Detecting Concepts and Generating Captions from Medical Images: Contributions of the VCMI Team to ImageCLEFmedical Caption 2023

Notebook for the ImageCLEFmedical Caption Lab at CLEF 2023

Isabel Rio-Torto[1,3,*], Cristiano Patrício[1,4], Helena Montenegro[1,2], Tiago Gonçalves[1,2] and Jaime S. Cardoso[1,2]

[1] *INESC TEC, Campus da FEUP Rua Dr. Roberto Frias s/n, 4200-465 Porto, Portugal*

[2] *Faculdade de Engenharia, Universidade do Porto, Rua Dr. Roberto Frias s/n, 4200-465 Porto, Portugal*

[3] *Departamento de Ciência de Computadores, Faculdade de Ciências, Universidade do Porto, Rua do Campo Alegre s/n, 4169–007 Porto, Portugal*

[4] *Departamento de Informática, Universidade da Beira Interior, Rua Marquês de Ávila e Bolama, 6201-001 Covilhã, Portugal*

Abstract

This paper presents the main contributions of the VCMI Team to the ImageCLEFmedical Caption 2023 task. We addressed both the concept detection and caption prediction tasks. Regarding concept detection, our team employed different approaches to assign concepts to medical images: multi-label classification, adversarial training, autoregressive modelling, image retrieval, and concept retrieval. We also developed three model ensembles merging the results of some of the proposed methods. Our best submission obtained an F1-score of 0.4998, ranking 3rd among nine teams. Regarding the caption prediction task, our team explored two main approaches based on image retrieval and language generation. The language generation approaches, based on a vision model as the encoder and a language model as the decoder, yielded the best results, allowing us to rank 5th among thirteen teams, with a BERTScore of 0.6147.

Keywords

Concept Retrieval, Image Captioning, Medical Concept Detection, Multi-label Classification, Natural Language Generation, Vision Transformers

CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece

[*] Corresponding author.

isabel.riotorto@inesctec.pt (I. Rio-Torto); cristiano.p.patricio@inesctec.pt (C. Patrício); maria.h.sampaio@inesctec.pt (H. Montenegro); tiago.f.goncalves@inesctec.pt (T. Gonçalves); jaime.cardoso@fe.up.pt (J. S. Cardoso)

https://github.com/icrto (I. Rio-Torto); https://cristianopatricio.github.io (C. Patrício); https://github.com/helenaMontenegro (H. Montenegro); https://tiagofilipesousagoncalves.github.io (T. Gonçalves); https://web.fe.up.pt/~jsc/ (J. S. Cardoso)

0000-0002-2302-8597 (I. Rio-Torto); 0000-0003-2215-3334 (C. Patrício); 0000-0001-6237-3011 (H. Montenegro); 0000-0003-4744-9174 (T. Gonçalves); 0000-0002-3760-2473 (J. S. Cardoso)

CEUR Workshop Proceedings (CEUR-WS.org)

1. Introduction

ImageCLEF 2023 [1] is a multi-modal challenge organised as part of the CLEF Initiative Labs[1] (Conference and Labs of the Evaluation Forum, formerly known as Cross-Language Evaluation Forum) set to promote the evaluation of technologies for annotation, indexing, classification and retrieval of multi-modal data. The 2023 edition included four challenges from diverse applications (i.e. medical, social media and Internet, and content recommendation).

Similarly to last year [2], our team, composed of members of the Visual Computing and Machine Intelligence (VCMI) Research Group of the Institute for Systems and Computer Engineering, Technology and Science (INESC TEC) from Porto, Portugal, participated in the ImageCLEFmedical Caption 2023 task [3] where the goal is to challenge the scientific community to design and train automatic algorithms capable of interpreting and summarising the insights gained from medical images. Once again, this challenge consisted of two independent, but complementary, tasks: *concept detection*, which aims to identify the presence of relevant concepts in a large corpus of medical images; and *caption prediction*, which aims to generate coherent textual descriptions describing a medical image. We addressed both the concept detection and caption prediction tasks.

For the concept detection task, we developed five different approaches: (i) baseline multi-label classification, in which a convolutional neural network (CNN) simultaneously predicts all the concepts from an image; (ii) adversarial approach, in which a multi-label classifier and a concept discriminator are trained in an adversarial manner to promote the learning of admissible concept combinations by the multi-label classifier; (iii) autoregressive approach, that aims to model dependencies between concepts using autoregressive learning; (iv) image retrieval, in which a model assigns concepts to an image based on its most similar images from the training data; and (v) concept retrieval, in which a model learns to map concepts and images into a common latent space where images are closer to the concepts they contain. We also developed three model ensembles using the aforementioned approaches: (i) multi-label classification and concept retrieval, (ii) autoregressive model and image retrieval using autoregressive model, (iii) adversarial model and image retrieval using autoregressive model. Our best submission (i.e.ensemble with autoregressive model and image retrieval using autoregressive model) obtained an F1-score of 0.4998, ranking 3rd among nine teams.

