UMUTeam at ImageCLEF 2025: Fine-Tuning a Vision-Language Model for Medical Image Captioning and SapBERT-Based Reranking for Concept Detection

Notebook for the ImageCLEF Lab at CLEF 2025

Ronghao Pan[1], Tomás Bernal-Beltrán[1], José Antonio García-Díaz[1,*] and Rafael Valencia-García[1]

[1]*Facultad de Informática, Universidad de Murcia, Campus de Espinardo, 30100, Spain*

Abstract

In this paper, we present the participation of UMUTeam in the ImageCLEFmed Caption 2025 challenge. We participated in two subtasks: caption prediction and concept detection, for which we used a two-stage vision-language system. For caption prediction, we fine-tuned the BLIP model on a large-scale radiology dataset, optimizing with a composite relevance metric that combines BERTScore, ROUGE-1, and exact match similarity. Our approach achieved the highest overall score in this subtask, ranking first with a final score of 0.3771. For concept detection, we implemented a hybrid pipeline combining named entity recognition based on SciSpacy and SapBERT-based embedding retrieval with a BERT-based sequence classifier for reranking and filtering UMLS concept candidates. Although the semantic retrieval component ensured high recall, false positives caused by class imbalance and entity ambiguity limited the system's performance, resulting in a final F1-score of 0.2398.

Keywords

Language-Vision Model, Natural Language Processing, Medical Image Captioning, Medical Entity Linking,

1. Introduction

Interpreting medical images is a central task in clinical diagnostic and follow-up workflows. It remains one of the most time-consuming and resource-intensive tasks in healthcare. Imaging modalities such as X-rays, computed tomography (CT), and magnetic resonance imaging (MRI) produce large volumes of visual data that must be carefully examined by trained specialists. This process demands extensive medical training and clinical expertise, making it both costly and labor-intensive. Several studies have shown that manual analysis of medical images significantly slows clinical workflows, affecting multiple stages, from initial screening to generating final diagnostic reports [1, 2].

In this context, automatic systems designed to support medical image interpretation have emerged as a promising solution for enhancing clinical efficiency. In particular, the development of models based on Machine Learning (ML) and Deep Learning (DL) techniques capable of generating informative and clinically relevant textual descriptions from visual data, commonly referred to as medical image captioning, has attracted increasing attention in recent years. Unlike traditional image captioning in general computer vision, generating captions for medical images requires the integration of structured biomedical knowledge and a high degree of semantic accuracy, since even minor errors can have serious consequences [2]. This challenge is addressed by the ImageCLEFmed Caption task [3], a dedicated subtask of the broader ImageCLEF initiative [4], which is held annually as part of the Conference and Labs of the Evaluation Forum (CLEF).

The ImageCLEFmed Caption task directly addresses these challenges by proposing a benchmarking framework to evaluate automatic systems that associate medical images with semantically meaningful

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

*Corresponding author

ronghao.pan@um.es (R. Pan); tomas.bernalb@um.es (T. Bernal-Beltrán); joseantonio.garcia8@um.es (J. A. García-Díaz); valencia@um.es (R. Valencia-García)

0009-0008-7317-7145 (R. Pan); 0009-0006-6971-1435 (T. Bernal-Beltrán); 0000-0002-3651-2660 (J. A. García-Díaz); 0000-0003-2457-1791 (R. Valencia-García)

textual output. The task has evolved significantly over the years, incorporating lessons learned and responding to community feedback. In the first two editions, held in 2017 and 2018, the dataset featured a wide variety of medical content and imaging contexts. In 2019, the training set was restricted exclusively to radiology images, and in 2020, modality-specific metadata was introduced to support pre-processing and multi-modal approaches. The 2021 edition focused on enhancing the clinical relevance of the data by using real radiology images annotated by medical professionals. However, acquiring large numbers of such high-quality images remains a challenge. In 2022, an extended version of the 2020 dataset was adopted. New evaluation metrics were introduced for the caption prediction subtask, with the aim of improving evaluation fidelity. By 2023, the dataset had undergone several improvements, addressing issues such as excessive concept granularity, lemmatization errors, and duplicate captions. Based on the results of those editions, BERTScore was adopted as the primary evaluation metric for caption prediction.

