FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 14, Issue 2 | Pages 375--380 | e-ISSN: 2347-470X

CXR-CapsNet: CNN-RNN-RL based Caption Generation Model

Satendra Singh Bhadoriya[1*], **Palak Keshwani**[2], **and K. Nagaiah**[3]

[1,2]*Department of Computer Science and Engineering, Faculty of Engineering and Technology, The ICFAI University Raipur, CG, India; Email: satendrab.phd2024@iuraipur.edu.in*[1]*; palakkeshwani@iuraipur.edu .in*[2]

[3]*Department of Electronics and Communication Engineering, Faculty of Engineering and Technology, The ICFAI University Raipur, CG, India; Email: nagaiah.k@iuraipur.edu.in*

***Correspondence:** satendrab.phd2024@iuraipur.edu.in; Tel.: +91-9039634657

ABSTRACT—Automated reporting of chest radiographs is an emerging task at the intersection of medical image analysis and natural language generation. In this problem, a model receives a chest X-ray and produces a clinically meaningful textual description, including the presence or absence of respiratory diseases. Conventional systems rely on an encoder–decoder pipeline in which a convolutional neural network (CNN) encodes the image and a recurrent neural network (RNN) decodes the representation into a report word by word. Recent work has shown that reinforcement learning can further align generated reports with sequence-level objectives.

To overcome this limitation, the proposed CXR-CapsNet model adapts deep reinforcement learning with embedded rewards to the domain of chest radiographic report generation by modifying the underlying CNN for medical imaging and by introducing an enhanced reward mechanism based on clinical and linguistic evaluation metrics. The policy network provides local guidance for predicting the next token in the report, while the value network provides global guidance over possible continuations of the partially generated sequence. A reward network, built on visual–semantic embeddings, combines similarity in the joint image–text space with a linear combination of metrics such as BLEU, CIDEr, ROUGE, and METEOR to better reflect the quality of the report. The policy and value networks are first pre-trained separately, with the reward network trained independently. In a benchmark chest radiographic data set, the proposed CXR-CapsNet achieves performance comparable to or better than strong baselines, while offering a substantial improvement over previous reinforcement-learning-based approaches for the generation of medical reports.

Keywords: Chest X-ray, Respiratory disease, Medical report generation, Reinforcement learning, CNN, RNN, Visual–semantic embedding, Policy network, Value network, Reward network.

ARTICLE INFORMATION

Author(s): Satendra Singh Bhadoriya, Palak Keshwani, and K. Nagaiah
Received: 29/03/26; **Accepted:** 10/05/26; **Published:** 25/06/26
E-ISSN: 2347-470X
Paper Id: IJEER2904B07
Citation: 10.37391/ijeer.140214
Webpage-link:
https://ijeer.forexjournal.co.in/archive/volume-14/ijeer-140214.html
Publisher's Note: FOREX Publication stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1.INTRODUCTION

In recent years, automated interpretation of medical images has become a critical research focus, particularly for high-volume modalities such as chest radiography. Chest X-rays are widely used to screen for and monitor respiratory diseases, and radiology reports remain the primary channel for communicating imaging findings to clinical decision-making. Automatically generating coherent, clinically meaningful reports from chest X-rays can support radiologists by reducing workload, standardizing descriptions, and improving accessibility of expertise in resource-constrained settings.

To address this challenge, researchers have explored methods that integrate advances in deep learning [1], reinforcement learning, and sequence modeling. Convolutional Neural Networks (CNNs) [2] are commonly used to extract high-level visual features from chest radiographs, capturing patterns associated with conditions such as pneumonia, effusion, consolidation, and cardiomegaly. These features provide a compact representation of the underlying pathology and global anatomy.

Recurrent Neural Networks (RNNs) [3] and their variants have been employed to model the sequential nature of radiology reports, enabling the system to maintain contextual information across tokens and to generate descriptive findings sections. By exploiting temporal dependencies in language, RNN-based decoders can produce reports that are syntactically coherent and clinically structured.

