



北京航空航天大學
BEIHANG UNIVERSITY

基于 R2Gen 的 X 光片智能诊断报告生成 从复现到优化：基于医学先验的数据增强策略 项目结题汇报

小组成员：陈师尧 凌浩然 曹东润

汇报时间：2025.12.23

指导老师：刘超

目 录

CONTENTS



北京航空航天大學
BEIHANG UNIVERSITY

1 背景介绍

2 方法讲解

3 结果展示

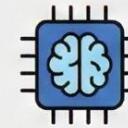
4 总结展望

任务介绍：医疗影像报告生成

研究任务与流程



输入 (Input):
放射影像 (如胸部 X-Ray)



AI 辅助系统
医疗影像报告生成



输出 (Output):
文本报告
(包含诊断发现和结论)

目标: 自动化生成诊断报告

痛点 (Pain Point: Why)



放射科医生工作量巨大 -> 容易疲劳导致误诊。迫切需要自动化辅助系统减负增效。

核心挑战



视觉-语言鸿沟
(Visual-Language Gap)

图像是像素，报告是离散符号，两者难以对齐。



长文本生成
(Long Text Generation)

相比普通看图说话，医疗报告很长，需描述多个病灶。



模式化严重
(Stereotypical Patterns)

报告中有很多套话模板，但也必须精准包含关键异常信息。



数据不平衡
(Data Imbalance)

正常样本极多，患病样本太少。

技术演进路线

1. 早期探索: CNN-RNN (Early Stage)



This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the version available on IEEE Xplore.

Learning to Read Chest X-Rays: Recurrent Neural Cascade Model for Automated Image Annotation

Hoo-Chang Shin¹, Kirk Roberts², Le Lu¹, Dina Demner-Fushman², Jianhua Yao¹, Ronald M Summers¹

¹Imaging Biomarkers and Computer-Aided Diagnosis Laboratory, Radiology and Imaging Sciences, Clinical Center

²Lister Hill National Center for Biomedical Communications, National Library of Medicine

National Institutes of Health, Bethesda, 20892-1182, USA

{hoochang.shin; le.lu; kirk.roberts; rmsj}@nih.gov; ddemner@mail.nih.gov; jyao@cc.nih.gov

Abstract

Despite the recent advances in automatically describing image contents, their applications have been mostly limited to image caption datasets containing natural images (e.g., Flickr 30k, MSCOCO). In this paper, we present a deep learning model to effectively detect diseases from an image and generate reports (i.e., concise sentences about affected organs). We employ a publicly available radiology dataset of chest x-rays and their reports, and use its image annotations to mine disease names to train convolutional neural networks (CNNs). In doing so, we train various recurrent neural networks (RNNs) to learn the larger context of disease names and their contexts. Recurrent neural networks (RNNs) are then trained to detect the contexts of a detected disease, based on the deep CNN features. Moreover, we introduce a novel approach to use the weights of the already trained RNNs to predict the context of a detected disease and to infer the joint image/text contexts for composite image labeling. Significantly improved image annotation results are demonstrated using the recurrent neural cascade model by taking the joint image/text contexts into account.



Figure 1: An example of OpenI chest x-ray image, report, and annotation.

improved performance, largely due to the introduction of image captioning [10]. Recently, there has been a deep learning trend in developing image captioning to recognize the images with a large pool of hierarchical representations. Recent work also adapt recurrent neural networks (RNNs), using the rich deep CNN features to generate captions. However, the applications of the previous studies were limited to general image datasets such as Flickr30k [23], Flickr30k [63], or MSCOCO [62] which can be generalized from ImageNet.

Likewise, there have been continuous efforts and progresses in the automatic recognition and localization of specific diseases and organs, mostly on datasets where target objects are explicitly annotated [20, 26, 35, 36, 30, 19]. Yet, learning from medical image text reports and generating annotations that describe diseases and their contexts

2487

2. 中期发展: Attention Mechanisms (Attention Stage)

On the Automatic Generation of Medical Imaging Reports

Baoyu Jing^{*}, Pengtao Xie^{*}, Eric P. Xing^{*}

^{*}Petuum Inc, USA
School of Computer Science, Carnegie Mellon University, USA
{baoyu.jing, pengtao.xie, eric.xing}@petuum.com

Abstract

Medical imaging is widely used in clinical practice for diagnosis and treatment. Report-writing can be error-prone for unexperienced physicians, and time-consuming and tedious for experienced ones. In this paper, we propose a framework to study the automatic generation of medical imaging reports. This task presents several challenges. First, a complete report contains multiple heterogeneous forms of information, including *findings* and *tags*. Second, abnormal regions in medical images are often small and scattered. Third, the reports are typically long, containing multiple sentences and cope with three challenges: (1) build a multi-task learning framework which jointly performs the prediction of tags and the generation of paragraphs; (2) propose a co-attention mechanism to locate regions and generate corresponding paragraphs for them; (3) develop a hierarchical LSTM model to generate long paragraphs. We demonstrate the effectiveness of the proposed methods on two publicly available datasets.

1 Introduction

Medical images, such as radiology and pathology images, are widely used in hospitals for the diagnosis and treatment of many diseases, such as pneumonia and pneumothorax. The reading and interpretation of medical images are usually conducted by specialized medical professionals. For example, radiologists are responsible for diagnosing and writing imaging reports. They write textual reports (Figure 1) to narrate the findings regarding each area of the body examined in the imaging study, specifically

for experienced radiologists and pathologists. Writing imaging reports is tedious and time-consuming. In nations with large population such as China, a radiologist may need to read hundreds



Figure 1: An exemplar chest x-ray report. In the *Impression* section, the radiologist provides a diagnosis. The *Findings* section lists the radiology observations regarding each area of the body examined in the imaging study. The *Tags* section lists the keywords which represent the critical information in the findings. These keywords are identified using the Medical Text Indexer (MTI).

whether each area was found to be normal, abnormal or potentially abnormal.

For less-experienced radiologists and pathologists, especially those working in the rural area where the quality of healthcare is relatively low, writing medical-imaging reports is demanding. For instance, to correctly read a chest x-ray image, the following skills are needed (Delue et al., 2011): (1) thorough knowledge of the anatomical structures of the thorax and the basic physiology of chest diseases; (2) skills of analyzing the radiograph through a fixed pattern; (3) ability of evaluating the evolution over time; (4) knowledge of clinical presentation and history; (5) knowledge of the correlation with other diagnostic results (laboratory results, electrocardiogram, and respiratory function tests).

