Distilling Visual Priors from Self-Supervised Learning

Visual Inductive Priors Workshop @ ECCV 2020

Bingchen Zhao[1,2] and Xin Wen[1]
1.Tongji University, Shanghai, China
2.Megvii Research Nanjing

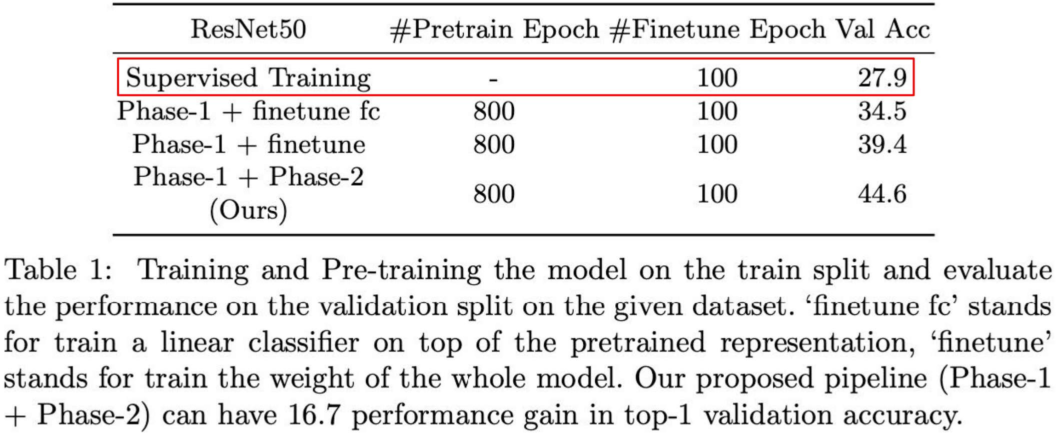
The problem we want to solve

- ➢ No pretrain, all model need to train from scrach
- ➢ Limited training data, each class only has 50 images.
- ➢ No external data or checkpoint can be used.
- ➢ Inject visual priors into the CNNs to save data and improve the performance.

Challenges with limited data

- ➢ Overfitting to the label information, small dataset can be simply memorized by CNNs.
- ➢ Insufficient data leads to a poor representation

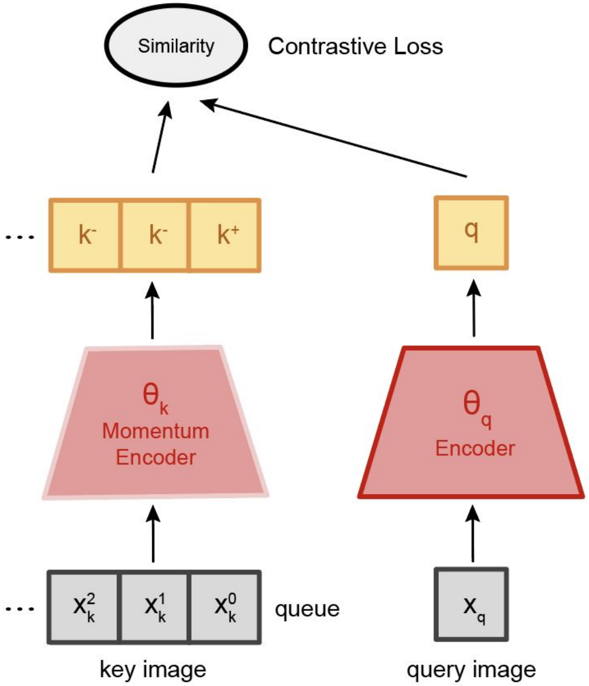
Supervised training only have a 27.9 validation acc!



ResNet50	#Pretrain Epoch	#Finetune Epoch	Val Acc
Supervised Training	-	100	27.9
Phase-1 + finetune fc	800	100	34.5
Phase-1 + finetune	800	100	39.4
Phase-1 + Phase-2 (Ours)	800	100	44.6

Table 1: Training and Pre-training the model on the train split and evaluate the performance on the validation split on the given dataset. 'finetune fc' stands for train a linear classifier on top of the pretrained representation, 'finetune' stands for train the weight of the whole model. Our proposed pipeline (Phase-1 + Phase-2) can have 16.7 performance gain in top-1 validation accuracy.

