AUEB NLP Group at ImageCLEFmedical Caption 2024

Notebook for the AUEB NLP Group at ImageCLEFmedical Caption 2024

Marina Samprovalaki[1,*,†], Anna Chatzipapadopoulou[1,*,†], Georgios Moschovis[1,2,*,†], Foivos Charalampakos[1,*,†], Panagiotis Kaliosis[1,*,†], John Pavlopoulos[1,2] and Ion Androutsopoulos[1,2]

[1] *Department of Informatics, Athens University of Economics and Business, 76, Patission Street, GR-104 34 Athens, Greece*
[2] *Archimedes Unit, Athena Research Center, 1, Artemidos Street, GR-151 25 Athens, Greece*

Abstract

This article describes the approaches that the AUEB NLP Group experimented with during its participation in the 8th edition of the ImageCLEFmedical Caption evaluation campaign, including both Concept Detection and Caption Prediction tasks. The objective of Concept Detection is to automatically categorize biomedical images into a set of one or more concepts. In contrast, the Caption Prediction task focuses on generating a precise and meaningful diagnostic caption that describes the medical conditions depicted in the image. Building on our prior research for the Concept Detection task, we utilized a diverse set of Convolutional Neural Network (CNN) encoders, followed by a Feed-Forward Neural Network. Additionally, we implemented two versions of the retrieval-based k-NN algorithm: a version that assigned concepts based on statistical frequency and a weighted version that took into account the order of the retrieved neighbors. Both models used the CNN image encoders to improve their retrieval capabilities. Regarding the Caption Prediction task, we fine-tuned the InstructBLIP model to generate initial captions and then enhanced it by employing rephrasing techniques with further pre-trained models. We also used synthesizing techniques that incorporated information from similar neighboring images in the training set to refine these captions. Additionally, we employed “Distance from Median Maximum Concept Similarity” (DMMCS), a novel guided-decoding approach that drives the model’s behaviour throughout the decoding process, aiming to integrate information from the predicted concepts of Concept Detection. We explored the application of DMMCS to all of our developed systems. Our group ranked 2nd in Concept Detection and 4th in Caption Prediction.

Keywords

Natural Language Processing, Computer Vision, Biomedical Images, Convolutional Neural Networks, Multi-Label Classification, Caption Generation, Generative Models, Transformers, Deep Learning,

1. Introduction

ImageCLEF [1] is an ongoing evaluation initiative, first run in 2003 as part of the Cross Language Evaluation Forum (CLEF)[1], that promotes the evaluation of technologies for annotation, indexing, classification, and retrieval of multi-modal data. ImageCLEFmedical is one of the four main tasks in this year’s ImageCLEF campaign. We participated in the ImageCLEFmedical Caption task, which was organized for the eigth time [2]. As in previous years, the task comprised two sub-tasks: Concept Detection and Caption Prediction.

The objective of Concept Detection is to accurately associate a biomedical image with one or more relevant medical concepts (tags), while in Caption Prediction, the goal is to automatically generate a

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

[*] Corresponding author.

[†] These authors contributed equally.

mar.samprovalaki@aueb.gr (M. Samprovalaki); annachatzipap@gmail.com (A. Chatzipapadopoulou); geomos@aueb.gr (G. Moschovis); phoebuschar@aueb.gr (F. Charalampakos); pkaliosis@aueb.gr (P. Kaliosis); annis@aueb.gr (J. Pavlopoulos); ion@aueb.gr (I. Androutsopoulos)

https://www.linkedin.com/in/marina-samprovalaki/ (M. Samprovalaki); https://www.linkedin.com/in/anna-chatzipapadopoulou/ (A. Chatzipapadopoulou); https://geomos.sites.aueb.gr/ (G. Moschovis); https://pkaliosis.github.io (P. Kaliosis); https://ipavlopoulos.github.io/ (J. Pavlopoulos); https://www.aueb.gr/users/ion/ (I. Androutsopoulos)

0000-0003-0547-0581 (G. Moschovis); 0000-0001-9188-742 (J. Pavlopoulos)

[1] https://www.imageclef.org/2024, Last accessed: 2024-06-20

preliminary diagnostic report that accurately describes the medical findings, as well as the anatomy of the body structures and organs shown in the image. Diagnostic Captioning remains a challenging research problem aimed at assisting the diagnostic process for patients by providing a preliminary report, rather than replacing medical professionals involved in the procedure [3]. It can thus be seen as an assistive tool, capable of producing an initial draft diagnosis regarding the patient's condition. Such a document would ideally allow doctors to focus on critical areas of the image [4] and help them produce more precise medical diagnoses at an increased speed [5]. Experienced clinicians could enhance their throughput by analyzing the large volume of daily medical examinations more quickly and efficiently. Less experienced clinicians could consider the automatically generated captions to reduce the likelihood of clinical errors [6]. Concept Detection can further improve Diagnostic Captioning by identifying key concepts that should be included in the draft report. We demonstrate the connection between the two sub-tasks by using "Distance from Median Maximum Concept Similarity" (DMMCS)[2] [7], which employs information derived from our Concept Detection systems in order to improve the performance of our Caption Prediction systems.

1.1. AUEB NLP Group contributions

In this work, we present the experiments conducted and the systems submitted as part of the AUEB NLP Group's participation in this year's Concept Detection and Caption Prediction tasks. We used a number of new approaches influenced by the remarkable progress in the field of NLP and based on instruction-tuned Large Language Models (LLMs) [8].

Our submissions to the Concept Detection sub-task are based on two distinct approaches. We used a Convolutional Neural Network (CNN) encoder to extract visual features from the medical images. In the first approach, these features were fed into a Feed-Forward Neural Network (FFNN) to classify the images into various medical concepts. In the second approach, we implemented a separate method using a k-nearest neighbors (k-NN) algorithm. In this approach, k neighbors are first retrieved, and the most frequently occurring concepts among these neighbors are selected.

Regarding the Caption Prediction sub-task,we tried five main approaches. First, we employed an InstructBLIP model [9] that was fine-tuned on the specified dataset [10] to generate an initial set of captions, which were then also used in the other four approaches. In the second approach, we enhanced the initial captions by drawing insights from captions of similar images and training a FLAN-T5 model [11] to refine them [12, 13]. The third approach was similar, but instead of FLAN-T5, we employed ClinicalT5 [14], which is pre-trained on numerous medical datasets, in order to rephrase and correct the initial captions produced by InstructBLIP. The fourth approach involved integrating the DMMCS algorithm [7] in the language model's decoding process in order to promote the inclusion of a given set of keywords, which in this case where predicted by one of our Concept Detection systems. Lastly, we also applied DMMCS decoding to ClinicalT5 in order to maximize their efficacy and improve the overall caption quality. In all our models we used CNN encoders, since there are signs that vision transformers [15] still have inferior performance in visual tasks, such as classification and semantic segmentation [16], especially in medical image tagging [5, 17].

Extending our history of successful entries [18, 19, 20, 21, 22] in the ImageCLEFmedical campaign, our submissions ranked 2nd among 9 participating groups in the Concept Detection sub-task and 4th among 11 participating groups in the Caption Prediction sub-task. In Section 2, we provide insight into this year's dataset, followed by a discussion of our approaches in Section 3. In Section 4, we present our experimental results for each sub-task. Finally, in Section 5, we summarize our findings and suggest directions for future research.

