Explainability for Medical Image Captioning

Djamila Romaissa Beddiar
University of Oulu, CMVS
Oulu, Finland
Djamila.Beddiar@oulu.fi

Mourad Oussalah
University of Oulu, CMVS
& MIPT, Faculty of Medicine
Oulu, Finland

Seppänen Tapio
University of Oulu, CMVS
Oulu, Finland

***Abstract*—Medical image captioning is the process of generating clinically significant descriptions to medical images, which has many applications among which medical report generation is the most frequent one. In general, automatic captioning of medical images is of great interest for medical experts since it offers assistance in diagnosis, disease treatment and automating the workflow of the health practitioners. Recently, many efforts have been put forward to obtain accurate descriptions but medical image captioning still provides weak and incorrect descriptions. To alleviate this issue, it is important to explain why the model produced a particular caption based on some specific features. This is performed through Artificial Intelligence Explainability (XAI), which aims to unfold the 'black-box' feature of deep-learning based models. We present in this paper an explainable module for medical image captioning that provides a sound interpretation of our attention-based encoder-decoder model by explaining the correspondence between visual features and semantic features. We exploit for that, self-attention to compute word importance of semantic features and visual attention to compute relevant regions of the image that correspond to each generated word of the caption in addition to visualization of visual features extracted at each layer of the Convolutional Neural Network (CNN) encoder. We finally evaluate our model using the ImageCLEF medical captioning dataset.**

***Index Terms*—Image Captioning, Medical images, Encoder-decoder, Attention-maps , Artificial Intelligence Explainability.**

I. Introduction

Image captioning is the process of describing the visual content of an image using natural language [1]. It involves two main fields: image processing to understand the visual content and natural language processing to generate meaningful and accurate descriptions. In general, image captioning is employed whenever textual description is required to be generated automatically from an image such as human like robot-robot interactions [2], visual question answering tasks [3] and medical report generation [4], [5]. Moreover, automatic captioning of medical images plays a vital role in patient care by promoting the workflow of the health practitioners, assisting them during diagnosis and disease treatment [6]. It is also essential in the development of computer-aided diagnosis systems [7].

Many research has been carried out so far, in the automatic image captioning field. However, the proposed systems still need to be improved, especially, in clinical setting where sentences, in addition to being grammatically correct, need to be clinically acceptable as well. One of the most frequent architectures that has been populated for this purpose is the generative networks [2], which relies on the use of encoder to extract visual features from the image and the decoder to generate the caption words. However, many existing techniques struggled in generating clinically correct sentences and even fail to create a new sentence that have never appeared in the training set [8]. To alleviate this challenge, it is important to scrutinize the functioning of the model to comprehend the feature engineering process and why the model reaches out a particular result. For that, explainability in artificial intelligence emerged to explain the predictions made by the models to give more useful cues about why these results were obtained.

Motivated by this, we extend our proposed model [9] for the ImageCLEFmedical 2021 [10], [11] by adding an explainability module. We present therefore, an attention-based encoder-decoder model for medical image captioning. Our model relies on two encoders: a CNN for visual features encoding, a Gated Recurrent Unit (GRU) with self-attention for semantic features encoding and, an attention-based Long Short-term Memory (LSTM) decoder for sentence generation. Therefore, we utilize both attention mechanisms for our explainability module by visualizing the attention maps and computing the word importance. These mechanisms exhibited the most significant regions and the considered word embeddings while extracting the features and generating new captions by the model. We evaluate the performance of our model on the ImageCLEF medical captioning dataset.

The rest of this paper is organized as follows. First, we briefly summarize some of the state-of-the-art on medical image captioning and explainability in Section. II. Section III outlines our approach. Then, we discuss the experimental results of our approach as well as the employed dataset in Section. IV. Finally, we conclude and provide some future directions for explainable medical image captioning in Section. V.

