Video Frame Prediction

Swapnil Sharma*
swapnil.sh@nyu.edu

Nikita Anand
email

Rishika Pabba
email

Abstract

Attempting Video frame prediction task. Things tried:

1. Semantic Segmentation Mask using ConvLSTM
2. Tried Segformer model for sematic segmentation mask
3. Using AutoEncoderDecoder for Semantic segmentation mask
4. Training convLSTM for video frame prediction in auto regressive manner
5. Tried convLSTM for video frame prediction in non auto regressive manner by predicting all frames at once
6. Tried convBiLSTM for video frame prediction in non auto regressive manner by predicting all frames at once
7. Implemented DataParallel for multiGPU training
8. Normalized images before feeding to convLSTM
9. Added skip connections in AutoEncoderDecoder for semantic segmentation mask
10. Using ResNet AutoEncoderDecoder for semantic segmentation mask
11. Using VPTR non auto regressive model for video frame prediction

Best results for semantic segmentation mask were obtained using ResNet AutoEncoders. Able to predict video frames using VPTR in auto regressive manner.

1 Introduction

This document describes the methodology and result analysis for the final Project in Deep Leanring Course. The task is to predict the semantic segmentation mask of 22^{nd} frame given initial 11 frames of video.

These videos have simple 3D shapes that interact with each other according to basic physics principles. Objects in videos have three shapes (cube, sphere, and cylinder), two materials (metal and rubber), and eight colors (gray, red, blue, green, brown, cyan, purple, and yellow). In each video, there is no identical objects, such that each combination of the three attributes uniquely identifies one object.

We tried breaking down the problem in semantic segmentation mask translation and future frame prediction tasks. To solve the first problem we tried using ConvLSTM, Segformer, and ResNet AutoEncoders. We got the best performance with ResNet AutoEncoders.

For the second task we tried using ConvLSTM, convBiLSTM and non autoregressive VPTR. Although we were not able to converge VPTR model on the entire dataset we got best performance on a subset with VPTR model.

2 Related Work

Most state-of-the-art (SOTA) models for video frame prediction use ConvLSTM-based AutoEncoders. These models were initially developed for predicting precipitation nowcasting, as introduced by Shi

*https://swappysh.github.io

Preprint. Under review.

et al. [2015]. They have been utilized in various studies, including Finn et al. [2016], Lotter et al. [2016], Xu et al. [2016], Ballas et al. [2015]. According to Jing and Tian [2019], these models also work with self-supervised tasks.

Although the ConvLSTM-based models are flexible and efficient, they are generally slow due to recurrent prediction. To address this issue, standard CNNs or 3D CNNs and VAE-based methods have been proposed, such as those by Mathieu et al. [2015] and Babaeizadeh et al. [2017].

State-of-the-art models commonly rely on complex ConvLSTM models integrating attention mechanisms or memory-augmented modules. For instance, the Long-term Motion Context Memory model by Lee et al. [2021] stores long-term motion context through a novel memory alignment learning, and the motion information is recalled during the test to facilitate long-term prediction. Chang [2021] proposed an attention-based motion-aware unit to increase the temporal receptive field of RNNs.

Almost all the state-of-the-art (SOTA) VFFP models are based on ConvLSTMs, i.e. convolutional short-term memory networks,which are efficient and powerful. Nevertheless as per Ye and Bilodeau [2022], they suffer from some inherent problems of recurrent neural networks(RNNs), such as slow training and inference speed, error accumulation during inference, gradient vanishing, and predicted frames quality degradation. Researchers keep improving the performance by developing more and more sophisticated ConvLSTM-based models.

With the introduction of transformers, they have also been applied in the Vision domain, including video frame prediction. The ConvTransformer model by Liu et al. [2020] follows the architecture of DETR introduced in Meinhardt et al. [2021], a classical neural machine translation (NMT) Transformer architecture. DETR also inspired the development of the VPTR-NAR model by Ye and Bilodeau [2022], a non-autoregressive model for video frame prediction.

3 Methodology

We started with trying to predict the semantic segmentation mask of the frames given Normalized images of the frames. We tried using ConvLSTM, Segformer and ResNet AutoEncoders

4 Results

5 Submission of papers to NeurIPS 2022

Please read the instructions below carefully and follow them faithfully.

