Multimodal Multihop Source Retrieval for Web Question Answering

Nikhil Yadala [*1] **Paritosh Mittal** [*1] **Saloni Mittal** [*1] **Shubham Gupta** [*1]

Abstract

Natural answers to general questions are often generated by aggregating information from multiple sources. However, most question answering approaches assume a single source of information (usually images or text). This is a weak assumption as information is often sparse and scattered. Next, the source of information can also be vision, text, speech, or any other modality. In this work, we move away from the simplistic assumptions of VQA and explore methods that can effectively capture the multimodal and multihop aspects of information retrieval. Specifically, we analyze the task of information source selection through the lens of Graph Convolution Neural Networks. We propose three independent methods which explore different graph arrangements and weighted connections across nodes. Our experiments and corresponding analysis highlight the prowess of graph-based methods in performing multihop reasoning even with primitive representations of input modalities. We also contrast our approach with existing baselines and transformer-based methods which by design fail to perform multihop reasoning for source selection task.

1. Introduction

Humans can learn to gather information from different modalities and collate them effectively to answer complex questions. However, current AI agents cannot still effectively perform this task. The advent of deep learning and large-scale data processing has led to some progress in combining inputs from two modalities. The success of neural models on Visual Question Answering (VQA) (Antol et al., 2015; Goyal et al., 2017a) is a testament to this. However, the problem setting here ensures that sufficient information is available in visual and textual cues. This is not a very strong approximation for real-world applications.

Figure 1. Top: Sample query; Mid: Possible Sources; Bottom: Desired response

In the real-world, the information is often scattered across multiple sources. Also, it is possible that to answer certain questions, we need context from multiple sources. The domain of multihop reasoning aims to tackle this problem. Here the AI agent is tasked to gather information from different sources and use the collated information to solve downstream tasks such as QA. A challenging and much more natural extension to multihop QA is multimodal multihop QA. Let us take the example of web search. Since the web has information in the form of text, vision, speech, etc. any query answering solution needs to be multimodal. On top of that complete information about a query is rarely found in a single source and an ideal system should learn to aggregate and compile information based on reasoning. Key technical challenges in this problem include (1) Extracting information from different modalities; (2) Selecting the relevant sources (from a set of sources); (3) Aggregating information from relevant sources; (4) producing a natural answer to the question based on the aggregated information. In this project, our team (mmmlX) aims to target the problem of multimodal multihop source retrieval (points 1&2 from before) as shown in Fig. 1. Specifically, we aim to develop an AI system which for a given text question, can select relevant sources of information required to generate a correct answer.

Graph Convolutional Networks (GCNs) are designed to share information across different nodes of a graph and effectively utilize them to make decisions. They are hence well suited for the task of multihop reasoning. However, to the best of our knowledge, there is no existing work that

[*]Equal contribution [1]Carnegie Mellon University. Correspondence to: Shubham Gupta <shubham2@andrew.cmu.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

aims to explore GCNs for solving multimodal multihop source retrieval. In this work, we propose three independent approaches each one designed to explore different aspects of GCNs and showcase their impact on our specific task. Through this work, we demonstrate the comparable performance of GCNs to complex transformer-based baselines (Chang et al., 2021) even with sub-optimal input representations. We suspect this is critically due to the inductive biases of GCNs making them well suited for information aggregation and multimodal retrieval tasks.

The primary contributions of this work are as follows:

- A novel graph-based framework to solve multimodal and multihop source retrieval that can scale to open-domain question-answering in the wild.
- Generate question-conditioned node representations of sources that are informed by the information present in other sources.
- Explore various graph architectures for the source retrieval problem.

The remaining part of the report is structured as follows; Sec. 2 includes the relevant past work; Sec. 3 introduces the formal problem formulation; Sec. 4 explains our three proposed approaches and their significance for source retrieval task. Sec. 5 & 6 provides details about the dataset, baseline methods, experiments conducted and significant insights from them. Sec. 7 finally concludes our project.

We will use this Github repository for the project: `https://github.com/shubham-gupta-iitr/mmmlX`

2. Related Work

2.1. multimodal visual Q/A

The seminal work of (Antol et al., 2015) released a large-scale dataset for Visual Question Answering. This work extracted uni-modal deep features and fused (point-wise multiplication) them for answering. (Das et al., 2017) introduced visual dialog task. The aim of this problem is to develop a conversational AI agent which can also process inputs from visual modality. However, there is a gap between the quantitative (reported) performance and the actual ability of models to generate realistic and diverse inputs.

The work from (Murahari et al., 2019) aims to improve the generative ability of models for real-life scenarios. They introduce two competing agents Q-BOT and A-BOT. Q-BOT is trained to ask diverse questions which forces A-BOT to explore a larger state space and jointly answer more informatively. The work of (Goyal et al., 2017a) was aimed to balance the VQA dataset by having almost twice the original image-question pairs. The pairs often contained conflicting question-answers for the same visual cues to prevent overfitting. Interestingly, extensive works still make a weak assumption that all relevant information is often limited to a single image. Although the context needed to answer many real-life questions can span across multiple images.

