Self-supervised Video Prediction

Lab Vision Systems: Learning Computer Vision on GPUs

Schmidt, Heiko
Akter, Mst Mahfuja

Universität Bonn
heiko@uni-bonn.de, Matrikelnummer: 2754786
s6msakte@cs.uni-bonn.de, Matrikelnummer: 3214647

Abstract. To solve self-supervised learning problems including video prediction by using deep learning algorithms is a non trivial task for many years. Video prediction problem requires analyzing the video frames, temporally and spatially, and constructing a model to predict how the environment changes over the time. Fully convolutional neural networks prevent the modeling of location-dependent patterns because of its spatially invariant behaviour. In this paper we described a model to predict future frames from a sequence of frames by using Convolutional Neural Networks in different ways. After that we classified the video's action using MLP. We have found that encoding location-dependent features is crucial for the task of video prediction. Our proposed methods significantly outperform spatially invariant models.

1 Introduction

A wide applications like scene understanding and prediction of future image in computer vision have triggered vast research in recent times. Visual recognition systems performing image classification, detection of objects and predict the future movement of objects have been popular and challenging tasks on this area of research. Significant developments in the area of neural networks have fueled the improved performance of future image prediction which is a sub-domain of visual recognition systems.

The Convolutional neural network(CNN) consists of a series of convolutional layers where the neurons are connected from one layer to the next layer. The main applications of these networks is image recognition, object detection, document analysis, historic and environmental collection, etc. For this project we are using fully convolutional Neural Network, location-dep Convolution Neural Networks and Convolutional Gated Recurrent Networks Unit (ConvGRU). ConvGRU model is a recursive model which consider the sequence of frames and their behavior of movement.

This project aims to predict three future frames from three seed frames from a video. Our implementation is developed to predict the future frames from input frames, classify the actions from the video, and tuned hyper parameters to achieve better results.

Rest of the paper is organized as follows: Section 2 describes the Related Work, Section 3 describes the detail explanation of different model we use in our project, Section 4 describes our model architecture and followed methods, Section 5 describes the outcome of our experiments and the last section 6 describes the concluding remarks.

2 Related Work

Our proposed model is inspired from the Video Ladder Network (VLN) Cricri et al., 2016 and "RoboCup 2019: Robot World Cup XXIII" 2019. VLN is a fully convolutional neural encoder-decoder network that is augmented by both recurrent and feed forward connections at all layers. These connections form a lateral recurrent residual block at each layer, where the feed forward connection works as a skip connection and the recurrent connection represents the residual connection.

On "RoboCup 2019: Robot World Cup XXIII" 2019, the model NimbRoNet2 have used pretrained ResNet18, each convBlock consists of two convolutional layers followed by batch-norm and ReLU activations. Instead of a convolutional layer they have used a location-dependent convolution in the last layer.

The fully convolutional neural network is spatially invariant which prevents to model location-dependent features. To overcome this problem we have introduced a combination of Location-dep convolution Azizi et al., 2018 and ConvGRU Tokmakov et al., 2017 layer in our model. This location dependency convolution layer can encode location features and learn location dependent biases. The convGRU Tokmakov et al., 2017 model extracts moving objects from a video. This model learns spatio-temporal features from intermediate visual representation, hence this part of model has a greatest impact on the future frame prediction in our model.

3 Literature

3.1 Video Ladder Network

Video Ladder Network (VLN) Cricri et al., 2016 is a fully convolutional neural network to generate future video frames efficiently. VLN is an encoder-decoder network where the decoder exploits temporal summaries generated from all layers of the encoder The encoder consists of dilated convolutional layers, whereas the decoder uses normal convolutions. Both encoder and decoder use batch-normalization (BN) and leaky-ReLU activation function followed by a convolutional layer. This way, the model provides fast inference and achieves outstanding results, while having a simple network structure.

3.2 Location Dependency Convolutional Neural Network

To predict video frame, Location dependency on deep convolution network Azizi et al., 2018 introduces a new location-biased convolutional layers to overcome

the spatial invariant behavior of CNN. This model encodes location features of the input in separate channels, and convolutional layers with learnable location-dependent biases. The effectiveness of location-dependent bias is evaluated on different architectures including VLN 3.1 model. This model significantly outperformed spatially invariant models in video prediction.

