Propose, Match, then Vote: Enhancing Robustness for Zero-shot Image Classification via Cross-modal Understanding Submitted to VizWiz-Classification Challenge 2024

Jialong Zuo, Hanyu Zhou, Dongyue Wu, Wenxiao Wu, Changxin Gao*, Nong Sang
National Key Laboratory of Multispectral Information Intelligent Processing Technology,
School of Artificial Intelligence and Automation, Huazhong University of Science and Technology
{jlongzuo, hanyzhou, dongyue_wu, wenxiaowu, cgao, nsang}@hust.edu.cn

Abstract

In this report, we propose our method for the task of Zero-shot image classification in VizWiz Grand Challenge 2024. Our approach can be systematically divided into three distinct phases. ***1) Propose*** *the candidate category label set with generated detailed image description;* ***2) Match*** *the candidate categories with the visual information and description of the information utilizing the extraordinary understanding ability of multi-modal large-scale models via a cross-modal validation manner;* ***3) Vote*** *for the final results to resolve matching conflicts using multiple visual classification models. We find that text-vision and text-caption dual-branch matching with cross-modal large-scale models can significantly enhance zero-shot classification performance. Moreover, for challenging samples where dual-branch matching results are inconsistent, employing a mixture of experts is surprisingly effective. Our method achieves an accuracy of* ***67.67%*** *on the testing set of the VizWiz-Classification dataset, ranking top-3 among all the competitors in this challenging task.*

1. The proposed method

In this section, we introduce the overall pipeline and components of our proposed method in details.

1.1. The overall pipeline

Our overall pipeline is shown in the Figure 1. It follows three steps to perform zero-shot classification: propose, match and vote. In the propose phase, we generate the potential categories for each image, together with its detailed description. All potential categories are stored in the Proposed Label Set, which are then fed into the Match phase. During Match phase, each element in the proposed label set serves as potential candidate. Based on the text of these candidates, a dual-branch matching is conducted to get text-vision matched category label and text-caption matched category label. Thereafter, if two matched category label is the same, then the final classification result is the corresponding category. If a conflict emerges, we employ an ensemble of multiple visual classification models to vote. The top-1 class with the most tickets is the output prediction.

1.2. Propose with VOLO

In this phase, we generate the proposed label set and the detailed description to extract text-modal information from the image. Transforming the vision-modal information to text-modal information helps to get rid of the noise and condense the semantic information. Firstly, the proposed label set contains top-20 categories with highest probability generated by VOLO [3]. We believe the generated top-20 categories could cover most potential items in the image, even though there are some image quality issues like blurring, obscured and framing. This procedure achieves a coarse but comprehensive understanding of the image. As the text of these categories can only provide highly condensed semantic information, we also employ the vision-language model QWEN-VL-Plus [1] to perform image caption task to generate the detailed description of all the items in the image. This detailed text-modal information are more robust for zero-shot classification. These text-modal information will be fully exploited during the next Match phase.

1.3. Match via LLaVA and Kimi

In the match phase, we exploit the text-modal representation, *i.e.* the text of category candidates in the proposed label set and detailed description of the image, to perform zero-shot classification. The proposed dual-branch matching mechanism performs a cross-modal validation of two matched categories, which mutually confirms the correctness of the matched classification results.

For the text-vision matching branch, we utilize multi-modal large model LLaVA [2] for matching the best candidate category with the visual information of the image.

[*] Corresponding author.

