Conceptualization of a GAN for future frame prediction

Nirvana Pillay[0000-0003-4999-1215] and Edgar Jembere[0000-0003-1776-1925]

University of KwaZulu-Natal, Durban, RSA
nirvanap02@gmail.com, jemberee@ukzn.ac.za

Abstract. The generation of future frames of a video involves the analysis of the previous *t-i* frames and the subsequent prediction of the following *t+j* frames. The majority of state of the art models are able to accurately predict a single future frame that exhibits a high degree of photorealism. The effectiveness of these models at generating quality results decreases as the number of frames generated increases due to the divergence of the solution space. The solution space is now multimodal and optimization of traditional loss functions, such as MSE loss, does not adequately model the multimodality and the resultant frames are blurred. The conceptualization of a GAN that generates several plausible future frames with adequate motion representation and a high degree of photorealism is presented.

Keywords: GANs, Transformation, ConvLSTM.

1 Introduction

The prediction of future frames has several applications in autonomous decision-making areas that include; self-driving cars, social robots and video completion [10]. For example, a SocialGAN [4] determines plausible and socially acceptable walking trajectories of people, thus, aiding in the navigation of human-centric environments. GANs ([1], [4], [5], [6], [9]) have been a popular approach to training spatio-temporal models for future frame prediction. The constituent components of a GAN is a generator and a discriminator, engaged in a minimax game [3]. GANs, however, are difficult to train; and are susceptible to mode collapse. In transformation space, the generator extracts transformations between adjacent input frames. It subsequently predicts a future transformation and applies it to the last frame of the input to generate the next frame and so forth. The source of variability is modelled and, thus, the need to store low level details of the input is eliminated. The resultant model requires fewer parameters which simplifies learning. Furthermore, the spatial data of the input is conserved.

2 Related Work

To model spatio-temporal relationships in video data, networks include either CNNs, RNNs or both. The standard for sequential modelling tasks is RNNs, such as LSTMs, due to its ability to represent long term temporal dependencies. A CNN that exhibits a similar efficacy is the Temporal Convolutional Network (TCN). A TCN in conjunction with a dilated CNN to model temporal and spatial dependencies respectively was implemented by [9]. A similar approach was undertaken by [1], with a PGGAN modelling spatial dependencies instead. Another attempt at sequential modelling utilizing CNNs [8] was an architecture in which a network was replicated through time. The resultant model was a 'peculiar RNN' as parameters were now shared across time whilst still convolving spatial data. A CNN-LSTM architecture was implemented by [6] to predict future frames of synthetic video data. These aspects were later united by [7] into a single network, a convolutional LSTM (ConvLSTM). A stacked ConvLSTM, coupled with a Spatial Transformer Network (STN) [2], addressed the problem of future frame prediction and determined the state of motion of a robot arm. The representation of motion is improved by models that operate in transformation space ([2], [8], [9]). Such a model, a CGAN [9] was evaluated using a Two-Alternative Forced Choice (2AFC) test. The generated video was preferred only 30.6% of the time over its ground-truth counterpart.

3 Proposed Model

In a bid to address the issues of motion representation, photorealism and plausibility of generated frames, this research proposes the implementation of a CGAN. The discriminator of the CGAN receives the context frames coupled with alternatively ground truth future frames or generated future frames and is only deceived by sequences of frames that exhibit plausibility. A mini-batch standard deviation layer is added to one of the last layers of the Progressively Growing Network (PGN) discriminator; aiding in the prevention of mode collapse. The generator comprises of 7 stacked ConvLSTMs, similar to [2], and preserves spatial data whilst modelling the complex dynamics of the data. Hidden Layer5 parameterizes a modified STN and the output of ConvLSTM5 is a predicted affine transformation matrix for each separate 'good feature' in the frame. The STN is modified to determined points by the *Shi-Tomasi Corner Detection* algorithm for which transformations are then predicted. The model also predicts a compositing mask over each transformation. The generated frame is reconstructed by applying predicted affine transformations, merged by masking, to the last input frame.

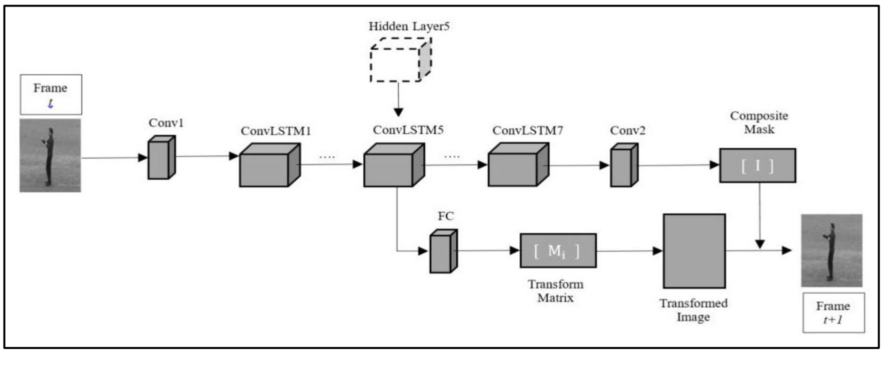

Fig. 1. Schematic of CGAN Generator

References

1. Aigner, S., Körner, M.: "FutureGAN: Anticipating the Future Frames of Video Sequences using Spatio-Temporal 3d Convolutions in Progressively Growing GANs.arXivpreprint," arXiv:1810.01325 (2018).
2. Finn, C., Goodfellow, I., Levine, S.: "Unsupervised Learning for Physical Interaction through Video Prediction.arXivpreprint" arXiv: 1605.07157 (2016).
3. Goodfellow, "NIPS 2016 Tutorial: Generative Adversarial Networks" in 2016
4. Gupta, A., Johnson, J., Fei-Fei, L., Savarese, S., Alahi, A.: "Social GAN: Socially Acceptable Trajectories with Generative Adversarial Networks.arXivpreprint" arXiv:1803.10892 (2018)
5. Lee, X., Zhang, R., Ebert, F., Abbeel, P., Finn, C., Levine, S.: "Stochastic Adversarial Video Prediction.arXivpreprint," arXiv:1804.01523 (2018).
6. Lotter, W., Kreiman, G., Cox, D.: "Unsupervised Learning of Visual Structure using Predictive Generative Networks.arXivpreprint," arXiv:1511.06380 (2016).
7. Shi, X., Chen, Z., Wang, H., Yeung, D.: "Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting.arXivpreprint" arXiv:506.04214 (2015).
8. van Amersfoort, J., Kannan, A., Ranzato, M.A, Szlam, A., Tran, D., Chintala, S.: "Transformation-based Models of Video Sequences.arXivpreprint," arXiv:1701.08435 (2017).
9. Vondrick, C., Torralba, A.: "Generating the Future with Adversarial Transformers," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 2992-3000 (2017).
10. Jai, YT., Hu, SM., Martin, R.: "Video Completion using Tracking and Fragment Merging," *The Visual Compute 21(8-10)*, pp. 601-610 (2005)