Assessing the Significance of Performance Differences on the PASCAL VOC Challenges via Bootstrapping

Mark Everingham, S. M. Ali Eslami, Luc Van Gool, Christopher K. I. Williams, John Winn, Andrew Zisserman

October 18, 2013

In the PASCAL VOC challenges, entrants in a particular competition are evaluated in terms of a specified metric. It can happen that some entrants will have similar scores, and it is of interest to assess the *significance* of these differences. For example, we might be interested to know if the highest-scoring entry is significantly better than some of the others. In this note we discuss the use of bootstrap sampling to address this question. We first came across the idea of bootstrapping precision-recall curves in the blog comment by O'Connor (2010), although bootstrapping of ROC curves has been discussed by many authors, e.g. Hall et al (2004); Bertail et al (2009).

In the bootstrap (see e.g. Wasserman, 2004, Ch. 8), the data points (images in our case) are sampled *with replacement* from the original n test points to produce B bootstrap replicates. To compare two methods A and B, we first compute the difference in scores on each bootstrap sample. We then obtain a confidence interval by sorting the differences, and then returning the $\alpha/2$ and $1-\alpha/2$ quantiles, where for example $\alpha = 0.05$ would yield a 95% confidence interval. (This is the percentile interval method described in Wasserman, 2004, Sec. 8.3.) The null hypothesis that A is equivalent to B (at the $1-\alpha$ level). This is rejected if zero is not contained in the obtained confidence interval, leading to the conclusion that method A is statistically significantly better than method B, or vice versa, depending on the result. This procedure is more informative than the unpaired bootstrap confidence intervals in determining whether two methods are significantly different; for example a variation in the hardness of the bootstrap replicates may give rise to overlapping score intervals, even if method A always beats method B.

In the challenge we can also determine the rank of each method on each bootstrap replicate, and thus a confidence interval for the rank of a method (using $\alpha/2$ and $1-\alpha/2$ quantiles as above). This can provide a useful summary of the relative strength of the methods without the need for pairwise comparisons.

For the classification, detection and action classification challenges the overall measure of performance is the average precision (AP), whose computation depends on all of the images. For segmentation, the per class accuracy is computed via the "intersection over union" measure (see Everingham and Winn 2012, sec. 5.4) accumulated over images.

Summary results for all 20 VOC classes highlighting methods that are not significantly different from the highest-scoring one are shown in Table 1 (for classification), Table 2 (for detection) and Table 3 (for segmentation). The entrant abbreviations used in these tables are decoded in Table 4. The results show that for clas-

Mark Everingham, who died in 2012, was the key member of the VOC project. His contribution was crucial and substantial. For these reasons he is included as the posthumous first author of this paper. An appreciation of his life and work can be found in Zisserman et al (2012).

Mark Everingham
University of Leeds, UK

S. M. Ali Eslami
University of Edinburgh, UK

Luc Van Gool
KU Leuven, Belgium and ETH, Switzerland

Christopher K. I. Williams (✉)
School of Informatics,
University of Edinburgh, UK
E-mail: ckiw@inf.ed.ac.uk

John Winn
Microsoft Research, Cambridge, UK

Andrew Zisserman
University of Oxford, UK

sification, NUS_SCM is significantly better than all of the other entrants on all cases except one. For detection there are often two entries in the winning equivalence class, and for segmentation there are often three or four entries in the winning equivalence class for each competition.

These results show that one should not over-interpret small differences in evaluation scores as constituting significant improvements in performance.

