Overcoming the Scarcity of Training Data

Hichem SAHBI

CNRS TELECOM ParisTech, University Paris Saclay

ImageCLEF/CLEF 2016 Scalable Concept Image Annotation (subtask 1)

Sep 7th 2016

Outline

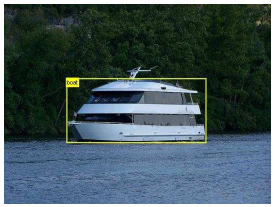
Visual Category Recognition

Goal : recognize categories (sky, sea, cars, trees, roads, persons,...)

Concept Detection

car, building, etc.

Concept Localization



Object Class Segmentation

Related work : Pascal VOC 05-12, ImageCLEF 04-., ImageNET 09-., MS COCO 14-., etc.

Motivation and Contribution (I)

Existing solutions and limitations

- **Step 1** (proposal suggestions) : sliding windows, pixel-based partitions, segmentation or multi-segmentation generation.
- **Step 2** (scoring/classification) : SVM, deep learning+ CRF, MRF, etc.
- Classification is challenging especially for classes with scarce data and high variability.

Transfer Learning

- How to adapt classifiers learned on some tasks (classes) in order to learn other tasks.
- Attribute and deep learning : learn common characteristics among different classes in order to benefit from all the available training data.
- Augmentation : use a priori knowledge to add more data.

Motivation and Contribution (II)

Proposed Solution in this Challenge (2-steps)

- Add more training data by querying the web.
- Once training data collected, **add more labels** by learning statistical dependencies between classes/labels.

Outline

Outline

Concept Detection with SVMs

- We trained "one-versus-all" SVM classifiers for each concept c ; we use many random folds (taken from training data) for multiple SVM training and we use these SVMs in order to predict the concepts on image (depending on $f_c(x)$)

$$f_c(x) = \sum_{\ell=1}^{N} 1_{\{g_{\ell(x)} \geq 0\}} - \sum_{\ell=1}^{N} 1_{\{g_{\ell(x)} < 0\}}, \qquad (N = 10 \text{ in practice})$$

$$g_\ell(x) = \sum_{x'} \alpha_{\ell,x'} \mathsf{K}_{x,x'} + b_\ell$$

- We used the coefficients of the pre-trained (not fine-tuned) VGG network.

"Chatfield, K. and Simonyan, K. and Vedaldi, A. and Zisserman, A.", Return of the Devil in the Details : Delving Deep into Convolutional Nets, "British Machine Vision Conference", 2014

Kernel Map Evaluation (I)

- About kernels : $\mathbf{K}_{x,x'}$, when (p.s.d) $\mathbf{K}_{x,x'} = \Phi'_x \Phi_{x'}$
- Linear kernel map : $\mathbf{K}_{x,x'} = \langle x, x' \rangle$ (just identity map).
- Polynomial kernel map : $\mathbf{K}_{x,x'} = \langle x, x' \rangle^p = \Phi'_x \Phi_{x'}$ with $\Phi_x = x \otimes ... \otimes x$ (p times).
- Histogram intersection map : $\mathbf{K}_{x,x'} = \sum_{d=1}^{s} \min(x^d, x'^d)$. Each dimension x^d of x is mapped using

$$\psi(x^d) = 2^0 + 2^1 + \cdots + 2^{k(x^d)}$$

$$k(x^d) = \left\lfloor Q \frac{x^d - \ell_d}{u_d - \ell_d} \right\rfloor$$

$\psi(.)$ is a "decimal-to-unary" map ; $\psi(x^d)$ is a Q dimensional vector with its $k(x^d)$ first dimensions equal to 1 and the remaining $Q - k(x^d)$ to 0, e.g., with $Q = 4$, 1 is mapped to 0001, 2 is mapped to 0011, 3 is mapped to 0111, and so on.