2.2. Multihop QA

There has been substantial work in recent years on building QA models that can reason over multiple sources of evidence. In 2018, (Yang et al., 2018b) introduced a text-based QA dataset, HotPotQA that required reasoning over multiple supporting documents. They used an RNN-based architecture that could produce "yes"/"no" or span-based answers. This seminal work is still uni-modal in nature.

Recent works have expanded multihop QA for multimodal inputs. The work from (Talmor et al., 2021) proposed MultiModalQA, which uses inputs image, text, and table to answer questions. They propose a multihop decomposition (ImplicitDecomp) method where the answer is iteratively generated by parsing through individual QA modules. However, they generate questions from a fixed template and hence task reduces to filling of missing entries once the template is identified. Hence it cannot scale well for zero-shot generations. MIMOQ (Singh et al., 2021) proposes a QA system that can reason and also respond in multiple modalities. The authors propose a novel multimodal framework called MExBERT (Multimodal Extractive BERT) that uses joint attention over input textual and visual streams for extracting multimodal answers given a question. They observe noticeable improvements when compared with methods that independently extract answers from unimodal QA modules. However, the key contribution of MIMOQ is still to generate multimodal answers and not information aggregation across modalities.

2.3. Cross modality representations

Significant advances in Visual QA and multimodal QA can be attributed to the advances in Transformer-based methods initially proposed by (Vaswani et al., 2017) and their large-scale pre-training methods. There are two major directions of pre-training: (a) Parallel streams of encoders one for each modality followed by fusion (Tan & Bansal, 2019)(Lu et al., 2019); (b) Unified encoder-decoder representations that can take both the language or image modalities (Zhou et al., 2019). (Anderson et al., 2017) proposed a new form of representation for images. They extract objects from the image and represent an image as a set of embeddings for the Regions of Interest (ROI). The spatial positional features (bounding box locations) are also added to the encoding. This representation is semantically closer to language representations. Hence refining these models and ways of

representations for multihop reasoning is expected to work well on WebQA (Chang et al., 2021) benchmark.

2.4. Graph Convolution Neural Networks

In recent years, there have been several convolutional neural networks for learning over graphs (Kipf & Welling, 2017; Niepert et al., 2016; Defferrard et al., 2016; Bruna et al., 2014). However, most of these approaches focused on whole graph classification problems and were designed specifically for semi-supervised learning objectives. However, the task of multihop multimodal source selection is essentially a node classification problem with dense labels. More recently, GraphSAGE (Hamilton et al., 2017) introduced a novel way of inductively learning node representations for large graphs. As opposed to conventional approaches, this work could scale to unseen graphs during inference making it suitable for our task. Inspired by the success of LSTMs, (Bresson & Laurent, 2018) introduced the concept of residual gated GCNs as a way to control message passing over edges. We leverage from these approaches to adopt GCNs for source retrieval task.

To the best of our knowledge, this is the first dataset that introduces open-domain question-answering in a multimodal and multihop setting. Success on WEBQA requires a system to retrieve relevant multimodal sources first and aggregate information from text and vision modality. Our approach significantly differs from all the prior work in this domain, where we leverage GNNs to solve the problem of multimodal retrieval and ranking. One major limitation of transformer-based models is that they cannot process all the sources together while making a decision during retrieval as the input length is limited. In an open-domain QA system that typically has hundreds of candidate sources to choose from, this problem is amplified manifold. Our approach solves this fundamental problem as it can leverage information from all the sources at the same time.

3. Problem Statement

For multimodal multihop QA the problem is defined such that the QA system gets a text question Q and a set of sources $s_i \in \mathcal{S}$ as input. It is important to note that source s_i can be either an image $\mathcal{I}_i$ or a text snippet $\mathcal{T}_i$ such that $\mathcal{S} \subseteq \mathcal{I} \cup \mathcal{T}$. The benchmark in (Chang et al., 2021) divides the problem into two separate tasks:

Task (A) is a source separation task. Here, for a given question Q system has to divide the sources $\mathcal{S}_Q$ into positive and distractors $(\mathcal{S}_Q^+, \mathcal{S}_Q^-)$. Hence formally the problem is defined as :

$$\mathcal{S}_Q^+ = f_{source}(Q, \mathcal{S}_Q|\phi) \quad \text{where } \mathcal{S}_Q^+ \subseteq \mathcal{S}_Q \subseteq \mathcal{S} \qquad (1)$$

Task (B) is the Question Answering task. Here for a given Q and $\mathcal{S}_Q^+$ the task is to learn a function to generate A_{pred} such that

$$A_{pred} = f_{qa}(Q, \mathcal{S}_Q^+|\theta) \qquad (2)$$

A_{pred} is the generated answer (not classification). The functions f_{qa} and f_{source} are parameterized by ϕ and θ which are the weights of neural models. This project primarily focus on Task (A) i.e. the problem of multimodal multihop source retrieval.

