Learning Image Classification and Retrieval Models

Thomas Mensink

26 October 2012

Jury members

Prof. Frédéric Jurie    Président
Prof. Christoph Lampert  Rapporteur
Dr. Barbara Caputo   Rapporteur
Dr. Cordelia Schmid    Advisor
Dr. Jakob Verbeek    Advisor
Dr. Gabriela Csurka    Advisor
Introduction

Image Understanding

Loved the colour of the sky and water against the dark tree reflections.

A long way to go to the top... Just above my house, close to Grenoble, in front of Belledonne mountains\textsuperscript{1}.

Bridge the semantic gap: the relation between low-level image features and semantic interpretation \textsuperscript{2}

\textsuperscript{1} Images from the Stony Brook University Captioned Photo data set
\textsuperscript{2} Smeulders \textit{et al.}, Content-based image retrieval at the end of the early years, PAMI 2000
Introduction

Image Classification

Giant Panda
fish; group; bush; claws.

Hippopotamus
fish; solitary; jungle; walks.

1. Images from Animals with Attributes data set
Introduction

Image Classification

Giant Panda
fish; group; bush; claws.

Hippopotamus
fish; solitary; jungle; walks.

1. Images from Animals with Attributes data set
## Introduction

### Image Retrieval

<table>
<thead>
<tr>
<th>Query</th>
<th>Top Ranked Retrieved Images</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paris</td>
<td></td>
</tr>
</tbody>
</table>

1. Query-by-text using Google Image Search
2. Query-by-example using the BigImbaz Visual Copy Detection Search
Introduction

Relation between Classification and Retrieval

- Retrieval as classification
  - Rank images according to confidence score from classifier

- Classification as retrieval
  - Obtain similar images and propagate labels
Introduction

Goals

1. Scaling to large data sets
2. Adapting to novel classes
3. Leverage user interaction
4. Modeling label dependencies
5. Exploiting multi-modal data
Introduction

Goal 1 - Scaling to large data sets

Flickr hosts over 6 billion photos

- Large volumes of images
- Large number of potential labels
- Efficient methods for representation, learning and testing
Introduction

Goal 2 - Adapting to novel classes

300 million photos are uploaded per day to Facebook

- Relevance of categories, classes or labels changes over time
  - new images, products and creations
Goal 3 - Leverage user interaction

- Incorporate the user set labels into predictions of other labels
- Find the middle ground between
  - automatic image labeling
  - manual image labeling
Introduction

Goal 4 - Modeling label dependencies

A tree structure defined over image labels

- Labels have an intrinsic structure or dependence
- To benefit from user input, structure is required
Introduction

Goal 5 - Exploiting multi-modal data

Data is often multi-modal
- Image is accompanied by title, place and textual description

Exploit complementarities of visual and textual information to improve image retrieval and annotation
Outline

1. Introduction
   - Motivation and Goals
   - Image Representations

2. Large Scale Classification and Adapting to Novel Classes

3. Leverage User Interaction using Label Dependencies

4. Exploiting Multi-Modal Data

5. Conclusion and Discussion
Introduction — Image Representations

Bag of Visual Words

- Successfully used for retrieval [1] and classification [2]
- Bag-of-Visual Words pipeline:
  1. Consider an image as a unordered set of patches
  2. Represent each patch with a descriptor, e.g. SIFT or LCS
  3. Assign each patch to “visual-word” from “visual-dictionary”
  4. Count the frequency of each visual-word

- Visual-dictionary: k-means clustering on large set of patches

---

1. Sivic and Zisserman, Video Google: A text retrieval approach to object matching, ICCV’03
2. Csurka et al., Visual categorization with bags of keypoints, ECCV’04
Fisher Vector


- Visual-dictionary: MoG in feature space \( p(x|\theta) \)
- Take the gradient for each patch \( \frac{1}{t} \nabla_\theta \ln p(x_t; \theta) \)

Encodes more information per visual word:
- frequency, mean and variance

Power and \( \ell_2 \) normalization:
- Improved Fisher Kernel for Large-Scale Image Classification
  Perronnin, Sánchez & Mensink, ECCV’10

