CSTRL: Context-driven Sequential Transfer Learning for Abstractive Radiology Report Summarization

Mst. Fahmida Sultana Naznin[1], Adnan Ibney Faruq[1], Mostafa Rifat Tazwar[1],
Md Jobayer[2], Md. Mehedi Hasan Shawon[2], Md Rakibul Hasan[2,3]

[1]Bangladesh University of Engineering and Technology, Dhaka 1000, Bangladesh
[2]BRAC University, Dhaka 1212, Bangladesh
[3]Curtin University, Bentley WA 6102, Australia
Correspondence: rakibul.hasan@curtin.edu.au

Abstract

A radiology report comprises several sections, including the Findings and Impression of the diagnosis. Automatically generating the Impression from the Findings is crucial for reducing radiologists' workload and improving diagnostic accuracy. Pretrained models that excel in common abstractive summarization problems encounter challenges when applied to specialized medical domains, largely due to the complex terminology and the necessity for accurate clinical context. Such tasks in medical domains demand extracting core information, avoiding context shifts, and maintaining proper flow. Misuse of medical terms can lead to drastic clinical errors. To address these issues, we introduce a sequential transfer learning that ensures key content extraction and coherent summarization. Sequential transfer learning often faces challenges like initial parameter decay and knowledge loss, which we resolve with the Fisher matrix regularization. Using MIMIC-CXR and Open-I datasets, our model, *CSTRL*—**C**ontext-driven **S**equential **TR**ansfer **L**earning—achieved state-of-the-art performance: 56.2% improvement in BLEU-1, 40.5% in BLEU-2, 84.3% in BLEU-3, 28.9% in ROUGE-1, 41.0% in ROUGE-2 and 26.5% in ROGUE-3 score over benchmark methods. We further analyze factual consistency scores while preserving the medical context. Our code is publicly available at https://github.com/fahmidahossain/Report_Summarization.

1 Introduction

A radiology report summarizes a radiologist's analysis of imaging data, detailing sections like Type of Exam, Clinical History, Technique, Findings, and Impression (Cai et al., 2023). Findings are observations from the examination of diagnosed body parts, classified as normal, abnormal, or potentially abnormal (Kareem, 2024). Impression is a summary of Findings, including possible causes (Kareem, 2024). Quality of contextual, informative, and factual correctness of generated Impression is subpar due to limited training for writing Impression (≤ 1 hour/year for 86% of radiologists) (Hartung et al., 2020). Synthesizing Impression from the Findings is, therefore, crucial in automated radiology report summarization.

Writing Impression from Findings falls under the *abstractive summarization* task that distills key clinical insights from diagnostic data (Herts and Berland, 2021). Extractive summarization selects sentences from the original text to create a summary, while abstractive summarization generates a new paragraph to summarize the content of the original document (Cai et al., 2023). Abstractive summarization is more complex than extractive summarization but yields more flexible and concise summaries (Cai et al., 2023). However, systematic evaluation of abstractive summarization across diverse domains is limited (Zhang et al., 2020a). One challenge of abstractive summarization in the clinical domain is managing the balance between providing concise, clinically relevant information and avoiding excessive technical language. For example, *"Liver metastases have enlarged"* is more straightforward and actionable than *"There is re-demonstration of multiple liver lesions consistent with metastases, which have increased in size in the interval"* (Hartung et al., 2020). Including irrelevant Findings can create ambiguity. For example, *"diverticulosis without diverticulitis"* (Hartung et al., 2020): diverticulosis indicates the presence of diverticula in the colon, which is typically benign, but, without proper context, it may raise concerns about potential complications and suggest that diverticulitis (inflammation of diverticula) could develop. To prioritize core sentences of Findings and preserve context, we propose a novel approach combining Sequential Transfer Learning with a Knowledge Distillation (KD) strategy, which automates the creation of Impression from the Find-

26526

ings of a radiology report.

Deep learning techniques for abstractive summarization face specific challenges with radiology corpora. A key limitation is the difficulty in evaluating performance on datasets dense with radiology-specific terminology, along with significant differences in word distributions between general and radiology report domains (Cai et al., 2023). Previous works (Section 2) have several shortcomings. Firstly, there is a lack of a proper methodology for generating Impression that focuses on primary observations (Hartung et al., 2020). We propose an optimized Gap Sentence Generation (GSG) technique to identify key Findings. Secondly, subtle differences in terminology can obscure the context and misrepresent the clinical scenario (Codish and Shiffman, 2005). In response, we propose a contextual tagging method. Thirdly, poor BLEU scores of recent studies indicate that the generated Impression section often fails to accurately reflect core points from the Findings section (Cai et al., 2023). It affects the readability, informativeness, and reliability of reports for clinical use. Lastly, reducing dimensionality and computational complexity for real-time production is a challenge (Meng et al., 2022; Che et al., 2015; Liu et al., 2020), for which we propose a KD method. Our key contributions in constructing Impression through proper clinical methodology with high performance and factual consistency are as follows:

- We propose an optimized GSG technique for pivotal sentence identification.
- We introduce a sequential transfer learning framework with KD using the T5 (Text-to-Text Transfer Transformer) model, fine-tuned on a radiology corpus. This approach aims to transfer knowledge from the GSG task to summarization effectively and gradually. To mitigate catastrophic forgetting during this sequential knowledge transfer, our model includes Fisher matrix regularization with penalty adjustment.
- We introduce a contextual tagging approach with a Named Entity Recognition (NER) system to capture and preserve essential contextual information by entity linking to the MR-CONSO database, a component of the Unified Medical Language System (UMLS) Metathesaurus. It provides a structured representation of medical concepts and terms from various sources.
- We compare our proposed summarization

model, CSTRL, against state-of-the-art models trained on the same dataset. We further train decoder-only models for comparison. CSTRL consistently outperforms all baselines across various settings. We further analyze factual consistency and conduct a comprehensive human evaluation to validate the quality and reliability of the generated summaries.

2 Related Work

In recent years, deep learning has significantly improved neural abstractive summarization tasks in NLP (Rush et al., 2015; Chopra et al., 2016; Tan et al., 2017; Zhou et al., 2017; Li et al., 2017). Traditional models are primarily trained on general datasets like CNN/Daily Mail (Nallapati et al., 2016) and Gigaword Corpus (Sutskever et al., 2014). The attention-based seq2seq model by Rush et al. (2015) and its extension by Chopra et al. (2016) laid the groundwork for these advancements. Recently, large-scale pre-trained language models have shown impressive summarization results (Karn et al., 2021; Raffel et al., 2020; Kieuvongngam et al., 2020; Chang et al., 2021). Kieuvongngam et al. (2020) employed BERT and GPT for summarizing COVID-19 research. However, unique word distributions in radiology corpora limit the applicability of these techniques (Cai et al., 2023).

