2023 IEEE Symposium Series on Computational Intelligence (SSCI)
Mexico City, Mexico. December 5-8, 2023

Image Caption Generation Based on Image-Text Matching Schema in Deep Reinforcement Learning

Elyas Rashno, Mahdieh Safarzadehvahed, Farhana Zulkernine, Sidney Givigi
School of Computing
Queen's University
Kingston, Canada
elyas.rashno@queensu.ca, 22ms70@queensu.ca, Farhana.zulkernine@queensu.ca, sidney.givigi@queensu.ca

***Abstract*—Image captioning applications require prompt and precise caption generation, which can improve the accessibility and understanding capabilities of images. We utilize an actor-critic approach based on deep reinforcement learning and propose a two-fold approach to enhance the performance of the actor-critic approach. First, we propose a novel image-text matching module to compute the reward in image-matching, for the actor-critic model. This module enables more accurate and meaningful evaluations, contributing to improved caption generation. Second, we apply various training scenarios in reinforcement learning, strategically updating both the policy and value networks. The scenarios ensure more effective learning dynamics and lead to enhanced overall performance. To assess the efficiency of our approach, we employ the Microsoft COCO dataset. The experiments demonstrate the superiority of our method in terms of both speed and precision compared to the existing techniques.**

***Index Terms*—Image Captioning, Actor-critic, Reinforcement Learning, Image-text matching**

I. Introduction

Vision-and-Language (V+L) tasks involve understanding and processing both visual and textual information simultaneously [1]. V+L tasks aim to bridge the gap between vision including images and videos and natural language descriptions by combining the two modalities to achieve a deeper understanding of the world. V+L tasks have attracted substantial interest in the domain of natural language processing (NLP) and computer vision due to their practical applications and challenges [1]–[8]. Some popular V+L tasks include caption generation [4], visual question answering [5], visual dialogue [6], visual grounding [7], text-to-image synthesis [8], and image-text matching [3]. Among them, caption generation, also known as image or video captioning, is the process of generating natural language descriptions of visual content such as images or videos [4]. This task involves analyzing and understanding the content of the visual input, selecting relevant information, and generating a coherent and descriptive sentence that accurately conveys the content of the image or video with an encoder-decoder framework. Caption generation holds significance as a task in both computer vision and natural language processing, and has a wide range of applications including the image or video retrieval [9], [10] and accessibility for people with visual impairments [11].

Researchers have proposed various approaches to image/video caption generation [12]–[20] based on encoder-decoder frameworks. An encoder-decoder framework requires a separate training phase, where the encoder functions as a feature extraction mechanism, and the decoder generates captions. In reinforcement learning, the encoder and decoder are trained jointly, where each generated word's output is used to compute a reward for updating the parameters of the encoder and decoder components. To address the caption generation problem dynamically, researchers have explored various approaches [15]–[18]. Furthermore, in the context of multimodal interactions, Chen et al. [1] and Sun et al. [21] utilize a multimodal foundation model to generate both images and texts.

Ren et al. [17] propose a novel image caption generation model based on reinforcement learning which uses an actor-critic learning strategy and a gradual training process. In this procedure, the model [17] obtains the corresponding caption for each image and removes a predetermined number of words from the end. The model is then tasked with identifying appropriate words to replace the removed ones. Through this approach, the model initially learns to predict the last words and then gradually predicts the entire caption. It trains the actor and critic jointly after each step. One of the main challenges in this scenario involves accurately calculating the reward based on the current state. To address this, a new value network is proposed to map both the image and generated caption into a unified space to calculate their similarity. Additionally, given the need for the real-time performance of image captioning applications, it is crucial to generate captions quickly. Therefore, this paper also aims to minimize the processing time. By employing reinforcement learning jointly train the actor and critic networks in each sequence, leading to both networks being trained simultaneously but with uncertainty. To overcome these mentioned challenges, we propose an efficient approach that enhances both efficiency and certainly during training.

