OOD-CV Challenge Report

September 18, 2023

1 Team details

- Challenge track:
  OOD-CV Challenge 2023 (Detection Track - ImageNet-1k leaderboard)

- Team name:
  Six Walnuts

- Team leader name:
  Rui Peng

- Team leader address, phone number, and email:
  Xi'an, Shaanxi, China
  +86-18710792770
  22171214876@stu.xidian.edu.cn

- Rest of the team members:
  Junpei Zhang
  Kexin Zhang
  Long Sun
  Licheng Jiao
  Lingling Li

- Team website URL: None

- Affiliation:
  School of Artificial Intelligence, Xidian University, Xi'an, China

- User names on the OOD-CV Codalab competitions:
  arui

- Link to the codes of the solution(s):
  https://github.com/SoftSisterRui/OOD-CV-Det-2023

2 Contribution details

- Title of the contribution :
  Object detection based on DINO and YOLOX

- General method description:
  The overall solution is implemented based on YOLOX and DINO.
  YOLOX adopts a multi-scale cascade detection structure with multiple detection heads, each of which is responsible for target detection at different scales. This helps detect targets of different sizes and resolutions. The backbone uses CSPDarknet, a deeply separable backbone network that helps extract features. YOLOX uses an adaptive learning rate strategy to improve the stability of the model during the training process. While maintaining high accuracy, it achieves faster inference speed through the optimization of the model structure and efficient inference.
  The core idea of DINO is to train the model through comparative learning. The model first maps the input data into a high-dimensional feature space, and then minimizes the similarity of the same samples in the same batch by maximizing the similarity of different samples in the same batch. This enables the model to better capture the structural and semantic information of the data. Backbone uses

SwinTransformer, which uses a hierarchical attention mechanism to decompose the image into different blocks and then perform self-attention calculations between these blocks to better capture long-distance contextual information.

We use different models for different categories in the data. Data augmentation includes multi-scale training, RandomAffine random affine transformation, MixUp, Mosaic, etc. Mosaic combines four images into one image, usually from different images in the training set, and stitches them into a larger image to increase the diversity of the data. MixUp data augmentation mixes two images together in a certain proportion to generate a new training sample. The loss function uses CrossEntropyLoss and IoULoss in YOLOX, and the loss function uses L1Loss and GIoULoss in DINO. In addition, for out-of-distribution data, we specially performed data augmentation operations such as adding occlusion and weather changes to the data. When using the DINO model test data, we found that tta can effectively improve the model's performance on the test set.

- Description of the particularities of the solutions deployed for each of the tracks :

  For different data categories, we choose different models. For example, our experiments found that the DINO model performed better than YOLOX in the IID and shape categories. We finally chose to use the DINO model in these two categories and the YOLOX model in other categories. We performed data augmentation separately for different data categories. In addition to basic data augmentation operations, we also added operations such as occlusion and different weather changes to the image.

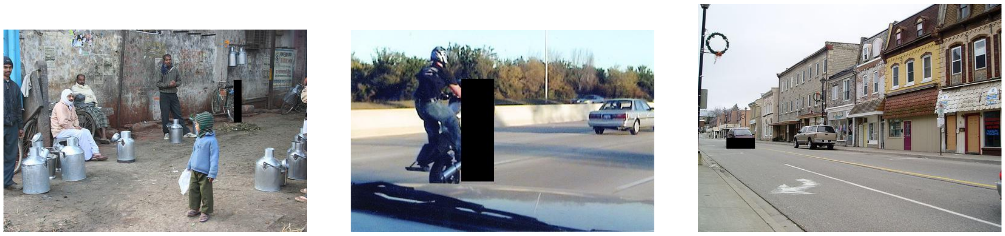
- References:

1. Sorting backbone analysis: A network-based method of extracting key actionable information from free-sorting task results[J]. Jacob Lahne. Food Quality and Preference.

2. Eye movement strategies in overall similarity and single-dimension sorting.[J]. Milton Fraser,Wills Andy. Proceedings of the Annual Meeting of the Cognitive Science Society.

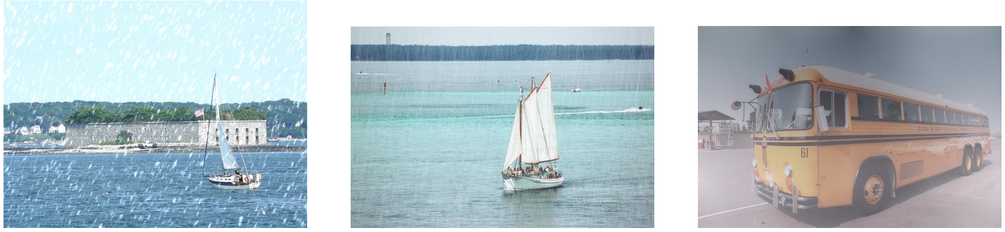
3. Approach based on high-performance liquid chromatography fingerprint coupled with multivariate statistical analysis for the quality evaluation of Gastrodia Rhizoma.[J]. Zhang Xuerong,Ning Ziwan,Ji De,Chen Yi,Mao Chunqin,Lu Tulin. Journal of separation science.
4. Occlusion aware underwater object tracking using hybrid adaptive deep SORT -YOLOv3 approach[J]. Mathias Ajisha,Dhanalakshmi Samiappan,Kumar R.. Multimedia Tools and Applications.

