Concept Detection and Caption Prediction of Radiology Images Using Convolutional Neural Networks

Prabavathy Balasundaram[1,†], Karthikeyan Swaminathan[1,†], Oviasree Sampath[1,†] and Pradeep Km[1,*,†]

[1]Department of CSE, SSN College of Engineering, Rajiv Gandhi Salai, Chennai, Tamil Nadu, India

Abstract

Automated interpretation of medical images promises to revolutionize traditional diagnostic approaches, rendering them not only more efficient but also significantly faster. For the Concept Detection task, a Multi-label CNN (Convolutional Neural Network) model is proposed which is capable of mapping a single image to multiple highly probable concepts. This model's versatility lies in its ability to discern various features within medical images, facilitating a nuanced understanding of complex visual data. For the Caption Prediction task, a CNN-LSTM (Convolutional Neural Network - Long Short-Term Memory) model is proposed to predict accurate captions for given images. This model harnesses the power of CNNs to extract salient visual features and combines it with the sequential processing capabilities of LSTM networks to generate contextually relevant and accurate descriptions. This working note paper presents the results of the Kaprov team at ImageCLEFmedical 2024 Image Captioning and its subtasks of concept detection and caption prediction.

Keywords

Multi-Label Classification, Multi-Label Binarizer, CNN (Convolutional Neural Network), CNN-LSTM (Convolutional Neural Network - Long Short-Term Memory), Radiology Images, UMLS (Unified Medical Language System) Concepts

1. Introduction

Radiology images, including X-rays, Computed Tomography scans (CT scans) and Magnetic Resonance Imaging (MRI), and ultrasounds, are fundamental tools in medical diagnostics. They provide detailed visual information that helps healthcare professionals diagnose, monitor, and treat various medical conditions. Interpreting these images requires extensive training and expertise, as radiologists must identify subtle patterns and anomalies that can indicate the presence of diseases or injuries.

Traditionally, classifying and captioning radiology images is a manual process carried out by experienced radiologists. This involves a systematic approach where radiologists *review images, analyze features, formulate impressions and generate reports.*

While traditional methods are highly accurate, they are also time-consuming and prone to variability based on individual expertise and experience. The growing number of radiology images from advanced imaging technologies and their widespread use in medicine is increasing the workload for radiologists, potentially causing delays in clinical workflows.

In response to these challenges, researchers have explored various traditional automated methods to assist in classifying and captioning radiology images. Some of these methods include:

- Pattern Recognition Algorithms: Employ algorithms to identify predefined patterns in images. These algorithms are often based on features like edges, textures, and shapes to detect abnormalities.
- Rule-Based Systems: Use a set of predefined rules and criteria to analyze images. These systems rely on expert knowledge encoded into rules that guide the interpretation process.

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

[*] Corresponding author.

[†] These authors contributed equally.

[‡] GitHub repository: https://github.com/pradeepkmaran/imageclef2024

prabavathyb@ssn.edu.in (P. Balasundaram); karthikeyan2210394@ssn.edu.in (K. Swaminathan); oviasree2210386@ssn.edu.in (O. Sampath); pradeep2210432@ssn.edu.in (P. Km)

- Statistical Models: Apply statistical techniques to model the relationships between image features and diagnostic outcomes. These models can help in identifying probable conditions based on image data

Recent advancements in the field of medical image captioning have made significant strides by leveraging deep learning techniques. Among these techniques, CNNs and LSTM networks are particularly notable [1]. These advanced methods have demonstrated impressive abilities to automatically generate descriptive captions for radiology images. They are adept in capturing the intricate details and contextual information embedded in medical images, thereby enhancing the interpretative process.

However, there are several challenges to address. One major issue is the lack of large, annotated datasets for training. Another is understanding how these complex models make decisions, known as model interpretability. Additionally, training and deploying these models require significant computational resources.