In this paper, we enhance the actor-critic model for image captioning proposed by Ren et. al [17] using deep reinforcement learning to jointly train the actor and the critic on multimodal embeddings of images and text to narrow down the semantic gap between visual and textual cues. The actor is trained to predict the caption text in a sequential decision-making process given an image. The critic evaluates the caption based on its similarity to the image and predicts the reward to the actor and moves to one of the potential next states. The rewards predicted by the critic are utilized to train the actor. Specifically, a policy network serves as the actor, and a value network functions as the critic. These networks are trained jointly through using deep reinforcement learning.

The main contributions of this study include the introduction of a novel image-text matching module as the value network for evaluating rewards in the actor-critic model and the experiment of the diverse training scenarios for reinforcement learning to simultaneously train both the policy and value networks. The paper is organized as follows: in Section II, an overview of the related works is presented. Section III provides a comprehensive description of the suggested methodology, including its key components and processes. Section IV showcases the outcomes from our experiments, including an analysis and interpretation of the results. Finally, we conclude the paper in Section VI with the list of future works.

II. Related Works

Zhang et al. [16] propose a technique for generating image captions that utilize reinforcement learning with an actor-critic model. The model is designed to address the problem of optimizing evaluation metrics that are not differentiable. The architectural design of the model encompasses two key components: a policy network, referred to as the actor, and a value network, known as the critic. The actor module functions as a sequential decision-maker problem and predicts the next token in the sequence, receiving a task-specific reward, an evaluation metrics score, at each stage of the sequence. The primary function of the critic is to forecast the anticipated reward. If the critic is able to successfully forecast the predicted reward, the actor will proceed to generate outputs by sampling from its probability distribution. They train the actor and critic jointly and one of the main challenges is to calculate the reward based on the current state accurately.

Chen et al. [1] propose a pre-trained model for joint multimodal embedding. The approach employs four tasks to pretrain the UNiversal Image-TExt Representation (UNITER). These tasks consist of Masked Language Modeling (MLM) conditioned on the image, Masked Region Modeling (MRM) conditioned on text, Image-Text Matching (ITM), and Word-Region Alignment (WRA). The encoding process involves mapping visual and bounding box properties of images as well as tokens and positions of textual words into a shared embedding space. Subsequently, a transformer module is employed to acquire contextualized embeddings for both regions and words by means of the pretraining job. Image-text matching helps to narrow down the semantic gap between visual and textual cues.

Liu et al. [22] propose a new method for creating image captions that take into account both language and visual policies, unlike the previous image captioning models based on reinforcement learning strategies that focus only on language policy. This new approach, called Context-Aware Visual Policy Network (CAVP), addresses the issue of capturing visual context which is important for understanding visual relationships and sequential visual reasoning. During each time step, the CAVP model takes into account the preceding visual attention as contextual information and determines its relevance in generating the current word, based on the current visual attention. The authors optimize CAVP and its subsequent language policy network end-to-end through an actor-critic policy gradient method, which is used to evaluate the generated captions against any caption evaluation metric. The visual policy, nonetheless, does not make use of spatial details concerning these regions in the image.

Zha et al. [23] propose a method for fine-grained image-to-language generation that considers image paragraph captioning in addition to image sentence captioning. They consider language and visual policy to capture visual context. When creating captions, the model takes into account visual attention from previous moments as context. Based on the current visual attention, it determines whether or not to use this context for generating the current word or sentence. To train the model they use the actor-critic policy gradient method. This approach enables the previous visual features generated as output, to be utilized as the visual context for the present action. It consists of four distinct sub-policy networks, including single, context, composition, and output policy which collectively complete the task of making decisions based on visual input. Each one of these four network policies is implemented using an RNN.

III. Proposed Model Architecture

In this section, we introduce our methodology by conceptualizing image captioning as a reinforcement learning problem. First, we provide a comprehensive explanation of the entire methodology and then a more intricate discussion on policy and value networks. Finally, the training strategy is explained to choose the best action in the language model.

A. Problem Modeling

The reinforcement learning framework comprises an agent that engages with an environment, wherein the agent perceives a state and decides on an action to accomplish a specific objective. As the agent transitions to a new state, it receives a reward based on the action taken. Within the realm of picture captioning, our primary aim is to produce a comprehensive and precise description that effectively portrays the provided image. We model the image captioning problem as shown in Figure 1.

