The Impact of Text Generation Techniques on Neural Image Captioning: An Empirical Study

Linna Ding[1], Mingyue Jiang[2]*, Liming Nie[3], Zuzhang Qing[4], Zuohua Ding[5]
Faculty of Computer Science and Technology, Zhejiang Sci-Tech University, Hangzhou, Zhejiang, China[1,2,5]
School of Computer Science and Technology, Nanyang Technological University, Singapore, Singapore[3]
Zhejiang Petroleum Integrated Energy Sales Co., Ltd, Hangzhou, Zhejiang, China[4]

***Abstract*—Image captioning is an advanced NLP task that has various practical applications. To meet the requirement of visual information understanding and textual information generation, the encoder-decoder framework has been widely adopted by image captioning models. In this context, the encoder is responsible for transforming an image into vector representation, and the decoder acts as a text generator for yielding an image caption. It is obvious and intuitive that the decoder is crucial for the entire image captioning model. However, there is a lack of comprehensive studies in which the impact of various aspects of the decoder on the image captioning is investigated. To advance the understanding of the impacts of text generation techniques employed by the decoder, we conduct an extensive empirical analysis of three types of language models, two types of decoding strategies and two types of training methods, based on four state-of-the-art image captioning models. Our experimental results demonstrate that the language model affects the performance of image captioning models, while different language models may benefit different image captioning models. In addition, it is also revealed that among the decoding and training strategies under investigation, the beam search, AOA mechanism and the reinforcement learning based training method can generally improve the performance of image captioning models. Moreover, the results also show that the combinational usage of these strategies always outperforms the use of single strategy for the task of image captioning.**

Keywords*—Image captioning; encoder-decoder; text generation techniques

I. Introduction

Image captioning aims to provide accurate and textual descriptions for a given image. It is a challenging task integrating visual as well as textual understanding, and it involves technologies from both computer vision and natural language processing. Automatic image captioning has found practical applications in various domains, including social media [1], remote sensing [2], robotics [3], and medical image report generation [4].

Automatic image captioning has been receiving much attention in recent years, and a variety of approaches and strategies have been proposed and studied [5]. Although deep learning models have made significant progress in image captioning, describing images correctly remains a challenge. Image captioning models need to understand image content, object recognition, and object relationships while capturing the interaction between images and language to generate natural language descriptions.

*Corresponding authors.

Inspired by the advances in neural machine translation, most state-of-the-art image captioning models follow the encoder-decoder pipeline , which consists of an encoder and a decoder. Specifically, an encoder is used to transform an image into vector representations, and a decoder is used for translating the information from the encoder into natural sentences, yielding a relevant caption. In the literature, different encoders, decoders, and varying strategies supporting the encoding or decoding process have been investigated [5], [6].

As one of the core elements of image captioning, the decoder that acts as a text generator has attracted lots of research focuses. At first, various different language models have been employed as the decoder. Under the encoder-decoder framework, a mainstream image captioning model is CNN-RNN [7], where convolutional networks (CNN) are employed as the encoder for feature learning, followed by recurrent neural networks (RNN) act as a decoder for caption generation. Apart from that, various different models have been proposed and developed, including CNN-Long Short-Term Memory (LSTM) [8], CNN-Gated Recurrent Unit (GRU) [9], and CNN-Transformer [10]. On the other hand, different decoding strategies have been proposed and studied. Firstly, beam search has been widely adopted by RNN-based decoders for improving the quality of the output caption [11]. Secondly, to enhance the attention-based decoder, the attention on attention (AOA) strategy [12] has been proposed to extend the conventional attention mechanism. Last but not least, in order to enhance the decoder's capability of learning to predict the words appearing in the caption, various training strategies have been explored, including cross-entropy loss and reinforcement learning.

Naturally, different image captioning approaches employing varying strategies or mechanisms may have varying capabilities of generating captions. As can be seen from Fig. 1, for the same given input image, four different image captioning approaches provide four different captions, which are of varying quality as revealed by the evaluation metrics. In other words, different image captioning approaches may exhibit different captioning performance. However, due to the complexity of the image captioning models, the difference in performance may originate from the encoder, the decoder, or the relevant strategies. Recent studies have comprehensively analyzed the effect of different encoders on the model performance from an empirical perspective [13]–[18]. For the decoder part, its impacts on the image captioning performance have also been revealed and studied [19]–[22]. Nevertheless, there is still a lack of detailed and comprehensive investigations

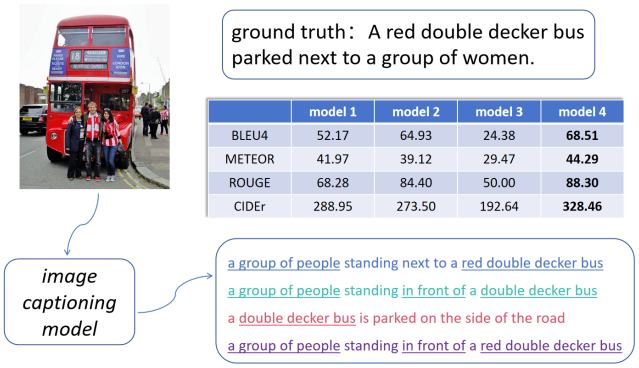


Fig. 1. For a given image, different image captioning approaches may yield varying captions.

of the impacts referring to various aspects of the decoder (including the language model, the decoding strategy, and training strategy).

To gain additional insights into the encoder-decoder based image captioning models, in this study, we conduct an extensive empirical study with the goal of comprehensively investigating the impacts of decoding related techniques on the performance of image captioning. We compared the impact of CNN-based, GRU-based, and LSTM-based decoders on image captioning models. In addition, we investigated the impacts of two types of decoding strategies, the search strategy (namely, the greedy search and the beam search) and the AOA mechanism.

We also analyzed the impacts of training methods on the performance of image captioning models and compared the impacts of two training methods, the Cross-Entropy Loss, and the reinforcement learning-based method. Furthermore, we investigated the impact of the combinational usage of these strategies.

We conducted experiments on the MSCOCO dataset [23], which is a widely-used dataset for the task of image captioning. We employed four state-of-the-art image captioning models as the basic models and further constructed a series of model variants from them by modifying the decoding parts of these basic models. To evaluate the performance of these image captioning models, we adopted six evaluation metrics, including BLUE1 [24], BLEU4 [24], METEOR [25], ROUGE [26], CIDEr [27] and SPICE [28]. Overall, our experimental results confirm the impacts of the language models, the decoding strategies, and the training strategies on the performance of image captioning models. More specifically, it is revealed that different language models may benefit different image captioning models, and the beam search, AOA mechanism, and the reinforcement learning based training method can generally improve the performance of image captioning models. In addition, it is also found that the combinational usage of various strategies can positively affect the captioning performance.

The contributions of this study are summarized as below.

