# Medical Image Captioning: A Comparative Study of Encoder Models

1st Luis Eduardo Robles Jiménez
*Facultad de Ingeniería*
*Universidad Panamericana*
Aguascalientes, México
0009-0003-4620-7306

2nd Francisco Alfredo Castrellón Carrillo
*Facultad de Ingeniería*
*Universidad Panamericana*
Aguascalientes, Mexico
0009-0004-5890-179X

3rd Maria Teresa Orvananos Guerrero
*Facultad de Ingeniería*
*Universidad Panamericana*
Aguascalientes, Mexico
torvananos@ags.up.mx

4th Ricardo Abel Espinosa Loera
*Facultad de Ingeniería*
*Universidad Panamericana*
Aguascalientes, Mexico
respinosa@up.edu.mx

5th Franz Rivera Tellez
*Facultad de Ingeniería*
*Universidad Panamericana*
Aguascalientes, Mexico
0000-0001-6836-1409

6th Luis Fernando Caro Reyna
*Facultad de Ingeniería*
*Universidad Panamericana*
Aguascalientes, Mexico
0000-0001-6765-7726

***Abstract*—This paper addresses the task of medical image captioning, which aims to generate accurate text descriptions for medical images. The importance of this task lies in its potential to assist in medical diagnosis, particularly in cases where the complexity of images requires expert analysis. The paper explores the use of two different feature extraction approaches, namely ResNet-50 and ResNet-101, and compares their performance in terms of training resources. The dataset used is the 2023 ImageCLEF competition dataset, and the evaluation metrics include training time, number of parameters, and BERT F1 score. The results demonstrate that both ResNet-50 and ResNet-101 produce coherent captions, with ResNet-101 requiring slightly more resources but yielding comparable performance. This study provides insights into choosing an appropriate feature extractor for medical image captioning tasks, considering hardware limitations.**

***Index Terms*—medical image captioning, encoder, decoder, attention, imageCLEF, ResNet, comparative study**

## I. Introduction

Image captioning is the task of automatically generating an accurate text description based on an image. Among its multiple uses, generating medical diagnosis using images has become one of the most important nowadays. Medical images are essential tools for clinical diagnosis since they allow doctors to diagnose various medical conditions; however, because of the complexity of these images, it requires experienced professionals to deliver a correct diagnosis and it's a very time consuming task. X-ray imaging is one of the most used methods because of its affordable and accessible nature and it is used to diagnose not only broken bones but a wide variety of lung diseases, making it hard to generalize exactly what to look for at first, since different diseases may have similar effects with very subtle differentiators, and this problem is common among medical imagery. Thus, there is a current need for a tool that reliably guides doctors in the areas of images they should focus on, as well as in interpreting and summarizing possible disease symptoms based on images; so, by fighting the world's growing shortage of medical doctors, the process could become more efficient and effective, more people could be diagnosed, and ultimately more illnesses could be treated. Because of the importance of this problem, there has been a lot of studies and approaches to develop tools that help professionals with this task [7, 15, 16, 17, 19]. From simple image segmentation to detect anomalous sections of an image to deep learning models to predict illnesses based on images; the specific application covered in this paper is automatic caption generation for medical images[2, 3, 4, 10], which is a core technology for automated diagnosis where the goal is to generate accurate, relevant and coherent descriptions of the images using medical terminology to provide doctors valuable information in a quickly manner. In general, these systems work by extracting features from images with a specific method so they can be later used to generate captions with a text-generation technique. This is a big challenge because of the nature of this area, since we are diagnosing people, the terminology needs to be very precise and the description coherent so it does not confuse the user or cause an incorrect diagnosis. Nowadays due to the success of deep learning in image detection and text generation it is being researched as a solution to this problem. This paper is intended to be a participating approach in the image captioning task of the 2023 ImageCLEF campaign. And the contributions are:

- Exploration of two different feature extraction approaches to compare their results against their demand of training resources.
- Train an existing model with a different dataset to see how it behaves with medical imagery.