All code used for our experiments is available on GitHub.[3]

[2] https://github.com/nlpaueb/dmmcs, Last accessed: 2024-06-20.

[3] https://github.com/nlpaueb/imageclef2024, Last accessed: 2024-06-20.

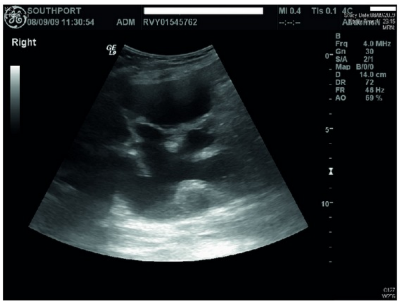
2. Data

In this year's edition of the ImageCLEFmedical Caption task, the dataset is an updated and extended version of the Radiology Objects in Context (ROCO) dataset [10], which originates from biomedical articles of the PubMed Open Access (PMC OA) subset.[4].

This dataset, which is common for both sub-tasks, consists of 80,080 biomedical images along with their respective medical concepts, in the form of UMLS [23] terms[5], and diagnostic captions. The dataset was originally split by the organizers into training and validation subsets, with 70,108 radiology images in the first set and 9,972 in the latter. After merging the provided data, we split them again, this time into three subsets, in order to also obtain a development (private test) subset for evaluation purposes. We used a 75%-10%-15% training-validation-development split, keeping relatively equal concept distributions in all three subsets. Consequently, we obtained 64,928 images as our training data, 7,179 images as our validation set, while the remaining 7,973 images constituted our held-out development set. All of our submissions were also evaluated on the hidden official test set (ROCOv2) [24]. The test dataset utilizes Radiology Objects in COntext Version 2 (ROCOv2) [24], an updated and extended version of the ROCO dataset [10]. This set includes 17,237 previously unseen images.

2.1. Concept Detection

Concept Detection is a multi-label classification problem covering a broad range of 1,945 distinct biomedical concepts, originating from the Unified Medical Language System (UMLS) [23]. In this sub-task, the goal is to identify (assign) the distinct medical concepts (tags) depicted in each image (e.g., particular medical conditions). Among the available concepts (tag set), four are specific imaging modalities: X-Ray Computed Tomography, Ultrasonography, Magnetic Resonance Imaging (MRI), PET/CT scans. All concepts are represented by Concept Unique Identifiers (CUIs) following the UMLS standard. Some examples of images and their ground truth concepts can be found in Figure 1.



CUI	UMLS Term
C0041618	Ultrasonography
C0018827	Heart Ventricle
C1510420	Cavitation
CC BY [Magdás et al. (2021)]	

Figure 1: CC BY [Magdás et al. (2021)] from the ImageCLEFmedical2024 dataset, along with the corresponding CUIs and UMLS terms.

The distribution of concepts is highly skewed. Some concepts are present in more than $25,000$ images, whereas others are associated with only 1 image. Figure 2(a) depicts the long-tail distribution of the entire (development + validation + train) dataset, as shown in the left plot, where the frequencies of the concepts (number of images each concept is associated with) are plotted in descending order against their respective class indices. After conducting a comprehensive exploratory analysis of this year's dataset, we found that certain concepts were more prevalent (Table 1); these mostly correspond

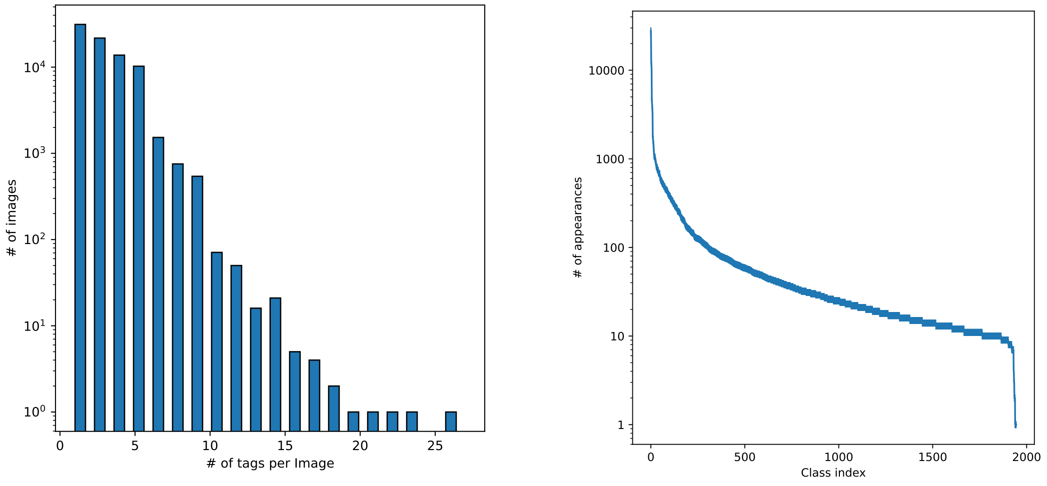
[4] PMC Open Access: https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/, Last accessed: 2024-06-20

[5] UMLS: https://www.nlm.nih.gov/research/umls/index.html, Last accessed: 2024-06-20

to kinds of medical examinations, such as X-Ray Computed Tomography or Plain x-ray. Most images are associated (in the ground truth) with at least one of these overarching concepts, alongside more specialized ones. The maximum and minimum number of concepts assigned to a single image are 27 and 1, occurring in 1 and 8,567 images respectively. The average number of assigned concepts per image is 3.1583. The aforementioned observations are outlined in the histogram in Figure 2(b).

Table 1
The ten most frequent concepts (CUIs) of the ImageCLEFmedical2024 dataset, along with their corresponding UMLS terms, and the number of images they are associated with.

Most Common Concepts			
Rank	**CUI**	**UMLS Term**	**Images**
1	C0040405	X-Ray Computed Tomography	27,852
2	C1306645	Plain x-ray	22,104
3	C0024485	Magnetic Resonance Imaging	12,733
4	C0041618	Ultrasonography	11,476
5	C0817096	Chest	10,323
6	C0002978	angiogram	4,808
7	C0000726	Abdomen	4,292
8	C0037303	Bone structure of cranium	4,130
9	C0030797	Pelvis	3,678
10	C0023216	Lower Extremity	3,254



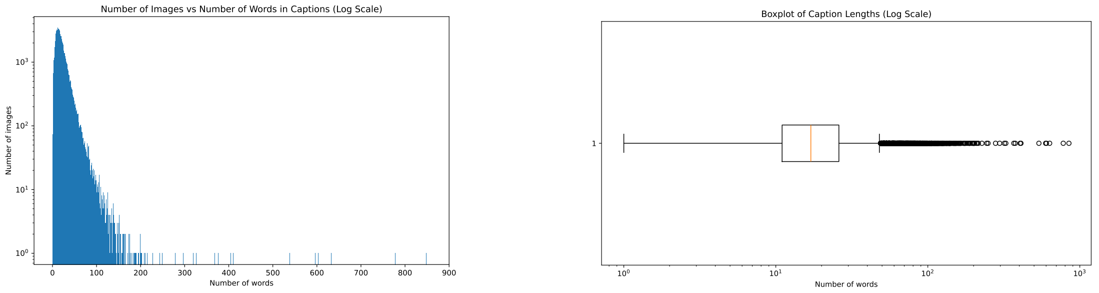
(a)

(b)

Figure 2: (a) Visualization of the dataset's long-tail distribution. The y-axis shows the number of occurrences of each concept, and the x-axis the concept's class index. (b) Histogram with 25 fixed-size bins (horizontal axis) depicting the number of gold concepts per image. Note that 13 concepts do not have corresponding UMLS terms.

