2020 25th International Conference on Pattern Recognition (ICPR)
Milan, Italy, Jan 10-15, 2021

Multi-modal Contextual Graph Neural Network for Text Visual Question Answering

Yaoyuan Liang†, Xin Wang‡§*, Xuguang Duan‡ and Wenwu Zhu‡§*
†Beijing University of Posts and Telecommunication, liangyaoyuan@bupt.edu.cn
‡Department of Computer Science and Technology, Tsinghua University §Peng Cheng Laboratory, China
{xin_wang, wwzhu}@tsinghua.edu.cn*, dxg18@mails.tsinghua.edu.cn

***Abstract*—Text visual question answering (TextVQA) targets at answering the question related to texts appearing in the given images, posing more challenges than VQA by requiring a deeper recognition and understanding of various shapes of human-readable scene texts as well as their meanings in different contexts. Existing works on TextVQA suffer from two weaknesses: i) scene texts and non-textual objects are processed separately and independently without considering their mutual interactions during the question understanding and answering process, ii) scene texts are encoded only through word embeddings without taking the corresponding visual appearance features as well as their potential relationships with other non-textual objects in the images into account. To overcome the weakness of existing works, we propose a novel multi-modal contextual graph neural network (MCG) model for TextVQA. The proposed MCG model can capture the relationships between visual features of scene texts and non-textual objects in the given images as well as utilize richer sources of multi-modal features to improve the model performance. In particular, we encode the scene texts into richer features containing textual, visual and positional features, then model the visual relations between scene texts and non-textual objects through a contextual graph neural network. Our extensive experiments on real-world dataset demonstrate the advantages of the proposed MCG model over baseline approaches.**

I. Introduction

Visual Question Answering [2], aiming to correctly answer natural language questions given images, has been a key problem towards image understanding and cross-modal intelligence. Besides, as shown in [29], the current VQA models do not own the ability to read scene text information in images like humans, limiting the applicability of VQA models in many real-world scenarios. For example, in visually-impaired assistant devices, most of the user requests may involve the scene text containing information that the users are interested in. The users may ask 'What time does my phone display?' or 'What number is my heartbeat recorder displaying?'. To handle these problems, text visual question answering (TextVQA) targets at exploring the ability to analyze images, questions and scene texts, as well as figuring out the correct answer based on the image objects information and scene text contents, which poses more challenges than VQA by requiring a deeper understanding of the meanings of various human-readable scene texts in different contexts.

Existing works on TextVQA [3], [19], [29] isolate the scene texts away from other non-textual objects by processing pre-extracted OCR tokens and visual features separately, ignoring the relationships between the scene texts and other non-textual objects in the given images. Take the scene in Fig. 1(a) as an example, existing TextVQA models will fail to predict the correct answer to the question such as "What is the license plate of the white car?". This is because current models process scene texts and objects separately, which means they could never distinguish the two *licence plates* without knowing the relationship between licence plates and cars. Besides, to correctly answer the aforementioned question, it is also necessary to get the position of the license plate and car as well as the color of the car, indicating that a good TextVQA model should obtain a comprehensive understanding about the scene texts, the non-textual objects and their relationships. However, given the complex and diverse contents in real-world images, precisely generating a comprehensive description for each of the scene texts and non-textual objects as well as capturing their relationships in these images is very challenging.

To tackle the challenges, we propose a novel multi-modal contextual graph neural network (MCG) model for TextVQA

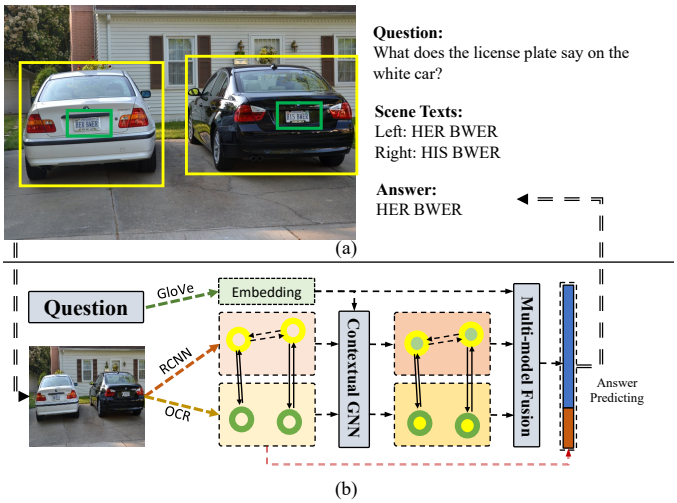


Fig. 1. Illustrations of the task, and model structure. On the top, the question asks the details about a **license plate** which requires a comprehensive understanding of objects and scene texts. In the bottom, we illustrate our model, which consists of a visual feature, scene texts feature extractor, and the GNN-based contextual information propagation mechanism. Besides, the dynamic answer prediction enables us to predict the answers which could only be seen in the specific image

* Xin Wang and Wenwu Zhu are corresponding authors.

978-1-7281-8808-9/20/$31.00 ©2020 IEEE

2020 25th International Conference on Pattern Recognition (ICPR) | 978-1-7281-8808-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICPR48806.2021.9412891

in this paper. Our proposed MCG model is capable of capturing the relationships between scene texts and non-textual objects to improve model performance. The MCG model also utilizes richer sources of multi-modal features including visual, textual and positional features for a better understanding of the scene texts and non-textual objects in the given images. Particularly, we encode the scene texts into richer features containing textual, visual and positional features, then model the visual relations between scene texts and non-textual objects through a contextual graph neural network. We aggregate visual features, textual features, positional features and contextual relation features together in a non-trivial way such that the information for both textual and visual objects in the images can be as comprehensive as possible. Fig. 1(b) illustrates the overview picture of our MCG model. The extensive experiments on real-world dataset validate the effectiveness of the proposed MCG model against existing TextVQA approaches. Our contributions can be summarized as follows,

- We propose a novel multi-modal contextual graph neural network (MCG) model for TextVQA, which is able to capture the relationships between visual features of scene texts and non-textual objects in the given images.
- We non-trivially aggregate richer sources of multi-modal features including visual, textual and positional features as well as the contextual relation features together for better understanding the scene texts and non-textual objects in the given images.
- We conduct extensive experiments on real-world TextVQA dataset and show that our proposed MCG model outperforms baseline methods on TextVQA in various aspects.