- We conduct extensive experiments to empirically analyze the impacts of the decoder involving various text generation techniques on the performance of the image captioning models. Our study considers the impacts of language models, decoding strategies, training strategies, and the combinational usage of decoding and training strategies, and accordingly evaluates the performance of 68 image captioning models (including 4 basic models and 64 model variants with varying usage of the language model, decoding strategy and training strategy).
- We highlight some practical findings. Our findings suggests that the performance of an image captioning model can be properly enhanced by configuring it with suitable language model as well as appropriate decoding and training strategies. This also provides a reference for further improving the performance of image captioning models.

The rest of the paper is structured as follows. Section II provides an in-depth discussion of previous research work We introduce some preliminary knowledge, including the commonly used language models, the decoding and training strategies for image captioning models, in Section III. In Section IV, we present our experimental design, including the research questions, the basic image captioning models employed in the experiments, the datasets and the evaluation metrics. Section V reports and discusses our experimental results to answer each of our research questions. Section VI concludes with a summary of this study and proposes directions for future research.

II. Related Work

Image caption : Image captioning [29], [30], [31], [12] achieves significant improvements over the neural encoder-decoder framework [6]. The Show-Tell model [30] uses convolutional neural networks (CNNs) [32] to encode images into fixed-length vectors, and a Long short-term memory (LSTM) [33] as a decoder to sequentially generate words. To capture fine-grained visual details, attention-to-image captioning models [29], [31], [12] have been proposed to dynamically pin words together with relevant image parts during generation. To reduce exposure bias in sequence training, Rennie et al. [34] use reinforcement learning to optimize non-differentiable metrics. In order to further improve the accuracy, transformer models [10], [35] were proposed, allowing the model to effectively capture the relationship between different positions in the input sequence.

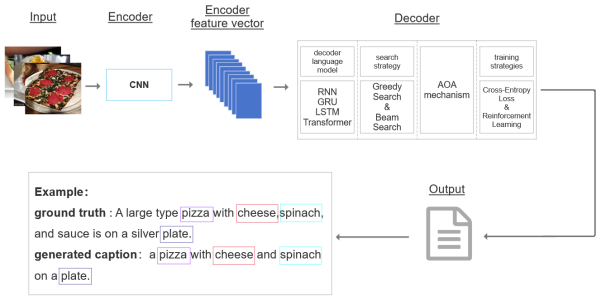


Fig. 2. Encoder-Decoder based image captioning.

Empirical study : The factors that affect the image caption model are roughly divided into two parts: encoder and decoder. In order to study the impact of encoders on image caption models, people began to use different CNN encoders, such as Inception-V3, VGG, Resnet, Densenet, etc., for empirical research. Among them, [13], [14], [16] used ordinary LSTM as the decoder, [17] used MSprop as the optimizer on this basis, and [18] changed the decoder part from LSTM to GRU. In order to ensure the comprehensiveness of the experiment, [15], [36] also used LSTM and a combination of LSTM and visual attention mechanisms as decoders.

The research on the decoder part mainly focuses on the decoder architecture, search strategy and visual attention mechanism. Among them, [37] mainly focused on the impact of search strategies. [20] considered the influence of one-way and two-way LSTM decoders and search strategies. [19] selected the injection model and conducted experiments using different search strategies. [21] and [22] mainly focued on the impact of visual attention mechanism on the model. In addition, [22] takes into account the Transformer model.

While other papers have analyzed only one aspect of the attention mechanism, or two types of decoder architectures, we have built on this foundation by experimenting with RNNs, GRUs, LSTMs, and Transformers using different types of strategies as well as combinations of strategies, in order to have a more comprehensive analysis of decoder language models.

III. Preliminaries

This section briefly introduces the commonly used language models, as well as the strategies that are applicable to text generators, in the context of encoder-decoder based image captioning.

A. Decoder Language Models

As shown in Fig. 2, the decoder of an image captioning model is responsible for translating the vector representation resulting from the encoder into a natural language caption. In the context of image captioning, the generation of captions can be formulated as a sequence to sequence learning task. Several language models have been employed to accomplish this task, including RNN, LSTM, GRU, and Transformer.

RNN [38] is used to process sequential data, but it does not handle long sequences and long-distance dependencies well due to vanishing or exploding gradient problems.

LSTM [33] is an improved RNN that effectively solves the gradient problem by introducing a gating mechanism, making it good at capturing long-term dependencies and achieving good results in tasks such as text generation and machine translation.

GRU [39] is an improved version of RNN. It uses a gating mechanism, has a simple structure and fewer parameters, and shows good performance in multiple sequence generation tasks, similar to LSTM.

Transformer [40] is a neural network based on a self-attention mechanism. It has global context modeling and parallel computing capabilities. It can comprehensively consider image features and subtitle sequences to generate accurate and coherent image subtitles.

B. Decoding Strategies

Apart from the language model, the decoder can be equipped with various different strategies. In this study, we mainly focus on the search strategy and the strategies relating to the attention mechanism.

1) Search Strategy: Greedy search and beam search are two search strategies for generating sequences that are commonly used in the task of image captioning.

Greedy search is a sequence generation method that selects the currently optimal option each time without considering the global optimal solution. It is usually computationally efficient but may sacrifice final performance.

Beam search is a sequence generation method that considers multiple alternative outputs and selects the set of alternative outputs with the highest probability score to improve the quality of the generated results, often used in natural language processing and machine translation tasks.

2) Attention on attention mechanism: For encoder-decoder framework based image captioning, the attention mechanism is commonly applied for guiding the decoding process. The Attention on Attention (AOA) approach aims at extending the traditional attention mechanism applied to image captioning tasks.

The AOA approach consists of two main parts: the first part is the global attention module, which is used to compute global attention weights between image features and context vectors; the second part is the local attention module, which is used to compute local attention weights based on the global attention weights.

C. Training Strategies

The training process aims to prepare the captioning model for learning to predict the probabilities of words that will appear in the caption. Two types of commonly adopted training strategies are Cross-Entropy Loss and reinforcement learning.

1) Cross-entropy loss: Traditional image captioning models are usually trained using maximum likelihood estimation (MLE) to optimize model parameters by minimizing cross-entropy loss. However, this method cannot directly measure subtitle quality and can easily lead to inaccurate or repeated subtitles.

2) Reinforcement learning: Training image captioning models using reinforcement learning has led to significant improvements. A typical method is self-criticism sequence training [34], which treats the generated subtitles as a sequence of actions, the quality is evaluated with the CIDEr-D metric, and the metric is maximized through reinforcement learning.

IV. Experimental Design

This section presents our research questions, basic image captioning models, datasets, and evaluation metrics.

A. Research Questions

We plan to investigate the following four research questions.

- **RQ1:** What is the impact of the language models on the performance of image captioning models?
- **RQ2:** How do different decoding strategies affect the performance of image captioning?
  RQ2.1: How does beam search compare to greedy search for the task of image captioning?
  RQ2.2: What is the impact of using the AOA mechanism with the language model on the performance of image captioning?
- **RQ3:** How do different training strategies used for the decoder impact the performance of image captioning?
- **RQ4:** What is the impact of the combinational usage of various strategies of the decoder on the performance of image captioning?

