IRLab@IIT_BHU at MEDIQA-MAGIC 2024: Medical Question Answering using Classification model

Arvind Agrawal[1,*], Sukomal Pal[2]

Abstract

This paper presents our submission to the MEDIQA2024 Multilingual and Multimodal Medical Answer Generation (M3G) shared task [1]. The paper presents two types of approaches: 1) Generation-based and 2) Classification-based. The generation-based model passed the title and content as text embeddings and images as visual embeddings as a prompt to a pre-trained LLM. The Classification model utilized the medically relevant NER tags obtained from the queries using pre-trained NER models and converted these tags and images to embeddings using CLIP text and vision encoders. These embeddings were passed through Bi-LSTM and an MLP to obtain final representations, which were combined to form query embeddings. The query and label embeddings were used to train the model using triplet loss. The answer label was predicted as the most similar label embedding to the query embedding using cosine similarity. The generative approach performs poorly compared to the classification because less training data is available. Our classification-based approach utilizes manually labeled data (160 labels) to predict the test set answers with a deltaBLEU-score of 4.829 and was ranked 2nd on the leaderboard.

Keywords

Med-VQA, MAGIC, deltaBLEU

1. Introduction

Telemedicine consultation for dermatology became very common during the pandemic to lessen the risk of human-human contact [2],[3]. Patients used to consult doctors by phone, which proved a viable solution. People have started to believe in telemedicine consultation. People have been integrating AI to assist doctors and telemedicine consultations through Medical VQA systems, which can assist both doctors and patients. Many medical-VQA systems have been developed utilizing both classification and generation-based approaches. However, these models have been developed with the medical VQA datasets available, which specifically cater to radiology[4], pathology[5], orthopedics, and the gastrointestinal datasets[6]. However, in the case of dermatology, not much exploration has been done due to the datasets available being very low-resource, and the traditional systems developed will not be able to perform well on the low-resource dataset provided by the organizers. The consumer answering systems developed for dermatology[7],[8] focused only on text (questions, textual context) and did not explore the vision modality (Images, Videos). This limits the model from considering the vision features that can provide fine visual details that are captured through images and are often difficult for the user to explain through text.

This paper focuses on developing a model capable of generating free-form text in response to a given question asked by the user, specifically focusing on clinical dermatology. The proposed model will be able to consider the vision modality but will not necessarily require it. The work described in this paper is presented as a participation of the ImageCLEF-2024-MEDIQA-MAGIC[1] shared task. The task focuses on the problem of Multimodal And Generative TelemedICine (MAGIC) in the area of dermatology. The challenge tackled the generation of an appropriate textual response to the query[9] asked by the user, along with the clinical context provided in the form of both text and images.

In this paper, we propose both classification and generation-based approaches. The generation-based approach is based on the work by Bazi et al. [10]. The classification-based approach was tried because the generation-based approach could not produce good results because of the low resource data provided

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

*Corresponding author.

arvind.agrawal.cse19@itbhu.ac.in (A. Agrawal); spal.cse@itbhu.ac.in (S. Pal)

by the organizers. The classification-based approach was used to predict answer classes corresponding to the answers that were manually labeled into 160 classes. The predicted class was later converted into long-form text based on a manually prepared label → Long-text answers mapping. Our approach performed 2nd best during the competition with a deltaBleu score of 4.829.

The paper is organized as follows; we present a literature review in Section 2. In Section 3, we provide details of the dataset provided. In Section 4 we explain how we pre-processed the data to fit our needs. In Section5, we provide the details of our participation in the ImageCLEF-2024 MAGIC shared task. In Section 6, we present the results and our corresponding analysis. Following this, the thesis is concluded with the future works in Section 7.

2. Related works

Telemedicine consultation became a go-to option for many during the pandemic. There were many studies describing the experiences of patients who had a consultation without human-human contact, and most found it satisfying. This not only decreased the importance of in-person visits but also opened many doors. Many people have started integrating AI with consumer answering systems in the medical domain. These systems can be divided into two categories: Classification-based and Generation-based approaches.

Classification-based Medical-VQA categorizes questions and answers for efficient responses, utilizing techniques like CNN, RNN, Bi-LSTM, and transformers to predict classes. Key contributions include using CNNs for visual feature extraction from medical images alongside RNNs for question processing [11]. Hierarchical deep multimodal networks enhance efficacy in classification and response generation through hierarchical attention mechanisms [12]. MMBert uses a transformer-style architecture for richer image and text representations [13]. Multimodal-Multihead Self Attention combines text and image embeddings for classification [14]. Caption Aware Medical-VQA integrates image-captioning models with BAN for superior performance [15]. A new dataset focusing on chest radiography images introduces relation graphs for improved reasoning [16].