We define the method as the agent A, the input is the image I, and the output is a sentence that includes a set of words $C = \{W_1, W_2, W_3, \ldots, W_T\}$. I and C comprise the environment $E = \{I, W_1, W_2, W_3, \ldots, W_T\}$. The $S_t = \{I, C_t\}$ action is to gradually select a word to complete a caption from dictionary D, which contains 10^3 words. The input image I and the partially generated caption C_t are considered as a state.

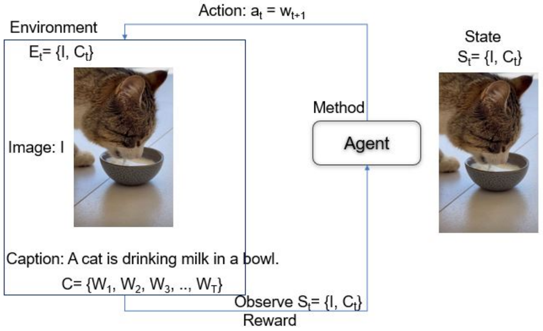
Action: $a_t = w_{t+1}$
Environment
$E_t = \{I, C_t\}$
Image: I
Caption: A cat is drinking milk in a bowl.
$C = \{W_1, W_2, W_3, .., W_T\}$
Method
Agent
State
$S_t = \{I, C_t\}$
Observe $S_t = \{I, C_t\}$
Reward

Fig. 1. Modeling Image captioning in terms of a reinforcement learning problem.

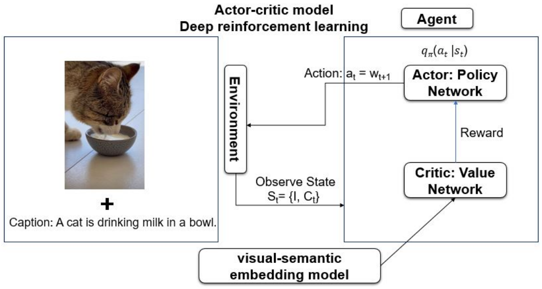
Actor-critic model
Deep reinforcement learning
Agent
Environment
+
Caption: A cat is drinking milk in a bowl.
$q_\pi(a_t | s_t)$
Action: $a_t = w_{t+1}$
Actor: Policy Network
Reward
Critic: Value Network
Observe State $S_t = \{I, C_t\}$
visual-semantic embedding model

Fig. 2. The architecture of the actor-critic model

To facilitate the training of the agent for generating descriptive captions, we employ an actor-critic model and employ deep reinforcement learning techniques. The actor tries to predict the next word to complete the caption, and the critic evaluates the word by calculating a reward. The actor and critic are modeled using a policy (q_π) and a value network (v_θ), respectively. The agent updates the policy network based on the feedback received from the value network. As shown in Figure 2, the policy network at each time t observes the state S_t and assigns probabilities to all the words in dictionary D based on their probability distribution. The most probable word is selected using a softmax function as the action a_t. Then, the value network calculates the similarity between the partially generated caption and the image's features between 0 and 1 and generates a reward for the agent.

B. Policy Network

The policy network q_π predicts the next action a_{t+1} of the agent at time t according to the current state $q_\pi(a_t|S_t)$. It comprises a CNN and an LSTM, which provide probabilities for the agent to make decisions at each given state. The architecture design of q_π is seen in Figure 3.

When an image is fed into the policy network, the CNN encodes the visual characteristics or features of the image. Then these features are inputted into the LSTM that provides the probability of selecting the action $a_t = W_{t+1}$ for the state $S_t = \{I, W_1, W_2, W_3, ..., W_t\}$ which is equal to $q_\pi(a_t|S_t)$. The policy network is trained using supervised learning based on a dataset of images and their captions that have been written by humans and optimizes with the cross-entropy loss.

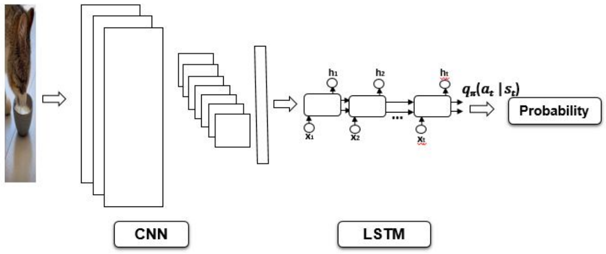

Fig. 3. The architecture of $q_\pi(a_t|S_t)$.

