**Variational Pretraining for Semi-supervised Text Classification**

会议：ACL2019

代码：<http://github.com/allenai/vampire>

摘要

我们介绍了VAMPIRE，这是一种在数据和计算资源有限的情况下用于有效文本分类的轻量级预培训框架。 我们预训练一个unigram文档模型作为域内未标记数据的变体自动编码器，并将其内部状态用作下游分类器中的特征。从经验上，我们显示了VAMPIRE在资源匮乏的情况下，相对于计算量大的上下文嵌入和其他流行的半监督基线的相对强度。 我们还发现，在有限的监管下，微调到域内数据对于从上下文嵌入中获得良好的性能至关重要。 我们在本文中随附代码以预训练并在下游任务中使用VAMPIRE嵌入。

**1简介**

长期以来，一种有效的半监督学习方法一直是NLP社区的目标，因为与标记数据相比，未标记数据趋向于丰富。 早期的工作强调使用与标签数据分布相同的分布获得的未标签数据（Nigam等，2000），但通过对语言模型（LM）目标进行训练的上下文嵌入，在大量数据中获得了更大，更可靠的收益。 来自维基百科或新闻领域的文本（Peters等，2018a; Devlin等，2019; Radford等，2018; Howard and Ruder，2018）。 后一种方法发挥了高资源设置的优势（例如访问Web级语料库和功能强大的机器），但是它们的计算和数据需求可能使它们在资源受限的环境中的用处不大。 在本文中，我们重点关注低资源设置（x2.1），并开发了一种轻量级的方法来进行半监督文本分类的预训练。

我们的模型称为VAMPIRE，将变分自动编码器（VAE）方法用于文档建模（Kingma和Welling，2013年； Miao等，2016年； Srivastava和Sutton，2017年）带有LM预训练的见解（Peters等， 2018a）。 通过对词袋的表示进行操作，我们避免了训练序列到序列VAE的时间复杂性和难度（Bowman等人，2016; Xu等人，2017; Yang等人，2017） 保留使用多层编码器的自由度，该多层编码器可以学习下游任务的有用表示形式。 因为VAMPIRE忽略了顺序信息，所以它导致模型的训练要便宜得多，并且在标记数据量较小时可以提供强大的性能。 最后，由于VAMPIRE是主题模型的后代，因此我们能够通过主题一致性而不是验证集复杂性来探索模型选择，从而获得更好的下游分类性能（x6.1）。

为了评估我们方法的有效性，我们尝试了四个文本分类数据集。 我们将我们的方法与传统的半监督基线（自我训练），可以访问域内数据的替代表示学习技术以及使用对域外数据进行训练的大型语言模型的全面替代进行比较（可选） 调整到任务域。

我们的结果表明，对于有限的资源设置，可以实现有效的半监督学习，而无需基于计算的基于序列的模型。虽然我们观察到将预训练的BERT模型微调到域可以提供最佳结果，但这取决于相关语言中这种模型的存在以及GPU对其进行微调。 如果这不是一种选择，我们的该模型以最少的计算要求即可提供与替代方案相当或更高的性能，尤其是在处理数量有限的标记数据时。

本文的主要贡献是：

1. 我们针对半监督文本分类（x3）将变体文档模型调整为现代的预训练方法，并强调模型选择（x3.2）的适当标准的重要性。
2. 我们通过实验证明，当数据和计算受到限制时（x5），我们的方法是一种有效的半监督文本分类方法。
3. 我们确认使用上下文嵌入进行文档分类时微调是必不可少的，并为希望在半监督文本分类（x8）中使用未标记数据的研究人员提供了实用建议的摘要。
4. 我们发布代码以预训练未标记数据上的变分模型，并在下游任务中使用学习的表示形式。

**2背景**

**2.1资源受限的环境**

在本文中，我们对低资源设置感兴趣，该设置要求对计算，标签和域外数据的访问受到限制。 对于某些任务，可以廉价获得标签数据，但对于另一些任务，标签可能需要昂贵且耗时的人工注释，这可能需要领域专家的人工注释，这将限制其可用性。