II. Related Work

Given that we utilize graph to capture the relations between scene texts and non-textual objects, we review related works on VQA, graph based VQA and TextVQA, three categories of works that are most relevant to our work in this section.

A. Visual Question Answering

Visual Question Answering (VQA) is a classical but important task towards visual and lingual co-understanding. Recent years, with the development of deep learning techniques, the research communities have been paying more attentions to this task in the domain of images [2], [10], [13] or videos [6], [8], [36]. The communities' interests include the image and question feature processing [14], [18], answer encoding [7], [21]. For better processing the features, we have tried attention mechanism [1], relation networks [17]. For answer encoding and prediction, we have tried to rank the answers with respect to each image, or to predict the answer from the huge candidate pools collected from the huge training set. These works have promised us a lot towards the visual and lingual co-understanding, and we will introduce more details about VQA in the followed several paragraphs.

B. Graph Based Visual Question Answering

The graph based VQA models can be categorized according to the type of the graph constructing methods, *i.e.*, either the graphs be constructed are scene graphs or knowledge graphs. There has been several works utilizing scene graphs [17], [22], [34]. In particular, Xiong *et.al.* [34] focuse on single scene VQA, relying on reasoning over entity graphs. Norcliffe *et.al.* [22] learn image representations that capture question specific relations via combining a graph learning module and graph convolutions. Li. *et.al.* [17] propose the ReGAT model which encodes images into graphs and uses the graph attention mechanism to model multi-type inter-object relations. On the other side, utilizing knowledge graph to enhance VQA performance is becoming a popular trending in VQA models [31], [20], [26], [32], [38]. These works utilize knowledge graph to get complementary and useful external knowledge with the help of memory network [33] and its variants.

C. Text Visual Question Answering (TextVQA)

As a relatively new task, TextVQA aims to overcome the insurmountable difficulties faced by traditional VQA models which do not have the ability to take the scene texts as extra visual clues when predicting answers. There are several works on TextVQA [3], [19], [27], [29]. To be specific, Singh. *et.al.* [29] proposed LoRRA which first takes care of the scene texts alone and then fuses it with results from other modalities such as textual questions and visual objects. This work extends the standard OCR token vocabulary obtained by the OCR extractor to a dynamic vocabulary. Similarly, Mishra. *et.al.* [19] first extract OCR tokens separately before fusing them with results from other modalities. Biten. *et.al.* [3] provide a series of tasks on TextVQA along with a baseline method which enlarges the vocabulary and employs the attention mechanism on image and OCR together. Singh. *et.al.* [27] take the external knowledge into TextVQA with the help of knowledge graph, and propose a new dataset named text-KVQA.

III. MCG:Multi-modal Contextual GNN for TextVQA

In this section, we will briefly describe the task formulation, followed by a detailed introduction on our proposed MCG model for TextVQA.

A. Task Formulation

In vanilla VQA task, the target is to infer the answer $a \in \mathcal{A}$, given an image $\boldsymbol{I}$ and a question $\boldsymbol{q}$ grounded in the image, where $\mathcal{A}$ is the candidate answer set shared by all images. However, in many real-world scenarios such as road navigation, completing a task or answering a question requires not only visual objects but also precise scene texts in the captured images, which leads to the fact that merely relying on results from implicit CNN based models becomes inadequate. This motivates the advent of TextVQA problem in which the answer set is expanded to $\mathcal{A} + \hat{\mathcal{A}}$, where $\mathcal{A}$ is defined the same as VQA while $\hat{\mathcal{A}}$ is an image-specific candidate answer set containing scene texts as well as their

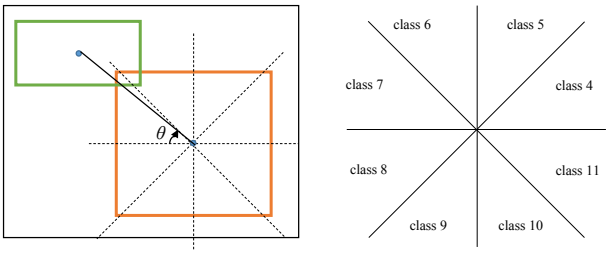


Fig. 2. The illustration of relation categories definition when IoU between two objects is less than 0.5 [35]. Under this situation, class index is relying on the size of relative angle θ_{ij}, set as $(\theta_{ij}/45°) + 3$ (class 4-11).

possible combinations appearing in that image. The TextVQA problem is formally defined as follows:

$$a = \arg \max_{a \in \widehat{\mathcal{A}} + \mathcal{A}} P_\theta(a|\boldsymbol{I}, \boldsymbol{q}), \quad (1)$$

where P_θ is the model to be learned. Inspired by [1], we explicitly model the non-textual objects $\boldsymbol{O}$ and scene texts $\boldsymbol{S}$ in the image simultaneously through expanding (1) as follows:

$$a = \arg \max_{a \in \widehat{\mathcal{A}} + \mathcal{A}} P_\theta(a|\boldsymbol{S}, \boldsymbol{O}, \boldsymbol{q}) \cdot Q_{\phi_1}(\boldsymbol{S}|\boldsymbol{I}) \cdot Q_{\phi_2}(\boldsymbol{O}|\boldsymbol{I}), \quad (2)$$

where Q_{ϕ_1} is a pre-trained OCR token extractor [5] and Q_{ϕ_2} is a pre-trained objects extractor bases on Faster-RCNN [25]. As such, we finish the representations of image, question, scene text and objects, and turn our focus towards the non-trivial relations between the decoupled scene texts, non-textual objects and questions.

