A Survey to Self-Supervised Learning

Naiyan Wang
Paradigm of Learning

• Supervised Learning & Unsupervised Learning
  • Given desired output vs. No guidance at all

http://oliviaklose.azurewebsites.net/content/images/2015/02/2-supervised-vs-unsupervised-1.png
Paradigm of Learning

• In Between…
  • Semi-Supervised Learning
    • Mix labeled and unlabeled data

Paradigm of Learning

• In Between...
  • Weakly-Supervised Learning
    • Use somewhat coarse or inaccurate supervision, e.g.
      • Given image level label, infer object level bounding box/ pixel level segmentation
      • Given video level label, infer image level label
      • Given scribble, infer the full pixel level segmentation
      • Given bounding box, infer the boundary of object

Paradigm of Learning

• In Between...
  • Transfer Learning
    • Train on one problem, but test on a different but related problem, e.g.
      • Multi-Task learning
      • Train on one domain, test on another domain (possibly unlabeled)

Paradigm of Learning

• More to mention...
  • Reinforcement Learning
  • Active Learning
  • Zero/One/Few-Shot Learning
Self-Supervised (Feature) Learning

• **What** is it?
  • Use naturally existed supervision signals for training.
  • (Almost) no human intervention

• **Why** do we need it?
  • The age of “representation learning”! (Pre-training – Fine-tune pipeline)
  • Self-Supervised learning can leverage self-labels for representation learning.

• **How** can we realize it?
  • That is in this talk!
Why not use construction?

- What is wrong with autoencoder?
  - Use pixel-wise loss, no structural loss incorporated
  - Reconstruction can hardly represent semantic information
- GAN may alleviate the first issue (e.g. BiGAN)
Outline

• Context
• Video
• Cross-Modality
• Exemplar Learning
Context

• Context is ubiquitous in CV/NLP
  • 管中窥豹 & 断章取义
  • Cat or hair?
  • Beyond using it to improve performance, can you use it as supervision directly?
Context

• Word2Vec: 1-dim context in NLP

https://deeplearning4j.org/img/word2vec_diagrams.png
Context

- Solving the Jigsaw
  - Predict relative positions of patches
  - You have to understand the object to solve this problem!
  - Be aware of trivial solution! CNN is especially good at it

Context

• Solving the Jigsaw
  • Use stronger supervision, solve the real jigsaw problem
  • Harder problem, better performance

Context

- Solving the Jigsaw
  - Visualization of filters

Context

• Why not directly predict the missing parts?
  • With the advancement of adversarial loss

Context

• Colorization
  • You have to know what the object is before you predict its color
  • E.g. Apple is red/green, sky is blue, etc.

Context

• Colorization
  • It is important how to interpret your work!
  • Example colorization of Ansel Adams’s B&W photos

Context

• Colorization
  • Stronger supervision, cross-supervision of different parts of data

Zhang, R., Isola, P., & Efros, A. A. Split-Brain Autoencoders: Unsupervised Learning by Cross-Channel Prediction. *In CVPR 2017*
Video

- Video can provide rich information
  - Temporal continuity
  - Motion consistency
  - Action order
Video

• Slow feature
  • Neighborhood frames should have similar features

\[ \mathcal{U}_2 = \{ \langle (j, k), p_{jk} \rangle : x_j, x_k \in \mathcal{U} \text{ and } p_{jk} = \mathbb{1}(0 \leq j - k \leq T) \}, \]

\[ R_2(\theta, \mathcal{U}) = \sum_{(j, k) \in \mathcal{U}_2} D_\delta(z_\theta(x_j), z_\theta(x_k), p_{jk}) \]

\[ = \sum_{(j, k) \in \mathcal{U}_2} p_{jk} d(z_\theta_j, z_\theta_k) + \overline{p_{jk}} \max(\delta - d(z_\theta_j, z_\theta_k), 0), \]


Video

• Slow and steady feature
  • Not only similar, but also smooth
  • Extend to triplet setting (Not triplet loss!)

\[ \mathcal{U}_3 = \{((l,m,n), p_{lmn}) : x_l, x_m, x_n \in \mathcal{U} \text{ and } p_{lmn} = 1 (0 \leq m - l = n - m \leq T) \}. \]

\[ R_3(\theta, \mathcal{U}) = \sum_{(l,m,n) \in \mathcal{U}_3} D_\delta(z_{\theta l} - z_{\theta m}, z_{\theta m} - z_{\theta n}, p_{lmn}), \]

