UIT-Oggy at ImageCLEFmedical 2025 Caption: CSRA-Enhanced Concept Detection and BLIP-Driven Vision-Language Captioning

Gia-Phuc Bui-Hoang[1,2], My-Huyen Dinh-Doan[1,2], Van-Minh Luong[1,2] and Thien B. Nguyen-Tat[1,2,*]

[1]University of Information Technology, Ho Chi Minh City, Vietnam
[2]Vietnam National University, Ho Chi Minh City, Vietnam

Abstract

This paper presents the development and evaluation of two deep learning models for the ImageCLEFmedical 2025 challenge, targeting automated medical concept detection and diagnostic captioning. The primary goal was to create practical models that enhance the accuracy and clinical relevance of automated radiological reporting. For the concept detection task, a novel dual-branch architecture named MedCSRA (Medical Class-Specific Residual Attention) was proposed. It integrates a Class-Specific Residual Attention branch with a Global branch, using a ResNet-101 backbone, to effectively balance localized and global visual features. For the caption prediction task, a pre-trained BLIP (Bootstrapping Language-Image Pre-training) model was fine-tuned on the competition's dataset to adapt its powerful vision-language capabilities to the medical domain. Both submissions achieved strong results in the official competition. The MedCSRA model secured a fourth-place ranking in the concept detection task, with a primary F1-score of 0.5613. In the captioning task, the fine-tuned BLIP model also achieved a fourth-place ranking, with a competitive overall score of 0.3554 and strong semantic coherence, as measured by a BERTScore (Recall) of 0.5951. Our results demonstrate two key findings. First, a custom dual-branch architecture that explicitly balances local and global context is a highly effective strategy for multi-label concept detection. Second, standard fine-tuning of a large pre-trained vision-language model like BLIP is sufficient to achieve a top-tier ranking in semantic captioning metrics, though challenges in clinical factuality, measured by metrics like UMLS (Unified Medical Language System) Concept F1, remain. These findings validate our approaches as competitive solutions for complex biomedical image analysis tasks.

Keywords

ImageCLEF 2025, Medical Image Processing, Concept Detection, Image Captioning, Radiology Images, CSRA, BLIP, Vision-Language Models, ResNet-101, DenseNet-121, EfficientNetB4, EfficientNetB5, Transformer Models, Multimodal Learning, Diagnostic Captioning, Multilabel Classification

1. Introduction

In recent years, machine learning—particularly deep learning—has been a driving force behind substantial advancements in biomedicine. As the volume of medical imaging data continues to grow rapidly, there is a rising need for intelligent systems that can extract meaningful information, assist with clinical decision-making, and streamline workflows in healthcare environments.

One key area in this domain is diagnostic captioning, which involves generating descriptive, diagnostic-level text based on medical images. This task holds great promise in assisting clinicians by improving reporting efficiency and reducing the risk of human error—especially for less experienced practitioners. Rather than replacing human expertise, these systems are designed to augment and support the diagnostic process by providing preliminary insights that guide medical professionals toward faster and more accurate decisions.

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain
*Corresponding author.
22521109@gm.uit.edu.vn (G. Bui-Hoang); 22520589@gm.uit.edu.vn (M. Dinh-Doan); 22520869@gm.uit.edu.vn (V. Luong); thiennttb@uit.edu.vn (T. B. Nguyen-Tat)
0009-0009-0642-7594 (G. Bui-Hoang); 0009-0006-6806-8134 (M. Dinh-Doan); 0009-0006-1910-7117 (V. Luong); 0000-0002-4809-7126 (T. B. Nguyen-Tat)

ImageCLEF [1], an annual evaluation campaign, provides a structured platform for advancing research in multimodal machine learning. Among its major tracks, ImageCLEFmedical addresses biomedical image analysis through a series of challenges, including diagnostic captioning [2].

Our team, UIT-Oggy, participated in the ImageCLEFmedical 2025 Caption task, which consists of two complementary subtasks: medical concept detection and caption prediction. In the concept detection task, the goal is to identify and extract medically relevant terms from an image, which can enhance indexing, retrieval, and diagnostic support. The caption prediction task—also referred to as diagnostic captioning—focuses on generating coherent, informative, and medically accurate descriptions of patient conditions and anatomical structures visible in the image.

While diagnostic captioning remains a challenging problem due to the complexity of medical language and visual interpretation, it represents a transformative tool in modern clinical workflows. By providing initial report drafts and highlighting important image features, these models can reduce report turnaround times and help clinicians manage increasing workloads. At the same time, concept detection ensures that critical medical terms are not overlooked and serves as a foundation for structured reporting and semantic image understanding.

In this paper, we detail our methods and results from the ImageCLEFmedical 2025 challenge [2]. We explore Transformer-based architectures for captioning and introduce a novel attention-based model for concept detection, aiming to contribute practical solutions that can enhance diagnostic support systems.

