

# Predicting Captions and Detecting Concepts for Medical Images: Contributions of the DBS-HHU Team to ImageCLEFmedical Caption 2024

Notebook for ImageCLEFmedical Caption Lab at CLEF 2024

Heiko Kauschke<sup>1,\*</sup>, Kirill Bogomasov<sup>1,†</sup> and Stefan Conrad<sup>1</sup>

<sup>1</sup>Heinrich-Heine-Universität Düsseldorf, 1 Universitätsstraße, Düsseldorf, 40225, Germany

## Abstract

This paper describes the work of the team DBS-HHU in the ImageCLEFmedical Caption 2024 in both sub-tasks Concept Detection and Caption Prediction. The goal of the Concept Detection sub-task is to extract the correct UMLS terms from medical images, while Caption Prediction aims to generate descriptions for them. For both sub-tasks are images from the Radiology Objects in COntext Version 2 dataset used. We preprocessed these images by removing their white borders and upscaling small images to improve the performance of our models. For Concept Detection we used two different architectures, the first one being an ensemble of four different Convolutional Neural Network (CNN) and the second being a hierarchical model consisting of two CNN. All models in this sub-task are compared by using the  $F_1$ -score. For Caption Prediction we experimented with two different version of the GIT architecture. These were compared to other models using the BERTScore as primary and ROUGE as secondary metric. Our ensemble scored the first place in Concept Detection with a  $F_1$ -score of 0.6374, while our GIT model placed tenth in Caption Prediction.

## Keywords

Multi-Label-Classification, Image Captioning, Deep Learning, CNN Ensemble, Hierarchical Model, GIT, ImageCLEFmedical 2024 Caption

## 1. Introduction

Analyzing and summarizing information derived from medical images, such as those produced in radiology, is a complex and time-intensive task requiring specialized expertise. This process often creates a bottleneck in clinical diagnosis workflows and therefore requires special attention.

As a result, there is a significant demand for automated methods that can translate visual data into concise textual descriptions. Improved knowledge of image features leads to more organized radiology scans, thereby enhancing the efficiency of radiologists in their interpretative work. Challenging tasks and unresolved issues in the field of visual analysis and interpretation often hold significant societal value and are rightfully of great interest to society, research, and industry. Particularly medical imaging is both demanding and valuable in interpretation due to the informational content. Challenging questions and the search for answers and solutions in image material is where ImageCLEF begins. ImageCLEF is the multimedia retrieval lab of CLEF (Conference and Labs of the Evaluation Forum). Since 2004 ImageCLEFmedical has consisted of various tasks. ImageCLEF2024 [1] included, among other tasks, the ImageCLEFmedical 2024 Caption [2] task. The task took place for the eighth time. On one hand, the fact that no satisfactory solution had been found in eight years (otherwise the challenge would be considered finished) suggests the complexity of the task. On the other hand, it indicates the significant interest of the research community in the problem, which has piqued our interest. The task itself is split into two sub-tasks: Concept Detection and Caption Prediction. The first sub-task can be considered as a multi-label classification problem. Each image is associated with at least one manually annotated Unified Medical Language System (UMLS) concept, which we will refer to as a concept or label throughout the

---

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

\*Corresponding author.

†These authors contributed equally.

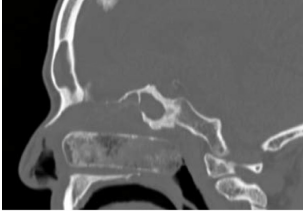
✉ hekau101@hhu.de (H. Kauschke); bogomasov@hhu.de (K. Bogomasov); stefan.conrad@hhu.de (S. Conrad)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

**Table 1**

Example of an image with corresponding CUIs and caption from the ImageCLEFmedical 2024 caption task dataset.

Image	Concepts	Caption
 <p>CC BY [Kelesidis et al. (2010)]</p>	<ul style="list-style-type: none"> <li>• C0040405 (X-Ray Computed Tomography)</li> <li>• C0332558 (Calcified nodule)</li> <li>• C0028429 (Nose)</li> </ul>	Sagittal view of the calcified nasal packing.

subsequent discussion. These need to be detected and further applied for information retrieval purposes or image analysis. The second sub-task can be viewed as an image captioning problem. Each image has a caption and the model is tasked with generating a comparable description of the images content.

Below, we detail our observations, considerations, and experiments.

## 2. Data

The annotated dataset ROCov2[3] was provided by the ImageCLEFmedical organizers and used for both sub-tasks. For an example see Tab. 1. The training set consists of 70,108 radiology images, of which 38,603 are in color, while 31,505 are gray-scale. While the validation set comprises 9,972 radiology images, with 6,671 in color and 3,301 in grayscale, the test set consists of 17,237 radiology images. Image sizes vary, ranging from  $119 \times 160$  to  $3799 \times 5842$  pixels. For more information, see Fig. 1.

