Predicting Doctor's Impression For Radiology Reports with Abstractive Text Summarization

Stanford CS224N Custom Project

Mentor: Zihan Wang

Xiyuan Chen
Department of Computer Science
Stanford University
xycshine@stanford.edu

Silvia Gong
Department of Computer Science
Stanford University
silvgong@stanford.edu

William Zhuk
Department of Computer Science
Stanford University
waz@stanford.edu

Abstract

Predicting doctor's impression (summarization) for radiology reports saves doctors and patients tremendous time from manually digging through the reports. But there are few pre-trained language models for summarization, especially for radiology datasets. We solve abstractive summarization for the free-text radiology reports in the MIMIC-CXR dataset [1] by building ClinicalBioBERTSum, which incorporates domain-specific BERT-based models into the state-of-the-art BERTSum architecture [2]. We give a well-rounded evaluation of our model performance utilizing both word-matching based metrics and semantic based metrics. Our best-performing model obtains a ROUGE-L F1 score of 57.37/100 and a ClinicalBioBERTScore of 0.55/1.00. With comprehensive experiments, we showcase that domain-specific pre-trained and fine-tuned encoders and sentence-aware embeddings could significantly boost the performance of abstractive summarization for radiology reports. Our work also provides a set of pre-trained transformer weights that could further facilitate practitioner's future research with radiology reports.

1 Introduction

Text summarization helps to direct people's attention to the most important contents and saves tremendous human labor for digging through the documents. Automatic summarization is especially significant for radiology reports since it alleviates patients' great burden of reading through the lengthy and obscure reports, and it often takes doctors years of training to accumulate enough expertise to write concise and informative radiology report summarization. Pre-trained BERT-based models greatly benefits various NLP tasks, but they are rarely applied to text summarization in existing works, and there is no known language models specially pretrained for radiology datasets. In this project, we aim to solve the problem of summarizing free-text radiology reports using pre-trained BERT-based models.

Specifically, using the "Indication" and "Findings" sections in the radiology reports, we aim to predict the "Impression" section, which corresponds to the doctors's summarization. (See Figure 3 for an example). We formulate this problem as an abstractive text summarization task, because doctors' summarization is generally compressed and comprehensive containing information from a wide coverage of a radiology report. Our model ClinicalBioBERTSum takes the free-text radiology reports from MIMIC-CXR dataset [1] as input and uses the seq2seq [3] architecture to generate

output abstractive summarizations. Inspired by BERTSum [2], our model ClinicalBioBERTSum utilizes the clinical medical domain-specific pre-trained language model ClinicalBioBERT to learn sentence-level representations and generates the predicted tokens with a transformer decoder. Based on the findings in the previous work that combining extractive and abstractive objectives improves the summarization [4], we also investigate how using the two objectives for a two-stage fine-tuning affects the performance of our model.

Additionally, we also experiment with Byte-Pair Encoding (BPE) [5] tokenizer to create a customized vocabulary tokenizer to better fit the radiology reports in our dataset. Moreover, it is challenging to evaluation abstractive summarization, since different words could convey the same underlying information. Using ROUGE score as the only metric which captures the n-gram matchings may not be sufficient. To give a better evaluation of our model, we propose another automatic evaluation metric ClinicalBioBERT-score based on BERTscore [6], which captures semantic similarities and is highly correlated with human evaluation.

Our main contributions are three-fold: we built ClinicalBioBERTSum, a domain-specific Bert-based language model for abstractive text summarization to save doctors' time to generate concise and informative summarization for radiology reports, we created a set of pre-trained transformer weights for future NLP works in the radiology field to build upon, and we explored a new semantics-based metric to better evaluate abstractive summarization for clinical text.

