Detecting Concepts and Generating Captions from Medical Images: Contributions of the VCMI Team to ImageCLEFmedical 2022 Caption

Isabel Rio-Torto[1,2], Cristiano Patrício[2,3], Helena Montenegro[2,4] and Tiago Gonçalves[2,4]

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

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

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

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

Abstract

This paper presents the main contributions of the VCMI Team to the ImageCLEFmedical 2022 Caption task. We addressed both the concept detection and caption prediction tasks. Regarding concept detection, our team employed three different strategies: multi-label classification, in which a convolutional neural network aims to simultaneously predict all the concepts from an image considering only the 100 most frequent concepts; 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; and semantic-based multi-label classification, which consists of training several models, each one specialised in predicting concepts from a given semantic type, and an aggregation operation to obtain the final prediction. Our best submission attained an F1-score of 0.433, placing 5th among 11 teams, and the best Secondary F1-score (0.863). Regarding the caption prediction task, our team designed two different approaches: a Vision Encoder-Decoder Transformer, that receives the input images as a sequence of 16×16 patches and is trained for next token prediction; and a modified Object-Semantics Aligned Pre-training for Vision-and-Language Tasks (OSCAR) model, i.e. an encoder-only Transformer, trained for masked language modelling, and modified to receive as input a sequence of image patches and the image concepts, besides the caption. Our best submission, the Vision Encoder-Decoder, attained a Bilingual Evaluation Understudy (BLEU) score of 0.306 and ranked 4th among 10 teams.

Keywords

concept retrieval, contrastive learning, image captioning, medical concept detection, multi-label classification, natural language generation, vision transformers

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

✉ 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)
🌐 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)
iD 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)

CEUR Workshop Proceedings (CEUR-WS.org)

1. Introduction

ImageCLEF 2022 [1] is an evaluation campaign organised as part of the CLEF Initiative[1] (Conference and Labs of the Evaluation Forum, formerly known as Cross-Language Evaluation Forum). The 2022 edition included four tasks related to different application domains (i.e., Internet, Medical, Nature, and Social Media). These challenges encompass the common objective of promoting the evaluation of technologies for annotation, indexing and retrieval of visual data, while contributing to the access to extensive collections of images in various usage scenarios and application domains.

Our team, composed by four 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 2022 Caption task [2] wherein the main motivation is to develop algorithms that can interpret and summarise insights gained from medical images. 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 of a medical image.

We addressed both the concept detection and caption prediction tasks. For the concept detection task, we developed three different approaches: multi-label classification, in which a convolutional neural network (CNN) aims to simultaneously predict all the concepts from an image; 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; and semantic-based multi-label classification, which consists of training several models, each one specialised in predicting concepts from a given semantic type, and an aggregation operation to obtain the final prediction. Our best submission attained an F1-score of 0.433, ranking 5th among 11 teams (the best team achieved an F1-score of 0.451), and the best Secondary F1-score (0.863).

For the caption prediction task, we designed two different approaches: a Vision Encoder-Decoder Transformer, that receives the input images as sequences of 16x16 patches and is trained for next token prediction; and a modified Object-Semantics Aligned Pre-training for Vision-and-Language Tasks (OSCAR) model [3], i.e. an encoder-only Transformer, trained for masked language modelling, and modified to receive as input a sequence of image patches and the image concepts, besides the caption. Our best submission, the Vision Encoder-Decoder, attained a Bilingual Evaluation Understudy (BLEU) score of 0.306 and ranked 4th among 10 teams (the best team achieved a BLEU score of 0.483).

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

[1] http://www.clef-initiative.eu (accessed on: 26-05-2022)

[2] https://github.com/icrto/ImageClefMedical

2. Data

2.1. Overview

The data set provided in this competition is a subset of the extended Radiology Objects in COntext (ROCO) data set [4]. As in previous editions, the data set originates from biomedical articles of the PMC OpenAccess subset [2]. The version provided to the participants is already divided into train (83,275 radiology images), validation (7,645 radiology images) and test (7,601 radiology images) sets.

2.2. Exploratory Data Analysis

We considered it important to do an exploratory data analysis step before delving into the development of technical strategies to tackle both the *Concept Detection* and *Caption Prediction* tasks.

2.2.1. Concept Detection Task

For the concept detection task, we analysed the data from two different perspectives: *image-based*, where we computed the average, minimum and maximum number of concepts per image, and the m most (Top-m) and the l least (Bottom-l) frequent concepts (see Table 1); and *concept-based*, where we computed the total number of concepts, the average, minimum and maximum number of images where each concept is present, the concepts that appear in n images or less, and the number of concepts that do not appear in any image (see Table 2).

Table 1
Exploratory data analysis for the concept detection task, from the image-based perspective. Note: "Avg.", "Min." and "Max." stand for "Average", "Minimum" and "Maximum" number of concepts per image, respectively. Top-3 corresponds to the 3 most frequent concepts, while Bottom-3 corresponds to the 3 least frequent concepts.

Subset	Total	Avg.	Min.	Max.	Top-3	Bottom-3
Training	83,275	4.7	1	50	C0040405, C1306645, C0024485	C0004760, C0398658, C0030847
Validation	7,645	4.7	1	27	C0040405, C1306645, C0041618	C0447028, C0272388, C1561643

Table 2
Exploratory data analysis for the concept detection task, from the concept-based perspective. Note: "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 to the number of concepts that do not appear in any image, respectively.

Subset	Total	Avg.	Min.	Max.	In 10 or less images	In 0 images
Training	8,374	47.2	2	25,989	4,923	0
Validation	4,357	4.3	0	2,896	7,842	4,017

Table 1 shows that every image has at least one concept and that there is an average of around 5 concepts per image. Furthermore, in the Top-3 column, we observe that two of the most predominant concepts in the training data are also the most common in the validation data. Figures 6 and 7 in the Appendix also show that 21 of the 31 concepts exposed as the most predominant in the training data are also the most common in the validation data. An important outcome from this analysis is that the most common concepts are present in both training and validation sets. Intuitively, and assuming that the distribution of concepts is similar on the test set, this observation allows us to use the Top-m (most frequent) concepts for the concept prediction task without losing too much information. Additionally, 83,110 (99.80%) of the training images and 7,617 (99.63%) of the validation images contain at least one of the Top-100 most frequent concepts, which further supports the hypothesis that removing the least-frequent concepts leads to a small loss of information, which might have little impact on the results of a model designed to predict these concepts.