2. Related Work

This section reviews prior work in two key areas relevant to our participation in the ImageCLEFmedical 2025 challenge. We begin with a broad overview of deep learning's impact on medical imaging before delving into task-specific advancements.

A comprehensive evaluation by Nguyen-Tat et al.[3] highlights the critical role of robust preprocessing pipelines and the effectiveness of deep learning methods across multiple modalities, establishing a baseline for developing high-performance models. Underpinning many advanced applications is the fundamental task of medical image segmentation, where recent innovations include hybrid architectures combining the U-Net architecture (so-named for its U-shape), attention mechanisms, and Transformer models for precise brain tumor segmentation [4], as well as weakly-supervised approaches like Qmaxvit-unet+ for scribble-based segmentation [5].

The task of medical image captioning is distinguished from its general-domain counterpart by its stringent requirements for clinical accuracy and domain-specific knowledge [6]. While early systems relied on template- or retrieval-based methods [7, 8], the field was revolutionized by deep learning. Foundational approaches employed a Convolutional Neural Network-Recurrent Neural Network (CNN-RNN) encoder-decoder architecture, using networks like DenseNet for image encoding and hierarchical Long Short-Term Memory (LSTM) networks for text generation, as demonstrated in influential works like TieNet [9].

A pivotal advancement was the integration of attention mechanisms, enabling models to focus on salient image regions. In an influential study, Jing et al.[10] proposed a hierarchical attention model that significantly improved caption coherence and accuracy. To address challenges with rare clinical findings, some studies have explored hybrid models, such as the reinforcement learning-based agent by Li et al.[11], which intelligently alternates between retrieving templates and generating novel sentences.

The current state-of-the-art is increasingly driven by Large Language Models (LLMs) and pre-trained vision-language models. Notably, models like Med-PaLM (Medical Pathways Language Model) have demonstrated impressive capabilities by achieving a passing score on the US Medical Licensing Examination [12]. Concurrently, models like BLIP (Bootstrapping Language-Image Pre-training) [13] leverage bootstrapping mechanisms to learn from noisy web-scale data, providing a strong foundation that can be fine-tuned for medical tasks.

Parallel to captioning, the automatic annotation of medical images with multiple concepts is a critical

multi-label classification task. As documented in the survey by Litjens et al. [14], this field has shifted from traditional methods using hand-crafted features to deep learning-based approaches.

A key development in this area was the release of the large-scale ChestX-ray14 dataset, which enabled the creation of high-performance models for thoracic disease detection [15]. The influential CheXNet model, a DenseNet-based architecture trained on this dataset, demonstrated the potential to achieve radiologist-level performance. However, this approach is tailored to a specific modality. More directly related to our work, several studies have explored advanced attention mechanisms. For instance, the Residual Attention Network introduced by Wang et al.[16] demonstrated how stacking attention modules within residual blocks can significantly improve classification performance. In parallel, Li et al.[17] proposed a framework with a class-specific attention module. While these methods show robust performance, they often do not explicitly balance global image context with localized attention—a key gap that our proposed MedCSRA model aims to address.

The ImageCLEF 2025 [1] challenge, part of the broader CLEF 2025 (Conference and Labs of the Evaluation Forum) initiative, serves as a crucial benchmark for advancing both tasks, providing standardized datasets and a rigorous evaluation framework [2]. Our work builds directly on this foundation, contributing a novel dual-branch architecture for concept detection that addresses the nuanced challenge of integrating global and local features, a limitation observed even in top-performing systems from previous years.

