GENERATION OF AUTOMATIC IMAGE CAPTIONING USING NATURAL LANGUAGE PROCESSING

Charitha pally
Charithapally425@gmail.com

Syeda Fathiha Buttal
sbutt3@unh.newhaven.edu

Vineeth Kumar Deepala
vineethd99@gmail.com

Abstract--- Image captioning is a job that combines computer vision and natural language processing. The features of NLP (Natural Language Processing) models are used in the research study. Models are typically evaluated according to a BLEU or CIDER metric, with the goal of generating meaningful legends for pictures.

1. Introduction

Image caption generation is based on CNN and RNN functionality. The development was done in Jupyter Notebook and the Keras Library was used. The Python programming language was used to implement this work.

This challenge was hypothetical even to the most advanced researchers in Computer Vision before the recent emergence of Deep Neural Networks. However, with the introduction of Deep Learning, this challenge may be readily handled provided the necessary dataset is available.

In addition to CNN, Natural Language Processing is employed to create image captions. LSTMs are specialized Recurrent Neural Networks that allow data to be retained. For picture classification, the VGG16 model has been used, which has been pre-trained on the ImageNet dataset.

2. Motivation

We must first comprehend the significance of this issue in real-world circumstances. Consider a few scenarios in which a solution to this challenge may be extremely beneficial.

Google Photos: Classify your photo into Mountains, sea, etc.

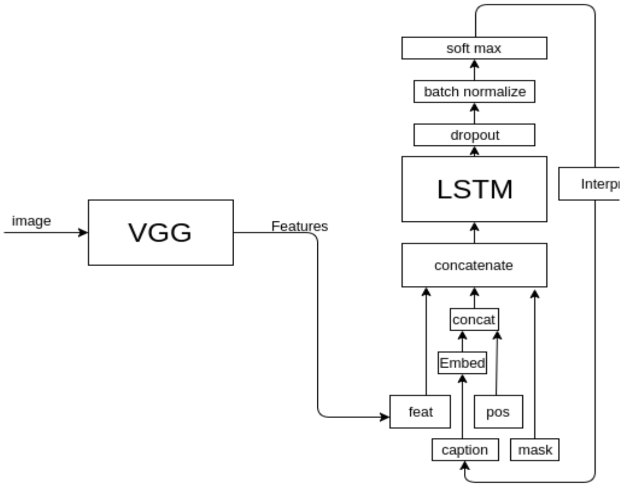
A U.S. company is predicting crop yield using images from satellites.

Self-driving cars: Automatic driving is one of the most difficult difficulties and captioning the area around the automobile can help the self-driving system.

CCTV cameras are already ubiquitous, but if we can provide appropriate captions in addition to watching the world, we can trigger warnings as soon as criminal behavior is detected someplace. This is likely to help minimize crime and/or accidents.

FedEx and other delivery firms have been employing handwritten digit recognition systems to accurately detect pin codes for quite some time now.

3. Workflow



To begin, the input image is run through a Convolutional Neural Network (CNN) to identify the items and scenes present. For the pre-processed model, transfer learning is also employed. CNN employs different ideas like as pooling, padding, and filtering. The CNN model will produce a set of words/objects as its output. Following that, Natural Language Processing (NLP) is used to assist us in communicating with the computer.

Finally, the Flickr8k text dataset is used to train the Recurrent Neural Network. The recognized items are sent into the RNN after some processing, and the RNN generates a suitable caption. To better understand the workflow, look at the graphic.

4. Neural Networks

Humans have the ability to think and remember information. Artificial intelligence is attempting to imitate this behavior. And it is on this foundation that Neural Networks are built. They may be thought of as a set of algorithms attempting to emulate the functions of the human brain. The human brain is made up of networks of neurons that operate as signal transmitters. Similarly, there are numerous layers in neural networks. They try to figure out what the underlying relationships are in the data. There are several levels to it. A perceptron is the name for each node in a neural network. A perceptron is a model of a single neuron that served as a forerunner to bigger neural networks. It feeds the signal generated by multiple linear regression into a nonlinear activation function.