For experienced radiologists and pathologists, writing imaging reports is tedious and time-consuming. In nations with large population such as China, a radiologist may need to read hundreds

3. Transformer 时代: Memory-driven (Project Core)

Cross-modal Memory Networks for Radiology Report Generation

Zhibong Chen^{*,†}, Yaling Shen^{*,‡}, Yan Song^{*,§}, Xiang Wan[¶]

^{*}The Chinese University of Hong Kong (Shenzhen)
[†]Shenzhen Research Institute of Big Data
[‡]{zhibongchen, yalingshen}@link.cuhk.edu.cn
[§]songyan@cuhk.edu.cn
[¶]wanxiang@sribd.cn

Abstract

Medical imaging plays a significant role in clinical practice of medical diagnosis, where the text reports of the images are essential in understanding them and facilitating later treatments. By generating the reports automatically, it is beneficial to help lighten the burden of radiologists. However, due to the lack of clinical automation, which already attracts much attention in applying artificial intelligence to medical diagnosis. Previous studies mainly follow the encoder-decoder paradigm and focus on the aspect of text generation, with few studies considering the importance of cross-modal mapping and explicitly exploit such mappings to facilitate radiology report generation. In this paper, we propose a cross-modal memory network (CMN) to overcome the encoder-decoder framework for radiology report generation, where a shared memory is designed to record the alignment between images and texts so as to facilitate the interaction and cooperation across modalities. Experimental results illustrate the effectiveness of our proposed model, where state-of-the-art performance is achieved on two widely used benchmarks, i.e., 14 X-Rays and 3D-MRI-CXR. Other analyses also prove that our model is able to better align information from radiology images and texts so as to help generating more accurate reports in terms of clinical indicators.¹

1 Introduction

Interpreting radiology images (e.g., chest X-ray) and writing diagnostic reports are essential operations in clinical practice and normally requires considerable manual workload. Therefore, radiology report generation, which aims to automatically generate a free-text description based on a radiograph, is highly desired to ease the burden of radiologists.

¹Corresponding author.

[†]Our code and the best performing models are released at

<https://github.com/shjohchan/R2GenCMN>.

[‡]Along this research track, recently there is only Jing et al. (2018) studying on a multi-task learning framework with a co-attention mechanism to explicitly explore information linking particular parts in a radiograph and its corresponding report.

解决“模式与长文本”问题 (引入记忆, 生成专业长报告)

4. 大模型时代: Multimodal LLM (SOTA / Future)

nature communications

Article

Towards a holistic framework for multimodal LLM in 3D brain CT radiology report generation

<https://doi.org/10.1038/s41467-025-57426-0>

Cheng-Yi Li^{1,2,3}, Kao-Jung Chang^{1,2,3,4,5}, Cheng-Fu Yang⁶, Hsin-Yu Wu^{1,2}, Wenting Chen⁷, Hritik Bansal⁸, Ling Chen⁹, Yi-Ping Yang¹⁰, Yu-Chun Chen^{1,2}, Shih-Pin Chen^{1,2,3,4,5}, Shih-Jen Chen^{1,2}, Jing-Feng Lin¹¹, Kai-Wei Chang¹², & Shih-Hwa Chiu^{1,2,3,12}

Multi-modal large language models (LLMs) have transformed the landscape of modern healthcare, with automated radiology report (RRG) emerging as a promising application. In 2020, the RRG has been well established in the 3D CT brain CT dataset. However, it remains largely unexplored. In this regard, we curate the 3D Brain-CT dataset (10,895 text-scan pairs) and develop BrainGPT, a clinically visual instruction-tuned (CMT) model designed for 3D CT RRG. While we notice that the traditional LLM metrics failed to gauge the diagnostic quality of the RRG, we propose feature-oriented radiology task evaluation (FORTE), an evaluation scheme that captures the clinical essence of the generated reports. Here we show that BrainGPT achieves an average FORTE F1 score of 0.71 (degree = 0.66; landmark = 0.706; feature = 0.693, and impression = 0.779) and 74% of BrainGPT generated reports were indistinguishable from human-written ground truth in a *Turing-like* test. Together, our work establishes a comprehensive framework encompassing dataset curation, analysis of model metrics and the design of robust evaluation metrics for the RRG. By sharing our experience in 3D LLM-based RRG, we aim to accelerate the expedition in human-machine collaboration for next-generation healthcare.

Artificial intelligence (AI) implementation in modern healthcare has revolutionized the way we diagnose diseases, distinguish normal from abnormal, and plan treatments. Although convolutional neural networks (CNNs) have conquered some major tasks in image classification and feature segmentation, the CNN outputs are relatively context

restrictive and were less apprehensive than a fully written diagnostic report. In order to close this clinical gap, early researches on models have been established^{1–3} for chest X-ray (CXR) interpretation^{4–6}. Whereof, the primary success of LLM-based CXR report generation had fueled interdisciplinary interest to explore human-computer interfaces,

¹School of Medicine, National Yang Ming Chiao Tung University, Taipei City, Taiwan. ²Department of Medical Research, Taipei Veterans General Hospital, Taipei City, Taiwan. ³Institute of Clinical Medicine, National Yang Ming Chiao Tung University, Taipei City, Taiwan. ⁴Department of Computer Science, National Yang Ming Chiao Tung University, Taipei City, Taiwan. ⁵Department of Electrical Engineering, University of California, Los Angeles, CA, USA. ⁶Department of Electrical Engineering, City University of Hong Kong, Hong Kong, SAR, China. ⁷Institute of Hospital and Health Care Administration, National Yang Ming Chiao Tung University, Taipei City, Taiwan. ⁸Jui-Brands, Taipei Veterans General Hospital, Taipei City, Taiwan. ⁹Department of Ophthalmology, School of Medicine, National Yang Ming Chiao Tung University, Taipei City, Taiwan. ¹⁰Department of Pharmacy, National Yang Ming Chiao Tung University, Taipei City, Taiwan. ¹¹Medi-Insight, Inc., Redwood City, CA, USA. ¹²Department of Optometry, National Yang Ming Chiao Tung University, Taipei City, Taiwan. ¹³Department of Pharmacology, National Yang Ming Chiao Tung University, Taipei City, Taiwan.

