

A Dual of Stacked Attention Networks (SAN's) and VGG-16 Model-Based Visual Question Answering Evaluation

Working notes for the ImageCLEFmed MEDVQA-GI Lab at CLEF 2023

Rohit Raj Gunti ^{1,2}, and Abebe Rorissa ³

^{1,2}University of Tennessee, 1345 Circle Park Drive, 412 Communications Building, Knoxville, TN, USA

²University of Tennessee, 1345 Circle Park Drive, 451 Communications Building, Knoxville, TN, USA

Abstract

The research aims to assess the number of open-source systems based on deep learning is growing, and various data preprocessing techniques are proposed. Considering the timeliness of Artificial Intelligence (AI) systems, particularly in terms of their immediate responsiveness and predictive capabilities, the evaluation of deep learning projects holds significant appeal for a diverse array of researchers, developers, and enthusiasts. The AI-based applications, like ChatGPT, draws interest due to their ability to rapidly generate informed responses and predictions, showcasing the potential for groundbreaking advancements in the field. The researchers can benefit from a wide selection of models and different data preprocessing methods without the need to start from scratch. Consequently, competitions like Medical Visual Question Answering for Gastrointestinal (MEDVQA-GI) Task, identifying Lesions in Colonoscopy images, organized by ImageCLEF Medical 2023, provide an opportunity for community-driven researchers to utilize multiple open-source algorithms such as Visual VGG-16 (Visual Geometry Group-16) Convolutional Neural Network model, and Long Short Term Memory (LSTM) models, enabling them to address complex challenges like identifying lesions in colonoscopy images effectively. Tasks like MEDVQA-GI allow participants to refine the literature and make the researchers work on new aspects of the field by adding multiple modalities to the picture. This study focuses on evaluating an open source system, namely Stacked Attention Networks for Image Question Answering, which utilizes a Task 1 approach, i.e., combines images and textual questions to generate textual answers, commonly applied in various research domains. The evaluation results, including assigned scores by the ImageClef MEDVQA-GI committee and other study observations, demonstrate that the selected system is highly suitable for Task 1. The system incorporates several preprocessing techniques, such as tokenization, word embedding using Word2Vec, preprocessing of questions and answers, question filtering, and feature extraction from images using the VGG16 model. Additionally, noteworthy observations were made regarding Task 2 (visual question generation) throughout the evaluation process. Overall, this research provides insights into the effectiveness of the Stacked Attention Networks for Image Question Answering open-source system for Task 1, highlighting the significance of the employed data preprocessing techniques and model selection. The findings contribute to the understanding of the capabilities of deep learning models and their applicability in addressing complex problems like identifying lesions in colonoscopy images. The results also offer valuable guidance to researchers, developers, and enthusiasts in choosing suitable open-source systems for their specific needs, saving them time and effort in model development.

Keywords

MEDVQA, Stacked Attention Networks, VGG-16, colonoscopy images, CNN classification


CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece

✉ rgunti@vols.utk.edu (R. R. G.); arorissa@utk.edu (A. R.)

🆔 0000-0002-5239-2419 (R. R. G.); 0000-0002-5300-617X (A. R.)



© 2023 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)

