SSN MLRG at ImageCLEFmedical Caption 2022: Medical Concept Detection and Caption Prediction using Transfer Learning and Transformer based Learning Approaches

Sheerin Sitara Noor Mohamed[a] and Kavitha Srinivasan[b]

[a, b] *Department of CSE, Sri Sivasubramaniya Nadar College of Engineering, Kalavakkam – 603110, India*

Abstract

The computer aided medical system for various applications is required now-a-days for an early and effective analysis. However most of the medical data are, publicly unavailable and exist in unstructured and unlabelled format are real challenges in developing the medical system. To address these issues, ImageCLEF forum is conducting many tasks on the medical domain from 2016 onwards. This year one of the tasks is medical concept detection and caption prediction. For this task, our team has proposed two concept detection techniques and caption prediction techniques. The concept detection models are developed using multi-label classification and information retrieval approaches resulted the F1-score and secondary F1-score as 0.418 and 0.654 respectively. The caption prediction models are implemented using ResNet with Bidirectional Encoder Representations from Transformers (BERT) and, Sparse Auto Encoder (SAE) with Multi-Layer Perceptron (MLP) and Gated Recurrent Unit (GRU), which resulted a BLEU and BERT score of 0.160 and 0.545 respectively.

Keywords 1

Concept Detection, Caption Prediction, Multi-label Classification, Information Retrieval, BERT, SAE, MLP, GRU, ResNet, DenseNet

1. Introduction

The advancement in medical domain is trying to improvise the quality and quantity of medical data. These medical data are available in different formats such as medical images, clinical reports and doctor transcriptions. Among these, medical images play a vital role in diagnosis. The different types of medical images are X-Ray, Computed Tomography (CT), Magnetic Resonance Imaging (MRI), f-MRI, mammogram, ultrasonography, Positron Emission Tomography (PET), Single Photon Emission Computerized Tomography (SPECT) and thermography [1]. The ImageCLEF forum is conducting various tasks related to medical images such as caption prediction, concept detection, Tuberculosis (TB) type detection, Multi-Drug Resistant (MDR) detection, TB severity score calculation, CT report generation and Visual Question Answering (VQA) from 2016 onwards. In this, we have participated in concept detection and caption prediction tasks during the current year.

In concept detection and caption prediction tasks [2], the dataset given by ImageCLEF consists of different modalities such as CT, XR, PET, angiogram and ultrasound images. The concepts and captions corresponds to these images are created by medical annotator from PubMED articles and Unified Medical Language System (UMLS) terms. Finally, these datasets are validated and verified by medical domain experts. These datasets are used to develop concept detection and caption prediction model using suitable techniques and evaluated performance metrics, are discussed in the following paragraphs.

The concept detection approaches are: multi-label classification approach using DenseNet [3], information retrieval approach using DenseNet [4] and EfficientNet [5], ensemble of classifier based

[1]CLEF 2022 – Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy
EMAIL: sheerinsitaran@ssn.edu.in (A. 1); kavithas@ssn.edu.in (A. 2)
ORCID: 0000-0003-1752-2107 (A. 1); 0000-0003-3439-2383 (A. 2)


on DenseNet and Feed-Forward Neural Network (FFNN) [6] and, Learned Perceptual Image Path Similarity (LPIPS) [7] based on VGGNet. Among these approaches, multi-label classification and information retrieval approaches along with DenseNet are used for this task execution. The multi-label classification approach (model 1) is chosen because it has an ability to find conditional dependencies between the labels and the independence between the labels are computed based on the specific condition. The information retrieval approach (model 2) is adapted because it effectively identifies the keyword in the query and retrieves the required answer from the dataset.