Generation-based Medical Visual Question Answering (MedVQA) approaches focus on generating precise, contextually appropriate free-form text responses using advanced deep-learning techniques. The Q2A transformer, though claimed generative, faces computational challenges as classes increase, utilizing learnable answer class embeddings and a SWIFT encoder for fine-grained features [17]. CGMVQA switches between classification and generation models based on the question [18]. Bazi and Yakoub's method employs an encoder-decoder transformer architecture, integrating image and text features for autoregressive answer generation [10]. MedfuseNet combines CNN and BERT embeddings with an MFB algorithm for feature fusion [19]. Zhou, Yuan, and Mei use a joint encoder for image and text embeddings without fusion, fine-tuning on the VQA dataset [20]. Van, Tom, and Derakhshani employ a pretrained language encoder and CLIP visual tokens for efficient training [21].

However, all these are limited to radiology, pathology, and orthopedics datasets but not dermatology. In the case of dermatology, we have simple classification tasks which do not concern with answer generation as a free-form text. We also found one task in which the authors[8] explored the performance of GPT-4v in differentiating between benign lesions and melanoma. The dataset provided by the organizers, however, includes a much larger domain problem set along with the difficulty of generating free-form text. Thus, this is a first-of-its-kind task.

3. Dataset Description

The MEDIQA-M3G task organizers provided their own dataset[9] to the participants, which we had to fetch by using the Reddit developer's API, as the dataset was part of a subreddit involving dermatologists answering the questions asked on it. The exact count of the dataset was not fixed and varied depending upon the time of fetching the dataset. The dataset was available in the English language. Each query in the dataset contained four main things:

1. Encounter_id: Query_id is used to score the results.
2. Image_ids: Containing a list of image IDs uploaded by the user asking the question.
3. query_title_en: As the name suggests, it was the query title in the English language
4. query_content_en: The query has some content, which can provide some extra context
5. responses: Three medical professionals/annotators answered the queries. Their responses, along with the annotator ID, are provided here.

Table 1
MEDIQA-M3G dataset details

Dataset	Expected Examples	Fetched Examples	Examples with images and non-deleted queries
Training	435	347	285
Validation	50	50	44
Test	100	93	78

The dataset was very noisy as the queries asked were on a subreddit; there was no particular format to ask questions. Some contained emojis; some had non-relevant information like *"I don't know how to upload more than one image in Reddit"*. Relevant statistics of the dataset received are shown in Table 1. Even after fetching the data, some queries were deleted, and some image URLs didn't exist, so the effective count was reduced to Table 1. The count mentioned is only for the English samples.

4. Data-Preprocessing

The dataset[9], as depicted in Table 1, was very small, and even if we trained the model by combining the training and validation data, the dataset was not sufficient. The dataset [9] was also quite noisy as it included emojis in some query titles, unwanted text statements like *"I don't know how to upload two images in Reddit"* which is not medically relevant and thus needed data preprocessing to extract medically relevant information before passing it to the model. This section includes the data-preprocessing steps we used to process the data before passing it to the main model.

4.1. Data Augmentation

We needed to augment the data by generating different titles and content as part of query content for the same image. We tried a few methods, such as word synonyms and translation methods, wherein we translated to another language and back-translated to English, but none were satisfactory. Because of this, we used the Textgenie repository to augment the data. Textgenie is a github repository for text data augmentations that facilitates the augmentation of text datasets and the creation of comparable samples. Moreover, it manages labeled datasets by retaining their labels in memory while generating analogous samples. The utilization of diverse Natural Language Processing techniques, including paraphrase generation, BERT mask filling, and converting passive voice constructions to active voice, is integral to its functionality. Presently, it is available only in the English language. Using this, we obtained at least three augmented titles and contents if they exist as per the condition stated in 4.2 for each query. For validation, we augmented the training queries only, but for the final test submission, we augmented the training and validation data and combined both of them.