Table 2 shows that 58.76% of the concepts available in the data appears only in 10 (0.012%) of the training images or less. These results suggest that the concept prediction task, interpreted as a multi-label classification task where multiple labels can be assigned to the same image, is highly imbalanced in the training data. Furthermore, we observe that, out of the 8,374 existing concepts, 7,842 (93.65%) are reflected in less than 10 (0.13%) of the validation images, of which 4,017 (47.97% out of all concepts) are not reflected in the validation data at all. These observations suggest not only that the validation data is imbalanced, but also that the validation set has a very limited capacity to verify the quality and the generalisation power of any model regarding the detection of those concepts. Please refer to Figures 8 and 9 for additional details.

2.2.2. Caption Prediction Task

For the caption prediction task, we analysed the length of the captions, i.e. the number of tokens obtained after tokenization[3], thus extracting the average, minimum, and maximum caption lengths (see Table 3). The minimum length is 3 in both training and validation sets, while the maximum length corresponds to 577 in the training set and 339 in the validation set. We can observe that the vast majority of the images have description lengths close to the average number of tokens. In fact, 89.4% of the training images have less than 50 tokens and 99,1% have less than 100 tokens. The same tendency can be verified in the validation set, where 85.8% of the images have less than 50 tokens and 98.4% have less than 100 tokens. This analysis is further complemented by Figures 10 and 11.

Table 3
Exploratory data analysis for the caption prediction task. We present the average, minimum and maximum number of tokens for the captions on the training and validation subsets.

Subset	Average	Minimum	Maximum
Training	29.73	3	577
Validation	32.37	3	339

[3] We used the distil-gpt2 tokenizer of the Hugging Face Transformers library (https://huggingface.co/docs/transformers/index - accessed on: 27-05-2022).

3. Methodology

This section describes all the strategies we used to tackle both the *Concept Detection* and *Caption Prediction* tasks, as well as the data processing steps applied.

3.1. Data Processing

All images were first resized to have 224 pixels of height, while keeping their aspect ratios, and were converted to greyscale. During training of the several approaches, random square crops of 224×224 were used. In the multi-label-based concept detection proposals the images were also rotated by at most 45° with 50% probability.

Following the conclusions derived in the exploratory data analysis phase, all captions were tokenized using the distil-gpt2 tokenizer from the Hugging Face Transformers library and truncated or padded to 100 tokens in the Vision Encoder-Decoder approach and to 50 tokens in the modified OSCAR architecture.

3.2. Concept Detection Task

The concept detection task consists of a multi-label classification problem, i.e., there are more than two classes and each data point may be labelled with more than one non-mutually exclusive class. To solve this task, we employed three different strategies: *multi-label classification*, *concept retrieval*, and *semantic-based multi-label classification*.

3.2.1. Multi-label Classification

A straightforward method to solve the task of concept detection is to use a multi-label classification model, since a single image can have multiple non-mutually exclusive concepts associated with it. The concepts are defined according to the Unified Medical Language System (UMLS) [5] 2020AB release, in which each concept has a unique identifier (CUI). Table 4 presents the Top-3 most frequent concepts in the training set and their frequency in both training and validation sets.

Table 4
Top-3 most frequent concepts (CUIs) in the training set alongside their correspondent UMLS term, and their respective frequency in the training and validation sets.

CUI	UMLS Term	Train	Validation
C0040405	X-Ray CT	25989	2896
C1306645	Plain X-Ray	24389	2023
C0024485	MRI	14622	1071

Based on the exploratory data analysis, we adopted two strategies for predicting the concepts: (i) train the model to predict all the 8,374 concepts, and (ii) train the model to predict only the m most frequent concepts, where $m = 100$. For this, a multi-label classification-based model was developed to predict the associated concepts for each image. Specifically, we adapted the

DenseNet-121 [6] by modifying the classification layer to have N outputs, where N is the number of concepts, and $N = 8,374$ or $N = 100$. Figure 1 illustrates the architecture diagram of the employed multi-label classification model.

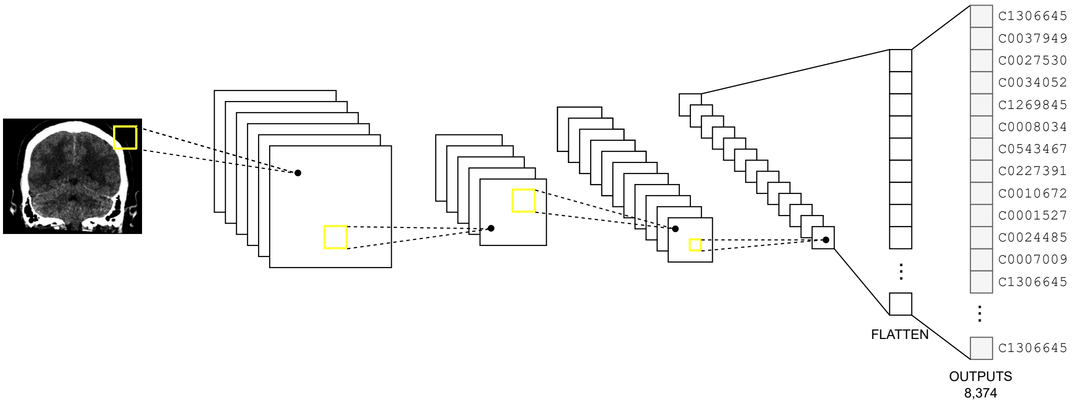

Figure 1: Diagram of the multi-label classification model. The input image is fed to the model, a DenseNet-121 [6], and the final output layer where the sigmoid activation function is employed predicts N values corresponding to the N concepts. A concept is assigned to the image if its predicted score is greater than a predefined decision threshold. Example image: CC BY [7].

Regarding the training process, the model was trained using the binary cross-entropy loss and the adaptive moment estimation (Adam) [8] optimiser with its default hyperparameters during 100 epochs with a learning rate of 10^{-3}. Concretely, we adopted three strategies for training the model: (i) we fine-tuned the classification layer of the model and kept the remaining layers frozen (“Frozen Backbone”), (ii) we trained the whole model with all layers unfrozen (“Whole Network”), and (iii) we froze the backbone layers for 5 epochs and then unfroze them for the remaining epochs (“2 Phases”). The model with the best validation loss was used for the testing phase.