3. Dataset

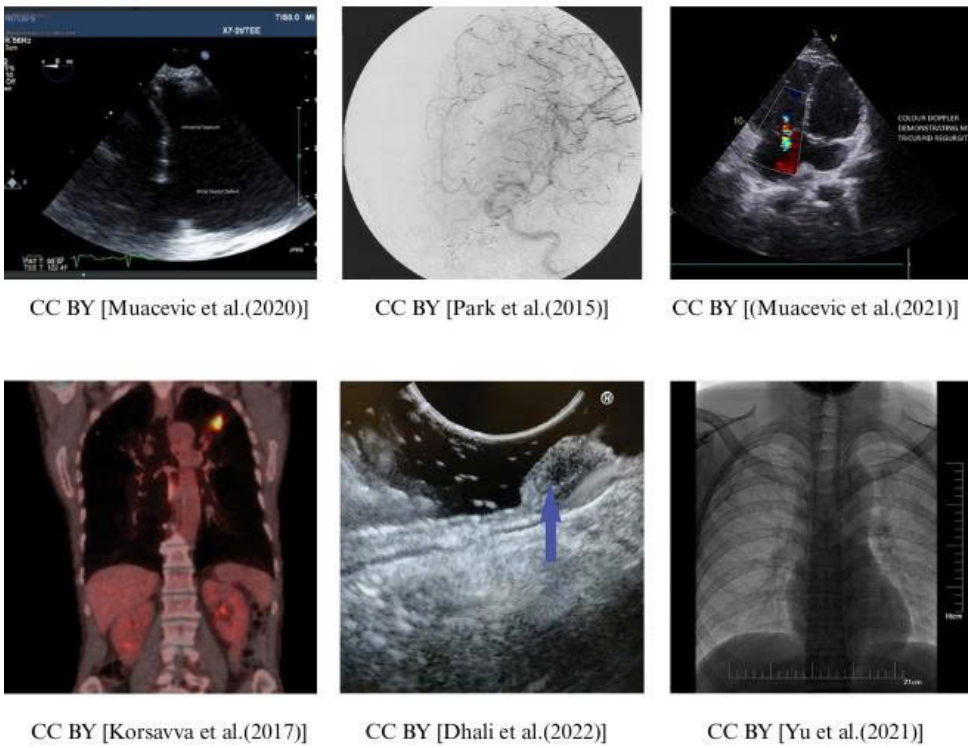

Figure 1: Several images from the ImageCLEFmedical2025 dataset.

Our work utilizes the official dataset provided for the ImageCLEFmedical 2025 Caption task. This dataset is a curated version of the Radiology Objects in Context (ROCO) v2 dataset[18], which serves as its foundation. The original ROCOv2 is a large-scale, "in-the-wild" benchmark comprising over 81,000

image-caption pairs sourced from biomedical publications in the PubMed Central Open Access corpus. For the 2025 competition, this dataset was specifically updated according to the established ImageCLEF procedure: a new, unseen test set was introduced, while the test set from the previous year became the current validation set, and the former validation set was integrated into the training data. This process and the final dataset composition are detailed in the official task overview paper [2]. The resulting 2025 training set consists of 70,108 medical images, primarily from radiology. A key feature of this dataset is its dual-annotation structure: each image is paired with both a set of UMLS (Unified Medical Language System) [19]-based medical concept labels and a free-text diagnostic caption. This design facilitates the joint training of models for concept detection and caption generation. The dataset encompasses various imaging modalities, such as X-rays and Computed Tomography (CT) scans, challenging models to generalize across different diagnostic scenarios. Examples of images from the dataset are presented in Figure 1.

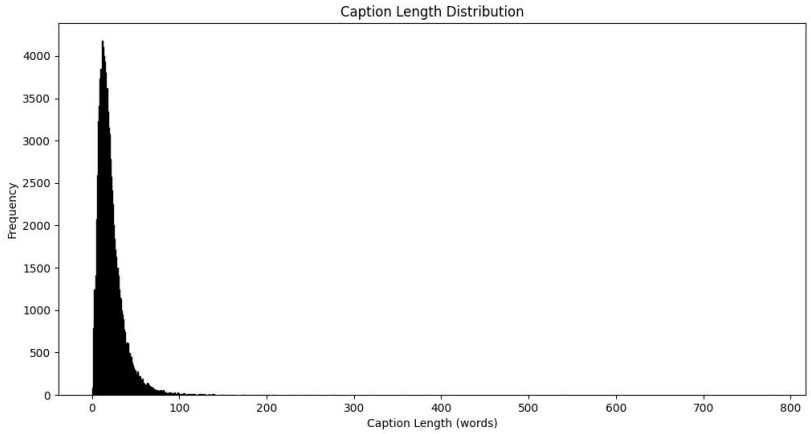

Figure 2: Distribution of caption lengths in the training set.

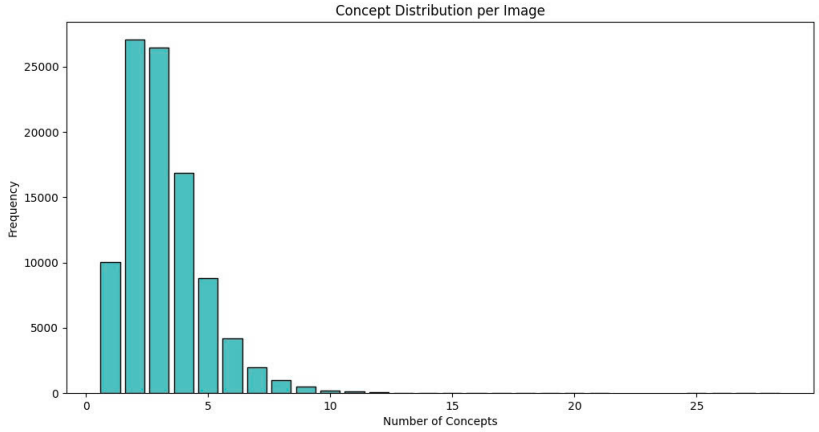

Figure 3: Concept Distribution per Image in the training set.

