nav-nlp at RadSum23: Abstractive Summarization of Radiology Reports using BART Finetuning

Kancharla Nikhilesh Bhagavan
Indian Institute of Technology Roorkee
kancharla_nb@cs.iitr.ac.in

Macharla Sri Vardhan
Indian Institute of Technology Roorkee
macharla_sv@cs.iitr.ac.in

Madamanchi Ashok Chowdhary
Indian Institute of Technology Roorkee
madamanchi_ac@cs.iitr.ac.in

Raksha Sharma[*]
Indian Institute of Technology Roorkee
raksha.sharma@cs.iitr.ac.in

Abstract

This paper describes the experiments undertaken and their results as part of the BioNLP 2023 workshop. We took part in Task 1B: Radiology Report Summarization (Delbrouck et al., 2023). Multiple runs were submitted for evaluation from solutions utilizing transfer learning from pre-trained transformer models, which were then fine-tuned on the MIMIC-III dataset, for abstractive report summarization. The task was evaluated using different evaluation metrics such as ROUGEL, Bertscore, and F1-RadGraph, and the corresponding scores of our best-performing system are 32.33, 54.49, and 32.68 respectively.

1 Introduction

BioNLP 2023 shared task, aims to attract researchers and students working on NLP across three summarization tasks in the bio-medical domain: problem list summarization, radiology report summarization and lay summarization of biomedical research articles. We took part in radiology report summarization and performed experiments. Documents that summarize and record radiological examinations are known as radiology reports. Three elements typically make up a radiology report: (1) a background section that contains general information about the patient and the exam, (2) a findings section that includes examination data, and (3) an impression section that summarizes the findings in relation to the background. Usually, medical professionals go through the finding section and pull out the valuable findings and try to summarize the whole radiology to make clinical decisions. But this manual process of summarization is time-consuming and inefficient. So, in this work we propose experiments to automate the generation of the impressions section from the findings of the radiology report, speeding up the radiology workflow and improving the efficiency of clinical communications.

A key feature of this task is that radiology reports used for training and evaluation are collected from various sources, e.g., training instances are sampled from the MIMIC-III database. Experiments were performed implementing sequence to sequence models with encoder-decoder architecture like BART, T5. These models were then further fine-tuned on a portion of the MIMIC-III Dataset to produce abstract summaries of the report's conclusions from the findings section. A portion of the MIMIC-III is used for validation purpose and standard ROUGEL metric is used for evaluation purpose.

2 Related Work

Summarization tasks were initially just focused on extractive summarization. It is a method of extracting noteworthy words from the text to form a summary. Advancement in neural networks models brought up the scope of abstractive summarization, which involves producing new words to convey the meaning of the text. It attempts to capture the meaning of the text and express it in a more concise and coherent form.

One of the earliest approaches to abstractive summarization was the use of rule-based systems that relied on handcrafted templates and heuristics to generate summaries. However, these systems were limited in their ability to capture the nuances of natural language and often produced summaries that were either too general or too specific.

In recent years, deep learning techniques have revolutionized the field of abstractive summarization, enabling models to learn from large amounts of data and generate summaries that are more accurate and fluent. One of the most influential models in this regard is the encoder-decoder framework with attention mechanism, which was first proposed by (Bahdanau et al., 2016) for machine

[*] Corresponding author

The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks, pages 541–544
July 13, 2023 ©2023 Association for Computational Linguistics

translation and later adapted for summarization by (See et al., 2017).

The encoder-decoder model involves training a neural network to encode the input text into a fixed-length vector representation, which is then decoded into the summary. The attention mechanism allows the model to focus on different parts of the input text during the decoding process, which improves the quality of the generated summary. Since the introduction of the encoder-decoder model, numerous other models have been proposed that build on this basic framework, including transformer-based models like BERT (Devlin et al., 2019) and GPT (Radford et al., 2018). These models have achieved state-of-the-art performance on a range of summarization tasks and have become increasingly popular in both industry and academia.

3 Task Description & Dataset

The goal of this research is to develop an effective algorithm for summarizing radiology reports. Given a radiology report, the algorithm should generate a concise summary of the key findings and conclusions of the study. The performance of the algorithm will be evaluated using standard metrics such as ROUGEL (Lin, 2004), Bertscore (Zhang* et al., 2020), and F1-RadGraph (Delbrouck et al., 2022).

The dataset used in this research is based on the MIMIC-III database, which is a publicly available clinical database containing de-identified health data for over 50,000 patients. The dataset includes radiology reports from six different modalities: CT Head, CT Abdomen, CT Chest, MR Head, CT Spine, CT Neck, and five different anatomies: head, abdomen, chest, spine, and neck.

The dataset contains a total of 79,779 radiology reports, with varying numbers of reports for each modality/anatomy pair. The dataset is divided into three subsets for training, validation, and testing, with the majority of reports (at least 10,000) in the training subset for each modality/anatomy pair.

The performance of the summarization algorithm will be evaluated using standard metrics such as ROUGEL, Bertscore, and F1-RadGraph, with the test subset of the dataset used for evaluation. The results of the evaluation will be reported in terms of the metrics and compared to state-of-the-art summarization algorithms.

4 Methods & Results

Pretrained summarization models are used in our suggested methodology. We fine-tuned two types of pretrained models for the radiology report summarization; BART and T5. We used Hugging Face Transformers libraries for fine-tuning.