II. Related Work

Understanding the content of images is of great interest in many fields such as video surveillance, image retrieval, automated vehicle systems and health care. Extracting relevant information from image content and then mapping them onto

978-1-6654-6964-7/22/$31.00 ©2022 IEEE

textual descriptions characterize the task of image captioning [12]. Many research was devoted to this, especially when dealing with natural images where describing the relationship between actors and objects of the image is straightforward. However, understanding and captioning medical images is more challenging and require involvement of medical experts because the relationship between visual features and semantic descriptions is not trivial and cannot be identified easily [12]. We review related work in medical image captioning, focusing on deep learning based techniques, and explainability in medical artificial intelligence in the following subsections.

A. Medical Image Captioning

Automatic captioning of medical images aims to build a bridge between visual observations and descriptive caption in natural language [13]. This can be useful for automatically generating medical reports and take part in the implementation of computer aided diagnosis systems [7]. Motivated by the Show and Tell model proposed by [14] for natural image captioning, based on generative networks and inspired from machine translation models, authors in [15], [16] achieved satisfying results for medical image captioning. The Show and Tell model relies on a CNN encoder and an LSTM decoder, which is further improved to capture highly relevant features by identifying the most important regions in image using attention mechanisms as in [17]. The LSTM decoder has been substituted in some cases with a GRU or a Recurrent Neural Network (RNN) for sentences generation as in [15], [18]. For instance, [19] proposed a hierarchical LSTM model to distinguish between normal and abnormal sentences and generate them using dual LSTM model. Likewise, [20] implemented the decoder in a multi-stage hierarchical manner to translate medical image features into textual description. In addition, [12] proposed a coarse-to-fine encoder-decoder ensemble model for ultrasound image captioning by first identifying the clinical organ, then the disease and, finally describing the content of the image.

B. Explainable Image Captioning

Inspired by what Jason Yosinski declared in Uber states [21]: "We build amazing models. But we don't quite understand them", and the fact that machine learning models are practically "black-boxes", a new trend emerged ultimately. It is about explainable artificial intelligence (XAI), which aims at explaining the outcomes of Artificial Intelligence (AI) models, providing interesting cues that help understand the reasoning and augment transparency of the models. Many explanation methods were used in the literature for classification and regression tasks such as gradient-based methods [22], decomposition-based methods [23], and sampling-based methods [24]. In addition, many research has been conducted in this regard and concept was adopted to clarify why captioning system produced a specific description for a given image. For instance, [25] employed XAI to explain predictions of a captioning model by depicting a part of image corresponding to a particular word and showing why the model generated this word. Similarly, [26] added an explanatory layer to the state-of-the-art Show, "Attend and Tell" model by augmenting the attention mechanism using additional bottom-up features. The attention is computed on the joint embedding space formed between the high-level and the low-level features of object salient regions identified with bounding boxes. Moreover, [27] used Gradient-weighted Class Activation Mapping (Grad-CAM) and Guided Grad-CAM to explain the captioning model based on Layer-wise Relevance Backpropagation (LRP) and attention mechanisms beyond visualizing the attentions. They showed how the explanations correlate with object relations in the image and identified at the same time words that are not supported by the image content. In general, using object detection for natural image captioning could help in the explanation process by identifying the objects of the image and their positions but could not be efficient in case of medical images where objects are not easy to identify and clinical findings is much more important than relationships between objects. So far, attention maps are most frequent explanatory cues in medical image captioning.

III. Methodology

We present in the this paper an explainable approach to generate accurate captions for medical images. Inspired by the *Show, Attend and Tell* [28], we propose an attention-based encoder-decoder model where two encoders are used to extract visual and semantic features separately. First, we start by pre-processing the data. The captions are pre-processed by tokenization, stop-word removal, lower-cased and stemming using the Natural Language Toolkit (NLTK) package [1] for the whole process. Then, two tokens: '$< start >$' and '$< end >$' are added to help the decoder identify the beginning and the end of each caption. Images are pre-processed by normalization, resizing and augmenting the training set with vertical and horizontal flipping and crop-centering.

