Concept Detection and Image Caption Generation in Medical Imaging

Notebook for the ImageCLEFmedical Caption Lab at CLEF 2023

Varsha Yeshwanth[1,*], Pranith P[1,*] and Lekshmi Kalinathan[1,*]

[1] *Sri Sivasubramaniya Nadar (SSN) College of Engineering, Chennai, India*

Abstract

Medical images help in the detection, diagnosis and prognosis of diseases. Extracting accurate insights from them is an indispensable factor in providing reliable healthcare. Currently, this task is performed manually, making it time consuming and labour intensive. The use of deep learning can automate this process of extracting concepts from medical images and turning them into reports. In this paper we propose two methods for the same as a part of the ImageCLEFmedical task. For concept detection the DenseNet121 architecture was used, and it achieved an F1 score of 0.49. We present a CNN encoder with an LSTM based decoder for caption prediction, with the model achieving a BERTScore 0.58.

Keywords

Multi-label classification, Image caption prediction, concept detection, LSTM, DenseNet121, CNN, ResNet-101

1. Introduction

Medical image analysis and interpretation is an important task perturbing healthcare professionals. Extracting insights from radiology scans and developing accurate interpretations of them is a crucial phase in the diagnosis process. This process is heavily dependent on human technicians who still physically carry out the analysis, which is both time consuming and labour intensive. This is a hindrance and calls for the development of automated techniques that can identify the major concepts in the scans and also translate them into condensed textual descriptions without compromising on the efficacy of the results. After all, the diagnosis can only be as good as the data.

This paper describes clef-CSE-Gan-Team's approach to the ImageCLEFmedical Caption task, which can broadly be classified into two, concept detection and caption prediction. The concept detection subtask is aimed at identifying the UMLS (Unified Medical Language System)[1] concepts embodied inside the image. UMLS concepts are unique identifiers assigned to various medical and health-related terms. This standardised vocabulary serves as a tool for effective communication and information exchange between healthcare professionals across various disciplines. Hence, it is important to accurately identify all the concepts present in an image,

CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece

*Corresponding author.

varshayeshwanth@gmail.com (V. Yeshwanth); pranithreddy13@gmail.com (P. P); lekshmik@ssn.edu.in (L. Kalinathan)

especially because this information can also be used in subsequent automation tasks like caption prediction or computerised prognosis.

The next subtask deals with the generation of condensed textual captions for the original images. Human-readable captions are essential to help medical professionals understand the insights contained in an image and to provide them with all the information they need to create a successful treatment plan.

2. Proposed Methodology

The following sections describe the methods used for the subtasks and the intuition behind their usage.

2.1. Concept Detection

The ImageCLEFmedical 2023 [2] Caption dataset [3] consisted of 60,918 training samples and had 2125 unique labels. Owing to the diversity and magnitude of data, we decided to deploy a deep neural network, specifically DenseNet, for the task. Figure 1 shows an overview of its architecture. DenseNets[4] or Densely Connected CNNs solve the vanishing gradient problem of traditional CNNs by connecting every layer in the network with each other, increasing feature reusability and flow of information. The model has L(L+1)/2 direct connections for L layers. DenseNets also have fewer parameters since the feature maps from previous layers are concatenated with those from subsequent layers instead of being summed. This makes them ideal candidates for tasks like multi-label classification, where the dimensionality of the label space is high. Furthermore, feature concatenation enables the model to access features at both local and global resolutions. This is advantageous for multi-label classification since multiple labels can be identified at the same time. The structure of DenseNet model is split into dense blocks. The feature map's dimensionality remains constant inside each block, while the number of filters varies. Transition layers are used between the blocks to reduce the number of channels. The last layer is a bottleneck layer to improve efficiency.

