## **Supplementary Material: Factor Graph Attention**

Idan Schwartz<sup>1</sup>, Seunghak Yu<sup>2,\*</sup>, Tamir Hazan<sup>1</sup>, Alexander Schwing<sup>3</sup>

<sup>1</sup>Technion

<sup>2</sup>MIT CSAIL

<sup>3</sup>UIUC

idansc@cs.technion.ac.il, seunghak@csail.mit.edu, tamir.hazan@technion.ac.il, aschwing@illinois.edu

In this supplementary material we provide additional quantitative (see Sec. 1) and qualitative (see Sec. 2) results.

## 1. Quantitative Evaluation

**Ensemble model for VisDial v0.9:** For VisDial v1.0, a simple ensemble technique has significantly improved the results as discussed in the main paper. We observe a similar effect for VisDial v0.9, pushing the current state of the art for MRR from 0.6525 to 0.6892 as summarized in Tab. 1. We achieve this result with an ensemble of 9 models which differ only by the initial seed. For VisDial v1.0 we report a 5 model ensemble score. Due to restriction of the number of submissions to the evaluation server we could not evaluate a larger ensemble model. The results in Tab. 1 suggest that the VisDial v1.0 score in the paper can be further improve with a larger ensemble model.

Analysis of Factor Graph Attention weights: To infer the attention belief for a utility, i.e.,  $b_i(u_i)$ , we aggregate marginalized joint and local interactions and also local-information and prior terms. To calibrate each cue, we use scalar weights, i.e.,  $\hat{w}_i, w_i, w_{i,j}$ . To obtain a better understanding of the reasoning process and analyze attention, we suggest an importance score:

$$S(\gamma) = \frac{|m_{\gamma} \cdot \gamma|}{\sum_{\delta \in \{\hat{w}_{i}, w_{i}, (w_{i,j})_{j \in \mathcal{U}}\}} |m_{\delta} \cdot \delta|},\tag{1}$$

where  $\gamma \in \{\hat{w}_i, w_i, (w_{i,j})_{j \in \mathcal{U}}\}$  is the weight of a cue and  $m_\gamma$  is the mean term of the corresponding cue  $\gamma$ , which was calculated over the entire validation set. Note that  $\gamma$  are the scalar weights.  $S(w_{i,j}) \ \forall j \in \mathcal{U}$  captures the importance of the j-th cue for utility i. A high score means the i-th utility attention belief heavily relies on cue j. Similarly,  $S(w_{i,i}), S(w_i), S(\hat{w}_i)$  capture the importance of local-interactions, local-information and prior cues for the i-th utility. We report the scores in Tab. 2. We observe that the answer utility relies mostly on local-interactions. The question heavily relies on the prior, but also makes use of history answers and question cues. The caption ignores all utilities other than the prior. The image question utility is the most important cue. Interestingly, we observe importance of priors. Image attention relies on the captions, while the caption ignores all the cues and preserves the prior behavior. The history question and answers rely on the question and the local factors.

Computation and insignificant interactions: Upon training interactions may be found to be unnecessary. Our model can be optimized easily: 1) The score in Eq. (1), can be used to omit less significant interactions. Previous multimodal attention doesn't model pairwise interaction scores, making it hard to eliminate computations. 2) For the same image but different question, we can re-use calculated joint interactions, such as local-interaction, image-caption, *etc*. This is impossible for approaches that pool cues since the question changes. 3) As mentioned in Sec. 4.2, it's possible to share weights between similar utilities, *e.g.*, different history questions/answers.

Currently, we don't consider most of those options, as the model trains quickly (8 hours vs. 33 hours of previous state-of-the-art) and fits into a single 12GB GPU.

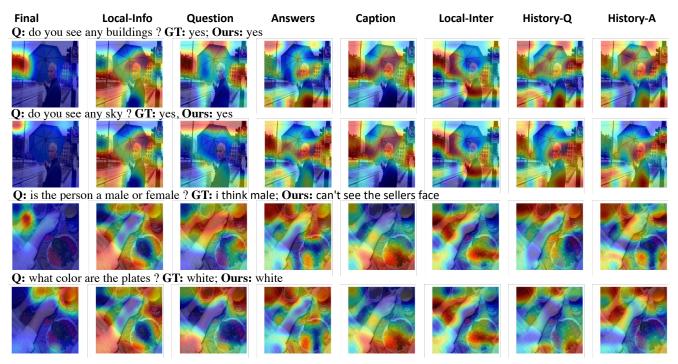


Figure 1: Two images each with two questions. We illustrate scores obtained from different types of factors. Local-info denotes 'Image-Local-Information,' Question refers to 'Image-Question,' *etc.* We observe 'Image-Question' to have the highest variance between different questions, since its heat map differs the most. 'Image-Question' also correlates the most with the final attention.

Table 1: Analysis of ensemble models for VisDial v0.9. With an ensemble of 9 models we observe an improvement of more than 3% over the single model.

Model	MRR	R@1	R@5	R@10	Mean
FGA	0.6525	51.43	82.08	89.56	4.35
Ensemble of 2 FGA	0.6711	53.56	83.83	90.97	3.92
Ensemble of 3 FGA	0.6786	54.28	84.71	91.69	3.73
Ensemble of 4 FGA	0.6819	54.56	85.19	92.10	3.62
Ensemble of 5 FGA	0.6848	54.82	85.57	92.38	3.55
Ensemble of 6 FGA	0.6860	54.95	85.71	92.52	3.50
Ensemble of 7 FGA	0.6869	54.97	85.91	92.67	3.47
Ensemble of 8 FGA	0.6881	55.10	86.04	92.77	3.44
Ensemble of 9 FGA	0.6892	55.16	86.26	92.95	3.39

Table 2: For each utility (column) we show the three most related cues based on the S score given in Eq. (1). We provide S in parenthesis. P,  $L_1$  and  $L_2$  indicate prior, local-information and local-interactions of the utility in the column.

A	Q	С	I	$H_Q$	$H_A$
$L_1$ (0.125)	P(0.851)	P(0.988)	Q(0.593)	Q(0.205)	$L_2(0.607)$
$H_Q$ (0.122)	$H_A(0.052)$	I (0.004)	$L_1$ (0.186)	$L_1$ (0.121)	Q(0.304)
Q(0.075)	$H_Q(0.031)$	A (0.001)	C(0.123)	$L_2 (0.085)$	$L_1$ (0.017)

## 2. Qualitative Evaluation

Factors visualization: We provide additional visualization in Fig. 1. We visualize scores for each image region obtained from different types of factors. 'Image-Local-Information,' 'Image-Caption' and 'Image-Local-Interaction' are constant for different questions, while 'Image-Question,' 'Image-Answer,' 'Image-History-Q' and 'Image-History-A' change for every question. We calculated the variance of interactions and observe that 'Image-Question' has the highest variance ( $\approx$  3), while 'Image-Answer,' 'Image-History-Q' and 'Image-History-A' have a variance of  $\approx$  1. Beyond the importance score, the high-variance also suggests that the 'Image-Question' cue is most important.

**Attention over dialogs:** In Fig. 2, we present a randomly-picked set of 50 images along with their corresponding dialogs. An automatic script is used to generate the figures. We highlight that image attention is aware of the scene in the question context, and able to attend to correct foreground or background regions. Question attention attends to informative words, and answer attention frequently correlates with the predicted answer. History attention emphasizes nuances.

Figure 2: 50 dialogs along with question, answers and history attention. The predicted answer (i.e., A) and ground-truth answer (i.e., GT) are also provided.

