RESEARCH ARTICLE **SPECIAL ISSUE**

RSP Science Hub

International Research Journal on Advanced Science Hub
2582-4376
Vol. 05, Issue 05S May
www.rspsciencehub.com

Check for updates

http://dx.doi.org/10.47392/irjash.2023.S047

International Conference on intelligent COMPUting TEchnologies and Research (i-COMPUTER) 2023

Multilingual Image caption Generator using Big data and Deep Learning

Naman Grover [1], Anchita Singh [1], Suganeshwari G [2]

[1]School of Computer Science and Engineering, Vellore Institute of Technology Chennai, Tamil Nadu, India.
[2]Assistant Professor(Sr.),School of Computer Science and Engineering, Vellore Institute of Technology Chennai, Tamil Nadu, India.

Emails: naman.grover2020@vitstudent.ac.in, anchita.singh2020@vitstudent.ac.in, suganeshwari.g@vitstudent.ac.in

Article History

Received: 2 March 2023
Accepted: 22 March 2023

Keywords:

LSTM;
CNN;
Big data;
BELU Score;
Deep learning

Abstract

Automatic image captioning aims to produce a descriptive sentence about a picture. For this task, we are creating a model that will spit out an English sentence when an image is given as input describing the image's subject. Scientists in the field of cognitive computing have paid much attention to it in recent years. The endeavor is challenging because it requires merging ideas from two distinct but related disciplines: natural language processing and computer vision. Using the integration of CNN with LSTM, we developed a model for generating image captions. The ideas behind a Convolutional Neural Network and a Long Short-Term Memory model were combined to create this model. The convolutional neural network serves as the encoder, extracting information from images. At the same time, the long short-term memory is responsible for the decoder role, coming up with words to describe the image. The problem arises when the dataset is significant, and it takes weeks for systems to have only CPU support to train the network to decrease the time it is required to train big data can be taken into accounts. After the caption generation phase, we use BLEU Scores to assess our model's performance. Using this information, our technology can help users find a fitting description for the uploaded photo with the desired language.

1. Introduction

Image captioning is a difficult, time-consuming procedure that frequently matches irrelevant phrases with a particular image. Some previously difficult Machine Learning tasks have become simple because of Deep Learning and Neural Networks, text processing techniques like Natural Language Processing, and more. They have a wide range of uses in artificial intelligence, particularly in the classification, captioning, and image identification fields. The practice of automatically creating written descriptions of the incidents seen in an input image is known as image captioning. A caption is created by the model using an input image. Its ability to automatically generate captions for photographs is improving at the same rate as technology. This Image Captioning will be very helpful for many emerging technologies, including self-driving cars. Several Machine Learning tasks for recommendation systems can benefit from image labeling. Object recognition models, visual attention-based picture captioning, and image captioning with deep learning are a few models that have been suggested for image captioning. The Inception model, the VGG model, the ResNet-LSTM model, and the more conventional CNN- RNN model are the only

deep learning models that are currently accessible.

In this study, we'll work with a sizable dataset, discuss the captioning process, and show how big data can reduce the time complexity by several hours. They are based on the ResNet-LSTM model.

2. Related Work

Simao Herdade's method involves running a picture through an object detector, which allows Armin Kappeler (Herdade et al.) to extract appearance and geometry data from everything in the image. After that, the caption text is created using the Object Relation Transformer. Zhongliang Yang and Yu-Jin Zhang (Elamri, De Planque, et al.) proposed a method that automatically generates a natural language description of an image. Image understanding will be substantially improved by this system. The suggested multimodal neural network approach closely resembles the capability of the human visual system to learn how to characterize the content of images on its own. It has modules for object localization and detection that address the challenging and inherently sequential nature of LSTM units over time.

