Introduction to CLIP and Application to Meta Data Extraction

Dr. Maram Akila | Zertifizierte KI, 2nd WS on Foundation Models | 27.09.2023

Agenda

Introduction to CLIP and Application to Meta Data Extraction

1. **Introduction to CLIP**
   - What constitutes a Foundation Model?
   - Basic concept of CLIP
   - Example applications
2. **Application to metadata extraction**
   - What and why of metadata
   - Intro to Semantic Testing
   - Example Results
3. **Conclusion**

Image via Easy Diffusion, Model by "realTang", Prompt: „Fraunhofer"

Part 01

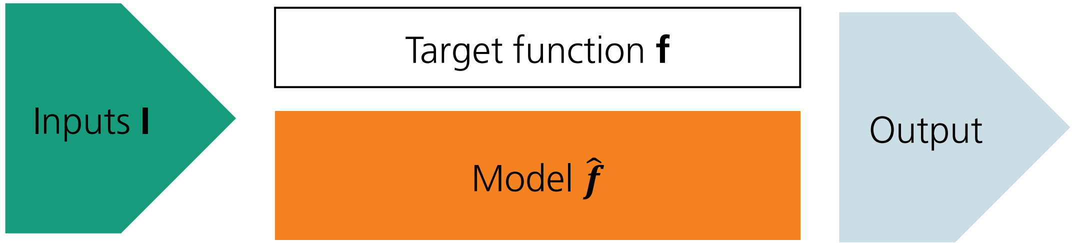
Introduction to CLIP

Caption: An Astronaut riding a horse in a photorealistic Style
(Image curtesy of Dall-E 2)

What constitutes a Foundation Model?

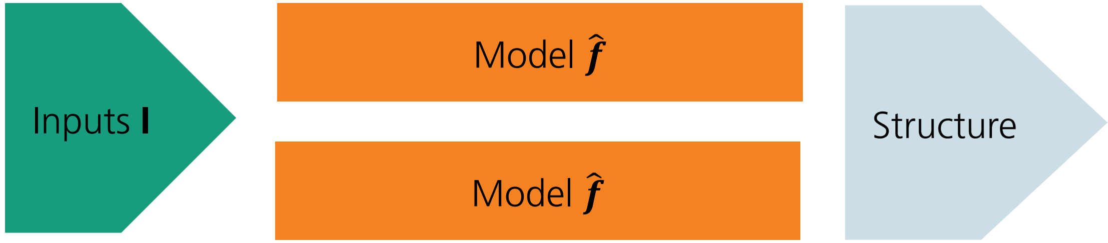
In contrast to other learning approaches

Supervised Learning



- **f** is unknown, but defined via **labelled examples** f(I) – „training set"
- Model training minimizes "loss" $|f(I) - \hat{f}(I)|$
- Predictive model: can be applied to new data

Unsupervised Learning



- There are internal **correlations *among*** the input data
- Examples: input value x is often close to / together with input value y
- $\hat{f}$ models these correlations explicitly (Clustering, itemset mining, graph mining, sketching, spatial analysis, visual analytics, …)
- "quality" : measured by correspondence between $\hat{f}$ and true correlations
- Descriptive model: gives insight into existing data

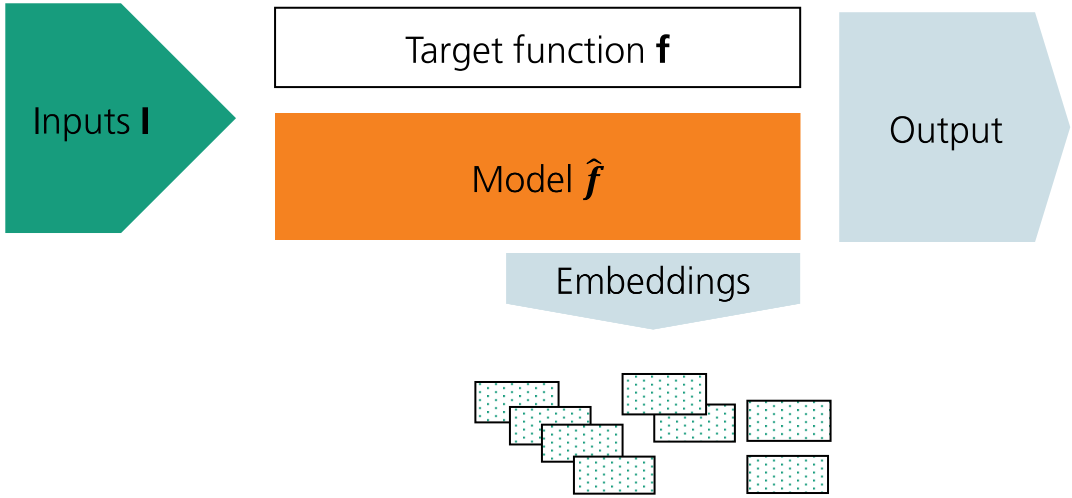
  Public

Fraunhofer IAIS

What constitutes a Foundation Model?

In contrast to other learning approaches

Self-Supervised Learning



- There are internal **correlations *within*** the input data
  - Example: Text "it was a sunny" often followed by "day
- **f** is derived from the input to exploit such correlation
  - Often as a form of reconstruction objective
  - Where parts of input are hidden from model
- Model training then optimizes reconstruction
  - E.g. the most probable next word
  - → Predictive: can be applied to new data
- As "side effect" $\hat{f}$ builds semantically meaningful **embeddings** of I in a latent space
  - → Descriptive: learns structural properties of (existing) data
  - Representations often useful for (related) downstream tasks (fine-tuning)

  Public

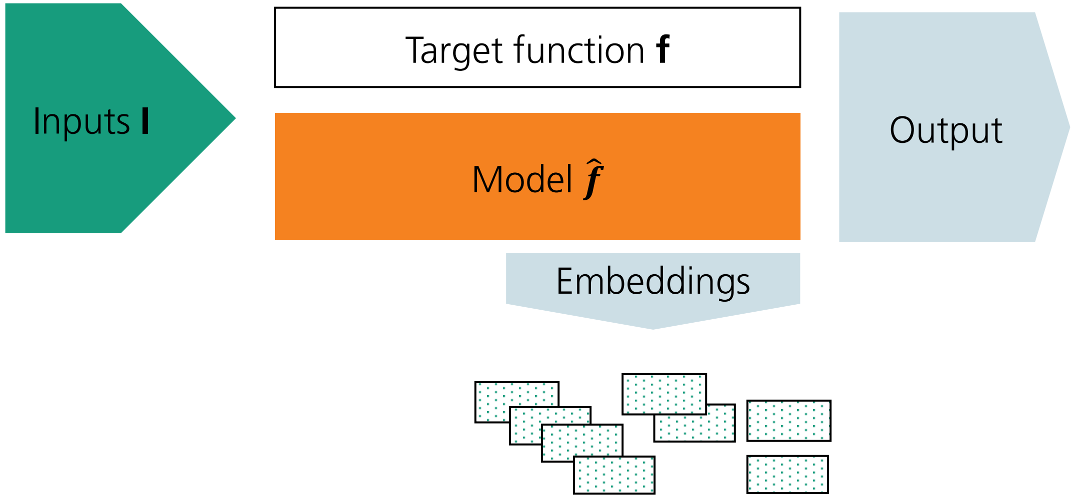
Fraunhofer IAIS

What constitutes a Foundation Model?

