

RBBA: ResNet - BERT - Bahdanau Attention for Image Caption Generator

Duc-Hieu Hoang <i>Faculty of Electrical & Electronics Eng. Ton Duc Thang University Ho Chi Minh City, Vietnam 41702056@student.tdtu.edu.vn</i>	Duc Ngoc Minh Dang* <i>Dept. of Computing Fundamental FPT University Ho Chi Minh City, Vietnam ducnm2@fe.edu.vn</i>	Hanh Dang-Ngoc <i>School of Electrical and Data Engineering University of Technology Sydney Sydney, NSW, Australia hanh.n.dang@student.uts.edu.au</i>
Anh-Khoa Tran <i>Faculty of Electrical and Electronics Ton Duc Thang University Ho Chi Minh City, Vietnam trananhkhoa@tdtu.edu.vn</i>	Phuong-Nam Tran <i>Dept. of Computing Fundamental FPT University Ho Chi Minh City, Vietnam namtpse150004@fpt.edu.vn</i>	Cuong Tuan Nguyen <i>Dept. of Information Technology Specialization FPT University Ho Chi Minh City, Vietnam cuongnt64@fe.edu.vn</i>

Abstract—In recent years, the topic of image caption generators has gained significant attention. Several successful projects have emerged in this field, showcasing notable advancements. Image caption generators automatically generate descriptive captions for images through the encoder and decoder mechanisms. The encoder leverages computer vision models, while the decoder utilizes natural language processing models. In this study, we aim to assess a comprehensive set of seven distinct methodologies, including six existing methods from prior research and one newly proposed. These methods are trained and evaluated with bilingual evaluation (BLEU) on the Flickr8K dataset. In our experiments, the proposed ResNet50 – BERT – Bahdanau Attention model outperforms the other models in terms of the BLEU-1 score of 0.532143 and BLEU-4 score of 0.126316.

Index Terms—Deep learning, Natural language processing, Encoder-Decoder, Flickr8K, BLEU, Image Caption.

I. INTRODUCTION

Generating textual descriptions or captions for images is one of the most challenging tasks for artificial intelligence (AI). Despite the difficulty, image caption generators have a wide variety of uses, from providing automatic image descriptions for the blind to enhancing image search outcomes and producing more interesting social media posts. The development of more precise and sophisticated image caption generators has advanced significantly in recent years, and they are now widely used across a variety of industries.

Image caption generators analyze the contents of an image using deep learning techniques and generate a description of what is happening in the image. Typically, image caption generators employ a combination of computer vision techniques to process the images and natural language processing (NLP) algorithms to generate the descriptions. These algorithms can be trained on enormous datasets composed of photos and their captions, teaching the system how to correlate various visual cues with relevant descriptions.

* Corresponding author: Duc Ngoc Minh Dang (ducnm2@fe.edu.vn)

II. RELATED WORKS

A. Convolutional neural networks

Convolutional neural networks (CNNs) have a crucial role in computer vision. The emergence of architectures such as VGGNet [1] and ResNet [2] has significantly enhanced the efficacy of computer vision models across various tasks.

VGGNet [1] is a well-known convolutional neural network architecture consisting of multiple layers with small convolutional filters. It gained popularity for its simplicity and effectiveness in image classification tasks. The network architecture typically follows a consistent pattern of stacking convolutional layers with 3x3 filters, followed by max-pooling layers to reduce the spatial dimensions. The VGGNet architecture offers various configurations, commonly known as VGG16 and VGG19, depending on the depth of the network. Its architecture has demonstrated impressive performance on various image classification benchmarks, achieving high-performance results and establishing itself as a reliable and effective choice for deep-learning tasks.

