Token Imbalance Adaptation for Radiology Report Generation

Yuexin Wu YWU10@MEMPHIS.EDU
University of Memphis

I-Chan Huang I-CHAN.HUANG@STJUDE.ORG
St. Jude Children's Research Hospital

Xiaolei Huang XIAOLEI.HUANG@MEMPHIS.EDU
University of Memphis

Abstract

Imbalanced token distributions naturally exist in text documents, leading neural language models to overfit on frequent tokens. The token imbalance may dampen the robustness of radiology report generators, as complex medical terms appear less frequently but reflect more medical information. In this study, we demonstrate how current state-of-the-art models fail to generate infrequent tokens on two standard benchmark datasets (IU X-RAY and MIMIC-CXR) of radiology report generation. To solve the challenge, we propose the **T**oken **Im**balance Adapt**er** (*TIMER*), aiming to improve generation robustness on infrequent tokens. The model automatically leverages token imbalance by an unlikelihood loss and dynamically optimizes generation processes to augment infrequent tokens. We compare our approach with multiple state-of-the-art methods on the two benchmarks. Experiments demonstrate the effectiveness of our approach in enhancing model robustness overall and infrequent tokens. Our ablation analysis shows that our reinforcement learning method has a major effect in adapting token imbalance for radiology report generation.

Data and Code Availability In this study, we conduct our experiments in two public datasets, IU X-RAY (Demner-Fushman et al., 2015) and MIMIC-CXR (Johnson et al., 2019). The datasets access to download from https://openi.nlm.nih.gov/ and https://physionet.org/content/mimic-cxr/2.0.0/. We publish our code at https://github.com/woqingdoua/TIMERf.

Institutional Review Board (IRB) In this study, we use the publicly available datasets after taking required training courses and signing data usage agreements. All the publicly available datasets have been de-identified and anonymized. Our study focuses on computational approaches and does not collect data from human subjects. We applied the institutional IRB determination that an IRB approval is not required for this study.

1. Introduction

Radiology report generation is to automatically generate a precise description in natural language given medical images, including computed tomography (CT) and X-RAY images. Increasing studies have deployed deep encoder-decoder neural architectures to encode medical images and decode the information to generate radiological reports (Jing et al., 2018, 2019; Chen et al., 2020, 2021; Qin and Song, 2022). Overfitting on frequent tokens is a common challenge in the text generation field that generators fail to predict infrequent tokens (Yu et al., 2022). Our empirical analysis has demonstrated that over 80% medical terms are infrequent tokens, while frequent tokens can count over 82% corpus (Section 2). The complex and lengthy tokens naturally occur less frequently than simple words (Nikkarinen et al., 2021), which is common across medical scenarios (Demner-Fushman et al., 2015; Johnson et al., 2019). However, existing studies have not explicitly consider infrequent tokens for radiology report generation, which can decrease model robustness and lead to imprecise reports.

Methods to reduce token-frequency-overfit commonly deploy post-processing (Mu and Viswanath, 2018) or regularization techniques (Welleck et al., 2020; Wang et al., 2020; Yu et al., 2022) for token embeddings. However, the approaches usually work for text-to-text generation given text sentences as inputs, while medical images are inputs in our study.

Nishino et al. proposes reinforcement learning approach for the class imbalance issue. However, the approach aims to solve image category imbalance instead of token imbalance. Modeling the token imbalance for the multimodal generation task is an unsolved challenge, especially for medical scenarios.

In this study, we propose the Token Imbalance Adapter (TIMER) model using reinforcement learning (Sutton et al., 1999) to adapt infrequent token generation and evaluate on radiology report generation by two publicly available datasets, IU X-RAY (Demner-Fushman et al., 2015) and MIMIC-CXR (Johnson et al., 2019). Our approach deploys an unlikelihood loss to penalize incorrect predictions for frequent tokens and develops a dynamic adaptation module to adjust the optimization process automatically. We compare the TIMER with three state-of-the-art baselines on overall performance and show overall improvements of our approach. By evaluating the performance of low and high-frequent tokens, our approach can significantly improve generation performance on infrequent token sets and maintain stable performance on frequent tokens.

Our major contributions are summarized as follows: 1) to our best knowledge, this is the first study that explicitly adapts token imbalance for radiology report generation. We have demonstrated the importance of infrequent tokens in radiology datasets, as most medical terms occur infrequently (Figure 1). 2) We propose a reinforcement learning method that effectively improves model overall performance as well as performance of infrequent tokens. 3) We conduct extensive ablation analysis to illustrate the effectiveness of adapting infrequent tokens. The ablation analysis proves that our method successfully reduces generation biases on frequent tokens and dynamically leverage token frequency.

2. Data

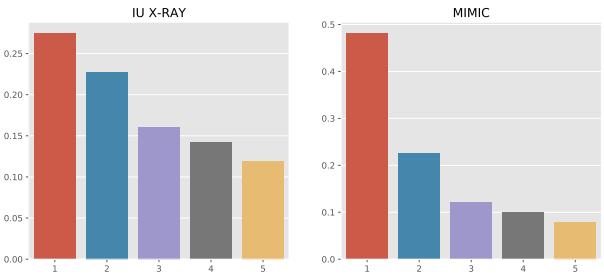
We retrieved two publicly available datasets of radiology report generation, IU X-RAY (Demner-Fushman et al., 2015) and MIMIC-CXR (Johnson et al., 2019). 1) *IU X-RAY*, collected by Indiana University, provides 3,955 reports and 7,470 X-RAY images from Indiana Network for Patient Care. 2) *MIMIC-CXR* (denote as MIMIC) is the largest radiography and publically available dataset to date. The dataset contains 227,835 reports with 377,110 images collecting between 2011 and 2016 at the Beth Israel Deaconess Medical Center. Each radiology report may associate with one or more front and side X-rays images. We summarize the data statistics in Table 1.

Table 1: Statistics of the two radiography datasets. Length is the report's average length, and Vocab is the vocabulary size.

