Concept Detection and Caption Prediction from Medical Images using Gradient Boosted Ensembles and Deep Learning

Mirunalini Palaniappan[1], Haricharan Bharathi[1], Eeswara Anvesh Chodisetty[1], Anirudh Bhaskar[1] and Karthik Desingu[1,*]

[1]*Sri Sivasubramaniya Nadar College of Engineering, Chennai, India*

Abstract

This paper outlines the contributions of our team in the annual ImageCLEFmedical Caption Task, which encompasses the Concept Detection and Caption Prediction sub-tasks. The concept detection sub-task focuses on automatically assigning appropriate medical concepts, based on Clinical Concept Unique Identifiers (CUIs), as tags of medical images. CUIs are unique identifiers assigned to medical concepts in the Unified Medical Language System (UMLS). They are based on a hierarchical structure and represent a standardized representation of various medical concepts, including diseases, anatomical structures, procedures, and more, while the caption prediction sub-task generates preliminary diagnostic captions for medical images, aiding medical professionals in preparing diagnostic reports. In the concept detection subtask, our approach involved using deep learning models to perform feature extraction, employing three distinct DenseNet models for feature extraction from the images. Subsequently, we utilized an XGBoost gradient boosting model to predict the Concept Unique Identifiers (CUIs) associated with a given image. In the caption prediction subtask, we used a model that utilizes a pre-trained InceptionV3 on the extended ROCO dataset to extract image features, which are then fed into a retrained LSTM model for caption generation. The method preprocesses the input image, extracts features using InceptionV3, and generates captions using the LSTM model through beam search.

Keywords

concept detection, caption prediction, natural language processing, computer vision ensemble, feature extraction, deep learning, automated image captioning, ResNet, EfficientNet, DenseNet, InceptionV3, LSTM, beam search.

1. Introduction

The ImageCLEFmedical Caption 2023 task [1] is the 7th edition of the caption task as a part of ImageCLEF 2023 [2]. Similar to previous the edition [3], the task consists of two subtasks: a Concept Detection task and a Caption Prediction task. The Concept Detection subtask involves recognizing and locating pertinent concepts within a vast collection of medical images. This is done by identifying the various Concept Unique Identifiers or CUIS from the Unified Medical

*Corresponding author.

miruna@ssn.edu.in (M. Palaniappan); haricharan2010267@ssn.edu.in (H. Bharathi); eeswaraanvesh2010038@ssn.edu.in (E. A. Chodisetty); anirudh2010094@ssn.edu.in (A. Bhaskar); karthik19047@cse.ssn.edu.in (K. Desingu)

0000-0001-6433-8842 (M. Palaniappan)

CEUR Workshop Proceedings (CEUR-WS.org)

Language System (UMLS) [4]. The Caption Prediction subtask involved generating cohesive captions that encompass the entire image. This is done by generating suitable captions based on the various CUIS generated from the first subtask.

Image captioning plays a crucial role in comprehending visual content. The groundbreaking research paper by Vinyals et al. [5] demonstrate the transformative potential of image captioning, showcasing how deep neural networks can generate accurate and coherent descriptions for visual content. With the surge in digital images and the need for automated image analysis, accurate and descriptive captions are essential. Image captioning enables improved searchability, context-aware information retrieval, and enhanced user experiences. The influential paper by Karpathy and Fei-Fe [6] demonstrates the effectiveness of the their proposed model, highlighting the power of deep visual-semantic alignments in generating high-quality image descriptions.