ResNet [2] is a groundbreaking convolutional neural network architecture that has revolutionized the field of computer vision. ResNet addresses a common challenge encountered in deep neural networks known as the vanishing gradient problem. As networks become deeper, the gradients can vanish, leading to difficulties in training and optimization. To address this issue, ResNet utilizes skip connections that allow the network to bypass certain layers. By doing so, ResNet enables the direct flow of information from earlier layers to subsequent layers, facilitating the learning process. ResNet has achieved remarkable success, outperforming previous models in various computer vision tasks, such as image classification, object detection, and image segmentation. Its deep architecture, with variants like ResNet-50, ResNet-101, and ResNet-152, has become a standard benchmark in the field. Furthermore, ResNet's impact extends beyond computer vision, the concept

of residual learning has inspired advancements in other domains, including natural language processing and audio signal processing.

B. Natural language processing

The field of NLP includes various activities, such as text generation, sentiment analysis, language translation, speech recognition, and natural language comprehension. Thanks to advancements in machine learning and deep learning techniques, NLP has experienced remarkable progress in recent years. With the appearance of Long Short-Term Memory networks [3] (LSTMs), Gated Recurrent Unit networks [4] (GRUs), and especially Transformer [5], the field of NLP witnessed significant advancements. These breakthroughs in sequence modeling and language understanding have revolutionized various NLP tasks, including machine translation, text generation, sentiment analysis, and question-answering.

LSTMs and GRUs are designed to address the vanishing gradient problem in traditional recurrent neural networks (RNNs). They employ gating mechanisms that enable the networks to selectively update and forget information over time, allowing them to capture long-range dependencies in sequential data. LSTMs have been widely used in various NLP applications, and GRUs, a simplified variant of LSTMs, have gained popularity due to their computational efficiency.

However, it was the introduction of the Transformer model in 2017 that truly revolutionized the field of NLP. The Transformer [5] model introduced a novel architecture based solely on self-attention mechanisms, doing away with recurrent connections entirely. This architecture enabled parallel processing of input sequences, making it highly scalable and efficient. The self-attention mechanism in Transformers allows the model to weigh the importance of different words or tokens within a sequence when processing each word. This attention mechanism provides a global context for each word, enabling the model to capture dependencies between words regardless of their position in the sequence. The use of self-attention also reduces the vanishing gradient problem, as information can flow directly from any word to any other word in the sequence.

C. Image caption generators

Image caption generators are essential tools that help improve accessibility for individuals with visual impairments. These systems automatically create descriptive captions for images, allowing visually impaired users to better understand visual content shared online, including posts on social media, articles in the news, and web pages. In image caption generators, two common approaches are used to generate descriptive captions for images: merged models and injected models [6]. The merged models combine image features with NLP techniques to generate captions. It processes the image through CNNs to extract visual features, which are then merged with textual features in a subsequent neural network to generate the caption. On the other hand, injected models incorporate the image features directly into RNNs that

generate the caption. Instead of merging the features at a later stage, the image features are injected into the RNNs during the caption generation process, influencing the output at each time step. Both approaches have their advantages and trade-offs, and their performance can vary depending on the dataset and specific requirements of the image captioning task.

In merged models, the recurrent neural networks (RNNs) never directly interact with the image feature vectors or any derived vector from the image. Instead, the image is added to the language model after the RNNs have encoded the full prefix. This architecture is known as late-binding, where the image representation remains constant throughout the decoding process and is not changed at each time step.

In injected models, the image feature vector or a derived vector from the image serves as input to the RNNs in parallel with the word feature vectors of the caption prefix, such that either RNNs take two separate inputs, or the word feature vector is combined with the image feature vector into a single input before being passed to the RNNs. The image feature vector does not have to be identical for every word, nor does it need to be associated with each word. This mixed binding architecture allows for some flexibility in the image representation. However, if the same image is repeatedly provided to the recurrent neural networks (RNNs) at each time step, modifying the image representation becomes more challenging as the RNNs' hidden state is refreshed with the original image during each iteration.

