

VisualT5: Multitasking Caption and Concept Prediction with Pre-trained ViT, T5 and Customized Spatial Attention in Radiological Images

Notebook for the Medical Image Computing Lab at CLEF 2024

Diedre Carmo^{1,*}, Letícia Rittner¹ and Roberto Lotufo¹

¹*School of Electrical and Computer Engineering, Universidade Estadual de Campinas, Campinas, Brazil*

Abstract

The development of more explainable and general deep learning-based predictive and generative models is of interest to the medical imaging processing field, largely due to the “black box” and often specialized nature of current models. This paper describes our participation in the ImageCLEF Caption Prediction and Concept Detection challenges with a multitasking, multimodal and explainable architecture named VisualT5. VisualT5 couples the embedding power of a frozen pre-trained Vision Transformer (ViT) with the clinical text generation capabilities of the pre-trained ClinicalT5. Moreover, we propose a modified spatial attention module that weights our visual encoder features in the token dimension, showcasing the spatial importance of each ViT token and permitting more interpretability regarding what parts of the image have more impact on the model’s conclusions. VisualT5-base-clinical as a single multitasking model achieved 0.61 BERTScore and 0.58 F1-score in the caption prediction and concept detection tasks, respectively, ranking 6/11 in the caption leaderboard and 6/9 in the concept leaderboard.

Keywords

vision transformer, t5, image captioning, image classification, medical imaging

1. Introduction

The success of deep learning for the creation of predictive and generative models is evident [1, 2], with success both in academic research and recently being integrated into real products such as ChatGPT [3] and other platform based LLMs [4]. Deep learning models have also been applied to medical imaging classification and caption generation [5]. However, the translation of such models to real applications in medicine is lagging behind, due to the complex nature of medical diagnosis and related signal processing. Some research has raised the potential problems of bias and other factors leading to the unfeasibility of translation to real clinic of many deep learning based methods [6, 7]. Medical information that leads to a diagnosis or disease understanding is presented in many modalities, either as different types of images acquisitions, structured and free text, and even 1D signals such as electrocardiograms. Moreover, the number of tasks involved in the pipeline of medical processes can’t be summarized into isolated academic tasks such as direct image classification, segmentation, or caption generation. Finally, explainability of key factors that led to decision making is paramount in the medical field [8]. This context has led current research into considering multimodality [9], multitasking [10], and explainability [11] as important aspects of automated medical imaging processing.

In terms of model architecture, current approaches for medical imaging classification mostly consist of using CNNs with fully connected layers or the vision transformer model, an state-of-the-art transformer for image classification [12]. In the context of image to caption generation, three methodologies are commonly used: encoder-decoder models, where an encoder generates image features which are decoded into text either by LSTMs or transformers [5]; visual language models, where transformer input

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

*Corresponding author.

✉ d211492@dac.unicamp.br (D. Carmo)

🌐 <https://miclab.fee.unicamp.br/> (D. Carmo)

🆔 0000-0002-5922-9120 (D. Carmo); 0000-0001-8182-5554 (L. Rittner); 0000-0002-5652-0852 (R. Lotufo)

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



tokens mix ViT-like image representations with text tokens [13]; and finally CLIP-like approaches, where image and text embeddings are aligned, and embedding alignment after training is used to perform various multimodal tasks [14].

In this preliminary work, we explore the multitasking of medical image classification and caption generation in various modalities of radiological images from two ImageCLEF [15] challenges at the same time: the medical imaging caption prediction and concept detection challenges [16]. Our participation in this challenge happens as a first step into exploring multitasking, multimodality and explainability in medical imaging processing for better generalization and usability in practice. Our proposal involves an encoder-decoder model marrying strong image representation from pre-trained ViT models together with pre-trained T5 as a text decoder for caption generation, including innovative uses of spatial attention to promote visual explainability.

2. Methodology

The proposed VisualT5 is an image-to-text encoder-decoder architecture coupling a vision transformer with an encoder-decoder T5 text transformer. VisualT5 is trained and evaluated using the ImageCLEF dataset, ROCov2.

