Improving Visual Question Answering through a Topic-Aware Selection Layer

Anonymous Author(s)

ABSTRACT

Visual Question Answering (VQA) has emerged as a promising area of research for developing AI-based systems for various interactive and immersive applications. Numerous VQA methods have been proposed, leveraging advancements in natural language processing and computer vision. However, in this paper, we argue that these methods may yield inconsistent predictions due to the lack of mechanisms to constrain the prediction category in relation to the asked question. For instance, asking "what color is the car?" might elicit a response like "No", which can create confusion and erode user trust in the system. To tackle these issues, we introduce a Topic-Aware Selection Output Layer (TASOL) that remains model-agnostic, seamlessly integrating with any existing VQA architecture. This layer embeds a data structure encoding the relationship between question categories and the answer output space. It effectively constrains the model's prediction output in alignment with the asked question, ensuring coherence and consistency. We evaluate the effectiveness of TASOL by employing two datasets, VQA v2.0 and SimpsonsVQA, six vanilla VQA models, and across various metrics. Our analysis reveals that when integrated with many VQA baseline methods, TASOL significantly enhances accuracy while also improving consistency and reliability.

KEYWORDS

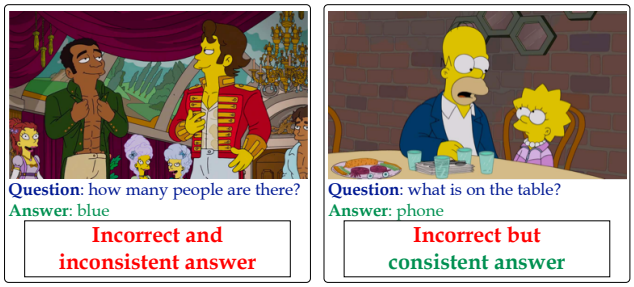
Visual Question Answering; Trust in AI Systems; Deep Learning.

ACM Reference Format:

Anonymous Author(s). 2024. Improving Visual Question Answering through a Topic-Aware Selection Layer. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Visual Question Answering (VQA) stands as a promising research field that bridges Computer Vision (CV) and Natural Language Processing (NLP). This innovative field allows machines to understand and respond to inquiries regarding visual content [1–5], marking a crucial step towards interaction between humans and artificial intelligence systems. Hence, over the past few year, numerous VQA algorithms and methods have been proposed, leveraging advancements in Artificial Intelligence to achieve this goal across a multitude of applications ranging from healthcare [6–8] and diagnosing medical images [9, 10], to cultural heritage [11, 12], aiding customer service [13], and enhancing entertainment experiences [14].



(a) (b)

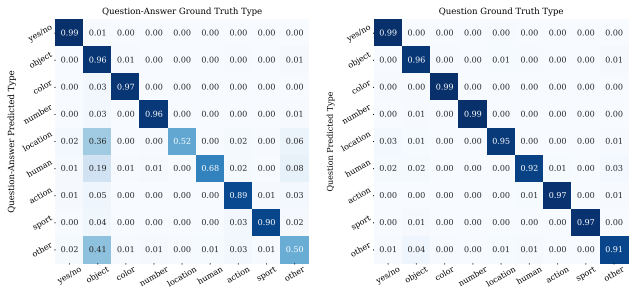
Figure 1: Examples of images, questions, and answers provided by the LSTM_VGG VQA model in [1].

However, in this paper, we posit that many established VQA methods are prone to generating inconsistent predictions. Consider the example shown in Figure 1a, where the user asks the question *"how many people are there?"*, for which the **LSTM_VGG** VQA model proposed in [1] erroneously responds with the answer *blue"*. Here, the answer is not only incorrect but also inconsistent with the question. Such discrepancies not only provide inaccurate information but can also create confusion and erode user trust in the system. On the other hand, in the example illustrated in Figure 1b, despite the incorrect response of *"phone"* to the question *"what is on the table?"*, it maintains at least some consistency as the answer type–an object–is relevant to the expected question type. Hence, in this paper, we argue that when providing an answer to a question, it is preferable to ensure consistency even if the answer is incorrect.

To assess the severity of inconsistent answers for the **LSTM_VGG** VQA model presented in [1], we refer to Figure 2a. Here, we present a confusion matrix comparing the model's predicted answer type with the expected true answer type. We observe that while the answer type frequently aligns with the question type for specific questions (such as "yes/no" questions, exhibiting consistency in 99% of cases), there exists a notable level of inconsistency with the question type for other answers. For example, in the case of "human", "location", or "other" type of questions, consistency is observed in only 68%, 52%, and 50% of cases, respectively. Often, these inconsistent predictions are due to the lack of mechanisms to constrain the prediction category in relation to the asked question. Yet, training a basic question type prediction classifier with a simple BiLSTM model, solely examining the question, yields fewer inconsistencies, as illustrated in Figure 2b. We note that inconsistency for "human" type questions improves respectively to 92%, for "location" to 95%, and for "other" to 91%. Therefore, the idea is that if we could instruct the VQA model to select the answer from a specific type only, this constraint could narrow down the search space and potentially assist the model in identifying the correct answer, or at

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn



(a) (b)

Figure 2: Confusion matrices comparing (a) predicted answer type and true answer type for the LSTM_VGG VQA model proposed in [1] and (b) predicted question type and true answer type using a simple BiLSTM model.

the very least, the correct answer type preserving confidence and ensuring trust in the system.

To tackle the issues discussed above, we introduce in this paper a Topic-Aware Selection Output Layer (TASOL) that remains model-agnostic, seamlessly integrating with any existing VQA architecture. This layer embeds a data structure encoding the relationship between question type categories and the answer space. It effectively constrains the model's prediction output in alignment with the asked question, ensuring coherence and consistency. We evaluate the effectiveness of TASOL by employing two datasets, VQA v2.0 and SimpsonsVQA, six vanilla VQA models, and across various metrics. Our analysis reveals that when integrated with most VQA baseline methods, TASOL significantly enhances accuracy while also improving consistency and reliability.

