Answering visual questions through representation learning

Sumit Agarwal [1] **Suraj Tripathi** [1] **Syeda Nahida Akter** [1] **Andrew Lyubovsky** [1] **Feng Xiang** [1]

1. Introduction

Visual Question Answering(Visual QA) is an emerging research direction involving answering questions correctly from given images. Over time, the task of visual QA has increased in difficulty as emerging datasets have more complex questions and need an understanding of not only the objects in the image but also the relationships between them (Hudson & Manning, 2019). Some questions even require models to compare relationships among objects between two images (Suhr et al., 2019), while others are specific to understanding the text within the image (Singh et al., 2019). Our goal in this project is to create a generalized model suited for multimodal QA which can perform competitively or better than state-of-the-art models in visual QA tasks. We will focus on extracting textual and visual features from data to answer questions by learning joint representations and using cross-attention modules between modalities. Our direction will also include designing task-agnostic and task-specific pre-training methods to learn better joint representations. The code for this project is publicly available on *GitHub* [1].

2. Experimental Setup

There are several visual QA datasets that are worth considering in addressing the general problem of designing a multimodal QA model, focusing on 3 diverse datasets.

VQA (Antol et al., 2015): The VQA dataset is a benchmark dataset for free-form visual QA. It contains roughly 250K images, 760K questions, and 10 million possible answers to the questions. While models have been found to achieve high accuracy on this dataset, some biases exist that allow for some models to achieve good results even in the absence of images. We chose to use VQA as it is a common benchmark for visual QA tasks.

GQA (Hudson & Manning, 2019): The GQA dataset was developed based on real-world visual reasoning and compositional QA. It uses a question engine to generate 22M multiple-choice questions from Visual Genome (Krishna et al., 2017) images and scene graphs. GQA also contains 113k images with a vocabulary size of 3097 words and a total of 1878 possible answers. In addition to standard accuracy, the paper introduces additional evaluation metrics which include consistency (similar responses across different questions), validity (answers that are theoretically correct or within the same theme), plausibility (answers that are reasonable and make sense), grounding (attending to relevant image regions pertaining to questions), and distribution (matching between ground truth answer and model prediction answer distributions). Many models that had performed well on previous visual QA datasets have been shown to perform poorly on GQA as it requires a more complex understanding of the scene and semantic relationships between objects.

NLVR2 (Suhr et al., 2019): The NLVR2 dataset requires deciding whether a caption is true or false while comparing two images. The creation of the dataset involved collecting images from google search queries and crowd-sourcing captions to mitigate human biases. Performing well in this task requires reasoning about sets of objects, counts, and spatial relations. NLVR2 contains 107k human-written sentences grounded in pairs of photographs. We have chosen this dataset because this requires understanding between pairs of images and is different from the previous datasets in design. Since it's a binary question answering task, the evaluation metric is the per-question prediction accuracy.

3. Prior Work

3.1. Pretrained Models

There has been a lot of work done on understanding visual and language features together through pretrained models by letting these models learn joint representations of both modalities as part of the pretraining task using an architecture similar to BERT. (Devlin et al., 2019).

VisualBERT (Li et al., 2019): This model presents a simple yet effective baseline for multimodal QA tasks. While being a significantly simpler architecture, VisualBERT shows promising results on four vision-and-language tasks: VQA, VCR, NLVR2, and Flickr30K. It makes use of a transformer (Vaswani et al., 2017) network that implicitly learns connections between visual and text modalities. First, it processes regions of interest (ROIs) into feature

[1] https://github.com/sumit-agrwl/multimodal-vqa

vectors, and then sends those feature vectors along with text representations into a stack of transformer layers that produces the output. They proposed two pretraining objectives: masked language modeling (MLM), and sentence-image prediction. The first task is the standard language modeling task which randomly masks words from the text input and predicts them using the remaining words and given image input. The second task makes use of multiple captions present in the COCO dataset (Lin et al., 2014). Each image is paired with two captions where one caption describes the image and another caption has a 0.5 probability that it belongs to this image or was randomly chosen. The model learns to distinguish between these two captions.

