MVAE: 用于检测假新闻的多模态变分自动编码器

Dhruv Khattar
International Institute of Information Technology
Hyderabad, India
dhruv.khattar@research.iiit.ac.in

Jaipal Singh Goud
International Institute of Information Technology
Hyderabad, India
jaipal.singh@research.iiit.ac.in

Manish Gupta[*]
International Institute of Information Technology
Hyderabad, India
manish.gupta@iiit.ac.in

Vasudeva Varma
International Institute of Information Technology
Hyderabad, India
vv@iiit.ac.in

摘要

近年来，假新闻和错误信息对我们的生活产生了破坏性的不利影响。由于微博网络作为大多数人的主要的新闻来源，现在假新闻的传播速度比以前更快，影响也更深远。这使得对假新闻的检测成为一个极其重要的挑战。假新闻文章，就像普通的新闻文章一样，利用多媒体内容来操纵用户的意见和传播错误信息。目前检测假新闻的方法的一个缺点是它们无法学习多模态（文本+视觉）信息的共享表示。我们提出了一个端到端的网络，多模态变分编码器（MVAE），它使用一个双模态变分自动编码器和一个二进制分类器来完成假新闻检测的任务。该模型由三个主要部分组成：一个编码器、一个解码器和一个假新闻检测器模块。变分自动编码器能够通过优化观察数据的边际可能性来学习概率潜变量模型。然后，假新闻检测器利用从双模态变分自动编码器获得的多模态表征，将帖子分类为假新闻或非假新闻。我们在两个标准的假新闻数据集上进行了广泛的实验，这些数据集是从流行的微博网站上收集的，包括微博和Twitter。实验结果表明，在这两个数据集中，我们的模型在平均准确率和F1分数上都超过了最先进的方法，幅度高达6%和5%。

CCS 概念

• **信息系统→社交网络；多媒体和多模态检索；**

• **计算方法学→神经网络。**

[*]作者也是微软的首席应用研究员。(gmanish@microsoft.com)

WWW'19，2019年5月13-17日，美国加利福尼亚州旧金山市。

ACM ISBN 978-1-4503-6674-8/19/05.
https://doi.org/10.1145/3308558.3313552

关键词

虚假新闻检测、多模态融合、变分自动编码器、微型博客

ACM 参考格式:
Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. 2019. MVAE: Multimodal Variational Autoencoder for Fake News Detection. In *Proceedings of the 2019 World Wide Web Conference (WWW '19), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3308558.3313552

1 介绍

近来，新闻消费方式的改变使假新闻和错误信息的问题成为讨论的焦点。由于每天都有数以千计的新新闻文章在社交媒体网络上涌现，而每篇文章都没有可信度或验证检查，一个由错误信息（无意中分享的虚假信息）和虚假信息（故意创造和分享已知的虚假信息）助长的生态系统已经建立。社会媒体网络已经使新闻文章从传统的纯文字新闻发展到带有图片和视频的新闻，这可以提供更好的故事讲述体验，并有能力吸引更多的读者。最近的假新闻文章正是利用了这种视觉背景辅助新闻的变化趋势。现在的假新闻文章可以通过包含错误的、不相关的和伪造的图片来误导读者。点对点网络（主要是社交媒体网络）允许假新闻（宣传）针对那些更有可能接收和分享特定信息的用户。近年来，假新闻因对公共事件产生广泛的负面影响而备受关注。实现的一个重要转折点是2016年的美国总统选举。据认为，在选举前的最后三个月里，有利于两位被提名人中任何一位的假新闻在Facebook上被接收和分享了3700多万次[1]。这使得检测假新闻的任务成为一项关键任务。

图 1 显示了Twitter数据集中的三个假新闻的例子。每条推文都有一定的文字内容和一个与之相关的图片。对于左边的推文，图片和文字都表明它可能是一条假新闻。在右边的推文中，图片并没有增加实质性的信息，但文字表明它可能是假新闻。在中间的推文中，很难从文本中得出结论，但被处理过的图像表明它可能是假新闻，这个例子反映了这样的假设。

*文本：*大象从坚硬的岩石上雕刻出来。壮观!

