Image Captioning and Generation Using Deep Learning

G. Venkateswarlu Reddy [1], P. Lakshmi Sai Priya[2], G. Siva Kumar[3], U. Megha Syam Chowdary[4], S. Aswini[5]

#1Assistant Professor in Department of CSE-AI, PBR Visvodaya Institute of Technology and Science, Kavali.

#2#3#4#5B.Tech with Specialization of Computer Science and Engineering-Artificial Intelligence in PBR Visvodaya Institute of Technology and Science, Kavali.

ABSTRACT_ This research combines computer vision (CV) and natural language processing (NLP) approaches with state-of-the-art Image Captioning and Generation tools. With transformer designs and attention processes, our model demonstrates a sophisticated comprehension of textual and visual information processing. Robust learning and autonomous caption creation during testing are ensured by the supervised learning architecture, which uses a Convolutional Neural Network (CNN) as an image encoder and a Recurrent Neural Network (RNN) as a language decoder.

We also investigate the Stable Diffusion model in KerasCV for text-driven image generation, where performance is improved via mixed precision support and XLA compilation. The results demonstrate adaptability to a variety of datasets, which adds to the changing field of multimodal AI applications. Applications include annotating photos and creating original images in response to textual cues, which could be advantageous for content development and accessibility solutions.

1.INTRODUCTION

In earlier days Image Captioning was a tough task and the captions that are generated for the given image are not much relevant. With the advancement of Neural Networks of Deep Learning and also text processing techniques like Natural Language Processing, Many tasks that were challenging and difficult using Machine Learning became easy to implement with the help of Deep Learning and Neural Networks. These are very much useful in image recognition, Image classification, Image Captioning and many other Artificial Intelligence applications. Image Captioning is basically generating descriptions about what is happening in the given input image. Basically ,this model takes image as input and gives caption for it. With the advancement of the technology the efficiency of image caption generation is also increasing. This Image

Captioning is very much useful for many applications like Self driving cars which are now talk of the town. Image captioning can be used in many Machine Learning tasks for Recommendation Systems. There are many models proposed for image captioning like object detection model, visual attention- based image captioning and Image Captioning using Deep Learning. In Deep Learning also there are different deep learning models like Inception model, VGG Model , ResNet-LSTM model, traditional CNNRNN Model. In this paper we are going to explain about the model we have followed for captioning the images .i.e; CNN And RNN model.

2.LITERATURE SURVEY

Machine Learning is an idea to learn from examples and experience, without being explicitly programmed. Instead of writing code, you feed data to the generic algorithm, and it builds logic based on the data given. Machine Learning is a field which is raised out of Artificial Intelligence(AI). Applying AI, we wanted to build better and intelligent machines. But except for few mere tasks such as finding the shortest path between point A and B, we were unable to program more complex and constantly evolving challenges. There was a realisation that theonly way to be able to achieve this task was to let machine learn from itself. This sounds similarto a child learning from its self. So machine learning was developed as a new capability for computers. And now machine learning is present in so many segments of technology, that didn‘t even realise it while using it.

Supervised learning is the most popular machine learning paradigm. It is easy to understand and very easy to use. It is a learning function that creates a map of output inputs based on the example of input-output pairs. It takes work from a training data lab that includes a set of training examples. In supervised reading, each example is a pair that includes an input object (usually vector) and the required output value (also called the directional signal). The supervised learning algorithm analyzes training data and generates targeted activity, which can be used to map new examples. Supervised Reading is very similar to teaching a child about the data provided and that data is in the form of labeled examples, we can feed the algorithm of learning with these pairs of individual model-labels, allowing the algorithm to predict the correct answer or not. Over time, the algorithm will learn to measure the exact nature of the relationship between models and their labels. When fully trained, the supervised learning algorithm will be able to detect a new, unprecedented model and predict its excellent label.

Unsupervised learning is a machine learning method, where you do not need to monitor the model. Instead, you need to let the model work on its own for information. Works great with non-labeled data and looks for patterns that were not previously found in a set of data that does not already have labels and has minimal human monitoring. In contrast to supervised reading that often uses personal name data, unchecked reading, also known as self-organizing, allows for the creation of a dynamic model over the input.

