Learning from Counting: Leveraging Temporal Classification for Weakly Supervised Object Localization and Detection

Chia-Yu Hsu
chsu53@asu.edu
Wenwen Li*
wenwen@asu.edu

School of Geographical Science and Urban Planning, Arizona State University
Tempe, AZ 85281 USA

Abstract

This paper reports a new solution of leveraging temporal classification to support weakly supervised object detection (WSOD). Specifically, we introduce raster scan-order techniques to serialize 2D images into 1D sequence data, and then leverage a combined LSTM (Long, Short-Term Memory) and CTC (Connectionist Temporal Classification) network to achieve object localization based on a total count (of interested objects). We term our proposed network LSTM-CCTC (Count-based CTC). This "learning from counting" strategy differs from existing WSOD methods in that our approach automatically identifies critical points on or near a target object. This strategy significantly reduces the need of generating a large number of candidate proposals for object localization. Experiments show that our method yields state-of-the-art performance based on an evaluation on PASCAL VOC datasets.

1 Introduction

Object detection (OD) using deep learning, more specifically, deep convolution neural networks (DCNN), has been broadly applied in vision tasks, such as detecting and tracking moving objects from remotely sensed images, surveillance videos, and autonomous robots [5, 10, 19, 20]. A great challenge in such tasks is the labor-intensive nature of preparing object-level labels, such as object class, which provides category information (image-level annotation), and object location – a bounding box (BBOX) showing the extent of each target object. This issue has drawn researchers' attention to developing Weakly Supervised Objection Detection (WSOD) approaches [29] that leverage weak labels (i.e., image-level annotation only) to achieve high-confidence object detection to alleviate the high cost associated with object labeling.

Like strongly supervised OD networks, such as Faster RCNN [21], a WSOD network typically consists of three key tasks in the object detection pipeline: (1) feature extraction: using a DCNN to extract low- to high-level features from the input images, (2) detection: relying on a region proposal network (RPN) to generate candidate region proposals containing

 *corresponding author: wenwen@asu.edu

the target objects, and (3) recognition: using a classifier to predict the object class. However, unlike the RPN used in Faster-RCNN, which was fed and trained with the ground-truth BBOX data, the challenge of WSOD is to predict region proposals that are highly likely to contain a target object without providing the BBOX information.

There are two research directions in advancing WSOD: making improvement on the classifiers or developing new RPNs to generate more accurate proposals. Many existing works were reported to leverage image-level annotation to develop and refine proposal classifiers[2, 6, 8, 9]. However, it is still very difficult for a WSOD network to achieve a level of prediction performance similar to that of strongly supervised approaches. One reason is that these works simply use the off-the-shelf region generation techniques such as selective search (SS) [26] or edge boxes (EB) [32], resulting in limited performance increase. Recent studies [18] have shown that the quality of proposals greatly affect the predictive performance of a network. Therefore, research taking the second direction – developing new RPNs has the potential to further boost the WSOD performance.

In this paper, we introduce a new solution in generating high-quality proposals by enabling a way of "learning from counting." Unlike existing networks that need to generate a large number of candidate proposals and then select a subset from them, our proposed network can achieve better detection performance by automatically locating critical points on or near a target object. By generating a small number of proposals around the critical points, a set of high-quality proposals can be obtained and sent to the next WSOD stage. Our research is motivated by the use of a combined LSTM [30] and CTC in its outstanding capability in segmenting sequential data without per-frame labels, an idea similar to weakly supervised learning. To leverage this temporal classification network, we further apply a dimension reduction strategy to serialize input image into 1D sequential data and identify the critical object location leveraging count-based learning.

To summarize, this work has made the following contributions: (1) It introduces for the first time the use of a Recurrent Neural Network (RNN)- LSTM as the proxy of a 'weak' RPN to improve WSOD performance. (2) The proposed RPN can easily be integrated into any WSOD network to generate high-quality proposals. (3) It enables a fully automated, end-to-end training framework with multiple independent data streams for region generation and classification to prevent the network from getting stuck in local optima. (4) Our method achieves state-of-the-art performance in WSOD.

