Attention-based CNN-GRU Model For Automatic Medical Images Captioning: ImageCLEF 2021

Djamila-Romaissa, Beddiar[1], Mourad, Oussalah[1,2] and Tapio, Seppänen[1]

[1]*Center for machine vision and signal analysis, University of Oulu, Finland*
[2]*MIPT, Faculty of Medicine, University of Oulu, Oulu, Finland*

Abstract

The action of understanding and interpretation of medical images is a very important task in the medical diagnosis generation. However, manual description of medical content is a major bottleneck in clinical diagnosis. Many research studies were devoted to develop automated alternatives to this process, which would have enormous impact in terms of efficiency, cost and accuracy in the clinical workflows. Different approaches and techniques have been presented in the literature ranging from traditional machine learning methods to deep learning based models. Inspired by the outperforming results of the later techniques, we present in the current paper, our team participation (RomiBed) to the ImageCLEF medical caption prediction task. We addressed the challenge of medical image captioning by combining a CNN encoder model with an attention-based GRU language generator model whereas a multi-label CNN classifier is used for the concept detection task. Using the provided data in the training, validation and test subsets, we obtain an average F_measure of 14.3% and a BLEU score of 0.243 on the ImageCLEF concept detection and the caption prediction challenges, respectively.

Keywords

Automatic image captioning, Medical images, Concept detection, Radiology, Multi-label Classification, Encoder-decoder, Attention Mechanism

1. Introduction

With the increasing number of medical images generated worldwide from different modalities in hospitals and health centers, the need to analyse and discover their content is crucial. Indeed, medical images offer a safe environment to explore patient's health state without the need for a surgery or any other invasive procedures [1]. Besides, this also helps clinicians in their daily routine by expediting clinical workflows and trigger automated alerts associated to potentially dangerous diseases. Recently, many research was devoted to the process of automatically generating clinically sound interpretations of medical images. Roughly speaking, generating clinically explainable and understandable analysis for medical images may enrich medical knowledge systems and facilitate the human-machine interactive diagnosis practice [2]. Therefore, automatic medical image captioning is one of the main focus of the interdisciplinary research in medical imaging field [2]. Especially, medical image captioning uses visual features of images to generate a concise textual description of the content of the medical image by

CLEF 2021 – Conference and Labs of the Evaluation Forum, September 21–24, 2021, Bucharest, Romania

Djamila.Beddiar@oulu.fi (D. Beddiar)
0000-0002-1371-3881 (D. Beddiar)

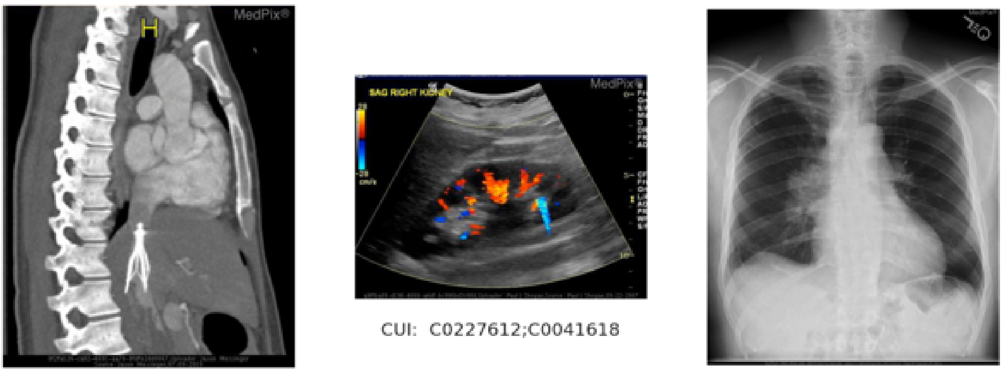
highlighting the clinically important observations. It represents a convergence of computer vision and natural language processing (NLP) with an emphasis on medical image processing [3]. In this regard, the ImageCLEFmedical task [4, 5] is organized each year as part of the CLEF initiative labs aiming at developing machine learning methods for medical image understanding and description. It includes two sub-tasks: the concept detection sub-task which aims to identify the UMLS Concept Unique Identifiers (CUIs) for a given medical image. Whereas the caption prediction sub-task aims to generate coherent caption based on the clinical concept vocabulary created in the first sub-task and the visual content of the image.

Motivated by the recent advances made in deep neural networks in different tasks of computer vision and NLP, especially due to their promising results in the machine language translation models, we present in this paper our contribution to the ImageCLEF 2021 medical task under the team name 'RomiBed'. We proposed a multi-label classification CNN model for the first sub-task after applying an augmentation technique based on the center cropping of the medical images. Features were extracted using a pre-trained model, while the classification is performed using a CNN network. For the second sub-task, we proposed an encoder-decoder model with an attention layer where the encoder is based on a CNN feature extractor and the decoder is composed of a GRU network with an attention mechanism.

