Object Retrieval using Segment Anything Model (SAM) and CLIP

Hasaan Maqsood

December 5, 2024

Course Information

- **University:** Skoltech University
- **Course:** MA030702: Transformers in Computer Vision
- **Instructor:** Sergey Zagoruyko, `S.Zagoruyko@skoltech.ru`
- **Teaching Assistants:**
  - Mohamad Al Mdfaa, `mohamad.almdfaa@skoltech.ru`
  - Ramil Khafizov, `Ramil.Khafizov@skoltech.ru`

Contents

1 Introduction

This project investigates the integration of the **Segment Anything Model (SAM)** with **CLIP** to perform object retrieval based on textual prompts. The objective is to accurately segment objects within images and identify masks corresponding to a specified object type (e.g., ”chair”) by leveraging both segmentation capabilities and semantic understanding.

2 Methodology

2.1 Data Preparation

- **Dataset Acquisition**: The dataset was downloaded from a Google Drive link and extracted into the `./Dataset` directory.
- **Data Structure**:
  - **Images**: Located in `./Dataset/images_0a7cc/`.
  - **Ground Truth Masks**:
    - `.jpg` masks in `./Dataset/gt_semantic_2d/render_semantic/`.
    - `.npy` masks in `./Dataset/gt_semantic_2d/render_semantic_npy/`.

2.2 Model Initialization

- **Segment Anything Model (SAM)**:
  - **Checkpoint**: `sam_vit_h_4b8939.pth`.
  - **Model Type**: `vit_h`.
  - **Functionality**: Generates potential masks for objects within an image.
- **CLIP Model**:
  - **Architecture**: `ViT-B/32`.
  - **Functionality**: Encodes textual prompts and image segments into a shared embedding space to compute semantic similarities.

2.3 Processing Pipeline

1. **Mask Generation**: For each image, SAM's automatic mask generator produces potential masks for objects present.
2. **Mask Filtering**: Each generated mask is cropped and preprocessed. CLIP encodes the cropped segments and the textual prompt. Cosine similarity determines the relevance of each mask to the specified object type.
3. **Mask Combination**: Masks exceeding a similarity threshold are combined to form the final predicted mask.

4. **Evaluation**: **Intersection over Union (IoU)** is computed between predicted masks and ground truth masks to assess accuracy.
5. **Visualization**: Results, including original images, ground truth masks, predicted masks, and overlays, are visualized for analysis.

3 Observations

3.1 Learning Outcomes

- **Model Synergy**: Combining SAM and CLIP effectively leverages segmentation and semantic understanding for precise object retrieval.
- **IoU as a Metric**: IoU provided a clear quantitative measure to evaluate mask accuracy.

3.2 Results Assessment

- **Mean IoU**: Achieved a mean IoU of **0.6824** for the class "chair", indicating a reasonable overlap between predicted and ground truth masks.

3.3 Potential Improvements

- **Threshold Optimization**: Fine-tuning the `similarity_threshold` can enhance the balance between precision and recall.
- **Multi-Class Handling**: Extending the approach to handle multiple object classes would increase system versatility.
- **Advanced Preprocessing**: Implementing techniques to enhance object contrast or reduce background clutter may improve mask accuracy.

4 Visualization of Results

4.1 IoU Calculation

The **Intersection over Union (IoU)** metric was used to evaluate the overlap between predicted and ground truth masks.

- **Mean IoU**: `0.6824`

4.2 Mask Visualization

For each processed image, the following were visualized:

1. **Original Image**
2. **Ground Truth Mask**
3. **Predicted Mask**
4. **Overlay of Predicted Mask on Original Image**

