Language-based Colonoscopy Image Analysis with Pretrained Neural Networks

Notebook for the Medical Visual Question Answering for GI Task - MEDVQA-GI Lab at CLEF 2023

Patrycja Cieplicka[1,*,†], Julia Kłos[1,†], Maciej Morawski[1,†] and Jarosław Opała[2,†]

[1]*Independent Researcher, Warsaw, Poland*
[2]*Independent Researcher, Wrocław, Poland*

Abstract

In this paper, our solutions for tasks from the ImageCLEF 2023 Challenge Medical Visual Question Answering for GI Task - MEDVQA-GI [1] are presented. The aim of the Visual Question Answering (VQA) task was to generate answers based on the given colonoscopy image and corresponding questions. For this problem, a multilabel classification approach was proposed. The solution included neural network training with the utilization of pretrained encoders used to generate embeddings of image and text. The Visual Question Generation (VQG) task was to generate text questions from a given colonoscopy image and answer. To solve this task, we also applied a multilabel classifier consisting of two pretrained encoders used to create embeddings of images and text. The Visual Location Question Answering (VLQA) task was focused on generating segmentation masks covering the area of abnormality based on a given colonoscopy image and specific question. For this purpose, two separate pretrained semantic segmentation models were fine-tuned. For the VQA task, the proposed model achieved an accuracy of 0.82 on the test dataset. For the VLQA task an accuracy of 0.95, a Jaccard index of 0.67, and a Dice coefficient of 0.68 were achieved on the test dataset.

Keywords

colonoscopy images, transfer learning, Visual Question Answering, Visual Question Generation, Visual Location Question Answering, ImageCLEF

1. Introduction

One of the most well-liked uses of artificial intelligence in medicine is the identification of abnormalities. The research has thus far mostly concentrated on single-image or video analysis. Through the addition of several modalities to the work, like text, we want to add a fresh perspective to the area of identification of lesions in colonoscopy images. The ImageCLEF 2023 Challenge [2] Medical Visual Question Answering for GI Task, MEDVQA-GI [1], was focused on generating answers to questions and questions to answers about colonoscopy images. The

CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece
*Corresponding author.
†These authors contributed equally.
patrycja.cieplicka@gmail.com (P. Cieplicka); julia.klos159@gmail.com (J. Kłos); mpfmorawski@gmail.com (M. Morawski); jaroslaw.edward.opala@gmail.com (J. Opała)
https://github.com/paatrycjaa/ (P. Cieplicka); https://github.com/julklos/ (J. Kłos); https://github.com/mpfmorawski (M. Morawski)

first and second part of the task was to generate text output, while the third part of the task was to locate the abnormalities providing segmentation masks. The intention was to make medical image analysis simpler for medical professionals by combining text and visual data.

2. Visual Question Answering

In the Visual Question Answering (VQA) task, the goal was to generate accurate answers based on the given image and corresponding questions. Our proposed approach involved the creation of a model designed to handle both image and text data.

2.1. Review of existing solutions

Existing solutions to the VQA task have seen significant progress, as evidenced by papers such as [3], [4], and [5]. These approaches leverage deep learning architectures, attention mechanisms, and reasoning techniques to effectively extract visual features and address challenges in VQA. According to Marino et al. [6], VQA can be classified into two categories. First, we can make use of symbolic knowledge, which can be represented using graphs. In this way implicit knowledge is encoded in the weights of a model trained using different datasets. The second case is supported by transformer-based language models like BERT [7]. However, further research is needed in order to progress and improve generalization.

2.1.1. Dataset

For both the VQA and Visual Question Generation (VQG) tasks, the training dataset contained 2,000 images. 500 images had 19 paired questions, and 1,500 images had 18 matching questions. For each question, at least one answer was provided, sometimes multiple answers were possible for a single question. The answers could be one of the possible types: number, text, yes/no and segmentation. Segmentation-type answers and questions were skipped in VQA and VQG sub-tasks as there are part of the Visual Location Question Answering (VLQA) task.

Both VQA and VQG were interpreted as multilabel classification tasks, where multiple outputs are possible as correct answers/questions. The dataset was split into training and validation sets using an 80:20 ratio. This division allowed for effective model training on the majority of the data while reserving a smaller portion for evaluating the model's performance.

2.2. Presented solution

The schema of our solution for the VQA task is shown in Figure 1. Details of the solution are presented in the following subsections.

