COSE474-2024F: Final Project
Enhancing Performance through Cropped Object-Focused Images

Soonhyeok Choi

1. Introduction

Recently, the groundbreaking framework Contrastive Language-Image Pretraining (CLIP) (Radford et al., 2021) has demonstrated exceptional performance across various vision-language tasks and has been widely adopted in numerous downstream applications. Its ability to generalize from base categories to novel categories has significantly advanced the paradigms of few-shot and zero-shot learning. Among the notable enhancements to CLIP is Conditional Context Optimization (CoCoOp) (Zhou et al., 2022), which improves generalization by leveraging learned prompts based on image features.

Despite these advancements, the reliance on full-image training introduces a potential drawback: the inclusion of irrelevant background elements can interfere with the model's ability to focus on the critical objects in an image. This may inadvertently hinder the model's overall performance by diverting attention away from the object of interest.

To address this limitation, we hypothesize that using object-centric, category-specific cropped images can enable CLIP-based models to better focus on the relevant objects, thereby enhancing generalization to novel categories. By eliminating distracting background elements, the model is encouraged to concentrate on object-centric features, facilitating better alignment between image features and textual prompts. Given that CoCoOp utilizes image features to optimize prompt learning, we anticipate a significant performance improvement with this approach. Additionally, this method can enhance the semantic alignment between images and text.

Furthermore, our approach requires no modifications to the underlying model architecture. By simply cropping images and using them as inputs, this technique can be easily applied to a variety of vision-language models, providing a practical and versatile solution for improving generalization capabilities.

Contributions are two part:

- I **introduce an object centric cropping strategy** to preprocess input images, focusing on category-specific objects. This preprocessing step is designed to enhance the alignment between image features and textual prompts
- This method can be **applied easily** in various model with no additional modification to the underlying model architecture.

2. Methods

The proposed method addresses a key limitation of the current CoCoOp-CLIP framework, which relies on the image encoder processing the entire image. This approach often includes background information that is less relevant to object learning and can negatively impact the model's generalization to novel categories. To overcome this issue, I introduce an object-centric cropping strategy, allowing model to fully focus on category-specific objects and thus enhance its ability to generalize effectively.

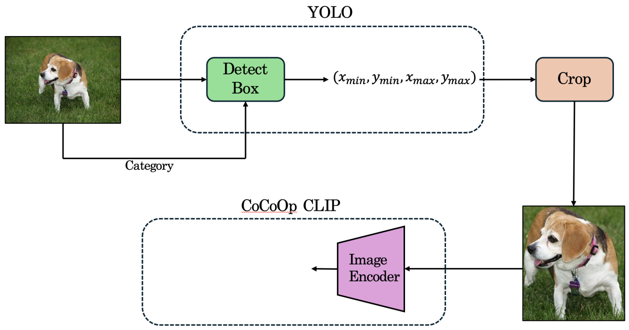


Figure 1. Overall pipeline of proposed method. YOLO detects the object, crops the image, and the cropped image is passed to the CoCoOp-CLIP image encoder for processing.

2.1. Pipeline

The model's pipeline is illustrated in Figure 1. It follows a structure that focuses on category-specific, object-centric cropping to eliminate distracting background information that could hinder learning. First, the input image is processed using YOLO(Redmon, 2016) to obtain the bounding box for the object. Next, the image is cropped according to the detected bounding box and saved. The cropped image is then passed through the CLIP's image encoder, enabling it to be used for training. This process ensures the model focuses solely on the relevant object, enhancing its learning efficiency.

2.2. Cropping Images

I performed object cropping on the images used in my experiments. To ensure consistency in the training and evaluation process, I used the same code for both the original and cropped images, with the only difference being the type of image. For this purpose, I saved the cropped images in the same directory structure as the original ones.