However, early encoder–decoder approaches based solely on CNN and RNN architectures often struggle to capture subtle radiographic signs, global consistency of the report, and clinically relevant emphasis. These limitations have motivated the integration of reinforcement learning techniques [4] into report generation frameworks [5]. In this setting, the model is viewed as an agent that learns to produce reports by maximizing a task-specific reward rather than only optimizing next-token likelihood.

A key advancement within this paradigm is the actor–critic approach, which employs a policy network, a value network, and a reward network to guide report generation. The policy network provides local guidance by determining the next word conditioned on previously generated tokens and the image representation. The value network estimates a global score for partially generated reports, encouraging sequences that lead to higher long-term reward. The reward network assigns a task-oriented score to image–report pairs, facilitating the training of both policy and value networks. This combination has shown promise in improving sequence-level quality in captioning settings.

The CXR-CapsNet proposes an enhanced framework for generating chest radiograph reports that combines CNN-based feature extraction, RNN-based decoding, reinforcement learning for sequential decision-making, and an actor–critic strategy. The objective is to better capture clinically important details and long-range report structure while maintaining linguistic fluency. Performance is evaluated on a standard chest radiograph corpus using standard metrics such as BLEU [6], CIDEr [7], ROUGE [8], and METEOR [9], and results are compared with strong medical and generic captioning baselines.

The key contributions of this work are as follows:

- Proposed *CXR-CapsNet*, a novel CNN–RNN–reinforcement learning–based framework for automated chest X-ray report generation that integrates medical image feature extraction with natural language generation to produce clinically meaningful reports.
- Developed an actor–critic learning strategy with an enhanced embedded reward mechanism, combining visual–semantic similarity and multiple evaluation metrics (BLEU, CIDEr, ROUGE-L, and METEOR) to provide both local and global guidance during sequential report generation.
- Validated the effectiveness of the proposed approach through extensive experiments, demonstrating competitive or improved performance over strong baseline methods in terms of word-level accuracy, report diversity, and overall report quality.

section 2 surveys related work in automatic image captioning and medical report generation, highlighting the transition from hand-crafted features to deep and reinforcement-learning-based methods. *section* 3 presents the problem formulation and details of the policy, value, and reward networks, including training and inference procedures. *section* 4 describes experimental settings, including dataset preparation, implementation details, hyperparameters, and quantitative results. Finally, *section* 5 summarizes the contributions and discusses future directions for improving respiratory-disease report generation.

2.RELATED WORK

Early work on visual description relied on hand-crafted features and traditional language models. For generic images, methods such as [10] combined low-level visual features with probabilistic language models to produce simple descriptions, but these features lacked the capacity to represent complex spatial and semantic relationships between regions and words. With the success of deep learning, CNN-based encoders and RNN-based decoders became the dominant paradigm for general image captioning, enabling richer feature representations and more fluent sentences [11, 12, 13, 14, 15, 16, 17].

Subsequent advances introduced attention mechanisms and transformer-style encoders to better focus on salient regions, leading to improved performance in [18, 19]. Despite these gains, many captioning systems still tend to produce generic, high-frequency descriptions that may miss fine-grained or rare details. This lack of diversity and specificity is particularly problematic in the medical domain, where small radiographic findings can have major clinical implications and where hallucination or omission of critical information is unacceptable.

Reinforcement learning was first popularized for visual tasks in domains such as game playing [20]. For captioning, [5] formulated sentence generation as a Markov Decision Process with an agent that sequentially selects words, optimizing rewards derived from visual–semantic embedding and evaluation metrics. Other work used variants of REINFORCE and actor–critic methods to directly optimize sequence-level metrics such as CIDEr [21, 22], or to combine multiple metrics through tailored reward functions [23, 24]. Common limitations across these approaches include limited diversity, exposure bias from teacher forcing pre-training, and difficulty in modeling long-range structure.