1. Jaakkola and Haussler, Exploiting generative models in discriminative classifiers, NIPS’99
2. Perronnin and Dance, Fisher kernels for image categorization. CVPR’07
Outline

1. Introduction

2. Large Scale Classification and Adapting to Novel Classes

3. Leverage User Interaction using Label Dependencies

4. Exploiting Multi-Modal Data

5. Conclusion and Discussion
Motivation

- Real-life data sets are evolving over time:
  - new images or items are added every second
  - new labels, tags and products are incorporated over time
  - for example, Flickr, Twitter, Facebook, Amazon...

- Need to index, retrieve, search and categorize these classes

Related publications

- *Large scale metric learning for distance-based image classification*, Mensink, Verbeek, Perronnin & Csurka, submitted
Large Scale Classification and Adapting to Novel Classes

Related Work

- Recent focus on large-scale image classification
  - ImageNet data set [1,2]
    ▶ currently over 14 million images
    ▶ multi-class with 20 thousand classes

- Good performance is usually obtained by using:
  - Linear 1-vs-Rest SVM classifiers
  - Stochastic Gradient Descent training

1. Deng et al., ImageNet: A large-scale hierarchical image database, CVPR’09
2. Deng et al., What does classifying 10,000 image categories tell us?, ECCV’10
3. Lin et al., Large-scale image classification: Fast feature extraction, CVPR’11
4. Perronnin et al., Good practice in large-scale image classification, CVPR’12
Adapting to New Images and Classes

- Limitations of 1-vs-Rest SVM for open-ended data sets:
  - Continued training when new images become available
  - For a new class training starts from scratch

**Our approach:**

- Distance based classifiers:
  - k-Nearest Neighbors
  - Nearest Class Mean Classification

- Trivial addition of new images or new classes

- Critically depends on the distance function
  - Introduce new metric learning approach for NCM
Large Scale Classification and Adapting to Novel Classes

Nearest Class Mean Classifier
Large Scale Classification and Adapting to Novel Classes

Nearest Class Mean Classifier

- Represent each class by its mean
  \[ \mu_c = \frac{1}{N_c} \sum_{i:y_i=c} x_i \]

- Assign an image \( i \) to the class with the closest class mean
  \[ c^* = \text{argmin}_c d(x, \mu_c) \]

- Very fast at test time: linear model
- Easy to integrate new images
- Easy to integrate new classes
- Class only represented with mean, less expressive than k-NN
Large Scale Classification and Adapting to Novel Classes

Nearest Class Mean Classifier

- Represent each class by its mean
  \[ \mu_c = \frac{1}{N_c} \sum_{i: y_i = c} x_i \]

- Assign an image \( i \) to the class with the closest class mean
  \[ c^* = \arg\min_c d(x, \mu_c) \]

- Very fast at test time: linear model
- Easy to integrate new images
- Easy to integrate new classes
- Class only represented with mean, less expressive than k-NN
NCM – Probabilistic Interpretation

- Define multinomial probability distribution over classes
- We use the soft-min formulation

\[
p(c|x) = \frac{\exp - d(x, \mu_c)}{\sum_{c'=1}^{C} \exp - d(x, \mu_{c'})}
\]

- Corresponds to class posterior in generative model
  - \( p(x|c) = \mathcal{N}(x; \mu_c, \Sigma) \), with shared covariance matrix
NCM – Metric Learning

- Replace the Euclidian distance \( d(x, \mu_c) \)
- Use Mahalanobis distance, parametrized by \( W \)
  \[
  d_W(x, \mu_c) = (x - \mu_c)^\top W^\top W (x - \mu_c)
  \]
- Learn low-rank projection matrix \( W : m \times D \)
- Discriminative maximum likelihood training:
  \[
  \mathcal{L}(W) = \sum_{i=1}^{N} \ln p(c_i|x_i)
  \]
Comparison to FDA

- Three not linearly separable classes
  - Find best projection in 1 dimension
Comparison to FDA

- Classical Fisher Discriminant
  - maximizes variance between all class means
Comparison to FDA