The caption prediction techniques are grouped into, (i). Techniques which performs both image and test processing like Visual Transformers [8], (ii). Combination of techniques for image processing and text processing. The image processing can be performed by pre-trained models like DenseNet [9], ResNet [10], VGGNet [11] or autoencoder like Sparse Auto Encoder (SAE) [12] and, the text processing can be performed by attention based encoder-decoder [13], Long Short Term Memory (LSTM) [14], Gated Recurrent Unit (GRU) [15] or, transformer-based architectures like BERT [16], GPT-2 [17]. Among the techniques, ResNet with BERT (model 3) and, SAE with MLP and GRU (model 4) are chosen for this task execution because, (i). The deep architecture and millions of trainable parameters in ResNet reduced the degradation problem and made it more suitable for image captioning problem. (ii). BERT predicts the context of the words from both left to right context and right to left context simultaneously. (iii). SAE supports dimensionality reduction and it has the capability to reconstruction the data from the latent space. (iv). GRU generates the next word in the sequence based on the previous sequence of words and it is better than LSTM in terms of memory and speed. (v). MLP has the capability to learn non-linear models.

The performance metrics given by ImageCLEF for evaluating concept detection and caption prediction tasks are: F1 Score, Bilingual Evaluation Understudy (BLEU) score, Recall-Oriented Understudy for Gisting Evaluation (ROUGE), Metric for Evaluation of Translation with Explicit Ordering (METEOR), Consensus-based Image Description Evaluation (CIDEr), Semantic Propositional Image Caption Evaluation (SPICE) and Bidirectional Encoder Representations from Transformers (BERT) Score [18]. The F1 Score is computed as the weighted average of precision and recall. It is more useful than accuracy especially if there is an uneven class distribution. The BLEU score compares the n-gram of the predicted caption with the n-gram of the reference caption to count the number of matches. The advantage of using BLEU over other score is, it provides the overall assessment of model quality. The ROUGE measures the longest common subsequence between predicted and reference caption. Like, BLEU score, ROUGE similarity value also lies between zero and one. It determines the overall quality of the predicted answer. The METEOR is based on harmonic mean of unigram recall and precision, with recall weighted higher than precision. It is used in a language-independent manner, and hence it has an ability to model features (especially synonyms, stem words and paraphrasing) of a specific language. The CIDEr measures the sentence similarity between the predicted and reference caption by inherently captures the notions of grammaticality, saliency, importance and accuracy. The SPICE measures the effectiveness of image captions in terms of recovered objects, attributes and the relationship between them. The BERT computes the similarity between each token in the predicted caption with each token in the reference caption. It overcomes the flaw of BLEU score by considering semantic and syntactic abilities.

The remaining part of the paper are discussed with following subsections. In Sect. 2, the concept detection and caption prediction datasets and the inference from these datasets are discussed. The design of the proposed system is explained in Section 3. A brief summary about the implementation, result and the evaluation of all runs for both tasks are given in Section 4 and, conclusion and future work are summarized at the end.

2. Dataset

The concept detection and caption prediction datasets comprises of training set, validation set and test set and it is represented in terms of number of images, concept IDs, concept names and captions in Table 1. In concept detection dataset, each image corresponds to one or more concept IDs and each concept ID represents one concept names. In caption prediction dataset, each image corresponds to only one caption.

Table 1
Dataset description for concept detection and caption prediction

	Datasets	Training Set	Validation Set	Test Set
Concept Detection [19]	**No. of images**	83275	7645	7601
	No. of concept IDs	83275	7645	-
	No. of concept name	8374	8374	-
Caption Prediction [19]	**No. of images**	83275	7645	7601
	No. of captions	83275	7645	-

Some of the interesting inferences from concept detection dataset is given in Table 2 and 3 and caption prediction dataset is shown in Figure 4. The maximum and minimum number of tags corresponds to the image along with the concept ID are mentioned in Table 2. Because each image is mapped with one or more tags and each tag corresponds to one concept name only. The length of the concept name ranges from one (minimum) to twelve (maximum) and it is listed in Table 2.

Table 2
Brief inference from concept detection dataset

Concept Detection	Value	Respective tags (or) concept name
Maximum number of tags	13	"C0002978;C0021102;C0085590;C0232180;C0205197; C0887842;C0004704;C0021398;C0441127;C2698651; C0333138;C0009924;C0036426"
Minimum number of tags	1	Most of the samples having only one tags
Maximum length of the concept	12	"Arterial Occlusive Disease, Progressive with Hypertension, Heart Defects, Bone Fragility and Brachysyndactyly"
Minimum length of the concept	1	Most of the concept size is one

In Table 3, the maximum occurrence of Concept IDs in the dataset are listed along with its frequency and concept name for better understanding.

