论文笔记《toward controlled generation of text》

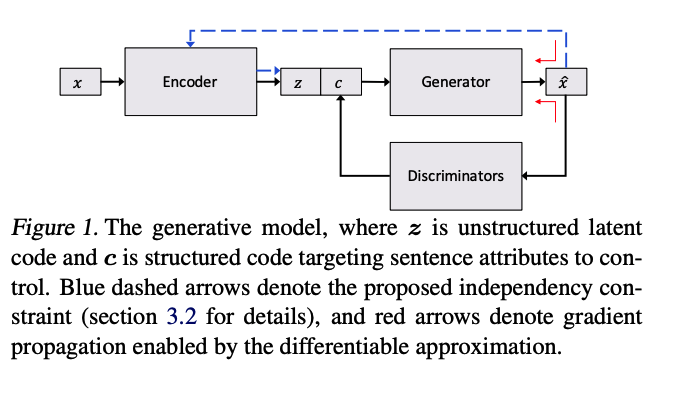
论文来源： 2107 ICML

论文主要内容：

这篇文章的主要内容是生成可控制方向的文本，其中方向就是文本的各种属性。

文中的框架是基于VAE提出的，将VAE与一个判别器相结合。为了实现定向文本生成，文中采用了一个新的变量c与原来的VAE中的隐变量z相结合，其中c关联着句子中想要控制的特定属性，比如情感，时态等，z代表其他的一些属性，通过变量的分布(z,c)来生成句子，再把生成的句子作为判别器D的输入，D再根据这些输入来训练改变c的分布。

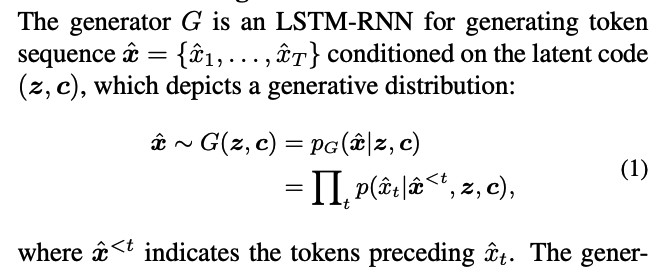
模型整体结构如下：



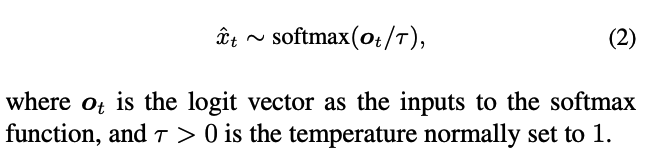
论文中没有具体介绍encode部分用的是什么，估计应该是LSTM等。下面对generator和discriminator部分进行介绍

**Generator**

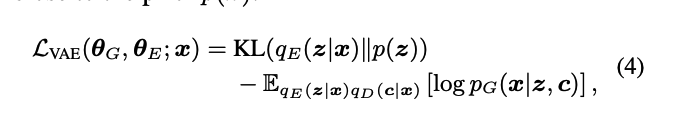
对于生成器的生成过程可以用如下的公式表示：



这里作者说由于每一步都直接取特定的单词作为输出，单词间的离散型将导致整个过程不可微，也就导致后续的判别器D无法优化，所以作者将每一步改为了一个softmax概率值，使它变成了一个可导的连续量【这里还没弄得很清楚】，如下：

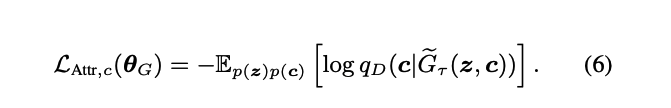


接下来就是标准的VAE的loss：

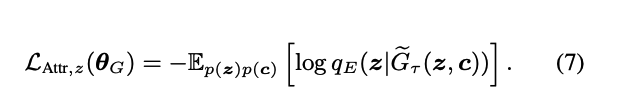


这里是将判别器D和c的分布qD(c|x)固定以后在进行计算的，也就是说这个过程不训练判别器D。

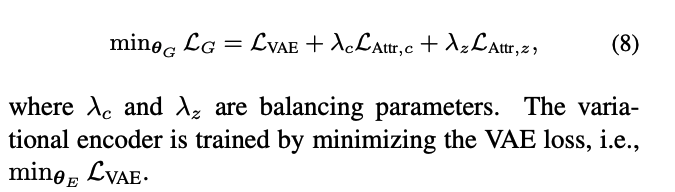
对于生成器G输出的文本，判别器D应当判别它是否与我们希望的属性相匹配，因此需要加入一项损失函数，如下：



同时为了保证属性之间的独立性，使其不产生相互影响，也就是判别器D只优化目标属性而不对其他属性进行优化。这里作者没有取训练一个新的判别器，而是直接将VAE中的编码器E拿过来使用。因此E对于z的作用和D对于c的作用可以视为一样的，都是将文本中的属性抽取出来得到其分布，因此，第三项loss为：

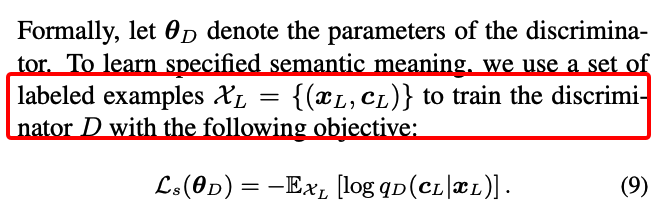


综上所述，编码器E和生成器G的损失函数为：

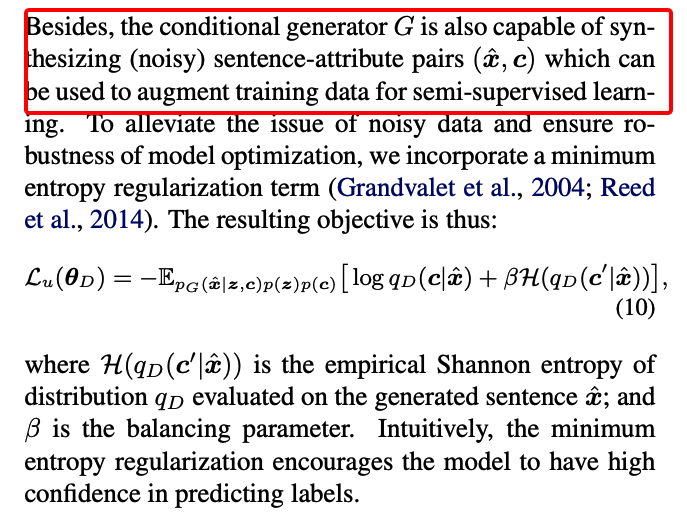


**Discriminator**

关于判别器的训练，是采用半监督学习的方法得到的，即少量的有属性的标签的样本和大量的生成器生成的无标签的样本进行训练，有监督的Loss如下：

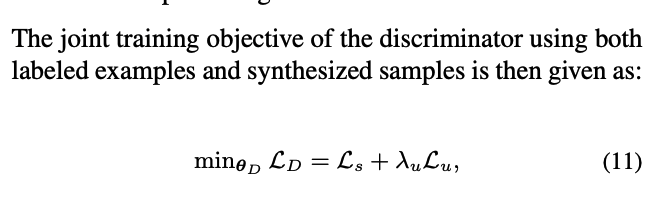


无监督的loss如下【无监督样本的生成方式如下图红框中所选的】：



其中H(qD(c’|x))是条件分布qD的熵。

因此判别器D的损失函数为：



总体的训练过程如下：

