Optimizing Intersection-Over-Union in Deep Neural Networks for Image Segmentation

Md Atiqur Rahman(✉) and Yang Wang

Department of Computer Science, University of Manitoba, Winnipeg, Canada
{atique,ywang}@cs.umanitoba.ca

Abstract. We consider the problem of learning deep neural networks (DNNs) for object category segmentation, where the goal is to label each pixel in an image as being part of a given object (foreground) or not (background). Deep neural networks are usually trained with simple loss functions (e.g., softmax loss). These loss functions are appropriate for standard classification problems where the performance is measured by the overall classification accuracy. For object category segmentation, the two classes (foreground and background) are very imbalanced. The intersection-over-union (IoU) is usually used to measure the performance of any object category segmentation method. In this paper, we propose an approach for directly optimizing this IoU measure in deep neural networks. Our experimental results on two object category segmentation datasets demonstrate that our approach outperforms DNNs trained with standard softmax loss.

1 Introduction

We consider the problem of object category segmentation using deep neural networks. The goal of object category segmentation is to label the pixels of a given image as being part of a given object (foreground) or not (background). In such a problem setting, the two classes (foreground and background) are often very imbalanced, as the majority of the pixels in an image usually belong to the background. Learning algorithms that are designed to optimize for overall accuracy may not be suitable in this problem setting, as they might end up predicting every pixel to be background in the worst case. For example, if 90% of the pixels belong to the background, a naive algorithm can achieve 90% overall classification accuracy simply by labeling every pixel as the background.

The standard performance measure that is commonly used for the object category segmentation problem is called intersection-over-union (IoU). Given an image, the IoU measure gives the similarity between the predicted region and the ground-truth region for an object present in the image, and is defined as the size of the intersection divided by the union of the two regions. The IoU measure can take into account of the class imbalance issue usually present in such a problem setting. For example, if a naive algorithm predicts every pixel of an image to be background, the IoU measure can effectively penalize for that, as

G. Bebis et al. (Eds.): ISVC 2016, Part I, LNCS 10072, pp. 234–244, 2016.
DOI: 10.1007/978-3-319-50835-1_22

the intersection between the predicted and ground-truth regions would be zero, thereby producing an IoU count of zero.

Most deep learning based methods address the image segmentation problem using simple loss functions, such as, softmax loss which actually optimizes for overall accuracy. Therefore, they are subject to the problem mentioned above. We argue that directly optimizing the IoU loss is superior to the methods optimizing for simple loss functions. In this paper, we address the object category segmentation problem by directly optimizing the IoU measure in a deep learning framework. To this end, we incorporate the IoU loss in the learning objective of the deep network.

2 Related Work

Our proposed approach for object category segmentation overlaps with two directions of research – one involves direct optimization of application specific performance measures (in this case, IoU measure), and the other line of research focuses on image semantic segmentation using DNNs. Below we briefly present some of the works most related to our proposed approach.

Direct loss optimization: Application specific performance measure optimization has been studied so far mainly for learning linear models. For example, Joachims [1] proposed a multi-variate SVM formulation for optimizing a range of nonlinear performance measures including F_1-Score and ROC-area. Other SVM-based methods include [2,3] that proposed approaches to directly optimize the mean average precision (mAP) measure. Very recently, there have been a few deep models proposed to directly optimize some application specific measures (e.g., [4] and [5] for mAP, [6] for ROC-area).

Regarding direct optimization of the IoU measure, the first work to address this problem was proposed by Blaschko et al. [7] with an application to object detection and localization. Based on a structured output regression model, they used joint-kernel map and proposed a constraint generation technique to efficiently solve the optimization problem of structural SVM framework. Ranjbar et al. [8] used structured Markov Random Field (MRF) model in an attempt to directly optimize the IoU measure. Tarlow and Zemel [9] addressed the problem using highly efficient special-purpose message passing algorithms. Based on Bayesian decision theory, Nowozin [10] used a Conditional Random Field (CRF) model and proposed a greedy heuristic to maximize the value of Expected-Intersection-over-Expected-Union (EIoEU). Premachandran et al. [11], on the other hand, optimizes exact Expected-IoU. A recent work by Ahmed et al. [12] draws the best from both of these approaches. Based on the fact that the EIoEU is exact for a delta distribution, they take the idea of approximating EIoU from [11] by taking the average of EIoEU as computed in [10].

