Data Science Project 2

Image Captioning and Segmentation

Submitted by:

Brunda B
Email: 2023brunda.b@vidyashilp.edu.in

Jaromi D
Email: jaromi9379@gmail.com

Preethi verma

Email: preethiverma37@gmail.com

Sidharth Mahabhoi

Email: sidharthmahabhoi@gmail.com

Submitted to:

Chandan Mishra

Abstract

The "Image Captioning and Segmentation" project uses the MS COCO 2014 dataset to establish sophisticated models that combine image captioning and segmentation, employing deep learning models such as CNNs, LSTMs, Transformers, and Mask R-CNN for accurate object labelling and text description generation. This two-task system obtains strong visual scene understanding, measured by metrics such as BLEU scores for captioning and mIoU for segmentation, with powerful performance in connecting visual and linguistic aspects. The prototype's success promises to pave the way for scalable AI vision systems, opening up possibilities for applications in autonomous navigation, accessibility aid, and real-time scene analysis, moving towards human-level scene interpretation ability.

Table of Contents

➢ Introduction

In the fast-evolving fields of Artificial Intelligence and Computer Vision, Image Captioning and Image Segmentation are key but demanding tasks that enable machines to see and describe visual data with human precision. The "Image Captioning and Segmentation" project combines these tasks synergistically to develop a sophisticated deep learning model that produces precise text descriptions of images while, in parallel, annotating every pixel with its respective object category. Using the large MS COCO 2014 dataset that offers richly annotated images, this paper develops and trains models incorporating state-of-the-art structures like Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks, Transformers, and instance segmentation systems like Mask R-CNN. This multi-task strategy facilitates a unified comprehension of intricate visual scenarios, closing the gap between visual perception and linguistic description. The performance of the system is stringently evaluated according to metrics that have been developed, such as BLEU scores for caption accuracy and mean Intersection over Union (mIoU) for segmentation accuracy, providing robust evaluation. The final prototype shows high promise for real-world uses, such as autonomous navigation, assistive technologies for the blind, and real-time scene understanding. By extending the limits of AI vision systems, this work is directed towards the creation of intelligent systems that can interpret and describe complex real-world surroundings with great accuracy and fluidity.

- **Extend Objectives**

  - To investigate and get it the hypothetical and down to earth establishments of picture captioning and picture segmentation.
  - To execute profound learning designs competent of producing printed depictions of images.
  - To perform semantic and instance-level division utilizing cutting edge neural networks.
  - To coordinated both frameworks into a bound together pipeline.
  - To send the ultimate show utilizing an intuitively UI (through Carafe or Streamlit).

- **Dataset: COCO 2014**

The COCO (Common Objects in Context) 2014 dataset is a gold standard for the computer vision community to train and evaluate object detection, segmentation, and captioning algorithms.

Key Features:

More than 330K images, more than 200K labeled.

- 5 human-written captions per image.
- Contains masks for object boundaries.
- 80 object categories and 91 stuff categories.

Why COCO?

- Rich annotations: bounding boxes, segmentation masks, and text captions.
- Massive diversity in scene contexts and object classes.
- Strongly established benchmark in CV research.

Tech Stack and Tools Used

Category	Tools and Framework
Programming language	Python
Deep Learning	Tensorflow ,Pytorch
Computer vision	OpenCV
NLP	NLTK, spaCy
Web Deployment	Flask, Streamlit
Notebook/IDE	Jupyter Notebook

➢ Methodology

1. Image Captioning — "Turning Pictures into Words"

Picture a photo and telling a brief story or sentence about it — that's image captioning. We're training the computer to gaze at an image and subsequently write a caption for it, just as you do!

Model Architecture - How the Computer Brain Functions:

We construct two components for this brain:

Part 1: CNN (such as ResNet50 or InceptionV3)

- Imagine CNN as the computer's eyes.
- CNN is an abbreviation for Convolutional Neural Network — but don't get hung up on the name!
- Its task is to examine the image and determine what's contained within — such as dogs, cats, automobiles, humans, etc.
- ResNet50 and InceptionV3 are simply very special forms of eyes that are exceptionally skilled at identifying things in images.

