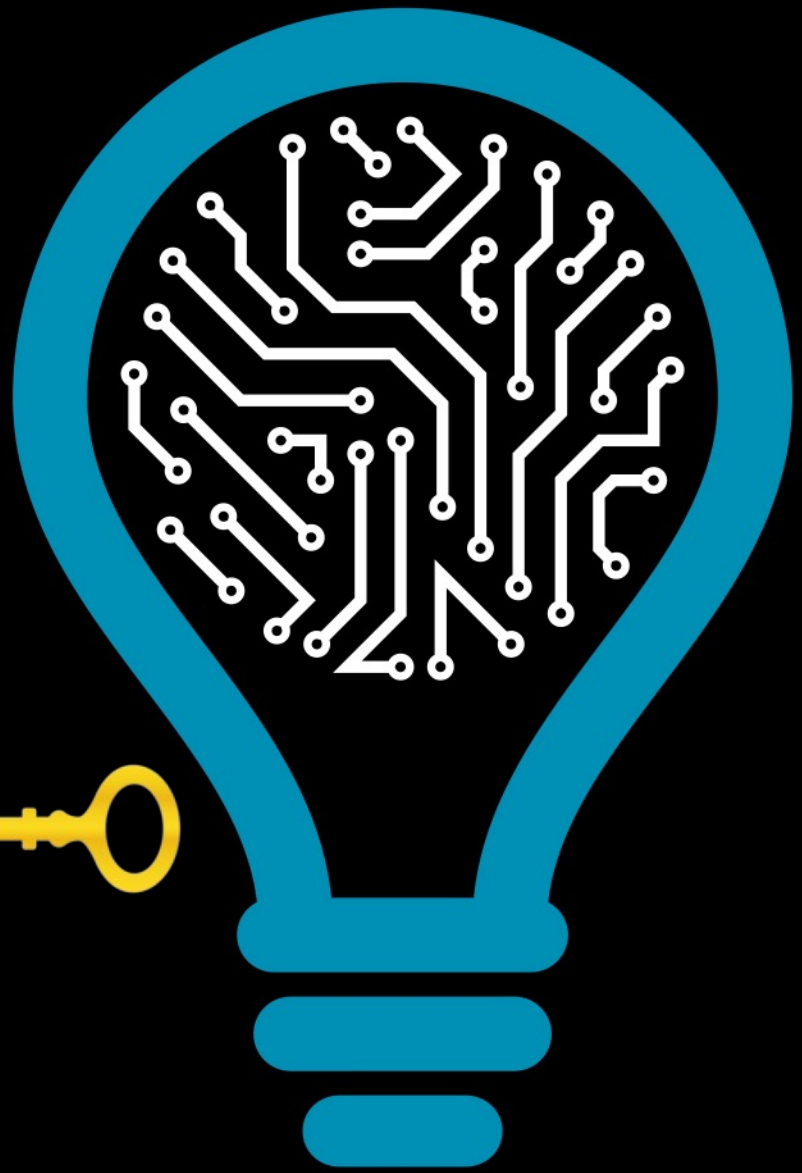
Detecting Concepts and Generating Captions from Medical Images: Contributions of the VCMI Team to ImageCLEFmedical 2022 Caption

Isabel Rio-Torto, Cristiano Patrício, Helena Montenegro and Tiago Gonçalves

isabel.riotorto@inesctec.pt

CLEF 2022 – Bologna, Italy
7 September 2022



VISUAL COMPUTING & MACHINE INTELLIGENCE RESEARCH GROUP

INESCTEC

U.PORTO

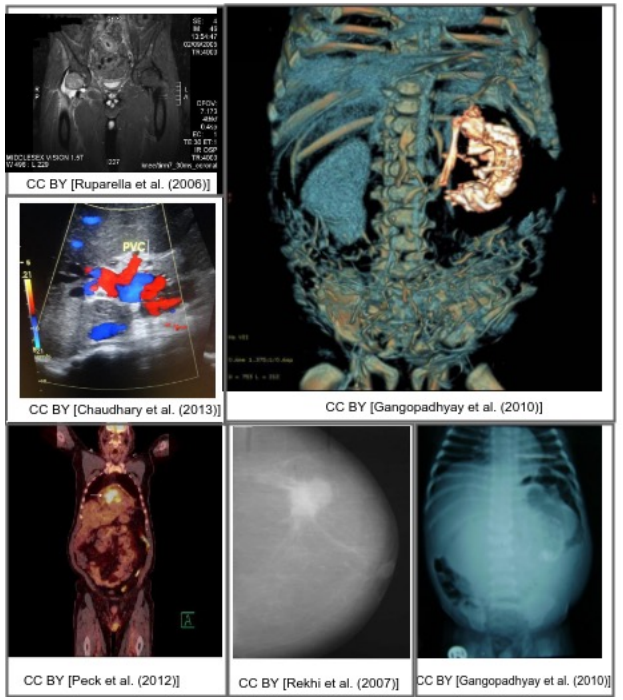
UNIVERSIDADE BEIRA INTERIOR

ImageCLEFmedical 2022

- Concept Detection
  - identify the presence of relevant concepts in a large corpus of medical images
  - multi-label multiclass classification

- Caption Prediction
  - generate coherent textual descriptions of a medical image

- Dataset: Radiology Objects in COntext (ROCO) [1]
  - Training Set: 83,275 radiology images
  - Validation Set: 7,645 radiology images
  - Test Set: 7,601 radiology images

