Unsupervised Learning of Visual Representation by Solving Jigsaw Puzzles, ECCV 16

2018/11/27

20173130 Jaeyoon Kim

CS688
Paper Presentation
KAIST

Image Retrieval with Mixed initiative and Multimodal Feedback, BMVC '18

- The system based on reinforcement learning **chooses an action** and **let users answer** their need or draw a sketch.
- The system Iteratively performs the action selection and finally gets adaptive retrieval result to users.

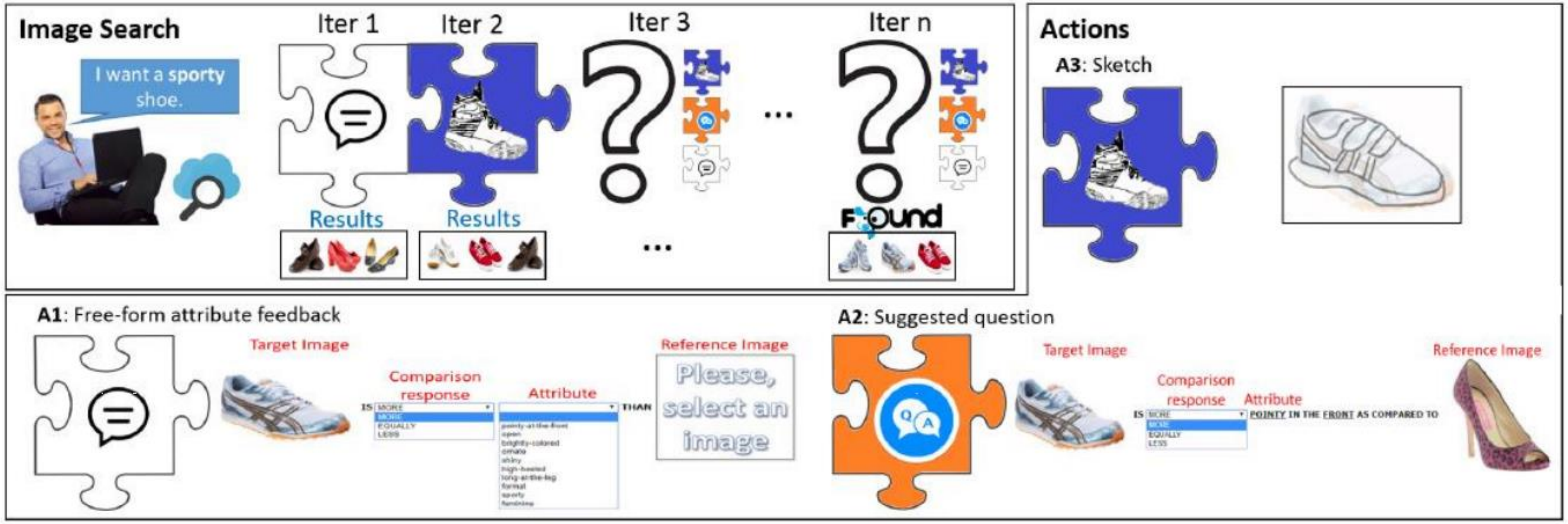


Table of Contents

- Introduction
  - Relationship with Image Retrieval
  - Context prediction task(relative position)
  - Its limitation
- Main Idea
- Experiment & Result