This paper is organized as follows. First, we briefly review the related medical image captioning studies from the literature in Section. 2. In Section. 3, we provide a brief description of the ImageCLEF dataset used in this study. Next, we detail the methodology we followed to construct the concept detection model as well as the caption prediction model in Section. 4. We discuss each step of the process and deliver the results in terms of F_measure for the concept detection and BLEU for the caption prediction. Finally, we finish with a conclusion where we highlight some key insights and future directions.

2. Related Work

Automatic image captioning (AIC) in the medical field has gained a particular attention from researchers due to its importance and its huge impact on health care centers by allowing instantaneous understanding of medical images for doctors as well as patients. In addition, the significant progress made to date in artificial intelligence due to deep learning models contributed greatly to the AIC task [2]. Therefore, different techniques ranging from traditional template-based and/or retrieval-based systems to generative models based on deep-neural networks passing through various hybrid models that combine different techniques [6] emerged. During the last years, many systems have been proposed to compete for the ImageCLEF medical challenge. For the first step towards medical image captioning, which consists of concept detection in ImageCLEF medical task, the multi-label classifications is found to play a leading role. For instance, [2, 7] exploited the transfer learning to perform a multi-label classification by extracting significant features from medical images using pre-trained models such as the Resnet50, InceptionV3 .... In addition, Wang et al. [2] explored a retrieval-based topic modelling method to extract the most relevant clinical concepts from images similar to the input image. Encoder-decoder (CNN-RNN) architectures were explored by many studies to generate appropriate captions. In many cases, attention-mechanism is added to the baseline



CUI: C0180860;C4085887;C0035092;C0040398

UMLS terms: Filters; Inspiratory vital capacity test; Structure of renal vein; Tomography, Emission-computed
ImageID: synpic53561

CUI: C0227612;C0041618

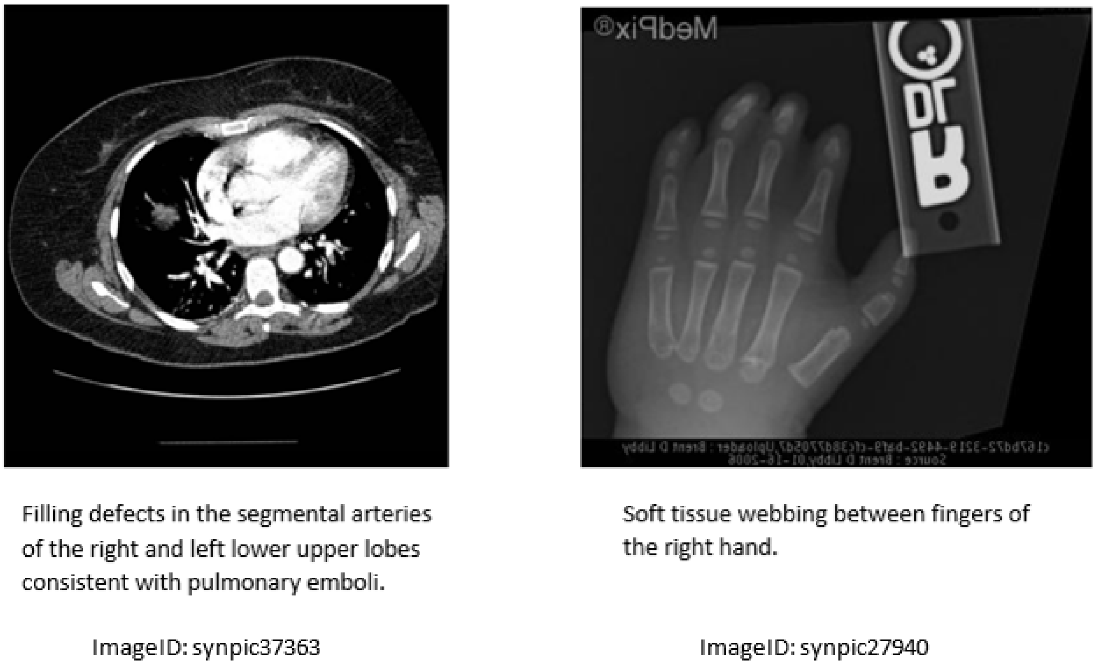
UMLS terms: Structure of inferior pole of kidney; Ultrasonography
ImageID: synpic38513

CUI: C0744903;C0264545;C1306645

UMLS terms: Hilar enlargement; Thickening of pleura; plain x-ray
ImageID: synpic58673

Figure 1: Radiology image samples from the ImageCLEF dataset where CUIs of each image and their respective UMLS terms are presented.

