Detecting Concepts and Generating Captions from Medical Images

Isabel Rio-Torto [1,2]
isabel.riotorto@inesctec.pt
Cristiano Patrício [2,3]
cristiano.p.patricio@inesctec.pt
Helena Montenegro [2,4]
maria.h.sampaio@inesctec.pt
Tiago Gonçalves [2,4]
tiago.f.goncalves@inesctec.pt
Jaime S. Cardoso [2,4]
jaime.s.cardoso@inesctec.pt

[1] Faculdade de Ciências da Universidade do Porto
Porto, Portugal
[2] INESC TEC
Porto, Portugal
[3] Universidade da Beira Interior
Covilhã, Portugal
[4] Faculdade de Engenharia da Universidade do Porto
Porto, Portugal

Abstract

The automatic extraction of concepts and clinical descriptions from medical images may facilitate the work of clinicians. Besides, it may also contribute towards an increase in the trust of clinicians in artificial intelligence methods since their learning depends on structured clinical information. In this work, we develop and compare various approaches to detect concepts and generate reports (i.e. perform image captioning) from medical images. Regarding concept detection, we explored 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. For the caption prediction task, we developed language generation models and compared them with a simple approach based on image retrieval.

1 Introduction

Developing algorithms capable of extracting concepts from medical images and subsequently generating summarised reports may contribute to the acceleration of clinical diagnosis pipelines, allowing clinicians to increase their time efficiency. Moreover, constraining algorithms to specific clinical data (such as concepts) may increase trust in these models and facilitate their use and adoption by the clinical community. In this work, we develop and compare various approaches to detect concepts and generate reports (i.e. perform image captioning) from medical images. Regarding concept detection, we explored multi-label classification, adversarial training, autoregressive modelling, image retrieval, and concept retrieval. For caption prediction, we develop language generation models, comparing them with a simpler approach based on image retrieval [6]. The developed methods are applied to the medical dataset of the caption prediction task of the medical track of the ImageCLEF Challenge 2023 [7]. The dataset contains a total of $81,828$ images from different modalities and body parts, annotated with $2,125$ concepts. Out of these images, $60,918$ constitute the training set, $10,437$ the validation set, and $10,473$ the testing set. We compare our results with those of the winners of the challenge in 2023 [7].

2 Methods

The following subsections describe the methods developed to tackle concept detection and caption prediction from medical images.

2.1 Concept Detection

We implement two types of approaches for concept detection: multi-label-based approaches and retrieval-based approaches.

2.1.1 Multi-label-based Approaches

The multi-label-based approaches model the concept detection task as a multi-label classification problem. We train a baseline multi-label classification network to predict the presence of the 2125 concepts. The network is trained by minimising the binary cross-entropy between predicted and ground-truth concepts. The main limitation of the baseline network lies in its assumption of independence between concepts. As there may be dependencies between the concepts (*e.g.*, concepts that refer to different body parts may never appear in the same image), we develop two other multi-label classification networks capable of capturing these dependencies using adversarial and autoregressive learning.

The adversarial approach aims to ensure that the model learns realistic combinations of concepts (*e.g.*, concept combinations related to opposite body parts are inadmissible). To achieve this, we built a model that consists of two components: a multi-label classifier trained to predict the *top-K* most frequent concepts ($K = 100$), using a ResNet50 as a feature extractor along with a multi-layer perceptron (MLP) with a sigmoid activation; and concept discriminator trained to distinguish between admissible and inadmissible combinations of concepts, using an MLP with two fully-connected layers followed by a ReLU activation and a fully-connected layer with sigmoid activation.

The autoregressive approach consists of a multi-label classification network that, instead of having a final classification layer to predict all concepts, contains 17 classification layers, each responsible for predicting 125 concepts. The layers are organised in a sequential manner, with each layer being conditioned on both the latent representation of the image and the predictions of the previous layers. Furthermore, the first classification layers of the model are responsible for predicting the most frequent concepts, as these are easier to predict. This model is also trained using the binary cross-entropy loss.

2.1.2 Retrieval-based Approaches

We develop two types of retrieval-based approaches: concept retrieval and image retrieval. Concept retrieval learns to map images and concepts into a common latent space and retrieves the concepts that are the closest to an image on inference. In this approach, we develop an image encoder network to obtain the images' latent representations and a concept encoder to obtain the concepts' latent representations. Then, we train these networks by minimising the Euclidean distance between the latent representations of each image and the concepts it contains, while maximising the distance between each image and the concepts it does not contain.