To tackle the aforementioned task, in this paper, we propose the **M**ulti-model **C**ontextual **G**NN model. Our MCG model consists of 3 components:

- **Encoding component** that encodes scene texts, non-textual objects and questions into appropriate latent representations
- **Relation modeling component** that employs contextual graph neural network to capture the relationships between visual features of non-textual objects and scene texts.
- **Multi-modal fusion and dynamic prediction component** that first aggregates visual, textual, positional, relational features together and then predicts the final answer through constructing a dynamic candidate answer set for each of the images.

Fig. 3 presents the details for the proposed MCG model, which will be detailedly introduced in the following sections.

B. Encoding Component

Given an image $\boldsymbol{I}$ and a question $\boldsymbol{q}$, as stated above, we encode non-textual objects and scene texts extracted by pre-trained extractors [5], [25], and encode questions with pre-trained word embedding model [24].

Specifically, the non-textual object features $\boldsymbol{O} = \{\boldsymbol{o}_i\}_{i=1}^K$ are extracted with a pre-trained Faster-RCNN model, where K is the number of objects. Each object feature $\boldsymbol{o}_i$ includes a visual feature vector $\boldsymbol{v}_i^{(o)} \in \mathbb{R}^{d_v}$ extracted from the RCNN fully-connected layer and a bounding box $\boldsymbol{b}_i^{(o)} = [x_{min}/W_{img}, y_{min}/H_{img}, x_{max}/W_{img}, y_{max}/H_{img}]$ indicating the top-left and bottom-right coordinates of the bounding box. *i.e.* $\boldsymbol{o}_i = \left[\boldsymbol{v}_i^{(o)}, \boldsymbol{b}_i^{(o)}\right]$.

TABLE I
DEFINITION OF SPATIAL RELATION BETWEEN EVERY TWO OBJECTS.

Class ID	Relation	IoU
1	inside	–
2	cover	–
3	overlap	IoU$\geq$0.5
4-11	rely on θ_{ij} (Fig. 2)	IoU$<$0.5

For the scene texts in the image, we apply scene text detector [5] to identify tokens in the image. We denote the extracted scene texts as $\boldsymbol{S} = \{\boldsymbol{s}_i\}_{i=1}^M$, where $\boldsymbol{s}_i = \left[\boldsymbol{t}_i, \boldsymbol{b}_i^{(s)}, \boldsymbol{v}_i^{(s)}\right]$ represent the tokens, visual bounding box, and visual feature, respectively. $\boldsymbol{t}_i$ and $\boldsymbol{b}_i^{(s)}$ are the output OCR tokens and region bounding boxes from the OCR extractor. $\boldsymbol{v}_i^{(s)}$ is extracted through feeding the bounding box $\boldsymbol{b}_i^{(s)}$ into the Faster-RCNN model. We extract the visual feature for each OCR region to capture the visual clues in that OCR region, *e.g.*, the color, texture and font of the text *etc.*

As for the question $\boldsymbol{q} = \{w_1, w_2, w_3 \dots w_n\}$, we follow the common practice as in other VQA works. We first project the words into an embedding space using a pre-trained word vector model (*e.g.,* GloVe [24]), then encode the resulting word embeddings together with a recurrent network (*e.g.,* LSTM [11]) into corresponding sentence representation $\boldsymbol{q} \in \mathbb{R}^{d_q}$.

C. Relation Modeling Component

Simply encoding non-textual objects and scene texts separately (as in [29]) is not enough to well understand the image in the context of the given question. This is because the relationships between non-textual objects and scene texts play an important role in comprehensively understanding the image and correctly answering the question. Moreover, it is also necessary to figure out the useful information in the context of different questions so that the answers to different questions can be correct (and different). Therefore, we propose our multi-modal contextual GNN to capture the relationship between non-textual objects and scene texts with in the context of different questions.

To start, we represent the union of the visual features of the non-textual objects and scene texts as $\boldsymbol{V} = \{\boldsymbol{v}_i\}_{i=1}^{K+M} = \{\boldsymbol{v}_i^{(o)}\}_{i=1}^K \bigcup \{\boldsymbol{v}_i^{(s)}\}_{i=1}^M$. We first build a graph whose nodes represent the non-textual objects and scene texts, and whose edges represent the relationships between them. Let $\boldsymbol{G} = (\boldsymbol{V}, \boldsymbol{E})$ denote the graph, where the nodes $\boldsymbol{V}$ denote the non-textual objects or scene texts, and the edges $\boldsymbol{E}$ capture the spatial relationships between the nodes. As the constructed graph does not distinguish between non-textual objects and scene texts, we refer *objects* to both non-textual objects and scene texts for succinctness.

1) Graph Construction: Spacial Relationship Modeling.