In terms of capability

Capable to train on large amount of unlabelled data

Self-Supervised Learning



- Useful as building block / "backbone" for multiple downstream tasks
- Emergent capabilities (especially in the text domain)

- There are internal **correlations *within*** the input data
  - Example: Text "it was a sunny" often followed by "day
- **f** is derived from the input to exploit such correlation
  - Often as a form of reconstruction objective
  - Where parts of input are hidden from model
- Model training then optimizes reconstruction
  - E.g. the most probable next word
  - → Predictive: can be applied to new data
- As „side effect" $\hat{f}$ builds semantically meaningful **embeddings** of I in a latent space
  - → Descriptive: learns structural properties of (existing) data
  - Representations often useful for (related) downstream tasks (fine-tuning)

Transformer Networks create Embeddings

Text and Image Transformers

Text Transformers

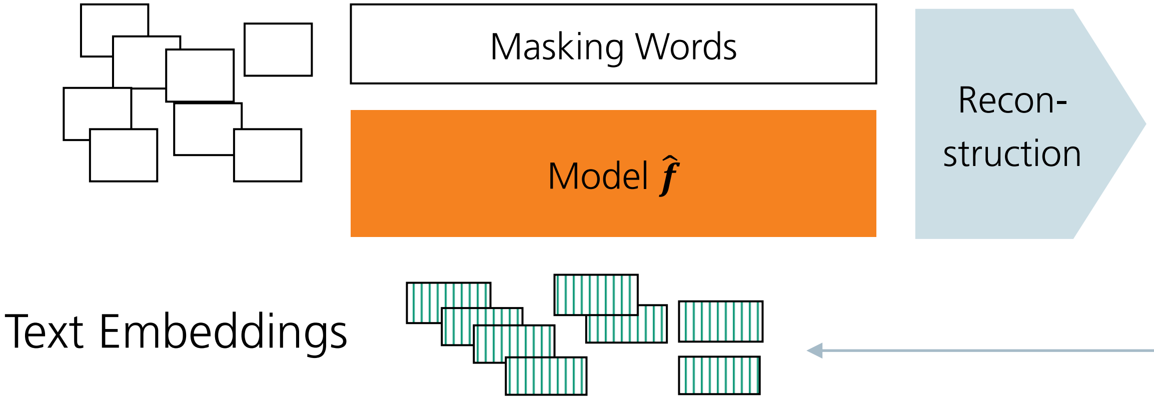
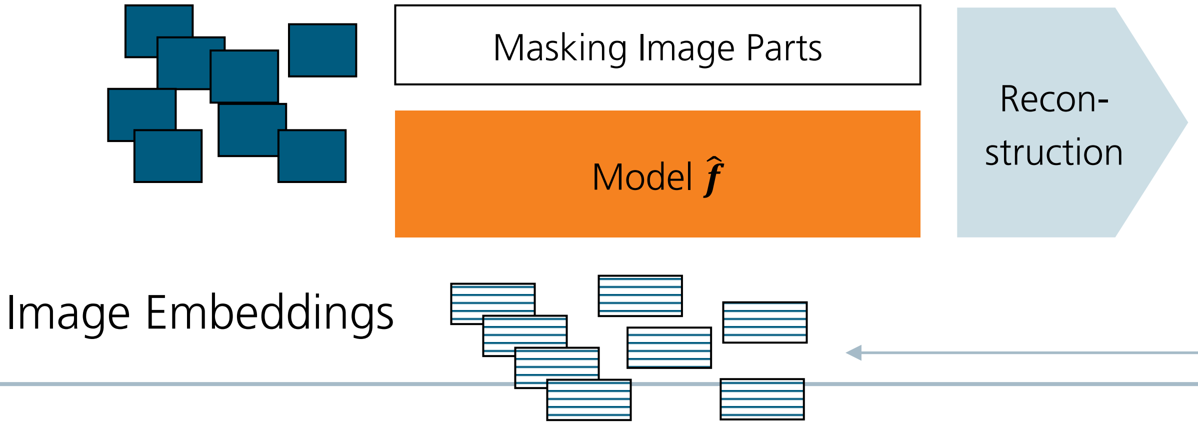
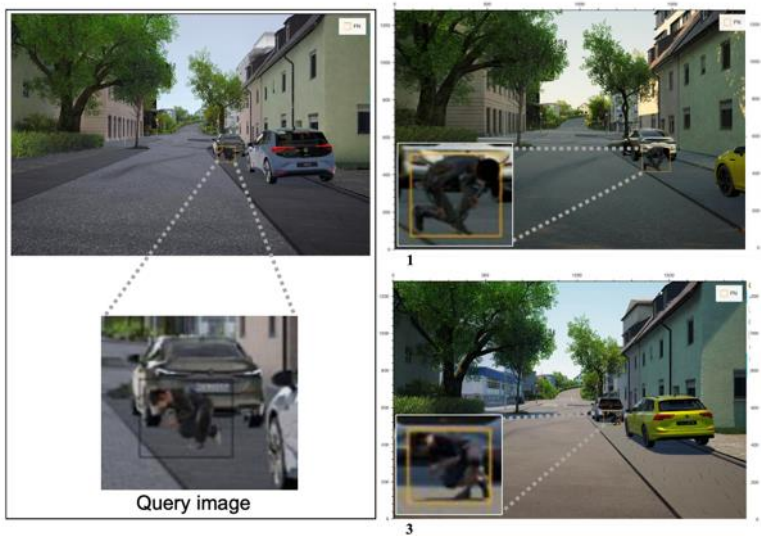


Image Transformers



- Transformer (i.e. attention based) representations proved good embeddings
- Embeddings are „short“ vectors
- Distance in the embedding space have a meaning

Famous example:
"Queen = King – Man" (based on word2vec)

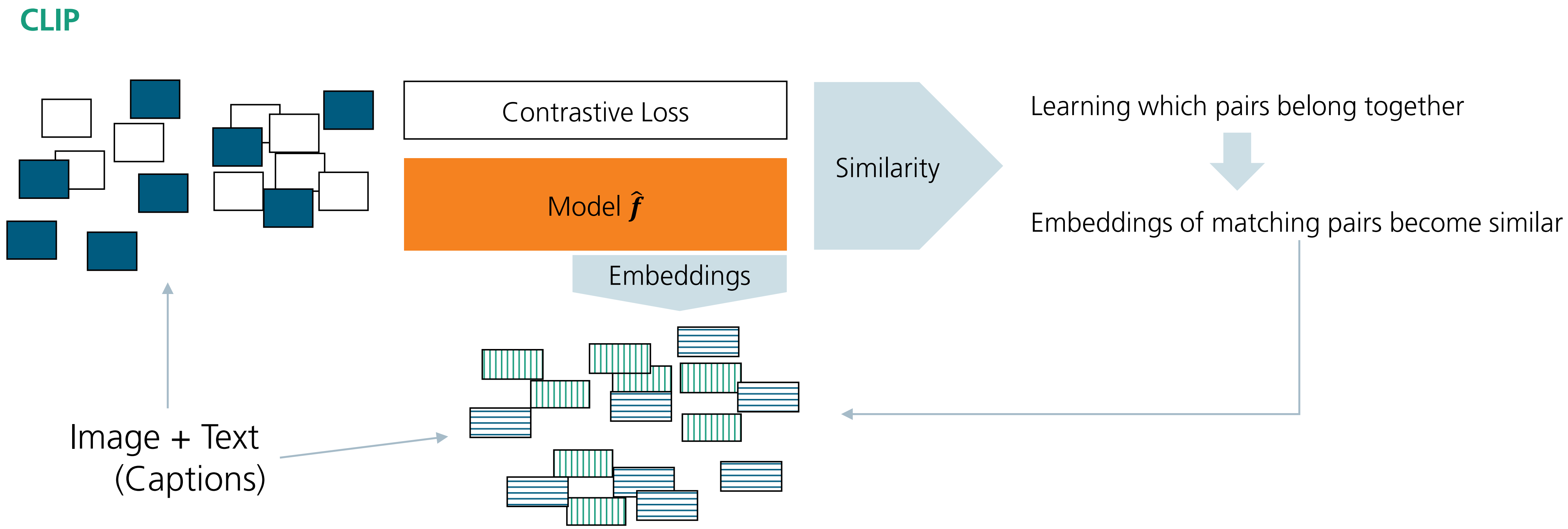


Similarity Search in "ScrutinAI", see
E. Haedecke et al., C&G 114, 265-275 (2023)