$q_\pi(a_t|S_t)$ is calculated using Equation1 where in $W^{x,v}$,ϕ and φ are the weights of the linear embedding model, input and output models of policy network's RNN, respectively.

$$\begin{aligned} x_0 &= W^{x,v} CNN(I) \\ h_t &= RNN(h_{t-1}, x_t) \\ x_t &= \phi(w_{t-1}), t > 0 \\ q_\pi(a_t|S_t) &= \varphi(h_t) \end{aligned} \tag{1}$$

C. Value Network

The value network $q_\pi(a_t|S_t)$ calculates a reward for action a_t in the current state S_t and helps to validate the selected action. The Reward system facilitates the training of two neural networks, which are responsible for translating distinct components of the image and the language into a common semantic space. As a result, it evaluates the similarity between the image and text on a word-by-word basis. This system comprises two distinct components referred to as the image encoder and the text encoder.

The text encoder employs a bi-directional LSTM [24] to extract meaning vectors from the textual description. In the context of this bi-directional LSTM, every word vector is associated with two distinct hidden states, each of which signifies a distinct direction. By concatenating these two vectors, a representation of the semantic meaning of each word is constructed. The matrix $e \subset R^{D \times n}$ is used to indicate the feature matrix of all words, where e_i represents the feature vector corresponding to the i^{th} word. In this context, the variable D symbolizes the dimension of the word vector, while n is employed to indicate the total number of words in the caption. Furthermore, the final hidden states of the bi-directional LSTM model are concatenated to form the global sentence vector, denoted as $\bar{e} \subset R^D$.

The utilization of a CNN in the image encoder facilitates the process of converting images into meaningful vectors. The intermediate layers of CNNs are responsible for capturing distinct features from different sub-regions of the input image, whilst the latter layers are primarily dedicated to learning global features. We use the Inception-v3 model [25], which was pre-trained on ImageNet [26], as the foundation for our image encoder (Figure 3).

The outputs from both encoders reside within a shared vector space. The reward is determined by evaluating the degree of similarity between the feature vectors produced

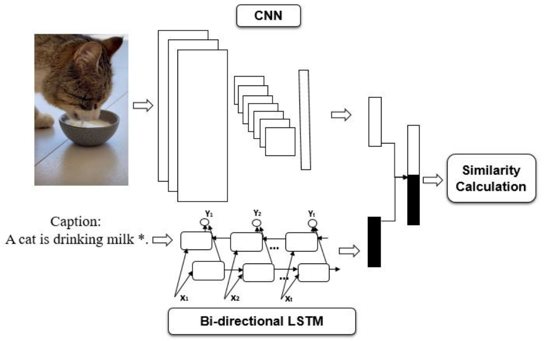

Fig. 4. The architecture of value network (v_θ)

by the text encoder and the image encoder. Notably, the model assigns a score of 1 when the feature vectors from both encoders are identical and a score of 0 when they are dissimilar.

D. Two-phase Training Strategy

The process of training occurs in two distinct phases. In the first phase, policy and value networks are trained separately. The policy network q_π is trained using supervised learning with cross-entropy loss according to Eq. 2.

$$L_{q'_t} = -log\; q(W_1, W_2, ..., W_t|I) = -\sum_{t=1}^{N} log\; q_\pi(a_t|S_t)] \tag{2}$$

In the second phase, the policy and value networks are concurrently trained using deep reinforcement learning. The model's parameters are learned through the process of maximizing the cumulative reward obtained by the agent during its interactions with the environment. The model is trained incrementally by gradually removing words from the end of the caption and predicting only the omitted words. Rather than predicting the entire caption at once, the model initially predicts a reduced number of words, and this prediction length is gradually increased until all words in the caption are included.

Once an image I is selected, the last two words of its corresponding caption are removed. Subsequently, the policy network is trained to predict these two missing words, while the value network provides feedback to the policy network by assigning rewards for each predicted word, aiding in the training process. Next, a great number of words are removed, and the training continues till predict the whole sentence.

