Supervised Contrastive Learning for Multimodal Unreliable News

Wenjia Zhang

摘要

随着数字新闻产业成为信息传播的主要渠道，假新闻的负面影响呈爆发性增长。新闻报道的可信度不应孤立地考虑。相反，以前发表的关于类似事件的新闻文章可以用来评估新闻报道的可信度。受此启发，我们提出了一个基于 bert 的多模态不可靠新闻检测框架，该框架利用对比学习策略从不可靠文章中获取文本和视觉信息。对比学习器与不可靠新闻分类器互动，将相似的可信新闻(或相似的不可靠新闻) 推得更近，同时在多模态嵌入空间中移动内容相似但可信度标签相反的新闻文章。在 covid -19 相关数据集 ReCOVery 上的实验结果表明，我们的模型在不可靠新闻检测方面优于 baseline。

关键词：虚假新闻检测；多模态；对比学习；

1 引言

近年来，作为即时信息传播渠道的在线网站越来越多。然而，它们也可以用来传播骗局和虚假信息。2020 年 3 月，世界卫生组织宣布新冠肺炎为全球大流行，这为假新闻提供了温床。关于诊断检测和免疫接种运动的错误信息可能会造成恐慌、社会反应分散，并使人们的生命处于危险之中。

传统上，新闻可信度是通过二分类，聚类或基于知识的方法识别给定新闻文章的真实性来评估的。虚假新闻检测的方法主要是基于文本的，包括基于手工特征的分类器和 CNN、RNN 等深度神经网络。最近，多模态方法试图通过注意机制捕捉模态之间的相互作用来学习文本和图像模态之间的不协调表达。现有的方法都孤立地考虑单个新闻的可信度，也就是只关注输入文章中文字与图像之间的相关性，而忽略了文章之间的微妙关系。在本文中，作者提出了一种新的对比学习框架，该框架利用在给定文章之前发布的最相似的可信/不可靠的新闻来提炼多模态嵌入空间中的学习指标，以克服上述问题。

具体来说，受 Vision Transformer[1]的启发，作者提出了一个统一的基于 bert 的学习框架，该框架基于多头注意力机制在潜在空间中比较文本和视觉信息，以更有效地从多种模态中提取特征。在此之上还加入了一种对比学习策略，以鼓励相似的可信新闻文章 (或相似的不可靠新闻) 在多模态嵌入空间中更加接近和相似，但具有相反可信度等级标签的新闻远离。同时，为了使用对比学习更有效率地训练模型，作者还使用了 memory bank 来缓存当前 epoch 输入文章的表征，以便在下一个 epoch 的对比损失中计算新闻文章的相似度。

2 相关工作

2.1 虚假新闻检测

早期的虚假新闻检测方法主要基于文本特征，并建立在各种神经网络架构上[2-3]。最近，Zhou 等人[4]提出了一种多模态虚假新闻检测方法，该方法利用预先训练好的图像字幕模型，首先为新闻文章

中的图像生成文本描述，然后将其与新闻文本连接起来，训练一个基于文本的假新闻分类器。然而，图像字幕模型生成的文本质量限制了虚假新闻检测的性能。此外，这样的方法不能捕捉到文本与图像特征之间的微妙交互。其他多模态方法首先将文本与图像映射到到它们各自的嵌入空间，然后通过简单的拼接，使用注意机制或张量网络[5-6]来融合多模态特征。然而，现有的方法只专注于建模单篇文章中的文本和图像之间的关系，忽略了文章之间的关系。而我们提出了一种新的对比学习框架，该框架利用在给定文章之前发布的最相似的可信/不可靠新闻来提炼多模态嵌入空间中的学习指标，以实现更好的不可靠新闻检测性能。

2.2 对比学习

对比学习的目的是学习一个嵌入空间，在这个嵌入空间中，正对相互靠近，负对相互远离。它也与度量学习密切相关，度量学习的目的是学习嵌入空间[7]中的距离函数。监督对比学习方法[8]同时采用了监督的分类损失和对比损失。他们在视觉表征学习[9]和少样本学习[10]等任务中取得了许多成功。然而，还没有研究利用监督对比学习进行多模态虚假新闻检测。一个可能的原因是，在典型的监督对比学习中，假设正样本和随机选择的负样本之间的不相关程度可以通过比较它们的类别标签的对比损失来捕捉。可是在虚假新闻检测中，即使文章共享相同的标签，也可能是不相关的，因为它们可能讨论不同的事件或主题。因此，想要将监督对比学习应用在虚假新闻检测中，需要制定一种不同的选择负样本进行对比学习的策略。

