Unsupervised Visual Representation Learning by Context Prediction, ICCV 15

2018/10/25

20173130 Jaeyoon Kim

CS688
Paper Presentation

Table of Contents

- Introduction
  - Self-Supervised Learning
  - Relationship with Image Retrieval

- Main Idea

- Experiment & Result

Introduction

-Self-Supervised Learning
-Relationship with Image Retrieval

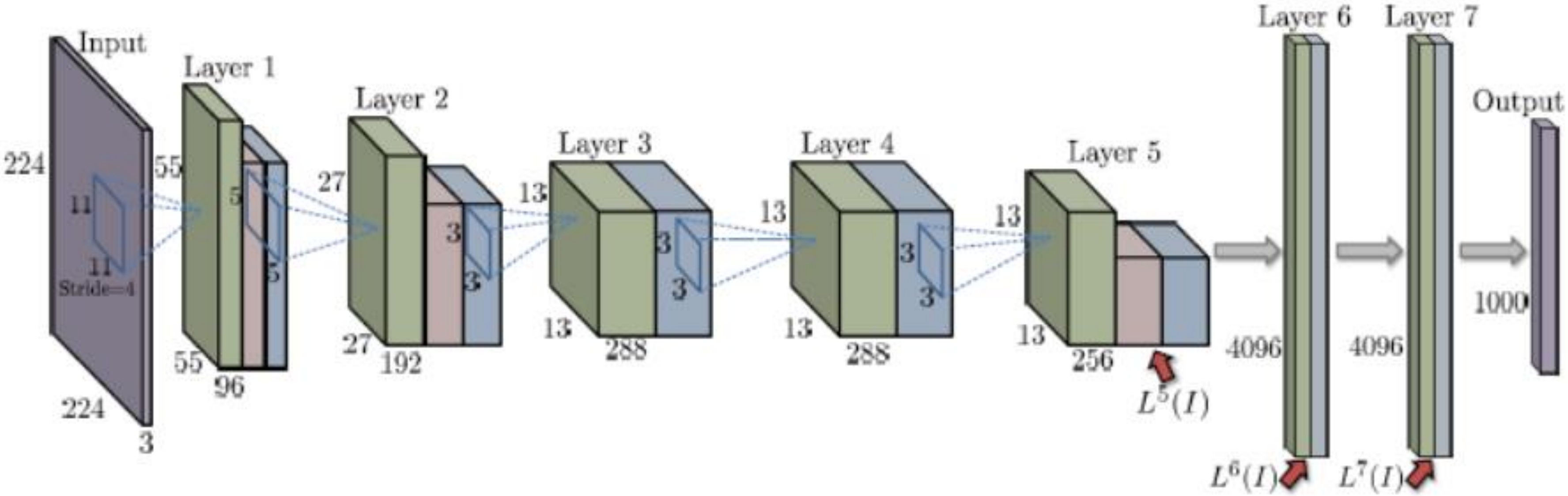
Self-Supervised Learning

- Supervised Learning(ImageNet)
  - Need labels for training the network.
  - The labels only can be obtained by human annotator.
  - So, annotating is very expensive or sometimes impossible.

- Self-Supervised Learning
  - A form of unsupervised learning where the data itself provides the supervision
  - Namely, it is able to automatically obtain labels for specific task.

