e-Health CSIRO at RadSum23: Adapting a Chest X-Ray Report Generator to Multimodal Radiology Report Summarisation

Aaron Nicolson, Jason Dowling, and **Bevan Koopman**
aaron.nicolson@csiro.au, jason.dowling@csiro.au, and bevan.koopman@csiro.au
Australian e-Health Research Centre,
Commonwealth Scientific and Industrial Research Organisation,
Herston, Queensland, 4006, Australia

Abstract

We describe the participation of team e-Health CSIRO in the BioNLP RadSum task of 2023. This task aims to develop automatic summarisation methods for radiology. The subtask that we participated in was multimodal; the impression section of a report was to be summarised from a given findings section and set of Chest X-rays (CXRs) of a subject's study. For our method, we adapted an encoder-to-decoder model for CXR report generation to the subtask. e-Health CSIRO placed seventh amongst the participating teams with a RadGraph ER F1 score of 23.9.

1 Introduction

The impression section of a radiology report provides an overview of the key findings from an imaging study. It is meant to convey important information in a concise manner that can be interpreted by the referring clinician. The impression is often written after the findings section of a report, which may include more detailed descriptions, measurements, and technical terms. The impression section should be structured consistently across reports, allowing for a comparison of studies over time on the same subject or subjects with similar conditions. Automated report summarisation is gaining more interest as it could reduce the amount of clinical documentation that clinicians must undertake (Zhang et al., 2018). This could be especially important in emergency, where it is predicted that the automation of clinical documentation could reduce the burden placed on clinicians (Dinggang, 2021).

Automated report summarisation is the focus of RadSum23 (Task 1B of BioNLP 2023), a challenge that requires participants to summarise a given findings section via the impression section (Delbrouck et al., 2022b). RadSum23 consisted of two subtasks: the first focused on summarising the findings sections of MIMIC-III (Johnson et al., 2016), while the second was a multimodal summarisation task that required participants to summarise the findings section and Chest X-Rays (CXRs) of a subject's study. The multimodal summarisation task involved two datasets: MIMIC-CXR for model development and a hidden test set of CheXpert. Evaluation of the submissions was based on the factual correctness between the predicted and ground truth impression sections. Participants were also permitted to use external data sources during model development. Methods that could be considered for both subtasks include Large Language Models (LLMs) (Zhang et al., 2023), or retrieval-based methods (An et al., 2021).

This paper details the participation of the e-Health CSIRO team in RadSum23. In particular, we participate in the multimodal summarisation subtask, and not the MIMIC-III subtask. Our background is in CXR report generation, where the the findings and impression sections for a subject's study are generated given the CXRs for the study. Hence, our aim was to adapt a CXR report generator to the task of multimodal summarisation and to determine if it could generalise.

2 Methodology

2.1 Competition dataset

Two datasets were used for the challenge: MIMIC-CXR and CheXpert (Johnson et al., 2019; Irvin et al., 2019). The training, validation, and test splits were derived from MIMIC-CXR, with 125 417, 991, and 1 624 studies in each, respectively. Each study included one findings and impression section, as well as one or more CXRs. The hidden test set was derived from CheXpert. It included 1 000 studies, each with one or more CXRs and a findings section. No impression section was provided for the CheXpert hidden test set; each participant was required to produce the impression section of each study given the CXRs and findings section.

The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks, pages 545–549
July 13, 2023 ©2023 Association for Computational Linguistics

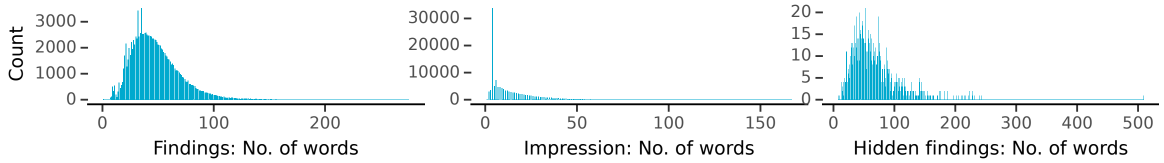

Figure 1: Number of words in the findings and impression sections of the MIMIC-CXR training set and the findings sections of the CheXpert hidden test set.

2.2 Pre-training task

CXR report generation with the MIMIC-CXR dataset was used as a pre-training task for some of the models (Johnson et al., 2019). Here, the task was to generate the findings and impression sections for a given set of CXRs from a subject's study. Sections from the ground-truth reports were obtained using the official text extraction tool.[1] Studies with either a missing findings or impression section, and studies with more than five CXRs per study were removed. This gave a training/validation/test split of 57 098/436/280 subjects, 125 395/991/1 624 studies, and 232 715/1 837/2 872 CXRs. Minimal formatting was applied to the ground-truth reports; newline characters, tab characters, and consecutive white spaces were replaced with a single whitespace character. The ground-truth reports were not truncated during training or testing.

