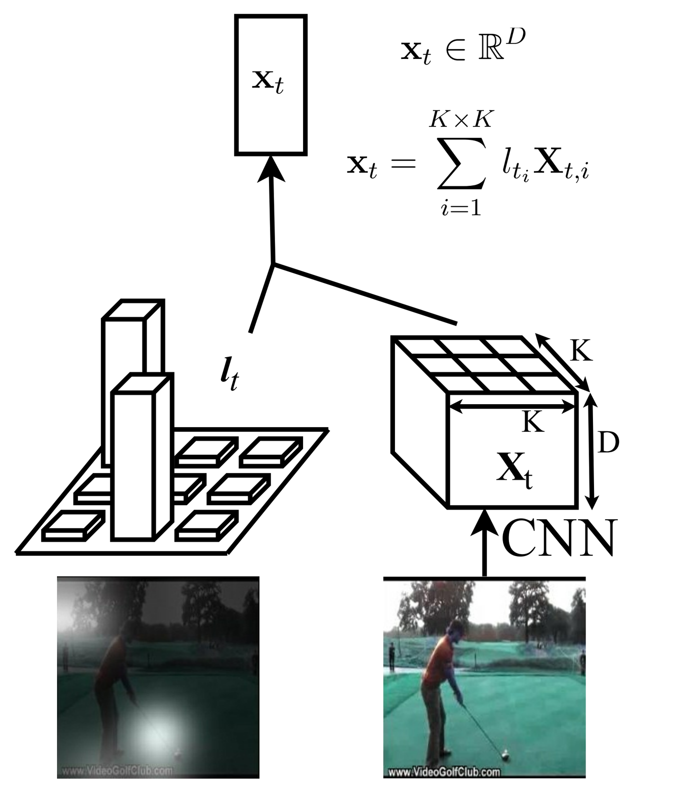
# Action Recognition using Visual Attention

Shikhar Sharma, Ryan Kiros and Ruslan Salakhutdinov
Department of Computer Science, University of Toronto

