IMAGE CAPTIONING USING DEEP LEANING

Piyush Singh
AIT-CSE in AIML
Chandigarh University
Vadodara,India
0piyush2003singh0@gmail.com

Bhudil Mallick
AIT-CSE in AIML
Chandigarh University
Vadodara,India line 4: City, Country
bhudil.mallick@gmail.com

Manan Khanna
AIT-CSE in AIML
Chandigarh University
Vadodara,India
manankhanna13@gmail.com

Nikhil Bansal
AIT-CSE in AIML
Chandigarh University
Vadodara,India
bansal562003@gmail.com

Ishant Kamboj
AIT-CSE in AIML
Chandigarh University
Vadodara,India
Kambojishant123@gmail.com

Abstract: Creating a description for an image is commonly achieved through the usage of image captioning. This strategy entails determining the key elements, their characteristics, and the connections between them in a picture. Image captioning has grown in importance recently and can be used for many different purposes. In order to fulfill this need, we present a deep learning model that creates English phrases that explain an image's content by combining computer vision and machine translation.

Keywords - Image, Caption, CNN, RNN, LSTM, Neural Networks

I. INTRODUCTION

Captions for photos, often called image captions, are used to convey the visual information of a picture in the language of the area. The creation of captions for photographs is a difficult task since it calls for the capacity to precisely identify the elements in the picture and their relationships, as well as to convey this information in plain words. Since recent advances in picture classification, object identification, and language modelling, the field of image captioning research has improved tremendously. These captions can help people who are visually impaired understand material found on the internet, which can be quite helpful for them. However, creating detailed captions in well-structured English is a difficult task that calls for language processing, object recognition, and image content recognition.

II. EASE OF USE

People with visual impairments can interact with and access internet content with ease because to user-friendly interfaces and efficient procedures. For users of all backgrounds and skill levels, the smooth creation of comprehensive captions in well-organized English enhances the surfing experience while also making comprehension easier.

For this activity, semantic information must be provided using a natural language model, such as English. [1]With a noteworthy outcome, the suggested Siamese Difference Captioning Model (SDCM) achieves competitive performance on the SpotThe-Diff baseline dataset, yielding understandable, meaningful, and concise textual interpretation.[2] A brand-new dual-attention picture caption generation approach has been put forth to take advantage of both textual and visual attention, where textual attention strengthens the information's integrity and visual attention improves comprehension of image features.[3] A brand-new

picture captioning model built on the foundation of GANs is introduced. It trains the model without using intermediary algorithms like policy gradient [4] They primarily concentrate on neural network-based techniques, which produce cutting-edge outcomes. due to the fact that neural network-based approaches employ various frameworks. They separated them into smaller groups and talked about each group separately.[5]They presented a brand-new image captioning model dubbed the domain-specific image captioning generator, which uses visual and semantic attention to generate a caption for a given image and creates a domain-specific caption with semantic ontology by substituting domain-specific words for the specific words in the general caption.[6] They introduced Inception, an innovative dual generator generative adversarial network designed to improve a generation-retrieval ensemble model. This allows for the mutual improvement of generation-based and retrieval-based picture captioning techniques. Layered networks are used in deep learning approaches to mimic the structure of the human brain and extract salient characteristics from an image. [1] Olowakandi, A., Baagyere, E. Y., Komeda, B., Alabdulkreem, E., & Qin, Z. (2019).Captionnet: An automatic, end-to-end, attention-based Siamese difference captioning model. IEEE Access 7, 106773–106783.CaptionNet: Autonomous Complete SDCM Paying Close Attention 2019; A. Oluwasanmi et al. The suggested multi-modal end-to-end encoder-decoder architecture makes use of a deep neural network to produce a natural language characterisation for contrast inside the image pairs. In order to generate a wide range of coherent language model opportunities, their suggested supervised model employs numerous deep learning algorithms to evaluate the viability of photography, alignment, and computer-assisted variance between the two image elements. Model tests are conducted using a standard spot-the-difference baseline dataset consisting of pairs of similar images and descriptions. variables, but not Greek symbols By using the picture labels produced by a Fully Convolutional Network (FCN) to create image captions, their suggested approach leverages visual attention to enhance comprehension of the image. Additionally, the approach makes use of textual attention to improve the information's integrity. Ultimately, the creation of labels, linked to the textual attention mechanism, and the creation of image captions have been combined to provide a trainable framework that can be used from start to finish. The effectiveness and feasibility of their suggested approach are demonstrated by the experimental findings obtained from the AIC-ICC image Chinese caption benchmark dataset.[3] Seydi, V., Madadi, Y., and Dehaqi, A. M. (2021). Adversarial Network for Image Captioning. SN Computer Science, 2(3), 1–14. The best practices for each category were outlined, and the advantages and disadvantages of each kind of labour were discussed. the early work on picture captioning, which focuses mostly on retrieval and template-based methods.