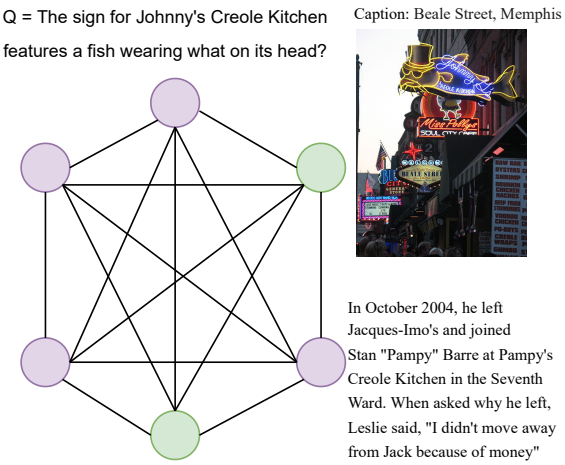

Figure 2. Each purple node indicates a negative source while green indicates a positive source. The example indicates the multimodal heterogeneous nodes on which we do classification

4. Proposed Approach

The problem of source selection critically depends on aggregating information from multiple sources and utilizing this knowledge to make informed decisions. Graph Convolution neural Networks (GCNs) are designed to effectively share information within nodes and are hence well suited for multihop reasoning. In this work, we leverage the inductive biases of GCNs for multimodal multihop source retrieval. Specifically, graphs learn to make decision on information stored in *nodes*. The information is shared by *edges* between nodes. We design three different and innovative approaches which explore different aspects of GCNs for this task.

4.1. Dense Super-Node Graph

For a specific question Q, the WebQnA dataset (Chang et al., 2021) contains a set of sources $\mathcal{S}_\mathcal{Q}$ which can be either from Images ($\mathcal{S}_\mathcal{I}$) or Text ($\mathcal{S}_\mathcal{T}$) modality. We define two domain specific Encoders $E_\mathcal{I}$ and $E_\mathcal{T}$ for images and text modalities, such that $E_\mathcal{I}(\mathcal{S}_{\mathcal{I}_i})$ is the embedding for image i. Similarly we can compute the embedding for Q and $\mathcal{S}_\mathcal{T}$ or captions for Image using $E_\mathcal{T}$.

For this approach, we construct a graph such that all sources

$\mathcal{S}_Q$ are connected to each other. We reinforce the embedding of each source by concatenating question embedding with each source. The following graph hence contains two kinds of nodes:

- **Node I** is an image node whose embedding is a combination of $< E_{\mathcal{T}}(Q), E_{\mathcal{I}}(\mathcal{S}_{\mathcal{I}}), E_{\mathcal{T}}(\mathcal{S}_{\mathcal{I}_{caption}}) >$
- **Node T** is a text node whose embedding is a combination of $< E_{\mathcal{T}}(Q), E_{\mathcal{T}}(\mathcal{S}_{\mathcal{T}}) >$

We denote each such node as a super node because of presence of question embedding $E_{\mathcal{T}}(Q)$ in each node of the graph. Full connections between nodes ensure that information relevant to make decision for a node is available in a single hop.

To learn the graph embedding we leverage from the learning of GraphSAGE (Hamilton et al., 2017). This work proposed a method to inductively learn node embeddings and the method generalizes well to previously unseen data. Specifically we use the following function to aggregate information across nodes:

$$x_i' = W_1 x_i + W_2 \cdot mean_{j \in \mathcal{N}(i)} x_j \tag{3}$$

Here $\mathcal{N}(i)$ denotes the neighborhood of node with index i. The task of source selection can then be reduced to binary node classification for a given graph. We use the binary weighted cross entropy loss function

$$loss_n = -\sum_{c=1}^{2} w_c \cdot y_{n,c} \cdot log\left(\frac{e^{x_{n,c}}}{\sum_{i=1}^{2} e^{x_{n,i}}}\right) \tag{4}$$

where w_c is the class weight and is used to handle class imbalance in the dataset. This approach has quadratic relation between number of edges and nodes. Because there are considerably more negative sources as compared to a positive ones, seemingly irrelevant connections and redundant information flow can impact performance or slow down training.

4.2. Star Graph

This approach is based on dis-entangling question embedding $E_{\mathcal{T}}(Q)$ from the source nodes. Together with this, we adopt a star-like graph node arrangement such that each source node is now only connected to the question node. Specifically, the graph contains three kinds of nodes:

- **Node Q** is a question node whose embedding is $< E_{\mathcal{T}}(Q) >$
- **Node I** is an image node whose embedding is a combination of $< E_{\mathcal{I}}(\mathcal{S}_{\mathcal{I}}), E_{\mathcal{T}}(\mathcal{S}_{\mathcal{I}_{caption}}) >$

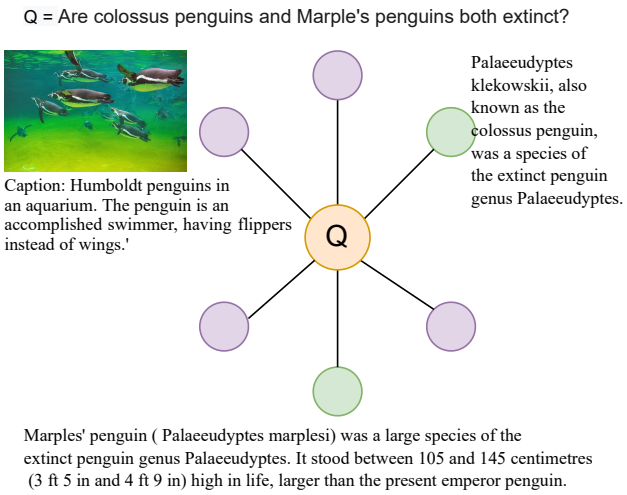

Figure 3. We disentangle the question node Q in yellow. There are considerably less edges in this architecture. This example suggests positive text snippets containing relevant information

