DISA at ImageCLEF 2014 Annotation Task: results, better results, and future work

Petra Budíková, Michal Batko

DISA seminar, 30. 9. 2014

Outline

- A brief history of MUFIN Annotations
- DISA at ImageCLEF 2014
  - ImageCLEF Scalable Concept Annotation Task
  - DISA solution
  - Competition results
- DISA annotation with DeCAF
  - New results
- Others at ImageCLEF
- Future work

A brief history of MUFIN Annotations

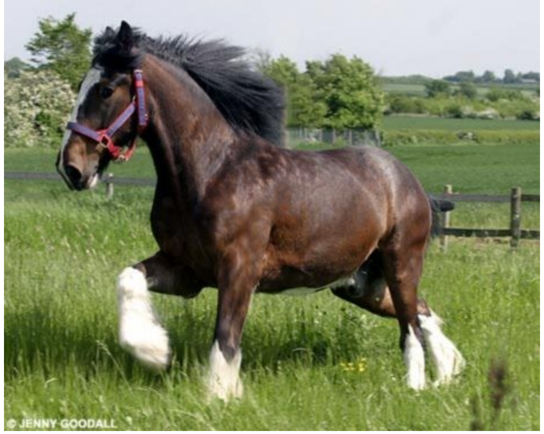
- Sometime in 2010:
  - We now have a reasonably working image search in large collections. How about using it for search-based image annotation?
- 2011:
  - Budikova, Batko, Zezula: *Online Image Annotation*. Demo SISAP 2011.
    - Very first implementation – take top N most similar images, return top K most frequent words from their descriptions. Merge synonyms using WordNet.
  - Budikova, Batko, Zezula: *MUFIN at ImageCLEF 2011: Success or Failure?*. ImageCLEF 2011.
    - Basic annotation implementation combined with face recognition and EXIF tag processing.
    - Ranked 13th out of 18 participants. Others mostly used machine learning, which was well applicable since manually preprocessed training data was available.
- 2012:
  - MUFIN Image Annotation software – extension for Firefox
    - Provides keyword annotation for arbitrary web image. Still the basic frequency-based annotation.

A brief history of MUFIN Annotations (cont.)

- 2013:
  - Batko, Botorek, Budikova, Zezula: *Content-Based Annotation and Classification Framework: A General Multi-Purpose Approach*. IDEAS 2013
    - A new generic architecture for search-based annotation processing. Multiple modules can be combined to create, expand and clean the annotation.
- 2014
  - Batko, Budikova, Elias Zezula: *CLAN Photo Presenter: Multi-modal Summarization Tool for Image Collections*. Demo ICMR 2014
    - Annotation tool used to assign keyword summaries to image clusters.
  - Budikova, Botorek, Batko, Zezula: *DISA at ImageCLEF 2014: The search-based solution for scalable image annotation*. ImageCLEF 2014
    - Exploits a new idea of conceptRank to estimate the probabilities of individual candidate concepts.
    - Ranked 5th out of 11 participants.
  - Budikova, Botorek, Batko, Zezula: *DISA at ImageCLEF 2014 Revised: Search-based Image Annotation with DeCAF Features*. Technical Report.
    - Annotations with conceptRank and DeCAF features.
    - Would have ranked 2nd out of 11 participants!

ImageCLEF 2014 Scalable Image Annotation Task

- Annotation task definition
  - Input: image + set of candidate concepts (40 to 207)
  - Expected result: set of relevant concepts



aerial airplane baby beach bicycle bird boat bridge building car cartoon castle cat chair child church cityscape closeup cloud cloudless coast **countryside** **daytime** desert diagram dog drink drum elder embroidery fire firework fish flower fog food footwear furniture garden **grass** guitar harbor hat helicopter highway **horse** indoor instrument lake lightning logo monument moon motorcycle mountain nighttime overcast painting park person **plant** portrait protest rain rainbow reflection river road sand sculpture sea shadow sign silhouette smoke snow soil space spectacles sport sun sunrise/sunset table teenager toy traffic train tricycle truck underwater unpaved wagon water

- 2 datasets
  - Development data: 1940 images, ground truth available
  - Test data: 7291 images, ground truth not available

