Finding one image among a collection of images gives a free supervisory signal. Thus we propose the following system:

- 2 agents – one has several images and the other has only one of those images
- They communicate via questions and answers. In the end, first agent outputs its guess for second agent’s image
- Can learn different visual tasks based on restriction on communication. For example: VQA, Dense Captioning, Attribute Prediction (this work)

**Evaluation Metrics**

- **Metrics for A**: We’ll study quality of A’s questions qualitatively
- **Metrics for B**: Average precision across all images. Precision for single image is fraction of its top-k attributes that are also in the top-k attributes of ground truth.
- **Metrics for the whole system**: Accuracy of disambiguation

**Baseline**:
Our baseline is Deep Carving model (Shankar et al, CVPR 2015).

**Dataset**:
We use SUN Weakly Supervised Dataset
- 42 attributes of types shape, color, texture etc.
- 22084 training images having 1 attribute strongly indicative of it
- 5618 test Images with entire ground truth attribute vectors

**Future Work**
- More than two images
- Multiround communication

**Other modalities**: VQA
**Other restrictions**: Dense Captioning