Unsupervised & Supervised Visual Codes with Restricted Boltzmann Machines

Hanlin Goh[1,2,3], Nicolas Thome[1], Matthieu Cord[1], Joo-Hwee Lim[2,3]

1. Laboratoire d'Informatique de Paris 6, UPMC – Sorbonne Universités, Paris, France
2. Institute for Infocomm Research, A*STAR, Singapore
3. Image and Pervasive Access Laboratory, CNRS UMI 2955, Singapore & France

✉ hlgoh@i2r.a-star.edu.sg, nicolas.thome@lip6.fr, matthieu.cord@lip6.fr, joohwee@i2r.a-star.edu.sg

ECCV 2012

Problem

Given a set of local descriptors extracted from images in a dataset, can we construct an accurate, small and fast visual codebook through unsupervised & supervised learning?

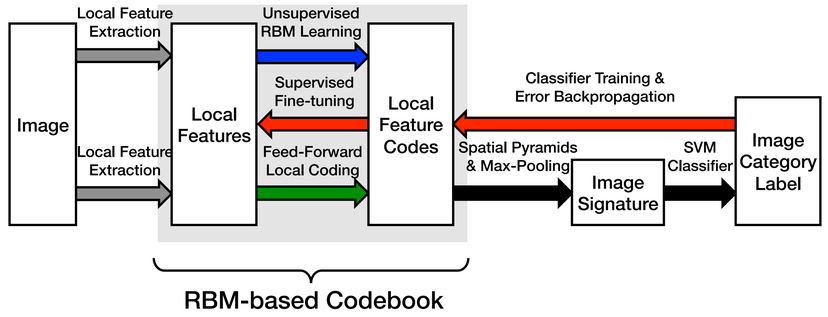
Previous Work

- *Non-Learned Assignment Coding*
  - Hard assignment [Lazebnik et al.]
  - Kernel codebooks [van Germert et al.]
  - Soft assignment [Liu et al.]
- *Sparse Coding*
  - ScSPM [Yang et al.]
  - LLC [Wang et al.]
  - SC & max pooling [Boureau et al.]
  - Multi-way local pool [Boureau et al.]
- *Restricted Boltzmann Machine (RBM)*
  - CDBN [Lee et al.]
  - Sparse RBM [Lee et al. / Sohn et al.]
  - CRBM [Sohn et al.]
- *Supervised Learning*
  - Discriminative codes [Boureau et al.]
  - LC-KSVD [Jiang et al.]

Overall Framework – Bag of Words

Our BoW architecture consists of five layers of representations, with the following operations:

1) Local feature extraction
2) Unsupervised RBM learning,
3) Supervised fine-tuning,
4) Low-level inference,
5) Spatial pooling, and
6) SVM training & classification.



RBM Regularization

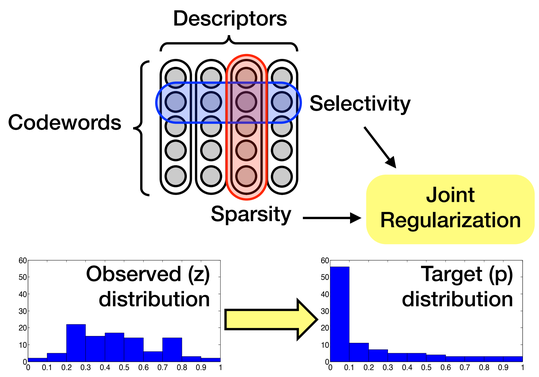
$$\underset{\{\mathbf{W},\mathbf{c},\mathbf{b}\}}{\arg\min} \; -\sum_{k=1}^{K} \log \sum_{\mathbf{z}} \Pr\left(\mathbf{x}_k, \mathbf{z}_k\right) + \lambda h(\mathbf{z})$$

RBM (max. likelihood approx.)

$$h(\mathbf{z}) = -\sum_{j=1}^{J}\sum_{k=1}^{K} p_{jk} \log z_{jk} + (1-p_{jk}) \log\left(1-z_{jk}\right)$$

Cross-Entropy Penalty (per descriptor & codeword)

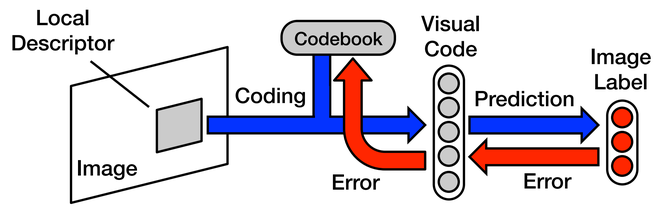
Regularize RBM with Selectivity and Sparsity



- Selectivity – each codeword should respond to only a small subset of input descriptors.
- Sparsity – each input descriptor should only have a small subset of codewords responding to it.
- We map the distributions of observed activations into long-tailed target distributions.
- Target matrix is jointly sparse and selective by mapping every column then remapping every row.
- Promotes diversity between codewords and discrimination between descriptors.

Local Supervised Fine-Tuning

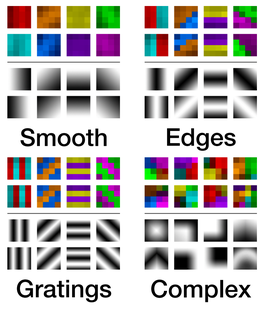
- Supervised learning is performed on the codebook initialized by the unsupervised reguarized RBM.
- The error backpropagation algorithm is used to fine-tune the local descriptor codebook using image labels.



Our Approach

- Train RBMs as visual codeboks.
- Regularize RBMs for desired coding – jointly selective & sparse for codebook conciseness.
- Fine-tune codebook with supervision using image labels.
- Evaluate on accuracy, codebook size and inference speed.

Visual Codewords Discovered



- Codewords capture local gradient structure of images
- Features are generic, diverse & discriminative

Image Categorization Results

	Unsupervised	Supervised
Caltech-101 (15 tr.)	70.2%	71.1%
Caltech-101 (30 tr.)	78.0%	78.9%
15 Scenes (100 tr.)	85.7%	86.0%

- Codebook size: 1024
- Local features: Macro features from SIFT

Performance Analysis & Summary

- We achieved **high accuracy**, reaching state-of-the-art among the family of feature coding methods using a single descriptor-type.
- Supervised fine-tuning improved performances slightly.
- The **codebooks are small** and concise, and codewords are diverse and discriminative.
- **Inference is fast** since RBMs by nature are encoders, unlike sparse coding where reoptimization is needed.