Relationship with Image Retrieval

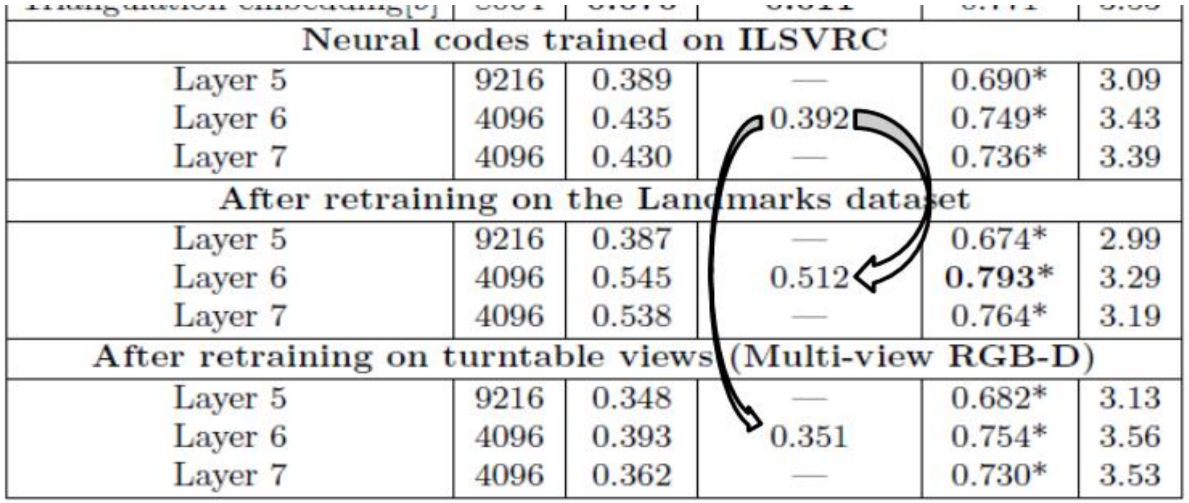
- These days, Deep features are widely used for Image Retrieval thanks to its performance
  - Ex) Neural codes for Image Retrieval(ECCV 14).



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…

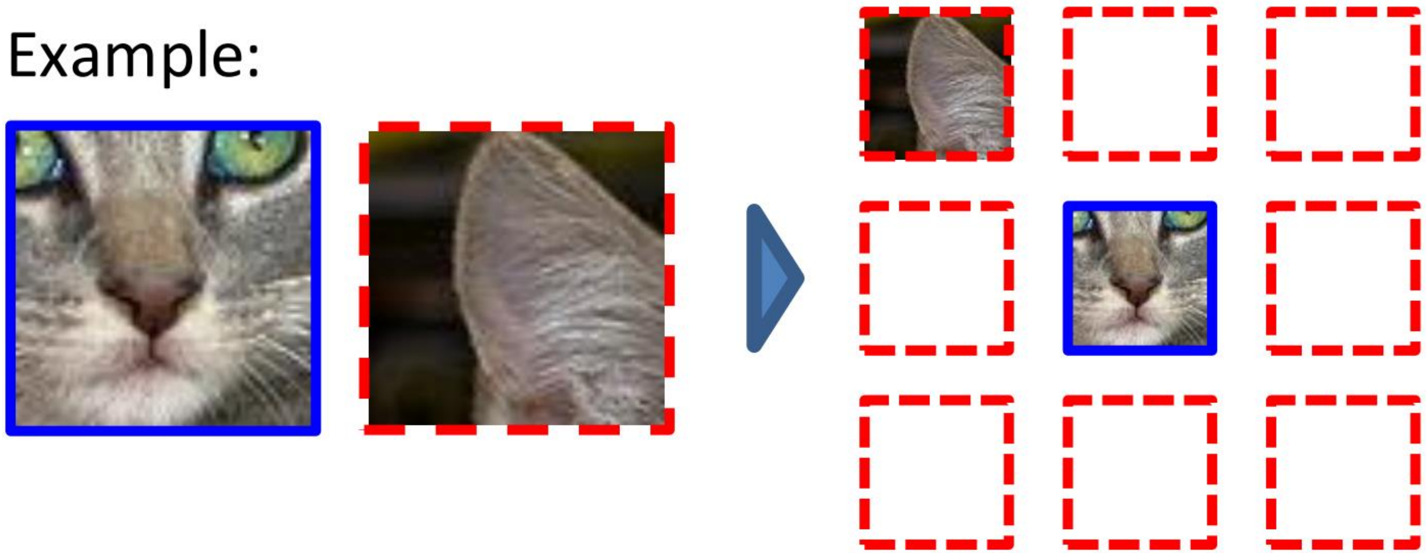


Triangulation embedding[5]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Neural codes trained on ILSVRC					
Layer 5	9216	0.389	—	0.690*	3.09
Layer 6	4096	0.435	0.392	0.749*	3.43
Layer 7	4096	0.430	—	0.736*	3.39
After retraining on the Landmarks dataset					
Layer 5	9216	0.387	—	0.674*	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