	Images	Reports	Length	Vocab
IU X-RAY	7,470	3,955	35.99	1,517
MIMIC	377,110	227,835	59.70	13,876

Overfitting on frequent patterns is a common challenge for deep neural models — radiology report generators can easily overfit on frequent tokens than infrequent tokens. A recent text generation study (Yu et al., 2022) has found that text generators commonly perform much worse on infrequent tokens. The low performance on infrequent tokens can significantly impact radiological report generation, where many important medical terms do not appear frequently. In this study, we argue that *infrequent tokens matter* in radiology report generation.

Figure 1: Ratios of medical terms across five equal splits of vocabulary. 1 represents the most infrequent token set, and 5 refers to the most frequent token split.



To verify our claim, we conduct a quantitative analysis of medical terms across token distributions. First, we deploy a named entity recognition (NER) model (en_ner_bc5cdr_md) from `Sci-SpaCy` to extract disease-related terms.[1] Next, we split the vocabulary of each dataset into five equal segments. Our empirical counts show that the first 20% (segment) of vocabulary types account for over 82% tokens in each

1. Package details refer to the link: https://allenai.github.io/scispacy/. We choose the NER model because it achieves the best performance on the medical data and covers larger medical terms than the other available models.

dataset while the rest of vocabulary types (80%) only account for less 18% of corpus tokens. We calculate frequency ratios of medical terms in each segment and visualize the results in Figure 1. The figure shows that infrequent tokens contain more medical terms than frequent tokens, especially for complex tokens (average character length per token).[2]

Ignoring the low performance on infrequent tokens can significantly harm robustness of radiological generators (i.e., baselines fail in infrequent token evaluations in Table 3). This is essentially important for medical scenarios where infrequent tokens are more likely as complex domain terms, which are not commonly as frequent words (Nikkarinen et al., 2021). However, there is few study in the text generation task adapting infrequent tokens, especially for the radiological report generation. Thus, the issue inspires us to propose our model, **T**oken **Im**balance Adapt**er** (*TIMER*).

3. Token Imbalance Adapter

In this section, we present our approach *TIMER* in Figure 2. TIMER consists of three major modules, 1) unlikelihood loss, 2) dynamic adaptation, and 3) joint optimization. We deploy unlikelihood loss to reduce overfit and token imbalance of the text generator by penalizing a frequent-token set. The dynamic adaptation module deploys reinforcement learning to allow adjust the frequent-token set automatically. Finally, We elaborate on how to jointly optimize the text generator and dynamic adaptation module.

3.1. Problem Statement

Radiology report generation is an image-to-text generation task. Given a radiology image $\mathbf{x}$ and the corresponding report $\boldsymbol{y} = (y_0, \ldots, y_L)$ with length L, the task aims to train a text generator $(p_\theta(\mathbf{y} \mid \mathbf{x}))$ by minimizing

$$\mathcal{L}_{NLG}(\theta) = -\sum_{l=1}^{L} \log p\left(y_l \mid y_1, \ldots, y_{l-1}, \mathbf{x}; \theta\right) \quad (1)$$

, where $p\left(y_l \mid y_1, \ldots, y_{l-1}, \mathbf{x}; \theta\right)$ is from the softmax function for the next token prediction, and θ refers to generation model parameters. Token imbalance can cause prediction overfit that leads to low performance on low-frequent tokens. To solve the challenge, we balance the performance by the unlikelihood loss.

3.2. Unlikelihood Loss

Inspired by the work (Welleck et al., 2020), we unitize unlikelihood loss to reduce over-fitting effects by penalizing predicted probabilities for frequent tokens. Firstly, we calculate the average predicted possibility $p(u)$ for each token u in a report,

$$p(u) = \frac{\sum_{l=1}^{L} \log\left(p\left(u \mid y_1, ..., y_l\right)\right)}{L} \quad (2)$$

Then, given a set of frequent tokens $\mathcal{U}_h$ in a corpus, the unlikelihood loss punishes each token $u \in \mathcal{U}_h$ by

$$\mathcal{L}_{\mathrm{UL}}(\mathcal{U}_h) = -\sum_{u \in \mathcal{U}_h} \log(1 - p\left(u\right)) \quad (3)$$

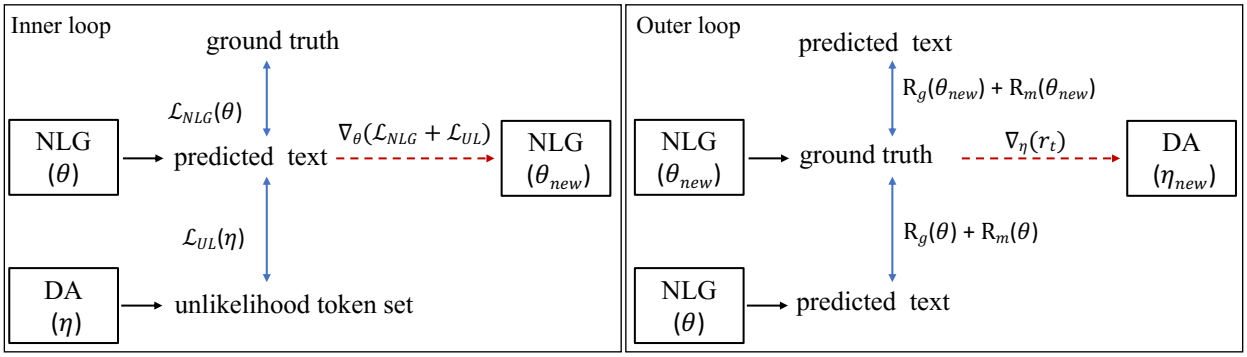
To decrease $\mathcal{L}_{\mathrm{UL}}$, the model predicts lower probabilities $p(u)$ on the frequent tokens, u. However, the unlikelihood loss has two issues: frequent token set and combining optimization objectives. Because statically fixing the frequent token set and equally combining two optimization objectives ($\mathcal{L}_{NLG}$ and $\mathcal{L}_{\mathrm{UL}}$) are not ideal for the report generation task. For example, Section 5.4 shows a static set of frequent tokens is not effective to reduce the token-frequency overfit of report generators. To enable dynamic adaptation, we deploy reinforcement learning to select a frequent token set. We leverage different training objectives by the joint optimization (Section 3.4).

