**走向文本的受控生成**

## 胡智婷 1 2杨子超 1梁晓楠 1 2Ruslan Salakhutdinov 1Eric P. Xing 1 2

**摘要**

arXiv：1703.00955v4 [cs.LG] 2018年9月13日

与最近的视觉领域深度生成建模相比，文本的通用生成和操作具有挑战性并且成功有限。本文旨在生成似是而非的文本句子，其属性通过学习具有指定语义的解缠结潜在表示来控制。我们提出了一种新的神经生成模型，它结合了变分自动编码器（VAE）和整体属性鉴别器，有效地强加了语义结构。该模型可替代地被视为利用唤醒 - 睡眠算法来增强VAE，以利用假样本作为额外训练数据。通过对离散文本样本的可微近似，对独立属性控制的显式约束，以及生成器和鉴别器的有效协作学习，我们的模型甚至仅从单词注释中学习可解释的表示，并产生具有期望的情感和时态属性的句子。使用训练分类器作为评估器的定量实验验证了短句和属性生成的准确性。

# 介绍

对深度生成模型的研究兴趣激增（[胡等人。](#_bookmark29), [2017](#_bookmark29)），如变分自动编码器（VAE）（[金马和威灵](#_bookmark31), [2013](#_bookmark31)），生成性对抗网（GAN）（[Goodfellow等人。](#_bookmark23), [2014](#_bookmark23)）和自回归模型（[van den Oord等。](#_bookmark54), [2016](#_bookmark54)).尽管在视觉领域取得了令人瞩目的进步，例如图像生成（[Radford等。](#_bookmark42), [2015](#_bookmark42)），学习可解释的图像表示（[陈等人。](#_bookmark35), [2016](#_bookmark35)）和图像编辑（[朱](#_bookmark48) [等。](#_bookmark48), [2016](#_bookmark48)），自然语言生成的应用研究相对较少。即使生成真实的句子也是具有挑战性的生成模型

1卡内基梅隆大学 2Petuum，Inc。通讯作者：Zhiting Hu <[zhitingh@cs.cmu.edu](mailto:zhitingh@cs.cmu.edu)*>*.

*会*议纪要34 *日* 国际机器学习会议，澳大利亚悉尼，PMLR 70,2017。版权所有2017年作者。

需要捕获基础句子的复杂语义结构。以前的工作主要限于受监督设置中的任务特定应用程序，包括机器翻译（[Bahdanau等。](#_bookmark32), [2014](#_bookmark32)）和图像字幕（[Vinyals等。](#_bookmark55), [2015](#_bookmark55)).但是，自动编码器框架（[Sutskever等人。](#_bookmark51), [2014](#_bookmark51)）和递归神经网络语言模型（[Mikolov等。](#_bookmark38), [2010](#_bookmark38)）由于有效隐藏代码的不平滑，不适用于任意隐藏表示的通用文本生成（[鲍曼等人。](#_bookmark33), [2015](#_bookmark33)).最近使用VAE的尝试很少（[鲍曼等人。](#_bookmark33), [2015](#_bookmark33); [唐等人。](#_bookmark53), [2016](#_bookmark53)）和GAN（[Yu等人。](#_bookmark44), [2017](#_bookmark44); [张等人。](#_bookmark46), [2016](#_bookmark46)已经被用来调查通用文本生成，而它们生成的文本在很大程度上是随机的和不可控制的。

在本文中，我们解决了文本的受控生成问题。也就是说，我们专注于生成逼真的句子，其属性可以通过学习解缠结的潜在表征来控制。为了能够处理生成的句子，需要解决一些挑战。

第一个挑战来自文本样本的离散性质。由此产生的不可微分性阻碍了全局鉴别器的使用，该鉴别器评估生成的样本和反向传播梯度以整体方式指导发生器的优化，如图所示在连续图像生成和表示建模中非常有效（[陈等人。](#_bookmark35), [2016](#_bookmark35); [拉森等人。](#_bookmark37), [2016](#_bookmark37); [多索维茨基](#_bookmark24) [是BrOx](#_bookmark24), [2016](#_bookmark24)).最近的一些方法试图通过政策学习解决不可区分问题（[宇](#_bookmark44) [等。](#_bookmark44), [2017](#_bookmark44)）在训练期间往往遭受高度变化或连续近似（[张等人。](#_bookmark46), [2016](#_bookmark46); [Kusner&Herndez-Lobato](#_bookmark36), [2016](#_bookmark36)）仅提供初步定性结果。作为基于鉴别器的学习的替代方案，半监督的VAE（[金马等人。](#_bookmark34), [2014](#_bookmark34)）最小化观察到的示例的元素重建误差，并适用于离散可见光。然而，这会丢失完整句子的整体视图，并且尤其对于建模全局抽象属性（例如，情感）而言可能较差。

可控生成的另一个挑战涉及学习解缠结的潜在表示。可解释性要求潜在表示的每个部分都要管理，并且只关注样本的一个方面。现有方法（[陈等人。](#_bookmark35), [2016](#_bookmark35); [Odena等。](#_bookmark40), [2016](#_bookmark40)）关于结构化代表性学习缺乏明确的实施

完全潜在表示的依赖属性，以及变化的单个代码可能导致除了期望的属性之外的其他未指定属性的意外变化。

在本文中，我们提出了一个新的文本生成模型，它解决了上述问题，允许使用指定的语义结构进行高度解密的表示，并生成具有动态指定属性的句子。我们将我们的生成器基于VAE与整体属性鉴别器相结合，以便在潜在代码上有效地施加结构。通过可微分的softmax近似实现端到端优化，可以平滑退火到离散情况，并有助于快速收敛。VAE的概率编码器还用作附加鉴别器以捕获隐式建模方面的变化，并引导生成器以避免在属性代码操纵期间的纠缠。

