

---

# Robust Factor Graph Attention Net

---

Akruti Kushwaha<sup>1</sup> Jeet Kanjani<sup>1</sup> Tanay Sharma<sup>1</sup>

## Abstract

Visual Dialog is a challenging problem where models not only need to interact with multiple modalities but also have to maintain context in the form of dialog history to provide an answer to the given query. This is a more natural form of the Visual Question Answering task as it allows for communication with the agent. Significant progress has been made in this domain. In this paper, we show the results of two baselines on the VisDial dataset and discuss their current challenges. We attempt to address them using a combination of methods and design choices such as a contrastive loss formulation, data-augmentation strategies, and generating unimodal and multi-modal heuristic scores while training. Thus, our goal is to make VisDial models more robust and accurate for general use.

## 1. Introduction

Recent years have seen a plethora of work in Artificial Intelligence (AI) involving vision and language. It not only tests the capability of AI to integrate computer vision, reasoning, and natural language understanding but also their importance in enhancing human-machine collaboration. Visual systems along with better understanding of the human language are able to aid in important tasks. For instance, helping the visually impaired to interact with visual content using language, enabling human-computer interaction, and improving visual search.

For this project, we focus on a specific vision and language task, the Visual Dialog [Das et al., 2017], an extension of the Visual Question Answering (VQA) [Agrawal et al., 2016] task. The problem of Visual Dialog has been defined as – given an image and dialog history consisting of a sequence of question-answer pairs, and a follow-up question about the image, predict a free-form natural language answer to

the question. Recent years have seen incredible progress in Visual Dialog. The state-of-the-art on the task has improved by more than 20% absolute (now  $\sim 74\%$  NDCG). However, the current models are far from perfect in their answer generation capability and are still affected with pressing issues that prevent their usage in real-world scenarios such as:

1. **Robustness:** We observe that the models trained on VisDial dataset are not robust to different types of variations pertaining to image and questions. If the images are augmented or we slightly paraphrase the queries, the model fails to output the correct answer. This is important as in real life we may not have the exact same queries and thus the model should be robust to such variations.
2. **Model uncertainty:** Models also suffer from varying confidence scores among the top relevant answer candidates. Ideally, all similar answers to the ground truth should be ranked higher. But we see that many times it does not happen in practice. Dissimilar/wrong answers are sometimes ranked higher than similar answers even when the ground-truth answer is predicted correctly. Metrics such as NDCG are also seriously affected due to the model uncertainty issue. This is mainly because the model is being optimized only towards the single ground truth answer, creating uncertainty.

Keeping the aforementioned issues in mind, we propose the following ideas as contributions to the improvement of visual dialog systems:

1. **Data augmentation + Contrastive Learning** Data augmentation techniques such as question paraphrasing and object-specific augmentations in the image can help to make the model more robust by stressing the model to learn important features from different input representations. Contrastive learning can be used by creating multi-modal positive and negative samples using various data augmentation strategies (Section 6.1) of images and questions. We add another loss term in the objective to improve the representations and to improve the robustness of the system.
2. **Training with heuristic scores** that represent the relevance scores of the related answer variations in the answer options. It can help the model be certain towards different answer variations and improve the model's

---

<sup>1</sup>Carnegie Mellon University, USA. Correspondence to: Akruti Kushwaha <akrutik.andrew.cmu.edu>, Jeet Kanjani <jkanjani@andrew.cmu.edu>, Tanay Sharma <tanays@andrew.cmu.edu>.

performance on the NDCG score. The recent update to the VisDial dataset has added dense annotations on the validation set. This contains the ground truth relevance score of the relevant answers. Generally, models are fine-tuned/trained on these annotations but we believe that the number of such annotations is often inadequate. Hence, we propose using a heuristic approach to calculate the normalized relevance score by the similarity of the ground-truth answer. We, then train our models using these relevance scores to see the final performance.

## 2. Related Work

### 2.1. Data Augmentation

A lot of work has been done to make the models more robust in the VQA domain wrt linguistic variations in the query and noise in the images. To tackle these variations, approaches like [Shah et al., 2019], [Tang et al., 2020], [Jiang et al., 2018] augment the input query with different variations of it with similar meaning while training. For instance [Shah et al., 2019], does this by creating a visual question generator model using image as input and question as output. Such model is trained using cycle consistency by making the model predict question from answer. [Kant et al., 2021] uses different NMT models to back translate the query and use that as the augmented query after appropriate confidence thresholds.

### 2.2. Contrastive Learning

Contrastive learning has been used in a variety of domains to improve the representations of the input data. In the visual domain, there has been great advancements to learn representations in a self supervised manner [Wu et al., 2018] [He et al., 2020] [Hénaff et al., 2020]. Apart from Image Classification, recently, [Gupta et al., 2020] used contrastive learning for phrase grounding. On the contrary, we want to learn representations which are robust to not only linguistic variations but also to visual representations. Some of the works [Khosla et al., 2021] samples random positive and negative pairs based on label information, whereas we use curated positive and negative pairs based on [Kant et al., 2021]. This [Kant et al., 2021] uses contrastive learning on VQA dataset. We extend this approach on VisDial dataset and show our analysis at the end.

