

Projekt: Wykorzystanie CLIP do generowania opisów zdjęć satelitarnych

Szymon Laszczyński

February 5, 2025

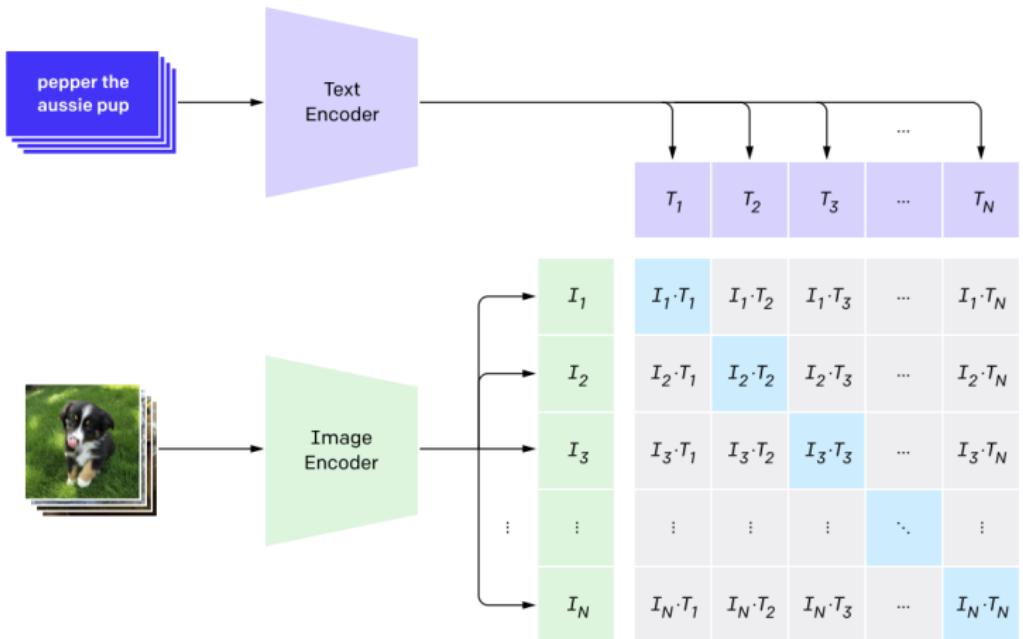
CLIP (Contrastive Language-Image Pretraining)

Łączenie osadzeń obrazów i tekstów w uwspomnionej przestrzeni poprzez uczenie różnicowe (ang. contrastive learning).

Kluczowe cechy:

- ▶ Trenowany na duzych zbiorach **par obraz-tekst**.
- ▶ realizuje zadania wizyjne **zero-shot** bez konieczności dostosowywania modelu.

CLIP - uczenie różnicowe



source: <https://openai.com/index/clip/>

Cele

Sukces częściowy:

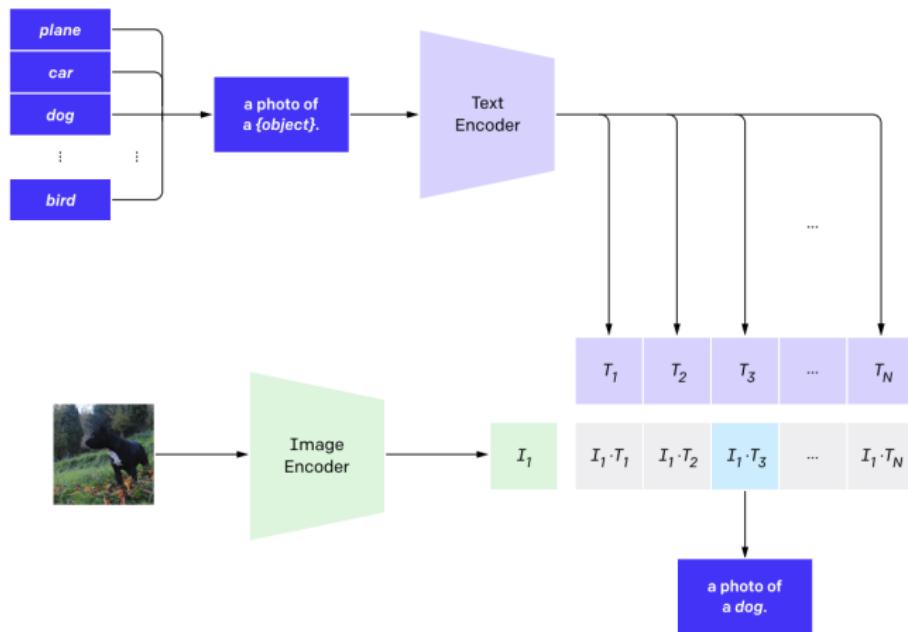
- ▶ rozpoznawanie obiektów na zdjęciach satelitarnych (zbiór etykiet).
- ▶ identyfikacja ograniczeń CLIP.