3.3. Dynamic Adaptation

We deploy a reinforcement learning (RL) (Sutton et al., 1999) method allowing for a dynamic unlikelihood token set instead of a fixed set. The dynamic adaptation includes three major components, a policy network, a value network, and a reward. The dynamic adaptation paradigm is to adjust the unlikelihood token set and reduce token frequency effects during the report generation. Specifically, the NLG model is jointly optimized by the generation and unlikelihood losses, where the DA module decides the unlikelihood set. Then our imbalance reward can evaluate the improvements between the trained and untrained model with the unlikelihood loss. This reward helps the DA module to dynamically tune a better unlikelihood set, which can guide the NLG model to better training by balancing performance between

2. In IU X-RAY, the top medical terms in frequent tokens are pleural, effusion, heart, lung, and focal, while the top ranks in the infrequent tokens are diverticula, shrapnel, hemorrhage, prosthesis, arthroplasty.

Figure 2: Illustration of our proposed TIMER model. We use arrows to indicate model workflow. Blue arrows refer to loss value calculations by Equations 1 and 3. Parameter update processes follow red dotted lines. Our learning progress has inner and outer loops. In the inner loop, we update the NLG model by $\mathcal{L}_{NLG} + \mathcal{L}_{UL}$. In the outer loop, we implement the dynamic adaptation (DA) via reinforcement learning and update DA module (η) by the updated parameters (θ_{new}) and the reward r_t in Eq. 4.



infrequent and frequent tokens. Such a setting increases flexibility and dynamics of adjusting balances between frequent and infrequent tokens.

Policy Network, $\pi_\theta(a_t \mid s_t)$, aims to predict a frequent token set as an action a_t from a state s_t at step t. We take a dot product of token distributions between prediction in Eq. 2 and ground truth in a sample as our state. The dot product can reflect total token-level prediction errors. The policy network feeds a state (s_t) into the fully connected layer and sigmoid function and outputs a possibility estimation for an action a_t. The fully connected layer has the same input and output sizes as the vocabulary size. Finally, the policy network adopts a Bernoulli sampling on the estimated possibility vector and obtains the final unlikelihood token set $\mathcal{U}$. The policy network identifies the frequent token set for each sample through learning stages. This setting can help the generation model dynamically adjust the tokens' weight to reduce overfitting during optimization.

Value Network, A2C algorithm (Sutton et al., 1999)utilizes a value network to guide the policy network's learning and reduce variance due to the environment's randomness. In our case, the NLG model may overfit to different tokens after each optimization step, which causes the same policy can have a different reward. To overcome such randomness, we also introduce a value network to help the policy network's stable learning. $Q(s_t, a_t)$, is to predict a reward by given a state (s_t) and a action (a_t). In our case, $Q(s, a)$ is a 1-D convolutional network with two input channels and one output channel. The kernel size was set to 10 in our experiment. The predicted reward is the average of the convolutional network output.

Reward We design a reward to combine generation report quality and leverage the performance variations between frequent and infrequent tokens. Our reward consists of a text generation reward (R_g) from Eq.1 and and imbalanced evaluation reward (R_m). The imbalanced evaluation divides the comparison count of performance variations into four sets according to the tokens' frequency and calculates the F1 score for each token's set, respectively. We denote $F(\mathcal{U}_l)$, $F(\mathcal{U}_m)$ and $F(\mathcal{U}_h)$ as the F1 score of the low, medium and high-frequency token sets, respectively. To avoid F1-score overestimation due to token repetitions, we restrict each token in the prediction only project into one token in the ground truth instead of the token repetition. For example, if the prediction text includes repetitive toke such as "bone is is intact" and the ground truth is "the heart size is abnormal", the correct prediction token number is 1 in this case according to our definition. Because the correct prediction token ("is") only occurs one time in the ground truth. Then, we average the F1 score of $F(\mathcal{U}_l)$, $F(\mathcal{U}_m)$, and $F(\mathcal{U}_h)$ as our imbalanced reward

R_m:

$$R_m = (|F(\mathcal{U}_h) - F(\mathcal{U}_l)| + |F(\mathcal{U}_m) - F(\mathcal{U}_l)| + |F(\mathcal{U}_h) - F(\mathcal{U}_m)|)/3$$

Compared to calculating the F1 score in one token set, this method can alleviate a biased evaluation due to imbalanced token distribution since R_m can balance performance across different frequency token sets. We can formulate the final reward as follows:

$$r_t = R_g(\theta_{new}) + R_m(\theta_{new}) - R_g(\theta) - R_m(\theta) \quad (4)$$

where R_g is text generation loss in Eq.1, and θ refers to model parameters. We include optimization steps in the following section to learn the policy network and value network with the reward.

3.4. Joint optimization

Our optimization includes *inner* and *outer* loops. In the inner loop, we update the natural language generation (NLG) model with parameters θ according to the loss $\mathcal{L}_{NLG}$ in Eq. 1) and the unlikelihood loss $\mathcal{L}_{\text{UL}}$ in Eq. 3 as follows,

$$\mathcal{L}_{inner} = \mathcal{L}_{NLG}(\theta) + \mathcal{L}_{\text{UL}}(\mathcal{U}; \eta) \quad (5)$$

$$\theta_{new} = \theta - \nabla_\theta \mathcal{L}_{inner} \quad (6)$$

In the outer loop, we update the dynamic adaptation module (DA) with parameters η by A2C algorithm (Sutton et al., 1999). The policy network learns how to interact with the environment within time t by minimizing the following expectation:

$$\mathcal{L}_{policy} = -(\mathbb{E}[\sum_t \log \pi_\theta(a_t \mid s_t) A(s_t, a_t)]),$$

where $A(s_t, a_t)$ is an advantage estimate and equals to the real advantage in expectation. A2C utilizes Temporal-Difference(TD) to calculate advantage $A(s_t, a_t)$,

$$A(s_t, a_t) = r_t + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t),$$

where r_t is a reward after taking the action a_t in the state s_t, and $Q(s_t, a_t)$ is a value function. γ is a discount factor that denotes the trade-off between immediate rewards and future returns. We predict an action for each sample. Each sample has an individual unlikelihood token set, therefore each step action is independent. Thus, we do not need to consider a future return and the $A(s_t, a_t)$ calculation can be simplified as follows,

$$A(s_t, a_t) = r_t - Q(s_t, a_t)$$

Next, we optimize value function $Q(s_t, a_t)$. A2C trains a value network by minimizing MSE loss of TD,

$$\mathcal{L}_{value} = (r_t - Q(s_t, a_t))^2$$

Then, we add an entropy loss as regularization term to promote action diversity of the policy function,

$$\mathcal{L}_{entropy} = -\sum \mathrm{P}(\pi_\theta) \log \mathrm{P}(\pi_\theta).$$

