Global Emancipation Network: GeoCatch
End-Quarter Report

Jamie Luna: jluna28@calpoly.edu
Rachel Roggenkemper: rroggenk@calpoly.edu
Matteo Shafer: mshafe01@calpoly.edu
Client: Global Emancipation Network (Sherrie Caltagirone)
Instructor: Dr. Hunter Glanz & Dr. Jonathan Ventura
DATA 452: Data Science Capstone II, Spring 2024
12 June 2024

Abstract

The Global Emancipation Network (GEN), led by Sherrie Caltagirone, combats human trafficking using Meta AI's Segment Anything Model (SAM) combined with StreetCLIP for image analysis and location prediction. While effective, the models lack explainability, posing challenges for practical applications. Our approach addresses this issue by providing textual explainability through Zero Shot Instance Segmentation and confidence metrics when serving predictions. By integrating features like customizable location lists, input prompts, and matching segments, we improved the model's accuracy and transparency. This enhances GEN's mission and empowers stakeholders with insights to aid in rescuing victims.

I. Introduction

Human trafficking, affecting approximately 21 million victims annually and generating nearly $50 billion for traffickers, remains a global crisis with fewer than 50,000 victims rescued each year. At the forefront of combating this issue is the Global Emancipation Network (GEN), led by Founder and Executive Director Sherrie Caltagirone, which serves as a premier data hub for trafficking information. GEN leverages an innovative model that utilizes Meta AI's Segment Anything Model (SAM) combined with StreetCLIP to analyze images and predict their locations, which is crucial in identifying and rescuing victims. However, this model currently provides predictions without any explanation, presenting a significant challenge. Our project aims to introduce a layer of explainability to this process, offering clear reasoning behind the model's location predictions. By doing so, we enhance GEN's mission to dismantle trafficking networks, empowering stakeholders—including law enforcement, government agencies, academia, and non-profits—with the information needed to make informed decisions in their efforts to save lives.

II. Background

Human traffickers are increasingly leveraging internet-based applications to conduct their operations, thereby exploiting technological advances to further their reach and complicate efforts to combat this global crisis. In response, the Global Emancipation Network (GEN) plays a pivotal role by facilitating the aggregation and analysis of extensive data related to human trafficking. This data, sourced through collaborations with various experts, spans the open web,

deep web, and dark web, encompassing a vast array of information with particular emphasis on photographic content.

These photographs, while rich in potential leads, often lack explicit information regarding the location of depicted scenes or individuals, which is critical for operational action. To bridge this gap, GEN has developed and implemented advanced modeling techniques, specifically utilizing Meta AI's Segment Anything Model (SAM) and StreetCLIP. This innovative combination segments images into smaller, analyzable components to ascertain the most probable locations associated with each segment. The challenge, however, lies in the models' current output which gives location predictions without underlying explanations.

Addressing this challenge, our project's central goal is to enhance this model by integrating explainability. Thus, not only pinpointing where an image might have been taken but also providing the rationale behind these predictions. This advancement is crucial for enabling various stakeholders, including law enforcement and non-profit organizations, to understand and act upon the insights generated by our model effectively. Through this project, we support GEN's overarching mission to dismantle trafficking networks and facilitate the rescue of victims, enhancing the efficacy and transparency of anti-trafficking efforts worldwide.

III. Methodology

Our initial exploration into adding explainability of location predictions in images began with two applets given to us by Sherrie.

Applet 1: Streamlit

The first applet, Streamlit, allowed for the specification of a list of locations and the use of a customizable prompt to inquire about these locations. Its output displays the likelihood of each specified location, showing the top 10 countries. The model was trained on Airbnb and hotel.com photos from around the world to ensure accuracy. However, this applet doesn't do object segmentation and analyzes the image as a whole. The work done with Streamlit was a result of a few hackathons so it included files for a few different projects.

Applet 2: Sesame

The second applet, Sesame, developed using Meta AI's Segment Anything Model (SAM), presented a more refined approach to location prediction. This applet segments the image into multiple segments, then uses SAM and runs StreetClip on each segment to see which objects contributed to the location decision. Therefore, the model outputs predictions for the image as a whole and predictions for each image segment. This model was also trained on Airbnb and hotel.com photos from around the world like Streamlit. However, unlike Streamlit, Sesame lacked a customizable locations list and customizable prompt. However, the detailed segmentation highlighted Sesame's superior analytical capabilities and its potential for deeper explainability, particularly in identifying specific image features that correlate with geographic locations.