虽然对于某些语言（例如英语）有大量未标记的文本可用，但并非所有语言都可以使用此数据规模。 当然，域内数据可用性因域而异。对于许多研究人员而言，尤其是在STEM领域之外，计算也可能是一种稀缺资源，因此从头开始训练上下文嵌入，甚至将它们整合到模型中都可能会非常昂贵。

此外，即使有这样的预训练模型，基于训练数据也不可避免地会产生潜在的不良偏见（Recasens等人，2013; Bolukbasi等人，2016; Zhao等人，2019 ）。特别是对于社会科学应用，可能更可取的是仅通过处理域内或策展的数据来排除此类混杂因素。

考虑到这些限制和局限性，我们寻求一种半监督学习的方法，该方法可以利用域内未标记的数据，仅使用少量标记的实例即可达到高精度，并且可以在CPU上高效运行。

**2.2半监督学习**

已经为NLP开发了许多半监督学习的方法，包括自举的变体（Charniak，1997； Blum和Mitchell，1998； Zhou和Li，2005； McClsky等，2006），以及使用生成模型或单词的表示学习（Mikolov等，2013； Pennington等，2014）。 上下文化嵌入最近已成为使用域外数据的有效方法（Peters等，2018a; Radford，2018），但是训练这些大型模型需要大量适当的数据（通常大约数百万个字）和行业规模的计算资源（在多个GPU上数百小时）。

也有人尝试利用VAE在NLP中进行半监督学习，主要采用序列到序列模型的形式（Xu等，2017; Yang等，2017），它们使用基于序列的编码器和解码器（ 参见x3）。 这些论文报告了出色的性能，但是还有许多悬而未决的问题需要进一步研究。首先，鉴于报道的训练序列到序列VAE的难度（Bowman等人，2016），令人怀疑的是这种方法是否在实践中有用。 此外，尚不清楚是否确实需要这种复杂的模型（训练成本很高）才能在诸如文本分类之类的任务上取得良好的性能。

在这里，我们改为基于神经文档模型建立框架（Miao等，2016; Srivastava和Sutton，2017; Card等，2018），它们提供更快的培训和主题形式的明确解释，并探索它们在半监督环境中的效用。

**3 Model**

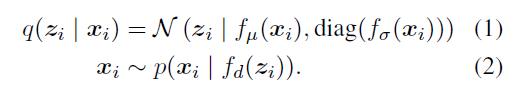
在这项工作中，我们假设我们有L个文档，，并且观察到类标签我们还假定可以访问从同一分布中抽取的较大的U文档集，但是对于这些文档而言，标签不可见，即。 我们的主要目标是学习概率分类器。

我们的方法大量借鉴了过去关于VAE的工作（Kingma and Welling，2013; Miao et al。，2016; Srivastava and Sutton，2017），我们采用了半监督文本分类（见图1）。 通过对未标记数据（x3.1）进行文档模型预训练，然后在下游分类器（x3.3）中使用学习的表示形式，可以做到这一点。 下游分类器利用了预训练文档模型的多个内部状态，如Peters（2018b）等人所述。我们还将探索如何以有利于下游任务（x3.2）的方式最好地进行模型选择。

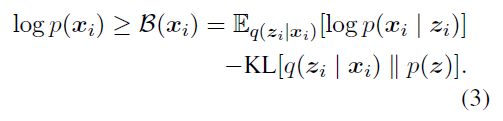
**3.1无监督预训练**

为了学习有用的表示形式，我们首先忽略标签，并假设每个文档都是由潜在变量z生成的。 然后，在估计此模型时学习的特征将提供表示形式，这些表示形式将用作监督学习中的特征。

使用变分自动编码器进行近似贝叶斯推断，我们同时学习了一个编码器和一个解码器，该编码器从观察到的文本映射到近似后验，并从潜在表示中重建文本。实际上，我们将编码器和解码器都实例化为神经网络，并假设编码器映射到正态分布的后验，即对于文档i，