Fraunhofer IAIS

Concept of CLIP

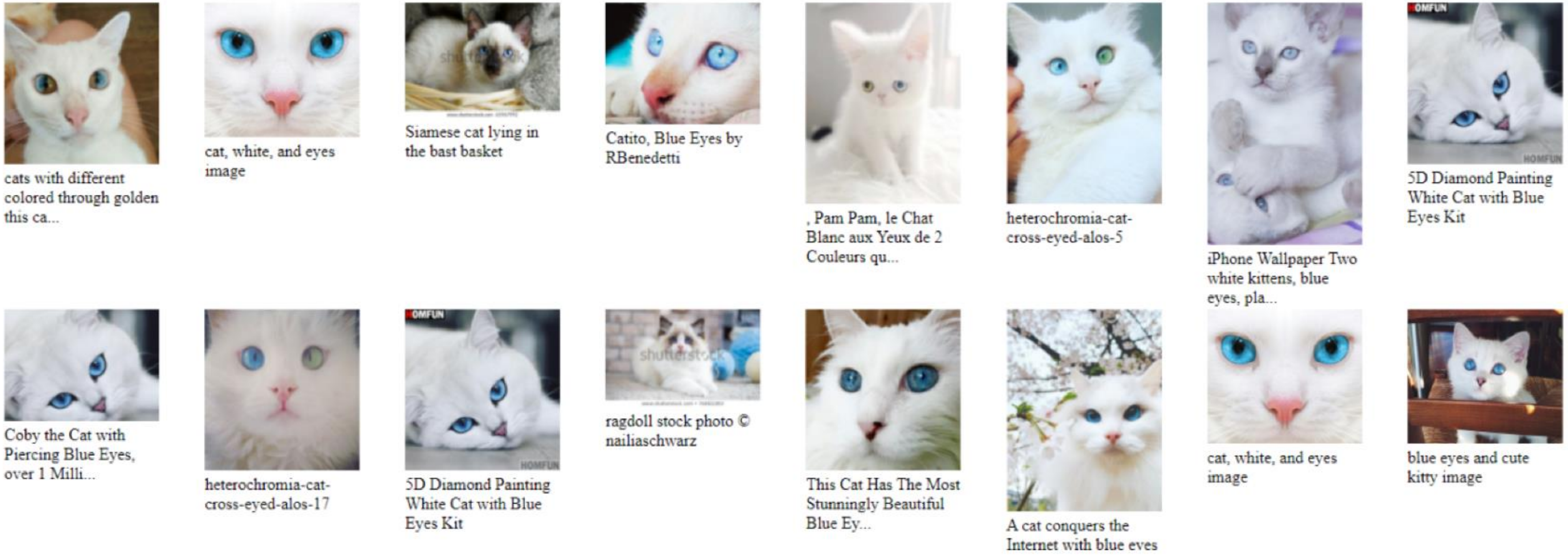
Create joined image and text embeddings



Web-scraped Example Data

Taking images and captions from the internet

"We have filtered all images and texts in the LAION-400M dataset with OpenAI's CLIP by calculating the cosine similarity between the text and image embeddings and dropping those with a similarity below 0.3." - https://laion.ai/blog/laion-400-open-dataset/



Retrieval prompt: "cat with blue eyes"

https://arxiv.org/abs/2111.02114, LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs, Schuhmann et. al, Data Centric AI NeurIPS Workshop 2021

Fraunhofer IAIS

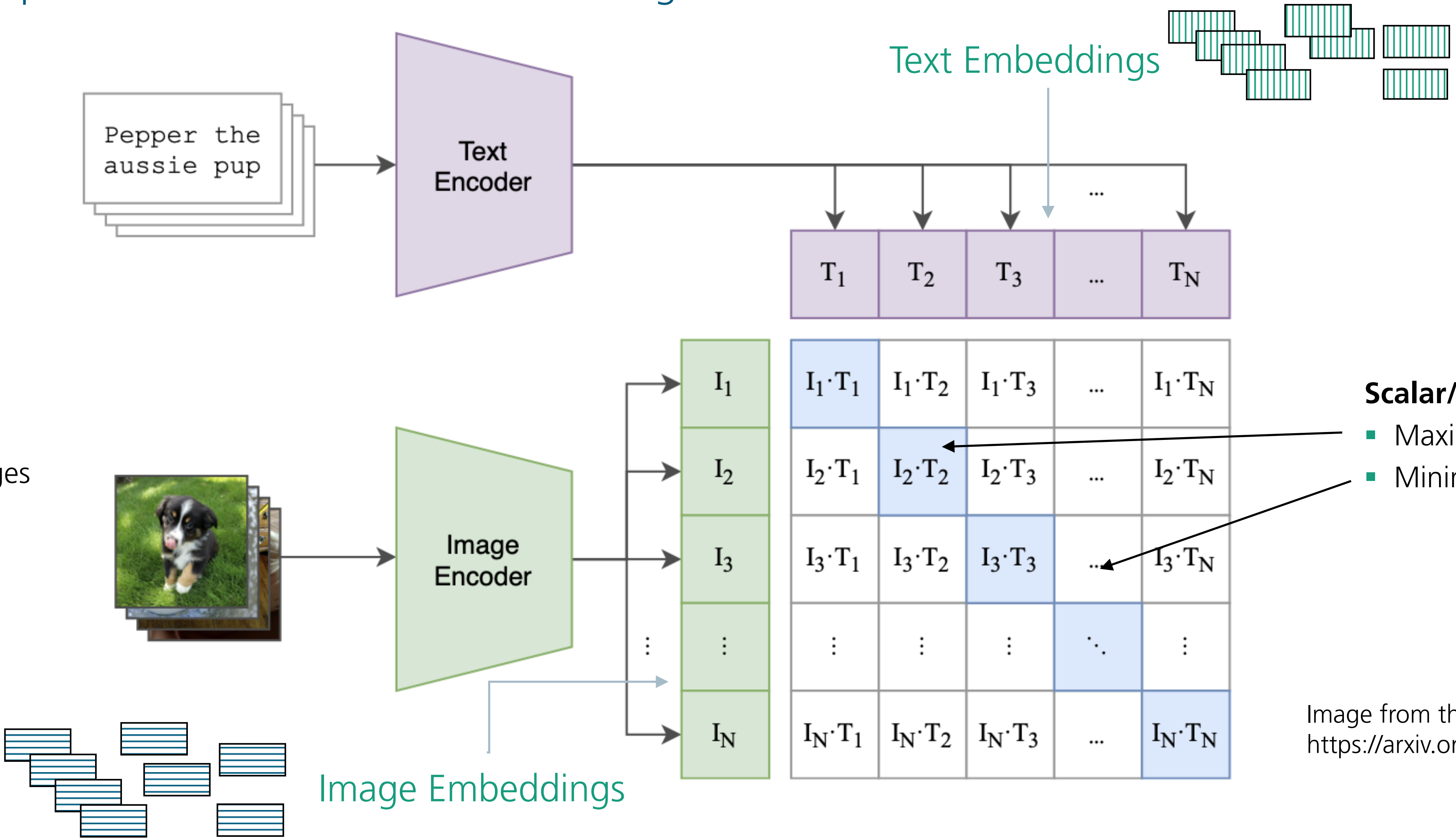
Using Contrastive Loss for training image/text pairs

Similar Inputs should have similar embeddings

400M Texts

x

400M Images



Scalar/Dot Product

- Maximise Diagonal
- Minimise Off-Diagonal

Image from the CLIP Paper
https://arxiv.org/abs/2103.00020

Summary on CLIP Model

Contrastive Language-Image Pre-training (CLIP)

- Trained on 400 million image + text pairs (not LAION)
- CLIP is class of models
  - with variations in encoder type (ResNet, ViT)
  - And latent space size (commonly 512 dimensions)
- Approximate Training Duration
  - ~2 weeks
  - 200-600 (V100) GPUs

Pre-Trained model(s) available

- *Remark:* Approach very general, learning on human described data

Learning Transferable Visual Models From Natural Language Supervision

Alec Radford [*1] Jong Wook Kim [*1] Chris Hallacy [1] Aditya Ramesh [1] Gabriel Goh [1] Sandhini Agarwal [1]
Girish Sastry [1] Amanda Askell [1] Pamela Mishkin [1] Jack Clark [1] Gretchen Krueger [1] Ilya Sutskever [1]

Abstract

State-of-the-art computer vision systems are trained to predict a fixed set of predetermined object categories. This restricted form of supervision limits their generality and usability since additional labeled data is needed to specify any other visual concept. Learning directly from raw text about images is a promising alternative which leverages a much broader source of supervision. We demonstrate that the simple pre-training task of predicting which caption goes with which image is an efficient and scalable way to learn SOTA image representations from scratch on a dataset of 400 million (image, text) pairs collected from the internet. After pre-training, natural language is used to reference learned visual concepts (or describe new ones) enabling zero-shot transfer of the model to downstream tasks. We study the performance of this approach by benchmarking on over 30 different existing computer vision datasets, spanning tasks such as OCR, action recognition in videos, geo-localization, and many types of fine-grained object classification. The model transfers non-trivially to most tasks and is often competitive with a fully supervised baseline without the need for any dataset specific training. For instance, we match the accuracy of the original ResNet-50 on ImageNet zero-shot without needing to use any of the 1.28 million training examples it was trained on. We release our code and pre-trained model weights at https://github.com/OpenAI/CLIP.

