International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Image Captioning Using ResNet RS and Attention Mechanism

[1]Roshani Raut, [2]Sonali Patil, [3] Pradnya Borkar, [4]Prasad Zore

Submitted: 25/03/2023 **Revised**: 28/05/2023 **Accepted**: 13/06/2023

Abstract: An enormous difficulty in the field of natural language processing is the creation of automatic image captions. The majority of researchers have used convolutional neural networks as encoders and recurrent neural networks as decoders. However, a model must be able to identify the semantic relationships between the many items visible in an image in order to correctly predict image captions. Encoder and decoder states of the attention mechanism are linearly integrated to emphasize both types of data by combining visual information from the image and semantic information from the caption. When attempting to predict captions for a particular image, we made use of a convolutional neural network that had been previously trained called ResNetRS101 in conjunction with the Bahdanau attention mechanism. To evaluate the manner in which the ResNetRS101 Model and the Bahdanau mechanism of attention perform on the same dataset in comparison to the conventional CNN-LSTM technique, this is our primary objective.

Keywords**:** *CNN (Convolutional Neural Network), Deep learning, Gradient descent, LSTM (Long Short Term Memory), Attention Mechanism, NLTK (Natural Language Toolkit), ResNet, RNN (Recurrent Neural Networks)*

1. Introduction

Image captioning, also known as the automatic production of descriptions for photographs depicting real-world settings, is gaining interest in the disciplines of computer vision [1, 2], natural language processing (NLP) [3, 4], picture indexing [5, 6], and assisting individuals who are visually impaired. The automatic and computer processing of human languages is referred to as "Natural Language Processing," which is abbreviated as "NLP." It is defined as the process of software and common speech automatically exchanging text and natural language with one another [7]. It is a difficult endeavour that calls for an in-depth grasp of two different kinds of media data, specifically language and visual data [8].

Although humans find captioning images to be simple, artificial intelligence (AI) finds the task to be very difficult. Because of this, many people believe that automatically captioning photographs based on a comprehensive grasp of real-world settings is a challenging and difficult process. Thanks to AI's semantic processing of sights, people who are blind or have low vision can take advantage of technologies such as the brain-machine interface, autonomous or aided driving, and intelligent navigation. The deep learning network's architecture is crucial to the success of captioning because the NLP creates captions based on visual data received from it. Deep learning techniques offer an efficient solution for data processing in neural networks, despite their extremely complex designs.

Deep learning has been applied in many industries, which involves healthcare, agriculture, and others, because of its broad variety of applications. Deep learning has been applied in healthcare to predict child disorders [9], [10], classify human diseases [11], [12], and solve classification difficulties in agriculture [13]–[15] and other image processing issues [16], [17]. Captioning uses deep learning networks to extract visual data from images. After that, they are given to NLP for caption creation. Several studies have used Long-Short Term Memory (LSTM) and Convolutional Neural Networks (CNN) to caption images, including those in [18]–[20]. The vast majority of researchers have relied on the CNN-LSTM framework when it comes to the captioning of images. The CNN performs the function of an image encoder, evaluating visual areas and programming them as area-specific characteristics, while the LSTM serves as the decoder and makes an effort to comprehend all of the words that are generated by the CNN. There are two methods that can be utilised in order to extract the areas from an image [8]: As opposed to splitting an image based on the model's architecture, an adaptive method makes use of a bounding box to capture the areas of a picture at the object level. This is done in place of dividing an image. Even though the production of non-visual words such as "the" and "were" and "itself" requires very little visual information, these neural models of attention are nevertheless obliged to incorporate image qualities, which leads to the generation being misled.

It is important to note that the method by which a model makes use of an image's features—both visual and non-

[1,2] *Department Of Information Technology Pimpri Chinchwad College Of Engineering*
Pune, India
[3]*Department of Computer Science and Engineering, Symbiosis Institute of Technology, Nagpur, Symbiosis International (Deemed University), MS ,India*
[4]*Department Of Computer Engineering, Pimpri Chinchwad College Of Engineering*
Pune, India

visual objects—is of utmost importance. The content of a picture should be focused while coming up with words for visual things, whereas the hints should be focused when coming up with captions for non-visual words. In order to avoid being tricked, the attention mechanism places a greater emphasis on the vector of non-visual characteristics than it does on the visual characteristic vector. As a result, attention models are employed to resolve confounding issues and produce appropriate captions for non-visual things shown in a picture. According to the published research, credible semantic descriptions of objects in all visual regions have attracted a significant amount of interest from researchers, particularly in the field of attention mechanisms. For example, some works, such as [21], blend the use of visual words with aspects drawn from earlier terminology that did not involve the use of visual words. The attention model ought to be taught non-visual cues so that it can contribute to the development of non-visual words and assist ease the problem of misleading information [8].