### 2.3. Heuristic Scores

Visual Dialog [Das et al., 2017] propose finetuning with dense annotations i.e. relevance scores for all 100 answer options corresponding to each question on a subset of the training set. Although they report a higher normalized discounted cumulative gain (NDCG) as compared to the baseline, the MRR score for their predictions suffers. [Agarwal

et al., 2020] perform a manual evaluation and find out that the relevance information for answers contains substantial noise. They note that the ground truth answers were marked as irrelevant for 20% of train and 10% of val set. Naturally the model gets confused by training on these annotations. They manually correct these relevance scores on the training subset of dense annotations and present that the model performs significantly better on single target metrics like MRR while still being comparable in NDCG. We extend the idea further to generate relevance scores for the whole dataset in a semi-supervised way.

## 3. Problem Statement

The problem of Visual Dialog task can be described as, an image  $I$ , the ‘ground-truth’ dialog history (including the image caption)  $H = (\underbrace{C}_{H_0}, \underbrace{(Q_1, A_1)}_{H_1}, \dots, \underbrace{(Q_{t-1}, A_{t-1})}_{H_{t-1}})$ ,

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  $P(a_i^c | H, Q_t)$  for each answer option  $i$ . The model is evaluated on retrieval metrics – (1) mean reciprocal rank (MRR) of the human response (higher is better), (2) recall@k, i.e. existence of the human response in top-k ranked responses, and (3) NDCG on a subset of answers for which the relevance scores are provided (the more relevant the better).

## 4. Baseline Models

### 4.0.1. VISDIAL-BERT

While prior work focused on training models for the task of VisDial in isolation, in [Murahari et al., 2020], the authors propose first pre-training on related vision language datasets before fine-tuning on the visual dialog task. The authors also leverage the, then, recently proposed ViLBERT [Murahari et al., 2020] model for multi-turn visually-grounded conversations.

ViLBERT uses two Transformer-based encoders, one for each of the two modalities - language and vision - and uses co-attention layers to attend over inputs from one modality conditioned on inputs from the other. First, The authors pretrain ViLBERT on Conceptual Captions [Sharma et al., 2018] ( $\leq 2$  sentences in length) and VQA [Agrawal et al., 2016] (image question answer pairs so more related to Visual Dialog) datasets. This is followed by finetuning for the Visual Dialog task which requires a different input representation to handle its long dialog history.

This model by itself outperforms prior work by  $> 1\%$  absolute and achieves the state-of-the-art on VisDial1.0. Additionally, the authors finetune the model further on the provided dense annotations that contain the relevance scores

for all 100 answer options corresponding to each question on a subset of the training set. They note a 10% improvement in the NDCG score while hurting the MRR by more than 17%.

#### 4.0.2. FACTOR GRAPH ATTENTION

The authors in this baseline have proposed a unified attention mechanism based on graph like interactions called factor graph attention. They argue that current attention mechanism face challenges when they need to attend to a lot of multimodal data (utilities). Their method outperforms state-of-the-art methods by 1.1% for VisDial0.9 and by 2% for VisDial1.0 on MRR. Their ensemble model improved the MRR score on VisDial1.0 by more than 6%. They propose a general factor graph that combines any number of utilities. The nodes in their graphs represent utilities and interactions between them are modelled by factors.

Mathematically, they define factor graph in terms of visual dialog setting. The different utilities in visual dialog are Image  $I$ , answers  $A$ , caption  $C$ , the history of past interactions  $(H_{Q_t}, H_{A_t})_{t \in \{1, \dots, T\}}$ . The whole set of utilities are then  $\mathcal{U} = \{I, A, C, (H_{Q_t}, H_{A_t})_{t \in \{1, \dots, T\}}\}$ . Each element in the utility set can further consist of basic entities. For instance, question can be divided into words, images in different regions or objects. Hence, each utility  $U_i$  is considered as a matrix of dimensions  $n_i \times d_i$  where  $n_i$  is the local entity and  $d_i$  is the dimension of the features. Once we have identified the utilities, the process of learning attention is three folds.

#### 4.0.3. LOCAL FACTORS

In this entity information and entity interactions are modelled within one utility. Entity information is the parameterization of the particular entity  $u_i \in U_i$  and thus is given by

$$\psi_i(u_i) = v_i^\top \text{relu}(V_i \hat{u}_i)$$

where  $V_i$  is the learned parameters.  $u_i$  is the result of the embedding model such as LSTM. This calculates the factor dependency in terms of cosine similarity between two transformed representations.

