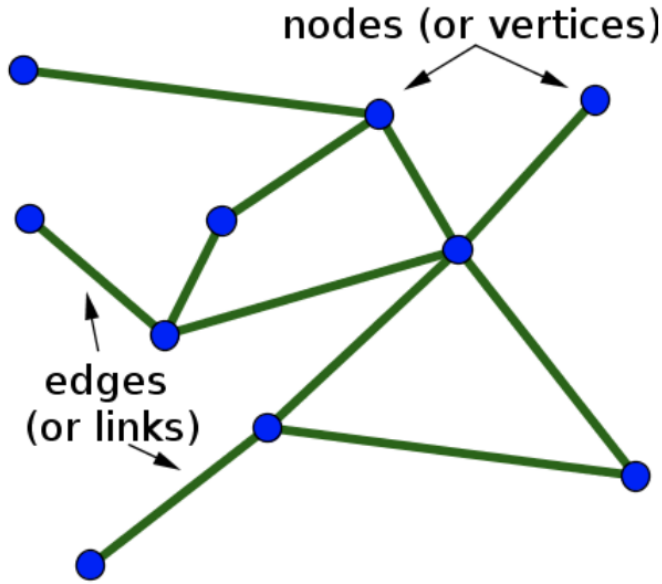
Generate the next few frames in a video using graph neural networks.