We integrate the multiple losses together as the outer loss and update DA module as follows:

$$\mathcal{L}_{outer} = \mathcal{L}_{policy} + \mathcal{L}_{value} + \mathcal{L}_{entropy} \quad (7)$$

DA module parameter is optimized by,

$$\eta_{new} = \eta - \nabla_\eta \mathcal{L}_{\text{outer}}\ (\theta_{new}) \quad (8)$$

We show the detailed optimization process of in 1.

Algorithm 1 Optimization Process of TIMER.

Require: The training set $\mathbf{x}_s$, maximum iteration I, the iteration of inner loop N; for $i = 1$; $i < I$; $i++$; for $n = 1$; $n < N$; $n++$

1: Samples a batch from $\mathbf{x}_s$;
2: Update θ via Eq. 5 and Eq. 6 ;
3: Samples a batch from $\mathbf{x}_s$;
4: Update η via Eq. 7 and Eq. 8;

4. Experiment

We follow the previous studies (Jing et al., 2018; Chen et al., 2020, 2021) for data preprocessing, data splits (training, development, and test splits), and model evaluations. We use natural language generation (NLG) metric and clinical efficacy to evaluate our model and baseline, such as BLEU (Papineni et al., 2002), METEOR (Denkowski and Lavie, 2011), and ROUGE-L (Lin, 2004). To evaluate clinical efficacy, we utilize CheXpert (Irvin et al., 2019) to annotate generated reports for MIMIC and IU X-RAY. While there are other evaluation methods (e.g., RadGraph (Jain et al., 2021) and CheXbert (Smit et al., 2020)), we deploy the same metrics as the baselines for consistence. More details of data preprocessing and our implementations are in the Appendix, which allows for experiment replications.

Table 2: Performance summary. Δ indicates averaged percentage improvements of TIMER over baselines. Clinical Metric calculates the F1 score.

Methods	BLEU_1	BLEU_2	BLEU_3	BLEU_4	Meteor	Rouge_L	Clinical Metric
IU X-RAY							
BiLSTM	41.83	29.30	21.27	15.49	18.75	34.26	65.06
R2GEN	48.80	31.93	23.24	17.72	20.21	37.10	63.62
CMN	45.53	29.50	21.47	16.53	18.99	36.78	64.83
CMM+RL	49.30	30.08	21.45	16.10	20.10	38.20	40.79
TIMER	**49.34**	**32.49**	**23.84**	**18.61**	**20.38**	**38.25**	**94.40**
Δ	6.42	7.57	9.07	13.06	4.45	4.55	61.16
MIMIC							
BiLSTM	26.81	15.77	10.12	7.00	11.26	26.00	49.50
R2GEN	35.42	21.99	14.50	10.30	13.75	27.24	34.77
CMN	35.60	21.41	14.07	9.91	14.18	27.14	41.21
CMM+RL	38.10	22.10	14.45	10.02	14.53	27.66	28.36
TIMER	**38.30**	**22.49**	**14.60**	**10.40**	**14.70**	**28.00**	**75.86**
Δ	11.27	9.66	9.01	10.50	8.64	3.54	49.30

4.1. Baselines

To demonstrate the effectiveness of our model, we compare our approach with four state-of-the-art (SOTA) baselines[3] that use the same experimental settings, BiLSTM (Jing et al., 2018), R2Gen (Chen et al., 2020), CMN (Chen et al., 2021), and CMM+RL (Qin and Song, 2022). To ensure comparisons, we utilized their open-sourced models and followed their experimental settings.

BiLSTM (Jing et al., 2018) incorporates semantic tags into visual feature representations to generate radiology reports. To obtan the semantic tags, BiLSTM utilizes the Convolutional Neural Networks (CNNs) to predict semantic tags and corresponding visual abnormality regions, which later will be fed to the Long Short-Term Memory (LSTM) for report generation. The model utilizes a co-attention mechanism to localize regions containing abnormalities and generate narrations.

R2Gen (Chen et al., 2020) designs a memory-driven Transformer, where a relational memory is used to capture critical information in the generation process. Then the model proposes a memory-driven conditional layer normalization to incorporate the memory into the decoder of the Transformer. We reuse the code and released models from the study.

CMN (Chen et al., 2021) proposes a shared memory mechanism to record the alignment between images and texts so as to facilitate the interaction and generation across modalities. The memory mechanism is an intermediate medium and contains querying and mapping processes that enhance and smooth mappings between text and image representations. We use the author's released code and model to reproduce the results.

CMM + RL (Qin and Song, 2022) proposes a method for enhancing text and image representation alignments using reinforcement learning (RL). The RL treats the generation model as the agent that interacts with an external environment (image and text representations). The method provides appropriate supervision from NLG evaluation metrics to search for better mappings between features from different modalities. Our TIMER model uses the RL in a different way that our RL is to dynamically adapt token imbalance instead of aligning modalities.

5. Results and Analysis

This section provides an overview of the performance by both natural language generation (NLG) metrics and clinical efficacy (Irvin et al., 2019). We also present an imbalanced evaluation focusing on high and low frequencies token sets. Furthermore, we per-

3. We chose the SOTA methods that achieved the best performance during our experimental steps. **Note** that this direction evolves rapidly in recent years, and therefore there might be newer methods published during our study's submission and review.