The Neural Network (or Artificial Neural Network) has the ability to learn by example. ANN is an information processing model inspired by a biological neuron system. ANN-biologically inspired images that are computergenerated to perform a specific set of tasks such as merging, segmentation, pattern recognition etc. It is made up of a large number of highly interconnected processing devices known as neurons to solve problems. It follows a non-linear approach and processes information uniformly across all nodes. The neural network is a complex flexible system. Adaptive means it has the ability to change its internal structure by adjusting the input weights

Deep learning is a branch of machine learning based entirely on neural networks that are practiced. In-depth learning is an artificial intelligence activity that mimics the functioning of the human brain in processing data and creating patterns that will be used in decision making. In-depth learning is a subset of machine learning in artificial intelligence (AI) with networks that can read without being monitored for random or unlabeled data. It has a large number of hidden layers and is known as deep neural learning or deep neural network. Deep learning has evolved in conjunction with the digital age, which has brought an explosion of data across all genres and regions of the world. This data, known as big data, is taken from sources such as social media, online search engines, e-commerce forums, and online cinemas, among others. This large amount of data is easily accessible and can be shared with fintech applications such as cloud computing. However, the details, often irregular, are so large that it can take decades for people to understand and produce the right information. Companies are recognizing the incredible power that can come from uncovering this wealth of information and are becoming increasingly familiar with AI systems for automated support. In-depth reading learns from large amounts of informal data that can often take decades for people to understand and process. In-depth learning also uses the hierarchical level of neural networks performed to perform the machine learning process. Nervous network networks are shaped like the human brain, with neuron nodes connected as a web. While traditional systems create data analytics in a straightforward manner, the hierarchical function of in-depth learning systems enables machines to process data indirectly

3.PROPOSED SYSTEM

Image Captioning:

Image captioning is the process of synthesizing coherent and descriptive textual descriptions for images in order to convert the visual content into a machine-readable format. Here, we'll create captions automatically using training models for deep learning. Transformer encoders, decoders, and various feature extractors are used in it..

Image Generation:

Send a text to One method to turn words into pictures is through image generation. Based on your description, it generates visually engaging material using intelligent computer models. Largely, special neural networks such as CNNs and RNNs are helpful; further improvements come with the addition of transformers and attention. It's not just tech stuff; it's a potent instrument that uses the combination of words and images to bring creative ideas to life. It generates photos in an automated manner.

3.1 IMPLEMENTATION

3.1.1 HOME PAGE MODULE:

Home Page Module is the first page that is going to open when we first run **streamlit_app.py** file. This page is all about telling us the introduction to Image Captioning and Image Generation from textual prompt and tips on how to get the accurate output for the models.

3.1.2 IMAGE CAPTIONING MODULE:

Image Captioning Module is the module where we are going to upload our image and get the caption. The entry page name is pages\1_📝_Image Captioning.py

Whenever we click on the generate button internally the script invokes different scripts. The flow will be as follows.

3.1.3 IMAGE CAPTIONING MODULE:

The invoking is as follows

image_captioning.py

model_handler.py

local_models

3.1.4 IMAGE GENERATION MODULE:

Image Generation Module is the module where we are going to give prompt and displays the generated image. The entry page name is as follows pages\2_🎨🖌□🖼□_Image Generation(Synthesis).py .

Whenever we click on the generate button internally the script invokes different scripts. The flow will be as follows.