文本： 在阿肯色州发现的新鱼类物种。

文本： 36岁女子生下14个不同父亲的14个孩子

图1：Twitter数据集中的假新闻例子

视觉和文字信息对可以给假新闻检测带来更好的帮助。

在概念层面上，检测假新闻的任务已经标注了包括错误信息和谣言的各种标签。我们研究的目标是检测那些捏造的、可以被证实为虚假的新闻内容。谣言和假新闻检测技术包括从传统学习方法到深度学习模型在内的多种方法。最初的方法[4] 试图仅使用从新闻故事的文本内容中提取的语言学特征来检测假新闻。Ma等人[14] 探讨了通过捕捉时间-语言特征用深度神经网络表示推文的可能性，随后Chen等人 [5] 在此基础上，将注意力机制引入RNNs。

最近的工作[26]在深度学习检测假新闻领域的研究表明，由于其提取相关特征的能力增强，性能比传统方法有所提高。Jin等人[9] 结合了视觉、文本和社会背景特征，使用注意力机制对假新闻进行预测。Wang等人[26]使用一个额外的事件判别器来学习所有事件中的共同特征，目的是消除不可转移的特定事件特征，并声称他们可以更好地处理新的和新出现的事件。现有模型的一个缺点是，它们没有任何明确的目标函数来发现*跨模态的相关性*。

我们的工作，受自动编码器的启发，试图在多模态环境中学习共享表征。Kingma等人 [12]提出潜在变量模型是生成复杂分布模型的有力方法，并提出了变分自动编码器（VAE）的想法。VAEs可以通过优化观测数据的边际似然性来学习概率性潜变量模型。我们克服了当前模型的局限性，提出了一个能够学习共享（视觉+文本）表征的多模态变分自动编码器，经过训练可以发现推文中各模态的相关性。然后，将VAE与一个分类器相结合以检测假新闻。

总而言之，我们的工作贡献如下:

- 我们提出了一种新颖的方法，使用帖子的内容，即文字和所附图片，对社交媒体帖子进行分类。
- 所提出的MVAE模型使用多模态变分自动编码器与假新闻检测器联合训练来检测一个帖子是否是假的。
- 我们在两个真实世界的数据集上广泛地评估了我们模型的性能。结果显示，我们提出的该模型可以学习到更好的多模态特征，并优于最先进的多模态假新闻检测模型。
- 实验表明，我们提出的模型能够发现不同模态之间的关联，从而得出更好的多模态共享表征。

本文的其余部分组织如下。在第2节，我们讨论了以前关于假新闻检测的工作和自动编码器的基础知识。在3这一节中，我们提出了我们的模型和它的各种组成部分。在4这一节中，我们描述了数据集、基线方法并提供了我们所提出的模型的实施细节。结果和分析在第5节展示，最后，我们在第6节一个简短的总结结尾。

2 相关工作

假新闻检测的任务与其他各种相互影响的挑战相似，从垃圾邮件检测[29]到谣言检测[24]到讽刺小说检测[20]。由于每个人都可能对这些相关概念有自己的直观定义，所以每篇论文都有自己的定义。按照以前的工作[21,22]，我们明确指出，我们研究的目标是检测捏造的、可以被验证为虚假的新闻内容。

一些早期的研究试图根据从新闻故事文本中提取的语言学特征来检测假新闻。Castillo等人[4] 采用了一组语言学特征，如特殊字符、感性的正、反义词、表情符号等来检测假新闻。Popat等人[19] 使用立场和语言风格特征，如as- sertive动词、话语标记等，来评估一个说法的可信度。在Feng等人的研究中，无语境语法规则被用来识别欺骗行为。[6]Ma等人[14] 率先探索了通过捕捉时间语言学特征用深度神经网络来表示推文的可能性。Chen等人[5] 将注意力机制纳入了循环神经网络（RNNs），以汇集具有特定焦点的独特时间语言特征。

由用户和推文之间的互动而建立的社会联系，为帖子诞生了丰富的社会背景。Wu等人[28] 推断出社交媒体用户资料的嵌入，并利用传播途径上的LSTM网络对假新闻进行分类。Liu等人[13] 将新闻故事的传播路径建模为多变量时间序列，并通过RNN和CNN的组合进行传播路径分类来检测假新闻。Ma等人[15] 使用基于树状结构神经网络的递归神经模型学习推文的表征。Jin等人[8] 使用手工制作的社会环境特征，如关注者的数量、转发量等。大多数的社会环境特征都是不成熟的，通常需要大量的手工劳动来收集。此外，它们不能为新出现的事件提供足够的信息。

最近的研究表明，视觉特征（图像）在检测假新闻方面起着非常重要的作用[27]然而，验证社交媒体上的多媒体内容的可信度得到的审查较少。对所附图像的基本特征的提取已经在[18]中进行了探讨。然而，这些特征仍然是手工制作的，很难代表图像内容的复杂分布。

深度神经网络在学习图像和文本表征方面取得了巨大的成功。它们已被成功地应用于各种任务，包括图像说明[10, 25]，视觉问题回答[2]，以及假新闻检测[8, 26]。

Jin等人[8]提出了一个提取视觉、文本和社会环境特征的模型，并通过注意力机制将其融合。Wang等人[26] 使用对抗网络和多模态特征提取器来学习事件多样特征。然而，这两个模型都没有明确的目标来发现不同*模态之间的相关性*。

