MARIC:M 多 A 范 R 理由用于 I 图 C 类别分类

Wonduk Seo*, Minhyeong Yu*, Hyunjin An, Seunghyun Lee†

AI Research, Enhans

ABSTRACT

图像分类传统上依赖于参数密集型模型训练，需要大规模标注数据集和广泛的微调以达到竞争性能。尽管最近的视觉-语言模型（VLMs）缓解了其中一些限制，但它们仍然受限于单次通过表示法，通常无法捕捉到视觉内容的互补方面。在本文中，我们引入了 M **超级-A 基于代理的 R 推理用于 I 图像 C 分类**(MARIC)，一个多智能体框架，它将图像分类重新表述为一种协作推理过程。**马里克**首先利用一个大纲代理分析图像的全局主题并生成有针对性的提示。基于这些提示，三个方面代理沿不同的视觉维度提取细粒度描述。最后，一个推理代理通过综合反思步骤合成这些互补输出，产生用于分类的统一表示。通过明确将任务分解为多个视角，并鼓励反思性综合，**马里克**缓解了参数密集型训练和单一化 VLM 推理的缺点。实验在 4 多样化的图像分类基准数据集上表明，**马里克**显著优于基线方法，突显了多智能体视觉推理对于稳健且可解释的图像分类的有效性。

Index Terms— 图像分类，多智能体推理，视觉语言模型，多媒体理解

代码可用性： 代码可在 https://github.com/gogoymh/MARIC [1] 获取

1 介绍

图像分类长期以来一直是计算机视觉中的基础任务，传统上由在大规模标注数据集上训练的深度学习模型主导 [1, 2]。早期的卷积神经网络（CNN）通过密集参数训练实现了强大的性能 [3, 4]，而最近的视觉变压器（ViT）推进了该领域的发展 [5]。然而，这些方法仍然依赖于大规模数据集、广泛的微调，并且提供有限的可解释性。

视觉-语言模型（VLMs）作为强大的替代方案出现 [6]，将视觉和文本模态结合，以实现无需重新训练的分类，并使视觉内容与自然语言标签对齐 [7]。然而，大多数基于 VLM 的方法集中在训练优化上，依赖单次通过表示，忽视了多角度线索 [8]。因此，VLMs 仍然落后于特定任务的分类器，并且需要更强的推理能力和证据聚合以实现可靠的分类 [9]。

为了解决这些挑战，我们提出了一个新颖的框架**多智能体基于的推理用于图像分类（马里克）**，将图像分类重新定义为专门代理之间的协作推理过程。与仅仅依赖参数繁重的训练或单一的 VLM 推断不同，**马里克**将任务分解成多个视角：一个大纲代理首先分析图像的全局上下文以生成有针对性的提示；然后方面代理提取沿着不同视觉维度的细粒度描述；最后，一个推理代理反思这些输出，批评其中的不一致之处，并强调重要的线索，在此基础上将它们合成一个由明确推理痕迹支持的统一分类决策。因此，**马里克**不仅提高了分类准确性，还增强了可解释性和在各个数据集上的鲁棒性。

广泛的实验在 4 多样化的基准数据集上表明，**马里克**显著优于竞争基线，突出了多智能体视觉推理作为一种可扩展且可解释的范式，在推进图像分类方面超越传统训练密集型或单次推断方法的潜力。

The authors denoted as * have contributed equally to this work.

† denotes corresponding author.

[1] 详细的实现细节在 github 仓库中提供。

2 相关工作

2.1 视觉-语言基础模型

图像分类传统上依赖于参数密集型模型训练，需要大量的标注数据集和广泛的微调以实现具有竞争力的表现 [1, 2, 3]。为缓解这些限制，早期的视觉-语言研究集中在联合嵌入学习上，用于如图像描述生成和图像-文本检索等任务。大规模图像-文本预训练与 Transformer 架构的结合进一步加速了进展 [6, 7]。代表性模型如 CLIP 和 Flamingo 学会了稳健的联合表示，缩小了视觉和语言之间的语义差距 [6, 10]，展示了在包括图像描述生成、视觉问题回答 (VQA)、文本到图像合成以及跨模态检索在内的多种任务中的强大零样本和少样本泛化能力。然而，这些模型仍然依赖于单次通过表示和静态对齐，常常无法捕捉互补的视觉线索，这促使后续研究集中在自适应零样本分类和增强推理的方法上。