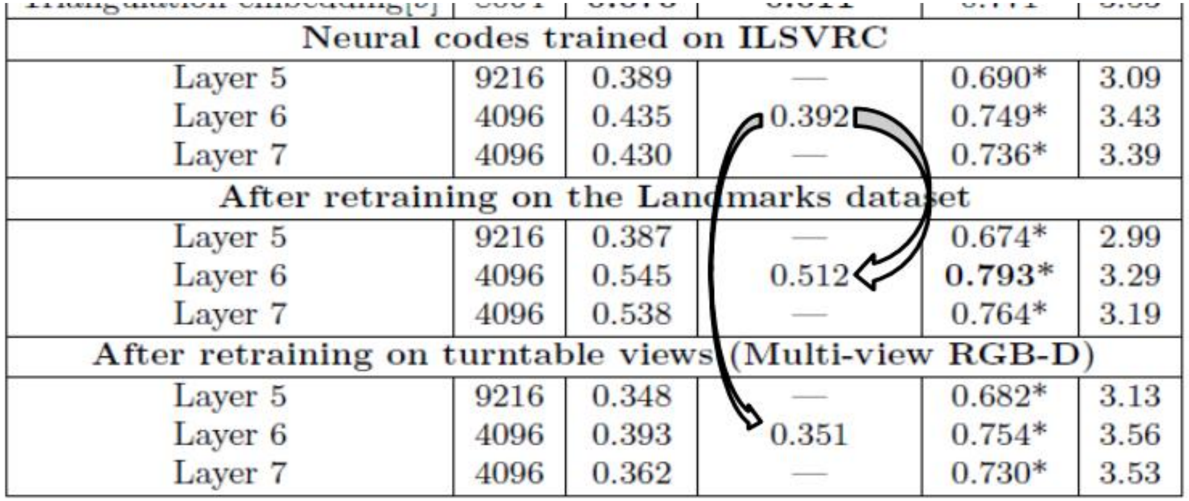
Introduction

- Relationship with Image Retrieval
- Context prediction task(relative position)
- Its limitation

Relationship with Image Retrieval

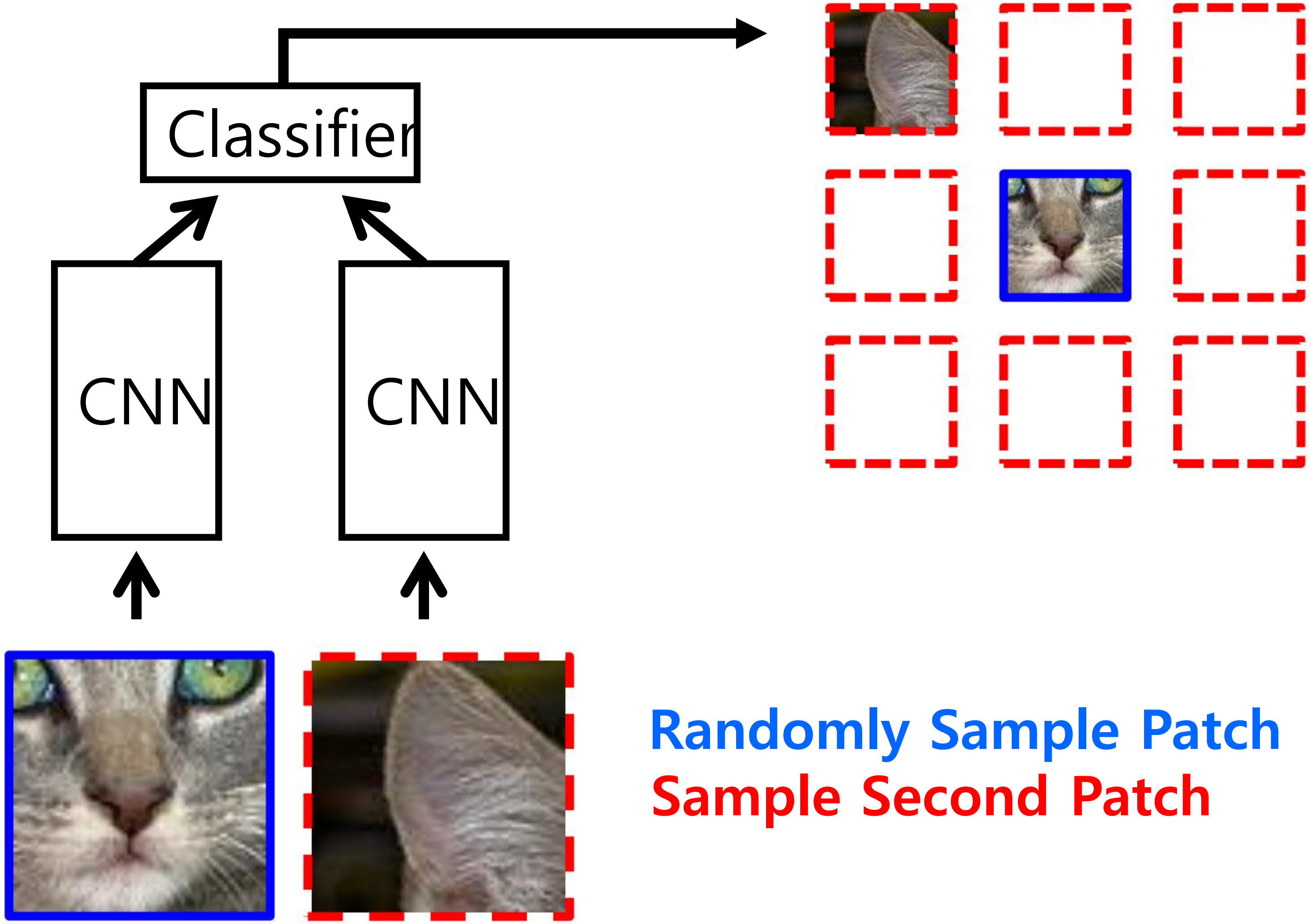
- In the class, we also saw performance improvement when fine-tuning with specific dataset.
- For fine-tuning with specific dataset, labels are necessary since it is performed in a supervised manner.
- Therefore, this unsupervised technique will be useful to cheap fine-tuning for image retrieval.

