Learning Semantic Segmentation Score in Weakly Supervised Convolutional Neural Network

Fariz Ikhwantri, Novian Habibie, Arie Rachmad Syulistyo, Aprinaldi, and Wisnu Jatmiko
Faculty of Computer Science
Universitas Indonesia
Depok, Indonesia
email: {fariz.ikhwantri,novian.habibie,arie.rachmad}@ui.ac.id; wisnuj@cs.ui.ac.id

***Abstract*-Semantic segmentation is an image labeling process for each pixels according to defined objects class and its presence in an image. Labeling process consists of recognizing, detecting location and labeling pixels that defines the object in the image. Annotation result of semantic segmentation needs ground truth to verify accuracy of score prediction. Therefore, this research propose a model to predict score of annotation accuracy. By casting the problem into constraining object boundary recognition, we described the annotation using foreground mask. To extract the feature, we used convolution neural network. We only used CNN trained on a image level annotation. In order to be able to infer the pixel instance, we adapt CNN architecture into weakly supervised learning. Experiments were conducted by finetuning Convolution Neural Network for object recognition using weakly supervised architecture for multilabel classification. In this paper we proposed to score semantic segmentation based on bag level information without the availability of pixel level annotation.**

Keywords-Semantic Segmentation, Convolutional Neural Networks, Jaccard Index, Regression, Weakly Supervised Learning, Multiple Instance Learning

I. INTRODUCTION

Semantic Segmentation is a major challenging task in scene understanding which aims to give label for each pixel in 2D image according to object presence, location, and shape. These segmentation result are applicable to many practical application such as robotic vision, embryo segmentation, recognition and tracking. Semantic segmentation process, which commonly consist of recognizing object presence, and object localization from homogeneous pixel (superpixels), use segmentation algorithm. Another labeling technique can be done by learning parameter in the process using structured prediction. This method needs groundtruth to measure algorithm performance.

Groundtruth is created by human expert (annotator) as benchmark of machine performance. Creating groundtruth for each pixel in image needs long time and hard effort. Hoewever, groundtruth significantly give contribution on the development of studying semantic segmentation problem [1].

Most of best performed and well-known approach in predicting label are trained by learning statistic priors of pixels or superpixels [2],[3] from groundtruth using Conditional Random Field model [4]. However in realistic scenario, there are no accurate groundtruth that describe accuracy of the result. Another strategy learn a pseudolikehood objective function to predict result score is done by Li et al [5]. Their approach learns a goodness of bottom-up figure ground hypothesis generated from graph segmentation. They rank segmented object and aggregate multiple segment to obtain highest scored object segment. These method use handcrafted feature extracted from local image features (SIFT, HOG). Li et al [5] use performance metric of semantic segmentation as target label. Performance metric in semantic segmentation are done by comparing intersection over union of result and groundtruth as document similarity (jaccard index). This objective function is non linear and intractable number of combination. Optimization of multilabel structured learning can be improved by increasing f-measure of classifier mentioned in [6].

Over past several years, Convolutional Neural Network (CNN) is state-of-the-art in scene understanding task. For example in image recognition [7]-[10], object detection [8],[11], and also in semantic segmentation. The motivation of using CNN than handcrafted features and conventional learning algorithm because of end-to-end learning capability and feature learning in hierarchical layer representation.

In classification problem such as object recognition, CNN achieved the best performance in very large dataset like ImageNet [12] dataset. However, CNN requires large set of training data sample in supervised learning multilabel and/or structured prediction such as object detection and semantic segmentation. There is also another obstacle of getting larger dataset such as obtaining groundtruth per pixel label that require hard effort. In order to overcome the groundtruth limitation, several works already propose to use weakly supervised learning. In this methodology, label prediction required only partial knowledge. These partial knowledge in semantic segmentation represented by its weak annotation such as bounding box even only image tag without groundtruth.

Our contribution in this paper is learning jaccard score of semantic segmentation per pixel-class accuracy from annotated image and convolutional neural network. Although there already previous predicting score of semantic segmentation such as in [13] based on knowledge based rely on prior statistics to its handcrafted features and scene oracles. We formulate CNN feature extraction, which only trained on weak annotation such as image presents tags. To

978-1-4673-9754-4/15/$31.00 ©2015 IEEE

predict annotation score we formulate it as regression. We use coefficient from joint probability between predicted label and extracted feature label. Our CNN architecture are based on [11] as multiple instance learning in semantic segmentation.