As shown in Figure 2, the distribution of caption lengths in the training set is highly right-skewed. Most captions are concise, typically fewer than 30 words, while a small number are extremely long, with the longest exceeding 800 words. This long-tailed behavior suggests that, while many image descriptions are brief, models must also be robust enough to handle verbose and complex medical text.

In Figure 3, we visualize the number of concepts annotated per image. The majority of images contain 1 to 4 concepts, with a sharp decline in frequency beyond that. Only a few cases involve 10 or more

concepts. This confirms the multilabel nature of the task and indicates that models must be capable of identifying both sparse and dense sets of medical concepts depending on the image complexity.

4. Image Pre-processing

Image preprocessing is a pivotal step in preparing medical images for captioning tasks, ensuring that input data is standardized and optimized for model performance. The approach focuses on transforming raw images into a format suitable for a vision-language model, balancing computational efficiency with the preservation of essential visual information.

4.1. Image Loading and Resizing

The images are loaded from the dataset directory, specifically handling JPEG files, using a computer vision library. Each image is resized to a uniform resolution of 224x224 pixels. This fixed size aligns with the input requirements of the vision-language model, ensuring compatibility with its pre-trained vision component. Standardizing image dimensions reduces computational complexity and facilitates consistent feature extraction in diverse medical images.

4.2. Data Encoding with Processor

A specialized processor, designed for the vision-language model, is used to prepare both images and their corresponding captions. This processor handles the following tasks:

- Image Normalization: Pixel values are scaled and normalized to match the expected input range of the pre-trained model, ensuring consistent processing across varied image sources.
- Text Tokenization: Captions are tokenized using the model's tokenizer, with padding applied to a maximum length of 200 tokens and truncation using a strategy that prioritizes longer sequences. This ensures that variable-length captions are uniformly formatted.
- Tensor Conversion: Both images and tokenized captions are converted into tensors, with unnecessary dimensions removed to match the model's input requirements. The resulting data includes image tensors, tokenized caption IDs, and attention masks to indicate valid tokens, which are critical for both training and inference.

5. Proposed Method

5.1. Caption Prediction

For the caption prediction task, the BLIP model was employed as the foundational architecture [13]. BLIP is a state-of-the-art vision-language framework that unifies understanding and generation tasks. Its robust performance, derived from pre-training on large-scale, noisy web data, makes it a strong candidate for adaptation to specialized domains like medical imaging.

5.1.1. Model Architecture

The BLIP architecture integrates a Vision Transformer (ViT) [20] as its image encoder and a BERT-based transformer [21] as its text encoder/decoder. The ViT processes an input image by dividing it into a sequence of patches and encoding them into rich visual representations, capturing both global and local features. This is particularly crucial for medical images where diagnostic clues can be subtle and localized. The text decoder, conditioned on these visual features, generates the caption autoregressively. A key strength of BLIP is its joint training strategy, which optimizes both the encoder and decoder to learn aligned multimodal representations. This pre-trained foundation is then leveraged during the fine-tuning stage, where the model adapts to the specific terminology and visual characteristics of the medical domain. The architecture is illustrated in Figure 4.

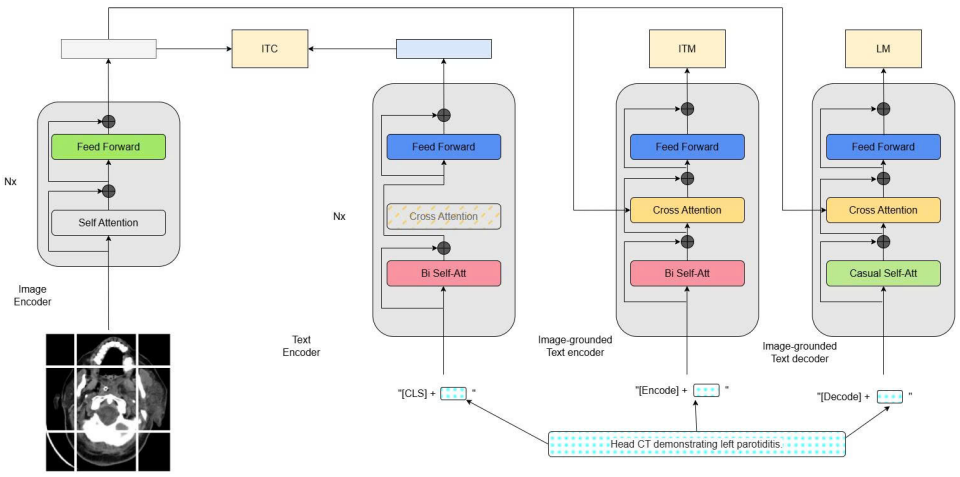

Figure 4: Architecture of a combined image and text encoding model for medical report generation.

5.1.2. Fine-tuning and Implementation Details

The captioning model was implemented using the PyTorch framework and the Hugging Face Transformers library [22]. The starting point for the experiments was the pre-trained `Salesforce/blip-image-captioning-large` checkpoint, publicly available on the Hugging Face Hub.