Table 3: The imbalanced evaluation in the high- and low-frequency token set. We evaluate the F1 score by dividing tokens into high and low-frequency set with three different bucket size, (e.g., 1/8 represents top 1/8 frequent tokens as a high-frequency set).

		IU X-RAY		MIMIC	
Ratio	Method	low	high	low	high
1/8	Bi-LSTM	3.85	47.31	1.27	37.48
	R2GEN	4.46	**62.73**	2.52	52.01
	CMN	5.88	55.86	2.23	45.60
	CMM + RL	5.19	49.36	0.21	43.64
	TIMER	**13.23**	61.89	**3.15**	**52.66**
1/6	Bi-LSTM	1.97	44.06	0.89	30.77
	R2GEN	2.80	61.62	**2.00**	49.86
	CMN	5.75	65.12	0.85	**52.02**
	CMM + RL	5.08	49.26	0.14	43.36
	TIMER	**5.93**	**67.79**	**2.02**	51.72
1/4	Bi-LSTM	0.00	44.41	0.28	30.09
	R2GEN	1.16	59.98	0.00	48.77
	CMN	2.60	63.92	0.33	51.09
	CMM + RL	2.19	47.21	0.07	43.05
	TIMER	**8.66**	**64.00**	**0.58**	**51.39**

form an ablation analysis to assess the impact of individual modules in our approach.

5.1. Overall Performance

Table 2 presents overall performance. The results show that our approach significantly outperforms baselines, ranging from 3.54% to 61.16% improvements. Compared to the BiLSTM, the transformer-based models (R2GEN, CMN, CMM+RL) consistently perform well on all datasets among the four baseline generators. However, while CMM+RL's performance is on par with other baselines (e.g.,). Its BLEU_4 scores are relatively lower, indicating inefficiency in generating fluent sentences. One reason could be that CMM+RL employs a greedy sampling approach in self-critic learning, which fails to consider long-term returns. In contrast, our model significantly improves language fluency by achieving a 13.06% and 10.50% increase in BLEU_4 scores on IU-XRAY and MIMIC, respectively. Our model's most significant improvement is in the clinical metric, which we attribute to its focus on improving the prediction of infrequent tokens, particularly clinical vocabulary.

5.2. Imbalanced Performance

Table 3 presents a performance evaluation on infrequent tokens, which demonstrates effectiveness of our approach to learning token imbalance. We use F1-score to evaluate model performance of high and low-frequency token sets. We define three high-frequency by three levels, 1/4, 1/6, and 1/8, which refers to the percentage of top frequent tokens as high-frequency in the vocabulary and the rest are low frequent tokens. The setting is to demonstrate effectiveness of our approach in adapting imbalanced tokens.

By comparing the baselines, we find that our method significantly outperforms them in the low-frequency token set, both on IU X-RAY and MIMIC datasets by 1% to 245.6%. Our model improves performance of rare tokens by over 100% on IU X-RAY dataset and 41% on MIMIC dataset, highlighting the importance of addressing token imbalance issue. Neural models trained on skewed token distributions tend to overfit on frequent tokens, resulting in a greater prediction error for rare tokens. Furthermore, our method does not sacrifice model performance on the frequent token set. Compared to CMN, our approach achieves a 10.78% improvement on IU X-RAY and a 15.49% improvement on MIMIC with a ratio of 1/8.

5.3. Ablation Analysis

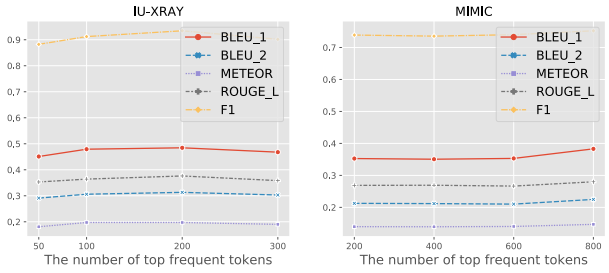
We performed an ablation analysis to evaluate the effectiveness of individual modules in our model. We used the notation "-DA+UL" to denote removing the DA module and replacing the module with a fixed frequent token set for the unlikelihood loss. We empirically selected the top 100 frequent tokens in IU X-RAY and the top 600 frequent tokens in MIMIC as the token set size. We used the same evaluation metrics as in the previous sections and summarized the performance results in Table 4.

Our results show that both the unlikelihood loss and DA modules can improve model performance across all metrics, proving the effectiveness of the proposed modules in generating radiology reports. However, we observed that adding the DA module has more performance improvements compared to the unlikelihood loss. This suggests that dynamic adaptation has a greater contribution in promoting model robustness overall and infrequent token adaptation.

Table 4: Ablation analysis. "-DA+UL" denotes removing DA and using unlikelihood loss with a fixed token set. "-DA" denotes the model is trained by negative log-likelihood loss 1 without unlikelihood loss. F1 is the clinical metric.

IU X-RAY			
	-DA + UL	-DA	full
BLEU_1	44.73	47.91	49.34
BLEU_2	29.08	30.60	32.49
BLEU_3	21.08	21.76	23.84
BLEU_4	16.36	15.98	18.61
METEOR	18.72	19.73	20.38
ROUGE_L	35.41	36.41	38.25
F1	82.43	89.72	94.40
MIMIC			
	-DA + UL	-DA	full
BLEU_1	34.02	35.31	38.30
BLEU_2	20.90	21.02	22.49
BLEU_3	13.99	13.73	14.60
BLEU_4	10.02	9.54	10.40
METEOR	13.76	14.03	14.69
ROUGE_L	27.43	26.66	28.00
F1	68.86	70.28	75.86

Figure 3: The performance comparison with different sizes of unlikelihood token sets. F1 is the clinical metric.



5.4. Dynamic Adaptation Analysis

TIMER dynamically adjusts the size of the unlikelihood token set to improve generator performance. This ablation analysis is to evaluate dynamic adaptation by comparing dynamic and static adjustments. For the static adaptation, the size of the unlikelihood token set is fixed at different sizes for the IU X-RAY and MIMIC datasets. The results of changes in the fixed sizes are visualized in Figure 3. The results show that different fixed sizes yield lower BLEU, METEOR, and ROUGE scores than dynamic adaptation, indicating that the size of the unlikelihood token set impacts generator performance. The effectiveness of dynamic adaptation provides further evidence of TIMER can improve overall performance and incorporate imbalance token adaptation.

