

Improved Fisher Vector for Large Scale Image Classification

XRCE's participation for ILSVRC

Jorge Sánchez, Florent Perronnin and Thomas Mensink

Xerox Research Centre Europe (XRCE)

Overview

- Fisher Vector
- Improved FV + results on VOC 07
- Compression
- Classification
- Results on VOC2010 & ILSVRC2010

Fisher Vector

- Exploiting Generative Models in discriminative classifiers [Jaakkola & Haussler 1999]
- Feature vector is derivative wrt probabilistic model
- Measure Similarity using the Fisher Kernel

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

- Fisher Information Matrix

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

- Learning a classifier on Fisher Kernel equals learning a linear classifier on $G_\lambda^X = L_\lambda G_\lambda^X$ with $F_\lambda = L_\lambda' L_\lambda$

Fisher Vector (2)

- Fisher Kernels on visual vocabularies for image categorization
[Perronnin & Dance 2007]
- $X = \{x_t, t = 1 \dots T\}$ D-dimensional local features from an image

- GMM:
$$u_\lambda(x) = \sum_{i=1}^K w_i u_i(x)$$

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

Fisher Vector (3)

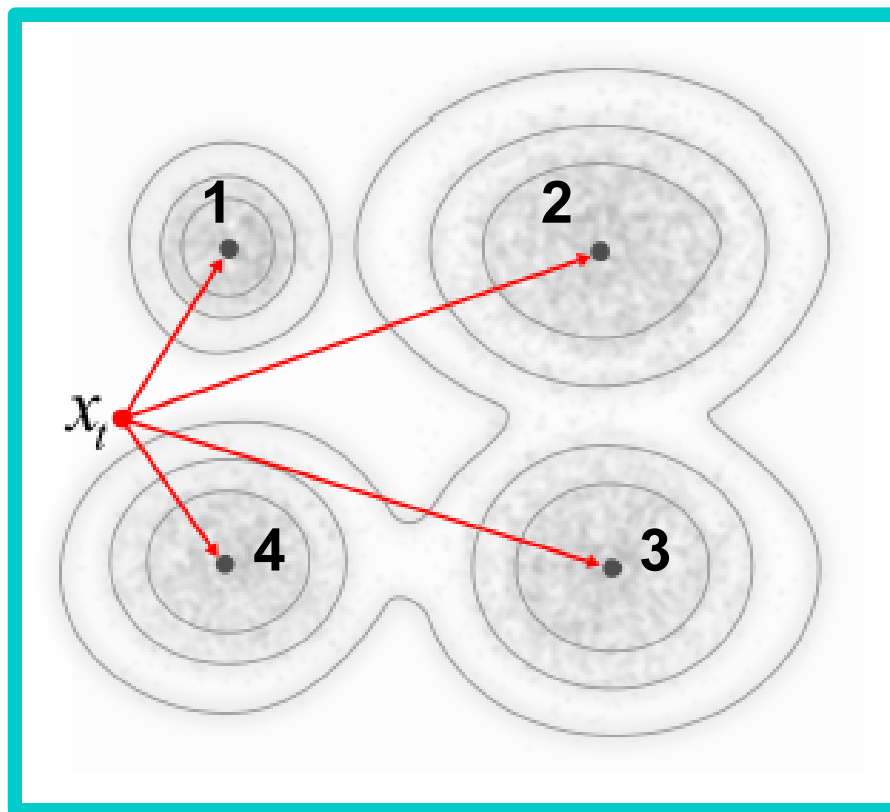
BOV

Hard Assignment

[0 0 0 1]

Soft Assignment

[.3 .1 .1 .5]



Fisher Vector

Gradient wrt w

[.15 -.2 -.35 .2]

Gradient wrt mean

[.8 -1.5 -3.7 -1.3 -3.8 1.2 -.9 1.4]

Gradient wrt var

[-1.2 -.9 1.4 -.8 1.5 -3.7 1.3 -3.8]