2.3 Metrics

Several metrics were used for evaluation: BLEU-4 (B-4) (Papineni et al., 2002), ROUGE-L (R-L) (Lin and Och, 2004), BERTScore (B) (Zhang et al., 2020), the micro-averaged CheXbert F1 score (CX) (Smit et al., 2020), and the RadGraph ER F1 score (RG) (Delbrouck et al., 2022a). Only five of the 14 CheXbert classes were used: 'cardiomegaly', 'edema', 'consolidation', 'atelectasis', and 'pleural effusion'. Scoring was performed between the generated impression and the ground truth impression for each study. The rank of the participants was determined by their RG scores.

2.4 Training

Two stages of training were performed: Teacher Forcing (TF) (Williams and Zipser, 1989) and Self-Critical Sequence Training (SCST) (Rennie et al., 2017). Gradient descent optimisation was performed with AdamW (Loshchilov and Hutter, 2022) at an initial learning rate of 5e-5 for TF and 5e-6 for SCST, with a mini-batch size of 32. Early stopping was used for TF, with a patience of four, and the validation RG score as the monitored metric. For SCST, one epoch was completed, and validation was performed every $\frac{1}{10}$ of an epoch. RG was used as the reward. The validation RG score was the monitored metric for checkpoint selection. For SCST, the baseline was generated with greedy search, while the sample was produced with top-k sampling ($k = 50$). The maximum number of tokens for the findings and impression sections was 384 and 128, respectively. This was based on the findings section and the impression section accounting for 75.4% and 24.6% of the tokens on average in the training set reports, as shown in Figure 1.

2.5 Encoder-to-decoder model

An encoder-to-decoder model was used to generate the impression section. Here, the generation of a study's impression section is conditioned on the features of all CXRs of the study via the cross-attention of the decoder, as well as the findings section of the study via a prompt, as shown in Figure 2.

The Convolutional vision Transformer (CvT) was the encoder (specifically, CvT-21 pre-trained on ImageNet-22K and fine-tuned on ImageNet-1K at a resolution of 384 × 384) (Wu et al., 2021; Nicolson et al., 2022). Layer normalisation was applied to its last hidden state, followed by a projection to the decoder's hidden size. The encoded features for each CXR were concatenated and fed to the cross-attention of the decoder. Each CXR was resized using bilinear interpolation so that its smallest side had a length of 384 and its largest side maintained the aspect ratio. Next, the resized CXR was cropped to a size of $\mathbb{R}^{3\times384\times384}$. The crop location was random during training and centred during testing. During training, the CXR was rotated around its centre where the angle of rotation was sampled from $\mathcal{U}[-5°, 5°]$. Finally, the CXR was standardised using the mean and standard deviation provided with the CvT-21 checkpoint.

For the decoder, a byte-level byte pair encoding

[1] https://github.com/MIT-LCP/mimic-cxr/tree/master/txt

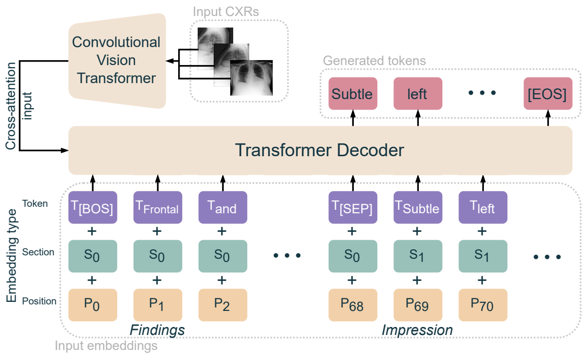

Figure 2: The encoder-to-decoder model tasked with generating the impression section. It is conditioned on the features of each CXR via the cross-attention and prompted by the findings section. `[BOS]`, `[SEP]`, and `[EOS]` denote the "beginning-of-sentence", "separator", and "end-of-sentence" special tokens, respectively.

tokeniser (Wang et al., 2020) was trained on all findings and impression sections of the training set of the pre-training dataset (before studies without a findings or impression section were dropped) and with a vocabulary size of 30 000. The $\text{BERT}_{\text{BASE}}$ architecture was employed as the decoder, as it is equipped with section embeddings. The decoder had 6 layers, was randomly initialised, included a language model head, and had a vocabulary size of 30 000. Greedy search and beam search with four beams were employed during validation and testing, respectively.

