Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles

School of Industrial and Management Engineering, Korea University

Hyeonji Kim

KOREA UNIVERSITY

DMQA hcai Human-Centered Artificial Intelligence Center

Contents

Research Purpose

❖ **Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles (ECCV 2016)**

- Institute for Informatiks, University of Bern에서 연구하였으며 2022년 4월 28일 기준으로 1694회 인용됨

Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles

Mehdi Noroozi and Paolo Favaro

Institute for Informatiks
University of Bern
{noroozi,paolo.favaro}@inf.unibe.ch

Abstract. In this paper we study the problem of image representation learning without human annotation. By following the principles of self-supervision, we build a convolutional neural network (CNN) that can be trained to solve Jigsaw puzzles as a *pretext* task, which requires no manual labeling, and then later repurposed to solve object classification and detection. To maintain the compatibility across tasks we introduce the *context-free network* (CFN), a siamese-ennead CNN. The CFN takes image tiles as input and explicitly limits the receptive field (or context) of its early processing units to one tile at a time. We show that the CFN includes fewer parameters than AlexNet while preserving the same semantic learning capabilities. By training the CFN to solve Jigsaw puzzles, we learn both a feature mapping of object parts as well as their correct spatial arrangement. Our experimental evaluations show that the learned features capture semantically relevant content. Our proposed method for learning visual representations outperforms state of the art methods in several transfer learning benchmarks.

Research Purpose

❖ **Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles (ECCV 2016)**

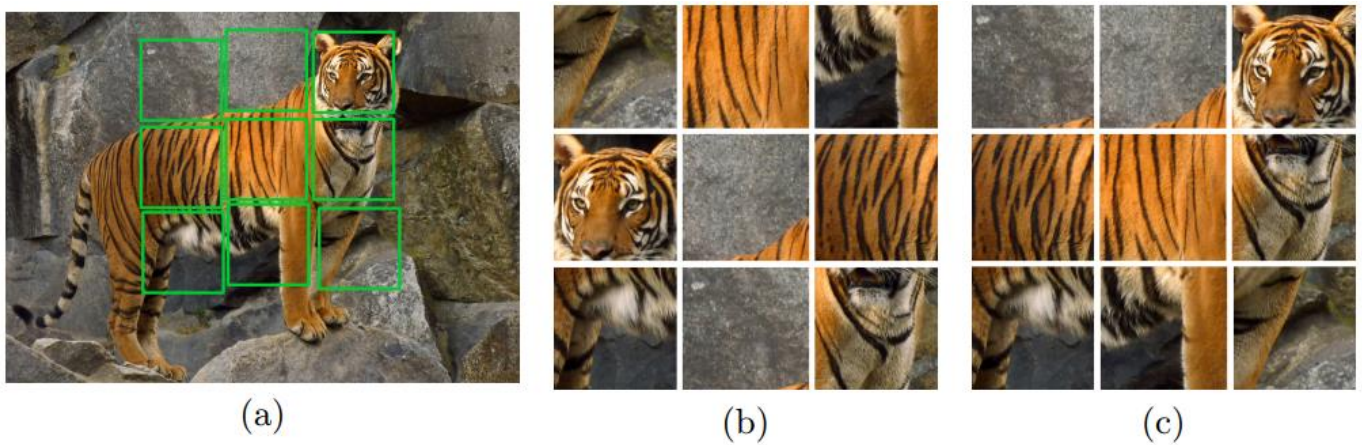
- Visual task를 위해 라벨링 된 데이터를 만드는 것은 비용이 많이 들기 때문에 라벨링이 되지 않은 데이터에 대한 **unsupervised leaning** 방법론이 각광 받고 있음
- 해당 논문에서는 **self-supervised** 방법론을 따라 **pretext task**로 **Jigsaw puzzle**을 풀도록 CNN을 학습하고, 이를 **transfer learning**하여 객체 분류 및 검출 task를 수행함

> **Pretext task:** 문제를 해결하도록 훈련된 네트워크가 다른 downstream task에 쉽게 적용할 수 있는 어떤 시각적 특징을 배우는 단계. 여기서 downstream task란 최종적으로 내가 해결하고자 하는 task.