The 2025 edition of the ImageCLEFmed Caption task includes the following subtasks:

- **Concept Detection Task**: This subtask focuses on identifying relevant medical concepts based on the visual content of images. It forms the foundation for understanding image scenes by detecting the components from which captions are composed. Detected concepts can also support downstream tasks such as context-aware image retrieval. Evaluation is performed using standard set coverage metrics, including precision, recall, and F1-score.
- **Caption Prediction Task**: In this subtask, the objective is to generate full, coherent captions for each image, using both the visual information and the detected concept vocabulary. In this case, BERTScore will be the primary evaluation metric, complemented by ROUGE as a secondary metric. Additional metrics such as MedBERTScore, MedBLEURT, and BLEU will also be reported.

For this edition, an automatic system was proposed by our research group for both caption prediction and concept detection. For the captioning subtask, the BLIP model[1] [5] was fine-tuned on the official training dataset, which consists of image–caption pairs derived from PubMed Central articles. During training, a composite relevance metric combining BERTScore, ROUGE, and lexical similarity was employed to guide model selection and evaluate semantic fidelity.

For the concept detection subtask, biomedical entities were first extracted from the generated captions using SciSpacy [6]. These entities were then normalized to UMLS Concept Unique Identifiers (CUIs) using semantic similarity based on SapBERT embeddings [7]. To improve precision, a reranking mechanism was introduced based on a BERT-based sequence classification model, fine-tuned to distinguish valid entity–concept pairs from unrelated matches. This model served as a semantic filter, refining the top-n CUI candidates by predicting their relevance to the original entity mention. The reranker helped reduce false positives and increased the accuracy of concept detection by ensuring that only clinically meaningful matches were retained.

These working notes are structured as follows. Section 2 provides a summary of key aspects related to the task setup and background. Section 3 describes our approach in detail, outlining the architecture and components of our system. The results of our experiments are presented and discussed in Section 4. Finally, Section 5 concludes the paper with a summary of our findings and directions for future work.

2. Related works

The task of generating structured textual descriptions from medical images intersects multiple research domains, including computer vision, Natural Language Processing (NLP), and multimodal representation learning. In recent years, this field has experienced significant growth due to the increasing availability of annotated medical data and the emergence of pretrained models for both vision and language.

Medical image captioning aims to automatically produce descriptive text that summarizes the visual content of medical images. This is particularly relevant in different medical domains, such as radiology, CT, and among others, where textual reports accompany each scan and provide critical information for

[1] https://huggingface.co/Salesforce/blip-image-captioning-base

diagnosis and treatment decisions. Unlike generic image captioning datasets, such as MSCOCO [8] or Flickr30k [9], medical captioning requires models to handle domain-specific terminology, high levels of semantic precision, and context-dependent information.

Initial approaches in this area relied on encoder-decoder architectures, typically using Convolutional Neural Networks (CNNs) for image encoding [10] and Recurrent Neural Networks (RNNs) or LSTM variants [11] for text generation. In [12], the authors proposed a hierarchical LSTM with co-attention mechanisms trained on the IU X-ray dataset, which became a benchmark in this field. In addition, [13] further improved fluency and accuracy by incorporating clinical entity constraints and reinforcement learning to guide caption generation.

As transformer-based models gained popularity, researchers began adapting pretrained architectures for radiology. The BLIP model [5] introduced a unified vision-language framework based on image transformers and vision-language pretraining, demonstrating promising results in general and medical domains. Recently, models such as GIT [14] and Flamingo [15], have also been explored in biomedical captioning tasks, although they often require domain-specific fine-tuning to maintain clinical validity. It should be noted that evaluation for this type of task remains a challenge. Traditional metrics such as BLEU or ROUGE often fail to capture clinical correctness. BERTScore, which leverages contextual embeddings, has been shown to correlate more closely with expert judgments in medical NLP tasks and has become a standard evaluation metric in benchmarks such as ImageCLEFmed since 2021.

The detection and normalization of biomedical concepts from image descriptions or free text is another critical subtask, particularly when aiming to build explainable or retrievable multimodal representations. Named Entity Recognition (NER) and Concept Normalization (CN), also referred to as Medical Entity Linking, are often addressed jointly in medical NLP pipelines.

