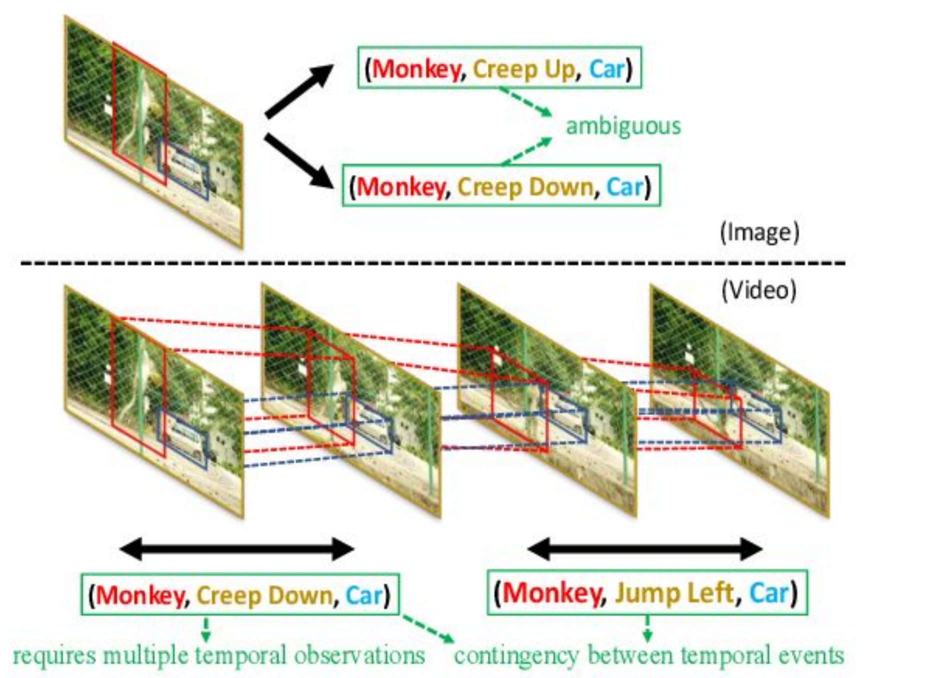
Visual Relationship Reasoning In Videos

Phalguna Mankar, Shubhankar Potdar
Electrical and Computer Engineering Department

Problem Statement

Temporal nature of Videos allows modeling of a more rich set of visual spatio-temporal features than images. This is useful for tasks like detecting relationships between entities interacting in a scene with each other. The led us to the task of VRRV. We analyse 2 methods for this task:
1. Two stream 3D ConvNet.
2. Gated energy graph on top of the baseline to model the spatial and temporal relations between the objects.



Related Work

The 2 main tasks in Visual Relationship in Videos:-

Action Recognition:
- Previous methods have used 2D ConvNets along with LSTMs to model the individual image features as well as the temporal features.
- 3D ConvNets were later used to model the spatial and motion features of the video.

Visual Relation Detection:
- The 1st approach detected individual objects in the video and tracked them across frames to extract object tracklet features as well as relative motion features.
- The tracklet and relative features were then concatenated and passed through individual classifiers for {subject, predicate, object} classification.