2 Related Work

Tokenizer. BERT Models rely on a word-slice tokenizer that has been trained on wikipedia and literature datasets. Custom tokenizers can also be created when datasets are large enough, and the tokenizers we discuss later on use a technique adapted froom BPE (Byte-Pair Encoding) [5]. This is a technique that starts with a base vocabulary of symbols, and then uses a greedy merging algorithm to form a vocabulary by merging existing vocabulary tokens into common larger tokens. This allows one to fit a tokenizer to a specific dataset, which can be useful when the vocabulary of a domain is different than the vocabulary of the pretrained model. We used BPE to create a clinical medical vocabulary to fit our radiology reports.

Pre-trained Language Model. Pre-trained Language models are models that learn the general language structure from a large amount of unannotated data before being applied to specific NLP tasks. The most widely used ones such as BERT [7] and DistilBERT [8] have not been pre-trained on specialty corpora such as clinical text, so they do not perform well on domain-specific texts. ClinicalBioBERT is a model created from BioBERT [9], which is a BERT model finetuned on PubMed abstracts and PMC articles. It is then fine-tuned on clinical texts from about 2M notes of all types in the MIMIC-III [10] to make ClinicalBioBERT, which fits the semantics of the radiology reports in our dataset well.

Text Summarization. Text summarization approaches have two categories. Extractive summarization is a binary classification task to select the most relevant sentences to summarize the document. Several extractive summarization models including NEUSUM [11], BanditSum [12] and STRASS [13] use different methods to compare and select sentences to capture the gist of the input document. On the other hand, BERTSum [2] introduces a novel document-level encoder to better embed different sentences on top of BERT. Inspired by BERTSum's encoder, we performed sentence-level embeddings to better capture multi-sentences features.

Abstractive summarization generates novel summaries containing sentences that may not appear in the source document. These models include T5 [14], BART [15], and BERT-based seq2seq [3].Liu et al. proposes the BERTSum [2] model which improves abstractive summarization performance by learning sentence-level representations and using a two-stage fine-tune scheme. We use T5 [14] and a word-level seq2seq [3] as our baselines, and extent BertSum with domain specific pre-trained language models to further improve radiology report summarization.

Evaluation Metrics. Standard text summarization evaluation techniques include ROUGE [16] and BERTscore [6]. ROUGE measures text overlaps whereas BERTscore computes token similarity using contextual embedding. Inspired by BERTscore [6], we use domain-specific encoders to learn contextual embeddings for the reference and predicted summarization for the radiology reports and compute their the embedding similarity accordingly for better evaluation.

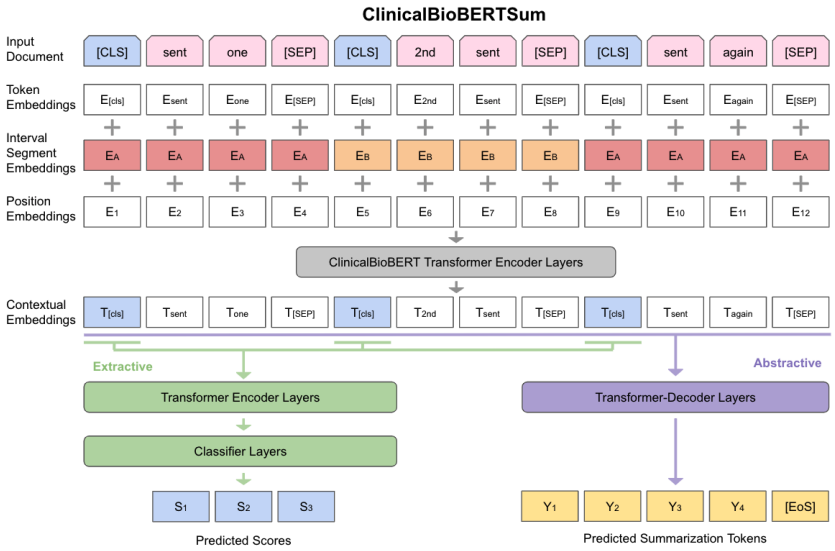

Figure 1: ClinicalBioBERTSum Pipeline. It inserts [CLS] tokens and uses interval segment embeddings to separate different sentences, learns sentence-level embedding with ClinicalBioBERT Transformer encoder. Only [CLS] token embeddings are used as inputs for the following layers with the extractive objective with S_n's being the predicted scores for each sentence. All token embeddings are used as inputs for the following layers with the abstractive objective with Y_n's being the predicted tokens for summarization.

