FROM RESEARCH TO INDUSTRY

cea tech

CEA LIST at ImageCLEF Scalable Image Concept Annotation 2013

Hervé Le Borgne, Adrian Popescu

list

INSTITUT CARNOT CEA LIST

digiteo

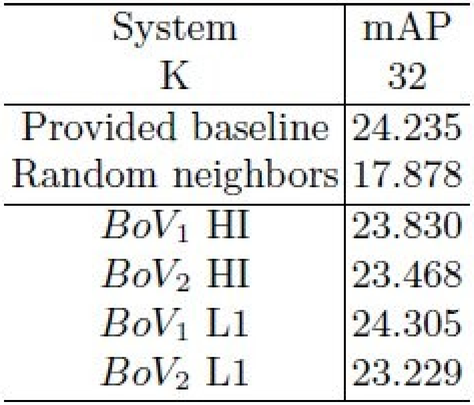
GENERAL PRINCIPLE

- **Successive « improvements » of the baseline provided**
- **We define different**
  - Visual models **V** + distance, learning models…
  - Textual models **M** (tag-based models)
- **Principle: for an image to annotate**
  - Find visual neighbors of the image into the learning database
  - These (neighboring) images have tags (with a confidence score)
  - The set of tags generates a textual description according to **M**
  - Concepts are described according to **M** as well → similarity
- **Late fusion with visual models**
- **Decision value (0/1) independant for each query**
  - Score at 1 when more tahn average + standard deviation
- **Use the development set to test the efficiency of differrent strategies**

FINDING VISUAL NEIGHBORS

- **Baseline uses C-SIFT based decriptors**
- **Our alternative:**
  - SIFT, gray-based, densely extracted every 3 pixels
  - Bag of visterm: local soft coding + max pooling
  - Two pyramids:
    - BoV_1: 1x1 +x 3x1 + 2x2
    - BoV_2 : 1x1 + 2x2 + 4x4
- **Two distances tested**
  - Histogram intersection
  - L_1
- **Similar results to the baseline (non significant improvement)**
  - K=32 visual neighbors
- **We'll use BoV_1 + L_1**

$$Dist_{HI}(x - y) = 1 - \frac{1}{D}\sum_{i=1}^{D}\frac{min(x_i, y_i)}{max(x_i, y_i)}$$



System	mAP
K	32
Provided baseline	24.235
Random neighbors	17.878
BoV_1 HI	23.830
BoV_2 HI	23.468
BoV_1 L1	24.305
BoV_2 L1	23.229

WIKIPEDIA-ESA MODEL

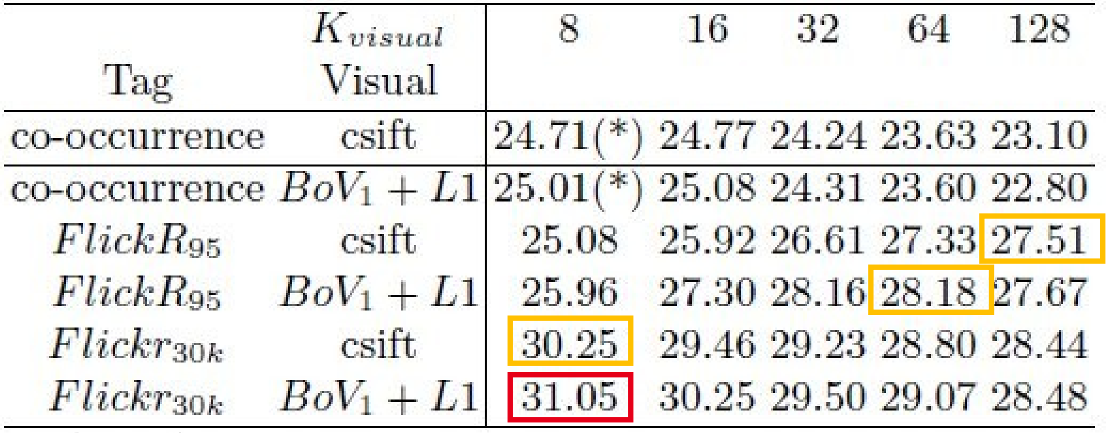
- **Explicit Semantic Analysis performed on top of Wikipedia content**
  - Map words onto Wikipedia concepts
  - 1187980 wikipedia concepts to start with
  - Concept selection using the inlink count to keep the most frequent concepts – experiments with top 5k concepts
- **In the task, map training image annotations to Wikipedia concepts**

