Sharing knowledge for large scale visual recognition

Lamberto Ballan
AI Lab, Stanford University
Standard computer vision paradigm

- Q: What objects are in the image?
Datasets drive computer vision progress

ImageNet
ImageNet: ILSVRC results

• Result in ILSVRC (classification) over the years

<table>
<thead>
<tr>
<th>Team Name</th>
<th>Error (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSRA</td>
<td>3.57</td>
</tr>
<tr>
<td>ReCeption</td>
<td>3.58</td>
</tr>
<tr>
<td>Trimps-Soushen</td>
<td>4.58</td>
</tr>
<tr>
<td>Qualcomm Research</td>
<td>4.87</td>
</tr>
<tr>
<td>VUNO</td>
<td>5.03</td>
</tr>
<tr>
<td>CIL</td>
<td>5.48</td>
</tr>
<tr>
<td>CUImage</td>
<td>5.86</td>
</tr>
<tr>
<td>MCG-ICT-CAS</td>
<td>6.31</td>
</tr>
<tr>
<td>HiVision</td>
<td>6.48</td>
</tr>
</tbody>
</table>
The long tail

- A small number of generic objects/entities/labels appear very often while most others appear rarely.
- There are a few real-world scenarios in which we have access to 1M+ images uniformly belonging to a set of 1000+ classes.

Q: How to scale up to very large vocabularies (infrequent labels) and a scenario where it is hard to collect ground truth data?
Automatic image annotation by exploiting image metadata and weak labels
Motivation

• Can you guess what’s in the image?

petal?
fruit?
tentacle?
Motivation

• Let’s try to add more context…

Tags:
flower
petal
closeup
water

GPS
groups

…
Motivation

• In the context of images which share similar metadata it is easier to give the right answer
Approach

• For an image \( x \in X \) and neighborhood \( z \in Z_x \), we use a function \( f \) parameterized by \( w \) to predict labels
  
  ‣ We compute hidden state representations for the image and its neighbors

  ‣ Then we operate on the concatenation of these two representations to compute label scores

• We demonstrate that our model can:
  
  ‣ handle different types of image metadata

  ‣ adapt to changing vocabularies
Approach

• (1) *non-parametric* step to build a neighborhood

[J.Johnson*, L.Ballan*, L.Fei-Fei - ICCV 2015]
Approach

• (2) *deep neural network* to blend visual information from the image and its neighbors

[J.Johnson*, L.Ballan*, L.Fei-Fei - ICCV 2015]
Approach

• In this way the model uses features from both the image and its neighbors
### Results

- Multi-label image annotation results on the NUS-WIDE dataset (~240K Flickr images)

<table>
<thead>
<tr>
<th>Method</th>
<th>mAP_L</th>
<th>mAP_I</th>
<th>Rec_L</th>
<th>Prec_L</th>
<th>Rec_I</th>
<th>Prec_I</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tag-only Model + linear SVM [37]</td>
<td>46.67</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Graphical Model (all metadata) [37]</td>
<td>49.00</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>CNN + softmax [15]</td>
<td>-</td>
<td>-</td>
<td>31.22</td>
<td>31.68</td>
<td>59.52</td>
<td>47.82</td>
</tr>
<tr>
<td>CNN + ranking [15]</td>
<td>-</td>
<td>-</td>
<td>26.83</td>
<td>31.93</td>
<td>58.00</td>
<td>46.59</td>
</tr>
<tr>
<td>CNN + WARP [15]</td>
<td>-</td>
<td>-</td>
<td>35.60</td>
<td>31.65</td>
<td>60.49</td>
<td>48.59</td>
</tr>
<tr>
<td>Upper bound</td>
<td>100.00±0.00</td>
<td>100.00±0.00</td>
<td>68.52±0.35</td>
<td>60.68±1.32</td>
<td>92.09±0.10</td>
<td>66.83±0.12</td>
</tr>
<tr>
<td>Tag-only + logistic</td>
<td>43.88±0.32</td>
<td>77.06±0.14</td>
<td>47.52±2.59</td>
<td>46.83±0.89</td>
<td>71.34±0.16</td>
<td>51.18±0.16</td>
</tr>
<tr>
<td>CNN [27] + kNN-voting [36]</td>
<td>44.03±0.26</td>
<td>73.72±0.10</td>
<td>30.83±0.37</td>
<td>44.41±1.05</td>
<td>68.06±0.15</td>
<td>49.49±0.11</td>
</tr>
<tr>
<td>CNN [27] + logistic (visual-only)</td>
<td>45.78±0.18</td>
<td>77.15±0.11</td>
<td>43.12±0.39</td>
<td>40.90±0.39</td>
<td>71.60±0.19</td>
<td>51.56±0.11</td>
</tr>
<tr>
<td>Image neighborhoods + CNN-voting</td>
<td>50.40±0.23</td>
<td>77.86±0.15</td>
<td>34.52±0.47</td>
<td>56.05±1.47</td>
<td>72.12±0.21</td>
<td>51.91±0.20</td>
</tr>
<tr>
<td>Our model: tag neighbors</td>
<td>52.78±0.34</td>
<td>80.34±0.07</td>
<td>43.61±0.47</td>
<td>46.98±1.01</td>
<td>74.72±0.16</td>
<td>53.69±0.13</td>
</tr>
<tr>
<td>Our model: tag neighbors + tag vector</td>
<td>61.88±0.36</td>
<td>80.27±0.08</td>
<td>57.30±0.44</td>
<td>54.74±0.63</td>
<td>75.10±0.20</td>
<td>53.46±0.09</td>
</tr>
</tbody>
</table>

