IIT Jodhpur

ARTIFICIAL INTELLIGENCE 2

Title: SME-Net: Sparse Motion Estimation for
Parametric Video Prediction through
Reinforcement Learning

REVIEW PAPER

Date: May 1, 2021

Instructor: Debarati Bhunia Chakraborty

Submitted by: Shristy Gupta

Roll Number: M20CS015

Abstract

The paper discusses about classical prediction technique which is parametric overlapped block motion compensation (POBMC) in Reinforcement Learning for video prediction. The traditional model uses large number of motion parameters with artificial regularization.

The authors proposed parametric video prediction on sparse motion-based prediction for video compression. They have used two neural networks under this new version of re-enforcement learning.

Introduction

In the computer vision task video prediction is very challenging part, both for motion dynamics and texture appearance.

Different class of approaches:

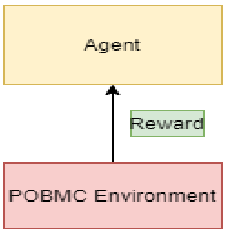
1. Use generative model: It synthesize future frames directly. It uses Long short-term Memory (LSTM). It captures motion dynamics, from past frames to produce future frame. This uses Convolutional neural network CNN. E.g., MCNet, PredNet, BeyondMSE etc. The issue with this approach is blurry synthesis.
2. Estimating the dense motion field: It uses pixel-adaptive kernels. However, this model requires complex models with artificial regularization.
3. Dense Motion Model: This is used to estimate a dense motion field that connects pixel values of future frames with past frames. E.g.: Flow based model, DBF.
4. Sparse Motion Model: Mostly used in video compression, only handful of target pixels are present.

To balance out the advantage of both the model a new model is proposed which is SME-Net aka parametric overlapped block motion compensation (POBMC) for video prediction.

This model uses one-step and multi-step prediction tests to elevate performance.

Discussion

Architecture model:



One-step Prediction:

Critical pixel determination is done as:

$$\{(S_{j,}\, v(s_j))\}^{k}_{j=1}\ j \in In\ and\ estimated\ vectors\ as\ v(s_j)$$

Positioning Network: It produces multi-nominal distribution over the location of the ith important pixel $s_{i.}$

$$S_i \sim P(s|\tau,\widetilde{I^{\iota-1};}\theta)$$

Multi-scale Motion Estimation-Network: It generates estimated motion vector (EMV) v(s_i) for the important pixel S_i

$$V\ (s_i) = M\ (P_c^{64}\ (s_i), P_c^{128}\ (s_i); \theta_m)$$

Parametric Frame Synthesis: This framework performs POBMC to the most recent past frame

$$\hat{I}_n^{(i)}(s) \ = \sum_{j\epsilon N(s)} w * \ I_{n-1}(s + v(s_j)), \forall s \in I_n$$

Target Pixel= critical pixel limit to two nearest neighbours in Euclidean distance

After, K iterations final prediction of the target frame with mean square error is:

$$\tau(\theta_p, \theta_m) \ = \ E(||I_n - \ \hat{I}_n\, ||\tfrac{2}{2})$$

Multi- Step Prediction: Let M be special time span where prediction is made. Then the multiple future frame prediction n to n + M -1. Authors have used one step prediction to reduce the recursive steps. Critical steps are derived from n+M-1 frames to form sliding window single frame of prediction.

Network Architecture: Here Authors have described how multi-scale motion estimation network works

Positioning Network: It uses six-layer convolutional neural network to capture coherence between context frames. The next frame prediction is done by understanding evolution of video frames in the temporal dimension. To mask the critical pixels that were identified before a binary mask is kept.

Multi-Scale Estimation Network: Let v(s) be the motion vector the pixel value I is identified as follows:

$$\hat{I}(s) = \ I_{n-1}(s \ + \ v(s))$$

Two scales of images are provided, one is 64 X 64 and other is 128 X 128. This is done so as both large scale and small-scale motion vectors can be captured. When s + v(s) is not in sampling grid, bilinear interpolation is used.

CONCLUSION

Results and Advantages:

Comparison on Caltech-Ped:

Used KITTI dataset by selecting random 1000 frames of city and roads. Then the FVD for multi-step prediction is:

	Proposed scheme	MC-Net	BMSE	DVF
Caltech-Ped	132	148	846	819

This is 40-50% better than other proposed schemes.

Comparison on UCF101: This is 10% better than existing models like GT, MCNet, BMSE, DVF

Comparison on CIF: Here one step MCNet and PSNR are little better.

Limitations and gaps: The automation of number of proper critical pixel to synthesize future video frames is open

This paper uses POBMC model for video prediction with a smaller number of motion vectors. Hence a smaller number of critical pixels and iterations to create state-of-the-art video prediction.

REFERENCES

Ho, Yung-Han, et al. "Sme-net: Sparse motion estimation for parametric video prediction through reinforcement learning." Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). 2019.

Url: https://openaccess.thecvf.com/content_ICCV_2019/papers/Ho_SME-Net_Sparse_Motion_Estimation_for_Parametric_Video_Prediction_Through_Reinforcement_ICCV_2019_paper.pdf