In the medical domain, similar challenges arise when adapting generic captioning models to report generation. Models trained only with cross-entropy often produce templated reports that overlook uncommon but clinically important patterns. Reinforcement-learning-based techniques can, in principle, encourage outputs that better align with clinical objectives, but require careful reward design and stable training strategies. The method proposed in CXR-CapsNet builds on the decision-making framework of [5] and the multi-metric optimization strategies of [23, 24], while introducing an off-policy actor–critic scheme and an embedding-based reward tailored to chest X-ray reporting.

3.PROPOSED METHODOLOGY

In this section, the problem formulation for reinforcement learning based image captioning is explained, followed by a detailed description of the networks involved. Then, the procedure for training the models, as well as the inference algorithm, is described.

The combination of the policy network with the value networks serves as the agent for the problem. The environment comprises the given image along with the current sequence of words. The set of actions is the set of all unique words that can be predicted, i.e. the vocabulary. The goal is to produce a correct image description for image I. A state is a representation of the environment at a time t. Reward is the feedback from the environment for the reinforcement learning.

Policy Network. The policy defines a conditional distribution over actions given states, represented as $q_\pi(b_t \mid r_t)$, where the state r_t comprises the visual feature F and the sequence of previously generated tokens $\{a_1, a_2, \ldots, a_t\}$, and the action b_t corresponds to the next token a_{t+1}.

The policy network is modeled using a convolutional and recurrent architecture, denoted as CNN_q and RNN_q, shown in *figure* 1. The visual feature F is encoded through $CNN_q(F)$ and projected into the initial input representation i_0 via $M_{i,v}CNN_q(F)$. The recurrent unit RNN_q maintains a hidden state z_t, which evolves over time through the transformation $RNN_q(z_{t-1}, i_t)$. For time steps $t > 0$, the input i_t is obtained via $\phi(a_{t-1})$. The hidden state z_t is mapped through $\varphi(z_t)$ to produce the policy distribution $\pi(b_t \mid r_t)$.

Reward Network. The reward network defines the optimization objective by assigning a scalar reward r to generated captions. It consists of CNN_r, RNN_r, and a linear mapping layer $lml(\cdot)$. The visual feature is represented as $k = CNN_r(F)$, while the sentence representation is given by the final hidden state $h'_T(C)$ of RNN_r.

The network is trained using a bidirectional ranking loss BRL_r, incorporating margin-based constraints involving γ, similarity terms $lml(k) \cdot h'_T(C)$, and negative samples C^- and k^-, ensuring correct image–caption pairs achieve higher similarity than incorrect ones.

For a predicted caption $\hat{C}$ and image feature k^*, the reward component r_1 is defined through the normalized similarity between $lml(k^*)$ and $h'_T(\hat{C})$. A second component r_2 aggregates evaluation metrics including $BLEU_1$, $BLEU_2$, $BLEU_3$, $BLEU_4$, and $ROUGE\text{-}L$, weighted by hyperparameters $e_1, e_2, \ldots, e_5$. The final reward r is obtained by combining r_1 and r_2 through averaging.

Value Network. The value network approximates the expected return associated with a given state under the policy. Its architecture is shown in *figure* 2. It consists of CNN_u, RNN_u, and MLP_u. The visual feature $CNN_u(F)$ and the encoded partial sequence $\{a_1, a_2, \ldots, a_t\}$ produced by RNN_u are combined using MLP_u to estimate the expected reward for state r_t.

Training. Training is conducted in two phases. In the first phase, the reward network is optimized using BRL_r, the policy network is trained using cross-entropy loss:

$$\mathscr{L}_{CE} = -\sum_{t=1}^{T} \log q_\pi(b_t \mid r_t), \tag{1}$$

and the value network is trained using mean squared error:

$$\mathscr{L}_{MSE} = \left\| u_\theta(r_i) - o \right\|^2, \tag{2}$$

where o denotes the observed reward and r_i represents a sampled state.