Pełen sukces:

- ▶ generowanie opisu do zdjęć satelitarnych (bez etykiet).

Opis na podstawie etykiet

Korzystamy z CLIP by zidentyfikować najbardziej prawdopodobne etykiety ze znanego zbioru.



source: <https://openai.com/index/clip/>

Wyniki - XView

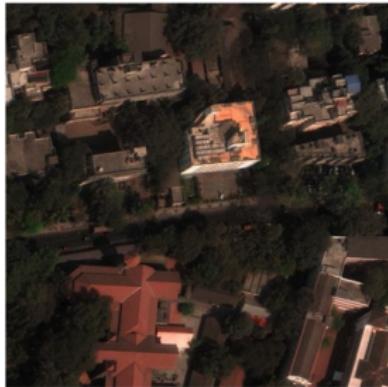
Testowanie na zbiorze z wieloma etykietami. Wszystkich etykiety jest 60.

Przykłady: *Building, Small Car, Cargo Truck, Utility Truck, Helipad, Damaged Building, Shipping container lot, Construction Site, Facility, Vehicle Lot, Small Aircraft, Storage Tank, Shipping Container.*

Wybieramy 10 najbardziej prawdopodobnych etykiet.

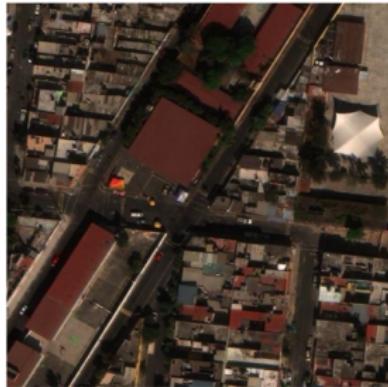
- ▶ 72% że przynajmniej jedna z etykiet została poprawnie zidentyfikowana.
- ▶ 14% że wszystkie etykiety zostały poprawnie zidentyfikowane.

Wyniki - XView



Bus, Small Car, Building

30% Helipad
14% Damaged Building
11% Shipping container lot
10% **Building**
7% Construction Site
6% Facility
4% Vehicle Lot
3% Small Aircraft



Small Car, Utility Truck,
Cargo Truck, Bus, Building

24% Shipping container lot
16% Helipad
10% Vehicle Lot
6% Storage Tank
3% Shipping Container
3% **Small Car**
3% Small Aircraft
2% **Cargo Truck**

Wyniki - EuroSAT

Testowanie na zbiorze z jedną etykietą dla zdjęcia.

Wszystkich etykiety jest 10: *Annual Crop, Forest, Herbaceous Vegetation, Highway, Industrial Buildings, Pasture, Permanent Crop, Residential Buildings, River, SeaLake*.

Wybieramy najbardziej prawdopodobną etykietę.