4.2. Question pre-processing

The objective of this step is to extract medically relevant information from the title and content so that it can be passed to the main model for proper learning. Each query contained a query title (Q_t) and a query content (Q_c). We first concatenated Q_t and Q_c to form $\mathbf{q}$, the only condition being that the query title was not deleted *(displayed as [deleted by user])*, and the query content was neither empty

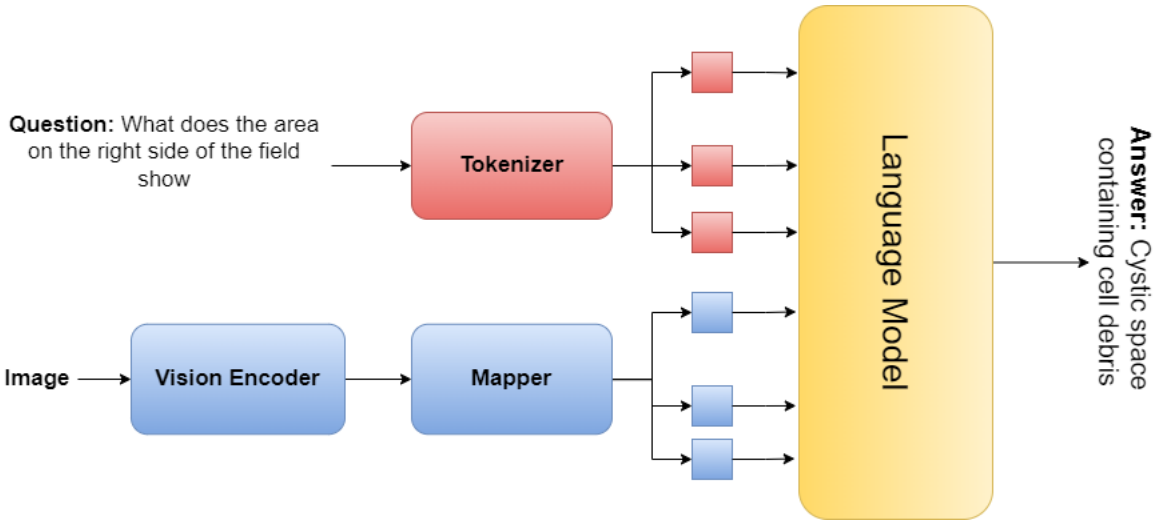


Figure 1: Model Architecture for Generation based Approach

nor deleted *(displayed as an empty string or [deleted] or [removed])*. Then **q** is passed through an emoji remover function to remove any emojis from it as it will not be of any medical relevance to give us **Q**.

5. Methodology

This section explains the 2 approaches and their model architectures we tried:- Generative and Classification based approach. The classification based approach is the top run submitted in the task [1].

5.1. Generation-based Model

We also tried a Generative-based Model and obtained results for the same. We fine-tune GPT2-xl[22] to accept questions and image information as a prompt. This model was an implementation of [21]. The text is encoded using GPT2-xl's [22] encoder, and the image is encoded by using the CLIP's [23] vision encoder. The model architecture is described in 1.

5.2. Data-Preprocessing for Classification Model

Due to the size and extreme-noisy nature of the dataset[9] we moved to the Classification approach but it required manually labeling the training and validation dataset and making a label-answer mapping.

5.2.1. Manual Labelling

For the classification approach we need to have classification labels so the model could be trained. For this, we did two things:-

1. Classify responses to Labels:- Manually classify the responses of training and validation to answer labels. Each query response can have multiple labels. Collectively, we form a set of 160 labels.
2. Make a $label \rightarrow Descriptive Answer mapping$:- For each of the 160 unique labels, we form a descriptive answer with the help of the responses of the train and valid queries and chatGPT[24].

This manual labeling helps us reduce the task complexity by making it a classification task. However, it will fails to answer some labels when the test set expects a response the model has never seen. The additional details of the labels is given in Appendix A.