3 Approach

Main Approach. Following BERTSum [2], we insert a [CLS] token at the beginning of each sentence and use interval segmentation embeddings to distinguish different sentences. Our model (Fig 1) learns sentence-level embeddings using a pre-tained Transformer Encoder [17] which captures bi-directional context and inter-sentence relationship. Compared to BERTSum [18], we propose to replace the BERT Encoder with the ClinicalBioBERT Encoder [17] pre-trained on the MIMIC-III dataset [10] containing various types of clinical notes, which better fits the semantics of the radiology reports in our MIMIC-CXR dataset and thus learns better representations.

To perform the abstractive summarization task, we feed the sentence-level embeddings generate by the encoder to a Transformer decoder (Fig 1) randomly initialized, train the decoder and fine-tune the encoder in an end-to-end fashion. (model #5 in Table 1) We use NMT loss function that computes word similarity as the distances in the embedding space.

$$l_{emb} = \sum_{i=0}^{I} \sum_{k=0}^{K} p(y_i|y_{<i}, X) d(E(V_k), E(y_i)) \tag{1}$$

It is a standard softmax cross-entropy loss with a weighted average of distances to a reference word in the continuous latent space. Y_i is the i-th word in the predicted summary and V_k is the k-th word in the target-side vocabulary. $E(w)$ is a vector word embedding of word w. d is the Euclidean distance function.

Previous work [4] indicates that combining extractive and abstractive objectives helps improve the summarization quality. Thus, we investigate two-stage fine-tuning for the ClinicalBioBERT encoder of our model. In the extractive setting(model #6 in Table 1), we use several Transformer Encoder layers after the ClinicalBioBERT encoder. The Transformer Encoder takes only the embeddings for the [CLS] tokens to capture bi-directional context and inter-sentence relationship, which enhances the expressiveness of our model.Then we use a sigmoid classifer to predict a score for each sentence, The sentences with scores larger than 0.5 are included in the extractive summarization output. To train the extractive model, we create the reference summarization by picking sentences from the input document using a greedy algorithm similar to [19]. Specifically, for each radiology report, we pick

the sentences from the indication and findings sections which maximize the ROUGE-2 score against the Impression section. We use standard BCEloss as the loss function. Then in the second stage, we perform another fine-tuning for the ClinicalBioBERT encoder with the abstractive objective (model #7 in Table 1). Specifically, we take the fine-tuned ClinicalBioBERT encoder from the model trained in the extractive setting, add a randomly initialized Transformer decoder and train the entire model end-to-end with the abstractive summarization task as described above.

Another technique we explore is training a custom model for this problem, paired with a custom tokenizer that is trained specifically for radiology reports (model #10 in Table 2). The tokenizer was promising, with tokens including "calcification", "interstitial", and "cardiomediastinal", which are very specific to our dataset. However, since this tokenizer will necessitate retraining a new model with a different embedding size, training a smaller model was necessary to comply with compute and time constraints. We trained MiniBERT [20], a smaller BERT model on our training data for 40 epochs with this custom tokenizer. We include the performance of this model in the appendix.