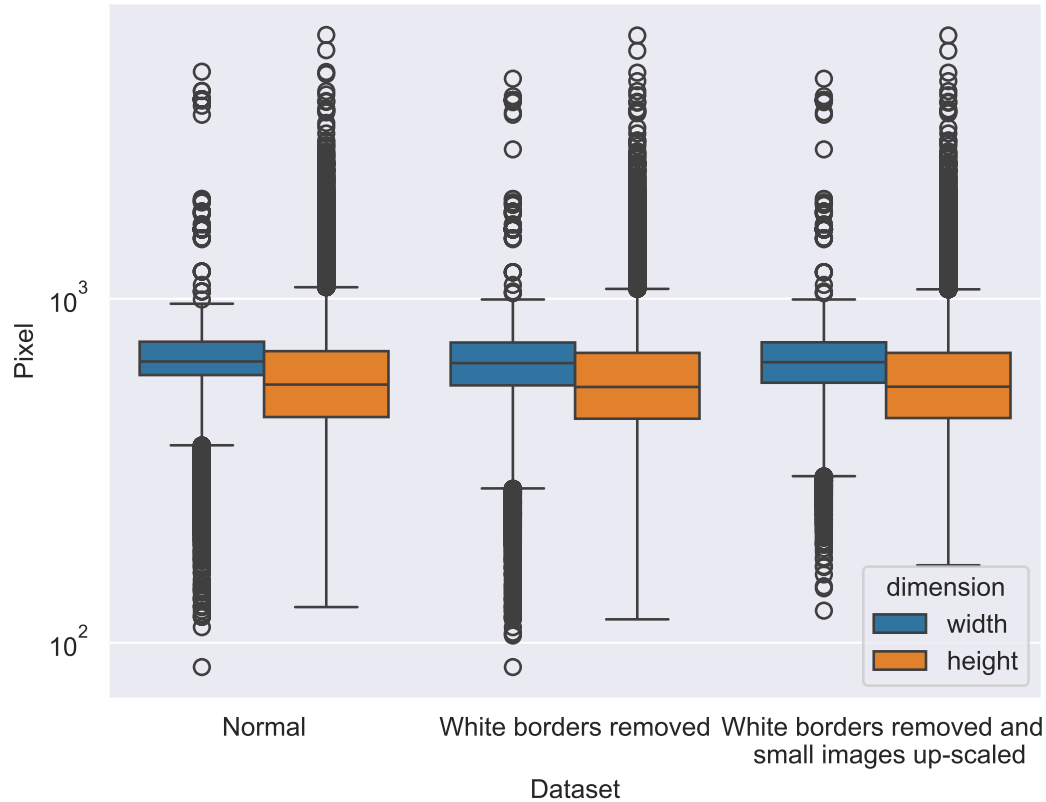
The number of unique concepts in the training dataset was reduced from 2125 to 1945 and in the validation dataset from 1945 to 1751. These were mainly concepts which were used very rarely. This was done by the organizers, because of the suggestions of last years participants. When considering the distribution of the number of labels within the images of the training dataset, it is observed that the majority of images can be assigned up to five labels, with the absolute majority having exactly two labels assigned (see Fig. 2).

The frequency of the concepts varies greatly. When combining the validation and test sets, the most frequently occurring concept is 'C0040405'/X-Ray Computed Tomography', used 27,852 times. Conversely, the least used concepts include 'C1962945'/Radiographic imaging procedure', 'C1690005'/MRI venography', 'C0243032'/Magnetic Resonance Angiography', 'C0412650'/Computed tomography of the cervical spine', 'C0011906'/Differential Diagnosis', and 'C0202657'/CT follow-up', each appearing only once (refer to Fig. 3).

As for the caption prediction task, nothing changed with the way the captions were handled in comparison to last year. The captions for the caption prediction task have already been preprocessed, resulting in the absence of any links within the captions. These captions exhibit significant variation in length. The median caption length for the training set is 17 tokens, with the largest caption containing 633 tokens and the smallest containing only one token (refer to Fig. 4).

## 3. Methodology

In this section, we will initially explore the preprocessing steps applied to the dataset, followed by an explanation of the various approaches utilized for the different tasks.