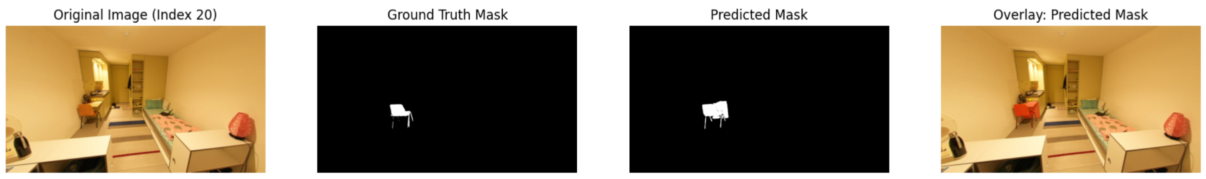

Figure 1

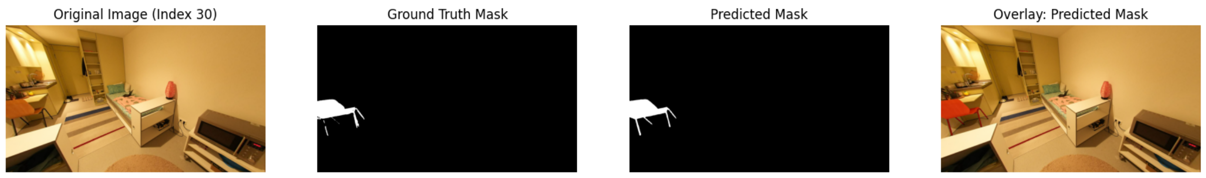

Figure 2

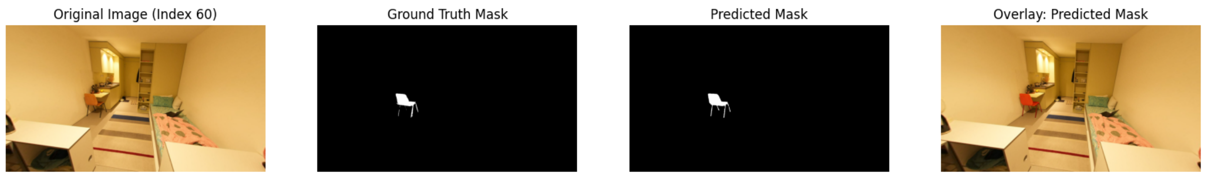

Figure 3

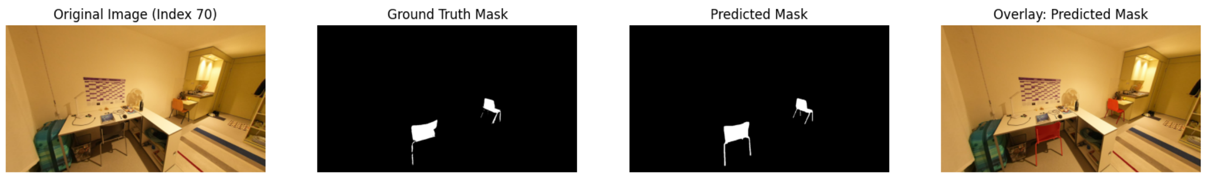

Figure 4

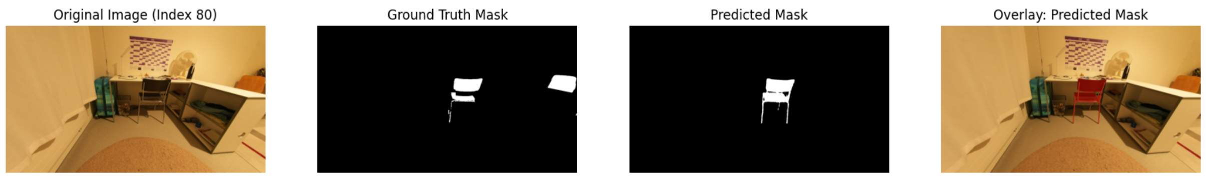

Figure 5

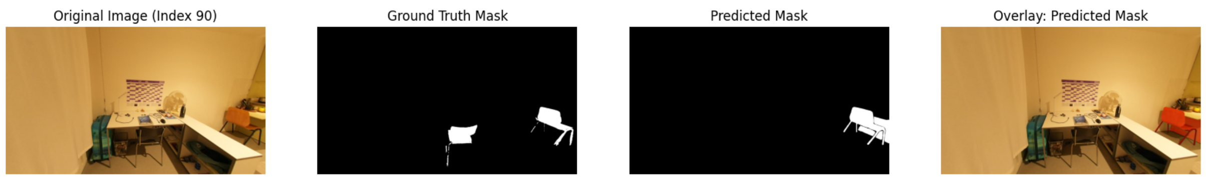

Figure 6

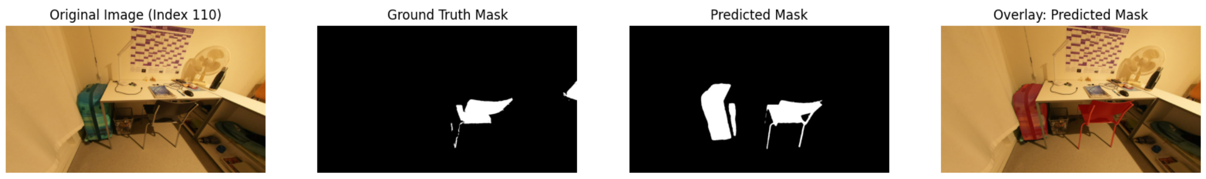

Figure 7

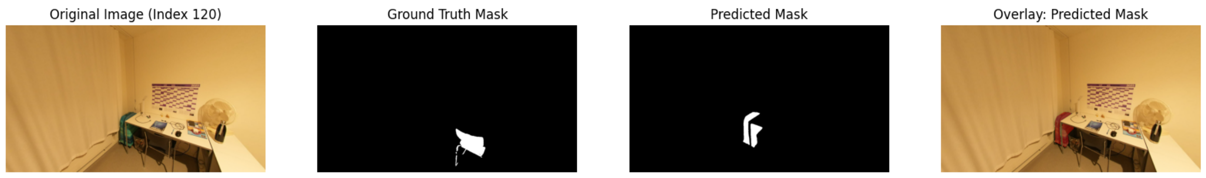

Figure 8

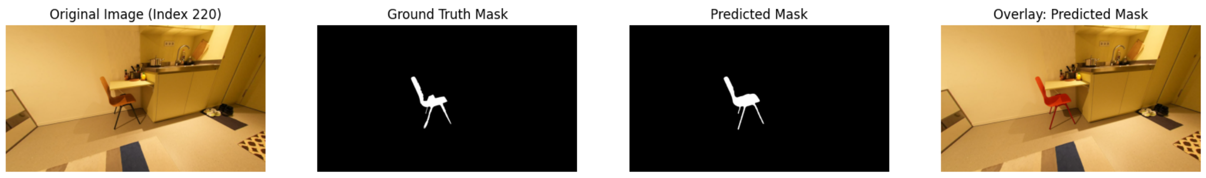

Figure 9

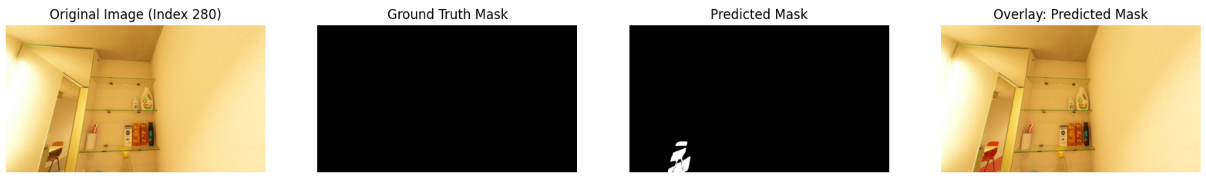

Figure 10

4.3 Failure Cases

Lower IoU scores were observed in cases involving:

- **Complex Backgrounds**: Difficulty in isolating objects from cluttered scenes.
- **Small Object Sizes**: Limited detail leading to inaccurate mask predictions.
- **Ambiguous Contexts**: Objects in unconventional settings or orientations causing misalignment.

5 Running the Solution

To replicate the results, follow the steps below within an IPython Notebook.

