AI2CYBER - GeoSpy

Phase 1: System Design & Model Selection

Introduction

Design and implement a prototype AI system that predicts the geographical region of an image.

System Design

System requirements

- Ingest images from publicly accessible URLs.
- Determine the geographical location (continent or country level) based on visual features.
- Handle cases where visual features are insufficient for accurate location identification.
- Evaluate performance on common and corner cases.
- Run inference on resource-constrained environments (no GPU).

Methodology

The most efficient way to prototype a system for the given objective is to use Deep Neural Networks and approach it as a classification task, since that the requested output labels (continents/countries) are readily definable. However, training and evaluating even the SOTA visual classifiers (EfficientNet, ViT, etc.) would require both computational power and processing time: we would have to build a huge visual database that would contain a large and diverse set of images, as well as train and evaluate computationally expensive models within a limited timeframe.

Therefore, we can leverage a visual model with zero-shot capabilities enabling inference on labels that were not explicitly used as targets during training. Such a model is OpenAI's CLIP (Radford et al., 2021) that is trained on a variety of (image, text) pairs and can be instructed in natural language to reference learned visual concepts (or describe new ones) facilitating zero-shot transfer of the model to downstream tasks. In this case, the task is image classification against continent/countries labels, also referenced as geo-location in Figure 21 of the CLIP paper.

The AI pipeline will be developed in a Jupyter Notebook for rapid development and debugging, and uploaded to GColab for ease of reproducibility and sharing. It will be based on the CLIP implementation from Hugging Face (openai/clip-vit-large-patch14) and, utilizing Pandas and Google StreetView, implement an inference function to process all test samples.

Model Selection

For zero-shot geolocation and rapid prototyping, CLIP is the superior choice. Its ability to understand the relationship between images and text allows for quick experimentation without extensive training data. We can describe locations textually, and CLIP will identify matching images, facilitating rapid development and testing of geolocation concepts. However, evaluating CLIP's performance can require custom metrics, and fine-tuning may be more involved than with ViT (Dosovitskiy et al., 2020).

For a mature geolocation product with robust evaluation capabilities, ViT is preferable. ViT excels at capturing detailed visual features and can be fine-tuned on labeled geolocation datasets for higher accuracy. Established evaluation metrics and benchmarks are readily available for ViT, simplifying performance assessment. While ViT requires more data to train from scratch, it's better suited for a production-ready solution where rigorous evaluation is paramount.

However, full training ViT would be more efficient and cost-effective than CLIP, as CLIP requires simultaneous finetuning of its text embedding branch based on image-text pairs rather than image-class pairs, which are easier to obtain or generate. Since CLIP uses ViT as its backbone for visual feature extraction, in the case of full retraining, it offers no significant scalability advantage.

Ultimately, the model selection for this assignment prioritized rapid prototyping. This will raise questions concerning the model's predictive confidence and its generalization capabilities, which will be addressed in the Phase 3 evaluation report.