我们的模型可以解释为通过扩展的唤醒睡眠程序增强VAE（[Hinton等人。](#_bookmark26), [1995](#_bookmark26)），其中睡眠阶段能够结合所生成的样本以交替方式学习发生器和鉴别器。发生器和鉴别器有效地相互提供反馈信号，从而产生有效的相互引导框架。我们展示一些监督（例如，100个带注释的句子）足以学习结构化表示。

除了有效的表示学习和启用半监督训练之外，与传统的基于条件重建的方法相比，使用鉴别器作为发生器的学习信号的另一个优点（[温等人。](#_bookmark37), [2015](#_bookmark37); [金马等人。](#_bookmark34), [2014](#_bookmark34)），是否可以独立训练不同属性的鉴别器。也就是说，对于每个属性，可以使用单独的标记数据来训练相应的鉴别器，并且可以任意组合训练的鉴别器以控制一组感兴趣的属性。相比之下，基于重建的方法通常要求训练数据的每个实例都用所有目标属性进行详尽的标记（[温等人。](#_bookmark37), [2015](#_bookmark37)），或边缘化任何缺失的属性（[金马等人。](#_bookmark34), [2014](#_bookmark34)）这可能是计算上昂贵的。

作为一个展示案例，我们应用我们的模型来生成具有受控情绪和时态的句子。虽然据我们所知，没有带有情感和时态标签的文本语料库，但我们的方法可以使用单独的数据集，一个具有注释情绪，另一个具有时态标签。定量实验证明了我们的方法的功效。我们的模型比以前的生成模型改进了生成指定属性的准确性以及使用生成的样本进行分类。我们展示了我们的方法只从单词级标签中学习高度解密的表示，并产生合理的短句。

# 相关工作

在深度生成建模方面取得了显着进展。[胡等人。](#_bookmark29) ([2017](#_bookmark29)）提供各种深层生成方法的统一视图。变分自动编码器（VAE）（[金马和威灵](#_bookmark31), [2013](#_bookmark31)）包括编码器和发生器网络，它们将数据示例编码为潜在表示并分别从潜在空间生成样本。通过在生成模型下最大化数据对数似然的变分下界来训练模型。KL散度损失被最小化以使潜在代码的后验与先验匹配，这使得从先前解码的每个潜在代码成为合理的句子。如果没有KL正则化，VAE会退化为自动编码器，并且不适用于通用编码器。香草VAE与离散的潜伏性不相容，因为它们阻碍了用于学习编码器的可区分参数化。唤醒睡眠算法（[Hinton等人。](#_bookmark26), [1995](#_bookmark26)）通过将推理网络与发生器相结合，引入用于学习深度定向图形模型与VAE共享相似性。唤醒阶段利用来自推理网络的训练数据生成的样本更新发生器，而睡眠阶段基于来自发生器的样本更新推断网络。我们的方法将VAE与扩展的唤醒睡眠相结合，其中睡眠过程更新发生器和推理网络（鉴别器），实现协作半监督学习。

除了在原始数据空间中的重建之外，基于鉴别器的度量为发生器学习提供了不同的方式，即鉴别器评估所生成的样本并反馈学习信号。例如，GAN（[好-](#_bookmark23) [同胞等人。](#_bookmark23), [2014](#_bookmark23)）使用鉴别器来反馈样本被识别为真实例子的概率。[拉森等人。](#_bookmark37)([2016](#_bookmark37)）将VAE与GAN相结合，以增强图像生成。[多索维茨基和布洛克](#_bookmark24)([2016](#_bookmark24)); [Taigman等。](#_bookmark52)([2017](#_bookmark52)）使用鉴别器来测量高水平的感知相似性。由于离散样本的不可微分性，将鉴别器应用于文本生成很困难（[Yu等人。](#_bookmark44), [2017](#_bookmark44); [张等人。](#_bookmark46), [2016](#_bookmark46); [库斯纳](#_bookmark36) [赫内德洛巴托](#_bookmark36), [2016](#_bookmark36)). [鲍曼等人。](#_bookmark33)([2015](#_bookmark33)); [唐](#_bookmark53) [等。](#_bookmark53)([2016](#_bookmark53)); [杨等人。](#_bookmark41)([2017](#_bookmark41)）而是使用没有鉴别器的VAE。所有这些文本生成方法都没有学习解缠结的潜在表示，导致随机化和不可控制的样本。相比之下，视觉领域的解缠绕生成取得了令人瞩目的进展。例如，InfoGAN（[陈等人。](#_bookmark35), [2016](#_bookmark35)），类似于我们的联合VAE /唤醒 - 睡眠算法的扩展睡眠过程，以无监督的方式解开潜在表示。在训练之后观察每个维度的语义，而不是由用户以受控方式指定。[Siddharth等人。](#_bookmark47)([2017](#_bookmark47)); [金马等人。](#_bookmark34) ([2014](#_bookmark34)）基于VAE并利用半监督学习获得解缠结的图像表示。[周＆Neu-](#_bookmark49)

[大](#_bookmark49)([2017](#_bookmark49)）扩展半监督的VAE用于文本转导。相比之下，我们的模型将VAE与鉴别器相结合，与元素重建相比，它提供了更好的整体度量。此外，这些方法中的大多数仅关注于潜在表示的结构化部分的解开，而忽略了结构化代码与未明确编码的属性的潜在依赖性。我们通过引入独立约束来解决这个问题，并展示其提高可解释性的有效性。

# 受控制的文本生成

我们的模型旨在生成以表示向量为条件的似是而非的句子，这些向量被赋予指定的语义结构。例如，为了控制句子情绪，我们的模型分配潜在表示的一个维度来编码“正”和“负”语义，并通过简单地指定特定代码来生成具有期望情绪的样本。受益于解开的结构，每个这样的代码能够捕获显着属性并且与其他特征独立。我们的深层文本生成模型与以前的工作相比具有几个优点，因为它1）通过使全局鉴别器能够指导离散文本生成器学习，促进潜在代码语义的有效强加;

2）通过明确地对独立属性控件实施约束来提高模型可解释性;3）通过使用定制的唤醒 - 睡眠方法合成变分自动编码器，允许有效的半监督学习和自举。我们首先介绍我们的框架概述（§[3.1](#_bookmark2)），然后详细描述模型（§[3.2](#_bookmark3)).

