Austin Narcomey
aon2@stanford.edu

Visual Relationships as Functions: Enabling Few-Shot Scene Graph Prediction

Task and Motivation

Motivation: Scene Graph Prediction

We approach the task of **visual scene graph prediction**, which requires classifying objects and relationships in an image. Scene graphs are useful for applications such as knowledge bases, image captioning and retrieval, visual question answering. We develop a scene graph model which can:

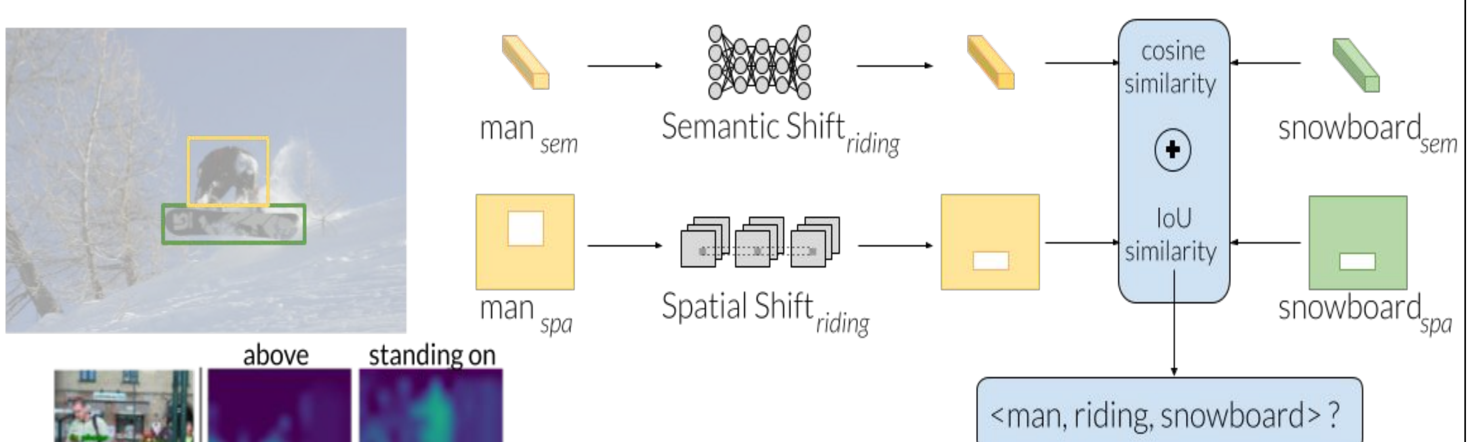
1) Learn new relationships with **only a few examples**
2) Learn **interpretable representations** of each relationship

Approach

Represent each object in an image as a spatial mask and semantic feature embedding.

Learn each relationship class as a spatial and a semantic shift function: *riding* transforms the spatial and semantic features of a person to approximate the snowboard they're riding. We also learn inverse functions, so *riding*$^{-1}$ transforms the snowboard to approximate the rider. See figure 1, 2.

Figure 1: Semantic and Spatial shift functions determine likelihood of a relationship



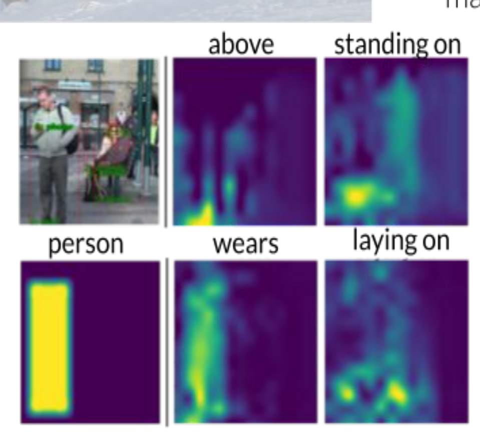


Figure 2: Visualization of interpretable spatial shift functions. The function for each relationship shifts the man's spatial mask to the predicted object of the relationship

Few-shot Learning Pipeline

1) **Fully train** Graph Convolution model and spatial and semantic shift functions on relationships with abundant data.
2) **Define** shift functions for new rare relationships with few examples using fully trained shift functions.
3) **Fine-tune** new shift functions with few training examples.