To address these challenges, the proposed approach integrates deep learning models with domain-specific knowledge bases, such as the Unified Medical Language System (UMLS) [2], which offers an extensive collection of medical terms and associations between various entities in this domain. By leveraging the strengths of both deep learning techniques and comprehensive medical knowledge bases, this approach aims to enhance the accuracy, contextual relevance, and interpretability of generated captions. Ultimately, this integration holds the promise of advancing the field of medical image captioning, making it a more reliable and efficient tool for clinical diagnostics and research.

2. Task and Dataset Description

There are two tasks that have been worked upon, the first one being the Images Concept Detection and the second one being the Caption Prediction Task.

The dataset used for these two tasks is from *ImageCLEF 2024* [3] [4]. The dataset for the concept detection task in *ImageCLEFmedical Caption 2024* [5] is structured into three subsets. The training set includes *70,108* radiology images used for developing and refining models. A separate validation set, comprising 9,972 radiology images, is employed to validate and optimize these models during development. Finally, the test set consists of *17,237* radiology images that are unseen during model development and validation, serving as the final evaluation benchmark for assessing the performance of concept detection and caption generation systems. The goal of the first task is to predict the concepts that the given radiology images belong to, out of the *1945* UMLS concepts provided. Evaluation is done using metrics like precision, recall, and their combinations to measure how well the set is covered. The goal of the second task in *ImageCLEFmedical Caption 2024* [5] is to develop systems that can automatically generate coherent and contextually relevant captions for medical images. Each image is accompanied by its respective captions. These captions accurately describe the visual content of the images, reflecting the medical scenarios shown.

This dataset dives deep into the intricate world of medical image analysis. It encompasses various imaging modalities such as X-rays, CT scans, MRIs, and ultrasounds. Each modality offers a unique perspective on the internal structures of the body, providing crucial information for diagnoses. This diversity not only reflects the breadth of clinical imaging techniques employed in medical practice but also challenges machine learning models to generalize across different imaging modalities. Moreover, many images in the dataset depict scenes where multiple medical concepts coexist, requiring models to discern and accurately label complex visual content. Furthermore, the captions linked with these images show a broad range in length and complexity, reflecting the detailed observations and nuanced interpretations by healthcare professionals.

Figure 1 depicts a representative image from the dataset, sourced from Muacevic et al. (2023) under CC BY license.

Figure 2 illustrates a frequency histogram of Concept Unique Identifiers (CUIs) used in the dataset, highlighting the distribution of concepts across the medical images.

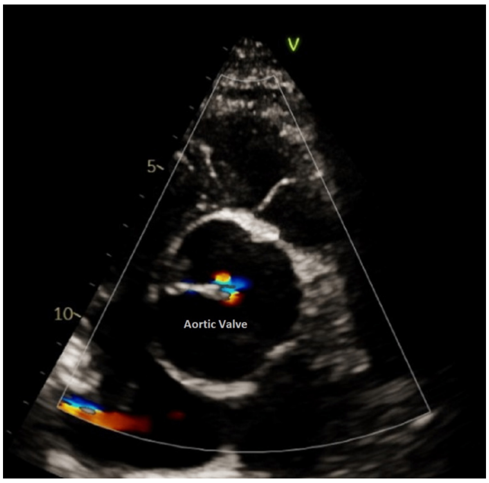

Figure 1: A sample image from the dataset. CC BY [Muacevic et al. (2023)].