Research Purpose

❖ **본 논문에서는 Jigsaw puzzle reassembly problem을 pretext task로 사용**

- Jigsaw puzzle을 풀면서 **이미지 표현**을 배우도록 함
- (a) 에서 타일을 만든 후 (b) 같이 섞어 퍼즐을 만들고, 네트워크가 이를 (c) 처럼 맞추도록 학습
- 이를 통해 객체가 어떤 요소로 만들어지고, 그 요소가 무엇인지 네트워크에게 가르칠 수 있음

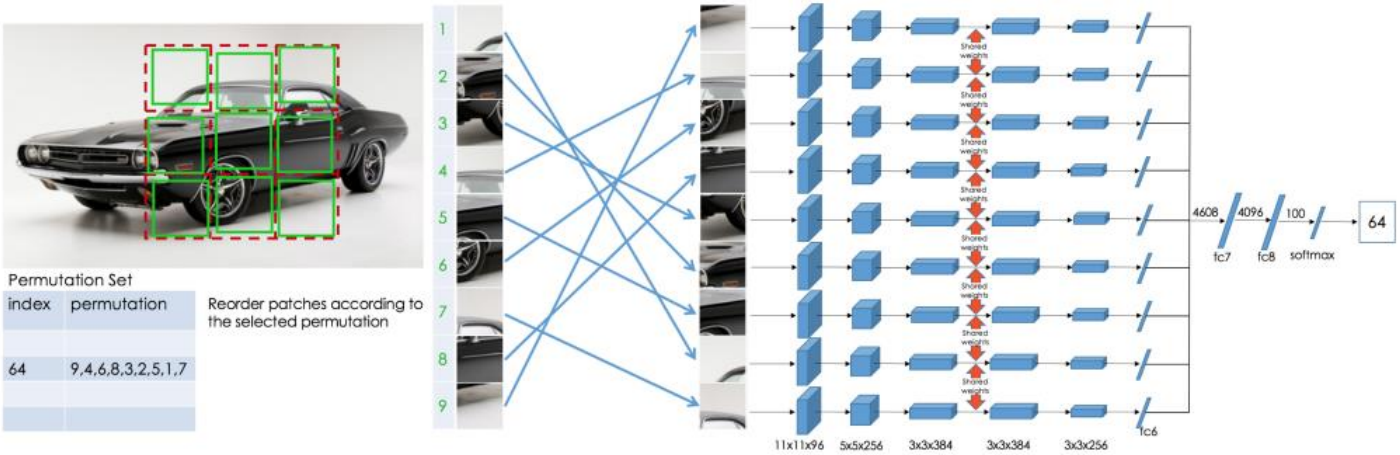


(a) (b) (c)

Proposed Method

❖ **CFN (Context Free Network)**

- 전체 이미지에서 225x225 크기로 random crop하고, 이를 3x3 grid로 분할한 후, 크기가 75x75인 각 grid에서 64x64 크기로 random crop하여 총 9개의 타일을 만듦
- 이 타일들을 사전에 정의한 permutation set으로부터 무작위로 선정한 permutation에 따라 정렬한 후 네트워크에 입력으로 넣어 줌



Proposed Method

❖ CFN (Context Free Network)

- 네트워크의 입력으로 9개의 타일을 채널에 따라 쌓아 사용할 수도 있지만(즉, 입력 데이터는 9x3=27 채널을 갖게 됨), 이 경우 타일간의 **저차원 특징**만 사용하는 문제가 발생
- 따라서 처음에는 각 타일들의 특징을 뽑아내고, 이를 이용해 각 부분의 배열을 맞춤
- 이 과정의 목표는 네트워크가 각 타일들의 **상대적인 위치를 결정**하기 위해 **각 부분의 대표적이고 차별적인 특징**을 학습하도록 하는 것

> **저차원 특징:** 유사한 구조적인 패턴이나 타일 경계면에서의 질감 등을 의미. 이는 Jigsaw puzzle을 푸는 데는 유용하지만 이를 통해 Jigsaw puzzle을 풀게 되면 객체 전체에 대한 이해가 필요하지 않게 됨.

