applied sciences

MDPI

Article

Contrastive Learning Penalized Cross-Entropy with Diversity Contrastive Search Decoding for Diagnostic Report Generation of Reduced Token Repetition

Taozheng Zhang [1,2], Jiajian Meng [2], Yuseng Yang [2] and Shaode Yu [1,2,*]

[1] State Key Laboratory of Media Convergence and Communication, Communication University of China, Beijing 100024, China; zhangtaozheng@cuc.edu.cn
[2] School of Information and Communication Engineering, Communication University of China, Beijing 100024, China; mengjiajian@cuc.edu.cn (J.M.); 2020211123015@cuc.edu.cn (Y.Y.)
* Correspondence: yushaodemia@163.com

Abstract: Medical imaging description and disease diagnosis are vitally important yet time-consuming. Automated diagnosis report generation (DRG) from medical imaging description can reduce clinicians' workload and improve their routine efficiency. To address this natural language generation task, fine-tuning a pre-trained large language model (LLM) is cost-effective and indispensable, and its success has been witnessed in many downstream applications. However, semantic inconsistency of sentence embeddings has been massively observed from undesirable repetitions or unnaturalness in text generation. To address the underlying issue of anisotropic distribution of token representation, in this study, a contrastive learning penalized cross-entropy (CLpCE) objective function is implemented to enhance the semantic consistency and accuracy of token representation by guiding the fine-tuning procedure towards a specific task. Furthermore, to improve the diversity of token generation in text summarization and to prevent sampling from unreliable tail of token distributions, a diversity contrastive search (DCS) decoding method is designed for restricting the report generation derived from a probable candidate set with maintained semantic coherence. Furthermore, a novel metric named the maximum of token repetition ratio (maxTRR) is proposed to estimate the token diversity and to help determine the candidate output. Based on the LLM of a generative pre-trained Transformer 2 (GPT-2) of Chinese version, the proposed CLpCE with DCS (CLpCEwDCS) decoding framework is validated on 30,000 desensitized text samples from the "Medical Imaging Diagnosis Report Generation" track of 2023 Global Artificial Intelligence Technology Innovation Competition. Using four kinds of metrics evaluated from *n*-gram word matching, semantic relevance, and content similarity as well as the maxTRR metric extensive experiments reveal that the proposed framework effectively maintains semantic coherence and accuracy (BLEU-1, 0.4937; BLEU-2, 0.4107; BLEU-3, 0.3461; BLEU-4, 0.2933; METEOR, 0.2612; ROUGE, 0.5182; CIDER, 1.4339) and improves text generation diversity and naturalness (maxTRR, 0.12). The phenomenon of dull or repetitive text generation is common when fine-tuning pre-trained LLMs for natural language processing applications. This study might shed some light on relieving this issue by developing comprehensive strategies to enhance semantic coherence, accuracy and diversity of sentence embeddings.

Keywords: diagnostic report generation; contrastive learning; cross entropy; diversity contrastive search; large language model

Citation: Zhang, T.; Meng, J.; Yang, Y.; Yu, S. Contrastive Learning Penalized Cross-Entropy with Diversity Contrastive Search Decoding for Diagnostic Report Generation of Reduced Token Repetition. *Appl. Sci.* **2024**, *14*, 2817. https://doi.org/10.3390/app14072817

Academic Editor: Juan A. Gómez-Pulido

Received: 26 January 2024
Revised: 16 March 2024
Accepted: 20 March 2024
Published: 27 March 2024

1. Introduction

Text summarization aims to compress a long text document into a short and human-readable form with the most important information of the source document [1]. There are two broad kinds of approaches, extractive and abstractive. The extractive approaches generate summaries through retrieving the most relevant and important phrases or sentences from the original text, while the abstractive approaches delve into the meaning

and semantics and utilizes natural language generation techniques to create a new and comprehensive text summary [2].

As a specific application of text summarization, diagnosis report generation (DRG) aims to make summaries and generate diagnostic reports according to the text description of medical imaging findings. It is part of the medical report generation [3] task which concentrates on using deep learning networks for generating diagnosis reports in terms of the medical image input. Clinically, medical imaging description and disease diagnosis are predominant in radiologists' daily works. These works are vitally important yet tedious and time-consuming. Accurate DRG from medical imaging description in an automated manner can decrease clinicians' workload dramatically, subsequently improving their routine efficiency. However, medical imaging diagnostic reports involve specific field vocabulary, complex organization structure and detailed visual description [4]. Due to the professionalism in disease diagnosis, treatment planning and therapeutic delivery, higher demands are required for the DRG quality, including precise comprehension of medical terminology, content understanding and reasoning capabilities, coherent diagnosis and ambiguity avoidance.

Abstractive approaches for high-quality DRG have been developed. Traditional methods mainly rely on statistics and shallow learning, such as using maximum entropy models to predict words or constructing feature engineering and classifiers to generate key sentences. These methods might be unable to handle large-scale text document inputs [5]. On the other hand, significant progress has been made in the field of natural language processing (NLP) by using deep learning networks [6], such as recurrent neural networks (RNNs) [7] and long short-term memory (LSTM) networks [8]. One milestone comes from the attention mechanisms of Transformers [9] that build encoder–decoder-based sequence-to-sequence models, and essential messages in the input text are concentrated on. Later, as a cost-effective approach, a great deal of attention has been paid to pre-trained large language models (LLMs), such as bidirectional encoder representations from Transformers (BERT) [10] and generative pre-trained Transformer (GPT) [11]. Through pre-training on large-scale corpora, LLMs can effectively improve the performance in massive downstream tasks, including but not limited to clinical notes summarization [12], biomedical natural language tasks [13] and text-to-image generation [14], and LLMs outperform medical experts in clinical text summarization [15] that could help clinicians to focus more on patient care.

Unfortunately, when transferring a pre-trained LLM to a specific application, semantic inconsistency of sentence embeddings has been massively witnessed from dull repetitions and undesirable text generation. It might be derived from the inconsistent representation of sentence embeddings, anisotropic distributions of token generation, and a narrow subset of the entire representation space [16–18]. When the distance between different tokens in a representation space is close, these tokens have high cosine similarity. A showcase reveals that cosine similarities between tokens within a sentence could be larger than 0.95, and therefore, duplicate tokens will be unavoidably generated at different stages [19].

To solve or to relieve this degradation problem, massive attempts have been made. One feasible way is mapping the generated sentence vectors into an isotropic and uniform distribution space. For instance, BERT-flow [20] turns the sentence representations from BERT encoder into a smooth and isotropic Gaussian distribution space using a reversible flow transformation. It achieves significant improvement on several semantic textural similarity tasks. Wang and his colleagues [21] design a dual-stream attention mechanism and use a positional residual strategy to improve the robustness of extractive summarization. A summary method based on two-layer Transformer in [22] employs BART (bidirectional and autoregressive Transformer) [23] and T5 (text-to-text transfer converter) [24] to ensure the summary coherence. Another promising way is from contrastive learning (CL). Traditional text augmentation is used to construct positive and negative sample pairs from the augmented sentence set. Its training objective becomes making the embeddings of positive sample pairs closer and the distance of negative pairs farther. Debiased CL [25] is this kind of approach that samples appropriate same-label data points, since negative pairs sampled

from different labels or classes improve performance [26], achieving consistent improvement on language, vision and reinforcement learning benchmarks. The contrastive learning for sentence representation (CLEAR) method [27] employs multiple sentence-level augmentation strategies, and during pre-training, different sentence augmentation strategies result in improvement on specific tasks. Token-aware CL (TaCL) [18] is a continual pre-training approach that is fully unsupervised and requires no additional samples. It embraces a teacher model and a student model, and both are initialized with the same pre-trained BERT. The objective function contains a masked language modeling term, a next sentence prediction term and a token-aware contrastive learning term for learning an isotropic and discriminative distribution of token representations. For reducing the impact of summary false negatives and effectively maintaining spatial consistency, a metric score is employed to dynamically penalize positive and negative samples during model training [28]. In extractive multi-document summarization, a contrastive hierarchical discourse graph is designed to capture complex discourse relationships and global topic coherence, and it shows excellent performance [29]. In any case, compared to greedy search (GS) [30] and nucleus search (NS) [31] decoding methods, some other decoding methods seem more promising to relieve this anisotropy problem [32]. For instance, a contrastive search (CS) decoding method injects CL into the text decoding stage, and its performance is verified to be better than traditional decoding methods [19]. On open-ended text generation, an empirical study [33] of CS and contrastive decoding indicates that CS substantially outperforms contrastive decoding in terms of the diversity and coherence metrics. The fidelity-enriched contrastive search (FECS) method [34] augments the CS framework with context-aware regularization terms, and in both abstractive summarization and dialogue generation tasks, it has been confirmed to improve semantic coherence among tokens, mitigate repetition, and strengthen fidelity to the provided source labels in the generated output. To reduce the number of repeated tokens in text generation when using encoder–decoder models, a repetition reduction module (RRM) [35] is proposed to supervise the training procedure by capturing the consistency of a sentence sample between the encoding and decoding sides.