Semantic segmentation using DNNs: The semantic segmentation problem is similar to the object category segmentation problem, but requires labeling each pixel of an image as being part of one of several semantic object categories (e.g., cow, bus etc.), instead of just foreground or background. Recently, several approaches have been proposed for semantic segmentation that take advantage of high-level representation of images obtained from DNNs. For example, Hariharan et al. [13] used a CNN architecture that can simultaneously perform object detection and semantic segmentation. Long et al. [14] proposed a novel DNN architecture that turns a classification CNN (e.g., AlexNet [15]) into fully convolutional net by replacing the fully connected layers of the CNN with convolution layers. Our proposed approach is based on this approach, but optimizes for application specific performance measure of IoU, instead of overall accuracy. Very recently, some deep encoder-decoder based models (e.g., SegNet [16], TransferNet [17]) have been proposed that mark the state-of-the-art for image semantic segmentation.

3 Proposed Approach

In this paper, we consider the problem of object category segmentation. Given an object category, the goal is to label the pixels of an image as being part of objects belonging to the category (foreground) or not (background). To this end, we convert a classification CNN into a fully-convolutional net as proposed in [14], and then train the deep network end-to-end and pixel-to-pixel with an objective to directly optimize the intersection-over-union (IoU) performance measure. The architecture of the deep network as well as details of the IoU loss function are discussed in the following subsections.

3.1 Network Architecture and Workflow

Following the recent novel work for semantic segmentation by Long et al. [14], we start with a classification CNN called AlexNet [15], and replace the last two fully connected layers (fc_6 and fc_7) of AlexNet with 1×1 convolution layers (C_6 and C_7, respectively in Fig. 1) to convert the CNN into a fully-convolutional network (FCN). We then add a scoring layer (C_8) which is also a 1×1 convolution layer. The sub-sampled output out of the scoring layer is then passed to a deconvolution layer (DC) that performs bilinear interpolation at a stride of 32 and produces an output equal to the size of the original input to the network. Up to this point, everything remains the same as the original 32-stride version of the FCN called "FCN-32s" [14].

Once an output equal to the size of the input is produced, we pass it through a sigmoid layer to convert the scores into class probabilities representing the likelihood of the pixels being part of the object. From this point forward, the proposed approach differs from [14] which computes softmax loss on each pixel score and trains the whole network based on this loss. We argue that this is not the right approach for a task like object category segmentation, where the ratio of object to background pixels is very small. The softmax loss is closely tied to the overall classification accuracy. If the number of examples in each class are balanced, minimizing the softmax loss will give high overall classification accuracy.

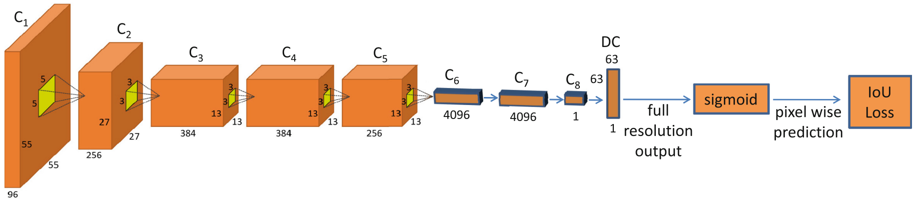

Fig. 1. Architecture of the proposed FCN. The first eight convolution layers ($C_1 - C_8$) and the deconvolution layer (DC) remain the same as the original FCN-32s proposed in [14]. For each layer, the number right at the bottom represents the depth, while the other two numbers represent the height and width of the layer output. The yellow boxes inside layers C_1 to C_5 represent the filters, while the numbers around them represent filter dimensions. The IoU loss layer at the end computes IoU loss on the full-resolution output representing object class probabilities of the pixels

