# Review of "Long-term recurrent convolutional networks for visual recognition and description"

Riyad Bin Rafiq

## 1. Paper summary

The paper [1] proposed a novel combination of CNN and LSTM architecture for video recognition and image description and retrieval tasks. The authors named the model as a long-term recurrent convolutional network (LRCN). Moreover, the model can be trained in an end-to-end approach. The spatially and temporally deep LRCN model outperformed the state-of-the-art models for recognition and generation tasks.

## 2. Contribution

### 2.1 Novel architecture

The authors proposed a novel model named LRCN. In this modeling approach, the authors used a CNN as a visual feature extractor and then jointly a recurrent model was used for processing the sequential data. The whole model was implemented for mapping pixels to sentence level natural language descriptions.

### 2.2 Three different problems

The proposed architecture was applied for three different types of vision problems. Firstly, the frames of video (activity recognition) were used to identify the label of the activity. The second one was an image description problem where a single image was used as input to generate the associated description. The last one was a video description task which was the first application of deep models. In my opinion, the research made a significant contribution by applying their proposed architecture to these three different kinds of problems with satisfactory results.

### 2.3 Evaluation: Datasets and existing methods

The evaluation of the LRCN was done on different datasets and the results were compared to various existing methods. For example, Flickr30k and COCO2014 datasets were used for image description tasks and the authors demonstrated the performance varying results among existing strategies such as m-RNN, DeFrag, ConvNet and so on.

## 3. Critique

### 3.1 Why that specific LSTM unit?

A lot of different LSTM units were available as the research on LSTMs had progressed. The authors chose a specific modified LSTM unit [2] but they didn't mention the reason why they particularly selected it. Would the results change if a different modified LSTM unit was chosen? This specific idea should be clear in the research.

### 3.2 T convolutional networks for activity recognition?

In the activity recognition task, the authors stated that T individual frames were inputs into T convolutional networks but it was not clear afterwards at all. The research used 16 frames in this work, so the number of convolutional networks would become 16 followed by the LSTM units and how did it work ? These things were not described thoroughly.

## Reference


1. Donahue J, Anne Hendricks L, Guadarrama S, Rohrbach M, Venugopalan S, Saenko K, et al. Long-term recurrent convolutional networks for visual recognition and description. Proceedings of the IEEE conference on computer vision and pattern recognition. 2015. pp. 2625–2634.

2. Zaremba W, Sutskever I. Learning to Execute. arXiv [cs.NE]. 2014. Available: http://arxiv.org/abs/1410.4615