Bilinear CNN Models for Fine-grained Visual Recognition Supplementary Material

Tsung-Yu Lin Aruni RoyChowdhury Subhransu Maji
University of Massachusetts, Amherst
{tsungyulin, arunirc, smaji}@cs.umass.edu

Overview

We study the effect of signed square-root and ℓ_2 normalization on the accuracy. We also present visualizations on the learned B-CNN models and the most confused classes for the cars and aircraft datasets.

1. Effect of normalization

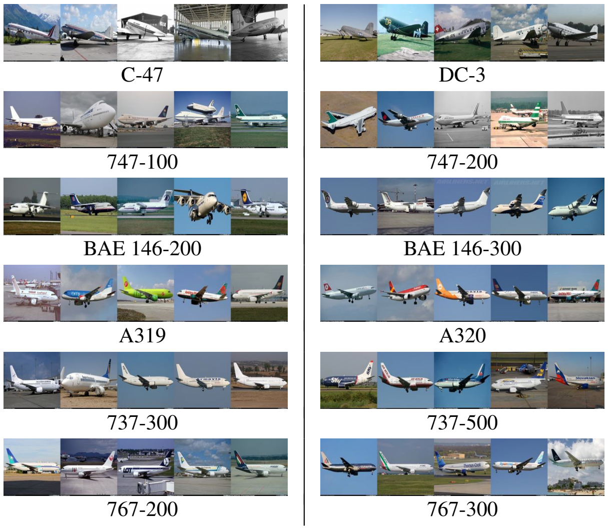
Tab. 1 shows the results of the B-CNN (D,M) model w/o fine-tuning using various normalizations of the bilinear vector. These experiments on the CUB-200-2011 dataset w/o bounding boxes during training and testing (Table 1, column "birds" in the main paper). The model with both square-root and ℓ_2 normalization achieves 80.1% accuracy. Only square-root normalization results in small drop in accuracy accuracy to 79.4%. Only ℓ_2 normalization causes a larger drop in accuracy to 77.3%. No normalization at all is significantly worse at 74.7% accuracy. This shows that both these normalizations are useful and square-root has a higher effect on the performance than ℓ_2.

normalization	**accuracy**	**mAP**
square-root + ℓ_2	80.1	81.3
square-root only	79.4	77.9
ℓ_2 only	77.3	79.6
none	74.7	70.9

Table 1. Effect of normalization on the B-CNN (D,M) model w/o fine-tuning on the CUB-200-2011 dataset ("birds" setting).

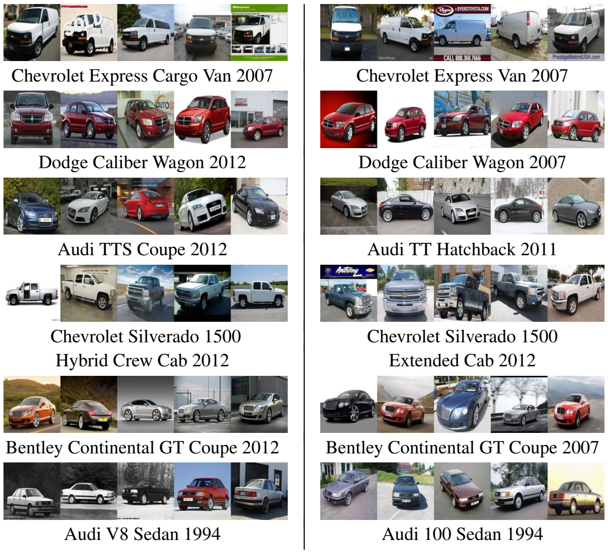
2. Common mistakes

Fig. 1 and 2 show the six most confused classes with images that are most confused by our B-CNN (D,M) model for the aircraft and cars datasets. Some of the confusions are among classes that are highly similar. For example, the Douglas C-47 is a military transport aircraft developed from the Douglas DC-3. Confusions in cars are among makes from different years and styles from the same manufacturer.



