

Multihop-Multilingual Co-attention Method for Visual Question Answering

Debajyoty Banik, Devashish Kumar Singh and Mohit Kumar Pandey

Kalinga Institute of Industrial Technology, Bhubaneswar,

Abstract

Our model revolves around expanding web-searching to the multi-domain, our project help to cover the gap present in today's research with regards to visual learning and harvesting it to gain knowledge out of it. we intend to mirror human behaviour with respect to gathering knowledge from multi domain sources. As the information matter not the source and web 2.0 and web 3.0 contain a lot of images and a picture speak a thousand words we intend to find way to harvest the info and make it efficient enough that the common people queries could be answered from information extracted from the pictures.

Keywords

VQA; Fast R-CNN; queries; Visual Genome

1. Introduction

In the present time, most of the in-use question-answering systems are text-based. If we want to understand this in simpler words, then we can rephrase it as when we search for something on the internet, the source in which system digs in to find the answers are most likely to be text-based. Though there is nothing wrong with this way of searching, it has some limitations and is not so relevant with time [1].


With the introduction of web 2.0 and web 3.0 on its way, the information present on the internet can be uploaded by anyone and everyone and a major portion of that information is photos, text snippets, etc which are also a gold mine of information [2]. But we at the present date are unable to harvest the data present in this multi-modal source and in this paper, we are going to be working on digging up a path of how we can give a tap on this source of information and use it to increase our search base and fulfill the user demands while being accurate.


We are going to use tokenize the question with the help of Bert based system, query for its text-based out and in relevant images in pairs to get information. we would rate them based on Bart-score, fluency, and accuracy and present it in front of the answer. while in the whole process we will hunt for answers from different sources making it multimodal and use scoring models to keep the answers accurate (see Figure 1).

International Workshop on Deep Learning for Question Answering-2022

✉ debajyoty.banik@gmail.com (D. Banik); furry7976@gmail.com (D. K. Singh); mohitpandeybgp@gmail.com (M. K. Pandey)

ORCID 0000-0002-3756-864X (D. Banik)

 © 2022 © 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

Our study uses faster R-CNN on the relevant images to make the process more accurate, faster and relevant. we try to mimic human behaviour and while looking for relevant info in the multi modal source pool where the current research tend to just rely just on the relevance of image based on their scores making less accurate.

2. Novelty

The Bert-base-cased tokenizer tokenizes all text segments, including the questions, answers, textual sources, and image captions. 100 regions from an object detection model, a Faster R-CNN variant, are used to represent each image. with a ResNeXt-101 FPN backbone and Visual Genome[3] pre-training. To satisfy the auto-regressive characteristic, attention masks are applied to tokens in A by the Transformer after we feed it [CLS], S, [SEP], Q, A, [SEP]>. During fine-tuning, we employ the usual Masked-Language-Modeling[4] loss. By repeatedly adding a [MASK] to the input's end, swapping it out for a predicted token, then adding a new [MASK] for the subsequent time step, we decode. After witnessing [SEP], [PAD], or when the length reaches a certain point, generation ends. To expose improvements and costs stemming from the complexity of providing models with data from both modalities, we additionally provide two modality-specific variants, VLPI and VLPT, which are skilled in answering text- or image-based inquiries alone as opposed to the whole data.¹

3. Task formulation

Let us say, somebody asked a question, then a sets of positive results will be produced which satisfy the condition of being either being a snippet or a pair of images and descriptions. has things like its location or another characteristic which will act as a reference to identify attached to it which serve as critical points in answering the question asked.

We accomplish the task in two stages. First, let us say the questions Q and s1, s2, ..., sn, The positive pairs found by searching the photos are identified by the model. The model uses question Q and the selected sources as context C in the second stage to produce answer A. However, Future research is needed because we are not aware of any modelling tools that can consume sufficiently large multimodal settings to accomplish this. A single-stage system would ideally combine the processing of Q, s1, s2,..., Sn to produce the determination of A and C.

