Learning CNNs from Web Data

Rodrigo Santa Cruz and Stephen Gould
Australian Centre for Robotic Vision, Australian National University, Canberra, Australia

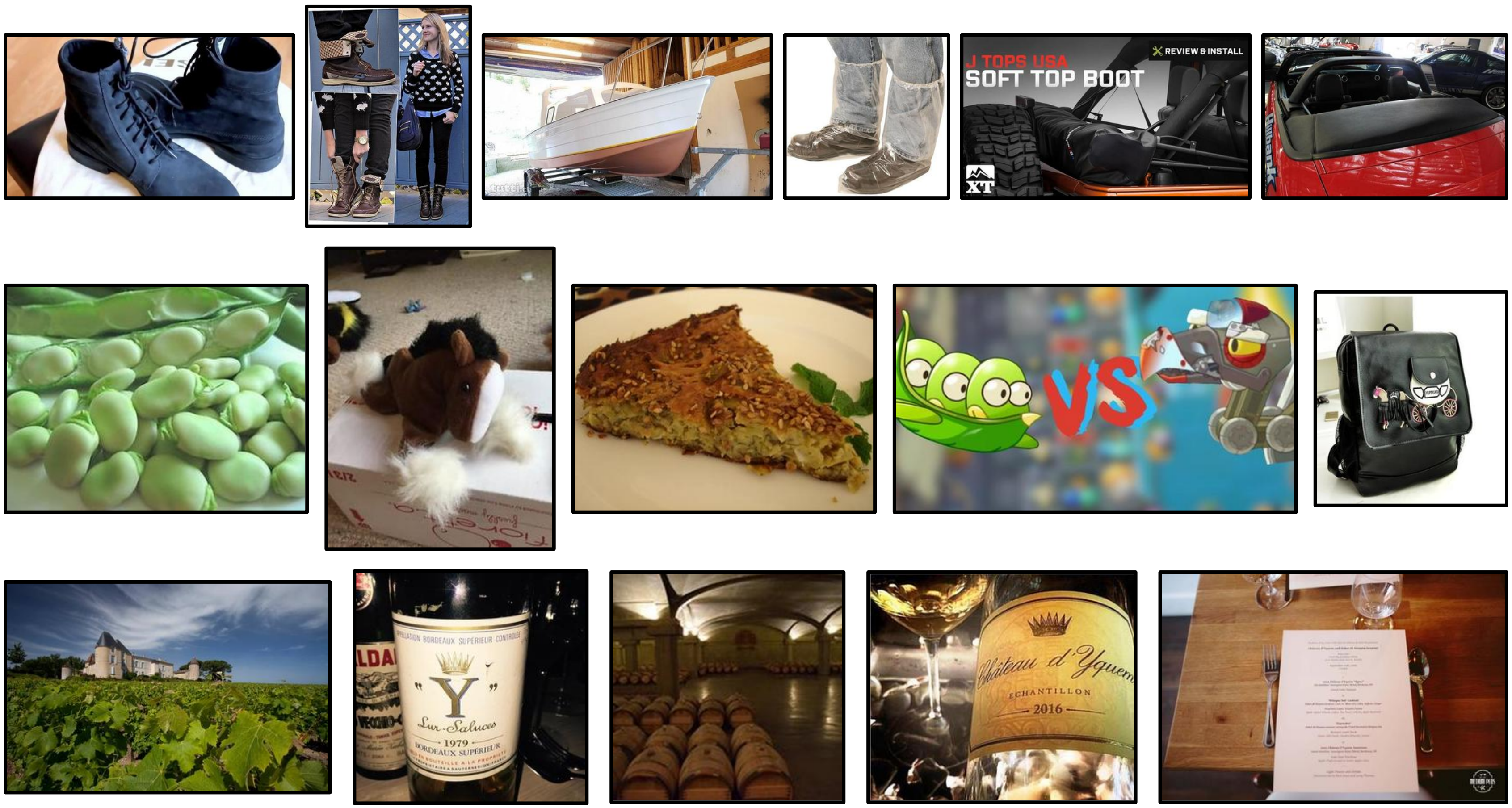
Outline

- WebVision Challenge
  - The large scale WebVision 2.0 dataset
  - Noisy labels
  - Imbalanced class distribution
- Approach
  - Intuition
  - Self-supervised pre-training
  - Base classifier
  - Weighted sampling
  - Dense prediction
- Results

