From Strings to Things: Knowledge-enabled VQA Model that can Read and Reason

Ajeet Kumar Singh[1] Anand Mishra[2,*] Shashank Shekhar[3] Anirban Chakraborty[3]

[1]TCS Research, Pune, India [2]IIT Jodhpur, India [3]Indian Institute of Science, Bangalore, India

Abstract

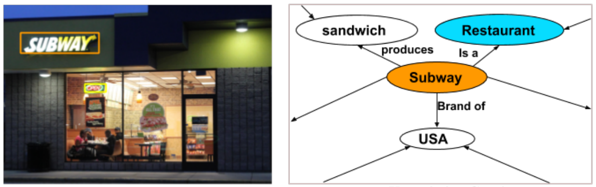
Text present in images are not merely strings, they provide useful cues about the image. Despite their utility in better image understanding, scene texts are not used in traditional visual question answering (VQA) models. In this work, we present a VQA model which can read scene texts and perform reasoning on a knowledge graph to arrive at an accurate answer. Our proposed model has three mutually interacting modules: (i) ***proposal module*** *to get word and visual content proposals from the image, (ii)* ***fusion module*** *to fuse these proposals, question and knowledge base to mine relevant facts, and represent these facts as multi-relational graph, (iii)* ***reasoning module*** *to perform a novel gated graph neural network based reasoning on this graph.*

The performance of our knowledge-enabled VQA model is evaluated on our newly introduced dataset, viz. text-KVQA. To the best of our knowledge, this is the first dataset which identifies the need for bridging text recognition with knowledge graph based reasoning. Through extensive experiments, we show that our proposed method outperforms traditional VQA as well as question-answering over knowledge base-based methods on text-KVQA.

1. Introduction

> "The more that you read, the more things you will know."
>
> *Dr. Seuss (in 'I Can Read With My Eyes Shut!')*

Texts appearing in images open the door to the world of knowledge and help gain a deeper and holistic understanding of the visual content. However, traditional visual question answering models do not leverage it. In this work, we introduce a novel task of knowledge-enabled visual question answering by reading text in images.



Image

Knowledge Graph

Word proposals [16]: Subway, open
Visual content proposals [55]: fast food restaurant, shop front

Q. Which restaurant is this?
A. Subway
Q. Can I get a sandwich here?
A. Yes
Q. Is this a French brand?
A. No

Figure 1. VQA model which only relies on visual cues may not be necessarily sufficient to answer many natural questions, for example, *Which restaurant is this?* for the given image. On the other hand, the text "subway" appearing on image and rich knowledge graph containing information like *Subway is a restaurant*, are indispensable cues for visual question answering. We present a VQA model that seamlessly integrates visual content (shown in cyan), recognized word (shown in orange), a question and knowledge facts to answer questions often asked in a real-world setting. **[Best viewed in color].**

Visual question answering (VQA) has emerged as an important problem spanning vision and language. Traditionally, VQA models [5, 20, 59] restrict themselves to analyze visual cues alone. It may not necessarily be sufficient, especially when the question asked demands deeper knowledge beyond the immediate visual content of the scene. For instance, consider an image shown in Figure 1, and a question *Which restaurant is this?*, visual cues are not enough to suggest the name of the restaurant. However, the fact that this

[*] Anand Mishra was associated with the Indian Institute of Science, Bangalore when this work was carried out.

image contains a word *Subway* and external knowledge that "Subway is a restaurant" provides an indispensable cue, and allows us to answer this question correctly. Further, with the access to rich open-source knowledge graphs such as Wikidata [44], we could ask a series of natural questions, such as, *Can I get a Sandwich here?*, *Is this a French brand?*, and so on, which are not possible to ask in traditional VQA [5] as well as knowledge-enabled VQA models [47, 48].

The need for development of VQA models that can read texts has been identified in a few very recent works [8, 32, 42]. However, the accompanying datasets are not backed up by rich world knowledge, and hence limited to questions which can be answered by visual and textual cues alone. Additionally, many questions in these datasets, such as, *What is the street name that starts with a color?*, *What is the word that comes after golden?* may pose computer vision challenges, but such questions are neither natural nor often asked in a real-world setting. This motivated us to come up with a novel task and accompanying dataset where access to world knowledge plays a crucial role in question answering, in addition to the ability to read scene texts. Our newly introduced dataset is much larger in scale as compared to the three aforementioned works [8, 32, 42] and more importantly, backed up by web-scale knowledge facts harvested from various sources, e.g., Wikidata [44], IMDb [1], a book catalogue [13]. Our dataset contains images of scene, movie posters, cover pages of books along with a series of natural and engaging questions, which may often be asked by people in real-world scenario. Our dataset named as text-KVQA, with associated knowledge bases can be downloaded from our project website: https://textkvqa.github.io/.

Our approach: Scene text recognition is graduating from research labs to academic demos as well as limited industrial applications ([31, 34, 45, 10, 49, 19, 7, 40, 9, 16]). However, only relying on scene text recognition method while developing a VQA model may not suffice. Hence, we propose to integrate multiple cues, i.e., visual contents, recognized words, question and knowledge facts, and perform a reasoning on a multi-relational graph using a novel gated graph neural network [27] formulation.