2.2 零样本视觉语言模型分类

零样本分类，其中类别名称被嵌入为文本提示并通过图像-文本相似性进行预测 [6]，已成为一种标准范式。通过提示调整、基于适配器的微调以及多模态提示对齐 [11]、集合提示和推理驱动的方法 [12]，准确性得到了提高。最近，角色差异化的多代理框架迭代地提出、批评并完善候选解决方案以实现更强大且可解释的分类 [8]。然而，许多方法仍然依赖于单次推断，限制了它们捕捉互补线索和产生透明推理的能力。这促使了诸如**马里克**等框架的发展，该框架将图像分类明确分解为异常检测、方面分析和推理代理，生成补充证据和结构化的推理轨迹，从而提高准确性和可解释性。

3 方法论

本节首先介绍了我们提出的**基于多智能体的图像分类推理 (马里克)** 框架 (第 3.1 节)。然后详细描述了每个代理的功能和公式，包括离群值代理 (第 3.2 节)、方面代理 (第 3.3 节) 和推理代理 (第 3.4 节)。**马里克**框架的概述如图 1 所示。

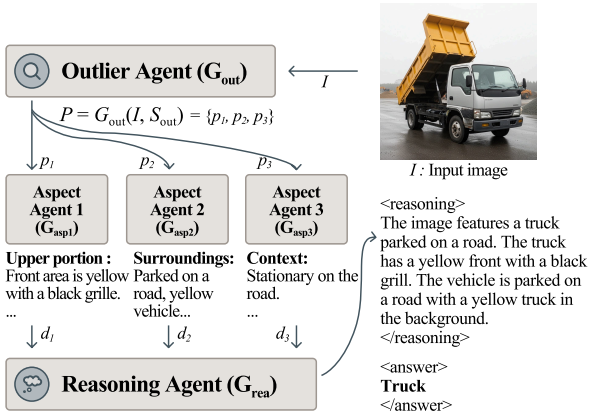

Fig. 1. **马里克**的概述。离群代理设置全局提示，方面代理生成互补描述，推理代理将它们整合成推理轨迹和最终标签。

3.1 概述 của 马里克

给定输入图像 I，*离群代理* G_{out} 首先识别图像的全局主题并生成指定焦点方面的提示。这些提示随后指导 3 *方面代理* G_{asp}，每个产生来自互补视觉角度的细粒度描述。*推理代理* G_{rea} 最终将这些输出综合成一个统一的推理轨迹和相应的分类结果。

3.2 大纲代理 G_{out}

大纲代理负责捕捉图像的整体主题。不是立即关注局部特征，G_{out} 建立了后续推理的全局背景。这是至关重要的，因为许多分类任务依赖于上下文线索。

形式上，G_{out} 生成一组提示 $\mathcal{P}$：

$$\mathcal{P} = G_{\text{out}}(I, S_{\text{out}}), \quad \text{where } \mathcal{P} = \{p_1, p_2, \ldots, p_n\} \quad (1)$$

其中每个 p_i 结合一个前缀（指定要关注的视觉区域或属性）和一个后缀（定义描述目标）。通过设置这些提示，离群值代理防止了后续阶段中的冗余或不集中的提取，确保随后的代理处理图像的不同且互补的方面。这里，S 特指生成系统提示。遵循之前的多代理推理工作 [8, 13]，我们默认设置 $n = 3$ 个方面代理，因为这种配置在多样性和冗余之间提供了平衡的权衡。

Table 1. 基准数据集上的分类准确性（%）。最佳结果在**粗体**，第二佳是 下划线的。

模型	方法	CIFAR-10	OOD-交叉验证	天气	皮肤癌
LLaVA 1.5-7B	Direct Generation	64.8	66.5	50.2	50.6
	Chain-of-Thought (CoT)	83.5	80.8	70.1	50.0
	SAVR	75.6	55.5	48.0	49.4
	马里克（我们的）	90.8	81.9	65.6	50.6
LLaVA 1.5-13B	Direct Generation	86.6	86.2	21.7	52.9
	Chain-of-Thought (CoT)	88.0	75.2	81.1	49.4
	SAVR	88.6	81.2	63.0	62.6
	马里克（ours）	93.5	89.9	85.2	56.3