Figure in the class…



[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Neural codes trained on ILSVRC					
Layer 5	9216	0.389	—	0.690*	3.09
Layer 6	4096	0.435	0.392	0.749*	3.43
Layer 7	4096	0.430	—	0.736*	3.39
After retraining on the Landmarks dataset					
Layer 5	9216	0.387	—	0.674*	2.99
Layer 6	4096	0.545	0.512	**0.793***	3.29
Layer 7	4096	0.538	—	0.764*	3.19
After retraining on turntable views (Multi-view RGB-D)					
Layer 5	9216	0.348	—	0.682*	3.13
Layer 6	4096	0.393	0.351	0.754*	3.56
Layer 7	4096	0.362	—	0.730*	3.53

Context Prediction, ICCV ‘15

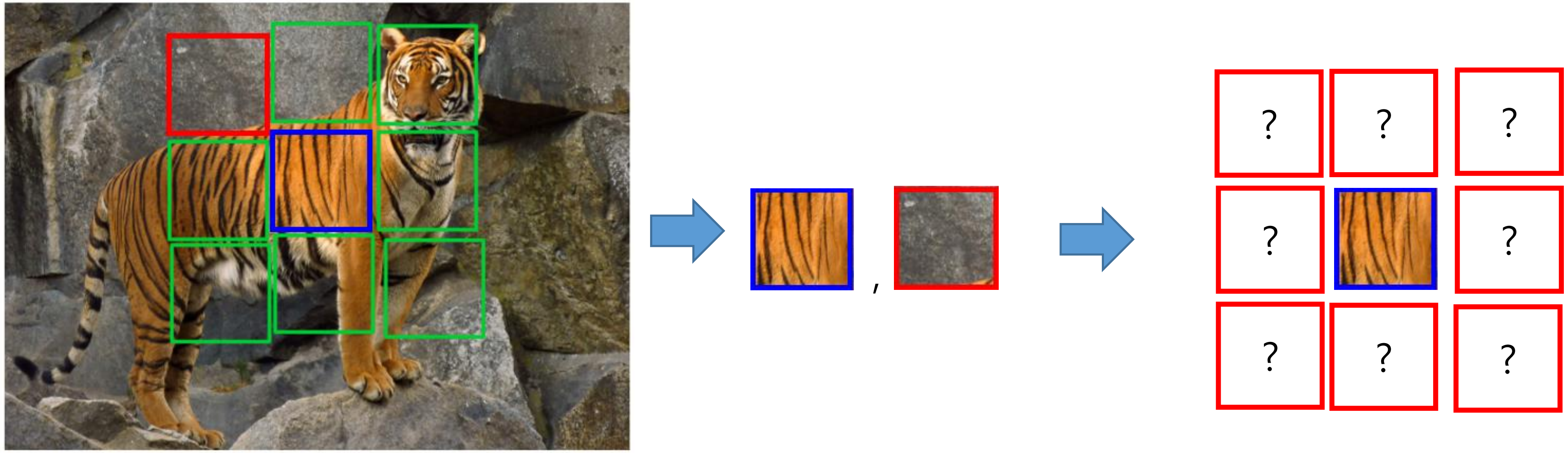


Randomly Sample Patch

Sample Second Patch

Critical Problem of Context Prediction

- If only two tiles are given, the machine might suffer from an ambiguity.
- Can you answer only if the blow blue and red patches are given?
  - There might be **ambiguity**.
  - As its negative effect, it **takes 4 weeks** to train the network with the task. -> very slow!



