Seeing is Knowing! Fact-based Visual Question Answering using Knowledge Graph Embeddings

Kiran Ramnath\textsuperscript{1} and Mark Hasegawa-Johnson\textsuperscript{1}

\textsuperscript{1}University of Illinois, Urbana-Champaign
\{kiranr2, jhasegaw\}@illinois.edu

Abstract

Fact-based Visual Question Answering (FVQA), a challenging variant of VQA, requires a QA-system to include facts from a diverse knowledge graph (KG) in its reasoning process to produce an answer. Large KGs, especially common-sense KGs, are known to be incomplete, i.e., not all non-existent facts are always incorrect. Therefore, being able to reason over incomplete KGs for QA is a critical requirement in real-world applications that has not been addressed extensively in the literature. We develop a novel QA architecture that allows us to reason over incomplete KGs, something current FVQA state-of-the-art (SOTA) approaches lack due to their critical reliance on fact retrieval. We use KG Embeddings, a technique widely used for KG completion, for the downstream task of FVQA. We also employ a new image representation technique we call ‘Image-as-Knowledge’ to enable this capability, alongside a simple one-step CoAttention mechanism to attend to text and image during QA. Our FVQA architecture is faster during inference time, being $O(m)$, as opposed to existing FVQA SOTA methods which are $O(N \log N)$, where $m =$ number of vertices, $N =$ number of edges $= O(m^2)$. KG embeddings are shown to hold complementary information to word embeddings: a combination of both metrics permits performance comparable to SOTA methods in the standard answer retrieval task, and significantly better (26% absolute) in the proposed missing-edge reasoning task.

1 Introduction

Multi-modal reasoning has been a prime focus in the pursuit of artificial general intelligence. One such task is Visual Question Answering (Antol et al., 2015; Goyal et al., 2017), which requires the system to answer a textual question based on an image. VQA is challenging because it requires many capabilities such as object detection, scene recognition, activity recognition in addition to language understanding and commonsense reasoning. (Wang et al., 2016) noted that most of the questions in the VQA dataset didn’t require commonsense knowledge residing outside of the image. Humans have a remarkable ability to blend in knowledge from their own prior experiences when answering a question about an image. (Wang et al., 2016) therefore introduced the FVQA benchmark, aiming to provide a more challenging set of questions which ensure that the answer to a given question requires some form of external knowledge, not present in the image or the question text. They provide this external information in the form of knowledge graphs, which are multi-relational graphs, storing relational representations between entities. The task of FVQA boils down to retrieving the correct entity as the answer most relevant to an image. For example, in Fig. 1 the fact triple – (Banana, HasProperty, rich in Potassium) is the external information required to answer the question; with the correct answer being ‘Banana’.

All published methods in the literature rely on
retaining the ground-truth fact for a question, whereas a real-world application will likely have questions asked about facts not present in the KG. The primary contribution of this paper is a method Seeing is Knowing (SiK) that permits FVQA to reason about common-sense facts that are absent from the knowledge graph using KG Embeddings. When used with a more general score function we define, the resultant framework works well in both complete and incomplete KG settings. KG embeddings permit us to offer two additional contributions to the SOTA in FVQA: an ‘image-as-knowledge’ representation of visual information, and a CoAttention method for combining visual and textual inputs. ‘Image-as-knowledge’ represents the image as the span of the KG embedding vectors for the entities found in it. Representing an image with textual semantics has been attempted before (Li et al., 2019), but not using KG embeddings; KG embeddings provide robustness to incomplete KGs by encoding information about the graph structure. CoAttention uses the words of the query to compute weighted combinations of the entity vectors, thus the answer is a vector within the span of the image entities. Two such identical models are trained separately, which are gated in order to query the entities in the KG during inference. Finally, SiK when used in a standalone fashion over incomplete KGs is more time-efficient during inference. It is $O(m)$ as it only needs to reason over existing nodes in the network. To the best of our knowledge, our work is the first one to apply KG embeddings to a VQA task.

2 Related Work

2.1 Knowledge Graphs

Knowledge graphs (Suchanek et al., 2007; Auer et al., 2007; Mitchell and Fredkin, 2014; Bollacker et al., 2008; Dong et al., 2014) have been studied as an effective way of representing objects or concepts and their inter-relationships. Such relational representations are formally defined in the RDF (Resource Description Framework) as triples $f = (\text{subject}, \text{predicate}, \text{object})$, where (subject, object) are entities, predicate is the relation connecting the two entities. Such linked representations have been found to correlate highly with how humans process cognitive information (Halford et al., 2010).

2.2 KG embeddings

Common-sense KGs extracted from web-scale texts are usually incomplete. KG embedding techniques (Bordes et al., 2013; Sun et al., 2019; Socher et al., 2013; Nickel et al., 2011; Dong et al., 2014; Dettmers et al., 2018) have been studied as a means to remediate incompleteness of large-scale KGs. Score-based latent feature models, which perform well on several benchmarks, learn a score mapping $\phi(h, r, t) : E \times R \times E \to \mathbb{R}$ where $E$ is the set of all entities, $R$ is the set of all relation-types. $h, t \in E$ are the head (subject) and tail (object), $r \in R$ is the relation. $G \subseteq E \times R \times E$ is the observed set of edges, while $G_o \supset G$ is the unknown set of true edges. The embeddings $(h, r, t)$ are learned so that the score $\phi(.)$ is high for edges in $G_o$, and low for edges in its complement.

We use Entity-Relation Multi-Layer Perceptron (ERMLP) (Dong et al., 2014) which uses an MLP to produce the score $\phi(h, r, t)$ for each fact triple. Distance-based models learn embeddings $h, r$ and $t$ in order to minimize the distance between $t$ and $f(h, r)$, for some projection function $f(.)$. We explore the use of two such distance-based models (results in Appendix): TransE (Bordes et al., 2013) and RotatE (Sun et al., 2019). TransE models $f(h, r) = h + r$. RotatE models $f(h, r)$ as a Hadamard product, $h \circ r$, where the embedding vectors $h, r$ and $t$ are in the complex plane and $r$ is enforced to be unit norm, so that $h \circ r$ is a rotation in the complex plane.

