用深度学习对恶意模因 (Hateful Memes) 进行分类

孟念 - Niklas Muennighoff
北京大学
Python 数据分析, 春 21
mengnian@stu.pku.edu.cn

1 引言

语言总是作为沟通媒介的一部分出现，例如视觉中的书面文字或声音中的口头文字。对于某些问题，提取语言用以建立有用的机器学习模型。因此，纯语言模型在机器翻译、文本分类任务或语言推理问题等领域对我们的生活产生了相当大的影响 [1]。然而，通常情况下，基础媒介是相关的，必须与语言结合起来理解。

其中一个领域是网络模因。它包含了图像和叠加的文本，用以提供一个细微的信息。通常情况下，图像或文字本身并不足以传达信息。这种细微的信息可能是恶意性的，但人工分类和删除恶意模因的成本很高。为了训练机器学习模型来解决这个问题，Facebook AI 发布了 The Hateful Memes Dataset [2]。 目前最先进的视觉 + 语言机器学习模型是基于注意力机制的方法 (transformers)[3]。其中，有两种流行的方法。单流模型，如 VisualBERT [4]、UNITER [5]、OSCAR [6]，使用一个 transformer 来同时处理图像和语言输入；双流模型，如 LXMERT [7], ERNIE-ViL [8], DeVLBERT [9], VilBERT [10], 依靠独立的两个 transformers 来处理视觉和语言，然后在模型的最后阶段进行合并。

下面，我将对数据集和机器学习问题进行分析。以及将阐述我用来解决问题的方法。最后，我将讨论结果和一些延伸想法。

2 问题和数据介绍

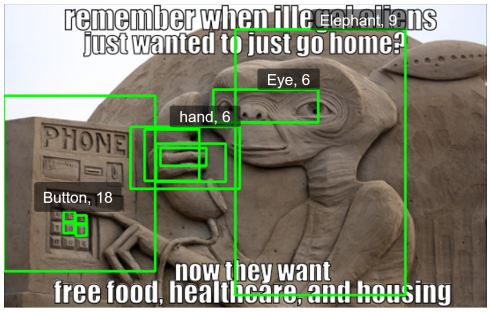
Hateful Memes 数据集有 8500 张图片的训练集、500 张图片的验证集和 1000 张图片的测试集组成。模因文本存在于图片上，但也在额外的 jsonl 文件中提供。 为了增加难度，该数据集包括文本和视觉混杂物。这种混杂物通过交换文本或图像，从恶意模因变为普通模因。图就是这样一个例子。他们确保模型必须对视觉和语言都进行推理。仅有视觉或仅有语言的模型无法成功完成任务。 因为这个数据集是比赛的一部分，测试集的标签是隐藏的。我们的目标是使用提供的训练和验证集来预测测试集。

$$AUC = \int_{x=0}^{1} \mathrm{TPR}(\mathrm{FPR}^{-1}(x))\, dx \tag{1}$$

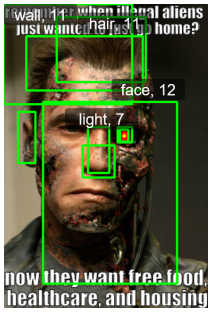
对测试集的预测由 AUC 来评分。直观地说，模型必须善于对模因的恶意性进行排序。概率值本身是高还是低并不重要，重要的是如何对模因进行排序。

3 3 具体方法

图中包含有多层次的概述。该通道由三个阶段组成：数据准备、模型和集成模型。



(a) 恶意模因



(b) 普通模因

图 1: Hateful Memes 数据集的一个混淆例子，有预测的 RoI (Regions of Interest, 重要的部分（绿框)) 和预测的对象 (白色的文字)。模因 a）是恶意性的，因为它引用了《E.T. 外星人》的电影来嘲笑美国的非法移民 (背景中的外星人来自《E.T. 外星人》的电影。在电影中，外星人 E.T. 总是说他想回家, 所以《还记得非法移民只想回家的时候，现在他们想要免费的食物、医疗和住房。》是恶意性的)。模因 b）不是恶意性的，因为背景中没有 E.T. 的形象，因此也没有关于移民的笑话。要注意，两张图片中的文字都是一样的，只有图片发生了变化。这说明了为什么模型必须能够同时处理图像和文字。©Getty Images