## 型号概述

我们从变分自动编码器开始构建我们的框架（[2](#_bookmark0)）已用于文本生成（[鲍曼等人。](#_bookmark33), [2015](#_bookmark33)），其中句子x是以潜码z为条件生成的。香草VAE采用非结构化矢量z，其中尺寸被缠结。为了以可解释的方式建模和控制感兴趣的属性，我们用一组结构化变量c来扩充非结构化变量z，每个结构化变量c都针对句子的显着且独立的语义特征。

*§*

我们希望我们的句子生成器对组合向量（z，c）进行条件化，并生成满足结构化代码c中指定的属性的样本。VAE背景下的条件生成（例如，半监督的VAE（[金马等人。](#_bookmark34), [2014](#_bookmark34)））通常通过在给定其特征代码的情况下重建观察到的示例来学习。然而，如在视觉领域中所证明的，与计算数据空间中的元素距离相比，在特征空间中计算距离允许不依赖于分散变换的不变性并且提供更好的整体度量。



𝑥$

𝑐

𝑧

鉴

𝑥

发电机

编码器

*图*1.生成模型，其中z是非结构化潜在代码，c是结构化代码，用于控制句子属性。蓝色虚线箭头表示建议的独立性约束（部分[3.2](#_bookmark3) 对于细节），红色箭头表示由可微分近似实现的梯度传播。

因此，对于c中的每个属性代码，我们设置一个单独的鉴别器来测量生成的样本与所需属性的匹配程度，并驱动生成器以产生改进的结果。在我们的上下文中应用鉴别器的困难在于文本样本是离散的且不可微分的，这破坏了从鉴别器到生成器的梯度传播。我们使用基于softmax的连续近似，其具有降低的温度，随着训练的进行退火到离散的情况。这种简单而有效的方法具有低方差和快速收敛。

直观地说，具有可解释的表示意味着c中的每个结构化代码可以独立地控制其目标特征，而不会与其他属性纠缠，尤其是那些未明确建模的属性。我们鼓励通过强制执行那些不相关的属性来完全捕获非结构化代码z中的独立性，从而将其与我们将要操作的c分开。为此，我们重用VAE编码器作为识别z中建模属性的附加鉴别器，并训练生成器，以便可以从生成的样本中恢复这些非结构化属性。因此，只要z不变，变化的不同属性代码将使非结构化属性保持不变。

数字[1](#_bookmark1) 显示整体模型结构。我们的完整模型包含VAE和属性鉴别器，其中VAE组件训练生成器重建真实句子以生成合理的文本，而鉴别器强制生成器生成与条件代码一致的属性。学习属性鉴别器以适合标记的示例以引起指定的语义，并且训练以解释来自生成器的样本。也就是说，发生器和鉴别器形成一对协作学习器并向彼此提供反馈信号。协作优化类似于唤醒睡眠算法。我们展示了组合的VAE /唤醒 - 睡眠学习能够实现高效的半监督框架，这需要一点点监督才能获得可解释的表示和生成。

## 模型结构

我们现在分别通过介绍生成器和鉴别器的学习来详细描述我们的模型。

## 发电机学习

生成器G是用于生成令牌序列x = x1的LSTM-RNN。..，x*t* 以潜在代码（z，c）为条件，描述了生成分布：

*{ }*

***x~G（z，c）= pG（x | z，c）***

= TT *p*(*x*ˆ *|****x***ˆ*<t,* ***z****,* ***c***)*,* (1)

*t*

*t*

其中x*<t* 表示x之前的标记。一般 -

目前描述的目标。除了驱动发生器产生逼真句子的重建损失之外，鉴别器还提供额外的学习信号，强制生成器产生与c中的结构化代码匹配的相干属性。然而，由于不可能通过离散样本从鉴别器传播梯度，我们求助于确定性连续近似。近似在每个步骤用等式中的概率向量替换采样的令牌x*t* （表示为单热矢量）（[2](#_bookmark5)）这与发电机的参数是可微分的。概率向量用作当前步骤的输出和沿决策序列的下一步骤的输入。该

ˆ *t* 得到的“软”生成句子，表示为Gτ（z，c），是

因此，涉及一系列离散决策，其在每个时间步骤t使用softmax函数从参数化的多项分布中采样令牌：

*x*t~softmax（ot /τ），(2)

其中o*t* 是logit向量作为softmax函数的输入，τ> 0是通常设置为1的温度。

表示的非结构化部分z被建模为具有标准高斯先验p（z）的连续变量，而结构化代码c可以包含连续和离散变量以编码具有适当的先验p的不同属性（例如，情感类别，形式）（ C）。给定观察x，基础VAE包括条件概率编码器E以推断潜在z：

***z***~E（x）= qE（z | x）。(3)

令θ*g* 和θ*e* 分别表示发生器G和编码器E的参数。然后对VAE进行优化以最小化观察到的真实句子的重建误差，同时使编码器规则化为接近先前的p（z）：

进入鉴别器[1](#_bookmark9) 测量目标属性的适应度，导致以下损失以改善G：

*L*页：1（θ*g*）= - E*p(z)p(c)* log q*d*（c | G*τ* （z，c））l。(6)