5.5. Qualitative Analysis

To further investigate the effectiveness of our model's generation, we performed a qualitative analysis by selecting three generated samples from IU X-RAY and MIMIC-CXR datasets and compared them with the ground truth, as shown in Figure 4. Our analysis revealed that TIMER can generate descriptions that closely match the ground truth. Furthermore, compared to CMN, TIMER generated more medical terms such as "pleural effusion," "pneumothorax," and "mediastinal silhouette." Additionally, TIMER accurately diagnosed the medical condition and produced a semantically similar sentence to the ground truth. For instance, the sentence "there are no acute bony findings" generated by TIMER was semantically similar to the ground truth sentence "bony structures are intact."

6. Related Work

Radiology report generation is to generate descriptive text from radiology images. An encoder-decoder network is the primary neural architecture of the task. For instance, (Jing et al., 2018) built a multi-task learning framework that employs a hierarchical LSTM model to generate long radiology reports, and a co-attention mechanism to jointly perform the prediction of tags. Recent studies deployed Transformer architecture (Vaswani et al., 2017) to improve the task performance (Lovelace and Mortazavi, 2020; Chen et al., 2020, 2021). For example, (Chen et al., 2020) proposed a relational memory to record key information in the generation process and applied a memory-driven conditional layer normalization to incorporate the memory into the decoder of the Transformer. (Chen et al., 2021) designed a shared memory between the encoder and decoder to record the alignment between images and texts. Several recent works (Dalla Serra et al., 2022; Li et al., 2022; Yang et al., 2022) have enhanced the quality of text generation by developing models that can learn clinical knowledge directly from reports. For

Figure 4: Qualitative comparison between TIMER and CMN. We highlight correct predictions of infrequent tokens. We set top 20% tokens in the vocabulary as frequent and the rest as infrequent tokens.

TIMER	CMN	Ground Truth
the cardiomediastinal silhouette and pulmonary vasculature are within limits in size. the lung are clear of focal airspace disease pneumothorax or pleural effusion. there are no acute bony findings.	studies may tubing calcified outside 13 vasculature levocurvature level sequelae represents slight. studies zone vasculature fracture reflect hyperinflation contours hypoinflated obstructive jugular lordotic diffuse. additional vasculature through left aortic elevated.	cardiomediastinal silhouette are within normal limits . the lung are clear . bony structures are intact .
the heart size and pulmonary vascularity appear within normal limits. the lungs are free of focal airspace disease . no pleural effusion or pneumothorax is seen .	studies borderline slight calcified outside collapse evidence levocurvature level sequelae. studies zone vasculature medial reflect hyperinflation contours hypoinflated. through lordotic diffuse jugular obstructive somewhat 12.	normal heart size and mediastinal silhouette are normal . stable calcification in the left upper lobe xxxx representing a granuloma . no focal airspace opacities . no pleural effusion or pneumothorax . visualized osseous structures are unremarkable in appearance.
the lung are clearly bilaterally . specifically no evidence of focal consolidation penumothorax or pleural effusion . cardio mediastinal silhouette is unremarkable . visualized osseous structures of the thorax are without acute adnormality.	studies zone vasculature fracture thorax . dextro through evaluation reflect hyperinflation surgical obstructive jugular lordotic diffuse . minimal exaggerated tubing somewhat dextrocurvature. scoliotic possible l1 reflect studies incidental vasculature attenuation left consolidation .	the lung are clearly bilaterally . cardio and mediastinal silhouette are normal . Pulmonary vasculature is normal . no penumothorax or pleural effusion . no acute bony abnormality .

instance, Yang et al. (2022) devised an automatic information extraction mechanism to extract clinical entities and relations directly from training reports, while Dalla et al.(Dalla Serra et al., 2022) extracted entities and relations from images and then generated complete reports based on the extracted entities and relations. To better evaluate the factualness of generated reports, (Smit et al., 2020) trained a BERT-based model to label the reports' diseases. More recently, Some works (Miura et al., 2021; Delbrouck et al., 2022; Qin and Song, 2022) have incorporated reinforcement learning (RL) into radiology reports to improve their quality. Delbrouck et al.(Delbrouck et al., 2022) improved the quality of report generation by predicting more precise entities, while Qin et al. (Qin and Song, 2022) improved report generation performance by better aligning the images and text with RL. However, none of these studies have explicitly worked on token imbalance, which is the main focus of our study.

Reinforcement Learning (RL) in NLG tasks is to improve sequence prediction (Shi et al., 2018; Bahdanau et al., 2016; Hao et al., 2022). The actor-critic approach proposed by Bahdanau et al. considers the task objective during training, improving maximum likelihood training but suffering from sparse reward. To address this, Shi et al. uses the maximum entropy Inverse Reinforcement Learning to mitigate the issues of reward sparsity. Wu and Huang (2022) employs reinforcement learning to address label imbalance issues across various domains. However, this is a classification task and it is challenging to apply it directly to generation tasks. In this study, we employ an actor-critic approach to optimize our model, but instead of applying it to the NLG model directly, we propose a novel learning strategy that updates the unlikelihood token set dynamically. While (Nishino et al., 2020) applies RL to radiology report generation, a key difference is that they focus on document label imbalance rather than the token imbalance, which is the primary target of our study.

Imbalance modeling refers to the task of modeling skewed distributions. Various strategies have been proposed to handle imbalanced data in natural language processing tasks, such as oversampling, under-sampling, and the few-shot technique (Tian et al., 2021; Yang et al., 2020). While existing solutions focus on classification tasks, those techniques may not apply to radiology report generation, a multimodal image-to-text generation in healthcare. In

radiology report generation, the imbalance between normal and abnormal samples has been addressed in prior studies through methods such as data augmentation with reinforcement learning (Nishino et al., 2020) and using separate LSTMs for abnormal and normal sentence generation (Harzig et al., 2019). However, no previous study has focused explicitly on imbalanced token distributions in medical report generation. In this work, we propose a novel approach that leverages reinforcement learning techniques to address this issue.