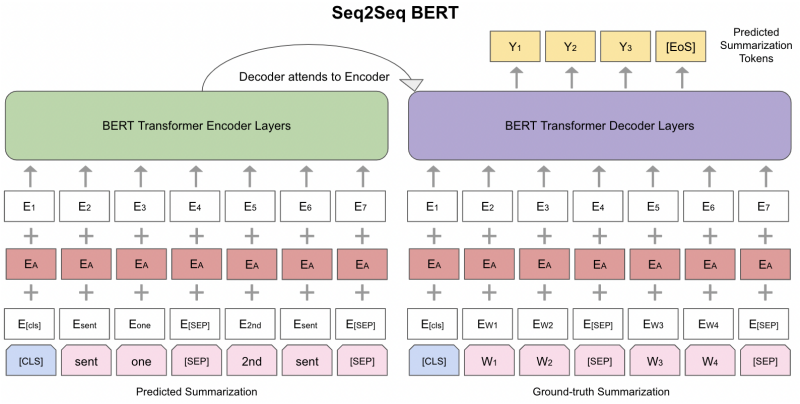

Figure 2: Seq2Seq BERT model (baseline) architecture.

Baselines. We use a T5 model (model #1 in Table 1) pretrained on the C4 dataset [14] in abstractive summarizer mode without fine-tuning as a baseline. T5 has been explicitly trained in the abstractive summarization task, so it can be applied without any fine-tuning to get a rough baseline. As a comparison, we also fine-tune some vanilla Bert-Based Seq2Seq architectures (Fig 2) with ClinicalBioBERT Transformers [17] (model #2 in Table 1) as the encoder and decoder to evaluate the efficacy of sentence-level representation learned with the BertSum[2] architecture. The last baseline is a pre-trained BERTSumExtAbs model [18] (model #3 in Table 1) without fine-tuning. It uses a BERT Transformer model as the encoder and a randomly initialized Transformer as the decoder, pre-trained with CNN/DailyMail [21] news articles.

Table 1 gives an overview of the models we considered for the main approach and baseline.

4 Experiments

Data. We use the MIMIC-CXR dataset [1] [22] with free-text reports of patients' DICOM scans. The "Findings" and the "Indication" (if present) sections are used as inputs for the summarization model, and the "Impression" section is the output prediction. We show an example radiology report in Fig 3. For data preprocessing, we filter out reports without "Findings" or "Impression" sections, or with "Impression" section longer than the "Findings" and "Indication". We then split the dataset into train/val/test sets with 95329/11995/11887 reports respectively. Note that our MIMIC-CXR dataset is mostly different from MIMIC-III ClinicalBioBERT pre-trained on. MIMIC-CXR features only chest x-ray radiology reports from 2011-2016 derived from 227,835 radiographic studies, whereas MIMIC-III has a variety of text-based information associated with roughly 40,000 patients between 2001 and 2012.

```
EXAMINATION:  CHEST (PA AND LAT)

INDICATION:  ___ year old woman with ?pleural effusion // ?pleural effusion

TECHNIQUE:  Chest PA and lateral

COMPARISON:  ___

FINDINGS:

Cardiac size cannot be evaluated. Large left pleural effusion is new.  Small
right effusion is new.  The upper lungs are clear.  Right lower lobe opacities
are better seen in prior CT.  There is no pneumothorax.  There are mild
degenerative changes in the thoracic spine

IMPRESSION:

Large left pleural effusion
```

Figure 3: An example data record from MIMIC-CXR [1]. We use "Indication" and "Findings" as input and "Impression" as output. Note that all PII information has been replaced with "___"

Evaluation method.
We evaluate the performance of our summarization model on the test set using Recall/Precision/F1 of ROUGE score[16]. ROUGE-1 and ROUGE-2 measure the unigram and bigram overlap, and ROUGE-L measures the common subsequence overlap between the predicted and reference summaries to evaluate the informativeness of the summary. For example, recall and precision of ROUGE-L are calculated as

$$R_{ROUGE-L} = \frac{LCS(X,Y)}{m} \quad P_{ROUGE-L} = \frac{LCS(X,Y)}{n} \tag{2}$$

where X of length m is reference summary and Y of length n is candidate summary. LCS is the maximum length of the Longest Common Sequence.

