ImageCLEF 2020: Image Caption Prediction using Multilabel Convolutional Neural Network

Sarada Devi Arul[0000−0001−5943−8494] , Kavitha Srinivasan[0000−0003−3439−2383]

Department of CSE, SSN College of Engineering, Kalavakkam–603110, India
sarada1806@cse.ssn.edu.in, kavithas@ssn.edu.in

Abstract. Radiology imaging encompasses different imaging modalities and the images are acquired from the human body for diagnostic and treatment purpose. The different imaging modalities are Computed Tomography (CT), Ultrasound, X-Ray, Positron Emission Tomography (PET), Magnetic Resonance Imaging (MRI), Angiography and Cardiac Output (CO). These images are used to identify the disease types and its stages. In this paper, an automatic caption detection technique for multi modality radiology images of various disease types and organs is implemented and explained for the task of ImageCLEF 2020. This research work includes dataset collection, preprocessing of the dataset and caption prediction using multilabel Convolutional Neural Network (CNN). The correctness of the predicted captions is validated using the metric, F1 Score. The result obtained from the proposed model for the test set is 13.46%. The achieved result is at 42nd position in the overall leaderboard of the ImageCLEF 2020 caption - concept detection for radiology images.

Keywords: Radiology images · Caption detection · Preprocessing · Multilabel CNN· F1 score.

1 Introduction

Medical imaging or radiology imaging can be acquired using various modalities like Computed Tomography (CT), Ultrasound, X-Ray, Positron Emission Tomography (PET), Magnetic Resonance Imaging (MRI), Angiography and Cardiac Output (CO) [3]. The applications of the radiology imaging include classification, prediction, information extraction, information retrieval, concept detection etc.

Image Caption identification is a kind of concept detection or prediction application. Captioning task can be carried out for natural images and medical images. In natural images, major features like colour and shape are extracted for caption identification [2]. However in case of medical images it is tedious to extract the important features, so the identification result is not accurate. Also,

interpreting and summarizing the insights gained from medical images such as radiology output is a time-consuming task and requires highly trained experts [1]. To address these issues, the automatic generation of captions for different modalities becomes an important task in reality [2, 8]. In this paper, a multi-label Convolutional Neural Network (CNN) approach for caption prediction is discussed with results. This work is a subtask of the medical tasks of ImageCLEF 2020 and establishes detection of captions for multimodality radiology images.

The sections includes the following: Section 1 gives a brief introduction about the necessity to perform caption prediction. Section 2 describes about the dataset which includes radiology images of various modalities and Section 2.1 details about the data preprocessing procedures. Section 3 explains the proposed model using multilabel convolutional neural network along with the parameters for analysis. In Section 4, the results are discussed. Finally, Section 5 concludes this paper with further refinement of the proposed model.

2 Dataset

In this edition of ImageCLEF 2020, for concept detection a total of 84,257 radiology images are given, out of which 64,753 are training images, 15970 are for validation and 3534 are for testing the model [13, 14]. All these images are present in any one of the seven modality folders namely, Computed Tomography (CT), Ultrasound, X-Ray, Positron Emission Tomography (PET), Magnetic Resonance Imaging (MRI), Angiography and Cardiac Output (CO). Similarly the captions of the images of seven modalities are given in seven excel sheets appropriately. In Figure 1, sample image for each modality is shown with its corresponding image id of the given dataset for each modality [13, 14].

Table 1. Distribution of dataset across 7 modalities

Modality	Training	Validation	Test	Total
Angiography	4713	1132	325	6170
Cardiac Output	487	73	49	609
CT	20031	4992	1140	26163
MRI	11,447	2848	562	14857
PET	502	74	38	614
Ultrasound	8629	2134	502	11265
X-Ray	18944	4717	918	24579

On further analysis of dataset, the maximum number of captions per image is nearly 140 and each image is of different size, are the challenges in generating the relevant captions.