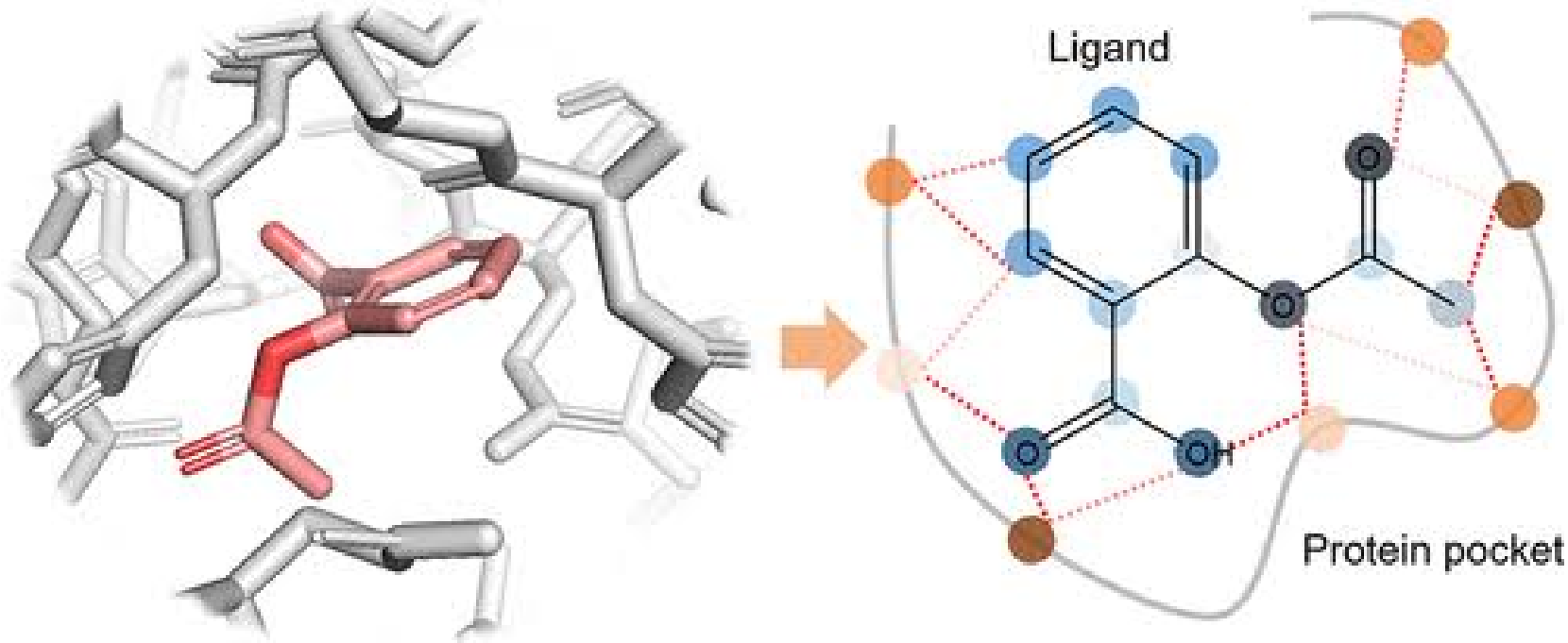
Shivam Raj
Shrishti Barethiya

22 October 2021

Goal

- To predict the next few frames of the video using a Graph Neural Network for spatio-temporal graphs.
- To know about the nodes a few frames later.


nodes (or vertices)
edges
(or links)


Ligand
Protein pocket

Relevent Papers

1. Bing Yu, Haoteng Yin, Zhanxing Zhu. Spatio -Temporal Graph Convolutional Networks: A Deep Learning Framework for Traffic Forecasting. 10.24963/ijcai.2018/505.
2. Bin Zhao, Haopeng Li, Xiaoqiang Lu, Xuelong Li. Reconstructive Sequence-Graph Network for Video Summarization. 10.1109/TPAMI.2021.3072117.
3. Rucha Bhalchandra Joshi, Subhankar Mishra. Learning Graph Representations.

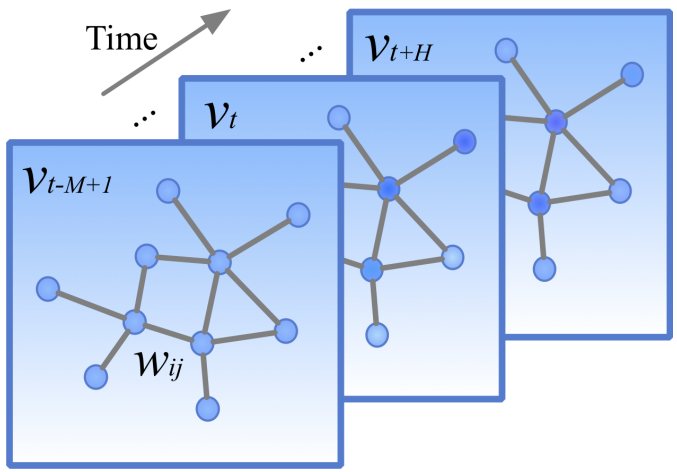
Frame Title

- We propose a novel deep learning framework, Spatio Temporal graph Convolutional network(STGCN), to tackle time series prediction problem.
- we purely apply convolutional structures to extract spatio-temporal features simultaneously from graph-structured in video prediction.
- The built model with complete convolutional structures, which enables much faster training speed with fewer parameters
- For spatial features, we uses convolutional neural network(CNN) to capture adjacent relation among the frame of the videos

Spatio -Temporal Graph Convolutional Networks

$$\hat{v}_{t+1}, ..., \hat{v}_{t+H} = \underset{v_{t+1},...,v_{t+H}}{\arg\max} \log P(v_{t+1}, ..., v_{t+H} | v_{t-M+1}, ..., v_t; G). \quad (1)$$

where $v_t \in \mathbb{R}^n$ is an observation vector of n image frames at time steps t.



Spatio -Temporal Graph Convolutional Networks

- Observation v_t not independent but connected by pairwise connection in graph.
- The data point v_t can be regarded as a graph signal that is defined on an undirected graph (or directed one) G with weights w_ij as shown in above Fig
- At the t-th time step, in graph $G_t = (V_t, \epsilon, \mathrm{W})$, V_t is a finite set of vertices, corresponding to the observations from n monitor stations in a traffic network; ϵ is a set of edges, indicating the connectedness between stations; while $\mathrm{W} \epsilon \mathbb{R}^{n \times n}$denotes the weighted adjacency matrix of G_t.