温度τ（Eq。[2](#_bookmark5)随着训练的进行，设置为τ0，产生越来越高峰的分布，最终模拟离散的情况。简单的确定性近似有效地导致训练期间的方差减小和快速收敛，这使得能够有效地学习条件生成器。生成结果的多样性得到保证，因为我们仅使用近似来进行属性建模，并且通过VAE学习基本句子生成。

*→*

目标是Eq。（[6](#_bookmark6)），生成句子的每个结构化属性通过c中的相应代码控制，并且与潜在表示中的其他变量无关。但是，未明确建模的其他属性仍然可能与c中的代码纠缠在一起，因此改变c的维度会产生我们不感兴趣的这些属性的意外变化。为了解决这个问题，我们引入了独立性约束。通过强制它们被非结构化部分z完全捕获，将这些属性与c分开。因此，

*L*VAE（9522）；G，9522；；E；X）=KL（QE（Z-124；X）P（z））

*-* E*qE时*（z | x）q*d*（c | x）[log p*g*（x | z，c）]，

(4)

除了在c中明确编码的属性之外，我们还训练生成器，以便可以从生成的样本中正确识别其他非显式属性并匹配

KL（哪里） ）是KL-分歧;和q*d*（cx）是

*· · |*

由鉴别器D为c中的每个结构化变量定义的条件分布：

*D*（x）= qD（c | x）。(5)

这里，为了简化符号，我们假设只有一个结构化变量，因此我们假设一个鉴别器，尽管我们的模型规范可以直接应用于许多属性。当我们学习解开的表示时，（z，c）因子上的因子分布为q*e* 和q*d* 。注意，这里鉴别器D和代码c不是在VAE丢失的情况下学习的，而是用

非结构化代码z。我们不使用新的鉴别器，而是重复使用变分编码器E，它精确地用于推断基础VAE中的潜在z。损失与方程式相同（[6](#_bookmark6)除了用编码器条件q*e*替换鉴别器条件q*d* ：

*L*属性，Z（θ*g*）= - E*p(z)p(c)* log q*e*（z | G*τ* （z，c））l。(7)

注意，作为方程中的鉴别器（[6](#_bookmark6)），编码器现在

1因此，概率向量用于对字嵌入矩阵求平均，以在每一步获得“软”字嵌入。

与在VAE中观察到的示例相反，对先前生成的样本进行推断。

结合Eqs。（[4](#_bookmark7))-([7](#_bookmark8)）我们获得了发电机目标：

**算**法1受控生成文本

**输**入：大量未标记的句子 = x几个句子属性标签L =（xL，cL）参数：λc，λz，λu，β - 平衡参数

*X { }*

*X { }*

*X*

Min§952；g lg=lve＋Clattr，c＋lz lattr，z，(8)

1：通过最小化方程初始化基础VAE（[4](#_bookmark7)） 上 与c

从先前的p（c）中抽样

2**：重复一遍**

其中λ*c* 和λ*z* 是平衡参数。通过最小化VAE损失来训练变分编码器，即min***952；e***L瓦斯。

3: 通过方程式训练鉴别器D（[11](#_bookmark14))

4: 通过方程式训练发电机G和编码器E.（[8](#_bookmark11)）和最小化方程（[4](#_bookmark7)）， 分别。

5：直到收敛

## 鉴别者学习

鉴别器D经过训练以准确地推断出

**输**出：句子生成器G以解开的表示为条件（z，c）

tence属性并评估恢复潜在代码中指定的所需特征的错误。例如，对于分类属性，可以将鉴别器表示为句子分类器;而对于连续目标，可以使用概率回归量。与VAE编码器相比，以不同的方式学习鉴别器，

𝑧 𝑐

𝑞2(𝒛|𝒙) 𝑞3(𝒄|𝒙)

𝑞-(𝒙|𝒛, 𝒄)

𝑝(𝒛) 𝑝(𝒄)

𝑧 𝑐



𝑞-(𝒙|𝒛, 𝒄)

𝑞3(𝒄|𝒙)

因为目标属性可以是离散的，而不是支持

𝑥 ∼ 𝜒

𝑥# ∼ 𝐺(𝑧, 𝑐)

移植到VAE框架中。此外，与以无监督方式学习的非结构化代码z相比，结构化变量c使用标记的示例来引入指定的语义。我们为鉴别器推导出一种有效的半监督学习方法。

正式地，让θ*d* 表示鉴别器的参数。为了学习指定的语义，我们使用一组标记的例子 *l* =（x*l*，c*l*）来训练鉴别器D，其目标如下：

*X { }*

*L*s（θD）= - EXL [log qD（cL | xL）]。(9)

此外，条件生成器G还能够合成（有噪声的）句子 - 属性对（x，c），其可用于增强用于半监督学习的训练数据。为了缓解噪声数据的问题并确保模型优化的稳健性，我们引入了最小熵正则化项（[Grandvalet等。](#_bookmark25), [2004](#_bookmark25); [芦苇](#_bookmark45) [等。](#_bookmark45), [2014](#_bookmark45)).由此产生的目标是：

*Lu*（θ*d*）= - E*的pG*（x | z，c）p（z）p（c）log q*d*（c | x）+βH（q*d*（c | X）），

(10)

其中（q*d*（c*/* x））是在生成的句子x上评估的分布q*d* 的经验Shannon熵;β是平衡参数。直观地，最小熵正则化促使模型在预测标签方面具有高可信度。

