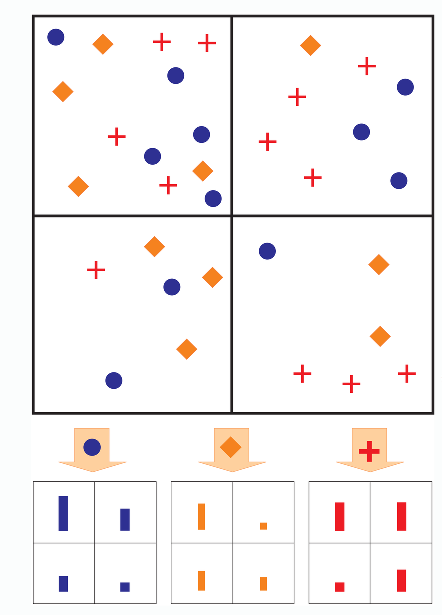
Learning Smooth Pooling Regions for Visual Recognition

Mateusz Malinowski and Mario Fritz

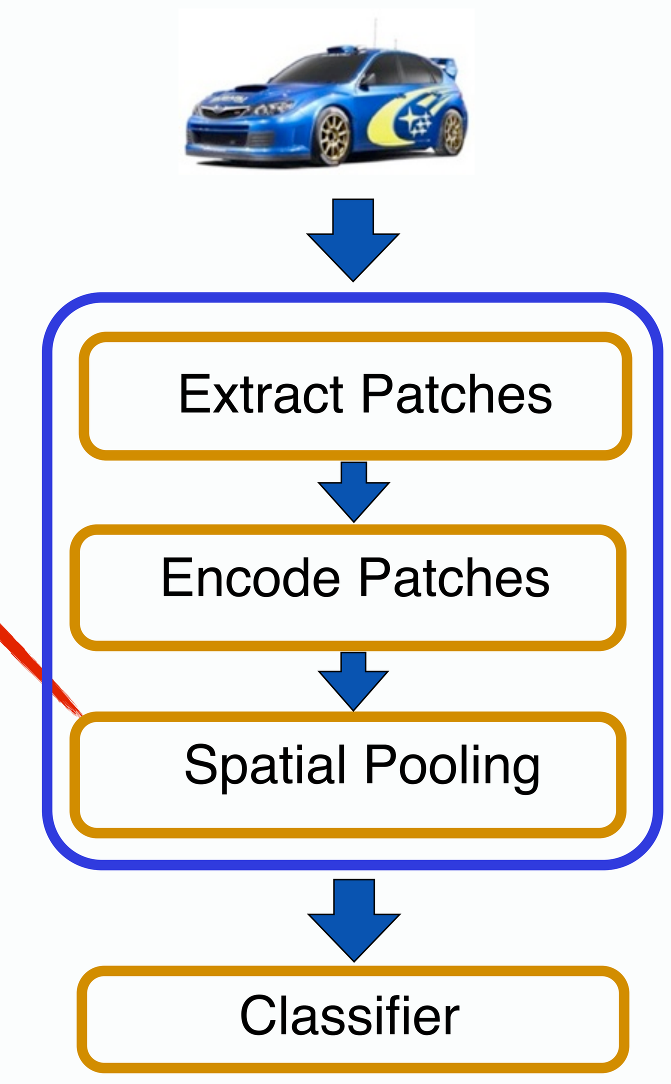
Motivation

- State-of-the-art object recognition algorithms are based on histograms of feature representations
- Spatial Pooling, in order to preserve some spatial information, aggregates statistics locally
- Current Spatial Pooling schemes are hand-crafted (e.g. SPM)

- Are such spatial regions optimal?
- Can we train jointly both the classifier and spatial regions?
- What assumptions on the Spatial Pooling scheme are needed to achieve best performance?



