Deep Learning from Temporal Coherence in Video

Hossein Mobahi, Ronan Collobert, Jason Weston
hmobahi2@uiuc.edu, collober@nec-labs.com, jasonw@nec-labs.com

NEC Laboratories America
The Goal

- Object classification using deep architectures
- Convolutional Neural Networks (CNNs)

 Lots of parameters in each layers require lot of training examples

 How can we leverage unlabeled data?
Exploit some structure in the data.
See DrLim (Chopra et al, 2005).
Related to: siamese networks, Laplacian Eigenmap, Isomap, LLE...
Embedding Algorithm: Applications

Language model: Positive pair: (the cat sat on the, mat).
Negative pair: (the cat sat on the, yesterday).
Ranking loss.

Negative pair: random (query, document).
Ranking loss.

Negative pair: random examples.
Euclidean distance.

Video: ?

See Jason Weston’s talk in the “Learning Feature Hierarchies” workshop
Video: Temporal Coherence

- Two consecutive frames likely to contain the same object(s)
- Temporal coherence information helps for learning invariance to pose, illumination, background, deformations...
Leveraging Temporal Coherence

Representation $z^l(\cdot)$ of frames in the $l^{th}$ deep layer
- are pushed **together** for consecutive frames
- are pulled **apart** for two random frames

Corresponds to minimize:

$$L_{coh}(x_1, x_2) = \begin{cases} 
||z^l(x_1) - z^l(x_2)||_1, & \text{if } x_1, x_2 \text{ consecutive} \\
\max(0, m - ||z^l(x_1) - z^l(x_2)||_1), & \text{otherwise}
\end{cases}$$
Algorithm

Given Data...

**Input:** Labeled data \((x_n, y_n), n = 1, \ldots N\),
unlabeled video data \(x_n, n = N + 1, \ldots N + U\)

Minimize...

\[
\frac{1}{N} \sum_{n} L(x_n, y_n) + \frac{1}{NM} \sum_{n,m} L_{coh}(x_m, x_n)
\]

With Stochastic Gradient...

repeat

Pick a random labeled example \((x_n, y_n)\)
Make a gradient step to decrease \(L(x_n, y_n)\)
Pick a random pair of consecutive images \(x_m, x_n\) in the video
Make a gradient step to decrease \(L_{coh}(x_m, x_n)\)
Pick a random pair of images \(x_m, x_n\) in the video
Make a gradient step to decrease \(L_{coh}(x_m, x_n)\)

until Stopping criterion is met
Previous Work: Semi-supervised Learning

Transduction:
- unlabeled must be from same distribution $p(x, y)$
- cluster assumption must be true
- kernel (if any) might be based on bad metric

Graph-based learning:
- k-nn: slow to construct, might be bad metric
- cluster assumption must be true
Bad Metric: Euclidean Distance

Lighting condition
Bad Metric: Euclidean Distance

Pose/Background
Bad Metric: Euclidean Distance

Occlusion
Bad Metric: Euclidean Distance

Temporal coherence defines a *natural* metric in the representation space $z^l(\cdot)$.
Cluster Assumption

No cluster assumption requirement in the original space
Previous Work: Semi-supervised Learning

Transduction:
- ✗ unlabeled must be from same distribution \( p(x, y) \)
- ✗ cluster assumption must be true
- ✗ kernel (if any) might be based on bad metric

Graph-based learning:
- ✗ k-nn: slow to construct, might be bad metric
- ✗ cluster assumption must be true

Learning from video:
- ✓ cluster assumption in representation space (not original space!)
- ✓ natural metric for pairs — Euclidean dist. might say they aren’t close
- ✓ no cost to collect pairs
- ✓ weak assumption on unlabeled distribution
Previous Work: Temporal Coherence

Many methods use video for “learning”...two related ones:

- **Slow Feature Analysis** [Wiskott & Sejnowski, 2002] Learn transformation functions invariant with time, s.t. no trivial solutions.
- In [Becker, 1999] temporal context is learnt with a special network: extra neurons (“contextual gating units”) + a Hebbian update rule for clustering based on context (“competitive learning”)
- **IMAX method** [Becker and Hinton, 1996]: maximizes the mutual information between different output units, applied to learning spatial or temporal coherency. Drawbacks [from authors]: “tendency to become trapped in poor local minima”, “learning is very slow”

Our method:
- **Simple**, highly scalable, easily trained on millions of examples
- We observe improved generalization whenever we applied it...
Experiments: COIL 100 Setup

Following [Wersing, 2003].