For the caption prediction task, we relied on Vision Encoder-Decoder Transformer-based architectures, since they worked well on last year's competition [2]. We explored two different categories of image feature extractors for the Encoder, namely a Vision Transformer and a CNN. Furthermore, we introduced a caption-to-concepts classification branch as an additional supervisory signal for the model, since the caption needs to contain enough information to allow, to some extent, for the prediction of the concepts. Our best submission (i.e. the Vision Transformer encoder model trained on both training and validation sets) achieved a BERTScore of 0.6147, ranking 5th among thirteen participating teams.

The remainder of this paper is organised as follows: Section 2 provides an overview of the data provided by the organisation to address the tasks and describes our exploratory data analysis; Section 3 details the different proposals developed to solve the aforementioned tasks;

[1] http://www.clef-initiative.eu (accessed on: 02-06-2023)

Section 4 presents the results and their discussion; and Section 5 concludes this paper and recommends future work directions. The code related to this paper is publicly available in a GitHub repository[2].

2. Data

The dataset provided in this competition is an extended version of the Radiology Objects in COntext (ROCO) dataset [4]. The data originates from biomedical articles of the PMC OpenAccess subset [3]. The images provided to the participants are divided into training (60,918 images), validation (10,437 images) and test (10,473 images) sets.

The concepts provided in the training and validation data were annotated according to the Unified Medical Language System (UMLS) [5] 2022 AB release, wherein each concept is uniquely identified through a Concept Unique Identifier (CUI). For additional details, please refer to [3].

Table 1 presents an analysis of the number of concepts contained in each training and validation image. On average, each image has 3.7 concepts, and while there are 4716 images with only 1 concept, there are also 233 images with more than 10 concepts, the maximum number of concepts per image being 24 in the training set.

Table 1
Data analysis for the concept detection task, from the image-based perspective. "Avg.", "Min." and "Max." stand for "Average", "Minimum" and "Maximum" number of concepts per image, respectively. The last two columns refer to the number of images with only 1 concept and with over 10 concepts.

Subset	Total Images	Avg.	Min.	Max.	With 1 concept	With over 10 concepts
Training	60918	3.7	1	24	4716	233
Validation	10437	3.8	1	33	771	65

Table 2 presents an analysis of the concepts and their frequency in the training and validation data. While the training set contains a total of 2125 concepts, only 1945 of these appear in the validation data. Furthermore, each concept appears on average on 106.9 images in the training set and on 20.6 images in the validation set. While there are concepts that appear in 1 or 2 images only, there are other concepts that appear in 20955 images, making this task highly imbalanced. Moreover, the validation data contains a significant amount of concepts that appear in 10 or less images. Since Deep Learning models require large amounts of training data to achieve good performance, the existence of these less-frequent concepts raises the difficulty of the concept detection task.

Regarding the data for the caption prediction task, which was obtained from the aforementioned biomedical articles, we present our analysis in Table 3. While the smallest captions contain only 2 tokens, the biggest ones can have up to 995 tokens. The average caption length is 33.6 and 36.6 for the training and validation sets, respectively. Looking at the 98% percentile, we conclude that only 2% of the dataset has captions with more than 100 tokens. This observation led us to choose a maximum length of 100 for all our developed methods described in the next section.

[2] https://github.com/TiagoFilipeSousaGoncalves/ImageCLEFmedical2023VCMI

Table 2
Data analysis for the concept detection task, from the concept-based perspective. "Avg.", "Min." and "Max." stand for "Average", "Minimum" and "Maximum" number of images per concept, respectively. The last two columns refer to the number of concepts that appear in 10 or less images and that appear in over 100 images, respectively.

Subset	Total Concepts	Avg.	Min.	Max.	In 10 or less images	In over 100 images
Training	2125	106.9	2	20955	397	272
Validation	1945	20.6	1	3740	1460	53

Table 3
Statistics of the training and validation data for the caption prediction task. "Avg.", "Min." and "Max." stand for "Average", "Minimum" and "Maximum" caption length (i.e. number of tokens), while the last three columns correspond to the 95%, 98%, and 99% percentiles, respectively.