## II. Related Work

Recently, multiple encoder-decoder systems have been implemented. Usually, Convolutional Neural Networks (CNN) or other type of Neural Networks perform encoder tasks to extract visual information as features, and Recurrent Neural Networks

(RNN) or other sequence Neural Network are used as decoders to generate the final text like Kelvin Xu demonstrated in [20]. Other works use RNNs to transform extracted features into shape-related information and then use long short term memory (LSTM) to get the final description [7]. Currently this is the basis for medical image captioning, but there are many works that by introducing different modifications achieve better results in specific areas, for example, Yuxuan Xiong et al. [19] in "Reinforced Transformer for Medical image Captioning" proposes a hierarchical transformer based model consisting on an image encoder that extracts heuristic visual features by a bottom-up attention mechanism and a non-recurrent captioning decoder that improves computational efficiency by parallel computing and achieving a BLEU-1 performance increase of 50% compared contemporary methods. Other state-of-the-art methods use attention and transformer models [15, 16, 14, 11, 5, 12, 20, 19], like "Medical image captioning via generative pretrained transformers" by Alexander Selivanov et al. [16] where they use the Show-Attend-Tell model and the GPT-3 model to generate a textual summary of radiology records containing essential information about the pathologies found, the location and 2Dheatmaps that localize each pathology and with that combination of language models they outperformed other models, introducing a new preprocessing pipeline that allows the obtention of higher metrics.

## III. Methodology

This task is intuitively approached by a decoder-encoder architecture [1, 13], where important features are first extracted from the images and then some coherent text is produced based on the processed data. Said Architecture is a framework commonly used in deep learning for various tasks, including sequence-to-sequence tasks such as machine translation, image captioning, text summarizing, and speech recognition. It was first used in by Ilya Sutskever in [18] and has been used for those different problems ever since. The encoder is responsible for processing the input data and capturing its relevant features or representations. It takes the input sequence, such as a sentence or an image, and applies a series of transformations to encode the information into a fixed-size representation or a context vector. The decoder, on the other hand, takes the encoded representation produced by the encoder and generates the output sequence. It operates in a step-by-step manner, where at each step, it predicts the next element of the output sequence based on the previous predictions and the context vector.

### A. Encoder

The Encoder part of the architecture consist in translating the input, which in this case consists of an image, into a consistent abstract representation. This task can be completed in many different ways. Some of the first approaches consisted on manual feature extraction by prepossessing the images with different image analysis techniques, but most of the modern approaches consist of using Neural Networks to automatically extract the features, which has proven effective enough to become the current state-of-the-art. The chosen architecture for this task was presented by Microsoft engineers in the paper Deep Residual Learning for Image Recognition [8] which is also known as ResNet and provides the ability to choose different depths. It addresses the problem of vanishing gradients and degradation in network performance in very deep neural networks training. ResNet accomplishes that by introducing residual connections that allow the output from a previous layer to be directly passed to a later layer. This means that the network can learn residual functions, i.e., the difference between the desired output and the current output, instead of learning the complete transformation at each layer. This enables the training of very deep neural networks with hundreds or even thousands of layers. This is because the residual connections mitigate the vanishing gradient problem, allowing the gradients to flow more easily during backpropagation. As a result, the network can learn more complex representations and achieve better accuracy. In this paper is analysed as a proof of concept of whether a deeper neural network is worth the resources, offering a comparison between ResNet-50 and ResNet-101, which are some of many ResNet variations.

### B. Decoder

For the decoder part we incorporate word embedding features along with sinusoidal positional embedding. These combined features, along with the additional information and encoder output features, serve as input for the decoder. The decoder itself consists of multiple identical Nd layers, with each layer comprising a mask multi-head self-attention sub-layer, a multi-head cross-attention sub-layer, and a positional feed-forward sub-layer arranged in that order. The output feature from the last decoder layer is passed through a linear layer, which has the same output dimension as the vocabulary size. This linear layer helps determine the meaning of the next word. To train the model we minimize the subsequent cross-entropy loss given the ground truth sentence $y^*_{1:T}$ and the prediction $y^*_t$ with the parameter $O$

$$L_{XE}(O) = -\sum_{t=1}^{T} log(p_O(y^*_t|y^*_{1:T})) \tag{1}$$

To perform word embedding in the decoding stage, we employed the BERT model from the Hugging Face library. Similar to the encoding stage, our decoding process utilized a multi-head self-attention sublayer, a multi-head cross-attention sublayer, and a positional feed-forward sublayer to fine-tune our model.

## IV. Experiments

As it was mentioned before, this work pretends to provide an insight on which neural network should be chosen when the hardware resources are limited. That being said, the experiments were performed as described in the following.

