Bachelorthesis

Comparing different state-of-the-art solutions for image prediction using time-series analysis

Sören Dittrich
soeren.dittrich@uni-hildesheim.de

September 2020

Abstract

This thesis compares different state-of-the-art solutions for image prediction. Key aspect of the work is the comparison of different versions of the ConvLSTM [7]. To be able to compare those different versions, an image prediction architecture, more explicit PredNet [4], is implemented as baseline architecture. This architecture uses the ConvLSTM module as recurrent sub-module. This sub-module is then changed during the experiments with another implementation (PredRNN [11]). Other comparisons are then performed theoretically The thesis introduces the reader about image prediction, convolutional LSTM's and other necessary parts to be able to follow. It then describes the PredNet architecture and idea and how this differs from other image prediction architectures. Then the performed experiments are described in-depth. Afterwards it gives a comprehensive discussion part, where the different sub-module performances are discussed very deeply, to understand the experimental results. Lastly there is a conclusion, which sums everything up.

Contents

1 Scientific questions

1. What different types of image prediction architectures exist?
2. What different types of recurrent modules exist?
3. How important is the choice of the recurrent module for the runtime and performance of the algorithm?

2 Introduction

The thesis will compare different state-of-the-art solutions for image prediction 2.2. The main module, which is a core aspect of this work, is the LSTM (Long short-term memory). 2.6. This module was invented by Hochreiter and Schmidhuber [3] in 1997 and is used heavily in the field of image prediction since then, e.g. in Srivastave et. al. [8]. During the time the module got many different add-ons and changes, which are described in different papers ([6], [4], [11], [10] and many more.). This work is implementing one specific network architecture (PredNet [4]). It uses the Shi et. al. „standard“ConvLSTM [7] as recurrent sub-module, which is changed during the experiments with other (more advanced) solutions. The PredNet algorithm is re-implemented in PyTorch [5], as well as the „standard“ConvLSTM and PredRNN. Practical experiments are performed on the PredNet using the „standard“ConvLSTM and on PredNet using PredRNN instead.

2.1 Deep Learning

2.2 Image Prediction

Image prediction is a field inside machine learning, where the key is to predict future images, given a sequence of image. The image sequence X is of length n, $(x_0, \ldots, x_{n-1})$. One possible use-case is the one-frame prediction, where one predicts x_n, given the the sequence X. Another common use-case is multi-frame prediction, where the key is to predict t many frames into the future. This is often performed using sequence-to-sequence learning [9]. Obviously the first frames look much better then the last frames, as ground-truth is missing, and the predicted frames are approximated, which means they contain a certain level of error.

2.3 Autoencoder

2.4 RNN

RNN (Recurrent neural network)

2.5 LSTM

LSTM (Long Short-term Memory) [3] is a form of RNN, which avoids a critical problem of standard RNN: Saving **Long-term dependencies** [2]. The architecture consists of different submodules, an inpute-gate, forget-gate, cell-state and output-gate.

$$i_t = \sigma(w_{x_i} x_t + w_{h_i} h_{t-1} + w_{c_i} c_{t-1} + b_i) \tag{1}$$

$$f_t = \sigma(w_{x_f} x_t + w_{h_f} h_{t-1} + w_{c_f} c_{t-1} + b_f) \tag{2}$$

$$c_t = f_t c_{t-1} + i_t tanh(w_{x_c} x_t + w_{h_c} h_{t-1} + b_c) \tag{3}$$

$$o_t = \sigma(w_{x_o} x_t + w_{h_o} h_{t-1} + w_{c_o} c_t + b_o) \tag{4}$$

$$h_t = o_t tanh(c_t) \tag{5}$$

w is the weight of the layer, σ the sigmoid function, b the layer bias. h_t is the output, in RNN's the output is often denoted as hidden.

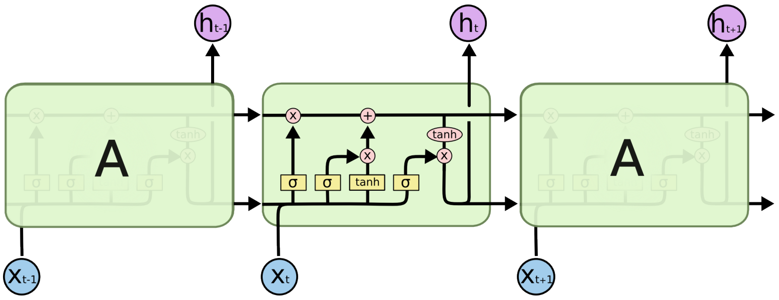

Figure 1: LSTM Architecture [1]

