



A Comparative Evaluation of Image Caption Synthesis Using Deep Neural Network

Sadia Nasrin Tisha^(✉), Md Shahidur Rahaman, and Pablo Rivas

Department of Computer Science, School of Engineering and Computer Science,
Baylor University, Waco, TX, USA
{Sadia_Tisha1,MdShahidur_Rahaman1,Pablo_Rivas}@Baylor.edu

Abstract. Image caption generation is a crucial challenge in deep learning and natural language processing, involving identifying the context of an image and providing appropriate captions. In this study, we aimed to evaluate and compare the performance of two different model architectures using pre-trained CNN models for image classification and sequential LSTM models for caption generation. Specifically, we used ResNet50 and inceptionV3 CNN models with word2Vec and GloVe word embeddings, respectively, to generate captions. We evaluated the models based on two criteria: calculating the BLEU score for each generated caption and comparing the BLEU score with the inceptionResNetV2 state-of-the-art model. Our results showed that the second model architecture with inceptionV3 and GloVe-based model outperformed the first model and closely followed the benchmark BLEU score of the state-of-the-art model. Therefore, our study provides evidence that the choice of pre-trained CNN model and word embedding technique can significantly impact the performance of image caption generation, with the proposed architecture offering an accurate and efficient solution.

Keywords: Image caption generation · ResNet50 · InceptionV3 · LSTM · GloVe · Word2vec · BLEU score

1 Introduction

Generating image captions is a fundamental challenge in computer vision and natural language processing. It involves creating descriptions for images that express the objects and scenes present in them and how those objects and backgrounds relate to one another. Recent advancements in machine learning, particularly in convolutional and recurrent neural networks, have significantly improved the quality of caption generation [3, 15]. The most successful models to date have been based on pre-trained convolutional neural networks, such as VGG-Net and inceptionResNetV2, and employ GloVe for word embedding [8].

However, selecting an appropriate configuration for generating accurate and efficient captions remains a significant challenge. This project aims to evaluate two different model architectures for image caption generation and provide

insight into which model is most appropriate to achieve state-of-the-art BLEU scores [10]. We compared the performance of two pre-trained CNN models, resNet50 and inceptionV3, for image classification, combined with long short-term memory (LSTM) sequential models as decoders. For the first model, we used ResNet50 CNN-based image feature extraction, word2vec word embedding, and LSTM to generate image captions. Finally, we employed the Inception-V3 model with GloVe word embedding and LSTM for sequential caption generation for the second model. We then evaluated both models based on their BLEU scores and compared the performance of the best-performing model with that of the state-of-the-art model, InceptionResNetV2.

The choice of an efficient model for generating image captions that balance execution time and accuracy is crucial. This study provides valuable insights into image caption generation's encoding and decoding phases and helps select the appropriate configuration for this task. With the availability of large classification datasets, such as COCO [4], Flickr [6], and Nocaps [1], and advanced training in deep neural networks, the generation of accurate and efficient captions for images has become more feasible. The results of this study demonstrate that the second model architecture, based on inceptionV3 and GloVe word embedding, is more effective in generating accurate and efficient image captions than the first model architecture.

The main contributions of our paper can be summarized as follows:

- Demonstration of the effectiveness of different image classification and word embedding techniques in image caption generation models.
- Identification of a more effective model, Model 2, which utilizes the inceptionV3 CNN model and GloVe word embedding technique to generate accurate and meaningful image captions.
- Findings that suggest the choice of image classification and word embedding techniques significantly impact the accuracy of image caption generation models.

2 Related Work

Recently, researchers extensively evaluated image caption generation using encoder-decoder-based models. Vinyals et al. [15] propose a neural and probabilistic framework that generates descriptions from images. In this study, the author encoded the variable length input into a fixed dimensional vector using an RNN model and obtained the desired output sentence by decoding the representation. The author uses arg max likelihood to increase the likelihood during training. Here, the author extracts the LSTM function and updates the memory block using a nonlinear function. The author employed CNN for image classification to feed a rich vector representation of the picture to RNN, which serves as a decoder. CNN is an innovative way to batch normalization. However, the researchers needed to provide which CNN model could best fit for encoding. Instead, the evaluation focused mainly on several datasets, e.g., Pascal [16], Flickr30k [6], COCO [4], and SBU [9].