Main Idea

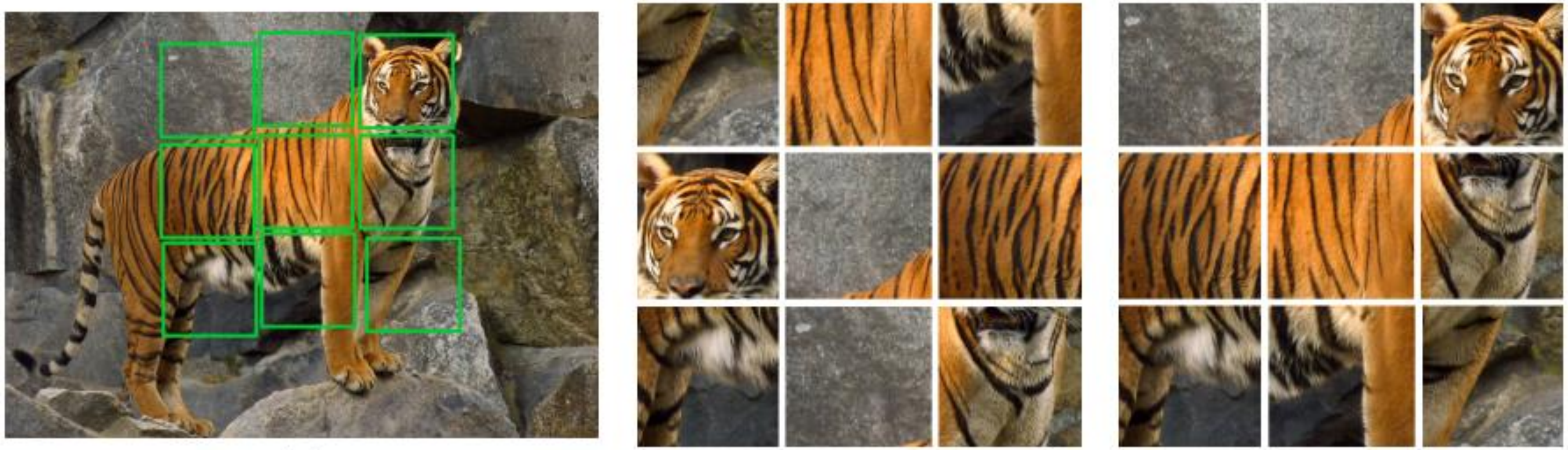
What is jigsaw puzzle?

- The task is to separate an object into several puzzles and put the puzzles together.
- It was introduced as a pretext task to help children learn geography.

An example of this task

1. Sample 9 neighbor tiles - figure (a).
2. Obtain a puzzle by randomly shuffling the sampled tiles – figure (b).
3. Determine all positions of the shuffled tiles - figure (c).

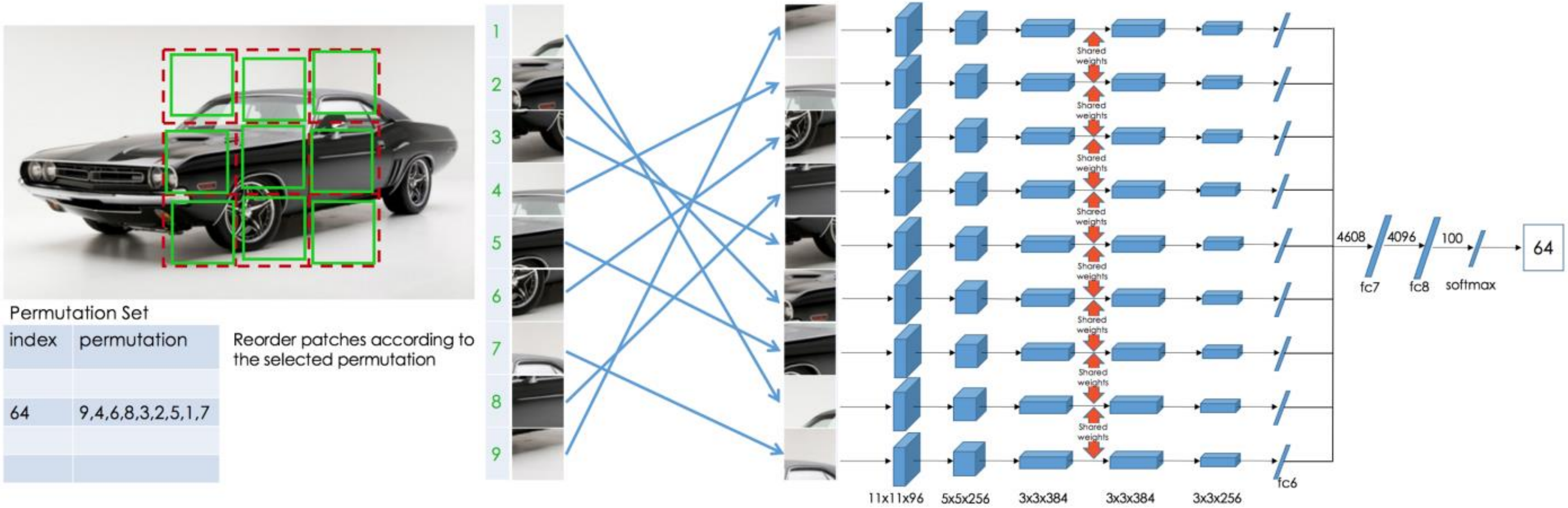
-> This work is **less ambiguous**, compared to previous method since all patches are given to network.



