# **AUEB NLP Group at ImageCLEFmedical Caption 2022**

Foivos Charalampakos<sup>1</sup>, Giorgos Zachariadis<sup>1</sup>, John Pavlopoulos<sup>1,2</sup>, Vasilis Karatzas<sup>1</sup>, Christoforos Trakas<sup>1</sup> and Ion Androutsopoulos<sup>1</sup>

<sup>1</sup>Department of Informatics, Athens University of Economics and Business, 76, Patission Street, 104 34 Athens, Greece <sup>2</sup>Department of Computer and Systems Sciences, DSV, Stockholm University, Postbox 7003, SE-164 07 Kista, Sweden

#### Abstract

We present the methods AUEB's NLP Group used to participate in the annual ImageCLEFmedical Caption Task. The task comprises of the Concept Detection and the Caption Prediction sub-tasks. Concept Detection aims to automatically tag medical images with relevant medical concepts, while Caption Prediction generates draft diagnostic captions of medical images to help medical experts prepare diagnostic reports. For Concept Detection, we employ CNN image encoders, combined with a feed-forward neural network classifier or a retrieval module, extending our previous work. For Caption Prediction, we also extend the retrieval approach of our previous work with a caption selection method; we also experiment with a state-of-the-art memory-enhanced caption generation method, a simpler CNN-RNN caption generation model, and a captions ensemble method, which combines predictions from our different models. We ranked 1st in Concept Detection and 2nd in Caption Prediction.

#### Keywords

Medical Images, Concept Detection, Image Captioning, Image Retrieval, Multi-label Classification, Convolutional Neural Network, Natural Language Processing, Machine Learning, Deep Learning

# 1. Introduction

ImageCLEF [1] is an evaluation campaign held annually since 2003 as part of CLEF, revolving around image analysis and retrieval tasks.<sup>1</sup> ImageCLEFmedical, which was one of the four main ImageCLEF 2022 tasks, consists of a series of challenges associated with the study and processing of medical images. The ImageCLEFmedical Caption Task [2] ran for the 6th year in 2022. As in the previous year, it included a Concept Detection sub-task, where the goal was to perform multi-label classification of medical images by automatically assigning medical terms (called concepts) to each image. The concepts stem from the Unified Medical Language System (UMLS) [3].<sup>2</sup> Selecting the appropriate medical terms can be a first step towards automatically generating image captions and/or assisting medical experts by reducing the time needed for a diagnosis [4]. Following the previous edition of the task, ImageCLEFmedical also included a

- (I. Androutsopoulos)
- http://www.aueb.gr/users/ion/ (I. Androutsopoulos)

D 0000-0001-9188-7425 (J. Pavlopoulos)

<sup>1</sup>http://www.clef-initiative.eu/

CLEF 2022: Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy

<sup>© 0 2022</sup> Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

<sup>&</sup>lt;sup>2</sup>https://www.nlm.nih.gov/research/umls/

Caption Prediction sub-task [2]. This sub-task aims to automatically generate draft diagnostic captions given medical images, which could potentially help medical experts prepare diagnostic reports and, more generally, help them analyze more efficiently large volumes of medical images (e.g., X-rays, MRI scans) they confront in their daily workflow [4].

In this paper, we describe the systems that were submitted to the ImageCLEFmedical Caption 2022 sub-tasks by AUEB's NLP Group, which extend prior work on medical image understanding [5, 6, 7, 8, 4]. For the Concept Detection sub-task, our submissions were based on two methods. The first method extends our classification system [5, 6, 7, 8], which uses a Convolutional Neural Network (CNN) image encoder [9] and a feed-forward neural network (FFNN) classifier on top. The second method was based on neural retrieval approaches of our previous work, again employing a CNN image encoder and a weighting scheme suitable for multi-label classification. For the Caption Prediction sub-task, our submissions were based on four methods. The first method was based on the neural retrieval approaches of our previous work, as in the Concept Detection sub-task, with a new mechanism [10] that uses the retrieved captions. The second method was based on a state-of-the-art caption generated captions. The third method was based on the well-known Show and Tell model [12], which adopts a simple CNN-RNN architecture. The fourth and last method is an ensemble of our previous approaches, which selects its caption from the ones produced by the other three methods.

Following our previous successful entries in the competition [5, 6, 7, 8], our best performing systems were ranked 1st among 11 participating teams in Concept Detection, and 2nd among the 10 participating teams in Caption Prediction, yet also 1st when ranked with the secondary evaluation metric. The remainder of this article first describes the dataset and our methods for the two subtasks, followed by our submissions and an experimental analysis. The article concludes with our findings and suggested future directions.

# 2. Data

The ImageCLEFmedical Caption 2022 dataset is a subset of the Radiology Objects in Context (ROCO) [13] dataset, which contains medical images extracted from open access biomedical journal articles of PubMed Central.<sup>3</sup> The organisers state that this year's ImageCLEFmedical Caption dataset is an extended version of the dataset used in ImageCLEFmedical Caption 2020. There are images of several different modalities in the dataset (e.g., X-rays, CT-Scans), but no information was provided regarding the modality of each image. The same set of images is used for both Concept Detection and Caption Prediction.

# 2.1. Concept Detection

All the images in the dataset are accompanied by relevant UMLS [3] medical concepts and, more specifically, by their unique UMLS identifiers (CUIs), which are the ground truth for the Concept Detection sub-task. These ground truth concepts, which are essentially medical terms, were obtained from the respective image captions using multiple concept extraction methods,

<sup>&</sup>lt;sup>3</sup>https://www.ncbi.nlm.nih.gov/pmc/



**Figure 1:** Three images of the dataset [14, 15, 16](1st row) with their corresponding ground truth tags (2nd row) and captions (3rd row).

and were then manually curated. An image can be associated with multiple CUIs (see Fig. 1). Table 1 shows the 5 most frequent ground truth concepts of the whole dataset (official training & official validation splits), the corresponding UMLS terms, and how many training images they are assigned to. As in previous years, the dataset is highly imbalanced. For example, the concept X-RAY COMPUTED TOMOGRAPHY occurs 28,885 times in the whole dataset (Table 1), while 1,149 other concepts appear only three times each (see the right long tail in Fig. 2). Furthermore, there is a large number of unique concepts (number of available classes), which increased from 1,585 last year [5] to 8,374 this year. The average number of concepts assigned to each image is 4.74, while the minimum number is 1. There are 4,316 images with only one assigned concept, and only one image with 50 concepts (the maximum number of assigned concepts).