Zhang et al. (2018) first explored automatic radiology Impression generation, followed by MacAvaney et al. (2019), who introduced an ontology-aware pointer-generator model to enhance summarization quality. A background-augmented pointer-generator network with copy and background-guided decoding was proposed in Zhang et al. (2020b), and a word graph captured critical words and relations in Hu et al. (2021). Anatomies were extracted, radiographs encoded, and fused with anatomy-enhanced co-attention in Hu et al. (2023). Sotudeh Gharebagh et al. (2020) enhanced clinical summarization by augmenting ontological terms, and Hu et al. (2022) integrated Findings by a unified framework with knowledge via text and graph encoders. Karn et al. (2022) proposed a two-step extractive-abstractive method using a Differentiable Multi-Agent Actor-Critic framework.

Li et al. (2018) proposed a hybrid retrieval-generation agent that integrates human knowledge with neural networks for medical report generation. Models like OpenAI GPT (Chang et al.,

$$F_{ij} = E[(\frac{\partial \log p(y \mid x;\theta)}{\partial \theta_i})(\frac{\partial \log p(y \mid x;\theta)}{\partial \theta_j})]$$

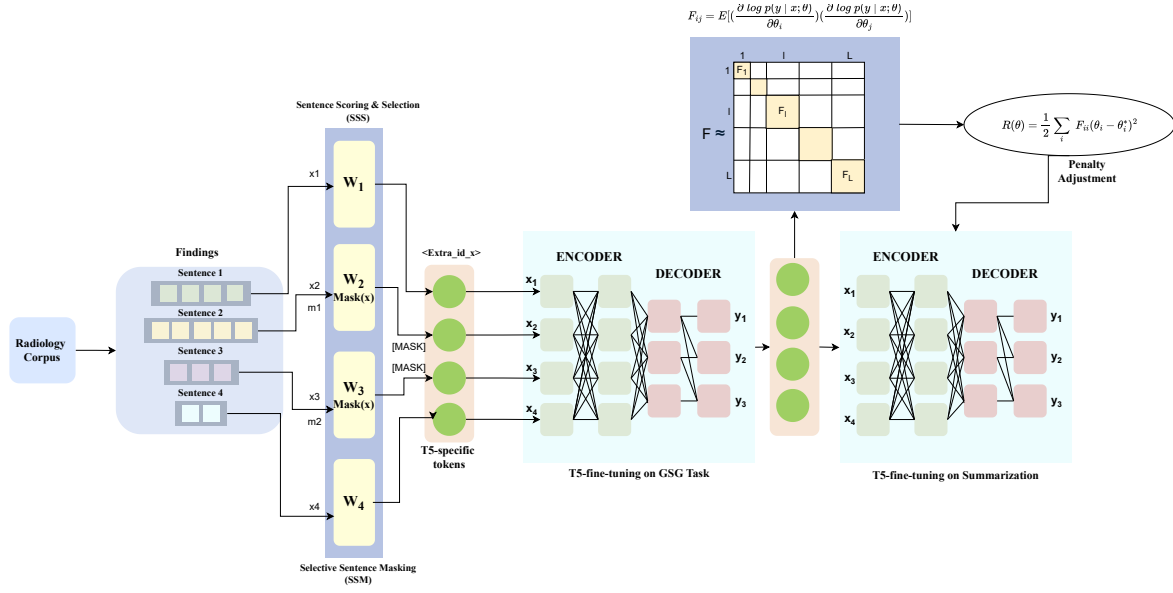$$R(\theta) = \frac{1}{2}\sum_i F_{ii}(\theta_i - \theta_i^*)^2$$

Figure 1: Workflow of the proposed scheme for Sequential Transfer Learning. Firstly, sentences from the radiology corpus are scored with ROUGE and BLEU metrics, and top-scoring sentences are masked. Secondly, the T5 model is trained on the GSG task to predict the masked sentences. The Fisher Information matrix is calculated to identify critical parameters. Lastly, the penalty term is adjusted during fine-tuning for summarization.

2021), BERT (Devlin et al., 2019), ELMo (Peters et al., 2018), and XLNet (Yang et al., 2019) have improved various NLP tasks through extensive external knowledge. BioBERT (Lee et al., 2020) captures semantic features in the biomedical field, pre-trained on large corpora like PubMed abstracts and PMC articles. Recently, Cai et al. (2023) proposed a pre-trained language model, ChestXrayBert, specifically designed for summarizing chest radiology reports. However, these studies may struggle with effective Impression generation methodologies.

3 Method

We adopt a sequential two-step approach to fine-tune the T5 model (Raffel et al., 2020) using a dataset on radiology corpus optimized by the GSG methodology (Zhang et al., 2020a). In the first step, the model is trained on a GSG task, where it predicts masked sentences, thereby gaining a deeper understanding of the crucial statements for crafting coherent summaries in the Impression section. In the second step, we fine-tune the model for summarization using the learned weights as initial parameters and employ Fisher matrix regularization for effective GSG-trained knowledge transfer to the summarization task. Then, the GSG fine-tuned model drives sequential knowledge distillation to achieve dimensionality reduction.

3.1 Proposed Sequential Transfer Learning

Optimized GSG Technique. We propose a novel pre-training methodology for GSG that extends BERT's Masked Language Modeling (MLM) (Cai et al., 2023) objective by integrating an enhanced version of the GSG strategy from PEGASUS (Zhang et al., 2020a). This approach uses sentence scoring and selection for identification of vital sentences.