For object category segmentation, the two classes are often very imbalanced, and therefore, the overall accuracy is not often a good performance measurement. For example, if 90% of the pixels belong to the background, a naive algorithm can achieve 90% overall classification accuracy simply by labeling every pixel as the background. In object category segmentation, the IoU score is often used as the standard performance measure, which takes into account of the class imbalance issue. Following this observation, instead of computing softmax loss, we pass the pixel probabilities out of the sigmoid layer to a loss layer that computes the IoU loss from the pixel probabilities and then train the whole FCN based on this loss. Figure 1 demonstrates the pipeline of the proposed approach.

3.2 Approximation to IoU and IoU Loss

The IoU score is a standard performance measure for the object category segmentation problem. Given a set of images, the IoU measure gives the similarity between the predicted region and the ground-truth region for an object present in the set of images and is defined by following equation.

$$IoU = \frac{TP}{FP + TP + FN} \tag{1}$$

where, TP, FP, and FN denote the true positive, false positive and false negative counts, respectively.

From Eq. 1, we see that IoU score is a count based measure, whereas, the outputs of the proposed FCN are probability values representing likelihood of the pixels being part of the object. Therefore, we cannot accurately measure the IoU score directly from the output of the network. We propose to approximate the IoU score using the probability values. More formally, let $V = \{1, 2, \ldots, N\}$ be the set of all pixels of all the images in the training set, X be the output of the network (out of the sigmoid layer) representing pixel probabilities over the set V, and $Y \in \{0, 1\}^V$ be the ground-truth assignment for the set V, where 0 represents background pixel and 1 represents object pixel. Then, the IoU count can be defined as:

$$IoU = \frac{I(X)}{U(X)} \tag{2}$$

where, $I(X)$ and $U(X)$ can be approximated as follows:

$$I(X) = \sum_{v \in V} X_v * Y_v \tag{3}$$

$$U(X) = \sum_{v \in V} (X_v + Y_v - X_v * Y_v) \tag{4}$$

Therefore, the IoU loss L_{IoU} can be defined as follows:

$$L_{IoU} = 1 - IoU = 1 - \frac{I(X)}{U(X)} \tag{5}$$

We then incorporate the IoU loss L_{IoU} into the objective function of the proposed FCN, which takes the following form:

$$\arg\min_{w} L_{IoU} = 1 - IoU \tag{6}$$

where, w is the set of parameters of the deep network.

In order to obtain the optimal set of parameters w, Eq. 6 is solved using stochastic gradient descent. The gradient of the objective function with respect to the output of the network can then be written as follows:

$$\begin{aligned} \frac{\partial L_{IoU}}{\partial X_v} &= -\frac{\partial}{\partial X_v}\left[\frac{I(X)}{U(X)}\right] \\ &= \frac{-U(X) * \frac{\partial I(X)}{\partial X_v} + I(X) * \frac{\partial U(X)}{\partial X_v}}{U(X)^2} \\ &= \frac{-U(X) * Y_v + I(X) * (1 - Y_v)}{U(X)^2} \end{aligned} \tag{7}$$

which can be further simplified as follows:

$$\frac{\partial L_{IoU}}{\partial X_v} = \begin{cases} -\frac{1}{U(X)} & \text{if } Y_v = 1 \\ \frac{I(X)}{U(X)^2} & \text{otherwise} \end{cases} \tag{8}$$

Once the gradients of the objective function with respect to the network output are computed, we can simply backpropagate the gradients using the chain rule of derivative in order to compute the derivatives of the objective function with respect to the network parameters w.

