Discriminative Filter Learning within a CNN

Deep Learning for Fine-grained Visual Classification

Denis Zavadski & Kevin Siswandi

Heidelberg Collaboratory for Image Processing (HCI)

Supervisors: `Prof Dr Bjoern Ommer, Timo Milbich` — *Deep Vision ST2019*

RUPRECHT-KARLS-
UNIVERSITAT HEIDELBERG
ZUKUNFT SEIT 1386

Abstract

We implement the recently proposed Discriminative Filter Learning (DFL) within a CNN [CVPR, 2018] for Fine-grained Visual Classification (FGVC), which produces state-of-the-art results on public fine-grained recognition datasets such as Stanford Cars and CUB-200-2011. Furthermore, we compare our implementation with the Bilinear CNN model [ICCV, 2015], which is widely regarded as the classical benchmark for FGVC. The DFL method is trained end-to-end without any bounding boxes or part annotations, highly interpretable, and works across different fine-grained visual domains. We also extend the basic DFL based on VGGNet as described in the original paper to use ResNet architecture as the base model, which is computationally more efficient.

Introduction

Fine-grained Visual Classification (FGVC) refers to the task of classifying objects that are visually very similar and belong to the same basic-level category. such as species of the bird, model of the car and type of the aircraft. Convolutional Neural Networks were proposed to solve FGVC, but the popular CNNs for coarse-grained classification (e.g. VGG, ResNet, Inception) can only achieve moderate performance, because it is relatively difficult for them to focus on the subtle differences of object parts – which can be easily overwhelmed by pose, viewpoint, or location – without special design.

In order to give more attention to few important object parts, classic applications of deep learning to FGVC rely on either part annotations or building weakly-supervised multi-stage CNN frameworks (e.g. classification network assisted by localization network). However, manual annotations of fine-grained dataset is expensive and requires specialized knowledge, while multi-stage frameworks are often time-consuming and tedious to train and tune. In this project, we address these issues using a bank of 1x1 filters in an asymmetric multi-stream architecture to learn discriminative patches directly within a CNN. Our main motivation is to educate ourselves on the latest ideas and development in this rapidly growing subfield of computer vision.

Key Strengths

1. Simple yet effective: involves only single-stage (end-to-end) learning and can achieve state-of-the-art results even in the absence of bounding boxes and other annotations.
2. Interpretable: learned discriminative patches can be mapped to the original image and visualized.
3. Annotation-free and consistent in performance across fine-grained visual domains.

Methods

The core component of the network in Figure 1 is a 1x1 convolutional layer (Conv6) followed by a GMP layer and then a classifier for discriminative patch learning (P-stream) , taking the feature map of a pre-trained model (e.g. ResNet-50) as the stream input. Meanwhile, another stream preserves the further convolutional and fully connected layers to focus on the global features (G-stream). For Conv6 to learn discriminative patch detectors, we impose supervision at the 1x1 filters by introducing a Cross-Channel Pooling layer followed by a softmax loss layer at the side branch.

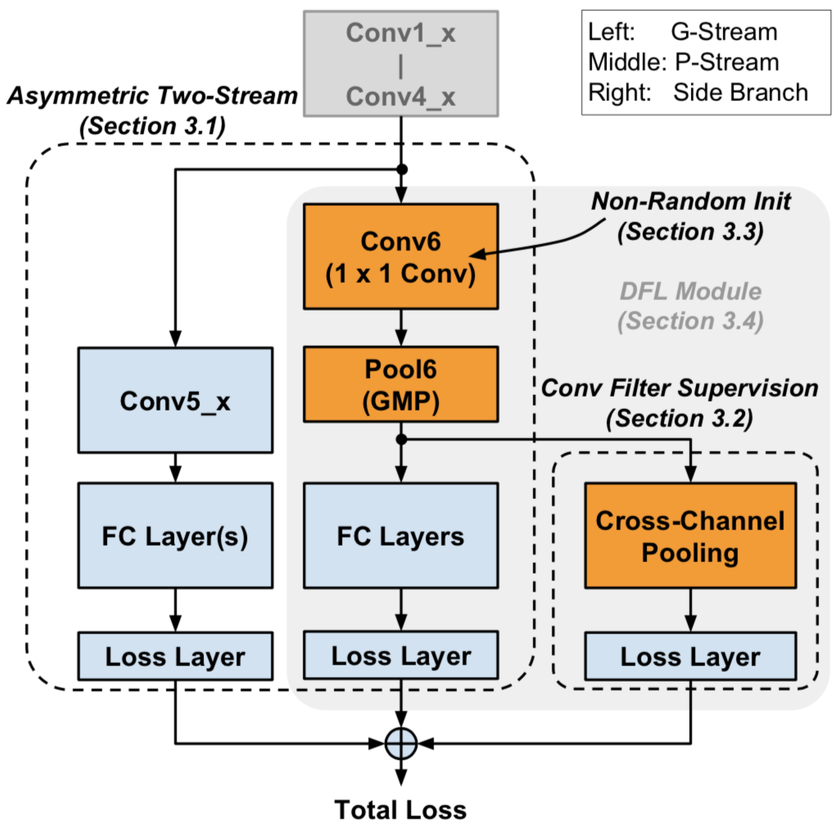

Figure 1: Overview of DFL-CNN: a) asymmetric two-stream architecture to learn both the discriminative patches and global features, b) supervision imposed to learn discriminative patch detectors and c) non-random layer initialization.