This model was subsequently fine-tuned on the official ImageCLEFmedical 2025 Caption [2] training set for 2 epochs. The AdamW optimizer [23] was employed with a learning rate of 1×10^{-5}. Due to hardware limitations, a batch size of 4 was used. All training was conducted on a single NVIDIA T4 GPU. During the fine-tuning process, all parameters of the pre-trained model were unfrozen and optimized to maximally adapt the model's representations to the medical domain.

5.2. Concept Detection

5.2.1. Overview of the MedCSRA Architecture

For the medical concept detection task, a novel dual-branch architecture named MedCSRA is proposed for multi-label image classification. The core innovation of MedCSRA lies in its hybrid approach, which integrates both global and local features to create a more comprehensive image representation. As illustrated in Figure 5, the model consists of a shared visual backbone and two parallel processing branches: a Global Branch and a Class-Specific Residual Attention (CSRA) Branch. The CSRA branch is inspired by the attention mechanism introduced by Fang et al. [24], which learns class-specific feature maps to focus on discriminative regions for each concept. This mechanism was adapted for the medical domain. Concurrently, a newly designed Global Branch processes the entire feature map to capture broader contextual information. The outputs (logits) from both branches are then fused, and a sigmoid function is applied to produce the final multi-label predictions.

5.2.2. Visual Feature Extraction

The MedCSRA model employs a ResNet-101 architecture as its visual feature extraction backbone [25]. ResNet-101, a 101-layer Residual Network, was chosen for its proven effectiveness in various computer vision tasks, including medical image analysis. Its core innovation lies in the use of "residual connections," which reformulate layers as learning residual functions with reference to the layer inputs. This design effectively mitigates the vanishing gradient problem in very deep networks and addresses the degradation issue where deeper models can perform worse than shallower ones. To leverage transfer learning, the ResNet-101 backbone was initialized with weights pre-trained on the large-scale ImageNet dataset [26]. For domain adaptation to our specific task, a fine-tuning strategy was applied: while the

majority of the network's early layers were kept frozen, the final convolutional block (layer4) was unfrozen and trained on the ImageCLEFmedical 2025 dataset. This approach allows the model to retain robust, low-level features from ImageNet while specializing its high-level feature representations for the unique patterns found in radiology images.

5.2.3. Dual-Branch Design

MedCSRA processes spatial features through two branches: the Class-Specific Residual Attention Branch, adapted from reference [24], and our novel Global Branch.

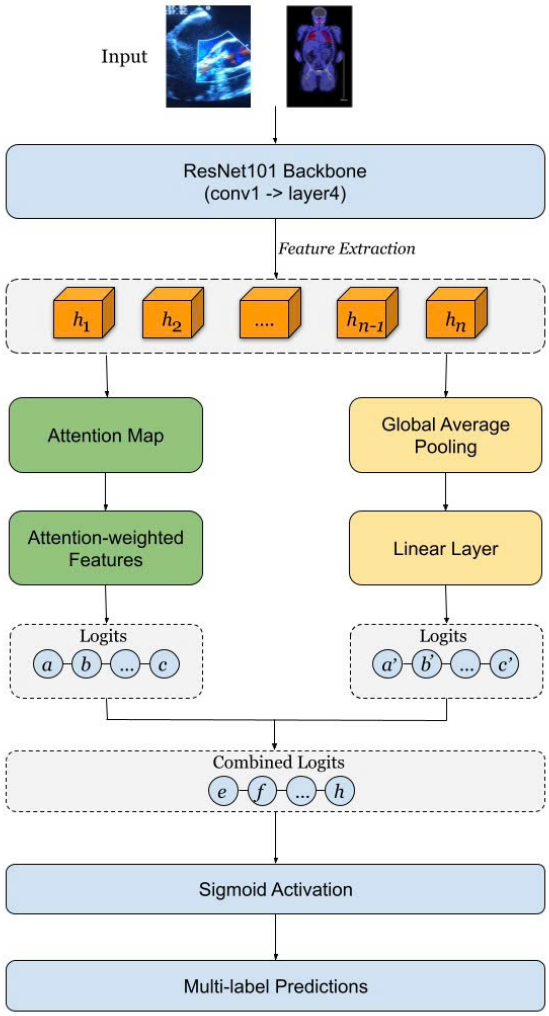

Figure 5: A diagram illustrating the MedCSRA architecture. The process begins with a ResNet-101 Backbone extracting spatial features, which are then processed by two branches: the Attention Map branch and the Global Average Pooling branch. The outputs are combined and passed through a Sigmoid Activation to produce multi-label predictions.