#### 4.0.4. JOINT FACTORS

Similar to the entity interactions, joint factor calculates interactions between entity of one utility and the other utility. It is worth noting that they perform batch-normalization during training and L2 normalization on  $u_i, u_j$  to prevent one utility element to negatively bias another one.

#### 4.0.5. EMBEDDINGS

To calculate image embedding they use the output of the final convolution layer of VGG network. The spatial dimension of that layer is  $7 \times 7 \times 512$ . For textual entities including questions, answers and history, they divide the sentence into max  $n_i$  words. Each word is represented by one-hot encoding of the word index and is linearly transformed. These are passed to LSTM layers to get a combined embedding of the utility.

They concat the attended vectors of all utilities to get a final vector  $L = d_I + d_Q + d_C + d_A + d_H$ . We use these concatenated features and pass it to the projection network for contrastive learning as described in section below. They combine all the answer choices separately with this vector and pass it to an MLP to get a probability distribution on all answers.

They do use multi-class cross entropy loss for training their model. One possible area of improvement in their model can be addition of more loss terms that do not penalize similar answers but penalize dissimilar more (Contrastive approach). We explain this approach in more detail as a new research direction to improve this model further.

#### 4.0.6. RANKING TOGETHER UNRELATED ANSWERS

For many of the binary answer questions, the FGA model is seen to give unrelated answer choices as the top rankings (e.g. one of each of ‘yes’, ‘no’, and ‘can’t tell’). Some examples can be seen in Figure [1]. This is seen in spite of the model being highly certain about its first answer choice and the dataset containing several variations of answers that are similar to the top ranked answer in the 100 answer choices provided for each question (e.g. ‘yes’, ‘yup’, ‘looks like it’, ‘yeah’. Similarly, ‘no’, ‘nope’, ‘i don’t think so’). The expectation is that similar answer variations should appear closer together in the ranking. This is observed in Visdial-BERT. In FGA, however, there are many examples of the top ranking answers being very different from each other. This is a possible explanation for its higher MRR score but significantly lower NDCG score as compared to VisDial-BERT.

## 5. Proposed Approach

### 5.1. Data Augmentation and Contrastive Learning

We aim to improve robustness of the model in two stages. First, we aim to use data augmentation strategies to increase the variations while training the model. Second, we use these augmented data to generate positive and negative samples and use contrastive loss functions. We describe these approaches below in detail.

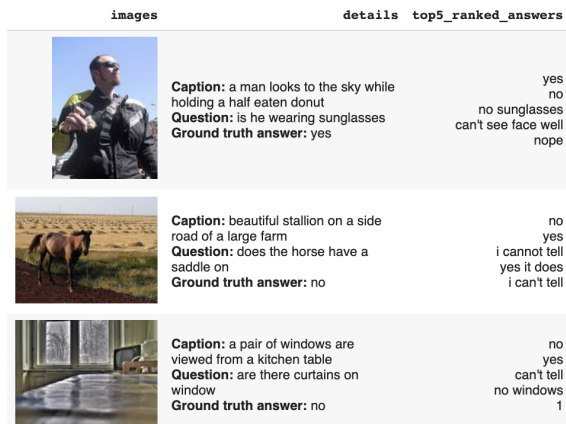


Figure 1. Qualitative examples of unrelated answer variations ranked close together in Factor Graph Attention

### 5.1.1. DATA AUGMENTATION

Data augmentation is a common and effective strategy in the deep learning community to make models more robust to variations of similar input. We plan to use this strategy in following ways :

1. **Query Paraphrasing:** We intend to build a query generator that outputs semantically similar but syntactically different queries to a given query. Prior work like [Shah et al., 2019], [Tang et al., 2020], [Jiang et al., 2018], [Kant et al., 2021] have used this technique to make the model robust to query perturbations. Recently, transformer based methods based on Hugging Face data has been used to generate paraphrased queries. We use one such model in our experiments to generate such augmentation on queries. Specifically, we generate 3 augmentations for each query in the dataset.
2. **Pyramid of Images:** As popular in training most of the models with image as the input modality, we also plan to input pyramid of images at different scales. Hence, we modify the image with a random scale while generating batches. The intuition is that different scales can capture features of objects appearing in different sizes better. It can then help in the downstream task of answer prediction.
3. **Random box blur:** This is another image augmentation in which we randomly select a box with a scale of (0.25, 0.4) of the image width and image height. We chose the corner of this box randomly and apply Gaussian blur on the image region enclosed by the box. The idea behind this is that the model should try to attend to different parts of the image to answer queries. Also, it will be inherently more robust to noise in images at the test time.

### 5.1.2. CONTRASTIVE LEARNING

FGA model as described above uses Cross Entropy (CE) Loss on the ground truth answer to train the model. We added another loss term specific to contrastive learning at the end as shown in the figure 2. The procedure to add this loss and the changes made in the base model FGA are outlined below. Contrastive Learning requires positive and negative pairs so that it can decrease the embedding distance between positive and reference samples and increase the embedding distance between negative and reference.

