VIDEOS AS SPACE-TIME REGION GRAPHS

Xiaolong Wang, Abhinav Gupta

Robotics Institute, Carnegie Mellon University

https://arxiv.org/pdf/1806.01810.pdf

PROBLEM OVERVIEW

- Try to classify videos
  - New datasets, Charades and Something Something, are more difficult and more video-oriented

CHARADES DATASET

SOMETHING SOMETHING

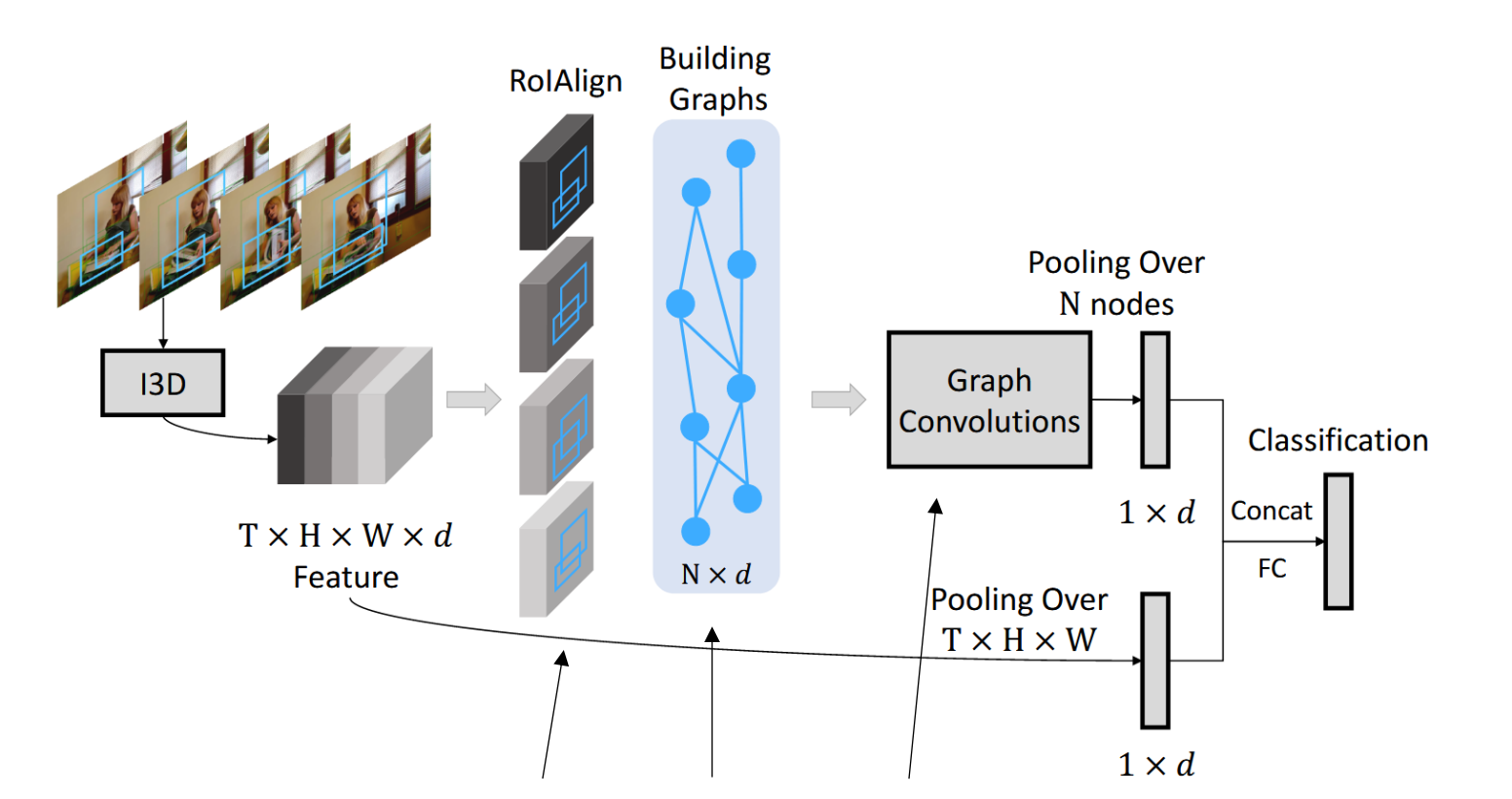
Putting *a white remote* into *a cardboard box*