4. Answers from Text

4.1. Hard Negative Mining

In the process of hard negative mining for text, we select the articles and sources that overlap based on the noun phrases extracted from the inquiry and mine sources like Wikipedia, articles, and other similar sources. though there is a lac

¹DLQ-2022: International Workshop on Deep Learning for Question Answering, Co-located with the KGSWC-2022, November 21-23, 2022, Madrid, Spain.

5. Answers from Images

5.1. Hard Negative Mining For Image

To respond to all questions and provide references, we develop pairs of text and image-based data during the hard negative mining process. which was produced while breaking down the question. The text is sourced from sources like articles, Wikipedia, magazines, comments, etc, and its chosen based on nouns present in the question.

For pictures, we use search engines API like bing Apis to find images relevant to the question on the basis of the description of the image and other factors. And we pair these text and images to form pairs.

5.2. Categorization

We divide the questions into yes or no or like which, why, how, and such nouns and we tend to compare such nouns on these pairs and classify them with the help of GQA and xGQA.

lack of clarity in question too sometimes we also simply sample randomly the sources and use all the available sources.²

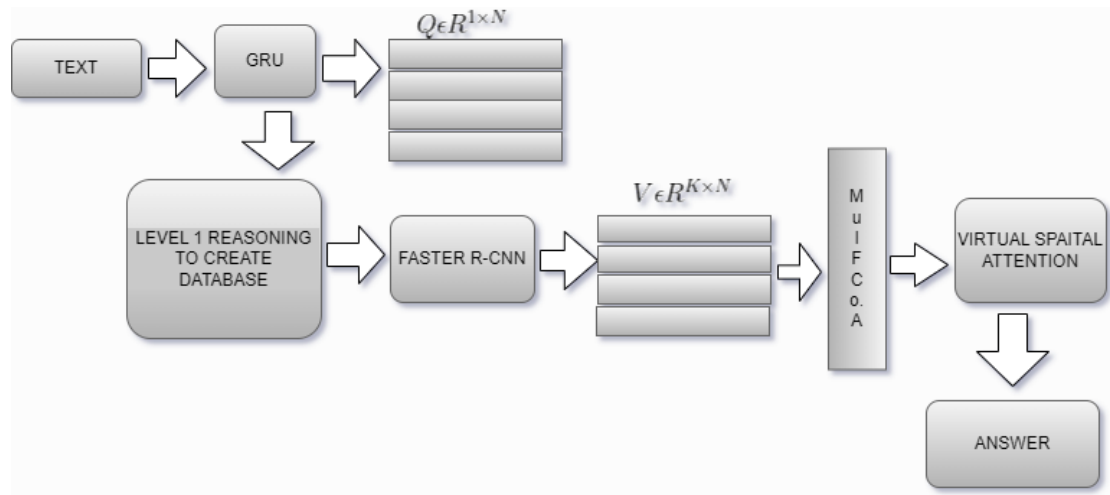


Figure 1: Multimodal searching here the user input the question in text format and the search engine does a text query to find the answer as well as does multimodal query of images, text snippets, etc pair it up, search for details asked in the question with help of aster RCN and then rank than up based on score and give top ones as output to the user.

²DLQ-2022: International Workshop on Deep Learning for Question Answering, Co-located with the KGSWC-2022, November 21-23, 2022, Madrid, Spain.

6. Answering & Quality Control

6.1. Quality Control

To ensure we give quality content in the answers, not something which is fake, we ensure quality control through 2 methods, one is crowd-sourcing[5] and the other is feedback on loops [2]. a group of annotators is trained and selected. after that, each batch is given data and a bonus for out-of-box thinking. each group is sent with data and constructive feedback looping to correct our mistakes through the help of these large numbers of human understanding.