Subset	Avg.	Min.	Max.	95% perc.	98% perc.	99% perc.
Training	33.6	2	715	78	101	122
Validation	36.6	2	995	87	115	142.6

3. Methodology

The following sections describe the methods developed to fulfill the concept detection and caption prediction tasks.

3.1. Concept Detection

The concept detection task was solved using two main approaches: modelling the concept detection task as a multi-label classification problem and as an information retrieval problem. We developed three models based on multi-label classification: a baseline model, a model trained in an adversarial manner and a model trained using autoregressive learning. Furthermore, we developed models to perform concept retrieval and image retrieval. The following subsections describe, in detail, each of the proposed methods.

3.1.1. Baseline Multi-Label Approach

A conventional approach to address the concept detection task involves employing a multi-label classification model, considering the inherent nature of images to encompass multiple non-mutually exclusive concepts.

Specifically, we adapted the DenseNet-121 [6] architecture by modifying the classification layer to have N outputs, where N is the number of concepts, i.e. 2125.

In the training phase, the model was trained using the binary cross-entropy loss function and the adaptive moment estimation (Adam) optimiser [7] with its default hyperparameters. The model was trained during 100 epochs with a learning rate of 1e-4. Concretely, we trained the classification layer of the model while keeping the remaining layers frozen. Subsequently, the model with the best validation loss was selected for the testing phase.

3.1.2. Adversarial Approach

Ensuring that the multi-label baseline approach learns the correct combination of concepts is not trivial (e.g. concepts related to different body parts should not be combined). Hence, we propose adversarial training to learn a realistic combination of concepts per image, according to the distribution of the training data. This model is composed of two blocks (see Figure 1):

- A multi-label classifier trained to predict the *top-K* most frequent concepts ($K = 100$) in the database. This block uses ResNet50 [8] as a feature extractor along with a multi-layer perceptron (MLP) with a sigmoid activation.
- A concept discriminator trained to distinguish between *real* (i.e. admissible) and *fake* (i.e. inadmissible) combinations of concepts. This block is an MLP with two fully-connected layers followed by a ReLU activation and a fully-connected layer with a sigmoid activation.

We trained this model for 20 epochs using binary cross-entropy as the loss function for both the multi-label classifier and the concept discriminator, and Adam [7] as the optimiser. The best model is saved according to the lowest validation loss.

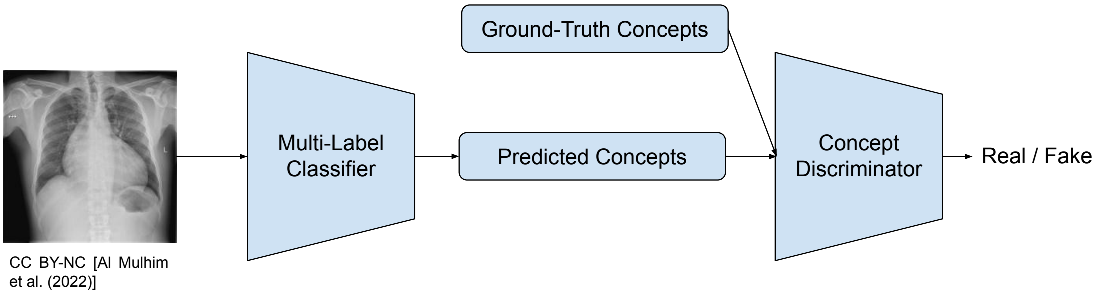

Figure 1: Overview of the multi-label adversarial model.

3.1.3. Autoregressive Approach

The main limitation of the baseline multi-label classification approach is that it assumes that the concepts are independent of each other. However, there may be dependencies between concepts, as there are concepts that never appear together in the training data, or concepts that only exist in the presence of other concepts. To overcome this limitation, we devised an approach to model dependencies between concepts based on autoregressive learning.

The proposed model is a multi-label classification network that, instead of having a final classification layer with 2125 units to predict all the concepts, contains several classification layers, each predicting a subset of concepts. To model dependencies, each layer is conditioned on the output of the previous layers. An overview of the autoregressive model is depicted in Figure 2. As the feature extractor of the network, we used a VGG16 [9] network pre-trained on ImageNet [10], followed by two fully-connected layers with LeakyReLU activations and Dropout. All of the classification layers are fully-connected layers with sigmoid activation.

