Visual Representation 2

Pattern Recognition & Machine Learning Laboratory
Tae-jin Woo
Aug 11, 2021

Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles [M. Noroozi et al., 2016] (1/7)

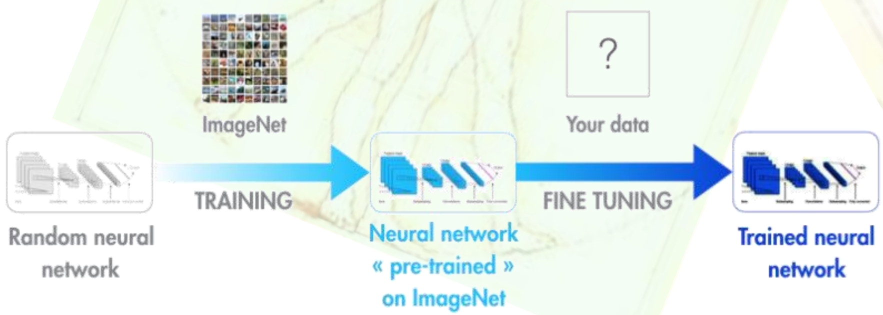
- **Goal**
  - Providing image representation learning without human annotation
  - Achieving encouraging performance comparable to supervised learning
- **Motivation**
  - Learning features of object parts and their correct spatial arrangement
    - By training a network to solve pretext task
  - Obtained features can be transferred to classification and detections tasks
- **Contribution**
  - Achieving State-of-the-Art (SOTA) in self-supervised learning method
  - Building a CNN that can be trained to solve jigsaw puzzles as a pretext task
  - Introduced Context-Free Network (CFN) to maintain the compatibility
    - CFN has fewer parameters than AlexNet

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%**

Results on PASCAL VOC 2007 detection and classification

Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles [M. Noroozi et al., 2016] (2/7)

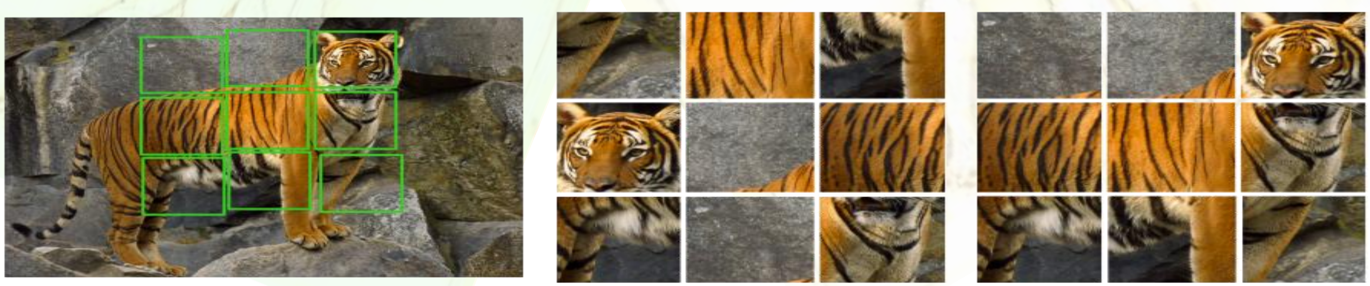
- **Self-supervised learning**
  - **Concept**
    - **Learning features of data through pretext task with unlabeled data**
      - **Learning supervision itself**
    - **Progress transfer learning of pre-trained model for downstream task**
      - **Both freezing pre-trained weights and fine-tuning are possible**
      - **Fewer labeled data would be used for transfer learning**
  - **Pros and cons**
    - **Pros**
      - **Enable learning with unlabeled data**
      - **Possible to get general features before fine-tuning of several downstream tasks**
    - **Cons**
      - **Lower performance than supervised learning in computer vision field**



Example of self-supervised learning 1

Example of self-supervised learning 2

- **Pretext task**
  - **Concept**
    - **Pre-designed problems for networks to solve**
      - **Visual features are learned through pretext task**
      - **Jigsaw puzzle reassembly problem is introduced in this paper**
    - **Only for efficient feature extracting applied to downstream tasks**
  - **Jigsaw puzzle**
    - **Solving the puzzle requires a good understanding of object features**
    - **Representative and distinguishable features of object part will be learnable**
    - **How to solve**
      - **(a) Image from which the tiles (marked with green lines) are extracted**
      - **(b) A puzzle obtained by shuffling the tiles**
      - **(c) Reassemble and determine the relative positions**



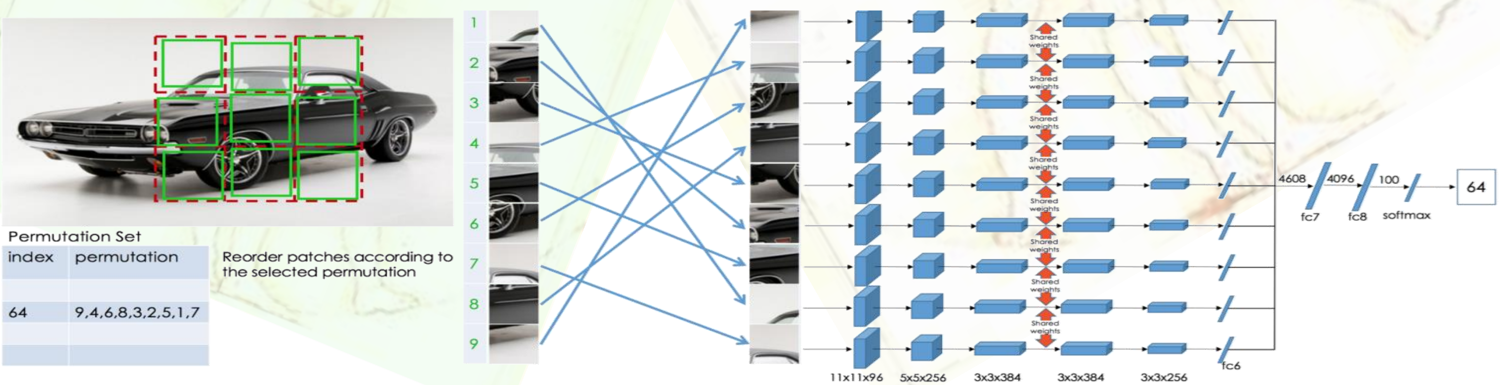
(a) (b) (c)