6.2. Fluency

Fluency is measured, with the help of Bart score, a newly proposed based on accurate measurement of paraphrase quality. The Bart-score[6] measure's the probability of generating B from [7].

This is calculated in our scenario as $\text{Bart-score}(r, c)$ [8], which can be understood as the likelihood of producing a candidate given a reference.

7. Faster R-CNN

Faster R-CNN is an extension of Fast R-CNN. It saves a lot of our time comparing it with Fast RCN. Faster R-CNN, as its name implies, is quicker than Fast R-CNN because of the region proposal network (RPN). Regions with convolutions neural networks (R-CNN) use a novel region proposal network(RPN) to generate a regional proposal, which compares with traditional algorithms like selective searching to save us time. Faster R-CNN combined with the RPN network is one of the best ways to detect R-CNN series based on deep learning.

The ROI Pooling layer, a CNN framework for successful end-to-end object identification, is intimately related to the proposal obtained by RPN[9]. Based on the implementation of Faster R-CNN models that can be obtained by training using the deep learning framework of Caffe, the viability of R-CNN works on the RwsNet101 network and PVANET network is examined.

7.0.1. DRAWBACKS OF R-CNN

The fact that RPN is trained so that all of its anchors in a mini-batch are 256 and taken from the same single image presents one potential disadvantage of the quicker R-CNN. As a result, samples may be correlated, which means that their features are likewise associated, delaying convergence.

From here, we could see that the pros are more than the cons, so it is not a bad idea to use it until a better and more advanced system sets its foot on the market.³

³DLQ-2022: International Workshop on Deep Learning for Question Answering, Co-located with the KGSWC-2022, November 21-23, 2022, Madrid, Spain.

8. Multimodal feature-wise attention module (MulFA)

Most of the currently used techniques take spatial into account cross-grounding. In other words, we could also say, they determine the relationship between each spatial object in an image and question. These models, however, entirely disregard the feature channel dimension's concentration in the image as well as the question representation and only concentrate on learning spatial attention. Some of the tasks of computer versions, for example, classification [8] and image caption [10], have proved that incorporating that feature channel attention mechanism has better performance than usual. Because it allows the model to learn effectively.

In this paper, we propose that the MulFA seeks to produce greater attention weights. to emphasize informative suppressing less significant aspects.. To generate attention weight, MulFA uses bilinear models. There are two types of MulFA, one for image modalities and the other for text, namely: IMulFA(see figure 2) and QMulFA(see figure 3). IMulFA is used for modulating images, and QMulFA is used for question or text modalities.⁴

