


Zero-shot Multimodal Deep Learning Models for Military Vehicle Detection – An Analysis

Philipp J. Rösch, **Fabian Deuser**, Konrad Habel, Norbert Oswald

https://go.unibw.de/vis-en

der Bundeswehr Universität München

Applications

- **Analyse large amounts of video and image data** (e.g. surveillance cameras or social media), without time-consuming data collection and training.
- In March 2022: Public traffic surveillance cameras in Germany have been switched off to **prevent possible analyses of movement of troops**

Two own Datasets

Military image dataset (mid)

- Web-scraped images
- Classes:
  - 3041 “military vehicles”
  - 977 “civilian vehicles”
  - 1475 “no vehicles”
- Object-of-interest in center and occupy large fraction


Military video image dataset (mvid)

- Images extracted from 10 YouTube videos
  - 5 “civilian videos”
  - 5 “military videos”
- Each type with
  - 3 dash camera-like videos
  - 2 surveillance camera-like videos
- Object-of-interest partially not in center and very small

From Classification to Zero-Shot Learning

Classical Classification task (Supervised Learning)

label=mil. truck

label=civ. truck

Drawbacks:

- We need (a lot of) labeled data
- We cannot extend it arbitrarily with more classes

Multimodal Zero-shot learning

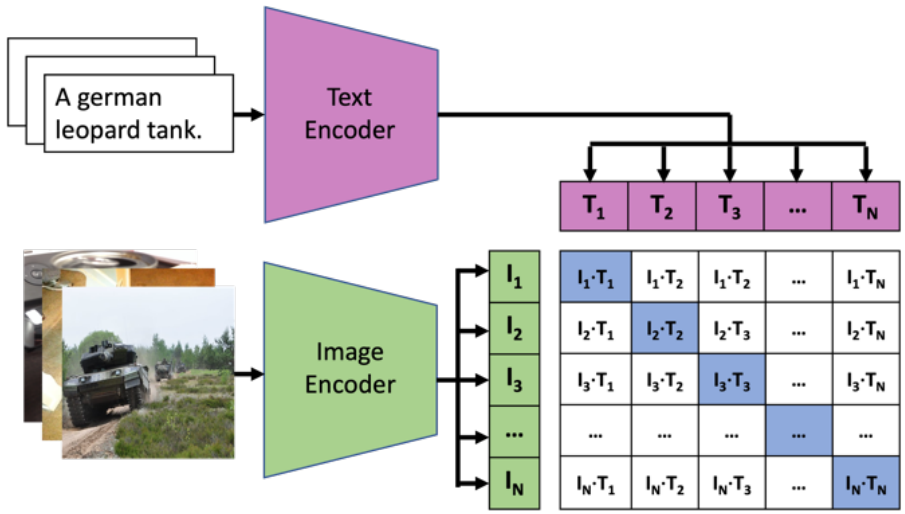
Learning of soft concepts without hard labels

“A military truck on a parking lot”

Enables multiple tasks like OCR, facial recognition and object classification

CLIP model

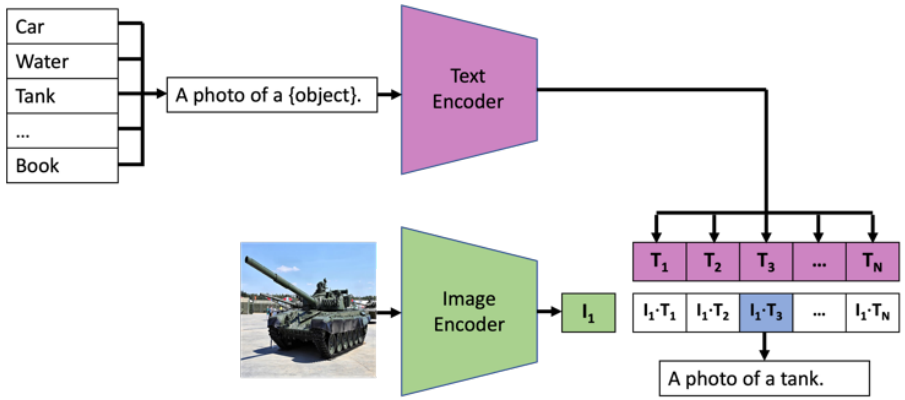
- Contrastive Language-Image Pre-Training (CLIP) model: two encoders which transforms data into an embedding space
- Trained on 400 million image-text pairs to learn concepts