Video

• Find corresponding pairs using visual tracking

Video

• Directly predict motion
  • Motion is not predictable by its nature
  • The ultimate goal is not to predict instance motion, but to learn common motion of visually similar objects

Walker, J., Gupta, A., & Hebert, M. Dense optical flow prediction from a static image. In ICCV 2015
Video

• Similar pose should have similar motion
  • Learning appearance transformation

• Is the temporal order of a video correct?
  • Encode the cause and effect of action

Video

• Is the temporal order of a video correct?
  • Find the odd sequence

Video

- Multi-view
  - Same action, but different view
  - View and pose invariant features

Video

• The world is rigid, or at least piecewise rigid
  • Motion provide evidence of how pixels move together
  • The pixels move together are likely to form an object

Cross-Modality

• In some applications, it is easy to collect and align the data from several modalities
  • Lidar & GPS/IMU & Camera
  • RGB & D
  • Image & Text

• How to utilize them for cross-supervision?
Cross-Modality

• Ego-motion
  • “We move in order to see and we see in order to move” - J.J Gibson
  • Ego-motion data is easy to collect
  • Siamese CNN to predict camera translation & Rotation along 3-axes. (Visual Odometry)

Agrawal, P., Carreira, J., & Malik, J. Learning to see by moving. In ICCV 2015
Cross-Modality

• Ego-motion
  • Learning features that are equivariant to ego-motion

Jayaraman, D., & Grauman, K. Learning image representations tied to ego-motion. In ICCV 2015
Cross-Modality

• Ego-motion
  • Siamese networks with contrastive loss
  • $M_g$ is the transformation matrix specified by the external sensors

\[
(\theta^*, M^*) = \arg \min_{\theta, M} \sum_{g, i, j} d_g(M_g z_\theta(x_i), z_\theta(x_j), p_{ij}),
\]

\[
d_g(a, b, c) = \mathbb{1}(c = g)d(a, b) + \mathbb{1}(c \neq g) \max(\delta - d(a, b), 0),
\]

Jayaraman, D., & Grauman, K. Learning image representations tied to ego-motion. In *ICCV 2015*
Cross-Modality

• Acoustics -> RGB
  • Similar events should have similar sound.
  • Naturally cluster the videos.

Cross-Modality

• Acoustics -> RGB
  • What does this CNN learn? Separation of baby and person :-D

Cross-Modality

- Features for road segmentation (Depth -> RGB)

Weiyue W., Naiyan W., Xiaomin W., Suya Y. and Ulrich N. Self-Paced Cross-Modality Transfer Learning for Efficient Road Segmentation. In ICRA2017
Cross-Modality

• Features for grasping
  • Verify whether we could grasp the center of a patch at a given angle

Pinto, L., & Gupta, A. Supersizing self-supervision: Learning to grasp from 50k tries and 700 robot hours. In ICRA 2016
Exemplar Learning

• Learning instance features
  • Each data sample as one class
  • Need strong augmentation

Exemplar Learning

• Learning instance features
  • The key is to avoid trivial solution. (Several tricks in this paper)
  • Project each sample on a random target uniformly samples on a unit ball

Evaluation

- Evaluate on general high-level vision tasks (classification, detection)
- Be cautious of different settings!