使用变分推论的标准原理，我们得出了观测数据的边际对数似然率的变分界，



直观上，可以将边界中的第一个项视为重构损失，从而确保生成的单词与原始文档相似。第二项是KL散度，它鼓励变分近似接近假定的先验p（z），我们将其作为球面正态分布。

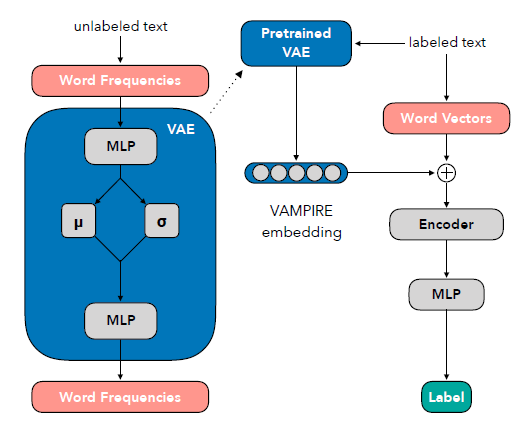
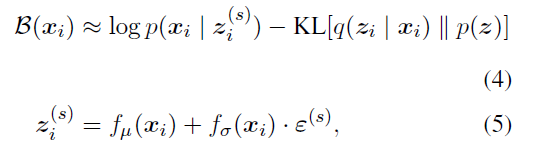


图1：VAMPIRE涉及在未标记的文本上预先训练深层变分自动编码器（VAE；显示在左侧）。 VAE完全由前馈网络组成，它学习用逻辑对数先验重构未标记文本的词频表示，由参数表示。 在下游，将预先训练的VAE的内部状态冻结并连接到特定于任务的单词向量，以改善资源短缺情况下的分类。

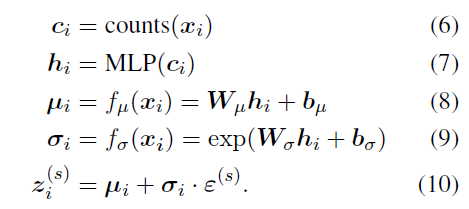
使用重新参数化技巧（Kingma and Welling，2013; Rezende et al，2014），我们将期望值替换为单样本近似值，即



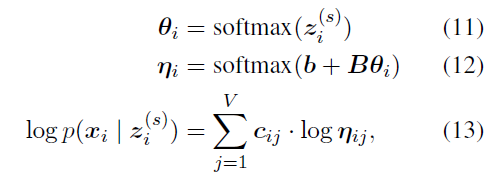
从独立正则采样。 然后可以通过对变化边界执行随机梯度上升来同时优化所有参数。

编码和解码文本的有效方法是使用序列模型。 也就是说，和会从令牌序列映射到一对向量，并且会类似地从z解码到令牌序列，使用递归、卷积或基于注意力的网络。一些作者采用了这种方法（Bowman等人，2016; Xu等人，2017; Yang等人，2017），但是如上所述（x2.2），它有很多缺点。

在本文中，我们采用了一种更轻量级且可直接解释的方法，并使用单词频率而不是单词序列。 使用与Miao（2016）等人相同的基本结构，但在解码器中采用softmax，我们使用在字数输入ci上运行的多层前馈神经网络对进行编码：



对于解码器，我们使用以下形式，该形式根据主题（词汇上的一致分布）重构输入：



j覆盖整个词汇表。

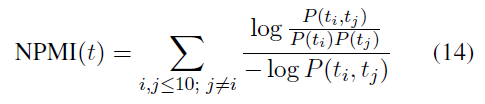
通过在z上放置softmax，我们可以将解释为潜在主题的分布（如主题模型（Blei等人，2003年）），而B则代表与背景b的正负主题偏差。 与基于序列的编码器和解码器相比，这种形式（本质上是一个字母组合LM）可以对z进行更有效的推断。

**3.2通过主题连贯性进行模型选择**

由于我们的预培训忽略了文档标签，因此优化文档以使其收敛将为下游分类产生最佳表示并不明显。 当使用LM物镜进行预训练时，通常会训练模型，直到模型拟合不再改善（即验证数据上的困惑）为止。 但是，在我们的案例中，自然地解释为模型（B）在潜在“主题”上分布的（针对文档i）。 因此，一种替代方法是将主题的质量用作早期停止的标准。

