Learning and Transferring Mid-Level Image Representations using Convolutional Neural Networks

1 Citation

Oquab, Maxime, et al. "Learning and transferring mid-level image representations using convolutional neural networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2014.

`https://www.cv-foundation.org/openaccess/content_cvpr_2014/papers/Oquab_Learning_and_Transferring_2014_CVPR_paper.pdf`

2 Abstract

We transfer an AlexNet pretrained on ImageNet to PASCAL VOC and set state of the art there.

3 Introduction

Traditional methods would beat CNNs because we didn't have large datasets before. ImageNet changed that, but we can't expect to collect datasets that large for every task. Re-using an ImageNet model on another dataset, like PASCAL VOC, is not trivial. ImageNet images tend to focus on the object, while PASCAL VOC has the object in a scene, for example. We present a technique for transferring an ImageNet model to PASCAL VOC.

4 Related Work

Others are working on transfer learning as well, but they transfer to different datasets or use different models.

5 Transferring CNN weights

We present a way to remap class labels between source and target tasks and show a sliding window approach to train the model on the new task.

Our model is AlexNet (5 conv layers with ReLU and max-pooling, 3 fully-connected layers, and a 1000-way softmax), which operates on 224×224 RBG images.

After pretraining AlexNet on ImageNet, we remove the last fully-connected layer and softmax. We then add two fully-connected layers and a softmax, call them FC_A and FC_B (FC_B contains the softmax). We will train only these two layers (not the rest of the network) on the target task.

Since ImageNet images focus on the object while PASCAL VOC images have an object in the scene, and since PASCAL VOC also has a "background" class, we train with a sliding window apprach. We

sample square 500 patches from each image of length $s = \min(w, l)/\lambda$ for various λ and rescale them to 224×224. To label the patch, we consider it P and the ground truth bounding boxes B. It's labeled as a positive example for class o if (1) $|P \cap B_o| \geq 0.2|P|$, (2) $|P \cap B_o| \geq 0.6|B_o|$, and (3) the patch overlaps with no more than one object. Most of the patches in the PASCAL VOC dataset come form teh background class, so we make sure to sample only 10% of all the background patches. At test time, we sample M patches and compute a score for each class as follows (where $y(C_n|P_i)$ is the score of class C_n given patch P_i and k is a parameter):

$$\text{score}(C_n) = \frac{1}{M} \sum_{i=1}^{M} y(C_n|P_i)^k \tag{1}$$

We cross validate to set $k = 5$.

6 Experiments

We use dropout and data augmentation as in the original AlexNet paper. We train on a big GPU instead of two small GPUs. We train with SGD and cut the learning rate by a factor of 10 when validation error stops improving. We train for 3 epochs (1 week).

We set state of the art on PASCAL VOC 2007 object recognition with our transfer. We also do pretty well on the 2012 challenge too. If we don't pretrain, we do a LOT worse. The ImageNet dataset has some classes in common with PASCAL VOC. If we just pick some ImageNet classes at random and train on those, we get a slight performance drop (so common classes between source and target helps, but isn't required). Using a larger dataset (with more classes) requires us to make the FC layers larger, but gives better performance.

We tried using one adaptation layer instead of FC_A and FC_B. This hurts performance. Adding adaptation layers also hurts.

We also do reasonably well at localization on PASCAL VOC. We just average scores for each class over all locations and pick the winning class. The location is just the sliding window location with the highest score for that class.

PASCAL VOC 2012 action recognition has 4K train (and 4K test) images of people doing actions like "walking" and "taking photo". Since the person bounding boxes are given, we just train on those boxes. Transferring does much better than directly training on the PASCAL VOC dataset. We also found that fine-tuning the FC_6 and FC_7 layers (instead of only the FC_A and FC_B layers) sets state of the art. We also do pretty well at localizing people, even if we don't have the ground truth bounding box.

Our model struggles when two objects are very close to each other (e.g. "person" sitting on "chair"). We also struggle with very large objects because it isn't fully captured in any patch.

7 Conclusion

Transferring from ImageNet to PASCAL VOC (with two FC layers replacing the last FC layer and a sliding window approach for training), is much better than training on PASCAL VOC directly.