Novel CNN Approach for Video Prediction Based on FitVid

Taiju Watanabe
School of FSE
Waseda University
Tokyo, Japan
lvpurin@fuji.waseda.jp

Takahiro Shindo
School of FSE
Waseda University
Tokyo, Japan
taka_s0265@ruri.waseda.jp

Hiroshi Watanabe
Graduate School of FSE
Waseda University
Tokyo, Japan
hiroshi.watanabe@waseda.jp

***Abstract*—Video prediction is a task in computer vision that predicts future frames from the past few frames of video. In video prediction, a simple CNN-based approach called "SimVP" has marked remarkable performance without using RNN or vison transformer (ViT). In this paper, we propose a model structure to improve performance of video prediction based on FitVid. For this goal, we introduce network modules used in SimVP to FitVid. Experimental results show that the proposed structure shows better prediction accuracy compared to SimVP.**

***Keywords*—SimVP, FitVid, RNN, ViT**

I. Introduction

Video prediction is a task in computer vision that predicts future frames from the past few frames of video. Video prediction is applied to scene understanding and high efficiency video coding. A recent trend of video prediction models is the combinations of convolutional and recurrent neural networks. SimVP [1] marks state-of-the-art accuracy on several video prediction datasets, despite using only a simple structure of convolutional layers. In this paper, we propose to further improve video prediction accuracy by introducing several modules used in SimVP based on FitVid [2].

II. Propsed Method

Video prediction models consists of an encoder, a translator and a decoder. As for the encoder and the decoder, we use similar structures to FitVid. FitVid is a video generation model that generates videos using images and actions. We enhance the performance of our model using skip connection. Moreover, we employ Squeeze-and-Excitation block (SE block) to acquire correlation among channels and intensify the effect of skip connection. As for upsampling in decoder layers, we use the nearest neighbor method instead of transposed convolution which is used in SimVP. Similar to SimVP, we use inception layers for the translator. Model structure of our final model is shown in Fig. 1.

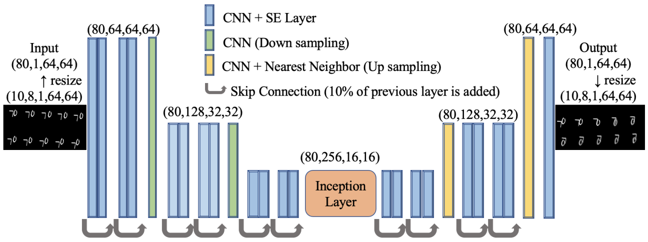

Fig. 1. Model structure of our final model (baseline model with skip connection, SE block and the nearest neighbor method)

TABLE I
Our model vs SimVP. THe optimal results are marked by **bold**

Method	Moving MNIST			TrafficBJ		
	MSE↓	MAE↓	SSIM↑	MSE× 100↓	MAE↓	SSIM↑
SimVP (CVPR 2022)	23.8	68.9	0.948	41.4	**16.2**	**0.982**
Baseline	23.7	71.0	0.948	41.4	**16.2**	**0.982**
Baseline + SC	23.3	73.6	**0.949**	**39.1**	16.3	0.981
Baseline + SC + SE	22.9	70.7	0.946	44.0	16.4	**0.982**
Baseline + SC + SE + NN	**22.5**	**68.3**	**0.949**	39.4	16.6	0.981

III. Experiment

To evaluate the performance of our model, we use two video prediction datasets, Moving MNIST [3] and TrafficBJ [4]. The task is to predict 10 future frames given 10 previous frames for Moving MNIST, 4 frames with 4 frames for TrafficBJ. We compare our model with SimVP on three metrics, MSE, MAE and Structural Similarity Index Measure (SSIM). We train our model for 1200 epochs for Moving MNIST and 80 epochs for TrafficBJ. SimVP is trained for 2000 epochs for Moving MNIST and 80 epochs for TrafficBJ. Result of our experiment is shown in Table I. SC stands for skip connection, SE stands for Squeeze-and-Excitation block and NN stands for the nearest neighbor method. From Table I, our models show better prediction accuracy than SimVP in terms of MSE, MAE and SSIM.

IV. Conclusion

We propose a novel CNN approach for video prediction based on FitVid by incorporating modules used in SimVP. Our model outperforms SimVP in prediction accuracy.

References

[1] Z. Gao, C. Tan, L. Wu and S. Z. Li,"SimVP: Simpler yet Better Video Prediction", CVPR, 2022.

[2] M. Babaeizadeh, M. T. Saffar, S. Nair, S. Levine, C. Finn and D. Erhan, "FitVid: Overfitting in Pixel-Level Video Prediction", ICLR, 2022.

[3] N. Srivastava, E. Mansimov and R. Salakhudinov, "Unsupervised learning of video representations using lstms", ICML, 2015.

[4] J. Zhang, Y. Zheng and D. Qi., "Deep spatio-temporal residual networks for citywide crowd flows prediction", AAAI, 2017.