Unsupervised Visual Representation Learning by Context Prediction

School of Industrial and Management Engineering, Korea University

Jihyun Kim

KOREA UNIVERSITY

DMQA hcai Human-Centered Artificial Intelligence Center

Contents

- Research Purpose
- Learning Visual Context Prediction
- Experiments and Results
- Conclusion

Research Purpose

❖ Unsupervised Visual Representation Learning by Context Prediction(arXiv, 2015)

- Carnegie Mellon University, UC Berkely에서 연구하였고 2022년 04월 26일 기준 1850회 인용됨

Unsupervised Visual Representation Learning by Context Prediction

Carl Doersch[1,2] Abhinav Gupta[1] Alexei A. Efros[2]

[1] School of Computer Science
Carnegie Mellon University

[2] Dept. of Electrical Engineering and Computer Science
University of California, Berkeley

Abstract

This work explores the use of spatial context as a source of free and plentiful supervisory signal for training a rich visual representation. Given only a large, unlabeled image collection, we extract random pairs of patches from each image and train a convolutional neural net to predict the position of the second patch relative to the first. We argue that doing well on this task requires the model to learn to recognize objects and their parts. We demonstrate that the feature representation learned using this within-image context indeed captures visual similarity across images. For example, this representation allows us to perform unsupervised visual discovery of objects like cats, people, and even birds from the Pascal VOC 2011 detection dataset. Furthermore, we show that the learned ConvNet can be used in the R-CNN framework [19] and provides a significant boost over a randomly-initialized ConvNet, resulting in state-of-the-art performance among algorithms which use only Pascal-provided training set annotations.

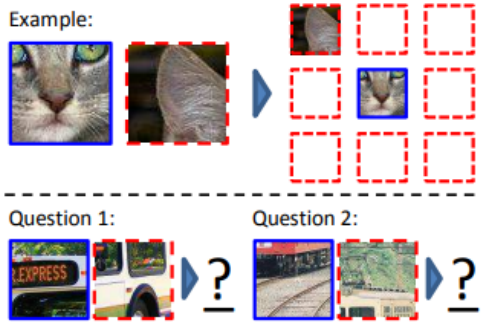


Figure 1. Our task for learning patch representations involves randomly sampling a patch (blue) and then one of eight possible neighbors (red). Can you guess the spatial configuration for the two pairs of patches? Note that the task is much easier once you have recognized the object!

Answer key: Q1: Bottom right Q2: Top center

in the context (i.e., a few words before and/or after) given

Research Purpose

❖ Unsupervised Visual Representation Learning by Context Prediction(arXiv, 2015)

- Visual Recognition Task에서 Self-Supervised Learning을 위한 방법론 제시
  - ✓ 레이블이 없는 대량의 이미지 데이터셋으로부터 유의미한 정보를 추출하기 위함
- **Spatial Context를 이용**하여 Visual Representation을 학습
  - ✓ 하나의 이미지로부터 9개의 이미지 Patch를 생성하고, **Patch의 위치 관계를 예측**
  - ✓ 이를 위해 **모델이 이미지의 문맥(Context)을 이해하도록 하는 것이 목표**

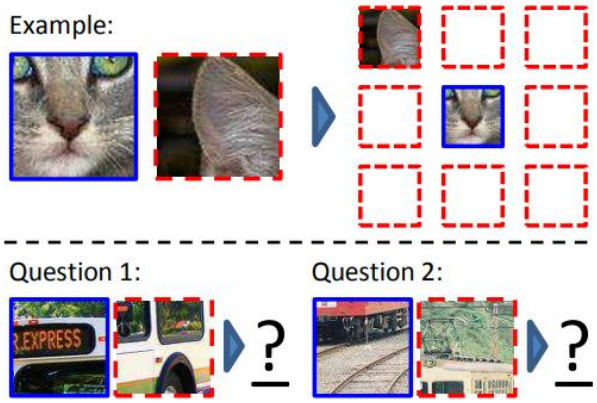


Figure 1. Our task for learning patch representations involves randomly sampling a patch (blue) and then one of eight possible neighbors (red). Can you guess the spatial configuration for the two pairs of patches? Note that the task is much easier once you have recognized the object!