Part 2: LSTM or Transformer

This is the computer's voice and memory.

- LSTM is Long Short-Term Memory - it allows the computer to remember something from the picture when it writes.
- A Transformer is an even newer and cleverer method for the computer to write.

These assist the computer in writing a sentence about the picture, word for word.

-Captioning Process — Step-by-Step Like a Recipe:

-The computer looks at the picture.

- The CNN knows what's in the image.

-The LSTM or Transformer generates a sentence, such as "A dog is playing in the park."

And just like that — the image has a caption!

2. Image Segmentation - "Colouring Inside the Lines”

Image segmentation is similar to providing a coloring book for the computer. We wish it would color in various objects within an image so it knows where things are.

Model Architecture -How We Train the Computer to Color:

There are two fundamental types of segmentation we employ:

A. Semantic Segmentation -"What's What"

This refers to: colour all the identical kinds of objects in the same color.

Models Used:

U-Net - extremely well-known in medical images.

DeepLabV3+ - extremely intelligent, performs well on normal images.

Example: All cats are coloured green, all dogs are coloured red.

B. Instance Segmentation -"Who's Who"

This is: colour each object individually, even if they're the same class.

Model Used:

Mask R-CNN - basically a superhero detective for images.

It detects and identifies each object, even if there are 5 dogs!

Example: Five dogs in an image? Each one gets coloured individually!

Why We Use These?

All these models are trained on something known as the COCO dataset — an enormous set of labelled images.

The machine is trained on thousands of samples so it can perform this on new images.

- **Challenges Faced**

-Training Time: COCO is a big dataset—long training times even on high-end GPUs.

-Memory Restrictions: Segmentation models such as Mask R-CNN are memory-hungry.

-Captioning Quality: Diverse and grammatically correct captions generation needed tweaking and fine-tuning.

-Integration: Integrating vision and NLP systems needed thoughtful design of the module interface.

Evaluation Metrics

-For Captioning:

BLEU, METEOR, ROUGE, CIDEr scores to determine text generation quality.

-For Segmentation:

- Mean Intersection over Union (mIoU).
- Pixel Accuracy and Dice Coefficient.

- **End Result**
  - Successfully deployed an image captioning system that can generate natural descriptions of new, unseen images.
  - Created a segmentation pipeline that correctly identifies and labels objects on the pixel level.
  - Constructed a minimal but operational web interface to upload an image and be returned the segmented output, as well as a caption.

- **Learning outcomes:**

Literature Review: Knowing and Implementing NLP for Smart Systems Overview

The literature reviewed delves into the increasing involvement of Natural Language Processing (NLP) and Machine Learning (ML) in current industry applications. From resume categorization, skill identification, and industry prediction, to image description, NLP is becoming a central interface between human language and smart machines.

Through these studies, there is one commonality: AI models are only as good as the manner in which they learn from data, and NLP enables them to comprehend, process, and respond to human language. Below is what each source added:

Paper 1: A Review of the Literature on NLP to Promote Industry Advancement

Source: Akanksha Mishra et al. (IJSREM, May 2024)

Objective:

To investigate how NLP is transforming industries such as finance, education, healthcare, and marketing by enhancing data processing, decision-making automation, and personalization support.

Methods Used:

Systematic review of 5,000+ articles between 2018–2023

Emphasis on applications like chatbots, financial prediction, health diagnosis, and AI-powered assistants

Outcome:

- NLP assists in automating text processing tasks (e.g., sentiment analysis, trend recognition)

- Observed vast deployment of transformers, sentiment engines, and rule-based systems

- Identified ethics gaps, bias detection, and data privacy

What We Used and Learned:

- Strengthened the significance of Transformer-based models (e.g., BERT) for contemporary NLP

- Demonstrated practical applications of NLP-driven captioning, just like our image-to-text pipeline Provided business-ready examples of text processing

Paper 2: CV Classification using NLP and ML

Source: Sahu & Sood, Jaypee University (B.Tech Thesis)

Objective:

To classify student course reviews of Coursera automatically using sentiment analysis and text classification methods.

