用于稳定变分自动编码器的球形潜在空间

徐嘉诚和Greg Durrett德克萨斯大学奥斯汀分校计算机科学系

{jcxu, gdurrett} @ cs.utexas.edu

摘要

用于文本处理的变分自动编码器（VAE）的标志是它们强大的编码器 - 解码器模型（例如LSTM）与简单的潜在分布（通常是多变量高斯分布）的组合。 这些模型提出了一个困难的优化问题：局部最优，其中变分后验总是等于先验，模型根本不使用潜在变量，这是一种“崩溃”，受到KL分歧项的鼓励。目标。 在这项工作中，我们尝试了另一种潜在分布选择，即von Mises-Fisher（vMF）分布，它将质量放置在单位超球面上。 通过这种先验和后验的选择，KL分歧项现在仅取决于vMF分布的方差，使我们能够将其视为固定的超参数。 我们表明，这样做不仅可以避免KL崩溃，而且在一系列建模条件下，包括循环语言建模和词袋文档建模，始终比高斯人提供更好的可能性。 对我们的vMF表示的属性的分析表明，他们在他们的潜在表征中比他们的高斯对应物学习更丰富和更细微的结构。[1]

1 介绍

最近的工作已经确定了NLP中一系列任务的深度生成模型的有效性，包括文本生成（胡等人。，2017; Yu等人。，2017），机器翻译（张等人。，2016）和风格转移（沉等人。，2017; 赵 等。，2017a）。 变分自动编码器，已在过去的文本建模工作中探索过（苗等人，2016; 鲍曼等人。，2016），设置一个连续的潜在变量，用于捕获数据中的潜在结构。 典型的VAE实现假设该潜在空间的先验是多元高斯；在训练期间，损失函数中的Kullback-Leibler（KL）发散项鼓励变分后验近似于先验。 过去工作中观察到的这种方法的一个主要限制是KL术语可能会促使潜在变量的后验分布“崩溃”到先前，有效地使潜在结构未被使用（弓- 男人等。，2016; 陈等人。，2016）.

在本文中，我们建议使用von Mises-Fisher（vMF）分布而不是Gaussian作为我们的潜在变量。 vMF在由平均参数 μ 和浓度参数 κ 控制的单位超球面上放置分布。我们的先验是单位超球面上的均匀分布（$\kappa = 0$），我们的后验分布族将 κ 视为固定模型超参数。 由于KL分歧仅取决于 κ，我们可以在结构上防止KL崩溃并使我们的模型的优化问题更容易。 我们证明这种方法实际上比试图灵活地学习 κ 更有效，并且固定 κ 的各种设置导致良好的性能。 我们的模型系统地实现了比模拟高斯模型更好的对数似然，同时具有更高的KL散度值，表明它在训练结束时更成功地利用了潜在变量。

过去的工作已经提出了其他几种处理高斯情况下KL崩溃的技术。 退出KL术语的重量（鲍曼等人。，2016）正如我们在Section中所示，在优化过程中仍然让我们处于脆弱状态 2。 其他先前的工作（杨等人。，2017; Semeniuta等。，2017）专注于使用CNN而不是RNN作为解码器，以削弱模型并鼓励使用

[1]代码和数据集可从以下网址获得： https: //开头github.com /阿联酋/ VMF _徐嘉诚_ NLP

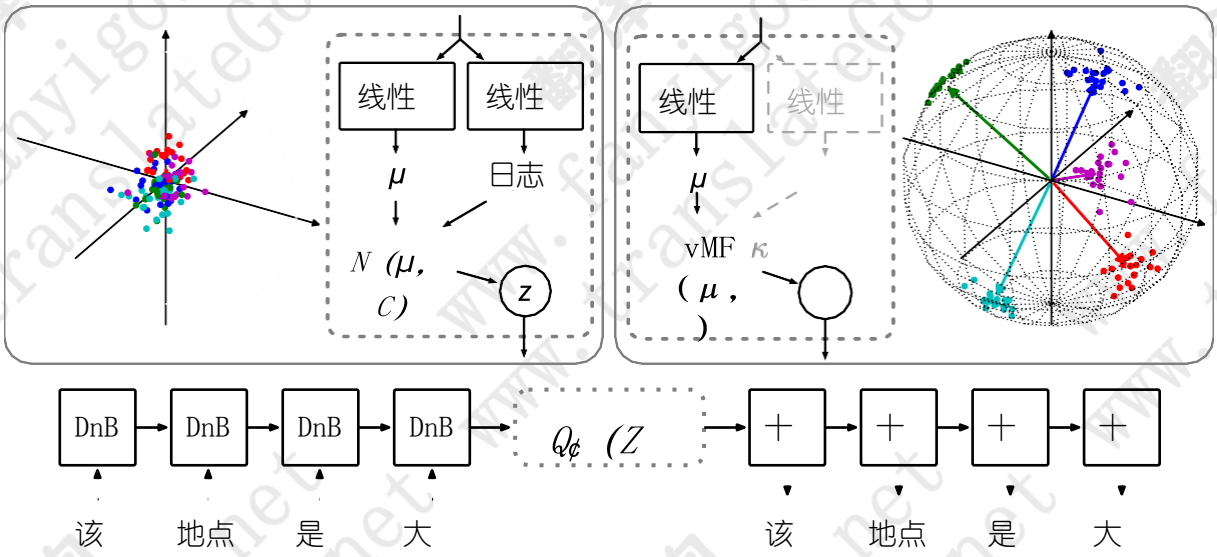


图1：基于高斯先验（左）和vMF先验（右）的神经变分RNN（NVRNN）语言模型。 编码器模型首先计算变分近似q_ϕ（zx）的参数（见虚线框）； 然后我们对z进行采样并生成给定z的单词x。 我们显示来自（0，I）和vMF（，$\kappa = 100$）的样本；后面的样品位于单位球体的表面上。 虽然可以从编码器网络预测κ，但我们通过实验发现，修复它可以实现更稳定的优化和更好的性能。

