FINE-GRAINED IMAGE CLASSIFICATION USING FISHER VECTORS

Akata Z., Sánchez J., Perronnin F., - Xerox Research Centre Europe
{Author1, Author2, Author3}@xrce.xerox.com

Abstract

-Fine-grained visual classification (FGVC) aims at the fine distinction of specific image categories (e.g fungus)

-We motivate Fisher Kernel framework for FGVC and show experimentally that it yields excellent results

Fisher Kernel Framework

-Model a sample X by its deviation from a distribution u_λ:

$$G_\lambda^X = \nabla_\lambda \log u_\lambda(X).$$

-Measure similarity using the **Fisher Kernel**:

$$K(X,Y) = G_\lambda^{X\prime} F_\lambda^{-1} G_\lambda^Y \text{ with}$$

$$F_\lambda = E_{x\sim u_\lambda}\left[\nabla_\lambda \log u_\lambda(x) \nabla_\lambda \log u_\lambda(x)'\right]$$

Application to Images

$X = \{x_t, t = 1 \dots T\}$ is a set of T i.i.d D-dim local descriptors (*e.g.* SIFT).

$$G_\lambda^X = \frac{1}{T}\sum_{t=1}^{T} \nabla_\lambda \log u_\lambda(x_t).$$

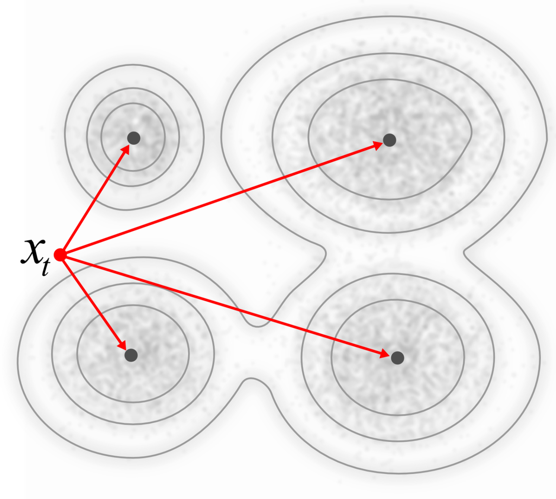
where $u_\lambda(x) = \sum_{i=1}^{K} w_i u_i(x)$ is a GMM with K Gaussians

→ We have a closed form diagonal approx. of F_λ

$$\mathcal{G}_{\mu,i}^X = \frac{1}{T\sqrt{w_i}} \sum_{t=1}^{T} \gamma_t(i) \left(\frac{x_t - \mu_i}{\sigma_i}\right)$$

$$\mathcal{G}_{\sigma,i}^X = \frac{1}{T\sqrt{2w_i}} \sum_{t=1}^{T} \gamma_t(i) \left[\frac{(x_t - \mu_i)^2}{\sigma_i^2} - 1\right]$$

Comparison with BOV: $\frac{1}{T}\sum_{t=1}^{T} \gamma_t(i)$

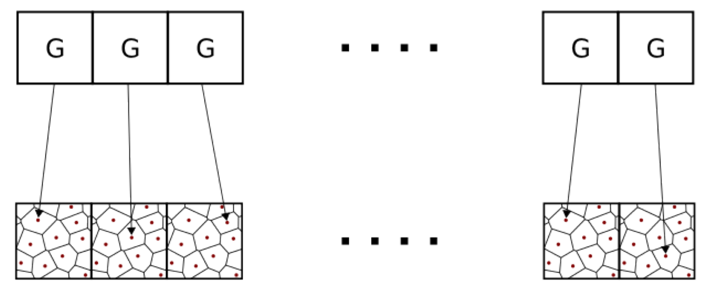


FV Compression

- When $D = 64$, $K = 256$ and $R = 8$, FV are $E = 262,144$-dim

- **PQ**: split FV into small sub-vectors of size G (*e.g.* $G = 8$) and perform VQ for each subvector.

→ A FV is represented as a vector of codebook indices.



SVM Training with SGD

Problem: predicting the unobserved output value y according to an observed input vector x

Goal: finding the label predictor by

$$\min E_{x,y} l(f(x), y) \text{ where } f(x) = w^T x + b$$

-SGD Learning: each sample is decompressed on the fly and fed to the SGD. In this way only one decompressed FV is "alive" at a time in RAM.

Advantages of using FV

-Scalability: FV is high dimensional therefore can be used with costless linear-SVMs, compressed FV are memory efficient.

-Discriminativity: Image is described by what makes it different from other images on average (tf-idf)

-Informativeness: The quantization process of BoW is lossy where only counting statistics is used. The FV extends BoW by employing higher order statistics.

Dataset

3 sub-branches of ImageNet:
-Fungus: 134 classes, ≈ 88K images
-Ungulate: 183 classes, ≈ 173K images
-Vehicle: 262 classes, ≈ 226.5K images
→ Half of the images are used for training and the other half as testing
-Different categories in fungus dataset:

Tricholoma vaccinum

Boletus chrysenteron

Experiments

-Image features: SIFT, 256 Gaussians, spacial pyramids, PQ compression on Fisher Vectors for both training and test images

-Top 1 Accuracy(%):

	[4]	ours
fungus	11.6	19.5
ungulate	14.5	27.9
vehicle	24.1	38.9

Current Work

Going beyond one-vs-all learning:
-Multiclass SVM
-Ranking SVM
-Tree structured SVM

→ preliminary results show limited improvement over one-vs-all.

References

[1] Perronnin F., Dance C., Fisher Kernels on Visual Vocabularies for Image Categorization, in *CVPR*, 2007

[2] Perronnin F., Sánchez J., Mensink T., Improving Fisher Kernel for Large-Scale Image Classification, in *ECCV*, 2010

[3] Sanchez J., Perronnin F., , High Dimensional Signature Compression for large-Scale Image Classification, in *CVPR*, 2011

[4] Deng J., Berg A.C., Li K., Fei-fei L., What does classifying more than 10,000 image categories tell us?, in *ECCV*, 2010

[5] Jaakkola T., Haussler D., Exploiting Generative Models in Discriminative Classifiers, in *NIPS*, 1998