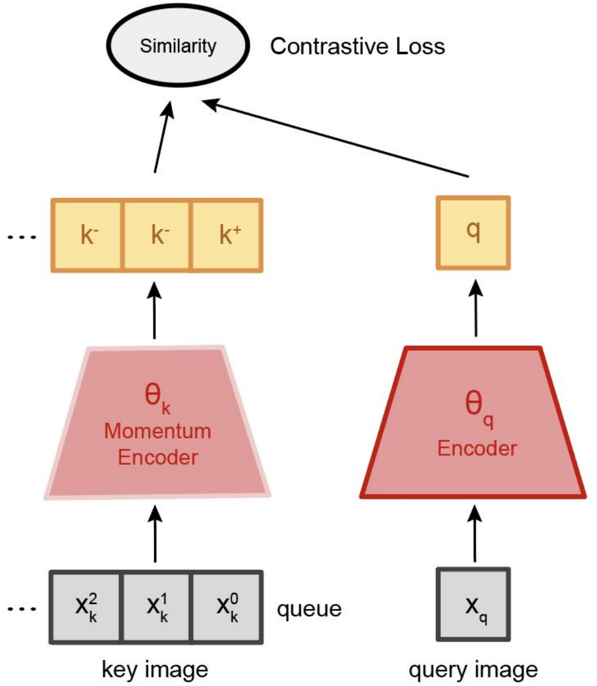
Our solution: self-supervised learning



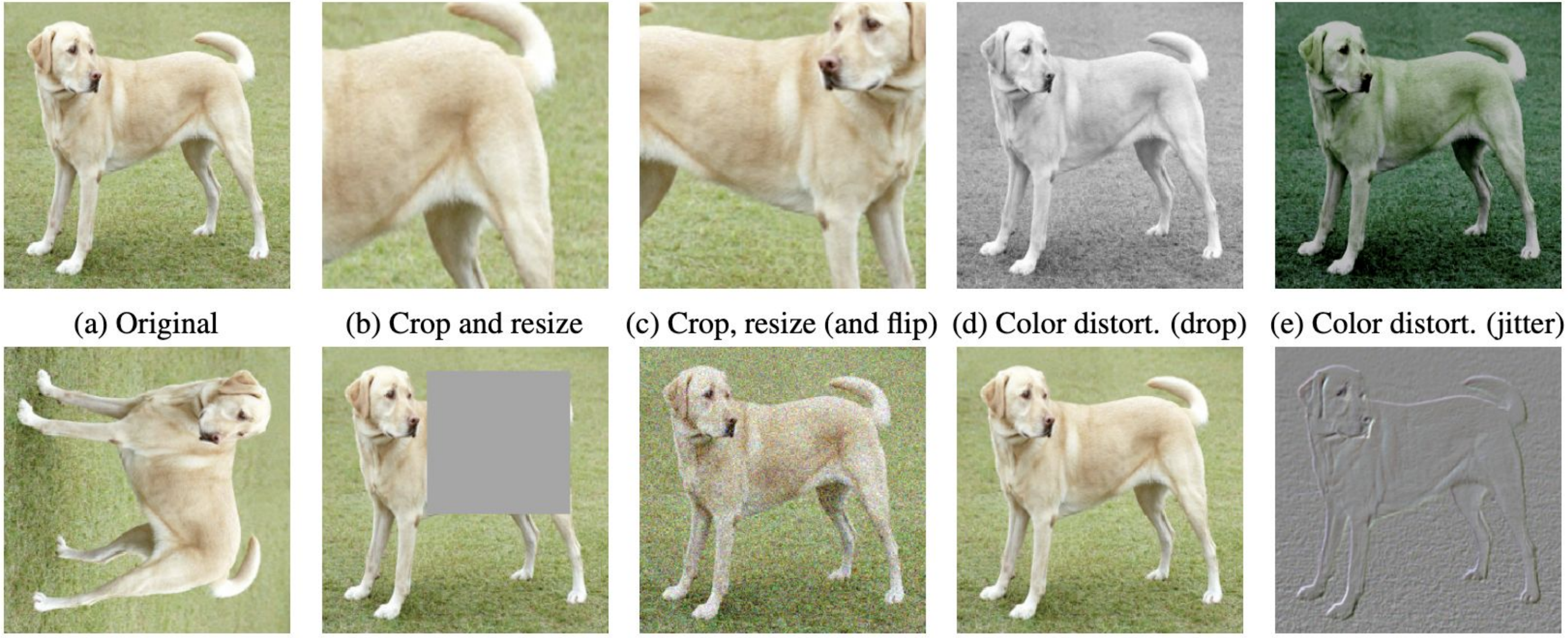
- ➢ Self-Supervised learning is not biased to human labels.
- ➢ Self-Supervised learning can learns a more diverse and more expressive representation