In the medical field, image captioning has emerged as a valuable tool facilitating comprehensive understanding and analysis of visual medical content. The Unified Medical Language System (UMLS) [4] plays a crucial role in solving the issue of concept detection in medical image analysis. By providing a comprehensive and standardized vocabulary of Clinical Concept Unique Identifiers (CUIs), UMLS enables accurate and consistent labeling of medical concepts within images. This allows for automatic assignment of appropriate medical concepts as tags to medical images, facilitating efficient retrieval, analysis, and interpretation of medical data. Also, the UMLS resolves inter-observer variability arising from differing concept identifications among doctors, providing a standardized vocabulary and unique identifiers. This promotes consistency, harmonization, and effective collaboration in medical image analysis, enhancing accuracy and reliability. Furthermore, generating error-free reports from medical images is of utmost importance in healthcare. Caption prediction models, when applied to medical images, can automatically generate textual descriptions or reports summarizing the content and findings within the images. The accuracy of these reports is vital for accurate diagnosis, treatment planning, and communication among healthcare professionals. By ensuring the generation of error-free reports, caption prediction models enhance patient care by reducing the potential for misinterpretation or miscommunication of critical information. This, in turn, improves the efficiency and effectiveness of medical decision-making, leading to better patient outcomes. Therefore, it is crucial for caption prediction models to generate error-free reports from medical images, contributing to enhanced healthcare delivery, and patient safety.

The accurate and descriptive captions generated by image captioning systems assist healthcare professionals in interpreting complex medical images and aid in diagnosis, treatment planning, and medical education. The paper by Alexander Selivanov et al. [7] contributed to the advancement of medical image captioning by demonstrating the efficacy of utilizing generative pretrained transformers, improving the generation of accurate and contextually relevant captions for medical images.

From the 2022 edition of the ImageCLEFMedical Caption task, we found that with respect to the first subtask, the AUEB-NLP-Group [8] achieved the highest primary F1-score with an ensemble of EfficientNetV2-B0 backbones, CMRE-UoG [9] proposed an image retrieval system with an ensemble of DenseNet-201, and the CSIRO group used an ensemble of DenseNet-161 with top 1% threshold optimization for multi-label classification. With respect to the Caption Prediction subtask, we found that the IUST_NLPLAB [10] team achieved the highest scores in caption prediction subtask, surpassing competitors significantly, using a multi-label

classification system, while the AUEB-NLP-Group [8] and CSIRO [11] teams also presented competitive results with their respective models based on Show and Tell and CvT-21 with DistilGPT2. A Python implementation of the experiments described in this paper will be made available at: https://github.com/karthik-d/ImageCLEFmedical-captioning-2023.

2. Related Work

In the field of caption prediction, several notable works have explored innovative approaches to generate descriptive and contextually relevant captions for images. The seminal paper by Vinyals [12] introduces an encoder-decoder framework using CNN and RNN for image captioning. It achieves state-of-the-art performance by learning the correlation between image features and descriptive captions. The paper's contribution lies in automating accurate and contextually relevant caption generation for images. The pioneering research paper by Karpathy and Fei-Fe [6] introduces a model that utilizes deep neural networks to generate accurate and coherent descriptions for images. It demonstrates the effectiveness of visual-semantic alignments in producing high-quality image captions. The approach improves the searchability, context-aware information retrieval, and user experiences related to image content. This paper showcases the transformative potential of deep learning techniques in advancing caption prediction for enhanced image understanding and analysis. The influential paper by Anderson et al. [13] introduces a two-stage attention mechanism for image captioning. The bottom-up attention identifies salient image regions, while the top-down attention generates contextually relevant words based on the visual and linguistic context. This approach improves the quality of generated captions by focusing on relevant image regions and incorporating both visual and language information. The model's effectiveness has been demonstrated through state-of-the-art performance on various captioning benchmarks, showcasing its application in generating more accurate and descriptive image captions. You et al. [14] propose a novel approach to image captioning by incorporating semantic attention mechanisms. The model dynamically attends to relevant regions in the image while generating captions, resulting in more accurate and contextually rich descriptions. This paper enhances the caption prediction process by improving the alignment between visual and semantic information. In the research done by Kelvin Xu [15], an attention-based model is introduced, leveraging techniques from machine translation and object detection, to automatically generate descriptive captions for images, achieving state-of-the-art performance on benchmark datasets and demonstrating the model's ability to focus on salient objects during caption generation.

