Machine Perception : summary of articles

mjazbec

April 2019

1 RNNs for human dynamics

- "prediction of human body pose in videos and motion capture. The ERD (encoder-recurrent-decoder) model is a recurrent neural network that incorporates nonlinear encoder and decoder networks before and after recurrent layers"
- article describes 3 tasks: **motion capture generation** (mocap), body pose labeling and body pose forecasting in videos. In Machine Perception project we are dealing with the mocap task.
- *Side note (NLU):* "For example, word embeddings were not crucial for RNNs to do well in text generation or machine translation, and the standard one hot encoding vocabulary representation also showed excellent results
- "We consider both deterministic and probabilistic predictions. In the probabilistic case, y_t parametrizes a Gaussian Mixture Model (**GMM**) over mocap vectors in the next frame [31]."
- Experiments with initializing weights in encoder and decoder parts of the ERD (significant improvements over random initialisation were not observed).
- "We regularize our mocap ERD with denoising: we provide mocap vectors corrupted with zero mean Gaussian noise [42]. **Curriculum learning**." (in the second article they present a better alternative to this - feed the network its own predicted samples).
- ERD vs LSTM-3LR

1.1 Questions

- Is the ERD model presented here equivalent/similar to the dummy model from code skeleton?

2 On human motion prediction using RNNs

- discontinuity problem
- feeding ground-truths vs predictions during training
- residual architecture that models first-order motion derivatives
- "deep networks are known to be hard to train when data is scarce (which is the data regime for action-specific motion models)."
- **seq2seq architectures**. " Moreover, there are multiple variations of seq2seq architectures (e.g., with attention mechanisms [4], or bi-directional encoders [35]), that could potentially improve motion prediction."
- "Striving for simplicity, during training we let the decoder produce a sequence by always taking as input its own samples".
- "...in our case residual connections help us model prior knowledge about the statistics of human motion."
- **Baselines:** "We also consider an agnostic zero-velocity baseline which constantly predicts the last observed frame. For completeness, we also consider running averages of the last two and four observed frames"

2.1 Questions:

- Is zero-velocity baseline (predicting the last observed frame) even possible in our setting?

3 General questions

- MSE vs Mean join angle difference. Can we use latter also during training?
- data: These matrices are all relative to their parent. This means that the given rotation specifies how the parent bone must be rotated to obtain the rotation of the current bone.