Traditional tools utilize rule-based systems and dictionary lookups over the UMLS to identify and normalize clinical terms. However, their rigidity and limited scalability have led to the development of embedding-based approaches. To address this, SciSpacy [16] has emerged as a modern NLP library built on spaCy and pretrained on biomedical corpora. It provides NER capabilities along with flexible entity linking modules and fast inference times. With the advent of contextualized language models and pretrained architectures such as BERT [17], RoBERTa [18], and XLM-RoBERTa [19], it is now possible to perform similarity-based retrieval more effectively. SapBERT [20] and PubMedBERT [21] are two domain-specific transformers trained on biomedical literature and concept alignment tasks. These models project entity mentions and concept definitions into a shared embedding space, enabling semantic retrieval of the most relevant UMLS concept for a given span.

Nevertheless, embedding similarity alone may produce noisy or semantically close but incorrect matches. To address this, recent work has introduced reranking mechanisms, where a secondary model (typically a sequence classifier) is trained to assess the compatibility between the entity and candidate concept definitions.

For this reason, a similar multi-stage approach was adopted in our system. The pipeline combined domain-adapted embeddings from SapBERT for initial concept retrieval with a BERT-based sequence classification model trained to rerank and validate candidate CUIs. This design allowed high recall to be maintained while improving concept relevance and overall precision in the concept detection subtask. Similarly, for caption generation, the system built upon state-of-the-art vision–language models such as BLIP, leveraging both pretrained image–text alignment and domain-specific fine-tuning to generate clinically meaningful descriptions.

3. Methodology

Our system addressed the two subtasks of the ImageCLEFmed Caption 2025 challenge: caption prediction and concept detection. We designed a modular architecture that integrates state-of-the-art vision–language models for caption generation with an entity linking pipeline based on biomedical embeddings and a reranking mechanism for concept normalization.

3.1. Dataset

For both subtasks, the ROCOv2 (Radiology Objects in Context Version 2) dataset [22], provided by the organizers, was used. This dataset is an updated and extended version of the original ROCO dataset. It was specifically designed for medical image understanding and is derived from open-access articles in the PubMed Central (PMC) OpenAccess subset. The dataset includes curated radiology images extracted from biomedical literature, each accompanied by a descriptive caption and manually controlled UMLS concept annotations. This structured metadata supports both semantic evaluation and downstream tasks such as information retrieval and concept linking.

The dataset is divided into the following subsets: 80,091 radiology images for training, 17,277 images for validation, and 19,267 images for testing. In this work, the dataset was specifically used for the first task, caption prediction, where the objective is to automatically generate descriptive and clinically meaningful captions for radiology images.

For the second subtask, concept detection, the same dataset (ROCOv2) was also used to extract biomedical entities from the image captions and associate them with UMLS concepts. To train a BERT-based reranking model for concept normalization, a labeled dataset was generated based on the original training and validation splits. The SciSpacy framework was employed to extract medical named entities from the captions, and for each entity, the top 20 most similar UMLS concepts were retrieved using cosine similarity over embeddings produced by a SapBERT model.

These candidate pairs were then labeled as positive or negative depending on whether the associated CUI was present in the gold standard annotations. Additionally, for concepts missed during the top-20 retrieval, semantic similarity between the UMLS term and the detected entities was used to construct additional positive examples. This process resulted in a balanced dataset for training and validating the BERT-based sequence classifier, which was employed to rerank and filter noisy candidate concepts. The resulting dataset statistics are presented in Table 1.

Table 1
Datasets statistics for training a reranked model based on BERT

label	train	val	total
positive	387.828	62.305	450.133
negative	2.275.520	508.996	2.784.516
total	2.663.348	571.301	3.234.649

3.2. Caption prediction task

To generate captions from medical images, we adopted a fine-tuning strategy based on the BLIP architecture. Specifically, we used the `Salesforce/blip-image-captioning-base` [23] model, which was selected due to its strong performance on general image captioning benchmarks and its design as a unified vision-language model. BLIP integrates a Vision Transformer (ViT) encoder with a language model decoder in a flexible framework that supports both generation and understanding tasks.

We fine-tuned this model on the dataset provided by the task organizers using the HuggingFace `Trainer` API [2]. The training set consisted of radiological images paired with expert-authored captions, providing strong supervision for clinical image description. Each image was preprocessed using the BLIP processor, which converts raw images and textual annotations into the appropriate tensor format through normalization, tokenization, and padding/truncation.