In this study, a contrastive learning penalized cross-entropy with diversity contrastive search (CLpCEwDCS) decoding framework is proposed. To improve the consistency of sentence embeddings and to relieve the anisotropy issue, CL is integrated into the fine-tuning stage and a novel objective function is formed as contrastive learning penalized cross-entropy (CLpCE). Moreover, in the decoding stage, a diversity contrastive search (DCS) decoding method is designed to balance the diversity and quality of report generation. For mitigating degenerative behaviors, the core idea of the DCS decoding method is different from the FECS method [34]. FECS promotes the diversity by augmenting a faithfulness reward term into the CS framework, while DCS determines the outcome via the estimation of the maximum token repetition ratio (maxTRR) of candidate outputs. Specifically, the proposed metric maxTRR estimates the token repetitions in the token space before the text generation, while the measure of word-, phrase-, and the sentence-level consecutive repetitions [36] or the subsentence-level consecutive repetition [35] is for performance evaluation after the text generation. Overall, the contributions of this study can be summarized as follows:

1. An objective function CLpCE is designed for balancing both unsupervised and supervised learning in the model fine-tuning stage to enhance the consistency of feature representation of sentence embeddings.
2. A novel decoding method DCS is developed to improve the representation diversity and to relieve anisotropic distributions of token generation with maintained quality of text summarization.
3. A supplementary metric named the maximum of token repetition ratio (maxTRR) is implemented which estimates the token repetition and determines the outcome of text generation.
4. The effectiveness of the proposed CLpCEwDSC decoding framework is verified, and competitive performance and better diversity are observed on the DRG task.

The remainder of this paper is organized as follows: Section 2 presents the relevant techniques of GPT-2 and contrastive learning of sentence embeddings. The data collection, the proposed framework, experiment design, implementation details and parameter settings are shown in Section 3. We then report the DRG accuracy and diversity and the effect of the diversity control in Section 4. After that, we discuss the results and some limitations of this work in Section 5, and conclude this work and future directions in Section 6.

2. Related Techniques

This section introduces related techniques and computing theories, including GPT-2 decoder block, contrastive learning of sentence embeddings in semantic representation, and contrastive search decoding.

2.1. GPT-2 Decoder Block

Figure 1 shows the diagram structure of Transformer decoder block and GPT-2 decoder block. In comparison to the Transformer decoder block, GPT-2 decoder block is simplified without multi-head self-attention module.

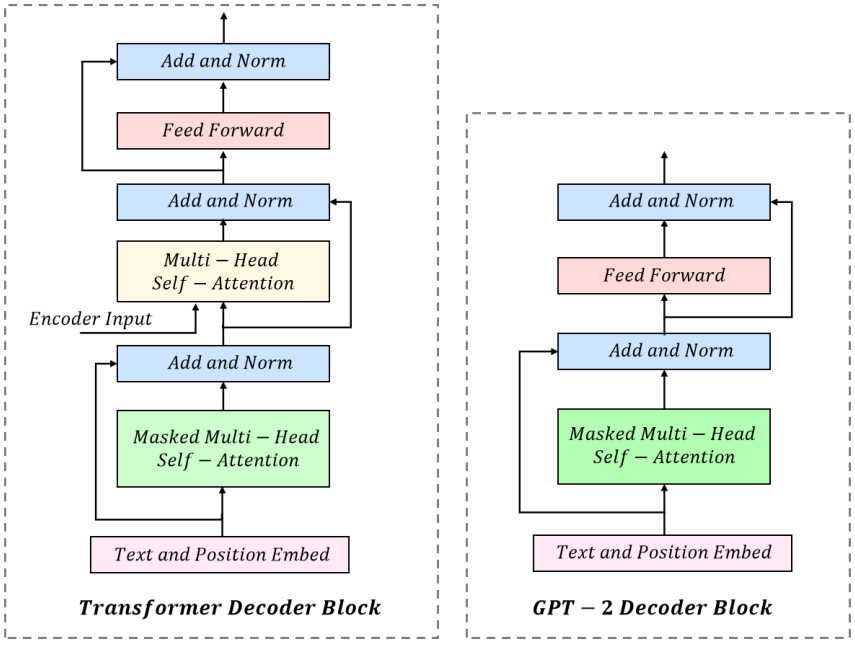

Figure 1. The structure of Transformer and GPT-2 decoder blocks.

GPT-2 is a large model trained in an unsupervised manner. For an unlabeled text sequence $\vec{t} = \{t_1, \cdots, t_i, \cdots, t_n\}$, it is trained by maximizing the likelihood function as below,

$$L^{PT}(\vec{t}) = \sum_i \log\{P(t_i|t_{i-k}, t_{i-k+1}, \cdots, t_{i-1}; \Theta)\}, \tag{1}$$

where PT stands for "pre-training", Θ denotes model parameters, and k historical tokens $\{t_{i-k}, t_{i-k+1}, \cdots, t_{i-1}\}$ are used to predict the current token t_i.

In the fine-tuning stage for a specific task, labeled samples are used for supervised learning. For an input sequence set $\{(\vec{t}_j, y^j)\}$ with $\vec{t}_j = \{t_1^j, \cdots, t_i^j, \cdots, t_n^j\}$ and label y^j, the fine-tuning of GPT-2 is by optimizing the loss function as

$$\begin{aligned} L^{FT}(\vec{t}, y) &= \sum_{(\vec{t},y)} \log\{P(y|\vec{t})\} \\ &= \sum_{(\vec{t},y)} \log\{softmax(h_n^{[L]} \cdot W^Y)\} \end{aligned} \tag{2}$$

in which FT stands for "fine-tuning", $h_n^{[L]}$ denotes the hidden state output of the last token in the $\vec{t}$ sequence from the last layer of the GPT-2 decoder block. $W^Y \in R^{d \times l}$ is the weight matrix of the fully connected layer, where d is the embedding dimension and l is the number of labels.

The GPT-2 embraces large and diverse multi-domain data for model pre-training, and the parameters are shared across different tasks, both enhancing the generalization capacity on specific downstream applications.

2.2. Contrastive Learning of Sentence Embeddings

Contrastive learning of sentence embeddings can improve semantic representation by minimizing the distance between similar samples and maximizing the distance between dissimilar samples [37]. For a small batch of sentence pairs $D = \{(x_i, x_i^+)\}$, where x_i^+ is the positive sample of x_i and they are semantically related sentence pairs, the training objective function of (x_i, x_i^+) is

$$\begin{aligned} L_i &= -\log \frac{e^{cos(h_i, h_i^+)/\tau}}{\sum e^{cos(h_i, h_i^+)/\tau}} \\ cos(h_i, h_i^+) &= \frac{h_i^T \cdot h_i^+}{\|h_i\| \cdot \|h_i^+\|} \end{aligned} \tag{3}$$

where τ is the temperature coefficient, (h_i, h_i^+) is the sentence vector representation of (x_i, x_i^+) obtained through pre-trained models $h = f_\Theta(\vec{x})$, and $cos(h_i, h_i^+)$ calculates cosine similarity between (h_i, h_i^+).

Simple contrastive learning of sentence embeddings (SimCSE) [38] is an efficient framework. Its core principle can be described as follows. For a small batch of sentence $\{x_i\}_{i=1}^N$, x_i^+ is set equal to x_i, and then, independently sampled dropout masks are used on (x_i, x_i^+) to obtain forward sentence pairs. In general, Transformers set dropout after the feed-forward layer and attention layer. Thus, $h_i^m = f_\Theta(x_i, m)$, and m is a random mask of dropout. By utilizing the random mask property of dropout, the same input is fed into the encoder twice to obtain two different dropout masks $\{m, m^+\}$.

In SimCSE [38], the embeddings of the forward sentence pairs and the training objective function can be expressed as

$$\begin{aligned} h_i^{m_i} &= f_\theta(x_i, m_i) \\ h_i^{m_i^+} &= f_\theta(x_i^+, m_i^+) \\ L_i &= -\log \frac{e^{cos(h_i^{m_i}, h_i^{m_i^+})/\tau}}{\sum e^{cos(h_i^{m_i}, h_i^{m_i^+})/\tau}} \end{aligned} \tag{4}$$

in which m is from the built-in dropout of Transformer. It should be noted that no dropout structures is added in our model, and the random noise brought by dropout can be viewed as a form of data augmentation.

