IUST_NLPLAB at ImageCLEFmedical Caption Tasks 2022

Malihe Hajihosseini[1], Yasaman Lotfollahi[1], Melika Nobakhtian[1], Mohammad Mahdi Javid[1], Fateme Omidi[1] and Sauleh Eetemadi[2]

[1] *Student at School of Computer Engineering, Iran University of Science and Technology, Tehran, Islamic Republic Of Iran.*
[2] *Assistant Professor of Computer Science, School of Computer Engineering, Iran University of Science and Technology, Tehran, Islamic Republic Of Iran.*

Abstract

We present models implemented by the IUST_NLPLAB group for ImageCLEFmedical Caption Task 2022. This task contains two subtasks: Concept Detection and Caption Prediction. Under the first subtask, the model should extract medical concepts contained in radiology images. These concepts can be used for context-based image and information retrieval. Under the second subtask, the model predicts the caption for a medical image. This can be used for improving the diagnosis and treatment of diseases by saving time, money and helping physicians. We used Retrieval Learning, Ensemble Learning, Multi Label Classification and Deep Learning techniques to rank 1st in the caption prediction subtask with 16 BLEU points over the second ranked group. We also ranked 8th in the concept detection task with a 5 percent gap from the top ranked group in F1 score.

Keywords

Medical Image Captioning, Concept Detection, Caption Prediction, Deep Learning, Multi-label Classification

1. Introduction

ImageCLEF[1] is part of CLEF[1]. ImageCLEF was launched in 2003 and added a medical task in 2004. Although it started with four participants, in 2020 was able to attract more than one hundred and ten participants from all around the world to participate in the competition. ImageCLEF includes various sections that retrieve and classify visual information using textual and visual data and their combinations.

In recent years, ImageCLEF has used the AIcrowd[2] platform to publish datasets and receive submissions. In 2022, one person from each group had to register with AIcrowd, then access

CLEF 2022: Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy
m_hajihosseini@comp.iust.ac.ir (M. Hajihosseini); y_lotfollahi@comp.iust.ac.ir (Y. Lotfollahi); m_nobakhtian@comp.iust.ac.ir (M. Nobakhtian); mahdijavid1380@yahoo.com (M. M. Javid); f_omidi97@comp.iust.ac.ir (F. Omidi); sauleh@iust.ac.ir (S. Eetemadi)
https://mohammadmahdijavid.ir/ (M. M. Javid); http://sauleh.github.io/ (S. Eetemadi)
0000-0002-0067-1184 (M. Hajihosseini); 0000-0002-4808-5141 (Y. Lotfollahi); 0000-0001-9788-2764 (M. Nobakhtian); 0000-0002-8447-7513 (M. M. Javid); 0000-0002-5886-7315 (F. Omidi); 0000-0003-1376-2023 (S. Eetemadi)
© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

[1]Conference and Labs of the Evaluation Forum
[2]https://www.aicrowd.com/ (last accessed: 2022-05-27)

CEUR-WS.org/Vol-3180/paper-106.pdf

the dataset and submit results on specified dates. Each group could register up to 10 successful submissions for each task. Five unsuccessful submissions for each group in each task were also allowed.

In ImageCLEFmedical 2022, two tasks were proposed: Image Captioning and Tuberculosis CT analysis. We selected the Image Captioning task from the ImageCLEFmedical section to participate in the competition. ImageCLEF medical Image Captioning task in 2022 contained two subtasks: Concepts Detection and Caption Prediction. These tasks have many uses, but their most important usage is to help physicians make accurate diagnoses and provide automatic descriptive reports of medical images which saves physicians's time. Each group could participate in one or both subtasks. In this paper, we present the methods our group, IUST_NLPLAB, from the Iran University of Science and Technology[3], School of Computer Engineering [4], Natural Language Pocessing Laboratory[5] used in both subtasks. This is our first time participating in the ImageCLEF competition. We participated in both subtasks and registered ten successful submissions in the concept detection and caption prediction subtasks [2]. We were able to win first place in the caption prediction task with a margin of 16 BLEU points from the second group. Also, in the concept detection task, we were able to win the eighth place in the competition with a gap of about five percent in F1 measure from the first ranked group. In the following sections, we will describe the datasets used, models developed, and the results we achieved in detail.

2. Task description

This year the ImageCLEF evaluation campaign hosted the 6th edition of the medical image caption task. Unlike some of the previous editions which only contained the caption prediction task (e.g., 2016 [3]) or only the concept detection task (e.g., 2019 [4]), the 6th edition contained both subtasks as described below.

2.1. Concept Detection

The goal for this task is to train a model based on the training data provided for extraction of UMLS[6][5] Concept Unique Identifiers (CUIs) from medical images. This helps to better understand the medical concepts contained in medical images and can be used in other jobs such as caption generation. Table 1 lists the top 15 most frequent concepts in the training data.

The 2022 dataset includes 8374 medical concepts, which is a significant increase compared to 2021.

2.2. Caption Prediction

The goal of caption prediction is to train a model based on the training data provided to predict a suitable caption for medical images. It is essential for the model to correctly diagnose and

[3]http://www.iust.ac.ir/en (last accessed: 2022-05-27)
[4]http://ce-inter.iust.ac.ir/ (last accessed: 2022-05-27)
[5]https://nlplab.iust.ac.ir (last accessed: 2022-05-27)
[6]Unified Medical Language System®

Table 1
Most frequent concepts in the training data

UMLS CUI	UMLS Meaning	frequency
C0040405	X-Ray Computed Tomography	25989
C1306645	Plain x-ray	24389
C0024485	Magnetic Resonance Imaging	14622
C0041618	Ultrasonography	11147
C0817096	Chest	7720
C0002978	angiogram	6027
C0000726	Abdomen	5772
C0037303	Bone structure of cranium	5144
C0221198	Lesion	3845
C0205131	Axial	3187
C0030797	Pelvis	3176
C0023216	Lower Extremity	2739
C0238767	Bilateral	2722
C0577559	Mass of body structure	2341
C0205129	Sagittal	2012

extract sufficient information from medical images to be able to correctly predict the appropriate caption. This task is inherently more complex since it requires combining image processing and natural language processing techniques to generate captions for medical images.