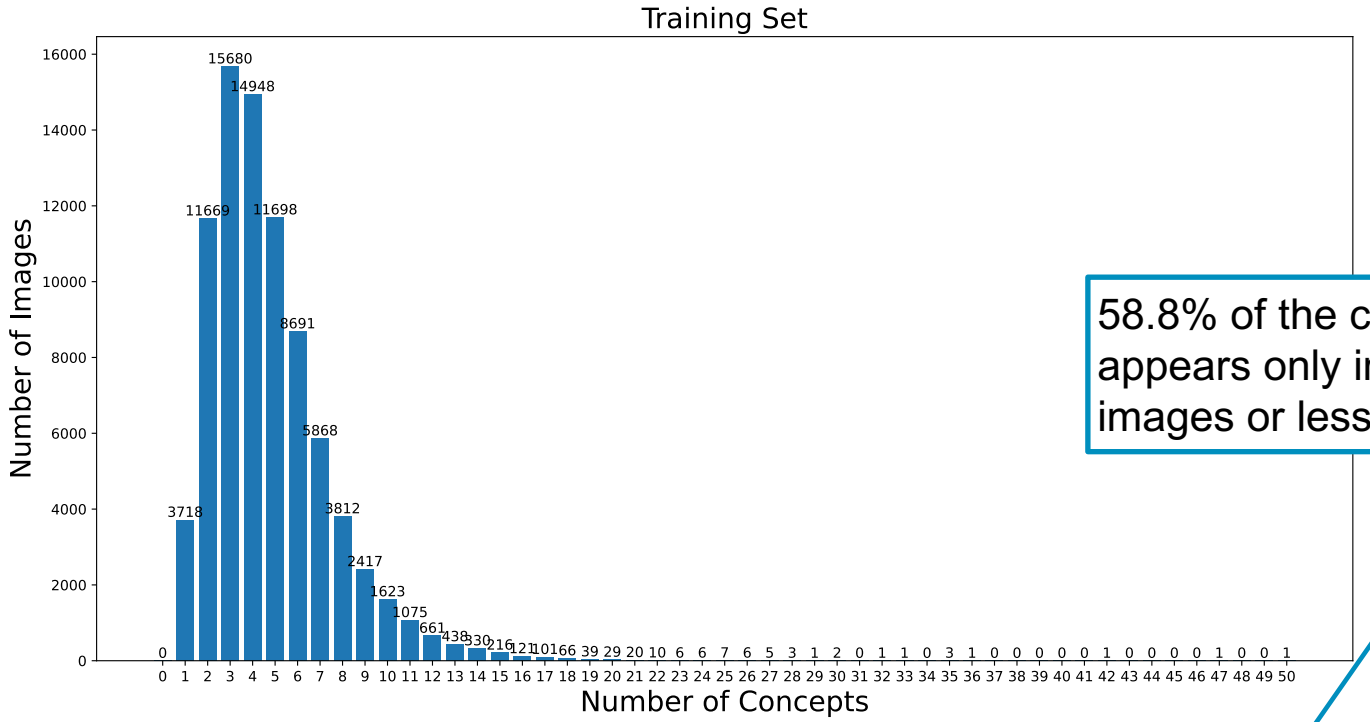


[1] O. Pelka, S. Koitka, J. Rückert, F. Nensa und C. M. Friedrich „Radiology Objects in COntext (ROCO): A Multimodal Image Dataset“, Proceedings of the MICCAI Workshop on Large-scale Annotation of Biomedical data and Expert Label Synthesis (MICCAI LABELS 2018), Granada, Spain, September 16, 2018, Lecture Notes in Computer Science (LNCS) Volume 11043, Page 180-189, DOI: 10.1007/978-3-030-01364-6_20, Springer Verlag, 2018.

INESCTEC

Concept Detection

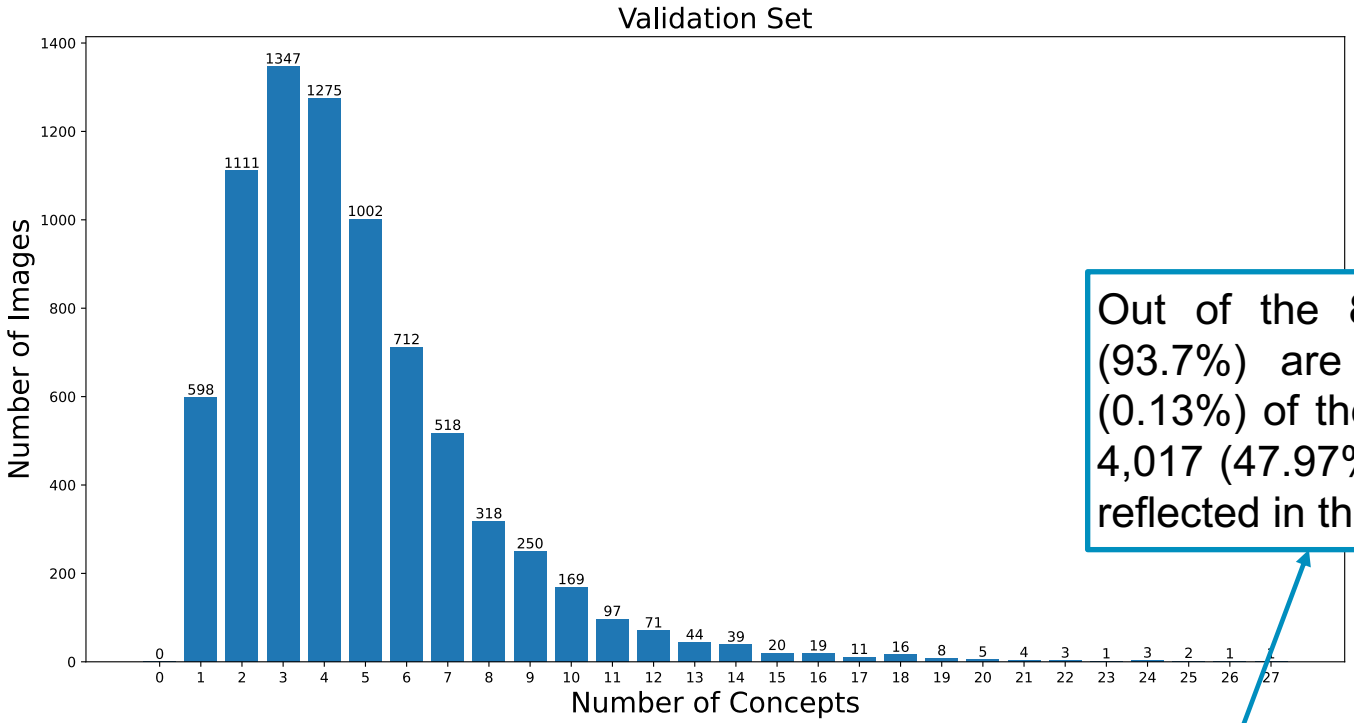
Data Analysis



58.8% of the concepts available in the data appears only in 10 (0.012%) of the training images or less