In the CL field, compared to traditional text augmentation methods, using the built-in dropout mask in pre-trained models leads to simpler implementation, higher-quality sentence embeddings and better performance on numerous unsupervised and supervised downstream tasks [38].

2.3. Contrastive Search Decoding

In order to ensure the generated output semantically coherent with these generated prefix texts, the key idea of CS decoding is to find out the most likely candidate set and to guarantee the output with sufficient discriminative capacity. Given the previous generated text $x_{<t}$, the choice of generating x_t at time t should satisfy

$$x_t^* = \arg\max_{v \in V^{(k)}} \{(1-\alpha) \times p_\theta(v|x_{<t}) - \alpha \times (\max\{cos(h_v, h_{x_j}) : 1 \leq j \leq t-1\})\} \quad (5)$$

in which $V^{(k)}$ denotes the prediction set of the probability distribution space $p_\Theta(\cdot|x_{<t})$, $p_\Theta(v|x_{<t})$ stands for the model confidence that presents the probability of the candidate v, and $\max\{cos(h_v, h_{x_j}) : 1 \leq j \leq t-1\}$ is the degradation penalty which measures the similarity of the candidate v and all tokens in the text set $x_{<t}$.

A larger degradation penalty means candidate v is more similar to the previous text $x_{<t}$, and thus, it is more likely to represent the previous content. The parameter $\alpha \in [0, 1]$ is used to adjust the importance between the components. When $\alpha = 0$, CS decoding degenerates into GS decoding.

3. Materials and Methods

This section presents the data collection and outlines the proposed framework. Subsequently, the experiment design, evaluation metrics, implementation details, and parameter settings are described for performance comparison.

3.1. Data Collection

The dataset comes from the "Medical Imaging Diagnosis Report Generation" track of a nationwide open competition "2023 Global Artificial Intelligence Technology Innovation Competition" (https://gaiic.caai.cn/ai2023/, accessed on 19 March 2024) hosted by the Chinese Association for Artificial Intelligence. It is the newest and highest-quality dataset with the purpose of generating medical diagnosis reports according to medical image descriptions.

The dataset consists of 30,000 plain-text data samples, including descriptions of patient scans and corresponding diagnostic reports in Chinese. For instance, a text sample shows "Image Description" as "There is a local bone defect in the left parietal bone. There are small areas of decreased density adjacent to the lateral ventricles on both sides. An arc-shaped cerebrospinal fluid density shadow is observed below the right frontal skull. The ventricular system is enlarged, and the sulci, fissures, and cisterns of the brain are widened. There is no displacement of the midline structures. Poor pneumatization is observed in both mastoids, with increased density inside." and its "Diagnosis Report" is as "There is a local defect in the left parietal bone, which may require surgical intervention. There are also scattered ischemic lesions in the brain. Additionally, there is a small amount of subdural effusion in the right frontal region, and the patient has bilateral mastoiditis.".

To avoid issues such as privacy leakage, the dataset provided for the competition is desensitized on a character-by-character basis. Thus, the aforementioned text sample becomes "Image Description" of the desensitized data "(14 108 28 30 15 13 294 29 20 18 23 21 25 32 16 14 39 27 14 47 46 69 70 11 24 42 26 37 61 24 10 79 46 62 19 13 31 95 19 28 20 18 10 22 12 38 41 17 23 21 36 53 25 10)" and "Diagnosis Report" of the desensitized data "(22 12 38 41 17 81 10)".

3.2. The Proposed CLpCEwDCS Decoding Framework

This sub-section gives the reasons for backbone network selection and then elaborates on the formulation of the CLpCE objective function and the DCS decoding procedure. During DCS decoding, we construct a set of candidate token sequence outputs and select the final outcome through the comparison of the maxTRR values.

3.2.1. The Backbone Network Selection

In this study, GPT-2 Chinese version [39] is used as the backbone network for further fine-tuning the DRG task. The reasons for using the GPT-2 model are manifold. Above all, this model holds promise in validating the effectiveness of the proposed framework, encompassing both the objective function and the DCS decoding method for the DRG task, all while accommodating our limited computing resources. Secondly, compared

to some other accessible models [23,40], GPT-2 was released earlier, and its pre-trained model is readily available and user-friendly. It should be noted that some other advanced models, such as GPT-4 [41], are powerful, while these models are not open-sourced, and using GPT-4 turbo token limit entails considerable expenses. Based on BERT tokenizer, the GPT-2 model can be re-trained for general language models and also support large training corpus. The pretrained model was downloaded from github (https://github.com/Morizeyao/GPT2-Chinese, accessed on 19 March 2024).

3.2.2. The CLpCE Objective Function

As an objective function, CE is widely used in the optimization procedure of text generation. For a text input x containing m sentences with length n, assuming the corresponding distribution is y and the predicted distribution is $\widehat{y}$, the CE loss is calculated as in Equation (6).

$$L_{CE} = \frac{1}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m}\{-y_{i,j} \times \log(\widehat{y_{i,j}}) - (1-y_{i,j}) \times \log(1-\widehat{y_{i,j}})\} \tag{6}$$

As to the same input as in Equation (6), the objective function of CL of text x can be calculated as in Equation (7), and notably, the parameters are defined the same as those in Equation (4).

$$L_{CL} = -\sum_{i=1}^{n}\log\left(\frac{e^{cos(h_i,h_i^+)/\tau}}{\sum_{j=1}^{N} e^{cos(h_i,h_i^+)/\tau}}\right) \tag{7}$$

Inspired by CL [37] and SimCSE [38], CLpCE is designed for guiding the fine-tuning process of GPT-2. The optimization goal of CLpCE can be defined as in Equation (8), where parameter $\beta \in [0,1]$ is used to adjust the proportion of the loss functions. It should be mentioned that when $\beta = 0$ and $\beta = 1$, the objective function CLpCE degenerates into CE and CL, respectively.

$$L_{CLpCE} = (1-\beta) \times L_{CE} + \beta \times L_{CL} \tag{8}$$

Figure 2 shows the model fine-tuning procedure. It consists of CE-based supervised learning and CL-based unsupervised learning parts, both of which are weighted by β in the CLpCE objective function.

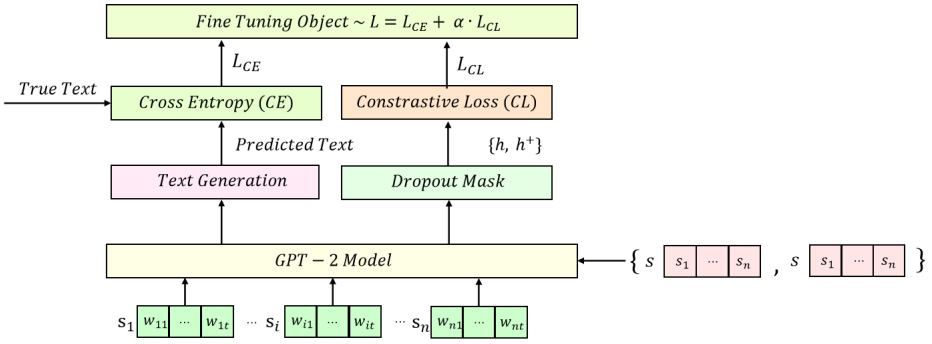

Figure 2. The CLpCE-based model fine-tuning procedure. L_{CE} guides the supervised learning and L_{CL} directs the unsupervised learning, both parts contributing to the fine-tuning of pre-trained LLMs for accurate feature representation towards a specific task.

3.2.3. The DCS Decoding

Essentially, CS is a GS decoding method with an additional degradation penalty term. When handling long texts, GS is prone to getting stuck in local optimal and generating duplicate tokens [32]. To overcome this anisotropy problem, a penalty term is added. It measures the similarity between the current candidate token and the previous tokens.

However, CS considers only the word with the highest probability at the current time, and the generated text lacks diversity.