II. Related Works

A. Semantic Segmentation

Semantic segmentation purposes to represents semantic class based on features in observed pixels. Semantic label is attached into region or pixel of object in image, which generated from learning and prediction of determined set of label. In general semantic segmentation will predict a label from set of label L = {l1, l2 , …ln} from every set of pixels or superpixels. S = {s1,s2,…,sn}. Overall problem can be formulated as y : S → L. These function give an output mn probability of prediction y = $\{y_1^n, y_2^n, \ldots, y_m^n\}$. Structured prediction labeling takes equation of interaction among variables from unary potential, pairwise potential and higher order potential. This formulation approaches usually use conditional random fields model, that use feature extraction from image. One of early works that uses conditional random field is joint feature of texture, color and shape as unary potential [14].

B. Multiple Instance Learning (MIL)

Multiple Instance Learning (MIL) is variation of supervised learning under constrain of weak label. These weak label is set of label from multiset (bags) $X_i \subset \boldsymbol{R^d}, i = 1,2,\ldots,N$. Every set of Xi are consist of instance level $X = x^1, x^{12}, \ldots, x^m$. In training dataset, there is only known bag level label Y_i. These label can be composed as binary {-1,1}, multiclass and multilabel Yi $\subset$ {1,2,…,c}. According to Babenko [15], instance level prediction for class Yi = 1 selected if there is exists an instance of i in bag of instance. Labeling strategy based on notation formulated on equation 1:

$$y_i = \begin{cases} 1 & \text{if } \exists j \in x_i \text{ s.t } y_{ij} = 1 \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

Prediction of instance in bag takes label from maximum value of every class label, it is formulated on equation 2 :

$$y_i = \max y_i^j \quad (2)$$

This method are based on key and door theory [15],[16]. MIL theorem succeed in modelling for drug discovery problem [17]. In computer vision, early idea of predicting bag level of data based on instance is from Keeler [18] for integrated segmentation and recognition of handwritten digit recognition.

C. Convolutional Neural Network

Convolutional Neural Network (CNN) is a model of artificial neural network inspired by research of visual neuron called visual cortex by Hubel and Wiesel [19]. CNN is an visual cortex representation which performs filtering process into image $I \in R^n$. Filtering process performs by shifting kernel into the image which assembled in deep hierarchical layers. Shifting process transforms input image using filter kernel by feed forward process:

$$h_l^k = pool(\phi(W_l^k h_{l-1} b_l^k)) \quad (3)$$

where $h_l \in R^n$ are feature maps from layer l-1 if h0 is input image, k is filter index for channel dimension, Φ is activation function e.g. sigmoid function, tanh function, and so on Wl is a weight value of learning rate of CNN.

In general, learning in CNN is a process of mapping input into prediction function $f : I \rightarrow \acute{L} \approx L$ (where $\acute{L} \in R^k$ is a prediction label into label target L = {l1, l12, …l1k}, L $\in$ RK with counting its error. Main goal of learning process is to reduce error in testing from training data $D = \{d_1, d_2, \ldots, d_n\}$ from observed label, which represents on equation:

$$E = \frac{1}{2} \sum_{j=1}^{n} \sum_{i=1}^{k} (f_i(d_i) - C_j(l_i))^2 \quad (4)$$

1) Transfer Learning Adaptation Layer: Transfer learning can be done in either unsupervised manner where no label available or similar data with different label and vice versa in supervised model [20]. In image domain dataset, transfer learning gain popular success in deep learning as it could increase performance.

Many previous work found that pre-trained and finetuning of deep neural network from larger dataset into another dataset improve large margin of performance [9][10][21]. Model trained on a large dataset tends to have good generalization performance. This adjusted model can be used for another difference dataset by transferring learned feature. Work on investigating transfer learning in deep neural network [22], revealed trained CNN in first and second layer of convolution from input tend to extract low level feature such as color and shape.