Knowledge-graph question answering (KGQA) is the task of answering questions using the facts in a KG. EmbedKGQA (Saxena et al., 2020) tries to reason over missing edges for text-only multi-hop questions using KG embeddings. Our task involves the visual modality in addition, as well as reasoning over a common-sense KG.

2.3 FVQA

The fact-based visual question answering (FVQA (Wang et al., 2016)) benchmark was designed using questions that obey the following condition: for each $(\text{question}, \text{image}, \text{answer})$ triplet in the dataset $((q_i, I_i, y_i) \in D)$, there exists exactly one supporting fact in the knowledge graph $(f_j = (h, r, t) \in G)$ such that the correct answer $y_i$ is either the head or the tail of $f_j$, and such that at least one of the two entities is visible in the image.

The accompanying KG is also diverse, comprising facts from three individual KGs: We-
Figure 2: Seeing is Knowing architecture. (1) Scenes, objects, & actions are detected in the image. (2) For detected entities, we retrieve their KG embedding representations. The span of these embedding vectors represents the ‘image-as-knowledge’. (3) Lexical semantic vectors for each word in the query are accumulated via an LSTM. (4) The joint image-question encoding is derived using a CoAttention mechanism described in Fig. 3, then (5) passed through a multi-layer perceptron, whose (6) last layer is used as a pair of queries that are (7) gated, and (8) used to retrieve the entity to answer the question.

bchild (Tandon et al., 2014), ConceptNet (Liu and Singh, 2004), and DBPedia (Auer et al., 2007). In total, the dataset contains 13 relations: \( \mathcal{R} \in \{\text{CATEGORY}, \text{HASPROPERTY}, \text{RELATEDTO}, \text{LOCATION}, \text{ISA}, \text{HASA}, \text{CAPABLEOF}, \text{USEDFOR}, \text{DESires}, \text{PARTOF}, \text{RECEIVESACTION}, \text{CREATEDBY}, \text{COMparATIVE}\} \). The dataset consists of 2190 images sampled from the ILSVRC (Rusakovsky et al., 2015) and the MSCOCO (Lin et al., 2014) datasets. The accompanying KG consists of roughly 194500 facts, concerning 88606 entities. Roughly 82% of the questions have a Key Visual Concept (KVC) as the answer, whereas 18% have information from the Knowledge Base (KB) as the answer.

Answering questions in FVQA is to solve for

\[
\hat{y} = \arg\max_{e \in \mathcal{E}} p(y = e | q, I, \mathcal{G}),
\]

i.e., finding the most probable entity as the answer given a question \( q \) and image \( I \), and given the graph \( \mathcal{G} \). Our formulation of the missing edge reasoning task considers \( \mathcal{G}_e \).

(Wang et al., 2016) attempted FVQA as a parsing and querying problem, constructing 32 different templates of queries, and classifying each image-question pair as requiring one of these templates. Simple keyword matching techniques further prune the retrieved facts. Straight to the Facts (STTF) (Narasimhan and Schwing, 2018) approached FVQA by directly learning to retrieve supporting facts, where each fact-entity was represented using lexical semantic representations. (Narasimhan et al., 2018) extended this approach in Out-of-the box (OOB), the current SOTA, by using local neighbourhood-based reasoning via a Graph Convolution Network (GCN) (Kipf and Welling, 2017) to answer each question. All of these current approaches rely on the ground-truth edge to be present in the KG.

3 Our approach - Seeing is Knowing

The proposed architecture for FVQA is shown in Fig. 2. As shown, a given image \( I \) and query \( q \) are combined via CoAttention to form two entity query vectors, \( f_{KVC}(q, I) \) and \( f_{KB}(q, I) \). The KG is then queried for the answer to the question, according to

\[
\hat{y}(q|I) = \begin{cases} 
\arg\max_{e \in \mathcal{E}} f_{KVC}(q, I)^T e & \text{if } g_{KVC}(q) = 1 \\
\arg\max_{e \in \mathcal{E}} f_{KB}(q, I)^T e & \text{if } g_{KVC}(q) = 0 
\end{cases}
\]

where the gating function \( g_{KVC}(q) \in \{0, 1\} \) is equal to 1 if the text of the question indicates that the answer is visible in the image, equal to 0 otherwise.
The rest of this section addresses representations of the entities, image, and query, the information fusion functions, the gating function, and the loss function.

### 3.1 KG representation

KG Embeddings are learned by training a surrogate binary classifier to discriminate true edges from false. Since the KG only contains positive examples, we perform hard negative mining to generate negative examples using self-adversarial negative sampling (Sun et al., 2019). A total of $n$ adversarial examples are generated for each positive example, and used to train discriminative embeddings using noise contrastive estimation (Gutmann and Hyvärinen, 2010). Thus the knowledge graph embedding loss $L_{KGE}$ includes the negative log probability that each observed edge is true ($\ln \sigma(\phi(f_i))$), and the expected log probability that the adversarial edges are false ($\ln \sigma(-\phi(f'_i)) = \ln \left(1 - \sigma(\phi(f'_i))\right)$):

$$L_{KGE} = -\sum_{i=1}^{\left|G\right|} \ln \sigma(\phi(f_i)) + E_i \ln \sigma(-\phi(f'_i)))$$

(3)

where expectation is with respect to the probability $p_i(f'_j)$, tuned using a temperature hyperparameter $\alpha$ as

$$p_i(f'_j) = \frac{\exp(\alpha \phi(f'_j))}{\sum_{k=1}^{n} \exp(\alpha \phi(f'_k))}.$$  

(4)

Eq. (3) is used to train embeddings of the head ($h$) and tail ($t$), which are applied to the FVQA task as described in the next several subsections. Eq. (3) also trains relation embeddings ($r$) and MLP weights for the ERMLP scoring function ($w_{MLP}$); these quantities are not used for the downstream FVQA task.

### 3.2 Image representation

(Narasimhan et al., 2018) found that providing raw feature maps from a standard convolutional network like ResNet (He et al., 2016) or VGG (Simonyan and Zisserman, 2015) actually hurt performance on FVQA. We therefore represent each image using visual concepts found in the image.

**Objects:** We use Torchvision’s Coco object-detector, a Faster RCNN detector (Ren et al., 2015) with ResNet50 backbone (He et al., 2016), and feature pyramid network (Lin et al., 2017), which detects 80 object classes. We also use a detector (ZFTurbo, 2018) trained on an OpenImages 600-class detection task. We include classes present in ImageNet 200 plus those in (Wu et al., 2016) to maximize overlap with the dataset used with FVQA.

