Social Media Image Detection: Classification of Military Social Media Videos for Geographical Location Approximation

Author: CDT Christian Preti

Advisors: LTC Bruce Chojnacki and Dr. Meyer Levy

Introduction

The United States has seen growing threats in the information security domain. The distribution and analysis of Publicly Available Information (PAI) on social media networks (such as TikTok) has allowed adversaries to create profiles on service members. This data is "used for marketing" and "can be used to target an individual in their home, commit identity fraud, harassment, or target groups as part of an influence campaign" [4]. Additionally, the information is cheap to acquire; a study completed through Duke University (and funded by the United States Military Academy) found that "sensitive data about active-duty members of the military" was found "for as low as $0.12 per record" [12]. The advancements in Large Language Models (LLMs) and Vision-Language Models (VLMs) have created the opportunity for adversaries to compile data from social media platforms and determine approximate unit locations. This severely impacts Operational Security, mostly due to social media, which has "increased the speed and range of information, diffused power over information, and shifted socio-cultural norms" [7]. Overall, the need to understand this spread of information—and possibly use it against adversaries—has been realized across the globe and has become a genuine concern for nation states.

Literature Review

The development of LLMs and VLMs has allowed monumental advancements in data collection and categorization to assist researchers in many projects. These machines, many of which are open-source and accessible by the public, include Llama 2 (developed by Meta) and Florence-2 (developed by Microsoft). Both have significantly impacted data categorization and collection, and (when combined with other models) have had success in these areas. However, they also provide an opportunity to be used in conjunction with the latest advancements in computer vision, such as GeoCLIP (developed by researchers at the University of Central Florida).

Llama 2 (which stands for "Large Language Model Meta AI") is similar to other LLMs in that it "works by taking a sequence of words as an input and predicts a next word to" generate text recursively [9]. These weighted predictions are what allow LLMs to be so effective. This process of establishing relationships and probability is the basis for neural networks, which "are machine learning models that mimic the complex functions of the human brain" because they "consist of interconnected nodes or neurons that process data, learn patterns, and enable tasks such as pattern recognition and decision-making" [13]. Llama 2 was used by members at Siemens Healthineers to "linearly represent anatomical landmarks" (which is "a biologically

meaningful point on an organism that aids in image navigation" and which is used in "radiology reporting") "in space with considerable robustness to different prompts" [1]. However, this modeling is not perfect. The main threats of LLMs are that they include "bias, toxic comments, and hallucinations" (being false information) that are difficult to analyze by researchers [9].

Florence-2 is a primary example of a VLM. Designed with the purpose to "take text-prompt as task instructions and generate desirable results in text forms," it can be used in "captioning, object detection, [and] grounding or segmentation" of images as a result of the "5.4 billion comprehensive visual annotations on 126 million images" [14]. The success of Florence-2 lies in its ability to "zero-shot:" this is a type of "machine learning…in which an AI model is trained to recognize and categorize objects or concepts without having seen any examples of those categories or concepts beforehand" [2]. An example of this was a study completed to "extract Geometric Dimensioning and Tolerancing information from 2D engineering drawings" by incorporating a "fine-tune[d] Florence-2," which was able to effectively complete the "extraction" and highlighting possible future "support [for] downstream manufacturing tasks" [8].

The prevalence of online information, artificial intelligence, and learning machines have only increased in both popularity and capabilities. PIGEON—"a deep learning model" that can "geolocate Google Street View images"—is a recent creation/usage of artificial intelligence to classify geolocation from an image [6]. PIGEON used "geocells" to classify images, which were able "to discretize the Earth's surface into a set number of classes" [5]. Similarly, PLANET AI has an "Intelligent Document Analysis—IDA—software suite [that] offers comprehensive capabilities…for short time-to-value automation and high-quality data capture, extraction and understanding" [10]. Companies have started to lean into data collection and have used AI models to do so.

However, many of these learning machines rely on extensive training sets. A novel advancement in artificial intelligence is GeoCLIP, an "Image-to-GPS retrieval" model that predicts the approximate GPS coordinates of a photo [3]. Normal "image-based geo-localization" models "focus on identifying the GPS locations of images within a specified region of interest" or try to "perform image-to-image retrieval" to match images directly to identify their location [3]. However, this is limited in that it requires an extensive data set of images along with the fact that locations outside of popular regions (i.e. outside cities and into rural areas) would not be as accurately documented. GeoCLIP uses a novel approach by "retriev[ing] the GPS coordinates of an unseen query image and matching it with the gallery of GPS coordinates" by utilizing "location and image encoders to project GPS and image in a shared embedding space" [3]. This allows the neural networks within the model to predict the location of images in any region across the globe. Additionally, this method allows "a much finer evaluation than previous techniques, resulting in better localization performance" and overall image geo-localization [3].