References

Alexiou I, Bharath A (2012) Efficient Kernels Couple Visual Words Through Categorical Opponency. In: Proceedings of British Machine Vision Conference

Bertail P, Clémençon SJ, Vayatis N (2009) On Bootstrapping the ROC Curve. In: Koller D, Schuurmans D, Bengio Y, Bottou L (eds) Advances in Neural Information Processing Systems 21, pp 137–144

Carreira J, Caseiro R, Batista J, Sminchisescu C (2012) Semantic segmentation with second-order pooling. In: Proceedings of European Conference on Computer Vision

Chen Q, Song Z, Hua Y, Huang Z, Yan S (2012) Generalized Hierarchical Matching for Image Classification. In: Proceedings of Conference on Computer Vision and Pattern Recognition

Everingham M, Winn J (2012) The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Development Kit, `http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2012/devkit_doc.pdf`

van Gemert J (2011) Exploiting photographic style for category-level image classification by generalizing the spatial pyramid. In: Proceedings of International Conference on Multimedia Retrieval

Hall P, Hyndman R, Fan Y (2004) Nonparametric confidence intervals for receiver operating characteristic curves. Biometrika 91:743–50

Hoai M, Ladicky L, Zisserman A (2012) Action Recognition from Still Images by Aligning Body Parts. URL `http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2012/workshop/segmentation_action_layout.pdf`, slides contained in the presentation by Luc van Gool on Overview and results of the segmentation challenge and action taster

Ion A, Carreira J, Sminchisescu C (2011a) Image segmentation by figure-ground composition into maximal cliques. In: Proceedings of International Conference on Computer Vision

Ion A, Carreira J, Sminchisescu C (2011b) Probabilistic Joint Image Segmentation and Labeling. In: Shawe-Taylor J, Zemel R, Bartlett P, Pereira F, Weinberger K (eds) Advances in Neural Information Processing Systems 24, pp 1827–1835

Karaoglu S, van Gemert J, Gevers T (2012) Object Reading: Text Recognition for Object Recognition. In: Proceedings of ECCV 2012 workshops and demonstrations

Khan F, Anwer R, van de Weijer J, Bagdanov A, Vanrell M, Lopez AM (2012a) Color Attributes for Object Detection. In: Proceedings of Conference on Computer Vision and Pattern Recognition

Khan F, Weijer J, Vanrell M (2012b) Modulating Shape Features by Color Attention for Object Recognition. International Journal of Computer Vision 98(1):49–64

Khosla A, Yao B, Fei-Fei L (2011) Combining Randomization and Discrimination for Fine-Grained Image Categorization. In: Proceedings of Conference on Computer Vision and Pattern Recognition

Li F, Carreira J, Lebanon G, Sminchisescu C (2013) Composite Statistical Inference for Semantic Segmentation. In: Proceedings of Conference on Computer Vision and Pattern Recognition

Nanni L, Lumini A (2013) Heterogeneous bag-of-features for object/scene recognition. Applied Soft Computing 13(4):2171–2178

O'Connor B (2010) A Response to "Comparing Precision-Recall Curves the Bayesian Way?". A comment on the blog post by Bob Carpenter on *Comparing Precision-Recall Curves the Bayesian Way?* at `http://lingpipe-blog.com/2010/01/29/comparing-precision-recall-curves-bayesian-way/`

Russakovsky O, Lin Y, Yu K, Fei-Fei L (2012) Object-centric spatial pooling for image classification. In: Proceedings of European Conference on Computer Vision

van de Sande K, Uijlings J, Gevers T, Smeulders A (2011) Segmentation As Selective Search for Object Recognition. In: Proceedings of International Conference on Computer Vision

Sener F, Bas C, Ikizler-Cinbis N (2012) On Recognizing Actions in Still Images via Multiple Features. In: Proceedings of ECCV Workshop on Action Recognition and Pose Estimation in Still Images

Song Z, Chen Q, Huang Z, Hua Y, Yan S (2011) Contextualizing Object Detection and Classification. In: Proceedings of Conference on Computer Vision and Pattern Recognition

Uijlings J, van de Sande K, Gevers T, Smeulders A (2013) Selective Search for Object Recognition. International Journal of Computer Vision 104(2):154–171

