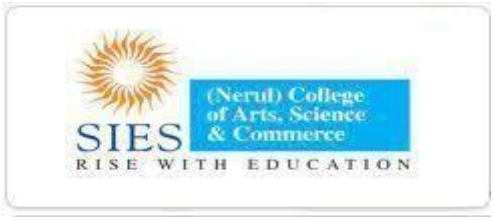


SIES (NERUL) COLLEGE OF ARTS, SCIENCE AND COMMERCE(AUTONOMOUS)

NAAC RE-ACCREDITED 'A' GRADE

SRI CHANDRASEKARENDRA SARASWATI VIDYAPURAM,

PLOT-E, SECTOR-5,

NERUL, NAVI-MUMBAI-400706

SUBJECT – ADVANCED DEEP LEARNING

PROJECT REPORT

IN

Image Captioning with Pretrained Model

SUBMITTED BY

Melwyn Titus John

MCS.23.11

Under the Esteemed Guidance

of

Dr. Rajeshri Shinkar

MCS.COMPUTER SCIENCE PART-2

SEMESTER 4 (2024-2025)

SIES (NERUL) COLLEGE OF ARTS, SCIENCE AND COMMERCE(AUTONOMOUS)

NAAC RE-ACCREDITED 'A' GRADE

SRI CHANDRASEKARENDRA SARASWATI VIDYAPURAM,

PLOT-E, SECTOR-5,

NERUL, NAVI-MUMBAI-400706

CERTIFICATE

This is to certify that the project entitled **Image Captioning with Pre-trained Model** is successfully completed by **Melwyn Titus John** of Part-II (Sem-4) Masters in Science (Computer Science) as per the requirement. It is also to certify that this is the original work of the candidate done during the academic year 2024-2025.

Roll no - MCS.23.11 **Date of submission: 23-03-2025**

Subject – Advanced Deep Learning

Dr. Rajeshri Shinkar

(project guide**)**

INDEX

Abstract

Deep Learning has advanced the field of Computer Vision by enabling models to generate textual descriptions for images, a process known as *Image Captioning*.

This project implements an image captioning system using a **pre-trained BLIP (Bootstrapping Language-Image Pre-training) model**. The model takes an image as input and generates a descriptive caption by leveraging **transformers and vision-language models**.

The **Salesforce/blip-image-captioning-base** model is used to generate captions for images. The implementation involves loading an image, passing it through a pre-trained model, and obtaining a caption without requiring explicit feature extraction or annotation. This approach can be applied in **assistive technology, automated image description, and content tagging systems**.

Keywords: Deep Learning, Image Captioning, BLIP Model, Transformer-based Vision Models, Pre-trained Model

Introduction

Image Captioning is a fundamental task in **computer vision and natural language processing (NLP)**, where a system generates a meaningful textual description of an image. It has applications in **assistive technologies for visually impaired individuals, autonomous systems, and content-based image retrieval**.

Traditional image captioning relied on **CNN-RNN architectures**, where Convolutional Neural Networks (CNNs) extracted features, and Recurrent Neural Networks (RNNs) generated text. With the advancement of **Transformer-based models**, vision-language models like **BLIP (Bootstrapping Language-Image Pre-training)** offer a powerful way to generate captions efficiently.

This project focuses on implementing **BLIP-based image captioning** using a **pre-trained model** without requiring extensive training on custom datasets.

Objectives

1. To implement an **image captioning model** using a pre-trained deep learning model.
2. To utilize the **BLIP model** for generating captions from images.
3. To process image inputs and obtain meaningful descriptions.
4. To evaluate the performance of the generated captions.
5. To demonstrate the effectiveness of pre-trained vision-language models.

Literature Review

Image Captioning has evolved from early **template-based methods** to deep learning techniques involving **CNNs and RNNs**. Early models used **encoder-decoder architectures**, where CNNs encoded image features and RNNs (such as LSTMs) generated captions. However, these methods suffered from **context loss and lack of generalization**.

Recent advancements in **Vision Transformers (ViTs) and multimodal learning** have improved captioning accuracy. The **BLIP (Bootstrapping Language-Image Pretraining) model**, developed by Salesforce, uses a vision transformer to understand images and a language model to generate captions. It outperforms traditional CNN-RNN architectures and offers better generalization across datasets.

Pretrained models like **BLIP, CLIP, and ViLT** have set new benchmarks in **zero-shot learning for vision-language tasks**, making them ideal for **image captioning** without domain-specific training.

Key Research Works:

- Vaswani et al. (2017): Introduced the **Transformer model** for NLP.
- Radford et al. (2021): Developed **CLIP**, a vision-language model.
- Li et al. (2022): Proposed **BLIP**, a model for bootstrapping vision-language learning.