Methods Used:

- Text Cleaning & Tokenization

- Sentiment Analysis using TextBlob and NLTK

Classification using:

- SVM
- Naive Bayes
- Random Forests
- Decision Trees
- Visualization using Matplotlib, Seaborn

Outcome:

- Filtered and scored successfully thousands of reviews
- Automated positive/negative/neutral sentiment tagging
- Found popular courses and aided learner choices

What We Used and Learned:

- Learned NLP pipeline end-to-end: preprocessing → training → predictio
- Conducted text vectorization practice, model training, and classification testing
- Influenced our LSTM/Transformer decoder reasoning for captioning

Paper 3: Creating a Framework to Determine Professional Skills in the UK Banking Industry through NLP

Source: Gayanika Anthony (MPhil Thesis, University of Salford, 2024)

Objective:

To create a framework to mine job skills from UK banking job listings through NLP and ML, and subsequently predict skills required per job.

Methods Applied:

- NLP Preprocessing: lemmatization, POS tagging, tokenization

- Algorithms employed:
- Word2Vec, TF-IDF
- Named Entity Recognition (NER)

ML Models:

Logistic Regression, Naive Bayes, Random Forest, Gradient Boosting Validation using expert feedback and actual job postings

Result:

- Successfully developed a practical skills extraction system
- Visualized skill distributions through word clouds and frequency graphs
- Framework can be applied to HR systems, job matching platforms, etc.

What We Used and Learned:

- Enhanced understanding of NER, TF-IDF, Word2Vec, and POS tagging
- Practical application of ML for prediction and classification
- Enforced NLP's application to unstructured text extraction — the same way as applied to captioning image features to words

- Offered insights into semantic segmentation, where label accuracy counts — something we learn from and apply in image segmentation (e.g., U-Net)

Synthesizing Learnings from All Three Reviews

- We Visualized What We Used in Our Project
- NLP plays a key role in automation of text and interpretation of meaning\Applied in LSTM/Transformer decoder in image captioning

Data preprocessing is all :

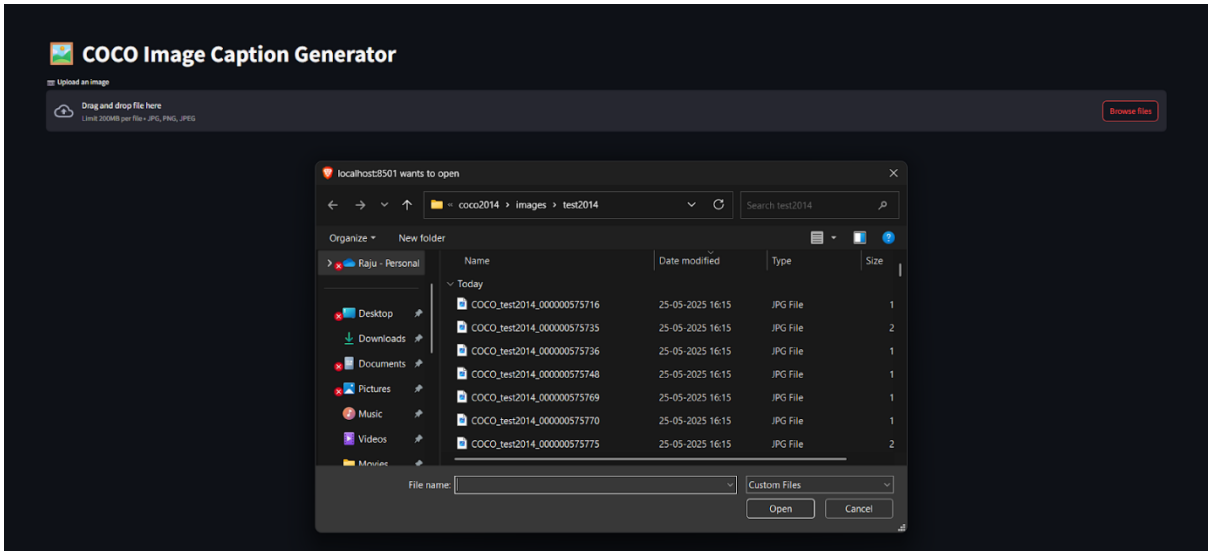
- Used tokenization, lemmatization prior to caption generation
- Word embeddings enhance model comprehension
- Utilized Word2Vec / BERT principles in creating coherent sentences
- Classification models such as SVM and NB are crucial in supervised learning