B. Basic Image Captioning Models

In this study, we employed four state-of-the-art image captioning models as the basic models, based on which we constructed various model variants (the details are elaborated in Section V). The information of these models is summarized in Table I, and further described below.

FC [34]: The FC model utilizes a deep CNN model ResNet101 to encode the input picture, and then a linear map is used for embedding. The model uses an LSTM-based decoder.

Att2in2 [34] : The Att2in2 model uses ResNet101 as an encoder and LSTM as a decoder. Particularly, it is an image captioning model involving the attention mechanism. The model is an improved version of the subtitle attention model [31].

Up-Down [29] : The up-Down model encoder part uses Faster R-CNN ResNet101, which is a classical target detection model, and local features from an image, and the decoder part employs a two-layer LSTM architecture (Top-Down Attention LSTM and Language LSTM), which utilizes both bottom-up and top-down attentional mechanisms, in order to generate natural language descriptions that match the content of the image.

Transformer [40] : The Transformer model uses Faster R-CNN ResNet101 as the encoder and employs a Transformer as the decoder. The Transformer decoder is an integral part of the Transformer model and is used to convert the input sequence generated by the encoder into a target sequence. It gradually generates target sequences through the self-attention mechanism and encoder-decoder attention mechanism.

As shown in Table I , the encoders of the above four models are either Faster R-CNN Resnet101 [41] or ResNet101 [32]. We further detail these two models as below.

ResNet101 [32]: ResNet101 is a CNN model having 101 layers. It is a variant of a Residual Network, and it introduces residual connections to solve the problem of gradient disappearance and gradient explosion in deep network training. Compared with traditional shallow networks, it can learn image features at a deeper level, thereby extracting more complex and advanced feature expressions.

Faster R-CNN ResNet101 [41]: Faster R-CNN ResNet101 combines the Faster R-CNN object detection algorithm and the ResNet101 feature extractor. As a feature extractor, ResNet101 can efficiently extract features from images. Faster R-CNN ResNet101 combines the efficiency of the target detection algorithm and the deep feature learning ability of ResNet101, making the model perform well in target detection tasks.

C. Dataset

Our experiments are conducted on the MSCOCO dataset [23], which is a popular benchmark for image captioning tasks, containing 123,287 images, each with 5 captions, for a total of 615,935 captions. We use the "Karpathy" data split [42], with 5,000 images for validation, 5,000 for testing, and the rest for training.

To preprocess the captions, we generated a vocabulary of 10,369 unique words by converting sentences to lowercase and removing words that appeared less than five times.

V. Experimental Results

This section analyzes and reports our experimental results. Specifically, for each RQ, we discuss the motivation, present the approach, and finally report the results.

A. RQ1: Impact of Language Models on the Task of Image Captioning

Motivation: For an encoder-decoder based image captioning model, the language model constitutes the key part of the decoder, and thus it is crucial to the overall performance of image captioning. Prior studies have proposed various language models for supporting the decoding stage of image captioning [5]. Yet, there is still a lack of empirical evidence revealing the extent of the impact, and it is also unclear how different language models affect the performance of image captioning models. Therefore, in this RQ, we investigated image captioning models with varying language models, in order to reveal the impact of language models on the performance of image captioning models.

Approach: First, we utilized the three baseline models employing an RNN-based language model (namely, FC, Att2in2, and Up-Down) by following their original configurations in the prior studies [12]. Secondly, we constructed six model variants from the three baseline models by modifying the decoder part. That is, for each model, two variants were constructed by

TABLE I. BASIC INFORMATION OF THE SELECTED MODEL

Model	Encoder	Decoder	Search Strategy	Attention
FC	ResNet101	LSTM	Greedy	
Att2in2	ResNet101	LSTM	Greedy	✓
Up-Down	Faster R-CNN ResNet101	Attention LSTM +Language LSTM	Greedy	✓
Transformer	Faster R-CNN ResNet101	Self-Attention mechanism+feed-forward neural network	Greedy	✓

replacing its default language model LSTM with an RNN and a GRU, respectively. For the sake of simplicity, we utilized $M\diamondsuit L$ to denote a model M supported with the specific language model L. For example, FC$\diamondsuit$RNN represents one variant of model FC where the default language model LSTM is replaced with an RNN. Thirdly, for each of the models obtained in the previous steps, we further constructed a variant for each of them by modifying its encoder model (i.e. from ResNet101 to Fast RCNN ResNet101, or vice versa). Finally, we evaluated these 18 models (including three baseline models and 15 model variants) and collected evaluation results on a series of evaluation metrics.

Results: Tables II and III, respectively report the evaluation results of nine models employing the same encoder model. Based on these results, we make the following observations.

1) ***The use of different language models leads to varying performance of the image captioning model.*** As shown in Table II, for each of the models, the use of RNN, GRU, or LSTM as the decoder model yields different values for each of the six evaluation metrics. For example, the BLEU1 values for the three models, that is, FC$\diamondsuit$RNN, FC$\diamondsuit$GRU, and the original FC model are 73.70, 73.71, and 74.06, respectively. Table III consistently reveals this point.

2) ***The language model affects different image captioning models in different ways.*** At first, it is observed that the language model may have opposite impacts on different image captioning models. Consider the FC and Att2in2 models as an example. According to Table II, compared to the use of RNN, the use of GRU positively contributes to the BLEU1 value of FC (the BLEU1 values of FC$\diamondsuit$RNN and FC$\diamondsuit$GRU are 73.70 and 73.71), while it negatively affects the BLEU1 value of Att2in2 (the BLEU1 values of Att2in2$\diamondsuit$RNN and Att2in2$\diamondsuit$GRU are 75.56 and 75.11). On the other hand, the degrees of the impacts of the language models may also be different when they are applied to different image captioning models. As can be observed from Table III, FC$\diamondsuit$GRU outperforms FC$\diamondsuit$RNN in terms of the CIDEr metric, exhibiting a discrepancy of 1.65 (97.72 vs. 96.07). Nevertheless, although Att2in2$\diamondsuit$GRU also outperforms Att2in2$\diamondsuit$RNN in terms of the CIDEr metric, the discrepancy in the performance is relatively tiny (0.14).

3) ***The best language model for different image captioning models may be different.*** Among the three language models under investigation (that is, RNN, GRU, and LSTM), they are beneficial to different image captioning models. For the models employing the faster R-CNN ResNet101 as the encoder, the best language model for the FC model is LSTM; while the Up-Down model exhibits the best performance with the GRU as the language model (as observed from Table II). Quite differently, for the models employing ResNet101 as the encoder, the FC model performs best with GRU, the Att2in2 model achieves the best performance with LSTM, while the Up-Down model performs best with RNN (as observed from Table III).

> **RQ1** : For the encoder-decoder based image captioning models, employing different language models as the decoder always leads to varying captioning performance. Nevertheless, the impact of the language models on different image captioning models may vary, and accordingly, the good language models may also be different from the perspective of different image captioning models.