2. Motivation

Automatic image captioning, particularly has been rendered possible by machine learning as well as deep learning models, is one of the most helpful visual resources. It can be found on many websites. Applications include remote sensing, mapping, social networking sites like Facebook, automatic inference from photos, and many others.

Building dynamic webpages, understanding medical images, and translating natural language to images are just a few examples. In addition, because to the widespread application of picture captioning, a variety of objects can be manually identified in photographs taken from a great distance by satellites as well as by medical professionals. Finding objects in photos is exceedingly difficult for experts and nearly impossible. It is imperative to make as much use as possible of machine learning and deep learning techniques in order to speed up the process of image interpretation. An efficient system for image captioning should be able automatically differentiate between both visual and non-visual things in a picture and label it in a way that is appropriate for each type of object.

3. Related Work

The creation of several models has been sparked by the increased interest of academics in picture captioning [22]–[24]. CNN has been adopted as an encoder and RNN has been adopted as a decoder in a large number of state-of-the-art techniques [25, 26] for evaluating images and producing captions as a result of the considerable breakthroughs made in Deep Neural Networks (DNNs). These advancements were made possible as a result of the significant progress made in DNNs. CNN is frequently utilised in the role of an encoder because of its ability to derive high-level contextual data from images. The retrieved features are then decoded using an RNN decoder to create captions. Since CNN is as well-liked as an approach to the processing of images, CNN-RNN, which is an amalgamation of CNN and RNN, has become quite popular in the field of computer vision. The vast majority of research organisations make use of pre-existing architectures such as ResNet-50 or DenseNet-121, but only a very small number of companies construct their very own networks from the bottom up. Since they are typically trained on huge, readily available datasets like ImageNet, these pre-trained neural networks can recognise visual elements. Pre-trained models can be used to new models to increase their precision while accelerating training. These models don't need to start from scratch because they are already familiar with the fundamental elements of an image.

When creating captions, a precise and dynamic decoding of both visual and non-visual elements is necessary. Anderson et al. proposed a ground-breaking method known as "bottom-up and top-down" in [22]. The crucial regions of an image that the bottom-up module determined are important are represented by a convolutional feature vector. Two LSTM networks make up the top-down module; the first serves as a top-down visual attention model and the second as a language model. Another paper [23] proposed a hybrid model that blends graph convolutional networks and LSTM in an effort to improve image representation by incorporating semantic and spatial relationships. The authors created two spatial and semantic graphs for this. Chen et al. created scene graphs using extensive captioning and structured language descriptions to get around the challenge of difficult image retrieval [24]. They introduced a scene graph matching technique and employed a customised CBIR dataset for in-depth research. For image retrieval, a rigorous caption reasoning method with two steps—creating dense captions and building and analysing scene graphs—was applied. The model occasionally returns meaningless data because the decoder has problems recognising non-visual words like "was," "here," "put," etc. In order to solve this issue, the attention mechanism is crucial. The attention mechanism has advanced significantly in many sequence learning problems. It generates an attention result or weighted average vector by computing candidate vector scores, normalising those scores into weights using the Softmax function, and then applying those weights to the candidates. Numerous more attentional mechanisms, including multi-level attention [25], multi-head and self-attention, as well as spatial and channel-specific attention, have also been theorised. The unique technique known as "Attention on Attention" [26], created by Huang et al., is utilised by both the encoder and the decoder. It increases the encoder's ability to more accurately simulate relationships between different objects and aids the decoder in avoiding wasting time on unneeded outputs. Yan et al. presented a hierarchical attention mechanism-based framework in [27] using a policy gradient approach and generative adversarial network (GAN). Khan et al.'s innovative method [28] collects information from an image

to create a feature vector using a pre-trained CNN model. They also contained Gated Recurrent Units (GRU) for picture decoding. The authors coupled the Bahdanau attention mechanism and GRU to enable the learning process to focus on particular areas of a picture. However, these studies have not addressed the concept of transfer learning.