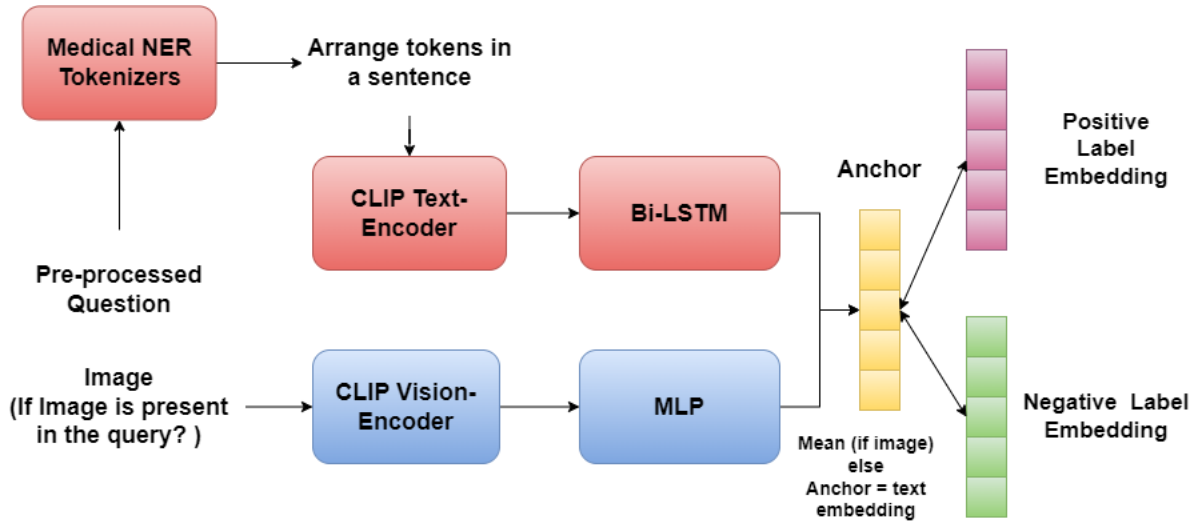


Figure 2: Model Architecture for Classification-based Approach

5.2.2. Preparing Answer Labels for Triplet Loss

The queries provided to us sometimes had more than one positive label from the set of 160 labels. We prepare the data such that each query has only one positive label and one negative label randomly selected from the list of remaining labels. Because each original query can have at least three augmented queries, we assign a single positive label to each one so that they get evenly distributed one at a time. For example, Suppose we have four augmented queries of the original query and three answer labels L_1, L_2, L_3 corresponding to it. In that case, we will assign L_1 to 2 queries, L_2 to another two queries, and L_3 to the remaining query, along with a randomly selected negative label for each. This is how we assign labels to each query and their augmentations.

5.3. Classification-based Model

The model converts the questions obtained in 4.2 to medically relevant NER tags. These tags are passed along with the image as a sentence through the CLIP encoders to obtain embeddings in the same latent space. The embeddings obtained are passed through separate Multi-Layer Perceptron (MLP) to obtain final embeddings. Separate MLPs are considered a way to make an ensemble model that benefits from both image and question as well as counter some training or validation examples that do not contain an image. The final embeddings are compared with the embeddings of the actual label (obtained by passing them through the CLIP[23] text-encoder model). The model (as shown in Fig 2) can be divided into five parts, each explained separately.

5.3.1. Question tokenization

When we obtained Q in Sec 4.2, it contains the actual text data entered by the user in the query, but it still contains some medically irrelevant information as specified in Sec 4. We thought of removing them manually, but that would not be justifiable for the task as it would create bias if I were a medical professional. So we extracted medical NER tags $\{t_1, t_2, ...t_{l_t}\}$ from these queries with the help of pre-trained Medical-NER models [25] and Clinical-AI-Apollo/Medical-NER model on hugging face. For the first NER tokenizer, we picked the tokens that belonged to ['Disease_disorder', 'Sign_symptom', 'Biological_structure', 'Coreference', 'Detailed_description', 'Color', 'Medication', 'Therapeutic_procedure', 'Shape'] token categories. For the second NER tokenizer, we picked the tokens belonging to ['DISEASE_DISORDER', 'BIOLOGICAL_STRUCTURE', 'SIGN_SYMPTOM', 'DETAILED_DESCRIPTION', 'MEDICATION'] token categories. After obtaining the NER tags from both the tokenizers, we removed

the duplicate tags and any stop-words if the tokenizers had picked them up. This list of NER tags forms the actual question tags, and this is done separately for each augmented data example obtained in 4.1. The tags provide essential medical information such as symptoms, description, color, shape, etc. Some training examples did not have any medical NER tags belonging to any token categories, so we took all the words of the query as the tokens for that example.

5.3.2. CLIP Encoders

The question tags obtained in 5.3.1 are combined to form a sentence with space as a delimiter; this sentence and the image corresponding to the query are passed separately from the pre-trained CLIP[23] model with ViT backbone to obtain embeddings for both of them separately. CLIP model with ViT backbone was chosen because it is a multimodal encoder model that can encode both text and images in the same latent space. As the embeddings belong to the same latent space, we need not pass it through an MLP to specifically convert them to another latent space. The answer labels are also passed through the CLIP encoder to obtain label embeddings, which are later used to calculate triplet loss.