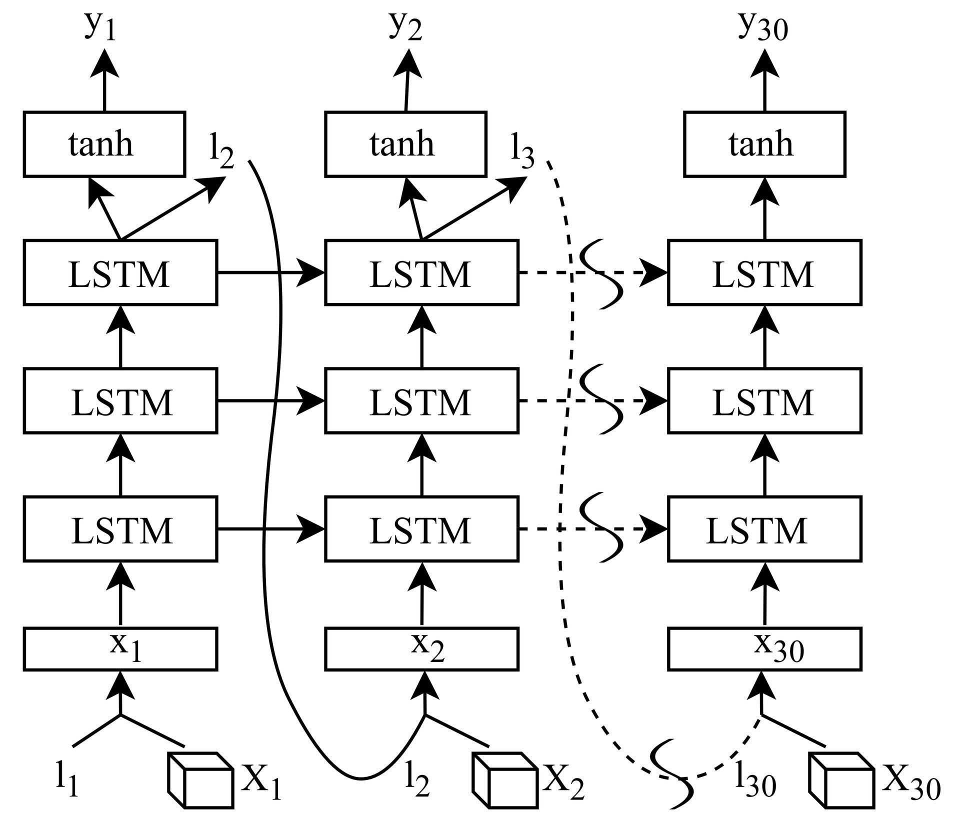
## Motivation

- Attention based models have been shown to achieve promising results on several challenging tasks, including caption generation [9], machine translation [1], game-playing and tracking [4].
- Attention based models can potentially infer the action happening in videos by focusing only on the relevant places in each video frame.
- Soft-attention models are deterministic and can be trained using backpropagation.
- We propose a soft-attention based model for action recognition in videos.
- We use multi-layered Recurrent Neural Networks (RNNs) with Long Short-Term Memory (LSTM).
- Our model tends to recognize important elements in video frames based on the activities it detects.

## The Attention Mechanism and the Model




(a) The soft-attention mechanism




(b) Our recurrent model

- We extract the last GoogLeNet [8] convolutional layer for the video frames.
- The last convolutional layer is a feature cube of shape $K \times K \times D$ ($7 \times 7 \times 1024$ here).
- Feature slices: the $K^2$ $D$-dimensional vectors within a feature cube.

$$\mathbf{X}_t = [\mathbf{X}_{t,1}, \ldots, \mathbf{X}_{t,K^2}], \qquad \mathbf{X}_{t,i} \in \mathbb{R}^D.$$

- Each of these $K^2$ vertical feature slices maps to different overlapping regions in the input space and our model chooses to focus its attention on these $K^2$ regions.
- The location softmax $\mathbf{l}_t$ over $K^2$ locations is:

$$l_{t,i} = p(\mathbf{L}_t = i|\mathbf{h}_{t-1}) = \frac{\exp(W_i^\top \mathbf{h}_{t-1})}{\sum_{j=1}^{K\times K}\exp(W_j^\top \mathbf{h}_{t-1})} \qquad i \in 1 \ldots K^2$$

where $W_i$ - the weights mapping to the $i^{th}$ element of the location softmax
$\mathbf{L}_t$ - a random variable which can take 1-of-$K^2$ values
$\mathbf{h}_{t-1}$ - the hidden state at time-step $t-1$

- The soft attention mechanism computes the expected value of the input at the next time-step $\mathbf{x}_t$:

$$\mathbf{x}_t = \mathbb{E}_{p(\mathbf{L}_t|\mathbf{h}_{t-1})}[\mathbf{X}_t] = \sum_{i=1}^{K\times K} l_{t,i}\mathbf{X}_{t,i}$$

where $\mathbf{X}_t$ - the feature cube at time-step $t$
$\mathbf{x}_t$ - the input to the LSTM at time-step $t$

## Loss Function

- Loss function: Cross-Entropy loss coupled with the doubly stochastic penalty introduced in [9]:

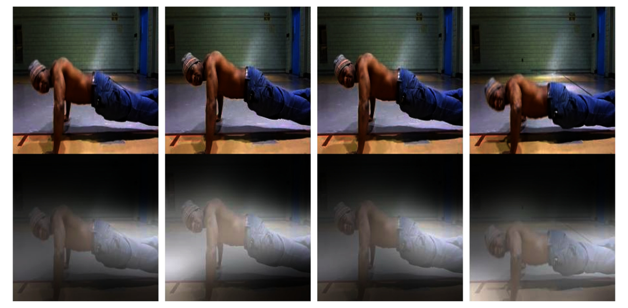
$$L = -\sum_{t=1}^{T}\sum_{i=1}^{C} y_{t,i}\log \hat{y}_{t,i} + \lambda \sum_{i=1}^{K\times K}(1 - \sum_{t=1}^{T} l_{t_i})^2 + \gamma \sum_i \sum_j \theta_{i,j}^2,$$

where $y_t$ - one hot label vector
$\hat{y}_t$ - vector of class probabilities at time-step $t$
$T$ - total number of time-steps
$C$ - number of output classes
$\lambda$ - attention penalty coefficient