We summarize the contributions of this article as follows: (i) we identify a significant challenge in existing VQA methods, namely the propensity to generate inconsistent predictions; (ii) we introduce TASOL, a model-agnostic layer designed to enhance coherence and consistency in VQA systems; (iii) we present a thorough evaluation of TASOL through experiments conducted on two datasets, VQA v2.0 and SimpsonsVQA, using various vanilla VQA models and metrics.

2 RELATED WORK

There is a substantial body of research related to visual question answering. Below, we provide a brief overview of existing VQA methods and highlight current research directions in the field.

Evolution and Trends : VQA gained significant attention following the release of the VQA v2.0 dataset [1] in 2014. Since then, there has been a notable surge in research interest, resulting in the development of various models ranging from initial small-scale models like LSTM+VGG [15], Stacked Attention Networks [16], and Co-Hierarchical Attention [17], to the transformative adoption of Transformers [18] in 2019, exemplified by the Modular Co-Attention Network (MCAN) [19]. The field underwent a significant paradigm shift with the introduction of Visual Language Pretrained (VLP) models such as VisualBERT [20] and LXMERT [21], which have set new benchmarks in visual-language tasks including VQA and Image Captioning [22]. The advent of these models, including large-scale architectures like Flamingo (with 80.2B parameters) [23] and PaLI (with 16.9B parameters) [24], which consist of billions of parameters and require extensive data, has predominantly been pursued by major AI corporations like Google and Microsoft.

Exploring question types in VQA: In the quest to refine the accuracy and efficiency of VQA systems, several innovative approaches have emerged, spearheaded by the introduction of the Task-Directed Image Understanding Challenge (TDIUC) dataset [25]. This seminal dataset categorizes questions into 12 distinct types, each reflecting a different aspect of reasoning skill, from identifying objects and colors to understanding spatial relationships and inferential reasoning. Hence, diverse VQA models have been developed to adeptly handle various question types, marking significant advancements with targeted approaches. Initially, the Question-Type Guided Attention model proposed in [26] employed question type features extracted from an LSTM to refine visual attention mechanisms. This was followed by the introduction of CAQT [27], which enhanced accuracy by concatenating question type features with co-attention features. However, considering each question type typically corresponds to a distinct answer vocabulary set—for instance, "Yes/No" questions are primarily answered with *"yes"* or *"no"*, making answers like *"1"* or *"sad"* contextually inappropriate—models like QC-VQA [28] and subsequent iterations [29] have been introduced to narrow down answer sets for each question type. Moreover, these models incorporate specialized sub-decoders, each designed to handle a specific question type, thus scaling the model's size based on the number of question types addressed. However, while these models integrate question types into the modeling process, they lack a clear mechanism to constraint and restrict the answer search space.

Enhancing VQA with Large Language Models (LLMs): The recent integration of LLMs like ChatGPT to VQA models represents a pivotal innovation, pushing the boundaries of what's achievable in the field. At the forefront of this advancement is Flamingo [23], an LLM boasting 80.2 billion parameters, setting the current benchmark for state-of-the-art performance. Despite its unparalleled capabilities, Flamingo's extensive resource requirements render it inaccessible for widespread use. In contrast, GPT-3, with its remarkable in-context few-shot learning abilities, offers a more accessible yet potent option for enhancing VQA systems. This model's capacity to rapidly adapt to new tasks makes it an invaluable asset for improving VQA models, as evidenced by recent studies [30–32].

3 TASOL: TOPIC-AWARE SELECTION LAYER

In this section we first present the VQA task, and then introduces our Topic-Aware Selection Layer.

3.1 Notation and the VQA task

Settings: Let's consider a set of images denoted as $\mathcal{I} = \{\mathbf{i}_1, \mathbf{i}_2, \cdots\}$, alongside a set of questions, represented as $\mathcal{Q} = \{\mathbf{q}_1, \mathbf{q}_2, \cdots\}$. Each question $\mathbf{q}_j$ corresponds to a specific question type (or category) t_i within a set of l question types, denoted as $\mathcal{T} = \{t_1, t_2, \cdots t_l\}$. Furthermore, the questions have a range of k possible answers, encompassed by $\mathcal{A} = \{a_1, a_2, \cdots a_k\}$. We define a dataset $\mathcal{D} =$