Algorithm 1 Object-Centric Cropping with Single Detection and Label Matching

```
Require: Dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$, where $x_i$ is the input
    image and $y_i$ is the category label, Object Detector $\mathcal{OD}$
Ensure: Cropped dataset $\mathcal{D}' = \{(x_i', y_i)\}_{i=1}^N$, where $x_i'$ is
    either the cropped image or the original image
for each image $x_i \in \mathcal{D}$ do
    Use $\mathcal{OD}$ to detect a single object in $x_i$, obtaining
    bounding box $b$ and detected class label $c$
    if $c = y_i$ then
        Crop $x_i$ using bounding box $b$ to obtain $x_i'$
    else
        Set $x_i' \leftarrow x_i$ {Use original image if label does not
        match}
    end if
    Add $(x_i', y_i)$ to $\mathcal{D}'$
end for
Output: Cropped dataset $\mathcal{D}'$
```

Algorithm 1 describes the process of cropping objects from the original images. To perform cropping, images with labeled categories are required. Object cropping is applied to all images in the dataset. For each image x_i, YOLO is used for object detection, with the image's category provided as input. If the detected object's class matches the given category, the bounding box for the corresponding object is used to crop the image, and the cropped result is stored in the cropped dataset $\mathcal{D}'$. If the detected object's class does not match the category, or if no object is detected, the original image is saved in $\mathcal{D}'$ instead. It is for category-specific cropping.

3. Experiments

3.1. Datasets

In this experiment, two datasets, Caltech-101(Fei-Fei et al., 2004) and Oxford Pets(Parkhi et al., 2012), were utilized. The Caltech-101 dataset comprises a total of 9,146 images spanning 101 object categories and one background category, making up 102 categories in total. On the other hand, the Oxford Pets dataset consists of 7,349 images across 37 categories, which include 12 breeds of cats and 25 breeds of dogs. These datasets were selected for their suitability for cropping using YOLO.
Since the YOLO model used for object detection was pretrained on the COCO dataset, it was crucial for the categories in the COCO dataset(Lin et al., 2014) to align with those in the datasets used in this study. The Oxford Pets dataset was particularly well-suited for this purpose, as its categories are divided into specific breeds of cats and dogs, and COCO also includes categories for cats and dogs. This allowed class labels to be reliably assigned during the cropping process for all images. Although Caltech-101 does not fully align with COCO categories, a substantial portion—11 categories—matched, making it a viable choice for this experiment.

3.2. Implementation details

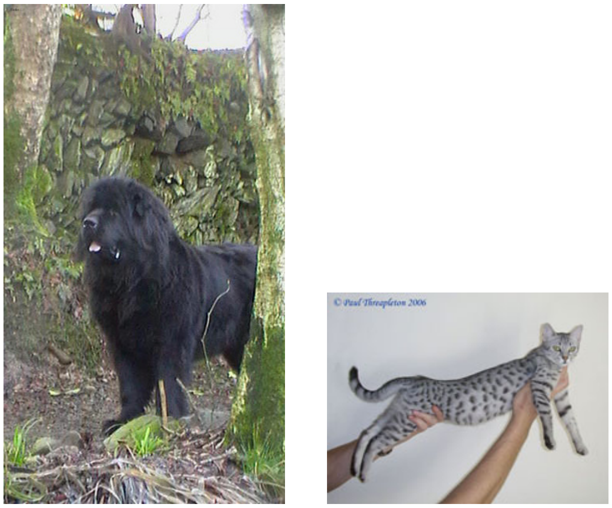
This study leverages a pre-trained CLIP model configured with a ViT-B/16 backbone for the image encoder, transformer-based text encoder. The weights for the CLIP model are sourced from its official implementation. During training, all parameters in the image and text encoders remain fixed, ensuring that feature extraction is consistent across all stages. To enhance generalization, the CoCoOp framework is employed, which trains a prompt learner tailored for the dataset. Object cropping is handled using the YOLOv8-nano model, pre-trained on the COCO dataset. The experiments are conducted using Google Colab, with PyTorch as the primary framework and an NVIDIA T4 GPU for computation. For training, the process is run with a batch size of 4 over 10 epochs. An Adam optimizer is employed with a learning rate initialized at 0.0005 Default hyperparameters for prompt tuning in CoCoOp are applied throughout.

