ISSN 1816-6075 (Print), 1818-0523 (Online)
Journal of System and Management Sciences
Vol. 12 (2022) No. 3, pp. 428-454
DOI:10.33168/JSMS.2022.0320

Zero-Shot Visual Question Answering based on DataSet Redistribution

Samar I. Zekrallah[1], Aboul Ella Hassanin[2], Nour Eldeen Mahmoud[1]

[1]Faculty of Computers and Information, Information Technology Department, Cairo University. Cairo, Egypt

[2]Scientific Research Group in Egypt, Cairo, Egypt

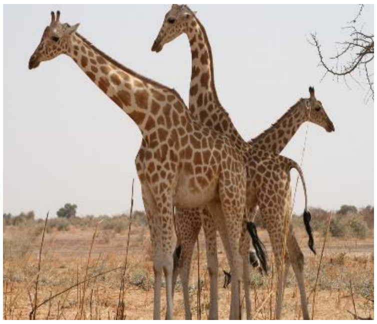
aboitcairo@gmail.com

Abstract. Visual Question Answering is an extremely active research area in which the computer is given an image, a question in natural language, and it is required to give a correct answer to the question according to the semantics of the input image. The ability of VQA system to answer new questions about unseen images during training process is one main measure of effectiveness of the VQA model and this capability is called Zero-Shot VQA, but VQA datasets suffer from some problems that hinder good evaluation on models trained on these datasets. Firstly, Testing instances are not chosen perfectly to address how much the trained model accomplish the task of asking about new concepts that is not presented during training process. Secondly, most of visual question answering datasets suffer from problems in their contents such as small dataset size, leakiness of explicitly defined question types, and question types have abused evaluation scores that makes it difficult to evaluate algorithms on them. So models are not perfectly evaluated on such datasets. In order to avoid those evaluation obstacles, experiment is done on TDIUC dataset which has explicitly defined 12 question types, data are redistributed for zero shot task by re-splitting it to new training, val, and test instances such that test instances contains new concepts that is not presented in training data. Evaluation is done using methods that give a more representative measure of accuracy over all question types(Simple Accuracy, AMPT, HMPT) and one more evaluation schema(GMPT) is proposed for evaluating accuracy which is more expressive. Experiment shows that evaluation results on TDIUC dataset before redistributing train, val, and test sets for Zero Shot purpose gives inaccurate indicator of model performance (around 20% higher performance)

Keywords: Zero shot, artificial intelligence, deep learning, image processing, natural language processing

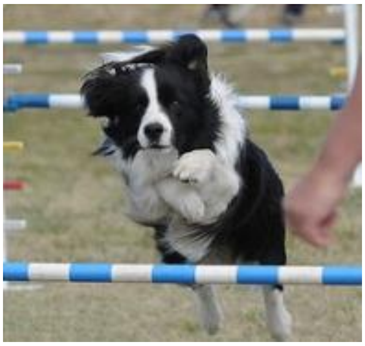
1. Introduction

Matured research in computer vision (CV) and Natural Language Processing (NLP) using deep learning approaches encouraged the researchers to advance from just solving low level AI tasks such as image classification, Object detection, and activity recognition to higher level AI tasks such as Image Captioning where the goal is to predict a one sentence description for the given image, answering reading comprehension questions by understanding short stories, visual question answering, text-to-image retrieval, and Visual dialog (in which given an image, a dialog history consists of questions, answers pairs, and followed by a question then the system generates a free form natural language answer) (Barra et al., 2021).VQA is more challenging problem than image captioning because VQA questions cannot be fixed and so the operations required to answer the question, VQA requires solving many computer vision subtasks, and asked questions may require common sense knowledge (Manmadhan et al., 2020). Figure 1 and 2 show the difference between VQA and image captioning.



Question: How many animals are in the picture?
Answer: Three

Fig. 1: Visual question answering



"Black and White Dog jumps over bar."

Fig. 2: Image captioning

Question Answering is one of the most challenging widely investigated problems in Natural Language Processing (NLP) and it is recently used to develop dialog systems, humanoid robots (Budiharto et al., 2020), and chat-bots. Recent developments in deep learning, neural network models have shown promising advancements for language modeling and moving from using sequential models to transformers models (Vaswani et al., 2017) and attention mechanisms like BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018) which was published at the end of 2018 by researchers of Google AI Language. BERT has outperformed other models in NLP field and reach state of the art for modeling language-based tasks.

On the other hand, Visual Question Answering (VQA) is one of the high level AI tasks and is considered a multi-discipline research problem in which given an image and a natural language question as input, the VQA System should be able to generate a natural language answer for any question type related to that image.

VQA differs from text-based Q&A. In Text-based Q&A, NLP is used for question categorization, extracting the objects in the question, and finding the right context for answering the question. After analyzing the question, the system builds a query and uses a knowledge base to get the answer. But A VQA system performs the reasoning and answer generation based on the content of the image. Questions can be arbitrary and contains many sub-problems needs to be solved in order to generate the appropriate answer such as Object recognition, Object detection, Attribute classification, Scene classification, Counting, common sense reasoning, Knowledge base reasoning, and spatial relationships between objects in image (Manmadhan et al., 2020). So computer vision (CV) is combined with Natural language processing (NLP) for accomplishing VQA tasks.

For example, if the image in Figure 3 is given to VQA System and the system is asked the following questions:

- Are there cars in the image?
- Is it raining?
- How many persons in the image?

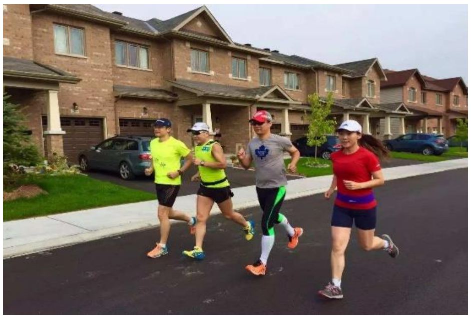


Fig. 3: Example of image and questions given to VQA system

The asked questions may require low level computer vision tasks like the first and third question or may require common sense knowledge like the second one. However, with the advancements in deep learning, systems became more capable of answering such questions. Solving the VQA problem is an important step toward human level understanding and it is considered “AI-Complete” task. In fact, the problem has also been suggested to be used as a Visual Turing Test by Geman et al., (2015).

There are many potential applications for VQA [2], including text-image retrieval (Zhang et al., 2020) which can be used in enhancing online shopping by selecting the most related images to the searched text, educational purposes, an aid to blind and visually impaired individuals as it provides user-specific information through scene understanding, medical tasks as in (Abacha et al., 2020) where answering questions from the visual content of radiology image and Video Surveillance.

The rest of this paper is organized as follows: Section II discusses background and some VQA related work, Section III shows the different datasets concerning VQA, Section IV clarifies the proposed Zero-Shot TDIUC dataset, Section V shows the measured accuracies after running experiment on original TDIUC and Zero-Shot TDIUC datasets, Section VI explains the used evaluation methods and a comparison between testing using zero- shot redistributed version of TDIUC dataset and the original dataset without re-distribution of data and Finally, conclusion of the paper.