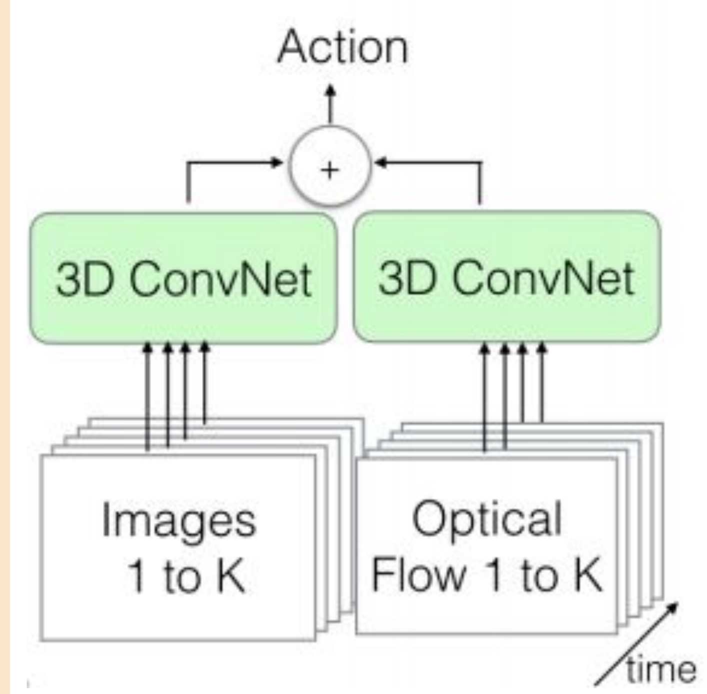
Dataset

- The original Charades dataset contains videos of human indoor activities (9, 848 in total with 7, 985 for training and the rest for evaluation,) . The visual relationship is defined on the triplet {verb, object, scene} or {object, verb, scene}. For example, {hold, blanket, bedroom}, {someone, cook, kitchen}, etc. It has 33 categories of verb, 38 objects and 16 scenes.
- Since the size of the dataset is huge, we have trained the model on 500 videos and tested on 200 videos.
- The results demonstrate the improvement a structured prediction model is able to obtain on the dataset over the baseline end to end deep learning based model.

Data Preprocessing

There are 2 sub parts of the datasets created:-
- The first is for fine grained relationship reasoning. This is also the dataset on which the model is trained. In this dataset each video is divided into segments of 10 overlapping frames. Each segments is tagged with its video ID, {object, verb} tuples along with their start and end frames.
- The second is for relationship reasoning on the entire scale of the video. In this, each video is divided into 25 segments of frames irrespective of the total number of frames per video.