In the second phase, an actor–critic framework is applied along with curriculum learning. The policy q_π is progressively refined using sequences of increasing length and complexity, enabling efficient exploration of the action space.

Inference. Beam search with beam size B is employed using both the policy network $q_\pi(b_t \mid r_t)$ and the value network. Starting from an initial token and visual feature F, candidate sequences are generated and evaluated. At each step, the top B candidates are retained based on their policy probabilities and value estimates. This continues until an end token is produced or a maximum sequence length is reached, resulting in accurate and contextually coherent captions.

4.RESULTS AND DISCUSSION

This section describes the details of the experiment and results to evaluate the performance of the model. It covers data preparation, network implementation, hyper-parameters chosen during training, and quantitative results.

4.1 Data Preparation

The National Institutes of Health made the Open-I Chest X-ray Dataset [25] available to the public for use in medical imaging research. It contains 7,000 chest X-ray images and structured radiology reports in XML format, which include findings and impressions.

Dataset images were passed through the InceptionResNetV2 pre-trained model using the Keras Applications module. A final feature vector of size 1536 is obtained after applying a global average pooling layer. The entire feature vector of all images is stored for use at multiple stages of training. Start and end tokens are added to each caption. The frequency of each word is calculated and the top 1004 words are kept; the rest are replaced by an unknown token. This reduces the total number of words considered for prediction. Words are then encoded as integers, with a decoding data structure retained for further reference.

4.2 Evaluation Metrics

The performance of the proposed model is evaluated using standard captioning metrics, including BLEU-1 to BLEU-4, ROUGE-L, METEOR, and CIDEr. A comprehensive comparison with existing state-of-the-art methods is presented in *table* 1. The proposed CXR-CapsNet achieves scores of 0.794 for BLEU-1, 0.639 for BLEU-2, 0.516 for BLEU-3, 0.381 for BLEU-4, 0.576 for ROUGE-L, 0.291 for METEOR, and 1.279 for CIDEr.

4.3 Quantitative Evaluation

From a quantitative perspective, the model demonstrates measurable improvements over past techniques, with gains of 0.25% in BLEU-1, 0.1% in BLEU-2, 2.7% in BLEU-3, and 0.3% in CIDEr recorded in *table* 1. These improvements indicate that the proposed approach is more effective in generating relevant and contextually appropriate words, particularly for shorter n-gram sequences. However, a slight decrease is observed in higher-order and structure-sensitive metrics, including a 2.8% reduction in BLEU-4, 2.2% in ROUGE-L, and 2% in METEOR.

4.4 Qualitative Evaluation

From a qualitative standpoint, the increase in BLEU-1, BLEU-2, and BLEU-3 suggests that the generated captions contain a higher proportion of correct and meaningful individual words and short phrases, reflecting improved lexical relevance and content coverage as reported in *table* 2. In contrast, the decrease in BLEU-4, ROUGE-L, and METEOR indicates limitations in capturing longer sequential dependencies and exact

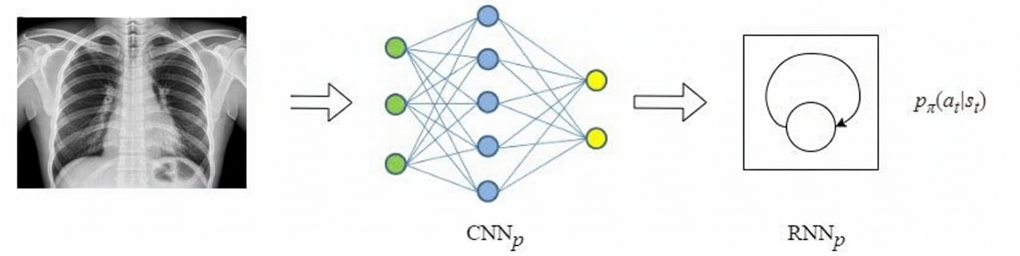

Figure 1. Diagram representing the structure of the policy network.