Table 3
Top 10 concept ID with its frequency

Concept ID	Concept Name	Frequency
C0040405	X-Ray Computed Tomography	28885
C1306645	Plain x-ray	26412
C0024485	Magnetic Resonance Imaging	15693
C0041618	Ultrasonography	12236
C0817096	Chest	8030
C0002978	Angiogram	6464
C0000726	Abdomen	6243
C0037303	Bone structure of cranium	5175
C0221198	Lesion	4094
C0205131	Axial	3528

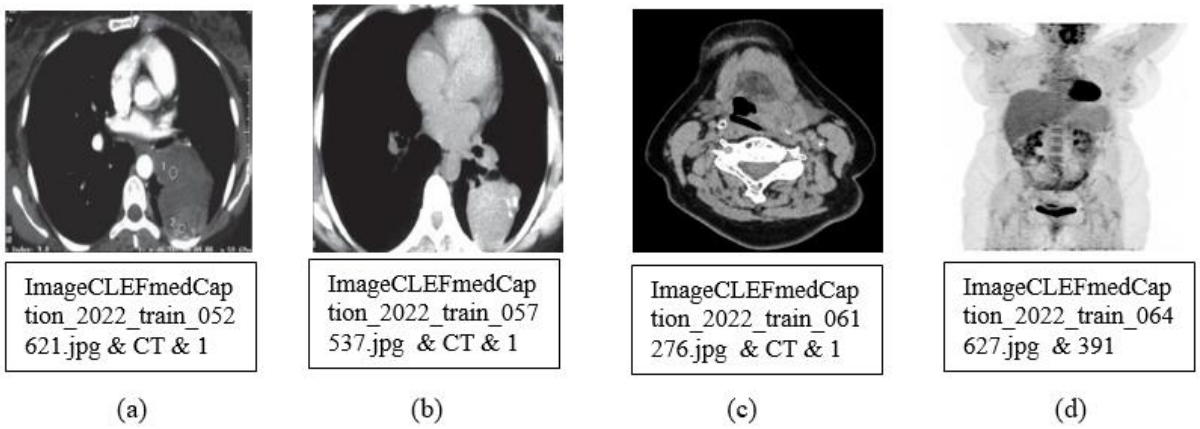

Figure 1: Minimum and maximum length caption's radiology images

The Figure 1 represents the radiology images corresponds to the minimum and maximum length of the caption. In the caption prediction dataset, the length of the caption varies from 1 (minimum) to 391 (maximum). The length of the caption and the words used in the caption varies for different human annotator, based on the annotator's domain knowledge and vocabulary skills and, the region of interest. The caption for the radiology images in Figure 1 are, (a) and (b) shows unusual cause of a lung mass [20], (c) represents a sudden-onset facial oedema [21] and, (d) illustrates chikungunya virus infection [22].

3. System Design

The system design of the proposed concept detection and caption prediction tasks are shown in Figure 2 and 3. In Figure 2, two concept detection models are developed based on images, concept IDs and concept names in the training phase and, the generated model is validated by detecting the concept for the radiology images in the test set. In Figure 3, two caption prediction models are developed based on the radiology images and captions in the training phase. The generated caption prediction model is validated by predicting the caption for the medical images in the test set.

3.1. Concept Detection

The concept detection model is developed in two ways namely, multi-label classification (model 1) and information retrieval (model 2) using DenseNet. The multi-label classification system will predict all the suitable concepts which has probability value greater than criterion value for each image in the test set. The information retrieval system will retrieve the suitable concept from the training set based on the nearest neighbour and cosine similarity for each image in the test set.