潜码，但增益有限，以这种方式改变解码器需要特殊的模型工程和仔细调整各种解码器容量参数。 我们的方法与解码器的选择正交，并且可以与这些方法中的任何一种组合。 在VAE中使用vMF分布也使我们能够灵活地以其他方式修改先验，例如使用统一的产品分布（Guu等人。，2018）或分段常数项（Serban等人。，2017a）.

我们在两种生成建模范例中评估我们的方法。对于RNN语言建模和词袋文档建模，我们发现vMF比高斯先验更健壮，并且我们的模型学习更多地依赖于潜在变量，同时实现更好的保持数据可能性。 为了更好地理解这些模型之间的对比，我们设计并进行了一系列实验来理解高斯和vMF潜码空间的特性，这些特征产生了不同的结构假设。 不出所料，这些潜在的代码分发捕获了大量相同的信息，但是我们表明vMF可以更容易地超越这一点，比高斯代码更有效地捕获排序信息。

2 文本的变分自动编码器

鲍曼等人。 (2016提出一个变化的自我启发式生成文本建模的编码器模型 金马和威灵 (2013）。 VAE不是像在香草语言模型中那样直接对p（x）进行建模，而是引入连续的潜在变量z并采用p（z）p（xz）的形式。 为了训练VAE，我们优化边际似然性p（x）= p_θ（z）p（xz）dz。 边际对数似然可以写成：

$$\log p_\theta(x) = KL(q_\phi(z \mid x) \,||\, p_\theta(z \mid x)) + L(\theta, \varphi; x)$$

$$L(\theta, \varphi; x) = -KL(q_\phi(z \mid x) \,||\, p_\theta(z)) + E_{q_\phi(z \mid x)} \log p_\theta(x \mid z) \quad (1)$$

q_ϕ（zx），对后验p（zx）的变分近似，可以不同地解释为识别模型或编码器，由神经网络参数化以将句子x编码成密集码z。 （θ，φ；x）通常称为证据下界（ELBO）。 ELBO的第一项是来自先前的近似后验的KL散度，第二项是预期的重建误差。

由于KL分歧总是非负的，我们可以使用L（θ，φ；x）作为边际似然log p_θ（x）的下界。 我们优化L（θ，φ；x），共同学习识别模型参数φ和生成模型参数θ。

作为先前p（z）的选择，大多数先前的工作使用中心多元高斯p_θ（z）=

$\mathcal{N}(z; 0, I)$。 由于高斯是位置分布的分布族，因此在先验和后验中使用它们允许我们应用重新参数化技巧并在实践中优化ELBO时通过采样阶段$z \sim E_{q_\varphi(z|x)}$进行区分（金马和威灵，2013）。

2.1 案例研究：NVRNN

用于语言建模的神经变分RNN（NVRNN）描述于 鲍曼等人。（2016）并在图中描绘 1。 NVRNN模型的目标是将句子的高级表示提取到z中，并用神经语言模型重构句子。

我们将单词序列表示为$x = x_1, x_2, x_n$。 与vanilla语言建模不同，NVRNN在生成$p_\theta(x \mid z) = \prod_{i=1}^{n} p(x_i|x_1, \ldots, x_{i-1}, z)$的每一步的条件下对潜在变量z进行条件$p_\theta(x_1|z)$。 使用像LSTM这样的周期性模型来建模能力分布（Hochreiter和Schmidhu- 误码率，1997）如图所示 1。 这个选择没有什么独特之处；其他循环序列模型，如CNN或变压器（Vaswani等。，2017） 可用于。

2.2 后塌陷

在训练VAE时，我们同时更新θ和φ。 优化方程 1 给出两个梯度项：重建损失的更新（正确标签的可能性）和KL分歧的更新。 虽然重建损失项鼓励z将有用信息传递给该模型，但KL项始终试图在每次梯度更新时将$q(z|x)$调整为先验。 这可能会将模型陷入一个糟糕的局部最优位置，其中$q_\varphi(z|x) = p_\theta(z)$对于所有x：在这种情况下，z只是一个噪声源，对模型没用，所以模型已经学会忽略它并且不会产生足够大的梯度更新以使$q(z \mid x)$超出此最佳值。

鲍曼等人。（2016称这个问题KL col- 失效并提出退火时间表来处理它，其中KL术语的重量在训练过程中增加。[2] 通过这种方式，模型最初学习使用潜在的代码，然后随着训练的进行而向先前规则化。 然而，这种技巧并不足以避免KL在所有情况下崩溃，特别是

	没有退火		Sigmoid退火3层	
		1层	3层	1层
在	0.00	3.37	1.05	**6.52**
NLL	135	129	132	**125**

表1：使用和不使用退火技术在Penn Treebank上训练的两个NVRNN模型的开发集KL和NLL值 鲍曼等人。（2016）。 当不使用退火时，较高容量的3层模型会崩溃，虽然退火可以提高性能，但仍然不如1层变体那样好。 相比之下，具有vMF的1层模型的变体给出了NLL值117和KL为18.6，更强烈的结果更依赖于潜在变量。