2.2. Caption Prediction

In the Caption Prediction data, each image is accompanied by a gold diagnostic caption that describes the medical conditions present in the image. There are $80,080$ gold captions across the whole dataset, one for each provided image. Similar to last year's campaign, the vast majority of the captions, specifically 99.47% ($79,658$ out of $80,080$ captions), are unique. The maximum number of words in a single caption is 848 (occurred once), while the minimum is 1 (encountered 73 times). The average caption length is 21.01 words. These statistics apply to the dataset as a whole, but we have carefully checked that they remain consistent in all three subsets (training, validation, development) we formed. The five most common captions, as well as the ten most popular words, excluding the stopwords, can be found in Tables 2 and 3, respectively. In Figure 3, we provide a histogram alongside a box plot, utilizing a logarithmic scale in our visualizations. This helps make smaller counts more visible and reduces the dominance of larger values, giving a more balanced view of how the data is distributed.



(a)

(b)

Figure 3: (a) Histogram visualizing the distribution of caption lengths. The y-axis, displayed on a logarithmic scale, represents the number of images falling into each bin, while the x-axis shows the number of words in the captions. (b) Box-plot illustrating the same distribution, with the y-axis displayed on a logarithmic scale, highlighting outliers in the range of 100 to 200 words.

Table 2
The five most common gold captions found in the ImageCLEFmedical2024 dataset [10] alongside the number of images they are associated with.

Most common captions		
Rank	**Caption**	**Occurrences**
1	Initial panoramic radiograph.	40
2	Final panoramic radiograph.	37
3	Chest X-ray.	20
4	Chest radiograph.	17
5	Preoperative CT scan.	9

According to the organizers, each caption is pre-processed before evaluated in the following manner:

- The caption is converted to lower-case.
- Numbers are replaced by words, e.g., number 10 becomes “ten”.
- Punctuation is removed.

3. Methods

In this section, we present the methods we used in our submissions for both the Concept Detection and the Caption Prediction sub-tasks.

3.1. Concept Detection

Our submissions for this year’s Concept Detection sub-task are built upon two frameworks. Initially, we extensively explored a CNN+FFNN framework, building upon our prior research [18, 19, 20, 21], experimenting with various image encoders. Additionally, we used a neural image retrieval approach by integrating a k-nearest neighbors (k-NN) algorithm, which selects k neighbors and aggregates tags based on their frequency among the neighbors. Furthermore, we submitted several ensembles of the aforementioned systems. The ensembles employed strategies such as union-based and intersection-based aggregation.

Table 3
The ten most common words (of gold captions) and their frequencies in the ImageCLEFmedical2024 dataset [10], after removing stop-words.

Most common words (excluding stop-words)	
Word	**Count**
showing	22,519
right	18,258
left	18,136
ct	15,167
image	10,245
chest	10,082
scan	9,296
computed	9,273
tomography	8,969
shows	8,600

3.1.1. CNN + FFNN

This system employs a CNN encoder as its backbone, followed by an FFNN classification head. We extract image features from the last convolutional layer of the image encoder and we condense these feature maps into a feature vector (an image embedding) using global pooling. More specifically, we used the Generalized-Mean (GeM) pooling [25] mechanism.

The FFNN component classifies the image into one or more concepts. Its output layer has $|C|$ neurons, where C represents the set of unique concepts in the dataset. Each neuron uses a sigmoid activation function to transform its value into a probability value in $[0, 1]$. This results in one probability per label, and if this probability exceeds a specific threshold value t, the corresponding concept is assigned to the image. The threshold, which is the same for all concepts, was chosen through a grid search procedure that optimized the primary metric of the competition, on our validation set. The model was trained by minimizing binary cross-entropy, treating each concept as a separate binary target and summing up the individual losses. We used the Adam optimizer [26], along with a decreasing learning rate strategy and early stopping based on the validation set loss with a patience value of 3 epochs. We used an initial learning rate of $\eta = 10^{-3}$ and decreasing factor of 10.

In order to form the ensembles, we trained several instances of the model, using different random initializations, and combined them using the UNION and the INTERSECTION of their predicted concept sets. More details about our submitted ensemble systems can be found in Section 4.1.

3.1.2. CNN + k-NN

For our k-nearest neighbors (k-NN) approach, we leveraged the image embeddings obtained from the encoder of the trained CNN+FFNN system (Section. 3.1.1). We discarded the dense classification head and used the last GeM pooling layer to extract embeddings (feature vectors) for all the training images. These embeddings served as the basis for the retrieval process in the k-NN algorithm. Given a test image, the goal of the system is to retrieve similar images from the training set and select concepts from the retrieved neighbors. For each test image, we used the same encoder to obtain its embedding and we retrieved the k closest neighbors from the training set, based on cosine similarity computed on the image embeddings. We tuned the value of k in the range from 1 to 100 using our validation set, which led to $k = 33$.

For each test image, having obtained its k neighbors from the training set, we formed the set of concepts associated with the neighbors. We then ranked the concepts of the set based on the number of retrieved neighbors associated with each concept, ordering them from highest to lowest frequency. The concept with the highest frequency was always included in the predictions of the k-NN method for the test image. We then used two thresholds, t_1 and t_2, which we tuned using grid search on our validation set, to select which other concepts of the neighborhood to include in the predictions of k-NN. We calculated the difference in frequency (**Fr**) between the first and second most frequent concepts, divided by the frequency of the first concept, and if the result exceeded t_1, we included the second concept in the prediction:

$$\frac{\mathbf{Fr}(\text{concept}_1) - \mathbf{Fr}(\text{concept}_2)}{\mathbf{Fr}(\text{concept}_1)} \geq t_1 \,. \tag{1}$$

Similarly, we determined whether to include in the prediction the third most frequent concept or not, based on a comparison involving the first and third most frequent concepts. We calculated the difference between the frequencies of the first and third concepts, dividing it by the frequency of the first concept, and if this ratio exceeded t_2, we included the third concept:

$$\frac{\mathbf{Fr}(\text{concept}_1) - \mathbf{Fr}(\text{concept}_3)}{\mathbf{Fr}(\text{concept}_1)} \geq t_2 \,. \tag{2}$$

The same approach was applied to the difference between the first and fourth most frequent concepts, checking again against t_2, to decide if the fourth most frequent concept should be predicted:

$$\frac{\mathbf{Fr}(\text{concept}_1) - \mathbf{Fr}(\text{concept}_4)}{\mathbf{Fr}(\text{concept}_1)} \geq t_2 \,. \tag{3}$$

We opted to predict at most four concepts due to the fact that the average number of concepts in the training split was 3.08. The rationale was to select concepts that have frequencies close to that of the highest frequency concept, while excluding concepts that show a significant drop in frequency compared to the preceding ones. We experimented with t_1, t_2 values ranging from 0.3 to 0.9. Validation results indicated that the best parameters were $t_1 = 0.58$ and $t_2 = 0.65$.