Contributions of this paper: (i) We draw attention to an important problem of visual question answering by reading text appearing in images, connecting it to knowledge graph and performing appropriate reasoning to arrive at an accurate answer. To this end, we introduce a large-scale dataset, namely text-KVQA. To the best of our knowledge, text-KVQA is the first dataset which identifies the need for bridging text recognition and knowledge graph based reasoning for VQA task. (ii) We present a VQA model which seamlessly integrates visual content, recognized words, question asked and knowledge facts, and performs reasoning on multi-relational graph using a novel GGNN formulation. (Section 4) (iii) Rigorous experiments and ablation studies are performed on the text-KVQA to validate effectiveness of our proposed approach. (Section 5)

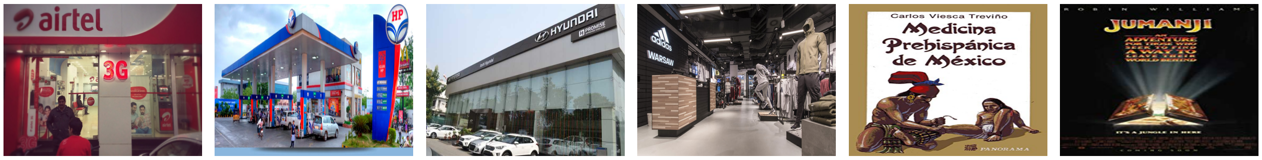
2. Related work

Visual Question Answering: VQA has gained huge interest in recent years. The traditional VQA methods can be grouped into following three broad categories: (i) joint embedding methods, (ii) attention mechanism, and (iii) compositional models. Learning image and language embeddings in a common space has been common practice in the vision and language communities. This has been leveraged in some of the earlier works in VQA such as [6, 11, 14, 15, 25, 29, 35, 41, 51, 52, 56, 60]. These methods typically use Bidirectional Long Short Term Memory and Convolutional Neural Networks for representing question and image respectively, and learn a joint model to predict the answer. More recently, VQA models also leveraged attention mechanism to improve further [24, 29, 11, 29, 36, 41, 51]. There has been growing interest in understanding compositional linguistic structure of the questions for VQA tasks. Methods, such as dynamic memory networks [26] and neural module networks [4] fall in this category. However, these methods are still mostly restricted to visual reasoning alone.

VQA over knowledge graph: Visual question answering over knowledge graphs is a recent trend in the VQA literature [48, 47, 37, 33, 46]. In these works, memory networks [50] and their variants have become the de facto baselines. However, as noted in [54], memory network treats knowledge graph as a flattened table of facts, making it hard to exploit the structural information in the graph and thus, is relatively weak on reasoning. To overcome this limitation, more recently graph representation learning has emerged as a natural choice to perform reasoning over large knowledge graphs [30, 53]. This motivated us to utilize the capability of graph representation learning in the form of gated graph neural networks (GGNN) [27]. Further, GGNN also allows us to integrate the visual and textual cues seamlessly.

Scene text localization and recognition: We have witnessed a significant boost in scene text localization and recognition performance over past years. Like many other areas in computer vision, deep neural nets have heavily influenced scene text localization research. Researchers have started approaching text localization along the lines of object localization tasks. Many works in text localization such as EAST [57], SSTD [17], TextBox++ [28] are influenced by seminal works in object localization. Once the text is localized, the next problem is to recognize the corresponding words. The modern methods [19, 7, 38, 16, 9] utilize the availability of large annotated datasets and deep CNN architecture to address this problem very effectively.

Combining visual and textual cues: Researchers have also shown interest in combining visual and textual cues,



(a) Q. Which mobile store is this? A. Airtel SF: Airtel is a telecommunication industry.

(b) Q. Can I fill petrol in my car here? A. Yes SF: HP is a petrolium industry.

(c) Q. Does this showroom sell car? A. Yes SF: Hyundai produces car.

(d) Q. Is this an American brand? A. No SF: Adidas is brand of Germany.

(e) Q. In which language this book is written? A. Spanish SF: Medicina Prehispanica De Mexico is written in Spanish.

(f) Q. Who is the director of this movie? A. Joe Johnston SF: Jumanji is directed by Joe Johnston

Figure 2. Sample images, question-ground truth answer pairs and a relevant supporting fact from our newly introduced text-KVQA dataset. Please note that supporting fact is not explicitly provided during training and inference of our method. Rather it is mined from the large-scale knowledge bases. Please refer to supplementary material for more examples.

e.g., improving scene text recognition using scene context [58], improving image classification using scene text [23], etc. Very recent works [8, 32, 42] highlight the need to combine visual and textual cues for visual question answering. However, despite early progress in knowledge-enabled VQA models and noticeable progress in the scene text recognition literature, the important and much needed task of combining these two research directions has not been explored so far. Our work aims to be the first attempt towards filling this gap.

3. Dataset

The traditional VQA models lack ability to read text in the images. Very recently, towards developing VQA models that can read, three datasets [8, 32, 42] have been introduced. However, these datasets do not allow knowledge-enabled questions to be asked. We identify the need for knowledge-enabled VQA model that can read and reason in knowledge, vision and text space. Towards this goal, a novel large-scale dataset namely, text-KVQA that contains 1.3 million question-answer pairs, 257K images and associated web-scale knowledge bases has been introduced in this work. We provide a table comparing text-KVQA with related datasets in the literature in the supplementary material.

The images of business brands, movie posters and book covers were collected as part of our dataset. Among these, movie posters and book cover images were obtained from [2] and [18] respectively. Further, we explicitly harvested scene images of business brands. To this end, we first prepare a list of 1000 business brands and use Google image search to obtain approximately 50 images per brand by applying filter to retrieve only licence free images. We use postfix like ‘store’, ‘showroom’, ‘building’ intelligently to maximize number of images containing relevant texts in the top retrieval. Subsequently, we give this collection of images to human annotators who remove all those images which do not contain any text of brand names (e.g., interior of a restaurant). These pruning stages end up retaining 500 brands and more than 10K scene images. The total number of images in our dataset including scenes of business brands, movie posters and book covers are 257K. Based on the content of images, we group our dataset into following three categories: text-KVQA (scene), text-KVQA (movie) and text-KVQA (book).

In order to allow knowledge-enabled questions to be asked, we construct three domain-specific knowledge bases for business brands, movies and books, namely, *KB-business*, *KB-movie* and *KB-book* respectively. To construct these three knowledge bases, we crawl open-source world knowledge bases, e.g., Wikidata [3], IMDb [1] and book catalogue provided by [18] around the anchor entities.[1] Each knowledge fact is a triplet connecting two entities with a relation. An example of these triplets is: *KFC, started in, 1930*.