3.3. Result

Dataset	Total Images	Selected Categories	Cropped Images	No Detection
Caltech-101	9,146	1,477	1,202	275
Oxford Pets	7,349	7,349	7,036	313

Table 1. Object detection and cropping results for Caltech-101 and Oxford Pets datasets.

Table 1 summarizes the number of cropped images obtained through object detection for the two datasets, Caltech-101 and Oxford Pets. In the case of Caltech-101, 11 categories were identified as corresponding to classes in COCO, allowing object detection to be performed on 1,477 images across these 11 categories. Among these, objects were successfully detected in 1,202 images, while 275 images failed to be detected. For Oxford Pets, since all categories correspond to either dogs or cats, object detection was attempted on all 7,349 images. As a result, 7,036 images were successfully detected and cropped, while 313 images either lacked detectable objects or were misclassified into unrelated categories, making cropping unsuccessful.



(a) Newfoundland

(b) Egyptian Mau

Figure 2. Examples of images detection error from the dataset.

Figure 2 illustrates examples of images where the detected object's class label did not match the given category, resulting in unsuccessful cropping. In Figure 2 -(a), although the image depicts a Newfoundland dog, the object detector misclassified it as a bear, which did not align with the provided category, "dog," and thus, the cropping was not performed. Similarly, in Figure 2 -(b), the image features an Egyptian Mau cat, but the detector classified a human arm in the image as a "person," which did not match the intended category, "cat," leading to a failure in cropping

Dataset	**Base (%)**	**Novel (%)**
Caltech-101	97.5	92.9
Caltech-101 + Crop	**97.7**	**93.2**
Oxford Pets	**94.7**	**97.7**
Oxford Pets + Crop	93.8	97.3

Table 2. Performance comparison of Base and Novel categories with original and cropped images.

Table 2 presents a performance comparison between original images and cropped images when processed through CLIP's image encoder for two datasets. For Caltech-101, the Base performance improved from 97.5% to 97.7%, reflecting a 0.2% increase, the Novel performance increased from 92.9% to 93.2%, showing a 0.3% increase. On the other hand, for Oxford Pets, the Base performance decreased from 94.7% to 93.8%, resulting in a 0.9% drop, and the Novel performance also declined from 97.7% to 97.3%.

Although the performance improvements for Caltech-101 were small, they are notable as they occurred at an already high baseline performance level for both Base and Novel categories. Additionally, the results suggest that the application of cropped images influenced the two datasets differently: in Caltech-101, where the proportion of cropped images was relatively low, performance improved, while in Oxford Pets, which had a higher proportion of cropped images, performance declined. This indicates that applying cropped images selectively may provide a balance between the advantages of diversity in training data and focusing on object-specific features. The results suggest that while cropping helps the model concentrate on relevant objects, a high proportion of cropped images can reduce training diversity and eliminate potentially useful background information, ultimately leading to performance degradation.

4. Future direction

This study compared the performance of CLIP when cropped images were used as input to its image encoder against the performance with original images. Experiments were conducted on two datasets, each with different proportions of cropped images, leading to varying performance outcomes. Future research is needed to determine whether these differences arise from the nature of the datasets themselves or from the varying ratios of cropped images.

Additionally, this study employed YOLO, pretrained on the COCO dataset, to perform object detection, which was then used to crop the images. For future work, we plan to enhance the accuracy of object cropping by utilizing object detectors pretrained on datasets specifically aligned with the selected datasets. This approach aims to improve detection precision and enable more robust evaluations of the proposed methodology.

References

Fei-Fei, L., Fergus, R., and Perona, P. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. Technical report, California Institute of Technology, 2004.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Parkhi, O. M., Vedaldi, A., and Zisserman, A. Cats and dogs. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3498–3505, 2012.

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

Redmon, J. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

Zhou, K., Yang, J., Loy, C. C., and Liu, Z. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16816–16825, 2022.