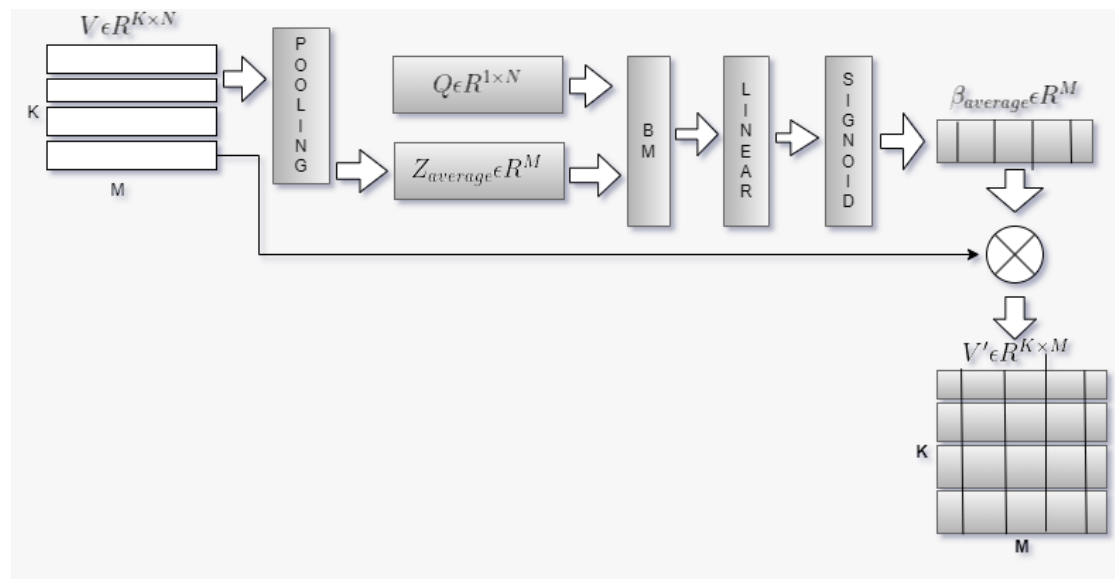


Figure 2: Image Multimodal feature-wise attention module

8.0.1. Image multimodal feature-wise attention module (IMulFA)

In this process, the image is understood by the AI. It happens or is processed in four steps. [11]:

1. Squeezing image feature,
2. Fusing feature-wise statics and question signals,
3. Computing feature-wise attention weight,
4. Feature-wise re-writing image

⁴DLQ-2022: International Workshop on Deep Learning for Question Answering, Co-located with the KGSWC-2022, November 21-23, 2022, Madrid, Spain.

The process is more briefly described with the help of fig.:2 In this fig.: we could clearly see the flow of the IMulFA. Here, V is the image in vector form whereas, k is the object feature and M is the feature channels.

8.0.2. Question multimodal feature-wise attention module(QMulFA)

In this process question or text is processed by the AI. The process is processed in just 3 simple steps[7].:

1. combining information from multiple sources to create feature-wise attention weight vectors,
2. Squeezing attention weight vector,
3. Re-calibrating question features.

The QMulFA figure could help the process run more smoothly. The question feature-wise attention weight vector is created in this case by fusing the signals from the visual and question feature channel statics. The ith object feature vector produces the ith weight vector.

$$V_i \in R^M$$

and the equation feature Q has :

$$f_i = BM(V_i, Q^T)$$

,

$$h_i = \text{sigmoid}(w_f^q f_i)$$

where $h_i \in R^N$, $w_f^q \in R^{N \times C}$, is a parameter matrix of the single linear layer, and

$$f_i \in R^C$$

denotes the fusion feature obtained by a bilinear model.⁵

The $V \in R^{K \times M}$ has K items in each of which can direct the question's feature-wise focus. Accordingly, To integrate, we use an average pooling operation. the following are all items effects:

$$a = \frac{1}{K} \sum_{i=1}^K h_i$$

, where $a \in R^N$ denotes the question feature-wise attention weight vector.

The question characteristics are then recalibrated using Q and the attention combined with element-wise multiplication. as

$$Q' = a^T \times Q$$

, whereas, $Q' \in R^{1 \times N}$ is the feature-wise attention feature. We define this QMulFA as

$$Q' = QMulFA(V, Q)$$

⁵DLQ-2022: International Workshop on Deep Learning for Question Answering, Co-located with the KGSwC-2022, November 21-23, 2022, Madrid, Spain.

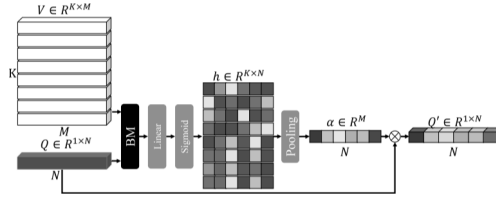


Figure 3: Question multimodal feature-wise attention module

9. Multimodal feature-wise co-attention module for VQA

With regard to picture and question modalities, we have created a feature-wise attention-learning module. We suggest three co-attention mechanisms to combine them, each of which has a different approach to how picture and question feature-wise attention is prioritized. the first two mechanisms, which we call alternate performing feature-based attention on the query and the image simultaneously, as below⁶

$$V' = \text{QMulFA}(V, Q), Q' = \text{QMulFA}(V', Q) \text{ or } Q' = \text{QMulFA}(V, Q), V' = \text{IMulFA}(V, Q')$$