反复观察到，不同类型的主题模型提供了困惑和主题质量之间的权衡（Chang等，2009； Srivastava和Sutton，2017）。 已经提出了几种自动评估主题一致性的方法（Newman等，2010； Mimno等，2011），例如Lau等人的归一化点向互信息（NPMI）。 （2014）被发现是与人类判断最相关的。 因此，我们考虑使用对数似然或NPMI作为VAMPIRE预训练（x6.1）的停止标准，并对其进行评估，以得出更好的下游分类器。

NPMI衡量两个单词在外部语料库中并置的概率（在我们的示例中为验证数据）。 对于B中的每个主题t，我们收集前十个最可能的单词并计算所有对之间的NPMI：

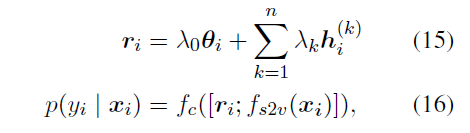


然后，我们通过平均所有主题的NPMI得出B的全局NPMI。 我们在预训练期间的每个时期结束时评估NPMI，并在NPMI在预定次数的时期停止增加时停止训练。

**3.3使用预训练的VAE进行文本分类**

金马等（2014年）提出了使用无监督VAE的潜在变量作为下游模型中图像分类的特征。 但是，为NLP进行预训练的工作，例如Peters等（2018a）发现，LM在不同的层中编码不同的信息，对于某些任务而言，每一个层或多或少都很有用。 在此，对于基于字数ci的n层MLP编码器，我们基于该思想，并使用和MLP的内部状态的加权和作为表示，权重由下游分类器学习。

也就是说，对于任何序列到向量的编码器，我们建议通过使用变分编码器内部状态的加权组合将每个文档的向量表示进行连接来增强它们的向量表示（Peters et al。，2018a） 。 然后，我们可以在加权组合上训练监督分类器，



其中fc是神经分类器，是softmax归一化的可训练参数。

**3.4优化**

在所有情况下，我们都使用Adam对模型进行优化（Kingma和Ba，2014）。 为了防止在预训练期间发散，我们在x的重建中使用了batchnorm层（Ioffe和Szegedy，2015）。 我们还使用KL退火（Bowman等人，2016），将等式（3）中的KL散度项标量权重，然后从零逐渐增加到1。 因为我们的模型完全由前馈神经网络组成，所以它很容易并行化，并且可以在CPU或GPU上高效运行。

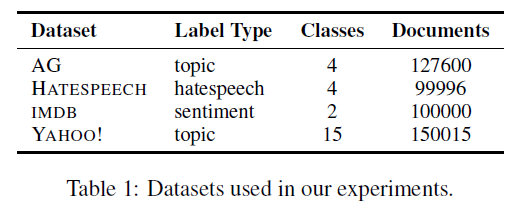
**4实验设置**

当我们改变标记数据的数量（从200到10,000个实例）时，我们将评估我们的方法在四个文本分类任务上的性能。在所有情况下，我们假设存在大约75,000至125,000个未标记的域内示例，这些示例来自未使用的训练数据和语料库提供的任何其他未标记的数据的并集。 因为我们正在处理少量的标记数据，所以我们用五个随机种子运行每个实验，每个种子都带有标记的训练实例的不同样本，并报告测试数据的平均性能。

**4.1数据集和预处理**

我们尝试使用跨越各种标签类型的文本分类数据集。 我们使用的数据集是熟悉的AG News（Zhang等，2015），IMDB（Maas等，2011）和YAHOO!。 答案数据集（Chang等人，2008），以及按四个HATESPEECH类别标记的推文数据集（Founta等人，2018）。 表1列出了摘要统计信息。在所有情况下，我们要么使用官方测试集，要么将25,000个文档的随机分层样本作为测试集。 我们还采样了5,000个实例作为验证集。

我们使用spaCy标记文档，并最多使用400个标记进行序列编码。 对于VAMPIRE预训练，我们将词汇表限制为数据集中的30,000个最常用词，然后排除短于三个字符的标记，带数字或标点的词和停用词。我们不限制用于下游分类的词表。