- Our proposed metric learning approach
  - maximizes variance between nearby class means
Large Scale Classification and Adapting to Novel Classes

**Illustration of Learned Distances**

![Illustration of Learned Distances](image)

<table>
<thead>
<tr>
<th>L2</th>
<th>crane</th>
<th>stupa</th>
<th>roller coaster</th>
<th>bell cote</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mah.</td>
<td>cabbage tree</td>
<td>pine</td>
<td>pandanus</td>
<td>iron tree</td>
</tr>
</tbody>
</table>
Relation to other linear classifiers

\[ f_c(x) = b_c + \mathbf{w}_c^\top x \]

- **Linear SVM**
  - Learn \( \{b_c, \mathbf{w}_c\} \) per class

- **WSABIE [1]**
  - \( \mathbf{w}_c = \mathbf{v}_c W \)
  - \( W \in \mathbb{R}^{m \times D} \)
  - Learn \( \{\mathbf{v}_c\} \) per class and shared \( W \)

- **Nearest Class Mean**
  - \( b_c = ||W\mu_c||_2^2 \)
  - \( \mathbf{w}_c = -2 (\mu_c^\top W^\top W) \)
  - Learn shared \( W \)

1. Weston et al., Scaling up to large vocabulary image annotation, IJCAI’11
Large Scale Classification and Adapting to Novel Classes

Experimental Evaluation
Large Scale Classification and Adapting to Novel Classes

Experimental Evaluation

- Data sets:
  - ILSVRC’10: 1.2M training images, 1,000 classes
  - ImageNet-10K: 4.5M training images, 10K classes

- Image features:
  - 4K and 64K dimensional Fisher Vectors
  - PQ Compression on 64K features \[1\]
    - Reduces memory usage from 320GB to 10GB

- Training:
  - Stochastic Gradient Descent
  - Learning rate and early stopping set by validation set

---

1. Jégou et al., Product quantization for nearest neighbor search, PAMI’11
ILSVRC’10 - Top 5 Accuracy

- k-NN & NCM improve with metric learning
- NCM outperforms more flexible k-NN

<table>
<thead>
<tr>
<th>Dimensionality</th>
<th>Eucl</th>
<th>Mahanalobis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>4K</td>
<td>256  512  1024</td>
</tr>
<tr>
<td>k-NN, LMNN [1] - dynamic</td>
<td>44.1</td>
<td>61.0  60.9  59.6</td>
</tr>
<tr>
<td>NCM, learned metric</td>
<td>32.0</td>
<td>62.6  <strong>63.0</strong>  <strong>63.0</strong></td>
</tr>
</tbody>
</table>

1. Weinberger & Saul, Distance Metric Learning for LMNN Classification, JMLR’09
Large Scale Classification and Adapting to Novel Classes

ILSVRC’10 - Top 5 Accuracy

- k-NN & NCM improve with metric learning
- NCM outperforms more flexible k-NN
- Distance based classifiers competitive with SVMs

<table>
<thead>
<tr>
<th>Dimensionality</th>
<th>Eucl</th>
<th>Mahanalobis</th>
</tr>
</thead>
<tbody>
<tr>
<td>4K</td>
<td>256</td>
<td>512</td>
</tr>
<tr>
<td>k-NN, LMNN [1] - dynamic</td>
<td>44.1</td>
<td>61.0</td>
</tr>
<tr>
<td>NCM, learned metric</td>
<td>32.0</td>
<td>62.6</td>
</tr>
<tr>
<td>WSABIE [2]</td>
<td></td>
<td>60.6</td>
</tr>
</tbody>
</table>

- Baseline: 1-vs-Rest SVM **61.8**

1. Weinberger & Saul, Distance Metric Learning for LMNN Classification, JMLR’09
2. Weston et al., Scaling up to large vocabulary image annotation, IJCAI’11
Large Scale Classification and Adapting to Novel Classes

Generalization to Novel Classes
Generalization to Novel Classes

- Nearest Class Mean Classifier
  - Compute means of ImageNet-10K classes: ± 1 CPU hour
  - Re-use metric learned on ILSVRC’10

- 1-vs-Rest SVM baseline
  - Train 10K SVM classifiers: ± 280 CPU days

- NCM is faster by a factor of 8500!