*H |*

然后，使用标记示例和合成样本的鉴别器的联合培训目标如下：

Min***952人d***L*d* =L*s* +Impact*u*L*u*，(11)

其中λ*u* 是平衡参数。

*图*2.左：VAE和唤醒程序，对应于Eq。（[4](#_bookmark7)).右：睡眠程序，对应于Eqs。（[6](#_bookmark6))-

([7](#_bookmark8)）和（[10](#_bookmark13)).黑色箭头表示推理和生成;红色虚线箭头表示梯度传播。睡眠过程中的两个步骤，即分别优化鉴别器和发生器，以交替方式执行。

## 总结与讨论

我们已经推导出我们的模型及其学习过程。首先通过在大量无标签句子上训练基础VAE来初始化生成器，目的是最小化方程（[4](#_bookmark7)此时使用潜在代码c从先前分布p（c）中采样。然后通过交替优化发生器和鉴别器来训练完整模型，如算法中所总结的[1](#_bookmark10).

我们的模型可视为将VAE框架与扩展的唤醒 - 睡眠方法相结合，如图所示[2](#_bookmark12).具体来说，在Eq。（[10](#_bookmark13)），样本由发生器产生，并用作鉴别器最大可能性训练的目标。这类似于唤醒 - 睡眠的睡眠阶段。方程（[6](#_bookmark6))-([7](#_bookmark8)）进一步利用生成的样本来改进发电机。我们可以将上述内容一起看作基于通过来自生成网络的祖先采样获得的“梦”样本的扩展睡眠过程。另一方面，Eq。（[4](#_bookmark7)）来自观测x上的鉴别器分布q*d*（cx）的样本c，以形成训练发生器的目标，其对应于尾流阶段。有效组合可实现离散潜码，整体鉴别器指标和高效的相互引导。

*|*

鉴别器的训练需要监督数据来强加指定的语义。可以在单独的标记集上独立地训练用于不同属性的鉴别器。也就是说，该模型不需要句子

使用所有属性进行注释，但每个属性只需要独立的标记数据。此外，由于标记数据仅用于学习属性语义而不是直接句子生成，因此允许我们将数据范围扩展到标记的句子之外，例如标记的单词或短语。如实验所示（部分[4](#_bookmark16)），我们的方法能够有效地将单词级别的知识提升到句子级别并生成令人信服的句子。最后，通过睡眠短语中增强的无监督训练，我们展示了一点监督就足以学习结构化表示。

# 实验

我们应用我们的模型来生成短句（长度

*≤*

15）具有受控的情绪和紧张。使用经过训练的分类器作为评估器的定量实验表明，我们的模型提高了生成精度。通过一些标签或仅使用单词注释来学习解开的表示。我们还验证了所提出的独立约束对可解释生成的影响。

## 数据集

**句**子语料库。我们使用大型IMDB文本语料库（[刁](#_bookmark22) [等。](#_bookmark22), [2014](#_bookmark22)）用于训练生成模型。这是350K电影评论的集合。我们选择包含最多15个字的句子，并更换与令牌“<UNK>”罕见的话。得到的数据集包含大约1.4M的句子，词汇量大小为16K。

**情**绪。为了控制生成句子的情绪（“正面”或“负面”），我们测试以下标记的情绪数据：（1）斯坦福情绪树库-2（SST-full）（[Socher等。](#_bookmark50), [2013](#_bookmark50)）由火车/开发/测试集中的二进制情绪注释分别包含6920/872/1821电影评论语句。我们使用句子长度为15的2837训练样例，并评估原始测试集的分类准确性。（2）SST-小。为了研究半监督学习中准确属性控制所需的标记数据的大小，我们从SST-full中抽取一小部分，仅包含250个标记的句子用于训练。（3）词典。我们还研究了我们的模型在使用单词级标签进行句子级控制方面的有效性。来自的词典（[威尔逊等人。](#_bookmark39), [2005](#_bookmark39)）包含带有情绪标签的2700个单词。我们通过将单词视为句子来使用词典进行训练，并在SST-full测试集上进行评估。

*≤*

1. **I**MDB。我们从IMDB语料库中收集数据集

通过随机选择正面和负面的电影评论。数据集在train / dev / test中有5K / 1K / 10K句子。

**紧**张。第二个属性是句子中主要动词的时态。虽然没有容易获得句子紧张注释的语料库，但我们的方法只能从标记的单词中学习并生成所需的句子。

Model 数据集

|  |  |  |  |
| --- | --- | --- | --- |
|  | SST-全 | SST-小 | 词汇 |
| S-VAE | 0.822 | 0.679 | 0.660 |
| 我们的 | **0.851** | **0.707** | **0.701** |

*表*1.生成句子的情绪准确性。S-vae[金马等人。](#_bookmark34), [2014](#_bookmark34)并且我们的模型在三个情感数据集上训练并分别生成30K句子。

我们从TimeBank（timeml.org）数据集中编译并获得5250个单词和短语的词典，标记为“过去”，“现在”，“未来”之一。词典主要由不同时态的动词（例如，“是”，“将是”）以及时间表达（例如，“将来”）组成。

*{ }*

请注意，我们的方法只需要为每个属性单独标记copora。并且对于时态属性，仅使用带注释的单词/短语。

## 参数设定

生成器和编码器设置为单层LSTM RNN，输入/隐藏维度为300，最大样本长度为15.鉴别器设置为ConvNets。详细配置在补充中。为了避免在VAE模块中消失小的KL术语（Eq。[4](#_bookmark7)) ([鲍曼等人。](#_bookmark33), [2015](#_bookmark33)），我们使用KL项权重在训练期间从0到1线性退火。平衡参数设置为λ*c* =λ*z* =λ*u* = 0.1，并且在开发集上选择β。在测试时，用方程生成句子（[1](#_bookmark4)).