Once data is prepared, we calculate word embeddings for the captions and extract visual features from images using our two encoders as illustrated by "Fig. 1". A self-attention GRU encoder is used for semantic feature extraction within a multi-label classification task, Whereas transfer learning is employed to extract visual features from images using a pre-trained Resnet50 model for which features correspond to convolutional layer activation instead of flatten layer activation. Weights of Resnet50 model pretrained on ImageNet are employed and all convolutional layers are considered. This helps us to construct relationships between image's parts and extracted features through a visual attention mechanism and self-attention is employed to highlight most important words in the caption. At each time step, the LSTM decoder exploit the previous hidden state, previous word and the calculated context vector (using the visual attention) to generate a new word of the new caption until the '$< end >$' token is met and, therefore, each newly generated word describes an important region of the image.

[1] https://www.nltk.org/

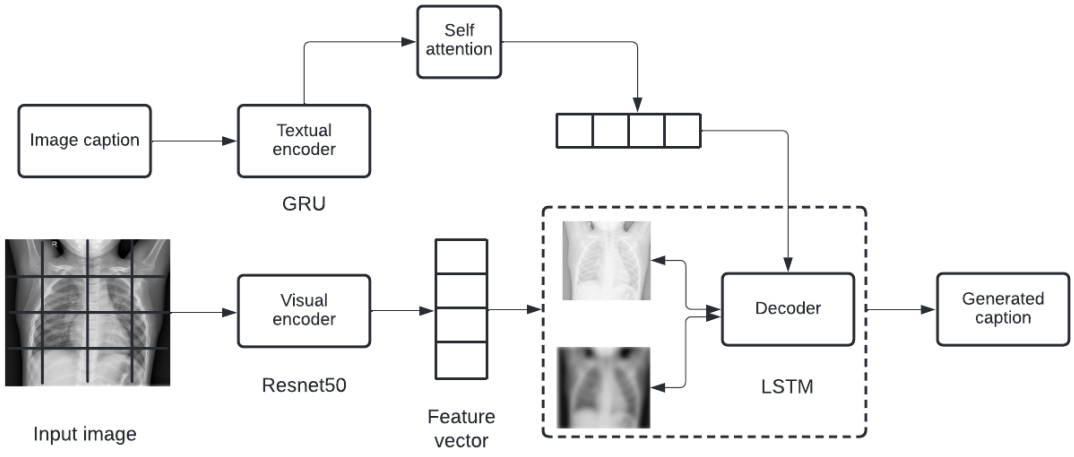


Fig. 1. General scheme of our encoder decoder model. Visual features are extracted using a Resnet50 encoder from the input image and textual features are extracted using a Self-attention GRU model and finally visual and textual features are fused and passed into an LSTM decoder through an attention mechanism.

Finally, the explanation module relies on the attention maps, feature visualization and word importance. Attention maps calculated through the attention-based encoder-decoder model are used to map relationships between image regions and generated words. We compute a weight matrix for input image regions and words of the original caption to illustrate the correspondence between words and regions. Word importance is used to highlight the most important words in the generated caption by leveraging the weights of the self-attention based multi-label classifier. These weights help us to visualize which word was mainly considered by the encoder when calculating the word embedding. We visualize as well features at different convolutional layers to show some of the considered visual features by the Resnet50 model.

IV. Results and Discussion

In this section, we provide our experimental results and describe our dataset, the evaluation metrics used for the model evaluation.

A. Dataset

We used for our medical image captioning model evaluation, the ImageCLEFmed 2021 dataset [10], [11], which includes three sets: the training set composed of 2756 medical images; the validation set and the test set consisting of 500 and 444 radiology images, respectively. For each medical image, the medical Concepts Unique Identifiers (CUIs) and caption consisting of one or more sentences are associated. The correspondences between the image ID and the CUIs and the image ID and the captions are stored into two excel files. This dataset is very challenging due to various image modalities it includes, in addition to different body parts captured and varying acquisition conditions.

B. Evaluation metrics

To evaluate the ability of our model in generating efficient captions, we calculate the BiLingual Evaluation Understudy (BLEU score) using the Python NLTK package. Each caption is assumed to be one single sentence.

