e-Health CSIRO at RRG24: Entropy-Augmented Self-Critical Sequence Training for Radiology Report Generation

Aaron Nicolson, Jinghui Liu, Jason Dowling, Anthony Nguyen, & Bevan Koopman
Australian e-Health Research Centre, CSIRO Health and Biosecurity, Brisbane, Australia
aaron.nicolson@csiro.au

Abstract

The Shared Task on Large-Scale Radiology Report Generation (RRG24) aims to expedite the development of assistive systems for interpreting and reporting on chest X-ray (CXR) images. This task challenges participants to develop models that generate the *findings* and *impression* sections of radiology reports from CXRs from a patient's study, using five different datasets. This paper outlines the e-Health CSIRO team's approach, which achieved multiple first-place finishes in RRG24. The core novelty of our approach lies in the addition of entropy regularisation to self-critical sequence training, to maintain a higher entropy in the token distribution. This prevents overfitting to common phrases and ensures a broader exploration of the vocabulary during training, essential for handling the diversity of the radiology reports in the RRG24 datasets. Our model is available on Hugging Face (https://huggingface.co/aehrc/cxrmate-rrg24).

1 Introduction

Machine learning holds the potential to significantly enhance diagnostic processes and clinical reporting, particularly within the field of radiology — a discipline characterised by high volumes of imaging data. Radiologists are often tasked with interpreting and reporting on hundreds of imaging studies daily, a repetitive process that is susceptible to fatigue and error. Automated systems capable of generating radiology reports from chest X-rays (CXRs) could greatly alleviate this burden by ensuring consistency and potentially reducing diagnostic turnaround times.

The Shared Task on Large-Scale Radiology Report Generation (RRG24) challenges participants to develop automated systems for producing textual reports from CXR images, with a particular focus on the findings and impression sections (Xu et al., 2024; Delbrouck et al., 2022b). These sections are crucial as they convey the diagnostic interpretation and clinical significance of a patient's study. The challenge provides a means to benchmark the various models under uniform conditions, offering insights into which approaches are most effective for CXR report generation. Participants were to train and evaluate their submissions on a dataset formed from five different sources, including MIMIC-CXR (Johnson et al., 2019), CheXpert (Chambon et al., 2024), PadChest (Bustos et al., 2020), BIMCV COVID-19 (Vayá et al., 2020), and Open-i IU X-ray (Demner-Fushman et al., 2016). This dataset consisted of four subsets, including the *training*, *validation*, *public-test*, and *hidden-test*, where the radiology reports were available for all except the hidden-test set. Finally, RRG24 presents participants with unique challenges to overcome, such as handling studies with missing sections and deciding whether to use a single model or separate models for each section.

This paper outlines the approach taken by team e-Health CSIRO in the RRG24 challenge. For this, we developed a multimodal language model that conditions report generation not only on previously generated words (or subwords), but also on the image embeddings of all the CXRs of a patient's study. We utilised a single model to generate both sections and incorporated special tokens to signify the absence of a section during training. These special tokens were also used to guide the model to generate specific sections during testing.

A key factor to the performance of our submissions was our modification to the self-critical sequence training (SCST) reinforcement learning (RL) algorithm (Rennie et al., 2017). A widely-used technique to enhance RL is to add entropy regularisation into the objective function. This approach boosts exploration and prevents the model from prematurely settling on less optimal actions (Mnih et al., 2016). Hence, we add entropy regularisation to SCST, forming Entropy-Augmented Self-

Proceedings of the 23rd Workshop on Biomedical Language Processing, pages 99–104
August 16, 2024.