7. Conclusion

In this study, we have proved the importance of infrequent tokens in radiology report generation and proposed a reinforcement-learning method to adapt token imbalance. We demonstrate effectiveness of our approach (TIMER) over three state-of-art baselines on radiology report generation. TIMER automatically penalizes overfitting on frequent tokens and dynamically adjusts rewards on infrequent token generations. Extensive experiments and ablation analysis show that TIMER can obtain significant improvements on infrequent token generations while maintaining performance on frequent tokens by multiple evaluation metrics. While we evaluate our approach on radiology report generation, we expect its broad applicability to text generation tasks and will extend applications in our future work.

Limitations

While we have proved the effectiveness of our proposed method, three primary limitations must be acknowledged to appropriately interpret our evaluation results, task and preprocessing.

Task. We primarily focus on the task of radiology report generation, while the task is only one of the downstream evaluations in the text generation field. Experiments on the radiological benchmarks may not generalize to all text generation tasks, and infrequent tokens may not contain critical and complex medical terms. In this study, to ensure consistency, we compare with the SOTA baselines on the same datasets of radiology report generation. **Note** that this task has evolved rapidly in recent years, and therefore there might be newer methods published during our study's publishing processes.

Preprocessing. We utilize the datasets and follow the same preprocessing steps from the previous study (Chen et al., 2020). Existing studies (Nguyen et al., 2021; Qin and Song, 2022) may have variations in the preprocessing steps that can directly impact the final results. For example, (Nguyen et al., 2021) selected the top 900 frequent tokens for performance evaluations, which shows a significant improvement in the frequent tokens. To ensure a consistent comparison, we keep the same experimental settings and deploy the released models from the baselines (Jing et al., 2018; Chen et al., 2020, 2021) that reported state-of-the-art performance on the same datasets. Our future work will develop more comprehensive evaluations of rare tokens across the existing models.

Human Evaluation. It is necessary to invite radiologists for human evaluations. However, we did not include the approach due to *subjectivity* and *domain* challenges. A common approach is to sample limited generated reports from a pair of methods (Miura et al., 2021; J Kurisinkel et al., 2021). However, the same baselines with different human evaluators may yield varied evaluation results (Liu et al., 2021b,a). It is a challenge to have enough certified radiologists to evaluate the task. We conducted preliminary human evaluations by the qualitative analysis in Table 4, though we miss enough support from radiologists. Having human evaluations will be our future work to provide more comprehensive perspectives.

Ethic and Privacy Concerns

We follow data agreement and training procedures to access the two radiology report datasets. To protect user privacy, we have followed corresponding data agreements to ensure proper data usage and experimented with the de-identified data. Our experiments do not store any data and only use available multimodal entries for research demonstrations. Due to privacy and ethical considerations, we will not release any clinical data associated with patient identities. Instead, we will release our code and provide detailed instructions to replicate our analysis and experiments. This study only uses publicly available and de-identified data.

Acknowledgments

The authors want to thank the anonymous reviewers for their constructive suggestions. This work was supported by a gift from West Cancer Foundation, Ralph E. Powe Junior Faculty Enhancement Award, and the National Science Foundation with award number IIS-2245920. With gifts from Adobe Research, we purchased a GPU workstation for this study.

References

Dzmitry Bahdanau, Philemon Brakel, Kelvin Xu, Anirudh Goyal, Ryan Lowe, Joelle Pineau, Aaron Courville, and Yoshua Bengio. An Actor-Critic Algorithm for Sequence Prediction. *arXiv e-prints*, art. arXiv:1607.07086, July 2016. URL https://arxiv.org/pdf/1607.07086.pdf.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1439–1449, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.112. URL https://aclanthology.org/2020.emnlp-main.112.

Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. Cross-modal memory networks for radiology report generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5904–5914, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.459. URL https://aclanthology.org/2021.acl-long.459.

Francesco Dalla Serra, William Clackett, Hamish MacKinnon, Chaoyang Wang, Fani Deligianni, Jeff Dalton, and Alison Q. O'Neil. Multimodal generation of radiology reports using knowledge-grounded extraction of entities and relations. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 615–624, Online only, November 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.aacl-main.47.

Jean-Benoit Delbrouck, Pierre Chambon, Christian Bluethgen, Emily Tsai, Omar Almusa, and Curtis Langlotz. Improving the factual correctness of radiology report generation with semantic rewards. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4348–4360, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.findings-emnlp.319.

Dina Demner-Fushman, Marc D. Kohli, Marc B. Rosenman, Sonya E. Shooshan, Laritza Rodriguez, Sameer Antani, George R. Thoma, and Clement J. McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 07 2015. ISSN 1067-5027. doi: 10.1093/jamia/ocv080. URL https://doi.org/10.1093/jamia/ocv080.

Michael Denkowski and Alon Lavie. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 85–91, Edinburgh, Scotland, July 2011. Association for Computational Linguistics. URL https://aclanthology.org/W11-2107.

Yongchang Hao, Yuxin Liu, and Lili Mou. Teacher forcing recovers reward functions for text generation. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=1_gypPuWUC3.

Philipp Harzig, Yan-Ying Chen, Francine Chen, and Rainer Lienhart. Addressing data bias problems for chest x-ray image report generation. *arXiv preprint arXiv:1908.02123*, 2019. URL https://arxiv.org/pdf/1908.02123.pdf.

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. Chexpert: A large chest radiograph dataset with uncertainty labels and expert

comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 590–597, Jul. 2019. doi: 10.1609/aaai.v33i01.3301590. URL https://ojs.aaai.org/index.php/AAAI/article/view/3834.

Litton J Kurisinkel, Ai Ti Aw, and Nancy F Chen. Coherent and concise radiology report generation via context specific image representations and orthogonal sentence states. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers*, pages 246–254, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-industry.31. URL https://aclanthology.org/2021.naacl-industry.31.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew Lungren, Andrew Ng, Curtis Langlotz, Pranav Rajpurkar, and Pranav Rajpurkar. Radgraph: Extracting clinical entities and relations from radiology reports. In J. Vanschoren and S. Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/file/c8ffe9a587b126f152ed3d89a146b445-Paper-round1.pdf.