Main Idea

Learning to Predict Relative Position

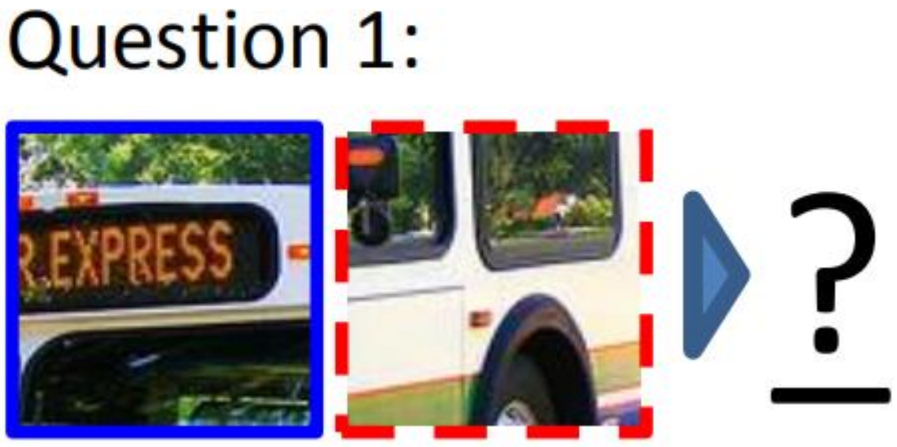
- What is the task of predicting relative position?



ition of red box based
upper-right?

Learning to Predict Relative Position

- What is the task of predicting relative position?
  - People can easily answer this relative position task.
  - This is hard if you don’t know what a cat is, but easy if you know its semantic.
  - So, If the machine do this task well, then we can think the machine is able to capture the semantic information.