We construct the graph to capture the spatial relationships between objects, *i.e.,* the edges between nodes (objects)

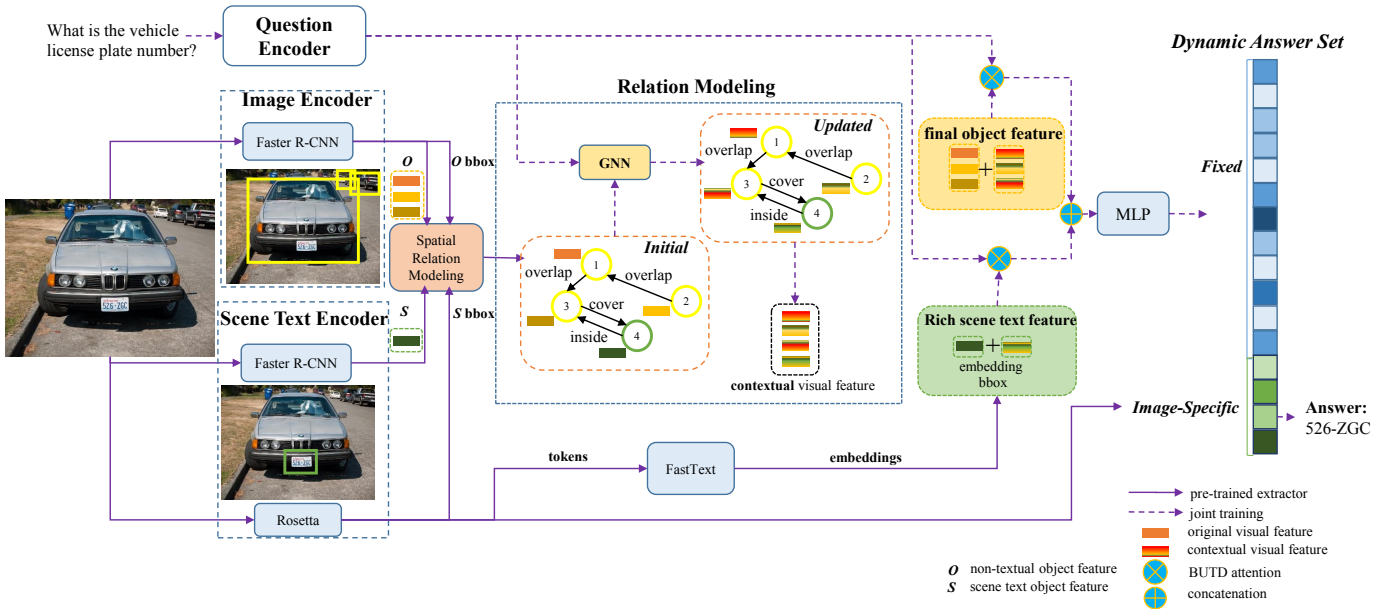


Fig. 3. Details about our MCG model which contains question encoder and image encoders. Question encoder extracts Fast-Text features and image encoder extracts features of non-textual objects as well as scene texts through aggregated results from question-contextual GNN. Various multi-modal features of non-textual objects, scene texts and questions are combined through the multi-modal fusion procedure. A dynamic candidate answer set tailored for each image is then constructed with a shared fixed answer set containing common candidate answers shared across all images, *e.g.*, Dog, Car, and an image-specific candidate answer set containing scene texts appearing in the target image, *e.g.*,the licence plate numbers. The final answer is predicted from the dynamic candidate answer set.

indicate the spatial relationships between objects. We define 11 categories of spatial relationships to describe the relative positions of objects [35]. For example, the relationships between $object_i$ and $object_j$ could be $\langle object_i - inside - object_j \rangle$, $\langle object_i - cover - object_j \rangle$, $\langle object_i - left - object_j \rangle$ (More relationships see Fig. 2 and Tab. I). Moreover, unlike Norcliffe *et.al.* [22] who consider the relationship between every two objects, we only consider those relationships whose head objects and tail objects are not far away from each other because the spatial relationships between objects would tend to be weak if their spatial distances are large. In Summary, $\boldsymbol{E} = \{e_{ij} | \boldsymbol{v_i}, \boldsymbol{v}_j \in \boldsymbol{V}, R_e(\boldsymbol{v_i}, \boldsymbol{v_j}) = 1 \wedge DIST(\boldsymbol{v_i}, \boldsymbol{v_j}) < \delta\}$ where R_e is the relation indication function for relation e and $DIST$ is the spatial distance calculation function.

2) Contextual Node Feature: Context adaptation.

Questions with no doubt provide significant guidance for extracting effective visual features in images and analyzing their relations in different contexts. Hence taking questions into consideration when conducting node feature aggregations in our contextual GNN model will help to increase the accuracies of answering questions. It is desirable that the semantic information in questions can exclusively and contextually guide the GNN aggregation process, so that those nodes having high relevance with the questions are assigned with high weights. Given a graph $\boldsymbol{G} = (\boldsymbol{V}, \boldsymbol{E})$ and question $\boldsymbol{q}$, we design a mechanism to obtain a contextual representation for each node with the target question $\boldsymbol{q}$ as the context:

$$\boldsymbol{v}_i^{(q)} = \sigma\left(\boldsymbol{q} \cdot \boldsymbol{W_q} \boldsymbol{v}_i\right), \ i = 1, 2, \cdots, K + M, \tag{3}$$

where $\boldsymbol{W_q} \in \mathbb{R}^{d_h \times d_v}$ is the context-adapted projection matrix, $\sigma(\cdot)$ is a non-linear function such as ReLU.

3) Graph Attention: Multi-head Attention.

