SelfGraphVQA: A Self-Supervised Graph Neural Network for Scene-based Question Answering

Bruno Souza*
University of Campinas
b234837@dac.unicamp.br

Marius Aasan
University of Oslo
mariuaas@uio.no

Helio Pedrini
University of Campinas
helio@ic.unicamp.br

Adín Ramírez Rivera
University of Oslo
adinr@uio.no

Abstract

The intersection of vision and language is of major interest due to the increased focus on seamless integration between recognition and reasoning. Scene graphs (SGs) have emerged as a useful tool for multimodal image analysis, showing impressive performance in tasks such as Visual Question Answering (VQA). In this work, we demonstrate that despite the effectiveness of scene graphs in VQA tasks, current methods that utilize idealized annotated scene graphs struggle to generalize when using predicted scene graphs extracted from images. To address this issue, we introduce the SelfGraphVQA framework. Our approach extracts a scene graph from an input image using a pre-trained scene graph generator and employs semantically-preserving augmentation with self-supervised techniques. This method improves the utilization of graph representations in VQA tasks by circumventing the need for costly and potentially biased annotated data. By creating alternative views of the extracted graphs through image augmentations, we can learn joint embeddings by optimizing the informational content in their representations using an un-normalized contrastive approach. As we work with SGs, we experiment with three distinct maximization strategies: node-wise, graph-wise, and permutation-equivariant regularization. We empirically showcase the effectiveness of the extracted scene graph for VQA and demonstrate that these approaches enhance overall performance by highlighting the significance of visual information. This offers a more practical solution for VQA tasks that rely on SGs for complex reasoning questions.

1. Introduction

The successful execution of Visual Question Answering (VQA) relies on a comprehensive understanding of the scene, including spatial interrelationships and reasoning inference capabilities [1, 13]. Incorporating scene graph (SG) representations in SG-VQA tasks has shown promising outcomes [12, 15, 17, 23, 28], providing concise representations of complex spatial and relational information.

Earlier investigations into SG-VQA demonstrated that successful models primarily rely on the utilization of manually annotated scene graphs for training [19, 20, 23], resulting in remarkably high levels of accuracy on the GQA dataset [13], surpassing human performance by a significant margin (see Table 1).

Despite the promising results, we argue that utilizing pre-annotated SGs in VQA is impractical in the real world due to its labor-intensive nature. Also, it permits a wide range of semantically corresponding SG [11] and when annotated it could potentially introduce questions-related biases, giving rise to concerns about its generalizability [2]. These issues may limit the model's ability to solve real-world problems beyond the dataset [21]. This is evident in a significant decline in accuracy, approximately 60% when models are confronted with automatically generated SGs. Additionally, studies assert that the main limitation in generalizing stems largely from linguistic correlations. [2, 16].

In this study, we address these challenges by extracting an SG from a given image using an unbiased, off-the-shelf scene graph generator [15], with the aim of removing any potential information leakage, as illustrated in Fig. 1's structure. Furthermore, our method employs semantically preserving augmentation, integrated with un-normalized contrastive framework, to further mitigate potential linguistic biases to enhance the visual cues translated as SG for VQA. We refer to it as the *SelfGraphVQA framework*, cf. Fig. 1.

Given its simplicity [7], our approach is trained using joint embeddings and a Siamese network architecture, inspired by the SimSiam model, which does not require negative samples [5, 9]. In this work, we explore three un-normalized contrastive approaches (node-wise, graph-wise, and regularization for permutation equivariance) and demonstrate its effectiveness by enhancing the visual information for the VQA task. A graph neural network (GNN) with a self-attention strategy (GAT) is employed to distill an SG representation relevant to the question by capturing

[*] Work carried out as Guest Researcher at UiO.

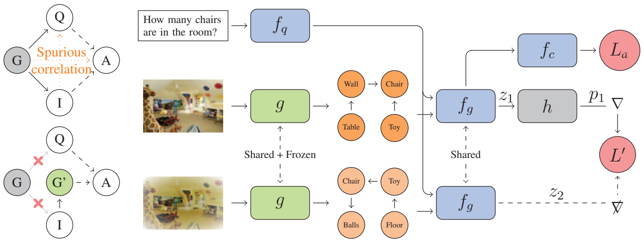


Figure 1: (Left) The statistical dependence of the task and the ideal graph, G. (Right) Our proposed framework removes data leakage by using the extracted SG G'. Our architecture comprises a question encoder f_q, a graph encoder f_g, and a classifier f_c. Two distinct views of one image are processed by the same pipeline. We use a frozen pre-trained SG generator g, and a prediction head h is applied through the top view with gradient backpropagation, while gradients are not propagated back from the lower view. We maximize the representation of the views using the similarity loss L'.

visual interaction content among objects in the scene [7].