encoder-decoder model as in [8] who contributed to the ImageCLEF 2017 edition. Similarly, Hasan et al. [9] enriched the soft attention-based encoder-decoder model by inputting, to the decoder, the output of the classification model on image modalities. This allowed them to supplement the decoder with more fine grained details on the data to make the generation process more focused. Indeed, supplementary information such as image modalities or body parts could enhance the classification model as reported in Lyndon et al. [10]. Furthermore, the RNN decoder [11] is replaced by different variants such as LSTM [12] Xu et al.[7], or GRU Ambati and Dudyala [13] who used the captioning module to resolve the task of visual question answering. Likewise, Benzarti et al. [14] employed the captioning model to medical retrieval systems in order to obtain the query terms. From another perspective, Rahman [15] proposed to extract textual and visual information using RNN-based and CNN-based networks respectively, and then merge the outputs of both models to generate relevant captions. Similarly, Mishra and Banerjee [16] adopted the same technique aiming to detect retinal diseases and to generate appropriate medical reports. In other techniques, generative models are combined with retrieval systems for AIC. For example, Kougia et al. [17] proposed to exploit the image visual features to retrieve similar images with their known concepts and then combine them to predict enhanced captions of the input image. The prediction is performed using an encoder-decoder generative model. Likewise, [18, 19] suggested to use a retrieval policy module that makes a choice between generating new sentence or retrieving a template sentence.



Filling defects in the segmental arteries of the right and left lower upper lobes consistent with pulmonary emboli.

ImageID: synpic37363

Soft tissue webbing between fingers of the right hand.

ImageID: synpic27940

Figure 2: Radiology image samples from the ImageCLEF dataset where caption describing each image is presented.

3. Data analysis

The data used for this year edition is shared between both the ImageCLEFmed Caption and the ImageCLEF-VQAMed tasks. The dataset include three sets: the training set composed of the VQA-Med 2020 training data with 2756 medical images; the validation set consisting of 500 radiology images and the test set consisting of 444 radiology images. In addition, for the concept detection task, an excel file containing the medical image ID and the corresponding concepts CUIs is given to map each medical image onto its related concepts. Similarly, an excel file containing the captions of each medical image is provided for the caption prediction task. We present in Fig. 1 sample images with their underlying concepts and in Fig. 2 samples of images with their captions.

On further analysis of the datsaet, we illustrate in Fig. 3 the number of images per each CUI concept. It is obvious that the most frequent CUI is the 'C0040398' corresponding to 'Tomography, Emission-Computed' with 1159 images. Moreover, Fig. 4 presents the number of images by the number of CUIs associated to each image. We can see from this image that most of the medical images are attached to 2 to 3 concepts whereas the maximum number of concept CUIs per image is 10.

Moreover, we noticed that the maximum number of sentences per image caption is 5 and the maximum length of any caption is 47 words before pre-processing whereas it is 33 words after the pre-processing.