BOV Histogram has size: K

Fisher Vector (wrt to mean and var): $2 * D * K$

Improving the Fisher Vector

- L2 Normalization
- Power Normalization
- Spatial Pyramid Matching

Improving the Fisher Kernel for Large-Scale Image Classification

Florent Perronnin, Jorge Sánchez and Thomas Mensink
Xerox Research Centre Europe (XRC-E)
email: {f.perronnin, j.sanchez, t.mensink}@xerox.com

xerox

Summary

- The Fisher kernel (FK) combines the benefits of generative and discriminative approaches and extends the popular bag-of-visual-words (BOV) by going beyond count statistics.
- However in image classification BOV still outperforms FK.
- We improve the FK, by L2 normalization, power normalization and spatial pyramids.
- On PASCAL VOC 2007 we increase the Average Precision from 47.9% to 53.3%, using these improve methods. Using only SIFT descriptors and linear classifiers.
- Large scale experiment: we learn classifiers from large datasets obtained from ImageNet and Flickr groups.
- Although not intended for that purpose, Flickr groups are great for training classifiers.
- Combining all these resources, using only SIFT features and linear classifiers, we obtain 53.5% Average Precision on PASCAL VOC 2007, which equals the current-state-of-the-art.

L2 Normalisation

- We can write the IV $\beta(x)$ as:

$$c_1^T = V_1 \int p(x) \log u_1(x) dx \quad (1)$$
- Decompose p into two parts: a background u_0 with λ estimated to maximize $\int u_0 \log u_0(x) dx$ and an image-specific part q :

$$c_1^T = \lambda V_1 \int q(x) \log u_1(x) dx + (1-\lambda) V_1 \int u_0(x) \log u_1(x) dx \quad (2)$$
- The Fisher vector focuses on image-specific content, but, depends on the proportion of image-specific information in S_0 , two images containing the same object at different scales will have different signatures.
- To remove the dependence on λ , we can L2-normalize c_1^T or equivalently q^T .

Experimental Setup

- Densely sampled local SIFT and color features, with PCA reduced to 64D.
- Train GMM with $K = 256$ using EM algorithm.
- Learn linear SVMs in primal using Stochastic Gradient Descent.
- Take pre-pair fusion of SIFT and color features.

PASCAL VOC 2007

- Around 25K images of 20 classes, evaluated using mean Average Precision (mAP).
- Best results up to date in ILSV, mAP obtained by combined Classification and Localization approach [Heinrich et al. 2009].
- The FK obtains 53.5% mAP, an improvement of over 30% compared to FK. FK is close to the best SIFT-only results (53.0%) of [Wang et al. 2010].
- Similarly, we demonstrate state-of-the-art accuracy on Cal Tech 256.

Table 1. Impact of the proposed modifications to the FK on PASCAL VOC 2007.

FK	L2	SP	SP+L2	Cal	Cal+L2
-	-	-	47.9	48.3	48.9
-	-	-	54.2	48.9	51.6
-	-	-	55.8	48.4	53.9
-	-	-	56.5	48.5	49.9
-	-	-	56.3	50.9	50.3

Power Normalization

- As the number of Gaussians increases, Fisher vector to some extent α .
- The dot product (L2 distance) is a poor measure of similarity on sparse vectors.
- Replace kernel with e.g. Laplacian kernel, or:
- L¹ to regularize the representation, advantage linear classification.
- Power Normalization to outsparsify representation:

$$\beta(x) = \text{sign}(x) |x|^\alpha \quad (3)$$
- Parameter $\alpha: 0 < \alpha \leq 1$, optimal value with the number K of Gaussians. But $\alpha = 0.5$ is a good value for $16 \leq K \leq 256$.
- When combined with L2 normalization, we first apply power normalization and then L2 normalization.

Figure 3. Effect of $K = \{16, 64, 256\}$ on sparsity and effect of power normalization.

