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

1 Citation

Chatfield, Ken, et al. "Return of the devil in the details: Delving deep into convolutional nets." arXiv preprint arXiv:1405.3531 (2014).

https://arxiv.org/pdf/1405.3531.pdf

2 Abstract

Convolutional Neural Networks are good at feature extraction, but they are trained with data augmentation. How does this compare to shallow feature extraction techniques like Improved Fisher Vectors? We find that preprocessing techniques like data augmentation work well with shallow methods but that CNNs do have an advantage over shallow methods.

3 Introduction

Shallow feature extraction methods like Improved Fisher Vectors (IFV) and Bag of Visual Words (BoVW) are shallow (compared to CNNs) and handcrafted. We consider three cases: (1) shallow method, (2) pre-trained CNN, (3) pre-trained CNN fine-tuned to target dataset. We also look at these preprocessing methods: (1) color information, (2) feature normalization, (3) data augmentation. We get near-state-of-the-art performance on PASCAL VOC with a simpler system that uses less data.

4 Scenarios

Our shallow representation is the Improved Fisher Vector (IFV). Here's how the IFV is computed:

1. Densely sample patches and compute SIFT descriptors for them
2. Soft-quantize the descriptors with a Gaussian Mixture Model
3. Compute first/second order differences between descriptor and its Gaussian center and aggregate them into a vector weighted by Gaussian soft assignment and covariance. This is the Fisher Vector.
4. Compute signed square root of scalar components and normalize to L2 unit norm. This is the Improved Fisher Vector.

We keep some things common across all three scenarios (i.e. shallow, CNN, CNN with fine-tuning). First, we do data augmentation with crops and flipping. Second, we train an SVM (minimizing a combination of hinge loss and a quadratic regularizer).

Our task is image classification on the PASCAL-VOC and Caltech-101 (and Caltech-256) datasets. CNNs are pretrained on ImageNet.

5 Details

First consider the Improved Fisher Vectors. We upscale by a factor of two, slide a window and compute IFVs at multiple scales. We decorrelated these vectors and reduce dimensionality with PCA. There are some other details, but we make three changes from the standard (1) intra-normalization between features to alleviate burstiness (2) use spatially-extended local descriptors because they reduce memory usage and (3) use Local Color Statistics.

We have three CNNs: AlexNet with a larger stride, a smaller version of Zeiler and Fergus's CNN, and a tweaked OverFeat accurate network.

We train on ImageNet ILSVRC-2012 with gradient descent + momentum, dropping learning rate by a factor of 10 when validation error stopped improving. Layers are initialized by sampling from a zero-mean Gaussian. We augment data with crops, flips, and color jittering. At test time, we randomly sample crops. Our fastest CNN (modified AlexNet) took 5 days to train and the slowest (modified OverFeat) took 3 weeks to train on an NVIDIA Titan.

For fine tuning, we slap on a fully connected layer intiailized with a zero-mean Gaussian and with 4096 units (so their feature vector is much smaller than the IFV).

PASCAL-VOC is multilabel classification, so we used one-vs-rest hinge loss and ranking hinge loss as our loss functions. For the Caltech dataset, we use the usual softmax regression loss. To avoid overfitting, we started the learning rate small.

We experimented with three data augmentaion strategies for the CNNs: (1) none, (2) flips, (3) crops and flips.

6 Analysis

Augmentation helps both IFV and the CNNs.

For IFV, using both SIFT and color descriptors gives gains.

Training a CNN on grayscale images hurts a lot.

CNNs consistently beat IFV.

We can cut CNN feature vector size from $4096D$ to $128D$ and performance isn't hurt much.

Fine-tuning, even on the tiny 5K image VOC dataset gives bbenefits.

Stacking IFV with CNN doesn't really help.

The CNNs get near state of the art on ImageNet, VOC, and Caltech.

7 Conclusion

CNNs are better than IFV. Using data augmentation can help IFV. CNNs pretrained on ImageNet do well on VOC and Caltech.