Image Captioning Using Xception-Long Short-Term Memory

Nisha Panchal[1,2(✉)] and Dweepna Garg[1]

[1] Department of Computer Engineering, Devang Patel Institute of Advance Technology and Research (DEPSTAR), Faculty of Technology and Engineering (FTE), Charotar University of Science and Technology (CHARUSAT), Changa, Anand, Gujarat, India
nishpanchal132@gmail.com

[2] U & P U Patel Department of Computer Engineering, Chandubhai S. Patel Institute of Technology (CSPIT), Faculty of Technology and Engineering (FTE), Charotar University of Science and Technology (CHARUSAT), Changa, Anand, Gujarat, India

Abstract. Deep Learning has shown great potential in developing applications capable of automatically generating captions or descriptions for images and video frames. The critical components of this process are image processing and natural language processing, which play a crucial role in captioning images and videos. These applications can be used in several areas, including robotic vision systems, assisting people with visual impairments, generating metadata for search engines, answering visual questions, visual grounding, and more. This paper discusses various working algorithms such as a combination of CNN-RNN, encoder-decoder, attention mechanisms, and transformation models with evaluation matrices, datasets, and limitations of existing models. Xception-LSTM shows great potential compared to the traditional encoding-decoding model using BLEU and METEOR evaluation matrics.

Keywords: Image captioning · Xception · Long-short term memory (LSTM) · CNN (convolution neural networks) · RNN (Recurrent neural networks)

1 Introduction

Image processing is an essential aspect of computer science and has substantial relevance across various fields, including object detection, scene interpretation, and visual recognition. Dedicated hardware was used by researchers for executing imaging techniques to get appropriate results, especially for rigid objects before the emergence of deep learning. However, CNN and RNN driven by deep learning have important influences on visual-to-text generation, demonstrating remarkable progress recently.

The task of describing a scenario depicted in an image or video clip comes naturally to humans, but it poses significant challenges for machines. To tackle this issue, computer scientists are exploring methods to integrate the ability to comprehend human language with the capability to automatically extract and analyze visual data, thereby enabling machines to perform similar tasks. Although, extracting objects with their actions from

S. Kadry and R. Prasath (Eds.): MIKE 2023, LNAI 13924, pp. 25–33, 2023.
https://doi.org/10.1007/978-3-031-44084-7_3

the image and producing crisp as well as relevant sentences needs much substantial work in comparison to a simple image recognition task.

Image and video caption generation primarily involve analyzing an image's features and generating a corresponding textual description. As this field demands visual and textual mastery proficiency, it utilizes a blend of CV and NLP techniques to translate image comprehension from feature vectors into words arranged in the proper sequence. The captioning method must capture the objects in the given scenario as well as their traits, actions, and interrelationships.

Therefore, the most common method for image captions is the encoder-decoder architecture, which combines a Convolutional Neural Network (CNN) to encode image features and a Recurrent Neural Network (RNN) to generate a caption.

It has a clear separation of tasks – The CNN is responsible for encoding the image features, while the RNN is responsible for generating the caption. This separation of tasks makes it easier to debug and analyze the model.

Overall, the Encoder-Decoder architecture is a popular choice for image captioning due to its effectiveness, flexibility, simplicity, and clear separation of tasks.

This paper mentions the following details in upcoming sections, which are related work of image captioning, the Proposed Methodology, Results, and Discussion, and at the end conclusion and future work.

2 Related Work

Our research has involved an in-depth exploration of numerous studies about image captioning, encompassing a range of techniques, datasets, and evaluation methodologies. CNN is often used to extract features from an image. These features are then used as input to a language model that creates the image caption. CNNs are trained on large image datasets and can learn to recognize patterns and features in images [1, 3–5, 7, 8, 12, 15, 17]. RNN takes as input the output of the previous step (which is a word embedding) and the visual features that the CNN extracted from the image. And then generates the next word in the caption [3, 8].

Encoder-decoder models are a type of neural network architecture that leverages an encoder component to extract features from an input and a decoder component to generate an output. In the context of image captioning, the encoder is typically implemented using a convolutional neural network (CNN), which extracts salient features from the input image. On the other hand, the decoder is usually implemented using a recurrent neural network (RNN), which generates the caption based on the features extracted by the encoder. [2, 17]. This innovative approach has served as a starting point for subsequent research in the area of image captioning in 2015 [18].

Subsequently, the author of [19] introduced a novel approach to simultaneously train a CNN and an RNN for generating captions by aligning image regions with their corresponding linguistic units. To facilitate their experimentation, the authors employed the COCO dataset, which has since emerged as a widely accepted benchmark for assessing the effectiveness of image captioning models. Of significance, this paper also introduced the CIDEr score, a widely used evaluation metric for image captioning models [19].

