Multimodal Ensembling for Zero-Shot Image Classification

Javon Hickmon

Department of Computer Science, University of Washington, Seattle WA
javonh@cs.washington.edu

Abstract

Artificial intelligence has made significant progress in image classification, an essential task for machine perception to achieve human-level image understanding. Despite recent advances in vision-language fields, multimodal image classification is still challenging, particularly for the following two reasons. First, models with low capacity often suffer from underfitting and thus underperform on fine-grained image classification. Second, it is important to ensure high-quality data with rich cross-modal representations of each class, which is often difficult to generate. Here, we utilize ensemble learning to reduce the impact of these issues on pre-trained models. We aim to create a meta-model that combines the predictions of multiple open-vocabulary multimodal models trained on different data to create more robust and accurate predictions. By utilizing ensemble learning and multimodal machine learning, we will achieve higher prediction accuracies without any additional training or fine-tuning, meaning that this method is completely zero-shot.

Introduction

Multimodal systems have been a promising new paradigm in the field of machine learning. Utilizing multiple modalities has resulted in machine learning models that are more accurate and generalizable on a broad range of tasks, including sentiment analysis (Wang, Wan, and Wan 2020) and cross-modal retrieval (Kim, Son, and Kim 2021). Despite this performance, issues such as data redundancy, noise, and class imbalance are just a few of the many difficulties that arise from collecting large amounts of training data. Despite this need for high-quality data, open-vocabulary models achieve high classification accuracy across many datasets without labeled training data. The utilization of open-vocabulary models is necessary because the system must be able to generalize and make predictions based off of words not seen in the training data in order to maintain zero-shot performance. These models leverage the massive amounts of image text pairs available online by learning to associate the images with their correct caption, leading to greater flexibility during inference (Pratt et al. 2023). In this work, we aim to create a meta-model that combines the predictions of separate open-vocabulary multimodal models trained on different data, in order to create more robust predictions.

Background

Pratt et al. (2023) introduces CuPL (Customized Prompts via Language models), the foundational work for this project. The resulting zero-shot accuracy was improved by generating descriptive prompts for each ImageNet class, then utilizing these descriptions to find the most likely class for a given image. The work solidifies the idea that prompt generation from a larger model trained on more high-quality data can lead to knowledge transfer, helping to emphasize the differences between classes. The authors also note that the multimodal model can still fail to distinguish between similar classes, despite the fact that the attention shifted when provided generative CuPL descriptions.

Breiman (1996) introduced bagging, the ensemble learning method that combines multiple models trained on different subsets of the training data. The predictions from each of the models is then aggregated to create the final prediction. Bagging both empirically and theoretically prove improves accuracy for a given set of weak classifiers or ”weak learners.”

Yang et al. (2018) is a state-of-the-art paper that attempts to reduce misclassification rates by developing a model called NTS-Net (Navigator-Teacher-Scrutinizer Network) to teach itself methods of identifying and scrutinizing fine-grained image details. Even though NTS-Net would likely achieve higher classification accuracy on a pre-defined dataset, our method is more generalizable due to its use of open-vocabulary models. NTS-Net must be trained to discriminate the fine-grained features of a given dataset to achieve the performance described in the paper, but it is not designed to perform well on unseen classes. Open-vocabulary models allow our method to generalize to unseen classes, so we infer this will result in higher zero-shot classification accuracies. We will know this inference is correct if our method can achieve higher top-1 accuracy than NTS-Net on a dataset they both were not trained on.

Approach

We aim to create an ensemble of models to improve multimodal image classification accuracy, especially for models that are trained on data with a class imbalance, as shown in Figure 1. CLIP (Contrastive Language-Image Pretraining) (Radford et al. 2021) will be used for image-to-text retrieval, and Stable Diffusion XL 1.0 (Podell et al. 2023) for image generation. Our approach is as follows:

Generating Prompts via CuPL

In this stage, we leverage the CuPL to create contextually relevant prompts for our multimodal model. CuPL is designed to generate prompts that effectively guide the model's attention and understanding, enabling better performance in zero-shot image classification. Fifty descriptive prompts are generated per class, then averaged and normalized for each category.

Text-to-Image Classification

Using the generated prompts, we initiate the text-to-image classification phase. CLIP ViT-L/14, our multimodal model, matches the descriptive text prompts for each class to the image embedding and results in a score that represents the likelihood that the class belongs to the image. This step allows us to bridge the semantic gap between text and images, facilitating improved image classification. This step is depicted by the green arrows in Figure 1.

Generate Class Images

We employ Stable Diffusion XL, a diffusion-based image generation model, to conditionally generate an image of each class in the dataset while providing negative prompts (Mokady et al. 2023) to steer the model away from every other class. Our result will be images that emphasizes the differences between the classes.