We use knowledge facts and ground truth scene text words to generate question-answer pairs of varying complexity for each image. Our questions are of diverse nature, such as factual questions (e.g., *Which petrol pump is this?*, *What does this store sell?*, *In which year this movie was released?*) and binary questions (e.g., *Can I get a sandwich here?*, *Is this a Dutch brand?*, *Is this a romantic movie?*). Here, we would like to highlight that unlike other recently introduced datasets, answers to the questions in our dataset

[1] We refer names of business brands and title of movies and books as anchor entities.

may not be directly answerable from visual and textual contents alone. Further, to add natural linguistic complexity, we paraphrase questions with the help of human annotators and randomly choose either original or paraphrased questions for each image.

We split the dataset images into train, test and validation sets by randomly dividing 80%, 10% and 10% of anchor entities for train, test, and validation, respectively. We make sure that these splits are disjoint, i.e., if an anchor entity belongs to train set, then it can neither belong to validation set nor to test set. It should be noted that this *zero-shot setting* is close to the real-word scenario where it is highly unlikely to have all anchor entities (e.g. business brands, movie titles, etc.) seen during training. Figure 2 shows some sample images, question- ground-truth answer pairs, and supporting fact from knowledge base. Please note that supporting fact is not explicitly provided during training and inference of our method rather it is mined from the large-scale knowledge bases.

Besides text detection and recognition *in the wild* setting, the major challenges present in text-KVQA are large-answer space, linguistic diversity and zero-shot setting. We firmly believe that our dataset will be useful for text recognition, VQA as well as QA on knowledge base communities.

4. Methodology

Our visual question answering model, which can read and reason, works as follows. We begin by generating word proposals and visual content proposals. These two modules leverage the best performing scene text recognition and image recognition methods. We, then, fuse these proposals, questions and knowledge-base triplets (facts), and obtain relevant facts. Subsequently, these relevant facts are used to construct a multi-relational graph.

Given this multi-relational graph, we intend to perform reasoning based on word proposals, visual concept proposals and question. A natural choice for this is 'gated graph neural network' (GGNN) [27] which is emerging as a powerful tool to perform reasoning over graphs. Note that GGNNs have been used in a variety of tasks including symbolic QA [27] to more complex visual reasoning [30]. We make appropriate changes to classical GGNN framework to seamlessly integrate cues coming from image, question and knowledge base to arrive at the intended answer. Figure 3 summarizes our proposed VQA scheme.

4.1. Proposal module

Given an image, first step of our knowledge-enabled VQA pipeline is to obtain a set of words recognized in it. Now, even the best performing scene text recognition methods do not work well "in the wild" setting due to the presence of occlusion, stylized fonts and different orientations of texts. Therefore, we take a different approach. Instead of just relying on exact text recognition, we perform a search in the list of KG entities, and take all those words as word proposals which are nearby to the recognized text in the normalized edit distance space. At the end of this step, we obtain a set W of n words and their respective confidence scores computed using normalized edit distance with KG entities. Each word in set W is represented using word2vec embeddings trained on Wikipedia [21], i.e., $W = \{\mathbf{w}_1, \mathbf{w}_2, \dots \mathbf{w}_n\}$. It should be noted that one or more of these words are often an anchor entity (e.g., brand name). In the experimental section, we evaluate four modern scene text recognition methods, and choose the best among them to be used along with the subsequent modules.

Next, we obtain the visual content proposals. It should be noted that OCRed texts in images can be noisy, and visual cues (e.g., scene) can boost overall performance. To this end, we rely on Places [55] for scene recognition and a fine-tuned VGG-16 model for representing visual contents from movie posters and book covers. Finally, we obtain a set of m visual content proposals V along with their confidence scores. Each visual content proposal in V is represented using word2vec embeddings trained on Wikipedia [21], i.e., $V = \{\mathbf{v}_1, \mathbf{v}_2, \dots \mathbf{v}_m\}$.

4.2. Fusion module

Once word and visual content proposals are obtained, the next step in our framework is to perform fusion of these multimodal cues. The primary objectives of this fusion module are two folds - (i) computational scalability, and (ii) obtaining relevant facts from web-scale knowledge base even in the case where word proposals are weak.

In this module, we have three cues: two coming from the image, namely, word proposals W, visual content proposals V; and one coming from language, i.e., average word2vec representation of words in question ($\mathbf{q}$). We combine these to find a set of relevant facts from our large-scale knowledge base. Let us denote i^{th} fact of our knowledge base as $f_i = (\mathbf{h}_i, \mathbf{r}_i, \mathbf{t}_i)$ denoting word2vec representations of head entity, relation and tail entity respectively. One example of our knowledge fact is Subway (head), is brand of (relation), United States of America (tail).

Given a set of word proposals W, visual content proposals V and question $\mathbf{q}$, we compute fusion score for i^{th} knowledge fact as follows.

$$\begin{aligned} F(\mathbf{h}_i, \mathbf{r}_i, \mathbf{t}_i) = \max_{j,k} \quad & \alpha_w s_{w_j} \langle \mathbf{w}_j, (\mathbf{h}_i, \mathbf{r}_i, \mathbf{t}_i) \rangle \\ + \quad & \alpha_v s_{v_k} \langle \mathbf{v}_k, (\mathbf{h}_i, \mathbf{r}_i, \mathbf{t}_i) \rangle \\ + \quad & \alpha_q \langle \mathbf{q}, (\mathbf{h}_i, \mathbf{r}_i, \mathbf{t}_i) \rangle . \end{aligned} \tag{1}$$