An attention function converts a query and a set of key-value pairs into an output. The compatibility between the query and relevant keys determines how much weight is assigned to each result in the output, which is calculated by computing a weighted sum of the values. Three different types of attentional methods can be distinguished from the decoder's output:

1) Attribute-based visual representation: A confidence vector is a matrix that not only includes objects and attributes, but also things, interactions, relations, and other elements. This matrix is used to describe the qualities of an image, and it is customary for this matrix to take the form of a vector. A text-based semantics attention model was presented by the researchers that worked on this project [22]. The encoder may learn which areas of the image the model should concentrate on using this attention model, which is based on the previously created text.

2) Without semantic labelling, a grid-based visual feature representation: It involves employing a CNN to extract features, which are then provided to the attention model. The attention model is concerned with particular spatial regions that need attention. A 1414 VGG-based end-to-end spatial attention model using both hard and soft methods was proposed by K. Xu et al. [23].

3) Object-based visual representation: Bidirectional RNN (BRNN) and Region-CNN (RCNN) were used in A. Karpathy et al.'s model [24] to connect picture areas with text segments and infer latent alignments. The sentences served as inadequate labels in this method. In order to generate visual descriptions, an end-to-end RNN model was employed.

4. Proposed Methodology

We use CNN, which provides the most accurate results possible for image processing, to enhance the estimation of captions. Figure 1 illustrates the suggested approach to the problem. Flattening an image and extracting its features are two of CNN's primary uses. It takes extensive knowledge of both visual and non-visual objects to accurately describe a picture. The many visual and non-visual objects interact with one another in different parts of an image in different ways. They only have these linkages under the specific terms that have been formed. It is designed specifically to create both visual and non-visual fine-grained semantic linkages. As can be seen in Figure 1, the feature vector is transformed into weights for each portion of the image to be analysed.

Within a CNN, the filter kernel of each convolutional layer performs the role of a semantic detector. The semantic detector is responsible for calculating the weighted sum of all of the features of both visually and non-visually perceivable objects. A semantic relationship between distinct items is provided by semantic knowledge representation. The attention mechanism receives its input from the semantic information. As a result, it aids in improving the accuracy of the attention mechanism's performance.

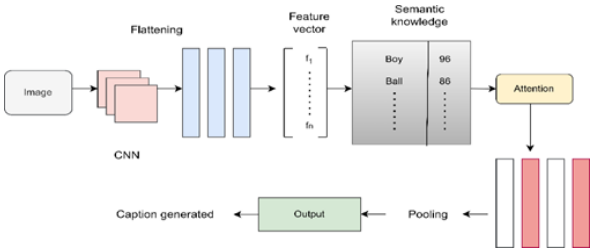

Fig 1. Proposed Framework

The attention process is then employed to emphasise the key aspects as higher-level semantic information is learned. Following feature learning, captions are generated using pooling. In order to extract information from images while accounting for their complex characteristics, convolutional neural networks (CNNs) use a convolutional layer. To cut down on computing expenses, we selected pre-trained CNN models that had already been trained on huge image datasets. Transfer learning appears to be extremely beneficial for fast updating models. Once we have trained the models to achieve very high accuracy, updating them with new data is straightforward. Specifically, we used the pre-trained ResNetRS101 model and referenced our research.

The ResNetRS101 model can identify key elements in an image and produce captions for them. The primary objective of this study is to determine whether a deep layered design is necessary to maintain optimal performance and how the model's architecture influences caption generation performance. Using a pre-trained deep neural network with an attention mechanism, we aim to identify the non-visual elements (such "here" and "with") that significantly contribute to the semantic meaning of image captions. Even in the instance of picture captioning, there is no need for retraining because pre-trained deep neural networks have already been trained on a huge number of photos to extract high-level properties. Pre-trained models are therefore expected to improve computing efficiency and reduce caption production costs. The attention technique is also applied to captioning photographs, when non-visual words are purposefully given greater attention than visual ones.