Based on that study, there are several practices of transfer learning in deep neural network. Most common practice is removing last layer (classification) that proceed by finetuning new layer based on class label target. This last layer is classifier for a new label. In recent years, an experiment find out that mid level layer of pre-trained CNN features are more common to train in object recognition task [21], such as action and approximate object location. Convolution layer is located at mid level layer before fully connected networks. In this paper, we add two convolution and global max pooling layer into adaptation layer in addition to initial five layer of convolution and pooling layer. Further details of our architecture illustrated in fig.1. In the fig.1 CNN architecture consist of 7 fully convolution layers ordered as 96 x 7 x 7, 256 x 5 x 5, 512 x 3 x 3, 512 x 3 x 3, 512 x 3 x 3, 4096 x 1 x 1, where is a number of feature maps x m kernel width x n kernel height.

2) Global Max Pooling Layer: Global max pooling aims to search all possible location of object in image. This pooling method also increases size invariant in extracting feature globally. Previous work uses global max pooling layer [11] applied for object recognition in extend to detection task. This layer in CNN aggregate $ n x m $ each feature maps into a 1 x 1 score for each of them respectively. These aggregation layer can also be seen as integrating convolutional neural network by doing inference from bag of feature maps as multiple instance learning .

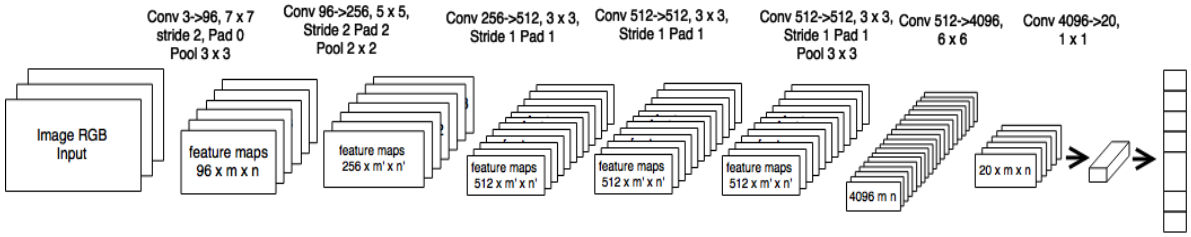

Figure 1. CNN Architecture Details

In this section, we describe our strategy to extract feature from annotated image. Our Convolutional Neural Network architecture is based on [10]. We choose VGG-CNN-S architecture over others because its configuration of small convolution of kernel size. These architecture generally consist of five convolution and two fully connected layer. In this paper, we aims to extracting bag of feature maps to learn object presence and segmentation in weakly supervised settings. Thus in our experiment we preserved five convolution layer and adding convolution adaptation layer similar in [11].

Weakly supervised learning for instance class prediction are based on maximum score map in bag of feature maps. Predicting a label for an individual from set of label and bag can be a hard decision. This problem also becomes more problematic if individual data is noisy. In multilabel classification setting, amount of label increase possibility of configuration. In order to maximize F measure of label probability, we trains feature extractor use one vs all classification approach. This learning strategy assumes independence relationship between label. Independences between label can be train in CNN using Multiclass Negative Log Likelihood as a loss function denoted in equation 5.

$$loss(f_k(x), c_k) = \sum_{i=1}^{k} \log(1 + \exp^{-f_k(x)}) \tag{5}$$

In this problem we also investigate different loss function. We observed that using sigmoid function 6 on top of global maximum pooling can increase accuracy of classification. Hence, we also use another loss function for training Weakly CNN.

$$f_k(x) = \frac{1}{1 + \exp(-x)} \tag{6}$$

III. Learning Semantic Segmentation Score as Weakly Supervised Learning

In this section we describe definition of jaccard index, and our methodology for learning semantic segmentation score. Semantic segmentation score obtains by count similarity between annotation and groundtruth. These performance metric defined as intersection over union between amount of true positive pixel (*Tp* = 1) pixel per class and sum of true positive, false positive and false negative (see equation 7) :

$$J(A, B) = \frac{\sum_{i=1}^{m}(A_i = 1 \cap B_i = 1)}{\sum_{i=1}^{m}(A_i = 1 \cup B_i = 1)} \tag{7}$$

In semantic segmentation for weakly supervised learning, that only known information from training images, are weak level annotation such as bounding boxes of objects presence. There are several advantages of learning with only weak annotation such as abundant training data available. For example in PASCAL [23] and ImageNet [12], object recognition consecutively are ~10K and ~400K training data, whereas there is only ~4K amount of fully annotated object segmentation in PASCAL dataset. Utilizing large amount of data also increases performance of prediction as demonstrated in ImageNet competition, most of them used CNN.