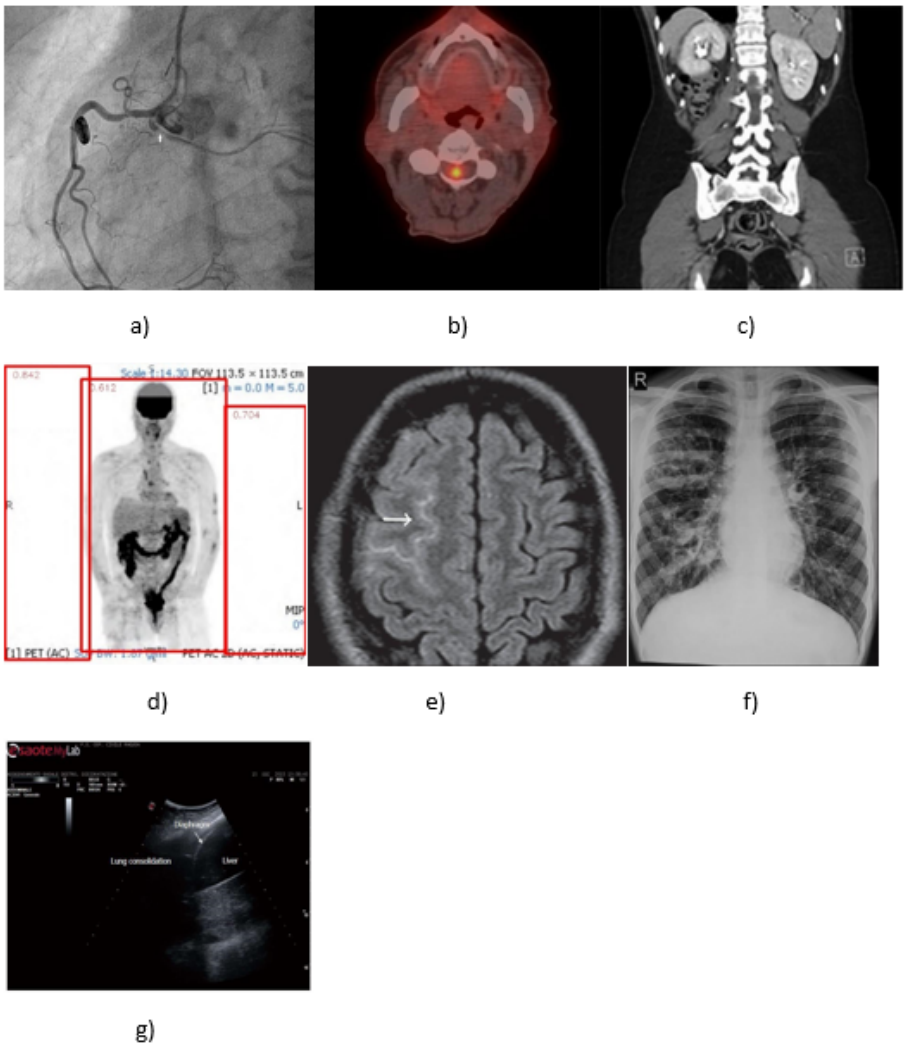

Fig. 1. Radiology images of seven modalities with image id and modality a) ROCO2_CLEF_00001 Angiography b) ROCO2_CLEF_05850 CO c) ROCO2_CLEF_06406 CT d) ROCO2_CLEF_45724 PET e) ROCO2_CLEF_31429 MRI f) ROCO2_CLEF_57063 XRAY g) ROCO2_CLEF_46300 Ultrasound

2.1 Data Preprocessing

Preprocessing of Text In the given dataset single file is present for text. This file includes image id and their corresponding caption unique id in sorted order. Using these labels one of the inputs for multilabel CNN is created. The input file is created in excel format, where the caption id as header and rows are filled with the image id of that particular modality alone. For each image id, one hot vector form [11] is created in such a way that the captions of specific columns are made as 1 and others as 0. Similarly, this is carried out for all the seven modalitites. Therefore, 7 different sized one hot vectors are derived. The one hot vector sizes for angiography, CO, CT, PET, MRI, X-Ray and Ultrasound modalities are 2578, 1675, 3013, 1491, 2980, 2986 and 2877 respectively.

Preprocessing of Images The radiology images of the given 7 modalities are of varying sizes. But, the images of same dimension must be given as input to the CNN. Therefore, resizing of the image is carried out in such a way that all the images are of same width and height i.e (600, 600), since most of the images in the dataset are of that size only.