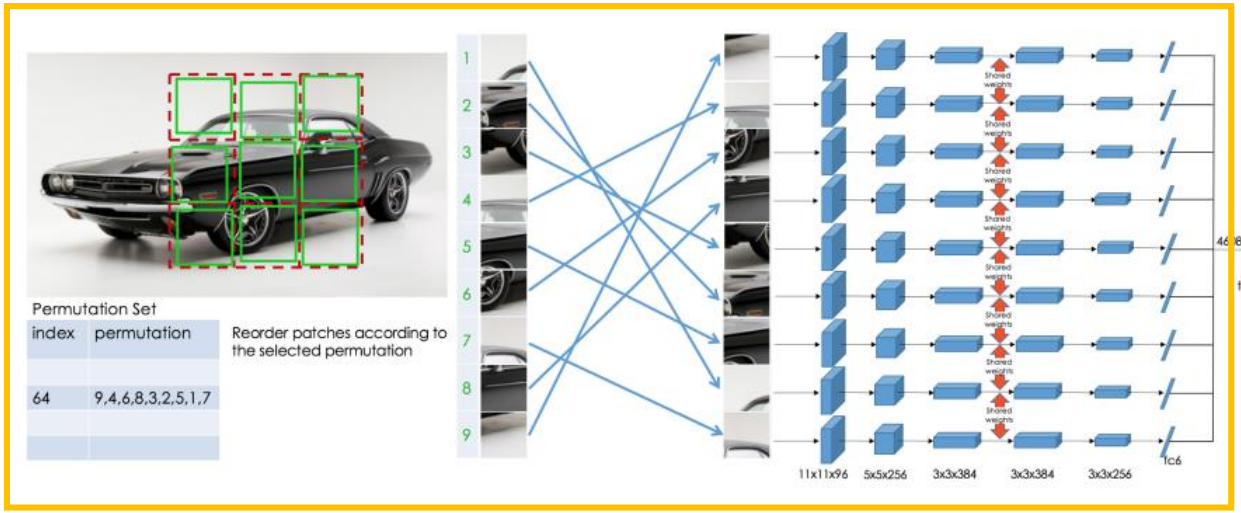
Proposed Method

❖ **CFN (Context Free Network)**

- fc6까지는 동일한 가중치를 공유하는 AlexNet 아키텍처를 사용
- 이 구간에서는 모든 타일이 따로 forward 되기 때문에 연산 과정에서 서로 관여하지 않음

→ Context Free

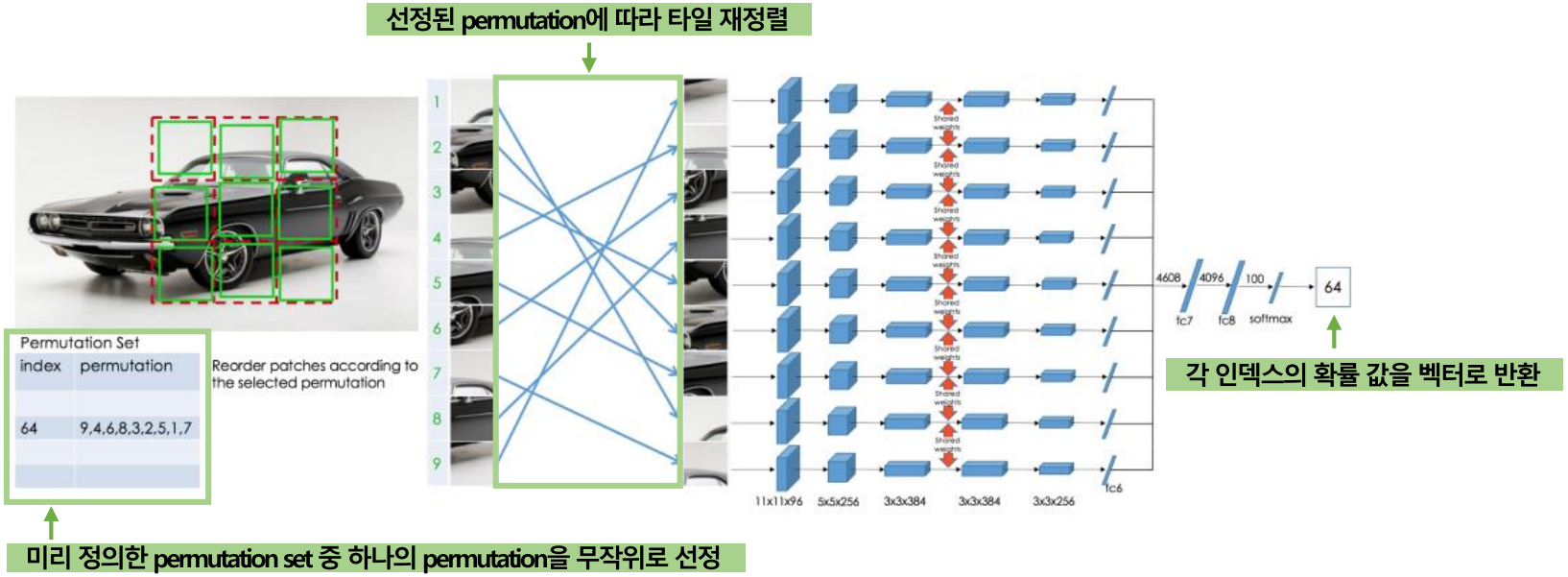
AlexNet 아키텍처 사용 → Context Free



Proposed Method

❖ The Jigsaw Puzzle Task

- 사전에 정의한 permutation set 중 하나의 permutation을 무작위로 선택하여, 그 permutation에 따라 9개의 타일을 재정렬하여 네트워크에 입력으로 넣어 줌
- 그리고 CFN은 각 인덱스의 확률 값을 벡터로 반환하도록 학습
- 9개의 타일이 가질 수 있는 순서의 경우의 수는 9! =362,880로 모든 경우를 따지기 어렵고, 후속 연구를 통해 이를 적절히 제한하는 것이 성능에 좋은 영향을 준다는 것을 확인