4 Experiments

We conducted training on individual object categories and learned segmentation models for each object category separately. In other words, when we train segmentation model for a particular object category, say dog, we assume pixels of all other categories as part of the background. During inference, we pass all test images through the learned models, one for each object category, and then segment the specific objects individually from the test images. In the following subsections, we describe the datasets and training setups used in the experiments, and also report and compare the experimental results of our approach and the baseline methods.

4.1 Experimental Setup

Datasets. To evaluate the proposed approach, we conducted experiments on three different datasets – PASCAL VOC 2010 [18] and PASCAL VOC 2011 [19] segmentation datasets, as well as the Cambridge-driving Labeled Video Database (CamVid) [20]. The PASCAL VOC is a highly challenging dataset containing images from 20 different object categories with the objects having severe variability in size, pose, illumination and occlusion. It also provides pixel-level annotations for the images. VOC 2010 includes 964 training and 964 validation images, while VOC 2011 includes 1,112 training and 1,111 validation images. We trained using 80% of the images in the training set, while the remaining 20% images were used for validation. We evaluated the different approaches on the dataset provided validation set.

CamVid is a road scene understanding dataset including over 10 min of high quality video footage and provides 701 high resolution images from 11 different object categories. It also provides pixel-level semantic segmentations for the images. Among the 701 images, 367 images were used for training, 233 for testing and the remaining 101 for validation.

Baselines. As a primary baseline, we compare our proposed approach to a method proposed in [14] that uses a fully convolutional net to address the semantic segmentation problem by optimizing for overall accuracy using softmax loss. We also perform comparison with [8] that tries to directly optimize the IoU measure based on an MRF model. For the rest of the paper, we refer to the proposed approach as FCN_{IoU}, the deep model optimizing for overall accuracy as FCN_{acc}, and the MRF-based model as MRF_{IoU}.

Implementation Details. We conducted training of the deep nets using stochastic gradient descent in mini batches. While preparing the mini batches, we made it sure that each batch contains at least one positive example (i.e., an image containing the object for which model is being trained). Training was initialized with pre-trained weights from AlexNet [15]. For PASCAL VOC, we

Table 1. Intersection-over-union (%) performance comparison for 6 different object categories on PASCAL VOC 2010 validation set

Method	Aeroplane	Bus	Car	Horse	Person	TV/Monitor
MRF_{IoU}	<20	<30	<30	<10	<25	<15
FCN_{acc}	71.07	72.85	71.67	60.46	**75.42**	64.03
FCN_{IoU}	**75.27**	**74.47**	**72.83**	**61.18**	72.65	**67.37**

resized the training images to 375×500, while testing was done on the original images without resizing. On the other hand, for CamVid, all images were resized to 360×480. We used a fixed learning rate of 10^{-4}, momentum of 0.99 and weight decay of 0.0005. We continued training until convergence and chose the model with the best IoU measure on the validation set. All the deep nets were implemented using a popular deep learning tool called MatConvNet [21].

4.2 Results on PASCAL VOC

For the PASCAL VOC 2010 dataset [18], Table 1 shows the results of the proposed approach and the baselines for 6 different object categories. Our proposed approach outperforms MRF_{IoU} by huge margin on all 6 categories. This performance boost is simply due to the use of deep features learned automatically by the proposed approach FCN_{IoU}, whereas, MRF_{IoU}, being a shallow model, lacks this ability. Please note that we could not report the exact IoU values for MRF_{IoU}, since [8] reports the results using a bar chart without using the exact numbers.