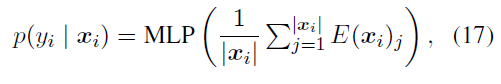


**4.2 VAMPIRE体系结构**

为了找到合理的VAMPIRE超参数，我们利用随机搜索策略进行预训练。 对于每个数据集，我们采用具有最佳NPMI的模型以供下游分类器使用。 我们在附录A.1的表5中详细说明了每个超参数的采样范围和最终赋值。

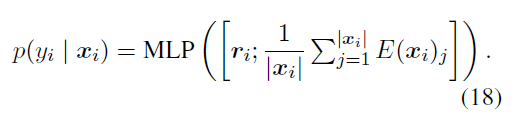
**4.3下游分类器**

对于所有实验，我们都使用深度平均网络（DAN）架构（Iyyer等人，2015）作为基线序列到矢量编码器。 即，将与每个令牌相对应的嵌入相加并通过多层感知器。



其中E（x）使用随机初始化的向量，现成的GLOVE嵌入（Pennington等人，2014）或上下文嵌入将标记序列转换为向量序列。

要将VAMPIRE学习到的文档表示形式并入下游分类器中，我们将其与随机初始化的可训练嵌入的平均值相连（即



初步实验发现，具有单层MLP和适度dropout的DAN在验证数据上的性能要比CNN或LSTM等更具表现力的模型更可靠，且超参数调整较少，尤其是在处理少数带标签实例的情况下（请参见附录A.2的详细信息）。 ）。

**4.4资源和基线**

在这些实验中，我们考虑低资源和高资源设置的基线，在高资源基线可以访问大量的计算资源以及大量的未标记数据或预训练的模型，例如ELMO或BERT的情况下。

**资源短缺** 在资源匮乏的情况下，我们假设计算资源非常宝贵，因此我们仅限于可以在CPU上高效运行的轻量级方法（例如VAMPIRE）。 作为基线，我们考虑：a）一个纯监督模型，具有随机初始化的50维嵌入，无法访问未标记的数据； b）使用300个维度的GLOVE向量初始化的同一模型，预训练了8400亿个单词； c）仅基于域内数据训练的300个维度的GLOVE向量； d）自我训练，可以访问域内未标记的数据。 对于自我训练，我们反复进行模型训练，在所有未标记的实例上预测标签，然后将训练后的所有未标记实例（其标签被预测为高可信度）添加到训练集中，最多重复五次，并使用验证精度最高的模型 。 在每次迭代中，给定标签的阈值等于带有相应标签的验证实例的预测概率的第90个百分位数。

**高资源** 在高资源环境中，我们假设可以访问大量的计算资源和大量的域外数据，这些数据可以通过预先训练的模型间接访问。 具体来说，我们评估了基于Transformer的ELMO（Peters等人，2018b）和BERT的性能，（a）现成的具有冻结嵌入的组件，以及（b）在半监督下微调至未标记和标记的域内数据。 为了执行半监督式微调，我们首先使用ELMO和BERT的原始物镜对未标记的数据进行微调。为了将ELMO微调到标记的数据，我们对LM状态进行平均并添加softmax分类层。 在此微调步骤中，我们采用倾斜的三角形学习率和逐渐解冻获得最佳结果（Howard and Ruder，2018）。 为了将BERT调整为标记数据，我们将与每个实例的[CLS]令牌相对应的隐藏状态馈送到softmax分类层。 我们使用AllenNLP对ELMO进行微调，并使用Pytorch预训练的BERT对BERT进行微调。

当没有可用的域外数据时，我们还尝试了仅对域内数据进行训练的ELMO作为高资源LM预训练方法的示例，例如Dai和Le（2015）。 具体来说，我们使用基于Transformer的ELMO生成上下文词表示。在下游分类期间，在等式（17）中求和之前，将所得的向量冻结并连接到随机初始化的单词向量。

**5结果**

在低资源设置中，我们发现VAMPIRE在我们考虑的所有低资源方法中都达到了最高的准确性，尤其是在标记数据量较小的情况下。 表2显示了随着标记数据量的变化，所有数据集上所有低资源模型的性能，为了便于比较，图2中也显示了其中的一部分。