Built our own video: COIL-like and Animal Set.

Show video learning improves error rate.

Show video helps even when from different source to task.

1Strongly engineered Neural Net (VTU): builds a hierarchy of biologically inspired feature detectors. It applies Gabor filters at four orientations, followed by spatial pooling, and learns receptive field profiles using a special type of sparse coding algorithm with invariance constraints.
Experiments: Coil 100

100 objects 72x72 pixels,
each of which has 72 different poses (5 degree turns).
4 views for train, 68 for test.
30 or 100 objects for train/test
40 objects, 4 types of objects in COIL100 (fruits, cars, cups, and cans). Each has 72 views, as a video stream.

Collected to provide similar sensory data as in the COIL dataset.
Experiments: Animal Set

60 animals such as horses, ducks, deer and rabbits. 72 views for each animal as a video stream.

Enable us to measure the success when the unlabeled video shares no objects in common with the supervised task.
## Experiments: COIL 100 Performance

<table>
<thead>
<tr>
<th>Method</th>
<th>30 objects</th>
<th>100 objects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nearest Neighbor</td>
<td>81.8</td>
<td>70.1</td>
</tr>
<tr>
<td>SVM</td>
<td>84.9</td>
<td>74.6</td>
</tr>
<tr>
<td>SpinGlass MRF</td>
<td>82.8</td>
<td>69.4</td>
</tr>
<tr>
<td>Eigen Spline</td>
<td>84.6</td>
<td>77.0</td>
</tr>
<tr>
<td>VTU</td>
<td>89.9</td>
<td>79.1</td>
</tr>
<tr>
<td>Standard CNN</td>
<td>84.88</td>
<td>71.49</td>
</tr>
<tr>
<td>videoCNN V:COIL100</td>
<td>-</td>
<td>92.25*</td>
</tr>
<tr>
<td>videoCNN V:COIL &quot;70&quot;</td>
<td>95.03†</td>
<td>-</td>
</tr>
<tr>
<td>videoCNN V:COIL-Like</td>
<td>-</td>
<td>79.77</td>
</tr>
<tr>
<td>videoCNN V:Animal</td>
<td>-</td>
<td>78.67</td>
</tr>
</tbody>
</table>

* Transductive setup with 100 objects
† Semi-supervised setup with 70 objects
Experiments: AT&T’s ORL Face Dataset

10 different gray scale images for each of the 40 distinct subjects.

Varying lighting and facial expressions (open / closed eyes, smiling / not smiling).
# Experiments: Simple ORL Experiment

Test Accuracy with magenta $k$ labeled examples per subject.

<table>
<thead>
<tr>
<th>Method</th>
<th>$k=1$</th>
<th>$k=2$</th>
<th>$k=5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nearest Neighbor</td>
<td>69.07</td>
<td>81.08</td>
<td>94.64</td>
</tr>
<tr>
<td>PCA</td>
<td>56.43</td>
<td>71.19</td>
<td>88.31</td>
</tr>
<tr>
<td>LDA</td>
<td>-</td>
<td>68.84</td>
<td>88.87</td>
</tr>
<tr>
<td>MRF</td>
<td>51.06</td>
<td>68.38</td>
<td>86.95</td>
</tr>
<tr>
<td>Standard CNN</td>
<td>71.83</td>
<td>82.58</td>
<td>94.05</td>
</tr>
<tr>
<td>videoCNN V:ORL</td>
<td>90.35</td>
<td>94.77</td>
<td>98.86</td>
</tr>
</tbody>
</table>

Images placed in a “video” sequence by concatenating 40 segments, one for each subject. Labeled train and test images are part of the video. [WARNING: “transductive” setup]
Conclusion

- Leverage structured data with embedding algorithm.

- Use of video coherence improves internal representation of images: potentially learn invariance to pose, illumination, background or clutter, deformations (e.g. facial expressions) or occlusions.

- Outperforms baselines with no engineered features.

- Weaker assumption than in semi-supervised learning.

- Huge collections of data can be obtained without human annotation.

- General idea: successfully applied to text, document retrieval, semi-supervised learning.