1. **Positive Samples :** To generate positive samples, we used data augmentation strategies as described above for each domain (image and query). The images after augmentation are sent to a VGG feature extraction pipeline to get a feature vector of dimension 512x49. The augmented query follows the same procedure as described in the FGA baseline to generate the features.
2. **Negative Samples :** We used image negatives and question negatives as the negative samples for the query part. Image negatives are queries which has the same image but different answers. Question negatives are those queries which are highly similar to each other but has different answers. For image domain, negative samples are randomly selected from the set of images excluding the reference image.

As described above for each ref data  $(I_{ref}, Q_{ref})$ , we randomly select one positive sample and 3 negative samples  $(I_{pos}, I_{neg1}, I_{neg2}, I_{neg3})$ . Similarly, for query we get  $(Q_{pos}, Q_{neg1}, Q_{neg2}, Q_{neg3})$ .

These data tuples are then sent to multiple forward passes as described in the FGA model. The features of history are concatenated into a single vector of dimension 2368 features. These features are then sent to a separate projection network consisting of 2 linear layers which outputs the final feature vector of size 200 for all the samples. The final supervised loss formulation can be given as follows:

$$\mathcal{L}_{SC}^i = - \sum_{p=1}^{|\mathcal{X}^+(x_i)|} \log \frac{\exp(\Phi(z_i \cdot z_p) / \tau)}{\sum_{k=1}^K \mathbb{1}_{k \neq i} \cdot \exp(\Phi(z_i \cdot z_k) / \tau)}$$

where,  $z_i$  is the reference,  $z_p$  is a positive sample,  $z_k$  are negative samples and  $\tau$  is a temperature parameter to control stability

## 5.2. Heuristic Scores

To our knowledge, models trained on the VisDial dataset are largely trained to optimize the ranking of a single ground truth answer. This is because while there exist several similar and relevant answers in the answer options, the VisDial

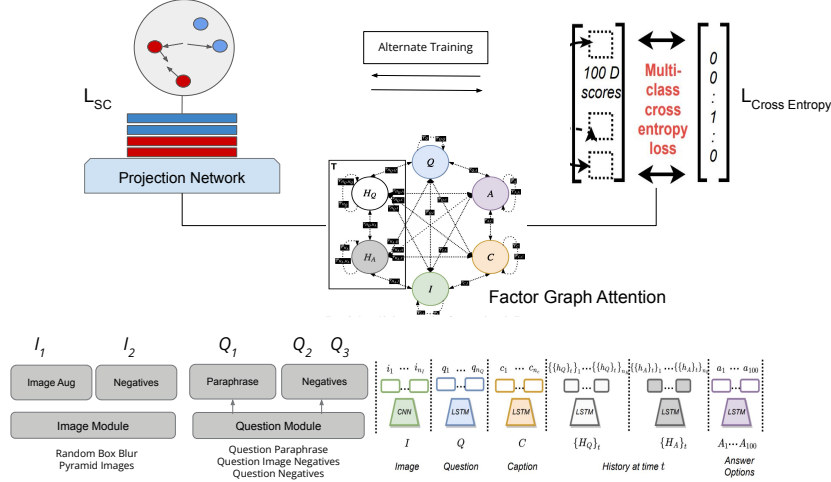


Figure 2. Proposed Model Architecture

training dataset does not contain the information of their relevance. The issue with this training approach is that the models are oblivious to the other variations present in the answer options that are similar to the ground truth answer and tend to rank them lower. In an ideal ranking, the model should be aware of the meaning of the various answer options and rank similar answers to the ground truth closer to the top. To provide this knowledge at the time of training, we propose the use of heuristic scores that provide the model with the information of answer similarity. We hypothesize that using this information, the model will be able to better learn the relations between different answer variations and exude certainty in its ranking. We argue that such a model would be more reliable and deploy-able in a real world setting. In this approach, we use a semi-supervised technique using uni-modal and multi-modal features to generate heuristic scores.

- **Uni-modal Heuristic Scores** - First, we obtain sentence embeddings for all answer option using the Sentence-BERT [Reimers & Gurevych, 2019] model that uses siamese and triplet network structures to derive semantically meaningful sentence embeddings of dimension 768. Using these embeddings, we compare the ground-truth answer to the other answer options provided using cosine-similarity and empirically select a threshold of 0.78 to select answer options that are semantically close to the ground truth answer. We assign relevant scores in proportion to the similarity to the ground truth answer.
- **Multi-modal Heuristic Scores** - For creation of the multi-modal heuristic scores, we use the ViL-BERT model to obtain multi-modal fused representations. The ViLBERT model we use is the same

architecture of the VisDial-BERT baseline explained above. It is pre-trained on the VQA and Conceptual Captions dataset and further fine-tuned on the VisDial dataset. The fused representation is obtained by passing the image  $I$ , caption and history  $H = (\underbrace{C}_{H_0}, \underbrace{(Q_1, A_1)}_{H_1}, \dots, \underbrace{(Q_{t-1}, A_{t-1})}_{H_{t-1}})$ , the current ques-

tion  $Q_t$ , and each answer option  $A_t^{(i)}$  through the model. Using the fused embeddings of 1024 dimension obtained from the model, we compare the various answer options to the ground-truth using cosine similarity and using follow the same procedure of assigning relevance scores to the related answer options as the Uni-model Heuristic Scores.

