Augmented Medical Image Captioning: Leveraging Class-Dependent Activation Maps for Enhanced Radiology Report Generation

Vincent Scheltjens
STADIUS Center for Dynamical Systems,
Signal Processing and Data Analytics,
KU Leuven

Lyse Naomi Wamba Momo
STADIUS Center for Dynamical Systems,
Signal Processing and Data Analytics,
KU Leuven

Wouter Verbeke
Faculty of Economics and Business
KU Leuven

Bart De Moor
STADIUS Center for Dynamical Systems,
Signal Processing and Data Analytics,
KU Leuven

Abstract

Radiologists face an increasing workload as a direct result of the growing number of medical images that need interpretation and reporting, especially for Chest X-Rays (CXRs). Therefore, automated medical image captioning has gained attention over recent years, driven by the latest advances in Large Language Models (LLMs). However, existing approaches often face challenges in accurately describing complex medical images, particularly in scenarios with multi-class pathology identification. In this work, we investigate whether, by augmenting the traditional 3-channel raw input image with the Grad-CAM activation maps corresponding to the multi-label classes, more accurate and interpretable reports can be generated. In doing so, we explicitly encode information with respect to the presence or absence of different pathology observations in the input for the image captioning model. A baseline is established based on the latest, state-of-the-art (SOTA), work for CXR image captioning. Specifically, the baseline constitutes a vision encoder-decoder model for which the Convolutional Vision Transformer (CvT) encoder expects a 3-channel input image. In our proposed approach we obtain the activation maps corresponding to all pathological classes for a given input image and enrich the image such that it contains 17 channels. To do so, a multi-label classifier is trained and the CvT is redefined to cope with 17 channels in the input. This work is implemented and evaluated on the MIMIC-CXR dataset for which the preliminary results show, based on the BLEU and Rouge metrics, that the generated reports are more accurate when explicit information on the presence of pathological classes is encoded in the input.

1. Introduction

Automated medical image captioning has emerged as a promising tool for more efficient and streamlined reporting radiological imaging. As the volume and complexity of medical imaging data continue to grow, there is an increasing demand for automated systems that can provide descriptive and actionable insights from these images. Radiology reports, which traditionally rely on textual descriptions provided by expert radiologists, often serve as critical documents guiding clinical decision-making processes. However, the manual generation of these reports can be time-consuming, labor-intensive, and prone to variability, especially in the context of multi-class pathology identification [5, 15, 19].

In recent years, deep learning techniques have shown success in various computer vision tasks, including image captioning [21, 26–28]. By learning to generate natural language descriptions directly from visual inputs, these models have the potential to automate and streamline the radiology reporting process. However, adapting image captioning models to the domain of medical imaging presents unique challenges. Medical images often exhibit complex anatomical structures and pathological manifestations, requiring a deeper understanding of the underlying visual features to generate accurate and clinically relevant captions [18].

In this work, we investigate and propose a method for enhanced medical image captioning by augmenting raw input images with class-dependent activation maps obtained from a multi-class classifier. Specifically, we leverage a

pre-trained (Densenet-121) [7] classifier to extract activation maps corresponding to 14 distinct pathological classes, enriching the visual representation with 14 additional channels. By incorporating these class-specific features into the image captioning pipeline, our approach aims to improve the quality and specificity of generated captions. In future work, we will evaluate whether, as a result of the richer latent space, better alignment between the visual features and generated text can be observed.

Based on the preliminary results reported in this work, we show that by explicitly encoding class-dependent information in the input space, the model learns to generate more accurate reports compared to the baseline model.

2. Related works

In recent years, various deep learning architectures, including Convolutional Vision Transformers (CvTs) [25], Vision Transformers (ViTs) [4], and Self-Attention Transformers (SAT) have been applied to medical image captioning tasks, particularly in the context of chest X-ray (CXR) data analysis [1, 2, 12, 16]. CvTs, which combine the benefits of convolutional neural networks (CNNs) and transformers, have shown promise in capturing both local and global features in medical images [4]. Similarly, ViTs leverage self-attention mechanisms to process visual inputs and have been applied successfully in medical imaging tasks [20].

Despite these advancements, existing approaches to medical image captioning often rely solely on raw input images and pretrained models for the extraction of relevant local and global features. Rather than relying only on the feature extraction, we propose to learn interpretable features that explicitly encode information on the presence or absence of certain pathological classes.