An official training set of 83,275 images and an official validation set of 7,645 images were provided by the organisers. The official test set comprised 7,601 images and its gold truth concepts were hidden. For our experiments, we merged the provided (official) training and validation sets, and used 15% of the merged data as our validation set, and another 20% of the merged data as our development set; the former was used for hyper-parameter tuning, whereas the latter was used to evaluate the performance of our models. The remaining 65% served as our training set.



**Figure 2:** Number of images (vertical axis) tagged with each concept in the training data. To save space, the horizontal axis shows the index of each concept (class index), instead of its CUI.

#### Table 1

The 5 most frequent concepts (CUIs and corresponding UMLS terms) in the official dataset of Image-CLEFmedical Caption 2022 and how many images they are assigned to.

CUI	UMLS term	Images
C0040405	X-Ray Computed Tomography	28,885
C1306645	Plain x-ray	26,412
C0024485	MAGNETIC RESONANCE IMAGING	15,693
C0041618	Ultrasonography	12,236
C0817096	Chest	8,030

## 2.2. Caption Prediction

The same images that were used for the Concept Detection sub-task (Sec. 2.2, Fig. 1) were also used for the Caption Prediction sub-task, but now each of these was also accompanied by a gold caption (the most frequent captions in the dataset are shown in Table 2). This year, approximately 97% (88,342 out of total 90,920) of the provided captions were unique (associated with only one image), whereas in last year's dataset only 65% of the captions were unique. This fact makes retrieval-based approaches less suitable for caption prediction this year. The maximum number of words (tokens) in a single caption was 391 (found in 1 image) and the average number of words per caption was 19.14. The histogram of the caption lengths (Fig. 3) indicates that the vast majority of captions are shorter than 50 words.

As in previous years, BLEU [17] was the main evaluation measure of the Caption Prediction sub-task. This year, ROUGE-1 [18] was also employed as a secondary measure. The organisers announced that the following pre-processing steps would be followed before computing the evaluation scores:

Table 2Most common captions in the dataset.

Caption	Occurrences
case a electrocardiogram with inferolateral early repolarization pattern with jpoint	261
elevation and qrs slur after hypothermia treatment red arrow	201
case telemetry tracing ventricular fibrillation precede by a ventricular extrasystole	179
case a a electrocardiogram with an aggressive inferiorlateral er pattern during	
hypothermia treatment red arrow b the electrocardiogram be completely normalise	115
after adminbetration of beoproterenol infusion	
the degenerative nuclear atypic area ancient modification he	61
chest xray	48



Captions' length distribution

Figure 3: Histogram of the length in tokens of all the official gold captions.

- Captions are converted to lower-case.
- All punctuation is removed and captions are tokenised using a particular tokeniser.<sup>4</sup>
- Stopwords are removed using NLTK's 'english' stop-word list.
- Spacy's lemmatiser is applied.<sup>5</sup>

We decided not to follow these pre-processing steps during the training of our models, to avoid discarding or distorting any potentially important words in the gold and generated captions, and to try to produce captions close to those medical experts generate.

Similarly to the Concept Detection sub-task (Sec. 2.2), we merged the official training and validation sets, creating our own data splits. Specifically, we used 6,000 instances for validation and development, 3,000 for each, and the remaining 84,920 instances were used for training. Again, we used our validation set to tune the hyper-parameters of our models, and the develop-

<sup>&</sup>lt;sup>4</sup>https://www.nltk.org/\_modules/nltk/tokenize/punkt.html#PunktLanguageVars.word\_tokenize. <sup>5</sup>https://spacy.io/api/lemmatizer.

### Table 3

Most common words found in the captions of the (whole) provided dataset, with and without stopwords. There are 19,217 words (w/ stopwords) with only 1 occurrence.

Most common words											
Word	the	of	shov	v and	d a	in	ı wi	th	be	arrow	right
Occurrences	129,758	84,428	3 41,36	4 40,00	,003 35,811 34,437 32,688 24,649 24		24,555	20,340			
Most common words (excluding stop-words)											
Word	show arrow right ct image left scan tomography chest						mass				
Occurrences	41,364	24,555	20,340	16,495	14,703	12,752	11,655		10,628	10,052	9,192

ment set for evaluation. In the training set, we applied a maximum threshold to remove the instances with captions longer than 80 words. This led to the removal of 517 training instances.

# 3. Methods

In this section, we describe the systems that were used in our submissions to the Concept Detection and Caption Prediction sub-tasks.

### 3.1. Concept Detection

We describe two Concept Detection systems, 2xCNN+FFNN, which follows the work of our past submissions [5, 6, 8] for the same task, and a retrieval-based system, dubbed CNN+wKNN.

#### 3.1.1. Concept Detection System 1: 2xCNN+FFNN

This system is an ensemble of two members that share the same architecture: an image encoder (e.g., a CNN), followed by a single dense layer with sigmoid activations (for multi-label classification).

