CNN Features off-the-shelf: an Astounding Baseline for Recognition

1 Citation

Sharif Razavian, Ali, et al. "CNN features off-the-shelf: an astounding baseline for recognition." Proceedings of the IEEE conference on computer vision and pattern recognition workshops. 2014.

https://www.cv-foundation.org/openaccess/content_cvpr_workshops_2014/W15/papers/Razavian_CNN_Features_Off-the-Shelf_2014_CVPR_paper.pdf

2 Abstract

We show that a pretrained (on ImageNet) OverFeat CNN can generalize quite well as a feature extractor for other datasets.

3 Introduction

Others have generalized CNNs to a few tasks. We combine Overfeat features with an SVM. We consider a variety of tasks, starting with PASCAL VOC and the MIT Scene Dataset (both are similar to ImageNet). Then, we look at fine-grained classification (i.e. identify flower species) and get good results with data augmentation and an SVM. Then, we consider attribute detection (e.g. has-glasses) of people in the H3D dataset and beat traditional approaches like poselets and the deformable parts model. Next, we consider instance retrieval, where our CNN features + SVM does quite well (and even better if we use PCA, whitening, and re-normalization), although not as well as models that explicitly model 3D geometry.

4 Background and Outline

We use the large version of the Overfeat network pretrained on ImageNet classification. The first fully-connected layer is our feature vector. Images are cropped to 221×221.

5 Visual Classification

We consider two settings. First, we L2 normalize the feature vector and feed it into our SVM. Second, we do data augmentation with crops/rotation and do component-wise power transform. Our SVM is trained with:

$$\text{minimize}_{\boldsymbol{w}} \frac{1}{2}||\boldsymbol{w}||^2 + C\sum_i \max(1 - y_i \boldsymbol{w}^T \boldsymbol{x}_i, 0) \tag{1}$$

Let's first look at image recognition. We consider the PASCAL VOC dataset, which has 10K images in 20 categories and the MIT Indoor scenes dataset which has 16K images from 67 classes. The indoor scenes are very different than ImageNet so we think this dataset will be harder. We measure mean average precision and get state-of-the-art on PASCAL VOC. Deeper layers are better feature extractors. We also do well on the indoor scene dataset, the network struggles on examples that are difficult for humans as well.

Next, consider object detection. Others have gotten good results.

For fine-grained image recognition we consider the Caltech birds and Oxford flowers datasets. The former has 12 thousand images from 200 categories, and the latter has 100 categories with between 40 to 260 images each. We beat all traditional methods.

For attribute recognition, we consider the UIUC dataset and the H3D dataset. We recognize attributes like having-glasses, is-furry, or has-head. Our model is competitive with traditional approaches like deformable parts.

6 Visual Instance Retrieval

In the instance retrieval problem, we aim to find images from a data set that resemble a query image. We look at five different data sets of buildings, sculptures, and more.

We represent images as L2 normalized feature vectors from Overfeat. Then, for a query image, we consider all $1 \leq i \leq h$ and extract i^2 overlapping subpatches (whose union covers the image). The distance between a query subpatch and reference image is the minimum L2 distance between the query subpatch and all reference subpatches. We also trie L2 normalizing, PCA (from 4096 length to 500 length), whitening, and L2 renormalizing our feature vectors. We do data augmentation. We do quite well.

7 Conclusion

CNN features (e.g. from OverFeat) are really effective feature extractors for images. When paired with SVM, they are competitive with traditional approaches.