While comparing the proposed approach FCN_{IoU} to the primary baseline FCN_{acc}, we see that FCN_{IoU} outperforms FCN_{acc} in almost all categories. It is particularly noteworthy that the performance improvements are more significant

Table 2. Background to object pixel ratio in PASCAL VOC 2010 and VOC 2011 datasets

Dataset	Aeroplane	Bicycle	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	Dining Table	Dog	Horse	Motorbike	Person	Potted Plant	Sheep	Sofa	Train	TV/Monitor
VOC 2010	152	319	107	142	150	64	66	40	97	152	82	75	117	91	25	182	111	99	85	104
VOC 2011	153	341	100	158	152	60	68	41	94	160	82	71	127	86	23	176	115	88	76	113

Table 3. Intersection-over-union (%) performance comparison on PASCAL VOC 2011 validation set

Method	Aeroplane	Bicycle	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	Dining Table	Dog	Horse	Motorbike	Person	Potted Plant	Sheep	Sofa	Train	TV/Monitor	Mean
FCN$_{acc}$	72.18	60.57	66.47	64.68	**65.03**	73.96	71.82	71.44	55.55	**64.22**	62.74	**67.03**	60.74	70.23	**76.78**	61.62	67.59	58.05	72.80	65.05	63.18
FCN$_{IoU}$	**75.07**	**62.00**	**67.45**	**67.64**	65.00	**75.37**	**72.87**	**71.94**	**56.01**	64.13	**63.91**	65.71	**60.92**	**70.90**	73.61	**63.78**	**68.83**	**58.56**	**72.66**	**66.81**	**63.82**

for object categories (e.g., "Aeroplane", "TV/Monitor" etc.) where the ratio of the background to object pixels is very large as shown in Table 2.

Table 3 shows IoU comparison of the proposed approach FCN$_{IoU}$ and the primary baseline FCN$_{acc}$ on the PASCAL VOC 2011 [19] validation set. The other method does not report any result on this dataset. We see that FCN$_{IoU}$ performs better than FCN$_{acc}$ in most cases and also on average. Specifically, the performance improvements are more significant for object categories with a larger ratio of background to object pixels.

We also show some qualitative results of the proposed approach FCN$_{IoU}$ and the primary baseline FCN$_{acc}$ in Fig. 2. Since softmax loss is tied to overall classification accuracy, the FCN$_{acc}$ model tends to misclassify object pixels as background (i.e., false negative), as there exist more background pixels. In contrast, FCN$_{IoU}$ tends to recover some of the false negative errors made by FCN$_{acc}$, as it directly optimizes for IoU score. This observation is supported by the example segmentations as shown in the figure.

4.3 Results on CamVid

For the CamVid dataset [20], among the 11 object categories, we report results on 5, namely, "Road", "Building", "Column-Pole", "Sign-Symbol", and "Fence". We chose the "Road" and "Building" categories for their relatively lower ratio of

Table 4. Intersection-over-union (%) performance comparison for 5 different object categories on CamVid val (validation) and test set

Method	Road		Building		Column-Pole		Sign-Symbol		Fence	
	val	test	val	test	val	test	val	test	val	test
FCN$_{acc}$	95.53	90.38	87.03	76.21	50.46	50.91	64.94	56.27	75.97	61.75
FCN$_{IoU}$	**95.58**	**90.69**	**88.30**	**76.72**	**53.48**	**52.79**	**67.78**	**57.78**	**80.68**	**62.23**