One problem with ROUGE is that it only measures exact words overlaps which may be imperfect for abstractive summarization, since a good abstractive summarization could include words with similar meanings that do not appear in the reference. Thus, we used another automatic evaluation metric BERTscore [6] that compares the pair-wise cosine similarity of the pre-trained BERT contextual embeddings of all the words in candidate and reference summarization. Since it measures soft overlap between token embeddings, it is good for paraphrase detection and effectively captures distant dependencies and ordering. It also correlates highly with human evaluations [6]. To better evaluate how similar reference and prediction summarizations are with respect to the medical semantics, we use ClinicalBioBERT instead of BERT to generate embeddings for the summarization and compute the ClinicalBioBERT-scores accordingly.

The numerical range of BERTScore is between -1 and 1, following [6], we rescaled it through a simple linear transformation of $score_{\text{rescaled}} = \frac{Score_{\text{original}} - base}{1 - base}$. The base is calculated by selecting a million sentences from WMT16 English text data, group two sentences randomly, compute ClinicalBioBERT-score on the random pairs and take the average. Intuitively, the base is the score the candidate-reference pair could get, where the candidate is generated by the baseline model that makes random prediction. The rescaled score is generally between 0 and 1 with a reliable baseline.

Experimental details.

Hyperparameters for ClinicalBioBERT Seq2Seq: (Training 8 hours, eval 15 minutes on a nvidia-T4)

```
max_length = 128
min_length = 2
no_repeat_ngram_size = 3
early_stopping = True
length_penalty = 2.0
num_beams = 4
```

```
epochs = 4
optimizer = AdamW
lr = 1e-3
lr_scheduling = linear warmup, linear decrease
iterations = 120,000
batch_size = 2
```

```
Hyperparameters for ClinicalBioBERTSumExt: (Training 5.5 hours, eval 8 minutes on a nvidia-tesla-v100 GPU)

    ClinicalBioBERT Encoder configuration: follows the pre-trained model
    Transformer Encoder Configuration:
    num_layer = 2; hidden_size = 768; num_attention_heads = 8; feed_forward_filter_size=2048; dropout_rate=0.1
    Training:
    optimizer: Adam(beta_1=0.09, beta_2=0.999); iterations = 50000; batch_size = 3000;
    warming up learning rate: lr=0.002; warmup_steps=10000
    accumulate the gradient every 2 steps


Hyperparameters for ClinicalBioBERTSumAbs: (Training 5.5 hours, eval 8 minutes on a nvidia-tesla-v100 GPU)

    ClinicalBioBERT Encoder configuration: follows the pre-trained model
    Decoder Configuration:
    num_layers = 6; hidden_size = 768; num_attention_heads = 8; feed-forward_filter_size = 2048; dropout_rate = 0.2
    Training:
    optimizer= Adam(beta_1=0.09, beta_2=0.999); iterations = 50000; batch_size = 140;
    warming up learning rate: lr_bert= 0.002; warmup_steps_bert 20000; lr_decoder=0.02; warmup_steps=10000
    note that the lr_bert is set to be much smaller than lr_decoder since it is pretrained
    accumulate the gradient every 5 steps
    Evaluation:
    beam search length normalization coefficient = 0.95; max_length=200; min_length=3
```

#	Model	Encoder	FT*	Decoder	FT*
1	T5	pretrained T5	×	pretrained T5	×
2	ClinicalBioBERT Seq2Seq	pretrained ClinicalBioBERT	✓	pretrained ClinicalBioBERT	✓
3	BERTSumExtAbs	BERT pretrained with CNNDM	×	Transformer pretrained with CNNDM	×
4	ClinicalBioBERTSumAbs	pretrained ClinicalBioBERT	×	Transformer trained from scratch	✓
5	ClinicalBioBERTSumAbs	pretrained ClinicalBioBERT	✓	Transformer trained from scratch	✓
6	ClinicalBioBERTSumExt	pretrained ClinicalBioBERT	✓	N/A	✓
7	ClinicalBioBERTSumExtAbs	pretrained ClinicalBioBERT	✓*	Transformer trained from scratch	✓

Table 1: Model architecture. FT* is Whether the encoder/decoder is fine-tuned on our MIMIC-CXR dataset. Model 1-3 are our baselines and model 4-7 are our main models. ✓* means the encoder is fine-tuned twice, first with the extractive objective then with the abstractive objective.