3.3 方面代理 G_{asp}

方面代理充当专门的观察者，每个代理的任务是根据提示生成图像的精细描述 $\mathcal{P}$。它们的角色是捕捉单一通过模型可能忽略的互补视角。对于每个 $p_i \in \mathcal{P}$，相应的方面代理 G_{asp} 生成一个描述：

$$d_i = G_{\text{asp}}(I, S_{\text{asp}} \mid p_i) \tag{2}$$

其中 d_i 是专注于特定维度（如颜色、纹理、形状或背景上下文）的文本描述。

显著的是，提示中的前缀 – 后缀结构至关重要，因为前缀确保了对指定区域或属性的关注；而后缀指定了描述目标。这种结构既保证了多样性和精确性，防止代理聚集在冗余细节上，并确保关键特征的全面覆盖。方面代理通过显式地跨多个维度分解任务来丰富整体表示，并提供可以追溯到特定视觉属性的证据。

3.4 推理代理 G_{rea}

推理代理是中心决策者，它将方面代理的输出综合成一个连贯的推理过程和最终分类。传统的分类头 $C_\theta(I)$ 仅依赖于向量嵌入，会被近似为生成包含解释性推理和在一组描述性表示 $D = \{d_1, d_2, ..., d_n\}$ 下预测标签的结构化输出的推理代理 G_{rea}：

$$C_\theta(I) \simeq G_{\text{rea}}(I, S_{\text{rea}}|D), \tag{3}$$

其输出格式为

⟨reasoning⟩ r ⟨/reasoning⟩ ⟨answer⟩ $\hat{y}$ ⟨/answer⟩.

这里，r 表示推理解释，而 $\hat{y}$ 是最终分类决定。重要的是，G_{rea} 包含一个集成反射步骤：在完成其推理之前，代理会明确地重新审视并批评来自方面代理的输出结果，过滤不一致之处并强调显著证据。因此，推理过程变得既透明又自我修正，确保分类得到逻辑上一致且可解释的证据支持。

4 实验

4.1 设置

4.1.1 实验数据集

我们评估**马里克**在四个图像基准数据集上：

CIFAR-10 [14]: 一个典型的 10 类基准数据集（飞机、汽车、鸟、猫、鹿、狗、青蛙、马、船、卡车），每类采样 100 张图像。

OOD-**交叉验证** [15]: 一个包含 10 个类别的分布外鲁棒性基准测试（类别包括飞机、自行车、船、公共汽车、汽车、椅子、餐桌、摩托车、沙发、火车），每个类别采样了 100 张图像。

天气数据集 [16]: 使用 1,125 张图像进行天气状况分类，涵盖 4 种类别（日出、晴天、雨天、多云）。

皮肤癌数据集 [17]: 来自 DermIS 和 DermQuest 的二分类黑色素瘤检测，每个类别（健康/癌变）有 87 张平衡图像。

4.1.2 使用的模型

为了评估**马里克**的有效性，我们采用了 2 个具有代表性的视觉-语言模型（VLMs），包括：(1) llava-1.5-7b-hf 模型，以及 (2) llava-1.5-13b-hf [18]。这些模型

被选中以代表参数规模和架构设计的范围，并配置了温度为 0 以确保输出精确且集中。

4.1.3 基线方法

我们将**马里克**与 3 具有代表性的基线进行比较，包括：(1) *直接生成*，其中 VLM 直接从输入图像生成分类结果，无需额外的提示或推理 [9]；(2) *思维链*（CoT），其中模型被明确指示进行逐步推理 [19]；(3) *单代理视觉推理*（SAVR）基线，其中单个手工制作的提示引导模型一次性生成推理和分类。

4.2 实验结果

4.2.1 主要结果

在所有 4 基准数据集（见表 1）中，基线方法显示出明显的优点但也存在显著的缺点。*直接生成*虽然简单且计算效率高，但往往产生浅层预测，忽视了图像中的互补线索。此外，*思维链*提供逐步推理，但倾向于生成冗长或不集中的解释，并不一定转化为更高的准确性。*单代理视觉推理*（SAVR）基线方法在一个步骤中捕捉推理和分类，而其依赖于手工编写的提示限制了多样性，经常遗漏对稳健决策至关重要的细微属性。

