CGRU

1. What is percepts? How authors used it in the network?

Solution: By utilizing Gated-Recurrent-Unit Recurrent Networks, we provide a method for learning spatio-temporal properties in films from preliminary visual representations we call to as "percepts" (GRUs).

Authors used it in the network such as-

- Used to extract from all levels of a deep convolutional network trained on the large ImageNet dataset.
- Can lead to high-dimensionality video representation.

2. How to define finer motion pattern using low level percepts?

Solution: We can build finer motion patterns from low-level senses because they maintain a higher spatial resolution. Yet, a high-dimensional representation of the video can be produced by using low-level perception. A GRU model variation was proposed that forces the sparse connectivity of model units via a convolution operation in order to lessen this impact and manage the amount of parameters.

3. What is the difference of GRU-RCN and Stacked GRU-RCN?

Solution: Stacked GRU-RCN sets each GRU-RNN to the output of the prior GRU-RNN in the current time step, whereas GRU-RCN applies each GRU-RNN level independently. A second input for GRU convolutional units is the prior hidden RNN representation. Further flexibility is provided by adding this extra link, which also enables the model to use representations with various resolutions. Performance of Stacked-GRU RCN is much worse than GRU-RCN. We contend that the Stacked-GRU RCN learning is challenging due to the bottom-up connection, deepening of the model, and scarcity of training data. The performance of the GRU-RCN is better for the Bi-directional GRU-RCN.

4. What the authors used to extract temporal pattern from visual?

Solution: The authors added recurrent convolutional units to pre-trained CNN convolution maps in order to distinguish temporal patterns from visual "perceptions" of varied spatial sizes. To combine the data obtained from these convolutional maps, they proposed GRU-RCN and Stacked GRU-RCN as two distinct RCN architectures.

5. What data-sets the authors used for the experiment listed them below.

Solution: Using the UCF-101 and YouTube2Text datasets, the authors empirically validated their strategy for human action identification and video captioning tasks.