Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart
Pfaffenwaldring 5B
D-70569 Stuttgart

Master Thesis Proposal

Multimodal Feature Fusion via Graph Neural Networks for Visual Dialog

Nishan Chatterjee

Course of Studies: M.Sc. Computational Linguistics
First Examiner: Prof. Dr. Andreas Bulling
Second Examiner: Frau Jun.-Prof. Dr. Carina Silberer
Supervisor: Mohamed Adnen Abdessaied
Beginning of the Thesis: 30.03.2022
End of thesis: 30.09.2022

1 Introduction and Motivation

The task of Visual Dialog requires an agent to hold a meaningful dialog with humans in a natural and conversational language about the content of an image. Or more specifically, given an image, a dialog history, and a question, an AI agent has to ground the question in the image while inferring from context history to answer questions accurately (Das et al., 2016). This task requires reasoning over both the context that has been stated in the Dialog, and a sense of world knowledge that underlies it.

Chen et al. (2021) in their Relation-Aware Graph-over-Graph Network for Visual Dialog showed how using three sequential graphs can result in the highest performance among all previous approaches except pretraining-based models (Wang et al., 2020). More specifically, the three sequential graphs that Chen et al. (2021) use are: (1) a History-graph (H-graph) which aims to capture the co-reference relations in Dialog history, which is then embedded into (2) the History-Aware Question-Graph (Q-Graph) by the use of a History-Aware attention mechanism (Vaswani et al., 2017) to inject semantic information of the history into this question graph, which then embeds this information multimodally to form a Question-Aware Image-Graph (I-graph) to perform the Visual Dialog Task.

This proposal highlights a new method of performing Visual Dialog. It shares similarities in the use of three sequential graphs, but there's no implicit embedding of the bottom sequential Graph layer into its adjacent layer. Rather, this method proposes the use of a master node which captures the co-reference relation of the dialog history for the H-graph, which is then passed up to generate a History-Aware Q-graph, and finally, the master node embedding of the History-Aware Q-graph is passed up to generate the Question-Aware I-Graph. As optional goals, the proposal also highlights other hyperparameter settings of using the master node to control the message passing between the graphs 4.5 for the Visual Dialog challenge task.

2 Related Work

While Pretrained Language Models (LM) have seen a lot of success in many Question Answering (QA) tasks as shown by Liu et al. (2019) and Raffel et al. (2019), LM's do not robustly capture the latent relationships between concepts, which is a key aspect of reasoning (McCoy et al., 2019). One of the workarounds to the problem of relationship modelling in unstructured data can be solved by the use of Knowledge Graphs (KG) like Freebase(Bollacker et al., 2008), Wikidata (Vrandečić and Krötzsch, 2014), and ConceptNet (Speer et al., 2017) that capture external knowledge between entities explicitly using triplets that model the relationship between various entities. Ren et al. (2020) show the significant role KG's can play in structured reasoning and query answering, while, Yasunaga et al. (2021) shows how QA tasks can see improved performance in relevance scoring, where LM's estimate the importance of nodes in a KG relative to a given QA context, and joint reasoning, where the QA context and the KG are connected to form a join graph by the use of Graph Neural Networks compared to simply using LM's.

However, extending these advantages in reasoning abilities of AI agents to general QA, i.e. when questions and answers are expressed in the form of natural language and therefore aren't easily mapped to strict logical queries that require proper integration of the information and constraints that a QA provides along with the knowledge from a KG. While Mihaylov and Frank (2018), Lin et al. (2019), and Feng et al. (2020) showcase a number of ways to leverage both modalities of structured and unstructured information to improve reasoning, the methods typically fuse the modalities in a shallow and non-interactive manner as both information is encoded separately and is fused at the final output step.

These methods showcase their advantages and disadvantages in uni-modal (text) data which is one part of the Visual Dialog Challenge. However, to perform Visual Dialog, an AI system must be capable of reasoning multi-modally, or in this case, over natural language (text) and images. Visual Question Answering (VQA) is a task similar to Visual Dialog without the use of Dialog History to carry out conversation. Recently, VQA has seen the use of Relation-Aware Attention Based Graph Neural

Networks where each image is encoded into a graph and models the multi-type inter-object relations to learn the question-adaptive relations which outperformed previous state-of-the-art approaches for the VQA challenge (Li et al., 2019). Liang et al. (2021) also showcase how in a language guided graph neural network, question answering can be translated into multiple iterations of a message passing operation among graph nodes. This brings us to the approach by Chen et al. (2021) and the alternate architecture highlighted under this proposal.

