AUEB NLP Group at ImageCLEFmed Caption Tasks 2021

Foivos Charalampakos[1], Vasilis Karatzas[1], Vasiliki Kougia[2,1], John Pavlopoulos[2,1] and Ion Androutsopoulos[1]

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

Abstract

This paper presents the systems that were implemented for the participation of AUEB's NLP Group in the sub-tasks of the 2021 ImageCLEFmed Caption task, which included a Concept Detection and a Caption Prediction sub-task. The goal of the Concept Detection sub-task was to identify medical terms that best describe each image, as a first step towards generating image captions or to improve the interpretation of medical images and help medical diagnosis. The objective of the Caption Prediction sub-task was to generate captions that describe medical images, which could assist medical experts in analyzing those images. The systems we implemented extend our previous work. For the Concept Detection sub-task, they employ convolutional neural network (CNN) image encoders, combined with an image retrieval module or a feed-forward neural network (FFNN) classifier. For the Caption Prediction we employed similar image encoders with image retrievals modules, and text generation models that either utilize the images or not. We ranked 1st in Concept Detection, and 2nd in Caption Prediction.

Keywords

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

1. Introduction

ImageCLEF [1] is an evaluation campaign that is annually organized as part of CLEF.[1] In the 2021 edition, ImageCLEF consisted of 4 main tasks with ImageCLEFmed being one of them. ImageCLEFmed consists of a series of tasks that are associated with the study of medical images. The ImageCLEFmed Caption Task [2] ran for the 5th year in 2021. It included a Concept Detection sub-task, that concerns multi-label classification of medical images by assigning medical terms (called concepts) to each image. The concepts stem from the Unified Medical Language System (UMLS) [3].[2] Selecting the appropriate concepts can be a first step towards automatically generating coherent image captions and assisting the medical experts by reducing

CLEF 2021 – Conference and Labs of the Evaluation Forum, September 21–24, 2021, Bucharest, Romania
phoebuschar@gmail.com (F. Charalampakos); karatzas.basil@gmail.com (V. Karatzas); kouyiav@aueb.gr (V. Kougia); annis@aueb.gr (J. Pavlopoulos); ion@aueb.gr (I. Androutsopoulos)
http://www.aueb.gr/users/ion/ (I. Androutsopoulos)
0000-0002-0172-6917 (V. Kougia); 0000-0001-9188-7425 (J. Pavlopoulos)

[1]http://www.clef-initiative.eu/
[2]https://www.nlm.nih.gov/research/umls/

the time needed for the diagnosis. This year, ImageCLEFmed also included a Caption Prediction sub-task [2, 4, 5], which was not included in the two previous years. This sub-task, that we also competed in, aims at assisting medical experts in interpreting radiology images, by automatically generating captions of the images. This kind of technology may eventually be able to generate draft diagnostic text from a medical image, which could help medical experts analyze more efficiently the large volumes of medical images (e.g., X-rays, MRI scans) they confront [6].

In this paper, we describe the systems of the AUEB NLP Group that were submitted to the ImageCLEFmed Caption sub-tasks, which extend our previous work [7, 8, 9, 10]. For Concept Detection, our submissions were based on three methods. The first method extended the retrieval component of [8, 10] and consisted of an ensemble of several 1-NN systems that use different image encoders. The second method extended the TagCXN classification system of [8, 9, 10], now using a ResNet-50 convolutional neural network (CNN) [11] to encode the images and a feed-forward neural network (FFNN) classifier on top. This year, we implemented a pre-training stage, where we trained the CNN encoder using supervised contrastive learning [12]. The third method consists of a combination of the previous two in which, for each test image, we used a similarity threshold to decide whether to employ the retrieval or the classification component. For Caption Prediction, we observed that image retrieval approaches were the best. We also tried text generation baselines that did not use information from the image and text generation models that did. For image retrieval, we used a k-NN model and experimented with different image encoders and ways of combining the captions returned. The Concept Detection task included the same images as the Caption Prediction task, so our team utilized the encoders trained for the Concept Detection task. As text generation baselines, we employed Bidirectional Encoder Representations from Transformers (BERT) [13], Generative Pre-trained Transformer 2 (GPT-2) [14], and GPT Neo, but the former one was dropped, because it was outperformed by the other two in early development stages. For the image-aware text generation models, we implemented both a simplistic architecture and an architecture based on Show, Attend and Tell [15].[3] Following the success of previous years [8, 9, 10], our best performing systems were ranked 1st among the best performing systems of 5 participating teams in Concept Detection, and 2nd among the best performing systems of 8 participating teams in Caption Prediction. The rest of the paper provides insights about the dataset, a description of the methods we used, experimental results and discussion.

2. Data

The ImageCLEFmed Caption 2021 dataset includes real clinical radiology images from original medical cases that were annotated by medical doctors. The organizers state that the dataset is the same with the one that was used in the ImageCLEF-VQAMed 2020 task [16]. The images stem from the Med-Pix database.[4] Additionally, a subset of the Radiology Objects in COntext (ROCO) dataset [17] is provided for training purposes [2]. The ROCO dataset consists of medical images extracted from open access biomedical journal articles of PubMed Central.[5] The same dataset is

[3]https://huggingface.co/transformers/master/model_doc/gpt_neo.html
[4]https://medpix.nlm.nih.gov/
[5]https://www.ncbi.nlm.nih.gov/pmc/