The DCS decoding enriches the diversity, and meanwhile, it ensures the probability difference to the best token acceptable. Given the previously generated text $x_{<t}$, the output of x_t at time t via DCS can be described as the token generation,

$$\begin{aligned} x_t^l = \{&(1-\psi) \times p_\theta(v|x_{<t}) \\ &- \psi \times (\max\{cos(h_v, h_{x_j}) : 1 \le j \le t-1\})\}. \end{aligned} \tag{9}$$

To enhance the text generation diversity, DCS decoding is designed which uses tokens with the highest probabilities to form a candidate set. Firstly, the token x_t^* with the highest probability ($p_{max} = p_{x_t^*}$) at the current time is added into the candidate set. Then, the probabilities (p) of the remaining tokens are compared to the highest probability. If the difference of the probabilities between tokens is less than threshold ρ, the token is added to the candidate set as well (Equation (10)).

$$x_t^m \in \{(p_{max} - p_{x_t^m}) \le \rho \times p_{max}\} \tag{10}$$

After that, selection of the candidate tokens will yield different outputs of token sequences as $\{seq_l\}_{l=1}^k$ and $seq_l = \{x_{<t}, x_t^l\}$ for text generation. In the end, among the generated outputs of token sequences ($\{seq_l\}_{l=1}^k$), the final outcome is determined by the maxTRR values as

$$out = \min\{maxTRR(seq_l)\}_{l=1}^k. \tag{11}$$

In Equations (9)–(11), "max" and "min" denote the operation of maximization and minimization, respectively. Since the parameter ρ dictates the quality of token generation, its value should be carefully defined. The metric maxTRR is defined in Equation (12), and it quantifies the token diversity in a candidate output of text summarization. In any case, when the token with p_{max} is selected, DCS is degenerated into the CS decoding strategy.

3.3. Experiment Design

Extensive experiments are conducted to validate the effectiveness of the proposed CLpCEwDCS decoding framework. In each experiment, the dataset is shuffled and randomly divided into a training set and a testing set with an 8:2 ratio for model building and validation.

Specifically, the effectiveness of the objective function CLpCE is validated with different β values ($\{0.0, 0.1, \cdots, 0.9, 1.0\}$), and different methods of DCS (ours), CS [38], GS [30], NS [31] and top-k search (TkS) [42] are used for decoding. The general trend and evaluation metric values are presented.

In addition, the diversity of the DCS decoding method is explored by using different control threshold ρ values. The generation accuracy, token candidate diversity, and visual perception of the output examples are illustrated.

3.4. Evaluation Metrics

Four kinds of evaluation metrics are used to quantify the text generation quality from various perspectives. The first metric is bilingual evaluation understudy (BLEU) [43], which is commonly used in machine translation evaluation. It measures the word overlap between generated and reference translations based on n-gram matching and fragment accuracy evaluation. This study involves BLEU-1, BLEU-2, BLEU-3, and BLEU-4, and higher scores indicate better text matching.

The second one is the evaluation of translation with explicit ordering (METEOR) [44]. It obtains the final score by exact word matching and semantic similarity at the word level via weighted fusion. A higher value reveals better word matching and semantic similarity.

The third one is recall-oriented understudy for gisting evaluation (ROUGE) [45]. It calculates the score based on the length of the longest common subsequence. A higher metric score denotes the generated summary more similar in content to the reference summary.

The fourth one is consensus-based image description evaluation (CIDER) [46]. It considers many factors such as consistency, semantic relevance and n-gram similarity comprehensively. A higher score shows better consistency and greater semantic similarity between the generated description and the reference description.

Besides, a supplementary metric (maxTRR) is implemented in this study for evaluating the token diversity. It is defined as the maximum repetition ratio of the tokens, and a lower value indicates higher representation diversity in text generation. Assuming s tokens are generated in a candidate output (seq_l), the j^{th} token T_j appears t_{T_j} times, and the maxTRR can be formulated as

$$maxTRR(seq_l) = \frac{\max\{(t_{T_j})\}_{j=1}^{s}}{\sum_{j=1}^{s}(t_{T_j})}, \tag{12}$$

in which the denominator represents the total number of all s tokens, and the numerator is the maximum number of times a token appears.

3.5. Implementation Details and Parameter Settings

The algorithms are implemented with python (version 3.10), pytorch (version 2.0.0 + cu118) and Transformers (version 4.28.1). The codes are deployed on a 64bit Win10 system (Intel(R) Core(TM) i9-12900K, 3.2 GHZ, and 128 GB RAM) with a 24GB GPU card (NVIDIA GeForce RTX 3080). The codes are available online (https://github.com/NicoYuCN/nlpMIDRG, accessed on 19 March 2024).

During model fine-tuning, the parameters of batch size (32), learning rate (0.0005), iteration number (10 epochs), maximum length of input text (230), maximum length of generated text (80) and optimizer (AdamW [47]) are defined, and the other parameters are set with default values.

For the decoding methods, the weighting parameter of CS is $\alpha = 0.70$ as suggested in [19], $k = 5$ is set for TkS, the probability threshold $\rho = 0.71$ is for NS, and the other parameters are set with default values.

4. Results

This section reports the DRG accuracy and diversity achieved through various decoding methods. It also includes ablation studies examining the impact of parameter β in the CLpCE objective function (Equation (8)) on DRG accuracy and parameter ρ in the DCS decoding (Equation (10)) on diversity control. In any case, achievement of the first-tier teams on the competition is summarized.

4.1. DRG Accuracy

Table 1 presents the text summarization accuracy. To each DRG model, the highest value of each metric is in boldface. It suggests that the optimal value of β in CLpCE is 0.60 regardless of decoding methods. On the other hand, no obvious difference is found among the highest metric values from DCS and CS decoding methods, and GS decoding achieves generally higher GOUGE and CIDER values.

Table 1 indicates the superiority of the objective function CLpCE over CE or CL. When using DCS for decoding, CLpCE improves the report generation performance with ≈ 0.03 increases on BLEU and METEOR, 0.015 on ROUGE and 0.09 on CIDER metrics when $\beta = 0.6$. This phenomenon can also be found when using other decoding methods.

Figure 3 shows the general trend of DRG accuracy when using different weighting values ($\beta \in \{0.0, \cdots, 1.0\}$) and decoding methods (DCS, CS, GS, NS, and TkS). From the perspective of β values, compared to $\beta = 0.0$, the other β values lead to a slight increase (≤ 0.03) on metric values, except for $\beta = 1.0$. From the perspective of decoding methods, TkS and NS cause inferior results, and CIDER values are less than 1.10 and 1.35, respectively.

The other decoding methods obtain slightly better performance, and the CIDER value from GS decoding is correspondingly higher.

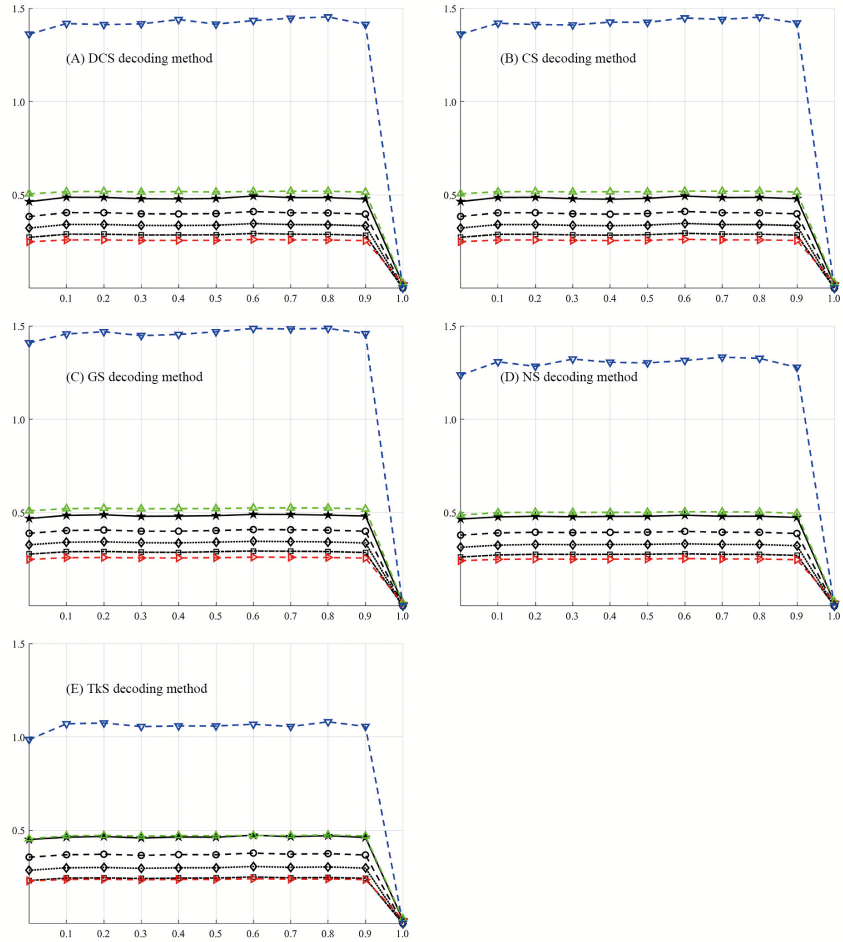

Figure 3. The effect of different β values and decoding methods on DRG text summarization. In the plot, the horizontal axis denotes the β values in the CLpCE objective function, and the vertical axis presents the values of evaluation metrics. Specifically, combinations of different types of lines, markers and colors are used for identifying different metric values of a DRG model (BLEU-1, solid black line with ★; BLEU-2, dashed black line with ○; BLEU-3, dotted black line with ♢; BLEU-4, dash-dotted black line with □; METEOR, dashed red line with ▷; ROUGE, dashed green line with △; and CIDER, dashed blue line with ▽).