## UCF-11, HMDB-51 and Hollywood2: Qualitative Analysis

**Success cases**: Figure (a)-(i)

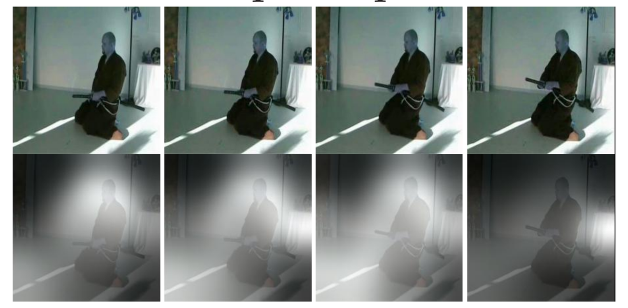
(a) cycling



(b) pushup

(c) DriveCar

(d) walking with a dog



(e) draw_sword

(f) Kiss

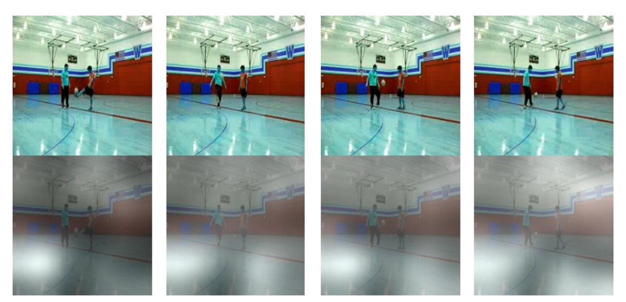
(g) push

(h) trampoline jumping

(i) golf swinging

**Failure cases**: Figure (j)-(l)

(j) "kick_ball" misclassified as "somersault"