2. Background and Related Work

In general, Approaches that are used to solve VQA problems work as follows :Given an Image *I*, Question *Q*, and Answer *A* : 1) Extract features from the image $I_{features}$, 2) Get sentence embedding for the question $Q_{features}$, 3) Combine $I_{features}$ and $Q_{features}$, and 4) Generate correct answer.

For image featurization, using pre-trained CNNs with their last layer removed has shown great results in extracting image features and doing it easily instead of training it from the scratch. ResNet (He et al., 2016) and VGG-Net (Simonyan et al., 2014) are mostly used for this task as indicated in Manmadhan et al., (2020).

Different approaches for text featurization have been used as (1) Count based methods like one-hot encoding, co-occurrence counts (Miller et al., 1991), and co-occurrence matrix with SVD (Eckart et al., 1936) to reduce dimensionality by k-rank approximation, (2) Prediction based methods like CBOW, and Skip-gram (Mikolov et al., 2013), (3) Hybrid methods like GloVe (Pennington et al., 2014), (4) CNNs like long short term memory (LSTM) which is mostly used by researchers as stated in Young et al., (2018) because it gives better results other than word sequence independent models like word2vec, gated recurrent unit (GRU) (Cho et al., 2014), and (5)Transformer based models like BERT, ELMO (Peters et al., 2018), GPT (Radford et al., 2018).

There are different methods used for combining image and question features ranging from simple baseline fusion models like concatenation (Huang et al., 2018), element-wise multiplication or addition (Antol et al., 2015; Goyal et al., 2017; Teney et al., 2018) to End-to-End neural network models and joint attention models like neural module networks (NMN) (Andreas et al., 2016), Multimodal Compact Bilinear Pooling(MCB) (Fukui et al., 2016), LXMERT (Tan et al., 2016) which has one additional novel cross modality encoder to connect vision and language semantics, ViLBERT (Lu et al., 2019), VisualBERT (Geman et al., 2015).

There are different paradigms for answer generation, it can be treated as a classification problem in which a set of the most frequent answers are defined as possible outputs and the model is trained to choose the most ranked answer as the solution to the given image and question, examples of models used this approach are in Ren et al., (2015), Fukui et al., (2016), Zhou et al., (2015), and Zhu etal., (2016).

The Second paradigm is to be treated as a generation model in which an open-ended question and image is given to the system and it is able to generate an answer as in Malinowski et al., (2015), these models (Wu et al., 2017) may use external knowledge bases to get answers for questions that is their answers are not presented in the image as in Wu et al., (2016), the authors used Pre-trained VGG16 for image processing, DBpedia (Auer et al., 2007) as external source. The external knowledge is encoded using Doc2Vec. Question vector along with textual information of image given to Encoder-Decoder based LSTM for answer generation.

Although the first approach is easier, it is restricted by the set of answers that has been seen during the training process and it is unable to generate new answers.

In Zhu et al., (2016), Chen et al., (2015), Kazemi et al., (2017), attention-based models were introduced and they recorded better performance. In these models instead of using global image features for predicting answer, only local image features that are related to question are used. This approach avoids using unnecessary data or noisy information from image and thus increasing performance. Cao et al. (2017) introduced semantic cross correlation between image and question besides attention. Also question type information has been used to extract image regions that is more related to the question as stated by Shi et al., (2018) and proved its usefulness in improving VQA. Question type is also used to narrow down answer search space by Misha et al., (2020)

Although proposed models showed promising performance results, the evaluation metrics have some biasness due to dataset language bias and also the answers in the entire dataset come from a small set of vocabulary (i.e. they follow long-tail distribution) as seen in Figure 4.

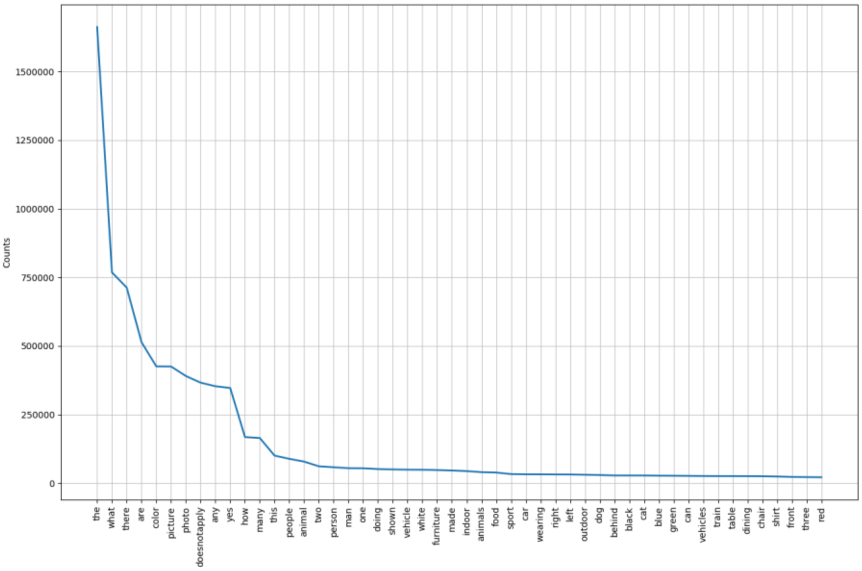
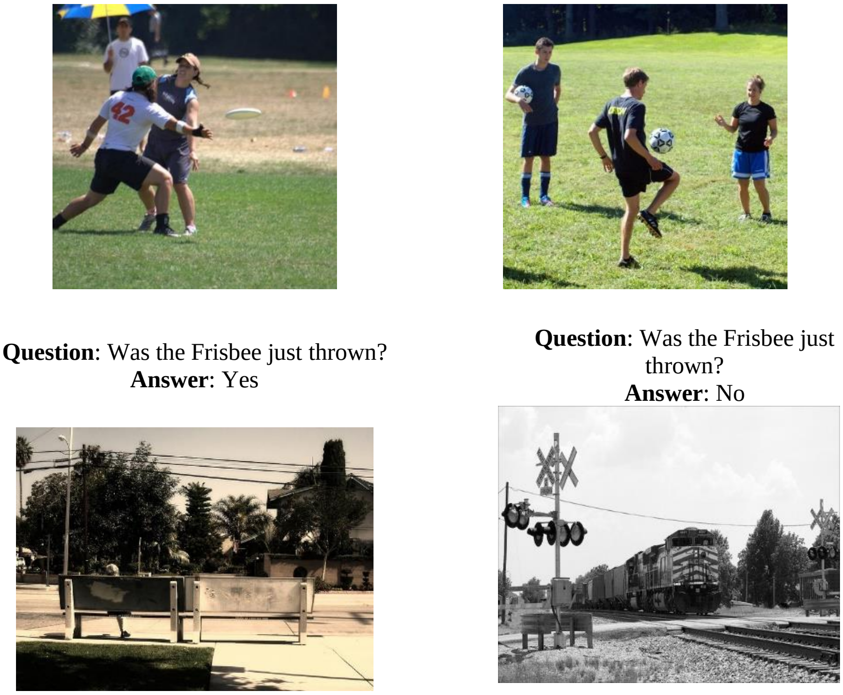


Fig. 4: Frequency distribution of top 50 words in TDIUC questions & answers

So there is a need to better evaluate models on more balanced datasets. Goyal et. al (201) introduced the VQA dataset in a new version, in which every question from the earlier version of VQA dataset is related to another similar image that differs a little from the second image such that the answer to the same question is different. Two examples of these complementary pairs are shown in Figure 5 with 2 different question types.