Answer key: Q1: Bottom right Q2: Top center

Learning Visual Context Prediction

❖ Unsupervised Visual Context Prediction

- **Pretext Task: Context Prediction** (Relative Patch Location)
- 한 장의 이미지에서 9개의 Patches를 얻은 뒤, 중앙 위치의 Patch를 기준으로 나머지 Patches가 어느 위치에 있는지 분류하도록 학습
  - ✓ 전체 이미지에서 지엽적인 두 개의 Patches만을 보고 상대적인 위치를 맞추는 문제
- 이를 위해서는 **이미지 전체의 Context를 이해할 수 있어야 함**

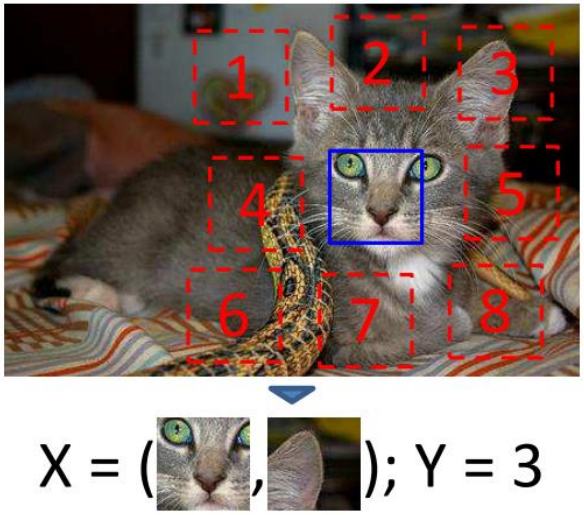


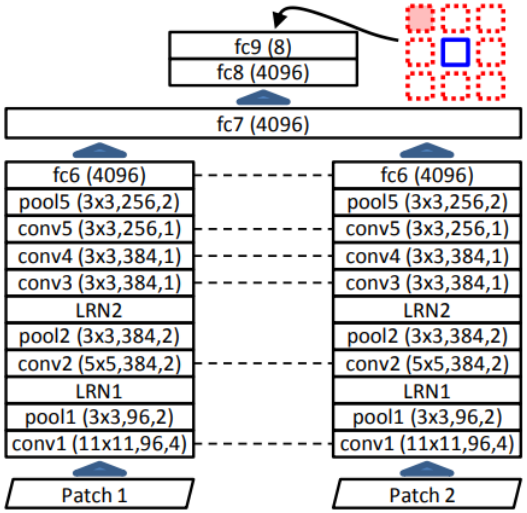
Figure 2. The algorithm receives two patches in one of these eight possible spatial arrangements, without any context, and must then classify which configuration was sampled.

Learning Visual Context Prediction

❖ Model Architecture

- 2개의 CNN 구조로 이루어진 Pair Classification을 위한 모델 제안
  - Patch 1: 9개 Patches 중 중앙의 Patch

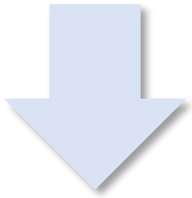
    Patch 2: 무작위로 선택한 나머지 8개 중 하나의 Patch
- 각 Patch는 각각의 CNN을 통과하고, 결과값이 fc7에서 합쳐져 최종적으로 fc9에서 위치를 예측하는 구조로 이루어짐
  - ✓ 2개의 CNN은 각 Patch에 대한 Representation을 Weight Sharing하여 학습 (Dotted Lines indicate shared weights)



Learning Visual Context Prediction

❖ Avoiding "Trivial" Solutions (1/2)

- 모델이 **Low-Level Cues를 이용하여 Patch 간 관계를 예측하는 것을 방지**하는 Solutions 제안
- 네트워크가 Trivial Shortcut 없이 High-Level Semantics를 이용하여 정보를 추출하도록 설계
  - ✓ **Low-Level Cues**: Boundary Pattern, Patch 사이의 Textures, etc.