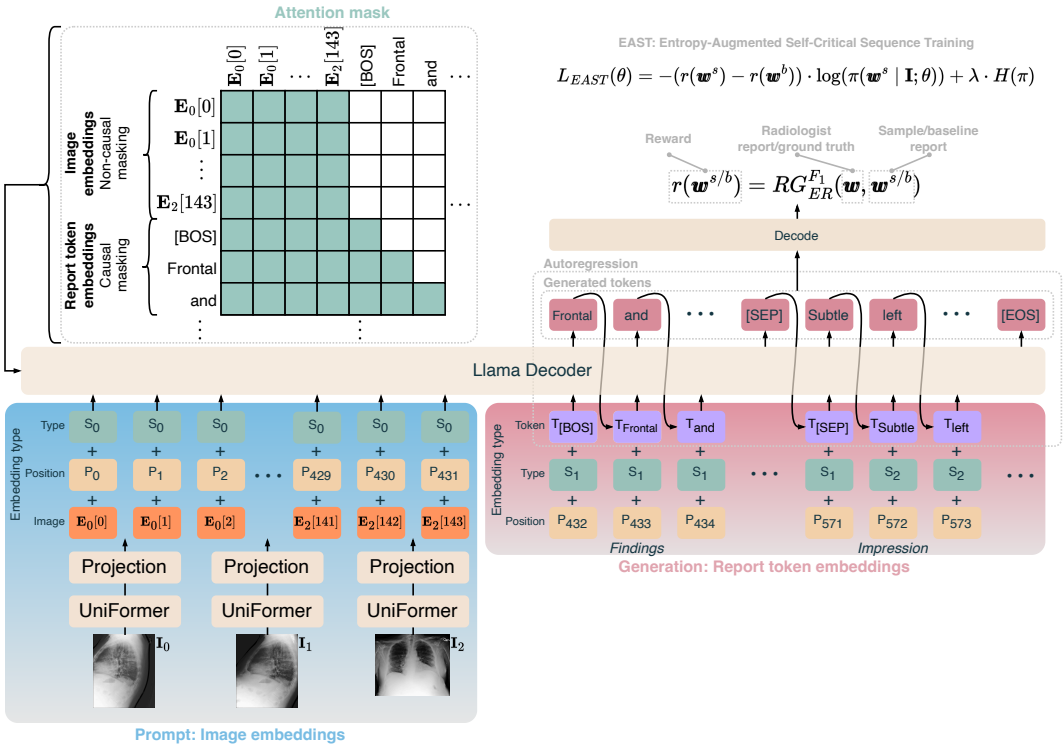

Figure 1: e-Health CSIRO's submission into RRG24, named CXRMate-RRG24. [BOS] denotes the *beginning-of-sentence* special token, [SEP] denotes the *separator* special token, and [EOS] denotes the *end-of-sentence* special token. $\mathbf{E}_k[i]$ is the i^{th} output of the projected last hidden state of the encoder for the k^{th} image of the study.

critical sequence Training (EAST). Using EAST, we optimised our model with RadGraph as the reward (Delbrouck et al., 2022a). RadGraph is the primary metric for RRG24; it evaluates the accuracy of a generated report by assessing how well the identified entities and their relationships align with those in a radiologist report. By optimising for this reward, we achieved multiple first-place finishes in RRG24.

2 Methodology

2.1 EAST: Entropy-Augmented Self-critical sequence Training

Entropy-Augmented Self-critical sequence Training (EAST) builds upon self-critical sequence training (SCST) by incorporating entropy regularisation. This encourages the model to maintain a higher entropy in its token distribution, thereby promoting diversity in token selection and preventing premature convergence on a smaller, selective set of tokens. The loss for SCST is as follows:

$$L_{SCST}(\theta) = -(r(\boldsymbol{w}^s) - r(\boldsymbol{w}^b)) \cdot \log(\pi(\boldsymbol{w}^s \mid \mathbf{I}; \theta)), \quad (1)$$

where $r(\boldsymbol{w}^s)$ is the reward for the sampled report ($\boldsymbol{w}^s = (w_1^s, ..., w_M^s)$ denotes the tokens of length M of the sampled report), $r(\boldsymbol{w}^b)$ is the reward for the baseline report ($\boldsymbol{w}^b = (w_1^b, ..., w_N^b)$ denotes the tokens of length N of the baseline report, where the baseline is generated with greedy search), $\mathbf{I} = [I_1, I_2, \ldots, I_K]$ denotes the images of a study (where K is the number of images in the study), θ represents the parameters of the model, and $\pi(\boldsymbol{w}^s \mid \mathbf{I}; \theta)$ denotes the policy under which $\boldsymbol{w}^s$ is sampled from. As illustrated in Figure 1, we utilise the RadGraph ER F1-score as the reward (Delbrouck et al., 2022a), where the generated report is either the sample or baseline report, both of which are compared to the radiologist report.