Teny et al., (2016), proposed Visual7W dataset in zero-shot aspect by re-splitting the data set into new training, validation, and test splits in order to have better evaluation of the VQA System's capability of generalization to questions. Zero-shot VQA as mentioned by Teney is inspired from Zero Shot Classification in order to measure the VQA model's capability of answering questions about unseen concepts during the training process making the validation/test splits contains at least one word in the question, answer, or both of them that has not been seen by the model during training.



Question: Was the Frisbee just thrown?
Answer: Yes

Question: Was the Frisbee just thrown?
Answer: No

Question: How many benches are there?
Answer: 2

Question: How many benches are there?
Answer: 1

Fig. 5: VQA v.2 complementary pairs

3. Datasets

The VQA field is so complex that a good dataset should be large enough to capture the long range of possibilities within questions and image content in real world scenarios.

In 2014, DAQUAR [40] (DAtaset for QUestion Answering on Real-world images), The First VQA dataset was released. It used images from the NYU-Depth v2 dataset and had 12,468 question/answer pairs on 1,449 images. However its limited size, It encouraged the research in image question answering area. VQA models are trained and evaluated on it. Many of the images used in VQA datasets are derived from Microsoft Common Objects in COntext (MSCOCO), which is a large-scale object detection, segmentation, and captioning dataset with 91 different object categories.

Many other datasets were introduced like COCO-QA (ren et al., 2015), VQA v.1 (Antol et al., 2015) which was released in 2015 and VQA v.2 (balanced version of VQA v.1) which was released in 2017 (Goyal et al., 2017) (There are 2 subsets of VQA dataset: one with real images which is called COCO-VQA, and the other with abstract scenes which is referred to as SYNTH-VQA), FM-IQA (Gao et al., 2015), VisualGenome (Krishna et al., 2016), Visual7w (Zhu et al., 2016) which is a subset of VisualGenome and has an additional visual groundings which is used in answering pointing QA and enable more visual reasoning.

In 2017, A Dataset called Task Driven Image Understanding Challenge (TDIUC) was introduced by Kushal Kafle and Christopher Kanan (2017) for better analysis of VQA models. TDIUC contains 1,654,167 open-ended question-answer pairs, of which 1,115,299 are to be used for training and 538,868 for validation/testing. These questions are derived from 3 distinct sources (COCO-VQA and Visual Genome, auto-generated from Visual Genome’s objects and attributes annotations and COCO’s semantic segmentation annotations and Human annotators) , and are organized into 12 different categories as the following:

- Object Presence (e.g., “Is there a hot dog in the picture?”)
- Subordinate Object Recognition (e.g., “what type of fruit is in the image?”)
- Counting (e.g., “How many people are there?”)
- Color Attributes (e.g., “What color is the man’s t-shirt?”)
- Activity Recognition (e.g., “What is the man doing?”)
- Sport Recognition (e.g., “What game is the man playing?”)
- Scene Classification (e.g., “Is the picture taken outdoors?”)
- Object Utilities and Affordances (e.g., “What object in the picture can be used for transportation?”)
- Positional Reasoning (e.g., “What is behind the man?”)
- Sentiment Understanding (e.g., “Is the player sad?”)
- Other Attributes (e.g., “What is the table made of?”)
- Absurd (i.e., is not answerable based on the image’s content)

TDIUC has an additional absurd question type (i.e., queries not related to image) which requires an algorithm to look at the image in order to determine if the question is appropriate for the image, So it enables judging the model’s ability to differentiate between questions that is related to image content or not.

The authors also proposed two new evaluation metrics to compensate for biasness in dataset to fairly compare different approaches and to determine the empirical limitations of the state-of-the-art and baselines.

In 2018, Samira Ebrahimi et al. (2017) introduced FIGUREQA dataset to study the ability of visual reasoning for scientific figures. Images in FIGUREQA are synthetic, scientific figures which model continuous and categorical information. It includes 5 types of plots (line, dot-line, vertical and horizontal bar graphs and pie charts). FIGUREQA has a balanced ratio of yes/no answers for each question type and figure in order to avoid exploiting biases in answers rather than learning to understand the visual content.

Another dataset which is called DVQA was introduced by Kushal Kafle et al. (2018) to encourage training models to learn extracting numeric and semantic information from bar charts. VQA models that is trained on real images datasets are not capable of answering many questions in DVQA. Easy-VQA is a “Hello World” for VQA, it is a simple dataset that uses simpler images and questions for easier training and evaluation. EasyVQA contains questions about shapes (Circle, Rectangle, Triangle) and colors (black, grey, red, green, blue, yellow, teal, brown), and it has only 13 fixed answers.

Full-Sentence Visual Question Answering (FSVQA) is built based on VQA dataset by using rules in order to convert answers to full sentence answers and there is also an augmented version of FSVQA by applying rules to generate questions that its answers are the MSCOCO captions. FSVQA is more complex in evaluation since other evaluation tools are working in short answers and measured accuracy may be misleading.

VizWiz-VQA dataset is a dataset collected from images which are taken by blind people and questions are recorded about it. This dataset raised one more challenge to predict answer if the visual question is unanswerable (Gurari et al., 2018).

VQA-Med dataset is firstly introduced in 2018 as a pilot task and it is considered the first VQA dataset in the medical domain to encourage exploration of models for automatic medical image interpretation (Hasan et al., 2018). A second edition is released in 2019 (Abacha et al., 2019) includes a training set of 3,200 medical images with 12,792 Question-Answer (QA) pairs, a validation set of 500 medical images with 2,000 QA pairs, and a test set of 500 medical images with 500 questions , it has four question categories(Modality, Plane, Organ System, and Abnormality). In 2021 (Abacha et al., 2021), dataset is released focusing on abnormality questions and contains one more challenging task to be explored which is Visual question generation(VQG) in which questions are generated based on radiology images.

The creation of large and less biased dataset is a key factor in order to assess proposed VQA models and their ability of solve VQA problems. Also the type of images used in the dataset showed different performance with different models.Table1 shows a comparison between some VQA datasets mentioning the ways of collecting question and answer pairs, the type of questions, dataset size in terms of questions and images, and some drawbacks. It is clear that most of the datasets contain biasness in their contents, and lack of well-annotated question types, so good evaluation metrics must compensate for these problems.