Sentence Scoring & Selection (SSS). Unlike PEGASUS, which relies solely on ROUGE, our optimized model employs a composite metric that combines ROUGE and BLEU for a more comprehensive assessment of fundamentals by quantifying the overlap and similarity between sentences. In the radiology domain, consistent terminology and phrasing is used in identical medical conditions (Khorasani et al., 2003; Panicek and Hricak, 2016; for Disease Control et al., 2016). This consistency makes n-grams essential for capturing the precise repetition of critical terms to ensure accurate contextual representation (for Disease Control et al., 2016). BLEU's emphasis on exact n-gram matches ensures clinical accuracy (Ibrahim et al., 2023; Santhosh, 2023). For each sentence x_i in the document D, the priority score W_i is computed as:

$$W_i = F1(\text{ROUGE}(x_i, D \setminus \{x_i\}) + \text{BLEU}(x_i, D \setminus \{x_i\})) \quad (1)$$

Selective Sentence Masking (SSM). We incorporate a Selective Sentence Masking (SSM) specifically designed for a radiology corpus. We applied three masking rules: (1) sentences ≥ 5 words received 3 masks, (2) sentences = 4 received 2 masks, and (3) sentences ≤ 3 words received 1 mask. The set of sentences selected for masking is denoted as $M = \{m_1, m_2, \ldots, m_k\}$, derived from a masking function $\text{Mask}(x)$ based on SSS. Our dataset's sentences are predominantly up to 9 sentences long, with 95% fitting this length. To create the masked text in the masking function, we replace each sentence in the set M with a placeholder token $[MASK]$ in the original text x. By applying this methodology to the dataset, we create a new dataset where selected sentences are systematically masked. One additional column contains the original sentences with the masked portions, while another column specifically lists the masked sentences alongside the existing columns.

Fine-Tuning for GSG Task. We replace the [MASK] in the GSG-optimized dataset with T5-specific placeholders (<extra_id_x>). The dataset $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ comprises input sequences x_i and their corresponding target sequences y_i which are derived from additional columns in the GSG-optimized dataset. Both x_i and y_i are tokenized with the pre-trained T5 tokenizer, applying truncation and padding for uniform length. The model is trained to predict the masked sentences.

3.2 Sequential Transfer Learning for Summarization

After fine-tuning the model on the GSG task, we utilize its learned weights as the initial parameters for the summarization task. However, we observe that direct sequential fine-tuning significantly updates the model parameters, which risks losing the valuable knowledge acquired during the GSG training. Our objective is to gradually transfer the knowledge from the GSG-trained model to enhance summarization without overwriting it. To resolve this issue, we implement Fisher Matrix Regularization (FMR), inspired by Elastic Weight Consolidation (Kirkpatrick et al., 2017) to diminish the problem of catastrophic forgetting. When a model erases prior knowledge while assimilating new tasks, compromising its ability to retain past information, it experiences catastrophic forgetting—a fundamental challenge in sequential learning. Deep neural networks, inherently vulnerable to this, require explicit regularization to mitigate its effects. The FMR is widely used as a pivotal mechanism to address this (Wang et al., 2022; Michel et al., 2019; Gupta et al., 2021; Kutalev, 2020; Liu et al., 2018; Ritter et al., 2018; Kemker et al., 2018) We compute the Fisher Information Matrix F (FIM) based on the GSG task to identify key parameters that should be preserved. The FIM is defined as:

$$F_{ij} = \mathbb{E}\left[\left(\frac{\partial \log p(y|x;\theta)}{\partial \theta_i}\right)\left(\frac{\partial \log p(y|x;\theta)}{\partial \theta_j}\right)\right] \tag{2}$$

where $p(y|x;\theta)$ represents the model's likelihood given the input data x and parameters θ. The matrix F quantifies how sensitive the performance is to changes in each parameter, highlighting those that are paramount for retaining the knowledge gained from the GSG task. During the fine-tuning process on the summarization task, we introduce a penalty term $R(\theta)$ to limit significant changes to these crucial parameters: $R(\theta) = \frac{1}{2}\sum_i F_{ii}(\theta_i - \theta_i^*)^2$ where θ^* denotes the optimal parameters obtained from the GSG task. The penalty term $R(\theta)$ is dynamically adjusted and gradually reduced over the training epochs. Figure 1 represents the overall architecture. This approach enables effective adaptation to summarization while retaining key knowledge from GSG training.

3.3 Contextual Tagging for Clinical Precision

We note that overgeneralization or overcomplication in abstractive summarization leads to significant misinterpretations. Such distortions, especially in clinical contexts, critically alter the intended meaning of Findings. To address this issue, we propose a methodology for developing an optimized NER system to obtain core contextual knowledge. We identify key medical terms from Impression for entity recognition using the Term Frequency-Inverse Document Frequency (TF-IDF) method (Christian et al., 2016). After applying TF-IDF vectorization, we generate a matrix representing each document in a high-dimensional space. The top n keywords, denoted as K, are extracted based on TF-IDF scores. Then, we filter data from the MRCONSO table to include only relevant English terms. The table maps medical terminologies by assigning each concept a unique CUI. It preserves context by linking terms to their source vocabulary and language. These identifiers ensure that variations in spelling or phrasing are
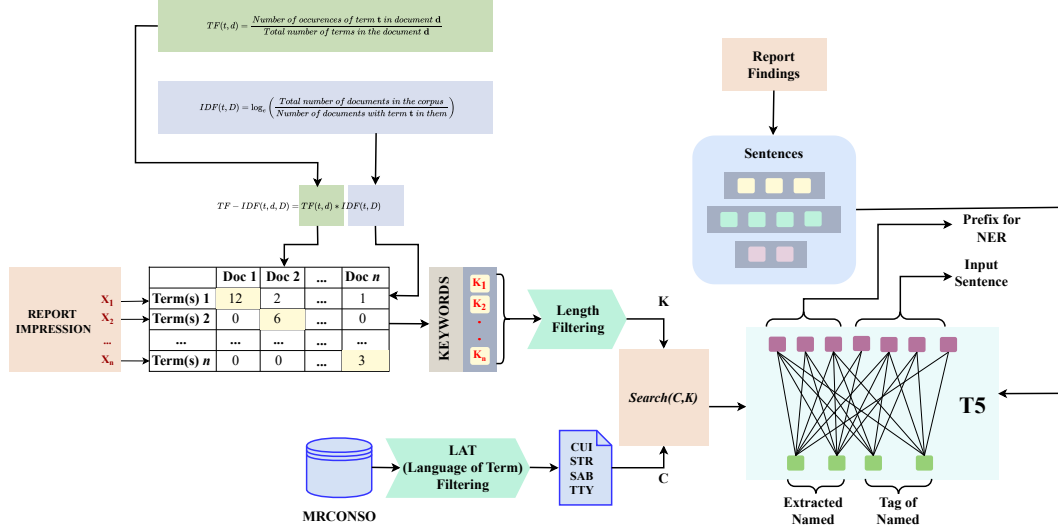
Figure 2: Workflow of the proposed scheme for contextual tagging. Firstly, keywords from the Impression section are extracted using TF-IDF vectorization. Secondly, relevant clinical terms are retrieved from the MRCONSO table, and tags are selected by searching keywords in the filtered table. Finally, the T5 model is trained to predict tags from the Findings section.

consistently mapped to the correct concept. It maintains semantic integrity across systems. The filtered dataset from MRCONSO includes the Concept Unique Identifier (CUI), String (STR), Source Abbreviation (SAB), and Term Type (TTY). They serve as a reference list of tags C. We search the extracted keywords K in C and make a new column of tags to their corresponding Impression. Then, we train the model using T5 to generate tags from Findings. Figure 2 shows the architecture of proposed contextual tagging.

