Exercise for MA-INF 2218 Video Analytics SS21
Submission on 20.07.2021
Weakly supervised action learning

In this sheet, you will build a weakly supervised action recognition system. The task is defined as follows:

Given an input video $\mathbf{x}_1^T = (x_1, \dots, x_T)$ with T frames either as raw video frames or as framewise features $x_t \in \mathbb{R}^D$, assign action label to each frame. During training, the only supervision available are action transcripts, i.e. for each video a sequence of actions that occur within the video.

Download the provided training data. Information for the specific files is provided in the `README`. Your system will be based on [1].

1. Train a weakly supervised system using HMMs+GMMs for five iterations. Start with a linear alignment of the action sequence to the video frames. You can find this alignment in `Data/uniform_labels`. The following steps are required:

   - Estimate the transition probabilities based on your current alignment.
   - Train the Gauss models using the current alignment as ground truth.
   - Realignment: compute the new alignment of hmm states to frames and use it as ground truth for the next iteration.
   - Repeat five times.

   Use 16 states for each HMM and initialize each action with the same number of HMM states. For the Gauss models, use a single Gaussian with a diagonal covariance matrix. *(4 Points)*

2. Replace the GMM by a simple multi-layer perceptron (MLP) with one hidden layer of 64 rectified linear units and a softmax output layer with one unit for each HMM state. Use the PyTorch framework and train the network for five epochs. Note that the network outputs are posterior probabilities of the form $p(s|x_t)$ where s is an HMM state. However, you need $p(x_t|s)$. Using Bayes rule, we have

   $$p(x_t|s) = \text{const} \cdot \frac{p(s|x_t)}{p(s)},$$

   so we can use the posteriors from the network if we remove the state prior (in log space) based on the current alignment. *(6 points)*

3. Replace the simple neural network by a recurrent network using gated recurrent units (GRUs). As input, the GRUs receive subsequences of 21 frames each (centered around the current frame). *(6 points)*

4. Train and evaluate the three systems (GMM, MLP, GRU) and fill your numbers in the table below. Compare your results with the results reported in Table 2 of [1]. What do you observe? What are possible explanations? (Be aware that [1] used 48 action classes and here, we have only 12 classes and a small subset of the dataset.) *(4 Points)*

	iter-1	iter-2	iter-3	iter-4	iter-5
GMM					
MLP					
GRU					

Please exclude the `data` folder from your submission due to its size. However, submit any file you created yourself. If you have any questions, feel free to contact me (**iqbalm@iai.uni-bonn.de**).

[1] A. Richard, H. Kuehne, J. Gall: Weakly supervised action learning with RNN based fine-to-coarse modeling, CVPR 2017.