arXiv:2303.01781v1 [cs.CV] 3 Mar 2023

采用多模态空间编码和人脸嵌入增强的表情包情感分析*

Muzhaffar Hazman[1][0000−0001−8262−2476]、
Susan McKeever[2][0000−0003−1766−2441]和
Josephine Griffith[1][0000−0002−1560−1867]

[1] 戈尔韦大学，戈尔韦，爱尔兰
{m.hazman1, josephine.griffith}@universityofgalway.ie[2]
都柏林理工大学，都柏林，爱尔兰susan.mckeever@TUDublin.ie

摘要。 互联网表情包的特点是在视觉元素之间穿插文本。尽管文本和视觉元素的排布位置蕴含着意义，当前最先进的多模态模因分类器却没有考虑这两种模态中各元素的相对位置。基于两个表情包情感分类数据集，我们系统地展示了将从表情包中提取的视觉对象、人脸和文本簇的空间位置纳入考量后，模型性能得到提升。此外，我们还展示了人脸嵌入可以显著增强多模态模因分类器的图像表征。最后，我们证明，纳入这些空间信息后，全自动方法的表现优于相应的基线；后者需要对光学字符识别提取的文本进行额外的人工验证。

关键词： 多模态深度学习 · 情感分析 · 互联网表情包。

1 引言

情感极性分类任务传统上是分析一段自然语言文本，将其情感分类为负面、正面或中性。情感分析最初是在文本上进行的。用户生成的多模态内容（例如视频、图像-标题对）的增长，推动了将情感计算技术扩展到文本之外的输入类型 [8]。多模态情感分析提出的问题与其纯文本前身相同，但扩展到了同时包含多个模态的输入。面对多模态输入时，Poria等人 [8] 将单模态编码器描述为多模态系统的关键构建模块，每个编码器都直接影响最终的

*本研究得到了爱尔兰科学基金会数字增强现实研究培训中心（d-real）资助（项目编号：18/CRT/6224）。

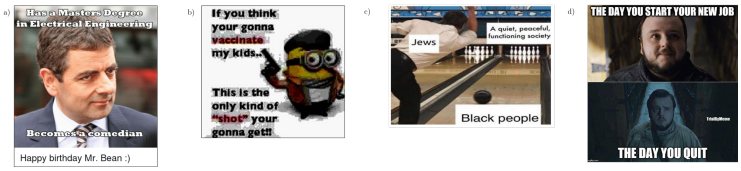

图1：模因示例，其中 a) 正面情感[7] 和 b) 负面情感[7],c) 仇恨性空间类比[12], 和 d) 空间片段[13]。

性能。此外，单模态表征的融合也发挥着关键作用，因为它能向分类器提供“额外信息” [8]。

随着用户生成内容中其他多模态格式的兴起，互联网表情包（或简称“模因”）不断涌现。模因常见于各种在线社群，用于传达想法、引发幽默并表达情绪。对模因进行自动化分析有助于：将模因纳入自动化观点挖掘流程 [8],；采取措施应对基于模因的仇恨言论 [12,6], ；识别虚假信息宣传活动 [1], ；以及研究社会与政治文化 [5]。本研究聚焦于一个基础问题：**模因的情感极性分类**：“给定一个视觉格式的模因，其中包含一幅图像 I 及嵌入文本 T，将该模因的整体情感分类为负面（例如图1b）、正面（例如图1a）或中性。”

网络迷因是自动情感分类问题中颇具挑战性的输入，因为它们通常包含极短的文本、对流行文化的引用、微妙的跨模态语义关系，以及对背景语境的依赖 [10,12,16,12]。因此，解决方案必须考虑每种模态的语义、文本和视觉模态，以及它们的组合 [6]。这一挑战涵盖各种情感目标，包括情感极性 [7,13], 冒犯性 [6,7,13],讽刺 [7,13], 和动机意图 [7,13]。

近期研究表明，纳入其他相关信息能够提升模因情感分类器的性能 [10], ，其中包括文本中词语的位置，以及图像中视觉对象的位置[12,16]。与许多其他形式的多模态内容不同，模因中的文本交错分布在图像中，通常叠加在图像上，或构成模因图像的一部分，从而形成共享的视觉媒介。模因创作者会有意安排词语组合（“文本簇”）的位置来传达含义，例如暗示仇恨类比 [12] （例如图 1c）；文本簇还可以与图像片段配对，每对都表达不同的情感（例如图 1d）。当前利用位置信息进行模因情感分类的方法往往忽略跨模态位置关系，即只考虑词语在文本中的位置，却不考虑其相对于模因图像的位置，反之亦然。

