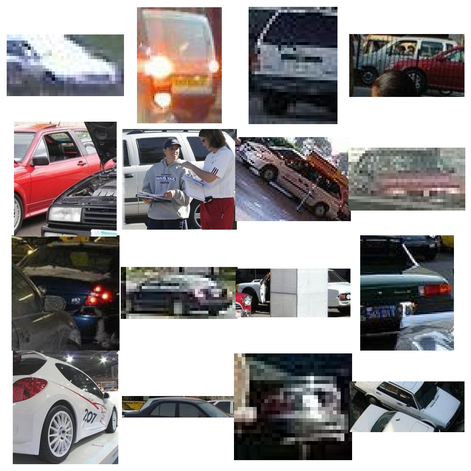
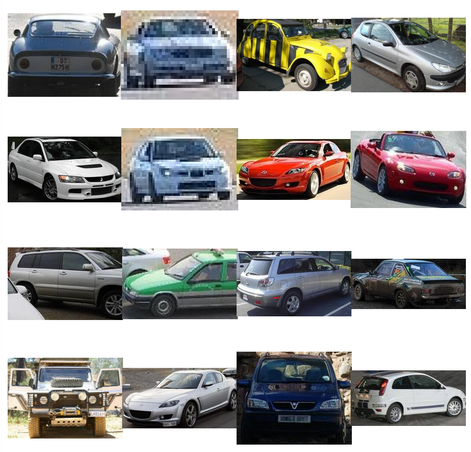
Datasets, Sample Selection, and Test Performance

Omid Aghazadeh Stefan Carlsson
Computer Vision and Active Perception laboratory, Stockholm, Sweden

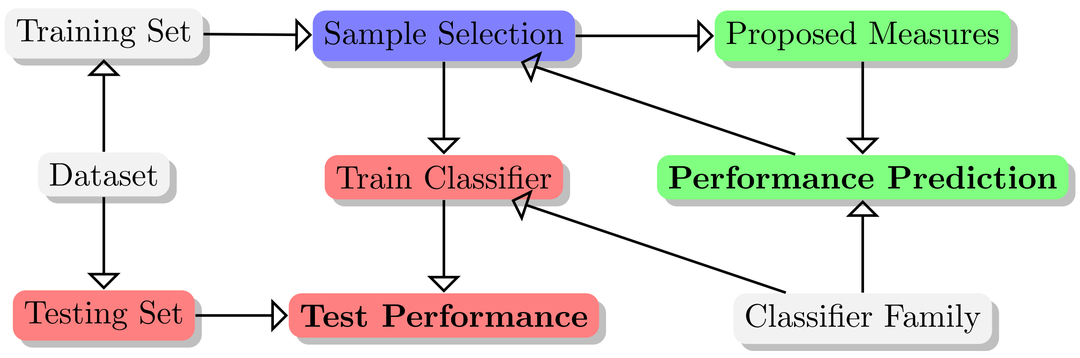
Introduction

- Quality of the training set → Test performance of classifiers

Low Quality Training Set	High Quality Training Set
	

Correspondences are much clearer in the 'High Quality' set

- This work discusses sample selection for ensuring a desired (quantified) quality of the training set



Training Data and Test Performance

- Traditional training+test procedure for class $\mathcal{C}$ and classifier family $\mathcal{M}$:

$$AP_{\mathcal{M}}^{(\mathcal{C})} = \tau\left(M(\mathcal{C}_{TR}), \mathcal{C}_{TST}\right)$$

- Semantics of the first order data describing measures proposed in [1]:

Measure	Scale	Semantic	Measure	Scale	Semantic
μ_L	Local	Connectivity	μ_S	Semi-Global	Lack of Variation
μ_G	Global	Intra-Class Variation	μ_P	Global	Connected Variation

- Let $\mu^{(\mathcal{C})}$ be a vector of some measures describing a (training/testing) set $\mathcal{C}$:

$$AP_{\mathcal{M}}^{(\mathcal{C})} \approx f_{\mathcal{M}}\left(\mu^{(\mathcal{C}_{TR})}, \mu^{(\mathcal{C}_{TST})}\right)$$

- When $\mu^{(\mathcal{C}_{TR})} \approx \mu^{(\mathcal{C}_{TST})}$

$$AP_{\mathcal{M}}^{(\mathcal{C})} \approx \tilde{f}_{\mathcal{M}}\left(\mu^{(\mathcal{C}_{TR})}\right)$$

$\tilde{f}_{\mathcal{M}}(.)$ *quantifies the quality of the training data*.