In addition to architecture-specific approaches, various metrics and evaluation methodologies have been proposed to assess the performance of medical image captioning models. Commonly used metrics include BLEU (Bilingual Evaluation Understudy) [13], Rouge (Recall-Oriented Understudy for Gisting Evaluation) [10], and CIDEr (Consensus-based Image Description Evaluation) [22], which measure the similarity between generated captions and ground truth annotations. Furthermore, datasets such as the IU X-ray [14] dataset and the MIMIC-CXR [9] dataset provide valuable resources for training and benchmarking medical image captioning models, offering diverse and annotated chest X-ray images along with corresponding radiological reports.

Overall, while significant progress has been made in the field of medical image captioning, there remains a need for more robust and interpretable models, particularly in the context of CXR data analysis. By exploring novel techniques such as augmenting input images with activation maps, we aim to improve performance and enhance interpretability.



Findings: A hazy opacity is present in the right lung which may represent aspiration, pleural effusion or hemorrhage. Retrocardiac opacity at the left base is unchanged. Moderate cardiomegaly is stable. Slight prominence of the pulmonary vasculature with cephalization and enlarged pulmonary arteries are consistent with mild pulmonary edema. Tracheostomy tube is in place. There are no displaced rib fractures.

Impression: 1. Hazy opacity in the right lung which may represent aspiration versus pleural effusion or hemorrhage. 2. Mild pulmonary edema. 3. No displaced rib fractures.

Figure 1. Example of an original, raw CXR image in the dataset with the corresponding findings and impression sections from the radiology report.

3. Methods

3.1. Data

The MIMIC-CXR (Medical Information Mart for Intensive Care - Chest X-ray) dataset [6, 9] is a publicly available database specifically for medical imaging research. The dataset contains a comprehensive collection of chest X-ray images accompanied by associated radiological reports and clinical metadata. With 377 110 CXR images from 227 835 unique studies, corresponding to 65 379 patients, the dataset provides a rich source of data for training and evaluating machine learning models in various medical imaging tasks. Each image is linked to one or more radiological reports, providing detailed textual descriptions of observed findings. In addition, the dataset includes clinical metadata for each patient, enabling research into critical care medicine and healthcare informatics. The meta data for the dataset contains a predefined train, validate and test split to which we adhere in this work. The MIMIC-CXR dataset, as indicated, is multi-modal, containing both images and text. For this work, we extract a single image per study and the corresponding written report. The written report is processed such that only the cleaned *findings* and *impression* sections remain, which are clinically of most relevance. Figure 1, depicts an example of an extracted, preprocessed, image-text pair.

3.2. Baseline

The baseline model defined in this work constitutes a custom vision encoder-decoder model that inherits from the vision encoder decoder model as defined in the Huggingface transformer package [24]. Similar to the work in [12], the encoder is comprised of a CvT that was pretrained on ImageNet-22K and later fine-tuned on ImageNet-1K [25]. On the last hidden state, layer normalization is applied prior to projecting the output to the hidden size of the decoder.

The decoder is subsequently defined as a byte-level byte

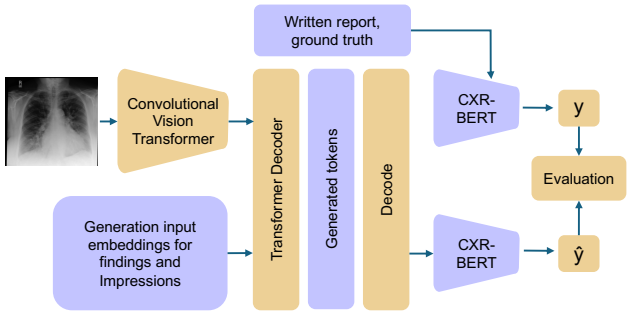


Figure 2. Model overview for the baseline vision encoder-decoder model. This implementation is adapted from the work presented in [12].

