A Tale of Three Summarizers: Investigating Unimodal Text, Vision and Multimodal Combinations for Medical Summarization

Aravind Dendukuri
Gurleen Kaur
Dhananjay Srivastava
Mansi Ranka

I. Introduction

Radiology reports play a crucial role in guiding a patient's treatment, and chest X-rays are often an essential component of these reports. However, generating a detailed and comprehensive report from an X-ray image can be a time-consuming and challenging task, requiring the expertise of highly experienced radiologists. Since generating an accurate impression requires technical expertise and knowledge of the medical domain, automating the impression generation process has become a priority for healthcare providers to reduce workload and improve patient outcomes.

Deep learning has emerged as a promising technology for automating this task. By training on a large dataset of radiological reports, these models can learn to generate impressions that are comparable to those generated by human radiologists. Most existing deep learning models for impression generation rely solely on the findings section of the radiological report. However, the use of multimodal data, including image features, can potentially improve the quality of the generated impressions.

Multimodality refers to the use of multiple data modalities, such as text, images, and other sensor data, to improve the performance of machine learning models. In radiology, the incorporation of multimodal data can be especially beneficial since it can provide additional information that is not captured by the text alone. By leveraging both textual and visual information, multimodal deep learning models can potentially generate more accurate and comprehensive impressions.

In this study, we seek to explore the potential benefits of incorporating image features into the impression generation process for chest X-ray reports. Specifically, we aim to understand how image feature embeddings can guide the text to create a more thorough impression. To achieve this goal, we develop several models for investigation, including monomodal text-based model, monomodal image captioning model, and multimodal models with varying degrees of text and image fusion. The results of the image captioning model were not satisfactory, suggesting that simply generating captions for the X-ray images does not provide sufficient information to create a comprehensive impression.

However, our multimodal models, despite being rudimentary in their approach, showed great promise for the incorporation of additional modalities in this field. We intend to experiment further into this and explore more sophisticated approaches in future research.

It is worth noting that this study was part of the BioNLP 23 Shared task b Challenge, which aimed to promote the development of clinical summarization models for generating radiology impression statements by summarizing textual findings written by radiologists. Our solution placed 6th in the challenge, demonstrating the potential of our approach for improving the accuracy and efficiency of radiological report generation.

In conclusion, our study highlights the potential benefits of incorporating multimodality in radiological report generation. With further development, this approach has the potential to automate and streamline the report generation process, reducing the workload of radiologists and improving patient outcomes.

II. Related research

The Multi-level Multi-Attention based encoder-decoder (MLMA) model given by [3] is a neural network architecture that integrates image and text-based features using a Bidirectional Long Short-Term Memory (LSTM) network. The MLMA model employs multiple levels of attention mechanisms to selectively focus on relevant information from both the image and text data. Specifically, the model applies attention over the textual input and image data separately to identify local features that are relevant to the task at hand. This approach allows the model to learn a more nuanced representation of the input data, enabling it to generate more accurate predictions. Overall, the MLMA model represents a powerful approach to multi-modal learning that combines the strengths of both image and text data, making it well-suited for a wide range of applications in natural language processing, computer vision, and beyond.

The paper [4] proposes a method for generating summaries of radiology reports and images using reinforcement learning techniques. The approach involves encoding the images into embedding maps, which are then used by a sentence decoder to generate topics. The word decoder then generates the summaries by applying attention mechanisms to the topics and the textual input. The reinforcement learning technique is used to train the model to optimize the summarization performance by maximizing a reward function that is based on the quality of the generated summaries. The proposed method demonstrates promising results in generating accurate and informative summaries for radiology reports and images, highlighting the potential of reinforcement learning techniques in addressing the challenges of summarizing complex medical data.

The approach presented in this article [5] uses a 16-layer VGG network to extract image features, which are then combined with textual information from radiology reports to generate image captions. The textual information is processed using a sequence processor that generates a word embedding layer, which is then passed through a Long Short-Term Memory (LSTM) layer to generate a contextual representation of the text. The image features and text embedding layer are fused together and passed through a decoder layer to generate summaries of the radiology reports. The fusion of image and text data enables the model to capture both visual and semantic information, resulting in more accurate and informative captions. The proposed approach demonstrates the potential of deep learning models for generating accurate and descriptive captions for complex medical data.