3.1.3. CNN + weighted k-NN

We also developed a weighted version of the k-NN algorithm, using the voting scheme that was described in [27]. More specifically, given a test image x, we calculate for each concept $c_i \in C$ a score $f_i(x; w_1, \ldots, w_k)$ from the k neighbors retrieved for x:

$$f_i(x; w_1, \ldots, w_k) = \frac{\sum_{j=1}^{k} w_j \cdot y_{i,j,x}}{\sum_{j=1}^{k} w_j} \tag{4}$$

where $y_{i,j,x} = 1$ if concept c_i is present in the ground truth of the j-th neighbor of x, otherwise $y_{i,j,x} = 0$, and w_j is the weight assigned to the j-th nearest neighbor position; we explain below how the weights w_j are learned. Concept c_i is predicted for the test image x if and only if $f_i(x; w_1, \ldots, w_k) \geq t$, yielding the predicted label set $H(x; w_1, \ldots, w_k) = \{c_i | f_i(x; w_1, \ldots, w_k) \geq t\}$. The classification threshold $t \in [0, 1]$ and the number of neighbors $k \in [1, 100]$ were tuned on our validation set, resulting in $t = 0.35$ and $k = 50$. The weights $w_1, \ldots, w_k$ are the same for all the concepts c_i and test images x. They are learned using a genetic algorithm (GA) [28] by maximizing the following objective, where V denotes the validation set, $Y(x)$ is the ground truth set of concepts of image x, and F_1 is the official evaluation measure of the Concept Detection task:

$$\begin{gathered} \max_{w_1, \ldots, w_k} \sum_{x \in V} F_1(Y(x), H(x; w_1, \ldots, w_k)) \\ \text{s.t.} \quad 1 \geq w_1 \geq \ldots \geq w_k \geq 0 \,. \end{gathered} \tag{5}$$

In detail, we created a population of 500 randomly initialized weight vectors, initial *chromosomes* in GA terminology. Each chromosome had the form $\langle w_1, \ldots, w_k \rangle$, with all weights $w_i \in [0, 1]$; we ensured that the monotonicity constraint $1 \geq w_1 \geq \ldots \geq w_k \geq 0$ was satisfied by all chromosomes. We then used a crossover mechanism where two chromosomes were combined to form two new ones. At each application of the crossover mechanism, we selected pairs of chromosomes (parents) out of the population and combined their values to form two new ones from each pair of parents. The crossover operator splits the two parent chromosomes at a random point and creates two children chromosomes by combining the values before the crossover point (or after) for one parent, and after (or before) the crossover point for the other parent. Furthermore, we used a mutation mechanism that perturbed the values of the resulting children chromosomes by adding a random value in $[-0.1, 0.1]$ to every gene, with a 0.1 mutation probability per gene (w_i). Both the crossover and the mutation operators paid respect to the range and monotonicity constraints; we added a clipping and a sorting operation that were applied if any of the constraints were violated in the resulting chromosomes. We used $F_1(Y(x), H(x))$ as the fitness function. The fitness function is used to select the chromosomes to be used as parents in the crossover mechanism at each iteration of the algorithm (fitter chromosomes are selected with higher probability as parents). At each generation (new population), we performed the crossover mechanism as many times as necessary to have a new generation with as many members as the previous one (and as many as the initial population, i.e., 500 chromosomes). We run the optimization process for 30 iterations (generations).

3.2. Caption Prediction

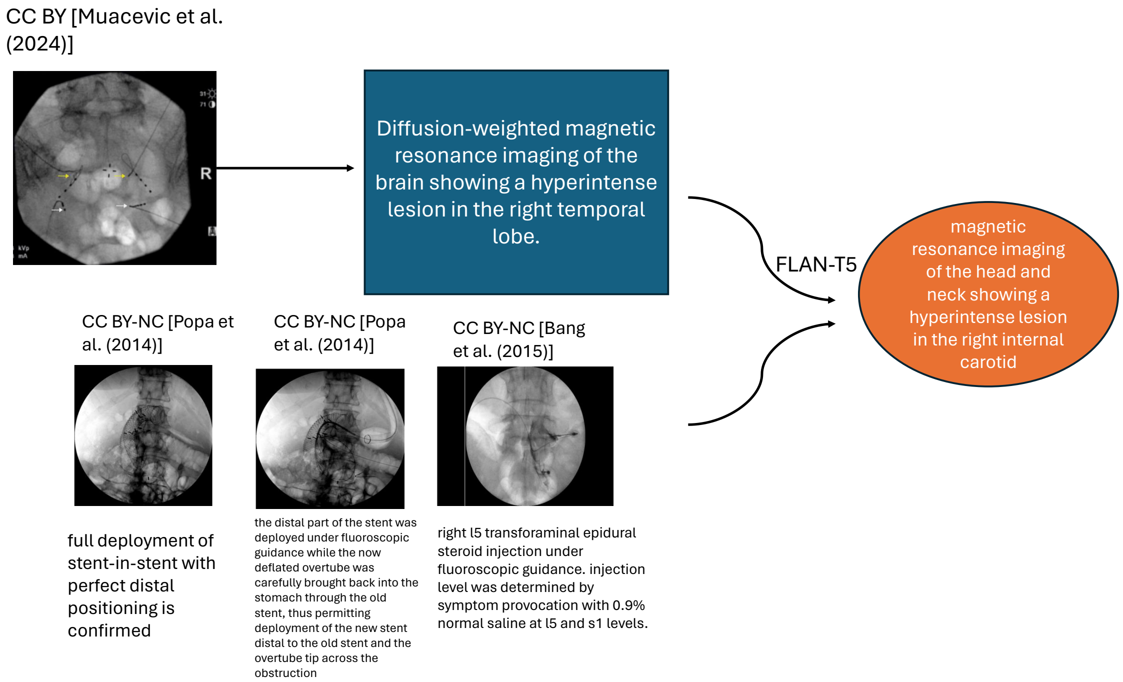
Our submissions for the Caption Prediction sub-task focused on four primary systems. The first system employs an InstructBLIP model [9] (Section 3.2.1), while the remaining submissions build on this model using techniques such as rephrasing [12, 13] (Section 3.2.3) and synthesizing [12] (Section 3.2.2). Finally, we implemented an innovative guided-decoding mechanism, DMMCS [7] (Section 3.2.4), which leverages information from the tags predicted by our CNN+k-NN classifier (Section 3.1.2) in the Concept Detection task to improve the generated caption.

3.2.1. InstructBLIP

The InstructBLIP model [9] is a sophisticated neural network designed to generate descriptive text for scientific images. It employs a technique known as instruction-tuning [29], which refines its behavior and responses based on user-provided instructions. This approach aims to enhance the model's controllability and its adaptability across different domains. The InstructBLIP model comprises three key components: an image encoder, a Q-Former [30], and an LLM. The frozen image encoder converts the image into a low-dimensional vector and generates image embeddings. The Q-Former then extracts instruction-aware visual features from these embeddings and can process the text prompt (instruction) to enhance this extraction. Through extensive training, the LLM learns to correlate textual prompts with relevant image features, thereby generating coherent and contextually appropriate descriptions. The InstructBLIP model played a crucial role in creating the initial captions, which were subsequently utilized in our other caption prediction methods.