Research Methodology

The data provided for this study consisted of approximately 80 GB of TikTok videos provided by ViralMoment, "a for-profit company that specializes in analytics related to social media" [11]. This project built on previous work completed by Leonardo Sabetta; the script can successfully "extract every 30th frame from every video provided," in which "Microsoft's Florence2-large model"—which has been extensively pretrained—creates a "text output of the objects detected" by using its zero-shot capabilities [11]. This script then utilizes "Facebook's Bart-large-MNLI" model in order "to classify the text label into 1 of 5 categories: Aircraft, Ground Vehicles, Maritime, Weapons, and Nonmilitary" [11]. If the majority of frames had been classified as being military-related, the video was then classified as being taken either indoors or outdoors.

```
#checking keywords
#print("[DEBUG] Starting indoor/outdoor classification...")
all_keywords = set(indoor_weights.keys()).union(outdoor_weights.keys())
indoor_score = 0
outdoor_score = 0
outdoor_frames = 0
indoor_frames = 0
for keyword in all_keywords:
    matches = text_lower.count(keyword.replace(" ", ""))
    indoor_score += matches * indoor_weights.get(keyword, 0)
    outdoor_score += matches * outdoor_weights.get(keyword, 0)
    if matches > 0:
        if keyword in outdoor_weights:
            outdoor_frames += matches
        elif keyword in indoor_weights:
            indoor_frames += matches



BIAS_MARGIN = 4 #margin for encouraging geoclip to run...

#decision; MORE FORGIVING for military and outdoors; if scores are close, will favor "outdoors"
if indoor_score > outdoor_score+BIAS_MARGIN:
    io_class = "indoors"
elif outdoor_score > indoor_score+BIAS_MARGIN:
    io_class = "outdoors"
else:
    io_class = "outdoors" #if the difference is within 2, will be outdoors since already military

#print(f"[DEBUG] Scores - Indoor: {indoor_score}, Outdoor: {outdoor_score}")

#adding to frame classification

frame_classifications.append(f"-> Indoor/Outdoor: {io_class} (I:{indoor_score}, O:{outdoor_score})")
```

Figure 1 – Classification Based on Indoor/Outdoor Frames

This process was completed by taking the raw initial output of the Florence2-large model and splicing up the output into individual phrases, which were in turn referenced against two separate dictionaries (one pertaining to objects that would be found indoors like a bed or ceiling, and another pertaining to objects that would be found outdoors like a fence or tree). If most of the objects (within a margin of four) pertained to being "outside," the overall video would be classified as outside.

If the video had followed the pipeline to be classified as outside, five frames were randomly selected to be passed to the GeoCLIP model.

```
if io_class == "outdoors":

    #print("[DEBUG] Starting GeoClip Process.")

    try:
        from geoclip import GeoCLIP
        geo_model = GeoCLIP()

        #getting 5 rand frames for geoclip
        sample_indices = random.sample(processed_frame_indices, k=min(5,len(processed_frame_indices)))

        gps_coords = []
        all_top_pred_gps = []
        all_top_pred_probs = []

        for i, target_frame in enumerate(sample_indices): #frames in rand sample

            if target_frame >= total_frames or target_frame < 0:
                #print(f"[DEBUG] Skipping target_frame={target_frame}, out of bounds")
                continue
            frame = processed_frames.get(target_frame)
            if frame is None:
                #print(f"[DEBUG] No saved frame for {target_frame}")
                continue
```

