A video consists of an ordered sequence of frames. Each frame contains *spatial* information, and the sequence of those frames contains *temporal* information. To model both of these aspects, we use a hybrid architecture that consists of convolutions (for spatial processing) as well as recurrent layers (for temporal processing). Specifically, we'll use a Convolutional Neural Network (CNN) and a Recurrent Neural Network (RNN) consisting of GRU layers. This kind of hybrid architecture is popularly known as a **CNN-RNN**.

A GRU layer learns dependencies between time steps in time series and sequence data.

GRUs and LSTMs utilize different approaches toward gating information to prevent the vanishing gradient problem. Here are the main points comparing the two:

- The GRU unit controls the flow of information like the LSTM unit, but without having to use a ***memory unit***. It just exposes the full hidden content without any control.
- GRUs are relatively new, and in my experience, their performance is on par with LSTMs, but computationally ***more efficient*** (*as pointed out, they have a less complex structure*). For that reason, we are seeing it being used more and more.