Nature Communications | (2025) 16:258 |

解决“有无”问题 (能生成了, 但很短)

解决“对齐”问题 (知道看哪里, 准确率提升)

解决“理解与交互”问题 (3D思维、像人一样思考)

方法讲解：X与Y

输入定义 (X: Input)

变量名：图像张量 (Image Tensor) (X)

张量维度 (Dimensions) :

[Batch_Size, 3, 224, 224]

(RGB, Resized至224分辨率)

数据集统计 (Dataset Statistics - Input X)



A. 学习阶段 (Training)

IU X-Ray Train Set: 5,229 张图像



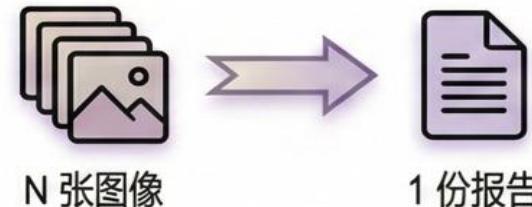
B. 评估阶段 (Evaluation) :

IU X-Ray Test Set: 1,494 张图像;

Private Set: 43 张图像 (老师提供)



多视图融合 (Multi-view Integration) :



输出定义 (Y: Output)

变量名：文本序列 (Text Sequence) (Y)

张量维度 (Dimensions) :

[Batch_Size, Seq_Len]

(Word Indices)

数据集统计 (Dataset Statistics - Output Y)



A. 学习阶段 (Training)

2,768 份报告 (对应 2,768 位患者)

(逻辑: 多张图对应一份报告)



B. 评估阶段 (Evaluation) :

IU X-Ray Test Set: 791 份报告

(Ground Truth);

Private Set: 22 份报告 (对应 22 位患者)

CMN：模型架构

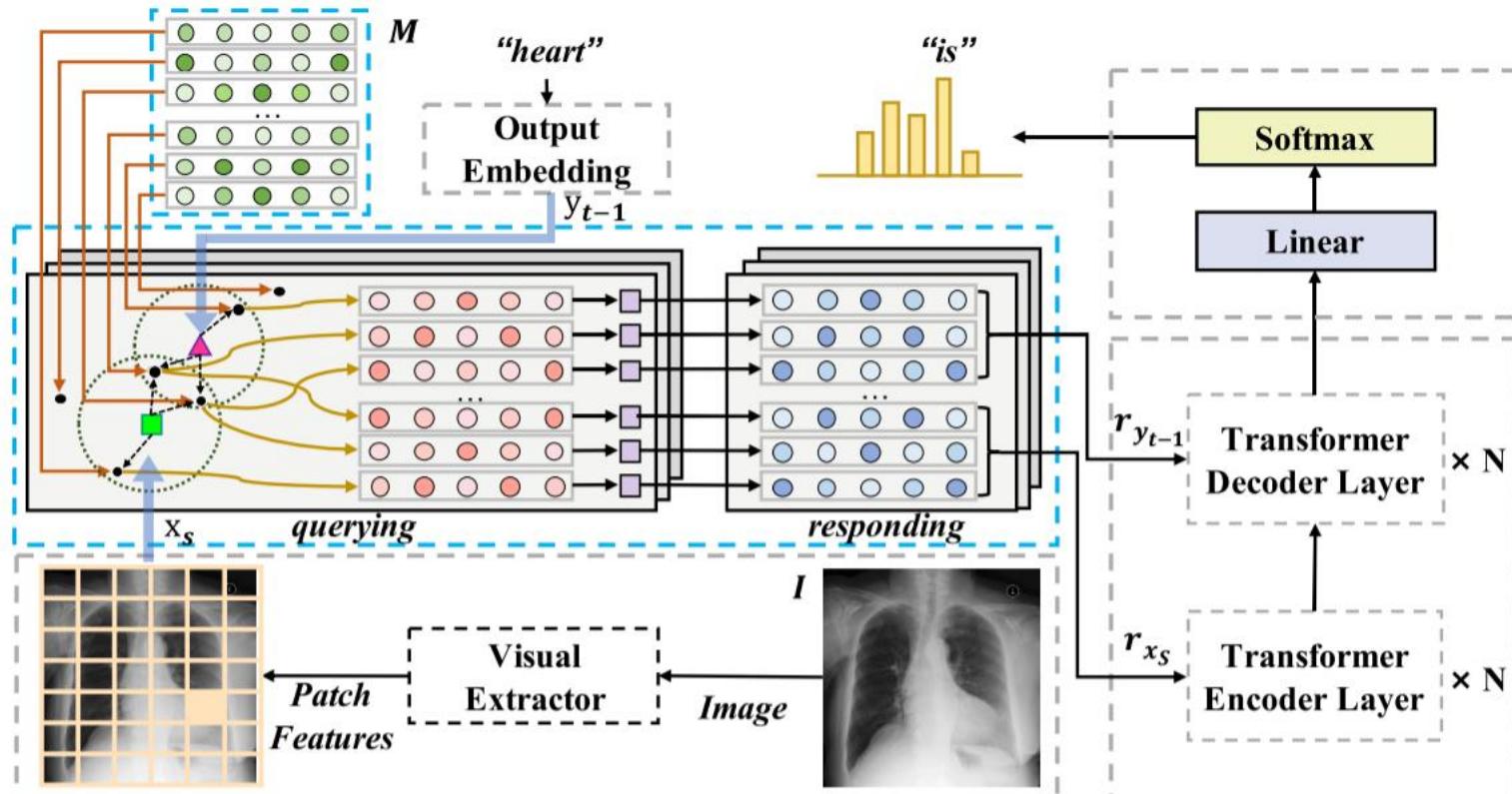
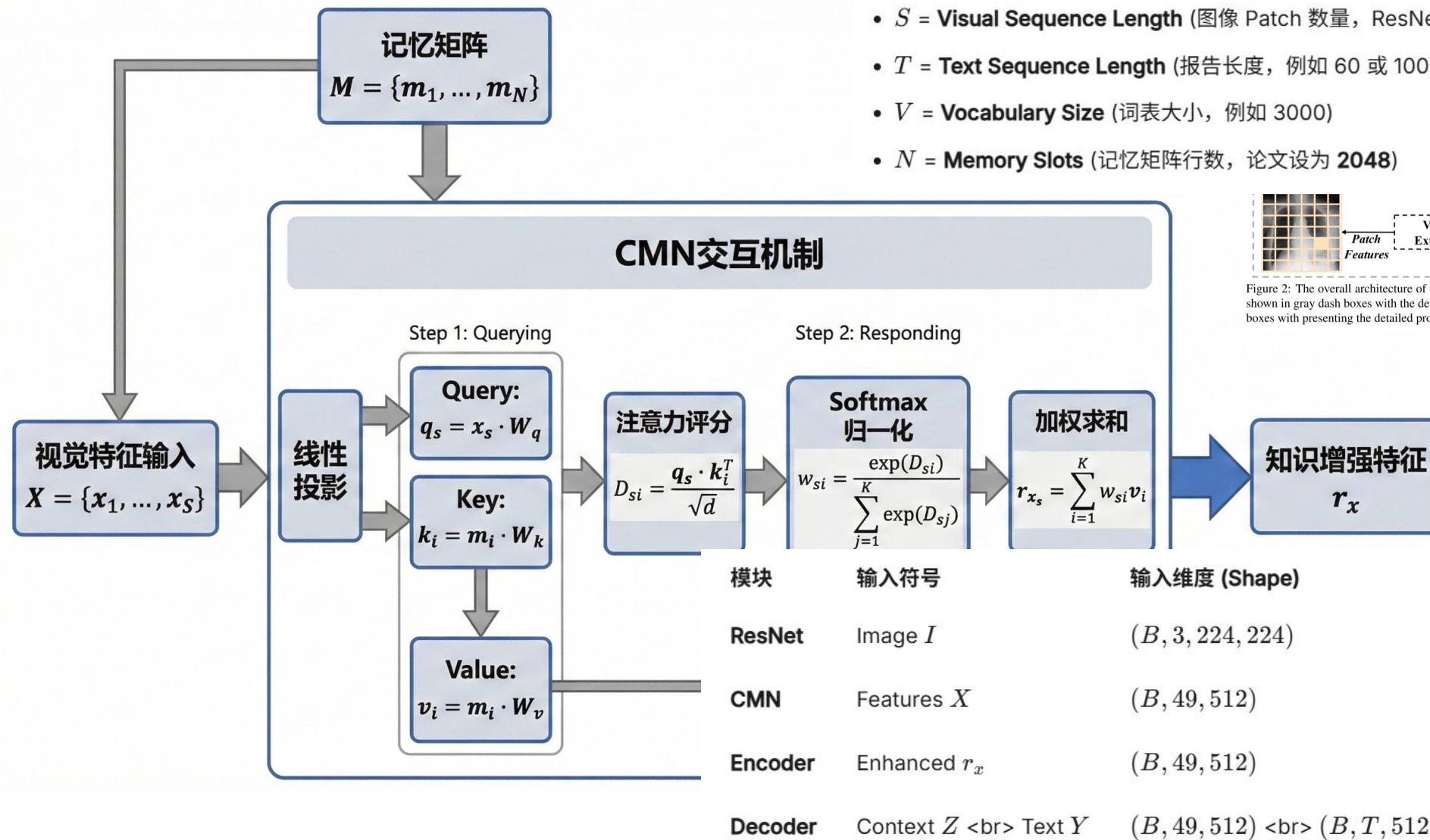