本文提出将模因两种模态中特征的空间信息注入深度学习多模态分类器，以提升情感分类性能。尤其是，我们考虑了视觉对象与



图2：模因示例 [7] a) 展示了数据集中给出的图像和文本模态， I和 T；b) 为局部特征生成的边界框：文本簇（t_1 和 t_2）、对象（o_1 检测为“狗”， o_2 检测为“人物”）以及人脸（f_1）；c) 用于为每个边界框生成空间编码的坐标系（例如， o_1、 p_{o_1}的顶点）。

文本簇交错排列的情况，通过在共享坐标系中表示各自的空间位置（“空间编码”）。在进行多模态融合和分类之前，我们将视觉对象（例如图2b中的 o_1， o_2 ）、人脸（例如图2b中的 f_1 ）和文本簇（例如图2b中的 t_1， t_2 ）的空间编码附加到其局部表示上。我们使用第3.2节所述的七种模型，在两个基准数据集上系统评估空间编码和局部表示对性能的影响。据我们所知，本文首次采用共享坐标空间编码和人脸的深层表示来处理网络迷因的情感分类问题。

2 相关工作

2.1 模因情感分类器

模因在若干关键方面不同于其他类型的多模态用户生成内容。首先，模因的文本和图像共享一种视觉媒介，这与更常见的图像-标题对不同。模因中的文本通常会被刻意放置在其他视觉内容之间，以创造意义 [12]。其次，模因使用简短的文本片段和少量前景视觉对象，依靠模态间关系来传达意义。Kiela 等人 [6] 展示了无害的图像和文本如何组合成仇恨模因。此外，对任一模态稍作改动，都可能使仇恨模因变为无害模因，反之亦然。因此，模因分类器必须能够在输入非常有限的情况下学习细微的模态间关系。

从架构角度来看，现有文献表明，各种情感分类任务均可应用于网络迷因，而无需采用截然不同的方法。最明显的例子是 Bucur 等人的 [3] Memotion

2022 年挑战赛 [7],获胜方案，该方案经过训练，可同时对情感极性、冒犯性、讽刺、幽默和动机意图进行分类。他们的研究结果表明，模因分类架构能够适应不同的情感计算任务。此外，Pramanick 等人 [9],报告了在 Memotion 1.0 数据集[13],上表现最佳的情感分类方案，并指出，当分别针对八个情感维度进行训练时，同一架构的表现优于所有竞争方案，或仅次于其中一个。

构建多模态模因分类器的一种典型方法，是先为每种模态生成单模态表征，再将这些表征融合为模因的多模态表征，例如 [3,10,12,9] 所示。此外，相关文献展示了用于视觉模态和文本模态的各种深度学习表征 [6,7,13], ，但没有明确证据表明其中任何一种方案能够始终优于其他所有方案。

2.2 位置编码

位置编码在 Transformer 架构中发挥着核心作用 [14] ，并已广泛应用于解决各种自然语言任务。它描述了词元在输入中的位置，例如词语在句子中的位置或区域在图像中的位置。然而，由于大多数多模态模因分类器采用单模态编码器，文本元素和视觉元素的位置会被分别编码。

据我们所知，在共同空间坐标系上共享文本模态与图像模态的位置编码（即“空间编码”）尚未应用于模因情感分类。据报道，用于学习模因情感分类的架构中，没有任何一种在 [13] 和 [7] 中使用了基于模态间共享坐标系的位置信息。此外，我们也未能找到能够直接支持这种共享编码的预训练多模态 Transformer。

在这项任务中，Pramanick 等人 [9] 通过将文本模态划分为文本簇，展示了性能提升，但没有明确表示每个文本簇的空间位置。为了对仇恨模因进行分类，Zhu [16] 采用图像块检测器，将每个模因划分为“图像区域”。随后，他们为每个文本词元附加其周围图像块的表示。然而，他们没有给出仅由这一方法带来的性能提升。此外，我们认为，在同一个图像块内放置多个文本簇（例如图 1c），或某个图像块仅由文本组成（例如图 1a）的情况下，这种基于图像块的位置定义并不适用。