In 2020, our best performing submission was an ensemble consisting of two instances of a CNN+FFNN classifier that employed a DenseNet-121 [19] image encoder, pre-trained on ImageNet [20]. We fine-tuned five instances on the task's data and kept the two that performed best on validation data. The two instances were combined by using the UNION and the INTERSECTION of their predicted concepts. The ensemble that used the INTERSECTION was ranked 1st. In 2021, we employed ResNet-50 [21] pre-trained on ImageNet [20] as the image encoder, and we integrated an extra pre-training step using supervised contrastive learning [22]. Then, we again fine-tuned and combined two instances using the UNION and the INTERSECTION of the predicted concepts of each instance. We had submitted the UNION combination as it performed better on our own test split. This system was ranked 8th in 2021, as it was outperformed by our retrieval-based winning systems.

This year's submissions employed an EfficientNetV2-B0 CNN [9], as the image encoder, pre-trained on ImageNet [20]. We extracted image embeddings (feature vectors) using the last convolutional layer of the backbone of the image encoder, followed by Generalized-Mean (GeM)

global pooling [23]. Given an input image, the output of the last convolutional layer is a 3D tensor X of  $W \times H \times K$  dimensions, where K is the number of feature maps (channels) in the layer. Let  $X_k$  be the k-th feature map of dimensions  $W \times H$  with  $k \in \{1, 2, ..., K\}$ . The GeM layer takes X as an input and returns a vector v as output:

$$v = [\tilde{v}_1 \dots \tilde{v}_k \dots \tilde{v}_K]^T, \ \tilde{v}_k = \left(\frac{1}{|X_k|} \sum_{x \in X_k} x^p\right)^{\frac{1}{p}}$$
(1)

where x is the value of the corresponding pixel of the k-th feature map  $X_k$ . v is the produced image embedding. The pooling parameter p can be trained along with the network or be manually set. The image embeddings were then passed through a dense layer with |C| outputs and sigmoid activations, where C is the set of all possible concepts, to produce a probability per label. The models were trained by minimizing the binary cross entropy loss of all concepts. We used Adam [24] as our optimizer and decreased the learning rate by a factor of 10 when the loss showed no improvement. We also used early stopping on the validation set, with patience of 3 epochs. For each instance of the system, a classification threshold for all the concepts was tuned by optimizing the F1 score on our validation data. Any concepts for which the respective model outputs exceeded that threshold, were assigned to the corresponding image during inference. We trained two instances of the same system by fine-tuning on the task's data and kept checkpoints from the two best (on validation data) epochs. Finally, in order to form the ensemble, we combined the two instances using the UNION and the INTERSECTION of their predicted concept sets. We call these models 2xCNN+FFNN@U and 2xCNN+FFNN@I, respectively.

#### 3.1.2. Concept Detection System 2: CNN+wKNN

Following our previous work [5, 6, 8, 7], this system employs a neural retrieval approach. Intuitively, 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. We used the image encoder of our fine-tuned CNN+FFNN system (see Sec. 3.1.1). We discarded the last dense layer of the classifier and used the last GeM pooling layer to extract embeddings (feature vectors) for all the training images. Given a test image, we used the same encoder to obtain its embedding (Fig. 4) and retrieved the (embeddings of the) k training images with the highest cosine similarity with the (embedding of the) test image. We tuned the value of k in the range from 5 to 100 with a step of 5 using our validation set, which led to k = 10. The voting scheme that we used was introduced in [25] and can be described as follows. Given a test image (query) x, for each concept  $c_i \in C$ , we calculate the weighted sum of k scores, from each of the k neighbors of x:

$$f_i(x) = \frac{\sum_{j=1}^k w_j \cdot y_i(N_k(x,j))}{\sum_{j=1}^k w_j}$$
(2)

where  $y_i(N_k(x, j))$  denotes the presence of the *i*-th concept in the multi-hot vector of the *j*-th nearest neighbor of *x*, and  $w_j$  is the weight assigned to the *j*-th nearest neighbor. The weights vector  $\langle w_1, \ldots, w_k \rangle$  can be learned or can be set manually with monotonically decreasing

weights, e.g., the top-ranked neighbor can be given weight k and the lowest-ranked neighbor weight 1. We used the latter simple linear assignment method in our experiments. We assigned concept  $c_i$  to test image x by the rule:

$$h_i(x) = \begin{cases} 1, & f_i(x) \ge 0.5\\ 0, & f_i(x) < 0.5 \end{cases}$$
(3)

yielding the predicted label set  $H(x) = \{c_i | h_i(x) = 1\} = \{c_i | f_i(x) \ge 0.5\}$ . We call this system CNN+wKNN.

Training images



**Figure 4:** Illustration of how the retrieval-based CNN+wKNN Concept Detection system (Sec. 3.1.2) works at inference time. The training image embeddings are computed offline.

### 3.2. Caption Prediction

In this subsection, we describe four Caption Prediction systems, Retrieval-based Approach, which follows the work of our past submissions [5, 6, 8], R2Gen & Image Clustering, which employs a memory-driven Transformer, CNN-RNN, an Encoder-Decoder model based on the Show&Tell model [26] and a Captions Ensemble system, which utilize the generated captions of the aforementioned systems.

### 3.2.1. Caption Prediction System 1: Retrieval-based Approach

Since our previous work had led to top performance with retrieval-based methods in Caption Prediction [5, 6, 8], we again explored retrieval-based methods, which are based on KNN. We used an image encoder, pre-trained on ImageNet [20], to obtain the embedding of each training image. The embedding was extracted from the last average pooling layer of the network. We experimented with several encoders. Specifically we experimented with DenseNet-121 [19], DenseNet-201 [19], EfficientNetB0 [27], ResNet50V2 [28], InceptionResNetV2 [29] and CotNet50 [30]. During inference, given a test image, we generate its embedding using the same encoder, and we retrieve the k most similar training images, based on the cosine similarity of their embeddings with the test image embedding. Following the Consensus Caption (CC) method

of [10], we then retrieve the captions of the k most similar training images, creating the set S. Among the captions in S, we select the caption  $c^*$  with the highest textual similarity with the other captions in S:

$$c^* = \operatorname*{argmax}_{c \in S} \sum_{c' \in S} \cos(c, c') \tag{4}$$

where  $\cos$  denotes the cosine similarity, calculated using the *TF-IDF* representations of the captions (the coefficients *TF* and *DF* were computed using only the retrieved captions). In effect, we select the caption closest to the centroid of the *k* retrieved captions as the prediction for the test image. An illustration of this approach can been seen in Fig. 5.