2.3 多模态模型

受到 ViT[1]的启发，Transformer 架构也能用于图像的特征提取，多模态领域开始逐渐转向采用端到端的纯 Transformer 架构。ViLT[11]首先做出了纯 Transformer 的尝试，其采用了与 ViT 类似的架构，将文本和图像向量化后直接拼接输入到模型中，直接提取多模态特征。CLIP[12]将对比学习引入多模态领域，利用大规模图文对比来训练模型的跨模态对齐和匹配能力，使其具备出色的检索性能。ALBEF[13]采用了类似于原始 Transformer 的 Encoder-Decoder 架构，但是对 Decoder 前半改成了 Encoder，后半的 cross-attention 用于融合多模态特征，这样能在特征融合之前优先进行图文对齐，在多个任务上达到了 SOTA。

3 本文方法

3.1 本文方法概述

本文方法的框架如图 1 所示。整个框架可以分为三部分：Image Encoder、Text Encoder 和 Joint Learning。Image Encoder 先将输入图像切片为一些小的图像块，对每个图像块使用 ResNet-50 提取特征，之后拼接起来作为整张图像的特征。Text Encoder 是一个经过预训练的 BERT，对整个输入文本进行特征提取。在得到图像和文本的特征后，将特征直接拼接，送入一个多头 Transformer 以提取新闻的多模态特征。Joint Learning 在多模态特征之上将新闻与以往发布的新闻做对比学习，得到对比损失，其与分类用的交叉熵损失构成了模型的总损失。

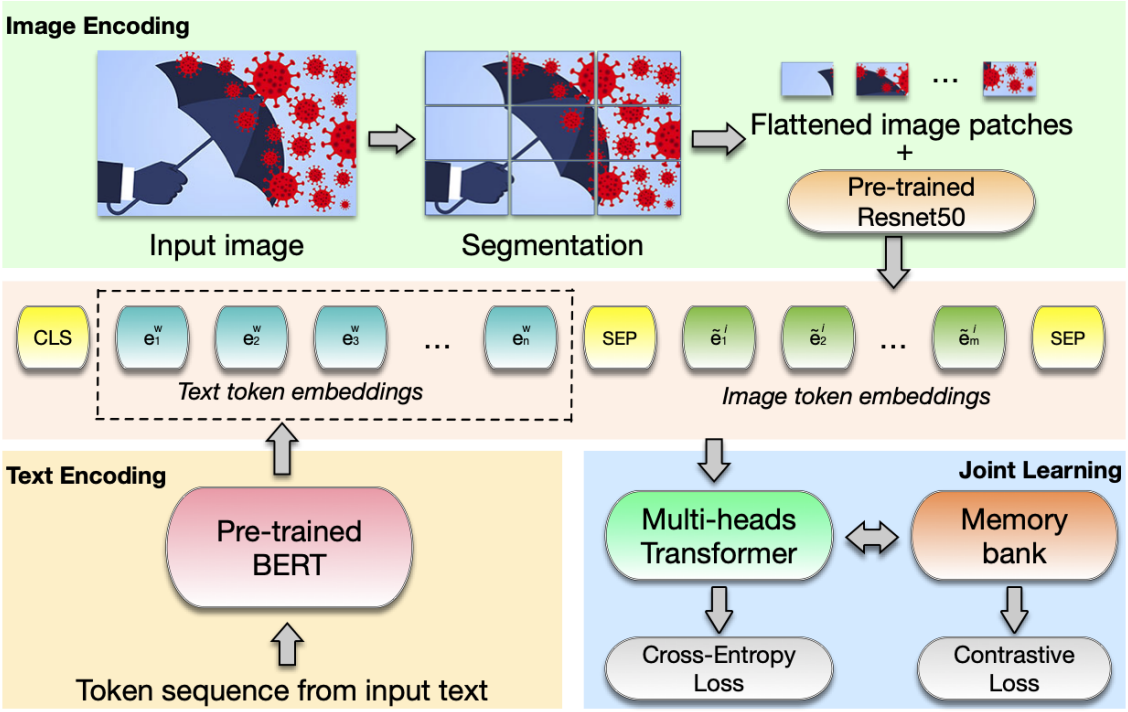

图 1: 方法示意图

3.2 损失函数

对新闻的最终表示通过 sigmoid 得到分类结果 $\hat{y} = \text{Sigmoid}(W_c x_d + b_c)$，分类结果与真实标签会计算交叉熵损失：