Under the constraints of this article, Oriol Vinyals' method (Vinyals et al.) suggests a neural and probabilistic framework for producing descriptions from images. Cutting-edge outcomes can be achieved by explicitly maximizing the likelihood of the accurate translation given an input sentence in an "end-to-end" manner, according to recent breakthroughs in statistical machine translation. Both training and logic are applicable here. Recent developments in statistical machine translation enabled this discovery.

The technique put out by Chaoyang Wang and Ziwei Zhou (Wang, Z. Zhou, L. Xu, et al.) is made up of two distinct functional components. First, picture feature data must be extracted, which primarily involves erasing object and location information from the image. The semantic information in the image description, which is created by combining the image features and semantic data, is examined in the following phase. It is a challenging task to automatically explain the substance of an image using English sentences that are correctly put together, as recommended by T. Planque and C. Elamri (Madankar, Chandak, Chavhan, et al.). Nonetheless, helping those who are blind could be necessary. With the ability to take pictures, modern cell phones can be helpful for visually challenged people to capture images of their surroundings. Here, captions for submitted photographs can be produced. Such captions may be audible to persons who are blind or visually challenged, improving their awareness of their surroundings. In this presentation, Christoper Elamri extracts features from a picture using a CNN model. Then, describe the environment using English sentences that are grammatically correct. These features are fed into an RNN or an LSTM model to provide a visual description.

In this method proposed by authors Kelvin Xu (K. Xu et al.) the encoder employs a convolutional neural network (CNN) to extract visual information, and the decoder uses a long short-term memory (LSTM) network to generate captions. They also added a visual attention mechanism that lets the decoder focus on different image regions when creating caption words. They train the model on a large collection of photos and captions and use beam search to produce captions for fresh images. The authors showed that their model beats state-of-the-art models on numerous criteria and that the visual attention mechanism generates more accurate and descriptive captions.

In this method proposed by authors Andrej Karpathy (Karpathy, Fei-Fei, et al.) A deep convolutional neural network (CNN) extracts visual information from the image, and a deep recurrent neural network (RNN) generates the natural language description. They also presented a novel alignment strategy that aligns visual characteristics and the language model at the word level, allowing the model to acquire a joint representation that correlates each word in the description with the relevant visual region in the image. They train the model on a large collection of photos and captions and use beam search to produce captions for fresh images. The authors showed that their model beats state-of-the-art models on numerous measures and that the acquired joint representations can be employed for picture retrieval and visual question answering.

The author Andre Araujo (Araujo, Carreira, Arandjelovic, et al.) suggested a method that ranks the most comparable instances in an input image to a query image. They extracted feature vectors from both the input and query images using a Faster

R-CNN object identification model, then computed the cosine similarity between the feature vectors to get a similarity score. The Faster R-CNN model is trained on a huge dataset of images and object annotations, then used to extract features for instance search. On many instance search benchmarks, the authors showed that Faster R-CNN features beat features from other object detection models.

The authors Peter Anderson, Xiaodong He, Chris Buehler (Anderson et al.) suggested a model that uses a bottom-up attention mechanism to extract a set of picture attributes at numerous spatial scales and a top-down attention mechanism to selectively attend to different portions of the image based on the generated caption or query. Top-down attention uses a recurrent neural network to dynamically attend to different image regions based on the current word or inquiry, whereas bottom-up attention uses a pre-trained object detector to select relevant image regions. They train the model on a huge dataset of photos and captions or questions and use beam search to create captions or responses for fresh images or queries. The authors showed that their model beats state-of-the-art models on several picture captioning and visual question-answering metrics and that bottom-up and top-down attention methods are complementary and can increase model performance.

The author's Steven J. Rennie, and Etienne Marcheret (Rennie et al.) suggested a caption model that samples words from a probability distribution conditioned on visual attributes. They then introduced self-critical training, which uses the model's captions as ground truth instead of human captions. The model generated captions for a set of photos, which they then evaluated using standard captioning assessment criteria (such as CIDEr or Bleu). Policy gradient approaches were utilized to update model parameters using the evaluation score as a reward signal.

