Hands on Advanced Bag-of-Words Models for Visual Recognition

Lamberto Ballan and Lorenzo Seidenari

MICC - University of Florence

Conclusion

- Final Remarks
  - How BoW models have evolved over time

- Implementation and practical details
  - Sampling and coding
  - Learning

- Open problems
  - Deep Learning *vs* “Feature Engineering”
  - Dataset Bias

BoW evolution

- After seminal work Video Google from *Sivic et al. 2003*, visual BoW models have drifted from their textual counterpart
- Spatial Pyramid Matching has been a major improvement in recovering the lost global information [*Lazebnik et al. 2006*]
- Other less rigid pooling schemes proved successful like Object centric pooling by *Russaskovsky et al. 2012* and Deformable spatial pyramid matching by *Kim et al. 2013*

BoW evolution

- Soft assignment technique inspired by kernel density estimation proposed to assign a feature to more than one word [*van Gemert et al. 2008*]

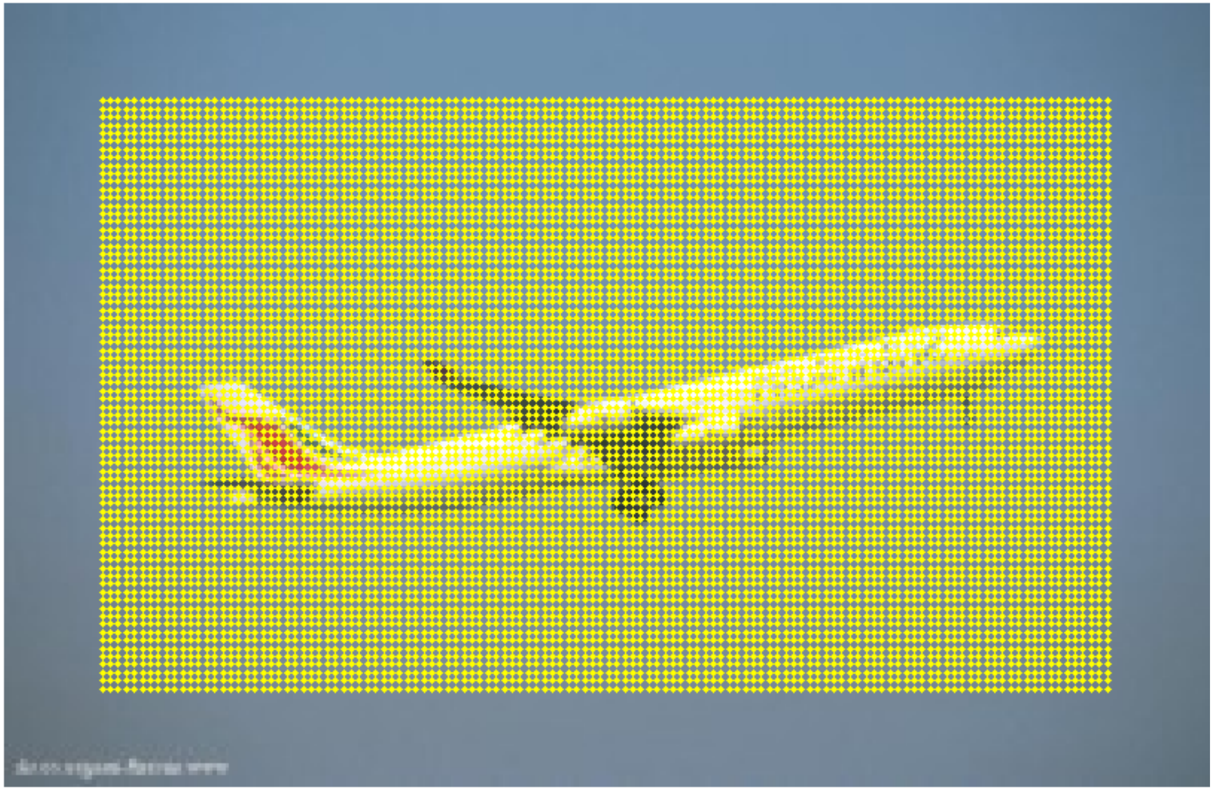
- Coding/Reconstruction based approaches have recently became popular
  - Local Linear Coding
  - Sparse coding
  - Truncated Soft Assignment
  - Fisher Vectors