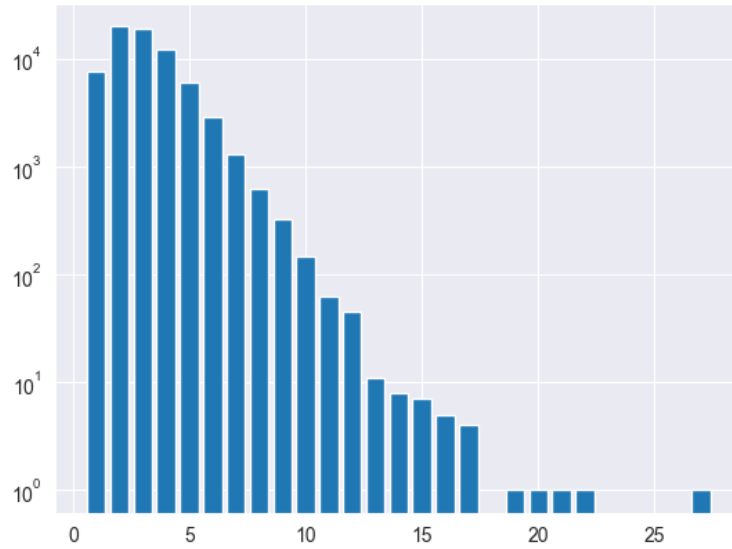


**Figure 1:** Boxplots of the pixel width and height of the train dataset

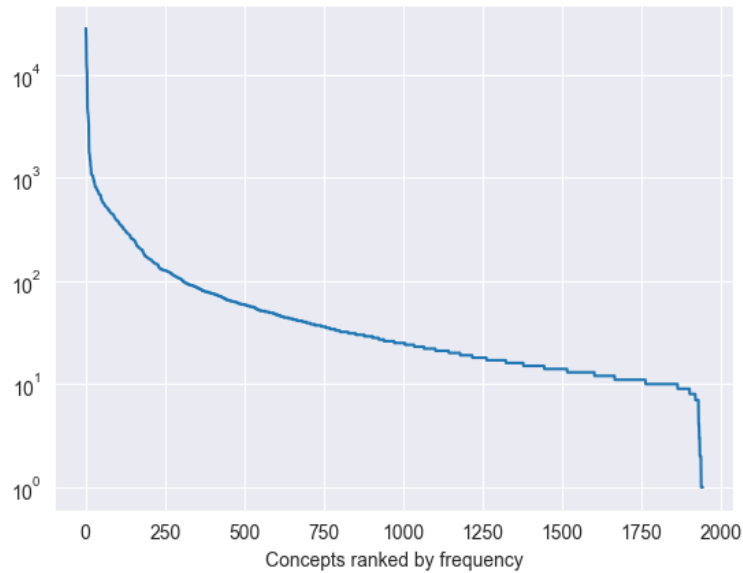
### 3.1. Preprocessing

After examining the dataset, we observed that many images feature a white border. Consequently, we decided to trim the white borders of all the images in the dataset, as we did not anticipate that our networks could extract significant information from them. Additionally, during our data analysis, we noted that there are 1251 images with dimensions less than  $300 \times 300$  pixels. For the most part, the resolution is way bigger with a mean of  $646.69 \times 593.50$  and a median of  $657 \times 563$  pixels. Experience has shown that an imbalance in sizes can negatively impact the performance of a deep learning architecture. Therefore, it was crucial to tackle this issue. This prompted us to consider leveraging a pre-trained network specialized in upscaling medical images. For this purpose, we utilized a feedback adaptive weighted dense network (FAWDN)[4]<sup>1</sup>. The architecture is visualized in 5. Since FAWDN utilizes a strict feedback mechanism, its implementation is based on recurrent neural networks (RNN) which means that the network consists of sub-networks equal to the used number of time steps. This feedback mechanism is used to produce better high-resolution images in each time step by correcting the errors from the preceding one. Another part of the feedback mechanism is that information needs to flow from the output to the input of the network. The networks consist of the input-, hidden- and output units, whose parameters are shared across time-steps. The hidden state receives the output of the previous hidden state and the current input state to enable a flow of information. A loss function is applied in every time step to make the hidden states contain information about the output image. An output image is created by adding the result of the output unit to a bilinear upsampled version of the input image. Ultimately, the image generated in the last time step is chosen as the final reconstructed high-resolution image. Another interesting aspect of the architecture is the design of the hidden unit. As previously

<sup>1</sup>code available at <https://github.com/Lihui-Chen/FAWDN>, last visited: 24.05.2024



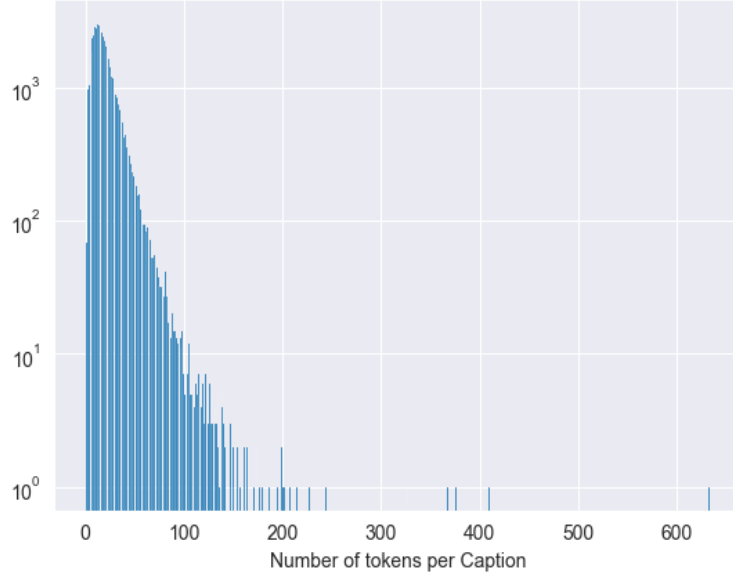
**Figure 2:** Distribution of the number of labels per image in the training dataset