Image retrieval retrieves the most similar images from the training data, assigning their concepts to the target image. To measure the distance between images, we calculate the Euclidean distance between their latent representations obtained using pre-trained networks. Specifically, we use a ResNet pre-trained on ImageNet, the previously mentioned image encoder of the concept retrieval network and the autoregressive multi-label classification network. Using these models, we retrieve the four most similar images and assign concepts that exist in at least two of these images to the target image. If no concept exists in at least two of the retrieved images, then all the concepts of the most similar image are assigned to the target image.

2.1.3 Ensemble

Since the multi-label-based approaches sometimes predict the absence of all concepts, we develop an ensemble strategy that ensures that all images are assigned at least one concept on inference. This strategy assigns the concepts predicted by one of the retrieval methods to the images for which the multi-label-based approaches fail to predict any concepts.

2.2 Caption Prediction

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

The language generation-based strategies employ an Encoder-Decoder framework, our best performing approach in last year's competition [5]. The Encoder (CNN or Vision Transformer) analyses the image and extracts relevant features, while the Decoder receives these features and generates the caption. Thus, the latter is usually an autoregressive model. We experimented with two different encoders: the small distilled version of the Data-efficient image Transformer (DeiT) [8], and DenseNet121 from

TorchXRayVision [1] pre-trained on all available datasets. The decoder consisted of the distilled version of GPT-2 [4].

The concepts of the concept detection task are tightly related to the captions of the captioning task. Thus, predicting the concepts from the captions might prove a good additional supervisory signal for training the captioning model. Therefore, we explored the inclusion of a text classifier that takes the caption of a given image and predicts its concepts. We added a fully connected layer directly on top of the latent representation of the Decoder's last token and trained the whole Encoder-Decoder plus classification layer together.

Finally, the image retrieval approach developed for concept detection was also applied to the caption prediction task, by retrieving the most similar image and assigning its caption to the target image.

3 Results

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

3.1 Concept Detection

The results for the concept detection task are presented in terms of example-based F1-score between the predicted and ground-truth concepts. Table 1 shows the achieved F1-Scores on both validation and test data. Additionally, it provides an extra Secondary F1-score metric (S. F1-Score) evaluated on a subset of manually validated concepts.

The baseline multi-label classification approach achieved an F1-score of 0.4469 on the test set. Surprisingly, the adversarial approach did not outperform the baseline, possibly due to its limited training on the top-100 concepts. Notably, the autoregressive approach demonstrated the highest performance among the multi-label models.

Regarding the retrieval-based approaches, using image retrieval with the autoregressive model yielded the best results. However, these results did not surpass the performance of the multi-label classification autoregressive model.

As expected, the ensemble methods demonstrate superior performance, with an F1-Score of 0.4998 and S. F1-Score of 0.9162 when combining the autoregressive multi-label classification network with the image retrieval approach using the autoregressive model.

Model	F1-score (Validation)	F1-score (Test)	S. F1-score (Test)
Baseline Multi-label*	0.4364	0.4469	0.8305
Adversarial* (Top-100)	0.2816	0.2803	0.5999
Autoregressive*	0.4905	0.4928	0.9062
Concept Retrieval	0.4523	0.4360	0.7582
IR (ResNet)	0.4693	0.4676	0.8305
IR (Autoregressive)	0.4793	0.4793	0.9014
IR (Concept Retrieval)	0.4379	0.4387	0.8394
Ensemble (Baseline+CR)	-	0.4728	0.8738
Ensemble (Autoregressive)	-	**0.4998**	**0.9162**
Ensemble (Adversarial)	-	0.3327	0.7049
Challenge Winners [2]	-	0.5223	0.9258

Table 1: Concept detection results in terms of F1-score and Secondary (S.) F1-score computed on a subset of manually validated concepts. The models identified with * were trained on the training and validation data. Validation results were obtained using models trained only on the training set. IR: Image Retrieval, CR: Concept Retrieval.

3.2 Caption Prediction

The caption prediction task is evaluated with BERTScore and ROUGE. The results are presented in Table 2. All retrieval-based approaches ranked below the language generation-based approaches. This confirms that solely relying on captions from similar images is insufficient to accurately describe a different image.

