Self-Supervised Learning Improves Agricultural Pest Classification

Soumyashree Kar, Koushik Nagasubramanian, Dinakaran Elango, Ajay Nair, Daren S. Mueller, Matthew E. O'Neal, Asheesh K. Singh, Soumik Sarkar, Baskar Ganapathysubramanian, Arti Singh*

Iowa State University, USA

skar@iastate.edu, koushikn@iastate.edu, delango@iastate.edu, nairajay@iastate.edu, dsmuelle@iastate.edu, oneal@iastate.edu, singhak@iastate.edu, soumiks@iastate.edu, baskarg@iastate.edu, arti@iastate.edu*

Abstract

Globally, crop insect pests lead to 10 – 40% yield loss. However, crop insect pest detection and mitigation remain an extremely challenging task for the farmers, due to several factors. While supervised learning has achieved a remarkable feat in insect detection, it requires significant human intervention in labeling the input data, thereby making the downstream tasks tedious and sometimes infeasible. This is particularly the case for identifying insects in the field, where labeling is tedious. Here, we present a self-supervised learning (SSL) approach – Bootstrap your own latent (BYOL) to classify 12 types of agricultural insect pests using minimal labeling. Both raw and segmented images were separately fed to the BYOL SSL method, and the linear classification accuracies from the representations learned were examined. The results indicate that using segmented images as input to BYOL could lead up to 94% classification accuracy.

Introduction

In agricultural fields insect pests pose a serious threat to yield quality and yield potential. Insect pests infestation is observed at all crop growth stages, from sowing to harvest causing serious biotic stresses in the plants. In USA, pest infestation is a major problem in soybean fields. However, it is extremely challenging to detect and identify insect pests in fields, owing to the similarities in their visual characteristics, as well as varied size and propensity to cluster together (leading to occlusions). This is exacerbated by the fact that most of these pests are not sessile and tend to hide under the leaves or fly away making trapping very difficult (Zhong et al. 2018). On many occasions, pests even colonize causing severe and widespread damage to the entire field. Therefore, early detection of the insect pests is needed to prevent such damage and facilitate precise pesticide application in the fields. Early identification of the insect pests and applying the apt pesticide in the right quantity and location would not only lower production costs and the adverse environmental impacts, but also help in contributing to better human health and food safety (Hao et al. 2020).

Several research projects have been conducted to detect the insect pests using imaging systems and computer vision techniques. Majority of them, however, rely on supervised learning approaches that require voluminous training and labeled datasets (Alliegro et al. 2020). For instance, in Tetila et al. (2020), Inception-v3, Resnet-50, VGG-16, VGG-19 and Xception models were evaluated across 5000 images for classifying 13 soybean insect pests, and the maximum classification accuracy reported is 93.82%. 98% accuracy has been reported in Li et al. (2020) in classifying ten types of crop insect pests using a manually collected dataset by fine-tuning the GoogLeNet model. However, this model implementation is both resource and time expensive. Further, large insect datasets have been published to aid insect pest classification task, e.g., IP102 dataset (Wu et al. 2019). However, even combined deep-CNN and saliency-based frameworks trained on these large datasets have failed to perform satisfactorily in case of insect pest images with large intra-class variation (Tetila et al. 2020).

While supervised methods look promising with very high classification accuracies, these involve intensive human intervention which is infeasible for large datasets. Additionally, supervised learning could hardly help with learning latent representations of the input images and enabling similarity measures between samples, in case of complex tasks like crop insect pests and disease detection and localization (Fang et al. 2021). Hence, self-supervised learning (SSL) is the state of the art (SOTA) that aims to learn useful representations from input data without any human annotations. Once these useful latent representations are learnt, downstream tasks can be performed with significantly reduced amount of labelled data. Pioneering studies in SSL have shown comparable performance between self-supervised and supervised representations (Caron et al. 2021, Grill et al. 2020). A key concept in SSL is augmenting the input to

learn the underlying representations that are invariant to the different distortions or augmentations (Misra and van der Maaten, 2019). There are several flavors of SSL based on how augmentation is performed, as well as how/which invariances and constraints are imposed. Successful approaches to SSL are broadly classified into contrastive learning, clustering, distillation and redundancy reduction-based approaches, with several SSL algorithms proposed.

