**GENERATION OF AUTOMATIC IMAGE CAPTIONING USING NATURAL LANGUAGE PROCESSING**

**Team Members**

**Farhan Shaik Mastanvali Shaik Siva Leela Yarmala**

**Abstract:**

Automatically creating the description or caption of an image using any natural language sentences is a very challenging task. It requires both methods from computer vision to understand the content of the image and a language model from the field of natural language processing to turn the understanding of the image into words in the right order. Our project aims to implement an Image caption generator that responds to the user to get the captions for a provided image. The ultimate purpose of Image caption generator is to make users experience better by generating automated captions. In this an Image caption generator, Basis on our provided image It will generate the caption from our trained model. The basic idea behind this is that users will get automated captions when we use or implement it on social media or on any applications.

**Introduction:**

Image caption generation is based on CNN and RNN functionality. The Keras Library is used, and the development was carried out in Jupyter Notebook. To carry out this task, the Python programming language is used.

Before the recent development of Deep Neural Networks, even the most accomplished researchers in computer vision considered this task to be hypothetical. But now that deep learning has been developed, this problem might be easily solved if the required dataset is accessible.

Natural Language Processing is used in addition to CNN to generate image captions. Retention of data is made possible by specialized recurrent neural networks called LSTMs. The VGG16 model, which was trained on the ImageNet dataset, has been used for picture classification.

**Motivation:**

In the past few years, computer vision in the image processing area has made significant progress, like image classification and object detection. Benefiting from the advances of image classification and object detection, it becomes possible to automatically generate one or more sentences to understand the visual content of an image, which is the problem known as Image Captioning. Generating complete and natural image descriptions automatically has large potential effects, such as titles attached to news images, descriptions associated with medical images, text-based image retrieval, information accessed for blind users, human- robot interaction. These applications in image captioning have important theoretical and practical research value.Image captioning is a more complicated but meaningful task in the age of artificial intelligence.

**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.

For a long time now, **FedEx** and other delivery companies have used handwritten digit recognition technologies to precisely detect pin codes.

**Workflow:**

**A picture containing text, diagram, map, screenshot

Description automatically generated**

To identify the objects and scenes present, the input image is first processed through a Convolutional Neural Network (CNN). Also used with the pre-processed model is transfer learning. CNN uses a variety of techniques, including filtering, padding, and pooling. The output of the CNN model is a collection of words or objects. The use of Natural Language Processing (NLP) to facilitate communication with the computer comes next.

The Recurrent Neural Network is then trained using the Flickr8k text dataset. After some processing, the identified objects are passed to the RNN, which creates an appropriate caption. View the graphic to better understand the workflow.

**Convolution Neural Networks:**

CNN stands for Convolutional Neural Network, a type of deep neural network commonly used in computer vision tasks, such as image classification, object detection, and image segmentation.

CNNs are inspired by the visual cortex of the brain, which is responsible for processing visual information. The network consists of multiple layers, including convolutional layers, pooling layers, and fully connected layers.

Convolutional layers apply a set of learnable filters to the input image, which helps to extract relevant features, such as edges, textures, and shapes. Pooling layers reduce the spatial dimensions of the feature maps, which helps to make the network more computationally efficient.

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.

**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.

A picture containing text, diagram, screenshot, plan

Description automatically generated

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.

**Training VGG16:**

The Kera’s 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.

**CNN WORKFLOW:**

Every layer of a convolutional neural network contains filters. When input is given, filters are responsible for spotting specific patterns or features in the data. Each layer's number of filters should have been mentioned. At first, the network filters are straight forward, identifying patterns like edges, circles, and so on, but as we move these filters get more effective as they are applied, being able to identify whole figures like mice, cats, and other animals. One way to conceptualize filters is as a matrix with a specific number of rows and columns. Any random integer may be used to initialize the matrix blocks. Although the repeating module in LSTMs is different, they too feature a chain-like structure.

**Padding:**

Padding in CNNs refers to adding extra layers of zeros around the input volume before performing the convolution operation. It is used to preserve the spatial dimensions of the input volume and prevent information loss at the edges. There are two types of padding: valid padding, where no padding is added and the output volume is smaller, and same padding, where padding is added so that the output volume has the same dimensions as the input volume. Same padding is preferred in CNNs to preserve spatial information. 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.

**Activation Function:**

Activation functions in convolutional neural networks (CNNs) are mathematical functions that are applied to the output of each convolutional layer to introduce non-linearity into the network. The purpose of activation functions is to transform the input signal into an output signal that can be used as input to the next layer of the network. 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.

A picture containing diagram, sketch, line, plan

Description automatically generated

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.

**RNN NETWORK:**

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 out perform 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.

**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.

A picture containing text, diagram, screenshot, design

Description automatically generated

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.

# **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.

A black text on a white background

Description automatically generated with low confidence

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.

**Process:**

**1.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.

**2.Data preprocessing:**

Preprocessed the data by resizing the images to a fixed size, and tokenizing the captions by converting them into a sequence of words.

**3. Feature Extraction:**

Used a pre-trained CNN such as VGG16 to extract visual features from the input images.The output of the CNN is a feature vector that represents the image.

**4.Text Preprocessing:**

Created a vocabulary of words that occur in the captions and convert the captions into sequences of integers representing the words.

**5.Model Architecture:**

Built a model that consists of an LSTM network that takes the visual features as input and generates a sequence of words as output. The LSTM network can be trained using teacher forcing, where the model is fed the correct sequence of words at each time step during training.

**6.Training:**

Trained the model using the preprocessed data and monitor the loss to prevent overfitting. The loss function we used the categorical cross-entropy loss and a Adam optimizer.

**7.Evaluation:** Evaluated the model on a held-out test set using metrics such as BLEU, which measure the similarity between the generated captions and the ground-truth captions.

**Applications:**

* E-commerce: Image captioning can help to automatically generate product descriptions for online shopping platform.
* Surveillance: Image captioning can be used to automatically identify and classify objects and events in surveillance footage.
* Content Moderation: Image captioning can be used to automatically detect and flag inappropriate or offensive content in images and videos.
* Robotics: Image captioning can be used to improve the visual perception capabilities of robots, enabling them to navigate and interact with their environment more effectively.
* Automated image analysis: Image caption generators can be used to automatically analyze images and provide insights into their contents. This can be useful for applications like surveillance, medical imaging, or autonomous vehicles.
* Education: Image caption generators can be used in educational settings to teach students how to describe images and understand their contents. They can also be used to generate quizzes or assessments based on visual content.

**Results:**

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

<https://github.com/farhan965/Image_captioning>A picture containing text, screenshot, sky, beach

Description automatically generatedA person on a snowmobile

Description automatically generated with low confidence

**Conclusion:**

An efficient neural network system called image captioning can read an image and produce captions. Convolution neural networks are the foundation of this system, and the model is trained to increase the likelihood of the text when an image is provided. In order to explain images to those who are blind or have low vision and rely on sounds and text to describe a scene, image captioning may be utilized. It's usual procedure in web construction to provide a description for each image that appears on the website, allowing users to read or listen to the image instead of just viewing it.

**References:**

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.
2. X. Chen, C. Lawrence Zitnick, "Mind's eye: A recurrent visual representation for image caption generation", CVPR, 2015
3. Yajurv, B., Aman, B., Deepanshu, R., and Himanshu, M.(2019). “Image Captioning using Google’s Inceptionresnetv2 and Recurrent Neural Network”,IEEE.
4. 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. Bahdanau, Dzmitry, Cho, Kyunghyun, and Bengio, Yoshua. Neural machine translation by jointly learning to align and translate. arXiv:1409.0473, September 2014