<table>
<thead>
<tr>
<th>Feat. dim.</th>
<th>64K</th>
<th>21K</th>
<th>128K</th>
<th>128K</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top-1</td>
<td>13.9</td>
<td><strong>21.9</strong></td>
<td>6.4</td>
<td>16.7</td>
</tr>
</tbody>
</table>

1. Deng et al., What does classifying 10,000 image categories tell us?, ECCV’10
2. Sánchez and Perronnin, High-dimensional signature compression, CVPR’11
3. Perronnin et al., Good practice in large-scale image classification, CVPR’12
Transfer Learning - Zero-Shot Prior

- Use ImageNet class hierarchy to estimate mean of new class [1]

---

1. Rohrbach et al., Knowledge transfer and zero-shot learning in large-scale, CVPR'11
Transfer Learning - Zero-Shot Prior

- Use ImageNet class hierarchy to estimate mean of new class [1]

1. Rohrbach et al., Knowledge transfer and zero-shot learning in large-scale, CVPR'11
Transfer Learning - Zero-Shot Prior

- Use ImageNet class hierarchy to estimate mean of new class [1]

1. Rohrbach et al., Knowledge transfer and zero-shot learning in large-scale, CVPR'11
Large Scale Classification and Adapting to Novel Classes

Transfer Learning - Results ILSVRC’10

- **Step 1** Metric learning on 800 classes
- **Step 2** Estimate means for remaining 200 for evaluation:
  - Data mean per class
  - Zero-Shot prior + data mean per class
  - Baseline — trained on all 1000 classes
Conclusion

- Nearest Class Mean Classification
  - We proposed metric learning by maximum-likelihood
  - Outperforms more flexible k-NN, on par with SVM

- Advantages of NCM over 1-vs-Rest SVMs
  - Allows adding new images and classes at near zero cost
  - Shows competitive results on unseen classes
  - Can benefit from class priors for small sample sizes

- More details in thesis
  - Extension using multiple class centroids
  - Different learning objectives to speed up training
  - Analysis on convergence of low-rank formulation
Outline

1. Introduction

2. Large Scale Classification and Adapting to Novel Classes

3. Leverage User Interaction using Label Dependencies

4. Exploiting Multi-Modal Data

5. Conclusion and Discussion
Leverage User Interaction using Label Dependencies

Motivation

- Multi-label image classification
- Interactive annotation to trade-off between
  - Fully automatic annotation – cheap but low accuracy
  - Fully manual annotation – expensive and high accuracy
- Need a structure to benefit from labels set by the user

Related publications

- *Learning structured prediction models for interactive image labeling*, Mensink, Verbeek & Csurka, CVPR 2011
- *Tree-structured CRF models for interactive image labeling*, Mensink, Verbeek & Csurka, PAMI 2012
- *Learning to rank and quadratic assignment*, Mensink, Verbeek & Caetano, DISCML’12
Leverage User Interaction using Label Dependencies

Related work

- Automatic image annotation
  - 1-vs-Rest SVM Classifiers [1]
  - Nearest Neighbors approaches [2]
  - Image ranking [3,4]

- No explicit modeling of label dependencies

- Interactive multi-class classification [5]

Our approach:

- Model explicitly structure in labels

- Interactive labeling scenario

1. Everingham et al., The PASCAL Visual Object Classes Challenge 2007-2011
2. Guillaumin et al., TagProp: metric learning in nearest neighbor models, ICCV’09
3. Grangier and Bengio, Kernel-based model to rank images from text queries, PAMI’08
4. Weston et al., Learning to rank with joint word-image embeddings, ECML’10
5. Branson et al., Visual Recognition with Humans in the Loop, ECCV’10
Leverage User Interaction using Label Dependencies

Tree structure over class labels
Leverage User Interaction using Label Dependencies

Tree structure over class labels

- All labels interact with each other in the structure
- Allows for efficient and exact inference
Leverage User Interaction using Label Dependencies