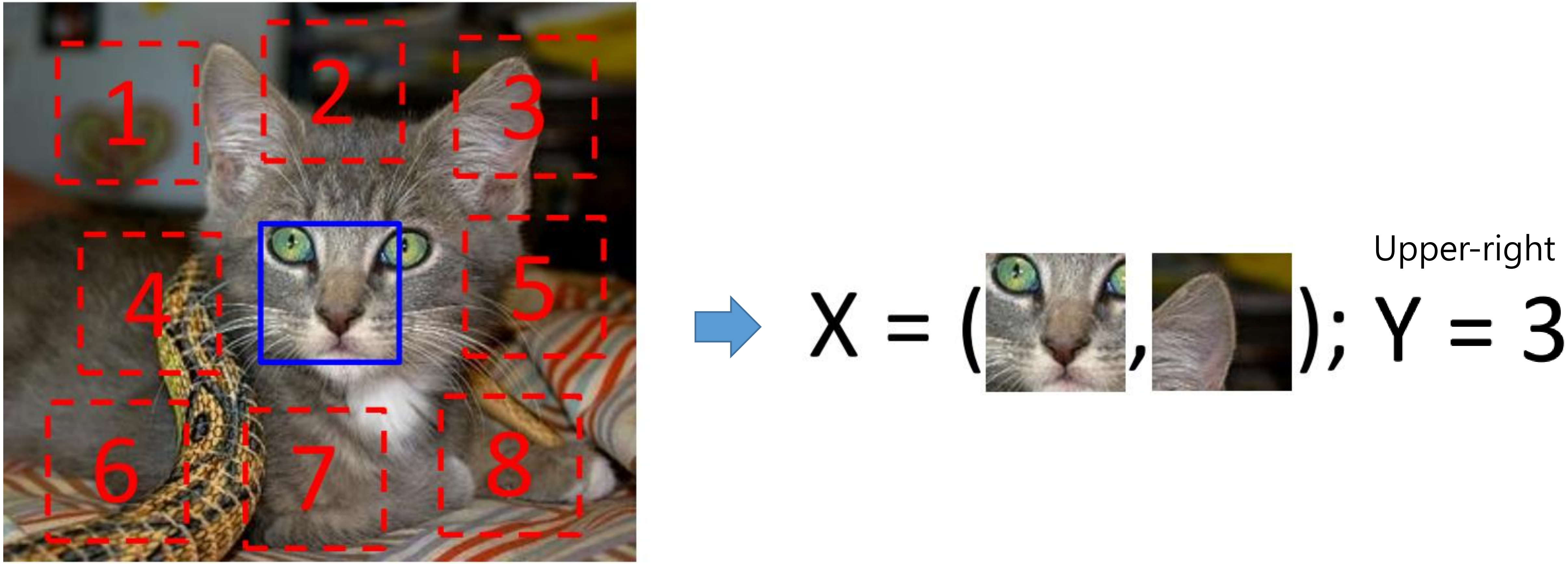


How does child learn from the puzzle game?

Cropped from 중앙일보

Problem Formulation for Machine