3.1 数据分析和准备

3.1.1 词语分析

为了更好地理解数据分布, 我研究了哪些词在恶意模因中最常见。使用了 TF-IDF 的改编版本, 我计算了每个词出现在恶意和非恶意模因中的可能性。然而，根据这个比例，最恶意的词本身往往不是恶意的，而是与经常被嘲弄的话题有关，比如“非法移民”。一个只用这个来预测测试集的简单模型取得了 60.6% 的 AUC。要注意，这个模型只是使用了一个单词比率的字典，而没有任何关于图像的信息。　　为了确定 transformer 模型的最佳序列长度，我对离群值进行了分析。数据中最长的模因文本约为 400 个字符。再将它转换为 transformer 的输入符号，大约是 80 个 transformer 符号。因此，我将最大长度设置为 128 个符号，以确保每个输入文本都能被 transformer 使用。

3.1.2 数据清洗

由于决定一个模因是否恶意性的是主观性的，恶意模因数据集是非常嘈杂的。存在 4 种类型的问题。1. 有 100% 的重复数据 2. 有相同标签的略微不同的重复数据 3. 有标签不同的重复数据 (模因一样，标签不同 - 所以一个标签绝对有错误) 4. 有错误标签的模因，可不是重复数据目标应该是去除其中的大部分，因为它们可能使模型的训练变得更糟 (Garbage in, garbage out!)。对于 1.-3.，我首先需要识别所有可能的重复内容。重复代表的意思是，两个模因的文字和图像都非常相似。为了识别相似的文本，我使用 Levenshtein 距离，定义为:

$$\operatorname{lev}(a,b) = \begin{cases} |a| & \text{if } |b| = 0, \\ |b| & \text{if } |a| = 0, \\ \operatorname{lev}(\operatorname{tail}(a), \operatorname{tail}(b)) & \text{if } a[0] = b[0] \\ 1 + \min \begin{cases} \operatorname{lev}(\operatorname{tail}(a), b) \\ \operatorname{lev}(a, \operatorname{tail}(b)) \\ \operatorname{lev}(\operatorname{tail}(a), \operatorname{tail}(b)) \end{cases} & \text{otherwise.} \end{cases}$$

我计算每个样本之间 Levenshtein 距离，并通过文本长度将其归一化，得到一个范围在 0-1 的比率。我保留所有比率大于 0.95 的例子作为可能的重复数据。接下来，我需要检查它们的图像是否也是相似的。为了做到这一点，我使用两个哈希函数对所有图像进行编码。首先，使用一个感知哈希函数 (perceptual hash function)，

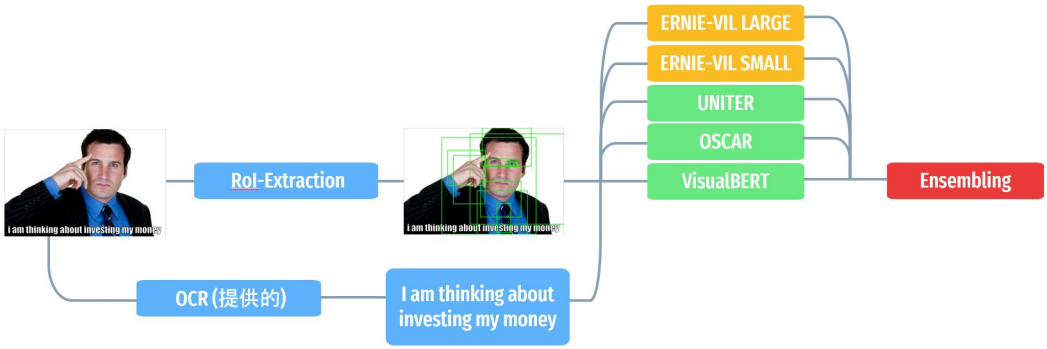


图 2: 我在 Hateful Memes 数据集上的模型管道分为三个阶段：数据准备、模型和集成学习。在开始时，我清洗数据并删除重复的数据和错误标记的模因。然后，我从每张图片中提取 RoIs，并将它们与所提供的图片文本相结合（蓝色）。接下来，我用 RoIs 和文本训练每个模型的多个随机种子（绿色标记的单流模型，黄色标记的双流模型）。最后，我将所有的模型进行集成学习（红色）。©Getty Images

它善于找到所有看起来相似的模因。然而，它漏掉了几个重复的图片，这些图片是相同的，但由于文字在图片上的位置不同，所以看起来不一样。所以我构建了第二个哈希函数，它比较了图像尺寸和角落的像素值，以找到这样的图像。剩下的数据有 1.-3. 种类型的分组重复。如果在一组重复中，有任何标签是不同的，我就把该组中的所有模因删除（3. 种类型）。对于 1.-2. 种类型，我保留其中一个重复的例子，并删除其余。