2.1. Dataset

Radiology Objects in COntext (ROCOv2) [17] is the main dataset used by both ImageCLEF challenges for caption prediction and concept detection labels. In summary, the authors of the dataset performed a semi-automatic pipeline to extract valid caption and radiological image pairs from publicly available medical papers. In this year’s version of the dataset, the training set consists of 70108 radiology images, with 9972 more for validation and 17237 for testing, with testing labels hidden from the participants.

Concepts classification is multilabel, and the main label used as primary groundtruth in the challenge uses concepts automatically extracted from captions, represented by 1934 Unified Medical Language System [18] Concept Unique Identifiers (CUIs). In addition, concepts are reduced into a manually curated subset containing only modality and body region CUIs for a secondary evaluation.

2.2. Architecture

In VisualT5 (Fig. 1), the frozen pre-trained ViT encoder from MEDSam [19, 21] is used to generate latent representations. To use their ViT-base [12] architecture, images are bilinearly interpolated to 1024x1024 while keeping the aspect ratio with 0 padding, by using their provided image processing pipeline¹. The resulting embedding with batch size 1 of [1, 4096, 768] reveals a hidden size of 768 and a 16x16 patch size, given that the sequence length is 4096, the number of 16x16 patches that fit in a 1024x1024 image. The last hidden state of same shape is used as a latent representation and weighted by a modified spatial attention mechanism. Instead of using convolutional layers as in Górriz et al. [20]’s 2D spatial attention, multiple linear layers with bias and LeakyReLU non-linear activations are used in the same fashion to compress the 768 hidden size into a single channel array of 4096 sigmoid activated values. Given that each of the 4096 values corresponds to one of the 64x64 patches, these values are used to weight (multiply) the contribution of each token, i.e, the importance of each region of the input image. These 4096 values can be visualized as a heatmap after a reshaping to 64x64 and bilinear interpolation to 1024x1024. Finally, the weighted latent space is used as visual encoder features for the subsequent tasks.

For concept detection, the visual encoder features are averaged in the sequence dimension and we train a projection through a linear layer into 1934 sigmoid activated neurons for multilabel concept detection, with each output neuron representing a CUI. The corresponding CUI strings are included in the prediction based on a multilabel activation threshold of 0.5. At the same time, for caption prediction,