Our training configuration was as follows: 5 training epochs, batch size of 4 for both training and validation, a learning rate of 2×10^{-5}, and evaluation at the end of each epoch. To evaluate model performance, we designed a custom metric called *relevance*, which is the average of:

[2] https://huggingface.co/docs/transformers/main_classes/trainer

- **BERTScore (F1)** [24], computed using the `microsoft/deberta-xlarge-mnli` [25] model, which captures semantic similarity between predicted and reference captions.
- **ROUGE-1** [26], which measures unigram overlap, serving as a proxy for lexical similarity.
- **Exact Match Similarity**, a binary score (1 or 0) that checks whether the predicted caption exactly matches the reference.

These metrics were computed at the sentence level, and their means were combined into a single scalar relevance score per batch. The final relevance score is computed as:

$$\text{Relevance} = \frac{1}{3}\left(\text{BERTScore}_{F1} + \text{ROUGE-1} + \text{ExactMatch}\right) \tag{1}$$

This relevance metric served as the **reference metric during training**. At the end of each epoch, the model checkpoint that achieved the highest relevance score on the validation set was saved and used as the best model for downstream tasks. This approach ensured that model selection was based on a comprehensive measure of clinical and semantic adequacy, rather than surface-level overlap alone.

3.3. Concept detection task

Following caption generation, the second subtask involved extracting standardized biomedical concepts—specifically UMLS CUIs—from the generated captions. This task was approached using a hybrid pipeline consisting of three components: entity recognition, candidate retrieval via embedding similarity, and reranking via classification.

Biomedical named entities were first extracted from each caption using the `en_core_sci_lg` model provided by SciSpacy [16], a high-performance NLP library designed for scientific and biomedical text. This model was selected due to its training on relevant corpora and its ability to provide accurate tokenization and entity segmentation for complex medical expressions.

Each extracted entity was then normalized by retrieving its top-n most semantically similar UMLS concepts. For this step, `SapBERT` [27], a transformer-based model pretrained for biomedical concept alignment, was employed. Embeddings were computed for both the entity mention and a dictionary of UMLS terms (from the 2022AB UMLS release), and similarity was calculated using cosine distance.

This embedding-based retrieval strategy was chosen for its ability to capture synonyms and variations in expression, which are common in medical language. Unlike simple string matching, SapBERT enabled robust retrieval of conceptually aligned terms even when there was no exact lexical overlap.

While embedding similarity was effective at achieving high recall, it often introduced false positives due to semantically adjacent but incorrect matches. To improve precision, a reranking module based on `bert-base-uncased` was introduced and trained as a binary sequence classifier. This model received the entity and a candidate concept term as input and predicted whether the candidate was a correct normalization for the entity. As a result, the BERT-based classifier acted as a semantic filter that eliminated noisy or ambiguous matches, yielding cleaner and more clinically accurate predictions.

At inference time, the pipeline proceeds as follows:

1. Extract named entities from the generated caption using SciSpacy.
2. For each entity, compute its embedding and retrieve the top-n closest UMLS concept candidates using SapBERT similarity.
3. Use the BERT-based reranker to classify each candidate as relevant or not.
4. Return the CUIs that are positively classified. If no valid CUIs are detected, assign the label UNKNOWN.

For the reranking stage, we fine-tuned a `bert-base-uncased` model as a binary classifier to distinguish correct from incorrect entity-to-concept pairs. The model was trained with a batch size of 32, a learning rate of 2e-5, and maximum input sequence length set to 512 tokens. Tokenization was performed with the corresponding pretrained tokenizer, and all inputs were padded to a fixed length. Only the best-performing checkpoint was saved based on validation performance. An overview of the full pipeline, including both caption generation and concept detection components, is shown in Figure 1.