We utilize these heuristic scores in the range of (0-1.0) to train our model. Concretely, we use the model's predicted likelihood scores  $\hat{\ell}_t^{(i)}$  for each answer option  $A_t^{(i)}$  at round  $t$ , normalize these to form a probability distribution over the 100 answers  $\hat{y}_t = [\hat{y}_t^{(1)}, \dots, \hat{y}_t^{(100)}]$ , and then compute a cross-entropy (CE) loss against the normalized ground-truth relevance scores  $\hat{y}_t$ , given by

$$-\sum_i y_t^{(i)} \log \hat{y}_t^{(i)}$$

## 6. Experimental Setup

### 6.1. Dataset

We use the Visual Dialog [Das et al., 2017] dataset for our project. VisDial is an AI task where the machine must hold dialog with a human about visual content. 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. When compared to VQA, VisDial tries to solve significantly richer task (dialog), overcomes a 'visual priming bias' in VQA (in VisDial, the questioner does not see the image), contains free-form longer answers, and is order of magnitude larger.

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

Table 1. Data distribution on VisDial 1.0

## 6.2. Evaluation metrics

The discriminative model is evaluated on retrieval metrics - (1) rank of ground-truth response (lower is better), (2) recall ( $R@1,5,10$ ), i.e. existence of the ground-truth response in top-k ranked responses, and (3) mean reciprocal rank (MRR) of the ground-truth response (higher is better). This evaluation protocol is compatible with both the discriminative models (that simply score the input candidates e.g. via a softmax over the options, and cannot generate new answers), and generative models.

Since some of the candidate options may be semantically identical (eg. 'yeah' and 'yes'), each candidate answer is assigned a relevance score based on human annotations. Authors of VisDial, recently released dense annotations i.e. relevance scores (0-1.0) for all 100 answer options from  $A_t$  corresponding to the question on a subset of the training set. Using this, the normalized discounted cumulative gain (NDCG) is reported over the top K ranked options, where K is the number of answers marked as correct by at least one annotator. NDCG is invariant to the order of options with identical relevance and to the order of options outside of the top K.

## 6.3. Experimental Methodology

### 6.3.1. CONTRASTIVE LEARNING

For this approach, we use the same official train and val split as described in section 7.1. We randomly sample the image related and query related positive and negative pairs with a probability of 0.4 and 0.6 respectively. The network along with the projection network as described in the section 6.3.2 is trained end-to-end. We use an alternate strategy to train the model. The parameters are back propagated with contrastive loss and Cross entropy loss with a ratio of 1:3. This means that 1/3 of the batches use contrastive loss while the rest use CE loss. It has shown to improve performance in [cite]. We experiment with multiple learning rates from  $1e-2$  to  $1e-4$  and train the model for 8 epochs. Finally, we show the results on the evaluation metrics as described in section 7.2.

### 6.3.2. TRAINING WITH HEURISTIC SCORES

To train the FGA model with custom heuristic scores, we also do an ablation study by first running the model with only one relevant answer, the ground-truth answer. Next, we create Uni-Modal heuristic Scores using just the language modality. We train the FGA model with these uni-modal heuristic scores. We experiment with three different thresholds of cosine similarity to the ground-truth answer and empirically choose 0.78 as the threshold looking at the model performance. We train the model using FastRCNN image features and LSTM text embeddings for the input modalities. We follow the original splits of the VisDial 1.0 dataset using  $\sim 120k$  for training and  $\sim 2k$  for validation. We replace the targets in the training set with our generated heuristic scores.

Next, we train the FGA model with Multi-modal Heuristic Scores. Since the creation of these scores require a forward pass for each of the 100 answer options in every round of every dialog, this was not feasible in the time-frame and limited resources that we were working with. We extracted the embeddings for 9350 dialogs where each dialog contains 10 rounds of question and answers. Empirically, we choose the cosine similarity of 0.8 to perform the training of the model. Yet again, we train the FGA model with the 9350 dialogs from the VisDial 1.0 training set. Here, we replace the targets with our Multi-modal Heuristic Scores. For a fair comparison with single target base model and the uni-modal heuristic scores, we train the model with the same subset of the data. We keep the various hyper-parameters such as learning rate constant throughout the experiment. We train the FGA model to convergence for  $\sim 9$  epochs. The validation of the method is done using the provided dense annotations for the validation set. The primary metric tracked for performance is NDCG, which gives us the information of cumulative gain of ranking the relevant answers as annotated by humans. Additionally, we also track MRR and  $R@1$ .