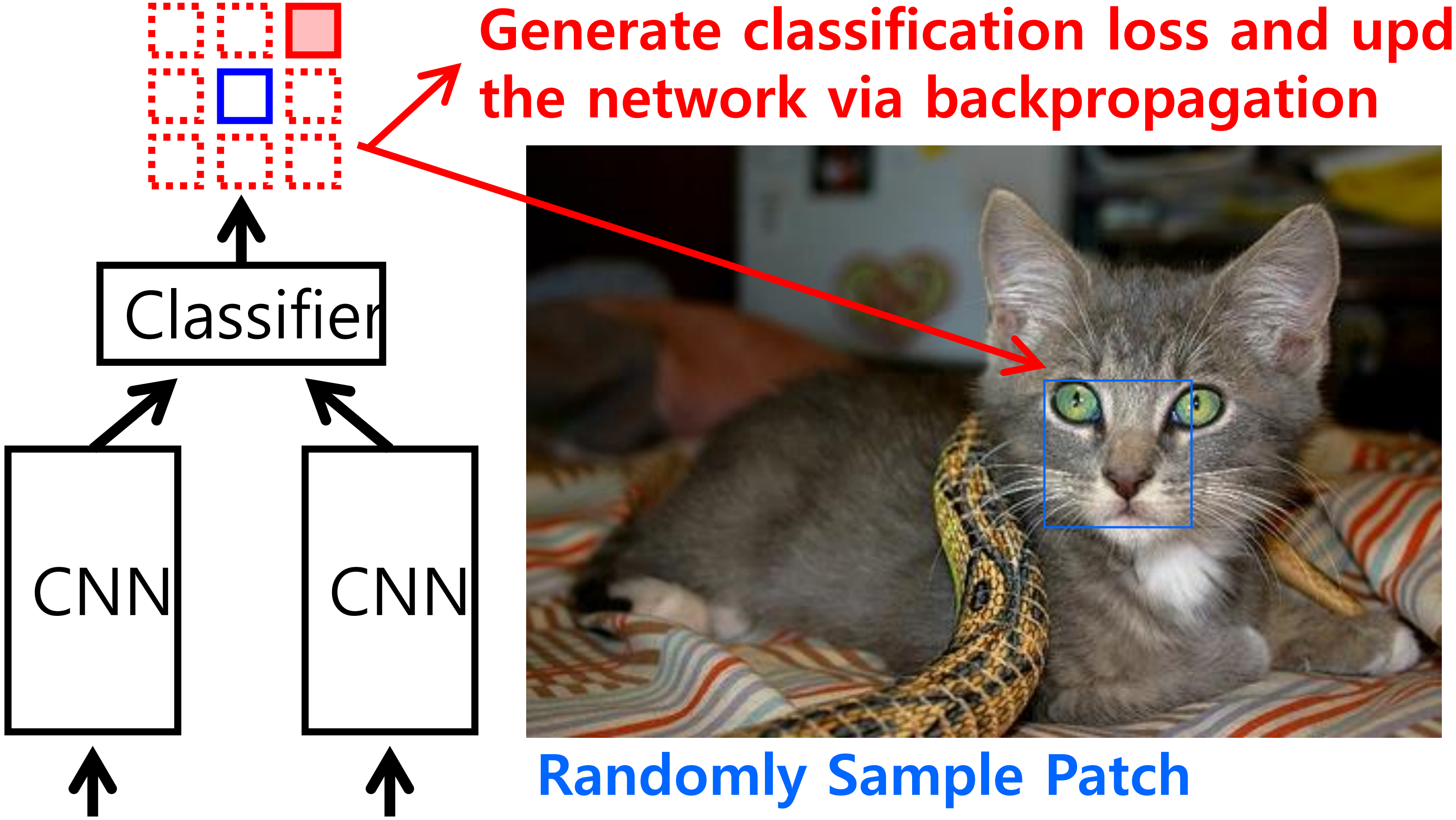
- Interpreting the relative position task as classification problem(8 classes)