(a) (b) (c)

Problem formulation as classification

- Given 9 tiles, there are 9! = 362,880 possible permutations.
- Due to **too many possible permutation**(classes), They quantize the possible permutation into **64 classes**.



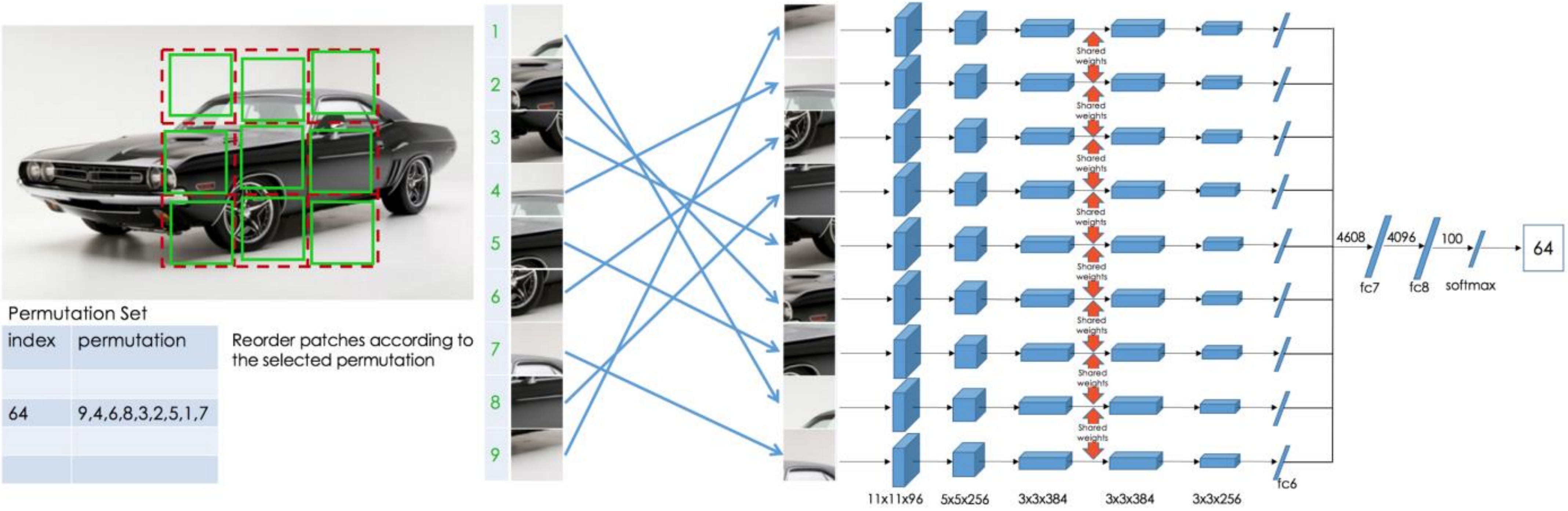
Permutation Set

index	permutation
64	9,4,6,8,3,2,5,1,7

Reorder patches according to the selected permutation

Problem formulation as classification

- The network takes 9 tiles as an input in a siamese manner
- And it predicts a specific sequence among 64 classes.
- Generate **classification loss** and update the network via backpropagation