FLICKR-ESA TAG MODEL

Inspired by Explicit Semantic Analysis but done with Flickr data

- Flickr 95 – map each image's annotation onto the set of 95 development concepts
- FlickR30k – map each image's annotation onto a set of 30k Wikipedia concepts

- **Results:**



	K_{visual}	8	16	32	64	128
Tag	Visual					
co-occurrence	csift	24.71(*)	24.77	24.24	23.63	23.10
co-occurrence	$BoV_1 + L1$	25.01(*)	25.08	24.31	23.60	22.80
$FlickR_{95}$	csift	25.08	25.92	26.61	27.33	27.51
$FlickR_{95}$	$BoV_1 + L1$	25.96	27.30	28.16	28.18	27.67
$Flickr_{30k}$	csift	30.25	29.46	29.23	28.80	28.44
$Flickr_{30k}$	$BoV_1 + L1$	31.05	30.25	29.50	29.07	28.48

- **Conclusion:**
  - Significant improvement due to the FlickR$_{30k}$-based tag model
  - The number of visual neighbors considered influences the results in conjunction with the complexity of the tag-based model

LEARNING VISUAL MODELS

- **Consider the learning database described with FlickR$_{30k}$**
- **Principle:**
  - Images ranked % score for each concept
  - Select positive and negative samples
  - Learn a SVM-based model for each concept
- **Different strategies to choose positive and negative samples**
  - S1: 100 most similar Versus 500 least similar
  - S2: two thresholds: positive>0.8 and negative < 0,1
  - S3: usage of visual coherency [Myoupo et al, 2010]
- **Results: the simpler, the better!**
  - S1: mAP = 0,219
  - S2: mAP = 0,212
  - S2: mAP = 0,209
- **Late fusion with tag-based results**

PARTICIPATION TO IMAGE CLEF 2013

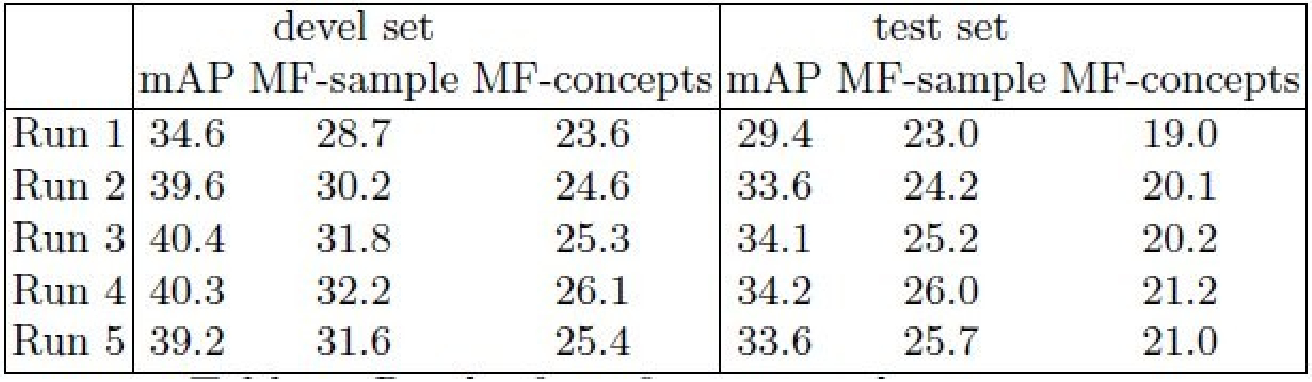
Run 1: 0.8 * FlickR$_{30k}$ + 0.2 * visual

Run 2 : 0.8 * (FlickR$_{30k}$+ESA$_{5k}$)+ 0.2 * visual

Run 3 : 0.8 * (FlickR$_{50k}$+ESA$_{5k}$)+ 0.2 * visual

Run 4 : 0.8 * (FlickR$_{30k}$+ESA$_{5k}$)+ 0.2 * VC$_{score}$

Run 4 : 0.8 * (FlickR$_{200k}$+ESA$_{5k}$)+ 0.2 * VC$_{score}$



	devel set			test set		
	mAP	MF-sample	MF-concepts	mAP	MF-sample	MF-concepts
Run 1	34.6	28.7	23.6	29.4	23.0	19.0
Run 2	39.6	30.2	24.6	33.6	24.2	20.1
Run 3	40.4	31.8	25.3	34.1	25.2	20.2
Run 4	40.3	32.2	26.1	34.2	26.0	21.2
Run 5	39.2	31.6	25.4	33.6	25.7	21.0