Object Detection with VICreg and RetinaNet

Haodong Wu [*1 2] **Zixiang Pei** [*1 3] **Zhehan Shi** [*1 3]

Abstract

For this project we aim at carrying out an object detection task with variable sized input. We first researched on recent state-of-the-art methods, and then performed our downstream task using VICreg (Variance-Invariance-Covariance Regularization) (Bardes et al., 2021) to pretrain our ResNet backbone and RetinaNet (Lin et al., 2017) to finetune on labeled images. Our team achieved a mAP score of 0.125 on validation dataset and we identified several ways to improve upon our current approach as well as proposed novel methods that we should try out.

1. Introduction

In this Deep Learning final competition we carried out an object detection task. We were provided with 30000 labeled images of variable size and 512000 unlabeled images of fixed (224 x 224) size. Our goal is to compete for the highest performance on a hidden dataset, that is, we need to predict bounding boxes, labels and confidence scores given an unseen image.

Object detection is one of the most heavily focused areas recently. Before 2020, pure supervised learning approaches such as YOLO (you only look once) (Redmon et al., 2016) and Faster R-CNN (Region-based Convolutional Network) (Ren et al., 2015) predominated. However, with the emergence of methods such as SimSiam (Chen & He, 2021) and Barlow Twins (Zbontar et al., 2021), self-supervised procedure for image representation learning is becoming more popular. Those methods achieve state-of-the-art results independent of batch size and can learn with fewer labels.

In this paper, since we were only provided with a relatively small number of labeled images in comparison to unlabeled images, we decided to make full use of our resource and adapted a self-supervised learning approach. VICreg, or joint embedding architecture with variance, invariance and covariance regularization (Bardes et al., 2021), is one such approach that came out in 2022 and the original paper reported very competitive results by transferring representations learned by VICreg on a ResNet backbone to object detection using Faster R-CNN. During our further research, we also found that using RetinaNet for our downstream task might result in better performance compared to Faster R-CNN, and one-stage detectors are typically faster than two-stage detectors. So in this paper, we propose this new combination of VICreg and RetinaNet which we believe is the best approach with our provided computing power.

2. Literature Review

2.1. Object Detection

2.1.1. Supervised Learning

There are two kinds of models for supervised object detection. Two stage detectors first generate a large set of potential objects and filtering out negative locations. Then, it classifies the objects as background or foreground classes. One representation is R-CNN (Girshick et al., 2014).

One-stage detectors such as YOLO and SSD (Liu et al., 2016) carry out image classificaiton and bounding box regression directly without doing region proposal.

2.1.2. Self-Supervised Learning

Sometimes the dataset is not ideal and does not have many labels. When we have large unlabeled dataset and small labeled dataset, we need models that belong to self-supervised learning. The mechanism is usually training the model with unlabeled dataset and clustering similar objects. Then, we use that model to continue training on labeled dataset. In such way, it is able to know the correct labels of clustered objects. A paradigm in this field is called SimCLR (Chen et al., 2020) and VICreg (Bardes et al., 2021).

[*]Equal contribution [1]New York University [2]Courant Institute of Mathematical Sciences [3]Center for Data Science. Correspondence to: Zhehan Shi <zs1113@nyu.edu>, Haodong Wu <hw1635@nyu.edu>, Zixiang Pei <zp2123@nyu.edu>.

2.2. Popular Models

2.2.1. YOLOv3

YOLOv3 is a better version of the model from the original YOLO one. It maintains the fast speed, with 22 ms at 0.282 mAP on 320 × 320 images (Redmon & Farhadi, 2018). The network predicts the bounding box with dimension clusters as anchor boxes and uses logistic regression to predict the objectness score. Regarding class prediction, since each bounding box may contain multiple labels, the researchers choose the logistic classifiers and remove the softmax activation function.

The model in detail is an extension of Dark-19 but with more successive 3×3 and 1×1 convolutional layers. The name is Dark-53 because it has 53 convolutional layers. Another interesting characteristic is that YOLOv3 generates anchor boxes for three scales in order to predict both large and small objects on the images. When upsampling the image, it will merge the feature maps from previous layers and combine it to predict the bounding boxes.

2.2.2. VICREG

Vicreg is a self-supervised method for training joint embedding architectures based on the principle of preserving the information content of the embeddings (Bardes et al., 2021). It aims at maximizing information captured in the output embedding vectors from two different views of the same image. In this case, image was processed using random crops followed by color distortion. It innovates in its conceptually simple but effective three-term loss function which include a variance term, a covariance term and an invariance term. The variance and the covariance term help prevent an informational collapse in which the variables are highly correlated. The invariance term makes the two output embeddings similar. The original paper reported object detection results on par with the state-of-the-art using ResNet-50 backbone pretrained using VICreg and finetuned using Faster R-CNN.