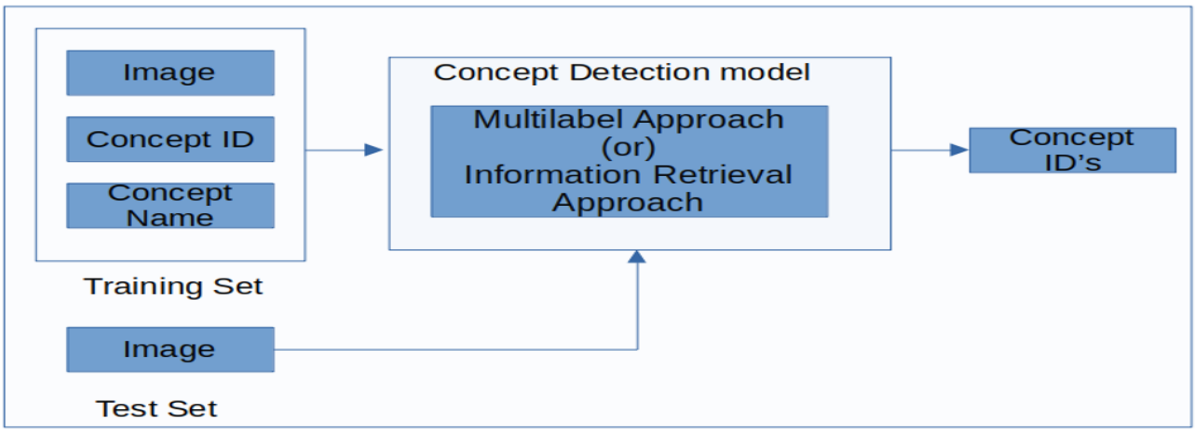

Figure 2: System design of proposed concept detection model

3.1.1. Model 1: Multi-label classification approach

In model 1, each image features are extracted using DenseNet and mapped with respective concepts in the training phase. For this, all layers except the last layer in DenseNet is freeze so weights remain same throughout the training and only the weights from added layers will update gradually. Then global average pooling, dense layer with sigmoid activation function are added after the last layer which predicts the probability value for each image. The global average pooling used in this model creation acts as a great alternative for CNN because it generates the one feature map for each corresponding concept category. The number of nodes in the dense layer is maintain to be equal to the number of concept names then only each node in this layer generates the probability value for each concept with respect to the image. Among the probability value, the concept which has probability value greater than criterion value is considered as the predicted concept. Moreover, the model is fine-tuned by minimizing the mean square error between the predicted and ground truth value. Then in the testing phase, the generated model is evaluated for radiology images in the test set which detects one or more concepts for each images. The concepts extracted for test set under different criteria are combined by (i). union of union (i.e., merging list of concepts from two results), (ii). Intersection of intersection (i.e., merging the common concepts from two results).

3.1.2. Model 2: Information retrieval approach

The model 2 generates feature vector for each image and mapped with concept name by DenseNet which is same as in multi-label classification approach. In the testing phase, the feature vector is extracted for each image, then the cosine similarity is computed between the test image and the training image feature vector. The training images which has the similarity score more than criterion value are retrieved. The respective concepts of these training images are considered as a resulted concept for the test images.

3.2. Caption Prediction

The caption prediction system is developed using ResNet followed by BERT (model 3) and, SAE followed by MLP and GRU (model 4). In model 3, based on the given radiology image, the system will generate the caption based on the context of the image. The SAE in model 4, extract the significant information based on the dimension of the latent space and then GRU and MLP will generate the sequence of next word in the caption based on previous word. In both models, caption prediction model is fine-tuned by minimizing mean square error loss, cross entropy loss and KL- divergence loss in the training phase.