- In all these approaches there is no more a unique feature word direct correspondence

- See [*Chatfield et al 2011*] for a comparison of recent coding, pooling and sampling techniques for BoW systems

Sampling

- Multi-scale **dense** sampling of **unoriented** SIFT descriptors has proven to be the best choice in several benchmarks (PASCAL VOC, Caltech-101, Caltech-256, Scene-15, …)
- Chatfield have shown that 2px step sampling produce the best results



Coding

- Coding techniques that consider multiple words either via sparsity (ScSPM) or via locality (LLC) or by using improved dictionaries (FV) perform best

- All this techniques are not just «counting word occurences» but add some additional information tipically accounting for the «divergence» between image and dictionary features distributions

- LLC is faster then sparsity based techniques but slower than Fisher Vectors

- Fisher Vectors produce very high dimensional signatures 500K+ if using spatial pyramids

Image Classification with the Fisher Vector: Theory and Practice, Perronin et al., IJCV 2013
Locality-constrained Linear Coding for Image Classification, Wang et al. , CVPR 2010
Linear Spatial Pyramid Matching using Sparse Coding for Image Classification, Yang et al., CVPR 2009

SVM Practicum

- SVM score can be use to rank examples
  - Decision value is the distance from margin
  - Farther element predicted labels are more reliable

- Kernels are a great way to add user domain knowledge
  - Software packages allow to: add a kernel function or use a pre-computed kernel matrix
  - Pre-computing the kernel is often more efficient (and easier)