2.2.1. Pretrained Transformers

Pretrained Transformers-based models were utilized in order to generate embeddings that were able to extract meaningful representations from both the image and text inputs. To achieve this, two separate pretrained transformer networks were employed: the microsoft/beit-base-patch16-224-pt22k-ft22k [8] model was used as an image encoder. This Vision Transformer model

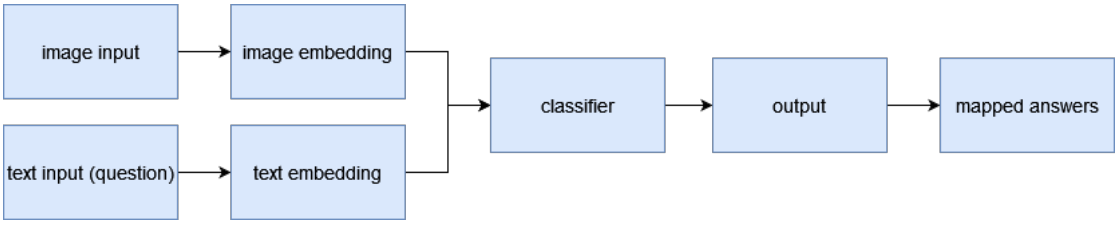

Figure 1: Schema of the proposed VQA solution.

had been specifically trained to process images and generate rich embeddings that capture intricate visual features. The albert-base-v2 [9] was chosen as the text encoder. This transformer-based architecture excels at encoding textual information and producing contextualized word embeddings. The choice of the models was based on experiments considering classification quality. Models were trained for 12 epochs, applying a mini-batch size of 32.

2.2.2. Concatenation

After obtaining the image and text embeddings, they were concatenated into a single representation. This step allowed us to fuse the visual and textual information, enabling the subsequent layers to effectively process and analyze the combined features.

2.2.3. Classification Layers

The concatenated embeddings were passed to the following layers for classification:

1. *Dense layer 1*: The combined embeddings were passed through a dense layer consisting of 4096 neurons. Rectified Linear Unit (ReLU) activation was applied to introduce non-linearity and enhance the network's expressive power. To prevent overfitting, a dropout rate of 0.5 was incorporated, which randomly omits connections during training.
2. *Dense layer 2*: Next, the outputs from the previous layer were fed into a second dense layer with 2048 neurons. Again, ReLU activation and dropout with a rate of 0.5 were applied to promote non-linearity and prevent overfitting.
3. *Final dense layer*: Finally, the processed embeddings were fed into a dense layer with a number of neurons equal to the total number of labels in the dataset, which in the described case was 63. This layer employed an appropriate activation function suitable for multi-label classification.
4. *Loss function*: In order to train the network, the binary cross entropy loss function was utilized. This loss function effectively measures the dissimilarity between predicted answers and ground-truth labels, allowing the network to optimize its parameters for accurate prediction during multi-label classification training.

Table 1
Results of the VQA model chosen for submission

Metric	Validation dataset	Final test dataset
Accuracy	0.8386	0.8193

Table 2
Results of the VQA model chosen for submission on final test dataset - by chosen questions

Question	Accuracy
What color is the anatomical landmark?	0.9990
What color is the abnormality?	0.5784
Where in the image is the instrument?	0.7585
Where in the image is the abnormality?	0.6615

2.3. Results

Models performance is compared with the use of accuracy metric defined as:

$$\text{accuracy} = \frac{\text{number of correct predictions}}{\text{total number of predictions}}$$

. Table 1 presents the performance of our proposed approach. On the validation dataset, our model achieved an impressive accuracy score of 0.8386. Furthermore, on the more challenging competition final test dataset, our model demonstrated robust performance, achieving an accuracy score of 0.8193.

In table 2, there are presented chosen questions with corresponding accuracy scores on the final test dataset. It is interesting that the model achieves a very high accuracy on one of the questions about color while struggling with the other. These results, while showing areas for enhancement, indicate the generalization capability of our approach and its potential for real-world applications. The achieved accuracy scores highlight the advancements made in VQA research and provide a strong foundation for further improvements in this field.