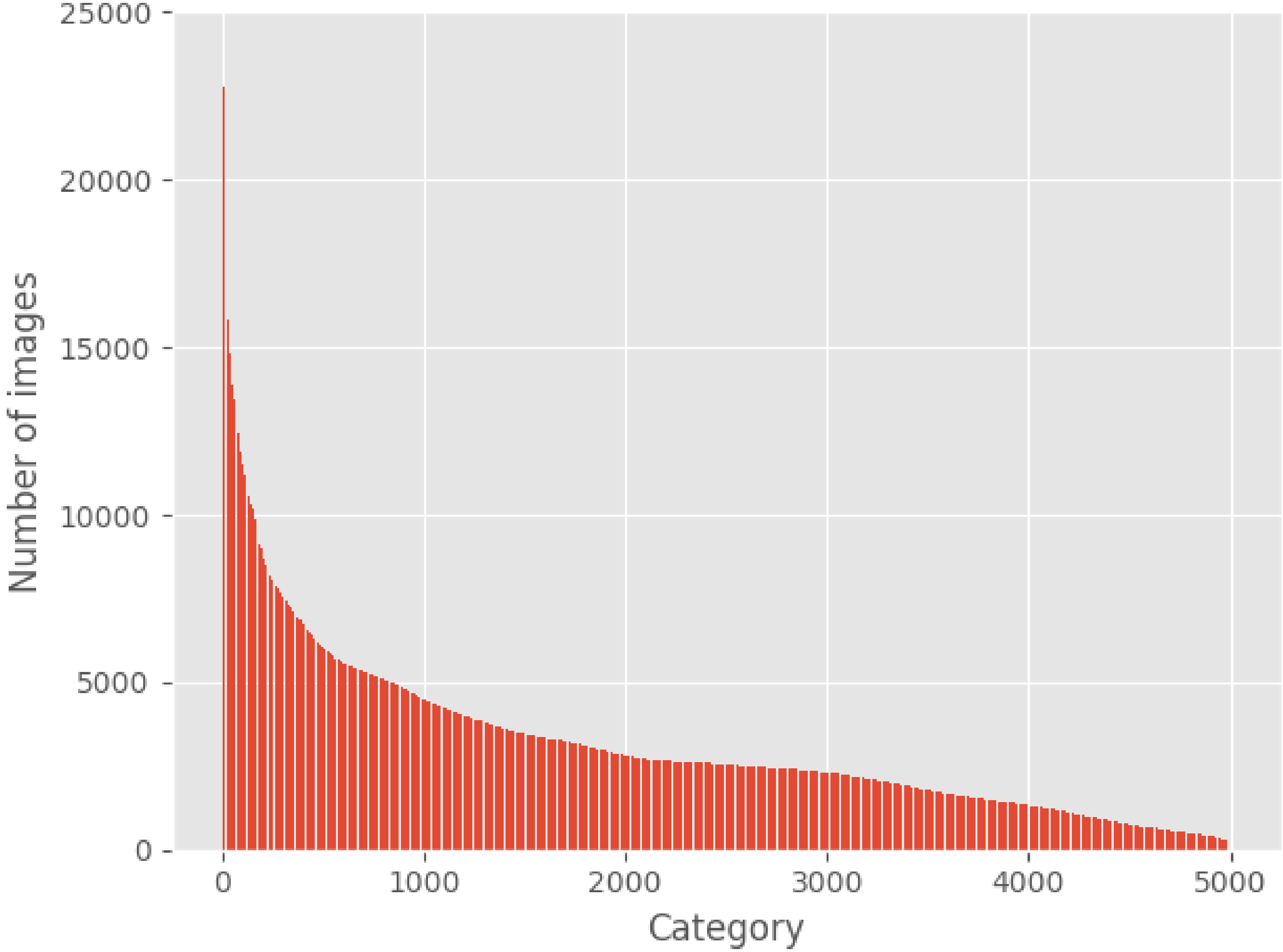
Large Scale Problem

- WebVision 2.0 Dataset:
  - Noyse dataset generated from more than 12000 queries to **Google image search** and **Flickr social media**.
  - It contains **5,000 visual concepts** associated to synsets in ImageNet.
  - It has more than **16 millions training images**, 250 thousands validation images and 250 thousands test images (no public labels).
  - It also provides additional information such as title and description for Google images and title, description and tags for flickr images.

Noisy Labels

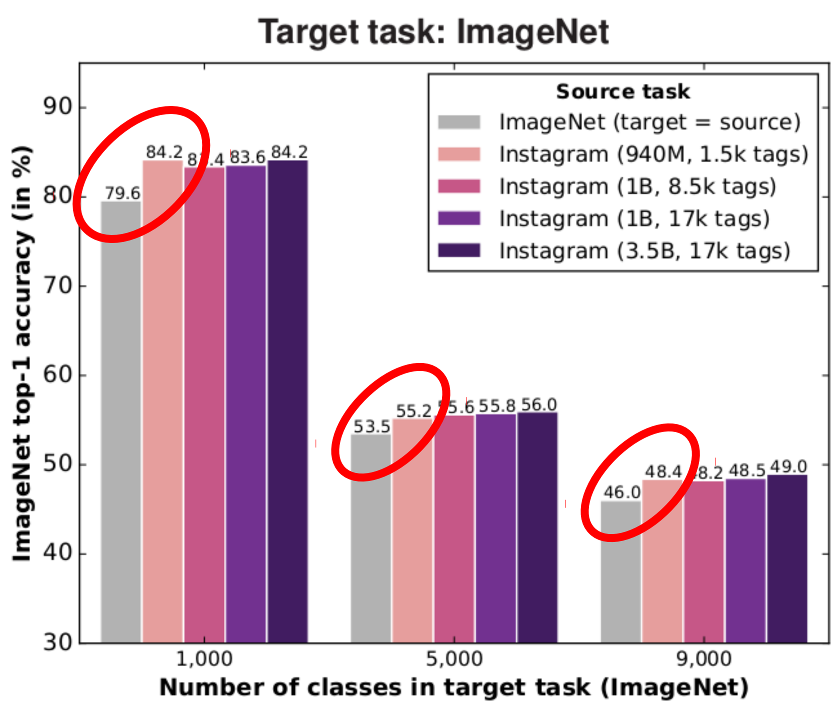


Imbalanced Class Distribution



Approach

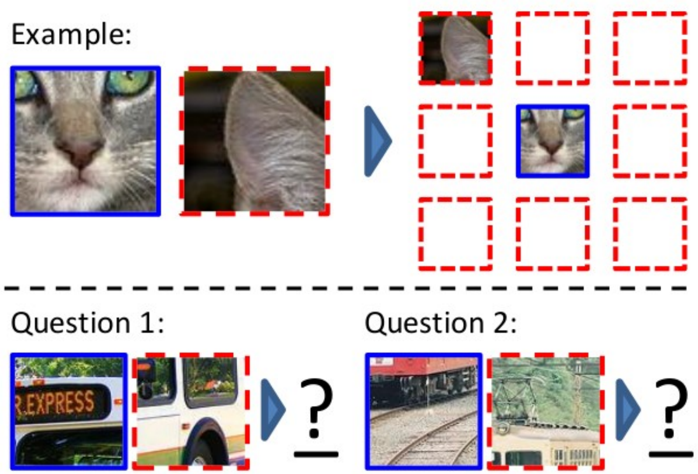
- Deep Learning Formula: **Pretrain** on large dataset and then **finetune** on a smaller task-specific dataset. Ex: object detection.
- Initialization is **crucial** to non-convex optimization problems such as learning deep models.
- Does a good start point provide **robustness** to noisy labels ?



[Exploring the Limits of Weakly Supervised Pretraining. Mahajan et al 2018]

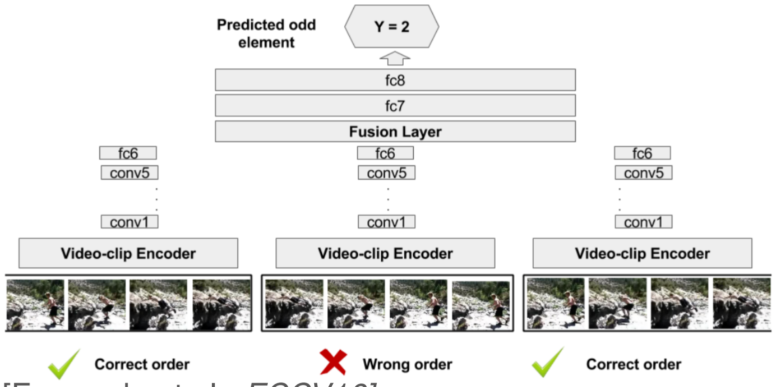
Self-supervised learning

- The main idea is to exploit supervisory signals, intrinsically in the data, to guide the learning process.
- In practice, we define a supervised proxy task, where labels are obtained with almost zero cost, to train the model before finetune for the target task.