3.4 Proposed Teacher-Student Framework for Knowledge Distillation

The GSG fine-tuned summarization model acts as a teacher, generating logits that serve as soft targets for the student model as shown in Figure 3. This allows the student to learn from the teacher's outputs rather than relying solely on hard labels. We employ a sequential training algorithm, where each batch undergoes multiple forward and backward passes. We use the T5 tokenizer to convert Findings and Impressions into token IDs and attention masks. These are fed into an embedding layer, mapping tokens to dense vector representations. The shared embedding framework between teacher and student models ensures consistency, improving performance in downstream tasks. To train the student model, we combine cross-entropy loss with distillation loss, allowing the student to learn from both ground-truth labels and softened outputs from
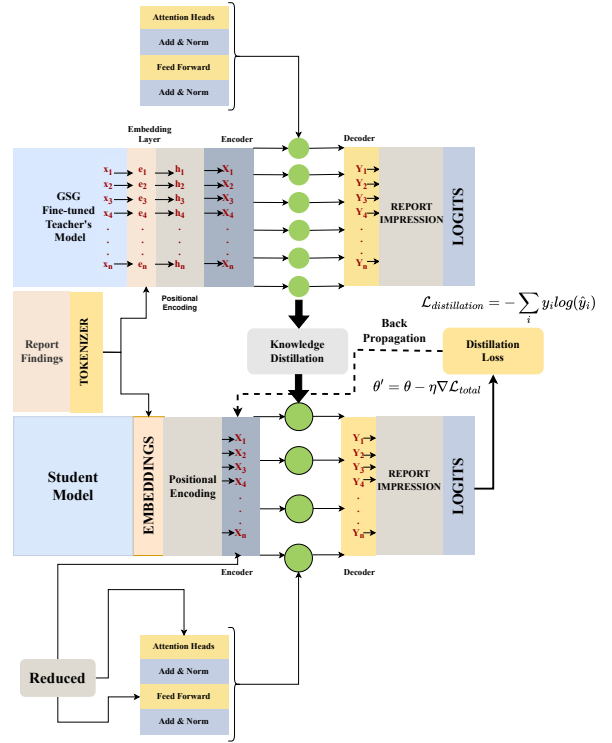

Figure 3: Schematic representation of the proposed knowledge distillation network. The combined loss function is calculated using the reduced feature layers and the Hard Labels loss. Effective back-propagation ensures the optimization of the Student Model.

the teacher. The combined loss is defined as:

$$\mathcal{L} = (1 - \alpha)\mathcal{L}_{CE} + \alpha\mathcal{L}_{KL} \qquad (3)$$

Dataset	# Reports	Findings (Avg.)			Impressions (Avg.)		
		# S	# W/S	# Source W	# S	# W/S	# Source W
MIMIC-CXR	121,975	5.47	10.08	55.09	1.94	8.50	16.46
Open-I	3,312	4.62	8.03	37.06	1.81	5.52	9.98

S: Sentence, W: Words

Table 1: Summary of dataset statistics of Findings and Impression

where $\mathcal{L}_{CE}$ is the cross-entropy loss, $\mathcal{L}_{KL}$ is the Kullback-Leibler (KL) divergence loss between teacher and student logits, and α is a tunable parameter. The cross-entropy loss $\mathcal{L}_{CE}$ is given by:

$$\mathcal{L}_{CE} = -\sum_{i=1}^{N} Y_i \log(\text{softmax}(S_i)) \qquad (4)$$

with S representing student logits, Y as ground-truth labels, and N as the number of samples. To compute the distillation loss, we apply temperature scaling to soften the teacher model's logits as follows:

$$\mathcal{L}_{KL} = T^2 \cdot \text{KLDiv}(\text{softmax}(S/T), \text{softmax}(T_t/T)) \qquad (5)$$

where T_t refers to teacher logits. T is the temperature. This softening enables the student to approximate the teacher's output distribution more effectively, and the T^2 term ensures appropriate gradient scaling during backpropagation. Model parameters are updated via backpropagation to minimize the combined loss. Figure 4 summarizes our CSTRL, which uses sequential transfer learning, knowledge distillation, and contextual tagging for generated impressions.

4 Experiment and Evaluation

4.1 Dataset Description

We evaluate our model using two public chest X-ray report datasets: MIMIC-CXR (Johnson et al., 2024) and Open-I (National Library of Medicine, 2024), as shown in Table 1. MIMIC-CXR contains 227,835 reports, while Open-I, a public radiography dataset from Indiana University, originally included 7,430 reports. Findings in the MIMIC-CXR dataset exhibit an average of 5.47 sentences per Findings, with each sentence containing approximately 10.08 words. It has an average of 55.09 words per Findings, which suggests this dataset is rich in detail and complexity. Conversely, Open-I has slightly fewer sentences per Findings, averaging 4.62 sentences, with each sentence comprising

8.03 words and a total of 37.06 words. Impression reveals similar trends. MIMIC-CXR has an average of 1.94 sentences per Impression, with an average of 8.50 words per sentence, leading to an average of 16.46 words. In contrast, Open-I averages 1.81 sentences per Impression, with 5.52 words per sentence and 9.98 source words.

To ensure the quality of the data for medical report summarization, we apply several filtering criteria inspired by a GPT-Based Radiology Report Optimization study (Ma et al., 2024). We remove (1) incomplete reports that lack either Findings or Impression, (2) reports with a Findings section of fewer than 10 words, and (3) reports with an Impression section of fewer than 2 words. We retain 121,975 reports from MIMIC-CXR and 3,312 reports from Open-I. We then merge these datasets, a total of 125,287 reports. This merged dataset is divided into training, validation, and test sets using an 8:1:1 split.

