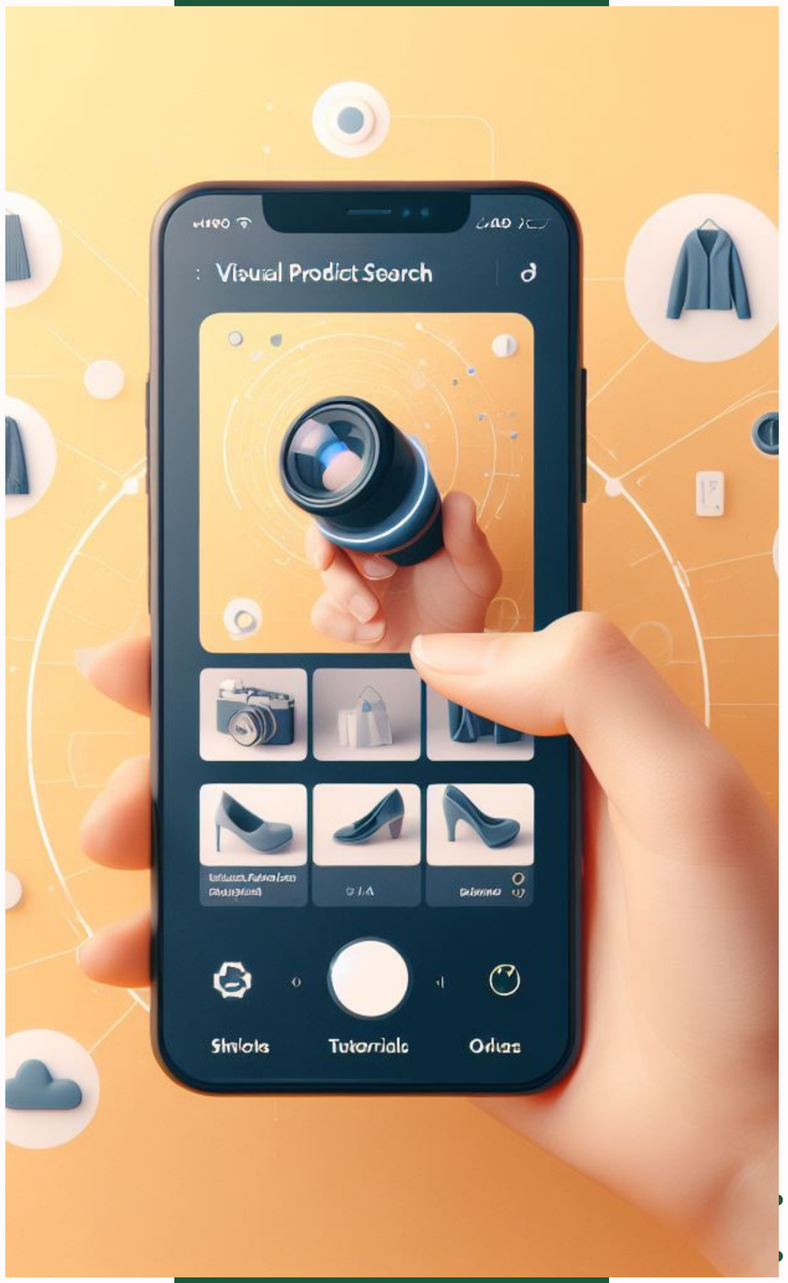
Presentation 2023

VISUAL PRODUCT RECOGNITION



Group id : 16

Content

01 Overview

02 Solution description

03 Data Preparation

04 Data Preprocessing

05 Model Training

06 Similarity Search

07 Model Validation

08 Assumptions

09 Results

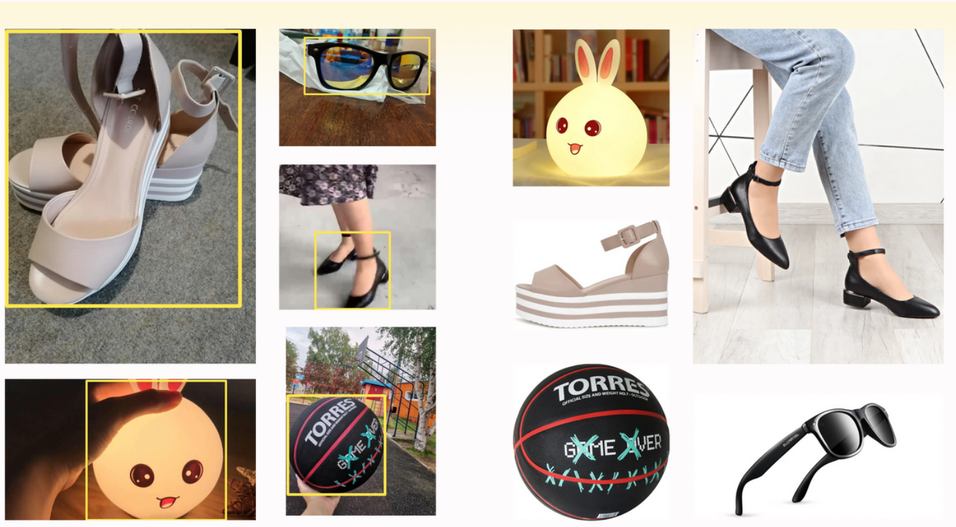
10 Challenges

Overview

- Enabling quick and precise search among millions of items on marketplaces is a key feature for e-commerce.
- The use of common text-based search engines often requires several iterations and can render unsuccessful unless exact product names are known.
- Image-based search provides a powerful alternative and can be particularly handy when a customer observes the desired product in real life, in movies or online media.
- This challenge aims to push existing computer vision methods to their limits in the context of image-based product search.

🔎 Problem Statement

In this challenge we separate product images into user and seller photos. User photos are typically snapshots of products taken with a phone camera in cluttered scenes. Such images differ substantially from seller photos that are intended to represent products on marketplaces. We provide object bounding boxes to indicate desired products on user photos and use such images and boxes as search queries. Given a search query, the goal of the algorithm is to find correct product matches in the gallery of seller photos.

