arXiv:2302.04156v1 [cs.CL] 8 Feb 2023

# 多模态仇恨表情包分类提示

**Rui Cao**[1], **Roy Ka-Wei Lee**[2],**Wen-Haw Chong**[1],**JingJiang**[11]新加坡管理大学[2]新加坡科技设计大学[1]
{ruicao.2020@phdcs.,whchong.2013@phdis.,jingjiang@}smu.edu.sg[2]
roy_lee@sutd.edu.sg

## 摘要

仇恨表情包分类是一项具有挑战性的多模态任务，需要复杂的推理和上下文背景知识。理想情况下，我们可以利用一个明确的外部知识库，为仇恨表情包补充上下文和文化信息。然而，目前尚无已知的明确外部知识库能够提供此类仇恨言论上下文信息。为弥补这一空白，我们提出了PromptHate，这是一种简单而有效的基于提示的模型，通过提示预训练语言模型（PLM）完成仇恨表情包分类。具体而言，我们构建了简单的提示，并提供少量上下文示例，以利用预训练 RoBERTa 语言模型中的隐式知识进行仇恨表情包分类。我们在两个公开可用的仇恨和冒犯性表情包数据集上开展了广泛的实验。实验结果表明，PromptHate 能够取得 90.96 的高曲线下面积，超越仇恨表情包分类任务上的先进基线模型。我们还针对不同的提示设置进行了细粒度分析和案例研究，并证明了这些提示在仇恨表情包分类中的有效性。

**免责声明**：本文包含歧视性内容，可能会令部分读者感到不适。

## 1 引言

互联网模因已发展成为社交媒体上最流行的交流形式之一。梗图以图像搭配文本的形式呈现，通常意在营造幽默或讽刺的效果。然而，恶意网络用户打着幽默的幌子制作并传播仇恨梗图（Kiela 等人，2020年；Pramanick 等人，2021年；Gomez 等人，2020年；Suryawanshi 等人，2020年）。许多仇恨表情包旨在针对种族、族裔和宗教等特定特征攻击和嘲讽他人。这些

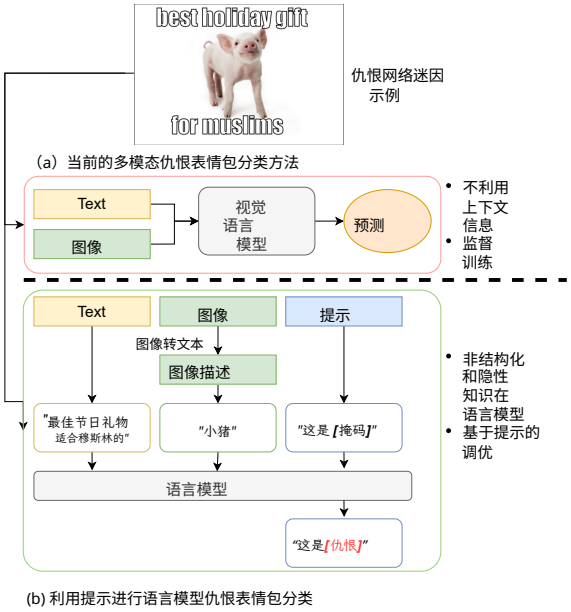



图1：（a）微调视觉语言模型方法与（b）基于提示的方法之间的比较。

仇恨表情包还可能比基于文本的网络仇恨言论对社会构成更大威胁，因为它们具有病毒式传播的特性；用户可以在多个对话和上下文中重新发布或分享这些仇恨表情包。

为了遏制仇恨梗图的传播，Facebook 等社交媒体平台最近发布了一个大型仇恨模因数据集，并将其作为一项挑战赛的一部分，以鼓励研究人员开发自动化解决方案，执行仇恨表情包分类（Kiela 等人，2020年）。然而，事实证明，对仇恨表情包进行分类是一项极具挑战性的任务，因为解决方案需要理解视觉和文本两种模态，并在二者之间进行推理。对这些模态进行推理还需要上下文背景知识。

近期研究提出了多模态仇恨表情包分类方法。例如，一些研究采用了预训练的视觉语言模型，如 VilBERT（Lu 等人，2019年）和 VisualBERT（Li 等人，2019年），并针对仇恨表情包分类任务对这些模型进行微调——

（Lippe 等人，2020年；Zhu，2020年；Zhou 和 Chen，2020年；Muennighoff，2020年；Ve- lioglu 和 Rose，2020年；Zhu 等人，2022）（如图1(a)所示）。然而，现有方法仍有局限性，因为理解仇恨梗图可能需要额外的背景知识。请看图1中的仇恨梗图示例。要推断该梗图具有仇恨性质，需要具备这样的背景知识：穆斯林认为猪是不洁之物，食用猪肉是一种罪过。

近期研究尝试对预训练语言模型（PLM）进行提示，并在单模态 NLP 任务中取得了良好表现（Schick 和 Schütze，2021年；Brown 等人，2020年；Schick 和 Schütze，2021年；Gao 等人，2021）。然而，鲜有研究尝试将提示用于 PLM 的多模态任务（Yao 等人，2021年；Zeng 等人，2022；Gui 等人，2021）。Yang 等人（2021年）探索了如何在视觉问答任务中提示 GPT-3 模型（Brown 等人，2020年）。然而，这种方法存在局限，因为 GPT-3 等大型模型的调优成本高昂。本研究针对这些研究空白，提出了一种新颖框架，利用 PLM 中的隐式、非结构化知识（Trinh 和 Le，2018年；Petroni 等人，2019年）来改进仇恨表情包分类。图 1 展示了现有多模态仇恨梗图分类方法与我们提出的基于提示的方法之间的对比。具体而言，在我们基于提示的方法中，首先将图像转换为 PL M 能够理解的文本描述，并设计专门的提示，以适配并利用 PLM 中的隐式知识。随后，给定一张梗图和一个提示，经过提示调优的 PLM 会生成文本输出，指示对输入梗图的预测标签。基于提示的方法背后的基本直觉是，PLM 会调用其大规模预训练数据中的隐式、非结构化知识来续写提示，即从“this is _”续写为“this is hateful.”。

我们将本文的贡献总结如下：（i）我们提出了一种名为 PromptHate 的多模态提示框架，通过提示并利用预训练语言模型中的隐式知识来执行仇恨表情包分类。（ii）我们在两个公开可用的数据集上进行了大量实验。实验结果表明，在仇恨表情包分类任务中，PromptHate 的表现优于最先进的方法。（iii）我们针对各种设置进行了细粒度分析和案例研究，以考察提示在仇恨梗图分类中的有效性。据我们所知，这是首篇探索通过提示预训练语言模型进行仇恨梗图分类的论文。[1]

## 2 相关工作

### 2.1 仇恨模因检测

仇恨表情包分类是一项新兴的多模态任务，随着近期多个仇恨梗图数据集的发布而受到广泛关注（Kiela 等人，2020年；Suryawanshi 等人，2020年；Gomez 等人，2020年）。例如，Facebook 组织了仇恨梗图挑战赛，鼓励研究人员提交解决方案，以执行仇恨表情包分类（Kiela 等人，2020年）。这些梗图经过特别构造，使得单模态方法无法在这一分类任务中取得良好表现。因此，现有研究采用了多模态方法来执行仇恨表情包分类。