In this paper the Bootstrap Your Own Latent (BYOL) method (Grill et al. 2020) is exploited to perform SSL on 12 insect classes consisting of total 9549 images. Considering the complexity of the input images SSL performance on raw and segmented images has also compared via linear evaluation of the representations learned in both the cases. Subsequently, supervised learning has also been performed to examine if the supervised and self-supervised results are comparable. The following sections, describe the dataset, methodology and the results.

Dataset

The RGB insect images were collected from the research fields of Iowa State University, USA using different mobile phones, both android and iPhones, between 8:00 AM to 5:00 PM over a period of two months. Thus, the 9549 images collected across 12 insect classes greatly varied in resolution. The images were taken from the different Iowa State University (ISU) experimental fields including soybean, mung bean, corn and various vegetable crops. Several classes were found to have large intra-class variability in terms of color, patterns, and texture of the images. Besides intra-class variability, two major challenges identified in the dataset were class-imbalance and large background with very small foreground. Due to varying illumination conditions in a day, shadow effects were also found. All the images were resized to 224 x 224 pixels, before being fed to the SSL framework.

Framework

The framework (Fig.1) comprises two major steps, pretraining and linear evaluation. In the pretraining stage, BYOL was utilized to derive the representations from the input images and in the linear evaluation stage, those representations were utilized as input to the linear classifier to assess the SSL efficiency. In many classes it was found that insects of the same class greatly differed in color and pattern, while insects from different classes looked similar. Hence, the framework was implemented on both raw and segmented images, to examine the quality of raw data as well as to study if segmentation helps in learning the representations better. Here, local entropy-based segmentation (Hržić et al. 2019) was chosen, since it segments an image based on the level of complexity in a certain section, majorly attributed to image texture than color. The initial values to the entropy function and the threshold values for masking after segmentation were empirically selected for each insect class. Fig.2(a) and (b) represent the intra-class variability and visual inter-class similarity, and Fig.3 shows the raw and segmented images of each class.

During pretraining, BYOL was implemented using both ResNet18 and ResNet50 models as the backbone and the results were compared. The ResNet models were initialized with weights trained on the Imagenet dataset (Deng et al. 2009), and then fine-tuned on the insect dataset i.e., 75% as the training and 15% validation data. The remaining 15% images were utilized as the test set for linear evaluation.

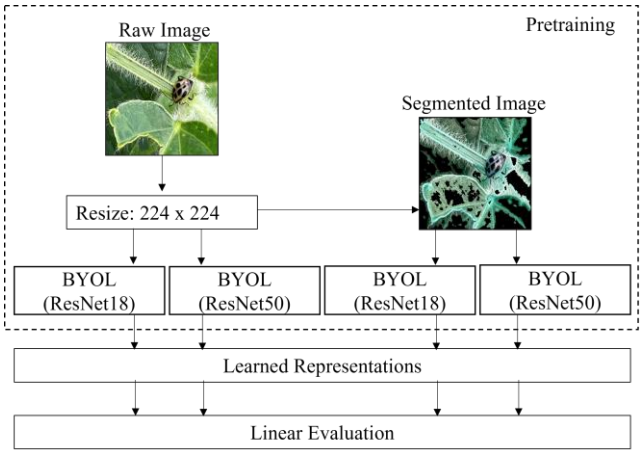

Figure 1. SSL framework for insect classification

BYOL was preferred in this study since it learns the representations from pretext tasks that can be used for downstream tasks, without relying on negative samples needed for contrastive methods. Its architecture comprises two same encoder networks, online and target that obtain representations from the image and its augmented view. Target network weights are essentially slow-moving average of the online network weights and help reduce the contrastive loss between the two representations. There are several hyperparameters, representing essential data augmentation attributes that need to be tuned for efficient SSL via BYOL. In this case, multi-cropping was enabled i.e., augmentations were performed on random crops of sizes 128,128,64 of the same image. Besides that, random grayscale conversion and random color distortion, consisting of a random sequence of color jitter, brightness, contrast, saturation, hue and gaussian blur adjustments were also applied. Pretraining was done for 800 epochs, followed by linear evaluation of the learned representations. The solo-learn library (Turrisi et al. 2021) was used for the entire implementation.