Shang 等人 [12] 提出了一种更通用的空间位置表示方法：在输入基于 [?] 设计的跨模态协同注意力池化模块之前，为提取出的视觉对象和文本簇附加空间编码。他们认为，该模型优于其他领先的仇恨模因分类器，是因为它“意识”到了冒犯性跨模态类比：有意将文本簇放置在视觉对象附近，以此表现冒犯性的概念比较。尽管他们的方法仅基于冒犯性空间类比，但我们认为，这种方法还可以捕捉

更广泛的跨模态空间关系类别，包括 Pramanick 等人 [9] 和 Zhu 的 [16] 方法所捕捉的关系。

2.3 视觉特征表示

尽管图像模态通常通过将整个表情包图像输入图像编码器 [7], 来表示，但事实证明，结合提取出的视觉对象的表示能够提升仇恨模因的分类效果[10,12,16]。一种方法是将表情包图像输入 Google Cloud Vision API 的网页实体检测功能，以创建相应的文本格式描述或属性集合 [10,16]。Zhu [16] 还证明，使用预训练的 FairFace 分类器为每张人脸添加种族和性别标签，能够进一步提升性能。Pramanick 等人 [10] 还表明，使用 VGG-19 表示裁剪后的视觉对象和人脸图像可以改善性能。Shang 等人 [12] 还发现，当同时提供全局和局部视觉特征表示时，他们的多模态分类器表现最佳。

利用人脸传达情感并非新颖做法，也并非网络迷因所独有。首先，视觉情感分析 [15] 指出，面部表情是对社交网络图像所传达情感进行分类时一种有价值的中层特征。其次，研究表明，面部表情表情符号能够为文本社交媒体内容的情感分类提供有用支持 [2]。在网络迷因中，Zhu [16]认为，鉴于当前迷因数据集的规模，期待全局图像编码充分识别能够预测仇恨性的面部特征是不合理的。尽管我们认同 Zhu 的观点，但我们认为其方法忽略了人脸所传达的其他可能表明迷因情感的信息，例如情绪、表情和身份。

3 方法论

在本研究中，我们评估了七种新型多模态分类模型的性能。这些模型分别在两个竞赛数据集 Mem- otion 1.0 [13] 和 Memotion 2.0, [7], 上训练，用于对网络迷因的情感极性进行分类。我们首先设计并评估了一种多模态深度学习模型，以建立基线性能。随后，我们逐步扩展该模型，以回答研究问题。扩展内容包括纳入人脸、视觉对象和文本簇的空间信息，各模型的具体情况见表 3。评估依据分别纳入和排除各项扩展的一对模型在宏平均和加权平均 F1 分数上的差异进行——这些指标由数据集作者规定[7,13]——本节将详细介绍所用的数据集和模型。

3.1 数据集与特征提取

本研究采用了 SemEval 2019 Memotion 1.0 [13]（“**Memo1**”）和 AAAI 2022 Memotion 2.0 [7] 挑战赛（“**Memo2**”）中介绍的数据集。两者

表1：每个数据集的样本数。

数据集	Memo1			Memo2		
	训练集	验证集	测试集	训练集	验证集	测试集
`Original`						
正面	4,156	–	1,099	1517	325	78
中性	2,204	–	584	4510	975	971
负面	631	–	172	973	200	451
总计	**6,991**	–	**1,855**	**7,000**	**1,500**	**1,500**
`Filtered & Filtered-OCR`						
正面	3,450	609	1,067	1,453	192	76
中性	1,837	324	572	4,363	951	939
负面	518	92	169	941	317	442
总计	**5,805**	**1,025**	**1,808**	**6,757**	**1,460**	**1,457**

都是由用户生成的模因组成的集合，并依据三种互斥的情感类别之一进行标注。数据集作者使用自动化 OCR 工具从每个模因中提取文本，然后手动纠正任何错误的文本提取结果。在我们的实验中，Memo1 和 Memo2 中的样本保持分开。未经筛选或预处理，这些样本构成了我们的 `Original`数据集，用于将我们的 `Baseline` 模型与领先的解决方案进行比较。

