Supplementary Material: Self-supervised Multi-actor Social Activity Understanding in Streaming Videos

Shubham Trehan[1] and Sathyanarayanan N. Aakur[1]

CSSE Department, Auburn University, Auburn, AL, 36849
{szt0113,san0028}@auburn.edu

Implementation Details

We use a DETR [1] model, pre-trained on MS COCO [3] as our ROI predictor. The CNN backbone is ResNet-50 [2]. Frames are resized to 224×224 before visual processing with DETR. The size of the global scene features F_t is $7 \times 7 \times 2048$. In Equation 1, $H = W = 7$. During training, all detections, regardless of class with a confidence score of more than 0.1, are taken as object candidates. The top $K = 25$ attention slots from the predictive learning error are used to select actors. We use a 2-layer LSTM network as our spatio-temporal predictor, defined in Section 3.2. The hidden size of each LSTM layer is set to 2048. The weights for the GCN layers for both spatial and temporal smoothing are set to 512 and a fully connected layer is used to project the features back to 2048 for multi-actor predictive learning. w_1 and w_2 in Equation 3 are set to 1. λ_1 and λ_2 are set to 1 in Equation 6. K_{OPT} is found to be $3 \times K_{GT}$ using the elbow method based on the intra-cluster variation as the metric. The learning rate for the prediction stack is set to be 1×10^{-4}, and the learning rate for the spatial and temporal smoothing layers is set to be 1×10^{-3} for all experiments. This is set based on grid search between 10^{-5} and 10^{-2}. All models are trained only for 1 epoch in a streaming manner. Training converges in 6 hours on a workstation server with a 64-core AMD ThreadRipper, an RTX5500 ($48GB$ VRAM) and 128GB RAM.

References

1. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: ECCV. pp. 213–229 (2020)
2. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
3. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV. pp. 740–755 (2014)