The joint submission of the TU Berlin and Fraunhofer FIRST (TUBFI) to the ImageCLEF2011 Photo Annotation Task

Machine Learning Group
Berlin Institute of Technology
www.ml.tu-berlin.de

www.first.fraunhofer.de

Klaus-Robert Müller
Joint Work with Alexander Binder, Wojciech Samek, Marius Kloft and Motoaki Kawanabe
The Task

• Annotate 10000 images into 99 overlapping concepts.
• Training Corpus of 8000 annotated images
• Evaluation by measures like MAP, AUC, EER.
• User Tags available.
Our History

• We started computer vision research from July 2007 (the project **THESEUS**).

• ImageCLEF2009 Photo Annotation (We were 3\(^{rd}\)/4\(^{th}\))
  – Inspired by other front-runners, we prepared many kernels based on various image features (e.g. SIFT Bag-of-Words).
  – We combined them by **non-sparse Multiple Kernel Learning** (Kloft et al. JMLR 2011).

• ImageCLEF2010 Photo Annotation (1\(^{st}\)/2\(^{nd}\)/3\(^{rd}\) in a post-challenge evaluation)
  – More and Improved image features
ImageCLEF2011 Photo Annotation

• Non-sparse Multiple Kernel Learning (MKL) to combine many kernels from various Bag-of-Words features (e.g. multiple color channels as in van de Sande et al. 2010 and spatial tilings as in Lazebnik et al. 2006)

• Multi-Task Learning (MTL) via non-sparse MKL (Samek et al. CAIP 2011) to deal with (semantic) relations between concepts

>>> the best MAP result in visual task

• Multi-modal classification via kernel combination (Kawanabe et al. WACV 2011)

>>> the best MAP result in multi-modal task

• Novel random biased sampling
• Modified soft mapping
Combining multiple descriptors

Multiple Feature Detectors
- Point sampling strategy
  - Harris-Laplace interest points
  - Dense sampling
  - Spatial pyramid (262)
  - Spatial pyramid (2K3)

Multiple Descriptors: SIFT, shape, color, ...
- Color descriptor extraction

VQ Coding and Spatial Pooling
- Codebook model
- Bag-of-words
- Multiple bags-of-words

Nonlinear SVM
- Machine learning
- Classifier

Diagram from SurreyUVA_SRKDA, winner team in PASCAL VOC 2008
Non-sparse MKL (Min-Max Problem)

$$\min_{\beta} \max_{\alpha} \quad y^\top \alpha - \frac{1}{2} \alpha^\top \sum_j \beta_j K_j \alpha$$

s.t. \quad 0 \leq \alpha \leq C; \quad y^\top \alpha = 1; \quad \beta \geq 0; \quad \|\beta\|_p = 1.$$
Plot taken from [Kloft, Nakajima, Brefeld, ECML 2009]
Multi-Task Learning via MKL (Samek et al. CAIP2011)

- Output kernels measuring similarity between classification scores of other concepts

\[ \tilde{K}_c(x_i, x_j) = \exp \left[ - \left( s_c(x_i) - s_c(x_j) \right)^2 \right] \]
Multi-Task Learning via MKL (Samek et al. CAIP2011)

- Advantages over classical MTL (e.g. Evgeniou et al. 2005):
  1. Sample size scalability: multiple $N \times N$ kernels instead of single $(CN) \times (CN)$ kernel ($C:=\#\text{Tasks}, N:=\#\text{Samples}$)
  2. Asymmetric interactions: bad performing concepts do not spoil better performing ones

We included the output kernels of the following concepts: *animals, food, no_persons, outdoor, indoor, building_sights, landscape_nature, single_person, sky, water, sea, trees*

Directional Interactions of the VOC 2009 Task
Multi-modal classification using visual and tag kernels (Kawanabe et al. WACV2011)

- User tags have potential to provide higher-order semantic information, but they are very noisy (e.g. personal variation)
- Smoothing tag BoW feature by Markov Random Walks (MRW) based on co-occurrence graph between selected tags
- Combining visual kernels and MRW tag kernels with different steps (smoothing levels) with uniform weights (infinity-norm MKL)
## Results

Results by MAP for the best three submissions.

<table>
<thead>
<tr>
<th>Submission</th>
<th>Modality</th>
<th>MAP on test data</th>
</tr>
</thead>
<tbody>
<tr>
<td>BPACAD 3</td>
<td>T(external)</td>
<td>34.56</td>
</tr>
<tr>
<td>IDMT 1</td>
<td>T</td>
<td>32.57</td>
</tr>
<tr>
<td>MLKD 1</td>
<td>T</td>
<td>32.56</td>
</tr>
<tr>
<td>TUBFI 1</td>
<td>V(visual)</td>
<td>38.27</td>
</tr>
<tr>
<td>TUBFI 2</td>
<td>V</td>
<td>37.15</td>
</tr>
<tr>
<td>TUBFI 4</td>
<td>V</td>
<td>38.85</td>
</tr>
<tr>
<td>TUBFI 5</td>
<td>V</td>
<td>38.33</td>
</tr>
<tr>
<td>CAEN 2</td>
<td>V</td>
<td>38.24</td>
</tr>
<tr>
<td>ISIS 3</td>
<td>V</td>
<td>37.52</td>
</tr>
<tr>
<td>TUBFI 3</td>
<td>V+T</td>
<td>44.34</td>
</tr>
<tr>
<td>LIRIS 5</td>
<td>V+T</td>
<td>43.70</td>
</tr>
<tr>
<td>BPACAD 5</td>
<td>V+T</td>
<td>43.63</td>
</tr>
</tbody>
</table>
Multi-modal classification using visual and tag kernels (Kawanabe et al. WACV2011)
ImageCLEF2011 Photo Annotation

- **Non-sparse Multiple Kernel Learning (MKL)** to combine many kernels from various Bag-of-Words features
  (e.g. multiple color channels as in van de Sande et al. 2010 and spatial tilings as in Lazebnik et al. 2006)

- **Multi-Task Learning (MTL) via non-sparse MKL** (Samek et al. CAIP 2011) to deal with (semantic) relations between concepts

  >>> the best MAP result in visual task

- **Multi-modal classification via kernel combination** (Kawanabe et al. WACV 2011)

  >>> the best MAP result in multi-modal task

- Novel random biased sampling
- Modified soft mapping
End

Thank You
TUBFI Submissions(1)

- **all**: tackle diversity of concepts by diversity of methods
- **TUBFI 1**: SVM over one big set of kernels
- **TUBFI 2**: TUBFI 1+ SVMs over subsets of TUBFI 1 kernel set
- **TUBFI 5**: TUBFI 2+ non-sparse Multiple Kernel Learning (MKL) *heuristically down-scaled*
- **TUBFI 4**: TUBFI 5+ Multi-task Learning (MTL) using heuristically down-scaled non-sparse MKL
- **TUBFI 3**: multimodal: TUBFI 1+ (BoW) over User Tags