Perspective	Total	Avg	Min	Max	< 10 imgs	0 imgs
Image-based	83,275	4.7	1	50	-	-
Concept-based	8,374	47.2	2	25,989	4,923	0

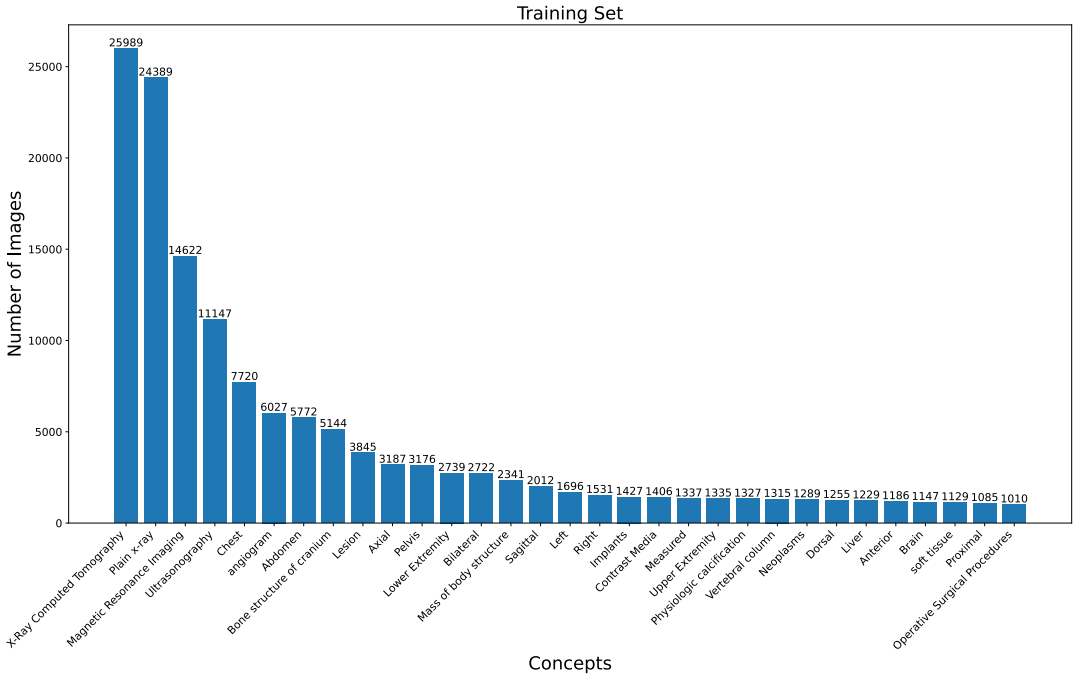
Data Analysis



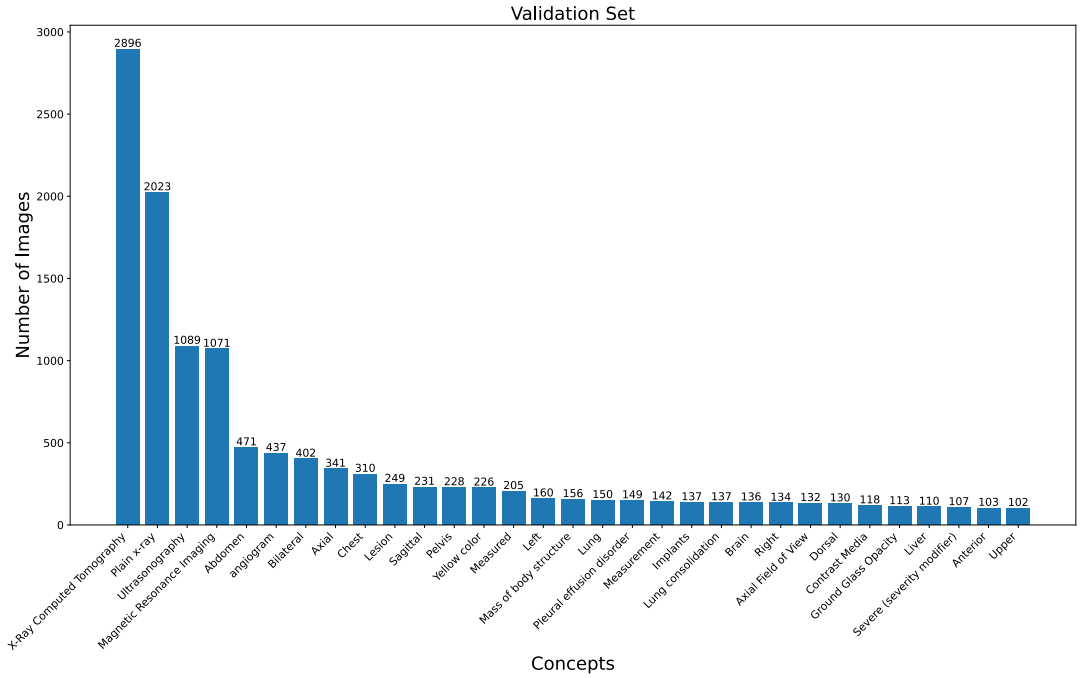
Out of the 8,374 total concepts, 7,842 (93.7%) are reflected in less than 10 (0.13%) of the validation images, of which 4,017 (47.97% out of all concepts) are not reflected in the validation data at all.