3 Methodolgy

Initially, CNN approach is applied to predict the image captions. The conventional CNN usually takes the folder name as captions, where as the given clef 2020 dataset comprises of more than one label for each image. Hence, it generates inappropriate captions and also only single caption per image. To address this issue, a multilabel CNN is proposed for this task. The given dataset has multiple modalities and maximum of 140 captions for each image [6]. The layers chosen for the proposed model are convolutional layer, max pooling layer, flattening layer and dense layer.

Import Keras and other packages that are required in building the CNN like Sequential, Convolution2D, MaxPooling2D, Flatten and Dense layer. Build the model using the Sequential.add() function. Four convolutional layers are added with the filter size as 16, 16, 32 and 32 respectively and the kernel size as (3,3). These layers are used to extract the high-level features such as edges and boundaries from the input image [5]. Add a pooling layer with a size of (2, 2), to reduce the spatial size of the representation of input image. One flatten layer is added to generate a vector from the fully connected layers and the last dense layer outputs as either 1 or 0.

Finally, the output nodes are fixed in the last layer based on the one hot vector size for each of the seven modality [11]. Each output node belongs to some class. Categorical_crossentropy loss function is used, since it is more suitable for multiclass classification [9, 12]. The sigmoid activation function used on the final layer converts each score of the final node between 0 to 1 independent of the other score. If the score of the particular class is more than 0.5, the data is classified into that class. And there could be multiple classes having a score of more than 0.5

independently. Thus the data could be classified into multiple classes. In Figure 2, the sample image (ROCO2_CLEF_05873) from Cardiac Output modality is given as input to the model for caption prediction.

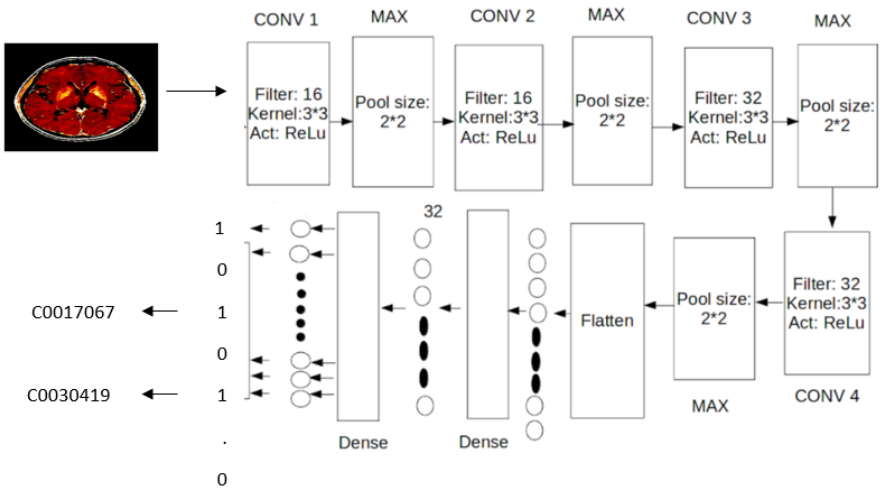

Fig. 2. Overall process of multilabel CNN

4 Experiment and Results

CNN model has many hyperparameters to build it efficiently. By fixing them appropriately, the results can be improved.

To make the decision on fixing the layers of convolution, visualisation can be carried out. After visualization, four convolutional layers are fixed, since the boundaries and edges of the image is more visible than the three layers. In Figure 3, the input image with image id ROCO2_CLEF_05865 from CO modality is given for understanding and visualizing the effect of convolution layers.

Table 2. CNN hyperparameters.

Hyperparameters	Values
No. of Convolutional layer	4
No. of filters in each layer	16, 16, 32, 32
Pooling function	max
Activation function	relu, sigmoid
No. of epochs	3
Loss type	categorical cross entrophy
Optimizer	adam

