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[1,2] and Tiago Gonçalves[1,2]

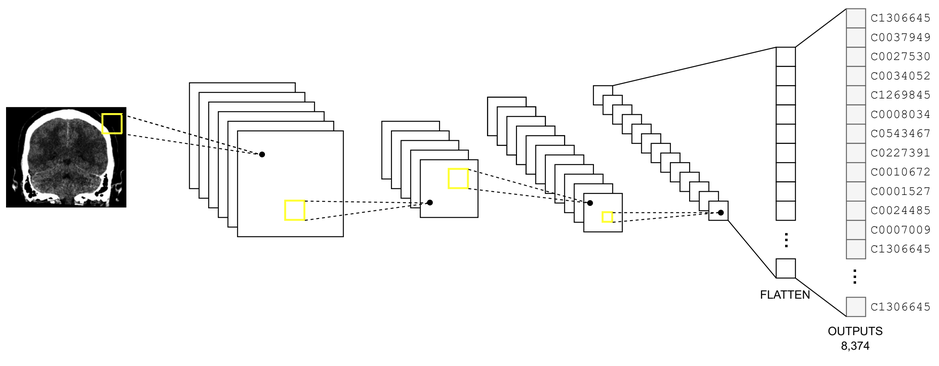
[1]Universidade do Porto, Porto, Portugal
[2]INESC TEC, Porto, Portugal
[3]Universidade da Beira Interior, Covilhã, Portugal

Concept Detection

The Challenge: identify relevant medical concepts in a large corpus of radiology images

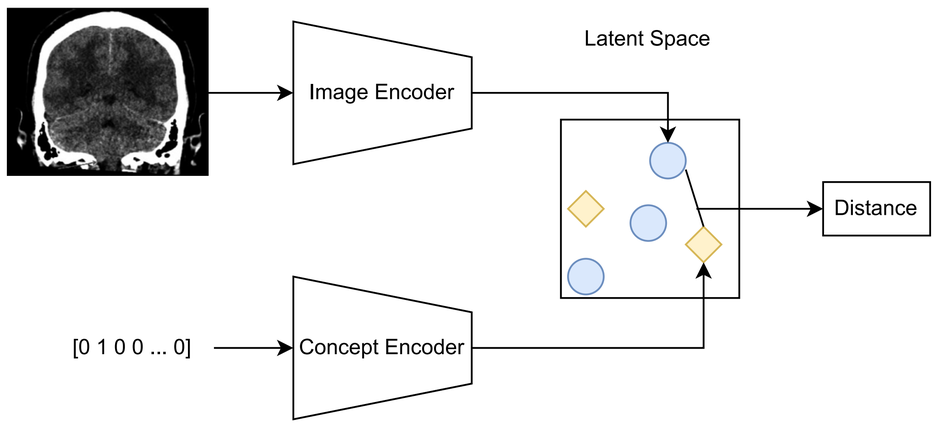
Approach 1: Multi-label Classification

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.



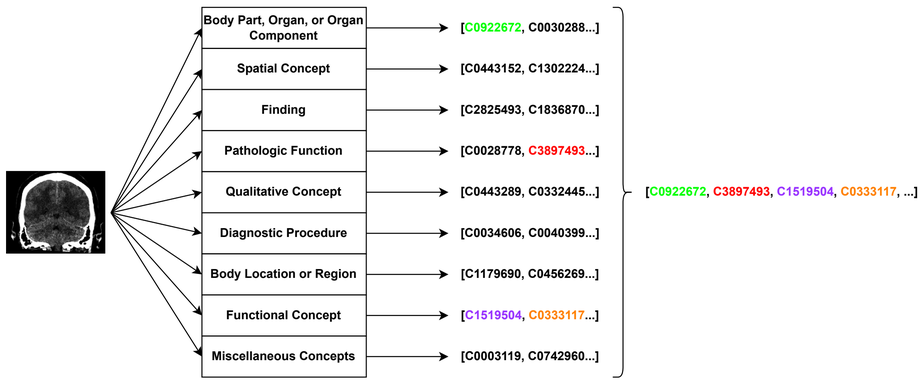
Approach 2: Concept Retrieval

Another approach is to perform concept retrieval: images and concepts are mapped into a common latent space (via a contrastive loss) where the images are expected to be closer to the concepts they contain.



Approach 3: Semantic Multi-label Classification

This approach consists in training one model per concept semantic type. During inference, each image is given to every model and the final set of predicted concepts corresponds to the union of the concepts predicted by each model.



Results

Table 1. 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.

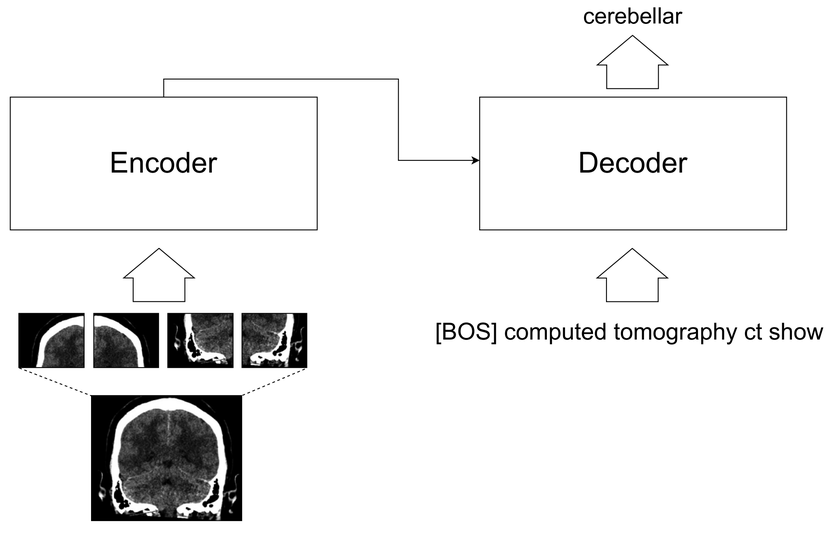
Model	Concepts	F1-score (Validation)	F1-score (Test)	Secondary F1-score (Test)
Multi-label (Frozen Backbone)	All	0.3710	-	-
Multi-label (Frozen Backbone)	Top-100	0.3740	-	
Multi-label (Whole Network)	Top-100	0.3947	0.430	0.861
Multi-label (2 Phases)	Top-100	0.3937	0.431	0.856
Euclidean Retrieval	All	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

Caption Prediction

The Challenge: generate coherent textual descriptions of radiology images

Approach 1: Vision Encoder-Decoder Transformer

This approach combines the original Transformer [4] architecture with the Vision Transformer (ViT) [2] as encoder. The model is trained autoregressively for next token prediction with causal (or unidirectional) self-attention.



Approach 2: Modified OSCAR

This architecture is based on the hypothesis that leveraging the medical concepts of each image might aid in generating the captions. The OSCAR model was modified, given that it uses object tags and region features obtained from an object detector, annotations that the provided data does not include. As such, this modified version receives as input the image divided into 16×16 patches similarly to what is done in the ViT model [2]. The model is trained for masked language modeling with causal self-attention.

Results

Table 2. 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

References

[1] Anna Curtis, Christian Lamb, Hussain Rao, Andrew Williams, and Amit Patel. Dialysis Disequilibrium Syndrome and Cerebellar Herniation with Successful Reversal Using Mannitol. *Case Reports in Nephrology*, 2020:1–4, aug 2020.

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[3] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 121–137, 2020.

[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. *Advances in Neural Information Processing Systems*, 30, 2017.