Low Supervision Visual Learning through Cooperative Agents

Abhishek Sinha, Ashish Bora

Our Idea

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)

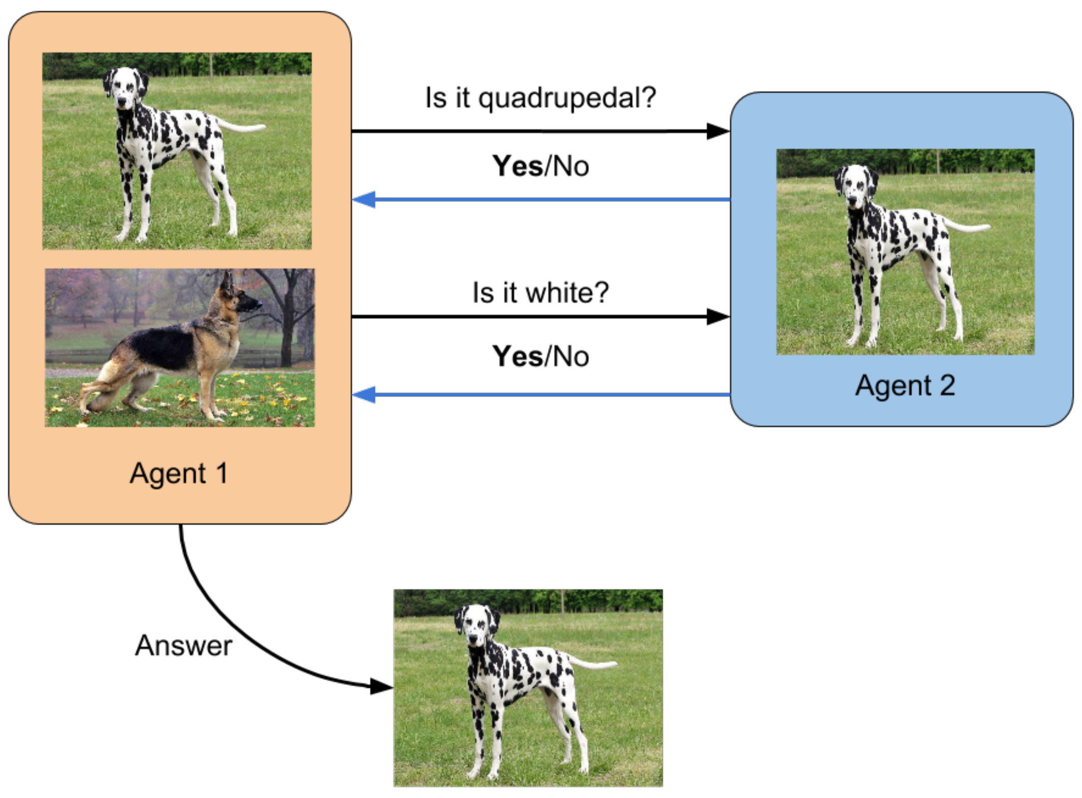


Figure 1: High Level Idea

Simplification

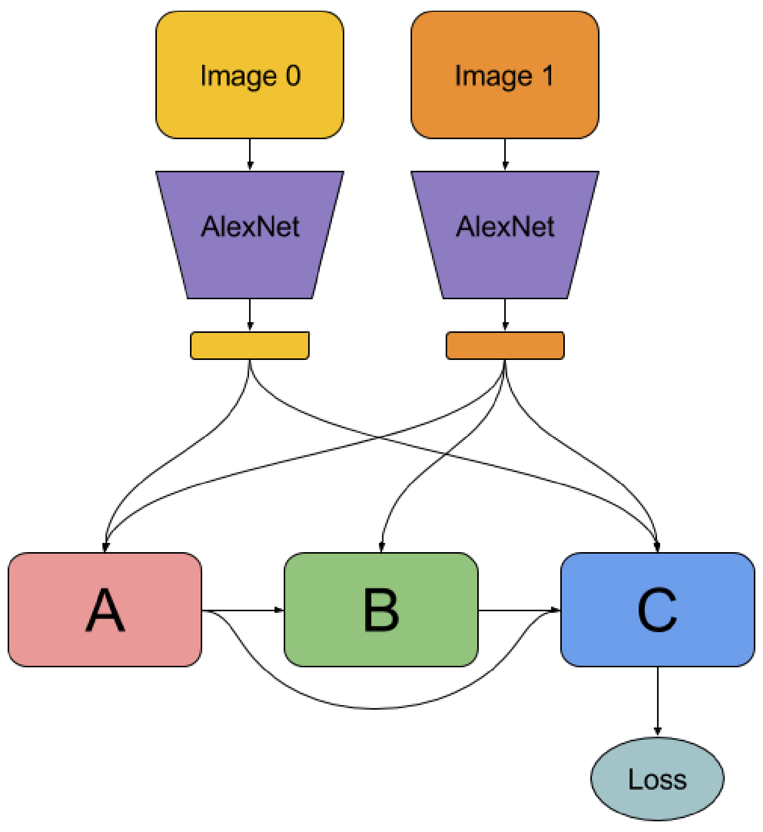


Figure 2: Simplified Architecture – Two Images, Single Question, Feed-forward

