Spatio-Temporal Attention for Manipulation Failure Detection

İpek Erdoğan, Assoc. Prof. Dr. Sanem Sarıel

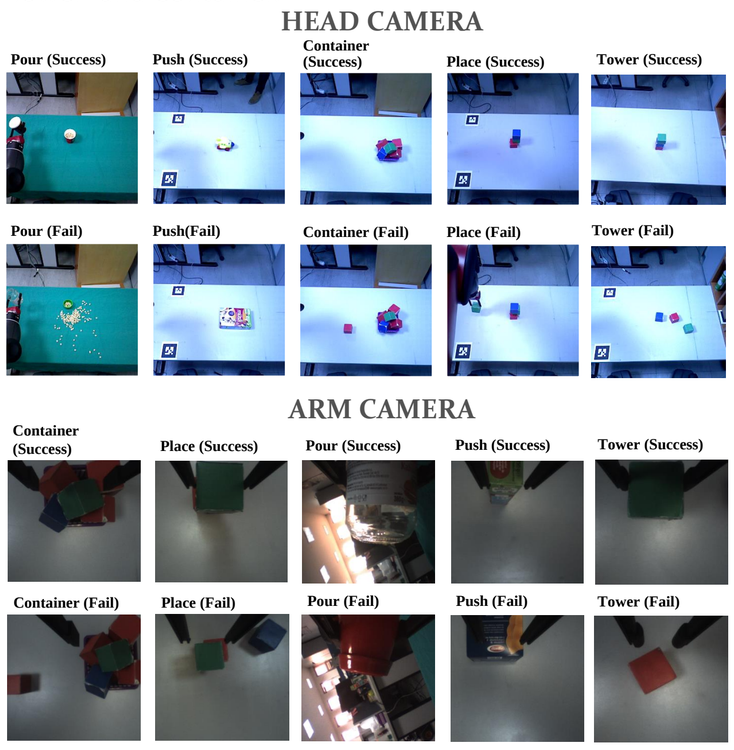
Istanbul Technical University, Computer Engineering Department

Abstract

When technological developments are taken into consideration, it is seen that scientists spend a lot of work in the field of robotics. It is one of these sub-topics to produce robots to accompany people in their daily lives and assist them in their work. However, the studies on humanoid robots that can perform human beings' actions in everyday life bring with them the possibility of making mistakes. These errors can negatively affect the lives of creatures in the environment that robot interacts. Therefore, it is necessary to give the robot a sense of understanding when it makes a mistake and even predicts how to avoid it. This project aims to detect the error when the robot makes a mistake by using deep learning models.

Dataset

The two datasets of this project consist of videos which have been taken by Baxter's cameras. These datasets are: Head Camera Dataset and Arm Camera Dataset. Both of the datasets have 5 subsets: Push, Place, Pour, Tower and Container.



Approaches

Data Processing

- Normalization
- Different Sampling Types
  - Random
  - Final
  - Equal distance
  - From beginnng and end
  - From middle
- Splitting the Dataset

Approaches to Reduce Overfitting

- Regularization
- Batch Normalization

Attention

- Temporal Attention
- Spatial Attention

Experiment Details

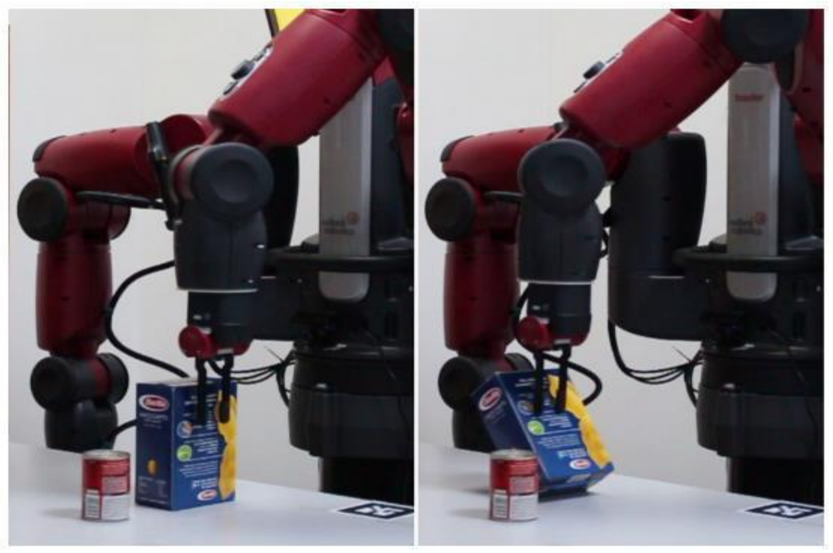
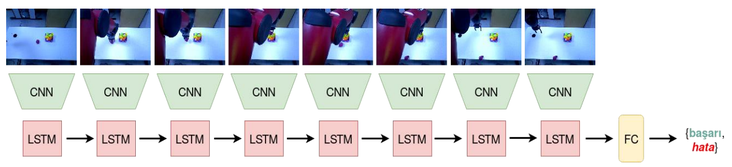