Figure 2. Examples of (a) intra-class variability showing same insect with different colors and patterns, and (b) inter-class similarity showing different similar looking insects

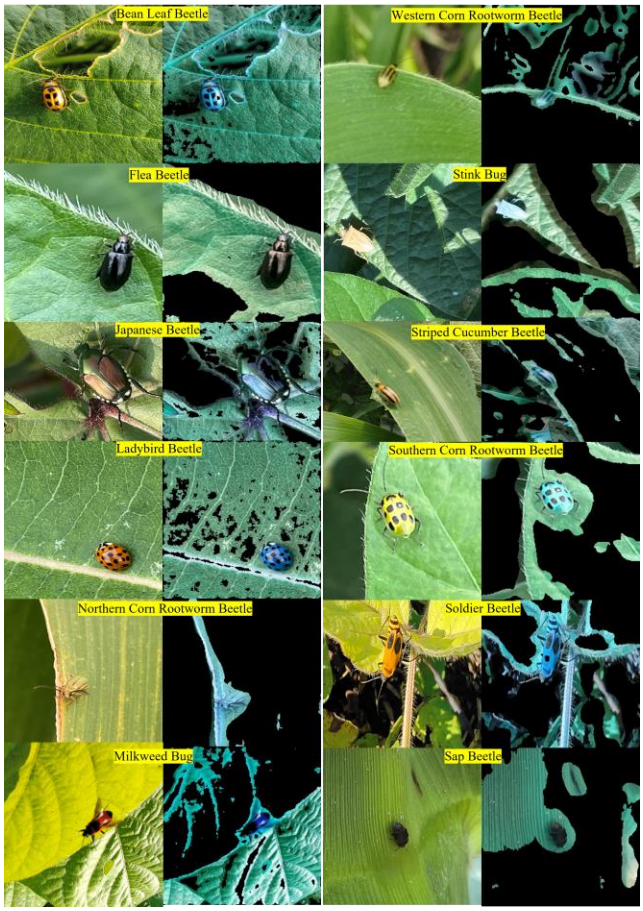

Figure 3. Raw and segmented images of each insect class.

Results and Discussion

In plant stress phenotyping studies, one of the major challenges in achieving desired efficiency from deep learning models, is caused due to the large and complex background compared to the foreground (e.g., an insect or a damaged portion of a diseased leaf). Therefore, many works have demonstrated the use of segmented images that yield superpixel regions (or superpixels), such that the networks can learn better representations of the foreground from those superpixels and enhance classification outcomes. Further, studies have also reported combining saliency networks to the deep learning frameworks, with an attempt to simultaneously explore model-interpretability and identify the visual features that helped in correct classification. However, all these analyses are performed using supervised learning, and it is shown (Nagasubramanian et al. 2020) that these interpretations are subjective to the interpretability methods used, and often spurious feature correlations were found to help in correctly classifying the images. The subjectivity and specificity of the results imply the problem of trivial solution (where the learned features do not generalize well for downstream tasks like classification, object detection, etc.) encountered in supervised learning methods that is overcome in the pretraining phase of SSL. Therefore, this study has attempted to leverage the possible benefits of segmentation and SSL in desirably classifying the insects.

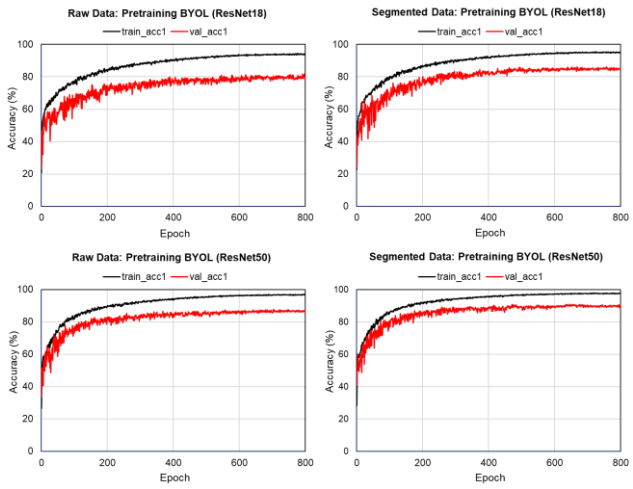

Figure 4. Training and validation accuracy curves during pretraining raw and segmented images with ResNet18 and ResNet50 architectures.

After pretraining the networks, the pretrained representations were used for linear evaluation, performed across 200 epochs. The pretraining results (Fig.4) show that, segmented images resulted in higher validation (val_acc1) as well as training accuracies (train_acc1), as suggested in previous works (Machado et al. 2016, Tetila et al. 2020). For instance,

Tetila et al. (2020) has utilized the Simple Linear Iterative Clustering (SLIC) superpixels to classify 5000 insect images. However, that method was not found beneficial for this case, owing to the large heterogeneity in the visual features within the same class, as described before. Therefore, entropy-based image subset selection was preferred, which has shown promising results in reducing the need of training data for deep learning-based segmentation in medical imaging (Gaonkar et al. 2021).