Our work differs from existing VQA models in three main aspects: (i) we generate as SG using a pre-trained, unbiased scene graph generator [15] in a more practical approach; (ii) we utilize un-normalized contrastive learning on the SG representation, along with augmentation, to eliminate any potential spurious correlations from annotated data and to heighten the visual information; and (iii) the use of a GAT encoder to enhance high-level semantic and spatial reasoning on the SG. We further investigate the behavior of visual enhancement when employing a more expressive language encoder, specifically BERT [14]. Importantly, our SelfGraphVQA framework does not require the costly pre-training strategy common to transformer-based models commonly used in vision-language tasks [8, 25, 28].

2. Related Work

Scene Graph and Visual Question Answering. Accurately assessment of VQA tasks, requiring a comprehensive understanding of visual perception and semantic reasoning, has gained substantial attention in the academic community, as these tasks holds significant practical value, particularly in enhancing accessibility for the visually impaired [4, 14, 18, 29, 30].

Several works have explored the information that SG representations may bring to VQA [19, 27], as opposed to the more data-hungry transformer-based visual language models [8, 18, 25]. However, existing SG-VQA approaches typically rely on idealized scene graphs and inherent dataset reasoning [19,20]. Obtaining such annotations can be costly without an end-to-end pipeline. Moreover, even SoTA methods in SG-VQA exhibit limited generalization capabilities, potentially due to spurious correlations [2].

Table 1: Our experiments revealed a notable accuracy reduction in top-notch methods on the GQA dataset when transitioning from well-annotated to extracted scene graphs. We categorize methods by data type (e.g., annotated data or purely image-question extraction) and SGG usage. All methods are trained and validated uniformly, except for the test extracted configuration, trained on ideal data and validated on extracted SGG data.

Method	Eval. Data	Acc (%)
Human [13]	–	89.3
GraphVQA [19]	Annotated/SGG	94.8
LRTA [20]	Annotated/SGG	93.1
Lightweight [23]	Annotated/SGG	77.9
CRF [22]	Annotated	72.1
LXMERT [25]	Extracted	59.8
GraphVQA (original pre-trained on ideal)	**Test Extracted/SGG**	29.7
SelfGraphVQA (Local)	Extracted/SGG	51.5
SelfGraphVQA (Global)	Extracted/SGG	52.3
SelfGraphVQA (SelfSim)	Extracted/SGG	54.0

Self-Supervised Learning. Broadly speaking, recent advancements in self-supervised learning can be categorized into normalized [3, 6] and maximization representation learning [7, 10, 26]. Contrastive methods aim to bring embeddings of identically labelled images closer together while separating embeddings generated from different versions. In visual-language data, the prevailing approach for self-supervised learning involves pretraining a transformer-based model on a large dataset to solve pretext tasks before fine-tuning for downstream tasks [8, 24, 25, 28]. However, these methods can be computationally expensive and complex due to the use of negative samples and masking techniques. Modern un-normalized contrastive learning methods, e.g., BYOL [10] and SimSiam [7], use architectures inspired by reinforcement learning to maximize the informational content of the representations. In our proposal, we adopt a similarity maximization approach using a Siamese architecture for visual scene graph representation.

3. Methodology

We refer the reader to the appendix for the implementation details. We experiment with the maximization strategy with three independent and distinct similarity losses over either a localized node representation (i.e., object-wise), a global pooled graph representation (i.e., scene-wise), or a regularization node representation term to induce permutation equivariance. We denote the graph representations $z_i = f_g\big(g(x_i), f_q(q)\big)$, and the predictor's output vectors $p_i = h(z_i)$. Generally, the representations are maximized by minimizing the generic cosine distance D loss.