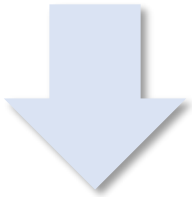


- **Sol 1) Including a Gap between patches** (approximately half the patch width)
  - ✓ 한계: It's possible that long lines spanning neighboring patches could give away the correct answer
- **Sol 2) Random Jittering 수행**: Each patch location by up to 7 pixels

Learning Visual Context Prediction

❖ Avoiding "Trivial" Solutions (2/2)

- **Chromatic Aberration을 방지**하기 위한 Solutions 제안
  - ✓ **Chromatic Aberration**: 렌즈에서 유리 굴절률이 빛의 파장에 따라 다르기 때문에 생기는 수차
    (e.g. 이미지에서 Green 채널이 다른 채널에 비해 카메라 중앙 쪽으로 치우쳐 나타남. 이때 ConvNet은 상대적인 green, magenta 채널의 분리를 인식하여 trivial solution을 생성할 수 있음)



- **Sol 1) Projection 수행:** To shift green and magenta towards gray

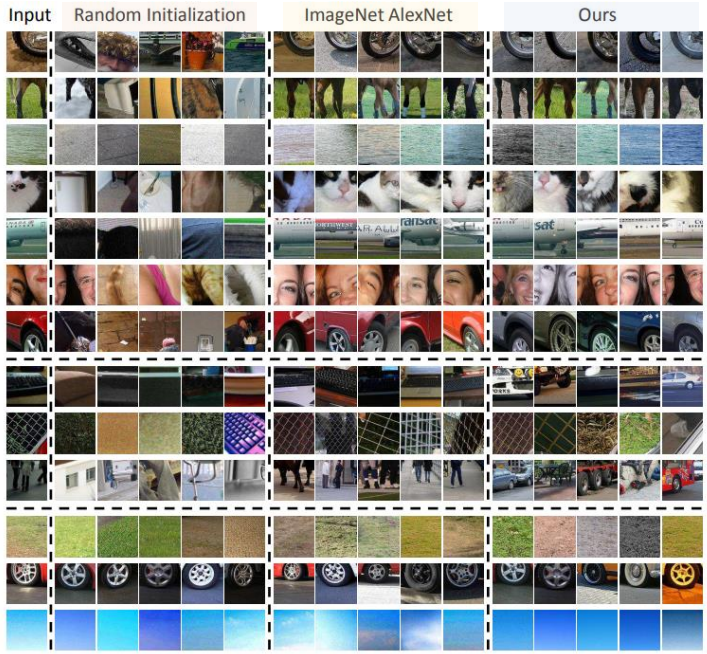
  (RGB 공간에서 Green-Magenta color axis가 a=[-1,2,-1로 정의된다면, B=I−a'a/(aa')를 정의하고, 모든 픽셀 값에 B를 곱하면 색 수차를 이미지에서 뺄 수 있음)

- **Sol 2) Color Dropping 수행:** Randomly drop 2 of the 3 color channels from each patch, replacing the dropped colors with Gaussian noise

Experiments and Results

❖ Nearest Neighbors

- Normalized Correlation 을 이용해 의미적으로 유사한 Patches를 찾아내도록 학습
- 비교 실험:
  - ✓ 학습이 되지 않은 제안 모델, ImageNet으로 학습된 AlexNet, 학습이 된 제안 모델 3가지에 대하여 Nearest Neighbors를 잘 찾아내는지를 확인 (** 제안 모델은 기존의 Architecture에서 fc6까지의 레이어만 이용)



Top Group: 제안 모델의 결과 우수

Middle Group: AlexNet의 결과 우수

Bottom Group: 세 모델 모두 좋은 결과

Experiments and Results

❖ Object Detection(1/2)

- 제안 모델을 Pre-trained Model으로 사용하여 한정된 학습 데이터만으로 Object Detection을 수행
  - ✓ Dataset: VOC 2007
- R-CNN을 Pipeline으로 이용하였으며, 이때 Object Proposal이 Patch 사이즈(96x96)로 줄어드는 것을 방지하고자 **모델의 구조를 변경**

