HIERARCHICAL PROMPTS WITH CONTEXT-AWARE CALIBRATION FOR OPEN-VOCABULARY OBJECT DETECTION

Anonymous authors
Paper under double-blind review

A APPENDIX

Visualization for visual-text similarity matrix We use the similarity matrix to analyze the discriminative ability of the hierarchical prompts. Figure 1 shows the similarity matrix between the hierarchical prompts embedding and the visual prototype of the category (48 base classes and 6 novel classes). Ideally, the matrix should have light colors on the diagonal (high similarity) and dark colors on the off-diagonal (low similarity). However, the matrix in Figure 1 is not maximal in the diagonal of the novel category (classes 48 to 53), which leads to a limited improvement in the detection performance of the novel class when only using hierarchical prompts. Therefore, context-aware calibration is needed to correct this similarity matrix. Although the context is related to the input and cannot be directly applied to schemas calculated using category prototypes, ablation studies show that context-aware calibration improves HiCA's performance by another 1.2% on novel classes, proving that it can effectively calibrate results with biased similarities.

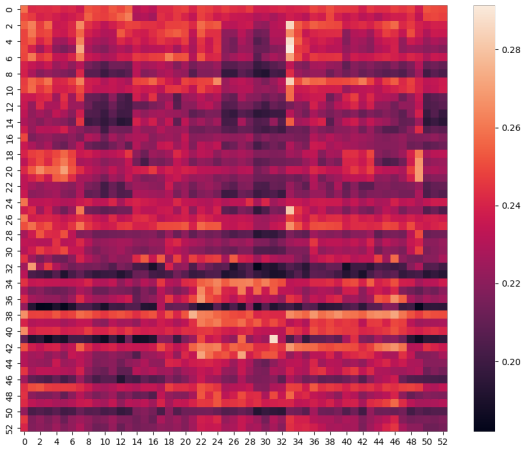
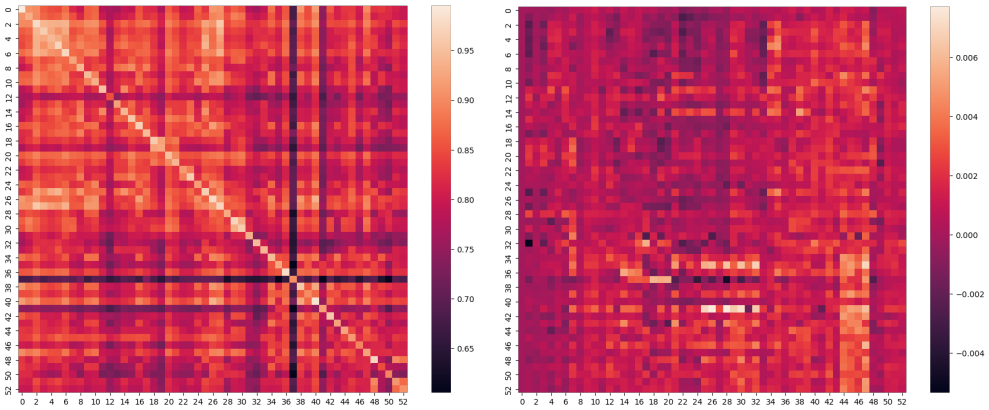


Figure 1: Visualization for visual-text similarity matrix with hierarchical prompts.

Quantitative analysis of hierarchical prompts We analyzed the discriminative power of hierarchical prompts for categories with similar appearances using a similarity matrix. We intercept some representative categories for analysis. Figure 2 (a) shows the similarity matrix of the visual features between different categories, which is obtained by the prototype of each category. The lighter the color, the more similar the appearance between categories. When text embedding is used to classify visual features, the optimal form of the visual-text similarity matrix should be light colors on the diagonal (high similarity) and dark colors on the off-diagonal (low similarity). Figure 2 (b) shows the result of the subtraction of the similarity matrix calculated using hierarchical prompts and single text prompts. The darker in off-diagonal position, the more effective the hierarchical prompt is (the gap between different categories of text and visual features is larger). For example, the similarity between categories 1 to 10 in the upper left corner is high, and the hierarchical prompts effectively improve the discrimination ability in this region, which proves its ability to distinguish categories with high similarity.



(a) Similarity matrix of the visual features.

(b) Subtraction of similarity calculated with hierarchical prompts and single text prompt.

Figure 2: Quantitative analysis for hierarchical prompts.

Detailed analysis of context-aware calibration As the results shown in Table 1, the performance of the model will decrease if the number of unsupervised context clusters is too large or too small. An increase in the cluster center of the context represents a further subdivision of the environment and is likely to result in more similar context embedding. This can lead to confusion when calculating the distribution matrix. However, if the number of context clusters is too small, some environments will be mixed and the distribution matrix will not be effective. The purpose of the DG layer is to map the context-superclass similarity matrix into a distribution matrix. A single fully connected layer for the DG layer cannot learn an effective mapping relationship, and too deep MLP may learn some bias in the training process. These reasons will lead to a degradation in performance.

Table 1: Ablation study of context clustering and the DG layer. "Number" represents the number of centers of the context clustering. "Depth" denotes the MLP depth of the DG layer.

Number	Depth	mAP_N	mAP_B	mAP_{50}
8	1	29.3	54.4	47.8
8	2	**31.2**	**57.2**	**50.4**
8	3	27.6	53.7	46.9
6	1	30.4	54.5	48.2
10	1	28.5	55.4	48.3