5.2.3.1. Global Branch

The Global Branch, our original contribution, enhances the model's ability to capture global semantic patterns across the entire image. Spatial features extracted from the ResNet-101 backbone are subjected to Global Average Pooling, reducing the spatial dimensions to a compact vector of size [batch_size, in_features]. This vector encapsulates overarching contextual information, which is then

passed through a fully connected layer to generate global logits of size [batch_size, num_classes]. The use of Global Average Pooling ensures computational efficiency while preserving critical global context, making it effective for detecting medical concepts with diffuse patterns. The design leverages the backbone's hierarchical features, complementing the localized focus of the CSRA Branch.

Let $\mathcal{H} = \{h_1, h_2, \ldots, h_n\}$ represent the spatial feature maps extracted by the backbone, where $h_i \in \mathbb{R}^{1024 \times H \times W}$ and H and W are the height and width, reflecting the output channels of ResNet-101's last convolutional block. Global Average Pooling is applied to compute a global feature vector:

$$F_{\text{global}} = \frac{1}{H \cdot W} \sum_{i=1}^{H} \sum_{j=1}^{W} h[:, i, j]$$

where $F_{\text{global}} \in \mathbb{R}^{1024}$ is the pooled vector for each sample in the batch. This vector is then passed through a fully connected layer to produce global logits:

$$L_{\text{global}} = W_{\text{fc}} \cdot F_{\text{global}} + b_{\text{fc}}$$

where $W_{\text{fc}} \in \mathbb{R}^{1024 \times K}$ and $b_{\text{fc}} \in \mathbb{R}^{K}$ are the weights and bias of the linear layer, and K is the number of classes.

5.2.3.2. Class-Specific Residual Attention Branch

Inspired by prior work [24], the Class-Specific Residual Attention Branch (CSRA) in MedCSRA focuses on localized pathological regions by applying a class-specific attention mechanism. Within the model, spatial feature maps H (with 1024 channels from the ResNet-101 backbone) are flattened into a tensor of shape $[\text{batch\_size}, 1024, H \times W]$. A linear projection, implemented as a fully connected layer with 1024 output features, generates an attention map, which is then normalized using the sigmoid function.

This attention map is used to compute attention-weighted features through a tensor operation, implemented via `torch.einsum`, resulting in a vector of shape $[\text{batch\_size}, 1024]$. These weighted features are then mapped to logits of size [batch_size, num_classes] using another fully connected layer. This process allows CSRA to emphasize class-specific regions, enhancing the model's ability to detect subtle pathological patterns. The resulting logits are subsequently combined with those from the Global Branch to produce the final predictions.

5.2.4. Fusion and Prediction

The final logits are computed by fusing the outputs of the two branches using a weighted combination:

$$L_{\text{combined}} = (1 - \lambda) \cdot L_{\text{global}} + \lambda \cdot L_{\text{csra}},$$

where $\lambda = 0.1$, a value set based on the model implementation. This weighted fusion balances the global context from the Global Branch and the localized attention from the CSRA Branch.

A threshold is applied to the fused logits to determine the predicted labels. This threshold is automatically selected by evaluating performance on the validation set across a range of values from 0.05 to 0.55. The optimal threshold determined during training was 0.35, corresponding to the highest achieved F1-score.

5.2.5. Loss Function

Given that this is a multilabel classification task, we use the Binary Cross Entropy Loss (BCE) computed independently for each class:

$$L = -\sum_{c=1}^{C} y_c \cdot \log(\hat{y}_c) + (1 - y_c) \cdot \log(1 - \hat{y}_c) \tag{1}$$

where C is the number of medical concepts, and $y_c \in \{0, 1\}$ is the ground truth label for class c. This loss function ensures the model learns to predict multiple labels accurately, aligning with the task's requirements.

5.2.6. Implementation and Training Details

The MedCSRA model was implemented in PyTorch. For optimization, the Adam optimizer was used with an initial learning rate of 1×10^{-4}, adjusted via a `CosineAnnealing` scheduler over a maximum of 100 epochs. A weight decay of 1×10^{-5} was applied to mitigate overfitting. To ensure reproducibility, all experiments were conducted with a fixed random seed of 42.

The model was trained with a batch size of 16 on a single NVIDIA A100 GPU with 40 GB of VRAM. For the ResNet-101 backbone, only the final convolutional block was unfrozen for fine-tuning, while earlier layers retained their pre-trained ImageNet weights. An early stopping mechanism was employed with a patience of 5 epochs, monitoring the F1-score on the validation set. Based on this criterion, the training process for the ResNet-101-based model concluded at epoch 34.

6. Experiment Results

This section details the performance of our proposed models on the official test sets of the ImageCLEFmedical 2025 challenge. We present the results for the concept detection and caption prediction subtasks separately.

6.1. Caption Prediction Results

For the caption prediction subtask, a fine-tuned BLIP model was submitted (Team: UIT-Oggy, Run ID: 1914). The comprehensive evaluation results, provided by the challenge organizers across multiple metrics, are summarized in Table 1.

