UCLA ENGINEERING
Henry Samueli School of Engineering and Applied Science
Birthplace of the Internet

An Empirical Evaluation of Current Convolutional Architectures' Ability to Manage Nuisance Location and Scale Variability.

Nikolaos Karianakis, Jingming Dong, Stefano Soatto

UCLA VISION LAB

IEEE 2016 Conference on Computer Vision and Pattern Recognition

CVPR 2016

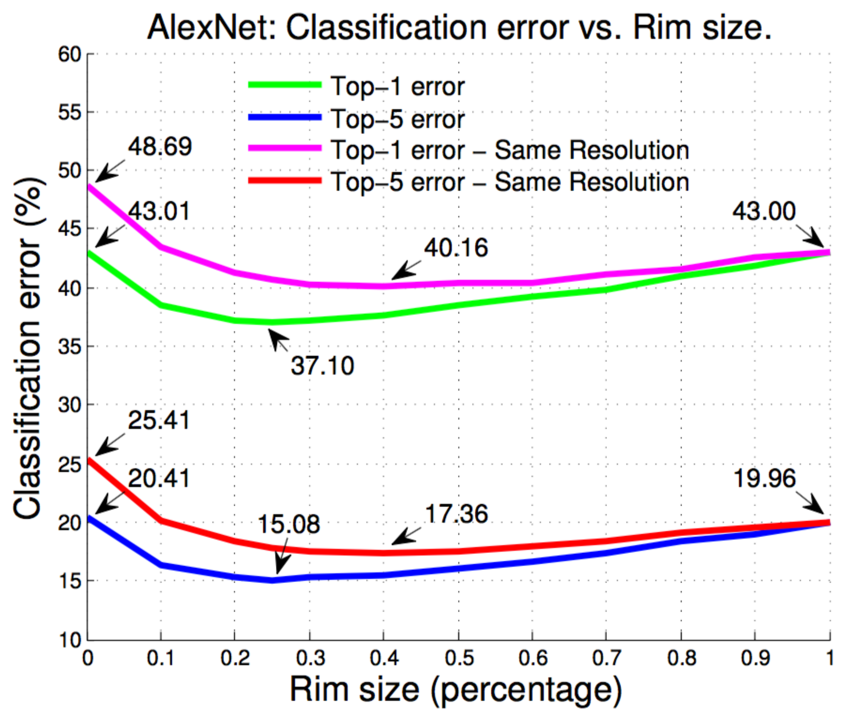
Abstract

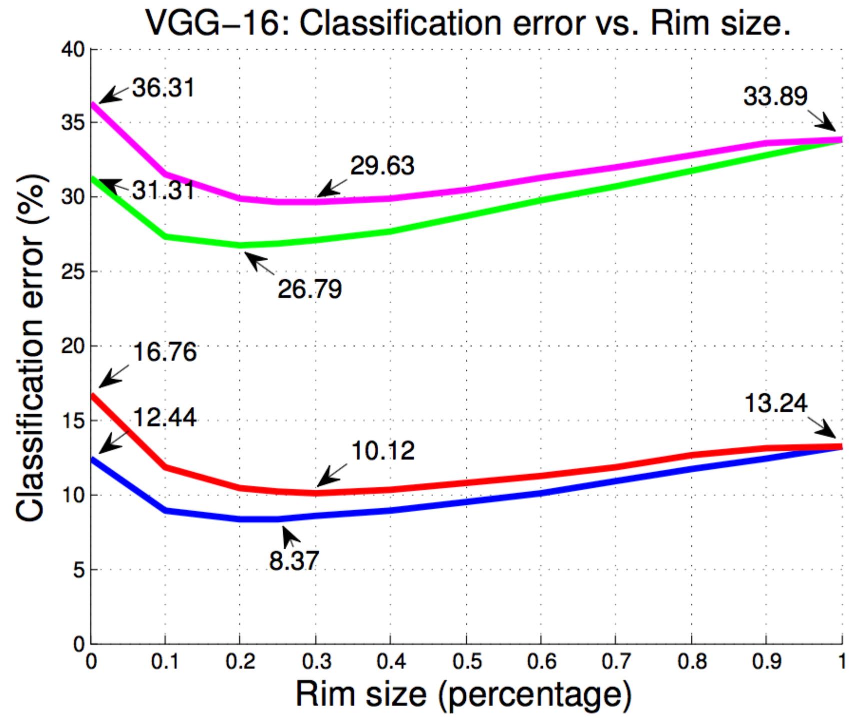
We conduct an empirical study to test the ability of Convolutional Neural Networks (CNNs) to reduce ***the effects of nuisance transformations*** *of the input data, such as location, scale and aspect ratio. We isolate factors by adopting a common convolutional architecture either deployed globally on the image to compute class posterior distributions, or* ***restricted locally to compute class conditional distributions given location, scale and aspect ratios of bounding boxes*** *determined by proposal heuristics. In theory, averaging the latter should yield inferior performance compared to proper marginalization. Yet empirical evidence suggests the converse, leading us to conclude that –* ***at the current level of complexity of convolutional architectures and scale of the data sets used to train them – CNNs are not very effective at marginalizing nuisance variability.***