Table 1. Evaluation of CLpCE guided DRG models by using different decoding methods. The values highlighted in bold represent the highest scores for each metric, while the β value underlined indicates the objective function.

	β	BLEU				METEOR	ROUGE	CIDER
		BLEU-1	BLEU-2	BLEU-3	BLEU-4			
DCS	0.0 (CE)	0.4638	0.3838	0.3223	0.2724	0.2487	0.5057	1.3607
	0.1	0.4870	0.4049	0.3414	0.2893	0.2585	0.5170	1.4179
	0.2	0.4864	0.4047	0.3414	0.2893	0.2586	0.5186	1.4118
	0.3	0.4794	0.3987	0.3363	0.2853	0.2561	0.5154	1.4162
	0.4	0.4784	0.3979	0.3358	0.2851	0.2560	0.5179	1.4391
	0.5	0.4805	0.3997	0.3372	0.2860	0.2564	0.5154	1.4147
	0.6 (CLpCE)	**0.4937**	**0.4107**	**0.3461**	**0.2933**	**0.2612**	0.5182	1.4339
	0.7	0.4855	0.4040	0.3409	0.2894	0.2586	**0.5199**	1.4459
	0.8	0.4854	0.4033	0.3400	0.2884	0.2582	0.5195	**1.4533**
	0.9	0.4780	0.3968	0.3342	0.2834	0.2549	0.5147	1.4132
	1.0 (CL)	0.0232	0.0013	0.0002	0.0000	0.0264	0.0284	0.0002

Table 1. *Cont.*

	β	BLEU				METEOR	ROUGE	CIDER
		BLEU-1	BLEU-2	BLEU-3	BLEU-4			
CS	0.0 (CE)	0.4645	0.3843	0.3227	0.2727	0.2491	0.5059	1.3611
	0.1	0.4858	0.4039	0.3406	0.2887	0.2579	0.5166	1.4196
	0.2	0.4866	0.4045	0.3410	0.2890	0.2585	0.5178	1.4125
	0.3	0.4793	0.3987	0.3364	0.2856	0.2562	0.5159	1.4101
	0.4	0.4767	0.3965	0.3346	0.2841	0.2552	0.5169	1.4255
	0.5	0.4810	0.4003	0.3378	0.2867	0.2568	0.5162	1.4240
	0.6 (CLpCE)	**0.4939**	**0.4112**	**0.3470**	**0.2943**	**0.2616**	**0.5198**	1.4477
	0.7	0.4856	0.4042	0.3410	0.2894	0.2587	0.5196	1.4395
	0.8	0.4864	0.4043	0.3408	0.2892	0.2586	**0.5198**	**1.4525**
	0.9	0.4798	0.3984	0.3355	0.2845	0.2558	0.5156	1.4208
	1.0 (CL)	0.0233	0.0012	0.0000	0.0000	0.0266	0.0286	0.0002
GS	0.0 (CE)	0.4681	0.3887	0.3274	0.2773	0.2489	0.5095	1.4090
	0.1	0.4846	0.4036	0.3410	0.2898	0.2580	0.5210	1.4567
	0.2	0.4881	0.4063	0.3431	0.2914	0.2592	0.5231	1.4684
	0.3	0.4796	0.3999	0.3381	0.2875	0.2568	0.5199	1.4477
	0.4	0.4809	0.4002	0.3376	0.2865	0.2567	0.5214	1.4542
	0.5	0.4834	0.4034	0.3413	0.2904	0.2580	0.5213	1.4682
	0.6 (CLpCE)	**0.4901**	**0.4088**	**0.3458**	**0.2941**	**0.2611**	0.5246	1.4861
	0.7	0.4894	0.4077	0.3443	0.2925	0.2602	**0.5247**	1.4835
	0.8	0.4865	0.4053	0.3424	0.2910	0.2591	0.5244	**1.4864**
	0.9	0.4812	0.3998	0.3370	0.2860	0.2559	0.5186	1.4583
	1.0 (CL)	0.0122	0.0009	0.0000	0.0000	0.0126	0.0169	0.0000
NS	0.0 (CE)	0.4654	0.3790	0.3136	0.2616	0.2422	0.4859	1.2368
	0.1	0.4765	0.3907	0.3254	0.2728	0.2492	0.4996	1.3073
	0.2	0.4800	0.3944	0.3290	0.2763	0.2511	0.5017	1.2831
	0.3	0.4775	0.3925	0.3278	0.2757	0.2501	0.5009	1.3221
	0.4	0.4793	0.3939	0.3285	0.2759	0.2504	0.5010	1.3049
	0.5	0.4798	0.3944	0.3292	0.2766	0.2512	0.5017	1.3014
	0.6 (CLpCE)	**0.4858**	**0.3991**	**0.3326**	**0.2789**	**0.2535**	**0.5044**	1.3143
	0.7	0.4799	0.3942	0.3288	0.2758	0.2511	0.5033	**1.3322**
	0.8	0.4803	0.3942	0.3286	0.2758	0.2511	0.5029	1.3259
	0.9	0.4737	0.3878	0.3226	0.2703	0.2473	0.4961	1.2776
	1.0 (CL)	0.0184	0.0007	0.0000	0.0000	0.0216	0.0226	0.0003
TkS	0.0 (CE)	0.4499	0.3554	0.2852	0.2304	0.2283	0.4542	0.9854
	0.1	0.4627	0.3686	0.2986	0.2436	0.2360	0.4701	1.0701
	0.2	0.4664	0.3712	0.3004	0.2447	0.2371	0.4718	1.0741
	0.3	0.4582	0.3651	0.2956	0.2410	0.2342	0.4681	1.0553
	0.4	0.4638	0.3695	0.2988	0.2434	0.2361	0.4710	1.0584
	0.5	0.4624	0.3687	0.2987	0.2437	0.2359	0.4700	1.0584
	0.6 (CLpCE)	**0.4730**	**0.3775**	**0.3059**	**0.2496**	**0.2402**	0.4715	1.0676
	0.7	0.4654	0.3713	0.3008	0.2449	0.2376	0.4712	1.0555
	0.8	0.4702	0.3745	0.3032	0.2470	0.2389	**0.4743**	**1.0807**
	0.9	0.4613	0.3672	0.2969	0.2421	0.2352	0.4689	1.0557
	1.0 (CL)	0.0206	0.0016	0.0000	0.0000	0.0225	0.0266	0.0002

4.2. DRG Diversity

Table 2 shows the representation diversity of text summarization using the CLpCE objective function ($\beta = 0.6$). It reveals that the proposed DCS decoding method achieves the lowest maxTRR value (0.12 ± 0.09), followed by CS and GS decoding methods. On the other hand, the maxTRR values of all the decoding methods indicates that more than 6 out of 50 generated tokens are the same, which cause unnaturalness or undesirable repetitions in text generation.

Table 2. Representation diversity of text summarization.

	DCS	CS	GS	NS	TkS
maxTRR	0.12 ± 0.09	0.22 ± 0.13	0.24 ± 0.15	0.27 ± 0.13	0.29 ± 0.16

To enhance the understanding of DRG diversity, two cases with CS and DCS decoding are shown in Table 3 for perception. The token with the maximum repeating times is underlined, and the maxTRR is shown at the end of the output (CS) or the candidate output (DCS). Case A is a relatively short desensitized data input, and text summarization seems good because of low token repetition ratio. CS decoding generates 11 tokens, and no tokens are the same. DCS decoding yields four candidates, while the fourth candidate has three identical tokens out of the twelve tokens (maxTRR, 25%). Case B is much longer. The CS decoding method yields seven tokens, and a token ("190") appears four times, and thus, maxTRR = 4/7. On the other hand, all four candidates from the DCS decoding method show much lower repetition token ratios, and the third and fourth output contains up to 30 tokens. Therefore, DCS decoding could provide more choices of text summarization output to balance both DRG accuracy and representation diversity for improved naturalness of diagnositic report generation.