对于这些数据集中的每个网络迷因，我们使用图 3 所列的工具对其中的文本簇、人脸和视觉对象进行定位、提取与表示。

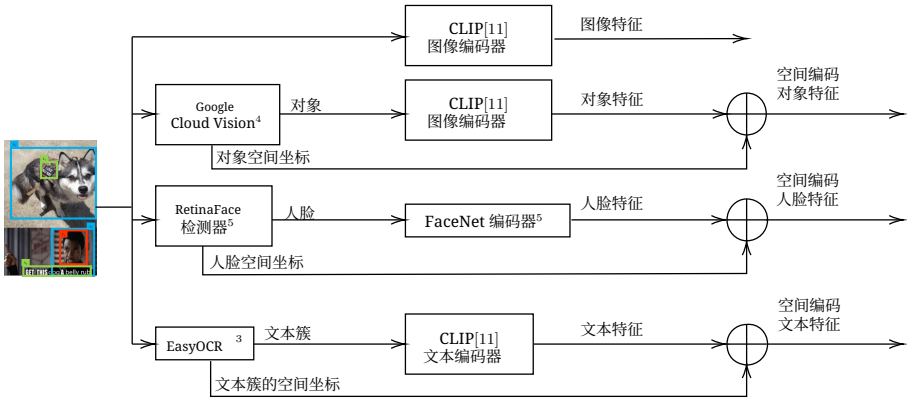

图3：对每个模因应用定位与表示过程，以提取其图像、对象、人脸和文本特征。

[3]https://github.com/JaidedAI/EasyOCR； `paragraph` 选项设为 true。[4] https://cloud.google.com/vision/docs/object-localizer [5]使用来自 https://github.com/serengil/deepface 的 DeepFace 封装

文本簇、视觉对象和人脸的最大数量分别设为 18、10 和 5；对于数量较少的网络迷因，则使用填充。文本簇的填充方式是将空字符串传入 CLIP 文本编码器，而视觉对象的填充方式是使用空白图像的 CLIP 编码；对于人脸，则使用零填充。

由于本研究适用于包含可识别视觉对象和文本簇的网络迷因，因此我们移除了不符合这些标准的网络迷因样本，以构成 `Filtered` 数据集。我们对 Memo1 和 Memo2 的所有子集都进行了此筛选。由于 Memo1 没有指定的验证集，我们按照 Memo1 数据集的作者报告并用于其竞赛投稿的方式 [13] ，以随机 85:15 的比例拆分训练集，并根据情感类别加权，以维持目标分布。我们保留了 Memo2[7]既定的训练集-验证集-测试集划分。对于包含可识别视觉对象但未检测到人脸的网络迷因样本，其人脸特征表示完全由填充内容构成。

最后， `Filtered-OCR` 数据集会将 `Filtered`中每个模因的文本替换为我们在特征提取的光学字符识别步骤中返回的内容。与 [7,9,13],不同，我们在光学字符识别提取过程中未进行任何额外的人工验证。所有模型都在生成的 `Filtered-OCR`数据集上进行训练、验证和测试。各数据集中模因的数量及情感标签见表1。

3.2 模型

本节介绍表 2 所列、图 4 所示模型的架构特征。所有模型均使用 PyTorch 构建，并采用三角循环学习率调度，学习率在 1e-4 至 1e-3 之间变化，步长为 52 个小批次，每批包含 512 个样本。训练期间，我们监控验证集表现，以判断是否出现过拟合；每个模型最多训练 100 个训练轮次。训练使用 AdamW 优化器，权重衰减为 $5e-1$，betas 为 0.1 和 0.25，并通过类别权重（与训练数据集中各类别的样本数量成反比）最小化负对数似然损失。所有非预训练权重均采用均值为零、标准差为 0.02 的高斯分布进行初始化，而预训练权重不作微调。所有模型均使用相同的超参数设置，并分别在各自的数据集上训练。

主流模因情感分类器采用了多种架构，但鲜有迹象表明哪种架构最优。对于我们的 `Baseline` 模型，我们借鉴了 Memotion 2.0 挑战赛 [7]中领先方案的典型整体方法：使用预训练编码器表示每种模态。随后融合这些表示（通常采用多模态注意力机制），最后将其输入全连接层。

为了在我们的 `Baseline`模型中对模因图像和文本进行编码（见图2a中的 I 和 T），我们分别选择使用 CLIP 的预训练图像编码器和文本编码器 [11]，其性能已被证明可与其他多模态方法相媲美 [10]。此外，已有研究表明，CLIP 图像编码的表现优于