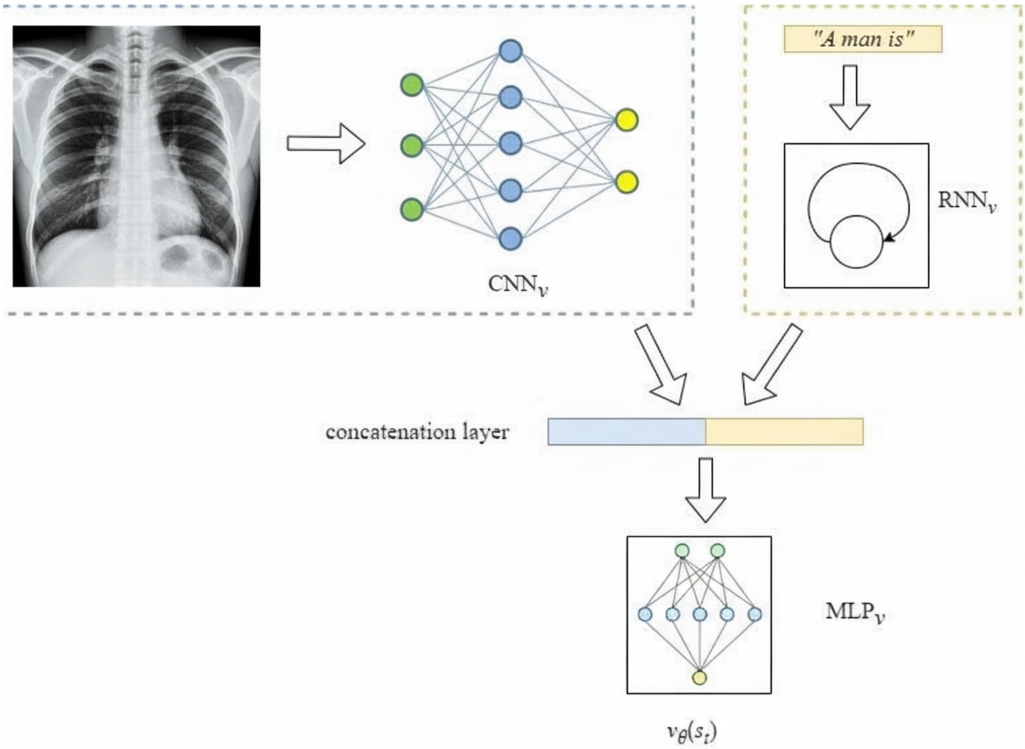

Figure 2. Diagram representing the structure of the value network.

phrase-level alignment with ground-truth captions. Since these metrics emphasize longer n-gram consistency and structural similarity, the observed decline suggests that the model may generate captions with slightly weaker grammatical coherence or reduced exact matching for longer sequences.

Nevertheless, this trade-off highlights an important characteristic of the proposed model. While strict n-gram matching is marginally reduced, the improvement in lower-order metrics and CIDEr score indicates enhanced diversity and semantic richness in the generated captions. The model demonstrates a stronger ability to capture key visual concepts and produce varied descriptions, rather than relying on rigid replication of reference sentences. Consequently, the proposed approach achieves a balance between accuracy and diversity, desirable for real-world caption generation tasks where multiple valid descriptions may exist for a single image.

4.5 Discussion

The policy network is implemented using an embedding layer for mapping given captions in lookup table form, a linear layer for mapping input features to the hidden layer, an LSTM for the RNN decoding, and finally a linear layer to produce words from output.

The reward network is implemented using an embedding followed by an LSTM to decode the captions, a linear layer for embedding the visual information, and another linear layer for embedding the semantic information of the decoded captions.

The value network is implemented using an embedding followed by an LSTM to decode the captions, a linear layer to connect the visual and semantic parts, and finally a linear layer to produce a single output scalar.