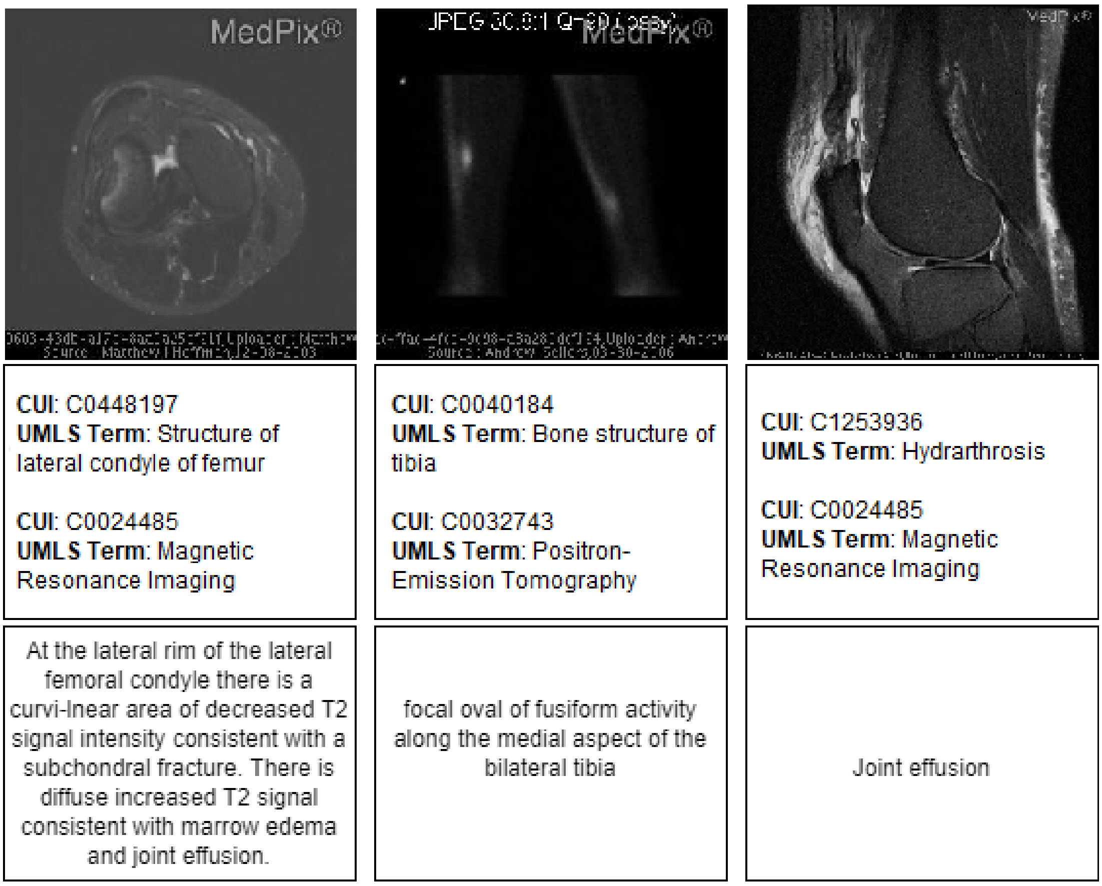

Figure 1: Three images of the dataset (1st row) with their corresponding tags (2nd row) and captions (3rd row).

used for the Caption Prediction sub-task. There are several different modalities present in the dataset (e.g., X-rays, CT-Scans), but unlike the previous year, this year's dataset was not provided classified into different modality categories.

2.1. Concept Detection

All images in the dataset are accompanied by the unique identifier (CUI) of the UMLS [3] concepts. These concepts, essentially medical terms, were extracted from the processed text of the respective image caption. An image can be associated with multiple CUIs (see Fig 1). Also, Table 1 shows the 5 most frequent concepts of the training set and how many training images they were assigned to.[6] It has to be noted that, although the dataset is imbalanced (e.g., 632 concepts appear only one time, 1 concept appears 1,400 times), there is no concept appearing in

[6]We used the REST API (https://documentation.uts.nlm.nih.gov/rest/home.html) of the UMLS Metathesaurus (uts.nlm.nih.gov/home.html) to map each CUI to its term.

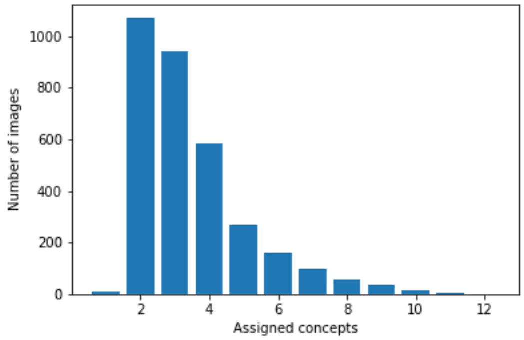

Figure 2: Distribution of assigned concepts in the data.

Table 1
The 5 most frequent concepts (CUIs) in the training set of ImageCLEFmed Caption 2021 and how many training images they are assigned to.

CUI	UMLS term	Images
C0040398	TOMOGRAPHY, EMISSION-COMPUTED	1,400
C0024485	MAGNETIC RESONANCE IMAGING	796
C1306645	PLAIN X-RAY	627
C0041618	ULTRASONOGRAPHY	373
C0009924	CONTRAST MEDIA	283

every image, unlike in the last year's dataset.

The number of unique concepts was reduced once again compared to previous years. There were 111,156 possible concepts in 2018 [18], 5,528 in 2019 [8], 3,047 in 2020 [10] and 1,585 in 2021. The average number of concepts assigned to each image was 3.48. The minimum number of concepts assigned to an image was 1 (found in 10 images), and the maximum was 12 (found in 2 images). Fig. 2 shows the distribution of the assigned concepts.