Table 3. Perception of DRG text summarization for diversity analysis. The token underlined shows the token with the maxTRR value.

	Desensitized Data Description
case A input	14 108 30 13 20 18 23 21 10 14 32 16 39 27 47 51 31 29 20 18 10 24 42 26 37 61 24 10 40 13 45 163 45 39 159 49 50 204 37 21 157 155 10
CS output	150 50 107 104 113 110 15 13 31 29 20 (maxTRR, 1/11)
DCS output	(1) 150 50 107 66 17 81 76 33 81 10 (maxTRR, 1/10) (2) 150 50 107 80 33 17 13 31 81 60 49 29 (maxTRR, 1/12) (3) 150 50 107 80 33 17 81 76 33 31 81 60 49 29 (maxTRR, 1/14) (4) 150 50 65 107 <u>29</u> 113 15 29 20 60 49 29 (maxTRR, 3/12)
case B input	83 12 38 41 17 1074 96 17 552 48 17 27 131 17 89 65 69 70 11 149 58 51 36 82 11 34 38 41 17 40 153 44 23 21 25 11 263 256 567 28 59 11 199 54 894 141 126 231 11 45 83 207 281 240 353 300 212 491 302 237 297 300 212 11 113 110 104 259 207 281 315 286 258 280 11 22 12 96 16 35 12 38 41 17 178 58 36 82 10 22 279 33 91 72 78 11 33 24 122 61 24 10 22 12 62 33 628 51 171 82 11 33 686 170 1119 11 22 12 119 17 143 175 105 744 26 37 72 78 11 22 12 38 41 17 210 143 170 179 10
CS output	<u>190</u> 57 190 190 190 79 10 (maxTRR, 4/7)
DCS output	(1) <u>49</u> 75 100 344 282 11 57 49 77 75 100 57 92 10 (maxTRR, 2/14) (2) <u>49</u> 75 100 344 282 49 57 49 77 75 100 57 92 10 (maxTRR, 3/14) (3) <u>49</u> 369 142 49 180 372 11 369 372 11 180 372 11 440 439 139 420 11 117 175 13 29 440 439 11 202 191 200 487 365 175 98 10 (maxTRR, 2/33) (4) <u>49</u> 369 142 49 180 372 11 369 372 11 180 372 11 440 439 139 420 11 117 487 384 440 439 11 202 191 175 98 278 10 (maxTRR, 2/30)

4.3. The Effect of the Diversity Control

Table 4 shows the effect of the control threshold ρ on the diverse text generation. Given the CLpCEwDCS decoding framework ($\beta = 0.60$), it is found that the evaluation metric values have no obvious difference when the ρ value increases, which indicates that the DCS decoding maintains the token generation quality along with ρ increase.

Table 4. DRG accuracy of DCS decoding by using different control threshold values.

ρ	**BLEU**				**METEOR**	**ROUGE**	**CIDER**
	BLEU-1	**BLEU-2**	**BLEU-3**	**BLEU-4**			
0.00	0.4939	0.4112	0.3470	0.2943	0.2616	0.5198	1.4477
0.01	0.4939	0.4111	0.3470	0.2942	0.2612	0.5190	1.4459
0.05	0.4939	0.4113	0.3466	0.2940	0.2613	0.5188	1.4445
0.10	0.4937	0.4107	0.3461	0.2933	0.2612	0.5182	1.4339

Figure 4 shows the average candidate numbers in fifty experiments. The dotted red line with ◊ shows $\rho = 0.01$, and the dashed blue line with ∘ indicates $\rho = 0.10$. It is found that more candidate outputs of text summarization are generated when the control threshold ρ values increase. When $\rho = 0.10$, the number of candidate outputs might be

larger than 1.4 that is potential to maintain text generation coherence and decrease token repetition ratio in DRG text summarization.

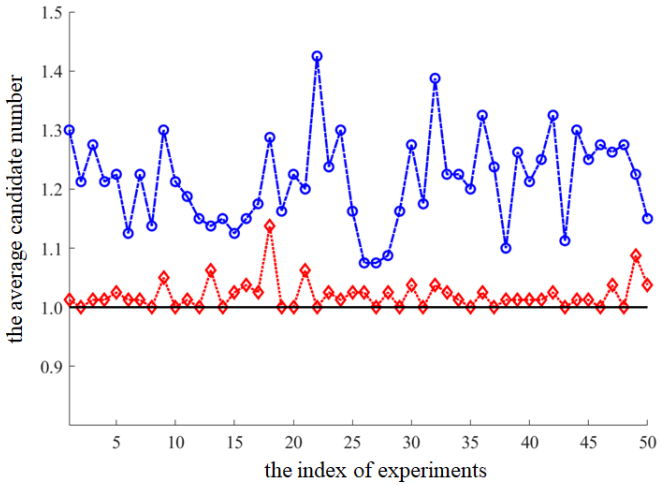

Figure 4. The effect of control threshold ρ on the text generation diversity ($\rho = 0.01$, dotted red line with $\Diamond$; $\rho = 0.10$, dashed blue line with $\circ$).

4.4. Achievement of the First-Tier Teams on the Competition

According to the report of the "Medical Imaging Diagnosis Report Generation" competition, achieved results from the first-tier teams are shown in Table 5. All the teams explore the tricks of exponential moving average of weights, fast gradient method and regularized dropout [48] for improved robustness and accuracy. Team B, C and D additionally use stochastic weight averaging [49] and label smoothing [50], and team E further integrates extract loss and sentence shuffle in the fine-tuning stage.

Table 5. Current achievement of the first-tier teams on the competition.

Team	Main Procedure in Diagnosis Report Generation	Score
A	CPT-base + noise-aware similarity bucketing + fine-tuning	2.327
B	BART-large + GBPQ + fine-tuning	2.297
C	(CPT-base + BART-base) + RAG + fine-tuning	2.285
D	BART-large + fine-tuning	2.272
E	BART-large + fine-tuning	2.263
F	BART-large + fine-tuning	2.249
ours	GPT2-Chinese + fine-tuning + CLpCEwDCS decoding	2.135

Based on the metric scores provided by the competition track, minor differences are observable among the results of the first-tier participants (Table 5). It is found that the teams focus on BART [23] and/or Chinese Pre-trained unbalanced Transformer (CPT) [40], either base or large models, for the DRG task. Team A proposes the noise-aware similarity bucketing [51] and generates the text summary output with the best prompt matching, team B designs the graph beam search with priority queue (GBPQ) for speeding up the reasoning procedure, and team C utilizes the retrieval augmented generation (RAG) [52] strategy. These models outperform the proposed framework from 0.114 to 0.192 on the score values. The score comparison also suggest that our framework dedicated to improving the diversity maintains DRG accuracy and coherence well.

5. Discussion

Accurate and automatic DRG improves clinical efficiency, and fine-tuning a pre-trained LLM is indispensable for realizing this specific application task. However, anisotropy degeneration or semantic inconsistency of sentence embeddings has been massively observed

from unnatural and undesirable text generation. To address this issue, a CLpCEwDCS decoding framework is proposed and evaluated on this challenging task. In any case, a supplementary metric (maxTRR) is designed to evaluate the token diversity in text summarization, which is also important in DCS decoding.

The CLpCE improves the consistency and accuracy of the sentence embeddings. In comparison to the CE objective function, the proposed CLpCE function leads to higher DRG quality regardless of the decoding methods. It increases the vales of evaluation metrics (Figure 3) and obtains superior performance when $\beta = 0.60$ (Table 1). Specifically, when using DCS for decoding, CLpCE ($\beta = 0.60$) enhances 0.03 on BLEU-1 and BLEU-2, 0.02 on BLEU-3 and BLEU-4, 0.01 on METEOR and ROUGE, and 0.09 on CIDER over the CE objective function ($\beta = 0.00$). Notably, this phenomenon can also be observed when using other decoding methods. It indicates that CLpCE can quantitatively improve the DRG quality from fragment accuracy, word matching, semantic similarity and content consistency. The main reason is the penalty term. CL is a self-supervised representation learning method by contrasting semantically similar and dissimilar pairs of samples [25]. Its purpose is to minimize the distance of samples from same distributions and to maximize the distance of samples from different distributions. Consequently, in the sentence embedding space, intra-class tokens could be close, and inter-class tokens could be kept a long distance. Thereby, the CL penalty term benefits LLM fine-tuning and guides the procedure towards a specific application task, and in this study, it improves DRG quality.