CLIP pre-training objective

CLIP model: Inference using Prompt Engineering

- Use the pre-trained model to run Zero-Shot Classification
- Take <u>one</u> image and <u>several</u> text prompts
- The final prediction is the prompt with the highest similarity score between image and text



CLIP inference

Our Experiments

Prompt Engineering for Zero-Shot Classification using military image dataset (mid)

Transfer insights to (real-world) video analysis

Zero-shot Evaluation on Video Streams using military video image dataset (mvid)

Used Metrics

- Precision
  - How many predictions are correct
- Recall (aka sensitivity)
  - How many true labels get detected
- F1 score
  - Harmonic mean of Precision and Recall
  - Often used for unbalanced datasets

Analysis 1: Prompt Engineering for Zero-Shot Classification

Testing which prompt syntax works best

- Word level:
  - "military vehicle"
  - "civilian vehicle"
  - "everyday object"
- Sentence level:
  - "an image of a military vehicle, a type of vehicle"
  - "an image of a civilian vehicle, a type of vehicle"
  - "an image of an everyday object"
- Ensemble level: 4 more prompts per class (see Appendix)

➡ **Very good results (without any training)!**

Prompt	F1 score	Precision	Recall
word	0.9125	0.8585	0.9736
sentence	0.9424	0.8934	0.9970
ensemble	0.9878	0.9888	0.9868

Results for color images

Prompt	F1 score	Precision	Recall
word	0.9152	0.8624	0.9750
sentence	0.9422	0.8941	0.9957
ensemble	0.9825	0.9832	0.9819

Results for grayscale images

Evaluation of Model on mvid Dataset

- Moderate to very good results
- Depends on video quality
- Small objects are hard to detect

Name	F1 score	Precision	Recall
mil_sumy	1.0000	1.0000	1.0000
mil_border	0.6471	0.9167	0.5000
mil_sun	0.8986	0.8378	0.9688
mil_tank	0.7473	1.0000	0.5965
mil_cctv	0.7778	1.0000	0.6364
civ_red	0.3810	0.4000	0.3636
civ_seoulday	0.5625	0.9000	0.4091
civ_seoulnight	0.6000	1.0000	0.4286
civ_street	0.7857	1.0000	0.6471
civ_ohiotraffic	0.8764	0.7800	1.0000

Typographic Attacks

- CLIP has strong Optical Character Recognition (OCR) capabilities.
- Adding text to images can distract CLIP from the real image content
- CLIP also shows a preference for the text within the image instead of the actual object

Military vehicle:	2.0%
Civilian vehicle:	97.5%
No vehicle:	0.5%

Military vehicle:	96.0%
Civilian vehicle:	4.0%
No vehicle:	0.0%

Where does CLIP look at?

- We use gScoreCAM to visualize the area the model looks at depending on the text prompt.
- The model is able to distinct object and written text.

“an image of a truck”

“an image of a military vehicle”

2016 Feldberg

2022 Wettiner Heide

2022 Kharkiv

Syria

German Autobahn

Implications and Conclusion

- Large scale annotation tool **without fine-tuning and specified training**
- Enabling **large scale search** in video and image data
- **Typographic attack** as possible defense and attack surface
- Defense mechanisms against typographic attacks would have to detect text and mask it
- Faster and more accurate open source intelligence analysis, through robustness to different domains

Thank you for your attention!

Philipp J. Rösch, M.Sc. *@phiyodr*
Research Associate

Fabian Deuser, M.Sc.
Research Associate

Konrad Habel, M.Sc.
Research Associate

Prof. Dr. Norbert Oswald
Professor for AI

https://go.unibw.de/vis-en

Appendix

Appendix

Prompts for “Ensemble level” (5 prompts per class)

- an image of a military vehicle, a type of vehicle
- an image of a military truck or military tank, a type of vehicle
- an image of a military lorry or panzer, a type of vehicle
- an image of a armored fighting vehicles, a type of vehicle
- an image of a military transporter or military tank, a type of vehicle
- an image of a vehicle, a type of vehicle
- an image of a civilian vehicle, a type of vehicle
- an image of a car or truck, a type of vehicle
- an image of a normal vehicle, a type of vehicle
- an image of a civilian car or civilian truck, a type of vehicle
- an image of an everyday object
- an image of a standard object
- an image of an empty street without cars
- an image of a random object
- an image of an empty street