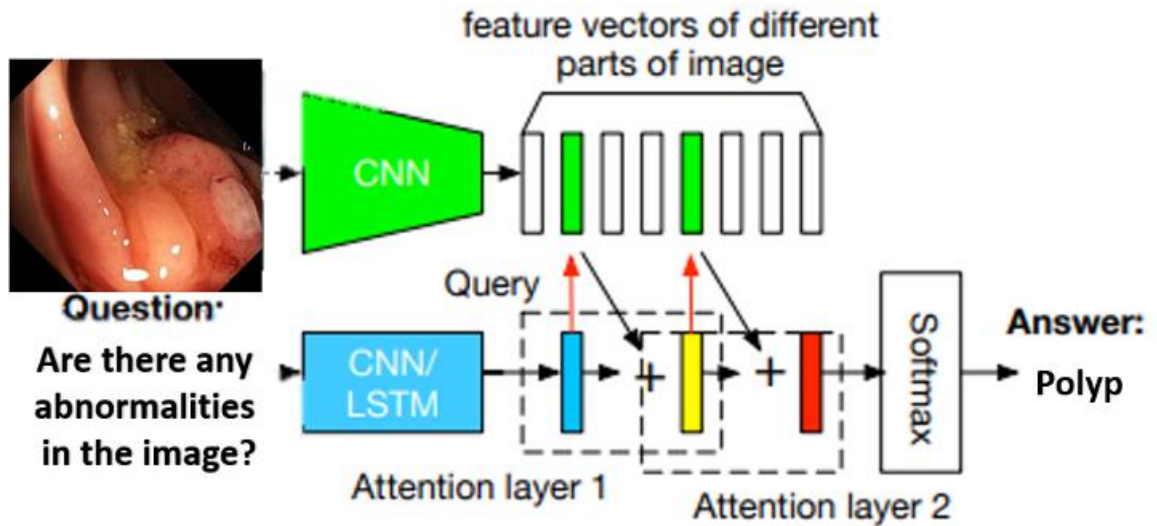
1. Introduction

Deep learning-trained open-source systems have gained significant popularity due to their effectiveness in various domains and applications [1]. These systems leverage different data preprocessing techniques to improve training and accuracy. Evaluating these systems is crucial to attracting researchers, developers, and enthusiasts, providing them with a diverse range of models and data preprocessing methods without starting from scratch. Competitions like the Medical Visual Question Answering for GI task (MEDVQA-GI) [2], organized by ImageCLEF medical [3] in 2023, offer a platform for community-driven researchers to utilize open-source algorithms, such as VGG-16, convolutional neural networks (CNNs), and LSTM models, to tackle complex challenges like identifying lesions in colonoscopy images.

This study focuses on the ImageCLEF MEDVQA-GI Task 1, which asks the participants to generate textual answers from image-question pairs [2]. The challenge primarily focuses on the application of image question-answering in the medical field, specifically in the domain of colonoscopy image analysis. The objective is to enhance the accuracy and usability of deep learning open-systems for identifying lesions in colonoscopy images. By incorporating multiple modalities such as visual question answering and visual question generation, the output of the analysis can be made more accessible to medical experts [2].

In addition to Task 1, the team participating in this study also proposes solutions for Task 2, where textual questions should be created from image-answer pairs. This expanded scope allows for a comprehensive exploration of the challenges associated with colonoscopy image analysis and facilitates advancements in the field [4, 5, 6]. Furthermore, similar approaches combining images and textual answers have proven successful in various research domains, including coral ecology for coral identification [7] and synthetic aperture radar imagery to identify natural disasters [8], among others. The paper [4] introduces SAN that extend the attention mechanism, successfully employed in image captioning and machine translation, to enable multi-step reasoning. The overall architecture of SAN is illustrated, consisting of three main components: the image model (utilizing a CNN to extract high-level image representations), the question model (employing a CNN or LSTM to extract a semantic vector of the question), and the stacked attention model (performing multi-step reasoning to locate image regions relevant to the question for answer prediction). The SAN operates by using the question vector to query the image vectors in the first visual attention layer. It combines the question vector and the retrieved image vectors to refine the query vector for querying the image vectors again in the second attention layer. The higher-level attention layer produces a more focused attention distribution, emphasizing regions more relevant to the answer. Finally, the image features from the highest attention layer are combined with the last query vector to predict the answer. This work makes three contributions. First, it proposes the stacked attention network as a solution for image QA tasks. Second, extensive evaluations on four image QA benchmarks demonstrate that the multiple-layer SAN outperforms previous state-of-the-art approaches by a significant margin. Third, the paper conducts a detailed analysis, visually showcasing the outputs of different attention layers of the SAN and illustrating the step-by-step process through which the SAN progressively focuses attention on relevant visual clues leading to the answer.

Yang et. al's work [4] (Stacked Attention Networks for Image Question Answering), shown in Figure 1, serves as the foundation for the study's investigation, training on different datasets, involving both text and images. In addition, the prior work [6] titled "A Dual Convolutional Neural Networks and regression model based Coral Reef Annotation and Localization" addresses the task of coral reef annotation and localization. Although the specific task differs from image question answering, it utilizes CNNs for image analysis and demonstrates the efficacy of combining CNNs with regression models.



