Adversarial Generative Grammars for Human Activity Prediction – Supplemental materials –

AJ Piergiovanni[1], Anelia Angelova[1], Alexander Toshev[1], and Michael S. Ryoo[1,2][0000−0002−5452−8332]

[1] Robotics at Google
[2] Stony Brook University
{ajpiergi,anelia,toshev,mryoo}@google.com

1 Implementation Details

Activity Prediction For activity prediction, the number of non-terminals ($\mathcal{N}$) was set to 64, the number of terminals ($\mathcal{T}$) was set to the number of classes in the dataset (e.g., 65 in MultiTHUMOS and 157 in Charades). We used 4 rules for each non-terminal (a total of 256 rules). G, f_R, f_T and f_N each used one fully connected layer with sizes matching the desired inputs/outputs. s is implemented as a two sequential temporal convolutional layers with 512 channels, followed by mean-pooling and a fully-connected layer to generate N_0.

3D Pose prediction For 3D pose prediction, the number of non-terminals ($\mathcal{N}$) was set to 1024, the number of terminals ($\mathcal{T}$) was set to 1024, where each terminal has size of 128 (32 joints in 4D quaternion representation). The number of rules was set to 2 per non-terminal (a total of 2048 rules). G was composed of 2 fully connected layers, f_R, f_T and f_N each used three fully connected layers with sizes matching the desired inputs/outputs. s was implemented as a 2-layer GRU using a representation size of 1024, followed by mean-pooling and a fully-connected layer to generate N_0.

1.1 Network Architecture

Here we provide full details on the structure of the networks.

CNN for Starting Non-terminal The function s (from Eq. 4) is implemented using I3D [1]. The input to the network is multiple frames with size 224×224. The number of frames varies based on how many seconds of video is shown to the network before future prediction. This is at least 16 frames and at most 256 frames. This feature is then used as input to the temporal convolution or GRU described above.

Discriminator Architecture The structure of D is relatively simple. We use 3 1D convolutional layers with a kernel size of 5 and a stride of 4. This gives a temporal receptive field size of 84, which captures long temporal durations (at 12fps, this is 7 seconds per-feature). These layers have 128, 256, and then 64 channels. This is followed by mean-pooling to obtain the feature used for binary classification by a fully-connected layer.

We also tried using an RNN for the discriminator, but found it had comparable performance, but was slower during training.

Training Details The model is trained for 5000 iterations using gradient descent with momentum of 0.9 and the initial learning rate set to 0.1. We follow the cosine learning rate decay schedule. Our implementation is in PyTorch and our models were trained on a single V100 GPU.

2 Supplemental results

Table 1 provides results of our approach for future 3D human pose prediction for all activities in the Human3.6M dataset. Figure 1 shows more examples of future predicted 3D pose at different timesteps.

Activity	80ms	160ms	320ms	400ms	560ms	640ms	720ms	1s	2s	3s	4s
Walking	0.25	0.43	0.65	0.75	0.79	0.85	0.92	0.96	1.37	1.34	1.87
Eating	0.2	0.34	0.53	0.67	0.79	0.92	1.01	1.23	1.66	2.01	2.14
Smoking	0.26	0.49	0.92	0.89	0.99	1.01	1.02	1.25	1.95	2.8	3.37
Discussion	0.29	0.65	0.91	1.00	1.23	1.52	1.68	1.93	2.32	2.58	2.65
Directions	0.39	0.59	0.78	0.87	0.99	1.01	1.25	1.46	1.88	2.37	2.19
Greeting	0.52	0.86	1.26	1.45	1.58	1.69	1.72	1.79	2.56	3.08	2.3
Phoning	0.59	1.15	1.51	1.65	1.47	1.71	1.78	1.84	2.63	2.97	3.71
Posing	0.25	0.54	1.19	1.43	1.86	2.10	2.15	2.66	3.46	4.04	4.49
Purchases	0.6	0.85	1.16	1.23	1.58	1.67	1.72	2.4	1.95	2.35	2.63
Sitting	0.39	0.62	1.02	1.17	1.24	1.42	1.48	1.65	2.73	3.09	3.47
SittingDown	0.39	0.75	1.10	1.23	1.35	1.48	1.65	1.88	2.71	3.88	4.81
TakePhoto	0.24	0.5	0.76	0.89	0.95	1.08	1.15	1.24	2.1	2.45	2.72
Waiting	0.31	0.61	1.13	1.37	1.75	1.92	2.12	2.55	2.82	3.18	3.53
WalkingDog	0.54	0.87	1.19	1.35	1.62	1.75	1.82	1.91	2.18	2.83	2.77
WalkTogether	0.25	0.51	0.7	0.74	0.82	0.88	0.91	1.33	1.4	1.62	2.14
Average	0.36	0.65	0.98	1.11	1.27	1.40	1.49	1.74	2.25	2.70	2.98

Table 1: Evaluation of future pose of our appoach for both short-term and long-term prediction horizons for all activities. Human3.6M benchmark.

References

1. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: CVPR. (2017) 1

Fig. 1: Various predicted 3D pose sequences for walking, greeting, taking photos, sitting, posing.