LXMERT (Tan & Bansal, 2019): LXMERT is a large-scale transformer model which makes use of three encoder modules: an object relationship encoder, a language encoder, and a cross-modality encoder. The object relationship encoder uses feature vectors from ROIs to encode their relationships, and the language encoder encodes the word embeddings from the input text sentence. Next, the cross-modality module takes both encodings and applies cross-modality attention to generate the output. LXMERT significantly outperforms other multimodal QA models on VQA, GQA, and NLVR2 datasets. This model also makes use of diverse pretraining tasks in order to improve its ability to connect language and vision semantics. The following pretraining tasks are proposed in this work which helps the model learn both intra- and inter-modality features:

- *Masked Language Modeling (MLM)*: This task masks words in the sentence with 0.15 probability. The model makes use of image region features along with the text input to predict the correct word.
- *Masked Object Prediction*: Similar to MLM, this task masks image region features by replacing them with zeros and makes use of remaining regions and the full input sentence to make its prediction. This prediction task can be performed in 2 ways: (i) using a regression model to generate region features, and (ii) predicting the object label corresponding to masked region features. Object annotation is not consistent in different datasets therefore most prior work makes use of the object detector's label as output for this task. This object output label could be noisy but empirically performs well.
- *Cross-modality matching*: This is a classification task that takes image-sentence pairs as input and tries to predict if the sentence belongs to the image.
- *Image question answering*: This task makes use of visual QA datasets to include question-image pairs in the pretraining dataset. Around 1/3 of the sentences in the pretraining set are questions. They observed that having questions in the pretraining helps the model perform better on downstream QA tasks.

UNITER (Chen et al., 2020) : To learn universal image-text representation for all downstream V+L tasks, UNITER proposes a transformer-based, large-scale, pre-trained model for joint multimodal embedding similar to that used in VisualBERT. In contrast to prior works, UNITER introduces:

- *Conditional Masking for MLM/Masked Region Modeling (MRM)*: Opposed to prior works where both modalities were masked randomly, UNITER masks one modality at a time and leaves the other modality intact. This can prevent potential misalignment of random masking where a masked region needs to be described by a masked word.
- *Optimal-Transport-based Word-Region Alignment*: By adapting an OT-based learning approach, the paper tries to minimize the embedding transportation cost between image regions and words in a sentence (vice versa) which enforces better cross-modal alignment.

Experimental results show that the inclusion of both conditional masking and OT-based WRA can ease the misalignment between images and text, which results in better joint embeddings for downstream tasks. Additionally, UNITER uses four large-scale V+L datasets for pretraining: (i) COCO; (ii) Visual Genome (VG); (iii) Conceptual Captions (CC) (Sharma et al., 2018); and (iv) SBU Captions (Ordonez et al., 2011) where insertion of both *in-domain* (COCO and VG) and *out-of-domain* (CC and SBU Captions) datasets lead to better generalizability. UNITER outperforms SOTA pretraining models (LXMERT, VisualBERT, VLBERT (Su et al., 2020), ViLBERT (Lu et al., 2019)) by a large margin on six V+L tasks across nine datasets.

12in1 (Lu et al., 2020): This multi-task model works upon the ViLBERT model which consists of two parallel BERT encoders for image and textual features respectively. Each encoder is a series of transformer blocks connected by co-attention transformer layers which enable information exchange between the two modalities. The model is pretrained on the Conceptual Caption dataset (Sharma et al., 2018) with two tasks: masked multi-modal modeling and multi-modal alignment prediction. The first task randomly masks approximately 15% of both words and image tokens and reconstructs them given the remaining inputs. The second task requires the model to predict whether an image and caption correspond or not. The 12 in 1 model then works on this pretrained model to learn parameters that minimize loss across 12 different datasets spread across 4 types of tasks including visual QA, caption-based image retrieval, identification of image regions based on natural language expression, and predict semantic relationships between images

Table 1. Question Answering accuracies for pre-trained models across various data sets.

