SSN MLRG at ImageCLEFmedical Caption 2023: Automatic Concept Detection and Caption Prediction using ConceptNet and Vision Transformer

Sheerin Sitara Noor Mohamed[1, *] and Kavitha Srinivasan[1]

[1] Department of CSE, Sri Sivasubramaniya Nadar College of Engineering, Kalavakkam – 603110, India

Abstract

An automatic image captioning attains tremendous advancement in the last few years. 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 techniques using ConceptNet and Vision Transformer (ViT). The concept detection models are developed using multi-label classification and resulted the F1-score and secondary F1-score as 0.464 and 0.860 respectively. The caption prediction models are implemented using Vision Transformer (ViT), which resulted a BERT and CLIPScore of 0.544 and 0.687 respectively.

Keywords

ImageCLEF, caption prediction, concept detection, ViT, multilabel classification, image captioning

1. Introduction

Images are extensively used for conveying enormous amounts of information over internet and social media and hence there is an increasing demand for image data analytics for designing efficient information processing systems. This leads to the development of systems with capability to automatically analyze the scenario contained in the image and to express it in meaningful natural language sentences. Image caption generation is an integral part of many useful systems and applications such as visual question answering, surveillance video analysis, video captioning, automatic image retrieval, assistance for visually impaired people, biomedical imaging, robotics and so on. A good captioning system is capable of highlighting the contextual information in the image similar to human cognitive system. In the recent years, several techniques for automatic caption generation in images have been proposed that can effectively solve many computer vision challenges related to medical image captioning. The different types of medical imaging modalities are Computed Tomography (CT), X-Ray (XR), Magnetic Resonance Imaging (MRI), Positron Emission Tomography (PET), angiogram, mammogram and ultrasound. The ImageCLEF forum [1] 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

[1] CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece
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)

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: information retrieval [3] and multi-label classification approaches. Among these approaches, multi-label classification using ConceptNet [4] 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 caption prediction techniques are grouped into, (i). Techniques which performs both image and test processing like Visual Transformers (ViT) [5], (ii). Combination of techniques for image processing and text processing using deep learning techniques [6]. Among the techniques, ViT is chosen because it can be applied to low-level vision, segmentation, and multi-modality.

The performance metrics given by ImageCLEF for evaluating concept detection and caption prediction tasks are: F1 Score [7], BiLingual Evaluation Understudy (BLEU) [8], Bilingual Evaluation Understudy (BLEURT) [9] score, Recall-Oriented Understudy for Gisting Evaluation (ROUGE) [10], Metric for Evaluation of Translation with Explicit Ordering (METEOR) [11], Consensus-based Image Description Evaluation (CIDEr) [12], Semantic Propositional Image Caption Evaluation (SPICE) [13], CLIPScore [14] and Bidirectional Encoder Representations from Transformers (BERT) Score [15]. Among these, except F1-score, all other performance metrics are used to evaluate the results of caption prediction tasks.

The remaining part of the paper are discussed with following subsections. In Section 2, the concept detection and caption prediction tasks and 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 is summarized at the end.

2. Task and Dataset Description

In this section, two sub tasks of image captioning and given datasets are discussed. The two sub tasks includes, ImageCLEF concept detection and caption prediction.

2.1 ImageCLEF Concept Detection and Caption Prediction Task

The automatic concept detection and caption prediction are identifying the presence and location of relevant concepts in a large corpus of medical images. Based on the visual image content, this sub task provides the building blocks for the scene understanding step by identifying the individual components from which captions are composed. The concepts can be further applied for context-based image and information retrieval purposes. On the basis of the concept vocabulary detected in the first sub task as well as the visual information of their interaction in the image, participating systems are tasked with composing coherent captions for the entirety of an image. In this step, rather than the mere coverage of visual concepts, detecting the interplay of visible elements is crucial for strong performance.