BLEU is an automatic metric used for evaluation of machine-translation systems by measuring the similarity between the machine translation and a set of reference translations. It varies between zero and one where zero refers to no overlap between translation and reference (low quality), and one refers to perfect overlap between them (high quality). BLEU is one of the most useful metrics in image captioning allowing us to compute the similarity between the original caption and the newly generated caption. Mathematically, the BLEU score is defined as:

$$BLEU = min(1, \exp(1 - \frac{r}{c}))(\prod_{i=1}^{4} P_i)^{1/4} \quad (1)$$

Where $BP = min(1, \exp(1 - \frac{r}{c}))$ refers to the brevity penalty, r refers to the original caption length, c to the generated caption and P refers to the modified precision. The Brevity Penalty allows us to choose the candidate caption which is most likely similar in length, word choice and word order to the original caption.

Modified precision is computed using "(2)", where $m^i_{cand}, m^i_{ref}, w^i_t$ correspond to the count of i-gram in candidate matching the original caption, the count of i-gram in the original caption and total number of i-grams in candidate caption, respectively.

$$p = \frac{\sum_{C \in \{Candidates\}} \sum_{i \in C} min(m^i_{cand}, m^i_{ref})}{w^i_t = \sum_{C' \in \{Candidates\}} \sum_{i' \in C'} m^{'i}_{cand}} \quad (2)$$

C. Results

First, we report the results of our caption generation model in terms of BLEU score compared to some existing techniques

in Table. I. We achieved good results compared to the-state-of-the-art even though [16] obtained the best BLEU score.

TABLE I
COMPARISON OF OUR RESULTS TO SOME STATE OF THE ART RESULTS

Method	*BLEU score*
ImageSem [29]	25.7%
"Show, Attend and Tell" [16]	43.7%
Attention-based encoder-decoder [9]	28.7%
Our proposal	40.29%

Afterwards, we visualize in "Fig. 2" some samples of visual features extracted from an input image at different layers using the pre-trained Resnet50 encoder. As mentioned before, we calculate the outputs of convolutional layers which distinguish global features and when going deeper, we obtain more detailed features.

Next, using the attention scores of the GRU classifier, we visualize in "Fig. 3", word importance of some image captions. The higher score, characterized by darker colour, refers to a high degree of relevance of the word in the calculation of semantic features of the image.

Finally, to illustrate the most relevant regions of the image, that have been considered by the decoder while generating the caption, we visualize some attention maps in "Fig. 4" and "Fig. 5" for correctly and wrongly generated captions. We highlight the words existing in both original and generated caption with red color. We can see that the model was able to generate exactly the same caption by focusing on some parts of the image only. The decoder was able to generate a word for each yellow region using the semantic and the visual feature extracted using both encoders. However, the decoder failed to generate some captions due to some pre-processing of the data or presence of unknown words.

V. CONCLUSION

We presented in the current paper an attention-based encoder-decoder inspired from the *Show, Attend and Tell* model for medical image caption generation with an explainability module. We fused two encoders to extract semantic and visual features separately. For semantic features extraction, we used a self-attention based GRU multi-label classifier whereas visual features are extracted using a pre-trained Resnet50 model. We employed visual attention mechanism to combine the visual encoder with the decoder allowing the decoder to focus on most relevant regions of the input image. For the explainability module, we exploited first the word importance derived from the self-attention scores to highlight the most efficient words considered while constructing the caption. Then, we visualized visual features which correspond to convolutional layers activation of the Resnet50 model. Finally, we employed the attention maps to illustrate the most important regions of the image exploited by the decoder when generating captions. We evaluated the performance of our model in terms of BLEU score and showed some visualizations of correctly and wrongly generated captions for the ImageCLEF dataset.

ACKNOWLEDGMENT

This work is supported by the Academy of Finland Profi5 DigiHealth project (#326291) and the European Young-sters Resilience through Serious Games, under the Internal Security Fund-Police action: 823701-ISFP-2017-AG-RAD grant, which are gratefully acknowledged.