## 生成属性的准确性

我们通过评估生成指定情绪的准确性以及使用样本训练分类器的效果来定量测量句子属性控制。我们与半监督的VAE（S-VAE）进行比较（[金玛](#_bookmark34) [等。](#_bookmark34), [2014](#_bookmark34)），为数不多的能够生成条件文本的深度模型之一。S-VAE学习在给定属性代码的情况下重建观察到的句子，并且不使用鉴别​​器。见§[2](#_bookmark0) 和[3.1](#_bookmark2) 进行更多讨论。

我们使用最先进的情绪分类器（[胡等人。](#_bookmark27), [2016a](#_bookmark27)）它在SST测试装置上达到90％的准确度，自动评估情绪产生的准确性。具体来说，我们生成给定情感代码c的句子，并使用预先训练的情感分类器为生成的句子分配情感标签。精度计算为与情感代码c匹配的预测百分比。表[1](#_bookmark15) 用SST-full，SST-small和Lexicon分别训练的两个模型显示了30K句子的结果。我们看到我们的方法在所有数据集上始终优于S-VAE。特别是，在SST-small中仅使用250个标记示例进行了训练，我们的模型实现了合理的生成准确性，证明了学习解缠的代表的能力

0.85

0.80

0.75

准确性

0.70

0.65

0.60



标准

H-注册

我们的

S-VAE

sst-完整 sst-小词汇 英姆德布

表[2](#_bookmark18) 比较分别具有和不具有约束项的模型生成的样本。在约束适用的左栏中，以不同情感代码为条件的每对句子在例如主题，语气和措辞方面高度相关，其在结构化代码c中未明确建模，而是隐含地编码在非结构化代码z。改变情绪代码可以精确地改变句子的情绪（并略微释义以确保流畅性），

*图*3.使用不同方法扩充的四个情绪数据集训练的分类器的测试集精度（详见文本）。前三个数据集使用SST完整测试集进行评估。

没有监督的发送。更重要的是，在Lexicon中只给出了单词级别的注释，我们的模型成功地将知识转移到句子级别并且合理地产生了期望的情绪。与通过直接评估生成的句子来推动学习的方法相比，S-VAE仅通过重建标记的单词来尝试捕获情感语义，这样的效率较低且性能较差。

接下来，我们使用生成的样本来增加情感数据集并训练情绪分类器。虽然不是为了在这些数据集上构建性能最佳的分类器，但分类准确性可以作为句子生成质量的辅助度量。也就是说，具有更准确的情感属性的更高质量的句子可以预测地帮助产生更强的情感分类器。数字[3](#_bookmark17) 显示了在具有不同增强的四个数据集上训练的分类器的准确性。“Std”是在标准原始数据集上训练的Con-vNet，其网络结构与我们模型中的情感鉴别器相同。“H-reg”还对生成的句子施加最小熵正则化。“我们的”包含了生成句子的最小熵正则化和情感属性代码c，如公式（[10](#_bookmark13)).S-VAE使用与我们的方法相同的协议来增加S-VAE模型生成的数据。比较图[3](#_bookmark17) 表明我们的方法始终在四个数据集上提供最佳性能。例如，在Lexicon上，与“Std”的0.701相比，我们的方法达到了0.733的准确度。“H-Reg”相对于“Std”的改进表明最小熵正则化对生成句子的积极影响。进一步结合所生成样本的条件情感码，如“我们的”和“S-VAE”，提供额外的性能增益，指示条件生成用于自动创建标记数据的优点。与上述实验一致，我们的模型优于S-VAE。

## 解开的表示

我们研究了生成的可解释性和显式的独立性约束（Eq。[7](#_bookmark8)）用于解开控制。

同时保持其他方面不变。相比之下，右列中未独立约束未激活的结果表明，改变情感代码不仅会改变样本的极性，还会改变意外控制的其他方面，使得生成结果更难以解释和预测。

我们通过一次改变一个属性变量来证明学习解开的表示的力量。表[3](#_bookmark19) 显示生成结果。我们看到模型中的每个属性变量都成功地控制了它的相应属性，并与其他属性代码解开。该表的右栏显示了紧张代码变化时句子时态的有意义变化。请注意，时态的语义仅从没有完整句子示例的词典中学习。我们的模型成功地捕获了关键成分（例如，动词“是”用于过去时和“将来”用于将来时）并且与精心构造的句子的知识相结合以生成具有指定时态属性的真实样本。表[4](#_bookmark20) 进一步示出了在结构化属性因子的不同设置中具有变化代码z的生成句子我们获得的内容多样，同时情绪和时态也一致。

我们偶尔也会在表格中观察失败案例[5](#_bookmark21)，例如难以置信的句子，不相关属性的意外变化以及不准确的属性生成。期望改进的建模，例如使用扩张的卷积作为解码器，以及利用波束搜索等进行解码。还需要更好的系统定量评估。

# 讨论

我们提出了一种深度生成模型，它可以学习可解释的潜在表征并生成具有指定属性的句子。我们获得了有限的句子长度的有意义的生成，并提高了情感和时态属性的准确性。在未来，我们希望如上所述改进建模和训练，并扩展以生成更长的句子/段落，并使用细粒度结构控制更多属性。