现有研究探索了经典的双流模型，这些模型利用基于注意力的机制和其他融合方法，将文本编码器和图像编码器学习到的文本与视觉特征结合起来，以执行仇恨表情包分类（Zhang 等人，2020；Kiela 等人，2020；Suryawanshi 等人，2020）。另一类常见方法是针对该任务对大规模预训练多模态模型进行调优（Lippe 等人，2020；Zhu，2020；Zhou 和 Chen，2020；Muennighoff，2020；Velioglu 和 Rose，2020；Pramanick 等人，2021年；Hee 等人，2022）。近期研究还尝试使用数据增强（Zhu，2020；Zhou 和 Chen，2020；Zhu 等人，2022）和集成方法（Zhu，2020；Velioglu 和 Rose，2020；Sandulescu，2020）来提升仇恨表情包分类的性能。在一项近期工作中，Lee 等人（2021）提出了 DisMultiHate 模型，旨在将梗图中的仇恨目标解耦。然而，现有研究在建模仇恨梗图中编码的上下文背景知识方面可能仍显不足。本文旨在通过提示预训练语言模型，利用其非结构化的隐式知识来进行仇恨表情包分类，从而弥补这一研究空白。

### 2.2 语言模型提示

GPT（Radford 等人，2019年；Brown 等人，2020年）等大规模预训练语言模型（PLM）日益普及，BERT（Devlin 等人，2019年）、RoBERTa（Liu 等人，

[1]代码：https://gitlab.com/bottle_shop/safe/prompthate

2019年）也推动了基于提示的学习的普及。现有的基于提示的学习研究探索了将预训练语言模型用作隐式、非结构化知识库（Talmor 等人，2020；Davison 等人，2019年；Schwartz 等人，2017年）。近期研究还通过提示预训练语言模型来执行自然语言推理和情感分类等自然语言处理任务，并在少样本设置中取得了良好表现（Gao 等人，2021年；Schick 和 Schütze，2021年，b）。近期也有研究通过提示视觉语言模型来执行计算机视觉任务（Zhou 等人，2022，2021年；Radford 等人，2021年）。

尽管如此，现有的大多数基于提示的学习研究都局限于单模态任务，针对多模态任务提示预训练语言模型的研究则较少（Yao 等人，2021年；Zeng 等人，2022；Gui 等人，2021年）。Yang 等人（2021年）探索了为视觉问答任务提示 GPT-3 模型的做法（Brown 等人，2020年）。然而，这种方法存在局限性：GPT-3 等大型模型的调优成本高昂。此外，输入长度的限制也限制了训练实例的数量。本文采用不同的方法，提出一种新颖的框架，PromptHate，通过提示 RoBERTa（Liu 等人，2019年）完成多模态仇恨梗图分类。PromptHate 与 GPT-3 相比是一个小得多的模型，而且可以使用训练实例进行微调。

## 3 预备知识

### 3.1 问题定义

我们将多模态仇恨梗图分类问题定义如下：给定一张包含图像 $\mathcal{I}$ 和文本 $\mathcal{O}$的梗图，分类模型将预测该多模态梗图的标签（仇恨或非仇恨）。传统上，这一二分类任务要求模型预测一个涵盖两个类别的概率向量$\mathbf{y} \in \mathbb{R}^2$。具体而言，$y_0$表示预测该梗图为非仇恨的概率，而 $y_1$ 表示预测该梗图为仇恨的概率。如果 $y_1 > y_0$，则预测该梗图为仇恨，否则预测为非仇恨。在该框架中，我们将仇恨表情包分类任务转化为掩码语言建模（MLM）问题。具体而言，我们提示预训练语言模型替换表示梗图标签（例如仇恨或非仇恨）的 `[MASK]` 词元。我们将在第4.1节讨论提示的具体细节。

### 3.2 图像描述

为了提示预训练语言模型执行多模态仇恨梗图分类，我们首先需要将梗图中的图像转换为预训练语言模型可接受的文本输入。提取图像语义并以文本描述呈现的一种常用方法是图像描述生成（Yang 等人，2021年；Gui 等人，2021年）。我们首先使用开源 Python 软件包 EasyOCR 提取梗图中的文本[2]，，然后使用 MMEditing 进行图像修复[3] 以移除文本。接着，我们应用预训练的图像描述生成模型 ClipCap（Mokady 等人，2021年）。ClipCap 能够为低分辨率网页图像生成高质量的图像描述。生成的图像描述通常会描述梗图图像中的主要物体或事件，我们将这些图像描述作为 PromptHate模型的输入。

除了为图像生成描述，我们还利用 Google Vision 网页实体检测 API[4] 和预训练的 FairFace 分类器（Kärkkäinen 和 Joo，2019年）提取梗图中的实体，以及在梗图中包含人物时提取人口统计信息。提取的实体和人口统计信息将作为补充信息，与图像描述结合后输入预训练语言模型。请注意，尽管提取的补充信息可能捕捉到梗图中的关键信息，但图像描述和补充信息中仍缺少背景知识。例如，利用实体信息，我们可以识别梗图中的一头猪，并从梗图文字中提取“穆斯林”一词。然而，穆斯林不吃猪肉这一背景知识并未包含在补充信息中。

## 4 方法论

图 2展示了我们提出的 PromptHate 模型的架构框架。 PromptHate的关键步骤是构建一个提示，其中包括一个正示例（即正常表情包）、一个负向示例（即仇恨梗图）和一个推理实例（即待预测的梗图）。我们首先使用第3.2节所述的数据预处理步骤，将这三个梗图转换为表情包文本和图像描述。随后，我们构建模板，即包含标签词的自然句子，用于

[2]https://github.com/JaidedAI/EasyOCR [3] https://github.com/open-mmlab/mmediting [4] https://cloud.google.com/vision/docs/detecting-web

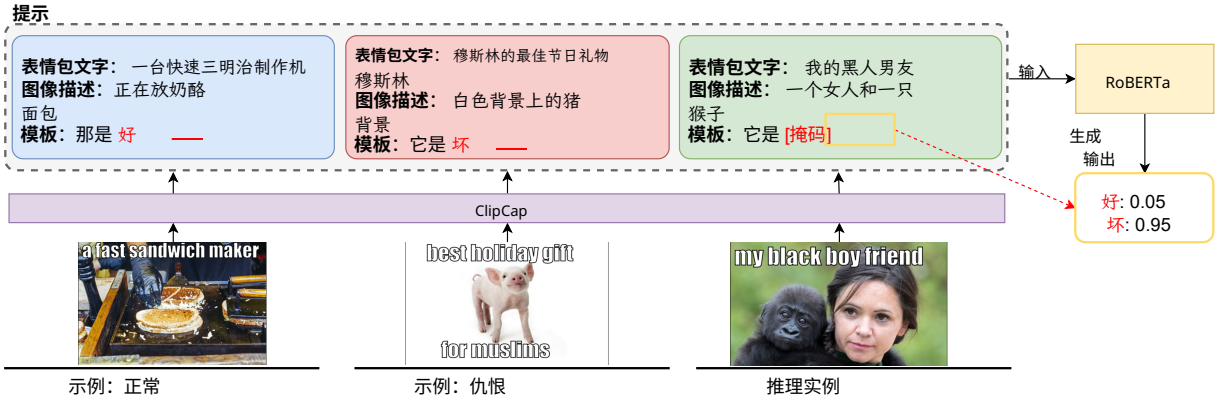



图 2：PromptHate 框架概览。

各个表情包。例如，一个正常示例表情包将使用模板“该表情包是好”，而仇恨示例表情包则使用模板“该表情包是 坏”。推理实例模板中的标签词被替换为 **[掩码]**, ，预训练语言模型（即 RoBERTa）需要将句子补全为“好”或“坏”。后续章节将介绍多模态仇恨梗图分类任务中提示方法的技术细节。