Wang X, Lin L, Huang L, Yan S (2013) Incorporating Structural Alternatives and Sharing into Hierarchy

	aeroplane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	diningtable	dog	horse	motorbike	person	pottedplant	sheep	sofa	train	tvmonitor
CVC	89.4	70.8	69.8	73.9	51.4	84.9	79.7	72.9	63.9	59.6	64.3	64.8	75.8	79.1	91.4	42.9	63.5	62.1	86.7	73.8
CVC_SP	92.1	74.2	73.1	77.5	54.4	85.2	81.9	76.4	65.3	63.6	68.7	69.0	78.3	80.9	91.6	56.1	69.6	65.5	86.7	77.4
IMPERIAL	73.3	33.6	31.1	45.0	17.3	57.8	34.7	46.0	41.3	18.7	30.7	34.6	23.3	39.5	57.3	12.1	23.7	25.6	51.4	36.5
ITI	89.1	62.4	60.1	68.2	33.6	79.8	67.0	70.3	57.5	51.3	55.3	59.4	68.7	74.5	83.2	26.0	57.4	54.1	83.4	64.9
ITI_FUSED	90.5	65.4	65.9	72.3	37.9	80.7	70.6	72.5	60.4	55.4	61.7	63.6	72.5	77.4	86.8	37.8	61.2	57.3	85.8	68.8
NUS_SCM	97.3	84.3	80.9	85.4	61.1	90.0	86.9	89.4	75.5	78.2	75.4	83.2	87.6	90.2	95.1	58.0	79.6	73.8	94.5	80.9
UP	–	–	–	–	–	–	–	–	–	–	–	–	–	–	88.7	–	–	–	–	–

Table 1: Bootstrapped classification results on all classes. The leading methods that are not statistically significantly different from each other are highlighted in gold.

	aeroplane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	diningtable	dog	horse	motorbike	person	pottedplant	sheep	sofa	train	tvmonitor
MISSOURI	51.4	53.7	18.3	15.6	31.7	56.5	47.1	38.7	19.5	32.0	22.1	25.1	50.4	51.9	44.9	12.0	37.8	30.8	50.9	39.4
NEC	65.0	46.8	25.1	24.7	16.1	50.9	44.9	51.6	13.0	26.7	31.0	40.2	39.8	51.6	32.8	12.8	35.8	33.7	48.0	44.7
OLB_R5	47.5	51.7	14.2	12.6	27.4	51.8	44.2	25.5	17.8	30.3	18.2	17.0	47.0	50.9	43.0	09.6	31.3	23.7	44.3	22.1
SYSU_DYNAMIC	50.0	47.0	07.9	03.8	24.9	47.2	42.7	31.3	17.5	24.4	10.1	21.4	43.7	46.4	37.5	07.9	26.4	21.6	43.2	36.5
OXFORD	59.5	54.5	21.9	21.7	32.1	52.6	49.3	40.8	19.1	35.3	28.9	37.2	51.0	49.9	46.1	15.7	39.4	35.7	49.0	42.8
UVA_HYBRID	61.6	52.0	24.6	24.9	20.2	57.1	44.5	53.7	17.4	33.1	38.1	42.9	49.0	59.5	35.8	22.8	40.3	39.7	51.1	49.4
UVA_MERGED	47.2	50.2	18.4	21.5	25.2	53.4	46.3	46.3	17.5	27.9	30.1	35.1	41.7	52.1	43.2	18.2	35.1	31.2	45.5	44.3

Table 2: Bootstrapped detection results on all classes. The leading methods that are not statistically significantly different from each other are highlighted in gold.