2 Literature Review

2.1 Weakly supervised object detection (WSOD)

Existing efforts to improve the WSOD performance depends mainly on two research directions - developing better proposal classifiers and developing new RPNs to generate more accurate proposals. [2] developed an effective, end-to-end deep network for WSOD, in which a pre-trained CNN is used for feature extraction and two data streams are developed to undertake detection and recognition in parallel. However, this model tends to assign a higher score to a proposal that contains the most discriminative part of an object rather than the entire object. [23] designed a strategy to assign the same image label for proposals that have significant overlaps with those receiving high scores during the weak supervision phase. An Online Instance Classifier Refinement (OICR) algorithm was then developed to use these proposals as pseudo ground-truths to classify the training images. Through continuous re-

finements, the proposed WSOD can achieve better instance recognition than the network in [2]. [28] developed a C-MIL (Continuation multiple instance learning) model to achieve WSOD using new loss functions. [12] developed a Count-guided Weakly Supervised Localization (C-WSL) network to achieve high-confidence OD. This work addressed the issue of the tendency to draw a proposal containing multiple objects in existing weakly supervised detectors. [13] leveraged segmentation maps with coupled multiple instance detection network (C-MIDN) to refine the proposals before sending them to the classifier, which also uses the OICR model. [31] integrated bounding-box regression (REG) into MIL as a single end-to-end network and enhanced the original feature map with implicit object location information by attention maps from images (guided attention module) (GAM). Two MIL approaches were adopted in this work: OICR and Proposal Cluster Learning (PCL)[24].

All the above approaches aim to improve one or more stages of a WSOD network. However, their proposal generation processes mostly rely on mature techniques, such as SS or EB. However, [18] and [25] argue that the quality of proposals has a significant impact on the overall OD performance. Therefore, in recent years, more studies have been undertaken to improve proposal generation in RPNs. [9] developed cascaded multiple networks with created class activation maps to infer better region proposals. [25] proposed a two-stage network to improve the quality of generated proposals. The refined proposals are then sent to another WSOD [23] to perform classification. Our research is also towards developing a RPN which can generate better proposals. But we take a very different approach - instead of relying on traditional object detection in a 2D realm, our approach converts the 2D object detection problem into a 1D sequence data segmentation problem and solves it by a novel use of LSTM and a count-based CTC.

2.2 Image labeling with LSTM and CTC

LSTM [17] is a type of RNN that was originally designed to model sequence data. LSTM can propagate information through lateral connections to model short- and long-term context dependencies. Regarding its usage in image processing tasks [1, 3], LSTM networks need to be extended from the temporal domain to the spatial domain and from single-dimensional learning to multi-dimensional learning [15]. [27] coupled four uni-dimensional RNNs to sweep across an image in four directions to replace the common convolutional-pooling layer. The network ensures that the output activation will appear in a specific location with respect to the whole image rather than a local context window of a CNN. In our work, we adopt the LSTM structure to identify critical points on a target object by taking advantage of its power in capturing global contexts and context dependencies in serialized data.

CTC [16] is a type of scoring function and a neural network output designed for training RNNs to tackle sequence learning problems such as speech recognition. Instead of the need for per-frame labels, a CTC only needs “phoneme”-level labels whereby one phoneme can be mapped to multiple timeframes in the original speech audio. A CTC achieves this by transforming the network outputs into conditional probabilities and calculating the overall probabilities of all possible alignments that yield the same label sequence. It then finds the most probable sequence and corresponding alignment and uses that as the final output. The alignment can be represented by a set of segmented positions that separate each “phoneme” in a speech recognition problem. CTC’s ability to handle time sequence data and make predictions without the per-frame labels significantly broadened the applicability of RNNs.

This paper combines the LSTM and CTC to enable a novel WSOD by leveraging temporal classification. Next section introduces our methodology in detail.

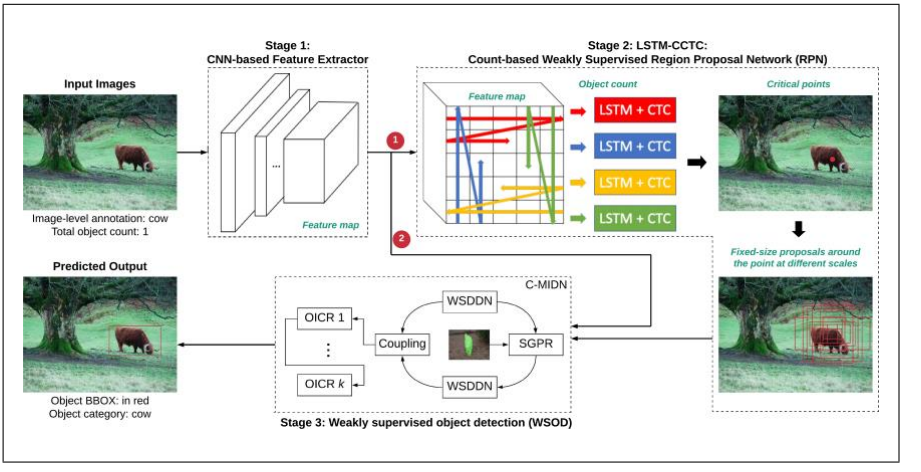

Figure 1: Architecture of the proposed LSTM-CCTC for count-based WSOD.