In the first phase of training, the policy network is trained using cross-entropy loss with a batch size of 256 for 100,000 epochs starting with a learning rate of 0.0001, achieving a final average loss of 0.8. Next, the reward network is trained on a custom bidirectional ranking loss with a batch size of 128 for about 50,000 epochs starting with a learning rate of 0.00001, achieving a final average loss of 0.4.

For the value network, a custom reward function is used. The

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 14, Issue 2 | Pages 375--380 | e-ISSN: 2347-470X

Table 1. Comparison of Image Captioning Methods on MSCOCO Dataset.

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR	CIDEr
NeuralTalk2 [26]	0.625	0.450	0.321	0.230	–	0.195	0.660
gLSTM [12]	0.670	0.491	0.358	0.264	–	0.227	0.812
Show and Tell [13]	–	–	–	0.277	–	0.237	0.855
Show, Attend & Tell [27]	0.718	0.504	0.357	0.250	–	0.230	–
Decision-Making [5]	0.713	0.539	0.403	0.304	0.525	0.251	0.937
SCST [21]	–	–	–	0.319	0.543	0.255	1.063
MLPR [24]	–	0.619	0.464	0.340	–	0.266	1.109
Up-Down [18]	0.772	–	–	0.362	0.564	0.270	1.135
COS-Net [19]	0.792	0.638	0.502	0.392	0.589	0.297	1.274
CGFN+GRN [28]	0.714	0.549	0.456	0.354	0.408	0.210	1.120
Proposed CXR-CapsNet	**0.794**	**0.639**	**0.516**	**0.381**	**0.576**	**0.291**	**1.279**

Table 2. Predicted Output Compared to Ground Truth.

Ground Truth Report	Predicted Report	Similarity
No acute cardiopulmonary abnormality. Lungs are clear.	Lungs are clear with no signs of acute disease.	0.92
Mild cardiomegaly with pulmonary edema.	Enlarged heart with signs of fluid accumulation in lungs.	0.87
Right lower lobe pneumonia suspected.	Possible infection in the right lower lung region.	0.81
No pleural effusion or pneumothorax detected.	No evidence of fluid or collapsed lung.	0.95

first part of the reward is calculated using the normalised distance between the visual and semantic parts of the reward network. For the second part, an evaluation-based reward is calculated as a weighted sum: 0.175 × BLEU-1, 0.075 × BLEU-2, 0.075 × BLEU-3, 0.175 × BLEU-4, 0.2 × ROUGE-L, and 0.3 × METEOR. The value network is then trained with a batch size of 128 using mean squared error loss for about 50,000 epochs at a learning rate of 0.00001, achieving a final average loss of 0.04. In the second phase, the previously trained policy and value networks are further trained using the actor–critic algorithm [29]. Curriculum learning is also applied to reduce variance. Training starts with predicting the last few words of the sentence and gradually increases in length until the entire sentence is predicted. The log probability of the executed action is calculated using the on-policy network. The difference between the reward and the value is the advantage. Actor loss is calculated from log probabilities and the advantage; critic loss is the squared advantage. The total loss is used to train both networks.

5.CONCLUSION

Image captioning, the problem of generating a human-understandable description of an image, has proven to be a challenging task. Previous encoder–decoder frameworks achieved great performance but were limited by lack of global guidance. Reinforcement learning has emerged as an effective tool for addressing this limitation. The approach employs three networks: a policy network that provides local guidance for generating the next word, a value network that provides global guidance by considering all possible extensions of the current partially generated sentence, and a reward network that guides the training of both previous networks by assigning a score to the partially generated sentence for a particular image.

The improvements to the CNN model and reward mechanism, along with the introduction of actor–critic training and policy–value scoring beam search inference, have taken this approach to a better overall performance. The proposed CXR-CapsNet demonstrates competitive performance on standard benchmarks and offers a strong foundation for automated chest X-ray reporting.

The approach can be further improved by incorporating attention mechanisms for focusing on specific parts of the image at each step. Additionally, incorporating a Large Language Model for token generation could overcome the vocabulary limitations inherent in frequency-based word selection, offering more diverse and clinically meaningful descriptions.