B. RQ2: Impact of Different Decoding Strategies on Image Captioning Models

Motivation: At present, the endoer-decoder based image captioning models have been extended and enhanced via a variety of decoding strategies [5]. Although these decoding strategies have been demonstrated to be able to positively contribute to captioning performance, they have not been comprehensively investigated on the same set of image captioning models and datasets. To fill this gap, in this RQ, we empirically studied the impacts of two types of decoding strategies, the search strategy and the AOA mechanism.

Approach: We first focus on the search strategies adopted by the decoder of the image captioning model. To this end, we conducted experiments on 20 models, including the 18 models constructed for RQ1, the basic Transformer model and its variant employing the RestNet101 instead of the Faster R-CNN ResNet101 as the encoder. It is noted that all of these 20 models adopt the greedy search (as reported in Table I). Based on these, we further constructed 20 model variants from them by replacing the greedy search with beam search. In particular, the latter set of models is configured with various beam sizes (in this study, we adopted four beam sizes, 2, 3, 4, and 5). As a result, there are 20 groups of models, each of which consists of two models sharing the same technical details except for the search strategy. We evaluated all of these models on the dataset and compared the performances of models within individual groups.

To study the impacts of the AOA mechanisms, the three base models, Att2in2, Up-Down, and Transformer, and their variants are utilized. The FC model and its variants are excluded because they do not employ the attention mechanism and thus the AOA mechanism is not applicable. For each of the models, we constructed a variant for it by additionally applying the AOA mechanism, and then conducted a comparison analysis of their performance.

Results: Fig. 3 reports the performance comparison results of image captioning models using or not using the beam search strategy. Particularly, Fig. 3 (a)-(f) reports the results for the ten

TABLE II. EVALUATION RESULTS OF THE NINE MODELS EMPLOYING FASTER R-CNN RESNET101 AS THE ENCODER. AMONG EACH BASIC MODEL AND ITS VARIANTS, THE BEST PERFORMANCE IN TERMS OF INDIVIDUAL METRICS IS HIGHLIGHTED WITH BOLD TYPE. FURTHERMORE, THE BEST PERFORMER IN TERMS OF INDIVIDUAL METRICS IS UNDERLINED

Model	Decoder Language Model	BLEU1	BLEU4	METEOR	ROUGE	CIDEr	SPICE
FC	RNN	73.70	30.86	25.82	53.93	100.83	19.10
	GRU	73.71	31.19	25.97	54.19	102.04	**19.26**
	LSTM	**74.06**	**31.42**	**26.07**	**54.38**	**102.53**	19.21
Att2in2	RNN	75.56	**33.51**	26.70	55.41	**108.95**	**20.12**
	GRU	75.11	32.98	**26.74**	55.40	107.57	20.05
	LSTM	**75.97**	33.49	26.67	**55.46**	108.14	20.10
Up-Down	RNN	75.60	33.81	27.16	55.72	110.60	20.34
	GRU	**76.17**	**34.34**	**27.37**	**56.18**	**112.21**	20.54
	LSTM	75.64	33.88	27.34	55.94	111.90	**20.60**

TABLE III. EVALUATION RESULTS OF THE NINE MODELS EMPLOYING RESNET101 AS THE ENCODER

Model	Decoder Language Model	BLEU1	BLEU4	METEOR	ROUGE	CIDEr	SPICE
FC	RNN	71.89	29.24	25.08	52.80	96.07	18.25
	GRU	**72.37**	**29.76**	**25.38**	**53.08**	**97.72**	**18.46**
	LSTM	72.29	29.51	25.25	53.01	96.68	18.30
Att2in2	RNN	74.77	32.47	**26.53**	54.87	105.44	19.74
	GRU	**74.83**	32.41	26.33	54.74	105.58	19.71
	LSTM	74.75	**32.89**	26.47	**54.95**	**106.67**	**19.96**
Up-Down	RNN	**74.81**	32.27	**26.72**	**55.03**	**107.31**	**20.03**
	GRU	74.54	**32.42**	26.71	54.87	106.73	19.80
	LSTM	74.46	31.98	26.60	54.81	106.43	19.84

groups of models employing the faster R-CNN ResNet101 as the encoder, where each subfigure focuses on the comparison of performance with respect to one of the evaluation metrics. Accordingly, the comparison results relating to the other ten groups of models that using the ResNet101 as the encoder are reported in 3 (g)-(l). Fig. 4 further reports the performance comparison results on seven groups of models applying or not applying the AOA mechanism. Based on these results, we have the following observations:

1) ***The use of different decoding strategies affects the performance of image captioning models.*** It can be observed from Fig. 3 that using greedy search or beam search leads to varying captioning performance of the relevant models. Similarly, Fig. 4 also shows that every image captioning model under investigation exhibits different performance with and without using the AoA mechanism.

2) ***Compared to greedy search, the use of beam search generally improves the captioning performance.*** Firstly, it can be observed from Fig. 3 that most of the models achieve better performance by using beam search. This indicates that the use of beam search is beneficial to image captioning models. On the other hand, it can also be found that the optimal beam size of the beam search for different models varies. Nevertheless, for the majority of models, the best performance is reached with a beam size of 2.