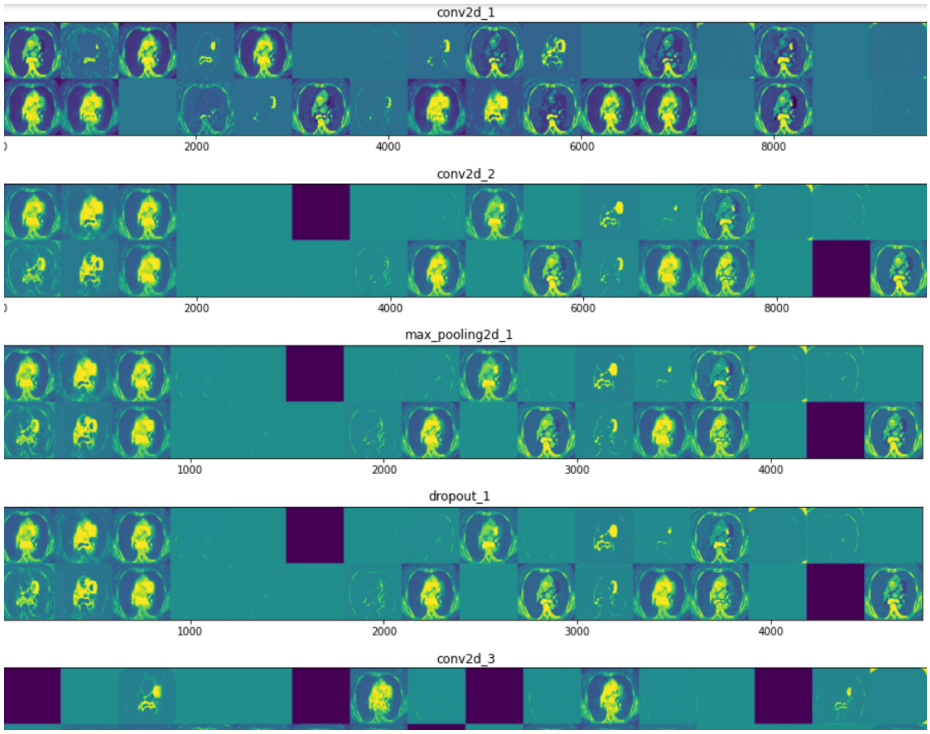

Fig. 3. Visualisation of Convolutional layers

The multilabel CNN model with the specific hyperparameters [3, 10] has been evaluated using the given dataset and appropriate performance metrics [12]. All the seven folders are trained with 3 epochs to build the seven different models. Testing is done with their respective models and the captions are obtained. The resulted accuracy of training and validation set are 0.4034 and 0.2478 respectively. For the test set, the F1 score obtained is 0.1346 in a single run and ranked as 42nd in the leaderboard of the ImageCLEF 2020 caption task. The F1 score is comparatively very less, because only 20 captions are used in the prediction of test set. The main challenges of this task are: large dataset with images of different characterstics, implementation of one hot vector with sparse data, maximum number of captions is around 140, training model needs more time, if the internet is used for execution it becomes still more tedious process, needs high requirements in terms of hardware like memory, processor etc for better computability.

5 Conclusion and Future Work

In this paper, an automatic caption prediction for multimodality radiology images is implemented and explained for the given ImageCLEF 2020 task dataset using multilabel CNN. The dataset is preprocessed in both text and image aspects, and the maximum number of captions per image is identified. From the number of captions identified, one hot vector is derived for every modality and training of the model is carried out. The model is evaluated using F1 metric for the test set (3534 images), which resulted in 13.46%. The limitations of the work are number of captions used in testing and hyperparameters of the multilabel CNN model.

In future, the prediction results can be improved further based on the dataset, methods to modify the one hot vector in an efficient way and advanced deep learning techniques.

6 Acknowledgement

Our profound gratitude to SSN College of Engineering, Department of CSE, for allowing us to utilize the High Performance Computing Laboratory and GPU Server for the execution of this challenge successfully.

References

1. Carsten Eickhoff, Immanuel Schwall, Alba G. Seco de Herrera, Henning Muller.: Overview of ImageCLEFcaption 2017- Image Caption Prediction and Concept Detection for Biomedical Images. In: CLEF 2017 Working Notes. CEUR Workshop Proceedings, Switzerland (2017).
2. Raffaella Bernardi, Ruket Cakici,Desmond Elliott, Aykut Erdem, Erkut Erdem, Frank Keller keller, Adrian MuscatBarbara Plank.: Automatic Description Generation from Images: A Survey of Models, Datasets, and Evaluation Measures. arXiv:1601.03896v2, pp.1–34, (2017).