III. IMAGE CAPTION GENERATOR

A. Merge-based Xception – Word2Vec (MXW2V)

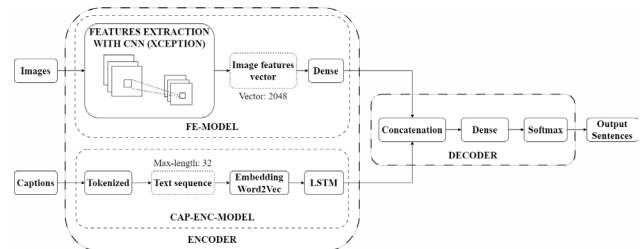


Fig. 1. Architecture of “Merge Xception - Word2Vec” method

Merge-based model is not exposed to the image feature vector at any point. Instead, the image is processed by CNNs and introduced into the language model after the prefix has been encoded by the RNNs in its entirety. This is a late binding architecture and it does not modify the image representation with every time step. The system architecture MXW2V is made up of three sub-models: the feature extraction model (FE-MODEL), the caption encoding model (CAP-ENC-MODEL), and finally the merged information decoding model. Xception architecture [7] is employed in the FE-MODEL, while the Word2Vec [8] technique is employed in the CAP-ENC-MODEL. The merged information decoding model simply concatenates both the feature extraction model and caption encoding model and forwards to a dense layer

using the ReLU activation function. Softmax is used as the activation function to predict the output word. The details of MXW2V are depicted in Figure 1.

B. Merge-based InceptionResnetV2 – GloVe (MIRG)

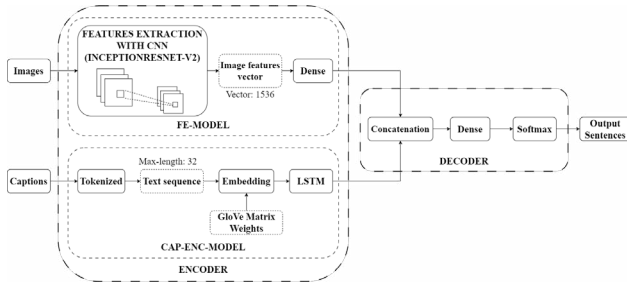


Fig. 2. Architecture of “Merge Inception ResnetV2 - GloVe” method

The MIRG employed in this approach builds upon the MXW2V, which utilized Xception and Word2Vec. However, it introduces notable improvements by leveraging the power of the InceptionResNetV2 [9] and the GloVe [10] technique whose architecture is illustrated in Figure 2. InceptionResNetV2 [9] is an exceptionally deep network, comprising a total of 164 layers. It combines the innovative ideas of the Inception module with the residual connections. These residual connections reduce the vanishing gradient problem commonly encountered in deep networks and facilitate the training of highly complex models. GloVe [10] technique changed the generation of word feature vectors by utilizing global word co-occurrence data. By capturing the semantic relationships between words, GloVe effectively merges the advantages of count-based methods. GloVe effectively merges the advantages of count-based Latent Semantic Analysis [11] and context-based Word2Vec [8].

C. Inject-based Xception – Word2Vec (IXW2V)

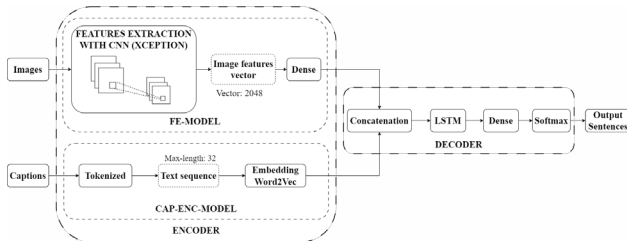


Fig. 3. Architecture of “Inject Xception - Word2Vec” method

The inject architecture, similar to the merge architecture, incorporates the combination of image characteristics and caption words into RNNs. In this approach, each caption word is processed alongside the image features, creating a new representation of the image for different parts of the phrase as it is generated. The IXW2V architecture specifically utilizes the inject architecture and is constructed based on it. Figure 3 illustrates the architecture of the IXW2V model. In IXW2V,

the output of the image and text features is concatenated together and then passed through LSMTs [3]. By feeding the concatenated representation through LSTMs, the model can effectively learn the relationships between the image and text features and capture the contextual information necessary for predicting the next word in the sentence.

