Visual Representaion 1

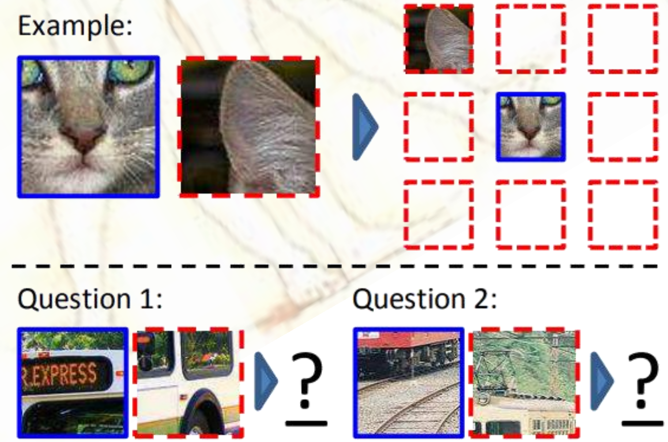
Pattern Recognition & Machine Learning Laboratory

Geon-jun Yang

Aug. 10th, 2021

Unsupervised Visual Representation Learning by Context Prediction [C. Doersch et al., 2015] (1/4)

- **Introduction**
  - **Unsupervised learning**
    - **Limitation of supervised learning**
      - Human annotation required
  - **Self-supervised learning**
    - **Text domain**
      - Context : powerful source of automatic supervision
      - Corpus ⟶ Feature vector ⟶ Predict words
      - Convert unsupervised problem into self-supervised one
  - **Self-supervised learning for image**
    - **Process**
      - Sample random pairs of patches
      - Provide two patches to network
      - Train to guess the position of the patches
    - **Contribution**
      - Good for object detection & unsupervised object discovery / visual data mining
      - Generalizes across images
      - Instance-level supervision

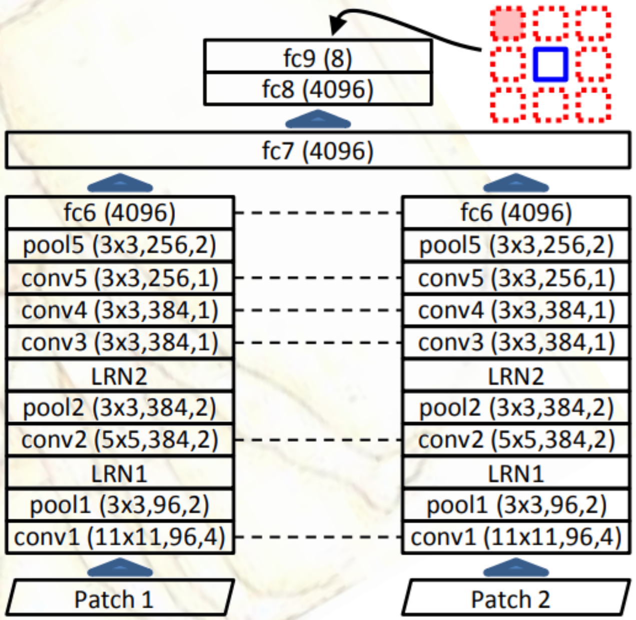


Types of object detection

C. Doersch, A. Gupta, and A. Efros, “Unsupervised Visual Representation Learning by Context Prediction ,” *ICCV*, 2015.

Unsupervised Visual Representation Learning by Context Prediction [C. Doersch et al., 2015] (2/4)

- **Learning visual context prediction**
  - **Architecture**
    - **Late-fusion architecture**
      - A pair of conv net that process separately
      - Must predict relative position of patches
      - Feed two input patches through conv layers
      - Produce output that assigns a probability
      - Feature embedding for individual patches
      - Semantic reasoning

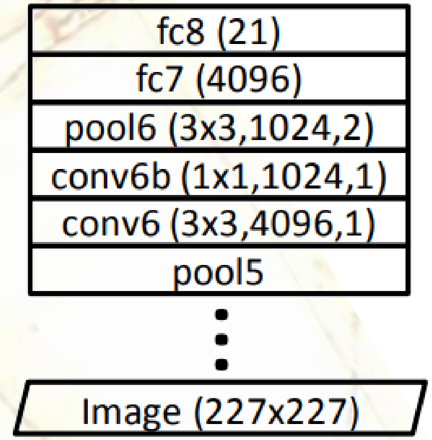


Architecture for pair classification

- **Avoiding trivial solutions**
  - **Extract the desired information**
    - **Use high-level semantic not texture or boundary**
    - **Include gap between patches**
    - **Randomly jitter each patch location**
  - **Chromatic aberration**
    - **Lens focuses light at different wavelengths**
    - **Conv net can learn to localize a patch relative to the lens itself**
      - Detecting the separation between green and magenta
      - Projection
      - Color dropping