Tree structured model

- Vector of (binary) labels:  $y = \{y_1, \ldots, y_L\}$
- Energy between a specific labeling and the image

$$E(y, x; w, v) = \sum_{i=1}^{L} \psi_i(y_i, x; w) + \sum_{(i,j) \in E} \psi_{ij}(y_i, y_j; v)$$

- Gibbs distribution for a specific configuration $y$:

$$p(y|x) = \frac{1}{Z(x)} \exp - E(y, x; w, v)$$

- Belief Propagation for label prediction, elicitation and parameter learning
Learning

- Learning \{w, v\}, using log-likelihood:

\[ \mathcal{L} = \sum_{n=1}^{N} \ln p(y_n|x_n). \]

- Energy is linear in parameters \(\rightarrow\) log-likelihood concave
- Maximize log-likelihood with gradient ascent

 Obtaining the tree structure

- Finding optimal tree for conditional models is intractable [1]

1. Bradley and Guestrin, Learning tree conditional random fields, ICML’10
2. Chow and Liu, Approximating probability distributions with trees, IT’68
Leverage User Interaction using Label Dependencies

Extensions
Leverage User Interaction using Label Dependencies

Extension 1 — Trees over groups of labels

- Increase the discriminative power by using nodes with $k$ labels
- Every state in a node is modeled explicitly
  - Each node has $2^k$ states
  - Label marginals read-off from state marginal table
- Trade-off: model expressiveness vs computational complexity
Leverage User Interaction using Label Dependencies

Extension 1 — Trees over groups of labels

<table>
<thead>
<tr>
<th>State</th>
<th>Marginal</th>
<th>Landscape</th>
<th>Sky</th>
<th>Clouds</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.4 %</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0.0 %</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>9.8 %</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>59.9 %</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>0.4 %</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>0.0 %</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>2.6 %</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>23.9 %</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Label marginal: 26.9% 96.2% 83.8%

- Increase the discriminative power by using nodes with \( k \) labels
- Every state in a node is modeled explicitly
  - Each node has \( 2^k \) states
  - Label marginals read-off from state marginal table
- Trade-off: model expressiveness vs computational complexity
Extension 2 — Mixture-of-trees

- Learn multiple trees
  - different group sizes $k$
  - different structures over a fixed set of nodes

- Define a mixture of $T$ trees as

$$p(y|x) = \sum_{t=1}^{T} \pi_t p_t(y|x)$$

- Learning a mixture-of-trees
  - Each tree is learned independently
  - Mixing weights are set uniformly
Leverage User Interaction using Label Dependencies

Interactive image labeling
Ask the user at test time to set a few labels
  • To improve the prediction performance

Iterative strategy: ask one label at the time
Leverage User Interaction using Label Dependencies

Label elicitation

- Determine which label should be set by the user
  - Objective: select label $y_i$ to minimize expected uncertainty of the remaining labels $H(y_i|y, x)$

- After label is set, update predictions on other labels
  - Information propagated in tree is now combination of visual information and user-provided information
Leverage User Interaction using Label Dependencies

Experimental evaluation
Leverage User Interaction using Label Dependencies

Experimental evaluation

- **Data sets**
  - ImageClef VCDT 2010 Challenge (imageClef) [1]
  - Scene Understanding (SUN’09) [2]
  - Animals with Attributes (AwA) [3]

- **Performance:** mean average precision (MAP)
  - retrieval performance per label

---

1. Nowak and Huiskes, New strategies for image annotation, ImageCLEF’10
2. Choi *et al.*, Hierarchical context on a large database of object categories, CVPR’10
3. Lampert *et al.*, Detect unseen object classes by between-class attribute transfer, CVPR’09
Leverage User Interaction using Label Dependencies

Results - Fully Automatic Labeling

- Baseline is state-of-the-art for ImageCLEF’10 and SUN’09.
Leverage User Interaction using Label Dependencies

Results - Interactive Labeling

- 10 labels are asked and set by the user

ImageCLEF’10  SUN09  AwA
Leverage User Interaction using Label Dependencies

Attribute-based image classification
Leverage User Interaction using Label Dependencies

Attribute-based image classification

- **Classes are defined by a given set of attributes**
- **Zero-shot**: Images of 10 test classes not used for training
- **Use structured models for attribute prediction**
- **Ask for attribute values to improve class prediction**
Any useful question eliminates at least 1 class
*Never more than 9 questions needed*
Leverage User Interaction using Label Dependencies

Conclusions

- Structured models for predicting image labels
  - Image annotation, and
  - Attribute-based classification.