	Validation Accuracy (%)			
	Raw		Segmented	
	Res-Net18	Res-Net50	Res-Net18	Res-Net50
Linear Evaluation	86.24	89.04	89.75	**93.54**
Supervised Learning	85.23	90.32	88.53	93.34
	Validation Loss			
	Raw		Segmented	
	Res-Net18	Res-Net50	Res-Net18	Res-Net50
Linear Evaluation	0.38	0.28	0.27	0.19
Supervised Learning	0.39	0.21	0.28	0.21

Table 1. Validation accuracy and loss comparison between linear evaluation and supervised learning (with ResNet18 and ResNet50 architectures) on raw and segmented images.

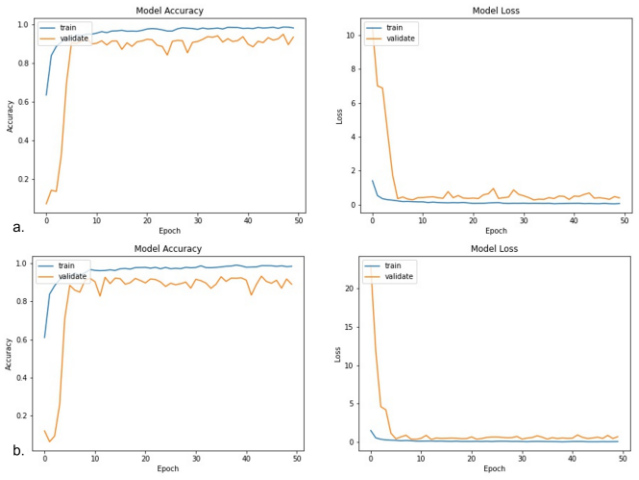

Figure 5. Validation and loss curves from supervised learning using ResNet50 on (a)segmented and (b)raw images.

The linear evaluation results (Table1) further showed that the texture-based superpixels helped in learning the distortion-invariant features better than the raw images. It was observed that the representations learned using the deeper ResNet50 architecture led to ~3% higher classification accuracy (93.539%) than ResNet18, and the validation loss was also greatly reduced (0.190). Additionally, SSL outcomes were found comparable to the supervised strategy. The validation accuracy and loss curves from supervised learning using ResNet50 on raw and segmented data is plotted in Fig.5. In Table1, the outcomes of transfer learning have also been shown, using the same ResNet18 and ResNet50 networks initialized with Imagenet weights. However, in this case the learning rate was 0.0001.

Throughout the analysis i.e., across the 4 cases, learning rate was maintained at 0.1. Given that, most of the image classes had high intra-class variability, as well as there were multiple instances of non-uniform illumination conditions (since the images were collected from real agricultural fields), multi-cropping with a wide range of augmentation parameters were selected before the runs. Each image was subjected to multiple crops i.e., 1, 1 and 6 of sizes 128, 128 and 64, followed by augmentation with brightness, color jitter, contrast, gaussian blur and hue transformations with probabilities, (0.4,0.4,0.4), (0.8,0.8,0.8), (0.4,0.4,0.4), (0.1,0.2,0.3) and (0.2,0.2,0.2), respectively. The multi-cropping feature of BYOL evidently helped in augmentation-invariant learning of the representations of each insect. In all the cases, stochastic gradient descent (sgd) optimizer was used, and the models were trained with a batch size of 64 and ReLU and softmax activations in the convolutional and dense layers, respectively.

Conclusion

This preliminary study on a dataset with real-world 9549 insect pest images of 12 insect classes, shows the potential of SSL in correctly identifying insect classes with up to 94% accuracy. In most cases, the results were either comparable or better than supervised learning. Thus, it could be deduced that the learned representations could be effectively used for further complex downstream tasks like object detection and extended to larger datasets with greater heterogeneity in the images. The significance of entropy-based image subset selection in learning distortion-invariant features has also been shown, which possibly achieved better performance by reducing intra-class variability. This is yet to be ascertained in future work by examining across the representations learned from various other SSL algorithms for different downstream tasks on plant stress phenotyping. Real-time application of such insect pests detection framework could not only help mitigation but also precise pesticide applications, thereby ensuring crop and food safety.