The authors Jiasen Lu (Lu et al.) devised a methodology that creates captions by focusing on image regions relevant to the current term. They then incorporated a visual sentinel technique, which requires learning a second "sentinel" vector to dynamically govern attention. Based on the prior attention weights and the current word being formed, the sentinel vector modulates the following time step's attention weights. This lets the model "know when to look" at different image parts based on the caption. The authors showed that their model beats state-of-the-art methods on numerous image captioning measures and that the visual sentinel mechanism may learn to attend to different image regions based on the caption context.

The authors Ruiyu Li, Jiwen Lu (Li et al.)suggested a model to caption films using geographical and temporal information. The spatial attention system attends to video frames, while the temporal attention mechanism attends to video frame segments. A new Temporal Segment Network (TSN) architecture was presented to extract temporal characteristics from video segments.

The author Ronghang Hu (Hu et al.) suggested a methodology that creates image captions by letting users adjust length, style, and substance. A "control unit" conditions caption creation on user-specified attributes, while a "grounding unit" grounds generated captions in the input image by attending to important image regions. Users can specify caption attributes during inference, and the model uses the control unit to alter caption creation. The grounding unit checks picture regions to ground-generated captions in the input image. The authors showed that their system works on numerous image captioning evaluation metrics and that the control unit can change caption creation based on user-specified variables. The grounding unit grounding captions for inappropriate image regions were also successful.

In this paper authors, Ashish Vaswani and Noam Shazeer (Vaswani et al.) introduce the Transformer model, a neural network architecture that leverages self-attention techniques to translate sequences without recurrent neural networks (RNNs). The Transformer is more parallelizable and computationally efficient than earlier neural machine translation models, enabling training on larger datasets. The Transformer performed well on numerous language translation benchmarks, including WMT 2014 English-German and English-French. The Transformer has been used for summarization, question answering, and language translation.

In this paper author, Dzmitry Bahdanau, and Kyunghyun Cho (Bahdanau et al.) introduce neural machine translation's attention mechanism (NMT). The attention method lets the model focus on different portions of the source sentence during trans-

lation, improving quality, especially for longer sentences. The authors suggested using a soft alignment between target and source words to calculate attention weights. The model learns translation and alignment tasks end-to-end without explicit alignment information. Their attention-based model outperformed state-of-the-art NMT models on multiple benchmark datasets, including the WMT 2014 English-German and English-French translation tasks. Several NMT models now include the attention mechanism, which has been applied for different natural language processing tasks.

In this paper authors, Guillaume Lample and Ludovic Denoyer (Lample et al.) introduce a unique unsupervised machine translation framework without parallel corpus for training. Instead, they use massive monolingual corpora in both languages. A shared encoder-decoder model is trained using unsupervised and supervised objectives after learning two language models, one for the source language and one for the destination language. The encoder is trained to provide a latent representation of the source sentence that the decoder can decode into the target language. Reconstruction loss, cycle-consistency loss, and language modeling loss enhance this process. The authors demonstrated competitive performance on English-French, English-German, and English-Spanish language pairs without training data. Machine translation's scarce and expensive parallel corpora may be reduced by the proposed approach.

In this paper the author (Johnson et al.) proves that a multilingual neural machine translation (NMT) model can translate untrained language pairings. The authors trained a single NMT model using a huge multilingual corpus of parallel sentences in 103 languages and showed that it can translate well not only for language pairs seen during training but also for new language pairs. The multilingual NMT model's shared representation captures language similarities and differences. The authors also proposed a novel zero-shot translation method that employs the learned parameters of the multilingual NMT model instead of fine-tuning on new language pair data. The authors demonstrated competitive performance in English-Spanish, English-French, and English-German without extra training data. This approach allows for multilingual communication and machine translation using a single model without the requirement for massive volumes of parallel data for each language pair.