5.3.3. Bi-LSTM and MLP layers

The text embedding obtained is passed through a Bi-LSTM layer, and the image embedding obtained is passed through an MLP layer. The text embeddings are obtained from a sentence with no semantic meaning; instead, it is the collection of medically relevant words. It is passed through a Bi-LSTM to make the embedding more comparable to the label embeddings when using triplet loss. We pass the image through an MLP to become comparable to the text embeddings obtained after passing through the Bi-LSTM.

5.3.4. Triplet Loss

We train the model by using Triplet loss[26] through cosine similarity. We obtain the final query embedding by taking the average of the text and the image embeddings obtained from Bi-LSTM and MLP, respectively. If the example does not have an image associated with it, we take the text embedding as the final query embedding. This query embedding will work as an Anchor. To obtain positive and negative embedding, we convert the positive label and negative label, as mentioned in 5.2.2, to label embeddings by passing them through a CLIP [23] encoder. The anchor, the positive, and the negative embeddings are then passed through the triplet loss function that calculates the triplet loss by taking into account the cosine similarity between the pair of embeddings. We also multiply the cosine similarity obtained by the class weight. The class weight for class i is calculated as:

$$w_i = \frac{n}{k \cdot n_i} \tag{1}$$

Where:

- n is the total number of samples after data augmentation,
- k is the total number of classes,
- n_i is the number of samples in class i.

5.3.5. Answer Generation

To generate the answer in the validation and test phase, we first obtain the query embeddings as explained in 5.3.4. The query embedding is used to calculate cosine similarity with each of the 160 label embeddings. The label with the highest cosine similarity is chosen as the answer label. The final descriptive answer is obtained through the **label → Descriptive Answer mapping** as explained in 5.2.1.

This is the model design for our top-performing run in the shared task. The second-best-performing model did not have any Bi-LSTM or MLP layer as mentioned in 5.3.3 and directly calculated the answer through 5.3.5.

6. Experiments and Results

This section contains information about the training setup, the experiments run, and the results obtained.

6.1. Training setup

We used the PyTorch framework and a pre-trained CLIP model [23] with ViT backbone as our text and image encoder. It gives us embedding of the size 512. Because we obtain the label encodings through CLIP model [23] with ViT backbone, they are also of size 512. The selection of hyperparameters was based initially on dataset analyses and later adjusted according to empirical observations. We opted for the Adams optimization algorithm for the training phase. The training began with a linear warm-up of over 500 steps, followed by a learning rate of 1e-4. All other Adam-optimizer settings were kept to a default. We have set the maximum query tag limit to 20 for the training, validation, and test phases. The model was trained using a batch size of 1, with gradient accumulation across up to 5 iterations because, after that, the model was overfitting. The following section details the test results as provided by the task organizers, providing insights into their effectiveness and applicability.

6.2. Results and Analysis

The results, as provided by the task organizers, are given in Table 2,3. For our top run, we were ranked second according to the Delta-Bleu score [27]. The 3rd ranked run was also ours, which was a simple CLIP encoder that gave embeddings to be compared with the label embeddings. The top-ranked run had an 8.6293 delta-bleu score. It fine-tuned a 1.86 B parameter Vision-Language model MoonDream2. Due to resource limitations, we could not fine-tune any large LLM. We tried the generative model approach as mentioned in 5, and its results are mentioned in Table 3.

Table 2
Test Run Scores for Classification Based Approaches

Type of Model	Features Used	Delta-BLEU Score	BERT Score	Test-Run Rank
Classification Based Model	Bi-LSTM, MLP+Triplet-Loss + Data Augmentation	4.829	0.839	2
	Only CLIP Model	4.819	0.838	3
	Bi-LSTM, MLP + Triplet-Loss	4.231	0.838	8

Table 3
Test Run Scores for Generation Based Approaches

Type of Model	Features Used	Delta-BLEU Score	BERT Score	Test-Run Rank
Generation Based Model	With Data Augmentation	2.525	0.829	12
	Without Data Augmentation	1.683	0.840	16

Analyzing results from Table 2,3 provides valuable scientific observations about the performance of our model on the dataset provided. The classification-based models performed better than the generative ones. In the classification-based approach, the model with Bi-LSTM layer and MLP performed slightly better than the pre-trained CLIP model with ViT backbone; this may suggest that the text embeddings obtained as a sentence embedding of the tokenized words may not correctly represent all the tokens due to the sentence not forming any semantic meaning. It may also suggest that the Bi-LSTM and MLP help the model learn text and image embeddings better to compare the labels.