Example of hand-crafted pooling scheme [1]



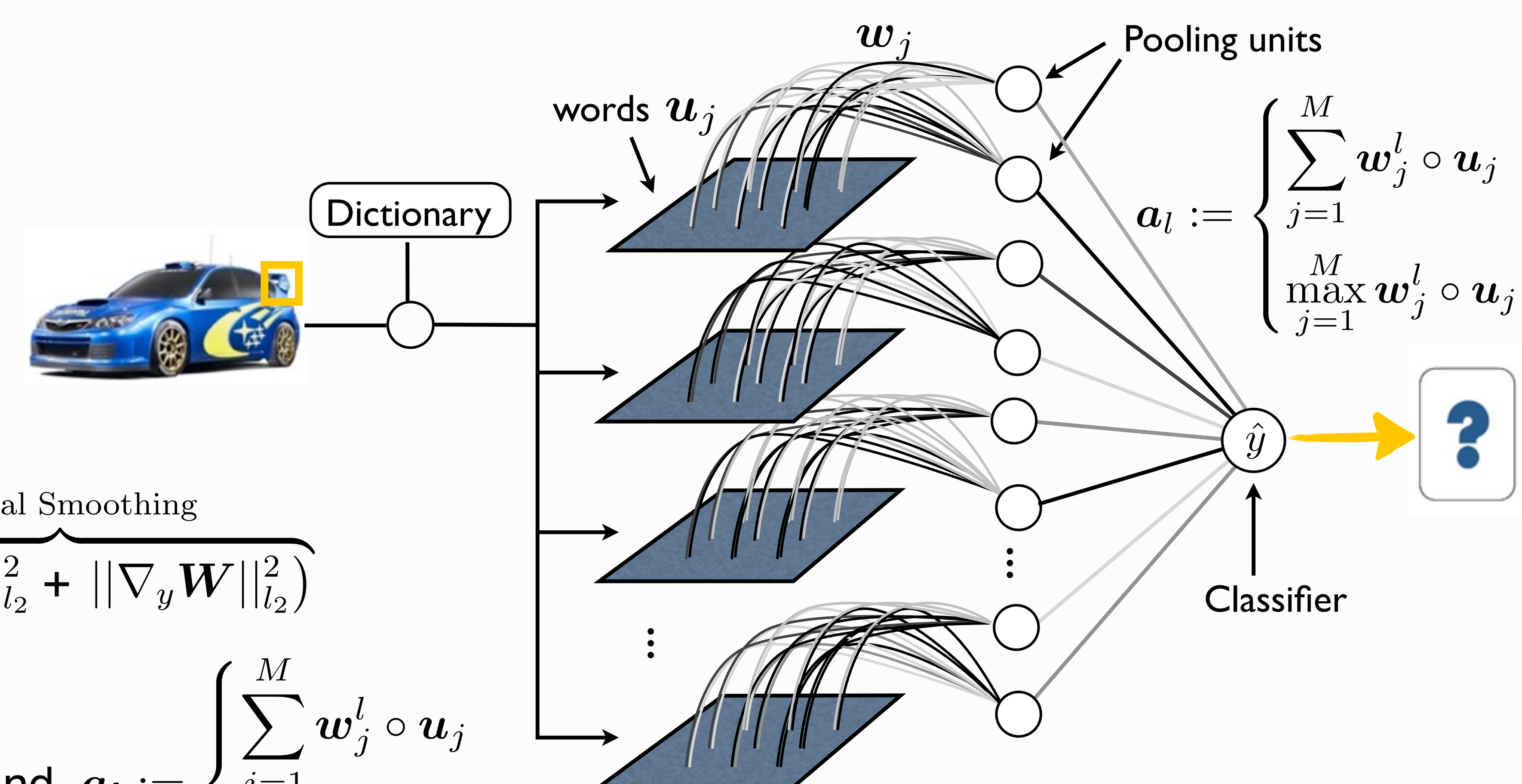
Our method

- Parameterized pooling operator
- Joint training of classifier and pooling regions
- Efficient and parallel approximation training
- Logistic regression as a classifier
- Our optimization problem:

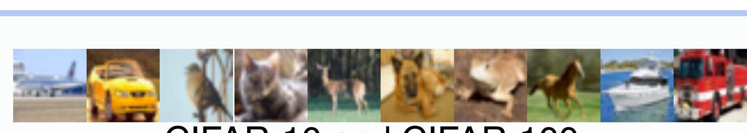
$$\text{minimize}_{\boldsymbol{W},\boldsymbol{\Theta}} \quad J(\boldsymbol{\Theta}) + \frac{\alpha_1}{2}\|\boldsymbol{\Theta}\|_{l_2}^2 + \frac{\alpha_2}{2}\|\boldsymbol{W}\|_{l_2}^2 + \frac{\alpha_3}{2}\overbrace{\left(\|\nabla_x \boldsymbol{W}\|_{l_2}^2 + \|\nabla_y \boldsymbol{W}\|_{l_2}^2\right)}^{\text{Spatial Smoothing}}$$