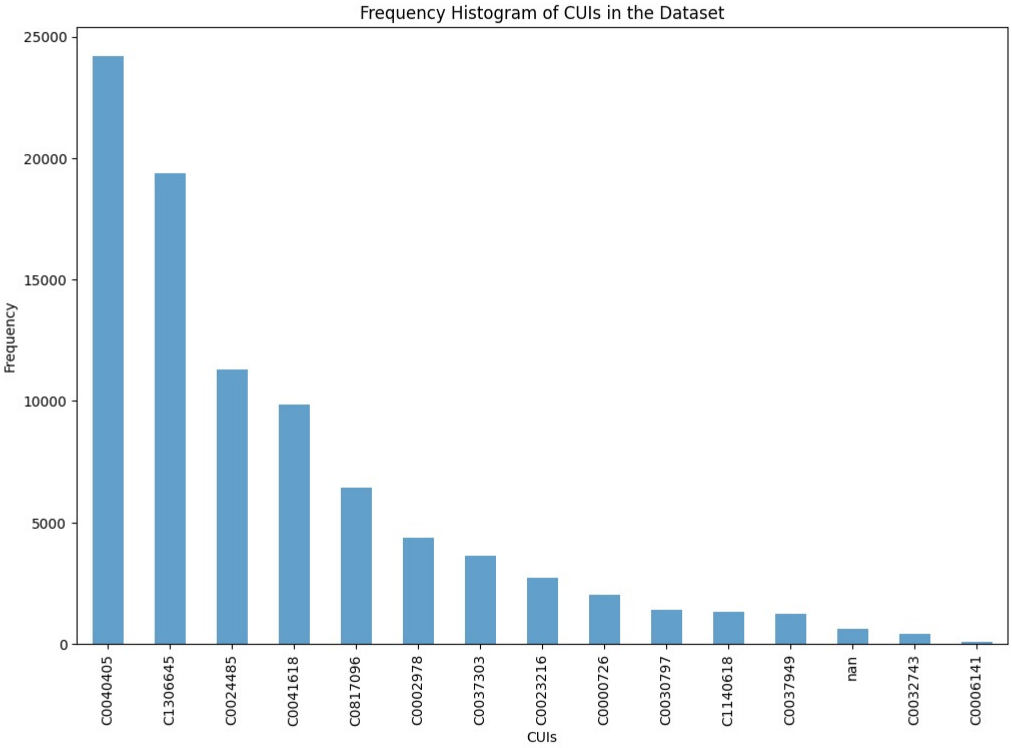

Figure 2: Frequency Histogram of CUIs in the dataset.

3. Data Pre-Processing

This section outlines the process of preparing data for concept detection and caption prediction.

3.1. Concept Detection

In concept detection, converting multi-label dataset into a machine learning-friendly format is crucial, as each image can be associated with multiple concepts.

Each instance can have multiple labels for different concepts, so it needs to be converted into a binary matrix format. The MultiLabelBinarizer from sklearn.preprocessing module is an effective tool for this task. It processes a list of lists, where each sublist contains the labels for a sample, and converts it into a binary matrix. In this matrix, each column represents a unique concept, and each row corresponds to a sample, with binary entries indicating the presence (1) or absence (0) of a concept.

3.2. Caption Prediction

The data pre-processing method for concept detection and image captioning involves several essential steps. Firstly, the VGG16 (Visual Geometry Group) [6] model, pre-trained on ImageNet, is used to extract features of the images using its penultimate layer. Training and test images are processed by loading, resizing to 224×224 pixels, converting to NumPy arrays, reshaping to fit the VGG16 input format, and preprocessed. The model then predicts features for each image, and these features are stored in dictionaries indexed by image IDs.

In parallel, the text captions related to these images are cleaned and standardized. This involves converting each caption to lowercase, removing non-alphabetical characters, and consolidating multiple spaces into one. Additionally, start and end tags are added to mark the beginning and end of each caption. The Tokenizer from keras.preprocessing.text module is used to tokenize the cleaned captions, convert words to unique integer indices and calculate the vocabulary size.

4. Methodologies Used

4.1. Multi-Label Image Classification using CNN with Batch Normalization

CNNs are well-suited for medical concept detection in image analysis [7] due to their ability to automatically learn and capture intricate patterns and structures from X-rays and CT scans. CNNs preserve spatial locality by using convolutional and pooling operations. These operations allow CNNs to capture how pixels in an image are related spatially, which is crucial for tasks like identifying patterns or structures in medical images such as X-rays and CT scans. CNNs are versatile in handling diverse input sizes and configurations, enabling accurate analysis of complex medical images and enhancing the reliability of concept detection systems in clinical settings.