为了克服现有多模态假新闻检测器的局限性，我们提出了一个多模态变分自动编码器（MVAE），它可以学习两种模态、文本和图像的共享表示。多模态变分自动编码器被训练为从所学的共享表示中重建两种模态，从而发现*跨模态的相关性*。我们将多模态变分自动编码器和一个分类器一起训练，以检测假新闻。此外，我们使用比基线方法更少的信息来检测假新闻，即我们不使用任何社会或事件相关信息。

3 提出的MVAE模型

3.1 MVAE 概述

我们提出了一种新型的深度多模态变分自动编码器（MVAE）来解决假新闻检测问题。MVAE的基本思想是学习一条推文内容的两种模态的统一表示。提出的MVAE的整体架构如2图所示 。它有三个主要组成部分。

- 编码器：它将来自文本和图像的信息编码为一个潜在的向量。
- 解码器：它从潜向量中重建原始图像和文本。
- 假新闻检测器。它使用学习到的共享表示（潜向量）来预测一条新闻是否是假的。

3.2 编码器

编码器的输入是帖子的文本和它所附的图片，它输出的是一个从两种模态中学习到的特征的共同表示。编码器可以被分解成两个子组件：文本编码器和视觉编码器。

3.2.1 文本编码器 文本编码器的输入是帖子中单词的顺序列表，$T = [T1\ T2 \dots Tn]$，其中n是文本中的单词数。文本中的每个单词 $T_i \in T$ 被表示为一个单词嵌入向量。每个词的嵌入向量都是通过一个深度网络获得的，该网络以无监督的方式对给定的数据集进行了预训练。

为了从文本内容中提取特征，我们使用带有长短时记忆（LSTM）单元的循环神经网络（RNNs）。RNNs是一类人工神经网络，它利用顺序信息，通过中间层保持历史。一个普通的RNN有一个内部状态，它在每个时间步的输出可以用前一个时间步来表示。然而，已经发现基本RNN存在梯度消失和爆炸的问题[7,17]。这导致模型在相隔几步的词之间学习低效的依赖关系。为了克服这个问题，LSTM扩展了基本的RNN，通过使用存储单元和有效的门控机制，在很长的时间段内存储信息。

让[$h1\ h2 \dots hn$]代表LSTM的状态，其状态更新满足以下公式。

$$\left[f_t,\ i_t,\ o_t\right] = \sigma\left(Wh_{t-1} + Ux_t + b\right) \tag{1}$$

$$c_t = f_t \circ c_{t-1} + i_t \circ \tanh\left(Wh_{t-1} + Ux_t + b\right) \tag{2}$$

$$h_t = o_t \circ \tanh\left(c_t\right) \tag{3}$$

其中，σ是Logistic sigmoid函数，f_t, i_t, o_t分别代表遗忘、输入和输出门，x_t表示输入，，h_t表示时间t的隐藏状态。遗忘门、输入门和输出门控制着整个序列的信息流。W, U 是代表权重的矩阵，b是与连接相关的偏差。

$$R_T = \phi\left(W_{tf} R_{lstm}\right) \tag{4}$$

我们采用堆叠的双向LSTM单元来提取文本特征。LSTM的最终隐藏状态是由前向和后向的状态连接起来得到的。最后，我们将LSTM的输出通过一个全连接层（enc text fc）来获得文本特征。

其中，R_{lstm} 是 LSTM的输出，W_{tf} 是全连接层的权重矩阵，ϕ 是使用的激活函数。

3.2.2 视觉编码器 视觉编码器的输入是与帖子相连的图像V。从各种视觉理解问题中可以看出，使用卷积神经网络（CNN）在大量数据上训练的图像描述符已经被证明是非常有效的。在网络的后几层中对空间布局和物体语义的隐性学习促成了这些特征的成功。

我们采用预先训练好的VGG-19结构的网络[23] 在ImageNet数据库中训练，并使用全连接层的输出（FC7）。在联合训练过程中，我们冻结了VGG网络的参数以避免参数爆炸。最后，我们将VGG的输出通过多个全连接层（enc vis fc*），以获得与文本相同大小的图像表示。

$$R_V = \phi\left(W_{vf} R_{vgg}\right) \tag{5}$$

其中，R_{vgg} 是由VGG-19得到的特征表示，W_{vf} 是全连接层的权重矩阵，ϕ 是使用的激活函数。

然后，文本特征表示法和视觉特征表示法被串联起来，并通过一个全连接层形成共享表示。然后，我们从共享表征中得到两个向量μ和σ。它们可以分别被视为共享表征分布的平均值和方差。此外，从先前的分布（例如高斯分布）中抽出一个随机变量ϵ。最终的重新参数化的多模态表示用$\mathrm{R_m}$ *表示*，可以计算如下：