Here, s_{w_j} and s_{v_k} denote confidence scores of j^{th} word

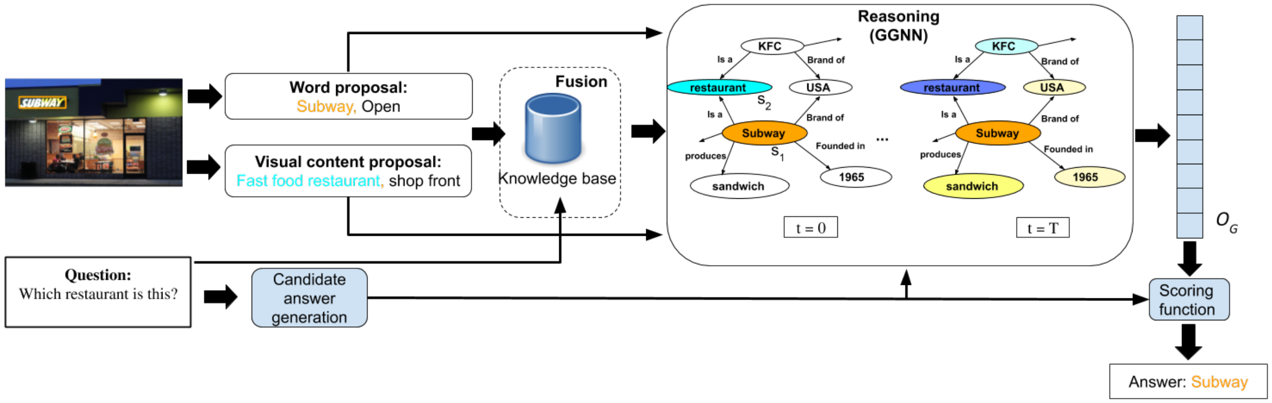


Figure 3. Proposed knowledge-enabled VQA model that can read and reason. For details please refer to Section 4.

proposal and k^{th} visual content proposal in the image, respectively. Further, $\langle \mathbf{x}, (\mathbf{h}_i, \mathbf{r}_i, \mathbf{t}_i) \rangle = \mathbf{x}.\mathbf{h}_i + \mathbf{x}.\mathbf{r}_i + \mathbf{x}.\mathbf{t}_i$. The parameters α_w, α_v and α_q are determined on a validation set with a constraint to maximize recall of relevant fact retrieval in top-K.

Now, by using fusion score for each knowledge fact, we retrieve top-K knowledge facts for each question and image pair, and construct a multi-relational graph.

4.3. GGNN formulation and reasoning

We obtain a multi-relational graph G from the above module. Now, our task is to perform reasoning on this graph to arrive at an accurate answer. We choose gated graph neural network (GGNN) [27] for this task. GGNN is a manifestation of graph neural networks for the sequential outputs. It uses gated recurrent units, and unroll the recurrence for a fixed number of steps and use backpropagation through time in order to compute the gradients. Our GGNN formulation works as follows.

Given a graph with N nodes, a task specific embedding of nodes $\mathbf{x}_u$ for each node u, word proposals W, visual content proposals V, and an answer candidate e_i, our goal is to produce graph-level embedding O_G for a graph classification task. It should be noted that here graph classification task is to determine if a candidate answer e_i is a ground truth answer or not. In order to obtain candidate answer, given a question, we first predict the answer-type. Predicting answer types has shown beneficial impact in VQA [22]. At coarser-level, answers in text-KVQA are either an entity in knowledge graph, e.g., Subway, Car showroom or obtained using reasoning over graph, e.g., Yes, No. Further, knowledge graph provides finer type of entities (e.g., brand name, year, country). We use these finer entity types along with yes-no and question-answer pairs in the training set for learning to predict an answer type for a given question. To this end, we train a simple multi-layer perceptron by representing each question using bi-directional long-short term memory (BLSTM), and posing answer prediction as multi-class classification problem. Once answer-type is predicted, we trivially generate a small set of c candidate answers $C = \{e_1, e_2, e_3, ..., e_c\}$ in one-hop of the anchor entity. Note that here e_i can either be yes-no or an anchor or non-anchor entity in knowledge graph.

Given above in hand, we define a scoring function such that maximum of which corresponds to answer (a^*).

$$a^* = \arg\max_{e_i \in C} \mathcal{S}(O_G, \mathbf{e}_i). \tag{2}$$

Here, O_G is a graph embedding obtained using GGNN explained in the subsequent paragraphs, and $\mathbf{e}_i$ is word2vec embedding of candidate answer e_i. It should be noted that scoring function is a binary classifier whose task is to determine whether a candidate answer e_i is correct answer or not. We train $\mathcal{S}$ using binary cross entropy loss on training set.

Graph-level embedding: Given a graph G = (vertices: $\mathcal{U}$, typed edges: $\mathcal{E}$), question $\mathbf{q}$, word proposals W, visual content proposals V and a candidate answer e_i, we obtain graph-level embedding. To this end, we first define initial node embedding for a node u as follows.

$$\mathbf{x}_u = \begin{cases} [\mathbf{n}_u,\ 0,\ 1,\ c_u]; & \text{if node } u \text{ is a word proposal,} \\ [\mathbf{n}_u,\ 1,\ 0,\ c_u]; & \text{if node } u \text{ is an answer candidate,} \\ [\mathbf{n}_u,\ 1,\ 0,\ c_u]; & \text{if node } u \text{ has highest embedding} \\ & \text{similarity with the question,} \\ [\mathbf{n}_u,\ 0,\ 0,\ c_u]; & \text{ Otherwise.} \end{cases} \tag{3}$$

Here $\mathbf{n}_u$ is a word2vec embedding for node u. If node u does not represent a word or visual content obtained using image (e.g., United States of America), the value c_u is set to

0, otherwise the value c_u is the confidence score for word or visual content proposals obtained from the image according to what node u represents. For example: if node u represents 'Subway' then c_u is the confidence score of recognizing text 'Subway' in the image.