Figure 2 – Pipeline to GeoCLIP Implementation

Each frame, in turn, would be returned with an estimated GPS location and a confidence probability value. The highest probability value would be stored along with the GPS coordinates in a file for later analysis.

```
            geo_image_path = os.path.join(output_video_dir, f"geo_clip_frame{target_frame}.png")
            Image.fromarray(cv2.cvtColor(frame, cv2.COLOR_BGR2RGB)).save(geo_image_path)
            top_pred_gps, top_pred_prob = geo_model.predict(geo_image_path, top_k = 5)
            all_top_pred_gps.extend(top_pred_gps)
            all_top_pred_probs.extend(top_pred_prob)

        #picking the best prediction
        if all_top_pred_gps and all_top_pred_probs:

            #find which one is the highest probability!!!
            best_index = all_top_pred_probs.index(max(all_top_pred_probs))
            best_coord = all_top_pred_gps[best_index]
            best_lat, best_lon = best_coord
            best_prob = all_top_pred_probs[best_index]


            with open(output_text_path, 'a') as f:
                f.write("\nGeoCLIP Top 5 Predictions:\n")
                f.write(f"Best Prediction (highest probability): ({best_lat:.6f},{best_lon:.6f})\n")
                for i in range(min(5, len(all_top_pred_gps))):
                    lat, lon = all_top_pred_gps[i]
                    prob = all_top_pred_probs[i]
                    f.write(f"Prediction {i+1}: ({lat:.6f}, {lon:.6f})\n")
                    f.write(f"Probability: {prob:.6f}\n")
            #Logging the BEST one and other info to location.txt file!!!
                with open(LOCATION_FILE, 'a') as f:
                    location_string = f"({best_lat:.6f}, {best_lon:.6f})"
                    f.write(f"{obj.object_name},outdoors,{location_string} BestProb={best_prob:.6f}, IF's={indoor_frames},OF's=
{outdoor_frames}\n")

                #print("[DEBUG] Location and frame counts successfully logged.")
        else:
            logging.warning("Failed to extract frame for GeoCLIP.")
```

Figure 3 – Determination of GPS Coordinates

Over fifteen thousand videos were processed over approximately four days and thirteen hours. Of the processed videos, over three thousand were classified as outdoors and thus ran through the GeoCLIP model to determine the approximate location of the video.

Analysis

Several trends emerged from the analysis of the location data. An initial point of emphasis is that the time to process videos through GeoCLIP takes significantly longer than normal classification of a nonmilitary video. This makes sense as—once a video is classified as being military-related—the video must process through another model entirely (being the GeoCLIP model).

Classification	Count of Files (Videos)	Average of Seconds
Military	3217	39.0976
Nonmilitary	12599	21.2242
Grand Total	15816	24.8599

Table 1. Aggregate Data of Videos

As seen within **Table 2**, most videos that GeoCLIP processed produced a confidence of 6-8% accuracy. Although this is not extraordinarily high, there are many that produced predictions of over 20% (which is remarkable considering the images are taken from frames within a TikTok video, in which the surrounding area is not the point of emphasis).

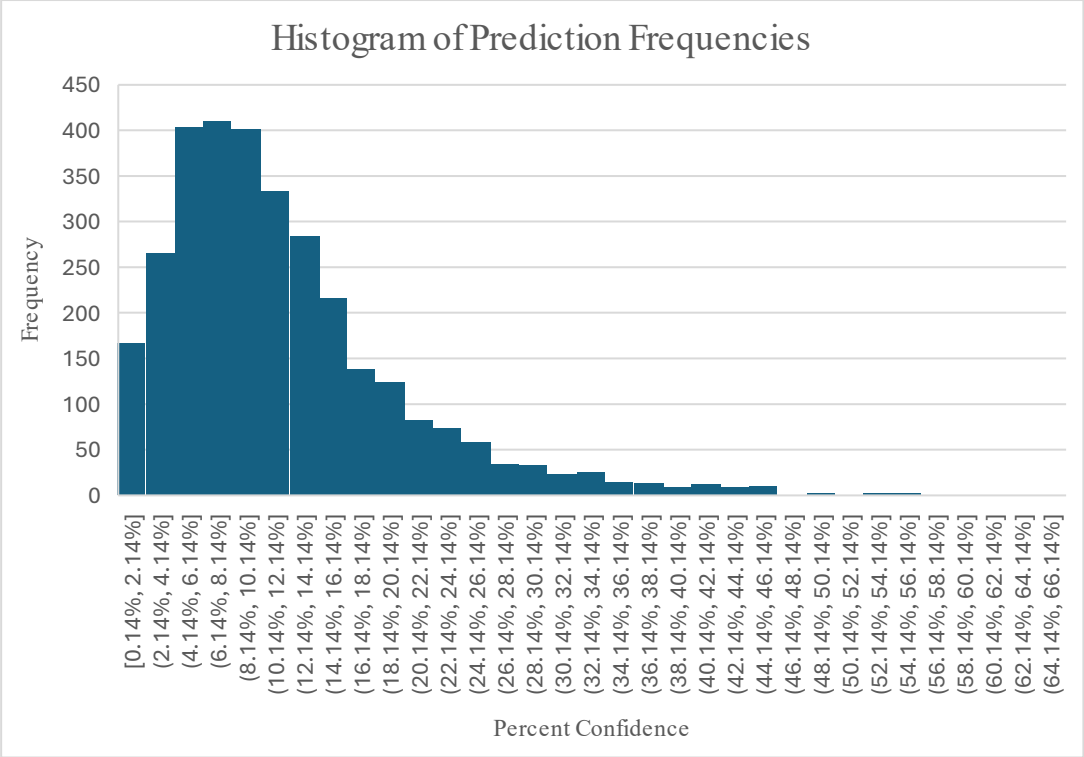

Table 2. Prediction Frequencies

Additionally, by aggregating the data and plotting the coordinates on a world map, **Figure 4** is created (by using Google's *My Maps*). This shows that most of the videos were located within the United States and Europe (with several areas supporting groups of videos with probabilities of over 20%).