IV. Results

In this section, we conduct comprehensive experiments to assess the effectiveness of our proposed framework. We begin by discussing the dataset and implementation specifics, followed by a comparison of our method with the state-of-the-art approaches in image captioning.

A. Dataset

We use the widely used Microsoft COCO dataset [27] and a well-known caption evaluation tool named BLEU [28], to compare our results with others. To enable a fair comparison with existing approaches in the literature, a standard dataset consisting of 82,783 training photos, 5,000 validation images, and 5,000 testing images is employed. Furthermore, it is worth noting that each image is accompanied by captions provided by a minimum of five distinct workers from the Amazon Mechanical Turk (AMT) platform. The preprocessing of the captions, which involves the construction of dictionaries and tokenization, is conducted using the approach outlined by Karpathy et al. [29].

B. Implementation Details

1) Network Architecture: Our policy network and value network as shown in Figures 3 and 4, incorporate a CNN and RNN which we train independently. For the policy network, we utilize VGG-16 [30] and LSTM [31] as our CNN and RNN architectures, respectively, with a node dimension and hidden state dimension of 512. While there are numerous CNN and RNN architectures available, such as ResNet [32] and GRU [33], which have reported better performance, we choose VGG-16 and LSTM for a fair comparison with the existing methods. In the value network, we employ a three-layer convolutional layer followed by a fully connected (FC) layer for image encoding. Additionally, we utilize a bi-directional LSTM for caption encoding of length based on the maximum word count size for captions in the dataset.

2) Training Details: During the training phase, we apply the Adam [34] optimization. It's essential to note that our network and embedding training exclusively used the dataset's provided images and captions. We achieved this by employing a pre-trained VGG-16 model without introducing external data. In the testing phase, we generate captions by sequentially selecting words until the suggested lookahead inference approach identifies a unique end token. In this paper, we introduce two methods to enhance analytical capabilities. Method A offers an alternative approach, grouping captions into three-word sequences to reduce training time. Method B suggests updating the policy and value networks after each step, rather than after each word.

C. Comparison Methods

Table I provides a comprehensive overview of the results obtained from our two methods as well as the published works. Method A is an alternative approach that groups captions into three-word sequences instead of two-word sequences to reduce the training time. Method B updates the policy and value networks after processing each group of words rather than a word. Our method surpasses all evaluation metrics and attains state-of-the-art performance on Microsoft COCO. We conduct a comparative analysis of our proposed method with a Deep Reinforcement Learning-based Image Captioning (RL) approach, as documented by Ren et al. [17]. Our results demonstrate superior performance for all Bleu metrics in our proposed method (I). This improvement is attributed to the utilization of an enhanced Image-text matching module, which

TABLE I
THE COMPARISON TABLE ON MICROSOFT COCO DATASET WITH OTHER STATE-OF-THE-ART MODELS

	Bleu-1	**Bleu-2**	**Bleu-3**	**Bleu-4**
Google NIC [39]	0.666	0.461	0.329	0.246
M-RNN [40]	0.67	0.49	0.35	0.25
BRNN [29]	0.642	0.451	0.304	0.21
LRCN [40]	0.628	0.442	0.304	0.21
MSR/CMU [41]	-	-	-	0.19
Spatial ATT [37]	0.718	0.504	0.357	0.25
gLSTM [41]	0.67	0.491	0.358	0.264
MIXER [38]	-	-	-	0.29
Semantic ATT [35]	0.713	0.539	0.403	0.304
DCC [36]	0.713	0.539	0.403	0.304
RL based [17]	0.713	0.539	0.403	0.304
Proposed Method A	0.709	0.517	0.385	0.264
Proposed Method B	**0.718**	**0.545**	**0.401**	**0.315**

yields more precise scoring as rewards for the model, leading to more effective updates in the policy network.

It should be noted that Semantic ATT [35] and DCC [36] employed additional data sources in their visual attribute predictor training, rendering their results cannot be compared to those achieved by other methodologies that did not use similar data. Interestingly, our method achieves better results than both Semantic ATT [35] and DCC [36], even without the use of external training data. The spatial attention [37] was proposed to replicate the human visual system as a mechanism to guide the model's focus towards significant and intricate details.