当使用强解码器并且z对$p_\theta(x \mid z)$具有轻微影响时。

表 1 显示了类似设置的实验那个 鲍曼等人。（2016）。 我们使用四种不同的超参数设置在Penn Treebank上训练NVRNN模型。 我们使用3层LSTM编码器或1层LSTM并使用或不使用S形退火计划（在前20个时期内将KL权重从0增加到1）。 我们使用具有退火的1层模型观察到最佳性能。 可以从该表中得出结论，退火技巧已经起作用，因为当使用退火时两种模型都获得了更好的性能。 但事实上，基于vMF的模型可以比任何一个（NLL为117）做得更好，而且，我们无法知道更好的退火方案在训练后可能无法实现更高的性能。 此外，更高容量的3层模型理论上可以执行1层模型所能做的任何事情，因此其较低的性能表明我们的训练通过过度拟合或陷入潜在变量未使用的局部最优而脱轨。[3]

因此，从VAE中获得最佳性能是一个具有挑战性的问题，需要仔细调整目标函数和优化程序（鲍曼等人。，2016; 赵 等。，2017b; 希金斯等人。，2017）。 除了记录良好的KL崩溃问题之外，优化器可能只是在训练期间陷入局部最优状态，因此无法找到最有效地利用潜在变量的模型。

我们在本文中提倡的解决方案是

[2] 重量增加KL术语也用于β-VAE等方法（希金斯等人。，2017）和InfoVAE（赵等人。，2017b）。

[3] 在我们的实验中，由于其他超参数（包括编码器是单向还是双向LSTM），我们发现崩溃频率存在显着差异。

改变潜在空间的分布并简化优化问题。 在下一节中，我们描述了von Mises-Fisher分布及其在VAE中的应用，其中它迫使模型将潜在表示放在单元超球面上而不是将所有东西压缩到原点。 重要的是，这种分布让我们通过固定分布浓度参数 κ 来固定KL项的值；这避免了KL的崩溃，并在两种生成模型范例中获得了良好的模型性能。

3 冯米塞斯 - 费希尔VAE

von Mises-Fisher分布是R^d中（d 1）维球面上的分布。 vMF分布由方向矢量定义 μ，$\mu=\|\mu\|$，浓度参数 κ 0. d维单位矢量x的vMF分布的PDF定义为：

$$f_d(x;\mu,\kappa) = C_d(\kappa)\exp(\kappa\mu^T x) \quad (2)$$

$$C_d(\kappa) = \frac{\kappa^{d/2-1}}{(2\pi)^{d/2} I_{d/2-1}(\kappa)} \quad (3)$$

其中I_v代表修改后的贝塞尔函数订单v中的第一种

数字 1显示来自具有各种μ向量（箭头），d = 3和 κ =的vMF分布的样品

这是一个高κ值，导致样本紧密聚集在μ周围，这是平均值

和分配方式。 当 κ = 0时，分布退化为均匀分布

在超球面上独立于μ。

过去的工作使用vMF作为排放分配无监督聚类模型中的bution（Baner— jee等。，2005），其他领域的VAE（戴维森 等。，2018; Hasnat等人。，2017），以及文本的生成编辑模型（Guu等人。，2018）。 我们专注于文本建模的vMF的经验属性，并系统地检查此先验如何影响VAE模型与使用高斯模型。

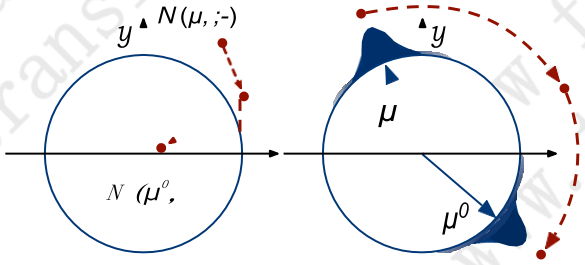
使用vMF的VAE我们将在我们的VAE模型中使用vMF作为我们的先验和变分后验。 否则，我们

（a）高斯 （b）vMF

图2：在学习期间针对单个示例的q如何随时间变化的可视化的可视化。 在高斯情形中，KL项倾向于将模型拉向先验（从 μ， σ 移动到 μ^t， σ^t），而在vMF情况下，没有这种压力朝向单个分布。

方向μ是编码神经网络的输出（图 1，右侧）和κ被视为常数。

在我们实施VAE之前，我们需要导出KL散度的表达式以优化ELBO（方程式 1并提供抽样



承认重新参数化技巧的算法（金马和威灵，2013).

KL分歧 使用vMF（, 0）作为我们的先验，KL分歧是：[4]

$$KL(vMF(\mu,\kappa)\,||\,vMF(\bullet,0)) = \kappa\frac{I_{d/2}(\kappa)}{I_{d/2-1}(\kappa)} + (d/2-1)\log\kappa - \frac{d}{2}\log(2\pi) - \log I_{d/2-1}(\kappa)$$