Since it is easier for a network to predict concepts that exist in more images, we organised the concepts in the layers according to how frequent they are among the training images. The

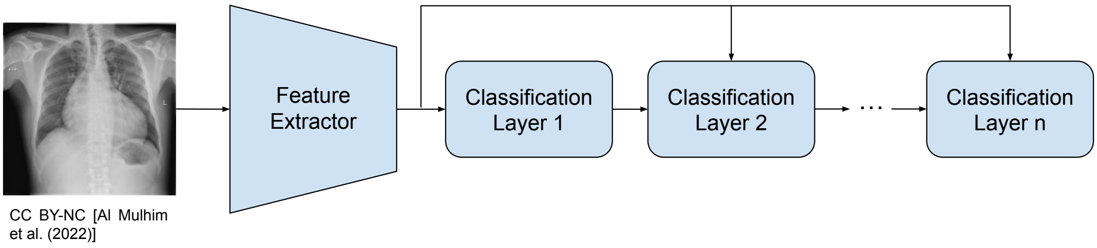

Figure 2: Overview of the multi-label autoregressive model.

most common concepts are predicted by the first while the rarest ones are predicted by the last classification layers. Since there is a total of 2125 concepts, we used 17 classification layers, each responsible for predicting 125 concepts.

In the training phase, the model was trained using binary cross-entropy as the loss function and the Adam optimiser with a learning rate of 1e-5. We trained the model in two phases. First, we trained the classification layers of the model for 50 epochs, with the feature extractor frozen. Then, we fine-tuned the entire network by training it for 20 epochs. We selected the best instance of the model by monitoring its loss on the validation data.

3.1.4. Retrieval Approaches

We implemented two main approaches to predict the concepts of an image based on information retrieval techniques: concept retrieval and image retrieval. In concept retrieval, the method maps images and concepts into a common latent space, retrieving the closest concepts to an image. In image retrieval, the method assigns concepts to an image based on its most similar images from the training data. Both these methods will be described in detail below.

In the concept retrieval approach, we use an image encoder and a concept encoder to map images and concepts into a common latent space. Then, we compute the Euclidean distance between the latent representations of the images and concepts, as depicted in Figure 3. During training, we minimise the Euclidean distance between an image and the concepts it contains, and we maximise the distance between the image and the concepts it does not contain.

In our implementation, the image encoder is a CNN with four blocks of convolutional layers with Batch Normalisation and Max Pooling, followed by a layer that performs Global Average Pooling and a fully-connected layer. The concept encoder is a Multi-Layer Perceptron (MLP) with one fully-connected layer with Dropout and LeakyReLU as the activation function, followed by a second fully-connected layer.

In addition to the Image-to-Concept (ITC) loss, we also performed some experiments where we added the following loss functions to the training of the networks:

- **Concept-to-Concept (CTC) loss**: Minimises the distance between two different concepts that exist in the same images, and maximises the distance between concepts that do not appear together in any image. We apply a weight to the loss function by multiplying it by the percentage of images that two concepts share (intersection over union).

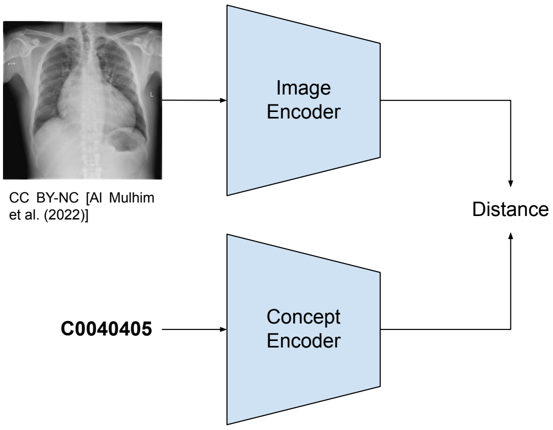

Figure 3: Overview of the concept retrieval model.

- **Image-to-Image (ITI) loss**: Minimises the distance between images that have some concepts in common, and maximises the distance between images that do not share any concepts. We apply a weight to the loss function by multiplying it by the percentage of concepts that two images have in common (intersection over union).

We performed three experiments: (i) training the concept retrieval networks only with the ITC loss for 2600 epochs, (ii) fine-tuning the network trained with the ITC loss using the CTC loss for 100 epochs, and (iii) fine-tuning the network trained with the ITC loss simultaneously using the CTC and ITI losses for 100 epochs. The networks were trained using the Adam optimiser with a learning rate of 1e-5. We monitored the loss on the validation data to obtain the best model.

In the image retrieval approach, we use pre-trained models to obtain latent representations of the images, which are then used to measure the distance between the target image whose concepts we want to predict and the images of the training data. We devised three strategies to assign concepts to the target image, based on its most similar images:

- **Strategy 1 (S1)**: Retrieve the closest image and assign its concepts to the target image.
- **Strategy 2 (S2)**: Retrieve the Top-N closest images and assign to the target image the concepts of the closest image that also exist in at least one other image from the Top-N retrieved images. If no concept of the closest image appears in another image of the Top-N, then all the concepts of the closest image are assigned to the target image.
- **Strategy 3 (S3)**: Retrieve the Top-N closest images and assign the concepts that exist in at least two of the Top-N retrieved images to the target image. Similarly to Strategy 2, if no concept appears in at least two images of the Top-N, then all the concepts of the closest image are assigned to the target image.