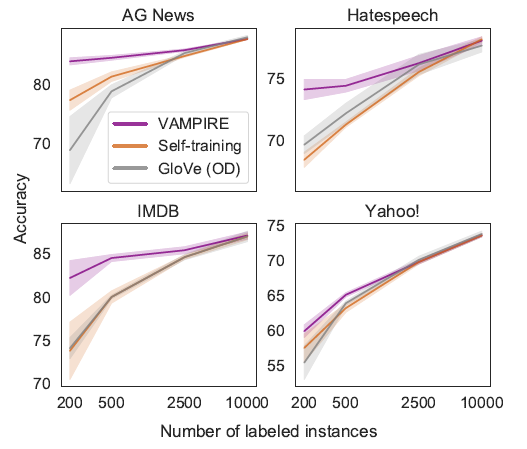


图2：低资源设置中所有数据集的学习曲线，显示了VAMPIRE，自训练和840B令牌GLOVE嵌入的五次运行的平均值（线）和一个标准差（带）。完整结果在表2中。

在高资源环境下，我们毫不奇怪地发现，将预训练的BERT模型微调到域内数据可提供最佳性能。 对于BERT和ELMO，我们发现使用冻结的现成矢量可得出结果，与微调到任务域相比，特别是对于HATESPEECH和IMDB而言，性能差得令人惊讶。对于这两个数据集，仅对域内数据进行训练的ELMO模型提供了比冻结的现成ELMO更好的性能（参见图3）。 但是这个差异较小，对于YAHOO和AG。 （有关完整结果，请参见附录B）。

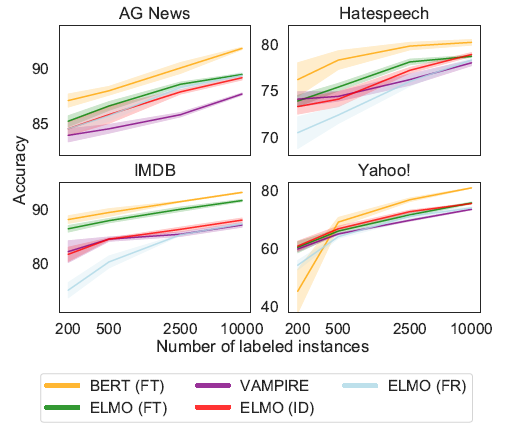


图3：四个数据集上的高资源方法（加上VAMPIRE）； ELMO的性能极大地受益于对域内数据进行（ID）或微调（FT）的训练（与BERT一样；附录B中的完整结果）。 密钥：FT（微调），FR（冻结），ID（域内）。

这些结果加在一起表明，尽管在大量Web文本上进行预培训相对于纯监督模型提供了很大的改进，但是访问未标记的域内数据对于在高资源环境中微调预培训的语言模型还是对在资源匮乏的环境中培训VAMPIRE都很重要。 Yogatama等人（2019）也报道了类似的发现处理自然语言推理和问题解答等任务。

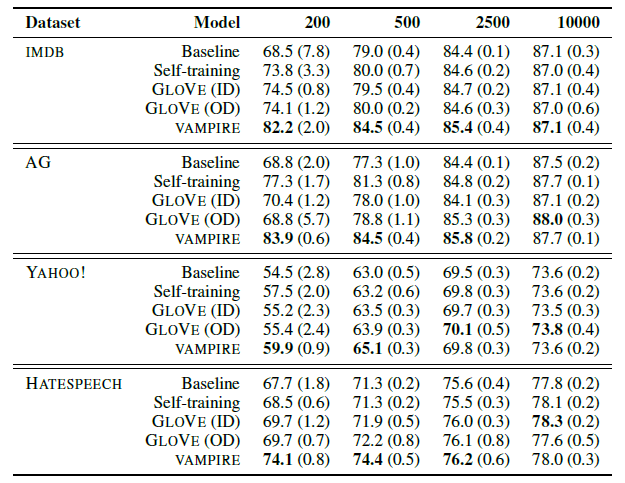


表2：在不同级别的标记训练数据（200、500、2500和10000文档）下，四个文本分类数据集在低资源设置中的测试准确性。 每个分数记录为五颗种子的平均值，括号中为标准差，每个设置中的最高平均结果以粗体显示。