A training set of 2,756 images and a validation set of 500 images were provided. A separate test set comprised of 444 images and the concepts for the test images were unknown. For our experiments, we merged the provided training and validation sets and used 10% of the merged data as our validation set, and another 10% of the merged data as our development set in which we evaluated the performance of our models. The remaining 80% served as the training set.

2.2. Caption Prediction

The maximum number of words in a caption was 43 (in 10 images), while the minimum number of words was 1 (also in 10 images). The distribution of the number of words per image caption can be seen in Figure 3. The total number of distinct words, after lower-casing, was 3,515. In Table 2 we can see the most common words in the merged dataset (training and validation sets combined) and in Table 3 the most common captions. Out of the 3,256 captions of the dataset,

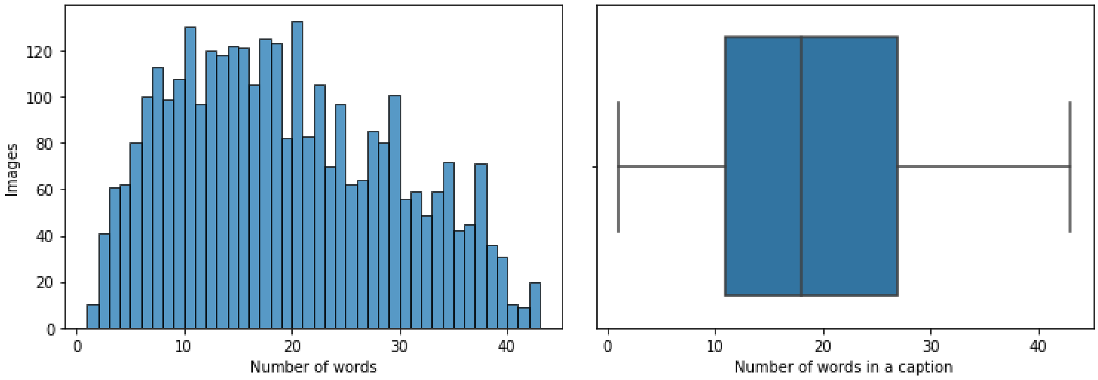

Figure 3: On the left, we show a histogram of the number of images that contain captions of a specific length. On the right, we show a boxplot of the number of words in the captions.

1,141 (about 35%) were duplicates (existed, exactly the same, as another image's caption), which indicated that retrieval-based approaches would probably be advantageous.

Table 2
The 10 most common words found in the captions of the (whole) dataset, w/ and w/o stopwords. With stopwords, there are 1,071 words with only 1 occurrence.

Most common words w/ stopword										
Word	the	of	with	and	a	right	in	left	to	mass
Occurrences	2,139	1,770	1,179	1,149	891	800	763	666	630	621

Most common words w/o stopwords										
Word	right	left	mass	ct	demonstrates	axial	images	image	contrast	within
Occurrences	800	666	621	616	511	451	385	379	365	302

For the calculation of the final score, which is calculated with the BLEU-4 metric, the organizers mention that the captions (both the predicted, and the gold ones) pass through these preprocessing steps:

1. Each text is lower-cased.
2. Punctuation is removed and the text is tokenised.[7]
3. Stopwords are removed, using the NLTK "english" stopword list.
4. Stemming is applied, using the Snowball Stemmer from NLTK.[8]

From now on, whenever we mention the preprocessing we will actually refer to bullets 2-4. This preprocessing raises a question: should we train our models on the original text or perform the preprocessing before the training? The second approach seems to be away from our main goal which is to help medical experts by giving them comprehensive captions that describe the

[7] http://www.nltk.org/_modules/nltk/tokenize/punkt.html#PunktLanguageVars.word_tokenize
[8] http://www.nltk.org/_modules/nltk/stem/snowball.html

Table 3
The 5 most common captions of the dataset.

Caption	Occurrences
fusion of multiple disc spaces squaring of the vertebral bodies fusion of si joints	14
extensive white matter lesions involving both cerebral hemispheres	11
fracture through the left c4 lateral mass and laminar arch with unilateral perched c45 facets on the right herniated and disrupted disk c45 torn intracapsular ligaments and ligament flavum	11
multilevel vertebral body lesions which are low signal on t1 and t2 scan sequences mediastinal adenopathy and perihilar nodular infiltrates on ct of chest	11
traumatic dislocation cervical spine at c1c2 level with marked widening of disc space and facet joints soft tissue edema anterior to spine and in posterior paraspinal locations edema and hemorrhage noted in lower medulla and upper cervical cord	10

Table 4
Example caption without (1st row) and with preprocessing (2nd row).

NORMAL	focal oval of fusiform activity along the medial aspect of the bilateral tibia
PREPROCESSED	focal oval fusiform activ along medial aspect bilater tibia

images. By predicting already preprocessed sentences, we may lose semantic value (see Table 4), thus one could argue that the preprocessing before the evaluation and the metric used for such a task should not reward models that generate preprocessed captions.

As in the other sub-task, we merged the given training and validation sets. We then used 60% of the merged data as our training set; 20% as our development set, to tune any hyperparameters; and 20% as our validation set, to test our models to decide which one we should submit.

3. Methods

In this section, we describe the systems that were used in our submissions.