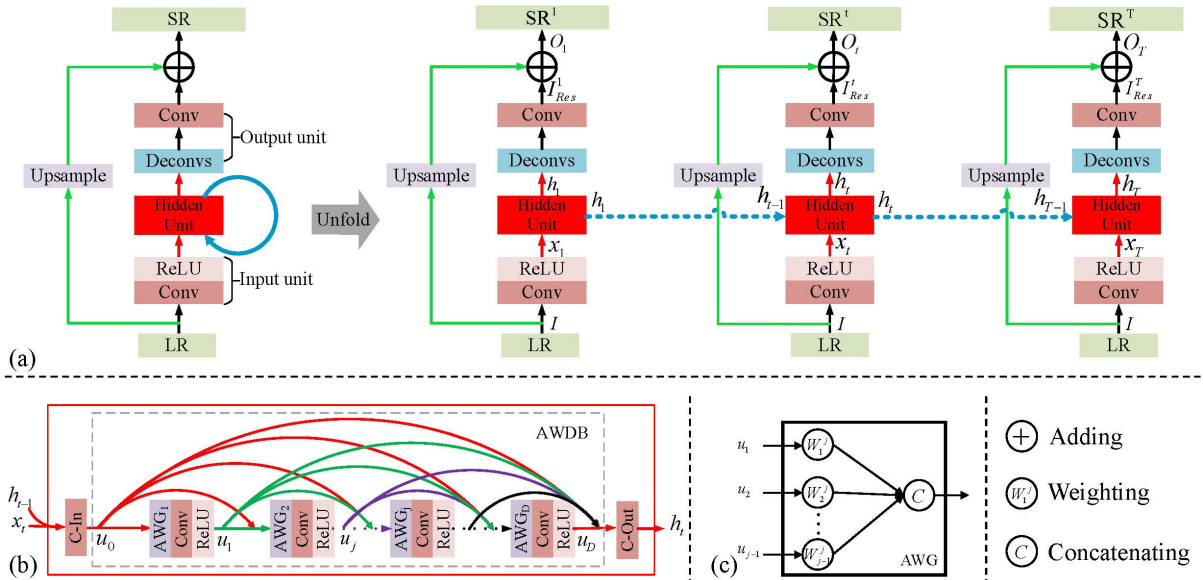
**Figure 3:** Usage frequency of concepts in the whole dataset

mentioned, it utilizes the preceding hidden state and the current input state by concatenating them, and it also includes two  $1 \times 1$  convolution layers, namely C-In and C-Out, and an adaptive weighted dense block (AWDB). In the hidden unit C-In modifies low-level features, from which the AWDB extracts more informative and advanced features. Finally, C-Out compresses the numerous features created previously. We chose FAWDN for the upscaling of the tiny images because of its convincing results presented in the original paper which demonstrated competitive performance across different datasets and resolutions. Additionally, a model trained on medical images was provided, which is crucial since we do not possess different-sized versions of the dataset to train a super-resolution network ourselves.

By applying FAWDN to the provided data, we created a new dataset in which those small images were upscaled to twice their size. Especially small images with a size of  $150 \times 150$  pixels and below were upscaled to triple their size. This would ensure that no classical upscaling methods would be needed when using a random crop size of  $224 \times 224$  for training. All of our models in both sub tasks were trained on the new dataset. One concept detection run was uploaded which used the original