EAST is formed by adding an entropy term to $L_{SCST}(\theta)$:

$$L_{EAST}(\theta) = L_{SCST}(\theta) + \lambda \cdot H(\pi) \quad (2)$$

where λ is a coefficient that determines the weight of the entropy term in the loss function. The entropy is as follows:

$$H(\pi) = -\sum_{v \in \mathcal{V}} \pi(v \mid x; \theta) \log \pi(v \mid x; \theta), \quad (3)$$

Table 1: **Public test set** scores for the findings and impression sections (**presented as findings/impression**). The order of the leaderboard for RRG24 was determined by RadGraph-F1. The best scores are indicated in boldface.

Team/Method	BLEU-4	ROUGE-L	BERTScore	CheXbert-F1	RadGraph-F1
e-Health CSIRO					
EAST	12.00/**9.43**	26.51/26.58	54.64/47.81	59.18/**57.73**	29.46/**27.01**
SCST	10.70/8.51	26.54/26.30	54.79/48.25	56.42/55.00	27.66/25.04
TF	11.63/7.52	25.92/23.34	51.34/41.46	50.73/47.27	23.12/20.08
Top three teams besides ours					
tartan	**21.59**/-	**42.03**/-	**64.34**/-	**59.70**/-	**38.05**/-
maira	12.26/8.68	28.00/**28.40**	55.76/**50.48**	59.71/56.46	26.33/25.89
airi	10.13/7.10	26.54/25.92	53.84/47.18	55.49/51.33	25.82/24.07

where x represents the current state (as determined by the image embeddings and the previously generated tokens) and v represents a token from the vocabulary $\mathcal{V}$. This discourages the policy from converging too quickly to deterministic actions, thus encouraging the exploration of a wider set of generated reports.

2.2 Special Tokens and Missing Sections

As illustrated in Figure 1, our model generates both sections. To delineate these sections within the generated text, we utilise a separator token, following CXRMate (Nicolson et al., 2024a).[1] To accommodate reports during training that have a missing section, we employ two special tokens: `[NF]` for 'no findings' section and `[NI]` for 'no impression' section. They are used in place of the missing sections. They also facilitate the generation of specific sections as needed. For example, if only the impression section is to be generated, `[BOS][NF][SEP]` can be fed to the decoder to signal that the findings section is not to be generated. Furthermore, to encourage the generation of the impression section, the probability of the `[NI]` token can be set to zero.

2.3 Model

Our model, CXRMate-RRG24, is an evolution of our previous model, CXRMate, and is illustrated in Figure 1. We utilised UniFormer as the encoder (in particular, the 384×384 base model warm started with its token labelling fine-tuned checkpoint) (Li et al., 2023), which, in preliminary testing, performed comparably to the convolutional vision Transformer (CvT) (which we found to be the best performing encoder for CXR report generation in our previous work (Nicolson et al., 2023)) but significantly reduced the training time. The image embedding prompt is formed by processing each image in the study separately with the encoder and then projecting the encoder's last hidden state to match the decoder's hidden size using a learnable weight matrix. Each image 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 image was cropped to a size of $\mathbb{R}^{3\times384\times384}$. The crop location was random during training and centred during testing. Following (Elgendi et al., 2021), the image was rotated around its centre during training, where the angle of rotation was sampled from $\mathcal{U}[-5°, 5°]$. Finally, the image was standardised using the statistics provided with the UniFormer checkpoint. A maximum of five images per study were used during training. If more were available, five were randomly sampled uniformly without replacement from the study.