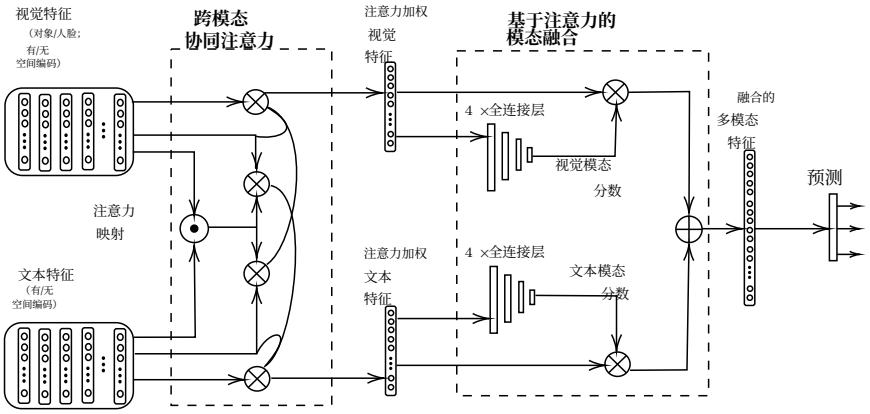

图 4：我们的 `Obj-NoSpatial`、 `Obj-Spatial`、 `Face-NoSpatial` 和 `Face-Spatial` 模型的架构。在 `Img-Obj-Spatial` 和`Img-Face-Spatial` 模型中使用的图像特征绕过了跨模态协同注意力模块，因此需要为基于注意力的模态融合模块增加另一组全连接层。本研究的 `Baseline` 模型不包含跨模态协同注意力模块。来源：跨模态协同注意力 [**?** ， 12]；基于注意力的模态融合 [4,9]

多种其他图像编码器，在仇恨模因的零样本分类任务中表现更佳 [11]，且 Memo2 挑战赛的获胜方案也采用了 CLIP [3]。我们选择了 CLIP 的 ViT–B/16 变体，而 Pramanick 等人 [10] 和 Bucur 等人 [3] 没有说明所选用的变体。

鉴于注意力融合已被证明在若干模因问题上表现良好 [9], ，我们在模型中加入了这种方法。我们的 `Baseline` 模型采用 Gu 的 [4] 注意力模态融合机制，对模因图像的 CLIP 表征与文本进行融合，该机制已用于 [10]。我们将四个全连接层的大小设为 256、64、8 和 1，从而为每个模态生成一个注意力分数。随后，将每种模态经注意力加权后的表征拼接起来，并输入一个由 GeLU 激活的全连接层，再经过对数 softmax 激活，输出各情感类别的预测 logits。

该模型在 `Original` 数据集上训练，以便与先前发表的研究进行性能比较。随后，我们在 `Filtered` 和 `Filtered-OCR` 数据集上评估了该模型。对于后者，所有文本簇 t_n 的内容会先拼接起来，再输入文本编码器。该模型在这两个数据集上的性能差异，使我们能够衡量相较于数据集作者采用的人工整理方法 [7,13]，我们基于 OCR 的文本提取结果对性能造成的影响。

`Obj-NoSpatial` 和 `Face-NoSpatial` 模型移除了模因图像和文本， I 和 T 则针对每个 `Baseline` 进行处理。作为输入，前者采用 CLIP 编码的视觉对象， o_1， o_2，..... o_j，以及从模因中提取的文本簇， t_1， t_2，..... t_{io} `Face-NoSpatial` 模型不使用对象，而是采用 FaceNet 对人脸的表示——

， f_1， f_2，..... f_k。接着，将 j 视觉对象或 k 人脸表示与 [12] 中使用的文本簇 i 进行协同注意力加权池化，但不使用空间编码。此步骤使模型能够学习每个对象/人脸与每个文本簇之间的注意力图，从而生成表示各个模态的一维向量。该表示取代图像模态表示，作为 Baseline 模型所述注意力融合机制的输入。

Obj-Spatial 和 **Face-Spatial** 模型分别为每个文本簇 p_{t_i}、视觉对象p_{o_j}以及人脸 p_{f_k}引入空间编码。我们将 Obj-NoSpatial和 Face-NoSpatial 中的协同注意力池化模块扩展为 [12]中提出的协同注意力类比对齐模块。具体而言，我们将每个对象及簇/人脸的表示向量与其空间编码拼接起来。空间编码的填充定义为所有坐标均为零。

Img-Obj-Spatial 和 **Img-Face-Spatial** 模型分别将模因图像的 CLIP 表示 I融合到 Obj-Spatial 和 Face-Spatial中。由于这些模型对每个模因使用了三种表示——图像、文本簇和对象/人脸——我们扩展了 Gu 的 [4] 融合机制，通过引入第三组全连接层来适应三个输入。