	mean	background	aeroplane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	diningtable	dog	horse	motorbike	person	pottedplant	sheep	sofa	train	tvmonitor
BONN_CSI	45.3	84.9	59.5	27.9	44.1	39.7	41.6	52.5	61.6	56.2	13.4	44.4	25.9	42.7	51.6	58.2	51.4	29.4	45.7	28.7	49.8	43.6
BONN_JOINT	46.9	85.1	65.9	29.3	51.7	33.3	43.8	60.1	60.5	52.2	13.5	54.0	32.6	40.3	57.7	57.0	49.0	33.1	53.6	29.1	47.3	37.6
BONN_LINEAR	44.8	83.9	60.3	27.3	46.5	39.9	42.0	57.5	59.2	50.2	09.9	41.5	21.7	42.9	51.8	57.1	50.1	33.4	44.0	29.1	47.8	44.8
NUS_SP	47.2	82.8	52.9	31.0	40.1	44.4	58.6	61.0	52.4	49.0	22.6	37.9	27.2	47.4	52.6	47.1	51.9	35.3	54.9	40.7	54.1	47.7
UVA_NBNN	11.2	63.2	10.4	02.3	02.9	02.9	00.9	30.2	14.7	14.9	00.2	06.0	02.2	05.0	12.2	15.2	23.4	00.5	08.8	03.4	10.7	05.2
METHODS BELOW ALSO TRAINED ON EXTERNAL DATA																						
BONN_CSI	46.7	85.0	64.0	26.7	45.9	42.0	47.1	54.3	58.8	55.1	14.4	48.9	30.6	46.1	52.7	58.4	53.4	31.7	44.4	34.5	45.5	42.6
BONN_JOINT	47.5	85.2	63.8	27.0	56.3	37.8	46.8	58.2	59.4	54.9	11.4	50.9	30.4	45.0	58.6	57.4	48.6	34.8	53.3	32.2	47.8	38.7
BONN_LINEAR	46.7	84.7	63.9	23.8	44.8	40.5	44.9	59.9	58.8	56.9	11.5	45.8	34.9	43.0	55.0	58.3	51.5	34.7	44.2	29.7	50.5	44.1

Table 3: Bootstrapped segmentation results on all classes. The leading methods that are not statistically significantly different from each other are highlighted in gold. The upper part of the has entries trained on the supplied VOC 2012 data only (competition 5); the lower part is for competition 6, which allowed external data to be used.

for Multiclass Object Recognition and Detection. In: Proceedings of Conference on Computer Vision and Pattern Recognition

Wasserman L (2004) All of Statistics. Springer

Xia W, Song Z, Feng J, Cheong LF, Yan S (2012) Segmentation over Detection by Coupled Global and Local Sparse Representations. In: Proceedings of European Conference on Computer Vision

Zisserman A, Winn J, Fitzgibbon A, Gool LV, Sivic J, Williams C, Hogg D (2012) In Memoriam: Mark Everingham. Transactions on Pattern Analysis and Machine Intelligence 34(11):2081–2082