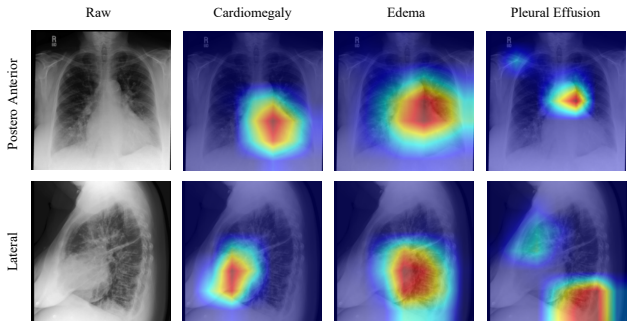


Figure 3. Exemplary condensed overview of the Grad-CAM activation maps obtained for 3 prominent pathology classes. The activation maps are plotted on top of the raw image for interpretation. In the input, the activation maps present a single channel image per class with the corresponding per pixel activations.

pair tokenizer [23] that we train on the impression and findings sections from the written reports in the training set. To this extent, the vocabulary size is set to 30 000 and tokens are supplied to the transformer decoder, initialized with 6 hidden layers and a language model head with the same vocabulary size. In Figure 2, the holistic reference architecture is shown, throughout the remainder of this work we refer to this model as the baseline. Following the encoder-decoder section of the model, both the generated report as well as the ground truth report are tokenized using the pretrained CXR-BERT [3] model. The tokenized results are passed to the evaluation step, which we describe in subsection 3.4.

As shown in 2, the encoder takes in the raw input images. Therefore, we preprocess, normalize and standardize the input with the statistics from the pretrained encoder such that the input fits the dimensions expected by the encoder $\mathbb{R}^{3\times384\times384}$.

3.3. Class-dependent image augmentation

To allow for augmented medical image captioning with class-dependent activation maps (AMICCAM), the baseline architecture shown in Figure 2, is altered such that the raw input image first passes through a classifier from which the per-class activation maps can be obtained. Subsequently, the raw input image is concatenated with the activation maps and passed to the vision encoder-decoder model.

3.3.1 Classifier

To obtain class-dependent activation maps, a single deep convolutional neural network, DenseNet-121 [7], is trained on the train set. The last layer of the model corresponds to the classification layer which is defined to yield the predictions for each of the 14 possible class labels.

The data is labeled such that for each observation, a corresponding vector of length 14 exists. Each entry in the vector constitutes the indication of whether the corresponding class is observed (1), not observed (0), or it is uncertain (-1) as to whether the pathology is present. Following the work presented in [8], the uncertainty labels are dealt with by replacing (-1) with (1), which has been shown to yield better predictive performance.

As such, the classifier is trained on this multi-label classification task until convergence and is subsequently used as the pretrained classifier to obtain per class activation maps for any given image in the training set for the vision encoder-decoder pipeline. In Table 1, a condensed overview of the performance for the classifier is shown on a selection of the most important pathological classes.

3.3.2 AMICCAM vision encoder-decoder model

Leveraging the trained classifier, the Gradient-weighted Class Activation Maps (Grad-CAMs) [17] can be computed for any given target class and any given input image. Essentially, Grad-CAM yields visual explanations with respect to the important regions in an image for a corresponding prediction. These visualizations, or activation maps, portray which areas of an image the model attributes importance to, with respect to a given outcome. In Figure 3, a condensed visual representation is shown in which a subset of the 14 activation maps for a specific image are plotted on top of the raw input, showing the activated regions for the corresponding class.

Now that we can obtain activation maps, we modify the encoder-decoder model to incorporate all 14 activation maps for each input. These maps are then concatenated to the preprocessed image, resulting in an input with dimension $\mathbb{R}^{17\times384\times384}$. It is worth noting that this input structure deviates from the dimensions expected by the CvT encoder. To address this, we extend the CvT model by introducing three sequential convolutional 2D layers as the input layer. These layers accommodate an input with 17 channels and project it to the original encoder input dimensions.

Model	Atelectasis	Cardiomegaly	Consolidation	Edema	Pleural Effusion
DenseNet-121	0.77 ± 0.01	0.77 ± 0.01	0.74 ± 0.00	0.84 ± 0.00	0.88 ± 0.01

Table 1. Classifier performance expressed in terms of area under the curve (AUC) scores for a selection of classes.