Semantic segmentation scoring purposes to predict an unknown score of annotation without using groundtruth. Class Accuracy of semantic segmentation for error profile represent as average of each class jaccard score that can be seen in equation 7. Multilabel problem in semantic segmentation is a function $f : XxX \rightarrow Y$To map function into Y which is a multilabel, we can optimize by $\mathbf{R^n} \rightarrow \mathbf{R}$ This optimization formulation $\mathbf{R^n} \rightarrow \mathbf{R}$ In order to generate an average accuracy of semantic segmentation, regression function represented as :

$$\hat{f}(X, \theta) = \underset{f}{\operatorname{argmin}} \sum_{i=1}(y, X_i) \tag{8}$$

$$\min \frac{1}{2}||\omega||^2 + C\sum_{i=1}^{n}\xi + C\sum_{i=1}^{n}\theta \tag{9}$$

A. Object Label Score Matching

Illustration of the proccess can be seen in fig.2 In our proposed pipeline (fig. 2), score of semantic segmentation

obtain from already annotated image. In the next step, we simply mask the annotation with the original image to obtain object or non object coefficient extracted from finetune weakly supervised learning CNN. After, we extract coefficient from feature maps, we calculate of gradient predicted label in annotation as relative object coefficient to regress it with average value of annotation score. Later this annotation score could help to optimize semantic segmentation, but it is not done yet in this paper.

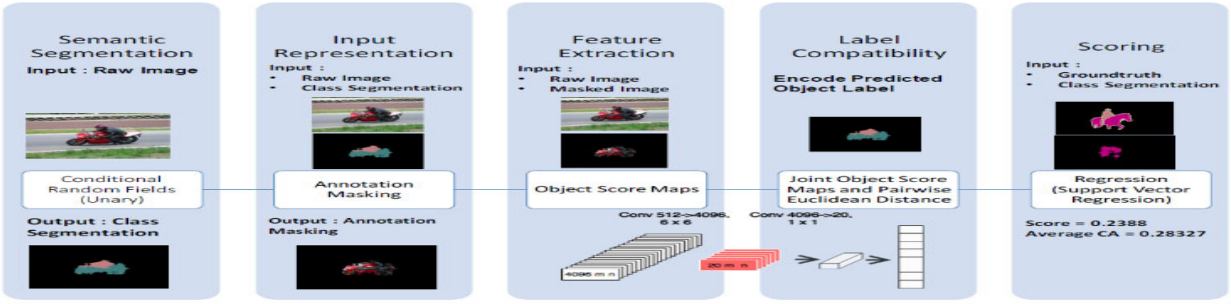

Figure 2. Scoring Segmentation Pipeline

IV. Feature Extraction Using Weakly CNN

In this section, we describe details of proposed method to learn semantic segmentation score from high order object unary. Extracted features from object score maps using global max pooling are probability of class label independence. This score maps layer output are label probability trained in backpropagation algorithm as one vs all classifier. From our preliminary experiments and others work [6], optimizing f-measure based on label indepence only cannot work properly. This because mainly score of annotation are depend on annotation. Another reason because while maximizing F-measure on binary label independently. It is different distribution between predicted label score and known label. Based on this experiment, joint feature of predicted score maps and given segmentation should be used. We formulate joint of feature as relative mutual information between two random variables. This joint features are based on [24]by hyvarinen as

$$\Psi(\xi,\theta) = \frac{\partial\ log\ p(\xi,\theta)}{\partial \xi_n} = \nabla_{\xi} log\ p\ (\xi,\theta) \quad (10)$$