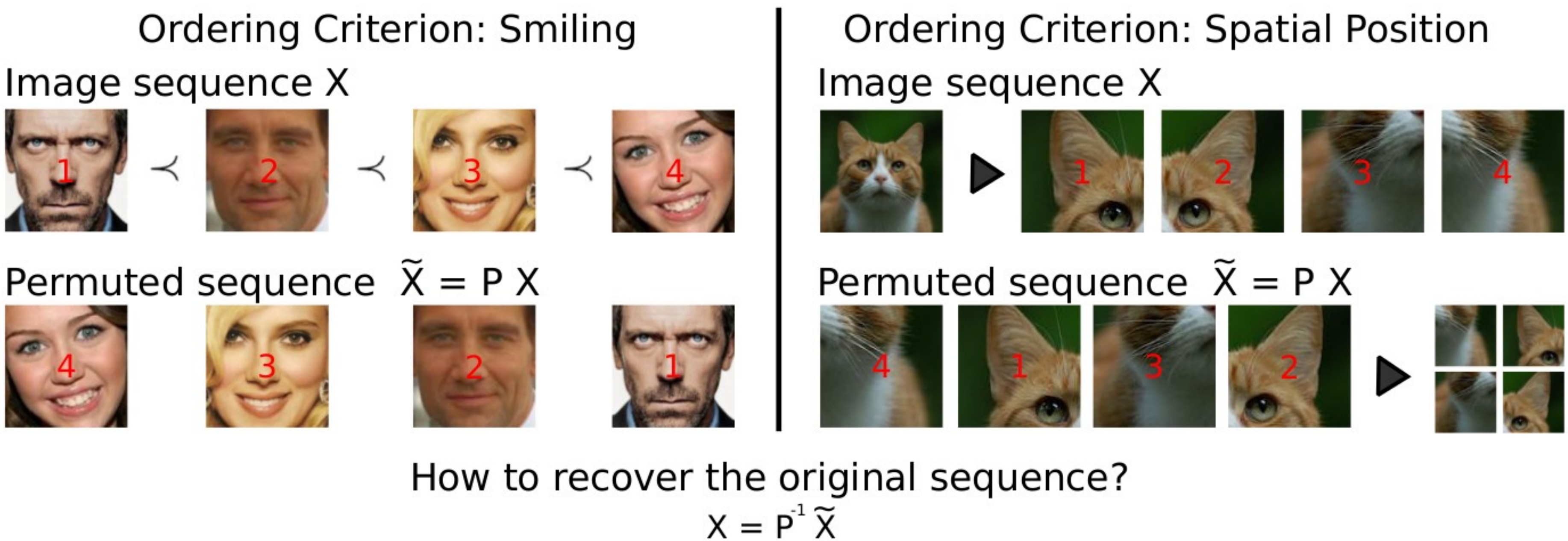
[Doersch et al., *ICCV 2015]*

[Zhang et al., *ECCV16]*



[Fernando et al., *ECCV16]*

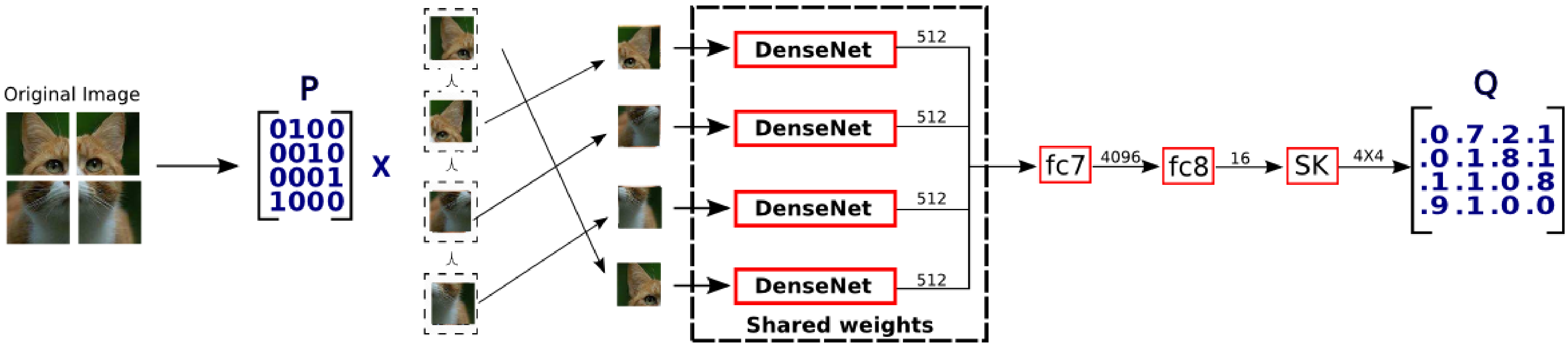
Visual Permutation Learning (VPL)



We hypothesize that the model trained to solve such task is able to capture high-level semantic concepts, structure and shared patterns in visual data.

[DeepPermNet: Visual Permutation Learning. Rodrigo Santa Cruz, Basura Fernando, Anoop Cherian, Stephen Gould. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.]

DeepPermNet



Remarks:

- We relax the inference over permutation matrices to inference over doubly-stochastic matrices (DSMs).
- We develop a neural network layer (Sinkhorn Layer) that approximates DSMs from CNN's outputs.
- Incorporating the DSM structure in our predictors can avoid the optimizer from searching over imposible solutions.

Sinkhorn Layer

- Sinkhorn Normalization*: Any non-negative square matrix can be converted to a DSM by repeatedly rescaling its rows and columns.
- Function:

$$R_{i,j}(Q) = \frac{Q_{i,j}}{\sum_{k=1}^{l} Q_{i,k}}; \quad C_{i,j}(Q) = \frac{Q_{i,j}}{\sum_{k=1}^{l} Q_{k,j}}$$

$$S^{n}(Q) = \begin{cases} Q, & \text{if } n = 0 \\ C\left(R\left(S^{n-1}(Q)\right)\right), & \text{otherwise.} \end{cases}$$

- Gradient (Row normalization):

$$\frac{\partial \Delta}{\partial Q_{p,q}} = \sum_{j=1}^{l} \frac{\partial \Delta}{\partial R_{p,j}} \left[\frac{[\![j = q]\!]}{\sum_{k=1}^{l} Q_{p,k}} - \frac{Q_{p,j}}{\left(\sum_{k=1}^{l} Q_{p,k}\right)^{2}} \right]$$

*[Sinkhorn and Knopp 1967][Knight 2008][Adams and Zamel 2011][Mena et al. 2018]