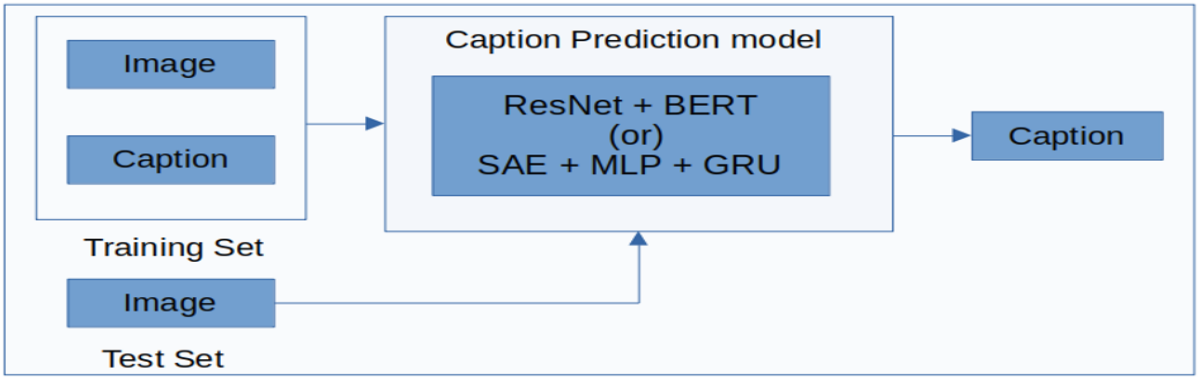

Figure 3: System design of proposed caption prediction model

3.2.1. Model 3: ResNet with BERT

The model 3 takes images and the respective captions as input and generates the image and text feature vectors by ResNet and BERT. Before that, glove embedding is used to generate the vocabulary list and length of the captions are normalized based on the maximum caption length by padding. In the training phase, the caption is predicted word by word from the vocabulary list in the glove embedding based on the feature vector and the length of the caption is maintained within the maximum caption length. This model is fine-tuned by minimizing the error rate and the obtained model is used for caption prediction in the test phase.

3.2.2. Model 4: SAE with MLP and GRU

In this model, the images and the respective captions in the training phase is taken as input by the SAE, MLP and GRU. The SAE consists of seven layers, the middle layer represents the latent space, the layer before the latent space represents the encoder and the layer after that, represents the decoder. In the encoder part, the nodes in each layer are in increasing fashion (power of two). The latent space has the maximum dimension and in decoder part, the node distribution are in decreasing fashion which is also power of two. Even though, many nodes are available only the particular sequence of nodes are adapted by each sample and it differs from sample to sample. In the encoder part, convolution2D layer, maxpooling2D and leuky ReLU are used to generate flatten sequence of features and in the decoder part, flatten features are unflatten then used convolutionTranspose2D, unmaxpooling2D and leuky ReLU to generate image feature vector. The multi-layer perceptron and Gated Recurrent Unit are used to extract the text feature vector based on the image feature vector and it is further supported by word embedding. In the training phase, the model is fine-tuned by reducing the loss value which we aforementioned and based on this obtained model, the captions are predicted for test set.

4. Implementation and Results

In this section, the proposed concept detection and caption prediction models are implemented and the results are compared and analysed using performance metrics.

4.1. System Specification

The hardware and software required for the implementation of concept detection and caption prediction model includes, (i). Intel i5 processor with NVIDIA graphics card, 4800M at 4.3GHZ clock speed, 16GB RAM, Graphical Processing Unit and 2TB disk space, (ii). Linux – Ubuntu 20.04 operating system, Python 3.7 package with required libraries like tensorflow, torch, sklearn, nltk, pickle, pandas, etc.,

4.2. Results of Concept Detection Models

The test set results obtained from concept detection model are measured by F1-score and secondary F1-score as given in Table 4, the first six runs are based on model 1 and the last one is based on model 2. The description about the runs are as follows, (i). In run1, the model is implemented based on the features learned in 50 epochs and early stopped if the continuous three epochs having same accuracy values. The particular concept ID are detected only if the obtained probability value for the predicted concept ID with respect to the image is greater than 0.4. (ii). Run2 is same as run1, but the probability value is modified from 0.4 to 0.1. (iii). In run3, the early stopping value is updated from three to five. (iv). Run4 is the union of results obtained from run2 and run3. The results are merged based on the test set image name which acts as a primary key. (v). Same as run1, but the number of epochs are increased from 50 to 100. (vi). Like run4, run6 is the intersection of results obtained from run2 and run3. The concept ID which is common in both results are considered for evaluation. (vii). In run7, the number of epochs is fixed to be 50, early stopped within three epoch and, based on the cosine similarity value between the particular test set image and training set images and the nearest neighbour of one is considered for evaluation.

From the results, it has been observed that, selection of appropriate hyperparameters plays a significant role in improving the model performance. They are, (i). Early stopping at third epoch performed better than fifth epoch and also avoids the overfitting problem. (ii). Samples with more than one label identifies all the concept names by fixing the probability value as 0.1 rather than the recommended value of 0.4. as per many research papers. (iii). Union of two independent results gives better performance than the intersection of result (iv). Cosine similarity value retrieves the more relevant images for each test image than the Jaccard similarity value.