5.1 1. Clone Repositories

```
# Clone the Segment Anything repository
!git clone https://github.com/facebookresearch/segment-anything.git

# Clone the CLIP repository
!git clone https://github.com/openai/CLIP.git
```

Listing 1: Clone SAM and CLIP Repositories

5.2 2. Install Required Libraries

```
# Install Segment Anything
!pip install git+https://github.com/facebookresearch/segment-anything.
    git

# Install CLIP
!pip install --upgrade git+https://github.com/openai/CLIP.git

# Install other dependencies
!pip install torch torchvision torchaudio --index-url https://download.
    pytorch.org/whl/cu118
!pip install scikit-learn matplotlib tqdm
```

Listing 2: Install Dependencies

5.3 3. Download the ViT-H Checkpoint

```
import os
import requests
from tqdm.notebook import tqdm

# URL for the vit_h checkpoint
checkpoint_url = "https://dl.fbaipublicfiles.com/segment_anything/
    sam_vit_h_4b8939.pth"

# Directory to save the checkpoint
checkpoint_dir = "./checkpoints"
os.makedirs(checkpoint_dir, exist_ok=True)

# File path for the checkpoint
checkpoint_path = os.path.join(checkpoint_dir, "sam_vit_h_4b8939.pth")

# Function to download the checkpoint with a progress bar
def download_checkpoint(url, path):
    if not os.path.exists(path):
        print("Downloading vit_h checkpoint...")

        # Stream the file with a progress bar
        response = requests.get(url, stream=True)
        total_size = int(response.headers.get('content-length', 0))  # Total file size in bytes
        block_size = 1024 * 1024  # Block size for tqdm (1 MB)

        with open(path, "wb") as f, tqdm(
            desc="Downloading",
            total=total_size,
            unit="B",
            unit_scale=True,
            unit_divisor=1024,
        ) as bar:
            for data in response.iter_content(block_size):
                f.write(data)
                bar.update(len(data))

        print(f"Checkpoint downloaded and saved to {path}")
    else:
        print(f"Checkpoint already exists at {path}")

# Download the checkpoint
download_checkpoint(checkpoint_url, checkpoint_path)
```

Listing 3: Download SAM Checkpoint

5.4 4. Execute the Processing Pipeline

Run the cells containing data loading, model initialization, processing, evaluation, and visualization as outlined in the **Methodology** and **Visualization of Results** sections above.

6 Conclusion

The integration of the **Segment Anything Model (SAM)** with **CLIP** has demonstrated a promising approach for object retrieval based on textual prompts. While the models effectively segmented and identified objects like **chair** with a reasonable IoU, challenges remain in handling complex backgrounds, small object sizes, and ambiguous contexts. Future work will focus on optimizing threshold settings, expanding to multiple object classes, and enhancing preprocessing techniques to further improve accuracy and robustness.

7 Future Work

- **Threshold Optimization**: Implement adaptive thresholding to balance precision and recall dynamically.
- **Multi-Class Handling**: Extend the approach to support multiple object classes simultaneously.
- **Advanced Preprocessing**: Develop preprocessing techniques to enhance object contrast and reduce background clutter.

- **Interactive Visualization**: Incorporate interactive tools for dynamic exploration of mask predictions.

8 Deliverables

1. **Running the Solution**: The solution is structured within an IPython Notebook, divided into clear sections for easy execution. Follow the steps in the **Running the Solution** section to clone repositories, install dependencies, download checkpoints, and execute the processing pipeline.

2. **Observations and Research**: The report above details observations from model performance, including successful mask generation and areas where models struggled. It also explores how factors such as prompt specificity, mask size, and background complexity impact performance.

9 Additional Resources

- **Segment Anything GitHub Repository**: `https://github.com/facebookresearch/segment-anything`
- **CLIP GitHub Repository**: `https://github.com/openai/CLIP`
- **TQDM Documentation**: `https://tqdm.github.io/`
- **PyTorch Documentation**: `https://pytorch.org/docs/stable/index.html`

10 Contact

Name:	Hasaan Maqsood
Degree:	Master in Data Science
Institution:	Skolkovo Institute of Science and Technology
Email:	Hasaan.Maqsood@skoltech.ru
Phone:	+7 993 624-92-56