3. Data

The dataset introduced for the ImageCLEFmedical Caption 2022 is a subset of the Radiology Objects in COntext (ROCO)[6] dataset. In this version of the dataset, imaging modality information is not mentioned. Also, as in previous versions, the dataset originates from biomedical articles of the PMC OpenAccess[7] subset. This dataset is used for both subtasks: Concept Detection and Caption Prediction. The published dataset consists of train, validation, and test images. Also, five Excel files were attached, including the names of concepts, concepts per train image, concepts per valid image, caption per train image, and caption per valid image. This dataset includes 83275 radiology images as training set, 7645 radiology images as validation set, and 7601 radiology images as test set. Figure 1 compares the data size presented in the last four years for the Medical Image Captioning task at the ImageCLEF[8, 9, 10] evaluation campaign. In 2021, the number of data has decreased significantly compared to previous years. This was due to only using radiology images described by medical experts[10].

Table 2 shows some training data examples with their corresponding concepts and captions.

3.1. Image Concepts

In this task, for each image, a number of concepts are defined by the Unified Medical Language System® (UMLS)[5] Concept Unique Identifiers (CUIs) are specified. The number of concepts is

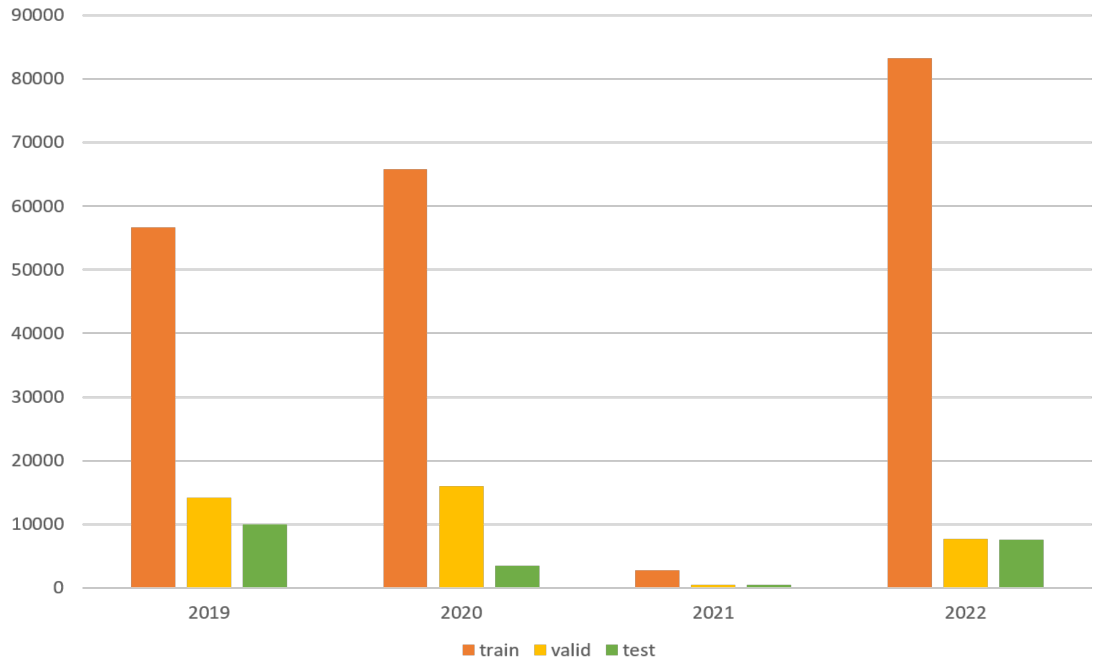


Figure 1: Comparison of ImageCLEFmedical Caption data in the last four years. As shown in the chart, the number of data in 2022 compared to 2021 has grown significantly.

different for each image. In training set, 3718 images have only one concept, while the maximum number of concepts for an image was 50. On average, five concepts are specified for each image.

3.2. Image Captions

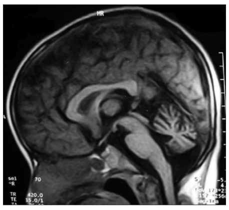
A caption is provided for each image in the training and validation sets in this task. The organizers mentioned that the captions are pre-processed in the following four steps:

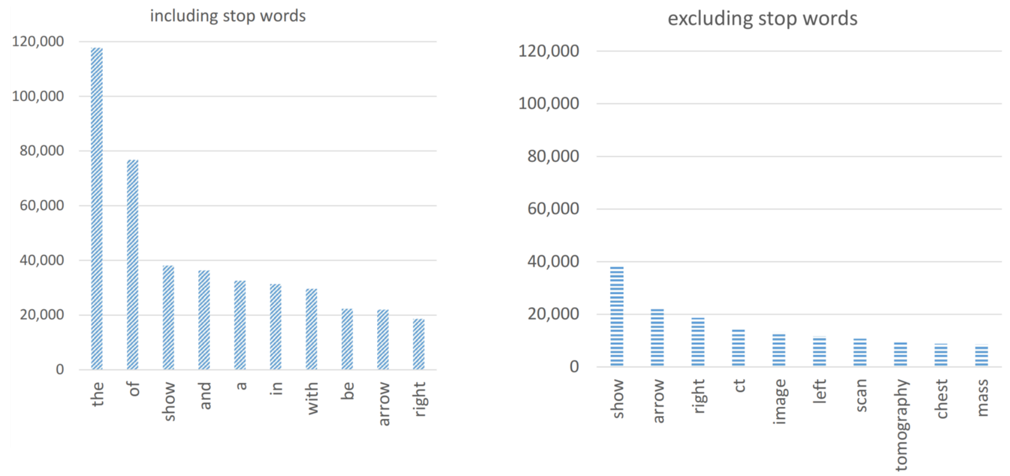
- Numbers and words containing numbers have been removed
- All punctuation was removed.
- Lemmatization was applied using spaCy.
- Captions were converted to lower-case.

The length of captions in the training set varies. According to surveys, in training set, 194 images have one-word captions, while the maximum caption length is 391. The most common caption length is ten words, of which 3771 images have captions of this length. The average length of a caption is 19 words. Figure 2 shows the most repetitive words in training set captions and their frequency with stop words, and without them. We also calculated the TTR[7] for this caption dataset. TTR is obtained by dividing the number of unique words by the size of the text and is a simple measure of lexical diversity[11]. Considering the stop words, the value of TTR in this dataset is 0.022, and without considering the stop words, it is 0.031.

[7]Type-token ratio

Table 2
Sample images from the training set along with their concepts and captions [12, 13, 14, 15].