的VAE的设置保持与章节中建立的高斯情况相同 2。 我们的先验是均匀分布vMF（， κ = 0）。 由于真后验$p_\theta(z|x)$在 -易处理的，我们将用变分后验$q_\phi(z|x) = vMF(z;\mu,\kappa)$来近似它

$$+ \frac{d}{2}\log\pi + \log 2 - \log\Gamma\left(\frac{d}{2}\right)$$

关键的是，这只取决于κ，而不是μ。 κ将被视为固定的超参数，因此该项对于我们的模型将是恒定的； 因此KL崩溃将变得不可能。

数字 2 显示了高斯和vMF VAE的学习轨迹的可视化。 对于高斯VAE，目标函数中的KL散度倾向于将后向拉向前置于原点的中心，因此，如前所述，使得优化变得困难。 对于vMF VAE，给定固定 κ，没有这种空态，μ 可以自由变化。

[4]我们的KL分歧同意 戴维森等人。(2018)（参见他们的附录中的推导），我们已经凭经验验证了它。 方程式 Guu等人。 (2018)给出略微不同的KL值，但是对于我们遇到的大多数 κ 和维度值，差异很小（<5%）。

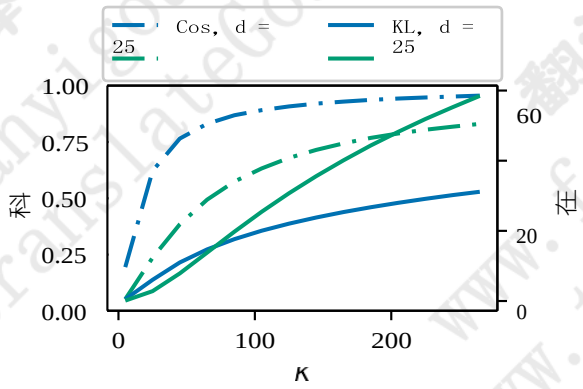


图3：vMF中 κ，KL和维度之间相互作用的可视化。Cos表示 μ 与来自 $\text{vMF}_d(\mu, \kappa)$ 的样本之间的余弦相似性，其反映了分布的分散程度。 KL被定义为具有均匀vMF事先KL的KL（$\text{vMF}_d(\mu, \kappa)$ vMF（，0））。 较高的 κ 值产生较高的余弦相似性，但KL成本较高。

数字 3 显示了两个不同维度的KL值和vMF（μ，κ）的浓度。 KL与 κ 单调增加，通过余弦相似性测量的浓度也是如此。 为了在维数增加时获得固定的余弦色散，需要更高的 κ 值，从而导致更高的KL值。

vMF的采样实施后 Guu等人。 (2018)，我们使用拒绝抽样方案 木 (1994) 采样“变化幅度”w。 然后给出我们的样本
通过z =wμ+ v $\sqrt{1 - w^2}$，其中v是随机的

采样单位向量与超球面相切
μ。 v和w都不依赖于 μ，因此我们现在可以根据需要采用相对于 μ 的z的梯度。

4 语言建模实验

我们首先在NVRNN设置中评估我们的vMF方法。我们将返回此模型并在章节中进一步分析其属性 6 和 7 在展示文档建模实验之后。

数据集对于NVRNN，我们使用Penn Treebank（马库斯等人。，1993），也用于 鲍曼等人。(2016) 和Yelp 2013（徐等人。，2016）。Yelp数据集中的示例比来自PTB的数据集更长且更多样化，需要更多地理解高级语义以生成连贯序列。 Yelp有很长的评论，所以我们将这些例子截断为最多50个字； 这仍然使得平均长度超过PTB设置的两倍。

名称	培养	开发	测试	莱恩	翻译
肺结核喊叫	42068	3370	3761	21.1	10K
20ng的	62522	7773	8671	49.5	15K
	11268	-	7505	96.1	2K

表2：我们实验中使用的数据集的统计数据。 Len代表一个例子的平均长度。
Vocab是词汇量；这些都是先前的工作。

关于本文中使用的所有数据集的统计数据是如表所示 2.

设置我们评估我们的NVRNN 弓- 男人等。(2016) 并探索两种不同的设置。 在标准设置中，每个时间步的RNN输入是潜在代码z和最后一个时间步的地面实况字的串联，而无输入设置不使用前一个字。 标准设置的功能更强大的解码器使潜在表示本身不太有用。 在无输入设置中，解码器需要仅在给定潜在代码的帮助下预测整个序列。 在这种情况下，迫切需要高质量的句子表示，并且驱动模型来学习它。

我们的VAE实现使用单层单向LSTM作为编码器和解码器。 我们在LSTM中使用100的嵌入大小和400的隐藏单位。 拉的尺寸
帐篷代码通过调整从{25, 50, 100}中选择

在开发集上。 我们使用SGD来优化所有具有衰减学习率和梯度削减的模型。 对于Yelp，情绪位（范围从1到5）也嵌入到50维向量中，并在解码阶段的每个时间步进输入。

结果NVRNN的实验结果如表所示 3。 我们报告负对数似然（NLL）[5] 和测试集上的困惑（PPL）。 我们遵循报告的实施 弓- 男人等。(2016) 对于高斯VAE，KL项权重退火； vMF VAE无需重量退火即可正常工作。 vMF分布为标准和无输入设置中的所有数据集提供了性能提升。 即使在标准设置中，我们的模型也能够成功地使用非零KL值来实现更好的效果

[5]报告的值实际上是真实NLL的下限，通过采样z从ELBO计算得出。

模型	标准NLL	肺结核 PPL	无输入 NLL	PPL	喊叫 标准NLL	PPL	无输入NLL	PPL
RNNLM（2016）	100（-）	116	135（-）	>600	–	–	–	–
G-VAE（2016）	101（2）	119	125（15）	380	–	–	–	–
RNNLM（我们的）	100（-）	114	134（-）	596	199（-）	55	300（-）	432
G-VAE（我们的）	99（4.4）	109	125（6.3）	379	199（0.5）	55	274（13.4）	256
vMF-VAE（我们的）	**96（5.7）**	**98**	**117（18.6）**	**262**	**198（6.4）**	**54**	**242（48.5）**	**134**