Specifically in the medical field, the aforementioned paper by Selivanov et al. [7] addresses the limitations of existing models in medical image captioning and proposes a new architecture that combines two language models with image and text attention mechanisms. The proposed approach outperforms current state-of-the-art models and introduces a new preprocessing pipeline for radiology reports, leading to higher natural language generation metrics. The results demonstrate the effectiveness of the proposed methods in generating descriptive and informative captions for medical images, particularly in chest X-Ray image captioning. The combination of language models and the use of the GPT-3 model show significant improvements in text generation scores. Furthermore, it suggests that large language models (LLMs) can play

a crucial role in enhancing the performance of report generation from image features detected through convolutional image models.

In conclusion, the field of caption prediction has witnessed significant advancements through innovative approaches proposed in several influential papers. These works have demonstrated the effectiveness of various techniques, such as encoder-decoder frameworks, deep neural networks, attention mechanisms, and semantic attention mechanisms, in generating descriptive and contextually relevant captions for images. The application of these methodologies extends to diverse domains, including image understanding, content retrieval, and multimedia captioning systems. These advancements pave the way for enhanced image understanding, content retrieval, and support medical decision-making through accurate and informative image captions.

2.1. Previous Iterations of ImageCLEF for Medical Image Captioning

Since the 2021 edition of ImageCLEFMedical caption task[16], the team had noticed in the concept detection subtask that different teams had employed deep learning models, such as DenseNet, Inception-V3, and MobileNet-v2, either as multi-label classifiers or in information retrieval-oriented solutions using image embeddings. In the 2021 edition, more modern architectures like EfficientNets and Visual Transformers (ViT) were introduced, resulting in improved F1-scores compared to previous years. In the caption prediction subtask, teams had utilized variations of the Show, Attend and Tell model, incorporated Transformer-based architectures, and explored the use of general language models like GPT-2, while finding that simple architectures outperformed pretraining with medically oriented datasets. With respect to the ImageCLEFMedical Caption task 2022 [3], the following results were inferred from the concept detection subtask :

- The AUEB-NLP-Group [8] achieved the best performance in the concept detection task with an ensemble of two EfficientNetV2-B0 backbones and a single classification layer, using the union of predicted concepts for the ensemble.
- The CMRE-UoG team [9] proposed an image retrieval system with an ensemble of five DenseNet-201 models, retrieving 100 different images each and assigning the union of predicted CUIs to each image.
- The CSIRO group [11] experimented with multiple backbones and their best approach involved an ensemble of 43 DenseNet-161 models with top-1% threshold optimization for multi-label classification.
- The SSN MLRG team [17] employed DenseNet for multi-label classification and an information retrieval system for their caption prediction model.

In the second subtask, namely Caption Prediction, the following were the major takeaways from the 2022 edition of the contest :

- The IUST_NLPLAB team [10] employed a multi-label classification system based on ResNet-50, treating each word as a label and assigning 26 words in order of their probability to each image, resulting in their superior performance with a BLEU score of 0.4828.

- The AUEB-NLP-Group [8] utilized the Show and Tell model, consisting of a CNN-RNN encoder-decoder with an EfficientNet-B0 backbone, achieving a BLEU score of 0.3222, demonstrating competitive performance in various evaluation metrics.
- The CSIRO group [11] experimented with different encoder-to-decoder models and achieved their best results using CvT-21 as the encoder and DistilGPT2 as the decoder.They obtained the overall best BERTScore of 0.6234, showcasing the effectiveness of their chosen model combination.
- The SSN MLRG team [17] employed a Sparse Auto Encoder (SAE) with a Multi-Layer Perceptron (MLP) and a Gated Recurrent Unit (GRU) for their caption prediction model.

Since the 2021 edition [16], the team had noticed different teams had employed deep learning models, such as DenseNet, Inception-V3, and MobileNet-v2, either as multi-label classifiers or in information retrieval-oriented solutions using image embeddings. In the 2021 edition, more modern architectures like EfficientNets and Visual Transformers (ViT) were introduced, resulting in improved F1-scores compared to previous years. Based on the above results, our team decided to use DenseNet models as feature extractors as it gave promising results in the previous year and feed the feature vectors to a gradient boosting model to detect the CUIS. For the caption prediction subtask, the team decided to go with an InceptionV3 model pretrained and then finetuned to extract the feature vectors. A LSTM-based caption generation model is employed to generate captions for the given images, incorporating the beam search algorithm to explore multiple possible captions and select the most probable one.