A. Deep Feature Extractor

Classical techniques, such as the histogram of oriented gradients (HOG) and the scale-invariant feature transform

(SIFT), have proven to be effective for performing a variety of computer vision tasks, such as object detection, classification, segmentation, and picture retrieval. These tasks have been successfully completed in the past. The past ten years have seen a large increase in the quantity of datasets that can be accessed, this has served as the contributing factor behind the production of learning-based descriptors such as AlexNet, ResNet, and GoogleNet [26–34]. This increase in the quantity of datasets that can be accessed has been seen as a motivating force driving the creation of learning-based descriptors. Deep learning models require a lot of computational power in addition to a large amount of training data. Machine learning feature extractors, on the other hand, only require specialised knowledge. The underlying principle of features representation-learning based on vast volumes of processed picture data has not altered, despite the fact that many different architectures for deep learning have been devised. Computer models are now able to learn interpretations of data at various degrees of abstraction with the assistance of numerous layers of neural network processing. Each layer abstracts the representation from the previous layer in order to eventually understand distinguishing characteristics. However, it has a high GPU, time, and processor computing cost. Although this makes it feasible to extract high-level characteristics from a large amount of data to characterize the original image. Transfer learning using pre-trained models is therefore considered to be a workable solution to these issues.

Deep learning has transformed not only computer vision but also related disciplines like natural language processing and image analysis in medicine. In this area, a number of pre-trained models have been introduced, taking their cue from convolutional neural networks' (CNNs') success with image captioning. However, the architectures of a number of pre-trained models based on deep learning and the Bahdanau mechanism for attention have not been compared to one another as of yet. For the purpose of this investigation, we make use of a model that has already been trained and is known as ResNetRS101 to generate captions for photos. The methodology for the captioning process as well as the fundamental structure of the pre-trained models are both detailed in this section.

B. ResnetRS Architecture

The term "residual networks," abbreviated as "Resnet," refers to a family of distinct deep neural networks that share the same topologies but employ varying levels of complexity. In order to forestall the gradual deterioration of deep neural networks over the course of their lifetimes, Resnet was designed with the presence of a structure known as a residual learning unit. When paired with a shortcut link, this unit's feedforward network architecture permits the development of fresh outputs as well as the insertion of entirely novel inputs to the network. This is made possible by the feedforward network structure. The most important advantage of including this component in the model is that it improves the reliability of categorization without introducing any new layers of complexity [23].

Following ResNet's enormous success, it was intensively explored, and a few more ResNet variations were developed. Let's examine a few of these ResNet variations. A new family of ResNet architectures called ResNet-RS is developed using enhanced scaling and training techniques. ResNet-RS is 1.7x–2.7x quicker on TPUs than EfficientNets while maintaining comparable accuracy on ImageNet. ResNet-RS is 4.7 times faster than EfficientNet NoisyStudent and achieves 86.2% top-1 ImageNet accuracy in a large-scale semi-supervised learning setting [24]. The training methods extend to video classification on Kinetics-400 and enhance transfer performance on a variety of downstream tasks, rivalling cutting-edge self-supervised algorithms.

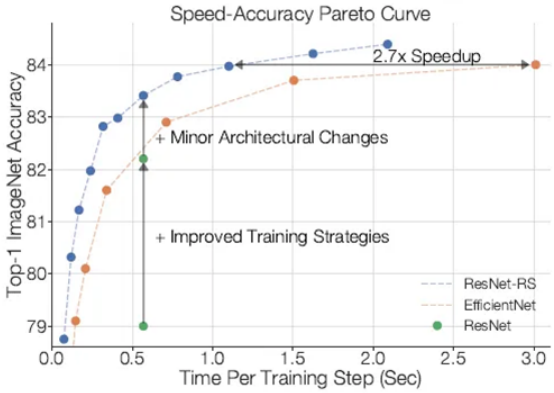

Fig 2. ResNet RS performance

C. Transfer Learning

Deep learning stands out due to the need for a large amount of data to function properly. The training procedure may result in under-fitting when the amount of data is insufficient. Researchers have created transfer learning methods to build deep learning networks with little training data in order to overcome this problem. The promise for accurate and effective picture classification is provided by transfer learning [30]. In this method, relevant features are found by pre-training a model on a dataset. The built-in model's pre-trained weights from ImageNet are initialised during implementation, ensuring that the previously discovered characteristics are taken into account and ultimately producing better results.

D. Attention Mechanism

Any source sentence, regardless of length, is encoded into a fixed-length vector in conventional machine translation encoder-decoder systems, which are then used by the decoder to produce a translation. However, when it comes to translating sentences, this method frequently produced subpar results. The difficulty for the decoder was to fit all of the incoming data into a single, fixed-size vector.