- Mixture-of-trees is a powerful yet tractable structured model
  - Efficiently transfers knowledge of labels set by the user
  - Allows to ask relevant labels to set by user

- More details in thesis
  - Comparison of joint and stage learning of unary potentials
  - Alternatives to obtain tree structures
  - Learning to rank with pairwise label interactions
Outline

1. Introduction

2. Large Scale Classification and Adapting to Novel Classes

3. Leverage User Interaction using Label Dependencies

4. Exploiting Multi-Modal Data

5. Conclusion and Discussion
Exploiting Multi-Modal Data

Motivation

- Different modalities as weak form of supervision
- Multiple modalities relatively cheap to obtain
- Combining text and image retrieval improves performance [1]

Related publications

- *Transmedia Relevance Feedback for Image Autoannotation*, Mensink, Verbeek & Csurka, BMVC 2010
Exploiting Multi-Modal Data

Related Work

- **Late Fusion** - Combine mono-modal similarities
  - ✓ Well studied problems, and well engineered solutions [1,2]
  - ✗ Unable to exploit the correlations between different modalities

- **Early Fusion** - Joint representation of different modalities
  - ✓ Exploit the correlations
  - ✗ Representation should allow for heterogeneity of the modalities
    - variations in semantic meaning (words vs. low level)
    - histogram concatenation, or topic models [3]

- **Intermediate Fusion**
  - • Transmedia relevance feedback

---

1. Manning *et al.*, Introduction to information retrieval, CUP’08
2. Datta *et al.*, Image retrieval: Ideas, influences, and trends of the new age, ACM’08
3. Barnard *et al.*, Matching words and pictures, JMLR’03.
Exploiting Multi-Modal Data

Transmedia Relevance Feedback

■ Pseudo-Relevance Feedback [1]
  • Textual query
  • Extract keywords from top $k$ documents retrieved
  • New query: query + extracted keywords

■ Transmedia Relevance Feedback [2,3]
  • Visual query
  • Rank using visual similarity
  • Swap modality
  • New query: textual description from top $k$ documents

1. Salton and Buckley, Improving retrieval performance by relevance feedback, ASIS'90
2. Chang and Chen, Using a word-image ontology for image retrieval, CLEF'06
3. Clinchant et al., XRCEs participation to ImageCLEFphoto 2007, CLEF'07
Exploiting Multi-Modal Data

Weighted Relevance Feedback

- **Equal Weighted** [1]
  \[ s(q, d) = \sum_{k=1}^{K} s_1(q, d_k) s_2(d_k, d) \]

- **Linear Weighted**
  \[ s(q, d; \gamma) = \sum_{k=1}^{K} \gamma_k s_1(q, d_k) s_2(d_k, d) \]
  - Constrain \( \gamma_k \) to be positive and ordered

- **Softmax Weighted**
  \[ s(q, d; \gamma) = \sum_{k=1}^{K} \tilde{s}_1(q, d_k; \gamma) s_2(d_k, d) \]
  - with \( \tilde{s}_1(q, d_k; \gamma) \propto \exp(\gamma s_1(d_k, q)) \)
  - Positive and ordered by construction

---

1. Ah-Pine et al., Leveraging image, text and cross-media similarities, Springer 2010.
Exploiting Multi-Modal Data

Learning Retrieval Functions

- Combine a set of (visual, textual, and transmedia) distances
  \[ f(q, d) = \sum_i w_i s_i(q, d; \gamma_i) \]

- Learn parameters \( \{w, \gamma\} \) using comparative classification [1]
  - score relevant document higher than negative document

