DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition

1 Citation

Donahue, Jeff, et al. "Decaf: A deep convolutional activation feature for generic visual recognition." International conference on machine learning. 2014.

https://arxiv.org/pdf/1310.1531.pdf

2 Abstract

We use an AlexNet pretrained on ImageNet as a feature extractor and get good performance on a variety of image recognition tasks.

3 Introduction

Neural nets trained on large supervised datasets extract better features than traditional gradient histograms. Thus, we use a pretrained AlexNet as a feature extractor (we call it DeCAF).

4 Related Work

People have tried transfer learning with unsupervised pretraining. We consider AlexNet which is supervised pretrained on ImageNet.

5 Deep Convolutional Activation Features

AlexNet won ImageNet 2012 (40.7% top-1 error). It operates on 224×224 RGB images and applies 5 conv layers with ReLU and pooling. It finishes with three fully-connected layers and a softmax. We pretrain this net (but we don't do the random color and illumination data augmentation trick).

To visualize how well these features work, we compute features for each image in a dataset, apply the t-SNE algorithm to embed the feature vectors down to a 2D space, and then plot them (or their classes) on a 2D graph. Since ImageNet has 1000 classes, which is hard to visualize, we consider higher levels on the WordNet hierarchy (e.g. outdoor vs. indoor) and observe that indoor and outdoor images are clustered separately.

Over half the compute time is taken by the fully-connected layers and 32% of the compute time are taken by the conv layers.

6 Experiments

Let $DECAF_n$ be the activations from the n^{th} layer, where $n = 7$ is right before the softmax. We resize/crop images as described in ImageNet, so we have 224×224 sized image inputs.

Our models are SVM and Logistic Regression, where we select hyperparameters with cross validation and apply dropout on the feature extractor output. We set state of the art on Caltech-101. $DECAF_6$ are our features. We also do quite well at one-shot recognition (i.e. with DeCAF we can learn the SVM weights for a new category with just a single positive training) example).

To see how DeCAF generalizes to other domains, we consider Amazon product images, office webcam images, and DSLR camera images. We plot the same categories from each of these datasets in a single t-SNE plot - the DeCAF features cluster together the same categories even if they come from different domains. Its clustering is better than SURF features and it beats SURF when used as features for classification.

To see how DeCAF generalizes to subcategories, we use a dataset where you must predict the species of a given bird (photo). We use the same classification approach as above and also try combining DeCAF with the deformable parts model (DPM). We get state of the art.

To see how DeCAF generalizes to scene recognition, we consider a dataset where we must predict the scene (e.g. Abbey, Diner, Mosque) from an image. $DECAF_7$ with SVM sets state of the art here.

7 Discussion

Using pretrained AlexNet as a feature extractor (i.e. DeCAF), you can just stack an SVM or Logistic Regression and and beat traditional hand engineered approaches on many kinds of image recognition tasks.