where and when accounts: action recognition in videos

Sha Hu, Yifang Fu

Abstract:

Videos usually contain redundant information that introduces disturbance for action recognition. The redundant information comes from two ways. The first is temporal-level redundancy: for a video, some frames have little relevance to the action. The second is spatial-level redundancy: in a frame, some regions have nothing to do with the action.
We believe decreasing the disturbance and only attending to the relevant information has the potential to boost accuracy for video action recognition. In this work, we plan to design a attention agent that can decide **where and when** to focus on.

Initial proposal for the spatiotemporal attention agent:

The proposed model leverages deep features extracted from Convolution NN's last layer. For every frame, we divide it into $K \times K$ grids. A spatial attentional agent will compute the weighted features for each frame, then the weighted features are put into a LSTM. A temporal attentional agent will compute the weighted output of the LSTM and obtain an action label.