After that, an attention-based approach to image captioning was introduced [20], where the model learns to selectively attend to different image regions when generating captions. The authors showed that attention mechanisms improve caption quality and reduce ambiguity, and proposed a novel "hard" attention mechanism that can be trained using backpropagation [20]. Attention mechanisms play a vital role in enabling image captioning models to focus on the most pertinent aspects of an image during caption generation. Rather than solely depending on the global image features, attention mechanisms allow these models to selectively concentrate on specific regions of the image that are most relevant to the current context of the caption being generated [2, 5, 13].

Thereafter bottom-up and top-down attention mechanism combines object-level features with region-level features to generate captions introduced [21]. This paper introduced a new dataset called Visual Genome, which contains more detailed object and attribute annotations than other datasets used for image captioning. This paper also introduced a new evaluation metric called SPICE, designed to measure the semantic similarity between generated and human captions [21].

Afterward, a new pre-training approach for image captioning that combines vision and language tasks to learn joint representations of images and captions was introduced [22]. The authors of this paper use a Transformer-based architecture that is pre-trained on a large corpus of image-caption pairs and shows that their method achieves state-of-the-art performance on several benchmark datasets.

Following that, a new Transformer-based architecture for image captioning that uses a meshed-memory mechanism to selectively attend to different regions of the image and the caption was introduced [23]. The authors show that their method outperforms other Transformer-based models and achieves state-of-the-art performance on the COCO dataset. Transformer Models Transformers are a relatively new development in the field of natural language processing and have proven to be highly successful in tasks such as text generation and machine translation. This is mainly because they use a self-attention mechanism that allows them to process input sequences simultaneously, making them well-suited for processing long input sequences such as captions. When creating captions, Transformer models are usually equipped with an encoder for extracting image features and a decoder for generating captions [11].

Visual Question answering is one of the major applications of image captioning that is mentioned in [24]. This paper proposes a new pre-training approach for image captioning using a single encoder to encode images and captions jointly.

A new approach was brought up in [25] for generating image captions by parallelizing the decoding process to improve efficiency. The authors propose a hierarchical structure for the caption that allows the model to generate the words in a parallel and efficient manner. The authors show that their method achieves state-of-the-art performance on the COCO dataset and is significantly faster than other models.

Throughout the years, numerous encoder-decoder techniques have emerged, employing different variations of Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN).

3 Methodology

A combination of CNN-LSTM is a commonly used Neural network in image captioning [3, 5–9, 12, 13, 15, 17]. In the proposed model Xception model is used for feature extraction. The Xception takes an input image and outputs a vector of visual features as the following equation.

$$V = Xception(I) \tag{1}$$

In the given context, before utilizing the Xception model, a series of image preprocessing techniques were employed to adequately prepare the image. Consequently, the Xception model was applied to extract the feature vector associated with the image. In the current context, In Eq. (1) the variable "I" represents the input image, while "V" refers to the vector of visual features extracted from it.

The Xception model's utilization of depthwise separable convolutions, in contrast to traditional CNN models, delivers notable improvements in computational efficiency and speed. This architectural choice enables the model to analyze spatial relationships and feature interactions more effectively while minimizing redundant computations. Therefore, the Xception model is used for encoding features from the image compared to the traditional one.

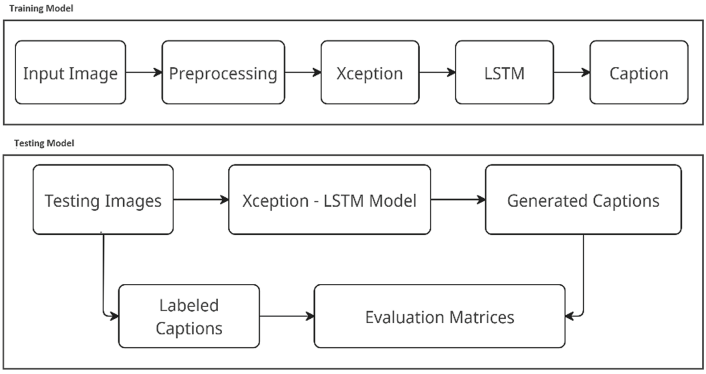

Fig. 1. Proposed Model

In the case of the decoding side, LSTM is a type of RNN that is frequently used for image captioning tasks because it is better suited for capturing long-term dependencies in sequential data such as natural language [3, 5–9, 12–13, 15, 17]. Figure 1 depicts the flow of the proposed model with training and testing bifurcations.