3. Visual Question Generation

In the VQG subtask, the aim was to generate text questions from a given colonoscopy image and text answer. In VQG, most images could have many questions generated, even if a model is guided by supporting second input such as an answer. Providing answer narrows the space of expected questions. Nevertheless, the evaluation of VQG is a non-trivial task, that requires checking grammatical coherence and relevance to the given image of generated question. Sometimes, especially in the area of medicine, deep domain knowledge is required.

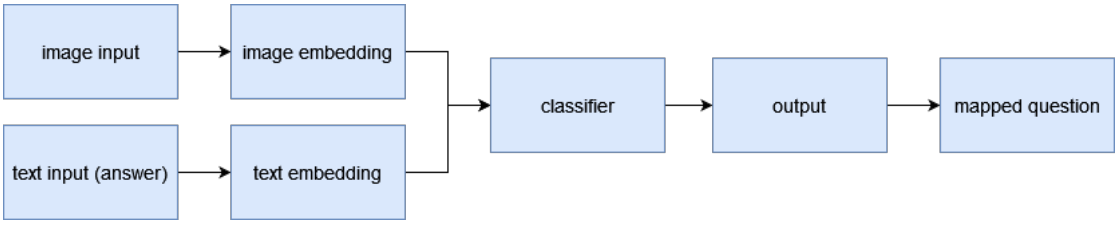

Figure 2: Schema of the proposed VQG solution.

3.1. Review of existing solutions

Recently, many multimodal tasks have been in the area of interest of the research community, such as VQA, multimodal translation, and image captioning. VQG domain remains under-researched despite the great focus on both text and image processing.

The pioneering paper [10] in the field of VQG used model-generated captions and an RNN-based encoder-decoder system to create questions. Only a few studies have looked into VQG since then. There was shown how well a Generative Adversarial Network (GAN) can be used in VQG systems, enabling non-deterministic and diverse outputs [11]. The model suggested by Jain et al. [12] used a Variational Autoencoder (VAE) rather than a GAN, however, their superior results necessitate the use of a target response during inference. In order to get around this unrealistic need, in the proposed solution by Krishna [13], answer categories were added to the VQA [14] dataset and suggested a model that does not require a response during inference. Considering that their design takes advantage of the target's information as input. More recently, Scialom et al. [15] improved a BERT ([7]) model using model-based object attributes and actual image captions to evaluate the cross-modal performance of pre-trained language models.

3.2. Presented solution

For the VQG task, the best results were obtained for the same encoders as for the VQA task: microsoft/beit-base-patch16-224-pt22k-ft22k as image encoder and albert-base-v2 used as text encoder. Features created by two embeddings were concatenated into a single representation. Then, they were passed to the following classification layers: two dense layers consisting of each of 8192 neurons. ReLU was applied as an activation function. The final dense layer consists of 16 neurons, which is equal to the total number of possible questions in the dataset. This layer employed an appropriate activation function suitable for multi-label classification.

Having a very limited dataset that includes only 16 unique questions, we decided to apply a multimodal classifier to solve this task. Given image **i** and answer **a**, we expect to receive one of 16 possible questions about colonoscopy. With two different inputs, our architecture makes use of two separate pretrained embeddings for text and image. Encoded data are joined to create an enormous vector of features including information about both the image and the answer. Such a vector of features is an input for fully connected layers. The schema of the proposed structure is shown in Figure 2, and the idea is identical to the one proposed for VQA, described in Section 2.

Table 3
Results of VQG model chosen for submission

Metric	Validation dataset
Accuracy	0.4610

3.3. Results

In Table 3, the performance of our VQG approach is presented. Using the validation dataset, our model reached an accuracy of 0.4610, which indicates that this approach might not be the best option for this task and that there is a lot of room for improvement. The results for the final test dataset have not been published so far, so they will be filled in the future.

4. Visual Location Question Answering

The VLQA task focus on generating visual outputs in response to textual queries, specifically through the segmentation of relevant regions in the image. In this particular VLQA task, the goal was to generate segmentation masks covering the area of abnormality based on a given colonoscopy image and question defined as "Where exactly in the image is the [ABNORMALITY] located?".

4.1. Review of existing solutions

VLQA task can be described as a language-based semantic segmentation that aims to generate pixel-level segmentation masks based on both image content and textual instructions (in our case, colonoscopy images and text questions).

One of the first papers that introduced this problem [16] proposed an end-to-end trainable recurrent long short-term memory (LSTM) and convolutional neural network (CNN) model that jointly learns to process visual and linguistic information. They indicated that this novel task of language-based segmentation differs from traditional semantic segmentation because it is not limited to a fixed set of categories and/or rectangular regions.