- **Node T** is a text node whose embedding is $< E_{\mathcal{T}}(\mathcal{S}_{\mathcal{T}}) >$

Fig 7 illustrates the graph structure wherein the only connections are of type **Node Q** - **Node I** or **Node Q** - **Node T**. Hence there is now a linear relation between the number of nodes and edges. However, for this approach, all relevant information to identify a particular source as positive is within two hops (as opposed to one hop in sec. 4.1). The task of source selection is again reducible to that of node classification. However, **Node Q** does not have any label and hence the loss is only computed for source nodes.

We train this approach using the binary-weighted cross-entropy loss described in eq. 4. Sparse connections enable the model to train faster and greatly reduces a large number of uninformative connections but may introduce a bottleneck as all nodes are only connected to **Node Q**. Like some other GCN algorithms, this approach does not discriminate between edges and hence they (or any associated parameter) do not change during the learning process.

4.3. Gated Graph

The previous approaches use a weighted combination of node embedding from neighborhood $\mathcal{N}(i)$ of a node i as explained in eq. 3. However, these approaches assume an equivariance over edges connecting two specific nodes. Because there are many more negative nodes as compared to positive ones (data imbalance) there can always be an irrelevant flow of information. We leverage from the learning of (Bresson & Laurent, 2018) to use a residual gating of graph edges. Hence this approach proposes to learn an edge

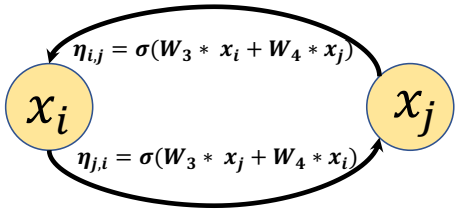

Figure 4. We unfold an undirected edge into two directed edges to highlight the differnce in η.

specific factor $\eta_{i,j}$ to control the impact of each neighbour j on a given node i. More specifically, $\eta_{i,j}$ is like an edge gate with value in range $[0, 1]$. The value is derived from the embedding of nodes i, j using

$$\eta_{i,j} = \sigma(W_3 x_i + W_4 x_j) \tag{5}$$

where σ is the sigmoid function. Hence the message passing function in graph learning now changes to

$$x'_i = W_1 x_i + \sum_{j \in \mathcal{N}(i)} \eta_{i,j} \cdot W_2 x_j \tag{6}$$

where W_1, W_2, W_3, W_4 are the trainable parameters in this approach. It is interesting to note that since an edge is mainly defined by the two nodes it connects hence in this approach the weight ($\eta_{i,j}$) of a particular edge is influenced by the embeddings of the corresponding nodes (x_i, x_j).

Fig. 4 visualizes this approach aimed at eliminating the equivariance of edges and simultaneously ensuring that only relevant sources contribute to source selection. The approach can be extended to both graph architectures defined in sections 4.1, 4.2. The models are trained for the task of node classification is trained using binary-weighted cross-entropy from eq. 4.

5. Experimental Setup

5.1. Dataset

We use the WebQA (Chang et al., 2021) dataset to solve the problem of multimodal multihop source retrieval. The dataset contains input from two different modalities: *Images* and *Text*. Table 1 summarises the number of instances of textual and visual source-based queries in the train, val, and test dataset. Each question Q is text-based and the dataset contains a set of sources $\mathcal{S}_Q$ associated with each question. Each element in the set $\mathcal{S}_Q$ can either be a positive source (with label $+1$) or a negative source (with label 0). Note that the number of sources ($|\mathcal{S}_Q|$) is not fixed and hence is different for each question Q. The questions with a positive source from visual modality are further sub-divided into six categories. Table 2 contains these question categories. We do not specifically use this sub-division while training however we do use them for quantitative analysis of performance.

Modality	Train	Dev	Test
Image	18,954	2,511	3,464
Text	3,464	2,455	4,076

Table 1. Statistics by the modality required to answer the question.

5.2. Input Representation

Our approach is aimed at exploring Graph Convolution Neural Networks to solve the task of multimodal multihop source retrieval. To encode each modality we use modality-specific encoders. For images, this encoder $E_{\mathcal{I}}$ is a ResNet-152 (He et al., 2015) network and hence each image is represented by a $2048d$ vector. For text, this encoder $E_{\mathcal{T}}$ is a sentence level representation from BERT (Devlin et al., 2019). This sentence-level representation ensures each input is represented by a $786d$ vector. We define these features as *primitive features* which is sufficient to demonstrate the efficacy of our GCN-based method. However, we suspect the final number might significantly improve if we switch to more advanced SOTA modality-specific encoders.

5.3. Multimodal baseline Models

The authors from WebQA (Chang et al., 2021) introduced two baselines and their variants for multimodal source retrieval.