We also quantify the effects of ***context*** *on the overall classification task and its impact on the performance of CNNs, and propose* ***improved sampling techniques*** *for heuristic proposal schemes that improve end-to-end performance to state-of-the-art levels. We test our hypothesis on a* ***classification*** *task using the ImageNet Challenge benchmark and on a* ***wide-baseline matching*** *task using the Oxford and Fischer's datasets.*

Trade-off between location-scale and visibility

- Restricting the support **does not just condition on the location-scale group, but also on visibility**.
- A 25% rim yields the lowest top-5 error on the ImageNet validation set for both AlexNet and VGG16.
- This indicates that **the context effectively leveraged by current CNN architectures is limited to a relatively small neighborhood of the object of interest.**





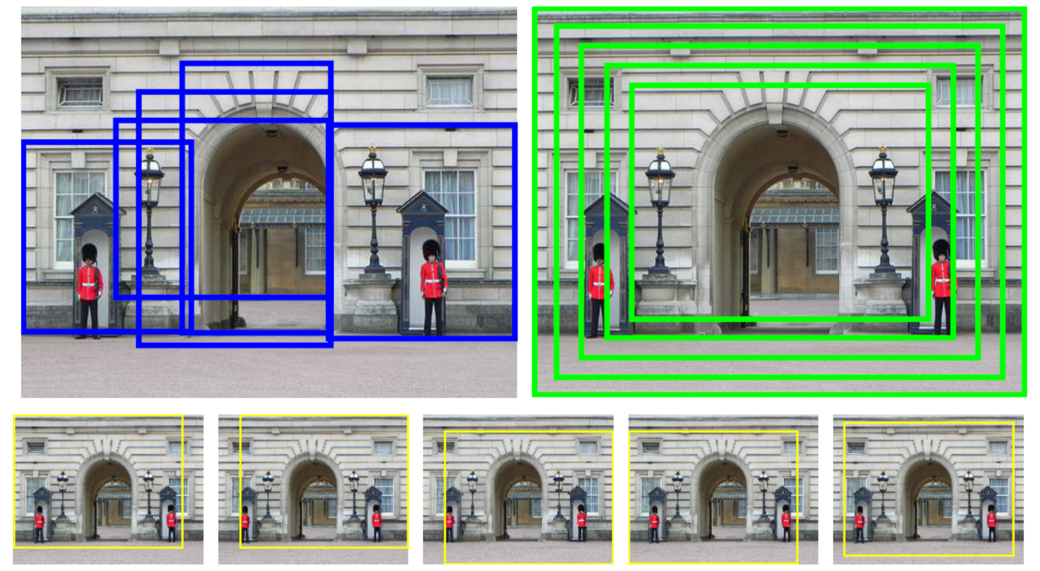
The effect of domain-size pooling

- In line with recent developments concerning **domain-size pooling** [2], averaging conditional densities from the scale group achieves higher classification accuracy than using any single domain size.