3. Methods

The following section explains in detail the systems that were utilized in our submissions for the Concept Detection and Caption Prediction sub-tasks. A Python implementation of the experiments described in this paper will be made available at: https://github.com/karthik-d/ImageCLEFmedical-captioning-2023.

3.1. Dataset

The current edition of the task consisted of a total of 60,918 images allocated to the training set, 10,437 images in the validation set, and 10,473 images in the testing set. These images were selected from the extended and revised version of the Radiology Objects in Context (ROCO) dataset, which is derived from biomedical articles available in the PMC OpenAccess subset.

The images used in the Concept Detection task were created using the UMLS 2022 AB release. These images were subsequently filtered based on their semantic type. Building on a suggestion from the previous year, concepts with low frequency were eliminated in the current implementation. The captions used in the study were obtained from annotated medical literature, and any hyperlinks present in the original text were excluded or removed.

3.2. Concept Detection

For Concept Detection, the team decided to use three different models of DenseNet [18] as feature extractors, namely DenseNet-121, DenseNet-169 and DenseNet-201. The feature vectors

where then given as input to an XGBoost ensemble classifier to predict the class labels for a given observation.

Feature extraction is a technique used in image processing to extract pertinent and distinct visual elements from photographs. In order to discover and record characteristic patterns, textures, shapes, or colours that are indicative of certain objects, structures, or concepts, image data must first be analysed. When performing concept recognition or classification tasks, these extracted features act as compact representations of the images and are put into a booster method like XGBoost. In order to understand complicated patterns and correlations, the booster uses the extracted features. This enables precise prediction and decision-making based on the visual properties present in the images.

Multiple deep learning architectures were considered for the concept detection task. Deep convolutional neural networks known as architectures were first presented by Huang et al. in 2017 [18] and are distinguished by their densely linked layers. By creating dense connections across layers, the DenseNet design tries to solve the vanishing gradient issue and encourage feature reuse. ResNet [19] , short for Residual Network, is a deep convolutional neural network architecture introduced by He et al. in 2015. ResNet designs are renowned for their creative use of residual connections, which, by overcoming the degradation issue that arises with adding more layers, allow for the training of very deep networks. EfficientNet [20] is a family of deep convolutional neural network architectures introduced by Tan et al. in 2019. The key idea behind EfficientNet is to achieve state-of-the-art performance with high efficiency in terms of both computational resources and model size. EfficientNet models have been designed using a compound scaling method that balances the network depth, width, and resolution to achieve optimal performance.

Ensemble methods [21] are techniques employed to combine multiple model to produce improved results. They boast higher accuracy scores than the individual models themselves. Boosting is a prominent ensembling technique used wherein new models are added to the existing features of the model to correct errors. Our solution adopted a gradient-boosting ensemble approach for concept detection to identify the various labels for a given image. XGBoost [22] is an implementation of gradient boosted decision trees designed for speed and performance. The authors decided to implement the XGBoost library package above all the other boosters because of its higher execution performance.

The choice of DenseNet models as the base learners for the XGBoost ensemble was motivated by several factors. We observed that EfficientNet and the ResNet models had significantly higher losses and lower accuracy as compared to the DenseNet models. The superior performance exhibited by the DenseNet models made them a compelling option for inclusion in the XGBoost ensemble. The aim of constructing an ensemble model is to leverage the strengths of individual base learners and mitigate their weaknesses through collective decision-making. The DenseNet models were favored due to its architecture's dense connectivity and its ability to encourage complementary learning make it well-suited for ensemble learning. Since each of the models have its own unique perspective on the data, the ensemble can exploit their diverse strengths and compensate for individual weaknesses.

Table 1
Model training scores.

Model	Epochs	Accuracy	Loss
DenseNet121	85	0.212	65.361
DenseNet169	50	0.198	6752.75
DenseNet201	31	0.186	36.022
EffecientNetB2	28	0.191	77810.516
ResNet101	10	0.062	11320.178

3.3. Caption Prediction

The UMLS utilizes standardized vocabulary (CUIs) to accurately label medical concepts in image analysis, enhancing retrieval and interpretation. Caption prediction models generating error-free reports improve patient care, facilitating precise diagnosis and treatment planning while minimizing miscommunication risks. By bridging concept detection gaps and ensuring reliable reports, UMLS and caption prediction models contribute to enhanced healthcare delivery and patient safety.