Perspective	Total	Avg	Min	Max	< 10 imgs	0 imgs
Image-based	7,645	4.7	1	27	-	-
Concept-based	4,357	4.3	0	2,896	7,842	4,017

Data Analysis

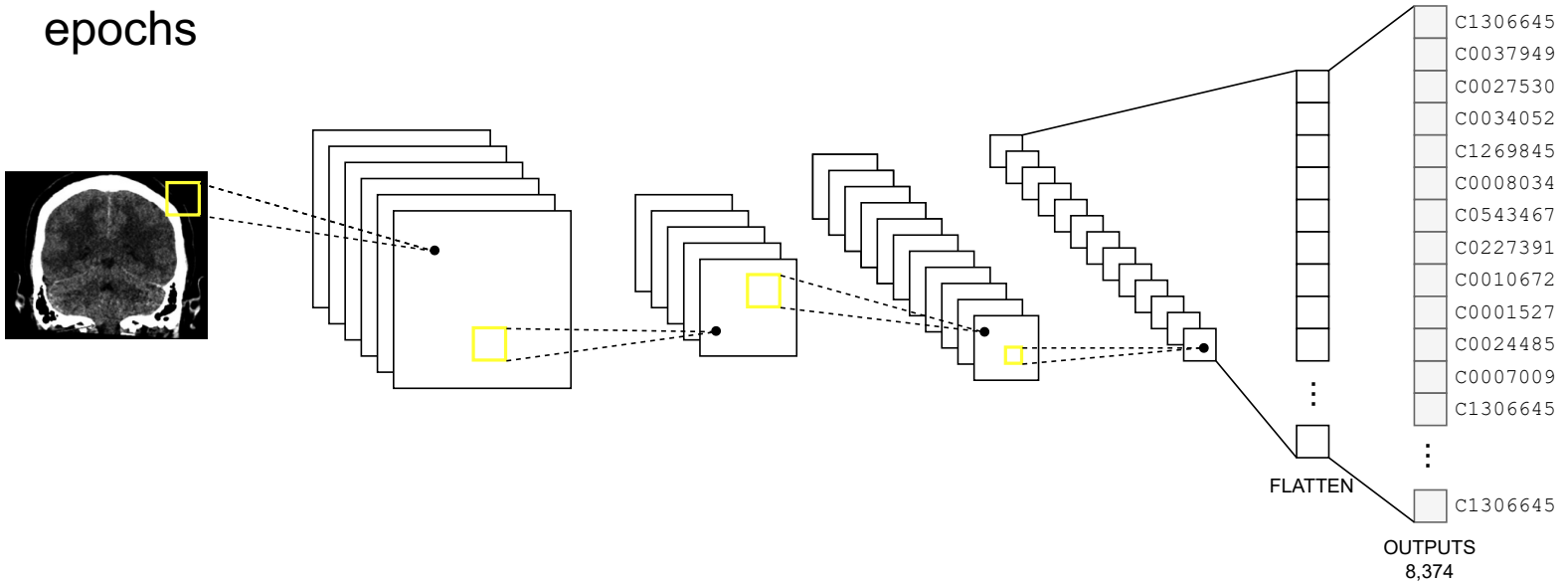


- 21 of the 31 most frequent concepts in the training data are also the most common in the validation data
- 83,110 (99.8%) of the training images and 7,617 (99.6%) of the validation images contain at least one of the Top-100 most frequent concepts



Approach 1: Multi-label Classification

- Straightforward multi-label classification approach
- DenseNet-121
- 2 alternatives:
  - predict all 8374 concepts
  - predict top-100 (most frequent) concepts
- Hyperparameters: 100 epochs, 1e-3 learning rate, Adam optimizer
- 3 training strategies:
  - “Frozen Backbone”
  - “Whole Network”
  - “2 Phases”: frozen backbone for 5 epochs and training whole network for the remaining epochs

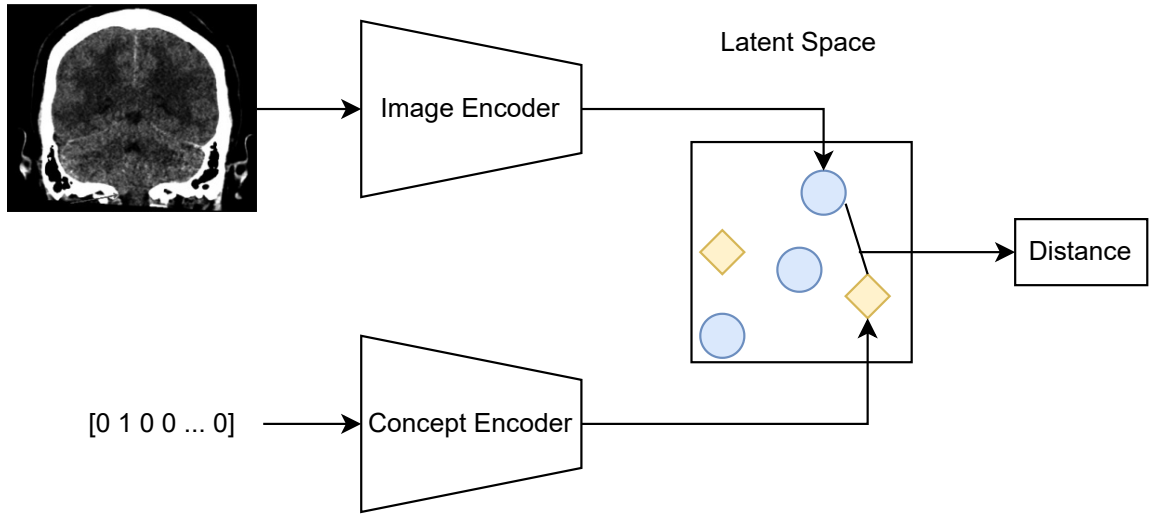


Approach 2: Concept Retrieval

- Concept retrieval + contrastive learning: images and concepts are mapped into a common latent space where the images are expected to be closer to the concepts they contain
- Image encoder: CNN with 4 blocks of convolutional
- Concept encoder: MLP with 2 fully-connected layers with LeakyReLU and Tanh activations

- Contrastive loss: minimise the distance between images and their corresponding concepts while maximising the distance between images and concepts they do not contain