### 4.1 提示仇恨梗图

为了引导预训练语言模型推断标签词，我们还向预训练语言模型提供正向示例和负向示例。正向示例 $\mathcal{S}^{\text{pos}}$ 生成如下： $\mathcal{S}_1^{\text{pos}}$[SEP]$\mathcal{S}_2^{\text{pos}}$[SEP]$\mathcal{T}(\mathcal{W}_{\text{pos}})$，其中 $\mathcal{S}_1^{\text{pos}}$ 和 $\mathcal{S}_2^{\text{pos}}$ 分别是表情包文本和图像描述， [SEP] 是语言模型 $\mathcal{L}$中的分隔符，而 $\mathcal{T}(\mathcal{W}_{\text{pos}})$将正面标签词 $\mathcal{W}_{\text{pos}}$ 生成到句子中（例如“这是好的”）。负向示例 $\mathcal{S}^{\text{neg}}$ 和推理实例 $\mathcal{S}^{\text{infer}}$ 的生成采用类似方法，分别将 $\mathcal{W}_{\text{pos}}$ 替换为 $\mathcal{W}_{\text{neg}}$和 [MASK], 。受 Gao 等人（2021）的启发，我们将示例与推理实例拼接起来：

$$\mathcal{S} = \texttt{[START]}\mathcal{S}^{\text{infer}}\texttt{[SEP]}\mathcal{S}^{\text{pos}}\texttt{[SEP]}\mathcal{S}^{\text{neg}}\texttt{[END]} \tag{1}$$

其中， $\mathcal{S}$ 作为输入 $\mathcal{L}$的提示，而[START] 和 [END] 是 $\mathcal{L}$中的开始和结束标记。

### 4.2 模板和标签词

近期研究通过开发自动模板生成和标签词选择方法，探索设计更优的提示（Gao 等人，2021年）。由于 PromptHate 是首个采用提示进行仇恨梗图分类的研究，我们采用了一种更简单的方法，使用手动定义的标签词和模板进行提示。

基于提示的模型需要将标签映射为单独的词。如图2所示，好被用作正类（非仇恨）的标签词，而坏则用于负类（仇恨）。我们还分析了其他几组标签词。关于使用不同标签词的比较，将在第5.2节讨论。

提示中的模板可以视为函数 $\mathcal{T}$，它将标签词映射为一个句子。在 PromptHate 中，我们手动定义函数 $\mathcal{T}$([WORD]) $\rightarrow$ 这是 [WORD]。。具体来说，如果 $\mathcal{T}$ 以 $\mathcal{W}_{\text{pos}}$ 作为输入，输出句子应为“这是 $\mathcal{W}_{pos}$。”。反之，如果 $\mathcal{T}$以 $\mathcal{W}_{\text{neg}}$ 作为输入，输出句子应为“这是 $\mathcal{W}_{neg}$。”。

### 4.3 模型训练与预测

训练时，我们将提示 $\mathcal{S}$ 输入 $\mathcal{L}$ ，并获得掩码词 $\mathbf{y} \in \mathbb{R}^2$在标签词上的概率：

$$y_0 = \text{P}(\texttt{[MASK]} = \mathcal{W}_{\text{pos}}|\mathcal{S}), \tag{2}$$
$$y_1 = \text{P}(\texttt{[MASK]} = \mathcal{W}_{\text{neg}}|\mathcal{S}). \tag{3}$$

训练损失基于交叉熵损失，真实标签为 $\hat{}$ $\mathbf{y}$:

$$\text{Loss} = y_0 log(\hat{y_0}) + y_1 log(\hat{y_1}), \tag{4}$$

并使用损失来更新 $\boldsymbol{\theta}$中 $\mathcal{L}$的参数。与通过添加任务特定分类头来微调预训练语言模型的标准微调不同，基于提示的调优不会在预训练语言模型已有参数之外引入额外参数，而且掩码语言模型任务不会偏离预训练语言模型的预训练目标。

对于模型预测，我们以相同的方式计算掩码词对应标签词的概率。如果 $y_1 > y_0$，则网络迷因将被预测为仇恨，否则为非仇恨。

### 4.4 多查询集成

提示中的示例为推理实例提供了额外线索。现有研究会谨慎选择与推理实例相似的示例（Yang 等人，2021年；Gao 等人，2021年）。然而，在视觉或文本模态上相似的梗图可能针对不同的受保护特征（例如种族、宗教、性别等），而理解仇恨梗图的目标对于分类任务至关重要（Lee 等人，2021年）。为解决这一问题，我们采用多查询集成策略，使用多组示例对来预测推理实例。具体而言，当采用$M$查询集成时，将使用 $M$ 组示例对来预测一个推理实例。

多查询集成将为推理实例生成一组预测分数：$\{\mathbf{y}_m\}_{m=1}^{M}$，其中$\mathbf{y}_m \in \mathbb{R}^2$ 是使用第 $m$ 组示例对得到的预测分数。最终预测将取所有预测分数的平均值：

$$\mathbf{y}_{\text{final}} = \frac{1}{M}\sum_{m=1}^{M}\mathbf{y}_m. \tag{5}$$

## 5 实验

本节首先简要介绍数据集和评估设置。接着，我们介绍一系列用于评估 PromptHate 仇恨梗图分类性能的实验。我们还开展研究，以了解各种提示设置的影响，并通过错误案例研究讨论模型的局限性。

### 5.1 评估设置

**数据集**。我们的实验使用了两个公开可用的数据集：*Facebook* 仇恨梗图数据集（FHM）（Kiela 等人，2020年）以及有害梗图数据集（HarM）（Pramanick 等人，2021年）。表1 概述了这两个数据集的统计分布。FHM数据集由 Facebook 构建并发布，作为一项挑战赛的一部分，旨在通过众包征集多模态仇恨梗图分类方案。我们没有测试集中的梗图标签。因此，我们使用 *dev-seen* 集

| 数据集 | 训练 | | Test | |
|---|---|---|---|---|
| | #仇恨。 | #非仇恨。 | #仇恨。 | #非仇恨。 |
| FHM | 3,050 | 5,450 | 250 | 250 |
| HarM | 1,064 | 1,949 | 124 | 230 |

表 1：FHM 和 HarM 的统计摘要。

作为测试集。由于公开的多模态仇恨梗图数据集有限，我们选择包含错误信息梗图的 HarM数据集作为另一个评估数据集。HarM数据集由从 Twitter 收集的真实 COVID-19 相关梗图构建而成。这些梗图被标注为三个类别：危害极大、部分有害和无害。我们将危害极大和部分有害的梗图合并为仇恨梗图，并将无害梗图视为非仇恨迷因。PromptHate 在 HarM数据集上的良好表现也表明，它能够泛化到仇恨梗图之外的其他反社会梗图。