¹<https://huggingface.co/flaviagammarino/medsam-vit-base>

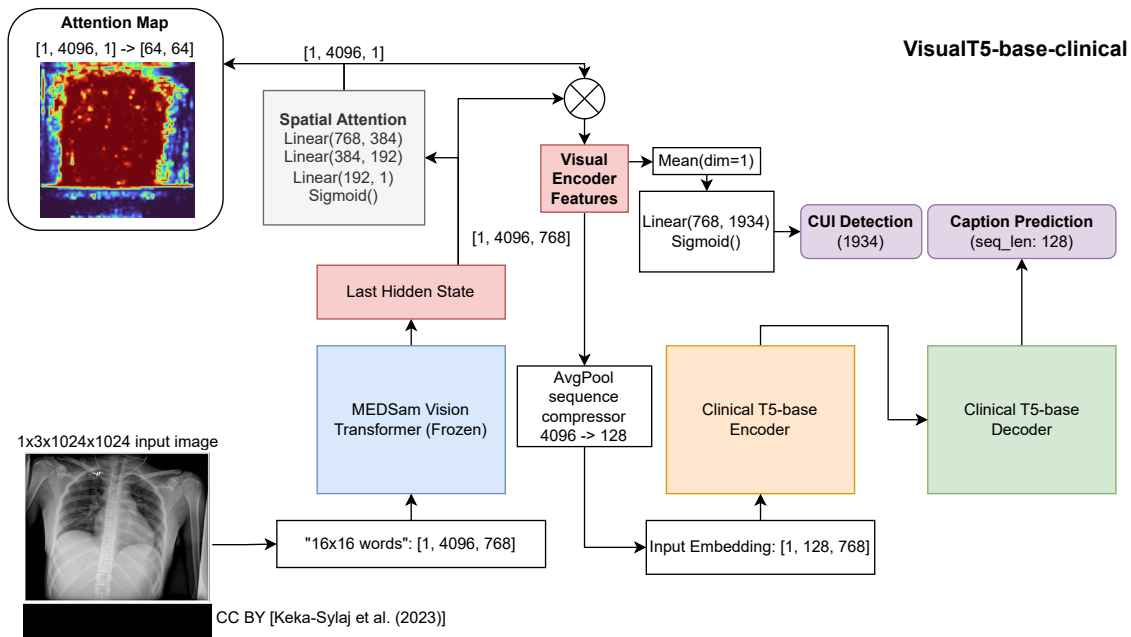


Figure 1: Overview of our final and best-performing architecture, VisualT5-base-clinical. The Vision Encoder is frozen with MEDSam’s [19] ViT-base weights. It’s last hidden state is weighted by our implementation of spatial attention [20] and used as visual encoder features, where both concept detection and caption prediction is derived. The T5 text decoder is initialized with Clinical T5 weights.

pre-trained ClinicalT5-base [22, 23] is used, including its text encoder, decoder and tokenizer. Note that the text encoder and decoder are not frozen, and are adjusted by our training. Our visual encoder features replace the input embeddings of ClinicalT5’s text encoder. We reduce the sequence length from 4096 to 128 using average pooling, due to limited GPU memory. To continue training and promoting the alignment of our visual encoder features as input embeddings for ClinicalT5, we follow original T5’s training procedure [24]. Text generation during evaluation consists of the computation of the visual encoder features once, and a Seq2Seq greedy decoding strategy for text generation with a 128 tokens maximum sequence length. Attempts at using only the ClinicalT5 decoder with visual encoder features as "encoder outputs" for T5’s encoder-decoder attention resulted in degraded quantitative performance with little computational efficiency benefits. With the multilabel concept detection and generated caption prediction being derived from the same visual encoder features, VisualT5 multitasks both tasks using the same model, and is trained in both tasks simultaneously.

2.3. Implementation Details

For implementation we used the Hugging Face Transformers library, PyTorch [25] and PyTorch Lightning [26]. MEDSam’s ViT pre-trained weights were also sourced from Hugging Face¹. Note that acquiring ClinicalT5-base weights for T5-base weight initialization required credentialing and ethical training through the PhysioNet platform [23]².

For validation, we employed the evaluation code provided by ImageCLEF organizers, which computes BERT [27] and ROUGE [28] scores for caption prediction and multilabel F1-scores for concept detection. BERTScore and F1-score for all provided concepts were the primary metrics used by challenge runners for ranking. During their test evaluation, they added and reported on additional metrics [16]. Both tasks are optimized at the same time using a 4090 24 GB GPU, with a batch size of 5, AdamW optimizer with 1e-5 initial learning rate and 1e-5 weight decay, and training for 100 epochs with an early stopping patience of 10 epochs since last validation BERTScore improvement.

²<https://physionet.org/content/clinical-t5/1.0.0/>

3. Results and Discussion

After early experiments defining some hyperparameters, four main experiments were performed and submitted to the ImageCLEF evaluation platform for testing. Results from the test phase were only revealed after the end of the challenge. These experiments aimed to evaluate the impact of the design variations over the previously described architecture (Tab. 1).

Table 1

Description of the four main experiments submitted to the challenge test phase.

Model	Description
VisualT5-small	Use a custom, not frozen ViT-small from scratch, and fine-tune the original T5-small pre-trained language model [24].
VisualT5-small-cls	Add a CLS token to ViT-small to derive concept results instead of linear projection of the average of visual encoder features in the sequence dimension.
VisualT5-base	Frozen pre-trained ViT-base from MEDSam [19] and fine-tune the original T5-base pre-trained language model [24].
VisualT5-base-clinical	Frozen pre-trained ViT from MEDSam [19] and fine tune pre-trained ClinicalT5-base [22].