The paper [6] describes two approaches for radiology report summarization, which were adopted by the authors to participate in a competition. The first approach is a monomodal architecture that uses an Encoder-Decoder model, with a bi-directional Gated Recurrent Unit (GRU) network as the encoder and a conditional GRU with a Bottleneck function as the decoder. The second approach is a multimodal architecture that fuses DenseNet121 for image classification with the monomodal architecture using three methods: encdecinit policy that initializes the encoder and decoder state with visual features, ctxmul policy that performs element-wise product of encoder annotations, and trgmul policy that performs element-wise product of target embedding. The proposed approaches demonstrate the potential of using deep learning models for radiology report summarization, with the multimodal architecture showing improved performance by incorporating both image and text information.

MedClip is a multimodal fusion model designed for medical applications, developed by researchers at UCLA. It comprises three main components: modality-specific feature extractors, which extract features from image and text data separately; modality-specific classifiers, which classify text or image independently, such as predicting whether a chest X-ray image shows signs of lung disease or whether a clinical note mentions symptoms of the disease; and a multimodal fusion model, which combines the feature representations generated by the feature extractors and the output probabilities generated by the classifiers using a multimodal fusion approach. The model's ability to integrate information from multiple modalities enables it to generate more accurate and informative predictions for medical applications, demonstrating the potential of multimodal fusion approaches in addressing the challenges of analyzing complex medical data.

III. Dataset

Our Text Model is based on the MIMIC CXR Dataset and our image Modle is based on the ChexPert Dataset.

The MIMIC Chest X-ray (MIMIC-CXR) Database v2.0.0 is a vast collection of chest radiographs in DICOM format, accompanied by free-text radiology reports, which is publicly accessible. This dataset comprises 377,110 images, which correspond to 227,835 radiographic studies conducted at the Beth Israel Deaconess Medical Center located in Boston, MA. In compliance with the US Health Insurance Portability and Accountability Act of 1996 (HIPAA) Safe Harbor requirements, the dataset has been de-identified by removing all protected health information (PHI). The main objective of this dataset is to facilitate research in medicine, particularly in areas such as image understanding, natural language processing, and decision support.

CheXpert is a publicly available dataset designed for the interpretation of chest radiographs. This dataset comprises 224,316 chest radiographs that correspond to 65,240 patients. We collected these chest radiographs retrospectively from Stanford Hospital, which were performed between October 2002 and July 2017 in both inpatient and outpatient centers, along with their associated radiology reports.

IV. Methodology

A. Image captioning on the xray images

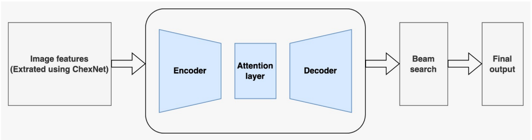


Fig. 1: Image Captioning model trained on the Chexpert based image embeddings. Based on the autoencoder architecture. Uses beam search for impression generation.

We have modeled this approach as an Image captioning pipeline. The main idea is to use an auto-encoder with attention to predict impressions in text format, given chest x-rays images.

To begin with, we generate fresh data points by augmenting the existing ones and supplement the text data with special markers such as "$\langle start \rangle$" and "$\langle end \rangle$" to aid the model in comprehending the sentence's initiation and conclusion. Using the Tokenizer provided by the TensorFlow deep learning library, we tokenise the text data into numerical data.

Next, we employ transfer learning to convert images to feature vectors, using the pre-trained weights of the CheXNet competition model. The CheXNet model is a multi-layered DenseNet121 model trained on a dataset of 112,120 chest X-ray images for the purpose of detecting 14 different diseases.

Our 112,120 data points could comprise multiple images each. For the sake of simplicity and computational expenses, we have chosen one image per datapoint and embedded it.

The next step involves feeding the image tensors to the encoder. The image features are first passed through a dense layer, followed by a dropout layer for fine-tuning.

We developed a one-step decoder layer for the decoder, which accepts the decoder input, encoder output, and state value. The decoder input is any character token number, and it is first passed through the embedding layer. Next, the embedding layer output and the encoder output are sent through the attention layer to generate the context vector. This context vector is then fed to the RNN (we used a GRU) with the initial state being that of the previous decoder.

In this approach, the attention mechanism plays a vital role by estimating the correlation between the elements of the input sequence and the target element. We use the Bahdanau additive attention mechanism here. The model takes both image vectors and text vectors as inputs.