As attention mechanisms [17] have been shown to be a powerful technique in deep learning models, particularly in natural language processing tasks, Gong et al. [18] proposed in 2019 a cross-modal self-attention (CMSA) module that effectively captures the long-range dependencies between linguistic and visual features to segment the object referred by the language expression in the image.

In 2022, the LSeg model [19] was proposed, which uses a text encoder to compute embeddings of descriptive input labels together with a transformer-based image encoder that computes dense per-pixel embeddings of the input image. This approach is described as a zero-shot semantic segmentation method, that aims to segment unseen objects without any additional samples of novel classes.

The text-guided image manipulation task is similar to the language-based semantic segmentation task in terms of input (text + image) and output (image). Since there are approaches that

are using generative models [20], [21] to solve this task, they can be also one of the solutions for language-based semantic segmentation task.

However, if the number of labels and associated questions are known, another approach can be traditional semantic segmentation, which is one of the fundamental topics in computer vision and it aims to assign semantic labels to every pixel in an image. Most of the solutions use fully convolutional networks, that have spatial pyramid pooling module [22] or encoder-decoder structure [23], [24]. The most known state-of-the-art solutions are Detectron2 [25], DeepLabv3+ [26], transformed-based model BEiT-3 [27] and the newest Meta AI model called Segment Anything (SAM) [28].

As the number of labels and corresponding questions was known, we decided to implement a traditional semantic segmentation model for the VLQA task.

4.2. Dataset

The training dataset provided includes 683 questions for which the answer is segmentation mask, i.e., 500 questions about the exact location of polyps and 183 questions about the exact location of instruments.

The provided masks are binary images, which size was the same as the colonoscopy image for which the question about the exact location of abnormality was asked. The images in the training dataset varied in size, i.e., with widths from 396 to 1920 pixels, and heights from 352 to 1072 pixels.

A total of 983 colonoscopy images were used for training and validation of the segmentation models, 683 images to which the appropriate mask was assigned (500 images containing polyps, 183 images containing instruments), and 300 images that contained neither polyps nor instruments. The images were then split into a training set (90% of the images for each category) and a validation set (10% of the images for each category) to make sure that the metrics on which the best models were selected were determined on images that were not included in the training process. Additionally, during training image augmentation techniques commonly used in computer vision were applied - random cropping, random horizontal and vertical flipping, random rotations, and random adjustment of brightness.

The test dataset contained 1,949 different colonoscopy images including images where polyps and instruments can be found. The test dataset did not specify which photos contained abnormalities.

4.3. Presented solution

Since the goal of the task was the segmentation of two specific abnormalities, i.e., instruments and polyps, we decided to use the current state-of-the-art solution in this field - Detectron2 [25] - a platform for object detection, segmentation and other visual recognition tasks developed by Meta AI Research.

Due to the fact that the images in the dataset did not always contain masks for both abnormalities, we trained two separate segmentation models, i.e., one for localizing polyps and one for localizing instruments (Figure 3). Each model was separately trained on images assigned for its category and images without any abnormality. In addition, hyperparameters (number

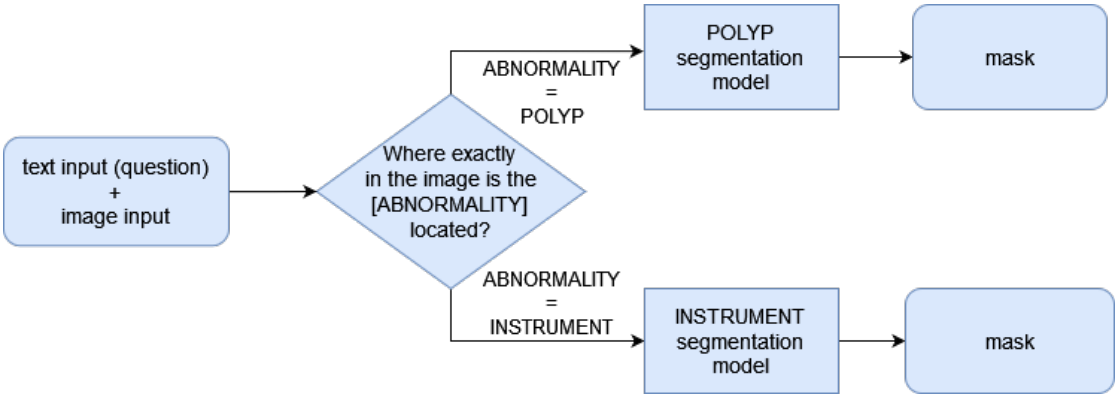

Figure 3: Schema of the proposed VLQA solution.