相比之下，我们的**马里克**框架明确将任务分解为多代理角色：*大纲代理*生成捕获全局上下文的目标提示，*方面代理*提取补充的细粒度细节，而*推理代理*通过反思整合这些输出。这种设计使**马里克**能够捕捉更广泛的视觉证据同时过滤冗余信息，在数据集上带来一致性的提升。此外，除了准确性之外，**马里克**提高了可解释性，因其结构化的推理轨迹为预测提供了透明的依据，解决了先前基准的关键限制。

4.3 代理组件的消融研究

为了评估**马里克**中方面代理的贡献，我们使用 LLaVA 1.5-13B 模型进行了一项消融研究 [18](表 2)。结果表明，在所有 4 基准测试中，完整框架都实现了最佳性能，在天气和皮肤癌分类方面取得了显著的收益。即使删除了*方面代理*，**马里克**仍然保持了很强的

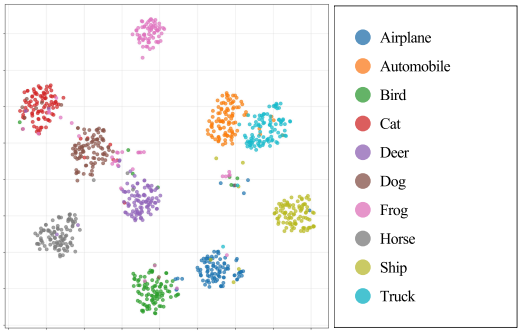

Fig. 2. t-SNE 可视化由**马里克**在 CIFAR-10 上生成的推理嵌入。每个类形成一个分离良好且紧凑的聚类，有助于鲁棒性和可解释性。

准确性，表明仅全局提示和反思推理就能提供具有竞争力的性能。

此外，图 2 展示了由**马里克**生成的用于 CIFAR-10 分类的推理文本嵌入的 t-SNE 可视化。每个点对应一个用 E5 [20] 编码的图像样本，并根据其真实类别着色。聚类揭示了有意义的关系：动物类别与车辆类别分离，而鸟类和飞机的聚类较为接近，反映了它们共享的“天空/飞行”语义。这表明**马里克**的推理捕捉到了超越分类边界的细微区别。

Table 2. 关于使用 LLava 1.5-13B 的**马里克**组件的消融研究。

配置	CIFAR-10	OOD-交叉验证	天气	皮肤癌
Full **马里克**	93.5	89.9	85.2	56.3
w/o Aspect Agents	93.4	89.4	84.5	52.9

Table 3. 定性研究结果（M±SD）。

方面相关性	方面多样性	描述 准确性
3.93 ± 1.08	3.97 ± 1.07	4.00 ± 1.05

5 定性分析

为了进一步评估**马里克**方面分解的质量，我们对来自 CIFAR-10 的 30 张随机抽样图像进行了人工研究，之所以选择 CIFAR-10 是因为它具有多类别特性。

11 名参与者（30 多岁的 AI 研究人员和数据科学家，独立于本研究；$M = 30.64$, $SD = 2.54$）评估了 Aspect Agent 生成的三个方面及其相应的描述。每组方面均使用 5 点 Likert 量表，根据三个标准进行评分——方面相关性、方面多样性和描述准确性。结果显示，平均相关性得分为 3.93，多样性得分为 3.97，准确性得分为 4，表明 Aspect Agent 生成了有意义、互补的方面，并忠实地描述了它们（参见表 3）。这些结果表明，这些方面有助于捕捉独特的特征并相互补充，而描述则忠实于视觉内容。

6 结论

本文介绍了**马里克**，一种新颖的图像分类多智能体框架，它将多智能体结合到协作推理过程中。通过生成有针对性的提示、提取互补的细粒度描述并在合成前反思这些描述，**马里克**能够产生稳健且可解释的分类决策。在四个多样化的基准数据集上的广泛实验确认了这种基于代理的合作能够带来一致性和显著性的准确性提升。

7 限制

虽然**马里克**通过协调大纲、方面和推理代理增强了图像分类，但在全局上下文被误指定或方面提示重叠时仍然面临残余错误，引入了额外的延迟和标记开销，并依赖于固定 n=3 方面代理设置，这可能无法跨领域推广。尽管其计算成本通常与较大的单次通过主干网络相当，但未来的工作将重点放在微调推理代理以减少冗余以及进行更大规模、更详细的定性研究以孤立故障模式并实现自适应代理调度。

8 References

[1] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel, "Backpropagation applied to handwritten zip code recognition," Neural computation, vol. 1, no. 4, pp. 541–551, 1989.