To obtain a deeper understanding of the input features, we employ a bi-directional GRU that extracts high-level features from the input. Additive attention is then applied to assign weight vectors to each sequence of words. Subsequently, the word-level features from each timestamp are aggregated into a sentence-level feature vector.

During each decoder time-step, the decoder invokes the one-step attention layer, which computes the scores and attention weights. The resulting outputs from each time-step are accumulated and saved in a variable. Each decoder step produces the subsequent word in the sequence, resulting in the final output sequence.

Finally, we use the Beam search algorithm to find the output sentence in the inference stage. At each time step in the decoder part, we select the top K-words (where K is the beam length) with the maximum probability of occurrence. The rationale for selecting Beam search instead of greedy search is that Beam search aims to maximize the conditional probability of all potential words and select the most likely result. It considers multiple possible solutions at each step, instead of just the most likely one.

Although we attempted to use the auto-encoder with attention model, it provided unsatisfactory results, as most of the predictions were identical. While it's possible that further model tuning could enhance its performance, we ultimately opted to move forward with other models that were yielding better outcomes. Our tests on a smaller dataset, Indiana Chest X Rays yielded significantly better results, with a BLEU score of 0.42, but that could be because the Indiana dataset is significantly smaller than this dataset and there isn't much variation in the impressions so the learnt vocabulary isn't very big.

B. *Dense auto encoder based multi modal feature mixing*

The inspiration for the model shown above is taken from machine translation model of pytorch and a few of the papers mentioned above. We have tried to train a short autoencoder with 13 layers. The aim was to combine the image and text embeddings and generate the final summarization. Since the text embedding generated in approach 3 given below was large, we decided to input the text summary generated by the approach 3 along with image embeddings generated by the DenseNet 121 model and train an autoencoder-based architecture to generate better summaries. The Rouge score for this model is very poor and all the summaries generated are the same. Of the many possible reasons some could be a smaller number of layers, Improper design of number of nodes in each layer, lack of attention mechanism, a poor loss function as we have used MSE only, use of poor embeddings like GLOVE.

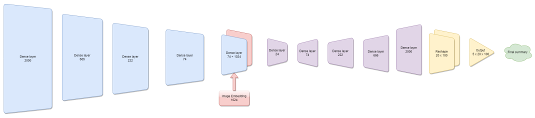


Fig. 2: Autoencoder model that takes in text embeddings as the input. The image embeddings are concatenated at the context layer.

C. *Naive feature combination methods*

1) *Approach 1:* The model used for training is a Bidirectional Autoregressive Transformer (BART), which is a state-of-the-art sequence-to-sequence model based on the Transformer architecture. The BART model has been specifically designed for text generation tasks, such as text summarization, by combining the strengths of both autoregressive and bidirectional models. The training process involved training the BART model for 3 epochs. The learning rate used during training was set to 5e-5, which is a common value used in natural language processing tasks. The training and evaluation batch sizes were set to 64, which refers to the number of samples used in each forward and backward pass during training and evaluation, respectively. This value was chosen to optimize the memory usage of the model while still maintaining a reasonable training speed. The training was performed on a single Nvidia A100 GPU with 40GB VRAM. To further optimize training speed, reduced precision training was used by setting the fp16 option to True. This option uses half-precision floating-point arithmetic, which reduces the memory usage and computation time required during training. The justification for using the BART model for automatic summarization is that it is a powerful and widely used model specifically designed for this task. The model has been trained on a robust sequence denoising objective, which helps to improve its accuracy in generating high-quality summaries. Therefore, the use of BART for automatic summarization is a well-established approach in the field of natural language processing.

2) *Approach 2:* To enhance the performance of the BART summarization model, we incorporated information from an image by using a pretrained DenseNet 121 model from torch xray vision. This model has been specifically designed for

image classification tasks and is trained on a dataset from the ChexPert challenge. The ChexPert challenge dataset consists of chest X-ray images and corresponding radiologist reports, which are labeled with 14 categories such as Lung opacity, Cardiomegaly, and others. The DenseNet 121 model is trained to predict these categories given a chest X-ray image. To incorporate this information into the summarization model, we added the output of the DenseNet 121 model as an additional input to the base summarization model. Specifically, we concatenated the text prediction output of the DenseNet 121 model with the input text for the base summarization model. We then ran the BART summarization model again with this augmented input. To evaluate the performance of this approach, we used the ROUGE score as a metric. ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is a set of metrics commonly used to evaluate the quality of automatic summarization. We observed a 2% increase in ROUGE score accuracy with this approach, which indicates that incorporating information from the image improves the quality of the generated summaries.

