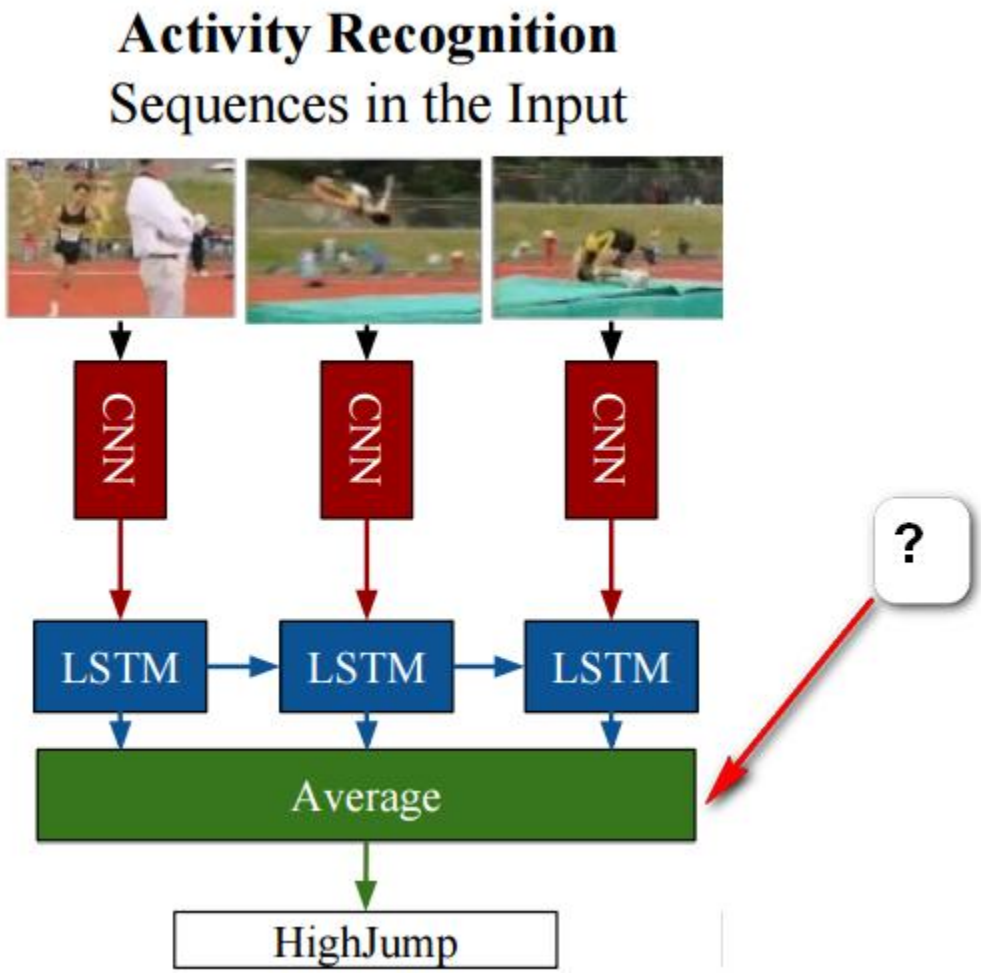
where and when accounts: action recognition in videos



In this model, they assumes that every frame contributes equally to the action label. Not reasonable.

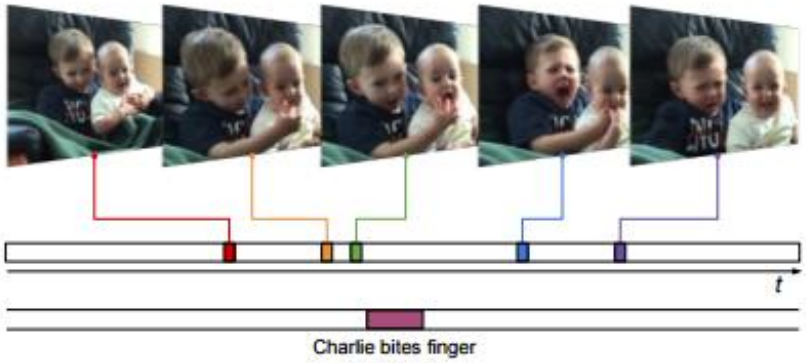
Model from [1]

[1]Donahue J, Anne Hendricks L, Guadarrama S, et al. Long-term recurrent convolutional networks for visual recognition and description[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 2625-2634.

A dog is standing on a hardwood floor.

A group of people sitting on a boat in the water.

Visualization of spatial attention from[2]



Visualization of temporal attention from[3]

[2]Xu K, Ba J, Kiros R, et al. Show, attend and tell: Neural image caption generation with visual attention[J]. arXiv preprint arXiv:1502.03044, 2015, 2(3): 5.
[3] Yeung S, Russakovsky O, Mori G, et al. End-to-end Learning of Action Detection from Frame Glimpses in Videos[J]. arXiv preprint arXiv:1511.06984, 2015.

Comparison

- Action localization for videos:

  Localize the action precisely in each frame
- Attention agent could be consider as an implicit focus on some frame and regions, while action localization can be considered as an explicit representation of the focus.