The MIXER technique, as described in [38], relies on metric-driven training. However, a model trained using the Bleu-4 metric from [38] struggles to generalize well across other evaluation metrics. In contrast, our approach, which employs embedding-driven decision-making, demonstrates strong performance across all metrics. Notably, our policy network, depicted in Figure 2, is inspired by the fundamental image captioning model similar to Google NIC [39]. The observed significant improvements in comparison to the findings presented in Vinyals et al. [39], provide validation for the efficacy of our proposed decision-making framework. This framework effectively integrates both policy and value networks. Moreover, the modularity of our framework allows for the seamless integration of other potent mechanisms, such as spatial attention and semantic attention, into the policy network, further enhancing overall performance.

D. Training Time

The experiments were conducted on a server equipped with 32 CPU cores, 256GB of RAM, and an A100 GPU with a 256 GB storage capacity. The server utilizes the Ubuntu operating system and is equipped with Python 3.8. Necessary data for the tests was stored on an expandable SSD storage medium. The hardware and software setup utilized in our studies proved to be adequate, facilitating the efficient execution of machine learning models. Table II displays the training duration per epoch for both our proposed technique and the RL-based method proposed by Ren et al. [17].

TABLE II
TRAINING TIME FOR THE BASE PAPER AND OUR PROPOSED METHOD

	Method	**Training Time**
1	RL based [17]	5 seconds
2	Proposed Method	**4 seconds**

V. DISCUSSION

Here, we provide a more detailed explanation of the rationale behind the results of our two proposed methods. In method A, we proposed an alternative approach that groups captions into three-word sequences to reduce training time. This method involves using [3, 6, 9, 12, 15] steps instead of [2, 4, 6, 8, 10, 12, 14, 16] during the training phase. However, our findings indicate that this method does not improve the accuracy of caption generation. The main issue with our approach is that when the model predicts the third word in the sequence, there is lower uncertainty compared to the first and second words. As a result, the policy and value networks are updated with lower uncertainty, leading to limited improvements in accuracy.

The second method entails updating the policy network at each step rather than for each word. Our approach is motivated by the observation that processing each word independently can lead to less effective value updates, and therefore lower accuracy. In contrast, by updating the value network after each step, the model can evaluate and update its understanding of the caption in a more holistic manner. For example, when the first predicted word is 'A' and the second predicted word is 'cat', the value network is updated based on these two words as a phrase rather than two individual words. This approach enables the model to better evaluate the caption as a whole and improve the validity of its predictions. In addition to improving accuracy, updating the value network after each step also reduces training time. This is because the model can learn from longer sequences of words at once, rather than processing each word independently.

VI. CONCLUSION

In this paper, we propose an actor-critic model based on deep reinforcement learning for image captioning by employing joint multimodal embeddings of images and text to narrow down the semantic gap between visual and textual cues. We used a novel image-text matching module to compute the reward in image-matching, for the actor-critic model. Then, we applied various training scenarios in reinforcement learning, strategically updating a policy network and a value network. In future investigations, we employ the suggested approach in the domain of video caption generation. To achieve this, it becomes imperative to instantiate a reinforcement learning agent predicated upon video frames, necessitating real-time processing.

REFERENCES

[1] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text

representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020.

[2] Akiyoshi Tomihari and Hitomi Yanaka. Logic-based inference with phrase abduction using vision-and-language models. *IEEE Access*, 11:45645–45656, 2023.

[3] Sungwoong Kim, Daejin Jo, Donghoon Lee, and Jongmin Kim. Magvlt: Masked generative vision-and-language transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23338–23348, 2023.

[4] AYUSH KUMAR GUPTA. *Attention Based Image Caption Generation*. PhD thesis, 2023.

[5] Usman Naseem, Matloob Khushi, and Jinman Kim. Vision-language transformer for interpretable pathology visual question answering. *IEEE Journal of Biomedical and Health Informatics*, 27(4):1681–1690, 2023.