ImageCLEF 2014 evaluation metrics

- Evaluation script provided by organizers
- Metrics
  - Concept-based: precision, recall, F-measure

$$F\text{-}measure = 2 * \frac{precision * recall}{precision + recall}$$

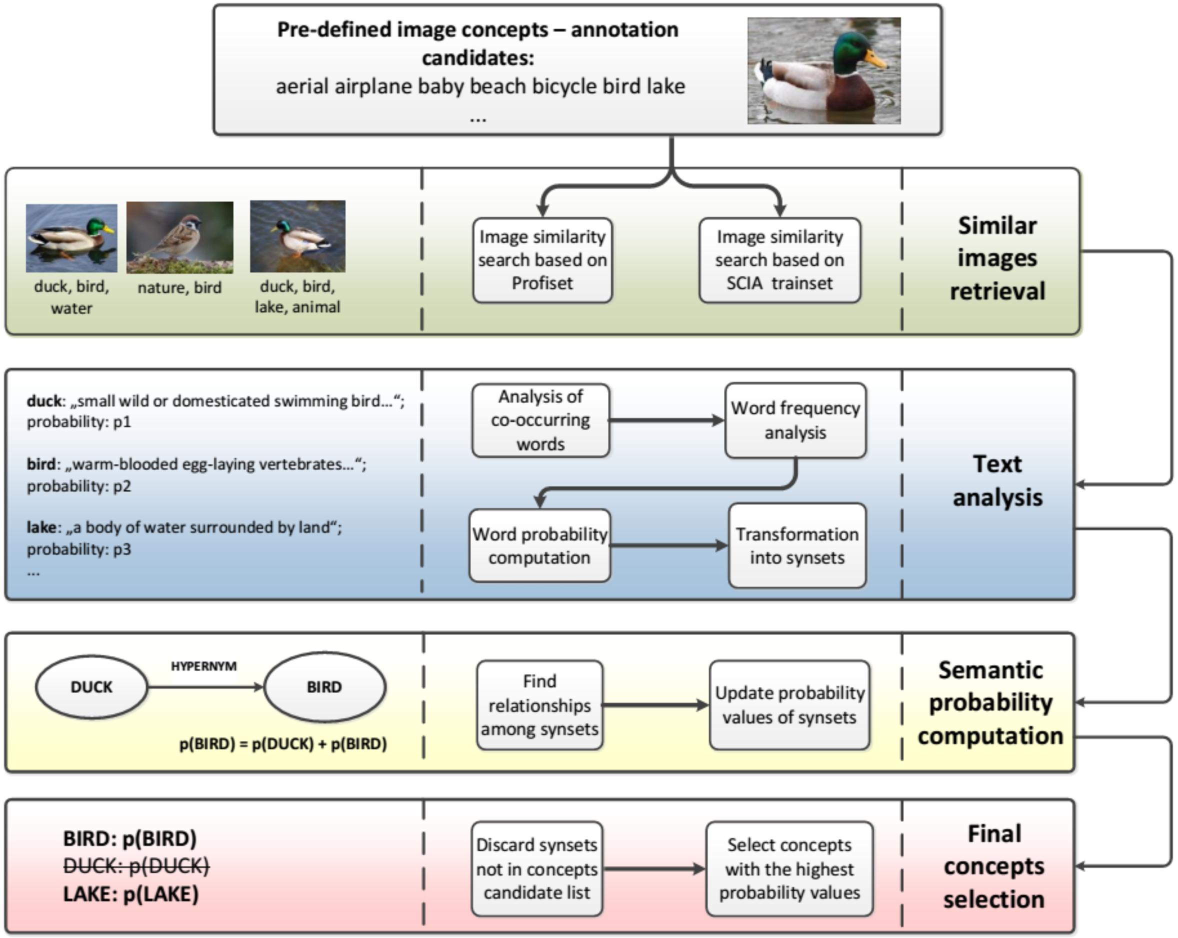
  - Sample-based: precision, recall, F-measure, AP

$$AP = \frac{1}{C} * \sum_{c=1}^{C} \frac{c}{rank(c)}$$

    - The AP uses the annotation scores, and it is computed for each image by sorting the concepts by these scores. 'C' is the number of ground truth concepts for the image and 'rank(c)' is the rank position of the c-th ranked ground truth concept.