2.6 ConvLSTM

The convolutional LSTM, invented by Shi et. al. [7] is a LSTM using convolutional layer instead of fully connected ones. Therefore the formulas looks very similar to the ones in section 2.5.

$$i_t = \sigma(x_t * w_{x_i} + h_{t-1} * w_{h_i} + w_{i_b}) \tag{6}$$

$$f_t = \sigma(x_t * w_{x_f} + h_{t-1} * w_{h_f} + w_{f_b}) \tag{7}$$

$$\tilde{c_t} = tanh(x_t * w_{x_{\tilde{c}}} + h_{t-1} * w_{h_{\tilde{c}}} + w_{c_{\tilde{b}}}) \tag{8}$$

$$c_t = \tilde{c_t} \odot i_t + c_{t-1} \odot f_t \tag{9}$$

$$o_t = \sigma(x_t * w_{x_o} + h_{t-1} * w_{h_o} + w_{o_b}) \tag{10}$$

$$h_t = o_t \odot tanh(c_t) \tag{11}$$

$*$ is the commonly used sign for the convolution operation.
$\odot$ is the hadamard product (point-wise multiplication).

2.7 PyTorch

3 Image Prediction Architectures

This section will describe a range of state-of-the-art architectures for image prediction. Image prediction is a very broad field, but almost all state-of-the-art solutions for image prediction share one common part, the recurrent module. LSTM's are the most used modules in image prediction, as they are able to store information over a long period of time, despite the standard RNN (recurrent neural network). All algorithms described here have a different way to perform image prediction, but all use a type of LSTM to store the time-series information.

3.1 LSTM Autoencoder

The paper „Unsupervised Learning of Video Representations using LSTMs“by Srivastava et. al. [8] is using the standard LSTM 2.5 in an autoencoder architecture for reconstruction and prediction. As this thesis topic is image prediction for future images, I will not cover the reconstruction architecture. The architecture is often used as a baseline in newer and more advanced architectures, because it consists of the standard LSTM. The model is typically trained using reconstruction error.

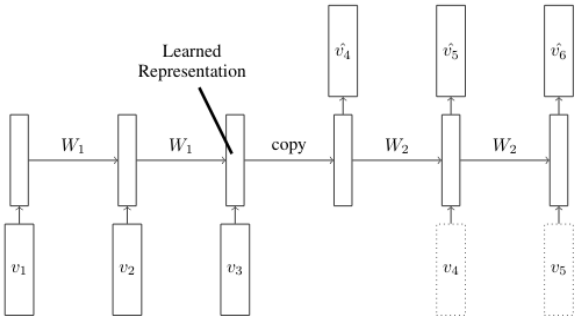

Figure 2: Future image prediction model [8]

3.2 ConvLSTM Autoencoder

The paper „Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting“by Shi et. al. [7] is using a similar architecture as Srivastava et. al. in section 3, but instead of using the standard LSTM, they use a ConvLSTM 2.6. This architecture outperforms the Srivastava et. al., because it „captures spatiotemporal correlations better“.

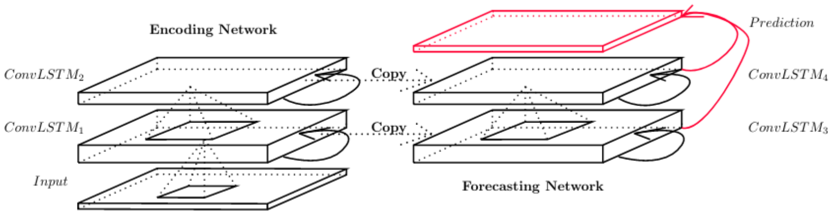

Figure 3: Future image prediction model [7]

3.3 Spatio-temporal Video Autoencoder

3.4 PredNet

The chosen implementation for the experiments 7.

4 Implementation

5 Methodology

6 Training

The training section will cover the aspects of different training types.

7 Experiments

The experiments performed on the implemented PredNet with ConvLSTM and PredNet with PredRNN will be described here. Also other theoretical comparisons will be covered in this section.

