Learning from Web Data and Adapting beyond It

Boqing Gong
BoqingGo@outlook.com
Learning based visual recognition

Annotator

Training images

Car

Car

Not car

Image features

New image

Car!

Courtesy K. Grauman
Learning-based visual recognition

Last 10+ years: impressive strides by learning appearance models (usually discriminative).

1. Web data with **noisy** labels
   ➔ Need different training techniques

Courtesy K. Grauman
Label correction & re-weighting

Label Correction

dog? ✓

dog? ✗

cat:

Re-weigh labels/data terms

0.2?
0.8?
0.3?

cat:

jacket:
Label correction & re-weighting removal

Label Correction

dog? ✓
dog? ✗

cat:

Hard to rectify wrong labels
Easier to just remove wrong labels

Semi-supervised learning?
Caveat: outlier images
A consistent term & its dual effect

[Stochastic augmentation & Gaussian noise]

Outlier still helps!

[Laine & Aila, ICLR 2017]
Noisy labels, no outlier

Result on CIFAR-10 and MNIST

Table 3. Comparison results on CIFAR-10 and MNIST

<table>
<thead>
<tr>
<th>Methods</th>
<th>CIFAR-10 14-layer ResNet</th>
<th>MNIST fully connected</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>p = 0 sy.p = 0.2 asy.p = 0.2 asy.p = 0.6</td>
<td>p = 0 sy.p = 0.2 asy.p = 0.2 asy.p = 0.6</td>
</tr>
<tr>
<td>cross-entropy [37]</td>
<td>87.8 83.7 85.0 57.6</td>
<td>97.9±0.0 96.9±0.1 97.5±0.0 53±0.6</td>
</tr>
<tr>
<td>unhinged (BN) [57]</td>
<td>86.9 84.1 83.8 52.1</td>
<td>97.6±0.0 96.9±0.1 97.0±0.1 71.2±1.0</td>
</tr>
<tr>
<td>sigmoid (BN) [12]</td>
<td>76.0 66.6 71.8 57.0</td>
<td>97.2±0.1 93.1±0.1 96.7±0.1 71.4±1.3</td>
</tr>
<tr>
<td>savage [30]</td>
<td>80.1 77.4 76.0 50.5</td>
<td>97.3±0.0 96.9±0.0 97.0±0.1 51.3±0.4</td>
</tr>
<tr>
<td>bootstrap soft [40]</td>
<td>87.7 84.3 84.6 57.8</td>
<td>97.9±0.0 96.9±0.0 97.5±0.0 53.0±0.4</td>
</tr>
<tr>
<td>bootstrap hard [40]</td>
<td>87.3 83.6 84.7 58.3</td>
<td>97.9±0.0 96.8±0.0 97.4±0.0 55.0±1.3</td>
</tr>
<tr>
<td>backward [37]</td>
<td>87.7 80.4 83.8 66.7</td>
<td>97.9±0.0 96.9±0.0 96.7±0.1 67.4±1.5</td>
</tr>
<tr>
<td>forward [37]</td>
<td>87.4 83.4 87.0 74.8</td>
<td>97.9±0.0 96.9±0.0 97.7±0.0 64.9±4.4</td>
</tr>
<tr>
<td>cross-entropy</td>
<td>87.9 82.4 85.5 56.2</td>
<td>98.0±0.1 97.1±0.1 97.6±0.2 52.9±0.6</td>
</tr>
<tr>
<td>improved baseline</td>
<td>87.8 83.6 85.2 74.1</td>
<td>98.0±0.1 97.1±0.1 97.7±0.1 76.7±1.6</td>
</tr>
<tr>
<td>ours</td>
<td>88.0 84.5 85.6 75.8</td>
<td>98.2±0.1 97.7±0.4 97.8±0.1 83.4±1.3</td>
</tr>
</tbody>
</table>