3 Key Novelty and Contributions

The novelty of this thesis lies in the Master Node usage to perform explicit relation-awareness from the History-Aware-Master-Node into the Q-Graph followed by the Question-Aware-Master-Node into the I-Graph as compared to Chen et al. (2021)'s implicit relation-awareness using History-aware concatenation to the Q-Graph and Question-Aware concatenation to the I-Graph before training the Multi-modal fusion model.

The optional settings as listed under 4.2 may also help to investigate the effectiveness of this Master-Node approach.

4 Methods

4.1 Primary Architecture

The main research goal is to develop and implement the architecture as highlighted in 1 which is closely modelled after the work of (Chen et al., 2021). This approach differs primarily in terms of having a master node representation from the bottom sequence graph for a more dynamic message passing protocol, which in this case is an explicitly modelled relation, as compared to Chen et al. (2021)'s approach where

they use a concatenation operation to form the three sequential graphs[1].

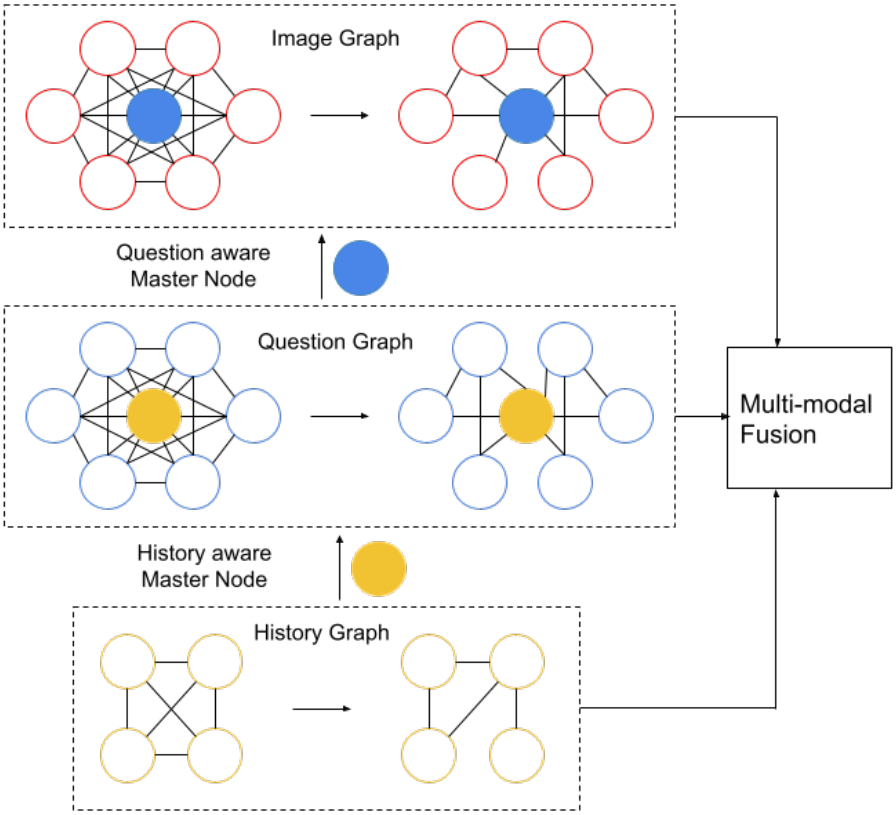

Figure 1: Primary Architecture using Master Node for Message Passing to create Multi-modal Fusion Graphs

4.2 Optional Architecture

An optional research goal is to investigate the performance of a different setting where the image and the History Dialog states graphs are embedded into a master node each and then fed into the Question-Graph. The architecture can be observed below [1].

[1]Note that the Master Nodes are replacing the implicit embeddings of Chen et al. (2021)'s GoG model, so the Master Nodes connect to all the nodes of the graph after the graph topology has been determined.