<table>
<thead>
<tr>
<th>Method</th>
<th>All</th>
<th>&gt;c1</th>
<th>&gt;c2</th>
<th>&gt;c3</th>
<th>&gt;c4</th>
<th>&gt;c5</th>
<th>All</th>
<th>&gt;c1</th>
<th>&gt;c2</th>
<th>&gt;c3</th>
<th>&gt;c4</th>
<th>&gt;c5</th>
<th>#wins</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supervised</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Imagenet</td>
<td>56.5</td>
<td>57.0</td>
<td>57.1</td>
<td>57.1</td>
<td>55.6</td>
<td>52.5</td>
<td>17.7</td>
<td>19.1</td>
<td>19.7</td>
<td>20.3</td>
<td>20.9</td>
<td>19.6</td>
<td>NA</td>
</tr>
<tr>
<td>Sup. Masks (Ours)</td>
<td>51.7</td>
<td>51.8</td>
<td>52.7</td>
<td>52.2</td>
<td>52.0</td>
<td>47.5</td>
<td>13.6</td>
<td>13.8</td>
<td>15.5</td>
<td>17.6</td>
<td>18.1</td>
<td>15.1</td>
<td>NA</td>
</tr>
<tr>
<td>Unsupervised</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jigsaw† [30]</td>
<td>49.0</td>
<td>50.0</td>
<td>48.9</td>
<td>47.7</td>
<td>45.8</td>
<td>37.1</td>
<td>5.9</td>
<td>8.7</td>
<td>8.8</td>
<td>10.1</td>
<td>9.9</td>
<td>7.9</td>
<td>NA</td>
</tr>
<tr>
<td>Kmeans [23]</td>
<td>42.8</td>
<td>42.2</td>
<td>40.3</td>
<td>37.1</td>
<td>32.4</td>
<td>26.0</td>
<td>4.1</td>
<td>4.9</td>
<td>5.0</td>
<td>4.5</td>
<td>4.2</td>
<td>4.0</td>
<td>0</td>
</tr>
<tr>
<td>Egomotion [2]</td>
<td>37.4</td>
<td>36.9</td>
<td>34.4</td>
<td>28.9</td>
<td>24.1</td>
<td>17.1</td>
<td>4.1</td>
<td>4.9</td>
<td>5.0</td>
<td>4.5</td>
<td>4.2</td>
<td>4.0</td>
<td>0</td>
</tr>
<tr>
<td>Inpainting [35]</td>
<td>39.1</td>
<td>36.4</td>
<td>34.1</td>
<td>29.4</td>
<td>24.8</td>
<td>13.4</td>
<td>4.1</td>
<td>4.9</td>
<td>5.0</td>
<td>4.5</td>
<td>4.2</td>
<td>4.0</td>
<td>0</td>
</tr>
<tr>
<td>Tracking-gray [46]</td>
<td>43.5</td>
<td>44.6</td>
<td>44.6</td>
<td>44.2</td>
<td>41.5</td>
<td>35.7</td>
<td>3.7</td>
<td>5.7</td>
<td>7.4</td>
<td>9.0</td>
<td>9.4</td>
<td>9.0</td>
<td>0</td>
</tr>
<tr>
<td>Sounds [33]</td>
<td>42.9</td>
<td>42.3</td>
<td>40.6</td>
<td>37.1</td>
<td>32.0</td>
<td>26.5</td>
<td>5.4</td>
<td>5.1</td>
<td>5.0</td>
<td>4.8</td>
<td>4.5</td>
<td>4.2</td>
<td>0</td>
</tr>
<tr>
<td>BiGAN [10]</td>
<td>44.9</td>
<td>44.6</td>
<td>44.7</td>
<td>42.4</td>
<td>38.4</td>
<td>29.4</td>
<td>4.9</td>
<td>6.1</td>
<td>7.3</td>
<td>7.6</td>
<td>7.1</td>
<td>4.6</td>
<td>0</td>
</tr>
<tr>
<td>Colorization [51]</td>
<td>44.5</td>
<td>44.9</td>
<td>44.7</td>
<td>44.4</td>
<td>42.6</td>
<td>38.0</td>
<td>6.1</td>
<td>7.9</td>
<td>8.6</td>
<td>10.6</td>
<td>10.7</td>
<td>9.9</td>
<td>0</td>
</tr>
<tr>
<td>Split-Brain Auto [52]</td>
<td>43.8</td>
<td>45.6</td>
<td>45.6</td>
<td>46.1</td>
<td>44.1</td>
<td>37.6</td>
<td>3.5</td>
<td>7.9</td>
<td>9.6</td>
<td>10.2</td>
<td>11.0</td>
<td>10.0</td>
<td>0</td>
</tr>
<tr>
<td>Context [8]</td>
<td>49.9</td>
<td>48.8</td>
<td>44.4</td>
<td>44.3</td>
<td>42.1</td>
<td>33.2</td>
<td>6.7</td>
<td>10.2</td>
<td>9.2</td>
<td>9.5</td>
<td>9.4</td>
<td>8.7</td>
<td>3</td>
</tr>
<tr>
<td>Context-videos† [8]</td>
<td>47.8</td>
<td>47.9</td>
<td>46.6</td>
<td>47.2</td>
<td>44.3</td>
<td>33.4</td>
<td>6.6</td>
<td>9.2</td>
<td>10.7</td>
<td>12.2</td>
<td>11.2</td>
<td>9.0</td>
<td>1</td>
</tr>
<tr>
<td>Motion Masks (Ours)</td>
<td>48.6</td>
<td>48.2</td>
<td>48.3</td>
<td>47.0</td>
<td>45.8</td>
<td><strong>40.3</strong></td>
<td><strong>10.2</strong></td>
<td><strong>10.2</strong></td>
<td><strong>11.7</strong></td>
<td><strong>12.5</strong></td>
<td><strong>13.3</strong></td>
<td><strong>11.0</strong></td>
<td><strong>9</strong></td>
</tr>
</tbody>
</table>