In image captioning, a critical task involves the model's ability to comprehend the context of the image and produce a fitting caption accordingly. To this end, the LSTM network is employed to process the image features extracted by an Xception. The LSTM network generates a sequential set of words, which are then combined to form a grammatically sound and semantically coherent sentence.

Compared to conventional RNNs, LSTM networks possess an added memory cell capable of preserving and retrieving information over extended periods. This feature

enables LSTMs to more effectively manage long-term dependencies, a challenge for traditional RNNs. Additionally; LSTMs overcome the issue of vanishing gradients that are commonly encountered in traditional RNNs, thereby increasing the effectiveness and efficiency of the network [5, 6, 9].

Moreover, LSTM networks can also selectively forget or remember information from the previous time step, making them well-suited for tasks where the model needs to maintain a context for a long period.

Therefore, we have used LSTM networks over traditional RNNs for image captioning tasks because of their ability to better capture the complex dependencies and long-term context of natural language data.

The LSTM takes the vector of visual features from the Xception and generates a sequence of words that form the image caption. The LSTM does this by processing each word in the sequence one at a time and updating its internal state based on the previous words in the sequence. This can be represented in Eq. (2)

$$h_t = LSTM\left(V, h_{\{t-1\}}\right) \tag{2}$$

$$y_t = Softmax\left(W_{\{hy\}h_t} + b_y\right) \tag{3}$$

where V is the visual features at time t, h_t is the internal state of the LSTM at time t, y_t is the output probability distribution over the vocabulary at time t, and $W_{\{hy\}}$ is the weight matrix connecting the LSTM output to the vocabulary, and b_y is the bias term.

The Softmax function is used to convert the output of the LSTM to a probability distribution over the vocabulary so that the network can predict the next word in the sequence based on the probability of each possible word as per Eq. (3).

In this paper, we have used the Flickr8k dataset. It contains 8000 images [26], each with five different captions provided by human annotators. The dataset is divided into the train, validation, and test sets, and is often used for evaluating image-captioning models.

Figure 2 depicts the outcome of our proposed model where we have used the start and end keywords to indicate the starting and ending of the captions.

4 Results and Discussion

Several evaluation methods are used for image captioning, including:

1. BLEU:
   It is an evaluation metric used in natural language processing to measure the quality of machine-generated translations. It compares the n-gram overlap between the sentences. BLEU scores range from 0 to 1, with higher scores indicating a better-quality translation. [2, 4, 5, 8, 11–17].
2. ROUGE:
   The ROUGE evaluation matrix consists of several metrics, including ROUGE-1, ROUGE-2, and ROUGE-L. ROUGE-1 calculates the overlap of unigrams (single words) between the generated and reference summaries. ROUGE-2 calculates the overlap of bigrams (pairs of adjacent words), while ROUGE-L measures the longest common subsequence between the generated and reference summaries [8, 10, 12, 15, 16].

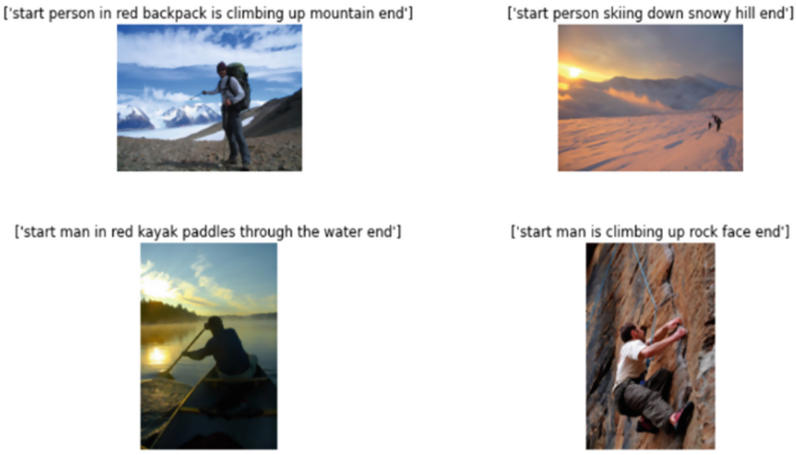

Fig. 2. Results of Xception-LSTM model on Flicker30K dataset

3. METEOR:

   It uses a combination of unigram precision, recall, and alignment-based metrics to evaluate the similarity between the sentences. METEOR is designed to handle nuances of natural language such as synonyms, paraphrases, and word order variations, [5, 6, 8, 10, 12, 13, 15–17].