4.2 Experimental Setup

Evaluation Metrics. To evaluate the quality of text summarization, we report F1 scores for ROUGE-1, ROUGE-2, and ROUGE-L (unigram, bigram, and longest common subsequence) (Lin, 2004). We also provide BLEU-1, BLEU-2, and BLEU-3 (unigram, bigram, and trigram) (Papineni et al., 2002). In addition, we measure the factual consistency of the generated summaries using SummaC-ZS and SummaC-Conv (Christian et al., 2016). We set `granularity` = "sentence" and `model_name` = "vitc".

Implementations and Model Details. We use PyTorch to implement our model on an NVIDIA A100 GPU. We configure the Teacher Model with 6 layers, 512 embedding dimensions, and 8 attention heads, while the Student Model has 3 layers, 128 dimensions, and 4 attention heads. Both use a vocabulary size of 32,128. Inputs are tokenized to 512 tokens, targets to 256. Models are trained for 20 epochs with a batch size of 32, using AdamW with a 0.003 learning rate. Logits are adjusted with a temperature of 20, and distillation loss is scaled by $\alpha = 0.7$. ROUGE and BLEU with the T5-small are used for key token identification, masking 3 sentences per Findings. We conduct a grid search throughout all experiments for tuning hyperparameters.
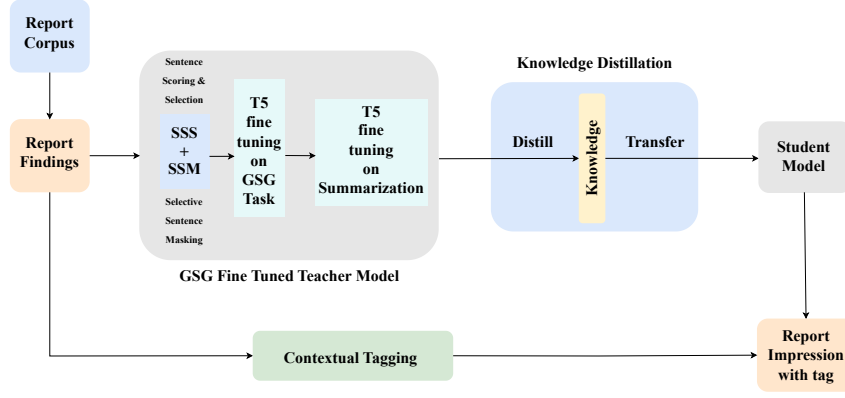
Figure 4: Summary of CSTRL: The GSG Finetuned Teacher model (Figure 1), using sequential transfer learning, is knowledge distilled, following the knowledge distillation process (Figure 3). Additionally, a contextual tagging method (Figure 2) is applied to the generated impressions.

Model	R-1	R-2	R-L	B-1	B-2	B-3
Content Selector (Sotudeh Gharebagh et al., 2020)	53.6	40.8	51.8	–	–	–
WGSUM (Trans+GAT) (Hu et al., 2021)	48.3	33.3	46.7	–	–	–
BASE+GRAPH+CL (Hu et al., 2022)	49.1	33.7	47.1	–	–	–
Seq2Seq (Cai et al., 2023)	35.1	23.7	35.4	24.3	11.5	4.1
PGN (Cai et al., 2023)	36.1	23.9	35.7	23.9	11.3	4.6
PGN(Cov) (Cai et al., 2023)	36.6	24.2	37.5	27.6	13.7	5.4
RadSum (Cai et al., 2023)	37.9	25.5	39.2	28.0	13.9	5.5
TransAbs (Cai et al., 2023)	37.7	26.9	38.7	28.1	14.1	5.7
BART (Cai et al., 2023)	39.4	27.3	39.7	28.3	14.2	5.8
ChestXRayBERT (Cai et al., 2023)	41.3	28.6	41.5	28.5	14.4	6.1
BASE+AP+DCA (Hu et al., 2023)	47.6	32.0	46.1	–	–	–
Meta-Llama-3-8B (Zhao et al., 2024)	–	–	29.0	–	–	9.4
Meta-Llama-3-1B (Zero-shot)	0.2	0.1	0.2	0.8	0.3	0.1
Meta-Llama-3-1B (Few-shot)	30.5	12.9	27.9	31.2	55.2	67.4
CSTRL (Ours)	**58.1**	**48.5**	**56.5**	**65.0**	**47.9**	**38.9**

Table 2: Performance comparison across various models on the combined OPEN-I and MIMIC-CXR test sets.

Ablation Settings			R-1	R-2	R-L	B-1	B-2	B-3
GSG	Fisher Matrix	Layer Unfreezing						
✗	✗	✗	55.9	45.2	54.2	63.2	45.4	35.4
✓	✗	✗	55.9	45.2	54.2	63.2	45.4	35.4
✓	✓	✗	58.2	48.5	56.5	65.0	47.9	38.9
✓	✗	✓	53.4	43.1	51.9	61.5	42.3	32.3

Table 3: Ablation study of the effect of varying techniques in the proposed CSTRL architecture.

Model	R1	R-2	R-L	B-1	B-2	B-3
CSTRL-T	58.1	48.5	56.5	65.0	47.9	38.9
CSTRL-S (×8)	49.8	37.9	48.8	61.0	37.1	26.3
CSTRL-S (×16)	47.8	36.3	46.7	58.5	35.9	25.2
CSTRL-S (×32)	46.0	34.9	44.9	56.4	34.6	24.3

Table 4: Performance of reduced models on different distillation schemes.

4.3 Results and Discussion

We compare our proposed summarization model, CSTRL, with state-of-the-art neural network models trained on the same dataset. We also train a decoder-only Meta-Llama-3-1B model (similar in size to T5's 770M). The results in Table 2 show that our model outperforms others in all evaluation metrics, including ROUGE-x (R-x) and BLEU-x (B-x). To simulate low-resource summarization, we randomly select subsets of 10^k samples (where $k = 2, 3, 4, 5$), including 40,000 and 80,000 sam-

ples, to train our model, selecting the best validation checkpoint after convergence. As shown in Figure 5, fine-tuning with 40,000 examples (32.8% of the dataset) achieves summarization quality similar to full-data training. As shown in Table 2, our model outperforms ChestXRayBERT by 56.2% in BLEU-1, 40.5% in BLEU-2, and an extraordinary 84.3% in BLEU-3. The performance gain in CSTRL comes from two factors: T5 base model and an effective sequential learning strategy. We
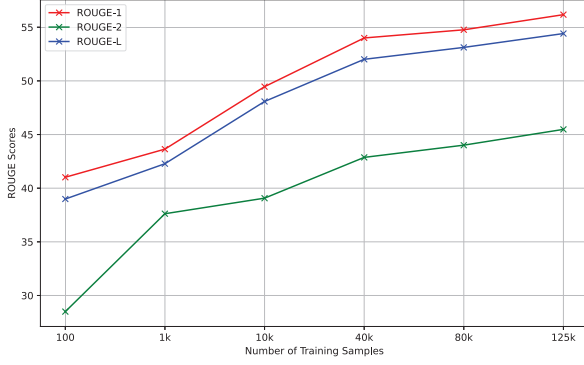
Figure 5: Fine-tuning with limited supervised examples. The solid lines are CSTRL fine-tuned on different numbers of datasets.