$$R_m = \mu + \sigma \circ \epsilon \tag{6}$$

我们把编码器表示为 $G_{enc}(M, \theta_{enc})$，其中θ_{enc} 表示编码器中要学习的所有参数，M表示多媒体帖子的集合。因此，对于一个多媒体帖子M，编码器的输出是多模态表示。

$$R_m = G_{enc}\left(m,\ \theta_{enc}\right) \tag{7}$$

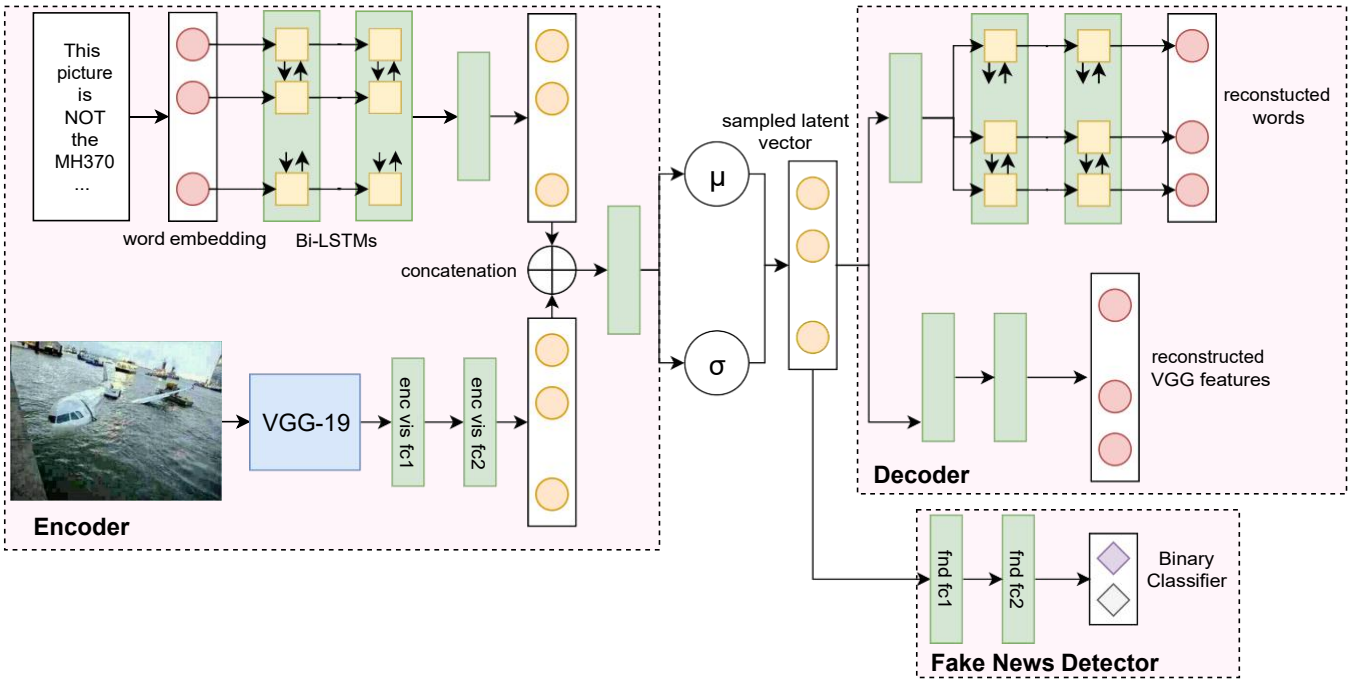

图2：提出的多模态变分自动编码器（MVAE）的网络结构。它有三个组成部分。编码器、解码器和假新闻检测器。

3.3 Decoder

解码器的结构与编码器的结构相似，但是是倒置的。解码器的目标是从采样的多模态表示中重建数据。就像编码器一样，它可以被分解成两个子组件：文本解码器和视觉解码器。

3.3.1 *文本编码器*

文本编码器将多模态表示作为输入，然后，我们有将他通过全连接层后得到一个双向LSTM网络的输入。之后我们使用一个类似解码器中的堆叠的双向LSTM单元，将输出结果经过一个由softmax激活的时间分布式的全连接层，得到每个词在该时间步中的概率。

3.3.2 *图像解码器*

视觉解码器的目标是重建图像特征，从多模态表示中构建VGG-19特征。视觉解码器子网络与视觉编码器的相应子网络正好相反。多模态表示通过多个全连接层（dec vis fc*）来重建VGG-19特征。

$$\left(\hat{t}_m, \hat{r}_{vgg_m}\right) = G_{dec}\left(R_m,\ \theta_{dec}\right) \tag{8}$$