Image_Generation(Synthesis).py

generation.py

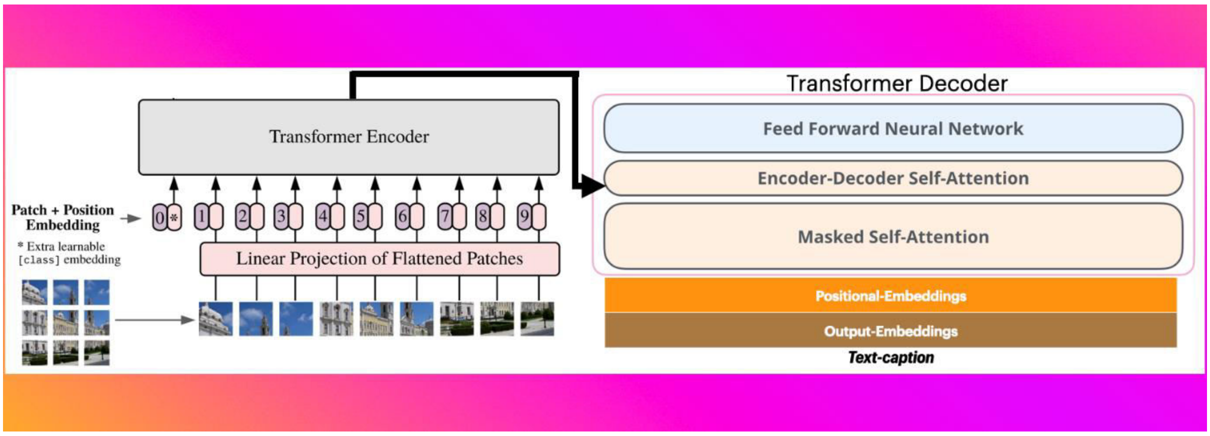
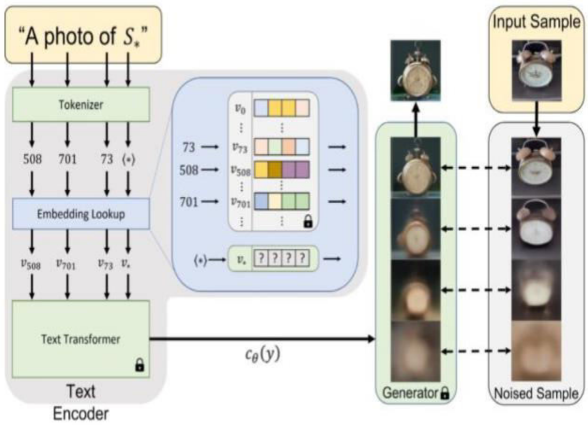

Transformer Encoder
Transformer Decoder
Feed Forward Neural Network
Encoder-Decoder Self-Attention
Masked Self-Attention
Patch + Position Embedding
* Extra learnable [class] embedding
0*
1
2
3
4
5
6
7
8
9
Linear Projection of Flattened Patches
Positional-Embeddings
Output-Embeddings
Text-caption

Fig 1:Architecture

The architecture for Image Generation using stable diffusion typically involves a sequence of transformations applied to noise vectors to generate images iteratively. Stable diffusion is a generative model framework used for image generation, which builds upon diffusion probabilistic models. Here's an architecture of the system for image generation using stable diffusion:



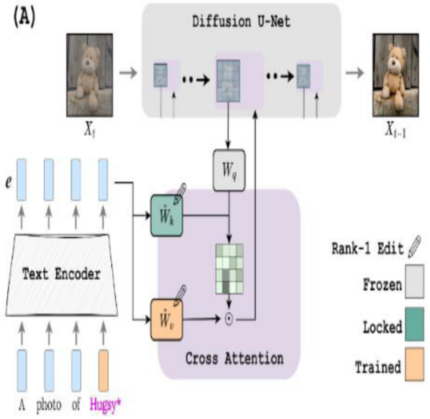


Fig 2: Image Generation from Text Modelling

4.RESULTS AND DISCUSSION

1. This page describes a Streamlit app that allows users to generate captions for images and generate images from text descriptions using machine learning models. This webpage includes instructions on how to use the app, including tips for better results. There is also some text about the models used in the app, including that they are based on state-of-the-art deep learning techniques.