**Scenes:** We use a wideresnet (Zagoruyko and Komodakis, 2016) detector trained on MIT365 places dataset (Zhou et al., 2017) and consider the 205 classes that were used in constructing the FVQA KG.

Overall, we detect 540 visual concepts.

Having detected visual concepts in each image, we represent the $i^{th}$ image as a collection of entities, $I_i = [e_i^1, \ldots, e_i^m] \in \mathbb{R}^{N_e \times m}$, where $N_e$ is the embedding dimension, and $m$ is the number of visual concepts detected in the image. We detect a maximum of $m = 14$ visual concepts in each image. Our findings imply that for FVQA reasoning, an image is best represented as a collection of KG entity embedding vectors $e_i^t$, apparently because these KG embeddings encode the graph structure and background information necessary to be able to answer questions. We think this is an important finding, and we show its effect on the answer prediction accuracy.

#### 3.2.1 Language representation

For representing the question words, we use the last layer of contextual ALBERT (Lan et al., 2020) embeddings without finetuning for FVQA training. After passing through an LSTM, we use the hidden state representation for each word $w_q$ for further processing as $q_i^t = h(w_q)$ as the question representation.
Table 1: Summary of all trainable parameters.  

| Parameters | Used for | Loss Function |
|------------|----------|---------------|
| $h, r, t, w_{MLP}$ | KG embeddings | $\mathcal{L}_{KGE}$ |
| $w_{\alpha q}, w_{\alpha t}, w_{KVC}, w_{KB}, \theta_{LSTM}$ | Answer-retrieval based on image and question | $\mathcal{L}_{FVQA}$ |
| $w_q, \theta_{LSTM}$ | Choose answer source $\in \{f_{KVC}, f_{KB}\}$ | Cross entropy |

Table 2: KG Embedding accuracy for ERMLP (see Appendix for TransE and RotatE).

| Method | Sampling | MR | MRR | Hits@1 | Hits@3 | Hits@10 |
|--------|----------|----|-----|--------|--------|---------|
| ERMLP  | adversarial | 11194 | 0.156 | 0.132 | 0.152 | 0.197 |
| ERMLP  | uniform | 14907 | 0.122 | 0.09 | 0.128 | 0.173 |

| Method (QA facts missing) | Sampling | MR | MRR | Hits@1 | Hits@3 | Hits@10 |
|---------------------------|----------|----|-----|--------|--------|---------|
| ERMLP (QA facts missing)  | adversarial | 11071 ± 479 | 0.162 ± 0.007 | 0.136 ± 0.003 | 0.162 ± 0.005 | 0.204 ± 0.007 |
| ERMLP (50% KG)            | adversarial | 12880 | 0.144 | 0.1 | 0.13 | 0.161 |

Incomplete KG results

3.3 Fusion Functions $f_{KVC}$ and $f_{KB}$

Image and query are fused using a CoAttention mechanism, as depicted in Fig. 3. First, we calculate attention weights for each word in the question, and subsequently a self-attention-weighted encoding for the question as:

$$A(q_i) = \sum_{t=1}^{|q_i|} \alpha_{q_i}^t q_i^t, \quad \alpha_{q_i}^t = \frac{\exp(w_{\alpha q}^T q_i^t)}{\sum_{t=1}^{|q_i|} \exp(w_{\alpha q}^T q_i^t)} \quad (5)$$

where $\alpha_{q_i}^t$, $w_{\alpha q}$ are respectively the attention paid to word $q_i$, and the weight vector from which it is computed.

Then, using $A(q_i)$ as a query, we compute an attention-weighted summary of the image:

$$A(I_i) = \sum_{j=1}^{m} \alpha_{I_i}^j e_j^i, \quad \alpha_{I_i}^j = \frac{\exp(w_{\alpha I}^T e_j^i)}{\sum_{k=1}^{m} \exp(w_{\alpha I}^T e_k^i)} \quad (6)$$

where $\alpha_{I_i}^j$, $w_{\alpha I}$, $e_j^i$ are respectively the attention paid to concept $j$ in the image, the weight vector from which it is computed, and the $j$th concept present in the image. $A(I_i)$ is the attention-weighted representation of the image. Both $A(I_i)$ and $A(q_i)$ learn a mapping: $R^{N_e \times m} \rightarrow R^{N_e}$.

These attention-weighted image and question encodings compute joint image-question encodings via late fusion, as:

$$f_{KVC}(q_i, I_i) = h(A(I_i), A(q_i); w_{KVC}) \quad (7)$$

$$f_{KB}(q_i, I_i) = h(A(I_i), A(q_i); w_{KB}) \quad (8)$$

where $h(\cdot)$ is a two-layer fully-connected network with ReLU activation functions. Using $f_{KVC}(q_i, I_i) = A(I_i)$, i.e., the attention weighted image encodings to directly retrieve the answer significantly reduces accuracy, suggesting the need for successive fully connected layers to add capacity.

3.4 Gating function $g_{KVC}$

The fusion functions $f_{KVC}$ and $f_{KB}$ are trained to retrieve two different entities from the KG, either of which might be the answer to the question. The gating function, $g_{KVC}$, selects one of these two. The gating function is a sigmoid fully-connected layer applied to the final output state of an LSTM, whose input is the query, and which is trained using binary cross entropy so that $g_{KVC} = 1$ if the correct answer is a key visual concept in the image.

3.5 Loss function

A summary of all parameters is provided in Table 1. Entity embeddings are learned in order to minimize the loss function in Eq. (3), then all parameters are jointly trained in order to minimize the cosine distance between the ground truth entity, $y_i$, and the network output:

$$\mathcal{L}_{FVQA} = \frac{1}{n} \sum_{i=1}^{n} (1 - y_i^T \hat{y}(q_i|I_i)) \quad (9)$$

where $\hat{y}(q_i|I_i)$ is as given in Eq. (2).