我们把解码器表示为G。其中包含解码器中的各种参数。因此，多个多媒体帖子的解码器的输出是每个词的概率矩阵，对于文本中的每个位置，重建为图像的VGG-19特征。

VAE模型是通过优化重构损失和KL发散损失之和来训练的。因此，我们采用分类交叉熵损失来重建文本，采用平均平方误差来重建图像特征。两个概率分布之间的KL发散只是衡量它们相互之间的发散程度。最小化KL发散意味着优化概率分布参数（μ和σ），使其与目标分布（正态分布）的参数接近。这些参数可以按以下方式计算。

$$\mathcal{L}_{rec_{vgg}} = \mathbb{E}_{m\sim M}\left[\frac{1}{n_v}\sum_{i=1}^{n_v}\left(\hat{r}_{vgg_m}^{(i)} - r_{vgg_m}^{(i)}\right)^2\right] \tag{9}$$

$$\mathcal{L}_{rec_t} = -\mathbb{E}_{m\sim M}\left[\sum_{i=1}^{n_t}\sum_{c=1}^{C}\mathbf{1}_{c=t_m^{(i)}}\log \hat{t}_m^{(i)}\right] \tag{10}$$

$$\mathcal{L}_{kl} = \frac{1}{2}\sum_{i=1}^{n_m}\left(\mu_i^2 + \sigma_i^2 - \log(\sigma_i) - 1\right) \tag{11}$$

其中，M是多媒体帖子的集合，n_v是VGG-19特征的维度，n_t是文本中的单词数，n_m是多模态特征的维度，C是词汇的数量量。我们通过寻求最佳的参数$\hat{\theta}_{enc}$和C是使VAE损失最小。$\hat{\theta}_{dec}$可以表示为如下。

$$\left(\theta_{enc}^*, \theta_{dec}^*\right) = \underset{\theta_{enc},\theta_{dec}}{\operatorname{argmin}}\left(\mathcal{L}_{rec_{vgg}} + \mathcal{L}_{rec_t} + \mathcal{L}_{kl}\right) \tag{12}$$

3.4 假新闻检测器

假新闻检测器将多模态表示作为输入，旨在将帖子分类为假的或真实的。它由多个具有相应激活函数的全连接层组成。我们

将假新闻检测器表示为$G_{fnd}(R_m, \theta_{fnd})$，其中$\theta_{fnd}$表示假新闻检测器的所有参数。对于一个多媒体帖子，$m$，假新闻检测器的输出是这个帖子是假新闻的概率。

$$\hat{y}_m = G_{fnd}\left(R_m, \theta_{fnd}\right) \tag{13}$$

我们可以把$\hat{y}_m$的值看作是一个标签，1意味着多媒体帖子m是假的，否则就是0。为了约束这些值在0和1之间，我们使用sigmoid logistic函数。因此，为了计算分类损失，我们采用交叉熵，具体如下。

$$\mathcal{L}_{fnd}\left(\theta_{enc}, \theta_{fnd}\right) = -\mathbb{E}_{(m,y)\sim(M,Y)}\left[y \log\left(\hat{y}_m\right) + \left(1-y\right)\log(1-\hat{y}_m)\right] \tag{14}$$

其中，M代表多媒体帖子的集合，Y代表实际真实标签的集合。我们最小化分类损失

通过寻求最佳参数$\hat{\theta}_{fnd}$和$\hat{\theta}_{enc}$，这可以被表示为如下的公式

$$\left(\theta^*_{enc}, \theta^*_{fnd}\right) = \underset{\theta_{enc}, \theta_{fnd}}{\operatorname{argmin}} \mathcal{L}_{fnd} \tag{15}$$

3.5 综合模型

提出的MVAE模型的完整架构如图2所示。编码器的输出被送入解码器以及假新闻检测器。解码器的目的是重建数据，而假新闻检测器的目的是将帖子分类为假新闻与否。我们联合训练VAE和假新闻检测器。因此，最终的损失可以写成如下。

$$\mathcal{L}_{final}\left(\theta_{enc}, \theta_{dec}, \theta_{fnd}\right) = \lambda_v \mathcal{L}_{rec_{vgg}} + \lambda_t \mathcal{L}_{rec_t} + \lambda_k \mathcal{L}_{kl} + \lambda_f \mathcal{L}_{fnd} \tag{16}$$

其中，λs可以用来平衡损失函数的各个项。在本文中，我们简单地将所有的λs设置为1，然后可以通过最小化最终损失来计算出最佳参数，具体右侧所示。

4 实验

在本节中，我们首先描述了实验中使用的两个社交媒体数据集。然后，我们简要讨论了最先进的假新闻检测方法，以及一些最先进的语言视觉模型。

4.1 数据集

鉴于结构化多媒体数据的稀缺性，我们利用两个标准数据集来评估我们的假新闻检测架构。这两个数据集包括从Twitter和微博收集的真实社会媒体信息。据我们所知，这两个数据集是唯一有图像和文本信息配对的可用数据集。