Image	Concepts	Caption
 CC BY-NC-ND [Jo et al. (2018)]	• C1306645 (Plain x-ray) • C0817096 (Chest) • C0225759 (Lung field)	• chest radiograph show multiple tiny nodule white arrow in both lung field.
 CC BY [Seshachalam et al. (2010)]	• C0024485 (Magnetic Resonance Imaging) • C0006104 (Brain) • C0740279 (Atrophy of cerebellum)	• mri brain show cerebellar atrophy.
 CC BY [Pereda et al. (2013)]	• C1306645 (Plain x-ray) • C1140618 (Upper Extremity) • C0018563 (Hand) • C0205082 (Severe (severity modifier)) • C5194734 (Tubular bones) • C0041600 (Bone structure of ulna) • C1441672 (Observed) • C0025526 (Metacarpal bone) • C0699952 (Fused)	• hand of a patient with acrodysostosbe and multihormonal resbetance severe and generalized brachydactyly through very short and broad tubular bone include ulna can be observe metacarpals iiv be proximally pointed and coneshape proximal phalangeal epiphysbe be prematurely fuse the general appearance of the hand be bulky and stocky courtesy of prof dr jess argente.
 CC BY [Sudulagunta et al. (2015)]	• C0024485 (Magnetic Resonance Imaging) • C0037949 (Vertebral column) • C0522510 (With intensity) • C3853028 (Thoracic Cord) • C0054967 (CD6 antigen)	• mri spine show hyperintensity in the thoracic cord till level



(a) Ten most frequent words in the training set with considered stop words. Given the widespread use of stop words in texts, it is natural for stop words to have a main place in the chart. However, a few non-stop words like "show" have a significant number, which seems natural considering the use of this word to describe images.

(b) Ten most frequent words in the training set without considered stop words. By looking at words, it is clear that most of the words are widely used in describing images or corrections in medical.

Figure 2: Ten most frequent words in the training set

4. Methods

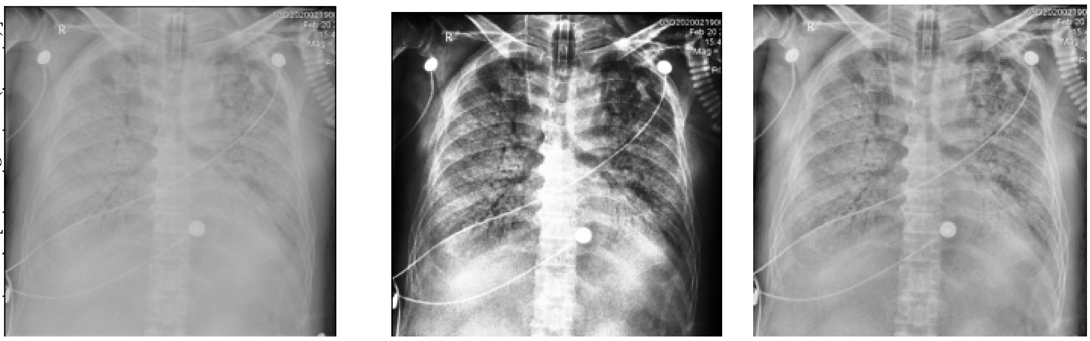
We first present image preprocessing techniques used for both subtasks, Next, we introduce models developed for the concept detection followed by caption prediction models.

4.1. Image Pre-processing

We used various techniques to improve the quality of medical images. Two of the most important are as follows.

- `Histogram Equalization:` Histogram Equalization is an image processing method that uses a contrast enhancement technique[16]. In this method, the image histogram is flattened as much as possible and the probability distribution is mapped to a uniform probability distribution. However, this is not the best way to improve image quality, and in some cases may not have a good output because the average brightness of the output image is significantly different from the input image.
- `Contrast Limited Adaptive Histogram Equalization (CLAHE):` CLAHE[17] is also a type of Histogram Equalization, in which contrast amplification is limited. In a typical Histogram Equalization, we see an increase in noise in near-constant regions. To

$CCBY[Xiangetal.(2020)]$



(a) Normal (b) Equalization histogram (c) CLAHE

Figure 3: The first column shows the normal image of the dataset. The second and third columns show the images after the Histogram Equalization and CLAHE technique applied on the normal image. As it is clear, the image quality has improved in some areas [18].

solve this problem and improve feature extraction, we use CLAHE. CLAHE equalizes the brightness and contrast of the images. This technique divides the image into sections and applies histogram equalization to each section. Then the contrast amplification limit, also known as clip limit, is applied. We use a clip limit of 2 in our models.

Cropping and flipping were also used for data augmentation. For models that do not use data augmentation techniques, the CLAHE method is used to improve image quality. Figure 3 shows the normal images of the dataset and images of histogram equalization and CLAHE operations on them.

4.2. Concept Detection

Concept detection is a classification problem. We examine the two main approaches we adopted to solve this problem.

4.2.1. Information Retrieval Approach

We studied and implemented the methods presented by Jacutprakart et al. [19]. Jacutprakart et al. received the second-best F1 score in the concept detection challenge in 2021. Their best approach was to extract image features from a CNN[8] and use k-NN[9] to extract the concepts. That is, for a test image, the closest k training images are found. The concepts of the closest images are then used to assign concepts to the test image. They got the best results by using cosine similarity to calculate the distance between two images, DenseNet121[20] to extract features, and setting k to 1. We attempted to implement this approach. However, because this year's dataset size is much larger than the previous year, we had trouble getting results from k-NN. We tried to train a similar model on a much smaller subset of the dataset, but the results were underwhelming. Using the output to extract labels, similar to a multilabel classification

[8] Convolutional Neural Network
[9] k-Nearest-Neighbor

task, produced significantly better results. Thus we stopped following this approach and moved on to trying a 1-NN ensemble.

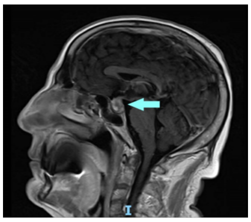
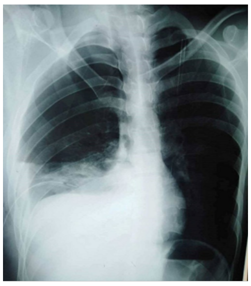
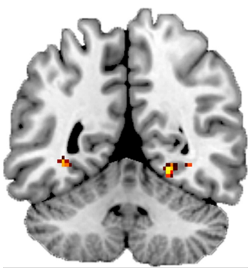
In the 1-NN ensemble, we used a retrieval approach based on the AUEB NLP group model in 2021[21] that achieved first place in the concept detection subtask. At first, we employed three kinds of different CNN encoders including a ResNet-50[22], a DenseNet-201[20], and an EfficientNet-B0[23] that made up our primary model. All encoders were pre-trained on ImageNet[24]. These encoders were fine-tuned on the training set for five epochs and the best weights were saved according to loss values on the training set. In the next step, we trained five models for each encoder as part of the validation set and we got 15 encoders. Each of these encoders were trained for three epochs. After working on encoders, we used trained encoders to get image embedding of training examples. Image embeddings were retrieved from the last average pooling layer of each encoder. To detect concepts in test images, we also find image embeddings of test images. After computing similarity between train embeddings and test embeddings, we find the most similar training image to the test image. In the end, 15 training images will be chosen and each image corresponds to an encoder. To assign concepts to test images, we used a majority-voting mechanism. In this mechanism, among the 15 chosen images, concepts that appear more than N times will be assigned to the test image. After trying different values for N and evaluating results with accuracy and F1 score, results showed the best value for N to be 8. Adding data augmentation to this model also improved its performance. Finally, after computing similarity between image embeddings with different methods, cosine similarity showed better results than other methods. In Table 3 we show a number of validation set images with the ground truth and predicted concepts along with their F1 score calculated by this approach.