3.2.2. Concept Retrieval

As a second approach, we attempted to solve the concept detection task through concept retrieval: images and concepts are mapped into a common latent space where the images are expected to be closer to the concepts they contain. Using this model, we retrieve the closest concepts to the images. Figure 2 presents an overview of the model.

The model is composed of an image encoder and a concept encoder, each responsible for translating images and concepts into their latent representations. The input concepts are represented as one-hot encodings. The image encoder is a CNN composed of four blocks of convolutional layers with max pooling and batch normalisation. The concept encoder is a multilayer perceptron composed of two fully-connected layers with LeakyReLU and Tanh activations, respectively.

During training, the model uses a contrastive loss [9] to minimise the distance between images and their corresponding concepts while maximising the distance between images and concepts they do not contain. The contrastive loss function used to train the model is expressed

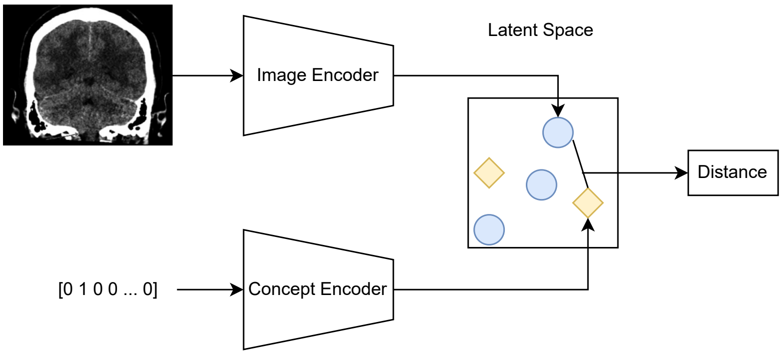

Figure 2: Diagram of the concept retrieval model trained using contrastive learning. The images and concepts are encoded into a common latent space. In the represented latent space, blue circles correspond to latent representations of images, while yellow diamonds correspond to latent representations of concepts. Example image: CC BY [7].

in Equation 1, where D is the distance between the input concept and image, and y is a binary value that represents the existence of the concept in the image. On separate experiments, we use two different distance measures: Euclidean distance and Cosine Similarity.

$$\mathcal{L}_{contrastive} = y \times D^2 + (1 - y)[\max(0, 1 - D)]^2 \tag{1}$$

During training, on each batch, the network receives as input a set of k existing concepts and k images. For each batch, the network only needs to process each concept and image once. As a result, for each training batch, the network predicts a square distance matrix where each line corresponds to an image, each column to a single one-hot encoded concept, and each cell represents the distance between the respective image and concept. This approach is more computationally efficient than providing independent image-concept pairs to the network, as, in that approach, the same concept and image would have to be processed multiple times in the same iteration.

During inference, we retrieve the concepts that are closer to the image and whose distance to the image is smaller than a threshold. This threshold corresponds to the average distance between an image and the closest concept on the validation data.

Using this methodology, we performed three separate experiments: (i) train the concept retrieval model using Euclidean distance and representing all $8,374$ concepts in the network's latent space; (ii) train the model using Euclidean distance and mapping only the 100 most frequent concepts to the latent space; and (iii) train the model using Cosine Similarity and considering only the 100 most frequent concepts. We trained the models using the Adam optimiser [8], with a learning rate of 10^{-6}. During the training process, we used a subset of the training data (15%) for validation, to obtain the epoch at which each model obtained the best results. As such, the models for the three experimental settings were trained for 1550, 703, and 279 epochs, respectively.

We also perform experiments where we merge the results of the concept retrieval and multi-label classification models, in an attempt to improve the respective results. In the first experiment, the images are labelled according to the multi-label classification model. In cases where the images are not assigned any label by the multi-label classification model, we retrieve its closer concepts according to the concept retrieval model (we call this the "Ensemble (NaN)" model). In the second experiment, we merge the predictions of the two models using an OR operation, assigning to the image all the concepts that were predicted by either the multi-label classification model or the concept retrieval model (we refer to this as the "Ensemble (OR)" model).

3.2.3. Semantic-based Multi-label Classification

To address this task from a "divide-and-conquer" perspective, we started with the conversion of each concept into its correspondent semantic type. We used the UMLS [5] Terminology Services REST API[4] to map each concept into a high-level semantic type. Afterwards, we computed the frequency of each high-level semantic type on the Top-100 concepts, thus getting to the following types: *Body Part, Organ, or Organ Component*; *Spatial Concept*; *Finding*; *Pathologic Function*; *Qualitative Concept*; *Diagnostic Procedure*; *Body Location or Region*; *Functional Concept* and *Miscellaneous Concepts* (i.e., the remaining semantic types which have lower frequencies). For each of these types, we trained a ResNet18 [10] on the images and their multi-labels (if present). To predict the final set of concepts per image, we run each of the previous models for the entire data set and perform an aggregation operation (i.e., the union). Regarding the training process, these models were trained during 10 epochs, using the Adam optimiser with a learning rate of 10^{-4}. These models also used the "2 Phases" strategy, where we froze the backbone layers for 5 epochs and then unfroze them for the remaining epochs. The best model is saved based on the lowest validation loss. Figure 3 illustrates this framework.