3.1. Concept Detection

3.1.1. System 1: 2xCNN+FFNN

This system constitutes an extension of our previous work [10] for the same task. Last year's implementation was an ensemble of two instances of a classifier that employed a DenseNet-121 [19] backbone encoder pre-trained on ImageNet [20]. The classifier was fine-tuned on the task's data and the two instances were combined using the UNION and the INTERSECTION of the predicted concepts. The ensemble that used the INTERSECTION was ranked 1st in 2020.

This year, a pre-training step that used supervised contrastive learning was integrated in the pipeline. Similarly to [12], we used a ResNet-50 [11] with weights initialized from ImageNet

[20] as our backbone encoder, which mapped each image to a dense representation vector of size $D_E = 2048$, and normalized it to the unit hypersphere before passing it through a projection (a single dense layer) of size $D_P = 128$. We trained the model for 300 epochs using supervised contrastive loss with a temperature value of $\tau = 0.1$ and Adam [21] with its default hyper-parameters as the optimizer. For the contrastive pre-training, we created a larger augmented version of the training set where, for each image, we added four additional variations: we split the image in four horizontal patches and applied Gaussian blur in each patch separately, resulting in four extra noisy images. Furthermore, the pre-training data included random horizontal flip and random rotation.[9] Contrastive pre-training aimed to bring visual representations belonging in the same class closer together than representations from different classes.

At the end of the pre-training phase, we discarded the projection layer, froze the encoder, and added a dense layer on top of it with $|C|$ outputs, where C is the number of all possible concepts. We trained the dense layer by minimizing the binary cross entropy loss. We once again used Adam [21] as our optimizer and decayed the learning rate by a factor of 10 when the loss showed no improvement, as in [8, 10]. We also used early stopping on the validation set, with patience of 3 epochs [8, 10]. A classification threshold for all the concepts was tuned by optimizing the F1 score. We used the same threshold for all the concepts. Any concepts for which the respective output values exceeded that threshold, were assigned to the corresponding image during inference.

Following [10], we trained 5 models using cross-validation and kept the 2 best performing ones, according to their F1 score. We then created an ensemble, using the UNION of the concepts returned by these two models. Hereafter, this ensemble will be called 2xCNN+FFNN@U. We also considered an ensemble with the INTERSECTION of the concepts that will be called 2xCNN+FFNN@I.

3.1.2. System 2: 1-NN ensemble

In this system, we followed a retrieval approach, extending the work of our previous systems in [7, 8, 9, 10]. We employed four different CNN encoders. One of the four encoders was the encoder used in System 1 (see Section 3.1.1), whereas the rest were a ResNet-50 [11], a DenseNet-201 [19], and an EfficientNet-B0 [22], all of which were pre-trained on ImageNet [20] and fine-tuned on our training set in a purely supervised setting.[10] For each encoder, we again trained 5 models using cross-validation, resulting in a number of 20 models in total. Having fine-tuned all the encoders, we used each one of them to obtain dense vector encodings, called *image embeddings*, of all the training images. The image embeddings are extracted from the last average pooling layer of the encoders. Following [8], we tuned the value of k in the range from 1 to 200 for each encoder separately, using the validation set, which led to $k = 1$. Therefore, given a test image, we retrieved the training image with the highest cosine similarity (computed on image embeddings) to the test image, resulting in a total of 20 images, one retrieved from each encoder. In order to make the submissions, we considered four different strategies for the concepts assignment. We submitted four systems using 20 encoders, one for each assignment strategy. We also submitted two additional 1-NN ensemble systems where we used a smaller

[9] Each image is rescaled to 224x224 and normalized with the mean and standard deviation of ImageNet.
[10] We did not use contrastive learning for the three extra encoders due to time restrictions.

number of models. In the first, we used four different encoders (i.e., the best out of the 5 models per encoder). In the second, we used three different encoders (discarding the encoder that yielded the lowest F1 score on the validation set). Our submitted systems will be discussed in Section 4.

3.1.3. System 3: Combination of 1-NN and 2xCNN+FFNN

This system is a combination of a retrieval and a classification method. We used the encoder that was trained in System 1 (see Section 3.1.1) in order to obtain the image embeddings for all training images as in System 2 (see Section 3.1.2). During inference, we retrieved the training image with the highest cosine similarity to the query image embedding and applied a similarity threshold s. If the similarity exceeded that threshold, the concepts of the closest neighbor were assigned to the query image. If not, we fed the query image to the classification component and assigned the predicted concepts to it. We tuned the value of s in the interval $[0.65, 1]$ with $step = 0.01$ which led to $s = 0.8$.

We also experimented with a variation where the 1-NN retrieval component was replaced by a 1-NN ensemble as in System 2 (see Section 3.1.2) using the version with the four different encoders. For each encoder, we retrieved one training image. If the similarity of each retrieved image to the query image exceeded the similarity threshold $s = 0.8$, the 1-NN ensemble was used for the assignment of the concepts. Otherwise, we fed the query image to the 2xCNN+FFNN@U ensemble and used its predictions for the concept assignment.

3.2. Caption Prediction

3.2.1. Text Generation Baselines

We trained each text generation baseline as a language model, on all the captions of the training set. Then, we simply used it to generate text, disregarding the image of the respective example. That is, these baselines generate a likely text per image, without actually considering the image. The training set for these models was preprocessed, since we observed that this led to better results. We used beam search, with the beam size for each step equal to 3. The baseline models we trained were GPT-2 [14] and GPT Neo.[11]

Table 5
The hyperparameters of the text generation baseline models. The batch size for GPT-2 and GPT Neo is dictated by their architectures.