D. Inject-based InceptionResnetV2 – GloVe (IIRG)

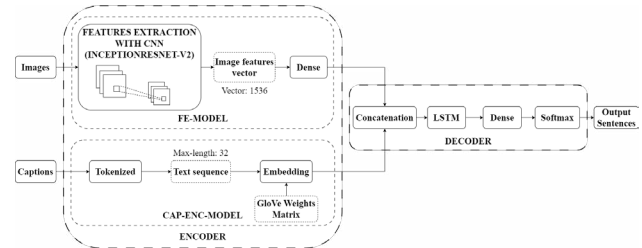


Fig. 4. Architecture of “Inject InceptionResnetV2 - GloVe” method

Similar to the IXW2V approach, this architecture leverages the InceptionResnetV2 [9] and GloVe [10] methods to extract features from images and text, respectively. These features are subsequently concatenated and fed into LSTMs to predict the next word. Figure 4 illustrates the architecture of the IIRG model.

E. VGG16 – GRU – Bahdanau attention (VGBA)

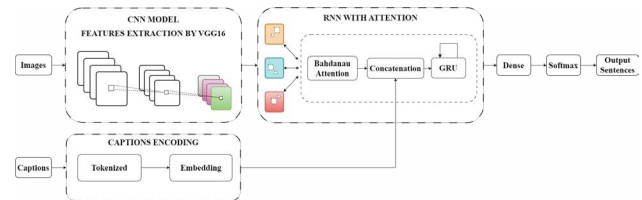


Fig. 5. Architecture of “VGG16 - GRU - Bahdanau attention” method

VGBA is based on the pioneering work of [12], which leverages an advanced encoding and decoding framework. Figure 5 illustrates the details of the VGBA architecture. The encoder incorporates the VGG16 [1] for visual feature extraction and the tokenized encoding method for caption processing. By including these components, the encoder effectively processes the input data. In the decoder section, both the GRU [4] network and the Bahdanau attention mechanism [13] are employed to enhance the output. The Bahdanau attention mechanism has demonstrated significant performance improvements. Its core concept involves assigning attention weights to prioritize specific feature vectors within the input sequence. These attention weights inform the decoder about the level of attention each input word should receive at different stages of decoding. By utilizing a set of attention weights, the decoder can focus on the most relevant portion of the image, guided by the alignment scores computed by a feed-forward neural network. This attention mechanism enables the

input and output sequences to concentrate on the most crucial elements, resulting in improved performance.

F. Xception – TT-LSTMs with Bi-LSTMs (XBi-LSTM)

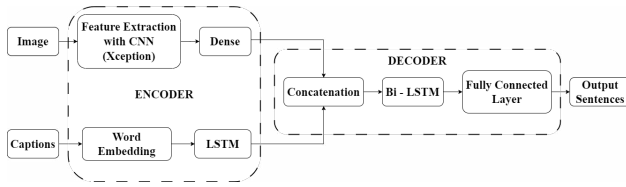


Fig. 6. Architecture of “Xception - TT-LSTMs with Bi-LSTMs” method

The encoder-decoder structure in Figure 6 is used by the approach known as TT-LSTM [14] which is built using a combination of the merge and inject models. For both text and image, TT-LSTM suggests creating two sub-encoder models. The two aforementioned procedures will then be merged. The Xception network is utilized in the image encoder model. Bi-LSTM is employed for the decoder, while LSTMs are applied to the language encoder.