Next, methods based on neural networks received much of the interest since they produced cutting-edge outcomes. Since neural network-based approaches employ a variety of frameworks, we further separated them into subcategories and addressed each one separately. Subsequently, cutting-edge techniques are contrasted using reference datasets. Provided a last talk about potential routes for future automatic image captioning research.[5]Han, S. H., and Choi, H. J. (February 2020). Semantically-ontological image caption generator that is domain-specific. (pp. 526–530) in the 2020 IEEE International Conference on Big Data and Smart Computing (BigComp). presented a novel EnsCaption model that uses a novel dual generator generative adversarial network to improve a collection of retrieval-based and generation-based picture captioning techniques. EnsCaption is made up of three main components: a caption generation model that creates customized captions for the query image, a caption re-ranking model that selects the best-matching caption from a pool of candidate captions that includes both generated and pre-retrieved captions, and a discriminator that comprehends the differences between the ground-truth captions and the generated/retrieved captions on multiple levels. The discriminator, which was trained using the multi-level ranking, was trained to assign low-ranking scores to the generated and retrieved candidate captions with high-ranking scores. Meanwhile, the caption re-ranking and generation models improved synthetic and retrieved candidate captions during the adversarial training process.

III . METHOD

For image caption generation, LRCN maximizes the probability of the description giving the image: $\theta * = \arg\max \theta$ X (I,S) log p(S|I; θ) (2) where θ are the parameters of the model, I is an image, and S is a sample sentence. Let the length of the sentence be N, the method applies the chain rule to model the joint probability over S0, · · · , SN : log p(S|I) = X N t=0 log p(St |I, S0, · · · , St−1) (3) where the θ is dropped for convenience, St is the word at step t. The model has two parts. The first part is a CNN which maps the image to a fixed-length visual feature. The visual feature is embedded to as the input v to the RNN. v = Wv(CNN(I)) (4) where Wv is the visual feature embedding. The visual feature is fixed for each step of the RNN. In the RNN, each word is represented a one-hot vector St of dimension equal to the size of the dictionary. S0 and SN are for special start and stop words. The word embedding parameter is Ws: xt = WtSt , t ∈ {0 · · · N − 1} (5)

In this way, the image and words are mapped to the same space. After the internal processing of the RNN, the features v, xt and internal hidden parameter ht are decoded into a probability to predict the word at current time: pt+1 = LSTM(v, xt , ht), t ∈ {0 · · · N − 1} (6) Because a sentence with higher probability does not necessary mean this sentence is more accurate than other candidate sentences, post-processing method such as Beam Search is used to generate more sentences and pick top-K sentences. An RGB image is mapped to a visual feature vector in this research using a convolutional neural network (CNN). The convolution, pooling, and fully-connected layers are the three most often utilized layers in a CNN. Rectified Linear Units (ReLU) f(x) = max(0, x) is another example of as the active function that is non-linear. Compared to the conventional methods, f(x) = tanh(x) or f(x) = (1 + e −x) −1, the ReLU is faster. The purpose of the dropout layer is to stop overfitting. Every hidden neuron's output is set to zero by the dropout with a probability of 0.5. The neurons that have "dropped out" are not involved in backpropagation or the forward pass.