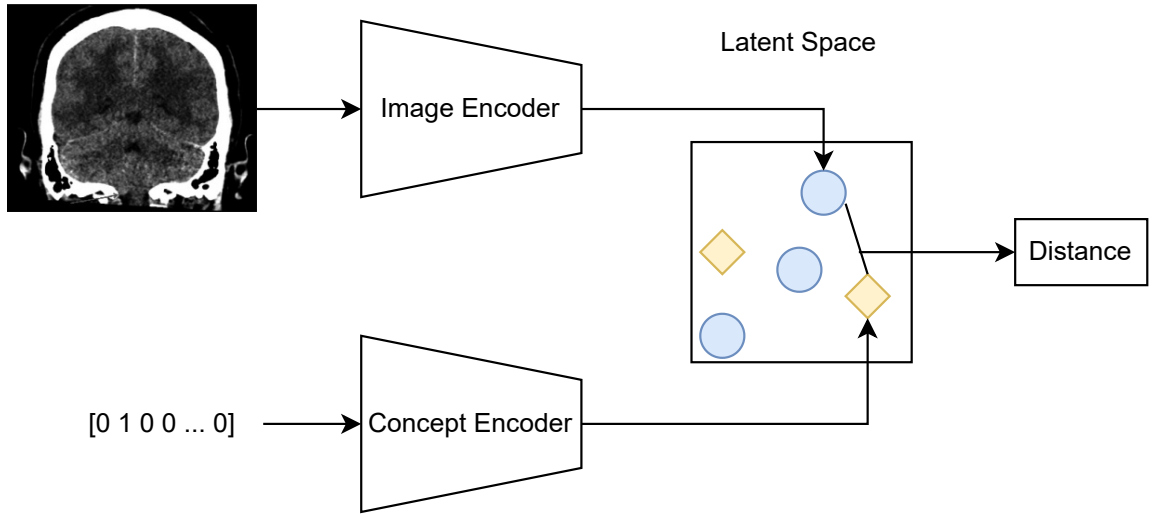
$$\mathcal{L}_{contrastive} = y \times D^2 + (1 - y)[\max(0, 1 - D)]^2$$



Approach 2: Concept Retrieval

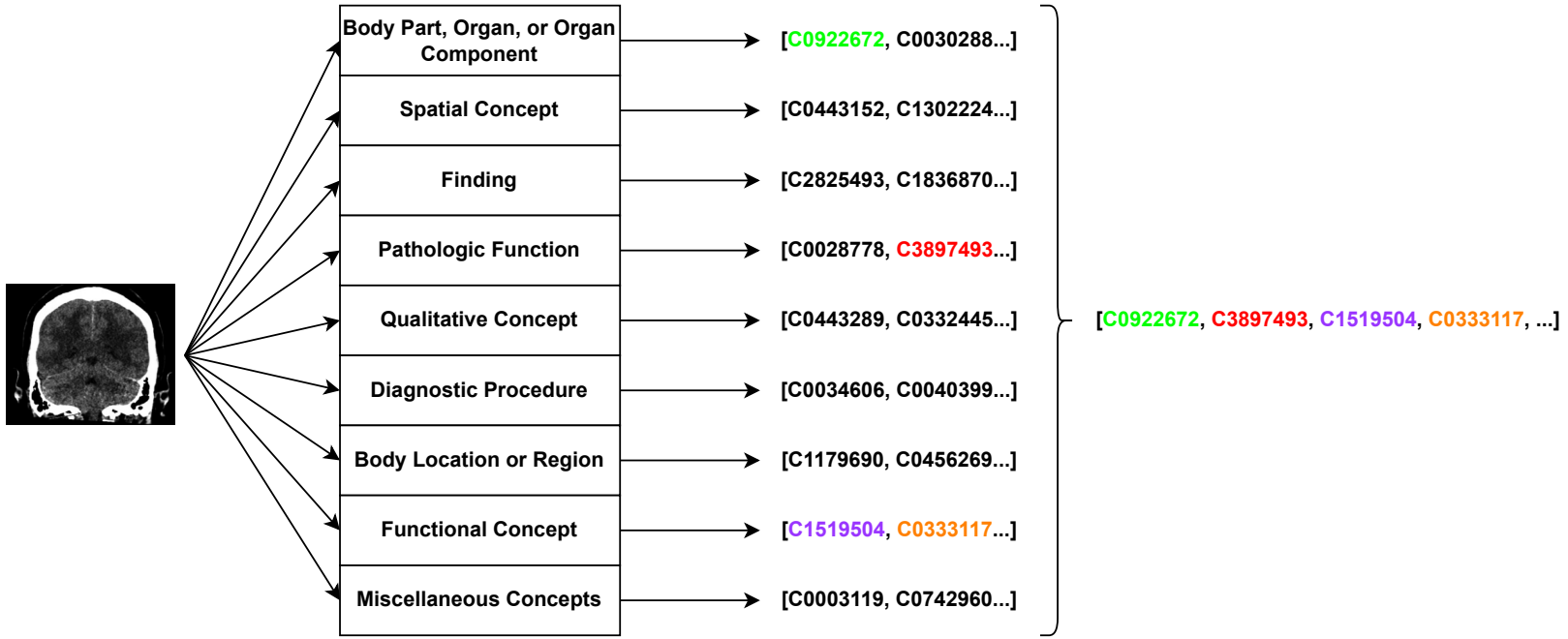
- Experiments (Adam with 1e-6 learning rate):
  - Euclidean distance + all 8, 374
  - Euclidean distance + top-100 most frequent concepts
  - Cosine Similarity + top-100 most frequent concepts

- Ensemble: Approach 1 + Approach 2:
  - “Ensemble (NaN)”: when no label is predicted by the multi-label model, the concept retrieval model is used
  - ”Ensemble (OR)”: merge the predictions of the two models using an OR operation



Approach 3: Semantic Multi-label Classification

- Hierarchical approach
- 1 multi-label classification model per semantic type
- 8 most-frequent semantic types in the top-100 concepts + Miscellaneous Concepts Category
- Inference: union of the outputs of every model
- Hyperparameters: ResNet18, 10 epochs, Adam optimizer, 1e-4 learning rate
- Used the “2 Phases” training process