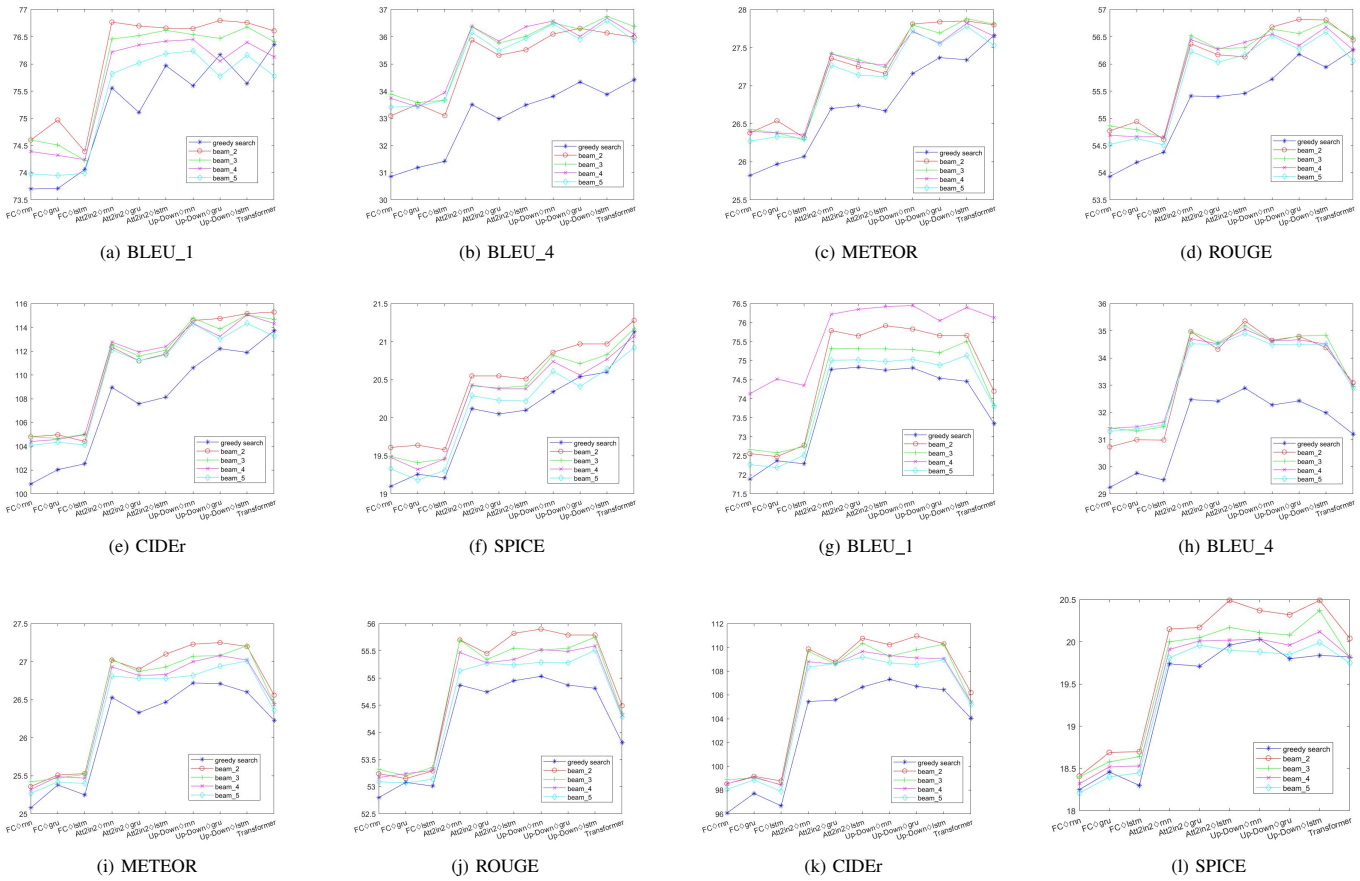
3) ***The application of the AoA mechanism benefits most of the image captioning models under investigation.*** Fig. 4 shows that after additional applying the AoA mechanism on the target image captioning models, the captioning performance has been improved in most cases (that is, for most of the models with respect to the majority of evaluation metrics). Although there are some models for which the application of the AoA mechanism leads to a decrease in captioning performance (i.e., the Up-Down model employing the encoder of faster R-CNN ResNet101), the extent of the decrease is relatively smaller than the extent of the increases resulted from using the AoA mechanism.

RQ2 : For encoder-decoder based image captioning models, the application of decoding strategies affects the captioning performance.Specifically, the use of beam search always outperforms the use of greedy search, and most models exhibit the best performance with the beam search configured with a beam size of 2. Moreover, the application of the AoA mechanism is beneficial to most of the image captioning models under investigation.

C. RQ3: Impact of the Training Strategies on Image Captioning Models

Motivation: Currently, encoder-decoder based image captioning models have emerged with various training methods. However, no prior study has focused on revealing the effect of training methods applied on the decoder part. Hence, in this RQ, we empirically studied two training approaches (Cross-Entropy Loss and Reinforcement Learning) and their impacts on captioning performance.

Approach: In the experiments, we reused the 20 image captioning models, including the FC, Att2in2, and Up-Down models employing the RNN, GRU, and LSTM in the decoder part as well as the Transformer model, and also their relevant variants using a different CNN (faster R-CNN ResNet101 or ResNet101) as the encoder. Noted that all of these models are trained by following their default method, namely, the cross-entropy loss method. Based on each of these models, we further constructed a model variant by training its decoder via a reinforcement learning based method, the self-critical sequence training method. These result in 20 groups of model, where each group consists of a model and its variant involving a decoder trained via reinforcement learning. We evaluated these newly constructed model variants and further conducted a comparison analysis with individual groups.



(a) BLEU_1 (b) BLEU_4 (c) METEOR (d) ROUGE

(e) CIDEr (f) SPICE (g) BLEU_1 (h) BLEU_4

(i) METEOR (j) ROUGE (k) CIDEr (l) SPICE

Fig. 3. Comparison of models employing the greedy search with those employing the beam search. For the latter, various beam sizes (2, 3, 4 and 5) have been investigated. A total number of 20 groups of models are studied, including ten groups of models using Faster R-CNN ResNet101 ((a) - (f)) and another ten groups of models using ResNet101 ((g) - (l)).

Results: Tables IV and V, respectively report the evaluation results of ten newly constructed models employing the same encoder model. For each newly constructed model (where the decoder is trained via reinforcement learning), we further compared its performance with the relevant model that trained via the cross-entropy loss method (as reported in Tables II and III). Accordingly, Tables IV and V further report the improvements made by applying the reinforcement learning based training method (the improvement is indicated by the ↑). Based on these results, we make the following observations:

1) ***Reinforcement learning is an effective training method for supporting the task of image captioning.*** Both Table IV and Table V reveal that training the decoder by the self-critical sequence training method leads performance improvement for all of the target models. For example, after applying the self-critical sequence training method, FC◇RNN exhibits 4.02 improvement in terms of the BLEU1 metric, while the improvement is 12.24 with respect to the CIDEr metic (as shown in the first row of Table IV).