(a) Stacked Attention Network for Image QA

Figure 1: Stacked Attention Network for ImageClef 2023 MedVQA-GI.

Additionally, Gunti and Rorissa's (2021) work [5] titled "A Convolutional Neural Networks based Coral Reef Annotation and Localization" explores the use of CNNs for coral reef annotation and localization tasks. While not directly related to image question answering, it aligns with the domain of image analysis and demonstrates the effectiveness of CNNs in similar contexts. Considering these works, our contribution lies in adapting the model architecture for multi-entry classification using categorical cross-entropy loss for Task 1 and Task 2. Furthermore, the preprocessing steps to train the model for virtual question generation tasks are updated to be compatible with the proposed multi-entry classification training.

By exploring these aspects and contributing to the existing literature [4, 5, 6], this study aims to gain insights into the performance of sparse categorical cross-entropy and categorical cross-entropy loss functions, as well as the potential of the proposed model for virtual question generation tasks. Previous works in related domains, such as coral reef annotation and localization utilizing CNNs and VGG, have been referenced to inform the training configuration adaptations and accuracy improvements made in this investigation - While the open-source system originally utilizes an Attention Network, the literature opts for a Dense network with a tanh activation function based on alternative implementations of the SAN.

Moreover, the research aims to investigate the feasibility of using the proposed model for training virtual question generators. Several contributions have been made to existing works in this regard. Firstly, the model architecture has been modified to enable training using "categorical_cross_entropy" for multi-entry classification in Task 1 and Task 2. Secondly, adjustments and attempts have been made to the data preprocessing steps to train the model for virtual question generation. Additionally, various combinations of learning rates, dropout rates, L1 and L2 normalization strengths, and model architectures have been

employed to train and save the model, which is available on GitHub².

In this working notes paper, we assess the selected system's performance on solving Task 1, as well as we make observations related to Task 2. The testing has been performed by the ImageClef MEDVQA-GI committee [2], assigning accuracy scores to each prediction and system based on its performance. Additionally, supporting observations are made throughout the evaluation process.

Research questions addressed.

RQ1 How effective is the proposed model for Task 1?

RQ2 Is data manipulation effective during training and testing?

RQ3 Is the investigation leading to the feasibility of the proposed model for the training virtual question generator and image segmentation, Task 2?

2. Materials and methods

In this study, two types of data were utilized: annotated data for developing the solution used for training and testing and test data without annotations for the final evaluation of the proposed solution.

To investigate the effectiveness of the system, the Hyper Kvasir dataset [12] (datasets.simula.no/hyper-Kvasir) was employed. This dataset was augmented with question-and-answer ground truth developed in collaboration with medical partners. It comprises a wide range of images covering the entire gastrointestinal tract, spanning from the mouth to the anus. The dataset encompasses various conditions, including abnormalities, surgical instruments, and normal findings, obtained from different procedures like gastroscopy, colonoscopy, and capsule endoscopy.