MODEL	VQA	NLVR2	GQA
12IN1	73.2%	78.9%	60.7%
UNITER	73.8%	79.1%	59.8%
LXMERT	72.5%	76.2%	60.3%
VISUALBERT	70.8%	67.4%	-

Table 2. Question Answering accuracy on GQA validation dataset.

MODEL	GQA VAL
MAC	61.9%
UNITER(WITHOUT SCENE GRAPH)	69.0%
SCENE GRAPH (GROUND TRUTH) + MAC	94.6%
GRAPHVQA	94.8%

given a question. The model shows that performing task-specific fine-tuning on the shared multi-task model achieves competitive or better results than SOTA models.

3.2. Scene Graphs

A scene graph, as defined in the Visual Genome dataset paper (Krishna et al., 2017), is a structured representation of an image, where nodes correspond to object bounding boxes with their object categories, and edges correspond to the pairwise relationships between objects. The information contained in scene graphs proves to be crucial for the performance of many modern visual QA models. GraphVQA is a graphical neural network-based model that takes question text and scene graph modalities. Tested against the GQA validation set, the model was able to beat SOTA models without the use of a visual image modality channel (Liang et al., 2021). Another high-performing scene graph model was is a MAC network model, which was the baseline model to the GQA dataset (Hudson & Manning, 2018), modified with an encoded scene graph input supplanted over the image representation channel (Damodaran et al., 2021). The modified MAC model was tested against the GQA validation set and surpassed what its original baseline performance was on the same validation set by a significant margin (Friedlander & Ng, 2019).

Scene graphs are evaluated through Recall@K scores which is the number of ground truth instances that were predicted in the top K predictions. There has been prior work in scene graph generation for images by analyzing the role of motifs: common substructures among scene graphs. Hence, the model tried to predict the most frequent relation between object pairs with the given labels, as seen in the training set (Zellers et al., 2018).

There has also been work around generating task-specific scene graphs like in VCTREE (Tang et al., 2019) which encodes the inherent parallel/hierarchical relationships among objects as a binary tree, e.g., “clothes” and “pants” usually co-occur and belong to “person”. This also allows the structure to vary from image to image and task to task, allowing more content-/task-specific message passing among objects. A Bidirectional Tree LSTM (BiTreeLSTM) is used to encode the contextual cues using the constructed VCTREE which is decoded for each specific end-task: Scene graph generation and visual QA. This method has been shown to achieve a mean Recall@50 score of 27.9% and should be well suited for the visual QA task.

4. Research Ideas

We would like to test multiple research directions that could improve the performance on visual QA tasks. These include training our model on multiple datasets and different tasks. Carefully analyzing the information that each modality contributes can also highlight existing biases that can then be removed. More specifically, we will explore the following directions: pretraining, scene graph generation, multi-task learning, image and text features, and no-answer scenarios.

Pretraining task: Along with making use of the prior task-agnostic and task-specific pretraining, we plan to propose new task-specific pretraining for our model. Prior work has shown that using the data of the task on masked language modeling with the image objective is beneficial for the task. Motivated by similar intuitions we will present our analysis on the following proposed pretraining tasks.