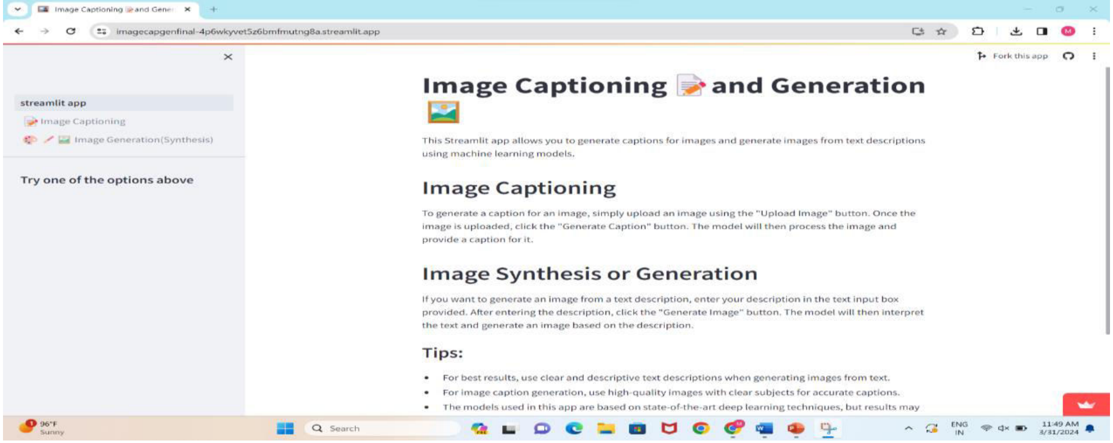


Image Captioning and Gene
imagecapgenfinal-4p6wkyvet5z6bmfmutng8a.streamlit.app
Fork this app
streamlit app
Image Captioning
Image Generation(Synthesis)
96°F
Sunny
Search
ENG
IN
11:49 AM
3/31/2024

Try one of the options above

Image Captioning and Generation

This Streamlit app allows you to generate captions for images and generate images from text descriptions using machine learning models.

Image Captioning

To generate a caption for an image, simply upload an image using the "Upload Image" button. Once the image is uploaded, click the "Generate Caption" button. The model will then process the image and provide a caption for it.

Image Synthesis or Generation

If you want to generate an image from a text description, enter your description in the text input box provided. After entering the description, click the "Generate Image" button. The model will then interpret the text and generate an image based on the description.

Tips:

- For best results, use clear and descriptive text descriptions when generating images from text.
- For image caption generation, use high-quality images with clear subjects for accurate captions.
- The models used in this app are based on state-of-the-art deep learning techniques, but results may

1. Fig.1 Home Page

2. In this section, there is a button that says "Choose an image…" and a file upload box where users can select an image from their device. There is also a button 6abelled "Generate Caption" , used to generate captions