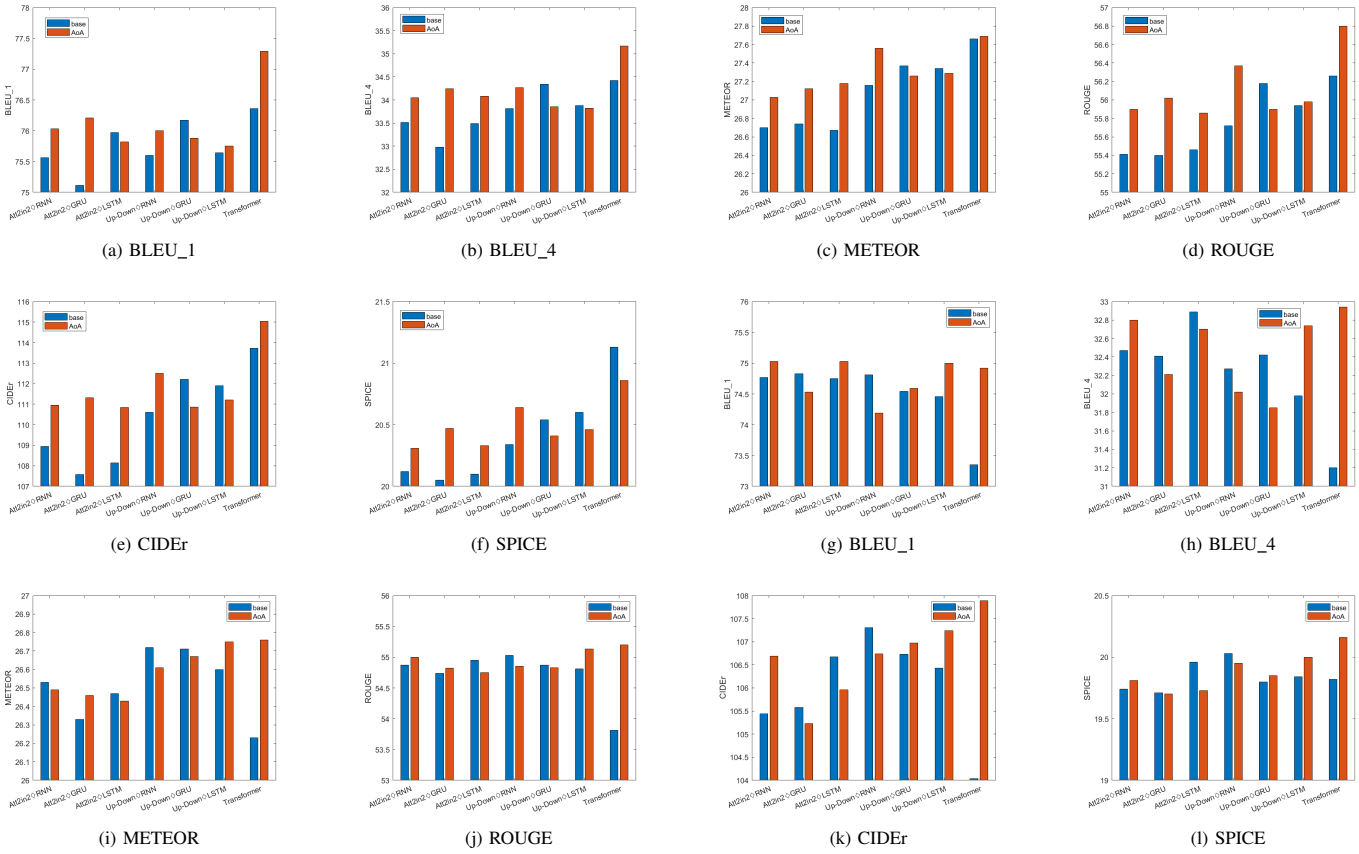
2) ***The performance improvements made by reinforcement learning are different for different image captioning models.*** According to Table IV, the performance improvements obtained via self-critical sequence training range from 0.59 to 14.16. Similarly, as shown in Table V, the highest increase in performance is 12.92, and the lowest increase is 0.46. Furthermore, it can be observed that the application of the reinforcement learning based training method leads to varying performance improvement for every of the target models.

> **RQ3** : For encoder-decoder based image captioning models, training the decoder with reinforcement learning will improve the model performance. Nevertheless, the degree of the improvements is different for different image captioning models.

D. Impact of the Combinational Usage of Decoding Strategies on the Performance of Image Captioning Models

Motivation: We have previously investigated the effect of every single strategy or mechanism on the performance of image captioning models. With the observation that these strategies and mechanisms can be applied to an image captioning model together, in this RQ, we further studied the effect of the application of various combinations of these strategies.

Approach: We utilized the four baseline models employing the faster R-CNN ResNet101 encoder as the basic model. We further considered the combination of the three strategies or methods, that is, the search strategy, the AoA mechanism, and the training method. We use P_{xyz} to denote the application



(a) BLEU_1 (b) BLEU_4 (c) METEOR (d) ROUGE

(e) CIDEr (f) SPICE (g) BLEU_1 (h) BLEU_4

(i) METEOR (j) ROUGE (k) CIDEr (l) SPICE

Fig. 4. Comparison of the 14 groups of models, including seven groups of models using Faster R-CNN ResNet101 ((a) - (f)) and another seven groups of models using ResNet101 ((g) - (l)). In each group, one model does not using AoA (denoted as *base*), while the other one applies AoA (denoted by AoA).

TABLE IV. PERFORMANCE OF MODELS WITH THE FASTER R-CNN RESNET101 AS ENCODER AND WITH THE DECODER TRAINED WITH SELF-CRITICAL SEQUENCE TRAINING METHOD. ↑ DENOTES THE RATE OF IMPROVEMENTS ACHIEVED BY APPLYING THE REINFORCEMENT LEARNING TRAINING METHOD. FOR EACH EVALUATION METRIC, THE LARGEST IMPROVEMENT IS HIGHLIGHTED BY UNDERLING

Model	Decoder Language Model	BLEU1	BLEU4	METEOR	ROUGE	CIDEr	SPICE
FC	RNN	77.72 ↑ 4.02	34.83 ↑ 3.97	26.89 ↑ 1.07	56.33 ↑ 2.41	113.07 ↑ 12.24	20.20 ↑ 1.10
	GRU	77.74 ↑ 4.03	34.62 ↑ 3.43	26.79 ↑ 0.81	56.16 ↑ 1.97	114.08 ↑ 12.05	20.05 ↑ 0.79
	LSTM	77.37 ↑ 3.31	34.69 ↑ 3.26	26.65 ↑ 0.59	56.17 ↑ 1.79	113.04 ↑ 10.51	19.99 ↑ 0.78
Att2in2	RNN	78.53 ↑ 2.97	36.46 ↑ 2.96	27.39 ↑ 0.68	57.17 ↑ 1.77	119.48 ↑ 10.53	20.84 ↑ 0.73
	GRU	78.58 ↑ 3.48	36.27 ↑ 3.29	27.45 ↑ 0.71	57.16 ↑ 1.75	119.25 ↑ 11.68	21.01 ↑ 0.97
	LSTM	78.46 ↑ 2.48	36.07 ↑ 2.59	27.38 ↑ 0.71	57.14 ↑ 1.67	119.13 ↑ 10.99	20.87↑ 0.77
Up-Down	RNN	79.88 ↑ 4.28	37.73 ↑ 3.92	28.20 ↑ 1.04	58.15 ↑ 2.42	124.76 ↑ 14.16	21.45 ↑ 1.11
	GRU	79.65 ↑ 3.48	37.26 ↑ 2.92	28.18 ↑ 0.80	57.89 ↑ 1.71	124.02 ↑ 11.81	21.55 ↑ 1.00
	LSTM	79.62 ↑ 3.97	37.24 ↑ 3.36	28.11 ↑ 0.77	57.83 ↑ 1.89	124.57 ↑ 12.67	21.52 ↑ 0.92
Transformer	Transformer	79.46 ↑ 3.09	38.04 ↑ 3.63	28.54 ↑ 0.88	57.94 ↑ 1.69	123.97 ↑ 10.26	22.20 ↑ 1.07