Figure 2: The overall architecture of our proposed approach, where the visual extractor, encoder and decoder are shown in gray dash boxes with the details omitted. The cross-modal memory networks are illustrated in blue dash boxes with presenting the detailed process of memory querying and responding.

1. Visual Extractor (视觉特征提取)
2. Transformer Encoder (编码 X 光图像特征)
3. Cross-modal Memory Networks (跨模态记忆网络)
 - ★ ← 本论文创新点
4. Transformer Decoder (根据记忆 + 之前的词生成下一个词)

CMN核心交互机制



- B = Batch Size (批量大小, 例如 16)
- d = Feature Dimension (特征维度, 论文设为 512)
- S = Visual Sequence Length (图像 Patch 数量, ResNet 输出 7×7 , 所以 $S = 49$)
- T = Text Sequence Length (报告长度, 例如 60 或 100)
- V = Vocabulary Size (词表大小, 例如 3000)
- N = Memory Slots (记忆矩阵行数, 论文设为 2048)

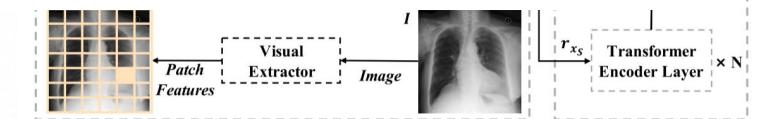
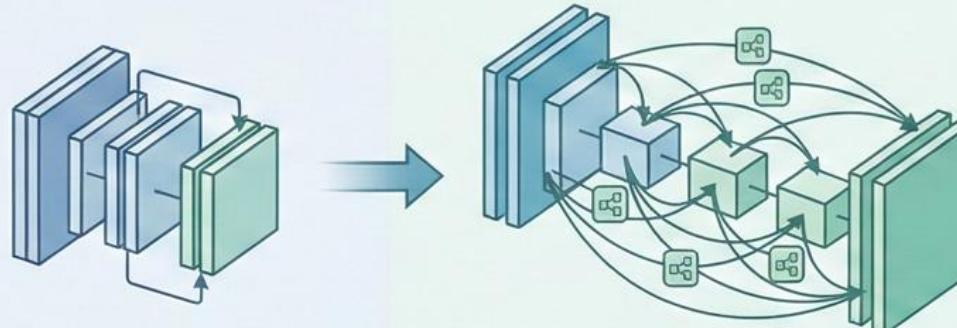


Figure 2: The overall architecture of our proposed approach, where the visual extractor, encoder and decoder are shown in gray dash boxes with the details omitted. The cross-modal memory networks are illustrated in blue dash boxes with presenting the detailed process of memory querying and responding.