3.3 Convolutional Gated Recurrent Networks Unit

The convolutional gated recurrent unit (ConvGRU) Tokmakov et al., 2017 that encodes the spatio-temporal evolution of objects in the input video sequence. This model extract moving objects from the input image sequences by learning patterns during the forward pass. ConvGRU model can learn from a small number of input video sequences. It takes a video frame as input, assigns each pixel an object or background label based on the learned spatio-temporal features as well as the "visual memory" specific to the video. The visual memory is implemented with convolutional gated recurrent units, which allows to propagate spatial information over time. This method have been very successful in the future image prediction tasks and proved to produce good results for object localization.

4 Methods

Like the model NimbRoNet2 "RoboCup 2019: Robot World Cup XXIII" 2019, we have used pretrained ResNet18. Each convBlock consists of two convolutional layers followed by batch-norm and ReLU activations. In parallel we have used ConvGRU block which contains a fully convolutional layer and three ConvGRU layer. Each ConvGRU block is a combination of two convolutional layers which is passed into gated units. Instead of a convolutional layer we used a location-dependent convolution in the first two layers. The set of convBlock would feed into convGRU block and finally they would be concatenated with two another convolutional layer which is followed by pixel shuffle and batch-norm ReLU activations respectively.

figure 1 shows the model architecture. To make the model simple, residual connections of ResNet are not depicted in figure.

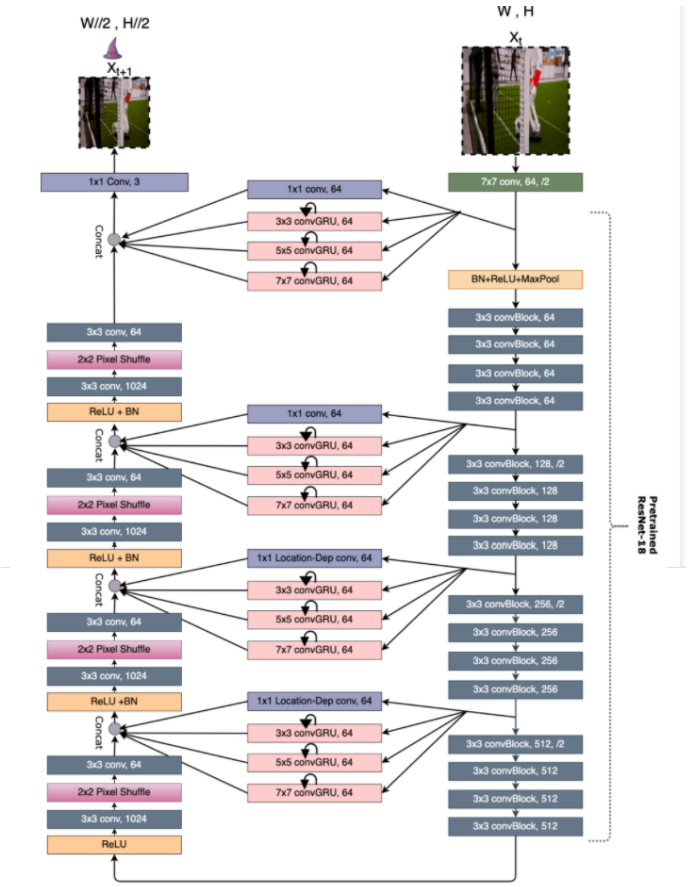

Fig. 1. Self-supervised Video Prediction model. Right: Input image is passed into convBlock. Middle: The outcome from each ConvBlock is passed into ConvGRU block. Right: The outcome from convGRU block is concatenated with two fully convolutional layer and feed into forward.

5 Results

6 Conclusion

References

Azizi, Niloofar et al. (2018). *Location Dependency in Video Prediction*. arXiv: 1810.04937 [cs.CV].

Cricri, Francesco et al. (2016). *Video Ladder Networks*. arXiv: 1612.01756 [cs.LG].

"RoboCup 2019: Robot World Cup XXIII" (2019). In: *Lecture Notes in Computer Science*. ISSN: 1611-3349. DOI: 10.1007/978-3-030-35699-6. URL: http://dx.doi.org/10.1007/978-3-030-35699-6.

Tokmakov, Pavel et al. (2017). *Learning Video Object Segmentation with Visual Memory*. arXiv: 1704.05737 [cs.CV].