1. Introduction and Motivating Work

Pre-training methods which learn directly from raw text have revolutionized NLP over the last few years (Dai & Le, 2015; Peters et al., 2018; Howard & Ruder, 2018; Radford et al., 2018; Devlin et al., 2018; Raffel et al., 2019).

*Equal contribution [1]OpenAI, San Francisco, CA 94110, USA. Correspondence to: <{alec, jongwook}@openai.com>.

Task-agnostic objectives such as autoregressive and masked language modeling have scaled across many orders of magnitude in compute, model capacity, and data, steadily improving capabilities. The development of "text-to-text" as a standardized input-output interface (McCann et al., 2018; Radford et al., 2019; Raffel et al., 2019) has enabled task-agnostic architectures to zero-shot transfer to downstream datasets removing the need for specialized output heads or dataset specific customization. Flagship systems like GPT-3 (Brown et al., 2020) are now competitive across many tasks with bespoke models while requiring little to no dataset specific training data.

These results suggest that the aggregate supervision accessible to modern pre-training methods within web-scale collections of text surpasses that of high-quality crowd-labeled NLP datasets. However, in other fields such as computer vision it is still standard practice to pre-train models on crowd-labeled datasets such as ImageNet (Deng et al., 2009). Could scalable pre-training methods which learn directly from web text result in a similar breakthrough in computer vision? Prior work is encouraging.

Over 20 years ago Mori et al. (1999) explored improving content based image retrieval by training a model to predict the nouns and adjectives in text documents paired with images. Quattoni et al. (2007) demonstrated it was possible to learn more data efficient image representations via manifold learning in the weight space of classifiers trained to predict words in captions associated with images. Srivastava & Salakhutdinov (2012) explored deep representation learning by training multimodal Deep Boltzmann Machines on top of low-level image and text tag features. Joulin et al. (2016) modernized this line of work and demonstrated that CNNs trained to predict words in image captions learn useful image representations. They converted the title, description, and hashtag metadata of images in the YFCC100M dataset (Thomee et al., 2016) into a bag-of-words multi-label classification task and showed that pre-training AlexNet (Krizhevsky et al., 2012) to predict these labels learned representations which preformed similarly to ImageNet-based pre-training on transfer tasks. Li et al. (2017) then extended this approach to predicting phrase n-grams in addition to individual words and demonstrated the ability of their system to zero-shot transfer to other image

https://arxiv.org/abs/2103.00020

  Public

Fraunhofer IAIS

What can CLIP be used for?

Example application show cases

Similarity Score

CLIP

CLIP can determine (semantic) similarity between caption and image

DALL·E 2

Caption: An Astronaut riding a horse in a photorealistic Style
(Image curtesy of Dall-E 2)

  Public

Fraunhofer IAIS

Image Retrieval via Queries

Example Use-Cases of CLIP model (1/3)

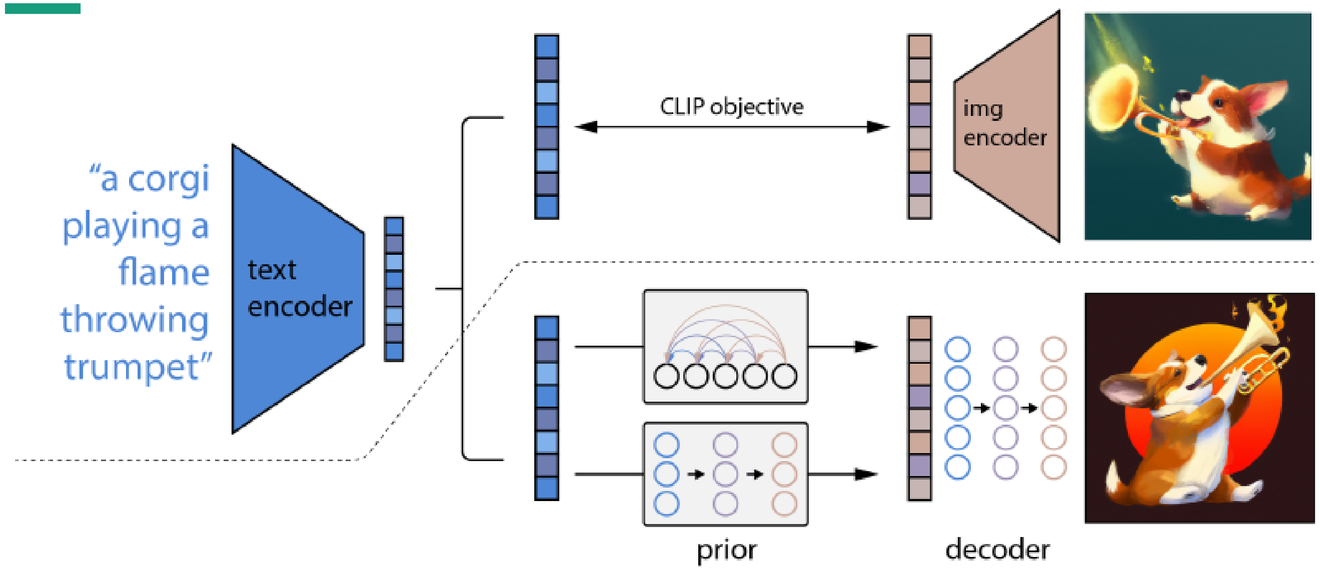
Figure 3: **LAION-5B examples.** Sample images from a nearest neighbor search in LAION-5B using CLIP embeddings. The image and caption (C) are the first results for the query (Q).

https://arxiv.org/abs/2210.08402, LAION-5B: An open large-scale dataset for training next generation image-text models, Schuhmann et. al, NeurIPS 2022, Track on Datasets and Benchmarks

Figures from:
- Ramesh, A. et al., 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*
- https://www.assemblyai.com/blog/how-dall-e-2-actually-works/

Image Generation via Stable Diffusion

Example Use-Cases of CLIP model (2/3)