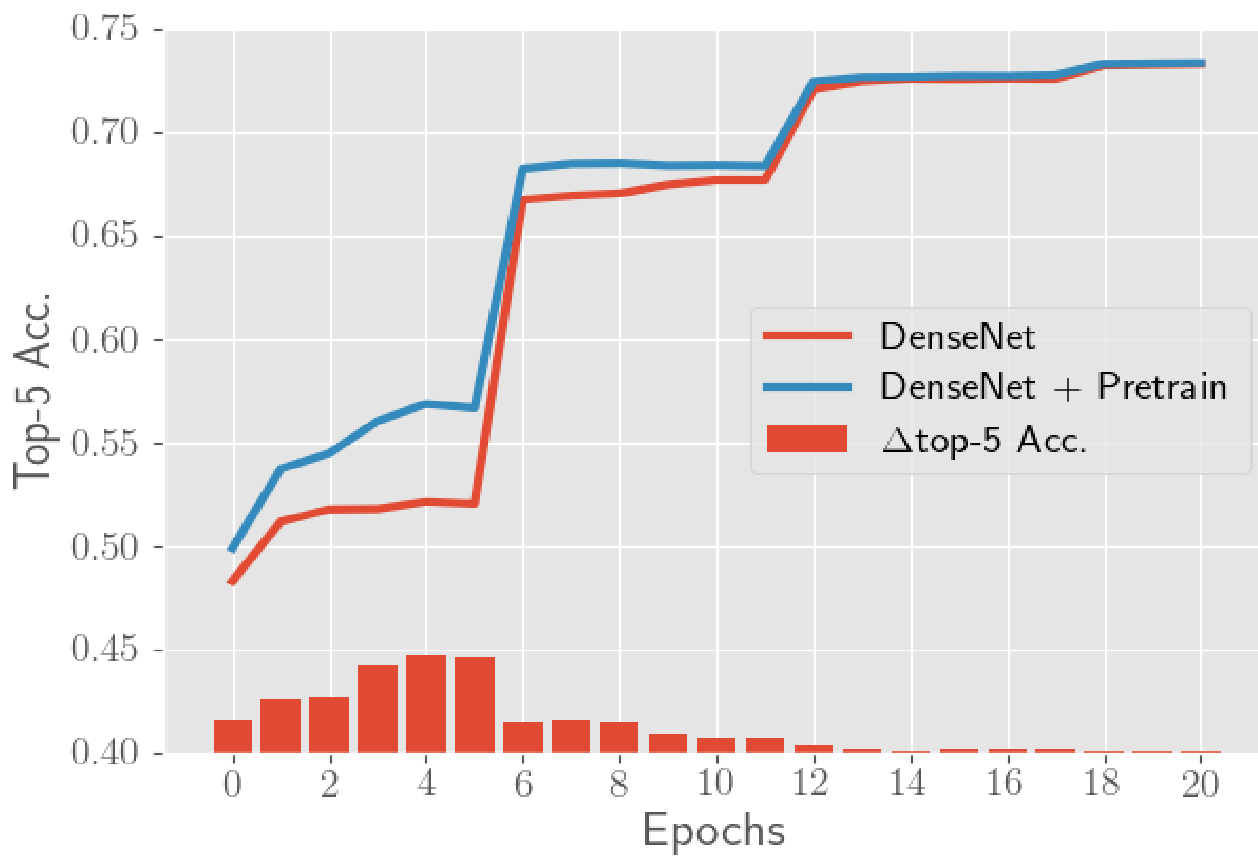
VPL Results

- It provides significant boost in performance compared to random initialization.
- This framework also presents good results for learning-to-rank problems such as image ranking.

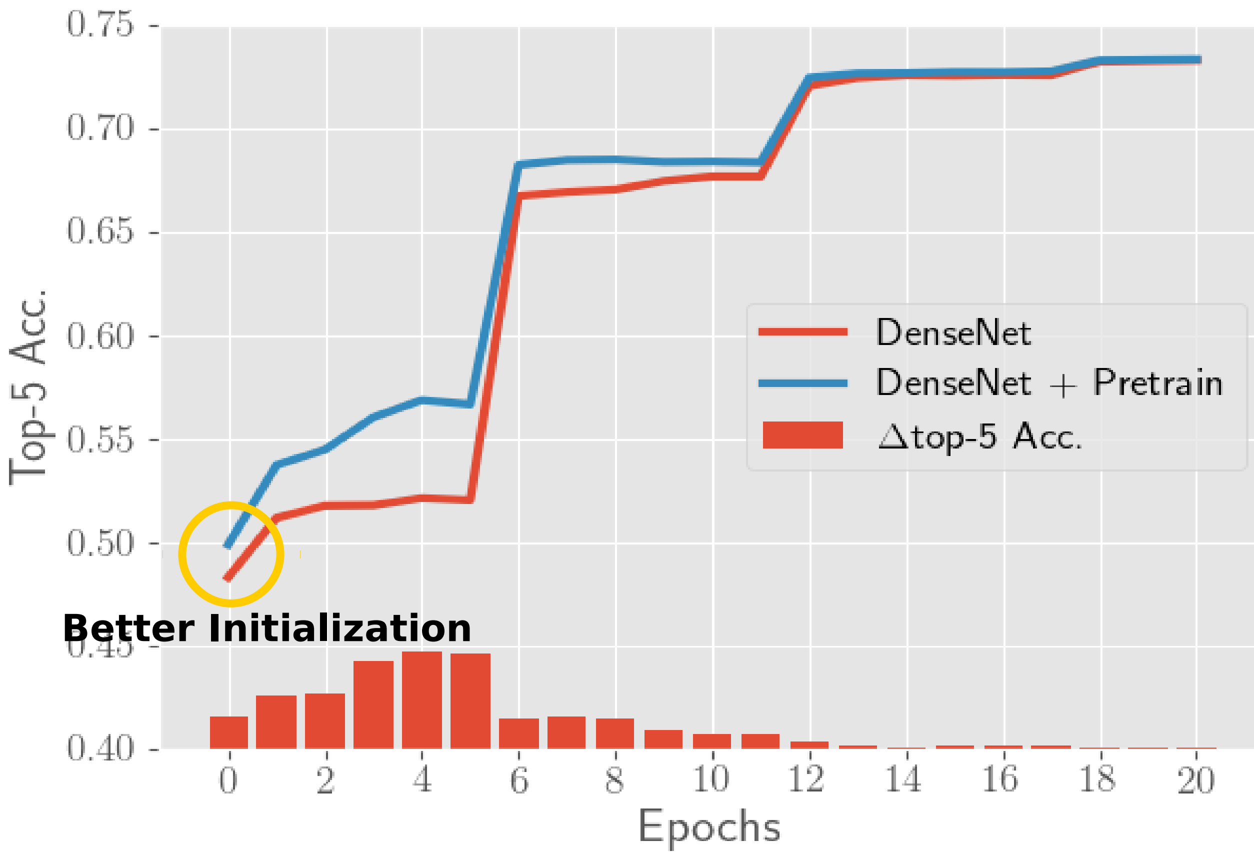
Method	Classification (mAP%)	FRCN Detection (mAP%)	FCN Segmentation (%mIU)
ImageNet	78.2	56.8	48.0
Random Gaussian	53.3	43.4	19.8
Context Prediction	55.3	46.6	-
Temporal coherence	58.4	44.0	-
In-painting	56.5	44.5	29.7
Colorization	65.6	47.9	35.6
Jigsaw Puzzle	68.6	**51.8**	36.1
DeepPermNet	**69.4**	49.5	**37.9**

