Human Motion Prediction

Maria Rozou
rozoum@student.ethz.ch

Rahel Straessle
strrahel@student.ethz.ch

ABSTRACT

Recurrent Neural Networks are commonly used in natural language translation and also generation. In the recent years, they have shown good performance when applying them to human motion prediction tasks.

0.1 Preprocesssing and representation

1 INTRODUCTION

Motion prediction is one of the tasks that is automatically done by humans but very hard to accomplish by machines. While even insects already have a physical model of the world, either learnt or predefined, we have to teach how to predict motion to machines. There are many applications nowadays that are based on human motion data analysis like human-computer interaction, motion synthesis, motion prediction for virtual and augmented reality, automated driving, biomedical engineering applications etc. Using a state of the art unrolled LSTM cell we build the most basic RNN network for human motion prediction and took this as our baseline for improvements.

2 RELATED WORK

Recent work focused on RNN based architectures to model human motion, with the goal of learning time-dependent representations that perform tasks such as short term motion prediction.

In 2014, an encoder-decoder (ERD) network based on LSTM was introduced, which is a type of recurrent neural network (RNN) model, that combines representation learning with learning temporal dynamics [2]. Basic preprocessing is done to increase performance and robustness of the network, like standarization of data and adding noise to the input data. Around the same time, a sequence to sequence architecture for language translation was proposed [5] that would give the field a new direction. The network consists of an encoder and a decoder and would take as input one whole sentence. An extended sequence to sequence model was recently published [3]. It is proposed that weight sharing between encoder and decoder and using Gated Rectified Units (GRU) instead of LSTM as base cell simplifies the model but does not hurt final performance. A residual connection between input and output allows to use velocities in the network - a advantages add-on as human motion is dynamic and not static.

Good long-term results were obtained with a dubbed Dropout Autoencoder LSTM (DAE-LSTM) that can extract both structural and temporal dependencies directly from the training data ETH: [**?**]. This network can produce long-term naturally looking poses without converging to mean. The autoencoder is trained to implicitly recover the spatial structure of the human skeleton via randomly removing information about joints during training.

A different approach shows that a feed-forward network can also compete with RNNs [1].

3 METHODOLOGY

3.1 Preprocesssing and representation

Preprocessing the data before feeding them into the network is understood to be advantageous [2]. However, for our RNN standardization of the input data would not lead to the best performing model. We standardized the data by subtracting the mean value of each joint and also divided by the standard deviation. One point worth mentioning is that some joints had zero standard deviation, meaning that the training data set were not changing at all for these angles, therefore we decided to discard them and not use them in the analysis. This led to an input with 51 angles instead of 75. If the test set is coming from the same distribution, we could expect that the same angles also in the test set have 0 std so they can be removed from the LSTM analysis and for the test purposes just keep the last frame of these angles as the predicted one. We did not see any improvements of the performance of our LSTM model compared to non-standardized inputs.

3.2 Network structure

In this project, the architecture we used to solve the problem of motion prediction is based on a RNN architecture using LSTM cells [4]. We created an unrolled LSTM network that in each time step is being trained to predict the next pose . Therefore we identified the problem as a regression task, where we construct the predicted poses in the continuous angles-space. This can be seen also in Figure 1, where the red poses are the predicted poses from the LSTM architecture and the green poses are the ground truth, which is in our case the real-true positions from the training data. For the prediction, 50 source frames are feed into the model. The model predicts the first next step using the last feed frame and continuous afterwards using the predicted frame as the input for the next prediction until frame 25. The process is shown in Figure 2.

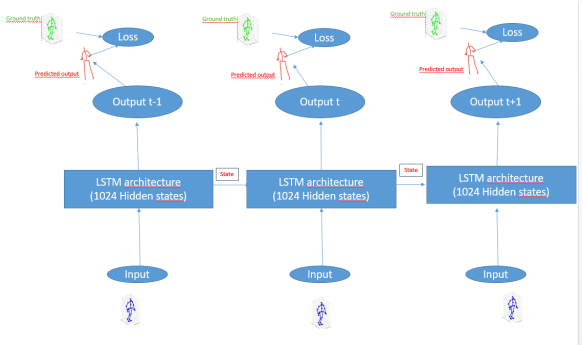

Figure 1: Training architecture

For language translation it was shown that deep LSTM's significantly outperform shallow LSTMs [5]. An optimum was reached with a sequence to sequence architecture with 4 layers. Stacking of

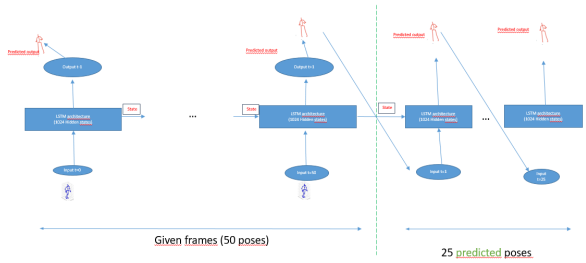

Figure 2: Prediction architecture

LSTM cells increases the model complexity and predictive power but at the expense of training times and difficulties. We combined the LSTM cell with a drop-out and stacked those. Dropout was added as it makes models in general more robust.