Results

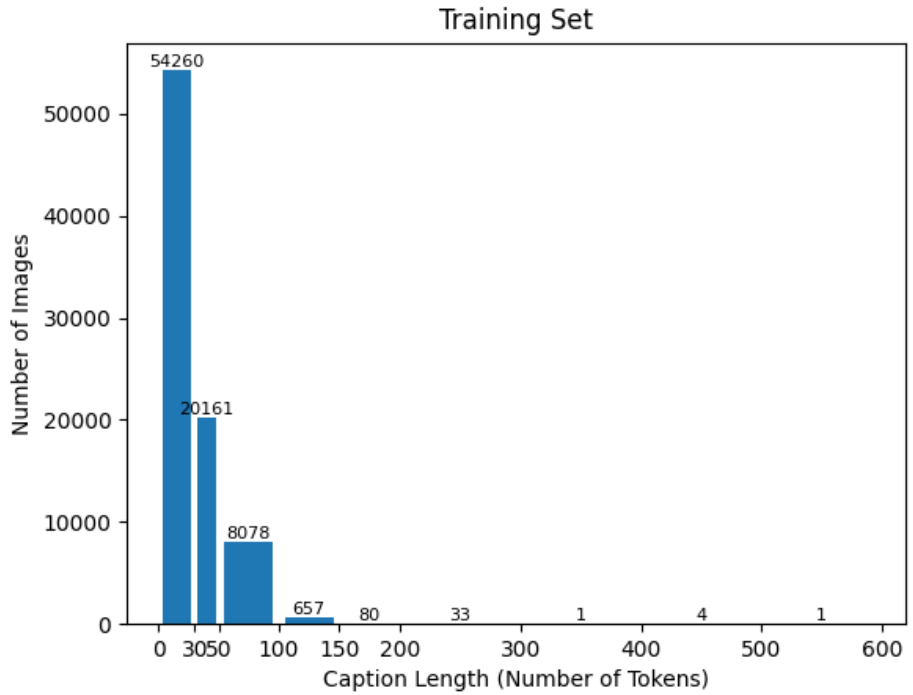
Results of the concept detection task in terms of F1-score and Secondary F1-score computed on a subset of manually validated concepts. “Top-100” and “All” refer to the (sub)set of concepts used to train the models.

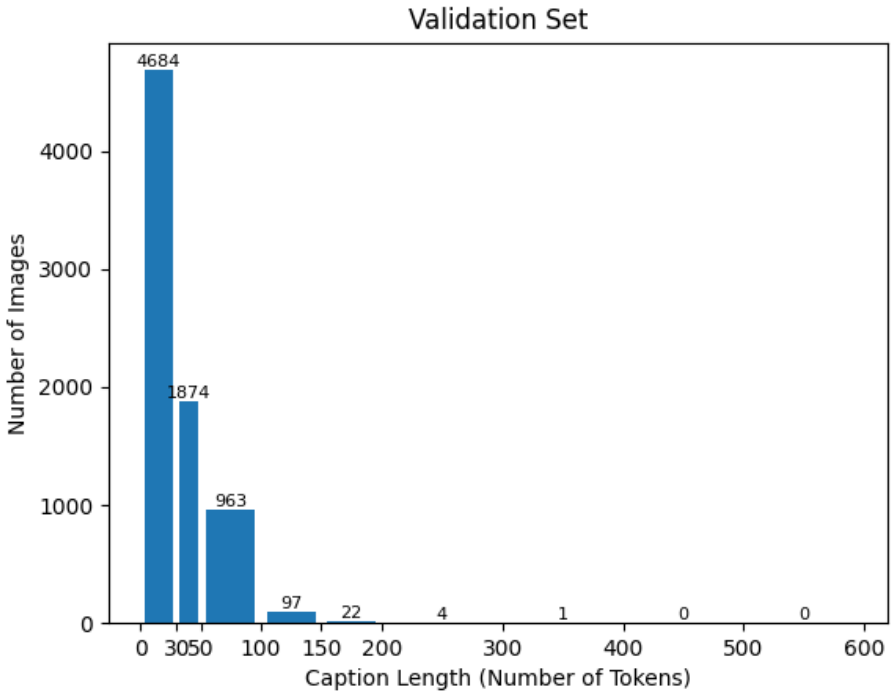
Model	Concepts	F1-score (Validation)	F1-score (Test)	Secondary F1-score (Test)
Multi-label (Frozen Backbone)	All	0.3710	-	-
Multi-label (Frozen Backbone)	Top-100	0.3740	-	
Multi-label (Whole Network)	Top-100	0.3947	0.430	0.861
Multi-label (2 Phases)	Top-100	0.3937	0.431	0.856
Euclidean Retrieval	All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