Upper-right

$X = (\; , \;); Y = 3$

Task Sequence for Training Network



Generate classification loss and update the network via backpropagation

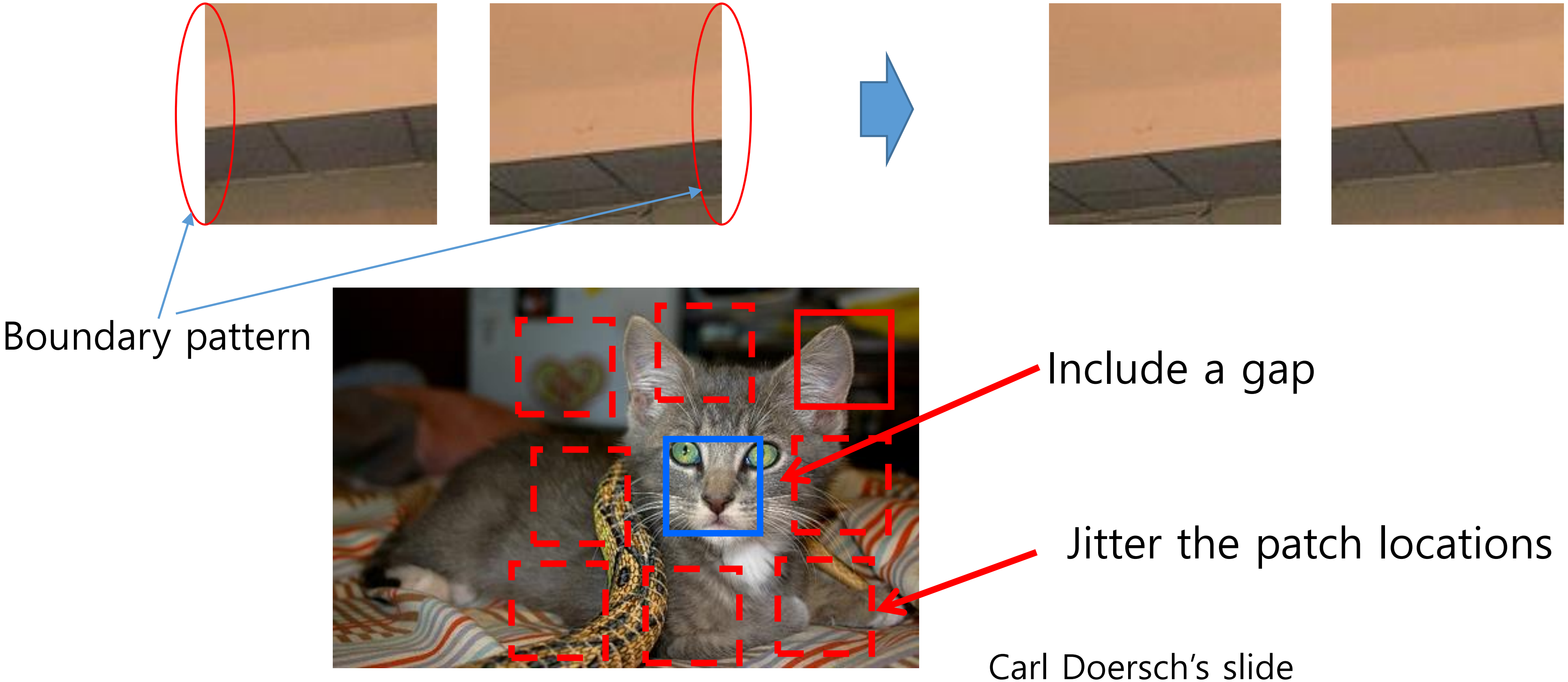
Randomly Sample Patch

Sample Second Patch

Carl Doersch's slide

Avoiding trivial solutions

- For easily solving this task, the machine is likely to capture boundary patterns or textures rather than semantic information as a cue.