Kesavan et al. [7] proposed a deep neural network to generate the caption with a pre-trained VGG-16 model. The researchers used CNN to create the thought vector for image extraction, which GRU uses as an RNN. The authors provide the state-of-the-art CNN configuration and generate the performance using the BLEU score. However, the model evaluation for the configuration is missing, as the aim was to achieve the state-of-the-art BLEU score from the image captions. In addition, Chen et al. [20] suggest a recurrent neural network to learn the bi-directional mapping between images and their sentence-based descriptions. Using a novel recurrent visual memory that automatically retains to recall long-term visual notions, they enable the development of innovative phrases given an image and utilize it to help in sentence generation and the reconstruction of visual features. These involve image retrieval, sentence production, and sentence retrieval. This paper presented state-of-the-art outcomes for creating innovative graphic descriptions for evaluation, where humans favor their automatically generated captions more frequently than 19.8% of the time when compared to captions that humans made. For techniques utilizing similar visual cues, performance on the picture and sentence retrieval tests is superior to or on par with state-of-the-art findings.

So we provided a comprehensive model analysis where the best-configured model can generate state-of-the-art captions, which will be measured by calculating BLEU scores.

3 Methodology

3.1 Dataset Description

We have used Flickr8K dataset [11]. The Flickr8k dataset is one of the standard datasets used in various image-related tasks. With 8,000 images paired with five different English captions that provide in-depth explanations of the key components and events, this dataset provides a collection of sentence-based image descriptions and searches. We added each image to five different English sentences that describe the image. We preprocessed the dataset by removing the noise so the model could detect the patterns easily in the text data. Here we cleaned the text's special characters, such as hashtags, punctuation, and numbers. The total dataset has a size of 1 GB. For training our model, we have divided our dataset into the training, testing, and validation sets, where 80% of images are in the training set, 10% in the testing set, and 10% in the validation set.

3.2 Model Description

In this project, we have generated a caption from an image using two model architectures. We have used convolutional neural networks (CNN) for image classification and a sequential LSTM model for developing captions. In the first architecture, we encoded the images using the CNN-based feature extraction architecture ResNet 50 and used word2vec embedding to obtain a vector representation for each corresponding sentence word. The LSTM layers would take

partial captions-generated vectors as input and output the following word in the caption sequence. In the second architecture, we encoded the images using a pre-trained CNN model (inceptionV3) and utilized GloVe embedding to produce a vector representation for each corresponding text word. The LSTM layers would receive the partially captioned vectors and output the following word in the caption sequence. The next section is a description of our two models and their architecture:

Model 1: We have generated our model architecture with ResNet 50 CNN-based architecture for image classification in the first model. We retrieved the image's features in this instance just before the final classification layer. A 2048-bit vector is created by converting an additional dense layer. For word vectorization, we have used word2vec word embedding techniques. We tokenized 8253 unique words from the training dataset to define the vocabulary. As computers do not understand English words, we have represented them with numbers and mapped each vocabulary word with a unique index value. We encoded each word into a fixed-sized vector and defined each word as a number. Then we used the LSTM layers that take the partial caption-generated vectors as input and the image feature vectors as output, and they output the following word in the caption sequence for each test image. Figure 1 shows the total architecture of Model 1.

Figure 2 shows the model architecture where we used a $224 \times 224 \times 3$ dimensional image as the first input layer. For better classification, we used max-pooling for extracting the image feature from the vector and padding the vectors. ReLu activation is used in LSTM to decode the vectors with the softmax activation function. Here, we used the categorical cross-entropy loss function to calculate the loss, and for a 0.2 dropout rate, we got the minimum error in our training set. After successfully training the model, we test the model by providing test image data as input, which gives us the upcoming word in the caption sequence.

Model 2: In this model, we have used InceptionV3 CNN-based architecture for image classification and encoded the image into a feature vector. Then we used the vector for the input layer of the sequential LSTM model. InceptionV3 is a pre-trained model that extracts image features using three different sizes (e.g., 1×1 , 3×3 , 5×5) convolution and one max pooling, which classify the images with deeper layers. We utilized Glove embeddings for encoding, creating a vector representation of each sentence's words. GloVe stands for global vectors for word representation that generates word embeddings by aggregating a global word-word co-occurrence matrix from a corpus [14]. LSTM layers receive vector image features and vocabulary inputs and predict the test image's caption as an output. Figure 4 shows the total architecture of model 2.