2.2. Caption Prediction

The dataset contained 60,918 training samples, each with an associated medical caption. The model proposed generates the caption word by word. This is thanks to the attention mechanism [5], which allows it to focus only on those parts of the image that are relevant to the next word it is going to produce. The model itself is an encoder-decoder architecture, as shown in Figure 2, combined with a beam search algorithm in the final phase. An encoder is used to convert all input images into a fixed-coded sequence. For the encoder, a ResNet-101 [6] [7] has been used because of its deep architecture, skip connections that help to learn the residual mappings, and the ability of the model to generalise well, thanks to the pretraining done on the ImageNet dataset. The encoder is needed to understand the semantics, extract the features, and also help in capturing the spatial and contextual information. A decoder is then used to construct a natural language representation by transforming the encoded information. Its job is to generate captions word by word by looking at the encoded images and paying attention to the important aspects.

Beam search used in the decoding process involves considering the top k candidates in the first step and generating k second words for each. Combinations of the first and second words are then chosen based on additive scores. This process is repeated for subsequent words, selecting the top k combinations. At each step, the best overall score is used to determine the terminating sequence among the k sequences generated. This helps in choosing the sequence of words with the highest overall score from a group of candidate sequences instead of letting the decoder rigidly choose the best word alone at each stage.

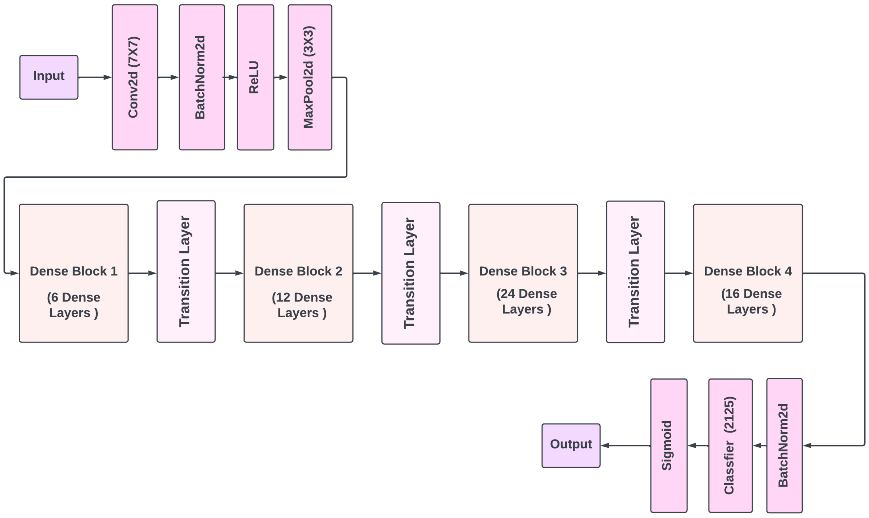


Figure 1: The DenseNet121 architecture.

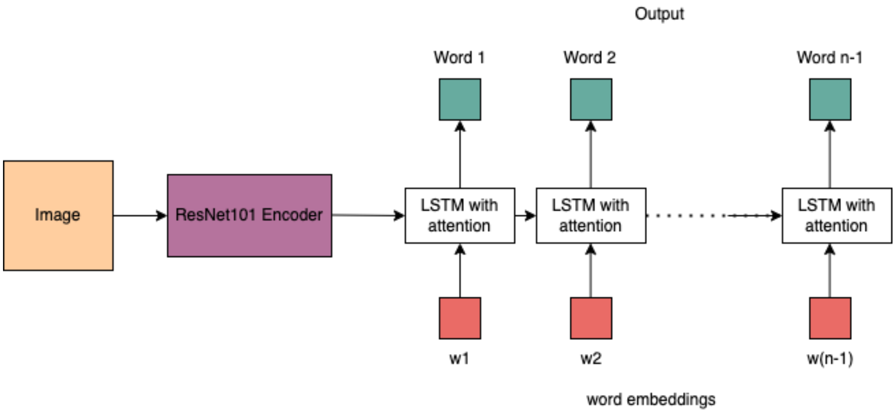


Figure 2: The encoder-decoder architecture for caption generation

3. Implementation

For both the subtasks, the PyTorch machine learning framework was used for the implementation. The code used, can be accessed at https://github.com/clefVP/medical-image-captioning.