### A. Dataset

The model was trained using the dataset provided for the 2023 ImageCLEF competition, which consists of 60,918 training images, 10,473 for validation and 10,473 for testing. However, as this is a current event, there's no access to the captions corresponding to the test set, thus, the validation set was trimmed 50% of its content in order to make our own test set and compare if the predicted sentences are similar to the reference ones.

Such dataset was preprocessed with three main operations listed below:

- Removal of punctuation
- Removal of single characters
- Removal of numeric symbols

## V. Evaluation

The proposals will be compared under several parameters, like training time, performance metrics and number of parameters. With this information is intended to provide guidance when choosing what feature extractor should be selected for a specific task.

### A. Metrics

BERT [6] is the main metric proposed for the competition in order to test whether a predicted sentence matches in meaning an reference caption. For this evaluation the native python implementation of BERT is used with the model type: microsoft/deberta-xlarge-mnli [9].

Due to the fact that this dataset is not public yet, it's not possible to compare with other achieved results in the state-of-the-art. Therefore, the metrics are reported as a comparison of the two proposed models.

### B. Quantitative analysis

The results of the multiple experiments are the following:

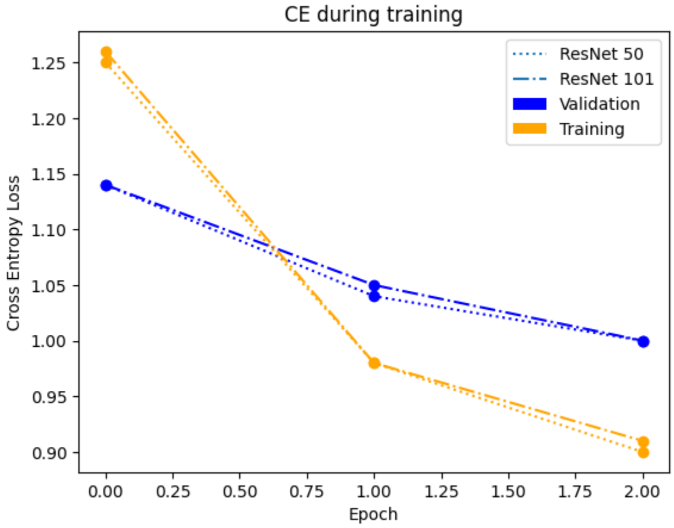


Fig. 1. Criterion throughout the training

| Encoder Model | Minutes per epoch | # of parameters | BERT F1 |
|---|---|---|---|
| ResNet-50 @ 3 | 197 | 120M | 0.5949 |
| ResNet-101 @ 3 | 244 | 138M | 0.6111 |

TABLE I
Quantitative Analysis of the ResNet models

As we can see the results with both architectures were very similar, so the main differential factor would be the computing requirements difference between them.

### C. Qualitative analysis

In this regard, the model came up with decent results, despite the fact that the hardware was a big limitation, in a few epochs the models were able to generate coherent captions to a point that the medical language did not seem to be a problem. Instead, it's more notorious a lack of concept detection, which was expected.

After an analysis, it's possible to see that the model is able to make understandable sentences and some of them are accurate regarding the content of an image. For example: see table II.

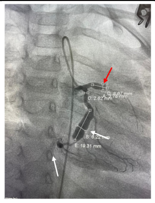
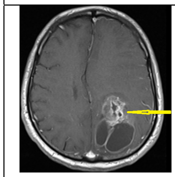
| Image | Reference | Predicted |
|---|---|---|
|  | Coronary angiography. Multiple coronary aneurysms on the anterior descending artery (stop flow, thrombosis process in progress: red arrow) and on the circumflex artery (white arrows). | the left coronary angiogram shows the left coronary artery ( arrow ). |
|  | brain magnetic resonance imaging (MRI) shows brain metastasis (arrow) | mri brain showing a large mass in the left frontal lobe. |

TABLE II
Images where the predicted captions were right using ResNet 50

And there are cases where the captions are understandable but the medical report is not exactly right: see table III.

## VI. Conclusion

The proposed model seems to be a good option for the task, since it is able to generate decent results from the very first epochs. However, based on the resources consumption, it's recommended to explore lighter approaches if the hardware is a problem when trying to reproduce the reported results, but as future work it would be interesting to explore the project with more computing power to truly see the potential of the architecture.