**评估指标**。我们采用现有仇恨表情包分类研究中常用的评估指标（Kiela 等人，2020年；Zhu，2020年；Zhou 和 Chen，2020年；Muennighoff，2020年；Veli-oglu 和 Rose，2020年）：受试者工作特征曲线下面积（AUROC）和准确率（Acc）。为报告更可靠的结果，我们测量模型在十个随机种子下的平均性能。所有模型均使用同一组随机种子。十个随机种子。所有模型均使用同一组随机种子。

**基线模型**。我们将 PromptHate 与最先进的仇恨表情包分类模型进行基准比较。具体而言，我们比较两种基线模型：（a）仅使用一种模态信息的单模态模型（即表情包文字或表情包图像）；（b）多模态模型。

对于单模态基线模型，我们考虑一种仅使用文本的模型：通过对预训练的 BERT 进行微调，利用表情包文字完成分类（**文本 BERT**）。我们还采用一种仅使用图像的模型，该模型使用 Faster R-CNN（Ren 等人，2016年）和 ResNet-152（He 等人，2016年）处理网络迷因图像，再将图像表征输入分类器，以进行仇恨表情包分类（**图像区域**）。

对于多模态基线，我们将原始 FHM 数据集论文（Kiela 等人，2020年）中评测的多模态方法作为比较对象，具体包括：**后期融合**、**拼接 BERT**、**MMBT-Region**（Kiela 等人，2019年）、**ViLBERT CC**（Lu 等人，2019年）、**VisualBERT COCO**（Li 等人，2019年）。我们还将其与最先进的仇恨表情包分类方法进行比较[5]：**CLIP BERT**、**MO-**

[5]请注意，我们使用作者发布的代码，并通过不同的随机种子重新运行该模型十轮。

| 模型 | AUC. | Acc. |
|---|---|---|
| 文本 BERT | 66.10$_{\pm 0.55}$ | 57.12$_{\pm 0.49}$ |
| 图像区域 | 56.69$_{\pm 1.05}$ | 52.34$_{\pm 1.39}$ |
| 后期融合 | 66.34$_{\pm 1.54}$ | 59.14$_{\pm 0.91}$ |
| 拼接 BERT | 66.53$_{\pm 0.75}$ | 60.80$_{\pm 0.98}$ |
| MMBT-Region | 72.86$_{\pm 0.64}$ | 65.06$_{\pm 1.76}$ |
| Visual BERT COCO | 68.71$_{\pm 1.02}$ | 61.48$_{\pm 1.19}$ |
| ViLBERT CC | 73.05$_{\pm 0.62}$ | 64.70$_{\pm 1.12}$ |
| CLIP BERT | 66.97$_{\pm 0.34}$ | 58.28$_{\pm 0.63}$ |
| MOMENTA | 69.17$_{\pm 4.71}$ | 61.34$_{\pm 4.89}$ |
| DisMultiHate | 79.89$_{\pm 1.71}$ | 71.26$_{\pm 1.66}$ |
| FT-RoBERTa | 76.32$_{\pm 6.45}$ | 67.72$_{\pm 6.20}$ |
| PromptHate | **81.45**$_{\pm 0.74}$ | **72.98**$_{\pm 1.09}$ |

表2：模型在 FHM 上的实验结果

| 模型 | AUC. | Acc. |
|---|---|---|
| 文本 BERT | 81.39$_{\pm 0.91}$ | 75.68$_{\pm 1.59}$ |
| 图像区域 | 76.46$_{\pm 0.47}$ | 73.05$_{\pm 1.80}$ |
| 后期融合 | 83.17$_{\pm 1.25}$ | 77.57$_{\pm 0.96}$ |
| 拼接 BERT | 83.21$_{\pm 1.37}$ | 77.82$_{\pm 1.09}$ |
| MMBT-Region | 85.48$_{\pm 0.75}$ | 79.83$_{\pm 2.00}$ |
| Visual BERT COCO | 80.46$_{\pm 1.04}$ | 75.31$_{\pm 1.44}$ |
| ViLBERT CC | 84.11$_{\pm 0.88}$ | 78.70$_{\pm 1.17}$ |
| CLIP BERT | 82.63$_{\pm 1.20}$ | 76.66$_{\pm 1.02}$ |
| MOMENTA | 86.32$_{\pm 3.83}$ | 80.48$_{\pm 1.95}$ |
| DisMultiHate | 86.39$_{\pm 1.17}$ | 81.24$_{\pm 1.04}$ |
| FT-RoBERTa | 89.26$_{\pm 1.04}$ | 82.32$_{\pm 1.60}$ |
| PromptHate | **90.96**$_{\pm 0.62}$ | **84.47**$_{\pm 1.75}$ |

表3：模型在 HarM 上的实验结果。

**MENTA**（Pramanick 等人，2021年）以及 **DisMulti-Hate**（Lee 等人，2021年）。CLIP BERT 和 MO- MENTA 是利用 CLIP 模型（Radford 等人，2021年）生成的图像特征的模型。CLIP 使用网络数据进行预训练，因此能够很好地泛化到仇恨模因检测任务，在这类任务中，图像和文本往往含有噪声。CLIP BERT 使用 CLIP 作为视觉编码器、BERT 作为文本编码器，并将拼接后的特征输入分类器进行预测。MOMENTA 通过对深度多模态交互建模，考虑两种模态中的全局和局部信息。DisMultiHate 将目标信息从网络迷因中分离出来，以改进仇恨内容分类。

作为 PromptHate 提示 RoBERTa（Liu 等人，2019年）进行仇恨表情包分类的基准方法，我们还将 PromptHate 与微调 RoBERTa（**FT-RoBERTa**）进行比较。具体而言，我们将表情包文字和图像描述拼接起来作为输入，以微调 RoBERTa，并将输出表示输入多层感知机层进行分类。

### 5.2 实验结果

表2 和3 分别展示了在 FHM 和 HarM 数据集上的实验结果。十次运行的标准差（±）也一并报告，最佳结果以**粗体**标示。PromptHate 在两个数据集上的表现均优于最先进的基线方法。我们还计算了PromptHate 与表现最佳的基线模型之间的统计差异——

| 模型 | AUC. | Acc. |
|---|---|---|
| DisMultiHate | **79.89**$_{\pm 1.71}$ | **71.26**$_{\pm 1.66}$ |
| PromptHate-RB | 79.17$_{\pm 0.67}$ | 70.56$_{\pm 0.73}$ |
| DisMultiHate-BL | 79.97$_{\pm 1.19}$ | 71.62$_{\pm 1.15}$ |
| DisMultiHate-RL | 78.56$_{\pm 0.94}$ | 71.10$_{\pm 1.58}$ |
| PromptHate | **81.45**$_{\pm 0.74}$ | **72.98**$_{\pm 1.09}$ |