- Correcting for inter-query variations
  \[ f'(q, d) = \alpha_q f(q, d) + \beta_q \]
  - Difference in distribution scores
  - Ranking is independent of \( \{\alpha_q, \beta_q\} \)

---

1. Joachims, SVMs for multivariate performance measures, ICML’05
Exploiting Multi-Modal Data

Experimental evaluation
Exploiting Multi-Modal Data

Experimental evaluation – Image retrieval

- **ImageClef’08 Retrieval Challenge**

- **Comparison to participants**

<table>
<thead>
<tr>
<th>Method</th>
<th>MAP</th>
<th>P@20</th>
</tr>
</thead>
<tbody>
<tr>
<td>AVEIR</td>
<td>31.8</td>
<td>43.5</td>
</tr>
<tr>
<td>UP-GPLSI</td>
<td>33.0</td>
<td>43.1</td>
</tr>
<tr>
<td>DCU</td>
<td>35.1</td>
<td>47.6</td>
</tr>
<tr>
<td>XRCE</td>
<td>41.1</td>
<td>57.3</td>
</tr>
<tr>
<td>Ours - 2 comp</td>
<td>42.7</td>
<td>59.7</td>
</tr>
<tr>
<td>Ours - 6 comp</td>
<td>43.1</td>
<td>59.9</td>
</tr>
</tbody>
</table>

- 2-Comp: combination of text and image-to-text distances
Experimental evaluation – Image annotation

- TagProp [1] is a weighted nearest neighbor labeling approach
  - Learns a weighting of different visual distances
  - Nearest neighbors also based on transmedia distance
  - Transmedia parameters learning integrated in TagProp

- Performance measured in MAP and iMAP
  - MAP measures keyword based retrieval performance
  - iMAP measures annotation performance

- Annotation results

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Method</th>
<th>MAP</th>
<th>iMAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corel-5K</td>
<td>TagProp</td>
<td>36.0</td>
<td>54.2</td>
</tr>
<tr>
<td></td>
<td>TagProp + Transmedia</td>
<td><strong>38.1</strong></td>
<td><strong>55.6</strong></td>
</tr>
<tr>
<td>IAPR-TC12</td>
<td>TagProp</td>
<td>35.4</td>
<td>47.0</td>
</tr>
<tr>
<td></td>
<td>TagProp + Transmedia</td>
<td><strong>35.9</strong></td>
<td><strong>48.0</strong></td>
</tr>
</tbody>
</table>

1. Guillaumin et al., TagProp: metric learning in nearest neighbor models, ICCV'09
Conclusions

- Transmedia relevance for combining modalities
  - Defines true multi-modal distance, e.g. from visual to text
  - Parametrized version allow to learn parameters from data

- Multi-modal image retrieval
  - Query dependent variables to learn better parameters

- Image Annotation
  - Transmedia distance incorporated into TagProp

- More details in thesis:
  - Comparison of learning objectives for retrieval
  - Comparison of different modalities for annotation
Outline

1. Introduction

2. Large Scale Classification and Adapting to Novel Classes

3. Leverage User Interaction using Label Dependencies

4. Exploiting Multi-Modal Data

5. Conclusion and Discussion
Conclusion and Discussion

Conclusions

Goals

1. Scaling to large data sets
2. Adapting to novel classes
3. Leverage user interaction
4. Modeling label dependencies
5. Exploiting multi-modal data
Conclusion and Discussion

Conclusions

Goals

1. Scaling to large data sets
2. Adapting to novel classes

Contributions

- Metric learning approach for Nearest Class Mean classifier
  - On par with state-of-the-art linear SVMs
  - Generalizes well to unseen classes
  - Proven scalability to data sets with millions of images
Conclusion and Discussion

Conclusions

Goals

3. Leverage user interaction
4. Modeling label dependencies

Contributions

- Mixtures-of-trees to model label dependencies
  - Moderate improvement in fully automatic setting
  - Efficiently leverages user interaction
  - Shown to work in different scenarios: image annotation and attribute-based classification
Conclusion and Discussion