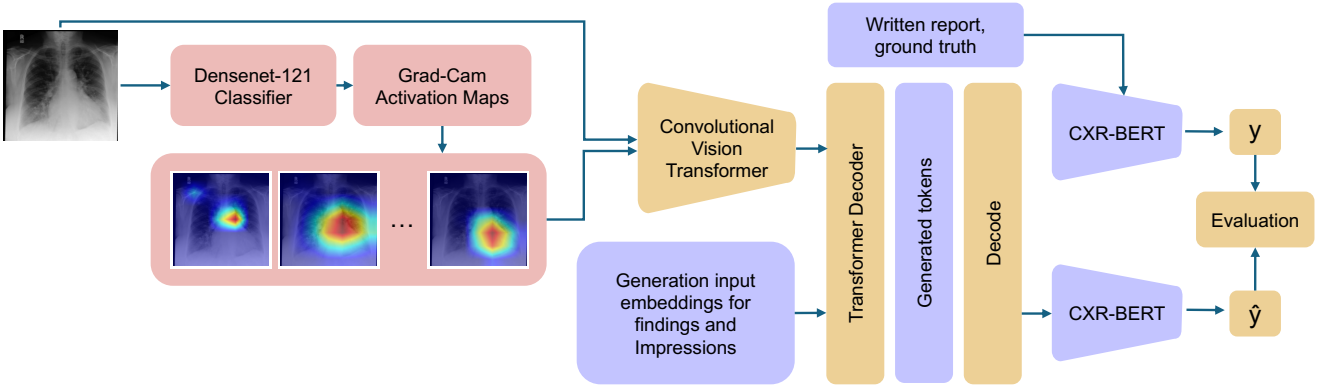


Figure 4. Proposed approach where the raw input is first augmented by obtaining the per class activation maps from the trained classifier

Given that these additional layers cannot leverage any of the pretrained weights, they are trained from scratch during the training procedure.

3.4. Evaluation

The generated reports are evaluated based on the BLEU [11, 13] and ROUGE [10] metrics, which are commonly used for evaluating the quality of generated text against reference text in natural language processing tasks. Together, these metrics provide a quantitative measure of how well the generated text matches the reference text in terms of precision and recall.

3.5. Training

Both the baseline and AMICCAM models are trained for 32 epochs with a mini-batch size of 16 and a learning rate of $5e-5$ for both the encoder and decoder models on $1\times$80GB NVIDIA A100 GPU. Checkpoints were saved every 100 steps such that, based on the validation metrics, the optimal checkpoint could be selected for inference and testing. At training time, the maximum length for the tokenized representation of the reports is limited to 256 tokens. Note that for this preliminary work, training was performed on a subset containing only 10% of the full dataset to verify whether the proposed approach shows improvement in terms of the reported metrics compared to the baseline, in future work, the full dataset will be considered for training, validation and testing.

Approach	BLEU	ROUGE-1	ROUGE-2	ROUGE-L
Baseline	9.62	36.12	16.30	31.17
AMICCAM	9.82	37.45	17.85	33.77

Table 2. Preliminary results for the BLEU, ROUGE-1, ROUGE-2 and ROUGE-L metrics. Results were obtained by running on a subset of the data only and are therefore only comparable relative to each other, not yet to the current SOTA.

4. Results and discussion

As indicated, the results reported in Table 2 are based on a training procedure that had access to 10% of the total dataset only. Therefore, the results are not comparable directly to the SOTA, nevertheless the results can be interpreted relative to each other. To that extent, we can compare the preliminary performance for the baseline with that of AMICCAM, the proposed approach. The results in 2 show that our proposed method, in which information with respect to the presence or absence of specific pathological observations is explicitly encoded in the input, improves the generated reports in terms of BLEU, ROUGE-1, ROUGE-2 and ROUGE-L, compared to the baseline. These initial results provide a first validation of the proposed approach, in subsequent work this will be further validated.

5. Conclusion and future work

The results reported herein are obtained based on a subset of the data only, therefore the results reported here are not yet comparable with the current SOTA. Nevertheless, the results do show improved performance resulting from explic-

itly encoding pathology-level information in the input for the image captioning model. In ongoing work, we extend this pipeline for training on the entire MIMIC-CXR dataset such that we can compare to the SOTA. In the extension, we no longer focus on a single image per study, but consider up to 10 images per study, if available. In addition, we investigate whether, apart from improved report generation, there is also better alignment between visual features and the generated text as a result of the enriched latent space learned in the encoder part of the captioning model. As an extension to this work, the proposed image captioning approach will be implemented in the distributed setting where, through federated learning, we will deal with privacy concerns and show how, practically, such complex models can be learnt over different silos, mitigating issues that would otherwise relate to data aggregation for training such complex models in a privacy sensitive environment.