Decision to Build off Sesame Applet:

After exploring both applets, we decided to build off the Sesame user interface. We were convinced by Sesame's inclusion of object segmentation. Additionally, we consulted Fred Lichtenstein from Camera Forensics, since he worked on adding explainability to this process in the past. He confirmed Sesame had been the main approach. Although we chose Sesame because of its object segmentation, we opted to integrate Streamlit's key functionalities—namely, the list of possible locations and the customizable prompt—into the Sesame model. This integration was aimed at using Sesame's analytical strengths while incorporating the interactivity, flexibility, and user-driven inquiry mode offered by Streamlit, thus allowing for more targeted location predictions. We believed this would allow us to create the most accurate model and lead to the greatest explainability.

Moving Forward:

In our pursuit to add explainability to the Sesame model's location predictions, we embraced the model's use of StreetClip combined with SAM, a pivotal machine-learning tool for object detection within images. This strategic choice was informed by a comprehensive evaluation of the available models.

To enhance Sesame's functionality and relevance to our project, we wanted to introduce the capabilities that Streamlit had, such as, manually inputting a list of predicted locations, offering significant flexibility in defining geographic boundaries for analysis. This feature, coupled with the modification of the input prompt, would allow us to create a more interactive dimension to the model. These enhancements are not only technical adjustments but strategic improvements designed to optimize the model's accuracy in predicting locations. By fine-tuning these parameters, we believed we could achieve more precise predictions.

We also wanted to add object recognition to the model and be able to detect objects within a specific image. We believed adding all these different features into our model that we chose to improve Sesame, would allow us to achieve the most accurate predictions and have explainability as to why the image was from the specified location.

Our methodology has also involved extensive research and hands-on experimentation with the applet's code, giving us the opportunity to implement the features mentioned above and navigate the challenges of adapting the code to meet our project's needs. This approach has deepened our understanding of the model's architecture and functionality, allowing us to refine the model for our specific needs.

In summary, our methodology is defined by a blend of strategic model selection, feature development, extensive research, and practical experimentation, guided by the initial insights provided by the Streamlit and Sesame applets and the goals of the Global Emancipation Network.

IV. Results

Our engagement with the Sesame applet has led to significant advancements in our final goal to add explainability to the model. As shown in Figure 1, our updated applets output the top 5 location predictions in addition to a labeled version of the input image.

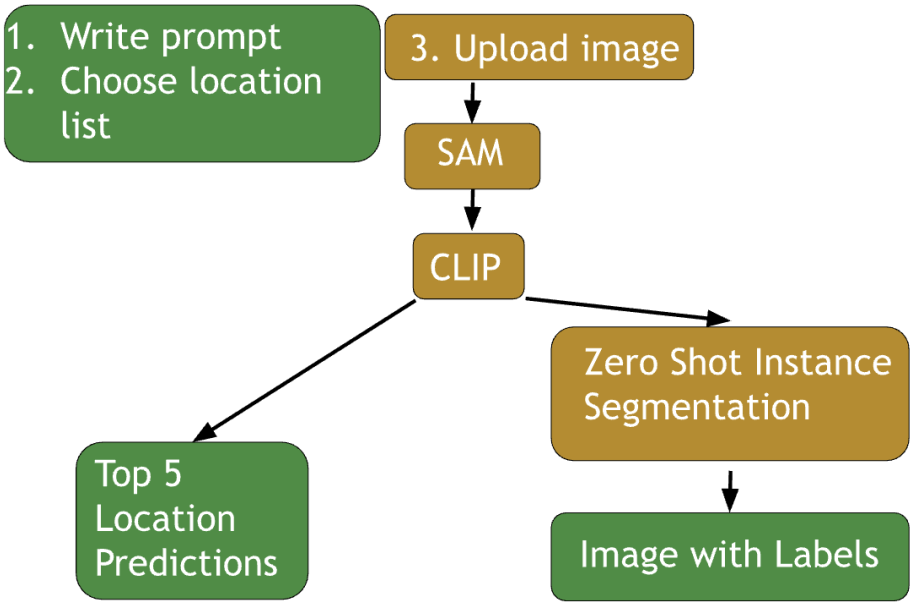

Figure 1: System Flowchart for Our Applet

Lessons Learned

We tried many things throughout the past two quarters. In particular, we learned two lessons about features we tried to implement because we thought they would enhance the explainability of the predictions, but they didn't result as expected.