Pretending to put *candy* onto *chair*

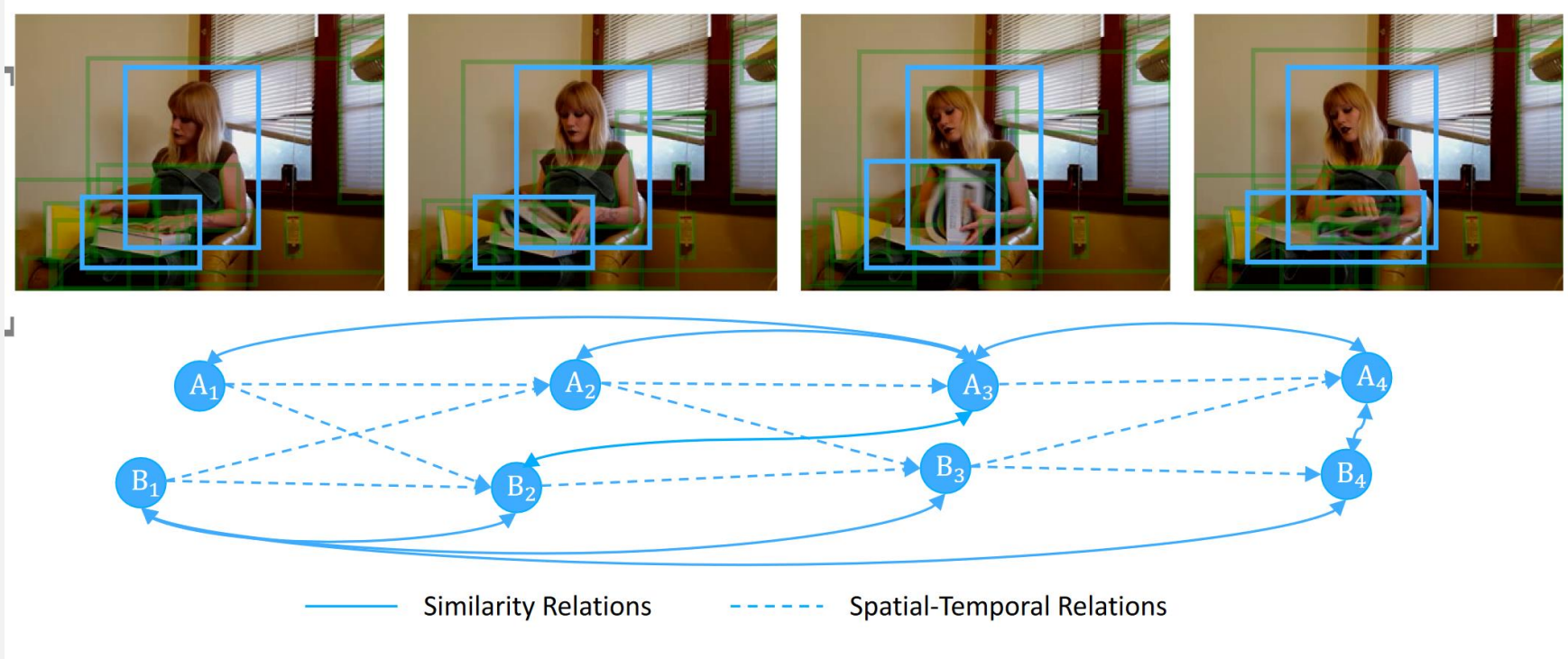
Pushing *a green chilli* so that it falls off the table