Conclusions

Goals

5. Exploiting multi-modal data

Contributions

- Parametrized transmedia relevance feedback
  - Effective way to combine multiple modalities
  - Learned parameters outperform manually set ones
  - Validated on multi-modal image retrieval and image annotation
Conclusion and Discussion

Future work
Conclusion and Discussion

Future work

Active and interactive learning of annotation models

- Learn classifiers and annotate large evolving set of images
- Balance the accuracy versus the annotation cost
  - For a label: select informative images to learn classifier
  - For a image: select informative labels

Research questions

1. Select between active and interactive
2. Model label dependencies in an evolving set of user labels
3. Incentives for high quality user input
Conclusion and Discussion

Future work

Structured-prediction in non-parametric models

- Combine kNN with structured models
  - Propagate pairwise marginals to the test image
- Research questions
  1. How to define structure over labels
  2. How to locally adapt the structure
Conclusion and Discussion

Future work

Structured-prediction in non-parametric models

- Combine kNN with structured models
  - Propagate pairwise marginals to the test image
- Research questions
  1. How to define structure over labels
  2. How to locally adapt the structure
Conclusion and Discussion

Future work

Address visual diversity in image classification

- Assumption: semantic class is a single coherent visual concept

- Richer class representations
  - using unsupervised discovery, e.g., Latent-SVM
  - using ideas of “visual phrases”
Learning Image Classification and Retrieval Models

Thomas Mensink

26 October 2012
Additional slides
Results - Number of questions

![Graph showing the relationship between MAP and the number of questions for different methods: Indep - Rand, Indep - Ent, Mixt - Rand, and Mixt - Ent.](image)
## Evaluation of Ridge Regression & FDA

- ILSVRC’10 dataset

<table>
<thead>
<tr>
<th></th>
<th>4K Fisher Vectors</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Projection dimensionality</strong></td>
<td>256</td>
</tr>
<tr>
<td>NCM, learned metric</td>
<td>62.6</td>
</tr>
<tr>
<td>Fisher Discriminant</td>
<td>48.0</td>
</tr>
<tr>
<td>PCA+Ridge-regression</td>
<td>31.9</td>
</tr>
</tbody>
</table>
Multiple Class Means

\[ p(c|x_i) = \sum_{j \in M_c} p(j|x_i) = \frac{1}{Z} \sum_{j \in M_c} \exp(-d_W(x_i, m_{cj})) \]

\[ Z = \sum_c \sum_{j \in M_c} \exp(-d_W(x, m_{cj})) , \]

- \( M_c \) is set of class centroids
- \( m_{cj} \) is a centroid \( j \) for class \( c \)
Learning to Rank - Summary

- **Assignment Problem** – assign a label to a rank
  - Large margin framework for structured output variables [1]
  - Loss-augmented prediction: \( \Pi^* = \arg\max_\Pi \{ f(\Pi, x; \theta) + \Delta(\Pi, z) \} \)

- **Linear score function** \( f \)

- **Loss function** \( \Delta \)
  - **linear**: e.g. Winner-takes-all, P@K [2]
    complexity \( O(N) \sim O(N^3) \)
  - **quadratic**: NP-hard in general
    Mean Average Precision, complexity \( O(N^2) \) [3]

- **Shortcoming**: label/item dependencies are not modelled

---

1. Tsochantaridis *et al.*, Large Margin Methods for Structured Output Variables, JMLR 2005
2. Le *et al.*, Optimization of Ranking Measures, JMLR 2009
3. Yue *et al.*, A SVM for Optimizing Average Precision, SIGIR 2007
Ranking with Pairwise Scores for P@K

- Label dependencies encoded in quadratic score functions
  - Results in QAP even if $\Delta$ is linear
  - General complexity: NP-hard

- C-Star models are tractable
  - Insight: given position of core $\rightarrow$ LAP
  - $N^C$ positions for core $\rightarrow O(N^{C+3})$

- P@K Loss - depends on items in top $k$
  - Insight: invariant of ordering
  - Permutation invariant score function
  - $2^C$ Rel/ Not Rel combinations in core $\rightarrow O(2^C(N + k \log k))$