The *Concept Detection Model* is designed for multi-label classification tasks [8], specifically handling images of size 32×32 with 3 color channels. Using 32×32 sized images in the *Concept Detection Model* balances computational efficiency with sufficient spatial detail for effective feature extraction by convolutional layers. It uses several convolutional layers with ReLU (Rectified Linear Unit) activation functions [9] to extract features, and batch normalization layers to improve training speed and stability by normalizing activations. Max pooling layers are used to reduce the size of the feature maps while keeping important information. Dropout layers help prevent overfitting. The model flattens the output from the convolutional layers and passes it through a dense layer with ReLU activation and batch normalization. The final layer uses sigmoid activation to predict probabilities for each of the 1945 classes. The model is compiled with the Adam (Adaptive Moment Estimation) optimizer and binary cross-entropy loss, suitable for multi-label classification where each label is independently predicted [10].

4.2. CNN-LSTM Fusion Caption Generator

The *Caption Prediction Model* combines image and text data to generate captions. First, it extracts image features using a dense layer with ReLU activation. At the same time, it converts text sequences into dense vectors to capture their meanings.

Next, it merges the image and text data by concatenating them, allowing the model to understand both inputs together. This combined data goes through an LSTM layer, which captures the sequence of words needed to create coherent captions. Dropout is used to prevent overfitting and improve the model's performance on new data.

The output from the LSTM is enhanced with the original image features, creating a strong link between the image and text.

This model is chosen for the seamless integration of visual and textual data. It begins with a dense layer that captures intricate details from medical images, ensuring effective feature extraction. The LSTM layers are essential for managing sequential data, creating coherent and logical text descriptions. Additionally, dropout regularization reduces overfitting, making the model more reliable with new data. This blend of techniques results in high-quality captions that are accurate and contextually meaningful.

5. Implementation

The *Concept Detection Model* is a Convolutional Neural Network (CNN) [Fig. 3], designed to classify images into 1945 different categories. The model utilizes a sequential architecture implemented with TensorFlow and Keras. It starts with a convolutional layer employing 16 filters of size 3×3, followed by a batch normalization layer to stabilize and accelerate the training process. This pattern of convolution, batch normalization, and pooling is repeated with more filters—32, 64, and 128 in the next convolutional layers—to capture increasingly complex features at each stage. Max pooling layers (2×2 size) reduce the size of the feature maps, making the model simpler and faster. Dropout layers with a rate of 0.25 are incorporated after each max pooling layer to prevent overfitting by randomly deactivating a fraction of neurons during training. After the convolutional and pooling layers, the model transitions to fully connected (dense) layers. The first dense layer has 64 units, followed by another batch normalization and dropout layer. This is followed by a dense layer with 128 units, again followed by batch normalization and dropout. The final dense layer uses a softmax activation function to calculate the probabilities for each of the 1945 classes. The model is compiled using the Adam optimizer and the categorical cross-entropy loss function. The model is trained with a batch size of 32 for 20 epochs. This extended training duration facilitates gradual adjustments to the model's weights using computed gradients from the training data. The EarlyStopping mechanism halts training if validation loss stagnates over three consecutive epochs, thereby mitigating overfitting and enhancing the model's ability to generalize to new data.

The *Caption Prediction Model* used for image caption prediction is the CNN-LSTM Fusion Caption Generator [Fig. 4], which integrates both dense and sequential data processing through a structured series of layers. It begins with two input layers: one for fixed-length dense data and another for variable-length sequential data. The dense data from the first layer passes through a dense layer with 256 units, serving as a feature extractor. Simultaneously, sequential data from the second layer undergoes embedding, converting each element into a 256-dimensional vector. These processed outputs are concatenated and fed into an LSTM layer with 256 units to capture temporal dependencies. Dropout layers are applied after the LSTM and subsequent dense layers to prevent overfitting. An additive connection merges the LSTM output with the initial dense layer's output to aid gradient flow. The final dense layer produces predictions tailored to the specific output requirements with 12,574 units. This model is implemented with a batch size of 32 and trained over 10 epochs. A batch size of 32 means that in each training iteration, the model processes 32 samples concurrently. This choice optimizes the utilization of computational resources like GPU memory while ensuring that the model's weight updates are stable and based on gradients computed from each batch.