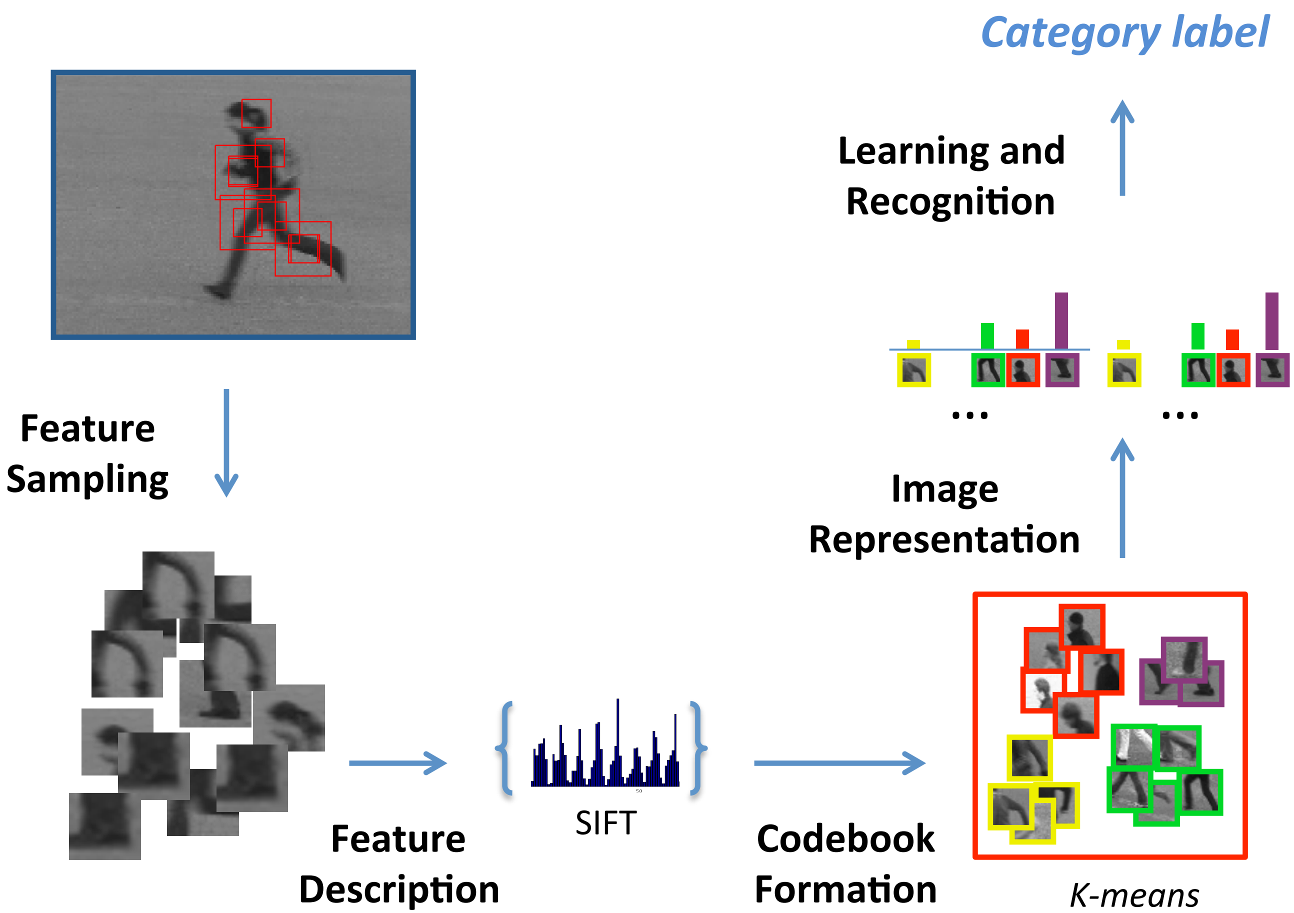
SVM Practicum

- Kernel evaluations are expensive
  - When original feature space is very high dimensional (100k+) use linear classifier
  - Linear classifier can be trained in linear time with iterative algorithms like SGD or dual coordinate descent
  - Feature embeddings can be approximated when not available (RBF, intersection)

- Software available online:
  - LibSVM and Liblinear: http://www.csie.ntu.edu.tw/~cjlin/
  - SGD: http://leon.bottou.org/projects/sgd

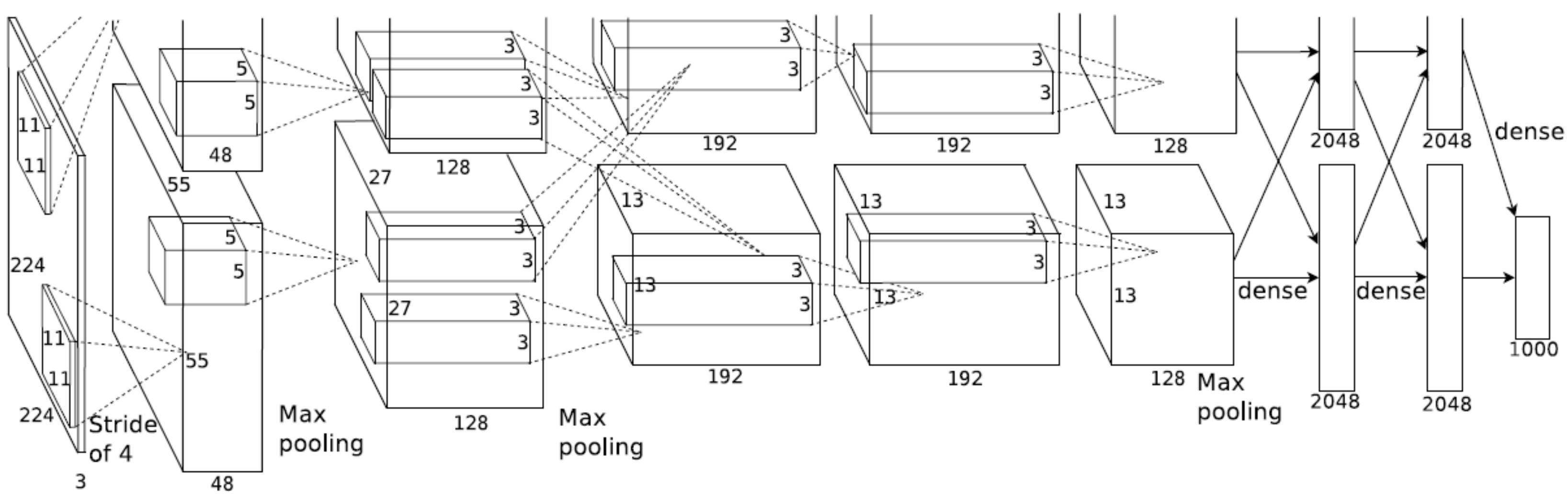
Deep Learning vs Feature Engineering

- You have learned how to engineer features using the BoW paradigm
- Roughly this can be seen as a filtering, coding and pooling stages followed by a supervised classification layer