BS	LR	R-1	R-2	R-L	B-1	B-2	B-3
8	0.001	39.4	25.5	35.9	54.6	34.5	26.8
16	0.003	39.5	25.7	35.1	55.1	34.1	27.2
32	**0.003**	**39.7**	**26.2**	**36.3**	**55.7**	**35.8**	**28.1**

BS: Batch size, LR: Learning rate

Table 5: Performance of GSG task with Different Hyperparameters.

select T5 due to its exploration of transfer learning with a unified text-to-text transformer. This transformer has a bidirectional encoder and a unidirectional decoder, which distinguishes it from traditional Seq2Seq models. The bidirectional encoder enhances context understanding by analyzing the entire input sequence, enabling better context capture and management of long-range dependencies. We achieve improvements of 28.9%, 41.0%, and 26.5% in ROUGE-1, ROUGE-2, and ROUGE-L, respectively. Fine-tuning on radiology corpora has greatly improved the accuracy of the model. Using GSG with Fisher Matrix Regularization facilitates impactful sequential learning, transitioning from key sentence identification to summarization, as detailed in the ablation study.

4.4 Performance Evaluation and Ablation Study

Effect of GSG Fine-Tunning. As shown in Table 5, we predict masked sentences and evaluate using ROUGE and BLEU values for different batch sizes. In abstractive summarization, capturing inter-sentence relationships is crucial. We employ GSG, where sentences are masked based on ROUGE and BLUE, akin to an extractive summary. This method effectively accommodates both large and small datasets. Across the three batch sizes, the ROUGE and BLEU scores are quite similar, showing consistent performance across various configurations. The best configuration appears to be a batch size of 32 and a learning rate of 0.003, which provides a BLEU-1 of 55.7, a BLEU-2 of 35.8, and a BLEU-3 of 28.1.

Effect of Fisher Matrix Regularization with Penalty Adjustment. We note that the perfor-

mance metrics show identical values with and without the GSG framework, as shown in Table 3, for facing catastrophic forgetting. However, our experiments reveal differences beyond the third decimal point. Direct fine-tuning on the summarization task results in substantial updates to the model parameters, potentially compromising the knowledge gained during GSG training. After applying Fisher matrix regularization, we observe significant improvements in the results. These enhancements are due to the gradual adjustment of key parameters encapsulated within the Fisher matrix. As a result, ROUGE-1, ROUGE-2, ROUGE-L, BLUE-1, BLUE-2 and BLUE-3 scores exhibit a notable 4.1%, 7.3%, 4.2%, 2.8%, 5.5% and 9.9% improvements, respectively. We explore an alternative to Fisher matrix regularization by implementing a gradual layer unfreezing strategy. Initially, we freeze certain layers to retain knowledge from the previous fine-tuned task (GSG). Then, we slowly unfreeze these layers to minimize the risk of losing critical information. This approach reduces performance, decreasing ROUGE-1, ROUGE-2, ROUGE-L, BLEU-1, BLEU-2, and BLEU-3 scores by 4.5%, 4.6%, 4.2%, 2.7%, 6.8%, and 8.7%, respectively (Table 3).

Effect of the Distillation Operation on Computation. As shown in Table 4, we scale the network parameters to investigate the CSTRL Student model's (CSTRL-S) performance compared to the CSTRL Teacher model (CSTRL-T). Reducing the parameter size significantly drops the performance from 20.8% to 15.3%, 28.04% to 21.9%, 20.5% to 13.6%, 13.2% to 6%, 27.8% to 22.5%, and 37.5% to 32.39% in ROUGE-1, ROUGE-2, ROUGE-L, BLEU-1, BLEU-2, and BLEU-3, respectively. In Table 7, we measure complexity and computational efficiency. Inference time is reduced by up to 63.1% compared to CSTRL-T. We compare our model, CSTRL, with different settings against widely used baselines. We observe that CSTRL models (T and

Model	Readability	Factual Correctness	Informativeness	Redundancy	Completeness
CSTRL-T	4.62	4.39	4.31	4.11	4.75
CSTRL-S (×8)	4.59	4.29	4.28	4.73	4.54

Table 6: Human evaluation results based on summarization quality. Higher scores indicate better performance.

S variants) are highly efficient. They outperform existing models by reducing computational demands (GFLOPS, GMACs) while maintaining compact parameter sizes and competitive performance. The CSTRL-S variants (×8, ×16, ×32) exhibit a scalable efficiency trade-off: as the model scales down (8× to 32×), computational load decreases, achieving 2.5 to 8 times faster inference than ClinicalBERT or BART and 7 to 20 times faster than XrayGPT, despite using 10 to 100 times fewer parameters. In contrast, larger models like Meta-Llama-3-1B and XrayGPT demand significantly higher computational resources—XrayGPT requires 45 times more GMACs than CSTRL-S 32×—and have considerably slower inference (4s vs. 0.19s). We achieve near-real-time performance with minimal resources, bridging lightweight deployment and clinical-grade AI.

4.5 Factual Consistency

We analyze how sequential learning strategies affect the factual consistency of CSTRL using two key models. We evaluate the alignment of the generated summaries with the original documents in the test set. We confirm that the samples used as train data are not present in the validation part. As shown in Table 8, the model trained with the GSG Framework and Fisher matrix regularization produces the most consistent Impression of Findings. Sequential learning enhances this consistency. For SummaC-Conv, the variance between Impressions in the original report and those generated without sequential learning (W/O STRL) is 0.020, corresponding to 5.54%. After implementing sequential learning (STRL), this variance decreases to 0.014, representing 3.87%. This change indicates a performance enhancement of 1.67%. In the case of SummaC-ZS, the GSG implementation achieves a variance of 0.002 (3.13%). After adjusting for sequential learning, the score was 0.028, reflecting 6.25%. This results in an improvement of 3.12%. We deem reduced variance between the original and predicted impressions indicative of a loss of nuance. This technique eliminates contextual errors caused by subtle changes.