Model	Huggingface name	Epochs	Batch Size	Block Size	Optimizer	Learning Rate
GPT-2	gpt2	15	12	52	Adam	3e-5
GPT Neo	EleutherAI/gpt-neo-125M	10	12	52	Adam	3e-5

The hyperparameters we used for each model, after light tuning, can be seen in Table 5. In all our models, it was better to merge the training captions. An example can be seen in Table 6,

[11] https://huggingface.co/transformers/master/model_doc/gpt_neo.html

where instead of feeding the model one caption at a time (and padding or truncating the input based on its length), we merged all the captions and then created equally-sized inputs.

Table 6
Two approaches (I, II) of creating the training input for the text generation baselines. Assuming a toy dataset with two captions and models with a maximum length of eight tokens.

	Caption 1: 'Left Upper lobe mass' **Caption 2:** 'Duplicated Right Renal System'
I	INPUT 1: [START], LEFT, UPPER, LOBE, MASS, [PAD], [PAD], [PAD] INPUT 2:[START], DUPLICATED, RIGHT, RENAL, PHRASE, [PAD], [PAD], [PAD]
II	INPUT 1: [START], LEFT, UPPER, LOBE, MASS, [START], DUPLICATED, RIGHT INPUT 2: RENAL, SYSTEM, [PAD], [PAD], [PAD], [PAD], [PAD], [PAD]

3.2.2. Image-Aware Text Generation

We experimented with two image captioning architectures that utilize the image. One was inspired by Show, Attend and Tell [15], and its architecture can be seen in Figure 4.[12] The second model was a token classifier, with image, and previous text inputs.[13] Models of the second architecture do not predict full sentences, but rather the next token for a given unfinished sentence and an image. In other words, a tuple consisting of an image and a caption of n words, will be used to create n different image-text inputs, with the text length varying from 1 to n (after adding a start token to the caption). The input images are passed through an encoder, then dropout is applied with 0.5 probability, and finally a dense layer yields the image representation. The text's words are transformed to word embeddings through a trainable embedding layer, then a dropout of 0.5 probability is applied, and the word embeddings are given to a Long Short-Term Memory (LSTM). The image representation and the text encoding (from the last output of the LSTM) are concatenated and then passed through a multilayer perceptron (MLP) with one hidden layer. In early stages of development, we saw that the model was predicting the next token correctly about 8% of the times (in the validation data), so later development stages only considered the Show, Attend and Tell inspired model.

3.2.3. Retrieval Methods

Following our previous work [7, 8, 10], we implemented a retrieval approach, based on k-NN. First, a pre-trained encoder generated an embedding for each training image. The images were reshaped beforehand according to the encoder used (Table 7). During inference, the same encoder generated an embedding for the test image, and the k training images with the most similar embeddings were retrieved (we used cosine similarity). The captions of the retrieved embedded images were then combined to yield a caption for the test image. We experimented

[12] https://colab.research.google.com/github/tensorflow/docs/blob/master/site/en/tutorials/text/image_captioning.ipynb

[13] Inspired by: https://towardsdatascience.com/image-captioning-with-keras-teaching-computers-to-describe-pictures-c88a46a311b8 (accessed on May 23rd, 2021)