Suppose $\mathbf{h}_u^{(t)}$ is a hidden state representation for node u at the GGNN time stamp t. We begin at $t = 0$, and initialize hidden states as $\mathbf{x}_u$. If needed, we do appropriate padding. Further, we use our graph structured encoding (i.e., adjacency matrix) A to retrieve the hidden states of adjacent nodes based on the *relation types* between them. The hidden states are then updated by a gated update module as follows.

$$\mathbf{h}_u^{(0)} = [\mathbf{x}_u^{\mathrm{T}}, \mathbf{0}]^{\mathrm{T}}; \;\; \mathbf{a}_u^{(t)} = A_u^{\mathrm{T}}\, [\mathbf{h}_1^{(t-1)\,\mathrm{T}} \dots \mathbf{h}_N^{(t-1)\,\mathrm{T}}]^{\mathrm{T}} + \mathbf{b}, \tag{4}$$

$$\mathbf{z}_u^t = \sigma(\mathbf{U}_1^z\, \mathbf{a}_u^{(t)} + \mathbf{U}_2^z\, \mathbf{h}_u^{(t-1)}), \tag{5}$$

$$\mathbf{r}_u^t = \sigma(\mathbf{U}_1^r\, \mathbf{a}_u^{(t)} + \mathbf{U}_2^r\, \mathbf{h}_u^{(t-1)}), \tag{6}$$

$$\tilde{\mathbf{h}}_u^{(t)} = \tanh(\mathbf{U}_1\, \mathbf{a}_u^{(t)} + \mathbf{U}_2\, (\mathbf{r}_u^t \odot \mathbf{h}_u^{(t-1)})), \tag{7}$$

$$\mathbf{h}_u^{(t)} = (1 - \mathbf{z}_u^t) \odot \mathbf{h}_u^{(t-1)} + \mathbf{z}_u^t \odot \tilde{\mathbf{h}}_u^{(t)}. \tag{8}$$

After T timesteps, we obtain the final hidden states. Here, A_u, $\mathbf{U}_1$ and $\mathbf{U}_2$ are the adjacency matrix of the graph for node u, and learned parameters respectively. Operator $\odot$ denotes element-wise multiplication. From above, graph-level embedding (O_G) is computed as follows.

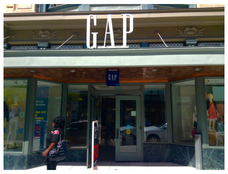
$$O_G = \tanh(\Sigma_{u \in \mathcal{U}}\; \sigma(f_\theta(\mathbf{h}_u^{(T)}, \mathbf{x}_u)) \odot \tanh(f_\phi(\mathbf{h}_u^{(T)}, \mathbf{x}_u))) \tag{9}$$

where, $\sigma()$ acts as an attention mechanism that decides relevant nodes for question-answering task. f_θ and f_ϕ are neural networks which take concatenation of hidden state and initial node embedding as input, and return real valued vector as output. Graph embedding O_G and an answer candidate is fed to a scoring function $\mathcal{S}$ to obtain a score for an answer candidate e_i. This scoring function is essentially a multi-layer perceptron trained on a training set to determine whether candidate answer e_i is correct answer or not.

5. Experiments and Results

In this section, we perform rigorous experimental analysis and show ablation studies to validate the effectiveness of our proposed model.

Evaluation of proposal module: Given an image, we first detect and recognize the texts appearing in it. We use combinations of four modern scene text detection and recognition methods as shown in Table 1. Once these methods provide recognized texts, we perform normalized edit distance (NED) based corrections using a list of candidate entities in our knowledge base to enhance entity recall. In Table 1, we report entity recall on all the three categories of text-KVQA without correction as well as with correction using $NED = 0.5$. Poor show of these state-of-the-art methods indicates the challenges associated with text detection and recognition in our dataset. We choose to use TextSpotter [16] and PixelLink [12]+CRNN [39] outputs in the next stage of our VQA module, owing to their relatively better performance. For the sake of simplicity, we refer to these methods as *photoOCR-1* and *photoOCR-2* respectively.



Word Recognition: {GALP}
Word Proposals:{GALP, GAP}
Visual Content Proposal: {clothing store, department store, gift shop}
Q. Can I get clothes here?
A (text only):. No
*A (full model):*Yes

Figure 4. Integration of visual content proposals help our full model to recover from noisy word recognition.

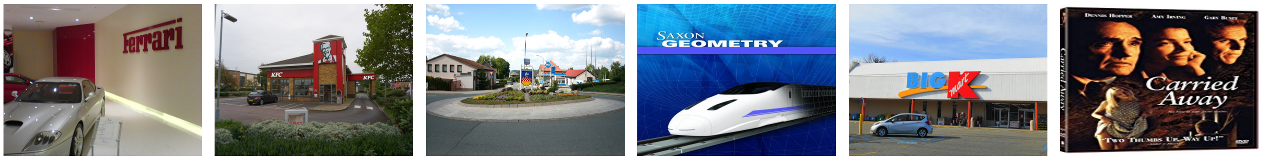
Two broad categories of visual contents can be observed throughout our dataset - (i) natural scene content for the images in text-KVQA (scene) subset of the dataset and (ii) artificially composed visual contents on movie posters and covers of books. We use Places [55] and a VGG-16 finetuned model for recognizing these visual contents for categories-(i) and (ii), respectively. Since category names in our dataset are not exactly the same in Places, we could not perform quantitative analysis on visual content evaluation of places. However, we evaluate the visual content classification module towards genre classification on movie posters and book covers and achieve 25% and 27% top-1 accuracy, and 58% and 59% top-5 accuracy, respectively.

Evaluation of fusion scheme: Once the word and visual content proposals have been extracted, they along with the question are fused with facts from knowledge bases, i.e., *KB-business*, *KB-movie* and *KB-book*, to obtain the top-*100* relevant facts. For the subsequent module (i.e., VQA using GGNN reasoning) higher recall of supporting fact is expected at this stage. To evaluate the contribution of each of these modalities, we perform ablation study as follows. (i) W: Only word proposals are fused with knowledge facts, i.e. ($\alpha_w = 1, \alpha_v = 0, \alpha_q = 0$). (ii) V: Only visual content proposals are fused with knowledge facts, i.e. ($\alpha_w = 0, \alpha_v = 1, \alpha_q = 0$) (iii) $\mathbf{q}$: Only question is fused with knowledge facts, i.e. ($\alpha_w = 0, \alpha_v = 0, \alpha_q = 1$) (iv) $W + V + \mathbf{q}$: optimal combination of word proposal, visual proposals and the question are fused with knowledge facts, i.e. ($\alpha_w = 0.7, \alpha_v = 0.2, \alpha_q = 0.1$). The values of hyper-parameters α_w, α_w and α_q are determined on a validation set.

