

Paper Presentation

Return of the Devil in the Details: Delving Deep into Convolutional Nets

Authors: Ken Chatfield, Karen Simonyan, Andrea Vedaldi,
and Andrew Zisserman Visual Geometry Group, Department
of Engineering Science, University of Oxford

Presentation by - Anubhooti Jain, Debalina Saha,
Baddepudi Venkata Naga Sri Sai Vineetha

Concept & Motivation

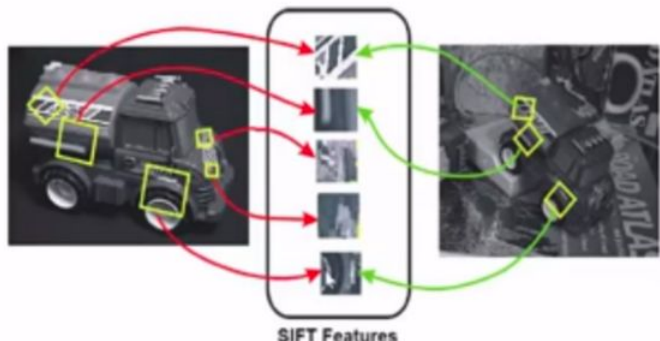
Concept - The paper draws a comparison between different CNN methods at a common ground and their performance on tasks like image classification and object recognition.

Focus - The paper focuses on constructing image representations, that is, encoding functions majorly comparing between traditional (shallow) methods like Improved fisher vector (IFV) and deep methods which use CNNs.

Shallow Vs Deep Representations

Shallow representation include image representation models like Bag of visual words (BoW) and IFV.

They have handcrafted parameters and can be described using descriptors like SIFT.



Deep representation have handcrafted parameters too but a large number are learnt from data. They use layers of non-linear feature extractors.

CNNs are used in order to learn these parameters from the data. A large dataset is used to train the model that further can be used for any other dataset smaller than which it was trained on.

Experiment Setup

The paper considered 3 scenarios for image representation before it was fed to a classifier -

Scenario 1 - Shallow image representations - IFV used in this case. The IFV is obtained by extracting a dense collection of patches and their corresponding local descriptors. These descriptors are soft-quantized using Gaussian Mixture model (GMMs)

Scenario 2 - Deep representations pre-trained on outside data - A large dataset is used to train the CNN which gives a powerful image descriptor.

Scenario 3 - Deep representations pre-trained and then fine-tuned on target dataset. - It is a modified case of Scenario 2. Before giving target dataset to the trained CNN (the image descriptor), the parameters of the trained model are fine-tuned on the target dataset.

Commonalities

They used generally applicable practices to find a common ground for comparison in all 3 scenarios. Like **color information**, **feature normalization**, and substantial **data augmentation**.

Data Augmentation - While it is largely used in the deep representation scenario, it can be applied to the shallow representation as well. Different transformations are used such that the underlying class remains same at the time of training (sometimes even during testing) like cropping and flipping.

Linear predictors - All three scenarios are image representations that are used to build linear predictors. The paper uses SVM (that uses hinge loss) in order to learn these predictors.

Datasets

- PASCAL-VOC
 - VOC-2007
 - VOC-2012
- ILSVRC-2012
- Caltech-101
- Caltech-256

Implementation details

Improved Fisher Vector

To standard formulation few modifications are added:

- Use of intra-normalisation of the descriptor blocks:

L2 normalisation is applied to individual sub-blocks $s(u_k, v_k)$ of the vector $\phi_{FV}(I)$

- Use of spatially-extended local descriptors instead of spatial pyramid
- Use of colour features in addition to SIFT descriptors

Local Colour Statistics (LCS) features are used.

Convolutional Neural Network

CNN Architectures

Arch.	conv1	conv2	conv3	conv4	conv5	full6	full7	full8
CNN-F	64x11x11 st. 4, pad 0 LRN, x2 pool	256x5x5 st. 1, pad 2 LRN, x2 pool	256x3x3 st. 1, pad 1 -	256x3x3 st. 1, pad 1 -	256x3x3 st. 1, pad 1 x2 pool	4096 drop- out	4096 drop- out	1000 soft- max
CNN-M	96x7x7 st. 2, pad 0 LRN, x2 pool	256x5x5 st. 2, pad 1 LRN, x2 pool	512x3x3 st. 1, pad 1 -	512x3x3 st. 1, pad 1 -	512x3x3 st. 1, pad 1 x2 pool	4096 drop- out	4096 drop- out	1000 soft- max
CNN-S	96x7x7 st. 2, pad 0 LRN, x3 pool	256x5x5 st. 1, pad 1 x2 pool	512x3x3 st. 1, pad 1 -	512x3x3 st. 1, pad 1 -	512x3x3 st. 1, pad 1 x3 pool	4096 drop- out	4096 drop- out	1000 soft- max

CNN training:

- Learning on ILSVRC-2012 using gradient descent with momentum
- The layers are initialized from Gaussian distribution with zero mean and variance equal to 10^{-2} . Data augmentation in the form of random crops, horizontal flips, RGB colour jittering is used.