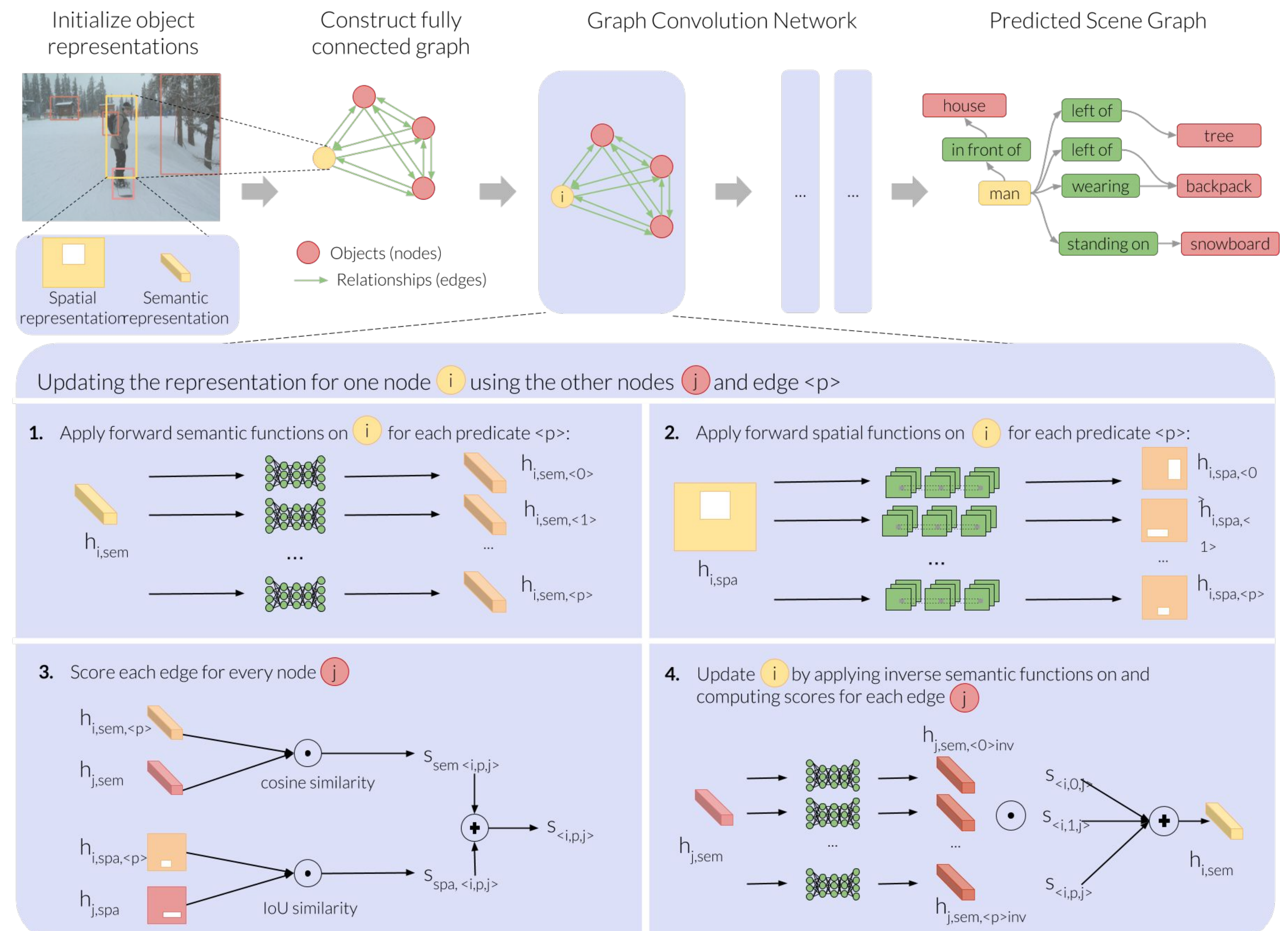
Data

Data: All data is from **Visual Genome**, which contains images annotated with bounding boxes, object classes, and relationships classes.

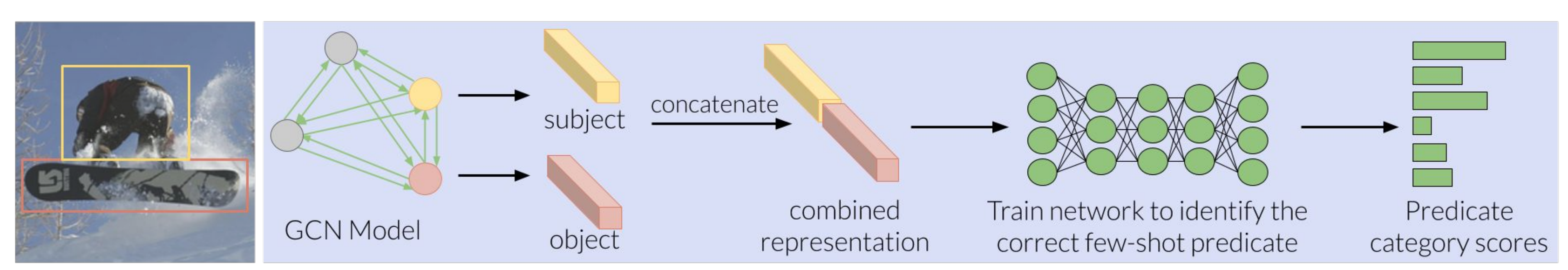
Architecture and Design

Models

Fully Trained Scene Graph Prediction



Few-Shot Scene Graph Prediction



We use a **Graph Convolution Network** (GCN) with the spatial and semantic features of each object as nodes in the graph, connecting all object nodes to all relationship edges. In each layer, the GCN applies **spatial and semantic shifts** to a node and measures similarity of the shifted nodes to the original node (high similarity, highly likely relationship). The node is updated using similarity scores for each relationship as weights.