- Implications:
  1. Possible to model and analyze the behavior of different families of classifiers as a function of various aspects of the training data
  2. **Improve the quality of the training set → Improve the test performance**

Sample Selection

- Given a desired criterion $g(\mu)$, search for $\mathcal{S} \subseteq \mathcal{C}$ which optimizes

$$s(\mathcal{S}) = g\left(\mu^{(\mathcal{S})}\right)$$

- Combinatorial → Resort to greedy optimization
- Consider two types of problems:
  1. *Fixed Cardinality*:

$$\mathcal{S}_F = \arg\max_{\mathcal{S}\subseteq\mathcal{C}} s(\mathcal{S}) \quad \text{s.t.} \quad |\mathcal{S}| = n_f$$

  2. *Largest Set*:

$$\mathcal{S}_L = \arg\max_{\mathcal{S}\subseteq\mathcal{C}} |\mathcal{S}| \quad \text{s.t.} \quad s(\mathcal{S}) \geq \tau$$

Subsets that Maximize the Predicted Test Performance

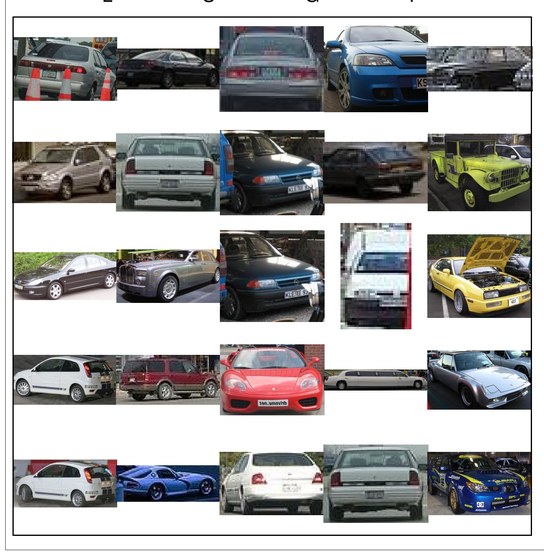
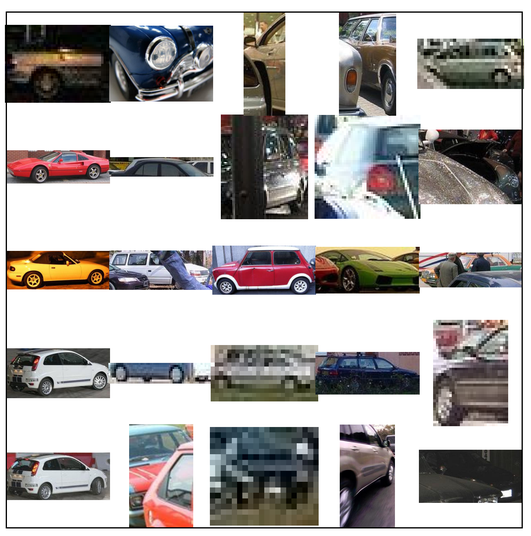
- The test set cannot be visited → have to assume $\mu^{(\mathcal{C}_{TST})} \approx \mu^{(\mathcal{C}_{TR})}$
- Simplifying assumption: $\mu^{(\mathcal{S})} \approx \mu^{(\mathcal{C}_{TR})}$, *i.e.* small modifications to the training set
- *Largest Set* + constraint: the predicted test performance should improve by ϵ_P:

$$\tau = \tilde{f}_{\mathcal{M}}\left(\mu^{(\mathcal{C}_{TR})}\right) + \epsilon_P$$
$$g\left(\mu^{(\mathcal{S})}\right) = \tilde{f}_{\mathcal{M}}\left(\mu^{(\mathcal{S})}\right)$$