2.1 Data

Radiology images [9] play a crucial role in clinical decision-making and population screening, particularly for conditions like cancer. To assist clinicians in managing large volumes of images, automated systems that can answer questions about image contents have gained prominence. Visual Question Answering (VQA) in the medical domain is an emerging field of artificial intelligence that explores approaches to this form of clinical decision support. Before the competition challenge, the VQA-RAD dataset [8] was experimented with as the first manually constructed dataset, where clinicians asked naturally occurring questions about radiology images and provided reference answers. The images and questions were manually categorized, offering insights into clinically relevant tasks and the appropriate natural language to phrase them. Through evaluation with well-known algorithms, the superior quality of this dataset over automatically constructed ones is demonstrated.

```

gt.json
1 [{"ImageID": "clb0lbwzadoyc086u0brshvx5",
2   "Labels": [
3     {
4       "Question": "Are there any
5       abnormalities in the image?",
6       "AnswerType": "Text",
7       "Answer": [
8         "Polyp"
9       ]
10    }
11  ]
}

train.json
1 [{"qid": 1, "image_name": "clb0lbwzadoyc086u0brshvx5.jpg",
2   "Question": "Are there any abnormalities in the image?",
3   "AnswerType": "Text", "Answer": "Polyp"}, {"qid": 2,
4   "image_name": "clb0lbwzadoyc086u0brshvx5.jpg", "Question":
5   "Are there any anatomical landmarks in the image?",
6   "AnswerType": "Text", "Answer": "No"}, {"qid": 3,
7   "image_name": "clb0lbwzadoyc086u0brshvx5.jpg", "Question":
8   "Are there any instruments in the image?", "AnswerType":
9   "Text", "Answer": "Biopsy forceps"}, {"qid": 4,
10  "image_name": "clb0lbwzadoyc086u0brshvx5.jpg", "Question":
11  "Have all polyps been removed?", "AnswerType": "Yes/No",
12  "Answer": "No"}, {"qid": 5, "image_name":
13  "clb0lbwzadoyc086u0brshvx5.jpg", "Question": "How many
14  findings are present?", "AnswerType": "Number", "Answer":
15  "2"}, {"qid": 6, "image_name":

```

Figure 2: left side shows the gt.json format (multi-entry) right side shows the train.json format (single entry).

VQA tools that focus on improving patient care. By utilizing this dataset, the study can develop and refine algorithms that effectively address clinical/colon challenges and enhance medical decision-making.

2.2 Approach

Trained the models with two different preprocessing approaches:

1. Extracted the image, question, and answer vectors from the input JSON, as shown in the left side of Figure 2, format provided by the ImageCLEF MEDVQA - gt.json
 - gt.json is the JSON file with 2000SAN entries each with two values:
 - ImageID – example training Image name -"clb0lbwzadoyc086u0brshvx5."
 - Labels - consist of 18 questions with AnswerType and Answer values as represented in Figure 2.
2. Extracted the image, question, and answer vectors as the individual entries summing up to 36683 entries, as shown in the right side of Figure 2, from the manually manipulated JSON format as demonstrated by the open-source system - train.json:
 - train.json is the manipulated gt.json file annotated separately with 36863 entries separated by every question_id (qid) ranging from 1 - 36863 as represented in Figure 2.

The Stacked Attention Networks for Image Question Answering system incorporates several preprocessing techniques, such as tokenization, word embedding using Word2Vec, preprocessing of questions and answers [10], question filtering, and feature extraction from images using the VGG16 model [6]. The workflow of the chosen evaluated open-source system involves several steps, including loading GoogleNews vectors, creating an h5 file containing question vectors and labels, tokenizing questions and converting them into feature vectors, converting answers into labels, and storing the data in the h5 format. Furthermore, images are preprocessed using VGG16 preprocessing layers to obtain dimensions of ~~14~~ 14×512, which are subsequently reshaped to 196×512 [4].

The model architecture, shown in Figure 1, consists of passing the question layer through a Long Short-Term Memory (LSTM) and the preprocessed image through a dense network with a

tanh/ReLU activation function [11]. The resulting vectors are concatenated and passed through additional dense layers, followed by a final layer with a softmax activation function. For better performance [11] in multi-class classification (training the data as group entries - gt.json) using categorical cross-entropy loss function, the softmax activation function is replaced with ReLU, where the categorical cross-entropy loss function is applied, the softmax activation function is replaced with relu [11].

The sparse categorical cross-entropy loss function for single-entry classification (train.json) and categorical cross-entropy loss function for multi-entry classification (gt.json) tasks with 901 unique answer labels was considered for training in Task 1. Additionally, a comparison of training history is made between models trained with “sparse_categorical_crossentropy” and “categorical_crossentropy” loss functions¹.

¹<https://stats.stackexchange.com/questions/326065/cross-entropy-vs-sparse-cross-entropy-when-to-use-one-over-the-other>

Table 1

MedVQA-GI committee evaluated scores.

Task	Metric	Score
Global metrics	Accuracy	0.441
Question-based metric	Average Accuracy of all 18 questions	0.463
Image-based metric	Average Accuracy of images	0.441

Table 2

The evaluation of a single-entry classification by employing sparse categorical cross-entropy.