Figure 3 shows the model architecture. Here, we have used $299 \times 299 \times 3$ dimensional image as the input layer. We have done the max pooling with a

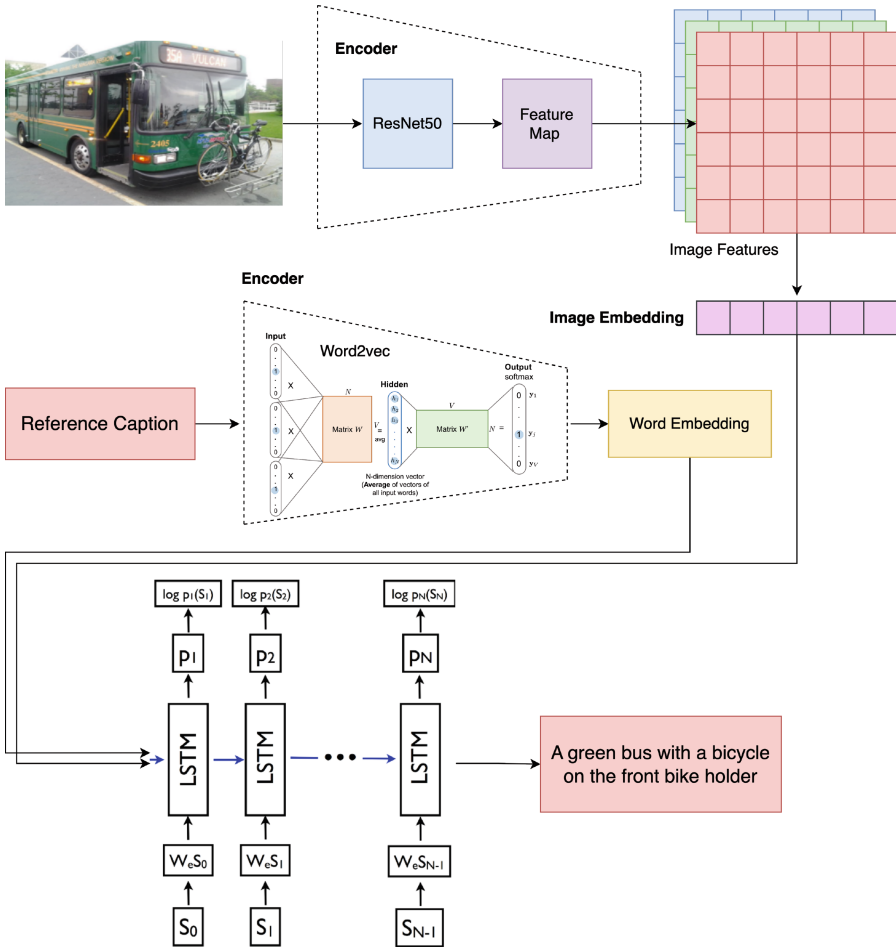


Fig. 1. Overall model architecture of Model 1.

2048 dense layer for extracting the image feature to a vector with batch normalization for better training. For word embedding, we have used Glove, with a total vocabulary size of 8763 with 256 dense layers and a 0.2 dropout rate. In addition, reLu activation and softmax activation functions are used in LSTM to decode the vectors, which improves the model error and training loss. Finally, after training, we put the model to the test by giving it test image data as input, and the model generates sequential sentences as captions.

3.3 Methods

Convolutional Neural Networks (CNN) for Image Classification: Convolutional neural networks are specialized deep neural networks that can process

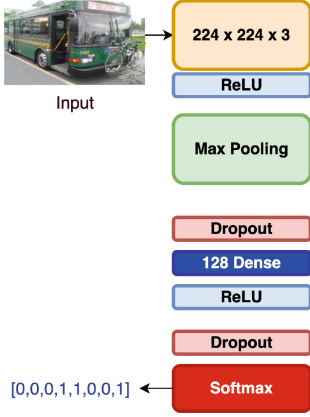


Fig. 2. Model architecture of ResNet50.

