Protocol 19: 2D CNN Training

17.05.2019

Convolutional Neural Networks (CNN)

The skeletal joints were used to construct images for training 2D CNN in the previous protocol. As mentioned in the document, the poor result may be caused by noisy data from pose estimation, or the skeleton sequence does not contain enough spatio-temporal information. To deal with the issues of skeletal data, motion sequence from raw videos is used as the input data in this protocol.

Figure 1: Motion sequence as an image.

1. Experiment

As shown in Figure 1, ten frames of a cow are combined into a motion sequence image. The sequence contains around one gait cycle. The number of frames depends on the amount of information: the more frames are included, the more spatio-temporal information is extracted but with less intensity of the moving cow.

- **Dataset:** Each data sample is a motion sequence image from 10 frames uniformly sampled from a gait cycle with a locomotion score as the label. The dataset has 501 samples, divided into training and validation sets with a 70/30 ratio.
- **Result:** The training accuracy improves gradually while the validation accuracy oscillates between 20% and 40%.
- **Notes:**
  - The method allows more data generated from raw video data.
  - There is a overfitting issue for thr training with or without pre-trained networks. Of all the available networks, ResNet has the highest training accuracy and thus the most obvious overfitting.

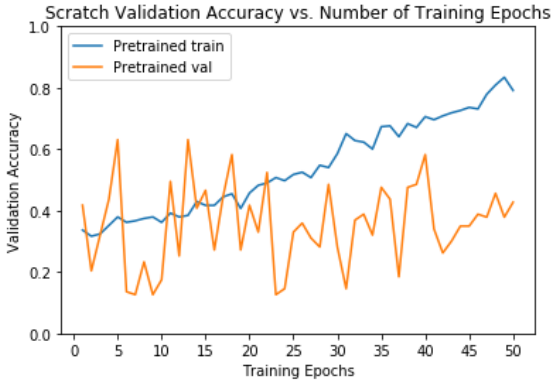

Figure 2: Training/validation loss using ResNet for 50 epochs with batch size of 16.

2. Discussion and Next Step

It turns out that the training of CNN with motion sequence images does not work either. There can be many reasons: the important spatio-temporal information is lost from the sampling of the frames; the amount of images is still too little for training a multi-layered CNN; the labels are not consistent for all the data. For the following weeks, LSTMs and traditional machine learning methods will be applied.