4 Text-augmented composite score

4.1 Text-augmented score metric

To enhance SiK’s answer retrieval, we define additionally a text-augmented metric to leverage lexical similarity between the facts and the question in the
5 Experimental Setup

5.1 KG Embeddings training

For training of each KG embedding algorithm, we split all the facts in the KG as 80% train, 20% test. Entity and relation embeddings have a dimension of $N_e = 300$. Batch size is 1000. Each network is trained for 25000 steps using the Adam optimizer with learning rate initialized as 0.01 shrunk every 10000 steps by 0.1. Learning rates for all experiments were decided by choosing from the set {0.1, 0.01, 0.001, 0.0001}. We set $\alpha = 1$ for all experiments. We use $n = 16$ negative samples for each positive sample. ERMLP uses a three layer neural network with 3 $N_e$ (taking in head, rel, and tail embeddings), 2 $N_e = 600$ and $N_e = 300$ units respectively, and the output contains one sigmoid unit (the score $\phi(h, r, t)$). Hidden activations are ReLU. No dropout was used.

KGE accuracy is measured using standard metrics: Hits @1, Hits @3, Hits @10 determine how often each correct tail/head gets ranked in the top 1, 3, or 10 ranked facts for each ground-truth $(h, r, t)$ pair. Mean Rank measures the mean rank of each true fact in the dataset, Mean Reciprocal Rank $= \frac{1}{\text{mean rank}}$. While we report these metrics, we do note that these metrics are not best suited for common-sense KGs for reasons mentioned before. However we leave a more refined analysis of these methods for future work.

5.2 FVQA training

We report Hits @1 and Hits @3 for each algorithm. All numbers are based on averaging results across five train - test splits provided with the dataset. The number of questions varies slightly, but roughly fall half into both training and testing sets for each split. Stochastic gradient descent with a batch size of 64 trains $f_{KVC}$ and $f_{KB}$ for 250 steps with a learning rate of 0.01, reduced by 0.1 every 100 epochs, and a weight decay of 1e-3. Fully-connected layers use a dropout probability of 0.3. The gating function $g_{KVC}$ is trained for 20 steps with step-size of 0.1.

GPUs provided by Google Colab are used to train all models. Our heaviest KG embedding technique (ERMLP) takes roughly 3 hours to train, while one train split for $f_{KVC}$ takes roughly 30 minutes. Subgraph construction for each question takes about 1s, and 250 epochs to train our implementation of OOB took roughly 3 hours.
5.3 Baselines

Results in the standard QA task for the models FVQA, STTF, and OOB are taken from the respective papers since the codes for these systems have not been made available. HiCoAtt denotes using a Hierarchical CoAttention network (Lu et al., 2016) as implemented in (Wang et al., 2016). AvgEmbed denotes a method which compares STTF’s and OOB’s fact representation methodology with our architecture to correctly perform missing-edge reasoning. We use averaged 300-dimensional GLoVE embeddings (Pennington et al., 2014) to represent each entity. We also report results from our OOB implementation in the incomplete KG setting.

6 Results and Discussion

The gating function achieves an accuracy of 96%, same as reported in (Narasimhan and Schwing, 2018). We discuss below the performance of other modules of our architecture.

6.1 $F_{100}$ fact recall and impact on FVQA

We only observe 63% fact recall within $F_{100}$ across the entire dataset as opposed to 84.8% as reported by (Narasimhan et al., 2018). This further goes down to 53% if filtering for top-3 relationships. This difference is significant, and FVQA accuracy reported by OOB at such levels of fact recall are lower. One factor for lower recall is noisy visual detections, whose keywords impact fact retrieval. Therefore, with access to similar visual detections as the other methods, SiK could potentially achieve even higher accuracy. Furthermore, owing to the lower fact retrieval recall, our OOB implementation could not replicate the reported performance in the standard QA task. This also highlights the crucial reliance of OOB on correct fact retrieval and further substantiates our observation of its low performance when the KG is incomplete.

6.2 Incomplete KGs

The robustness of SiK is demonstrated in the incomplete KG setting. Here we discuss two of them - one where only the QA-related facts are missing and another where 50% of the KG is missing (more settings in Appendix). In both cases, we see standalone SiK and text-augmented SiK only lose some of their accuracy. Our implementation of OOB, and AvgEmbed both underperform SiK by over 25%. Setting $\lambda_1 = 0$ in the score also leads to similar

---

| Technique  | Hits@1  | Hits@3  |
|------------|---------|---------|
| FVQA top-1 qq-mapping | 52.56 ± 1.03 | 59.72 ± 0.82 |
| STTF       | 62.2 ± NR | 75.6 ± NR |
| OOB top-1 rel | 65.8 ± NR | 77.32 ± NR |
| OOB top-3 rel | 69.35 ± NR | 80.25 ± NR |

### Incomplete KG

| Method | Only QA facts missing | 50% KG missing |
|--------|----------------------|----------------|
| AvgEmbed, IaK | 27.77 ± 0.72 | 32.43 ± 0.94 |
| OOB top-3 rel | 28.58 ± 0.01 | 43.56 ± 0.01 |
| ERLMP - Adv, IaK | 53.45 ± 0.77 | 65.1 ± 1.41 |
| ERLMP - Adv, IaK (0.3, 0.03, 0.33) | 55.13 ± 0.97 | 73.04 ± 0.54 |
| ERLMP - Adv, IaK (0.3, 0.03, 0.33) | 26.8 ± 0.90 | 49.24 ± 0.62 |

Table 3: FVQA accuracy (IaK - Image as Knowledge, Adv - self-adversarial negative sampling, NR - Not reported)
Question 1: What is the difference between the instrument & the violin?
**SPO triple:** \{Cello, HasProperty, like a violin but larger\}
**Answer Source:** KG
**Answer:** like a violin but larger
**Answer predicted:** like a violin but larger

Question 2: Which object in this image belongs to the category Herbivorous animals?
**SPO triple:** \{Giraffe, Category, Herbivorous animals\}
**Answer Source:** Image
**Answer:** Giraffe
**Answer predicted:** Giraffe

Question 3: What popular pet is in this image?
**SPO triple:** \{Goldfish, ISA, popular pet\}
**Answer Source:** Image
**Answer:** Goldfish
**Answer predicted:** Fish

Question 4: What object in this image is commonly eaten for lunch?
**SPO triple:** \{Sandwich, ISA, meal commonly eaten for lunch\}
**Answer Source:** Image
**Answer:** Sandwich
**Answer predicted:** Bread

Question 5: Where can you find the large object in the back of this image?
**SPO triple:** \{Bus, AtLocation, Bus stop\}
**Answer Source:** KG
**Answer:** Bus stop
**Answer predicted:** Wait place

Question 6: What can we find in the place shown in this image?
**SPO triple:** \{Furniture, AtLocation, Bedroom\}
**Answer Source:** KG
**Answer:** Furniture
**Answer predicted:** Your House

Figure 4: Success and Failure cases of Seeing is Knowing

performance, highlighting the importance of KG features for FVQA. For a more detailed discussion of the missing edges experiments, the interested reader is directed to the Appendix.