3 Methodology

Figure 1 illustrates the overall WSOD network with our proposed LSTM-CCTC as the RPN. The deep learning pipeline is divided into three main stages. First, a pre-trained CNN is used as the feature extractor. Second, we propose a new architecture for proposal generation. In this RPN, the resulting feature map will first be serialized into a 1D vector. This vector together with the total-class count will then be sent to the proposed LSTM-CCTC network to identify critical points falling on or near the target objects. A number of proposals with different aspect ratios and size will be generated around the critical points and sent to the last-stage of the WSOD network for proposal classification. While our proposed RPN (LSTM-CCTC) can be integrated into any WSOD, in this paper we used the one reported in [13] for the classification task in Stage 3. We introduce details of each component in the following sections.

3.1 Feature extractor

Our feature extractor is built on a pretrained VGG16 neural network [22] that is trained on ImageNet [7] using image-level labels. The feature map after the last convolutional block (convolution + ReLU) is fed into the proposed LSTM+CCTC for proposal generation and then the C-MIDN [13] for classification.

3.2 Region proposal network (RPN)

Our PRN consists of four LSTMs that process the 1D feature map serialized by scan orders in four different directions (the four colored arrows on the feature map in Stage 2 of Figure 1). Each LSTM is connected to a CTC layer. The original CTC framework is designed for segmenting sequence data, while here we apply it to WSOD. To do so, we perform a transformation of the feature map with the size of $n \times n \times k$ to a 1D vector with a length of n^2. Each element is at $1 \times k$ in dimension. This conversion makes the feature map suitable to

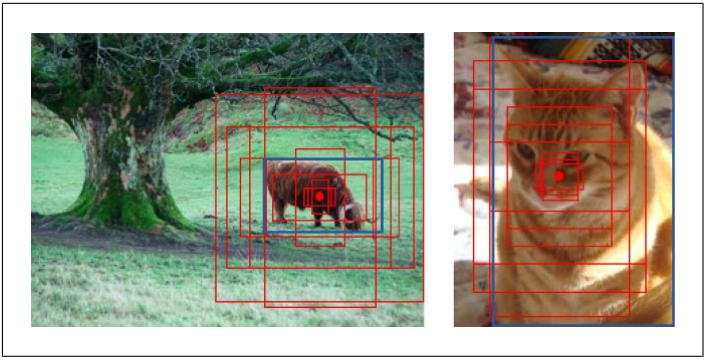

Figure 2: Output from the proposed RPN (red point: segmented critical point by LSTM-CCTC, blue box: ground truth, red box: sample proposals.

serve as the input of the LSTM + CTC network. Different from the traditional CTC model which trains on and predicts a sequence of (different) labels, our model is based on a count-based CTC, or CCTC, the goal of which is to inspect objects without the need to differentiate object type. So this LSTM-CCTC can also be considered as a binary segmentation problem. The object type is identified by the classifier at the next stage of the WSOD pipeline. After critical points are located, initial proposals with different aspect ratios (1:1, 2:1, 1:2, 1:3 and 3:1) are generated. Then, proposals are gradually enlarged while keeping at the same shape (Figure 2) until they hit the border of the feature map.

3.3 Weakly supervised object detection (WSOD)

In this stage, a region-based CNN is trained to classify proposals into different classes. In this work, we adopt C-MIDN [13] as our WSOD network (Figure 1). This network couples two WSDDN [2] and contains iterative refinement in its instance classifier [23]. The original WSDDN map proposal scores in a given image to an image-level classification confidence. Therefore, it can be trained solely under image-level supervision by optimizing a multi-class cross-entropy loss. However, the WSDDN often localizes the most discriminative object parts instead of the entire object because of the non-convex optimization. To address this issue, C-MIDN leverages a pair of WSDDNs, and the top-scoring proposals of the first WSDDN are removed from the input of the second WSDDN, preventing the second WSDDN from localizing the same proposal again. Furthermore, a segmentation map of the image is used to improve the robustness of the proposal removal process (Segmentation Guided Proposal Removal) (SGPR) to ensure that the full-context proposals will be kept. If the coverage rate between the segmentation map and the proposal is too small, it is more likely that there exist other tight proposals. Finally, proposals from the two WSDDNs are coupled to generate the final detection results. Next, the OICR training is adopted. Each stage is trained under the supervision of instance labels obtained from the previous stage. To obtain instance labels for supervision, given an image with a class label c, the proposal j with the highest score for class c will be used as the pseudo ground-truth BBOX. Besides labeling j, other proposals which have a high spatial overlap with j will also be labeled as class c. The refinement strategy will give a preference to select the BBOX that contains the entire object instead of a part of it. It is very convenient to generate the pseudo ground-truth BBOX from

our proposed RPN because these BBOXes (shown in Figure 2) share the same center, and many of them have a large overlap.