BART: BART (Bidirectional and Auto-Regressive Transformers) is a pre-trained neural network architecture that is based on the transformer model, which is a popular architecture for natural language processing (NLP) tasks. BART was introduced by (Lewis et al., 2019) and has since been shown to achieve state-of-the-art results on a variety of NLP tasks, including text summarization.BART consists of two main components: a masked language model (MLM) and a sequence-to-sequence (Seq2Seq) model. The MLM component is trained to predict masked tokens in an input text, while the Seq2Seq component is trained to generate a target sequence given an input sequence. The combination of these two components allows BART to be fine-tuned on a range of NLP tasks, including text summarization.

In the context of summarization, BART takes an input document and generates a summary by conditioning it on a special token indicating the start of the summary. During fine-tuning, the model is trained to generate a summary that maximizes a similarity metric such as ROUGE (Recall-Oriented Understudy for Gisting Evaluation) with respect to a reference summary.BART's ability to generate high-quality summaries can be attributed to its pre-training on large amounts of text data, which allows it to capture the general patterns and structures of natural language. Fine-tuning specific summarization tasks further refines the model's ability to generate accurate and concise summaries. In addition to text summarization, BART has also been shown to perform well on other NLP tasks, including machine translation, question answering, and dialogue generation. Its flexibility and effectiveness make it a valuable tool for a wide range of NLP applications.

T5: T5 (Text to Text Transfer Transformer) is a powerful language model that has demonstrated impressive results on a wide range of NLP tasks, including language translation, summarizing, question answering, and text classification, among others. Unlike traditional NLP models that are de-

Run	ROUGEL	Bertscore	F1-RadGraph
1	32.33	54.49	32.68
2	31.22	52.39	31.89

Table 1: Evaluation of Radiology Report Summarization on mimic-iii-open-test set

Run	ROUGEL	Bertscore	F1-RadGraph
1	32.39	55.34	33.37
2	30.96	52.60	32.39

Table 2: Evaluation of Radiology Report Summarization on mimic-iii-hidden-test set

signed to perform a specific task, T5 is designed to perform a variety of tasks by converting the input text into a desired output text. This makes it a highly versatile model that can be used for a wide range of NLP applications.

T5 is based on the transformer architecture, which was originally proposed by (Vaswani et al., 2017). The transformer architecture is a type of neural network that is particularly well-suited for processing sequential data such as natural language text. Unlike traditional recurrent neural networks (RNNs), which process input data one element at a time, transformers are able to process entire sequences of input data in parallel. This allows them to process longer sequences more efficiently, which is particularly important for NLP tasks that often involve long sequences of text. Additionally, transformers use a self-attention mechanism that allows them to selectively focus on different parts of the input sequence, which helps to capture long-range dependencies and improve performance on tasks such as language translation and summarization.

5 Experiments

We propose different runs for this task. Table 1 and Table 2 show the evaluation of different models tested with mimic-iii-open-test and mimic-iii-hidden-test sets respectively. We discussed different versions of Bart and T5 on Hugging Face transformers. We ended up using Bart-base and T5-base due to resource constraints. The standard set of hyperparameters used for the different runs are Learning rate=5e-05, number of epochs=15, and gradient accumulation steps=5. The evaluation results of various runs of this task have been shown in Table 1 and Table 2.

Team	ROUGEL	Bertscore	F1-RadGraph
utsa-nlp	34.07	56.30	35.25
shs-te-dti-mai	33.93	55.49	34.93
nav-nlp	32.33	54.49	32.68
sinai	31.62	54.33	32.65
knowlab	32.22	54.91	32.49
elirf	29.57	52.24	31.40
aimi	24.45	45.54	21.24

Table 3: mimic-iii-open-test set leaderboard

Different runs of this task are:
1. Our first proposed method is Bart-base. We fine-tuned the Bart-base model using the mimic-iii-train set and chose the best Bart-base model by testing with the mimic-iii-validation set and estimating validation loss. We then tested it with mimic-iii-test and mimic-iii-hidden-test sets respectively and the results are shown in the above tables. We have used a batch size of 20 for the training set and 4 for validation set.
2. This run is the same as the first run instead we used T5-base as the base model. We have used a batch size of 10 for the training set and 4 for the validation set.

6 Conclusion

In this paper, we presented all our experiments of fine-tuning the pre-trained hugging face model for abstractive radiology report summarization. Our approach aimed to identify the most effective architecture for this task, and after experimenting with various architectures, we found that a BART-like architecture outperformed the others. We then provided detailed results and analysis of our experiments, which demonstrated the effectiveness of our methods for radiology report summarization. Rankings for the mimic-iii-open-test set and mimic-iii-hidden-test set are shown in Tables 3 and 4 respectively. Our findings suggest that fine-tuning pre-trained models is a promising approach for improving the performance of NLP models in medical settings, with potential implications for enhancing the efficiency and accuracy of medical record analysis and ultimately improving patient care.

Team	ROUGEL	Bertscore	F1-RadGraph
utsa-nlp	35.32	57.26	36.94
shs-te-dti-mai	34.41	57.08	36.31
aimi	33.43	55.54	35.12
sinai	32.32	55.04	33.96
knowlab	32.02	55.64	33.39
nav-nlp	32.39	55.34	33.37
elirf	30.19	53.94	32.58

Table 4: mimic-iii-hidden-test set leaderboard

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2016. Neural machine translation by jointly learning to align and translate.

Jean-Benoit Delbrouck, Pierre Chambon, Christian Bluethgen, Emily Tsai, Omar Almusa, and Curtis Langlotz. 2022. Improving the factual correctness of radiology report generation with semantic rewards. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4348–4360, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jean-Benoit Delbrouck, Maya Varma, Pierre Chambon, and Curtis Langlotz. 2023. Overview of the radsum23 shared task on multi-modal and multi-anatomical radiology report summarization. In *Proceedings of the 22st Workshop on Biomedical Language Processing*, Toronto, Canada. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

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

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.