convolution on graph

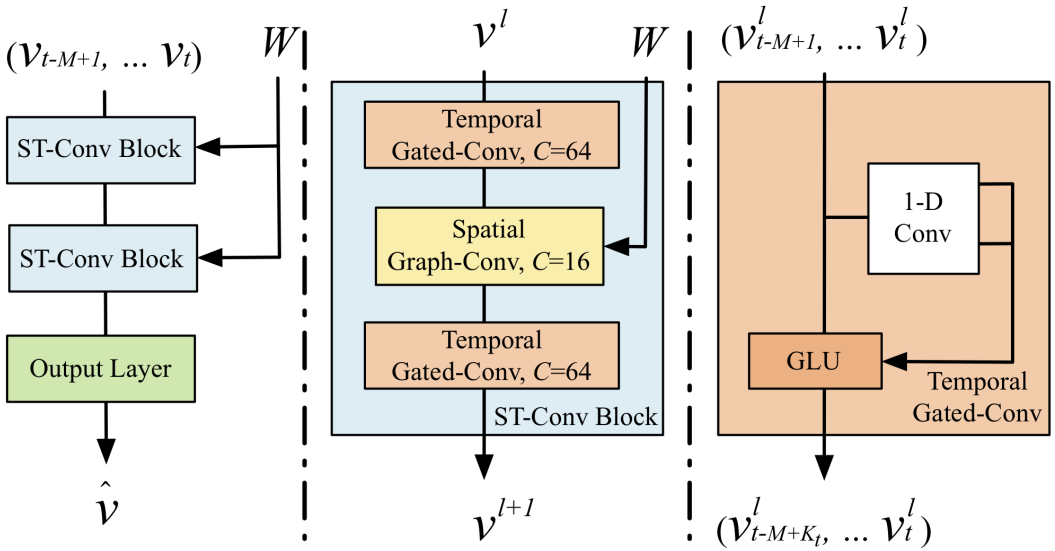
- Two basic approaches to generalize CNNs for structured data :-
  - spatial definition of a convolution
  - spectral domain with graph Fourier transforms
- Here notion of graph convolution operator "*G" , " based on the conception of spectral graph convolution, as the multiplication of a signal x $\epsilon \mathbb{R}^n$ with a kernel Θ,

$$\Theta_{*G} x = \Theta(L)x = \Theta(U \Lambda U^T)x = (U \Theta(\Lambda) U^T)x$$

where graph Fourier basis $U \epsilon \mathbb{R}^{n \times n}$ is the matrxi of eigenvectors of the normalized graph laplacian $L = I_n - D^{-\frac{1}{2}} W D^{-\frac{1}{2}} = U \Lambda U^T \epsilon \mathbb{R}^{n \times n}$

Spatio -Temporal Graph Convolutional Networks.

- STGCN is composed of several spatiotemporal convolutional blocks.



- The input v_{t-M+1}, ..., v_t is uniformly processed by ST-Conv blocks to explore spatial and temporal dependencies

STGCN

Main characteristics of our model STGCN are:-

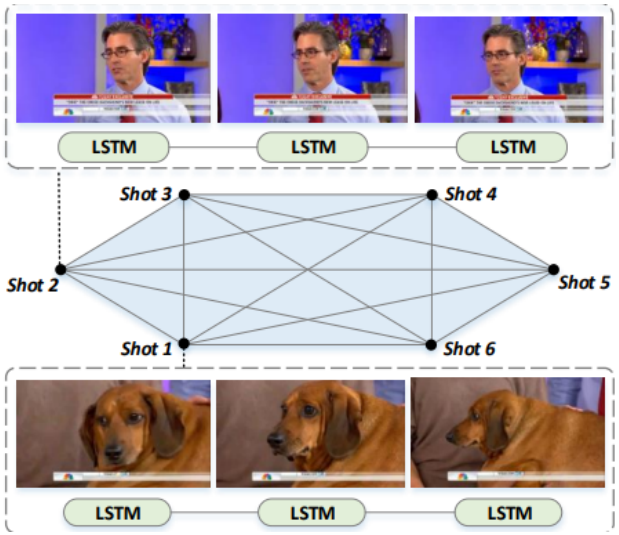
- STGCN is a universal framework to process structured time series. It is not only able to tackle image modeling and prediction issues but also to be applied to more general spatio-temporal sequence learning tasks.
- The spatio-temporal block combines graph convolutions and gated temporal convolutions, which can extract the most useful spatial features and capture the most essential temporal features coherently.
- The model is entirely composed of convolutional structures and therefore achieves parallelization over input with fewer parameters and faster training speed. More importantly, this economic architecture allows the model to handle large-scale networks with more efficiency.