$$\text{subject to } \boldsymbol{W} \in [0,1]^{K\times M\times L}$$

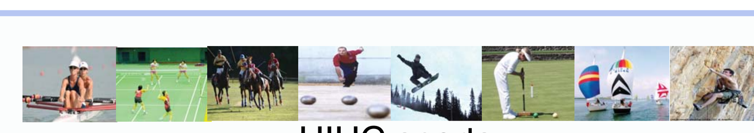
Where $J(\boldsymbol{\Theta}) := -\frac{1}{D}\sum_{i=1}^{D}\sum_{j=1}^{L} \mathbf{1}\{y^{(i)}=j\}\log p(y^{(i)}=j|\boldsymbol{a}^{(i,2)};\boldsymbol{\Theta})$ and $\boldsymbol{a}_l := \begin{cases} \sum_{j=1}^{M} \boldsymbol{w}_j^l \circ \boldsymbol{u}_j \\ \max_{j=1}^{M} \boldsymbol{w}_j^l \circ \boldsymbol{u}_j \end{cases}$



Results



CIFAR-10 and CIFAR-100

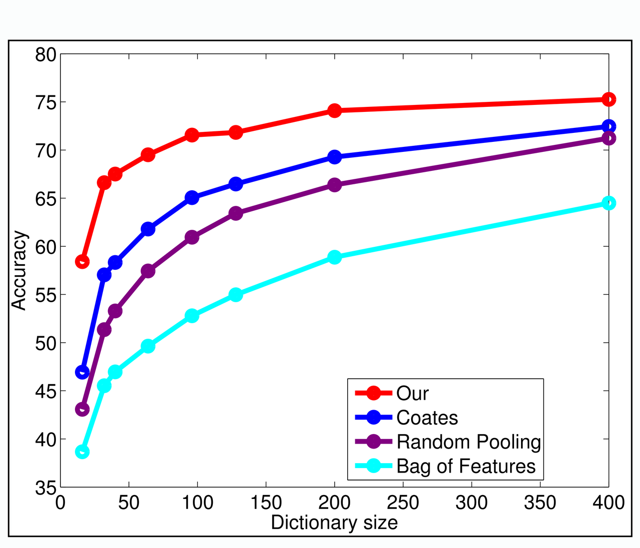


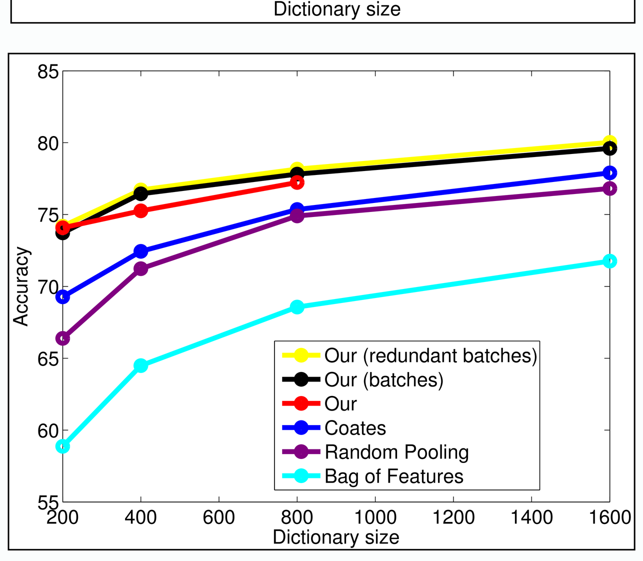
UIUC sports

- Evaluation on Object and Event recognition tasks
- Hand-crafted Spatial Pooling as a baseline[4, 5]
- Strong improvement over hand-crafted Spatial Pooling[4, 5] 3% on Event and up to 10% on Object recognition
- State-of-the-Art on CIFAR-100 given SPM