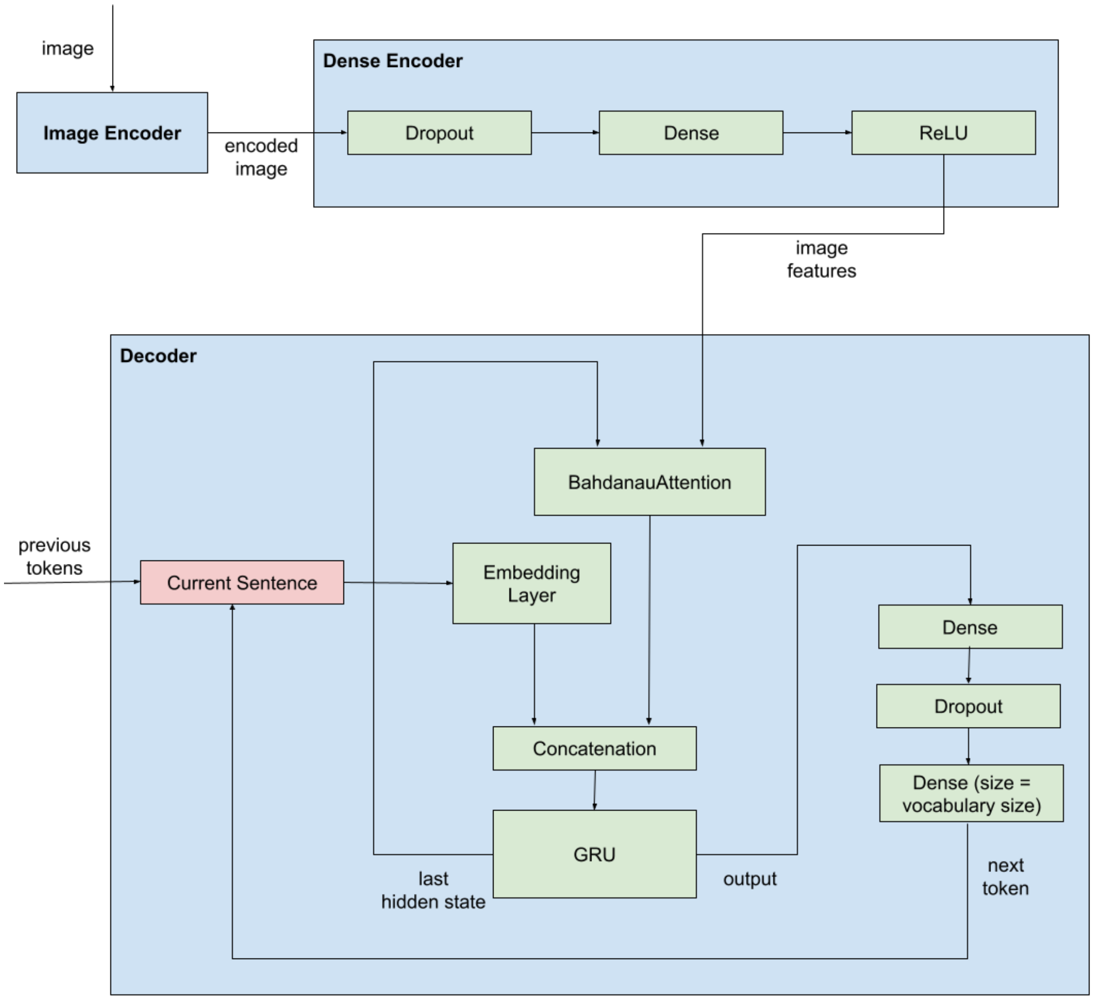

Figure 4: The architecture of the Show, Attend and Tell inspired model. The image encoder we used was InceptionV3 and the image attention mechanism was based on [23].

both with $k = 1$ and $k \neq 1$. For the latter, we split each retrieved caption into sentences and then concatenated the r most frequent sentences (the most frequent first) to form a caption. Alternatively, we summarized the k retrieved captions using an off-the-shelf summarizer.[14] 1-NN models performed best, but we plan to further investigate the effect of summarization in the future, and more specifically by using a model that utilizes Bidirectional and Auto-Regressive Transformers (BART), [24], inspired by its application in Retrieval-Augmented Generation (RAG) [25].

The 1-NN models were outperforming other competing methods, so we focused on experi-

[14]https://www.geeksforgeeks.org/python-text-summarizer/

Table 7
Image encoders and image input shapes in the image-aware text generation models.

Model	Input Shape
EfficientNetB0	224x224
EfficientNetB7	600x600
DenseNet121 DenseNet201	224x224
InceptionV3	299x299
ResNet50 ResNet152V2	224x224
NASNetLarge	331x331
InceptionResNetV2	299x299
Xception	299x299

menting with pre-trained encoders to better represent the images. We used encoders from the trained models of Section 3.1.1 (we named them Tag-Trained encoders), and ensembles of 1-NNs with different encoders. Since combining the predictions did not seem to work well, for the ensembles we just predicted the caption that most of the 1-NN methods of the ensemble predicted for each image. If every prediction was different for an image, i.e., if all the 1-NNs disagreed on it, we either used the prediction of the best 1-NN model (evaluated on validation data), or the prediction of the GPT-2 text generation model. GPT-2 was our next best model, after 1-NNs, but we observed that it always generated the same sentence (not one that appears in the dataset though).

4. Submissions and results

4.1. Concept Detection

We used our development set to evaluate all our models and submitted those that performed best. Six out of our ten submissions used the 1-NN ensemble (System 2). We considered four different concept assignment strategies and made four submissions, one for each strategy, using an ensemble of twenty encoders:

- **MAJORITY VOTING:** For each concept, we count how many of the 20 images(one retrieved per encoder) are associated with it. If the concept is present in the majority (i.e., more than 10), we assign it to the test image. We will refer to this system as 1-NN@20xMJ.
- **UNION OF INTERSECTIONS**: We used the INTERSECTION of the concepts of the 5 images that were retrieved from the 5 models that used the same backbone encoder. Then, we used the UNION of the above INTERSECTION. We will refer to this system as 1-NN@20xUoI.
- **INTERSECTION OF UNIONS**: Here, we reversed the calculation. We used the UNION of the concepts of the 5 images that were retrieved from the 5 models that used the same backbone encoder and then used the INTERSECTION of the UNION. This system will be referred to as 1-NN@20xIoU.

- **UNION OF UNIONS**: We used the UNION of the concepts of the 5 images that were retrieved from the 5 models that used the same backbone encoder and then used the UNION of the UNION. This system will be referred to as 1-NN@20xUoU.

The two additional 1-NN ensemble submissions, that used three and four encoders respectively, were made using the MAJORITY VOTING strategy. For the concept assignment, we used 2 and 3 votes as the majority for the three and four encoders respectively. We will refer to these systems as 1-NN@3xMJ and 1-NN@4xMJ.

Three submissions were based on the combination of a retrieval and a classification method (System 3). In the first submission, the classification component was the 2xCNN+FFNN@U ensemble, while in the second submisson we used the 2xCNN+FFNN@I as the classifier. Both of these methods used 1-NN as the retrieval component. The third submission, discussed in Section 3.1.3, was the variation that used a 1-NN ensemble as the retrieval component and the 2xCNN+FFNN@U as the classifier.

Additionally, one submission was made using an ensemble classification system (System 1). We only submitted the 2xCNN+FFNN@U system as it performed better than 2xCNN+FFNN@I in our development set.

The official measure of the competition was F1, calculated by comparing the binary multi-hot vectors y_{true} and y_{pred} of each test image and then macro-averaging over all test images. To generate the predictions for the test set, we merged the training with the development set. We used a held-out set (15% of the merged data) to tune the threshold of the 2xCNN+FFNN model. This resulted in the values $t_1 = 0.268$ and $t_2 = 0.304$.