Unfortunately, dealing with excessively long words or circumstances where the traits differed across multiple photos, like image captioning, caused problems. In these circumstances, it is necessary to think about whether a single vector can sufficiently capture all the crucial data related to the image. Additionally, in some cases, non-visual elements should be given precedence over visual elements.

The objective is to highlight the crucial words inside the sentences rather than concentrating on the entire vector. The performance bottleneck present in traditional encoder-decoder systems was addressed by the Bahdanau focus [35], resulting in a significant improvement over the traditional approach. This attention mechanism, also known as additive attention, computes the encoder and decoder states linearly. The fundamental tenet of the Bahdanau attention mechanism is the weighted prioritisation of specific input vectors within a sequence. These attention weights, which describe how much "attention" should be given to each input word at each decoding level, are supplied to the decoder.

The method developed by Bahdanau involves combining the forward and backward hidden states derived from the bi-directional encoder with the undetectable states derived from the target that came before it in the case of a non-stacking bidirectional decoder. In contrast to earlier models, which lacked attention and used just the most current encoder hidden state, the context vector is constructed by including both the encoder and the decoder's hidden states. Earlier models relied solely on the most recent encoder hidden state. A feed-forward neural network is used to calculate an alignment score, which is then used by the attention mechanism to direct both the input and the output sequences to concentrate on the most significant part of the image. The model generates captions for photos by forecasting context vectors linked to the source position and previously produced target words. Figure 3 and the following description provide a detailed illustration of the Bahdanau attention model's design:

- S_{t-1} is the secret decoder state that existed at time step t-1.
- Each decoder step produces a distinct context vector c_t at time step t in order to produce a target word y_t.
- An annotation h_i that concentrates on the i-th word out of all the words and captures the important information about that word.
- Each annotation's weight value at the current time step t is equal to $t,_i$.
- The attention score et,i. produced by the given model a (.) demonstrates how well S_{t1} and h_i match.

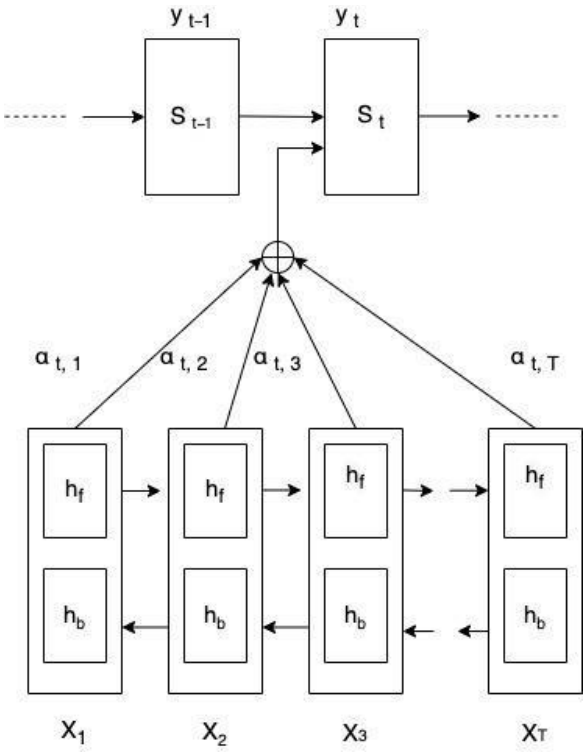

Fig 3. Attention Mechanism

The Bahdanau architecture is represented in Figure 3 by a bidirectional recurrent neural network (BI-RNN), which combines an attention mechanism and functions as both an encoder and a decoder.

5. Experimental Setup

In the part that came before this one, we discussed the several approaches for extracting pre-trained deep learning features. This section provides an explanation of both the information being collected and the evaluation of the experiment.

E. Dataset

The most used dataset for picture captioning is Flickr8k. There are 8092 total photos in it, of which 6000 are utilised to train the model and 2000 to verify its effectiveness. There are five captions that are relevant to each image. The various captions are used to both train and test the model. In this effort, the data was filtered by removing superfluous words from an image's caption. The most and least commonly used words were initially calculated. Following that, we eliminated superfluous words as depicted in Figures 4 and 5.