Moving *puncher* closer to *scissor*

ARCHITECTURE OVERVIEW



SPACE-TIME GRAPH



NODES OF GRAPH

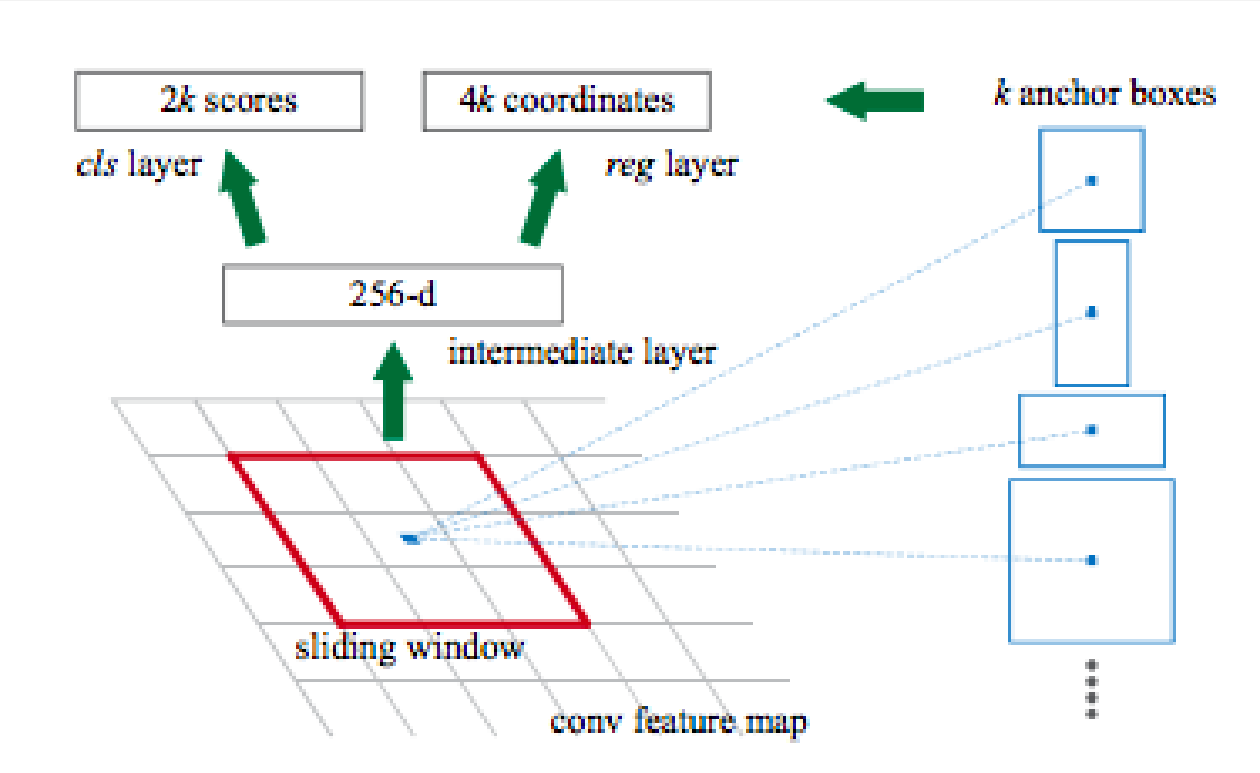
Create nodes:

- Region Proposal Network

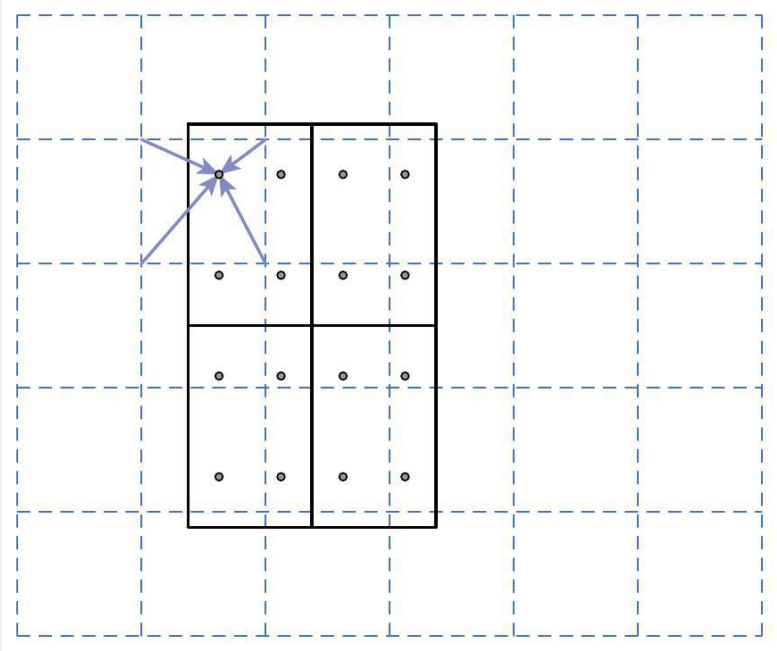
Fill nodes with data

- I3D Backbone based on resnet
- Crop features to bounding boxes
- ROIAlign

REGION PROPOSAL NETWORK



ROIALIGN



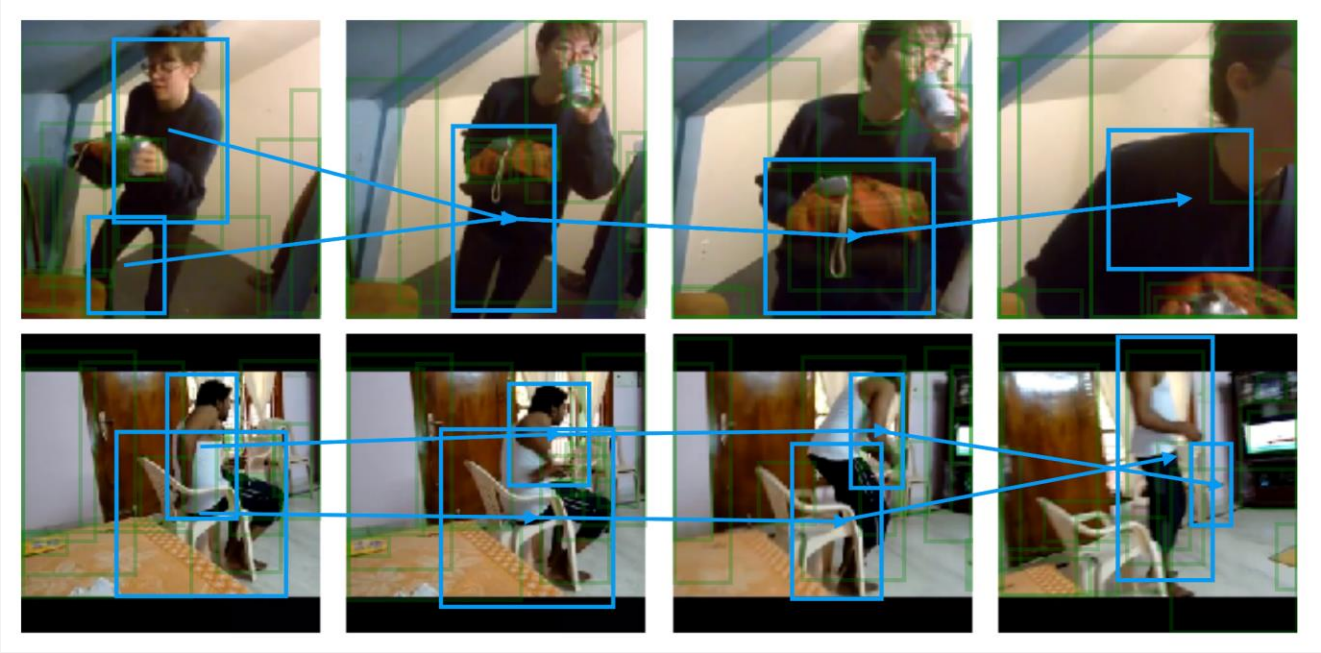
EDGES OF GRAPH

- Similarity graph
  - This is just repackaged transformer again
- Spatial-Temporal Graph
  - Backwards and forward edges weighted by bounding box overlap

SIMILARITY GRAPH

- Transformer: Key, Query, Value
- Non Local network: θ, ϕ, g
- Space Time Region Graphs: ϕ, ϕ', W

SPATIAL-TEMPORAL GRAPH



SPATIAL-TEMPORAL GRAPH

- Link objects in frame t to objects in frame t+1
- $\sigma_{ij} = \frac{Intersection(i,j)}{Union(i,j)}$
- $G_{ij}^{front} = \frac{\sigma_{ij}}{\sum_{j=1}^{N} \sigma_{ij}}$