4.2.2. Multi-Label Classification Method

In this method, we used image concepts as labels and built a multi-label classification model. We used CNNs with ImageNet[24] pre-trained weights, removed their last layer, and added a classification layer. The output layer has 8374 units and uses sigmoid as the activation function. The final model was then fine-tuned on the target dataset. We experimented with different pre-trained models and different configurations. Models were compiled with Adam[25] as the optimizer. Results on validation and test set were generated after every five epochs. We tried different thresholds on the output layer's activation function to classify the image concepts and used their F1 score to evaluate and find the best one. Table 4 shows the details of the implemented MLC[10] methods. Figure 4 also shows the architecture of the MLC-based concept detection model.

[10]Multi-Label Classification

Table 3
Example of concept detection with ground truth and predicted concepts with F1 scores of them [26, 27, 28, 29].

Image	Ground Truth	Prediction	F1 score
 CC BY [Rai et al. (2020)]	• C0024485 (Magnetic Resonance Imaging) • C0346308 (Pituitary macroadenoma) • C0205129 (Sagittal)	• C0024485 (Magnetic Resonance Imaging) • C0346308 (Pituitary macroadenoma) • C0205129 (Sagittal)	• 1.0
 CC BY [Davar et al. (2020)]	• C0041618 (Ultrasonography) • C0016823 (Structure of fundus of eye)	• C0041618 (Ultrasonography) • C0016823 (Structure of fundus of eye) • C0439828 (Variable (uniformity)) • C0205396 (Identified) • C0013938 (Embryo transfer (procedure))	• 0.571
 CC BY [Jazia et al. (2021)]	• C1306645 (Plain x-ray) • C0442808 (Increasing) • C0032227 (Pleural effusion disorder) • C0008034 (Chest Tubes)	• C1306645 (Plain x-ray) • C0817096 (Chest) • C0032326 (Pneumothorax)	• 0.285
 CC BY [van Kesteren et al. (2020)]	• C1707489 (Connectivity) • C0205202 (Corrected)	• C0221198 (Lesion) • C0024485 (Magnetic Resonance Imaging)	• 0.0

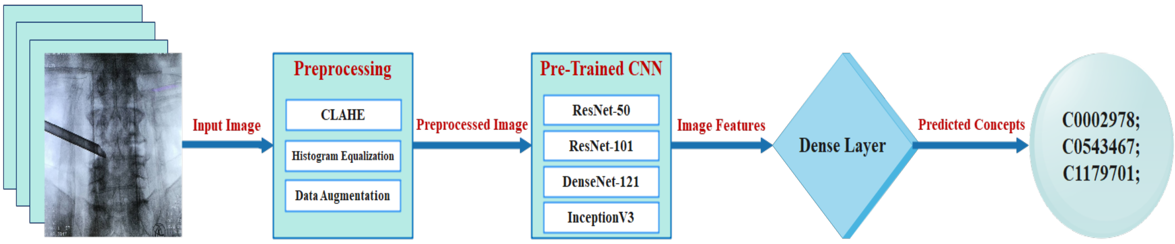


Figure 4: Architecture of the MLC-based concept detection model

Table 4
Description of our concept detection models

Name	Base model	Regularization	Learning rate	Data augmentation
v2.1	Resnet50	None	0.001	None
v2.2	Resnet50	None	0.001	CLAHE, equalizeHist, hflip, original
v2.3	Resnet101	None	0.001	None
v2.4	Resnet50	Dropout(0.5)	0.001	None
v2.5	DenseNet121	None	0.001	None
v3.1	InceptionV3	None	0.009	None
v3.2	InceptionV3	None	0.009	None
v3.3	InceptionV3	None	0.009	CLAHE, equalizeHist, random crop
v3.4	InceptionV3	L2	0.009	equalizeHist, random crop
v3.5	InceptionV3	Dropout (0.5)	0.009	CLAHE, random crop

4.3. Caption Prediction

For the caption prediction subtask, we studied the method implemented in [30], which achieved first place in last year's caption prediction challenge. This team's best approach was to use a multi-label classification model. In this approach, each word is considered to be a label. A classification model is trained to predict the words that will later create a caption for the given image.

We used a CNN pre-trained on ImageNet[24] and fine-tuned it on the subtask training set to extract image features, similar to the multi-label classification method used in the concept detection subtask. For fine-tuning, the last layer of the CNN was removed. A dropout layer, an activation layer, and a dense layer were added. We tried different CNN models and different configurations.

To generate a caption for an image, the model will predict its corresponding words. Probability of each word is calculated in the output layer using sigmoid activation function. Then, the top N words with the highest probability are chosen. N is a hyper-parameter that will define the length of captions. Different values of N in the range of 15 to 27 were tested on the validation set, and the best N for each model was chosen using the BLEU score [31]. Two methods were used to turn the generated words into full captions:

1. Words are ordered from highest to lowest probability.
2. Words are ordered based on their statistics in the training set. Each word is assigned to its most common position in the caption.

Overall, we focused on predicting the correct words rather than finding their correct order. These two methods were not able to find the right order of words. That is why the final prediction may not be grammatically correct. But this approach was able to predict words well, which led to a high BLEU score.

Table 5 shows the details of the implemented classification models for the caption generation subtask. Figure 5 also shows the architecture of the MLC-based caption prediction models.

Table 6 shows some examples of validation images with ground truth caption and predicted caption by our best submission. Their BLEU and ROUGE scores are also mentioned in the table. Because we used stemming while creating our vocabulary, some words, like "image", are stemmed in the final caption.