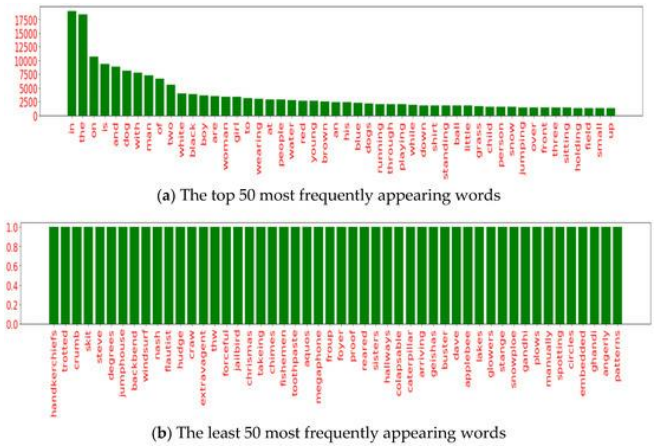

Fig 4. Top 50 words in the dataset by frequency of use.

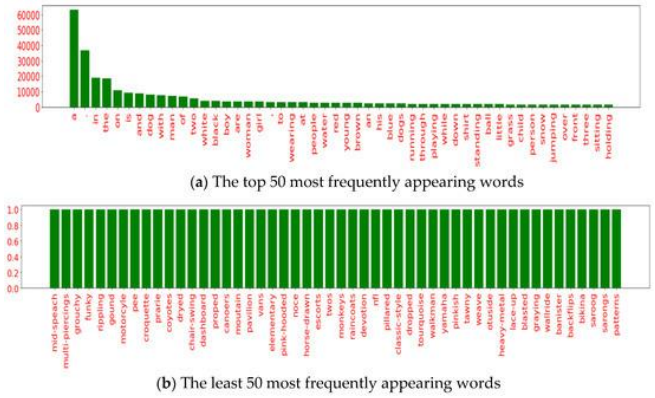

Fig 5. After captions have been cleaned, the top 50 most and least used words in the dataset.

F. Performance Metrics:

We employed a variety of criteria, which are explained below, to evaluate the effectiveness of model prediction:

Bilingual evaluation understudy or BLEU: This well-known machine translation statistic gauges the degree to which one sentence resembles other sentences in a collection. Papineni et al. made the suggestion in [25]. It is possible to return a value, with a higher value indicating a greater degree of resemblance. By counting the n-grams that are present in the reference sentence, this methodology arrives at an estimate of the total number of n-grams that are found in each sentence. A token is referred to as a unigram, which can also be abbreviated as a one-gram. On the other hand, a bi-gram is an abbreviation for a pair of words.

G. Experiment Settings

The ideal collection of parameters for each model was our goal. In order to do that, we split the data into two sets: a training set of 6000 data points, a testing set of 1000, and a validation set of 1000. The frequency of the most and least utilised words in our sample was initially measured. All of the captions were given and labelled at the pre-processing stage. The goal is for the models to be able to distinguish between the beginning and finish of captions. Utilising VGG16, a pre-trained model developed using the ImageNet dataset, is the next stage. It has fully connected layers and convolutional layers. We did tokenization and built language to produce image captions. Vector notations are created for each word in the caption.

Encoder and decoder: To create captions, we use RNN as the decoder and ResNetRS101 as the encoder. Although CNN's final layer, known as Softmax, is employed for classification, we deleted it for this study in order to feed the decoder's features. The batch size, the number of units, the learning rate, and the embedding dimension are all set to 64 for ResNetRS101 implementation. It employs Adam optimizer. The error is computed using the sparse cross-entropy. The dropout was set to 0.5. ResNetRS101 has over 60 million trainable parameters in total.

II. RESULTS

The following are the results and the BLEU Score produced for sample images:

Fig 6. Sample Image

Fig 7. Sample Image

Fig 8. Sample Image

6. Conclusion

The goal of this effort was to create a model for automatically creating image captions using different CNN. The use of transfer learning techniques is the overall theme of this work. As a result, in this research, we examined ResNetRS101's performance as a pre-trained deep learning model. Additionally, we added the Bahdanau attention mechanism to the two pre-trained deep learning CNNs. We demonstrate the critical significance that semantic knowledge plays in producing accurate captions using the Bahdanau attention mechanism. The effectiveness of attention-based learning in comparison to conventional image captioning techniques was tested in experiments. Our research has produced promising findings. Due to Flickr8K's limited dataset, deeper models perform better on the retrieval task than shallow ones. This research can also be strengthened by using multiple attention strategies with various pre-trained CNN architectures. In order to test the local and limited attention mechanism so that the model is able to create captions for many various sorts of inputs, such as images, videos, and audios, etc., additional algorithmic techniques for deep learning, such as Transformers, can be introduced into the model. This allows for the evaluation of these mechanisms so that the model can generate captions for as many different types of inputs as possible. These approaches allow the model to provide captions for a wide variety of inputs.