Given the fact that the influences of neighbor nodes on the target node may origin from multiple aspects, we incorporate the multi-head attention mechanism [30] into the neighborhood feature aggregation process of our contextual GNN. Given Graph $\boldsymbol{G} = (\boldsymbol{V}, \boldsymbol{E})$, the feature aggregation for layer h of a vanilla GNN containing several layers can be written as follows:

$$\boldsymbol{v}_i^{h+1} = \sigma\left(\sum_{j \in \mathcal{N}_i} \alpha_{ij} \cdot \boldsymbol{W_h} \boldsymbol{v}_j^h\right), \tag{4}$$

where $\boldsymbol{W}_h \in \mathbb{R}^{d_{h+1} \times d_h}$ is the learnable projection matrix and α_{ij} is the attention weight. To incorporate the multi-head attention mechanism, we set the attention weight for the l-th head α_{ij}^l as:

$$\alpha_{ij}^l = \frac{\exp\left(\left(\boldsymbol{U}^l \boldsymbol{v}_i^h\right)^\top \cdot \boldsymbol{V}^l \boldsymbol{v}_j^h\right)}{\sum_{j \in \mathcal{N}_i} \exp\left(\left(\boldsymbol{U}^l \boldsymbol{v}_i^h\right)^\top \cdot \boldsymbol{V}^l \boldsymbol{v}_j^h\right)}, \tag{5}$$

where $\boldsymbol{U}^l$ and $\boldsymbol{V}^l$ are the learnable projection matrices, and $(\cdot)$ refers to the scaled dot production used to compute the inner product between $\boldsymbol{U}^l v_i^h$ and $\boldsymbol{V}^l v_j^h$ to measure their similarity. The denominator is used to normalize the attention weights. Totally L independent attention heads are executed and the outputs from these heads are concatenated, resulting in the

following aggregated feature representation for node i in the $h+1$ layer:

$$\boldsymbol{v}_i^{h+1} = \Big\Vert_{l=1}^{L} \sigma\left(\sum_{j\in\mathcal{N}_i} \alpha_{ij}^{l} \cdot \boldsymbol{W_h}^{l} \boldsymbol{v}_j^{h}\right), \tag{6}$$

where $||$ is the concatenation operation. Suppose the graph attention module has K layers, we then rewrite the K-th layer aggregated node feature $\{\boldsymbol{v}_i^K\}_{i=1}^{K+M}$ as $\{\boldsymbol{v}_i^*\}_{i=1}^{K+M}$ to represent the final results from our multi-head graph attention operation. Also, we decouple the non-textual object features and scene text features from $\{\boldsymbol{v}_i^*\}_{i=1}^{K+M}$ and denote them as $\{\boldsymbol{v}_i^{(o),*}\}_{i=1}^{K}$ and $\{\boldsymbol{v}_i^{(s),*}\}_{i=1}^{M}$, respectively.

D. Multi-modal Fusion and Dynamic Prediction Component

Besides the visual features and their contextual relations, semantic meanings carried in scene texts and their positions are also important for answering questions related to scene texts. Instead of modeling the semantic contents separately, we propose to combine them with with positional features and contextual aggregated features together in a multi-modal way to further improve the prediction accuracy.

In detail, given the textual feature of the token sequence in scene text $\boldsymbol{t}_i$, the raw visual feature of scene text $\boldsymbol{v}_i^{(s)}$, the contextual aggregated visual feature $\boldsymbol{v}_i^{(s),*}$ and the positional feature $\boldsymbol{b}_i^{(s)}$ obtained from bounding box of the scene text, we concatenate these feature to form the final representation for scene text s_i as follows:

$$\boldsymbol{s}_i = \left[\boldsymbol{t}_i \;\Big\Vert\; \left(\boldsymbol{v}_i^{(s)} + \boldsymbol{v}_i^{(s),*}\right) \;\Big\Vert\; \boldsymbol{b}_i\right], \quad \text{for } i = 1, \ldots, M, \tag{7}$$

where $[\cdot\|\cdot]$ means concatenation and M is the number of scene texts in the image. Specially, to prevent the dimension from becoming uncontrollable, we employ a residual connection representation $(\boldsymbol{v}_i + \boldsymbol{v}_i^*)$ for visual features.

To summarize, the representations for non-textual objects are obtained from our contextual GNN and the representations for scene texts contain textual information, visual information, influencing information obtained from contextual GNN and positional information from the bounding boxes of scene texts.

Given the non-textual object $\boldsymbol{v}^{(o)}$ and scene text $\boldsymbol{s}$, we pass them to the effective top-down attention module [1], denoted as $TD(\cdot)$, to obtain their final representations:

$$\boldsymbol{O} = \left[TD\left(\{\boldsymbol{v}_i^{(o)}\}_{i=1}^{K}, \boldsymbol{q}\right) \;\Big\Vert\; TD\left(\{\boldsymbol{v}_i^{(o),*}\}_{i=1}^{K}, \boldsymbol{q}\right)\right], \tag{8}$$

$$\boldsymbol{S} = TD\left(\{\boldsymbol{s}\}_{i=1}^{M}, \boldsymbol{q}\right), \tag{9}$$

where $\boldsymbol{O}$ and $\boldsymbol{S}$ denote the final representations of non-textual object and scene text respectively.

Note that, the ordering information of scene text token gets lost after passing through the top-down attention module because the features are weighted averaged. To provide the ordering information for the answer predicting procedure, we concatenate the attention weights and the output features from top-down attention module together, allowing the answer predicting module to know the original attention weights for each token in order.

Differing from vanilla VQA problem where the candidate answer set is fixed, the candidate answer set in TextVQA is dynamically changing for different images because different images may have different scene texts that could also be the answer potentially). Therefore, we follow previous work [29] to construct a dynamic candidate answer set for each image. We extend the fixed candidate answer set of size F to a dynamic candidate answer set of size F+M, where the first F slots contain the common candidate answer set shared across all images and the last M slots contain the scene texts appearing in a particular image.