Fisher Kernel Framework

- Model a sample X of T iid. local descriptors x_i by its deviation from a Gaussian mixture model $u_1(x) = \sum_{k=1}^K u_k(x)$:

$$c_1^T = \sum_{k=1}^K \sum_{i=1}^T \log u_k(x_i) \quad (4)$$
- Assume diagonal covariance matrix, and consider the gradient w.r.t. means and covariances, then gradient vector is $2K \cdot D$ -dimensional.

- Measure the similarity using the Fisher Kernel:

$$K(X, Y) = c_1^T F_1^{-1} c_2 \quad (5)$$
- with $F_1 = \text{tr}_x \sum_{k=1}^K \sum_{i=1}^T \log u_k(x_i) \log u_k(x_i)^T$.
- This equals a dot product on normalized Fisher Vectors (FV):

$$c_1^T = F_1 c_1^{\text{norm}}, \quad \text{with } F_1^{-1} = F_1^{-1} L_1 \quad (6)$$
- Learning a classifier on the Fisher Kernel is equivalent to learning a linear classifier on the Fisher Vectors c_1^{norm} .

Spatial Pyramids

- Introduced by Lawick et al. 2006, to take into account the rough geometry.
- Follow the splitting of the coding systems of INSCAL VOC 2006.
- Power normalisation alone even more important since PV becomes even sparser.

Figure 4. Spatial Pyramid Matching.

Large-Scale Experiments

- ImageNet: 10 groups, up to 25K per class, total of 250K images.
- Flickr Groups: 10 groups, up to 25K per class, total of 250K images.

Table 2. Learning from different training resources using SIFT only. Fv+L2 is combination of the classifiers, evaluation on VOC 2007 "test" set.

Train	Class	Color	SIFT	Color+SIFT	Color+SIFT+L2	Color+SIFT+L2+SP	Color+SIFT+L2+SP+L1
F	36.0	36.9	46.8	71.8	74.0	74.7	74.7
F	46.3	72.7	75.0	76.7	76.5	78.0	77.8
F	73.7	58.8	52.8	53.6	54.1	54.1	53.5
Fv+L2	45.5	73.0	43.0	38.5	38.5	71.4	77.8
F	77.2	68.0	56.3	66.0	61.0	64.1	64.4

Table 3. Learning from different training resources using SIFT only. Fv+L2 is combination of the classifiers, evaluation on VOC 2007 "test" set.

Train	Class	Color	SIFT	Color+SIFT	Color+SIFT+L2	Color+SIFT+L2+SP	Color+SIFT+L2+SP+L1
F	36.0	43.0	74.4	61.1	-	51.7	58.4
F	31.4	47.7	56.3	49.8	75.4	28.1	46.5
F	56.9	72.7	74.3	64.4	62.7	24.3	56.4
Fv+L2	56.2	55.0	74.7	39.2	42.9	32.7	56.1
F	62.1	45.2	74.8	69.7	36.1	52.4	54.4

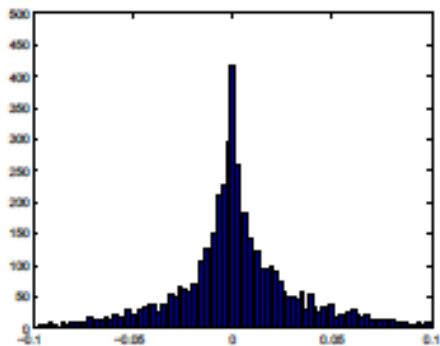
- According to Pascal VOC Competition 3 using any data including the test set.
- We use only SIFT features.
- Flickr Groups are a great resource for training classifiers.
- Adding more data, and combining different data sources improve classification.
- Linear classifiers using FK trained on large datasets performs equally to costly classification and localization approach of [Heinrich et al. 2009].