Local Similarity. To account for permutation invariance in the node representations, we compute cosine distances over all object pairs from the two views and use the maximally similar node embedding pairs to compute the local loss by

$$L^*_\ell(p_1, z_2) = \frac{1}{O} \sum_i^O \arg\min_{z_{2,j}} D(p_{1,i}, z_{2,j}), \tag{1}$$

where O is the number of objects in the scene. Symmetrically, we compute $L^*_\ell(p_2, z_1)$, to obtain the overall local loss

$$L_\ell(z_1, z_2) = \frac{1}{2}\big(L^*_\ell(p_1, z_2) + L^*_\ell(p_2, z_1)\big). \tag{2}$$

Global Similarity. After obtaining a graph representation, we follow an approach similar to cosine similarity maximization for image classification [7, 10]. Along with the intuition that contrasting between global representations may enhance the visual cues, we assume that the global representation contains the full information about the scene. Similar to the local representation, we minimize the cosine distance, yielding a loss on the form

$$L_g(z_1, z_2) = \frac{1}{2}\big(D(p_1, z_2) + D(p_2, z_1)\big). \tag{3}$$

Regularization for Permutation Equivariance. We employ an *anchor*, where the SG of an unmodified image guides the SG of the augmented image, allowing us to obtain a more accurate representation of the original scene. Our assumption is that the local similarity loss decreases the global performance, while global similarity provides a contextual representation but loses local details. This technique aligns similar nodes and encourages regularization, making augmented scene representations closer to the original, thus mitigating permutation invariance in graph representations.

Denote the anchored representation by z_1, and the augmented representation by z_2. We determine intra-similarities of the anchors $s_{1,i} = \arg\min_{z_{1,j}} D(z_{1,i}, z_{1,j})$ and similarities of augmented views $s_{2,ij} = D(z_{2,i}, z_{2,j})$. We then compute cross-entropy (CE) between anchors and augmentations

$$J(z_1, z_2) = \mathrm{CE}(s_1, s_2), \tag{4}$$

which acts as a regularizer to constrain permutation equivariance for the augmentations in addition to the local loss. We combine these losses using

$$L_s(z_1, z_2) = L_\ell(z_1, z_2) + J(z_1, z_2), \tag{5}$$

which we refer to as a local self-similarity loss (SelfSim).

Distribution Link Representation Regularization. Similarly to the regularization for permutation equivariance, we apply link regularization *in conjunction with one of the other three similarity strategies*. The edges of the *anchor* SG guide the edges of the augmented SG. Denote the anchored edge score representation by r_1, and the augmented edge score representation by r_2. These scores characterize the relationship between the objects in the scene, and we aim to make the link distribution more robust to perturbation. *In this case, the scene graph generator [15] is trainable.* We compute the cross-entropy between the anchored edge scores and the augmented edge scores $J_e(r_1, r_2) = \mathrm{CE}(r_1, r_2)$, which acts as a regularizer to constrain the link prediction distribution, yielding

$$L_e(z_1, z_2) = L_\ell(z_1, z_2) + J_e(r_1, r_2). \tag{6}$$

All models utilizing this added link distribution regularizer are characterized by the inclusion of the term "link."

Overall Optimization Objective. Lastly, we outline the overall loss for optimizing the VQA objective. To identify the correct answer $a \in A$ given an example (x, q, A), where x represents the input image, and q is the associated question, we extract a point estimate of probabilities

$$p(a \mid x, q) = \sigma\left(\mathrm{logit}(x)\right), \tag{7}$$

where σ is the softmax function, and $\mathrm{logit}(x) = f(x, q)$ are the logits for all possible answers produced by our encoder. We calculate the cross-entropy loss for each instance,

$$L_{sup}(x) = \mathrm{CE}\left(p(a \mid x, q), a\right). \tag{8}$$

Our combined training loss is then given by

$$L(x) = \alpha L_{sup}(x) + \beta L'(z_1, z_2), \tag{9}$$

where L' can be any of the aforementioned similarity loss strategies: L_ℓ, L_g, or L_s, with or without L_e. The α and β are controlled hyperparameters that balance the contribution of the various components in the total loss.

