# Video action classification

In this test case, your task is to perform video action classification on the Breakfast actions dataset. This dataset includes 1712 videos and shows activities related to breakfast preparation. Each video is composed of multiple sub-actions. Overall, there are 48 different sub-actions with in total approximately 11.3K samples, including around 3.6K "silence (SIL)" samples. For further information about the dataset, please refer to [KAS14]. In Figure 1, you can see an example video from this dataset, where a person is preparing coffee. There are 5 sub-actions, including the "silence (SIL)", with varying lengths of frames.

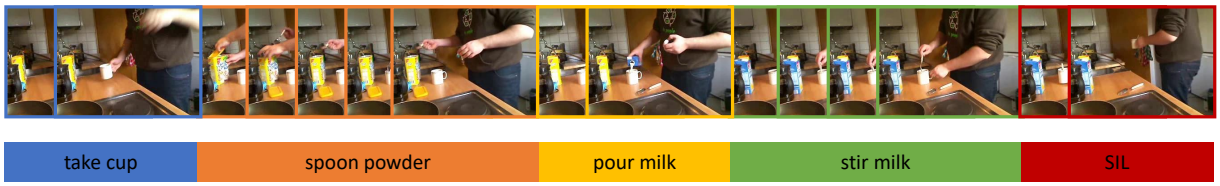


Figure 1: A person is making coffee. There are four sub-actions of varying number of frames, as well as one SIL (silence) segment.

For the Breakfast actions dataset, the visual features are computed per frame. Each frame has an associated sub-action label. You can download video frame features, ground truth labels and dataset splits from here. In the `''data''` folder you can find the I3D [CZ17] features that are computed for each frame. In the `''groundTruth''` folder you can find the frame-wise sub-action labels for each video. In the `''splits''` folder you can find the dataset splits. The dataset is composed of four splits and you are asked to train your models on the training set of split1 (`''train.split1.bundle''`) and test them on the test set (`''test.split1.bundle''`) of split1. You can use the `''read_datasetBreakfast.py''` script for processing the data.

Let denote $D$-dimensional frame features $\mathbf{x}_i^n \in \mathbf{R}^D$, for all $N_i$ frames, indexed by $n = 1, \ldots, N_i$ for video $i$. Frames are always represented by their feature vectors and we use the terms "frame", "frame feature" or "feature vector" interchangeably. Also, let $s_i^k$, $k = 1, \ldots, K_i$, denote the $K_i$ consecutive sub-actions of video $i$. For example, in Figure 1, there are 5 sub-actions including "SIL". Each sub-action, $s_i^k$, has a start $\mathbf{x}_i^{n_k}$ and end frame $\mathbf{x}_i^{n_{k+1}-1}$. We refer to the group of frames between the start and end frames as a "sub-action segment". Each segment $s_i^k$ corresponds to a sub-action label like "pour milk". The goal of this test case is to classify each segment to any of the 48 sub-actions. Your task is to experiment with different ways of segment representations and to perform action classification on top of these representations. You should report your classification accuracy averaged over all sub-actions in the test set. Note that $D = 400$ in this test case as we are using I3D features.

1. Given a segment, use a Long Short-Term Memory (LSTM) network for representing the segment frames. Use the BiLSTM (Bidirectional LSTM) with max-pooling architecture proposed by Conneau et. al. [Con+17] in your solution. The BiLSTM will input $D$-dimensional feature vectors for all the frames in a segment and output a $D$-dimensional

feature vector for the entire segment. You should use this final feature vector for classifying the segment. You should select the number of LSTM layers and the number of neurons experimentally.

2. Given a segment, apply max-pooling over the frame features to obtain a $D$-dimensional feature vector and perform sub-action classification over this feature vector.

3. Optional: you can propose your own feature representation.

4. Write a report of at most 1 page about your experiments and present your results. There are no restrictions regarding the framework/language that you can use for developing your models.

# References

[1] Joao Carreira and Andrew Zisserman. "Quo vadis, action recognition? a new model and the kinetics dataset". In: *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 6299–6308 (cit. on p. 1).

[2] Alexis Conneau et al. "Supervised learning of universal sentence representations from natural language inference data". In: *arXiv preprint arXiv:1705.02364* (2017) (cit. on p. 1).

[3] Hilde Kuehne, Ali Arslan, and Thomas Serre. "The language of actions: Recovering the syntax and semantics of goal-directed human activities". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014, pp. 780–787 (cit. on p. 1).