The DFL method is able to accurately localize discriminative regions or parts because:

1. it works well with off-the-shelf modern feature extractors such as VGG-16 and ResNet-50 due to the small receptive field of their feature maps.
2. the discrimative patches do not need to be shared across categories, thus avoiding the tradeoff between recognition and localization (in contrast to previous multi-stage approaches that first find the corresponding parts and then compare their appearance).

Implementation Details

- Filter supervision: with k filters per class and M classes, Cross-Channel Pooling averages the output of Pool6 ($kM \times 1 \times 1$) across every group of k dimensions, resulting in an M-dimensional vector that is fed into a softmax layer to encourage learning of discriminative patches.
- Layer initialization: the patch detectors of a certain class are initialized by patch representations from the samples of that class, using cluster centers of the patches with largest L2-norm across channels in the feature map.

Results

Table 1 shows how our own implementation of DFL and B-CNN compare with fine-tuned VGGNet, fine-tuned ResNet, and a annotation-based approach using bounding boxes. It should be noted that the performance of DFL is insensitive to bounding boxes or any extra annotations because the method is able to utilize discriminative patches without image localization.

Method	Dataset	Accuracy (top1)
DFL*/DFL (original)	Stanford Cars	87.0%/93.8%
DFL*/DFL (original)	CUB-200-2011	–/87.4%
B-CNN*/B-CNN (original)	Stanford Cars	82.5%/91.3%
B-CNN*/B-CNN (original)	CUB-200-2011	–/84.1%
Fine-tuned VGG-19	Stanford Cars	84.9%
Fine-tuned VGG-19	CUB-200-2011	77.8%
Fine-tuned ResNet-50	Stanford Cars	91.7%
Fine-tuned ResNet-50	CUB-200-2011	84.1%
CoSeg(+BBox)	Stanford Cars	92.8%
CoSeg(+BBox)	CUB-200-2011	82.6 %

Table 1: Our simple, quick-and-dirty DFL and B-CNN implementations demonstrate competitive performance compared to other state-of-the-art results in the literature. We expect to achieve even higher performance after hyperparameter tuning and other improvements in the remaining two weeks. [*] indicates our own implementation

The basic idea behind DFL is that each filter acts as a discriminative patch detector. Figure 2 shows the visualization of the part with the highest activation for some classes, obtained by remapping the learned patches in the feature map back to the original images.

Figure 2: Visualization of the top discriminative patch for each image learned by DFL. The results are consistent with human perception: headlight (BMW 1 Series), air intake (Bugatti Veyron), frontal face (Mercedes-Benz 300 Class), and the distinctive tail (Tesla Model S).

Conclusions

- Fine-grained Visual Classification (FGVC) is challenging because discriminative local features are difficult to define, vary from object to object, and costly to label.
- Discriminative Filter Learning (DFL) is an approach to Fine-grained Visual Classification that can be trained in an end-to-end fashion with only category labels and without any extra annotations.
- DFL learns high-quality discriminative parts in an unsupervised manner and can obtain state-of-the-art results on both rigid/non-rigid fine-grained datasets.

Future Work

Because there are still two weeks after the poster session before the submission deadline, we will fine-tune our DFL implementation to achieve results closer to the state-of-the-art, as well as work on some improvements to our implementation of Bilinear CNN.

Figure 3: For the same fine-grained class, one can predict the wrong category when focusing only on one part/region. Selectively erasing one discriminative part can encourage the CNN to extract discriminative features from other parts. A similar idea applies to selective cropping and this can be used to guide data augmentation procedure for FGVC.

Furthermore, since DFL can find discriminative parts by weakly-supervised learning, it is possible to incorporate a data augmentation strategy similar to [1] that selectively crops and drops certain discriminative regions in an image (see Figure 3). This can potentially improve the generalization of the model and has been recently shown to be highly effective for FGVC. In FGVC, random data augmentation such as random image cropping, rotation, and color distortion often introduces significant noises rather than increase training efficiency.

References

[1] T. Hu, H. Qi, Q. Huang, and Lu Y. See better before looking closer: Weakly supervised data augmentation network for fine-grained visual classification. *ArXiv Preprint*, 2019.

[2] T. Lin, A. RoyChowdury, and S. Maji. Bilinear CNN models for fine-grained visual recognition. *International Conference on Computer Vision (ICCV)*, 2015.

[3] Y. Wang, V.I. Morariu, and L.S. Davis. Learning a discriminative filter bank within a CNN for fine-grained recognition. *Conference on Computer Vision and Pattern Recognition*, June 2018.