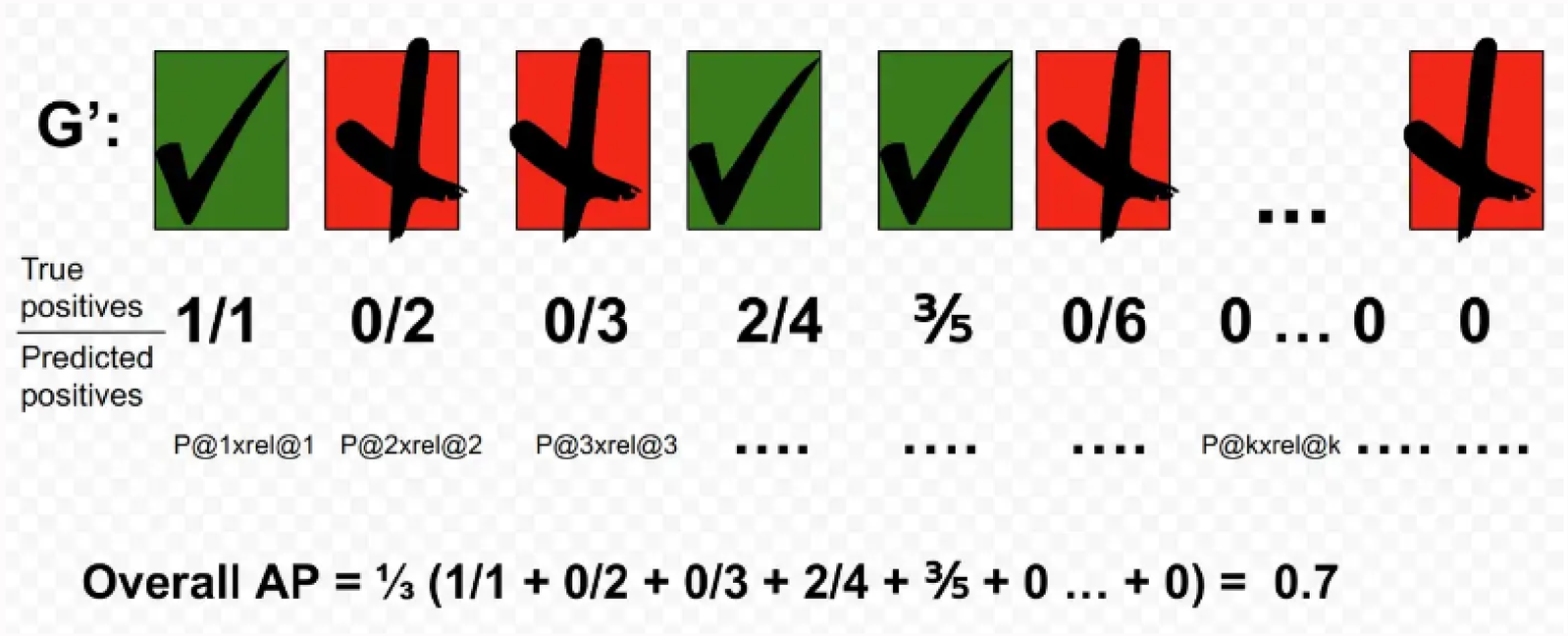


User images

Seller images

Evaluation Metric

- The competition's evaluation metric is the mean Average Precision (mAP).
- The organizers have provided a clear explanation, detailing how each image is scored based on its location and the index of the Ground-Truth (GT) image.
- To achieve the highest score, the location should precisely match the index of the GT image.



Solution

- Evaluation
- Loss Function and Optimization
- Model Architecture and Training Configurations
- Data Preparation and Training

Datasets

- For most time of the competition, We were just using Product10K pointed out by the organizers.
- Here are some key statistics about the dataset:

1. ~200k images
2. ~9700 classes

- Later in the project, I made the decision to incorporate the H&M dataset for further experimentation. This choice was made following earlier attempts with several other datasets throughout the project.

Data Preparation

- "products-10k" images are grouped by class labels, and samples from each class are collected for training. Classes with a sufficient number of samples are chosen.
- Ensured that each class had enough samples for training.

Data Preprocessing

- Diverse data augmentation techniques, such as AutoAugment, AugMix, and ColorJitter, were applied to enhance the diversity and quality of the training dataset.

- Images were resized, converted to tensors, and normalized to ensure consistent data formatting.

```
    final_transform = T.Compose([
            T.Resize(
                size=(CFG.image_size, CFG.image_size),
                interpolation=T.InterpolationMode.BICUBIC,
                antialias=True),
            T.ToTensor(),
            T.Normalize(
                mean=(0.48145466, 0.4578275, 0.40821073),
                std=(0.26862954, 0.26130258, 0.27577711)
            )
        ])
```

```
    if data_aug == 'image_net':
        transform = T.Compose([
            T.ToPILImage(),
            T.AutoAugment(T.AutoAugmentPolicy.IMAGENET)
        ])
```