4. Experiments and Ablations

We evaluate our framework on the GQA dataset [13]. Our study aims to establish a practical foundation for demonstrating the potential of SG along with an un-normalized contrasting approach to improve visual cues for VQA. Despite the noise data in the extracted SG, we demonstrate its effectiveness, Fig. 2, by highlighting the importance of further exploration. The utilization of non-idealized SG-VQA methods with un-normalized contrastive learning leads to improvements across all metrics, Table 2. Furthermore, our framework demonstrates faster convergence during training, approximately 20% faster in epochs compared to baselines. However, further investigation is required to validate them.

The un-normalized contrastive approach universally enhances results across question categories (Fig. 2), with specific types of approaches further improving the model's performance based on the query type.

Table 2: Results (%) on GQA by standard metrics.

Method	Binary (↑)	Open (↑)	Consist. (↑)	Validity (↑)	Plausab. (↑)	Distr. (↓)	Acc (↑)
Baseline	65.8	29.7	58.2	94.9	90.5	11.7	50.1
Baseline+BERT	68.0	32.2	62.6	95.0	90.9	7.7	53.8
Local	66.8	30.2	59.4	94.9	90.6	8.8	51.5
Global	67.7	30.8	62.5	94.9	90.6	6.7	52.3
SelfSim	**68.4**	31.3	**65.9**	94.9	90.7	**2.1**	54.0
Global+BERT+link	68.0	**33.0**	63.9	95.0	**91.2**	8.9	**54.5**
SelfSim+BERT+link	68.2	32.8	64.3	**95.0**	91.0	8.0	**54.5**

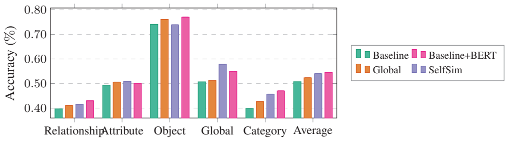


Figure 2: Accuracy on different question types.

Table 3: Change in accuracy under potentially disruptive augmentations and perturbations.

Question Type	Augmentation	Baseline	Global	Local	SelfSim
Relation	Flip	−1.6	−3.4	−3.2	−3.9
Attribute	Strong Color Jitter	+1.14	−3.7	−0.8	−1.2
Global	Gaussian Noise + Crop	−5.6	−7.7	−5.5	−8.1

Table 4: Results(%) of the Aug. Baseline and SelfSim.

Method	Binary	Open	Validity	Plausibility	Acc
Baseline Aug	65.1	28.7	94.6	90.1	50.1
SelfSim	68.4	31.3	94.9	90.7	54.0

We conducted ablations to demonstrate the functionality of our approach and carried out detailed observations that go beyond mere reliance on metrics using the GQA dataset.

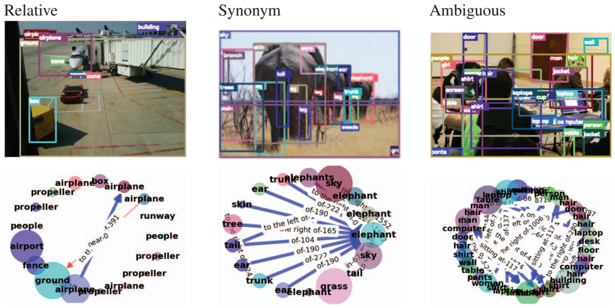
Does the Scene Graph Really Matter? Through a perturbation study where images were augmented based on question types, we introduced disruptive noise such as image flipping to challenge the model's ability to answer spatial relational questions. The goal was to observe mistakes in the model's answers. The results, compared to the baseline (Table 3), showed greater variation in our model's performance, indicating that it pays more attention to visual information, whereas the baseline appears to rely on other sources of information.

Are Performance Gains Mainly Due to Augmentations? We compared our approach with the baseline architecture, training solely with data augmentation techniques to evaluate their influence on overall performance. Table 4 provides evidence that data augmentation techniques actually impair the performance of the architecture.

Are Our Models Less Biased? Our initial hypothesis was that current top-performing models might incorporate biases present in the questions into their weights. We conducted experiments to analyze this issue, introducing random noise to features in the scene graph while preserving its topology, and perturbing the language in up to 50% of the words in the questions. The results in Table 5 demonstrate that our approach relies less on linguistic features, prioritizing overall information and reducing linguistic bias. Additionally, we explored visual enhancement, even when trained with a more expressive language module such as BERT. The experiments in Table 5 examine the impact of using BERT and its effect on enhancing visual information.