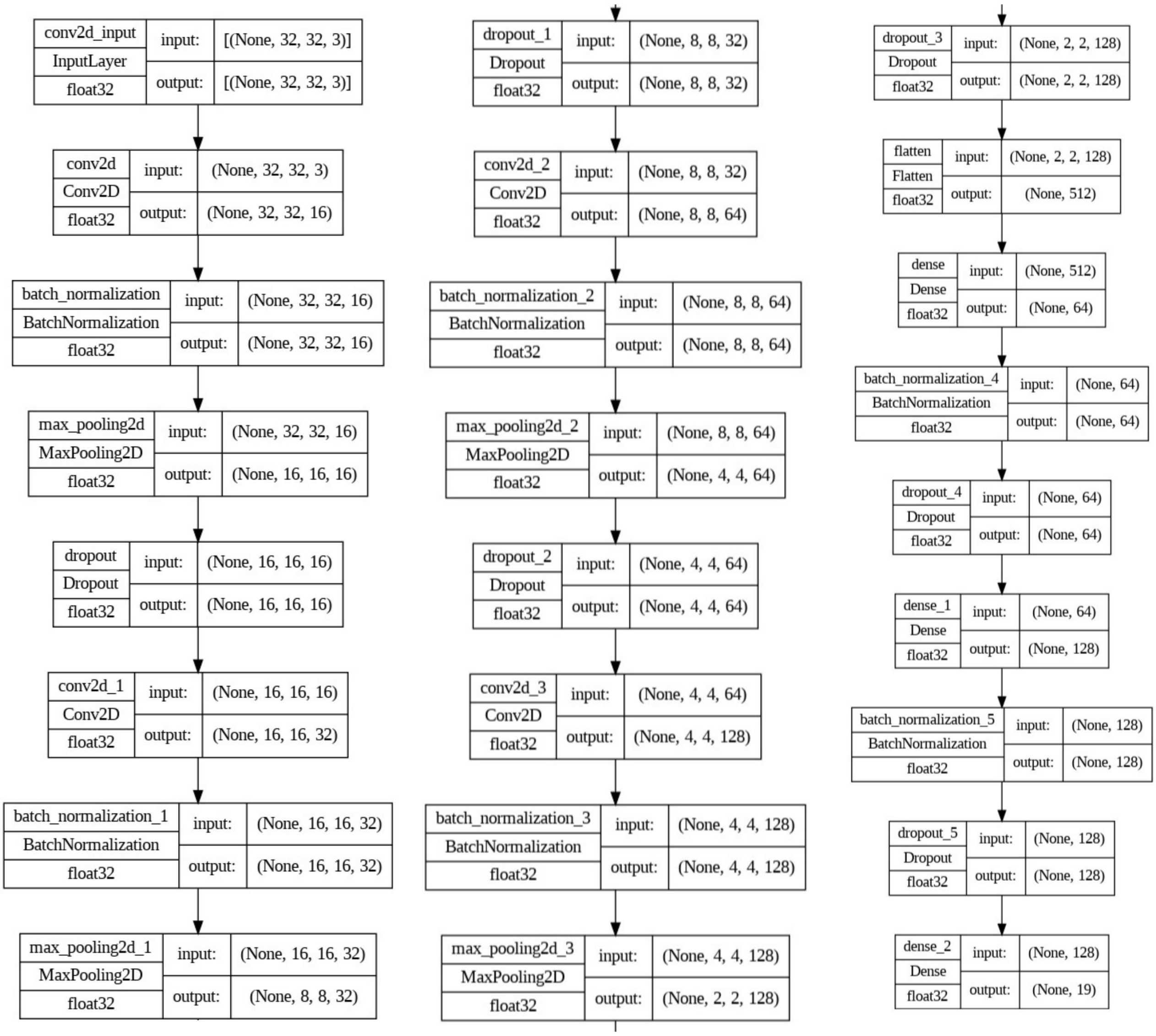

Figure 3: *Concept Detection Model* architecture.

