Appendix for "End-to-End Instance Segmentation with Recurrent Attention"

Anonymous CVPR submission

Paper ID 3051

A. Training procedure specification

We used the Adam optimizer [2] with learning rate 0.001 and batch size of 8. The learning rate is multiplied by 0.85 for every 5000 steps of training.

A.1. Scheduled sampling

We denote θ_t as the probability of feeding in ground-truth segmentation that has the greatest overlap with the previous prediction, as opposed to model output. θ_t decays exponentially as training proceeds, and for larger t, the decay occurs later:

$$\theta_t = \min\left(\Gamma_t \exp\left(-\frac{epoch - S}{S_2}\right), 1\right) \quad (1)$$

$$\Gamma_t = 1 + \log(1 + Kt) \quad (2)$$

where $epoch$ is the training index, S, S_2, and K are constants. In the experiments reported here, these values are 10000, 2885, and 3.

B. Evaluation metrics

We include the details of evaluation metrics here. Symmetric best dice (SBD) (Eq. 3-5) is used on the CVPPP dataset. Mean (un)weighted coverage (MUCov, MWCov) (Eq. 6-7) is used on the KITTI dataset. Average precision (AP) (Eq. 9) is used on the Cityscapes dataset.

$$\text{DICE}(A, B) = \frac{2|A\hat{B}|}{|A| + |B|} \quad (3)$$

$$\text{BD}(\{A_i\}, B) = \max_i \text{DICE}(A_i, B) \quad (4)$$

$$\text{SBD}(y_i\}, \{y_j^*\}) = \min\Big(\frac{1}{N}\sum_j \text{BD}(\{y_i\}, y_j^*), \frac{1}{M}\sum_i \text{BD}(\{y_j^*\}, y_i)\Big) \quad (5)$$

Table 1: MS-COCO Zebra Results

	MWCov ↑	MUCov ↑	|DiC| ↓	Acc. ↑
detect [1]	-	-	2.56	-
aso-sub [1]	-	-	1.03	-
Ours	69.2	64.2	**0.79**	0.57

$$\text{MUCov}(\{y_i\}, \{y_j^*\}) = \sum_i \frac{1}{N} \max_j \text{IoU}(y_i, y_j^*) \quad (6)$$

$$\text{MWCov}(\{y_i\}, \{y_j^*\}) = \sum_i w_{\text{cov},i} \max_j \text{IoU}(y_i, y_j^*) \quad (7)$$

$$w_{\text{cov},i} = \frac{|y_i|}{\sum_i |y_i|} \quad (8)$$

$$\text{AP}(\{y_i\}, \{y_j^*\}) = \max_s \sum_\theta \sum_j Pr(y_{s(i)}, y_j) \cdot \mathbb{1}[\text{IoU}(y_{s(i)}, y_j^*) \geq \theta], \quad (9)$$

C. More experimental results

We include the segmentation and counting performance on the MS-COCO zebra images in Table 1. In terms of counting, our model out-performs a baseline method that runs an object detector and then non-maximal suppression, and a new associative-subitizing method [1].

D. Model architecture

D.1. Foreground + Orientation FCN

We resize the image to uniform size. For CVPPP and MS-COCO dataset, we adopt a uniform size of 224×224, for KITTI, we adopt 128×448, and for Cityscapes 256×512 (4x downsampling). Table 2 lists the specification of all layers.