Image-to-Image Classification

Finally, we get the embedding of the query image by passing it through the CLIP image encoder, then do the same for each of the generated classes. We can find the predicted image classification by performing a similar matching step as in our Text-to-Image Classification to obtain the highest score and assign that class as the prediction. This step is depicted by the blue arrows within Figure 1.

Evaluation

The baseline will be the standard method of top-1 image classification accuracy with the CLIP model.

Datasets

Various datasets will be used to assess the accuracy of the method. The evaluation will include ImageNet (Deng et al. 2009), DTD (Cimpoi et al. 2014), FGVC Aircraft (Maji et al. 2013), and Flowers 102 (Nilsback and Zisserman 2008). This aims to balance fine-grained classification with general image classification.

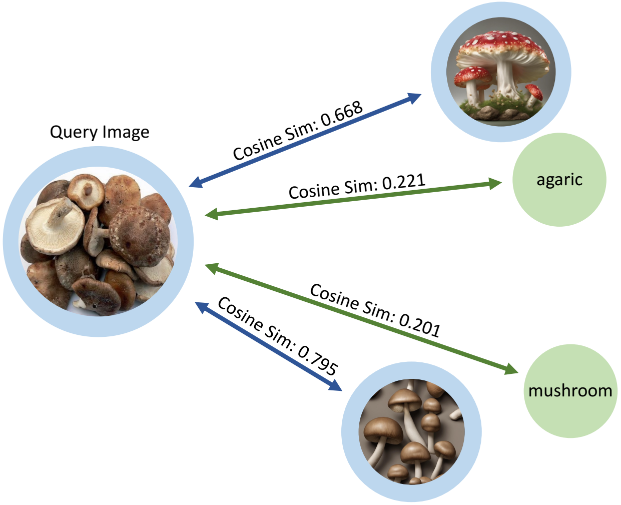

Figure 1: The Multimodal Ensemble Method. This generates an image for every class in the dataset, then takes a weighted sum of the image-to-image similarity scores and the image-to-text similarity scores.

Metrics

The metrics will primarily include traditional accuracy, precision, recall and F1 score; however, we aim to also include more specialized metrics such as confusion matrix analysis. Additionally, we hope to investigate the robustness of our approach to class imbalance, object size, data redundancy, and noise levels.

Discussion

Two prominent assumptions made in this proposal:

1. The generative model has a better learned representation of the true distribution of the data (due to its increased complexity and data diversity).
2. The base multimodal model can distinguish between similar classes. Our method will not improve performance if this is not the case.

Along with this, it is important to note that despite aiming to improve multimodal fairness our method will not produce fair results; rather, it offsets learned biases. This means that the method can either reduce or accentuate human bias, and should not be used as a universal architecture to improve multimodal model fairness.

Conclusion

Traditional multimodal classification methods can struggle to achieve high fine-grained image classification accuracies when trained on data with class imbalances. These challenges can lead to underfitting and misclassification. To mitigate these issues, the proposed approach utilizes generative prompting to emphasize differences between similar classes by leveraging open-vocabulary models. This approach is expected to reduce the misclassification rate and enhance accuracy in various classification tasks.

Acknowledgements

I would like to thank Dr. Hyo Lee and Dr. Rebecca Williams for their many helpful discussions and feedback on this proposal. Along with this, many thanks to Sarah Pratt for her continual guidance and mentorship in my research.

References

Breiman, L. 1996. Bagging predictors. *Machine learning*, 24: 123–140.

Cimpoi, M.; Maji, S.; Kokkinos, I.; Mohamed, S.; ; and Vedaldi, A. 2014. Describing Textures in the Wild. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Gellevij, M.; Van Der Meij, H.; De Jong, T.; and Pieters, J. 2002. Multimodal versus unimodal instruction in a complex learning context. *The Journal of Experimental Education*, 70(3): 215–239.

Kim, W.; Son, B.; and Kim, I. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, 5583–5594. PMLR.

Ma, M.; Ren, J.; Zhao, L.; Testuggine, D.; and Peng, X. 2022. Are multimodal transformers robust to missing modality? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18177–18186.

Maji, S.; Rahtu, E.; Kannala, J.; Blaschko, M.; and Vedaldi, A. 2013. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*.

Mokady, R.; Hertz, A.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2023. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6038–6047.

Nilsback, M.-E.; and Zisserman, A. 2008. Automated Flower Classification over a Large Number of Classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

Pratt, S.; Covert, I.; Liu, R.; and Farhadi, A. 2023. What does a platypus look like? generating customized prompts for zero-shot image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15691–15701.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Wang, Z.; Wan, Z.; and Wan, X. 2020. Transmodality: An end2end fusion method with transformer for multimodal sentiment analysis. In *Proceedings of The Web Conference 2020*, 2514–2520.

Yang, Z.; Luo, T.; Wang, D.; Hu, Z.; Gao, J.; and Wang, L. 2018. Learning to navigate for fine-grained classification. In *Proceedings of the European conference on computer vision (ECCV)*, 420–435.