**Figure 4:** Distribution of the number of tokens per caption in the training dataset

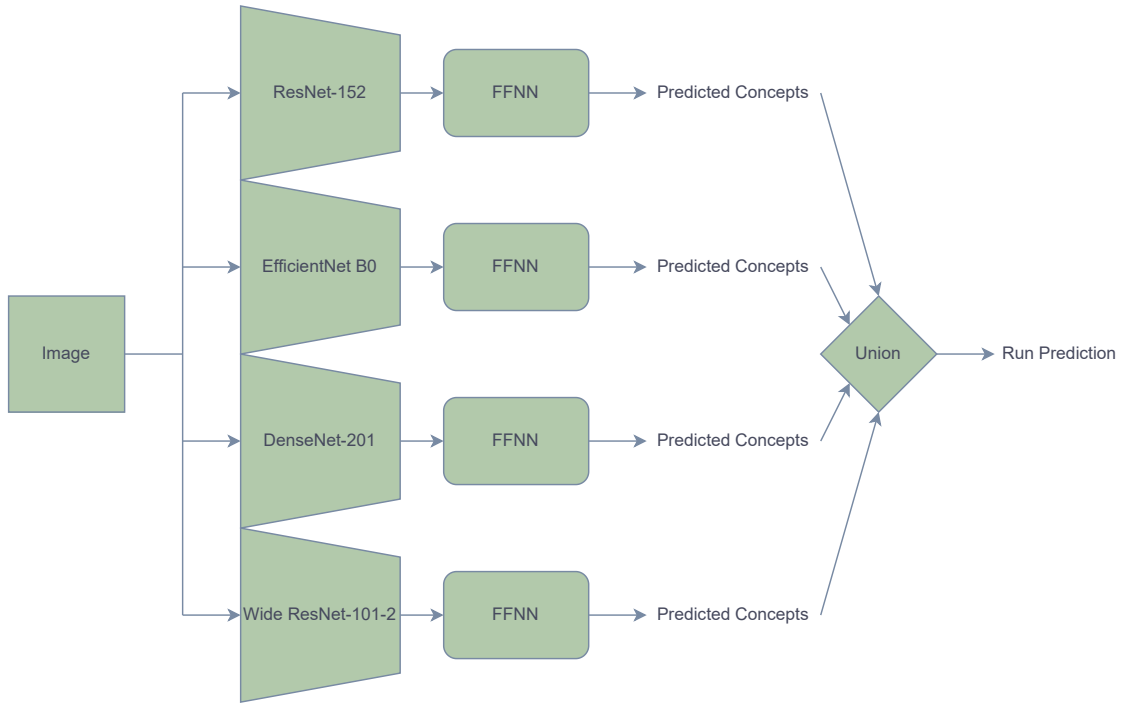


**Figure 5:** Architecture of the FAWDN[4]

dataset for comparison.

### 3.2. Concept Detection

In the last years, Convolutional Neural Networks (CNNs) proved outstanding results in multi-label classification problems. However, to get the most out of the models, ensembles are commonly created. Several studies have demonstrated the benefits of using ensemble methods for improving performance on computer vision tasks. By combining predictions from multiple models, the variance errors can be reduced, the generalization increased and the overall accuracy improved. In medical image analysis, ensemble learning helps to address the variability in annotations (caused by the inconsistency of annotators) and observer interpretations and build more robust diagnostic predictions. [5]. Finally, ensemble learning can also improve generalization across different datasets, which in particular is important while working on computer vision challenges since commonly the data originates from



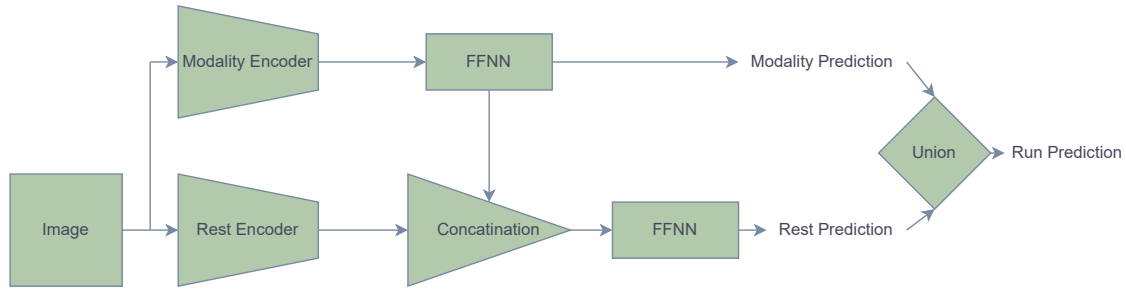
**Figure 6:** Schema of the ensemble architecture

different sources. The benefits of model ensembles also apply to this challenge. This confirms the fact that the winning team of last year relied on an ensemble model [6]. Another way to leverage the strengths of multiple modules is to build a complex model in a hierarchical way. In particular, this is often beneficial when working with imbalanced, distributed data.

In the following, we describe our two approaches the CNN ensemble and the hierarchical model.

### 3.2.1. CNN Ensemble

The ensemble we built consisted of four different CNNs: ResNet152 [5], EfficientNetB0 [6], DenseNet201 [7], and Wide ResNet-101-2 [8]. All models utilized pre-trained weights from ImageNet and were followed by different feed forward neural networks (FFNNs) composed of fully connected layers, dropout layers, and ReLU layers. We re-trained each model separately either with binary cross-entropy or multi-label soft margin loss. During training, we normalized the images with the channel-wise mean and standard deviation of the used dataset and applied a random crop with size  $224 \times 224$ , random horizontal flip with 50% probability and random rotations up to  $10^\circ$  as transformation steps. An Adam optimizer with an initial learning rate of  $1 \times 10^{-4}$  was used, but the rate was reduced when the loss reached a plateau. During training we used the validation set to monitor the F1-score so we could use the model which had the best metric score after training. The training was capped at 50 epochs, with early stopping employed to save computational time, if the validation metric did not make any changes above  $5 \times 10^{-3}$  for ten consecutive epochs. For the final prediction, we used the union of concepts predicted by each model, meaning every predicted concept was included. To properly evaluate the prediction, we included concepts that were predicted by more than one model only once. The architecture that demonstrated the most outstanding performance, achieving first place, is schematically depicted in Figure 6.