of epochs, learning rate, and the minimum confidence score required for an object detection called threshold) were optimized separately for both models. The best model was selected by the highest IoU value on the validation dataset. For instrument segmentation, a model trained for 1,000 epochs was chosen, with an initial learning rate of 0.00025 and a threshold of 0.8. For polyp segmentation, a model trained for 1,500 epochs was chosen, with an initial learning rate of 0.00025 and a threshold of 0.85.

To train the models, we have used transfer learning techniques instead of starting from scratch and training a model on a new dataset. By utilizing this technique, we could significantly reduce the amount of data and time required for training. This was especially important because of the small size of the provided dataset. During training, we have used a Mask R-CNN object detection model [29] pretrained on the COCO Dataset for Semantic Image Segmentation [30].

4.4. Results

On the validation dataset, metrics were calculated separately for both models. Accuracy, mean IoU and Dice coefficient were calculated only for questions for which the ground-truth masks contained an abnormality. In addition, for each model, it was determined for what portion of the images, that did not contain abnormality, the model incorrectly detected its presence. The results of the models chosen for submission are shown in Table 4.

The results on the test dataset are shown in Table 5. It is worth mentioning that the following assumptions were added in the calculation of these metrics compared to the calculation of metrics on the validation dataset: if there was no abnormality in the generated mask and the ground truth mask, then all metrics were equal to 1, and if the generated mask contained an object and the ground truth mask did not, then all metrics (except accuracy) were equal to 0.

Figure 4 shows example images from the test dataset and masks that were generated from the question "Where exactly in the image is the polyp located?". Example A shows the correct segmentation of a polyp. The generated mask is close to the ground-truth mask. Example B shows that the polyp was detected, but the mask is larger than it should be. Besides the polyp, the mask also includes a yellow substance. Example C, on the other hand, shows the detection of a polyp that was not marked in the ground-truth mask from the test dataset. However, it is

Table 4
Results of the VLQA models chosen for submission on validation datasets

Model	Accuracy	IoU	Dice	Unexpected abnormality detection
INSTRUMENT	0.9911	0.8766	0.9320	0% of images without abnormalities
POLYP	0.9427	0.7252	0.8022	6.667% of images without abnormalities

Table 5
VLQA submission results on test dataset

Metrics	Accuracy	Jaccard	F1-score	Recall	Precision	Dice
Results	0.9521	0.6660	0.6769	0.6810	0.6796	0.6769