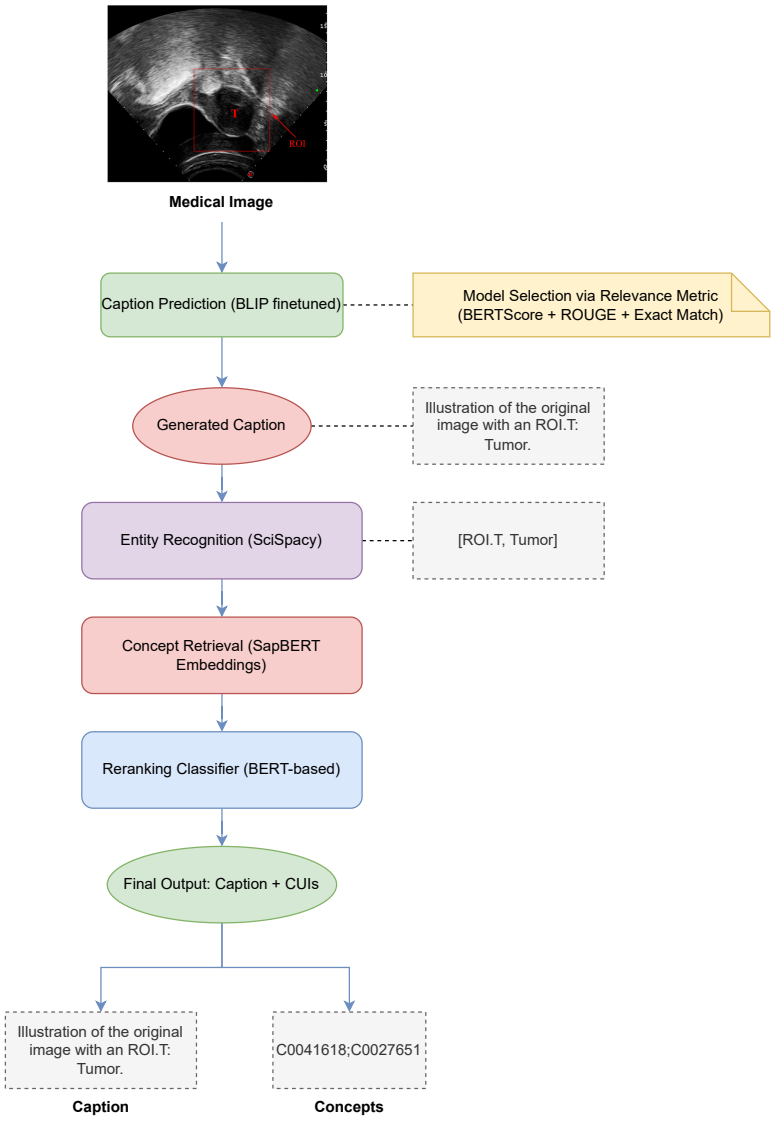

Figure 1: Overview of our two-stage system pipeline for caption prediction and concept detection. The model first generates a clinically relevant caption using a fine-tuned BLIP model. Biomedical entities are then extracted and normalized to UMLS CUIs via embedding-based retrieval and re-ranking. Image source: ImageCLEFmedical_Caption_2025_valid_0, CC BY, Shahzad et al., 2023.

4. Results

We submitted runs for both subtasks of the ImageCLEFmed Caption 2025 challenge and evaluated the performance of our system according to the official metrics provided by the organizers.

For the Concept Detection subtask, performance was evaluated using the F1 score. Each prediction was compared against the ground truth concepts using binary arrays, and the F1 score was calculated per instance and averaged across the test set. Two types of F1 scores were reported: a primary score considering all predicted and ground truth concepts, and a secondary score restricted to manually annotated concepts.

For the Caption Prediction subtask, evaluation was based on two key aspects: relevance and factuality. Relevance was assessed using semantic similarity (via embeddings), BERTScore, ROUGE-1, and BLEURT. Factuality was evaluated through UMLS Concept F1 (using MedCAT and QuickUMLS) and AlignScore

(based on RoBERTa for factual alignment). The final system ranking was determined by the average score across all six metrics, providing a balanced evaluation of both linguistic relevance and clinical factual accuracy.

4.1. Caption prediction results

In the caption prediction subtask, our submission achieved the best overall performance among all teams, with an overall score of **0.3432**. Key evaluation metrics included:

- **Overall**: 0.3432
- **Similarity**: 0.9271
- **BERTScore (Recall)**: 0.5977
- **ROUGE-1**: 0.2594
- **BLEURT**: 0.3230
- **Relevance Average**: 0.5268
- **UMLS Concept F1**: 0.1816
- **AlignScore**: 0.1375
- **Factuality Average**: 0.1596

To better understand the behavior of the model across epochs during training, we recorded the evolution of the `eval_relevance` score. As shown in Table 2, the best checkpoint (epoch 4) achieved a relevance score of **0.3581**, which was selected as the final model for test submission.