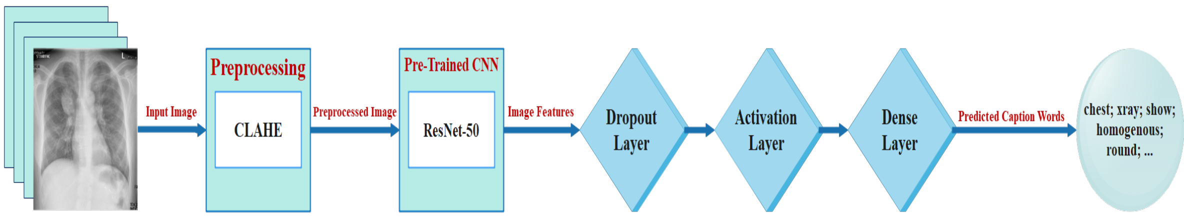
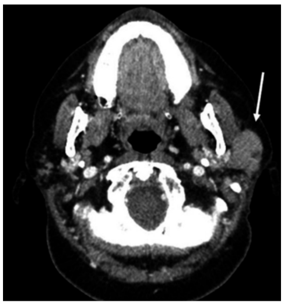
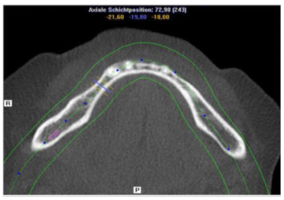
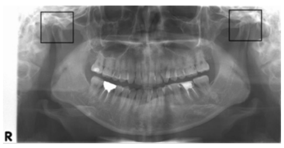


Figure 5: Architecture of the MLC-based caption prediction model

Table 5
Description of our caption prediction models

Name	CNN	Data Augmentation	Activation	Freeze CNN	Learning rate
v1.1	ResNet50	None	PReLU	No	5e-4
v1.2	ResNet50	CLAHE	PReLU	No	5e-4
v1.3	ResNet50	None	PReLU	Yes	5e-4
v1.4	ResNet50	None	ReLU	No	5e-4
v1.5	ResNet50	None	PReLU	Yes	5e-4, decay every 2500 steps

Table 6
Example of caption prediction with different scores. In first row model had a good score but in second row it could not predict well [32, 33, 34, 35].

Image	Ground Truth	Prediction	BLEU	ROUGE
 CC BY [Povlow et al. (2021)]	axial ct image of the neck with intravenous contrast at the level of the parotid gland show asymmetric left parotid gland enlargement with replacement by a soft tissue mass white arrow	arrow show ct axial imag tomographi enhanc scan comput left right mass muscl lesion enlarg gland red neck view contrast tissu soft demonstr white nerv tumor	0.806	0.468
 CC BY [Wimmer et al. (2021)]	horizontal section show bony deficit at implant site	axial show imag ct comput scan right tomographi lesion left patient arrow bone view cortic measur treatment fractur month later head area cbct plate section margin	0.526	0.121
 CC BY [Trisnawati et al. (2020)]	chest xray show bilateral pneumonia	chest xray show left patient leav right arrow lung pleural effus hemithorax admiss radiograph lobe mediastin mass tip day mediastinum postop elev upper enlarg tube imag	0.140	0.193
 CC BY [Choi et al. (2021)]	result of roi extraction pixel	arrow show right imag panoram left maxillari radiograph lesion coron bone patient bilater impact mandibular molar later view side white cortic case area fractur sinus first	0.0	0.0

5. Results

In this part, we review the results of the models implemented in the two subtasks of concept detection and caption prediction. In the concept detection subtask, F1 score was used to evaluate the models and the ranking was based on this metric. Also reported is the Secondary F1 score, which is calculated using only a subset of manual validated concepts. In the caption prediction subtask, BLEU[31], ROUGE[36], METEOR[37], CIDR[38], SPICE[39] and BERTScore[40] were used to evaluate the models. The ranking was based on BLEU score. ROUGE scores were also reported during the competition. Other metrics were reported after the challenge.

Before presenting the results of our implemented models, we review the results obtained in the last six years in this task. Table 7 shows information about the size of the dataset, the number of concepts, and the results of the first three groups in each subtask[2, 10, 9, 8, 41, 42, 43, 44]. Note that the purpose of presenting this table is to express statistical information on this task in the last few years. Due to the differences in the data sets of different years, it is not correct to compare their results with each other.

Table 7
This table shows information about the datasets and the results obtained in ImageCLEFmedical Caption in the last six years. In the data set section, the number of training data, validation and test is specified. The number of concepts in the dataset per year is also shown. In the section on concept detection and caption prediction, the results of the top three groups are mentioned. As mentioned in the text, the purpose of presenting this table is to show the statistical information of the imageCLEFmedical caption task in recent years. The datasets of the years denoted by * are different, so comparing the results of the years with each other does not provide accurate information[2, 10, 9, 8, 41, 42, 43, 44].

Year	Dataset				Concept Detection (F1)			Caption Prediction (BLEU)		
	train	**valid**	**test**	**concepts**	**1st**	**2nd**	**3rd**	**1st**	**2nd**	**3rd**
2022*	83275	7645	7601	8374	0.451	0.450	0.447	0.482	0.322	0.311
2021	2756	500	444	1586	0.505	0.468	0.419	0.509	0.461	0.431
2020*	65753	15970	3534	3047	0.394	0.392	0.380	-	-	-
2019*	56629	14157	10000	5528	0.282	0.265	0.223	-	-	-
2018	222305	-	10000	111156	0.110	0.009	0.050	0.250	0.179	0.172
2017	164614	10000	10000	20464	0.171	0.164	0.143	0.563	0.321	0.260

5.1. Concept Detection

We chose our best models based on F1 score on the validation set and results with the highest score were submitted. One of our submissions used information retrieval method (submission v1) and other nine submissions followed multi-label classification approach. Table 4 has information about different MLC models. Table 8 shows our scores on the test set and the number of epochs and threshold for each model submission. The best result had 0.398 as F1 score. It also achieved 0.673 for secondary F1. Secondary F1 score was calculated using a subset of manually validated concepts (anatomy and image modality) only. This submission ranked 8 among all submitted group results. This system used ResNet50 as its base model with dropout[45] and no data augmentation. Information retrieval model had close scores to best MLC methods but at last

Table 8
IUST_NLPLAB concept detection submissions details and test results. ES stands for "Extra submission". These submissions were sent after the competition's deadline.

Run ID	Name	Epochs	Threshold	F1 score	Secondary score
181667	v1	-	-	0.394	0.750
181948	v2.1	16	0.1	0.281	0.355
182279	v2.2	16	0.1	0.252	0.352
182280	v2.3	12	0.1	0.255	0.352
182291	v2.4	48	0.1	0.387	0.611
182292	v2.5	4	0.12	0.242	0.332
182293	v2.5	8	0.1	0.244	0.318
182302	v2.3	48	0.1	0.243	0.305
182304	v2.4	48	0.12	0.394	0.656
182307	v2.4	48	0.13	0.398	0.673
ES1	v2.4	60	0.4	0.348	0.730
ES2	v2.4	96	0.25	0.411	0.785
ES3	v3.1	20	0.3	0.240	0.356
ES4	v3.2	20	0.4	0397	0.668
ES5	v3.3	40	0.4	0.385	0.623
ES6	v3.4	40	0.4	0.302	0.634
ES7	v3.5	20	0.4	0.419	0.721

MLCs ranked higher. It's interesting that information retrieval earned higher secondary F1 than MLC models.