Method	AlexNet		VGG16	
Whole image	19.96		13.24	
Ground-Truth Bounding Box (GT)	20.41		12.44	
	Isotropically	Anisotropically	Isotropically	Anisotropically
GT padded with 10 px	17.66	17.65	10.91	10.30
Ave-GT, 4 domain sizes (padded with [0,30] px)	15.96	16.00	9.65	8.90
Ave-GT, 8 domain sizes (padded with [0,70] px)	14.43	14.22	8.66	7.84

Data augmentation with regular and adaptive sampling

- The whole-image classification performance is evaluated with various proposal schemes.
- We test the following strategies:
  - C regular crops, as customary (*e.g.* C=10 or C=50 in 1 or 3 scales).
  - D concentric domain sizes (with their horizontal flip).
  - Generic object proposals (e.g. Edge Boxes).



- We show that **data augmentation strategies with regularly sampled crops do not suffice**.

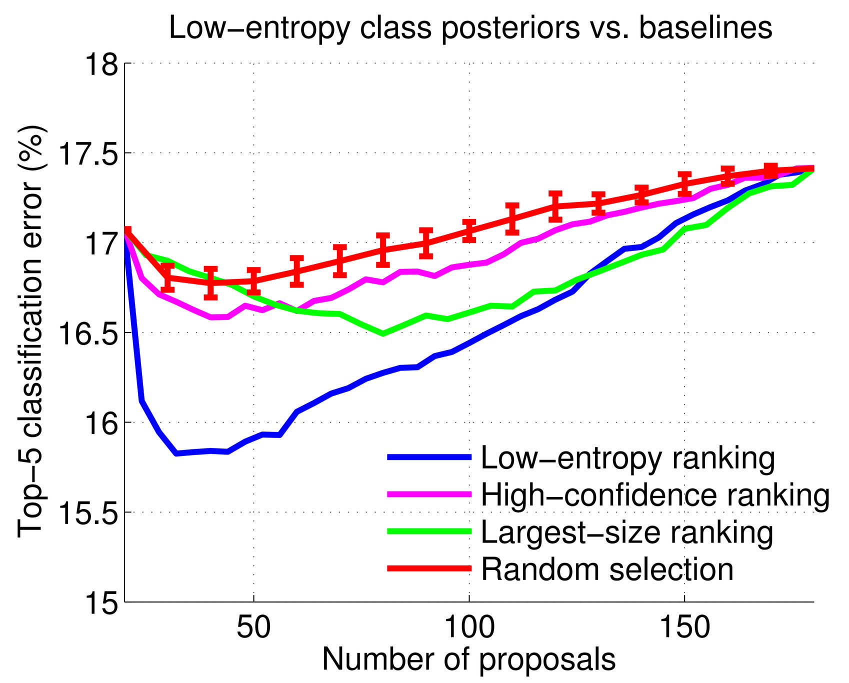
Posterior selection with Rényi entropy

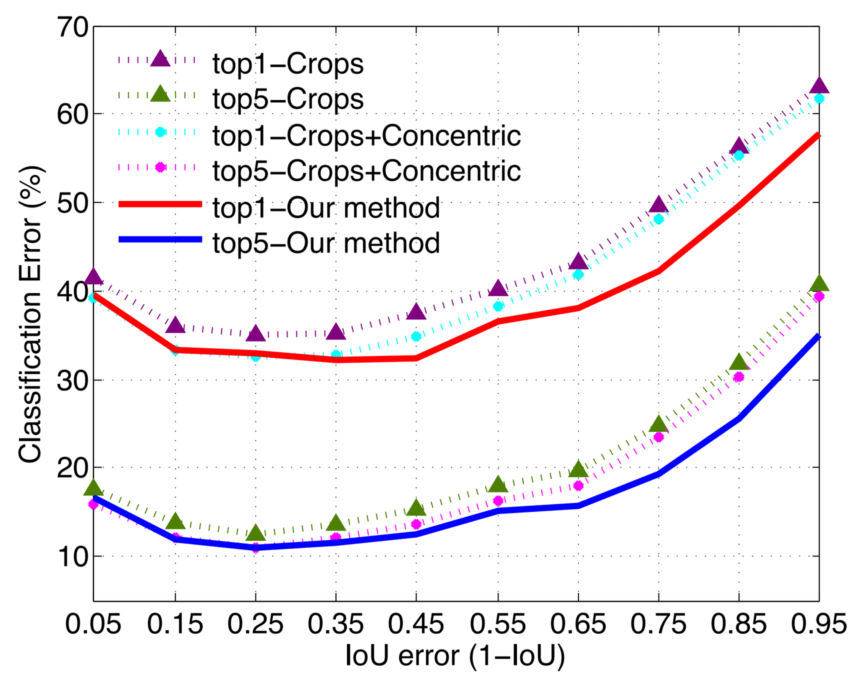
- We choose a small subset (E) of proposals, whose class posterior has the lowest **Rényi entropy** (alpha is set to 0.35).
