Image Caption Generator using Reinforcement Learning

Arun Madhusudhanan*, Tejaswini Dilip Deore*

Abstract— **Generating meaningful captions for an image is a challenging task in machine learning. Researchers have come up with various methods to perform this task. Most of the existing approaches are based on deep learning models. These models normally have a Convolutional neural network (CNN) as encoder and Recurrent Neural Network (RNN) as a decoder. In this project, we explore image captioning as a decision-making process and train the models using reinforcement learning. We implement the "Deep Reinforcement Learning-based Image Captioning with Embedding Reward" approach proposed by Zhou Ren et al. [1], which presents an innovative approach towards image captioning using deep reinforcement learning techniques enhanced by an embedding reward mechanism. We focus on comparing the reinforcement learning based model with the results generated by deep learning based model. We evaluate the results using metrics like BLEU [2] score, ROUGE [3] score, METEOR [4] score, and CIDEr [5] score on Flickr8k dataset. Our findings indicate that the results generated using reinforcement learning based model were slightly better than the results from deep learning based model.**

I. Introduction

Image captioning is the task of automatically generating natural language descriptions for images, and it has caught significant attention in the computer vision and natural language processing communities due to its potential applications in various domains, including assistive technologies for visually impaired individuals to give them better understanding of their surroundings, content retrieval and management to enhance user experience on social media platforms, and human-computer interaction systems.

Traditional approaches to image captioning have relied on retrieval-based methods or template-based methods. Retrieval-based methods involve retrieving and ranking pre-existing captions based on their similarity to the input image. Whereas, template-based methods generate captions by filling in slots in predefined templates with object labels and attributes. However, these approaches sometimes struggle to capture the rich semantics and relationships present in complex visual scenes and end up generating very simple or inaccurate captions.

In recent years, there have been various advancements in deep learning and reinforcement learning techniques [6], [7], which have shown promising results in addressing the challenges of image captioning. Deep learning architectures, such as encoder-decoder models with attention mechanisms, have proven effective in capturing the correlation between visual and textual data. These models use encoder to extract image features and decoder to translate these features into captions describing image content.

Reinforcement learning techniques, on the other hand, have been employed to optimize caption generation by taking advantage of reward signals that measure the quality or relevance of the generated captions.

In this project, we implement the "Deep Reinforcement Learning-based Image Captioning with Embedding Reward" approach proposed by Zhou Ren et al. [1]. This approach considers a decision-making framework for image-captioning. It utilizes a policy network and a value network for predicting the next best word of the caption being generated, unlike the deep learning approach where a RNN greedily selects the next prediction.

This approach utilizes actor-critic [8]- a reinforcement learning framework, to optimize the caption generation process by incorporating an embedding-based reward signal that measures the semantic similarity between the generated captions and ground-truth captions.

II. Related Work

A. Template-based and Retrieval-based Methods:

Traditional image captioning methods relied on template-based or retrieval-based approaches. Template-based methods involve detecting objects, attributes, and relationships within an image and generating captions by filling in predefined templates with the detected components. But these methods can produce grammatically correct captions, but they lack the flexibility and expressiveness to capture the complexity and minor differences of visual scenes. Retrieval-based methods involve retrieving and ranking pre-existing captions based on their similarity with the input image. These approaches require large databases of image-caption pairs and depend on effective similarity measures to retrieve relevant captions. However, they may fail to generate novel or context relevant captions.

B. Deep Learning-based Methods:

Encoder-decoder architectures [6] have been widely used and these models typically consist of a convolutional neural network (CNN) encoder that extracts visual features from the input image, and a recurrent neural network (RNN) decoder that generates captions based on the visual features from encoder.

Vinyals et al. [6] proposed one of the earliest CNN-RNN encoder-decoder architecture and Xu et al. [9] extended this work by incorporating an attention mechanism, allowing the decoder to focus on different regions of the image dynamically during caption generation.

*Equal contribution to the work

C. Reinforcement Learning for Image Captioning:

Ren et al. [1], whose work is the focus of our implementation, proposed the ”Deep Reinforcement Learning-based Image Captioning with Embedding Reward” approach. This method employs an actor-critic reinforcement learning framework to optimize caption generation using an embedding-based reward that measures the semantic similarity between generated and ground-truth captions. The actor component learns to generate captions based on visual features and previous words, while the critic component evaluates the quality of generated captions and provides feedback for policy improvement.

III. Project Description

Image captioning can be formulated as a decision-making process and can be represented in reinforcement learning paradigm as depicted in Figure 1. The *agent* interacts with *environment*, taking certain *actions*. In this model proposed by Ren et al. [1] the policy network p_π and value network v_θ can be viewed as the agent.