The third mechanism, which we call parallel co-attention, generate images, and question attention simultaneously, defined as

$$V' = \text{IMulFA}(V, Q) \quad Q' = \text{QMulFA}(V, Q)$$

10. Multimodal spatial attention module

The issue of visual question answering (VQA) in computer vision is widely recognized. Due to how crucial it is to comprehend an image, text-based VQA assignments have recently attracted a lot of attention. In this area of research, we suggest a cutting-edge encoder-decoder framework to specifically predict complicated responses. We use the attention mechanism, which can choose characteristics based on the questions, to obtain the more pertinent features for the inquiry.

In order to answer correctly or even relevant to the question, we need to focus on the region which is related to our question, and hence, we take i n use of multimodal spatial attention module. Unlike its name, it focuses on the important part of the image and suppresses the rest of them. In order to that we first fuse the visual feature

$$V' \in R^{K \times M}$$

and the question features

$$Q' \in R^{1 \times N}$$

The attention distribution over the area of the image is produced by the bilinear model's computations, which are then fed with the fusion feature to a softmax function as illustrated in the figure.:1.

⁶DLQ-2022: International Workshop on Deep Learning for Question Answering, Co-located with the KGSWC-2022, November 21-23, 2022, Madrid, Spain.

11. VQA 1.0 and VQA 2.0

Bottom-up[12] applies the quicker R-CNN-based bottom-up attention approach suggested in [13], which enables the region related to a question.

MLB (Multimodal Low-Rank Bilinear Pooling) [14] is a solution to the issue of the computational cost of the bilinear model while utilizing its acceptable capacity for representation.

MLB is extended by MFH (Multimodal Factorized High-Order Pooling) to incorporate multimodal characteristics. High-order pooling is used. It utilizes high-level characteristics and image convolutions features.

The BAN (bilinear attention network) adopts a bottom-up focus on image attributes and task-level question features. An attention map is produced by BAN by computing the bilinear interaction between each pair of picture and question features.

The counter is specifically designed to handle hard counting questions, which call for a model to specify which types of objects need to be counted.

12. Datasets

VQA 1.0. In this update, there are more than 204k images from the Microsoft common object in context (MS coco) dataset, more than 600k questions (at least 3 questions per image), and around 6 million of answers (10 answers per question). The datasets are of three types namely :

1. Train : consists of 80k images and 240k question-answer pairs
2. val : consists of 40k images and 120k question-answer pairs
3. Test: consists of 80k images and 240k question-answer pairs

VQA 2.0. This is the updated version of the previous VQA which is VQA 1.0. VQA 2.0 has a longer scale, having 240k images from Microsoft's common object in context (MS coco), more than 1 million questions, and 11 million's of answers. It is composed of 4,43,757 pairs of image, questions, and answer for training, whereas, 2,14,354 for validating and 4,47,793 for testing.[15]

Our evaluation findings for the VQA 1.0 test set are displayed in table 1 We contrast the outcomes of our models with those from a number of cutting-edge models, including the VQA 1.0 Challenge's reigning champion, the MFH model. Table 1 demonstrates that our model UFSCAN outperforms every method, including the winner of the VQA 1.0 challenge, MFH[16]. With the exception of the three MFH-based models, it greatly outperforms the rest. The most recent model, MFH+CoAtt+Glove (bottom-up), is trained using the same train set and validation set as UFSCAN and uses the same bottom-up attention features. Notably, MFH uses more question characteristics than UFSCAN and adds the question spatial attention method. Nevertheless, UFSCAN exceeds the top-performing MFH model, MFH+CoAtt+Glove (bottom-up), highlighting the benefits of our suggested MulFA. Additionally, with data augmentation utilizing the Visual Genome, our model UFSCAN + VG achieves the best overall accuracy of 70.19% and 70.24% on the test-dev set and test-standard set, respectively. UFSCAN performs at the cutting edge on VQA 1.0 as a consequence.[17]

⁷DLQ-2022: International Workshop on Deep Learning for Question Answering, Co-located with the KGSWC-2022, November 21-23, 2022, Madrid, Spain.