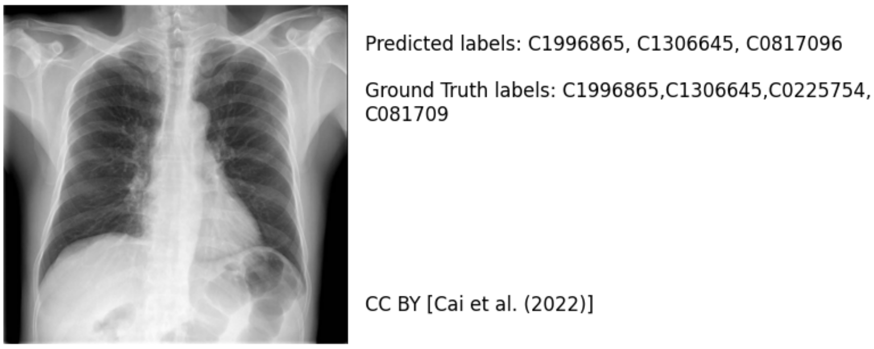
3.1. Concept Detection

The DenseNet121 model was implemented for the concept detection subtask. The PyTorch implementation of the model was used, with certain changes to adapt the default definition for the current task. The model consisted of four dense blocks and had a growth rate of 32. Every layer increased the number of features added by the growth rate. The Softmax function in the final classification layer was replaced with the Sigmoid function to refine the model for multi-label classification. The Sigmoid function is not constrained to producing probabilities that sum up to one, thus making it more suitable for multi-output tasks. For training, all the images were resized to 256 X 256 pixels and normalised. Transfer learning was leveraged by using the pretrained *imagenet1k_v1* [8] weights. These weights were obtained by training the DenseNet121 model on the ImageNet dataset, which contains 14,197,122 manually annotated images. The following parameters were utilised - a batch size of 32, a learning rate of 1×10^{-4} and the Adam optimizer[9], and the Binary Cross Entropy (BCE) loss function. The BCE loss function was chosen since it treats each label independently and outputs a probability for each label. A threshold value of 0.5 for the predictions was fixed after considering the F1 scores of different values in a fixed search space

3.2. Caption Prediction

An encoder-decoder architecture has been used for the caption prediction task. The input images were encoded into a fixed form using a prevalent CNN architecture like ResNet-101 as the encoder. The last two layers of the architecture were removed owing to the fact they would be unnecessary in an encoding pipeline. The model creates smaller representations of the input images, with each ensuing representation having more number of learned channels than the previous. The encoder produces an encoded sequence for each image having 2,048 learned channels. A batch size of 16 was used along with a encoder learning rate of 1e-4.

The encoded images are subsequently fed into an LSTM [10] based decoder which computes a weighted average across all pixels of the encoded image (important pixels being given more weight- Attention) and this weighted representation can be fed to the decoder as the initial hidden state. At each subsequent decoding step, for each pixel in the Attention network, weights are created using the encoded image and the preceding hidden state. For the training, overly long captions were truncated to fit a caption length that was greater than or equal to the lengths of 99 percent of the data. All the remaining captions were padded to uniformly have this caption length. The LSTM Decoder receives the previous word produced and the weighted average of the encoding to produce the subsequent word. The decoder used a learning rate of 4×10^{-4}. In the implementation of the network, the Adam optimiser was used as the optimiser and cross-entropy as the primary loss function. Gradient clipping was also carried out to a magnitude of 5.



Predicted labels: C1996865, C1306645, C0817096

Ground Truth labels: C1996865,C1306645,C0225754, C081709

CC BY [Cai et al. (2022)]

Figure 3: An example of a prediction

Finally, the decoder's output is passed through a linear layer that transforms the output into a score for each word in the vocabulary. Using beam search, the sequence of words with the highest overall score is chosen from a group of candidate sequences. Figure 5 and the subsequent description depicts the generated caption for an image from the validation set and shows how it compares with the actual caption.