阶段一：视觉骨干替换与训练策略调整

视觉提取器升级



ResNet-101

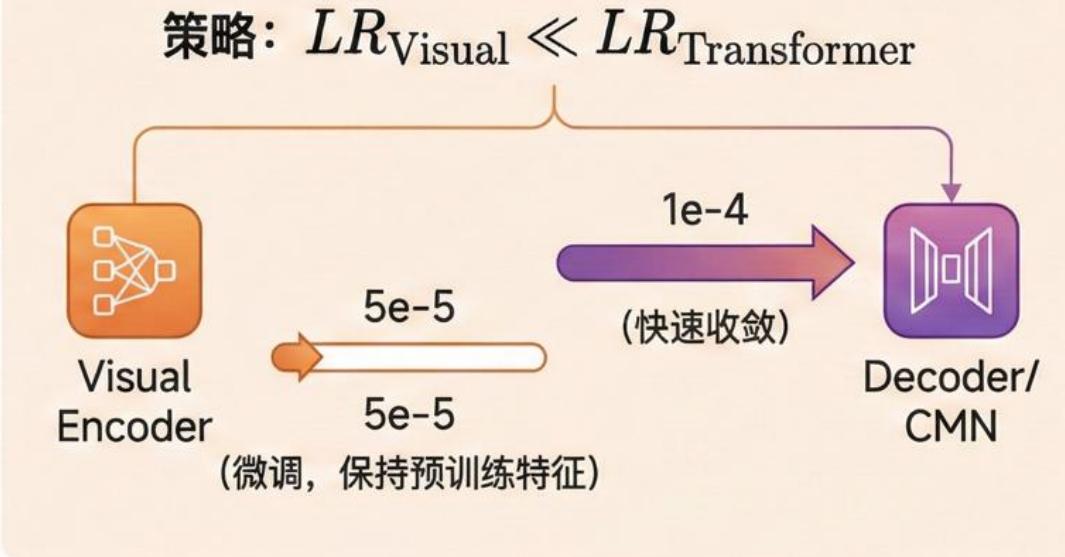
经典残差结构

DenseNet-121

密集连接，特征重用，高效

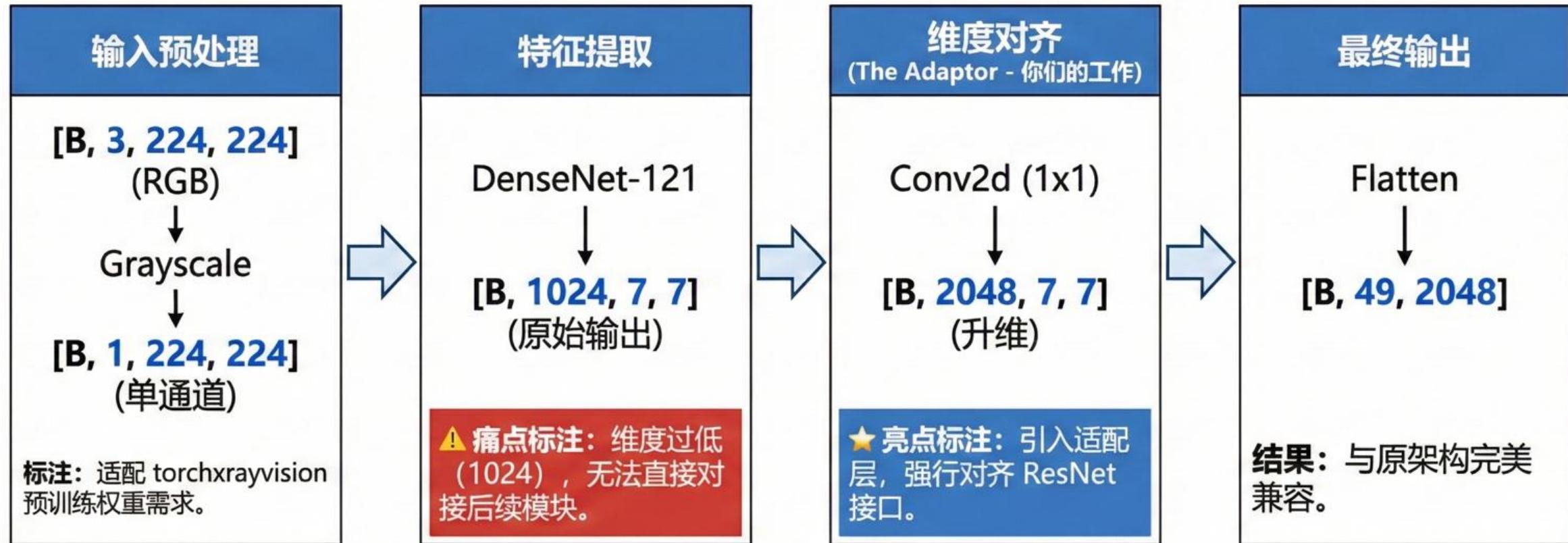
理由：DenseNet 的密集连接特性使其更擅长捕捉医学图像中细微的纹理特征（Feature Reuse），且参数利用率更高。

差异化学习率



目的：防止在训练初期破坏 CNN 提取图像特征的能力。

阶段一：视觉骨干替换与训练策略调整



阶段二：基于相似病例检索的增强生成



痛点：单纯看图说话容易产生“幻觉”，且难以生成规范的医学术语



灵感：医生在诊断疑难杂症时，往往会参考既往相似病例的报告



优势：引入了额外的先验知识，显著提升了报告的专业度和规范性

阶段二：基于相似病例检索的增强生成

输入分支 A：当前视觉特征

来自 DenseNet

Tensor: [Batch, 49, 2048]

说明：原始图片的 49 个 Patch。

输入分支 B：检索到的知识 —— [新增]

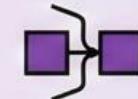
操作：Top-K 检索 → Embedding 映射

Tensor: [Batch, K, 2048]

说明：K 代表检索到的相似报告数量（或关键特征数），例如 K=5。

！**关键标注：**维度对齐：必须映射到 2048 维才能融合。

融合操作



动作：`torch.cat(dim=1)`
(沿序列长度维度拼接)

标注：Sequence
Concatenation

最终输入：

Tensor:

[Batch, 49 + K, 2048]