Table 1: Comparison between some existing VQA datasets

Dataset	Question/answer Collection	Question Types	# of Questions	# of Images	Drawbacks
DAQUAR	–Generated automatically using predefined templates. –Human Annotators	Colors, Numbers, Objects, or sets of those	12,468	1,449	∗Focus on few prominent objects (Tables, Chairs). ∗Restricted answers (only 16 colors and 894 object categories). ∗Only indoor scenes. ∗Clutter in images and some extreme lightening conditions. ∗Evaluation on humans shows 50.2% accuracy ∗Dataset size is relatively small for training complex models
COCO-QA	Automatically generated from COCO image caption.	Object, Color, Number, or Location	117,684	123,287	∗High repetition rate of questions. ∗Questions may be formulated incorrectly or have grammatical errors. ∗Questions are not equally distributed (69.84% of questions are about objects in image.
COCO-VQA v.1	Amazon Mechanical Truck(AMT)	Yes/No, Number, Other.	614,163	204,721	∗Questions can be answered without looking at image content due to language bias. ∗Stronger dataset bias than SYNTH-VQA. ∗38% of questions' answers are yes/no and around 59% of answers are 'yes' (bias in answers). ∗Questions are too subjective to have a single right answer. ∗Contains questions with unclear answer. ∗Difficult to be used to assess VQA Models. ∗Questions Categories are not explicitly assigned, so performance on each category cannot be measured.
SYNTH-VQA v.1	Amazon Mechanical Truck(AMT)	Yes/No, Number, Other.	150,000	50,000	
COCO-VQA v.2	Amazon Mechanical Truck(AMT)	Yes/No, Number, Other.	1,105,904	204,721	∗Although it is a balanced dataset, it has bias in the distribution of question types and answers within every question type.

Dataset	Question/answer Collection	Question Types	# of Questions	# of Images	Drawbacks
Binary SYNTH-VQA v.2	Amazon Mechanical Truck(AMT)	Yes/No	33,383	31,325	*Models used cannot be extended to real pictures rather than cartoon images.
FM-IQA	From COCO Dataset provided by Amazon Mechanical Truck(AMT)	Action Recognition, Object Recognition, Position, Attributes, Common Sense.	316,193	158,392	*Difficult automatic evaluation because of full sentence answers rather than few words.
Visual Genome	Human Annotators	What, Where, How, When, Who, Why, and Which	1,773,258	101,174	*43 % of answers are more than one word, making evaluation more challenging. *Questions Categories are not explicitly assigned, so performance on each category cannot be measured.
Visual7w	Amazon Mechanical Truck(AMT)	*Telling QA*: What, Where, How, When, Who, and Why. *Pointing QA*: Which	327,939	47,300	*Questions Categories are not explicitly assigned, so performance on each category cannot be measured.
TDIUC	–Exported from COCO-VQA and Visual Genome. –Automatically generated Using Templates. –Manual	Object Presence, Object Recognition, Counting, Color, Other attributes, Sport Recognition, Activity Recognition, Positional Reasoning, Scene Classification, Sentiment Understanding, Utilities& Affordance.	1,654,167	167,437	*Dataset bias, but evaluation metrics are designed to compensate for that bias. *Unbalanced distribution of questions between question types and answers within question types except for Object Presence type
FIGUREQA	Generated from 15 templates.	Yes/No	1,550,000	120,000	*Only one question type. *No numerical value answers. *Dataset contains only bar charts. *Plots are Synthetic
DVQA	Auto Generated from templates.	Structure Understanding, Data Retrieval, and	3,487,194	300,000	*Dataset contains only bar charts.

Dataset	Question/answer Collection	Question Types	# of Questions	# of Images	Drawbacks
		Reasoning.			
easy-VQA	Generated using algorithm	Yes/No, Shapes, Colors	48,248	5,000	Since the dataset is made for easier training and evaluation: - It has only 13 possible answers. - The images and questions are much simpler
FSVQA	Rule-based NLP techniques applied on VQA dataset answers to generate full sentence answer.	Yes/No, Number, Other.	369,861	62,292	*Additional Challenges regarding not only creating answers but also making it full sentence answer. *measured accuracy may be misleading
FSVQA aug	Set of rules applied to MS-COCO captions to generate question for captions as answers.	Yes/No, Number, Other.	986,628	64,060	
VizWiz	Images are captured by blind people and questions are recorded.	Yes/No, Number, Other, Unanswerable	32,842	32,842	*Difficult to have good results on it because it has blurred images, questions almost start with rare words or 'what' so question type cannot be identified easily, contains long multi sentence and noisy content unrelated to question.
VQAMed(2021)	–Medical images from the MedPix database –Test set is validated by medical doctors	Modality, Plane, Organ System, and Abnormality	5,000	5,000	*Small dataset size

4. Zero-Shot TDIUC Dataset

Here the TDIUC dataset is redistributed in Zero-Shot aspect with respect to question and answer in order to enable testing the VQA model's capability of generalization to questions without exploiting language priors and at the same time to get the advantage of TDIUC dataset's bigger size, wider range of explicitly defined question types, and good evaluation metrics for better evaluation of the effectiveness of VQA systems.

VQA datasets have 2 major problems. ***The first*** is that they are unbalanced towards some question types so good performance on less frequent kinds of questions has negligible impact on overall performance (i.e., some questions types are more common than others so performing well on less common question types will not show good impact on overall performance and thus model performance cannot be judged well), TDIUC's performance metrics compensate for this bias, so achieving good performance on TDIUC dataset requires having good accuracy across all kinds of questions. ***The Second*** problem is that questions can be answered without reasoning from the image, TDIUC has an additional absurd question type (i.e., queries not related to image) which requires an algorithm to look at the image to be able to determine if the question is appropriate for the image or not.

The entire TDIUC dataset is split into 70% train and 30% validation/test, there is no overlap between images in training and validation/test sets so as not to encourage overfitting. Each training/test instance is a triple of an image, a question, and an answer.

The dataset is re-organized for zero-shot purpose as the following and is indicated in Figure 6:

(1) For every question type *i*, a list of all distinct words *total_words_i* is generated such that the following word classes are excluded:

- ***Coordinating Conjunctions***: that join two elements of equal status like (and, or, but, if, while, although).
- ***Determiners or Articles***: that mark the beginning of a noun phrase like (the, a, some, most, every, which)
- ***Personal Pronouns***: that refer to persons or entities (you, she, I, it, me … etc.).
- ***Particles***: that resemble a preposition or an adverb and is used in combination with a verb like (at, on, out, over, that, up, with).
- ***Possessive Pronouns***: like (my, your, his, her, its, one's, our, their).
- ***Modals***: like (can, should).

(2) Choose a subset of the resulted distinct words set *Zero_Shot_i* such that it contains the least frequent words in *total_words_i* that will be used later as zero-shot words.

(3) Dividing *Zero_Shot_i* into two equally sized lists one for validation set *Zero_Shot_val_i* and the other for test set *Zero_Shot_test_i*.
(4) Search for all instances that contain at least one word from *Zero_Shot_val_i* in question, answer, or both of them as validation set instances.
(5) Search for all instances that contain at least one word from *Zero_Shot_test_i* in question, answer, or both of them as test set instances.
(6) Choose the training instances such that questions and answers do not contain any word in Zero_Shot_val_i or Zero_Shot_test_i and only take instances that have images not related to images used in any validation or test instance in order not to encourage model overfitting.

In each question type, there are different frequencies for each word. So the least frequent word is chosen for each type by different thresholds. As shown in Table 2, for example in Scene Recognition question type, all words that is repeated less than or equal to 75 (threshold value) are chosen as the unique words for this category to further be used to assess model's capability of generalization to unseen words by question type. Table 2 lists the different word frequency thresholds that is chosen for every question type in the experiment, the selected words to be unique words including POS words, and the number of unique words after excluding POS words for each category.

Also Images in the training set are chosen to be disjoint from those in validation/test set in order to avoid model overfitting. A detailed analysis of the original TDIUC split and the proposed TDIUC Zero-Shot split is shown in Table 3. The same percentage of instances in Zero-Shot TDIUC split (70% training instances, and 30% test/validation instances) is kept like the original TDIUC Split. Also, the distribution of question types in training, validation, and test is approximately the same distribution.

