R-C3D: Region Convolutional 3D Network for Temporal Activity Detection Supplementary Material

Huijuan Xu Abir Das Kate Saenko
Boston University
Boston, MA
{hxu, dasabir, saenko}@bu.edu

Overview

This supplementary material contains the following:

- Additional results with different settings of anchor scales for THUMOS'14 [1] and Charades datsets [4] are given. Table 1 shows the statistics for three sets of anchor scales and the corresponding activity detection results in terms of mAP at different thresholds α. We take the one-way buffer setting in these experiments and everything else is kept the same as the experiments in the main paper except the anchor scales. The first set of anchor scales is used in table 1 of the main paper, and contains 10 different anchor scales with the longest anchor segment duration of 5.12 seconds which covers around 78.13% of all the ground truth activities in the training data. We try two additional sets of anchor scales with longer maximum anchor segment duration of 17.92 seconds and coverage of over 99% of all the ground truth activities in training data. One set of anchor scales has gradually increasing steps and the other has uniform steps of length two. As seen in the table, the mAP for all these sets are close. However, compared to the first set of anchor scales, these two sets have slightly improved mAP at all the five thresholds. In general the improvement at threshold 0.5 is less than 1% showing our model's robustness to this hyperparameter. A possible reason of the slight improvement with these anchor scales could be that these scales cover more ground truth activities, resulting in better detection performance. However, this marginal improvement comes at the cost of more computation as number of classification and regression tasks in both the subnets increase. Table 2 shows the mAP on 25 equidistant frames as introduced in [3] before and after post-processing. The last column of the table shows mAP@0.5 for two different sets of anchor scales on Charades. The first set of anchor scales is used in table 4 of the main paper. Both sets of scales have a high coverage of ground truth activities over 99%, and the detection performance is quite close. Similar to THUMOS'14 dataset, it shows that our model R-C3D is robust to the anchor scales when the maximum anchor segment covers most ground truth activities in training data.

- The detailed configuration of R-C3D is presented in Table 3. For each filter, the kernel size, stride and output sizes are shown. The input video length is denoted by the letter L. The letters n, l, h, w and c denote the number of input channels, temporal length, height, width and the number of output channels of each filter respectively. For all the layers, padding of size $1\times1\times1$ is used. K is the number of anchor scales while P denotes the number of activity proposals retained after the NMS step. C denotes the number of activity classes.

- Finally, we provide some screenshots with analysis of the qualitative results. Two videos containing visual results for several test videos in THUMOS'14 and validation videos in ActivityNet datasets are provided along with the supplementary material. The visualizations for four videos in THUMOS'14 are contained in "00_THUMOS14_vis.mp4". In addition to the segments detected by R-C3D , we also show the segments detected by SCNN [2], for which the code is available publicly. The groundtruth labels are shown as white text while the predicted action labels by R-C3D and SCNN are shown in blue and green respectively.

  Figure 1 shows five frames each for 3 different test videos from "00_THUMOS14_vis.mp4". Figure 1(a) shows five frames that appear between 1:36 and 1:40 minutes in the video "00_THUMOS14_vis.mp4". It can be seen that R-C3D successfully detects the activity 'CliffDiving' which is not captured by SCNN. Similarly, Figure 1(b) shows five frames from another set of results that appear between 2:06 and 2:10 minutes. It can be seen that R-C3D successfully detects the

activity ‘CleanAndJerk’ while SCNN predicts it to be ‘HammerThrow’. Figure 1(c) shows five frames from another test video which contains the activity ‘BaseballPitch’. These frames appear between 2:06 and 2:10 minutes in the supplementary video. In this case, both R-C3D and SCNN successfully detect the activity ‘BaseballPitch’ with some redundancy.

The visual results for five videos in ActivityNet are provided in “01_ActivityNet_vis.mp4”. Figure 2 shows five frames each for 2 different validation videos from “01_ActivityNet_vis.mp4”. Figure 2(a) shows five frames that appear between 0:26 and 0:30 minutes in the video “01_ActivityNet_vis.mp4”. It can be seen that R-C3D successfully detects the activity ‘Using the balance beam’. Similarly, Figure 2(b) shows five frames that appear between 3:53 and 3:57 minutes in the video “01_ActivityNet_vis.mp4”. R-C3D successfully detects the activity ‘Starting a campfire’.