Proposed Method

❖ Pretext task 해결을 위한 shortcut 학습 방지

- 이미지의 고차원 특징들을 학습하는 대신 Jigsaw puzzle을 풀기 위한 shortcut만 학습할 수 있음
- 이를 해결하기 위한 여러 방법들을 사용
  - ✓ 이미지가 절대적인 위치를 학습하는 것을 피하기 위해 이미지 당 가능한 1,000개의 puzzle 구성 중 평균적으로 69개의 구성을 사용하였고, 평균 hamming distance가 충분히 큰 구성들을 선택함

    비슷한 permutation이 적으므로 타일이 비슷한 순서로 나열되는 경우 감소
  - ✓ Pixelation/color distribution, edge의 연속성을 피하기 위해, 각 타일 당 평균적으로 11 pixel 씩 간격을 두고 crop함
  - ✓ chromatic aberration을 피하기 위해 grayscale 이미지를 이용한 color channel jittering 이용함

> **hamming distance:** 집합 사이의 거리를 나타내는 지표 중 하나. 예를 들어 {1, 2, 3, 4}와 {1, 3, 2, 4}는 두 번째와 세 번째 요소가 다르기 때문에 둘 사이의 hamming distance는 2임.

> **chromatic aberration:** 렌즈가 모든 색상을 동일한 지점에 초점을 맞추지 못하는 현상으로 각 색상 채널에서 올바르게 위치하지 못하고 조금씩 밀려 있게 됨.

Experiment

❖ Transfer Learning

- Self-supervised task로 CFN을 학습한 후 CFN 가중치를 사용해 standard AlexNet 네트워크의 컨볼루션 레이어를 초기화
- 그런 다음 실험 데이터셋을 사용해 downstream task를 위한 fine-tunning을 진행함
- Unsupervised learning 방법 중에서는 가장 좋은 성능을 보였고, supervised learning 방법의 모델보다는 성능이 떨어지지만 그 간격을 좁힘

Table 1: Results on PASCAL VOC 2007 Detection and Classification. The results of the other methods are taken from Pathak *et al.* [30].

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%**

Experiment

❖ Ablation Study

- Permutation Set
  - ✓ **Cardinality:** Permutation set의 수가 많을 수록 Jigsaw puzzle task를 위한 학습은 더 어려워지지만 transfer learning task에 대한 성능은 더 좋아짐
  - ✓ **Average Hamming distance:** hamming distance가 클수록 Jigsaw puzzle의 정확도와 transfer learning의 성능이 좋아짐
  - ✓ **Minimum hamming distance:** 서로 유사한 순서는 Jigsaw puzzle task를 어렵게 만들기 때문에 permutation들 간 최소 거리를 조정하며 비교

Table 4: Ablation study on the impact of the permutation set.

Number of permutations	Average hamming distance	Minimum hamming distance	Jigsaw task accuracy	Detection performance
1000	8.00	2	71	**53.2**
1000	6.35	2	62	51.3
1000	3.99	2	54	50.2
100	8.08	2	88	52.6
95	8.08	3	90	52.4
85	8.07	4	91	52.7
71	8.07	5	92	52.8
35	8.13	6	94	52.6
10	8.57	7	97	49.2
7	8.95	8	98	49.6
6	9	9	99	49.7

Experiment

❖ Ablation Study

- Preventing Shortcuts
  - ✓ **Low Level Statistics:** 근접한 타일끼리는 pixel의 평균과 표준편차가 비슷할 것이므로 각 타일마다 nomalize 해 줌
  - ✓ **Edge continuity:** 85x85 픽셀 셀들에서 64x64 픽셀 타일들로 crop하여 타일 간 21pixel 간격이 생기도록 함
  - ✓ **Chromatic Aberration:**
    1. 원본 이미지의 가운데를 잘라서 255x255로 resize
    2. 컬러 이미지와 grayscale이미지를 적절히 섞어 학습에 이용
    3. 각 타일의 컬러 채널을 ±0, ±1, ±2 pixels 씩 무작위로 jittering

Table 5: Ablation study on the impact of the shortcuts.

Gap	Normalization	Color jittering	Jigsaw task accuracy	Detection performance
✗	✓	✓	98	47.7
✓	✗	✓	90	43.5
✓	✓	✗	89	51.1
✓	✓	✓	88	52.6

Conclusion

❖ Conclusion

- 본 논문은 Jigsaw puzzle reassembly task를 detection/classification task로 쉽게 transfer 할 수 있는 CFN을 도입함
- CFN은 Jigsaw puzzle를 풀면서 각 타일을 객체의 부분으로 식별하고, 각 부분이 하나의 객체로 조합되는 방법을 학습함
- 이를 통해 CFN은 이미지의 유의미한 특징들을 학습할 수 있었고, downstream task에서도 좋은 성능을 보임
- 기존 unsupervised learning 방법론들보다 좋은 성능을 보였으며, supervised learning 방법론에도 근접한 성능을 보임


Thank You