The different models for our submissions were based off the aforementioned encoder-to-decoder model by using different pre-training data, using different combinations of TF and SCST, and different generation configurations on the test sets:

1. **TF**: This is the encoder-to-decoder model fine-tuned on the competition training set using TF.
2. **MIMIC-CXR**: This is the encoder-to-decoder model fine-tuned with TF on the aforementioned pre-training task to generate the findings and impression sections. It is not fine-tuned on the competition training set.
3. **MIMIC-CXR SCST**: This is **MIMIC-CXR** fine-tuned on the competition training set with SCST (i.e., no TF was performed on the competition training set).
4. **MIMIC-CXR TF**: This is **MIMIC-CXR** fine-tuned on the competition training set with TF.
5. **MIMIC-CXR TF SCST**: This is **MIMIC-CXR FT** additionally fine-tuned on the competition training set with SCST.
6. **Length penalty 1**: The average number of words for the findings section of the competition MIMIC-CXR training set was 48, as compared to 68 for the CheXpert hidden test set. Hence, we made the assumption that the length of the impression sections for the CheXpert hidden test set were also longer on average than the impression sections of the competition MIMIC-CXR training set. Hence, we applied a length penalty α to the conditional probability of **MIMIC-CXR SCST** during generation to encourage impressions with more words to be generated, via $\log(P(I|F,R))/|Y|^{\alpha}$, where I and F denote the tokens for the impression and findings section, respectively, and R denotes the concatenated features of the CXRs. Here, we use $\alpha = 1.2$.
7. **Length penalty 2**: This is **Length penalty 1** with a more aggressive length penalty of $\alpha = 5$, as well as a penalty applied to the probability of tokens to prevent an n-gram from appearing more than once in a caption (the penalty was realised by setting a token's probability to zero). An n-gram size of three was used. This configuration was to encourage long impression sections without repetitions.

3 Results & Discussion

The results for each of the different models on the MIMIC-CXR test set and the CheXpert hidden test set are given in Tables 1 and 2, respectively. MIMIC-CXR TF produced better scores than TF and MIMIC-CXR on both test sets, showing that the pre-training task of CXR report generation, fol-

Table 1: Results for the different models on the MIMIC-CXR test set.

Model	B-4	R-L	B	CX	RG
TF	9.5	39.9	56.2	61.7	38.7
MIMIC-CXR	16.1	42.1	59.5	69.1	41.7
↳SCST	18.0	44.1	61.5	71.7	**45.0**
MIMIC-CXR TF	15.8	42.1	59.7	69.5	42.2
↳SCST	17.6	43.8	61.2	**73.1**	44.3
Length penalty 1	**18.1**	**44.2**	**61.5**	71.8	**45.0**
Length penalty 2	10.0	36.8	55.7	68.5	40.0

Table 2: Results for the different models on the CheXpert hidden test set.

Model	B-4	R-L	B	CX	RG
TF	3.5	18.8	40.1	53.1	21.8
MIMIC-CXR	0.9	16.5	35.6	35.7	16.5
↳SCST	1.5	18.5	39.5	46.7	20.3
MIMIC-CXR TF	2.8	19.2	40.9	50.8	20.3
↳SCST	**4.1**	**21.6**	**43.9**	**53.5**	**23.9**
Length penalty 1	1.5	18.5	39.5	47.0	20.3
Length penalty 2	2.9	15.5	36.9	49.0	17.6

Table 3: Best result for each participant on the MIMIC-CXR test set (ordered by RG).

Team	B-4	R-L	B	CX	RG
`utsa-nlp`	**25.9**	**47.9**	64.7	**77.9**	**51.8**
`dmis-msra`	25.6	47.8	**64.8**	76.3	51.0
`shs-te-dti-mai`	25.3	47.5	63.6	74.3	49.0
`knowlab`	23.0	46.2	63.4	75.1	48.0
`e-health csiro`	18.0	44.1	61.5	71.7	45.0
`iuteam1`	10.1	40.4	56.4	58.0	39.5
`nlpaueb`	11.7	36.8	55.5	59.5	36.9

Table 4: Best result for each participant on the CheXpert hidden test set (ordered by RG).

Team	B-4	R-L	B	CX	RG
`dmis-msra`	**18.6**	34.6	**55.9**	**72.4**	**43.2**
`utsa-nlp`	16.3	**35.0**	55.5	69.4	42.7
`knowlab`	14.4	33.6	54.7	67.2	40.0
`shs-te-dti-mai`	14.6	32.4	54.0	69.0	38.4
`aimi`	5.2	31.8	47.8	64.2	32.1
`iuteam1`	2.0	26.1	46.8	40.3	27.4
`e-health csiro`	4.1	21.6	43.9	53.5	23.9
`nlpaueb`	5.0	19.9	41.8	50.7	23.3

lowed by fine-tuning on the competition training set was best. SCST (with RG as the reward) improved the scores of MIMIC-CXR and MIMIC-CXR TF for all metrics on both test sets, especially on RG, showing the benefit of avoiding the *exposure bias* problem (Rennie et al., 2017). Length penalty 1 unexpectedly benefited the MIMIC-CXR test set, rather than the CheXpert hidden test set. In fact, Length penalty 1 offered minimal improvement (over MIMIC-CXR SCST) on the CheXpert hidden test set. The more extreme Length penalty 2 improved the scores for some metrics (B-4 and CX), while worsening the scores for others (R-L, B, and RG) on the CheXpert hidden test set. The model that performed best was MIMIC-CXR TF SCST, and was the model that was compared to the methods of the other participants.