3.4 Implementation details

In the feature extractor implementation, we replaced the last pooling layer with a SPP layer and combined it with the proposed RPN and a WSOD network ([13]). In the RPN, the LSTMs are unidirectional with 2 layers, with the size at 512 for the input layer, and 256 for the hidden layer. The outputs are connected to the same fully-connected layer with the output size of 2 for binary object detection. We followed the original settings in [13] for training and testing the WSOD network. During training, since the proposals from the first several iterations are noisy, we trained the proposed LSTM-CCTC for 20 epochs and then connected it with the WSOD network for an end-to-end training. The number of objects in each image is transferred into a sequence (e.g. 3 is transferred into 'ooo', where 'o' refers to a target object of any class) to train the LSTM-CCTC. All the initialization for newly added layers used Gaussian distributions with 0-mean and standard deviation 0.01. For the optimization, we used Stochastic Gradient Descent (SGD) with a min-batch size of 2. The learning rate is 0.001 for the first 200 epochs and 0.0001 for the following epochs. The momentum and weight decay are set to 0.9 and 0.0005 respectively.

We also trained Fast R-CNN [14] with top-scoring proposals generated by our proposed approach as the pseudo ground-truth. This is a common practice to improve the detection performance [9, 13, 23, 25].

4 Experiments

In this section, we demonstrate the outstanding performance of our method through a series of experiments using benchmark datasets. We will also illustrate how different factors in the proposed method affect the prediction performance through ablation experiments.

4.1 Experimental setup

Datasets We evaluated our method on PASCAL VOC 2007 and VOC 2012 datasets [11] which are two widely used benchmarks in WSOD. For both datasets, we combined training and validation images as the *trainval* set for training and used test images for testing. We generated the object count for each image from the BBOX annotations to train our proposed RPN and used image-level labels to train the Stage 3 WSOD network.

Evaluation metrics We used two performance metrics for evaluation: mean average precision (mAP) [11] and correct location (CorLoc) [8]. mAP is a standard metric to evaluate the prediction accuracy and CorLoc measures the localization accuracy of a trained model.

Benchmarks We compare our method with seven other popular WSOD networks, including (1) WSDDN [2], one of the phenomenal CNN-based networks for WSOD; (2) OICR [23], an improved WSDDN with a classifier refinement algorithm OICR; (3) WSRPN [25], a CNN network focusing on generating high-quality proposals. This work shares the same goal as our proposed method; (4) C-WSL [12], which uses per-class count as the weak labels for WSOD; and three most recent state-of-the-art WSOD solutions: (5) C-MIDN [13], (6) C-MIL [28], and (7) MIL-OICR(PCL)+GAM+REG [31].