5.1 Style

Papers to be submitted to NeurIPS 2022 must be prepared according to the instructions presented here. Papers may only be up to **nine** pages long, including figures. Additional pages *containing only acknowledgments and references* are allowed. Papers that exceed the page limit will not be reviewed, or in any other way considered for presentation at the conference.

The margins in 2022 are the same as those in 2007, which allow for $\sim$15% more words in the paper compared to earlier years.

Authors are required to use the NeurIPS LaTeX style files obtainable at the NeurIPS website as indicated below. Please make sure you use the current files and not previous versions. Tweaking the style files may be grounds for rejection.

5.2 Retrieval of style files

The style files for NeurIPS and other conference information are available on the World Wide Web at

`http://www.neurips.cc/`

The file `neurips_2022.pdf` contains these instructions and illustrates the various formatting requirements your NeurIPS paper must satisfy.

The only supported style file for NeurIPS 2022 is `neurips_2022.sty`, rewritten for LaTeX 2_ε. **Previous style files for LaTeX 2.09, Microsoft Word, and RTF are no longer supported!**

The LaTeX style file contains three optional arguments: `final`, which creates a camera-ready copy, `preprint`, which creates a preprint for submission to, e.g., arXiv, and `nonatbib`, which will not load the `natbib` package for you in case of package clash.

Preprint option If you wish to post a preprint of your work online, e.g., on arXiv, using the NeurIPS style, please use the `preprint` option. This will create a nonanonymized version of your work with the text "Preprint. Work in progress." in the footer. This version may be distributed as you see fit. Please **do not** use the `final` option, which should **only** be used for papers accepted to NeurIPS.

At submission time, please omit the `final` and `preprint` options. This will anonymize your submission and add line numbers to aid review. Please do *not* refer to these line numbers in your paper as they will be removed during generation of camera-ready copies.

The file `neurips_2022.tex` may be used as a "shell" for writing your paper. All you have to do is replace the author, title, abstract, and text of the paper with your own.

The formatting instructions contained in these style files are summarized in Sections 6, 7, and 8 below.

6 General formatting instructions

The text must be confined within a rectangle 5.5 inches (33 picas) wide and 9 inches (54 picas) long. The left margin is 1.5 inch (9 picas). Use 10 point type with a vertical spacing (leading) of 11 points. Times New Roman is the preferred typeface throughout, and will be selected for you by default. Paragraphs are separated by ½ line space (5.5 points), with no indentation.

The paper title should be 17 point, initial caps/lower case, bold, centered between two horizontal rules. The top rule should be 4 points thick and the bottom rule should be 1 point thick. Allow ¼ inch space above and below the title to rules. All pages should start at 1 inch (6 picas) from the top of the page.

For the final version, authors' names are set in boldface, and each name is centered above the corresponding address. The lead author's name is to be listed first (left-most), and the co-authors' names (if different address) are set to follow. If there is only one co-author, list both author and co-author side by side.

Please pay special attention to the instructions in Section 8 regarding figures, tables, acknowledgments, and references.

7 Headings: first level

All headings should be lower case (except for first word and proper nouns), flush left, and bold.

First-level headings should be in 12-point type.

7.1 Headings: second level

Second-level headings should be in 10-point type.

7.1.1 Headings: third level

Third-level headings should be in 10-point type.

Paragraphs There is also a `\paragraph` command available, which sets the heading in bold, flush left, and inline with the text, with the heading followed by 1 em of space.

8 Citations, figures, tables, references

These instructions apply to everyone.

8.1 Citations within the text

The `natbib` package will be loaded for you by default. Citations may be author/year or numeric, as long as you maintain internal consistency. As to the format of the references themselves, any style is acceptable as long as it is used consistently.

The documentation for `natbib` may be found at

`http://mirrors.ctan.org/macros/latex/contrib/natbib/natnotes.pdf`

Of note is the command `\citet`, which produces citations appropriate for use in inline text. For example,

```
\citet{hasselmo} investigated\dots
```

produces

> Hasselmo, et al. (1995) investigated...