Model	GFLOPS	GMACs	Params	Inference Time
CSTRL-T	71.7	35.8	60.50M	0.521s
CSTRL-S (×8)	17.9	9.0	11.20M	0.330s
CSTRL-S (×16)	10.6	5.3	6.40M	0.220s
CSTRL-S (×32)	**5.6**	**2.8**	**3.63M**	**0.192s**
Meta-Llama-3-1B	1270	632.7	1.24B	1.527s
ClinicalBERT (Huang et al., 2020)	50	40	110M	1s
BART (Lewis et al., 2019)	100	200	140M	2s
XrayGPT (Thawakar et al., 2024)	250	500	350M	4s

Table 7: Comparison of models with original and distilled versions based on FLOPs, MACs, and Parameters.

Summary Type	SummaC-ZS	SummaC-Conv
Original	0.361	0.064
STRL	0.341	0.068
W/O STRL	0.375	0.066

Table 8: Performance comparison of factual consistency.

4.6 Human Evaluation

We conduct expert evaluations to correlate ROUGE and BLEU improvements with human judgments. We randomly selected 50 samples from the dataset. We compare summaries generated by our teacher model against its distilled students. Five volunteers participated, including three radiology researchers and two clinical radiologists. They rate the samples on a scale from 1 (very poor) to 5 (very good). Table 6 shows the close performance of both models. Upon inspection, we find that student summaries are rather telegraphic in terms of factual correctness.

5 Conclusion

This paper proposes CSTRL—Context-driven Sequential Transfer Learning—a novel fine-tuning framework for generating factually correct and coherent radiology report summaries. CSTRL follows a dual-stage process, where it is fine-tuned on both the GSG and summarization tasks. We show in the ablation study that it not only improves summarization quality, as measured by ROUGE and BLEU scores, but also enhances factual correctness. To the best of our knowledge, we first introduce a fusion of methodologies for generating accurate impressions and preserving context within the clinical domain.

Limitations

Our proposed model is evaluated on two datasets, both containing predominantly short sentences in the Impression section. A key limitation is that the model's predicted Impressions are similarly concise, likely due to the bias in the training data. As a result, the model may encounter difficulties in generating longer, more detailed impressions.

Acknowledgement

References

Xiaoyan Cai, Sen Liu, Junwei Han, Libin Yang, Zhenguo Liu, and Tianming Liu. 2023. ChestXRayBERT: A Pretrained Language Model for Chest Radiology Report Summarization. *IEEE Transactions on Multimedia*, 25:845–855.

Ernie Chang, Vera Demberg, and Alex Marin. 2021. Jointly Improving Language Understanding and Generation with Quality-Weighted Weak Supervision of Automatic Labeling. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 818–829, Online. Association for Computational Linguistics.

Zhengping Che, Sanjay Purushotham, Robinder Khemani, and Yan Liu. 2015. Distilling knowledge from deep networks with applications to healthcare domain. *arXiv preprint arXiv:1512.03542*.

Sumit Chopra, Michael Auli, and Alexander M. Rush. 2016. Abstractive Sentence Summarization with Attentive Recurrent Neural Networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 93–98, San Diego, California. Association for Computational Linguistics.

Hans Christian, Mikhael Pramodana Agus, and Derwin Suhartono. 2016. Single Document Automatic Text Summarization using Term Frequency-Inverse Document Frequency (TF-IDF). *ComTech: Computer, Mathematics and Engineering Applications*, 7(4):285–294.

Shlomi Codish and Richard N. Shiffman. 2005. A model of ambiguity and vagueness in clinical practice guideline recommendations. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2005:146–150.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Centers for Disease Control, Prevention, et al. 2016. Everyday words for public health communication. *Atlanta: CDC*.

Sharut Gupta, Praveer Singh, Ken Chang, Liangqiong Qu, Mehak Aggarwal, Nishanth Arun, Ashwin Vaswani, Shruti Raghavan, Vibha Agarwal, Mishka Gidwani, et al. 2021. Addressing catastrophic forgetting for medical domain expansion. *arXiv preprint arXiv:2103.13511*.

Michael P. Hartung, Ian C. Bickle, Frank Gaillard, and Jeffrey P. Kanne. 2020. How to Create a Great Radiology Report. *RadioGraphics*, 40(6):1658–1670.

Brian R. Herts and Lincoln L. Berland. 2021. Make Even Greater Radiology Reports. *RadioGraphics*, 41(3):E92–E93.

Jinpeng Hu, Zhihong Chen, Yang Liu, Xiang Wan, and Tsung-Hui Chang. 2023. Improving radiology summarization with radiograph and anatomy prompts. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12066–12080, Toronto, Canada. Association for Computational Linguistics.

Jinpeng Hu, Jianling Li, Zhihong Chen, Yaling Shen, Yan Song, Xiang Wan, and Tsung-Hui Chang. 2021. Word graph guided summarization for radiology findings. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4980–4990, Online. Association for Computational Linguistics.

Jinpeng Hu, Zhuo Li, Zhihong Chen, Zhen Li, Xiang Wan, and Tsung-Hui Chang. 2022. Graph enhanced contrastive learning for radiology findings summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4677–4688, Dublin, Ireland. Association for Computational Linguistics.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2020. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *Preprint*, arXiv:1904.05342.

Alshimaa M Ibrahim, Marco Alfonse, and M Aref. 2023. A systematic review on text summarization of medical research articles. *International Journal of Intelligent Computing and Information Sciences*, 23(2):50–61.

Alistair Johnson, Tom Pollard, Roger Mark, Seth Berkowitz, and Steven Horng. 2024. MIMIC-CXR Database.

Ashitha Kareem. 2024. Understanding Medical Reports and Results from Indian Hospitals.

Sanjeev Kumar Karn, Francine Chen, Yan-Ying Chen, Ulli Waltinger, and Hinrich Schütze. 2021. Few-Shot Learning of an Interleaved Text Summarization Model by Pretraining with Synthetic Data. In *Proceedings of the Second Workshop on Domain Adaptation for NLP*, pages 245–254, Kyiv, Ukraine. Association for Computational Linguistics.