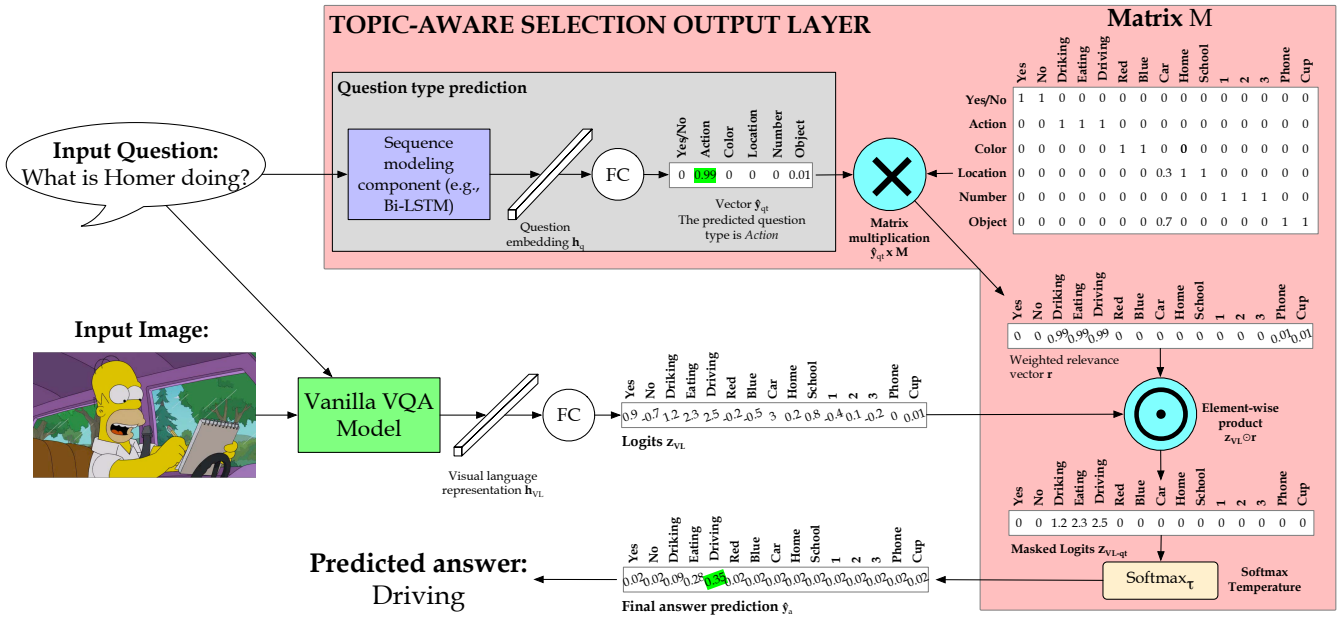


Figure 3: Architecture of the Topic-Aware Selection Output Layer (TASOL) with toy example.

$\{(\mathbf{i}^{(i)}, (\mathbf{q}^{(i)}, \mathbf{t}^{(i)}), a^{(i)})\}_{i=1}^{m}$ with m instances, where each instance represents a unique combination of an image, a question with its corresponding type, and the correct answer.

The VQA task: The objective of the traditional VQA task is to design a classification algorithm that learns a mapping function $f : (\mathcal{I}, \mathcal{Q}) \rightarrow \mathcal{A}$, that maps each image-question pair $(\mathbf{i}^{(i)}, \mathbf{q}^{(i)})$ to its corresponding correct answer $a^{(i)}$. The task represents the scenario where a user poses a question to which the system is expected to respond accurately. We note that often, the question type $t^{(i)}$ associate with a question $\mathbf{q}^{(i)}$ is ignored in the process. Hence, our aim in this work is to use the question type $t^{(i)}$ as a constraint for the model, guiding it to select answers that are probable matches for this type.

3.2 Proposed *TASOL* Layer

Figure 3 illustrates the architecture of the proposed Topic-Aware Selection Output Layer (TASOL) using a toy example. As previously mentioned, TASOL serves as a layer that can be integrated with any VQA model to refine the answer selection process and narrow down the search space. To briefly outline, the vanilla VQA model first processes both the input image and question, resulting in the generation of an unnormalized prediction vector of logits denoted as $\mathbf{z}_{VL}$. On the other hand, the TASOL layer initially categorizes the question based on its type, employing a sequence model to identify its category. Subsequently, TASOL combines the logit vector $\mathbf{z}_{VL}$ with the predicted question type to discern the most relevant answers tailored to the specific question type. This crucial step ensures that the final answer not only accurately corresponds to the image content but also aligns with the nature of the question posed, thereby enhancing consistency and performance in the VQA task.

Further details regarding the functionality of TASOL are provided in the subsequent subsections.

3.2.1 Question Type Classification: This component in TASOL learns a mapping function $f : \mathcal{Q} \rightarrow \mathcal{T}$, that maps each question $\mathbf{q}^{(i)}$ to its corresponding type $t^{(i)}$. Specifically, a sequence model, like a basic Recurrent Neural Network (RNN) such as LSTM or GRU, or a more sophisticated model like BERT [33] or ROBERTA [34], is employed to analyze and comprehend the intricacies embedded within the posed question. This sequence model will extract deep, contextual features from the text, providing an embedding representation of the question's intent denoted $\mathbf{h}_q$. Upon extracting these features, the next step involves passing $\mathbf{h}_q$ through a Fully Connected (FC) layer with a softmax activation to obtain $\hat{\mathbf{y}}_{qt}$, an l-dimensional probability distribution vector where l represents the number of predefined question types.

3.2.2 Answer-Question Type Relationship Encoding: TASOL embeds a data structure that encodes the relationship between question types and answers, which can be learned from the training data. Specifically, given a dataset $\mathcal{D} = \{(\mathbf{i}^{(i)}, (\mathbf{q}^{(i)}, \mathbf{t}^{(i)}), a^{(i)})\}_{i=1}^{m}$ at hand, we can define a matrix $M \in \mathbb{R}^{l \times k}$ to capture the relationship between question types and answers. Here, each entry $M_{i,j}$ in the matrix denotes the conditional probability $p(t_i|a_j)$ of encountering a question type t_i given an answer a_j within the dataset $\mathcal{D}$. For instance, it is anticipated that $p(t_i = \text{“color”}|a_j = \text{“red”}) \approx 1$, indicating a high likelihood of the question type being related to color when the answer is “red”, whereas $p(t_i = \text{“action”}|a_j = \text{“red”}) \approx 0$, suggesting a very low probability of the question type being associated with “action” in the presence of the same answer. On the other hand, as illustrated in Figure 3, one might anticipate $p(t_i = \text{object”}|a_j = \text{car”}) \approx 0.7$ and $p(t_i = \text{location”}|a_j = \text{car”}) \approx 0.3$, as the answer “car” could plausibly be associated with both object and

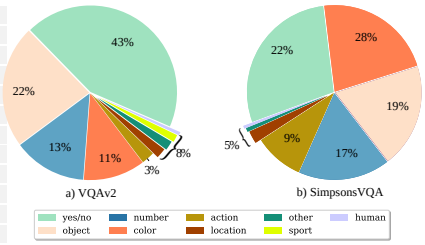
Dataset	VQA 2.0	SimpsonsVQA
Training set		
#images	62,456	13,936
#QA Pairs	335,454	60,643
Validation set		
#images	40,566	3,451
#QA Pairs	214,099	9,764
Test set		
#images	20,432	5,409
#QA Pairs	108,443	10,507

(a) Datasets statistics.