3.2.2. Synthesizer

Our goal was to the captions obtained from the InstructBLIP model (Section 3.2.1) by leveraging information from similar training images, based on the intuition that similar images may have similar captions [31, 32]. To achieve this, we computed embeddings for all images in the dataset using the CCN + FFNN model, which was developed for Concept Detection (Section 3.1.1). A cosine similarity threshold was then applied to decide if an image qualified as a neighbor of the test image. Images exceeding this threshold were considered neighbors [33]. For each image in the test set [24], we identified the k most similar images from the entire dataset [10], which includes training, validation, and development images, to retrieve their corresponding captions. We experimented with $k \in \{1, 3, 5\}$; the best results in our validation set were obtained for $k = 5$, so we used that value. The Synthesizer, a FLAN-T5 model [11], was trained to refine the captions generated by InstructBLIP by considering also the captions of the neighbors, which are concatenated to the caption of InstructBLIP, similarly in spirit to [13]. We also experimented with different beam sizes m, for the beam search decoding of the Synthesizer during inference; setting $m = 5$ yielded the best validation scores, so we used that value. Figure 4 illustrates the process (for $m = 3$), starting with the caption generated by InstructBLIP, merging it with the captions of the neighbors, and using FLAN-T5 to obtain a refined caption.


CC BY [Muacevic et al. (2024)]
Diffusion-weighted magnetic resonance imaging of the brain showing a hyperintense lesion in the right temporal lobe.
FLAN-T5
magnetic resonance imaging of the head and neck showing a hyperintense lesion in the right internal carotid
CC BY-NC [Popa et al. (2014)]
CC BY-NC [Popa et al. (2014)]
CC BY-NC [Bang et al. (2015)]

full deployment of stent-in-stent with perfect distal positioning is confirmed

the distal part of the stent was deployed under fluoroscopic guidance while the now deflated overtube was carefully brought back into the stomach through the old stent, thus permitting deployment of the new stent distal to the old stent and the overtube tip across the obstruction

right l5 transforaminal epidural steroid injection under fluoroscopic guidance. injection level was determined by symptom provocation with 0.9% normal saline at l5 and s1 levels.

Figure 4: Illustration of a radiology image (CC BY [Muacevic et al., 2024]), accompanied by similar neighbor images (CC BY-NC [Popa et al., 2014], CC BY-NC [Popa et al., 2014], CC BY-NC [Bang et al., 2015]) and their corresponding captions from the 2024 ImageCLEFmedical caption task [10, 24]. The initial caption, generated by InstructBLIP, is concatenated with the captions of the neighbors and is then fed to a FLAN-T5 Synthesizer, which generates a refined caption.

3.2.3. Rephraser

Furthermore, we experimented with a domain-specific variation of T5, namely ClinicalT5. This is an encoder-decoder transformer, which is pre-trained in a series of both supervised and unsupervised tasks [34], including denoising tasks, and then further pre-trained on the union of MIMIC-III and IV clinical notes, to which we were granted access through PhysioNet[6]. Following our previous work [35], we created a corrective text-to-text training set, consisting of noisy and ground truth caption pairs, with the former having been generated by our captioning systems. Therefore, we treated our original system as a noise-insertion function, then we further fine-tuned ClinicalT5, in order to rephrase the noisy captions to approximate the gold ones, hoping it would acquire knowledge of the medical domain, use medical terms more accurately and therefore generate more medically fluent text captions. Specifically, we fine-tuned ClinicalT5 to rephrase the captions of InstructBlip (Section 3.2.1), InstructBlip with FLAN-T5 Synthesizer (Section 3.2.2) on top and InstructBlip with DMMCS (Section 3.2.4) using $\alpha = 0.10$. Performance in terms of the primary metric in our development set improved, but test-time performance (in the official evaluation) deteriorated.

3.2.4. DMMCS

In this section, we present "Distance from Median Maximum Concept Similarity" (DMMCS) [7], a novel data-driven guided decoding mechanism designed to incorporate domain-specific information (in the form of keywords) into the text generation process. The intuition behind this guided decoding algorithm lies in the observation that an accurate diagnostic caption should mention the key medical conditions

[6] https://www.physionet.org/content/clinical-t5/1.0.0/, Last accessed: 2024-06-20

depicted in the given image. For example, if a radiology image is assigned the tag “Pneumonia”, but the generated caption does not refer to this medical condition either explicitly or implicitly, then the caption is potentially inaccurate. Such conditions are typically represented by the medical tags provided in the ImageCLEF2024 dataset, which the Concept Detection task is also trying to predict. Therefore we use tags predicted by one of our Concept Detection systems (Section 3.1), in order to guide our Caption Prediction models towards captions that express the tags appropriately. We achieve this by imposing a new penalty at each decoding step, aiming to prioritize the generation of words semantically similar to the (predicted) medical tags. This penalty also considers the frequency with which each tag is explicitly or implicitly expressed in the dataset’s gold captions.

In more detail, recent work examining DC datasets [22, 7] has shown that some tags are more prominently expressed than others in the corresponding diagnostic captions. More specifically, Kaliosis et al. [7] performed an exploratory analysis on the ImageCLEF2023 and MIMIC-CXR datasets, where they investigated the relationship between each tag and the gold captions of the images that are associated with the tag in the ground truth. This was achieved by calculating the cosine similarity between the word embeddings of each caption’s tokens and each tag. The results showed that some tags are always explicitly expressed in the gold captions of the images the tags are associated with, while other tags are mentioned more implicitly or even not at all. More concretely, the similarity between a tag t and a caption c is defined as the maximum cosine similarity (MCS) between the centroid $h(t)$ of the word embeddings of t and the embedding $h(c_i)$ of each token in c, i.e.,

$$\mathit{MCS}(t, c) = \max_{1 \leq i \leq |c|} \mathit{sim}(h(t), h(c_i)). \tag{7}$$

A high *MCS* score between a tag t and a caption c implies that t is strongly expressed in the caption, while a low *MCS* score indicates that it was rather implicitly (or not at all) mentioned. The *MCS* similarity is also calculated for all the gold captions of the images a tag t is associated with in the training data. Specifically, for each tag t and the set C containing its associated captions, the distribution $R(t, C)$ is calculated as:

$$R(t, C) = \{\mathit{MCS}(t, c) | c \in C\}. \tag{8}$$

The median value of the distribution $R(t, C)$, hereafter called Median Maximum Cosine Similarity (MMCS), indicates how strongly t is expressed on average in the training captions it is associated with.

$$\mathit{MMCS}(t, C) = \mathit{median}(R(t, C)). \tag{9}$$

During inference, when generating the caption for an image with a single tag t, the $\mathit{MCS}(t, c)$ of the tag t and each candidate (possibly still incomplete) caption c of the beam search is calculated (Eq. 7). The penalty, imposed at each decoding step, is then defined as the squared difference between $\mathit{MCS}(t, c)$ and $\mathit{MMCS}(t, C)$. The former shows how strongly the tag is mentioned in the candidate caption, while the latter indicates how strongly the tag is expressed on average in the gold training captions associated with the tag. When more than one tags are assigned to an image, a distinct penalty is calculated for each tag, and the overall penalty is the average of the individual penalties. Thus, given a candidate caption c, the set of its associated training captions C, and a set of tags T, the penalty is calculated as:

$$\mathit{DMMCS}_{\mathit{pen}}(T, C, c) = \frac{1}{|T|} \sum_{t \in T} (\mathit{MCS}(t, c) - \mathit{MMCS}(t, C))^2. \tag{10}$$

Intuitively, the objective of the DMMCS algorithm is to guide the model to generate captions that express each associated tag as explicitly (or implicitly) as it is expressed in the training corpus. Overall, at each decoding step, each candidate caption c generated through the beam search process is scored by the following formula:

$$DMMCS(c) = \alpha \cdot DMMCS_{pen}(T, C, c) + (1 - \alpha) \cdot (1 - D_{score}), \tag{11}$$

where T is a given set of predicted tags, α is a tunable weighting factor, while D_{score} is the score that the decoder assigns to the candidate caption c.

