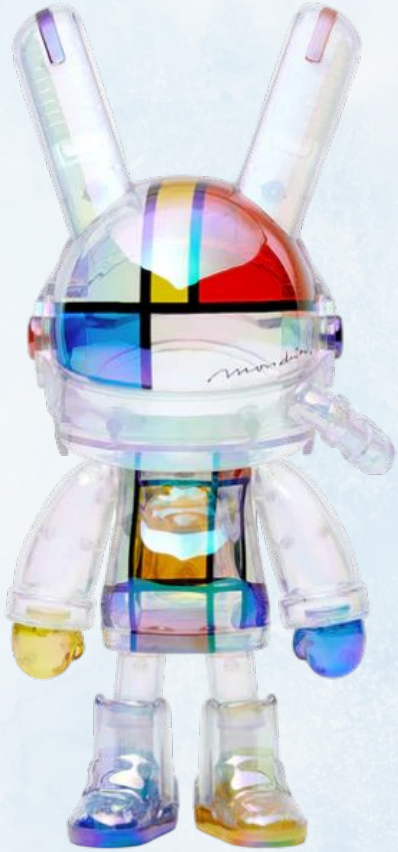
Модель классификации изображений игрушек Robbi



Данные

Датасет:

- 643 картинок, 36 класса
- на каждый класс в среднем 15-18 изображений
- удалены плохие кейсы, где видно только часть игрушки или далеко расположены
- отказались от Segmentation API в сторону ручной

Модель	оригинал	автосегментация	ручная
Clip	70%	73.92%	-
Maws	71%	73.73%	79.96%
Clip ViT large	73.7%	75.8%	89.27%

* % топ 5 accuracy

Датасет: Оригинал - Segmentation API - Ручная

Промты

Промты:

- **написаны вручную, использовалась Chat GPT4**
- **до 70% accuracy на изменениях промта**
- **на 10% увеличилось accuracy после указания конкретного предмета**

Было: "Inside the head there is a triangle with a ball in the middle, a transparent head and a blue bottom, a gradient."

Стало: "The robot features a transparent head with a triangle and a ball in the center, creating focal point. Its color appears to change depending on the background and lighting, leading to varying light distortions. The body has a blue gradient towards the bottom."

Примеры

"The figure features a white and yellow color scheme, with yellow accents on its ears, arms, and boots. It has a transparent helmet with a hat on top, gloves draped like a scarf on the chest, and warm, padded boots on its feet.",

"The robot has a glossy black finish with red accents and displays a 'B' logo on its chest.",

"This figure is transparent with a gradient of pastel rainbow colors. Its helmet is filled with small colorful balls.",

"The robot is white with a metallic finish on its helmet. It is depicted next to a sleek, futuristic bicycle.",

"This robot is purple with a metallic sheen and is positioned beside a matching futuristic chair.",

