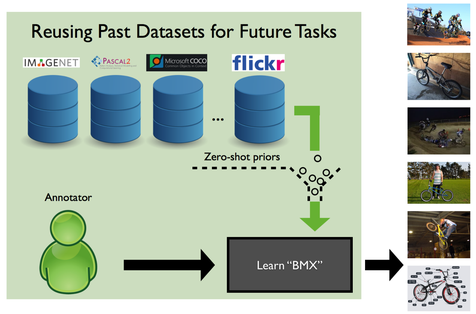
Active Transfer Learning with Zero-Shot Priors: Reusing Past Datasets for Future Tasks

E. Gavves[1], T. Mensink[2], T. Tommasi[3], C.G.M. Snoek[1], T. Tuytelaars[4]

[1]QUVA Lab, University of Amsterdam, [2]University of Amsterdam, [3]UNC Chapel Hill, [4]KU Leuven

Motivation

- Existing image datasets contain much more information than what is conveyed by their labels.
- Need to search for new manually annotated images when learning the visual model of a new class.
- Avoid discarding useful information and make the learning process as cheap as possible.



transfer learning
- new task: small amount of annotated data
- discover relation between old and new tasks

zero-shot learning
- new task: no annotated data
- known relation between old and new tasks

active learning
- new task: no annotated data

Active Learning

- Unlabeled samples $\{\boldsymbol{x}_i\}_{i=0}^N \in \mathbb{R}^d$ belonging to $\mathcal{C}$ classes. Binary setting $y_i \in \{-1, +1\}$.
- Model vector: $\boldsymbol{w} \in \mathbb{R}^d$.

Prediction score: $f(\boldsymbol{x}) = \boldsymbol{w} \cdot \boldsymbol{x}$.
Final annotation: $y = sign(f(\boldsymbol{x}))$.

- $\gamma_i^t \in \{0, 1\}$ indicates whether at time step t the label y_i has been queried. B: annotation budget.

$$\max_{\alpha^t, \gamma^t} \sum_i \gamma_i^t \lambda_i^t \alpha_i^t - \frac{1}{2} \sum_{i,j} \alpha_i^t \alpha_j^t \gamma_i^t \gamma_j^t y_i y_j \boldsymbol{x}_i \cdot \boldsymbol{x}_j$$

$$\text{s.t.} \quad \sum \gamma_i^t \alpha_i^t y_i = 0$$
$$0 \leq \alpha_i^t \leq C, \ \forall i \ ,$$
$$\gamma_i^t \geq \gamma_i^{t-1}, \ \forall i \ ,$$
$$\sum_i \gamma_i^t = \sum_i \gamma_i^{t-1} + B \ .$$

Conditions & Priors

Maximum Conflict
Query the sample i^* such that *(a)* its label y_{i^*} has an opposite sign from its classification score at $(t-1)$, while *(b)* the classifier score is as high as possible.

Label Equality
The number of positive and negative examples in the training set should be balanced, *i.e.* $\sum_i \gamma_i^t [y_i = 1] = \sum_i \gamma_i^t [y_i = -1]$.

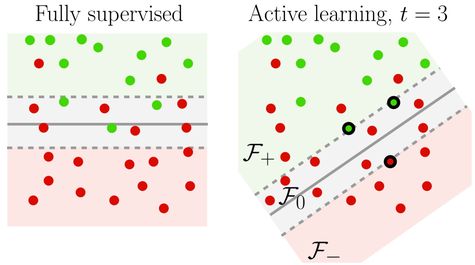
Zero-Shot Priors
Given a set of known concepts $\mathcal{K}$, the zero-shot prediction for a new class can be written as

$$f^{zs}(\boldsymbol{x}) = \sum_{k \in \mathcal{K}} \beta_{ck} \, \boldsymbol{w}_k \cdot \boldsymbol{x}_i \ ,$$

We modify the prediction score of active learning

$$f^t(\boldsymbol{x}) = \eta^t f^{zs}(\boldsymbol{x}) + \boldsymbol{w}^t \cdot \boldsymbol{x}.$$

Query Sampling



- At each iteration t compute the likelihood of sampling positive and negative labels from $\mathcal{F}_+$ and $\mathcal{F}_0$:

$$p(l|\mathcal{F}_+, t) \sim p(l|\mathcal{F}_+, t-1) + \frac{\sum_{r=1}^{t-1} [l_r = 1] \cdot [\boldsymbol{x}^r \in \mathcal{F}_+^r]}{t-1},$$