Figure 4 – Geographical Prediction

Discussion

Although highly promising, this data is challenging to analyze for several reasons. Without knowing geographical locations, confirming the predictions offers a major limitation to this process. Additionally, these models had been processed on an RTX A6000 1; if more (and stronger) GPUs had been used to analyze these videos, the videos would be processed in a fraction of the time. The dictionaries used to classify the images as indoors/outdoors could also be advanced further; they currently contain a limited number of elements, but a broader list could ensure more accurate classification.

However, assuming that the predictions are as accurate as GeoCLIP's probabilities presume, these results highlight artificial intelligence's impact on OPSEC—especially within social media platforms. These videos were processed in the background over several days, meaning that data can be continuously scraped from the internet, passed through object detection models, and input into models like GeoCLIP, which could approximate the location of the video. This means that—if troop units are not careful with what they post online—locations of movements can easily be determined by adversaries.

REFERENCES

[1] Abdi, Mohamad, Gerardo Hermosillo Valadez, and Halid Ziya Yerebakan. *Automatic Mapping of Anatomical Landmarks from Free-Text Using Large Language Models: Insights from Llama-2*. Cornell University, 2024. https://arxiv.org/pdf/2410.12686. Accessed 25 February 2025.

[2] Bergmann, Dave. "What is zero-shot learning?" *IBM*. 2024. https://www.ibm.com/think/topics/zero-shot-learning. Accessed 25 February 2025.

[3] Cepeda, Vicente Vivanco et al. "GeoCLIP: Clip-Inspired Alignment between Locations and Images for Effective Worldwide Geo-localization." *arXiv*, 2023. https://arxiv.org/abs/2309.16020. Accessed 7 May 2025.

[4] "Death by a Thousand Cuts: Commercial Data Risks in the Army." (n.d.). *The Army Cyber Institute*. United States Military Academy, West Point, NY. Accessed 7 February 2025.

[5] Haas, Lukas et al. *PIGEON: Predicting Image Geolocations*. Cornell University, 2024. https://arxiv.org/pdf/2307.05845. Accessed 25 February 2025.

[6] Haas, Lukas et al. "PIGEON: Predicting Image Geolocations." *GitHub Pages*, Stanford University, 2024. https://lukashaas.github.io/PIGEON-CVPR24/. Accessed 7 February 2025.

[7] *Information Environment—Opportunities and Threats to DOD's National Security Mission*. U.S. Government Accountability Office. https://www.gao.gov/assets/730/722922.pdf. Accessed 7 February 2025.

[8] Khan, Muhammad Tayyab et al. *Fine-Tuning Vision-Language Model for Automated Engineering Drawing Information Extraction*. Cornell University, 2024. https://arxiv.org/pdf/2411.03707. Accessed 25 February 2025.

[9] "Introducing LLaMA: A foundational, 65-billion-parameter large language model." *Meta*. Meta, 2023. https://ai.meta.com/blog/large-language-model-llama-meta-ai/. Accessed 25 February 2025.

[10] "Planet AI: The Power of Intelligence." (n.d.) Planet, PLANET AI GmbH, 2025. https://planet-ai.com/. Accessed 25 February 2025.

[11] Sabetta, Leonardo. "Machine Learning Object Detection for Open-Source Intelligence: Using Zero-Shot Object Detection & Natural Language Processing Text Classifiers to Identify Military Equipment." West Point, 2024.

[12] Sherman, Justin et al. *Data Brokers and the Sale of Data on U.S. Military Personnel*. November 2023.

[13] "What is a Neural Network?" *GeeksforGeeks*. 2024. https://www.geeksforgeeks.org/neural-networks-a-beginners-guide/. Accessed on 25 February 2025.

[14] Xiao, Bin et al. *Florence-2: Advancing a Unified Representation for a Variety of Vision Tasks*. Cornell University, 2023. https://arxiv.org/pdf/2311.06242. Accessed 25 February 2025.