表3：NVRNN在PTB和Yelp测试组上的实验结果。 上部RNNLM和G-VAE显示结果 鲍曼等人。（2016）。 KL分歧显示在括号中，同时显示总NLL。 最佳结果以粗体显示。 vMF始终使用更高的KL术语权重，但在所有四个设置中实现了相当或更好的NLL和困惑值。

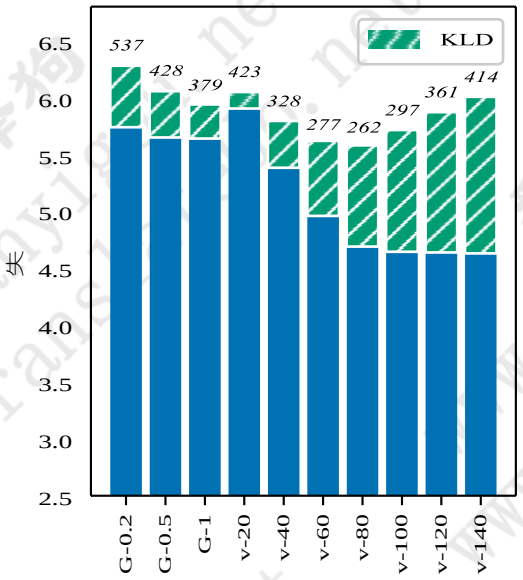


图4：具有不同超参数的高斯和vMF-NVRNN的比较。在无输入设置中对所有模型进行PTB训练，其中潜在维度为50.G-α表示高斯VAE，其中KL通过给定常数 α 退火，并且V-κ表示VAE，其中 κ 设置为给定值。 绿色条反映KL损失量，而总高度反映整个目标。 上面的数字是困惑。 vMF具有更高的可调性，并且在各种 κ 值范围内都能获得更强的结果。

困惑，甚至当KL崩溃似乎不是这种情况时（例如，PTB标准设置上的G-VAE），高斯分布族导致较低的KL和较差的对数似然，可能是由于优化挑战。 在无输入设置中，我们看到了很大的收益：与高斯VAE相比，vMF VAE在PTB中将PPL从379降低到262，在Yelp中降低到256到134。

权衡比较除了整体困惑之外，我们也对以下权衡感兴趣：

重建损失和KL，以及KL对整个目标的贡献。数字 4 显示了我们的模型明确控制KL和重建术语之间平衡的能力。 首先，我们通过将KL项的权重设置为小于1的常数（在我们的情况下为0.2和0.5）来“永久地”退火高斯VAE。我们发现这个技巧确实可以缓解KL崩溃，但整体表现更差。 因此，这不仅是关于KL与NLL权衡的数字游戏，而且是如何构建模型以学习有效潜在表征的更深层次的挑战。

对于vMF VAE，当我们逐渐增加 κ 的值时，平均方向 μ 周围的分布浓度更高，并且来自vMF的样本更接近 μ。 当 $\kappa = 80$时，该模型实现了最佳的困惑。由于任务的难度和LSTM解码器的有限容量，重建误差约为4.5。虽然 κ 是需要调整的超参数，但模型总体上对它不是很敏感，我们在Section中显示 7 合理的κ值在类似的任务中传递。

5 文档建模实验

我们还研究了vMF VAE在不同环境中的表现，不受KL崩溃问题的困扰。 具体而言，提出了神经变分文档模型（NVDM） 苗 等。（2016），是一种基于VAE的无监督文档模型。 该模型遵循Section中介绍的VAE框架 2。 我们的文档表示是文档中单词存在或不存在的指示符向量x。由于这是一个固定大小的表示，我们使用2层MLP

编码器q（z | x）的400个隐藏单元

和解码器p（x | z）；解码器放置简单

型号fDARN	暗淡	20ng 的	RCV1
(2014)	50	917	724
	200	-	598
G-NVDM (2016)	50	836	563
	200	852	550
v-NVDM (我们的)	25	**793**	558
	50	830	**529**
	200	851	[illegible]

表4：文档建模任务的测试集困惑。 前馈深度自回归神经网络（fDARN）由实现 Mnih和 格雷戈尔 (2014）。 提出了基于高斯的NVDM（G-NVDM） 苗等人 (2016）。 Dim表示潜码的维度。 我们的v-NVDM模型大大优于过去的模型。

词汇中单词的多项式分布，文档的概率是其词语概率的乘积。

数据集对于NVDM，我们使用两个标准新闻语料库，20个新闻组（20NG）和路透社RCV1-v2，它们被用于 苗等人 (2016).[6]

结果实验结果[7] 如表所示 4。 与NVRNN相比，NVDM完全依赖潜在代码的功能来预测字分布，因此我们从未观察到KL崩溃，但vMF仍然比高斯更好。 如图所示 3为了在变分后部的样本中保持相同的色散量，较大的潜在维度需要较大的κ值和相应较大的KL项值。 对于比RCV1小得多的20NG，因此尺寸更小\ 提供更好的表现。 对于两个数据集，κ = 100，dim = 25和 κ = 150，dim∈的设置 {50，200}运作良好。