Table 1
Performance of the BLIP-based Model on the Official ImageCLEF Test Set.

Model	Overall (%)	Similarity (%)	BERTScore (Recall) (%)	ROUGE-1 (%)	BLEURT (%)	Relevance Avg (%)	UMLS Concept F1 (%)	AlignScore (%)	Factuality Avg (%)
BLIP (ID:1914)	**32.11**	**87.98**	**59.51**	**25.35**	**30.20**	**50.76**	**16.72**	**10.21**	**13.46**

The model achieved an overall score of 32.11%, indicating a moderate level of performance. While the high Similarity (87.98%) and BERTScore (59.51%) suggest a good semantic alignment with the reference captions, other metrics revealed significant challenges. The low UMLS Concept F1 (16.72%) and Factuality Average (13.46%) scores highlight that the model struggled to maintain clinical accuracy and factual correctness, which is a critical limitation. This suggests that while pre-trained vision–language models like `BLIP` provide a strong foundation, extensive domain-specific adaptation is required to overcome the nuances of medical reporting.

6.2. Concept Detection Results

For the concept detection subtask, our primary submission utilized the MedCSRA model with a ResNet-101 backbone. The official performance of this submission (Team: UIT-Oggy, Run ID: 1892) on the final test set, as evaluated by the challenge organizers, is presented in Table 2.

Table 2
Performance of MedCSRA on the Official ImageCLEF Test Set.

Model	F1-Score (%)	F1-Secondary (%)
MedCSRA (ID:1892)	**56.13**	**91.04**

The model achieved a primary F1-score of 0.5613, securing a fourth-place ranking among all international teams participating in the challenge. This strong competitive result serves as a direct validation of our proposed dual-branch architecture. The high performance suggests that the model's ability to fuse global contextual features (from the Global Branch) with localized, class-specific details (from the CSRA Branch) is a highly effective strategy for this complex multi-label classification task. It demonstrates that explicitly balancing these two types of information allows the model to successfully identify a wide range of medical concepts, from diffuse abnormalities to subtle, localized findings, within a single unified framework.

7. Conclusions and Future Work

7.1. Conclusions

In this paper, we presented the methods and results of the UIT-Oggy team's participation in the ImageCLEFmedical 2025 challenge.

For the medical concept detection task, we introduced MedCSRA, a novel dual-branch architecture designed to integrate both global and localized visual features. Our experiments demonstrated that this approach is highly effective, with the MedCSRA model using a ResNet-101 backbone achieving a fourth-place ranking in the official competition (F1-score: 0.5613). This result validates that balancing class-specific attention with global context is a robust strategy for multi-label medical image classification.

For the caption prediction task, we adapted and fine-tuned a pre-trained BLIP model. While the model showed a strong capability for generating semantically coherent text (Similarity: 87.98%), our analysis revealed significant limitations in its clinical and factual accuracy (UMLS Concept F1: 16.72%). This finding underscores a key challenge: despite the power of large-scale pre-training, achieving clinical reliability in generative models requires more advanced domain-specific adaptation techniques.

Overall, our work contributes an effective architecture for concept detection and provides a critical analysis of the current state of vision-language models in the context of diagnostic captioning, paving the way for future research directions.

7.2. Future Work

Based on the findings of this study, several promising research directions are identified. For our medical concept detection model, MedCSRA, future work will focus on addressing the challenge of class imbalance, a common issue in medical datasets where rare but critical concepts are underrepresented. We plan to incorporate Focal Loss [27] into the training process. This will compel the model to pay more attention to hard-to-classify, infrequent concepts, aiming to enhance its diagnostic utility while building upon the successful dual-branch architecture.

For medical image captioning, our analysis revealed that while the fine-tuned BLIP model achieves strong semantic coherence, its clinical and factual accuracy remains a significant limitation. To bridge this gap, a primary direction is to infuse external medical knowledge into the model. Furthermore, to improve the low factuality score, we propose introducing a fact-checking module, inspired by claim verification systems [28], which would validate generated statements against a trusted medical knowledge base.

Finally, to enhance the model's ability to ground its descriptions in visual evidence and improve the AlignScore, we intend to explore multi-scale attention mechanisms like the Convolutional Block Attention Module CBAM [29]. This would allow the model to simultaneously focus on both broad anatomical structures and fine-grained pathological details. By pursuing these targeted enhancements, we aim to develop more robust and clinically reliable models for automated medical image analysis, addressing the specific challenges identified in each task.

Declaration on Generative AI

During the preparation of this work, we used Gemini 2.5 Pro and ChatGPT-4o in order to check grammar and sentence structure. After using these tools, we reviewed and edited the content as needed and take full responsibility for the publication's content.

Acknowledgments

This research is funded by University of Information Technology-Vietnam National University HoChiMinh City under grant number D4-2025-04.