Since ViT-small is not defined in the original ViT publication [12], we designed it with 512 hidden size, image size of 256x256, patch size of 16x16, 8 heads and layers, and 1024 MLP dimensions. With these parameters, ViT-small analyzes 256 tokens (patches), providing full input embedding alignment with a 256 sequence length T5-small, without the necessity of sequence length compression through average pooling. VisualT5-small trains the ViT-small visual encoder from scratch, in contrast with VisualT5-base where the pre-trained ViT is kept frozen due to memory limitations. Experimental results showcase the variations in performance resulting from these differences in VisualT5 design (Tab. 2).

Table 2

Primary metrics for caption prediction (BERTScore), and multilabel concept detection (F1-score) for each proposed multitasking VisualT5 model variation, with the respective Run IDs for the submissions on the challenge platform.

Run ID		Model	Validation		Test	
Caption	Concept	VisualT5	BERTScore	F1-score	BERTScore	F1-score
274	275	VisualT5-small	0.61	0.52	0.59	0.53
676	679	VisualT5-small-cls	0.61	0.50	0.61	0.53
677	680	VisualT5-base	0.61	0.52	0.37	0.56
678	681	VisualT5-base-clinical	0.61	0.54	0.61	0.58

It is noticeable that caption prediction performance did not change significantly during validation according to BERTScore. Using a CLS token strategy for concept detection resulted in the worst F1-score in validation, with the full VisualT5-base-clinical method being the best overall. This also translated to testing computed by the challenge runners, where the full base models with related pre-trained weights performed best. Of notice is the apparent lack of generalization to the test set of VisualT5-base, which used a general T5-base text decoder. This overfitting did not happen when training from the ClinicalT5-base text decoder weights, suggesting using pre-trained encoders and decoders from the medical domain is beneficial. In the overall test leaderboard [16], our multitask method placed 6/9 in concept detection and 6/11 in caption prediction.

In addition to quantitative performance, qualitative evaluation through random visual inspection of around a hundred test cases reveals that the model can ascertain modality and anatomical information well in the generated captions and concepts. However, the model is often unable to predict associated symptoms and diagnostic-related details, which are sometimes present in the target. Those are commonly related to clinical context or the reason for the examination, information outside of the image scope

(Fig. 2). We believe including more clinical information such as the reason for the image acquisition as input to these types of methods would lead to improved performance in these tasks.

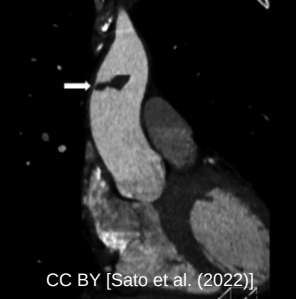
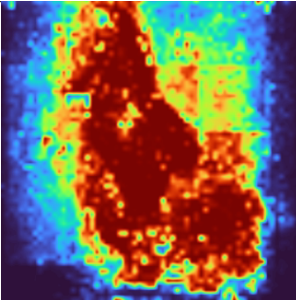
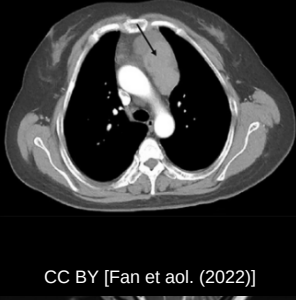
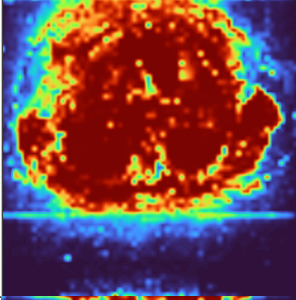