References

[1] P. Wang *et al.*, "OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework," Feb. 2022, Accessed: May 23, 2023. [Online]. Available: https://arxiv.org/abs/2202.03052v2

[2] T. Y. Hsu, C. L. Giles, and T. H. Huang, "SciCap: Generating Captions for Scientific Figures," *Findings of the Association for Computational Linguistics, Findings of ACL: EMNLP 2021*, pp. 3258–3264, 2021, doi: 10.18653/V1/2021.FINDINGS-EMNLP.277.

[3] M. Z. Hossain, F. Sohel, M. F. Shiratuddin, H. Laga, and M. Bennamoun, "Text to Image Synthesis for Improved Image Captioning," *IEEE Access*, vol. 9, pp. 64918–64928, 2021, doi: 10.1109/ACCESS.2021.3075579.

[4] S. Sehgal, J. Sharma, and N. Chaudhary, "Generating Image Captions based on Deep Learning and Natural language Processing," *ICRITO 2020 - IEEE 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)*, pp. 165–169, Jun. 2020, doi: 10.1109/ICRITO48877.2020.9197977.

[5] H. Jain, J. Zepeda, P. Perez, and R. Gribonval, "Learning a Complete Image Indexing Pipeline," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 4933–4941, Dec. 2018, doi: 10.1109/CVPR.2018.00518.

[6] S. Pang, M. A. Orgun, and Z. Yu, "A novel biomedical image indexing and retrieval system via deep preference learning," *Comput Methods Programs Biomed*, vol. 158, pp. 53–69, May 2018, doi: 10.1016/J.CMPB.2018.02.003.

[7] B. Makav and V. Kilic, "A New Image Captioning Approach for Visually Impaired People," *ELECO 2019 - 11th International Conference on Electrical and Electronics Engineering*, pp. 945–949, Nov. 2019, doi: 10.23919/ELECO47770.2019.8990630.

[8] Z. Zhang, Q. Wu, Y. Wang, and F. Chen, "High-Quality Image Captioning with Fine-Grained and Semantic-Guided Visual Attention," *IEEE Trans Multimedia*, vol. 21, no. 7, pp. 1681–1693, Jul. 2019, doi: 10.1109/TMM.2018.2888822.

[9] S. Alam, P. Raja, and Y. Gulzar, "Investigation of Machine Learning Methods for Early Prediction of Neurodevelopmental Disorders in Children," *Wirel Commun Mob Comput*, vol. 2022, 2022, doi: 10.1155/2022/5766386.

[10] F. Sahlan, F. Hamidi, M. Z. Misrat, M. H. Adli, S. Wani, and Y. Gulzar, "Prediction of Mental Health Among University Students," *International Journal on Perceptive and Cognitive Computing*, vol. 7, no. 1, pp. 85–91, Jul. 2021, Accessed: May 23, 2023. [Online]. Available: https://journals.iium.edu.my/kict/index.php/IJPCC/article/view/225

[11] [11] S. A. Khan, Y. Gulzar, S. Turaev, and Y. S. Peng, "A Modified HSIFT Descriptor for Medical Image Classification of Anatomy Objects," *Symmetry 2021, Vol. 13, Page 1987*, vol. 13, no. 11, p. 1987, Oct. 2021, doi: 10.3390/SYM13111987.

[12] Y. Gulzar and S. A. Khan, "Skin Lesion Segmentation Based on Vision Transformers and Convolutional Neural Networks—A Comparative Study," *Applied Sciences 2022, Vol. 12, Page 5990*, vol. 12, no. 12, p. 5990, Jun. 2022, doi: 10.3390/APP12125990.