5.3.1. Lexical Overlap

This is a trivial baseline that considers the overlap between a given question Q and the snippet/caption to make a decision for a particular course. This baseline returns the Top-2 sources with the highest overlap. Interestingly, the dataset is designed to suppress Lexical Overlap and force neural networks to not overfit on this overlap.

5.3.2. VLP + VinVL

VLP (Zhou et al., 2019) is a pre-trained multimodal transformer well suited for both understanding and generative tasks. The source retrieval baseline is finetuned on the VLP

Question Category	Train	Val
YesNo	6492	828
Number	1859	259
Color	1651	179
Choose	3718	502
Others	4743	669
Shape	491	74

Table 2. Statistics by the Question Category for the visual modality.

Table 3. The table contains the hyper-parameter choices we found to work best for our experiments. Interestingly, similar hyper-parameters worked well for all three approaches in sec. 4.1,4.2 and 4.3

Epochs	Batch	Base LR	LR Scheduler	Sh. Gamma	Optimizer	Loss	Class weight
200	32	0.00002	StepLR	0.9	AdamW	CrossEntropy	10 (+ve)

checkpoints. Text modality inputs are tokenized by Bert-base-cased tokenizers. For images, the baseline extracts ~ 100 region proposals using latest state-of-the-art model VinVL (Zhang et al., 2021). The authors conducted experiments with various image encoders like x101fpn etc. however demonstrated VinVL (Zhang et al., 2021) to work best. In this work, we hence draw comparisons with the baseline's best VLP + VinVL architecture.

For a given question Q, every source $\mathcal{S}_{Q,i} \in \mathcal{S}_Q$ is fed into the baseline model one by one. The input to the model is hence $< [CLS], \mathcal{S}_{Q,i}, [SEP], Q, [SEP] >$ and the model predicts the probability of it being a positive source. This approach hence makes a critical assumption that prediction over a source is independent of other sources in the dataset. We find this to be a weak assumption and not in the spirit of multihop reasoning which the dataset (Chang et al., 2021) critically demands.

5.4. Evaluation Metrics

The performance for source retrieval is measured using the F1-score.

5.4.1. F1-Score

There is a severe imbalance in the dataset and hence accuracy of source selection is not the best metric. For example, the base model can achieve $\sim 92\%$ accuracy by predicting every source as negative. Hence we use the F1-score to measure performance. F1-score is measured as a harmonic mean of Precision and Recall.

$$F1score = 2 * \frac{Pr * Recall}{Pr + Recall} \tag{7}$$

where Precision (Pr) is defined as $\frac{TP}{TP+FP}$ and $Recall$ is defined as $\frac{TP}{TP+FN}$. Further, TP are the true positives, FP are false positives and FN are false negatives. F-score enables us to measure of well the model is in predicting positive and negative sources even with imbalance.

5.5. Experiment details

This section explores the hyper-parameters and model architectures used to realize the above three approaches described in sec. 4. Table. 3 contains the final learning rates, batch sizes, and other hyper-parameters that worked for our experiments. It is important to note that we conducted several experiments with different hyper-parameters and found these as the optimal ones. For example, we found that Graph models do not converge when we use larger batch sizes of > 64. We also found these models to be sensitive to learning rate (lr) choices and found sub-optimal convergence with lrs like $\{0.001, 0.0001, 0.0004, \cdots\}$. Interestingly, when compared with FocalLoss we found weighted cross-entropy with $+10$ weight for positive source and $+1$ for negative sources to work better.

Layers	Input dim	Output dim
GraphConv	-1	2048
GraphConv	2048	1024
GraphConv	1024	512
GraphConv	512	256
GraphConv	256	128
Linear	128	128
ReLU	128	128
Linear	128	64
ReLU	64	64
Linear	64	2

Table 4. The following table contains the model architecture used for our experiments. **GraphConv is a placeholder** which is replaced by SAGEConv (Hamilton et al., 2017) for approaches in sec. 4.1,4.2 and by ResGatedGraphConv (Bresson & Laurent, 2018) for approach in sec. 4.3. -1 indicates lazy initialization and helps accommodate heterogeneous graphs like ours.

Table 4 contains our common model architecture for the experiments. Each of our approaches assumes that edges between nodes are undirected. It is important to note that for approaches in sec. 4.1, 4.2 we change the node arrangements while neural model remains the same. For approach in sec. 4.3 we modify the neural network to use $\eta_{i,j}$ and use them for message passing. Hence in sec. 4.3 the GraphConv layer changes accordingly.

6. Results and Discussion

In this section, we will present the results of qualitative and quantitative analysis which compares and contrasts our proposed approaches with each other and with existing baselines.

6.1. Quantitative Results

Table 5 showcases the F1 scores of our proposed approaches on validation data from WebQnA. The table also compares these numbers to that from SOTA baselines (Chang et al.,

Model	Image	Text	Combined
Lexical Overlap	44.83	33.78	-
VLP-VinVL	68.13	69.48	68.86
Dense SuperNode (ours)	**67.15**	59.74	62.81
Star Graph (ours)	66.62	60.73	63.15
SuperNode Gated Graph (ours)	49.32	57.97	54.48
Star Gated Graph (ours)	66.89	**60.87**	**63.44**

Table 5. F1-score comparison of baselines with our methods.