Baoyu Jing, Pengtao Xie, and Eric Xing. On the automatic generation of medical imaging reports. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2577–2586, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1240. URL https://aclanthology.org/P18-1240.

Baoyu Jing, Zeya Wang, and Eric Xing. Show, describe and conclude: On exploiting the structure information of chest X-ray reports. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6570–6580, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1657. URL https://aclanthology.org/P19-1657.

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 de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1): 317, dec 2019. ISSN 2052-4463. doi: 10.1038/s41597-019-0322-0. URL https://doi.org/10.1038/s41597-019-0322-0.

M. Li, W. Cai, K. Verspoor, S. Pan, X. Liang, and X. Chang. Cross-modal clinical graph transformer for ophthalmic report generation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20624–20633, Los Alamitos, CA, USA, jun 2022. IEEE Computer Society. doi: 10.1109/CVPR52688.2022.02000. URL https://doi.ieeecomputersociety.org/10.1109/CVPR52688.2022.02000.

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. URL https://aclanthology.org/W04-1013.

Fenglin Liu, Shen Ge, and Xian Wu. Competence-based multimodal curriculum learning for medical report generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3001–3012, Online, August 2021a. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.234. URL https://aclanthology.org/2021.acl-long.234.

Fenglin Liu, Changchang Yin, Xian Wu, Shen Ge, Ping Zhang, and Xu Sun. Contrastive attention for automatic chest X-ray report generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 269–280, Online, August 2021b. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.23. URL https://aclanthology.org/2021.findings-acl.23.

Justin Lovelace and Bobak Mortazavi. Learning to generate clinically coherent chest X-ray reports. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1235–1243, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.110. URL https://aclanthology.org/2020.findings-emnlp.110.

Yasuhide Miura, Yuhao Zhang, Emily Tsai, Curtis Langlotz, and Dan Jurafsky. Improving factual completeness and consistency of image-to-text radiology report generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5288–5304, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.416. URL https://aclanthology.org/2021.naacl-main.416.

Jiaqi Mu and Pramod Viswanath. All-but-the-top: Simple and effective postprocessing for word representations. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=HkuGJ3kCb.

Hoang Nguyen, Dong Nie, Taivanbat Badamdorj, Yujie Liu, Yingying Zhu, Jason Truong, and Li Cheng. Automated generation of accurate & fluent medical X-ray reports. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3552–3569, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.288. URL https://aclanthology.org/2021.emnlp-main.288.

Irene Nikkarinen, Tiago Pimentel, Damián Blasi, and Ryan Cotterell. Modeling the unigram distribution. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3721–3729, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.326. URL https://aclanthology.org/2021.findings-acl.326.

Toru Nishino, Ryota Ozaki, Yohei Momoki, Tomoki Taniguchi, Ryuji Kano, Norihisa Nakano, Yuki Tagawa, Motoki Taniguchi, Tomoko Ohkuma, and Keigo Nakamura. Reinforcement learning with imbalanced dataset for data-to-text medical report generation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2223–2236, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.202. URL https://aclanthology.org/2020.findings-emnlp.202.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA, 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://doi.org/10.3115/1073083.1073135.

Han Qin and Yan Song. Reinforced cross-modal alignment for radiology report generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 448–458, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.38. URL https://aclanthology.org/2022.findings-acl.38.

Zhan Shi, Xinchi Chen, Xipeng Qiu, and Xuanjing Huang. Toward diverse text generation with inverse reinforcement learning. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, IJCAI'18, page 4361–4367. AAAI Press, 2018. ISBN 9780999241127. URL https://arxiv.org/abs/1804.11258.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Ng, and Matthew Lungren. Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1500–1519, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.117. URL https://aclanthology.org/2020.emnlp-main.117.

Richard S. Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*, NIPS'99, page 1057–1063, Cambridge, MA, USA, 1999. MIT Press. URL https://proceedings.neurips.cc/paper/1999/file/464d828b85b0bed98e80ade0a5c43b0f-Paper.pdf.

Jiachen Tian, Shizhan Chen, Xiaowang Zhang, Zhiyong Feng, Deyi Xiong, Shaojuan Wu, and Chunliu Dou. Re-embedding difficult samples via mutual information constrained semantically oversampling for imbalanced text classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3148–3161, Online and Punta Cana, Dominican

Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.252. URL https://aclanthology.org/2021.emnlp-main.252.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964. URL https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Lingxiao Wang, Jing Huang, Kevin Huang, Ziniu Hu, Guangtao Wang, and Quanquan Gu. Improving neural language generation with spectrum control. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=ByxY8CNtvr.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. Neural text generation with unlikelihood training. In *ICLR*, 2020. URL https://openreview.net/forum?id=SJeYe0NtvH.

Yuexin Wu and Xiaolei Huang. Unsupervised reinforcement adaptation for class-imbalanced text classification. In *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics*, pages 311–322, Seattle, Washington, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.starsem-1.27. URL https://aclanthology.org/2022.starsem-1.27.

Shuxin Yang, Xian Wu, Shen Ge, S Kevin Zhou, and Li Xiao. Knowledge matters: Chest radiology report generation with general and specific knowledge. *Medical image analysis*, 80:102510, August 2022. ISSN 1361-8415. doi: 10.1016/j.media.2022.102510. URL https://doi.org/10.1016/j.media.2022.102510.

Wenshuo Yang, Jiyi Li, Fumiyo Fukumoto, and Yanming Ye. HSCNN: A hybrid-Siamese convolutional neural network for extremely imbalanced multi-label text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6716–6722, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.545. URL https://aclanthology.org/2020.emnlp-main.545.

Sangwon Yu, Jongyoon Song, Heeseung Kim, Seongmin Lee, Woo-Jong Ryu, and Sungroh Yoon. Rare tokens degenerate all tokens: Improving neural text generation via adaptive gradient gating for rare token embeddings. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29–45, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.3. URL https://aclanthology.org/2022.acl-long.3.