Table 2: Results on NUS-WIDE. Precision and recall are measured using $n = 3$ labels per image. Metrics are reported both per-label (mAP_L) and per-image (mAP_I). We run on 5 splits of the data and report mean and standard deviation.
Results: ours vs CNN baseline

• Experiment 1: evaluates AP for each label of our model vs the visual-only CNN baseline

![Graph showing mAP difference vs label frequency and visual-only scores for different labels.]
Qualitative results
Qualitative results

V-only
animal
water
flowers

Ours
water
swimmers
person

V-only
sky
clouds
person

Ours
police
person
military
Qualitative results
Qualitative results
Results: generalization

• Experiment 2: vocabulary generalization

Performance as we vary overlap between tag vocabularies used for training and testing: strong results even in the case of disjoint vocabularies
Results: generalization

• Experiment 3: metadata generalization

<table>
<thead>
<tr>
<th>Train:</th>
<th>Test:</th>
<th>Tags</th>
<th>Sets</th>
<th>Groups</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tags</td>
<td>Tags</td>
<td>52.78±0.34</td>
<td>47.12±0.35</td>
<td>48.14±0.33</td>
</tr>
<tr>
<td>Sets</td>
<td>Tags</td>
<td>52.21±0.29</td>
<td>48.02±0.33</td>
<td>48.49±0.16</td>
</tr>
<tr>
<td>Groups</td>
<td>Tags</td>
<td>50.32±0.28</td>
<td>47.82±0.24</td>
<td>48.87±0.22</td>
</tr>
</tbody>
</table>

Results using different types of metadata for training and testing

Probability that the k-th neighbor of an image has a label given that the image has the label
Knowledge transfer for scene-specific motion prediction
Humans in crowded spaces

• When humans navigate a crowded space their motion is influenced by the scene and the other active agents

• Stanford Campus Dataset: videos of various agents that navigate in a real world outdoor environment

[A.Robicquet, A.Alahi, A.Sadeghian, B.Anenberg, J.Doherty, E.Wu, S.Savarese - arXiv 2016]
Goal: motion (*trajectory*) prediction

- Given a single picture and an observed agent, humans are able to predict the most likely future
Motivation

• We believe this ability is mostly driven by two factors
Motivation

• (1) the *dynamics* of previously observed targets
Motivation

• (2) the *semantic of the scene*
Challenges

• Our model should be able to exploit the interplay between scene semantics and agents

• Data collection is hard and expensive

• Q: how to scale to large dataset / new scenes?
Approach

- This knowledge can be transferred to a new scene

Knowledge Transfer

["L. Ballan, F. Castaldo, A. Alahi, F. Palmieri, S. Savarese - arXiv 2016"]
Approach

• Given an input scene we build a navigation map \( M \) which collects the navigation statistics

• For each patch in the map we collect:
  
  ‣ Popularity score, Routing score, Histogram of Directions and Histogram of Speeds
Prediction model

- Navigation Map: two qualitative examples

Column (a) visualizes the most common paths for both classes. Columns (b,c) show the corresponding popularity and routing maps.
Prediction model