L2 Normalization

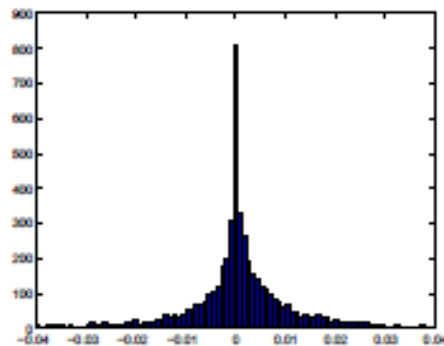
- By construction the Fisher Vector discards descriptors which are likely to occur in any image
- The FV **focus on image specific features**
- However, the FV depends on the **amount** of image specific information / background information
 - 2 images with same object on a different scale will have a different feature vector
- L2 Normalization to remove this dependence

Power Normalization

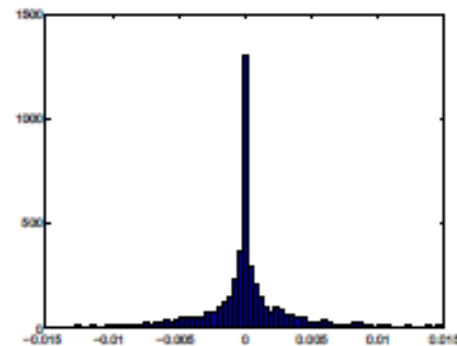
- As the number of Gaussians increase, the FV becomes sparser
 - Replace dot-product with other kernel
 - Unsparsify the representation
- Power normalization to unsparsify: $f(z) = \text{sign}(z)|z|^\alpha$