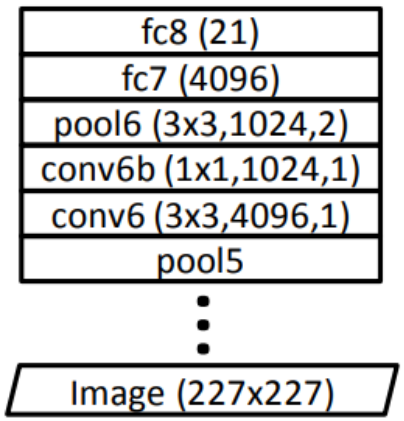


Figure 6. Our architecture for Pascal VOC detection. Layers from conv1 through pool5 are copied from our patch-based network (Figure 3). The new 'conv6' layer is created by converting the fc6 layer into a convolution layer. Kernel sizes, output units, and stride are given in parentheses, as in Figure 3.

Experiments and Results

❖ Object Detection(2/2)

- 제안 모델을 Pre-trained Model으로 사용하여 한정된 학습 데이터만으로 Object Detection을 수행
  - ✓ Dataset: VOC 2007
- R-CNN을 Pipeline으로 이용하였으며, 이때 Object Proposal이 Patch 사이즈(96x96)로 줄어드는 것을 방지하고자 모델의 구조를 변경
- 학습이 잘 된 제안 모델이 학습되지 않은 경우 대비 MAP가 6% 높았고, AlexNet-Style 모델 대비 MAP가 5% 높았음

VOC-2007 Test	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
DPM-v5[17]	33.2	60.3	10.2	16.1	**27.3**	54.3	58.2	23.0	20.0	24.1	26.7	12.7	58.1	48.2	43.2	12.0	21.1	36.1	46.0	43.5	33.7
[8] w/o context	52.6	52.6	19.2	25.4	18.7	47.3	56.9	42.1	16.6	41.4	41.9	27.7	47.9	51.5	29.9	20.0	41.1	36.4	48.6	53.2	38.5
Regionlets[55]	54.2	52.0	20.3	24.0	20.1	55.5	68.7	42.6	19.2	44.2	**49.1**	26.6	57.0	54.5	43.4	16.4	36.6	37.7	**59.4**	52.3	41.7
Scratch-R-CNN[2]	49.9	60.6	24.7	23.7	20.3	52.5	64.8	32.9	20.4	43.5	34.2	29.9	49.0	60.4	47.5	28.0	42.3	28.6	51.2	50.0	40.7
Scratch-Ours	52.6	60.5	23.8	24.3	18.1	50.6	65.9	29.2	19.5	43.5	35.2	27.6	46.5	59.4	46.5	25.6	42.4	23.5	50.0	50.6	39.8
Ours-projection	58.4	62.8	**33.5**	27.7	24.4	**58.5**	68.5	41.2	**26.3**	**49.5**	42.6	**37.3**	**55.7**	**62.5**	49.4	**29.0**	**47.5**	28.4	54.7	56.8	45.7
Ours-color-dropping	**60.5**	**66.5**	29.6	**28.5**	26.3	56.1	**70.4**	**44.8**	24.6	45.5	45.4	35.1	52.2	60.2	**50.0**	28.1	46.7	**42.6**	54.8	**58.6**	**46.3**
Ours-Yahoo100m	56.2	63.9	29.8	27.8	23.9	57.4	69.8	35.6	23.7	47.4	43.0	29.5	52.9	62.0	48.7	28.4	45.1	33.6	49.0	55.5	44.2
Ours-VGG	63.6	64.4	42.0	42.9	18.9	67.9	69.5	65.9	28.2	48.1	58.4	58.5	66.2	64.9	54.1	26.1	43.9	55.9	69.8	50.9	53.0
ImageNet-R-CNN[19]	64.2	69.7	50	41.9	32.0	62.6	71.0	60.7	32.7	58.5	46.5	56.1	60.6	66.8	54.2	31.5	52.8	48.9	57.9	64.7	54.2

Table 1. Results on VOC-2007. R-CNN performance with our unsupervised pre-training is 5% MAP better than training from scratch, but still 8% below pre-training with ImageNet label supervision.


Thank You