CS 231N Project

Prototypical Pre-Training for Visual Representation Learning

Rohan Sikand
Department of Computer Science, Stanford University

Background & Introduction

- Visual representation learning is an important subtopic in deep learning for computer vision which revolves around pre-training an encoder to learn a set of representations that are useful for downstream tasks.
- Recent approaches such as SimCLR and MoCo.
- We propose a novel method called “prototypical representation learning” (PRL) which harnesses the underlying concept of prototypical networks (Snell 2017) , but adapt it for pre-training for representation learning.
- We test our method on STL-10 image classification via downstream fine-tuning.

Problem Setup

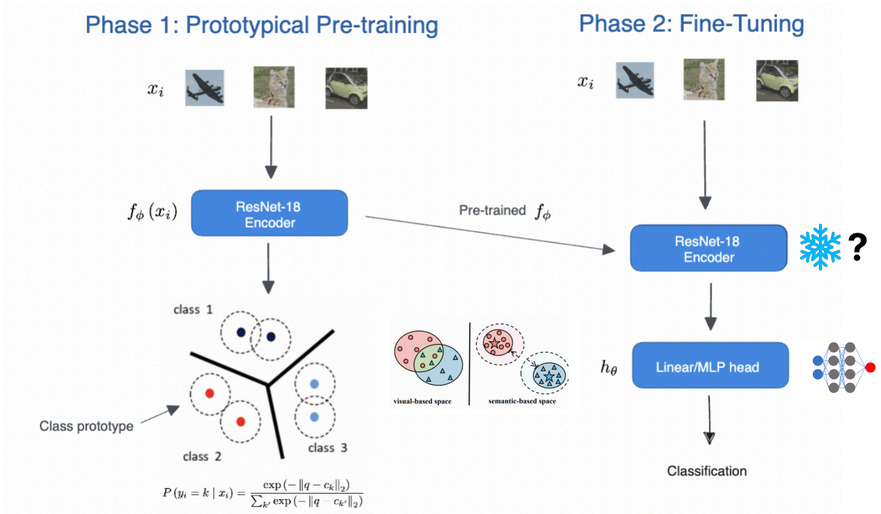
- Representation learning: pre-train feature extractor and attach shallow network during downstream tasks.
- To test how well our approach learns features, we perform image classification on the STL-10 dataset.
- Our representation learning approach is *supervised* and self-contained: we use the downstream training set for both pre-training and fine-tuning.
- The goal is to improve performance on the downstream task by harnessing the representations learning during pre-training.

Experiments

	Description
MLP supervised baseline	- train randomly initialized MLP neural network
ResNet-18 supervised baseline	- train randomly initialized ResNet-18 convolutional neural network
SimCLR representation learning baseline	- perform SimCLR pre-training and fine-tuning which will serve as a representation learning baseline
Query Prototypical Pre-training	- PRL variant in which we divide the input batch into 1/4 "query" and 3/4 "support" - prototypes are calculated with the support set, but the loss is applied with the query set. - We hypothesize that this encourages generalization.
Frozen vs. unfrozen f_ϕ parameters	- after pre-training, should we keep the ResNet-18 encoder's parameters ϕ frozen or unfrozen during fine-tuning?
MLP vs. linear head h	- for the fine-tuning head that maps latent vectors to class probabilities, we experiment with multilayer perceptron network with *nonlinearities* (ReLU) vs. a single linear layer.

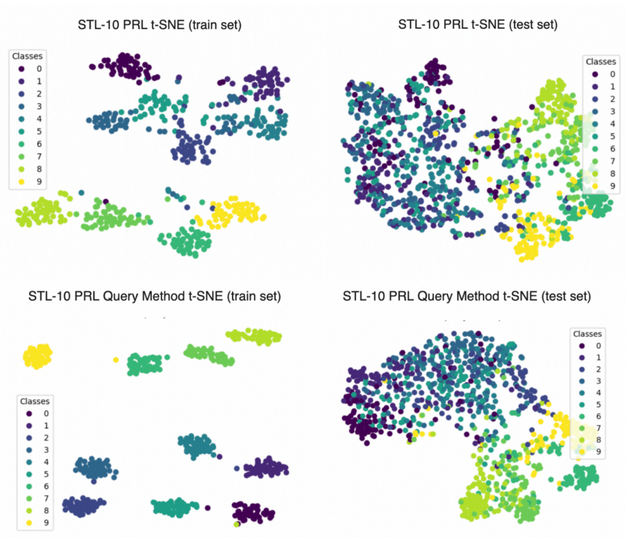
Methods

Prototypical Representation Learning:



Analysis

- To qualitatively gauge the performance of PRL, we plotted the latent embeddings from 100 train and test images using t-SNE.
- Clearly, the method is able to produce well defined clusters for the train set (i.e., the same set that was used for pre-training), but not for the test set.
- The query variant of PRL forms less ambiguous clusters



Results

Method	Dataset	Test Accuracy	Train Accuracy
Supervised Baseline MLP	STL-10	41.4%	59.1%
Supervised Baseline ResNet-18	STL-10	62.5%	98.4%
SimCLR Baseline	STL-10	64.6%	96.5%
PRL Standard Frozen Linear	STL-10	60.7%	99.7%
PRL Standard Frozen MLP	STL-10	60.6%	99.8%
PRL Standard Unfrozen Linear	STL-10	63.8%	99.9%
PRL Standard Unfrozen MLP	STL-10	60.2%	98.9%
PRL Query Unfrozen Linear	STL-10	63.6%	99.7%
PRL Query Frozen Linear	STL-10	64.2%	99.9%
PRL Query Unfrozen MLP	STL-10	63.8%	99.9%
PRL Query Frozen MLP	STL-10	63.1%	99.6%

Discussion and Conclusion

- It is clear that prototypical pre-training produces useful representations as visually depicted in the t-SNE plots.
- Quantitatively, we saw performance (+2%) increase in the best performing PRL approach compared to the ResNet-18 CNN baseline. Our approach is competitive with SimCLR (-.4%).
- Experimentally, the “query” variant of PRL outperforms the standard formulation. We hypothesize that this is because the query variant encourages generalization.
- In this project, we proposed a novel method for visual representation learning which is performant for the downstream task of image classification.
- Future work would include refining the method to get the test set t-SNE clusters to be more well defined. In addition, we hope to test our method for the tasks of domain generalization, few-shot learning, and transfer learning.

References

[1] 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. 1, 2, 4

[2] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33:22243–22255, 2020. 2

[3] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 215–223. JMLR Workshop and Conference Proceedings, 2011. 2, 5

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5

[5] 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. 1, 2

[6] Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. A survey on contrastive self-supervised learning. *Technologies*, 9(1):2, 2020. 1

[8] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017. 1, 2, 4

[9] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 7

[10] Chaofan Zheng, Xinyu Lyu, Lianli Gao, Bo Dai, and Jingkuan Song. Prototype-based embedding network for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22783–22792, 2023. 6