Codename	Cls	Det	Seg	Act	Institutions	Contributors	References
BONN_CSI	·	·	•	·	University of Bonn, Georgia Institute of Technology, University of Coimbra	Joao Carreira, Fuxin Li, Guy Lebanon, Cristian Sminchisescu	Li et al (2013)
BONN_JOINT	·	·	•	·	University of Bonn, Georgia Institute of Technology, University of Coimbra, Vienna University of Technology	Joao Carreira, Adrian Ion, Fuxin Li, Cristian Sminchisescu	Ion et al (2011a,b)
BONN_LINEAR	·	·	•	·	University of Bonn, University of Coimbra	Joao Carreira, Rui Caseiro, Jorge Batista, Cristian Sminchisescu	Carreira et al (2012)
CVC	•	·	·	·	Computer Vision Barcelona	Fahad Khan, Camp Davesa, Joost van de Weijer, Rao Muhammad Anwer, Albert Gordo, Pep Gonfaus, Ramon Baldrich, Antonio Lopez	Khan et al (2012a)
CVC_CLS	·	•	·	·	Computer Vision Barcelona	Albert Gordo, Camp Davesa, Fahad Khan, Pep Gonfaus, Joost van de Weijer, Rao Muhammad Anwer, Ramon Baldrich, Jordi Gonzalez, Ernest Valveny	Khan et al (2012a,b)
CVC_SP	•	·	·	·	Computer Vision Barcelona, University of Amsterdam, University of Trento	Fahad Khan, Jan van Gemert, Camp Davesa, Jasper Uijlings , Albert Gordo, Sezer Karaoglu, Koen van de Sande, Pep Gonfaus, Rao Muhammad Anwer, Joost van de Weijer, Cees Snoek, Ramon Baldrich, Nicu Sebe, Theo Gevers	Khan et al (2012a,b); Karaoglu et al (2012); van Gemert (2011)
HU	·	·	·	•	Hacettepe University, Bilkent University	Cagdas Bas, Fadime Sener, Nazli Ikizler-Cinbis	Sener et al (2012)
IMPERIAL	•	·	·	·	Imperial College London	Ioannis Alexiou, Anil A. Bharath	Alexiou and Bharath (2012)
ITI, ITI_ENTROPY, ITI_FUSED	•	·	·	·	ITI-CERTH, University of Surrey, Queen Mary University of London	Elisavet Chatzilari, Spiros Nikolopoulos, Yiannis Kompatsiaris, Joseph Kittler	-
MISSOURI	·	•	·	·	University of Missouri Columbia	Guang Chen, Miao Sun, Xutao Lv, Yan Li, Tony Han	-
NEC	·	•	·	·	NEC Laboratories America, Stanford University	Olga Russakovsky, Xiaoyu Wang, Shenghuo Zhu, Li Fei-Fei, Yuanqing Lin	Russakovsky et al (2012)
NUS_SCM	•	·	·	·	National University of Singapore, Panasonic Singapore Laboratories, Sun Yat-sen University	Dong Jian, Chen Qiang, Song Zheng, Pan Yan, Xia Wei, Yan Shuicheng, Hua Yang, Huang Zhongyang, Shen Shengmei	Song et al (2011); Chen et al (2012)
NUS_SP	·	·	•	·	National University of Singapore, Panasonic Singapore Laboratories	Wei Xia, Csaba Domokos, Jian Dong, Shuicheng Yan, Loong Fah Cheong, Zhongyang Huang, Shengmei Shen	Xia et al (2012)
OLB_R5	·	•	·	·	Orange Labs Beijing, France Telecom	Zhao Feng	-
OXFORD	·	•	·	·	University of Oxford	Ross Girshick, Andrea Vedaldi, Karen Simonyan	-
OXFORD_ACT	·	·	·	•	University of Oxford	Minh Hoai, Lubor Ladicky, Andrew Zisserman	Hoai et al (2012)
STANFORD	·	·	·	•	Stanford University, MIT	Aditya Khosla, Rui Zhang, Bangpeng Yao, and Li Fei-Fei	Khosla et al (2011)
SYSU_DYNAMIC	·	•	•	·	Sun Yat-Sen University	Xiaolong Wang, Liang Lin, Lichao Huang, Xinhui Zhang, Zechao Yang	Wang et al (2013)
SZU	·	·	·	•	Shenzhen University	Shiqi Yu, Shengyin Wu, Wensheng Chen	-
UP	•	·	·	·	University of Padova	Loris Nanni	Nanni and Lumini (2013)
UVA_HYBRID	·	•	·	·	University of Amsterdam	Koen van de Sande, Jasper Uijlings, Cees Snoek, Arnold Smeulders	van de Sande et al (2011); Uijlings et al (2013)
UVA_MERGED	·	•	·	·	University of Amsterdam	Sezer Karaoglu, Fahad Khan, Koen van de Sande, Jan van Gemert, Rao Muhammad Anwer, Jasper Uijlings, Camp Davesa, Joost van de Weijer, Theo Gevers, Cees Snoek	Khan et al (2012a); Uijlings et al (2013)
UVA_NBNN	·	·	•	·	University of Amsterdam	Carsten van Weelden, Maarten van der Velden, Jan van Gemert	-

Table 4: Participation in the 2012 challenge. Each method is assigned an abbreviation used in the text, and identified as a classification (Cls), detection (Det), segmentation (Seg), or action classification (Act) method. The contributors to each method are listed with references to publications describing the method, where available.