INESCTEC

Caption Prediction

Data Analysis

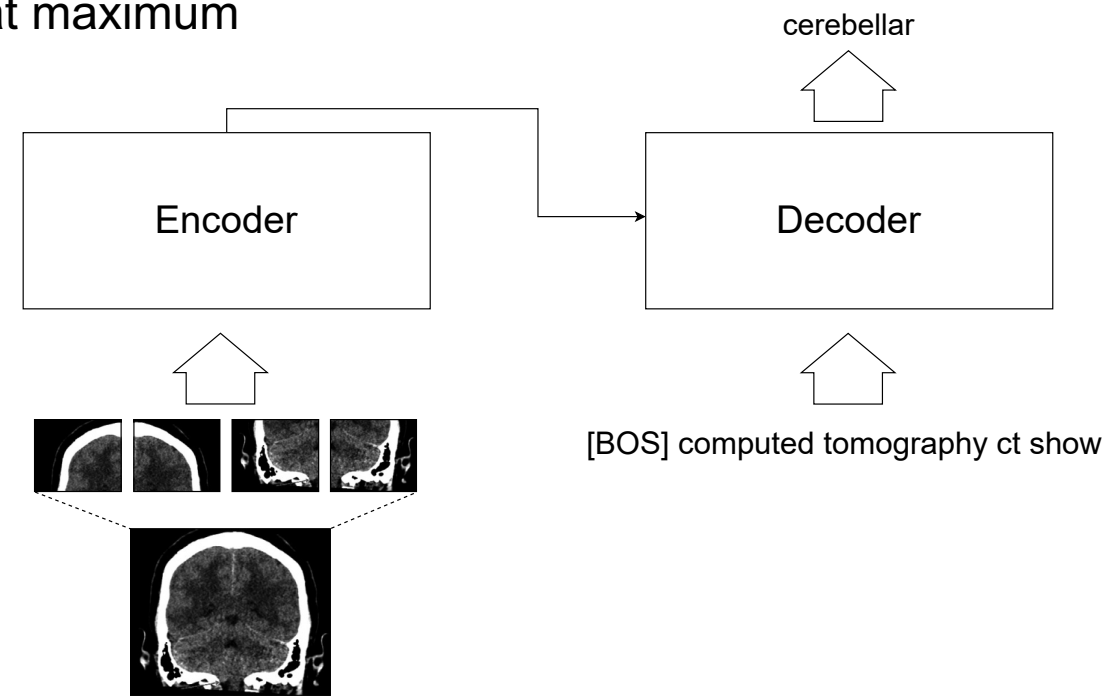




Subset	Avg	Min	Max	< 50 tokens	< 100 tokens
Train	29.73	3	577	89.4%	99.1%
Val	32.37	3	339	85.8%	98.4%

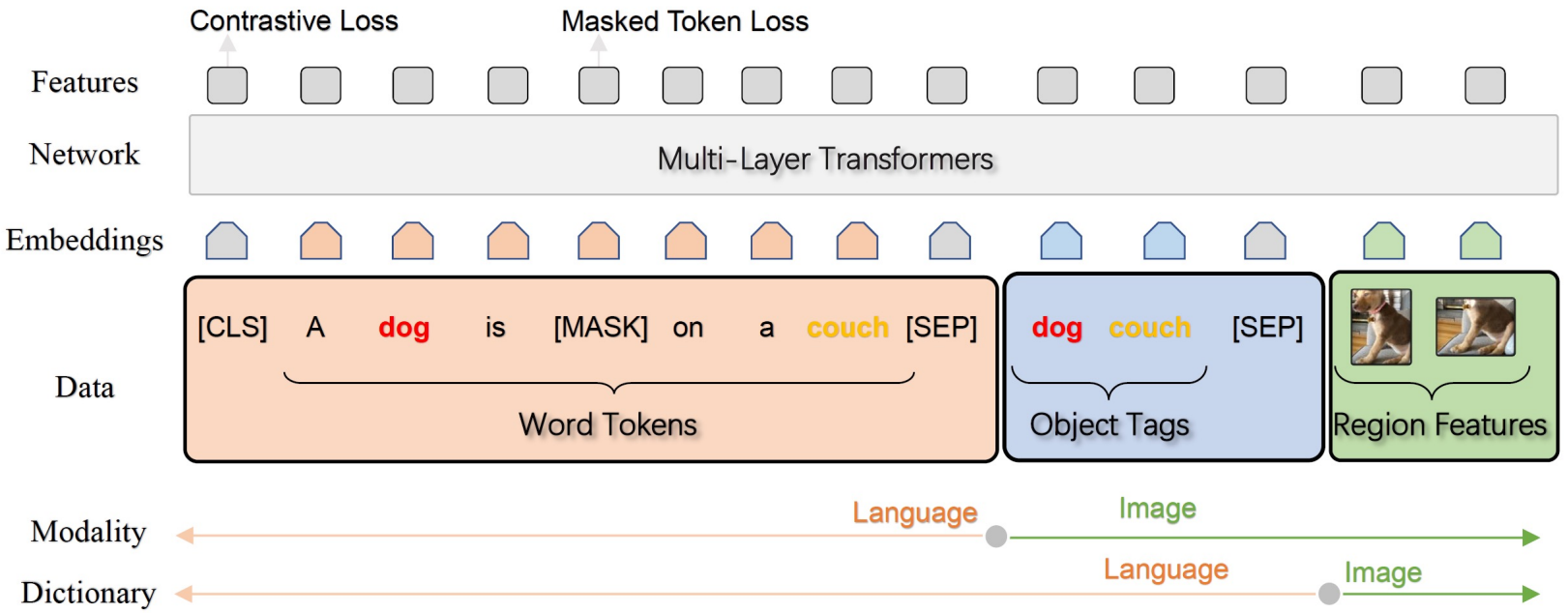
Approach 1: Vision Encoder-Decoder

- Original Transformer w/ Vision Transformer (ViT) [1] as encoder
- Next token prediction w/ causal self-attention
- Encoder: Data-eficient image Transformer (DeiT) [2] pretrained on ImageNet
- Decoder: Distilled-GPT2 [3]
- Hyperparameters: 20 + 20 epochs, AdamW optimizer, 5e-5 linearly decayed learning rate, 100 tokens at maximum



[1] 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
[2] 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
[3] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language Models are Unsupervised Multitask Learners, OpenAI Blog 1 (2019) 9