IV. SYSTEM ARCHITECTURE

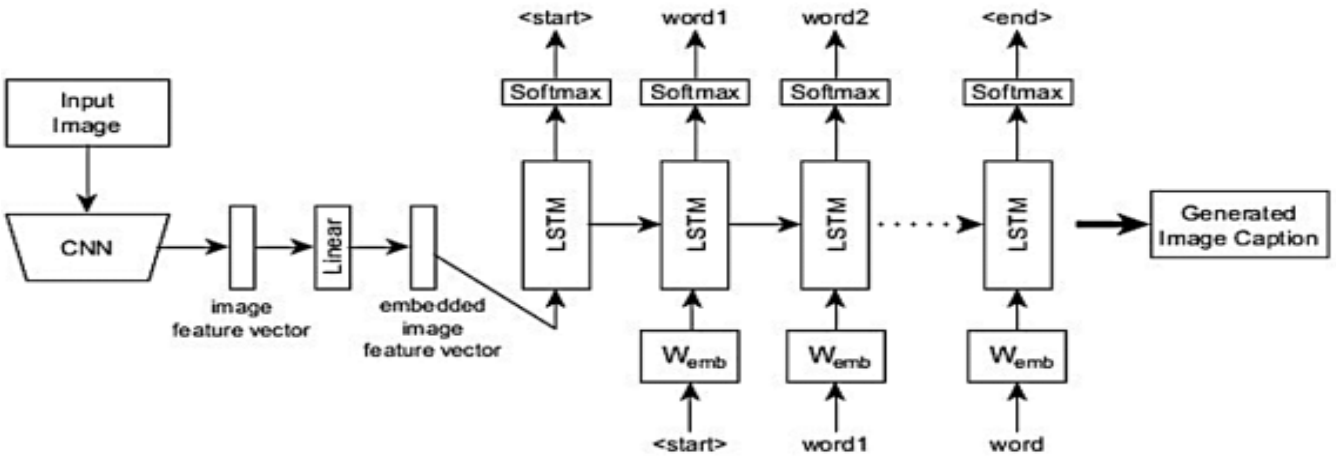


Fig1. Model Architecture

V . METHODOLOGY

i. DATASET

The largest dataset is called MS COCO (Microsoft Common Objects in Context); it consists of over 300,000 images with at least five descriptions, along with class labels, picture segment labels, and an annotation set for each image caption. Microsoft has released a visual dataset that includes several pictures of everyday objects seen on challenging forums.

This puts MSCOCO apart from other object detection datasets that may pertain to particular fields of artificial intelligence. This dataset serves as an example of all 80-item classification masks, pixelclass classification by 91 categories, captioning algorithms, and full panoptic group classification for 80 thing categories. It is frequently used for training and benchmark detection tasks.

ii. PREPROCESSING

Various annotated picture collections are accessible for this kind of preliminary processing. Preprocessing entails loading the dataset and saving the InceptionV3 model's output to a disk cache.

iii. INCEPTION V3

We are now creating a tf.keras model in the inceptionV3 architecture, using the exit layer as the conversion output layer. The size of the output layer is 8x8x2048. Every image is passed over the network, and the dictionary records the resulting vector—which is shown as image name -> feature vector. Save the result to the CD once each image has been processed using InceptionV3.

IV. ALGORITHMS

a. FRAMEWORK OF ENCODER-DECODER

Initially, in the encoding step, the model that

finds a context in an image that is provided as input and uses convolutional neural networks to extract positions by indicating a relationship between spatial variables. Convolutional neural networks process images and provide all the information of a close-up picture, including light, length, width, edges, and so on.

All of the data will be represented as annotation vectors, which are collections of feature vectors.

The code's encoding part produces L annotation vectors, each of which represents a Ddimensional related object and its spatial area within the input image.

VI. MODEL TRAINING

The lowest convolutional layer's picture features are extracted using the Inception-V3 model, producing a vector with the size (8, 8, 2048), which is subsequently reshaped to (64, 2048). The CNN Encoder receives this reshaped vector and uses it to separate and assign weights to various image parts in order to comprehend the information of the image. An RNN is utilized to iterate over the image and create the caption. To create a phrase word by word, the model first transforms the image into a word vector, which is then given to LSTM cells. To create the next word in the phrase, this procedure makes use of the context vector, the prior concealed state, and pre-existing words.

The following are the implementation steps:

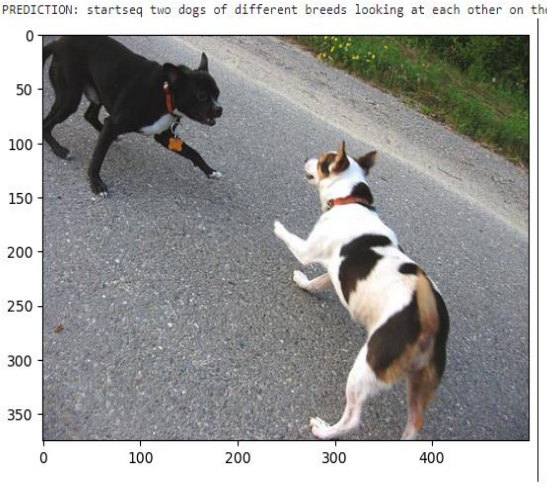
Select a set of terms that could appear in the caption for the picture. We employ the weak monitoring method in Mult instance learning (MIL) to find the words from the provided vocabulary that correspond to the content of the related image in order to train the detectors repeatedly. By using a fully convolutional network on an image, we may create a draft spatial response graph. Every location on the response map denotes a response that was found by using the original CNN on the portion of the input image where the shift was applied (basically, searching different regions of the image for possible objects). After up sampling the image to create a response map on the last completely linked layer, we apply the noisy-OR version of MIL to each response map. Each word has its own unique probability.

```
generate_caption("101669240_b2d3e7f17b.jpg")
```

```
PREDICTION: startseq man looks at framed pictures in the snow next to trees endseq
```


```
generate_caption("1001773457_577c3a7d70.jpg")
```

```
PREDICTION: startseq two dogs of different breeds looking at each other on the road endseq
```



Words are visually spotted during the caption creation process, and the most likely sentence is subsequently located. The 0e language model, which describes the probability distribution of a word sequence, is the fundamental component of this process. The maximum entropy language model (ME), although being a statistical model, has the ability to capture tremendously meaningful data.

For example, "running" is more likely to come after the word "horse" than "speaking." This information can be used to distinguish the appropriate words and possibly encode common sense knowledge.

VII. CONCLUSION AND FUTURE WORK

The suggested model makes use of an LSTM decoder to generate phrases based on the content of the image and a convolutional neural network (CNN) encoder for feature extraction. Using BLEU measures, the model improves performance on the MS COCO dataset by including attention-based approaches, with the goal of increasing the likelihood of producing a correct description for a given image. There may be uses for the photo caption generator for people with vision impairments. Subsequent studies could examine how subsampling affects CNNs and look at using auto-encoders to produce feature vectors that are more accurate. The model may also be used for jobs like video captioning and visual question answering, and it can generate long descriptions and dense captions for different languages.

REFERENCES

[1] Oluwasanmi, A., Aftab, M. U., Alabdulkreem, E., Kumeda, B., Baagyere, E. Y., & Qin, Z. (2019). Captionnet: Automatic end-to-end siamese difference captioning model with attention. IEEE Access, 7, 106773-106783.

[2] Liu, M., Li, L., Hu, H., Guan, W., & Tian, J. (2020). Image caption generation with a dual attention mechanism. Information Processing & Management, 57(2), 102178.

[3] Dehaqi, A. M., Seydi, V., & Madadi, Y. (2021).

Adversarial Image Caption Generator Network. SNComputer Science, 2(3), 1-14.

[4] Bai, S., & An, S. (2019). A survey on automatic image caption generation. Neurocomputing, 311,291-304.

[5]Han, S. H., & Choi, H. J. (2020, February).Domain-specific image caption generator withsemantic ontology. In 2020 IEEE InternationalFor papers published in translation journals, please give the English citation first, followed by the original foreign-language citation Conference on Big Data and Smart Computing(BigComp) (pp. 526-530). IEEE.

[6] Yang, M., Liu, J., Shen, Y., Zhao, Z., Chen, X.,Wu, Q., & Li, C. (2020). An ensemble ofgeneration-and retrieval-based image captioningwith dual generator generative adversarialnetwork. IEEE Transactions on Image Processing, 29, 9627-9640.

[7]Anu, M., & Divya, S. (2021, May). Building a voice-based image caption generator with deep learning. In 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS) (pp. 943-948). IEEE.

[8] Dehaqi, A. M., Seydi, V., & Madadi, Y. (2021). Adversarial Image Caption Generator Network. SNComputer Science, 2(3), 1-14.

[9]Wang, Y., Zhang, C., Wang, Z., & Li, Z. (2022, July). Image Captioning According to User's Intention and Style. In 2022 International Joint Conference on Neural Networks (IJCNN) (pp. 1-9).IEEE.

[10]Verma, A., Saxena, H., Jaiswal, M., & Tanwar, P. (2021, July). Intelligence Embedded Image Caption Generator using LSTM-based RNN Model. In 2021 6th International Conference on Communication and Electronics Systems

(ICCES) (pp. 963-967). IEEE