REFERENCES

[1] M. Stefanini, M. Cornia, L. Baraldi, S. Cascianelli, G. Fiameni, and R. Cucchiara, "From show to tell: A survey on image captioning," *arXiv preprint arXiv:2107.06912*, 2021.

[2] H. Ayesha, S. Iqbal, M. Tariq, M. Abrar, M. Sanaullah, I. Abbas, A. Rehman, M. F. K. Niazi, and S. Hussain, "Automatic medical image interpretation: State of the art and future directions," *Pattern Recognition*, p. 107856, 2021.

[3] H. Sharma and A. S. Jalal, "Image captioning improved visual question answering," *Multimedia Tools and Applications*, pp. 1–22, 2021.

[4] B. Jing, P. Xie, and E. Xing, "On the automatic generation of medical imaging reports," *arXiv preprint arXiv:1711.08195*, 2017.

[5] X. Yang, M. Ye, Q. You, and F. Ma, "Writing by memorizing: Hierarchical retrieval-based medical report generation," *arXiv preprint arXiv:2106.06471*, 2021.

[6] X. Huang, F. Yan, W. Xu, and M. Li, "Multi-attention and incorporating background information model for chest x-ray image report generation," *IEEE Access*, vol. 7, pp. 154 808–154 817, 2019.

[7] X. Zeng, L. Wen, Y. Xu, and C. Ji, "Generating diagnostic report for medical image by high-middle-level visual information incorporation on double deep learning models," *Computer Methods and Programs in Biomedicine*, vol. 197, 2020.

[8] Y. Xue, T. Xu, L. Rodney Long, Z. Xue, S. Antani, G. Thoma, X. Huang, S. J.A., D. C., A.-L. C., F. G., and F. A. and, "Multimodal recurrent model with attention for automated radiology report generation," *21st International Conference on Medical Image Computing and Computer Assisted Intervention, MICCAI 2018*, vol. 11070, pp. 457–466, 2018.

[9] D. Beddiar, M. Oussalah, and T. Seppänen, "Attention-based CNN-GRU model for automatic medical images captioning : ImageCLEF 2021," in *Working Notes of CLEF - Conference and Labs of the Evaluation Forum,CLEF-WN 2021, 21-24 September, Bucharest, Romania*, 2021, pp. 1160–1173.

[10] O. Pelka, A. Abacha, A. De Herrera, J. Jacutprakart, C. Friedrich, and H. Müller, "Overview of the ImageCLEFmed 2021 Concept & Caption Prediction Task," in *CLEF2021 Working Notes*, ser. CEUR Workshop Proceedings. Bucharest, Romania: CEUR-WS.org, September 21-24 2021.

[11] B. Ionescu, H. Müller, R. Péteri, A. Ben-Abacha, M. Sarrouti, D. Demner-Fushman, S. Hasan, S. Kozlovski, V. Liauchuk, Y. Dicente, V. Kovalev, O. Pelka, A. Seco de Herrera, J. Jacutprakart, C. Friedrich, R. Berari, A. Tauteanu, D. Fichou, B. Brie, M. Dogariu, L. Ştefan, M. Constantin, J. Chamberlain, A. Campello, A. Clark, T. Oliver, H. Moustahfid, A. Popescu, and J. Deshayes-Chossart, "Overview of the ImageCLEF 2021: Multimedia retrieval in medical, nature, internet and social media applications," in *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, ser. Proceedings of the 12th International Conference of the CLEF Association (CLEF 2021). Bucharest, Romania: LNCS Lecture Notes in Computer Science, Springer, September 21-24 2021.

[12] X.-H. Zeng, B.-G. Liu, and M. Zhou, "Understanding and generating ultrasound image description," *Journal of Computer Science and Technology*, vol. 33, no. 5, pp. 1086–1100, 2018.

[13] Y. Xiong, B. Du, P. Yan, S. H., L. M., L. C., and Y. P., "Reinforced transformer for medical image captioning," *10th International Workshop on Machine Learning in Medical Imaging, MLMI 2019 held in conjunction with the 22nd International Conference on Medical Image Computing and Computer-Assisted Intervention, MICCAI 2019*, vol. 11861, pp. 673–680, 2019.