First, a key area of exploration involved the application of image filters, such as adjustments to contrast, exposure, and saturation, to investigate their impact on the model's object recognition and location prediction accuracy. We thought this approach would offer an indirect method to enhance the model's performance without necessitating alterations to its foundational architecture. However, through testing many images with different types of filters and seeing how the predictions of location and object recognition would change, we noticed the opposite happened. The filters altered the image to where the predictions of location and object recognition were no longer as accurate.

Secondly, we were able to add the customizable prompt feature to the applet like the Streamlit applet had, seen below in Figure 2. However, we learned that the location list overrides the customizable prompt. For example, if the prompt asks for a country but the location list only includes cities, since the model is limited to choose locations just from the location list, it will only be able to output cities even though the prompt is asking for countries. Currently the prompt does not have an affect on the output, although it was left in the applet as future work.

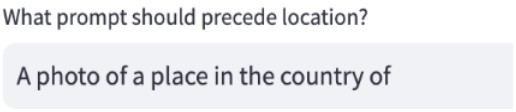

Figure 2: Addition of Customizable Prompt Feature

We learned valuable lessons through the ideas that didn't work out as we expected. However, we completed tasks that did enhance the explainability of the model predictions.

Customizable Location List

As mentioned above, we built off the Sesame applet and chose to implement features from the Streamlit applet. We added the capability to customize the location list and offered several default location lists, seen below in Figure 3. These location lists are vital because the model chooses only from these locations for the predictions of the image. We started off by adding the countries list which was the default from the Streamlit applet. We then tested it using the fifty U.S. states, and to our surprise, with the locations becoming more specific, we actually saw that location predictions got more accurate. Since we saw improved predictions going from countries to states, we tested going even more specific and moving towards cities. At first, we tested the major cities around the world, where we saw the continuation of accurate predictions. However, our most updated list just includes cities from the 9 countries that GEN has active projects in. Through testing, we saw that when the location list has multiple cities from the same area, although the top predictions might not always include the exact city, the top predictions will always include cities in the nearby vicinity. These city-level predictions are more useful than the state-level predictions because although not entirely accurate, it gives a much more specific location to the overall area of where the photo is taken. Additionally, these location lists are only a starting point. The user is still able to edit the locations, including deleting any of them or adding any new ones.

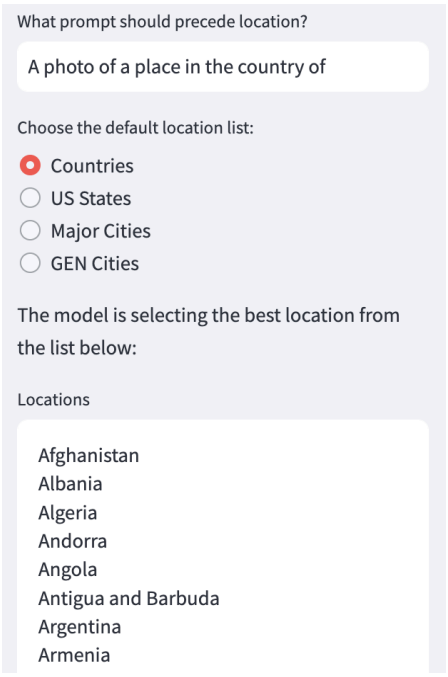

Figure 3: Addition of Customizable Location List.

One of the primary results observed from the implementation of these features is a substantial increase in the specificity and reliability of location predictions. By enabling the specification of potential locations, the model can now operate within a more defined geographic scope, drastically reducing the ambiguity of its predictions. This specificity is crucial for pinpointing the probable locations of trafficking incidents, thereby improving the potential for intervention and support.