[Ding et al., WACV’18]
Noisy labels, & outlier images

Results on Clothing1M

<table>
<thead>
<tr>
<th>#</th>
<th>model</th>
<th>loss / method</th>
<th>initialization</th>
<th>training set</th>
<th>accuracy (reported)</th>
<th>accuracy (our impl.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>AlexNet</td>
<td>pseudo-label [25]</td>
<td>#9</td>
<td>1M, 50K</td>
<td>73.04</td>
<td>–</td>
</tr>
<tr>
<td>2</td>
<td>AlexNet</td>
<td>bottom-up [47]</td>
<td>#9</td>
<td>1M, 50K</td>
<td>76.22</td>
<td>–</td>
</tr>
<tr>
<td>3</td>
<td>AlexNet</td>
<td>label noise model [59]</td>
<td>#9</td>
<td>1M, 50K</td>
<td>78.24</td>
<td>–</td>
</tr>
<tr>
<td>4</td>
<td>50-ResNet</td>
<td>cross-entropy</td>
<td>ImageNet</td>
<td>1M</td>
<td>68.94</td>
<td>69.03</td>
</tr>
<tr>
<td>5</td>
<td>50-ResNet</td>
<td>backward [37]</td>
<td>ImageNet</td>
<td>1M</td>
<td>69.13</td>
<td>–</td>
</tr>
<tr>
<td>6</td>
<td>50-ResNet</td>
<td>forward [37]</td>
<td>ImageNet</td>
<td>1M</td>
<td>69.84</td>
<td>–</td>
</tr>
<tr>
<td>7</td>
<td>50-ResNet</td>
<td>ours</td>
<td>ImageNet</td>
<td>1M</td>
<td>–</td>
<td>77.34</td>
</tr>
<tr>
<td>8</td>
<td>50-ResNet</td>
<td>ours</td>
<td>ImageNet</td>
<td>1M, 50K</td>
<td>–</td>
<td>79.38</td>
</tr>
<tr>
<td>9</td>
<td>AlexNet</td>
<td>cross-entropy</td>
<td>ImageNet</td>
<td>50K</td>
<td>72.63</td>
<td>–</td>
</tr>
<tr>
<td>10</td>
<td>50-ResNet</td>
<td>cross-entropy</td>
<td>ImageNet</td>
<td>50K</td>
<td>75.19</td>
<td>74.84</td>
</tr>
<tr>
<td>11</td>
<td>50-ResNet</td>
<td>cross-entropy</td>
<td>#6</td>
<td>50K</td>
<td>80.38</td>
<td>–</td>
</tr>
<tr>
<td>12</td>
<td>50-ResNet</td>
<td>cross-entropy</td>
<td>#7</td>
<td>50K</td>
<td>–</td>
<td>80.44</td>
</tr>
<tr>
<td>13</td>
<td>50-ResNet</td>
<td>cross-entropy</td>
<td>#8</td>
<td>50K</td>
<td>–</td>
<td>80.53</td>
</tr>
</tbody>
</table>

[Ding et al., WACV’18]
Augment the same example twice

\[ \frac{d_Y(f(x_1), f(x_2))}{d_X(x_1, x_2)} \leq K. \]

Two data points around that example

\[ \text{Lipschitz continuity in Wasserstein GAN} \]

Detour: a consistent term & its dual effect

\[ \gamma_l \]

x_l / x_u

stochastic augmentation

dropout and Gaussian noise

cross-entropy

squared difference

sum

loss

[Xiang*, Gong*, et al., ICLR 2018]
Outline

Web data with noisy labels
Hard to rectify wrong labels
Easier to just remove wrong labels

Web data with accurate labels
3D movies

Web data of multi-modalities
Web images vs. Web videos

Semi-supervised learning
3D movies
Geometry & semantics

Shape from dense views
geometric problem

Shape from one view
semantic problem

[Snavely et al, CVPR ‘06]

[Sinha et al, ICCV’93]

Courtesy K. Grauman & D. Jayaraman
3D movies

Start with synthetic imagery and precise geometry cues

Followed by 3D movies to incorporate reality cues
Results on UCF101

It is important to follow the right curriculum!