G. ResNet50 – BERT – Bahdanau attention (RBBA)

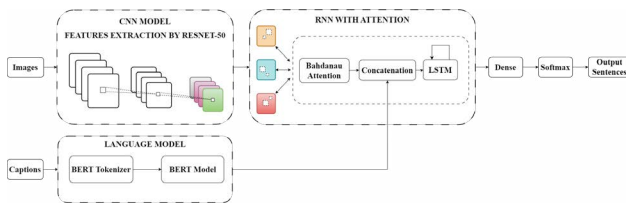


Fig. 7. Architecture of “ResNet50 - BERT - Bahdanau attention” method

This architecture utilizes the ResNet50 [2] and BERT [15] methods, the same approach as mentioned in VGBA, to extract features from images and text, respectively. In this architecture, the Bahdanua attention mechanism is combined with LSTMs [3] to generate the final output. The detailed architecture is illustrated in Figure 7.

IV. PERFORMANCE EVALUATION

A. Experiment setup

In this research, we utilized the Flickr8K dataset, which is a widely used and diverse dataset for image captioning tasks. The dataset consists of 8,000 high-quality images sourced from the popular online platform Flickr. These images cover a wide range of life themes, including captivating scenes of animals such as dogs, cats, and people engaging in various activities. The dataset also includes images depicting fun and entertainment activities, sports events, and daily life routines.

The diversity of themes and subjects within the Flickr8K dataset makes it a suitable choice for training and testing the image caption generator models. By incorporating such a diverse collection of images, we aimed to enhance the model’s ability to generate accurate and meaningful captions for a wide range of visual content.

To assess the performance of our model, we employed the BLEU metric [16], which is a widely used and established evaluation measure in the field of natural language processing. BLEU is commonly utilized to evaluate the quality of machine-generated text by comparing it to one or more reference translations or human-generated text. It calculates a score ranging from 0 to 1, with a higher score indicating a better match between the generated text and the reference text. The BLEU metric takes into account various factors such as precision, n-gram matches, and brevity penalty. It considers both the presence and the ordering of words, thereby capturing the fluency and correctness of the generated captions. By employing the BLEU metric, we aimed to quantitatively evaluate the performance of our model and compare it with other methods in the field. The use of this standard metric allows for a fair and objective assessment of the quality of the captions generated by our model.

We carefully selected specific parameter configurations to optimize our research outcomes. These include utilizing the Adam optimizer with a learning rate of 0.001 to guide the training process. For the cost function, we employed the cross-entropy loss function, a commonly used measure for multi-class classification tasks. To initialize the LSTMs/GRUs, we applied the glorot-uniform initializer, which helps ensure effective information flow within the model. During training, we used a batch size of 32, which determines the number of samples processed in each iteration. The model was trained for a total of 10 epochs. To prevent overfitting and enhance generalization, we incorporated a dropout rate of 0.5, which randomly deactivates a portion of the neural network units during training. This regularization technique encourages the model to learn more robust and generalized features.

B. Experiment results

Tables I and II present the experimental results of various architectures for image captioning. Among these architectures, the VGBA demonstrates exceptional performance in terms of both speed and accuracy. With a relatively short training time of 826.8s, this architecture achieves the best score in terms of BLEU-2 and BLEU-3 (Table I). The inclusion of Bahdanau attention yields a significant enhancement in the output, as it allows the model to comprehend the image context more accurately and consistently.




Table II illustrates these methods in action for captioning the image. Among them, the RBBA attention stands out as it describes the image with the highest level of detail and accuracy compared to the other methods. It effectively leverages the ResNet50 and BERT models, along with the inclusion of Bahdanau attention, resulting in more precise and comprehensive image captions.