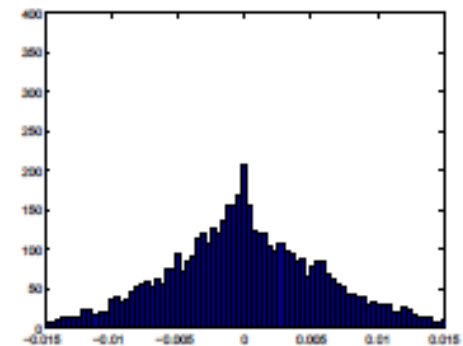
K = 16



K = 64



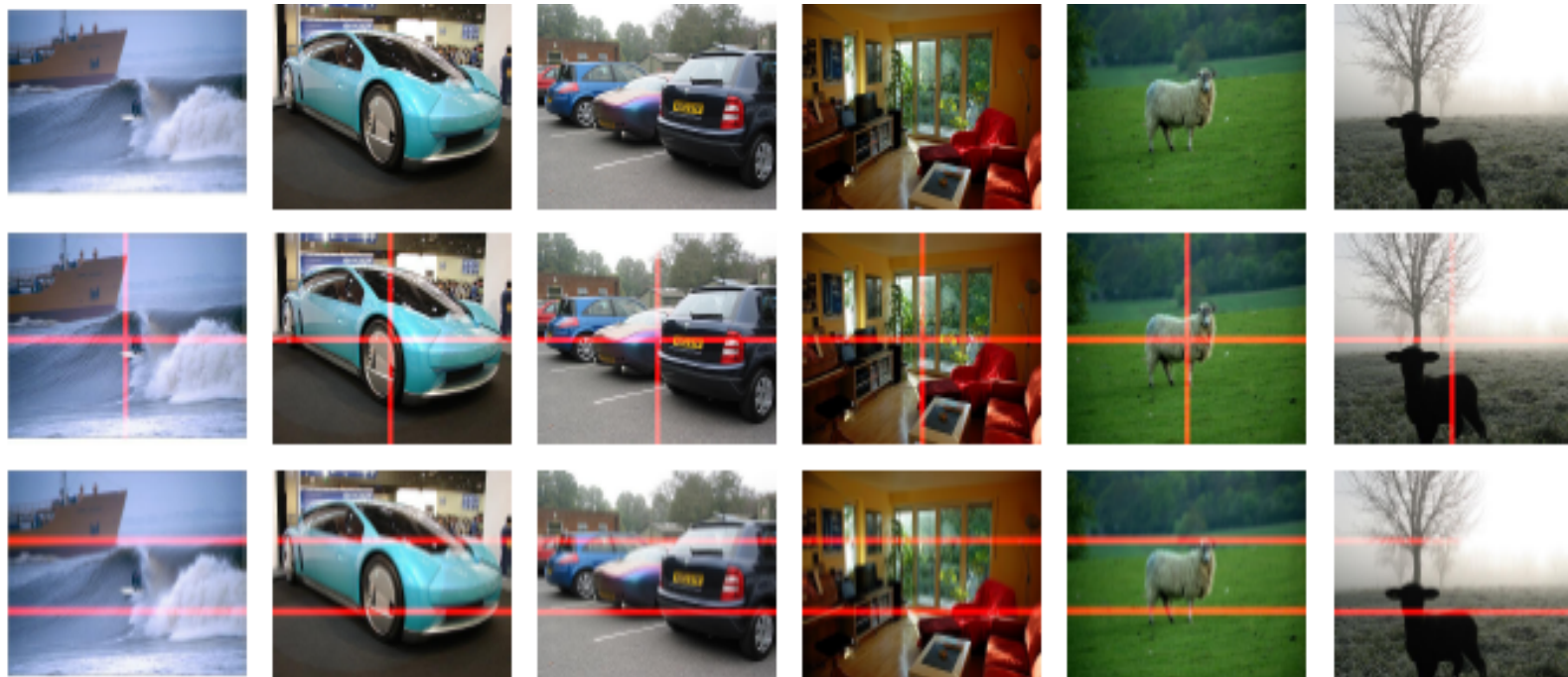
K = 256



K = 256
Power Normalized

Spatial Pyramids

- Take rough geometry into account [Lazebnik 2006]



- Power normalization becomes even more important (FV is sparser)

Experiments VOC 2007

- Improved Fisher Vector [ECCV 2010]
- Dense multiscale sampling, PCA, K=256
- Linear SVM

PN	L2	SP	SIFT	Col	S+C
-	-	-	47.9	34.2	45.9
✓	-	-	54.2	45.9	57.6
-	✓	-	51.8	40.6	53.9
-	-	✓	50.3	37.5	49.0
✓	✓	✓	58.3	50.9	60.3

Experiments VOC 2007 (2)

- Improved Fisher Vector [ECCV 2010]
- **Larger Scale**
 - Flickr Group Images
 - Up to 25k per class / 350k in total
 - Late fusion with VOC07 trainset
 - 63.3% (SIFT only)
- Best results 63.5% Localization and Classification [Harzallah et al. 2009]
- Flickr Groups are a great resource for labelled images
 - No additional labelling used!
- **More training data improves performance**

So far...

- FV is a rich representation, extends BOV.
- High dimensional ($2 \times D \times K \times S$) but allows for linear SVM
- Performance is compatible to state of the art

However...

- $2 * 64 * 256 * 8 = 262,144$ dimensions
- Almost dense feature
- $\sim 1\text{MB}$ per image / per modality
- ImageNet Train/Test/Val $\rightarrow 1.4\text{ TB}$ (per modality)

Compression [unpublished]

Two options

A) Dimension Reduction

- PCA / Dense Random projections
 - is costly in high dimensional dense space
- Hash Kernels
 - Observation: performance decreases rapidly (already by factor 4)
- Can improve learning speed (not necessary)