Table 1

On the VQA 1.0 Test-dev and Test-standard accuracy(%) of single module set compared to state-of-the-art module. Whereas “-” represents that the result is currently unavailable, “Att” represents the visual spatial attention mechanism, “CoAtt” stands for the question and visual co-attention mechanism, “GloVe” indicates the word embedding method and “VG” indicates the Visual Genome for data augmentation.

Model	Test-dev	Test-standard
LSTM Q + I	57.8	58.2
SMem	58.0	58.2
SAN	58.7	58.9
FDA	59.2	59.5
DMN+	60.3	60.4
HieCoAtt	61.8	62.1
RAU	-	64.1
MCB + Att + GloVe + VG	65.4	-
MLB + Att + StV + VG	65.8	-
MFH + CoAtt + GloVe	66.8	66.9
MFH + CoAtt + GloVe + VG	67.70	67.5
MFH + CoAtt + GloVe (bottom-up)	68.78	-
UFSCAN	69.06	69.34
UFSCAN + VG	70.19	70.24

Table 2

Test-dev and test-standard accuracy of single-model on the VQA 2.0 data-set, whereas “-” indicates the result is not available

	Model	Test-dev accuracy (%)				Test-standard accuracy (%)			
		ALL	Yes/No	Number	Other	All	Yes/No	Number	Other
W/o counte	Bottom-up	65.32	81.82	44.21	56.05	65.67	82.2	43.9	56.26
	MFH	66.12	-	-	-	-	-	-	-
	MFH+Bottom-up	68.76	84.27	49.56	59.89	-	-	-	-
	BAN	69.66	85.46	50.66	60.50	-	-	-	-
	UFSCAN	69.83	85.21	50.98	60.98	70.09	85.51	50.21	61.22
Counter	Counter	68.09	83.14	51.62	58.97	68.41	83.56	51.39	59.11
	BAN + counter	70.04	85.42	54.04	60.26	70.35	-	-	-
	UFSCAN + counter	70.46	85.52	54.99	61.08	70.73	85.87	54.37	61.30

Table [2] contrasts our model’s performance on the VQA 2.0 data set with that of the most recent cutting-edge models. A counting module called Counter was proposed by Zhang et al. [18, 19] with the goal of addressing counting-related issues. Significantly improve the accuracy of answering counting questions. . Table 2 is divided into two sections for easier comparison: the first section lists the techniques that do not use the counting module, and the second section lists the methods that do. Visual Genome [3] is used for data augmentation, and all the models are trained using the identical training and validation splits.

13. Conclusions

In the model, we create a new model for⁸ answering the question in a multi-modal way, which is a great challenge in these changing times when we are changing from web 2.0 to web 3.0. design to simulate the environment, one is going to face in the real world while searching for information. Our model searches in multiple domains for the answers rather than just being dependent on a text query[20].

At the same time, we also focus on the fluency and accuracy of the answer. In this paper, For the purpose of bridging multimodal QA and IR research, we have offered both a restricted and a complete retrieval setting. In addition to reflecting our daily web experience, this data set offers the community a playground to investigate significant sub-challenges with the goal of developing a single mode. For knowledge aggregation, multimodal reasoning, and open-domain visual comprehension. Our project's ultimate objective is to gather pertinent data from the multi-domain mode, combine it above a sizable context window, and produce fluent, natural answers.⁹