Learning image representations by solving Jigsaw puzzles

M. Noroozi, P. Fabaro, "Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles", *ECCV,* 2016.

Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles [M. Noroozi et al., 2016] (4/7)

- **Learning method**
  - **Architecture**
    - **Shuffling the order of each tile and use it as input to the CFN**
      - Learning through average 69 permutation set each input image
    - **Features are extracted from the input image first and the order is set last**
      - To solve the problem of learn low-dimensional features between tiles
        - » Low-dimensional features mean similar structural patterns or textures
    - **Building a siamese-ennead convolutional network**
      - Weights of convolutional network are shared up to $fc6$ layer
      - CFN architecture is more compact than AlexNet
        - » $fc6$ layer of CFN includes $18M$ parameters,
          while $fc6$ layer of AlexNet includes $37.5M$ parameters



Context-Free Network

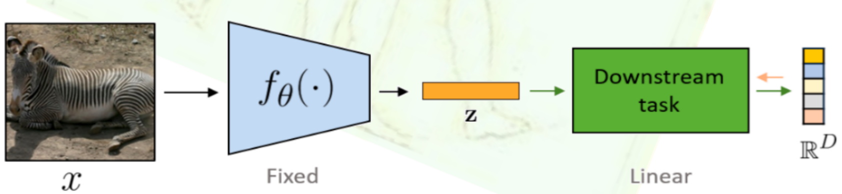
Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles [M. Noroozi et al., 2016] (5/7)

Training

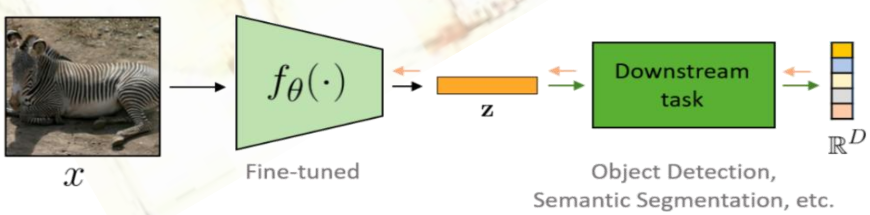
- **Output**
  - CFN can be seen as the conditional pdf
  - $p(S|A_1, A_2, \dots, A_9) = p(S|F_1, F_2, \dots, F_9) \prod_{i=1}^{9} p(F_i|A_i)$
    - S is the configuration of the tiles
    - A_i is the $i-th$ part appearance of the object
    - F_i is the intermediate feature representation
  - $p(L_1, L_2, \dots, L_9|F_1, F_2, \dots, F_9) = \prod_{i=1}^{9} p(L_i|F_i)$
    - If S can be as a list of tile positions $S = (L_1, L_2, \dots, L_9)$
    - CFN learns only spatial arrangement if S is a single per image
  - Learning is making F_i become a meaningful feature

Transfer learning

- **Freezing pre-trained weights**
  - Ability to evaluate the performance of feature extraction
- **Fine-tuning pre-trained weights**
  - Ability to conduct downstream task



Transfer learning with fixed pre-trained weights



Transfer learning with fine-tuning pre-trained weights

Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles [M. Noroozi et al., 2016] (6/7)

- **Experiments**
  - Transfer learning
    - Fine-tuning pre-trained features by using AlexNet on PASCAL VOC 2007
      - Initialized all the $conv$ layers with CFN weights of a standard AlexNet
      - Retrained the rest of the network with Gaussian noise as initial weights
    - Performance evaluation
      - Outperformed all other unsupervised methods
      - Closing the gap with features obtained with supervision
  - ImageNet classification
    - Finding a layer extracting features of the network
      - Method: Fix parameters of a specific network and retrain
    - Checking result
      - $conv5$ layer starts to be specialized on the pretext task
        - » Significant improvement when the $conv5$ layer is also trained

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%**

Results on PASCAL VOC 2007 detection and classification

	conv1	conv2	conv3	conv4	conv5
CFN	**54.7**	**52.8**	**49.7**	45.3	**34.6**
Doersch *et al.* [10]	53.1	47.6	48.7	**45.6**	30.4
Wang and Gupta [39]	51.8	46.9	42.8	38.8	29.8
Random	48.5	41.0	34.8	27.1	12.0

Comparison of classification results on ImageNet 2012

M. Noroozi, P. Fabaro, "Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles", *ECCV*, 2016.

Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles [M. Noroozi et al., 2016] (6/7)

- **Ablation studies**
  - Permutation set
    - Cardinality
      - Performance of the downstream task increased as the permutation set increased
    - Average hamming distance
      - The higher distance, the higher the performance of the downstream task
  - Preventing shortcuts
    - Low level statistics
      - Solution: Normalized pixel mean and standard deviation independently
    - Edge continuity
      - Solution: Making 21 pixel gap between tiles by selecting tiles randomly
    - Chromatic aberration
      - Solution: Use resize, 30% of greyscale input images, and color jittering

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

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

Results on PASCAL VOC 2007 detection and classification