Your task is to categorize the last question below into one of the following question types: {color, number, object, yes/no, action, location, sport, human, other}.

Question: What color are the clouds?
Question Type: color
Question: What is the person doing?
Question Type: action
Question: How many people are there?
Question Type: number
Question: is the person male or female?
Question Type:

(b) Example of question classification Prompt.



(c) Question type distribution.

Figure 4: Description of datasets.

location categories within the dataset. This probability is estimated as the ratio of the occurrences where a_j is paired with t_i in $\mathcal{D}$ to the total occurrences of a_j in $\mathcal{D}$.

3.2.3 Answer relevance estimation. The matrix M serves to filter and select answers according to their relevance to the identified question type. To achieve this, we employ the probability distribution vector $\hat{\mathbf{y}}_{qt}$ and project it onto an answer selection and relevance vector $\mathbf{r}$ as follows:

$$\mathbf{r} = \hat{\mathbf{y}}_{qt} \times M \tag{1}$$

where $\mathbf{r} \in \mathbb{R}^k$ is a k-dimentional vector, indicating the weighted relevance of each answer to the question, based on the classified question type. Hence, the higher the value of r_i is, the more relevant the i-th answer is to the predicted question type.

3.2.4 Topic-aware Selection. Given the unnormalized prediction vector of logits $\mathbf{z}_{VL}$ from the vanilla VQA network, where each element reflects the initial confidence in the corresponding answer, we integrate the contextual relevance vector $\mathbf{r}$ by performing a Hadamard product (element-wise multiplication), resulting in a modified logit vector $\mathbf{z}_{VL-qt}$, as follows:

$$\mathbf{z}_{VL-qt} = \mathbf{z}_{VL} \odot \mathbf{r} \tag{2}$$

which will ensure that the influence of each answer's relevance to the question type is accounted for in the final confidence levels. Finally, to transform these modified confidence levels into a probabilistic distribution, conducive to answer selection, we apply a temperature softmax function to the vector $\mathbf{z}_{VL-qt}$ in order to adjust the sensitivity of the distribution, resulting in:

$$\hat{\mathbf{y}}_a = Softmax_\tau(\mathbf{z}_{VL-qt}) \tag{3}$$

where, $\hat{\mathbf{y}}_a$ is the resultant probability vector, with each element $\hat{y}_{ai}$ representing the final probability that the i-th answer is the correct response to the input question. The softmax temperature parameter τ is a hyperparameter that needs to be tuned on the validation set in order to control the degree of smoothing in the probability distribution.

3.2.5 Loss function. The model is trained by jointly minimizing the two objective functions for answer prediction and question type prediction as follows:

$$\frac{1}{m}\sum_{j=1}^{m}\left[\mathcal{L}_1(\hat{\mathbf{y}}_a^{(j)}, \mathbf{y}_a^{(j)}) + \mathcal{L}_2(\hat{\mathbf{y}}_{qt}^{(j)}, \mathbf{y}_{qt}^{(j)})\right] \tag{4}$$

where $\mathcal{L}_1(\bullet, \bullet)$ denotes the binary cross-entropy loss function to train a k-way classifier [19, 35]; $\mathcal{L}_2(\bullet, \bullet)$ denotes the cross-entropy function.

In summary, TASOL introduces a novel approach to enhance answer selection in VQA systems by incorporating contextual relevance through question type classification and answer relevance estimation.

4 EXPERIMENTAL SETUP

In this section, we describe the experimental setup we use in our evaluations, including a description of the datasets, the metrics used, and the details of our implementation.

4.1 Datasets

We evaluate the proposed TASOL layer on two VQA datasets, which we briefly describe below.

VQA v2.0 [5]: is a popular dataset for evaluating VQA models. It comprises human-annotated question-answer pairs related to images from the MS-COCO dataset [36]. Given that the test set of VQA v2.0 is not publicly available, we partition the training set into a new training set and a new test set. From the original training set, we select 108,124 questions associated with 20,000 unique images to build the test set. For training, evaluation, and test, we exclusively utilize QA pairs comprising the 1,000 most frequent answers, which encapsulates 85% of the dataset.

SimpsonsVQA [37]: is a novel VQA dataset, fully automated in its creation. Images for the dataset are extracted from one of the most popular cartoon series, *The Simpsons*. The question-answer pairs are generated using VLP and GPT-3 [38]. Each question-answer pair is judged by three different individuals on AMT.

Details of train/val/test sets used for each dataset are given in Table 4a.

4.2 Question Type Extraction

Determining the format of a question is crucial in guiding the expected answer in TASOL. Language biases set aside, the nature of

the question significantly influences the answer. For instance, questions focused on quantity, typically phrased as *"how many..."*, predominantly receive numerical responses. Similarly, questions that starts with *"is/are..."* are usually answered with binary responses like "yes" or *"no"*. However, this paradigm shifts for high-context questions like "is this the man or the woman wearing the red hat?", where the expected answer transitions from a simple binary choice to a more descriptive "man" or "woman". This nuanced approach to question analysis aims to address the limitations inherent in datasets such as the TDIUC, where question types are rigidly defined by templates, and in VQA v2.0, which categorizes questions based solely on their initial words. Such methodologies may not capture the complexities of varied question formats adequately.

We leverage GPT-3's [38] (version *gpt-3.5-turbo-0125*) few-shot learning capabilities, particularly 3-shot learning, to surpass existing limitations in question type categorization for VQA systems. This method allows us to consider the broader context of queries, moving beyond rigid templates and simple cues to classify questions into nine types: *color*, *number*, *object*, *yes/no*, *action*, *location*, *sport*, *human*, and *other*. This strategy, detailed in Figure 4b, aims to enhance VQA system efficacy and applicability by achieving more nuanced and accurate question type extraction. Finally, the distribution of question types in our two datasets is given in Figure 4c.