## 7. Results and Discussion

We list findings from all our experiments in this section.

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*	0.554	0.4687	41.09
FGA (VGG) + SC loss	0.54	0.47	0.41
FGA (F-RCNNx101) + UHS	0.602	<b>0.611</b>	50.60

Table 2. Supervised Contrastive loss (SC) and Heuristic Scores (UHS) experiments results on VisDial 1.0 val set

### 7.1. Contrastive Learning

We trained the FGA model with the contrastive loss as per the architecture shown in figure 2. The results can be seen in the table 2 at row "FGA (VGG) + SC loss". We used VGG features for images. We can see that MRR reached to 0.54 while NDCG reached to 0.47. This model was not able to outperform the baseline VGG model. We attribute this to the following possible reasons. First, we observe that the loss that we used was quickly able to learn the positive and negative cases. Hence, the loss did not contribute after a few iterations. This could be because of the difficulty in selecting hard-negatives. Second, it may be possible that the data is not enough to learn good representations using contrastive learning approach. Pretraining the network with some large data on similar problem would have helped. But due to time constraints, we could not validate this hypothesis.

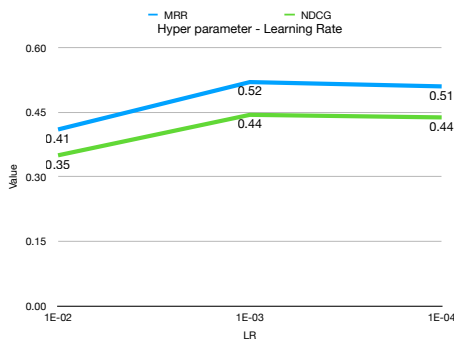


Figure 3. Effect of changing LR in contrastive learning on evaluation metrics

To further improve the accuracy of this approach, we tried various hyper-parameter tuning. The effect of learning rate change can be seen in the figure 7.1. We found 1e-3 to work best in our case. We also try different temperature parameters for the contrastive loss function. We find 1.0 to work best for us as can be seen in the figure 4

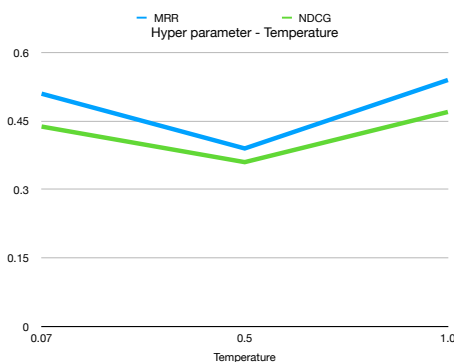


Figure 4. Effect of changing Temperature in contrastive learning on evaluation metrics

We also show the model performance qualitatively on paraphrased queries. As can be seen in the figure 5, the baseline model fails to answer the query correctly in one of the cases

inspite of the fact that the grass is visible in the image and the query is similar. This shows that such models tend to overfit on queries or image features.



Figure 5. Comparison of baseline and our model on query paraphrasing (robustness)

### 7.2. Heuristic Scores



Figure 6. Qualitative results of using Uni-modal Heuristic Scores (UHS) for FGA training

From Table 2, we see that training the FGA model using Heuristic Scores has been able to improve upon the NDCG significantly though the MRR takes a hit as compared to the baseline FGA model. The FGA model trained on Uni-modal Heuristic Scores (UHS) has an NDCG of 0.611 and an MRR of 0.602 which signifies an improvement of absolute  $\sim 5$  in the NDCG using just the language modality to compare the different answer options. The MRR decreases by absolute  $\sim 6$  but this is expected since our aim is to improve upon all the relevant answers and not just the ground-truth answer. We note the improvement in NDCG in our qualitative analysis as well. As seen in figure 6, the FGA + UHS model is able to rank the related answer options close to each other

with more certainty. For the question, "Is it raining?", the base model has ranked really diverse answer options such as "Yes", "It may have stopped", and "No". The FGA+UHS model on the other hand has been able to rank more complicated but still related answer options to the top. For the question "What color is the ground?", it is able to rank the related colors to the top "Reddish brown", "red and white", and "brown". While the base FGA model ranks together unrelated answer options.



Figure 7. Qualitative results showing reason for low MRR in the Uni-modal Heuristic Scores (UHS) approach

The decrease in the MRR is misleading because while the model is better able to rank the related answers to the top, it may not necessary rank the ground-truth answer specified by the dataset at the top. In figure 7, we see that for the question, "Can you see any grass?", the FGA+UHS model ranks "Yep" to the top while the ground-truth answer "Yes" appears at the second rank. Additionally, the base FGA model has a high MRR score also owing to the fact that it ranks the popular answer choices in the dataset such as "Yes", "No", and "I can't tell" to the top. Even though it isn't certain about the correct answer for a question, it could get a high MRR just for ranking the popular answer options to the top.