Boundary pattern

Include a gap

Jitter the patch locations

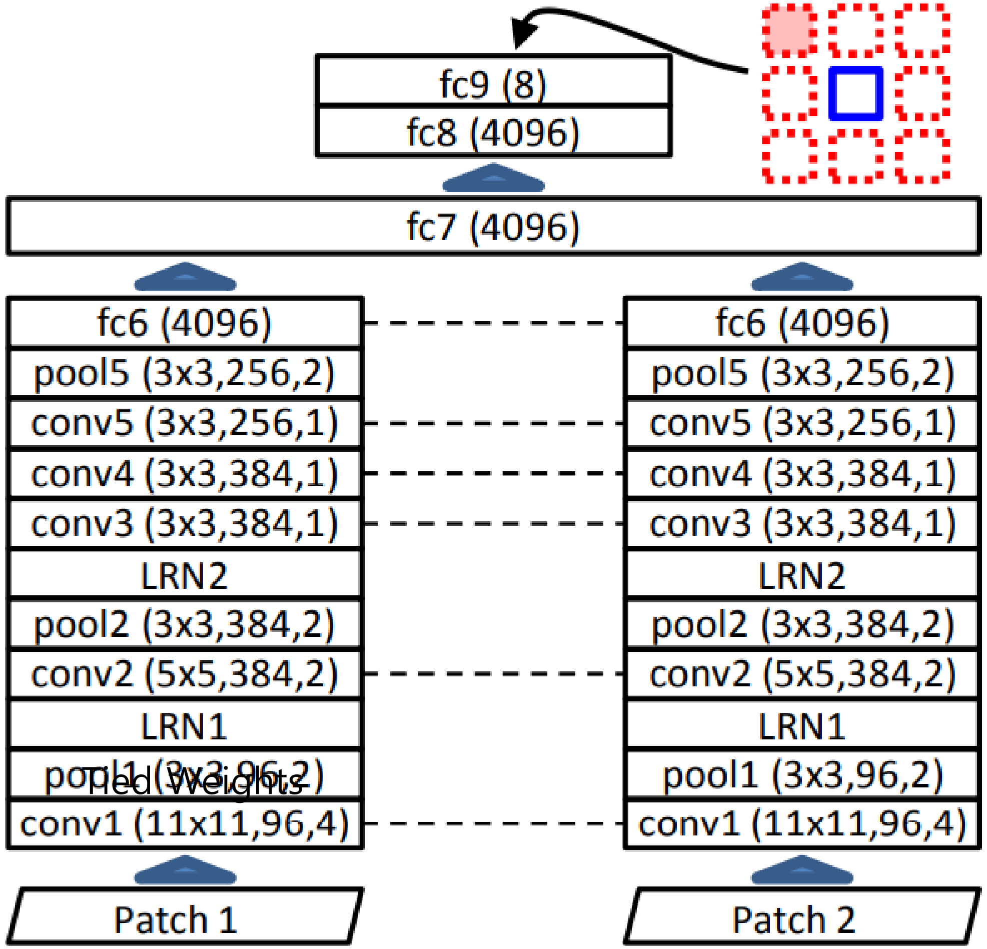
Carl Doersch's slide

Experiments & Results

Network Architecture

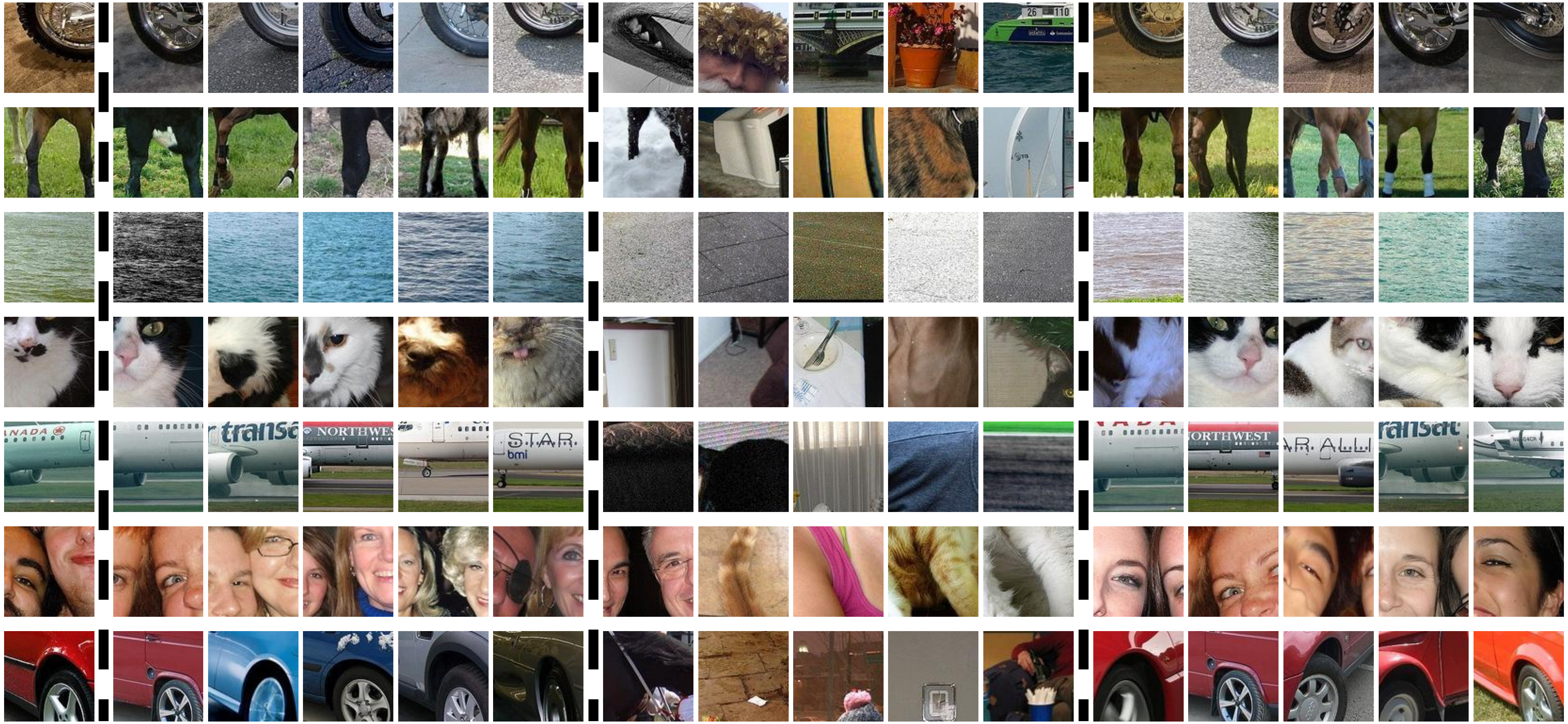
- They use quite simple architecture that is manually designed
- Network is learned from scratch without any pre-training
- Training with ImageNet
- Remove domain-specific layer when applying to other domain.