4.3 Metrics

The performance is measured using two evaluation metrics: (i) *Recall*, employed to assess answer consistency with respect to the asked question, indicating the proportion of image-question pairs for which the provided answer aligns with the expected question type out of all true answers belonging to that type; and (ii) *Accuracy*, utilized to evaluate the overall performance of the evaluated models. It is worth noting that for VQA v2.0, we employ the accuracy metric proposed in [15].

4.4 Baseline methods

We evaluate TASOL using 6 vanilla models with varying levels of complexity.

LSTM_CNN: This model utilizes a deep LSTM network comprising two hidden layers to convert questions into intricate 1024-dimensional embeddings, coupled with a CNN for extracting visual features from images. Subsequently, the model combines two crucial multimodal features through multiplication.

LSTM_VGG [15] : This model utilizes a deep LSTM network to convert questions into comprehensive 1024-dimensional embeddings and employs a pre-trained, frozen VGG16 [39] network on on the ImageNet dataset to extract visual features from images. Following their independent processing, the feature from the question embedding and the visual feature from the image analysis are combined with an element-wise multiplication.

LSTM_RESNET : This model is similar to **LSTM_VGG** , butwe replace VGG16 with a pre-trained, frozen ResNet network [40].

Stacked Attention Networks (SAN) [16]: A pre-trained VGG16 model is utilized to extract visual features from input images, while questions are encoded into semantic vectors using an LSTM network. This implementation incorporates two sequential attention layers, which iteratively refine the model's focus on image regions most pertinent to the question by combining the LSTM-encoded question vector with the VGG16-extracted visual features. Subsequently, the refined visual context, refined through the application of two stacked attention mechanisms, is integrated with the question representation to facilitate predictions.

Bottom-Up Top-Down Attention (BUTD) [41]: In this model, a Faster R-CNN model [42] is employed, pre-trained on the Visual Genome [43] dataset and frozen to detect up to 36 distinct objects within an image, effectively capturing rich, object-level visual features. For encoding the question, a GRU (Gated Recurrent Unit) [44] is used, which processes the textual input to produce a semantic representation. Subsequently, Top-Down attention leverages the question to dynamically focus on specific objects among the 36 extracted ones that are most relevant to the query.

Deep modular co-attention networks (MCAN) [19]: Drawing inspiration from Transformers [18], MCAN introduces a novel approach by incorporating a series of Modular Co-Attention layers. These layers employ both self-attention and guided-attention mechanisms to effectively analyze and integrate multimodal inputs. Each Modular Co-Attention layer processes the inputs through two basic attention units, allowing the model to simultaneously attend to relevant parts of the image and question, thus enhancing the overall accuracy of the system. This paper implements only the encoder-decoder architecture. Furthermore, similar to BUTD, we utilize only 36 predefined fixed objects extracted from a pretrained Faster R-CNN.

4.5 Implementation details

In TASOL, we employ a Bi-LSTM for Question Type prediction. Also, all question encoders in the backbone VQA models utilize GloVE embeddings [45] with 300 dimensions as the initial weights. These models are trained from scratch, without relying on pre-trained models, using the Adam optimizer [46] for 50 epochs. A linear learning rate scheduler is employed to manage learning rate adjustments throughout training. The scheduler initializes with a starting learning rate of 10^-4 and gradually reduces it to 10^-5 by the end of the training period. The implementation is based on PyTorch version 1.12.1 and conducted on a Linux Ubuntu 18.04.1 LTS system with a Dual Intel(R) Xeon(R)350 350 Silver CPU @2.20GHz and a NVIDIA Tesla V100 GPU.

Finally, in the case of VQA v2.0, we consider 9 question types: *color*, *number*, *object*, *yes/no*, *action*, *location*, *sport*, *human*, and *other*. However, for SimpsonsVQA, we exclude the *sport* category due to the scarcity of questions falling into this type.

5 EXPERIMENTAL EVALUATION

In this section, we report and discuss the main results we obtained in an offline evaluation and then a comparison with the baselines.

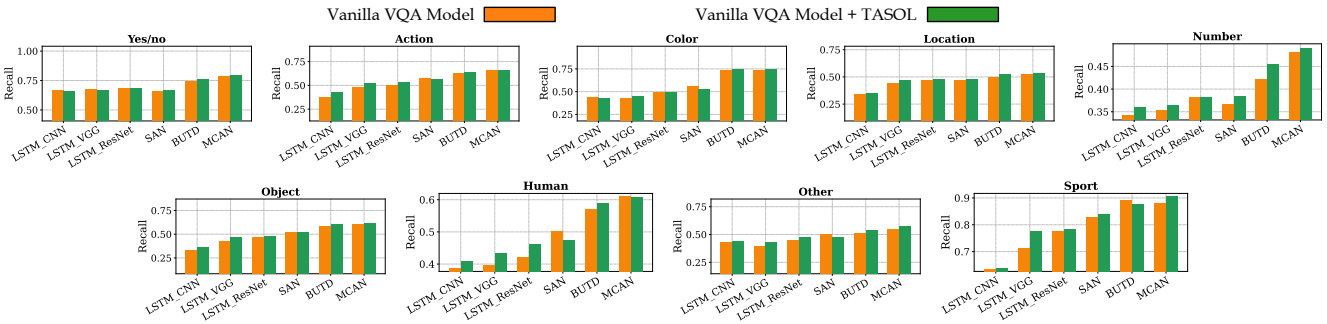


Figure 5: Recall for each question type, indicating the assessment of answer consistency with respect to the asked question on VQA v2.0.

