Self-supervised learning for Object Detection

Haresh Rengaraj Rajamohan
hrr288@nyu.edu

Xinhao Liu
xl3136@nyu.edu

Zecheng Wang
zw2374@nyu.edu

Abstract

In this work, we applied self supervised methods like Barlow Twins and VICReg to pre-train a Resnet50 image processing backbone and then used this backbone in Faster R-CNN and fine tuned for the task of object detection. The Faster R-CNN model trained with Barlow Twins pretrained initialization achieved an mAP of 0.265 on the testset

1 Literature Review

1.1 Self-Supervised Pretraining

A popular class of self supervised pretraining involves the use of Joint Embedding Predictive Architectures (JEPA), where the model is trained to be invariant to a set of data augmentation transforms applied to the unlabeled images. This approach is mainly of two types (i) Contrastive and (ii) Non- contrastive learning.

1.1.1 Contrastive learning

In contrastive methods, positive pairs and negative pairs are used for learning representations. Positive pairs here are two augmented views of the same image and negative pairs are two augmented views for different images. The model is trained by minimizing the energy between positive pairs and maximizing the energy between negative pairs maximised.

Chen et al. (2020) proposed SimCLR, where a mini-batch with both positive, negative pairs are sampled and the cosine similarity between the outputs of positive pairs minimised and cosine similarity between negative pair outputs maximized. Major problem with the SimCLR is that it requires large batch sizes with large number of negative samples to learn good representations. He et al. (2020) proposed MoCo, which has a query encoder, a momentum key encoder and maintains a dictionary of these key representations, that can be reused to compute the contrastive loss. This removes the need for large mini-batch computations, as the key representations can be reused.

1.1.2 Non - Contrastive learning

Non-contrastive methods minimize the energy between positive pairs and use regularization or architectural constraints to avoid model collapse (flat energy surface). Caron et al. (2021) proposed DINO, a non-contrastive SSL framework which uses a teacher student framework to learn representations. They use the momentum encoder idea from MoCo and stop gradient to prevent model collapse.

Barlow Twins was proposed by Zbontar et al. (2021). As opposed to DINO, it uses symmetric encoders (i.e, the two encoders are identical here). It uses regularization to prevent model collapse. Here, the outputs from two augmented views of the image are obtained and then the cross correlation

Preprint. Under review.

matrix over a batch between them is pushed to identity. So the diagonal terms being pushed to 1 leads to invariance to augmentations. The off-diagonal terms being pushed to 0 ensures that there's no redundancy in the learned representations (dimensional collapse).

Recently Bardes et al. (2022) proposed VICReg, which uses three terms to prevent model and dimensional collapse. The invariance term, where the mean squared distance between the two outputs from the two augmented views to ensure invariance to augmentations. Variance of the outputs over the batch is pushed to be greater than a threshold to ensure that the model predicts different outputs. The covariance between all pairs of variables over the batch output is pushed to 0 to prevent redundancy. VICReg also has symmetric structure. Another class of SSL methods is clustering based- Caron et al. (2018) used kmeans clustering to get pseudo labels which were used for learning representations. Caron et al. (2020) proposed swav, a combination of clustering approach and contrastive instance learning.

1.2 Object Detection

Girshick et al. (2014) proposed R-CNN, which used selective search to generate region proposals and a CNN is used to extract feature representations, predict bounding box offsets from these regions and then a SVM is used to predict the object class. To make it faster, Girshick (2015) proposed Fast R-CNN, where the whole input is passed through the CNN to get a feature map and then region proposals are extracted from this feature map with selective search, RoIpooling and then classified using a fully connected layer. This was further improved in Faster R-CNN (Ren et al., 2015), where instead of selective search, another network is used to predict region proposals.

Recently, DETR (Carion et al., 2020) was proposed, that uses a transformer decoder encoder with a bipartite matching loss to achieve end to end object detection. To pretrain transformer encoder decoder, up-detr (Dai et al., 2021) was proposed which uses "random query patch detection" where patches are randomly cropped from the input and then passed as queries to the decoder and the model is trained to detect these patches from the original input.

2 Method

Contrastive methods typically require large batches with lots of negative pairs to work. Due to compute, time constraints only non-contrastive methods were used in this work- Barlow Twins and VICReg. From the default setting, the scale range for randomized resized crop was changed from [8%,100%] to [40%,100%] to ensure that representations retain global visual features of the image. The images were normalized with mean (0.49, 0.468, 0.414) and standard deviation (0.286, 0.278, 0.297) computed from the dataset. Further, the dimension of MLP projector in both the methods was decreased from 8192 to 2048 to reduce the complexity (and avoid overfitting) , as the number of unlabeled images here (512K) is significantly less than Imagenet. Resnet50)He et al. (2016)) backbone was used for the barlow twins and Vicreg pretraining. Additionally a swin transformer (Liu et al., 2021) was trained using barlow twins. The barlow twins model was trained for 150 epochs, VICreg ones were trained for 100 epochs and their losses had stabilized at this point.

These backbones were added to a Faster R-CNN model and trained on the provided labeled training daatset (30K images) for object detection. The Resnet50 models were loaded and fine tuned on the labeled dataset. Since the default SGD led to poor mAPs (< 0.1), Adam optimizer was used. The model was trained for 20 epochs with various learning rates and we observe best performance for a learning rate of 1×10^{-4} in both Barlow Twins and VICReg pretrained backbones. Since the models didn't converge in 20 epochs, they were trained for another 20 epochs with a lower learning rate of 5×10^{-5}. Everything else was maintained same as it was in the demo code.

