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

Ans : Percept is an approach to learn the spatio-temporal features that are present in the video by using intermediate visual representations & GRU.

1. The authors method relies on percepts that are extracted from all levels of a deep convolutional network trained on the large ImageNet dataset.

2. The approach proposes to leverage visual “percepts” extracted from different layers in the 2D-CNN.

3. They use recurrent convolutional units on pre-trained CNN convolutional maps, to extract temporal patterns from visual “percepts” with different spatial sizes.

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

Ans : Low-level percepts preserve a higher spatial resolution from which we can model finer motion patterns.

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

Ans : While GRU-RCN applies each layer-wise GRU-RNN in an independent fashion, Stacked GRU-RCN preconditions each GRU-RNN on the output of the previous GRU-RNN at the current time step. The previous RNN hidden representation is given as an extra input to the GRU convolutional units. Adding this extra-connection brings more flexibility and gives the opportunity for the model to leverage 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 convolutional maps, to extract temporal patterns from visual “percepts” with different spatial sizes.

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

Ans : They have empirically validated our approach on the Human Action Recognition and Video Captioning tasks using the UCF-101 and YouTube2Text datasets.