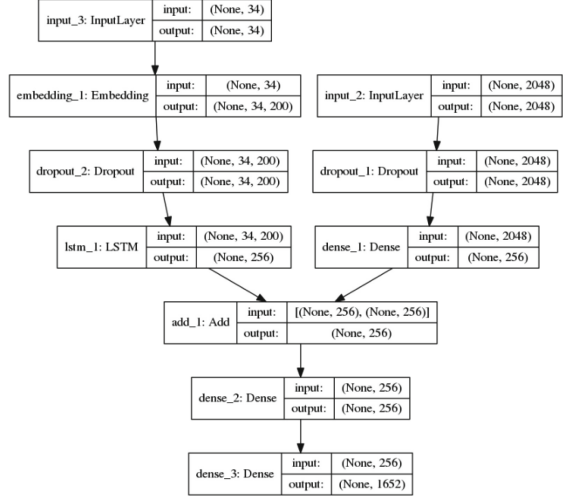


Fig. 3. Model architecture of model 2.

data in the form of a 2D matrix as input. It scans images from left to right and top to bottom to extract significant elements before combining them to classify them [17]. In addition, we have used CNN-based feature extraction architectures, ResNet 50, and Inception v3 architectures in this project.

1. **ResNet50:** The encoder step of image feature extraction using CNN has a specific expression ability, which is crucial in determining the quality of the image caption model. ResNet50 introduced a residual learning framework that is simple to optimize and has a low computing impact. We calculate the residual error to design and handle degradation and gradient problems, improving the network's performance as the depth grows. The model takes an image and produces a caption encoded as a sequence of $1 - k$ coded words.

$$y = y_1, y_2, y_3 \dots y_c, y_i \in R^k$$

Here, k is the size of the dictionary, and c is the caption length. We use CNN, particularly ResNet50, to obtain set annotation vectors like the feature vectors. The extractor produces L -vectors, all of which are a D -dimensional representation of the corresponding part of an image.

2. **Inception v3:** In image classification tasks, the researchers commonly utilized Inception v3. The Inception module typically contains three different sizes of convolution and one maximum pooling. The channel is aggregated after the convolution process for the preceding layer's network output, and then nonlinear fusion is conducted [18]. Finally, the Inception v3 network structure uses a convolution kernel splitting method to divide large volume integrals into small convolutions.