4. Experiments, Submissions and Results

In this section, we provide details about our experiments regarding this year's evaluation campaign [1]. Moreover, we share details about our submissions and the scores achieved in our held-out development set, as well as the official test set of the competition [24] for both sub-tasks.

4.1. Concept Detection

In the Concept Detection sub-task we submitted our ten best performing models, after evaluating them on our held-out development set. We submitted two instances with different image encoders of our CNN + FFNN model (Section 3.1.1), one instance of our CNN + k-NN model (Section 3.1.2), and a single instance of our CNN + weighted k-NN model (Section 3.1.3). In our subsequent submissions, we employed ensemble systems. These involved exploring the integration of predictions from multiple instances by computing either the union or the intersection of their predicted concept sets. Our submitted ensemble systems consisted of various combinations of CNN-based architectures paired with different classifiers, specifically CNN + FFNN, CNN + k-NN (KNN), and CNN + weighted k-NN (wKNN). To enhance the diversity and robustness of our ensembles, we incorporated different architectures for the CNN component.

The primary evaluation metric for this year's Concept Detection sub-task was the F_1-score, calculated between the predicted and ground truth captions. It is calculated as the sum of the F_1-scores for each test image, divided by the total number of test images. Each partial score is derived from the binary multi-hot candidate vector compared to the corresponding ground truth vector. Specifically, let F_1 represent the overall F_1-score, and $\hat{f}_1$ denote the individual F_1-score for each test image. Additionally, let p_t and g_t be the predicted and ground truth concepts for an image t, respectively. Finally, let T be the test set [24].

$$F_1 = \frac{1}{|T|} \sum_{t \in T} \hat{f}_1(p_t, g_t) \tag{6}$$

Moreover, a secondary evaluation metric (again an F_1 score) was calculated, which only considered manually selected concepts, such as anatomy, topography, and modality.

For our first system (CNN+FFNN), we experimented with a variety of CNN encoders as their backbone components. Specifically, we trained the networks using state-of-the-art CNN architectures, including EfficientNet and DenseNet. Furthermore, we extended our experiments by incorporating these CNN encoders into our k-NN models.

During testing on our held-out development set, we observed a slightly higher F1 score in models utilizing the EfficientNet image encoder.

Our ensembling approaches did not show significant improvement over our individual models, with minimal differences observed in both the development and test set [24].

Table 4
Summary of the scores of our individual experiments (ensembles included) in the ImageCLEFmedical2024 Concept Detection sub-task. This table presents the highest scores of our systems on our held-out development set for each method.

Individual Concept Detection Experiments		
Run ID	**Method**	**Development**
619	CNN+FFNN (DenseNet)	0.6007
624	CNN+KNN	0.6007
640	INTERSECTION(UNION(3xCNN+FFNN),624)	0.6022
642	UNION(2xCNN+FFNN)	0.6047
644	CNN+FFNN (EfficientNet)	0.6042
648	UNION(644,624)	0.6045
651	CNN+wKNN	0.5961
654	UNION(651,644)	0.6008
655	UNION(651,624)	0.5970
656	UNION(651,619)	0.5981

Table 5
Summary of our submissions to the ImageCLEFmedical2024 Concept Detection sub-task. The table presents the scores of our systems on both our held-out development set and the official test set [24]. It also includes the rankings of these systems among all submissions from the 9 participating teams.

Individual Concept Detection Experiments					
Run ID	**Method**	**Primary F1**		**Secondary F1**	**Rank**
		Dev	**Test**		
619	CNN+FFNN (DenseNet)	0.6007	0.6240	0.9339	12
624	CNN+KNN	0.6007	0.6274	0.9375	8
640	INTERSECTION(UNION(3xCNN+FFNN),624)	0.6022	0.6272	**0.9415**	10
642	UNION(2xCNN+FFNN)	**0.6047**	0.6304	0.9332	7
644	CNN+FFNN (EfficientNet)	0.6042	**0.6319**	0.9392	**4**
648	UNION(644,624)	0.6045	0.6308	0.9321	6
651	CNN+wKNN	0.5961	0.6135	0.9238	17
654	UNION(651,644)	0.6008	0.6207	0.9243	13
655	UNION(651,624)	0.5970	0.6155	0.9233	16
656	UNION(651,619)	0.5981	0.6162	0.9217	15

4.2. Caption Prediction

For the Caption Prediction sub-task, we submitted nine systems based on their performance on our development set. Our submissions included InstructBLIP (Section 3.2.1), a synthesizer variant combining InstructBLIP with FLAN-T5 (Section 3.2.2), and a rephrasing variant that employs ClinicalT5 (Section 3.2.3). Additionally, we explored combinations of all three approaches, aiming to refine the captions generated by InstructBLIP and FLAN-T5 (Section 3.2.2) using our ClinicalT5 rephraser on top. Furthermore, we submitted three variations of InstructBLIP and DMMCS, each with a different α value (Section 3.2.4). Finally, we provided two instances where we employed ClinicalT5 to rephrase the results generated by the combination of InstructBLIP and DMMCS, in this case using a $\alpha = 0.10$.

In this year's campaign, BERTScore [36] was the primary evaluation metric in the Caption Prediction task, while ROUGE [37] was the secondary metric. Other metrics utilized include, for example, BLEU-1 [38], BLEURT [39], and METEOR [40]. Table 6 shows captions produced by each of our submissions for the test image CC BY [Muacevic et al. (2024)], extracted from the test dataset [24].

Finally, Table 7 provides an overview of our models, detailing their performance across fundamental campaign metrics in both our development set and the provided test set [24], along with our attained

Table 6
Captions generated by our submitted models for the test image [24] CC BY [Muacevic et al. (2024)]

Generated captions	
InstructBLIP	Diffusion-weighted magnetic resonance imaging of the brain showing a hyperintense lesion in the right temporal lobe.
InstructBLIP + Synthesizer	magnetic resonance imaging of the head and neck showing a hyperintense lesion in the right internal carotid.
InstructBLIP + Rephraser	Axial computed tomography scan of the head showing a mass in the left maxillary sinus (arrow).
InstructBLIP + Synthesizer + Rephraser	Computed tomography scan of the head and neck showing a mass in the right parotid gland.
InstructBLIP + DMMCS (alpha 0.1)	Chest X-ray showing bilateral pulmonary edema.
InstructBLIP + DMMCS (alpha 0.1) + Rephraser	Computed tomography scan of the head and neck showing a mass in the right parotid gland.
InstructBLIP + DMMCS (alpha 0.1) + Rephraser (random restart)	Anteroposterior radiograph of the pelvis showing a large right-sided pleural effusion.

Table 7
Summary of the scores of our submissions to the ImageCLEFmedical2024 Caption Prediction sub-task.