5. Convolution Neural Networks

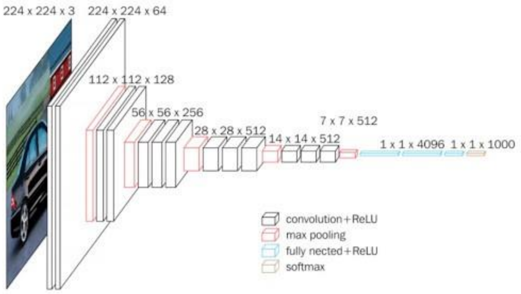
Convolutional neural networks, often known as CNN or ConvNet, are a type of artificial neural network commonly used for image processing. It may also be utilized to solve classification and data analysis issues. A CNN is an Artificial Neural Network that detects patterns using classification and attempts to deduce meaning from them.

CNN is frequently utilized since it has addressed image annotation difficulties effectively and with high accuracy. For feature extraction from picture datasets, we trained and evaluated two alternative models. The two models have varied capacities when it comes to extracting image features, with both models using 224 224 3 input pictures and VGG using 4096 convolutional features.

In more technical terms, CNN is a deep learning technique that allows a user to enter a picture and the algorithm to assign learnable biases and weights to distinct objects or characteristics in the image, allowing it to distinguish one thing from another.

A. VGG16

The VGG-16 is a 16-layer deep convolutional neural network. The ImageNet database contains a pre-trained version of the network that has been trained on over a million photos. The network can identify photos into 1000 different item categories, including keyboard, mouse, pencil, and a variety of animals. As a result, the network has picked up rich feature representations for a variety of pictures.



On the ImageNet dataset, VGG16 was shown to be the highest performing model out of all the setups. Let's have a look at the architecture of this setup.

A fixed-size 224 by 224 picture with three channels – R, G, and B – is regarded as the input to any of the network setups. The only pre-processing done is to normalize each pixel's RGB values. Every pixel is subtracted from the mean value to achieve this.

Following ReLU activations, the image is sent through the first stack of two convolution layers with a very tiny receptive area of 3 × 3. There are 64 filters in each of these two levels. The padding is 1 pixel, while the convolution stride is fixed at 1 pixel. The spatial resolution is preserved in this arrangement, and the output activation map is the same size as the input picture dimensions. The activation maps are then run via spatial max pooling with a stride of 2 pixels over a 2 x 2-pixel frame. The size of the activations is reduced by half. The activations at the end of the first stack are thus 112 x 112 x 64.

The activations are then sent through a second stack, this time with 128 filters instead of 64 in the first. As a result, after the second layer, the dimensions are 56 x 56 x 128. The third stack has three convolutional layers and a max pool layer. The stack's output size is 28 x 28 x 256 due to the 256 filters used. Then there are two stacks of three convolutional layers, each with 512 filters. Both stacks' output will be 7 × 7 x 512.

Following the convolutional layer, stacks are three fully linked layers separated by a flattening layer. The first two layers each contain 4,096 neurons, while the output layer has 1,000 neurons, matching the 1,000 possible classes in the ImageNet dataset. The SoftMax activation layer is used for categorical categorization after the output layer.

B. Training VGG16

The Keras Applications library also includes a pre-trained VGG16 model. The ImageNet weights are included in the pre-trained model. We may employ transfer learning methods to train on your custom photos while using the pre-trained model.

C. CNN Workflow

Filters are incorporated into every layer of a Convolutional Neural Network. Filters are in charge of identifying certain patterns or features in data when input is provided. The number of filters that each layer should have been specified. The network filters are simple at first, detecting patterns such as edges, circles, and so on, but as we progress through the layers, these filters grow more powerful, able to recognize full figures such as mice, cats, and so on. Filters may be thought of as a matrix with a certain number of rows and columns. The matrix blocks can be initialized with any random integer. LSTMs have a chain-like structure as well, but the repeating module is different.