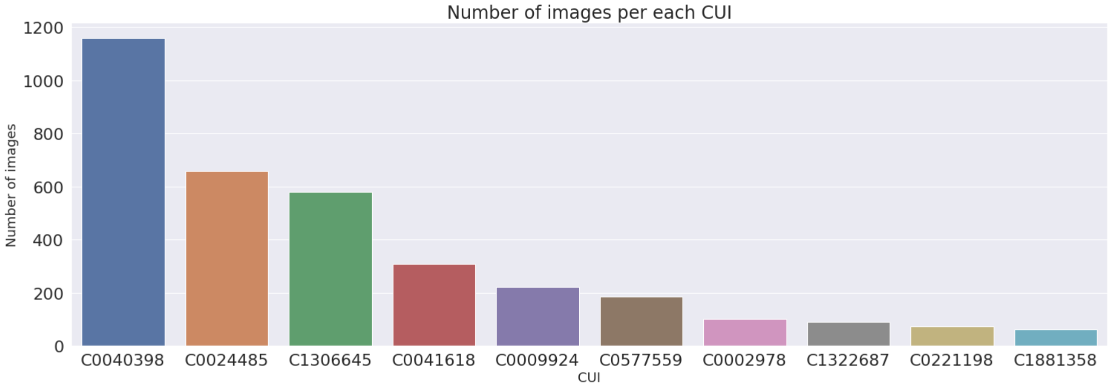


Figure 3: Number of medical images attached to each clinical concept

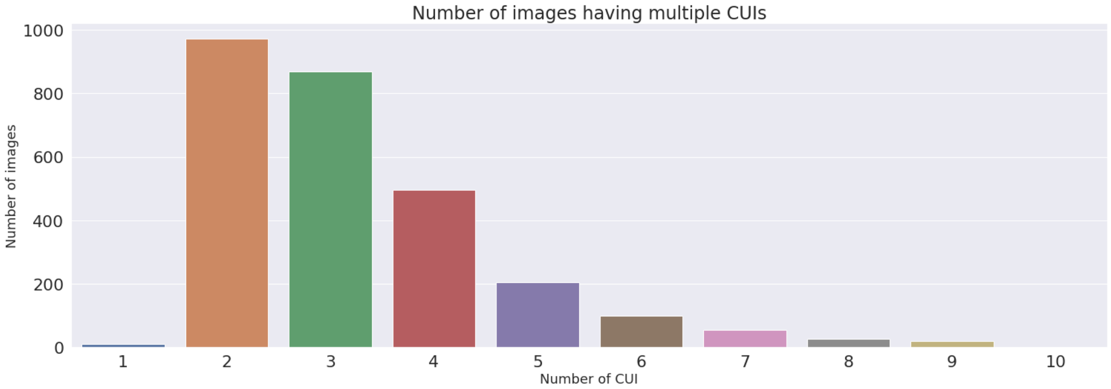


Figure 4: Distribution of the number of clinical concepts attached to each medical image

4. Methodology

For this year edition of the ImageCLEFmed Caption challenge [5], two subtasks are put forward: the concept detection and the caption prediction. To resolve each of these challenges, we present in the current paper a model for the concept detection based on a multi-label classification using a CNN architecture [20] and an encoder-decoder-based model for caption prediction. First, data is pre-processed to transform the text into understandable units. Images are as well pre-processed by performing a data augmentation technique on the training set. The detail is provided in the next subsections.

4.1. Data Pre-processing

Text Pre-processing: We apply a pre-processing scheme on the image to concept matching file to organize the concepts of each medical image into a list and create data-frames from image IDs and their underlying CUIs. In addition, we pre-process the captions using the NLTK package, by performing tokenization, punctuation and stop-words removal (using the default NLTK's "english" stopword list), lower casing each token and finally applying the stemming to obtain filtered sentences for each medical image using the NLTK's Snowball stemmer. Morover, for the RNN decoder, we add two tokens: '<start>' and '<end>' to identify the beginning and the end of each caption.

Image Pre-processing: Image data generators are created for the three sets of data to pre-process the images before feature extraction. These generators iterate over the data subsets and normalize the images to facilitate the features calculation. Then we apply horizontal and vertical flip in addition to a crop-center based data augmentation technique that we implement with a fraction of 87.5%. The implemented data augmentation approach allowed us to expand the training data without altering the visual content of the image. Then, the images are resized to fit in the feature extractor size which is 224 × 224 × 3 in our case.

4.2. Concept Detection

As we mentioned before, the first task of concept detection aims at identifying and localizing the relevant concepts present in each medical image. Therefore, we exploit the visual image content to extract significant visual features that allow us to distinguish the underlying concepts. These concepts are used further to construct image captions and could as well be utilized for the context-based images and information retrieval purposes.

To achieve the first step towards caption prediction, we performed a multi label classification. In the first run, we extracted image features using the pre-trained MobileNet-V2 and then performed the classification using a GRU network [21]. In the second run, we performed feature extraction using the pre-trained Inception-V3 model followed by a classification using a CNN network. ImageNet weights were used for both models and features were extracted from the last convolutional layer.

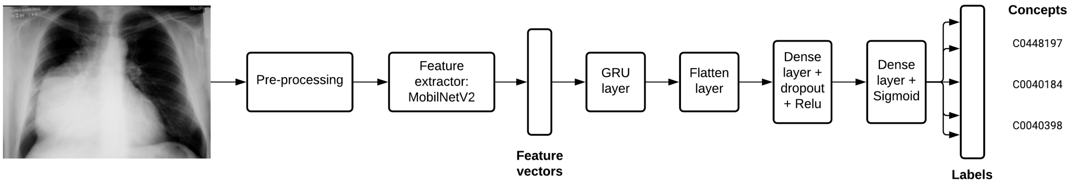


Figure 5: GRU based multi-label classification of features extracted from the MobilNetV2 pre-trained model into medical concepts.

As illustrated by Fig. 5, the features extracted from the radiology images using the pre-trained MobileNet-V2 model are passed through a GRU layer, a flatten layer and then a fully connected layer with a Relu activation function and dropout. Finally, the labels of each medical image are predicted using a fully connected layer with a Sigmoid activation function.

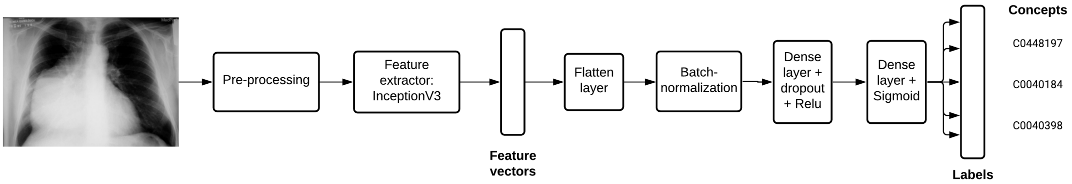


Figure 6: CNN based multi-label classification of features extracted from the InceptionV3 pre-trained model into medical concepts.

