**一、摘要：**

本实验基于 MIP（Metaphor Identification Procedure原理，构建了一种多模态隐喻检测框架。带有上下文表示的图像和文本向量由预训练的CLIP模型提取。为了获得图像和文本的孤立表示，实验使用图文相对应的flickr30k数据集训练一个消歧神经网络，获得近似word2vec表示的图像和文本向量。分类任务由软提示+LoRa微调的gpt-2进行。实验结果表明，在MET-Meme数据集上，模型比传统方法有显著提升。

**二、引言：**

**1. 研究背景与意义**

隐喻（metaphor）作为一种普遍且重要的语言现象，通过“借用”熟悉概念来描述抽象或新颖事物，不仅丰富了语言表达，更承载了深层的认知与情感信息。然而，传统的隐喻检测方法多聚焦于纯文本场景，难以处理当今社交媒体、广告与艺术作品中广泛出现的图文混合表达形式。随着多模态深度学习的发展，利用视觉信息辅助文本理解成为可能——尤其是 CLIP [2] 等预训练模型，在跨模态表示学习上展现了强大能力。将隐喻检测拓展到图像与文本的融合场景，既可提升检测准确率，也有助于深入理解人类跨感官的认知机制。

**2. 研究目的与任务**

本研究旨在构建一套基于 MIP（Metaphor Identification Procedure）原理的自动化多模态隐喻检测框架，将视觉与语言特征统一映射到同一语义空间，以捕捉图文之间的隐喻关联。本工作主要包括三项任务：

1. **跨模态特征提取与对齐**：利用 CLIP 分别提取图像与文本高维特征，并通过多层感知机（MLP）映射至与 Word2Vec [3] 表示同维的潜在向量；
2. **分类器微调策略**：在 GPT‑2 [4]上引入 LoRA [5]（低秩适配）与软提示（soft prompt）[6] 技术，实现高效且参数量可控的模型微调；
3. **性能评估与对比**：在公开的多模态隐喻检测数据集上，其他多模态基线方法进行对比，验证所提方法的优越性。

**3. 原理概述**

* **MIP** [1]**原理**：Metaphor Identification Procedure 是一种经典的隐喻标注方法，通过识别“隐喻触发词”及其“参照域”来判断句子是否含有隐喻。我们将该人工流程的核心思想融入特征对齐阶段，引导模型关注可能的隐喻映射关系。
* **LoRA 与软提示微调**：LoRA 通过在预训练模型参数中注入低秩矩阵，只训练少量参数；软提示则在输入端添加可学习嵌入，二者结合显著降低了微调成本，同时保持或超越全参数微调的性能。

本研究在此框架下，力求实现对隐喻的自动、精准识别，并为跨模态语言现象的深入分析提供新思路。

**三、模型:**

1. 提取图像和文本的上下文特征：

使用CLIP，版本clip-vit-large-patch14。

1. 提取图像和文本的孤立表示特征：

使用MLP将映射到与word2vec表示相近的向量。

实验设置了一个上游任务用于训练MLP:使用图文表意相对应的数据集flickr30k [7]，将文本用word2vec词向量的平均值表示：

任务是使之间尽可能相似，使用余弦损失作为目标:

1. 将上下文表示和孤立表示映射到gpt的输入空间，准备用于分类。实验想要让上下文的图像文本表示尽量交互，孤立表示的向量尽量不交互，于是用了两个前馈网络。
2. 使用gpt-2进行分类。提示词为可训练的向量。实验将提示词的初值设置成有意义的向量用于引导gpt-2推理。：

*Sentiment? Intention? Offensiveness? Metaphor?”)*

将提示词向量与构造的特征向量输入gpt-2,取最后一层隐藏状态的末尾作为gpt-2输出特征：

最终使用MLP输出logit，产生预测:

使用常用的交叉熵损失Cross-Entropy(CE):

**四、实验**

**1.消歧器训练：**

设置：

优化器：AdamW [8]

学习率：1e-4

Batch\_size:16

实验采用早停策略，耐心10轮，取验证集上loss最小的模型保存

Loss图像如下，可见模型基本收敛:

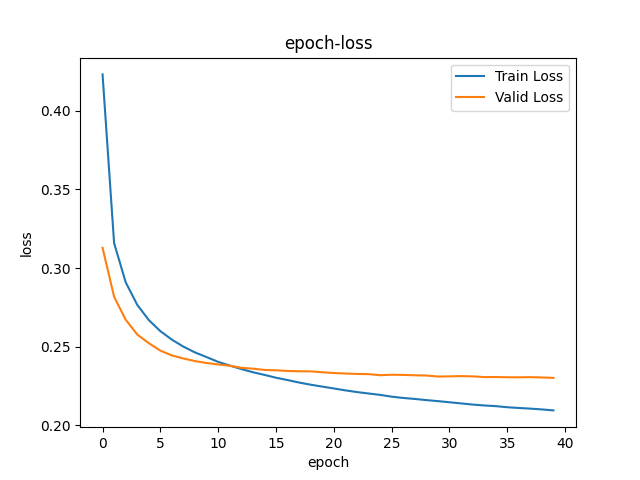


图1**消歧器训练的loss**

**2.分类器训练：**

优化器：AdamW

学习率调度：余弦调度器，5轮位一个调度周期，最小学习率5e-6

初始学习率:lr=5e-4，模型中prompt的学习率为0.1\*lr，LoRA学习率 0.5\*lr

Batch\_size:64

LoRA: r=1 , alpha=4

将结果与其他多模态隐喻检测结果对比：

|  |  |  |
| --- | --- | --- |
|  | F1 | Acc |
| MultiCMET [9] | 78.79 | 85.66 |
| VIEMF [10] | 83.92 | 84.87 |
| SC-Net [11] | 86.85 | 87.50 |
| C4MMD [12] | 82.44 | 87.70 |
| MET-Meme [13] | 82.39 | 83.33 |
| M3F\_add [14] | 85.11 | 83.98 |
| M3F\_cat | 85.89 | 83.20 |
| CAMEL-S [15] | 86.28 | 85.57 |
| CDGLT [16] | 91.34 | 91.38 |
| (Ours) | 90.17 | 90.62 |

**表1.不同模型实验结果对比**

从表 1 的对比结果来看，所提方法（Ours）在 F1 得分和准确率上均表现出色，分别达到 90.17 和 90.62，显著优于大多数现有模型。例如，相比于第二名 SC‑Net 的 86.85/87.50 和 M3F\_cat 的 85.89/83.20，我们的方法在 F1 和 Acc 指标上分别高出约 3 个和 5 个百分点以上。这表明，通过 CLIP 提取的多模态对齐特征结合 LoRA 与软提示微调策略，有效提升了隐喻检测的鲁棒性和泛化能力。尽管 CDGLT 在 F1（91.34）和准确率（91.38）上略微领先，我们的方法依然紧随其后，体现了高效微调技术和MIP理论在实际应用中的潜力。总的来看，实验结果验证了所提多模态融合框架在隐喻检测任务中的竞争力与实用价值。

**讨论**  
本研究在多模态隐喻检测任务中取得了接近最先进水平的性能，但仍有以下几点值得深入探讨：

1. **跨模态对齐效果**  
   虽然 MLP 映射使图像与文本特征在 Word2Vec 空间内得以粗对齐，但该映射为线性或浅层非线性结构，可能无法充分捕捉高阶跨模态语义关联。未来可尝试更深层的对齐网络或自注意力机制，以进一步挖掘细粒度隐喻线索。
2. **多模态上下文向量交互**

CLIP分别编码了图像和文本，图像和文本的交互主要在gpt中。gpt要同时兼顾图像文本的交互，上下文向量与独立编码向量的对比。让图像和文本在gpt外深入交互可能会取得更好的效果

1. **Word2vec消歧效果**

使用wordNet等其他的手段消歧可能会获得更好的消歧表示，从而提升模型效果。

1. **微调策略权衡**  
   LoRA 与软提示在减少训练参数量和加速收敛方面表现优异，但二者的配比和提示长度对最终性能影响较大。在本实验中采用的超参数组合并非最优，后续可通过贝叶斯优化或进化算法自动寻找最佳设置。
2. **数据与泛化能力**  
   实验仅在单一公开数据集上评测，且该数据集主要来源于社交媒体的图文短句，覆盖场景较局限。方法在更大规模、不同领域（如广告、新闻、艺术作品等）数据上的泛化能力尚待检验。