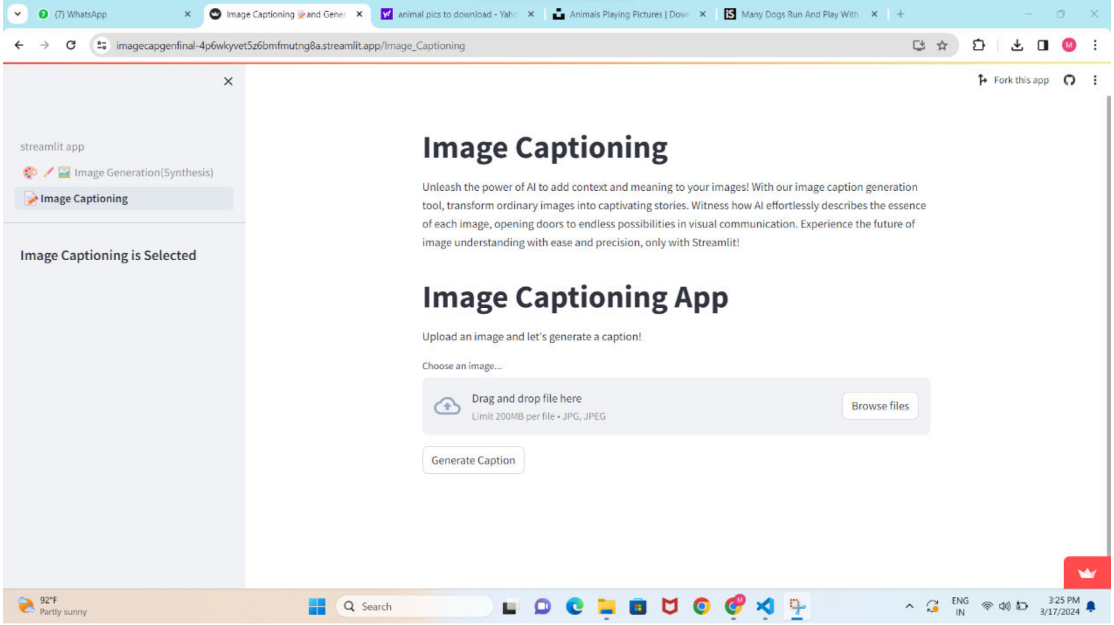

(7) WhatsApp
Image Captioning and Gener
animal pics to download - Yah
Animals Playing Pictures | Dow
Many Dogs Run And Play With
imagecapgenfinal-4p6wkyvet5z6bmfmutng8a.streamlit.app/Image_Captioning
Fork this app
streamlit app
Image Generation(Synthesis)
Image Captioning
Drag and drop file here
Limit 200MB per file • JPG, JPEG
Browse files
Generate Caption
92°F
Partly sunny
Search
ENG
IN
3:25 PM
3/17/2024

Image Captioning is Selected

Image Captioning

Unleash the power of AI to add context and meaning to your images! With our image caption generation tool, transform ordinary images into captivating stories. Witness how AI effortlessly describes the essence of each image, opening doors to endless possibilities in visual communication. Experience the future of image understanding with ease and precision, only with Streamlit!

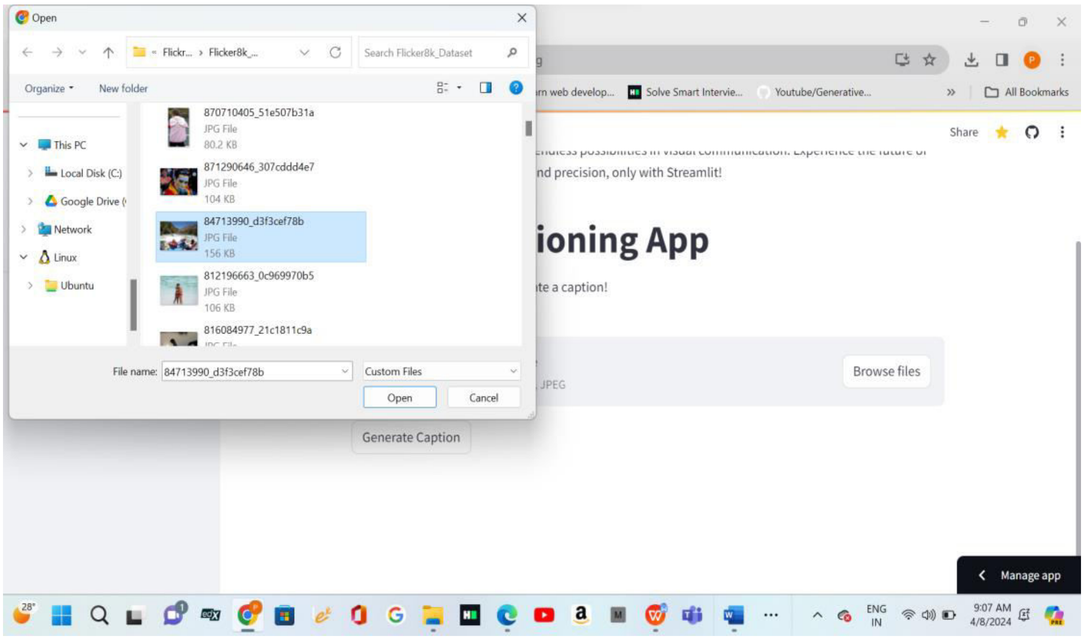
Image Captioning App

Upload an image and let's generate a caption!

Choose an image...