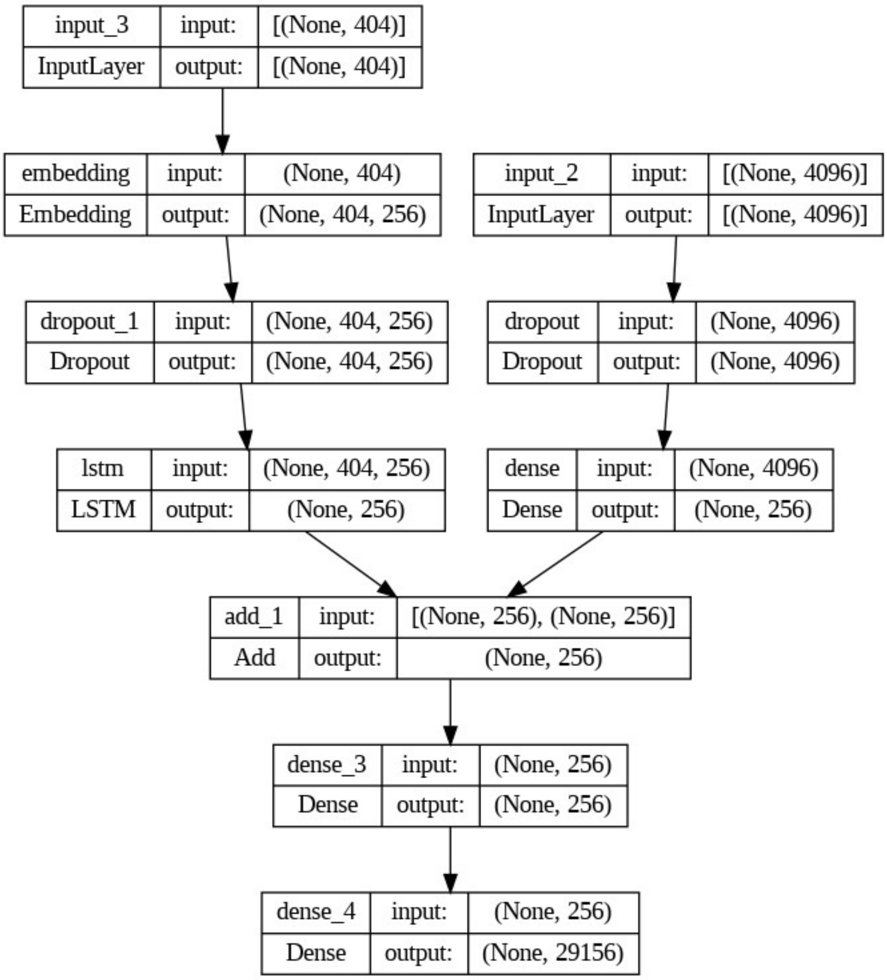

Figure 4: *Caption Prediction Model* architecture.

6. Result and Analysis

The *Concept Detection Model*, as shown in Table 1, demonstrated significant accuracy in identifying relevant concepts within radiology images. Evaluation metrics, including F1-score and F1-score manual, indicated that the model effectively handled the multi-label classification challenge, accurately mapping images to multiple concepts.

The *Caption Prediction Model*, detailed in Table 2 , exhibited strong capabilities in generating coherent and contextually relevant captions. Performance metrics such as BERTScore and ROUGE scores were used to evaluate the quality of the generated captions. The results showed that the model produced captions with high linguistic richness and relevance, closely aligning with human annotations.

Table 1
Task 1 (Concept Detection).

Metric	Value
F1-Score	0.4609
F1-Score Manual	0.7301

Table 2
Task 2 (Caption Prediction).

Metric	Value
BERTScore	0.5964
ROUGE	0.1905
BLEU-1	0.169
BLEURT	0.259
METEOR	0.06
CIDEr	0.107
CLIPScore	0.792
RefCLIPScore	0.787
ClinicalBLEURT	0.440
MedBERTScore	0.609

7. Conclusions

In both the Concept Detection and Caption Prediction tasks, the models showcased strong performance, demonstrating its ability to effectively analyze radiology images. For the Concept Detection task, it accurately identified relevant concepts, and endeavors to provide valuable support in medical diagnostics by mapping multiple concepts to each image. In the Caption Prediction task, the model generated meaningful and contextually appropriate descriptions, aligning well with the visual content of the images. Various metrics, including F1-Score, BERTScore, ClinicalBLEURT, CLIPScore, RefCLIPScore, BLEU-1, BLEURT, CIDEr, METEOR, and ROUGE underscored the models' comprehensive performance in producing linguistically rich and accurate captions.

