Supplementary Material

Kyle Min Jason J. Corso
University of Michigan
Ann Arbor, MI 48109
{kylemin,jjcorso}@umich.edu

A. Clarification of the architecture

The input-output sizes of our proposed architecture are summarized in Table 1. As mentioned in the paper, we use the same convolutional blocks of the I3D (`Mixed_5b-c`) and add three convolutional layers (kernel size=[(1,3,3), (1,3,3), (1,1,1)], stride=[(1,1,1), (1,1,1), (1,1,1)]) on top of it to model the gaze distribution $q_\phi(\mathbf{z}|\mathbf{x})$. The details of each convolutional block are described in the original I3D paper [1].

Type	$p_\theta(\mathbf{y}\|\mathbf{x},\mathbf{z})$	$q_\phi(\mathbf{z}\|\mathbf{x})$	Input size	Output size
Convolution	$7 \times 7 \times 7$, 64, stride (2, 2, 2)	-	$3(2) \times 24 \times 224 \times 224$	$64 \times 12 \times 112 \times 112$
MaxPooling (2a)	$1 \times 3 \times 3$, stride (1, 2, 2)	-	$64 \times 12 \times 112 \times 112$	$64 \times 12 \times 56 \times 56$
Convolution	$1 \times 1 \times 1$, 64, stride (1, 1, 1)	-	$64 \times 12 \times 56 \times 56$	$64 \times 12 \times 56 \times 56$
Convolution	$3 \times 3 \times 3$, 192, stride (1, 1, 1)	-	$64 \times 12 \times 56 \times 56$	$192 \times 12 \times 56 \times 56$
MaxPooling (3a)	$1 \times 3 \times 3$, stride (1, 2, 2)	-	$192 \times 12 \times 56 \times 56$	$192 \times 12 \times 28 \times 28$
Inception	`Mixed_3b`, 256, stride (1, 1, 1)	-	$192 \times 12 \times 28 \times 28$	$256 \times 12 \times 28 \times 28$
Inception	`Mixed_3c`, 480, stride (1, 1, 1)	-	$256 \times 12 \times 28 \times 28$	$480 \times 12 \times 28 \times 28$
MaxPooling (4a)	$3 \times 3 \times 3$, stride (2, 2, 2)	-	$480 \times 12 \times 28 \times 28$	$480 \times 6 \times 14 \times 14$
Inception	`Mixed_4b`, 512, stride (1, 1, 1)	-	$480 \times 6 \times 14 \times 14$	$512 \times 6 \times 14 \times 14$
Inception	`Mixed_4c`, 512, stride (1, 1, 1)	-	$512 \times 6 \times 14 \times 14$	$512 \times 6 \times 14 \times 14$
Inception	`Mixed_4d`, 512, stride (1, 1, 1)	-	$512 \times 6 \times 14 \times 14$	$512 \times 6 \times 14 \times 14$
Inception	`Mixed_4e`, 528, stride (1, 1, 1)	-	$512 \times 6 \times 14 \times 14$	$528 \times 6 \times 14 \times 14$
Inception	`Mixed_4f`, 832, stride (1, 1, 1)	-	$528 \times 6 \times 14 \times 14$	$832 \times 6 \times 14 \times 14$
MaxPooling (5a)	$2 \times 2 \times 2$, stride (2, 2, 2)	-	$832 \times 6 \times 14 \times 14$	$832 \times 3 \times 7 \times 7$
Inception	`Mixed_5b`, 832, stride (1, 1, 1)	`Mixed_5b'`, 832, stride (1, 1, 1)	$832 \times 3 \times 7 \times 7$	$832 \times 3 \times 7 \times 7$
Inception	`Mixed_5c`, 1024, stride (1, 1, 1)	`Mixed_5c'`, 1024, stride (1, 1, 1)	$832 \times 3 \times 7 \times 7$	$1024 \times 3 \times 7 \times 7$
Convolution	-	$1 \times 3 \times 3$, 256, stride (1, 1, 1)	$1024 \times 3 \times 7 \times 7$	$256 \times 3 \times 7 \times 7$
Convolution	-	$1 \times 3 \times 3$, 64, stride (1, 1, 1)	$256 \times 3 \times 7 \times 7$	$64 \times 3 \times 7 \times 7$
Convolution	-	$1 \times 1 \times 1$, 1, stride (1, 1, 1)	$64 \times 3 \times 7 \times 7$	$1 \times 3 \times 7 \times 7$
Activation	-	Sigmoid function	$1 \times 3 \times 7 \times 7$	$1 \times 3 \times 7 \times 7$
Fully-connected	-	3*7*7, 3*7*7	$1 \times 3 \times 7 \times 7$	$1 \times 3 \times 7 \times 7$
Avg-pooling	$3 \times 7 \times 7$, stride (1, 1, 1)	-	$1024 \times 3 \times 7 \times 7$	$1024 \times 1 \times 1 \times 1$
Fully-connected	1024, #Classes	-	$1024 \times 1 \times 1 \times 1$	#Classes $\times 1 \times 1 \times 1$
(#Params, FLOPs)	(24.7M, 80.2G)	(7.2M, 1.1G)	-	-

Table 1: Detailed input-output sizes of our proposed network architecture. The input of the first convolutional layer has 3 channels for the RGB stream and 2 for the optical flow stream. Batch normalization [2] and ReLU [4] follow after each convolutional layer except the last one. The number of activity classes is 106 for the EGTEA dataset [3].

References

[1] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 4724–4733. IEEE, 2017.

[2] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[3] Yin Li, Miao Liu, and James M Rehg. In the eye of beholder: Joint learning of gaze and actions in first person video. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 619–635, 2018.

[4] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.