Model	MRR	NDCG	R@1
Baseline FGA (VGG)	0.412	0.347	50.58
FGA (VGG) + UHS	0.3815	0.456	52.75
FGA (VGG) + MHS	0.3750	0.407	48.49

Table 3. Comparison of Uni-modal Heuristic Scores and Multi-modal Heuristic Scores on 9350 dialogs containing 10 rounds each

In Figure 3, we see that while Multi-modal Heuristic Scores have shown an improvement on the NDCG over the subset of the data which is  $\sim 9k$  samples, the improvement is not significantly greater than the Uni-modal Heuristic Scores. We attribute this to the fact that language modality has most of the relevant information needed to be able to judge which of the answer options provided are closer to the ground-truth answer. The image, the caption, and the dialog history are not able to significantly enrich this information to be able to give us better heuristic scores in the multi-modal setting.

In Figure 8, we see that while the multi-modal heuristic scores are able to give relevant answers that have more



Figure 8. Qualitative results showing improvement of answer options ranking using Multi-modal Heuristic Scores (MHS)

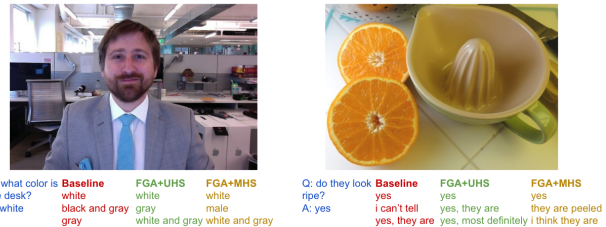


Figure 9. Qualitative results showing noise in ranking using Multi-modal Heuristic Scores (MHS)

information than just the ground-truth answers for example, the answer "I think there is some syrup bottle on the tale" for the question "Are there any drinks on the table?" when the ground-truth answer is just "Yes, some". However, there are several cases where the answers ranks in the Multi-modal Heuristic Scores case are irrelevant too. For example, the answer "Yes, they are peeled" for the question "Do they look ripe?" in Figure 9.

## 8. Conclusion and Future Work

In this paper, we have proposed two approaches to improve upon the overall robustness and certainty in the answer rankings of our baseline FGA model using multi-modal approaches such as Data Augmentation, Supervised Contrastive loss, and creation of Heuristic Scores. We discuss the learnings from these approaches and show improvements on the FGA model performance on the NDCG metric.

For future work, we suggest pre-training the FGA model with large amount of VQA data before fine-tuning on the VisDial dataset. Tuning the hyper-parameters for Supervised Contrastive (SC) loss can improve the model's performance further. Additionally, we plan to experiment with joint training of Multi-modal Heuristic Scores along with the alternate training with SC loss.



## References

- Agarwal, S., Bui, T., Lee, J.-Y., Konstas, I., and Rieser, V. History for visual dialog: Do we really need it?, 2020.
- Agrawal, A., Lu, J., Antol, S., Mitchell, M., Zitnick, C. L., Batra, D., and Parikh, D. Vqa: Visual question answering, 2016.
- Das, A., Kottur, S., Gupta, K., Singh, A., Yadav, D., Moura, J. M. F., Parikh, D., and Batra, D. Visual dialog, 2017.
- Gupta, T., Vahdat, A., Chechik, G., Yang, X., Kautz, J., and Hoiem, D. Contrastive learning for weakly supervised phrase grounding, 2020.
- He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning, 2020.
- Hénaff, O. J., Srinivas, A., Fauw, J. D., Razavi, A., Doersch, C., Eslami, S. M. A., and van den Oord, A. Data-efficient image recognition with contrastive predictive coding, 2020.
- Jiang, Y., Natarajan, V., Chen, X., Rohrbach, M., Batra, D., and Parikh, D. Pythia v0.1: the winning entry to the vqa challenge 2018, 2018.
- Kant, Y., Moudgil, A., Batra, D., Parikh, D., and Agrawal, H. Contrast and classify: Training robust vqa models, 2021.
- Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., and Krishnan, D. Supervised contrastive learning, 2021.
- Murahari, V., Batra, D., Parikh, D., and Das, A. Large-scale pretraining for visual dialog: A simple state-of-the-art baseline, 2020.
- Reimers, N. and Gurevych, I. Sentence-bert: Sentence embeddings using siamese bert-networks, 2019.
- Shah, M., Chen, X., Rohrbach, M., and Parikh, D. Cycle-consistency for robust visual question answering, 2019.
- Sharma, P., Ding, N., Goodman, S., and Soricut, R. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. 2018.
- Tang, R., Ma, C., Zhang, W. E., Wu, Q., and Yang, X. Semantic equivalent adversarial data augmentation for visual question answering, 2020.
- Wu, Z., Xiong, Y., Yu, S., and Lin, D. Unsupervised feature learning via non-parametric instance-level discrimination, 2018.