Although our information retrieval model had good results among our different submissions, but we submitted just one model from this type. It happened because the information retrieval model was more complex than our MLC models, so it needed more time and resources to train. For example, MLC models approximately required three hours to train on *server 3*, but the information retrieval model took seven days to train on *server 2*. We tried to improve this model with different methods, but we were unable to get the final results due to time and resource constraints. Thus we decided to train simpler models with fewer parameters in both subtasks. This enabled us to have a faster turn-around time and iterate on more model improvement ideas.

After trying different models and configurations, we focused on different settings of v2.4 before the deadline, which produced the best result. The last submissions of v2 were from this particular version.

The submission limit allows us to submit only 10 runs but we still had some results that were not evaluated. After the submission deadline we asked ImageCLEF organizers to evaluate few more models for us which they generously accepted. These extra models showed improvement in concept detection results both in F1 score and secondary F1. Submission ES7, which used InceptionV3[46] as its base model and dropout for regularization, achieved 0.419 F1 score, which was the best score for us. This system also used data augmentation techniques including CLAHE and random crop. The best secondary F1 belongs to submission ES2. This system is like our best model in this competition but it trained for 96 epochs and set its threshold to 0.25.

Table 9
Caption prediction submissions' details

Run ID	Name	Epochs	N	Sorting method
181670	v1.1	20	20	1
181951	v1.2	10	27	1
182249	v1.1	10	26	1
182250	v1.3	10	26	1
182275	v1.4	5	26	1
182290	v1.5	10	25	2
182314	v1.5	10	17	1
182315	v1.4	5	26	2
182319	v1.1	15	26	1
182327	v1.5	15	15	1

5.2. Caption Prediction

Our MLC approach could predict captions well, as all our ten submissions had BLEU scores higher than 0.430 and ranked from 1 to 10 on the released leaderboard. Using different settings, like using a different activation function, could improve the BLEU score.

Overall, choosing the parameter N, which determined the length of predicted captions, was a trade-off between the BLEU score and the ROUGE score. A higher N resulted in a higher BLEU score and lower ROUGE score, and vice versa. As the primary score in this competition was BLEU, we focused on using a setting that would give us the highest BLEU score.

As we mentioned in the Methods sections, we used two different methods to sort the predicted words. In the first method, words were sorted from highest to lowest probability. For the second method, we studied the training set and noted the positions each word appears in. Then, we tried sorting the generated words using this data, but it did not improve the scores. The first method had better results. So, we focused on using this method in most submissions.

The best result was achieved by setting N to 26 and using ReLU as the activation function. In this submission, words were ordered from highest to lowest probability. Table 5 shows the details of the submitted runs and Table 10 shows our submission scores on the test set.

6. AIOps

Running a large number of experiments in a short amount of time with limited resources requires meticulous planning and operations. Given our limitations in time and resources, we believe our operations' strategy played a significant role in our success. This section elaborates on what worked and did not work for training models faster with fewer resources, consisting of (GPU, CPU, RAM, and Disk Space).

6.1. Memory Optimization

The first challenge our team faced was running out of memory. One of the bottlenecks in Artificial Intelligence projects is large datasets, which do not fit into memory. Data Generators

Table 10
Caption prediction submissions' test results

Run ID	BLEU	ROUGE	METEOR	CIDEr	SPICE	BERTScore
181670	0.457	0.140	0.082	0.045	0.013	0.570
181951	0.474	0.138	0.092	0.026	0.006	0.554
182249	0.480	0.138	0.090	0.027	0.005	0.553
182250	0.482	0.139	0.089	0.026	0.005	0.557
182275	0.483	0.142	0.092	0.030	0.007	0.561
182290	0.481	0.142	0.091	0.031	0.014	0.567
182314	0.462	0.158	0.085	0.062	0.010	0.578
182315	0.480	0.142	0.093	0.030	0.013	0.570
182319	0.469	0.136	0.089	0.029	0.006	0.551
182327	0.440	0.162	0.083	0.071	0.013	0.574

can help; They will generate values lazy (on demand). It is not efficient or sometimes feasible to load all the data into memory at once. Another benefit is that our Model does not have to wait until all the data is processed before using them.

Generators save the internal state without holding the entire data in memory. When the new data is requested, they continue from their previous saved state by providing the next batch of data, which are tiny portions of a larger dataset, to the requestor.

There are multiple ways to achieve this. We have used Sequence[47] from Keras API[48] because it is safer in multiprocessor environments. According to Keras documentation[47]: "This structure guarantees that the network will only train once on each sample per epoch, which is not the case with generators."

There are some pitfalls when using a generator. For example, using a mutable global variable in a data generator, which is called multiple times, can result in unexpected behavior, and some anti-patterns can invert the purpose of generators and cause incremental memory growth.

6.2. Faster Execution

To accomplish faster code execution, first, we need to determine why our code is time-consuming and which sections have the most significant impact on it. We used the TensorBoard Profiling tool[49] to analyze our model.

Profiling is the study of hardware resource consumption based on information gathered during the program's execution to identify which part of our program needs optimization and how we can speed up the overall program while minimizing resources.

After running the Profiler, TensorBoard[49] will offer us a visual representation of the gathered information, which consists of:

1. Recommendations for next steps in model improvement. These suggestions range from determining whether our Model is input-bound, how much time is spent on Kernel Launch, and what percentage of the operations performed are 16 or 32-bit.
2. to figure out which GPU operations take the longest, TensorFlow Stats are used, and then We should improve the most time-consuming parts to notice substantial changes in

execution time.

To achieve higher throughput and better GPU utilization, we can raise the batch size sufficiently without exhausting the resources and running Out of Memory (OOM). To prevent the Model's accuracy from decreasing, we should scale the Model by tuning hyperparameters.

Parallel execution and multi-threading can also be used, depending on whether the tasks are CPU-Bound or I/O-Bound.

6.2.1. I/O-Bound

1. Pre-fetching and caching the data at the cost of higher memory usage will enhance the Model's throughput. The input pipeline prepares the data for the next phase before the data is requested.
2. For I/O operations, multi-threading is highly recommended. It involves adding a new thread to an existing process, and memory is shared among them. Because of the shared memory, we need to use locks to control access to the shared data and prevent race conditions.