[52.4, 52.6, 52.9, 54.1]

[Gan et al., CVPR’18]
Detour: curriculum learning

Feed a learning system “easy” examples first
Gradually introduce more difficult ones

[Bengio et al., ICML’09]
Detour: curriculum domain adaptation

Feed a learning system “easy” tasks first. Their solutions find better local optima, and act as a regularizer, i.e., focusing on the test set.
Detour: curriculum domain adaptation

Feed a learning system “easy” tasks first. Their solutions find better local optima, and act as a regularizer, i.e., focusing on the test.

Input: An urban scene image
Algorithm: Super-pixel + Logistic regression
Output: Labels of some super-pixels
Detour: curriculum domain adaptation

Feed a learning system “easy” tasks first. Their solutions find better local optima, and act as a regularizer, i.e., focusing on the test set.

**Input:** An urban scene image

**Algorithm:** Logistic regression

**Output:** Label distributions
Detour: curriculum domain adaptation

Feed a learning system "easy" tasks first. Their solutions find better local optima, and act as a regularizer, i.e., focusing on the test.

\[
\min_{\Theta} \mathcal{L}(Y_s, \hat{Y}_s) + d(p_t, p_t(\hat{Y}_t))
\]

\(s: \text{Source, } t: \text{Target}\)

\(p_t : \text{Perturbation function}\)
Detour: curriculum domain adaptation

- Sim --> Sim: 60
- Sim --> Real: 22
- Sim --> Real w/DA: 40
- Half sim half real: 53

[Yang et al., ICCV’17]
Outline

Web data with noisy labels
- Hard to rectify wrong labels
- Easier to just remove wrong labels

Web data with accurate labels
- Geometry from 3D movies
- Geometry encodes semantics

Web data of multi-modalities
- Web images vs. Web videos

Semi-supervised learning
Curriculum learning & curriculum adaptation
A comment on self-supervised learning

Geometry Guided Convolutional Neural Networks for Self-Supervised Video Representation Learning

Self-supervised learning??

Supervised learning from self-labeled data

[Chuang Gan et al., CVPR’18]
Geometry Guided Self-Supervised Learning

Abstract

Self-supervised learning of video representations. In particular, we extract pixel-wise geometry information such as flow fields and disparity maps from synthetic imagery and real 3D movies, respectively. Although the geometry and high-level semantics are seemingly distant topics, surprisingly, we find that they can be effectively adapted to semantic understanding tasks. Empirical results show that our geometry guided networks significantly outperform the competing methods that are trained with other types of labeling-free supervision signals.

1. Introduction

It is often laborious and costly to manually annotate videos for training high-quality video recognition models. The actions of interest, for instance, the temporal ordering used in "cutting in kitchen", may last for only several seconds in an hour-long video. In order to obtain a training example of this action, the annotator needs to watch through the lengthy video, manually localize those positive frames, and then trim the video. Even with sophisticated GUIs, the labor cost for obtaining one training video sequence is still prohibitively large. In this paper, we instead explore geometric signals in the previous work. Intuitively, the signal that can be used to assist semantic understanding has to be strongly correlated with semantics. Our experimental results therefore also indicate the intrinsic correlation between the geometric and semantic features.
Outline

Web data with noisy labels
- Hard to rectify wrong labels
- Easier to just remove wrong labels

Web data with accurate labels
- Geometry from 3D movies
- Geometry encodes semantics

Web data of multi-modalities
- Web images vs. Web videos

Semi-supervised learning

Curriculum learning & curriculum adaptation
Web images vs. Web videos

Given a query,

**Relevant** Web images & video frames are alike

An **irrelevant** Web image or video frame is irrelevant in its own way
Web images vs. Web videos

Given a query,

**Relevant** Web images & video frames are alike

An **irrelevant** Web image or video frame is irrelevant in its own way

(b) Bench Press
Web images vs. Web videos

Given a query,

**Relevant** Web images & video frames are alike

An **irrelevant** Web image or video frame is irrelevant in its own way

(c) Pizza Tossing
Web images vs. Web videos

Mutually vote for commonness
to select training examples
Kernel mean embedding

\[ \mu[P] \triangleq \mathbb{E}_x[\phi(x)] \]