Sanjeev Kumar Karn, Ning Liu, Hinrich Schuetze, and Oladimeji Farri. 2022. Differentiable multi-agent actor-critic for multi-step radiology report summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1542–1553, Dublin, Ireland. Association for Computational Linguistics.

Ronald Kemker, Marc McClure, Angelina Abitino, Tyler Hayes, and Christopher Kanan. 2018. Measuring catastrophic forgetting in neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Ramin Khorasani, David W Bates, Susan Teeger, Jeffrey M Rothschild, Douglas F Adams, and Steven E Seltzer. 2003. Is terminology used effectively to convey diagnostic certainty in radiology reports? *Journal of the American College of Radiology*, 1(7):457–460.

Virapat Kieuvongngam, Bowen Tan, and Yiming Niu. 2020. Automatic Text Summarization of COVID-19 Medical Research Articles using BERT and GPT-2.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526.

Alexey Kutalev. 2020. Natural way to overcome the catastrophic forgetting in neural networks. *arXiv preprint arXiv:2005.07107*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

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. *Preprint*, arXiv:1910.13461.

Christy Y. Li, Xiaodan Liang, Zhiting Hu, and Eric P. Xing. 2018. Hybrid retrieval-generation reinforced agent for medical image report generation. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pages 1537–1547, Red Hook, NY, USA. Curran Associates Inc.

Piji Li, Wai Lam, Lidong Bing, and Zihao Wang. 2017. Deep Recurrent Generative Decoder for Abstractive Text Summarization. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2091–2100, Copenhagen, Denmark. Association for Computational Linguistics.

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.

Xialei Liu, Marc Masana, Luis Herranz, Joost Van de Weijer, Antonio M Lopez, and Andrew D Bagdanov. 2018. Rotate your networks: Better weight consolidation and less catastrophic forgetting. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 2262–2268. IEEE.

Yang Liu, Sheng Shen, and Mirella Lapata. 2020. Noisy self-knowledge distillation for text summarization. *arXiv preprint arXiv:2009.07032*.

Chong Ma, Zihao Wu, Jiaqi Wang, Shaochen Xu, Yaonai Wei, Zhengliang Liu, Fang Zeng, Xi Jiang, Lei Guo, Xiaoyan Cai, Shu Zhang, Tuo Zhang, Dajiang Zhu, Dinggang Shen, Tianming Liu, and Xiang Li. 2024. An Iterative Optimizing Framework for Radiology Report Summarization With ChatGPT. *IEEE Transactions on Artificial Intelligence*, 5(8):4163–4175.

Sean MacAvaney, Sajad Sotudeh, Arman Cohan, Nazli Goharian, Ish Talati, and Ross W. Filice. 2019. Ontology-Aware Clinical Abstractive Summarization. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, pages 1013–1016, New York, NY, USA. Association for Computing Machinery.

Hefeng Meng, Zhiqiang Lin, Fan Yang, Yonghui Xu, and Lizhen Cui. 2022. Knowledge distillation in medical data mining: A survey. In *5th International Conference on Crowd Science and Engineering*, ICCSE '21, page 175–182, New York, NY, USA. Association for Computing Machinery.

Paul Michel, Elisabeth Salesky, and Graham Neubig. 2019. Regularizing trajectories to mitigate catastrophic forgetting.

Ramesh Nallapati, Bowen Zhou, Caglar Gulcehre, Bing Xiang, et al. 2016. Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

National Library of Medicine. 2024. Open-i: An open access biomedical search engine.

David M Panicek and Hedvig Hricak. 2016. How sure are you, doctor? a standardized lexicon to describe the radiologist's level of certainty. *AJR American Journal of Roentgenology*, 207(1):2–3.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1):140:5485–140:5551.

Hippolyt Ritter, Aleksandar Botev, and David Barber. 2018. Online structured laplace approximations for overcoming catastrophic forgetting. *Advances in Neural Information Processing Systems*, 31.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A Neural Attention Model for Abstractive Sentence Summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389, Lisbon, Portugal. Association for Computational Linguistics.

S Santhosh. 2023. Understanding bleu and rouge score for nlp evaluation. *Medium. URL: https://medium. com/@ sthanikamsanthosh1994/understanding-bleu-and-rouge-sc ore-for-nlp-evaluation-1ab334ecadcb (date of access: 15.05. 2024)*.

Sajad Sotudeh Gharebagh, Nazli Goharian, and Ross Filice. 2020. Attend to medical ontologies: Content selection for clinical abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1899–1905, Online. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pages 3104–3112, Cambridge, MA, USA. MIT Press.

Jiwei Tan, Xiaojun Wan, and Jianguo Xiao. 2017. Abstractive Document Summarization with a Graph-Based Attentional Neural Model. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1171–1181, Vancouver, Canada. Association for Computational Linguistics.

Omkar Chakradhar Thawakar, Abdelrahman M Shaker, Sahal Shaji Mullappilly, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, Jorma Laaksonen, and Fahad Khan. 2024. Xraygpt: Chest radiographs summarization using large medical vision-language models. In *Proceedings of the 23rd workshop on biomedical natural language processing*, pages 440–448.

Meng Wang, Qiang Wang, and Haipeng Liu. 2022. Clfm: Few-shot object detection via low-resource contrastive learning and fisher matrix updating for overcoming catastrophic forgetting. *IEEE Access*, 10:115307–115321.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. XLNet: generalized autoregressive pretraining for language understanding. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 517, pages 5753–5763. Curran Associates Inc., Red Hook, NY, USA.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020a. PEGASUS: pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *ICML'20*, pages 11328–11339. JMLR.org.

Yuhao Zhang, Daisy Yi Ding, Tianpei Qian, Christopher D. Manning, and Curtis P. Langlotz. 2018. Learning to Summarize Radiology Findings. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, pages 204–213, Brussels, Belgium. Association for Computational Linguistics.

Yuhao Zhang, Derek Merck, Emily Tsai, Christopher D. Manning, and Curtis Langlotz. 2020b. 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.

Xingmeng Zhao, Tongnian Wang, and Anthony Rios. 2024. Improving expert radiology report summarization by prompting large language models with a layperson summary. *Preprint*, arXiv:2406.14500.

Qingyu Zhou, Nan Yang, Furu Wei, and Ming Zhou. 2017. Selective Encoding for Abstractive Sentence Summarization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1095–1104, Vancouver, Canada. Association for Computational Linguistics.