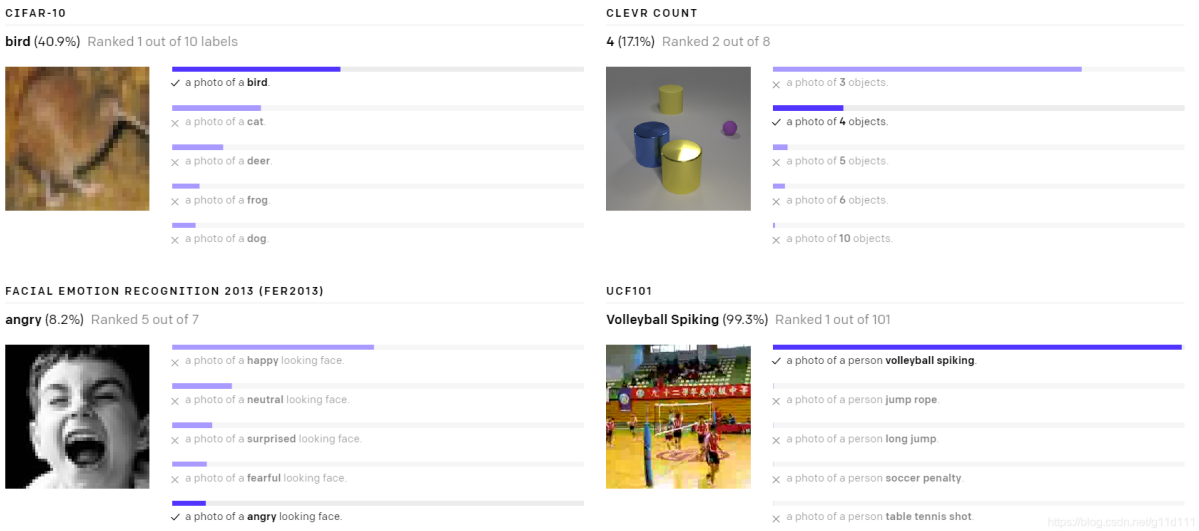
Exploring Rejection Strategies for Zero-Shot Classification



Zero-shot learning is a challenging task that requires models to classify samples that belong to classes unseen during training. State-of-the-art contrastive learning models such as CLIP have shown impressive performance in image-text classification, but they still struggle with handling out-of-distribution samples.

In this project, we aim to explore state-of-the-art rejection strategies in the zero-shot setting for popular models such as CLIP. The objective is to investigate different approaches to reject or abstain from making a classification decision when the model is uncertain or when the input is out-of-distribution. We will first review existing rejection strategies such as uncertainty-based, entropy-based, and margin-based methods. Then, we will experiment with these strategies and compare their performance on standard zero-shot classification benchmarks. Finally, we will propose novel ideas for rejection strategies and evaluate their effectiveness. The ultimate goal is to improve the robustness and reliability of zero-shot classification models, making them more suitable for real-world applications.

References:

- Zhang, Xu-Yao, et al. "A survey on learning to reject." *Proceedings of the IEEE* 111.2 (2023): 185-215.
- Hendrickx, Kilian, et al. "Machine learning with a reject option: A survey." *arXiv preprint arXiv:2107.11277* (2021).
- Hendrycks, Dan, and Kevin Gimpel. "A baseline for detecting misclassified and out-of-distribution examples in neural networks." *arXiv preprint arXiv:1610.02136* (2016).
- Liang, Shiyu, Yixuan Li, and Rayadurgam Srikant. "Enhancing the reliability of out-of-distribution image detection in neural networks." *arXiv preprint arXiv:1706.02690* (2017).