2021). The first major observation is that all our proposed methods outperform the trivial Lexical Overlap baseline indicating that they do in fact learn something meaningful. While it is true that the VLP-VinVL baseline has better F1 scores across the board, we wish to stress here the complexity and resource requirements for this baseline.

For example, our graph-based approaches are trained from scratch, whereas the VLP-VinVL is heavily pre-trained with several multimodal task objectives on billions of image-text pairs. It is is also finetuned on VQA and Image captioning on MSCOCO captions (Chen et al., 2015), VQA 2.0 (Goyal et al., 2017b) and Flickr30k (Yang et al., 2018a). Moreover, VinVL (Zhang et al., 2021) pre-training focuses on improving the object-centric visual representations that can be used by the downstream VLP-based multimodal fusion models. It is interesting to see even with much simpler input representations, the graph-based approach has comparable performance to SOTA due to inherent 'multihop' reasoning ability.

In table 5 we observe that the dense version of Gated Graph does not provide comparable results. We believe that dense connections and limited data make learning optimal parameters difficult, and further hyperparameter tuning can lead to improvements. More specifically for image modality, we see the Dense SuperNode variant yields the best performance. However, the performance suffers in text modality, this could be because while text snippets are very long we only consider snippet-level embedding, in the interest of time and computing. We expect significant gains if we switch to word-level embeddings.

Figure 5 compares the performances of each of our graph-based approaches across sub-categories in the data. These categories are ordered in the decreasing order of the data distribution in the training set. We observe that the performance of the dense super node graph is comparable or better than the others in the categories with more samples in training data. This could be because the approach has exponentially more edges per source and hence needs to better filter out distracting information (particularly, given that majority of the nodes represent the negative sources). Another interesting observation is that while the *Number*, *Color* and *Shape* categories have relatively less data but

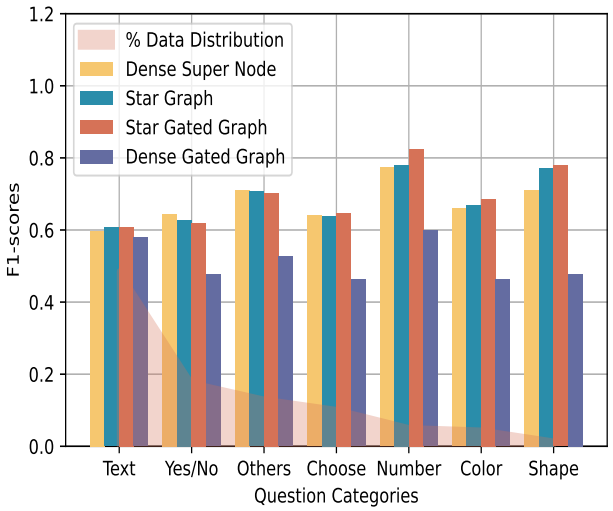

Figure 5. F1 score across question categories against their data distribution in the dataset

our approaches still work much better. This highlights the model's ability to generalize well in low-resource settings and simultaneously highlights the complexity of the problem. Example, *Yes/No*, *Others*, and *Choose* by nature are open-ended and hence sources may be difficult to identify.

We further observe from figure 5 that even with sparse connections in star graph can outperform dense graph in almost all categories. This could be because deeper layers of GCNs enable multihop (two-hop in this case) reasoning and hence do not need exhaustive connections between nodes. Additionally, gated star-graph can outperform vanilla star and dense graphs in most categories. This highlights the importance of edges weights and non-symmetric pair-wise interactions between nodes.

For the next section, we perform a qualitative comparison for Dense SuperNode, StarGraph, and Star Gated Graphs as they are all comparable with baselines and different in their unique ways.

6.2. Qualitative Results

In this section we visualize some key examples from the validation set of WebQnA (Chang et al., 2021) dataset. For each such example we draw meaningful insights that shed light on the performance of our proposed approach. For fair comparison we showcase both success and failure cases which motivate our future work.

6.2.1. Dense Super Node

As shown in Fig. 6 we can see that for the question in row 1, our model is able to capture the correct image-based sources which contain the nearby regions of the "Wonders

Figure 6. Retrieval Results for dense graph approach: Green box indicate predictions which were correct while red indicates incorrect

