**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.

---

**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.

**Models**

**Fully Trained** Scene Graph Prediction

![Diagram of Fully Trained Scene Graph Prediction](image)

**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.

**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.

---

**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 \(k \in \{1,2,3,4,5\}\) examples of each of the few-shot relationships, and the testing dataset consists of only few-shot relationships.

**Performance Against Existing Models**

<table>
<thead>
<tr>
<th>Model</th>
<th>SG GEN</th>
<th>SGCLS</th>
<th>PGGNet</th>
<th>PG-Net</th>
<th>MotifNet</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recall@50</td>
<td>Recall@100</td>
<td>Recall@50</td>
<td>Recall@100</td>
<td>Recall@50</td>
<td>Recall@100</td>
</tr>
<tr>
<td>BNP [5]</td>
<td>26.00</td>
<td>30.00</td>
<td>25.00</td>
<td>30.00</td>
<td>25.00</td>
</tr>
<tr>
<td>MIDN [35]</td>
<td>30.50</td>
<td>35.50</td>
<td>30.50</td>
<td>35.50</td>
<td>30.50</td>
</tr>
<tr>
<td>MotifNet [16]</td>
<td>29.00</td>
<td>32.00</td>
<td>29.00</td>
<td>32.00</td>
<td>29.00</td>
</tr>
<tr>
<td>Our full model</td>
<td>30.50</td>
<td>34.00</td>
<td>30.50</td>
<td>34.00</td>
<td>30.50</td>
</tr>
</tbody>
</table>

**Training set**: 36,662 images 134,642 relationships

**Testing set**: 15,983 images 60,834 relationships

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

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.