**Paper: Long-Term Recurrent Convolutional Networks for Visual Recognition and Description**

Name: Milon Hossain

ID: 1610497042

**Abstract:** Recent image interpretation tasks have been controlled by models centered on deep convolutional networks. Author of the project investigate whether models that are also recurrent or temporarily deep are successful for tasks involving sequences, visual interpretation and so on. Built a modern, recurrent and convolutional architecture that end-to-end training optimal for large-scale visual learning and demonstrate the importance of these models on benchmark video recognition task, image definition problems with retrieval and video narration and challenges. Such model may have advantages when target concepts are complex and training data are limited. Learning long term dependencies is possible when nonlinearities are incorporated into the network state updates. Long-term RNN models are appealing in that they directly can map variable-length inputs to variable length outputs and can model complex temporal dynamics, yet they can be optimized with backpropagation.

**Introduction:** Recognition and description of images and videos is a fundamental challenge of computer vision. A supervised convolutional model on image recognition tasks have made dramatic improvement, and a variety of modifications have recently been suggested to process imagery. Ideally, a video model should allow input sequence of variable length to be processed, and also account for outputs of variable length. This involves the generation of full-length explanations of sentences that go beyond traditional one-on-all prediction tasks. Suggested a Long-Term recurrent convolutional network (LRCNs), a modern architecture. Recognition and definition that incorporates convolutional and long-range temporal recursion layers and its trainable end-to-end. RNN models are well known to be “deep in time” explicitly so when unrolled, and form implicit compositional representations in the time domain. Such deep models predated deep spatial convolutional models in the literature.

**Recurrent Neural Network (RNN):** Traditional RNN can learn complex temporal dynamics by mapping input sequences to a sequence of hidden states, and hidden states to output via the following recurrence equation:

ht = g(Wxhxt +Whhht\_1 + bh)

zt = g(Whzht + bz)

Where g is an element-wise non-linearity, such as a sigmoid or hyperbolic tangent, Xt is the input ht belongs to Rn is the hidden state with N hidden units, and Yt is the output at a time. For a length T input sequence, The updates above are compared sequentially as h1,y1,h2y2……..hTyT. Though RNNs have proven successfully on tasks such as speech recognition and text generation, it can be difficult to train them to learn long-term dynamics.

**Long-Term Recurrent Convolutional Network (LRCN):** This work proposes a Long-Term Recurrent Convolutional Network (LRCN) model combining a deep hierarchical visual feature extractor (Such as CNN) with a model that can learn to recognize and synthesize temporal dynamics for tasks involving sequential data visual, linguistical and otherwise. Here the LRCN model works by passing each visual input through a feature transformation. Three vision problems which instantiate one of the following broad classes sequentially learning tasks are considered in this project.

1. Sequential inputs, fixed outputs.
2. Fixed inputs, Sequential outputs.
3. Sequential inputs and sequential outputs.

**Long-Short Term Memory (LSTM**): The LSTM framework allows to model the video as a variable length input stream as discussed. Mainly three different approaches are distinguished in this project.

1. LSTM encoder and decoder with CRF max.
2. LSTM decoder with CRF max.
3. LSTM decoder with CRF Prob.

**Dataset:** Flicker30k is used as a dataset here. Additionally, report results on the new COCO2014 dataset which has 80,000 training images and 40,000 validation images. Similar to flicker30k each image annotated with 5 or more image annotation.

**Conclusion:** The LRCN, a class of models that is both the spatially and temporally deep, and has the flexibility to be applied to a variety of vision tasks involving sequential inputs and outputs. As the field of computer vision mature beyond tasks with static input and predictions, they envision “doubly deep” sequence modeling tools like LRCN will soon become central pieces of most vision systems and convolutional architecture.

Reference:

1. <https://openaccess.thecvf.com/content_cvpr_2015/html/Donahue_Long-Term_Recurrent_Convolutional_2015_CVPR_paper.html>