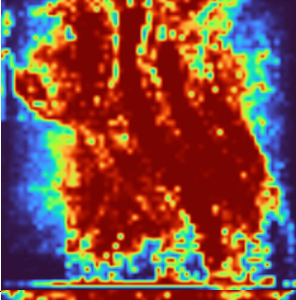
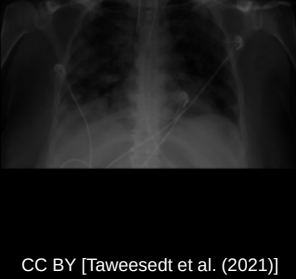
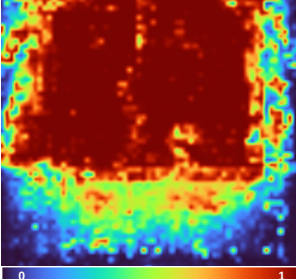
Input	Target	Prediction	Spatial Attention
 <p>CC BY [Sato et al. (2022)]</p>	<p>Computed tomography (CT) shows floating thrombosis (white arrow)</p> <p>Concept: ['C0040405', 'C0040053'], ['X-Ray Computed Tomography', 'Thrombosis']</p>	<p>Aortic root CT angiogram showing aortic root dilation.</p> <p>Concept: ['C0040405'], ['X-Ray Computed Tomography']</p>	
 <p>CC BY [Fan et al. (2022)]</p>	<p>Enhanced CT scan of the chest revealed an anterior mediastinal tumor (black arrow).</p> <p>Concept: ['C0025066', 'C0027651', 'C0040405'], ['Mediastinum', 'Neoplasms', 'X-Ray Computed Tomography']</p>	<p>Axial CT scan of the chest showing a large mass in the anterior mediastinum.</p> <p>Concept: ['C0040405'], ['X-Ray Computed Tomography']</p>	
 <p>CC BY [Trowbridge et al. (2022)]</p>	<p>Early sagittal T2-weighted MRI.</p> <p>Concept: ['C0024485'], ['Magnetic Resonance Imaging']</p>	<p>Sagittal T2-weighted MRI of the cervical spine showing a hyperintense signal in the spinal cord at the C3-C4 level (arrow).</p> <p>Concept: ['C0024485'], ['Magnetic Resonance Imaging']</p>	
 <p>CC BY [Taweeseedt et al. (2021)]</p>	<p>The typical chest X-ray finding of a patient with coronavirus disease 2019 infection showing bilateral infiltration.</p> <p>Concept: ['C0332448', 'C1306645', 'C0817096', 'C0009450'], ['Infiltration', 'Plain x-ray', 'Chest', 'Communicable Diseases']</p>	<p>Chest X-ray showing bilateral infiltrates</p> <p>Concept: ['C1306645;C0817096'], ['Plain x-ray;Chest']</p>	

Figure 2: Validation samples with target, prediction, and spatial attention from VisualT5-base-clinical.

The proposed spatial attention scheme seems to work well empirically, when rendering the generated 4096 sigmoid weights as heatmaps using the Turbo colormap (Fig 3). The ViT tokens related to foreground parts of the image are being weighted more than background regions. This type of layer has the potential to improve the readability of ViT-derived transformers, which are notable for having difficult to visualize output attentions [29]. Note, however, that there is no specific highlight of the abnormal region. Our spatial attention seems to converge to a state where most foreground tokens are “important”, with values close to 1. More exploration of this type of module in future work might lead to improved contrast and more specific indication of abnormality localization on the generated heatmaps. Possibilities include experimenting with different activations and colormaps for visualization.