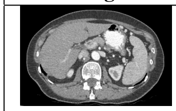
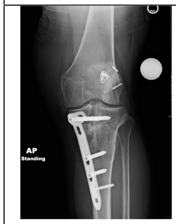
| Image | Reference | Predicted |
|---|---|---|
|  | Posttreatment CT showing no residual pancreatic tumor | ct scan of the abdomen showing a large mass in the right kidney. |
|  | An AP radiograph showing the healed osteotomy | x - ray of the left knee showing a well - defined radiolucent lesion in the left femur. |

TABLE III
IMAGES WHERE THE PREDICTED CAPTIONS WERE NOT RIGHT USING RESNET 50 BUT SEMANTICALLY CORRECT

Also, based on the analysis, it's possible to conclude that if one tries to improve the results of this model, the first step should not be to change the backbone of the encoder model (ResNet), since, in this scenario there was no big difference between ResNet-50 and ResNet-101 but the execution time. So, a greater improvement might be achieved if another stage of the model is changed.

## References

[1] Kyle Aitken et al. *Understanding How Encoder-Decoder Architectures Attend*. 2021. arXiv: 2110.15253v1 [cs.CL].

[2] Peter Anderson et al. *Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering*. 2018. arXiv: 1707.07998v3 [cs.CL].

[3] Raffaella Bernardi et al. *Automatic Description Generation from Images: A Survey of Models, Datasets, and Evaluation Measures*. 2017. arXiv: 1601.03896v2 [cs.CL].

[4] Hanting Chen et al. *Pre-Trained Image Processing Transformer*. 2021. arXiv: 2012.00364v4 [cs.CL].

[5] Marcella Cornia et al. *Meshed-Memory Transformer for Image Captioning*. 2020. arXiv: 1912.08226 [cs.CV].

[6] Jacob Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2019. arXiv: 1810.04805 [cs.CL].

[7] Yue Zhang Haoran Wang and Xiaosheng Yu. "An Overview of Image Caption Generation Methods". In: *Hindawi* (2020).

[8] Kaiming He et al. "Deep Residual Learning for Image Recognition". In: *CoRR* abs/1512.03385 (2015). arXiv: 1512.03385. URL: http://arxiv.org/abs/1512.03385.

[9] Pengcheng He et al. *DeBERTa: Decoding-enhanced BERT with Disentangled Attention*. 2021. arXiv: 2006.03654 [cs.CL].

[10] Simao Herdade et al. *Image Captioning: Transforming Objects into Words*. 2020. arXiv: 1906.05963v2 [cs.CL].

[11] Jia Cheng Hu, Roberto Cavicchioli, and Alessandro Capotondi. *ExpansionNet v2: Block Static Expansion in fast end to end training for Image Captioning*. 2022. arXiv: 2208.06551 [cs.CV].

[12] Chenliang Li et al. *mPLUG: Effective and Efficient Vision-Language Learning by Cross-modal Skip-connections*. 2022. arXiv: 2205.12005 [cs.CL].

[13] Wei Liu et al. *CPTR: FULL TRANSFORMER NETWORK FOR IMAGE CAPTIONING*. 2021. arXiv: 2101.10804v3 [cs.CL].

[14] Ron Mokady, Amir Hertz, and Amit H. Bermano. *ClipCap: CLIP Prefix for Image Captioning*. 2021. arXiv: 2111.09734 [cs.CV].

[15] Md Aminul Haque Palash et al. *Bangla Image Caption Generation through CNN-Transformer based Encoder-Decoder Network*. 2021. arXiv: 2110.12442 [cs.CV].

[16] Alexander Selivanov et al. "Medical image captioning via generative pretrained transformers". In: *Scientific Reports* 13.1 (2023), p. 4171.

[17] Faisal Shah et al. "Bornon: Bengali Image Captioning with Transformer-Based Deep Learning Approach". In: *SN Computer Science* 3 (Jan. 2022). DOI: 10.1007/s42979-021-00975-0.

[18] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. *Sequence to Sequence Learning with Neural Networks*. 2014. arXiv: 1409.3215v3 [cs.CL].

[19] Yuxuan Xiong, Bo Du, and Pingkun Yan. "Reinforced transformer for medical image captioning". In: *Machine Learning in Medical Imaging: 10th International Workshop, MLMI 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13, 2019, Proceedings 10*. Springer. 2019, pp. 673–680.

[20] Kelvin Xu et al. *Show, Attend and Tell: Neural Image Caption Generation with Visual Attention*. 2016. arXiv: 1502.03044 [cs.LG].