物理意义：Transformer 看到的“输入序列”变长了。它不仅看到了当前的图，还看到了 K 个历史参考信息。

Transformer

阶段三：引入强化学习优化指标

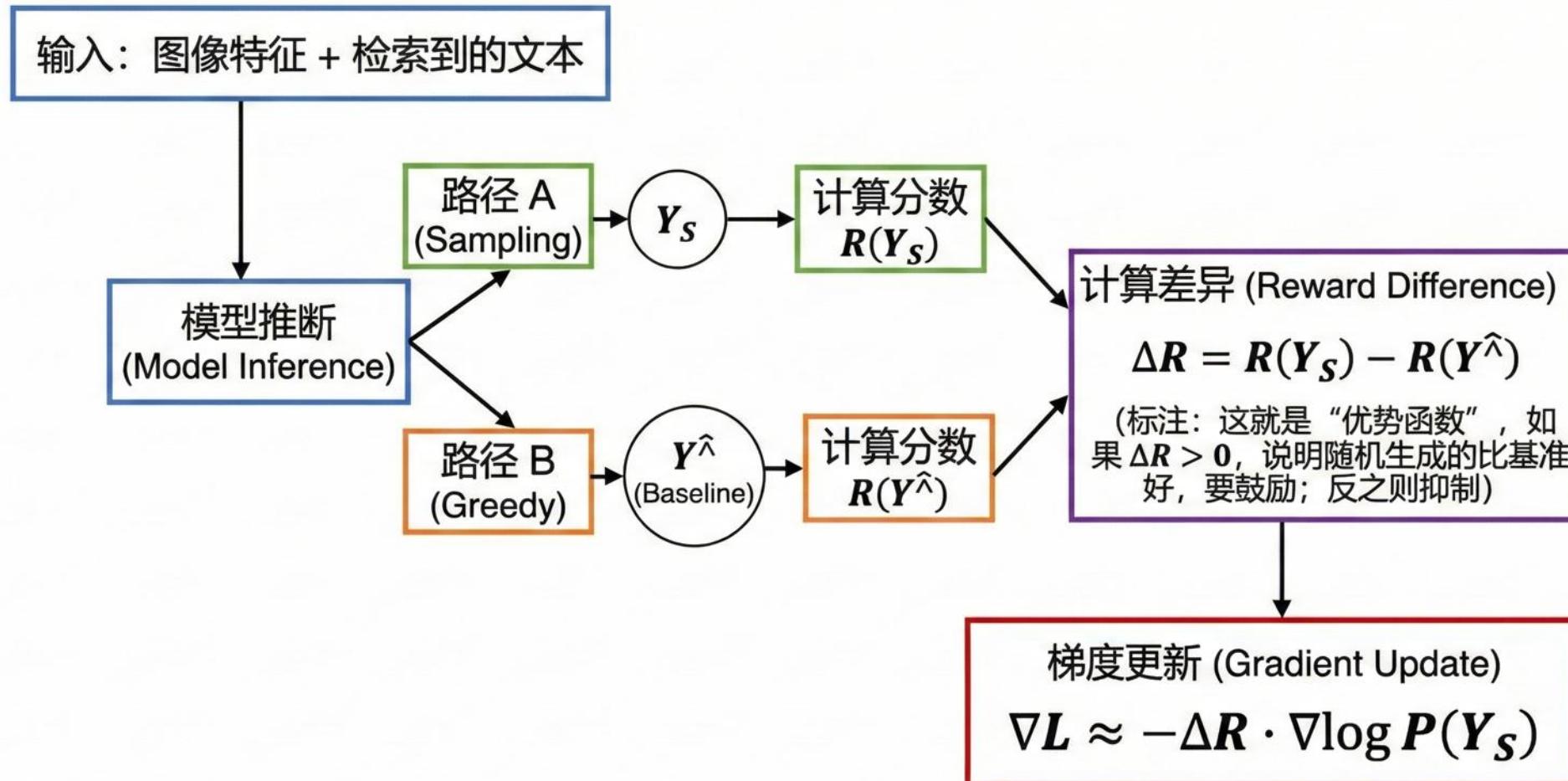
- 方法: Self-Critical Sequence Training (SCST)
- 原理:



- 实施方式: 在 Cross-Entropy训练收敛的模型基础上, 开启SCST 进行微调。

阶段三：引入强化学习优化指标

Self-Critical Sequence Training (SCST) 工作原理



阶段三：引入强化学习优化指标

特性	交叉熵损失 (Cross-Entropy)	SCST 损失 (Our Method)
优化目标	最大化“下一个词”的概率	最大化整句的评价指标 (如 CIDEr)
数值范围	$[0, +\infty)$ (概率对数)	取决于 Reward 差值 (可正可负, 非 0-1)
反向传播	链式法则 (直接求导)	REINFORCE 算法 (策略梯度)
核心优势	训练稳定, 快速收敛	直接优化测试指标, 解决暴露偏差