6.2.2. CPU-Bound

Multiprocessing is used for intensive CPU-limited tasks to achieve full CPU utilization. Each process has its own address space, and it is only applicable when we have multiple CPU cores. Inter-process communication (IPC) with Pipes and Queues is also possible. Since multiprocessing comes with full CPU utilization, it is ideal for the jobs with the least amount of data but the most operations.

"The more processes or threads we have, the faster it is." Another vital observation to note is that this sentence is not entirely accurate. This is because OS has to manage all these processes, and when there are too many, it may face scheduling overhead and reduce its overall speed.

6.3. Disk Usage Best Practices

Models will be trained regularly, and for having reproducible projects, we should version our Data and Models so that they can be easily shared, compared, and repeatedly reconstructed in our experiments. These tasks will be more manageable using the Version Control System.

- When choosing VC[11], it should support both on-premises and remote Cloud Storage Services (Azure, S3, GC)
- When facing a lack of storage, try to use Symlinks Instead of duplicate files, we can compress files that currently are not needed to free up some space and keep the files simultaneously.

[11]Version Control

6.4. Hardware

Most of our development was done on a system with GTX 1080 Ti GPU and another system with a RTX 2060 GPU provided by the computer engineering department. We also had limited access to an A100 GPU for final runs provided by the Simorgh Cloud.

Table 11
Information about the hardware used.

Server ID	GPU			CPU	RAM	Disk	Duration
	Model	Memory Size	Memory Type	VCores	Memory Size	Disk Space	Days
ID : 1							
	GeForce RTX 2060 Super	8 GB	GDDR6	6	24 GB	200 GB	10
ID : 2							
	GeForce GTX 1080 Ti	11 GB	GDDR5X	6	32 GB	200 GB	20
ID : 3							
	A100	40 GB	HBM2e	8	96 GB	200 GB	7

7. Conclusion

This paper describes the participation of IUST_NLPLAB at Iran University of Science and Technology at ImageCLEF caption 2022 task. In the Concept Detection subtask, we ranked 8 among 11 participating teams. We used MLC and information retrieval approaches in this subtask. Our MLC methods with adding dropout had better overall score. In the Caption Prediction subtask, all of our 10 submissions ranked higher than other groups' submissions and we achieved first rank in this subtask. We performed multi-label classification in this subtask as well. We used a classification model to predict the constructing words of a caption. Then, we used two different methods to create a caption.

We hope to be able to participate in this competition in the future and achieve better results.

Acknowledgments

This work has been supported by the Simorgh Supercomputer - Amirkabir University of Technology under Contract No ISI-DCE-DOD-Cloud-900808-1700. We also thank the support of the School of Computer Engineering of Iran University of Science and Technology and Iran's National Elites Foundation[12] for participating in this competition.

References

[1] B. Ionescu, H. Müller, R. Peteri, J. Rückert, A. Ben Abacha, A. G. S. de Herrera, C. M. Friedrich, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, S. Kozlovski, Y. D. Cid, V. Kovalev, L.-D. Ştefan, M. G. Constantin, M. Dogariu, A. Popescu, J. Deshayes-Chossart,

[12] https://en.bmn.ir/ (last accessed: 2022-05-27)

H. Schindler, J. Chamberlain, A. Campello, A. Clark, Overview of the ImageCLEF 2022: Multimedia retrieval in medical, social media and nature applications, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the 13th International Conference of the CLEF Association (CLEF 2022), LNCS Lecture Notes in Computer Science, Springer, Bologna, Italy, 2022.

[2] J. Rückert, A. Ben Abacha, A. García Seco de Herrera, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, H. Müller, C. M. Friedrich, Overview of ImageCLEFmedical 2022 – caption prediction and concept detection, in: CLEF2022 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Bologna, Italy, 2022.

[3] A. García Seco de Herrera, R. Schaer, S. Bromuri, H. Müller, Overview of the ImageCLEF 2016 medical task, in: Working Notes of CLEF 2016 (Cross Language Evaluation Forum), 2016.

[4] B. Ionescu, H. Müller, R. Péteri, Y. D. Cid, V. Liauchuk, V. Kovalev, D. Klimuk, A. Tarasau, A. B. Abacha, S. A. Hasan, V. Datla, J. Liu, D. Demner-Fushman, D.-T. Dang-Nguyen, L. Piras, M. Riegler, M.-T. Tran, M. Lux, C. Gurrin, O. Pelka, C. M. Friedrich, A. G. S. de Herrera, N. Garcia, E. Kavallieratou, C. R. del Blanco, C. C. Rodríguez, N. Vasillopoulos, K. Karampidis, J. Chamberlain, A. Clark, A. Campello, ImageCLEF 2019: Multimedia retrieval in medicine, lifelogging, security and nature, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the Tenth International Conference of the CLEF Association (CLEF 2019), LNCS Lecture Notes in Computer Science, Springer, Lugano, Switzerland, 2019.

[5] O. Bodenreider, The unified medical language system (umls): integrating biomedical terminology, Nucleic acids research 32 (2004) D267–D270.

[6] O. Pelka, S. Koitka, J. Rückert, F. Nensa, C. M. Friedrich, Radiology objects in context (ROCO): a multimodal image dataset, in: Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis, Springer, 2018, pp. 180–189.

[7] R. J. Roberts, Pubmed central: The genbank of the published literature, 2001.

[8] O. Pelka, C. M. Friedrich, A. Seco De Herrera, H. Müller, Overview of the ImageCLEFmed 2019 concept detection task, in: CEUR Workshop Proceedings, volume 2380, CEUR Workshop Proceedings, 2019.

[9] O. Pelka, C. M. Friedrich, A. García Seco de Herrera, H. Müller, Overview of the ImageCLEFmed 2020 concept prediction task, in: Proceedings of the CLEF 2020-Conference and labs of the evaluation forum, CONFERENCE, 22-25 September 2020, 2020.

[10] O. Pelka, A. B. Abacha, A. G. S. 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.

[11] K. Kettunen, Can type-token ratio be used to show morphological complexity of languages?, Journal of Quantitative Linguistics 21 (2014) 223–245.

[12] H. Jo, J. Baek, Case of pulmonary benign metastasizing leiomyoma from synchronous uterine leiomyoma in a postmenopausal woman, Gynecologic oncology reports 26 (2018) 33–36.

[13] A. Seshachalam, S. Cyriac, N. Reddy, S. T. Gnana, Ataxia telangiectasia: Family management, Indian Journal of Human Genetics 16 (2010) 39.