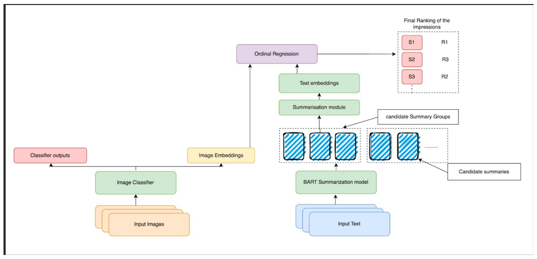
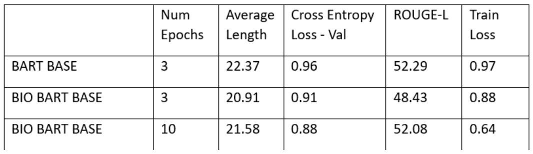


Fig. 3: A Bidirectional AutoRegressive Transformer (BART) is trained for generating summaries using diverse beam search, and a DenseNet-121 model is used to extract image features to rank these summaries. An SGD Regressor model is then trained using a proxy output developed by a proxy BERTScore output to rank the combined embeddings.

3) Approach 3: The first step in generating multiple candidate summaries is modifying the beam search of the model's generate function. Beam search is a popular decoding algorithm used in natural language processing to generate text. It works by keeping track of the top-k most likely sequences of tokens, where k is the beam width. However, this method can be limited in generating diverse outputs. To address this limitation, the ”num_beam_groups” feature is used to generate multiple diverse groups of summaries in the model's generate function. The result is a larger pool of candidate summaries. This is achieved by using the ”num_beam_groups” feature to generate 3 diverse groups of summaries in the model's generate function, with a total of 9 beams being used to generate all the candidate summaries. Once the candidate summaries are generated, the next step is to evaluate their quality. BERTScore is used as a metric to calculate the similarity between the candidate summaries and the ground truth. BERTScore uses the BERT encoder to calculate embeddings for the candidate summaries and the ground truth. These embeddings are then used to calculate the cosine similarity between the candidate summaries and the ground truth. BERTScore has been shown to be a reliable metric for evaluating the quality of natural language generation models. Next, the embeddings from the BART model outputs are concatenated with embeddings from the image. These embeddings are generated from the Chexpert model's last layer, applying a ReLU activation function and average pooling the output to get a final embedding of shape 1024x1 for the image. Concatenating these embeddings provides a more comprehensive representation of the input data. To rank the candidate summaries, an SGDRegressor model is trained. The SGDRegressor is a linear model that is trained using stochastic gradient descent. It is used to predict the score of each candidate summary based on the concatenated embeddings of the BART model outputs and the image. The SGDRegressor is trained using the partial fit method because all of the data cannot be fit in memory at once.

Link to the github repo for the above discussed methodologies: https://github.iu.edu/cs-b657-sp2023/final-project

V. Result Discussion



	Num Epochs	Average Length	Cross Entropy Loss - Val	ROUGE-L	Train Loss
BART BASE	3	22.37	0.96	52.29	0.97
BIO BART BASE	3	20.91	0.91	48.43	0.88
BIO BART BASE	10	21.58	0.88	52.08	0.64

Fig. 4: Base Text model scores comparison of different models.

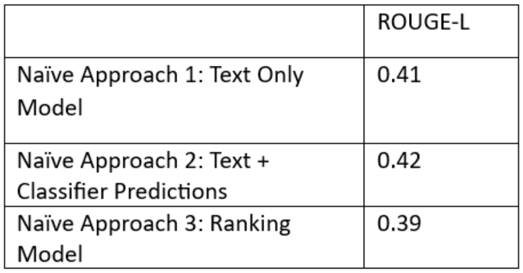
To initiate the process, our first step is to train our text summarization models, which are essential for summarizing large volumes of text into shorter and more manageable forms. For our training, we rely on two models, namely BART base and BioBART base.

BioBART base is a variant of BART base that has been pre-trained on the PubMed corpus, a massive collection of medical research articles. Due to this pre-training, BioBART base is expected to perform better on medical domain datasets compared to the standard BART model.

In our experimentation, we have observed that training BioBART base for a longer duration gives us the best validation loss. Validation loss is a metric used to evaluate the performance of a model by measuring the difference between predicted and actual output on a validation set. Thus, by training BioBART base for an extended period, we can achieve a more accurate and efficient text summarization model for medical domain datasets.