[14] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3156–3164.

[15] O. Pelka, C. Friedrich, M. T., F. N., G. L., and C. L. and, "Keyword generation for biomedical image retrieval with recurrent neural networks," *18th Working Notes of CLEF Conference and Labs of the Evaluation Forum, CLEF 2017*, vol. 1866, 2017.

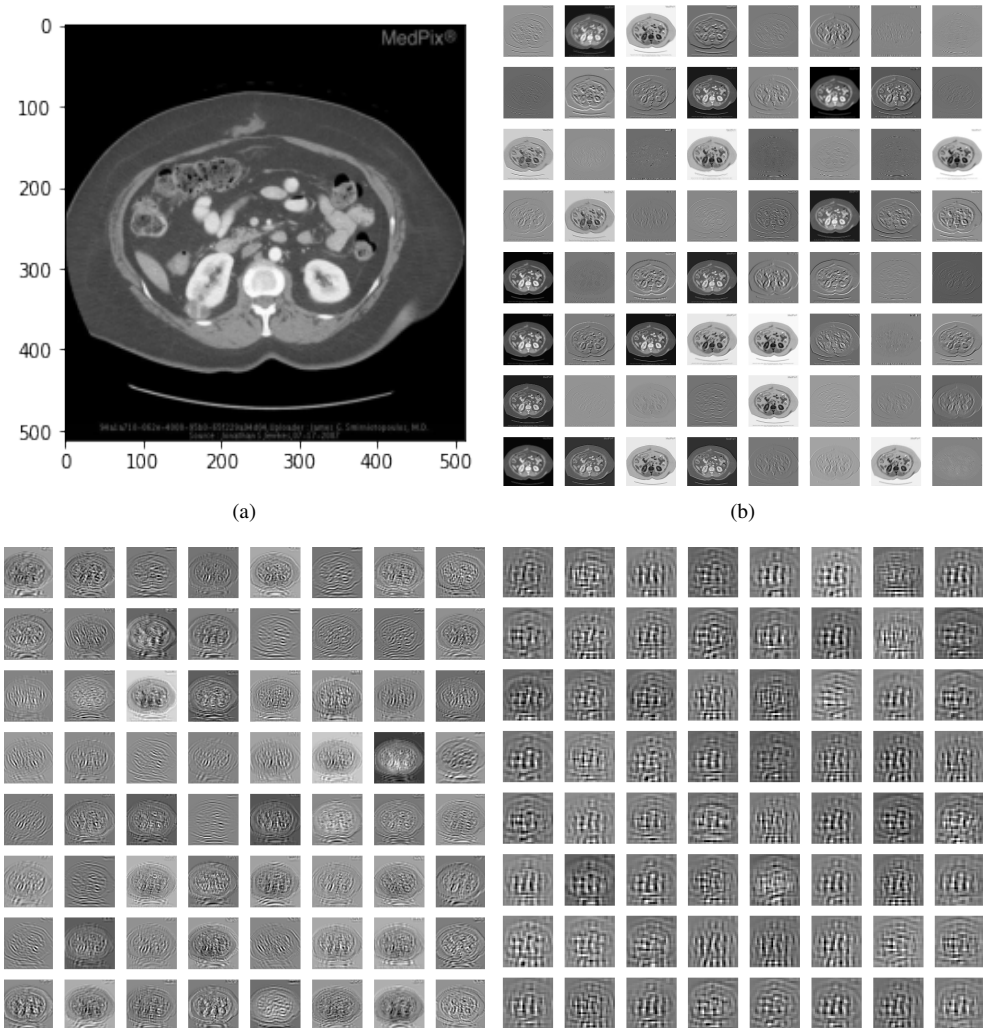


Fig. 2. Visualization of features extracted from the ResNet50 encoder at different layers for one sample from the ImageCLEF medical captioning dataset. a) original image, b) visual features at layer 0, c) visual features at layer 20, d) visual features at layer 40.