- The class conditionals for the whole image are approximated as

$$\sum_r p(c|x_{|_r})p(x|_r)$$

where $p(x|_r)$ is defined as 0 for most proposals and equal to the inverse Rényi entropy ($\mathbb{H}^{-1}\{p(c|x_{|_r})\}$) for the E most discriminative samples based on the entropy criterion.

- The entropy criterion chooses more discriminative samples, as opposed to other strategies such as selecting high-confidence proposals or max-out.
- The proposals help the classifier when the regular crops have small overlap with the objects.





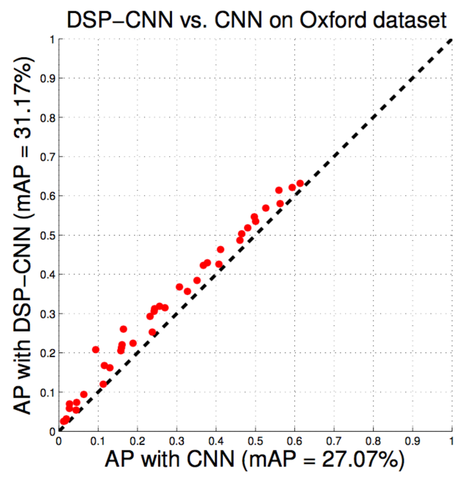
Comparisons

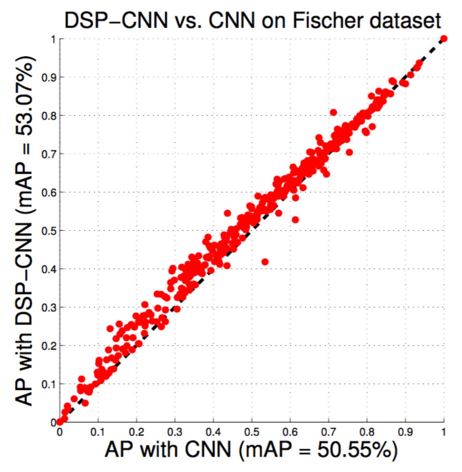
- We evaluate various sampling and inference strategies at the ILSVRC 2014 classification challenge.
- Our method achieves a top-5 classification error of **15.82%** and **7.91%** for AlexNet and VGG16, as opposed to 17.55% and 8.85% respectively using a multi-crop scheme (**~10% relative error reduction**).
- Adaptive sampling introduces a *modest* computational overhead.
- Weighted marginalization based on inverse entropy ("W") further improves the performance.