2.2.3. RETINANET

One stage detectors have been considered faster yet less accurate than two stage detectors, (Lin et al., 2017) attributes it to the class imbalance problem which happens when a detector predicts $10^4 - 10^5$ boxes but only few contain objects. RetinaNet is proposed to solve this problem. Specifically, a new loss function named "Focal Loss" is applied in RetinaNet. The researchers find that the loss will resolve the class imbalance issue and performs a score the same as the two-stage detectors while keeping evaluation speed fast. Previously, the loss function is usually cross entropy loss for classification. Define p_t as the estimated probability of the class, the focal loss is defined as

$$\mathrm{FL}(p_t) = -(1-p_t)^{\gamma}\log(p_t)$$

According to this definition of loss, the well-classified examples yield nearly 0 loss but the poorly classified ones are penalized for large loss. The modulating factor γ also adds flexibility for hypertuning.

3. Method

3.1. Pretraining

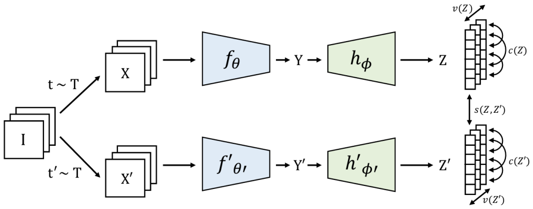

Figure 1. VICreg Structure (Bardes et al., 2021)

Initially, we attempted to implement YOLOv3. Unfortunately, we failed at the converting bounding boxes, labels and scores to proper formats. Additionally, our chosen implementation of NMS (non-maximum suppression) function was inefficient, whose time complexity was $O(N^3)$. Consequently, we pivoted to another approach which also resulted in a greater use of unlabeled images.

We pretrained ResNet backbone for 31 epochs of batch size 256 on 512,000 unlabeled images of fixed size. We trained it on 2 GPUs with 2 workers on each GPU. As displayed in Figure 1, raw images are first transformed into two branches and encoded into two representations. The representations are then fed into expanders producing two embeddings. VICreg minimizes the distance between the two embeddings with its innovative loss function and the encoder backbones are then used for finetuning.

3.2. Finetuning

We choose RetinaNet instead of Faster R-CNN because it has better object detection results (Lin et al., 2017). RetinaNet outperforms Faster R-CNN in AP, AP_{50}, and AP_{75} when using the same ResNet-101-FPN backbone. We load our pretrained backbones into RetineNet and trained on the 30000 labeled images. The RetinaNet detector was trained for 24 epochs and its performance plateaued.

4. Results

We have achieved the result of 0.125 on mAP (IoU 0.5:0.95) during our evaluation on 20,000 labeled validation images.

Epochs	Detector	Backbone	mAP (IoU 0.5:0.95)
18	RetinaNet	ResNet50 (VICreg)	0.125
24	RetinaNet	ResNet50 (VICreg)	0.125

Table 1. mAP results

4.1. Sample Predictions

Although the the predicted bounding boxes correctly identify the ground truth as a dog, it also includes many irrelevant boxes. However, all the boxes, whether relevant or not, by and large bound the correct locations (coordinates) of the objects.

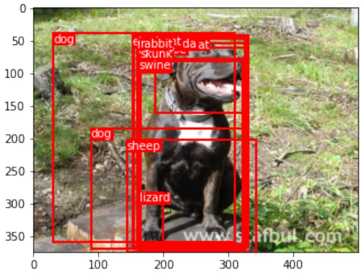

Figure 2. Sample Prediction

4.2. Feature map analysis

The feature maps (Figure 3) show the result of three images after being passed into the last Conv2d network of the respective 4 layers of the trained RetinaNet backbone. We displayed an example of a dog, a mushroom and a race car. Take the example of the dog (Figure 2) in the previous section. From its feature map (row 1), at layer 1 the model learns about the black parts of its body; at layer 2 the model learns about the contour and the white parts of the body; at layer 3 the bottom half of the dog is learned; at layer 4 we can only tell the model is learning many details on and around the dog. From all three images we can conclude that as the image moves deeper into the neural network, it is harder to interpret what the model is learning.