Results. We report the performance of the baselines and our main models in Fig 4 and include results for all the models we experimented with in the appendix. We show the training curves in Fig 5.

#	ROUGE-1			ROUGE-2			ROUGE-L			ClinicalBioBERT-score		
	R	P	F1	R	P	F1	R	P	F1	R	P	F1
1	20.86	19.10	17.07	9.11	7.19	6.94	18.45	17.58	15.4			
2	58.20	46.96	48.84	43.44	36.00	36.94	55.33	45.02	46.67	0.50	0.55	0.52
3	31.69	18.82	21.00	13.17	8.66	9.31	30.01	17.60	19.70			
4	52.79	65.28	55.18	42.05	51.35	43.71	51.53	63.63	53.85			
5	57.96	**66.95**	**58.97**	46.44	**53.37**	**47.06**	56.38	**65.09**	**57.37**	0.51	**0.61**	**0.55**
6	48.31	20.97	26.60	23.36	11.07	13.70	45.64	19.60	24.92	0.33	0.09	0.20
7	**58.75**	65.87	58.91	**46.70**	51.89	46.52	**57.07**	63.84	57.18	**0.51**	0.60	0.54

Figure 4: Recall, Precision and F1 of ROUGE and ClinicalBioBERT-scores

The moderate performance of the pre-trained T5 model (#1) in abstractive summarizer mode without finetuning indicates that it is reasonable to formulate predicting doctor's impressions for radiology reports as an abstractive text summarization problem. It also serves as a baseline for how a summarization model pre-trained without specifically any medical data may perform on this niche task. Compared to the model trained in extractive setting (#6), the significantly better performance of the model trained in the abstractive setting (#5) further validates our formulation for the abstrative summarization task.

Notice that all of the models using domain-specific pre-trained ClinicalBioBERT (#2, #4,#5,#6,#7) perform significantly better than the other baselines pre-trained only on common English texts (#1,

#3). This reveals the efficacy of domain-specific pre-training for language models. Besides, fine-tuning the pre-trained models on our dataset greatly helps to improve the summarization quality, as seen in the difference between model #4 and model #5.

Our best performing model (#5) obtains significantly higher precision than the baseline of finetuned ClinicalBioBERT Seq2Seq model (#2). We think this could be because instead of learning word embeddings as in the Seq2Seq model, ClinicalBioBERTSumAbs distinguishes different sentences in the input by adding extra [CLS] tokens and using interval segment embeddings. The sentence-level embeddings it learns helps to better capture the inter-sentence relationships and thus improves the overall quality of summarization.

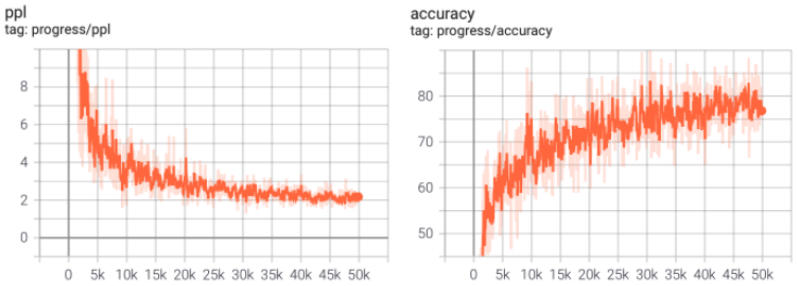

Figure 5: ClinicalBioBERTSumAbs's training curve, perplexity (left), accuracy (right).

Among models using the pre-trained ClinicalBioBERT[17] encoder, the purely extractive model (#6), unsurprisingly, struggles to perform well in this task. This is because the reference impressions are often not sentences directly extracted from the findings or indication. This is especially true in the examples where the summarization contains doctor's conclusion based on findings, such as "No acute cardiopulmonary pathology".