D. Padding

When moving from one layer to the next, padding is utilized to safeguard the length and width of the input picture. Padding allows for the creation of a deeper network. Because the information is retained at the boundaries, the performance is believed to be better.

E. Activation Function

The activation function is a node that is preserved in the middle or at the end of a neural network. An activation function aids in determining whether to activate the neuron. It's a non-linear function that's applied to the input signal before it's modified and passed to the next layer of neurons, where it's treated as input. The activation functions RELU and SoftMax were utilized.

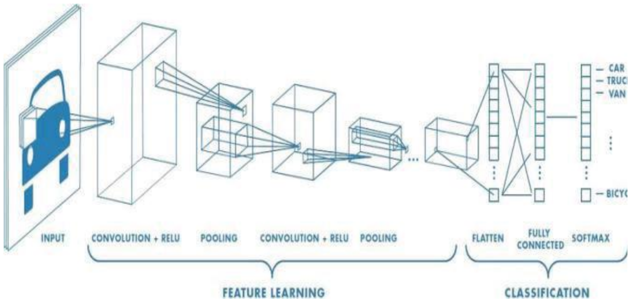


Fig: Workflow of CNN

The input layer is used to provide input to our network, which should be a three-dimensional picture. It might be colorful or black and white.

After that, the picture is transmitted to the convolutional layer. Filters are applied to the input picture in this layer.

The number of filters is set. Convolution is the term for this procedure. After that, the activation function is used. Here, the RELU activation function is utilized. Then our output becomes an input to the pooling layer, which helps to reduce the image's resolution. The convolutional layer, on the other hand, goes through the same procedure. More than one convolutional layer is possible.

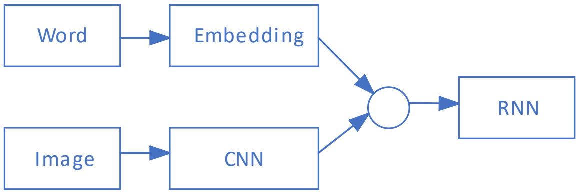
As we progress through the convolutional layers, the patterns get more complex, such as eyes, faces, and birds. The matrix is then flattened before the completely linked layer is added. This layer aids in classifying the items in the supplied picture. The probability of classes is calculated using the SoftMax formula. Finally, the output layer is created, which contains the items that were present in the picture.

6. RNN Networks

Humans do not start thinking all over again every second. You comprehend each word in this essay depending on your grasp of prior words. You don't chuck everything away and start from the beginning. Your ideas are persistent.

This is something that traditional neural networks can't achieve, and it appears to be a fundamental flaw. Consider how you would categorize the type of event that occurs at each moment in a movie. It's unclear how a typical neural network might utilize prior events in the movie to guide subsequent ones.

Although using RNN as a language model is not as prevalent as the prior method, it has produced some excellent results that outperform the previous method. Using this strategy, we will create an image captioning model. The word embeddings are sent into the RNN, and the RNN's final state is merged with picture data and fed into another neural network to predict the caption's next word.

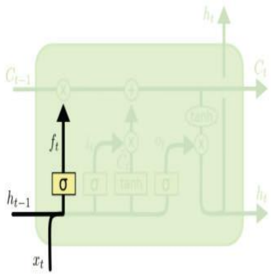


LSTMs, a particular specific type of recurrent neural network that performs far better than the normal version for many tasks. Recurrent neural networks are used to create almost all fascinating results.