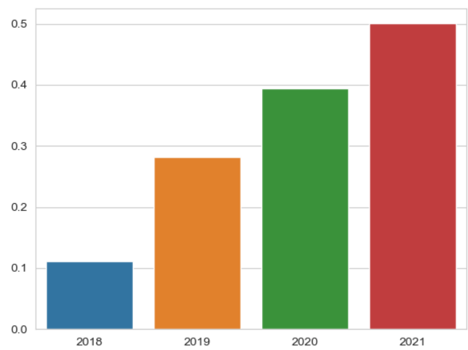

Figure 5: The top F1 score achieved each year in the Concept Detection sub-task of the ImageCLEFmed Caption tasks.

Table 8 presents the scores of our systems on the development and the official test set, as well their position according to the official ranking. 1-NN@20xMJ had the best results on the development and test set. As noted in Section 2, the number of the unique concepts present in the data is 1,585. Furthermore, the number of the total assigned concepts is 11,361. There is a large overlap between the assigned concepts, which means that many images have almost, or

Table 8
The results and rankings of our systems on the development and test set.

ID	Approach	F1 Score		Rank
		Development	Test	
cd1	1-NN@20xMJ	**61.99**	**50.5**	**1**
cd2	1-NN@20xUoI	55.73	45.6	9
cd3	1-NN@20xIoU	60.85	49.5	2
cd4	1-NN@20xUoU	43.33	34.8	23
cd5	1-NN@4xMJ	59.08	49.3	3
cd6	1-NN@3xMJ	60.76	49.0	5
cd7	2xCNN+FFNN@U	57.24	45.9	8
cd8	1-NN/2xCNN+FFNN@U	59.53	46.6	7
cd9	1-NN/2xCNN+FFNN@I	59.46	45.1	11
cd10	1-NN@4xMJ/2xCNN+FFNN@U	59.08	49.3	4
cd11	MOST FREQUENT BASELINE	27.92	–	–

completely, the same concepts. This fact can explain why our retrieval approaches worked very well in this task.

Furthermore, the reduced (in comparison with previous years) number of unique concepts and the slight improvement regarding the class imbalance, probably help the systems achieve better results. This is indicated by the increasing score of the winning systems (see Fig. 5) and also supported by the low performance of the MOST FREQUENT BASELINE on our development set. This baseline assigned to every test image the same three (as the average number of concepts is 3.48) most frequent concepts of the training data (see Table 8).

We also have to note that we only made use of the data that was provided by the organizers, without using external medical datasets to avail our models.

4.1.1. Caption Prediction

The official evaluation measure for caption prediction was BLEU-4, so we evaluated our models using that measure on our development set to decide which ones we should submit (each participant was allowed at most 10 submissions). For the 1-NN submissions, we combined the training, validation and development sets, since there were no hyperparameters to tune for those models. In Table 9, we can see the scores on the development set, and in Table 10 we can see the final scores of our submitted models in the unknown test captions. Some interesting points to be made are the following.

1. Encoders do not always benefit from an increase in their architecture complexity. We see that some of the 1-NN encoders actually scored higher in our development set when they had fewer trainable parameters.
2. The encoders that were further trained for the Concept Detection task (where our team took the first place) did not seem to have any significant lead over the others (our best model still remained an Ensemble without the Tag-Trained encoders).

Table 9
The scores of all of our Caption Prediction systems on our development set.

ID	**Approach**	**BLEU-4 Score**
cp1	GPT-2 (117M parameters)	34.923
cp2	GPT Neo (125M parameters)	25.540
cp3	Show, Attend and Tell inspired	20.471
cp4	DenseNet121 1-NN	51.405
cp5	DenseNet201 1-NN	52.755
cp6	ResNet50 1-NN	52.256
cp7	ResNet152V2 1-NN	42.120
cp8	InceptionV3 1-NN	49.342
cp9	InceptionResNetV2 1-NN	49.250
cp10	Xception 1-NN	48.963
cp11	NASNetLarge 1-NN	45.728
cp12	EfficientNetB0 1-NN	51.747
cp13	EfficientNetB7 1-NN	51.099
cp14	Tag-Trained ResNet50 1-NN	50.988
cp15	Tag-Trained DenseNet201 1-NN	53.381
cp16	Tag-Trained EfficientNetB0 1-NN	52.641
cp17	Ensemble of cp5, cp8 and cp10	53.634
cp18	Ensemble of cp4, cp5, cp8, cp9 and cp10	54.153
cp19	Ensemble of cp4, cp5, cp8, cp9 and cp10 GPT-2 on non-Agreement	**55.342**
cp20	Ensemble of cp14, cp15 and cp16 GPT-2 on non-Agreement	54.877
cp21	Ensemble of cp5, cp6, cp14, cp15 and cp16 GPT-2 on non-Agreement	55.023
cp22	cp23 with 2 most frequent sentences instead of most frequent caption	51.161

Table 10
The final scores of our 6 submissions, along with their rank on the test set based on submissions from all participants.

ID	**BLEU-4 Score**		**Rank**
	Development	**Test**	
cp19	**55.342**	**46.1**	**3**
cp21	55.023	45.2	4
cp22	52.161	44.8	5
cp17	53.634	44	7
cp4	51.405	37.5	18
cp3	20.471	19.9	38

3. Surprisingly, the models that were *not* aware of the image were better than the image-aware models (an image-aware model was even dropped, as mentioned in Section 3.2.2, due

to very low scores). Could this mean that the task and data made it difficult to extract information from the images, and thus the images became noise, or were the image-aware models we tested not good enough? Due to time constraints, we couldn't investigate this issue further, but we intend to do so in future work.