In this paper authors, Armand Joulin, and Edouard Grav (Joulin et al.) describe the creation of a massively multilingual neural machine translation (NMT) model that can translate between several languages without separate models. The scientists trained a single model on a huge parallel corpus of 25 languages and showed that it can perform well on translation tasks for all 25 languages even between pairs of languages not seen during training.The authors suggested using language codes and shared embedding layers to help the model manage so many languages. A novel training objective enables the model to share parameters across languages, improving performance for low-resource languages with less training data. The authors tested various language pairs, including low-resource languages like Swahili and Tamil, to prove their method works. Their model performed well on zero-shot translation tasks in numerous benchmarks. This approach allows machine translation systems to handle many languages without distinct models for each language pair.

Some of the most crucial aspects of information retrieval are covered in the paper by Mangala et al. (Madankar, Chandak, Chavhan, et al.), including cross-lingual information retrieval (CLIR), multilingual information retrieval (MLIR), and methodologies and techniques for machine translation. The methods include morphological analysis, bilingual dictionaries, and machine translation. In their proposal for a CLIR system for bilingual publications in Japanese and English, Atsushi et al. (Fujii, Ishikawa, et al.) emphasize the need of translating technical words. Also, one of our objectives is to combine many elements into a unified framework. Before collecting relevant documents to process a given query that uses technical terms, a system must translate it into the target language. The results demonstrate that while utilizing transliteration and compound word translation procedures improves baseline performance, the benefit is significantly larger when used in tandem.

In this paper (Sanchez-Martinez, Carrasco, et al.), the authors chose the best query phrases when seeking certain documents translation. This study investigated three alternative approaches based on statistical machine translation techniques. The strategy

that yields the best results among all those put to the test is the one in which the source text is broken down into a list of single words. In order to maximize the relationship between the likelihood that a source word will be translated into a phrase containing the query word and the frequency of the query word in the source document, queries are built from the words in the phrases that appear from the translation.

3. Materials and Methods

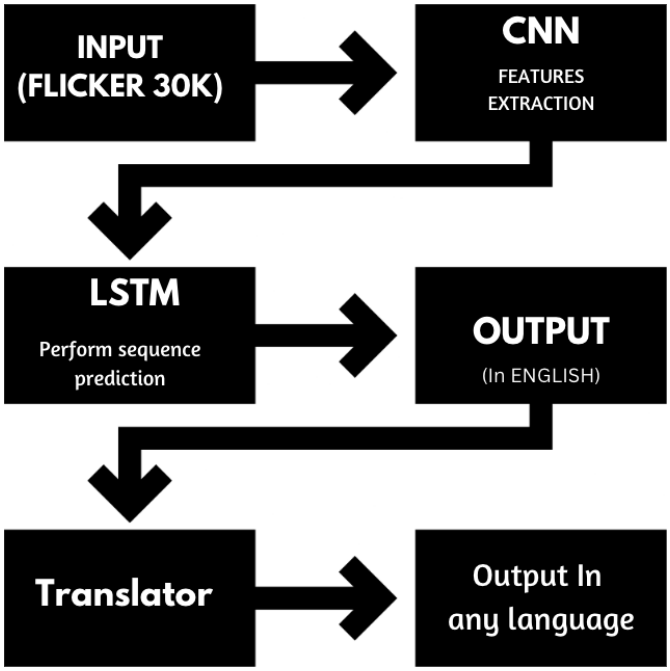

FIGURE 1. Design of the proposed Method

This research presents a model [Figure1] that retrieves information from incoming visual data and transforms it into a selected language. A conventional system requires a powerful GPU and considerable processing time to acquire and handle such voluminous unstructured data. We recommend employing Big Data ideas to solve this issue, which efficiently manage unstructured and enormous volumes of data. For this model, we will be using deep learning in the spark environment

3.1. Convolution Neural Network

A subtype of specialized neural networks called deep Convolutional neural networks processes data presented as a two-dimensional matrix. These neural networks are a subset of the CNN family. Due to the ease with which images may be represented as a two-dimensional matrix, CNN is particularly successful when working with photographs. With the help of CNN, image categorization and identification may be done very quickly.