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
WSDDN [2]	39.4	50.1	31.5	16.3	12.6	64.5	42.8	42.6	10.1	35.7	24.9	38.2	34.4	55.6	09.4	14.7	30.2	40.7	54.7	46.9	34.8
OICR [23]	58.0	62.4	31.1	19.4	13.0	65.1	62.2	28.4	24.8	44.7	30.6	25.3	37.8	65.5	15.7	24.1	41.7	46.9	64.3	62.6	41.2
WSRPN [25]	57.9	70.5	37.8	05.7	21.0	66.1	69.2	59.4	03.4	57.1	**57.3**	35.2	64.2	68.6	**32.8**	28.6	50.8	49.5	41.1	30.0	45.3
C-WSL+ODR [12]	62.7	63.7	40.0	25.5	17.7	70.1	68.3	38.9	25.4	54.5	41.6	29.9	37.9	**64.2**	11.3	27.4	49.3	54.7	61.4	**67.4**	45.6
C-WSL+ODR* [12]	**62.9**	64.8	39.8	28.1	16.4	69.5	68.2	47.0	27.9	55.8	43.7	31.2	43.8	65.0	10.9	26.1	52.7	55.3	60.2	66.6	46.8
C-MIL [28]	62.5	58.4	49.5	32.1	19.8	70.5	66.1	63.4	20.0	60.5	52.9	53.5	57.4	68.9	08.4	24.6	51.8	58.7	66.7	63.5	50.5
MIL-OICR+GAM+REG [31]	55.2	66.5	40.1	31.1	16.9	69.8	64.3	67.8	27.8	52.9	47.0	33.0	60.8	64.4	13.8	26.0	44.0	55.7	68.9	65.5	48.6
MIL-PCL+GAM+REG [31]	57.6	70.8	**50.7**	28.3	**27.2**	**72.5**	69.1	65.0	26.9	64.5	47.4	47.7	53.5	66.9	13.7	**29.3**	56.0	54.9	63.4	65.2	51.5
C-MIDN [13]	53.3	71.5	49.8	26.1	20.3	70.3	**69.9**	68.3	**28.7**	**65.3**	45.1	**64.6**	58.0	71.2	20.0	27.5	54.9	54.9	69.4	63.5	52.6
LSTM-CCTC+OICR (Ours)	56.9	**73.4**	40.7	26.3	16.4	66.8	65.3	68.3	23.5	58.3	40.9	39.6	40.4	62.8	24.6	25.1	48.3	51.2	57.7	62.9	47.5
LSTM-CCTC+C-MIDN (Ours)	60.2	71.7	47.6	**29.4**	25.2	72.1	66.8	**69.6**	27.4	62.9	45.8	58.3	54.6	67.8	25.4	26.8	**60.1**	**60.2**	65.4	64.2	**53.1**
OICR-Ens.+FRCNN [23]	**65.5**	67.2	47.2	21.6	22.1	68.0	68.5	35.9	05.7	63.1	49.5	30.3	64.7	66.1	13.0	25.6	50.0	57.1	60.2	59.0	47.0
C-WSL+ODR+FRCNN [12]	61.9	61.9	48.6	28.7	23.3	71.1	71.3	38.7	**28.5**	60.6	45.4	26.3	49.7	65.5	07.2	27.3	54.7	61.6	63.2	59.5	47.8
C-WSL+ODR*+FRCNN [12]	62.9	68.3	52.9	25.8	16.5	71.1	69.5	48.2	26.0	58.6	44.5	28.2	49.6	66.4	10.2	26.4	55.3	59.9	61.6	62.2	48.2
WSRPN-Ens.+FRCNN [25]	63.0	69.7	40.8	11.6	**27.7**	70.5	**74.1**	58.5	10.0	66.7	**60.6**	34.7	**75.7**	**70.3**	25.7	26.5	55.4	56.4	55.5	54.9	50.4
C-MIL+FRCNN [28]	61.8	60.9	56.2	28.9	18.9	68.2	69.6	71.4	18.5	64.3	57.2	66.9	65.9	65.7	13.8	22.9	54.1	**61.9**	68.2	**66.1**	53.1
C-MIDN+FRCNN [13]	54.1	**74.5**	**56.9**	26.4	22.2	68.7	68.9	**74.8**	25.2	64.8	46.4	**70.3**	66.3	67.5	21.6	24.4	53.0	59.7	**68.7**	58.9	53.6
LSTM-CCTC+FRCNN (Ours)	63.2	73.6	50.2	**31.7**	24.6	**73.4**	69.3	72.6	28.3	**67.1**	53.9	55.7	64.3	66.1	**26.8**	**27.4**	**61.2**	60.8	62.5	59.4	**54.6**

Table 1: Result comparison in terms of AP (%) and mAP (%) on the PASCAL VOC 2007 *test* set