Surprisingly, unlike the findings in [2] that combining extractive and abstrative objetives helps to improve the summarization quality, we find that the two-stage fine-tuning barely improve the performance of the abstractive summarization model, as seen by comparing model #7 to model #5. We think this is because the worddings of the summarization vary a lot from those in the original radiology report and the performance of the extractive model is relatively bad on its own. The interesting trend we notice is that for both ROUGE-Based scores and ClinicalBioBERT-Score, the recall of the ext-abs model with two-stage fine-tuning is higher than the abstractive model. We think this is because when the model is first trained in the extractive setting, the encoder learns to pay higher attention to the sentences in the input with the largest overlap with the reference summarization. Therefore, when the encoder is then trained with the abstractive objective, it would be geared towards generating summarization mostly based on the sentences in the input that maximally cover the contents in the reference summarization, and thus yielding high recall rate.

5 Qualitative Analysis

Example Input.

```
INDICATION: Right lower quadrant pain and recent pneumonia, evaluate for pneumonia.

FINDINGS:   PA and lateral views of the chest.
            Previously seen pneumonia in the right lower and mid lung are no longer apparent.
            The lungs are clear. The cardiac, mediastinal, and hilar contours are normal.
            There is no pleural effusion or pneumothorax.
            There is no pulmonary vascular congestion.

IMPRESSION: Resolution of previously seen pneumonia. No new consolidations.
```

Example Output.

```
Ext Model:  Previously seen pneumonia in the right lower and mid lung are no longer apparent.
            There is no pulmonary vascular congestion.
```

```
Abs Model:  No acute cardiopulmonary abnormality.
Ext + Abs:  Previously seen pneumonia in the right lower and mid lung are no longer apparent.
```

As seen in the example above, all three models gives reasonable summarization predictions. The extractive model picks the sentences in INDICATION and FINDINGS that are most similar to the reference IMPRESSION. In contrast, the abstractive model gives concise and informative summarization of the INDICATION and FINDINGS with words like "abnormality", which do not appear in the input text but capture the essence of the information included in the input context. It is interesting that the abstractive model generates the well-defined medical word "cardiopulmonary" that accurately captures descriptions for both cardiac and pulmonary conditions in the Findings section. It is the only one among the three models that manage to do so. Comparing with the abstrative model, the Ext-Abs model with two-stage fine-tuning generates more words from the original INDICATION and FINDINGS.

6 Conclusion

In this project, we build ClinicalBioBERTSum by fine-tuning the ClinicalBioBERT [17] Transformer encoder and training a Transformer decoder end-to-end. It takes the Finds and Indications sections of a radiology report as input and predicts the Impression section via abstractive summarization, which greatly helps reduce the human labor for doctors and patients to manually dig through the documents. To better evaluate the abstractive summarization generated by our model, we use ROUGE scores capturing exact word-matching and BERT-based scores taking account of synonyms and different word-orderings. Our model generates concise and informative abstractive summarization, achieving a ROUGE-F1 score of up to 57.37/100 and a ClinicalBioBERTScore of up to 0.55/1.00.

With comprehensive experiments, we showcase that pre-trained domain-specific language models and sentence-aware embeddings significantly improves the quality of summarizing radiology reports. Our transformer model trained on the radiology reports may serve as good checkpoints for other NLP models dealing with radiology reports, such as auto-labeling models or models that highlight relevant parts of lengthier findings sections.

Future work includes using our custom tokenizer to train a BERT-scale model specifically for radiology reports. Another future direction is to fully evaluate the performance of our model by using the resulting summarization in downstream NLP tasks. Alternatively, incorporating human evaluation, especially from doctors, could offer further insights into the quality of the summarizations our model generates.