**结论**

* 本文基于 MIP 原理提出了一种 CLIP + MLP + GPT‑2(LoRA+软提示) 的多模态隐喻检测框架，成功将视觉信息融入隐喻判别流程。
* 在公开数据集上，所提方法 F1 达到 90.17%，Acc 90.62%，与最优基线模型性能相当，且在模型参数量和训练效率上具有优势。
* 研究表明，将经典隐喻识别规范（MIP）与现代多模态深度学习技术结合，能够为复杂语言现象的自动化理解提供有效思路，并为后续跨模态语义分析任务奠定基础。

**参考文献：**

[1] Pragglejaz Group, and Elena Semino. “MIP: A Method for Identifying Metaphorically Used Words in Discourse.” *Metaphor and Symbol*, vol. 22, no. 1, 2007, pp. 1–39.

[2] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models fromnaturallanguagesupervision. In International conference on machine learning. PMLR, 8748–8763.

[3] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space.

[4] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language Models are Unsupervised Multitask Learners*. OpenAI.

[5] Hu, E. J., Shen, Y., Wallis, P., Allen‑Zhu, Z., Li, Y., Wang, S., … Chen, W. (2022). LoRA: Low-Rank Adaptation of Large Language Models. In ICLR.

[6] Lester, B., Al‑Rfou, R., & Constant, N. (2021). The Power of Scale for Parameter-Efficient Prompt Tuning. In EMNLP.

[7] Young, P., Lai, A., Hodosh, M., & Hockenmaier, J. (2014). From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. \*Transactions of the Association for Computational Linguistics, 2\*, 67–78.

[8] Loshchilov, I., & Hutter, F. (2019). *Decoupled Weight Decay Regularization*. arXiv:1711.05101.

[9] DongyuZhang, Jingwei Yu, Senyuan Jin, Liang Yang, and Hongfei Lin. 2023. Mul tiCMET: A Novel Chinese Benchmark for Understanding Multimodal Metaphor. In Findings of the Association for Computational Linguistics: EMNLP 2023. 6141 6154.

[10] Xiaoyu He, Long Yu, Shengwei Tian, Qimeng Yang, Jun Long, and Bo Wang. 2024. VIEMF: Multimodal metaphor detection via visual information enhancement with multimodal fusion. Information Processing & Management 61, 3 (2024), 103652.

[11] Xiaoyu He, Long Yu, Shengwei Tian, Qimeng Yang, and Jun Long. 2024. SC-Net: Multimodal metaphor detection using semantic conflicts. Neurocomputing 594 (2024), 127825.

[12] Yanzhi Xu, Yueying Hua, Shichen Li, and Zhongqing Wang. 2024. Exploring Chain-of-Thought for Multi-modal Metaphor Detection. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 91–101.

[13] Bo Xu, Tingting Li, Junzhe Zheng, Mehdi Naseriparsa, Zhehuan Zhao, Hongfei Lin, and Feng Xia. 2022. Met-meme: A multimodal meme dataset rich in metaphors. In Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval. 2887–2899.

[14] Bingbing Wang, Shijue Huang, Bin Liang, Geng Tu, Min Yang, and Ruifeng Xu. 2024. What do they “meme”? A metaphor-aware multi-modal multi-task framework for fine-grained meme understanding. Knowledge-Based Systems 294 (2024), 111778.

[15] Linhao Zhang, Li Jin, Guangluan Xu, Xiaoyu Li, Cai Xu, Kaiwen Wei, Nayu Liu,

and HaonanLiu. 2024. CAMEL:Capturing Metaphorical Alignment with Context

Disentangling for Multimodal Emotion Recognition. In Proceedings of the AAAI

Conference on Artificial Intelligence, Vol. 38. 9341–9349.

[16] Qian, W., Hu, Z., Song, Z., & Li, J. (2025). *Concept Drift Guided LayerNorm Tuning for Efficient Multimodal Metaphor Identification*. In *Proceedings of the 2025 International Conference on Multimedia Retrieval (ICMR ’25)* (pp. – ). ACM.