In image captioning, we can view image and its captions as environment, caption prediction as agent's actions and rewards as correctness of the captions. After each action, environment returns the image and its captions generated so far.

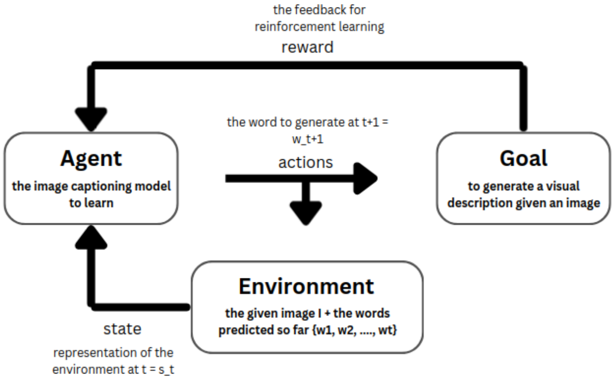

the feedback for
reinforcement learning
reward
the word to generate at t+1 =
w_t+1
actions
Agent
the image captioning model to learn
Goal
to generate a visual description given an image
Environment
the given image I + the words predicted so far {w1, w2, ..., wt}
state
representation of the environment at t = s_t

Fig. 1: Image captioning as a decision making process.

A. Models

The ”Deep Reinforcement Learning-based Image Captioning with Embedding Reward” approach proposed by Zhou Ren et al. [1] employs a reinforcement learning framework to optimize the caption generation process. Our implementation of this approach consists of three key components: a policy network, a value network, and a reward network.

The model architecture follows an encoder-decoder structure, where the encoder extracts visual features from the input image, and the decoder generates the caption based on these features.

1) Policy Network: We used a convolutional neural network (CNN) for visual feature extraction coupled with a long short-term memory (LSTM) network for language modeling. The CNN encoder is a ResNet-50 architecture pre-trained on the Flickr8k dataset. It is used to extract visual features from the input image, which serve as the initial state for the LSTM decoder.

The ResNet50 CNN (2) takes a 224x224 pixel input image and outputs a 2048-dimensional feature vector encoding the visual content of the image. These visual features are then passed as input to the LSTM network, which is responsible for generating the corresponding caption word-by-word.

In our experiments, we evaluated the CNN-LSTM model using both greedy search and beam search with k=3. The beam search approach showed improved results over greedy search across all evaluation metrics.

2) Value Network: The value network is responsible for estimating the expected reward or value of a given state-action pair in the reinforcement learning setting. It helps in assessing the quality of the generated captions so far. The value network considered consists of a CNN module, LSTM module and an MLP module. CNN module, which is a pre-trained ResNet50, will extract the visual features. LSTM module will extract the features from the current state i.e the captions generated so far. These features are combined and fed through an MLP network to produce a single value for the current state.

3) Reward Network: Since we are considering image captioning as a decision making process, we need to define a reward for learning. We can use visual semantic embedding similarities as reward.

The reward network considered consist of a CNN module and a RNN module. The CNN module is used to extract the visual embedding and RNN module is used to extract the semantic text embedding. We can then use the cosine similarity between the two embeddings to estimate the reward. High similarity will result in high reward and low similarity will result in low reward.

B. Training by reinforcement learning

First we trained policy network using supervised learning. Images and its labelled captions were fed through policy network to tune network parameters.

Then we trained reward network to predict a reward given an image and a caption associated with it. We used Cosine Embedding loss function to train the reward network.

After that we trained value network to predict the value of the current state i.e the captions generated till a given time t. The network was trained by minimising a mean square loss error between the value output of network and reward value predicted by reward network.

Once this training phase is over, we trained the policy network and value network together using deep reinforcement learning. We used actor-critic method for the training where value network will act as the critic and policy network as the actor. However the action space of the policy network, which is the length of vocabulary, is quite high. Hence we are using curriculum learning for actor-critic training.