Methodology

The development of the Custom NER Model follows these steps:

Pretrained Model: `Salesforce/blip-image-captioning-base`

Libraries Used:

- torch: For deep learning operations
- transformers: For loading the BLIP model
- PIL (Pillow): For image processing

1. **Load the BLIP model and processor**

   from transformers import BlipProcessor, BlipForConditionalGeneration

   import torch

   from PIL import Image

2. **Load an image and preprocess it**

   image_path = "sample/sample.jpg" # Replace with your image path
   image = Image.open(image_path).convert("RGB")
   processor = BlipProcessor.from_pretrained("Salesforce/blip-image-captioning-base", use_fast=True)

3. **Generate the caption using the model**

   model = BlipForConditionalGeneration.from_pretrained("Salesforce/blip-image-captioning-base")

   inputs = processor(image, return_tensors="pt")

   with torch.no_grad():

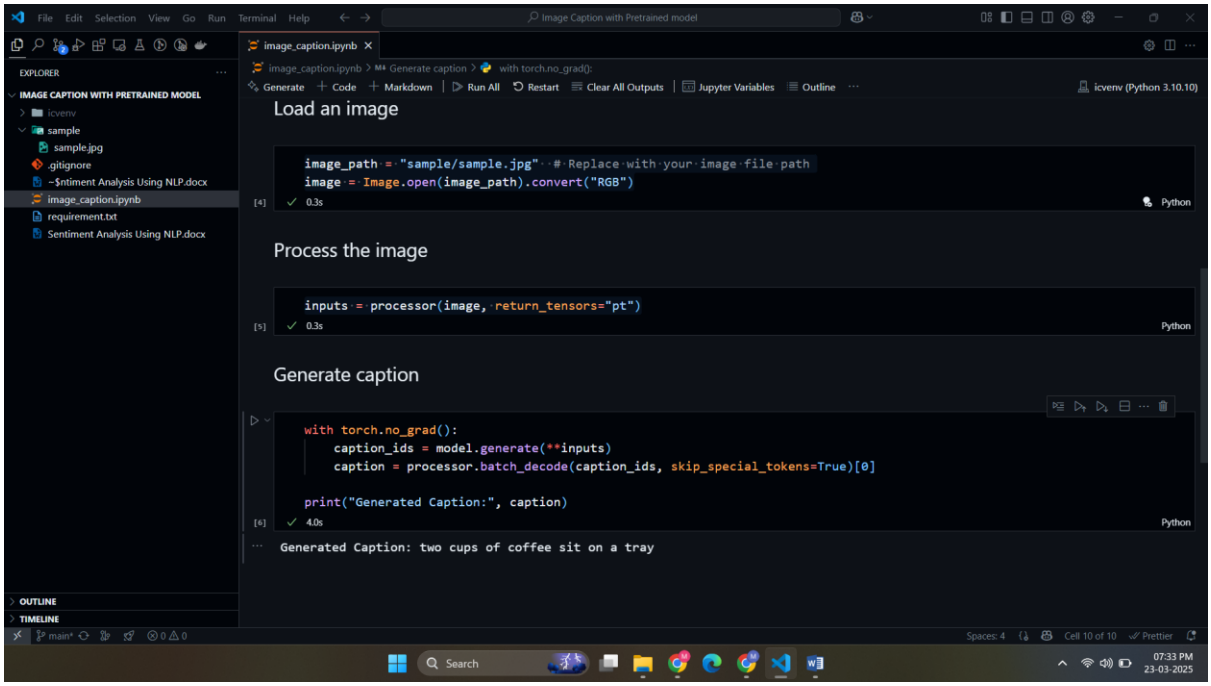
```
    caption_ids = model.generate(**inputs)
caption = processor.batch_decode(caption_ids,
skip_special_tokens=True)[0]
print("Generated Caption:", caption)    return text    for ent in
annotations["entities"]:
        ner.add_label(ent[2])
```

Implementation

The flow of execution follows:

1. Load the **pre-trained BLIP model**.
2. Read and process an image.
3. Convert the image into a tensor format for input to the model.
4. Generate a **text caption** using the model.
5. Display the generated caption.

Results



Conclusion

The **BLIP-based image captioning system** efficiently generates textual descriptions for images using a **pretrained vision-language model**. This approach eliminates the need for extensive training on custom datasets and provides **high-quality captions** using **transformer-based deep learning models**.

Future improvements include:

- Fine-tuning on **domain-specific datasets** for better contextual captions.
- Implementing **interactive applications** for real-time image captioning.
- Exploring **other multimodal learning models** like **CLIP and GPT-4 Vision**.

References

Vaswani, A. et al. (2017). Attention Is All You Need. Advances in Neural Information Processing Systems.

Radford, A. et al. (2021). Learning Transferable Visual Models From Natural Language Supervision. arXiv preprint arXiv:2103.00020.

Li, J. et al. (2022). BLIP: Bootstrapped Language-Image Pretraining for Unified Vision-Language Understanding. arXiv preprint arXiv:2201.12086.