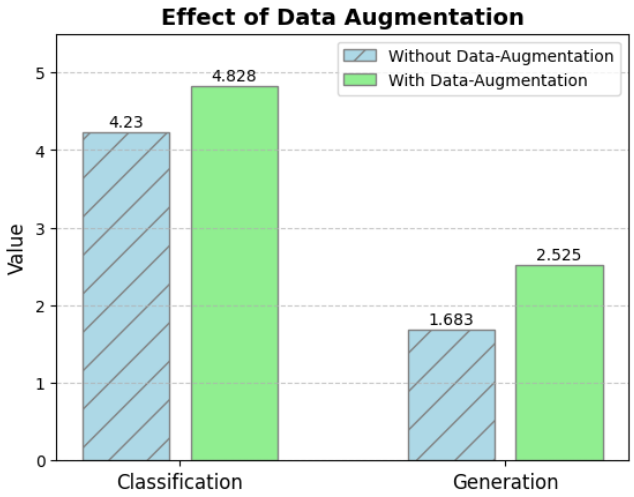
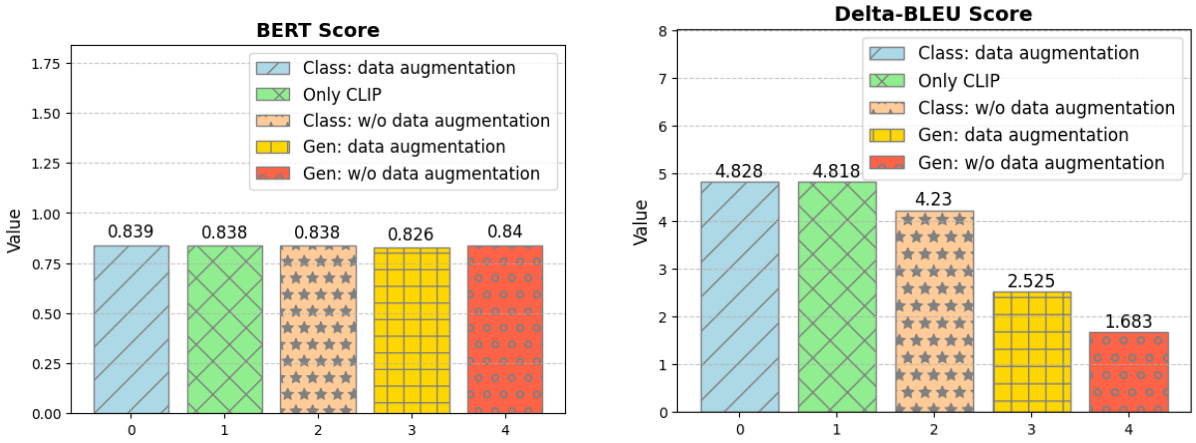


Figure 3: Effect of Data Augmentation



(a) BERT Score

(b) BLEU Score

Figure 4: Comparison of Scores

The results in Table 2, 3 also suggest that data augmentation is helping as the run without data augmentation in the classification-based model achieved a delta-bleu score of 4.231. In contrast, with data augmentation, it achieved a score of 4.829. As we can see in Fig 3, data augmentation also helped in the case of the generative approach as the delta-bleu score increased from 1.684 to 2.525.

7. Conclusion and Future works

The paper described our participation in ImageCLEF-2024-MEDIQA-MAGIC, which resorted to classification-based Med-VQA. We began with the generation-based model, but initial results were not promising due to the extremely noisy nature of the training data, and because of this, we quickly shifted to a classification-based approach. The classification-based approach worked in this case but reached its limit as we tried several models, and the score did not increase further. This was mainly due to a high number of answer classes (160), which will only increase further as we expand our answer domain and cause computational overhead. Thus, generative Med-VQA is the way to go, as the domain of answers does not limit it. In future works, we can fine-tune pre-trained multimodal generative models with the help of compute-efficient techniques and explore the feasibility of pre-training the model with other vibrant dermatology datasets through mask training.

References

[1] W. Yim, A. Ben Abacha, Y. Fu, Z. Sun, M. Yetisgen, F. Xia, Overview of the mediqa-magic task at imageclef 2024: Multimodal and generative telemedicine in dermatology, in: CLEF 2024 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Grenoble, France, 2024.

[2] K. Pogorzelska, L. Marcinowicz, S. Chlabicz, Understanding satisfaction and dissatisfaction of patients with telemedicine during the covid-19 pandemic: An exploratory qualitative study in primary care, Plos one 18 (2023) e0293089.