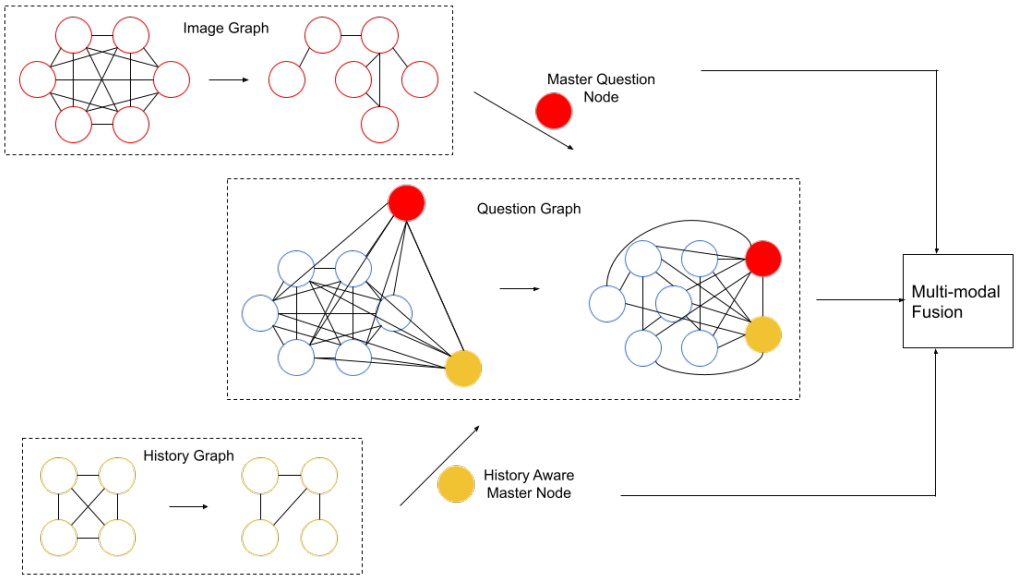

Figure 2: Optional Architecture using Master Node for Message Passing to create Multi-modal Fusion Graphs

4.3 Dataset

4.3.1 Primary Dataset

The primary Datasets to be used in this thesis are Vis-Dial v0.9 and v1.0 for Experiments and Ablation Studies (Das et al., 2016).

4.4 Optional Dataset

The optional Dataset used for further evaluation is a special case of the Vis-Dial dataset called VisDialConv which consists of 97 dialogs where the crowd-workers identified single turns (with dense annotations) requiring historical information (Agarwal et al., 2020).

4.5 Ablation Studies

This is another optional part of the thesis proposal which aims to investigate the performance of the Multi-modal Fusion Model by:

- Ablating the I-Graph.
- Ablating the H-Graph.
- Partial Connectedness of the nodes of a graph with the master node which occurs when the master node has been added to the H-Graph or the I-Graph before the respective graph topology has been identified. This experiment of partial connectedness with the master node will be carried out with the same ablation study structure carried out by Guo et al. (2020) where each node selects the K most relevant nodes in every message passing step and receives messages from them [2].

5 Intended Outcomes

The thesis is intended to discover the benefits of the proposed method of multi-modal fusion 1 in the task of Visual Dialog, and to compare the model performance to the previous state-of-the-art baselines (including Chen et al. (2021)'s GoG model) and outperform them. The task would be carried out on Vis-Dial v0.9 and v1.0 (Das et al., 2016).

As the Optional Goals highlight, the thesis also intends to study the outcome of the performances of the alternative methods of multi-modal fusion outlined 4.2.

The Ablation Studies, as highlighted under both 4 and 6.2 are used to study the model performance when either the Dialog History, or the Image Graph is ablated, both of whose outcomes should result in lower performance, but could help understand the importance and relevance of each of these graphs. The Ablation Study

[2]K-neighbour settings: k = {1,2,4,8,16,36} nodes, similar to the settings used by Guo et al. (2020)

for Dialog History and the performance drop can also be confirmed with the work of Agarwal et al. (2020) as we should expect to see similar performances since the proposed model explicitly encodes Dialog History as well.

Finally, the thesis also as an optional goal intends to benchmark another dataset called VisDialConv and compare the performance of the models developed with this dataset (Agarwal et al., 2020).

6 Goal

6.1 Mandatory Goals

- Reimplement Chen et al. (2021)'s GoG baseline for comparison.
- Create the Multi-Modal Fusion Graph Neural Network for Visual Dialog using the Architecture listed in 1 and measure its performance against previous state-of-the-art baselines.

6.2 Optional Goals

- Create the Multi-modal Fusion Graph Neural Network for Visual Dialog using the Architecture listed in 2 and measure its performance against the primary model and the previous state-of-the-art baselines.
- Perform Ablation Studies as listed under 4.5.
- Compare primary (and/or) optional model performances with the Optional Dataset.
- Visualization of Attention (word highlights for text, bounding boxes for images) in the Graphs

7 Schedule with Milestones

An estimated timeline for the thesis is shown below. The start date of the thesis is on the 1st of April 2022 and the end is 30th September 2022.