References

Falcon, W. and Cho, K., 2020. A framework for contrastive self-supervised learning and designing a new approach. arXiv preprint arXiv:2009.00104.

Fang, U., Li, J., Lu, X., Gao, L., Ali, M. and Xiang, Y., 2021. Self-supervised cross-iterative clustering for unlabeled plant disease images. *Neurocomputing*, 456, pp.36-48. https://doi.org/10.1016/j.neucom.2021.05.066.

Nagasubramanian, K., Singh, A.K., Singh, A., Sarkar, S. and Ganapathysubramanian, B., 2020. Usefulness of interpretability methods to explain deep learning based plant stress phenotyping. arXiv preprint arXiv:2007.05729.

Machado, B.B., Orue, J.P., Arruda, M.S., Santos, C.V., Sarath, D.S., Goncalves, W.N., Silva, G.G., Pistori, H., Roel, A.R. and Rodrigues-Jr, J.F., 2016. BioLeaf: A professional mobile application to measure foliar damage caused by insect herbivory. Computers and electronics in agriculture, 129,pp.44-55.https://doi.org/10.1016/j.copag.2016.09.007.

Gaonkar, B., Beckett, J., Attiah, M., Ahn, C., Edwards, M., Wilson, B., Laiwalla, A., Salehi, B., Yoo, B., Bui, A.A. and Macyszyn, L., 2021. Eigenrank by committee: Von-Neumann entropy based data subset selection and failure prediction for deep learning based medical image segmentation. *Medical Image Analysis*, 67, p.101834. https://doi.org/10.1016/j.media.2020.101834.

Zhong, Y., Gao, J., Lei, Q. and Zhou, Y., 2018. A vision-based counting and recognition system for flying insects in intelligent agriculture. *Sensors*, 18(5), p.1489. https://doi.org/10.3390/s18051489.

Hao, G.F., Zhao, W. and Song, B.A., 2020. Big Data Platform: An Emerging Opportunity for Precision Pesticides. Journal of Agriculture Food Chemistry. v68, 41. p. 11317-11319. https://doi.org/10.1021/acs.jafc.0c05584.

Alliegro, A., Boscaini, D. and Tommasi, T., 2021. Joint Supervised and Self-Supervised Learning for 3D Real World Challenges. In 2020 25th International Conference on Pattern Recognition (ICPR) (pp. 6718-6725). IEEE.

Tetila, E.C., Machado, B.B., Astolfi, G., de Souza Belete, N.A., Amorim, W.P., Roel, A.R. and Pistori, H., 2020. Detection and classification of soybean pests using deep learning with UAV images. Computers and Electronics in Agriculture, 179, p.105836. https://doi.org/10.1016/j.compag.2020.105836.

Li, W., Chen, P., Wang, B. and Xie, C., 2019. Automatic localization and count of agricultural crop pests based on an improved deep learning pipeline. *Scientific reports*, 9(1), pp.1-11. https://doi.org/10.1038/s41598-019-43171-0.

Wu, X., Zhan, C., Lai, Y.K., Cheng, M.M. and Yang, J., 2019. Ip102: A large-scale benchmark dataset for insect pest recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 8787-8796).

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P. and Joulin, A., 2021. Emerging properties in self-supervised vision transformers. arXiv preprint arXiv:2104.14294.

Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P.H., Buchatskaya, E., Doersch, C., Pires, B.A., Guo, Z.D., Azar, M.G. and Piot, B., 2020. Bootstrap your own latent: A new approach to self-supervised learning. arXiv preprint arXiv:2006.07733.

Misra, I. and Maaten, L.V.D., 2020. Self-supervised learning of pretext-invariant representations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 6707-6717). arXiv:1912.01991.

Hržić, F., Štajduhar, I., Tschauner, S., Sorantin, E. and Lerga, J., 2019. Local-entropy based approach for X-ray image segmentation and fracture detection. *Entropy*, 21(4), p.338. https://doi.org/10.3390/e21040338.

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K. and Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition (pp. 248-255). IEEE. https://doi.org/10.1109/CVPR.2009.5206848.

Turrisi da Costa, V.G., Fini, E., Nabi, M., Sebe, N. and Ricci, E., 2021. Solo-learn: A Library of Self-supervised Methods for Visual Representation Learning. arXiv e-prints, pp.arXiv-2108.