Likewise, the features extracted from the radiology images using the pre-trained Inception-V3 model are passed through a flatten layer, a fully connected layer with batch normalization, Relu activation function, and dropout. Then, the probability of each class is calculated using a fully connected layer with a Sigmoid activation function (as shown by Fig. 6). If the probability is greater than 20%, we assert the input image belongs to that class. This probability was fixed to 20% after experimenting with different thresholds. If this threshold is fixed to a higher value, we would get a lot of false negatives where many images are not classified in their correct classes. However, if it is fixed to a smaller value, we would get a lot of false positives where many images are categorized into incorrect classes. Finally, we map the labels to their corresponding concepts.

4.3. Caption Prediction

The second sub-task relies on the concept vocabulary detected in the first sub-task in addition to the visual features extracted from the medical images to establish relationships between them and predict descriptive caption for each medical image. We attempted to address the issue of caption generation using an encoder-decoder architecture with attention mechanism.

The visual features are extracted from the medical images using a pre-trained model 'InceptionV3' where weights from the ImageNet were employed. Then, these features are passed through a CNN model that is composed of a fully connected layer to flatten the feature vector. Next, an attention mechanism is employed to focus on the most important parts of the image and a context vector is constructed. Captions are pre-processed as we mentioned before and passed to an embedding layer. A concatenation layer is farther used to merge the context vector with the resulting embedding vector and the output is passed to a GRU layer. A flatten layer and two fully connected layers with dropout and a Relu for the first one and a Sigmoid activation function for the second were employed. Finally, relevant captions are generated word by word until the '<end>' token is met. Figure. 7 illustrates the attention-based encoder-decoder architecture we used to construct new captions for the medical images. Moreover, we exploit the teacher forcing during the training by using the ground truth sequences at every step rather than the sequence of newly generated words at previous steps.