For Caption Prediction, our team decided to use a fine-tuned InceptionV3 model to extract the features, which were given as input to an LSTM model that generated captions based on those features. Beam search was used to explore multiple possible sequences of words and select the most likely caption based on the model's predictions and the specified beam index.

InceptionV3 [23] is a convolutional neural network architecture that was introduced by Google in 2015. Its primary purpose is to perform image classification tasks and it has gained significant popularity in various computer vision applications. The fundamental concept behind InceptionV3 revolves around utilizing inception modules, which enable the network to effectively capture features at different spatial scales. These modules comprise parallel convolutional layers with varying sizes, enabling the network to learn both local and global features. Additionally, InceptionV3 incorporates techniques like batch normalization and regularization to enhance training and generalization. The model is pretrained on a large dataset, such as ImageNet [24], and can be fine-tuned for specific tasks by replacing the final classification layer. InceptionLSTM is a neural network architecture used for image captioning tasks. It combines the InceptionV3 convolutional neural network with a Long Short-Term Memory (LSTM) recurrent neural network. The InceptionV3 model extracts visual features from input images, while the LSTM generates a sequence of words as captions based on those features. This architecture enables the model to capture both visual and semantic information, resulting in meaningful and contextually relevant image captions.

Beam search is a decoding algorithm commonly used in sequence generation tasks, such as machine translation and image captioning. It explores multiple possible sequences of words by maintaining a beam of the most likely candidates at each decoding step. The beam width or beam size determines the number of candidates retained at each step. Beam search is applied during the caption generation process to select the most likely captions based on the model's predictions. It helps to generate more diverse and accurate captions by considering multiple hypotheses simultaneously and choosing the one with the highest probability. The above code

used beam search to generate captions based on the fine-tuned InceptionV3 and LSTM model's predictions, allowing it to produce more accurate and contextually relevant captions for the given images.

4. Experiments

4.1. Concept Detection

The detection process was divided into two stages: Feature Extraction and Boosting. In Feature Extraction, transfer learning techniques were used with pre-trained models such as ResNet-101, EfficientNet-B2, DenseNet-121, DenseNet-169, and DenseNet-201. The Adam Optimizer was used for all models, and the validation loss was monitored. Since each image had multiple labels, a Label Encoder was used to assign a unique label to each CUI ID. This was then converted into a multi-hot encoded array using a Multi-Label Binarizer. Both the input image and encoded labels were given as input to all the models. However, ResNet-101 and EfficientNet-B2 had high validation loss, so they were excluded as feature extractors. We continued with the DenseNet models, where we extracted the core architecture and added an additional dense and flatten layer with 4096 nodes to represent the feature vectors. The initial DenseNet models were trained with input images, and the weights of the core architecture were frozen. Then, the models were retrained after adding the new dense layer for all three DenseNet architectures. Categorical Cross Entropy loss was used to monitor model performance.

In the second stage, XGBoost was used. The feature vectors extracted from DenseNet models (DenseNet-121, DenseNet-169, and DenseNet-201) were fed into XGBoost. The predicted outputs were plotted on a Receiver Operator Characteristics (ROC) curve, and a probability threshold was determined. This threshold was used to truncate the predicted labels to obtain outputs.

Categorical cross entropy is a metric that quantifies the disparity between two discrete probability distributions. The Softmax activation function is employed at the output layer to generate a probability distribution across all classes. Softmax is a mathematical function that transforms a vector of numbers into a vector of probabilities. Each probability corresponds to the relative magnitude of its corresponding value in the vector. In essence, it normalizes the outputs, converting them from weighted sums to probabilities that add up to one.

Table 2
Model training parameters used to train each of the convolutional neural networks used for this classification task.