如果不手动识别图像，要找到标签错误的 4 种类型模因是很困难的。幸运的是，其他用户报告了一个对可能不正确的例子的简短清单。我据此删除了该清单中的所有模因。

总的来说，我删除了 1.98% 的训练数据。在验证和测试数据中没有问题。

3.1.3 输入数据

第一步，使用 detectron2 框架从图像中提取出 RoIs（Regions of Interest）。使用 RoIs 而不是整个图像可以加快训练过程，并且仍然包括最重要的内容。事实证明，使用不同的 RoIs 集有利于训练。因此，我用不同的 detectron2 模型（在 VisualGenome 上预先训练过）来提取不同的 RoIs。图 1 显示了一个预测 RoI 的例子，以及两个模因上最常预测的 4 个对象。

与使用光学字符识别（OCR, Optical Character Recognition）提取并在数据集中提供的模因文本一起，我将 RoI 送入模型。

3.2 模型

以下是我对个别模型的一些设置和变化的具体解释。

3.2.1 一般设置

我使用由视觉 + 语言模型的原作者提供的预训练权值。我在 Hateful Memes 数据集上使用 MLM(Masked Language Modelling) 对 VisualBERT 和 OSCAR 再进行了预训练。这是一种无监督的训练方法，输入的部分内容被屏蔽掉，模型试图预测把他们预测出来。 然后，我使用有标签的数据来微调所有的模型，使用 Binary Cross Entropy 的损失函数和 8 个批次大小。我使用 Adam 优化器，学习率为 1e-5，预热步骤率为 10%。为了避免梯度爆炸，我将 VisualBERT、OSCAR 和 UNITER 的最大梯度设置为 5，将 ERNIE-ViL 设置为 1。VisualBERT, OSCAR 和 UNITER 训练了 5 个 epochs，在训练的最后 25% 期间使用了随机权值平均法 (SWA, Stochastic Weight Averaging)。ERNIE-ViL 模型训练了 5000 步。我从所有模型的最后一步提取权值，并将其用于测试集的预测。 总的来说，在超参数优化上花费的时间不多，因为基本架构的变化影响更大。

来源	模型	**验证集** AUC	**测试集** AUC
Facebook AI 提供的 Hateful Memes 基线	人类	-	82.65
	ViLBERT	71.13	70.45
	VisualBERT	70.60	71.33
	ViLBERT CC	70.07	70.03
	VisualBERT COCO	73.97	71.41
我的方案	VisualBERT	75.49	75.75
	OSCAR	77.16	77.30
	UNITER	77.75	78.65
	ERNIE-ViL Base	78.18	77.02
	ERNIE-ViL Large	78.76	80.59
	Ensemble	81.56	82.52

表 1: 模型性能。

3.2.2 ERNIE-ViL

用飞桨 (PaddlePaddle, 百度的深度学习框架) 编写的 ERNIE-ViL 模型是基于 VilBERT 和 ERNIE 的。除了提取的 RoI 之外, 原始的 ERNIE-ViL 还在地面真理标记 (ground truth labels) 的 RoI 上进行了预训练。由于 Hateful Memes 数据集没有提供 ground-truth 的 RoI，所以我用 10-100 个提取的 RoI 作为假 ground-truth。除了对 CC（Conceptual Captions）进行预训练的权值外，还使用了对 VCR 进行预训练的权值来增加多样性。

3.2.3 UNITER & OSCAR

根据 huggingface/transformers 的变化更新了 UNITER 和 OSCAR，例如更新了激活函数和 embedding 的计算。OSCAR 还使用图像-文本匹配 (ITM, Image-Text Matching) 和屏蔽语言 (MLM, Masked Language Modelling) 对 Hateful Memes 进行了预训练。在预训练过程中，像原来的实现那样添加预测的对象是没有好处的。我使用了 LXMERT 的分类层与 GeLU 的激活函数。OSCAR 和 UNITER 的预训练权值是基于 BERT transformer 的。我还用 RoBERTa 和 ALBERT 做了实验，然而由于没有预训练的权值，它们的预测结果比 BERT 差。

3.2.4 VisualBERT

与 UNITER 和 OSCAR 类似，我基于 huggingface/transformers 更新了 VisualBERT。与 OSCAR 一样，VisualBERT 模型也是使用 MLM 进行预训练的。虽然原始的 VisualBERT 对语言和视觉输入使用相同的 token type(标记类型)，但我创建了一个单独的视觉 token type。因此，我从头开始初始化和重新训练 token type 权值。这使模型的 AUC 提高了 1.2%。Multi-sample Dropout 和对 transformer 层进行平均化，进一步改善了模型。我在 transformer 之后使用的分类层的学习率是 transformer 学习率的 500 倍。