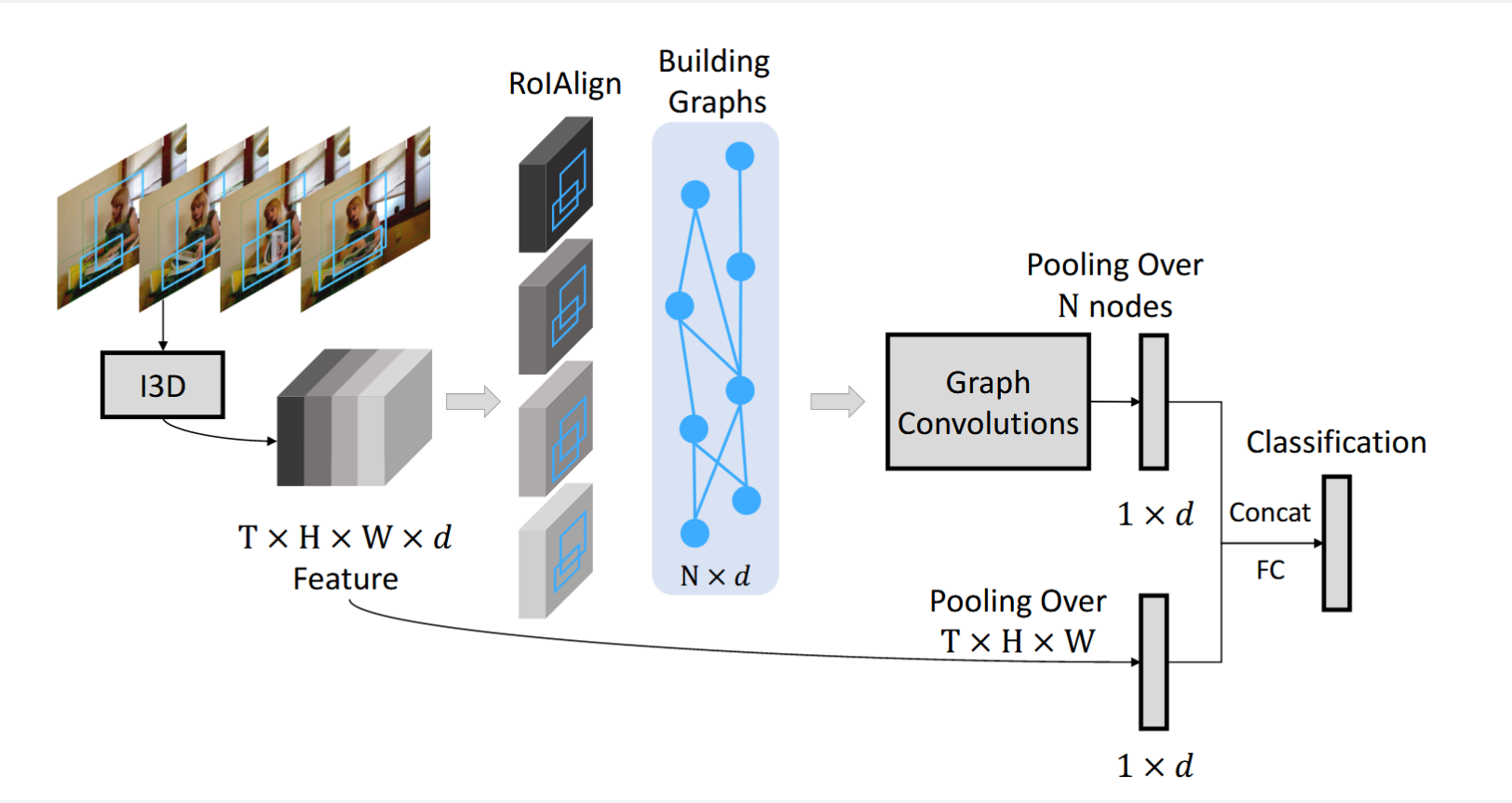
GRAPH CONVOLUTION

- Single Graph:
  - $Z = GXW + X$
- Multiple graphs:
  - $Z = \sum_n G_n XW_n + X$