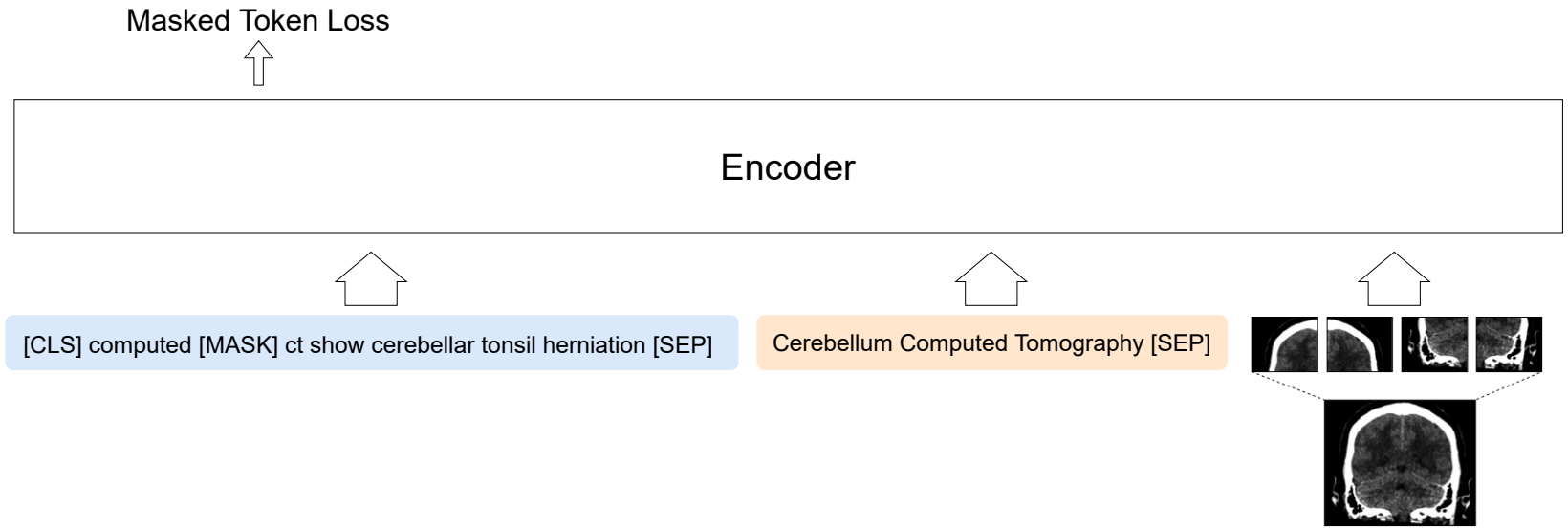
Approach 2: Modified OSCAR



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

Approach 2: Modified OSCAR

- Hypothesis: leveraging the information present in the concepts might aid in generating the captions
- Masked Language Modelling w/ causal self-attention
- Training: used ground-truth concepts
- Inference: used concepts predicted by “Ensemble (NaN)” model
- Hyperparameters: 20 epochs, AdamW optimizer, 1e-4 linearly decayed learning rate,
- Max caption length: 50 tokens
- Max concept sequence length: 10 tokens



Results

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

Conclusions and Future Work

- Concept Detection:
  - 3 approaches + ensemble
  - used subset of 100 most frequent concepts
  - 5th place out of 11 (primary F1-score)
  - best secondary F1-score

- Caption Prediction:
  - 2 Transformer-based approaches
  - 4th place out of 10

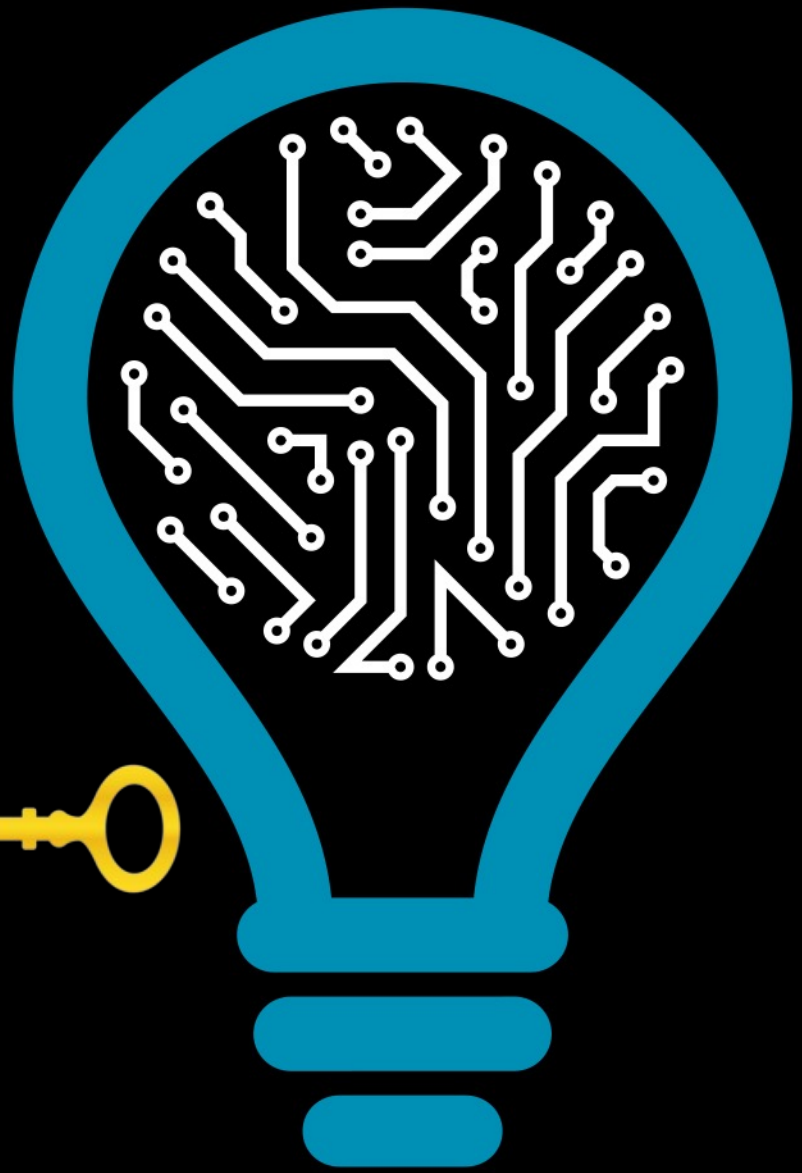
- Future Work:
  - include predicted concepts during training of the modified OSCAR model
  - ablation study: train the modified OSCAR model without concepts

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

Isabel Rio-Torto, Cristiano Patrício, Helena Montenegro and Tiago Gonçalves

isabel.riotorto@inesctec.pt

CLEF 2022 – Bologna, Italy
7 September 2022



VISUAL COMPUTING & MACHINE INTELLIGENCE RESEARCH GROUP

INESCTEC

U.PORTO

UNIVERSIDADE BEIRA INTERIOR