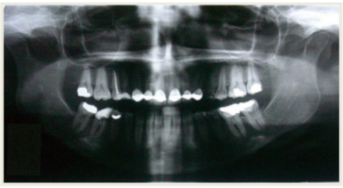
4. Results

The DenseNet121 model implemented for the concept detection subtask achieved an F1 score of 0.4957 and a manual F1 score of 0.9105. The F1-score[11] is an evaluation metric that combines the precision and recall metrics. In classification tasks where the dataset is imbalanced, accuracy alone is an ineffective metric since it is biased towards the majority class. Using a metric like the F1 score that considers both precision and recall can thus help mitigate the impact of class imbalance. Thus for the task of concept detection, the F1 score was used as the primary metric. An F1-score Manual, calculated using a subset of manually validated concepts (anatomy, directionality, and image modality) was also provided by the organizers [3]. A few examples of the model's predictions for the task are shown in Figure 3 and Figure 4. Table 1 depicts the results obtained:

Table 1
Concept Detection on the test set.

Model	F1 score	Manual F1 score
DenseNet121	0.495730	0.910585

We employed DenseNet121 for concept detection as it proved effective in handling the task of identifying multiple labels associated with each image. The dense architecture of DenseNet121 facilitated comprehensive feature extraction, allowing us to accurately detect various concepts within the images. However, when it came to caption prediction, our requirements differed. We didn't require intricate feature identification, but rather the association of images with given



Predicted labels: C0034579, C0037303, C1306645

Ground Truth labels: C0034579,C0037303,C0011334, C130664

CC BY-NC [Almehmadi et al. (2021)]

Figure 4: An example of a prediction

captions during the encoded phase. Hence, we opted for ResNet-101, which excels at capturing high-level features and proved more suitable for our caption prediction needs. By leveraging ResNet-101, we achieved efficient and accurate matching of images with their corresponding captions.

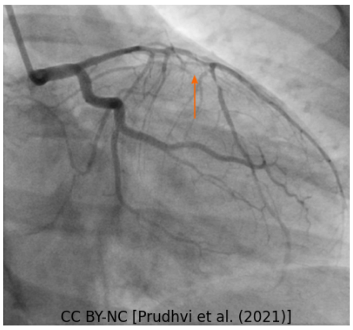
For the caption prediction task, the model accomplished a BERTScore of 0.58 and a ROUGE score of 0.22. A summarised report of the results obtained on the test set is shown below in Table 2. BERTScore is used as the primary metric instead of BLEU, and ROUGE is used as a secondary metric.

BERTScore [12] is an evaluation metric based on the BERT (Bidirectional Encoder Representations from Transformers) language model. It utilizes word embeddings to measure the similarity between the reference text and generated text. Cosine similarity is used to make the comparison and outputs a score between 0 and 1. ROUGE-1 (Recall-Oriented Understudy for Gisting Evaluation)[13] is an evaluation metric that measures the overlap of unigrams between the reference and generated text. It outputs a score between 0 and 1 indicating the extent of overlap.

The performance of the model could be attributed to the encoder-decoder architecture as it is able to understand the context better and focus its attention on the parts of the image that matter the most. Furthermore, truncating overly long captions in the training set has helped remove the outliers (taken to be captions that had lengths representing the top 1 percent) and train the model faster.

Table 2
Caption Prediction on the test set

Model	BERTScore	ROUGE	BLEURT	BLEU	METEOR	CIDEr	CLIPScore
ResNet-101 Encoder + LSTM Decoder	0.581625	0.218103	0.269043	0.145035	0.070155	0.173664	0.789327



Actual Caption: Perioperative transoesophageal echocardiography mid-oesophageal aortic valve long-axis view showing a stent protruding from the right coronary artery almost 1.cm into the Sinus of Valsalva (arrow).
Predicted Caption: Coronary angiography showing the right coronary artery.

Figure 5: An example of caption prediction.