REFERENCES

[1] Warren S McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5:115–133, 1943.

[2] Kunihiko Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics*, 36(4):193–202, 1980.

[3] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

International Journal of
Electrical and Electronics Research (IJEER)
Research Article | Volume 14, Issue 2 | Pages 375--380 | e-ISSN: 2347-470X

[4] P Read Montague. Reinforcement learning: an introduction, by sutton, rs and barto, ag. *Trends in cognitive sciences*, 3(9):360, 1999.

[5] Zhou Ren, Xiaoyu Wang, Ning Zhang, Xutao Lv, and Li-Jia Li. Deep reinforcement learning-based image captioning with embedding reward. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 290–298, 2017.

[6] 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 of the Association for Computational Linguistics*, pages 311–318, 2002.

[7] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 4566–4575, 2015.

[8] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[9] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.

[10] Kobus Barnard, Pinar Duygulu, David Forsyth, Nando De Freitas, David M Blei, and Michael I Jordan. Matching words and pictures. *The Journal of Machine Learning Research*, 3:1107–1135, 2003.

[11] Girish Kulkarni, Visruth Premraj, Vicente Ordonez, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. Babytalk: Understanding and generating simple image descriptions. *IEEE transactions on pattern analysis and machine intelligence*, 35(12):2891–2903, 2013.

[12] Xu Jia, Efstratios Gavves, Basura Fernando, and Tinne Tuytelaars. Guiding the long-short term memory model for image caption generation. In *Proceedings of the IEEE international conference on computer vision*, pages 2407–2415, 2015.

[13] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 3156–3164, 2015.

[14] Konisha Kar, Shivam Nishad, Jayanti Rout, Ashutosh Soni, and Surendra Kumar Nanda. Medical image captioning using cvt and distillgpt2. In *2024 Second International Conference on Advances in Information Technology (ICAIT)*, volume 1, pages 1–6. IEEE, 2024.

[15] Prateek Singh and Sudhakar Singh. Chestx-transcribe: a multimodal transformer for automated radiology report generation from chest x-rays. *Frontiers in Digital Health*, 7:1535168, 2025.

[16] Zhanyu Wang, Lei Wang, Xiu Li, and Luping Zhou. Diagnostic captioning by cooperative task interactions and sample-graph consistency. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[17] Ting Yu, Wangwen Lu, Yan Yang, Weidong Han, Qingming Huang, Jun Yu, and Ke Zhang. Adapter-enhanced hierarchical cross-modal pretraining for lightweight medical report generation. *IEEE Journal of Biomedical and Health Informatics*, 2025.

[18] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on Computer Vision and Pattern Pecognition*, pages 6077–6086, 2018.

[19] Yehao Li, Yingwei Pan, Ting Yao, and Tao Mei. Comprehending and ordering semantics for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17990–17999, 2022.

[20] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

[21] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 7008–7024, 2017.

[22] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Reinforcement learning*, pages 5–32, 1992.

[23] PR Devi, V Thrivikraman, D Kashyap, and SS Shylaja. Image captioning using reinforcement learning with bluder optimization. *Pattern Recognition and Image Analysis*, 30:607–613, 2020.

[24] Ning Xu, Hanwang Zhang, An-An Liu, Weizhi Nie, Yuting Su, Jie Nie, and Yongdong Zhang. Multi-level policy and reward-based deep reinforcement learning framework for image captioning. *IEEE Transactions on Multimedia*, 22(5):1372–1383, 2019.

[25] 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, 2016.

[26] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 3128–3137, 2015.

[27] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015.

[28] Junsan Zhang, Ming Cheng, Xiangyang Li, Xiuxuan Shen, Yuxue Liu, and Yao Wan. Generating medical report via joint probability graph reasoning. *Tsinghua Science and Technology*, 30(4):1685–1699, August 2025.

[29] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR*, 2015.