对原有模型的改进总结

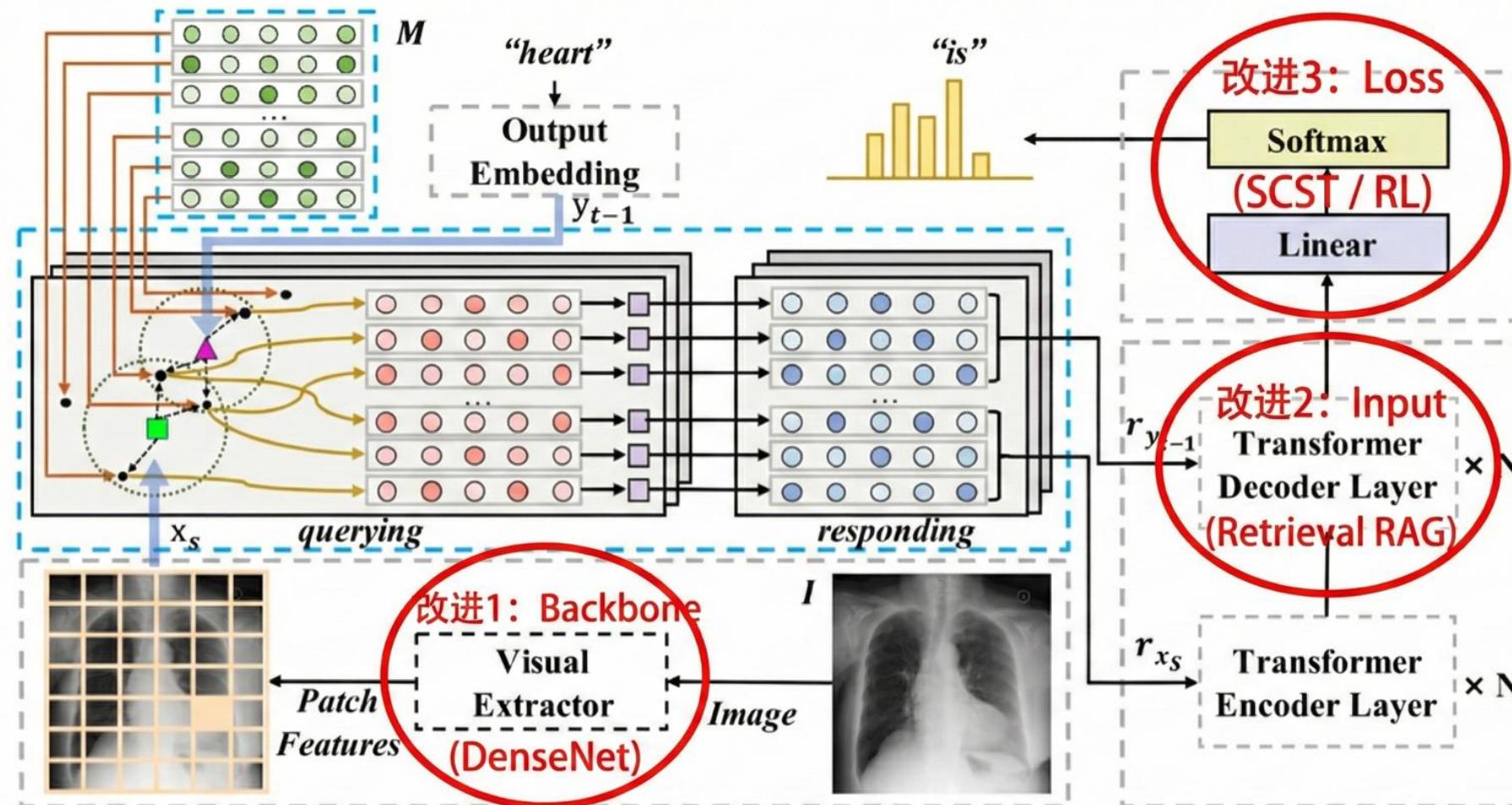


Figure 2: The overall architecture of our proposed approach, where the visual extractor, encoder and decoder are shown in gray dash boxes with the details omitted. The cross-modal memory networks are illustrated in blue dash boxes with presenting the detailed process of memory querying and responding.