Model Training

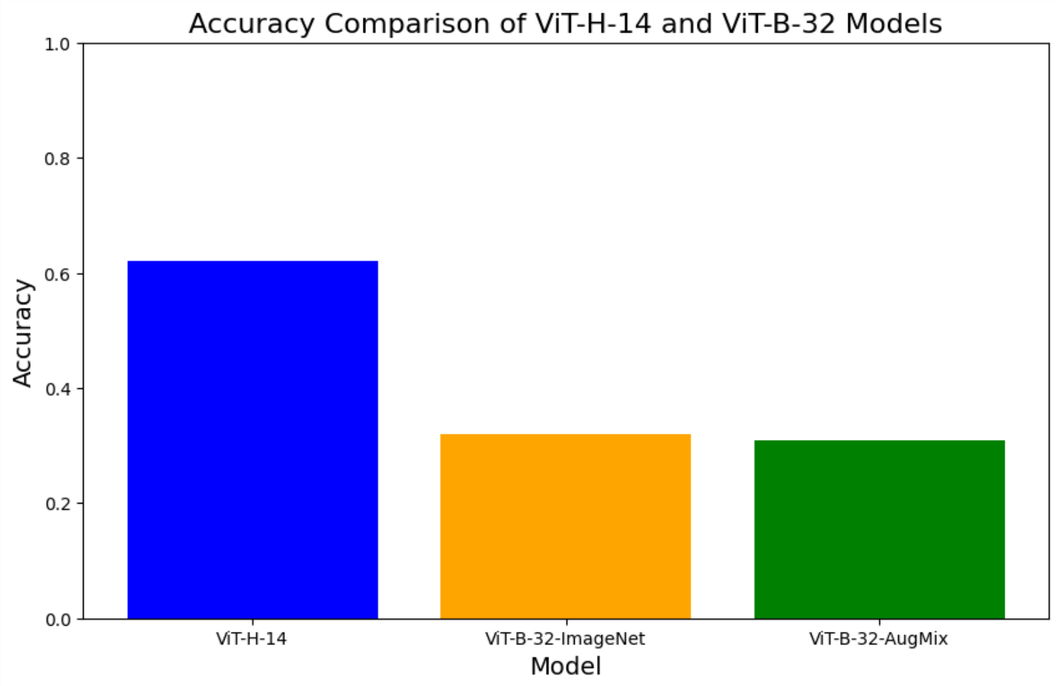
We utilized two versions of the pipeline

- ViT-H-14: Leveraged for higher-resolution image processing and intricate feature extraction.
- ViT-B-32-quickgelu : Utilized for efficient computation and faster inference times, suitable for large-scale datasets.

Tailored data augmentation strategies to optimize training data preprocessing for each CLIP model variant.

Loss Function : ArcFace with Adaptive margin

Optimizer : AdamW + CosineScheduler with a warmup.



Similarity Search

- Leveraging a k-nearest neighbors (kNN) approach, 1000 nearest image embeddings were computed for each query image.

- The Euclidean distance was utilized to measure the dissimilarity between feature vectors, enabling efficient similarity calculation.

- Feature vector extraction, kNN search based on Euclidean distances, and retrieval of the most visually similar images for each query image were performed.