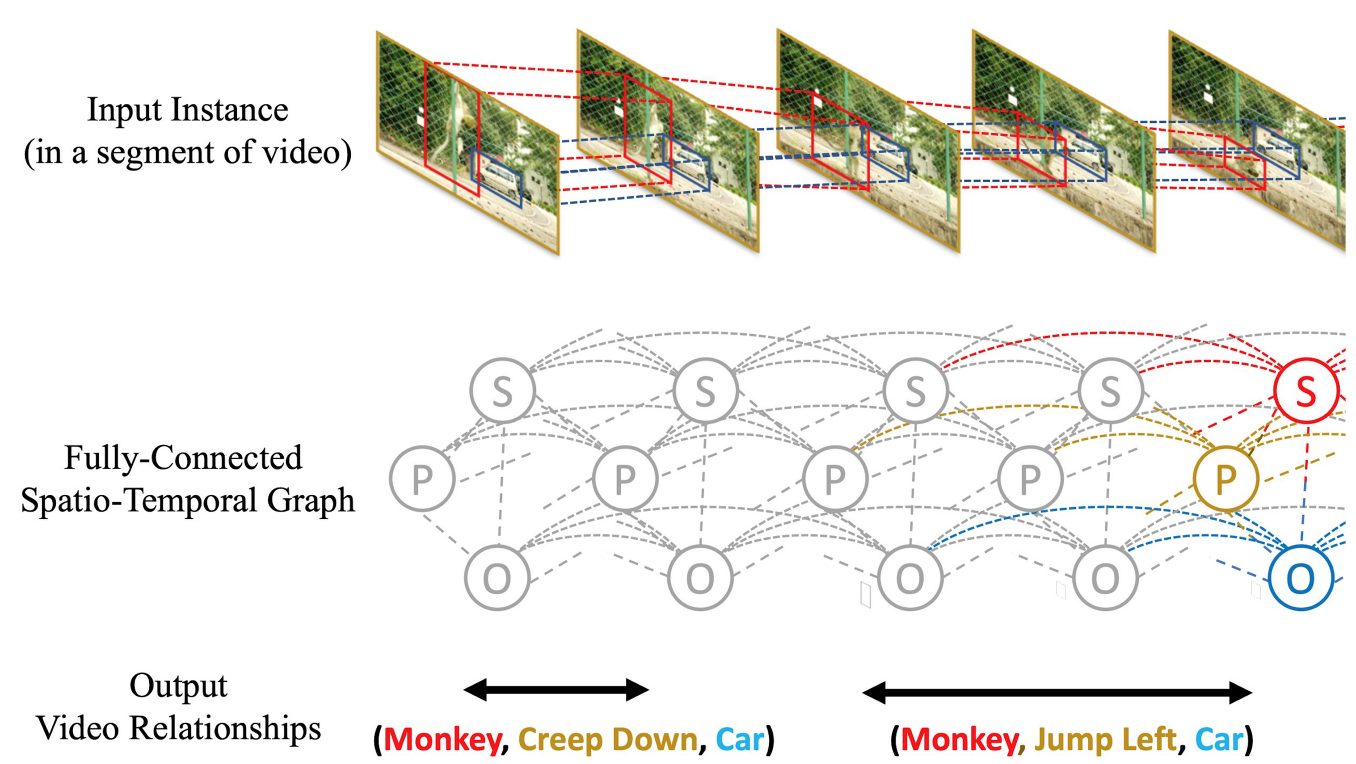
Baseline Approach



Inception I3D [4]

- Different from the technique used in other visual relationship works as in [2] and [3].
- It is inspired by the action recognition architecture suggested in [4].
- It consists of using a 2 stream network with one stream applying a 3D ConvNet on the video frames, and other is a optical flow stream.
- While the 3D ConvNet is itself good at learning motion features from RGB inputs directly, optical flow algorithms are recurrent and hence help in capturing temporal features.

Structured Prediction Approach



- We model the spatial and temporal structure of relationships in videos by constructing a fully-connected spatio-temporal graph.
- Each node in the graph represents an entity and the edges represent statistical relations between them.
- Much of the previous work assumes a predefined or globally-learned pairwise energy function. Replicating the work done by [3], we make use of a observation-gated version called the Gated Spatio-Temporal Energy Graph(GSTEG) that allows us to make the statistical dependency between entities (conditioned on the observation). We adopt mean field algorithm for doing inference on the CRF.

- **(CRF) parametrized by a Gibbs distribution:**

$$P\Big(Y = y|X\Big) = \frac{1}{Z(X)}\exp\Big(-\mathbf{E}(y|X)\Big)$$

- **The Energy function:**

$$\mathbf{E}_{\psi,\varphi}(y|X) = \sum_{t,k}\psi_{t,k}(y_t^k|X) + \sum_{\{t,k\}\neq\{t',k'\}}\varphi_{t,k,t',k'}(y_t^k, y_{t'}^{k'}|X)$$

- **New Pairwise Energy Function:**

$$\varphi_{t,k,t',k'}(y_t^k, y_{t'}^{k'}|X) := \langle f^{\varphi}\rangle_{X,t,t',k,k',y_t^k,y_{t'}^{k'}}$$

- **Observation Gated Pairwise Energy Function :**

$$\langle f^{\varphi}\rangle_{X,t,t',k,k',y_t^k,y_{t'}^{k'}} \approx f_\theta^{\varphi}(X_t^k, t, t', k, k', y_t^k, y_{t'}^{k'})$$

$$= \begin{cases} \Big\langle g_\theta^{kk'}(X_t^k) \otimes h_\theta^{kk'}(X_t^k)\Big\rangle_{y_t^k, y_{t'}^{k'}} & t = t' \\ K_\sigma\Big(t,t'\Big)\Big\langle r_\theta^{kk'}(X_t^k) \otimes s_\theta^{kk'}(X_t^k)\Big\rangle_{y_t^k, y_{t'}^{k'}} & t \neq t' \end{cases}$$

Outcomes (Object, Verb, Scene)