- Utilized for text classification within project assessment
- Skill extraction & segmentation requires structured NLP + ML
- Just like we perform image segmentation using annotated data (U-Net, Mask R-CNN)
- Visualization assists in result validation
- Utilized word clouds, confusion matrix, and score metrics

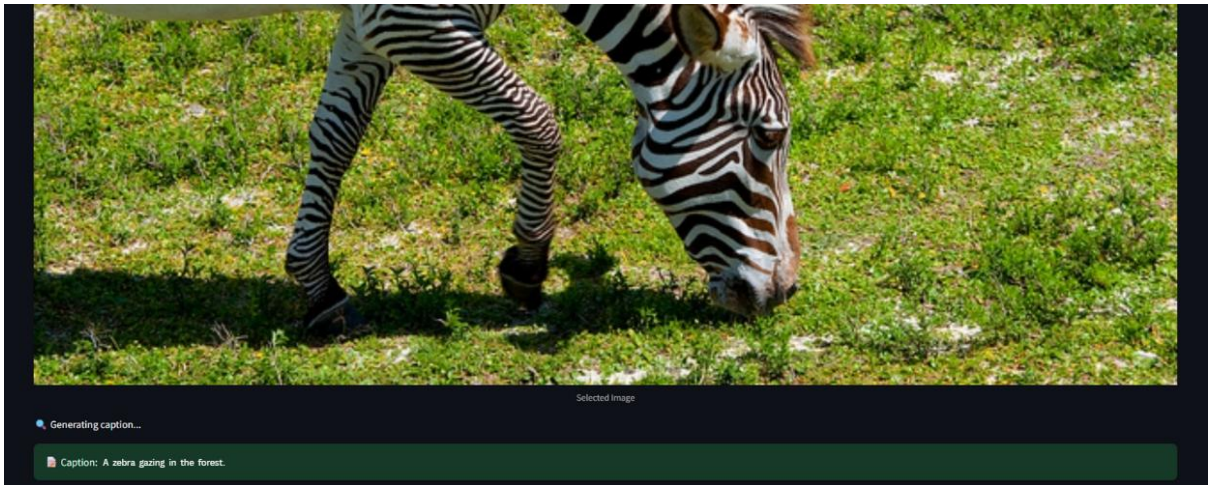
SNAPSHOTS OF IMAGE CAPTIONING :