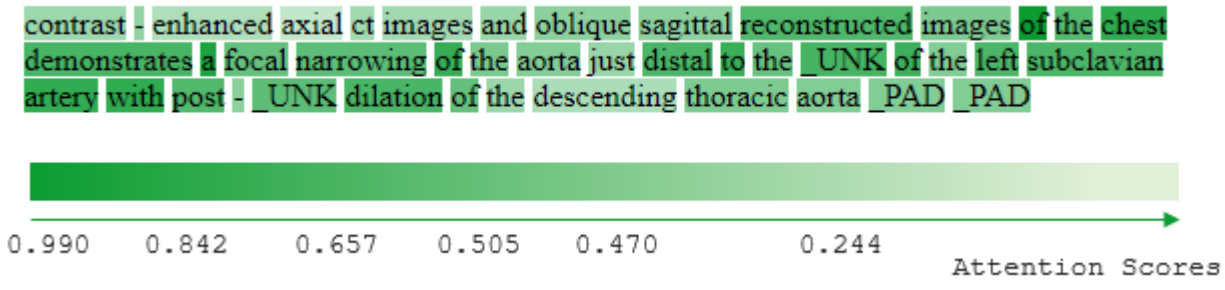
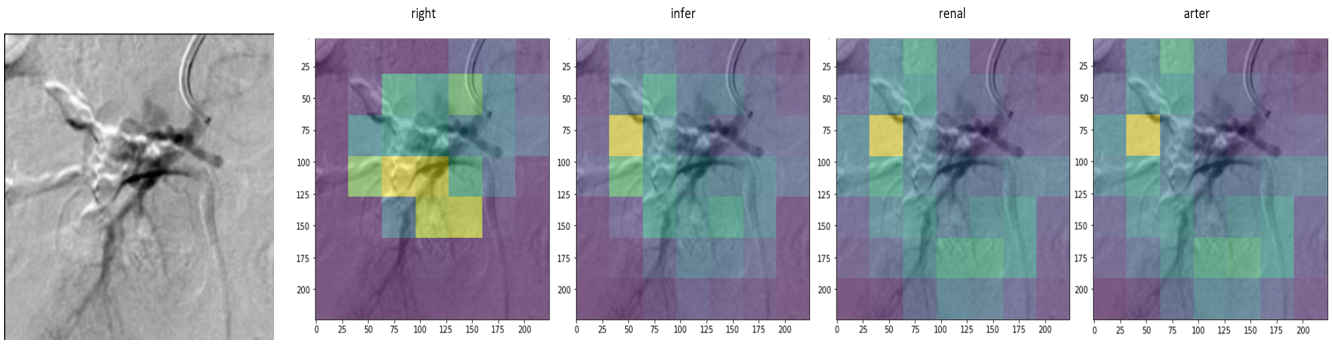
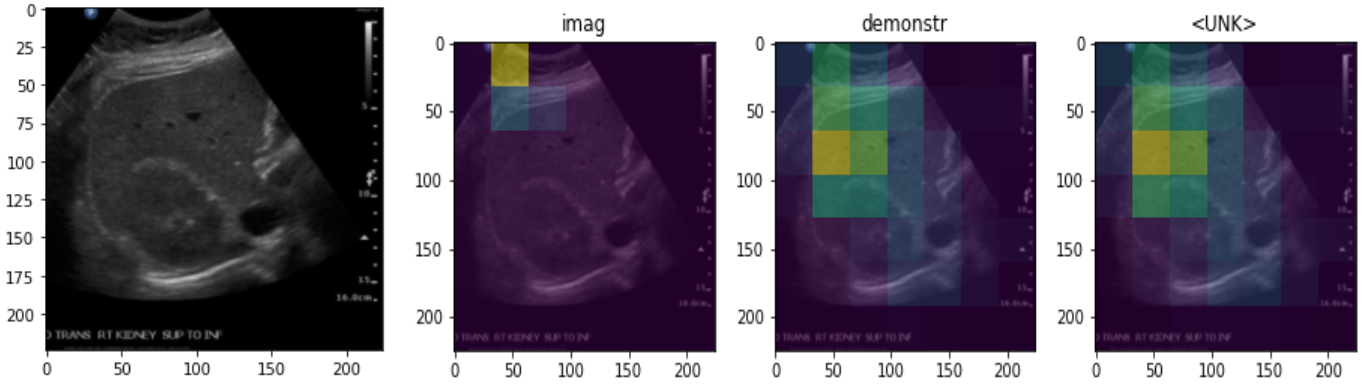