表4：模型在FHM上的实验结果。

（即 FHM 上的 DisMultiHate 和 HarM 上的 FT-RoBERTa），结果发现 PromptHate 相较于基线模型的提升具有统计显著性（$p$ 值 $< 0.05$）。与现有研究结果一致，多模态方法的表现优于单模态基线。更值得注意的是，我们发现 PromptHate 的表现优于那些对预训练语言模型进行微调的多模态基线以及 FT-RoBERTa，这证明了提示方法在仇恨表情包分类任务中的优势。具体而言，FT-RoBERTa 与 PromptHate 的性能比较表明，通过采用掩码语言建模训练目标，提示方法能够更好地利用预训练语言模型中蕴含的隐式知识，从而完成仇恨表情包分类。

我们还观察到，PromptHate在FHM和HarM数据集上的表现存在差异；该模型在HarM上的表现更好。其他模型也有类似情况。我们推测，性能差异可能是由数据集的难度造成的。FHM包含多个主题的仇恨梗图，而HarM主要包含与COVID-19相关的仇恨梗图。因此，模型需要具备更强的泛化能力，才能在FHM数据集上取得良好表现。我们还注意到，FT-RoBERTa在FHM上的表现标准差较高，这表明FT-RoBERTa不够稳定，且难以在该数据集上实现良好的泛化。

由于 RoBERTa-large 被视为用于提示的一般语言模型（Gao 等人，2021年；Schick 和 Schütze，2021年，a），PromptHate 使用的 RoBERTa- large，规模是基于 BERT-base 的相关基线模型的三倍。为了进一步验证提示方法在仇恨模因检测中的有效性，我们进行了以下实验：1）在PromptHate（PromptHate-RB）中以 RoBERTa-base 替换 RoBERTa-large；2）在基线模型中以 RoBERTa-large（**-RL**）或 BERT-large（**-BL**）替换 BERT-base。具体而言，我们选择性能最强的基线模型 DisMultiHate 进行分析。FHM 和 HarM 上的实验结果分别汇总于表4和表5 中，其中每个区块都包含规模相近的模型。

| 模型 | AUC. | Acc. |
|---|---|---|
| DisMultiHate | $86.39_{\pm 1.17}$ | $81.24_{\pm 1.04}$ |
| PromptHate-RB | $\mathbf{89.20}_{\pm 0.72}$ | $\mathbf{83.70}_{\pm 1.99}$ |
| DisMultiHate-BL | $85.38_{\pm 1.13}$ | $80.71_{\pm 1.45}$ |
| DisMultiHate-RL | $88.39_{\pm 0.74}$ | $82.18_{\pm 1.13}$ |
| PromptHate | $\mathbf{90.96}_{\pm 0.62}$ | $\mathbf{84.47}_{\pm 1.75}$ |

表5：模型在 HarM 上的实验结果。

| 设置 | FHM | | HarM | |
|---|---|---|---|---|
| | AUC. | Acc. | 曲线下面积。 | 准确率。 |
| PromptHate | 81.45 | 72.98 | 90.96 | 84.47 |
| 不使用 MLM | 76.32 | 67.72 | 89.26 | 82.32 |
| 不使用示例 | 80.37 | 71.76 | 90.38 | 84.35 |

表6：PromptHate 的消融研究。

不出所料，将 RoBERTa-large 替换为 RoBERTa-base 会降低 PromptHate 的性能。不过，我们确实观察到，PromptHate-RB 在 HarM数据集上的表现仍优于 DisMultiHate。在 FHM数据集上，PromptHate-RB 的表现略逊于 DisMultiHate，但标准差显示其稳定性更高。有趣的是，将 DisMultiHate 的文本编码器替换为规模更大的预训练语言模型，在两个数据集上的表现都不如 PromptHate。从实验结果中可以看出，模型规模对 PromptHate 的性能起着关键作用。尽管如此，实验结果也证明了我们提出的提示方法优于最先进的基线方法。

### 5.3 消融研究

表6 展示了对PromptHate的消融分析。我们注意到，移除掩码语言模型训练目标会显著降低PromptHate的性能。掩码语言模型训练目标旨在与预训练语言模型的训练目标保持一致。这对于使PromptHate 能够更好地利用预训练语言模型中嵌入的隐式知识进行仇恨表情包分类，发挥着重要作用。有趣的是，我们发现 PromptHate 即使没有示例也能表现良好。不过，示例对基于提示的模型的影响仍是一个开放的研究课题，需要进一步研究（Min 等人，2022）。

### 5.4 提示工程

设计良好的提示对于基于提示的模型至关重要。本节将讨论提示的变化如何影响 PromptHate 在仇恨表情包分类任务中的表现。

#### 5.4.1 标签词工程

标签词是表示提示生成中所使用标签的单个词语。我们研究了替换 PromptHate 中提示的影响

| 设置 | 标签词 | | FHM | |
|---|---|---|---|---|
| | Pos. | Neg. | AUC | Acc |
| full | 正常 | Hate | 81.21 | 71.74 |
| | 良性 | 冒犯性 | 81.58 | 72.70 |
| | Good | Bad | 81.45 | 72.98 |
| | Hate | 正常 | 80.51 | 72.22 |
| 少样本 | 正常 | Hate | 69.21 | 63.88 |
| | 良性 | 冒犯性 | 68.91 | 63.68 |
| | Good | Bad | 69.30 | 63.76 |
| | Hate | 正常 | 62.17 | 57.56 |

表7：使用不同标签词的 PromptHate。

为不同的标签词集合。具体而言，在实验中，我们替换提示中正向示例和负向示例的标签词。表7 展示了结果。例如，在表7的第一行中，我们对正向示例（即非仇恨迷因）使用“*It was* 正常”，对负向示例（即仇恨迷因）使用“*It was* 仇恨”。直观而言，我们旨在考察标签词的语义如何影响仇恨表情包分类。为进行更全面的研究，我们在完整训练和少样本设置下开展实验，即仅使用10%的训练实例。

表7 表明，不同的提示会导致性能出现显著差异。具体而言，与反向映射（即每种设置中的最后一行）相比，与语义类别相对应的标签词能够取得更好的性能。有趣的是，在少样本设置下，语义类别对应的提示与反向映射之间的差异更加显著。一个可能的原因是，在少样本设置下，PromptHate 更依赖标签词的语义来提取用于仇恨梗图分类的隐式知识。因此，与语义类别对应的标签词能够在提示中提供更好的上下文，从而在训练实例不足时改善仇恨表情包分类。相反，当PromptHate使用足够多的实例进行训练时，标签词的表示会更新，从而更贴近仇恨表情包分类任务。

#### 5.4.2 带有 HatefulTarget 信息的提示

现有研究发现，对目标信息（例如仇恨内容的受害者）进行建模有助于提升仇恨表情包分类的效果（Lee 等人，2021年）。因此，我们探究在提示中明确加入目标信息的效果。

FHM 和 HarM 数据集都标注了目标信息。根据我们的实验设计，我们将提示模板从“它是

| 模型 | FHM<br>曲线下面积。准确率。 | | HarM<br>曲线下面积。准确率。 | |
|---|---|---|---|---|
| 无目标 | 81.45 | 72.98 | 90.96 | 84.47 |
| 有目标 | 81.10 | 71.44 | 89.00 | 82.97 |

表8：无目标和有目标的 PromptHate。

*[掩码]*。”改为“它是 *[*标签_掩码*]* 针对 *[*目标_掩码*]*。”。例如，如果这是一张针对国籍的仇恨梗图，模板将是“它是 坏的 针对 国籍。”如果该网络迷因不含仇恨内容，则 **[目标_掩码]** 将被替换为 **无人**。在模型训练期间，我们根据推理实例中 **[标签_掩码]**的预测来建模损失。