Table 2 shows the individual fact recall@top-100 (i.e., percentage of supporting facts contained in top-100 relevant facts) results for above four variants. We observe that word proposals are the most helpful in obtaining higher recall of supporting facts which is further boosted by optimally com-

Method	text-KVQA (scene)		text-KVQA (book)		text-KVQA (movie)	
	Original	NED=0.5	Original	NED=0.5	Original	NED=0.5
CTPN [43] + CRNN [39]	0.16	0.38	0.15	0.27	0.22	0.37
EAST [57] + CRNN [39]	0.36	0.60	0.43	0.66	0.24	0.42
Text Spotter [16]	0.38	0.58	**0.53**	**0.70**	**0.35**	**0.48**
PixelLink [12] + CRNN [39]	**0.43**	**0.64**	0.38	0.56	0.14	0.27

Table 1. We report recall statistics without edit distance correction (original) as well as recall after normalized edit distance (NED=0.5) correction for state-of-the-art scene text detection and recognition methods. We refer methods in row-3 and row-4 as *PhotoOCR-1* and *PhotoOCR-2* respectively from here on.



(a)

Words: {Ferrari}
Visuals: {auto showroom, garage}
Q. What is this?
A. Auto showroom.
SF: Ferrari is an auto showroom.

(b)

Words: {KFC}
Visuals: {fastfood restaurant, food court}
Q. Is this an American brand?
A. Yes.
SF: KFC is a brand of USA.
(**Observation:** Identifies American as USA.)

(c)

Words: {Amul, Aral}
Visuals: {gas station, highway}
Q. Can I fill fuel in my car?
A. Yes
SF: Aral produces gas.
(**Observation:** Amul is a restaurant, but visual proposal helps here.)

(d)

Words: {Saxon Geometry}
Visuals: {Travel, Maths}
Q. Who published of this book?
A. Saxon publishers
SF: Saxon Geometry is published by Saxon publishers.

(e)

Words: {bligh, CVS}
Visuals: {motel, pharmacy}
Q. What is this?
A. Pharmacy
SF: CVS is pharmacy.
(**Observation:** Both word and visual proposal mislead here.)

(f)

Words: {Carried Away}
Visuals: {Drama, Thriller}
Q. In which year this movie was released?
A. 1996
SF: Carried Away was released in 1996.

Figure 5. A set of qualitative results obtained using proposed method. Answer in red indicate failure cases. **[Best viewed in color]**.

Fusions	Fact recall (in %)
W (*photoOCR1*)	55.8
W (*photoOCR2*)	59.9
V	20.8
q	5.3
W (*photoOCR1*)+V+**q**	58.9
W (*photoOCR2*)+V+**q**	**60.7**

Table 2. Relevant fact recall (in %) at top-100 retrieval for the text-KVQA (scene) dataset based on different combination of word proposals (W), visual content proposals (V) and question (**q**).

bining it with visual proposals and question.

Evaluation of GGNN reasoning: We evaluate GGNN reasoning by reporting visual question answering accuracy on text-KVQA. The quantitative results obtained by our method are shown in Table 3. We use three ways of obtaining word proposals: (i) using *photoOCR-1*, (ii) using *photoOCR-2*, and (iii) ideal text recognition (oracle) setting. We compare variants of our methods with the following traditional and KB-based VQA models.

(i) Traditional VQA models: These methods rely on visual cues alone and are not designed to read texts in images. We have chosen - (a) BoW + CNN (b) BLSTM (language Only), (c) BLSTM + CNN, (d) Hierarchical Co-Attention [29], and (e) Bilinear Attention Network [24] as traditional VQA baselines to compare. Among the above five baselines, first three are basic VQA models used in early works [5]. Method-(d) reasons jointly about the visual and question attention, and Method-(e) builds two separate attention distribution for both the images and the questions, and then uses bilinear attention to predict the answers.

(ii) QA over KB based method: The task of question answering over knowledge bases has gained attention in the NLP community over the past few years, and many approaches have been proposed. One of the well-established approaches among these is memory network [50]. Therefore, we evaluate performance of memory network baseline on text-KVQA by replacing our GGNN module with memory units, while keeping the remaining modules of our proposed framework identical. To this end, we represent relevant facts obtained after our fusion module as memory units and train the memory network. The hyper-parameters of this network are chosen using a validation set.

We observe that the traditional VQA methods perform poorly on all the three categories of our dataset. This poor show indicates the importance of reading text for VQA task which these methods are not capable of. Secondly, these are

Method	text-KVQA (scene)	text-KVQA (book)	text-KVQA (movie)
Traditional VQA methods			
BoW + CNN	11.5	8.7	7.0
BLSTM (language Only)	17.0	12.4	11.3
BLSTM + CNN [5]	19.8	17.3	15.7
HiCoAttenVQA [29]	22.2	20.2	18.4
BAN [24]	23.5	22.3	20.3
QA over KB based method			
Memory network [50](with *photoOCR-1*)	49.0	57.2	42.0
Memory network [50](with *photoOCR-2*)	52.6	47.8	22.2
Our variants			
Vision only	21.8	19.8	18.2
Text only (with *photoOCR-1*)	48.9	55.0	41.4
Text only (with *photoOCR-2*)	52.2	48.6	20.5
Full model (with *photoOCR-1*)	52.2	**62.7**	**45.2**
Full model (with *photoOCR-2*)	**54.5**	49.8	23.0
Oracle (ideal text recognition)	80.1	71.3	76.2

Table 3. Comparison of variants of our proposed framework with traditional VQA methods and a QA over KB method [50] (in %). Methods *PhotoOCR-1* and *PhotoOCR-2* use TextSpotter [16] and PixelLink [12] + CRNN [39] respectively for obtaining word proposals.

fully-supervised models, and do not cope well with the challenges arising due to *zero-shot nature* of the text-KVQA. Our proposed knowledge-enabled VQA model which is capable of reading text in images significantly outperforms these baselines methods. Moreover, our GNNN-based full model also achieves improved VQA performance over the memory network based KB-QA baseline. As mentioned earlier, memory network treats KG as a flattened table of facts. Therefore, it is hard to exploit the structural information present in the graph which weakens the reasoning performance.