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mean
WSDDN [2]	65.1	58.8	58.5	33.1	39.8	68.3	60.2	59.6	34.8	64.5	30.5	43.0	56.8	82.4	25.5	41.6	61.5	55.9	65.9	63.7	53.5
OICR [23]	81.7	80.4	48.7	49.5	32.8	81.7	85.4	40.1	40.6	79.5	35.7	33.7	60.5	88.8	21.8	57.9	76.3	59.9	75.3	81.4	60.6
C-WSL+ODR [12]	**86.3**	80.4	58.3	50.0	36.6	**85.8**	86.2	47.1	42.7	81.5	42.2	42.6	50.7	90.0	14.3	61.9	85.6	64.2	77.2	82.4	63.3
C-WSL+ODR* [12]	85.8	81.2	64.9	50.5	32.1	84.3	85.9	54.7	**43.4**	80.1	42.2	42.6	60.5	90.4	13.7	57.5	82.5	61.8	74.1	82.4	63.5
WSRPN [25]	77.5	81.2	55.3	19.7	44.3	80.2	**86.6**	69.5	10.1	87.7	**68.4**	52.1	**84.4**	**91.6**	**57.4**	**63.4**	77.3	58.1	57.0	53.8	63.8
C-MIDN [13]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	65.0
MIL-OICR+GAM+REG [31]	81.7	81.2	58.9	54.3	37.8	83.2	86.2	77.0	42.1	83.6	51.3	44.9	78.2	90.8	20.5	56.8	74.2	**66.1**	81.0	**86.0**	66.8
MIL-PCL+GAM+REG [31]	80.0	83.9	**74.2**	53.2	**48.5**	82.7	86.2	69.5	39.3	82.9	53.6	61.4	72.4	91.2	22.4	57.5	83.5	64.8	75.7	77.1	68.0
C-MIDN [13]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	68.7
LSTM-CCTC+OICR (Ours)	76.9	**85.7**	60.0	52.5	32.1	82.6	84.2	80.5	37.7	**88.1**	45.5	55.2	56.4	85.3	37.1	55.6	78.3	59.6	69.4	82.3	65.3
LSTM-CCTC+C-MIDN (Ours)	78.5	83.3	63.9	**57.6**	43.8	85.5	84.4	**83.2**	39.0	87.8	50.4	**67.5**	67.8	90.2	42.5	46.3	**87.9**	62.5	**84.4**	84.6	**70.0**
OICR-Ens.+FRCNN [23]	85.8	82.7	62.8	45.2	43.5	84.8	87.0	46.8	15.7	82.2	51.0	45.6	83.7	91.2	22.2	59.7	75.3	65.1	76.8	78.1	64.3
C-WSL+ODR+FRCNN [12]	85.8	78.0	61.6	52.1	44.7	81.7	88.4	49.1	50.0	82.9	44.1	44.4	63.9	92.4	14.3	60.4	86.6	68.3	**80.6**	82.8	65.6
C-WSL+ODR*+FRCNN [12]	**87.5**	81.6	65.5	52.1	37.4	83.8	87.9	57.6	50.3	80.8	44.9	44.4	65.6	92.8	14.9	**61.2**	83.5	**68.5**	77.6	**83.5**	66.1
WSRPN-Ens.+FRCNN [25]	83.8	82.7	60.7	35.1	**53.8**	82.7	**88.6**	67.4	22.0	86.3	**68.8**	50.9	**90.8**	**93.6**	44.0	**61.2**	82.5	65.9	71.1	76.7	68.4
C-MIDN+FRCNN [13]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	71.9
LSTM-CCTC+FRCNN (Ours)	84.5	**84.7**	**66.4**	**59.1**	43.3	**86.8**	85.3	**88.5**	**52.4**	**87.0**	60.3	**71.4**	80.3	89.9	**46.3**	58.1	**88.4**	59.4	78.8	81.3	**72.6**

Table 2: Result comparison in terms of CorLoc (%) on the PASCAL VOC 2007 *trainval* set

4.2 Comparison with the state-of-the-arts

Table 1 shows the AP and mAP evaluated on PASCAL VOC 2007 dataset. It can be seen that our method (LSTM-CCTC+C-MIDN) outperforms all other related works. In particular, our proposed model achieves better performance (7.8% higher in mAP) than WSRPN [25], which also aims at improving the quality of generated proposals. Our proposed method also beats the count-guided C-WSL (C-WSL+ODR) [12] by 7.5% even with the use of weaker labels (total object count) than the per-class count used in C-WSL. We also integrated our proposed RPN into two well-performed WSOD frameworks: OICR[23] and C-MIDN[13]. The integrated networks yield a 6.3% and a 0.6% increase in mAP than [23] and [13], respectively. This owes solely to the introduction of LSTM-CCTC for proposal generation.

Our model also achieves the state-of-the-art performance by training Fast R-CNN with pseudo ground-truths. The statistics on CorLoc are shown in Table 2. The results show that our proposed WSOD achieves the best CorLoc among all methods compared, including those utilizing ensemble learning, such as WSRPN-Ens.+FRCNN[25].

The same experiments were conducted on VOC 2012 and results are shown in Table 3. The results verify that our proposed model achieves better performance than the other popular WSOD models.

Method	mAP	CorLoc
OICR [23]	37.9	62.1
WSRPN [25]	40.8	65.6
C-MIL [28]	46.7	67.4
MIL-OICR+GAM+REG [31]	46.8	69.5
MIL-PCL+GAM+REG [31]	45.6	68.7
C-MIDN [13]	50.2	71.2
LSTM-CCTC+OICR (Ours)	42.3	66.2
LSTM-CCTC+C-MIDN (Ours)	50.5	72.5
OICR-Ens.+FRCNN [23]	42.5	65.6
WSRPN-Ens.+FRCNN [25]	45.7	69.3
C-MIDN+FRCNN [13]	50.3	73.3
LSTM-CCTC+FRCNN (Ours)	51.8	75.1

Table 3: Result comparison on the PASCAL VOC 2012 *test* set

4.3 Ablation Study

We also conducted ablation experiments on PASCAL VOC 2007 to analyze the impact of different factors on the performance of our proposed network.