   It should be emphasized that a holistic assessment of image captioning systems cannot rely solely on a single metric. Instead, a blend of multiple evaluation techniques is usually employed to achieve a more comprehensive and precise evaluation of image captioning system performance.

The Xception-LSTM model evaluates the quality of the captions generated using BLEU and METEOR matrices shown in Fig. 3. BLEU evaluation works on n-gram overlapping words where the n value changes from 1 to 4. Based on the value of n results decrease. With that METEOR works on the ordering of the generated caption compare to the labeled caption.

5 Limitation

Despite significant advancements in the field of image captioning in recent years, there remain several challenges and issues that require attention and resolution. Here are a few examples:

- Context – It can be challenging to generate an accurate caption that conveys the intended message of an image. A single image can be perceived in different ways, leading to ambiguity in generating a descriptive caption. For example, a picture of a person riding a bicycle might be captioned differently depending on the specific details in the picture. The caption could vary from “a man riding a bicycle”, “a woman riding a bicycle” or “men riding a vehicle” depending on the contextual information in the image.

Evaluation matrices	Results
BLEU-1	85.4
BLEU-2	67.6
BLEU-3	51.0
BLEU-4	38.8
METEOR	50.8

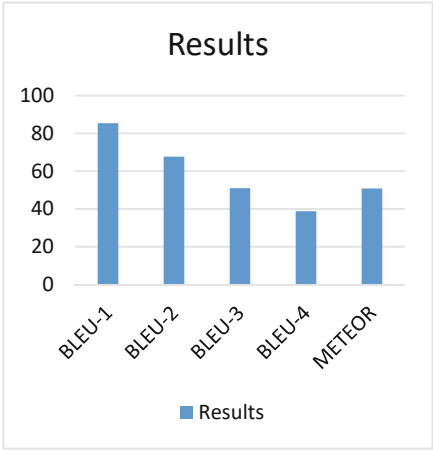

Fig. 3. Evaluation matrices on Flicker30k dataset using Xception-LSTM

- Ambiguity – Creating a precise and descriptive caption for an image can be a difficult task due to the ambiguity and subjectivity of visual content. Images can be interpreted in various ways, making it difficult to generate a single caption that accurately represents the content. Additionally, there may be more than one valid caption for a single image due to the different interpretations that people may have.
- Rare or Unseen Words – In some cases, image captioning models may generate captions that contain infrequent or unfamiliar words, making them difficult for people to comprehend. This issue can be particularly troublesome for individuals who do not have expertise in the language utilized in the caption.
- Data Bias – The process of training image captioning models involves using extensive datasets of image-caption pairs. However, these datasets may occasionally exhibit a bias towards particular types of images or captions. Consequently, the trained model may generate less accurate or less descriptive captions for certain types of images due to this bias.
- Evaluation – Assessing the quality of image captions lacks a single standard metric, and the suitability of various metrics varies based on the specific application. For instance, certain metrics may prioritize accuracy, whereas others may prioritize the diversity or originality of the generated captions.

6 Conclusion and Future Work

Compared to other models like VGG-LSTM and ResNet-LSTM, the Xception-LSTM model offers several advantages. For one, it boasts greater computational and memory efficiency, which makes it more suitable for training on larger datasets. Additionally, the LSTM-based language decoder employed by the Xception-LSTM model is capable of modeling long-term dependencies during the caption generation process. This is a crucial factor in generating coherent and semantically meaningful captions. Moreover, the Xception-LSTM model can be fine-tuned on other tasks such as visual question answering and image retrieval, which demonstrates its versatility and effectiveness in

various applications. However, the Xception-LSTM model still faces some challenges such as handling rare words and dealing with the ambiguity and diversity in the caption generation process. Future research can focus on addressing these challenges and improving the performance of the Xception-LSTM model on image captioning.