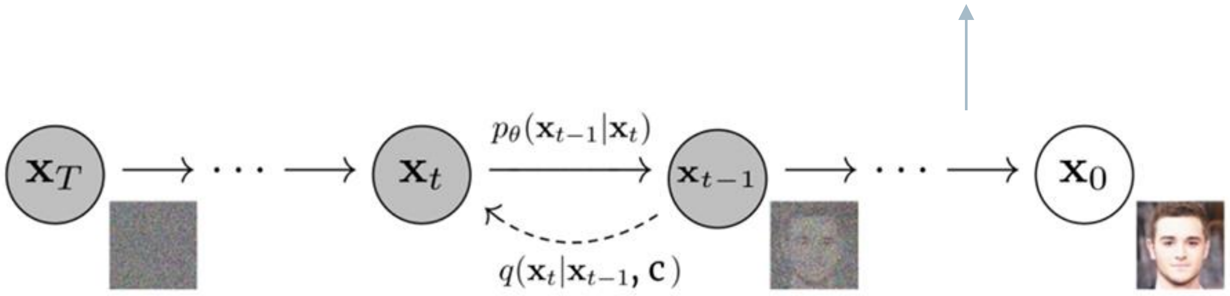


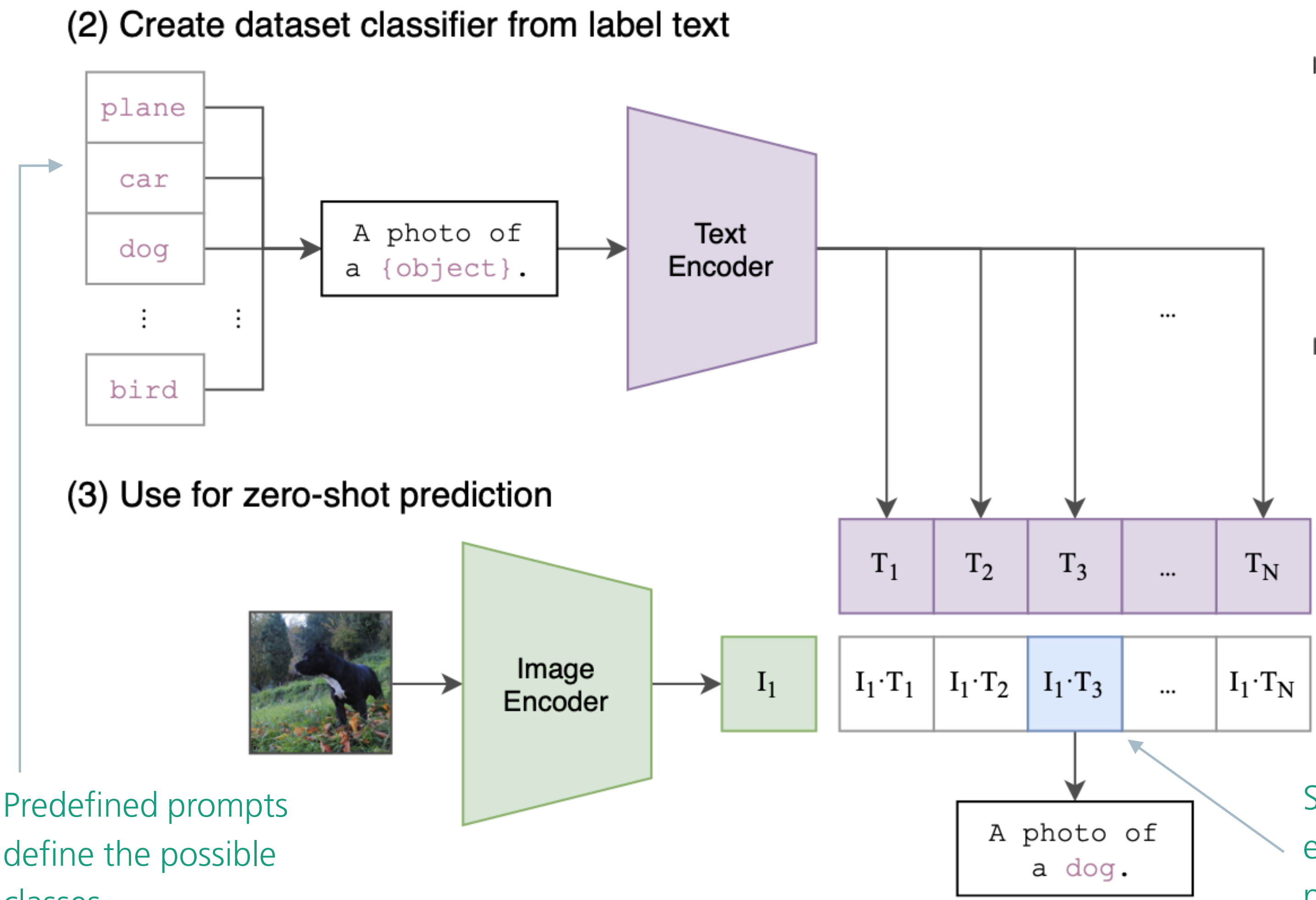
Figure 3: Variations of an input image by encoding with CLIP and then decoding with a diffusion model. The variations preserve both semantic information like presence of a clock in the painting and the overlapping strokes in the logo, as well as stylistic elements like the surrealism in the painting and the color gradients in the logo, while varying the non-essential details.

Simplified Steps of Dall-E

- Prior converts text to image embedding (Improves Quality)
- Image Embedding is decoded into Image
- For this iterative / diffusive "de-noising" process is learned from sample data (i.e. by noising known images)
- Process is conditioned on:
  - (dim. reduced) image embedding (**C**)
  - "time" step **t** of the diffusion
  - previous iteration

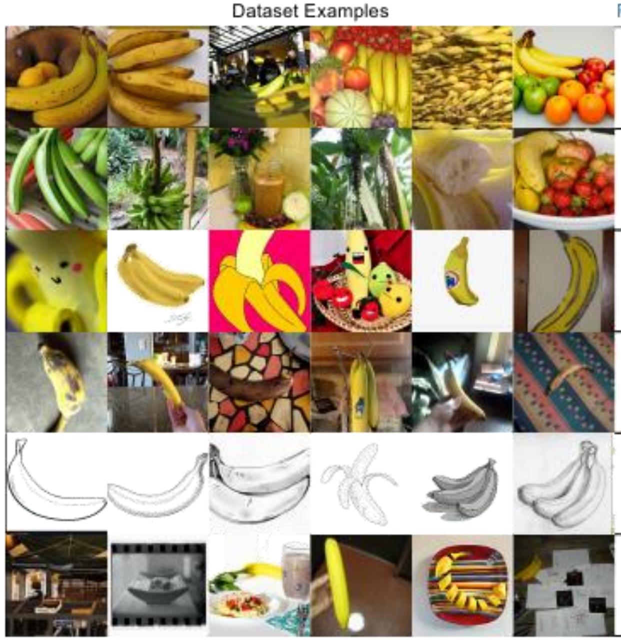
Zero-Shot Image Classification

Example Use-Cases of CLIP model (3/3)



Predefined prompts define the possible classes

Smallest distance in the embedding space gives the prediction



	ImageNet ResNet101	Zero-Shot CLIP	Δ Score
ImageNet	76.2	76.2	0%
ImageNetV2	64.3	70.1	+5.8%
ImageNet-R	37.7	88.9	+51.2%
ObjectNet	32.6	72.3	+39.7%
ImageNet Sketch	25.2	60.2	+35.0%
ImageNet-A	2.7	77.1	+74.4%

Beats pre-trained models and generalizes better

Images from the CLIP Paper
https://arxiv.org/abs/2103.00020

Part 02

Application to Metadata Extraction

Example from: Karkkainen, K., & Joo, J. (2021). FairFace: Face Attribute Dataset […]. IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 1548-1558).

Given values are predictions only

What is Metadata?

Introduction to Metadata Extraction

Definition of Metadata

Metadata is "data that provides information about other data" *(Merriam-Webster dictionary)*

Typically, such information is seen as, e.g.,

- Time-Stamps, Author information, Keywords

Here, we use it in a broader sense:

- Structural information about a given datum
- Especially, information descriptive of the datums content

Examples for Metadata

Race: Asian
Gender: Female
Age: 30-39

Race: Asian
Gender: Female
Age: 30-39

Race: Black
Gender: Male
Age: 3-9

Race: White
Gender: Male
Age: 60-69

Why Metadata?

Introduction to Metadata Extraction

Example from: Karkkainen, K., & Joo, J. (2021). FairFace: Face Attribute Dataset [...]. IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 1548-1558).

Given values are predictions only

Use cases for Metadata

- Structuring retrieval of data
  - Especially for otherwise unstructured data (e.g. images)
- Advanced Labelling
  - Depending on application, other attributes might be of interest
  - Data description / specification
- **Analysis of data-space**
  - Testing of data coverage and performance

Example: Fairness Investigations

- Are specific groups less reflected?
- Are groups discriminated against?