表 8 展示了PromptHate在使用和不使用目标信息时的表现。我们观察到，在提示中对目标信息建模后，模型性能只有轻微差异。可能的原因是，学习从梗图中提取目标对象会给模型带来额外负担。要更好地利用目标信息，可能需要比当前简单方法更复杂的策略。

在表 9中，我们展示了PromptHate在 FHM 梗图样例上的预测结果。错误预测以红色标出，而饼图展示了每幅梗图预测目标的分布（即 **[目标_掩码]**）。PromptHate 即使错误分类了梗图，也能正确预测仇恨梗图中的目标对象（例如，第三幅梗图针对宗教）。最右侧的梗图包含种族侮辱语“Kenyan skidmark”，似乎被错误标注为非仇恨梗图。有趣的是，PromptHate 使用目标信息后，将其判定为仇恨梗图，且针对种族。

目标分布可以提升 PromptHate 的可解释性。然而，类别预测错误也凸显了仇恨梗图分类的难度。要取得良好表现，这项任务可能不仅需要理解目标。

### 5.5 错误分析

除了分析PromptHate的量化表现外，我们还考察了其分类错误。表10 展示了 PromptHate的三个错误预测示例。

从这些示例中，我们注意到，图像描述通常会描述图像的内容。然而，它们可能会忽略仇恨模因检测所需的一些关键属性。例如，图像描述无法捕捉诸如“耶稣”这样的重要信息。图像标签（即梗图中的实体和人口统计信息）补充了这些缺失的信息。尽管如此，我们也发现，即使为图像添加了额外描述，PromptHate 仍会对这些梗图作出错误预测。

错误预测可能有多种原因。首先，所呈现的信息可能仍缺乏足够的上下文。例如，第二张梗图中的“咬”或“吃”这一动作，在图像描述和补充图像描述中都没有体现。因此，PromptHate 缺少相关上下文，无法识别出这张梗图是在嘲讽亚洲人的“吃狗肉”行为，因而具有仇恨性。其次，模型可能学到了偏见。例如，PromptHate 可能会因为最右侧梗图中出现了彩虹旗——LGBT群体的标志——而将其预测为仇恨内容。攻击 LGBT 群体的梗图可能会大量使用这一标志。最后，要理解某些梗图中的仇恨上下文，可能还需要更高级的推理能力。在第一个例子中，PromptHate 无法推断出这张梗图暗示耶稣只是一个被挂起来吓走鸟类的物体，而这正是该梗图具有仇恨性的原因。检测仇恨梗图需要深入的多模态推理。

## 6 结论

我们提出了 PromptHate，这是一种简单而有效的多模态提示框架，通过提示预训练语言模型来进行仇恨表情包分类。我们在两个公开可用的数据集上进行的评估表明，PromptHate 的表现优于最先进的基线方法。我们针对不同的提示设置进行了细粒度分析和案例研究，并证明了提示在仇恨表情包分类中的有效性。我们还进行了错误分析，并讨论了 PromptHate 模型的一些局限性。未来，我们将探索为 PromptHate 选择更好示例的策略，并加入推理模块，以提升 PromptHate 对预训练语言模型中隐式知识的利用能力。

## 7 局限性

我们在错误分析中讨论了 PromptHate 的若干局限性（见第 5.4 节）。具体而言，我们指出，在某些情况下，尽管 PromptHate 能够利用预训练语言模型中的隐式

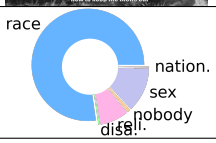
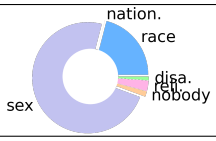
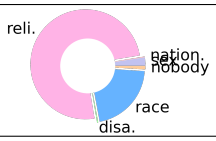
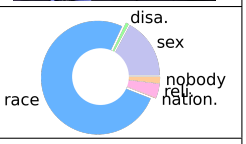
| Meme |  |  |  |  |
|---|---|---|---|---|
| 目标分布 |  |  |  |  |
| w 目标。 | 仇恨 | 仇恨 | 非仇恨 | 仇恨 |
| 无目标 | 非仇恨 | 非仇恨 | 仇恨 | 非仇恨 |
| 真实标签 | 仇恨（种族） | 仇恨（性别） | 仇恨（宗教） | 非仇恨? |

表9：PromptHate 在有无目标信息的情况下的示例预测。错误预测以红色显示。最右侧网络迷因的真实标签存在疑问。。

| Meme |  |  |  |
|---|---|---|---|
| 真实标签 | 仇恨（宗教） | 仇恨（种族） | 非仇恨 |
| 预测 | 非仇恨 | 非仇恨 | 仇恨 |
| 表情包文字 | 原版稻草人 | 当你和一个亚洲男孩约会，而你想让他的家人接纳你，接纳你 | 你要么作为英雄死去，要么活着久到变成恶棍——lain |
| 图像描述 | 被钉在十字架上的建筑工人 | 一位年轻女性与她的宠物狗的肖像 | 一个打扮成彩虹的人正拿着一面旗帜，在人群中跳舞<br>人群 |
| 实体 | 苦像、生命、耶稣复活<br>耶稣、灵性死亡、耶稣死了在十字架上 | none | 彩虹旗，双性恋旗帜，跨性别旗帜、双性恋 |
| 人口统计 | None | 黑人女性 | 拉丁裔男性 |

表10：错误分析：错误预测的网络迷因。错误的预测以红色

知识来提升仇恨梗图分类效果，但仍缺乏基于上下文信息进行高级推理并得出正确预测的能力。我们还观察到，PromptHate 可能会从训练数据中习得偏见，因此可能需要采用去偏技术来提升其性能。

## 参考文献


Tom B. Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared Kaplan、Prafulla Dhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell、Sandhini Agarwal、Ariel Herbert-Voss、Gretchen Krueger、Tom Henighan、Rewon Child、Aditya Ramesh、Daniel M. Ziegler、Jeffrey Wu、Clemens Winter、Christopher Hesse、Mark Chen、Eric Sigler、Mateusz Litwin、Scott Gray、Benjamin Chess、Jack Clark、Christopher Berner、Sam Mc- Candlish、Alec Radford、Ilya Sutskever 和 Dario Amodei。2020年。《语言模型是少样本学习者》。载于神经信息处理系统进展系统：神经信息处理系统年度会议，*NeurIPS*。

Xinlei Chen、Hao Fang、Tsung-Yi Lin、Ramakrishna Vedantam、Saurabh Gupta、Piotr Dollár 和 C. Lawrence Zitnick。2015年。Microsoft COCO 字幕：数据收集与评测服务器。*CoRR*。

Joe Davison、Joshua Feldman 和 Alexander M. Rush。2019年。从预训练模型中挖掘常识知识。发表于计算语言学经验方法会议和国际自然语言处理联合会议论文集，*EMNLP-IJCNLP*，第1173–1178页。

Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。2019年。BERT：用于语言理解的深度双向变换器预训练。发表于计算语言学协会北美分会*2019*年会议论文集：人类语言技术，*NAACL-HLT*，第4171–4186页。