$$\mathcal{L}_c = -\sum_{c=1}^{C} y \log(\hat{y}) \tag{1}$$

其中，y 为真实标签，C 为类别数目。

原始的对比损失没法直接应用于此处，对于每条新闻，需要找到 k 个正样本和 k 个负样本，这里的样本选择策略是比较特殊的。具体来说，需要找到发布时间早于本新闻的所有新闻，将可信度相同的视为正样本，不同的视为负样本。对新闻标题计算彼此的余弦相似度，各取相似度最高的 $top-k$ 个。最终的对比损失可以表示为：

$$\mathcal{L}_s = -\frac{1}{2k}\left(\sum_{s_{pos}\in S^x_{pos}} \log \frac{\exp(\cos(x, s_{pos}))}{\sum_{s_{neg}\in S^x_{neg}} \exp(\cos(x, s_{neg}))}\right) \tag{2}$$

其中，s_{pos} 为正样本，s_{neg} 为负样本。

总损失为交叉熵和对比损失的和，由参数 α 控制权重：

$$\mathcal{L} = (1-\alpha)\mathcal{L}_c + \alpha\mathcal{L}_s \tag{3}$$

4 复现细节

4.1 与已有开源代码对比

此处使用了作者发布的代码（https://github.com/wenjiazh/btic）来作为复现工作的基础，但作者发布的是 Jupyter 代码文件，各种代码都分散在文件各处，难以进行改进优化。在最终实验的代码中，保留了原作者 memory bank 的代码逻辑，其余部分包括训练、数据处理、数据集、模型等都进行了重写。由于作者所使用的数据集 ReCOVery 无法完整访问，改用了 Fakeddit[14]作为训练数据集，它们都是类

似格式的多模态虚假新闻检测数据集。

除了对原文的复现，还在这之上进行了三个改进实验。分别是：1) Swin Transformer[15]替换 Image Encoder；2) ConvNeXt[16]替换 Image Encoder；3) ALBEF[13]作为 Multimodal Encoder。

4.2 Swin Transformer 替换 Image Encoder

在原文中，作者使用 ResNet-50 单独提取每个图像块的特征作为图像的特征序列，这样做并没有什么意义并且非常耗时，因为提取出的图像块特征之间是独立的，这导致了最后的图像特征中，每一个 token 都只有局部信息。另一方面，划分图像块的操作在 ViT 类模型中应用地十分广泛，并且取得了不错的性能。因此，如图 2 所示，本实验使用 Swin Transformer 来替换 Image Encoder 以实验其性能提升。Swin Transformer 是在 ViT 的基础上改进而来的一个模型，它借用了 CNN 的思想来改造 ViT，解决了很多缺点，并且同参数性能上是在当时最佳的。

