Video frame Interpolation and extrapolation

Zibo Gong[1], Ziyi Yang[1]
[1]Department of Electrical Engineering, Stanford University
{zibo, zy99}@stanford.edu

Abstract

In our project, we use auto-encoder network and feature fusion model to interpolate and extrapolate video frames. The auto-encoder network is used to extract features from frame 1 and frame 2. Each layer generated by auto-encoder of two input frames are fused by convolutional neural networks, in order to predict object motion and frame shift. Finally, predicted frame is generated by 5X5 convolutional layer.

Problem Statement

Predicting video frames from existing ones involves accurately modelling the content and dynamics of image evolution. Also generating in-betweens frame and subsequent frames can achieve high video frame rates and predict the motions, which is of great commercial prospect. Meanwhile, video frame interpolation is also a quite challenging problem, since it requires accurate modelling of clear motion and clear reconstruction of pixel reconstructions.

Data Sets

We trained and evaluate our networks on UCF101 dataset. We sampled frame triplets from some videos, creating a training set of approximately 30,000 triplets. The pixel values are normalized into the range of [−1, 1]. For vanilla model, we crop the images in our training set to 32*32 images. We use 180,000 small triplets to train the the vanilla model. For both two model, we use the same test set for evaluation. We generate the 100 triplets from unused videos in UCF101 dataset using the same method.

References

[1] Mathieu, M., Couprie, C., & LeCun, Y. (2015). Deep multi- scale video prediction beyond mean square error. arXiv preprint arXiv:1511.05440.

[2] Liu, Z., Yeh, R., Tang, X., Liu, Y., & Agarwala, A. (2017). Video Frame Synthesis using Deep Voxel Flow.arXiv preprint arXiv:1702.02463.

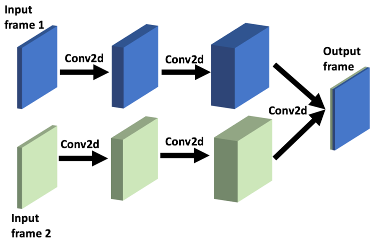
[3] Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A. P., Bishop, R., ... & Wang, Z. (2016). Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1874-1883).

[4] Niklaus, S., Mai, L., & Liu, F. (2017). Video Frame Interpolation via Adaptive Convolution. arXiv preprint arXiv:1703.07514.

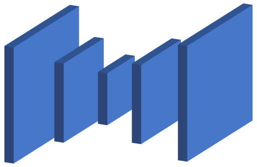
[5] Meyer, S., Wang, O., Zimmer, H., Grosse, M., & Sorkine- Hornung, A. (2015). Phase-based frame interpolation for video. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1410-1418).

Model

Vanilla Model

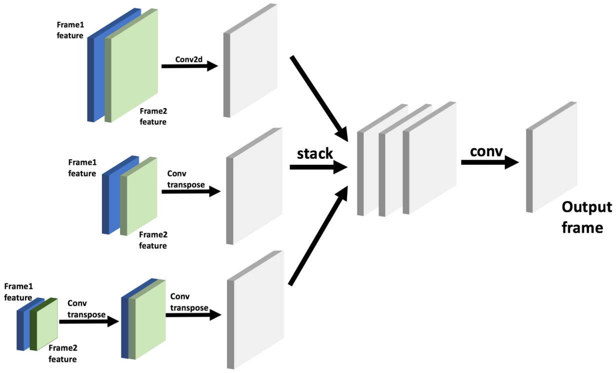


Refined Model: Feature Extraction



Pre-train an auto-encoder network to extract features from the first input frame and the second one.

Refined Model: Prediction

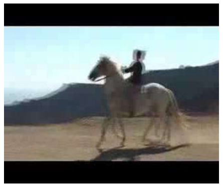


Using features extracted from two input frames by auto-encoder, we predict the frame by convolutional network

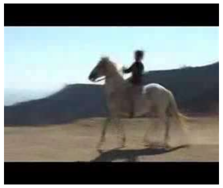
Evaluation and Results

Frame1

True Frame2



Beyond MSE



Ours

Frame3

UCF-101 Interpolation evaluation

Method	Vanilla Model	Refined Model	Beyond MSE	Deep Voxel Flow
PSNR	26.2	29.8	28.8	30.9
SSIM	0.84	0.92	0.90	0.94

Future Work

We plan to train model on frame extrapolation in the future. Also we have other options on loss function. What's more, the data example selection can be improved, which means we can choose input frames with significant differences and obvious motions.