TagProp: Discriminative Metric Learning in Nearest Neighbor Models for Image Annotation

Matthieu Guillaumin, Thomas Mensink, Jakob Verbeek, Cordelia Schmid

LEAR team, INRIA Rhône-Alpes, Grenoble, France
TagProp

- **Goal:** predict relevance of keywords for images
TagProp

- **Application 1:** Image annotation
  - Propose a list of relevant keywords to assist human annotator
Corel Annotation - Examples

**BEP: 100%**  
Ground Truth: **sun** (1.00), **sky** (1.00), **tree** (1.00), **clouds** (0.99)  
Predictions: **sun** (1.00), **sky** (1.00), **tree** (1.00), **clouds** (0.99)

**BEP: 100%**  
Ground Truth: **mosque** (1.00), **temple** (1.00), **stone** (1.00), **pillar** (1.00)  
Predictions: **mosque** (1.00), **temple** (1.00), **stone** (1.00), **pillar** (1.00)

**BEP: 50%**  
Ground Truth: **grass** (0.98), **tree** (0.98), **bush** (0.54), **truck** (0.05)  
Predictions: **flowers** (1.00), **grass** (0.98), **tree** (0.98), **moose** (0.95)

**BEP: 50%**  
Ground Truth: **herd** (0.99), **grass** (0.98), **tundra** (0.96), **caribou** (0.13)  
Predictions: **sky** (0.99), **herd** (0.99), **grass** (0.98), **hills** (0.97)

**BEP: 50%**  
Ground Truth: **mountain** (1.00), **tree** (0.99), **sky** (0.98), **clouds** (0.94)  
Predictions: **hillside** (1.00), **mountain** (1.00), **valley** (0.99), **tree** (0.99)
TagProp

- **Application 2:** Keyword based image search
  - Given one or more keywords, propose a list of relevant images

  ![Image of a jet](image.png)
Corel Retrieval - Examples

tiger 100.00 (10)

garden 60.00 (10)

town 22.22 (9)

water, pool 90.00 (10)

beach, sand 25.00 (8)
TagProp

- **Approach:** generalize from a data base of annotated images
1. Related work
2. Metric learning for nearest neighbors
3. Data sets & Feature extraction
4. Results
5. Conclusion
Related Work

- **Latent topic models**
  - Inspired from text-analysis models (pLSA, LDA)
  - Generative model over keywords and image regions
  - Trade-off: overfitting & capacity limited by nr. of topics

- **Other mixture models**
  - Non-parametric KDE over image features

- **Binary classifiers**
  - One classifier for each keyword:
    many classifiers, (no) parameter sharing
  - Many terms with very few examples
Related Work

• Many approaches for image annotation
  - Seen as machine translation [Duygulu et al. '02]
  - Extensions of LDA [Barnard et al. '03]
  - Multiple Bernoulli relevance model [Feng et al. '04]
  - Supervised multiclass labeling [Carneiro et al. '07]
  - Kernel-based ranking [Grangier & Bengio '08]

• Local learning use most similar images to predict keywords: state-of-the-art image annotation results
  - Diffusion of labels over similarity graph [Liu et al. '09]
  - Ad hoc nearest neighbor model [Makadia et al. '08]
    • Simple model, combination of many visual features
Nearest Neighbor Image Annotation

- How to choose the visual distance to define neighbors?
- How many neighbors to consider?
- How to transfer the tags of neighboring images?
Presentation Outline

1. Related work

2. Metric learning for nearest neighbors

3. Data sets & Feature extraction

4. Results

5. Conclusion
A predictive model for keyword relevance

- **Notation**
  - relevance of keyword $w$ for image $i$, $y_{iw} \in \{+1,-1\}$
  - visual distance between images $d_{ij} \geq 0$

- Use $d_{ij}$ to define weights $\pi_{ij}$ for a nearest neighbor model

- The model outputs probability of keyword relevance $p(y_{iw}=+1)$
A predictive model for keyword relevance

- **Predictions**: weighted sum over neighbor images

\[ p(y_{iw} = +1) = \sum_j \pi_{ij} p(y_{iw} = +1 \mid j) \]
A predictive model for keyword relevance

- Imagine we select image $j$ to predict keyword $w$ of image $i$

$$p(y_{iw} = +1 | j) = \begin{cases} 1 - \varepsilon & \text{if } y_{iw} = +1 \\ \varepsilon & \text{otherwise} \end{cases}$$
A predictive model for keyword relevance