If you wish to load the `natbib` package with options, you may add the following before loading the `neurips_2022` package:

```
\PassOptionsToPackage{options}{natbib}
```

If `natbib` clashes with another package you load, you can add the optional argument `nonatbib` when loading the style file:

```
\usepackage[nonatbib]{neurips_2022}
```

As submission is double blind, refer to your own published work in the third person. That is, use "In the previous work of Jones et al. [4]," not "In our previous work [4]." If you cite your other papers that are not widely available (e.g., a journal paper under review), use anonymous author names in the citation, e.g., an author of the form "A. Anonymous."

8.2 Footnotes

Footnotes should be used sparingly. If you do require a footnote, indicate footnotes with a number[2] in the text. Place the footnotes at the bottom of the page on which they appear. Precede the footnote with a horizontal rule of 2 inches (12 picas).

Note that footnotes are properly typeset *after* punctuation marks.[3]

8.3 Figures

All artwork must be neat, clean, and legible. Lines should be dark enough for purposes of reproduction. The figure number and caption always appear after the figure. Place one line space before the figure caption and one line space after the figure. The figure caption should be lower case (except for first word and proper nouns); figures are numbered consecutively.

You may use color figures. However, it is best for the figure captions and the paper body to be legible if the paper is printed in either black/white or in color.

8.4 Tables

All tables must be centered, neat, clean and legible. The table number and title always appear before the table. See Table 1.

[2] Sample of the first footnote.
[3] As in this example.

Figure 1: Sample figure caption.

Table 1: Sample table title

Part		
Name	Description	Size (μm)
Dendrite	Input terminal	$\sim$100
Axon	Output terminal	$\sim$10
Soma	Cell body	up to 10^6

Place one line space before the table title, one line space after the table title, and one line space after the table. The table title must be lower case (except for first word and proper nouns); tables are numbered consecutively.

Note that publication-quality tables *do not contain vertical rules.* We strongly suggest the use of the `booktabs` package, which allows for typesetting high-quality, professional tables:

`https://www.ctan.org/pkg/booktabs`

This package was used to typeset Table 1.

References

Mohammad Babaeizadeh, Chelsea Finn, D. Erhan, Roy H. Campbell, and Sergey Levine. Stochastic variational video prediction. *ArXiv*, abs/1710.11252, 2017.

Nicolas Ballas, L. Yao, Christopher Joseph Pal, and Aaron C. Courville. Delving deeper into convolutional networks for learning video representations. *CoRR*, abs/1511.06432, 2015.

Zheng Chang. Mau: A motion-aware unit for video prediction and beyond. In *Neural Information Processing Systems*, 2021.

Chelsea Finn, Ian J. Goodfellow, and Sergey Levine. Unsupervised learning for physical interaction through video prediction. *ArXiv*, abs/1605.07157, 2016.

Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43:4037–4058, 2019.

Sangmin Lee, Hak Gu Kim, Dae Hwi Choi, Hyungil Kim, and Yong Man Ro. Video prediction recalling long-term motion context via memory alignment learning. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3053–3062, 2021.

Zhouyong Liu, Shun Nian Luo, Wubin Li, Jingben Lu, Yufan Wu, Chunguo Li, and Luxi Yang. Convtransformer: A convolutional transformer network for video frame synthesis. *ArXiv*, abs/2011.10185, 2020.

William Lotter, Gabriel Kreiman, and David D. Cox. Deep predictive coding networks for video prediction and unsupervised learning. *ArXiv*, abs/1605.08104, 2016.

Michaël Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. *CoRR*, abs/1511.05440, 2015.

Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixé, and Christoph Feichtenhofer. Trackformer: Multi-object tracking with transformers. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8834–8844, 2021.

Xingjian Shi, Zhourong Chen, Hao Wang, D. Y. Yeung, Wai-Kin Wong, and Wang chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *NIPS*, 2015.

Huazhe Xu, Yang Gao, Fisher Yu, and Trevor Darrell. End-to-end learning of driving models from large-scale video datasets. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3530–3538, 2016.

Xiutiao Ye and Guillaume-Alexandre Bilodeau. Vptr: Efficient transformers for video prediction. *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 3492–3499, 2022.