CNN fine-tuning on the target dataset:

- Fine-tuned CNN-S using VOC-2007, VOC-2012, or Caltech-101 as the target data.
- VOC-2007 and VOC-2012 - softmax regression loss is replaced with one-vs-rest classification hinge loss or ranking hinge loss
- In Caltech-101 fine-tuning was performed using softmax regression loss.

Low-dimensional CNN feature training:

- Same dimensionality of 4096 is used in the last hidden layer.
- 3 modifications of CNN-M are trained with lower dimensional full7 layers of 2048, 1024, 128 respectively.

Data Augmentation

3 Strategies:

- No augmentation - '-'
- Flip Augmentation - 'F'

Mirroring images about y-axis

- C+F Augmentation - 'C'

Combines cropping and flipping

Augmented images are used as - standalone samples (f), fusing corresponding descriptors using sum (s) or max (m) pooling or stacking (t)

Color Information

- Shallow networks -
 - Color information added by **replacing** SIFT with color descriptors. Denoted by method COL in analysis table.
 - Color information added by **combining** SIFT with color descriptors (done by stacking the corresponding IFVs). Denoted by method COL+.
- Deep networks -
 - Color information subtracted by retraining CNN with all input image converted to grayscale. Denoted by method GS.

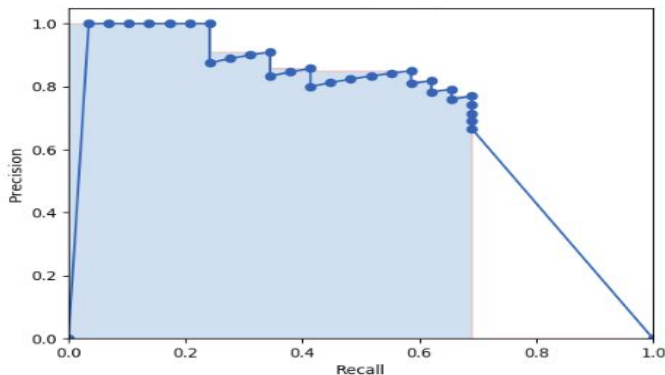
Model Evaluation Metrics

- mAP(mean average Precision)

- Used for VOC 2007 and VOC 2012

- $$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$



		True/Actual	
		Positive (🐶)	Negative
Predicted	Positive (🐶)	5 (TP)	1 (FP)
	Negative	2 (FN)	2 (TN)

Precision = 5/6, Recall = 5/7

		True/Actual		
		Cat (🐱)	Fish (🐟)	Hen (🐔)
Predicted	Cat (🐱)	4	6	3
	Fish (🐟)	1	2	0
	Hen (🐔)	1	2	6

Class: Cat - Precision = 4/13, Recall = 4/6

Other metrics:

- Top-5 error rates used for ILSVRC-2012
- Accuracy used for Caltech-101 and Caltech-256

Observations







Method	SPool	Image Aug.			Dim	mAP						
(I) FK BL	spm	–			327K	61.69	79.0	67.4	51.9	70.9	30.8	72.2
(II) DECAF	–	(C)	t	t	327K	73.41	87.4	79.3	84.1	78.4	42.3	73.7
(a) FK	spm	–			327K	63.66	83.4	68.8	59.6	74.1	35.7	71.2
(b) FK IN	spm	–			327K	64.18	82.1	69.7	59.7	75.2	35.7	71.3
(c) FK	(x,y)	–			42K	63.51	83.2	69.4	60.6	73.9	36.3	68.6
(d) FK IN	(x,y)	–			42K	64.36	83.1	70.4	62.4	75.2	37.1	69.1
(e) FK IN	(x,y)	(F)	f	–	42K	64.35	83.1	70.5	62.3	75.4	37.1	69.1
(f) FK IN	(x,y)	(C)	f	s	42K	67.17	85.5	71.6	64.6	77.2	39.0	70.8
(g) FK IN	(x,y)	(C)	s	s	42K	66.68	84.9	70.1	64.7	76.3	39.2	69.8
(h) FK IN 512	(x,y)	–			84K	65.36	84.1	70.4	65.0	76.7	37.2	71.3
(i) FK IN 512	(x,y)	(C)	f	s	84K	68.02	85.9	71.8	67.1	77.1	38.8	72.3
(j) FK IN COL 512	–	–			82K	52.18	69.5	52.1	47.5	64.0	24.6	49.8
(k) FK IN 512 COL+	(x,y)	–			166K	66.37	82.9	70.1	67.0	77.0	36.1	70.0
(l) FK IN 512 COL+	(x,y)	(C)	f	s	166K	67.93	85.1	70.5	67.5	77.4	35.7	71.2
(m) CNN F	–	(C)	f	s	4K	77.38	88.7	83.9	87.0	84.7	46.9	77.5
(n) CNN S	–	(C)	f	s	4K	79.74	90.7	85.7	88.9	86.6	50.5	80.1
(o) CNN M	–	–			4K	76.97	89.5	84.3	88.8	83.2	48.4	77.0
(p) CNN M	–	(C)	f	s	4K	79.89	91.7	85.4	89.5	86.6	51.6	79.3
(q) CNN M	–	(C)	f	m	4K	79.50	90.9	84.6	89.4	85.8	50.3	78.4
(r) CNN M	–	(C)	s	s	4K	79.44	91.4	85.2	89.1	86.1	52.1	78.0
(s) CNN M	–	(C)	t	t	41K	78.77	90.7	85.0	89.2	85.8	51.0	77.8
(t) CNN M	–	(C)	f	–	4K	77.78	90.5	84.3	88.8	84.5	47.9	78.0
(u) CNN M	–	(F)	f	–	4K	76.99	90.1	84.2	89.0	83.5	48.1	77.2
(v) CNN M GS	–	–			4K	73.59	87.4	80.8	82.4	82.1	44.5	73.5
(w) CNN M GS	–	(C)	f	s	4K	77.00	89.4	83.8	85.1	84.4	49.4	77.6
(x) CNN M 2048	–	(C)	f	s	2K	80.10	91.3	85.8	89.9	86.7	52.4	79.7
(y) CNN M 1024	–	(C)	f	s	1K	79.91	91.4	86.9	89.3	85.8	53.3	79.8
(z) CNN M 128	–	(C)	f	s	128	78.60	91.3	83.9	89.2	86.9	52.1	81.0
(α) FK+CNN F	(x,y)	(C)	f	s	88K	77.95	89.6	83.1	87.1	84.5	48.0	79.4
(β) FK+CNN M 2048	(x,y)	(C)	f	s	86K	80.14	90.9	85.9	88.8	85.5	52.3	81.4
(γ) CNN S TUNE-RNK	–	(C)	f	s	4K	82.42	95.3	90.4	92.5	89.6	54.4	81.9