### 6.3 Composite metric performance

When the required edge is present in the graph, standalone SiK underperforms OOB by 11%. This is enhanced using the composite score, which sees a 6% point improvement. Quite remarkably, text-augmented SiK works best when roughly equal coefficients are used for all three similarities. However, the strongest similarity metric is still that obtained from SiK (see Table 3 and Appendix).

### 6.4 Image as Knowledge

To understand how useful the Image-as-Knowledge representation is compared to other variants, we compare it to the multihot variant proposed by (Narasimhan et al., 2018) along with our best-performing KG embedding technique. ‘Image-as-knowledge’ provides a 3-point performance improvement (Table 3), apparently because the retrieval of an entity happens in the entity space spanned by the vectors of the IaK representation.

### 6.5 Success and Failure cases

To investigate the only learning module in the score, we look at the failure modes for the standalone SiK
architecture. Since the gating function performs with almost perfect accuracy, our main source of error is incorrect entity prediction. Since a prediction is considered correct only if there’s an exact match in the entities, a lot of inaccuracies stem from cases where the network predicts a semantically valid but different entity. Entities sharing parent-child relationships (Question 3, Fig. 1) or similar attributes through multi-hop relationships (Questions 4-6) are commonly mistaken by the SiK network. We discuss more failure cases in the appendix.

7 Conclusion

An architecture is proposed that solves the FVQA task even when the required edge is missing from the knowledge graph. In the process, we present the first approach to use KG embeddings for FVQA. A composite answer retrieval method augments its accuracy by incorporating complementary lexical features and KG semantic features. Serendipitous benefits of the standalone approach include: (1) improved computational complexity, and (2) an improved representation of each image, as the span of KG embeddings of the visible entities.

8 Ethical impact

We now examine the ethical implications of our work. A prominent issue could be that of different biases known to exist in our data sources. (Shankar et al., 2017) show population / representation bias existing in OpenImages and ImageNet. (Fisher et al., 2019) showed web-scale commonsense KGs can be tough to curate and can allow biases to creep in. (Janowicz et al., 2018) note how the density of world locations generating DBPedia data (extracted from Wikipedia) is at odds with world population density. Fig. 5 shows how this manifests in our system. The answer for a place relevant to giraffes in wildlife, comes up as Zimbabwe, even though the only edges present in the KG concerning Zimbabwe were {Person, AtLocation, in Zimbabwe}, {Dog, AtLocation, in Zimbabwe}, {Tree, AtLocation, in Zimbabwe}, {Elephant, AtLocation, in Zimbabwe}. Such an error is quite informative and humbling – this means that different modalities (image, language, graph) working in tandem could still amplify their individual biases. For deploying such a system, we recommend debiasing of parameters learned in our architecture. While ConceptNet has been active in debiasing their representations, debiasing KG embeddings has not received as much attention, and it could pose subtle problems given that most entities would have low node-degree. A multimodal approach to debias representations could be an interesting research direction.

References

S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. 2015. Vqa: Visual question answering. In 2015 IEEE International Conference on Computer Vision (ICCV), pages 2425–2433.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Zachary Ives, and et al. 2007. Dbpedia: A nucleus for a web of open data. In PROC. 6TH INT’L SEMANTIC WEB CONF. Springer.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In SIGMOD Conference, pages 1247–1250.

Antoine Bordes, Nicolas Usunier, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In In Advances in Neural Information Processing Systems 26. Curran Associates, Inc, pages 2787–2795.

T. Dettmers, P. Minervini, P. Stenetorp, and S. Riedel. 2018. Convolutional 2D knowledge graph embeddings. In Proc. AAAI, pages 1811–1818.

Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Patrick Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. 2014. Knowledge vault: a web-scale approach to probabilistic knowledge fusion. In Knowledge Distillation and Data Mining, pages 601–610.
Joseph Fisher, Dave Palfrey, Christos Christodoulopoulos, and Arpit Mittal. 2019. Measuring social bias in knowledge graph embeddings.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Michael Gutmann and Aapo Hyvärinen. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. Proceedings of Machine Learning Research, pages 297–304, Chia Laguna Resort, Sardinia, Italy.

Graeme S. Halford, William H. Wilson, and Steven Phillips. 2010. Review relational knowledge: the foundation of higher cognition.

Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 770–778.

Krzysztof Janowicz, Bo Yan, Blake Regalia, Rui Zhu, and Gengchen Mai. 2018. Debiasing knowledge graphs: Why female presidents are not like female popes. In Proceedings of the ISWC 2018 Posters & Demonstrations.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In International Conference on Learning Representations (ICLR).

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. In International Conference on Learning Representations.

Hui Li, Peng Wang, Chunhua Shen, and A. V. D. Hengel. 2019. Visual question answering as reading comprehension. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 6312–6321.

T-Y Lin, M Maire, S Belongie, J Hays, P Perona, D Ramanan, P Dollár, and CL Zitnick. 2014. Microsoft COCO: Common objects in context. Lecture Notes in Computer Science, 8693:740–755.

Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. 2017. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Hugo Liu and Push Singh. 2004. Conceptnet: A practical commonsense reasoning toolkit. BT TECHNOLOGY JOURNAL, 22:211–226.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. In Advances in Neural Information Processing Systems 29, pages 289–297.

T. Mitchell and E. Fredkin. 2014. Never-ending language learning. In 2014 IEEE International Conference on Big Data (Big Data), pages 1–1.

Medhini Narasimhan, Svetlana Lazebnik, and Alexander Schwing. 2018. Out of the box: Reasoning with graph convolution nets for factual visual question answering. In Advances in Neural Information Processing Systems 31, pages 2654–2665.