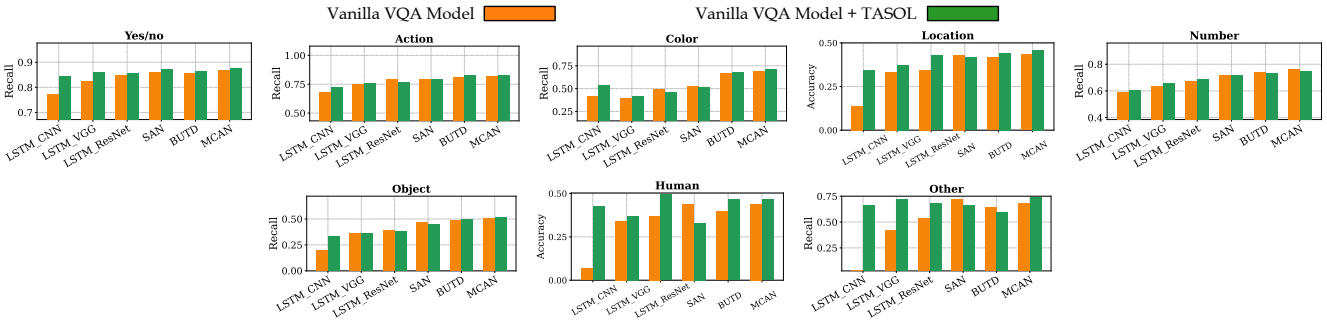
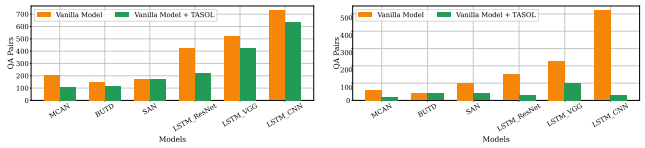


Figure 6: Recall for each question type, indicating the assessment of answer consistency with respect to the asked question on SimpsonsVQA .



(a) VQA v2.0. **(b) SimpsonsVQA.**

Figure 7: Number of inconsistent QA pairs.

5.1 Consistency analysis

In Figures 5 and 6, respectively for VQA v2.0 and SimpsonsVQA, the Recall values are displayed for each question type, indicating the assessment of answer consistency with respect to the asked question. Briefly, from the obtained results we make the following observations:

- First, we note that for both datasets and nearly all question types, consistency, as measured by recall, exhibits improvement.
- Second, we observe that in some cases, there is a substantial improvement in consistency. For instance, simple and lightwise models like **LSTM_CN**, **LSTM_VGG**, and **LSTM_ResNet** demonstrate significant enhancements for "Human" type questions on VQA v2.0. These models offer efficiency, making them suitable for deployment on resource-constrained devices where simpler models are more practical, albeit with slightly reduced prediction accuracy. In such scenarios, TASOL can serve as a valuable tool to enhance their accuracy and overall performance.
- Third, we also observe that even sophisticated models like **MCAN** or **BUTD** benefit significantly from TASOL in enhancing their answer consistency.

Finally, we present the raw number of inconsistent QA pairs in Figure 7 for each dataset and Vanilla VQA model, both with and without TASOL. Overall, we observe that TASOL significantly reduces the number of inconsistent responses, especially evident in the SimpsonsVQA dataset. For instance, the total number of inconsistent answers drops from about 500 to less than 50 for the **LSTM_CNN** model, from about 200 to 100 for **LSTM_VGG**, and from 100 to 50 for **SAN**.

5.2 Performance analysis

The overall accuracy performance is presented in Table 1. It is noteworthy that in numerous instances, the accuracy improves when TASOL is integrated with the vanilla VQA models, with some cases showing substantial enhancements. For example, there is a notable improvement observed for **LSTM_CNN** on the SimpsonsVQA dataset. These results demonstrate that TASOL not only enhances model consistency but also improves the accuracy of the final answers provided.

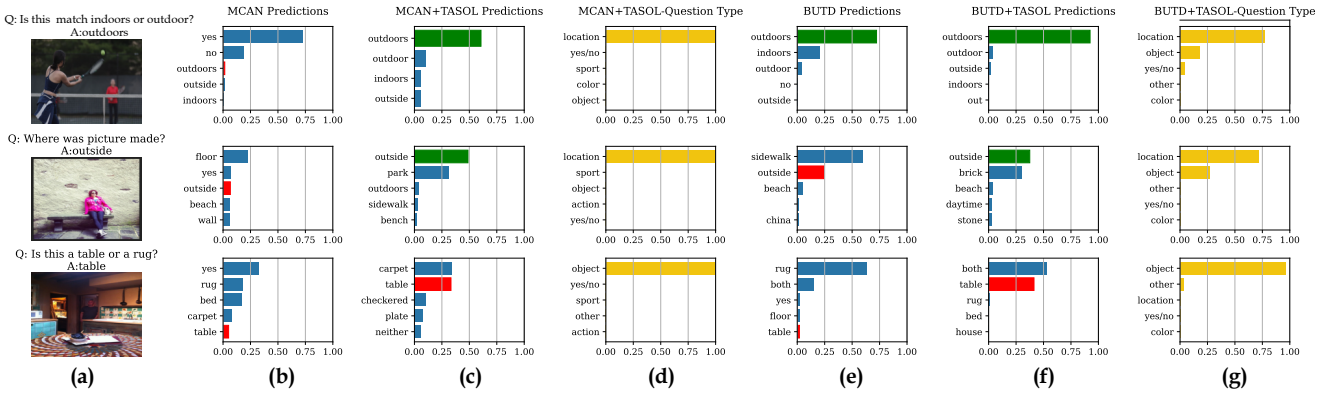


Figure 8: Case studies on VQA v2.0. Each row represents a testing case. Column (a): test question and image with ground truth answer. Column (b): top 5 guesses from the MCAN. Column (c): top 5 guesses from MCAN + TASOL. Column (d): top 5 guesses from TASOL for the question type. Columns (e)-(g): same as columns (b)-(d) using BUTD.