3.3 Experiments

Our models are trained on 162 samples with different length. Each sample is a dictionary with an array with 3x25 entries for the joint angles per frame, the action label and the id's. The validation was done on a separate set of 18 samples. For training and validation set together, there were 12 samples of 14 different activities. All the samples are cut at a maximum sequence length and shorter sequences padded with zeros. The test set consists of 516 samples with 50 frames each.

The loss function we used was the mean squared error function (L2 loss) to calculate the prediction error for the training sequence. As comparison, we also tested another distance function, the cosine loss.

To solve the optimization problem in the training process, we used stochastic gradient descent (SGD). For the optimizer, we looked at different choices: the ADAM optimizer, the Gradient Descent Optimizer and the RMSPropOptimizer. Moreover, next to the learning rate more parameters of the optimizer can be tuned as well.

Parameters than can be tuned:

4 RESULTS AND DISCUSSION

The best performing model was the most basic LSTM model with one layer, trained with the L2 loss and optimized with the Adam Optimizer. We clipped the gradient norm to a factor of 5, so that we limit the magnitude of the gradients. With such a configuration, the Adam optimizer descends indeed to a minimum. In the following table the parameters of the LSTM architecture are summarized. The training progress shows that after 10 epochs there was no substantial decrease in validation error. To avoid overfitting, our final submission is with 6 epochs. A small batch size can help to step away from local minima. In our best run, we had a batch size of 10.

Adding more layers would not yield better results compared to a single layer. It is well possible that we couldn't find the convergence point for this model. Neither did it help to add a drop-out layer. Dropout is more useful for deep networks, which can be the reason that dropout would not improve our validation error, even after

Table 1: Parameters for our best performing model

Parameter	Value
Learning rate	0.0015 fixed
Batch Size	10
Optimizer	Adam
Epochs	9
Max Length size	600
Clipping gradient	5
Layer	1

12 epochs, for our rather shallow networks. The predictions were very human like, also for the longer term prediction, but they were flickering or noisy. Even though in this model we could not see an improvement, we believe that both stacking of more layers and dropout could increase model performance and add robustness to the model if another architecture was chosen.

Other distance functions, like the cosine loss, did not improve the performance of the model on the validation set and instead the results were poorer. As well for the optimizer, changing to another optimizer or adjusting the settings, such as epsilon, would not improve our model performance.

5 CONCLUSION AND OUTLOOK

Our network performed reasonably well for its simplicity. Tuning the hyper-parameters, optimizer or loss function would not have a significant influence on the performance. Adding more layers and dropout could not improve the model. This is in line what we could expect when consulting the literature.

The movements of human body are sequential, meaning in each frame there is a dependency from the previous. Therefore, the obvious architecture that is outperforming the LSTM network is a sequence to sequence (seq2seq) architecture, where two networks are trained: an encoder generates an internal representation of the input, which is fed to the decoder. The decoder produces a maximum likelihood estimation for the prediction. One of the obvious advantages is that the model is trained for exactly the task that it should carry out in the test. As the seq2seq architecture is heavily used in machine translation, there are several options to improve the architecture also for motion prediction, such as attention or bidirectional architectures.

Other distance functions, like the cosine loss, did not improve the performance of the model on the validation set or even gave worse results.

Another preprocessing step would be to add Gaussian noise to the angles as has been done in previous work [2]. This makes the LSTM more robust and can help to avoid overfitting. However, the robustness comes with the price of having another hyperparameter to tune: the standard deviation of the Gaussian noise. Another solution to increase robustness of the model is to feed into the decoder its own produced samples. This comes without any parameter to tune and yields good results [3].

Human motion is dynamic and not static and human motion can be better expressed in velocities than in positions. Previous work has taken this into account and used a residual connection to have the prediction continuous [3]. Many architectures suffer from

a unreasonable first step when the prediction starts which can be avoided with the residual connection. A add-on to this model could include a further connection from the last input to the output that would teach the model about acceleration too.

ACKNOWLEDGMENTS

The authors would like to thank Prof. Dr. Otmar Hilliges for providing the theoretical background needed for this work and his assistants for the skeleton code. The work was carried out in the frame of the lecture "Machine Perception" at ETH Zurich in the spring semester 2018.

REFERENCES

[1] Judith Bütepage, Michael J. Black, Danica Kragic, and Hedvig Kjellström. 2017. Deep representation learning for human motion prediction and classification. *CoRR* abs/1702.07486 (2017). arXiv:1702.07486 http://arxiv.org/abs/1702.07486

[2] Katerina Fragkiadaki, Sergey Levine, and Jitendra Malik. 2015. Recurrent Network Models for Kinematic Tracking. *CoRR* abs/1508.00271 (2015). http://arxiv.org/abs/1508.00271

[3] Julieta Martinez, Michael J. Black, and Javier Romero. 2017. On human motion prediction using recurrent neural networks. In *CVPR*.

[4] Christopher Olah. [n. d.]. Understanding LSTM Networks. http://colah.github.io/posts/2015-08-Understanding-LSTMs/

[5] I. Sutskever, O. Vinyals, and Q. V. Le. [n. d.]. Sequence to Sequence Learning with Neural Networks. *ArXiv e-prints* ([n. d.]). arXiv:1409