# CONTRASTIVE FACTOR GRAPH ATTENTION FOR VISUAL DIALOG

Akruti Kushwaha†, Jeet Kanjani†, Tanay Sharma†

†Carnegie Mellon University

## Motivation

Visual Dialog is a challenging and natural form of the visual question answering (VQA) problem where, we need to interact with multiple modalities and maintain context in the form of dialog history to provide an answer to the given query.

Our contributions:

1. Visual Dialog systems including our baseline model are not robust to minor linguistic variations and environmental conditions that they may be employed in. To solve this issue, we propose to **train our baseline model by augmenting the training data** with different variations of the input dialog and image.
2. Crossentropy(CE) loss treats every image-question pair independently and fails to exploit the information that some questions and images in the augmented dataset are variations of each other. We propose to **use supervised contrastive loss** with CE loss to tackle this issue.
3. The FGA model incorrectly ranks diverse answer options higher than related variations of the ground truth. We propose the **creation of heuristic relevance scores** to improve on the NDCG metric.

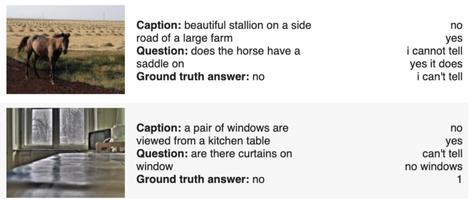



Fig. 1: FGA model tends to rank diverse answer options together

## Dataset

We use the Visual Dialog [1] dataset for our project. VisDial contains 1 dialog each (with 10 question-answer pairs) on 140k images from COCO dataset, for a total of 1.4M dialog question-answer pairs.

The problem of Visual Dialog task can be described as, an image I, the 'ground-truth' dialog history (including the image caption) $H = (\underbrace{C}_{H_0}, \underbrace{(Q_1, A_1)}_{H_1}, \ldots, \underbrace{(Q_{t-1}, A_{t-1})}_{H_{t-1}})$ , the question $Q_t$, and a list of N = 100 candidate answers, the model is expected to return a sorting of the candidate answers. The model outputs the $P(a_t^i|H, Q_t)$ for each answer option i.

| Type | Train | Val | Test |
|---|---|---|---|
| Images | 123287 | 2064 | 8000 |
| Dialogs | 1232870 | 20640 | 8000 |

Fig. 2: Distribution of VisDial 1.0 dataset

## Experimental Methodology

We build upon the Factor Graph Attention [4] model that has a unified attention mechanism based on graph like interactions. The nodes in the graph represent utilities and interactions between them are modelled by factors.

**Training on VisualDialog and Improved Joint representations** - The original data is passed through an image and question module to create augmented images and paraphrased questions as positive samples.

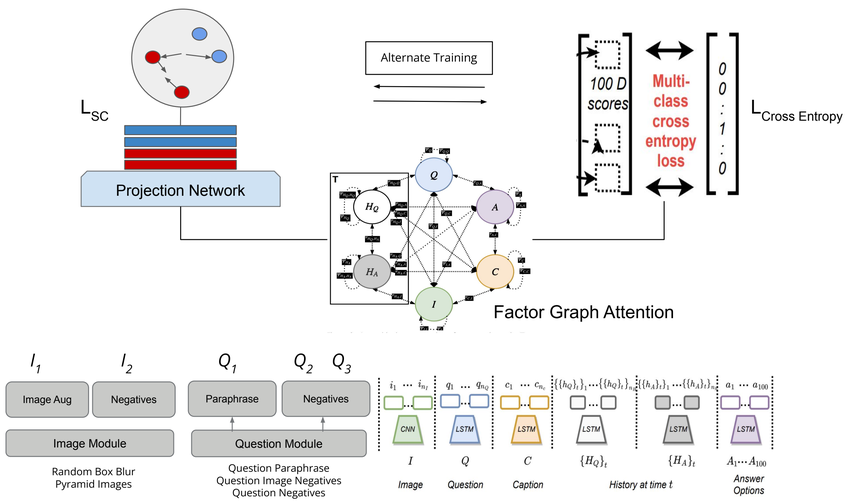



Fig. 3: FGA model with alternate training with Supervised Contrastive (SC) loss

We use multitask learning strategy to train our models. Taking inspiration from [2], for the first task, we use Supervised Contrastive (SC) loss between data augmented positive and negative samples using Equation (1). We aim to improve input representations and increase robustness of the model by this loss.