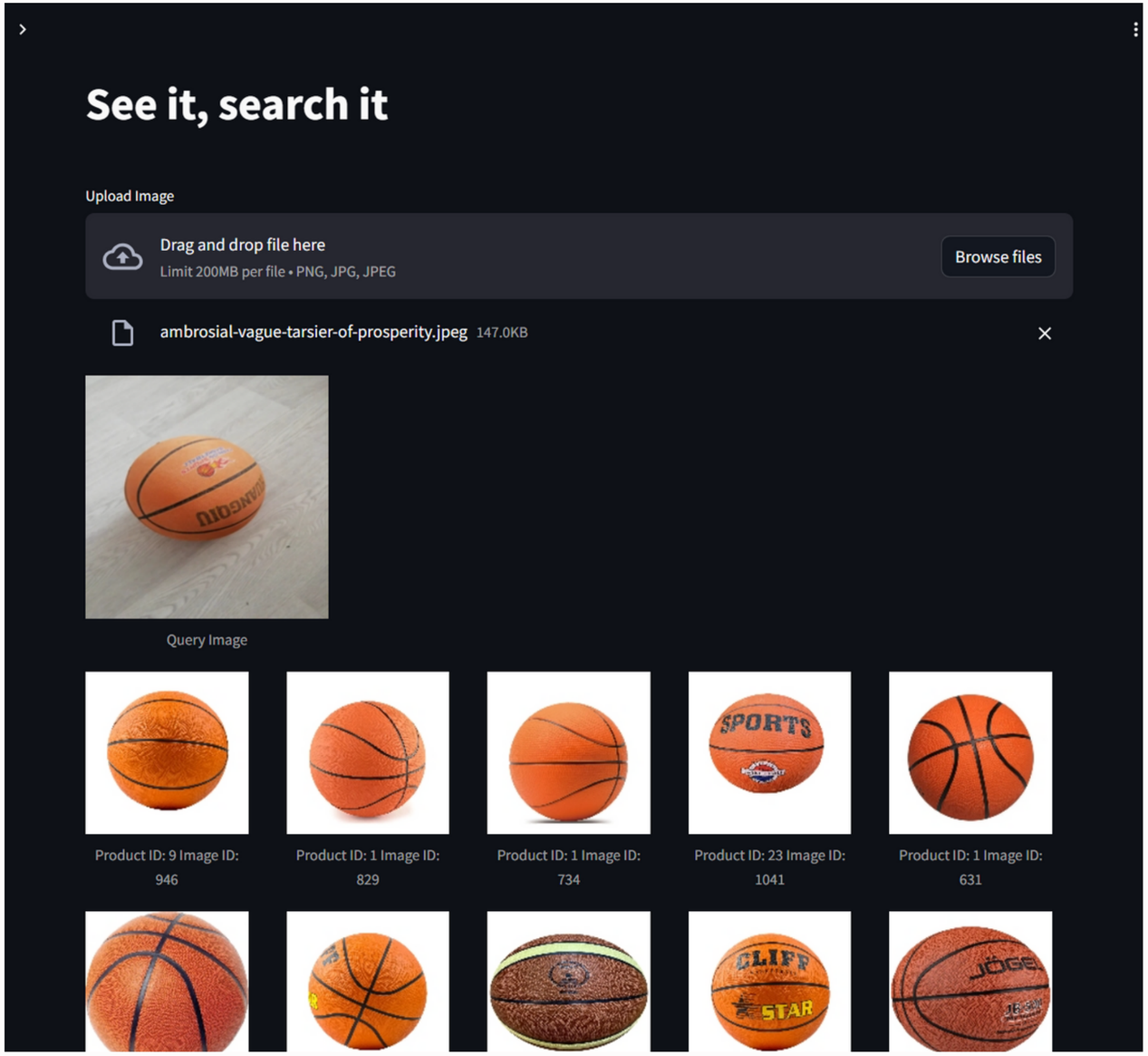
Model Validation

- Utilized the development test dataset to perform similarity searches between query and gallery images.
- Identified the top 1000 similar images from the gallery dataset for each query image.
- Retrieved product labels corresponding to the identified similar images for further analysis.
- Calculated the mAP (mean Average Precision) score as an accuracy metric for the retrieval process.
- Achieved an impressive mAP score of 0.62, showcasing the efficacy of the ViT-H-14 CLIP model in accurately retrieving relevant product images based on the queries.

Assumptions

- The quality and availability of the product images are sufficient for accurate similarity search and retrieval.
- The product categories and labels provided in the dataset are accurate and comprehensive.
- The trained model's performance is indicative of its generalization capabilities to real-world scenarios.
- Assumed user-uploaded images have no background distractions and solely focus on the relevant product.
- The chosen image preprocessing and augmentation techniques are appropriate for improving the model's performance without introducing significant noise or bias.

Results


See it, search it
Upload Image
Drag and drop file here
Limit 200MB per file • PNG, JPG, JPEG
Browse files
ambrosial-vague-tarsier-of-prosperity.jpeg 147.0KB
Query Image
Product ID: 9 Image ID: 946
Product ID: 1 Image ID: 829
Product ID: 1 Image ID: 734
Product ID: 23 Image ID: 1041
Product ID: 1 Image ID: 631
SPORTS
CLIFF
STAR
JÖGE

Challenges

- Managing memory limitations and computational resources when processing and analyzing large-scale image datasets.

- ViT-B-32 exhibited comparatively lower performance in terms of accuracy, potentially due to its limited capacity for intricate feature extraction. However, ViT-B-32 was noted for its efficient memory usage, making it a suitable option for environments with resource constraints. Conversely, ViT-H-14 demonstrated higher accuracy but at the cost of increased memory consumption, making it more suitable for scenarios with ample computational resources.

- Various experiments were conducted, but no significant improvement was observed.

Team Members

1. PAIRAVI . T (200441F)

2. PATHIRANA A.S.W (200448H)

3. PATHIRAJE P.M.H.K (200447E)

THANK YOU