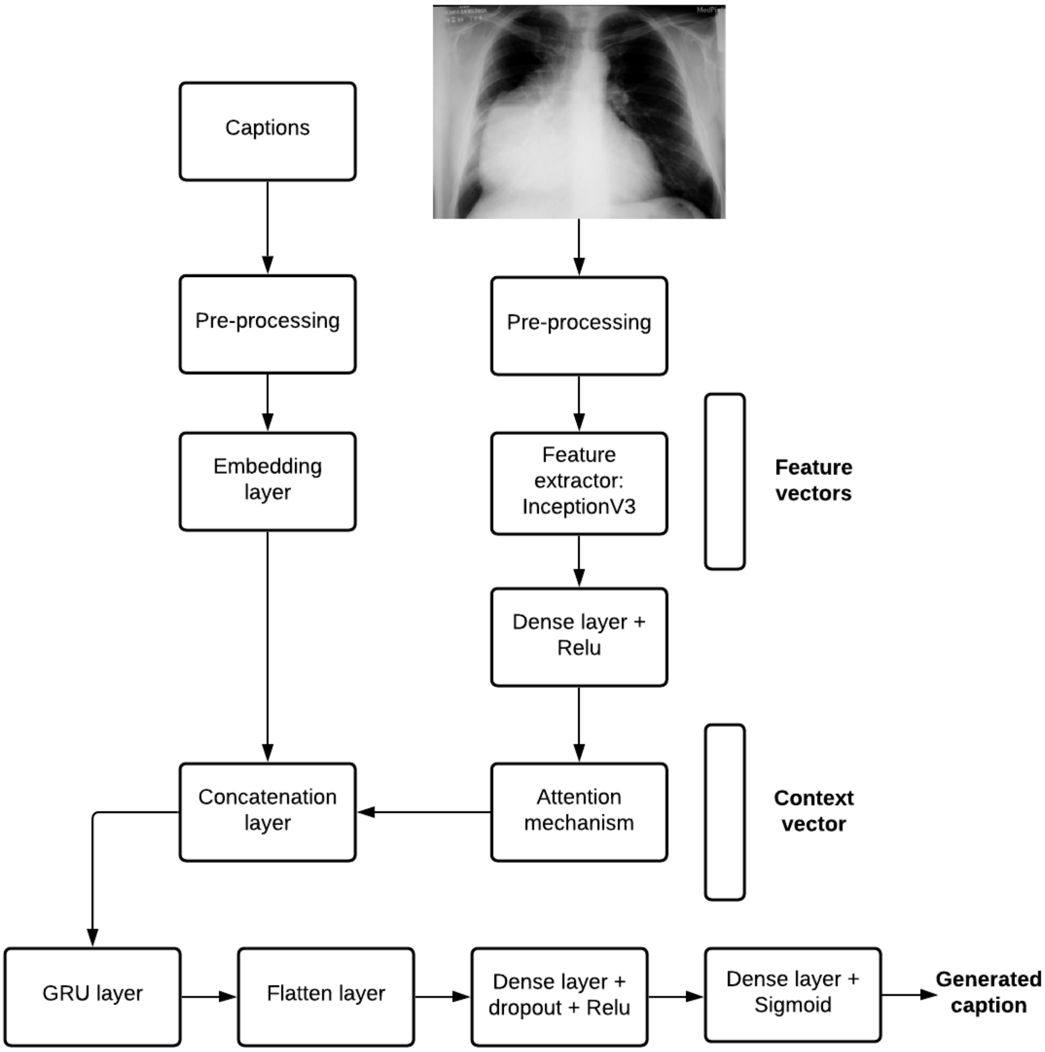


Figure 7: The attention-based encoder decoder architecture used for caption prediction. Features are extracted from the medical images and passed to an attention mechanism to select the most important parts of the image. Then, the constructed context vector is concatenated with the embedding vector obtained from the captions and inputted to a GRU layer and captions are finally generated word by word.

5. Experiment and Results

We used the data provided by the ImageCLEF medical task to evaluate the performance of our models. Three subsets are used for the training, the validation and the test respectively. We report in this section, the performance metrics calculated as well as the results we obtained for both models.

5.1. Performance metrics

We calculate the F_Measure, using the default 'binary' averaging method, for the concept detection task as follows:

$$F\_Measure = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{1}$$

Where the recall and the precision are calculated as follow and TP, FN, TN, FP correspond to true positive, false negative, true negative and false positive respectively.

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

For the caption prediction task, we calculate the BLEU score by assuming that each caption is a single sentence even if it is actually composed of several sentences. For that we use the default implementation of the Python NLTK based on [22]:

$$BLEU = BP \cdot \exp(\sum_{n=1}^{N} w_n \cdot \log p_n) \tag{4}$$

Where BP refers to the brevity penalty, N refers to the number of n_grams (uni-gram, bi-gram, 3-gram and 4-gram), W_n refers to the weight of each modified precision and P_n refers to the modified precision. By default N=4 and W_n= 1/N = 1/4.

Brevity Penalty (BP) allows us to pick the candidate caption which is most likely close in length, word choice and word order to the reference caption. It is an exponential decay and is calculated as follows:

$$BP = \begin{cases} 1 & c > r \\ \exp^{(1-r/c)} & c \leqslant r \end{cases} \tag{5}$$

Where r refers to the count of words in the reference caption and c refers to the count of words in the candidate caption.

Modified precision is computed for each n_gram as the sum of clipped n_gram counts of the candidate sentences in the corpus divided by the number of candidate n_grams as shows "(6)" [22]. It allows us to compute the adequacy and the fluency of the candidate translation to the reference translation.

$$p_n = \frac{\sum\limits_{C \in \{Candidates\}} \sum\limits_{n\_gram \in C} Count_{clip}(n\_gram)}{\sum\limits_{C' \in \{Candidates\}} \sum\limits_{n\_gram' \in C'} Count(n\_gram')} \tag{6}$$

5.2. Results

We obtained an average F_measure of 24.49% during the training process and a value of 14.3% during the inference process for the concept detection task using the MobileNetV2 as a feature extractor and the GRU network as a classifier.

Table 1
F_measure results for the concept detection task

TeamName	Run ID	Validation set	Test set
RomiBed	136025	23.65%	13.7%
	136011	24.49%	14.3%

Similarly, we obtained an average F_measure of 23.28% during the training process and a value of 13.7% during the inference process using the InceptionV3 model as a feature extractor and a CNN network as a classifier. However, our F_measure results are comparatively lower compared to the leading group (IALab_PUC) with a score of 50.5 %. Results are illustrated by Table. 1 where the run ID 136011 corresponds to the first configuration and the 136025 corresponds to the second configuration. Figure. 8 shows the evolution of the multi-label classification model accuracy across the epochs.

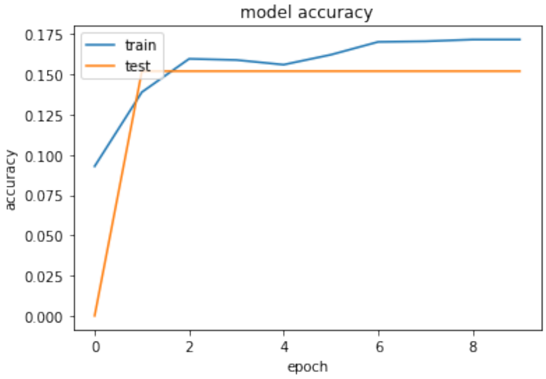


Figure 8: Evolution of the accuracy per epoch for the concept detection task

For the caption prediction task, we obtained a BLEU score of 0.287 during the training process and a value of 0.243 during the inference. Results are illustrated by Table. 2. In addition, Figure. 9 shows the loss calculated during the training process of the encoder-decoder model. We noticed the decrease of the cross-entropy loss across the epochs.