We empirically chose to retrieve the Top-4 closest images in strategies 2 and 3.

As the pre-trained models to obtain a latent representation of the images we used a ResNet50 [8] trained on ImageNet [10], and the image encoders of the previously described concept retrieval and autoregressive models.

3.1.5. Ensemble

The multi-label classification-based approaches (baseline, adversarial and autoregressive) often fail to predict any concepts for a given test image, leading to many images in the test dataset with no predicted concepts. As such, we devise an ensemble strategy where, for each image where the multi-label approaches fail to predict any concepts, we assign the concepts predicted by one of the retrieval approaches.

3.2. Caption Prediction

The caption prediction task involves generating text that describes an image. To tackle this task we considered two categories of approaches, retrieval and language generation, which we describe in more detail below.

3.2.1. Retrieval Approach

We applied the image retrieval approach developed for the concept detection task to obtain captions for the test images. We used the pre-trained ResNet [8], the concept retrieval network trained using the ITC loss and the autoregressive network to obtain latent representations of the images. These representations were then used to obtain the closest images from the training and validation data whose captions were assigned to the test samples.

3.2.2. Language Generation Approaches

The language generation-based strategies used to tackle this task employ an Encoder-Decoder framework, since it was our best performing approach in last year's competition [2]. The Encoder, typically a CNN or a Vision Transformer, is responsible for analysing the image and extracting relevant features. The Decoder then receives the encoded image features and generates the caption. Thus, it is usually an autoregressive model, such as GPT-2 [11].

We experimented with two different encoders: the small distilled version of the Data-efficient image Transformer (DeiT) [12] from the Huggingface Transformers library [13], and, inspired by the work of Hou et al. [14], DenseNet121 [6] from TorchXRayVision [15, 16] pre-trained on all available datasets (`densenet121-res224-all`). The decoder consisted of the distilled version of GPT-2 [11]. Both models were trained with an initial learning rate of 1e-4 using the AdamW optimiser [17] for 25 epochs. We monitored the BERTScore on the validation data to obtain the best model.

Since the UMLS concepts of the concept detection task are tightly related to the captions in the caption prediction task, we hypothesise that it should be possible to predict the concepts from the captions to some extent. Furthermore, predicting the concepts from the captions might prove a good additional supervisory signal for training the caption prediction model. Therefore, we explored the inclusion of a text classifier that takes the caption of a given image and predicts its concepts (see Figure 4).

To accomplish this we originally trained a DistilBERT [18] model for caption-to-concept multi-label classification. The model was trained with the binary cross-entropy loss on the CLS token for 20 epochs with an initial learning rate of 2e-5 and the AdamW optimiser. This caption-to-concept classifier was then used (but kept frozen) on top of the DenseNet-DistilGPT2 model to provide an extra loss function for training. However, since the output of the Encoder-Decoder module and the input of the caption-to-concept classifier (i.e. the generated text) is discrete, Reinforcement Learning (RL) is needed, similarly to what is done in Self-Critical Sequence Training [19]; thus, the whole sentence needs to be generated before classification can occur, making this approach much slower compared to teacher forcing-only training.

We also experimented with simply adding a fully connected layer directly on top of the latent representation of the Decoder's last token and training the whole Encoder-Decoder plus classification layer together. This approach has the advantage of not needing RL, thus making it faster and easier to train.

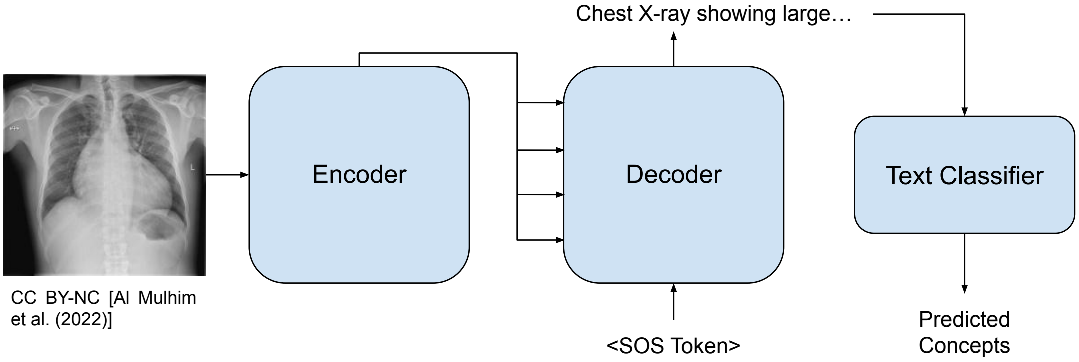
Chest X-ray showing large...
Encoder
Decoder
Text Classifier
CC BY-NC [Al Mulhim et al. (2022)]
<SOS Token>
Predicted Concepts

Figure 4: Overview of the captioning model.