B) Data Compression

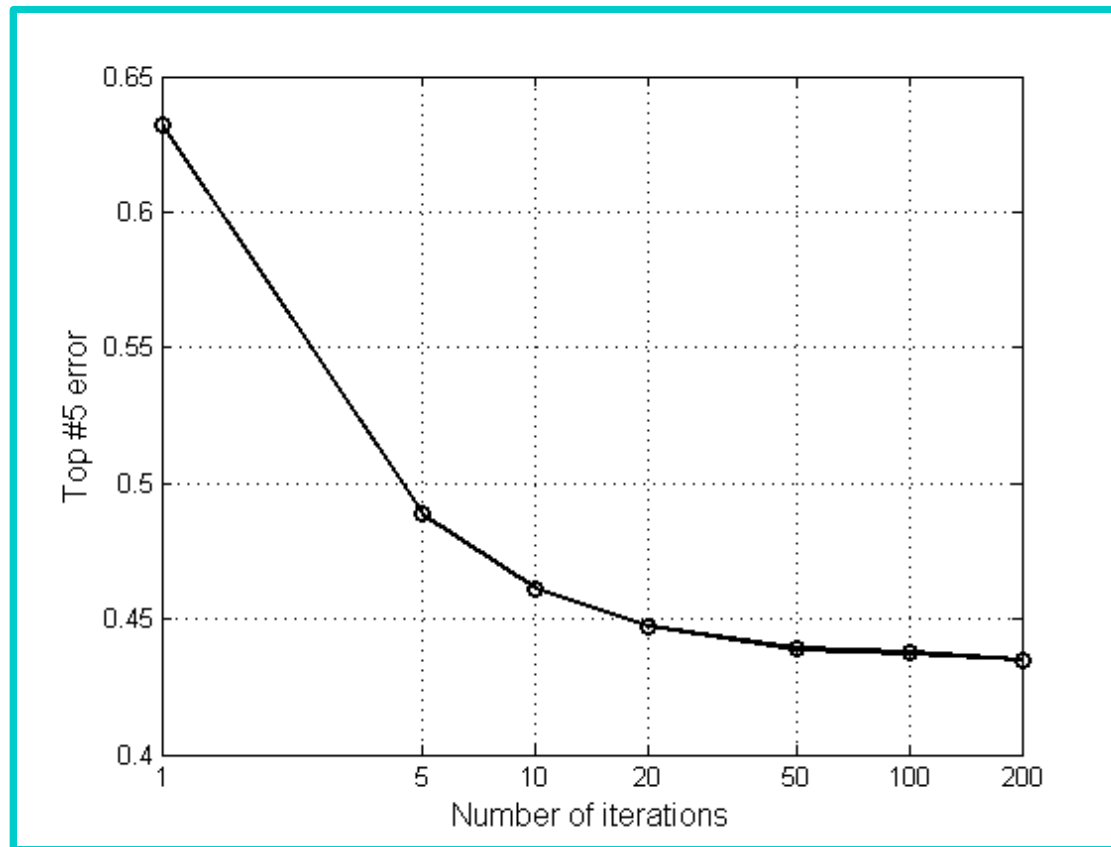
- Use same dimensionality
- But lossy compression up to factor 64 possible.
- 1.4TB → 20GB (per modality)
- Not able to learn in compressed space

Stochastic Gradient Descent (SGD)

- Learn Linear SVM in the primal, PEGASOS
[Shalev-Shwartz et al. 2007]
- SGD inspired on Pegasos by L. Bottou
[<http://leon.bottou.org>]
- Online algorithm, using one sample at the time
- Our approach is:
 1. Load compressed vector
 2. Decompress vector
 3. SGD Update

Stochastic Gradient Descent (2)

- Performance vs number of passes through data



Categorization Pipeline

- Extract dense sampled features (SIFT, Colour)
- Project (with PCA) to 64D
- Learn codebook with K (256) Gaussians on 1M features
- Learn Compressor (on small set of FV)
- Compute and compress FV
- Learn Linear Classifiers using SGD
- Classify test images

Categorization Pipeline (2)

- Computation time for Learning ImageNET

Intel Xeon double quadcore (16 proc) @ 2.53GHz, 32GB RAM

	CPU	Wall-Clock
Extract SIFT+Projection	36h	18h
GMM	minutes	
Learn Compressor	48h	3h
Extract FV + Compression	96h	6h
500 SGD Iterations*	960h	66h
Total (SIFT)	1140h	93h

1.2M train images, training ~ 4 CPU sec per image / modality

* Without significant loss of performance 500 → 50 iterations

Categorization Pipeline (3)