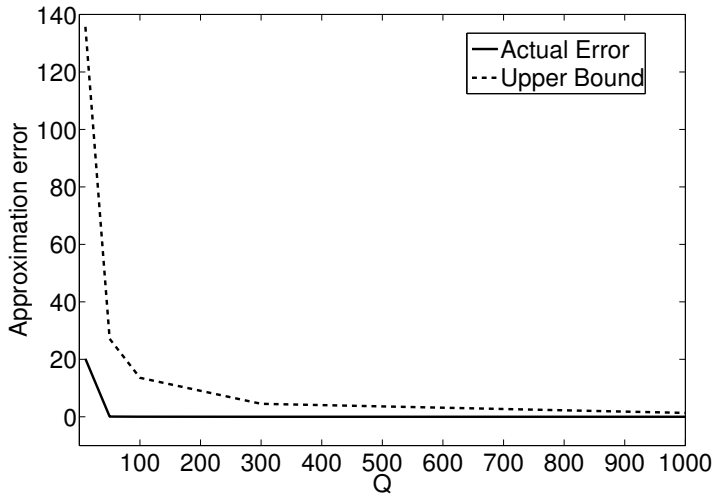
Kernel Map Evaluation (II)

Proposition

Given x, x' in $\mathcal{X}$, for sufficiently large Q, the inner product $\langle \Phi_x, \Phi_{x'} \rangle$, with $\Phi_x = \left(\psi(x^1)' \sqrt{\frac{u_1 - \ell_1}{Q}}, \sqrt{u_1}, \ldots, \psi(x^s)' \sqrt{\frac{\ell_s - u_s}{Q}}, \sqrt{u_s} \right)'$,
approximates the histogram intersection kernel $\sum_{d=1}^{s} \min(x^d, x'^d)$.

Proof shows that $\left| \langle \Phi_x, \Phi_{x'} \rangle - HI(x, x') \right| \leq \frac{1}{Q} \sum_{d=1}^{s} u_d - \ell_d \rightsquigarrow 0$ as $Q \nearrow$

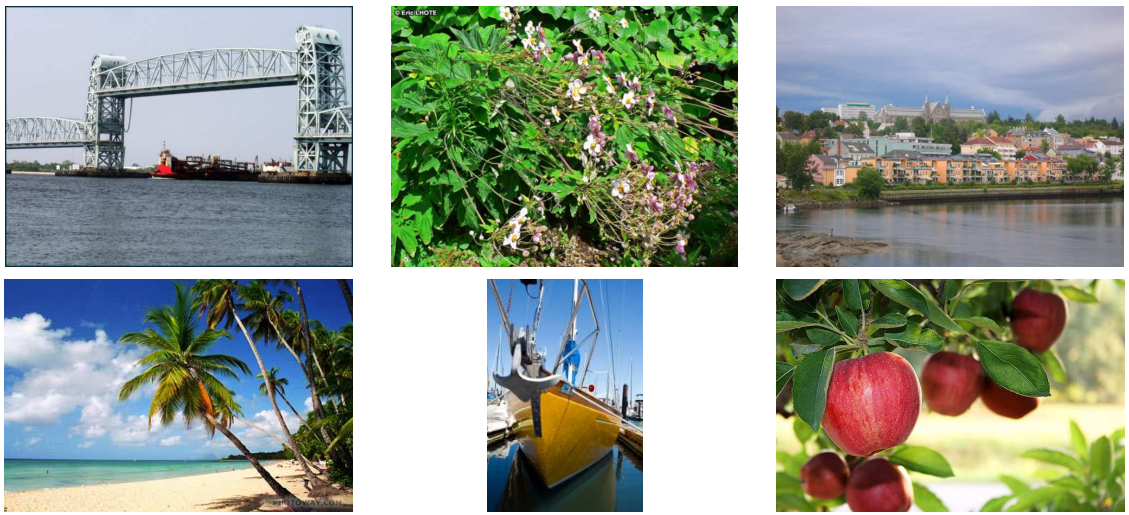
(Sahbi, ICPR 2014)



Outline

Training Set Enrichment (I)

- Internal (ImageCLEF16) Set
  - More than 500k images : with only 2k images being labeled.
- External (2-steps)
  - Augmentation : we use the crawler "googlebot-image", and we collect 42,272 images by querying all the 251 concepts.
  - Label enrichment : using dependency statistics between concepts.