TABLE V. PERFORMANCE OF MODELS WITH THE RESNET101 AS ENCODER AND TRAINED WITH SELF-CRITICAL SEQUENCE TRAINING METHOD

Model	Decoder Language Model	BLEU1	BLEU4	METEOR	ROUGE	CIDEr	SPICE
FC	RNN	75.89 ↑ 4.00	33.10 ↑ 3.86	25.95 ↑ 0.87	55.10 ↑ 2.03	107.78 ↑ 11.71	19.32 ↑ 1.08
	GRU	76.16 ↑ 3.79	33.15 ↑ 3.39	26.14 ↑ 0.76	55.11 ↑ 2.03	108.98 ↑ 11.26	19.47 ↑ 1.01
	LSTM	76.33 ↑ 4.05	33.46 ↑ 3.94	26.19 ↑ 0.94	55.27 ↑ 2.25	107.91 ↑ 11.23	19.40 ↑ 1.10
Att2in2	RNN	77.71 ↑ 2.94	35.51 ↑ 3.04	26.99 ↑ 0.46	56.64 ↑ 1.77	116.59 ↑ 11.15	20.44 ↑ 0.71
	GRU	77.70 ↑ 2.86	35.34 ↑ 2.93	27.04 ↑ 0.71	56.56 ↑ 1.82	116.79 ↑ 11.21	20.50 ↑ 0.79
	LSTM	77.85 ↑ 3.10	35.35 ↑ 2.46	26.96 ↑ 0.49	56.63 ↑ 1.68	116.09 ↑ 9.42	20.44 ↑ 0.48
Up-Down	RNN	78.57 ↑ 3.76	36.03 ↑ 3.76	27.54 ↑ 0.82	57.11 ↑ 2.07	119.56 ↑ 12.24	20.98 ↑ 0.96
	GRU	78.81 ↑ 4.27	35.79 ↑ 3.36	27.58 ↑ 0.86	57.06 ↑ 2.19	119.42 ↑ 12.69	21.00 ↑ 1.20
	LSTM	78.76 ↑ 4.30	35.82 ↑ 3.84	27.56 ↑ 0.95	56.99 ↑ 2.18	119.35 ↑ 12.92	21.14 ↑ 1.30
Transformer	Transformer	76.77 ↑ 3.42	34.90 ↑ 3.69	27.25 ↑ 1.02	55.78 ↑ 1.96	115.37 ↑ 11.33	21.31 ↑ 1.48

of one combination of these three strategies, where $x = 1$, $y = 1$, and $z = 1$ respectively represent the use of the beam search, AoA mechanism, and the self-critical sequence training methods. For example, P_{110} denotes that the beam search and

AoA mechanism are applied together, while P_{011} represents that the AoA mechanism and the self-critical sequence training method are applied together.

We further applied various combinations of different strategies on every basic model to construct some model variants. For the FC model, since it does not support the attention mechanism, only P_{101} (that is, beam search and the self-critical sequence training) is applicable. Accordingly, one model variant was constructed from the FC model. For the other three basic models, four different combinations of these strategies are applicable (namely, P_{101}, P_{110}, P_{011}, and P_{111}), and thus four model variants were constructed from each of them. At last, these model variants were evaluated on the dataset.

Results: Table VI reports the evaluation results of 13 model variants employing some combination of the strategies or methods applied on the decoder part. Noted that for each model variant, its relevant models employing one of these strategies have already been evaluated and investigated in the previous RQs, we thus compare it with the one exhibiting the best performance in order to report the performance improvement achieved via the application of combined strategies (the performance improvement is shown in Table VI). Based on these results, we make the following observations:

1) ***The combination of various strategies helps to improve the performance of image captioning models in most cases.*** Table VI shows that the captioning performance is improved in most cases (i.e., most of the evaluated metrics for most models) after using the combination strategy on the target image captioning models. Although the application of the combination strategy to some models leads to a decrease in their captioning performance (e.g., the Att2in2 model with the usage of the beam search and self-critical sequence training method), the decrease is relatively small.

2) ***Different combinations of methods have different effects on the performance enhancement.*** As can be seen from the Table VI, the model performance improvement is different for different models using the same combination of strategies, and the model performance improvement is also different for the same model using different combinations of strategies. Nevertheless, it is observed that for the three models to which variuos combinations of strategies have been applied, they exhibit the best performance with P_{111}. That is, by applying the beam search, the AoA mechanism, and the reinforcement learning based training method together, these models perform better than those equipped with only parts of these strategies.

RQ4 : For encoder-decoder based image captioning models, applying various strategies to the decoder is helpful for improving the overall captioning performance. For the image captioning models under investigation, they exhibit the best performance with the use of the beam search, AoA mechanism and the reinforcement learning based training method.

VI. Conclusion

In this work, we focus on the impact of various aspects of the decoder on image captioning. In order to understand the impact of the text generation technique employed by the decoder on the results, we have conducted an extensive empirical analysis involving three different language models, two different decoding strategies, and two different training methods. The results of the research and analysis show that different language models have different impacts on the performance of the generated subtitles. Meanwhile, the use of two different decoding strategies as well as the training method of reinforcement learning helps to improve the model performance. In addition, it was found that using a combination of these strategies is usually better than using only a single strategy in image subtitle generation tasks. Future research directions can consider expanding our research to more complex datasets, especially exploring in cross-cultural environments. In addition, further research on how to integrate other machine learning technologies, such as transfer learning, to further improve model performance is also an important direction. The development of these future works will help expand our research and have a broader impact.

Acknowledgments

This work was supported by the National Nature Science Foundation of China (Grant No.61802349, No. 62132014, and No. 61972359), the Zhejiang Provincial Natural Science Foundation of China (Grant No. LY20F020021), and the Zhejiang Provincial Key Research and Development Program of China (No.2022C01045)

References

[1] Y.-T. Chen, F. Chen, M. Cooper, and D. Joshi, "Using business-aware latent topics for image captioning in social media," in *2016 IEEE International Conference on Multimedia and Expo (ICME)*, 2016, pp. 1–6.

[2] Q. Wang, W. Huang, X. Zhang, and X. Li, "Word–sentence framework for remote sensing image captioning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 12, pp. 10 532–10 543, 2020.

[3] R. C. Luo, Y.-T. Hsu, Y.-C. Wen, and H.-J. Ye, "Visual image caption generation for service robotics and industrial applications," in *2019 IEEE International Conference on Industrial Cyber Physical Systems (ICPS)*, 2019, pp. 827–832.

[4] C. Yin, B. Qian, J. Wei, X. Li, X. Zhang, Y. Li, and Q. Zheng, "Automatic generation of medical imaging diagnostic report with hierarchical recurrent neural network," in *2019 IEEE international conference on data mining (ICDM)*, 2019, pp. 728–737.

[5] M. Stefanini, M. Cornia, L. Baraldi, S. Cascianelli, G. Fiameni, and R. Cucchiara, "From show to tell: A survey on deep learning-based image captioning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 1, pp. 539–559, 2022.

[6] M. Z. Hossain, F. Sohel, M. F. Shiratuddin, and H. Laga, "A comprehensive survey of deep learning for image captioning," *ACM Computing Surveys (CsUR)*, vol. 51, no. 6, pp. 1–36, 2019.

[7] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4651–4659.

[8] C. Wang, H. Yang, C. Bartz, and C. Meinel, "Image captioning with deep bidirectional lstms," in *Proceedings of the 24th ACM international conference on Multimedia*, 2016, pp. 988–997.

[9] J.-B. Delbrouck and S. Dupont, "Bringing back simplicity and lightliness into neural image captioning," *arXiv preprint arXiv:1810.06245*, 2018.

[10] S. Herdade, A. Kappeler, K. Boakye, and J. Soares, "Image captioning: Transforming objects into words," *Advances in neural information processing systems*, vol. 32, 2019.