References

[1] Omar Alfarghaly, Rana Khaled, Abeer Elkorany, Maha Helal, and Aly Fahmy. Automated radiology report generation using conditioned transformers. *Informatics in Medicine Unlocked*, 24:100557, 2021. 2

[2] William Boag, Tzu-Ming Harry Hsu, Matthew Mcdermott, Gabriela Berner, Emily Alesentzer, and Peter Szolovits. Baselines for Chest X-Ray Report Generation. In *Proceedings of the Machine Learning for Health NeurIPS Workshop*, pages 126–140. PMLR, 2020. 2

[3] Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C. Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, Hoifung Poon, and Ozan Oktay. Making the most of text semantics to improve biomedical vision-language processing, 2022. 3

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. 2

[5] Dhakshinamoorthy Ganeshan, Andrew B Rosenkrantz, Roland L Bassett Jr, Lori Williams, Leon Lenchik, and Wei Yang. Burnout in academic radiologists in the united states. *Academic radiology*, 27(9):1274–1281, 2020. 1

[6] A L Goldberger, L A Amaral, L Glass, J M Hausdorff, P C Ivanov, R G Mark, J E Mietus, G B Moody, C K Peng, and H E Stanley. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation*, 101(23):E215–20, 2000. 2

[7] Gao Huang, Zhuang Liu, and Kilian Q. Weinberger. Densely connected convolutional networks. *CoRR*, abs/1608.06993, 2016. 2, 3

[8] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Christopher Chute, Henrik Marklund, Behzad Haghgoo, Robyn L. Ball, Katie S. Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *CoRR*, abs/1901.07031, 2019. 3

[9] Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. MIMIC-CXR: A large publicly available database of labeled chest radiographs. *CoRR*, abs/1901.07042, 2019. 2

[10] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, 2004. Association for Computational Linguistics. 2, 4

[11] Chin-Yew Lin and Franz Josef Och. ORANGE: a method for evaluating automatic evaluation metrics for machine translation. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 501–507, Geneva, Switzerland, 2004. COLING. 4

[12] Aaron Nicolson, Jason Dowling, and Bevan Koopman. Longitudinal data and a semantic similarity reward for chest x-ray report generation, 2024. 2, 3

[13] Kishore Papineni, Salim Roukos, Todd Ward, and Wei jing Zhu. Bleu: a method for automatic evaluation of machine translation. pages 311–318, 2002. 2, 4

[14] John Pavlopoulos, Vasiliki Kougia, and Ion Androutsopoulos. A survey on biomedical image captioning. In *Proceedings of the Second Workshop on Shortcomings in Vision and Language*, pages 26–36, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. 2

[15] Bruce I Reiner, Nancy Knight, and Eliot L Siegel. Radiology reporting, past, present, and future: the radiologist's perspective. *Journal of the American College of Radiology*, 4(5):313–319, 2007. 1

[16] Alexander Selivanov, Oleg Y. Rogov, Daniil Chesakov, Artem Shelmanov, Irina Fedulova, and Dmitry V. Dylov. Medical image captioning via generative pretrained transformers, 2022. 2

[17] Ramprasaath R. Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. *CoRR*, abs/1610.02391, 2016. 3

[18] Harsh Sharma, Jai Sethia Jain, Priti Bansal, and Sumit Gupta. Feature extraction and classification of chest x-ray images using cnn to detect pneumonia. In *2020 10th international conference on cloud computing, data science & engineering (Confluence)*, pages 227–231. IEEE, 2020. 1

[19] Andrew F. Simon, John H. Holmes, and Erin Simon Schwartz. Decreasing radiologist burnout through informatics-based solutions. *Clinical Imaging*, 59(2):167–171, 2020. 1

[20] Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli, Giuseppe Fiameni, and Rita Cucchiara. From show to tell: A survey on deep learning-based image captioning, 2021. 2

[21] Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel

Shu Wei Ting. Large language models in medicine. *Nature medicine*, 29(8):1930–1940, 2023. 1

[22] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation, 2015. 2

[23] Changhan Wang, Kyunghyun Cho, and Jiatao Gu. Neural machine translation with byte-level subwords. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05): 9154–9160, 2020. 3

[24] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Huggingface's transformers: State-of-the-art natural language processing, 2020. 2

[25] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers, 2021. 2

[26] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5579–5588, 2021. 1

[27] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.

[28] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 1