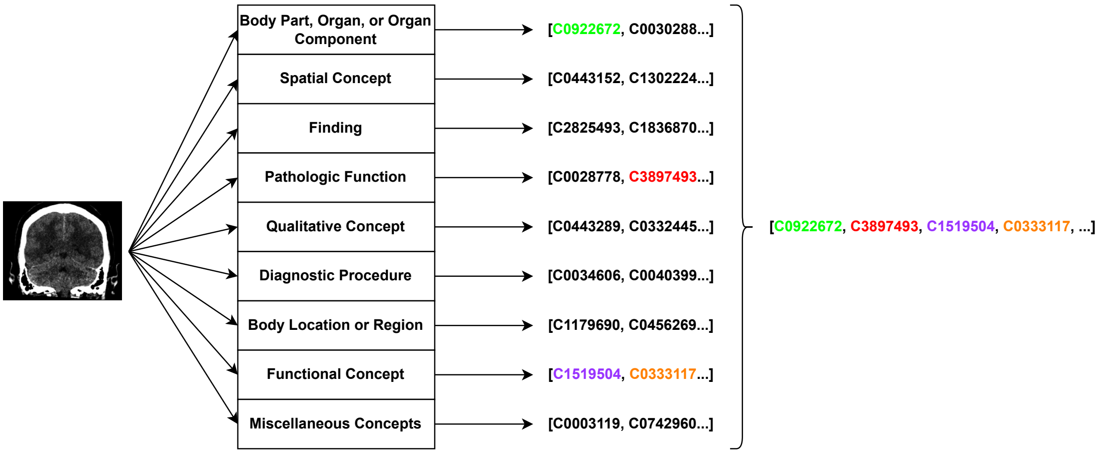

Figure 3: Diagram of the semantic-based multi-label classification model. Each model is trained on a set of concepts of the same semantic type. During inference, the input image is given to all the models and an aggregation operation (i.e., union) is performed. An example of the aggregation of the concepts detected by different modules of the network is shown by the different coloured concepts. Example image: CC BY [7].

[4] https://www.nlm.nih.gov/research/umls/index.html (accessed on 26-05-2022)

3.3. Caption Prediction Task

The caption prediction task consists of a natural language generation problem, more specifically, an image captioning problem where a textual description of the images must be generated. To tackle this problem we developed two strategies, both based on the Transformer architecture [11]: a *Vision Encoder-Decoder* Transformer and a *modified OSCAR* Transformer [3].

3.3.1. Vision Encoder-Decoder

The Vision Encoder-Decoder architecture combines the original Transformer [11] with the Vision Transformer (ViT) [12], i.e. while it keeps the original Transformer's encoder-decoder structure, the encoder receives as input an image divided into patches of 16×16 pixels. The decoder receives the ground-truth caption as input (i.e., training is done using teacher forcing) and the encoder hidden states as inputs to the cross attention layers. The model is trained autoregressively for next token prediction using causal (or unidirectional) self-attention, which means that a given token can only attend to previous tokens. Figure 4 depicts this architecture.

The model was implemented using the Vision Encoder Decoder class from the Hugging Face Transformers library and we chose a tiny Data-efficient image Transformer (DeiT) [13] pretrained on ImageNet for the encoder. For the decoder we leverage pretrained weights from the Distilled-GPT2, a distilled version of the GPT-2 architecture [14]. We trained this model initially for 20 epochs, and then for an additional 20 epochs starting from the checkpoint with the lowest validation loss. We used the AdamW [15] optimiser with an initial learning rate of 5×10^{-5}, linearly decayed. Due to limitations of the computational resources available we were not able to fine tune the model using self-critical sequence training [16].

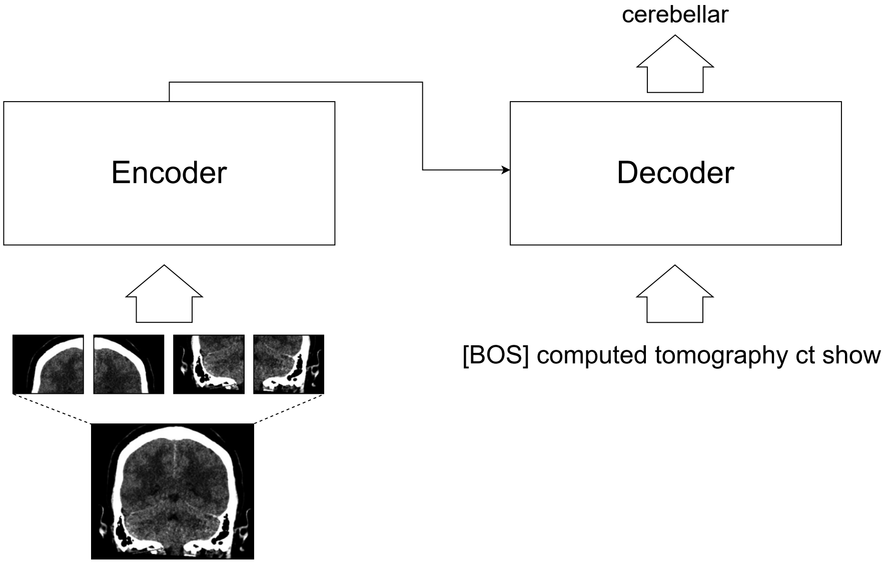

Figure 4: Diagram of the Vision Encoder-Decoder captioning model. The encoder receives the input image divided into patches of 16×16 pixels, while the decoder receives the ground-truth caption and the encoder hidden states, and predicts the next word in the sentence. Example image: CC BY [7].

3.3.2. Modified OSCAR

We hypothesised that leveraging the information present in the concepts might aid in generating the captions. Therefore, we developed a modified OSCAR [3] architecture, as depicted in Figure 5. Since the original architecture uses object tags and region features obtained from an object detector (e.g. Faster R-CNN) and we do not have access to bounding box annotations for our data, we modified OSCAR to receive as input the image divided into 16×16 patches similarly to what is done in the ViT model. Furthermore, instead of object detection tags the model receives the ground-truth concepts for each image.

Contrary to the Vision Encoder-Decoder model, OSCAR is trained for masked language modelling, so the objective is to predict the masked input tokens. We adopted the same masking strategy as in the original OSCAR [3]. However, instead of using bidirectional attention like in the original Bidirectional Encoder Representations from Transformers (BERT) encoder model [17], OSCAR (as well as our modified version) adopts causal self-attention, since our target goal is text generation. Therefore, when predicting the masked token, it can attend to every concept and image patch, but it can only attend to previous tokens from the caption.

During inference on the test set, we obtained the concepts from our best concept detection model, i.e. the "Ensemble (NaN)" model. We start by passing a MASK token as the textual input and the model then generates the rest of the sequence autoregressively.

We leveraged weights from an OSCAR model pretrained on the MSCOCO Captions data set [18]. The model was fine tuned on the competition data set for 20 epochs, with the AdamW [15] optimiser and a learning rate of 10^{-4}, linearly decayed. Maximum caption length was defined as 50, while the maximum length for the sequence of concepts was 10. As before, we planned on fine tuning the model using self-critical sequence training [16], but that was not possible due to limitations in computational resources.