[3] D. Giansanti, Advancing dermatological care: A comprehensive narrative review of tele-dermatology and mhealth for bridging gaps and expanding opportunities beyond the covid-19 pandemic, in: Healthcare, volume 11, MDPI, 2023, p. 1911.

[4] B. Liu, L.-M. Zhan, L. Xu, L. Ma, Y. Yang, X.-M. Wu, Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering, in: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), IEEE, 2021, pp. 1650–1654.

[5] X. He, Y. Zhang, L. Mou, E. Xing, P. Xie, Pathvqa: 30000+ questions for medical visual question answering, arXiv preprint arXiv:2003.10286 (2020).

[6] A. Ben Abacha, S. A. Hasan, V. V. Datla, J. Liu, D. Demner-Fushman, H. Müller, Vqa-med: Overview of the medical visual question answering task at imageclef 2019, in: Working Notes of CLEF 2019, volume 2380 of *CEUR Workshop Proceedings*, CEUR-WS.org, Lugano, Switzerland, 2019. URL: https://ceur-ws.org/Vol-2380/paper_272.pdf.

[7] Z. Li, K. C. Koban, T. L. Schenck, R. E. Giunta, Q. Li, Y. Sun, Artificial intelligence in dermatology image analysis: current developments and future trends, Journal of clinical medicine 11 (2022) 6826.

[8] K. Cirone, M. Akrout, L. Abid, A. Oakley, Assessing the utility of multimodal large language models (gpt-4 vision and large language and vision assistant) in identifying melanoma across different skin tones, JMIR dermatology 7 (2024) e55508.

[9] W. Yim, Y. Fu, Z. Sun, A. Ben Abacha, M. Yetisgen, F. Xia, Dermavqa: A multilingual visual question answering dataset for dermatology, CoRR (2024).

[10] Y. Bazi, M. M. A. Rahhal, L. Bashmal, M. Zuair, Vision–language model for visual question answering in medical imagery, Bioengineering 10 (2023) 380.

[11] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[12] D. Gupta, S. Suman, A. Ekbal, Hierarchical deep multi-modal network for medical visual question answering, Expert Systems with Applications 164 (2021) 113993.

[13] Y. Khare, V. Bagal, M. Mathew, A. Devi, U. D. Priyakumar, C. Jawahar, Mmbert: Multimodal bert pretraining for improved medical vqa, in: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), IEEE, 2021, pp. 1033–1036.

[14] V. Joshi, P. Mitra, S. Bose, Multi-modal multi-head self-attention for medical vqa, Multimedia Tools and Applications (2023) 1–24.

[15] F. Cong, S. Xu, L. Guo, Y. Tian, Caption-aware medical vqa via semantic focusing and progressive cross-modality comprehension, in: Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 3569–3577.

[16] X. Hu, L. Gu, K. Kobayashi, Q. An, Q. Chen, Z. Lu, C. Su, T. Harada, Y. Zhu, Interpretable medical image visual question answering via multi-modal relationship graph learning, arXiv preprint arXiv:2302.09636 (2023).

[17] Y. Liu, Z. Wang, D. Xu, L. Zhou, Q2atransformer: Improving medical vqa via an answer querying decoder, in: International Conference on Information Processing in Medical Imaging, Springer, 2023, pp. 445–456.

[18] F. Ren, Y. Zhou, Cgmvqa: A new classification and generative model for medical visual question answering, IEEE Access 8 (2020) 50626–50636.

[19] D. Sharma, S. Purushotham, C. K. Reddy, Medfusenet: An attention-based multimodal deep learning model for visual question answering in the medical domain, Scientific Reports 11 (2021)

19826.
[20] Y. Zhou, J. Mei, Y. Yu, T. Syeda-Mahmood, Medical visual question answering using joint self-supervised learning, arXiv preprint arXiv:2302.13069 (2023).
[21] T. Van Sonsbeek, M. M. Derakhshani, I. Najdenkoska, C. G. Snoek, M. Worring, Open-ended medical visual question answering through prefix tuning of language models, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2023, pp. 726–736.
[22] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., Language models are unsupervised multitask learners, OpenAI blog 1 (2019) 9.
[23] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International conference on machine learning, PMLR, 2021, pp. 8748–8763.
[24] OpenAI, Chatgpt: Conversational agent, https://www.openai.com/chatgpt, 2024. Accessed: 2024-05-17.
[25] S. Raza, D. J. Reji, F. Shajan, S. R. Bashir, Large-scale application of named entity recognition to biomedicine and epidemiology, PLOS Digital Health 1 (2022) e0000152.
[26] K. Q. Weinberger, L. K. Saul, Distance metric learning for large margin nearest neighbor classification., Journal of machine learning research 10 (2009).
[27] M. Galley, C. Brockett, A. Sordoni, Y. Ji, M. Auli, C. Quirk, M. Mitchell, J. Gao, B. Dolan, deltableu: A discriminative metric for generation tasks with intrinsically diverse targets, arXiv preprint arXiv:1506.06863 (2015).