TABLE 2

VOC 2007 results

Analysis of observations

Scenario 1: Shallow representation-

- Impact of data augmentation:
 - Improves performance significantly by ~3% (d vs f)
- Impact of color information:
 - COL (replacing) yields worse performance (j vs h).
 - COL+ (combining) yields improved performance in non-augmented (h vs k), nonsignificant impact in augmented (i vs l)
- Impact of Intra-normalization:
 - Improving performance by 1% (c vs d)
- Increasing number of GMM centers
 - Improving performance (d vs h) (but at the cost of increase in dimensionality)

Scenario 2: Deep representation-

- Comparison with shallow:
 - Outperforms shallow encodings even after modifications by a large difference of ~10% (i vs p)
 - Works better than all shallow representations even without augmentation (o vs a-l)
- Impact of augmentation:
 - Improves performance significantly by ~3% (o vs p)
 - Using additional samples as standalone for training and sum-pooling for combining works best (p)
- Impact of color information:
 - GS results in performance drop of 3% (w vs p).
- Fast vs Medium and slow:
 - Medium and slow outperforms fast by 2~3% (m vs p and n)
- Dimensionality reduction:
 - Dimensionality reduction of final layer in x, y, z. Marginal performance boost with 2x smaller dimension x vs p.
 - In z 32x smaller dimension than CNN of p (~650x smaller than best performing IFV i) without huge performance drop

Scenario 3: Pre-trained CNN with fine-tuning-

- Comparison with scenario 2:
 - Improvement of 2.7% in performance (γ vs n)

Conclusion

- This paper presented an empirical evaluation of CNN based methods for image classification, and comparison with traditional shallow feature representations.
- Shown that data augmentation can improve both shallow and deep representation.
- Shown that deep networks outperform the shallow methods by large margin and fine-tuning further improves performance.

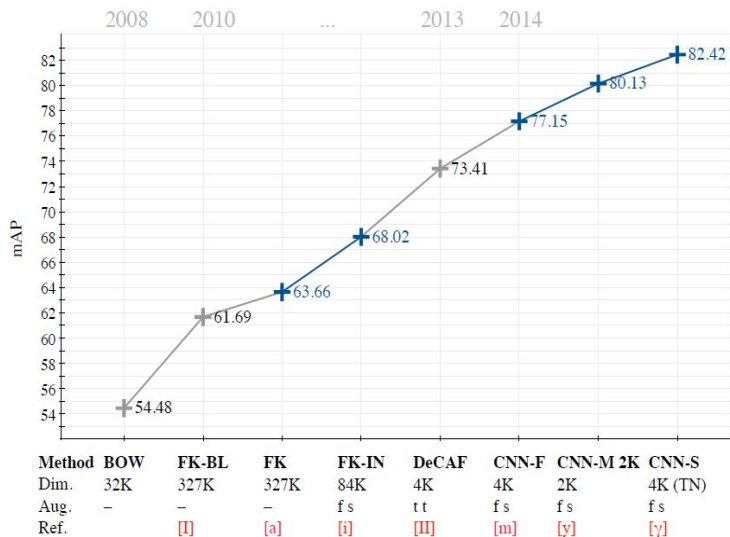


Fig. 1. Evolution of Performance on PASCAL VOC-2007 over the recent years. Please refer to Table 2

Questions?