4. Results and Discussion

This section details the results obtained by the methods developed for the concept detection and caption prediction tasks.

4.1. Concept Detection

The concept detection task is evaluated using the example-based F1-score between the predicted and ground-truth concepts. Table 4 presents the results in terms of F1-Score obtained by each proposed method on the validation and test data. Furthermore, it presents a secondary F1-score metric (F1-Score Manual) that compares the concepts predicted on the test data with a subset of manually validated concepts.

The baseline multi-label classification approach obtained an F1-score of 0.4469 on the test set. Contrary to our expectations, the adversarial approach did not improve upon the baseline. This might be explained by the fact that this adversarial model was only trained on the top-100

Table 4
Results of the concept detection task in terms of F1-score and Secondary F1-score computed on a subset of manually validated concepts. The models identified with * were trained on the training and validation data before being applied to the test data. The results on the validation dataset were obtained using models trained only on the training set.

Run	Model	F1-score (Validation)	F1-score (Test)	F1-score Manual (Test)
1	Baseline Multi-label*	0.4364	0.4469	0.8305
4	Adversarial* (Top-100)	0.2816	0.2803	0.5999
5	Autoregressive*	0.4905	0.4928	0.9062
2	Concept Retrieval (ITC)	0.4523	0.4360	0.7582
-	Concept Retrieval (ITC + CTC)	0.4404	-	-
-	Concept Retrieval (ITC + CTC + ITI)	0.2446	-	-
3	Image Retrieval - S3 (ResNet)	0.4693	0.4676	0.8305
7	Image Retrieval - S3 (Autoregressive)	0.4793	0.4793	0.9014
10	Image Retrieval - S3 (Concept Retrieval)	0.4379	0.4387	0.8394
6	Ensemble (Runs 1 and 2)	-	0.4728	0.8738
8	Ensemble (Runs 5 and 7)	-	**0.4998**	**0.9162**
9	Ensemble (Runs 4 and 7)	-	0.3327	0.7049
-	Task Winners	-	0.5223	0.9258

concepts. Thus, we leave as future work a more in-depth exploration of this approach. The autoregressive approach achieved the highest performance among the multi-label-based models.

In the concept retrieval approach, we verify that adding the CTC and ITI loss functions to the network trained only with the ITC loss leads to a lower F1-score.

Regarding the image retrieval method, we empirically found that Strategy 3 (S3) produced the best results. This ablation study can be found in Table 5, that compares the different image retrieval strategies on the validation data, using the concept retrieval model trained with ITC loss as the base. We verify that assigning concepts that exist in at least two of the Top-4 most similar images (Strategy 3) leads to the highest F1-Score. Among the three different base models used (ResNet, autoregressive and concept retrieval), the best results were obtained by using the autoregressive model. Nevertheless, these results do not surpass the values obtained by the multi-label classification-based autoregressive model.

However, the retrieval-based approaches proved very useful as complements to the classification-based methods. As expected, the ensemble methods, which combine both techniques, improved the results of all three multi-label classification networks (baseline, adversarial and autoregressive). We obtained the best results by merging our two best models from each category, the autoregressive multi-label classification network and the image retrieval approach using the autoregressive model, achieving an F1-Score of 0.4998 and a Manual F1-Score of 0.9162. This allowed us to rank 3rd in the competition among nine teams.

Table 5
Comparison (on the validation set) between the different strategies of image retrieval using the concept retrieval model with the Image-To-Concept loss.

Strategy	Retrieved Images	F1-score
S1	1	0.3314
S2	3	0.4030
S2	4	0.4232
S2	5	0.4234
S2	10	0.4176
S3	3	0.4354
S3	4	**0.4379**
S3	5	0.4345
S3	10	0.3887

4.2. Caption Prediction

The caption prediction task is evaluated in terms of BERTScore [20] and ROUGE [21]. We present the obtained results in Table 6 for both retrieval and language generation-based approaches.

All retrieval approaches ranked below the language generation-based approaches, which confirms that simply using the captions from similar images is not enough to accurately describe a different image.