The DCS decoding method relieves the anisotropy degeneration issue by decreasing the frequency of token repetition. It achieves competitive DRG quality with the CS and the GS decoding methods (Table 1). Most importantly, it leads to more candidate outputs of text summarization in the token space (Figure 4) and decreases token repetition ratio (Table 2) by using the minimum of the maxTRR values, while the generation cohesion and accuracy are maintained well (Table 4). Of particular concern is the proposed metric maxTRR (Equation (12)). Its value is applied to determine the final token sequence output (Equation (11)) in an automated fashion. Additionally, two case examples further reveal that DCS decoding provides more candidate outputs of text generation with lower repeats and frequent tokens (Table 3). It should be admitted that there is discrepancy between the human and model word distributions, and further training on more data could not rectify this discrepancy [26,53]. Interestingly, the DCS decoding shows the potential to decrease the discrepancy by improving the output diversity. It keeps the accuracy and coherence as the CS decoding method and outperforms other traditional methods [19]. Therefore, using a small control threshold value ($\rho = 0.10$) could keep these dissimilar tokens with the top-high probabilities and generate diverse text summarization.

According to the track report, our framework achieves state-of-the-art performance for the DRG competition (Table 5). A close look into these models reveals that BART and CPT models are preferred due to their focus on text summarization tasks. Conversely, as a general generation model, GPT-2 supports a broad spectrum of downstream applications, and a slight drop on the score value becomes understandable. Meanwhile, the first-tier teams utilize NLP tricks, including but not limited to exponential moving average of parameters, fast gradient method and regularized dropout, and these tricks contribute to the improved performance of text generation. The proposed framework stands to benefit from these techniques if they are appropriately integrated into the fine-tuning stage.

There are several limitations in the current study. On the DRG task, the proposed framework has been verified effectively relieving the anisotropy degeneration problem, and its feasibility and generalizability on other NLP applications becomes desirable. However, it definitely involves large-scale data processing and massive time cost that is beyond our budget due to limited funding and computing resources. Secondly, as a result of technological evaluation, more powerful LLMs [41,54,55] with hundreds of billions of parameters are now available, while utilizing these models requires additional expenses and heavy computing resources. The investigation into whether the proposed framework, employing advanced models, would enhance the DRG task is currently underway. Thirdly,

besides contrastive learning [37], other fine-tuning and decoding strategies, such as fidelity-enriched contrastive search [34], self-supervised learning [56], and reinforcement learning with human feedback [57], could reduce the dependency on the labeled data samples. Last but not the least, combining with other data sources, such as dialogues, images, videos and human feedback [58], could broaden the application fields of the proposed framework.

6. Conclusions

When fine-tuning pre-trained LLMs for some specific downstream application tasks, the anisotropy degeneration problem has been massively witnessed. To address this problem, the CLpCEwDSC decoding framework is implemented that promotes the objective function of CE with a CL penalty term for accurate representation of sentence embeddings and designs a DCS decoding method for improving output diversity via selecting the candidate token sequence with the minimum maxTRR value. It has been verified effective on the DRG task with five types of evaluation metrics, and further improvement of the framework could be conducted by using more advanced models, proper fine-tuning strategies, multi-modal data learning and generalizability verification.

In the field of medical imaging, there is a long way to go before a fully automated medical image report generator can be used to facilitate clinical decision making. The proposed framework, aimed at generating accurate and natural diagnostic reports from medical image descriptions, could be further enhanced by integrating more powerful LLMs and effective fine-tuning strategies. On the other hand, most attention should be directed towards addressing other challenges, such as medical image understanding, vision–language alignment, and interpretation of diagnosis reports, in order to expedite the realization of automated and precise medical imaging diagnostic report generation.

Author Contributions: Conceptualization, T.Z., J.M. and S.Y.; methodology, J.M. and Y.Y.; software, J.M. and Y.Y.; validation, J.M. and Y.Y.; formal analysis, T.Z. and S.Y.; investigation, S.Y.; resources, J.M.; data curation, J.M. and Y.Y.; writing—original draft preparation, J.M.; writing—review and editing, T.Z. and S.Y.; visualization, J.M. and Y.Y.; supervision, S.Y.; project administration, S.Y.; funding acquisition, T.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the National Key R&D Program of China (Grant No. 2023YFF0904604) and the Fundamental Research Funds for the Central Universities (Grant No. CUC23ZDTJ014).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset supporting the current study is available online at https://gaiic.caai.cn/ai2023/, accessed on 19 March 2024.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

DRG	Diagnostic report generation
LLM	Large language model
CLpCE	Contrastive learning penalized cross-entropy
DCS	Diversity contrastive search
maxTRR	Maximum of token repetition ratio
GPT	Generative pre-trained Transformer
CLpCEwDCS	CLpCE with DCS
BLEU	Bilingual evaluation understudy
METEOR	Evaluation of translation with explicit ordering
ROUGE	Recall-oriented understudy for gisting evaluation

CIDER	Consensus-based image description evaluation
NLP	Natural language processing
RNN	Recurrent neural network
LSTM	Long-short term memory
BERT	Bidirectional encoder representations from Transformers
BART	Bidirectional and autoregressive Transformer
T5	Text-to-text transfer converter
CL	Contrastive learning
CLEAR	Contrastive learning for sentence representation
TaCL	Token-aware contrastive learning
GS	Greedy search
NS	Nucleus search
CS	Contrastive search
FECS	Fidelity-enriched contrastive search
RRM	repetition reduction module
PT	Pre-training
FT	Fine-tuning
SimCSE	Simple contrastive learning of sentence embeddings
TkS	Top-k search
GBPQ	Graph beamsearch with prioirity queue
RAG	Retrival augmented generation

References

1. Kryscinski, W.; Keskar, N.S.; McCann, B.; Xiong, C.; Socher, R. Neural text summarization: A critical evaluation. *arXiv* **2019**, arXiv:1908.08960.
2. Allahyari, M.; Pouriyeh, S.; Assefi, M.; Safaei, S.; Trippe, E.D.; Gutierrez, J.B.; Kochut, K. Text summarization techniques: A brief survey. *arXiv* **2017**, arXiv:1707.02268.
3. Pang, T.; Li, P.; Zhao, L. A survey on automatic generation of medical imaging reports based on deep learning. *Biomed. Eng. Online* **2022**, *22*, 48. [CrossRef] [PubMed]
4. Chen, Z.; Varma, M.; Delbrouck, J.; Paschali, M.; Blankemeier, L.; Van Veen, D.; Valanarasu, J.; Youssef, A.; Cohen, J.; Reis, E. CheXagent: Towards a Foundation Model for Chest X-Ray Interpretation. *arXiv* **2024**, arXiv:2401.12208.
5. Jones, K.S. Automatic summarising: The state of the art. *Inf. Process. Manag.* **2007**, *43*, 1449–1481. [CrossRef]
6. Minaee, S.; Kalchbrenner, N.; Cambria, E.; Nikzad, N.; Chenaghlu, M.; Gao, J. Deep learning–based text classification: A comprehensive review. *ACM Comput. Surv.* **2021**, *54*, 1–40. [CrossRef]
7. Schuster, M.; Paliwal, K.K. Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.* **1997**, *45*, 2673–2681. [CrossRef]
8. Van Houdt, G.; Mosquera, C.; Napoles, G. A review on the long short-term memory model. *Artif. Intell. Rev.* **2020**, *53*, 5929–5955. [CrossRef]
9. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 6000–6010.
10. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
11. Paulus, R.; Xiong, C.; Socher, R. A deep reinforced model for abstractive summarization. *arXiv* **2017**, arXiv:1705.04304.
12. Chuang, Y.; Tang, R.; Jiang, X.; Hu, X. SPeC: A soft prompt-based calibration on performance variability of large language model in clinical notes summarization. *J. Biomed. Inform.* **2024**, *151*, 104606. [CrossRef] [PubMed]
13. Tian, S.; Jin, Q.; Yeganova, L.; Lai, P.; Zhu, Q.; Chen, X.; Yang, Y.; Chen, Q.; Kim, W.; Comeau, D. Opportunities and challenges for ChatGPT and large language models in biomedicine and health. *Briefings Bioinform.* **2024**, *25*, bbad493. [CrossRef] [PubMed]
14. Li, J.; Li, D.; Savarese, S.; Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv* **2023**, arXiv:2301.12597.
15. Van Veen, D.; Van Uden, C.; Blankemeier, L.; Delbrouck, J.; Aali, A.; Bluethgen, C.; Pareek, A.; Polacin, M.; Reis, E.; Seehofnerová, A. Adapted large language models can outperform medical experts in clinical text summarization. *Nat. Med.* **2024**. [CrossRef]
16. Dong, Y.; Cordonnier, J.-B.; Loukas, A. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In Proceedings of the 38th International Conference on Machine Learning, Virtual Event, 18–24 July 2021; pp. 2793–2803.
17. Ethayarajh, K. How contextual are contextualized word representations? comparing the geometry of BERT, ELMO, and GPT-2 embeddings. *arXiv* **2019**, arXiv:1909.00512.
18. Su, Y.; Liu, F.; Meng, Z.; Lan, T.; Shu, L.; Shareghi, E.; Collier, N. Tacl: Improving bert pre-training with token-aware contrastive learning. *arXiv* **2021**, arXiv:2111.04198.
19. Su, Y.; Lan, T.; Wang, Y.; Yogatama, D.; Kong, L.; Collier, N. A contrastive framework for neural text generation. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 21548–21561.