Examples for Metadata

Race: Asian
Gender: Female
Age: 30-39

Race: Asian
Gender: Female
Age: 30-39

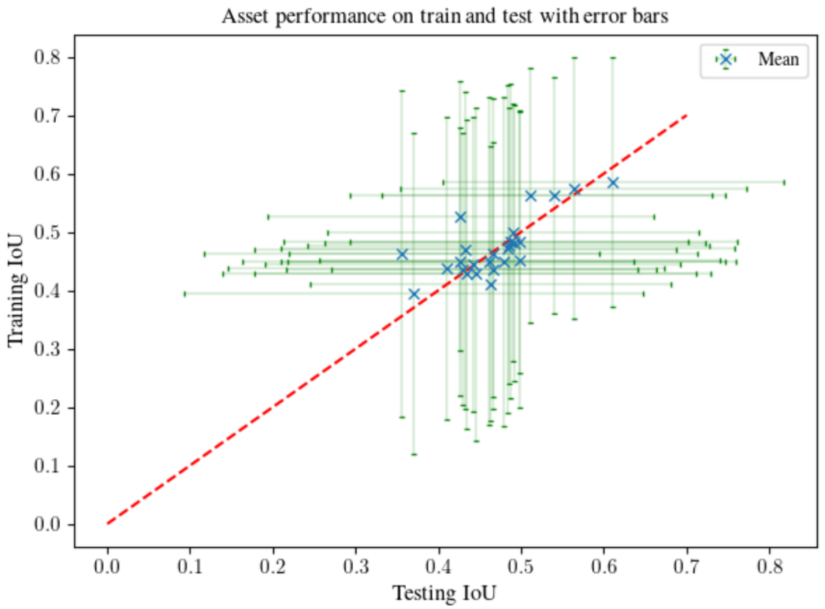
Race: Black
Gender: Male
Age: 3-9

Race: White
Gender: Male
Age: 60-69

ODD based Testing using Metadata

Safety Concerns beyond Fairness

- ODD = Operational Design Domain
  - Set/Domain of inputs on which the AI system is supposed to work
- But does it work on all sub-domains?
  - Fairness: ethnicity, gender, …
  - Outside: more general (brands, situations, …)



Example: Semantic Testing on Carla

- Carla = synthetic image generator from AD domain
- Can provide metadata (with add-on, see paper)

- "Pedestrian assets" show systematically different performance
- → Potential systematic risk (not statistically hedged)

S. Gannamaneni, S. Houben, and M. Akila. "Semantic concept testing in autonomous driving by extraction of object-level annotations from Carla." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021.

Usage of ODD Descriptions / Metadata

From unstructured to structured data

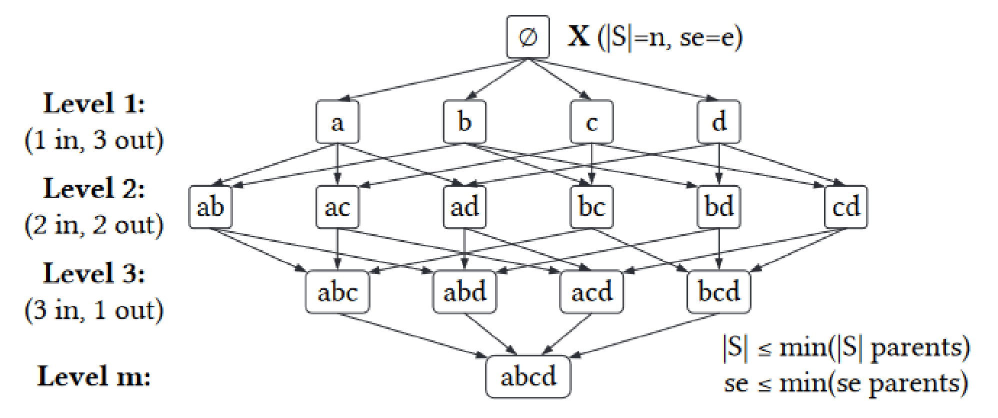
General

- Testing of unstructured data challenging / open problem
  - Verifying / Checking Specification, Weakspots
- Contrast: structured data known / treated since long time
- → Multiple Algorithms available
  - K-point coverage / density estimation
  - Search algorithms within the space

Example: Sliceline (r.h.s.)

- Analysis of single elements
- Sub-Division into further slices (more attributes)
- Recurse as long as error signal "exists"
  - *Scoring function to the right*

$$sc = \alpha\left(\frac{\overline{se}}{\overline{e}} - 1\right) - (1-\alpha)\left(\frac{n}{|S|} - 1\right)$$



Sagadeeva, S., & Boehm, M. (2021, June). Sliceline: Fast, linear-algebra-based slice finding for ml model debugging. In Proceedings of the 2021 International Conference on Management of Data (pp. 2290-2299).

Using CLIP to label meta-data

Challenges and Example

Image samples
CelebA Dataset

Metadata often not available "for free"

Using CLIP

- Metadata extraction can be seen as form of
  - Captioning of images
  - Zero/Few-Shot classification w.r.t. multiple classes
- Challenges for CLIP
  - Extraction of non-dominant features
  - Training domain of CLIP
  - Prompt engineering (e.g. negations)

Important: shown results are zero-shot

CLIP performance

Semantics	Attribute	Counts	Naive				Ensemble			
			Accuracy	Precision	Recall	F1Score	Accuracy	Precision	Recall	F1Score
Age	Young	156734	0.78	0.80	0.95	0.87	0.86	0.91	0.91	0.91
	Not-young	45865		0.53	0.21	0.30		0.70	0.70	0.70
Gender	Male	84434	0.95	0.95	0.91	0.93	0.99	0.99	0.98	0.99
	Not-male	118165		0.94	0.97	0.95		0.99	0.99	0.99
Skin-color	Pale	8701	0.84	0.11	0.41	0.18	0.44	0.07	0.92	0.12
	Not-Pale	193898		0.97	0.86	0.91		0.99	0.42	0.59
Hair-color	Black	47323	0.77	0.93	0.64	0.76	0.78	0.94	0.65	0.77
	Blond	28252		0.81	0.93	0.87		0.83	0.93	0.87
	Gray	7928		0.76	0.69	0.72		0.81	0.65	0.72
	Brown	39167		0.65	0.83	0.73		0.64	0.86	0.73
Misc.	Eyeglasses	13193	0.97	0.86	0.55	0.67	0.99	0.94	0.90	0.92
	No eyeglasses	189406		0.97	0.99	0.98		0.99	1.00	0.99
	Hat	9818	0.92	0.35	0.73	0.47	0.96	0.56	0.74	0.64
	No Hat	192781		0.99	0.93	0.96		0.99	0.97	0.98
	Bald	4547	0.87	0.07	0.39	0.11	0.93	0.19	0.60	0.29
	Not Bald	198052		0.98	0.88	0.93		0.99	0.94	0.96
	Goatee	12716	0.53	0.05	0.37	0.09	0.90	0.26	0.30	0.28
	No Goatee	189883		0.93	0.54	0.68		0.95	0.94	0.95
	Beard	33441	0.81	0.23	0.06	0.10	0.84	0.69	0.10	0.18
	No Beard	169158		0.84	0.96	0.89		0.85	0.99	0.91
	Smiling	97669	0.86	0.86	0.84	0.85	0.87	0.88	0.86	0.87
	Not-smiling	104930		0.86	0.87	0.86		0.87	0.89	0.88

S. Gannamaneni et al., (2023). Investigating CLIP Performance for Meta-Data Generation in AD Datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 3839-3849).

CelebA: Ziwei L. Ziwei et al., (2015), Deep Learning Face Attributes in the Wild, In *Proceedings of the IEEE ICCV* (pp. 3730-3738)