C-47 | DC-3
747-100 | 747-200
BAE 146-200 | BAE 146-300
A319 | A320
737-300 | 737-500
767-200 | 767-300

Figure 1. Top confused classes along with the most confused images for each class in the aircraft variant dataset.



Chevrolet Express Cargo Van 2007 | Chevrolet Express Van 2007
Dodge Caliber Wagon 2012 | Dodge Caliber Wagon 2007
Audi TTS Coupe 2012 | Audi TT Hatchback 2011
Chevrolet Silverado 1500 Hybrid Crew Cab 2012 | Chevrolet Silverado 1500 Extended Cab 2012
Bentley Continental GT Coupe 2012 | Bentley Continental GT Coupe 2007
Audi V8 Sedan 1994 | Audi 100 Sedan 1994

Figure 2. Top confused classes along with the most confused images for each class in the cars dataset.

3. Visualization of the learned B-CNN models

We present visualizations of the fine-tuned B-CNN (D,M) model on the aircraft (Fig. 3) and cars (Fig. 4) datasets. The top activations of several filters of the D-Net and M-Net are shown on the training set. One can see from these figures that the bilinear model learns to recognize various highly localized attributes.

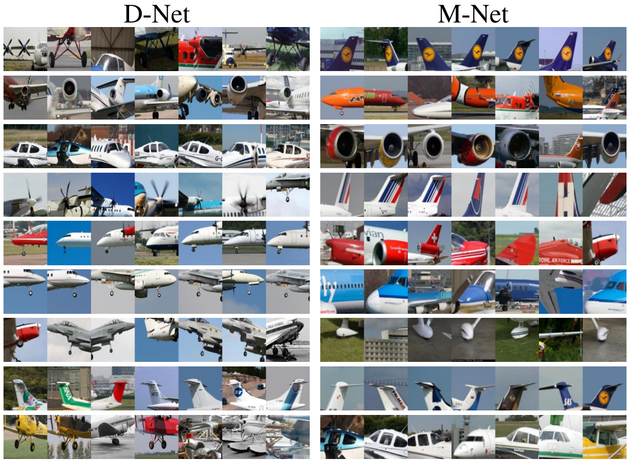


Figure 3. Patches with the highest activations for several filters of the fine-tuned B-CNN (D, M) model on the training images of the aircraft variant dataset. The model learns to recognize highly localized attributes such as engines, propeller, undercarriage, tail stabilizers, cockpit, sometimes even those of a specific color.

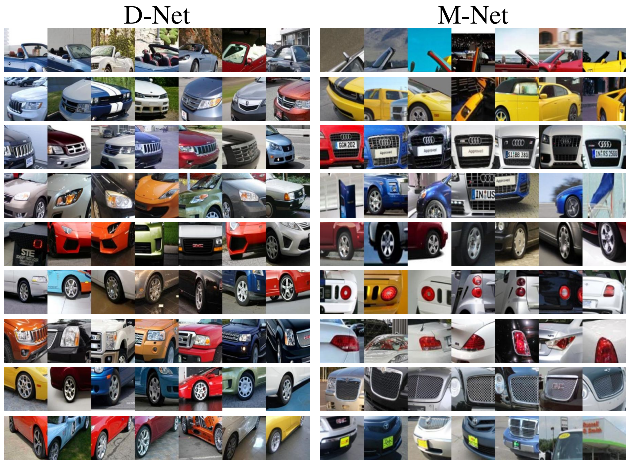


Figure 4. Patches with the highest activations for several filters of the fine-tuned B-CNN (D, M) model on the training images of the cars dataset. These filters learn to recognize attributes such as convertibles (top row, D-Net and M-Net), front bumpers and wheels (D-Net rows 2, 3, 8 and 9), various styles of tail lights (M-Net, rows 6 and 7), specific company logos as a pattern (M-Net row 3), of a specific color (M-Net rows 2 and 4).