AUEB NLP Group - Submission Table						
Run ID	**Approach**	**BERTScore**		**ROUGE-1**		**Rank**
		Dev	**Test**	**Dev**	**Test**	
564	InstructBLIP	0.6164	0.6152	0.1931	0.2052	22
577	InstructBLIP + Rephraser	0.7651	0.6106	0.1840	0.1837	26
605	InstructBLIP + Synthesizer	0.6194	0.6113	0.1898	0.1889	24
630	InstructBLIP + DMMCS ($\alpha = 0.1$)	0.6564	**0.6211**	**0.2027**	**0.2048**	**10**
635	InstructBLIP + DMMCS ($\alpha = 0.05$)	0.6534	0.6210	0.2025	0.2047	11
639	InstructBLIP + Synthesizer + Rephraser	0.7603	0.6111	0.1840	0.1827	25
647	InstructBLIP + DMMCS ($\alpha = 0.1$) + ClinicalT5	0.7981	0.6209	0.1928	0.1807	13
650	InstructBLIP + DMMCS ($\alpha = 0.1$) + ClinicalT5 (random restart)	**0.8012**	0.6159	0.1932	0.1936	20
646	InstructBLIP + DMMCS ($\alpha = 0.15$)	0.6530	0.6209	0.2024	0.2044	12

rankings. Additionally, Table 8 presents a summary of all the metrics utilized in this year's campaign, offering a comprehensive view of the experiments.

5. Conclusion

Our participation in the ImageCLEFmedical Caption task provided an opportunity to explore innovative NLP approaches for medical image captioning. Utilizing state-of-the-art models, we demonstrated competitive performance in both the Concept Detection and Caption Prediction sub-tasks.

In the Concept Detection sub-task, we achieved a 2nd place ranking among the participating groups. Our top-performing system was a CNN+FFNN pipeline (Section 3.1.1), while our remaining submissions included a CNN+KNN (Section 3.1.2) and a CNN+wKNN (Section 3.1.3), which also produced competitive

Table 8
Summary of our submissions regarding the Caption Prediction sub-task. The table contains each system's performance on all officially reported measures.

AUEB NLP Group Submissions - Evaluation on All Metrics											
Run ID	**BERTScore**	**ROUGE**	**BLEU-1**	**BLEURT**	**METEOR**	**CIDEr**	**CLIPscore**	**RefCLIPscore**	**ClinicalBLEURT**	**MedBERTScore**	**Rank**
630	0.6211	0.2049	0.1110	0.2899	0.0680	0.1769	0.8041	0.7987	0.4866	0.6261	10
635	0.6210	0.2047	0.1108	0.2895	0.0680	0.1762	0.8040	0.7986	0.4870	0.6260	11
646	0.6210	0.2044	0.1107	0.2900	0.0678	0.1758	0.8041	0.7988	0.4872	0.6261	12
647	0.6210	0.1807	0.0860	0.2846	0.0580	0.1459	0.7936	0.7912	0.5021	0.6291	13
650	0.6160	0.1936	0.1050	0.2859	0.0638	0.1597	0.7980	0.7948	0.4874	0.6212	20
564	0.6153	0.2052	0.1274	0.2920	0.0698	0.1728	0.8045	0.7968	0.4844	0.6197	22
605	0.6114	0.1889	0.1147	0.2796	0.0616	0.1305	0.8037	0.7962	0.4834	0.6174	24
639	0.6111	0.1827	0.0744	0.2717	0.0515	0.1293	0.7858	0.7845	0.5212	0.6141	25
577	0.6107	0.1838	0.0751	0.2706	0.0513	0.1292	0.7832	0.7826	0.5158	0.6134	26

results. We also employed ensembles that combined these approaches using union and intersection (of predicted tags) approaches.

In the Caption Prediction sub-task, we were ranked 4th among all participating groups, by both extending our previous work [22, 21, 17] and exploiting the state-of-the-art in NLP, such as instruction-tuned Large Language Models. Our approach involved the initial generation of captions using the InstructBLIP model [9], followed by their enrichment through the synthesis of information from the captions of similar images [12, 13] and the utilization of a model further pre-trained in the medical domain [14] to improve the originally generated captions.

In future work, we plan to further investigate and improve biomedical LLMs and further explore their reasoning capabilities through instruction tuning and, more generally, alignment with medical professionals needs [41]. We also plan to utilize a model capable of processing both image and text inputs in our Synthesizer approach (Section 3.2.2) to combine information not only from the captions of the neighbors, but also from the images themselves. Furthermore, we plan to exploit Retrieval-Augmented Generation [42] algorithms to combine prior knowledge with new medical cases. Finally, the generated captions need to be evaluated in collaboration with medical experts, to assess their medical accuracy and usefulness.

Acknowledgments

This work has been partially supported by project MIS 5154714 of the National Recovery and Resilience Plan Greece 2.0 funded by the European Union under the NextGenerationEU Program.

References

[1] B. Ionescu, H. Müller, A. Drăgulinescu, J. Rückert, A. Ben Abacha, A. Garcıa Seco 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. Andrei, Y. Prokopchuk, D. Karpenka, A. Radzhabov, V. Kovalev, C. Macaire, D. Schwab, B. Lecouteux, E. Esperança-Rodier, W. 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.

[2] 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.

[3] J. Pavlopoulos, V. Kougia, I. Androutsopoulos, D. Papamichail, Diagnostic Captioning: A Survey, Knowledge and Information Systems 64 (2022) 1–32. doi:10.48550/arXiv.2101.07299.

[4] H.-C. Shin, K. Roberts, L. Lu, D. Demner-Fushman, J. Yao, R. M. Summers, Learning to Read Chest X-Rays: Recurrent Neural Cascade Model for Automated Image Annotation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2497–2506. doi:10.48550/arXiv.1603.08486.

[5] G. Moschovis, Medical image captioning based on Deep Architectures, Master's thesis, KTH Royal Institute of Technology, Stockholm, Sweden, 2022. URL: http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-323528, Last accessed: 2024-06-20.

[6] J. Pavlopoulos, V. Kougia, I. Androutsopoulos, A Survey on Biomedical Image Captioning, in: R. Bernardi, R. Fernandez, S. Gella, K. Kafle, C. Kanan, S. Lee, M. Nabi (Eds.), Proceedings of the Second Workshop on Shortcomings in Vision and Language, Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 26–36. URL: https://aclanthology.org/W19-1803. doi:10.18653/v1/W19-1803, Last accessed: 2024-06-20.

[7] P. Kaliosis, J. Pavlopoulos, F. Charalampakos, G. Moschovis, I. Androutsopoulos, A data-driven guided decoding mechanism for diagnostic captioning, in: Findings of the Association for Computational Linguistics: ACL 2024, 2024.

[8] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du, C. Yang, Y. Chen, Z. Chen, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu, P. Liu, J.-Y. Nie, J.-R. Wen, A Survey of Large Language Models, 2023. doi:10.48550/arXiv.2303.18223. arXiv:2303.18223.

[9] W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. N. Fung, S. Hoi, InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning, Advances in Neural Information Processing Systems 36 (2024). doi:10.48550/arXiv.2305.06500.

[10] O. Pelka, S. Koitka, J. Rückert, F. Nensa, C. Friedrich, "Radiology Objects in COntext (ROCO): A Multimodal Image Dataset: 7th Joint International Workshop, CVII-STENT 2018 and Third International Workshop, LABELS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings", 2018, pp. 180–189. doi:10.1007/978-3-030-01364-6_20.