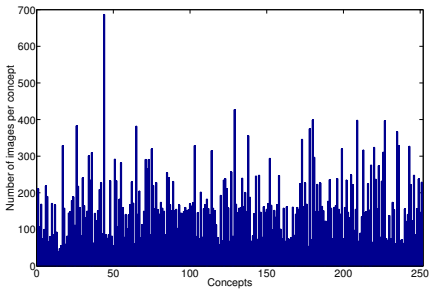


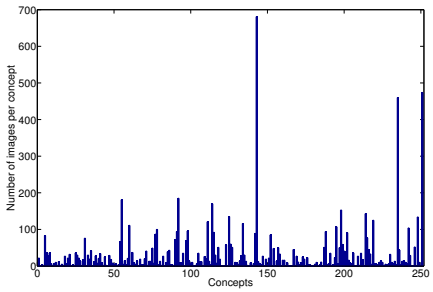
Training Set Enrichment (Step 1) : Augmentation

- We query each concept c (in "concept.txt").
- We define $\mathcal{V}(c)$: a set of expansions of c into multiple keywords (morphological, synonyms, translations, compound word definitions with "+" and "|", etc.)

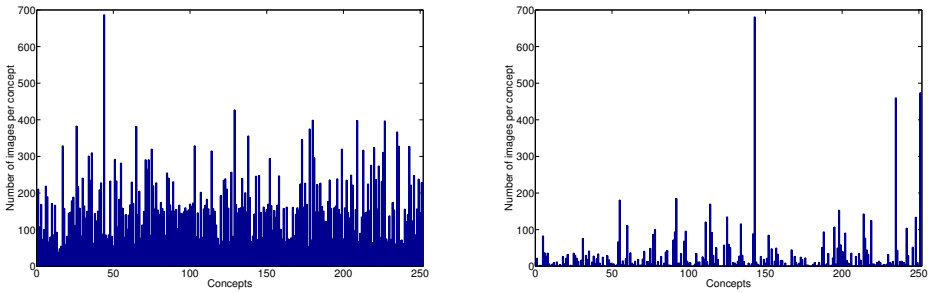
  n02691156 airplane.n.01 airplane,aeroplane,plane ["an aircraft that has a fixed wing and is powered by propellers or jets"]

  $\mathcal{V}$("airplane.n.01") = {airplane,aeroplane,plane,aircraft,propellers,wings+engines,jet,avion}
- We submit each keyword in $\mathcal{V}$ to googlebot-image crawler.
- No concept localization -> objects are usually centered.





Training Set Enrichment (Step 2) : Label Enrichment



- Still some concepts remain rare.
- To get more images for those concepts, our idea is to transfer the knowledge about the co-occurrence of some labels using a simple principle :
- *Given two concepts c and c', if c, c' are highly correlated, then the presence of one of these two concepts in a given training image implies the presence of the other concept.*

Training Set Enrichment (Step 2) : Label Enrichment

- To implement this statement, we define

$$\mathbf{C}(c'|c) = \frac{\sum_{i=1}^{N} \mathbf{Y}_{ic}\mathbf{Y}_{ic'}}{\sum_{i=1}^{N} \mathbf{Y}_{ic}}, \qquad c, c' = 1 \dots K,$$

  $\mathbf{Y}_{ic} = 1$ iff c is present into image $\mathcal{I}_i$ and $\mathbf{Y}_{ic} = 0$ otherwise.
- As external images are collected using "individual keywords" as queries, they have a single label per image and cannot be used to learn these co-occurrences (while dev set can be used).
- Labels in the external set are enriched as follows
  $\forall c, c' \in \{1, \dots, K\}$, $\forall i \in \{N+1, \dots, N+N'\}$ $(N' = 42,272)$,
  if $\mathbf{Y}_{ic} = 1$ and $\mathbf{C}(c'|c) \geq \sigma$ then $\mathbf{Y}_{ic'} \leftarrow 1$.

