Research Proposal

Candidate: Boxuan Zhang

Title: Image-Object Co-Assistance for Self-Supervised Training

1 Introduction

In recent years, deep learning has achieved satisfactory performance for a range of computer vision tasks. Generally, large-scale, high-quality labeled data is a prerequisite for obtaining a high-performance deep learning model. However, there are significant impediments to large-scale labeling tasks, especially object-level labeling and even pixel-level labeling for object detection and semantic segmentation, mainly because of the requirement of enormous human, material, and financial resources. Self-supervised learning(SSL) has become the core technique to construct models with strong generalization ability with limited labeled data, thereby addressing the complex and variable scenarios in real-world applications. Recent SSL methods mainly build the unsupervised representation of images by introducing pretext tasks, ranging from generative learning and contrastive learning. The training goals are mainly to improve the feature representation and discrimination performance of the model, which in turn can be used for a number of downstream tasks.

Despite the substantial progress, the features extracted by the existing SSL methods still have a low correlation with the discriminative features of the objects on the image. This stems from the fact that they will primarily be pre-trained on the object-centric ImageNet dataset, which results in their low discriminative capability for feature learning in complex real-world scenarios (i.e., numerous objects, complicated structures and backgrounds), thereby failing to obtain representations that have a gain on the performance of image decoding models. Therefore, it is crucial to design an effective learning paradigm for object-level representations in order to thoroughly leverage the large number of unlabeled real-world images. In fact, the latest work, called ORL, has proposed a solution for object-level unsupervised representation learning, which is a multi-stage framework based on several contrastive learning modules. After a standard image-level BYOL pre-training, ORL leverages the pre-trained model to discover the corresponding object-instance pairs that are finally used for object-level BYOL representation learning.

However, there are some non-negligible deficiencies. ORL utilizes the unsupervised pre-trained model from image-level tasks, which may introduce inconsistent learning signals since random crops of the same image may contain different objects. At the same time, the object-level contrastive learning module adopted by ORL needs both positive and negative RoIs for representation learning, relying on the precision of the discovered corresponding object-instance pairs, which is challenging to warrant by the unsupervised pre-trained model above. In addition, due to the relatively roughness of the discovered RoIs, directly learning object-level representations from them may lose some possibly existing small objects. Hence, it is imperative to propose a method with both a more effective image-level pre-training model and more precise object-level representation learning.

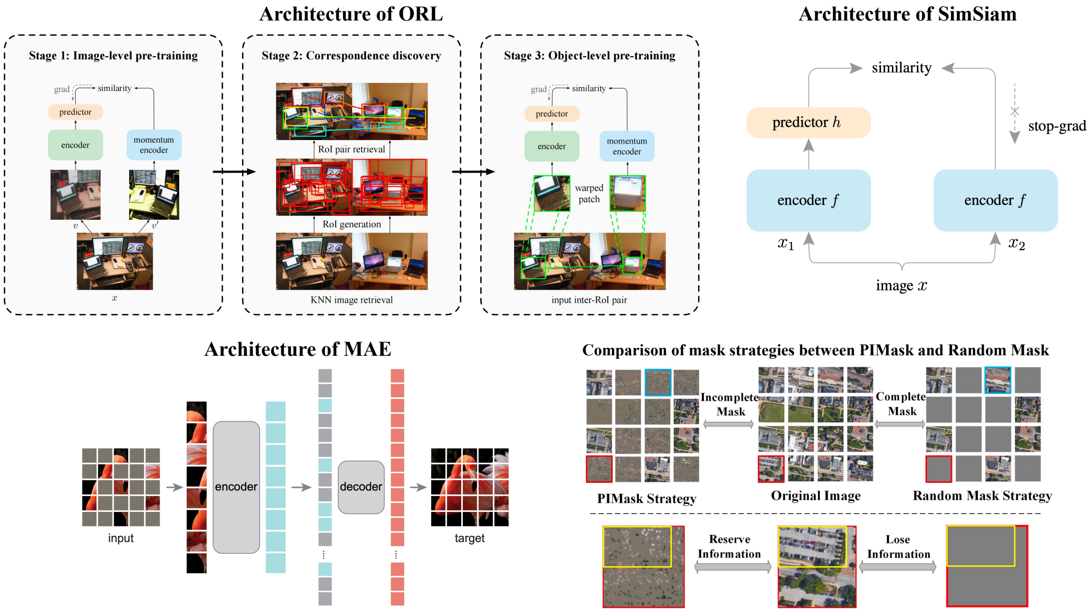

Figure 1: Overview of the highlighted architectures.

