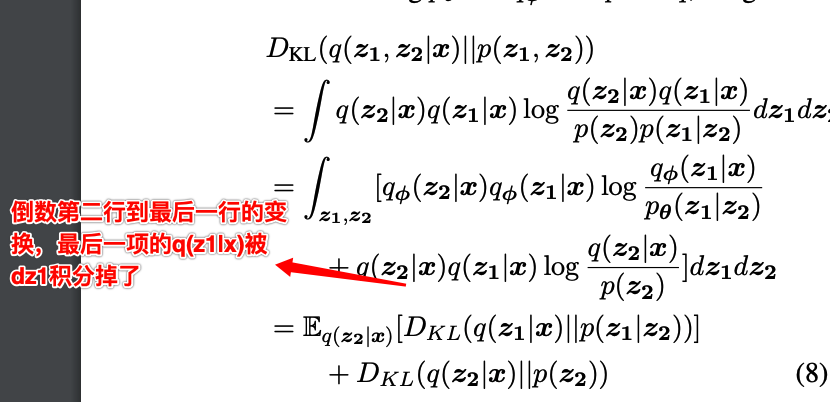
在生成网络中（此时参数为），是由在给定的情况下抽样得到的，因此有：



因此VAE的下届ELBO可以写成：



将第二项展开为：



后续则是在数据集上的一些实验结果。