Our solution at ImageCLEF 2014

- Search-based annotation with utilization of semantic relationships defined by WordNet



Pre-defined image concepts – annotation candidates:
aerial airplane baby beach bicycle bird lake
...

duck, bird, water

nature, bird

duck, bird, lake, animal

Image similarity search based on Profiset

Image similarity search based on SCIA trainset

Similar images retrieval

duck: „small wild or domesticated swimming bird…“; probability: p1

bird: „warm-blooded egg-laying vertebrates…“; probability: p2

lake: „a body of water surrounded by land“; probability: p3

...

Analysis of co-occurring words

Word frequency analysis

Word probability computation

Transformation into synsets

Text analysis

DUCK

HYPERNYM

BIRD

p(BIRD) = p(DUCK) + p(BIRD)

Find relationships among synsets

Update probability values of synsets

Semantic probability computation

BIRD: p(BIRD)
~~DUCK: p(DUCK)~~
LAKE: p(LAKE)

Discard synsets not in concepts candidate list

Select concepts with the highest probability values

Final concepts selection

Our solution (cont.)

- Image datasets for similarity-based searching:
  - Profiset: 20M images with high-quality keywords
  - Dataset provided by ImageCLEF organizers ("SCIA trainset"): 500K images from internet, descriptions more noisy, but covers all topics in the contest

- Image content extraction:
  - Combination of 5 MPEG7 global features

- Exploitation of semantic relationships:
  - Synonyms
  - Probability ranking of possible meanings of each word
  - Hypernymy/hyponymy
  - Holonymy/meronymy

Other group approaches (first three)

- KDEVIR - Computer Science and Engineering department of the Toyohashi University of Technology (Aichi, Japan)
  - Used features provided by organizers
  - Automatic ontology built per concept using WordNet and Wikipedia
  - Training positive and negative samples selected by exploiting ontologies
- MIL - Machine Intelligence Lab of the University of Tokyo (Tokyo, Japan).
  - Used combination of various descriptors, including FisherVectors & DeCAF
  - Linear multi-label classifier by machine learning
- MindLab - Machine learning, perception and discovery Lab from the Universidad Nacional de Colombia (Bogotá, Columbia)
  - Used DeCAF features
  - A logistic regression (soft-max) mode machine learning to classification

ImageCLEF 2014 results

- Our solution ranked 5th of the 11 groups
- We are more successful in “sample” metrics
  - The “concept” metric require that we find the

System	MAP-samples				MF-samples				MF-concepts				
	all	ani.	food	207	all	ani.	food	207	all	ani.	food	207	unseen
KDEVIR 9	36.8	33.1	67.1	28.9	37.70	29.9	64.9	32.0	54.7	67.1	65.1	31.6	66.1
MIL 3	36.9	30.9	68.6	23.3	27.50	20.6	53.1	18.0	34.7	34.7	50.4	16.9	36.7
MindLab 1	37.0	43.1	63.0	22.1	25.80	17.0	45.2	18.3	30.7	35.1	35.3	16.7	34.7
MLIA 9	27.8	18.8	53.6	16.7	24.80	12.1	46.0	16.4	33.2	32.7	37.3	16.9	34.8
DISA 4	34.3	46.6	39.6	19.0	29.70	40.6	31.2	16.9	19.1	23.0	22.3	7.3	19.0
RUC 7	27.5	25.2	44.2	15.1	29.30	28.0	28.2	20.7	25.3	20.1	23.1	10.0	18.7
IPL 9	23.4	30.0	48.5	18.9	18.40	20.2	29.8	17.5	15.8	15.8	33.3	12.5	22.0
IMC 1	25.1	35.7	35.6	12.9	16.30	14.3	21.0	10.9	12.5	10.2	15.1	6.1	11.2
INAOE 5	9.6	6.9	15.0	8.5	5.30	0.4	0.5	6.4	10.3	1.0	0.8	17.9	19.0
NII 1	14.7	23.2	22.0	4.6	13.00	18.9	18.7	4.9	2.3	3.0	2.1	0.9	1.8
FINKI 1	6.9	N/A	N/A	N/A	7.20	8.1	12.3	4.1	4.7	6.3	9.0	2.9	4.7

New features for image retrieval

- DeCAF$_7$ visual features
  - Utilization of deep convolutional network
  - Outperformed all participants at ImageNet large scale visual recognition challenge ILSVRC-2012 (Krizhevsky et. al. 2012)
  - Adopted as visual descriptor (Donahue et. al. 2013)
    - Result from the last hidden layer used as 4096-dimensional visual descriptor
    - Similarity using classical L$_p$ metric
    - Gives better results than traditional features on benchmarks from other domains

- Easily used by our similarity-search framework
  - PPP-Codes technique able to index 20M collection of data
  - Real-time response on a common server hardware
    - 8 cores, 8GB RAM, 256GB SSD

- Improved results of our annotation!