Permutation Set

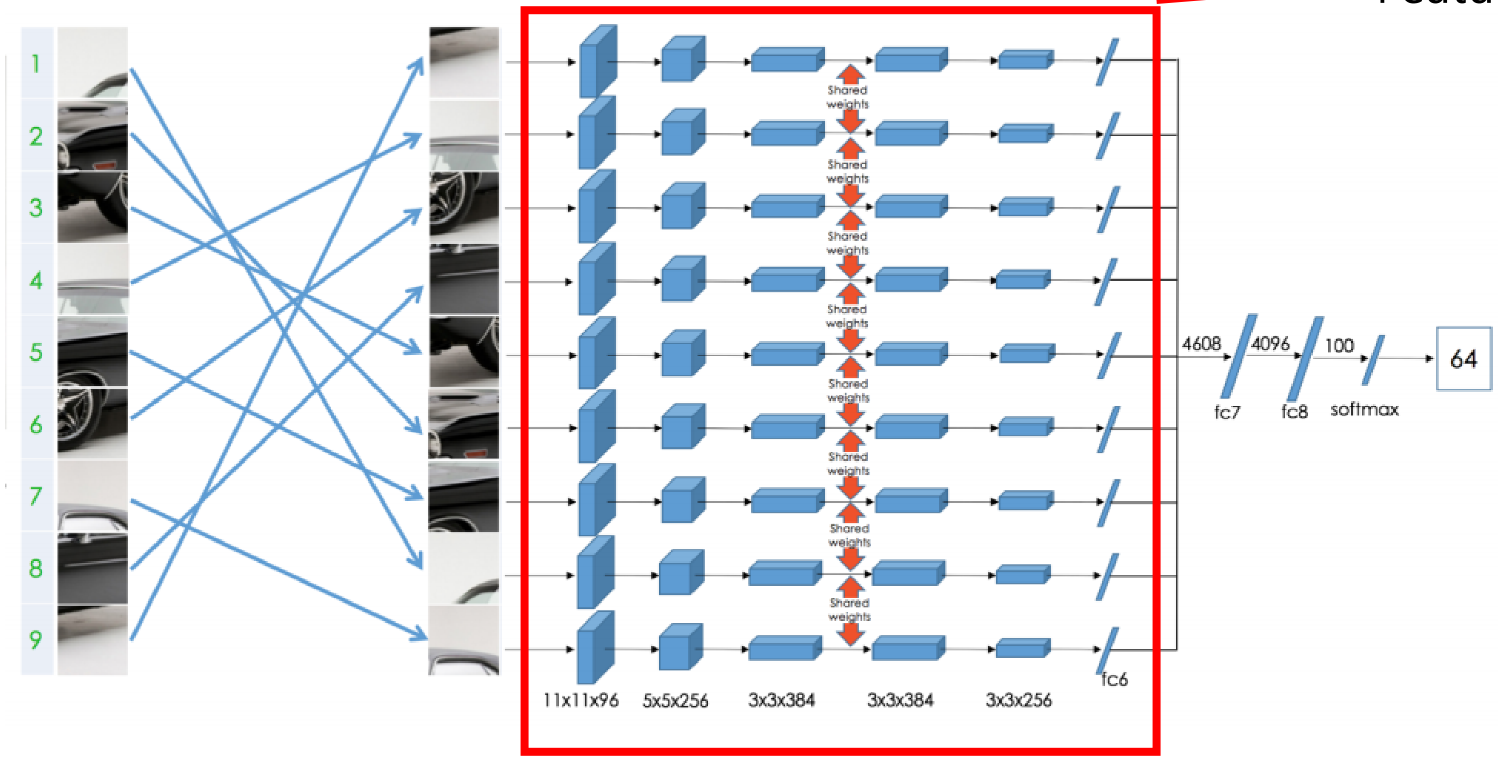
index	permutation
64	9,4,6,8,3,2,5,1,7

Reorder patches according to the selected permutation

Experiments & Results

Transfer learning for evaluation

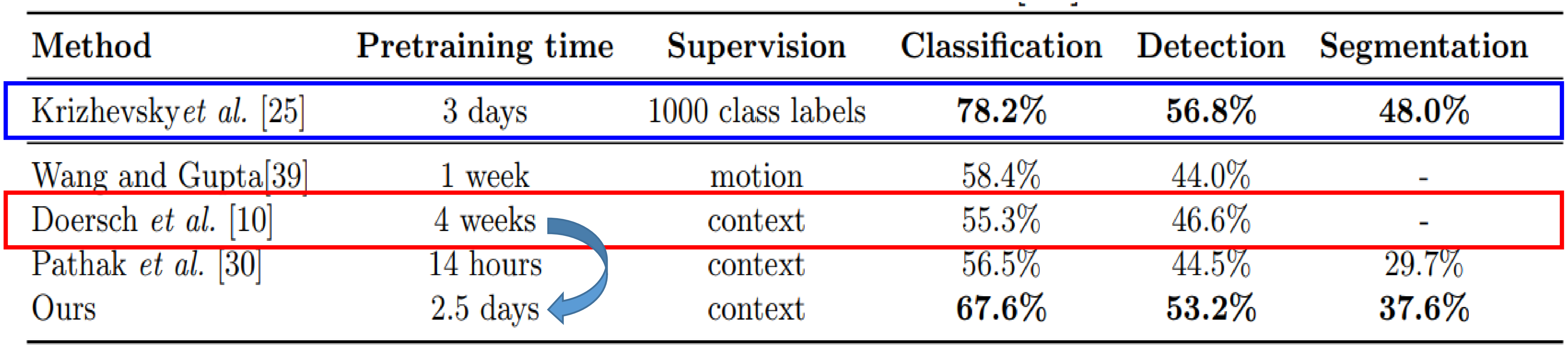
- They use the feature extractor which is in below red box for evaluating the network.
- They perform transfer learning for each task such as classification, detection and semantic segmentation.