表 2：每个实验模型的目标。

模型	**数据集**	**Goal**
Baseline	Original	基准测试我们选定的模态编码和 融合机制与领先解决方案的对比。
	Filtered	在样本上建立基线性能 这些样本中包含可检测的文本簇和视觉对象。
	Filtered-OCR	衡量替换.....的影响 人工- 替换为人工筛选的文本 包含由自动化 OCR 识别出的文本簇， 此外，还建立了 我们全自动方法的基线。
Obj-NoSpatial	Filtered-OCR	衡量仅使用 CLIP 编码的局部 化视觉对象来表示 图像模态且不使用空间编码时的性能。
Obj-Spatial	Filtered-OCR	衡量纳入 对象和文本簇的空间编码所带来的性能影响。
Img-Obj-Spatial	Filtered-OCR	通过补充 对象和文本簇来增强图像输入， 以及它们各自的空间编码。
Face-NoSpatial	Filtered-OCR	衡量表示.....的性能 仅使用本地化人脸嵌入的图像模态 ，不包含空间编码。
Face-Spatial	Filtered-OCR	衡量纳入 人脸和文本簇的空间编码所带来的性能影响。
Img-Face-Spatial	Filtered-OCR	在图像输入中加入人脸、文本簇 及其相应的空间编码。

表3：我们的Baseline 模型在Memo1 数据集上与领先方案的性能比较。来源： [13,9]。

方案	宏平均 F1
MHA-模因[6] [9]	**0.3762**
Vkeswani IITK	0.3547
我们的 Baseline	0.3546
Guoym	0.3520
Aihaihara	0.3502
Sourya Diptadas	0.3476
Irina Bejan	0.3469

表4：我们的Baseline 模型在Memo2 数据集上与领先解决方案的性能对比。来源： [7]。

解决方案	加权 F1
BLUE	**0.5318**
BROWALLIA	0.5255
Yeti	0.5088
Little Flower	0.5081
Greeny	0.5037
我们的 Baseline	0.5035
Amazon PARS	0.5025

4 结果

在 Original 数据集上评估 Baseline 模型时，该模型在每个相应数据集上均跻身性能最高的六种解决方案之列；参见表 3 和表 4。

表 5 展示了 Baseline 模型在 Original、 Filtered 和Filtered-OCR数据集上的性能。该模型在 Filtered 数据集上的性能低于在 Original 数据集上的性能，这很可能是因为移除了仅包含文本且位于无物体背景上的样本。对这类样本进行分类类似于辨别单模态文本输入的情感，超出了本研究的范围。我们认为， Baseline 模型在 Filtered-OCR 数据集与 Filtered 数据集上的性能差异，源于与人工整理的文本相比，利用我们的自动 OCR 流程提取的文本质量较低。尽管如此，我们的空间感知模型仍能够克服这一性能损失。如表 6 所示，在每个数据集上表现最佳的模型构成了**全自动方法**，其性能优于各自使用来自 Filtered 数据集的人工整理文本训练的 Baseline 模型。通过消除人工干预的需要，全自动

表 5： Baseline 模型在所有数据集上的加权 F1 (F1-W) 和宏平均 F1 (F1-M)。

数据集	Memo1		Memo2	
	F1-W	F1-M	F1-W	F1-M
Original	**0.481**	**0.355**	**0.504**	**0.325**
Filtered	0.475	0.327	0.503	0.314
Filtered-OCR	0.462	0.326	0.439	0.283

[6]非竞赛提交；结果基于原始数据集的子集

表 6：所有模型在 Memo1 和 Memo2 Filtered-OCR 数据集上的加权 F1（F1-W）和宏平均 F1（F1-M）。**相对**表示相对于各数据集**比较**列中所列模型的相对性能。

模型	比较	Memo1			Memo2		
		F1-W	F1-M	相对	F1-W	F1-M	相对
Baseline	-	0.462	0.326	-	0.439	0.283	-
Obj-NoSpatial	对比 Baseline	0.452	0.307	↓	0.427	0.271	↓
Obj-Spatial	对比 Obj-NoSpatial	0.481	0.317	↑	0.482	0.305	↑
Img-Obj-Spatial	对比 Obj-Spatial	**0.489**	0.336	↑	0.499	0.300	↑↓
Face-NoSpatial	对比 Baseline 对比 Obj-NoSpatial	0.476	0.340	↑ ↑	0.471	0.298	↑ ↑
Face-Spatial	对比 Face-NoSpatial 对比 Obj-Spatial	**0.485**	**0.341**	↑ ↑	0.496	0.310	↑ ↑
Img-Face-Spatial	对比 Face-Spatial 对比 Img-Obj-Spatial	0.473	0.332	↓ ↓	**0.509**	**0.314**	↑ ↑

