Fakultät für Mathematik, Informatik und Naturwissenschaften
Lehr- und Forschungsgebiet Informatik VIII
Computer Vision
Prof. Dr. Bastian Leibe

Seminar Report

Unsupervised Learning of Video Representations using LSTMs

Cun Wang
Matriculation Number: 362024

July 2017

Advisor: Moritz Wolter

Abstract

In this lab, Long Short Term Memory (LSTM) networks are used to learn the representation of video sequences. It adpots the Encoder-Decoder framework. An encoder LSTM is used to map the inputs sequences to a fixed length representation. Then an decoder LSTM decodes this representation to reconstruct the input sequences. In order to capture spatiotemporal correlation better, the convolutional LSTM (ConvLSTM) is applied besides fully-connected LSTM (FC-LSTM).

Contents

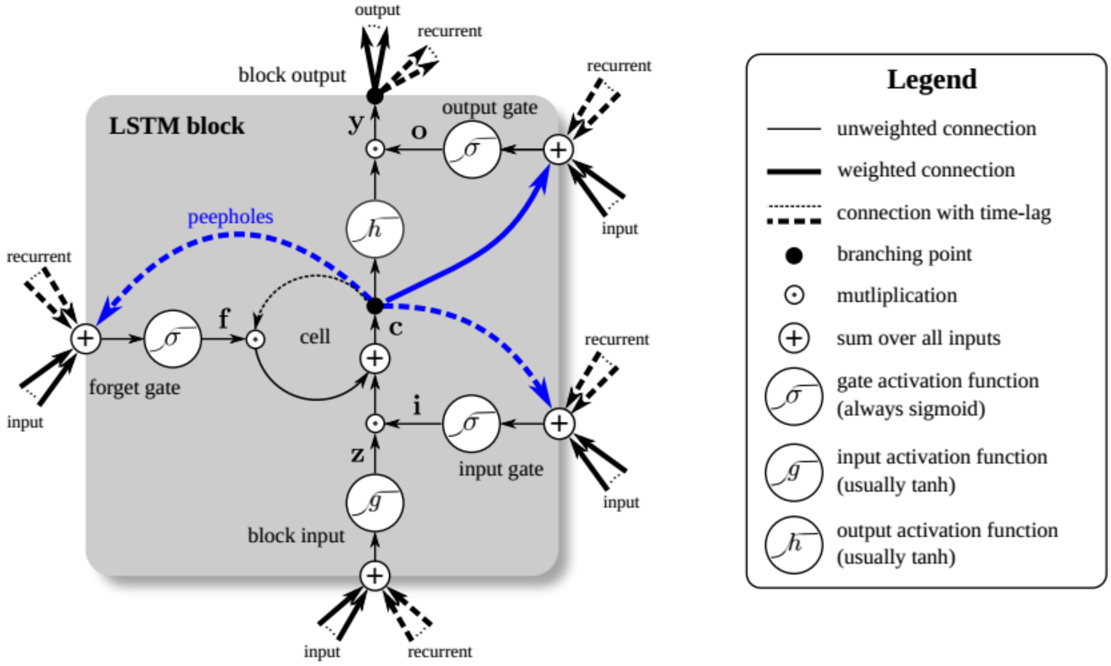

Figure 1: vanilla LSTM block architecture

Source: [GSK$^+$15]

1 Introduction

Understanding temporal sequences is important for solving many problems, such as speech recognition, caption generation. Srivastava *et al.* use the LSTM Encoder-Decoder framework to learn video representations [SMS15]. A sequence of frames are fed in encoder LSTM to generate a representation. This representation is then decoded through another LSTM to produce a target sequence. Srivastava *et al.* consider two choices of the target. One choice is to use the inputs as the target sequence. The other is to predict the future frames. They also evaluate on two kinds of inputs. One is image patches, the other is the high-level "percepts" extracted by applying a convolutional net.

LSTM is an important part of Encoder-Decoder framework. LSTM is a specical kind of recurrent neural network, which can capture long-term temporal dependencies without suffering from the optimization problems such as gradient vanishing. In order to make use of frame structure, convolutional LSTM introduced in [SCW$^+$15] is applied in Encoder-Decoder framework.

2 Methodology

2.1 LSTM architecture

As mentioned above, Encoder-Decoder framework is composed of LSTMs. We'll introduce vanilla LSTM architecture [GSK$^+$15]. A schematic of the vanilla LSTM block is illustrated in figure 1. The key to LSTMs is the cell state, which is controlled by three kinds of gates, adding information to or removing it from the cell state. The gates are composed of a sigmoid layer and a pointwise multiplication operation. The output of sigmoid layer ranges from zero to one, describing how much of each component should be let through. The forgate gate decides what information is going to be throwed away from the cell state. The input gate decides what information is going to be stored in the cell state. The output gate decides what is going to output.