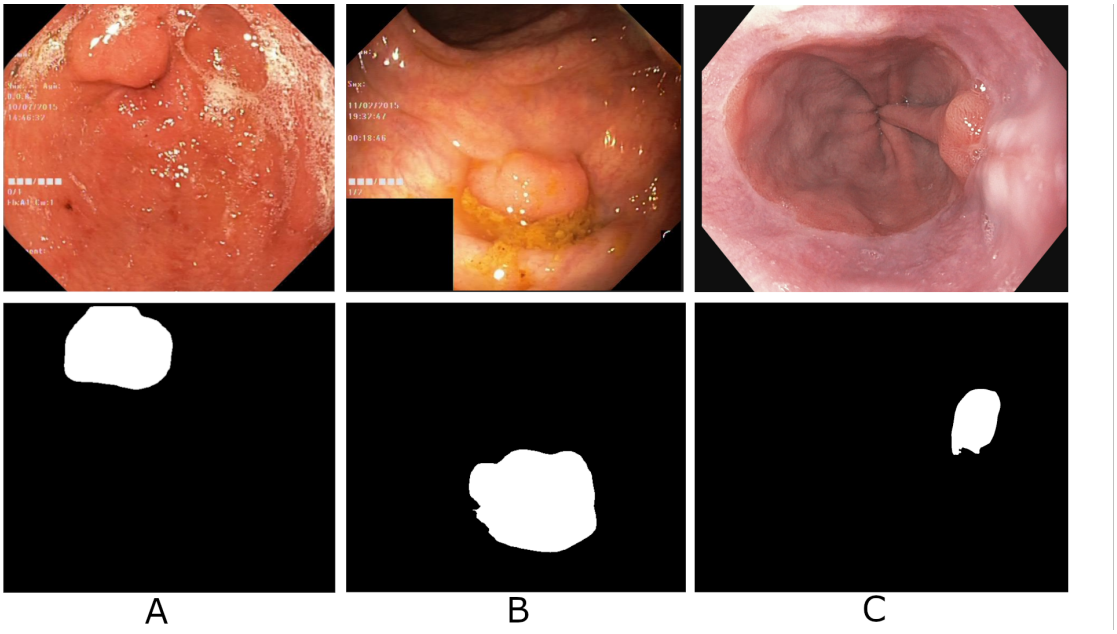

Figure 4: Examples of masks generated by the model trained for polyp segmentation. All input images (A, B, C) from the test dataset.

important to notice that the model's discovery is similar to polyp.

5. Conclusions

In conclusion, this paper presented our solutions for each of the subtasks in the ImageCLEF 2023 MEDVQA-GI challenge. Our solution to the VQA task relied heavily on two crucial factors: the selection of pretrained encoders, which facilitated efficient feature extraction, and the careful tuning of the dense layer sizes and dropout rates, ensuring optimal information processing. These factors greatly influenced the performance and accuracy of the approach. The experiments

showed that an inaccurate choice of encoders may notably lower the quality of the solution. The achieved accuracy on the challenge test dataset did not vary significantly from the accuracy on the development test dataset, which shows the generalization capability of the model.

The VQG system presented generates questions based on images and answers using a combination of text transformers, vision transformers, and a multi-label classifier. However, there is still room for improvement. Future work could focus on further improving the text and image encoding models and exploring different classification models. There are various approaches to VQG problems, including the use of generative adversarial networks. Although, due to the fact that in medical problems the precision of solution is crucial, we considered a predefined set of questions, thus our approach was multi-label classification model training.

In our solution to the VLQA task, we used the state-of-the-art platform in the field of object detection and segmentation, Detectron2 [25], to tackle the task of segmenting two specific abnormalities: instruments and polyps. Due to the provided dataset, where not all images contained masks for both abnormalities, we decided to train two separate segmentation models. Future work could focus on expanding and completing the dataset. A larger dataset used for training could probably significantly improve the obtained results. Moreover, if masks of both instruments and polyps were available for all colonoscopy images, it would be possible to train and test a multiclass segmentation model. In addition, it would be worth extending the study to check the results for other models, particularly language-based semantic segmentation models.

Acknowledgments

We would like to express our sincere gratitude to the organizers of Medical Visual Question Answering for GI Task for their support and assistance throughout the research and publication process. In particular, to Steven A. Hicks and Michael A. Riegler for their guidance and prompt responses to our inquiries.

References

[1] S. A. Hicks, A. Storås, P. Halvorsen, T. de Lange, M. A. Riegler, V. Thambawita, Overview of ImageCLEFmedical 2023 – Medical Visual Question Answering for Gastrointestinal Tract, in: CLEF2023 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece, 2023.