This equation proposed to gradient log output of independent label model, in our case is score map, respect to model value (unary segment model). In scoring segmentation based on score maps, we used gradient log of given output with segment higher order presence or absence. To obtain smooth difference between masked image I' and full image I, we compute pairwise euclidean. This enforce smooth function between predicted annotation and image.

$$||\ \ (x(I),\Theta) -\ \ (x(I'),\Theta)||_D^2 \quad (11)$$

Thus our regression function can be noted as

$$\min \frac{1}{2} \|\omega.\psi(x(I),\Theta) - \psi(x(I'),\Theta)\|_2^2 + C\sum_{i=1}^{n}\xi + C\sum_{i=1}^{n}\theta \quad (12)$$

V. Implementation and Experiment

A. Experiment Environment

We conduct experiment in Personal Computer (PC). Specification of PC in this experiment is using processor Intel(R) Core(TM) i7-3770 CPU @ 3.40 GHx, graphics card NVIDIATM GTX 970, memory 16GB RAM, and operation system Linux Ubuntu 14.04. In this research we also use libraries and frameworks. Some of them are: 1) Torch7 [25], a deep learning framework for implementation of Convolutional Neural Network. 2) OpenCV [26], computer vision library for masking from result of unary prediction into raw image. 3) Textonboost [14], library for generating unary term of input image. 4) Open Graphical Model (OpenGM) [27], library for modelling a discrete graph to do a structured prediction using conditional random field. 5) VOCdevKit [23] - Matlab, tools for evaluate accuracy of semantic segmentation. 6) Liblinear/Libsvm, library for classification and regression of data into target label.

B. Dataset

This research use dataset PASCAL VOC 2012. This dataset contains several types of images for different purpose, such as: object classification, object detection, human pose estimation, and semantic segmentation. For research purpose we only use images of semantic segmentation and object classification problem.

C. *Experiment* Step

Our experiment consists of two step. Our first scenario is to finetuning CNN in order to learn target class label domain. In the next phase, we perform learning semantic segmetation score using feature extracted from feedforward CNN.

1) Finetuning of Convolutional Neural Network: Finetuning is a process to re-train a trained model with new parameters with transfer learning method. In this scenario we re-train classification model from PASCAL dataset images with a weak labels. This scenario runs in two phases, training and testing phase. Traning phase conducts experiment different loss function, negative log likelihood, with label {-1,1} and sigmoid cross entropy with label {1,0} Testing phase conducts to measure the performance of finetuned model. In finetune training scenario of Weakly CNN, we use measurement of classification based on standard PASCAL competition of object recognition [23].

2) Semantic Segmentation Scoring: In order to obtains average score of annotation, we use unary and pairwise potential data from textonboost [14]. Our primary goal learns an agnostic framework of learning semantic segmentation score. This means that it could handle scoring generally. In order to do so regression score had to be balanced and spreading from 0 to 1. Based on previous work [6] average of multilabel classification score are macro F_β score. This formulation is known to have non linear function. In scoring segmentation we use extracted object score maps for regression. Object score maps formulation as predicting exact score allow us to use object presence and absence as measuring and optimizing macro F_β with label independence and joint mode of predicted label and given label based on [6].

$$|y - f(x,\omega)|_\varepsilon = \begin{cases} 0 & if\ |y - \langle x,\omega\rangle + \beta| - \varepsilon < \varepsilon \\ |y - \langle x,\omega\rangle + \beta| - \varepsilon & otherwise \end{cases} \quad (13)$$

VI. Result and Analysis

In this section we will describe the experiment result based on implementation setup that describe in previous section. Then, we will give an analysis based on the experiment result.

A. Experiment Result and Analysis

1) Convolutional Neural Network Finetuning Experiment and Analysis: Finetuning conducted in this experiment aims to train the adaptation layer by using different loss function and label. In this experiment we use hyperparameter that can be seen on Table I. Pretrain CNN is finetune using batch gradient descent optimization. This batch will be updated after every 10 iteration. We also use L2 regularization in order to prevent overfitting given our large number of parameter. Learning rate are set in 1e-3 as we want to iteration converge faster, however it could also lead into trap on local maximum. Table II shows accuracy of our finetuned network. Result of experiment shows that we get moderate 70% average accuracy of finetuned Weakly CNN. We argue that our model are not too deep enough to generalize. example illustration input and output of finetuned network can be seen in fig.3(a)(above, input image), fig.3(b)(below left side, output feature maps), fig.3(c)(below right side, output feature maps). In the middle above, example of input image containing object labeled as aeroplane, while pictures below are output of feature maps from negative log likelihood and binary cross entropy consecutively.