## 9. Appendix




images	details	top5_ranked_answers
	<b>Caption:</b> a man looks to the sky while holding a half eaten donut <b>Question:</b> is he wearing sunglasses <b>Ground truth answer:</b> yes	yes no no sunglasses can't see face well nope
	<b>Caption:</b> beautiful stallion on a side road of a large farm <b>Question:</b> does the horse have a saddle on <b>Ground truth answer:</b> no	no yes i cannot tell yes it does i can't tell
	<b>Caption:</b> a pair of windows are viewed from a kitchen table <b>Question:</b> are there curtains on window <b>Ground truth answer:</b> no	no yes can't tell no windows 1

Figure 10. Qualitative examples of unrelated answer variations ranked close together in Factor Graph Attention





images	details	top5_ranked_answers
	<b>Caption:</b> the fruits and vegetables are on dishes on the table <b>Question:</b> how many dishes are on the table <b>Ground truth answer:</b> 5	2 3 4 1 5
	<b>Caption:</b> 2 baseball teams playing on a baseball field <b>Question:</b> how many players can you see <b>Ground truth answer:</b> 5	2 3 4 8 5
	<b>Caption:</b> the man is sitting on the sofa ready to eat his pizza <b>Question:</b> how many people are there <b>Ground truth answer:</b> 2	1 2 3 i cannot tell 0
	<b>Caption:</b> a living room with black furniture and a dining room behind it <b>Question:</b> how many chairs are there in the dining room <b>Ground truth answer:</b> about 10	4 2 5 3 i cannot tell

Figure 11. Qualitative examples of count answer type issues in Factor Graph Attention





images	details	top5_ranked_answers
	<b>Caption:</b> a woman flying a brightly colored kite in a cloudless sky <b>Question:</b> is the kite red <b>Ground truth answer:</b> yes, partially	nope no not that i can see yes, it is yes it is
	<b>Caption:</b> 2 small red planes are in the sky <b>Question:</b> are the planes big <b>Ground truth answer:</b> no, on the small side	yes yes they are it looks like it yes, they are i think so
	<b>Caption:</b> a person holding a peeled and overly ripe banana <b>Question:</b> do you see any logos <b>Ground truth answer:</b> chiquita	no nope not that i can see yes yes ,one
	<b>Caption:</b> a snowboarder is sitting on the snow with a snowboard on their feet <b>Question:</b> are they wearing goggles <b>Ground truth answer:</b> i can't see their face	yes yes they are i can't tell can't tell i cannot tell

Figure 12. Qualitative examples of incorrect answers by Visdial-BERT





images	details	top5_ranked_answers
	<b>Caption:</b> 2 canadian geese take their 3 goslings for a walk in the wet parking lot <b>Question:</b> are the geese walking in a straight line <b>Ground truth answer:</b> nope	yes no nope yes they are they are running
	<b>Caption:</b> a snowboarder makes a jump with his snowboard at a ski resort <b>Question:</b> are there any trees <b>Ground truth answer:</b> yes	no yes in the far background in the background yes in the background
	<b>Caption:</b> a train moving down the tracks while a car waits <b>Question:</b> is the train moving <b>Ground truth answer:</b> yes	i cannot tell no yes i don't think so i can't tell
	<b>Caption:</b> city outdoor marketplace with people shopping and a bike loaded with bananas <b>Question:</b> are the bananas ripe or green <b>Ground truth answer:</b> both	green both i cannot tell mostly green some are ripe and some are not

Figure 13. Qualitative examples of incorrect answers by Factor Graph Attention

## Robust Factor Graph Attention Net





images	details	gt_answer_text	top5_ranked_answers
	<p><b>Caption:</b> a close up of a red fire hydrant with chinese writing on it</p> <p><b>Question:</b> are the people chinese</p>	yes	<p>i cannot tell</p> <p>no</p> <p>yes</p> <p>i can't tell</p> <p>i can't tell for sure, there hands are behind there back or in their pockets</p>
	<p><b>Caption:</b> 2 cows standing my water bins in a field</p> <p><b>Question:</b> how many people are there</p>	2	<p>i cannot tell</p> <p>0</p> <p>no people</p> <p>there are 0</p> <p>no</p>
	<p><b>Caption:</b> a man has his mouth wide open eating a hotdog</p> <p><b>Question:</b> does it look good</p>	sure i guess so	<p>yes</p> <p>no</p> <p>i cannot tell</p> <p>not really</p> <p>sure does!</p>
	<p><b>Caption:</b> a pitcher is captured in mid-pitch on the mound</p> <p><b>Question:</b> is there a lot of dirt on the ground</p>	no	<p>yes</p> <p>yes there is</p> <p>no</p> <p>just an inch or so,</p> <p>yes, it is</p>

Figure 14. Examples of dataset issues