- *Region Extraction*: Following the intuition that only a small subset of regions are relevant for a given question, we will analyze the effectiveness of predicting relevant image regions given input image-question pairs. To generate data for this task, we will extract key phrases from the question and match them with the object detector’s labels to create a set of regions relevant to the given keyphrases. We will pose this task as a multilabel classification problem during pretraining. Also, we will see its impact as an auxiliary task while finetuning.
- *Question Generation*: This task will make use of images from the downstream task and generate captions for it using the SOTA models. Further, these captions will be converted into what, where, which, who, when type questions by using NER tagging on the captions and replacing them with relevant question keywords based on the predicted tags. Finally, the model will take the image-question pair as input to predict the correct answer. As our downstream tasks are based on QA, we believe that this novel direction might help improve the performance of our models.
- *Text-Image Autoencoder*: Unlike recent prior works, we will analyze the effectiveness of learning joint representation of texts and images using the autoencoder-based

learning objective. After learning a joint representation model, we will further train it to predict a missing modality given the other modality. We plan to make use of the image caption dataset for this work and we will analyze the use of raw image pixels and sentence representations.

Scene graphs: Scene graphs can be used along with the image features, object representations identified by Faster R-CNNs, and ROIs. There has been recent work that shows that using scene graph information on visual question answering generally has better performance than other visual-text baselines (Liang et al., 2021). We plan to use the GQA dataset for ground-truth scene graphs and for datasets that do not have scene graphs associated with them, we can use state-of-the-art scene graph generation techniques (Tang et al., 2019), (Tang et al., 2020). We can then encode the scene graph information using graph neural networks (Zhou et al., 2020), and pass them along with the image features to allow better co-attention between the different modalities. As a pretraining task, we can mask node/relation representations and expect the model to predict them, in a way learning to attend to parts of the scene graph necessary to answer the question.

Multi task learning: As lots of prior work has shown that multi-task learning on different datasets has enabled the model to capture the essence of visual and language correlation, none of the prior work has addressed multi-task learning only on Visual Question Answering datasets. We want to pretrain our model to perform well on different types of visual QA datasets and tasks like GQA, which require understanding the relationship between objects, VQA, which require a simple understanding of the image for the question, NLVR2, which require the knowledge of the relationship between images, and TextVQA, which requires the understanding of the text within the image. The intention of doing so is that the image and textual features can co-attend well to one another to answer the question at hand. Since all the datasets will belong to the visual QA domain, the model will be able to capture a larger amount of domain-specific information.

Image Captions along with Question Answering: Sometimes just asking the model to answer the question bypassing the question features and just the image features can be quite difficult. Instead, we can generate captions from the image with the help of pre-trained caption generation models and use them together with the question during the QA task in an attempt to look more at the textual features as well while answering the question.

Image Features: Using a strong model to encode visual images is still an important component for visual QA to this day. The use of different Convolutional Neural Networks (CNN), pretrained on either ImageNet or COCO datasets, affects performance by significant margins. When developing Hierarchical Co-Attention models, for example, the use of a pretrained ResNet model over a pretrained VGGNet CNN increased accuracy performance by about 1.5 percent when tested against the VQA dataset (Lu et al., 2016). Nowadays, there are other models emerging that could perform just as well in image encoding at less computational cost than CNNs. For example, on some tests in the task of image classification, Visual Transformers (ViT) performed better than deep CNN models at much smaller computational costs (Dosovitskiy et al., 2020). It would be worth investigating the performance of different neural network models against the end performance of the overall visual QA model.

Enhancing Image and Text Features: To address the inherent linguistic biases present in the dataset, prior works proposed two major directions: (1) modifying pretraining tasks and (2) data augmentation. As a part of modified pretraining tasks, previous works mostly focused on MLM, variants of MRM, image-text matching (ITM). But in terms of question answering, the focus should be primarily on the objects mentioned in both question text and image. Additionally, it is important for the model to identify question types and learn which answers are valid for each one, regardless of how often each question type occurs in training. This would lead to better generalization. Instead of masking random words in the question text, we can mask the whole question type (e.g., masking “How many” vs “How” or “many”). This will enforce models to learn about the question pattern instead of learning dataset-specific spurious correlations. Furthermore, previous works suggested different data augmentation techniques such as object removal, color change, negation, etc. which enforce models to focus on the relevant regions in the image.These augmentation techniques often require additional answer formulation which can be time-consuming. Rather we want to design perturbations that do not necessarily change the answer but can bewilder models. For example, (1) rephrasing questions with semantically similar words, (2) changing question types, (3) applying different image augmentation techniques, (4) adding irrelevant image segments to the original image, (5) adding unnecessary verbose to the question text/ captions. These experiments will enhance the robustness of the models against out-of-distribution questions and adversarial attacks.