TABLE I Experiment Parameter

Parameter	Value
Batchsize	10
learning rate	1e-3
Regularization	1e-5 (L2)
Epoch	100

2) Experiment and Analysis of Learning Semantic Segmentation Score: This experiment aims to predict score of annotation label on semantic segmentation. Score measurement is done by learning based on the image data masking with annotation label. Feature is represented by model convolutional neural network which has been trained on finetuning experiment. Feature representation is gained by feedforwarding image and masked image into Convolutional Neural Network and learnin score by regression. Hyperparameters setting used in this experiment can be seen on Table III.

The experiment result with those parameter setting can be seen on Table IV. Based on the experiment result, the mean value of MSE on object score map and feature map are ~0.015%. Table IV shows scoring value. Result of R2 correlation is around 0.5. However we believe this correlation are not well representative as there is limited number of training and imbalanced average precision. Example of score prediction of our proposed method illustrated in fig.4(a), 4(b), 4(c), 4(d), 5(a), 5(b), 5(c), 5(d), 6(a), 6(b), 6(c), 6(d). In this illustration, group of figure {4,5,6}.a shows image, group of figure {4,5,6}.b shows masked image, group of figure {4,5,6}.c shows groundtruth, group of figure {4,5,6}.d shows annotation

TABLE II CNN Finetuning Result

No	Classes	NLL	BCE
1	aeroplane	88.218	91.864
2	bicycle	54.828	53.765
3	bird	81.720	83.247
4	boat	66.939	68.537
5	bottle	34.000	33.287
6	bus	68.182	69.307
7	car	68.667	69.912
8	cat	78.318	79.092
9	chair	38.690	38.229
10	cow	73.288	81.164
11	diningtable	35.088	29.630
12	dog	71.577	74.529
13	horse	54.825	54.709
14	motorbike	62.903	63.317
15	person	91.808	91.056
16	pottedplant	78.652	80.702
17	sheep	87.692	89.683
18	sofa	64.407	66.667
19	train	93.846	94.385
20	tvmonitor	89.474	90.449
21	background	—-	—-
	Averaged CA	69.156	70.176
	Union CA	55.360	56.454
	Gobal Accuracy	70.221	70.138

TABLE III Experiment Parameter

Parameter	Value
E (margin)	0.1
v (support vector)	0.5 (n * total sample)
Kernel	RBF (Gaussian)

TABLE IV Score Prediction Experiment Result of Semantic Segmentation

SVR Type	Kernel	Feature	MSE	R^2
v-SVR	Linear	Object (NLL)	0.0154377	0.54747
E-SVR	Linear	Object (NLL)	0.0153997	0.544074
v-SVR	RBF	Object (NLL)	0.0145547	0.570143
E-SVR	RBF	Object (NLL)	0.014722	0.564973
v-SVR	Linear	Object (BCE)	0.015272	0.550995
v-SVR	RBF	Object (BCE)	0.0144629	0.572937
E-SVR	Linear	Object (BCE)	0.0152194	0.549448
E-SVR	RBF	Object (BCE)	0.014674	0.566681

Note :
NLL = Negative Log Likelihood
BCE = sigmoid Binary Cross Entropy
E = margin error tolerance in SVR
ν = number of support vector used as training
RBF = Radial Basis Function

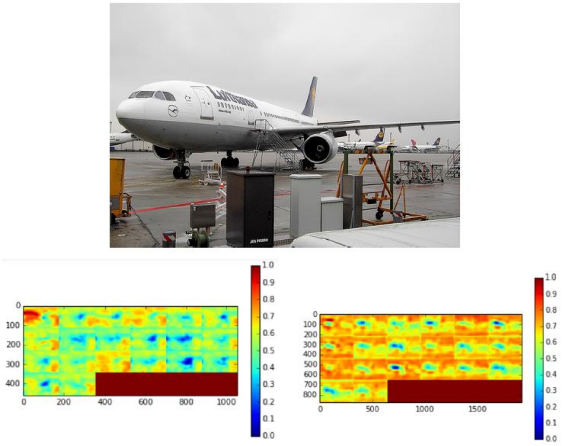

Figure 3. Original Images and Feature Maps Result

Above image is original image, left below score maps output using negative log likelihood loss, right below score maps output using sigmoid + binary cross entropy.