[14] A. Pereda, I. Garin, M. Garcia-Barcina, B. Gener, E. Beristain, A. M. Ibañez, G. Perez de Nanclares, Brachydactyly e: isolated or as a feature of a syndrome, Orphanet journal of rare diseases 8 (2013) 1–14.
[15] S. R. Sudulagunta, M. B. Sodalagunta, H. Khorram, M. Sepehrar, J. Gonivada, Z. Noroozpour, N. Prasad, Autoimmune thyroiditis associated with neuromyelitis optica (nmo), GMS German Medical Science 13 (2015).
[16] O. Patel, Y. P. Maravi, S. Sharma, A comparative study of histogram equalization based image enhancement techniques for brightness preservation and contrast enhancement, arXiv preprint arXiv:1311.4033 (2013).
[17] K. Zuiderveld, Contrast limited adaptive histogram equalization, Graphics gems (1994) 474–485.
[18] C. Xiang, L. Huang, L. Xia, Mobile chest x-ray manifestations of 54 deceased patients with coronavirus disease 2019: Retrospective study, Medicine 99 (2020).
[19] J. Jacutprakart, F. P. Andrade, R. Cuan, A. A. Compean, G. Papanastasiou, A. G. S. de Herrera, Nlip-essex-itesm at imageclefcaption 2021 task: deep learning-based information retrieval and multi-label classification towards improving medical image understanding, in: CLEF2021 Working Notes, CEUR Workshop Proceedings, CEUR-WS. org, 2021.
[20] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4700–4708.
[21] F. Charalampakos, V. Karatzas, V. Kougia, J. Pavlopoulos, I. Androutsopoulos, Aueb nlp group at imageclefmed caption tasks 2021, in: CLEF2021 Working Notes, CEUR Workshop Proceedings, CEUR-WS. org, Bucharest, Romania, 2021.
[22] 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.
[23] M. Tan, Q. Le, Efficientnet: Rethinking model scaling for convolutional neural networks, in: International conference on machine learning, PMLR, 2019, pp. 6105–6114.
[24] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE conference on computer vision and pattern recognition, Ieee, 2009, pp. 248–255.
[25] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).
[26] H. K. Rai, G. John, M. Anton, Atypical presentation of panhypopituitarism, Cureus 12 (2020).
[27] R. Davar, S. M. Poormoosavi, F. Mohseni, S. Janati, Effect of embryo transfer depth on ivf/icsi outcomes: A randomized clinical trial, International Journal of Reproductive BioMedicine 18 (2020) 723.
[28] R. B. Jazia, J. Ayachi, F. Chatbouri, A. Kacem, A. Faidi, D. B. Braiek, A. Maatallah, Unusual case of spontaneous hemopneumothorax in a tunisian pulmonology department: a case report, The Pan African Medical Journal 38 (2021).
[29] M. T. van Kesteren, P. Rignanese, P. G. Gianferrara, L. Krabbendam, M. Meeter, Congruency and reactivation aid memory integration through reinstatement of prior knowledge, Scientific Reports 10 (2020) 1–13.

[30] V. Castro, P. Pino, D. Parra, H. Lobel, Puc chile team at caption prediction: Resnet visual encoding and caption classification with parametric relu, in: CLEF2021 Working Notes, CEUR Workshop Proceedings, CEUR-WS. org, Bucharest, Romania, 2021.

[31] 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.

[32] M. R. Povlow, M. Streiff, S. Madireddi, C. Jaramillo, A primary parotid mucosa-associated lymphoid tissue non-hodgkin lymphoma in a patient with sjogren syndrome, Cureus 13 (2021).

[33] I. Trisnawati, R. El Khair, D. A. Puspitarani, A. R. Fauzi, et al., Prolonged nucleic acid conversion and false-negative rt-pcr results in patients with covid-19: A case series, Annals of Medicine and Surgery 59 (2020) 224–228.

[34] L. Wimmer, P. Petrakakis, K. El-Mahdy, S. Herrmann, D. Nolte, Implant-prosthetic rehabilitation of patients with severe horizontal bone deficit on mini-implants with two-piece design—retrospective analysis after a mean follow-up of 5 years, International Journal of Implant Dentistry 7 (2021) 1–14.

[35] E. Choi, D. Kim, J.-Y. Lee, H.-K. Park, Artificial intelligence in detecting temporomandibular joint osteoarthritis on orthopantomogram, Scientific Reports 11 (2021) 1–7.

[36] C.-Y. Lin, Rouge: A package for automatic evaluation of summaries, in: Text summarization branches out, 2004, pp. 74–81.

[37] M. Denkowski, A. Lavie, Meteor universal: Language specific translation evaluation for any target language, in: Proceedings of the ninth workshop on statistical machine translation, 2014, pp. 376–380.

[38] R. Vedantam, C. Lawrence Zitnick, D. Parikh, Cider: Consensus-based image description evaluation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 4566–4575.

[39] P. Anderson, B. Fernando, M. Johnson, S. Gould, Spice: Semantic propositional image caption evaluation, in: European conference on computer vision, Springer, 2016, pp. 382–398.

[40] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with bert, arXiv preprint arXiv:1904.09675 (2019).

[41] A. García Seco de Herrera, C. Eickhof, V. Andrearczyk, H. Müller, Overview of the ImageCLEF 2018 caption prediction tasks, in: Working Notes of CLEF 2018-Conference and Labs of the Evaluation Forum (CLEF 2018), Avignon, France, September 10-14, 2018., volume 2125, CEUR Workshop Proceedings, 2018.

[42] C. Eickhoff, I. Schwall, A. Garcia Seco De Herrera, H. Müller, Overview of ImageCLEFcaption 2017–image caption prediction and concept detection for biomedical images, CLEF 2017 working Notes 1866 (2017).

[43] K. Dimitris, K. Ergina, Concept detection on medical images using deep residual learning network, Working Notes CLEF (2017).

[44] Y. Zhang, X. Wang, Z. Guo, J. Li, Imagesem at imageclef 2018 caption task: Image retrieval and transfer learning, in: CLEF CEUR Workshop, Avignon, France, 2018.

[45] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, The journal of machine learning research

15 (2014) 1929–1958.
[46] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2818–2826.
[47] Tensorflow, Sequence class from keras api, https://www.tensorflow.org/api_docs/python/tf/keras/utils/Sequence, 2022. Last Accessed: 2022-05-27.
[48] K. API, Keras api documentation, https://keras.io/, 2022. Last Accessed: 2022-05-27.
[49] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, X. Zheng, TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL: https://www.tensorflow.org/, last Accessed: 2022-05-27, Software available from tensorflow.org.