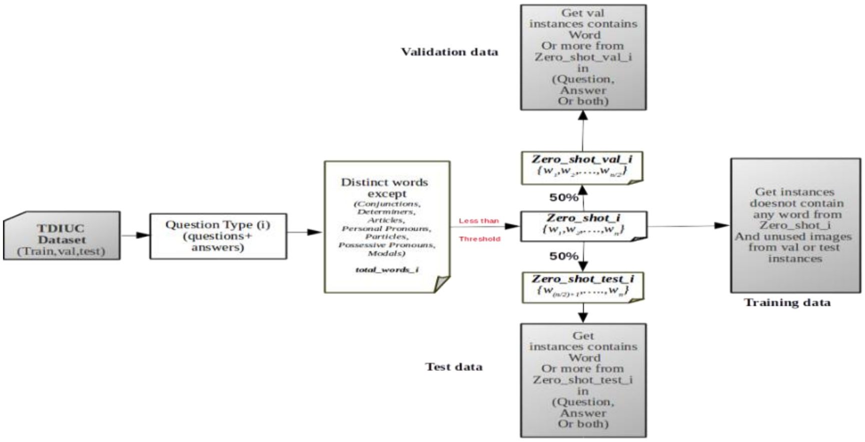

Validation data
Get val instances contains Word Or more from Zero_shot_val_i in (Question, Answer Or both)
Zero_shot_val_i $\{w_1,w_2,\dots,w_{n/2}\}$
50%
TDIUC Dataset (Train,val,test)
Question Type (i) (questions+ answers)
Distinct words except (Conjunctions, Determiners, Articles, Personal Pronouns, Particles, Possessive Pronouns, Modals)
total_words_i
Less than
Threshold
Zero_shot_i $\{w_1,w_2,\dots,w_n\}$
Get instances doesnot contain any word from Zero_shot_i And unused images from val or test instances
Training data
50%
Zero_shot_test_i $\{w_{(n/2)+1},\dots,w_n\}$
Test data
Get instances contains Word Or more from Zero_shot_test_i in (Question, Answer Or both)

Fig. 6: ZeroShot TDIUC splits generation

By examining the original TDIUC split, Figure 7 shows that the words frequency distribution in each category are following a long-tailed distribution so small part of words has many occurrences (the head) while a large part of the words has comparatively few occurrences (the tail). This means that if the part of words which have few occurrences are not asked about in validation/test set, we may have more bias to specific words which is represented by the words in the head of distribution. This is an advantage in the zero-shot TDIUC split by good choice of validation/test instances in order to get more expressive accuracy results.

In table 3, it is obvious that the percentage of instances which have at least one word of the selected unseen words in the validation set of the original TDIUC split is 15.47 % which is very small when it is compared to the percentage in the validation/test set in Zero-Shot TDIUC which is 100%, Also the training set of the original TDIUC Split has some instances which contain the selected unseen words (16.37%).

Table 2: Word frequency threshold for question types

Question Type	Word Frequency Threshold	Unique Words including POS Words	Total Unique Words
Scene Recognition	75	451	397
Sport Recognition	30	522	437
Color	65	6008	5365
Other Attributes	30	2642	2335
Activity Recognition	10	627	505
Positional Reasoning	20	3149	2629
Object Recognition	120	1477	1387
Absurd	500	1137	790
Utility / Affordance	1	543	282
Object Presence	2000	180	50
Counting	100	4099	3800
Sentiment Understanding	3	477	336

Table 3 Comparison between original TDIUC splits and the proposed zero-shot TDIUC split

	Original TDIUC Split		Zero-Shot TDIUC Split		
	Training	Validation	Training	Validation	Test
Number of Questions	1,115,299	538,868	403,787	80,155	92,898
Number Of Images	113,830	53,607	82,743	51,105	56,352
Question Types (scene, sport, color, attributes, activity, positional reasoning, object recognition, absurd, utility, object presence, counting, sentiment)	4.01% 1.94% 11.93% 1.75% 0.52% 2.33% 5.64% 22.08% 0.03% 39.61% 10.03% 0.13%	4.09% 1.86% 11.60% 1.71% 0.50% 2.28% 5.70% 22.35% 0.03% 39.96% 9.82% 0.12%	0.83% 0.82% 18.13% 3.23% 0.37% 2.82% 2.31% 40.39% 0.06% 23.06% 7.87%0.11%	0.72% 0.83% 18.68% 3.06% 0.38% 2.63% 2.10% 38.67% 0.06% 24.98% 7.78% 0.10%	0.92% 0.81% 17.65% 3.37% 0.36% 2.98% 2.49% 41.87% 0.08% 21.40% 7.96% 0.12%
				(disjoint sets)	
Number of instances with >0 unseen word	182,571 (16.37%)	84,835 (15.74%)	------------------------	80,155 (100.00%)	92,898(100.00%)
–unseen word in question	176,727 (15.85%)	176,727 (15.85%)	------------------------	77,825 (97.09%)	90,124 (97.01%)
–unseen word in answer	3,461 (0.31%)	1,548 (0.29%)	------------------------	1,807 (2.25%)	2,075 (2.23%)
–unseen word in question and answer	2,383 (0.21%)	1,083 (0.20%)	------------------------	523 (0.65%)	699 (0.75%)

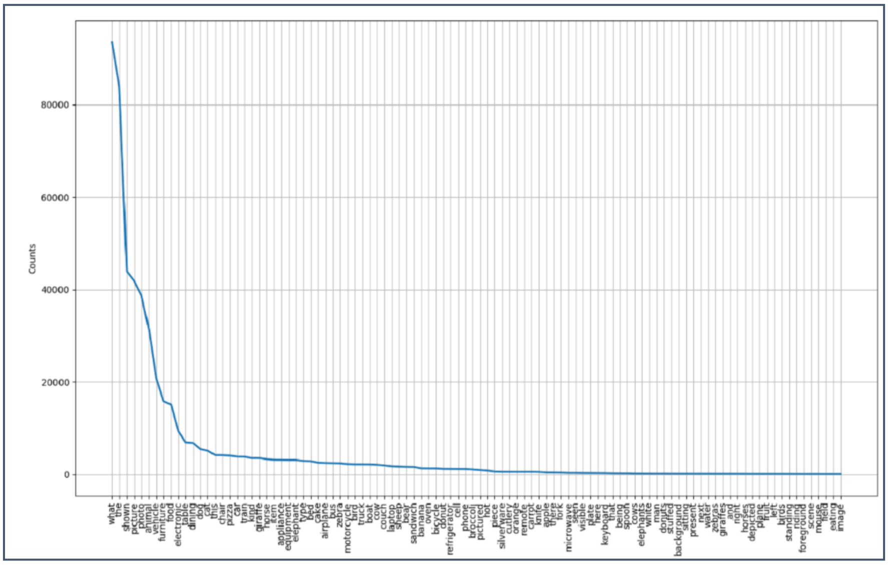


Fig. 7: Word frequency distribution in object recognition question type

5. Experiments using Deep Learning

Different models have been trained and evaluated on the proposed Zero-Shot TDIUC splits using different evaluation metrics. The results are reported in table 4. The following models are used to show how much the choice of test instances affects evaluation results:

(1) LSTM+VGG-19 VQA Model (Lu et al., 2015):

Getting the image encoding using the last hidden layer of VGG-19 layers network to get 4096 dimensional features (VGG is mostly used by researchers for image featurization because as mentioned by Manmadhan et al., (2020), VGGNet extracts features that are slightly more general, better performance for datasets other than ImageNet on which different CNN models are trained, and simply implemented, NLTK is used to tokenize the question then obtain the question's encoding (512 dimensional) using LSTM with 2 hidden layers each layer have 512 hidden nodes, then Multi Modal transformation is applied to features of both question and image to get common embedding size (1024), the embedding of image and question are combined together using element-wise multiplication, the feature vector then fed to a SoftMax layer to generate one answer from 1074 different answers using zero shot TDIUC splits' trained model or one answer from 1480 answers using standard TDIUC splits' trained model. As shown in Table 4, it is obvious that accuracy percentage for every question type based on standard test split is higher than the percentage reported using zero shot test split (e.g., sport recognition, scene recognition), which means that the measured accuracy on a model trained on original TDIUC splits gives a misleading indicator of the model's ability of generalization to new words that is presented in test questions.

(2) BOW+ResNet-18:

During training the model, validation data is used to choose the best epoch number and get the best model by calculating the best acquired accuracy over validation data from every epoch. This Model extracts image features using CNN network (here ResNet-18 is used), question and answer features is extracted using Bag Of Words. A BOW of question will be of size 9349 because there is 9349 different vocabularies in questions. The BOW of question is then concatenated to the CNN extracted features (512 length) and passed to the SoftMax layer to predict the answer class.

(3) BOW:

When question's encodings is used alone to generate an answer without seeing the image, some question types have little better accuracy than using (image + question) encodings or even the image's encoding only.

(4) IMG:

Image features are also used individually to find an answer to the question without using question features during training to check biasness or overfitting. Answering questions using only image as an input generates almost zero accuracy and this means that the model depends on question features more than image features to generate an answer. Only absurd category outputs high accuracy because all questions have the same answer (doesn't apply) and the question does not relate to the image.

Table 4: Accuracy of VQA Models on every question type and for overall question types

	Standard Splits				Zero Shot Splits			
	LSTM+VGG19	BOW	IMG	BOW + ResNet	LSTM+VGG19	BOW	IMG	BOW + ResNet
Scene Recognition	89.53	52.07	0.24	49.57	19.25	23.22	0.00	22.87
Sport Recognition	86.47	24.31	0.00	22.66	48.41	8.46	0.00	8.60
Color	53.96	40.91	0.00	39.06	26.77	20.70	0.00	20.20
Other Attributes	50.71	41.12	0.00	33.18	29.12	23.20	0.00	22.68
Activity Recognition	44.71	10.70	0.00	9.81	22.62	10.14	0.00	9.36
Positional Reasoning	31.05	18.10	0.00	14.94	11.31	7.16	0.00	7.18
Object Recognition	76.93	24.29	0.00	21.90	28.30	9.54	0.00	9.34
Absurd	76.97	91.95	99.26	94.59	96.34	96.68	99.89	97.22
Utility Affordance	29.24	13.45	0.00	7.60	11.02	8.47	0.00	9.32
Object Presence	90.51	68.98	0.37	68.85	74.67	69.27	0.01	67.20
Counting	49.19	45.10	0.00	44.31	42.96	43.06	0.00	42.71
Sentiment Understanding	64.04	53.15	0.47	49.53	28.12	28.12	0.00	27.08
Simple Accuracy	75.98	62.48	22.34	62.24	66.95	63.67	40.34	63.27
Overall (Arithmetic MPT)	61.94	40.34	8.36	38.00	36.57	29.00	8.32	28.65
Overall (Geometric MPT)	58.01	33.57	0.00	29.89	29.87	20.15	0.00	19.95
Overall (Harmonic MPT)	53.87	27.29	0.00	22.47	24.71	15.05	0.00	15.01
Overall (Arithmetic N-MPT)	42.29	26.03	8.35	25.58	29.02	21.26	8.32	20.17
Overall (Geometric N-MPT)	36.53	18.93	0.00	17.40	21.84	14.11	0.00	12.72
Overall (Harmonic N-MPT)	17.82	15.47	0.00	13.05	17.82	11.50	0.00	10.27

6. Evaluation Method

After getting the Zero-Shot TDIUC split, accuracies are reported over the trained models using 5 different evaluation metrics that were proposed by Kafle et al., (2017) and one more additional metric is used:

- *Simple Accuracy*: calculates accuracy over all 12 question types.

$$Simple\ accuracy = \frac{\text{\# of correct test instances}}{\text{total \# of test instances}} \tag{1}$$

This metric suffers from bias because the number of instances in every question type are not equally distributed. But as shown in Figure 8, some question types have few number of instances like activity recognition, sentiment understanding when compared to other question types like absurd, and color. So overall accuracy might not represent accuracy per question type efficiently. To get more representative measure, Mean Per Type Accuracy is calculated. Before calculating Mean Per Type Accuracy, a very small value $\varepsilon = 10^{-10}$ is added to accuracy of every question type individually in order to avoid dividing by zero when calculating harmonic mean. Accuracy of every question type is calculated using equation (2).

$$acc(questiontype)_i = 100 * \frac{\text{\# of correct instances(questiontype)}_i}{\text{total \# of instances(questiontype)}_i} \tag{2}$$

- *Arithmetic Mean Per Type(AMPT)*: is used to have more representative measure of accuracy using mean of accuracies over all question types as in equation (3)

$$AMPT = \frac{\sum_{i=1}^{12} acc(questiontype)_i}{12} \tag{3}$$

The arithmetic mean works well to produce an average number of a set of values when there is an additive relationship between the numbers (linear relationship) but when the values (accuracy per question type) in TDIUC dataset were graphed in ascending order as seen in figure 9, the values resemble more of a curve than a straight line. So in this situation, the arithmetic mean is ill-suited to produce an “average” number to summarize this data.

- *Geometric Mean Per Type (GMPT):* Since the relationship approximately tends to be multiplicative, Geometric mean is proposed here as a new metric using equation 4. Geometric mean performs better if the numbers are in different ranges entirely and we do not want one very large number to affect he result that much.

$$GMPT = \sqrt[12]{acc(questiontype)_1 \times \cdots \times acc(questiontype)_{12}} \tag{4}$$

As shown in Figure 9, Geometric mean is very much resembling the middle value of accuracies. In fact, it is the nearest to the median. The geometric mean will equal the median, only in cases where there is an exact consistent multiplicative relationship between all numbers.