Ways of feature map serialization The first factor is ways of feature-map serialization, or more specifically, the number of scan orders used in serializing feature map at Stage 2. As described in Section 3.2, we made a point that applying LSTM to serialized data derived from multiple scanning directions will lead to identification of more (and accurate) critical points falling on or near a target object. This is based on the assumption that the "temporal" and contextual patterns exerted by different objects may be more predominantly shown in data serialized by different scan orders instead of a fixed one. We conducted experiments to apply only one way of serialization (direction in red in Figure 1), and two ways of serialization (directions in both red and black in Figure 1) and all four ways of serialization (Figure 1). The mAP and CorLoc we obtained for the three scenarios are 38.1% mAP and 52.4% CorLoc for scenario 1, 44.6% mAP and 63.1% CorLoc for scenario 2 and 53.1% mAP and 70.0% CorLoc for scenario 3. This result verifies our assumption.

Different proposal generation methods Previous studies [18] argued that the quality of proposals plays an important role in affecting the OD performance. This statement is especially true in a WSOD context when the exact BBOX labels are not available. Here, we compare our region generation method with typical methods such as SS and EB. To make a fair comparison, we integrated these comparing approaches into our learning framework (LSTM-CCTC+C-MIDN), with the replacement of Stage 2 by SS and EB, respectively. To ensure other networks to achieve their best possible performance within a reasonable training time, 2k proposals were generated for each method. Only an average of 200 proposals were generated by our proposed method instead. The results are: 52.6% mAP and 68.7% CorLoc for SS and 49.5% mAP and 66.4% CorLoc by EB. Our method (53.1% mAP and 70.0% CorLoc) clearly outperforms commonly used proposal generation techniques.

Proposal recall Figure 3 shows the proposal recall with ground truth bounding boxes at different IoU levels [25]. According to [4, 18], a high recall is not a necessary condition of high detection mAP. We simply use this result to measure the quantity and quality of the generated proposals by our proposed network. In figure 3, we observe that our recall is not as high as other methods when the IoU level is low. It is because our algorithm highly relies on finding objects in one-time scanning instead of an exhaustive search. Compared to an exhaustive search, our network might not be able to locate as many objects as they do,

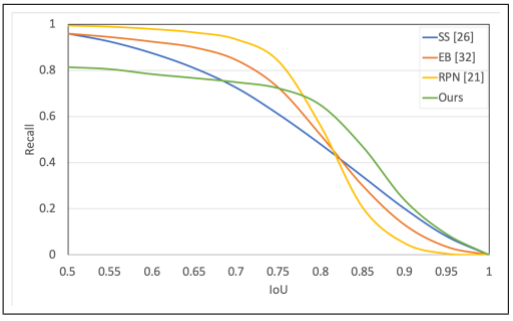

Figure 3: Recall vs. IoU for different proposal methods on the VOC 2007 test set.

however, once an object is located, our network generates a proposal of much better quality. This can be seen in the figure that our curve decays slower than other methods when IoU level gets higher. Our algorithm proposes multiple candidate boxes with different sizes and ratios around an object location such that it is more likely to have a better-quality proposal. The recall curve of RPN [21] is used as a reference here showing the difference between strong supervision (with bounding box information) and weak supervision.

This result proves that our network can correctly label the position of objects and once we have correct aspect ratios of objects, the network generates tight proposals pretty well. One concern of increasing the number of proposals is that it also increases the computation effort. However, our total number of proposals is still far less than traditional methods like SS [26] and EB [32]. Only an average of 200 proposals per image were generated by our proposed method. In addition, unlike other works which generate a fixed number of proposals for each image, our total number of proposals for each image is proportional to the number of interested objects in the image. Our network achieves a good trade-off between computation effort and detection accuracy.

5 Conclusion and future work

This paper introduces a new solution in developing a proposal generation network LSTM-CCTC to achieve high-confidence WSOD. We converted the 2D object recognition problem into a 1D sequential data segmentation problem by leveraging the power of a combined LSTM and CTC network. Specifically, we take advantage of LSTM in its ability of capturing temporal patterns and context dependencies in sequence data, and the CTC in segmenting sequence data without the need of providing frame-wise labels. An improvement is made upon the LSTM-CTC network to create a count-based CTC (CCTC) which will enable weak supervision through a total object count. Multiple data serialization methods are introduced to help more accurately identifying the segmented locations –critical points falling on or near a target object in the original image. Experimental results show that our proposed region generation method has achieved the state-of-the-art performance for WSOD. In the future, we will explore ways to further improve the location accuracy of the identified critical points in the LSTM-CCTC network for better proposal generation.