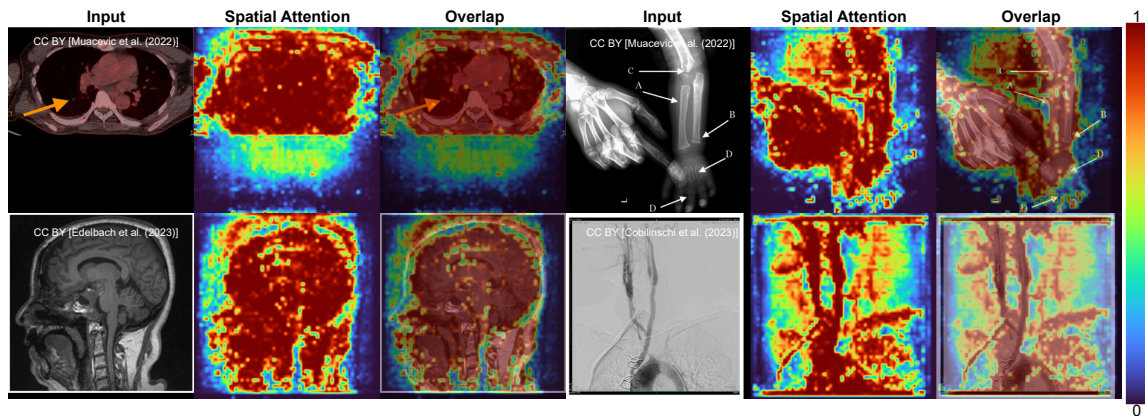


Figure 3: Some selected VisualT5-base-clinical test outputs showcasing the highlight of the most important tokens by our proposed spatial attention.

4. Conclusion

We proposed VisualT5, an encoder-decoder model based on coupling pre-trained Vision Transformers with pre-trained T5 transformers. Better performance in multitasking the ImageCLEF Caption Prediction and Concept Detection tasks was observed when using models pre-trained on the medical domain. The same multitasking weight placed in the middle of the leaderboard for both tasks in the challenge’s test phase. Moreover, the proposed modified spatial attention successfully highlighted areas of medical interest. Future work will experiment with more general promptable visual language models including prior information outside of the scope of the radiological acquisition, adding more tasks and modalities, towards a lightweight open-source, multitasking, multimodal, and explainable model.

Acknowledgments

D. Carmo was partially supported by Sao Paulo Research Foundation (FAPESP) grant #2019/21964-4. R Lotufo is partially supported by CNPq (The Brazilian National Council for Scientific and Technological Development) under grant 313047/2022-7. L Rittner is partially supported by CNPQ grant 317133/2023-3, and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) grant 506728/2020-00.

References

- [1] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, *Nature* 521 (2015) 436–444.
- [2] S. Feuerriegel, J. Hartmann, C. Janiesch, P. Zschech, Generative ai, *Business & Information Systems Engineering* 66 (2024) 111–126.
- [3] OpenAI, Chatgpt, 2024. URL: <https://chat.openai.com/chat>, accessed: 2024-06-18.
- [4] R. Pires, H. Abonizio, T. S. Almeida, R. Nogueira, Sabiá: Portuguese large language models, in: *Brazilian Conference on Intelligent Systems*, Springer, 2023, pp. 226–240.
- [5] D.-R. Beddiar, M. Oussalah, T. Seppanen, Automatic captioning for medical imaging (mic): a rapid review of literature, *Artificial Intelligence Review* 56 (2023) 4019–4076.
- [6] L. Wynants, B. Van Calster, G. S. Collins, R. D. Riley, G. Heinze, E. Schuit, E. Albu, B. Arshi, V. Bellou, M. M. Bonten, et al., Prediction models for diagnosis and prognosis of COVID-19: systematic review and critical appraisal, *bmj* 369 (2020).
- [7] M. Roberts, D. Driggs, M. Thorpe, J. Gilbey, M. Yeung, S. Ursprung, A. I. Aviles-Rivero, C. Etmann, C. McCague, L. Beer, et al., Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans, *Nature Machine Intelligence* 3 (2021) 199–217.