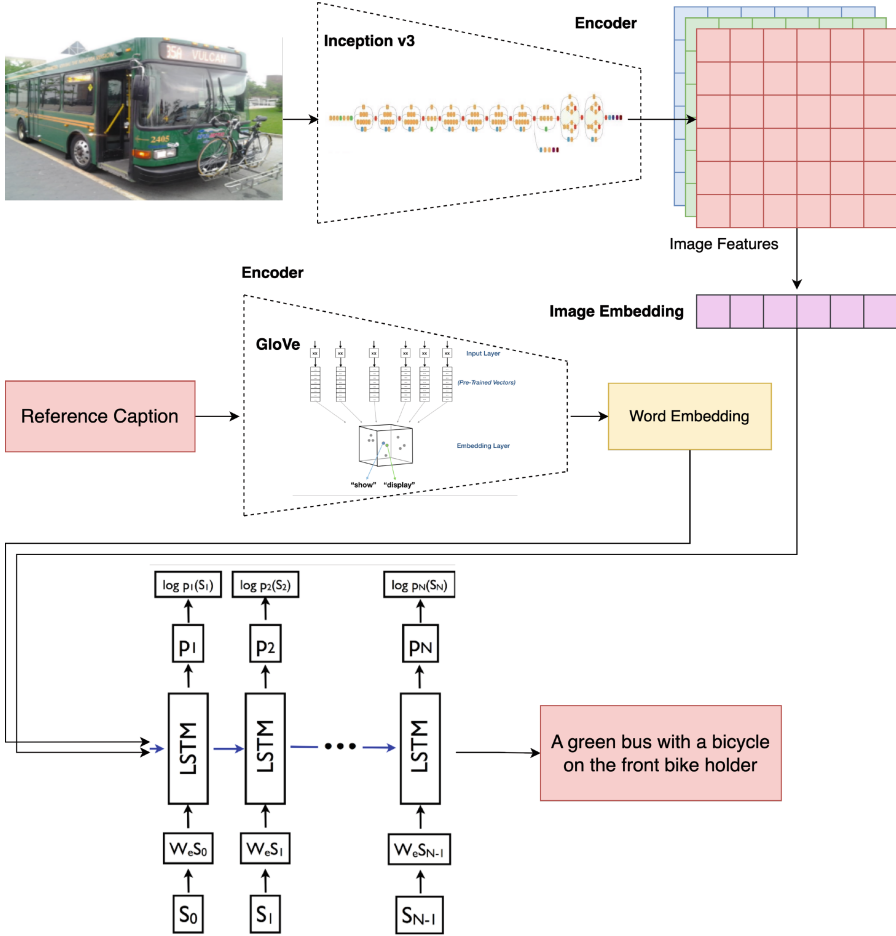


Fig. 4. Overall model architecture of model 2.

LSTM (Long Short-Term Memory): The Long Short-Term Memory Network (LSTMN) is an enhanced RNN (sequential network) that permits information to be stored indefinitely. LSTMs vary from standard feedforward neural networks in that they feature feedback connections. This trait allows LSTMs to process entire data sequences without having to handle each point separately instead of preserving necessary knowledge about primary data in the sequence to aid in processing incoming data points [19]. This project uses the LSTM language model to generate proper captions based on the input vector from the ResNet50 and inception v3 output.

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f)$$

The output vector of the previous cell h_{t-1} with the new element of the sequence x_t are concatenated and passed as one vector through the layer with the softmax activation function.

4 Result Analysis

4.1 Model Result

For Model 1, we have trained the model with 300 epochs. Here Fig. 5 and 6 show after 150 epochs, the loss function is minimized to 0 where the training error is 0, and we got 85% accuracy for training. Finally, we generate the caption of test image data. The image on Table 1 top row shows the output caption of the image generated by Model 1, where the caption tells the correct sentence which describes the image.

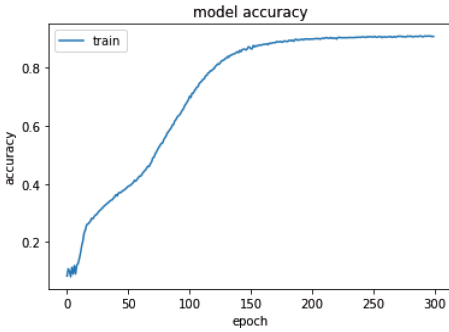


Fig. 5. The Accuracy graph of Model 1.

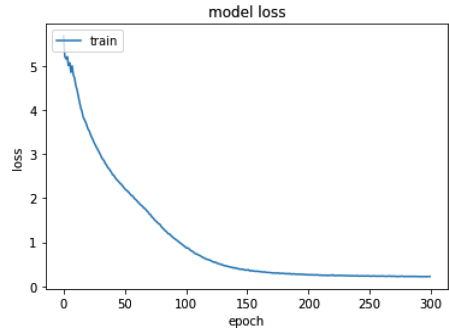


Fig. 6. The loss graph of Model 1.



In addition, for Model 2, we trained the model with 30 epochs, and Fig. 7 shows that the validation and training errors are minimized to 0. We got higher accuracy of 98% in both the training and validation sets. The image in the bottom row of Table 1, shows the output caption of the image generated by Model 2, which correctly identifies the image.

4.2 Model Evaluation

In this project, to evaluate the two models, first, we have calculated the BLEU score for each generated caption of the testing images and second, we compare the BLEU score with state-of-the-art model (e.g., inceptionResNetV2) results.