C. Inferencing using a policy network and value network

Most common methods for predicting words during text generation are greedy search or beam search. One of the

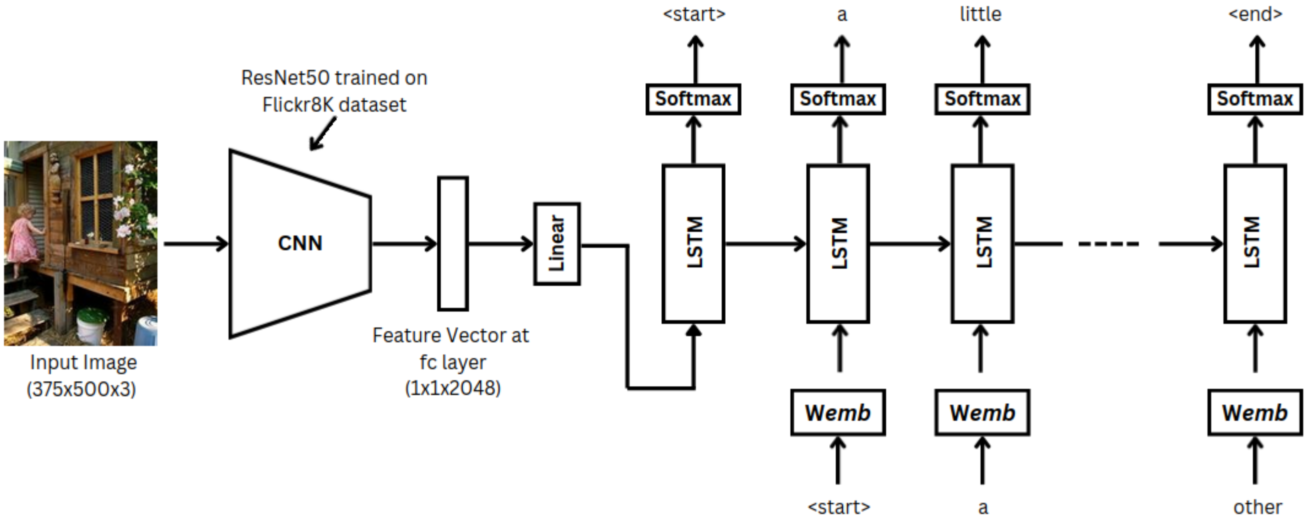


Fig. 2: Policy Network: CNN-LSTM model. This figure illustrates the architecture of an image caption generator model, which combines a ResNet50 convolutional neural network (CNN) for image feature extraction with a long short-term memory (LSTM) recurrent neural network (RNN) for generating captions. The CNN encodes the input image into fixed-size visual representations, while the LSTM decoder generates word-by-word captions based on the encoded image features.

main contribution of our reference research paper is a look ahead inference mechanism combining results from both policy network and value network. The policy network will provide a local guidance and value network will acts a global guidance during the caption generation. This mechanism is inspired by MCTS implementation in AlphaGo [10]. The lookahead inference method is formulated in equation x. We will experiment with different hyper parameter λ values and use the one which gives the best results.

We will be comparing the captions generated using greedy search, beam search and lookahead inference method using the evaluation metrics discussed in the further sections.

D. Dataset

We selected the Flickr8k dataset due to its suitability for training deep learning models for image captioning given that we had limited computational resources. This dataset contains of 8000 images and each image in is paired with 5 descriptive captions annotated by human annotators.

We used 75% of our dataset for training our model and 25% of dataset for testing.

E. Evaluation Metrics

We evaluated implemented caption generator on the Flickr8K dataset using standard image captioning metrics, including BLEU [2], ROUGE [3], METEOR [4], and CIDEr [5]. These metrics measure different aspects of the generated captions, such as n-gram precision, recall, semantic similarity, and consensus-based image description evaluation. In general, higher scores across these metrics indicate better captioning performance.

1) BLEU (Bilingual Evaluation Understudy) Score: BLEU [2] is a metric that measures the n-gram overlap between the generated caption and the reference captions. It is calculated based on the precision of n-grams in the candidate caption against the n-grams in the reference captions. We reported the BLEU-1 and BLEU-2 score, which measures the overlap of unigrams and bigrams.

2) ROUGE (Recall-Oriented Understudy for Gisting Evaluation) Score: ROUGE [3] is a set of metrics used to evaluate automatic summarization and machine translation systems. We used the ROUGE-1 variant, which calculates the F-score based on the count of unigrams and ROUGE-L variant, which calculates the longest common subsequence-based F-score between the generated caption and reference captions. It captures sentence-level structure similarity.

3) METEOR (Metric for Evaluation of Translation with Explicit Ordering) Score: METEOR [4] is a metric that goes beyond just measuring unigram precision and recall. It calculates alignment between the generated caption and references based on exact, stem, synonym, and paraphrase matches between words and phrases. It aims to incorporate more linguistic intelligence into the evaluation.