Furthermore, the ability to customize input prompts has opened new avenues for extracting meaningful insights from the model's predictions. This feature allows users to guide the model's focus more precisely, adding a layer of interactivity to the model, leading to outputs that are not only more relevant but also accompanied by a greater degree of explainability. The results from these enhancements have been encapsulated in the updated functionalities of the Sesame applet. The applet now has an enriched user experience, offering insights into the model's prediction process that were previously unclear.

Output with Probabilities
At the start, our model outputted one prediction for the entire image and five predictions for each segment. Now, we output five predictions for the entire image along with the corresponding probabilities, and five predictions for each segment along with the corresponding probabilities. This updated output can be seen below in Figure 4.

This allowed us to have more insight into the confidence of each prediction. Now the user can see how far apart each prediction is and know where to narrow their geographic scope.

Top Predictions (with Corresponding Probability):

1. Hawaii: 98.97%
2. Florida: 0.28%
3. California: 0.12%
4. Oregon: 0.11%
5. Arizona: 0.10%

91 segments found.

Figure 4: Output with probabilities

Matching Segments
To begin, there would always be zero matching segments in the final location prediction. Initially, we thought we would have to find the influential segments manually, by identifying

segments that had a top 5 location prediction match with one of the top 5 location predictions of the entire image as an influential and thus “matching” segment. However, we discovered that now with the updated location lists and thus more accurate predictions, the model is automatically now outputting matching segments.

This feature is vital because it tells us the influential segments in the image that are contributing most to the final location prediction. This also adds to the explainability of the model, because if the contents in the segment are influential to the final location prediction, that means knowing what is in the specific segment can help us understand why the model is choosing a specific location.

Object Recognition

After we identified the influential segments, our next ambition revolved around achieving a textual level of explainability. To this end, we employed Zero Shot Instance Segmentation. This approach leverages both SAM and CLIP again, to identify key objects within images. We are able to annotate images with precise outlines of the identified objects, coupled with a textual list of labels found in the image. In this way, the user can visually see what objects were important in creating a location prediction while also knowing the labels the model provides those objects.

A notable challenge in Zero Shot Instance Segmentation arose from the model's reliance on a predefined list of objects for detection and classification. To address this, we worked with Sherrie to develop a comprehensive list of key objects. This list guides the model in accurately recognizing and labeling the elements deemed significant in our analysis.

An example of an application of Zero Shot Instance Segmentation can be seen in Figure 5. The image is annotated with outlines of identified objects and there is a list of labels returned corresponding to those objects.

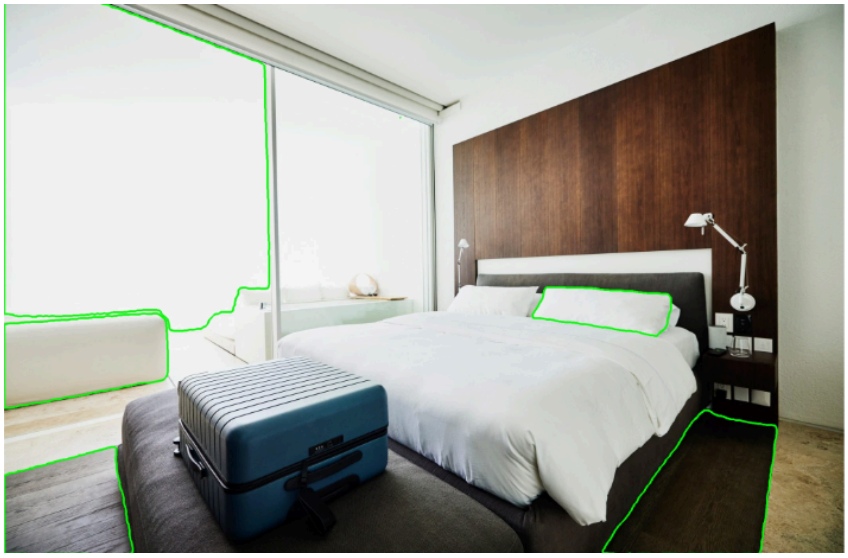

Figure 5: Object Recognition

Overall, the refinements made to the Sesame model highlight the importance of adaptability and user-centric design in developing software solutions for complex social issues. Our progress highlights the potential of AI in aiding the fight against human trafficking and the critical role of iterative improvement in maximizing the impact of such technologies. As we continue to evolve the model's capabilities, our results offer a promising glimpse into the future of data-driven interventions in global human rights efforts.