Metric	Score
F1-Score	0.9097830620656927
Accuracy	0.9152173913043479
Recall	0.9152173913043479
MCC	0.9040461706444576
MIOU	0.8516313376491703
Dice	328.0456197082703

3. Results and discussion

The evaluation results demonstrate that the selected system is highly suitable for Task 1 single-entry training. From the below ImageCLEF MEDVQA-GI evaluation, the model evaluated has decent scores overall score as shown in Table 1, which satisfies RQ1.

Overall, this study sheds light on the capabilities of deep learning models and their applicability in addressing complex challenges [2, 7]. It highlights the significance of data preprocessing techniques and model selection in achieving high performance. The results not only contribute

²<https://github.com/rohitgunti/MEDVQA-GI>

to the understanding of image analysis but also offer practical guidance for those interested in utilizing open-source systems for similar tasks, ultimately facilitating advancements in the field.

The findings of this research provide guidance to researchers, developers, and enthusiasts in selecting appropriate open-source systems for their specific needs. By leveraging pre-existing models available on Drive, they can save time and effort in model development. The effectiveness of the Stacked Attention Networks for the Image Question Answering system in Task 1 underscores the importance of utilizing proper data preprocessing techniques and training histories. The evaluation metrics presented in Table 2 demonstrate promising performance across various measures, including F1-score, accuracy, recall, Matthews correlation coefficient (MCC), and mean Intersection over Union (mIOU). These metrics collectively indicate a positive outcome for the model under evaluation. However, upon closer examination, it is apparent that the reported Dice coefficient, which is typically utilized to assess segmentation algorithms, deviates significantly from the expected range of 0 to 1. The Dice coefficient is calculated using a formula that compares the predicted segmentation to the ground truth segmentation, and a value of 1 signifies a perfect match.

The Dice coefficient value of 328.0456197082703 reported in the results is clearly outside the acceptable range and raises concerns about the accuracy of its computation. Consequently, this discrepancy necessitates further investigation to ascertain the reason behind this anomalous calculation, such as potential errors or inaccuracies in the implementation. Although the other metrics indicate favorable performance, the unreliable Dice coefficient warrants a comprehensive examination and potential refinement of the calculation method. It is essential to address this discrepancy in upcoming competitions of MEDVQA to ensure the reliability and validity of the reported results.

The data preprocessing techniques (train.json) and model selection play a significant role in achieving this performance. Moreover, valuable insights are gained regarding Task 2 and Task 3, reflecting on RQ3, and contributing to a better understanding of the capabilities of deep learning models in addressing complex problems like identifying lesions in colonoscopy images.

$$de = \text{Model}([e, \text{age}], [= a]) \quad (1)$$

$$de = \text{Model}([a, \text{age}], [= e]) \quad (2)$$

$$de = \text{Model}([a, \text{age}], [= e]) \quad (3)$$

The model equation for training, as represented by $\text{model} = \text{Model}([\text{ques}, \text{images}], [\text{out}])$ as represented in equation (1), signifies that the model is being trained by optimizing the parameters of the neural network based on the provided inputs (questions and images) to predict the output (answer probabilities).

Assumptions and attempts are made for training Task 2 using equation (2)

- It assumes that the answers (ans) are provided as input, along with the images (images), to train the model.
- The model aims to predict the questions (ques) corresponding to the given answers and images.

- This implies that there is a relationship between the provided answers and the target questions, which the model is expected to learn during training.

Assumptions and attempts are made for training Task 3 using equation (3)

- It assumes that masks (masks) are provided as input along with the images (images) for training the model.
- The model is designed to predict the questions (ques) based on the given masks and images.
- This implies that there is an underlying connection between the provided masks and the target questions, which the model is intended to capture during the training process.

In both cases, the model architecture, combined with the provided inputs and outputs, aims to learn the associations between the given data and the target questions. The training process involves adjusting the model's parameters to minimize the discrepancy between the predicted questions and the ground truth questions for the given inputs. To facilitate further implementation, the source code attempts for Task 2 and Task 3 is readily accessible on GitHub³.