Finally, we learn a 2-layer MLP over the concatenated $\boldsymbol{O}$ and $\boldsymbol{S}$ to generate the prediction of our MCG model for each answer to the question:

$$p = MLP\left([\boldsymbol{O} \parallel \boldsymbol{S}]\right), \tag{10}$$

where p is the binary probability as logits for each answer and binary cross entropy is adopted as the loss function to train the model [1].

IV. Experiments

We evaluate our approach on a challenging dataset TextVQA [29], and our model outperforms previous work on this dataset.

A. Datasets

TextVQA is a new dataset collected in order to address the task of answering questions that requires analyzing scene text in images. TextVQA contains 28,408 images selected from Open Images v3 dataset [16]. The questions require reasoning about the scene text in the image, and there are 10 answers collected for each question. The challenge of TextVQA dataset is that 26,263 (49.2%) answers are unique leading to the answer space a high diversity, including brand names, cities people's names, *etc.*, resulting in the difficulty of having a fixed answer space that is used in the most existing VQA datasets.

B. Implementation Details

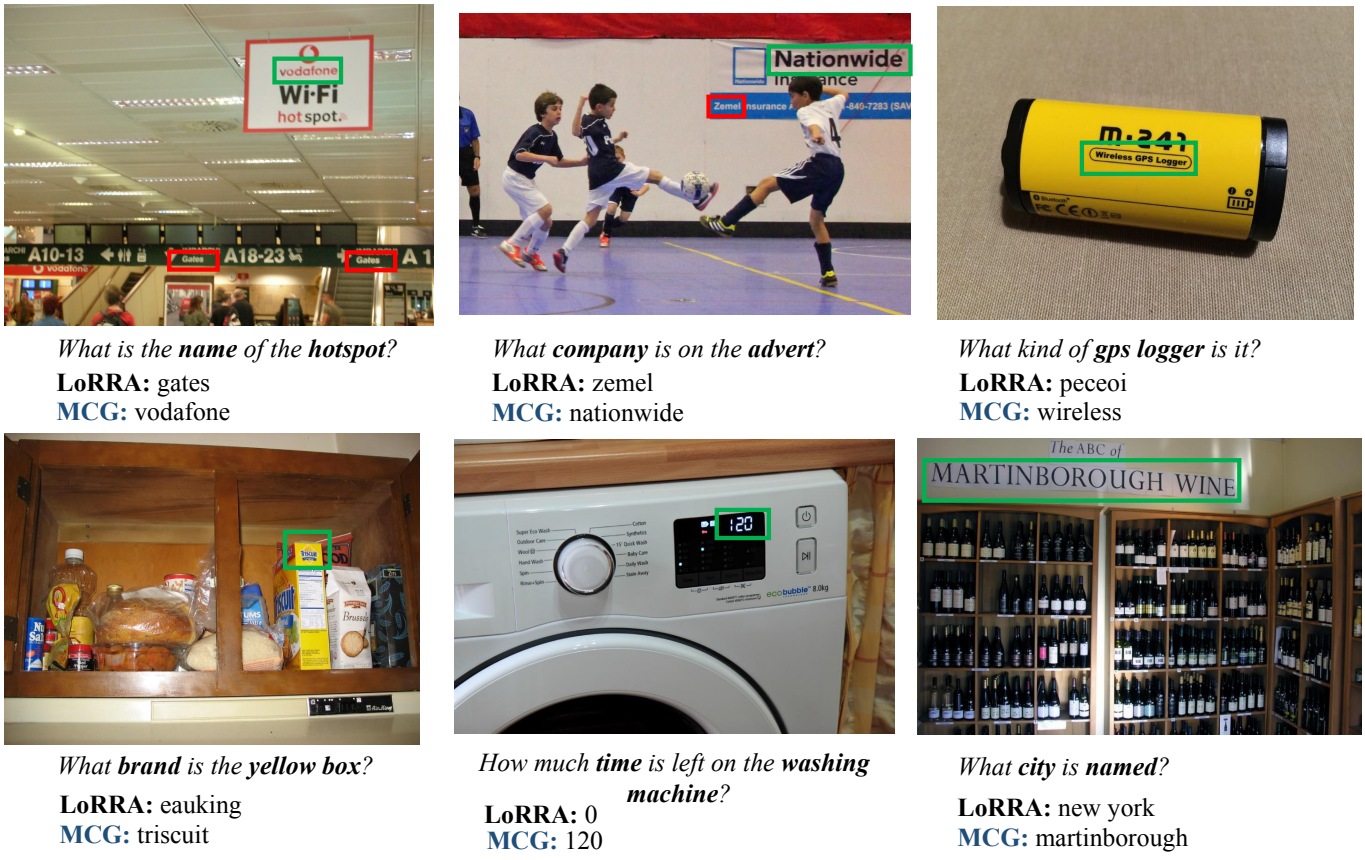
To generate question embedding, every word in the question is tokenized and embedded by a 300-dimensional GloVe [24] word embedding, then the sequence of embedded words is fed into a LSTM with self-attention [37].

For the non-textual object extraction, following [28] and [29], we extract visual features from fc6 layer of an improved Faster R-CNN detector [25] pre-trained on the Visual Genome dataset [15], and we fine-tune the fc7 weights during training [12] to get the 2048-dimension object visual features.

For the scene texts, we use text tokens and bounding boxes extracted by Rosetta OCR system [5] provided in TextVQA dataset. Furthermore, in order to obtain the visual feature of the scene text object, we feed the bounding box of each scene texts into ROI-pooling layer of a pretrained faster RCNN. We also extract FastText [4] embedding feature for each scene text

TABLE II
OVERALL MODEL PERFORMANCE COMPARISION. THE VALIDATION SET ACCURACY (VAL)IS COMPUTED LOCALLY, WHILE THE TEST SET ACCURACY (TEST) IS OBTAINED THROUGH THE ONLINE JUDGING SYSTEM.