**Figure 5:** Illustration of how the retrieval-based Caption Prediction method (Sec. 3.2.1) works at inference time. The training image embeddings are again computed offline.

### 3.2.2. Caption Prediction System 2: R2Gen & Image Clustering

This system is based on a memory-driven Transformer and on image clustering. The former, R2Gen [11], is a captioning system that has been reported to achieve competitive performance on several medical captioning datasets. Originally, it employed a ResNet-101 [21], to extract image (patch) embeddings. However, we substituted it with a DenseNet-121 [19], pre-trained on ImageNet [20], which showed better performance in preliminary experiments. The representations resulted per image are then passed to a Transformer encoder-decoder [31], whose decoder is enhanced with a relational memory and a (memory-driven conditional) layer normalization.

As an extra step before training we clustered the embeddings of all the images, as these were generated by our DenseNet-121. We used k-Means,<sup>6</sup> with k = 8, defined by varying k from 5 to 9 and evaluating each clustering with Silhouette [32]. For each cluster, we used the training images to fit a separate R2Gen instance (Fig. 6). Provided a test image, we retrieve the R2Gen instance trained on the cluster the test image belongs to, and we use it to generate a caption. We experimented both, with and without image clustering.

<sup>&</sup>lt;sup>6</sup>https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html.



**Figure 6:** Illustration of the R2Gen Caption Prediction method with image clustering added, at inference time (Sec. 3.2.2).

#### 3.2.3. Caption Prediction System 3: CNN-RNN

This system is based on the Show&Tell model [12], which consists of a CNN-RNN encoderdecoder. For the CNN encoder, we experimented with DenseNet-121, DenseNet-201, Efficient-NetB0, ResNet50V2, InceptionResNetV2, and CotNet50. The RNN decoder is fed with the encoded image representation, and generates a caption word by word. In specific, the encoded image representation is concatenated with the hidden states of an encoding GRU [33], which operates on the unfinished generated caption. The result is fed onto a decoding GRU, whose last hidden state is followed by a FFNN, in order to yield a probability distribution over the vocabulary and decide the next word to output.

We pre-processed each training caption by adding a start and end token, at the beginning and end of the text, respectively. Then we created a vocabulary, keeping only the words that appeared 3 or more times in the training captions. The out-of-vocabulary words (OOV) were replaced by an <UNK> token. The maximum length was set to 40 tokens, based on preliminary experiments.

#### Table 4

Example of a text sequence input before and after the preprocessing used in the Caption Prediction System 3: CNN-RNN (see Sec. 3.2.3).

Before	digital subtraction angiography of the left pedal vessel
After	startsequence digital subtraction angiography of the left <unk> vessel endsequence</unk>

We experimented with both greedy decoding (selecting the vocabulary word with the highest predicted probability at each decoding step) and beam search decoding (searching for the most probable sequence of output words, as in machine translation) with different beam sizes (1, 2, 3, 5, 7). In addition, we implemented three different ensembles consisted of Show & Tell models with different encoders (the same we used in Sec. 3.2.3), which all used greedy decoding.

For the first ensemble, dubbed Maximum Probability (MP), at each decoding step we select the single word with the highest probability from the k probability distributions produced by the k ensemble members. For the second ensemble, Maximum Voting Probability (MVP) [34], at each decoding step we keep the most probable word from the probability distribution of each ensemble member as the word selected by that ensemble member, and we output the word that was selected by most ensemble members. For the third ensemble, Average Probability (AP) [34], at each decoding step, each word is assigned the average of the probabilities assigned to that word by the ensemble members. We then output the word with the highest average probability. An extension of this approach could use weighted probability averaging, with weights reflecting how well each ensemble member performs on its own on validation data, so that better ensemble members would influence more the predictions of the ensemble than worse ones.

In our experiments, beam search performed better (specifically with beam size 3 and 5) than all three ensembles (which all used greedy decoding; we did not use beam search in the ensembles).

### 3.2.4. Caption Prediction System 4: Captions Ensemble

In this approach, we utilised the captions generated by each one of the previous systems for a given image, to create an ensemble model. Specifically, for each test image, we gathered all the captions assigned to it by the aforementioned systems, and we applied the Consensus Caption (CC) technique (see Sec. 3.2.1) to select one of the gathered captions. For this system, we used ClinicalBERT [35] to extract caption embeddings instead of *TF-IDF* representations.



**Figure 7:** Illustration of the Captions Ensemble method of Caption Prediction, which utilises the Consensus Caption technique [10].

# 4. Submissions and results

In this Section, we provide the results of our experiments for both sub-tasks as well as some insight and comments upon those.