**Figure 7:** Schema of the hierarchical architecture

### 3.2.2. Hierarchical Model

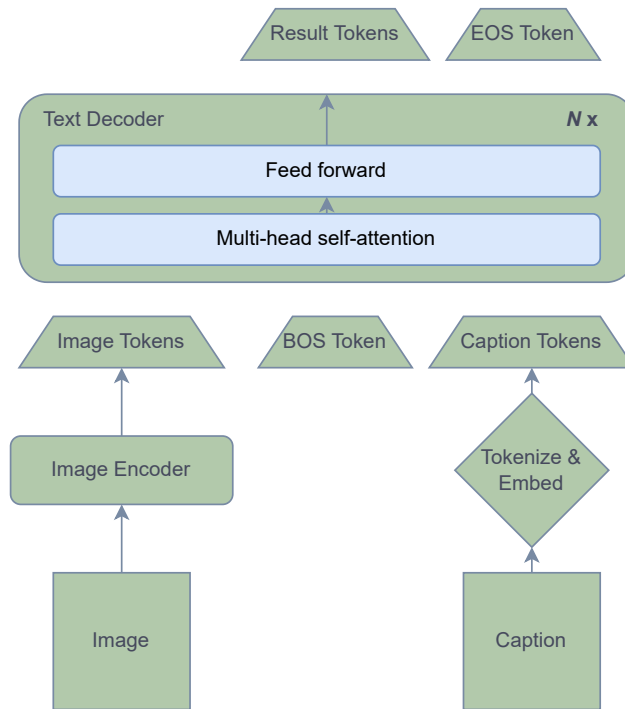
In this approach, we aimed to improve the design of last year’s CNN+FFNN-based Multi-task Classifier from the AUEB NLP Group [6] by expanding the architecture. We also hypothesized that utilizing the hierarchical relationship between concepts could lead to better results. To enhance last year’s design we used two separate backbones instead of two task-specific classification heads, as illustrated in 7. The used backbones are ResNet152 and the FFNN constructed as described in the the previous subsection. One network is responsible for predicting the image modalities and the other for the remaining concepts, with the connection that the output of the modality network is concatenated into the network for the remaining concepts before its FFNN. The modality model is trained with cross entropy loss and the other model with multi-label soft margin loss. To utilize all available images, we introduced an ‘empty’ label during training, because some images did not have a modality concept and others did not have concepts but modalities. However, we also experimented without the empty labels and discarded the images that would have been labeled as empty. The training parameters remained the same as in the previous approach. By implementing these modifications, we sought to leverage the strengths of both CNN and FFNN architectures and the hierarchical relationships between different concepts to improve the overall performance of the multi-task classifier.

### 3.3. Caption Prediction

The field of image captioning is currently dominated by transformer architectures. They are commonly known to have exceptional capabilities when it comes to handling language tasks. Last year’s competition also underlined the strength and variety of these kind of architectures. This prompted us to experiment with a new transformer architecture to examine how it would perform in the medical context and if it would yield any new significant results. The network we chose is called Generative Image-to-text Transformer (GIT)[7]. Its architecture is designed to handle both image/video captioning and visual question-answering tasks. Despite its versatile applications, GIT is fundamentally composed of an image encoder and a text decoder, as illustrated in Figure 8.

At a high level, GIT processes an image using the image encoder, transforming it into a 2D feature map that is then flattened into a list of features. An additional linear layer and a layernorm layer[8] project these image features so that they can be used as input for the text decoder. The pre-training involves first using a contrastive task to pre-train the image encoder, followed by a generation task to pre-train both the image encoder and the text decoder. The choice of the image encoder depends on the specific model variant. In the original GIT model, a Florence/CoSwin image encoder is used [9]. We experimented with the GIT-base and GIT-large variants. GIT-base employs a CLIP/ViT-B/16 encoder [10], while GIT-large uses a CLIP/ViT-L/14 encoder [10]. Another difference between these variants is the datasets used for pre-training. GIT-base is pre-trained on 10 million image-text pairs or 4 million images, sourced from a combination of COCO[11], SBU[12], CC3M[13], and VG[14] datasets. GIT-large is pre-trained on 20 million image-text pairs or 14 million images, which includes the 10 million image-text pairs from GIT-base supplemented with the CC12M[15] dataset. The text decoder is





**Figure 8:** Schema of the Generative Image-to-text Transformer architecture, derived from the original [7]

consistent across all variants and consists of a transformer module with multiple transformer blocks. Each block includes a self-attention layer and a feed-forward layer. First the text needs to be tokenized and embedded in the same number of dimensions as the image features. Then follows the addition of the positional embedding and a layernorm layer. To finalize the input for the text decoder the image features are concatenated with the text embeddings with a BOS token between them. Now the decoder can start from the BOS token and decode the next token in an auto-regressive way until the EOS token or reaching the maximum steps. The sequence-to-sequence attention mask is configured in such a way that a text token only depends on its predecessor and all image tokens, while image tokens can attend to each other. We fine-tuned the two variants in the same way, using an initial learning rate of  $5 \times 10^{-5}$  for 50 epochs. We used AdamW as optimizer with standard parameters and trained with 16-bit (mixed) precision training instead of 32-bit training. Because of the size of the model we could not evaluate the model during training which is why we used the one we obtained after the last epoch.

In our experiments, we aimed to leverage GIT's capabilities to generate meaningful and accurate medical image captions, hypothesizing that the transformer-based approach would enhance performance over traditional methods. The results of these experiments could provide insights into the applicability of advanced transformer architectures in the specialized field of medical image captioning, potentially setting a new benchmark for future research and applications.

## 4. Evaluation

In this section, we will present the results of our submissions and explain the used metrics for each sub task.