- [8] N. Burkart, M. F. Huber, A survey on the explainability of supervised machine learning, *Journal of Artificial Intelligence Research* 70 (2021) 245–317.
- [9] L. Heiliger, A. Sekuboyina, B. Menze, J. Egger, J. Kleesiek, Beyond medical imaging—a review of multimodal deep learning in radiology, *Authorea Preprints* (2023).
- [10] Y. Zhao, X. Wang, T. Che, G. Bao, S. Li, Multi-task deep learning for medical image computing and analysis: A review, *Computers in Biology and Medicine* 153 (2023) 106496.
- [11] T. Dhar, N. Dey, S. Borra, R. S. Sherratt, Challenges of deep learning in medical image analysis—improving explainability and trust, *IEEE Transactions on Technology and Society* 4 (2023) 68–75.
- [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, *International Conference on Learning Representations abs/2010.11929* (2020).
- [13] K. Zhang, J. Yu, Z. Yan, Y. Liu, E. Adhikarla, S. Fu, X. Chen, C. Chen, Y. Zhou, X. Li, et al., Biomedgpt: A unified and generalist biomedical generative pre-trained transformer for vision, language, and multimodal tasks, *arXiv preprint arXiv:2305.17100* (2023).
- [14] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: *International conference on machine learning*, PMLR, 2021, pp. 8748–8763.
- [15] B. Ionescu, H. Müller, A. Drăgulescu, J. Rückert, A. Ben Abacha, A. García Seco 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. Andrei, Y. Prokopchuk, 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.
- [16] 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*.
- [17] 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.
- [18] O. Bodenreider, The unified medical language system (umls): integrating biomedical terminology, *Nucleic acids research* 32 (2004) D267–D270.
- [19] J. Ma, Y. He, F. Li, L. Han, C. You, B. Wang, Segment anything in medical images, *Nature Communications* 15 (2024) 654.
- [20] M. Górriz, J. Antony, K. McGuinness, X. Giró-i Nieto, N. E. O’Connor, Assessing knee OA severity with CNN attention-based end-to-end architectures, in: *International conference on medical imaging with deep learning*, PMLR, 2019, pp. 197–214.
- [21] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, et al., Segment anything, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4015–4026.
- [22] E. Hernandez, D. Mahajan, J. Wulff, M. J. Smith, Z. Ziegler, D. Nadler, P. Szolovits, A. Johnson, E. Alsentzer, et al., Do we still need clinical language models?, in: *Conference on Health, Inference, and Learning*, PMLR, 2023, pp. 578–597.
- [23] E. Lehman, A. Johnson, Clinical-t5: Large language models built using mimic clinical text, *PhysioNet* (2023).
- [24] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring

the limits of transfer learning with a unified text-to-text transformer, *Journal of machine learning research* 21 (2020) 1–67.

- [25] J. Ansel, E. Yang, H. He, N. Gimelshein, A. Jain, M. Voznesensky, B. Bao, P. Bell, D. Berard, E. Burovski, G. Chauhan, A. Chourdia, W. Constable, A. Desmaison, Z. DeVito, E. Ellison, W. Feng, J. Gong, M. Gschwind, B. Hirsh, S. Huang, K. Kalambarkar, L. Kirsch, M. Lazos, M. Lezcano, Y. Liang, J. Liang, Y. Lu, C. Luk, B. Maher, Y. Pan, C. Puhersch, M. Reso, M. Saroufim, M. Y. Siraichi, H. Suk, M. Suo, P. Tillet, E. Wang, X. Wang, W. Wen, S. Zhang, X. Zhao, K. Zhou, R. Zou, A. Mathews, G. Chanan, P. Wu, S. Chintala, *PyTorch 2: Faster Machine Learning Through Dynamic Python Bytecode Transformation and Graph Compilation*, in: *29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2 (ASPLOS '24)*, ACM, 2024. URL: <https://pytorch.org/assets/pytorch2-2.pdf>. doi:10.1145/3620665.3640366.
- [26] W. Falcon, The PyTorch Lightning team, *PyTorch Lightning*, 2024. URL: <https://github.com/Lightning-AI/lightning>. doi:10.5281/zenodo.10779019.
- [27] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, *BERTScore: Evaluating text generation with BERT*, *International Conference on Learning Representations* abs/1904.09675 (2019).
- [28] L. Chin-Yew, *Rouge: A package for automatic evaluation of summaries*, in: *Proceedings of the Workshop on Text Summarization Branches Out*, 2004, 2004.
- [29] T. Darcet, M. Oquab, J. Mairal, P. Bojanowski, *Vision transformers need registers*, *arXiv preprint arXiv:2309.16588* (2023).