7. LSTM Networks

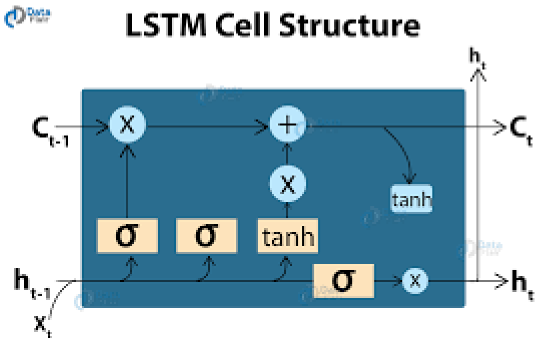
Long Short-Term Memory networks, or "LSTMs," are a kind of RNN that can learn long-term dependencies. Hoch Reiter & Schmid Huber (1997) introduced them, and numerous individuals developed and popularized them in subsequent work. 1 They are currently frequently utilized and function exceptionally effectively in a wide range of situations.

LSTMs are specifically developed to prevent the problem of long-term reliance. They don't have to work hard to remember knowledge for lengthy periods of time; it's nearly second nature to them.



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

LSTMs have a chain-like structure as well, but the repeating module is different. Instead of one neural network layer, there are four, each interacting in a unique way.



8. Natural Language Processing

Natural Language Processing, or NLP, is a branch of artificial intelligence that aids in computer communication. Computers can now perform activities that were previously impossible, like reading tests, hearing and understanding what humans are trying to say, and even interpreting the crucial bits, thanks to NLP. NLP may be used to translate text from one language to another, among other things. It may also be used in programs like Grammarly and others to fix grammatical errors, as well as in call centers to reply to diverse consumers. NLP is used to control personal assistant apps like Alexa and Siri.

Natural language processing is employed after the CNN technique in our picture captioning research study. We use NLP to use algorithms that detect natural language norms so that our language may be transformed into a format that computers can understand. As a result, NLP is employed to assist humans in communicating with the computer.

9. Evaluation metrics

To check the quality of the translated text generated from input images, we used the BLEU algorithm, which stands for Bilingual Evaluation Understudy. BLEU compares the candidate translation with existing human-generated translations, known as reference translations. Even two good human translations of the identical text may only score in the 0.6 or 0.7 range since their language and phrasing are likely to differ. The BLEU score ranges from 0 to 1. A perfect match receives a 1.0 score, whereas a perfect mismatch receives a 0.0 value. The score was created to assess the accuracy of automatic machine translation systems' predictions. So, the higher the score the higher will be the quality of generated translated text. Let us discuss more BLEU.

BLEU:

The BLEU metric is based on the candidate's accuracy value. The accuracy is calculated by dividing the total number of words in the proposed translation by the number of unigrams that appear in the reference.

The number of instances of a candidate word is clipped by the number of times it occurs in the reference translation and then divided by the total number of (unclipped) words in the candidate translation by BLEU's modified n-gram accuracy.

The averaged ratio of n-gram matches is called bleu. We calculate the ratio of the number of i-gram tuples in the candidate that also appear in the reference for each i-gram, where i=1, 2…. N.

$$p(i) = \frac{matched(i)}{H(i)}$$

where H(i) is the number of i-gram tuples in the candidate.

The BLEU metric solely considers the model's adjusted precision. Some models have extremely high BLEU ratings yet would be considered poor performers by a human.

10. Applications

- Allows you to determine whether a skin issue is skin cancer or not.
- Aids in the automation of the picture tagging.
- Helps you organize files without becoming bogged down in the job of image captioning.
- Visually challenged persons who can read the material in a much bigger font may find it useful.
- Facebook uses image captioning and find the patterns in between different photos.

11. Process

A. *Data collection*

For the Data set, we used the Flickr 8k dataset. This dataset contains 8000 images, and each image has 5 captions which are the given captions for the model to train and learn the generated translated text.

B. *Data Preparation*

In the process of data preparation, first, the captions will be tokenized (for example, by splitting into spaces). This will assist us in creating a lexicon of all the data's unique terms.

C. *Data Cleaning*

Prior to the application of the NLP algorithms to the data, the data must first be cleaned and readied for analysis. If this step is neglected, the analysis phase will be completely ruined. Only once the data is cleaned, it make sense to extract features.