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in 2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009, pp. 248–255.

[3] Dominik Scherer, Andreas Müller, and Sven Behnke, "Evaluation of pooling operations in convolutional architectures for object recognition," in International conference on artificial neural networks. Springer, 2010, pp. 92–101.

[4] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," Advances in neural information processing systems, vol. 25, 2012.

[5] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran, "Image transformer," in International conference on machine learning. PMLR, 2018, pp. 4055–4064.

[6] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al., "Learning transferable visual models from natural language supervision," in International conference on machine learning. PmLR, 2021, pp. 8748–8763.

[7] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu, "Learning to prompt for vision-language models," International Journal of Computer Vision, vol. 130, no. 9, pp. 2337–2348, 2022.

[8] Wonduk Seo, Seungyong Lee, Daye Kang, Zonghao Yuan, and Seunghyun Lee, "Vispath: Automated visualization code synthesis via multi-path reasoning and feedback-driven optimization," arXiv e-prints, pp. arXiv–2502, 2025.

[9] Yuhui Zhang, Alyssa Unell, Xiaohan Wang, Dhruba Ghosh, Yuchang Su, Ludwig Schmidt, and Serena Yeung-Levy, "Why are visually-grounded language models bad at image classification?," Advances in Neural Information Processing Systems, vol. 37, pp. 51727–51753, 2024.

[10] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al., "Flamingo: a visual language model for few-shot learning," Advances in neural information processing systems, vol. 35, pp. 23716–23736, 2022.

[11] Jameel Abdul Samadh, Mohammad Hanan Gani, Noor Hussein, Muhammad Uzair Khattak, Muhammad Muzammal Naseer, Fahad Shahbaz Khan, and Salman H Khan, "Align your prompts: Test-time prompting with distribution alignment for zero-shot generalization," Advances in Neural Information Processing Systems, vol. 36, pp. 80396–80413, 2023.

[12] Mahmoud Masoud, Ahmed Abdelhay, and Mohammed Elhenawy, "Exploring combinatorial problem solving with large language models: A case study on the travelling salesman problem using gpt-3.5 turbo," arXiv preprint arXiv:2405.01997, 2024.

[13] Wonduk Seo and Seunghyun Lee, "Qa-expand: Multi-question answer generation for enhanced query expansion in information retrieval," arXiv preprint arXiv:2502.08557, 2025.

[14] Alex Krizhevsky, Geoffrey Hinton, et al., "Learning multiple layers of features from tiny images," 2009.

[15] Bingchen Zhao, Shaozuo Yu, Wufei Ma, Mingxin Yu, Shenxiao Mei, Angtian Wang, Ju He, Alan Yuille, and Adam Kortylewski, "Ood-cv: A benchmark for robustness to out-of-distribution shifts of individual nuisances in natural images," in European conference on computer vision. Springer, 2022, pp. 163–180.

[16] Gbeminiyi Ajayi and Prateek Srivastava, "Multi-class weather dataset," Kaggle, 2023, Licensed under CC BY 4.0.

[17] Vipin Venugopal, Justin Joseph, M. Vipin Das, and Malaya Kumar Nath, "Skin cancer detection dataset," Kaggle, 2023, University of Waterloo.

[18] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee, "Improved baselines with visual instruction tuning," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2024, pp. 26296–26306.

[19] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al., "Chain-of-thought prompting elicits reasoning in large language models," Advances in neural information processing systems, vol. 35, pp. 24824–24837, 2022.

[20] Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei, "Text embeddings by weakly-supervised contrastive pre-training," arXiv preprint arXiv:2212.03533, 2022.