Evaluation

• Best so far

<table>
<thead>
<tr>
<th>Initialization</th>
<th>Architecture</th>
<th>Class. %mAP</th>
<th>Seg. %mIU</th>
</tr>
</thead>
<tbody>
<tr>
<td>ImageNet (+FoV)</td>
<td>VGG-16</td>
<td>86.9</td>
<td>69.5</td>
</tr>
<tr>
<td>Random (ours)</td>
<td>AlexNet</td>
<td>46.2</td>
<td>23.5</td>
</tr>
<tr>
<td>Random [31]</td>
<td>AlexNet</td>
<td>53.3</td>
<td>19.8</td>
</tr>
<tr>
<td>k-means [19, 5]</td>
<td>AlexNet</td>
<td>56.6</td>
<td>32.6</td>
</tr>
<tr>
<td>k-means [19]</td>
<td>VGG-16</td>
<td>56.5</td>
<td>-</td>
</tr>
<tr>
<td>k-means [19]</td>
<td>GoogleLeNet</td>
<td>55.0</td>
<td>-</td>
</tr>
<tr>
<td>Pathak et al. [31]</td>
<td>AlexNet</td>
<td>56.5</td>
<td>29.7</td>
</tr>
<tr>
<td>Wang &amp; Gupta [38]</td>
<td>AlexNet</td>
<td>58.7</td>
<td>-</td>
</tr>
<tr>
<td>Donahue et al. [5]</td>
<td>AlexNet</td>
<td>60.1</td>
<td>35.2</td>
</tr>
<tr>
<td>Doelesch et al. [4, 5]</td>
<td>AlexNet</td>
<td>65.3</td>
<td>-</td>
</tr>
<tr>
<td>Zhang et al. (col) [42]</td>
<td>AlexNet</td>
<td>65.6</td>
<td>35.6</td>
</tr>
<tr>
<td>Zhang et al. (s-b) [43]</td>
<td>AlexNet</td>
<td>67.1</td>
<td>36.0</td>
</tr>
<tr>
<td>Noroozi &amp; Favaro [28]</td>
<td>Mod. AlexNet</td>
<td>68.6</td>
<td>-</td>
</tr>
<tr>
<td>Larsson et al. [20]</td>
<td>VGG-16</td>
<td>-</td>
<td>50.2</td>
</tr>
<tr>
<td>Our method</td>
<td>AlexNet</td>
<td>65.9</td>
<td>38.4</td>
</tr>
<tr>
<td>(+FoV) VGG-16</td>
<td></td>
<td>77.2</td>
<td>56.0</td>
</tr>
<tr>
<td>(+FoV) ResNet-152</td>
<td></td>
<td>77.3</td>
<td>60.0</td>
</tr>
</tbody>
</table>

### Action Recognition

<table>
<thead>
<tr>
<th>Method</th>
<th>UCF101-split1</th>
<th>HMDB51-split1</th>
</tr>
</thead>
<tbody>
<tr>
<td>DrLim [17]</td>
<td>45.7</td>
<td>16.3</td>
</tr>
<tr>
<td>TempCoh [32]</td>
<td>45.4</td>
<td>15.9</td>
</tr>
<tr>
<td>Obj. Patch [44]</td>
<td>40.7</td>
<td>15.6</td>
</tr>
<tr>
<td>Seq. Ver. [31]</td>
<td>50.9</td>
<td>19.8</td>
</tr>
<tr>
<td>Our - Stack-of-Diff.</td>
<td><strong>60.3</strong></td>
<td><strong>32.5</strong></td>
</tr>
<tr>
<td>Rand weights - Stack-of-Diff.</td>
<td>51.3</td>
<td>28.3</td>
</tr>
<tr>
<td>ImageNet weights - Stack-of-Diff.</td>
<td>70.1</td>
<td>40.8</td>
</tr>
</tbody>
</table>
Discussion

• How to cross the semantic gap between low-level and high-level?
  • Utilize high-level/global context
  • Explore piece-wise rigidity in real-life
  • More to discover...

• What is a useful self-supervised learning?
  • Improve the performance of subsequent task.
  • Task Related Self-Supervised Learning
Active Research Groups

• Alexei Efros (Berkeley)

• Abhinav Gupta (CMU)

• Martial Hebert (CMU)
Uncovered Papers

- **Colorization:**

- **Optical Flow**

- **Others**
  - Pinto, L., Gandhi, D., Han, Y., Park, Y. L., & Gupta, A. The curious robot: Learning visual representations via physical interactions. In *ECCVW 2016*.