Pose Aware Fine-Grained Visual Classification using Pose Experts

Kushagra Mahajan[1]; Tarasha Khurana[1]; Ayush Chopra[1]; Isha Gupta[1]; Chetan Arora[1]; Atul Rai[2]
[1]IIIT Delhi, [2]Staqu Technologies

Abstract

We focus on the problem of fine-grained visual classification (FGVC). We posit that unreasonable effectiveness of the state-of-the-art in this area is because of similar object categories present in the ImageNet dataset, which allows such models to be pretrained on a much larger set of samples and learn generic features for those object categories.

We observe that in FGVC problems, the objects are captured from a small set of viewing angles only. We notice that subtle differences between object categories are difficult to pick from an arbitrary angle but easier to identify from a similar pose. We show in this paper that training specialized pose experts, focusing on classification from a single, fixed pose, and combining them in an ensemble style framework successfully exploits the structure in the problem. To highlight the contribution when the target category features may not be available in a pretrained network, we test on footwear class.

Datasets Description

Footwear Dataset: Contribution of ~1000 scraped images' dataset corresponding to 12 classes for four different poses. The classes spanned across: Ankle Boots, Knee High Boots, Formal Shoes, Casual Shoes, Sandals, Slippers, Ballerinas, Boat Shoes, Clogs, Ethnic Chappal, Ethnic Juti, Heels. The four poses used were: Facing Left, Facing Right, Diagonal Facing Left and Diagonal Facing Right.

Fig 1. Representative images from the contributed Footwear dataset.

Fig 2. Representative images from the CUB dataset (Left), Stanford Cars dataset (Middle) and FGVC-Aircrafts dataset (Right) divided into poses.

Our Approach

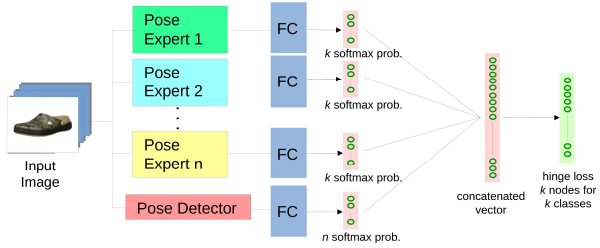

Fig 3. Proposed Network Architecture

Experimental Results

Table 1. Performance of Pose Experts v/s Single Network for all poses. 'PE Network' denotes Pose Ensemble Network

Classes	Single Network			PE Network		
	LeNet	AlexNet	VGG16	LeNet	AlexNet	VGG16
4	72.2	87.3	88.1	80.7	90.5	90.8
8	63.7	74.2	73.2	71.3	82.3	82.7
12	52.1	73.4	72.1	59.6	79.1	79.3

Table 2. Usefulness of pose experts with reduced (R) networks.

Classes	Single Network		PE Network	
	R-AlexNet	R-VGG16	R-AlexNet	R-VGG16
4	88.3	88.8	93.1	94.1
8	77.5	78.6	84.5	86.3
12	76.2	77.5	82.6	83.5

The tables indicate the viability of replacing a state-of-the-art single deep network with multiple smaller pose experts. An ensemble of shallower networks with less number of trainable parameters is thus able to outperform the single deeper networks.

Table 3. Performance comparison with state-of-the-art on standard datasets

	birds		cars		aircrafts	
	\bbox	bbox	\bbox	bbox	\bbox	bbox
Proposed	76.3	78.4	87.9	92.0	82.5	83.9
MixDCNN	-	81.1	-	-	-	-
BCNN	84.1	85.1	91.3	-	84.1	-
BGL	75.9	80.4	86.0	90.5	-	-
SCDA	80.5	-	85.9	-	79.5	-

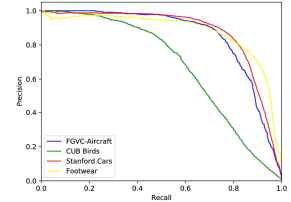

Fig 4. Precision-Recall curves for the four datasets.

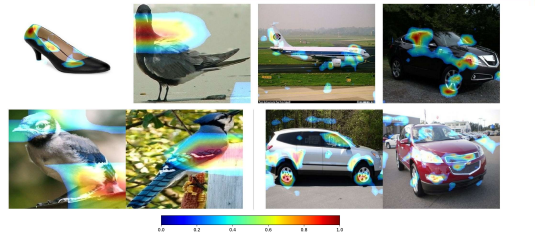

Fig 5. Activation Maps. Row 1 shows maps from the 4 datasets. Row 2 shows 2 pairs of images, each belonging to the same class with different viewpoints & discriminative regions.

Fig 6. Top two most confused class pairs in each of the 4 datasets

Conclusion

We posit that it's harder for a single network, deep or shallow, to overcome large intra-class variance and small inter-class variance, as observed from an arbitrary view, in a data scarce FGVC problem. The classification problem gets significantly simplified when viewing objects from similar pose.

References

[1] Ge *et al.* "*Fine-grained classification via mixture of deep convolutional neural networks.*" IEEE WACV 2016.

[2] Shen *et al.* "*Selective convolutional descriptor aggregation for fine-grained image retrieval,*" IEEE TIP 2017.

[3] Lin *et al.* "*Bilinear CNN models for fine-grained visual recognition.*" IEEE ICCV 2015.