[6] An-An Liu, Chenxi Huang, Ning Xu, Hongshuo Tian, Jing Liu, and Yongdong Zhang. Counterfactual visual dialog: Robust commonsense knowledge learning from unbiased training. *IEEE Transactions on Multimedia*, pages 1–13, 2023.

[7] Zhihong Chen, Ruifei Zhang, Yibing Song, Xiang Wan, and Guanbin Li. Advancing visual grounding with scene knowledge: Benchmark and method. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15039–15049, June 2023.

[8] Guokai Zhang, Ning Xu, Chenggang Yan, Bolun Zheng, Yulong Duan, Bo Lv, and An-An Liu. Cd-gan: Commonsense-driven generative adversarial network with hierarchical refinement for text-to-image synthesis. *Intelligent Computing*, 2:0017, 2023.

[9] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

[10] Avinash Madasu, Estelle Aflalo, Gabriela Ben Melech Stan, Shao-Yen Tseng, Gedas Bertasius, and Vasudev Lal. Improving video retrieval using multilingual knowledge transfer. In *European Conference on Information Retrieval*, pages 669–684. Springer, 2023.

[11] Simran Sharma, Madhulika Sharma, Arpit Yadav, and Rajkumar Balmeeki. Audio and image caption generator from image using vgg-16.

[12] MD Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys (CsUR)*, 51(6):1–36, 2019.

[13] Himanshu Sharma, Manmohan Agrahari, Sujeet Kumar Singh, Mohd Firoj, and Ravi Kumar Mishra. Image captioning: a comprehensive survey. In *2020 International Conference on Power Electronics & IoT Applications in Renewable Energy and its Control (PARC)*, pages 325–328. IEEE, 2020.

[14] Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli, Giuseppe Fiameni, and Rita Cucchiara. From show to tell: A survey on deep learning-based image captioning. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):539–559, 2022.

[15] Haichao Shi, Peng Li, Bo Wang, and Zhenyu Wang. Image captioning based on deep reinforcement learning. In *Proceedings of the 10th International Conference on Internet Multimedia Computing and Service*, pages 1–5, 2018.

[16] Li Zhang, Flood Sung, Feng Liu, Tao Xiang, Shaogang Gong, Yongxin Yang, and Timothy M Hospedales. Actor-critic sequence training for image captioning. *arXiv preprint arXiv:1706.09601*, 2017.

[17] 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.

[18] Anan Liu, Ning Xu, Hanwang Zhang, Weizhi Nie, Yuting Su, and Yongdong Zhang. Multi-level policy and reward reinforcement learning for image captioning. In *IJCAI*, pages 821–827, 2018.

[19] Andrew G Barto, Richard S Sutton, and Charles W Anderson. Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE transactions on systems, man, and cybernetics*, (5):834–846, 1983.

[20] Elyas Rashno and Farhana Zulkernine. Efficient video captioning with frame similarity-based filtering. In *International Conference on Database and Expert Systems Applications*, pages 98–112. Springer, 2023.

[21] Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative pretraining in multimodality. *arXiv preprint arXiv:2307.05222*, 2023.

[22] Daqing Liu, Zheng-Jun Zha, Hanwang Zhang, Yongdong Zhang, and Feng Wu. Context-aware visual policy network for sequence-level image captioning. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1416–1424, 2018.

[23] Zheng-Jun Zha, Daqing Liu, Hanwang Zhang, Yongdong Zhang, and Feng Wu. Context-aware visual policy network for fine-grained image captioning. *IEEE transactions on pattern analysis and machine intelligence*, 44(2):710–722, 2019.

[24] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681, 1997.

[25] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[26] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

[27] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[28] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

[29] 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.

[30] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[31] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[32] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[33] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[34] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[35] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4651–4659, 2016.

[36] Lisa Anne Hendricks, Subhashini Venugopalan, Marcus Rohrbach, Raymond Mooney, Kate Saenko, and Trevor Darrell. Deep compositional captioning: Describing novel object categories without paired training data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–10, 2016.

[37] 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.

[38] Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. *arXiv preprint arXiv:1511.06732*, 2015.

[39] 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.

[40] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015.

[41] Xinlei Chen and C Lawrence Zitnick. Mind's eye: A recurrent visual representation for image caption generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2422–2431, 2015.