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

Ans : Percept is an approach to learn the spatio-temporal functions present in the video through intermediate visual representations and GRU.

1. The creation method is based on observations extracted from all layers of a deep convolution network trained on the large ImageNet dataset.

2. The approach proposes to use visual "percepts" extracted from different levels in CNN 2D.

3. They use recurrent convolutional units on pre-trained CNN convolutional maps to extract temporal models from visual "percepts" of different spatial dimensions.

1. **How to define finer motion patterns using low level precepts?**

Ans : Low-level perceptions preserve a higher spatial resolution from which we can model finer motion patterns. However, using low-level perception can result in a high-dimensional representation of the video. To mitigate this effect and control the number of parameters, a variant of the GRU model was introduced that utilizes a convolution operation to force sparse connectivity of model units.

1. **What is the difference between GRU-RCN and Stacked GRU-RCN?**

Ans : While GRU-RCN applies each GRU-RNN level independently, Stacked GRU-RCN sets each GRU-RNN to the output of the previous GRU-RNN in the current time step. The previous hidden RNN representation is provided as an additional input for GRU convolutional units. Adding this additional connection offers more flexibility and allows the model to use representations with different resolutions.

1. **What did the authors use to extract temporal patterns from visuals?**

Ans: The authors used recurrent convolutional units on pre-trained CNN convolution maps to extract temporal patterns from visual "perceptions" of different spatial sizes.They proposed two alternative RCN architectures, GRU-RCN and Stacked GRU-RCN, to combine information extracted from these convolutional maps.

1. **What datasets the authors used for the experiment listed them below.**

Ans : The authors empirically validated their approach to human action recognition and video caption activities using the UCF-101 and YouTube2Text datasets.