Medhini Narasimhan and Alexander G. Schwing. 2018. Straight to the facts: Learning knowledge base retrieval for factual visual question answering. CoRR.

M. Nickel, V. Tresp, and H.P. Kriegel. 2011. A three-way model for collective learning on multi-relational data. In Proc. International Conference on Machine Learning.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1532–1543.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In Advances in Neural Information Processing Systems 28, pages 91–99.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV).

Apoorv Saxena, Adity Tripathi, and Partha Talukdar. 2020. Improving multi-hop question answering over knowledge graphs using knowledge base embeddings. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4498–4507, Online. Association for Computational Linguistics.

Shreya Shankar, Yoni Halpern, Eric Breck, James Artwood, Jimbo Wilson, and D. Sculley. 2017. No classification without representation: Assessing geodiversity issues in open data sets for the developing world. In NIPS 2017 workshop: Machine Learning for the Developing World.

K Simonyan and A Zisserman. 2015. Very deep convolutional networks for large-scale image classification. In Proc. International Conference on Learning Representations.
Quite remarkably however, there is only a very slight decrease in SiK’s performance when using any of these KG embeddings for the downstream FVQA task (see Table 5). Only at 50% KG occlusion, we see a statistically significant yet still small drop in performance. This highlights the robustness of using KG embeddings for downstream semantic tasks.

B. Characterizing inaccuracies of Seeing is Knowing Network: Qualitative discussion

In order to analyze the types of errors SiK is prone to make, we now discuss some more failure cases. We observe that there are three broad categories of errors -

1. Semantically correct answer, but answer is a different entity
2. Incorrect answer, explainable relevance to question and image
3. Entirely incorrect answers

An empirical study of these errors shows a significant fraction of them fall in the first two categories. Errors in the first category can arise when the model has implicitly pinned down the correct key visual entity and relationship being asked about, but then chose a different entity (see Question 4-6 Fig.6). A bed is an object one can lose things in, chairs come in many styles, and a flute can be used for practice. These errors can also arise if the model has chosen to place visual emphasis on an entity that is different from ground truth, but is equally valid given a question (see Question 7 Fig.6). Here, the model has retrieved the answer ‘build a treehouse’ that appears in a fact with ‘tree’, but the ground truth of the question asks about the entity ‘lake’.

Errors of the second category can arise in mainly two cases - incorrect language and / or visual grounding. In cases where the visual grounding is off, it could be due to false positives in detections (see Questions 1-3 Fig.6). Giraffes, tomatoes, and foxes are mistaken to be camels, bell peppers, and lions respectively; entities relevant to the mistaken entities are then chosen as answers.

In cases where the language grounding is off, one case could be that the model has predicted the correct symbolism for a question (i.e. correct entity and relationship), but the retrieved entity comes...
**Question 1:** What is the animal famous for?

**SPO triple:** \{Giraffe, HasProperty, Long necked\}

*Answer Source: KG*

*Answer:* long necked

*Answer predicted:* long necked

**Question 2:** Which object in this image is used in spaghetti sauce?

**SPO triple:** \{Sphagetti sauce, RelatedTo, Tomato\}

*Answer Source: Image*

*Answer:* Tomato

*Answer predicted:* Bell Pepper

**Question 3:** What is the family of the animal in this image?

**SPO triple:** \{Fox, Category, Canidae\}

*Answer Source: KG*

*Answer:* Canidae

*Answer predicted:* Burundian Culture

**Question 4:** What is the wooden thing used for?

**SPO triple:** \{Flute, UsedFor, making music\}

*Answer Source: KG*

*Answer:* making music

*Answer predicted:* Practice

**Question 5:** What is the material of the chair?

**SPO triple:** \{Chair, RelatedTo, Wooden\}

*Answer Source: KG*

*Answer:* Wooden

*Answer predicted:* come in many style

**Question 6:** What is the object in the left of the image used for?

**SPO triple:** \{Bed, UsedFor, Laying on\}

*Answer Source: KG*

*Answer:* Laying on

*Answer predicted:* Loose (sic) thing in

**Question 7:** What can we do in this place shown in this image?

**SPO triple:** \{Lake, UsedFor, row a boat\}

*Answer Source: KG*

*Answer:* row a boat

*Answer predicted:* build a tree-house

**Question 8:** This game is most popular in which country?

**SPO triple:** \{Table Tennis, HasProperty, Popular in China\}

*Answer Source: KG*

*Answer:* Table Tennis

*Answer predicted:* Game Person Play

**Question 9:** What is the baseball bat looks like?

**SPO triple:** \{Baseball, IsA, Long round tapered object\}

*Answer Source: KG*

*Answer:* long round tapered object

*Answer predicted:* aggressive animal

---

*Figure 6: Performance of standalone Seeing is Knowing network without the composite score-based retrieval*
| Method | KG Occlusion | MR   | MRR | Hits@1 | Hits@3 | Hits@10 |
|--------|--------------|------|-----|--------|--------|---------|
| ERMLP  | Only QA facts | 11071| 0.162| 0.136  | 0.162  | 0.204   |
| ERMLP  | 10% of KG     | 9690 | 0.175| 0.148  | 0.1779 | 0.22    |
| ERMLP  | 20% of KG     | 10733| 0.1613| 0.133  | 0.162  | 0.213   |
| ERMLP  | 30% of KG     | 10437| 0.154| 0.127  | 0.156  | 0.199   |
| ERMLP  | 40% of KG     | 11467| 0.139| 0.116  | 0.138  | 0.179   |
| ERMLP  | 50% of KG     | 12880| 0.144| 0.1  | 0.13  | 0.161   |

| Method | KG Occlusion | Hits @ 1 | Hits @ 3 |
|--------|--------------|----------|----------|
| ERMLP  | QA facts     | 53.45 ± 0.77 | 65.1 ± 1.41 |
| ERMLP  | 10% of KG    | 53.19 ± 0.47 | 65.67 ± 0.31 |
| ERMLP  | 20% of KG    | 53.62 ± 1.01 | 64.72 ± 0.98 |
| ERMLP  | 30% of KG    | 53.15 ± 1.62 | 64.55 ± 1.23 |
| ERMLP  | 40% of KG    | 53.14 ± 0.47 | 64.77 ± 0.55 |
| ERMLP  | 50% of KG    | 52.29 ± 0.86 | 62.74 ± 0.69 |

Table 4: KG Embedding performance of ERMLP when varying levels of the KG is occluded.