Artificial neurons are layered on top of one another to create convolutional neural networks. Because mathematical functions produce an activation value by computing the weighted sum of many inputs, you might conceive of them as approximate models of the biological neurons they copy. Artificial neurons are sloppy copies of their biological counterparts. Each layer of a ConvNet creates a variety of activation functions after receiving a picture as input, which are subsequently passed on to the subsequent layer.

The first layer typically gets rid of basic elements like edges that run horizontally or diagonally. The next layer uses this result to identify more intricate features like corners and combinational edges. The network can recognize increasingly complex properties, such as objects, faces, and other features, as we probe deeper into it. The categorization layer generates a series of confidence scores (numbers ranging from 0 to 1) for each classification based on the activation map of the final convolution layer. These confidence scores indicate that it is more or less likely that the image belongs to a particular ”class.”

3.2. LSTM

A recurrent neural network, or RNN, is an algorithm that is particularly good at addressing sequence prediction problems. An RNN subtype is the LSTM. We can make assumptions about the words that will follow it based on the material that came before it. By getting beyond the restrictions that the RNN places on its use, it has shown to be more successful than the RNN. The LSTM algorithm has the ability to keep track of pertinent information while simultaneously eliminating irrelevant data. The LSTM model utilized in the suggested model is shown in Figure 2.

3.2.1. Working of LSTM

The memory cell and the gates—both the input gate and the ignore gate—are the two most important parts of the LSTM. The internal components of the memory cell are modified by the input gates and forget gates. As long as both segues are closed, the information in the memory cell won't change from one time step to the next gradients. This is due to the segue's function as a time-travel gate. The usage of gates allows for the maintenance of information

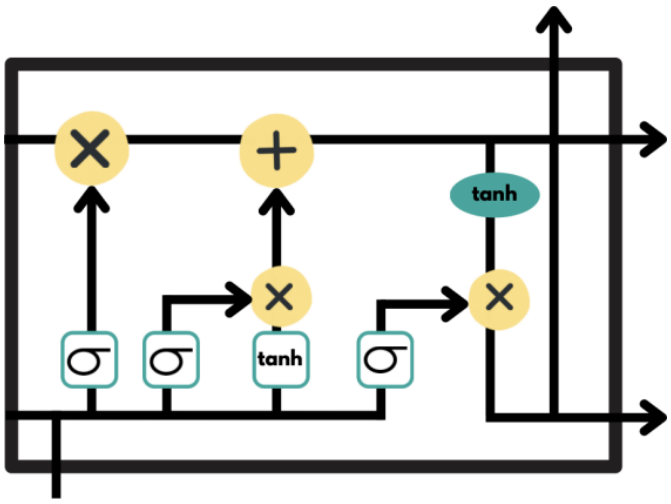

FIGURE 2. LSTM Model

across a number of time steps and the flow of group information over a number of time steps. As a result, the LSTM model is able to avoid the vanishing gradient issue that affects the majority of recurrent neural network models. As will be seen in the following, an LSTM Network is composed of four unique gates, each of which serves a specific function.

A percentage between 0 and 1 is created by adding the input and previous output at the forget gate to determine how much of the prior state must be preserved. This fraction indicates how much of the former state should be forgotten and how much must be preserved. The state that came before it is then multiplied by this output. For your information, an activation output of 1.0 signifies ”remember everything,” while a value of 0.0 means ”forget everything.” The ”forget gate” would perhaps be better referred to be the ”remember gate” if one were to think about things differently.

The input gate’s objective is to select which fresh information will be added to the state of the LSTM. The forget gate and input gate both use the same signals. The output of the input gate, which is once more a fraction between 0 and 1, is multiplied by the output of the tan h block to create the new values that must be added to the initial state. The gated vector is then added to the prior state to produce the new one.