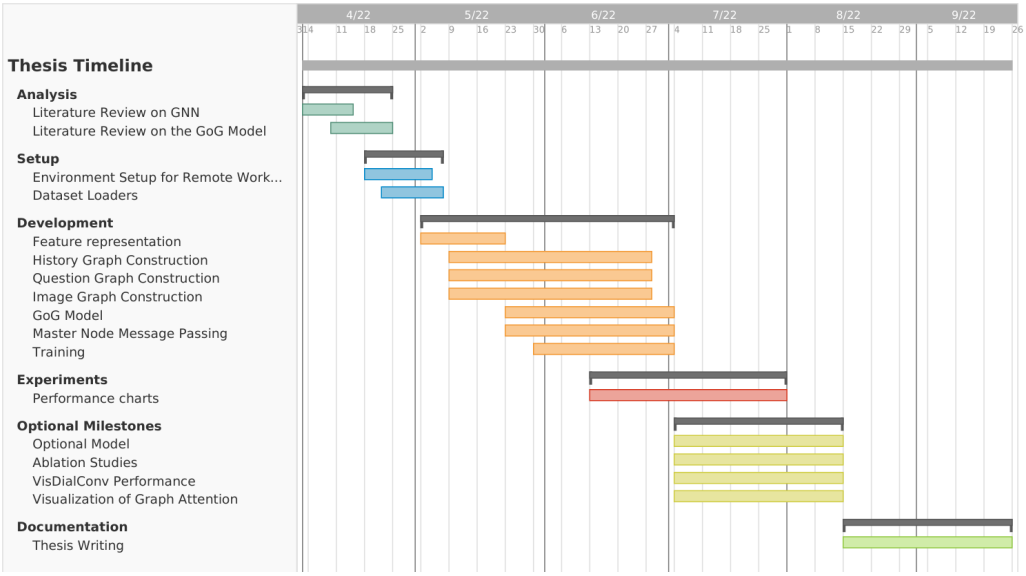

Figure 3: Timeline for the Master's Thesis

References

Shubham Agarwal, Trung Bui, Joon-Young Lee, Ioannis Konstas, and Verena Rieser. History for visual dialog: Do we really need it? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8182–8197, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.728. URL https://aclanthology.org/2020.acl-main.728.

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

Feilong Chen, Xiuyi Chen, Fandong Meng, Peng Li, and Jie Zhou. Gog: Relation-aware graph-over-graph network for visual dialog. *CoRR*, abs/2109.08475, 2021. URL https://arxiv.org/abs/2109.08475.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M. F. Moura, Devi Parikh, and Dhruv Batra. Visual dialog. *CoRR*, abs/1611.08669, 2016. URL http://arxiv.org/abs/1611.08669.

Yanlin Feng, Xinyue Chen, Bill Yuchen Lin, Peifeng Wang, Jun Yan, and Xiang Ren. Scalable multi-hop relational reasoning for knowledge-aware question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1295–1309, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.99. URL https://aclanthology.org/2020.emnlp-main.99.

Dan Guo, Hui Wang, Hanwang Zhang, Zheng-Jun Zha, and Meng Wang. Iterative context-aware graph inference for visual dialog. *CoRR*, abs/2004.02194, 2020. URL https://arxiv.org/abs/2004.02194.

Linjie Li, Zhe Gan, Yu Cheng, and Jingjing Liu. Relation-aware graph attention network for visual question answering. *CoRR*, abs/1903.12314, 2019. URL http://arxiv.org/abs/1903.12314.

Weixin Liang, Yanhao Jiang, and Zixuan Liu. Graghvqa: Language-guided graph neural networks for graph-based visual question answering. *CoRR*, abs/2104.10283, 2021. URL https://arxiv.org/abs/2104.10283.

Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. KagNet: Knowledge-aware graph networks for commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2829–2839, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1282. URL https://aclanthology.org/D19-1282.

Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the variance of the adaptive learning rate and beyond. *CoRR*, abs/1908.03265, 2019. URL http://arxiv.org/abs/1908.03265.

Tom McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1334. URL https://aclanthology.org/P19-1334.

Todor Mihaylov and Anette Frank. Knowledgeable reader: Enhancing cloze-style reading comprehension with external commonsense knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 821–832, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1076. URL https://aclanthology.org/P18-1076.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683, 2019. URL http://arxiv.org/abs/1910.10683.

Hongyu Ren, Weihua Hu, and Jure Leskovec. Query2box: Reasoning over knowledge graphs in vector space using box embeddings. *CoRR*, abs/2002.05969, 2020. URL https://arxiv.org/abs/2002.05969.

Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1), Feb. 2017. URL https://ojs.aaai.org/index.php/AAAI/article/view/11164.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017. URL http://arxiv.org/abs/1706.03762.

Denny Vrandečić and Markus Krötzsch. Wikidata: A free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85, sep 2014. ISSN 0001-0782. doi: 10.1145/2629489. URL https://doi.org/10.1145/2629489.

Yue Wang, Shafiq R. Joty, Michael R. Lyu, Irwin King, Caiming Xiong, and Steven C. H. Hoi. VD-BERT: A unified vision and dialog transformer with BERT. *CoRR*, abs/2004.13278, 2020. URL https://arxiv.org/abs/2004.13278.

Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. QA-GNN: reasoning with language models and knowledge graphs for question answering. *CoRR*, abs/2104.06378, 2021. URL https://arxiv.org/abs/2104.06378.