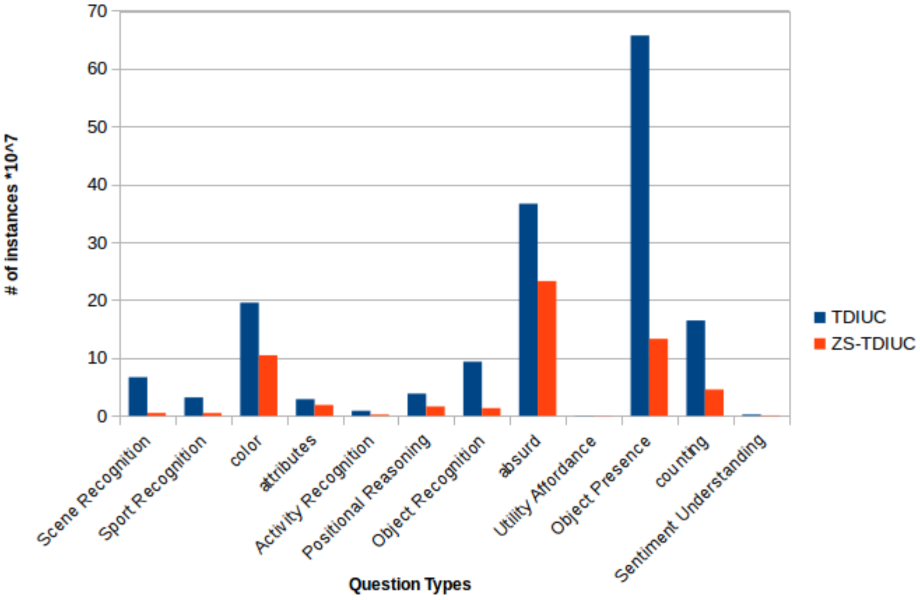


Fig. 8: Question types distribution in standard and zero-shot split of TDIUC

- *Harmonic Mean Per Type (HMPT):* Since arithmetic mean value isn't particularly close to most of the different question types' accuracies, Harmonic mean is also used as a measure of accuracy. Harmonic mean is the reciprocal of the arithmetic mean of the reciprocals as in equation 5. The advantage of harmonic mean is that it is more sensitive to small than large numbers and works better with fractions and here mean of percentage values is calculated.

$$\text{HMPT}=\frac{12}{\sum_{i=1}^{12}\frac{1}{acc(questiontype)_i}} \tag{5}$$

- *Normalized Arithmetic Mean Per Type (N-AMPT):* Since there is an imbalance in the distribution of answers within every question type, accuracy of every distinct answer in every question type is calculated using equation (6) and accuracy for every question type based on answers' accuracy is calculated using equation (7). So we get more normalized accuracy measure.

$$\text{AccAnswer}_j\ (\text{questiontype}_i) = \frac{\#\ \text{of j correct answers}}{\#\text{of instances with answer j}} \tag{6}$$

$$\text{AccAns}(\text{questiontype}_i) = \frac{\sum_{j=1}^{n} \text{AccAnswer}_j\ (\text{question type}_i)}{n} \times 100 \qquad (7)$$

where n = number of distinct answers in question type.

$$NAMPT = \frac{\sum_{i=1}^{12} \text{AccAns}(\text{questiontype}_i)}{12}$$

- Normalized Geometric Mean Per Type(N-GMPT):

$$NGMPT = \sqrt[12]{\text{AccAns}(\text{questiontype}_1) \times \cdots \times \text{AccAns}(\text{questiontype}_{12})} \qquad (8)$$

- Normalized Harmonic Mean Per Type(N-HMPT):

$$NHMPT = \frac{12}{\sum_{i=1}^{12} \frac{1}{\text{AccAns}(\text{questiontype}_i)}} \qquad (9)$$

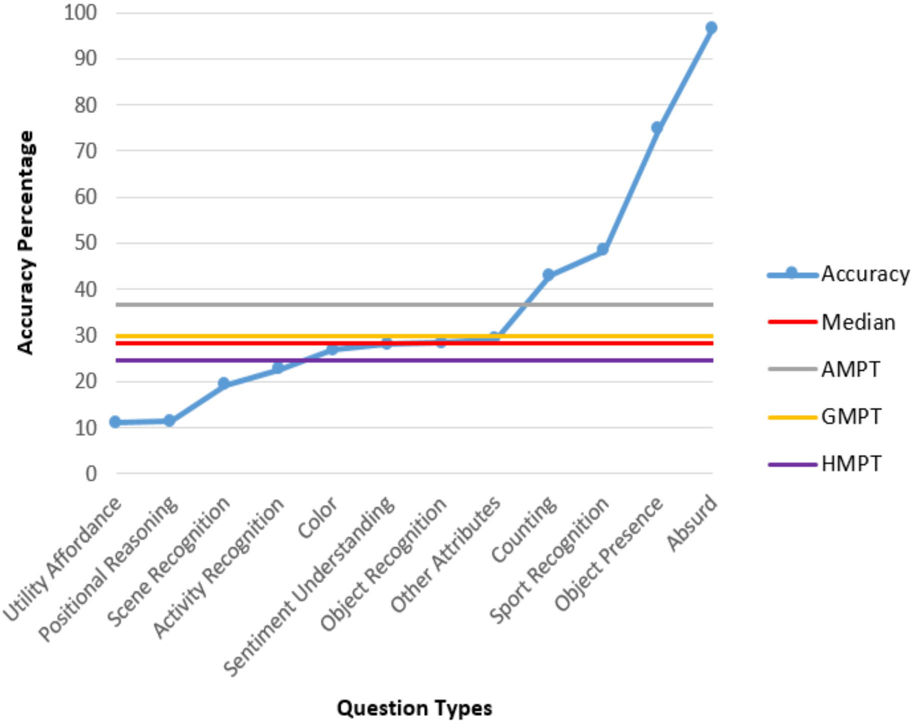


Fig. 9: LSTM-VGG per question type accuracy

7. Conclusion

In This paper, a redistributed version of TDIUC dataset is proposed to discuss how much the choice of test instances can affect the honesty of model's evaluation. By re-arranging TDIUC dataset such that test instances have new concepts that is not shown before in training instances, biasness is recorded in reported accuracies on the original split of TDIUC dataset. So the model’s capability of generalization can be measured better using Zero Shot redistribution of dataset and this must encourage researchers to take in account when collecting datasets to choose test instances in zero shot aspect in order to help generating good VQA models.

Also using explicitly defined question types provides an additional capability by evaluating models on every question type separately. Accuracy is measured by metrics that is not affected by very high/low accuracy in specific question types. GMPT is introduced as a new metric and it is considered as a more expressive measure because it is the closest to the median of question type’s accuracies.

References

Abacha, A. B., Datla, V. V., Hasan, S. A., Demner-Fushman, D., & M¨uller, H. (2020). Overview of the vqa-med task at imageclef 2020: Visual question answering and generation in the medical domain. in CLEF (Working Notes).

Andreas, J., Rohrbach, M., Darrell, T., & Klein, D. (2016). Neural module networks. *in Proceedings of the IEEE conference on computer vision and pattern recognition*, 39-48.

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., & Parikh, D. (2015). Vqa: Visual question answering. *in Proceedings of the IEEE international conference on computer vision*, 2425-2433.

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., & Ives, Z. (2007). Dbpedia: A nucleus for a web of open data. *in The semantic web*, 722–735, Springer.

Barra, S., Bisogni, C., De Marsico, M., & Ricciardi, S. (2021). Visual question answering: Which investigated applications?, arXiv preprint arXiv:2103.02937.