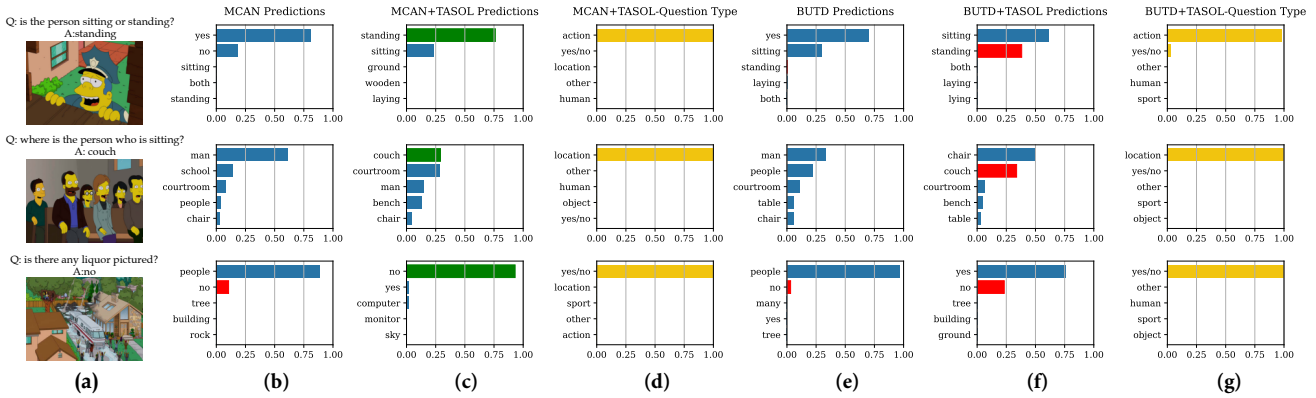


Figure 9: Case studies on SimpsonsVQA. Column (a): test question and image with ground truth answer. Column (b): top 5 guesses from the MCAN. Column (c): top 5 guesses from MCAN + TASOL. Column (d): top 5 guesses from TASOL for the question type. Columns (e)-(g): same as columns (b)-(d) using BUTD.

Table 1: Accuracy on VQA v2.0 and SimpsonsVQA.

	VQA v2.0		SimpsonsVQA	
Name	**Vanilla Model**	**Vanilla Model + TASOL**	**Vanilla Model**	**Vanilla Model + TASOL**
MCAN	0.6860	**0.6948**	0.72095	**0.7317**
BUTD	0.6525	**0.6720**	0.7037	**0.7115**
SAN	0.5690	**0.5693**	**0.6654**	0.6626
LSTM_ResNet	0.5578	**0.5608**	**0.6280**	0.6250
LSTM_VGG	0.5324	**0.5450**	0.5838	**0.6052**
LSTM_CNN	0.5012	**0.5092**	0.5173	**0.6108**

5.3 Case studies

We aim to explore how TASOL rectifies the errors made by vanilla VQA models. In Figures 8 and 9, we present four testing cases for the VQA v2.0 and SimpsonsVQA datasets, respectively. For instance, in the first case depicted in Figure 8, MCAN fails to provide the correct answer, responding with "yes" to the question "Is this match indoors or outdoors?" This inconsistency could undoubtedly confuse the user, as the answer does not pertain to a location. In contrast, the model integrated with TASOL successfully discerns the expected location-based answer and accurately provides "outdoor" as the response to the question. Furthermore, in the final case illustrated in Figure 9, when asked "Is there liquor pictured?" BUTD provides both an inconsistent and incorrect answer. However, when coupled with TASOL, while the answer remains incorrect, it maintains consistency, offering a more coherent response. We argue that coherence and consistency are essential aspects of an effective VQA system, as they enhance user comprehension and trust in the system's outputs.

6 TASOL USAGE

TASOL[1] is a straightforward output layer that seamlessly integrates into existing models, making it easy to incorporate any architecture for improved performance. In Figures 10 and 11, we present a code snippet demonstrating how TASOL can be seamlessly integrated into a vanilla VQA model using respectively Keras and PyTorch.

[1]The code and implementation are hidden to maintain submission anonymity but will be made publicly available upon publication of the paper.

```
# Define the input shapes
q_input = Input(shape=(100,), name='q_input')  # assuming question is represented as a sequence of word embeddings
img_input = Input(shape=(224, 224, 3), name='img_input')  # assuming image is RGB with size 224x224

# Vanilla VQA Model processing
embedding = VanillaVQA()([q_input,img_input])
output_logits = Dense(num_classes, name='output_logits')(embedding)

# Instantiate the TASOL layer which takes as input the matrix M and the softmax temperature paramter t
tasol_output = TASOL(name='tasol_output', matrix_m, temperature_value=t)(output_logits, question_input)

# Create the model
model = Model(inputs=[q_input, img_input], outputs=[tasol_output])

# Compile the model
model.compile(optimizer='adam', loss='categorical_crossentropy', metrics=['accuracy'])
```

Figure 10: TASOL Usage example in Keras.

```
# Instantiate the Vanilla VQA model
vanilla_vqa_model = VanillaVQA()

# Define the input shapes
q_input = torch.Tensor(100)  # Assuming question is represented as a sequence of word embeddings
img_input = torch.Tensor(3, 224, 224)  # Assuming image is RGB with size 224x224

# Forward pass through Vanilla VQA model
embedding = vanilla_vqa_model(q_input, img_input)

# Define the output logits layer
output_logits = nn.Linear(in_features=num_classes, out_features=num_classes)(embedding)

# Instantiate the TASOL layer which takes as input the matrix M and the softmax temperature paramter t
tasol_layer = TASOL(matrix_m, temperature_value=t)

# Forward pass through TASOL layer
tasol_output = tasol_layer(output_logits, q_input)

# Define loss function and optimizer
criterion = nn.CrossEntropyLoss()
optimizer = torch.optim.Adam(model.parameters())
```

Figure 11: TASOL Usage example in PyTorch.