Table 2
BLEU score results for the caption prediction task

TeamName	Run ID	Validation set	Test set
RomiBed	135896	0.287	0.243

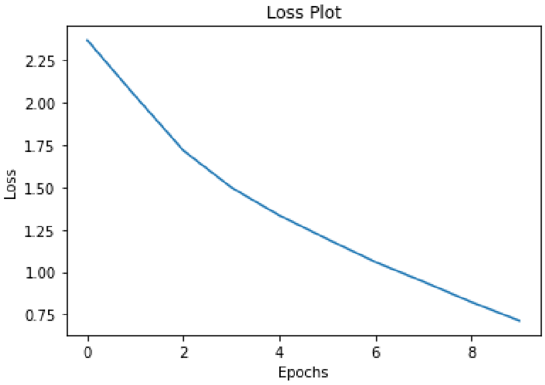
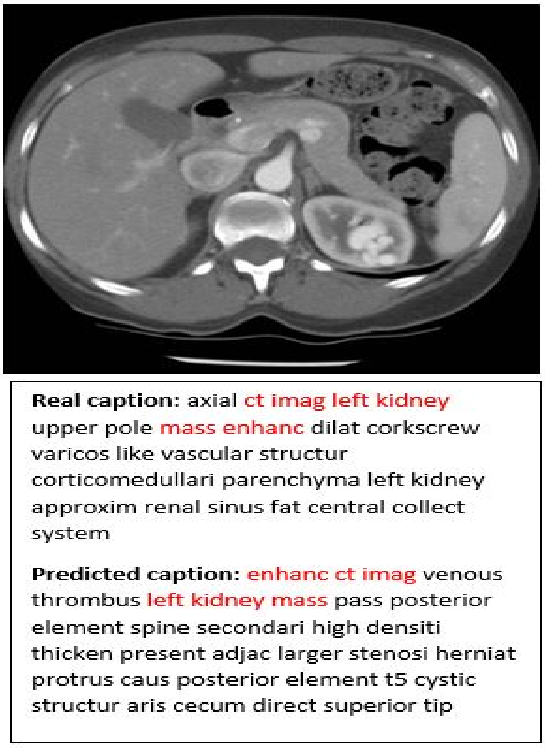


Figure 9: Evolution of the training loss per epoch for the caption prediction task



Real caption: axial ct imag left kidney upper pole mass enhanc dilat corkscrew varicos like vascular structur corticomedullari parenchyma left kidney approxim renal sinus fat central collect system

Predicted caption: enhanc ct imag venous thrombus left kidney mass pass posterior element spine secondari high densiti thicken present adjac larger stenosi herniat protrus caus posterior element t5 cystic structur aris cecum direct superior tip

Figure 10: A sample of medical image with its real caption and the caption generated by our system.

Finally, we show an example of a random medical image from the validation set with its real caption and the newly generated caption in Fig. 10. We observed a BLEU score of 0.339 for this image, where 8 words were correctly generated but the order of the words in the generated caption is different.

6. Conclusion and Future Work

We presented in this paper our contribution to the ImageCLEF 2021 medical task where we proposed a CNN based multi-label classification model for the concept detection task and an attention-based encoder-decoder model for the caption prediction task. For both models, a transfer learning is used to extract significant features from the real radiology images and a data augmentation based on center-cropping is applied to expand the used training subset. The evaluation of the caption detection task is conducted using the mean F_measure for which we obtained a score of 14.3%. Furthermore, BLEU score is used to evaluate the reliability of the generated captions for the caption prediction task for which we obtained a score of 0.243. We believe that we did not obtain promising results due to the small amount of data used and the fact that we did not explore more fine-tuned parameters for both models for time constraints. In addition, we did not include the textual features substituted by the medical concepts to generate the new captions. In future work, we will integrate the textual features of the images to the visual information to obtain more relevant captions. We will also investigate more advanced deep learning algorithms inline with more fine-tuned parameters. In addition, we will investigate our model performance on larger scale dataset for medical image captioning.

Acknowledgments

This work is supported by the Academy of Finland Profi5 DigiHealth project (#326291), which is gratefully acknowledged.

References

[1] I. Allaouzi, M. Ben Ahmed, B. Benamrou, M. Ouardouz, Automatic Caption Generation for Medical Images, in: Proceedings of the 3rd International Conference on Smart City Applications, 2018, pp. 1–6.

[2] X. Wang, Z. Guo, Y. Zhang, J. Li, Medical Image Labelling and Semantic Understanding for Clinical Applications, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2019, pp. 260–270.