Table 2
Validation relevance scores per epoch during BLIP fine-tuning on the validation split.

Epoch	Eval Loss	Relevance Score
1	0.1889	0.3462
2	0.1799	0.3539
3	0.1775	0.3573
4	0.1793	**0.3581**
5	0.1823	0.3569

This consistent improvement confirms the effectiveness of using a relevance-driven training criterion, combining semantic and lexical measures to guide model selection.

4.2. Concept detection results

To improve concept selection precision on the validation split, we trained a reranking classifier using a BERT-based sequence classification model. On the validation set of the reranked dataset, the classifier achieved an overall accuracy of **90.42%** and a macro-averaged F1-Score of **0.8002**, as shown in Table 3.

Table 3
Validation results of the reranking model on the concept detection task on the validation split.

Label	Precision	Recall	F1-score
Negative (0)	0.9787	0.9124	0.9444
Positive (1)	0.5392	0.8377	0.6561
Macro avg	0.7589	0.8751	0.8002
Weighted avg	0.9308	0.9042	0.9129

Finally, our approach obtained an official F1-Score of **0.2398**, with a secondary F1-Score of **0.5377**. Although our score was lower than the top-ranking systems (F1 up to 0.5888), our pipeline produced

interpretable predictions based on semantic similarity and concept verification. However, a key limitation of our method is that it lacks visual interpretability. This means that the model's outputs cannot be directly traced back to specific regions in the image, making it harder to understand which visual features influenced the generated captions or extracted concepts.

4.3. Discussion

The overall performance of our system in the caption prediction subtask was strong, achieving the best score across multiple semantic and relevance-based metrics. The use of a fine-tuned BLIP model, coupled with relevance-driven model selection, allowed for the generation of coherent and clinically meaningful image captions. In contrast, the concept detection subtask yielded more modest results. Although our pipeline integrated advanced techniques such as SapBERT-based semantic retrieval and BERT-based reranking, the final F1-Score on the test set was relatively low (0.2398). One of the main limitations observed was the high number of false positives introduced during the retrieval stage.

As shown in our validation set statistics (Table 1), only 62,305 positive instances (correct UMLS concepts) were available, compared to over 2.27 million negative instances. This extreme class imbalance caused the reranking model to be highly sensitive to noise in the top-n retrieved concepts, especially when entities were ambiguous or not well aligned semantically with UMLS entries. Despite achieving high precision on the majority class (negative), the model struggled to confidently identify correct concepts among the candidates. It should be noted that, in cases where no CUI could be retained, we applied a fallback strategy by assigning the label UNKNOWN. This occurred in 10.81% of the test instances. While this strategy ensured the robustness of the system, it affected the performance of concept detection, especially in borderline cases where concepts were partially recognized, as shown in the results obtained in the ranking.

5. Conclusions and further work

In this paper, we described the participation of UMUTeam in both subtasks of the ImageCLEFmed 2025 shared task in CLEF, using a modular system for concept detection and caption prediction.

For caption prediction, we fine-tuned the BLIP model using a relevance-based metric that combines semantic, lexical, and exact-match evaluations. This resulted in strong performance, achieving the highest overall score across all participants in this subtask.

For concept detection, we implemented a multi-stage pipeline combining entity recognition based on SciSpacy with SapBERT-based semantic retrieval and a BERT-based reranker to filter noisy candidates. While the system design is robust and interpretable, its performance was limited by a high number of false positives and a strong class imbalance in the training data. The system showed difficulty in confidently detecting correct concepts, especially when dealing with ambiguous or underrepresented entities.

Thus, in future work, we aim to reduce both false negatives and false positives by training the reranking model on a more balanced dataset. In addition, we plan to explore recent transformer architectures capable of fusing image and text inputs (e.g., LLaVA) to enable end-to-end training for captioning and concept detection within a shared embedding space. For example, in [28] explores different approaches for fusing embeddings from multiple modalities, such as audio and image, for a classification task. Furthermore, we intend to incorporate LLMs as selectors, leveraging their strong performance in various NLP tasks, such as hate speech classification, as demonstrated in [29] and [30].