Model	Object Combine	OCR Combine	No.of GNN Layer	Rich OCR Feature	Acc. on Val	Acc. on Test
LoRRA [29]	–	–	–	–	26.56%	27.63%
MCG(max-pooling)	–	–	1	yes	17.85%	17.34%
MCG	residual	residual	1	yes	29.29%	29.29%
MCG	2 att.	concat.	1	yes	27.68%	27.91%
MCG	2 att.	residual	1	no	27.81%	27.98%
MCG	2 att.	residual	2	yes	28.71%	29.06%
MCG	2 att.	residual	1	yes	**29.40%**	**29.61%**



*What is the **name** of the **hotspot**?*
LoRRA: gates
MCG: vodafone

*What **company** is on the **advert**?*
LoRRA: zemel
MCG: nationwide

*What kind of **gps logger** is it?*
LoRRA: peceoi
MCG: wireless

*What **brand** is the **yellow box**?*
LoRRA: eauking
MCG: triscuit

*How much **time** is left on the **washing machine**?*
LoRRA: 0
MCG: 120

*What **city** is **named**?*
LoRRA: new york
MCG: martinborough

Fig. 4. Qualitative examples from our MCG model on TextVQA test set. We circle the **resulting visual clue** of OCR tokens predicted by our MCG model in **green**, and predicted by LoRRA [29] in **red**. Comparing to the previous model LoRRA, MCG model is better-preformed in building the relation among question, visual clue of image and scene text tokens.

token. In the contextual GNN process, we employ multi-head attention with 16 heads, and the dimension of each hidden layer of GNN is 2048.

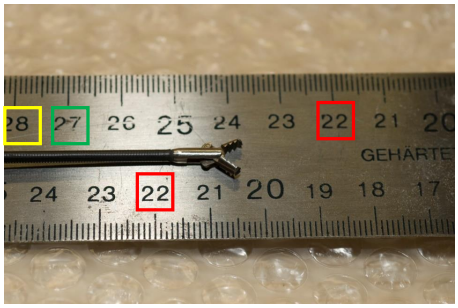
Our model is implemented based on PyTorch [23] and Pythia [28] framework. During training, we use Adamax optimizer with mini-batch size as 128. We deploy the warm-up strategy [9] and the initial learning rate is 0.002, then we linearly increase it at each iteration till it reaches 0.01 at iteration 1000. We also use staircase learning rate schedule, where we reduce the learning rate by a factor 0.1 at the 14000-th and 19000-th iterations.

C. Result and Ablation Study

The evaluation result of our MCG model is shown in Tab. II, and our model outperforms previous work LoRRA[29] by about 2% accuracy improvement on the online test set of TextVQA. The success sample and failure sample are demonstrated in Fig. 4 and Fig. 5 respectively. By analyzing and comparing success samples, we can draw the conclusion that our MCG model is better-performed than LoRRA [29] in modeling the relation among question, objects visual feature and OCR tokens. Particularly, MCG is more capable of spotting visual clue of corresponding scene text on the basis of question demand. For example, for one of the questions

*How many **way stop** is this **sign** for?*

LoRRA: 3
MCG: all
Human: 4



*What is the **largest number** on the **top row** of this **ruler**?*

LoRRA: 22
MCG: 27
Human: 28

*What does it say in **blue**?*

LoRRA: kullik
MCG: ilihakvik
Human: kullik ilihakvik

Fig. 5. Faulty examples from our MCG model on TextVQA test set. We further circle the vital visual clue for predicting the correct answers in **yellow**. Excluding the effect of imperfections of scene text token extractor, we can infer that previous work LoRRA and our MCG model are weak in 2 aspects: 1) do not have the ability in predicting answers that require more than 1 token. 2) unable to split extracted OCR tokens according to the semantic clue given in question (*e.g.* **4-WAY** in sample 1).

in Fig. 4 "What city is named?", MCG is able to relate the key word "city" with the specific visual clue and then predict the correct answer "martinborough" rather than select a city name from the fixed answer set. For the failure samples, if we exclude the negative effect of imperfections of extracted OCR tokens, it can be deduced that MCG and LoRRA are weak in 2 aspects: 1) can only produce one OCR token as the predicted answer. 2) are not capable of splitting extracted OCR tokens on the demand of semantic clues given in question.

In Tab. II, we also compare four ablated instances of MCG with is complete form. To analyze the performance of each instance, we compare the predicting accuracy on test set of TextVQA.

1) Ablations on object feature concatenation: Firstly, we validate the effectiveness of the concatenation strategy of combining original and attented visual feature of non-textual object. There are two relatively effective combination strategy, one is as mentioned in Sec. III-D, and the alternative one introduces residual between original and contextual visual feature. The comparison between line 4 and line 8 shows a gain of +0.32% for concatenation rather than residual. Secondly, we validate the effectiveness of the residual strategy of combining scene text object visual feature before and after GNN. In contrast, between line 5 and line 8, we see a gain of approximately +1.7% for the residual strategy. Hence we can conclude that residual combination of scene text object is critical for MCG model.

2) Ablations on fusion between GNN output and question: To verify the significant improvement on MCG's performance of applying question attention over GNN outputs. We cancel the top-down attention for GNN output nodes, instead, similar to [22], we apply max-pooling on GNN output nodes, then multiply the max-pooling result with question to make question involved. To enrich the representation of scene text token in this condition, we concatenate the visual feature and FastText [4] embedding of OCR tokens. In the comparison between line3 and line8, we observe a huge gain of +12.27% for applying question attention over the output nodes of GNN. Besides, we also ablate the influenced MCG performance on number of GNN hidden layers. Comparing line 7 and line 8, there is a +0.55% gain for one-hidden-layer GNN.