Table 2: FCN specification

Name	Type	Input	Spec (size/stride)	Size CVPPP/MS-COCO	Size KITTI	Size Cityscapes
input	input	-	-	$224 \times 224 \times 3$	$128 \times 448 \times 3$	$256 \times 512 \times 3$
conv1-1	conv	input	$3 \times 3 \times 3 \times 32$	$224 \times 224 \times 32$	$128 \times 448 \times 64$	$256 \times 512 \times 64$
conv1-2	conv	conv1-1	$3 \times 3 \times 32 \times 64$	$224 \times 224 \times 64$	$128 \times 448 \times 32$	$256 \times 512 \times 32$
pool1	pool	conv1-2	max 2×2	$112 \times 112 \times 64$	$64 \times 224 \times 64$	$128 \times 256 \times 64$
conv2-1	conv	pool1	$3 \times 3 \times 64 \times 64$	$112 \times 112 \times 64$	$64 \times 224 \times 64$	$128 \times 256 \times 64$
conv2-2	conv	conv2-1	$3 \times 3 \times 64 \times 96$	$112 \times 112 \times 96$	$64 \times 224 \times 96$	$128 \times 256 \times 96$
pool2	pool	conv2-2	max 2×2	$56 \times 56 \times 96$	$32 \times 112 \times 96$	$64 \times 128 \times 96$
conv3-1	conv	pool2	$3 \times 3 \times 96 \times 96$	$56 \times 56 \times 96$	$32 \times 112 \times 96$	$64 \times 128 \times 96$
conv3-2	conv	conv3-1	$3 \times 3 \times 96 \times 128$	$56 \times 56 \times 128$	$32 \times 112 \times 128$	$64 \times 128 \times 128$
pool3	pool	conv3-2	max 2×2	$28 \times 28 \times 128$	$16 \times 56 \times 128$	$32 \times 64 \times 128$
conv4-1	conv	pool3	$3 \times 3 \times 128 \times 128$	$28 \times 28 \times 128$	$16 \times 56 \times 128$	$32 \times 64 \times 128$
conv4-2	conv	conv4-1	$3 \times 3 \times 128 \times 128$	$28 \times 28 \times 128$	$16 \times 56 \times 128$	$32 \times 64 \times 128$
conv4-3	conv	conv4-2	$3 \times 3 \times 128 \times 128$	$28 \times 28 \times 128$	$16 \times 56 \times 128$	$32 \times 64 \times 128$
conv4-4	conv	conv4-3	$3 \times 3 \times 128 \times 128$	$28 \times 28 \times 128$	$16 \times 56 \times 128$	$32 \times 64 \times 128$
conv4-5	conv	conv4-4	$3 \times 3 \times 128 \times 128$	$28 \times 28 \times 128$	$16 \times 56 \times 128$	$32 \times 64 \times 128$
conv4-6	conv	conv4-5	$3 \times 3 \times 128 \times 128$	$28 \times 28 \times 128$	$16 \times 56 \times 128$	$32 \times 64 \times 128$
conv4-7	conv	conv4-6	$3 \times 3 \times 128 \times 128$	$28 \times 28 \times 128$	$16 \times 56 \times 128$	$32 \times 64 \times 128$
conv4-8	conv	conv4-7	$3 \times 3 \times 128 \times 256$	$28 \times 28 \times 256$	$16 \times 56 \times 256$	$32 \times 64 \times 256$