Outline

ImageCLEF 2016 Benchmark

- >500k images, 251 categories, only 2k labeled.



- Images are described using VGG deep features.

- Performance measured using MAP based at different percentages of bounding box overlaps.

ImageCLEF 2016 : Submitted Runs

- # submitted runs : 10 based on a combination of 2 criteria
  - **Criterion 1 :** whether enrichment is used or not (5 strategies).
  - **Criterion 2 :** whether external or external+internal data are used (2 sets).
  - No bounding box heuristics are considered for this edition.

Datasets \ Enrichment	No $\tau = 0.00$	Yes $\sigma = 0.01$ τ (RA)	Yes $\sigma = 0.01$ $\tau = 8.00$	Yes $\sigma = 0.1$ $\tau = 0.00$	Yes $\sigma = 0.75$ $\tau = 0.00$
External	TAB.0.1.res	TAB.0.4.res	TAB.0.4.1.res	TAB.0.3.res	TAB.0.5.res
External +ImageCLEF16 (dev)	TAB.1.1.res	TAB.1.4.res	TAB.1.4.1.res	TAB.1.3.res	TAB.1.5.res

c exists in x iff $f_c(x) = \sum_{\ell=1}^{N} 1_{\{g_{\ell(x)} \geq 0\}} - \sum_{\ell=1}^{N} 1_{\{g_{\ell(x)} < 0\}} \geq \tau$

ImageCLEF 2016 : External vs. External+Internal

Runs \ Overlap	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100 %
External Only											
CNRS/TAB.0.1.res	**19.62**	15.67	12.01	9.78	8.13	6.73	5.77	4.83	3.86	2.81	**1.65**
CNRS/TAB.0.5.res	19.39	**15.89**	**12.38**	**10.08**	**8.39**	**6.88**	**5.90**	**4.95**	3.90	2.75	1.64
CNRS/TAB.0.3.res	17.31	14.27	11.53	9.53	8.00	6.77	5.89	4.93	**3.97**	**2.84**	**1.81**
CNRS/TAB.0.4.res	10.59	7.43	5.71	4.64	3.73	3.05	2.67	2.25	1.87	1.37	0.91
CNRS/TAB.0.4.1.res	10.25	7.12	5.41	4.33	3.48	2.85	2.49	2.10	1.72	1.25	0.82
External + CLEF16 dev											
CNRS/TAB.1.1.res	**24.75**	**21.89**	**18.32**	**15.14**	**12.83**	**11.11**	**9.62**	**7.71**	**6.13**	**4.13**	**2.42**
CNRS/TAB.1.5.res	21.53	19.44	16.48	13.66	11.56	9.96	8.66	6.85	5.41	3.38	2.06
CNRS/TAB.1.3.res	16.85	13.72	10.98	9.06	7.58	6.36	5.63	4.74	3.79	2.67	1.64
CNRS/TAB.1.4.res	10.29	7.03	5.06	3.99	3.30	2.70	2.35	2.01	1.67	1.29	0.83
CNRS/TAB.1.4.1.res	9.79	6.55	4.80	3.74	3.09	2.53	2.17	1.91	1.57	1.19	0.74

- combining external data with the ImageCLEF16 dev set provides a clear gain compared to the use of external data only.
- ImageCLEF16 dev set has (possibly) a similar distribution w.r.t ImageCLEF16 test set.
- and this makes it possible to adapt training parameters (SVM weights) to test data.