Unsupervised Visual Representation Learning by Context Prediction [C. Doersch et al., 2015] (3/4)

- **Experiments**
  - **Nearest neighbors**
    - **Use normalized correlation**
    - **Repeat the experiment using fc7 & fc6**
      - **Fc7: feature from AlexNet trained on ImageNet**
      - **Fc6: feature from authors' architecture without training**
    - **In a few cases, untrained ConvNet does reasonably well**
  - **Object detection**
    - **None of unsupervised pre-training provide such a performance boost**
    - **Adopt R-CNN pipeline**
    - **Use only one stack**
    - **Resize the conv layer 227x227**
    - **Reduce dimensionality to 1024**
    - **5% better than training from scratch**
    - **8% below label supervision**



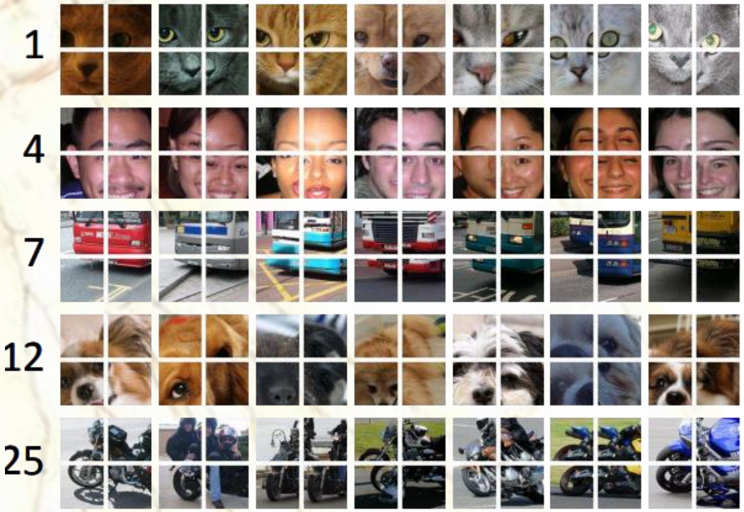
Architecture for Pascal VOC detection

Results on VOC-2007

VOC-2007 Test	person	plant	sheep	sofa	train	tv	mAP
DPM-v5[17]	43.2	12.0	21.1	36.1	46.0	43.5	33.7
[8] w/o context	29.9	20.0	41.1	36.4	48.6	53.2	38.5
Regionlets[55]	43.4	16.4	36.6	37.7	**59.4**	52.3	41.7
Scratch-R-CNN[2]	47.5	28.0	42.3	28.6	51.2	50.0	40.7
Scratch-Ours	46.5	25.6	42.4	23.5	50.0	50.6	39.8
Ours-projection	49.4	**29.0**	**47.5**	28.4	54.7	56.8	45.7
Ours-color-dropping	**50.0**	28.1	46.7	**42.6**	54.8	**58.6**	**46.3**
Ours-Yahoo100m	48.7	28.4	45.1	33.6	49.0	55.5	44.2
Ours-VGG	54.1	26.1	43.9	55.9	69.8	50.9	53.0
ImageNet-R-CNN[19]	54.2	31.5	52.8	48.9	57.9	64.7	54.2

Unsupervised Visual Representation Learning by Context Prediction [C. Doersch et al., 2015] (4/4)

- **Visual data mining**
  - **Definition**
    - **Collect images that depict the same semantic objects**
    - **Dataset visualization, image search**
    - **Connect visual data to unstructured data**
  - **Method**
    - **Sample four adjacent patches from an image**
    - **Find the top 100 images**
    - **Use geometric verification**
    - **Rank the different constellations**
- **Accuracy on the relative prediction task**
  - **Improve the representation**
    - **Analyze classification performance on pretext task**
      - **Sample 500 random images from Pascal VOC**
    - **Accuracy of 38.4%**
    - **Pretext task is difficult**
      - **Large fraction of patches within each image**
      - **The task is almost impossible**



Object cluster discovered by algorithm