Classification with CLIP

A second look on the mechnics

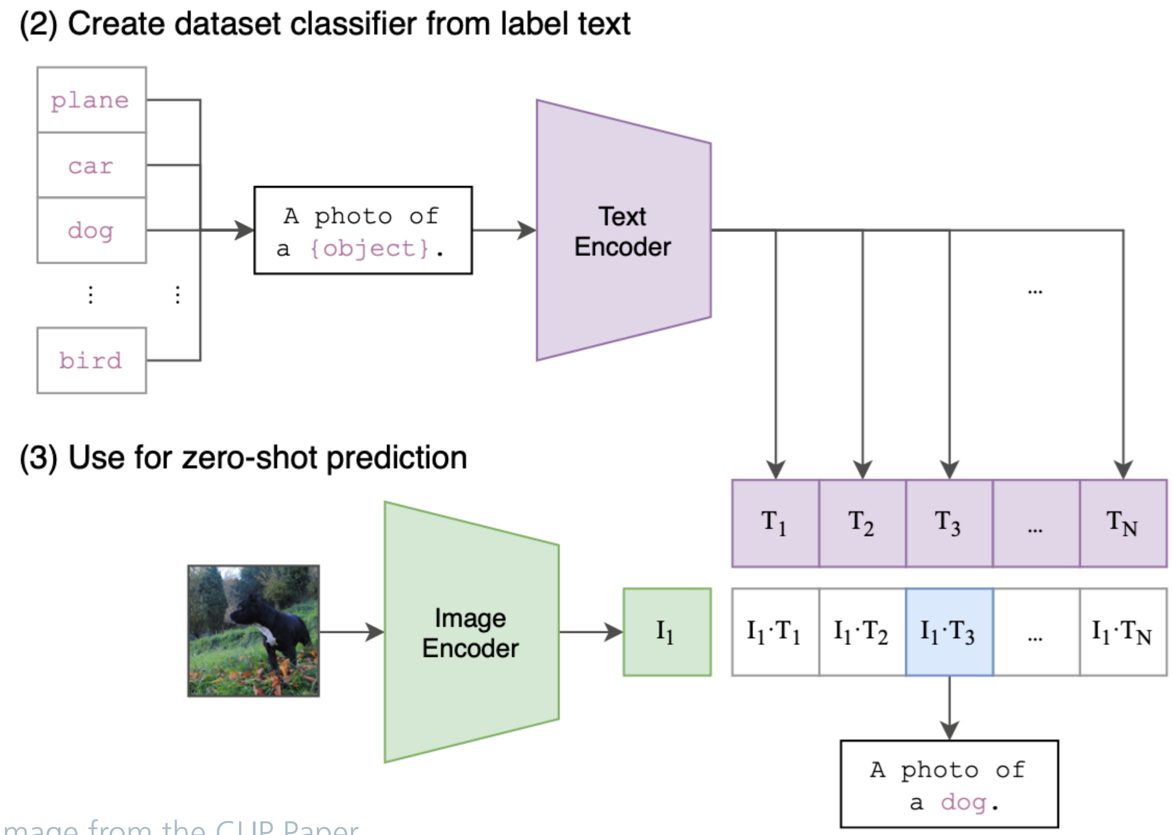
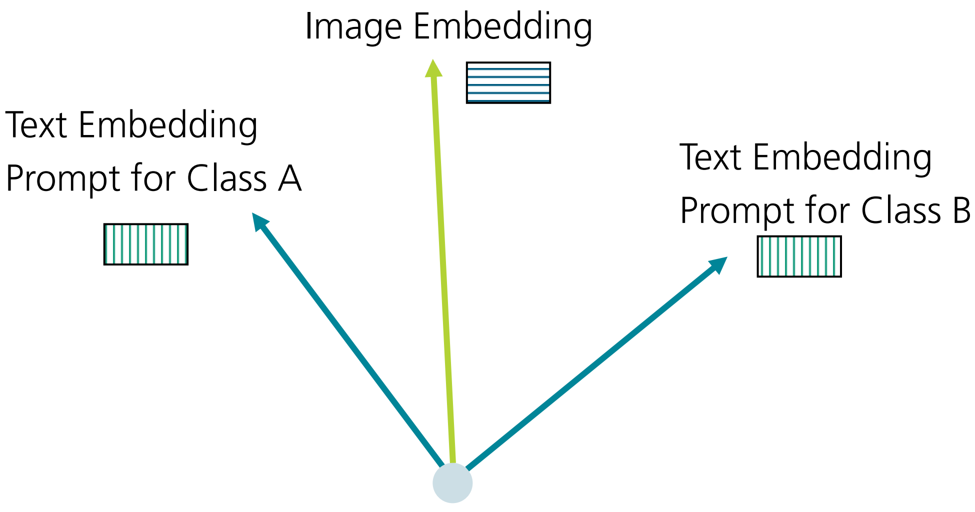


Image from the CLIP Paper
https://arxiv.org/abs/2103.00020



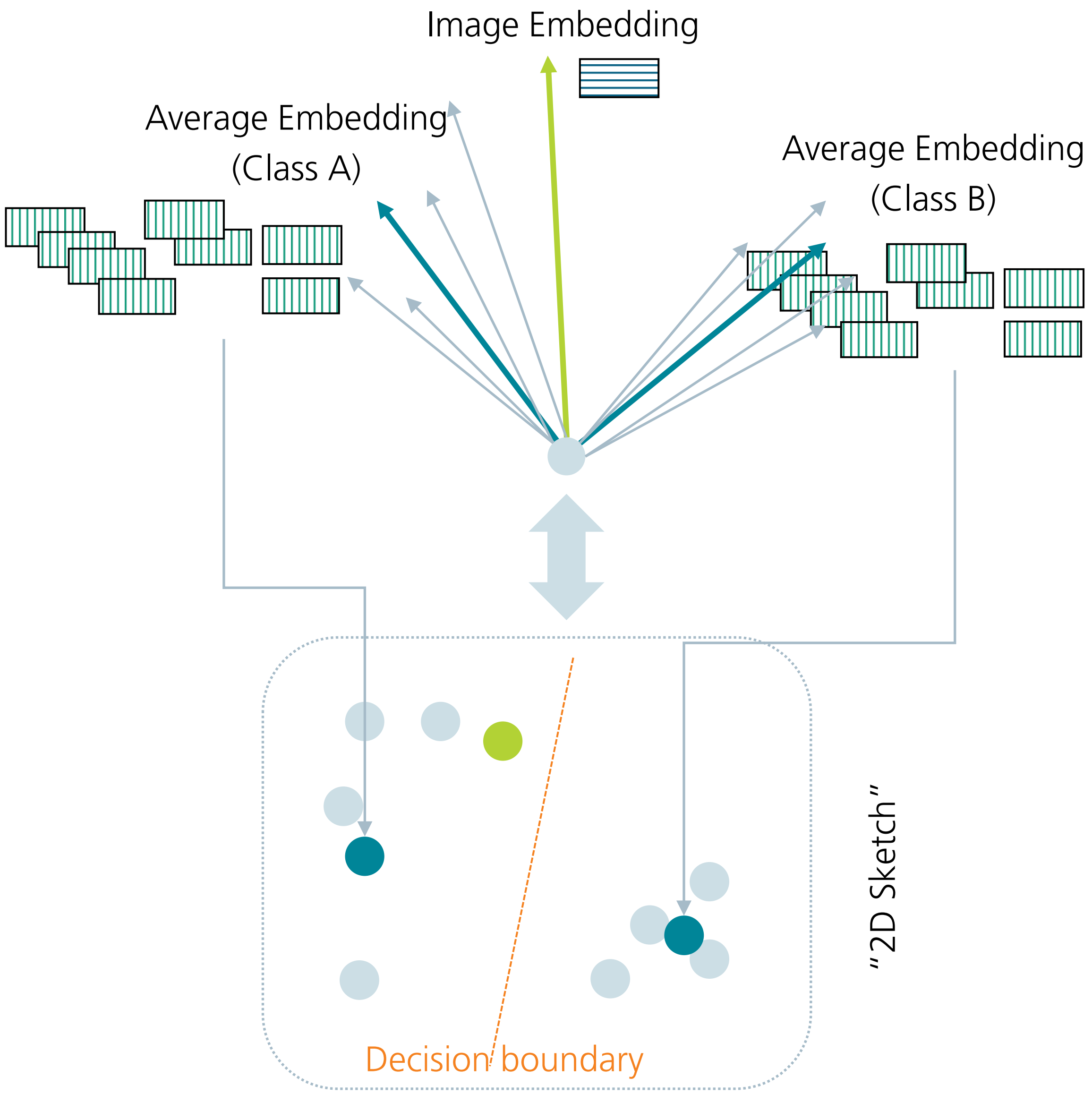
- Obtain list of distances (e.g. cosine distance)
- Select closest match or build softmax classifier

- But, different prompts have different embeddings, e.g.,
  - "A photo of a woman"
  - "A photo of a lady"
- Different embeddings give different results

Classification with CLIP

Using ensembles of prompts

- Often multiple prompts match parts of data
- → useful to include "all"/multiple prompts
- Can be seen as compensating undesired artifacts
- E.g. for a dog vs cat classifier one might include
  - "big dog" / "hairy dogs" / "dog fetching a stick" / …
  - In parts, prompt engineering task
- Technically, multiple embeddings can be averaged over
  - Where average is then equivalent to average over decisions
  - *Detail:* Average of linear distance and average of embeddings commutes
- Remark: If embeddings (within classes) have strong spread non-linear averaging might be beneficial
  - See, e.g., S. Gannamaneni et al., (2023). Investigating CLIP Performance for Meta-Data Generation in AD Datasets