6 我们的VAE编码什么？

我们设计了更多探测任务来演示由vMF VAE诱导的潜在表示中编码的内容。 我们在这里探索的另一个模型变体是NVRNN-BoW模型。 这是NVRNN的变体，其中解码器另外

[6]可以从中下载预处理版本 http：／／／／nvdm github.com ysmiao

[7]我们没有比较结果 Serban等人。 (2017a）。 与我们目前的结果相比，该工作报告了20NG的非常强劲的表现和RCV1的非常弱的表现；我们认为他们要么使用不同的预处理，要么在报告结果时犯了错误，但无法与作者确认。

P (x | z, BoW)	标准		无输入NLL	
		PPL	NLL	PPL
RNNLM	79 (–)	43	106 (–)	152
的G -阿联酋	79 (0.0)	43	106 (0.4)	153
V的阿联酋	**73 (0.2)**	**33**	**93 (11.4)**	**82**

表5：NVRNN-BoW在PTB上的实验结果；也就是说，解码器还根据要生成的句子的一个单词表示条件。 在这种情况下，高斯模型表现出KL崩溃，但vMF仍然可以有效地学习。

向量上的条件BoW $= \frac{1}{n} e(x_i)$，句子x的平均单词嵌入值。 在人工设置的同时，通过使这种信息的形式独立可用，我们可以看到潜在代码如何有效地捕获除简单单词选择之外的信息。 表 5 在此设置中显示结果，我们再次看到高斯模型的KL崩溃问题以及vMF在标准和无输入设置中的困惑性方面的更好性能。

潜在的代码不仅仅是一袋字吗？ 对于所有这些模型，一个假设是编码器可能正在学习记忆单词包，然后优先从解码器生成该包中的单词。 为了验证这一点，我们研究是否可以相互重建BoW表示和学习的潜在代码。 具体地，给定句子x，我们可以计算如上定义的BoW和 μ = enc（x），x的潜在编码由编码器输出的平均向量表示。 我们可以使用一个简单的多层感知器来尝试从单词包到潜在代码：

μ MLP(BoW)，然后学习参数

通过最小化样品上的$\mu\mu^2$来实现MLP。 可以使用相同的过程来学习从 μ 回到单词包的映射。

表 6 显示了在Gaussian和vMF模型下我们的重建的平均余弦相似性。 对于vMF， μ 可以比词袋重建 μ 更准确地重建词袋，表明vMF中的潜码捕获了超出词袋的更多信息。

我们在单独的NVRNN模型中重复该实验，其中解码器可以明确地对上述BoW矢量进行调节。 结果显示在表的右栏中 6。 我们的模型v-VAE实现了较低的余弦

模型	NVRNN		NVRNN弓
设置	$\mu \rightarrow$BoW	BoW$\rightarrow \mu$	$\mu \rightarrow$BoW
的G -阿联酋	0.74	0.74	0.32
V的阿联酋	0.77	0.57	0.23

表6：尝试从单词包中重建潜在代码 μ 时的平均余弦相似度，反之亦然。 在vMF中，潜在代码包含超出单词包的更多信息，如预测BoWμ（0.57）时的较低余弦相似性所示。 当潜在代码在以词袋（右栏）为条件的模型中学习时，它预测词袋不太好，表明该模型成功地学习了正交信息。

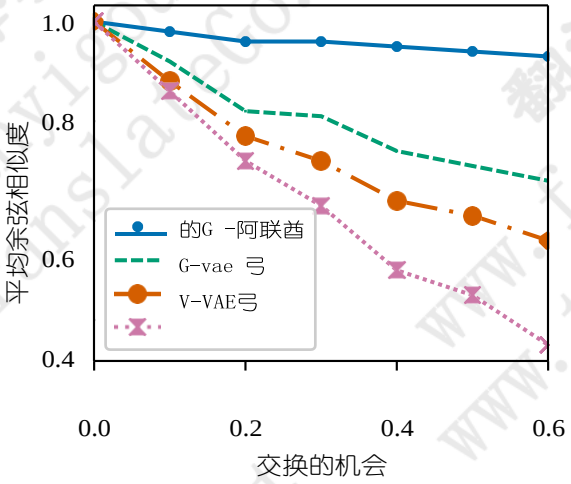


图5：潜码对交换编码序列的相邻字的敏感性。在原始句子的潜码（编码平均向量）和应用交换之后的句子之间测量余弦相似度。 我们看到vMF对NVRNN和NVRNN-BoW设置中的交换更加敏感，表明其潜在空间可能编码更多的订购信息。

相似性比G-VAE（0.23对0.32），表明它捕获较少的冗余信息并使用潜在空间更有效地建模数据的其他属性。

对词序表的敏感性 6 表明带有vMF的NVRNN对信息包之外的信息进行编码；一个自然的假设是它正在编码单词顺序。 我们可以在NVRNN和NVRNN-BoW设置的上下文中更直接地研究这一点。 灵感来自 赵等人。(2017a)，我们提出了一个实验，探测在PTB上的无输入设置中随机交换相邻词对的灵敏度。 我们改变交换每个单词对的概率，并查看潜在代码随着交换数量的增加而变化的方式。 理想情况下，我们的模型应捕获订购信息，因此对此更改敏感。

数字 5 显示结果。 v-VAE的表示比G-VAE更敏感：随着交换变得更可能，它们的变化更快。[8] 在NVRNN-BoW设置中，我们看到模型更加敏感。 vMF使我们能够更容易地在句子编码中学习这种理想的信息。