### 4.1. Concept Detection

We used our development set to evaluate all our models and submitted those that performed best. Two of our six submissions were the ensemble systems (2xCNN+FFNN@U & 2xCNN+FFNN@I) described in Sec. 3.1.1. Additionally, we submitted a CNN+wKNN system (Sec. 3.1.2) as well as another ensemble system that employed two instances of our CNN+FFNN (Sec. 3.1.1) and an instance of the CNN+wKNN system. This ensemble used a majority voting rule to make the concept predictions. That is, for every test image, if a concept is predicted at least by two of the three systems, it is assigned to the image. The sixth assignment consisted of a CNN+FFNN that was trained using the Sharpness-Aware Minimization (SAM) algorithm [36] and the Gradient Centralization (GC) [37] optimization technique in order to achieve better generalization performance. In principle, SAM tries to find parameter values whose entire neighborhoods have uniformly low training loss, rather than seeking parameter values that

simply have a low training loss. Meanwhile, GC centralizing the gradient vectors of weight matrices to have zero mean and acts as a regularizer [37]. We also experimented with a loss function that aimed to optimise  $F_1$ , which was the official evaluation measure of the task. Similarly to [38], we employed the 'soft  $F_1$  score' ( $sF_1$ ), a differentiable version of the  $F_1$  measure that computes true positives, false positives, and false negatives using the output probability distributions of the model, without applying any threshold to round them to binary decisions. In practice,  $1 - sF_1$  is used for minimization and the total loss that was used is:

$$L_{F_1} = (1 - sF_1) \cdot bce \tag{5}$$

where *bce* is the standard binary cross-entropy loss used in the multi-label setting (i.e., summed for all labels).

As mentioned, the primary measure of the competition was  $F_1$ , calculated by comparing the binary multi-hot vectors  $y_{true}$  (the ground truth) and  $y_{pred}$  (the predicted concepts) of each test image and then macro-averaging over all test images. In addition to the primary measure, this year's task included a secondary  $F_1$  score that was calculated using a subset of manually validated concepts (only anatomy and image modality ones). To generate the predictions for the official test set, we merged our training, validation and development data. We used a held-out set (15% of the merged data) as our final development set for hyper-parameter tuning and trained the models using the rest of the data.

#### Table 5

Scores and rankings of experiments on our final development and the official test sets. Systems that were not submitted do not have test scores and rankings available. cd3's development score is also not available, as its predictions were created using the output files of the ensemble's members.

ID Run ID		Approach	Primary I	71	Secondary F1	Rank	
		Арргоасн	Development	Test	Secondary 11	ITAIIK	
cd1	182358	2xCNN[9]+FFNN@U	47.01	45.11	79.07	1	
cd2	182356	2xCNN[9]+FFNN@I	46.32	44.27	84.42	12	
cd3	182359	2xCNN[9]+FFNN/wKNN	-	44.63	84.30	5	
cd4	182340	CNN[9]+FFNN	45.71	44.39	81.20	9	
cd5	182354	$CNN[9]+FFNN(SAM+GC+L_{F1})$	46.06	45.02	82.14	3	
cd6	182333	CNN[9]+wKNN	45.25	43.05	84.40	28	
cd7	-	CNN[9]+FFNN(BAYES OPT.)	45.79	_	-	_	
cd8	-	CNN[39]+FFNN	45.27	-	-	-	
cd9	-	$CNN[39]+FFNN(SAM+GC+L_{F1})$	45.81	-	-	-	
cd10	-	VIT-B[40]+FFNN	42.09	-	-	_	
cd11	-	$VIT-B[40]+FFNN(SAM+GC+L_{F1})$	44.70	-	-	_	
cd12	-	CoAtNet-0[41]+FFNN	43.65	-	-	_	
cd13	-	CNN[9]+FFNN(+KNN)	43.62	_	_	_	

Table 5 presents the scores of our submitted systems on our final development and the official test sets, as well as their official rankings. 2xCNN+FFNN@U had the best results. The table also includes systems that were not submitted and are parts of further experiments that we conducted. We experimented with several image encoders ranging from pure CNNs [39] to Vision Transformers (ViTs) [40] and hybrid versions of the two [41]. Overall, pure convolutional

encoders yielded better results and more particularly, EfficicientNetV2-B0 had the best score, so we only used this specific image encoder for our ensemble and CNN+wKNN systems.

We also experimented with tuning the decision thresholds of the classifiers (probability thresholds for assigning each concept or not). Instead of tuning a single decision threshold value (the same for all concepts), we tried to tune a different threshold value for each concept. Due to the large size of the concept set (|C| = 8,374) and the large search space for each threshold, we employed Bayesian Optimization [42] instead of a full parameter sweep and used the task's primary measure as the evaluation function.

Furthermore, we observed that, despite of their high performance in terms of precision, our systems yielded low recall scores, which can be explained by the fact that the number of the predicted concepts was very low compared to the total number of possible concepts (i.e., only a few hundreds vs. thousands). In order to alleviate this problem, we experimented with a retrieval-augmented classification system. We employed a simple CNN+FFNN model and added to its predicted concepts the concepts of a simple KNN system. That is, every test image was passed through CNN+FFNN and the predicted concepts were added (set union) to the concepts predicted by the KNN system. To generate the latter predictions, we retrieved the top-k closest training images (w.r.t. cosine similarity computed on the image embeddings) of the test image and returned the 2 concepts that were most frequently assigned to the k images. We used a kof value 10. In spite of producing a higher recall score, the retrieval-augmented classification system scored much lower precision-wise leading to a worse  $F_1$  score than our other systems and thus, it was not submitted. In general, we aimed to deal with the class imbalance of the dataset and experimented with additional loss functions suitable for imbalanced multi-label tasks, such as the ASL [43] and Focal [44] losses, but these experiments did not yield better results either.

### 4.2. Caption Prediction

For Caption Prediction, in addition to the BLEU measure [17] of previous years, this year the organisers added ROUGE-1 [18] as a secondary measure. For the former, the organizers clarified that BLEU is calculated for up to 4-grams, using uniform weights (this is called BLEU-4). For the latter, they used ROUGE-1, which considers the overlap of unigrams between the generated caption and the gold-truth caption.<sup>7</sup> Therefore, we used these two measures to evaluate our models and decide which ones we were going to submit.

For the ensemble KNN model (Sec. 3.2.1), we combined our training and validation sets, to have more images to retrieve from. The k hyper-parameter was tuned in the range [1, 100] on our development set, and the best k was 18. All k values can been seen in Table 6. It is worth mentioning that this model did not yield the anticipated results for the secondary ROUGE-1 measure. For our first submission, we used an ensemble of KNNs (Caption Prediction System 1: Retrieval-based Approach Sec. 3.2.1) with different image encoders (the same ones we used in Sec. 3.2.3). During inference, for each test image, we collected the captions produced by each ensemble member (for the test image) and selected a single caption from them using the CC method (Sec. 3.2.1).