[13] K. Albarrak, Y. Gulzar, Y. Hamid, A. Mehmood, and A. B. Soomro, "A Deep Learning-Based Model for Date Fruit Classification," *Sustainability 2022, Vol. 14, Page 6339*, vol. 14, no. 10, p. 6339, May 2022, doi: 10.3390/SU14106339.

[14] Y. Gulzar, Y. Hamid, A. B. Soomro, A. A. Alwan, and L. Journaux, "A Convolution Neural Network-Based Seed Classification System," *Symmetry 2020, Vol. 12, Page 2018*, vol. 12, no. 12, p. 2018, Dec. 2020, doi: 10.3390/SYM12122018.

[15] Y. Hamid, S. Wani, A. B. Soomro, A. A. Alwan, and Y. Gulzar, "Smart Seed Classification System based on MobileNetV2 Architecture," *Proceedings of 2022 2nd International Conference on Computing and*

Information Technology, ICCIT 2022, pp. 217–222, 2022, doi: 10.1109/ICCIT52419.2022.9711662.

[16] Y. Hamid, S. Elyassami, Y. Gulzar, V. R. Balasaraswathi, T. Habuza, and S. Wani, "An improvised CNN model for fake image detection," *International Journal of Information Technology (Singapore)*, vol. 15, no. 1, pp. 5–15, Jan. 2023, doi: 10.1007/S41870-022-01130-5.

[17] Gummadi, A. ., & Rao, K. R. . (2023). EECLA: A Novel Clustering Model for Improvement of Localization and Energy Efficient Routing Protocols in Vehicle Tracking Using Wireless Sensor Networks. International Journal on Recent and Innovation Trends in Computing and Communication, 11(2s), 188–197. https://doi.org/10.17762/ijritcc.v11i2s.6044

[18] M. Faris *et al.*, "A Real Time Deep Learning Based Driver Monitoring System," *International Journal on Perceptive and Cognitive Computing*, vol. 7, no. 1, pp. 79–84, Jul. 2021, Accessed: May 31, 2023. [Online]. Available: https://journals.iium.edu.my/kict/index.php/IJPCC/article/view/224

[19] H. Sharma and A. S. Jalal, "Incorporating external knowledge for image captioning using CNN and LSTM," *Modern Physics Letters B*, vol. 34, no. 28, Oct. 2020, doi: 10.1142/S0217984920503157.

[20] C. Wang, H. Yang, C. Bartz, and C. Meinel, "Image captioning with deep bidirectional LSTMs," *MM 2016 - Proceedings of the 2016 ACM Multimedia Conference*, pp. 988–997, Oct. 2016, doi: 10.1145/2964284.2964299.

[21] J. Aneja, A. Deshpande, and A. G. Schwing, "Convolutional Image Captioning," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 5561–5570, Dec. 2018, doi: 10.1109/CVPR.2018.00583.

[22] X. Yang, H. Zhang, and J. Cai, "Learning to collocate neural modules for image captioning," *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2019-October, pp. 4249–4259, Oct. 2019, doi: 10.1109/ICCV.2019.00435.

[23] L. Zhou, C. Xu, P. Koch, and J. J. Corso, "Watch what you just said: Image captioning with text-conditional attention," *Thematic Workshops 2017 - Proceedings of the Thematic Workshops of ACM Multimedia 2017, co-located with MM 2017*, pp. 305–313, Oct. 2017, doi: 10.1145/3126686.3126717.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-December, pp. 770–778, Dec. 2016, doi: 10.1109/CVPR.2016.90.

[25] I. Bello *et al.*, "Revisiting ResNets: Improved Training and Scaling Strategies," *Adv Neural Inf Process Syst*, vol. 27, pp. 22614–22627, Mar. 2021, Accessed: May 31, 2023. [Online]. Available: https://arxiv.org/abs/2103.07579v1

[26] Ms. Nora Zilam Runera. (2014). Performance Analysis On Knowledge Management System on Project Management. International Journal of New Practices in Management and Engineering, 3(02), 08 - 13. Retrieved from http://ijnpme.org/index.php/IJNPME/article/view/28

[27] H. Maru, T. S. S. Chandana, and D. Naik, "Comparison of Image Encoder Architectures for Image Captioning," *Proceedings - 5th International Conference on Computing Methodologies and Communication, ICCMC 2021*, pp. 740–744, Apr. 2021, doi: 10.1109/ICCMC51019.2021.9418234.