The vector formulas for a vanilla LSTM layer forward pass are given below. $\boldsymbol{x}^t$ is the input vector at time t, the $\boldsymbol{W}$ are rectangular input weight matrices, the $\boldsymbol{p}$ are peephole weight vectors and $\boldsymbol{b}$ are bias vectors.

Functions σ, g and h are pointwise nonlinear activation functions. The pointwise multiplication of two vectors is denoted with ⊙:

2.2 Convolutional LSTM

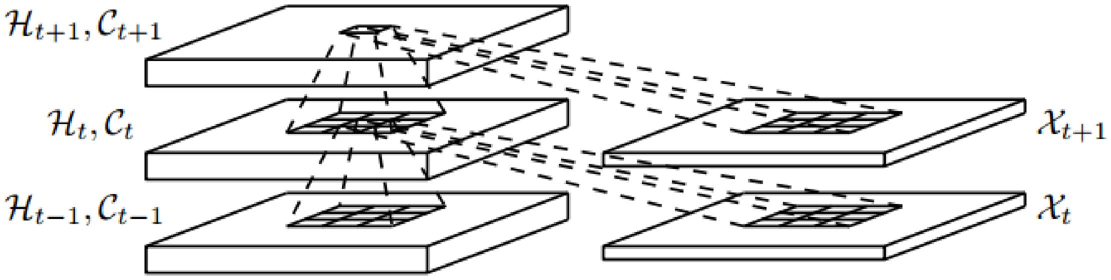
$\mathcal{H}_{t+1}, \mathcal{C}_{t+1}$
$\mathcal{H}_{t}, \mathcal{C}_{t}$
$\mathcal{H}_{t-1}, \mathcal{C}_{t-1}$
$\mathcal{X}_{t+1}$
$\mathcal{X}_{t}$

Figure 2: inner structure of convolutional LSTM

Source: adopted from slides of Kaiming He

2.3 LSTM autoencoder model

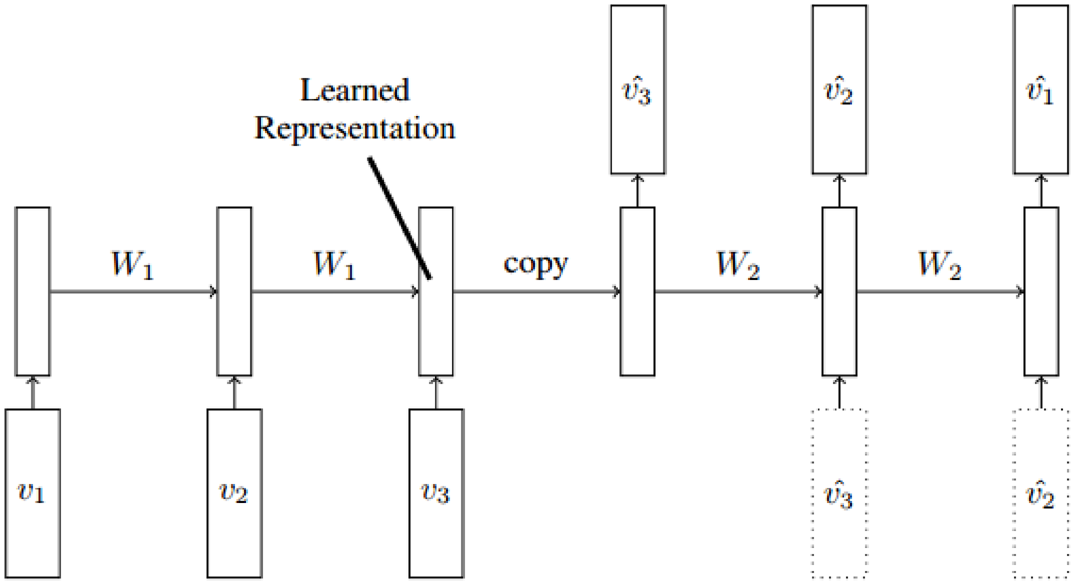

Figure 3: LSTM Autoencoder Model

Source: adopted from slides of Kaiming He

3 Experiments

3.1 Datasets

The model is trained on a dataset of moving MNIST digits. Each video is 20 frames long and consist of 2 digits moving inside 64×64 patch.

References

[GSK+15] Klaus Greff, Rupesh Kumar Srivastava, Jan Koutník, Bas R. Steunebrink, and Jürgen Schmidhuber. LSTM: A search space odyssey. *CoRR*, abs/1503.04069, 2015.

[SCW+15] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 802–810, 2015.

[SMS15] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhutdinov. Unsupervised learning of video representations using lstms. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pages 843–852, 2015.