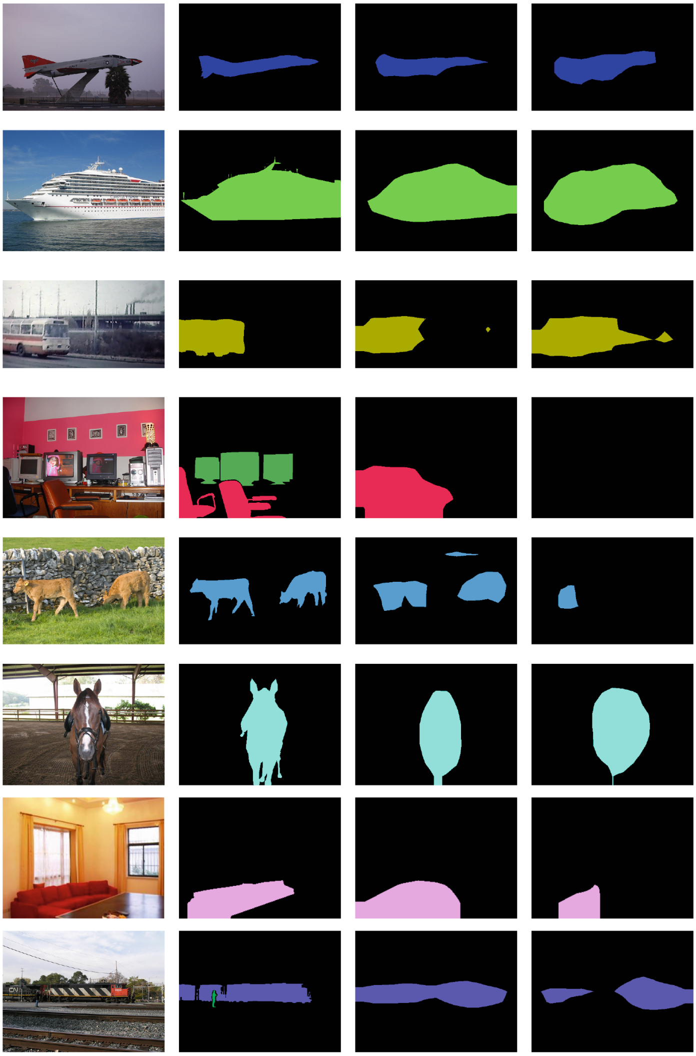

Fig. 2. Sample segmentations results on the PASCAL VOC 2011 validation set. Columns (left to right): original images, ground-truth segmentations, segmentations produced by FCN_{IoU}, and segmentations produced by FCN_{acc}

background to object pixels compared to other object categories in the dataset, while the other 3 categories were chosen for the opposite reason. We do this to investigate how the proposed approach performs as the ratio of background to object pixels varies. Table 4 reports IoU scores on the 5 object categories. The results show that FCN_{IoU} outperforms FCN_{acc} in all 5 categories. More importantly, performance improvements are more significant for smaller object categories (e.g., “Column-Pole”, “Sing-Symbol”, and “Fence”) having higher class imbalance than those that are relatively balanced (e.g., "Road" and "Building").

We also show some qualitative results on the CamVid test set in Fig. 3. The results show that FCN$_{IoU}$ performs better than FCN$_{acc}$, specially for smaller object categories (e.g., Column-Pole) where there exists large imbalance in the number of object and background pixels.

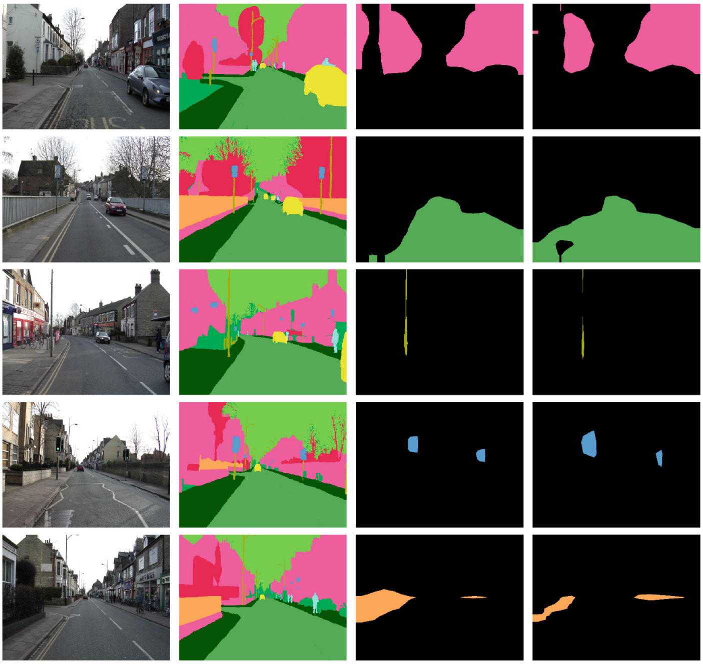

Fig. 3. Sample segmentation results on the CamVid test set. Rows (top to bottom): segmentations for "Building", "Road", "Column-Pole", "Sign-Symbol", and "Fence". Columns (left to right): original images, ground-truth segmentations, segmentations produced by FCN$_{IoU}$, and segmentations produced by FCN$_{acc}$