Deep Learning vs Feature Engineering

- In deep learning the layers performing these operations are stacked by forming a deep architecture that learns the representation and the discriminative function at the same time
- This requires large amounts of data and more computational power (GPUs are the norm)
- Recently Krizhevsky et al. “aced” the ImageNet competition beating competing BoW methods based on SIFT and Fisher Vectors by 10% on Top-5 error



ImageNet classification with Deep Convolutional Neural Networks, Krizhevsky et al. NIPS 2012

Dataset Bias

- If you are working in image recognition this game is easy!
- This means that dataset are actually highly biased…
- This is a major issue that stands between academic work and real world systems

An unbiased look at dataset bias, Torralba et al. CVPR 2011

Dataset Bias

- When compiling a dataset some bias will be introduced inevitably:
  - Capture bias, e.g. all objects centered and portrayed at the same scale with no clutter
  - Negative Set bias. The negative set is astronomically large, dataset are restricted to sample a (proportionally) very small subset of it

- Here what happens when you try to cross-test learning algorithms:

task	Test on: / Train on:	SUN09	LabelMe	PASCAL	ImageNet	Caltech101	MSRC	Self	Mean others	Percent drop
"car" classification	SUN09	**28.2**	29.5	16.3	14.6	16.9	21.9	28.2	19.8	**30%**
	LabelMe	14.7	**34.0**	16.7	22.9	43.6	24.5	34.0	24.5	**28%**
	PASCAL	10.1	25.5	**35.2**	43.9	44.2	39.4	35.2	32.6	**7%**
	ImageNet	11.4	29.6	36.0	**57.4**	52.3	42.7	57.4	34.4	**40%**
	Caltech101	7.5	31.1	19.5	33.1	**96.9**	42.1	96.9	26.7	**73%**
	MSRC	9.3	27.0	24.9	32.6	40.3	**68.4**	68.4	26.8	**61%**
	Mean others	10.6	28.5	22.7	29.4	39.4	34.1	53.4	27.5	48%

An unbiased look at dataset bias, Torralba et al. CVPR 2011

Conclusion

- Today you have learned some of the fundamentals aspects of a BoW pipeline
- You are now able to implement a full visual recognition pipeline: **from the image pixels to the class label**
- We gave you a brief overview of the more recent evolution of these methods and a peek of other promising techniques
- BoW models are easy to understand and implement and can be employed as a first step in many computer vision tasks

References

Papers

- Distinctive Image Features from Scale-Invariant Keypoints, David G. Lowe, IJCV 2004
- Evaluating Color Descriptors for Object and Scene Recognition, van de Sande et al. TPAMI 2011.
- Video Google: A Text Retrieval Approach to Object Matching in Videos Sivic et al. ICCV 2003.
- Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories, Lazebnik et al., CVPR 2006.
- Object-centric spatial pooling for image classification, Russakovsky et al. ECCV 2012.
- Deformable Spatial Pyramid Matching for Fast Dense Correspondences, Kim et al. CVPR 2013.
- Kernel codebooks for scene categorization, van Gemert, ECCV 2008
- Image Classification with the Fisher Vector: Theory and Practice, Perronin et al., IJCV 2013.
- Locality-constrained Linear Coding for Image Classification, Wang et al. , CVPR 2010.
- Linear Spatial Pyramid Matching using Sparse Coding for Image Classification, Yang et al., CVPR 2009.
- An unbiased look at dataset bias, Torralba et al. CVPR 2011
- ImageNet classification with Deep Convolutional Neural Networks, Krizhevsky et al. NIPS 2012.
- The devil is in the details: an evaluation of recent feature encoding methods, Chatfield et al., BMVC 2011.

Software

- LibSVM and Liblinear: http://www.csie.ntu.edu.tw/~cjlin/
- SGD: http://leon.bottou.org/projects/sgd
- VLFEAT: http://www.vlfeat.org/