■ 运行：硬件平台（算力）

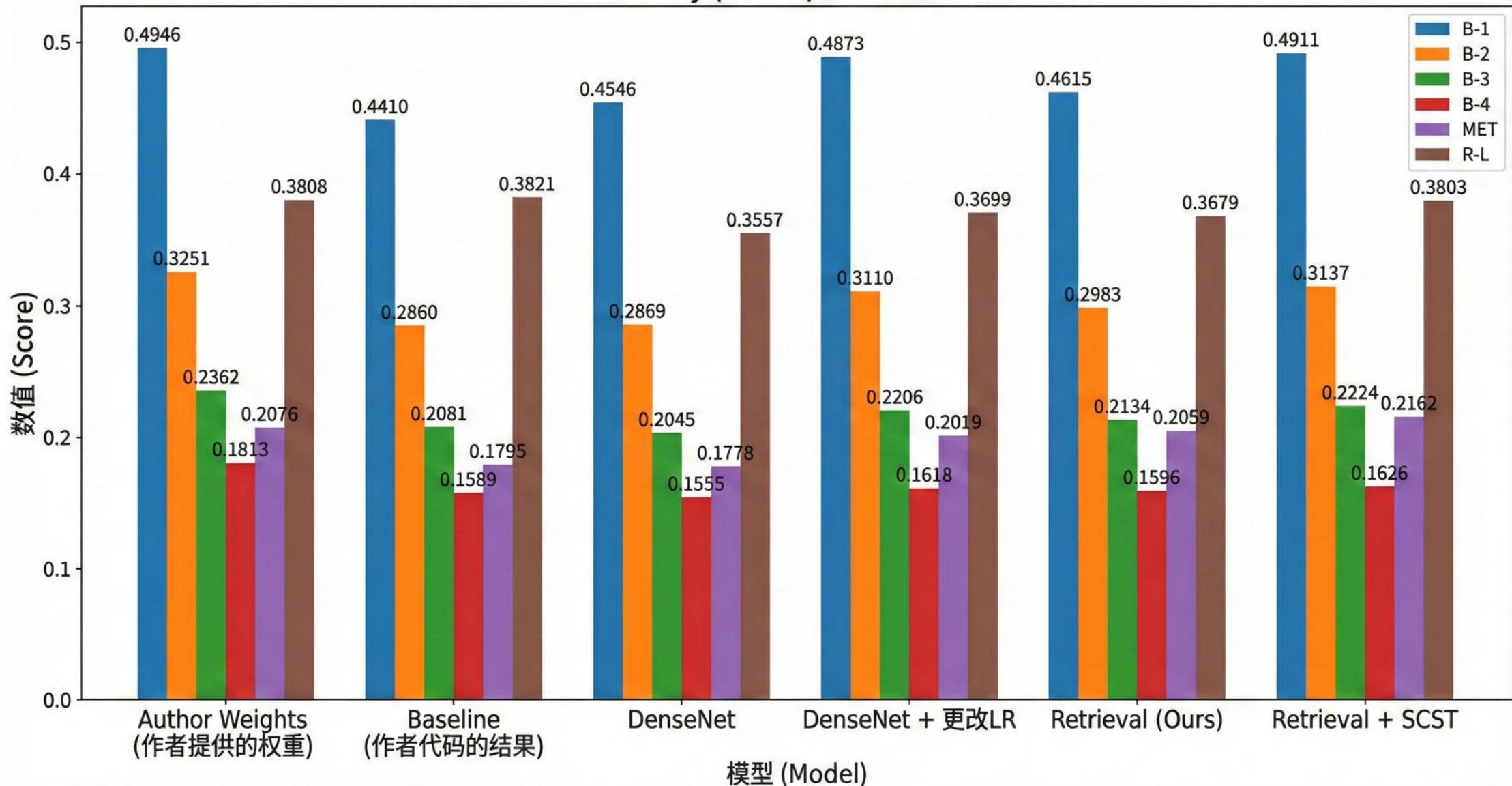
平台：AutoDL云平台 NVIDIA RTX5090

运行速度：一轮（100epoch）约1.5小时

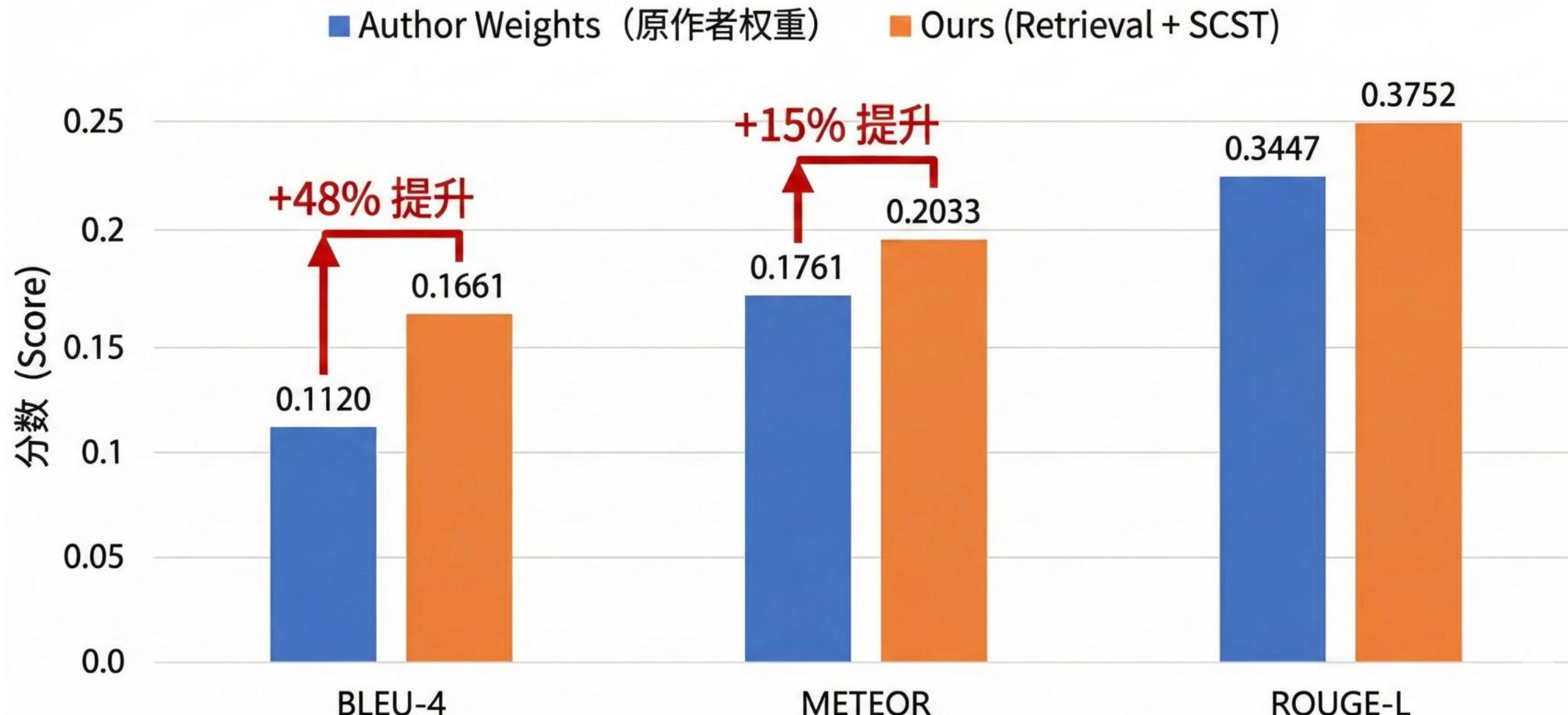
结果展示

模型 (Model)	IU X-ray (公开集)						老师私有集 (泛化能力)					
	B-1	B-2	B-3	B-4	MET	R-L	B-1	B-2	B-3	B-4	MET	R-L
Author Weights (作者提供的权重)	0.4946	0.3251	0.2362	0.1813	0.2076	0.3808	0.3974	0.2510	0.1635	0.1120	0.1761	0.3447
Baseline (作者代码的结果)	0.4410	0.2860	0.2081	0.1589	0.1795	0.3821	-	-	-	-	-	-
DenseNet	0.4546	0.2869	0.2045	0.1555	0.1778	0.3557	-	-	-	-	-	-
DenseNet + 更改LR	0.4873	0.3110	0.2206	0.1618	0.2019	0.3699	-	-	-	-	-	-
Retrieval (Ours)	0.4615	0.2983	0.2134	0.1596	0.2059	0.3679	0.4116	0.2800	0.2090	0.1661	0.1901	0.3602
Retrieval + SCST	0.4911	0.3137	0.2224	0.1626	0.2162	0.3803	0.4634	0.3076	0.2204	0.1661	0.2033	0.3752

结果展示：IU X-ray公开集



结果展示：老师的私有集



项目核心总结

关键问题	我们的实现与答案
1. X 与 Y 的定义	<ul style="list-style-type: none">• 输入 X: 图像张量, Shape = [B, 3, 224, 224]• 输出 Y: 文本序列, Shape = [B, Seq_Len]
2. 模型架构	<ul style="list-style-type: none">• 基座: ResNet-101 (Encoder) + CMN (Memory) + Transformer (Decoder)• 改进: 引入 Retrieval (检索增强) + SCST (强化学习)
3. 训练集数据	<ul style="list-style-type: none">• 来源: IU X-Ray 公开数据集• 规模: 5229 张图像 (对应2768份报告)
4. 测试集表现	<ul style="list-style-type: none">• 来源: IU X-Ray 测试集 (20% 切分)• 结果: 改进后 BLEU-4 提升 45% (0.112 -> 0.163)
5. 私有集泛化	<ul style="list-style-type: none">• 来源: 老师提供的 43 张盲测图片 (对应 22 份报告)• 结果: 成功生成流畅报告, BLEU-4 相对 Baseline 提升 48%

项目核心总结

关键问题

我们的实现与答案

6. 硬件平台

- **设备:** 单卡 NVIDIA GeForce RTX 5090 (24GB)

7. Loss 如何收敛

- **阶段一 (Warm-up):** 使用交叉熵 (CE), Loss 快速平稳下降
- **阶段二 (Fine-tuning):** 开启 SCST, Loss (负奖励) 虽有震荡, 但 CIDEr 指标持续上升并收敛

8. 对比了哪些算法

- **SOTA 参照:** 原作者提供的预训练权重 (Author Weights)
- **内部基线:** 复现的原始代码 (Baseline)
- **消融对象:** 替换骨干网络的变体 (DenseNet)

局限性与反思

局限性

- **细微病灶遗漏：** 虽然句式像医生了，但在面对极细微特征（如“主动脉轻微扭曲”）时，模型仍存在漏诊（Recall 问题）。
- **评价指标的偏差：** SCST 虽然刷高了 BLEU/CIDEr 分数，但高分不完全等于“临床正确”。有时候模型为了凑分数，可能会生成一些“万金油”式的废话。
- **依赖训练集分布：** 检索增强非常依赖数据库。如果训练集里根本没有某种罕见病，检索机制也无能为力。

未来方向

- **从 2D 到 3D：** 尝试处理 CT 数据（如 BrainGPT）。
- **引入 LLM：** 利用大语言模型（如 LLaMA-Med）强大的推理能力，修正“幻觉”，不仅仅是模仿，而是进行推理。



北京航空航天大學
BEIHANG UNIVERSITY

Thanks for listening!

请老师批评指正