3.3 集成学习

对于每个模型，我用不同的提取 RoI 来平均 3-5 个随机种子。然后，我将每个模型的平均预测值送入一个集成学习循环 (ensembling loop)，应用 simple averaging(简单平均法)、ranked averaging(等级平均法)、power averaging 和 Simplex (单纯形法, 用的是下山单纯形法 (Nelder-Mead)) 来创建最终预测。Simplex 的权值是根据验证集的预测来学习的，然后应用于在完整数据集（train+val）上训练的模型的测试集预测。

4 结果和讨论

4.1 性能和限制

各个模型在验证集和测试集上的 AUC 可以在表 1 中看到。我的 ensemble（集成学习后的模型）缩小了所提供的基线模型和人类在 Hateful Memes 上的 AUC 之间的差距。然而，Facebook AI 指出，为人类的 AUC 测试的人并不是专家，人类的实际 AUC 能更接近于 100%。调查这一点并建立一个新的人类基准线是一个有趣的研究方向。 虽然没有报告，但我的模型作标签预测，而不是概率预测的集合准确率只有 75.40%，而人类在测试集上的准确率是 84%。其原因是，我优化了模型，以便用概率（AUC）对模因进行排名。然而，人类在二进制预测（准确率）方面比概率预测更好。准确率为未来的研究留下了相当大的差距，需要弥补。

包括平均的随机种子，我的 ensemble 是由 19 个训练好的模型组成的。特别是 ERNIE-ViL 模型往往是不稳定的，因此要对五个随机种子进行平均化。为 ERNIE-ViL 增加随机权值平均法，而不仅仅是其他模型，可以帮助解决这个问题。用更少的随机种子和更少的计算量实现同样的结果是值得研究的。

4.2 未来的研究想法

关于未来研究方向的四个想法如下：

- 有解决恶意性的 GIF 或者恶意性的短视频的方案吗?
- 在 Hateful Memes 数据集中，模因的标题是标准化的（它们都使用白色字体（图 1），但在现实中，它们的字体和大小可能有所不同。这个是有用的信息，因为它可能有助于模型确定恶意性。因此，我们能否将光学字符识别整合到训练模型中?
- 随着最近 transformer 应用于视觉方面的进展，我们是否可以跳过 RoI 提取，从头开始就应用单流或双流的 transformer?
- 单流 transformer，如 VisualBERT，似乎超过了双流 transformer，如 VilBERT，但 ERNIE-ViL 除外。ERNIE-ViL 复制了 VilBERT 的 transformer，可是以 Scene Graph Parser(场景图解析) 与众不同。用单流 transformer 创建的 ERNIE-VisualBERT 模型能否胜过双流的 ERNIE-ViL?

参考文献

[1] TB Brown, B Mann, N Ryder, M Subbiah, J Kaplan, P Dhariwal, A Neelakantan, P Shyam, G Sastry, A Askell, et al. Language models are few-shot learners. arxiv 2020. arXiv preprint arXiv:2005.14165, 4.

[2] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. arXiv preprint arXiv:2005.04790, 2020.

[3] A Vaswani, N Shazeer, N Parmar, J Uszkoreit, L Jones, AN Gomez, L Kaiser, and I Polosukhin. Attention is all you need. arxiv 2017. arXiv preprint arXiv:1706.03762, 2017.

[4] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. arXiv preprint arXiv:1908.03557, 2019.

[5] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Learning universal image-text representations. arXiv preprint arXiv:1909.11740, 2019.

[6] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In European Conference on Computer Vision, pages 121–137. Springer, 2020.

[7] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. arXiv preprint arXiv:1908.07490, 2019.

[8] Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. Ernie-vil: Knowledge enhanced vision-language representations through scene graph. arXiv preprint arXiv:2006.16934, 2020.

[9] Shengyu Zhang, Tan Jiang, Tan Wang, Kun Kuang, Zhou Zhao, Jianke Zhu, Jin Yu, Hongxia Yang, and Fei Wu. Devlbert: Learning deconfounded visio-linguistic representations. In Proceedings of the 28th ACM International Conference on Multimedia, pages 4373–4382, 2020.

[10] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In Advances in Neural Information Processing Systems, pages 13–23, 2019.