Regarding the generation-based approaches, using the DeiT encoder yielded slightly improved results when compared to using DenseNet-121. As expected, adding the classification loss improved the corresponding base architecture, but it was not enough to surpass the DeiT + DistilGPT2 model. This suggests that, had time permitted, adding the classification loss to the DeiT instead of the DenseNet-based model would have further improved our results. We would like to point out that we do not report the results obtained by our model with the RL concept-to-caption classifier because we were not able to train it in a reasonable amount of time given the computational resources available.

Thus, our best results were obtained by the DeiT + DistilGPT2 model trained on both training and validation sets. This also suggests that our other developed methods could have better results if trained on both sets, something we leave as future work. In the end, these results awarded us the 5th place in the competition among thirteen participating teams.

5. Conclusions and Future Work

This work described the methods developed by the VCMI team in the ImageCLEFmedical Caption 2023 task. We developed approaches based on multi-label classification and retrieval to assign concepts to medical images, obtaining an F1-Score of 0.4998 that granted us 3rd place among the nine teams that participated in the challenge. For caption generation, we focused on encoder-decoder approaches with Transformers, obtaining a 5th place among thirteen teams, with a BERTScore of 0.6147.

Table 6
Results of the caption prediction task on the validation and test sets in terms of BERTScore and ROUGE. The models identified with * were trained on the training and validation data before being applied to the test data. The results on the validation dataset were obtained using models trained only on the training set.

Run	Model	BERTScore (Validation)	ROUGE (Validation)	BERTScore (Test)	ROUGE (Test)
1	Image Retrieval (ResNet)	0.5738	0.1417	0.5734	0.1427
2	Image Retrieval (Concept Retrieval)	0.5653	0.1268	0.5647	0.1284
8	Image Retrieval (Autoregressive)	0.5756	0.1464	0.5750	0.1464
3	DeiT + DistilGPT2	0.6133	0.2167	0.6138	**0.2181**
5	DeiT + DistilGPT2*	0.6133	0.2167	**0.6147**	0.2175
4	DenseNet-121 + DistilGPT2	0.6108	0.1935	0.6096	0.1938
6	DenseNet-121 + DistilGPT2 + Clf loss	0.6113	0.1947	0.6103	0.1948
-	Task Winners	-	-	0.6425	0.2446

In the concept detection task, the experiments show that training an autoregressive multi-label classification network to model dependencies between concepts is a promising approach capable of achieving high performance. As such, future work includes the further development of autoregressive models, potentially with the integration of more advanced autoregressive networks from the literature, such as Transformers [22]. We also intend to continue developing the concept retrieval approach by pre-training the concept encoder using the concept-to-concept loss before training the whole model. Finally, we consider the application of the adversarial approach to predict all concepts, rather than only the Top-100 most frequent concepts, and the potential integration between the adversarial and the autoregressive approaches into one model.

In the caption prediction task, future work involves exploring different and more powerful image encoders, as well as more recent language models. We also intend to explore more in-depth the inclusion of the concept classification loss into our base encoder-decoder approach, not only by applying it to all our model configurations, but also by investigating the best way of integrating it during training, e.g. only after the captioning module is sufficiently trained.

Acknowledgments

We would like to thank our colleague Pedro Neto for his valuable feedback and suggestions.

This work was partially funded by the Project TAMI - Transparent Artificial Medical Intelligence (NORTE-01-0247-FEDER-045905) financed by ERDF - European Regional Fund through the North Portugal Regional Operational Program - NORTE 2020 and by the Portuguese Foundation for Science and Technology - FCT under the CMU - Portugal International Partnership, and also by the Portuguese Foundation for Science and Technology (FCT) within PhD grants 2022.14516.BD, 2022.11566.BD, 2020.06434.BD and 2020.07034.BD.

References

[1] B. Ionescu, H. Müller, A. Drăgulinescu, W. Yim, A. Ben Abacha, N. Snider, G. Adams, M. Yetisgen, J. Rückert, A. García Seco de Herrera, C. M. Friedrich, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, S. A. Hicks, M. A. Riegler, V. Thambawita, A. Storås, P. Halvorsen, N. Papachrysos, J. Schöler, D. Jha, A. Andrei, A. Radzhabov, I. Coman, V. Kovalev, A. Stan, G. Ioannidis, H. Manguinhas, L. Ştefan, M. G. Constantin, M. Dogariu, J. Deshayes, A. Popescu, Overview of ImageCLEF 2023: Multimedia retrieval in medical, socialmedia and recommender systems applications, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the 14th International Conference of the CLEF Association (CLEF 2023), Springer Lecture Notes in Computer Science LNCS, Thessaloniki, Greece, 2023.