Parameter	Optimizer	Learning rate	Batch size	Epochs
DenseNet121	Adam	$1e{-}8$	8	85
DenseNet169	Adam	$1e{-}8$	8	50
DenseNet201	Adam	$1e{-}8$	32	31
EffecientNetB2	Adam	$1e{-}5$	8	28

4.2. Caption Prediction

For the Caption Prediction subtask, the proposed idea involed using a pretrained InceptionV3 model which was fine-tuned as feature extractor to give the feature vectors as input to an InceptionLSTM model to generate the captions. Furthermore, the beam search algorithm was employed to explore and evaluate multiple potential sequences of words for generating captions. It involved considering different word combinations based on the model's predictions and a specified beam index. The aim was to identify the most probable caption by considering a limited number of top-scoring candidates at each step of the caption generation process.

The training workflow began by resizing them to a fixed size of 299x299 pixels to ensure consistency. The captions were tokenized into individual words and encoded into numerical sequences using a tokenizer. Two models were utilized in the training procedure: a pre-trained InceptionV3 convolutional neural network (CNN) and a Long Short-Term Memory (LSTM) recurrent neural networks. The pre-trained InceptionV3 model was employed to extract visual features from the input images. Each image was passed through the InceptionV3 model, and the output of the second-to-last layer was extracted as the visual feature representation. The training process involved generating captions for the input images. Initially, a start token was provided as the input to the LSTM model, which then predicted the next word in the caption sequence. This process was repeated iteratively, with the predicted word being fed back into the LSTM model as input for the next iteration. The objective was to maximize the probability of generating the ground truth captions for the given images. To train the LSTM model, a loss function was defined to measure the discrepancy between the predicted captions and the ground truth captions. The loss function utilized was typically the cross-entropy loss. The weights of the LSTM model were updated through backpropagation using an optimizer, such as Adam, to minimize the loss. The training procedure involved iterating over the entire dataset multiple times, known as epochs. In each epoch, the dataset was randomly shuffled to introduce diversity during training. The model was trained in mini-batches, where a subset of images and their corresponding captions were fed into the model simultaneously. The gradients computed during backpropagation were accumulated over the mini-batches, and the model's weights were updated after each batch. Various hyperparameters were tuned to optimize the performance of the model. These included the learning rate, batch size, number of LSTM units, and the maximum length of the generated captions. Different combinations of hyperparameters were experimented with to find the optimal configuration. After each epoch, the model's performance on the validation set was evaluated and early stopping was employed to avoid overfitting.

5. Conclusion

The concept detection sub-task gave a F1-score of 0.0173 and F1-score manual of 0.1172. The gradient boosting approach has a very low F1-score as the model predicted multiple CUI's for a single image hinting at the fact that all feature vectors extracted from the images were given equal importance. This can be overcome by choosing the most important feature vectors that represent the data and generating the appropriate CUI's by feeding them to the XGBoost model. The approach for caption prediction gave a BERTScore of 0.6019 which ranked seventh on the leaderboards. The authors were able to achieve a very high CLIPScore of 0.7759. CLIPScore

tried to mimic human judgement and gives a score based on the compatibility of image and caption pair. A very low METEORScore of 0.0615 indicated that the generated captions had many Grammatical errors and were not fluent. These shortcomings can be overcome by using a more well trained Language model which has been trained on more data and by fine tuning the architecture. Post editing and human feedback can also be given to train better Language models. To better the approach more complex deep learning architectures can be used for feature extraction and transformers can be deployed in the future which would be more robust.

Acknowledgments

The authors would like to express their gratitude to the Machine Learning Research Group (MLRG), Sri Sivasubramaniya Nadar College of Engineering, Chennai, India (https://www.ssn.edu.in/) for providing the GPU resources for model training and testing.

References

[1] J. Rückert, A. Ben Abacha, A. G. Seco de Herrera, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, H. Müller, C. M. Friedrich, Overview of ImageCLEFmedical 2023 – Caption Prediction and Concept Detection, in: CLEF2023 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece, 2023.