**Table 2**

Evaluation results: DBS-HHU Concept Detection Task

Affiliation	ID	$F_1$ -score (Dev)	$F_1$ -score (Test)	$F_1$ -score manual
DBS-HHU	603	0.5969	0.6375	0.9534
DBS-HHU	625	0.5928	0.6309	0.9488
DBS-HHU	604	0.5938	0.6269	0.9461
DBS-HHU	610	0.3300	0.3417	0.4477
DBS-HHU	616	0.2332	0.3413	0.4340

**Table 3**

DBS-HHU: Best run on the Caption Prediction Task

Team	BERTScore (Dev)	BERTScore (Test)	ROUGE	BLEU-1	BLEURT	METEOR	CIDEr	CLIPScore	RefCLIPScore
DBS-HHU	0.5917	0.5769	0.1531	0.1493	0.2710	0.0559	0.0644	0.7842	0.7750

## 4.1. Concept Detection

For this task, the  $F_1$ -score between the predictions and the ground truth is used as a primary evaluation metric. It is calculated by averaging over all  $F_1$ -scores for every image. The score for an image is calculated by creating multi-one hot encoded vectors for the prediction and ground truth and calculating a harmonic mean of the precision and recall. As a secondary metric, the  $F_1$ -score is calculated with a ground truth set of manually validated concepts.

We submitted three different versions of our ensemble model and two different versions of our hierarchical model. Our best model, which also won this year’s challenge, is an ensemble trained on our preprocessed dataset using a multi-label soft margin loss (ID 603). Following this, the next best was our ensemble trained on the preprocessed dataset with binary cross-entropy (BCE) loss (ID 625), and then the ensemble trained on the normal dataset with BCE loss (ID 604). Our proposed hierarchical models did not perform well. This is probably due to the way the information of the modality part of the model is fed into the model for the remaining classes. We firstly suspected that the mass of empty labels led the model (ID 610) to primarily classify the images as empty, but our run without the empty labels (ID 616) performed worse. The results can be seen in Table 2 with an additional comparison to our validation results.

## 4.2. Caption Prediction

The primary evaluation metric for this task is the BERTScore. As preprocessing for the evaluation all captions were turned into lowercase, had their punctuation removed and their numbers replaced by the token number so that the focus of the evaluation lies on the linguistic content. The metric uses the contextualized word embeddings of the Microsoft/deberta-xlarge-mnli model. The BERTScore for a single sentence is calculated by matching each token in the candidate sentence to the most similar token in the reference sentence in terms of cosine similarity, and vice versa, to compute Recall and Precision, which are then combined to calculate the  $F_1$  score. The final score is the sum of all sentence scores divided by the number of captions. Since the BERTScore is more focused on imitating human judgment the ROUGE score was used as a secondary metric. This metric is computed by comparing which n-grams can be found in one sentence in the other and vice versa. This combination of a more human-oriented and a classical metric should give a good comparison between models. Outside of the primary and secondary metrics were other metrics calculated, for further comparison, as seen in Table 3

We submitted two models for this task: a fine-tuned version of the GIT-base model and a fine-tuned version of the GIT-large model. Our best run, the GIT-large model, achieved tenth place. The performance difference between the GIT-large and GIT-base models is negligible, as indicated by a BERTScore difference of only  $1 \times 10^{-10}$ .

## 5. Conclusion

At the end of this paper, we will summarize the insights gained from our experiments and their results, and we will also propose ideas for possible future work.

### 5.1. Discussion

Starting with the concept detection sub task, even though our ensemble approach performed very well, it needs a considerable amount of resources since four different networks need to be trained. This also slows down the evaluation process since an image must pass through all four networks. While very effective, it still is a time-intensive approach. Our hierarchical model did not perform well, due to potentially not optimal network design. The information is only available in the FFNN and does not get back-propagated to the CNN, which is the reason why it does not learn a connection between the modalities and their related concepts. Nevertheless, we remain convinced that an approach in this direction has the potential to achieve good results. This conviction stems from the fact that a model that utilizes the concept of hierarchy works with more information than just images, which should confer an advantage.

As noted in the previous section, our models for the caption prediction sub task did perform the same. Since both were trained for 50 epochs, both models may be equally overfitted. Both of these models were pre-trained on a large amount of data that is not medical related which may cause them to have problems to adapt to the development dataset which is small in comparison. GIT's strength seems to lie in its versatile use cases and not in its ability to perform highly specialised tasks like medical image captioning.

### 5.2. Future Work

In the previous discussion, we highlighted the potential of hierarchical models for concept detection. A different method to transfer the information from the modality network into the the network for the remaining classes could make big improvements to the model. Another idea would be to further split up the model and use a sub-network for every modality. That means a modality network predicts the image modality. This prediction determines to which network the image is passed next so that it can predict the remaining concepts.