[2] I. Rio-Torto, C. Patrício, H. Montenegro, T. Gonçalves, Detecting Concepts and Generating Captions from Medical Images: Contributions of the VCMI Team to ImageCLEFmedical 2022 Caption, in: Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS.org, Bologna, Italy, 2022, pp. 1535–1553.

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

[4] O. Pelka, S. Koitka, J. Rückert, F. Nensa, C. M. Friedrich, Radiology Objects in COntext (ROCO): A Multimodal Image Dataset, in: Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis, Springer International Publishing, Cham, 2018, pp. 180–189.

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

[6] G. Huang, Z. Liu, L. V. D. Maaten, K. Q. Weinberger, Densely connected convolutional networks, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society, Los Alamitos, CA, USA, 2017, pp. 2261–2269. doi:10.1109/CVPR.2017.243.

[7] D. P. Kingma, J. Ba, Adam: A Method for Stochastic Optimization, in: 3rd International Conference on Learning Representations (ICLR), San Diego, CA, USA, 2015.

[8] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Las Vegas, NV, USA, 2016, pp. 770–778. doi:10.1109/CVPR.2016.90.

[9] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: 3rd International Conference on Learning Representations (ICLR), San Diego, CA, USA, 2015.

[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Miami, FL, USA, 2009, pp. 248–255. doi:10.1109/CVPR.2009.5206848.

[11] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language Models are

Unsupervised Multitask Learners, OpenAI Blog 1 (2019) 9.

[12] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, H. Jegou, Training data-efficient image transformers & distillation through attention, in: Proceedings of the 38th International Conference on Machine Learning (ICML), volume 139 of *Proceedings of Machine Learning Research*, PMLR, Online, 2021, pp. 10347–10357. URL: https://proceedings.mlr.press/v139/touvron21a.html.

[13] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, A. Rush, Transformers: State-of-the-Art Natural Language Processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP): System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 38–45. doi:10.18653/v1/2020.emnlp-demos.6.

[14] B. Hou, G. Kaissis, R. M. Summers, B. Kainz, RATCHET: Medical Transformer for Chest X-ray Diagnosis and Reporting, in: Medical Image Computing and Computer Assisted Intervention (MICCAI), Springer International Publishing, Cham, 2021, pp. 293–303. doi:https://doi.org/10.1007/978-3-030-87234-2_28.

[15] J. P. Cohen, M. Hashir, R. Brooks, H. Bertrand, On the limits of cross-domain generalization in automated X-ray prediction, in: Proceedings of the Third Conference on Medical Imaging with Deep Learning (MIDL), volume 121 of *Proceedings of Machine Learning Research*, PMLR, Online, 2020, pp. 136–155. URL: https://proceedings.mlr.press/v121/cohen20a.html.

[16] J. P. Cohen, J. D. Viviano, P. Bertin, P. Morrison, P. Torabian, M. Guarrera, M. P. Lungren, A. Chaudhari, R. Brooks, M. Hashir, H. Bertrand, TorchXRayVision: A library of chest X-ray datasets and models, in: Proceedings of The 5th International Conference on Medical Imaging with Deep Learning (MIDL), volume 172 of *Proceedings of Machine Learning Research*, PMLR, Online, 2022, pp. 231–249. URL: https://proceedings.mlr.press/v172/cohen22a.html.

[17] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, in: 7th International Conference on Learning Representations (ICLR), New Orleans, LA, USA, 2019.

[18] V. Sanh, L. Debut, J. Chaumond, T. Wolf, DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, in: 5th Edition of EMC2: Energy Efficient Machine Learning and Cognitive Computing Workshop at Neural Information Processing Systems (NeurIPS), 2019.

[19] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, V. Goel, Self-Critical Sequence Training for Image Captioning, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society, Los Alamitos, CA, USA, 2017, pp. 1179–1195. doi:10.1109/CVPR.2017.131.

[20] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, BERTScore: Evaluating Text Generation with BERT, in: International Conference on Learning Representations (ICLR), Addis Ababa, Ethiopia, 2020.

[21] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text Summarization Branches Out, Association for Computational Linguistics (ACL), Barcelona, Spain, 2004, pp. 74–81. URL: https://aclanthology.org/W04-1013.

[22] J. Lanchantin, T. Wang, V. Ordonez, Y. Qi, General Multi-label Image Classification with Transformers, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern

Recognition (CVPR), IEEE Computer Society, Los Alamitos, CA, USA, 2021, pp. 16473–16483. doi:10.1109/CVPR46437.2021.01621.