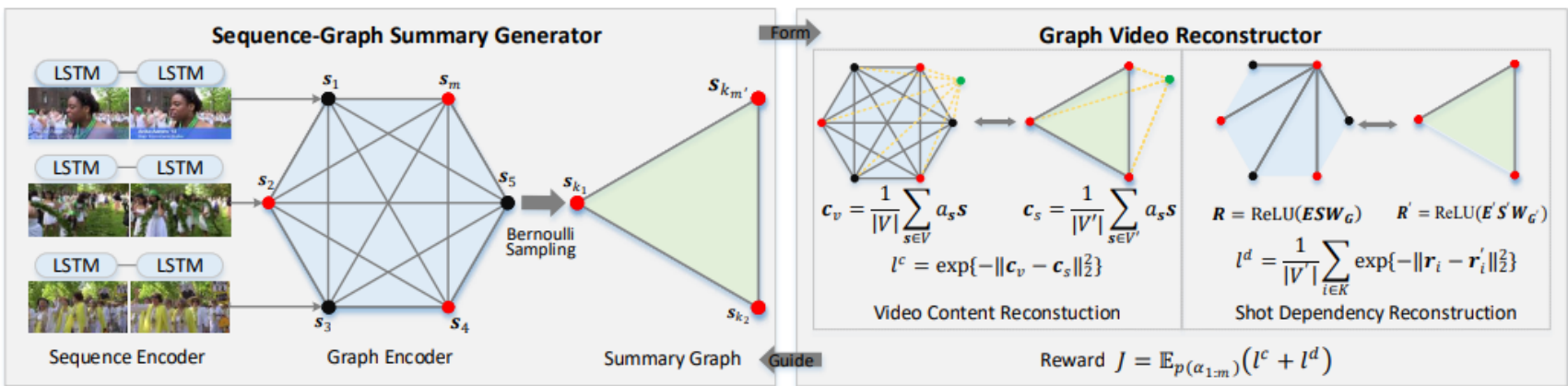
Future Frame Prediction of a Video Sequence

Here there are problem that there may be multiple future sequences possible for the same input video Two different approaches have attempted to address this problem:

- Use of latent variable models that represent underlying stochasticity.
- Adversarially trained models to aim to produce sharper images.

Reconstructive Sequence-Graph Network for Video Summarization





Datasets

We will be using these datasets:

- MNIST datasets: This dataset has randomly sampled two digits from the original MNIST dataset, floating and bouncing at boundaries at constant velocity and angle inside a 64×64 patch. New sequences can be generated as and when required making the dataset an almost infinite source of video sequences. 8000 videos are used for training and 2000 videos for testing
- UCF101: : This dataset contains 13320 annotated action videos. The videos have been accumulated from YouTube and has 101 different categories of action. The videos have a resolution of 320×240 pixels and a frame rate of 25 fps. 10000 videos are used for training and 3320 videos for testing. Data can be downloaded from here `https://www.crcv.ucf.edu/data/UCF101.php`
- Sports-1M Dataset: The Sports-1M dataset is licensed under Creative Commons 3.0 and contains 1,133,158 video URLs which have been annotated automatically with 487 Sports labels using the YouTube Topics API.

Future Work

- To implement these datasets to predict the next frame of the video using the proposed model and also to find the position of the nodes after some number of frames.

Thank You.