- Normalize the probabilities to sum up to one, and measure the label equality

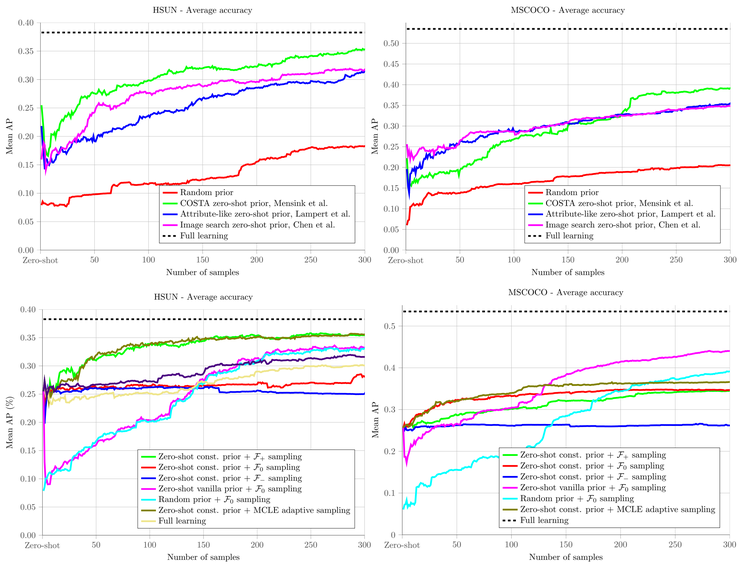
$$\rho^{t-1} = \frac{\sum_{r=1}^{t-1} [l_r = 1]}{\sum_{r=1}^{t-1} [l_r = 1] + [l_r = -1]} \ .$$

- Sample such that

$$\rho^{t-1} < \rho' \Rightarrow \boldsymbol{x}_t \sim \mathcal{F}_+^{t-1} \ ,$$
$$\rho^{t-1} \geq \rho' \Rightarrow \boldsymbol{x}_t \sim \mathcal{F}_0^{t-1}$$

Results

Datasets
HSUN: multi-class, multi-label dataset, covering object and scene categories. 107 classes, 8634 images, split into training/test set of 4367/4317 images.
MS-COCO: multi-class, multi-label dataset containing 80 object categories. 123,287 images in total. We used the full dataset and a subset with 4000 images in training and test.

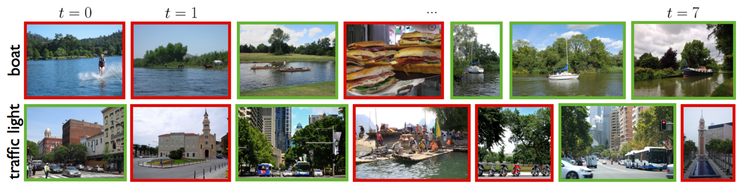


Choosing the Zero-Shot Prior
COSTA: object co-occurrence statistics;
Attribute-like: binarized class-to-class relations from class attributes;
Image search: model trained with the first 12 images returned by Google image search as positive examples.

Maximum Conflict - Label Equality
constant prior: $\eta^t = 1, \forall t$
vanilla prior: $\eta^0 = 1, \eta^{t \neq 0} = 0$
Our method automatically adapts to the different conditions providing always higher or equal results than selecting only from F_+ or F_0.

	HSUN (All samples: 0.383 mAP)							*Small COCO (All samples: 0.460 mAP)*						
No. of queries	0	50	100	150	200	250	300	0	50	100	150	200	250	300
MCLE (Dataset prior)	0.255	0.315	0.337	0.348	0.346	0.355	0.361	0.250	0.350	0.381	0.383	0.444	0.448	0.460
MCLE (External prior)	0.270	0.289	0.327	0.341	0.336	0.348	0.358	0.197	0.293	0.391	0.427	0.436	0.442	0.457
BBAL [1]	0.158	0.241	0.276	0.309	0.322	0.328	0.325	0.168	0.283	0.335	0.364	0.380	0.395	0.408
Hiearchical Sampling [3]	0.089	0.156	0.199	0.221	0.234	0.230	0.246	0.076	0.182	0.250	0.287	0.309	0.331	0.365
GP Mean [4]	0.154	0.282	0.319	0.340	0.350	0.361	0.365	0.186	0.344	0.394	0.412	0.421	0.431	0.438
GP Variance [4]	0.154	0.201	0.206	0.216	0.226	0.240	0.244	0.186	0.233	0.263	0.284	0.291	0.309	0.326
GP Impact Bayes [2]	0.154	0.251	0.286	0.305	0.316	0.327	0.345	0.186	0.298	0.346	0.393	0.417	0.430	0.436
GP EMOC Bayes [2]	0.154	0.277	0.310	0.320	0.328	0.332	0.337	0.186	0.336	0.375	0.388	0.397	0.399	0.405

Comparison with state-of-the-art active learning methods. For HSUN the external prior comes from COSTA learned on MSCOCO, whereas for MSCOCO the external COSTA prior is learned on HSUN. For Small MSCOCO, MCLE reaches the full mAP within 300 samples, even with external priors.



Active learning queried samples on MSCOCO with HSUN zero-shot priors. Positive/Negative samples are in green/red. Observe the balanced label distribution, while the visual similarity reveals the maximum conflict selection of the next sample.

References

[1] S. Vijayanarasimhan, et al. Far-sighted active learning on a budget for image and video recognition. CVPR, 2010.
[2] A. Freytag, et al. Selecting influential examples: Active learning with expected model output changes. ECCV, 2014.
[3] S. Dasgupta and D. Hsu. Hierarchical sampling for active learning. ICML, 2008.
[4] A. Kapoor et al. Gaussian processes for object categorization. IJCV, 2010.