"TAB.1.1.res" vs. "TAB.1.5.res" Perfs with 0% overlap

- The correlation factor of c : $F_\sigma(c) = \sum_{c'=1}^{K} 1_{\{C(c|c') \geq \sigma\}}$.

concepts	$F_\sigma(.)$	# initial set	# enriched set	perfs % (before enrichment)	perfs % (after enrichment)
wolf	8	103	1325	2.08	4.55
deer	3	149	149	47.27	50.00
airplane	8	413	644	47.22	48.89
apron	**9**	**173**	**318**	**8.00**	**12.50**
bathtub	13	172	2794	8.33	14.29
computer	1	142	142	14.63	17.65
cup	**6**	**80**	**80**	**0**	**13.04**
drum	7	369	1405	8.51	16.28
fork	1	150	150	0	4.17
helmet	**32**	**176**	**1156**	**11.11**	**18.80**
keyboard	25	177	770	3.70	6.67
kitchen	**8**	**174**	**433**	**10.71**	**17.65**
microphone	**8**	**187**	**769**	**10.00**	**21.43**
mirror	12	304	8928	2.04	3.39
motorcycle	**8**	**170**	**340**	**43.33**	**75.00**
painting	**10**	121	121	10.53	16.67
piano	**7**	**70**	**469**	**7.41**	**14.29**
ramp	17	308	4143	0.82	1.18
shirt	**10**	**188**	**1098**	**48.98**	**58.77**
stadium	1	68	68	28.00	35.29
towel	19	148	148	9.09	13.64
tractor	13	280	2219	12.90	15.38
tray	**18**	**91**	**650**	**0**	**10.00**
vase	22	171	1134	0	3.03
vest	1	75	75	3.70	7.09
wall	**2**	**79**	**248**	**66.67**	**95.00**
mouth	1	227	227	50.00	75.36
foot	**11**	**302**	**550**	**12.50**	**32.76**
arm	**18**	**159**	**2361**	**61.11**	**77.94**
newspaper	1	163	163	6.67	9.09
book	8	126	126	10.00	12.50
beer	19	141	195	9.09	16.67
ribbon	6	331	331	3.23	10.00
valley	2	313	313	17.74	21.43
male_child	**26**	**75**	**1152**	**11.78**	**20.00**

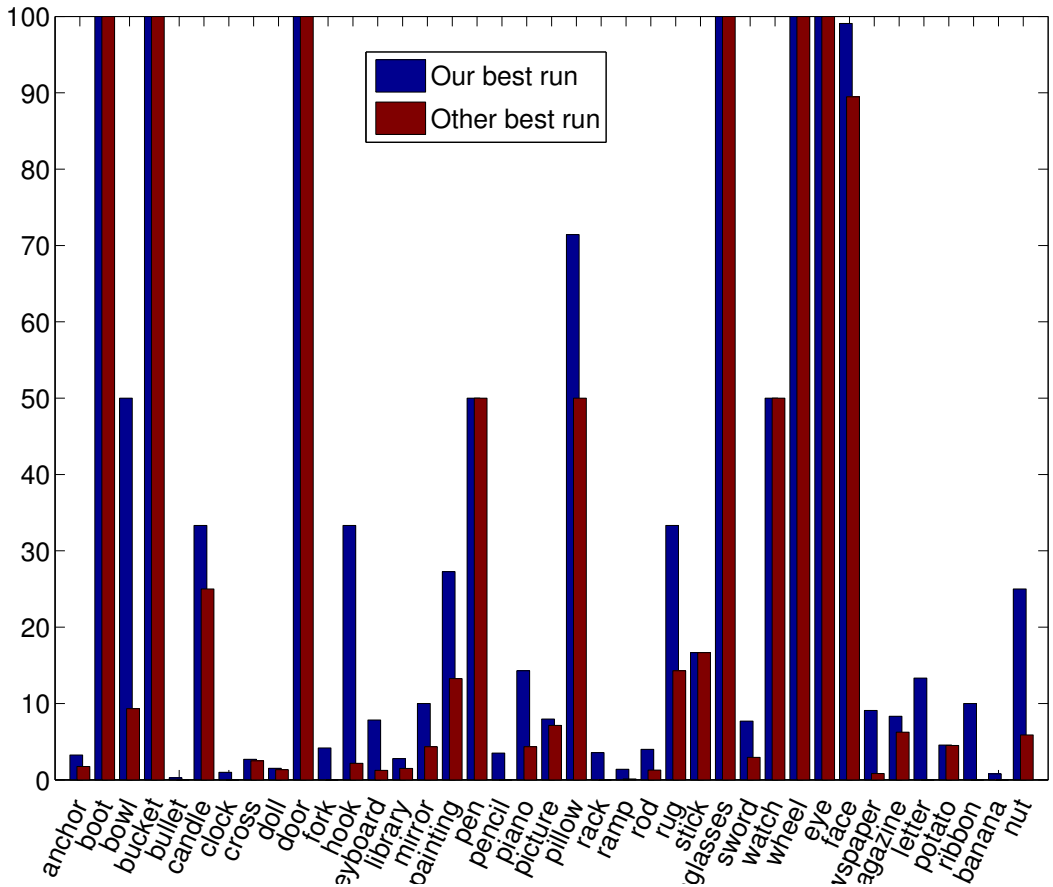
"TAB.1.1.res" vs. "TAB.1.5.res" Perfs with 50% overlap