Self-Supervision is not biased to label information

- Taking each image and its augmented version as a class.
- No human label information is required.
- Preventing the model from overfitting to the human label on the small dataset.



Self-Supervision can learn a diverse representation



(a) Original

(b) Crop and resize

(c) Crop, resize (and flip)

(d) Color distort. (drop)

(e) Color distort. (jitter)

(f) Rotate {90°, 180°, 270°}

(g) Cutout

(h) Gaussian noise

(i) Gaussian blur

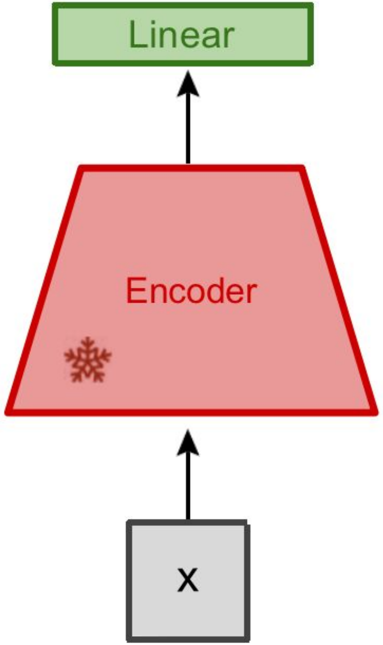
(j) Sobel filtering

Chen, Ting et al. “A Simple Framework for Contrastive Learning of Visual Representations.”

Classification: Linear readout

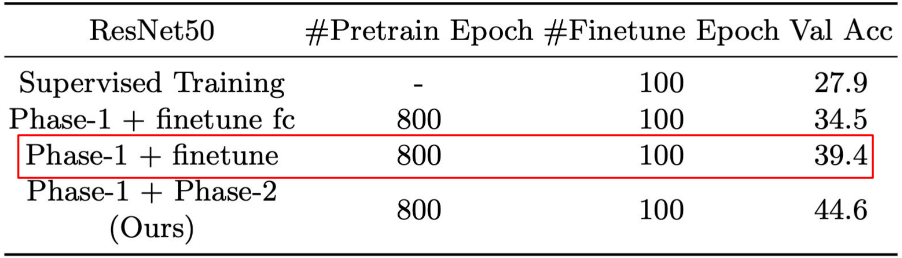
- Common practice for evaluating self-supervised representations is **linear readout** which freezes the encoder and train a MLP to get the result.
- We show that for performance proposes on a small dataset, **linear readout** could outperform supervised train but is not the best.

ResNet50	#Pretrain Epoch	#Finetune Epoch	Val Acc
Supervised Training	-	100	27.9
Phase-1 + finetune fc	800	100	34.5
Phase-1 + finetune	800	100	39.4
Phase-1 + Phase-2 (Ours)	800	100	44.6