7 CONCLUSION AND FUTURE WORK

In this paper we proposed TASOL, a model agnostic Topic-Aware Selection Output Layer that aims to reduce inconsistent predictions in VQA models, which can undermine user trust and comprehension. By embedding a structure that aligns question types with answer space, TASOL effectively mitigates inconsistency thereby enhancing coherence and reliability in VQA systems. Through rigorous experimentation on diverse datasets, including VQA2.0 and SimpsonsVQA, we have demonstrated TASOL's efficacy in improving accuracy and consistency across various metrics. Thus, this work not only highlights an important issue in existing VQA methodologies but also presents a robust solution that contributes to advancing the field towards more reliable and trustworthy interactions between humans and AI systems. Future work could focus on refining TASOL through additional experimentation and exploration of its applicability in more diverse and complex VQA scenarios.

REFERENCES

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[2] Mengye Ren, Ryan Kiros, and Richard Zemel. Exploring models and data for image question answering. *Advances in neural information processing systems*, 28, 2015.

[3] Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Yin and yang: Balancing and answering binary visual questions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[4] Lin Ma, Zhengdong Lu, and Hang Li. Learning to answer questions from image using convolutional neural network. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[5] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017.

[6] Sadid A Hasan, Yuan Ling, Oladimeji Farri, Joey Liu, Henning Müller, and Matthew P Lungren. Overview of imageclef 2018 medical domain visual question answering task. In *CLEF (Working Notes)*, 2018.

[7] Jason J. Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific Data*, 5, 2018.

[8] Asma Ben Abacha, Vivek V Datla, Sadid A Hasan, Dina Demner-Fushman, and Henning Müller. Overview of the vqa-med task at imageclef 2020: Visual question answering and generation in the medical domain. In *CLEF (Working Notes)*, 2020.

[9] Zhihong Lin, Donghao Zhang, Qingyi Tac, Danli Shi, Gholamreza Haffari, Qi Wu, Mingguang He, and Zongyuan Ge. Medical visual question answering: A survey. *arXiv preprint arXiv:2111.10056*, 2021.

[10] Haifan Gong, Guanqi Chen, Sishuo Liu, Yizhou Yu, and Guanbin Li. Cross-modal self-attention with multi-task pre-training for medical visual question answering. In *Proceedings of the 2021 International Conference on Multimedia Retrieval*, ICMR '21, page 456–460, New York, NY, USA, 2021. Association for Computing Machinery.

[11] Shurong Sheng, Luc Van Gool, and Marie-Francine Moens. A dataset for multimodal question answering in the cultural heritage domain. In *Proceedings of the COLING 2016 Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 10–17. ACL, 2016.

[12] Noa Garcia, Chentao Ye, Zihua Liu, Qingtao Hu, Mayu Otani, Chenhui Chu, Yuta Nakashima, and Teruko Mitamura. A dataset and baselines for visual question answering on art. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 92–108. Springer, 2020.

[13] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443, 2019.

[14] Daniel Gordon, Aniruddha Kembhavi, Mohammad Rastegari, Joseph Redmon, Dieter Fox, and Ali Farhadi. Iqa: Visual question answering in interactive environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[15] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*, 2015.

[16] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[17] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. *Advances in neural information processing systems*, 29, 2016.

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

[19] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6281–6290, 2019.

[2] https://creativecommons.org/licenses/by-nc-sa/4.0/

[20] 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.

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

[22] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[23] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022.

[24] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022.

[25] Kushal Kafle and Christopher Kanan. An analysis of visual question answering algorithms. In *Proceedings of the IEEE international conference on computer vision*, pages 1965–1973, 2017.

[26] Yang Shi, Tommaso Furlanello, Sheng Zha, and Animashree Anandkumar. Question type guided attention in visual question answering. In *proceedings of the European conference on computer vision (ECCV)*, pages 151–166, 2018.

[27] Chao Yang, Mengqi Jiang, Bin Jiang, Weixin Zhou, and Keqin Li. Co-attention network with question type for visual question answering. *IEEE Access*, 7:40771–40781, 2019.

[28] Aakansha Mishra, Ashish Anand, and Prithwijit Guha. Cq-vqa: Visual question answering on categorized questions. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2020.

[29] Aakansha Mishra, Ashish Anand, and Prithwijit Guha. Dual attention and question categorization-based visual question answering. *IEEE Transactions on Artificial Intelligence*, 4(1):81–91, 2022.

[30] Jiaxian Guo, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Boyang Li, Dacheng Tao, and Steven Hoi. From images to textual prompts: Zero-shot visual question answering with frozen large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10867–10877, 2023.

[31] Yueting Yang, Xintong Zhang, and Wenjuan Han. Enhance reasoning ability of visual-language models via large language models. *arXiv preprint arXiv:2305.13267*, 2023.

[32] Zhenwei Shao, Zhou Yu, Meng Wang, and Jun Yu. Prompting large language models with answer heuristics for knowledge-based visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14974–14983, 2023.

[33] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[34] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[35] Damien Teney, Peter Anderson, Xiaodong He, and Anton Van Den Hengel. Tips and tricks for visual question answering: Learnings from the 2017 challenge. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4223–4232, 2018.

[36] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[37] Ngoc Dung Huynh, Mohamed Reda Bouadjenek, Sunil Aryal, Imran Razzak, and Hakim Hacid. SimpsonsVQA: Enhancing Inquiry-Based Learning with a Tailored Dataset. Technical report, Deakin University, Jan 2024.

[38] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[39] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[40] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

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

[42] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

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

[44] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[45] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics.

[46] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.