Addressing no-answer scenario: To the best of our knowledge, no work has been done on a no-answer question scenario for visual QA which is very common in contextual extractive QA (Rajpurkar et al., 2018; Kamath et al., 2020; Jiang et al., 2021). In this study, we will explore no-answer cases for visual QA and will try to design a calibrator using the text-image features and output distributions to produce a confidence score against each answer.

References

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., and Parikh, D. VQA: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015.

Chen, Y.-C., Li, L., Yu, L., Kholy, A. E., Ahmed, F., Gan, Z., Cheng, Y., and Liu, J. UNITER: Universal image-text representation learning. In *Proceedings of the European conference on computer vision*, 2020.

Damodaran, V., Chakravarthy, S., Kumar, A., Umapathy, A., Mitamura, T., Nakashima, Y., Garcia, N., and Chu, C. Understanding the role of scene graphs in visual question answering, 2021.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 4171–4186, 2019.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16× 16 words: Transformers for image recognition at scale. 2020.

Friedlander, H. and Ng, P. Real world graphical reasoning and compositional question answering. 2019.

Hudson, D. A. and Manning, C. D. Compositional attention networks for machine reasoning. In *International Conference on Learning Representations*, 2018.

Hudson, D. A. and Manning, C. D. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6700–6709, 2019.

Jiang, Z., Araki, J., Ding, H., and Neubig, G. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977, 2021.

Kamath, A., Jia, R., and Liang, P. Selective question answering under domain shift. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5684–5696, 2020.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.

Li, L. H., Yatskar, M., Yin, D., Hsieh, C.-J., and Chang, K.-W. VisualBERT: A simple and performant baseline for vision and language, 2019.

Liang, W., Jiang, Y., and Liu, Z. GraphVQA: Language-guided graph neural networks for graph-based visual question answering. In *Proceedings of the Third Workshop on Multimodal Artificial Intelligence*, pp. 79–86, 2021.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *Proceedings of the European conference on computer vision*, pp. 740–755, 2014.

Lu, J., Yang, J., Batra, D., and Parikh, D. Hierarchical question-image co-attention for visual question answering. *Advances in neural information processing systems*, 29:289–297, 2016.

Lu, J., Batra, D., Parikh, D., and Lee, S. ViLBERT: Pre-training task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pp. 13–23, 2019.

Lu, J., Goswami, V., Rohrbach, M., Parikh, D., and Lee, S. 12-in-1: Multi-task vision and language representation learning. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Ordonez, V., Kulkarni, G., and Berg, T. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24:1143–1151, 2011.

Rajpurkar, P., Jia, R., and Liang, P. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pp. 784–789, 2018.

Sharma, P., Ding, N., Goodman, S., and Soricut, R. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pp. 2556–2565, 2018.

Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., and Rohrbach, M. Towards vqa models that can read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., and Dai, J. VL-BERT: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*, 2020.

Suhr, A., Zhou, S., Zhang, A., Zhang, I., Bai, H., and Artzi, Y. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6418–6428, 2019.

Tan, H. and Bansal, M. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019.

Tang, K., Zhang, H., Wu, B., Luo, W., and Liu, W. Learning to compose dynamic tree structures for visual contexts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019.

Tang, K., Niu, Y., Huang, J., Shi, J., and Zhang, H. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3716–3725, 2020.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Zellers, R., Yatskar, M., Thomson, S., and Choi, Y. Neural motifs: Scene graph parsing with global context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5831–5840, 2018.

Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., and Sun, M. Graph neural networks: A review of methods and applications. *AI Open*, 1:57–81, 2020.