It is often believed to be a part of the input gate, and much of the published information on LSTMs either completely ignores it or thinks that it is a part of the input gate. By having the Internal State Cell zero-mean and non-linear, the information that the Input gate will record onto it is changed. To achieve the goal of modifying the information, this is done. The learning process is sped up by doing this since zero-mean data converges more quickly. The inclusion of this gate in the construction of the LSTM unit is still advised since it is good practise, despite the fact that its actions are less significant than those of the other gates and despite the common misconception that it offers finesse, which is inaccurate.

At the output gate, the input and the previous state are controlled similarly to produce a new scaling fraction. The current state is created by combining this fraction and the output of the tanh block. The output and system status are delivered to the LSTM block..

Three inputs and two outputs, represented by the letters ht and Ct, are present in each LSTM cell. When t occurs, ht represents the status of the private information, while Ct stands for the condition of the cell or its memory. It is the input, or the point at which information is now available. Two inputs with the labels ht-1 and xt make up the first sigmoid layer. The state that was hidden in the cell before it is represented by the ht-1 input. It is also known by its moniker ”forget gate,” since its output is a selection of the amount of data from the preceding cell that should be included. This explains why it has several names. The result of it will be a value in the [0,1] range multiplied (pointwise) by the previous cell’s state.

Drawback

They became more well-known as a result of their solution to the issue of gradient vanishing. Despite this, it seems like they are unable to resolve the problem. The issue arises because data must be moved between cells in order to be analyzed. The cell is also becoming significantly more complex as new features (like the forget gate) are consistently added.

They require a large amount of time and money to train and prepare for applications in the real world. Each cell contains linear layers, which demand a high memory bandwidth, which the system is often unable to provide. As a result, in terms of hardware usage, LSTMs become relatively inefficient.

3.2.2. Dataset

The Flickr dataset was utilized to create an image caption generator. Other large datasets exist, such as Flickr 30K and the MSCOCO dataset. However, training the network with those datasets might take several weeks for computers with CPU support,

so we utilized the more compact. Flickr8k dataset. While constructing a more accurate model, using a large dataset is helpful.

3.2.3. Units

The total process can be broken down into the following primary stages:

3.2.4. Read the Caption File:

Reading the text and tokens stored in the flickr8k file, determining the overall file length, and then separating the file.

3.2.5. Data Cleaning:

Data cleaning is fixing or removing inaccurate, corrupted, improperly formatted, duplicate, or absent data from a dataset. The reliability of the findings and methods is compromised if the underlying data are incorrect, even though they appear correct on the surface.

3.2.6. Loading Data:

The procedure entails training the Pictures File, putting it through its paces, and assembling a train description dictionary containing beginning and ending sequences.

3.2.7. The pre-processing of data involving images and captions:

The procedure begins by loading the image and preprocessing, encoding, and testing the image. The captions are being loaded, the beginning and closing sequences are being appended, and the maximum length of the caption is being determined.

3.2.8. The preparation of data consists of:

Data preparation is preparing unprocessed data for processing and analysis by cleaning and making changes to it. It is a crucial step before processing and usually entails reformatting, editing, and merging various data sets to improve the data.

3.2.9. Embedding of Words:

Words into vectors are being converted here (Embedding Layer Output)

3.2.10. The architecture of the model:

Constructing an image feature extractor model and a partial caption sequence model and then combining the two networks, Figure 3 shows the model's architecture.

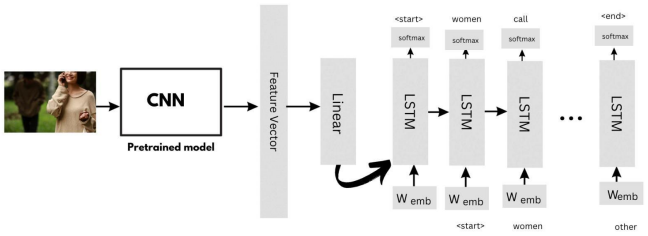

FIGURE 3. LSTM and CNN architecture for caption generator

3.2.11. Train Our Model:

The term "training model" refers to a dataset that is used to teach a machine learning algorithm. It consists of some instances of output data and the groups of input data that pertain to those instances of output data and affect the output.