Ben Abacha, A., Hasan, S. A., Datla, V. V., Liu, J., Demner-Fushman, D., & M¨uller, H. (2019). VQA-Med: Overview of the medical visual question answering task at imageclef 2019. in CLEF2019 Working Notes, CEUR Workshop Proceedings, (Lugano, Switzerland), CEUR-WS.org <http://ceur-ws.org>, September 09-12.

Ben Abacha, A., Sarrouti, M., Demner-Fushman, D., Hasan, S. A., & M¨uller, H. (2021). Overview of the vqa-med task at imageclef 2021: Visual question answering and generation in the medical domain. in CLEF 2021 Working Notes, CEUR Workshop Proceedings, (Bucharest, Romania), CEUR-WS.org, September 21-24.

Budiharto, W., Andreas, V., & Gunawan, A. A. S. (2020). Deep learning-based question answering system for intelligent humanoid robot. *Journal of Big Data*, 7(1), 1-10.

Cao, L., Gao, L., Song, J., Xu, X., & Shen, H. T. (2017). Jointly learning attentions with semantic cross-modal correlation for visual question answering. *in Australasian Database Conference*, 248-260, Springer.

Chen, K., Wang, J., Chen, L.-C., Gao, H., Xu, W., & Nevatia, R. (2015). Abc-cnn: An attention based convolutional neural network for visual question answering. arXiv preprint arXiv:1511.05960.

Cho, K,. Van Merri ̈enboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Eckart, C. & Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3), 211-218, 1936.

Fukui, A., Park, D. H., Yang, D., Rohrbach, A., Darrell, T., & Rohrbach, M. (2016). Multimodal compact bilinear pooling for visual question answering and visual grounding. arXiv preprint arXiv:1606.01847.

Gao, H., Mao, J., Zhou, J., Huang, Z., Wang, L., & Xu, W. (2015). Are you talking to a machine? dataset and methods for multilingual image question. *in Advances in neural information processing systems*, 2296–2304.

Geman, D., Geman, S., Hallonquist, N., & Younes, L. (2015). Visual turing test for computer vision systems. *Proceedings of the National Academy of Sciences*, 112(12), 3618–3623.

Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., & Parikh, D. (2017). 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*, 6904–6913.

Gurari, D., Li, Q., Stangl, A. J., Guo, A., Lin, C., Grauman, K., Luo, J., & Bigham, J. P. (2018). Vizwiz grand challenge: Answering visual questions from blind people. *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3608–3617.

Hasan, S. A., Ling, Y., Farri, O., Liu, J., M ̈uller, H., & Lungren, M. P. (2018). Overview of imageclef 2018 medical domain visual question answering task. *in CLEF (Working Notes).*

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *in Proceedings of the IEEE conference on computer vision and pattern recognition*, 770-778.

Huang, L. -C., Kulkarni, K., Jha, A., Lohit, S., Jayasuriya, S., & Turaga, P. (2018). Cs-vqa: Visual question answering with compressively sensed images. in 2018 25th IEEE International Conference on Image Processing (ICIP), 1283–1287, IEEE.

Kazemi, V. & Elqursh, A. (2017). Show, ask, attend, and answer: A strong baseline for visual question answering. arXiv preprint arXiv:1704.03162.

Kafle, K. & Kanan, C. (201). An analysis of visual question answering algorithms. *in 2017 IEEE International Conference on Computer Vision (ICCV)*, 1983–1991, IEEE.

Kafle, K., Cohen, S., Price, B., & Kanan, C. (2018). Dvqa: Understanding data visualizations via question answering. arXiv preprint arXiv:1801.08163.

Kahou, S. E., Atkinson, A., Michalski, V., Kadar, A., Trischler, A., & Bengio, Y. (2017). Figureqa: An annotated figure dataset for visual reasoning. arXiv preprint arXiv:1710.07300.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., Bernstein, M., & Fei-Fei, L. (2016). Visual genome: Connecting language and vision using crowdsourced dense image annotations.

Lu, J., Lin, X., Batra, D., & Parikh, D. (2015). Deeper lstm and normalized cnn visual question answering model. https://github.com/VT-vision-lab/VQA LSTM CNN.

Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. arXiv preprint arXiv:1908.02265.

Manmadhan. S. & Kovoor, B. C. (2020). Visual question answering: a state-of-the-art review. *Artificial Intelligence Review*, 53(8), 5705-5745.

Malinowski, M. & Fritz, M. (2014). A multi-world approach to question answering about real-world scenes based on uncertain input. *in Advances in Neural Information Processing Systems 27* (Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, eds.), 1682–1690, Curran Associates, Inc.

Malinowski, M., Rohrbach, M., & Fritz, M. (2015). Ask your neurons: A neural-based approach to answering questions about images. *in Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV),* 1–9, IEEE Computer Society.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Miller, G. A. & Charles, W. G. (1991). Contextual correlates of semantic similarity. *Language and cognitive processes*, 6(1), 1-28.

Mishra, A., Anand, A., & Guha, P. (2020). Cq-vqa: Visual question answering on categorized questions. *in 2020 International Joint Conference on Neural Networks (IJCNN)*, 1–8, IEEE.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. *in Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. arXiv preprint arXiv:1802.05365.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding with unsupervised learning.

Ren, M., Kiros, R., & Zemel, R. (2015). Exploring models and data for image question answering, *in Advances in neural information processing systems*, 2953–2961.

Ren, M., Kiros, R., & Zemel, R. (2015). Image question answering: A visual semantic embedding model and a new dataset. *Proc. Advances in Neural Inf. Process. Syst*, 1(2), 5.

Simonyan, K. & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556.

Shi, Y., Furlanello, T., Zha, S., & Anandkumar, A. (2018). Question type guided attention in visual question answering. *in Proceedings of the European Conference on Computer Vision (ECCV)*, 151–166.

Tan, H. & Bansal, M. (2019). Lxmert: Learning cross-modality encoder representations from transformers. arXiv preprint arXiv:1908.07490.

Teney, D. & van den Hengel, A. (2018). Visual question answering as a meta learning task. *in Proceedings of the European Conference on Computer Vision (ECCV)*, 219–235.

Teney, D. & v. d. Hengel, A. (2016). Zero-shot visual question answering. arXiv preprint arXiv:1611.05546.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N. Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *in Advances in neural information processing systems*, 5998-6008.

Wu, Q., Shen, C., Wang, P., Dick, A., & van den Hengel, A. (2017). Image captioning and visual question answering based on attributes and external knowledge. *IEEE transactions on pattern analysis and machine intelligence*.

Wu, Q., Wang, P., Shen, C., Dick, A., & van den Hengel, A. (2016). Ask me anything: Free-form visual question answering based on knowledge from external sources. *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4622–4630.

Young, T., Hazarika, D., Poria, S., & Cambria, E. (2018). Recent trends in deep learning based natural language processing. *IEEE Computational intelligenCe magazine*, 13(3), 55-75.

Zhang, Q., Lei, Z., Zhang, Z., & Li, S. Z. (2020). Context-aware attention network for image-text retrieval. *in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3536–3545.

Zhou, B., Tian, Y., Sukhbaatar, S., Szlam, A., & Fergus, R. (2015). Simple baseline for visual question answering. arXiv preprint arXiv:1512.02167.

Zhu, Y., Groth, O., Bernstein, M., & Fei-Fei, L. (2016). Visual7w: Grounded question answering in images. *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4995–5004.