Based on our findings, we conclude that from Task 1, we propose to apply the following two first equations as future work for improving the results on Task 1 and 2. Our team did not participate in Task 3 of the challenge, which asked to generate image segmentations from pairs of images and textual questions. For future work, we would still test equation 3 for solving Task 3.

Acknowledgments

The authors would like to express their deepest gratitude to Akarsh⁴, whose invaluable contributions and expertise have been instrumental in the understanding and implementation of the model.

References

- [1] I. H. Sarker, Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions, *SN Computer Science* 2 (2021) 420. doi:10.1007/s42979-021-00815-1.
- [2] S. A. Hicks, A. Storås, P. Halvorsen, T. de Lange, M. A. Riegler, V. Thambawita, Overview of imageclef medical 2023 – medical visual question answering for gastrointestinal tract, in: *CLEF2023 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece, 2023*.
- [3] B. Ionescu, H. Müller, A. Drăgulinescu, W. Yim, A. Ben Abacha, N. Snider, G. Adams, M. Yetisgen, J. Rückert, A. García Seco de Herrera, C. M. Friedrich, L. Bloch, R. Brün- gel, A. Idrissi-Yaghir, H. Schäfer, S. A. Hicks, M. A. Riegler, V. Thambawita, A. Storås, P. Halvorsen, N. Papachrysos, J. Schöler, D. Jha, A. Andrei, A. Radzhabov, I. Coman, V. Ko- valev, A. Stan, G. Ioannidis, H. Manguinhas, L. Ştefan, M. G. Constantin, M. Dogariu, J. Deshayes, A. Popescu, Overview of ImageCLEF 2023: Multimedia retrieval in medical, socialmedia and recommender systems applications, in: *Experimental IR Meets Multilin- guality, Multimodality, and Interaction, Proceedings of the 14th International Conference of the CLEF Association (CLEF 2023), Springer Lecture Notes in Computer Science LNCS, Thessaloniki, Greece, 2023*.

³<https://github.com/rohitgunti/MEDVQA-GI>

⁴<https://github.com/uakarsh/med-vqa>

- [4] Z. Yang, X. He, J. Gao, L. Deng, A. Smola, Stacked attention networks for image question answering, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 21–29.
- [5] R. Gunti, A. Rorissa, A convolutional neural networks based coral reef annotation and localization., in: CLEF (Working Notes), 2021, pp. 1229–1238.
- [6] R. R. Guntia, A. Rorissaa, A dual convolutional neural networks and regression model based coral reef annotation and localization, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the 13th International Conference of the CLEF Association (CLEF 2022), LNCS Lecture Notes in Computer Science, Springer, Bologna, Italy, 2022.
- [7] J. Chamberlain, A. Garcia Seco De Herrera, A. Campello, A. Clark, Imageclefcoral task: coral reef image annotation and localisation, in: CEUR Workshop Proceedings, volume 3180, 2022, pp. 1318–1328.
- [8] E. Nemni, J. Bullock, S. Belabbes, L. Bromley, Fully convolutional neural network for rapid flood segmentation in synthetic aperture radar imagery, *Remote Sensing* 12 (2020) 2532.
- [9] J. J. Lau, S. Gayen, A. Ben Abacha, D. Demner-Fushman, A dataset of clinically generated visual questions and answers about radiology images, *Scientific Data* 5 (2018) 180251. doi:10.1038/sdata.2018.251.
- [10] Z. Rahimi, M. M. Homayounpour, The impact of preprocessing on word embedding quality: A comparative study, *Lang. Resour. Eval.* 57 (2022) 257–291. doi:10.1007/s10579-022-09620-5.
- [11] A. F. Agarap, Deep learning using rectified linear units (relu), arXiv preprint arXiv:1803.08375 (2018).
- [12] Borgli, H., Thambawita, V., Smedsrud, P. H., Hicks, S., Jha, D., Eskeland, S. L., Randel, K. R., Pogorelov, K., Lux, M., Nguyen, D. T. D., & others. (2020). HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. *Scientific data*, 7(1). doi:10.1038/s41597-020-00622-y