7 用κ控制方差

到目前为止，我们方法的一个核心方面是将κ视为固定的超参数。 从优化的角度来看，修复 κ 是有益的：它使模型更难以陷入局部最优。 但它也降低了模型的灵活性，因为我们不能再预测每个例子的 κ 值，并且它引入了另一个参数系统设计师必须调整。

幸运的是，各种κ值对于我们考虑的任务似乎都很有效。 数字 6 表示当潜在维度和其他超参数保持固定时，浓度参数κ如何改变PTB上的结果。 在潜在表示的必要性方面，我们已经从左到右将任务从“最难”命令到“最简单”：无输入设置需要来自潜在代码的重信息来重建句子，而标准-BoW设置具有极其重要性强解码器来预测下一个字。 我们看到，在每种情况下，各种 κ 值都有效，而且两个标准之间和两个无输入设置之间的合理 κ 值传递，表明整体方法对这些超参数值不是非常敏感。

学习κ的脆弱性在整个这项工作中，我们将κ视为固定参数。 但是，我们可以在高斯情况下以与 σ 相同的方式处理 κ，并在每个实例的基础上学习它。 对于给定第一类修正贝塞尔函数的梯度的 κ，vMF的KL偏差是可微分的，[9] 允许我们根据每个实例更改浓度。 然而，这再次引入了KL崩溃的问题：KL术语将鼓励 κ 低

[8] 这里的高斯VAE几乎没有使用潜在的变量，因此表示变化很小的原因。

[9] $\nabla_{\kappa} I_d(\kappa) = \frac{1}{2}(I_{d-1}(\kappa) + I_{d+1}(\kappa))$

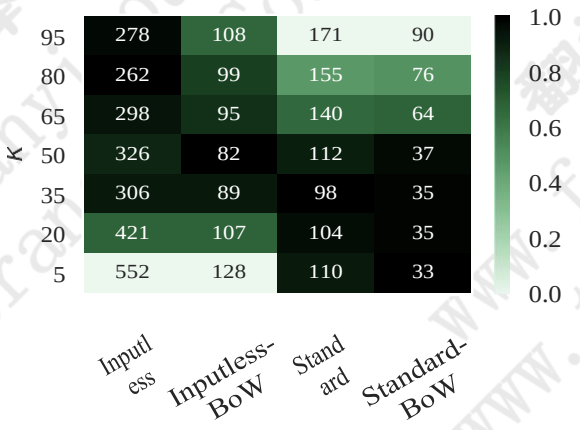


图6：当潜在尺寸为50时，具有不同 κ 值的不同设置中的v-VAE的困惑度。较暗的颜色对应于更接近于该设置观察到的最佳值的困惑值。 对于每个任务，我们看到有一系列 κ 值运行良好，并且这些值在可比任务之间传递。

尽可能使潜在的变量变空。

在实践中，我们观察到由于数值原因有必要将κ值剪辑到某个范围。 在此范围内，模型倾向于最小的 κ 值，并且比使用我们的固定 κ 方法训练的模型表现得更差。 这表明即使使用vMF模型，ELBO提出的优化问题也只是一个难点，而固定KL分歧的方法是一种令人惊讶的优秀技术。

8 相关工作

VAE在NLP中的应用深度生成模型在NLP附近的域中取得了令人瞩目的成功，例如图像生成（格雷戈尔等人。，2015; Oord等人。，2016a）和语音生成（Chung等人。，2015; Oord等人。，2016b）。 特别是VAE（金马和威灵，2013; Rezende等人。，2014）一直是NLP中流行的模型变体。 它们已应用于包括文档建模在内的任务（苗等人，2016），语言建模（鲍曼等人。，2016）和对话生成（Serban等人。，2017b）。 VAE也可用于半监督分类（徐等人。，2017）。 最近关于标准VAE方法的曲折包括将VAE和整体属性鉴别器结合起来用于条件生成（胡等人。，2017并使用通过对抗方法正规化的更灵活的潜在空间（赵等人。，2017a).

VAE目标最近的几项工作突出了优化VAE目标的问题。 Alemi等。 (2018从信息论的角度阐明问题。 赵等人。 (2017b）和 希金斯等人。 (2017）两者都提出了目标的各种权重以及理论和经验证明。

VAE Priors的选择过去的一些工作已经探索了VAE的各种先驱。 Serban等人。 (2017a）提出了一种分段常数分布，它处理多种模式，但牺牲了连续插值的性质。 Guu等人。 (2018）也在VAE模型中应用了vMF，但在句子编辑案例中特别使用了它们。 戴维森等人。 (2018）在MNIST的VAE模型和链路预测任务中探索了vMF。 Hasnat等人。 (2017）应用vMF分布进行面部识别。 其他过去的工作使用了不同的解码器，包括CNN（杨 等。，2017）和CNN-RNN混合动力车（Semeniuta 等。，2017）。 改变解码器是一个很大程度上改变先前的变化：它可以减轻KL消失的问题，但它不一定能扩展到新的设置，也不能明确控制潜在代码的利用。

9 结论

在本文中，我们建议使用von Mises-Fisher VAE来解决变量自动编码器中文本的优化问题。这种分布选择使我们能够以原则的方式明确地控制解码器的容量和潜在表示的利用之间的平衡。 实验结果表明，所提出的模型在一系列设置中具有比高斯VAE更好的性能。 进一步分析表明，vMF VAE对字序信息更敏感，可以更有效地利用潜码空间。