pool4	pool	conv4-8	max 2×2	$14 \times 14 \times 256$	$8 \times 28 \times 256$	$16 \times 32 \times 256$
conv5-1	conv	pool4	$3 \times 3 \times 256 \times 256$	$14 \times 14 \times 256$	$8 \times 28 \times 256$	$16 \times 32 \times 256$
conv5-2	conv	conv5-1	$3 \times 3 \times 256 \times 256$	$14 \times 14 \times 256$	$8 \times 28 \times 256$	$16 \times 32 \times 256$
conv5-3	conv	conv5-2	$3 \times 3 \times 256 \times 256$	$14 \times 14 \times 256$	$8 \times 28 \times 256$	$16 \times 32 \times 256$
conv5-4	conv	conv5-3	$3 \times 3 \times 256 \times 512$	$14 \times 14 \times 512$	$8 \times 28 \times 512$	$16 \times 32 \times 512$
pool5	pool	conv5-4	max 2×2	$7 \times 7 \times 512$	$4 \times 14 \times 512$	$8 \times 16 \times 512$
deconv6-1	deconv	pool5	$3 \times 3 \times 256 \times 512/2$	$14 \times 14 \times 256$	$8 \times 28 \times 256$	$16 \times 32 \times 256$
deconv6-2	deconv	deconv6-1 + conv5-3	$3 \times 3 \times 256 \times 512$	$14 \times 14 \times 256$	$8 \times 28 \times 256$	$16 \times 32 \times 256$
deconv7-1	deconv	deconv6-2	$3 \times 3 \times 128 \times 256/2$	$28 \times 28 \times 128$	$16 \times 56 \times 128$	$32 \times 64 \times 128$
deconv7-2	deconv	deconv7-1 + conv4-7	$3 \times 3 \times 128 \times 256$	$28 \times 28 \times 128$	$16 \times 56 \times 128$	$32 \times 64 \times 128$
deconv8-1	deconv	deconv7-2	$3 \times 3 \times 96 \times 128/2$	$56 \times 56 \times 96$	$32 \times 112 \times 96$	$64 \times 128 \times 96$
deconv8-2	deconv	deconv8-1 + conv3-1	$3 \times 3 \times 96 \times 192$	$56 \times 56 \times 96$	$32 \times 112 \times 96$	$64 \times 128 \times 96$
deconv9-1	deconv	deconv8-2	$3 \times 3 \times 64 \times 96/2$	$112 \times 112 \times 64$	$64 \times 224 \times 64$	$128 \times 256 \times 64$
deconv9-2	deconv	deconv9-1	$3 \times 3 \times 64 \times 64$	$112 \times 112 \times 64$	$64 \times 224 \times 64$	$128 \times 256 \times 64$
deconv10-1	deconv	deconv9-2	$3 \times 3 \times 32 \times 64/2$	$224 \times 224 \times 32$	$128 \times 448 \times 32$	$256 \times 512 \times 32$
deconv10-2	deconv	deconv10-1	$3 \times 3 \times 32 \times 32$	$224 \times 224 \times 32$	$128 \times 448 \times 32$	$256 \times 512 \times 32$
deconv10-3	deconv	deconv10-2 + input	$3 \times 3 \times 9 \times 35$	$224 \times 224 \times 9$	$128 \times 448 \times 9$	$256 \times 512 \times 9$