5 Conclusion

We have presented a new approach for direct optimization of the IoU measure in deep neural networks. We have applied our approach to the problem of object category segmentation. Our experimental results demonstrate that optimizing the IoU loss leads to better performance compared with traditional softmax loss commonly used for learning DNNs. In this paper, we have focused on binary segmentation problems. As for future work, we would like to extend our approach to directly handle multi-class image segmentation.

References

1. Joachims, T.: A support vector method for multivariate performance measures. In: Proceedings of ICML (2005)
2. Yue, Y., Finley, T., Radlinski, F., Joachims, T.: A support vector method for optimizing average precision. In: Proceedings of SIGIR (2007)
3. Behl, A., Jawahar, C.V., Kumar, M.P.: Optimizing average precision using weakly supervised data. In: CVPR (2014)
4. Song, Y., Schwing, A.G., Zemel, R.S., Urtasun, R.: Direct loss minimization for training deep neural nets. CoRR abs/1511.06411 (2015)
5. Henderson, P., Ferrari, V.: End-to-end training of object class detectors for mean average precision. CoRR abs/1607.03476 (2016)
6. Rahman, M.A., Wang, Y.: Learning neural networks with ranking-based losses for action retrieval. In: Proceedings of CRV (2016)
7. Blaschko, M.B., Lampert, C.H.: Learning to localize objects with structured output regression. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008. LNCS, vol. 5302, pp. 2–15. Springer, Heidelberg (2008). doi:10.1007/978-3-540-88682-2_2
8. Ranjbar, M., Lan, T., Wang, Y., Robinovitch, S., Mori, G.: Optimizing non-decomposable loss functions in structured prediction. IEEE Trans. Pattern Anal. Mach. Intell. **35**(4), 911–924 (2013)
9. Tarlow, D., Zemel, R.S.: Structured output learning with high order loss functions. In: Proceedings of AISTATS (2012)
10. Nowozin, S.: Optimal decisions from probabilistic models: the intersection-over-union case. In: CVPR (2014)
11. Premachandran, V., Tarlow, D., Batra, D.: Empirical minimum bayes risk prediction: how to extract an extra few % performance from vision models with just three more parameters. In: CVPR (2014)
12. Ahmed, F., Tarlow, D., Batra, D.: Optimizing expected intersection-over-union with candidate-constrained CRFS. In: ICCV (2015)
13. Hariharan, B., Arbeláez, P., Girshick, R., Malik, J.: Simultaneous detection and segmentation. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8695, pp. 297–312. Springer, Heidelberg (2014). doi:10.1007/978-3-319-10584-0_20
14. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR (2015)
15. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS (2012)
16. Badrinarayanan, V., Kendall, A., Cipolla, R.: SegNet: a deep convolutional encoder-decoder architecture for image segmentation. arXiv preprint (2015). arXiv:1511.00561
17. Hong, S., Oh, J., Lee, H., Han, B.: Learning transferrable knowledge for semantic segmentation with deep convolutional neural network. CoRR abs/1512.07928 (2015)
18. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The pascal VOC2010 challenge results (2010)
19. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The pascal VOC2011 challenge results (2011)
20. Brostow, G.J., Fauqueur, J., Cipolla, R.: Semantic object classes in video: a high-definition ground truth database. Pattern Recogn. Lett. **20**(2), 88–97 (2009)
21. Vedaldi, A., Lenc, K.: Matconvnet - convolutional neural networks for matlab. In: Proceedings of ACM International Conference on Multimedia (2015)