The superior performance of our method can be attributed to the seamless integration of visual and text recognition cues and powerful reasoning over graph using GGNN. In order to study the effect of different modalities towards the overall performance of the proposed framework, we perform an ablation study. In Table 3 we report results of variants of our method with text only, vision only and full model which seamlessly integrates visual content and word proposals. Our text only and vision only methods use word proposals and visual content proposals respectively, along with GGNN reasoning. As expected, since questions in our dataset are often connected to the text appearing in images, the vision only variant fails to perform well, especially in comparison to the text only baselines. However, the utility of the visual content proposals in the overall VQA performance can not be underestimated. This is primarily because even the best text recognition methods are not perfect. Adding visual content proposals to the framework provides a way to rectify errors due to noisy text recognition, and add robustness. This can also be understood via an example in Figure 4. Due to noisy word recognition, the text-only model has led to incorrect answer, whereas, the full model (visual content + text) is able to correctly answer the question. Therefore, our final model is designed to integrate both these modalities and subsequently perform reasoning over graph using GGNN, which helps it achieve improved performance in comparison to these ablations and baselines.

Figure 5 shows a set of example results obtained by the proposed method. We observe that our method recovers well even from weak word proposals utilizing visual cues, e.g., Figure 5 (c). However, if both text and visual recognition mislead the method, it fails to generate correct answer, e.g., Figure 5 (e). Please refer to the supplementary material for more qualitative results and their detailed analysis.

6. Summary and future work

In this work, we have taken the first step towards knowledge-enabled VQA model that can read and reason. We approached this problem by seamlessly integrating visual cues, textual cues and rich knowledge bases, and performed reasoning using a novel GGNN formulation. Our approach significantly outperformed traditional VQA models as they are not designed to read text in images as well as a QA over KB-based method. Also, as part of our work, we have introduced a large-scale challenging dataset, namely, text-KVQA containing a series of natural and engaging questions about images. The current method and dataset, however, are limited to questions which can be answerable from one-hop reasoning on the knowledge graph. As future research, we would like to develop models that perform multi-hop and more complex reasoning on knowledge graphs, as well as pose the VQA task in our dataset as visual dialogue.

References

[1] https://www.imdb.com/. IMDB, accessed March 10, 2019. 2, 3

[2] https://www.kaggle.com/neha1703/movie-genre-from-its-poster. Kaggle, accessed August 11, 2019. 3

[3] https://www.wikidata.org/. Wikidata Knowledge Graph, accessed March 10, 2019. 3

[4] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Deep compositional question answering with neural module networks. *CoRR*, 2015. 2

[5] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: visual question answering. In *ICCV*, 2015. 1, 2, 7, 8

[6] Hedi Ben-Younes, Rémi Cadène, Nicolas Thome, and Matthieu Cord. Mutan: Multimodal tucker fusion for visual question answering. *ICCV*, 2017. 2

[7] Alessandro Bissacco, Mark Cummins, Yuval Netzer, and Hartmut Neven. PhotoOCR: Reading Text in Uncontrolled Conditions. In *ICCV*, 2013. 2

[8] Ali Furkan Biten, Ruben Tito, Andrés Mafla, Lluís Gómez, Marçal Rusiñol, Ernest Valveny, C. V. Jawahar, and Dimosthenis Karatzas. Scene text visual question answering. *CoRR*, abs/1905.13648, 2019. 2, 3

[9] Fedor Borisyuk, Albert Gordo, and Viswanath Sivakumar. Rosetta: Large scale system for text detection and recognition in images. In *KDD*, 2018. 2

[10] Adam Coates, Blake Carpenter, Carl Case, Sanjeev Satheesh, Bipin Suresh, Tao Wang, David J. Wu, and Andrew Y. Ng. Text Detection and Character Recognition in Scene Images with Unsupervised Feature Learning. In *ICDAR*, 2011. 2

[11] Abhishek Das, Harsh Agrawal, C. Lawrence Zitnick, Devi Parikh, and Dhruv Batra. Human Attention in Visual Question Answering: Do Humans and Deep Networks Look at the Same Regions? In *EMNLP*, 2016. 2

[12] Dan Deng, Haifeng Liu, Xuelong Li, and Deng Cai. Pixellink: Detecting scene text via instance segmentation. In *AAAI*, 2018. 6, 7, 8

[13] Khaoula Elagouni, Christophe Garcia, Franck Mamalet, and Pascale Sébillot. Combining multi-scale character recognition and linguistic knowledge for natural scene text OCR. In *DAS*, 2012. 2

[14] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv:1606.01847*, 2016. 2

[15] Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu. Are you talking to a machine? dataset and methods for multilingual image question answering. In *NIPS*, 2015. 2

[16] Ankush Gupta, Andrea Vedaldi, and Andrew Zisserman. Synthetic data for text localisation in natural images. In *CVPR*, 2016. 1, 2, 6, 7, 8

[17] Pan He, Weilin Huang, Tong He, Qile Zhu, Yu Qiao, and Xiaolin Li. Single shot text detector with regional attention. In *ICCV*, 2017. 2

[18] Brian Kenji Iwana, Syed Tahseen Raza Rizvi, Sheraz Ahmed, Andreas Dengel, and Seiichi Uchida. Judging a book by its cover. *arXiv preprint arXiv:1610.09204*, 2016. 3

[19] Max Jaderberg, Andrea Vedaldi, and Andrew Zisserman. Deep Features for Text Spotting. In *ECCV*, 2014. 2

[20] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, 2017. 1

[21] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. In *EACL*, April 2017. 4

[22] Kushal Kafle and Christopher Kanan. Answer-type prediction for visual question answering. In *CVPR*, 2016. 5

[23] Sezer Karaoglu, Ran Tao, Theo Gevers, and Arnold WM Smeulders. Words matter: Scene text for image classification and retrieval. *IEEE transactions on multimedia*, 19(5):1063–1076, 2017. 3