- Computation time for Testing on ImageNET

	CPU
Feature Extraction + Projection	2.5h
FV Extraction	30m
Classifiers	12h

- Total SIFT + Col = 30h
- 150K images / 1000 classes

Classification \ll 1ms per image/class/modality

Results

- Pascal VOC 2010 and ImageNet Challenge
- Same approach and settings
- $K = 256$
- FV + L2 & Power Norm + pyramids
- Compression
 - Except for VOC train/val set
- Linear SVM in primal
 - Number of SGD iterations are different

Pascal VOC 2010

- 10K test images / 20 classes / multi-label
- Challenge 1: Only provided train/val data

Rank		MAP
1	NUSPSL	73.8
2	NLPR	71.2
3	NEC	70.9
9	XRCE Improved FV	61.2

Pascal VOC 2010 (2)

- 10K test images / 20 classes / multi-label
- Challenge 2: Any data except test data
- 1M Flickr Group images of 18 classes
 - Tv/monitor and sofa are missing
 - No additional labelling, just the group labels

Rank		MAP	# Classes
4	BIT	26.9	20
3	UCI	51.7	9
2	XRCE Flickr 1M	65.5	18
1	XRCE Optimal Fuse F & V	68.3	20

More data helps: 61.2 → 68.3 MAP

ImageNet Challenge

- 150k test images / 1000 classes / single labelled.
 - Flat cost: is the correct label in the top 5?
 - Hierarchical cost: distance to lowest common ancestor

	flat cost	hie cost
NEC-UIUC	0.28191	2.1144
XRCE	0.33649	2.5553
ISIL	0.44558	3.6536
UCI	0.46624	3.6288

Conclusions

- Improved Fisher Vector for Image Classification
- Linear Classification → scales to **larger** scale
- More data (Flickr Groups) helps
 - Pascal VOC 2010: 61.2 → 68.3 MAP
- Using compression → scales to **LARGE** scale
- **Very very fast:** training (8s/i) & classifying (2s/i)

Questions?

Improving the FV

Lp normalization

Descriptors which are likely to occur in any image are automatically discounted in :
$$\mathcal{G}_i^X = \frac{1}{\sqrt{w_i}} \sum_{t=1}^T \gamma_t(i) \left(\frac{x_t - \mu_i}{\sigma_i} \right)$$

- hard assignment assumption $\gamma_t(i) \approx 1 \Rightarrow p(x_t) \approx w_i p_i(x_t)$
 - large w_i value: corresponds to frequent visual word
 - large $p_i(x_t)$ value, i.e. small $\left\| \frac{x_t - \mu_i}{\sigma_i} \right\|^2$ value
- ⇒ extension of tf-idf to continuous variables

More generally, assuming that the features can be modeled by a mixture of an image-independent distribution (p) and an image-specific distribution:

- the image-independent part is automatically discarded from the FV
- Lp normalizing the FV removes dependence on amount of image-independent information

L2 Normalization (2)

- Fisher Vector

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

$$G_{\lambda}^X \approx \nabla_{\lambda} E_{x \sim p} \log u_{\lambda}(x) = \nabla_{\lambda} \int_x p(x) \log u_{\lambda}(x) dx. \quad 2$$

$$G_{\lambda}^X \approx \omega \nabla_{\lambda} \int_x q(x) \log u_{\lambda}(x) dx + (1 - \omega) \nabla_{\lambda} \int_x u_{\lambda}(x) \log u_{\lambda}(x) dx \quad 3$$

- GMM Trained with Maximum Likelihood, ie maximize

$$E_{x \sim u_{\lambda}} \log u_{\lambda}(x)$$

$$\nabla_{\lambda} \int_x u_{\lambda}(x) \log u_{\lambda}(x) dx = \nabla_{\lambda} E_{x \sim u_{\lambda}} \log u_{\lambda}(x) \approx 0$$

- Fisher Vectors automatically **focus on image specific features** and discard image independent/background features
- L2 Normalization to remove dependence on ω