Supplementary: Temporal Coherence based Criteria for Predicting the Future using Deep Multi-stage Generative Adversarial Networks

Anonymous Author(s)
Affiliation
Address
email

A Smoothed Normalized Cross-Correlation Loss (S-NCCL)

In this section, we provide a modification of the Normalized Cross-Correlation Loss (NCCL) presented in section 3 of the paper. This modification assumes that, while comparing two local patches between the previous frame at timestamp $t-1$ and the current frame at timestamp t, majority of the motion similar to both the frames occur around the central pixel of the patches. This assumption makes the system more *robust* to sudden small variation in motion occuring at the boundaries of the local patches.

To accomplish this heuristic in practical terms, a weight function can be learned whose parameters are learned adaptively. This requires learning these parametrs along with those of the multi-stage GAN discussed in sec. 2, which is a non-trivial problem. For the sake of simplicity, we approximate this weight function using a two-dimensional mean-centered Gaussian lowpass filter (2D-GLPF) and experiment with varying amount of standard deviation of the filter. The algorithm for calculating the smoothed normalized cross-correlation score is outlined in algo. A.1. The w_{xy} values are the smoothing weights that we apply while convolving two image patches for calculating the similarity score. We show results obtained using using this version of the NCCL in fig. [INSERT FIGURE]

B Higher Order Pairwise Contrastive Divergence Loss

The Pairwise Contrastive Divergence Loss (PCDL) discussed in sec. 4 of the paper takes into account (dis)similarities between two consecutive frames to separate or bring them closer in the spatio-temporal feature space. This idea can be extended for higher order situations involving three or more consecutive frames.

For $n = 3$, where n is the number of consecutive frames considered, the PCDL can be defined as:

$$\begin{aligned}\mathcal{L}_{3-PCDL} &= \sum_{i=0}^{T} D_\delta(abs(\hat{Y}_i - \hat{Y}_{i+1}), abs(\hat{Y}_{i+1} - \hat{Y}_{i+2}), p_{i,i+1,i+2}) \\ &= \sum_{i=0}^{T} p_{i,i+1,i+2} d(abs(\hat{Y}_i - \hat{Y}_{i+1}), abs(\hat{Y}_{i+1} - \hat{Y}_{i+2})) \\ &\quad + (1 - p_{i,i+1,i+2}) max(0, \delta - d(abs(\hat{Y}_i - \hat{Y}_{i+1}), abs(\hat{Y}_{i+1} - \hat{Y}_{i+2})))\end{aligned} \tag{I}$$

wher, $p_{i,i+1,i+2} = 1$ only if p_i, p_{i+1} and p_{i+2}- all are simultaneously 1, *i.e.*, the discriminator is very sure about the predicted frames that, they are from the original data distribution. All the other symbols bear standard representations defined in the paper.

Submitted to 31st Conference on Neural Information Processing Systems (NIPS 2017). Do not distribute.

Algorithm A.1: Calculation of the smoothed normalized cross-correlation score for finding similarity between a set of predicted frame(s) and a set of ground-truth frame(s).

Input: Ground-truth frames (GT), Predicted frames $(PRED)$, Gaussian filter $(GLPF)$ [Dimension = $h \times h$]
Output: Smoothed Cross-correlation score $(Score_{SNCC})$

Variables:
w_{xy} = entry of the $x-$th row and $y-$th column of GLPF
h = height and width of an image patch
t = current time;
Initialize: $Score_{SNCC} = 0$;
for $t = 1$ *upto* T **do**
  for $i = 0$ *upto* H, $i \leftarrow i + h$ **do**
    for $j = 0$ *upto* H, $j \leftarrow j + h$ **do**
      $P_t \leftarrow extract\_patch(PRED_t, i, j, h)$;
      \\ Extracts a patch from the predicted frame at time t of dimension $h \times h$ starting from the top-left pixel index (i, j);
      $\hat{P}_{t-1} \leftarrow extract\_patch(GT_{t-1}, i-2, j-2, h+4)$;
      \\ Extracts a patch from the ground-truth frame at time $(t-1)$ of dimension $(h+4) \times (h+4)$ starting from the top-left pixel index $(i-2, j-2)$;
      $\mu_{P_t} \leftarrow avg(P_t)$;
      $\mu_{\hat{P}_{t-1}} \leftarrow avg(\hat{P}_{t-1})$;
      $\sigma_{P_t} \leftarrow standard\_deviation(P_t)$;
      $\sigma_{\hat{P}_{t-1}} \leftarrow standard\_deviation(\hat{P}_{t-1})$;
      $Score_{SNCC} \leftarrow Score_{SNCC} + Absolute\big(\sum_{x,y} \frac{w_{xy}(P_t(x,y)-\mu_{P_t})(\hat{P}_{t-1}(x,y)-\mu_{\hat{P}_{t-1}})}{\sigma_{P_t}\sigma_{\hat{P_{t-1}}}}\big)$;
    end
  end
end
$Score_{SNCC} \leftarrow avg(Score_{SNCC})$;

This version of the objective function, in essence, shrinks the distance between the predicted frames occuring in sequence in a temporal neighborhood, thereby increasing their similarity and maintaining the temporal coherency.

C Results on KITTI Dataset

Due to space restrictions, we present the results of training the multi-stage GAN using the proposed objective functions on the KITTI dataset in table I of this supplementary document.

As the videos in this dataset have significant movement in most of the parts of the frames, the results slightly deteriorate from that of UCF-101. In spite of this, the trend is still visible, as the rate of fall in the quality of the frames acrross time is significantly less, thus ensuring the superiority of our method.

Table I: Experimental results on KITTI datset. PCDL refers to the version defined in eqn. 8 in the paper and SNCCL is the smoothed normalized cross correlation (refer to sec. A).

	1st frame prediction scores		2nd frame prediction scores		4th frame prediction scores	
Methods	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Adv + NCCL	31.1	0.89	29.4	0.88	20.1	0.62
Adv + NCCL + PCDL	31.8	0.89	30.2	0.89	20.9	0.62
Adv + NCCL + PCDL + L1	31.9	0.89	30.4	0.89	21.1	0.62
Adv + SNCCL + PCDL + 3-PCDL	32.6	0.90	31.3	0.89	21.8	0.63