**6分析**

**6.1 NPMI与NLL作为终止标准**

为了分析VAMPIRE中不同停止准则的有效性，我们在IMDB上预训练了200个VAMPIRE模型：在验证数据上通过NPMI选择了100个模型，并通过负对数似然（NLL）选择了100个模型。 有趣的是，我们观察到VAMPIRE NPMI和NLL值呈负相关（= –0.72；图4A），这表明在收敛之后，训练得更好的数据拟合模型也趋向于具有更连贯的主题。 然后，我们在IMDB数据集的固定200个文档随机子集中训练200个具有相同超参数的下游分类器，并在NPMI和NLL选择的VAMPIRE模型上统一采样作为附加功能。 在图4B和图4C中，我们观察到更好的预训练VAMPIRE模型（根据任一标准）往往会产生更好的下游性能。 （分别为NPMI和NLL的= 0.55和= –0.53）。

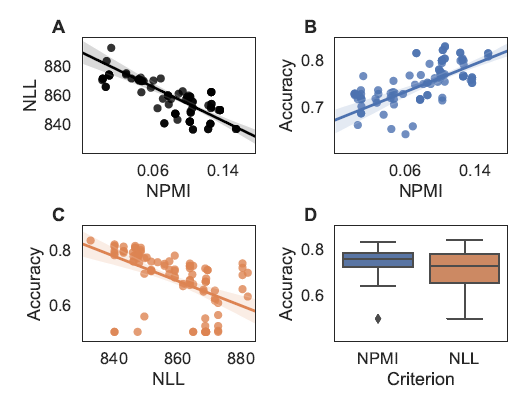


图4：比较NPMI和NLL作为选择VAMPIRE模型的早期停止标准。 NPMI和NLL是模型拟合的相关度量，但是NPMI选择的VAMPIRE模型在200个带有IMDB标签的文档中，对下级分类性能的差异较小。 在验证数据上报告准确性。 有关更多详细信息，请参见x6.1。

但是，我们还观察到使用NLL作为停止标准而获得的VAMPIRE模型之间在准确性上存在较高的差异（图4D）。 通过NLL选择的此类模型的主题连贯性和下游性能均较差。 因此，使用NPMI进行模型选择是首选方法，表2中的所有VAMPIRE结果均基于使用此准则选择的预训练模型。

丁等人（2018）的实验对这种行为提供了一些见解。 他们发现，在训练神经主题模型时，模型拟合和NPMI最初倾向于在每个时期得到改善。 但是，在某些时候，困惑不断改善，而NPMI开始下降，有时会急剧下降。

在训练VAMPIRE时，我们也会观察到这种现象（请参阅附录C）。 如我们建议的那样，使用NPMI作为停止标准有助于避免因训练时间过长而导致的退化模型。

在一些初步实验中，我们还观察到由于主题冗余而导致NPMI虚高的情况。 Srivastava和Sutton也指出，将batchnorm应用于重构可以显着提高主题间并置单词的多样性。未来的工作可能会探索为NPMI指标分配一个单词多样性调整器，以鼓励跨主题具有更强一致性和单词多样性的模型。

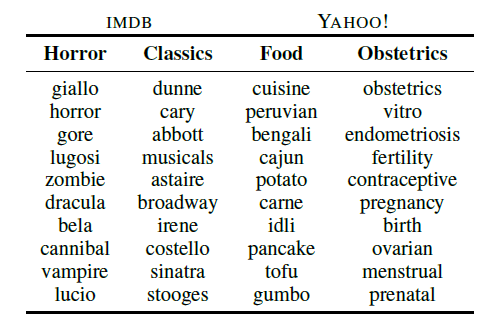


表3：VAMPIRE在IMDB和YAHOO中学习的示例主题数据集。 有关更多示例，请参见附录D。

**6.2学习的潜在主题**

除了轻量级之外，VAMPIRE的一个优点是它可以生成可以按主题明确解释的文档表示形式。 尽管我们输入到下游分类器中的输入将这种表示形式与编码器的内部状态结合在一起，但是主题解释有助于总结预训练的知识。 表3和附录D中提供了VAMPIRE学习的主题示例。