2. **Fig .2 Image Captioning Page**

3. Below screens represents the uploading the image from computer using Drag and drop down box, make sure that the image must be a .jpeg or .jpg format only. Any other file formats will not be considered by the model.


endless possibilities in visual communication. Experience the future of
nd precision, only with Streamlit!

ioning App

te a caption!

File name: 84713990_d3f3cef78b

Custom Files

3. **Fig.3 Image Uploading**

4.After successfully uploading the image and click “Generate Caption”, The Model will generate the caption and displayed as below.

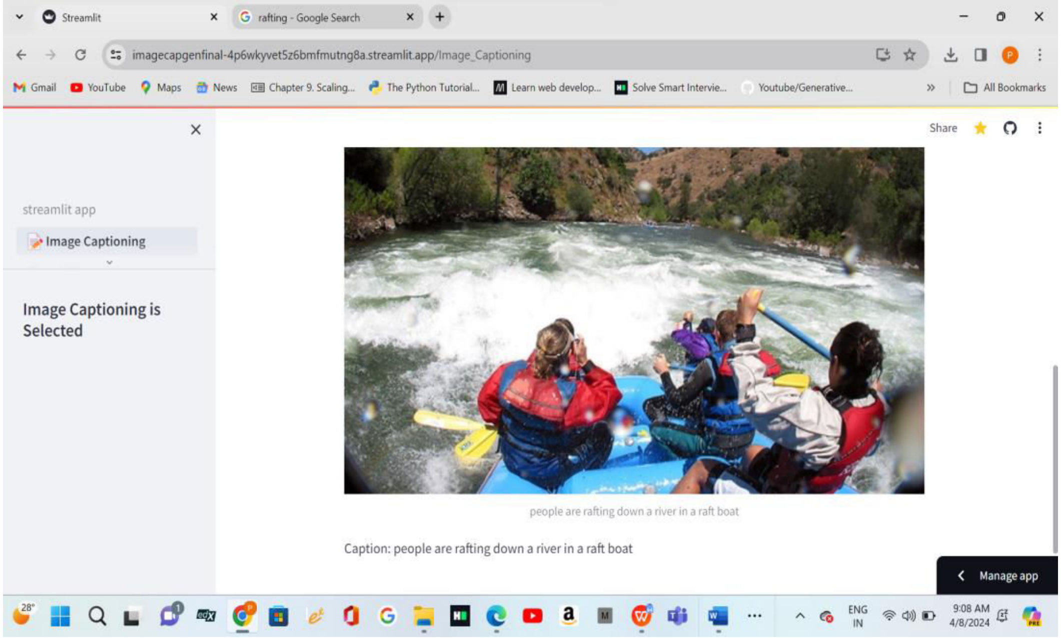

imagecapgenfinal-4p6wkyvet5z6bmfmutng8a.streamlit.app/Image_Captioning

Image Captioning is Selected

people are rafting down a river in a raft boat

Caption: people are rafting down a river in a raft boat

4. **Fig .4 Caption Generated**

5. In the Image Generation Module Page, A text box is available where users can enter a description of the image they want to create. There is also a button labeled “Generate Image", used to generate image.