[11] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma, et al., Scaling Instruction-Finetuned Language Models, Journal of Machine Learning Research 25 (2024) 1–53. doi:10.48550/arXiv.2210.11416.

[12] Y. Li, X. Liang, Z. Hu, E. Xing, Knowledge-Driven Encode, Retrieve, Paraphrase for Medical Image Report Generation, in: AAAI Conference on Artificial Intelligence, volume abs/1903.10122, 2019. doi:10.1609/aaai.v33i01.33016666.

[13] G. Vernikos, A. Brazinskas, J. Adamek, J. Mallinson, A. Severyn, E. Malmi, Small Language Models Improve Giants by Rewriting Their Outputs, in: Y. Graham, M. Purver (Eds.), Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, St. Julians, Malta, 2024, pp. 2703–2718. URL: https://aclanthology.org/2024.eacl-long.165. doi:10.48550/arXiv.2305.13514, Last accessed: 2024-06-20.

[14] Q. Lu, D. Dou, T. Nguyen, ClinicalT5: A Generative Language Model for Clinical Text, in: Findings of the Association for Computational Linguistics: EMNLP 2022, 2022, pp. 5436–5443. doi:10.18653/v1/2022.findings-emnlp.398.

[15] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, in: International Conference on Learning Representations, 2021. URL: https://openreview.net/forum?id=YicbFdNTTy. doi:10.48550/arXiv.2010.11929, Last accessed: 2024-06-20.

[16] I. Athanasiadis, G. Moschovis, A. Tuoma, Weakly-Supervised Semantic Segmentation via Transformer Explainability, in: ML Reproducibility Challenge 2021 (Fall Edition), 2022. doi:10.5281/zenodo.6574631.

[17] G. Moschovis, E. Fransén, NeuralDynamicsLab at ImageCLEF Medical 2022, in: CLEF2022 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Bologna, Italy, 2022.

[18] V. Kougia, J. Pavlopoulos, I. Androutsopoulos, AUEB NLP Group at ImageCLEFmed Caption 2019, in: Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, Lugano,

Switzerland, September 9-12, volume 2380 of *CEUR Workshop Proceedings*, 2019.

[19] B. Karatzas, J. Pavlopoulos, V. Kougia, I. Androutsopoulos, AUEB NLP Group at ImageCLEFmed Caption 2020, in: Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, volume 2696 of *CEUR Workshop Proceedings*, 2020.

[20] F. Charalampakos, V. Karatzas, V. Kougia, J. Pavlopoulos, I. Androutsopoulos, AUEB NLP Group at ImageCLEFmed Caption Tasks 2021, in: Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucharest, Romania, September 21-24, volume 2936 of *CEUR Workshop Proceedings*, 2021, pp. 1184–1200.

[21] F. Charalampakos, G. Zachariadis, J. Pavlopoulos, V. Karatzas, C. Trakas, I. Androutsopoulos, AUEB NLP Group at ImageCLEFmedical Caption 2022, in: CLEF2022 Working Notes, CEUR Workshop Proceedings, CEUR-WS.or, Bologna, Italy, 2022, pp. 1355–1373.

[22] P. Kaliosis, G. Moschovis, F. Charalampakos, J. Pavlopoulos, I. Androutsopoulos, AUEB NLP Group at ImageCLEFmedical Caption 2023, in: CLEF2023 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece, 2023.

[23] O. Bodenreider, The Unified Medical Language System (UMLS): integrating biomedical terminology, Nucleic acids research 32 (2004) D267–D270. doi:10.1093/nar/gkh061.

[24] 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.

[25] F. Radenović, G. Tolias, O. Chum, Fine-Tuning CNN Image Retrieval with No Human Annotation, IEEE Transactions on Pattern Analysis and Machine Intelligence 41 (2019) 1655–1668. doi:10.1109/TPAMI.2018.2846566.

[26] D. P. Kingma, J. L. Ba, Adam: A Method for Stochastic Optimization, in: 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015.

[27] T.-H. Chiang, H.-Y. Lo, S.-D. Lin, A Ranking-based KNN Approach for Multi-Label Classification, in: Proceedings of the Asian Conference on Machine Learning, volume 25, Singapore Management University, Singapore, 2012, pp. 81–96.

[28] A. Eiben, J. E. Smith, Introduction to Evolutionary Computing, 2nd ed., Springer Publishing Company, Incorporated, 2015. doi:10.1007/978-3-662-44874-8.

[29] J. Wei, M. Bosma, V. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, Q. V. Le, Finetuned Language Models Are Zero-Shot Learners, International Conference on Learning Representations abs/2109.01652 (2021). doi:10.48550/arXiv.2109.01652.

[30] J. Li, D. Li, S. Savarese, S. C. H. Hoi, BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models, in: International Conference on Machine Learning, 2023. URL: https://api.semanticscholar.org/CorpusID:256390509. doi:10.48550/arXiv.2301.12597, Last accessed: 2024-06-20.

[31] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, H. Wang, Retrieval-Augmented Generation for Large Language Models: A Survey, 2024. doi:10.48550/arXiv.2312.10997. arXiv:2312.10997.

[32] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Kuttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, Neural Information Processing Systems abs/2005.11401 (2020).

[33] Y. Huang, J. Huang, A Survey on Retrieval-Augmented Text Generation for Large Language Models, 2024. doi:10.48550/arXiv.2404.10981. arXiv:2404.10981.

[34] C. Raffel, N. M. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer, Journal of machine learning research 21 (2019) 140:1 – 140:67. doi:10.48550/arXiv.1910.10683.

[35] P. Kaliosis, Exploring Uni-modal, Multi-modal and Few-Shot Deep Learning Methods for Diagnostic Captioning, 2023. M.Sc. thesis, Department of Informatics, Athens University of Economics and Business.

[36] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, BERTScore: Evaluating text generation with BERT, International Conference on Learning Representations abs/1904.09675 (2019).

[37] C.-Y. Lin, ROUGE: A Package for Automatic Evaluation of Summaries, in: Text Summarization Branches Out, Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 74–81. URL: https://aclanthology.org/W04-1013, Last accessed: 2024-06-20.

[38] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, BLEU: a Method for Automatic Evaluation of Machine Translation, in: P. Isabelle, E. Charniak, D. Lin (Eds.), Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 2002, pp. 311–318. URL: https://aclanthology.org/P02-1040. doi:10.3115/1073083.1073135, Last accessed: 2024-06-20.

[39] T. Sellam, D. Das, A. Parikh, BLEURT: Learning Robust Metrics for Text Generation, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 7881–7892. URL: https://aclanthology.org/2020.acl-main.704. doi:10.18653/v1/2020.acl-main.704, Last accessed: 2024-06-20.

[40] S. Banerjee, A. Lavie, METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments, in: J. Goldstein, A. Lavie, C.-Y. Lin, C. Voss (Eds.), Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, Association for Computational Linguistics, Ann Arbor, Michigan, 2005, pp. 65–72. URL: https://aclanthology.org/W05-0909. doi:10.3115/1626355.1626389, Last accessed: 2024-06-20.

[41] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. E. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, R. J. Lowe, Training language models to follow instructions with human feedback, Neural Information Processing Systems abs/2203.02155 (2022). doi:10.48550/arXiv.2203.02155.

[42] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, in: Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 9459–9474. doi:10.48550/arXiv.2005.11401.