论文笔记Generating Sentence from a continuous space

来源：2016 ICLR Stanford NLP

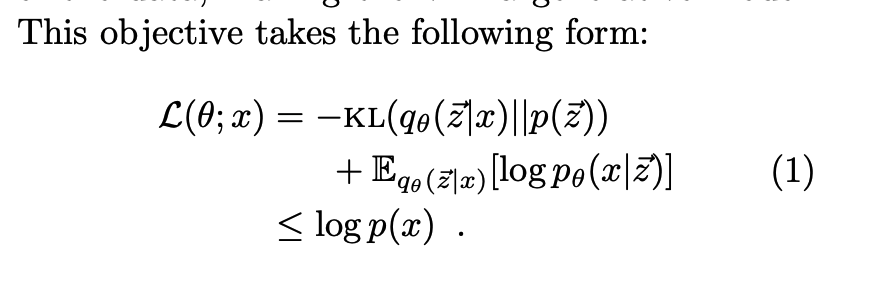
文章主要内容：

优化目标1：极大似然p(x|z)

优化目标2：z对x的后验分布接近其先验分布，即：p(z)和p(z|x)的KL散度尽可能的小。

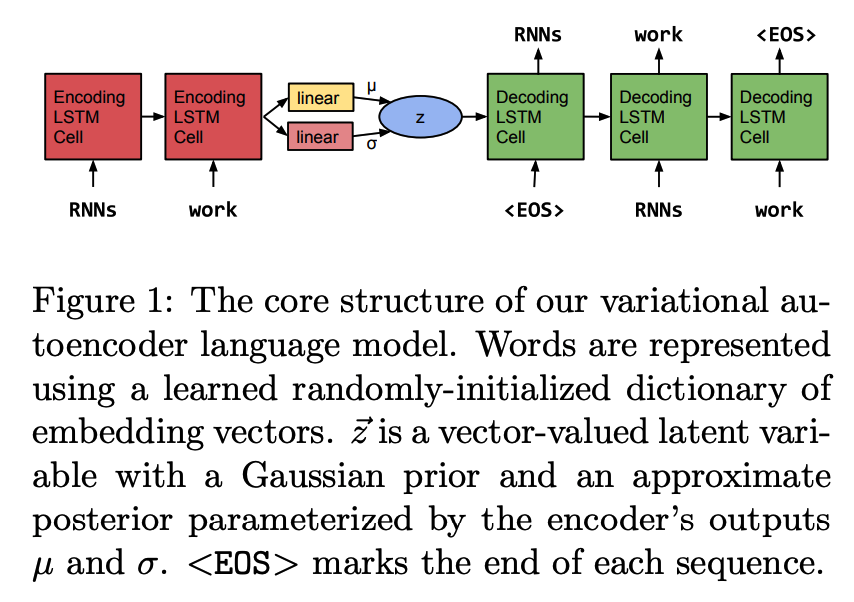
标准的RNN语言模型基于前一个词和一直在演变的隐层状态来预测每一个词，这样虽然有效，但它没有学习得到一个全句的表示。因此为了包含一个连续的潜在的句子表示，我们首先需要一种句子到分布式表示的隐射方法，这种分布式表示可以通过无监督学习的设定来训练。

VAE(变分自编码器)的训练目标如下：



其中是reconstruction quality，而是regularization term。

下面是论文的总体框架：

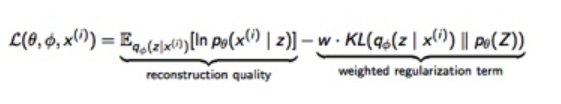


Loss的组成还是和VAE一样，具体到模型上，encoder和decoder都采用单层LSTM，decoder可以看做是特殊的RNN，它的初始状态【initial state】是z【隐藏变量】，z采样自Gaussian分布G，G的参数由encoder后面加一层linear layer得到，这里的z就是作者希望得到的global latent sentence representation。

模型很简单，但是在训练时有一个严重的问题：KL会迅速降到0，那么后验就失效了。原因在于，由于RNN的decoder有着非常强的modeling power，直接导致即使依赖很少的历史信息【隐层变量z】也可以让reconstruction errors降到很低，换句话说，decoder不依赖encoder提供的这个z了，模型等同于退化成RNN了。

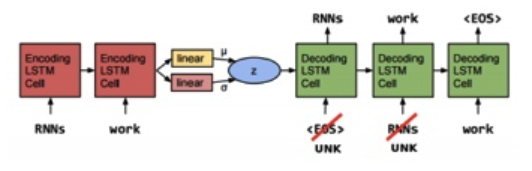
文中提出了两种方法来解决这种问题：KL cost annealing 和 word dropout。

* KL cost annealing：



作者引入一个权重w来控制这个KL项，并让w从0开始随着训练慢慢增大【从0到1】。作者的意思是一开始让模型学会encoder更多信息到z里，然后随着w增大再smooth encoding。其实从工程的角度看，因为KL这项其实更容易降低，模型会先去优化KL，于是KL很快就降为0。但是如果我们一开始乘以很小的w，模型就会选择忽视KL（这项整体很小不用再降低了），选择优先去降低reconstruction errors。当w慢慢增大时，模型也慢慢开始关注降低KL这项了。这个技巧在调参中也非常实用。

* Word dropout:



既然RNN-based的decoder能力太强了，那我们就来弱化它就好了。具体方法是把input的词替换成UNK，模型被迫只能去多多依赖z。当然保留多少input也需要尝试，我们把全部不保留的叫做inputless decoder。