<sup>&</sup>lt;sup>7</sup>https://github.com/google-research/google-research/tree/master/rouge.

#### Table 6

ID	Approach	Development			
ID	Арргоасн	BLEU	ROUGE-1		
cp1	DenseNet121 KNN (best k=28)	0.3166	0.1117		
cp2	DenseNet201 KNN(best k=22)	0.3019	0.1139		
ср3	EfficientNetB0 KNN (best k=16)	0.3165	0.1276		
cp4	ResNet50V2 KNN (best k=26)	0.3189	0.1226		
cp5	InceptionResNetV2 KNN (best k=31)	0.2964	0.0981		
cp6	DenseNet121@CNN-RNN - BS3	0.3029	0.1567		
cp7	DenseNet201@CNN-RNN - BS3	0.3054	0.1578		
cp8	EfficientNetB0@CNN-RNN - BS3	0.3109	0.1587		
cp10	ResNet50V2@CNN-RNN - BS3	0.3002	0.1589		
cp11	InceptionResNet50V2@CNN-RNN - BS3	0.2987	0.1382		
cp12	DenseNet121@CNN-RNN - BS5	0.3189	0.1467		
cp13	DenseNet201@CNN-RNN - BS5	0.3116	0.1598		
cp14	EfficientNetB0@CNN-RNN - BS5	0.3280	0.1678		
cp15	ResNet50V2@CNN-RNN - BS5	0.3145	0.1532		
cp16	InceptionResNet50V2@CNN-RNN - BS5	0.3021	0.1243		
cp17	ResNet101@R2Gen (Authors)	0.2885	0.1490		
cp18	DenseNet121@R2Gen (Best split)	0.3089	0.1938		
cp19	DenseNet121@R2Gen (2nd Best split)	0.3021	0.1939		
cp20	DenseNet121@R2Gen (3rd Best Split)	0.3000	0.1998		
cp21	DenseNet121@ImageClustering + R2Gen	0.3183	0.1892		
cp22	Ensembles KNN (best k=18) (CC)	0.3196	0.1267		
cp23	Ensembles Greedy Search (MP)	0.2938	0.1660		
cp24	Ensembles Greedy Search (MVP)	0.2848	0.1629		
cp25	Ensembles Greedy Search (AP)	0.2854	0.1647		

The BLEU and ROUGE-1 scores from all of our experiments on our development set.

For the R2Gen model without the image clustering step (see Sec. 3.2.2), we performed a K-fold Cross Validation (CV) on our sets. Then, we kept the best 3 models from the Fold-splits with the best BLEU scores, on their corresponding Fold-split development sets. Two of our submissions consisted of model instances that were trained on the training set of the two best fold-splits. One of our submissions is an ensemble that utilizes the models from the three best fold-splits, and other submissions (see Table 7 for more detail). It uses the Captions Ensemble method (see Sec. 3.2.4) to decide the final predictions. These submissions are shown in Table 7.

For models based on the CNN-RNN system (Sec. 3.2.3), we only considered beam search decoding for the submissions (BS*m* for short, where *m* is the beam size), as well as and ensembles with greedy decoding (MP, MVP, AP – see Sec. 3.2.3). An interesting point about CNN-RNN models is that whenever we observed an increase in the primary BLEU score, a decrease was detected in the secondary ROUGE-1 score. Consequently, we plan to further investigate the generated captions step by step to conduct an exploratory error analysis and shed more light on this phenomenon.

For the Captions Ensemble method (Sec. 3.2.4, Fig. 7), we used ClinicalBERT instead of *TF-IDF*, as already noted, which worked better. This may be due to the fact that ClinicalBERT is a BERT model [45] pre-trained on numerous medical texts. The Captions Ensemble models we submitted are listed in Table 7.

Lastly, an interesting observation, is that encoders with complex architectures performed worse than encoders with fewer parameters. Hence, we did not use CotNet50 [30] as the backbone encoder in our systems and Ensemble models.

#### Table 7

Our 9 submissions to the ImageCLEFmedical Caption Prediction sub-task, along with their rank on all submission runs. The development scores of submissions cp26, cp27 and cp28 are not available due to the fact that their caption predictions were created using other submission files. All these submissions are based on System 4 (Captions Ensemble method, Sec. 3.2.4)

ID Run	Dam ID	in ID Approach	Deve	lopment		Donk	
	Kun ID		BLEU	ROUGE-1	BLEU	ROUGE-1	Nalik
cp14	181853	EfficientNetB0@CNN-RNN - BS5	0.3280	0.1678	0.3221	0.1664	11
cp26	182129	Ensemble of cp8, cp14, cp22, cp28 (CC)	-	-	0.3195	0.1817	12
cp27	182100	Ensemble of cp8, cp14, cp18, cp19, cp20, cp22 (CC)	-	-	0.3166	0.1991	13
cp22	181285	Ensemble KNN (best k=18) (CC)	0.3196	0.1267	0.3126	0.1177	14
cp8	181488	EfficientNetB0@CNN-RNN - BS3	0.3109	0.1587	0.3086	0.1741	21
cp28	182287	Ensemble of cp18, cp19, cp20, cp27 (CC)	-	-	0.3084	0.2062	22
cp18	182052	DenseNet121@R2Gen (Best split)	0.3089	0.1938	0.2960	0.2013	29
cp19	181357	DenseNet121@R2Gen (2nd Best split)	0.3021	0.1939	0.2895	0.2051	32
cp21	181536	DenseNet121@ImageClustering + R2Gen	0.3183	0.1892	0.2741	0.1760	42