References

[1] Sean Bell, C Lawrence Zitnick, Kavita Bala, and Ross Girshick. Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2874–2883, 2016.

[2] Hakan Bilen and Andrea Vedaldi. Weakly supervised deep detection networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2846–2854, 2016.

[3] Wonmin Byeon, Thomas M Breuel, Federico Raue, and Marcus Liwicki. Scene labeling with lstm recurrent neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3547–3555, 2015.

[4] Neelima Chavali, Harsh Agrawal, Aroma Mahendru, and Dhruv Batra. Object-proposal evaluation protocol is' gameable'. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 835–844, 2016.

[5] Xueyun Chen, Shiming Xiang, Cheng-Lin Liu, and Chun-Hong Pan. Vehicle detection in satellite images by hybrid deep convolutional neural networks. *IEEE Geoscience and remote sensing letters*, 11(10):1797–1801, 2014.

[6] Ramazan Gokberk Cinbis, Jakob Verbeek, and Cordelia Schmid. Weakly supervised object localization with multi-fold multiple instance learning. *IEEE transactions on pattern analysis and machine intelligence*, 39(1):189–203, 2016.

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[8] Thomas Deselaers, Bogdan Alexe, and Vittorio Ferrari. Weakly supervised localization and learning with generic knowledge. *International journal of computer vision*, 100(3): 275–293, 2012.

[9] Ali Diba, Vivek Sharma, Ali Pazandeh, Hamed Pirsiavash, and Luc Van Gool. Weakly supervised cascaded convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 914–922, 2017.

[10] Dumitru Erhan, Christian Szegedy, Alexander Toshev, and Dragomir Anguelov. Scalable object detection using deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2147–2154, 2014.

[11] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[12] Mingfei Gao, Ang Li, Ruichi Yu, Vlad I Morariu, and Larry S Davis. C-wsl: Count-guided weakly supervised localization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 152–168, 2018.

[13] Yan Gao, Boxiao Liu, Nan Guo, Xiaochun Ye, Fang Wan, Haihang You, and Dongrui Fan. C-midn: Coupled multiple instance detection network with segmentation guidance for weakly supervised object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9834–9843, 2019.

[14] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[15] Alex Graves and Jürgen Schmidhuber. Offline handwriting recognition with multidimensional recurrent neural networks. In *Advances in neural information processing systems*, pages 545–552, 2009.

[16] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376. ACM, 2006.

[17] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[18] Jan Hosang, Rodrigo Benenson, Piotr Dollár, and Bernt Schiele. What makes for effective detection proposals? *IEEE transactions on pattern analysis and machine intelligence*, 38(4):814–830, 2015.

[19] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553): 436, 2015.

[20] Guanbin Li and Yizhou Yu. Deep contrast learning for salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 478–487, 2016.

[21] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[22] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[23] Peng Tang, Xinggang Wang, Xiang Bai, and Wenyu Liu. Multiple instance detection network with online instance classifier refinement. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2843–2851, 2017.

[24] Peng Tang, Xinggang Wang, Song Bai, Wei Shen, Xiang Bai, Wenyu Liu, and Alan Yuille. Pcl: Proposal cluster learning for weakly supervised object detection. *IEEE transactions on pattern analysis and machine intelligence*, 42(1):176–191, 2018.

[25] Peng Tang, Xinggang Wang, Angtian Wang, Yongluan Yan, Wenyu Liu, Junzhou Huang, and Alan Yuille. Weakly supervised region proposal network and object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 352–368, 2018.

[26] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International journal of computer vision*, 104 (2):154–171, 2013.

[27] Francesco Visin, Kyle Kastner, Kyunghyun Cho, Matteo Matteucci, Aaron Courville, and Yoshua Bengio. Renet: A recurrent neural network based alternative to convolutional networks. *arXiv preprint arXiv:1505.00393*, 2015.

[28] Fang Wan, Chang Liu, Wei Ke, Xiangyang Ji, Jianbin Jiao, and Qixiang Ye. C-mil: Continuation multiple instance learning for weakly supervised object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2199–2208, 2019.

[29] Chong Wang, Weiqiang Ren, Kaiqi Huang, and Tieniu Tan. Weakly supervised object localization with latent category learning. In *European Conference on Computer Vision*, pages 431–445. Springer, 2014.

[30] SHI Xingjian, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *Advances in neural information processing systems*, pages 802–810, 2015.

[31] Ke Yang, Dongsheng Li, and Yong Dou. Towards precise end-to-end weakly supervised object detection network. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8372–8381, 2019.

[32] C Lawrence Zitnick and Piotr Dollár. Edge boxes: Locating object proposals from edges. In *European conference on computer vision*, pages 391–405. Springer, 2014.