concepts	$F_\sigma(.)$	# initial set	# enriched set	perfs % (before enrichment)	perfs % (after enrichment)
cat	**8**	**173**	**679**	**13.64**	**22.22**
deer	3	149	149	21.82	22.92
fish	**10**	**148**	**1562**	**6.25**	**18.18**
airplane	8	413	644	40.28	46.67
apron	**9**	**173**	**318**	**4.00**	**12.50**
bathtub	13	172	2794	8.33	14.29
bottle	28	115	5489	0	0.45
box	30	90	338	0	3.12
computer	1	142	142	3.66	3.92
cup	6	80	80	0	0.39
drum	7	369	1405	4.26	6.98
helmet	32	176	1156	0	0.85
kitchen	**8**	**174**	**433**	**10.71**	**17.65**
motorcycle	**8**	**170**	**340**	**43.33**	**75.00**
necktie	9	485	15719	0	0.31
picture	6	236	319	2.17	2.70
pillow	8	297	2023	0	0.90
ramp	17	308	4143	0.82	1.18
scarf	16	151	1018	0.61	1.32
shirt	**10**	**188**	**1098**	**12.24**	**20.18**
shoe	8	356	6901	0	0.40
stadium	1	68	68	26.00	35.29
stick	9	232	6042	0	0.27
tractor	13	280	2219	3.23	3.85
train	3	102	102	9.09	9.30
tray	**18**	**91**	**650**	**0**	**10.00**
vest	1	75	75	0	1.57
wheel	9	169	753	0	0.56
ear	23	58	593	0	0.38
head	10	378	378	4.94	8.06
book	8	126	126	0	0.51
letter	20	271	4560	0	0.28
valley	2	313	313	9.68	14.29
femalechild	12	79	79	2.30	3.85
male_child	**26**	**75**	**1152**	**2.17**	**5.00**