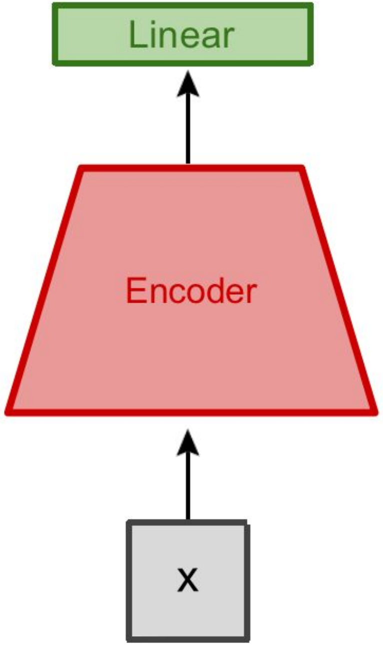


Classification: Finetune

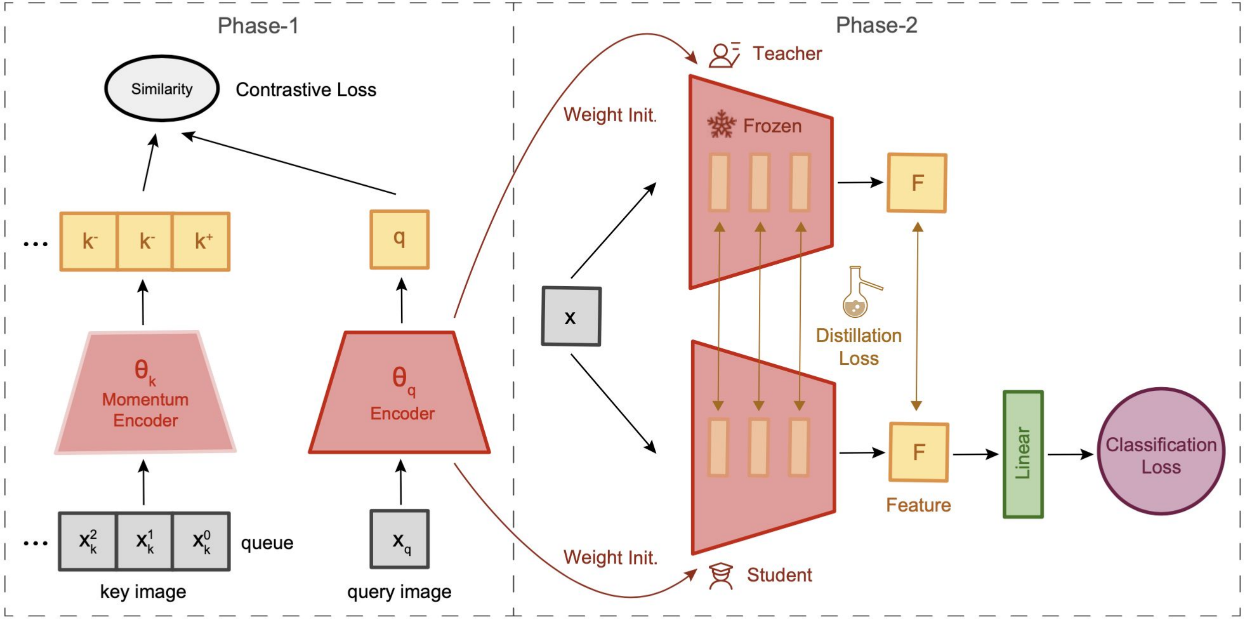
- Use the self-supervise trained encoder as a pretrained checkpoint and then perform finetune leads to a performance gain.
- But the finetune process still face the risk of overfitting.



ResNet50	#Pretrain Epoch	#Finetune Epoch	Val Acc
Supervised Training	-	100	27.9
Phase-1 + finetune fc	800	100	34.5
Phase-1 + finetune	800	100	39.4
Phase-1 + Phase-2 (Ours)	800	100	44.6



Classification: Self-distillation

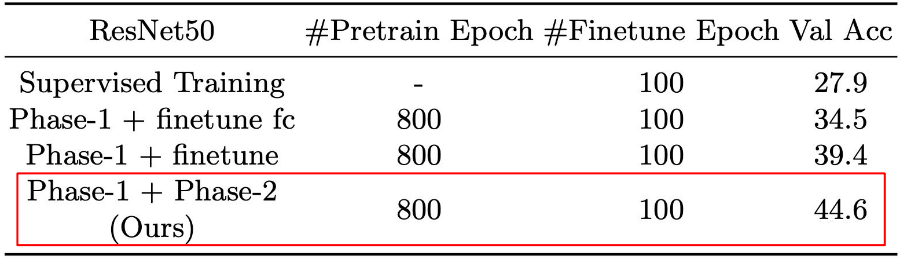


Our proposed two phase method,

- The first phase is to use self-supervision to learn a good representation
- The second phase uses self-distillation to prevent the model from overfitting during the finetuning process.