[2] B. Ionescu, H. Müller, A. Drăgulinescu, W. Yim, A. Ben Abacha, N. Snider, G. Adams, M. Yetisgen, J. Rückert, A. Garcıa Seco de Herrera, C. M. Friedrich, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, S. A. Hicks, M. A. Riegler, V. Thambawita, A. Storås, P. Halvorsen, N. Papachrysos, J. Schöler, D. Jha, A. Andrei, A. Radzhabov, I. Coman, V. Kovalev, A. Stan, G. Ioannidis, H. Manguinhas, L. Ştefan, M. G. Constantin, M. Dogariu, J. Deshayes, A. Popescu, Overview of ImageCLEF 2023: Multimedia retrieval in medical, socialmedia 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, 2023.

[3] J. Rückert, A. Ben Abacha, A. Garcia 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: CEUR Workshop Proceedings, volume 3180, CEUR Workshop Proceedings, 2022, pp. 1294–1307.

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

[5] O. Vinyals, A. Toshev, S. Bengio, D. Erhan, Show and tell: A neural image caption generator, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3156–3164.

[6] A. Karpathy, L. Fei-Fei, Deep visual-semantic alignments for generating image descriptions, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3128–3137.

[7] A. Selivanov, O. Y. Rogov, D. Chesakov, A. Shelmanov, I. Fedulova, D. V. Dylov, Medical image captioning via generative pretrained transformers, Scientific Reports 13 (2023) 4171.
[8] F. Charalampakos, G. Zachariadis, J. Pavlopoulos, V. Karatzas, C. Trakas, I. Androutsopoulos, Aueb nlp group at imageclefmedical caption 2022 (2022).
[9] F. Dalla Serra, F. Deligianni, J. Dalton, A. Q. O'Neil, Cmre-uog team at imageclefmedical caption 2022: Concept detection and image captioning (2022).
[10] M. Hajihosseini, Y. Lotfollahi, M. Nobakhtian, M. M. Javid, F. Omidi, S. Eetemadi, Iust_nlplab at imageclefmedical caption tasks (2022).
[11] L. Lebrat, A. Nicolson, R. Santa Cruz, G. Belous, B. Koopman, J. Dowling, Csiro at imageclefmedical caption 2022 (2022).
[12] O. Vinyals, A. Toshev, S. Bengio, D. Erhan, Show and tell: A neural image caption generator, 2015. arXiv:1411.4555.
[13] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, L. Zhang, Bottom-up and top-down attention for image captioning and visual question answering, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 6077–6086.
[14] Q. You, H. Jin, Z. Wang, C. Fang, J. Luo, Image captioning with semantic attention, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 4651–4659.
[15] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, Y. Bengio, Show, attend and tell: Neural image caption generation with visual attention, in: International conference on machine learning, PMLR, 2015, pp. 2048–2057.
[16] O. Pelka, A. Ben Abacha, A. G Seco de Herrera, J. Jacutprakart, C. M. Friedrich, H. Müller, Overview of the imageclefmed 2021 concept & caption prediction task, in: Proceedings of the CLEF 2021 Conference and Labs of the Evaluation Forum-working notes, 21-24 September 2021, 2021.
[17] S. S. N. Mohameda, K. Srinivasanb, Ssn mlrg at imageclefmedical caption 2022: Medical concept detection and caption prediction using transfer learning and transformer based learning approaches (????).
[18] G. Huang, Z. Liu, L. van der Maaten, K. Q. Weinberger, Densely connected convolutional networks, 2018. arXiv:1608.06993.
[19] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, 2015. URL: https://arxiv.org/abs/1512.03385. doi:10.48550/ARXIV.1512.03385.
[20] M. Tan, Q. V. Le, Efficientnet: Rethinking model scaling for convolutional neural networks (2019). URL: https://arxiv.org/abs/1905.11946. doi:10.48550/ARXIV.1905.11946.
[21] Z.-H. Zhou, J. Wu, W. Tang, Ensembling neural networks: many could be better than all, Artificial intelligence 137 (2002) 239–263.
[22] T. Chen, C. Guestrin, XGBoost, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2016. URL: https://doi.org/10.1145%2F2939672.2939785. doi:10.1145/2939672.2939785.
[23] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, 2015. arXiv:1512.00567.
[24] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, Communications of the ACM 60 (2017) 84–90.