2.2 ImageCLEF Concept Detection and Caption Prediction 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, concepts and captions in Table 1. The number of images are equally distributed as 75%, 12.5% and 12.5% for training set, validation set and test set for both datasets. 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

Task	Input/Output	Training Set	Validation Set	Test Set	Total
Concept Detection/	**#Images**	60918	10437	10473	81828
Caption Prediction	**#Concept/#Captions**	60918	10437	-	71355

3. System Design

The system design of the proposed concept detection and caption prediction tasks are shown in Figure 1 and 2. In Figure 1, multilabel classification based 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 2, caption prediction models are developed using ViT, 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 by multi-label classification using ConceptNet [4]. The multi-label classification system predicts all the suitable concepts which has probability value greater than criterion value for each image in the test set. In the training phase, based on the frequency, the high-level semantic type on the Top-100 concepts are extracted. Then image features are extracted using ConceptNet and mapped with respective concepts in the training phase. For this, all layers except the last layer in ConceptNet is freeze so weights remain same throughout the training and only the weights from added layers updates 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 Convolutional Neural network (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). Regarding the training process, these models were trained during 10 epochs, using the Adam optimizer with a learning rate of 10−4. These models also used the “2 Phases” strategy, where the backbone layers are frozen for 5 epochs and then unfroze them for the remaining epochs. The best model is saved based on the lowest validation loss and its system design is shown in Figure 1.

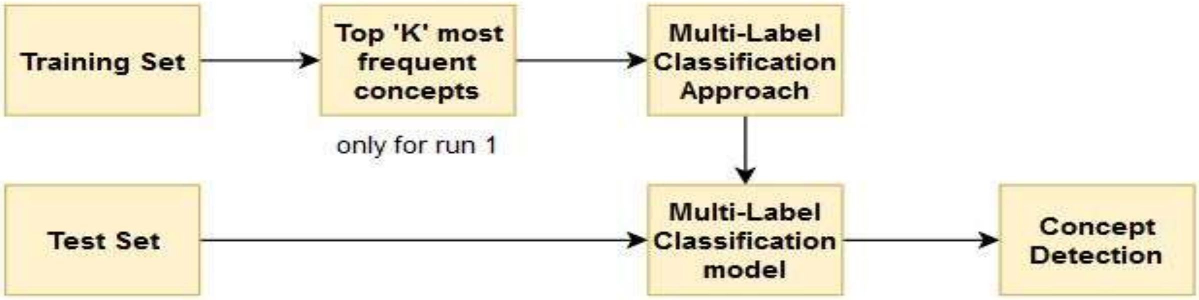
Training Set
Top 'K' most frequent concepts
only for run 1
Multi-Label Classification Approach
Test Set
Multi-Label Classification model
Concept Detection

Figure 1: System design of proposed concept detection model

3.2 Caption Prediction

The caption prediction system is developed using Vision Transformer (ViT). Because ViT uses self-attention mechanism and it derives information from whole image. This pre-trained transformer based model generates the caption based on the context of the image. The system design is shown in Figure 2 for better understanding. The encoder in ViT receives as input an image divided into patches of 16×16 pixels. The decoder receives the ground-truth caption as input (i.e., training is done using teacher forcing) and the encoder hidden states as inputs to the cross attention layers. The model is trained autoregressively for next token prediction using causal (or unidirectional) self-attention, which means that a given token can only attend to previous tokens. The model was implemented using the Vision Encoder Decoder class from the Hugging Face Transformers library and chose a tiny Data-efficient image Transformer (DeiT) pretrained on ImageNet for the encoder. For the decode, leverage pretrained weights from the Distilled-GPT2, a distilled version of the GPT-2 architecture are used. This model is initially trained for 20 epochs, and then for an additional 20 epochs starting from the checkpoint with the lowest validation loss. The AdamW optimiser with an initial learning rate of 5×10−5, linearly decayed are chosen. Due to limitations of the computational resources available unable to fine tune the model using self-critical sequence training.