The results of all participants on the MIMIC-CXR test set and the CheXpert hidden test set are given in Tables 3 and 4, respectively. Compared to the other participants, our method (e-Health CSIRO) could not produce impressions that were as factually correct as those of the other participants, placing fifth and seventh on the MIMIC-CXR test set and CheXpert hidden test set, respectively. For the MIMIC-CXR test set, we attained the fifth highest score for each metric, while on the CheXpert hidden test set, we attained the seventh highest score for each metric, except CX, where we attained the sixth highest score. Comparing our placement on the MIMIC-CXR test set to that on the CheXpert hidden test set, our method did not generalise well, losing a position to `iuteam1`.

4 Conclusion

For BioNLP RadSum 2023, the performance of our best submission placed the AEHRC CSIRO team seventh, with a RadGraph ER F1 score of 23.9. Even when optimising for the metric used to rank each participant with SCST, our encoder-to-decoder model developed for CXR report generation and adapted to the summarisation task could not produce impressions that were as factually correct as those produced by the other participants. We will be investigating more appropriate methods for multimodal radiology summarisation in the future, such as LLMs specifically trained for summarisation and information retrieval-based methods.

References

Chenxin An, Ming Zhong, Zhichao Geng, Jianqiang Yang, and Xipeng Qiu. 2021. RetrievalSum: A Retrieval Enhanced Framework for Abstractive Summarization. ArXiv:2109.07943 [cs.CL].

Jean-Benoit Delbrouck, Pierre Chambon, Christian Bluethgen, Emily Tsai, Omar Almusa, and Curtis Langlotz. 2022a. Improving the Factual Correctness of Radiology Report Generation with Semantic Rewards. In *EMNLP*, pages 4348–4360.

Jean-benoit Delbrouck, Khaled Saab, Maya Varma, Sabri Eyuboglu, Pierre Chambon, Jared Dunnmon, Juan Zambrano, Akshay Chaudhari, and Curtis Langlotz. 2022b. ViLMedic: A framework for research at the intersection of vision and language in medical AI. In *ACL: System Demonstrations*, pages 23–34.

Shen Dinggang. 2021. Grand Challenges in Radiology. *Frontiers in Radiology*, 1.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. 2019. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. In *AAAI*, volume 33, pages 590–597.

Alistair E. W. Johnson, Tom J. Pollard, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G. Mark, Seth J. Berkowitz, and Steven Horng. 2019. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. PhysioNet.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Liwei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1):160035.

Chin-Yew Lin and Franz Josef Och. 2004. Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statistics. In *ACL*, pages 605–612.

Ilya Loshchilov and Frank Hutter. 2022. Decoupled Weight Decay Regularization. In *ICLR*.

Aaron Nicolson, Jason Dowling, and Bevan Koopman. 2022. Improving Chest X-Ray Report Generation by Leveraging Warm-Starting. ArXiv:2201.09405 [cs.CV].

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A Method for Automatic Evaluation of Machine Translation. In *ACL*, pages 311–318.

Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-Critical Sequence Training for Image Captioning. In *CVPR*, pages 1179–1195.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Ng, and Matthew Lungren. 2020. Combining Automatic Labelers and Expert Annotations for Accurate Radiology Report Labeling Using BERT. In *EMNLP*, pages 1500–1519.

Changhan Wang, Kyunghyun Cho, and Jiatao Gu. 2020. Neural machine translation with byte-level subwords. In *AAAI*, pages 9154–9160.

Ronald J. Williams and David Zipser. 1989. A Learning Algorithm for Continually Running Fully Recurrent Neural Networks. *Neural Computation*, 1(2):270–280.

Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. 2021. CvT: Introducing Convolutions to Vision Transformers. In *ICCV*, pages 22–31.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *ICLR*.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2023. Benchmarking Large Language Models for News Summarization. ArXiv:2301.13848 [cs.CL].

Yuhao Zhang, Daisy Yi Ding, Tianpei Qian, Christopher D. Manning, and Curtis P. Langlotz. 2018. Learning to Summarize Radiology Findings. In *LOUHI*, pages 204–213.