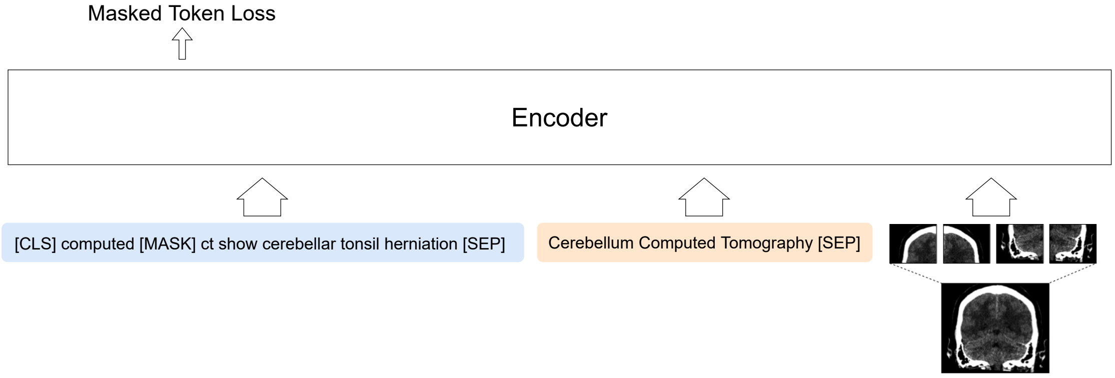
Masked Token Loss
Encoder
[CLS] computed [MASK] ct show cerebellar tonsil herniation [SEP]
Cerebellum Computed Tomography [SEP]

Figure 5: Diagram of the modified OSCAR captioning model, trained for masked language modelling using causal self-attention layers. Instead of receiving object regions and tags from an object detector, this modified version takes the image concepts and 16×16 patches as input alongside the caption. Example image: CC BY [7].

4. Results and Discussion

This section presents and discusses the results of the conducted experiments in the *Concept Detection* and *Caption Prediction* tasks.

4.1. Concept Detection Task

The evaluation of the concept detection task is conducted in terms of the example-based F1-score between the predicted and ground truth concepts. Additionally, a variant of the F1-score (Secondary F1-score) was computed using only a subset of manually validated concepts related to anatomy and image modality. Table 5 presents the results obtained in the experiments relative to this task. In this table, the first two columns refer to the model used and to the concepts that were considered during the respective model's training. The table also includes the results in terms of F1-score in the validation and test sets and Secondary F1-score in the test set.

Table 5
Results of the concept detection task in terms of F1-score and Secondary F1-score computed on a subset of manually validated concepts. "Top-100" and "All" refer to the (sub)set of concepts used to train the models.

Model	Concepts	F1-score (Validation)	F1-score (Test)	Secondary F1-score (Test)
Multi-label (Frozen Backbone)	All	0.3710	-	-
Multi-label (Frozen Backbone)	Top-100	0.3740	-	
Multi-label (Whole Network)	Top-100	0.3947	0.430	0.861
Multi-label (2 Phases)	Top-100	0.3937	0.431	0.856
Euclidean Retrieval	All	0.3367	-	-
Euclidean Retrieval	Top-100	**0.3973**	0.368	0.778
Cosine Similarity Retrieval	Top-100	0.3184	-	-
Ensemble (NaN)	Top-100	0.3959	**0.433**	**0.863**
Ensemble (OR)	Top-100	0.3956	-	-
Semantic	Top-100	-	0.418	0.838
Task Winners	-	-	0.451	0.791

In the multi-label classification model, a concept is associated with the image if its predicted score is greater than the defined decision threshold value (0.5). As evidenced by the results presented in Table 5 concerning the multi-label classification model, the best performance is obtained when the model is trained to predict only the Top-100 concepts, achieving an F1-score of 0.3740 when using the Multi-label ("Frozen Backbone") approach, and 0.3947 when using the Multi-label ("Whole Network") model. On the other hand, when the model is trained to predict the $8,734$ concepts, performance slightly decreases (0.3740 to 0.3710). We can also conclude that training the whole network improves the results in terms of F1-score compared with the results obtained when freezing the weights of the feature extraction layers of the DenseNet-121 (0.3740 to 0.3947).

Alternatively, when adopting the “2 Phases” strategy, we verify a marginal decrease of the F1-score on the validation set (0.3937) compared to the best result (0.3947) obtained with the “Whole Network” strategy. However, the F1-score on the test set is marginally higher (0.431) than the result obtained with the “Whole Network” strategy (0.430).

Regarding the retrieval task using contrastive learning, we verify that the model that uses Euclidean distance in the contrastive loss yields better results than the model that uses Cosine Similarity. We also observe that training the model to recognise only the Top-100 concepts leads to higher F1-score than when all the existing concepts are used, which is consistent with the results obtained with the multi-label approach. Despite the Euclidean Retrieval model trained with the Top-100 concepts achieving higher F1-scores than the multi-label classification model on the validation set, we verify that its results are considerably worse in the test set. Nevertheless, when we merge the results of the multi-label classification and the Euclidean retrieval models in the ensemble model referred to as Ensemble (NaN), the results in the test set improve slightly. As such, we were able to improve the results of the multi-label classification model by using retrieval to detect concepts for images where the multi-label classification network failed to detect any concepts. Interestingly, the Ensemble with the OR operation is not able to surpass the previous Ensemble.

Regarding the semantic-based multi-label classification, we highlight the proximity of the scores obtained by this approach to the previous ones. However, it is important to mention that the results do not confirm our initial intuition that using prior knowledge (i.e., in this case, optimising different models for a specific semantic type of concepts) would improve the predictive performance.

Although our best submission ranked 5th (see Table 5) in terms of F1-score among all submissions from 11 teams, the difference between the winner’s F1-score and ours is relatively small (0.018). On the other hand, we note the results obtained on the Secondary F1-score, which is computed with a specific subset of manually validated concepts (anatomy and image modality). In this metric, our team achieved the best score (0.863) of the whole competition and by a considerable margin compared to the result of the task winners (0.791).