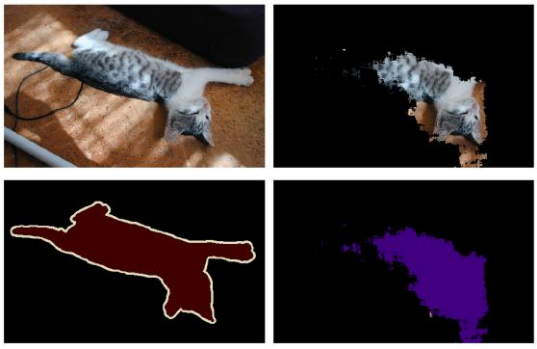

Figure 4. Groundtruth {cat}, annotation {dog}

From left to right image, groundtruth, annotation, masked image, Score = 0.16653, Average CA = 0.16908

Figure 5. Groundtruth {person, horse}, annotation{horse}

From left to right image, groundtruth, annotation, masked image, Score = 0.2388, Average CA = 0.28327

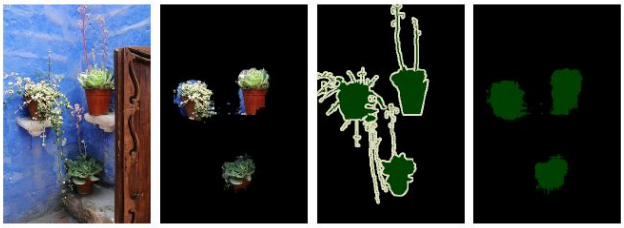

Figure 6. Groundtruth{potted plant}, annotation{potted plant}

From left to right image, groundtruth, annotation, masked image, Score = 0.13535, Average CA = 0.56039.

VII. Conclusion

In this paper we proposed learning semantic segmentation score, namely jaccard index. This performance metric is similarity of annotation and groundtruth. We learn score of annotation score by using average jaccard score over class by using regression. By training max score of object score maps as multiple instance learning, in this paper we formulate joint mutual information of predicted label from weakly supervised learning and given label of annotated image. Result from semantic segmentation experiment of mean square error is 0.015 and R2 correlation is around 0.55. This correlation is mainly unary of relative information from predicted foreground label, full image given annotated label. Another way to improve scoring segmentation performance of r2 is create more instance accuracy in balanced or online training. This result encouraging weakly supervised learning to improve semantic segmentation performance without groundtruth.

Acknowledgement

This research was supported by Universitas Indonesia and Directorate General of Higher Education, under Grant Research Collaboration and Scientific Publication No: 0403/UN2.RI2/HKP.05.00/2015.

References

[1] J. Shotton, J. Winn, C. Rother, and A. Criminisi, “Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context,” Int. J. Comput. Vision, vol. 81, no. 1, pp. 2–23, Jan. 2009. [Online]. Available: http://dx.doi.org/10.1007/s11263-007-0109-1

[2] P. Krahenb¨uhl¨ and V. Koltun, “Efficient inference in fully connected crfs with gaussian edge potentials,” in Advances in Neural Information Processing Systems 24, J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, Eds. Curran Associates, Inc., 2011, pp. 109–117. [Online]. Available: http://papers.nips.cc/paper/4296-efficient-inference-in-fully-connected-crfs-with-gaussian-edge-potentials.pdf

[3] P. Kohli, L. Ladicky, and P. Torr, “Robust higher order potentials for en-forcing label consistency,” in Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, June 2008, pp. 1–8.