A. Manual Labelling

This section contains the list of all the labels we decided manually based on the training and validation answers. There are a total of 160 labels that we found manually by ourselves without any professional help or prior experience. Some labels were easy to identify as they were straight forward mentioned in the answer but some were difficult because the answer text contained multiple possibilities due to lack of information due to single image and less query content. In the case where there was not a single label and multiple possibilities we either had multiple labels or just a single label asking them to refer to a dermatologist as suggested by the annotator in the answer texts provided. The method fails when test set contains a label which was not a part of the manually picked labels.

The complete list of labels is as mentioned below:- cyst, blind pimple, pimple, folliculitis, Solar lentigo, contact eczema, eczema, common wart, sun exposure, coarse wrinkle, eczema due to dry skin, lip dryness, itchy scalp, keratoacanthoma, Pityriasis versicolor, rosacea, tan, callus around heel, leukoderma, hormonal acne, acne, fungal infection, ringworm, pityriasis rosea, ingrown hair, mole, skin cancer, cherry angioma, fungal infection due to nail thickening, pupuric spot, solar keratosis, keratosis, seborrheic keratosis, scratching, itching, birthmark, nevus, nodular melanoma, melanoma, urticaria, bug bite, insect bite, dyshidrotic eczema, contact dermatitis, heat rash, lipoma, cat and dog fleas, angiofibroma, ecchymosis, HTD (Habit-tic deformity), dermatologist for lesion examination, dermatologist consultation, alopecia areata, nevus sebaceous, spider veins, photodermatoses, lymphatic malformation, comedones, healing, milia, xanthelasma, chalazia, hyperpigmentation, longitudinal melanonychia, Pyogenic granuloma, post inflammatory hyperpigmentation, dermatitis artifacts, dermatitis, atrophoderma, athlete's foot, pseudofolliculitis, subungual hematoma, Neutrophilic dermatoses, Discoid eczema, atopic dermatitis, acneiform eruptions, keratosis pilaris, tinea versicolor, dermatofibroma, viral rash, angular cheilitis, flushing skin because of alcohol, compund nevus, rash, morphea, inflammatory rash, shingles, dandruff, psoriasis, trauma, lip licker's dermatitis, aquagenic wrinkle, cystic fibrosis, eclipse nevi, nummular eczema, eczema due to working in water, hairy tongue, syringomas, lip biting, irritated hair follicle, cyst under skin, spider angioma, inflammatory acne, Schamberg's purpuric dermatosis, vasculitis, sebaceous hyperplasia, tinea corporis, granuloma annular, viral infection, hive, mast cells in

Label	Frequency
eczema	50
mole	18
fungal infection	17
acne	17
dryness	17
cyst	14
dermatologist for lesion examination	12
skin cancer	12
dermatitis	12
atopic dermatitis	11

Table 4
Training labels details

Label	Frequency
eczema	8
insect bite	4
fungal infection	4
mole	3
dryness	3
post inflammatory hyperpigmentation	2
itching	2
psoriasis	2
bruise	2
bug bite	2

Table 5
Validation labels details

body, herpes, herpetic whitlow, comedonal acne, angioma, drug reaction, syphilis, infection on skin, trichostasis spinulosa, erosive pustular dermatosis, retention hyperkeratosis, inflammation of the nail fold, dermatitis herpetiformis, sebaceous cyst, observe, skin tag, nail trauma, dryness, molluscum, friction, blood collection, tick bites, irritated mole, Opthalmologist consultation, hand sweating, hidradenitis suppurativa, corn, cyst due to mucus, herpes simplex, seborrheic dermatitis, abscess, allergy due to sun, bruise, Keratolysis exfoliativa, idiopathic guttate hypomelanosis, infection at hair follicle, eye dark circles, erythema, diabetes, atrophic scars, periorificial dermatitis, normal, HIV, scar, skin peeling, telangiectasia, genetic predisposition, irritated skin, chicken pox, furuncle.

The tables 4 and 5 presents the most frequent labels encountered in training and validation data respectively.