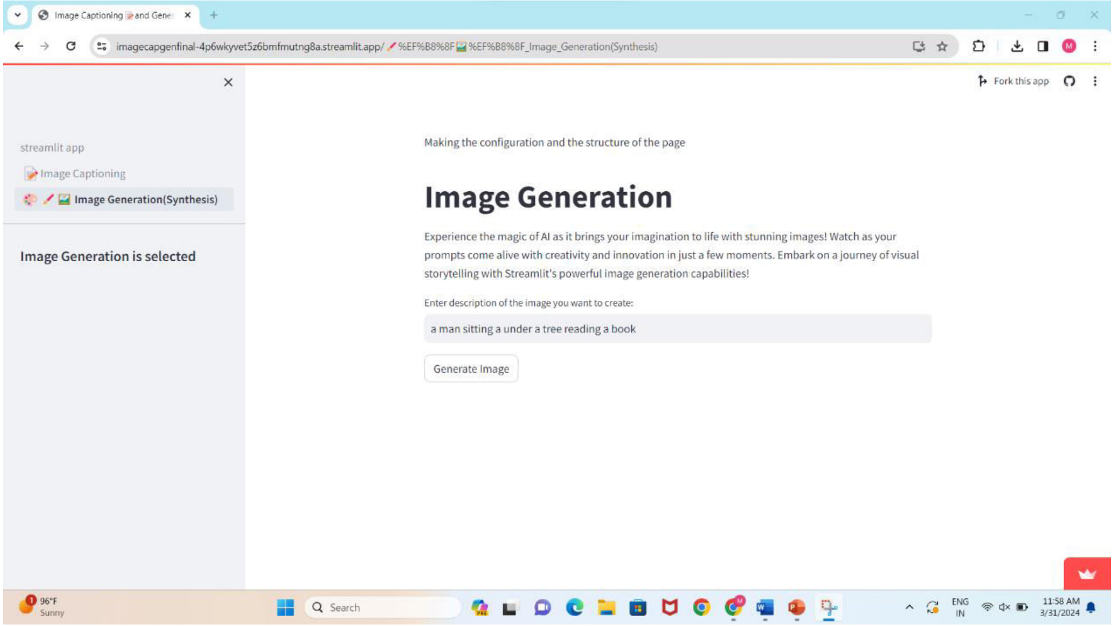


Image Captioning and Gene
imagecapgenfinal-4p6wkyvet5z6bmfmutng8a.streamlit.app/%EF%B8%8F%EF%B8%8F_Image_Generation(Synthesis)
Fork this app
96°F
Sunny
Search
ENG
IN
11:58 AM
3/31/2024

streamlit app

Image Captioning

Image Generation(Synthesis)

Image Generation is selected

Making the configuration and the structure of the page

Image Generation

Experience the magic of AI as it brings your imagination to life with stunning images! Watch as your prompts come alive with creativity and innovation in just a few moments. Embark on a journey of visual storytelling with Streamlit's powerful image generation capabilities!

Enter description of the image you want to create:

a man sitting a under a tree reading a book

Generate Image

5. **Fig .5 Image Generation Page**

6. Based on the given user input text and after clicking “Generate Image” Button, Image Generation Model generated the image which is relevant to the user input text.