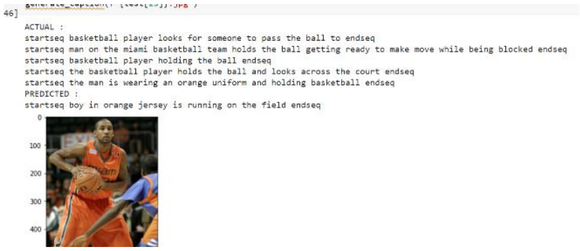
D. Data Extraction

After text cleaning, some features will be extracted since they are more useful at this point. We must aim to extract as many features as possible throughout this procedure, as more features may give helpful information during text analysis. We don't have to be concerned about whether the features will be beneficial in the future.

12. Results

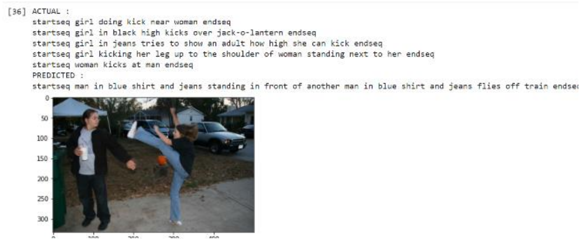
Here are the results from the code that was uploaded on My GitHub. I am sharing the link below:

https://github.com/dvk69/Image-Captioning.git



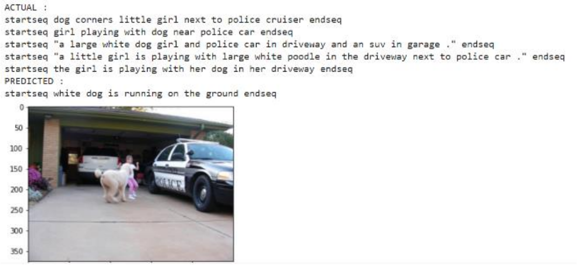
```
    generate_caption(
46]
    ACTUAL :
    startseq basketball player looks for someone to pass the ball to endseq
    startseq man on the miami basketball team holds the ball getting ready to make move while being blocked endseq
    startseq basketball player holding the ball endseq
    startseq the basketball player holds the ball and looks across the court endseq
    startseq the man is wearing an orange uniform and holding basketball endseq
    PREDICTED :
    startseq boy in orange jersey is running on the field endseq
```

The Above image gives us that the boy in orange jersey is running on the field which is almost same but not exact.



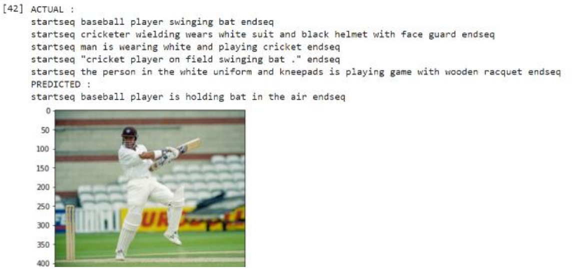
```
[36] ACTUAL :
     startseq girl doing kick near woman endseq
     startseq girl in black high kicks over jack-o-lantern endseq
     startseq girl in jeans tries to show an adult how high she can kick endseq
     startseq girl kicking her leg up to the shoulder of woman standing next to her endseq
     startseq woman kicks at man endseq
     PREDICTED :
     startseq man in blue shirt and jeans standing in front of another man in blue shirt and jeans flies off train endse
```

We have a less appropriate prediction but still our model is able to find out the colors and actions.



```
ACTUAL :
startseq dog corners little girl next to police cruiser endseq
startseq girl playing with dog near police car endseq
startseq "a large white dog girl and police car in driveway and an suv in garage ." endseq
startseq "a little girl is playing with large white poodle in the driveway next to police car ." endseq
startseq the girl is playing with her dog in her driveway endseq
PREDICTED :
startseq white dog is running on the ground endseq
```

This image above gives correct prediction as the dog runs on the ground



```
[42] ACTUAL :
     startseq baseball player swinging bat endseq
     startseq cricketer wielding wears white suit and black helmet with face guard endseq
     startseq man is wearing white and playing cricket endseq
     startseq "cricket player on field swinging bat ." endseq
     startseq the person in the white uniform and kneepads is playing game with wooden racquet endseq
     PREDICTED :
     startseq baseball player is holding bat in the air endseq
```

This result is not as accurate as it should be, but it gives the prediction as holding the bat.