Table 5: FVQA accuracy of Seeing is Knowing using ERMLP when varying levels of the KG is occluded.

from a fact that is not grounded in the language of the question. In Question 8 Fig. 6, ‘game person play’, a property about the game table tennis is chosen, but it is not grounded in the question-text which asks about the country where it is popular. A fix for this would be to ensure that entities’ representations contain both lexical and graph-based features. One approach could be to use contextual word embeddings to encode entities before doing graph learning. Another approach involving RotatE, could be to fix real part of the entities as lexical representations during training, so that the imaginary parts could be enforced to encode graph-based features for fixed lexical semantic features.

Lastly, entirely incorrect answers, such as Question 9 Fig.6 can occur in case of sparse node connections (only one edge is present for ‘aggressive animal’). SiK is indeed prone to making such unexplainable errors, highlighting the scope for improvement to be gained from incorporating better lexical semantic features as used in our composite score function.

### C. Attention maps

To investigate the performance of the Image-as-Knowledge module, Fig.7 depicts both the visual and textual attention maps learnt by the network. We can see that the CoAttention module is learning to pay attention to contextually relevant cues based on the input word embeddings in the question and the KG embeddings of the visual concepts. For Question 1 the textual attention correctly attends to the word ‘amber and green’ and the image attention correctly attends to the entity ‘traffic light’. In Question 2 similarly, the textual attention attends to the word ‘jazz club’ while the image attention attends to the entity ‘trumpet’. Likewise in Question 3, words ‘centre of the image’ and entity ‘runway’ are correctly attended to.

In cases where the answer is incorrectly predicted, the learnt attention map throws light on the reasoning of the network.

### D. Choosing hyperparameters for composite score-based retrieval

To investigate the importance of each metric in the composite score, we vary the respective coefficient $\lambda_k$ to each metric across a range of values.

The first three rows of Table 6 demonstrate that the KG similarity metric individually is strongest by some margin. Considering only Jaccard similarity or GLoVe similarity is not enough to produce the accurate answer. This further highlights the crucial involvement that a KG reasoning module has in performance of any FVQA system. A standalone KG similarity metric, however, falls short of the highest achieved accuracy, owing to its relative inability to leverage lexical similarity between entity words and words in the question. (Note that the system obtained by assigning all the weight to
only SiK metric is not the same as directly using SiK to retrieve the answer, since the former still uses the subgraph construction method to prune the fact search space. This is also demonstrated in the slightly different accuracy value of the two methods.)

The results in the next section in the table demonstrate that lexical and distributional semantic features indeed hold complementary information to that of KG embeddings. To remedy the trend of SiK ignoring lexical similarity, roughly equal importance is assigned to the Jaccard and GLoVE similarity metrics. Interestingly, the convex combination works best when each distance metric is weighted roughly equally. Our experiments indicate that, the majority of the correct answers are produced due to the similarity score arising from SiK, but an equal weighting is required so as to nudge the correct answer ahead of the incorrect ones.

Lastly, we observe that weighting the SiK metric more than 0.6 starts to yield lesser and lesser benefits, and the performance monotonically decreases.

In both the incomplete KG settings (Table 7), we observe that pruning the fact search space still improves accuracy over the standalone networks. But the contribution of the textual similarity metrics becomes almost negligible compared to its impact when required facts are present in the network.

E. Performance of all KG Embedding algorithms

ERMLP has better mean-rank, while RotatE performs better in all other standard metrics for KG embeddings. (Table 8). However, ERMLP-based embeddings are more successful for FVQA (Table 9). Apparently ERMLP adds more capacity to the embedding model, which also helps the downstream task of FVQA.

Using self-adversarial negative sampling is beneficial to all KG Embedding algorithms. The FVQA accuracy improves by 1.67%, 2.47 %, and 1.25% respectively for TransE, RotatE, and ERMLP.
### Table 6: FVQA performance for different convex combinations of $\lambda_k$ in the standard answer prediction task. KG Embedding algorithm used is ERMLP and image representation is Image-as-Knowledge. $\lambda_1$ is the coefficient of KG similarity, $\lambda_2$ is the coefficient of Jaccard similarity and $\lambda_3$ is the coefficient for GLoVE similarity

| Coefficient values | Top 100 facts | Top 500 facts |
|--------------------|---------------|---------------|
|                    | $\lambda_1$   | $\lambda_2$   | $\lambda_3$   | Hits @1         | Hits @3         | Hits @1         | Hits @3         |
| Single similarity metric being used |               |               |               |                 |                 |                 |
| 1                  | 0             | 0             | 0             | 56.05 ± 0.57   | 71.84 ± 0.44   | 55.43 ± 0.57   | 69.55 ± 0.58   |
| 0                  | 1             | 0             | 0             | 17.44 ± 0.55   | 42.98 ± 0.86   | 13.83 ± 0.57   | 34.64 ± 1.00   |
| 0                  | 0             | 1             | 0             | 13.38 ± 0.75   | 30.67 ± 0.80   | 11.75 ± 0.62   | 28.82 ± 0.75   |
| Equal importance to all similarity metrics |               |               |               |                 |                 |                 |
| 0.34               | 0.33          | 0.33          |               | 60.73 ± 0.82   | 77.03 ± 0.46   | 60.63 ± 0.73   | 76.9 ± 0.51    |
| 0.4                | 0.3           | 0.3           |               | 60.53 ± 0.94   | 77.14 ± 0.50   | 60.33 ± 0.92   | 77.01 ± 0.55   |
| 0.5                | 0.5           | 0             |               | **60.82 ± 0.98** | 75.83 ± 0.42   | 59.96 ± 0.89   | 74.48 ± 0.47   |
| 0.5                | 0.25          | 0.25          |               | 59.96 ± 0.72   | 76.73 ± 0.37   | 59.57 ± 0.68   | 76.31 ± 0.57   |
| 0.5                | 0             | 0.5           |               | 54.59 ± 71.76  | 71.76 ± 0.46   | 52.52 ± 0.74   | 71.03 ± 0.68   |
| 0                  | 0.5           | 0.5           |               | 40.18 ± 1.13   | 61.62 ± 0.47   | 39.76 ± 1.08   | 60.84 ± 0.67   |
| Highest importance to KG similarity |               |               |               |                 |                 |                 |
| 0.6                | 0.2           | 0.2           |               | 58.89 ± 0.54   | 75.89 ± 0.44   | 58.5 ± 0.51    | 75.1 ± 0.658   |
| 0.7                | 0.15          | 0.15          |               | 58.05 ± 0.64   | 74.87 ± 0.49   | 57.6 ± 0.5     | 73.84 ± 0.49   |
| 0.8                | 0.1           | 0.1           |               | 57.23 ± 0.54   | 73.9 ± 0.56    | 56.73 ± 0.49   | 72.41 ± 0.40   |

