Ensembling Downstream Image Classification Models Pretrained on Self-Supervised Models

Introduction

Computer vision has utilized many objective functions for pretext tasks to learn representations of unlabeled images and videos for self-supervised learning [1]. These objective functions include methods for utilizing metadata within the images and knowledge about operations performed on the images to learn models. Some literature includes combining several objective functions together to help strengthen the representations expressed in the learned models [2]. Once a model is produced from these self-supervised techniques it is then often tested by utilizing it to pretrain models for downstream image classification and object detection tasks.

Of the pretext task and downstream tasks models generated in literature to date involve the generation of a single pretext task model which is used to train a single downstream task model. Yet, ensembling methods have been a proven framework for enhancing predictive model performance [3]. Therefore we propose to 1) develop a series self-supervised models utilizing several pretext task learning objectives 2) create a series of downstream image classification models pretrained on our self-supervised models and 3) ensemble the downstream models to produce a final cohesive image classification prediction.

In addition to developing a novel pipeline driven by self-supervised learning, this method will also enable us to perform important analysis. Because we will be ensembling many different models together pretrained on different pretext tasks, we will be able to analyze how different pretext task learning objectives compare in enhancing the performance of the downstream image classification tasks. Time permitting, we can also analyze how different datasets impact the performance of downstream image classification tasks as well.

Prior Work

Most work done utilizing pretext tasks for self-supervised learning in computer vision involves comparing or combining learning objectives to measure how they impact performance of downstream supervised learning tasks. Learning objectives can include prediction of where one patch is in an image relative to another [4], to what degree an image was rotated [5], or placing a scrambled set of image patches back in order like a jigsaw puzzle [6]. Some literature has also combined several learning objectives together for the generation of a single model such as augmenting jigsaw puzzle pieces [7], but this still a relatively new area with potential for future directions [1]. Also, while learning objectives have been combined to produce a stronger pretext tasks model, as mentioned above, there is no literature that ensambles downstream models pretrained on pretext tasks model to compare how they impact image classification performance.

Methods

The main dataset to drive our experimentation will be the ImageNet dataset. This dataset contains the highest volume of training examples and variety in object classes and therefore will produce the most robust pretext representation models for our transfer learning experiments. We will then proceed to utilize additional datasets including Places205, CIFAR 10, STL-10, and Pascal VOC at time permits to expand the versatility of our overall transfer learning classification pipeline.

Our pretext tasks will be performed using a series of learning objectives on a neural network with a ResNet 50 backbone. Learning objectives will include methods such as rotation, relative patch location, and jigsaw techniques. Learning objectives may also be combined to create a single model. A subset of images from our targeted dataset(s) will be utilized without labels to perform the training of our models in this step.

Our downstream models will be neural network image classification models also using a ResNet 50 backbone to match our pretext architecture. All downstream models will first be provided a pretext task model for pretraining. All downstream models will receive supervised training on a distinct set of images from the ones used in the prior pretext tasks. In this way we hope to prevent any bias from images used across multiple types of training.

Lastly, all of the inputs from the downstream model will be entered into an ensembling algorithm to produce a final inference for image classification. This algorithm will initially implement a bagging ensembling approach where each downstream model will have an equal vote in the image classification prediction. We will then later change weights on voting to see how that impacts the model performance and how models pretrained with different pretext task learning objects and/or datasets impact the performance. Performance will be measured utilizing F1 score, accuracy, precision, and recall. These experiments will provide us the information required to analyze how much each learning objective helps our pipeline.

Timeline

Date	Task
Feb 14	Download ImageNet dataset and set it up for training and testing
Feb 18	Begin development of pretext tasks, downstream tasks, and ensembling algorithms
Mar 3	Complete initial development and begin model training and testing
Mar 17	Complete initial testing, perform write-up for first project presentation
Mar 24	First project presentation

Mar 24	Download new dataset and/or add new pretext tasks learning objectives to enhance the pipeline
Apr 14	Complete testing and perform final write-up of project
Apr 21	Second project presentation

References

1. L Jing, Y Tian. “Self-supervised Visual Feature Learning with Deep Neural Networks: A Survey”. arXiv:1902.06162v1 [cs.CV], Feb 2019
2. C Doersch, A Zisserman. “Multi-task Self-Supervised Visual Learning”. In ICCV, 2017.
3. Wikipedia contributors. (2020, January 20). Netflix Prize. In Wikipedia, The Free Encyclopedia. Retrieved 18:24, February 11, 2020, from https://en.wikipedia.org/w/index.php?title=Netflix_Prize&oldid=936701180
4. C. Doersch, A. Gupta, and A. A. Efros, “Unsupervised visual rep-resentation learning by context prediction,” inICCV , pp. 1422– 1430, 2015
5. S. Gidaris, P. Singh, and N. Komodakis, “Unsupervised represen-tation learning by predicting image rotations,” inICLR , 2018.
6. M. Noroozi and P. Favaro, “Unsupervised learning of visual representations by solving jigsaw puzzles,” inECCV , 2016.
7. D. Kim, D. Cho, D. Yoo, and I. S. Kweon, “Learning image representations by completing damaged jigsaw puzzles,” arXiv preprint arXiv:1802.01880 , 2018.