4.2. Caption Prediction Task

The caption prediction task is evaluated in terms of natural language generation metrics. The BLEU score was chosen as the primary competition metric, but ROUGE, METEOR, CIDEr, SPICE and BERTScore are also computed. Table 6 presents the results obtained in the competition test set.

Our best performing model was, surprisingly, the vanilla Vision Encoder-Decoder trained for 40 epochs, achieving a BLEU score of 0.306, and placing in 4th place out of 10 participating teams, while the modified OSCAR only achieved 0.230. This might be due to the fact that the modified OSCAR was trained on ground-truth concepts, but during inference the model used the predicted concepts from our best concept detection model. Perhaps it would have been beneficial to introduce some uncertainty during training by alternating between feeding ground-truth concepts and feeding predicted concepts from that same concept detection model, in an effort to teach the Transformer the bias introduced by the concept detection model.

Furthermore, had the time permitted it, the Vision Encoder-Decoder could have been trained for more epochs, since the training loss was still decreasing, which is corroborated by the fact that training an additional 20 epochs improved the BLEU score from 0.300 to 0.306.

Finally, it is interesting to note that although in terms of BLEU score there is a considerable difference between our best performing model (0.306) and the task winners (0.483), our model largely outperforms the winner in terms of CIDEr (by 0.175), a natural language generation metric that tries to better correlate with human judgement.

Table 6
Results of the caption prediction task on the test set in terms of BLEU, ROUGE, METEOR, CIDEr, SPICE, and BERTScore.

Model	BLEU	ROUGE	METEOR	CIDEr	SPICE	BERTScore
Vision Encoder-Decoder (20 epochs)	0.300	0.172	0.073	**0.210**	**0.039**	**0.604**
Vision Encoder-Decoder (40 epochs)	**0.306**	**0.174**	**0.075**	0.205	0.036	**0.604**
Modified OSCAR	0.230	0.111	0.047	0.088	0.023	0.551
Task Winners	0.483	0.142	0.0928	0.030	0.007	0.561

5. Conclusions and Future Work

This paper described the work developed by the VCMI team in the ImageCLEFmedical 2022 Caption task. Regarding the concept detection task, three different strategies were adopted: (i) a multi-label classification model, (ii) a retrieval-based approach, and (iii) a semantic-based multi-label classification model. The experimental results indicated that merging the multi-label classification model with the retrieval approach (ensemble model) was the best strategy for the concept detection task, achieving the highest F1-score (0.433) on the test set among all our submissions, and ranking 5th among all the 11 participating teams. In terms of Secondary F1-score, we achieved the best value (0.863) among all the participating teams. Concerning the caption prediction task, we explored two strategies based on Transformer architectures, a Vision Encoder-Decoder Transformer and a modified OSCAR. Our best submission, the Vision Encoder-Decoder, obtained a BLEU score of 0.306, thus achieving the 4th place among all the 10 participating teams.

Future work should be devoted to improving the developed methods for both the concept detection and caption prediction tasks. Regarding the first task, we believe we could improve our results by building the ensemble model with the ”2 Phases” multi-label classification model. In terms of improving the generation of captions, we would start by training the Vision Encoder-Decoder for more epochs, as the training loss was still decreasing. Furthermore, we also believe that improving the concept detection phase would boost the performance of the modified OSCAR approach and that including non ground-truth concepts during training would teach the Transformer the bias introduced by the concept detection model and, consequently, be more adapted to the inference scenario. Finally, we would also like to perform an ablation study and compare both Transformer approaches more directly, by training the modified OSCAR without the concepts.

Acknowledgments

This work, developed within the scope of the project “TAMI - Transparent Artificial Medical Intelligence” (NORTE-01-0247-FEDER-045905), is co-financed by ERDF - European Regional Fund through the Operational Program for Competitiveness and Internationalisation - COMPETE2020, 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 the Ph.D. grants “2020.06434.BD” and “2020.07034.BD”.

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, A. Popescu, 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] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei, et al., Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks, in: Proceedings of the European Conference on Computer Vision (ECCV), 2020, pp. 121–137.

[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, 2018, pp. 180–189.

[5] O. Bodenreider, The Unified Medical Language System (UMLS): integrating biomedical terminology, Nucleic Acids Research 32 (2004) D267–D270.

[6] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely Connected Convolutional Networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4700–4708.

[7] A. Curtis, C. Lamb, H. Rao, A. Williams, A. Patel, Dialysis Disequilibrium Syndrome and Cerebellar Herniation with Successful Reversal Using Mannitol, Case Reports in Nephrology 2020 (2020) 1–4. doi:10.1155/2020/8850850.

[8] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, in: Y. Bengio, Y. LeCun (Eds.), 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015.

[9] R. Hadsell, S. Chopra, Y. LeCun, Dimensionality Reduction by Learning an Invariant

Mapping, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, volume 2, 2006, pp. 1735–1742.

[10] 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 (CVPR), 2016, pp. 770–778.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention Is All You Need, Advances in Neural Information Processing Systems 30 (2017).

[12] 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: Proceedings of the International Conference on Learning Representations (ICLR), 2021.

[13] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, H. Jégou, Training data-efficient image transformers & distillation through attention, in: Proceedings of the International Conference on Machine Learning (ICML), 2021, pp. 10347–10357.

[14] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language Models are Unsupervised Multitask Learners, OpenAI Blog 1 (2019) 9.

[15] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, in: 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019, OpenReview.net, 2019.

[16] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, V. Goel, Self-critical Sequence Training for Image Captioning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 7008–7024.

[17] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019, pp. 4171–4186.

[18] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, C. L. Zitnick, Microsoft COCO Captions: Data Collection and Evaluation Server, arXiv preprint arXiv:1504.00325 (2015).