References

[1] B. Ionescu, H. Müller, D. Stanciu, A. Andrei, A. Radzhabov, Y. Prokopchuk, L. Ştefan, M. 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. 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.

[2] 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.

[3] T. B. Nguyen-Tat, T. Q. Hung, P. T. Nam, V. M. Ngo, Evaluating pre-processing and deep learning methods in medical imaging: Combined effectiveness across multiple modalities, Alexandria Engineering Journal 119 (2025) 558–586. doi:10.1016/j.aej.2025.01.090.

[4] T. B. Nguyen-Tat, T. Q. T. Nguyen, H. N. Nguyen, V. M. Ngo, Enhancing brain tumor segmentation in mri images: A hybrid approach using unet, attention mechanisms, and transformers, Egyptian Informatics Journal 27 (2024). doi:10.1016/j.eij.2024.100528.

[5] T. B. Nguyen-Tat, H.-A. Vo, P.-S. Dang, Qmaxvit-UNet+: A query-based maxvit-unet with edge enhancement for scribble-supervised segmentation of medical images, Computers in Biology and Medicine 187 (2025) 109762. doi:10.1016/j.compbiomed.2025.109762.

[6] Y. Li, S. Wang, L. Li, A survey on deep learning for medical image captioning, arXiv preprint arXiv:2303.01151 (2023). doi:10.48550/ARXIV.2303.01151.

[7] H.-C. Shin, K. Roberts, L. Lu, D. Demner-Fushman, J. Yao, R. M. Summers, Learning to read chest x-rays: Recurrent neural cascade model for automated image annotation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2497–2506. doi:10.1109/CVPR.2016.273.

[8] V. Datla, J. Wang, J. M. Sierra, S. Delisle, H. Liu, Intelligent word embeddings of free-text radiology reports, in: Proceedings of the 2014 IEEE International Conference on Healthcare Informatics, IEEE, 2014, pp. 112–119. doi:10.1109/ICHI.2014.20.

[9] X. Wang, Y. Peng, L. Lu, Z. Lu, R. M. Summers, Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 9049–9058.

[10] B. Jing, P. Xie, E. P. Xing, On the automatic generation of medical imaging reports, Proceedings of

the 56th Annual Meeting of the Association for Computational Linguistics (ACL) (2018) 2577–2586. doi:10.18653/v1/P18-1240.

[11] Y. Li, X. Liang, Z. Hu, E. P. Xing, Hybrid retrieval-generation reinforced agent for medical image report generation, in: Advances in neural information processing systems (NeurIPS), volume 31, 2018.

[12] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, ..., G. S. Corrado, Large language models encode clinical knowledge, Nature 620 (2023) 172–180.

[13] 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 (ICML), PMLR, 2022, pp. 12763–12779.

[14] G. Litjens, et al., A survey on deep learning in medical image analysis, Medical image analysis 42 (2017) 60–88. URL: https://www.sciencedirect.com/science/article/abs/pii/S1361841517301135.

[15] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, R. M. Summers, ChestX-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases, in: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), 2017, pp. 2097–2106. doi:10.1109/CVPR.2017.369.

[16] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, X. Tang, Residual attention network for image classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3156–3164. doi:10.1109/CVPR.2017.337.

[17] B. Li, S. Wang, Z. Guo, Y. Liu, D. Liu, Learning to see through the haze: A novel framework for multi-label classification of radiology images, in: 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, 2021, pp. 1162–1167.

[18] 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.

[19] O. Bodenreider, The unified medical language system (umls): integrating biomedical terminology, Nucleic acids research 32 (2004) D267–D270.

[20] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, in: International Conference on Learning Representations (ICLR), 2021.

[21] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019, pp. 4171–4186.

[22] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, A. M. Rush, Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 38–45. URL: https://www.aclweb.org/anthology/2020.emnlp-demos.6.

[23] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, International Conference on Learning Representations (2017).

[24] K. Fang, Z. Wu, Y. Zhao, Y. Wang, M. Wu, Residual attention: A simple but effective method for multi-label recognition, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 15962–15971.

[25] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[26] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE conference on computer vision and pattern recognition, Ieee, 2009, pp. 248–255.

[27] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: Proceedings

of the IEEE International Conference on Computer Vision (ICCV), 2017.
[28] F. Petroni, T. Rocktäschel, S. Riedel, P. Lewis, A. Bakhtin, Y. Wu, A. Miller, Language models as knowledge bases?, in: Conference on Empirical Methods in Natural Language Processing (EMNLP), 2019.
[29] S. Woo, J. Park, J.-Y. Lee, I. S. Kweon, Cbam: Convolutional block attention module, ECCV (2018).

A. Online Resources

To ensure full reproducibility of our results, the source code for both models developed in this study is publicly available on GitHub. The repositories are organized as follows:

- BLIP Fine-tuning,
- MedCSRA.