In this section, we will discuss the various Naïve approaches we have implemented for feature mixing in our text summarization model.

Approach 1: In this approach, we have utilized the BART BASE model to generate the summaries, which serves as the baseline for our model. We only use the text as input for generating the summaries.



	ROUGE-L
Naïve Approach 1: Text Only Model	0.41
Naïve Approach 2: Text + Classifier Predictions	0.42
Naïve Approach 3: Ranking Model	0.39

Fig. 5: ROUGE scores comparison of different models.

Approach 2: To improve the accuracy of our model, we added the text predictions from the Chexpert model to our input text. Chexpert is a model that predicts various medical conditions based on chest X-ray images. By including its predictions in our input text, we observed a marginal bump in accuracy.

Approach 3: In this approach, we ranked the beam search outputs generated by the model and selected the summary with the highest score. However, this approach yielded a lower accuracy compared to the previous approaches.

While analyzing the poor results of the third approach, we discovered that we were suppressing a warning related to an update in the transformers library. This update changed the maximum length parameter to maximum new tokens in the generate function, which was causing all our outputs to be truncated to 20 tokens, negatively impacting our accuracy. We fixed this issue by simply using the BioBART base model, which helped us improve our ROUGE Scores up to 0.45. However, we need to further inspect our model to identify the potential bottlenecks and improve its accuracy.

VI. Conclusion and Future Work

In this project we aimed to identify and implement multiple approaches for multimodal summarization of radiological reports. Our Naive feature mixing approaches were able to achieve an accuracy of around 0.4 on the test set. We submitted these approaches to Shared task 1B of the 23rd BioNLP Workshop at the ACL Conference. We ranked 6th within the eight teams that participated with a ROUGE L score of 26.08 on a hidden test set prepared by the organizers.

There are a lot of ways in which we would like to improve our current results such as

- Fixing simple token length bugs and complete the image captioning and autoencoder models and train them longer to get the final results.
- Study variation across multiple Text and Image Backbones such as Big Bird and Pegasus for text and EfficientNet or Vision Transformers for the Images.
- Understand and visualize the image text feature combinations deeper for example in a phrase grounding type of way.

We hope this work goes on to show that there is merit in these simple feature mixing approaches and we hope to explore them in detail further.

References

[1] https://physionet.org/content/mimic-cxr/2.0.0/

[2] https://kundan-jha.medium.com/impression-generation-from-x-ray-images-case-study-341d25af6edf#0535-6053f58966b1

[3] https://link.springer.com/chapter/10.1007/978-981-15-4015-8_15

[4] https://arxiv.org/abs/1904.02633

[5] https://machinelearningmastery.com/develop-a-deep-learning-caption-generation-model-in-python/

[6] https://aclanthology.org/2021.bionlp-1.33.pdfMethodology

[7] Jean-benoit Delbrouck, Khaled Saab, Maya Varma, Sabri Eyuboglu, Pierre Chambon, Jared Dunnmon, Juan Zambrano, Akshay Chaudhari, and Curtis Langlotz. 2022. ViLMedic: a framework for research at the intersection of vision and language in medical AI. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pages 23–34, Dublin, Ireland. Association for Computational Linguistics.

[8] Delbrouck, Jean-Benoit, et al. ”Improving the factual correctness of radiology report generation with semantic rewards.” EMNLP 2022 findings.

[9] Yuhao Zhang, Derek Merck, Emily Tsai, Christopher D. Manning, and Curtis Langlotz. 2020. Optimizing the Factual Correctness of a Summary: A Study of Summarizing Radiology Reports. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5108–5120, Online. Association for Computational Linguistics.

[10] BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension (Lewis et al., ACL 2020)

[11] Yuan, H., Yuan, Z., Gan, R., Zhang, J., Xie, Y., Yu, S. (2022). BioBART: Pretraining and Evaluation of A Biomedical Generative Language Model. ArXiv. /abs/2204.03905

[12] Irvin et al, CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison.

[13] Joseph Paul Cohen and Mohammad Hashir and Rupert Brooks and Hadrien Bertrand On the limits of cross-domain generalization in automated X-ray prediction. Medical Imaging with Deep Learning 2020 (Online: https://arxiv.org/abs/2002.02497)

[14] https://arxiv.org/pdf/2210.10163.pdf