模型提高了大规模开展网络迷因情感分类的可行性，并减少了创建未来迷因数据集所需的工作量。

结果表明，空间编码能够提升性能。 Obj-Spatial和 Face-Spatial 分别优于 Obj-NoSpatial 和 Face-NoSpatial 。这些结果说明，跨模态空间信息对该任务具有信息价值，而 CLIP 对整张表情包图像的编码未能充分表示这类信息。这一发现对于将深度学习解决方案应用于模因尤为重要，因为文本模态被整合到图像中并与图像交织在一起。尽管词元位置对于主流解决方案架构的重要性早已得到充分证实，但在许多其他视觉语言任务中，图像模态和文本模态缺乏共享的视觉媒介，因此主流多模态架构会为每种模态分别采用独立的位置表示。根据我们的研究结果，我们认为，对于视觉对象与文本共享同一视觉媒介的其他视觉语言任务，也应考虑采用空间编码。

使用局部视觉特征表示图像模态所带来的性能提升，取决于特征是定义为对象还是人脸。经 CLIP 编码的对象表示性能不如 Baseline。这是因为图像编码器可用的视觉信息有所减少。然而，使用 FaceNet 嵌入表示人脸的 Face-NoSpatial，表现优于 Obj-NoSpatial 和 Baseline ，尽管它同样面临可用视觉信息减少的问题，甚至减少得更多。此外，Obj-Spatial 相较于 Baseline 的结果好坏参半，而 Face-Spatial 在两个数据集上的表现均优于 Baseline 。值得注意的是，基于视觉对象的模型并未完全排除人脸，因为许多模因样本检测出的对象中都包含“人物”。因此，我们认为，这两种方法之间的性能差异源于 FaceNet 提供了粒度更细的人脸嵌入，以及这类方法固有地排除了非人脸视觉对象，从而突显了人脸对模因

这两种方法的差异源自 FaceNet 提供的更细粒度的人脸嵌入，以及其固有地排除非人脸视觉对象的特性；这突显了人脸对模因情感的贡献。

我们发现，为表情包图像添加对象或人脸的局部表示及其空间编码，始终优于仅依赖图像的模型。然而，选择以 CLIP 编码的对象还是以 FaceNet 编码的人脸来增强表情包图像，结果并不一致，且取决于数据集。尽管在 Memo1 和 Memo2 数据集中， `Img-Obj-Spatial` 和`Img-Face-Spatial`分别表现最佳，但它们相较于 `Obj-Spatial` 和 `Face-Spatial` 的表现似乎取决于数据集。此处性能下降可能源于冗余的模态间信息（例如，全局图像表示与基于对象的表示之间的信息）。与 [9], 不同，我们没有采用任何形式的跨模态学习过滤。

5 结论

在这项工作中，我们探讨了在互联网表情包情感极性分类中应用空间编码和人脸嵌入。我们开发了七种新颖的架构，并在两个具有挑战性的数据集上分别进行了评估。在这两个数据集上，我们提出的基线多模态分类器均跻身领先的最先进解决方案前六名之列。我们发现，仅用视觉对象来表示图像模态并不能始终带来性能提升，而基于人脸的表示则可以。此外，将这些视觉特征的空间信息纳入模型，相较于仅使用图像或仅使用人脸／对象的方法，能够提升性能。对于每个 Memotion 数据集，我们表现最佳的方案都是用经过空间编码的视觉特征和文本簇来增强图像模态。我们提出，这些方案可作为具有竞争力的全自动替代方案，取代当前依赖人工验证基于 OCR 的文本提取的最先进解决方案。