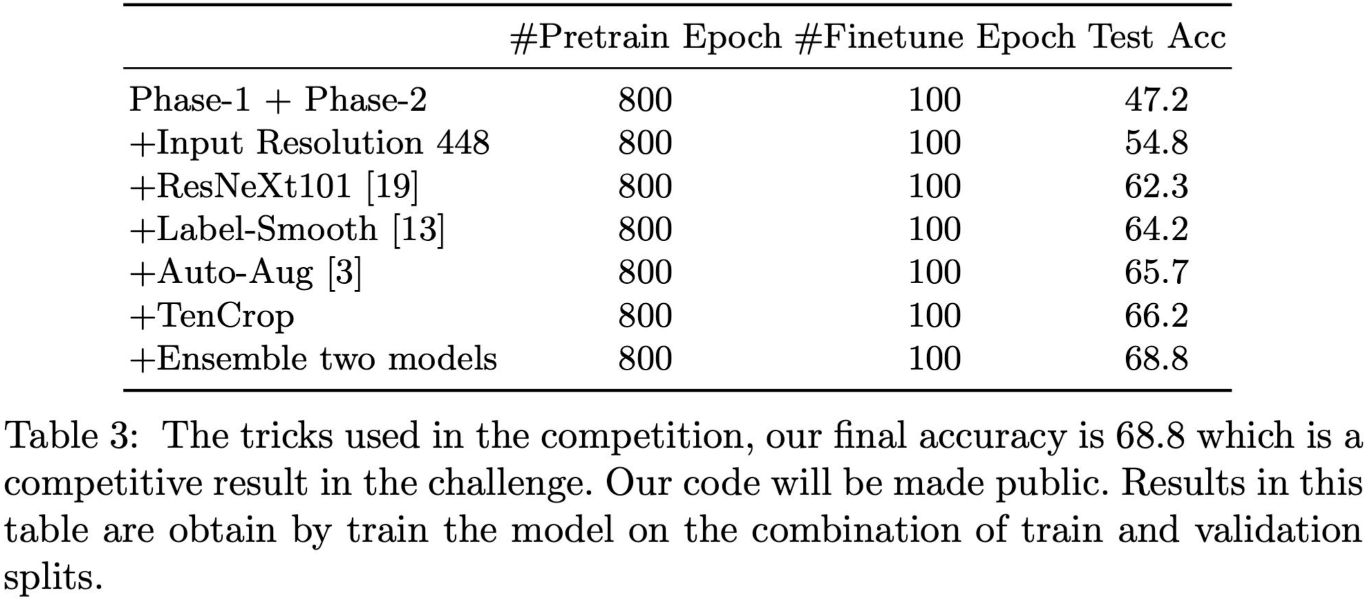
Classification: Self-distillation

Our Method outperforms the baseline by 16.7 in validation accuracy!!



ResNet50	#Pretrain Epoch	#Finetune Epoch	Val Acc
Supervised Training	-	100	27.9
Phase-1 + finetune fc	800	100	34.5
Phase-1 + finetune	800	100	39.4
Phase-1 + Phase-2 (Ours)	800	100	44.6

Competition: Other tricks



	#Pretrain Epoch	#Finetune Epoch	Test Acc
Phase-1 + Phase-2	800	100	47.2
+Input Resolution 448	800	100	54.8
+ResNeXt101 [19]	800	100	62.3
+Label-Smooth [13]	800	100	64.2
+Auto-Aug [3]	800	100	65.7
+TenCrop	800	100	66.2
+Ensemble two models	800	100	68.8

Table 3: The tricks used in the competition, our final accuracy is 68.8 which is a competitive result in the challenge. Our code will be made public. Results in this table are obtain by train the model on the combination of train and validation splits.

We ranked 2nd place in the Image Classification Competition by ensemble only **two** models.

Thanks

https://github.com/DTennant/distill_visual_priors

Our code is open sourced on GitHub.

Bingchen Zhao and Xin Wen are both undergraduate students at Tongji University.
Bingchen Zhao is currently an intern at Megvii Research Nanjing.
Xin Wen is currently an intern at ByteDance AI Lab.