我们的方法将VAE与属性鉴别器结合起来，并对属性控制强加显式的独立约束，从而实现解密的潜在代码。联合VAE /唤醒 - 睡眠中的半监督学习

## w /独立性约束 没有独立性约束

这部电影是严格的常规！ 表演很糟糕。

这部电影充满了想象力。 这部电影非常有趣。

看完这部电影后，我感到很失望。 这一切都不是很原始。

看完这部电影之后，我就是粉丝。 强烈推荐观看它的勇气和想法。

表演也是一致的。 太平淡了

表演一般都很好。 高度可观

这太可怕了。 我可以在不超过三个单词的情况下分析这部电影。

这是纯粹的天才。 我强烈推荐这部电影给任何欣赏音乐的人。

*表*2.来自具有或不具有属性控制的独立性约束的模型的样本（即，等式1）。[7](#_bookmark8)).在修复非结构化代码z的同时，分别将情感代码设置为“负”和“正”来生成每对句子。SST完整数据集用于学习情感表示。

## 改变时态代码

我觉得这部电影太平淡无奇了 这是过去十年中最出色的惊悚片之一，我觉得这部电影过于平淡无奇 这是有史以来最杰出的惊悚片之一

我猜这部电影太平淡了 这将是有史以来最伟大的惊悚片之一

*表*3.通过在修复情感代码和z时改变时态代码来生成每个三个句子。

## 改变非结构化代码z

*（“消极”，“过去”） （“积极的”，“过去的”）*

表演也有点受伤。 他的演技无可挑剔

我希望自己永远不会见到它 这是壮观的，我在剧院看了两次到最后我是如此迷失，我只是不在乎了 这很有趣

*（“负面”，“现在”） （“正面”，“现在”）*

电影与情节和人物的表演非常接近 这是一部比较好的舞蹈电影

这个时代似乎遥不可及 我一直是智能对话的忠实粉丝。

我认为在电影结束时，它已经混淆了自己 我建议你去看看，特别是如果你受伤了

*（“消极”，“未来”） （“积极”，“未来”）*

我不会看电影 我希望他将来会拍更多电影

那将是毁灭性的！ 我肯定会在DVD上买这个

我不会进入故事，因为真的没有 事后你会想到的，我向你保证

*表*4.通过改变非结构化代码z的样本给定情绪（“正”/“负”）和时态（“过去”/“存在”/“未来”）代码。

## 失败的情况

情节不是那么原始 其他舞蹈电影没有任何好转

情节编织我们进入<UNK>它没有达到它们，但故事看起来

他是一个可怕的演员 我只是这么认为

他是一个比站立更好的演员 我只是想 ！

*表*5.修改了其他代码固定的情绪代码时的失败情况。

框架在很少或不完整的监督下有效。[胡等人。](#_bookmark29)([2017](#_bookmark29)）建立一系列深层生成范式的统一视图，包括GAN，VAE和唤醒睡眠算法。我们的模型可以在视图中被激励，即通过延长睡眠阶段和利用生成的样本来增强VAE。

潜在表示的可解释性不仅允许动态控制生成的属性，还提供将端到端神经模型与传统结构化方法连接的接口。例如，我们可以对可解释的潜在代码编码结构化约束（例如，逻辑规则或概率结构化模型），以结合先验知识或人类意图（[胡等人。](#_bookmark27), [2016a](#_bookmark27);[b](#_bookmark28));或者将解开的生成模型插入对话系统，以从结构化对话状态生成自然语言响应（[Young等人。](#_bookmark43), [2013](#_bookmark43)).

虽然我们专注于模型的生成能力，但是提出的协作半监督学习框架还通过生成用于数据增强的标记样本来帮助改进鉴别器（例如，参见图[3](#_bookmark17)).更一般地，对于任何判别性任务，我们可以构建条件生成模型以合成附加的标记数据。与以前的生成方法相比，我们的方法的准确属性生成可以提供更大的性能提升。

**实**现我们已经发布了所提算法的改编版本的代码：[https://github.com/asyml/texar/tree/master/](https://github.com/asyml/texar/tree/master/examples/text_style_transfer) [实例/ text\_style\_transfer](https://github.com/asyml/texar/tree/master/examples/text_style_transfer).

实施基于Texar（[胡等人。](#_bookmark30), [2018](#_bookmark30)），通用文本生成工具包。

**致**谢本研究得到NSF IIS1447676，ONR N000141410684和ONR N000141712463的支持。

# 参考

Bahdanau，Dzmitry，Cho，Kyunghyun和Bengio，Yoshua。神经机器翻译通过联合学习对齐和翻译。arXiv preprint arXiv：1409.0473,2014。

Bowman，Samuel R，Vilnis，Luke，Vinyals，Oriol，Dai，Andrew M，Jozefowicz，Rafal和Bengio，Samy。从连续空间生成句子。arXiv preprint arXiv：1511.06349,2015。

Chen，Xi，Duan，Yan，Houthooft，Rein，Schulman，John，Sutskever，Ilya和Abbeel，Pieter。InfoGAN：通过信息最大化生成对抗网络的可解释表示学习。在神经信息处理系统的进展，第2172-2180页，

2016.

刁，启明，邱明辉，吴超元，斯莫拉，亚历山大J，江，京，王冲。联合建模电影推荐（JMARS）的方面，评级和情绪。在第20届ACM SIGKDD知识发现和数据挖掘国际会议论文集，第193-202页。ACM，2014年。

Dosovitskiy，Alexey和Brox，Thomas。基于深度网络生成具有感知相似性度量的图像。arXiv preprint arXiv：1602.02644,2016。

Goodfellow，Ian，Pouget-Abadie，Jean，Mirza，Mehdi，Xu，Bing，Warde-Farley，David，Ozair，Sherjil，Courville，Aaron和Bengio，Yoshua。生成对抗网。在神经信息处理系统的进展，第2672-2680页，2014年。

Grandvalet，Yves，Bengio，Yoshua，et al。熵最小化的半监督学习。在NIPS，第17卷，第529-536页，2004年。

Hinton，Geoffrey E，Dayan，Peter，Frey，Brendan J和Neal，Radford M.无监督神经网络的“唤醒 - 睡眠”算法。Science，268（5214）：1158,1995。