- ▶ 44% dla oryginalnego zestawu etykiet.
- ▶ 58% dla ograniczonego zestawu etykiet.
(field, forest, highway, buildings, body of water)

Wyniki - EuroSAT



Pasture

34% Pasture

75% field



Highway

42% Residential Buildings

17% Highway

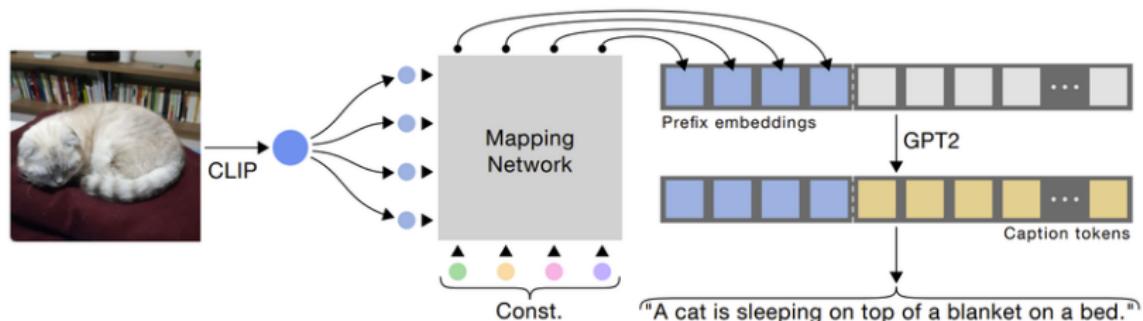
45% highway

Ograniczenia CLIP

- ▶ Problemy z precyzyjnym identyfikowaniem obiektów
- ▶ Wrażliwość na dobór etykiet
- ▶ Bias wywołany przez zbiór treningowy

Opis bez etykiet

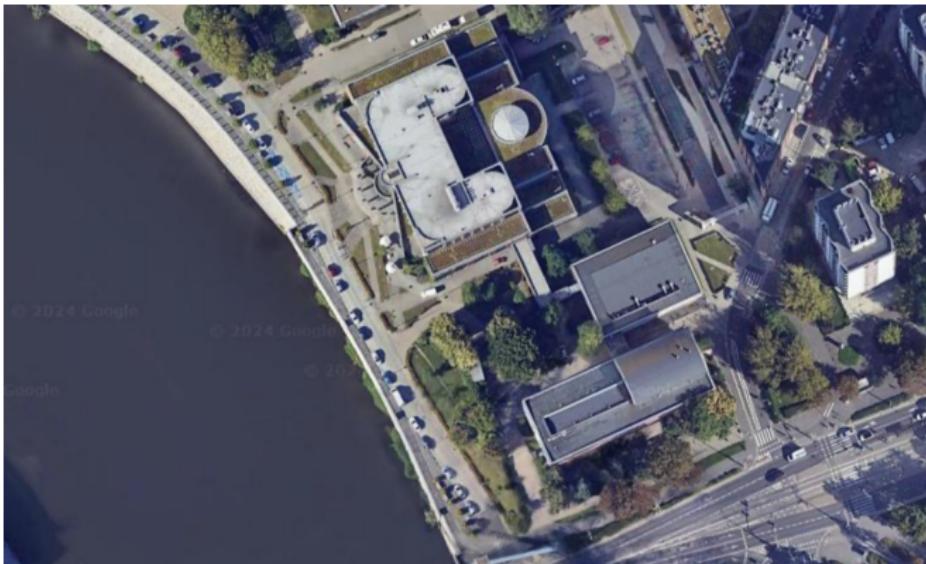
Trenujemy prostą sieć mapującą osadzenia obrazów na prefix dla prompta gpt-2.