TABLE VI. EVALUATION RESULTS OF MODELS APPLYING MULTIPLE (COMBINED) STRATEGIES. BY COMPARING EACH MODEL (THAT APPLY AT LEAST TWO TYPES OF STRATEGIES TOGETHER) WITH THE BEST ONE APPLYING ONLY ONE OF SUCH STRATEGIES, THE IMPROVEMENT ACHIEVED VIA COMBINATIONAL USAGE OF VARIOUS STRATEGIES IS REPORTED (↑ AND ↓ REPRESENT THE POSITIVE AND NEGATIVE IMPROVEMENTS, RESPECTIVELY

	FC	Att2in2				Up-Down				Transformer			
	$P_{(101)}$	$P_{(101)}$	$P_{(110)}$	$P_{(011)}$	$P_{(111)}$	$P_{(101)}$	$P_{(110)}$	$P_{(011)}$	$P_{(111)}$	$P_{(101)}$	$P_{(110)}$	$P_{(011)}$	$P_{(111)}$
BLEU1	77.40	78.46	76.52	78.49	78.57	79.86	76.54	79.80	80.09	80.15	76.85	80.47	80.75
	↑0.03	↑0.00	↑0.70	↑0.03	↑0.12	↑0.24	↑0.79	↑0.18	↑0.48	↑0.70	↓0.44	↑1.01	↑1.29
BLEU4	34.83	36.09	36.36	36.30	36.43	37.60	35.86	37.54	37.93	38.76	37.18	38.97	39.19
	↑0.14	↑0.01	↑2.28	↑0.23	↑0.36	↑0.36	↑2.04	↑0.30	↑0.69	↑0.72	↑2.01	↑0.93	↑1.15
METEOR	26.71	27.29	27.55	27.44	27.47	28.16	27.71	28.30	28.36	28.79	28.03	28.97	29.00
	↑0.06	↓0.09	↑0.36	↑0.06	↑0.09	↑0.05	↑0.42	↑0.19	↑0.25	↑0.25	↑0.34	↑0.43	↑0.46
ROUGE	56.27	57.10	56.54	57.23	57.27	57.94	56.59	58.09	58.22	58.36	57.16	58.79	58.88
	↑0.11	↓0.04	↑0.69	↑0.09	↑0.14	↑0.11	↑0.61	↑0.26	↑0.40	↑0.42	↑0.36	↑0.85	↑0.94
CIDEr	113.50	118.75	113.17	119.76	119.94	124.88	113.86	124.91	125.52	127.33	116.58	127.41	128.85
	↑0.46	↓0.38	↑2.32	↑0.63	↑0.81	↑0.31	↑2.66	↑0.34	↑0.95	↑3.35	↑1.53	↑3.44	↑4.87
SPICE	20.05	20.78	20.41	20.86	20.86	21.56	20.80	21.81	21.88	22.55	21.08	22.45	22.58
	↑0.06	↓0.09	↑0.08	↓0.01	↓0.01	↑0.04	↑0.34	↑0.29	↑0.36	↑0.35	↑0.21	↑0.24	↑0.37

[11] M. Freitag and Y. Al-Onaizan, "Beam search strategies for neural machine translation," *arXiv preprint arXiv:1702.01806*, 2017.

[12] 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.

[13] A. Pal, S. Kar, A. Taneja, and V. K. Jadoun, "Image captioning and comparison of different encoders," in *Journal of Physics: Conference Series*, vol. 1478. IOP Publishing, 2020, p. 012004.

[14] V. Atliha and D. Šešok, "Comparison of vgg and resnet used as encoders for image captioning," in *2020 IEEE Open Conference of Electrical, Electronic and Information Sciences (eStream)*, 2020, pp. 1–4.

[15] S. Katiyar and S. K. Borgohain, "Comparative evaluation of cnn architectures for image caption generation," *arXiv preprint arXiv:2102.11506*, 2021.

[16] V. Sri Neha, B. Nikhila, K. Deepika, and T. Subetha, "A comparative analysis on image caption generator using deep learning architecture—resnet and vgg16," in *Computational Vision and Bio-Inspired Computing: Proceedings of ICCVBIC 2021*. Springer, 2022, pp. 209–218.

[17] M. S. Alam, M. S. Rahman, M. I. Hosen, K. A. Mubin, S. Hossen, and M. Mridha, "Comparison of different cnn model used as encoders for image captioning," in *2021 International Conference on Data Analytics for Business and Industry (ICDABI)*, 2021, pp. 523–526.

[18] Gaurav and P. Mathur, "Empirical study of image captioning models using various deep learning encoders," in *International Conference on Machine Intelligence and Signal Processing*. Springer, 2022, pp. 305–316.

[19] K. R. Suresh, A. Jarapala, and P. Sudeep, "Image captioning encoder–decoder models using cnn-rnn architectures: A comparative study," *Circuits, Systems, and Signal Processing*, vol. 41, no. 10, pp. 5719–5742, 2022.

[20] S. Takkar, A. Jain, and P. Adlakha, "Comparative study of different image captioning models," in *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*, 2021, pp. 1366–1371.

[21] P. P. Khaing *et al.*, "Attention-based deep learning model for image captioning: a comparative study," *International Journal of Image, Graphics and Signal Processing*, vol. 11, no. 6, p. 1, 2019.

[22] P. Dandwate, C. Shahane, V. Jagtap, and S. C. Karande, "Comparative study of transformer and lstm network with attention mechanism on image captioning," *arXiv preprint arXiv:2303.02648*, 2023.

[23] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. 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*. Springer, 2014, pp. 740–755.

[24] 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.

[25] 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.

[26] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.

[27] 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.

[28] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "Spice: Semantic propositional image caption evaluation," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*. Springer, 2016, pp. 382–398.

[29] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. 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 recognition*, 2018, pp. 6077–6086.

[30] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3156–3164.

[31] 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.

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[33] A. Graves and A. Graves, "Long short-term memory," *Supervised sequence labelling with recurrent neural networks*, pp. 37–45, 2012.

[34] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7008–7024.

[35] L. Guo, J. Liu, X. Zhu, P. Yao, S. Lu, and H. Lu, "Normalized and geometry-aware self-attention network for image captioning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 327–10 336.

[36] C. S. NagaDurga and T. Anuradha, "Attention-based comparison of automatic image caption generation encoders," in *Advances in Micro-Electronics, Embedded Systems and IoT: Proceedings of Sixth International Conference on Microelectronics, Electromagnetics and Telecommunications (ICMEET 2021), Volume 1*. Springer, 2022, pp. 157–167.

[37] C. L. Chowdhary, A. Goyal, B. K. Vasnani *et al.*, "Experimental assessment of beam search algorithm for improvement in image caption generation," *Journal of Applied Science and Engineering*, vol. 22, no. 4, pp. 691–698, 2019.

[38] Z. C. Lipton, J. Berkowitz, and C. Elkan, "A critical review of recurrent neural networks for sequence learning," *arXiv preprint arXiv:1506.00019*, 2015.

[39] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[40] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[41] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

[42] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3128–3137.