5. Conclusion and future work

This article described the submissions of AUEB's NLP Group to the 2021 ImageCLEFmed Caption sub-tasks. In the Concept Detection sub-task, our top system, ranked 1st amongst all submissions of 5 teams. It was a 1-NN retrieval-based ensemble using majority voting and several different image encoders. Our submissions also included an ensemble of classifiers trained using supervised contrastive learning [12], as well as combinations of the retrieval and classification modules that were implemented. In the Caption Prediction sub-task, by mainly focusing on retrieval methods, we managed to take the 2nd place amongst the 8 competing teams. Although we tried different approaches, including both image-aware and image-unaware generation models, nothing was able to beat our retrieval models, showing how powerful they can be, as they also helped us win the Concept Detection sub-task of the same campaign in previous years. Finally, we also observed that encoders do not always benefit from larger architectures with more trainable parameters.

In future work, we aim to assess our models on additional medical datasets and experiment more with retrieval-based methods that have proved promising. Our future plans also include a further analysis of image-aware captioning models, with the addition of image-aware pretrained Transformer models like GPT-2 and BERT, which we are currently working on. We also want to improve our retrieval models for captioning by studying methods of summarizing or combining text. Through these model experiments, and the analysis of more datasets, our research on diagnostic captioning [6] will also continue.

References

[1] B. Ionescu, H. Müller, R. Peteri, A. Ben Abacha, D. Demner-Fushman, S. Hasan, M. Sarrouti, O. Pelka, C. Friedrich, A. Herrera, J. Jacutprakart, V. Kovalev, S. Kozlovski, V. Liauchuk, Y. Dicente Cid, J. Chamberlain, A. Clark, A. Campello, H. Moustahfid, A. Popescu, The 2021 ImageCLEF Benchmark: Multimedia Retrieval in Medical, Nature, Internet and Social Media Applications, Lecture Notes in Computer Science (2021).

[2] O. Pelka, A. B. Abacha, A. G. S. de Herrera, J. Jacutprakart, C. M. Friedrich, H. Müller, Overview of the ImageCLEFmed 2021 Concept and Caption Prediction Task, in: CLEF2020 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Bucharest - Romania, 2021.

[3] O. Bodenreider, The Unified Medical Language System (UMLS): integrating biomedical terminology, Nucleic acids research 32 (2004) 4.

[4] C. Eickhoff, I. Schwall, A. G. S. de Herrera, H. Müller, Overview of ImageCLEFcaption 2017 – Image Caption Prediction and Concept Detection for Biomedical Images, in: CLEF2017 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Dublin, Ireland, 2017.

[5] A. G. S. de Herrera, C. Eickhoff, V. Andrearczyk, H. Müller, Overview of the ImageCLEF 2018 caption prediction tasks, in: CLEF2018 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Avignon, France, 2018.

[6] J. Pavlopoulos, V. Kougia, I. Androutsopoulos, D. Papamichail, Diagnostic captioning: A survey, CoRR abs/2101.07299 (2021).

[7] V. Kougia, J. Pavlopoulos, I. Androutsopoulos, A Survey on Biomedical Image Captioning, in: Workshop on Shortcomings in Vision and Language of the Annual Conference of the North American Chapter of the Association for Computational Linguistics, Minneapolis, MN, USA, 2019, pp. 26–36.

[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.

[9] 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.

[10] B. Karatzas, V. Kougia, J. Pavlopoulos, I. Androutsopoulos, AUEB NLP Group at ImageCLEFmed Caption 2020, in: CLEF2020 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece, 2020.

[11] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition (2016).

[12] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A.Maschinot, C. Liu, D. Krishnan, Supervised Contrastive Learning, Advances in Neural Information Processing Systems 33 (2020) 18661–18673.

[13] J. Devlin, M. 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), Minneapolis, Minnesota, 2019, pp. 4171–4186.

[14] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners, volume 1(8), 2019.

[15] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, Y. Bengio, Show, attend and tell: Neural image caption generation with visual attention, in: Proceedings of the 32nd International Conference on Machine Learning, Lille, France, 2015, pp. 2048–2057.

[16] A. B. Abacha, V. V. Datla, S. A. Hasan, D. Demner-Fushman, H. Müller, Overview of the VQA-Med Task at ImageCLEF 2020: Visual Question Answering and Generation in the Medical Domain, in: CLEF (Working Notes), 2020.

[17] 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.

[18] Y. Zhang, X. Wang, Z. Guo, J. Li, ImageSem at ImageCLEF 2018 Caption Task: Image Retrieval and Transfer Learning, in: CLEF2018 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Avignon, France, 2018.

[19] G. Huang, Z. Liu, L. V. D. Maaten, K. Q. Weinberger, Densely connected convolutional networks (2017) 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] D. P. Kingma, J. Ba, Adam: A Method for Stochastic Optimization, in: 3rd International Conference on Learning Representations (ICLR), San Diego, CA, USA, 2014.
[22] M. Tan, Q. V. Le, Efficientnet: Rethinking model scaling for convolutional neural networks (2019) 6105–6114.
[23] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, arXiv preprint arXiv:1409.0473 (2014).
[24] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension (2020) 7871–7880.
[25] P. S. H. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-augmented generation for knowledge-intensive NLP tasks, CoRR abs/2005.11401 (2020).