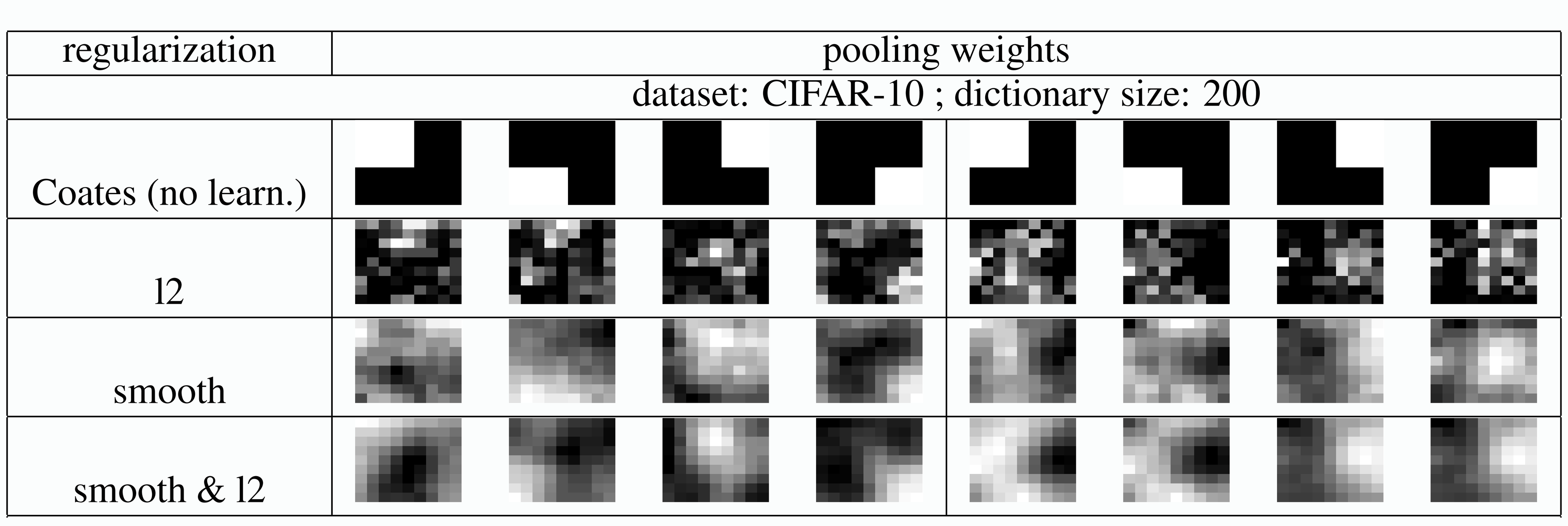
Conclusion

- Importance of learnt Spatial Pooling regions
- Scalable algorithm for larger dictionaries
- Discovery of new pooling schemes
- Importance of Spatial Smoothness prior
- Applicable to sum- and max-pooling





CIFAR-10 dataset



Learnt pooling regions

Regularization	CV Acc.	Test Acc.
free	68.48%	69.59%
l2	67.86%	68.39%
smooth	73.36%	73.96%
l2 + smooth	70.42%	70.32%

CIFAR-10; dictionary size 200

	UIUC sports
Object Banks + SPM [5]	76.3%
Object Banks + our method	79.4%

Event recognition with object banks

Method	Dict. size	Features	Acc.
Jia	1600	6400	80.17%
Coates	1600	6400	77.9%
Our (batches)	1600	6400	79.6%
Our (redundant)	1600	12800	80.02%

Object recognition on CIFAR-10

Method	Dict. size	Features	Acc.
Jia	1600	6400	54.88%
Coates	1600	6400	51.66%
Our (batches)	1600	6400	56.29%

Object recognition on CIFAR-100

Source	Target	Acc.
CIFAR-10	CIFAR-100	52.86%
CIFAR-100	CIFAR-10	80.35%

Results of transfer of learnt pooling regions

1.Lazebnik, S., Schmid, C., Ponce, J.: Beyond bag of features: Spatial pyramid matching for recognizing natural scene categories. CVPR 2006.
2.Yang, J., Yu, K., Gong Y., Huang T.: Linear spatial pyramid matching using sparse coding for image classification. CVPR 2009.
3.Jia, Y., Huang, C.: Beyond spatial pyramid: Receptive field learning for pooled image features. NIPS Workshop on Deep Learning 2011.
4.Coates, A., Ng, A.: The importance of encoding versus training with sparse coding and vector quantization. ICML 2011.
5.Li, L., Su. H., Xing E., Fei-Fei, L.: Object Bank: A high-level image representation for scene classification & semantic feature sparsification. NIPS 2010.

max planck institut
informatik

UNIVERSITÄT DES SAARLANDES