Feature extractor

Results on PASCAL VOC 2007

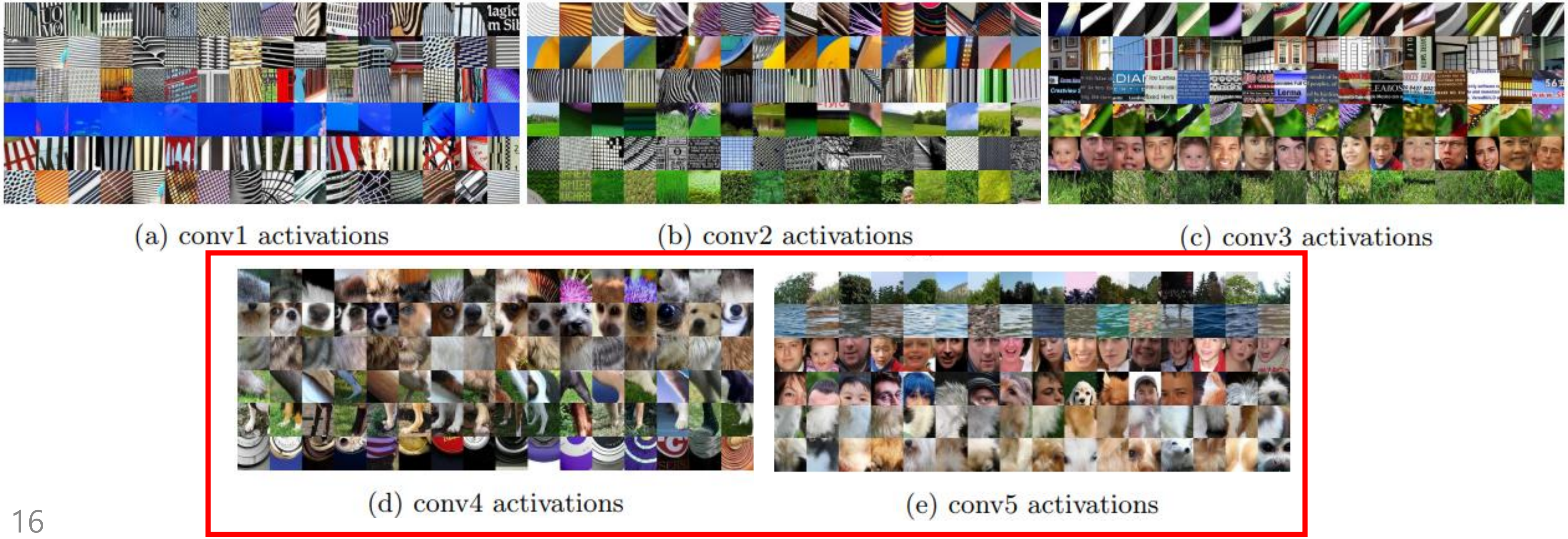
- They fine-tuned the pre-trained network with PASCAL VOC training data.
- **Blue box** is a supervised method and **red box** is Context Prediction method.
- This method is much superior to Context Prediction in terms of **pre-training time** as well as accuracy thanks to **less ambiguity** of the task.