16% 6% 18%

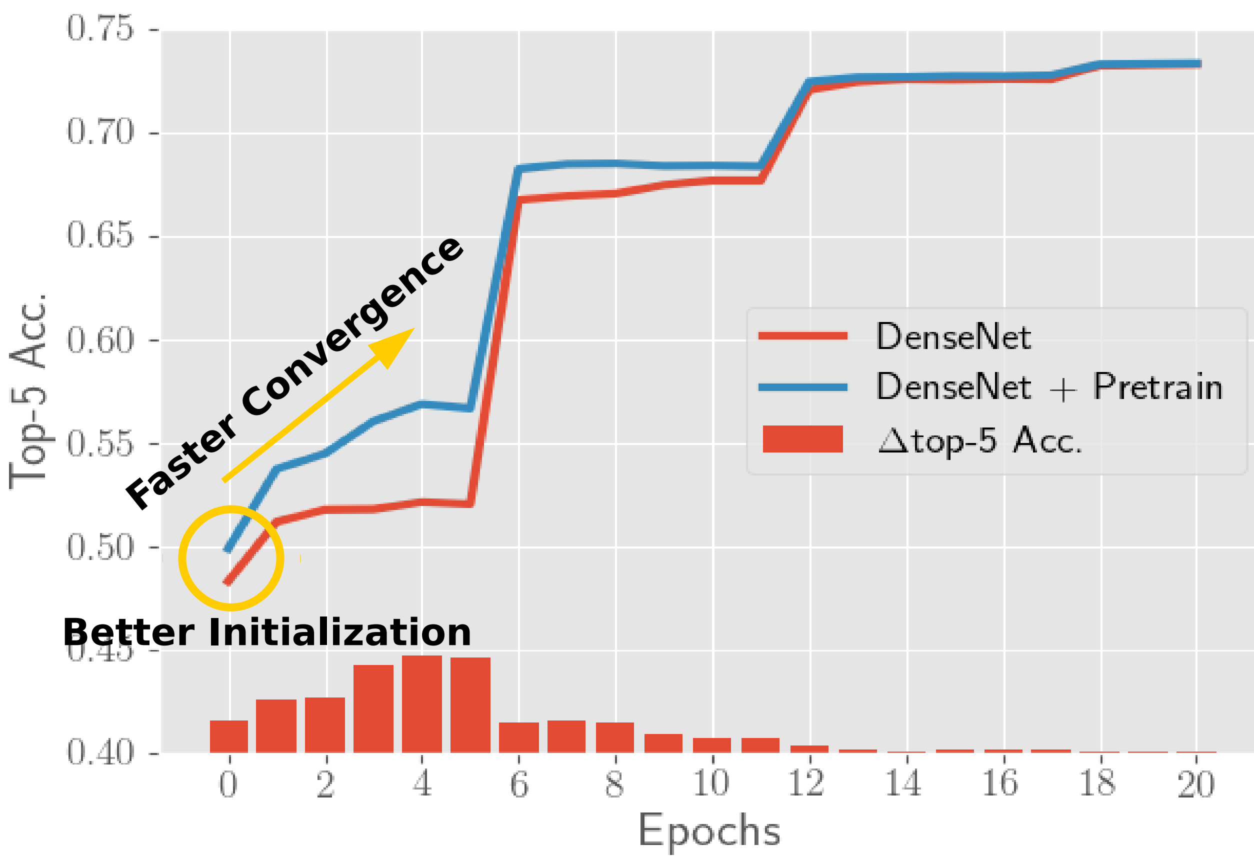
VPL on the WebVision



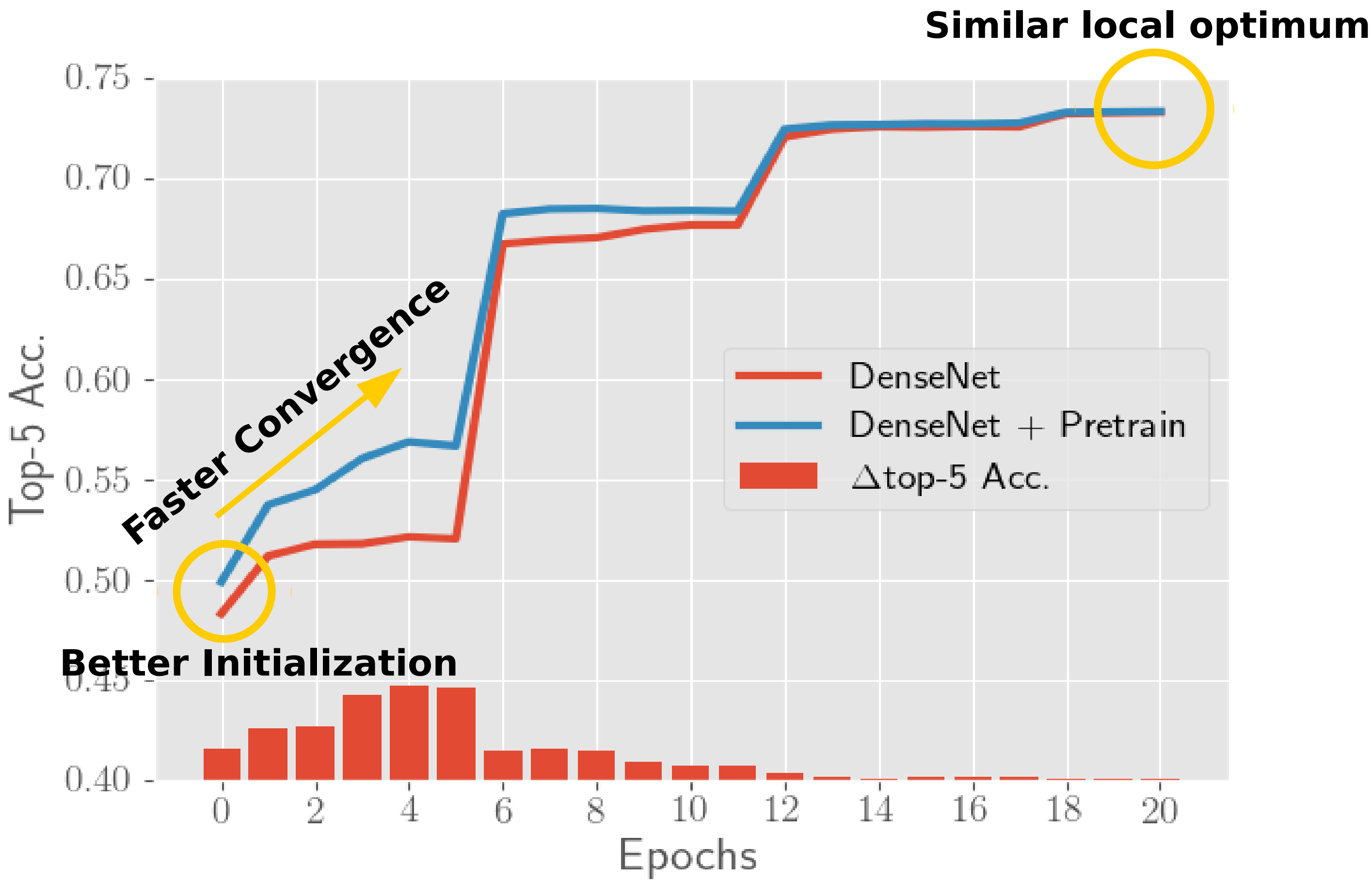
VPL on the WebVision



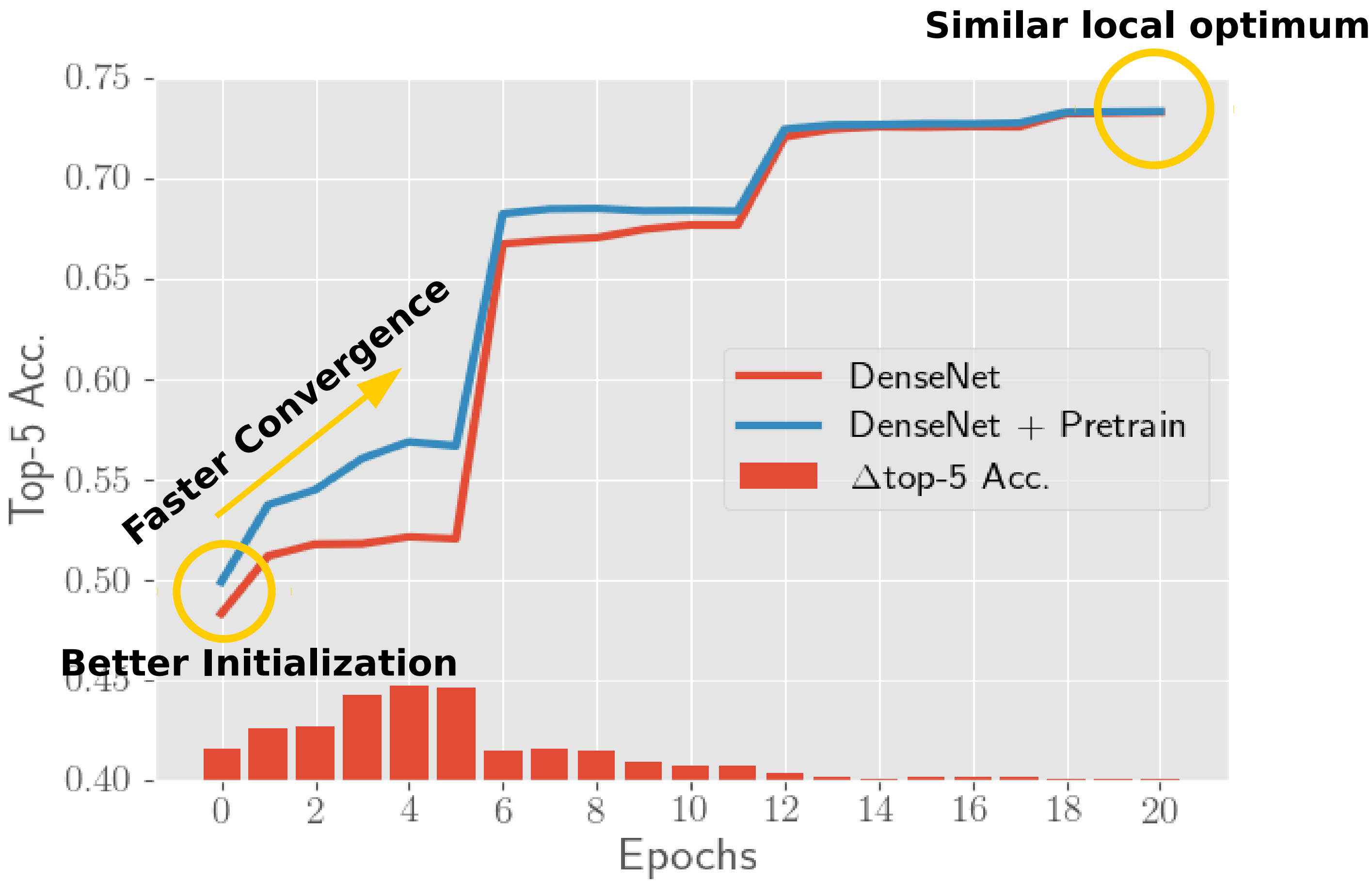
VPL on the WebVision



VPL on the WebVision



VPL on the WebVision



Similar behavior is observed under different learning rate schedule.

VPL on the WebVision

- We only achieved **marginal improvements (< 1%)** using visual permutation learning as pretraining procedure for the WebVision task.
- We can see significant improvements at the **beginning of the training** which is **diluted** as the training progresses, reaching **similar performance after convergence**.
- It is still useful when you we need to train a model for few epochs.