References

[1] Y.-G. Jiang, J. Liu, A. R. Zamir, G. Toderici, I. Laptev, M. Shah, and R. Sukthankar. THUMOS Challenge: Action Recognition with a Large Number of Classes. http://crcv.ucf.edu/THUMOS14/, 2014. 1

[2] Z. Shou, D. Wang, and S.-F. Chang. Temporal Action Localization in Untrimmed Videos via Multi-stage CNNs. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 1

[3] G. A. Sigurdsson, S. Divvala, A. Farhadi, and A. Gupta. Asynchronous Temporal Fields for Action Recognition. *arXiv preprint arXiv:1612.06371*, 2017. 1, 3

[4] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta. Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding. In *European Conference on Computer Vision*, 2016. 1

Table 1. Activity detection results with different anchor scales on THUMOS'14 (in percentage). mAP at different IoU thresholds α is reported.

anchor scale	statistics			α				
	number	longest	coverage	0.1	0.2	0.3	0.4	0.5
[2,4,5,6,8,9,10,12,14,16]	10	5.12s	78.13	51.6	49.2	42.8	33.4	27.0
[2,4,6,8,10,12,14,16,20,24,28,32,36,40,48,56]	16	17.92s	99.19	51.9	50.4	43.6	35.2	28.0
[2,4,6,8,10,12,14,16,18,20,22,24,26,28,30,32,34, 36,38,40,42,44,46,48,50,52,54,56]	28	17.92s	99.19	51.7	49.0	42.8	34.8	27.6

Table 2. Activity detection results with different anchor scales on Charades (in percentage). We report the results using the same evaluation metric as in [3] and mAP at IoU threshold 0.5.

anchor scale	statistics			mAP		mAP@0.5
	number	longest	coverage	standard	post-process	
[1,2,3,4,5,6,7,8,10,12,14,16,20,24,28,32,40,48]	18	76.8s	99.96	12.4	12.7	9.3
[1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24]	24	38.4s	99.07	12.3	12.6	9.1

(a)

(b)

(c)

Figure 1. Qualitative visualization of the predicted activities in THUMOS'14 from the supplementary video "00_THUMOS14_vis.mp4". In each subfigure five frames from different test videos are shown. (a) The frames appear between 1:36 to 1:40 minutes in the video "00_THUMOS14_vis.mp4". It can be seen that R-C3D successfully detects the activity 'CliffDiving' which is not captured by SCNN. (b) The frames appear between 2:06 to 2:10 minutes in the video "00_THUMOS14_vis.mp4". It can be seen that R-C3D successfully detects the activity 'CleanAndJerk' while SCNN predicts it as 'HammerThrow'. (c) The frames appear between 3:17 to 3:21 minutes in the video "00_THUMOS14_vis.mp4". In this case both R-C3D and SCNN successfully detect the activity 'BaseballPitch' with some redundancy.

Table 3. Detailed configuration of R-C3D . The letters n, l, h, w and c denote the number of input channels, temporal length, height, width and the number of output channels of each filter respectively. K and P denote the number of anchor scales and the number of activity proposals retained after the NMS step. C is the number of activity classes.