Table 4
Brief description about each runs

Run number	Approach/Techniques	F1 Score	Secondary F1 Score
1	Model 1	0.385	0.524
2	***Model 1***	***0.418***	***0.654***
3	Model 1	0.408	0.829
4	***Model 1***	***0.418***	***0.654***
5	Model 1	0.412	0.661
6	Model 1	0.406	0.614
7	Model 2	0.316	0.412

Among the results, run2 and run4 obtained better performance value in terms of F1 score and secondary F1 score and it is italicized in Table 4. From run2, it has been inferred that probability value greater than 0.1 gives considerably better result not only for this case but also for most of the cases. Moreover, the early stopping with a patience of three epochs also achieved better performance value than five epochs. The run4 exhibit the characteristics of run2 and run3 because run4 is the union of the results of these two runs.

Table 5
Top 10 ranking of ImageCLEF 2022 concept detection task

Rank	Team Name	F1 Score	Secondary F1 Score	No. of runs submitted
1	AUEB-NLP-Group	0.451	0.791	6
2	fdallaserra	0.450	0.822	5
3	CSIRO	0.447	0.794	10
4	eecs-kth	0.436	0.855	10

5	vcmi	0.433	0.863	9
6	PoliMi-ImageClef	0.432	0.851	10
7	***SSNSheerinKavitha***	***0.418***	***0.654***	***7***
8	IUST-NLPLAB	0.398	0.673	6
9	Morgan_CS	0.351	0.628	8
10	kdelab	0.310	0.411	10

In concept detection task, totally 61 teams were registered, 11 teams were participated and they 104 submissions were recorded. Among this, we have submitted seven successful submissions and achieved seventh rank in ImageCLEF 2022 Concept Detection task. The overall ranking achieved by top 10 teams are listed in Table 5.

4.3. Results of Caption Prediction Models

The brief description about each run in terms of techniques used and the performance obtained for caption prediction model are listed in Table 6. Among seven runs, model 3 is used in three runs and, model 4 is used in four runs. The description about each run are as follows. In run1, model 4 is used, in which the number of epochs is fixed to be 50, early stopped with three epochs, used adam optimizer and maintain learning rate to be 0.002. The run2 is same as run1, but the learning rate is reduced to 0.001 and used RMSProp optimizer. The model 3 is used in run3, in which the number of epochs is fixed to be 50, early stopping with three epochs, used "adam" optimizer and maintain learning rate to be 0.004. The run4 is same as run2, but the latent size dimension is doubled. In run 5, the batch size is increased to 64 and remaining is as same as run4. The batch size can't be increased further of memory insufficiency problem. The run6 is same as run3 but the optimizer here used is Stochastic Gradient Descent. In run7, the learning rate is reduced to 0.001, early stopped in fifth epoch and remaining is as same as run6.

The analysis of results, leads to some important observations: (ii). The learning rate value 0.004 gives better performance than 0.001 or 0.002 (ii). In terms of optimizer, adam optimizer improves the performance of the model because few parameters are required for tuning and reduces computation time (iii). Early stopping at the third epoch performs better than the fifth epoch and avoids overfitting problem (iv). The model is trained for the batch size of 64 only, since the higher batch size leads to data insufficiency problem.

Table 6
Brief description about each runs

Run Number	Approach/ Techniques	BLEU	Rouge	METEOR	CIDEr	SPICE	BERTScore
1	Model 4	0.159	0.042	0.023	0.017	0.007	0.545
2	Model 4	0.141	0.039	0.020	0.015	0.006	0.550
3	***Model 3***	***0.160***	***0.043***	***0.023***	***0.017***	***0.007***	***0.545***
4	Model 4	0.154	0.039	0.022	0.015	0.006	0.550
5	Model 4	0.153	0.040	0.021	0.015	0.007	0.552
6	Model 3	0.155	0.039	0.022	0.014	0.006	0.550
7	Model 3	0.142	0.038	0.020	0.014	0.006	0.549

From the results, it has been inferred that run3 achieved better performance value and it is italicized in Table 6. The overall results show that the lowest learning rate and early stopping gives better result. In caption prediction task, totally 43 teams were registered, 10 teams were participated and 81 submissions were recorded. Among this, we have made seven successful submissions and achieved tenth rank in ImageCLEF 2022 caption prediction task. The overall ranking achieved by top 10 teams are given in Table 7.