\( \mu \) maps distribution \( P \) to Reproducing Kernel Hilbert Space

\( \mu \) is injective if \( \phi(\cdot) \) is characteristic

[Müller’97, Gretton et al.’07, Sriperumbudur et al.’10]
Empirical kernel mean estimation

\[ \mu[P] \triangleq \mathbb{E}_x[\phi(x)] \]

Empirical kernel embedding:

\[ \hat{\mu}[P] = \frac{1}{n} \sum_{i=1}^{n} \phi(x_i), \quad x_i \sim P \]
Mutually vote by matching kernel means

\[ \mu[P] \triangleq \mathbb{E}_x[\phi(x)] \]

Empirical kernel embedding:

\[ \hat{\mu}[P] = \frac{1}{n} \sum_{i=1}^{n} \phi(x_i), \quad x_i \sim P \]
Mutually vote by matching kernel means

\[
\min_{\alpha, \beta \in \{0, 1\}} \left\| \frac{1}{\sum_m \alpha_m} \sum_{m'} \alpha_{m'} \phi(I_m) - \frac{1}{\sum_n \beta_n} \sum_{m'} \beta_{m'} \phi(F_m) \right\| + \mathcal{R}(\beta)
\]

\[
\alpha_m = \begin{cases} 
1 & \text{if } I_m \text{ is similar to selected video frames} \\
0 & \text{else}
\end{cases}
\]

\[
\mathcal{R}(\beta) = \text{Reconstruct video from the selected video frames}
\]
Table 6. Comparisons with state of the arts results using fully labeled data on UCF101.

<table>
<thead>
<tr>
<th>Method</th>
<th>Acc (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LRCN [7]</td>
<td>71.1</td>
</tr>
<tr>
<td>LSTM composite model [34]</td>
<td>75.8</td>
</tr>
<tr>
<td>IDT + FV [41]</td>
<td>87.9</td>
</tr>
<tr>
<td>C3D [40]</td>
<td>82.3</td>
</tr>
<tr>
<td>Karpathy et al. [20]</td>
<td>65.4</td>
</tr>
<tr>
<td>Spatial stream network [29]</td>
<td>73.0</td>
</tr>
<tr>
<td>Ours (spatial)</td>
<td>69.3</td>
</tr>
</tbody>
</table>
Outline

Web data with noisy labels
- Hard to rectify wrong labels
- Easier to just remove wrong labels

Semi-supervised learning

Web data with accurate labels
- Geometry from 3D movies
- Geometry encodes semantics

Curriculum learning & curriculum adaptation

Web data of multi-modalities
- Web images vs. Web videos

Mutually vote by kernel means
Future work: The Web is rich & inspiring

Web data with noisy labels
Hard to rectify wrong labels
Easier to just remove wrong labels

Web data with accurate labels
Geometry from 3D movies
Geometry encodes semantics

Web data of multi-modalities
Web images vs. Web videos

Query, tags, news, audio, etc.

Semi-supervised learning
Curriculum learning & curriculum adaptation
Mutually vote by kernel means
Future work: The Web is rich & inspiring

Query-focused video summarization

[Sharghi et al., ECCV’16, CVPR’17, ECCV’18]
Future work: The Web is rich & inspiring

Web data with **noisy** labels
- Hard to rectify wrong labels
- Easier to just remove wrong labels

Web data with **accurate** labels
- Geometry from 3D movies
- Geometry encodes semantics

Web data of **multi-modalities**
- Web images vs. Web videos

Query, tags, news, audio, etc.

Semi-supervised learning

Curriculum learning & curriculum adaptation

Mutually vote by kernel means
Future work: The Web is rich & inspiring

Web data with noisy labels
Hard to rectify wrong labels
Easier to just remove wrong labels

Semi-supervised learning

Web data with accurate labels
Geometry from 3D movies
Geometry encodes semantics

Curriculum learning & curriculum adaptation

Web data of multi-modalities
Web images vs. Web videos

Mutually vote by kernel means

Query, tags, news, audio, etc.

Multi-modal methods
Domain adaptation
References