4.1.1 推特数据集 作为MediaEval的一部分[3]的一部分，Twitter数据集被发布用于验证多媒体使用任务。该任务的目的是检测社交媒体上的虚假多媒体内容。该数据集由推特（在推特上发布的短消息）组成，每条推特都有文本内容、图像、视频和社会背景信息。

表1：MVAE与其他方法在两个不同的数据集上的性能对比

数据集	方法	准确度	虚假新闻			真实新闻		
			精度	召回	F_1	精度	召回	F_1
推特	文本	0.526	0.586	0.553	0.569	0.469	0.526	0.496
	视觉	0.596	0.695	0.518	0.593	0.524	0.7	0.599
	VQA	0.631	0.765	0.509	0.611	0.55	0.794	0.65
	Neural-talk	0.610	0.728	0.504	0.595	0.534	0.752	0.625
	att-RNN	0.664	0.749	0.615	0.676	0.589	0.728	0.651
	EANN	0.648	**0.810**	0.498	0.617	0.584	0.759	0.660
	MVAE	**0.745**	0.801	**0.719**	**0.758**	**0.689**	**0.777**	**0.730**
微博	文本	0.643	0.662	0.578	0.617	0.609	0.685	0.647
	视觉	0.608	0.610	0.605	0.607	0.607	0.611	0.609
	VQA	0.736	0.797	0.634	0.706	0.695	0.838	0.760
	Neural-talk	0.726	0.794	0.713	0.692	0.684	0.840	0.754
	att-RNN	0.772	0.854	0.656	0.742	0.72	**0.889**	0.795
	EANN	0.782	0.827	0.697	0.756	0.752	0.863	0.804
	MVAE	**0.824**	**0.854**	**0.769**	**0.809**	**0.802**	0.875	**0.837**

与它相关的。该数据集有大约17000条独特的推文，跨越了不同的事件。该数据集被分成两部分：训练集（9000条假新闻推文，6000条真新闻推文）和测试集（2000条推文）。它们是以这样的方式分割的，推文没有重叠的事件。考虑到我们对图像和文本信息的关注，我们过滤掉了所有附有视频的推文。我们使用开发集进行训练，使用测试集进行测试，以保持与基线方法相同的数据分割方案。

4.1.2 微博数据集 微博数据集，在[8]中用于检测假新闻的数据集包括从新华社（中国的一个权威新闻来源）和微博（中国的一个微博网站）收集的数据。

$$\left(\theta^*_{enc}, \theta^*_{dec}, \theta^*_{fnd}\right) = \underset{\theta_{enc}, \theta_{dec}, \theta_{fnd}}{\operatorname{argmin}} \mathcal{L}_{final}\left(\theta_{enc}, \theta_{dec}, \theta_{fnd}\right) \tag{17}$$

这些数据是在2012年5月至2016年1月的时间跨度中获得的，并由微博的官方辟谣系统验证。该系统鼓励普通用户在微博上报告可疑的推文，然后由知名用户组成的委员会进行审查。将可疑的推文分类为假的或真的。根据以前的工作[14,27]，这个系统也作为收集谣言新闻的权威来源。非谣言推文是由新华社核实的推文。数据集的预处理采用与[8]类似的方法。初步步骤包括去除重复的图像（使用位置敏感散列）和低质量的图像，以确保整个数据集的同质性。然后，数据集被分成训练集和测试集，与Jin等人的方法一样，推特比例约为4：1。

4.2 实验设置

对于词的嵌入，我们采用分布式的Word2Vec rep-resentation for words [16]。对于Twitter数据集，有一些不是英文的帖子，所以我们把它们翻译成英文以保持数据的一致性。我们对Twitter数据集进行了标准的文本预处理。对于微博数据集，文本全是中文的。中文文本的书写在单词之间没有空格。我们使用斯坦福单词分割器将中文文本标记为单词。

我们以无监督的方式在训练数据集上对Word2Vec模型进行预训练，维度大小为32。

对于视觉特征，我们使用在ImageNet集上预先训练的19层VGGNet的第二至最后一层的输出[23]。从VGG-19得到的特征尺寸为4096。我们不对VGG的权重进行微调，也就是说，VGG网络的权重是冻结的。

文本编码器由LSTM组成，隐藏层的尺寸为32。文本编码器中使用的全连接层的尺寸为32。视觉编码器由两个大小为1024和32的全连接层组成。解码器由与编码器尺寸相同的层组成。假新闻检测器有两个大小为64和32的全连接层。

