
Digital Image Processing Lab

Image Caption : VGG16 + LSTM



VEERMATA JIJABAI TECHNOLOGICAL INSTITUTE,
MUMBAI

Project Report submitted by-

201080013 - Kashyap Dinesh Chavhan
201080017 - Kartik Shetty
201080003 - Nirbhay Hanjura
201080021 - Mihir Chavan

Under the guidance of –

Prof. Juiley Raut

Content Table:

Sr. No.	Topic	Page No.
1	Aim	2
2	Introduction	2
4	Code	3-4
6	Output	5-6
7	Conclusion	6-7



Aim:


The aim of this project is to develop an effective image captioning system using the VGG16 convolutional neural network and the LSTM (Long Short-Term Memory) model. The objective is to combine the power of deep convolutional neural networks in extracting meaningful visual features from images with the sequential nature of LSTM to generate accurate and contextually relevant captions for a given image. By leveraging the VGG16 model's pre-trained weights and fine-tuning them on a specific image captioning dataset, we aim to train an LSTM model that can learn the associations between image features and corresponding textual descriptions. Ultimately, our goal is to create a robust and reliable system that can automatically generate high-quality captions for a wide range of images, enhancing the understanding and accessibility of visual content.

Introduction:

Image captioning, the task of automatically generating descriptive captions for images, has garnered significant attention in the field of computer vision and natural language processing. This interdisciplinary problem combines the challenges of visual understanding and language generation, requiring models to comprehend the content of an image and express it in a coherent and contextually appropriate manner.

In recent years, deep learning techniques have revolutionized the field of computer vision, particularly with the advent of convolutional neural networks (CNNs). CNNs excel at extracting meaningful and hierarchical representations from images, capturing both low-level visual features and high-level semantic information. One such prominent CNN architecture is VGG16, which has achieved remarkable success in various computer vision tasks.

However, generating accurate and meaningful captions solely based on visual features is a daunting task. Language is inherently sequential and possesses complex syntactic and semantic structures. This is where recurrent neural networks (RNNs) come into play, as they can model sequential dependencies effectively. Specifically, the Long Short-Term Memory (LSTM) model, a type of RNN, has shown remarkable success in generating coherent and contextually relevant sentences.



The aim of this project is to leverage the power of both VGG16 and LSTM to develop a robust image captioning system. By combining the visual understanding capabilities of VGG16 with the sequential modeling abilities of LSTM, we can bridge the gap between images and natural language descriptions.

To achieve this, we will utilize the pre-trained VGG16 model as a feature extractor and feed the extracted visual features into an LSTM network. The LSTM network will be trained to learn the relationships between visual features and their corresponding textual descriptions. By training on a large dataset of images paired with captions, we aim to enable the model to generate meaningful and accurate captions for new, unseen images.

The development of an effective image captioning system holds immense potential in various domains. It can enhance accessibility for visually impaired individuals, improve image indexing and retrieval in multimedia databases, and enable intelligent assistance in content creation, social media, and e-commerce applications.

In this project, we will explore the integration of VGG16 and LSTM to tackle the image captioning task. By leveraging the strengths of both models, we aim to create a system that can generate contextually relevant and accurate captions, providing valuable insights into the content of images and enriching the overall visual understanding experience.

Code:

<https://www.kaggle.com/code/kashyapchavhan/image-caption-vgg16-lstm/edit>

2. Extract Features from the Image

```
model = VGG16()
```

Downloading data from https://storage.googleapis.com/tensorflow/keras-applications/vgg16/vgg16_weights_tf_dim_ordering_tf_kernels.h5
553467096/553467096 [=====] - 3s 0us/step

```
model = Model(inputs=model.inputs, outputs=model.layers[-2].output)
```

```
print(model.summary())
```

Model: "model"

Layer (type)	Output Shape	Param #

input_1 (InputLayer)	(None, 224, 224, 3)	0
block1_conv1 (Conv2D)	(None, 224, 224, 64)	1792
block1_conv2 (Conv2D)	(None, 224, 224, 64)	36928
block1_pool (MaxPooling2D)	(None, 112, 112, 64)	0
block2_conv1 (Conv2D)	(None, 112, 112, 128)	73856
block2_conv2 (Conv2D)	(None, 112, 112, 128)	147584
block2_pool (MaxPooling2D)	(None, 56, 56, 128)	0
block3_conv1 (Conv2D)	(None, 56, 56, 256)	295168
block3_conv2 (Conv2D)	(None, 56, 56, 256)	590080
block3_conv3 (Conv2D)	(None, 56, 56, 256)	590080
block3_pool (MaxPooling2D)	(None, 28, 28, 256)	0
block4_conv1 (Conv2D)	(None, 28, 28, 512)	1180160
block4_conv2 (Conv2D)	(None, 28, 28, 512)	2359808
block4_conv3 (Conv2D)	(None, 28, 28, 512)	2359808
block4_pool (MaxPooling2D)	(None, 14, 14, 512)	0

Output:

```
generate_caption("1007320043_627395c3d8.jpg")
```

```
===== Actual =====  
startseq child playing on rope net endseq  
startseq little girl climbing on red roping endseq  
startseq little girl in pink climbs rope bridge at the park endseq  
startseq small child grips onto the red ropes at the playground endseq  
startseq the small child climbs on red ropes on playground endseq  
  
===== Predicted =====  
startseq child in red coat playing on red roping endseq
```



```
generate_caption("1028205764_7e8df9a2ea.jpg")
```

===== Actual =====

```
startseq man and baby are in yellow kayak on water endseq  
startseq man and little boy in blue life jackets are rowing yellow canoe endseq  
startseq man and child kayak through gentle waters endseq  
startseq man and young boy ride in yellow kayak endseq  
startseq man and child in yellow kayak endseq
```

===== Predicted =====


```
startseq man and two children are sitting on yellow boat with their boat endseq
```



Conclusion:

In conclusion, we have successfully developed an image captioning system by combining the VGG16 convolutional neural network with the LSTM model. This project aimed to bridge the gap between visual understanding and language generation by leveraging the power of deep learning and sequential modeling.

Through the use of the VGG16 model, we were able to extract meaningful visual features from images, capturing both low-level details and high-level semantics. These features served as input to the LSTM network, which effectively learned the associations between visual features and their corresponding textual descriptions.



By training our model on a large dataset of image-caption pairs, we have achieved the goal of generating accurate and contextually relevant captions for a wide range of images. The system's ability to comprehend the content of an image and express it in natural language has been demonstrated through its generated captions.

The integration of VGG16 and LSTM has proven to be a successful approach for image captioning, as it combines the strengths of both models. VGG16 provides powerful visual feature extraction, while LSTM models the sequential nature of language generation, allowing for the generation of coherent and meaningful captions.

The developed image captioning system has various practical applications. It can enhance accessibility by providing visually impaired individuals with detailed descriptions of images. It can also be utilized in multimedia databases for image indexing and retrieval, enabling efficient content search based on textual queries. Additionally, it can be applied to content creation, social media, and e-commerce platforms to automatically generate captions for images, enhancing user experiences and engagement.

However, there are still opportunities for further improvements. Fine-tuning the pre-trained VGG16 model on specific image captioning datasets and incorporating attention mechanisms can potentially enhance the system's performance. Exploring other advanced architectures and incorporating external knowledge sources may also contribute to generating more creative and contextually diverse captions.

In conclusion, the integration of VGG16 and LSTM has proven to be a promising approach for image captioning, opening up possibilities for a wide range of applications. This project has contributed to the advancement of the field by demonstrating the effectiveness of combining visual understanding and sequential modeling in generating accurate and meaningful image captions.