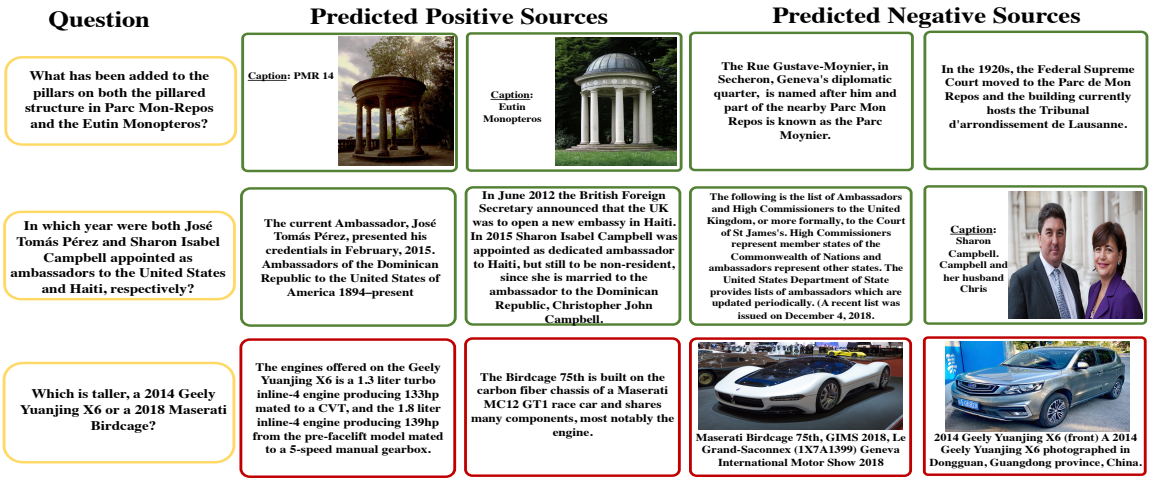

Figure 7. Retrieval Results for star graph approach: Green box indicate predictions which were correct while red indicates incorrect

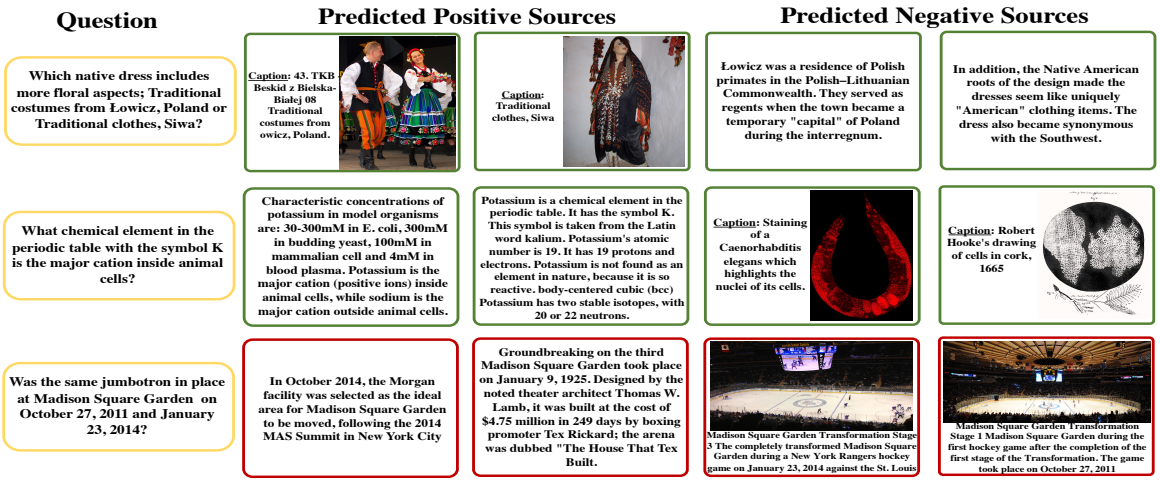

Figure 8. Retrieval Results for gated graph approach: Green box indicate predictions which were correct while red indicates incorrect

of Life" and "Children's Fairyland". This shows that the model has developed the contextual understanding of identifying the object and nearby regions of the object in the image. Even though the text sources contain the similar words "EPCOT", "Fairyland", the model is able to discard these sources. Similarly, for row 2, we can observe that the model is able to identify the correct sources that contain the information regarding the fire of the machine gun and discard the image-based sources even though the caption contains the words "Light machine Gun" and the image is also a gun. This shows us that the model can reason about what is the information being asked and discard the sources which don't contain that information. Row 3 shows a failure case of our model. We can observe that the model is predicting wrong text sources compared to the expected image-based sources. We believe that is happening due to the very high lexical overlap of the question with these text sources. Moreover, we can observe that the question contains some spanish words which may not be interpreted correctly in the embedding space of our BERT model which is trained in english.

6.2.2. Star Graph

As shown in Fig. 7 we can see that for the question in row 1, our model can capture the correct image-based sources which contain the pillars. Even though pillar is not an explicit class in the image-based model, we see that the model can identify the correct image sources. We believe that is partly due to the fact that caption text matches with the query. At the same time, we can also observe that the text facts which contain overlapping words with a query are correctly discarded. This shows that the model has developed the understanding that it requires an image to answer a given count-based query, even when the text might have a high overlap. Similarly, for row 2, we can observe that the model is able to identify the correct sources that contain the information regarding the ambassadors and discard the image-based sources even though the image belongs to the person asked in the query. This shows us that model can reason about what is the information asked regarding the 2 ambassadors and reason about it correctly to identify the positive sources. Row 3 shows a failure case of our model. We see that the model is predicting wrong text sources compared to the expected image-based sources. We believe that since the query is regarding the comparison of size, the model is wrongly associating it with certain measuring quantities such as "liter", in spite of having high lexical overlap such as "Maserati Birdcage", "Geely Yuanjing X6".

