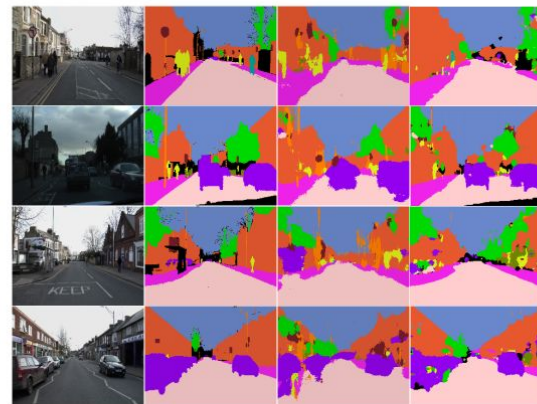
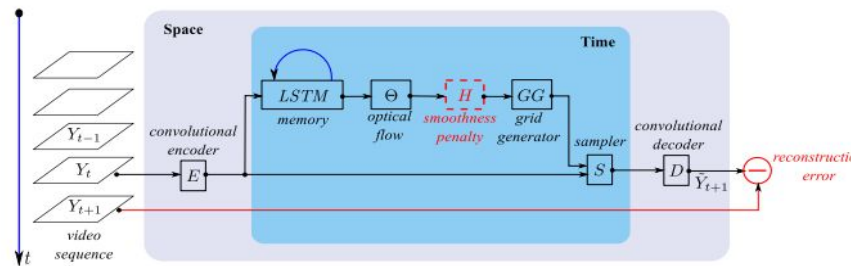


# Deep learning on Video Analysis

# Spatio-temporal video autoencoder with differentiable memory

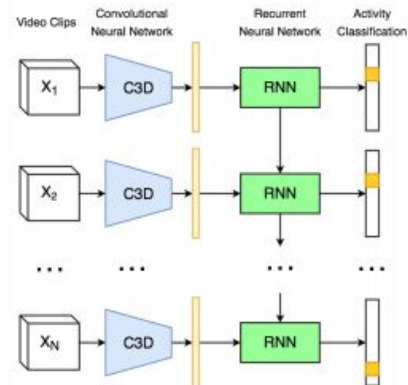
- The autoencoder is based on a classic spatial image autoencoder and a novel nested temporal autoencoder
- The temporal encoder is composed of convolutional long short-term memory (LSTM) cells that integrate changes over time.
- The system predicts the optical flow based on the current observation and the LSTM memory state and applies it to the current frame to generate the next frame.
- By minimising the reconstruction error between the predicted next frame and the corresponding ground truth next frame, the whole system is trained in unsupervised manner.



Code: <https://github.com/viorik/ConvLSTM>

# Temporal Activity Detection in Untrimmed Videos with Recurrent Neural Networks

- A simple pipeline to classify and temporally localize activities in untrimmed videos.
- It uses features from a 3D Convolutional Neural Network (C3D) as input to train a recurrent neural network (RNN) that learns to classify video clips of 16 frames.
- After clip prediction, the output of the RNN is post-processed to assign a single activity label to each video, and determine the temporal boundaries of the activity within the video



Video ID: ArzhjEk4j\_Y  
Ground Truth: Building sandcastles

Prediction:  
0.7896 Building sandcastles  
0.0073 Doing motocross  
0.0049 Beach soccer

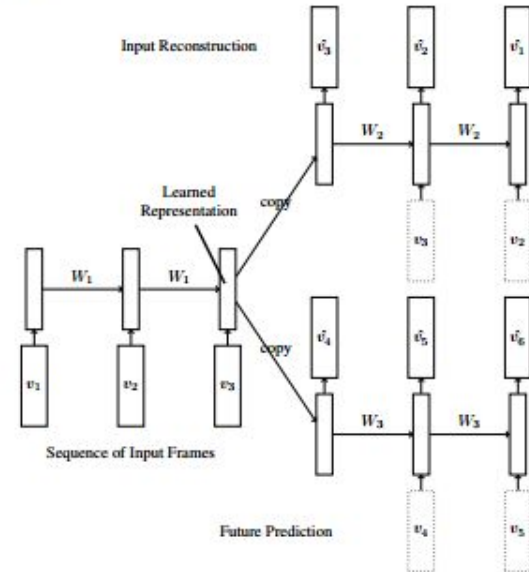


Video ID: AimG8xzchfI  
Activity: Curling

Prediction:  
0.3843 Shoveling snow  
0.1181 Ice fishing  
0.0633 Waterskiing

# Unsupervised Learning of Video Representations using LSTMs

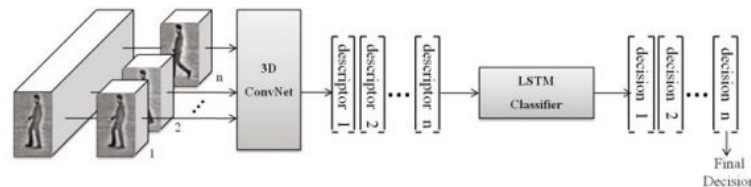
- LSTM network is used to learn representations of video sequences.
- The model uses an encoder LSTM to map an input sequence into a fixed length representation.
- This representation is decoded using single or multiple decoder LSTMs to perform different tasks, such as reconstructing the input sequence, or predicting the future sequence.
- 2 kinds of input sequences – patches of image pixels and high-level representations (“percepts”) of video frames extracted using a pretrained convolutional net.



Code: <https://github.com/emansim/unsupervised-videos>

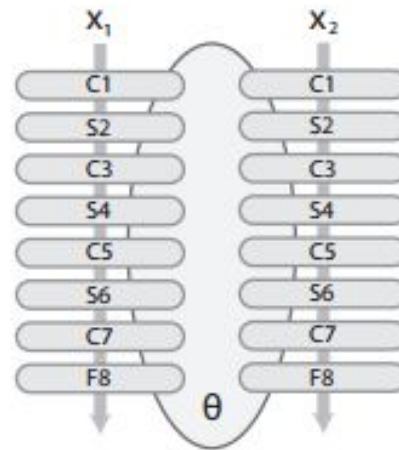
# Sequential Deep Learning for Human Action Recognition

- Convolutional Neural Networks to 3D, automatically learns spatio-temporal features.
- The features are automatically constructed with the 3D-ConvNet and the entire sequence is labeled based on the accumulation of several individual decisions corresponding each to a small temporal neighbourhood,
- An LSTM is then trained to classify each sequence considering the temporal evolution of the learned features for each timestep



# Deep Learning from Temporal Coherence in Video

- The system takes advantage of the fact that two successive frames are likely to contain the same object or objects.
- This coherence is used as a supervisory signal over the unlabeled data, and is used to improve the performance on a supervised task of interest.
- The cost(L1 norm) is minimized for all pairs of images by stochastic gradient descent over a “siamese network” architecture.



# Two-Stream Convolutional Networks for Action Recognition in Videos

- A two-stream ConvNet architecture which incorporates spatial and temporal networks.
- Bidirectional optical flow based on optical flow stacking and trajectory stacking is used as input.
- SVM-based fusion of softmax scores outperforms fusion by averaging.