4) CIDEr (Consensus-based Image Description Evaluation) Score: CIDEr [5] computes the similarity between a generated caption and a set of reference captions by representing them as vectors of n-grams. The cosine similarity between these two vectors gives a score for the generated caption. CIDEr measures how well the generated captions match consensus in the human annotations.

IV. Experiment

A. Training Details:

.

Setting	Value
Epoch	30
Batch Size	32
Learning Rate	0.0003
Optimizer	Adam
Loss Function	Cross Entropy Loss

TABLE I: Training parameters for Policy Network

Setting	Value
Epoch	30
Batch Size	32
Learning Rate	0.0003
Optimizer	Adam
Loss Function	Cosine Embedding Loss

TABLE II: training arameters for Reward Network

Setting	Value
Epoch	30
Batch Size	32
Learning Rate	0.0003
Optimizer	Adam
Loss Function	Mean Square Error Loss

TABLE III: Training parameters for Value Network

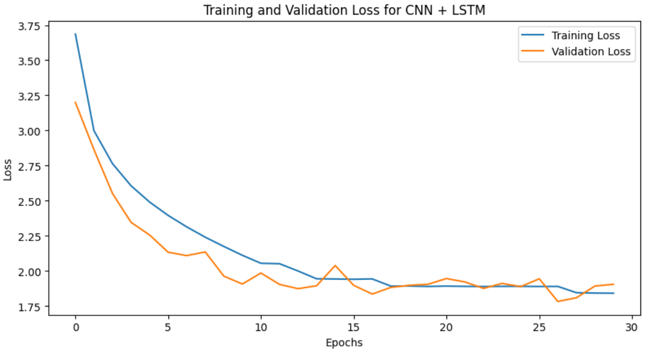


Fig. 3: Training and Validation Loss for Policy Network

V. Results

VI. Conclusion

Acknowledgement

The successful completion of this project is attributed to the exceptional guidance and support of Professor Lawson Wong, without whom this work would not have been feasible. We also want to extend our gratitude to the Teaching Assistants whose invaluable assistance and support facilitated our progress throughout the course and project.

References

[1] Z. Ren, X. Wang, N. Zhang, X. Lv, and L.-J. Li, “Deep reinforcement learning-based image captioning with embedding reward,” 2017.

[2] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “Bleu: a method for automatic evaluation of machine translation,” in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[3] C.-Y. Lin, “ROUGE: A package for automatic evaluation of summaries,” in *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81. [Online]. Available: https://aclanthology.org/W04-1013

[4] S. Banerjee and A. 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*, 2005, pp. 65–72.

[5] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, “Cider: Consensus-based image description evaluation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4566–4575.

[6] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, “Show and tell: A neural image caption generator,” 2015.

[7] L. Huang, W. Wang, J. Chen, and X.-Y. Wei, “Attention on attention for image captioning,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 4634–4643.