Модели

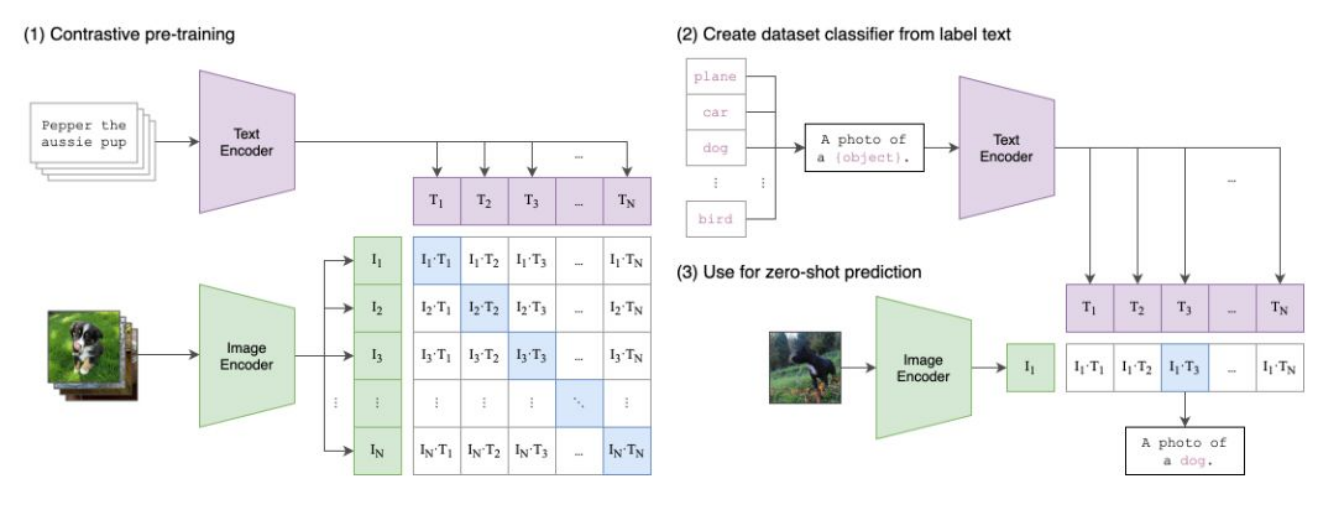
Maws zero shot

MAWS (Multilingual Augmented Word Spaces) - это модель для мультимодального понимания, разработанная Facebook AI Research, которая использует данные изображений и текстов на различных языках

- **Мультиязычность - обучена на разных языках**
- **Механизм внимания**

Архитектура:
- Vision Transformer (ViT): Использует трансформеры для обработки изображений, разбивая их на патчи и обрабатывая последовательности патчей.
- XLM-RoBERTa: Мощная языковая модель, использующая трансформеры для обработки текстов на различных языках.

Clip zero shot



Это многоязычная версия модели OpenAI CLIP, отображает текст и изображения в общее векторное пространство. Модель можно использовать для поиска изображений и многоязычной классификации изображений с нулевым кадром (метки изображений определяются как текст).

- **CLIP ViT-B/32 использует патчи размером 32x32, быстрая и легкая.**
- **CLIP ViT Large обладает более крупной архитектурой, что позволяет ей обрабатывать сложные изображения с большей точностью.**

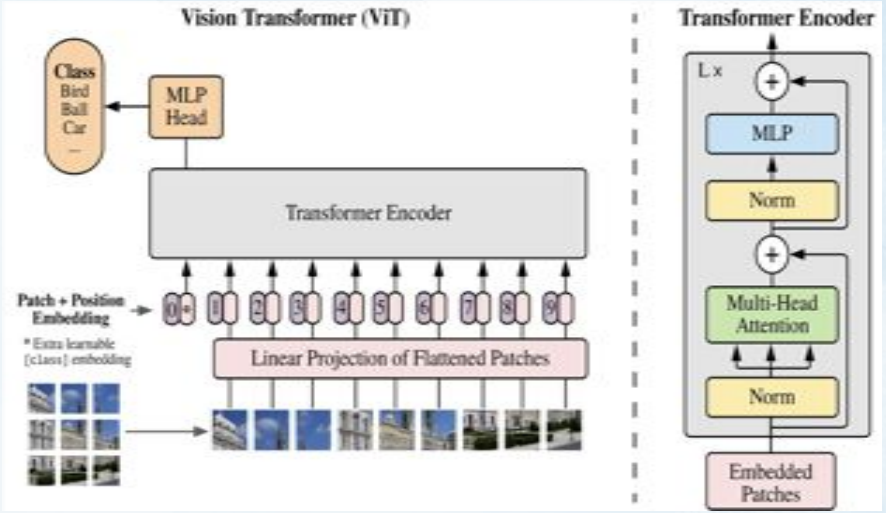
Выбранная модель - VIT

Vision Transformer - основана на трансформерах, адаптирована для обработки изображений

Используется механизм внимания

Основная идея - изображение в виде последовательности патчей

Далее применяется архитектура трансформера



Результаты

Clip ViT Large - оказалось лучшей моделью - accuracy 89%

Модель	оригинал	автосегментация	ручная
Clip	70%	73.92%	-
Maws	71%	73.73%	79.96%
Clip ViT large	73.7%	75.8%	89.27%

* % топ 5 accuracy