Attribute Prediction Problem

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

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

Agent Architectures

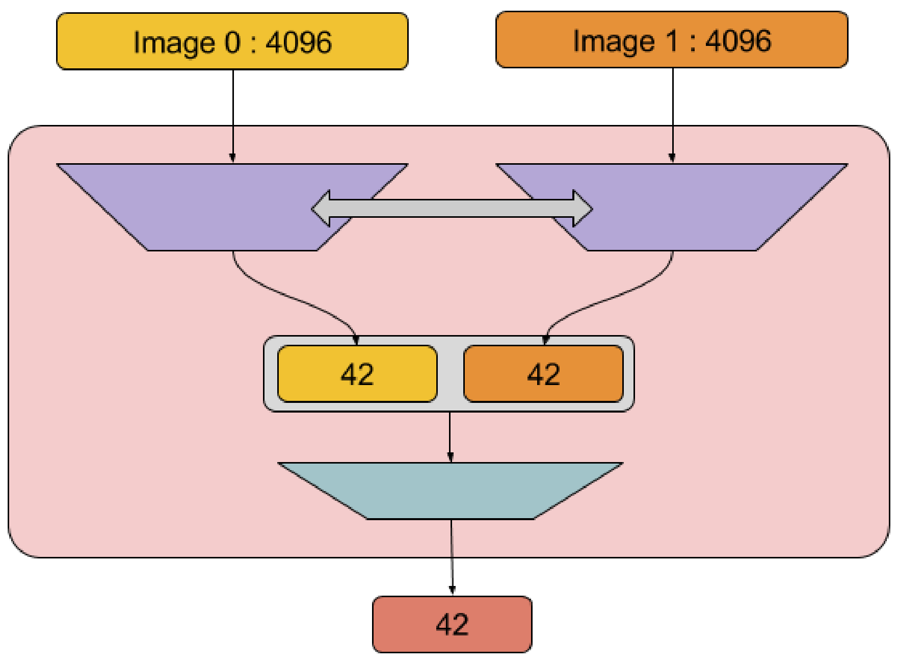


Figure 3: Questioning Network (A). Double Arrow indicates weight sharing and Trapezoids indicate fully connected layers. Non-Linearities are not shown to avoid clutter.