20. Li, B.; Zhou, H.; He, J.; Wang, M.; Yang, Y.; Li, L. On the sentence embeddings from pre-trained language models. *arXiv* **2020**, arXiv:2011.05864.
21. Wang, Z.; Zeng, J.; Tao, H.; Zhong, L. RBPSum: An extractive summarization approach using Bi-stream attention and position residual connection. In Proceedings of the 2023 International Joint Conference on Neural Networks (IJCNN), Gold Coast, Australia, 18–23 June 2023; pp. 1–8.
22. Abanoub, G.E.; Fawzy, A.M.; Waly, R.R.; Gomaa, W.H. Generate descriptions of medical dialogues through two-layers Transformer-based summarization. *Intell. Method Syst. Appl.* **2023**, 32–37. [CrossRef]
23. Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; Zettlemoyer, L. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv* **2019**, arXiv:1910.13461.
24. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* **2020**, *21*, 5485–5551.
25. Chuang, C.-Y.; Robinson, J.; Lin, Y.-C.; Torralba, A.; Jegelka, S. Debiased contrastive learning. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 8765–8775.
26. Welleck, S.; Kulikov, I.; Roller, S.; Dinan, E.; Cho, K.; Weston, J. Neural text generation with unlikelihood training. *arXiv* **2019**, arXiv:1908.04319.
27. Wu, Z.; Wang, S.; Gu, J.; Khabsa, M.; Sun, F.; Ma, H. CLEAR: Contrastive learning for sentence representation. *arXiv* **2020**, arXiv:2012.15466.
28. Tan, C.; Sun, X. CoLRP: A contrastive learning abstractive text summarization method with ROUGE penalty. In Proceedings of the 2023 International Joint Conference on Neural Networks (IJCNN), Gold Coast, Australia, 18–23 June 2023; pp. 1–7.
29. Mai, T.P.; Nguyen, Q.A.; Can, D.C.; Le, H.Q. Contrastive hierarchical discourse graph for vietnamese extractive multi-document summarization. In Proceedings of the 2023 International Conference on Asian Language Processing (IALP), Singapore, 18–20 November 2023; pp. 118–123.
30. Klein, G.; Kim, Y.; Deng, Y.; Senellart, J.; Rush, A. OpenNMT: Open-Source Toolkit for Neural Machine Translation. *Annu. Meet. Assoc. Comput. Linguist. Syst. Demonstr.* **2017**, *35*, 67–72.
31. Holtzman, A.; Buys, J.; Du, L.; Forbes, M.; Choi, Y. The curious case of neural text degeneration. *arXiv* **2019**, arXiv:1904.09751.
32. Fu, Z.; Lam, W.; So, A.; Shi, B. A theoretical analysis of the repetition problem in text generation. *Proc. AAAI Conf. Artif. Intell.* **2021**, *35*, 12848–12856. [CrossRef]
33. Su, Y.; Xu, J. An empirical study on contrastive search and contrastive decoding for open-ended text generation. *arXiv* **2022**, arXiv:2211.10797.
34. Chen, W.L.; Wu, C.K.; Chen, H.H.; Chen, C.C. Fidelity-enriched contrastive search: Reconciling the faithfulness-diversity trade-off in text generation. *arXiv* **2023**, arXiv:2310.14981.
35. Zhang, Y.; Kamigaito, H.; Aoki, T.; Takamura, H.; Okumura, M. Generic Mechanism for Reducing Repetitions in Encoder-Decoder Models. *J. Nat. Lang. Process.* **2023**, *30*, 401–431. [CrossRef]
36. Xu, J.; Liu, X.; Yan, J.; Cai, D.; Li, H.; Li, J. Learning to break the loop: Analyzing and mitigating repetitions for neural text generation. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 3082–3095.
37. Hadsell, R.; Chopra, S.; LeCun, Y. Dimensionality reduction by learning an invariant mapping. *IEEE Comput. Vis. Pattern Recognit.* **2006**, *2*, 1735–1742.
38. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A simple framework for contrastive learning of visual representations. *Int. Conf. Mach. Learn.* **2020**, *119*, 1597–1607.
39. Du, Z. *GPT2-Chinese: Tools for Training GPT2 Model in Chinese Language*; GitHub Repository, 2019.
40. Shao, Y.; Geng, Z.; Liu, Y.; Dai, J.; Yan, H.; Yang, F.; Zhe, L.; Bao, H.; Qiu, X. CPT: A pre-trained unbalanced transformer for both chinese language understanding and generation. *arXiv* **2021**, arXiv:2109.05729.
41. Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F.L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. GPT-4 technical report. *arXiv* **2023**, arXiv:2303.08774.
42. Fan, A.; Lewis, M.; Dauphin, Y. Hierarchical neural story generation. *arXiv* **2018**, arXiv:1805.04833.
43. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.-J. BLEU: A method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics, Philadelphia, PA, USA, 6–12 July 2002; pp. 311–318.
44. Banerjee, S.; Lavie, A. METEOR: An automatic metric for mt evaluation with improved correlation with human judgments. In *Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*; Association for Computational Linguistics: Toronto, ON, Canada, 2005; pp. 65–72.
45. Lin, C.-Y. ROUGE: A Package for Automatic Evaluation of Summaries. Text Summarization Branches Out. 2004; pp. 74–81. Available online: https://aclanthology.org/W04-1013.pdf (accessed on 19 March 2024).
46. Vedantam, R.; Lawrence Zitnick, C.; Parikh, D. Cider: Consensus-based image description evaluation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 4566–4575.
47. Loshchilov, I.; Hutter, F. Decoupled weight decay regularization. *arXiv* **2017**, arXiv:1711.05101.
48. Wu, L.; Li, J.; Wang, Y.; Meng, Q.; Qin, T.; Chen, W.; Zhang, M.; Liu, T. R-drop: Regularized dropout for neural networks. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 10890–10905.
49. Izmailov, P.; Podoprikhin, D.; Garipov, T.; Vetrov, D.; Wilson, A. Averaging weights leads to wider optima and better generalization. *arXiv* **2018**, arXiv:1803.05407.

50. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826.
51. Wu, X.; Gao, Y.; Zhang, H.; Yang, Y.; Guo, W.; Lu, J. The Solution for the CVPR2023 NICE Image Captioning Challenge. *arXiv* **2023**, arXiv:2310.06879.
52. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.; Rocktäschel, T. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 9459–9474.
53. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language models are unsupervised multitask learners. *OpenAI Blog* **2019**, *1*, 9.
54. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901.
55. Du, Z.; Qian, Y.; Liu, X.; Ding, M.; Qiu, J.; Yang, Z.; Tang, J. GLM: General language model pretraining with autoregressive blank infilling. *arXiv* **2022** arXiv:2103.10360.
56. Baevski, A.; Hsu, W.-N.; Xu, Q.; Babu, A.; Gu, J.; Auli, M. Data2vec: A general framework for self-supervised learning in speech, vision and language. In Proceedings of the 39th International Conference on Machine Learning, Baltimore, MD, USA, 17–23 July 2022; pp. 1298–1312.
57. Uc-Cetina, V.; Navarro-Guerrero, N.; Martin-Gonzalez, A.; Weber, C.; Wermter, S. Survey on reinforcement learning for language processing. *Artif. Intell. Rev.* **2023**, *56*, 1543–1575. [CrossRef]
58. Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A. Training language models to follow instructions with human feedback. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 27730–27744.