3. Baoyu Jing, Pengtao Xie, Eric P.Xing.: On the Automatic Generation of Medical Imaging Reports. arXiv:1711.08195v3, pp.2577–2586, (2018)
4. Alba G. Seco de Herrera, Carsten Eickhoff, Vincent Andrearczyk, Henning Muller.: Overview of the ImageCLEF 2018 Caption Prediction Tasks. In: CLEF 2018 Working Notes. CEUR Workshop Proceedings, Switzerland (2018).
5. Jeel Sukhadiya, Harsh Pandya, Vedant Singh.: Comparison of Image Captioning Methods. In: IJEDR, Volume 6, Issue 4, pp.43–48, (2018)
6. Yu Zhang, Xuwen Wang, Zhen Guo, Jiao Li.: ImageSem at ImageCLEF 2018 Caption Task: Image Retrieval and Transfer Learning. In: CLEF 2018 Working Notes. CEUR Workshop Proceedings, Switzerland (2018).
7. Md Mahmudur Rahman.:A Cross Modal Deep Learning Based Approach for Caption Prediction and Concept Detection by CS Morgan State. In: CLEF 2018 Working Notes. CEUR Workshop Proceedings, Switzerland (2018).
8. Obioma Pelka, Christoph M.Friedrich, Alba G. Seco de Herrera, Henning Muller.: Overview of the ImageCLEFmed 2019 Concept Detection Task. In: CLEF 2019 Working Notes. CEUR Workshop Proceedings, Switzerland (2019).
9. Vasiliki Kougia, John Pavlopoulos, Ion Androutsopoulos.: AUEB NLP Group at ImageCLEFmed Caption 2019. In: CLEF 2019 Working Notes. CEUR Workshop Proceedings, Switzerland (2019).
10. Zhen Guo, Xuwen Wang, Yu Zhang, Jiao Li.: ImageSem at ImageCLEFmed Caption 2019 Task: a Two-stage Medical Concept Detection Strategy. In: CLEF 2019 Working Notes. CEUR Workshop Proceedings, Switzerland (2019).
11. Jing Xu, Wei Liu, Chao Liu, Yu Wang, Ying Chi, Xuansong Xie, Xiansheng Hua.: Concept detection based on Multi-label Classification and Image Captioning Approach DAMO at ImageCLEF 2019. In: CLEF 2019 Working Notes. CEUR Workshop Proceedings, Switzerland (2019).
12. Sonit Singh, Sarvnaz Karimi, Kevin Ho-Shon, Len Hamey.: Biomedical Concept Detection in Medical Images: MQ-CSIRO at 2019 ImageCLEFmed Caption Task. In: CLEF 2019 Working Notes. CEUR Workshop Proceedings, Switzerland (2019).
13. Obioma Pelka, Christoph M. Friedrich, Alba G. Seco de Herrera Henning Müller, Overview of the ImageCLEFmed 2020 Concept Prediction Task: Medical Image Understanding. In: CLEF2020 Working Notes. CEUR Workshop Proceedings(2020), Thessaloniki, Greece, CEUR-WS.org $$, Springer (September 22-25, 2020).
14. Bogdan Ionescu, Henning Müller, Renaud Péteri, Asma Ben Abacha, Vivek Datla, Sadid A. Hasan, Dina Demner-Fushman, Serge Kozlovski, Vitali Liauchuk, Yashin Dicente Cid, Vassili Kovalev, Obioma Pelka, Christoph M. Friedrich, Alba García Seco de Herrera, Van-Tu Ninh, Tu-Khiem Le, Liting Zhou, Luca Piras, Michael Riegler, Pål Halvorsen, Minh-Triet Tran, Mathias Lux, Cathal Gurrin, Duc-Tien Dang-Nguyen, Jon Chamberlain, Adrian Clark, Antonio Campello, Dimitri Fichou, Raul Berari, Paul Brie, Mihai Dogariu, Liviu Daniel Ștefan, Mihai Gabriel Constantin, Overview of the ImageCLEF 2020: Multimedia Retrieval in Lifelogging, Medical, Nature, and Internet Applications In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the 11th International Conference of the CLEF Association (CLEF 2020), Thessaloniki, Greece, LNCS Lecture Notes in Computer Science, 12260, Springer (September 22-25, 2020).
15. Kavitha, S., Nandhinee, P.R., Harshana, S., Srividya, J.S., Harrinei, K.: ImageCLEF 2019: A 2D Convolutional Neural Network Approach for Severity Scoring of Lung Tuberculosis using CT Images. In: CLEF 2019 Working Notes. CEUR Workshop Proceedings, Switzerland (2019).