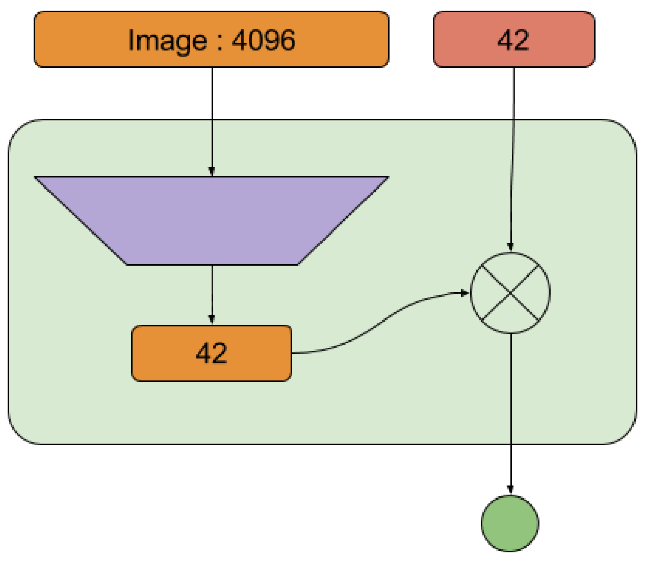
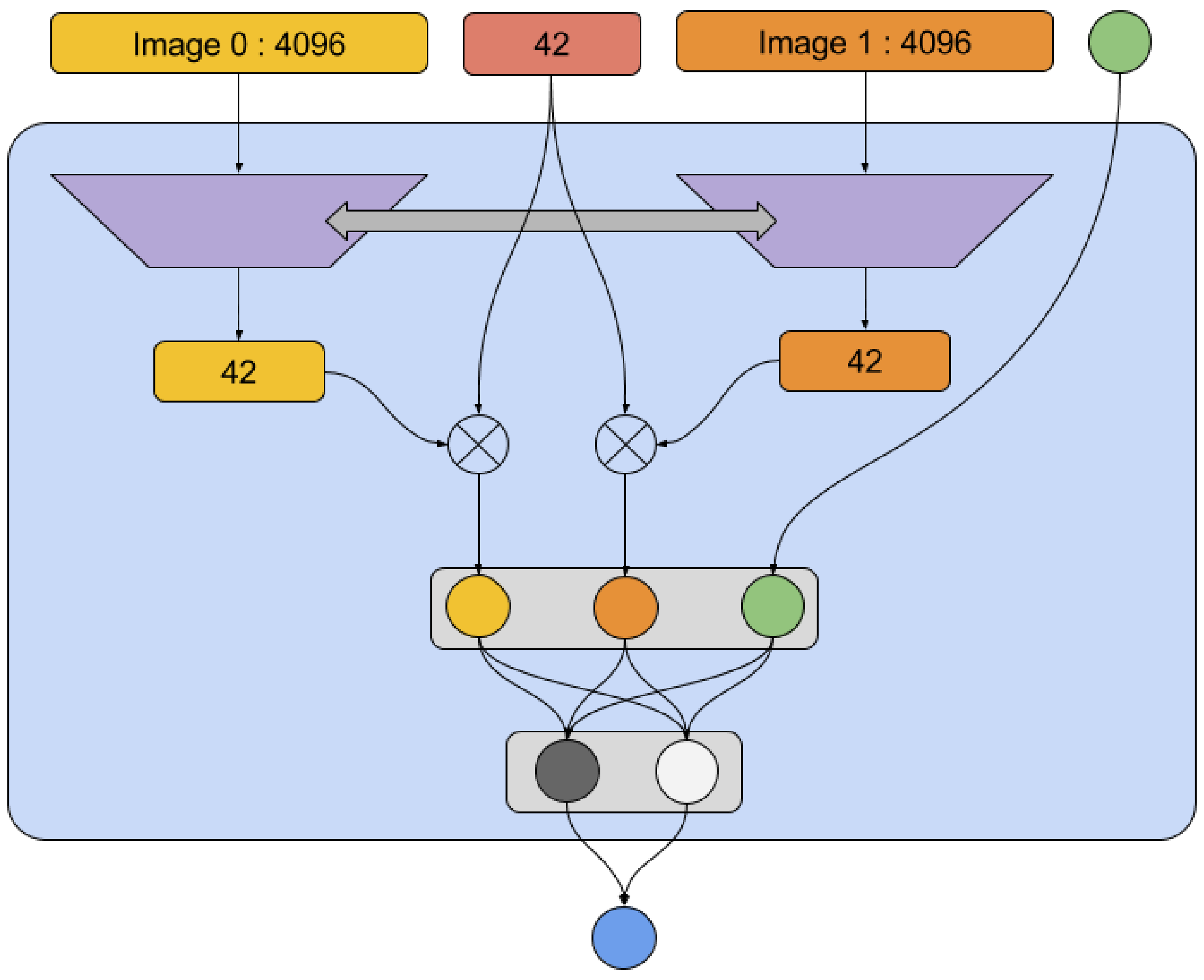


Figure 4: Figure on the Top is Answering Network (B). ⊗ indicates dot product

Figure 5: Figure on the right is Judgement Network (C).



Challenges and Tricks

1. **Mimicking :** Agents learn to mimick each other. No visual learning.
   Solutions : Dropout, Different architectures, random crops, add noise to data
2. **Interpretability :** Agents communicate in a strange code language. Fundamental problem since any permutation of meaning works.
   Solution : Supervised pretraining, tune B on original task intermittently
3. **Low Supervisory Signal :** At most single bit of information per example.
   Solution : Large batch size, low learning rate

Training

Loss function : Binary Cross Entropy

$$L(\theta) = \sum_{i=1}^{m} (y_i \log(p(x_i;\theta)) \\ + (1-y_i)\log(1-p(x_i;\theta)))$$

We train the network end-to-end using standard backpropagation. We train in following order

1. Fix B to be pretrained model
2. Train C with synthetic questions and answers given by B
3. Fix B and C, train A.
4. Finetune the whole system.

Results

Future Work

- More than two images
- Multiround communication
- Other modalities: VQA
- Other restrictions: Dense Captioning