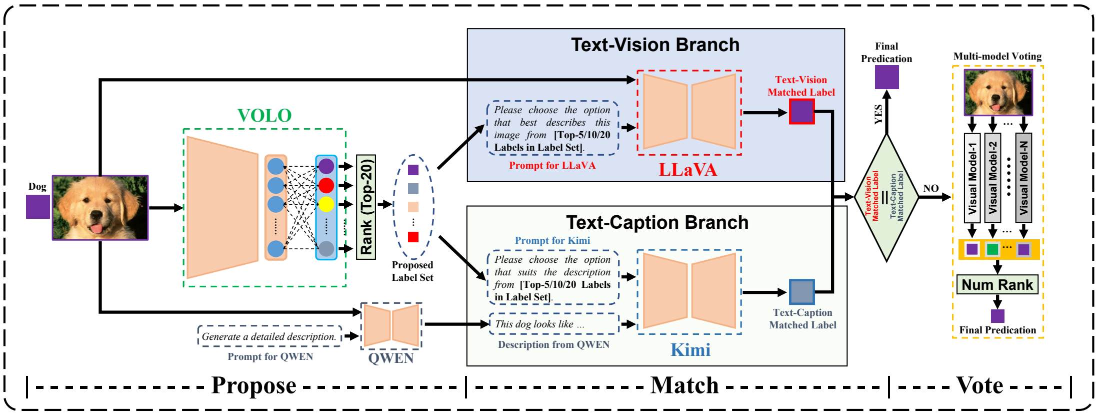
Text-Vision Branch
VOLO
Dog
Rank (Top-20)
Proposed Label Set
Please choose the option that best describes this image from [Top-5/10/20 Labels in Label Set].
Prompt for LLaVA
LLaVA
Text-Vision Matched Label
Final Predication
YES
NO
Text-Vision Matched Label
Text-Caption Matched Label
Multi-model Voting
Visual Model-1
Visual Model-2
Visual Model-N
Num Rank
Final Predication
Text-Caption Branch
Prompt for Kimi
Please choose the option that suits the description from [Top-5/10/20 Labels in Label Set].
This dog looks like ...
Description from QWEN
Kimi
Text-Caption Matched Label
Generate a detailed description.
Prompt for QWEN
QWEN
Propose
Match
Vote

Figure 1. The overall pipeline of the proposed method.

LLaVA takes the prompt with the text of all candidate categories and the original image as input. The prompt is: “Please carefully observe this image carefully, especially the item presented in it. Please choose the option that best describes this image from the following options: [*text of categories*]”. The “[*text of categories*]” can be the text of top-5, top-10, top-20 categories, which are provided for cascaded matching. If the best option is not included in the top-5, then we add top-10 classes. If still not included, we provide top-20 classes and finally select the highest confidence option.

The text-caption matching branch adopts Kimi to perform a question-answering task. The input of Kimi is the prompt with the text of categories, like “Please read this description carefully, especially the item present in it. Please choose the option that suits the description from the following options: [*text of categories*]”. Since the capacity of the large language model (LLM) to reason from text is usually better than multi-modal large models, we utilize Kimi instead of other vision-language models.

1.4. Vote by multi classification models

In most cases when those two matched category labels are consistent with each other, the final output will be the corresponding category. However, for some difficult samples, there are occasions when the text-vision matched label and text-caption matched label are not the same. To address the conflicts, we employ a voting mechanism. As the matched result of the two branches conflicts with each other, we directly resort to a mixture of expert models to perform classification. We employ multiple pre-trained visual and visual-language models to predict the category most likely shown in the image, respectively. Then, the top category with the most tickets is the final prediction.

2. Experiments

In this section, we provide the experimental results in Table 1. To utilize the complementary benefits of different modalities, we validate the proposed phase, matching phase, and vote phase respectively. The results show that the capacity of LLM in T-C to understand and perform reasoning significantly improves the performance by 2.55% compared with only adopting T-V. The dual-branch matching and voting further improves the result by 3.3%.

Method	T-V (LLaVA)	T-C (Kimi)	Vote	VizWiz								
				BLR	BRT	FRM	ROT	OBS	DRK	Corr	Clean	Total
VOLO(base)				56.88	53.12	54.67	49.55	40.82	55.4	51.74	59.71	57.13
base w/ T-V	✓			58.82	53.16	57.14	51.16	57.14	67.04	57.41	63.75	60.65
base w/ T-C		✓		61.34	64.57	61.46	58.47	57.14	55.68	59.78	65.35	63.20
Ours	✓	✓	✓	64.99	64.06	63.97	58.36	52.66	63.41	61.24	68.80	66.56
Ours†	✓	✓	✓	-	-	-	-	-	-	-	-	**67.67**

Table 1. Comparison with different methods on VizWiz. ‘T-V’ denotes the text-vision branch. ‘T-C’ represents the text-caption branch. We use ‘†’ to denote that a classifier fine-tuned on ImageNet is added to some visual-language model in vote phase.

3. Conclusion

In this work, we propose an effective solution for zero-shot image classification. The proposed method adopts a cross-modal and dual-branch manner to enhance the robustness based on the extraordinary ability of visual-language understanding and text-modal reasoning. Moreover, the dual-branch matching ensures consistency, with a voting mechanism dealing with matching conflicts. Based on these designs, the proposed method achieves high performance and robustness in the competition.

References

[1] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 1

[2] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 1

[3] Li Yuan, Qibin Hou, Zihang Jiang, Jiashi Feng, and Shuicheng Yan. Volo: Vision outlooker for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 45 (5):6575–6586, 2022. 1