Our team officially ranked 2nd among 10 teams in the Caption Prediction sub-task. Our best model was EFFICIENTNETB0@CNN-RNN - BS5, which is based on System 3 (Sec. 3.2.3) and employed EfficientNet-B0 [27] as the image encoder. We also ranked 1st in the secondary metric (ROUGE-1) according to the official results, by using an ensemble of cp18, cp19, cp20 and cp27 with the Consensus Caption method (Sec. 3.2.4). The table with all official measures is provided in the Appendix A

# 5. Conclusions and future work

We described the submissions of AUEB's NLP Group to the 2022 ImageCLEFmedical Caption subtasks, Concept Detection and Caption Prediction. In Concept Detection, we ranked 1st amongst 11 teams. Our top system was an ensemble of two CNN+FFNN multi-label classifiers, which employed an EfficientNetV2-B0 [9] image encoder. Our submissions also included classifiers trained with different optimization techniques [36, 37] and objectives, as well as a neural retrieval approach that was again competitive, as in previous years. In Caption Prediction, we ranked 2nd amongst 10 teams, by using Show and Tell [12], with EfficientNet-B0 [27] for image encoding and a GRU [33] for text decoding. Our analysis included experiments with R2Gen [11], combined with image clustering, and with a neural retrieval approach that was based on prior work [5, 6, 8]. In future work, we aim to investigate more neural retrieval methods and to explore multi-modal approaches that incorporate information from both images and text [46].

# References

- [1] B. Ionescu, H. Müller, R. Péteri, J. Rückert, A. Ben Abacha, A. G. S. de Herrera, C. M. Friedrich, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, S. Kozlovski, Y. D. Cid, V. Kovalev, L. D. Ştefan, M. G. Constantin, M. Dogariu, P. A, J. Deshayes-Chossart, H. Schindler, J. Chamberlain, A. Campello, A. Clark, Overview of the ImageCLEF 2022: Multimedia retrieval in medical, social media and nature applications, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the 13th International Conference of the CLEF Association (CLEF 2022), LNCS Lecture Notes in Computer Science, Springer, Bologna, Italy, 2022.
- [2] J. Rückert, A. Ben Abacha, A. García 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: CLEF2022 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Bologna, Italy, 2022.
- [3] O. Bodenreider, The Unified Medical Language System (UMLS): integrating biomedical terminology, Nucleic acids research 32 (2004) 4.
- [4] J. Pavlopoulos, V. Kougia, I. Androutsopoulos, D. Papamichail, Diagnostic captioning: A survey, 2021. arXiv:2101.07299.
- [5] F. Charalampakos, V. Karatzas, V. Kougia, J. Pavlopoulos, I. Androutsopoulos, AUEB NLP Group at ImageCLEFmed Caption Tasks 2021, in: CLEF2021 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Bucharest, Romania, 2021, pp. 1184–1200.
- [6] B. Karatzas, V. Kougia, J. Pavlopoulos, I. Androutsopoulos, AUEB NLP Group at Image-CLEFmed Caption 2020, in: CLEF2020 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece, 2020, p. 10.
- [7] V. Kougia, J. Pavlopoulos, I. Androutsopoulos, Medical Image Tagging by Deep Learning and Retrieval, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction Proceedings of the Eleventh International Conference of the CLEF Association (CLEF 2020), Thessaloniki, Greece, 2020, pp. 154–166.
- [8] V. Kougia, J. Pavlopoulos, I. Androutsopoulos, AUEB NLP Group at ImageCLEFmed Caption 2019, in: CLEF2019 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Lugano, Switzerland, 2019, p. 8.
- [9] M. Tan, Q. Le, Efficientnetv2: Smaller models and faster training, in: Proceedings of the 38th International Conference on Machine Learning, volume 139, Online, 2021, pp. 10096–10106.
- [10] J. Devlin, S. Gupta, R. Girshick, M. Mitchell, C. L. Zitnick, Exploring nearest neighbor approaches for image captioning, arXiv preprint arXiv:1505.04467 (2015).
- [11] Z. Chen, Y. Song, T. H. Chang, X. Wan, Generating radiology reports via memory-driven transformer, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 2020, pp. 1439–1449.
- [12] 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, Boston, MA, USA, 2015, pp. 3156–3164.
- [13] O. Pelka, S. Koitka, J. Rückert, F. Nensa, C. M. Friedrich, Radiology Objects in COntext (ROCO): A Multimodal Image Dataset, volume 11043, Lecture Notes in Computer Science

(LNCS), 2018, pp. 180-189.

- [14] I. Martinovic, H. Greve, Surgical reconstruction of the left main coronary artery with patch-angioplasty, J Cardiothorac Surg 6 (2011) 24.
- [15] B. Shetty, P. K. Shetty, P. Sharma, Ascariasis cholecystitis: An unusual cause, J Minim Access Surg 4 (2008) 108–110.
- [16] T. Phulambrikar, S. V. Kant, M. Kode, S. Magar, Cone Beam Computed Tomography Findings in Calcifying Cystic Odontogenic Tumor Associated with Odontome: A Case Report, J Dent (Shiraz) 16 (2015) 374–379.
- [17] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th annual meeting of the Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 2002, pp. 311–318.
- [18] C. Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text Summarization Branches Iut, Barcelona, Spain, 2004, pp. 74–81.
- [19] G. Huang, Z. Liu, L. V. D. Maaten, K. Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 2017, pp. 4700–4708.
- [20] J. Deng, W. Dong, R. Socher, L. Li, L. Kai, F. Li, Imagenet: A large-scale hierarchical image database, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Miami Beach, FL, USA, 2009, pp. 248–255.
- [21] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 2016, pp. 770–778.
- [22] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, D. Krishnan, Supervised contrastive learning, in: Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 18661–18673.
- [23] 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.
- [24] D. P. Kingma, J. Ba, Adam: A Method for Stochastic Optimization, in: 3rd International Conference on Learning Representations (ICLR), San Diego, CA, USA, 2015.
- [25] T. Chiang, H. Lo, S. 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.
- [26] 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, Lille, France, 2015, pp. 2048–2057.
- [27] M. Tan, Q. Le, Efficientnet: Rethinking model scaling for convolutional neural networks, in: International conference on machine learning, Long Beach, CA, USA, 2019, pp. 6105–6114.
- [28] K. He, X. Zhang, S. Ren, J. Sun, Identity mappings in deep residual networks, in: European Conference on Computer Vision, Amsterdam, NL, 2016, pp. 630–645.
- [29] C. Szegedy, S. Ioffe, V. Vanhoucke, A. A. Alemi, Inception-v4, inception-resnet and the impact of residual connections on learning, in: AAAI Conference on Artificial Intelligence, San Fransicso, CA, USA, 2017.
- [30] Y. Li, T. Yao, Y. Pan, T. Mei, Contextual transformer networks for visual recognition, IEEE