- **Objective:**
  Maximize likelihood of leave-one-out predictions on training data

\[ L = \sum_{i,w} c_{iw} \log p(y_{iw}) \]

- **Optimization:**
  Gradient descent with constraints on parameters to enforce:

\[ \pi_{ij} \geq 0, \quad \sum_j \pi_{ij} = 1 \]

- **What about the weights?**
Rank-based weights

- Fixed weight for k-th neighbour: $\gamma_k$
- K parameters
- Effective neighborhood size set automatically
Distance-based weights

- Weights $\pi_{ij}$ depend smoothly on $d_{ij}$, exponential decrease
  \[ \pi_{ij} = \frac{\exp(-\lambda d_{ij})}{\sum_k \exp(-\lambda d_{ik})} \]
  - Single parameter $\lambda$: decay rate, effective neighborhood size

- What is the right visual distance?

- Metric learning to linearly combine distances
  \[ d_{ij} = w^{(1)} d_{ij}^{(1)} + w^{(2)} d_{ij}^{(2)} + \ldots + w^{(n)} d_{ij}^{(n)} \]
  - One parameter for each ‘base’ distance
Increasing the recall of rare words

- **Keywords with low frequency in database have low recall**
  - Neighbors that have the keyword do not account for enough mass in the weighted sum
  - Systematic low relevance: need to boost it

- **Adjust ‘dynamic range’ per keyword using sigmoid**
  
  \[ x_{iw} = \sum_{j} \pi_{ij} p(y_{iw} = +1 | j) \]
  
  \[ p(y_{iw} = +1) = \sigma(\alpha_w x_{iw} + \beta_w) \]
Some practical issues

- **Objective function and gradient are quadratic in the number of images**
  - We limit the size of neighborhoods to $K$. The exponential decrease justifies to forget images that are too far anyway.

- **The neighborhood is not fixed when learning metric**
  - We include as many neighbors from each distance as possible.
  - Overlap of neighborhoods allow to use approx. $2K/D$.

- **We use annotation costs to compensate for noisier keyword absences**
  - Balance weight assigned to absences and presences.
Optimization

- **Rank-based:**
  - $K$ parameters
  - concave objective
  - Gradient descent with convex constraints

- **Distance-based weights:**
  - $\#$ parameters = $\#$ base distances
  - Gradient descent with convex constraints

- **Sigmoidal modulation: iterative optimization**
  - Optimize $\{\alpha_w, \beta_w\}$ for all words, concave
  - Optimize the $\pi_{ij}$
Presentation Outline

1. Related work
2. Metric learning for nearest neighbors
3. Data sets & Feature extraction
4. Results
5. Conclusion
Data set 1: Corel 5k

- 5000 images: landscape, animals, cities
- Vocabulary of 260 words
- Annotations designed for retrieval
Data set 2: ESP Game

- 20,000 images: photos, drawings, graphs
- Vocabulary of 268 words
- Annotations generated by players of online game
Data set 3: IAPR TC-12

- 20,000 images: touristic photos, sports
- Vocabulary of 291 words
- Annotations extracted from descriptive text (nouns)
Feature extraction

- **Bag-of-words histograms**
  - SIFT [Lowe ’04] and Hue [van de Weijer & Schmid ’06]
  - Dense grid and Harris interest points
  - K-means quantization

- **Global color histograms**
  - Color spaces: RGB, HSV, LAB
  - Each channel quantized in 16 bins

- **Global GIST descriptor** [Oliva & Torralba ’01]

- **Spatial 3x1 partitioning** [Lazebnik et al. ’06]
  - Concatenate histograms from regions
  - Done for all features except GIST
Presentation Outline

1. Related work
2. Metric learning for nearest neighbors
3. Data sets & Feature extraction
4. Results
5. Conclusion
Evaluation measures

Measures computed per keyword, then averaged

- Annotate images with the 5 most likely keywords
  - **Recall**: # ims. correctly annotated / # ims. in ground truth
  - **Precision**: # ims. correctly annotated / # ims. annotated

- **Mean average precision**
# Variants of TagProp

<table>
<thead>
<tr>
<th></th>
<th>COREL 5K</th>
<th>IAPR TC-12</th>
<th>ESP Game</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P  R</td>
<td>P  R</td>
<td>P  R</td>
</tr>
<tr>
<td>Rank-based</td>
<td>28% 32%</td>
<td>35% 22%</td>
<td>27% 20%</td>
</tr>
<tr>
<td>Fixed distance</td>
<td>30% 33%</td>
<td>50% 20%</td>
<td>48% 19%</td>
</tr>
</tbody>
</table>