Acknowledgments

We thank the CSE department of SSN College of Engineering for letting us utilize the GPU machine extensively to implement this task.

References

[1] Hartatik, H. Al Fatta, U. Fajar, Captioning image using convolutional neural network (cnn) and long-short term memory (lstm), in: 2019 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), 2019, pp. 263–268. doi:10.1109/ISRITI48646.2019.9034562.

[2] O. Bodenreider, The Unified Medical Language System (UMLS): integrating biomedical terminology, Nucleic Acids Research 32 (2004) D267–D270. URL: https://doi.org/10.1093/nar/gkh061. doi:10.1093/nar/gkh061.

[3] B. Ionescu, H. Müller, A.-M. Drăgulinescu, J. Rückert, A. B. Abacha, A. G. S. de Herrera, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, C. S. Schmidt, T. M. G. Pakull, H. Damm, B. Bracke, C. M. Friedrich, A.-G. Andrei, Y. Prokopchuk, D. Karpenka, A. Radzhabov, V. Kovalev, C. Macaire, D. Schwab, B. Lecouteux, E. Esperança-Rodier, W. wai Yim, Y. Fu, Z. Sun, M. Yetisgen, F. Xia, S. A. Hicks, M. A. Riegler, V. Thambawita, A. Storås, P. Halvorsen, M. Heinrich, J. Kiesel, M. Potthast, B. Stein, Overview of ImageCLEF 2024: Multimedia retrieval in medical applications, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the 15th International Conference of the CLEF Association (CLEF 2024), Springer Lecture Notes in Computer Science LNCS, Grenoble, France, 2024.

[4] J. Rückert, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, C. S. Schmidt, S. Koitka, O. Pelka, A. B. Abacha, A. G. S. de Herrera, H. Müller, P. A. Horn, F. Nensa, C. M. Friedrich, ROCOv2: Radiology Objects in COntext version 2, an updated multimodal image dataset, Scientific Data (2024). URL: https://arxiv.org/abs/2405.10004v1. doi:10.1038/s41597-024-03496-6.

[5] J. Rückert, A. Ben Abacha, A. G. Seco de Herrera, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, B. Bracke, H. Damm, T. M. G. Pakull, C. S. Schmidt, H. Müller, C. M. Friedrich, Overview of ImageCLEFmedical 2024 – Caption Prediction and Concept Detection, in: CLEF2024 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Grenoble, France, 2024.

[6] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, CoRR abs/1409.1556 (2014). URL: http://arxiv.org/abs/1409.1556.

[7] M. A. Hossain, M. S. A. Sajib, Classification of image using convolutional neural network (cnn), Global Journal of Computer Science and Technology 19 (2019) 13–14.

[8] Q. Wang, N. Jia, T. P. Breckon, A baseline for multi-label image classification using an ensemble of deep convolutional neural networks, in: 2019 IEEE International Conference on Image Processing (ICIP), 2019, pp. 644–648. doi:10.1109/ICIP.2019.8803793.

[9] C. Banerjee, T. Mukherjee, E. Pasiliao, An empirical study on generalizations of the relu activation function, in: Proceedings of the 2019 ACM Southeast Conference, ACM SE '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 164–167. URL: https://doi.org/10.1145/3299815.3314450. doi:10.1145/3299815.3314450.

[10] R. O. Ogundokun, R. Maskeliunas, S. Misra, R. Damaševičius, Improved cnn based on batch normalization and adam optimizer, in: O. Gervasi, B. Murgante, S. Misra, A. M. A. C. Rocha, C. Garau (Eds.), Computational Science and Its Applications – ICCSA 2022 Workshops, Springer International Publishing, Cham, 2022, pp. 593–604.