Figure 8 demonstrates the performance of various methods in generating image captions, as measured by the BLEU scores. The bar chart clearly illustrates the distinct differences in performance among these methods. Notably, the Bahdanau attention mechanism still stands out as a particularly effective approach. The Bahdanau attention mechanism has proven to be

TABLE I
PERFORMANCE COMPARISON ON FLICKR8K DATASET

Methods	Parameter	Time training (per epochs)	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Inject Xception – Word2Vec	5,002,649	807.6 s	0.364886	0.190213	0.121723	0.046814
Merge Xception – Word2Vec	5,002,649	732.8 s	0.375879	0.196083	0.118743	0.042923
Inject InceptionResnetV2 – GloVe	5,512,165	3,387.1 s	0.360748	0.201883	0.116986	0.062958
Merge InceptionResnetV2 – GloVe	5,315,557	2,096.8 s	0.381128	0.221911	0.155406	0.069800
VGG16 – GRU – Bahdanau attention	4,872,345	826.8 s	0.461538	0.339683	0.254815	0.039086
Xception – TT_LSTM with Bi-LSTM	8,517,273	3,139.8 s	0.426407	0.265075	0.191381	0.089676
ResNet50–BERT–Bahdanau attention	119,818,810	7,488.0 s	0.532143	0.227003	0.175572	0.126316

TABLE II
CAPTION COMPARISON RESULTS OF METHODS FOR IMAGES

Test image	Real caption	Method	Predict caption
	<ol style="list-style-type: none"> girl with a black swimsuit plays in the sprinkler young girl is playing in fountain of water young girl plays in fountain water little girl crouches to splash fountain water young girl in a bathing suit playing with water shooting out of the ground 	Xception – Word2Vec (Inject)	young boy is playing in the water
		Xception – Word2Vec (Merge)	young girl is playing in the water
		InceptionResnetV2 – GloVe (Inject)	young girl in pink shirt is playing into the water
		InceptionResnetV2 – GloVe (Merge)	girl in bathtub spits water from water fountain
		VGG16 – GRU – Bahdanau attention	young girl playing in fountain water
		Xception – TT_LSTM with Bi-LSTM	boy in blue shirt is playing in the water
		ResNet50 – BERT – Bahdanau attention	the little girl wearing the swimsuit is playing into the water fountain
	<ol style="list-style-type: none"> black dog jumps up to catch white ball dog catches ball in the air dog catches toy outside of brick house dog leaps to catch ball the big black dog is jumping up in the air to catch ball 	Xception – Word2Vec (Inject)	dog is running through the grass
		Xception – Word2Vec (Merge)	dog is running through the grass
		InceptionResnetV2 – GloVe (Inject)	dog is jumping over hurdler
		InceptionResnetV2 – GloVe (Merge)	black dog is running through the grass
		VGG16 – GRU – Bahdanau attention	black dog jumps up to catch the ball
		Xception – TT_LSTM with Bi-LSTM	black dog jumps to catch ball
		ResNet50 – BERT – Bahdanau attention	dog catch ball outside
	<ol style="list-style-type: none"> boy plays basketball boy wearing blue shorts is bouncing basketball in front of the net little boy plays with basketball and toy basketball hoop little boy playing basketball in the grass the child in the blue shorts dribbled the basketball 	Xception – Word2Vec (Inject)	two men are playing in the grass
		Xception – Word2Vec (Merge)	two men are playing in the grass
		InceptionResnetV2 – GloVe (Inject)	the man in the red shirt is playing the basketball
		InceptionResnetV2 – GloVe (Merge)	the basketball player in the red strip is trying to get the ball
		VGG16 – GRU – Bahdanau attention	boy is holding basketball
		Xception – TT_LSTM with Bi-LSTM	the boy is playing basketball
		ResNet50 – BERT – Bahdanau attention	little boy wearing blue short is playing basketball in the grass

highly successful in generating accurate and contextually relevant image captions. It encompasses a sophisticated attention mechanism that allows the model to focus on different regions of the image while generating the corresponding captions.