For the decoder, we employed the Llama architecture, which is notable for features such as its rotary positional encoding (RoPE), root mean square normalisation (RMSNorm), and SwiGLU activation function (Touvron et al., 2023). The decoder was initialised randomly and used the CXRMate vocabulary, which was derived from the MIMIC-CXR training set. The hyperparameters of the Llama decoder mirror that of the CXRMate decoder, with six hidden layers, a hidden size of 768, 12 attention heads per layer, and an intermediate size of 3 072. Following CXRMate, we added source type embeddings to the input of the decoder to differentiate between findings and impression section tokens, as well as image embeddings. The max number of position embeddings was set to 2048 to accommodate both the image embeddings and the report token embeddings. The maximum number of tokens that could be generated was set to 512, which was also the limit for the radiologist reports during training. During testing, a beam size of four was utilised. Another factor that led to the use of the Llama decoder was the ease of providing a cus-

[1] https://huggingface.co/aehrc/cxrmate

Table 2: **Hidden test set** scores for the findings and impression sections (**presented as findings/impression**). The order of the leaderboard for RRG24 was determined by RadGraph-F1. The best scores are indicated in boldface.

Team/Method	BLEU-4	ROUGE-L	BERTScore	CheXbert-F1	RadGraph-F1
e-Health CSIRO					
EAST	**11.68/12.33**	26.16/28.32	53.80/50.94	57.49/**56.97**	**28.67/27.83**
SCST	10.25/10.95	26.10/27.34	53.88/50.07	55.78/54.79	27.29/24.97
TF	11.12/9.89	25.43/24.94	51.10/42.49	50.02/47.24	22.99/21.27
Top three teams besides ours					
maira	11.24/11.66	**26.58/28.48**	**54.22/51.62**	**57.87**/53.27	25.48/25.26
airi	9.97/10.91	25.82/27.46	52.42/49.55	54.25/52.32	25.29/24.67
gla-ai4biomedic	7.65/9.60	24.35/25.27	52.69/48.60	46.21/46.74	24.13/22.10

tom attention mask to current implementations.[2] This enabled non-causal masking to be utilised for the prompt and causal masking for the report token embeddings, as shown in Figure 1. This ensured that the self-attention heads were able to attend to all of the image embeddings at each position.

2.4 Training

Two stages of training were performed; teacher forcing (TF) (Williams and Zipser, 1989), followed by RL (either EAST or SCST). *AdamW* (Loshchilov and Hutter, 2022) was used for mini-batch gradient descent optimisation with an initial learning rate of 5e-5 for TF and 5e-6 for RL, a mini-batch size of 16 for TF and 8 for RL, a maximum of 32 epochs for TF and 1 epoch for RL, executed on a 94GB NVIDIA H100 GPU with FP32. For RL, validation was performed every $\frac{1}{50}$ of an epoch. The validation macro-averaged CheXbert F1 was the monitored metric for checkpoint selection. For RL, the sample report was generated with top-k sampling ($k = 50$). During RL, the encoder was frozen. For EAST, the entropy weight (λ) was set to 0.05.

3 Results and Discussion

The results for our key submissions on the public and hidden test sets are shown in Tables 1 and 2, respectively. The metrics utilised for RRG24 include BLEU-4 (Papineni et al., 2001), ROUGE-L (Lin and Och, 2004), BERTScore (Zhang et al., 2020), CheXbert-F1 (Smit et al., 2020), and RadGraph-F1 (Delbrouck et al., 2022a), the later of which is the primary metric used to rank the teams. Here, we compare TF, to SCST, and to our proposed method, EAST. EAST attained a higher score than TF for each metric, something SCST was not able to do (TF attained a higher BLEU-4 score than SCST for the findings section of both test datasets).

Comparing EAST to SCST, SCST attained a higher ROUGE-L score on the public-test findings sections, and a higher BERTScore on the public-test findings and impression sections, as well as the hidden-test findings sections. For all other cases, EAST demonstrated an improvement over SCST. Policies trained with entropy regularisation often have improved generalisation, as they have learnt to consider a broader set of possible actions. This may have led EAST to be more robust to the differing characteristics of each of the datasets used in the public and hidden test sets. With EAST, team e-Health CSIRO achieved a first-place finish amongst participants for the public-test impression sections and the hidden-test findings and impression sections. We also also achieved a second-place finish for the public-test findings sections. For a comparison of CXRMate-RRG24 to state-of-the-art methods in the literature, please see Nicolson et al. (2024b).