**6.3学习的标量层重量**

由于ri中的标量重量参数是可训练的，因此我们能够研究分类器倾向于选择预训练的VAE的哪一层。 我们一致发现，该模型倾向于增加VAE编码器的第一层h（1）和，并降低编码器的其他层的重量。 为了改善学习效果，特别是在资源设置较低的情况下，我们初始化了应用于第一个编码器层并具有较高值的​​标量权重，并降低了中间层的权重，从而提高了验证性能。 但是，我们还观察到在VAMPIRE中使用多层编码器会导致较大的下游增益。

**6.4计算要求**

VAMPIRE的一个吸引人的方面是它的紧凑性。表4显示了在单个GPU或CPU上训练VAMPIRE所涉及的计算要求，与从头开始在GPU相同数据上训练ELMO模型相比。 即使没有昂贵的硬件，也可以比ELMO更快地训练VAMPIRE数量级，使其特别适合在资源有限的情况下获得快速结果。

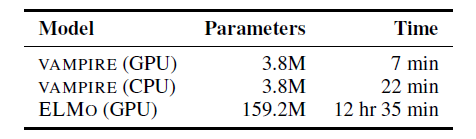


表4：VAMPIRE比基于Transformer的ELMO紧凑得多，但在资源匮乏的情况下仍具有竞争力。 在这里，我们显示了对IMDB数据集中的域内未标记文本预训练VAMPIRE和ELMO的计算要求。 我们报告了在GeForce GTX 1080 Ti GPU上训练VAMPIRE（具有附录A.1中列出的超参数）和ELMO（具有默认配置）以及在2.60GHz Intel Xeon CPU上训练VAMPIRE的结果。 VAMPIRE在GPU上使用大约750MB的内存，而ELMO需要大约8.5GB。

**7相关工作**

除全文中提供的参考文献外，许多其他文献还探索了在处理数量有限的带标签数据时提高性能的方法。 语音识别的早期工作证明了在半监督环境中对深度模型进行预训练和微调的重要性（Yu等，2010）。 Chang等。 （2008年）考虑了“无数据”分类，其中类别名称提供了唯一的监督。Miyato等。 （2016年）表明，对抗性预训练可以带来很大的收获，有效地增加了可用的数据量。 主动学习中的一长串工作类似地尝试在获得标签的成本很高时最大化性能（Settles，2012）。 谢等。 （2019）描述了利用反向翻译和TF-IDF单词替换的新型数据增强技术。 所有这些方法都可以与本文提出的方法有效地结合。

**8建议**

根据我们在本文中的发现，我们为希望进行有效的半监督文本分类的人提供以下实用建议。

当资源无限时，当前可以通过使用诸如BERT之类的预训练模型来获得最佳结果，但是对域内数据进行微调至关重要（另请参见Howard和Ruder，2018年）。

当计算资源和注释受到限制，但是有大量未标记的数据时，与其他低资源方法相比，VAMPIRE可获得更大的收益。

仅在域内数据上训练诸如ELMO之类的语言模型可以提供与VAMPIRE相当的性能，或者在某种程度上要好于VAMPIRE，但是除非使用GPU，否则其代价可能是过高的。

另外，可以投入资源来获得更多注释。 如果具有足够的标记数据（成千上万个实例），则其他未标记数据所提供的优势将变得微不足道。 当然，其他NLP任务可能涉及数据，速度和准确性之间的不同权衡。

**9结论**

ELMO和BERT等模型的出现使半监督NLP复活了，这表明在大量数据上对大型模型进行预训练可以提供对大量NLP任务有益的表示形式。 在本文中，我们确认当标记实例的数量较少时，这些模型对于文本分类很有用，但证明对域内数据进行微调也至关重要。 在由于计算限制或由于不存在相关语言的合适的预训练模型而无法轻松使用BERT的环境中，VAMPIRE提供了一种竞争性的轻量级替代方案，可用于在资源较少的情况下从未标记的数据进行预训练。 当使用有限数量的标记数据时，我们可以达到优于基线的性能，例如自我训练，或使用对域外数据进行预训练的词向量，并可以达到仅对域内数据进行了训练的ELMO的性能只是计算成本的一小部分。