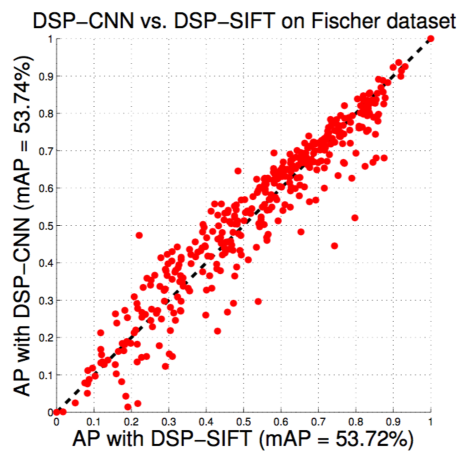
Method			AlexNet			VGG16			#eval	#ave
# crops	# sizes	# proposals	top-1	top-5	t (s/im)	top-1	top-5	t (s/im)		
–	$D=1$	–	43.00	19.96	0.01	33.89	13.24	0.06	1	1
$C=10$	–	–	41.50	18.69	0.06	27.55	9.29	0.48	10	10
$C=50$	–	–	41.01	18.05	0.66	27.44	9.12	1.34	50	50
$C=10\times 3$	–	–	40.58	17.97	0.16	27.23	8.88	1.26	30	30
$C=50\times 3$	–	–	**40.41**	**17.55**	**0.82**	**27.14**	**8.85**	**3.48**	150	150
–	$D=10$	–	40.00	17.86	0.08	28.16	9.46	0.60	10	10
$C=10$	$D=10$	–	39.38	17.08	0.22	26.94	8.83	1.08	20	20
$C=10\times 3$	$D=10$	–	39.36	17.07	0.46	26.76	8.68	1.88	40	40
–	–	$E=40$	40.18	17.53	1.26	25.60	8.24	3.02	160	40
$C=10$	–	$E=20$	38.91	16.63		25.28	7.91		170	30
–	$D=10$	$E=12$	38.05	16.19	1.34	25.19	8.11	4.38	170	22
$C=10$	$D=10$	$E=12$	37.69	15.83		25.11	8.01		180	32
$C=10$	$D=10$	$E=12$ (fast)	37.71	15.88	0.94	25.12	8.08	3.70	180	32
$C=10$	$D=10$	$E=12$ (W, fast)	**37.57**	**15.82**	**1.28**	**25.11**	**8.02**	**3.80**	180	32
$C=10$	$D=10$	$E=12$ (*test set*)	37.417	16.018	–	25.117	7.909	–	180	32

Wide-baseline correspondence

- To test the effect of domain-size pooling on CNN features absent the knowledge of ground-truth location, we develop a domain-size pooled CNN and test it in a wide-baseline correspondence task.
- The regions are selected by a generic low-level detector (Maximally Stable Extremal Regions).
- The DSP-CNN outperforms its counterpart CNN by 5–15% mean AP and performs comparably to the state-of-the-art local descriptor DSP-SIFT [2].







Method	Dim	mAP
Raw patch	4,761	34.79
SIFT	128	45.32
DSP-SIFT	128	**53.72**
CNN-L3	9,216	48.99
CNN-L4	8,192	50.55
DSP-CNN-L3	9,216	52.76
DSP-CNN-L4	8,192	53.07
DSP-CNN-L3-L4	17,408	**53.74**
DSP-CNN-L3 (PCA128)	128	51.45
DSP-CNN-L4 (PCA128)	128	52.33
DSP-CNN-L34 (concat. PCA128)	256	**52.69**

Acknowledgements

• Supported by ARO, ONR, AFOSR.

References

1. *An Empirical Evaluation of Current Convolutional Architectures' Ability to Manage Nuisance Location and Scale Variability.* N. Karianakis, J. Dong and S. Soatto (CVPR 2016)
2. *Domain-Size Pooling in Local Descriptors: DSP-SIFT.* J. Dong and S. Soatto (CVPR 2015)
3. Source code available at http://vision.ucla.edu/~nick/proj/cnn_nuisances/