致谢

这项工作得到了NSF Grant IIS-1814522，彭博数据科学基金和NVIDIA设备资助的部分支持。 作者承认德克萨斯大学奥斯汀分校的德克萨斯高级计算中心（TACC）提供用于进行此项研究的HPC资源。 还要感谢匿名审稿人的有益评论。

参考

Alexander Alemi, Ben Poole, Ian Fischer, Joshua Dillon, Rif Saurous和Kevin Murphy。 2018.修复损坏的ELBO。 国际机器学习会议。

Arindam Banerjee, Inderjit S Dhillon, Joydeep Ghosh和Suvrit Sra。 利用von Mises-Fisher分布对单位超球面进行聚类。 机器学习研究杂志，6。

Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz和Samy Bengio。 2016.从连续空间生成句子。 第20届SIGNLL计算自然语言学习会议论文集。

Xi Chen, Diederik P Kingma, Tim Salimans, Yan Duan, Prafulla Dhariwal, John Schulman, Ilya Sutskever和Pieter Abbeel。 2016.变分有损自动编码器。 arXiv preprint arXiv: 1611.02731。

Junyoung Chung, Kyle Kastner, Laurent Dinh, Kratarth Goel, Aaron C Courville和Yoshua Bengio。 2015.顺序数据的循环潜变量模型。 神经信息处理系统的进展。

Tim Davidson, Luca Falorsi, Nicola Cao, Thomas Kipf和Jakub Tomczak。 2018.超球面变分自动编码器。 第34届人工智能不确定性会议。

Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Rezende和Daan Wierstra。 绘图：用于图像生成的递归神经网络。 国际机器学习会议。

Kelvin Guu, Tatsunori B. Hashimoto, Yonatan Oren和Percy Liang。 2018.通过编辑原型来生成句子。 计算语言学协会的交易，6：437-450。

Md Hasnat, Julien Bohn, Jonathan Milgram, Stphane Gentric和Liming Chen。 2017. von Mises-Fisher混合模型的深度学习：面部验证的应用。 arXiv preprint arXiv: 1706.04264。

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed 和 Alexander Lerchner。 2017.β-VAE：使用约束变分框架学习基本视觉概念。 学习代表国际会议。

Sepp Hochreiter和Ju ̈rgenSchmidhuber。 1997.长短期记忆。 神经计算，9。

胡志婷，杨子超，梁晓丹，Ruslan Salakhutdinov和Eric P Xing。 2017.走向受控的文本生成。 国际机器学习会议。

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

Mitchell P. Marcus, Mary Ann Marcinkiewicz和Beatrice Santorini。 1993.建立一个大型注释英语语料库：宾州树库。 COM-放。 语言学家，19（2）：313-330。

Yishu Miao, Lei Yu和Phil Blunsom。 2016.文本处理的神经变分推理。 国际机器学习会议。

Andriy Mnih和Karol Gregor。 2014.信念网络中的神经变分推理和学习。 国际机器学习会议。

Aaron Oord, Nal Kalchbrenner 和 Koray Kavukcuoglu。 2016a。 像素回归神经网络。 国际机器学习会议。

Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior和Koray Kavukcuoglu。 2016B。 WaveNet：原始音频的生成模型。 arXiv preprint arXiv: 1609.03499。

Danilo Rezende, Shakir Mohamed 和 Daan Wierstra。 深层生成模型中的随机反向传播和近似推断。 国际机器学习会议。

Stanislau Semeniuta, Aliaksei Severyn 和 Erhardt Barth。 用于文本生成的混合卷积变分自动编码器。 2017年自然语言处理经验方法会议记录。

Iulian Serban, Alexander G Ororbia, Joelle Pineau和Aaron Courville。 2017A。 神经变分文本处理的分段潜变量。 2017年自然语言处理经验方法会议记录。

Iulian Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville和Yoshua Bengio。 2017b。 一种生成对话的分层潜变量编码器 - 解码器模型。 AAAI人工智能会议。

沉小霄，陶磊，Regina Barzilay 和 Tommi Jaakkola。 2017.通过交叉对齐从非平行文本转换样式。 神经信息处理系统的进展。

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, ukasz Kaiser和Illia Polosukhin。 2017.注意力就是你所需要的。 神经信息处理系统的进展。

安德鲁伍德。 1994.冯米塞斯费希尔分布的模拟。 统计通信 - 模拟与计算，23（1）：157-164。

徐家成，陈丹露，邱锡鹏，黄玄静。 2016.用于文档级情感分类的缓存长短期记忆神经网络。2016年自然语言处理经验方法会议记录。

徐伟迪，孙浩泽，邓超，谭莹。 2017.用于半监督文本分类的变分自动编码器。 AAAI人工智能会议。

Zichao Yang，Zhiting Hu，Ruslan Salakhutdinov和Taylor Berg-Kirkpatrick。 2017.使用扩张卷积进行文本建模的改进的变分自动编码器。 国际机器学习会议。

俞涛涛，张渭南，王军，雍宇。 2017. SeqGAN：具有策略梯度的序列生成性对抗网。 AAAI人工智能会议。

张彪，戴德雄，洪端，张敏。 2016.变分神经机器翻译。 2016年自然语言处理经验方法会议记录。

赵俊波，金尹，张凯莉，亚历山大M拉什和Yann LeCun。 2017A。 对抗正规化自动编码器。 的arXiv。

赵胜嘉，宋嘉明和斯特凡诺 埃尔蒙。 2017b。 InfoVAE：最大化变分自动编码器的信息。 的arXiv。