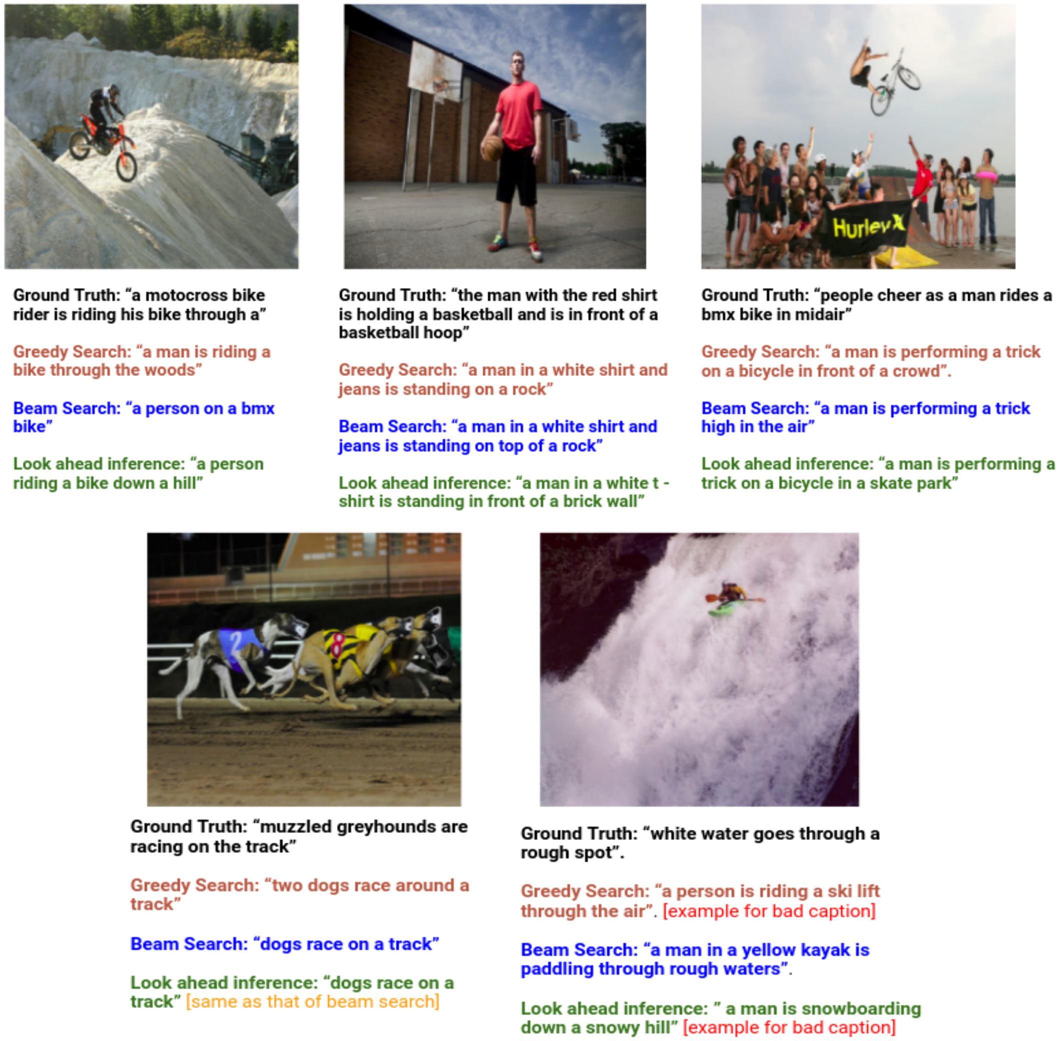
[8] V. Konda and J. Tsitsiklis, “Actor-critic algorithms,” *Advances in neural information processing systems*, vol. 12, 1999.

[9] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, “Show, attend and tell: Neural image caption generation with visual attention,” in *International conference on machine learning*. PMLR, 2015, pp. 2048–2057.

[10] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot *et al.*, “Mastering the game of go with deep neural networks and tree search,” *nature*, vol. 529, no. 7587, pp. 484–489, 2016.

Method	BLEU-1	BLEU-2	ROUGE-1	ROUGE-L	METEOR	CIDEr
Greedy Search	51.717	21.812	48.714	45.147	42.729	6.016
Beam Search (with beam size k=3)	65.399	28.446	50.121	46.809	**45.998**	8.756
Lookahead Inference	**65.753**	**29.424**	**50.428**	**47**	45.662	**8.773**

TABLE IV: Quantitative Analysis: Performance measurement of models with evaluation metrics scores. The values highlighted in bold indicate the highest score obtained for its respective metric. Lookahead Inference appears to be outperforming other approaches except for METEOR score.



Ground Truth: "a motocross bike rider is riding his bike through a"

Greedy Search: "a man is riding a bike through the woods"

Beam Search: "a person on a bmx bike"

Look ahead inference: "a person riding a bike down a hill"

Ground Truth: "the man with the red shirt is holding a basketball and is in front of a basketball hoop"

Greedy Search: "a man in a white shirt and jeans is standing on a rock"

Beam Search: "a man in a white shirt and jeans is standing on top of a rock"

Look ahead inference: "a man in a white t - shirt is standing in front of a brick wall"

Ground Truth: "people cheer as a man rides a bmx bike in midair"

Greedy Search: "a man is performing a trick on a bicycle in front of a crowd".

Beam Search: "a man is performing a trick high in the air"

Look ahead inference: "a man is performing a trick on a bicycle in a skate park"

Ground Truth: "muzzled greyhounds are racing on the track"

Greedy Search: "two dogs race around a track"

Beam Search: "dogs race on a track"

Look ahead inference: "dogs race on a track" [same as that of beam search]

Ground Truth: "white water goes through a rough spot".

Greedy Search: "a person is riding a ski lift through the air". [example for bad caption]

Beam Search: "a man in a yellow kayak is paddling through rough waters".

Look ahead inference: " a man is snowboarding down a snowy hill" [example for bad caption]

Fig. 4: Qualitative Results: The figure displays a few images from the dataset. The caption in black color is true caption, the one in red color is caption generated using Policy Network with Greedy Search, the one in blue color is caption generated using Policy network with Beam Search and the one in green color is generated by Look-ahead Inference.