Evaluation results

- Development data

	mP-concept	mR-concept	mF-concept	mP-sample	mR-sample	mF-sample	mAP-sample
Baseline (random)	*0.0775*	*0.0641*	*0.0498*	*0.0730*	*0.0969*	*0.0722*	*0.1578*
DISA-best with MPEG and Profiset data	0.2954	0.2746	0.2184	0.3044	0.4516	0.3352	0.4268
DISA-best with MPEG and Profiset+SCIA data	**0.2919**	**0.2778**	**0.2202**	**0.3052**	**0.4533**	**0.3369**	**0.4281**
DISA-best with DeCAF and Profiset data	0.4768	0.4899	0.4165	0.4466	0.6152	0.4825	0.6105
DISA-best with DeCAF and Profiset+SCIA data	**0.4928**	**0.5085**	**0.4315**	**0.4534**	**0.6252**	**0.4901**	**0.6196**

- Test data

	mF-concept	mF-sample	mAP-sample
Baseline (random)	*0.026*	*0.035*	*0.088*
DISA-best with MPEG and Profiset data	0.154	0.279	0.316
DISA-best with MPEG and Profiset+SCIA data	0.191	0.297	0.343
Competition best	**0.547**	**0.377**	**0.368**
DISA-best with DeCAF and Profiset+SCIA data	**0.411**	**0.399**	**0.486**

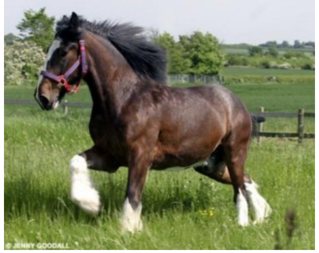
Evaluated by ImageCLEF organizers as a favor after competition deadline

New result evaluation – details

	mF-concept	mF-sample	mAP-sample
DISA-MU 04 (DISA best in competition)	19.1 [17.5–21.8]	29.7 [29.2–30.3]	34.3 [33.8–35.0]
KDEVIR 09 (competition winner)	54.7 [50.9–58.3]	37.7 [37.0–38.5]	36.8 [36.1–37.5]
DISA-MU NEW	41.1 [38.3–44.2]	39.9 [39.3–40.5]	48.6 [47.9–49.3]

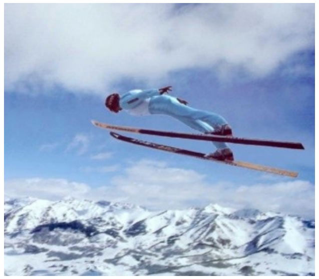
System	MAP-samples				MF-samples				MF-concepts				
	all	ani.	food	207	all	ani.	food	207	all	ani.	food	207	unseen
KDEVIR 9	36.8	33.1	67.1	28.9	37.70	29.9	64.9	32.0	54.7	67.1	65.1	31.6	66.1
DISA NEW	48.6	51.0	67.2	32.3	39.90	44.4	48.5	26.7	41.1	N/A	N/A	N/A	44.9
MIL 3	36.9	30.9	68.6	23.3	27.50	20.6	53.1	18.0	34.7	34.7	50.4	16.9	36.7
MindLab 1	37.0	43.1	63.0	22.1	25.80	17.0	45.2	18.3	30.7	35.1	35.3	16.7	34.7
MLIA 9	27.8	18.8	53.6	16.7	24.80	12.1	46.0	16.4	33.2	32.7	37.3	16.9	34.8
DISA 4	34.3	46.6	39.6	19.0	29.70	40.6	31.2	16.9	19.1	23.0	22.3	7.3	19.0
RUC 7	27.5	25.2	44.2	15.1	29.30	28.0	28.2	20.7	25.3	20.1	23.1	10.0	18.7
IPL 9	23.4	30.0	48.5	18.9	18.40	20.2	29.8	17.5	15.8	15.8	33.3	12.5	22.0
IMC 1	25.1	35.7	35.6	12.9	16.30	14.3	21.0	10.9	12.5	10.2	15.1	6.1	11.2
INAOE 5	9.6	6.9	15.0	8.5	5.30	0.4	0.5	6.4	10.3	1.0	0.8	17.9	19.0
NII 1	14.7	23.2	22.0	4.6	13.00	18.9	18.7	4.9	2.3	3.0	2.1	0.9	1.8
FINKI 1	6.9	N/A	N/A	N/A	7.20	8.1	12.3	4.1	4.7	6.3	9.0	2.9	4.7