5. Discussion

For this competition we did not score very high on the leaderboard and we identified the following problems:

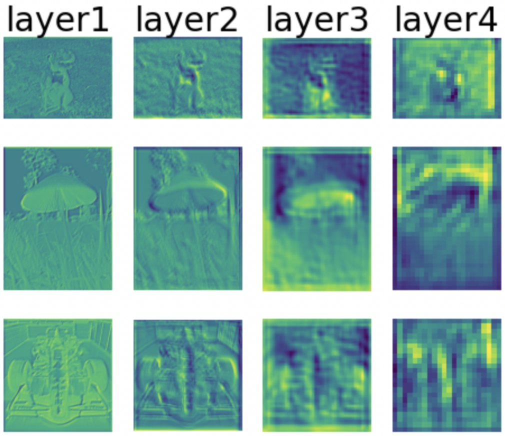

Figure 3. Feature Map of RetinaNet Backbone

Nonconvergence: This is a major downside of our approach. As we only had limited time and we devoted part of our time to early research and other approaches such as YOLOv3, we were not able to pretrain our backbone until convergence. We have only pretrained our backbone for 30 epochs and from the original VICreg paper, it typically suggested 100 epochs of pretraining. If more time allowed, we would continue pretrain until convergence.

Small Learning Rate and Batch Size: For our finetuning stage we used a learning rate of 5e-5. However, the original paper used a learning rate of 0.01 on all its downstream tasks. We picked such a small learning rate because otherwise we constantly run out of memory on our GPU, even when reducing our batch size to only 4. Nevertheless, our model did not improve after 18 epochs of training and we are not sure whether and how much using a large combination of batch size and learning rate would improve model's performance.

Although our model did not converge fully and that affected its performance, our model is in general simple to understand and it is relatively small compared to other teams' models. We believe in the model's competence and we will continue exploring its power and potential.

6. After Competition Experiment

We have trained VICreg to see how long they would converge. In the end, it took around 100 epochs to converge.

Later, we used the converged ResNet50 Backbone to finetune on the 30,000 labeled training images. The finding confirms our suspicion that the convergence of the backbone greatly improves the precision. The unconverged backbone has plateaued at 18 epochs, whereas the converged backbone continued to improve after 18 epochs. The converged backbone is able to reach 0.154 mAP (IoU 0.5:0.95) in its 24th epoch, exceeding the performance of unconverged backbone.

Epochs	Detector	Backbone	mAP (IoU 0.5:0.95)
18	RetinaNet	ResNet50 (Unconverged)	0.125
24	RetinaNet	ResNet50 (Unconverged)	**0.125**
18	RetinaNet	ResNet50 (Converged)	0.125
24	RetinaNet	ResNet50 (Converged)	**0.154**

Table 2. mAP results (experiments)

7. Future Approach

During the presentation and our further research, we have also discovered new and interesting approaches from our classmates.

7.1. Vision Transformer

In the future, we would be more inclined to replace the traditional ResNet backbone with ViT (Vision Transformer) backbone (Li et al., 2022). We would first pretrain ViT using MAE (Masked Autoencoders) (He et al., 2022), and then put the pretrained ViT into the RetinaNet.

7.2. Consistent Teacher

Inspired by the Unbiased Teacher (Liu et al., 2021) framework addressing the inherent class imbalance in the labeled training set, we did some further research. As a result, we discovered Consistent Teacher (Wang et al., 2022), where it outperforms the Unbiased Teacher.

7.3. Next Time Approach

In the future, we would do the following to be better at object detection.

1. Pretrain ViT using MAE
2. Put ViT Backbone into RetinaNet
3. Use Consistent Teacher to finetune

We believe this particular combination is our future to-go combination for similar objection detection tasks.

References

Bardes, A., Ponce, J., and LeCun, Y. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Chen, X. and He, K. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15750–15758, 2021.

Girshick, R., Donahue, J., Darrell, T., and Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 580–587, 2014.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022.

Li, Y., Mao, H., Girshick, R., and He, K. Exploring plain vision transformer backbones for object detection. *arXiv preprint arXiv:2203.16527*, 2022.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., and Berg, A. C. Ssd: Single shot multibox detector. In *European conference on computer vision*, pp. 21–37. Springer, 2016.

Liu, Y.-C., Ma, C.-Y., He, Z., Kuo, C.-W., Chen, K., Zhang, P., Wu, B., Kira, Z., and Vajda, P. Unbiased teacher for semi-supervised object detection. *arXiv preprint arXiv:2102.09480*, 2021.

Redmon, J. and Farhadi, A. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779–788, 2016.

Ren, S., He, K., Girshick, R., and Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

Wang, X., Yang, X., Zhang, S., Li, Y., Feng, L., Fang, S., Lyu, C., Chen, K., and Zhang, W. Consistent targets provide better supervision in semi-supervised object detection. *arXiv preprint arXiv:2209.01589*, 2022.

Zbontar, J., Jing, L., Misra, I., LeCun, Y., and Deny, S. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pp. 12310–12320. PMLR, 2021.