3.1 Conclusion

Our proposed approach, EAST, was able to generate reports that were quantitatively more aligned with radiologist reports than those generated using SCST. By incorporating entropy regularisation, EAST is able to maintain a higher diversity in token selection and mitigate overfitting to maintain generalisability. This was likely crucial in handling the varied characteristics of the datasets used in RRG24. While EAST shows promise, a more thorough investigation is required to validate its potential, including the impact of varying the entropy coefficient.

References

Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria De La Iglesia-Vaya. 2020. Padchest: A

[2] https://huggingface.co/blog/poedator/4d-masks

large chest x-ray image dataset with multi-label annotated reports. *Medical image analysis*, 66:101797.

Pierre Chambon, Jean-Benoit Delbrouck, Thomas Sounack, Shih-Cheng Huang, Zhihong Chen, Maya Varma, Steven QH Truong, Chu The Chuong, and Curtis P. Langlotz. 2024. CheXpert Plus: Augmenting a Large Chest X-ray Dataset with Text Radiology Reports, Patient Demographics and Additional Image Formats.

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 *Findings of the Association for Computational Linguistics: 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.

Dina Demner-Fushman, Marc D. Kohli, Marc B. Rosenman, Sonya E. Shooshan, Laritza Rodriguez, Sameer Antani, George R. Thoma, and Clement J. McDonald. 2016. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310.

Mohamed Elgendi, Muhammad Umer Nasir, Qunfeng Tang, David Smith, John-Paul Grenier, Catherine Batte, Bradley Spieler, William Donald Leslie, Carlo Menon, Richard Ribbon Fletcher, Newton Howard, Rabab Ward, William Parker, and Savvas Nicolaou. 2021. The Effectiveness of Image Augmentation in Deep Learning Networks for Detecting COVID-19: A Geometric Transformation Perspective. *Frontiers in Medicine*, 8.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. 2019. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317.

Kunchang Li, Yali Wang, Junhao Zhang, Peng Gao, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. 2023. UniFormer: Unifying Convolution and Self-Attention for Visual Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12581–12600.

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*.

Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. 2016. Asynchronous Methods for Deep Reinforcement Learning. In *ICLR*, pages 1928–1937.

Aaron Nicolson, Jason Dowling, and Bevan Koopman. 2023. Improving chest X-ray report generation by leveraging warm starting. *Artificial Intelligence in Medicine*, 144:102633.

Aaron Nicolson, Jason Dowling, and Bevan Koopman. 2024a. Longitudinal Data and a Semantic Similarity Reward for Chest X-Ray Report Generation. arXiv:2307.09758 [cs].

Aaron Nicolson, Shengyao Zhuang, Jason Dowling, and Bevan Koopman. 2024b. The Impact of Auxiliary Patient Data on Automated Chest X-Ray Report Generation and How to Incorporate It. ArXiv:2406.13181 [cs].

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. BLEU: a method for automatic evaluation of machine translation. In *ACL*, page 311.

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.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971 [cs].

Maria De La Iglesia Vayá, Jose Manuel Saborit, Joaquim Angel Montell, Antonio Pertusa, Aurelia Bustos, Miguel Cazorla, Joaquin Galant, Xavier Barber, Domingo Orozco-Beltrán, Francisco García-García, et al. 2020. BIMCV COVID-19+: a large annotated dataset of RX and CT images from COVID-19 patients. *arXiv:2006.01174 [eess.IV]*.

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

Justin Xu, Zhihong Chen, Andrew Johnston, Louis Blankemeier, Maya Varma, Jason Hom, William J. Collins, Ankit Modi, Robert Lloyd, Benjamin Hopkins, Curtis Langlotz, and Jean-Benoit Delbrouck. 2024. Overview of the First Shared Task on Clinical Text Generation: RRG24 and "Discharge Me!". In *The 23rd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, Bangkok, Thailand. Association for Computational Linguistics.

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