Results illustration – the top 5 concepts (DeCAF-best method, random selection of queries)



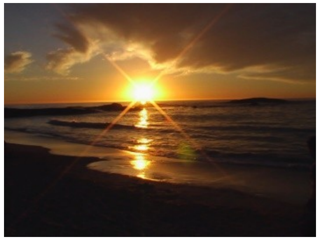
Horse, ~~person~~, grass, daytime, plant

Missing: countryside



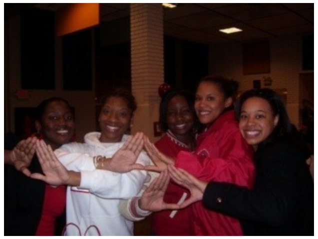
Sport, ~~male~~, ~~water~~, ~~sea~~, ~~sky~~

Missing: cloud, ski, snow



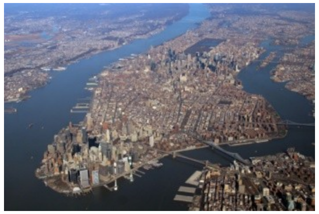
Sunrise/sunset, water, beach, sea, coast

Missing: cloud, reflection, sand, sky, sun



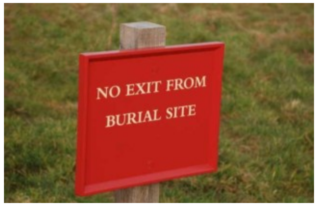
Person, ~~food~~, ~~child~~, indoor, ~~teenager~~

Missing: --



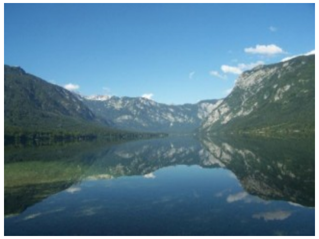
~~Sea~~, cityscape, water, ~~coast~~, aerial

Missing: bridge, outdoor, river



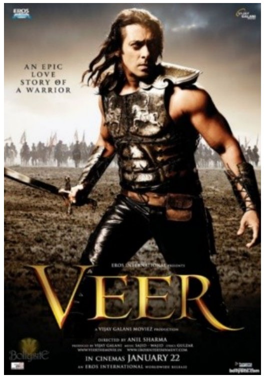
Outdoor, ~~person~~, ~~tree~~, ~~food~~, ~~forest~~

Missing: grass, sign



Mountain, water, lake, ~~river~~, ~~sea~~

Missing: cloud, daytime, reflection



Person, ~~female~~, male, ~~motorcycle~~, poster

Missing: cloud, horse, overcast, soil

DeCAF vs. MPEG

- Out of 1940 development queries
  - AP_{MPEG} is higher than AP_{DeCAF} in 357 cases
  - $\text{Precision}_{\text{MPEG}}$ is higher than $\text{Precision}_{\text{DeCAF}}$ in 201 cases
  - $\text{Recall}_{\text{MPEG}}$ is higher than $\text{Recall}_{\text{DeCAF}}$ in 158 cases

- When MPEG results are better, typically
  - the query image is difficult
  - neither MPEG nor DeCAF provide good results
  - MPEG-based results often better by small margin
  - MPEG-based results often probably better by chance

- With very few exceptions, DeCAF-based visual similarity is better

Conclusions

- Presented modular architecture of DISA annotation tool
  - allows easy replacement of any component

- Our approach is based on nearest-neighbor search not training
  - completely scalable – crawled data can be directly indexed
  - no need for ground truth
  - generic vocabulary (keyword) annotation – no need to hit predefined classes

- New visual similarity by DeCAF features
  - The new similarity-search component enabled us to increase the quality of annotations by approximately 10-20 % (depending on the quality measure)
  - New DISA results outperform the best results submitted to ImageCLEF 2014 Annotation Challenge in 2 out of 3 quality measures

Future work

- With CVUT: other descriptors/neural network descriptors trained on different data
- Refinement of conceptRank algorithm
- Relevance feedback

- Experiments with other queries+GT

- Journal paper (by December)