

Relationship Prediction for Scene Graph Generation



Uzair Navid Iftikhar (unavid), Aamnah Khalid (aamnah)
CS229 Stanford University

Introduction

Scene graphs offer powerful representations of images, aimed at preserving information relating to semantic and spatial relationships, as well as salient attributes of objects. We explore relationship prediction on the GQA dataset.

Top Relations (left/right excluded)

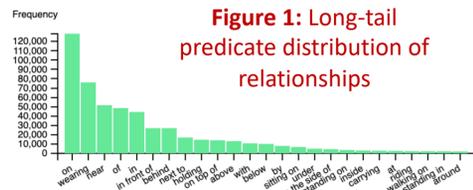


Figure 1: Long-tail predicate distribution of relationships

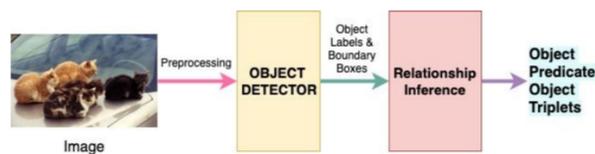


Figure 2: Example of an image and scene graph [1]

Methodology

- Used subset of 12k images out of 113k.
- Scene graph has characteristic long tail distribution of relationship predicates [Figure 1]
- 75 - 12.5 - 12.5 train-val-test split. 99-1 split does not capture variance with rare predicates.
- Pre-trained object detector YOLOv3 to predict nodes in graph.
- Maximum Likelihood estimator for baseline.
- Custom Neural network for relationship classification with architecture [Figure 3]
- DenseNet Relationship Prediction Model: accepts cropped images containing source, target, and union of the bounding boxes. Outputs a relationship class. [Figure 4]

Figure 10: Relationship Prediction Pipeline



Methodology (cntd.)

- Hyperparameter tuning on validation set first with a random search, and then an organized search of the hyperparameter space.
- Cross-entropy loss: $-\frac{1}{N} \left(\sum_{i=1}^N y_i \cdot \log(\hat{y}_i) \right)$

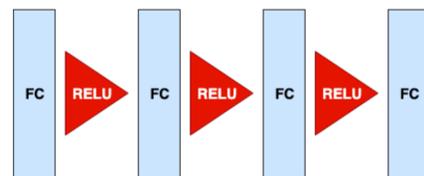


Figure 3: Custom FC-ReLU Model (left)

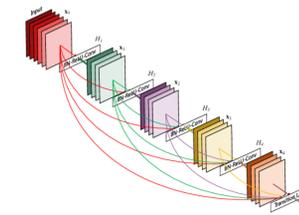


Figure 4: DenseNet Architecture [3]

Results

- Evaluation metric for each image was accuracy:

$$\sum_{\mathcal{R} \in \text{Image}} \left(\frac{1_{[\text{predictedRelationship} == \mathcal{R}]}}{\text{total relations in image}} \right)$$

- Maximum Likelihood Estimator: Laplace smoothing with $\lambda = 1$. Given source object and target object names, we estimated conditional distribution of relationship label. 37% accuracy to establish baseline. Accuracy on predicates with top-20 accuracy shown below:

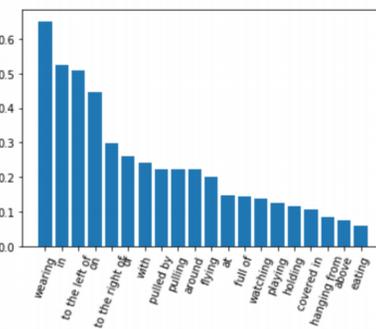


Figure 5: Top-20 MLE Accuracy Predicates

Model	Accuracy
MLE	37%
FC-ReLU 4 Layer	4.2%
DenseNet	38.5%

Figure 6: Accuracy comparisons for different models on validation set

Results (cntd.)

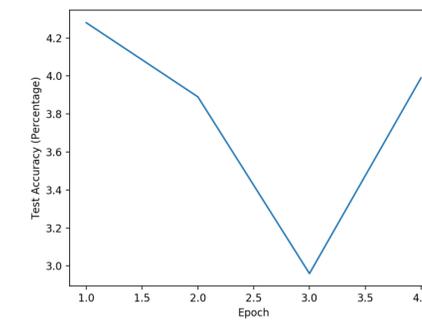


Figure 7: Validation set accuracy over epochs during training of FC-ReLU 4-Layer network

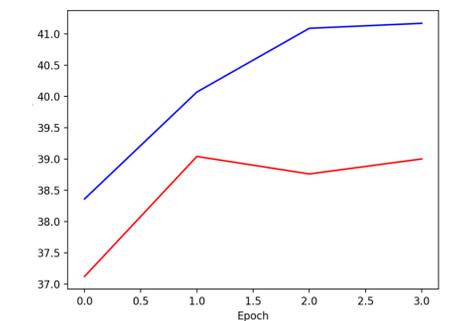
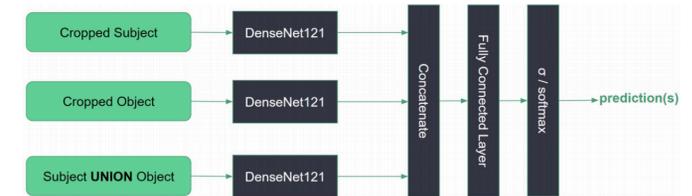


Figure 8: Accuracy over epochs during training of DenseNet model (blue: training, red: val)

Figure 9: Architecture for DenseNet-based model [2]



- DenseNet performed the best: 41% training acc, and 38.5% validation accuracy. **Beat baseline.**
- Poor performance of FC-ReLU network suggests that the shapes of bounding boxes are not enough to give relationship context. DenseNets extract more semantic information while reducing the learning of redundant feature maps. [3]

Future Work

- Training with more images / epochs.
- Incorporating Q/A data, which is a part of the GQA dataset.
- Comparing loss functions: New loss functions recently proposed by Zhang et al. [4] to reduce penalty for negative relationship direction and better handle multiple instances of same object class in an image.

REFERENCES

- [1] Hudson, D. A., Manning, C. D. (2019). GQA: a new dataset for compositional question answering over realworld images. arXiv preprint arXiv:1902.09506.
[2] Dornadula et al. "Graphr: Scene Graph Generation using Deep Variation-structured Reinforcement Learning." CS234 Project.
[3] Huang, Gao, et al. "Densely connected convolutional networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.
[4] Zhang, Ji, et al. "Graphical Contrastive Losses for Scene Graph Parsing." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019.