References

- [1] E. M. Bender, A. Koller, Climbing towards nlu: On meaning, form, and understanding in the age of data, in: Proceedings of the 58th annual meeting of the association for computational linguistics, 2020, pp. 5185–5198.
- [2] Y. Bisk, A. Holtzman, J. Thomason, J. Andreas, Y. Bengio, J. Chai, M. Lapata, A. Lazaridou, J. May, A. Nisnevich, et al., Experience grounds language, arXiv preprint arXiv:2004.10151 (2020).
- [3] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, et al., Visual genome: Connecting language and vision using crowdsourced dense image annotations, *International journal of computer vision* 123 (2017) 32–73.
- [4] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).
- [5] N. Nangia, S. Sugawara, H. Trivedi, A. Warstadt, C. Vania, S. R. Bowman, What ingredients make for an effective crowdsourcing protocol for difficult nlu data collection tasks?, arXiv preprint arXiv:2106.00794 (2021).
- [6] W. Yuan, G. Neubig, P. Liu, Bartscore: Evaluating generated text as text generation, *Advances in Neural Information Processing Systems* 34 (2021) 27263–27277.
- [7] W. Yuan, G. Neubig, P. Liu, Bartscore: Evaluating generated text as text generation, *Advances in Neural Information Processing Systems* 34 (2021) 27263–27277.
- [8] K. J. Shih, S. Singh, D. Hoiem, Where to look: Focus regions for visual question answering, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 4613–4621.

⁸DLQ-2022: International Workshop on Deep Learning for Question Answering, Co-located with the KGSWC-2022, November 21-23, 2022, Madrid, Spain.

⁹Conference - IWDLQ, Co-located with the KGSWC-2022

- [9] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, *Advances in neural information processing systems* 28 (2015).
- [10] H. Xu, K. Saenko, Ask, attend and answer: Exploring question-guided spatial attention for visual question answering, in: *European conference on computer vision*, Springer, 2016, pp. 451–466.
- [11] S. Zhang, M. Chen, J. Chen, F. Zou, Y.-F. Li, P. Lu, Multimodal feature-wise co-attention method for visual question answering, *Information Fusion* 73 (2021) 1–10.
- [12] D. Teney, P. Anderson, X. He, A. 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*, 2018, pp. 4223–4232.
- [13] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, L. 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*, 2018, pp. 6077–6086.
- [14] J.-H. Kim, K.-W. On, W. Lim, J. Kim, J.-W. Ha, B.-T. Zhang, Hadamard product for low-rank bilinear pooling, *arXiv preprint arXiv:1610.04325* (2016).
- [15] F. Ortiz-Rodriguez, S. Tiwari, R. Panchal, J. M. Medina-Quintero, R. Barrera, Mexin: multidialectal ontology supporting nlp approach to improve government electronic communication with the mexican ethnic groups, in: *DG. O 2022: The 23rd Annual International Conference on Digital Government Research*, 2022, pp. 461–463.
- [16] Z. Yu, J. Yu, C. Xiang, J. Fan, D. Tao, Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering, *IEEE transactions on neural networks and learning systems* 29 (2018) 5947–5959.
- [17] D. Gaurav, F. O. Rodriguez, S. Tiwari, M. Jabbar, Review of machine learning approach for drug development process, in: *Deep Learning in Biomedical and Health Informatics*, CRC Press, 2021, pp. 53–77.
- [18] Y. Zhang, J. Hare, A. Prügél-Bennett, Learning to count objects in natural images for visual question answering, *arXiv preprint arXiv:1802.05766* (2018).
- [19] J.-H. Kim, J. Jun, B.-T. Zhang, Bilinear attention networks, in: S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, R. Garnett (Eds.), *Advances in Neural Information Processing Systems*, volume 31, Curran Associates, Inc., 2018.
- [20] S. Gupta, S. Tiwari, F. Ortiz-Rodriguez, R. Panchal, Kg4astra: question answering over indian missiles knowledge graph, *Soft Computing* 25 (2021) 13841–13855.