[24] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. In *NeurIPS*, 2018. 2, 7, 8

[25] Jin-Hwa Kim, Sang-Woo Lee, Dong-Hyun Kwak, Min-Oh Heo, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. Multimodal residual learning for visual QA. *NeurIPS*, 2016. 2

[26] Ankit Kumar, Ozan Irsoy, Peter Ondruska, Mohit Iyyer, James Bradbury, Ishaan Gulrajani, Victor Zhong, Romain Paulus, and Richard Socher. Ask me anything: Dynamic memory networks for natural language processing. In *ICML*, 2016. 2

[27] Yujia Li, Daniel Tarlow, Marc Brockschmidt, and Richard S. Zemel. Gated graph sequence neural networks. *CoRR*, abs/1511.05493, 2015. 2, 4, 5

[28] Minghui Liao, Baoguang Shi, and Xiang Bai. Textboxes++: A single-shot oriented scene text detector. *IEEE Trans. Image Processing*, 27(8):3676–3690, 2018. 2

[29] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering, 2015. 2, 7, 8

[30] Kenneth Marino, Ruslan Salakhutdinov, and Abhinav Gupta. The more you know: Using knowledge graphs for image classification. In *CVPR*, 2017. 2, 4

[31] Anand Mishra, Karteek Alahari, and C. V. Jawahar. Top-Down and Bottom-Up Cues for Scene Text Recognition. In *CVPR*, 2012. 2

[32] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. OCR-VQA: Visual question answering by reading text in images. In *ICDAR*, 2019. 2, 3

[33] Medhini Narasimhan, Svetlana Lazebnik, and Alexander G. Schwing. Out of the box: Reasoning with graph convolution nets for factual visual question answering. *NeurIPS*, 2018. 2

[34] Lukas Neumann and Jiri Matas. Real-time scene text localization and recognition. In *CVPR*, 2012. 2

[35] Mengye Ren, Ryan Kiros, and Richard S. Zemel. Exploring models and data for image question answering. In *NeurIPS*, 2015. 2

[36] Idan Schwartz, Alexander G. Schwing, and Tamir Hazan. High-order attention models for visual question answering. *NeurIPS*, 2017. 2

[37] Sanket Shah, Anand Mishra, Naganand Yadati, and Partha Pratim Talukdar. KVQA: Knowledge-aware visual question answering. In *AAAI*, 2019. 2

[38] Baoguang Shi, Xiang Bai, and Cong Yao. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(11):2298–2304, 2017. 2

[39] Baoguang Shi, Xiang Bai, and Cong Yao. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(11):2298–2304, 2017. 6, 7, 8

[40] Cunzhao Shi, Chunheng Wang, Baihua Xiao, Yang Zhang, Song Gao, and Zhong Zhang. Scene Text Recognition Using Part-Based Tree-Structured Character Detection. In *CVPR*, 2013. 2

[41] Kevin J. Shih, Saurabh Singh, and Derek Hoeim. Where to look: Focus regions for visual question answering. In *CVPR*, 2016. 2

[42] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards VQA models that can read. In *CVPR*, 2019. 2, 3

[43] Zhi Tian, Weilin Huang, Tong He, Pan He, and Yu Qiao. Detecting text in natural image with connectionist text proposal network. In *ECCV*, 2016. 7

[44] Denny Vrandecic and Markus Krötzsch. Wikidata: a free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85, 2014. 2

[45] Kai Wang, Boris Babenko, and Serge Belongie. End-to-End Scene Text Recognition. In *ICCV*, 2011. 2

[46] Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton van den Hengel. Explicit knowledge-based reasoning for visual question answering. In *IJCAI*, 2017. 2

[47] Peng Wang, Qi Wu, Chunhua Shen, Anthony R. Dick, and Anton van den Hengel. Explicit knowledge-based reasoning for visual question answering. In *IJCAI*, 2017. 2

[48] Peng Wang, Qi Wu, Chunhua Shen, Anthony R. Dick, and Anton van den Hengel. FVQA: fact-based visual question answering. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(10):2413–2427, 2018. 2

[49] Jerod J. Weinman, Zachary Butler, Dugan Knoll, and Jacqueline Feild. Toward Integrated Scene Text Reading. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(2):375–387, 2014. 2

[50] Jason Weston, Sumit Chopra, and Antoine Bordes. Memory networks. *CoRR*, abs/1410.3916, 2014. 2, 7, 8

[51] Huijuan Xu and Kate Saenko. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. *ECCV*, 2016. 2

[52] Licheng Yu, Eunbyung Park, Alexander C. Berg, and Tamara L. Berg. Visual madlibs: Fill in the blank image generation and question answering. *ICCV*, 2015. 2

[53] Muhan Zhang, Zhicheng Cui, Marion Neumann, and Yixin Chen. An end-to-end deep learning architecture for graph classification, 2018. 2

[54] Yuyu Zhang, Hanjun Dai, Zornitsa Kozareva, Alexander J. Smola, and Le Song. Variational reasoning for question answering with knowledge graph. In *AAAI*, 2018. 2

[55] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 1, 4, 6

[56] Bolei Zhou, Yuandong Tian, Sainbayar Sukhbaatar, Arthur Szlam, and Rob Fergus. Simple baseline for visual question answering. abs/1512.02167, 2015. 2

[57] Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang. EAST: an efficient and accurate scene text detector. In *CVPR*, 2017. 2, 7

[58] Anna Zhu, Renwu Gao, and Seiichi Uchida. Could scene context be beneficial for scene text detection? *Pattern Recognition*, 58:204–215, 2016. 3

[59] Yuke Zhu, Oliver Groth, Michael S. Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. In *CVPR*, 2016. 1

[60] C. Lawrence Zitnick, Aishwarya Agrawal, Stanislaw Antol, Margaret Mitchell, Dhruv Batra, and Devi Parikh. Measuring machine intelligence through visual question answering. *AI Magazine*, 37(1):63–72, 2016. 2