V. CONCLUSIONS

In conclusion, the comparison of these methods sheds light on the advantages and trade-offs inherent in different model and attention mechanism combinations. The RBBA approach excels in generating accurate and descriptive captions, which is demonstrated by its impressive BLEU-1 and BLEU-4 scores, as well as the meaningfulness of the captions produced for sample images. On the other hand, although the VGBA

approach falls short of the RBBA approach in terms of BLEU-1 and BLEU-4 scores, it outshines the leading model when it comes to BLEU-2 and BLEU-3 scores. In terms of speed, the VGBA model appears to be significantly faster than the RBBA model. VGBA model achieves the training of an epoch within only 826.8 seconds, whereas the RBBA model, on the other hand, takes up to 7,488 seconds.

REFERENCES

- [1] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.
- [2] 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 (CVPR)*, 2016, pp. 770–778.

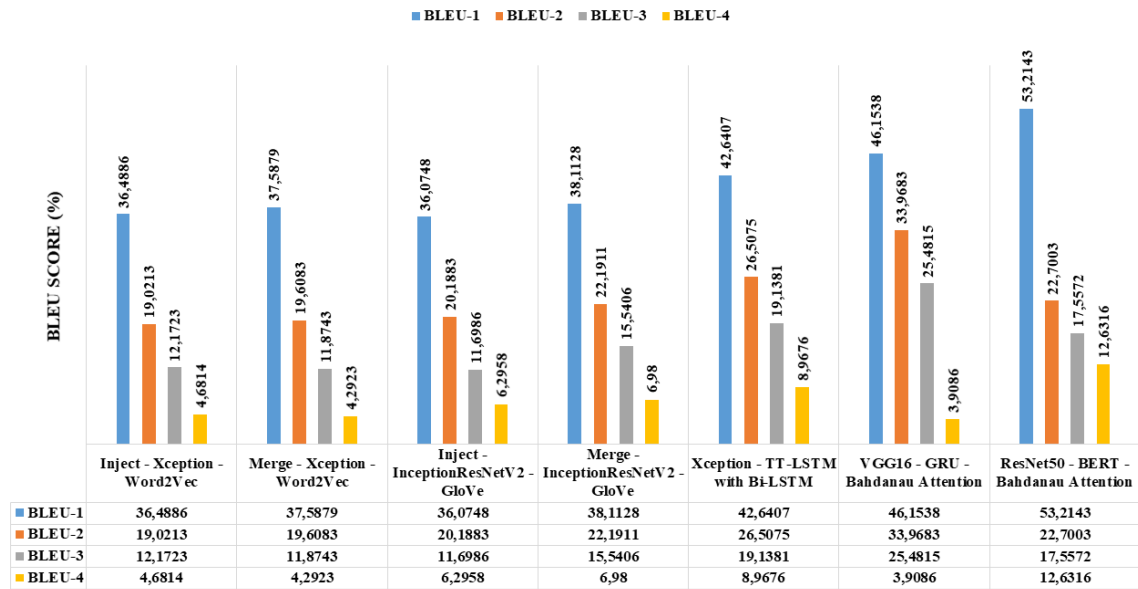


Fig. 8. BLUE-score comparison on Flickr8K dataset

- [3] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [4] 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.
- [5] 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.
- [6] M. Tanti, A. Gatt, and K. P. Camilleri, "Where to put the image in an image caption generator," *Natural Language Engineering*, vol. 24, no. 3, pp. 467–489, 2018.
- [7] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.
- [8] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.
- [9] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 31, no. 1, 2017.
- [10] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.
- [11] S. T. Dumais, "Latent semantic analysis," *Annual Review of Information Science and Technology (ARIST)*, vol. 38, pp. 189–230, 2004.
- [12] 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.
- [13] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.
- [14] P. P. Khaing *et al.*, "Two-tier lstm model for image caption generation," *International Journal of Intelligent Engineering & Systems*, vol. 14, no. 4, 2021.
- [15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.
- [16] 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.