A. Exploratory data analysis

A.1. Concept detection

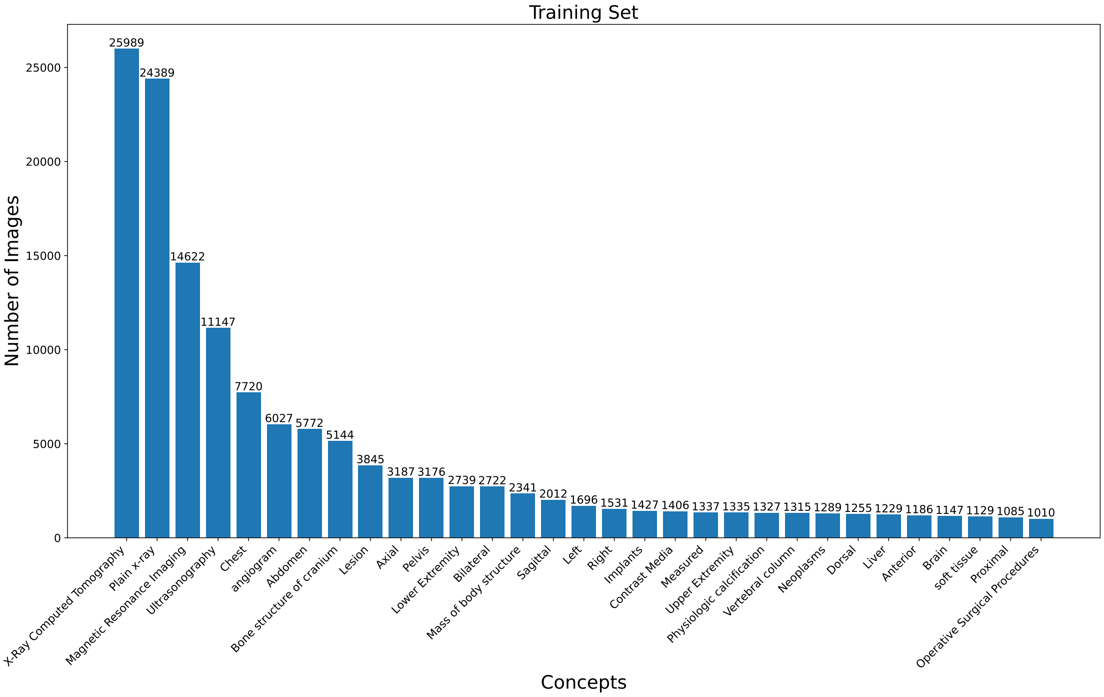

Figure 6: Top-31 most frequent concepts in the training set.

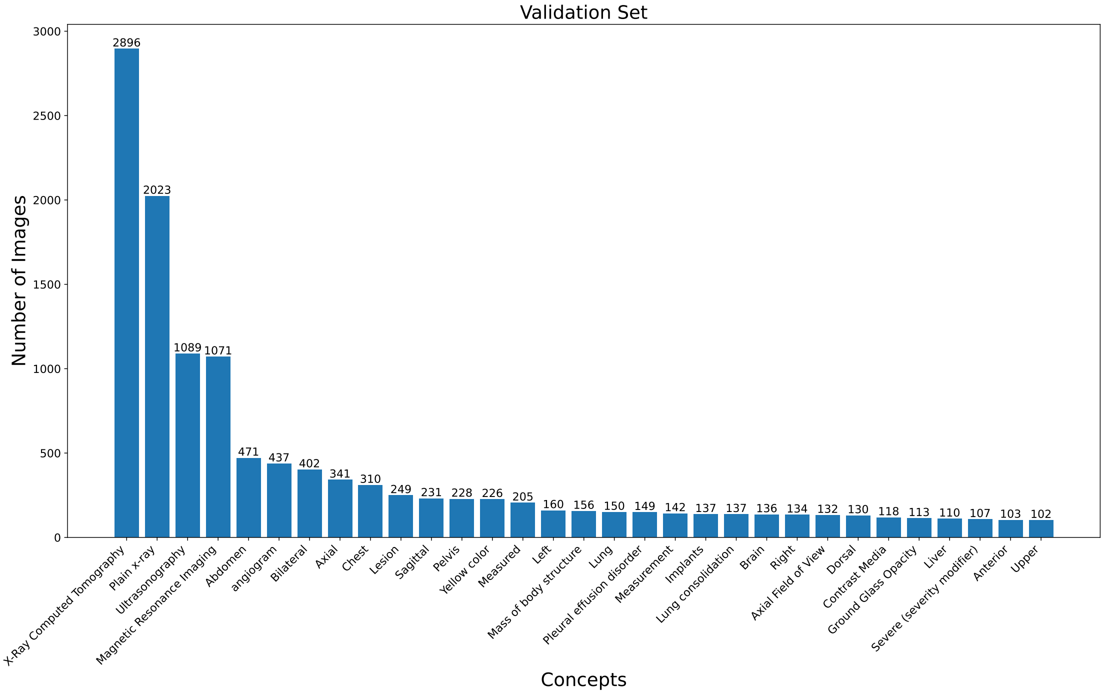

Figure 7: Top-31 most frequent concepts in the validation set.

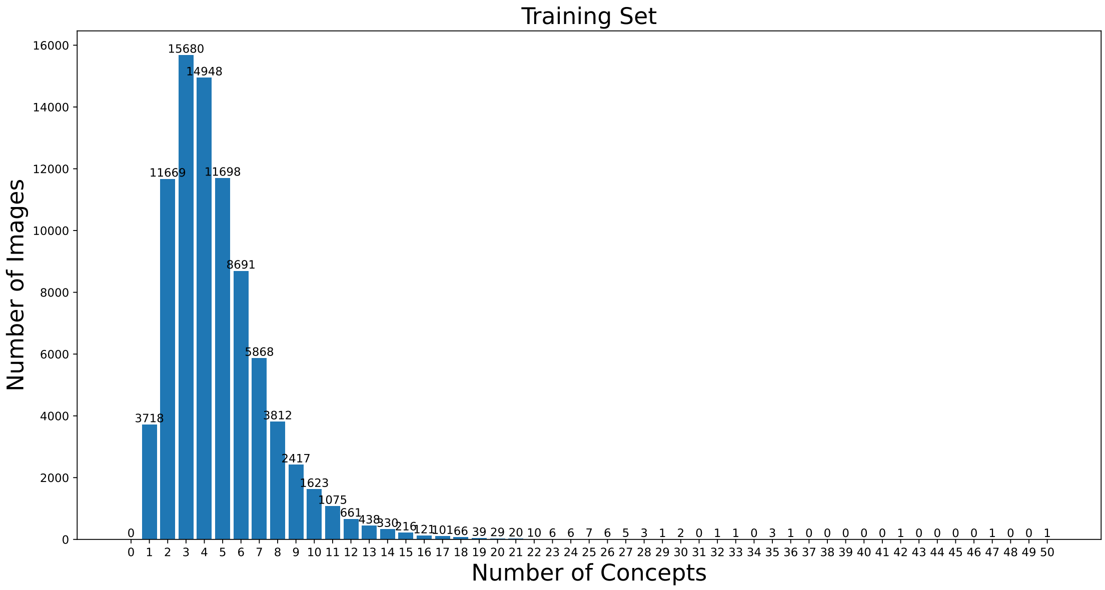

Figure 8: Distribution of the concepts in the training set.