5. Conclusion

The image captioning task in ImageCLEFmedical 2023 had two subtasks - concept detection and caption prediction. The dataset had 60,918 training samples, 10437 validation samples and 2,125 unique labels. To efficiently handle the diverse data, a deep neural network was chosen for the concept detection subtask. The DenseNet121 model was implemented for the concept detection task and it yielded an F1 score of 0.49. Further improvements can be made by accounting for the imbalance in the data, since the maximum representation of a label was 20,000 samples while the minimum representation was just 3. Additionally the classes could be clustered according to type as well.

To deal with the caption prediction task, an encoder-decoder architecture was chosen. This was important to take advantage of transfer learning in the encoding phase and attention mechanisms in the decoding stage. A beam search was performed to optimally predict the captions, keeping in mind the correctness of the whole sequence. It yielded a BERTScore of 0.58 and a ROUGE score of 0.22. Additional improvements can be made by including the text embedded inside the images for caption generation, and furthermore, the definitions of the concepts detected in the concept detection sub-task may also be employed to frame more relevant captions.

6. Acknowledgements

The authors would like to thank the Department of Computer Science and Engineering, Sri Sivasubramaniya Nadar College of Engineering, Chennai, India (https://www.ssn.edu.in/) for providing the computational resources for model training and testing.

References

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

[2] B. Ionescu, H. Müller, A. Drăgulinescu, W. Yim, A. Ben Abacha, N. Snider, G. Adams, M. Yetisgen, J. Rückert, A. García Seco de Herrera, C. M. Friedrich, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, S. A. Hicks, M. A. Riegler, V. Thambawita, A. Storås, P. Halvorsen, N. Papachrysos, J. Schöler, D. Jha, A. Andrei, A. Radzhabov, I. Coman, V. Kovalev, A. Stan, G. Ioannidis, H. Manguinhas, L. Ştefan, M. G. Constantin, M. Dogariu, J. Deshayes, A. Popescu, Overview of ImageCLEF 2023: Multimedia retrieval in medical, socialmedia and recommender systems applications, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the 14th International Conference of the CLEF Association (CLEF 2023), Springer Lecture Notes in Computer Science LNCS, Thessaloniki, Greece, 2023.

[3] J. Rückert, A. Ben Abacha, A. G. Seco de Herrera, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, H. Müller, C. M. Friedrich, Overview of ImageCLEFmedical 2023 – Caption Prediction and Concept Detection, in: CLEF2023 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece, 2023.

[4] G. Huang, Z. Liu, L. van der Maaten, K. Q. Weinberger, Densely connected convolutional networks, Proceedings of the IEEE conference on computer vision and pattern recognition (2017). doi:10.1109/CVPR.2017.243.

[5] K. Xu, J. L. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. S. Zemel, Y. Bengio, Show, attend and tell: Neural image caption generation with visual attention, in: Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37, ICML'15, JMLR.org, 2015, p. 2048–2057.

[6] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, Proceedings of the IEEE conference on computer vision and pattern recognition (2016). doi:10.1109/CVPR.2016.90.

[7] V. Atliha, D. Šešok, Comparison of VGG and ResNet used as encoders for image captioning, IEEE Open Conference of Electrical, Electronic and Information Sciences (eStream) 22 (2020). doi:10.1109/eStream50540.2020.9108880.

[8] H. E. Kim, A. Cosa-Linan, N. Santhanam, M. Jannesari, M. E. Maros, T. Ganslandt, Transfer learning for medical image classification: a literature review, BMC medical imaging 22 (2022). doi:10.1186/s12880-022-00793-7.

[9] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, CoRR abs/1412.6980 (2014).

[10] Y. Yu, X. Si, C. Hu, J. Zhang, A review of recurrent neural networks: LSTM cells and network architectures, Neural computation 31 (2019). doi:10.1162/neco_a_01199.

[11] M. Grandini, E. Bagli, G. Visani, Metrics for multi-class classification: an overview, ArXiv abs/2008.05756 (2020).

[12] 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.

[13] C.-Y. Lin, Rouge: A package for automatic evaluation of summaries, in: Text summarization branches out, 2004, pp. 74–81.