2 Related Work

As a branch of unsupervised learning, SSL commonly expects to learn a generalized feature representation for downstream tasks. In recent years, the mainstream methods can be classified into generative-based and contrastive-based. Generative self-supervised methods learn representations by reconstructing or generating input data. A prominent set of methods in this category are autoencoders(AE), which was first introduced by [1] to serve as pre-training for artificial neural networks. After that, some well-acclaimed approaches, such as denoising autoencoder(DAE) [2] and variational autoencoder(VAE) [3], were proposed. In particular, DAE is trained to enhance robustness against noise in data, while VAE decouples the encoder and the decoder by encoding a latent distribution. Recently, Masked image modeling(MIM) [4], [5] has gained great interest as a self-supervised learning method applying the generative model. It aims to reconstruct the original image from the masked pixels and learn the general feature representation among the data distribution, which can be classified as a DAE model. Among the latest MIM-based methods, Masked Autoencoder(MAE) [4] raised great attention in the computer vision community with a breakthrough in autoencoding self-supervised pre-training of vision transformers. Inspired by DAE, MAE masks out random patches of the input image, sends visible patches to the encoder, and reconstructs the missing patches from the latent representation and masked tokens. This work proves the potential of transformer-based autoencoders for self-supervised visual representation learning. Meanwhile, SimMIM [5] designed a simple masked image modeling method and replaces masked patches with learnable mask token vectors. Masked and visible tokens are fed together as input to predict the original pixel values of masked patches by the encoder decoder, which allows SimMIM to be applied to the Swin Transformer as well. Considering the previous mask strategies(i.e., random masking) tended to ignore dense and small objects in complicated scene images, RingMo [6] designed a Patch Incomplete Mask (PIMask) strategy, where the objects are preserved by reserving some pixels in masked patches while keeping the overall mask ratio.

Another category of SSL is contrastive learning, which allows the network not to rely on a single pretext task, thus learning a high-level representation. Generally, contrastive-based methods train a model by contrasting semantically identical inputs((e.g., two augmented views of the same image) and pushing them to be close-by in the representation space, which can be regarded as a common Siamese-like architecture design [7]. A distinguished set of methods for contrastive SSL is based on knowledge distillation, which is commonly found in the teacher-student framework and optimizes a similarity metric of two augmented views of the same input image. Grill et al.[8] proposed the first milestone called BYOL for self-supervised learning based on knowledge distillation. The general architecture is similar to MoCo [9] but without the usage of negative samples. It was shown that, if a fixed randomly initialized network (which would not collapse because it is not trained) is used to serve as the key encoder, the representation produced by the query encoder would still be improved during training. If then the target encoder is set to be the trained query encoder and iterates this procedure, it will progressively achieve better performance. Therefore, BYOL proposed an architecture with an exponential moving average strategy to update the target encoder just as MoCo does, and used mean square error as the similarity measurement. SimSiam [10] presented a systematic study on the importance of different tricks in avoiding collapse and proposed a simplified version of the previous self-supervised contrastive methods, arguing that the additional predictor of BYOL is helpful but not necessary to prevent model collapse. Instead, the stop gradient operation of the teacher (target) network is the most critical component to make target representation stable. In addition, the authors showed the relationship between SimSiam and other popular contrastive methods: (1) SimSiam can be thought of as “SimCLR [11] without negatives”; (2) SimSiam is conceptually analogous to “SwAV [12] without online clustering”; (3) SimSiam can be seen as ”BYOL without the momentum encoder”.

3 Research Objectives

The objectives of this research proposal are listed as follows:

- Objective 1: By designing an efficient self-supervised learning paradigm based on ORl, we make full use of the enormous unlabeled data to improve the performance of decoding the high-level semantics of images, thus significantly reducing the cost of labeling in practical applications.
- Objective 2: The candidate plans to propose a more excellent self-supervised learning scheme assisted by Image-Object collaboration. It adopts a more effective unsupervised region discovery method and can effectively capture the small objects that may be missed in the region to realize more precisely object-level discriminative representation learning.

4 Research Methodology

Although the existing approaches are generally efficient, there still needs to be a higher correlation between the representations learned by most of them and object-level discriminative features. Despite some methods, such as ORL, have proposed a object-level representation learning solution, they may still ignore small objects that may be contained in the discovered regions as they are obtained by an unsupervised pre-training model. Finally, this research proposal is thus aiming to address the issues of existing methods, which may be detailed as follows:

- Stage 1. Efficient unsupervised target region discovery: In the first stage, the candidate employs an generative-based image-level pre-trained model to explore pairs of regions of interest in KNN neighbor images for object-level representation learning. Firstly, the candidate uses a random-masked autoencoder(MAE) to construct the image-level SSL model. The loss function $\mathcal{L}_{\text{image}}$ can be calculated by the mean square error of reconstructed and raw images in pixel space:

$$\mathcal{L}_{\text{image}} = \frac{1}{\Omega\left(x_M\right)}\left\|y_M - x_M\right\|_2^2 \tag{1}$$