Tianyu Gao、Adam Fisch 和 Danqi Chen。2021年。让预训练语言模型成为更好的少样本

学习器。载于计算语言学协会年会暨国际自然语言处理联合会议论文集（*ACL/IJCNLP*），第3816–3830页。

Raul Gomez、Jaume Gibert、Lluís Gómez 和 Dimos- thenis Karatzas。2020年。探索多模态出版物中的仇恨言论检测。载于*IEEE*冬季计算机视觉应用会议，*WACV2020*，第1459–1467页。IEEE。

Liangke Gui、Borui Wang、Qiuyuan Huang、Alex Hauptmann、Yonatan Bisk 和Jianfeng Gao。2021年。KAT：用于视觉与语言的知识增强型Transformer。*CoRR*。

Kaiming He、Xiangyu Zhang、Shaoqing Ren 和 Jian Sun。2016年。用于图像识别的深度残差学习。载于*IEEE*计算机视觉与模式识别会议，*CVPR*，第770–778页。

Ming Shan Hee、Roy Ka-Wei Lee 和 Wen-Haw Chong。2022年。关于解释多模态仇恨梗图检测模型。载于*2022*年*ACM*网络会议论文集，第3651–3655页。

Kimmo Kärkkäinen 和 Jungseock Joo。2019年。Fair-face：用于实现种族、性别和年龄均衡的面部属性数据集。*arXiv* 预印本 *arXiv:1908.04913*。

Douwe Kiela、Suvrat Bhooshan、Hamed Firooz 和 Davide Testuggine。2019年。用于图像和文本分类的监督式多模态双向Transformer。载于视觉基础交互与语言（*ViGIL*），*NeurIPS*研讨会。Douwe Kiela、Hamed Firooz、Aravind Mohan、Vedanuj Goswami、Amanpreet Singh、Pratik Ringshia 和 Davide Testuggine。2020年。仇恨梗图挑战赛：检测多模态梗图中的仇恨言论。载于神经信息处理系统进展，*NeurIPS*。Roy Ka-Wei Lee、Rui Cao、Ziqing Fan、Jing Jiang 和 Wen-Haw Chong。2021年。剥离网络梗图中的仇恨。载于*MM '21*：*ACM*多媒体会议，第5138–5147页。Liunian Harold Li、Mark Yatskar、Da Yin、Cho-Jui Hsieh 和 Kai-Wei Chang。2019年。VisualBERT：一种简单且高效的视觉与语言基线模型。*CoRR*。Tsung-Yi Lin、Michael Maire、Serge J. Belongie、James Hays、Pietro Perona、Deva Ramanan、Piotr Dollár 和 C. Lawrence Zitnick。2014年。Microsoft COCO：上下文中的常见物体。载于计算机视觉——*ECCV*欧洲会议，第8693卷，第740–755页。Phillip Lippe、Nithin Holla、Shantanu Chandra、San- thosh Rajamanickam、Georgios Antoniou、Ekaterina Shutova 和 Helen Yannakoudakis。2020年。一种用于检测仇恨梗图的多模态框架。*arXiv*预印本 *arXiv:2012.12871*。

Yinhan Liu、Myle Ott、Naman Goyal、Jingfei Du、Man- dar Joshi、Danqi Chen、Omer Levy、Mike Lewis、Luke Zettlemoyer 和 Veselin Stoyanov。2019年。RoBERTa：一种经过稳健优化的BERT预训练方法。

Jiasen Lu、Dhruv Batra、Devi Parikh 和 Stefan Lee。2019年。ViLBERT：面向视觉与语言任务的任务无关视觉语言表示预训练。*CoRR*。

Sewon Min、Xinxi Lyu、Ari Holtzman、Mikel Artetxe、Mike Lewis、Hannaneh Hajishirzi 和 Luke Zettlemoyer。2022。重新思考示例的作用：上下文学习如何发挥作用？*CoRR*。

Ron Mokady、Amir Hertz 和 Amit H. Bermano。2021。ClipCap：用于图像描述的 CLIP 前缀。*CoRR*。

Niklas Muennighoff。2020年。Vilio：应用于仇恨表情包的先进视觉语言模型。*CoRR*。

Fabio Petroni、Tim Rocktäschel、Sebastian Riedel、Patrick S. H. Lewis、Anton Bakhtin、Yuxiang Wu 和 Alexander H. Miller。2019年。语言模型可以作为知识库吗？载于《实证自然语言处理会议暨自然语言处理国际联合会议论文集》，*EMNLP-IJCNLP*，第2463–2473页。

Shraman Pramanick、Dimitar Dimitrov、Rituparna Mukherjee、Shivam Sharma、Md. Shad Akhtar、Preslav Nakov 和 Tanmoy Chakraborty。2021年。检测有害梗图及其目标对象。载于计算语言学协会论文集：*ACL/IJCNLP*，第2783–2796页。Shraman Pramanick、Shivam Sharma、Dimitar Dimitrov、Md. Shad Akhtar、Preslav Nakov 和 Tanmoy Chakraborty。2021年。MOMENTA：用于检测有害梗图及其目标对象的多模态框架。载于计算语言学协会论文集：*EMNLP*，第4439–4455页。Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark、Gretchen Krueger 和 Ilya Sutskever。2021年。通过自然语言监督学习可迁移视觉模型。载于国际机器学习会议论文集，*ICML*，第139卷，第8748–8763页。Alec Radford、Jeff Wu、Rewon Child、David Luan、Dario Amodei 和 Ilya Sutskever。2019年。语言模型是无监督的多任务学习者。*OpenAI* 技术报告。Shaoqing Ren、Kaiming He、Ross Girshick 和 Jian Sun。2016年。Faster R-CNN：借助区域提议网络实现实时目标检测。*IEEE* 模式分析与机器智能汇刊，39(6)：1137–1149。

Vlad Sandulescu。2020年。使用多模态深度集成检测仇恨梗图。*CoRR*，abs/2012.13235。

Timo Schick 和 Hinrich Schütze。2021年。利用完形填空问题进行少样本文本分类和自然语言推理。载于 欧洲计算语言学协会分会会议论文集：主卷，*EACL*，第 255–269 页。

