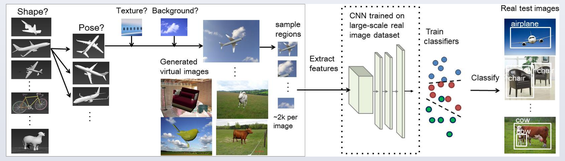
What Do Deep CNNs Learn About Objects?

Xingchao Peng, Baochen Sun, Karim Ali, and Kate Saenko

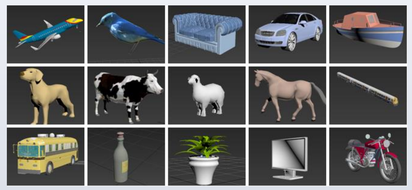
University of Massachusetts Lowell

OVERVIEW



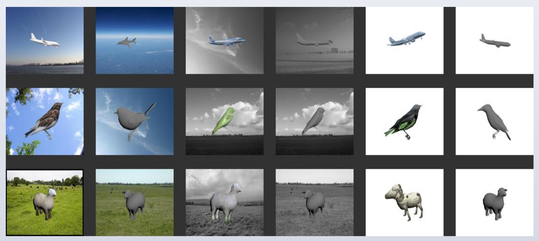
- Deep convolutional neural networks learn extremely powerful image representations, yet most of that power is hidden in the millions of deep-layer parameters.
- We explore the invariance of CNNs to various intra-class variations by simulating different rendering conditions

3D MODELS



- Collected models for 20 objects in PASCAL dataset
- Selected about 20 models per category manually from 3D warehouse
- 15 poses per model were generated by randomly rotating the original model from 0 to 15 degrees in each of the three axes

VIRTUAL IMAGE EXAMPLES

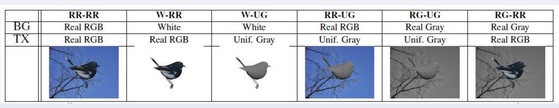


EXPLORING THE INVARIANCES

- Select image formation parameters
- Generate a batch of synthetic 2D images with those parameters
- Sample positive and negative patches for each class
- Extract hidden CNN layer activations from the patches as features
- Train a classifier for each object category
- Test the classifiers on real images.

MAIN IDEA

	RR-RR	W-RR	W-UG	RR-UG	RG-UG	RG-RR
BG	Real RGB	White	White	Real RGB	Real Gray	Real Gray
TX	Real RGB	Real RGB	Unif. Gray	Unif. Gray	Unif. Gray	Real RGB



- To explore the representation of CNN features, we choose a subset of factors that can easily be modeled using simple computer graphics techniques, namely, **object texture** and **color**, **context/background** appearance and **color**, **3D pose** and **3D shape**.
- We use these synthetic data to train deep convolutional neural networks and test on PASCAL test set.

EFFECT OF REAL VIEWPOINT

We are interested here in objects whose frontal view presentation differs significantly (eg: the side-view of a horse vs a frontal view).

$Pool_5$ Activations

Top 10 regions with strongest activations for pool5 units using the method of RCNN. For each unit we show results on (top to bottom): real PASCAL images, RR-RR, W-RR, W-UG.

RESULTS

PASC-FT	aero	bike	bird	boat	botl	bus	car	cat	chr	cow	tab	dog	hse	mbik	pers	plt	shp	sofa	trn	tv	mAP
RR-RR	50.9	57.5	28.3	20.3	17.8	50.1	37.7	26.1	11.5	27.1	2.4	25.3	40.2	52.2	14.3	11.9	40.4	16.3	15.2	32.2	28.9
W-RR	46.5	55.8	28.6	21.7	21.3	50.6	46.6	28.9	14.9	38.1	0.7	27.3	42.5	53.0	17.4	22.8	30.4	16.4	16.7	43.5	31.2
W-UG	54.4	49.6	31.5	24.8	27.0	42.3	62.9	6.6	21.2	34.6	0.3	18.2	35.4	51.3	33.9	15.0	8.3	33.9	2.6	49.0	30.1
RR-UG	55.2	57.8	24.8	17.1	11.5	29.9	39.3	16.9	9.9	35.1	4.7	30.1	37.5	53.1	18.1	9.5	12.4	18.2	2.1	21.1	25.2
RG-UG	49.8	56.9	20.9	15.6	10.8	25.6	42.1	14.7	4.1	32.4	9.3	20.4	28.0	51.2	14.7	10.3	12.6	14.2	9.5	28.0	23.6
RG-RR	46.5	55.8	28.6	21.7	21.3	50.6	46.6	28.9	14.9	38.1	0.7	27.3	42.5	53.0	17.4	22.8	30.4	16.4	16.7	43.5	31.2
IMGNET	aero	bike	bird	boat	botl	bus	car	cat	chr	cow	tab	dog	hse	mbik	pers	plt	shp	sofa	trn	tv	mAP
RR-RR	34.3	34.6	19.9	17.1	10.8	30.0	33.0	18.4	9.7	13.7	1.4	17.6	17.7	34.7	13.9	11.8	15.2	12.7	6.3	26.0	18.9
W-RR	35.9	23.3	16.9	15.0	11.8	24.9	35.2	20.9	11.2	15.5	0.1	15.9	15.6	28.7	13.4	8.9	3.7	10.3	0.6	28.8	16.8
W-UG	38.6	32.5	18.7	14.1	9.7	21.2	36.0	9.9	11.3	13.6	0.9	15.7	15.5	32.3	15.9	9.9	9.7	19.9	0.1	17.4	17.1
RR-UG	26.4	36.3	9.5	9.6	9.4	5.8	24.9	0.4	1.2	12.8	4.7	14.4	9.2	28.8	11.7	9.6	0.7	4.9	0.1	12.2	11.6
RG-UG	32.7	34.5	20.2	14.6	9.4	7.5	30.1	12.1	2.3	14.6	9.3	15.2	11.2	30.2	12.3	11.4	2.2	9.9	0.5	13.1	14.7
RG-RR	26.4	38.2	21.0	15.4	12.1	26.7	34.5	18.0	8.8	16.4	0.4	17.0	20.9	32.1	11.0	14.7	18.4	14.8	6.7	32.0	19.3