- Representative image / diagram of the method(s):



(a) bicycle (b) person (c) car

Figure 1: Add occlusion



(a) snow (b) rain (c) fog

Figure 2: Add weather changes

3 Global Method Description

[* Indicates method used in competition test results.]

- Total method complexity:
  YOLOX-X:
  Model complexity: YOLOX-X uses CSPDarknet as the backbone network, which is a deeply separable network structure for feature extraction. A multi-scale cascade detection structure is adopted, including multiple detection heads, each detection head is responsible for target detection at different scales, which increases the complexity of the model.
  Data processing and preprocessing complexity: YOLOX uses data enhancement methods such as multi-scale training, RandomAffine random affine transformation, MixUp, Mosaic, etc., which increase complexity in the data processing and preprocessing stages.
  Training strategy and skill complexity: YOLOX uses an adaptive learning rate strategy to improve the stability of training, and also optimizes the model structure to achieve faster inference speed. These strategies and techniques also add to the complexity of the overall approach.
  Post-processing and evaluation complexity: YOLOX uses post-processing methods such as non-maximum suppression to improve detection performance, which requires additional calculations and processing.
  DINO-SwinLarge:
  Model complexity: DINO-SwinLarge uses the Swin Transformer backbone network, which is a larger model with multiple layers and parameters, so the model complexity is high.
  Data processing and preprocessing complexity: DINO requires extensive data augmentation and preprocessing to generate image pairs with rich visual features. These complex data processing methods add to the complexity of the overall approach.
  Training strategies and trick complexity: DINO uses contrastive loss and large-batch training strategies, which require large-scale data and computing resources.

- Model Parameters:
  YOLOX Flops:0.141T Params:99.004M
  DINO Flops:0.245T Params:47.559M

- Run Time:
  After different data augmentations, the time required for training is different. It takes about seven days when there is a large amount of data, and about three days when the amount of data is small.

- Which pre-trained or external methods / models have been used:
  ImageNet-1K Pretrained Swin-V1 Models

- Training description :
  During YOLOX training, the batchsize is 8, and the total number of training epochs is 300. Multi-scale training is used during training, and data preprocessing includes Mosaic, RandomAffine, MixUp, YOLOXHSVRandomAug, etc. During DINO training, the batchsize is 1, the total number of training epochs is 36, and multi-scale training is used during training. Retraining is performed after different data augmentations.

- Testing description:
  When testing, first divide the test set into each category according to the given nuisance2filenames.json, and test each category in turn. At the same time, when testing the DINO model, we found that using tta will improve the performance of the model. When setting the specific parameters of tta, the non-maximum suppression (NMS) threshold is set to 0.65, and the maximum number of targets per image is set to 1,000. During the tta test, operations such as scaling the image to different scales, randomly flipping, and filling the image were performed.

- Quantitative and qualitative advantages of the proposed solution :
  Quantitative advantages:
  Improved Object Detection Accuracy: YOLOX-X, an advanced object detection model, provides state-of-the-art accuracy in detecting objects in images and videos. It achieves high mAP (mean average precision) scores, indicating superior performance in terms of precision and recall.

High parameter efficiency: YOLOX-X is known for its efficiency in model parameters. Compared with some other detection models, it can achieve high accuracy with relatively few parameters, which makes it suitable for resource-constrained environments.
Fast inference speed: YOLOX-X is designed for real-time or fast inference, suitable for applications requiring fast object detection, such as autonomous driving and video surveillance.
Robust to scales and resolutions: YOLOX-X?s multi-scale detection head enables it to detect objects of different sizes and resolutions in the same image, allowing it to handle a variety of object scales.
Adaptive learning rate strategy: YOLOX-X adopts an adaptive learning rate strategy to improve the stability of the model while maintaining high accuracy during the training process. This helps to converge faster and improve training efficiency.
Data Augmentation: Both YOLOX-X and DINO-SwinL benefit from data augmentation technologies such as Mosaic, RandomAffine and MixUp. These enhancements increase the diversity of training data, thereby improving generalization and accuracy.
Qualitative advantages:
Cross-domain generalization: DINO-SwinL learns rich visual representations from data. This allows it to generalize well across different domains and datasets.
Powerful feature learning: DINO-SwinL encourages models to learn powerful and discriminative features, which is very beneficial for tasks such as target detection. It captures high-level semantic information of images.
Cross-modal learning: DINO-SwinL is capable of cross-modal learning and can learn from different data modalities, such as images and text. This ability can be used for tasks that require understanding and correlating visual and textual information.

- Results of the comparison to other approaches (if any) : None

- Novelty of the solution and if it has been previously published:
  We performed specific data enhancement operations for different data categories, and selected different models for different categories based

on experimental results.

4 Ensembles and fusion strategies

- Describe in detail the use of ensembles and/or fusion strategies (if any).:None

- What was the benefit over the single method? :
  Different models often perform well on different data distributions or tasks. By using multiple models, you can better adapt to different data types.

- What were the baseline and the fused methods? : YOLOX and DINO

5 Technical details

- Language and implementation details (including platform, memory, parallelization requirements) :
  python
  2 Nvidia V100 GPUs

- Human effort required for implementation, training and validation?: Yes

- Training/testing time? Runtime at test per image :
  Depending on the data augmentation, the training time is also different. It takes about 7 days when the amount of data is large, and about 3 days when the amount of data is small.

- Comment the efficiency of the proposed solution(s)? : good

6 Other details

- General comments and impressions of the OOD-CV challenge. : satisfactory

- Other comments: None