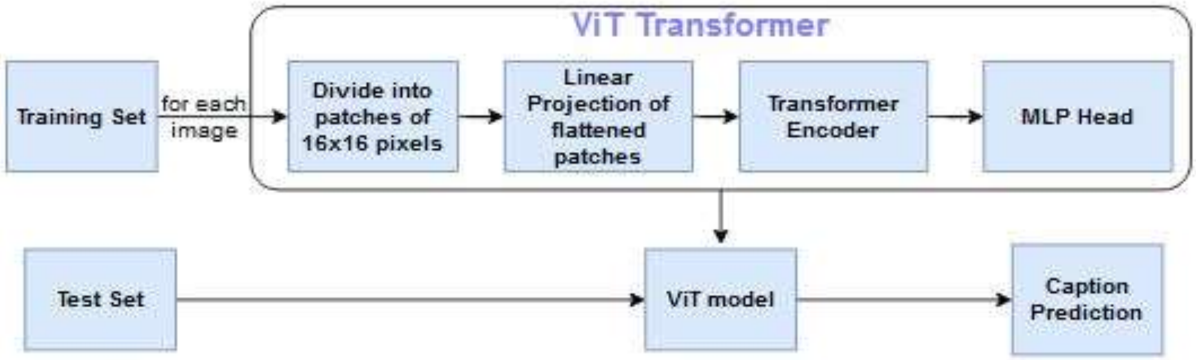

Figure 2: System design of proposed caption prediction model

4. Experiments and Results

The hardware and software required for the implementation of concept detection and caption prediction model includes, (i). Intel i5 processor with NVIDIA GeForce Ti 4800 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 2.13.0, torch 2.0, sklearn 0.20, nltk, pickle, pandas, etc., Among the results, run1 obtained better performance value in terms of F1 score and F1 score manual and it is italicized in Table 2. From run2, it has been inferred that considering top-100 concepts gives better result. In concept detection task, we have submitted two successful submissions and achieved seventh rank in ImageCLEF 2023 Concept Detection task.

Table 2
Brief description about each runs

Run Number	F1-Score	F1-Score Manual
1	***0.464***	***0.860***
2	0.461	0.856

Table 3
Top ranking of ImageCLEF 2023 concept detection task

Team Name	F1-Score	F1-Score Manual	Rank
AUEB-NLP-Group	0.522	0.925	1
KDE-Lab_Med	0.507	0.932	2
VCMI	0.499	0.916	3
IUST_NLPLAB	0.495	0.880	4
Clef-CSE-GAN-Team	0.495	0.910	5
CS_Morgan	0.483	0.890	6
SSNSheerinKavitha	***0.464***	***0.860***	***7***
closeAI2023	0.448	0.856	8
SSN_MLRG	0.017	0.112	9

Among the results, run1 obtained better performance value in terms of F1 score and F1 score manual and it is italicized in Table 2. From run2, it has been inferred that considering top-100 concepts gives better result. In concept detection task, we have submitted two successful submissions and achieved seventh rank in ImageCLEF 2023 Concept Detection task. The overall ranking achieved by top teams are listed in Table 3.

The brief description about each run are listed in Table 4. The description about each run are as follows. In run1, the number of epochs is fixed to be 20, used adam optimizer and maintain learning rate to be 0.005, batch size is 32. The run2 is same as run1, but the learning rate is reduced to 0.001 and used Stochastic Gradient Descent (SGD) optimizer and number of epochs fixed to be 40. From the results, it has been inferred that run4 achieved better performance value and it is italicized in Table 4. The overall results show that the lowest learning rate, epochs and SGD gives better result. The overall ranking achieved by top teams are given in Table 5.

Table 4
Brief description about each runs

Run Number	BERTScore	ROUGE	BLEURT	BLEU	METEOR	CIDEr	CLIPScore
3	0.543	0.086	0.215	0.074	0.025	0.014	0.685
4	***0.544***	***0.086***	***0.215***	***0.074***	***0.025***	***0.014***	***0.687***

Table 5
Top ranking of ImageCLEF 2023 caption prediction task

Team Name	BERT Score	ROUGE	BLEURT	BLEU	METEOR	CIDEr	CLIP Score	Rank
CSIRO	0.642	0.244	0.313	0.161	0.079	0.202	0.814	1
closeAI2023	0.628	0.240	0.320	0.184	0.087	0.237	0.807	2
AUEB-NLP-Group	0.617	0.213	0.295	0.169	0.071	0.146	0.803	3
PCLmed	0.615	0.252	0.316	0.217	0.092	0.231	0.802	4
VCMI	0.614	0.217	0.308	0.165	0.073	0.172	0.808	5
KDE-Lab_Med	0.614	0.222	0.301	0.156	0.072	0.181	0.806	6
SSN_MLRG	0.601	0.211	0.277	0.141	0.061	0.128	0.775	7
DLNU_CCSE	0.600	0.202	0.262	0.105	0.055	0.133	0.772	8
CS_Morgan	0.581	0.156	0.224	0.056	0.043	0.083	0.759	9
Clef-CSE-GAN-Team	0.581	0.218	0.269	0.145	0.070	0.173	0.789	10
Bluefield-2023	0.577	0.153	0.271	0.154	0.060	0.100	0.783	11
IUST_NLPLAB	0.566	0.289	0.222	0.268	0.100	0.177	0.806	12
SSNSheerinKavitha	***0.544***	***0.086***	***0.215***	***0.074***	***0.025***	***0.014***	***0.687***	***13***