参考文献

1. Al-Rawi, A.：Instagram 上的政治表情包与假新闻话语。Media and Communication **9**3533 2 1 276 290 2021 10 17645 9 1 ( ), – ( )。https://doi.org/ . /mac.v i . . Ayvaz, S., Shiha, M.：表情符号对情感分析的影响。International Journal of Computer and Electrical Engineering **9**, 360–369 (01 2017)。https://doi.org/10 369 3 17706 2017 9 1 360 . /IJCEE. . . . - . Bucur, A.M., Cosma, A., Iordache, I.：BLUE 在 Memotion 2.0 2022 中的表现：你有我的图像、我的文本和我的 Transformer。载于：De-Factify @ AAAI 2022。首届多模态事实核查与仇恨言论检测研讨会。CEUR研讨会论文集，AAAI (02 2022) 4. Gu, Y., Yang, K., Fu, S., Chen, S., Li, X., Marsic, I.：基于混合注意力的口语语言分类多模态网络。载于：第27届国际计算语言学会议论文集。第2379–2390页。计算语言学协会，美国新墨西哥州圣菲 (8 2018)

5. Joshi, A., Buntain, C.：利用右翼：使用共享图像推断在线影响力宣传活动中的意识形态立场。载于：第16届国际AAAI网络与社交媒体会议研讨会论文集。(06 2022)。https://doi.org/10 45 6 36190 2022. / . . Kiela, D., Firooz, H., Mohan, A., Goswami, V., Singh, A., Fitzpatrick, C.A., 等：仇恨迷因挑战赛：竞赛报告。载于：NeurIPS 2020竞赛与演示专题论文集。机器学习研究会议论文集，第133卷，第344–360页。PMLR (06–12 Dec 2021) 7. Patwa, P., Ramamoorthy, S., Gunti, N., Mishra, S., Suryavardan, S., Reganti, A., 等：Memotion 2 的研究结果：迷因情感与情绪分析。载于：De-Factify @ AAAI 2022。首届多模态事实核查与仇恨言论检测研讨会。CEUR研讨会论文集，AAAI (02 2022) 8. Poria, S., Cambria, E., Bajpai, R., Hussain, A.：情感计算综述：从单模态分析到多模态融合。Information Fusion **37**, 98 9 2017 125 10 1016 2017 02 – ( )。https://doi.org/ . /J.INFFUS. . .003 9. Pramanick, S., Akhtar, M.S., Chakraborty, T.：锻炼？我还以为你说的是“额外薯条”：利用句子边界标记和多跳注意力进行迷因情感分析。国际AAAI网络与社交媒体会议论文集 **15**(1), 513–524 (2021年5月) 10. Pramanick, S., Sharma, S., Dimitrov, D., Akhtar, M.S., Nakov, P., Chakraborty, T.：MOMENTA：用于检测有害迷因及其目标的多模态框架。载于：计算语言学协会论文集：EMNLP 2021。第4439–4455页。计算语言学协会，多米尼加共和国蓬塔卡纳 (11 2021 379 11 10 18653 1 2021)。https://doi.org/ . /v / . findings-emnlp. . Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., 等：通过自然语言监督学习可迁移的视觉模型。载于：国际机器学习会议。第8748–8763页。PMLR (2021) 12. Shang, L., Zhang, Y., Zha, Y., Chen, Y., Youn, C., Wang, D.：AOMD：一种面向社交媒体冒犯性迷因检测的类比感知方法。Inf. Process. Man- age. **58**102664 13 5 9 2021 10 1016 2021 ( ) ( )。https://doi.org/ . /j.ipm. . . Sharma, C., Bhageria, D., Scott, W., PYKL, S., Das, A., Chakraborty, T., 等：SemEval-2020 任务8：Memotion 分析——视觉语言隐喻！载于：第十四届语义评测研讨会论文集。第759–773页。国际计算语言学委员会，巴塞罗那（线上）(Dec 2020)。https://doi.org/10 99 14 18653 1 2020 1 . /v / .semeval- . . Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.：Attention Is All You Need。载于：第31届国际神经信息处理系统会议论文集。第6000–6010页。NIPS’17，Curran Associates Inc.，美国纽约州纽约市雷德胡克 (2017) 15. Yuan, J., Mcdonough, S., You, Q., Luo, J.：Sentribute：从中层视角进行图像情感分析。载于：第二届情感发现与观点挖掘问题国际研讨会论文集。WISDOM ’13，计算机协会，美国纽约州纽约市 (2013)。https://doi.org/10 2502079 16 1145 2502069 . / . . Zhu, R.：通过外部标签和域内预训练增强多模态Transformer：仇恨迷因挑战赛获奖方案 (12 2020)。https://doi.org/10.48550/arxiv.2012.08290