## References

- [1] B. Ionescu, H. Müller, A.-M. Drăgulescu, J. Rückert, A. B. Abacha, A. G. S. de Herrera, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, C. S. Schmidt, T. M. G. Pakull, H. Damm, B. Bracke, C. M. Friedrich, A.-G. Andrei, Y. Prokopcuk, D. Karpenka, A. Radzhabov, V. Kovalev, C. Macaire, D. Schwab, B. Lecouteux, E. Esperança-Rodier, W. Yim, Y. Fu, Z. Sun, M. Yetisgen, F. Xia, S. A. Hicks, M. A. Riegler, V. Thambawita, A. Storås, P. Halvorsen, M. Heinrich, J. Kiesel, M. Potthast, B. Stein, Overview of ImageCLEF 2024: Multimedia retrieval in medical applications, in: *Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the 15th International Conference of the CLEF Association (CLEF 2024)*, Springer Lecture Notes in Computer Science LNCS, Grenoble, France, 2024.
- [2] J. Rückert, A. Ben Abacha, A. G. Seco de Herrera, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, B. Bracke, H. Damm, T. M. G. Pakull, C. S. Schmidt, H. Müller, C. M. Friedrich, Overview of ImageCLEFmedical 2024 – Caption Prediction and Concept Detection, in: *CLEF2024 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org*, Grenoble, France, 2024.
- [3] J. Rückert, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, C. S. Schmidt, S. Koitka, O. Pelka, A. B. Abacha, A. G. S. de Herrera, H. Müller, P. A. Horn, F. Nensa, C. M. Friedrich, ROCov2: Radiology Objects in COntext version 2, an updated multimodal image dataset, *Scientific Data* (2024). URL: <https://arxiv.org/abs/2405.10004v1>. doi:10.1038/s41597-024-03496-6.

- [4] L. Chen, X. Yang, G. Jeon, M. Anisetti, K. Liu, A trusted medical image super-resolution method based on feedback adaptive weighted dense network, *Artificial Intelligence in Medicine* 106 (2020) 101857. URL: <https://www.sciencedirect.com/science/article/pii/S0933365719310073>. doi:<https://doi.org/10.1016/j.artmed.2020.101857>.
- [5] S. Rajaraman, S. Sornapudi, P. O. Alderson, L. R. Folio, S. K. Antani, Analyzing inter-reader variability affecting deep ensemble learning for covid-19 detection in chest radiographs, *PloS one* 15 (2020) e0242301.
- [6] P. Kaliosis, G. Moschovis, F. Charalampakos, J. Pavlopoulos, I. Androutsopoulos, AUEB NLP group at ImageCLEFmedical Caption 2023, in: M. Aliannejadi, G. Faggioli, N. Ferro, M. Vlachos (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), Thessaloniki, Greece, September 18th to 21st, 2023, volume 3497 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 1524–1548. URL: <https://ceur-ws.org/Vol-3497/paper-126.pdf>.
- [7] J. Wang, Z. Yang, X. Hu, L. Li, K. Lin, Z. Gan, Z. Liu, C. Liu, L. Wang, GIT: A generative image-to-text transformer for vision and language, *Transactions on Machine Learning Research* (2022). URL: <https://openreview.net/forum?id=b4tMhpN0JC>.
- [8] J. L. Ba, J. R. Kiros, G. E. Hinton, Layer normalization, 2016. [arXiv:1607.06450](https://arxiv.org/abs/1607.06450).
- [9] L. Yuan, D. Chen, Y.-L. Chen, N. Codella, X. Dai, J. Gao, H. Hu, X. Huang, B. Li, C. Li, C. Liu, M. Liu, Z. Liu, Y. Lu, Y. Shi, L. Wang, J. Wang, B. Xiao, Z. Xiao, J. Yang, M. Zeng, L. Zhou, P. Zhang, Florence: A new foundation model for computer vision, 2021. [arXiv:2111.11432](https://arxiv.org/abs/2111.11432).
- [10] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, in: *International Conference on Machine Learning*, 2021, pp. 8748 – 8763.
- [11] T.-Y. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft COCO: Common objects in context, in: *European Conference on Computer Vision*, 2014, pp. 740 – 755. doi:[10.1007/978-3-319-10602-1\\_48](https://doi.org/10.1007/978-3-319-10602-1_48).
- [12] V. Ordonez, G. Kulkarni, T. Berg, Im2text: Describing images using 1 million captioned photographs, in: J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, K. Weinberger (Eds.), *Advances in Neural Information Processing Systems*, volume 24, Curran Associates, Inc., 2011. URL: [https://proceedings.neurips.cc/paper\\_files/paper/2011/file/5dd9db5e033da9c6fb5ba83c7a7e9-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2011/file/5dd9db5e033da9c6fb5ba83c7a7e9-Paper.pdf).
- [13] P. Sharma, N. Ding, S. Goodman, R. Soricut, Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning, in: *Annual Meeting of the Association for Computational Linguistics*, 2018. URL: <https://api.semanticscholar.org/CorpusID:51876975>.
- [14] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. S. Bernstein, L. Fei-Fei, Visual genome: Connecting language and vision using crowdsourced dense image annotations, *International Journal of Computer Vision* 123 (2016) 32 – 73. doi:[10.1007/s11263-016-0981-7](https://doi.org/10.1007/s11263-016-0981-7).
- [15] S. Changpinyo, P. K. Sharma, N. Ding, R. Soricut, Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts, 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021) 3557–3567. URL: <https://api.semanticscholar.org/CorpusID:231951742>.