BLEU Score Calculation: We use the Bilingual Evaluation Understudy (BLEU) to evaluate the generated captions. The BLEU method is widely used to assess Natural Language Processing (NLP) systems that create language, particularly in natural language creation and machine translation [12]. The number in BLEU indicates the n-gram that the BLEU analyzes. This measure computes

Table 1. Captions generated by the two models

Model	Image	Generated caption
1		Two girls are lying upside down on a white bed.
2		Two football players are fighting over the ball.

the similarity score between produced and target text, which ranges from 0 to 1, with 1 indicating similarity and 0 indicating no resemblance [8]. In this project, for both Model 1 and Model 2, we have calculated the BLUE score for each generated caption for images with its reference captions from text data. Table 2 shows the BLEU score generated from Model 1, and the table also shows the BLUE score for Model 2. The table indicates that Model 2, with inception v3 image classification and GloVe word embedding model, gives a higher BLEU score than Model 1.

State-of-the-Art Model Evaluation: According to Keras application [13], inceptionResNetV2 is the most accurate image classification model. inceptionResNetV2 is another pre-trained model of convolutional neural architecture that builds on the Inception family of architectures but incorporates residual connections. Using our method, we utilized this pre-trained model and its image captions [2] and then computed the BLEU score for each caption. Finally, we compared the BLEU score with our calculated BLEU score for both models. Table 3 shows that, for this image, Model 2 gives a 0.6012 BLEU score, which is close to the inceptionResNetV2(state-of-the-art) model. On the other hand, Model 1 also shows a closer BLEU score than inceptionResNetV2.

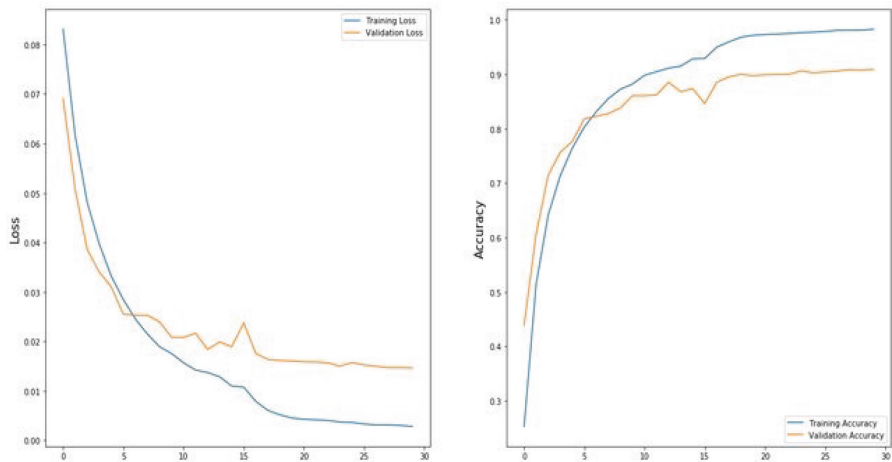




Fig. 7. The loss and accuracy graph of Model 2


Table 2. Captions generated by the two models

Image	Caption Model 1	BLEU	Caption Model 2	BLEU
	A boy in a life jacket is jacket is top on the side of a boat	0.44	Young boy in swim trunks is	0.66
	A boy players with a stringless racket in his backyard	0.40	Little girl in pink sweat-shirt is holding racket	0.50

5 Discussion

The results of this project suggest that the combination of the inceptionV3 CNN model and GloVe word embedding technique used in Model 2 is more effective in generating accurate and meaningful captions for images than Model 1. The higher accuracy of Model 2 can be attributed to using a more advanced image classification model, which can identify more intricate features in the images, as well as a more sophisticated word embedding technique. These findings suggest that the choice of image classification and word embedding techniques can significantly impact the accuracy of image caption generation models.

Table 3. Comparative analysis of the model with its caption and BLEU score

Image	Model	Generated Caption	BLEU
	Inception-ResNetV2 (State of the art)	Young boy jumps on bed	0.623
	ResNet50 (Model 1)	A boy off a green, shirt jumps	0.549
	InceptionV3 (Model 2)	Boy jumps off bed	0.601

Moreover, comparing the calculated BLEU score with state-of-the-art model results indicates that the generated captions of Model 2 are closer to the benchmark solution, suggesting that this model can outperform other state-of-the-art models. However, it is essential to note that the results may vary depending on the dataset and image types used.

Another interesting finding from this project is that even though Model 1 has a lower accuracy and BLEU score than Model 2, the generated captions are still relatively close to the benchmark solution. This suggests that even a less advanced image caption generation model can still generate meaningful captions for images.