Table 1 Results of synthetic data

The generation settings RR-RR, W-RR, W-UG, RG-RR with PASC-FT all achieve comparable performance, despite the fact that W-UG has no texture and no context. For the IMGNET network, the trend is similar

IMGNET	aero	bike	bird	boat	botl	bus	car	cat	chr	cow	tab	dog	hse	mbik	pers	plt	shp	sofa	trn	tv	mAP
front	24.9	38.7	12.5	9.3	9.4	18.8	33.6	13.8	9.7	12.5	2.1	18.0	19.6	27.8	13.3	7.5	10.2	9.6	13.8	28.8	16.7
front,side	24.3	36.8	19.0	17.7	11.9	26.6	36.0	10.8	9.7	15.5	0.9	21.6	21.1	32.8	14.2	12.0	14.3	12.7	10.1	32.6	19.0
front,side,intra	33.1	40.2	19.4	19.6	12.4	29.8	35.3	16.1	5.2	16.5	0.9	19.7	19.0	34.9	15.8	11.8	19.7	16.6	14.3	29.8	20.5
PASC-FT	aero	bike	bird	boat	botl	bus	car	cat	chr	cow	tab	dog	hse	mbik	pers	plt	shp	sofa	trn	tv	mAP
front	41.8	53.7	14.5	19.1	11.6	42.5	40.4	25.5	9.9	24.5	0.2	29.4	37.4	47.1	14.0	11.9	18.9	12.7	22.6	38.8	25.8
front,side	45.6	50.2	24.4	28.8	17.4	51.9	41.8	24.5	7.2	27.9	9.2	23.1	37.0	51.3	17.8	13.2	28.6	18.9	9.3	37.8	28.3
front,side,intra	54.2	55.5	22.7	27.0	20.5	52.6	40.1	26.8	8.1	27.3	2.3	30.6	36.6	53.3	17.8	14.2	34.1	26.4	19.3	37.5	30.3

Table 2 Results of Synthetic Pose

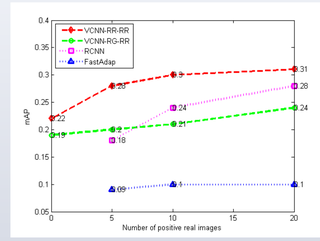
We change the number of views used in each experiment, but keep the total number of synthetic training images (RR-RR) exactly the same, by generating random small perturbations around the main view. Results indicate that for both networks adding side view to front view gives a boost, but improvement from adding the third view is marginal

Net	Views	aero	bike	bird	bus	car	cow	dog	hrs	mbik	shp	trn	tv	mAP
PASC-FT	**all**	64.2	69.7	50	62.6	71	58.5	56.1	60.6	66.8	52.8	57.9	64.7	61.2
PASC-FT	**-random**	62.1	70.3	49.7	61.1	70.2	54.7	55.4	61.7	67.4	55.7	57.9	64.2	60.9
PASC-FT	**-front**	61.7	67.3	45.1	58.6	70.9	56.1	55.1	59.0	66.1	54.2	53.3	61.6	59.1
PASC-FT	**-side**	62.0	70.2	48.9	61.2	70.8	57.0	53.6	59.9	65.7	53.7	58.1	64.2	60.4
PASC-FT(-front)	**-front**	59.7	63.1	42.7	55.3	64.9	54.4	54.0	56.1	64.2	55.1	47.4	60.1	56.4

Table 3 Result of Realistic Pose

Results point to important and surprising conclusions regarding the representational power of the CNN features. Note that mAP drops by less than 2% when detectors exclusively trained by removing either view are tested on the PASCAL VOC test set. In a last experiment, we reduce the fine-tuning training set by removing front-view objects, and note a larger mAP drop of 5 points (8%), but much less than one may expect.

VCNN MODEL



When the real annotated images are limited or not available, eg. for a novel category, VCNN performs much better than RCNN and the Fast Adaptation method.

SAMPLE DETECTIONS

Sample detections of our VCNN model

CONCLUSION

We investigated the representation of ConvNet to various factors in the training data: 3D pose, foreground texture and color, back-ground image and color. To simulate these factors we used synthetic data generated from 3D CAD models and a few real images. Our results demonstrate that the popular deep ConvNet fine-tuned for detection on real images for a set of categories is indeed invariant to these factors. Training on synthetic images with those variations leads to similar performance as training on synthetic images without those variations. However, if the network is not fine-tuned for the task on real images, its invariance is diminished.

ACKNOWLEDGMENTS

This research was supported by NSF award #1212928 and by DARPA.