Table 7
Top 10 ranking of ImageCLEF 2022 caption prediction task

Rank	Team Name	BLEU	Rouge	METEOR	CIDEr	SPICE	BERTScore
1	IUST-NLPLAB	0.483	0.142	0.093	0.030	0.007	0.561
2	AUEB-NLP-Group	0.322	0.167	0.074	0.190	0.031	0.599
3	CSIRO	0.311	0.197	0.084	0.269	0.046	0.623
4	Vcmi	0.306	0.174	0.075	0.205	0.036	0.604
5	eecs-kth	0.292	0.116	0.062	0.132	0.022	0.573
6	Fdallaserra	0.291	0.201	0.082	0.256	0.046	0.610
7	Kdelab	0.278	0.158	0.074	0.411	0.051	0.600
8	Morgan_CS	0.255	0.144	0.056	0.148	0.023	0.583
9	MAI_ImageSem	0.221	0.185	0.068	0.251	0.039	0.606
10	***SSNSheerinKavitha***	***0.160***	***0.043***	***0.023***	***0.017***	***0.007***	***0.545***

5. Conclusion and Future Works

This paper describes four different approaches to solve ImageCLEF 2022 medical concept detection and caption prediction tasks. The concept detection tasks are implemented using two approaches namely, multi-label classification and information retrieval approach using DenseNet pre-trained model. The caption prediction tasks are implemented in two ways as, ResNet followed by BERT and, SAE followed by MLP and GRU. From the results of these models, it has been inferred that multi-label classification gives better result for concept detection and, ResNet followed by BERT gives better results for caption prediction task. As compared with the best scores given by ImageCLEF, the proposed concept detection model lacks only by 0.033 and 0.131 in terms of F1-score and secondary F1-score respectively. And the proposed caption prediction model lacks only by 0.323 and 0.016 in terms of BLEU score and BERT score respectively.

In future work, the performance of the concept detection system can be improved by majority voting instead of union by union or intersection by intersection. For caption prediction system, the performance can be enhanced by Generative Pre-trained Transformer instead of BERT. The overall performance of the system can be improved by reducing the irrelevant samples, increasing the number of epochs and maintaining the minimum learning rate.

6. Acknowledgements

Our profound gratitude to Sri Sivasubramaniya Nadar 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.

7. References

[1] H. Kasban, M. A. M. El-Bendary, D. H. Salama, A comparative study of medical imaging techniques, International Journal of Information Science and Intelligent System, 4(2) (2015) 37-58. URL: https://ilearn.th-deg.de/pluginfile.php/480243/mod_book/chapter/8248/updated_JXIJSIS2015.pdf

[2] B. Ionescu, H. Müller, R. Péteri, J. Rückert, A. B. Abacha, A. G. S. D. Herrera, C. M. Friedrich, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, S.Kozlovski, Y. D. Cid, V. Kovalev, L. Ștefan, M. G. Constantin, M. Dogariu, A. Popescu, J. Deshayes-Chossart, 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, in: Proceedings of the Thirteen International Conference of the CLEF Association (CLEF 2022), Springer Lecture Notes in Computer Science, LNCS, Bologna, Italy, September 5-8, 2022.

[3] 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. URL: https://www.computer.org/csdl/proceedings-article/cvpr/2017/0457c261/12OmNBDQbld

[4] N. Thakur, D. Mehrotra, A. Bansal, M. Bala, Analysis and Implementation of the Bray-Curtis Distance-Based Similarity Measure for Retrieving Information from the Medical Repository, in: International Conference on Innovative Computing and Communications, Springer, 2019, pp. 117–125. URL: https://link.springer.com/chapter/10.1007/978-981-13-2354-6_14

[5] M. Tan, Q. Le, EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks, in: K. Chaudhuri, R. Salakhutdinov (Eds.), in: Proceedings of the 36th International Conference on Machine Learning, California, USA, 2019, pp. 6105–6114. URL: http://proceedings.mlr.press/v97/tan19a.html.