After that, as illustrated in the Stage 2 of ORL, for each input image, the K nearest neighbor image pairs set N_K containing similar visual contexts are retrieved in the embedding space by cosine distance from the feature representations learned from the first stage. Next, the candidate proposes to apply an unsupervised region proposal algorithm to the images and used some predefined thresholds to filter the proposed regions(RoI), remaining top-m RoIs for each image. Finally, different from ORL, for each RoI from the query image, the candidate plans to search all the RoIs from images in N_K and compute their cosine similarity in embedding space. With the obtained similarity matrix $M_k \in R^{m\times m\times K}$, the top n nearest-neighbor RoI pairs for each image in N_K to construct the set of object pairs, denoted as $\{B_k^n\}$. The RoIs in the query image and $\{B_k^n\}$ are saved for the next stage.

- Stage 2. SSL of collaborative image-level and object-level: In the second stage, the candidate proposes to combine contrastive and generative learning methods for object-level representation learning. At the last of Stage 1, we obtain the RoIs from the query image and $\{B_k^n\}$. In Stage 2, RoIs in the query image are used to construct a self-supervised pre-training module based on object-level augmented reconstruction and to guide the generation of masks, while RoIs in $\{B_k^n\}$ are used for similarity contrastive learning based on SimSiam to update the object-level encoder.

  Considering the possible small objects missed in the region proposals, the candidate proposes to conduct *PIMask(Patch Incomplete Mask)* on RoIs in the query images, which has been proposed in [6], thus applying it to the complex real-world scenarios. In addition, the encoder in *PIMask* will be connected to a prediction MLP head [8], thereby being updated by contrastive learning with the RoIs in $\{B_k^n\}$. The candidate proposes to use the simple and effective SimSiam as the basic contrastive learning module. Hence, the loss function of object-level representation learning $\mathcal{L}_{\text{image}}$ can be calculated by the contrastive learning loss and RoI reconstructing loss:

$$\mathcal{L}_{\text{object}} = \mathcal{L}_{\text{SimSiam}} + \frac{1}{\Omega\left(o_M\right)}\left\|o_M' - o_M\right\|_1 \tag{2}$$

  where o,$o^{'}$ denotes the original and reconstructed pixel values of the target, ω is the number of elements and M denotes the set of mask pixels. Finally, the objective loss function is calculated by integrating the image-level and object-level representation learning loss, denoted as follows:

$$\mathcal{L} = \lambda_1\mathcal{L}_{\text{image}} + \lambda_2\mathcal{L}_{\text{object}} \tag{3}$$

5 Candidate Experience

The candidate has received a bachelor's degree majoring in computer science and technology, and has expected to receive a master's degree majoring in artificial intelligence in the School of Computing, Wuhan University. During the education experience, the candidate develops a solid mathematics foundation and has a good command of signal processing knowledge.

During the master's study, the candidate focuses on artificial intelligence and computer vision, and has conducted extensive research and submitted a paper as the first author. In the direction of semi-supervised object detection, one journal paper has been submitted to ***IEEE Transactions on Geoscience and Remote Sensing*** (TGRS, 2023; **IF: 8.2, RANK: Q1**). In addition, the research on "Semantic Segmentation for Open Set Domain Adaptation", which was the candidate's Undergraduate Graduation Design, has gained an Outstanding Thesis Awards in the School of Computing. These research experience has made the candidate quite **self-motivated** in doing research and has gained experience and insights in many fields such as **machine learning methods(e.g., semi-supervised learning)** and **downstream tasks of computer vision(e.g., object detection)**, which support the proposed research proposal.

In general, the candidate believes that this research proposal can be conducted successfully and make his own contributions with the help of the supervisor.

References

[1] Dana H Ballard. Modular learning in neural networks. In *Proceedings of the sixth National Conference on artificial intelligence-volume 1*, pages 279–284, 1987.

[2] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, and Léon Bottou. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(12), 2010.

[3] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[4] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[5] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9653–9663, 2022.

[6] Xian Sun, Peijin Wang, Wanxuan Lu, Zicong Zhu, Xiaonan Lu, Qibin He, Junxi Li, Xuee Rong, Zhujun Yang, Hao Chang, et al. Ringmo: A remote sensing foundation model with masked image modeling. *IEEE Transactions on Geoscience and Remote Sensing*, 2022.

[7] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature verification using a" siamese" time delay neural network. *Advances in neural information processing systems*, 6, 1993.

[8] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

[9] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[10] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758, 2021.

[11] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[12] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020.