A relatively small # of layers compared to VGG and Alexnet.



Nearest Neighbors

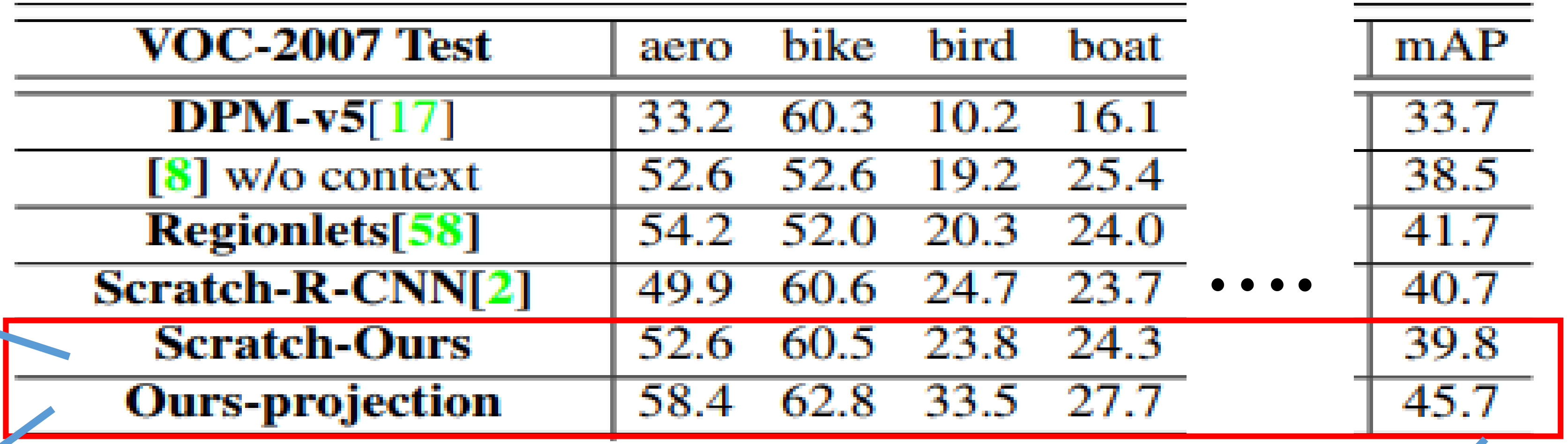
- Nearest neighbors of specific patches.(Thanks to capturing semantics)



Input Theirs Random Initialization ImageNet AlexNet

Object Detection

- Pascal VOC-2007 dataset



VOC-2007 Test	aero	bike	bird	boat		mAP
DPM-v5[17]	33.2	60.3	10.2	16.1		33.7
[8] w/o context	52.6	52.6	19.2	25.4		38.5
Regionlets[58]	54.2	52.0	20.3	24.0		41.7
Scratch-R-CNN[2]	49.9	60.6	24.7	23.7	....	40.7
Scratch-Ours	52.6	60.5	23.8	24.3		39.8
Ours-projection	58.4	62.8	33.5	27.7		45.7

Only supervised training from scratch

Unsupervised pre-training and supervised fine-tuning for Pascal VOC

Boosting by 6%

Thank you!!