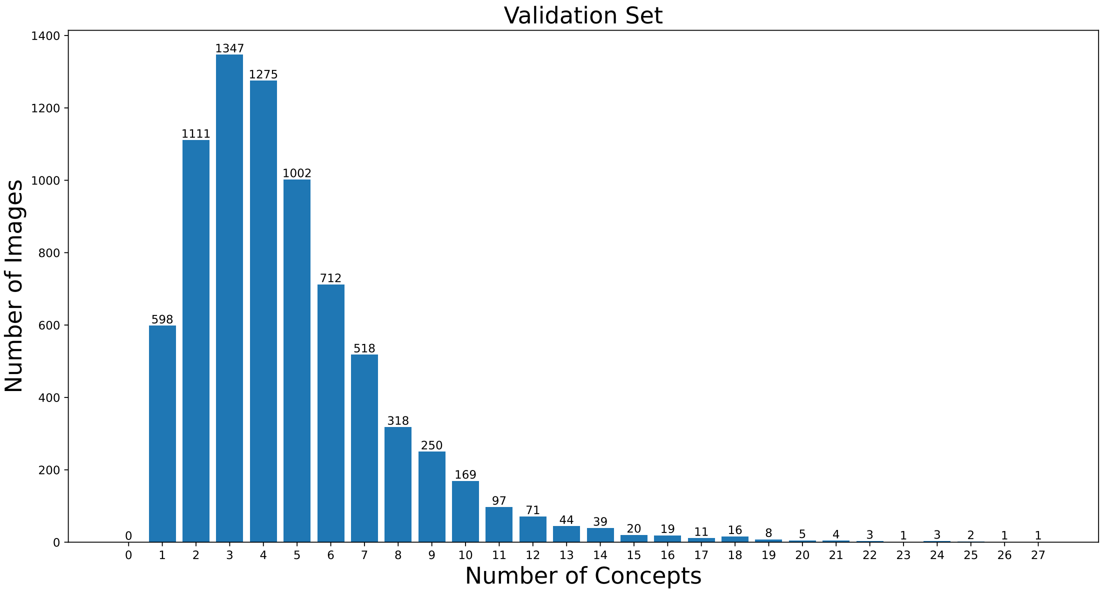

Figure 9: Distribution of the concepts in the validation set.

A.2. Caption prediction

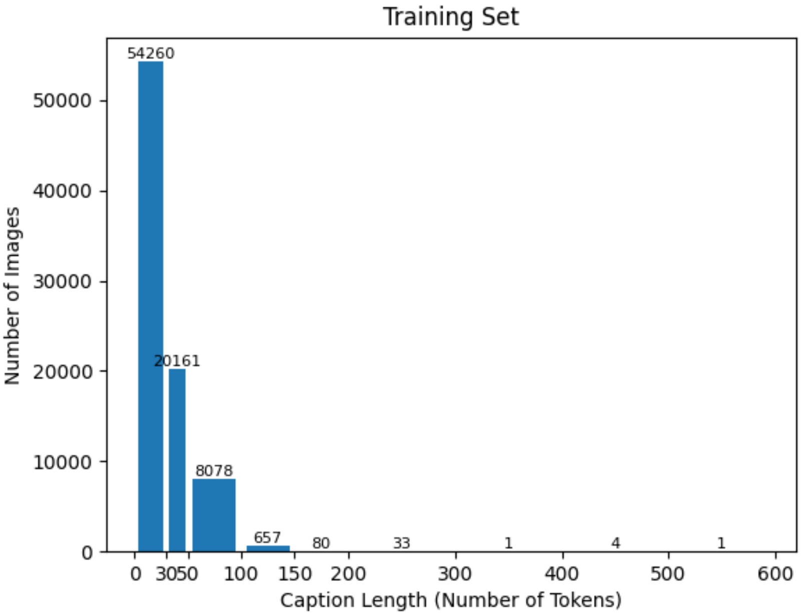

Figure 10: Distribution of the lengths of the captions in the training set.

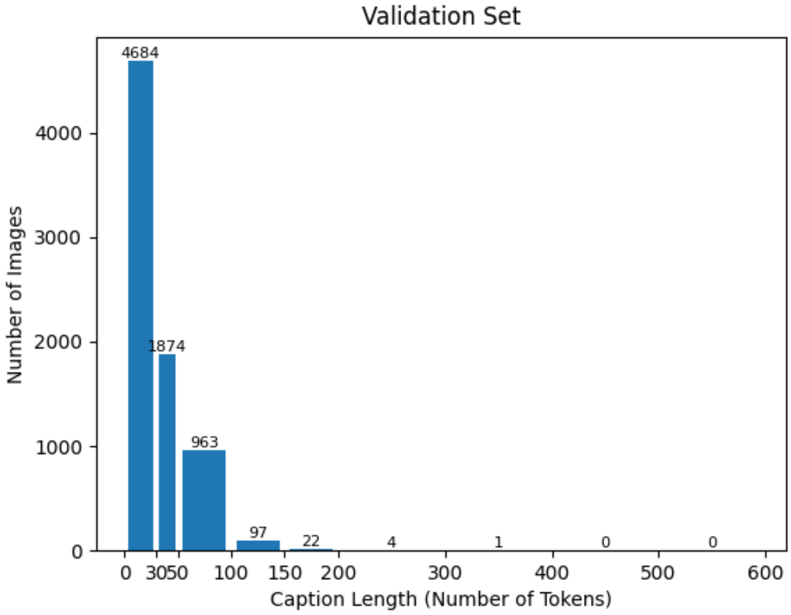

Figure 11: Distribution of the lengths of the captions in the validation set.