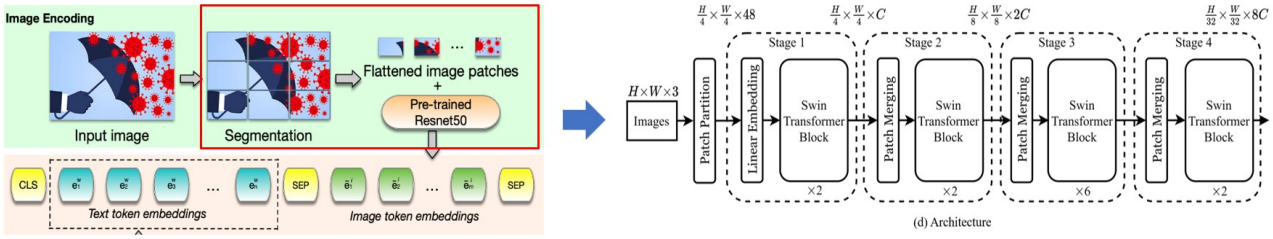

图 2: 实验一示意图

4.3 ConvNeXt 替换 Image Encoder

在原文中，文本端的特征提取器是 Transformer，能直接得到特征序列，而视觉端是 CNN，需要将矩形特征展开成序列形式。而在多模态任务中，一个十分重要的操作就是模态间的对齐，原模型没有做到这一点。实验一的改进也让两侧的架构变得相同了，这样无法判断是同构带来的提升，还是模型本身性能带来的提升，所以需要有一个对比实验。如图 3 所示，本实验使用 ConvNeXt 来替换 Image Encoder 来与实验一做对比。ConvNeXt 是一个仿照 Swin Transformer 来改造 CNN 的模型，它在思想、结构、性能等方面都与 Swin Transformer 十分相似，最大的区别就是 ConvNeXt 用卷积实现，属于 CNN。

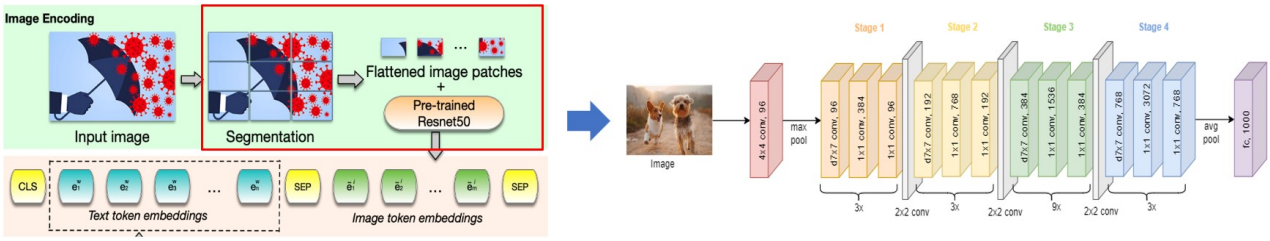

图 3: 实验二示意图

4.4 ALBEF 作为 Multimodal Encoder

原文中的模态融合部分做的比较简单，直接将特征进行拼接，这样其实并不能得到很好的多模态特征，因此考虑使用真正的多模态模型来完成。如图 4 所示，本实验使用 ALBEF 来作为多模态的特征提取器，将原模型中的整个特征提取部分都换成了多模态模型，保留最后的对比学习。

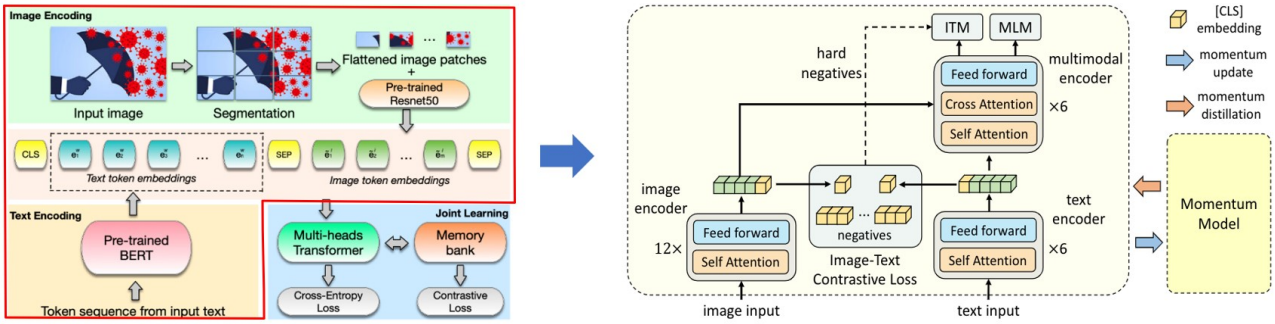

图 4: 实验三示意图

5 实验结果分析

实验结果如表 1 所示，实验的评价指标采用了分类问题常用的四个指标：准确率、精确度、召回率以及 F1 分数，其中不同的实验使用不同的 Encoder 名称区分。ConvNeXt 的结果与 Res50 处于同一水平，此处没提升是因为 ConvNeXt 作用在原图上而 Res50 作用在切割出的 patch 上，推理速度会快很多。Swin-T 比 ConvNeXt 高 2%多，二者在图像分类上是相近的性能，但是 Swin-T + BERT 都是 Transformer 架构，相同架构能很大地缓解多模态不对齐的问题，证明了同构的作用。ALBEF 作为多模态模型，效果显著强于其他的方法，证明了多模态模型所提取到的特征要更好。不过这里所使用的 ALBEF 参数量要高于其他的模型，并没有相近参数的模型可供使用，所以只能当作参考。