Method	Pretraining time	Supervision	Classification	Detection	Segmentation
Krizhevsky*et al.* [25]	3 days	1000 class labels	**78.2%**	**56.8%**	**48.0%**
Wang and Gupta[39]	1 week	motion	58.4%	44.0%	-
Doersch *et al.* [10]	4 weeks	context	55.3%	46.6%	-
Pathak *et al.* [30]	14 hours	context	56.5%	44.5%	29.7%
Ours	2.5 days	context	**67.6%**	**53.2%**	**37.6%**

Visualization of top activations

- We can see that the network is able to **capture semantic information** as going to higher layer even though any semantic label is not given during training.



(a) conv1 activations

(b) conv2 activations

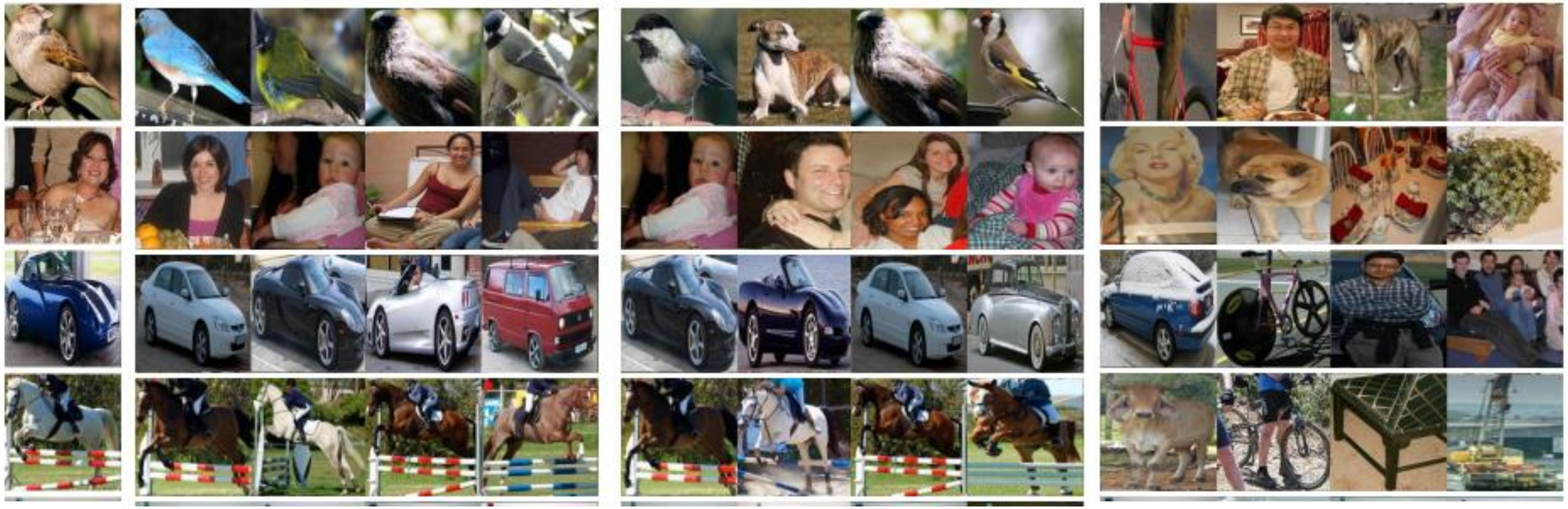
(c) conv3 activations

(d) conv4 activations

(e) conv5 activations

Image Retrieval Results

- They found nearest neighbor results on the PASCAL VOC dataset



query

This method

Supervised method

Random weight

Thank you!!