Regarding the generation-based approaches, using the DeiT encoder resulted in marginally better results when compared to using DenseNet-121. As expected, we verify that adding the classification loss to the corresponding base architecture slightly improves the results. However, using DeiT jointly with DistilGPT2 remains the best combination, particularly when trained on both training and validation sets. Furthermore, we hypothesize that adding the classification loss to the DeiT would have further improved our results.

Model	BERTScore (Validation)	ROUGE (Validation)	BERTScore (Test)	ROUGE (Test)
IR (ResNet)	0.5738	0.1417	0.5734	0.1427
IR (CR)	0.5653	0.1268	0.5647	0.1284
IR (AR)	0.5756	0.1464	0.5750	0.1464
DeiT + DistilGPT2	0.6133	0.2167	0.6138	**0.2181**
DeiT + DistilGPT2*	0.6133	0.2167	**0.6147**	0.2175
DenseNet + DistilGPT2	0.6108	0.1935	0.6096	0.1938
DenseNet + DistilGPT2 + Clf loss	0.6113	0.1947	0.6103	0.1948
Challenge Winners [3]	-	-	0.6425	0.2446

Table 2: Captioning results on the validation and test sets in terms of BERTScore and ROUGE. The models identified with * were trained on the training and validation data. Validation results were obtained using models trained only on the training set. IR: Image Retrieval, CR: Concept Retrieval, and AR: Autoregressive.

4 Conclusions

This paper introduces the developed approaches for the tasks of concept detection and caption prediction from medical images. In the concept detection task, the experimental results show that the ensemble of an autoregressive multi-label classification network with the image retrieval approach using an autoregressive model obtains the best F1-Score. Regarding the caption prediction task, we conclude that generation-based approaches outperform the retrieval-based approaches, and that using DeiT as the Encoder instead of DenseNet produced our best results.

Acknowledgements

We would like to thank our colleague Pedro Neto for his valuable feedback and suggestions. This work is co-financed by Component 5 - Capitalization and Business Innovation, integrated in the Resilience Dimension of the Recovery and Resilience Plan within the scope of the Recovery and Resilience Mechanism (MRR) of the European Union (EU), framed in the Next Generation EU, for the period 2021 - 2026, within project HfPT, with reference 41, and by National Funds through the Portuguese funding agency, FCT-Foundation for Science and Technology Portugal, within PhD grants 2022.14516.BD, 2022.11566.BD, 2020.06434.BD and 2020.07034.BD.

References

[1] Joseph Paul Cohen, Joseph D. Viviano, Paul Bertin, Paul Morrison, Parsa Torabian, Matteo Guarrera, Matthew P Lungren, Akshay Chaudhari, Rupert Brooks, Mohammad Hashir, and Hadrien Bertrand. TorchXRayVision: A library of chest X-ray datasets and models. In *MIDL*, volume 172, pages 231–249, Online, 06–08 Jul 2022. PMLR.

[2] Panagiotis Kaliosis et al. Aueb nlp group at imageclefmedical caption 2023. In *CLEF2023 Working Notes*, pages 1524–1548, 2023.

[3] Aaron Nicolson et al. A concise model for medical image captioning. In *CLEF2023 Working Notes*, pages 1611–1619, 2023.

[4] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language Models are Unsupervised Multitask Learners. *OpenAI Blog*, 1(8):9, 2019.

[5] Isabel Rio-Torto, Cristiano Patrício, Helena Montenegro, and Tiago 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*, CEUR Workshop Proceedings, pages 1535–1553. CEUR-WS.org, 2022.

[6] Isabel Rio-Torto et al. Detecting concepts and generating captions from medical images: Contributions of the VCMI team to ImageCLEFmedical Caption 2023. In *CLEF2023 Working Notes*, CEUR Workshop Proceedings, pages 1653–1667, 2023.

[7] Johannes Rückert et al. Overview of ImageCLEFmedical 2023 – Caption Prediction and Concept Detection. In *CLEF2023 Working Notes*, CEUR Workshop Proceedings, pages 1328–1346, 2023.

[8] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers & distillation through attention. In *ICML*, volume 139, pages 10347–10357, Online, 18–24 Jul 2021. PMLR.