3) Ablations on rich representation of scene text: In order to validate the effectiveness of diverse feature representation on scene text, which is an important strategy in our model, similar to LoRRA [29], we remove the feature other than FastText [4] embedding, and the result is presented in line 6. Between line6 and line8, we can observe a relatively large gain of +1.63% for rich representation.

V. Conclusion

In this paper, we study the problem of TextVQA. We propose a multi-modal contextual graph neural network (MCG) model which utilizes a much richer multi-modal representation for scene texts containing visual, textual, positional and relational features to make full use of the available information in the given images. To capture the relational features, we construct a relation graph for non-textual objects and scene texts and employ a question-contextual graph neural network (GNN) to generate the contextual aggregated features for both non-textual objects and scene texts. Our extensive experiments demonstrate the efficacy of the proposed MCG model against existing literature.

Acknowledgement

This work is supported by the National Key Research and Development Program of China (No. 2020AAA0107800, 2018AAA0102000) National Natural Science Foundation of China Major Project (No. U1611461).

REFERENCES

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018.

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[3] Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. Scene text visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4291–4301, 2019.

[4] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

[5] Fedor Borisyuk, Albert Gordo, and Viswanath Sivakumar. Rosetta: Large scale system for text detection and recognition in images. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 71–79, 2018.

[6] Xuguang Duan, Wenbing Huang, Chuang Gan, Jingdong Wang, Wenwu Zhu, and Junzhou Huang. Weakly supervised dense event captioning in videos. In *Advances in Neural Information Processing Systems*, pages 3059–3069, 2018.

[7] 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 preprint arXiv:1606.01847*, 2016.

[8] Lianli Gao, Zhao Guo, Hanwang Zhang, Xing Xu, and Heng Tao Shen. Video captioning with attention-based lstm and semantic consistency. *IEEE Transactions on Multimedia*, 19(9):2045–2055, 2017.

[9] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

[10] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 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*, pages 6904–6913, 2017.

[11] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[12] Yu Jiang, Vivek Natarajan, Xinlei Chen, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. Pythia v0. 1: the winning entry to the vqa challenge 2018. *arXiv preprint arXiv:1807.09956*, 2018.

[13] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2901–2910, 2017.

[14] Jin-Hwa Kim, Kyoung-Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. Hadamard product for low-rank bilinear pooling. *arXiv preprint arXiv:1610.04325*, 2016.

[15] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017.

[16] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Tom Duerig, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *arXiv preprint arXiv:1811.00982*, 2018.

[17] Linjie Li, Zhe Gan, Yu Cheng, and Jingjing Liu. Relation-aware graph attention network for visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10313–10322, 2019.

[18] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. In *Advances in neural information processing systems*, pages 289–297, 2016.

[19] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *Proceedings of the International Conference on Document Analysis and Recognition*, volume 1, page 5, 2019.

[20] Medhini Narasimhan, Svetlana Lazebnik, and Alexander Schwing. Out of the box: Reasoning with graph convolution nets for factual visual question answering. In *Advances in neural information processing systems*, pages 2654–2665, 2018.

[21] Duy-Kien Nguyen and Takayuki Okatani. Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6087–6096, 2018.

[22] Will Norcliffe-Brown, Stathis Vafeias, and Sarah Parisot. Learning conditioned graph structures for interpretable visual question answering. In *Advances in Neural Information Processing Systems*, pages 8334–8343, 2018.

[23] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[24] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[25] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[26] Sanket Shah, Anand Mishra, Naganand Yadati, and Partha Pratim Talukdar. Kvqa: Knowledge-aware visual question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8876–8884, 2019.

[27] Ajeet Kumar Singh, Anand Mishra, Shashank Shekhar, and Anirban Chakraborty. From strings to things: Knowledge-enabled vqa model that can read and reason. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4602–4612, 2019.

[28] Amanpreet Singh, Vivek Natarajan, Yu Jiang, Xinlei Chen, Meet Shah, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. Pythia-a platform for vision & language research. In *SysML Workshop, NeurIPS 2019*, 2018.

[29] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8317–8326, 2019.

[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[31] Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton van den Hengel. Fvqa: Fact-based visual question answering. *IEEE transactions on pattern analysis and machine intelligence*, 40(10):2413–2427, 2018.

[32] Peng Wang, Qi Wu, Chunhua Shen, Anton van den Hengel, and Anthony Dick. Explicit knowledge-based reasoning for visual question answering. *arXiv preprint arXiv:1511.02570*, 2015.

[33] Jason Weston, Sumit Chopra, and Antoine Bordes. Memory networks. *arXiv preprint arXiv:1410.3916*, 2014.

[34] Peixi Xiong, Huayi Zhan, Xin Wang, Baivab Sinha, and Ying Wu. Visual query answering by entity-attribute graph matching and reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8357–8366, 2019.

[35] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring visual relationship for image captioning. In *Proceedings of the European conference on computer vision (ECCV)*, pages 684–699, 2018.

[36] Haonan Yu, Jiang Wang, Zhiheng Huang, Yi Yang, and Wei Xu. Video paragraph captioning using hierarchical recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4584–4593, 2016.

[37] Zhou Yu, Jun Yu, Chenchao Xiang, Jianping Fan, and Dacheng Tao. Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. *IEEE transactions on neural networks and learning systems*, 29(12):5947–5959, 2018.

[38] Yuyu Zhang, Hanjun Dai, Zornitsa Kozareva, Alexander J Smola, and Le Song. Variational reasoning for question answering with knowledge graph. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.