3.2.12. Predictions:

When predicting the probability of a specific outcome, such as predicting a caption for a picture, an algorithm's output that has been trained on historical data and applied to new data is referred to as prediction. In this context, prediction refers to the process of predicting a caption for a photo.

3.2.13. Translating:

We are translating the English caption into other languages.

4. Results and Discussion

We can understand from the data that it is very large and will take approximately a week to get trained using our LSTM model, so to decrease its time complexity, we will use HDFS in conjunction with Map Reduce to solve this issue. That way, we may shorten the time or speed up the processing. The process of thoroughly training the network took the system one week. Thus, in order to cut the time down to hours, we will be using Hadoop.

5. Conclusion

This research presents a mechanism that retrieves visual input and converts it into a chosen language (visual to text and text to multiple languages). The proposed approach's purpose is to manage the large volume of unstructured data efficiently. To satisfy this problem, we use deep learning in the spark environment.

6. Authors' Note

The authors declare that there is no conflict of interest regarding the publication of this article. The authors confirmed that the paper was free of plagiarism

References

Anderson, Peter, et al. "Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering". (2018).

Bahdanau, et al. "Neural Machine Translation by Jointly Learning to Align and Translate". *Dzmitry Bahdanau* (2014).

Fujii, Atsushi, Tetsuya Ishikawa, et al. "Cross-language information retrieval for technical documents". (2011).

Herdade, Simao, et al. "Image captioning: Transforming objects into words". *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems*. 2019. 11135–11145.

Hu, Ronghang, et al. "Show, Control and Tell: A Framework for Generating Controllable and Grounded Captions". (2019).

Johnson, Melvin, et al. "Zero-Shot Translation with Google's Multilingual Neural Machine Translation System". (2018).

Joulin, Armand, et al. "Massively Multilingual Neural Machine Translation". (2019).

Karpathy, Andrej, Li Fei-Fei, et al. "Deep Visual-Semantic Alignments for Generating Image Descriptions ". (2015).

Lample, Guillaume, et al. "Unsupervised Machine Translation Using Monolingual Corpora Only". (2017).

Li, Ruiyu, et al. "Multimodal Attention-based Video Captioning with Temporal Segment Networks". (2019).

Lu, Jiasen, et al. "Knowing When to Look: Adaptive Attention via A Visual Sentinel for Image Captioning ". (2017).

Madankar, Mangala, M B Chandak, Nekita Chavhan, et al. "Information Retrieval System and Machine Translation: A Review". *Procedia Computer Science* 78 (2016): 845–850.

Rennie, Steven J., et al. "Self-critical Sequence Training for Image Captioning". (2017).

Sanchez-Martinez, Felipe, Rafael C Carrasco, et al. "Document translation retrieval based on statistical machine translation techniques". *Applied Artificial Intelligence* 25.5 (2011): 329–340.

Vaswani, Ashish, et al. "Attention Is All You Need". (2017).

Vinyals, Oriol, et al. "Show and tell: A neural image caption generator". *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015): 3156–3164.

Wang, Chaoyang, Ziwei Zhou, Liang Xu, et al. "An Integrative Review of Image Captioning Research". *Journal of Physics: Conference Series* 1748.4 (2021): 042060–042060.

Xu, Kelvin, et al. "Neural Image Captioning with Visual Attention". (2015).

Embargo period: The article has no embargo period.

To cite this Article: , Naman Grover, Anchita Singh , and Suganeshwari G . "**Multilingual Image caption Generator using Big data and Deep Learning** ." International Research Journal on Advanced Science Hub 05.05S May (2023): 345–352. http://dx.doi.org/10.47392/irjash.2023.S047