Hu，Zhiting，Ma，Xuezhe，Liu，Zhengzhong，Hovy，Eduard，Xing，Eric。利用逻辑规则利用深度神经网络。在ACL中，2016a。

Hu，Zhiting，Yang，Zichao，Salakhutdinov，Ruslan和Xing，Eric P.具有大量学识的深层神经网络。在EMNLP，2016b。

Hu，Zhiting，Yang，Zichao，Salakhutdinov，Ruslan和Xing，Eric P.关于统一深层生成模型。arXiv preprint arXiv：1706.00550,2017。

胡志智，石浩然，杨子超，谭博文，赵天成，何俊贤，王文涛，俞，新江，秦连辉，王迪，马，学哲，刘，赫克托，梁， Xiaodan，Zhu，Wanrong，Sachan，Devendra Singh和Xing，Eric。Texar：用于文本生成的模块化，通用且可扩展的工具包。2018。

Kingma，Diederik P和Welling，Max。自动编码变分贝叶斯。arXiv preprint arXiv：1312.6114,2013。

Kingma，Diederik P，Mohamed，Shakir，Rezende，Danilo Jimenez和Welling，Max。具有深层生成模型的半监督学习。在神经信息处理系统的进展，第3581-3589页，

2014.

Kusner，Matt和Hernndez-Lobato，Jos.GANs用于具有Gumbel-softmax分布的离散元素序列。arXiv preprint arXiv：1611.04051,2016。

Larsen，Anders Boesen Lindbo，Sønderby，SørenKaae和Winther，Ole。使用学习的相似性度量自动编码超出像素。在ICML，2016年。

Mikolov，Tomas，Karafia't，Martin，Burget，Lukas，Cer-nocky`，Jan和Khudanpur，Sanjeev。基于递归神经网络的语言模型。在Interspeech，第2卷，第3页，2010年。

Odena，Augustus，Olah，Christopher和Shlens，Jonathon。使用辅助分类器GAN进行条件图像合成。arXiv preprint arXiv：1610.09585,2016。

Radford，Alec，Metz，Luke和Chintala，Soumith。使用深度卷积生成对抗网络的无监督表示学习。arXiv preprint arXiv：1511.06434,2015。

Reed，Scott，Lee，Honglak，Anguelov，Dragomir，Szegedy，Christian，Erhan，Dumitru和Rabinovich，Andrew。用自举训练在嘈杂标签上训练深度神经网络。arXiv preprint arXiv：1412.6596,2014。

Siddharth，N.，Paige，Brooks，Desmaison，Alban，Meent，Jan-Willem van de，Wood，Frank，Goodman，Noah D.，Kohli，Pushmeet和Torr，Philip HS学习深层生成模型中的解构表达。2017年。

Socher，Richard，Perelygin，Alex，Wu，Jean Y，Chuang，Jason，Manning，Christopher D，Ng，Andrew Y，Potts，Christopher，et al。针对情感树库的语义组合的递归深度模型。在自然语言处理经验方法会议论文集（EMNLP），第1631卷，第1642页.Citeseer，2013年。

Sutskever，Ilya，Vinyals，Oriol和Le，Quoc V.使用神经网络对序列学习进行排序。在神经信息处理系统的进展中，第3104-3112页，2014年。

Lig，Taigman，Yaniv，Polyak，Adam和Wolf，Lior。无监督的跨域图像生成。在ICLR，2017年。

Tang，Shuai，Jin，Hailin，Fang，Chen和Wang，Zhaowen。使用对抗性自动编码器学习无监督的句子表示。2016年

van den Oord，Aaron，Kalchbrenner，Nal和Kavuk的Kavukcuoglu。像素递归神经网络。在ICML，2016年。

Vinyals，Oriol，Toshev，Alexander，Bengio，Samy和Erhan，Dumitru。显示和告诉：神经图像标题生成器。“IEEE计算机视觉与模式识别会议论文集”，第3156-3164页，2015年。

Wen，Tsung-Hsien，Gasic，Milica，Mrksic，Nikola，Su，Pei-Hao，Vandyke，David和Young，Steve。基于语义的基于lstm的自然语言生成用于语音对话系统。在EMNLP，2015年。

Wilson，Theresa，Wiebe，Janyce和Hoffmann，Paul。识别短语级情绪分析中的语境极性。在人类语言技术会议论文集和自然语言处理中的经验方法，第347-354页。计算语言学协会，2005年。

Yang，Zichao，Hu，Zhiting，Salakhutdinov，Ruslan和Berg-Kirkpatrick，Taylor。改进的变分自动编码器用于使用扩张卷积的文本建模。在ICML，2017年。

Young，Steve，Gasic'，Milica，Thomson，Blaise和Williams，Jason D. POMDP统计口语对话系统：综述。IEEE，101（5）：1160-1179,2013的会议记录。

Yu，Lantao，Zhang，Weinan，Wang，Jun和Yu，Yong。Se-qGAN：具有策略梯度的序列生成对抗网。在AAAI，2017年。

张一哲，甘，哲和卡林，劳伦斯。通过对抗训练生成文本。在NIPS对抗训练研讨会上，2016年。

周，春婷和Neubig，格雷厄姆。用于半监督标记序列转导的多空间变分编码器 - 解码器。在ACL中，2017年。

Zhu，Jun-Yan，Kra¨henbu¨hl，Philipp，Shechtman，Eli和Efros，Alexei A.对自然图像流形的生成性视觉操纵。在欧洲计算机视觉会议上，第597-613页。施普林格，2016年。