6.2.3. Star Gated Graph

As shown in Fig. 8 we can see that for the question in row 1, our model is able to capture the correct sources that contain the images of the clothes of the "Lowicz, Poland" and "Traditional clothes, Siwa". Interestingly, we can observe that even though there was a spelling mistake in the image source ("owicz" compared to "Lowicz"), our model was able to correctly identify it as the positive source. This shows that model has developed the contextual understanding of the given text in captions and questions and is able to correlate that with the image. Even though the text sources contain the similar words "Lowicz", "clothes", "Poland", "dresses" etc, the model is able to discard these sources. Similarly, for row 2, we can observe that the model is able to identify the correct sources that contain the information regarding the chemical element in the periodic table and discard the image-based sources even though the caption contains the words "cell" and image is also representing a cell. This shows us that model can reason about what is the information asked regarding the chemical element and it is able to combine the information in both positive sources to infer them as the positive sources. Row 3 shows a failure case of our model. We can observe that the model is predicting wrong text sources compared to the expected image-based sources. We believe that is happening due to the very high lexical overlap of the question with these text sources. Moreover, we can observe that the question contains many numbers in the form of dates. We believe that correlating these dates with the original image is quite difficult based on the primitive representations that we use for text and images.

7. Conclusions and Future Directions

In this report, we proposed novel graph-based methods for the multimodal source retrieval task using simple features. We implement various graph structures such as Dense, Star, etc., and also explore ways to control information flow using pair-wise 'gating' operation over edges. Our approach and experiments are driven by the belief that graph structures are ideally suited for performing multi-hop reasoning which is a critical component for realistic QA systems. As a primary part of this project, we analyze the performance of each of our proposed approaches and highlight their benefits over existing state-of-the-art baselines. We believe that our ideas can be further extended, by using more complex input feature representations or by using attention mechanism with graphs along with bipartite optimization. While this the first exploratory work to estimate the prowess of GCNs in this task, we hope that our work will start a new paradigm of graph-based methods for multimodal multihop source retrieval in the wild problems.

References

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., and Zhang, L. Bottom-up and top-down attention for image captioning and VQA. *CoRR*,

abs/1707.07998, 2017. URL http://arxiv.org/abs/1707.07998.

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., and Parikh, D. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*, 2015.

Bresson, X. and Laurent, T. Residual gated graph convnets, 2018.

Bruna, J., Zaremba, W., Szlam, A., and Lecun, Y. Spectral networks and locally connected networks on graphs. In *International Conference on Learning Representations (ICLR2014), CBLS, April 2014*, 2014.

Chang, Y., Narang, M., Suzuki, H., Cao, G., Gao, J., and Bisk, Y. WebQA: Multihop and Multimodal QA. 2021. URL https://arxiv.org/abs/2109.00590.

Chen, X., Fang, H., Lin, T.-Y., Vedantam, R., Gupta, S., Dollar, P., and Zitnick, C. L. Microsoft coco captions: Data collection and evaluation server, 2015.

Das, A., Kottur, S., Gupta, K., Singh, A., Yadav, D., Moura, J. M., Parikh, D., and Batra, D. Visual Dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Defferrard, M., Bresson, X., and Vandergheynst, P. Convolutional neural networks on graphs with fast localized spectral filtering. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pp. 3844–3852, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., and Parikh, D. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017a.

Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., and Parikh, D. Making the v in vqa matter: Elevating the role of image understanding in visual question answering, 2017b.

Hamilton, W. L., Ying, R., and Leskovec, J. Inductive representation learning on large graphs. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 1025–1035, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.

Kipf, T. N. and Welling, M. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2017.

Lu, J., Batra, D., Parikh, D., and Lee, S. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *CoRR*, abs/1908.02265, 2019. URL http://arxiv.org/abs/1908.02265.

Murahari, V., Chattopadhyay, P., Batra, D., Parikh, D., and Das, A. Improving generative visual dialog by answering diverse questions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

Niepert, M., Ahmed, M., and Kutzkov, K. Learning convolutional neural networks for graphs. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, pp. 2014–2023. JMLR.org, 2016.

Singh, H., Nasery, A., Mehta, D., Agarwal, A., Lamba, J., and Srinivasan, B. V. Mimoqa: Multimodal input multimodal output question answering. In *NAACL*, 2021.

Talmor, A., Yoran, O., Catav, A., Lahav, D., Wang, Y., Asai, A., Ilharco, G., Hajishirzi, H., and Berant, J. Multimodalqa: Complex question answering over text, tables and images. *ArXiv*, abs/2104.06039, 2021.

Tan, H. and Bansal, M. LXMERT: learning cross-modality encoder representations from transformers. *CoRR*, abs/1908.07490, 2019. URL http://arxiv.org/abs/1908.07490.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Yang, X., Tang, K., Zhang, H., and Cai, J. Auto-encoding scene graphs for image captioning, 2018a.

Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W. W., Salakhutdinov, R., and Manning, C. D. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *EMNLP*, 2018b.

Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., Choi, Y., and Gao, J. Vinvl: Revisiting visual representations in vision-language models, 2021.

Zhou, L., Palangi, H., Zhang, L., Hu, H., Corso, J. J., and Gao, J. Unified vision-language pre-training for image captioning and VQA. *CoRR*, abs/1909.11059, 2019. URL http://arxiv.org/abs/1909.11059.