source: <https://arxiv.org/pdf/2111.09734>

Trening przeprowadzamy na zbirze: arampacha/rsicd

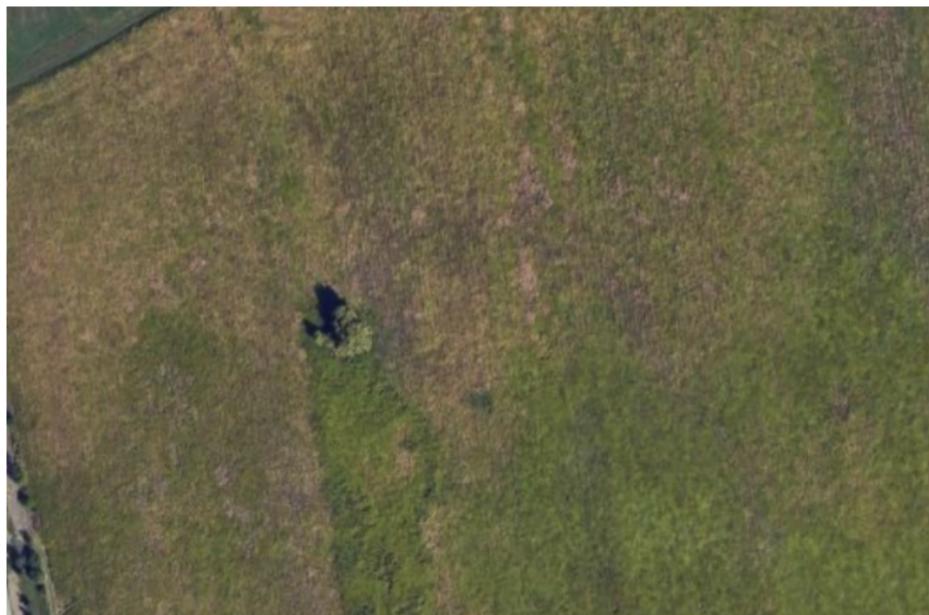
Wyniki



several buildings and green trees are near a river in a seaside resort.

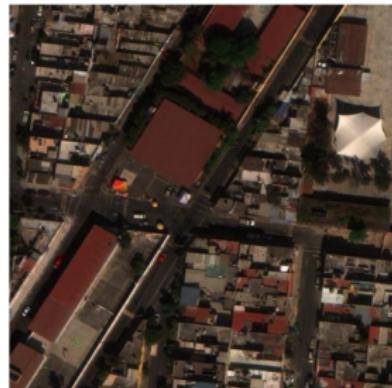
a church is close to a river with many green trees and a pond.

Wyniki



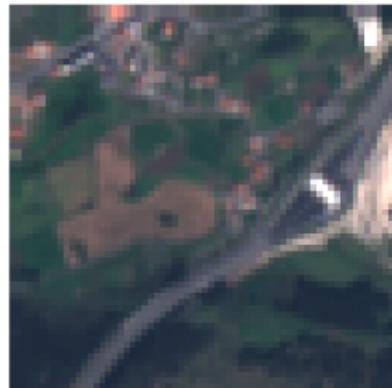
Many green trees are in a piece of green meadow.

Wyniki



XView

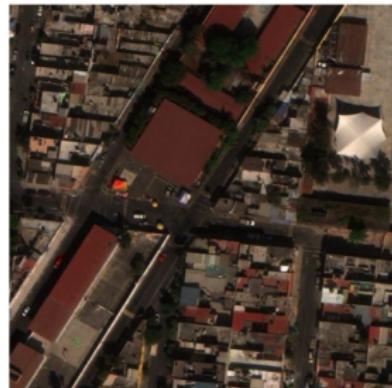
Many buildings and green trees are located in a dense residential area.



EuroSAT

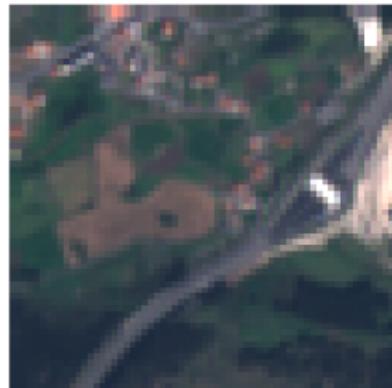
Many green trees are planted on both sides of a piece of green meadow.

Wyniki



XView

Many buildings and green trees are located in a dense residential area.



EuroSAT

Many green trees are planted on both sides of a piece of green meadow.

Repozytorium i źródła

projekt: <https://github.com/simon-joseph/uwr-llm-project>
źródła:

- ▶ CLIP : <https://openai.com/index/clip/>
- ▶ CLIPCAP : <https://arxiv.org/pdf/2111.09734>
- ▶ CLIPCAP :
https://github.com/rmokady/CLIP_prefix_caption

zbiory danych:

- ▶ huggingface.co/datasets/CHichTala/xview
- ▶ huggingface.co/datasets/blanchon/EuroSAT_RGB
- ▶ huggingface.co/datasets/arampacha/rsicd