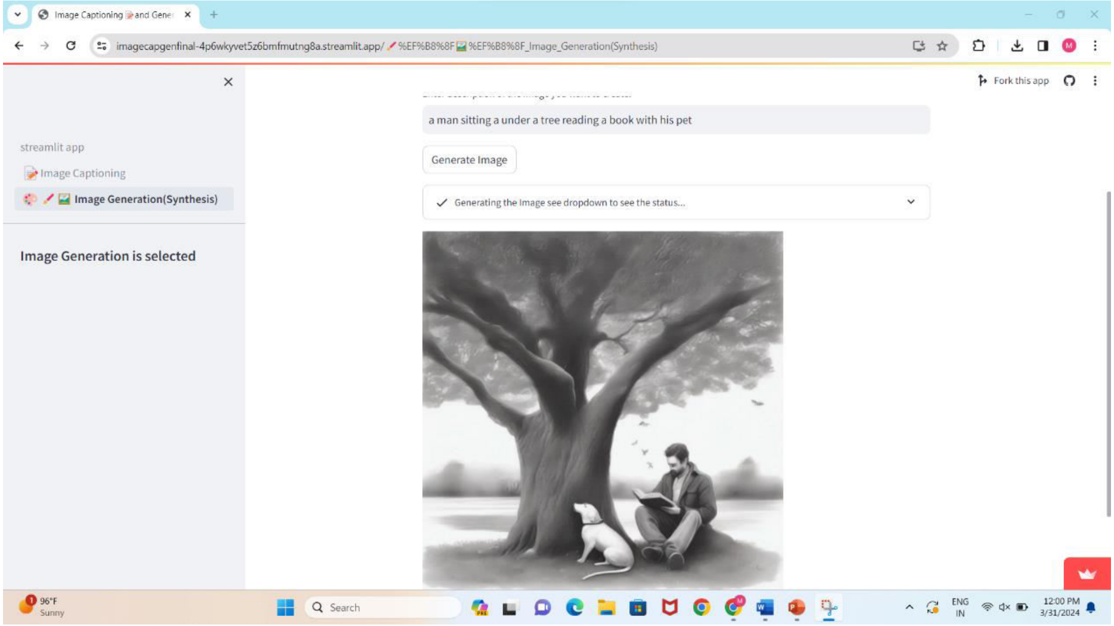


Image Captioning and Gene
imagecapgenfinal-4p6wkyvet5z6bmfmutng8a.streamlit.app/%EF%B8%8F%EF%B8%8F_Image_Generation(Synthesis)
Fork this app
96°F
Sunny
Search
ENG
IN
12:00 PM
3/31/2024

streamlit app

Image Captioning

Image Generation(Synthesis)

Image Generation is selected

a man sitting a under a tree reading a book with his pet

Generate Image

✓ Generating the image see dropdown to see the status...

6. **Fig .6 Generated Image**

5.CONCLUSION

To sum up, our investigation into image production and captioning has demonstrated the revolutionary power of sophisticated deep learning architectures, especially transformers, encoders, and decoders. By utilizing methods such as transformer-based language models and generative adversarial networks (GANs), picture synthesis and captioning have advanced, pushing the limits of semantic understanding and realism. These models provide a window into the future of AI-assisted creativity and communication because of their capacity to grasp minute visual details and subtle verbal subtleties.

Future studies on transformers, encoders, and decoders have a great deal of potential to improve the caliber and adaptability of captioning and picture generating systems. We can unlock new degrees of contextuality and realism in generated content by investigating multimodal frameworks, optimizing attention mechanisms, and scaling up model sizes. By guiding responsible innovation and ethical considerations as we navigate the complexities of this multidisciplinary field, we are paving the way for a future in which machines will seamlessly integrate with human creativity, enhancing our digital experiences and profoundly influencing our visual storytelling.

REFERENCES

1. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S. & Bengio, Y. (2014). Generative adversarial nets. In “Advances in neural information processing systems” (pp. 2672-2680).

2. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R. & Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. “International Conference on Machine Learning”.

3. Johnson, J., Alahi, A., & Fei-Fei, L. (2016). Perceptual losses for real-time style transfer and super-resolution. In “European Conference on Computer Vision” (pp. 694-711).

4. Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. In “Proceedings of the IEEE conference on computer vision and pattern recognition” (pp. 3156-3164).

5. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., & Zhang, L. (2018). Bottom-up and top-down attention for image captioning and visual question answering. In “Proceedings of the IEEE conference on computer vision and pattern recognition” (pp. 6077-6086).

6. Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In “Proceedings of the IEEE international conference on computer vision” (pp. 2223-2232).

Author’s Profiles

G.Venkateswarlu Reddy working as Assistant Professor in Department of CSE, PBRVITS, Kavali. He completed his B.Tech in Computer Science and Engineering from Gokula Krishna College of Engineering, Sullurupeta, completed his M.Tech in Computer Science and Engineering from Sree Kavitha Engineering College, Karepalli, and pursuing Phd in Shri Venkateswara University, UttarPradesh. He has 15 years of Teaching experience in various engineering colleges.

P. Lakshmi Sai Priya, B.Tech with Specialization of Computer Science and Engineering-Artificial Intelligence in PBR Visvodaya Institute of Technology and Science, Kavali.

G. Siva Kumar, B.Tech with Specialization of Computer Science and Engineering-Artificial Intelligence in PBR Visvodaya Institute of Technology and Science, Kavali.

U. Megha Syam Chowdary, B.Tech with Specialization of Computer Science and Engineering-Artificial Intelligence in PBR Visvodaya Institute of Technology and Science, Kavali.

S. Aswini, B.Tech with Specialization of Computer Science and Engineering-Artificial Intelligence in PBR Visvodaya Institute of Technology and Science, Kavali.