We implement **Few-Shot Learning** by concatenating the object representation from the GCN with the GCN representation of another object, and train an MLP to predict scores for each relationship type between the two objects. This uses the object representations built by the shift functions of fully trained relationships.

Future Work

The next steps are to experiment with few-shot learning on other scene graph prediction architectures. This model uses a generative approach which typically outperforms classifiers when data is limited, but classifier models should be more thoroughly investigated and optimized for few-shot learning.

Results

Evaluation Methods

Metric: Scene graph prediction is typically evaluated with **Recall@n**: how many of the ground truth relationships in an image are within the top n scoring relationships. Predictions are constrained to one relationship per pair of objects.

Few-shot constraints: For few-shot learning we constrain predictions to only the relationships learned in few-shot rather than fully trained. For our evaluations of few-shot learning, the training dataset consists of all available examples of fully trained relationships and k ϵ {1,2,3,4,5} examples of each of the few-shot relationships, and the testing dataset consists of only few-shot relationships

Performance Against Existing Models

Fully Trained Scene Graph Prediction

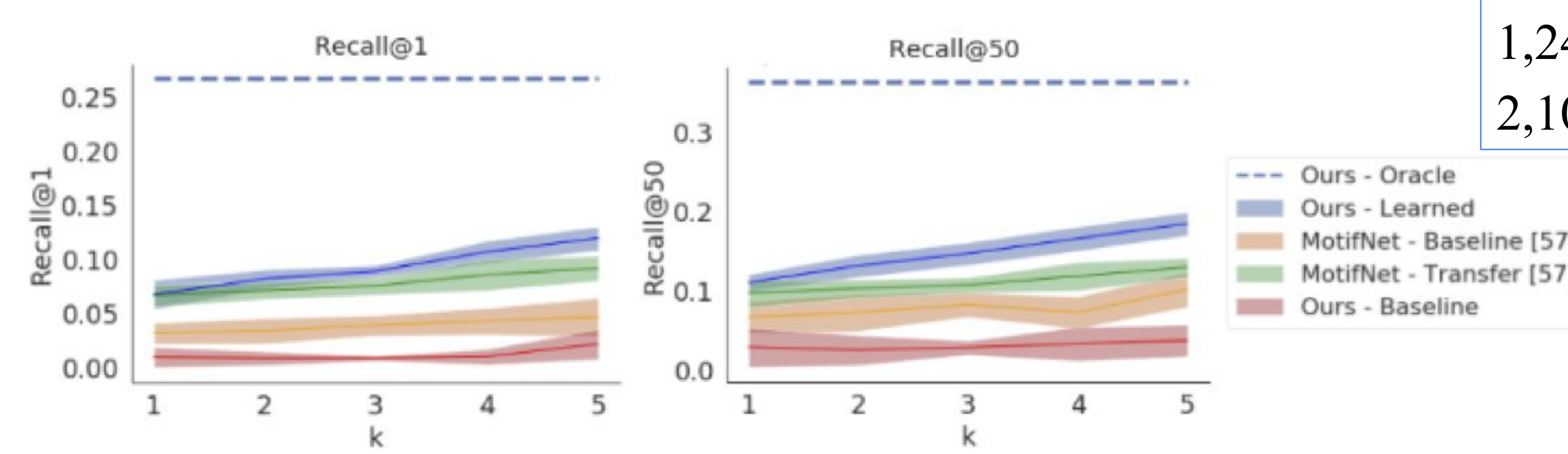
		SG GEN		SG CLS		PRED CLS	
	Metric	recall@50	recall@100	recall@50	recall@100	recall@50	recall@100
vision only	IMP [55]	06.40	08.00	20.60	22.40	40.80	45.20
	MSDN [35]	07.00	09.10	27.60	29.90	53.20	57.90
	MotifNet-freq [58]	06.90	09.10	23.80	27.20	41.80	48.80
	Graph R-CNN [56]	11.40	**13.70**	**29.60**	**31.60**	54.20	**59.10**
	Our full model	**13.18**	13.45	23.71	24.66	**56.65**	57.21
external	Factorizable Net [34]	13.06	16.47	-	-	-	-
	KB-GAN [18]	13.65	17.57	-	-	-	-
	MotifNet [58]	**27.20**	**30.30**	**35.80**	**36.50**	**65.20**	**67.10**
Ablation	Our spatial only	02.05	02.32	03.92	04.54	04.19	04.50
	Our semantic only	12.92	12.39	23.35	24.00	56.02	56.67
	Our full model	**13.18**	**13.45**	**23.71**	**24.66**	**56.65**	**57.21**

Training set:
36,662 images
134,642 relationships

Testing set:
15,983 images
60,83 relationships

Few-shot testing set:
1,245 images
2,107 relationships

Few-Shot Scene Graph Prediction



Our model achieves **near state-of-the-art performance on fully trained scene graph prediction**, while also performing strongly on few-shot prediction. We see from ablations that semantic information is the primary driver of scene graph prediction performance rather than spatial information.

In few-shot learning we see that our model which learns few-shot relationships as an MLP over GCN object representations **outperforms all baselines**, including a state-of-the-art classifier MotifNet.