### Table 7: FVQA performance for different convex combinations of $\lambda_k$ in the incomplete KG FVQA task for top-100 facts retrieved. KG Embedding algorithm used is ERMLP and image representation is Image-as-Knowledge. $\lambda_1$ is the coefficient of KG similarity, $\lambda_2$ is the coefficient of Jaccard similarity and $\lambda_3$ is the coefficient for GLoVE similarity

| Coefficient values | Only QA facts missing | 50% KG missing |
|--------------------|-----------------------|----------------|
|                    | $\lambda_1$   | $\lambda_2$   | $\lambda_3$   | Hits @1         | Hits @3         | Hits @1         | Hits @3         |
| Single similarity metric being used |               |               |               |                 |                 |                 |
| 1                  | 0             | 0             | 0             | 55.23 ± 0.44   | 70.56 ± 0.41   | 55.20 ± 0.77   | 70.47 ± 0.36   |
| 0                  | 1             | 0             | 0             | 9.91 ± 0.58    | 27.44 ± 1.09   | 9.92 ± 0.61    | 27.48 ± 1.06   |
| 0                  | 0             | 1             | 0             | 12.42 ± 0.72   | 29.43 ± 0.72   | 12.44 ± 0.76   | 29.47 ± 0.72   |
| Equal importance to all similarity metrics |               |               |               |                 |                 |                 |
| 0.34               | 0.33          | 0.33          |               | 55.13 ± 0.98   | 73.04 ± 0.54   | 55.13 ± 1.09   | 72.95 ± 0.38   |
| 0.4                | 0.3           | 0.3           |               | 56.35 ± 1.08   | 73.56 ± 0.58   | 56.38 ± 1.08   | 73.47 ± 0.44   |
| 0.5                | 0.5           | 0             |               | 54.78 ± 0.78   | 70.53 ± 0.56   | 54.79 ± 0.98   | 70.48 ± 0.33   |
| 0.5                | 0.25          | 0.25          |               | **56.80 ± 0.82** | **73.62 ± 0.42** | **56.81 ± 0.84** | **73.54 ± 0.25** |
| 0.5                | 0             | 0.5           |               | 53.68 ± 0.82   | 70.77 ± 0.67   | 53.63 ± 0.87   | 70.67 ± 0.35   |
| 0                  | 0.5           | 0.5           |               | 26.8 ± 0.90    | 49.24 ± 0.62   | 26.79 ± 0.85   | 49.21 ± 0.63   |
| Highest importance to KG similarity |               |               |               |                 |                 |                 |
| 0.6                | 0.2           | 0.2           |               | 56.46 ± 0.51   | 73.36 ± 0.51   | 56.45 ± 0.63   | 73.31 ± 0.22   |
| 0.7                | 0.15          | 0.15          |               | 56.23 ± 0.36   | 72.7 ± 0.39    | 56.17 ± 0.66   | 72.63 ± 0.34   |
| 0.8                | 0.1           | 0.1           |               | 55.88 ± 0.25   | 72.01 ± 0.46   | 55.85 ± 0.65   | 71.91 ± 0.31   |
Question 1: Which object in this image is colored red, amber and green?
SPO triple: (Traffic Light, HasProperty, colour red amber and green)
Answer Source: Image
Answer: Traffic light
Answer predicted: Traffic light

Question 2: Which object can be found in a jazz club?
SPO triple: (Trumpet, AtLocation, jazz club)
Answer Source: Image
Answer: Trumpet
Answer predicted: Trumpet

Question 3: Where can object in the center of the image be found?
SPO triple: (Airplane, AtLocation, airport)
Answer Source: KG
Answer: Airport
Answer predicted: Runway

Figure 7: Visualizing CoAttention weights learnt by Seeing is Knowing
| Method   | Sampling  | MR   | MRR   | Hits@1 | Hits@3 | Hits@10 |
|----------|-----------|------|-------|--------|--------|---------|
| TransE   | adversarial | 14692 | 0.1035 | 0.05   | 0.13   | 0.19    |
| TransE   | uniform   | 13736 | 0.0967 | 0.04   | 0.112  | 0.194   |
| RotatE   | adversarial | 17660 | 0.173  | 0.14   | 0.18   | 0.22    |
| RotatE   | uniform   | 18468 | 0.144  | 0.122  | 0.152  | 0.179   |
| ERMLP    | adversarial | 11194 | 0.156  | 0.132  | 0.152  | 0.197   |
| ERMLP    | uniform   | 14907 | 0.122  | 0.09   | 0.128  | 0.173   |

Table 8: KG Embedding accuracy for standard retrieval task

| Technique                  | Hits@1       | Hits@3       |
|----------------------------|--------------|--------------|
| TransE, IaK                | 38.64 ± 1.51 | 52.58 ± 1.93 |
| TransE – Adv, IaK          | 40.37 ± 0.87 | 53.1 ± 1.12  |
| RotatE, IaK                | 45.59 ± 0.52 | 56.55 ± 1.18 |
| RotatE – Adv, IaK          | 47.92 ± 1.15 | 61.3 ± 0.97  |
| ERMLP, IaK                 | 53.16 ± 0.79 | 63.9 ± .63   |
| ERMLP - Adv, IaK           | **54.38 ± 0.94** | **65.76 ± 0.5** |

Table 9: FVQA accuracy for standard retrieval task (IaK - Image as Knowledge, Adv - self-adversarial negative sampling)