在整个网络的训练中，我们使用了128个实例的批次大小。该模型被训练了300个epochs，学习率为10^{-5}，采取早停训练策略得到结果。我们使用双曲正切函数作为非线性激活函数。为了防止过拟合，我们对模型的权重使用了L2-正则化。我们试验了[0, 0.05, 0.1, 03, 0.5]的权重惩罚，并对编码器和解码器设置为0.05，对假新闻检测器设置为0.3。为了寻求我们模型的最佳参数，我们使用Adam [11] 作为优化器。我们公开了该代码[1]。

4.3 基线

为了验证所提出的多模态变分自动编码器的性能，我们将其与两类基线模型进行比较：单模态模型和多模态模型。

4.3.1 单模态的模型 MVAE利用视觉和文本数据的信息来识别潜在的假新闻。针对这样的多模态方法，我们还实验了两个单模态模型，如下文所述。

- **文本**: 该模型仅使用帖子中的文本信息，将其分类为假的或不是。每个词都被看作是一个32维的向量。词嵌入是根据帖子的文本内容进行训练的。然后，单个帖子被送入Bi-LSTM以提取文本特征。然后，文本特征被送入一个32维的全连接层，该层有一个与Bi-LSTM连接的softmax函数，负责进行最终预测。
- **视觉:**视觉模型只使用帖子中的图像来分类它们是否是假的。图像被送入一个预先训练好的VGG-19网络，该网络有一个全连接层来提取视觉特征。与文本模型类似，视觉特征随后被送入一个32维的全连接层进行预测。

4.3.2 多模态模型 多模态方法利用多模态的信息来完成假新闻分类的任务。

- **VQA** [2]:视觉问题回答的目的是回答问题。
  关于给定图像的意见。我们调整了视觉模型保证质量
  将原来为多级分类任务设计的数据转为我们的二进制分类任务。这是通过用二元分类层替换最后的多类层来实现的。我们使用一个单层的LSTM，隐藏单元的数量设置为32。
- **Neural Talk** [25]: Vinyals等人[25]在图像标题领域的工作，提出使用深度递归框架生成描述图像的自然语言句子。按照与他们类似的结构，我们通过在每个时间点平均RNN的输出来获得潜在的表征，作为推文中图像和文字的联合表征。然后，这些表征被送入一个全连接层，然后是熵损失层，以进行预测。
- **att-RNN** [8]: att-RNN使用注意力机制来结合文本、视觉和社会背景特征。在这个端到端终端网络，图像特征被纳入文本和社会背景的联合表示中，使用LSTM网络获得。来自LSTM输出的神经注意力机制是融合视觉特征的一个组成部分。为了进行公平的比较，在我们的实验中，我们删除了处理社会背景信息的部分。
- **EANN** [26]: 事件对抗性神经网络（EANN）
  由三个主要部分组成：多模态特征提取器、假新闻检测器和事件判别器。多模态特征提取器从帖子中提取文本和视觉特征。它与假新闻检测器一起工作，学习用于检测假新闻的鉴别性表示。事件判别器负责去除任何特定事件的特征。只用两个组件，即多模态特征提取器和假新闻检测器，也可以检测假新闻。因此，为了进行公平的比较，在我们的实验中，我们使用的是不包括事件判别器的EANN变体。

请注意，与[13,15]中提出的方法进行公平的比较是不可能的，因为他们使用的是额外的信息，如推特的宣传数据。另外，这些方法并不是为多模态数据集开发的。我们在表中报告了所有基线方法和我们提出的模型的准确率、召回率、精确度和 F_1 得分。

5 结果和分析

表1中 显示了基线方法和我们提出的方法在两个数据集上的结果。我们报告了我们的假新闻检测器的准确性，以及我们的方法对假新闻和真新闻的精确度、召回率和F1得分。我们可以清楚地看到，我们提出的方法比基线方法的表现要好得多。

在Twitter数据集上，在单一模态的模型中，视觉模型比文本模型表现得更好。这可能是由于在VGG-19的帮助下学习的图像特征与文本特征相比，有更多的可共享的模态来对新闻进行分类。尽管视觉特征比文本特征表现更好，但单一模态模型比多模态模型表现更差。

在多模态模型中，att-RNN击败了EANN，这告诉我们，注意力机制可以通过考虑图像中与文本相关的部分来帮助提高模型的性能。我们提出的模型MVAE在很大程度上胜过了基线模型，并将准确率从66.4%提升到了 74.5%，并将F1得分从66%提高到了 73%。

在微博数据集上，我们看到结果有类似的趋势。在单模态模型结果中，文本模型战胜了视觉模型。

[1] https://github.com/dhruvkhattar/MVAE

在多模态方法中，EANN和att-RNN在这项任务中的表现比Neural Talk和VQA更好。MVAE的表现优于所有的基线方法，在准确率方面从78.2%提高到82.4%，与之前的最佳基线方法相比，F1分数增加了5%。这验证了我们提出的MVAE方法在检测社交媒体上假新闻方面的有效性