5. Conclusion

In this paper, medical concept detection and caption prediction models are developed for ImageCLEF image captioning task. The concept detection tasks are implemented by multi-label classification approach by ConceptNet by considering only the top 100 concepts. The caption prediction task is implemented by ViT. From the results of these models, it has been inferred that multi-label classification approach using ConceptNet by considering only the top 100 concepts gives better results. As compared with the best scores given by ImageCLEF, the proposed concept detection model lacks only by 0.058 and 0.065 in terms of F1-score and F1-score manual respectively. And the proposed caption prediction model lacks only by 0.098 in terms of BLEU score. In future work, 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] Bogdan Ionescu, Henning Müller, Ana-Maria Drăgulinescu, Wen-wai Yim, Asma Ben Abacha, Neal Snider, Griffin Adams, Meliha Yetisgen, Johannes Rückert, Alba García Seco de Herrera, Christoph M. Friedrich, Louise Bloch, Raphael Brüngel, Ahmad Idrissi-Yaghir, Henning Schäfer, Steven A. Hicks, Michael A. Riegler, Vajira Thambawita, Andrea Storås, Pål Halvorsen, Nikolaos Papachrysos, Johanna Schöler, Debesh Jha, Alexandra-Georgiana Andrei, Ahmedkhan Radzhabov, Ioan Coman, Vassili Kovalev, Alexandru Stan, George Ioannidis, Hugo Manguinhas, Liviu-Daniel Ștefan, Mihai Gabriel Constantin, Mihai Dogariu, Jérôme Deshayes, Adrian Popescu, Overview of the ImageCLEF 2023: Multimedia Retrieval in

Medical, Social Media and Recommender Systems Applications, in Experimental IR Meets Multilinguality, Multimodality, and Interaction.Proceedings of the 14th International Conference of the CLEF Association (CLEF 2023), Springer Lecture Notes in Computer Science LNCS, Thessaloniki, Greece, September 18-21, 2023.

[2] Johannes Rückert, Asma Ben Abacha, Alba G. Seco de Herrera, Louise Bloch, Raphael Brüngel, Ahmad Idrissi-Yaghir, Henning Schäfer, Henning Müller and Christoph M. Friedrich. Overview of ImageCLEFmedical 2023 – Caption Prediction and Concept Detection, in Experimental IR Meets Multilinguality, Multimodality, and Interaction. CEUR Workshop Proceedings (CEUR-WS.org), Thessaloniki, Greece, September 18-21, 2023.

[3] 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

[4] Q. Chen, A. Allot, R. Leaman, R. Islamaj, J. Du, L. Fang, and Z. Lu. Multi-label classification for biomedical literature: an overview of the BioCreative VII LitCovid Track for COVID-19 literature topic annotations. Database, 2022.

[5] Z. Xia, X. Pan, S. Song, L. E. Li and G. Huang. Vision transformer with deformable attention. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 4794-4803, 2022.

[6] P. K. Chaubey, T. K. Arora, K. B. Raj, G. R. Asha, G. Mishra, S. C. Guptav and M. Alhassan, M. Sentiment Analysis of Image with Text Caption using Deep Learning Techniques. Computational Intelligence and Neuroscience, 2022.

[7] S. S. N. Mohameda and K. Srinivasanb, SSN MLRG at ImageCLEFmedical caption 2022: Medical concept detection and caption prediction using transfer learning and transformer based learning approaches, in: Proceedings of the working notes of CLEF 2022, Bologna, Italy, September 5 – 8 2022.

[8] K. Papineni, S. Roukos, T. Ward, and 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 (pp. 311-318).

[9] T. Sellam, D. Das and A. P. Parikh, BLEURT: Learning robust metrics for text generation. arXiv preprint arXiv:2004.04696.

[10] Lin, C. Y. (2004, July). Rouge: A package for automatic evaluation of summaries. In Text summarization branches out (pp. 74-81).

[11] S. Banerjee and A. Lavie (2005, June). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization (pp. 65-72).

[12] G. O.D. Santos, E. L. Colombini, and S.Avila, S. (2021). CIDEr-R: Robust consensus-based image description evaluation. arXiv preprint arXiv:2109.13701.

[13] P. Anderson, B. Fernando, M. Johnson, S. Gould, S. (2016). Spice: Semantic propositional image caption evaluation. In Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14 (pp. 382-398). Springer International Publishing.

[14] J. Hessel, A. Holtzman, M. Forbes, R. L.Bras and Y. Choi, (2021). Clipscore: A reference-free evaluation metric for image captioning. arXiv preprint arXiv:2104.08718.

[15] https://iclr.cc/virtual_2020/poster_SkeHuCVFDr.html