Transactions on Pattern Analysis and Machine Intelligence (2022).

- [31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).
- [32] P. J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, Journal of Computational and Applied Mathematics 20 (1987) 53–65.
- [33] J. Chung, C. Gulcehre, K. Cho, Y. Bengio, Empirical evaluation of gated recurrent neural networks on sequence modeling, in: NIPS 2014 Workshop on Deep Learning, Montreal, Canada, 2014.
- [34] H. Katpally, A. Bansal, Ensemble learning on deep neural networks for image caption generation, in: 2020 IEEE 14th International Conference on Semantic Computing (ICSC), San Diego, CA, USA, 2020, pp. 61–68.
- [35] E. Alsentzer, J. Murphy, W. Boag, W. H. Weng, D. Jindi, T. Naumann, M. McDermott, Publicly available clinical BERT embeddings, in: Proceedings of the 2nd Clinical Natural Language Processing Workshop, Minneapolis, Minnesota, USA, 2019, pp. 72–78.
- [36] P. Foret, A. Kleiner, H. Mobahi, B. Neyshabur, Sharpness-aware minimization for efficiently improving generalization, in: International Conference on Learning Representations, 2021.
- [37] H. Yong, J. Huang, X. Hua, L. Zhang, Gradient-centralization: A new optimization technique for deep neural networks (2020) 635–652.
- [38] M. Kalimuthu, F. Nunnari, D. Sonntag, A Competitive Deep Neural Network Approach for the ImageCLEFmed Caption 2020 Task, in: CLEF2020 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece, 2020.
- [39] Z. Liu, H. Mao, C. Wu, C. Feichtenhofer, T. Darrell, S. Xie, A convnet for the 2020s, CoRR abs/2201.03545 (2022). arXiv:2201.03545.
- [40] 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: 9th International Conference on Learning Representations, ICLR 2021, Online, 2021.
- [41] Z. Dai, H. Liu, Q. V. Le, M. Tan, Coatnet: Marrying convolution and attention for all data sizes, in: Advances in Neural Information Processing Systems, volume 34, Curran Associates, Inc., 2021, pp. 3965–3977.
- [42] J. Snoek, H. Larochelle, R. P. Adams, Practical bayesian optimization of machine learning algorithms, in: Advances in Neural Information Processing Systems, volume 25, Lake Tahoe, NV, USA, 2012.
- [43] E. B. Baruch, T. Ridnik, N. Zamir, A. Noy, I. Friedman, M. Protter, L. Zelnik-Manor, Asymmetric loss for multi-label classification, 2021 IEEE/CVF International Conference on Computer Vision (ICCV) (2021) 82–91.
- [44] T. Lin, P. Goyal, R. B. Girshick, K. He, P. Dollár, Focal loss for dense object detection, 2017 IEEE International Conference on Computer Vision (ICCV) (2017) 2999–3007.
- [45] J. Devlin, M. W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 4171–4186.
- [46] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, I. Sutskever, Zero-

shot text-to-image generation, in: International Conference on Machine Learning (ICML), Online, 2021, pp. 8821–8831.

# A. Caption Prediction results on all official measures

## Table 8

The submissions to the ImageCLEFmedical Caption Prediction Task, along with their results on all official measures.

ID Run ID	Pup ID	Approach	Test						Dank
	Арргоасн	BLEU	ROUGE-1	METEOR	CIDEr	SPICE	BERTScore	Kalik	
cp14	181853	EfficientNetB0@CNN-RNN - BS5	0.3221	0.1664	0.0737	0.1902	0.0312	0.5988	11
cp26	182129	Ensemble of cp8, cp14, cp22, cp28 (CC)	0.3195	0.1817	0.0777	0.2235	0.0344	0.6089	12
cp27	182100	Ensemble of cp8, cp14, cp18, cp19, cp20, cp22 (CC)	0.3166	0.1991	0.0834	0.2658	0.0427	0.6163	13
cp22	181285	Ensemble KNN (best k=18) (CC)	0.3126	0.1177	0.0621	0.0923	0.0199	0.5732	14
cp8	181488	EfficientNetB0@CNN-RNN - BS3	0.3086	0.1741	0.0729	0.2123	0.0308	0.6035	21
cp28	182287	Ensemble of cp18, cp19, cp20, cp27 (CC)	0.3084	0.2062	0.0846	0.2815	0.0467	0.6187	22
cp18	182052	DenseNet121@R2Gen (Best split)	0.2960	0.2013	0.0822	0.2709	0.0470	0.6130	29
cp19	181357	DenseNet121@R2Gen (2nd Best split)	0.2895	0.2051	0.0823	0.2802	0.0487	0.6156	32
cp21	181536	DenseNet121@ImageClustering + R2Gen	0.2741	0.1760	0.0700	0.2064	0.0352	0.5957	42