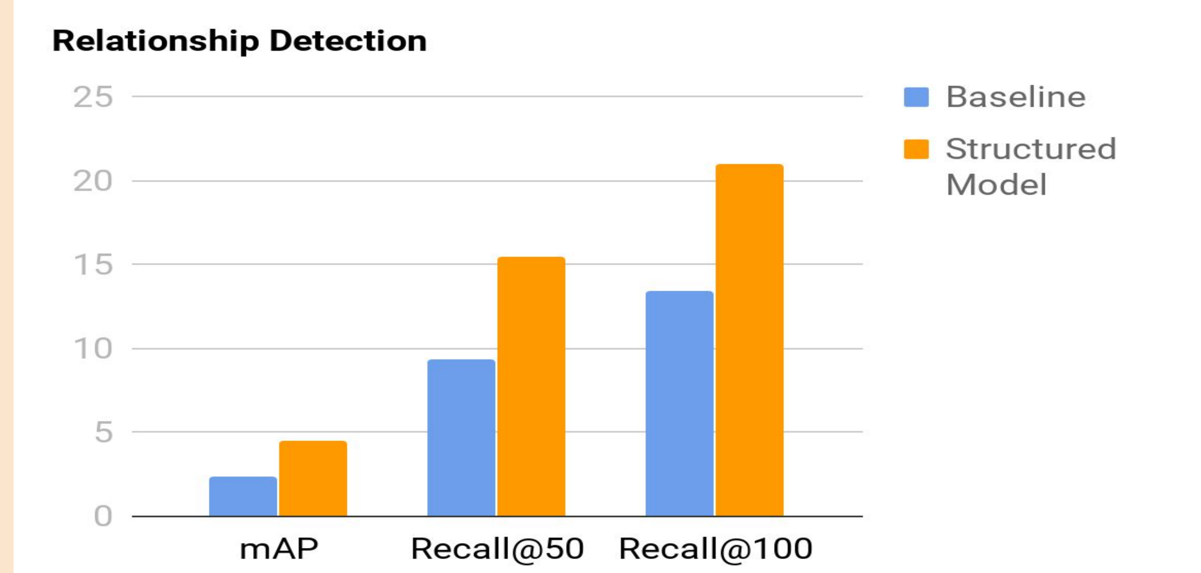
Baseline:	Food Hold Basement
GSTEG:	Food Eat Kitchen
True LB:	Food Eat Kitchen

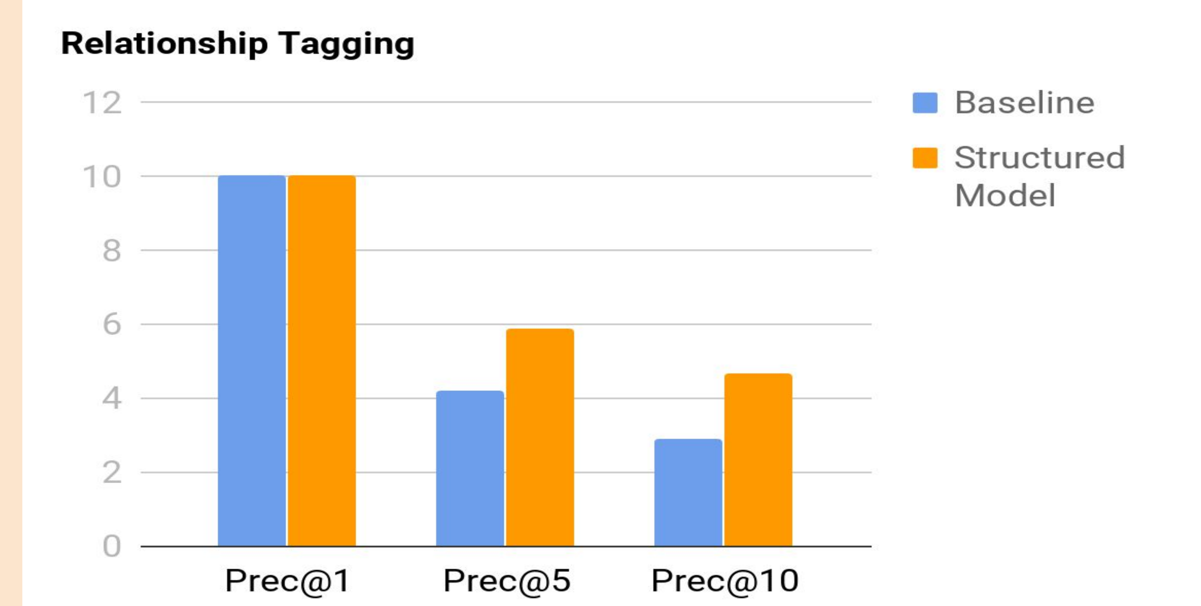
Baseline:	Food Hold Dining Room
GSTEG:	Dish Hold Kitchen
True LB:	Clothes Hold Kitchen