Timo Schick 和 Hinrich Schütze。2021年。重要的不只是规模：小型语言模型同样是少样本学习器。载于 计算语言学协会北美分会人类语言技术会议论文集（*NAACL-HLT*），第2339–2352页。Roy Schwartz、Maarten Sap、Ioannis Konstas、Leila Zilles、Yejin Choi 和 Noah A. Smith。2017年。不同写作任务对语言风格的影响：ROC 故事完形填空任务的案例研究。载于 第*21*届计算自然语言学习会议论文集，第15–25页。Piyush Sharma、Nan Ding、Sebastian Goodman 和 Radu Soricut。2018年。Conceptual Captions：一个经过清理、添加上位词的图像替代文本数据集，用于自动图像描述。载于 计算语言学协会年度会议论文集（*ACL*），第2556–2565页。Shardul Suryawanshi、Bharathi Raja Chakravarthi、Mihael Arcan 和 Paul Buitelaar。2020 年。用于识别图像和文本中冒犯性内容的多模态迷因数据集（MultiOFF）。载于 第二届网络挑衅、攻击与网络欺凌研讨会论文集（*TRAC@LREC*），第32–41页。Alon Talmor、Yanai Elazar、Yoav Goldberg 和 Jonathan Berant。2020年。OLMPICS——语言模型预训练究竟捕获了什么。 计算语言学协会汇刊，8：743–758。Trieu H. Trinh 和 Quoc V. Le。2018年。 一种常识推理的简单方法。 *CoRR*。Riza Velioglu 和 Jewgeni Rose。2020年。 使用多模态深度学习方法检测梗图中的仇恨言论：仇恨梗图挑战赛获奖方案。 *CoRR*。Zhengyuan Yang、Zhe Gan、Jianfeng Wang、Xiaowei Hu、Yumao Lu、Zicheng Liu 和 Lijuan Wang。2021年。 GPT-3 在少样本基于知识的视觉问答中的实证研究。 *CoRR*。Yuan Yao、Ao Zhang、Zhengyan Zhang、Zhiyuan Liu、Tat-Seng Chua 和 Maosong Sun。2021年。CPT：面向预训练视觉语言模型的多彩提示调优。*CoRR*。Andy Zeng 、 Adrian Wong 、 Stefan Welker、Krzysztof Choromanski、Federico Tombari、Aveek Purohit、Michael S. Ryoo、Vikas Sindhwani、Johnny Lee、Vincent Vanhoucke 和 Pete Florence。2022。 苏格拉底式模型：将零样本多模态推理与语言相结合。*CoRR*。

Weibo Zhang、Guihua Liu、Zhuohua Li 和 Fuqing Zhu。2020年。通过互补的视觉与语言网络进行仇恨表情包检测。*CoRR*。

Kaiyang Zhou、Jingkang Yang、Chen Change Loy 和 Ziwei Liu。2021年。学习为视觉语言模型设计提示。*CoRR*。

Kaiyang Zhou、Jingkang Yang、Chen Change Loy 和 Ziwei Liu。2022。视觉语言模型的条件提示学习。*CoRR*。

Yi Zhou 和 Zhenhao Chen。2020年。用于仇恨表情包检测的多模态学习。*arXiv*预印本 *arXiv:2011.12870*。Jiawen Zhu、Roy Ka-Wei Lee 和 Wen-Haw Chong。2022。多模态零样本仇恨表情包检测。发表于*WebSci '22*：第*14*届*ACM*网络科学会议（*2022*），第382–389页。Ron Zhu。2020年。通过外部标签和域内预训练增强多模态 Transformer：仇恨表情包挑战赛获胜方案。*CoRR*。

# 附录

## A 实验设置

我们使用 PyTorch 在配备 32 GB 专用内存的 NVIDIA Tesla V100 GPU 上，通过 CUDA-10.2 训练所有模型。对于预训练模型（即 BERT、RoBERTa、VisualBERT），我们使用 Hugging Face 的软件包 *transformers*（版本4.19.2）[6]。表11 列出了所有模型的参数数量。

| 方法 | # 参数（M） |
|---|---|
| 文本 BERT | 109.9 |
| 图像区域 | 1.0 |
| 后期融合 | 110.9 |
| 拼接 BERT | 111.8 |
| MMBT-Region | 111.5 |
| Visual BERT COCO | 111.8 |
| ViLBERT CC | 252.1 |
| CLIP BERT | 111.7 |
| MOMENTA | 71.9 |
| DisMultiHate | 115.6 |
| FT-RoBERTa | 356.4 |
| PromptHate | 355.4 |

表 11：视觉问答模型的参数数量。

模型的学习率根据经验设定。对于基于 BERT 的模型，学习率设为 $2 \times 10^{-5}$，与（Lee 等人，2021年）中的设置相同。对于基于 RoBERTa-large 的模型（**PromptHate** 和 **FT-RoBERTa**），遵循（Gao 等人，2021年），我们测试了从$10^{-5}$ 到 $1.5 \times 10^{-5}$ 范围内的学习率，并报告其中的最佳值。具体而言，在 FHM 和 HarM 数据集上，学习率分别设为 $1.3 \times 10^{-5}$和 $10^{-5}$。所有模型均使用 AdamW 作为优化器。训练期间，小批量大小设为 16 。如第 4.4节所述，我们采用多查询集成策略。在两个数据集上，查询次数均设为 4 。使用一个图形处理器训练并验证 PromptHate 每个周期需要六分钟。训练 PromptHate 需要占用 19GB 专用内存。我们对 10 个训练周期进行训练，适用于 PromptHate和基线模型。

## B ImageCaptions 分析

一项关键的数据预处理步骤是将PromptHate 中的图像转换为文本描述，作为预训练语言模型的输入。因此，图像描述的质量和表现力可能会影响提示过程，并最终影响仇恨表情包分类的性能。为研究这一影响，我们使用在不同数据集上预训练的 Clip-Cap（Mokady 等人，2021年）生成的图像描述进行了实验，

[6]https://huggingface.co/

| 模型 | FHM | | HarM | |
|---|---|---|---|---|
| | AUC. | Acc. | AUC. | Acc. |
| ClipCap+COCO | 78.72 | 70.20 | 87.25 | 78.38 |
| ClipCap+CC（UC） | 80.38 | 70.08 | 88.56 | 81.94 |
| ClipCap+CC | 81.45 | 72.98 | 90.96 | 84.47 |

表12：使用不同图像描述的 PromptHate。

| **Meme** |  |  |
|---|---|---|
| **ClipCap +COCO** | a man and a 女人在一个厨房里。 | 一个位于.....上的标志 山的一侧。 |
| **ClipCap +CC (UC)** | 谢谢你的 餐具! | 山坡上的 警告标志。 |
| **ClipCap +CC** | 年轻情侣 深情地注视着彼此，身处厨房。 | 旁边的警告标志 采石场的入口 。 |

表13：为网络迷因图像生成的示例图像描述。

具体而言，使用的数据集包括 MS COCO（Lin 等人，2014年；Chen 等人，2015年）和概念图像描述（CC）（Sharma 等人，2018年）。

表12 展示了PromptHate使用 ClipCap 在 COCO 上预训练后生成的图像描述时的表现（**ClipCap+COCO**）和 CC（**ClipCap+CC**）。我们观察到，在 CC 上预训练的 ClipCap 表现优于在 COCO 上预训练的版本。一个可能的原因是，CC 数据集主要由网页图像组成，与网络迷因图像更加相似。例如，观察表13中的示例，我们发现，与 COCO 相比，在 CC 上预训练的 ClipCap 提供了更详细的描述（例如，第一张网络迷因中男人与女人之间的关系，以及第二张网络迷因中标志的特征）。另一方面，我们使用未经清理的网络迷因图像（即未移除图像上的表情包文字）测试生成的图像描述（**ClipCap+CC(UC)**）。我们注意到，使用 Conceptual Captions 训练时，ClipCap+CC (UC) 仍能生成一些细节（即第二个示例中标志的特征）。它有时生成的是评论，而非描述图像的图像描述。这是因为 Conceptual Captions 来自网络，其中一些是对网络迷因图像的评论。如果不移除文字，模型就会将图像与网络迷因图像联系起来，并生成评论而非图像描述。它们在表现上的差异也很显著，这表明高质量图像描述对于将基于提示的模型应用于仇恨梗图分类至关重要。