6 总结

在这项工作中，我们探讨了多模态假新闻检测的任务。我们克服了当前模型的局限性，解决了在推文中学习不同模态之间的相关性的挑战，并为此提出了一个多模态变分自动编码器，学习共享（视觉+文本）表征来帮助假新闻的检测。我们的模型由三个主要部分组成，一个编码器、一个解码器和一个假新闻检测器。我们提出的模型，MVAE是通过联合学习编码器、解码器和假新闻检测器来训练的。我们提出的架构的性能在两个真实世界的数据集上进行了评估，MVAE模型的性能优于目前最先进的架构。在未来，我们计划使用推特传播数据和用户特征来扩展MVAE。

引用

[1] Hunt Allcott and Matthew Gentzkow. 2017. Social Media and Fake News in the 2016 Election. 31 (05 2017), 211–236.

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*. 2425–2433.

[3] Christina Boididou, Symeon Papadopoulos, Duc-Tien Dang-Nguyen, Giulia Boato, Michael Riegler, Stuart E Middleton, Andreas Petlund, and Yiannis Kompatsiaris. [n. d.]. Verifying Multimedia Use at MediaEval 2016. ([n. d.]).

[4] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*. ACM, 675–684.

[5] Tong Chen, Lin Wu, Xue Li, Jun Zhang, Hongzhi Yin, and Yang Wang. 2017. Call Attention to Rumors: Deep Attention Based Recurrent Neural Networks for Early Rumor Detection. *arXiv preprint arXiv:1704.05973* (2017).

[6] Song Feng, Ritwik Banerjee, and Yejin Choi. 2012. Syntactic stylometry for deception detection. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*. Association for Computational Linguistics, 171–175.

[7] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[8] Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, and Jiebo Luo. 2017. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the 2017 ACM on Multimedia Conference*. ACM, 795–816.

[9] Zhiwei Jin, Juan Cao, Yongdong Zhang, Jianshe Zhou, and Qi Tian. 2017. Novel visual and statistical image features for microblogs news verification. *IEEE transactions on multimedia* 19, 3 (2017), 598–608.

[10] Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3128–3137.

[11] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[12] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

[13] Yang Liu and Yi fang Brook Wu. 2018. Early Detection of Fake News on Social Media Through Propagation Path Classification with Recurrent and Convolutional Networks. In *AAAI*.

[14] Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J Jansen, Kam-Fai Wong, and Meeyoung Cha. 2016. Detecting Rumors from Microblogs with Recurrent Neural Networks.. In *IJCAI*. 3818–3824.

[15] Jing Ma, Wei Gao, and Kam-Fai Wong. 2018. Rumor detection on twitter with tree-structured recursive neural networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1. 1980–1989.

[16] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

[17] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. 2013. On the difficulty of training recurrent neural networks. In *International Conference on Machine Learning*. 1310–1318.

[18] Dong ping Tian et al. 2013. A review on image feature extraction and representation techniques. *International Journal of Multimedia and Ubiquitous Engineering* 8, 4 (2013), 385–396.

[19] Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2016. Credibility assessment of textual claims on the web. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. ACM, 2173–2178.

[20] Victoria Rubin, Niall Conroy, Yimin Chen, and Sarah Cornwell. 2016. Fake news or truth? using satirical cues to detect potentially misleading news. In *Proceedings of the Second Workshop on Computational Approaches to Deception Detection*. 7–17.

[21] Natali Ruchansky, Sungyong Seo, and Yan Liu. 2017. Csi: A hybrid deep model for fake news detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. ACM, 797–806.

[22] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter* 19, 1 (2017), 22–36.

[23] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[24] Tetsuro Takahashi and Nobuyuki Igata. 2012. Rumor detection on twitter. In *Soft Computing and Intelligent Systems (SCIS) and 13th International Symposium on Advanced Intelligent Systems (ISIS), 2012 Joint 6th International Conference on*. IEEE, 452–457.

[25] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3156–3164.

[26] Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. EANN: Event Adversarial Neural Networks for Multi-Modal Fake News Detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 849–857.

[27] Ke Wu, Song Yang, and Kenny Q Zhu. 2015. False rumors detection on sina weibo by propagation structures. In *Data Engineering (ICDE), 2015 IEEE 31st International Conference on*. IEEE, 651–662.

[28] Liang Wu and Huan Liu. 2018. Tracing fake-news footprints: Characterizing social media messages by how they propagate. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. ACM, 637–645.

[29] Yin Zhu, Xiao Wang, Erheng Zhong, Nathan Nan Liu, He Li, and Qiang Yang. 2012. Discovering Spammers in Social Networks.. In *AAAI*.