8 Discussion

9 Conclusion

10 Explanation

Erklärung über das selbstständige Verfassen von „Comparing different state-of-the-art solutions for image prediction using time-series analysis“
Ich versichere hiermit, dass ich die vorstehende Bachelorarbeit selbständig verfasst und keine anderen als die angegebenen Hilfsmittel benutzt habe. Die Stellen der obigen Arbeit, die anderen Werken dem Wortlaut oder dem Sinn nach entnommen wurden, habe ich in jedem einzelnen Fall durch die Angabe der Quelle bzw. der Herkunft, auch der benutzten Sekundärliteratur, als Entlehnung kenntlich gemacht. Dies gilt auch für Zeichnungen, Skizzen, bildliche Darstellungen sowie für Quellen aus dem Internet und anderen elektronischen Text-und Datensammlungen und dergleichen. Die eingereichte Arbeit ist nicht anderweitig als Prüfungsleistung verwendet worden oder in deutscher oder in einer anderen Sprache als Veröffentlichung erschienen. Mir ist bewusst, dass wahrheitswidrige Angaben als Täuschung behandelt werden.

Datum, Ort Unterschrift

References

[1] *Understanding LSTM Networks.* https://colah.github.io/posts/2015-08-Understanding-LSTMs/. – Accessed: 2020-07-13

[2] Goodfellow, Ian ; Bengio, Yoshua ; Courville, Aaron: *Deep Learning.* MIT Press, 2016. – http://www.deeplearningbook.org

[3] Hochreiter, Sepp ; Schmidhuber, Jürgen: Long Short-term Memory. In: *Neural computation* 9 (1997), 12, S. 1735–80

[4] Lotter, William ; Kreiman, Gabriel ; Cox, David D.: Deep Predictive Coding Networks for Video Prediction and Unsupervised Learning. In: *CoRR* abs/1605.08104 (2016). – URL http://arxiv.org/abs/1605.08104

[5] Paszke, Adam ; Gross, Sam ; Massa, Francisco ; Lerer, Adam ; Bradbury, James ; Chanan, Gregory ; Killeen, Trevor ; Lin, Zeming ; Gimelshein, Natalia ; Antiga, Luca ; Desmaison, Alban ; Kopf, Andreas ; Yang, Edward ; DeVito, Zachary ; Raison, Martin ; Tejani, Alykhan ; Chilamkurthy, Sasank ; Steiner, Benoit ; Fang, Lu ; Bai, Junjie ; Chintala, Soumith: PyTorch: An Imperative Style, High-Performance Deep Learning Library. In: Wallach, H. (Hrsg.) ; Larochelle, H. (Hrsg.) ; Beygelzimer, A. (Hrsg.) ; Alché-Buc, F. d(Hrsg.) ; Fox, E. (Hrsg.) ; Garnett, R. (Hrsg.): *Advances in Neural Information Processing Systems 32.* Curran Associates, Inc., 2019, S. 8024–8035. – URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[6] Patraucean, Viorica ; Handa, Ankur ; Cipolla, Roberto: Spatio-temporal video autoencoder with differentiable memory. In: *CoRR* abs/1511.06309 (2015). – URL http://arxiv.org/abs/1511.06309

[7] SHI, Xingjian ; CHEN, Zhourong ; WANG, Hao ; YEUNG, Dit-Yan ; WONG, Wai-kin ; WOO, Wang-chun: Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. In: CORTES, C. (Hrsg.) ; LAWRENCE, N. D. (Hrsg.) ; LEE, D. D. (Hrsg.) ; SUGIYAMA, M. (Hrsg.) ; GARNETT, R. (Hrsg.): *Advances in Neural Information Processing Systems 28*. Curran Associates, Inc., 2015, S. 802–810. – URL http://papers.nips.cc/paper/5955-convolutional-lstm-network-a-machine-learning-approach-for-precipitation-nowca pdf

[8] SRIVASTAVA, Nitish ; MANSIMOV, Elman ; SALAKHUTDINOV, Ruslan: Unsupervised Learning of Video Representations using LSTMs. In: *CoRR* abs/1502.04681 (2015). – URL http://arxiv.org/abs/1502.04681

[9] SUTSKEVER, Ilya ; VINYALS, Oriol ; LE, Quoc V.: Sequence to Sequence Learning with Neural Networks. In: *CoRR* abs/1409.3215 (2014). – URL http://arxiv.org/abs/1409.3215

[10] WANG, Yunbo ; GAO, Zhifeng ; LONG, Mingsheng ; WANG, Jianmin ; YU, Philip S.: PredRNN++: Towards A Resolution of the Deep-in-Time Dilemma in Spatiotemporal Predictive Learning. In: *CoRR* abs/1804.06300 (2018). – URL http://arxiv.org/abs/1804.06300

[11] WANG, Yunbo ; LONG, Mingsheng ; WANG, Jianmin ; GAO, Zhifeng ; YU, Philip S.: PredRNN: Recurrent Neural Networks for Predictive Learning using Spatiotemporal LSTMs. In: *NIPS*, 2017