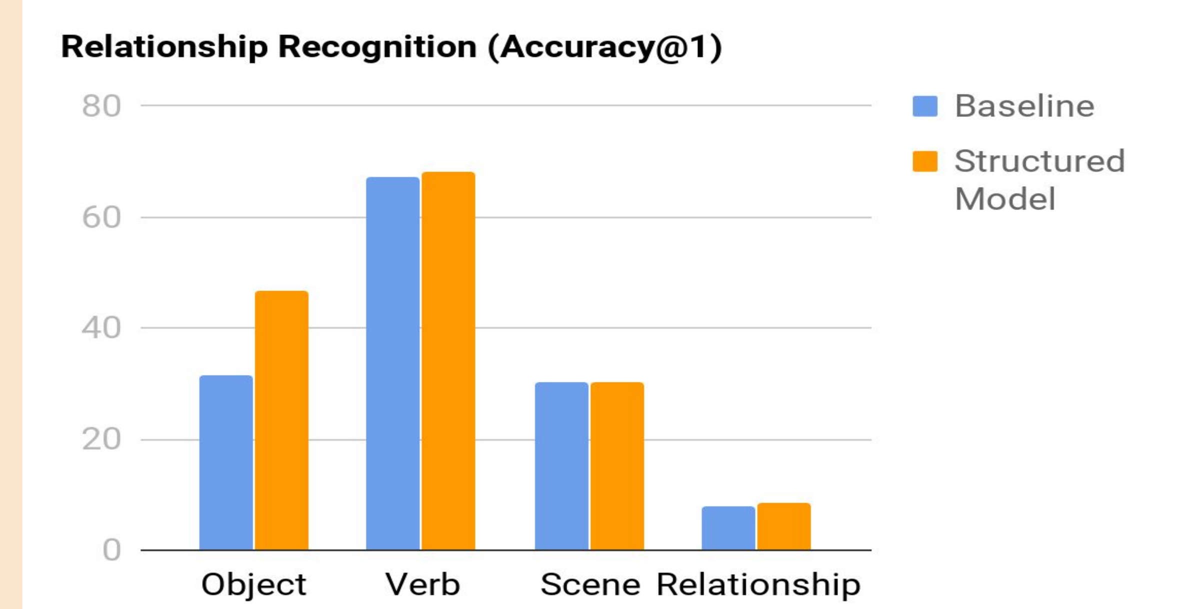
Baseline:	Food Hold Basement
GSTEG:	Dish Wash Kitchen
True LB:	Dish Wash Kitchen

Experiments and Evaluation

1. Relationship Detection: Detect Visual Relationships without object localization for the scale of the entire video. For evaluation we adopt mAP and Recall@K metrics where mAP measures the average of the precisions at different recall values and Recall@K measures the fraction of the positives detected in the top K detection results.
2. Relationship Tagging: The relationship tagging task [2] focuses only on precision. We adopt Precision@K to measure the accuracy of tagging results.
3. Relationship Recognition: The difference between this task and the previous task is that here we focus on fine grained relationship recognition, that is relationship tagging in each segment of the video.







References

[1]Xindi Shange et.al. Video visual relation detection. In proceedings of the 2017 ACM on Multimedia Conference, pages 1300–1308.ACM, 2017.
[2]Tsai, Yao-Hung Hubert et.al. Video Relationship Reasoning using Gated Spatio-Temporal Energy Graph. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2019.
[3] Carreira J, Zisserman A. Quo vadis, action recognition? a new model and the kinetics dataset. Inproceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2017 (pp. 6299-6308).