Faster R-CNN with Barlow Twins pretrained Swin Transformer was trained for 35 epochs, when the loss converged. Additionally, DETR was used in this work with Barlow Twins pretrained Resnet50 and transformer encoder decoder pretrained on the random patch query detection task for twenty epochs.

3 Results

We evaluate the performance of our models on the validation dataset (20K images). Using the methods from above, we achieve the best result of 0.281 mAP using Faster R-CNN with ResNet50 backbone pretrained with Barlow Twins. The Swin transformer and DETR approaches perform poorly achieving mAP close to 0. The DETR model predominantly seems to predict no-object classes for all the object queries and the train loss stagnating.

Table 1: Results of our experiments.

Backbone	Pretraining	Detector	mAP (IoU 0.5:0.95)
ResNet50	Barlow Twins	Faster R-CNN	**0.281**
ResNet50	VICReg	Faster R-CNN	**0.273**
ResNet50	Barlow Twins	DETR	0
ResNet50	Barlow Twins	DETR (UP-DETR)	0.002
Swin Transformer	Barlow Twins	Faster R-CNN	0.012

4 Visualization and Analysis

4.1 Detection result

Fig. 1 shows the detection result of some images in the validation dataset. When the object is big, and the object is not cluttered, the model is able to detect the object correctly with high confidence (high prediction score). On the other hand, when the input is cluttered, the model correctly predicts the bounding box, but fails in classification. For example, in pair (g), the water blurs the person and make sweater looks like a bear's fur. In pair (h), the model may misinterpret the person's hair as a dog's fur.

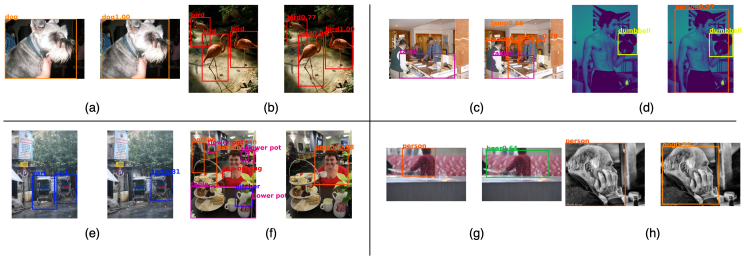
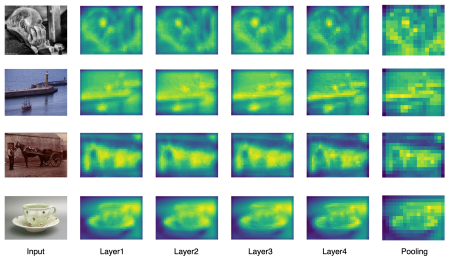


Figure 1: Detection results of our method. For each pair of images, the left shows the GT bounding box and labels, and the right one shows the model's prediction and score. The top-left quadrant shows correct predictions. The top-right quadrant shows examples where missing labels in the ground are being detected. In the bottom-left quadrant, some objects are missing in the prediction. The bottom-right quadrant shows incorrect predictions.

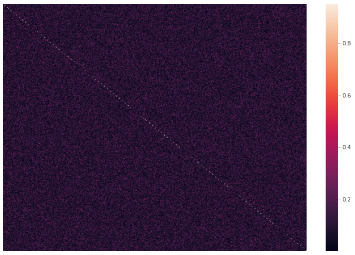
When there are multiple objects with little overlap, and their sizes are similar, the model is able to detect all of them correctly, such as in pair (b). Otherwise, the model prediction misses some objects in the image, especially the smaller ones, such as in pair (f). We also observe that the model is able to detect some objects that are not labeled in the ground truth - probably due to human labeling error. It shows that the image processing backbone has learned to retain and focus on the object of interest.

4.2 Feature map

Figure 2a is a visualization of the feature maps with different inputs. The objects of interest seem to be highlighted in these feature maps well. Figure 2b shows cross correlation matrix between a batch of samples sampled randomly from the unlabeled dataset. It seems that the diagonal elements are close to 1 and off-diagonal elements close to 0 as expected. It seems that not all of the diagonal elements are close to 1 and imilarly not all of off-diagonal elements close to 0 as expected. This trend is surprising as during training, the loss stagnated after 110 epochs and did not go down further. To understand whats happening here,m odels with different embedding dimensions need to be trained and their cross-correlation matrices need to be visualized.



(a) Visualization of feature maps from each layer of the feature pyramid network



(b) Cross correlation matrix computed over batch of 64 augmented positive pairs from the unlabeled dataset

References

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

He, K.; Fan, H.; Wu, Y.; Xie, S.; Girshick, R. Momentum contrast for unsupervised visual representation learning. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020; pp 9729–9738.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; Joulin, A. Emerging properties in self-supervised vision transformers. Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021; pp 9650–9660.

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

Bardes, A.; Ponce, J.; LeCun, Y. VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning. International Conference on Learning Representations. 2022.

Caron, M.; Bojanowski, P.; Joulin, A.; Douze, M. Deep clustering for unsupervised learning of visual features. Proceedings of the European conference on computer vision (ECCV). 2018; pp 132–149.

Caron, M.; Misra, I.; Mairal, J.; Goyal, P.; Bojanowski, P.; Joulin, A. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems* **2020**, *33*, 9912–9924.

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

Girshick, R. Fast r-cnn. Proceedings of the IEEE international conference on computer vision. 2015; pp 1440–1448.

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

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. European conference on computer vision. 2020; pp 213–229.

Dai, Z.; Cai, B.; Lin, Y.; Chen, J. Up-detr: Unsupervised pre-training for object detection with transformers. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021; pp 1601–1610.

He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016; pp 770–778.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021; pp 10012–10022.