A Appendix (optional)

#	Model	Encoder	FT*	Decoder	FT*
1	T5	pretrained T5	×	pretrained T5	×
2	BERTSumExtAbs	BERT pretrained with XSUM	×	Transformer pretrained with XSUM	×
3	BERTSumExtAbs	BERT pretrained with CNNDM	×	Transformer pretrained with CNNDM	×
4	ClinicalBioBERTSumAbs	pretrained ClinicalBioBERT	×	Transformer trained from scratch	✓
5	BERT Seq2Seq	pretrained BERT	✓	pretrained BERT	✓
6	ClinicalBioBERT Seq2Seq	pretrained ClinicalBioBERT	✓	pretrained ClinicalBioBERT	✓
7	ClinicalBioBERTSumAbs	pretrained ClinicalBioBERT	✓	Transformer trained from scratch	✓
8	ClinicalBioBERTSumExt	pretrained ClinicalBioBERT	✓	not decoder(n/a)	✓
9	ClinicalBioBERTSumAbsExt	pretrained ClinicalBioBERT	✓*	Transformer trained from scratch	✓
10	custom MiniBERT	MiniBERT from scratch	✓	MiniBERT from scratch	✓

Table 2: The architecture of all the models we experimented with

Model #	ROUGE-1			ROUGE-2			ROUGE-L		
	R	P	F1	R	P	F1	R	P	F1
1	20.86	19.10	17.07	9.11	7.19	6.94	18.45	17.58	15.4
2	6.93	6.93	5.99	0.88	0.87	0.76	6.37	6.23	5.42
3	31.69	18.82	21.00	13.17	8.66	9.31	30.01	17.60	19.70
4	52.79	65.28	55.18	42.05	51.35	43.71	51.53	63.63	53.85
5	57.15	47.41	48.76	40.82	34.32	35.03	53.84	44.99	46.16
6	58.20	46.96	48.84	43.44	36.00	36.94	55.33	45.02	46.67
7	57.96	**66.95**	**58.97**	46.44	**53.37**	**47.06**	56.38	**65.09**	**57.37**
8	48.31	20.97	26.60	23.36	11.07	13.70	45.64	19.60	24.92
9	58.75	65.87	58.91	**46.70**	51.89	46.52	**57.07**	63.84	57.18
10	**67.26**	10.35	16.5	47.98	5.62	9.18	60.38	8.45	13.69

Table 3: The performance of all the models we experimented with

References

[1] 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] Yang Liu and Mirella Lapata. Text summarization with pretrained encoders. *arXiv preprint arXiv:1908.08345*, 2019.

[3] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks, 2014.

[4] Sebastian Gehrmann, Yuntian Deng, and Alexander M. Rush. Bottom-up abstractive summarization, 2018.

[5] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units, 2016.

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

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

[8] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2020.

[9] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, Sep 2019.

[10] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035, 2016.

[11] Qingyu Zhou, Nan Yang, Furu Wei, Shaohan Huang, Ming Zhou, and Tiejun Zhao. Neural document summarization by jointly learning to score and select sentences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 654–663, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[12] Yue Dong, Yikang Shen, Eric Crawford, Herke van Hoof, and Jackie Chi Kit Cheung. Banditsum: Extractive summarization as a contextual bandit, 2019.

[13] Thomas Peel Léo Bouscarrat, Antoine Bonnefoy and Cécile Pereira. Strass: A light and effective method for extractive summarization based on sentence embeddings. 2019.

[14] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, 2020.

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

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

[17] Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics.

[18] Bertsum. `https://github.com/nlpyang/BertSum`.

[19] Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

[20] Henry Tsai, Jason Riesa, Melvin Johnson, Naveen Arivazhagan, Xin Li, and Amelia Archer. Small and practical bert models for sequence labeling. *arXiv preprint arXiv:1909.00100*, 2019.

[21] Karl Moritz Hermann, Tomáš Kočiskỳ, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. *arXiv preprint arXiv:1506.03340*, 2015.

[22] Mimic-cxr database. `https://physionet.org/content/mimic-cxr/2.0.0/`.