HOW TO COMBINE N GRAPHS

- Combining forward + backward + similarity at each step doesn't work!
  - Hand-wavey explanation
- Run forward + backward in one branch, similarity in another
  - Average at end

ARCHITECTURE OVERVIEW REDUX



TRAINING PROCEDURE

- Trained in steps
- I: Train region proposal on COCO
- II: Train remaining weights as follows:
  - Train Resnet on Imagenet
  - → Inflate to I3D
  - → Train I3D on Kinetics
  - → Freeze I3D weights, train GCN on target dataset
  - → Unfreeze I3D weights, continue training

ABLATION: CHARADES

model, R50, I3D	mAP
baseline	31.8
Proposal+AvgPool	32.1
Spatial-Temporal GCN	34.2
Similarity GCN	35.0
Joint GCN	**36.2**

model, R50, I3D	mAP
baseline	31.8
Non-local	33.5
Joint GCN	36.2
Non-local + Joint GCN	**37.5**

BENCHMARKS: CHARADES

model	backbone	modality	mAP
2-Stream [93]	VGG16	RGB + flow	18.6
2-Stream +LSTM [93]	VGG16	RGB + flow	17.8
Asyn-TF [93]	VGG16	RGB + flow	22.4
MultiScale TRN [36]	Inception	RGB	25.2
I3D [8]	Inception	RGB	32.9
I3D [58]	ResNet-101	RGB	35.5
NL I3D [58]	ResNet-101	RGB	37.5
NL I3D + GCN	ResNet-50	RGB	37.5
I3D + GCN	ResNet-101	RGB	39.1
NL I3D + GCN	ResNet-101	RGB	**39.7**

BENCHMARKS: CHARADES

- Results table from "Videos as Space-time region graphs"

model	backbone	modality	mAP
2-Stream [93]	VGG16	RGB + flow	18.6
2-Stream +LSTM [93]	VGG16	RGB + flow	17.8
Asyn-TF [93]	VGG16	RGB + flow	22.4
MultiScale TRN [36]	Inception	RGB	25.2
I3D [8]	Inception	RGB	32.9
I3D [58]	ResNet-101	RGB	35.5
NL I3D [58]	ResNet-101	RGB	37.5
NL I3D + GCN	ResNet-50	RGB	37.5
I3D + GCN	ResNet-101	RGB	39.1
NL I3D + GCN	ResNet-101	RGB	**39.7**

Results table from "Nonlocal Neural Networks"

model	modality	*train/val*	*trainval/test*
2-Stream [43]	RGB + flow	18.6	-
2-Stream +LSTM [43]	RGB + flow	17.8	-
Asyn-TF [43]	RGB + flow	22.4	-
I3D [7]	RGB	32.9	34.4
I3D [ours]	RGB	35.5	37.2
NL I3D [ours]	RGB	**37.5**	**39.5**

BENCHMARKS: SOMETHING SOMETHING

model	backbone	*val* top-1	*val* top-5	*test* top-1
C3D [21]	C3D[7]	-	-	27.2
MultiScale TRN [36]	Inception	34.4	63.2	33.6
I3D	ResNet-50	41.6	72.2	-
I3D + Spatial-Temporal GCN	ResNet-50	42.8	74.7	-
I3D + Similarity GCN	ResNet-50	42.7	74.6	-
I3D + Joint GCN	ResNet-50	43.3	75.1	-
NL I3D	ResNet-50	44.4	76.0	-
NL I3D + Joint GCN	ResNet-50	46.1	76.8	45.0

SUMMARY

- Network is essentially nonlocal network, but with attention between bounding boxes instead of attention between downsampled pixels- more efficient
- Gets a few points better than nonlocal network on one dataset, and combines with nonlocal network for even better results

STRENGTHS

- COCO pretraining provides higher dimensional guidance which is extremely valuable
- Good benchmarks- especially combined with Non Local

WEAKNESSES

- To train in a domain you need:
  - Large number of image-level annotations for first pretrain
  - Large number of bounding box level annotations for second pretrain
  - Large number of video annotations for pretrain
  - Medium number of video annotations for final train
- Red flag: Spatial-Temporal Graph and Similarity Graph must be isolated from each other to train?

QUESTIONS