Acknowledgments

This work is part of the research project LaTe4PoliticES (PID2022-138099OB-I00) funded by MCIN/AEI/10.13039/501100011033 and the European Regional Development Fund (ERDF)-a way to

make Europe. Mr. Tomás Bernal-Beltrán is supported by University of Murcia through the predoctoral programme.

Declaration on Generative AI

During the preparation of this work, the author(s) used DeepL in order to: Grammar and spelling check.

Online Resources

The source code is available in the repository: https://github.com/NLP-UMUTeam/ImageCLEFmed-Caption-2025

References

[1] H. Ayesha, S. Iqbal, M. Tariq, M. Abrar, M. Sanaullah, I. Abbas, A. Rehman, M. F. K. Niazi, S. Hussain, Automatic medical image interpretation: State of the art and future directions, Pattern Recognition 114 (2021) 107856. URL: https://www.sciencedirect.com/science/article/pii/S0031320321000431. doi:https://doi.org/10.1016/j.patcog.2021.107856.

[2] P. Savadjiev, J. Chong, A. Dohan, M. Vakalopoulou, C. Reinhold, N. Paragios, B. Gallix, Demystification of ai-driven medical image interpretation: past, present and future, European radiology 29 (2019) 1616–1624.

[3] H. Damm, T. M. G. Pakull, H. Becker, B. Bracke, B. Eryilmaz, L. Bloch, R. Brüngel, C. S. Schmidt, J. Rückert, O. Pelka, H. Schäfer, A. Idrissi-Yaghir, A. B. Abacha, A. G. S. de Herrera, H. Müller, C. M. Friedrich, Overview of ImageCLEFmedical 2025 – medical concept detection and interpretable caption generation, in: CLEF 2025 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Madrid, Spain, 2025.

[4] B. Ionescu, H. Müller, D.-C. Stanciu, A.-G. Andrei, A. Radzhabov, Y. Prokopchuk, Ştefan, Liviu-Daniel, M.-G. Constantin, M. Dogariu, V. Kovalev, H. Damm, J. Rückert, A. Ben Abacha, A. García Seco de Herrera, C. M. Friedrich, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, C. S. Schmidt, T. M. G. Pakull, B. Bracke, O. Pelka, B. Eryilmaz, H. Becker, W.-W. Yim, N. Codella, R. A. Novoa, J. Malvehy, D. Dimitrov, R. J. Das, Z. Xie, H. M. Shan, P. Nakov, I. Koychev, S. A. Hicks, S. Gautam, M. A. Riegler, V. Thambawita, P. Halvorsen, D. Fabre, C. Macaire, B. Lecouteux, D. Schwab, M. Potthast, M. Heinrich, J. Kiesel, M. Wolter, B. Stein, Overview of ImageCLEF 2025: Multimedia retrieval in medical, social media and content recommendation applications, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the 16th International Conference of the CLEF Association (CLEF 2025), Springer Lecture Notes in Computer Science LNCS, Madrid, Spain, 2025.

[5] J. Li, D. Li, C. Xiong, S. Hoi, Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, in: International conference on machine learning, PMLR, 2022, pp. 12888–12900.

[6] M. Neumann, D. King, I. Beltagy, W. Ammar, ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing, in: Proceedings of the 18th BioNLP Workshop and Shared Task, Association for Computational Linguistics, Florence, Italy, 2019, pp. 319–327. URL: https://www.aclweb.org/anthology/W19-5034. doi:10.18653/v1/W19-5034. arXiv:arXiv:1902.07669.

[7] F. Liu, E. Shareghi, Z. Meng, M. Basaldella, N. Collier, Self-alignment pretraining for biomedical entity representations, in: K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, Y. Zhou (Eds.), Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 4228–4238. URL: https://aclanthology.org/2021.naacl-main.334/. doi:10.18653/v1/2021.naacl-main.334.

[8] O. Vinyals, A. Toshev, S. Bengio, D. Erhan, Show and tell: Lessons learned from the 2015 mscoco image captioning challenge, IEEE transactions on pattern analysis and machine intelligence 39 (2016) 652–663.

[9] H. Liu, Y. Song, X. Wang, X. Zhu, Z. Li, W. Song, T. Li, Flickr30k-cfq: A compact and fragmented query dataset for text-image retrieval, in: International Conference on Database Systems for Advanced Applications, Springer, 2024, pp. 419–434.