Figure 1.1: A "fail" situation during a push operation

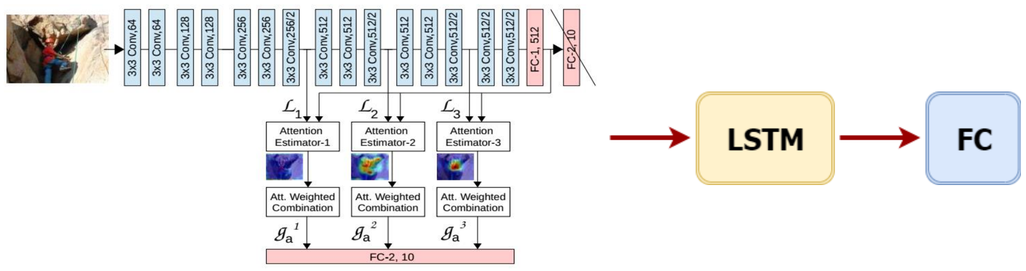
- L1 Regularization (Lambda = 0.001)
- Adam Optimizer
- 8 Frames
- Learning Rate: 0.0001
- Epochs: 50
- Iterations: 5

Models

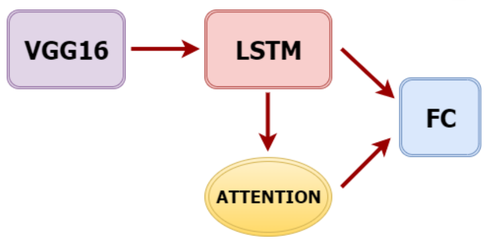
- **VGG16 + LSTM**



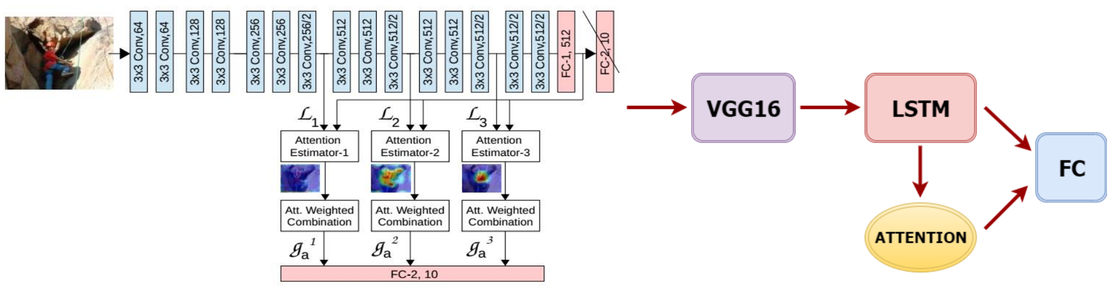
- **VGG16 + Spatial Attention +LSTM**



- **VGG16 + LSTM + Temporal Attention**



- **VGG16 + LSTM + Spatial Attention + Temporal Attention**



Results

Test Accuracies	All Dataset	Container	Place	Tower	Pour	Push
VGG16 + LSTM	0.80±0.01	0.59±0.02	0.74±0.03	0.97±0.04	0.85±0.01	0.69±0.01
VGG16 + Spatial Attention + LSTM	0.62±0.03	0.58±0.02	0.64±0.06	0.70±0.06	0.73±0.08	0.65±0.10
VGG16 + LSTM + Temporal Attention (1)	0.78±0.02	**0.61±0.03**	0.62±0.07	0.93±0.05	0.84±0.02	0.63±0.04
VGG16 + LSTM + Temporal Attention (2)	0.79±0.01	0.59±0.02	**0.75±0.06**	0.91±0.05	**0.86±0.02**	0.69±0.03
VGG16 + LSTM + Spatial & Temporal Attention	0.64±0.03	0.57±0.06	0.73±0.09	0.66±0.09	0.74±0.08	0.59±0.05

Conclusion

In this project, our aim was to increase our model's accuracy with help of both spatial and temporal attention mechanisms; but results are not enough consistent. One of the biggest reason of this result was our dataset. As it was not big enough, we couldn't train our models with our own datasets end-to-end, we had to use pre-trained models. Also, dataset's heterogeneity restricted us to reach a convergent result.

Future Work

- Creating a new dataset
- Multimodal sensor fusion
- Use ConvLSTM instead of LSTM

QR Code