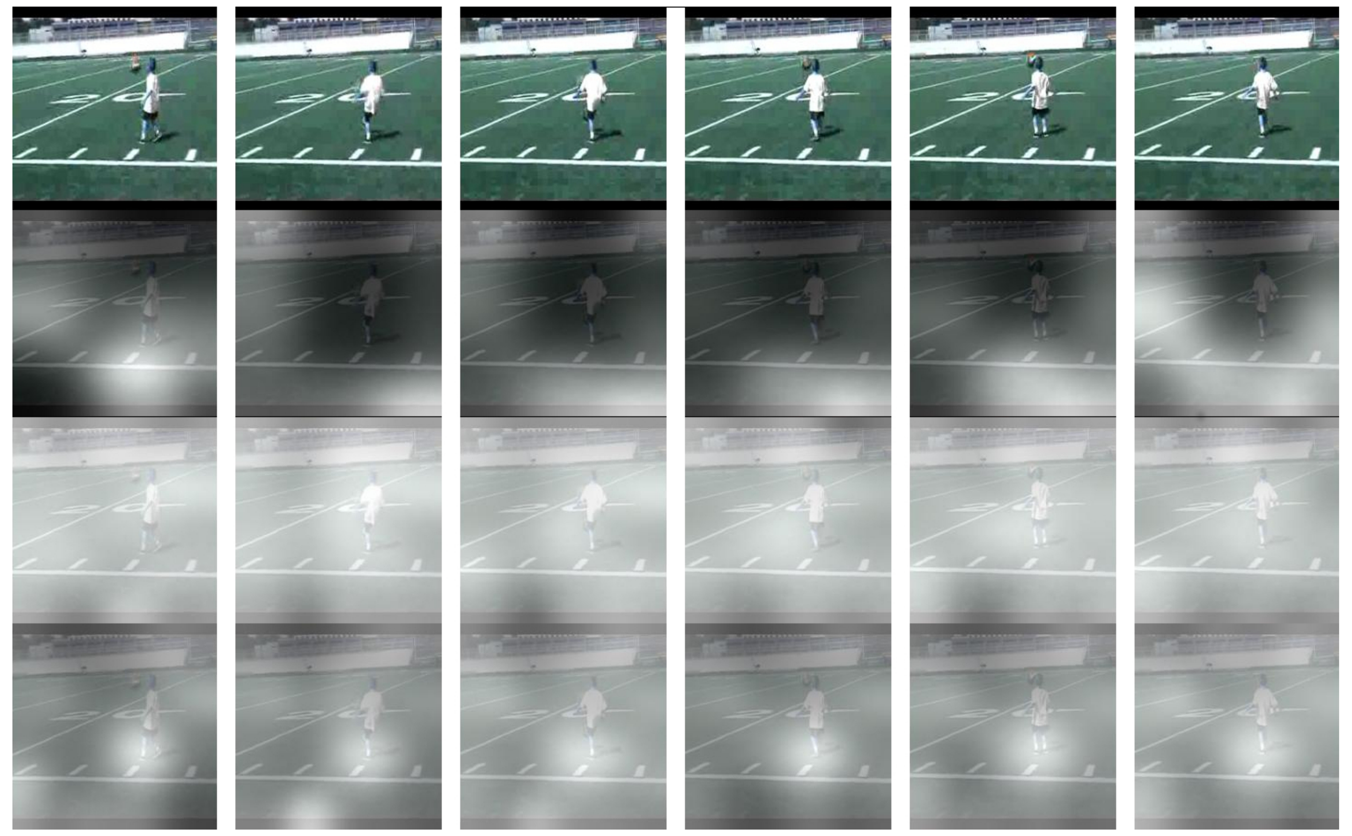


(k) "soccer juggling" misclassified as "diving"

(l) "flic_flac" misclassified as "hit"

- We can see that to classify the corresponding activities correctly, the model focuses on
  - Fig.(a): parts of the cycle
  - Fig.(b): the person doing push-ups
  - Fig.(c): the steering wheel, the rear-view mirror
  - Fig.(d): the dogs and the person
- Among failure cases:
  - Fig.(j): the model misclassifies the example despite attending to the relevat location
  - Fig.(l): the model misclassifies the example and does not even attend to the relevant location

## Optimizing Attention with the Correct Labels



**(First)**
The original video frames

**(Second)**
Failure case of model
Prediction: tennis swinging

**(Third)**
Random initialization
Prediction: tennis swinging

**(Fourth)**
Attention after optimization
Prediction: soccer juggling

## UCF-11, HMDB-51 and Hollywood2: Quantitative Analysis

**Table 1:** Performance on UCF-11 (acc %), HMDB-51 (acc %) and Hollywood2 (mAP %)

| Model | UCF-11 | HMDB-51 | Hollywood2 |
|---|---|---|---|
| Softmax Regression (full CNN feature cube) | 82.37 | 33.46 | 34.62 |
| Avg pooled LSTM (@ 30 fps) | 82.56 | 40.52 | 43.19 |
| Max pooled LSTM (@ 30 fps) | 81.60 | 37.58 | 43.22 |
| Soft attention model (@ 30 fps) | 84.96 | 41.31 | 43.91 |

**Table 2:** Comparison of performance on HMDB-51 and Hollywood2 with state-of-the-art models

| Model | | HMDB-51 (acc %) | Hollywood2 (mAP %) |
|---|---|---|---|
| Spatial stream ConvNet | [5] | 40.5 | - |
| Soft attention model | (Our model) | 41.3 | 43.9 |
| Composite LSTM Model | [6] | 44.0 | - |
| DL-SFA | [7] | - | 48.1 |
| VideoDarwin | [2] | 63.7 | 73.7 |
| Objects+Traditional+Stacked Fisher Vectors | [3] | 71.3 | 66.4 |

- We have divided Table 2 into three sections:
  - First section: models using only RGB data
  - Second section: models using both RGB and optical flow data
  - Third section: models using RGB, optical flow and object responses on some ImageNet categories
- Using hybrid soft and hard attention models in the future can potentially reduce computation cost and allow scaling to larger datasets like Sports-1M.

## References

[1] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *ICLR*, 2015.
[2] B. Fernando, E. Gavves, J. Oramas, A. Ghodrati, and T. Tuytelaars. Modeling video evolution for action recognition. In *CVPR*, 2015.
[3] M. Jain, J. C. v. Gemert, and C. G. M. Snoek. What do 15,000 object categories tell us about classifying and localizing actions? In *CVPR*, June 2015.
[4] V. Mnih, N. Heess, A. Graves, and K. Kavukcuoglu. Recurrent models of visual attention. In *NIPS*, 2014.
[5] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*. 2014.
[6] N. Srivastava, E. Mansimov, and R. Salakhutdinov. Unsupervised learning of video representations using LSTMs. *ICML*, 2015.
[7] L. Sun, K. Jia, T. Chan, Y. Fang, G. Wang, and S. Yan. DL-SFA: deeply-learned slow feature analysis for action recognition. In *CVPR*, 2014.
[8] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, 2015.
[9] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. *ICML*, 2015.