	Fakeddit			
Method	Accuracy	Precision	Recall	F1-score
Res50	86.12	86.09	86.12	86.10
ConvNeXt	86.05	86.3	86.05	86.11
Swin-T	88.45	88.55	88.45	88.48
ALBEF	**94.52**	**94.54**	**94.52**	**94.53**

表 1: 实验结果

6 总结与展望

本文提出了一种新的虚假新闻检测框架，其将文本和视觉信息都被输入到一个基于 bert 的多模态模型中，以生成多模态特征，更好地编码文本和图像之间的交互。在此之上还结合了对比学习，通过使用过去发表的文章和报道的类似事件来更好地学习多模态表示。

不过该框架依然存在一定的问题，在图像编码部分的处理策略不够好，所以设计了三个改进实验来进行对比。前两个实验相互对比，将框架中的图像编码器部分进行替换，取得了一定的提升，并得出了同构有利于模态融合的结论。第三个实验用了一个著名的多模态模型，证明了真正的多模态模型在多模态特征提取上有着独到之处。在复现工作之中也存在着一些不足之处，例如切换模型需要改很多代码，实验方案较为简单等，还可继续改进。

参考文献

[1] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale[J]. Learning, 2020.

[2] CHEN T, WU L, LI X, et al. Call Attention to Rumors: Deep Attention Based Recurrent Neural Networks for Early Rumor Detection[J]. knowledge discovery and data mining, 2017.

[3] LU Y J, LI C T. GCAN: Graph-aware Co-Attention Networks for Explainable Fake News Detection on Social Media[J]. meeting of the association for computational linguistics, 2020.

[4] ZHOU X, MULAY A, FERRARA E, et al. ReCOVery: A Multimodal Repository for COVID-19 News Credibility Research[J]. conference on information and knowledge management, 2020.

[5] WANG Y, MA F, JIN Z, et al. EANN: Event Adversarial Neural Networks for Multi-Modal Fake News Detection[J]. knowledge discovery and data mining, 2018.

[6] JIN Z, CAO J, GUO H, et al. Multimodal Fusion with Recurrent Neural Networks for Rumor Detection on Microblogs[J]. acm multimedia, 2017.

[7] WEINBERGER K Q, SAUL L K. Distance Metric Learning for Large Margin Nearest Neighbor Classification[J]. Journal of Machine Learning Research, 2009.

[8] KHOSLA P, TETERWAK P, WANG C, et al. Supervised Contrastive Learning[J]. neural information processing systems, 2020.

[9] DWIBEDI D, AYTAR Y, TOMPSON J, et al. With a Little Help From My Friends: Nearest-Neighbor Contrastive Learning of Visual Representations[J]. international conference on computer vision, 2021.

[10] MAJUMDER O, RAVICHANDRAN A, MAJI S, et al. Revisiting Contrastive Learning for Few-Shot Classification.[J]. arXiv: Computer Vision and Pattern Recognition, 2021.

[11] KIM W, SON B, KIM I. ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision[J]. arXiv: Machine Learning, 2021.

[12] RADFORD A, KIM J W, HALLACY C, et al. Learning Transferable Visual Models From Natural Language Supervision[J]. international conference on machine learning, 2021.

[13] LI J, SELVARAJU R R, GOTMARE A, et al. Align before Fuse: Vision and Language Representation Learning with Momentum Distillation[J]. neural information processing systems, 2021.

[14] NAKAMURA K, LEVY S, WANG W Y. Fakeddit: A New Multimodal Benchmark Dataset for Fine-grained Fake News Detection.[J]. Language Resources and Evaluation, 2020.

[15] LIU Z, LIN Y, CAO Y, et al. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows.[J]. international conference on computer vision, 2021.

[16] LIU Z, MAO H, WU C Y, et al. A ConvNet for the 2020s[C]/ /. 2022.