input ($3\times L\times 112\times 112$ frame sequence)					
Subnet	Filter Name	Filter Type	Kernel Size ($n\times l\times h\times w/l\times h\times w$)	Kernel Stride ($l\times h\times w$)	Output Size ($c\times l\times h\times w$)
3D ConvNet	conv1a	Convolution	$64\times 3\times 3\times 3$	$1\times 1\times 1$	$64\times L\times 112\times 112$
	relu1a	ReLU			$64\times L\times 112\times 112$
	pool1	MAX Pooling	$1\times 2\times 2$	$1\times 2\times 2$	$64\times L\times 56\times 56$
	conv2a	Convolution	$128\times 3\times 3\times 3$	$1\times 1\times 1$	$128\times L\times 56\times 56$
	relu2a	ReLU			$128\times L\times 56\times 56$
	pool2	MAX Pooling	$2\times 2\times 2$	$2\times 2\times 2$	$128\times L/2\times 28\times 28$
	conv3a	Convolution	$256\times 3\times 3\times 3$	$1\times 1\times 1$	$256\times L/2\times 28\times 28$
	relu3a	ReLU			$256\times L/2\times 28\times 28$
	conv3b	Convolution	$256\times 3\times 3\times 3$	$1\times 1\times 1$	$256\times L/2\times 28\times 28$
	relu3b	ReLU			$256\times L/2\times 28\times 28$
	pool3	MAX Pooling	$2\times 2\times 2$	$2\times 2\times 2$	$256\times L/4\times 14\times 14$
	conv4a	Convolution	$512\times 3\times 3\times 3$	$1\times 1\times 1$	$512\times L/4\times 14\times 14$
	relu4a	ReLU			$512\times L/4\times 14\times 14$
	conv4b	Convolution	$512\times 3\times 3\times 3$	$1\times 1\times 1$	$512\times L/4\times 14\times 14$
	relu4b	ReLU			$512\times L/4\times 14\times 14$
	pool4	MAX Pooling	$2\times 2\times 2$	$2\times 2\times 2$	$512\times L/8\times 7\times 7$
	conv5a	Convolution	$512\times 3\times 3\times 3$	$1\times 1\times 1$	$512\times L/8\times 7\times 7$
	relu5a	ReLU			$512\times L/8\times 7\times 7$
	conv5b	Convolution	$512\times 3\times 3\times 3$	$1\times 1\times 1$	$512\times L/8\times 7\times 7$
	relu5b	ReLU			$512\times L/8\times 7\times 7$
Proposal subnet	tpn_conv	Convolution	$512\times 3\times 3\times 3$	$1\times 1\times 1$	$512\times L/8\times 7\times 7$
	tpn_relu	ReLU			$512\times L/8\times 7\times 7$
	tpn_pool	MAX Pooling	$1\times 7\times 7$	$1\times 1\times 1$	$512\times L/8\times 1\times 1$
	tpn_cls	Convolution	$2K\times 1\times 1\times 1$	$1\times 1\times 1$	$2K\times L/8\times 1\times 1$
	tpn_twin_pred	Convolution	$2K\times 1\times 1\times 1$	$1\times 1\times 1$	$2K\times L/8\times 1\times 1$
	tpn_cls_reshape	Reshape			$2\times KL/8\times 1\times 1$
	tpn_loss_cls	SoftmaxWithLoss			$1\times KL/8\times 1\times 1$
	tpn_loss_twin	SmoothL1Loss			$2K\times L/8\times 1\times 1$
Classification subnet	tpn_cls_prob	Softmax			$2\times KL/8\times 1\times 1$
	tpn_cls_reshape	Reshape			$2K\times L/8\times 1\times 1$
	proposal	Python NMS			$P\times 3$
	roi_pool5	ROIPooling			$P\times 512\times 1\times 4\times 4$
	fc6	InnerProduct	num_output: 4096		$P\times 4096$
	relu6	ReLU			$P\times 4096$
	drop6	Dropout	dropout_ratio: 0.5		$P\times 4096$
	fc7	InnerProduct	num_output: 4096		$P\times 4096$
	relu7	ReLU			$P\times 4096$
	drop7	Dropout	dropout_ratio: 0.5		$P\times 4096$
	cls_score	InnerProduct	num_output: (C+1)		$P\times (C+1)$
	twin_pred	InnerProduct	num_output: 2(C+1)		$P\times 2(C+1)$
	loss_cls	SoftmaxWithLoss			$P\times 1$
	loss_twin	SmoothL1Loss			$P\times 2(C+1)$

(a)

(b)

Figure 2. Qualitative visualization of the predicted activities in ActivityNet from the supplementary video "01_ActivityNet_vis.mp4". In each subfigure five frames from different test videos are shown. (a) The frames appear between 0:26 and 0:30 minutes in the video. It can be seen that R-C3D successfully detects the activity 'Using the balance beam'. (b) The frames appear between 3:53 and 3:57 minutes in the video "01_ActivityNet_vis.mp4". The R-C3D successfully detects the activity 'Starting a campfire' in it.