Table 5: Sensitivity of accuracy (%) for bias question analyzes of SelfGraphVQA and SelfGraphVQABERT.

Setup	**Methods**			
Scene Graph + Question	Baseline	Local	Global	SelfSim
Noise + SG	16.2	16.6	28.6	26.6
Question + Noise	39.9	38.3	37.4	39.8
Noise + Noise	12.7	14.6	18.9	21.0
Question + Scene Graph	BERT Baseline	BERTGlobal+link	BERTSelfSim+link	
Noise + SG	21.0	23.2	24.5	
Question + Noise	42.4	41.8	42.8	
Noise + Noise	19.8	21.7	21.3	



Q: Is there an airplane in the picture that is not small?
Answer: Yes
Prediction: No

Q: Where are the weeds?
Answer: Plain
Prediction: Field

Q: Is the man to the right or to the left of the cup?
Answer: Right
Prediction: Left

Figure 3: Examples to demonstrate the complexity of VQA.

Examples. Given the wide range of acceptable answers, we argue that solely relying on standard evaluation metrics may not provide a fair comparison, thus presenting additional challenges to the field. Fig. 3 demonstrates the utility of SG for interpretability, as they enable a graphical analysis of objects and the overall composition of the scene.

5. Conclusions

Despite promising results in VQA tasks with idealized SG, our study revealed that models relying on manually annotated and expensive SG struggle with real-world data. To address this, we proposed SelfGraphVQA, a more practical SG-VQA framework that breaks the spurious correlation of annotated SG and learns to answer questions using extracted SG from a pre-trained SG generator. We employed un-normalized contrastive learning to maximize similar graph representations in different views. All approaches utilizing self-supervision showed improvement over their baselines. Overall, we demonstrated the effectiveness of extracted SG in VQA, underscoring the significance of continued exploration of the potential of SG for complex tasks. We also showed that self-supervision over the SG representation improved the results by enhancing the visual information within the task. We hope that this work raises awareness of the challenges of accentuating the role of the scene in answering questions from images.

Acknowledgements

This work was supported in part by the FAPESP (São Paulo Research Foundation) grant no. 2022/09849-8 and BEPE grant no. 2022/09849-8. The computations were performed in part on resources provided by Sigma2—the National Infrastructure for High-Performance Computing and Data Storage in Norway—through Project NN8104K. This work was funded in part by the Research Council of Norway, via the Visual Intelligence Centre for Research-based Innovation (grant no. 309439), and Consortium Partners.

References

[1] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *IEEE/CVF Inter. Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 4971–4980, 2018. 1

[2] Aishwarya Agrawal, Ivana Kajic, Emanuele Bugliarello, Elnaz Davoodi, Anita Gergely, Phil Blunsom, and Aida Nematzadeh. Reassessing evaluation practices in visual question answering: A case study on out-of-distribution generalization. In *Conf. European Ch. Assoc. Comput. Ling. (EACL)*, pages 1171–1196, 2023. 1, 2

[3] Laurence Aitchison. InfoNCE is a variational autoencoder. *arXiv preprint arXiv:2107.02495*, 2021. 2

[4] Hedi Ben-Younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome. Mutan: Multimodal tucker fusion for visual question answering. In *IEEE/CVF Inter. Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 2612–2620, 2017. 2

[5] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature verification using a "siamese" time delay neural network. *Adv. Neural Inf. Process. Sys. (NeurIPS)*, 6, 1993. 1

[6] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 2

[7] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *IEEE/CVF Inter. Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 15750–15758, 2021. 1, 2, 3

[8] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European Conf. Comput. Vis. (ECCV)*, pages 104–120. Springer, 2020. 2

[9] Quentin Garrido, Yubei Chen, Adrien Bardes, Laurent Najman, and Yann Lecun. On the duality between contrastive and non-contrastive self-supervised learning. *arXiv preprint arXiv:2206.02574*, 2022. 1

[10] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Adv. Neural Inf. Process. Sys. (NeurIPS)*, 33:21271–21284, 2020. 2, 3

[11] Xuanli He, Quan Hung Tran, Gholamreza Haffari, Walter Chang, Trung Bui, Zhe Lin, Franck Dernoncourt, and Nhan Dam. Scene graph modification based on natural language commands. *arXiv preprint arXiv:2010.02591*, 2020. 1

