# Video Prediction

February 16, 2022

Matthew Coleman 23'  MAE 340
**Advisor:** Professor Olga Russakovsky

## 1 Background

In many machine learning fields, it is common practice to provide data to a model in the form of media labeled by large groups of people, for example the popular ImageNet dataset was put together in part by Amazon's Mechanical Turk program, a crowdsourcing website that employs thousands of on-demand workers to contribute to data validation and research tasks. [5]

While this method works well for practical tasks like image classification, detection, segmentation, etc., datasets collected in such a way naturally reflect human biases, and this directly drives the training of models that perpetuate that bias in their real-world implementations. [8] Among these human biases are the most egregious: race, gender, sexuality, etc., but also the most innocuous, for example in classification tasks, niche, cultural, or otherwise "nonconforming" items may be labeled incorrectly or miscategorized as belonging to a more well-known group. In labeling segmentation tasks, researchers must make countless "judgement calls" about whether something should be considered part of something else or an object in its own right, and of course, it is notoriously difficult to get thousands of researchers to agree on one way of doing things. [2]

The goal of teaching a machine everything about the world—all the while pretending to know everything about the world—presents a unique challenge. On one hand, it is desirable and necessary to produce high-classification-accuracy models to carry out tasks as soon as possible, but on the other hand it is also wise to seek out machine learning paradigms that don't suffer as much from human biases, even if they don't yet yield high percentages in classical tasks.

To combat this drawback, some computer vision tasks focus instead on learning directly from the world, rather than from humans, for example, by using only real-world observations as ground-truths. An example is the task of video prediction, in which the goal is to predict a future frame of a video stream given only the sequence of preceding frames. [4] This project will focus on implementing and experimenting on video prediction tasks with different model architectures.

## 2 Research Goals

The goals for this project are primarily exploratory: to survey video prediction models by implementing the most notable architectures, experimenting on different datasets, and making meaningful conclusions from the results. Also, if applicable, to label and determine sources of bias in models and datasets used.

### 2.1 Datasets

Datasets may include KTH [6] (human actions), BAIR action-free and action-conditioned (robotic pushing movements with and without robot state), [3], and Human 3.6m (human poses and actions) [1], among others.

### 2.2 Architectures

#### 2.2.1 Convolutional RNN

A Convolutional RNN is the absolute most basic model architecture capable of performing video prediction as well as the easiest to implement. Therefore, it is the natural choice for a baseline model with which to compare proceeding models on the same datasets.

#### 2.2.2 LSTM

An LSTM (Long-Short-Term-Memory) is a modification to the convolutional RNN which will likely perform better by being able to make use of data over longer time periods [7], however it is more difficult to implement, and may suffer from classical video prediction errors such as blurring.

#### 2.2.3 SAVP

SAVP (Stochastic Adversarial Video Prediction) is one of the most advanced methods of video prediction, involving a GAN (Generative Adversarial Network) as well as a probabilistic model. [3]

### 2.3 Methodology

Since the plan for this project involves a substantial amount of exploratory research and experimentation, the best way of accomplishing as much as possible will be to take an iterative approach: by first exploring a subject mentioned above, conferring with advisors such as Professor Russakovsky or Professor Majumdar to help draw meaningful conclusions, and then to choose a branching direction to continue exploration. In order to make sure that something substantive is accomplished even if exploration fails, it will be necessary to code as much as possible so that finished pipelines are completed and results can be discussed in the final report.

Also, while I plan to implement all aforementioned architectures, it may be necessary to use researched components either in part or in their entirety for implementing SAVP, as it is the most difficult architecture to implement.

## 3 Timeline

| Week | School Events | Project Todo |
|---|---|---|
| Feb. 14 | | Submit proposal and apply for funds |
| Feb. 21 | | Research, implement, and test real-world learning methods |
| Feb. 28 | Midterm Week | |
| Mar. 7 | Spring Break | |
| Mar. 14 | | |
| Mar. 21 | | Design and plan adaptations to learning methods |
| Mar. 28 | | Implement adaptations and experiment |
| Apr. 4 | | |
| Apr. 11 | | |
| Apr. 18 | Last Week of Classes | Write-up results and make poster |
| Apr. 25 | | Written Report Deadline (Apr. 26) and Poster Session (Apr. 27) |

## References


[1] C. S. Catalin Ionescu, Fuxin Li. Latent structured models for human pose estimation. In *International Conference on Computer Vision*, 2011.

[2] W. Ji, S. Yu, J. Wu, K. Ma, C. Bian, Q. Bi, J. Li, H. Liu, L. Cheng, and Y. Zheng. Learning calibrated medical image segmentation via multi-rater agreement modeling. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12336–12346, 2021. doi: 10.1109/CVPR46437.2021.01216.

[3] A. X. Lee, R. Zhang, F. Ebert, P. Abbeel, C. Finn, and S. Levine. Stochastic adversarial video prediction. *CoRR*, abs/1804.01523, 2018. URL `http://arxiv.org/abs/1804.01523`.

[4] S. Oprea, P. Martinez-Gonzalez, A. Garcia-Garcia, J. A. Castro-Vargas, S. Orts-Escolano, J. Garcia-Rodriguez, and A. Argyros. A review on deep learning techniques for video prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 1–1, 2020. ISSN 1939-3539. doi: 10.1109/tpami.2020.3045007. URL `http://dx.doi.org/10.1109/TPAMI.2020.3045007`.

[5] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

[6] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local svm approach. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 3, pages 32–36 Vol.3, 2004. doi: 10.1109/ICPR.2004.1334462.

[7] R. C. Staudemeyer and E. R. Morris. Understanding LSTM - a tutorial into long short-term memory recurrent neural networks. *CoRR*, abs/1909.09586, 2019. URL http://arxiv.org/abs/1909.09586.

[8] D. Zhao, A. Wang, and O. Russakovsky. Understanding and Evaluating Racial Biases in Image Captioning. *arXiv e-prints*, art. arXiv:2106.08503, June 2021.