13.Conclusion

Image Captioning is an effective neural network system that will read an image and generate the captions. It is based on Convolution Neural Networks where the model is trained to improve the likelihood of the text when an image is given. Image Captioning may be used to communicate visuals to blind or low-vision persons who rely on noises and text to describe a sight. It's standard practice in web development to offer a description for each picture that appears on the website, so that the image may be read or heard rather than just viewed.

REFERENCES & CITE SOURCES

1.Quanzeng You, HailinJin, Zhaowen Wang, Chen Fang, and JieboLuo. Image captioning with semantic attention. CoRR,abs/1603.03925, 2016.

2.M. Malinowski, M. Rohrbach, M. Fritz, Ask your neurons:a neural based approach to answering questions about images, in International Conference on Computer Vision, 2015

3.X. Chen, C. Lawrence Zitnick, "Mind's eye: A recurrent visual representation for image caption generation", CVPR, 2015

4.O. Vinyals, A. Toshev, S. Bengio, D. Erhan, “Show and Tell: Lessons learned from the 2015 MSCOCO Image Captioning Challenge IEEE transactions on Pattern Analysis and Machine Intelligence,2016.

5.Aditya, A. N., Anditya, A. and Suyanto, (2019). “Generating Image Description in the Indonesian Language using Convolutional Neural Network and Gated Recurrent Unit”, 7th International Conference on Information and Communication Technology (ICoICT)

6. Yajurv, B., Aman, B., Deepanshu, R., and Himanshu, M.(2019). “Image Captioning using Google’s Inceptionresnetv2 and Recurrent Neural Network”,IEEE.

7. Zakir, H., Ferdous S., and Mohd F. S. (2018). “A Comprehensive Survey of Deep Learning for Image Captioning”, ACM Computing Surveys

8.Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. Every picture tells a story: Generating sentences from images. InProceedings of the 11th European Conference on Computer Vision:Part IV, ECCV’10, pages 15–29, Berlin, Heidelberg, 2010. Springer-Verlag.

9. Bahdanau, Dzmitry, Cho, Kyunghyun, and Bengio, Yoshua. Neural machine translation by jointly learning to align and translate. arXiv:1409.0473, September 2014

[1] Md. Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. 2018. A Comprehensive Survey of Deep Learning for Image Captioning. ACM Comput. Surv. 0, 0, Article 0 (October2018), 36 pages. Computing methodologies Machine learning; Neural networks.

10.Denil, Misha, Bazzani, Loris, Larochelle, Hugo, and de Freitas, Nando. Learning where to attend with deep architectures for image tracking. Neural Computation, 2012.

11.X. Chen, C. Lawrence Zitnick, "Mind's eye: A recurrent visual representation for image caption generation", CVPR, 2015 12.Tieleman, Tijmen and Hinton, Geoffrey. Lecture 6.5 - rmsprop. Technical report, 2012

12. Karen Simonyan, Andrew Zisserman, Very Deep Convolutional Networks For Large scale image recognition, in International Conference on Learning Representation, 2015.

13. Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet

classication with deep convolutional neural networks. Advances in neural information processing systems. 2012.

14. Karpathy, Andrej, and Li Fei-fei. visual-semantic alignments for generating image descriptions. arXiv preprint arXiv: 1412.2306. (2014).

15. Wei, Yunchao, et al. CNN: Single-label to multi-label. arXiv preprint arXiv:1406.5726 (2014).

16. Deng, Jia, et al. Imagenet: A large-scale hierarchical image database. Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference

17. xiangyutang2 Image captioning online blog.

18.My great learning Introduction to VGG16 online blog.