Table 3: External memory specification

Name	Filter spec	Size CVPPP/MS-COCO	Size KITTI	Size Cityscapes
ConvLSTM	3×3	$224 \times 224 \times 9$	$128 \times 448 \times 9$	$256 \times 512 \times 9$

D.2. External memory

D.3. Box network

The box network takes in 9 channels of input directly from the output of the FCN. It goes through a CNN structure again and uses the attention vector predicted by the LSTM to perform dynamic pooling in the last layer. The CNN hyperparameters are listed in Table 4 and the LSTM and glimpse MLP hyperparameters are listed in Table 5. The glimpse MLP takes input from the hidden state of the LSTM and outputs a vector of normalized weighting over all the box CNN feature map spatial grids.

D.4. Segmentation network

The segmentation networks takes in a patch of size 48×48 with multiple channels. The first three channels are the original image R, G, B channels. Then there are 8 channels of orientation angles, and then 1 channel of foreground heat map, all predicted by FCN. Full details are listed in Table 6. Constant β is chosen to be 5.

References

[1] P. Chattopadhyay, R. Vedantam, R. S. Ramprasaath, D. Batra, and D. Parikh. Counting everyday objects in everyday scenes. *CoRR*, abs/1604.03505, 2016. 1

Table 4: Box network CNN specification

Name	Type	Input	Spec (size/stride)	Size CVPPP/MS-COCO	Size KITTI	Size Cityscapes
input	input	-	-	$224 \times 224 \times 9$	$128 \times 448 \times 9$	$256 \times 512 \times 9$
conv1-1	conv	input	$3 \times 3 \times 9 \times 16$	$224 \times 224 \times 16$	$128 \times 448 \times 16$	$256 \times 512 \times 16$
pool1	pool	conv1-2	max 2×2	$112 \times 112 \times 16$	$64 \times 224 \times 16$	$128 \times 256 \times 16$
conv1-2	conv	conv1-1	$3 \times 3 \times 16 \times 16$	$112 \times 112 \times 16$	$64 \times 224 \times 16$	$128 \times 256 \times 16$
pool1	pool	conv1-2	max 2×2	$56 \times 56 \times 16$	$32 \times 112 \times 16$	$64 \times 128 \times 16$
conv2-1	conv	pool1	$3 \times 3 \times 16 \times 32$	$56 \times 56 \times 32$	$32 \times 112 \times 32$	$64 \times 128 \times 32$
conv2-2	conv	conv2-1	$3 \times 3 \times 32 \times 32$	$56 \times 56 \times 32$	$32 \times 112 \times 32$	$64 \times 128 \times 32$
pool2	pool	conv2-2	max 2×2	$28 \times 28 \times 32$	$16 \times 56 \times 32$	$32 \times 64 \times 32$
conv3-1	conv	pool2	$3 \times 3 \times 32 \times 64$	$28 \times 28 \times 64$	$16 \times 56 \times 64$	$32 \times 64 \times 64$
conv3-2	conv	conv3-1	$3 \times 3 \times 64 \times 64$	$28 \times 28 \times 64$	$16 \times 56 \times 64$	$32 \times 64 \times 64$
pool3	pool	conv3-2	max 2×2	$14 \times 14 \times 64$	$8 \times 28 \times 64$	$16 \times 32 \times 64$
conv3-1	conv	pool2	$3 \times 3 \times 64 \times 64$	$14 \times 14 \times 64$	$8 \times 28 \times 64$	$16 \times 32 \times 64$
conv3-2	conv	conv3-1	$3 \times 3 \times 64 \times 64$	$14 \times 14 \times 64$	$8 \times 28 \times 64$	$16 \times 32 \times 64$
pool3	pool	conv3-2	max 2×2	$7 \times 7 \times 64$	$4 \times 14 \times 64$	$8 \times 16 \times 64$

Table 5: Box network LSTM specification

Name	Size CVPPP/MS-COCO	Size KITTI	Size Cityscapes
LSTM	256	256	256
GlimpseMLP1	256	256	256
GlimpseMLP2	7×7	4×14	8×16

[2] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 1

Table 6: Segmentation network specification

Name	Type	Input	Spec (size/stride)	Size
input	input	-	-	$48 \times 48 \times 13$
conv1-1	conv	input	$3 \times 3 \times 13 \times 16$	$48 \times 48 \times 16$
conv1-2	conv	conv1-1	$3 \times 3 \times 16 \times 32$	$48 \times 48 \times 32$
pool1	pool	conv1-2	max 2×2	$24 \times 24 \times 32$
conv2-1	conv	pool1	$3 \times 3 \times 32 \times 32$	$24 \times 24 \times 32$
conv2-2	conv	conv2-1	$3 \times 3 \times 32 \times 64$	$24 \times 24 \times 64$
pool3	pool	conv2-2	max 2×2	$12 \times 12 \times 64$
conv3-1	conv	pool2	$3 \times 3 \times 64 \times 64$	$12 \times 12 \times 64$
conv3-2	conv	conv3-1	$3 \times 3 \times 64 \times 96$	$12 \times 12 \times 96$
pool3	pool	conv3-2	max 2×2	$6 \times 6 \times 96$
deconv4-1	deconv	pool3	$3 \times 3 \times 64 \times 96/2$	$12 \times 12 \times 64$
deconv4-2	deconv	deconv4-1 + conv3-1	$3 \times 3 \times 64 \times 128$	$12 \times 12 \times 64$
deconv5-1	deconv	deconv4-2 + conv2-2	$3 \times 3 \times 32 \times 128/2$	$24 \times 24 \times 32$
deconv5-2	deconv	deconv5-1 + conv2-1	$3 \times 3 \times 32 \times 64$	$24 \times 24 \times 32$
deconv6-1	deconv	deconv5-2 + conv1-2	$3 \times 3 \times 16 \times 64/2$	$48 \times 48 \times 16$
deconv6-2	deconv	deconv6-1 + conv1-1	$3 \times 3 \times 16 \times 32$	$48 \times 48 \times 16$
deconv6-3	deconv	deconv6-2 + input	$3 \times 3 \times 1 \times 29$	$48 \times 48 \times 1$