References

1. Mathews, A.: Captioning images using different styles. In: MM 2015 – Proceedings of the 2015 ACM Multimedia Conference, Oct 2015, pp. 665–668 (2015). https://doi.org/10.1145/2733373.2807998
2. Cho, K., Courville, A., Bengio, Y.: Describing multimedia content using attention-based encoder-decoder networks. IEEE Trans. Multimedia **17**(11), 1875–1886 (2015). https://doi.org/10.1109/TMM.2015.2477044
3. Donahue, J., et al.: Long-term recurrent convolutional networks for visual recognition and description. IEEE Trans. Pattern Anal. Mach. Intell. **39**(4), 677–691 (2017). https://doi.org/10.1109/TPAMI.2016.2599174
4. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. IEEE Trans. Pattern Anal. Mach. Intell. **39**(4), 664–676 (2017). https://doi.org/10.1109/TPAMI.2016.2598339
5. Gao, L., Guo, Z., Zhang, H., Xu, X., Shen, H.T.: Video captioning with attention-based LSTM and semantic consistency. IEEE Trans. Multimedia **19**(9), 2045–2055 (2017). https://doi.org/10.1109/TMM.2017.2729019
6. Yang, Y., et al.: Video captioning by adversarial LSTM. IEEE Trans. Image Process. **27**(11), 5600–5611 (2018). https://doi.org/10.1109/TIP.2018.2855422
7. Wu, Q., Shen, C., Wang, P., Dick, A., van den Hengel, A.: Image captioning and visual question answering based on attributes and external knowledge. IEEE Trans. Pattern Anal. Mach. Intell. **40**(6), 1367–1381 (2018). https://doi.org/10.1109/TPAMI.2017.2708709
8. Lu, X., Wang, B., Zheng, X., Li, X.: Exploring models and data for remote sensing image caption generation. IEEE Trans. Geosci. Remote Sens. **56**(4), 2183–2195 (2018). https://doi.org/10.1109/TGRS.2017.2776321
9. Han, M., Chen, W., Moges, A.D.: Fast image captioning using LSTM. Clust. Comput. **22**, 6143–6155 (2019). https://doi.org/10.1007/s10586-018-1885-9
10. Xu, N., et al.: Multi-level policy and reward-based deep reinforcement learning framework for image captioning. IEEE Trans. Multimedia **22**(5), 1372–1383 (2020). https://doi.org/10.1109/TMM.2019.2941820
11. Yu, J., Li, J., Yu, Z., Huang, Q.: Multimodal transformer with multi-view visual representation for image captioning. IEEE Trans. Circuits Syst. Video Technol. **30**(12), 4467–4480 (2020). https://doi.org/10.1109/TCSVT.2019.2947482
12. Turkerud, I.R., Mengshoel, O.J.: Image captioning using deep learning: text augmentation by paraphrasing via back translation (2021). https://doi.org/10.1109/SSCI50451.2021.9659834
13. Chen, N., et al.: Distributed attention for grounded image captioning. In: MM 2021 – Proceedings of the 29th ACM International Conference on Multimedia, Oct 2021, pp. 1966–1975. https://doi.org/10.1145/3474085.3475354
14. Mahalakshmi, P., Fatima, N.S.: Summarization of text and image captioning in information retrieval using deep learning techniques. IEEE Access **10**, 18289–18297 (2022). https://doi.org/10.1109/ACCESS.2022.3150414
15. Ji, J., Ma, Y., Sun, X., Zhou, Y., Wu, Y., Ji, R.: Knowing what to learn: a metric-oriented focal mechanism for image captioning. IEEE Trans. Image Process. **31**, 4321–4335 (2022). https://doi.org/10.1109/tip.2022.3183434

16. Bae, J.W., Lee, S.H., Kim, W.Y., Seong, J.H., Seo, D.H.: Image captioning model using part-of-speech guidance module for description with diverse vocabulary. IEEE Access **10**, 45219–45229 (2022). https://doi.org/10.1109/ACCESS.2022.3169781
17. Ramos, R., Martins, B.: Using neural encoder-decoder models with continuous outputs for remote sensing image captioning. IEEE Access **10**, 24852–24863 (2022). https://doi.org/10.1109/ACCESS.2022.3151874
18. 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 (CVPR), pp. 3156–3164 (2015)
19. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3128–3137 (2015)
20. Xu, K., et al.: Neural image caption generation with visual attention. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3121–3129 (2015)
21. Anderson, P., et al.: 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 (CVPR), pp. 6077–6086 (2018)
22. Li, X., Yin, X., Li, C., Hu, Z., Zhang, H., Sun, F.: VLP: vision-language pre-training for image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10976–10985 (2020)
23. Cornia, M., Stefanini, M., Baraldi, L., Cucchiara, R.: Meshed-memory transformer for image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1635–1644 (2020)
24. Lu, J., Batra, D., Parikh, D., Lee, S.: Unicoder-VL: a universal encoder for vision and language. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2238–2247 (2021)
25. Huang, L., Li, Y., Shen, J., Wu, J.: Parallel decoding of hierarchical structure for image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11168–11177 (2021)
26. Hodosh, M., Young, P., Hockenmaier, J.: Framing image description as a ranking task: data, models and evaluation metrics. J. Artif. Intell. Res. **47**, 853–899 (2013)