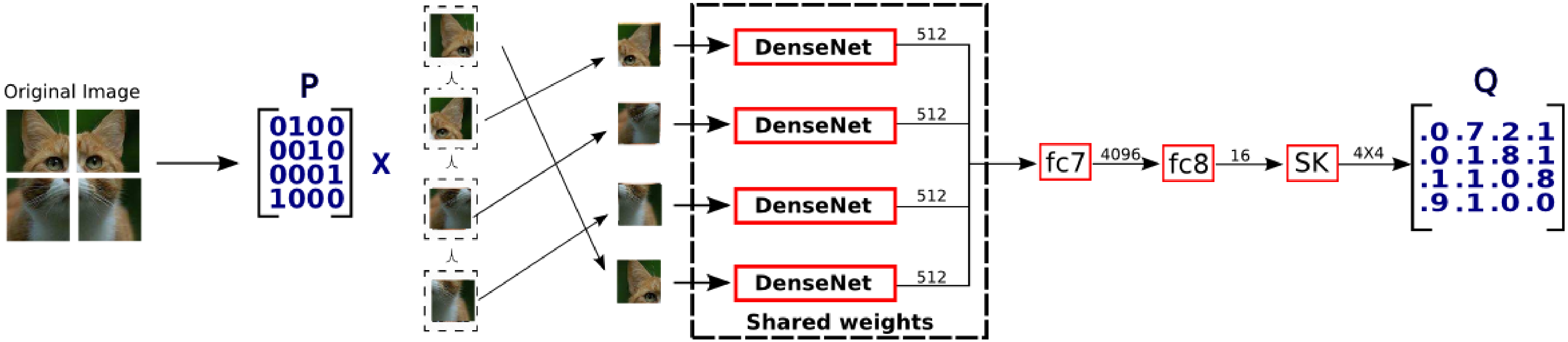
VPL Regularizer

VPL Regularizer

- Pretrain in the visual permutation learning:

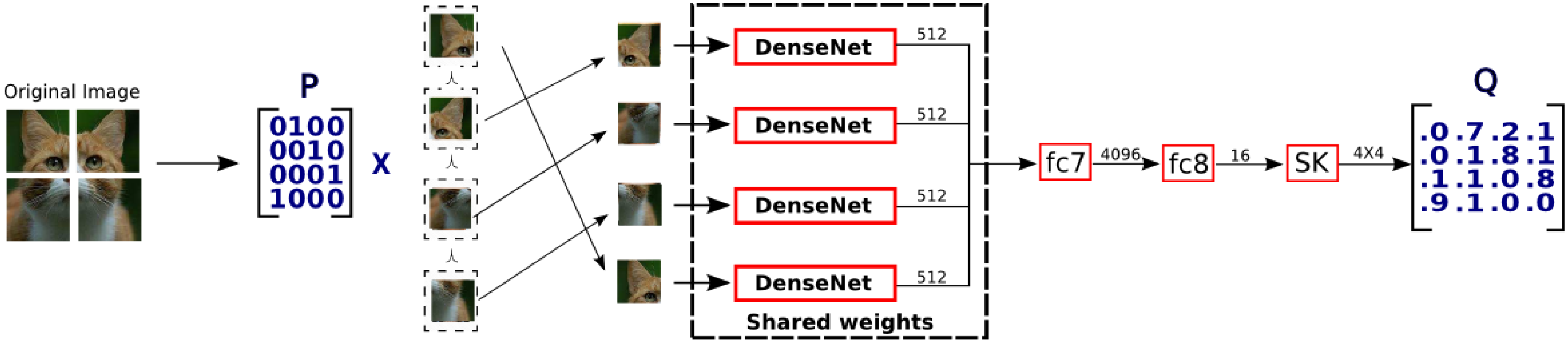
VPL Regularizer

- Pretrain in the visual permutation learning:



VPL Regularizer

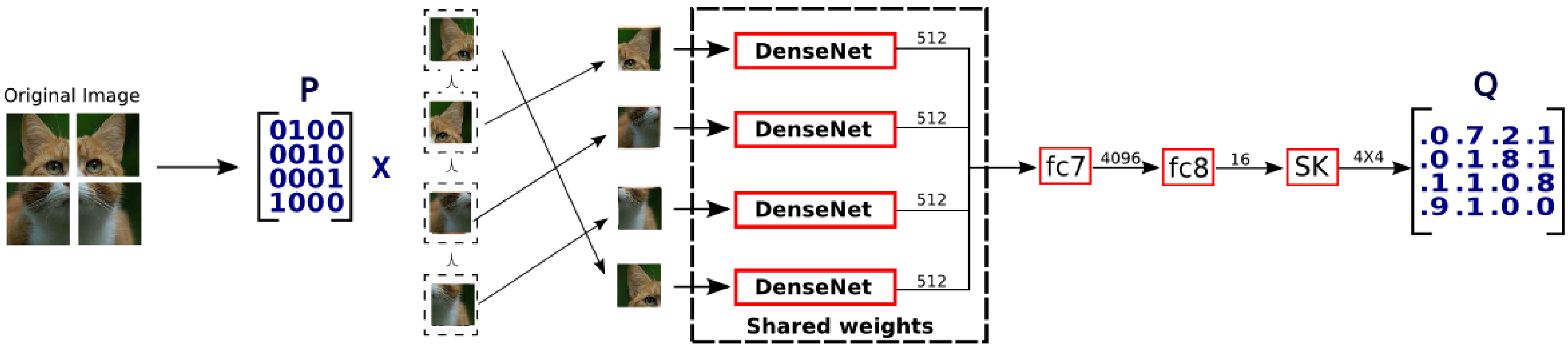
- Pretrain in the visual permutation learning:



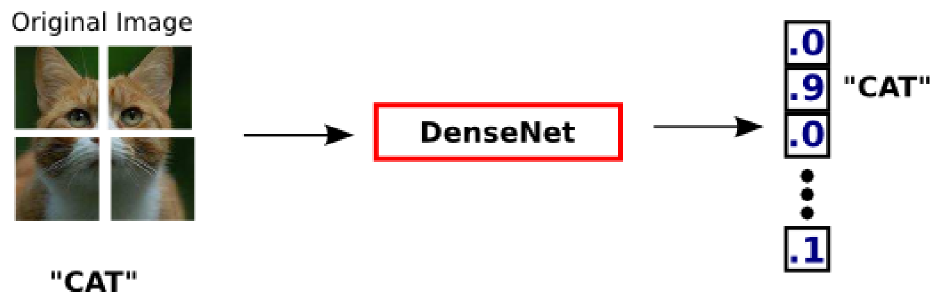
- Finetune on the target task:

VPL Regularizer

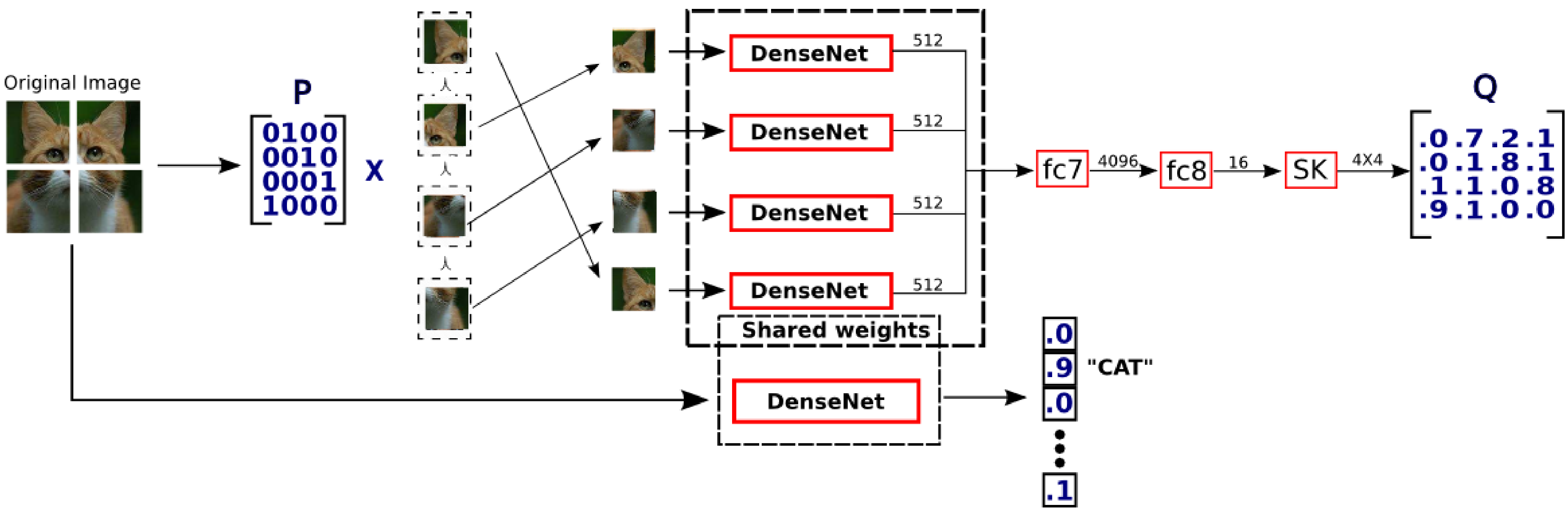
- Pretrain in the visual permutation learning:



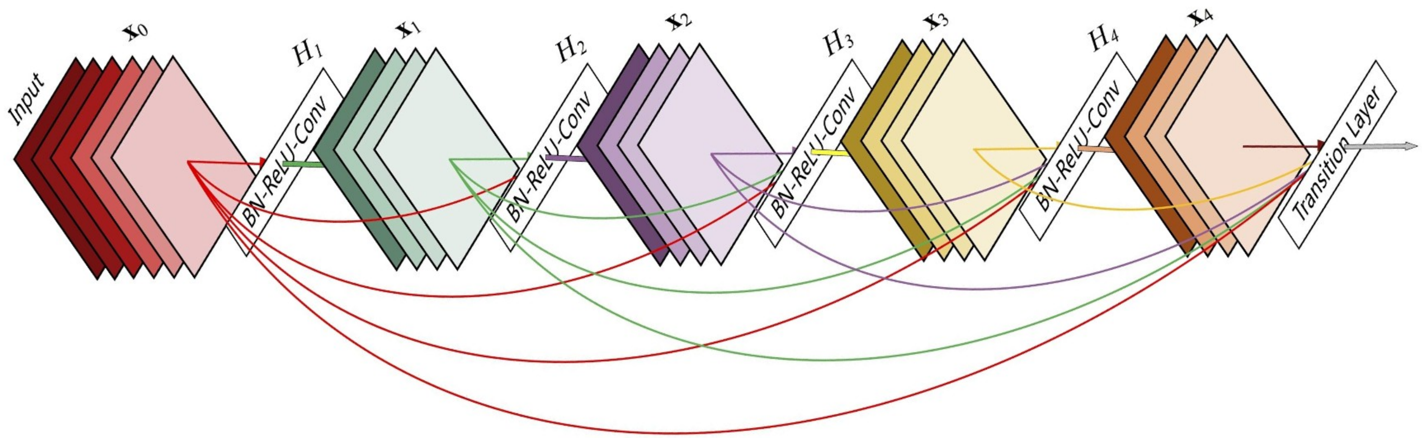
- Finetune on the target task:



VPL Regularizer



Base Classifier - DenseNet121



Training hyper-parameters			
Hyper-parameter	**value**	**Hyper-parameter**	**value**
Learning rate	0.01	Optimizer	SGD
Lr. schedule	Decay by 0.1 every 6 epochs	Momentum	0.9
Batch size	320	Weight Decay	1e-4
Num. epochs	20	Framework	PyTorch

Weighted Random Sampling

"The network often memorizes the category with more instances when trained on an extremely imbalanced dataset."

Then, we adopted a weighted sampling strategy which the probability of a a image *i* been sampled is proportional to the inverse of its frequency,

$$W_i = \frac{N}{N_{c_i}}$$

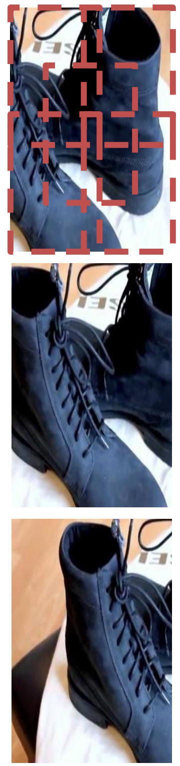
where N is the total number of images and N_{ci} is the number of images belonging to the same class of image *i*.

Multiple Crops Prediction

1) Multiple Dimensions

2) Multiple Regions

3) Multiple Crops



5 image crops

+ Horizontal Flips + Resize (224px)

= 4 x 3 x (5 + 1) x 2 = **144 image crops**

[Christian Szegedy et al. “Going deeper with convolutions”. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015]

Results

Results on the Validation Set	
Model Variation	**Top-5 Acc. (%)**
DenseNet 121 + Center Crop	0.733
DenseNet 121 + 10 Crops	0.748
DenseNet 121 + 144 Crops	0.750
DenseNet 121 + Pretrain + 144 Crops	**0.753**

Remarks

- We got the **third-place** in the competition scoring **69.56%** in top-5 accuracy on the **test set**.
- We are the only team in the top three **not using ensemble** of networks.
- We investigated **Self-supervised pre-training** as a tool to provide robust initialization for deep learning models.
- As recent papers suggest*, deep learning models seems to be reasonably **robust to some types of label noise**.

*[Rolnick et al. “Deep Learning is Robust to Massive Label Noise”.https://arxiv.org/abs/1705.10694]
*[Drory et al. “On the Resistance of Neural Nets to Label Noise”. https://arxiv.org/abs/1803.11410]

Learning CNNs from Web Data

Rodrigo Santa Cruz and Stephen Gould
Australian Centre for Robotic Vision, Australian National University, Canberra, Australia