[3] H. Park, K. Kim, J. Yoon, S. Park, J. Choi, Feature Difference Makes Sense: A Medical Image Captioning Model Exploiting Feature Difference and Tag Information, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, 2020, pp. 95–102.

[4] O. Pelka, A. Ben Abacha, A. García Seco de Herrera, J. Jacutprakart, C. M. Friedrich, H. Müller, Overview of the ImageCLEFmed 2021 Concept & Caption Prediction Task, in: CLEF2021 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Bucharest, Romania, 2021.

[5] B. Ionescu, H. Müller, R. Peteri, A. B. Abacha, M. Sarrouti, D. Demner-Fushman, S. A. Hasan, S. Kozlovski, V. Liauchuk, Y. Dicente, V. Kovalev, O. Pelka, A. G. S. de Herrera, J. Jacutprakart, C. M. Friedrich, R. Berari, A. Tauteanu, D. Fichou, P. Brie, M. Dogariu, L. D. Ştefan, M. G. Constantin, J. Chamberlain, A. Campello, A. Clark, T. A. Oliver, H. Moustahfid,

A. Popescu, 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, Proceedings of the 12th International Conference of the CLEF Association (CLEF 2021), LNCS Lecture Notes in Computer Science, Springer, Bucharest, Romania, 2021.

[6] M. Alsharid, H. Sharma, L. Drukker, P. Chatelain, A. T. Papageorghiou, J. A. Noble, Captioning Ultrasound Images Automatically, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2019, pp. 338–346.

[7] J. Xu, W. Liu, C. Liu, Y. Wang, Y. Chi, X. Xie, X.-S. Hua, Concept Detection based on Multi-label Classification and Image Captioning Approach-DAMO at ImageCLEF 2019., in: CLEF (Working Notes), 2019.

[8] S. A. Hasan, Y. Ling, J. Liu, R. Sreenivasan, S. Anand, T. R. Arora, V. V. Datla, K. Lee, A. Qadir, C. Swisher, et al., PRNA at ImageCLEF 2017 Caption Prediction and Concept Detection Tasks., in: CLEF (Working Notes), 2017.

[9] S. A. Hasan, Y. Ling, J. Liu, R. Sreenivasan, S. Anand, T. R. Arora, V. Datla, K. Lee, A. Qadir, C. Swisher, et al., Attention-based Medical Caption Generation with Image Modality Classification and Clinical Concept Mapping, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2018, pp. 224–230.

[10] D. Lyndon, A. Kumar, J. Kim, Neural Captioning for the ImageCLEF 2017 Medical Image Challenges., in: CLEF (Working Notes), 2017.

[11] D. E. Rumelhart, G. E. Hinton, R. J. Williams, Learning representations by back-propagating errors, nature 323 (1986) 533–536.

[12] S. Hochreiter, J. Schmidhuber, Long Short-term Memory, Neural computation 9 (1997) 1735–1780.

[13] R. Ambati, C. R. Dudyala, A Sequence-to-Sequence Model Approach for ImageCLEF 2018 Medical Domain Visual Question Answering, in: 2018 15th IEEE India Council International Conference (INDICON), IEEE, 2018, pp. 1–6.

[14] S. Benzarti, W. B. A. Karaa, H. H. B. Ghezala, Cross-Model Retrieval Via Automatic Medical Image Diagnosis Generation, in: International Conference on Intelligent Systems Design and Applications, Springer, 2019, pp. 561–571.

[15] M. Rahman, A Cross Modal Deep Learning Based Approach for Caption Prediction and Concept Detection by CS Morgan State., in: CLEF (Working Notes), 2018.

[16] S. Mishra, M. Banerjee, Automatic Caption Generation of Retinal Diseases with Self-trained RNN Merge Model, in: Advanced Computing and Systems for Security, Springer, 2020, pp. 1–10.

[17] V. Kougia, J. Pavlopoulos, I. Androutsopoulos, AUEB NLP Group at ImageCLEFmed Caption 2019., in: CLEF (Working Notes), 2019.

[18] C. Y. Li, X. Liang, Z. Hu, E. P. Xing, Hybrid Retrieval-Generation Reinforced Agent for Medical Image Report Generation, in: Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS 2018), NIPS'18, Curran Associates Inc., 2018, p. 1537–1547.

[19] F. Wang, X. Liang, L. Xu, L. Lin, Unifying Relational Sentence Generation and Retrieval for Medical Image Report Composition, IEEE Transactions on Cybernetics (2020). doi:10.1109/TCYB.2020.3026098.

[20] Y. LeCun, Y. Bengio, et al., Convolutional networks for images, speech, and time series, The handbook of brain theory and neural networks 3361 (1995) 1995.

[21] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using RNN encoder-decoder for statistical machine translation, arXiv preprint arXiv:1406.1078 (2014).

[22] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a Method for Automatic Evaluation of Machine Translation, in: Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311–318.