Fig. 3. Word importance using images captions and self-attention scores.



(a) (b)

Fig. 4. Visualization of attention maps of a correctly generated caption. a) original image, b) attention maps. Original caption: Right Inferior Renal Artery; Generated caption: right infer renal arter.



(a) (b)

Fig. 5. Visualization of attention maps of a wrongly generated caption, the model was not able to generate a caption. a) original image, b) attention maps.

[16] R. Tsuneda, T. Asakawa, and M. Aono, "Kdelab at imageclef 2021: Medical caption prediction with effective data pre-processing and deep learning," in *CLEF2021 Working Notes, CEUR Workshop Proceedings, CEUR-WS. org, Bucharest, Romania*, 2021.

[17] Y. Xue, T. Xu, L. Rodney Long, Z. Xue, S. Antani, G. R. Thoma, and X. Huang, "Multimodal recurrent model with attention for automated radiology report generation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 457–466.

[18] D. Lyndon, A. Kumar, J. Kim, M. T., F. N., G. L., and C. L. and, "Neural captioning for the imageclef 2017 medical image challenges," *18th Working Notes of CLEF Conference and Labs of the Evaluation Forum, CLEF 2017*, vol. 1866, 2017.

[19] P. Harzig, Y.-Y. Chen, F. Chen, R. Lienhart, and A. A. et al.; facebook; Intel; Microsoft, "Addressing data bias problems for chest x-ray image report generation," *30th British Machine Vision Conference, BMVC 2019*, 2020.

[20] S. Singh, S. Karimi, K. Ho-Shon, L. Hamey, S. Singh, S. Karimi, K. Ho-Shon, and L. Hamey, "From chest x-rays to radiology reports: A multimodal machine learning approach," *2019 Digital Image Computing: Techniques and Applications (DICTA)*, pp. 462–469, 2019.

[21] P. Voosen, "How ai detectives are craking open the black box of deep learning," *Science*, 2017.

[22] R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[23] W. Murdoch, P. Liu, and B. Yu, "Beyond word importance: Contextual decomposition to extract interactions from lstms," *arXiv preprint arXiv:1801.05453*, 2018.

[24] R. Luss, P. Y. Chen, A. Dhurandhar, P. Sattigeri, Y. Zhang, K. Shanmugam, and C. Tu, "Generating contrastive explanations with monotonic attribute functions," *arXiv preprint arXiv:1905.12698*, 2019.

[25] S. Sahay, N. Omare, and K. Shukla, "An approach to identify captioning keywords in an image using lime," in *2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*. IEEE, 2021, pp. 648–651.

[26] R. Biswas, M. Barz, and D. Sonntag, "Towards explanatory interactive image captioning using top-down and bottom-up features, beam search and re-ranking," *KI-Künstliche Intelligenz*, vol. 34, no. 4, pp. 571–584, 2020.

[27] J. Sun, S. Lapuschkin, W. Samek, and A. Binder, "Understanding image captioning models beyond visualizing attention," *arXiv preprint arXiv:2001.01037*, 2020.

[28] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*. PMLR, 2015, pp. 2048–2057.

[29] X. Wang, Z. Guo, C. Xu, L. Sun, and J. Li, "Imagesem group at imageclefmed caption 2021 task: exploring the clinical significance of the textual descriptions derived from medical images," in *CLEF2021 Working Notes, CEUR Workshop Proceedings, CEUR-WS. org, Bucharest, Romania*, 2021.