[6] G. Huang, Z. Liu, L. V. D. Maaten, K. Q. Weinberger, Densely Connected Convolutional Networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, Honolulu, USA, 2017, pp. 2261–2269. doi:10.1109/ CVPR.2017.243.

[7] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, O. Wang, The unreasonable effectiveness of deep features as a perceptual metric, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 586–595. doi:10.1109/CVPR.2018.00068.

[8] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

[9] G. Huang, Z. Liu, L. V. D. Maaten, K. Q. Weinberger, Densely Connected Convolutional Networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, Honolulu, USA, 2017, pp. 2261–2269. doi:10.1109/CVPR.2017.243.

[10] 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, Paper - 95, pp. 1-8. URL: http://ceur-ws.org/Vol-2936/paper-95.pdf

[11] S. S. N. Mohamed, K. Srinivasan, ImageCLEF 2020: An approach for Visual Question Answering using VGG-LSTM for Different Datasets, in: CLEF (Working Notes). CEUR Workshop Proceedings, Greece, September 22-25, 2020, Paper – 94, pp. 1 – 10. URL: http://ceur-ws.org/Vol-2696/paper_94.pdf

[12] A. Tasissa, E. Theodosis, B. Tolooshams, D. Ba. (2020). Towards improving discriminative reconstruction via simultaneous dense and sparse coding. arXiv preprint arXiv:2006.09534.

[13] D. R. Beddiar, M. Oussalah, T. Seppänen, Attention-based CNN-GRU model for automatic medical images captioning: ImageCLEF 2021, in: CLEF2021 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Bucharest, Romania, 2021, Paper - 94, pp. 1-11. URL: http://ceur-ws.org/Vol-2936/paper-94.pdf

[14] N. M. S. Sitara, K. Srinivasan, SSN MLRG at VQA-MED 2021: An Approach for VQA to Solve Abnormality Related Queries using Improved Datasets, in: CLEF (Working Notes) CEUR Workshop Proceedings, Romania, 2021, Paper – 110, pp. 1329 - 1335. URL: http://ceur-ws.org/Vol-2936/paper-110.pdf

[15] R. Dey, F. M. Salem, Gate-variants of gated recurrent unit (GRU) neural networks, in: IEEE sixty international midwest symposium on circuits and systems, 2017, pp. 1597-1600. doi: 10.1109/MWSCAS.2017.8053243

[16] J. Devlin, M. W. Chang, K. Lee, K. Toutanova. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[17] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners, Technical Report, Open-AI, 2019.

[18] https://www.imageclef.org/2022/medical/caption

[19] J. Rückert, A. B. Abacha, A. G. S. D. 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: Experimental IR Meets Multilinguality, Multimodality, and Interaction, in: CLEF2022 Working Notes, CEUR Workshop Proceedings (CEUR-WS.org), Bologna, Italy, September 5-8, 2022.
[20] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2732071/
[21] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3812827/
[22] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5617949/
[23] A. G. S. D. 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, pp. 1-13. URL: http://ceur-ws.org/Vol-1609/16090219.pdf
[24] J. Singh, R. Banerjee, A Study on Single and Multi-layer Perceptron Neural Network, in: Third International Conference on Computing Methodologies and Communication, 2019, pp. 35-40, doi: 10.1109/ICCMC.2019.8819775.
[25] S. Targ, D. Almeida, K. Lyman. (2016). Resnet in resnet: Generalizing residual architectures. arXiv preprint arXiv:1603.08029.
[26] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, D. Krishnan, Supervised contrastive learning, in: H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, 33, 2020, pp. 18661–18673. URL: https://proceedings.neurips.cc/paper/2020/file/d89a66c7c80a29b1bdbab0f2a1a94af8-Paper.pdf.
[27] E. Loza Mencía, F. Janssen, Learning rules for multi-label classification: a stacking and a separate-and-conquer approach, Machine Learning 105 (2016) 77–126. URL: https://doi.org/10.1007/s10994-016-5552-1