[2] B. Ionescu, H. Müller, A. Drăgulinescu, W. Yim, A. Ben Abacha, N. Snider, G. Adams, M. Yetisgen, J. Rückert, A. Garcia Seco de Herrera, C. M. Friedrich, L. Bloch, R. Brüngel, 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. Kovalev, 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 Multilinguality, 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] V. Kazemi, A. Elqursh, Show, ask, attend, and answer: A strong baseline for visual question answering, ArXiv abs/1704.03162 (2017).
[4] H. Nam, J.-W. Ha, J. Kim, Dual Attention Networks for Multimodal Reasoning and Matching, 2017. arXiv:1611.00471.
[5] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, M. Rohrbach, Multimodal compact bilinear pooling for visual question answering and visual grounding, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Austin, Texas, 2016, pp. 457–468.
[6] K. Marino, X. Chen, D. Parikh, A. Gupta, M. Rohrbach, KRISP: Integrating Implicit and Symbolic Knowledge for Open-Domain Knowledge-Based VQA, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14106–14116.
[7] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2019. arXiv:1810.04805.
[8] H. Bao, L. Dong, S. Piao, F. Wei, BEiT: BERT Pre-Training of Image Transformers, 2022. arXiv:2106.08254.
[9] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, R. Soricut, ALBERT: A Lite BERT for Self-supervised Learning of Language Representations, 2020. arXiv:1909.11942.
[10] S. Zhang, L. Qu, S. You, Z. Yang, J. Zhang, Automatic Generation of Grounded Visual Questions, 2017. arXiv:1612.06530.
[11] Z. Fan, Z. Wei, S. Wang, Y. Liu, X. Huang, A Reinforcement Learning Framework for Natural Question Generation using Bi-discriminators, in: Proceedings of the 27th International Conference on Computational Linguistics, Association for Computational Linguistics, Santa Fe, New Mexico, USA, 2018, pp. 1763–1774.
[12] U. Jain, Z. Zhang, A. Schwing, Creativity: Generating Diverse Questions using Variational Autoencoders, 2017. arXiv:1704.03493.
[13] R. Krishna, M. Bernstein, L. Fei-Fei, Information Maximizing Visual Question Generation, 2019. arXiv:1903.11207.
[14] A. Agrawal, J. Lu, S. Antol, M. Mitchell, C. L. Zitnick, D. Batra, D. Parikh, VQA: Visual Question Answering, 2016. arXiv:1505.00468.
[15] T. Scialom, P. Bordes, P.-A. Dray, J. Staiano, P. Gallinari, What BERT Sees: Cross-Modal Transfer for Visual Question Generation, 2020. arXiv:2002.10832.
[16] R. Hu, M. Rohrbach, T. Darrell, Segmentation from Natural Language Expressions, in: Computer Vision – ECCV 2016, Springer International Publishing, Cham, 2016, pp. 108–124.
[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, I. Polosukhin, Attention is All you Need, in: Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017.
[18] H. Gong, G. Chen, S. Liu, Y. Yu, G. Li, Cross-Modal Self-Attention with Multi-Task Pre-Training for Medical Visual Question Answering, in: Proceedings of the 2021 International Conference on Multimedia Retrieval, ICMR '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 456–460.
[19] B. Li, K. Q. Weinberger, S. Belongie, V. Koltun, R. Ranftl, Language-driven Semantic Segmentation, in: International Conference on Learning Representations, 2022.
[20] R. Togo, M. Kotera, T. Ogawa, M. Haseyama, Text-guided style transfer-based image

manipulation using multimodal generative models, IEEE Access 9 (2021) 64860–64870.
[21] T. Zhang, H.-Y. Tseng, L. Jiang, W. Yang, H. Lee, I. Essa, Text as Neural Operator:Image Manipulation by Text Instruction, in: Proceedings of the 29th ACM International Conference on Multimedia, MM '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 1893–1902.
[22] K. He, X. Zhang, S. Ren, J. Sun, Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition, in: Computer Vision – ECCV 2014, Springer International Publishing, Cham, 2014, pp. 346–361.
[23] O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional Networks for Biomedical Image Segmentation, in: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, Springer International Publishing, Cham, 2015, pp. 234–241.
[24] V. Badrinarayanan, A. Kendall, R. Cipolla, SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence 39 (2017) 2481–2495.
[25] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, R. Girshick, Detectron2, https://github.com/facebookresearch/detectron2, 2019.
[26] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation, in: Computer Vision – ECCV 2018, Springer International Publishing, Cham, 2018, pp. 833–851.
[27] W. Wang, H. Bao, L. Dong, K. Aggarwal, S. Singhal, S. Som, F. Wei, J. Bjorck, Z. Peng, Q. Liu, O. K. Mohammed, Image as a Foreign Language: BEiT Pretraining for All Vision and Vision-Language Tasks, 2022.
[28] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, R. Girshick, Segment Anything, arXiv:2304.02643 (2023).
[29] K. He, G. Gkioxari, P. Dollár, R. B. Girshick, Mask R-CNN, CoRR abs/1703.06870 (2017). URL: http://arxiv.org/abs/1703.06870. arXiv:1703.06870.
[30] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Doll'a r, C. L. Zitnick, Microsoft COCO: Common Objects in Context, CoRR abs/1405.0312 (2014). URL: http://arxiv.org/abs/1405.0312. arXiv:1405.0312.

A. Online Resources

The source code is available via GitHub:

- https://github.com/paatrycjaa/ImageCLEF2023-MEDVQA-GI-VisionQAries,