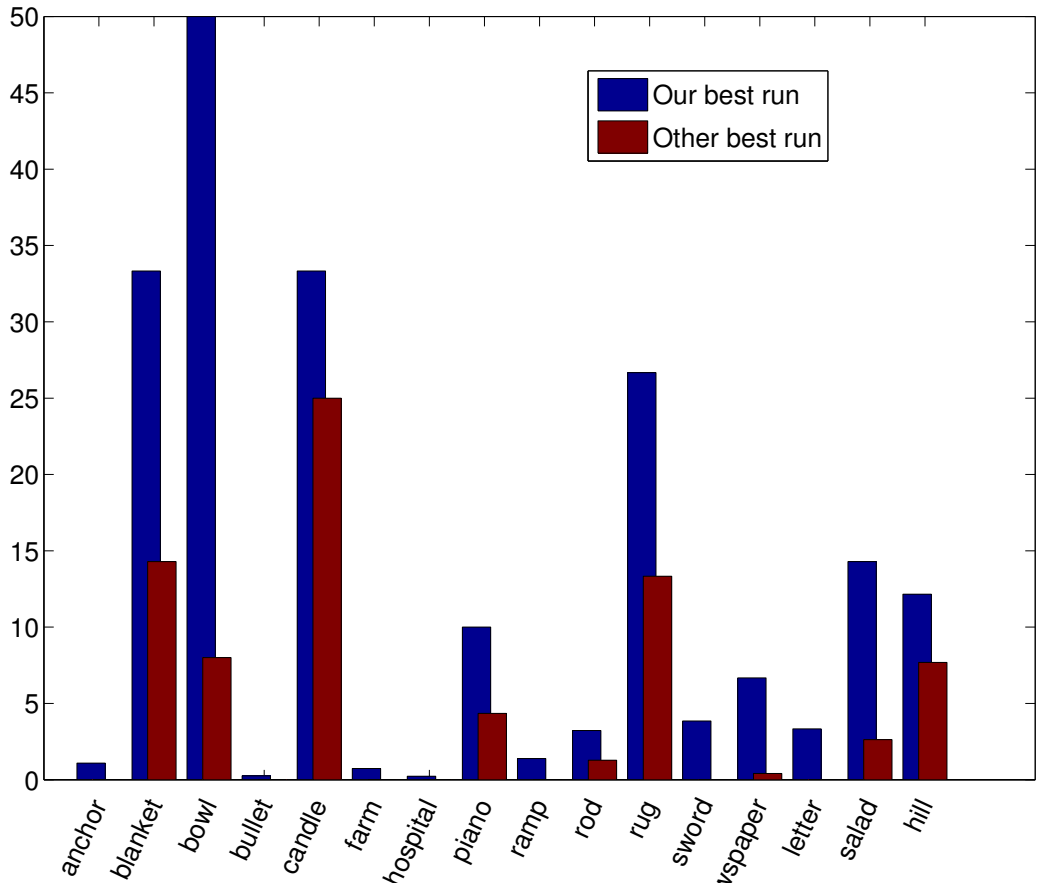
Remarks

- Better concept detection (at 0% overlap) implies better localization (at 50% overlap), even with simple (blind) localization heuristics.
- Clear gain especially for concepts with a high correlation factor $F_\sigma(c)$: as these concepts (such as "arm", "shirt", "shoe") co-occur with many other concepts and hence inherit larger training subsets.
- Some concepts even with small correlation factors (such as "apron", "cup") also benefit from the enrichment process, with a relatively smaller gain.
- Concepts which are usually centered in pictures (such as "motorcycle", "kitchen", "shirt") are relatively well localized using our simple blind localization.
- Difficult concepts (such as "cat") get substantial improvement.
- "TAB.1.5.res" is better than "TAB.1.3.res" as the former is more conservative (σ is high).

Best Performing Concepts (at 0% overlap)



Best Performing Concepts (at 50% overlap)



Outline

Conclusion and Extensions

- Submitted runs are based on SVMs built on top of enriched training sets.
- This enrichment makes it possible to share images between concepts and allows us to enhance the performances of concepts with scarce training images.
- Observed results also show that the correlation factor of a given concept has an impact on the resulting performances after enrichment.
- Future extensions : (i) how to make concept localization non-blind and also coupled with concept detection, consider dependency statistics (not only for concepts but also their locations). (ii) How to find the expansion sets $\mathcal{V}()$ automatically. (iii) How to use label enrichment as a post-processing step.