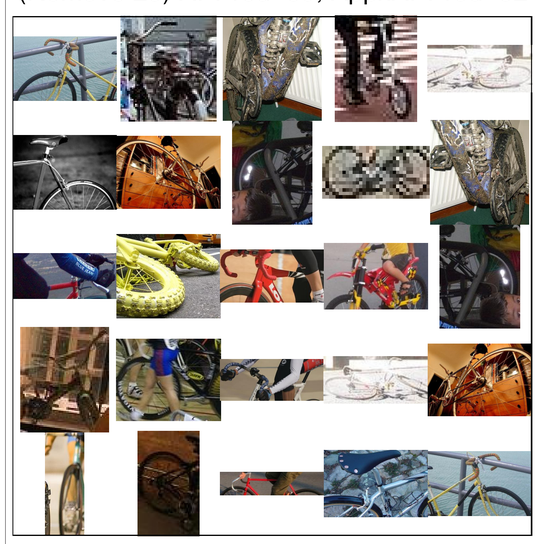
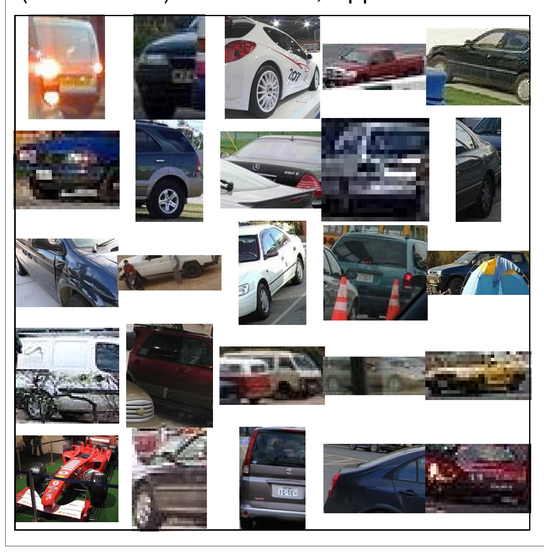
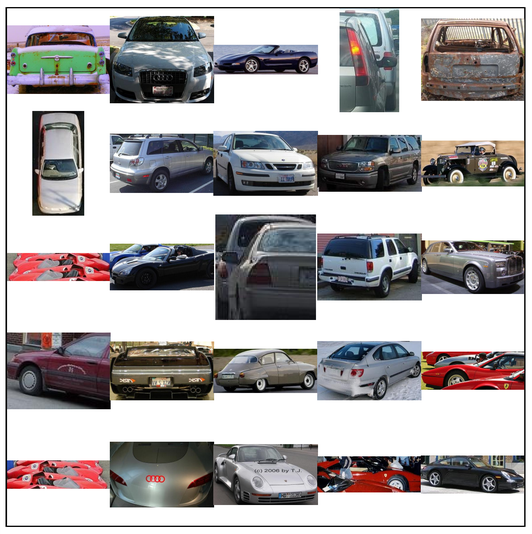
- Large ϵ_P invalidates our simplifying assumption.
- The assumption can be avoided by using a richer estimator $f_{\mathcal{M}}\left(\mu^{(\mathcal{C}_{TR})}, \mu^{(\mathcal{C}_{TST})}\right)$

Experiments

- (Fixed Cardinality) μ_P on Pascal VOC 2007

High μ_P	Low μ_P
mu_L=92, mu_S=44, mu_G=29, mu_P=655	mu_L=32, mu_S=14, mu_G=85, mu_P=96
	

- (Largest Set) Predicted AP on Pascal VOC 2007

(Worst Outliers) Remove to gain 1% AP	(Best Inliers) Remove to loose 1% AP
(Remove 25) APPred=59, AppxAPPred=52	(Remove 15) APPred=58, AppxAPPred=50
	
(Remove 97) APPred=56, AppxAPPred=55	(Remove 72) APPred=55, AppxAPPred=53
	

- Qualitatively, most gross outliers are either:
  1. significantly truncated
  2. significantly occluded
  3. taken from a significantly low quality image, are noisy or too small
  4. captured from viewpoints without enough "support" in the training set.

  The latter is related to photographer and selection biases discussed in [2].

Conclusion

- Modelling test performance as a function of the training data enables us to devise rules for selection of training data.
- Exemplars without enough "support" in the training set decrease global connectivity μ_P, which was shown to be strongly correlated to the test performance [1]. Removing such exemplars can be expected to make the data 'cleaner' [3] thereby improving test performance.
- "Big Data" is not necessarily connected. Connectivity of the training data seems to be too significant to ignore.

References

O. Aghazadeh and S. Carlsson. Properties of Datasets Predict the Performance of Classifiers, BMVC, 2013.
A. Torralba and A. A. Efros. Unbiased Look at Dataset Bias, CVPR, 2011.
X. Zhu, C. Vondrick, D. Ramanan and C. C. Fowlkes. Do we need more training data or better models for object detection? BMVC, 2012.

This work was supported by the Swedish Foundation for Strategic Research and by the European Commission KIC: EIT ICT labs.