V. Discussion

Our journey in enhancing the explainability of location predictions through the use of Meta's Segment Anything Model (SAM) and CLIP documentation has been both challenging and enlightening. The SAM model, as highlighted in an article by Akruti Acharya on Encord[1], was developed to tackle the scarcity of segmentation masks on the Internet. This led to the creation of the SA-1B dataset, a groundbreaking compilation of over 1.1 billion masks, representing the largest labeled segmentation dataset available today. This extensive diversity of segments, ranging from natural scenes and urban environments to medical imagery and satellite images, is instrumental in the model's ability to recognize and categorize a wide array of objects and scenes.

The utilization of SAM, coupled with the innovative capabilities of CLIP, supports our project's methodology. CLIP's sophisticated approach to understanding images in the context of natural language descriptions complements SAM's segmentation process, making the combination a powerful tool. This synergy allows for the interpretation of images, enhancing our model's predictive accuracy and the richness of its outputs.

However, one of the primary difficulties we've encountered stems from the nature of SAM's predetermined segmentation. While the segmentation facilitated by SAM is impressively comprehensive, it does limit our ability to manually select or highlight specific image segments that we believe are critical for enhancing explainability. The ideal scenario would allow us to box entire objects—such as an outlet or a street sign—to directly associate the presence of these objects with the predicted location, based on their native habitats or common occurrences. This level of specificity would significantly bolster our model's explainability, offering clear, tangible reasons for its geographic predictions.

Despite these challenges, our engagement with the SAM model remains a cornerstone of our project. Its advanced capabilities and the breadth of the SA-1B dataset provide a solid foundation for our efforts. Through discussions with Sherrie and Fred, we've identified promising strategies to navigate the constraints imposed by predetermined segments. These conversations have sparked innovative ideas for future enhancements, reinforcing our commitment to evolving our model's functionality.

Moreover, our comparative analysis of SAM and the StreetClip model through extensive documentation review and empirical testing with a variety of images has reinforced our conviction in SAM's superiority. Its advanced analytical capabilities, coupled with the added depth from CLIP's contextual understanding, offer a compelling advantage over alternative models.

---

[1] https://encord.com/blog/segment-anything-model-explained/

Our project's exploration of the "black box" of location prediction using SAM and CLIP is a testament to the potential of combining advanced AI technologies to address complex challenges. Despite the hurdles presented by predetermined segmentation, our ongoing efforts to adapt our approach highlight the dynamic nature of data science work. We remain excited about the possibilities that SAM and CLIP present, not just for enhancing the explainability of our model, but also for contributing to the broader mission of leveraging technology in the fight against human trafficking.

VI. Conclusion

As we conclude the final phase of our project, we've enhanced the Sesame model's accuracy and explainability in location prediction. Our focus was on integrating features that allow for using varied location specificity and the ability to label objects within the image. These enhancements, designed to refine the model's precision and user interactivity, have been incorporated into the applet. The model now outputs the top five predicted locations for each segment with the confidence for each prediction and a list of labels identified within the image.

VII. Future Work

Although we were able to accomplish all our goals these past two quarters, in becoming familiar with the code and processes, we do recognize that there is always more work to be done. Thus, here are two main aspects that could be done if this project were to continue in the future.

The first would be to integrate the object recognition abilities into the applet. Currently, the object recognition capabilities are housed in a separate file. To streamline the user experience, we recommend integrating this file into the prebuilt Sesame applet. This would give the user one space to get location predictions and also a list of objects in the image they are trying to locate.

Another task is to match objects with the location predictions. Currently, we are separately providing location predictions and a list of objects that are in the given image. However, we know that there is likely a connection between the location predicted and the objects that are in the image, and this would also add to the explainability of the model. So, we think trying to match objects that coincide with the location prediction would also be a vital step toward adding explainability.