[10] M. B. S. Reddy, P. Rana, Biomedical image classification using deep convolutional neural networks–overview, in: IOP Conference Series: Materials Science and Engineering, volume 1022, IOP Publishing, 2021, p. 012020.

[11] Y. Tatsunami, M. Taki, Sequencer: Deep LSTM for image classification, Advances in Neural Information Processing Systems 35 (2022) 38204–38217.

[12] B. Jing, P. Xie, E. Xing, On the automatic generation of medical imaging reports, in: I. Gurevych, Y. Miyao (Eds.), Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 2577–2586. URL: https://aclanthology.org/P18-1240/. doi:10.18653/v1/P18-1240.

[13] G. Liu, T.-M. H. Hsu, M. McDermott, W. Boag, W.-H. Weng, P. Szolovits, M. Ghassemi, Clinically accurate chest x-ray report generation, in: Machine Learning for Healthcare Conference, PMLR, 2019, pp. 249–269.

[14] 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, arXiv preprint arXiv:2205.14100 (2022).

[15] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, et al., Flamingo: a visual language model for few-shot learning, Advances in neural information processing systems 35 (2022) 23716–23736.

[16] M. Neumann, D. King, I. Beltagy, W. Ammar, Scispacy: Fast and robust models for biomedical natural language processing, in: Proceedings of the 18th BioNLP Workshop and Shared Task, 2019, pp. 319–327.

[17] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, CoRR abs/1810.04805 (2018). URL: http://arxiv.org/abs/1810.04805. arXiv:1810.04805.

[18] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized BERT pretraining approach, CoRR abs/1907.11692 (2019). URL: http://arxiv.org/abs/1907.11692. arXiv:1907.11692.

[19] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, CoRR abs/1911.02116 (2019). URL: http://arxiv.org/abs/1911.02116. arXiv:1911.02116.

[20] F. Liu, E. Shareghi, Z. Meng, M. Basaldella, N. Collier, Self-alignment pretraining for biomedical entity representations, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021, pp. 4228–4238.

[21] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, H. Poon, Domain-specific language model pretraining for biomedical natural language processing, 2020. arXiv:arXiv:2007.15779.

[22] 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. Horn, F. Nensa, C. M. Friedrich, ROCOv2: Radiology objects in context version 2, an updated multimodal image dataset, Scientific Data 11 (2024). doi:10.1038/s41597-024-03496-6.

[23] J. Li, D. Li, C. Xiong, S. Hoi, Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022. URL: https://arxiv.org/abs/2201.12086. doi:10.48550/ARXIV.2201.12086.

[24] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with BERT, in: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020, OpenReview.net, 2020. URL: https://openreview.net/forum?id=

SkeHuCVFDr.
[25] P. He, X. Liu, J. Gao, W. Chen, Deberta: Decoding-enhanced bert with disentangled attention, in: International Conference on Learning Representations, 2021. URL: https://openreview.net/forum?id=XPZIaotutsD.
[26] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text Summarization Branches Out, Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 74–81. URL: https://aclanthology.org/W04-1013/.
[27] F. Liu, E. Shareghi, Z. Meng, M. Basaldella, N. Collier, Self-alignment pretraining for biomedical entity representations, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 4228–4238. URL: https://www.aclweb.org/anthology/2021.naacl-main.334.
[28] R. Pan, J. A. García-Díaz, M. Ángel Rodríguez-García, R. Valencia-García, Spanish meacorpus 2023: A multimodal speech–text corpus for emotion analysis in spanish from natural environments, Computer Standards & Interfaces 90 (2024) 103856. URL: https://www.sciencedirect.com/science/article/pii/S0920548924000254. doi:https://doi.org/10.1016/j.csi.2024.103856.
[29] R. Pan, J. A. García-Díaz, R. Valencia-García, Optimizing few-shot learning through a consistent retrieval extraction system for hate speech detection, Procesamiento del Lenguaje Natural 74 (2025) 241–252.
[30] R. Pan, J. A. García-Díaz, R. Valencia-García, Spanish mtlhatecorpus 2023: Multi-task learning for hate speech detection to identify speech type, target, target group and intensity, Computer Standards & Interfaces 94 (2025) 103990.