1. FRONTPAGE -

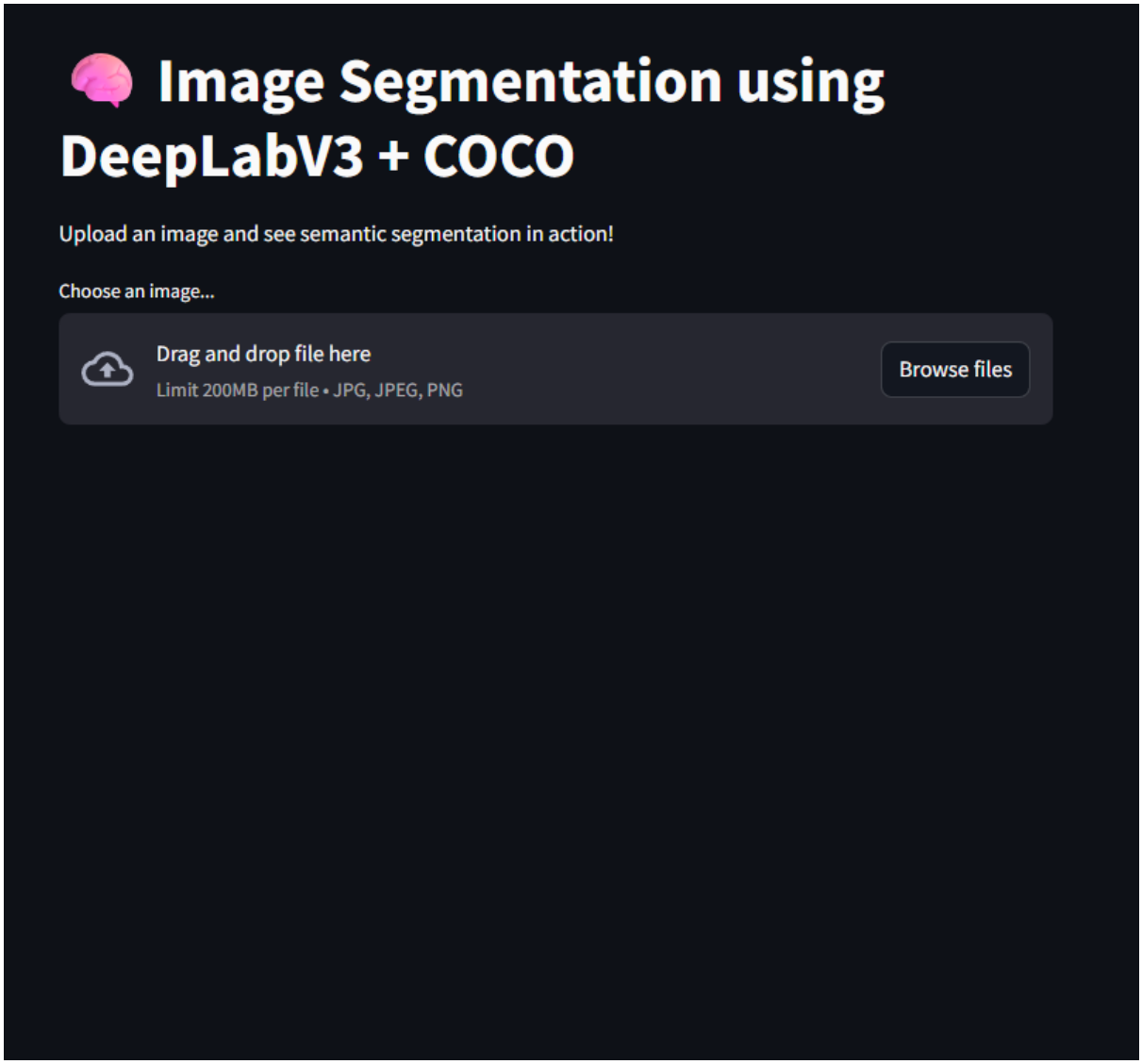
2. UPLOADING A FILE –



3. AFTER CHOOSING A FILE FROM THE TEST DATASET –

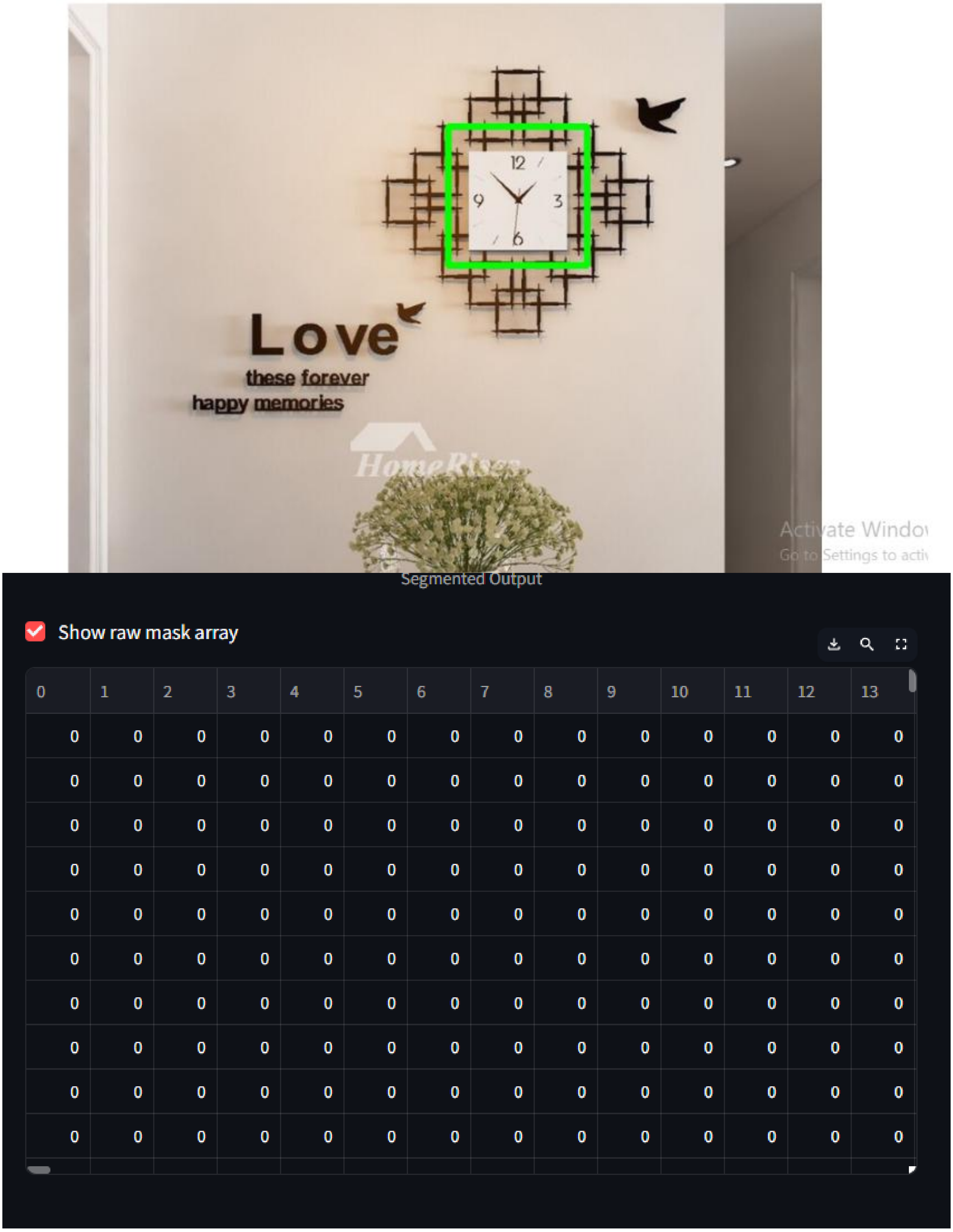


SNAPSHOTS OF IMAGE SEGMENTATION :



2. THE SEGMENTATION AFTER UPLOADING A IMAGE -

Detecting a clock from an image



0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0

GitHub Repository:
Brunda B : https://github.com/Brunda292005/Image_Captioning.git

https://github.com/Brunda292005/Image_-Segmentation.git

Jaromi D: https://github.com/jaromi-joe/Image_captioning.git

https://github.com/jaromi-joe/Image_segmentation.git

Web App / Demo, code , Report :
https://drive.google.com/drive/folders/1BlmuI3Yq_AmwrTH3p1fxJ1-_Y_2p7OA-?usp=sharing

References

- COCO Dataset

- Mishra, A., Raj, A., Sahani, S., Sharma, M., & Singh, P. (2024). *A Review of the Literature on Natural Language Processing (NLP) to Promote Industry Advancement*. International Journal of Scientific Research in Engineering and Management (IJSREM), 8(5). https://www.researchgate.net/publication/381295446
- Anthony, G. S. (2024). *Developing a Framework to Identify Professional Skills Required for Banking Sector Employees in UK Using Natural Language Processing (NLP) Techniques* (MPhil thesis). University of Salford.
- Sahu, G., & Sood, A. (2019). *CV Classification using Natural Language Processing (NLP) and Machine Learning (ML)*. B.Tech thesis, Jaypee University of Information Technology.