- Distance-based > Rank-based, comparable for COREL
## Variants of TagProp

<table>
<thead>
<tr>
<th></th>
<th>COREL 5K</th>
<th>IAPR TC-12</th>
<th>ESP Game</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
<td>P</td>
</tr>
<tr>
<td>Rank-based</td>
<td>28%</td>
<td>32%</td>
<td>35%</td>
</tr>
<tr>
<td>Fixed distance</td>
<td>30%</td>
<td>33%</td>
<td>50%</td>
</tr>
<tr>
<td>Fixed with sigmoid</td>
<td>28%</td>
<td>35%</td>
<td>41%</td>
</tr>
</tbody>
</table>

- **Distance-based > Rank-based**, comparable for COREL
- **Sigmoid**: trades precision for recall
## Variants of TagProp

<table>
<thead>
<tr>
<th></th>
<th>COREL 5K</th>
<th>IAPR TC-12</th>
<th>ESP Game</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
<td>P</td>
</tr>
<tr>
<td>Rank-based</td>
<td>28%</td>
<td>32%</td>
<td>35%</td>
</tr>
<tr>
<td>Fixed distance</td>
<td>30%</td>
<td>33%</td>
<td>50%</td>
</tr>
<tr>
<td>Fixed with sigmoid</td>
<td>28%</td>
<td>35%</td>
<td>41%</td>
</tr>
<tr>
<td>ML with sigmoid</td>
<td>33%</td>
<td>42%</td>
<td>46%</td>
</tr>
</tbody>
</table>

- **Distance-based** > **Rank-based**, comparable for COREL
- **Sigmoid**: trades precision for recall
- **Metric learning**: improves results significantly
Effect of sigmoid, detailed

- Mean recall of words
  - IAPR keywords binned by how many images they occur in
  - ML (light blue), and ML with sigmoid (dark blue)
Comparison to state-of-the-art

<table>
<thead>
<tr>
<th></th>
<th>Feng '04</th>
<th>Makadia '08</th>
<th>TagProp</th>
</tr>
</thead>
<tbody>
<tr>
<td>COREL 5K</td>
<td></td>
<td>+6</td>
<td>+10</td>
</tr>
<tr>
<td>IAPR TC-12</td>
<td></td>
<td>+18</td>
<td>+6</td>
</tr>
<tr>
<td>ESP Game</td>
<td></td>
<td>+17</td>
<td>+2</td>
</tr>
</tbody>
</table>
Retrieval performance on Corel 5k

<table>
<thead>
<tr>
<th></th>
<th>All</th>
<th>Single</th>
<th>Multi</th>
<th>Easy</th>
<th>Difficult</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grangier ’08</td>
<td>26%</td>
<td>34%</td>
<td>26%</td>
<td>43%</td>
<td>22%</td>
</tr>
<tr>
<td>Fixed distance</td>
<td>31%</td>
<td>41%</td>
<td>30%</td>
<td>49%</td>
<td>27%</td>
</tr>
<tr>
<td>with sigmoid</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ML with sigmoid</td>
<td>36%</td>
<td>46%</td>
<td>35%</td>
<td>55%</td>
<td>32%</td>
</tr>
</tbody>
</table>

- 2241 (multi-word) queries with at least one relevant image
- Easy queries: three or more relevant images
- [D. Gangier and S. Bengio, "A discriminative kernel-based model to rank images from text queries", PAMI 2008]

- **Mean average precision:** +10% overall
- **Metric learning:** improves results significantly
Distance-based weights

- Learned linear combinations are dataset-specific
Comparison to JEC and SVM

- Makadia’08, SVM & Fixed: the same combination of distances
- SVM: one per keyword, Gaussian kernel, $C=100$
- SVM, Fixed and ML: sigmoidal modulation on output scores
Conclusion

- **The main contributions**
  - Probabilistic nearest neighbor model
  - Metric learning to find optimal distance combination
  - Effective neighborhood size set automatically
  - Sigmoidal non-linearity to boost recall of rare words

- **State-of-the-art results**
  - Both on image annotation, and keyword-based retrieval
  - On three different data sets and two evaluation protocols

- **Future work**
  - Learn models specifically for annotation / retrieval
  - Use localization of (some) concepts to improve performance
TagProp: Discriminative Metric Learning in Nearest Neighbor Models for Image Annotation

Matthieu Guillaumin, Thomas Mensink, Jakob Verbeek, Cordelia Schmid

LEAR Team, INRIA Grenoble, France

Online Demo – http://lear.inrialpes.fr/~verbeek