[4] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, “Conditional random fields: Probabilistic models for segmenting and labeling sequence data,” in Proceedings of the Eighteenth International Conference on Machine Learning, ser. ICML ’01. San Francisco, CA,

USA: Morgan Kaufmann Publishers Inc., 2001, pp. 282–289. [Online]. Available: http://dl.acm.org/citation.cfm?id=645530.655813

[5] F. Li, J. Carreira, and C. Sminchisescu, "Object recognition as ranking holistic figure-ground hypotheses," in Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, June 2010, pp. 1712– 1719.

[6] W. Waegeman, K. Dembczynski,´ A. Jachnik, W. Cheng, and E. Hullermeier,¨ "On the bayes-optimality of f-measure maximizers," Journal of Machine Learning Research, vol. 15, pp. 3333–3388, 2014. [Online]. Available: http://jmlr.org/papers/v15/waegeman14a.html

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Advances in Neural Information Processing Systems 25, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097– 1105. [Online]. Available: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

[8] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," CoRR, vol. abs/1312.6229, 2013. [Online]. Available: http://arxiv.org/abs/1312.6229

[9] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: an astounding baseline for recognition," CoRR, vol. abs/1403.6382, 2014. [Online]. Available: http://arxiv.org/abs/1403.6382

[10] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets"," in British Machine Vision Conference, 2014.

[11] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Is object localization for free? weakly-supervised learning with convolutional neural networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015.

[12] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," International Journal of Computer Vision (IJCV), pp. 1–42, April 2015.

[13] D. Vektor, Aprinaldi, Z. Ian, and J. Wisnu., "A novel knowledge-compatibility benchmarker for semantic segmentation," Jun 2015.

[14] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "Textonboost: Joint appearance, shape and context modeling for mulit-class object recognition and segmentation," in European Conference on Computer Vision (ECCV), January 2006. [Online]. Available: http://research.microsoft.com/apps/pubs/default.aspx?id=6745

[15] B. Babenko, "Multiple instance learning: algorithms and applications," View Article PubMed/NCBI Google Scholar, 2008.

[16] F. Li and C. Sminchisescu, "Convex multiple-instance learning by estimating likelihood ratio," in Advances in Neural Information Processing Systems 23, J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, Eds. Curran Associates, Inc., 2010, pp. 1360–1368. [Online]. Available: http://papers.nips.cc/paper/3926- convex-multiple-instance-learning-by-estimating-likelihood-ratio.pdf

[17] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Perez,´ "Solving the multiple instance problem with axis-parallel rectangles," Artif. Intell., vol. 89, no. 1-2, pp. 31–71, Jan. 1997. [Online]. Available: http://dx.doi.org/10.1016/S0004-3702(96)00034-3

[18] J. D. Keeler, D. E. Rumelhart, and W.-K. Leow, "Integrated segmentation and recognition of hand-printed numerals," in Proceedings of the 1990 Conference on Advances in Neural Information Processing Systems 3, ser. NIPS-3. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1990, pp. 557–563. [Online]. Available: http://dl.acm.org/citation.cfm?id=118850.118942

[19] D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," The Journal of Physiology, vol. 160, no. 1, pp. 106–154, 1962. [Online]. Available: http://dx.doi.org/10.1113/jphysiol.1962.sp006837

[20] S. J. Pan and Q. Yang, "A survey on transfer learning," Knowledge and Data Engineering, IEEE Transactions on, vol. 22, no. 10, pp. 1345– 1359, Oct 2010.

[21] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in CVPR, 2014.

[22] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" CoRR, vol. abs/1411.1792, 2014. [Online]. Available: http://arxiv.org/abs/1411.1792

[23] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisser-man, "The pascal visual object classes (voc) challenge," International Journal of Computer Vision, vol. 88, no. 2, pp. 303–338, jun 2010.

[24] A. Hyvarinen,¨ "Estimation of non-normalized statistical models by score matching," in Journal of Machine Learning Research, 2005, pp. 695– 709.

[25] R. Collobert, K. Kavukcuoglu, and C. Farabet, "Torch7: A matlab-like environment for machine learning," in BigLearn, NIPS Workshop, 2011.

[26] G. Bradski, Dr. Dobb's Journal of Software Tools.

[27] B. Andres, B. T., and J. H. Kappes, "OpenGM: A C++ library for discrete graphical models," ArXiv e-prints, 2012. [Online]. Available: http://arxiv.org/abs/1206.0111