• The target state variable is defined as $X_k = (P_k, V_k)^T$
  
  ‣ $P_k = (X_k, Y_k)^T$ (position) and $V_k = (\Omega_k, \Theta_k)^T$ (velocity)

• The target interacts with the map $M$ by exploiting the navigation values for the patch he is occupying

• Given an initial condition $X_0$, our goal is to generate a sequence of future states $X_1, \ldots, X_T$, i.e. a path $\Psi_T$
The dynamic process describing the target motion is defined by:

\[ P_{k+1} = P_k + (\Omega_k \cos \Theta_k, \Omega_k \sin \Theta_k)' + w_k \] (constant velocity)

\[ V_{k+1} = \Phi(P_k, V_k; M) \]

The learned expected values in \( M \) allows our model to generate non-linear behaviors.

\( \Phi(\cdot) \) is defined in probabilistic terms by means of a Dynamic Bayesian Network (DBN)
Knowledge transfer

“The elements of the scene define a semantic context, and they might determine similar behaviors in scenes characterized by a similar context”

- Our data-driven approach uses scene similarity to transfer the functional properties to a new scene
- Scene parsing: we use a “non-parametric” algorithm (based on SIFT+LLC, GIST and MRF inference)

[J.Yang, B.Price, S.Cohen, M.Yang - CVPR 2014]
Knowledge transfer

- Context Descriptors: a weighted concatenation of the *global* and *local* semantic context components
  - *global context*: vector of distances between classes
  - *local context*: encodes the spatial configuration of nearby patches at multiple levels
Results: datasets

- **UCLA-courtyard**: 6 videos, 2 scenes, single-class (pedestrian), scene labeled with 8 semantic classes

- **Stanford-UAV**: 21 videos, 15 scenes, multi-class (pedestrian and cyclist), scene labeled with 10 semantic classes

- **Evaluation metric**: Modified Hausdorff Distance (MHD) to measure the pixel distance between ground-truth trajectories and predicted paths
Results: path prediction

• Experiment 1: evaluates the ability of the proposed model to predict long-term trajectories

<table>
<thead>
<tr>
<th>Method</th>
<th>UCLA-courtyard</th>
<th>Stanford-UAV</th>
</tr>
</thead>
<tbody>
<tr>
<td>LP</td>
<td>41.36±0.98</td>
<td>31.29±1.25</td>
</tr>
<tr>
<td>LP&lt;sub&gt;CA&lt;/sub&gt;</td>
<td>-</td>
<td>21.30±0.80</td>
</tr>
<tr>
<td>IOC [2]</td>
<td>14.47±0.77</td>
<td>14.02±1.13</td>
</tr>
<tr>
<td>SFM [14]</td>
<td>-</td>
<td>12.10±0.60</td>
</tr>
<tr>
<td>Ours</td>
<td>10.32±0.51</td>
<td>8.44±0.72</td>
</tr>
</tbody>
</table>

(a) MHD error for a given final destination  
(b) Path generation strategies (ours)
Results: path prediction

- Qualitative examples on the UCLA-courtyard dataset
  ((blue is ground-truth, cyan is LP, yellow is IOC, red is ours)
starting point

heatmap

path prediction
starting point

heatmap

path prediction
Results: knowledge transfer

- Experiment 2: evaluates the ability of our model to generalize and make predictions on novel scenes

<table>
<thead>
<tr>
<th>MHD error</th>
<th>Pedestrian</th>
<th>Cyclist</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>LP</td>
<td>34.07</td>
<td>26.15</td>
<td>31.29±1.25</td>
</tr>
<tr>
<td>IOC [2]</td>
<td>17.99</td>
<td>18.84</td>
<td>18.42±0.97</td>
</tr>
<tr>
<td>Ours</td>
<td>12.36</td>
<td>17.10</td>
<td>14.29±0.84</td>
</tr>
</tbody>
</table>

(a) Path prediction

(b) Impact of training data
Results: knowledge transfer

• Qualitative examples: (1) path forecasting vs. (2) knowledge transfer
Results: impact of the parameters

- How the performance obtained with knowledge transfer is influenced by the different parameters?

(a) image parsing

(b) context descriptors

(c) KNN / retrieval
Contact Info
lballan@cs.stanford.eu
www.lambertoballan.net

Thanks!
References