6 Conclusions

In conclusion, this project demonstrates the effectiveness of using different image classification and word embedding techniques in image caption generation models. Model 2, which utilizes the inceptionV3 CNN model and GloVe word embedding technique, outperforms Model 1 in terms of accuracy and BLEU score. The results also suggest that the generated captions of Model 2 are close to the benchmark solution, indicating that this model has the potential to outperform other state-of-the-art models.

However, there is still room for improvement in image caption generation models, and further research is needed to explore the use of other image classification and word embedding techniques. Nonetheless, this project contributes to the growing field of image caption generation and provides insights into the effectiveness of different techniques in generating accurate and meaningful captions for images.

References

1. Agrawal, H., et al.: Nocaps: novel object captioning at scale. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 8948–8957 (2019)
2. Bhatia, Y., Bajpayee, A., Raghuvanshi, D., Mittal, H.: Image captioning using Google’s inception-resnet-v2 and recurrent neural network. In: 2019 Twelfth International Conference on Contemporary Computing (IC3), pp. 1–6. IEEE (2019)
3. Chen, M., Ding, G., Zhao, S., Chen, H., Liu, Q., Han, J.: Reference based LSTM for image captioning. In: Thirty-First AAAI Conference on Artificial Intelligence (2017)
4. García, C.: MS-COCO-ES: Spanish coco captions (2019). <https://www.kaggle.com/datasets/colmejano/mscoco-es-spanish-coco-captions>
5. Hossain, M.Z., Sohel, F., Shiratuddin, M.F., Laga, H.: A comprehensive survey of deep learning for image captioning. *ACM Comput. Surv. (CsUR)* **51**(6), 1–36 (2019)
6. HSANKESARA. Flickr image dataset flickr (2021). <https://www.kaggle.com/datasets/hsankesara/flickr-image-dataset>
7. Kesavan, V., Muley, V., Kolhekar, M.: Deep learning based automatic image caption generation. In: 2019 Global Conference for Advancement in Technology (GCAT), pp. 1–6. IEEE (2019)
8. Mahadi, M.R.S., Arifianto, A., Ramadhani, K.N.: Adaptive attention generation for Indonesian image captioning. In: 2020 8th International Conference on Information and Communication Technology (ICoICT), pp. 1–6. IEEE (2020)
9. Ordonez, V., Kulkarni, G., Berg, T.: Im2text: describing images using 1 million captioned photographs. In: *Advances in Neural Information Processing Systems*, vol. 24 (2011)
10. Papineni, K., Roukos, S., Ward, T., Zhu, W.-J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311–318 (2002)
11. Rashtchian, C., Young, P., Hodosh, M., Hockenmaier, J.: Collecting image annotations using amazon’s mechanical turk. In: Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk, pp. 139–147 (2010)
12. Reiter, E.: A structured review of the validity of bleu. *Comput. Linguist.* **44**(3), 393–401 (2018)
13. Keras Team. Keras documentation: Keras applications
14. Tifrea, A., Bécigneul, G., Ganea, O.-E.: Poincaré glove: hyperbolic word embeddings. *arXiv preprint [arXiv:1810.06546](https://arxiv.org/abs/1810.06546)* (2018)
15. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: a neural image caption generator. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3156–3164 (2015)
16. ZARAK. Pascal voc 2007 (2018). <https://www.kaggle.com/datasets/zaraks/pascal-voc-2007>
17. Chauhan, R., Ghanshala, K.K., Joshi, R.C.: Convolutional neural network (CNN) for image detection and recognition. In: 2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC), pp. 278–282. IEEE (2018)
18. Dong, N., Zhao, L., Chun-Ho, W., Chang, J.-F.: Inception V3 based cervical cell classification combined with artificially extracted features. *Appl. Soft Comput.* **93**, 106311 (2020)

19. Yong, Yu., Si, X., Changhua, H., Zhang, J.: A review of recurrent neural networks: LSTM cells and network architectures. *Neural Comput.* **31**(7), 1235–1270 (2019)
20. Chen, X., Zitnick, C.L.: Learning a recurrent visual representation for image caption generation. arXiv preprint [arXiv:1411.5654](https://arxiv.org/abs/1411.5654) (2014)