$$\mathcal{L}_{\mathrm{SC}}^{i} = -\sum_{p=1}^{|\mathcal{X}^{+}(x_i)|} \log \frac{\exp\left(\Phi\left(\boldsymbol{z}_i \cdot \boldsymbol{z}_p\right)/\tau\right)}{\sum_{k=1}^{K} \mathbb{1}_{k \neq i} \cdot \exp\left(\Phi\left(\boldsymbol{z}_i \cdot \boldsymbol{z}_p\right)/\tau\right)} \tag{1}$$

We use Cross Entropy Loss (CE) on the original data to improve the discriminative power of the model. The network is trained end to end using an alternative training strategy.

**Heuristic Scores** - FGA model applies CE loss to ground truth answer only.

- Generate Unimodal Heuristic Scores (UHS) - by comparing the answer options to the ground truth answer using cosine similarity of their contextualized embeddings.
- Generate Multimodal Heuristic Scores (MHS) - by using embeddings from a ViLBERT model pretrained on the VQA task inspired from the VisDial-BERT [3] implementation. We will report these results in the final submission.

Using these scores, we optimize the model by computing the CE loss against the log likelihood probabilities outputted by the model.

## Comparison

**Q:** is it raining?
**A:** No

| Baseline | FGA + UHS |
|---|---|
| yes | no |
| it may have stopped | no it's not |
| no | no not really |

**Q:** what color is the ground?
**A:** reddish brown dirt

| Baseline | FGA + UHS |
|---|---|
| green | reddish brown dirt |
| reddish brown dirt | red and white |
| spring green | brown |

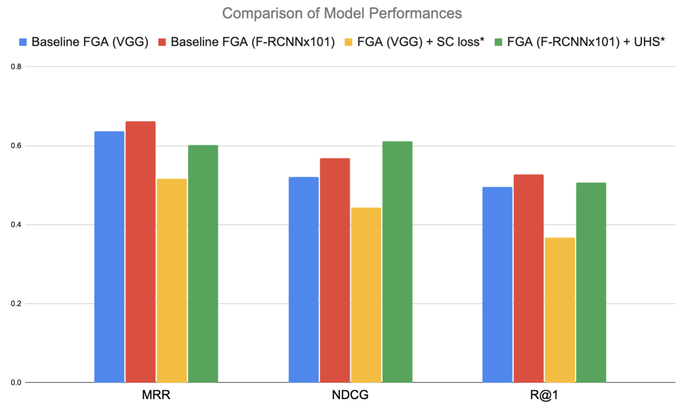
**Q:** Can you see any grass
**A:** yes

| Baseline | FGA + UHS |
|---|---|
| yes | yep |
| no | yes |
| i cannot tell | yes it is |


**Q:** Are there trees around
**A:** No

| Baseline | FGA + UHS |
|---|---|
| No | I don't think so |
| I can only see the water | Nope |
| I cannot tell | No |

## Results




| Model | MRR | NDCG | R@1 |
|---|---|---|---|
| Baseline FGA (VGG) | 0.637 | 0.521 | 49.58 |
| Baseline FGA (F-RCNNx101) | 0.662 | 0.569 | 52.75 |
| FGA (VGG) + SC loss* | 0.516 | 0.444 | 36.72 |
| FGA (F-RCNNx101) + UHS | 0.602 | **0.611** | 50.60 |
| FGA (VGG) + SC loss + UHS* | 0.585 | 0.486 | 43.70 |

Fig. 7: Supervised Contrastive loss (SC) and Unimodal HeuristicScores (UHS) experiments results on VisDial 1.0 val set

*signifies still in training

Using UHS with FGA, we see an improvement in the NDCG metric but the MRR takes a hit. This is because initially the model was trained to only optimize for the ground truth answer while in heuristic scores, the probabilities of the correct answers have been distributed across the relevant answers.

## References

[1] Abhishek Das et al. *Visual Dialog*. 2017. arXiv: 1611.08669 [cs.CV].

[2] Yash Kant et al. *Contrast and Classify: Training Robust VQA Models*. 2021. arXiv: 2010.06087 [cs.CV].

[3] Vishvak Murahari et al. *Large-scale Pretraining for Visual Dialog: A Simple State-of-the-Art Baseline*. 2020. arXiv: 1912.02379 [cs.LG].

[4] Idan Schwartz et al. *Factor Graph Attention*. 2020. arXiv: 1904.05880 [cs.CV].