[12] Ronghang Hu, Anna Rohrbach, Trevor Darrell, and Kate Saenko. Language-conditioned graph networks for relational reasoning. In *IEEE/CVF Inter. Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 10294–10303, 2019. 1

[13] Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. *10.48550/arxiv.1902.09506*, 2019. 1, 2, 3

[14] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. *Adv. Neural Inf. Process. Sys. (NeurIPS)*, 31, 2018. 2

[15] Boris Knyazev, Harm de Vries, Cătălina Cangea, Graham W Taylor, Aaron Courville, and Eugene Belilovsky. Graph density-aware losses for novel compositions in scene graph generation. *arXiv preprint arXiv:2005.08230*, 2020. 1, 2, 3

[16] Brenden Lake and Marco Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *Inter. Conf. Mach. Learn. (ICML)*, pages 2873–2882. PMLR, 2018. 1

[17] Linjie Li, Zhe Gan, Yu Cheng, and Jingjing Liu. Relation-Aware Graph Attention Network for Visual Question Answering. In *IEEE/CVF Inter. Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 10313–10322, 2019. 1

[18] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. VisualBERT: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 2

[19] Weixin Liang, Yanhao Jiang, and Zixuan Liu. GraghVQA: Language-guided graph neural networks for graph-based visual question answering. In *Wksp. Multimodal Artif. Intell. (ACLW)*, pages 79–86, Mexico City, Mexico, June 2021. Association for Computational Linguistics. 1, 2

[20] Weixin Liang, Feiyang Niu, Aishwarya Reganti, Govind Thattai, and Gokhan Tur. LRTA: a transparent neural-symbolic reasoning framework with modular supervision for visual question answering. *arXiv preprint arXiv:2011.10731*, 2020. 1, 2

[21] Man Luo, Shailaja Keyur Sampat, Riley Tallman, Yankai Zeng, Manuha Vancha, Akarshan Sajja, and Chitta Baral. 'just because you are right, doesn't mean i am wrong': Overcoming a bottleneck in the development and evaluation of open-ended visual question answering (VQA) tasks. *arXiv preprint arXiv:2103.15022*, 2021. 1

[22] Binh X Nguyen, Tuong Do, Huy Tran, Erman Tjiputra, Quang D Tran, and Anh Nguyen. Coarse-to-fine reasoning for visual question answering. In *IEEE/CVF Inter. Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 4558–4566, 2022. 2

[23] Sai Vidyaranya Nuthalapati, Ramraj Chandradevan, Eleonora Giunchiglia, Bowen Li, Maxime Kayser, Thomas Lukasiewicz, and Carl Yang. Lightweight visual question answering using scene graphs. In *ACM Inter. Conf. Inf. Knowl. Manag. (CIKM)*, pages 3353–3357, 2021. 1, 2

[24] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry,

Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Inter. Conf. Mach. Learn. (ICML)*, pages 8748–8763. PMLR, 2021. 2

[25] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019. 2

[26] Shantanu Thakoor, Corentin Tallec, Mohammad Gheshlaghi Azar, Remi Munos, Petar Veličković, and Michal Valko. Bootstrapped representation learning on graphs. In *Wksp. Geom. Topol. Represent. Learn. (ICLRW)*, 2021. 2

[27] Yanan Wang, Michihiro Yasunaga, Hongyu Ren, Shinya Wada, and Jure Leskovec. VQA-GNN: Reasoning with multimodal semantic graph for visual question answering. *arXiv preprint arXiv:2205.11501*, 2022. 2

[28] Zhecan Wang, Haoxuan You, Liunian Harold Li, Alireza Zareian, Suji Park, Yiqing Liang, Kai-Wei Chang, and Shih-Fu Chang. SGEITL: Scene graph enhanced image-text learning for visual commonsense reasoning. In *AAAI Conf. Artif. Intell. (AAAI)*, volume 36, pages 5914–5922, 2022. 1, 2

[29] Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. Merlot reserve: Neural script knowledge through vision and language and sound. In *IEEE/CVF Inter. Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 16375–16387, 2022. 2

[30] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. VinVL: Revisiting visual representations in vision-language models. In *IEEE/CVF Inter. Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 5579–5588, June 2021. 2