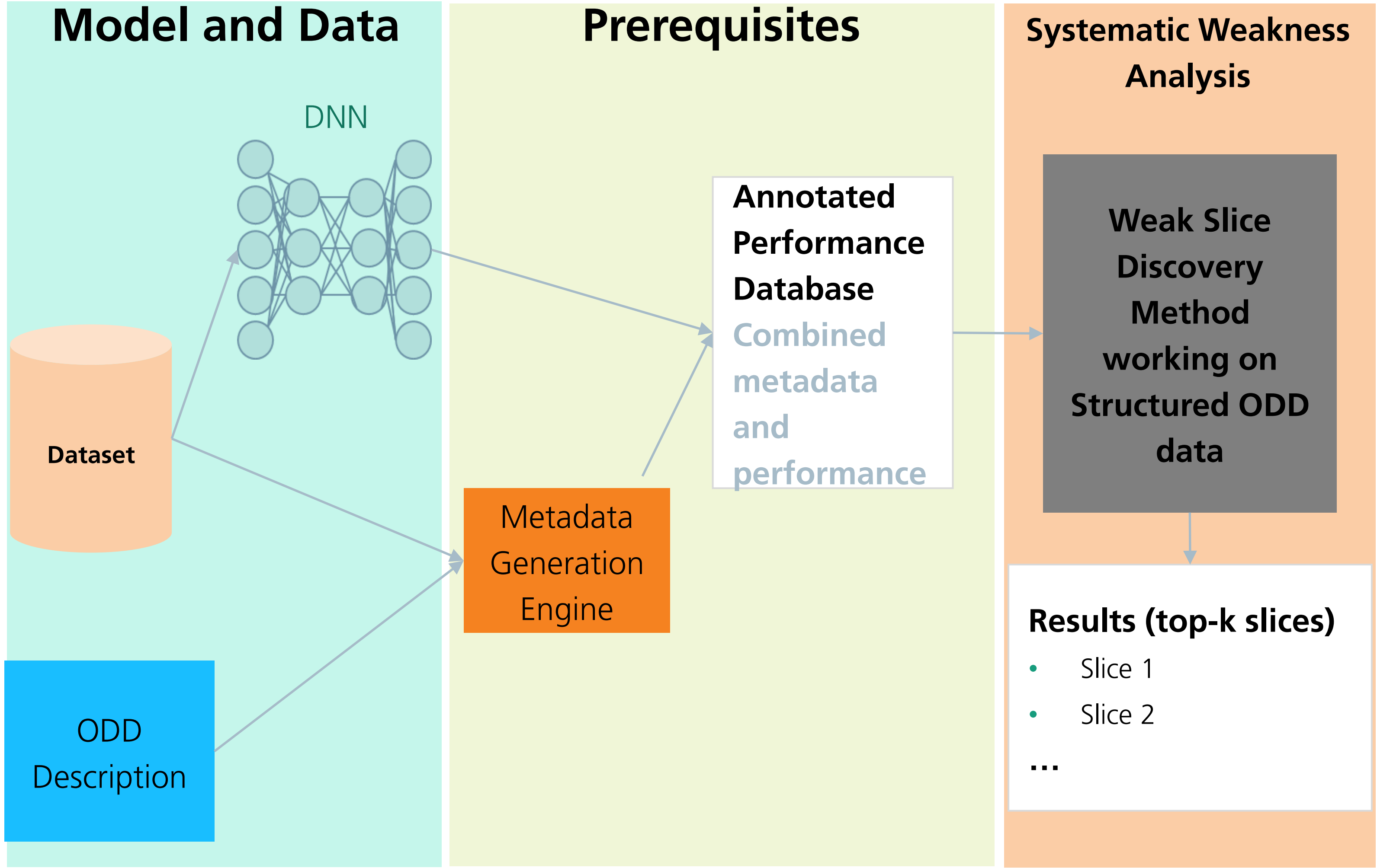


  Public

Fraunhofer IAIS

Autonomous discovery of weaknesses given model and data

Automated weak slice discovery based on automated ODD labelling



  Public

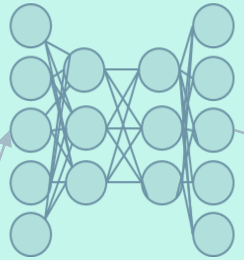
Fraunhofer IAIS

Autonomous discovery of weaknesses given model and data

Automated weak slice discovery based on automated ODD labelling

Model and Data

DNN



Dataset

ODD Description

Prerequisites

Annotated Performance Database

Combined metadata and performance

Metadata Generation Engine

Systematic Weakness Analysis

Weak Slice Discovery Method working on Structured ODD data

Results (top-k slices)

- Slice 1
- Slice 2

…

Example Weak Slice:
Gender: Female & Age: child

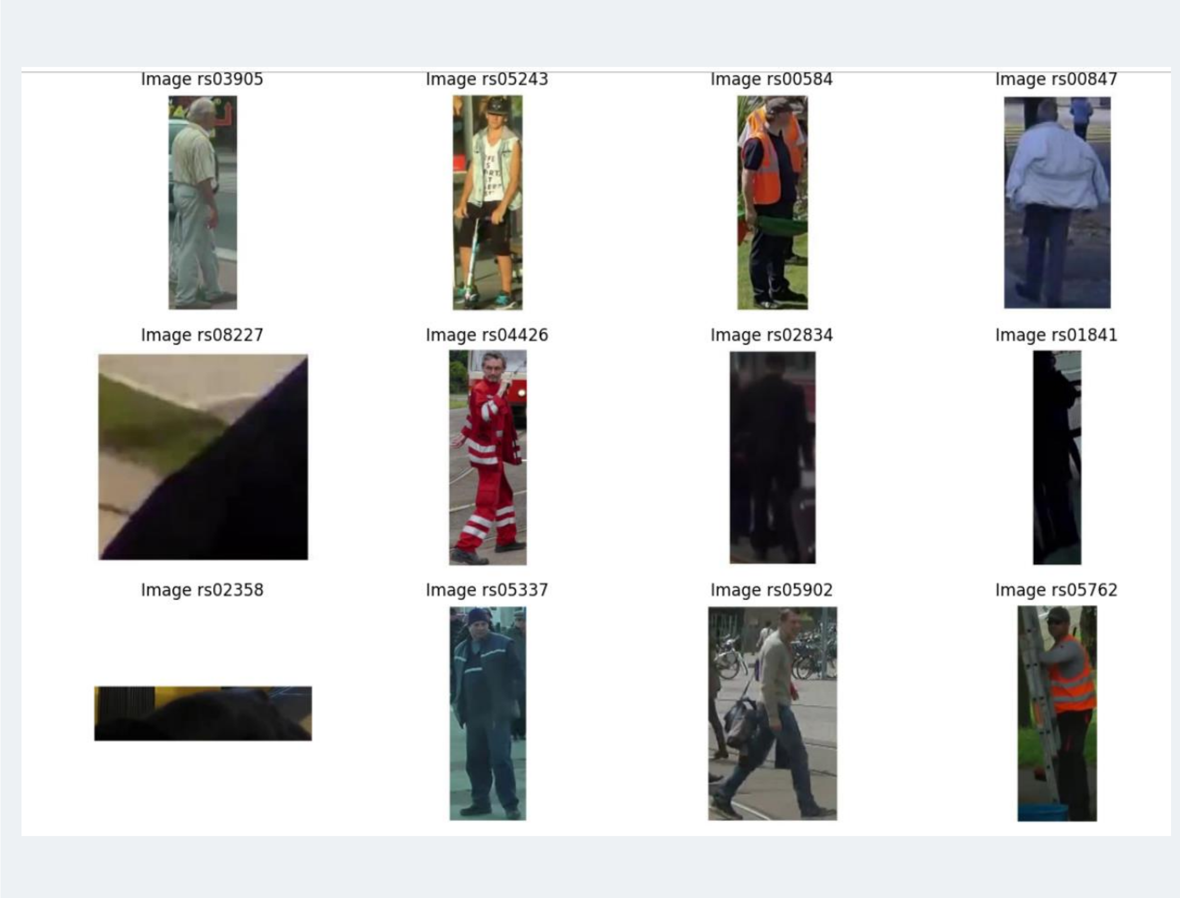
Public

Fraunhofer IAIS

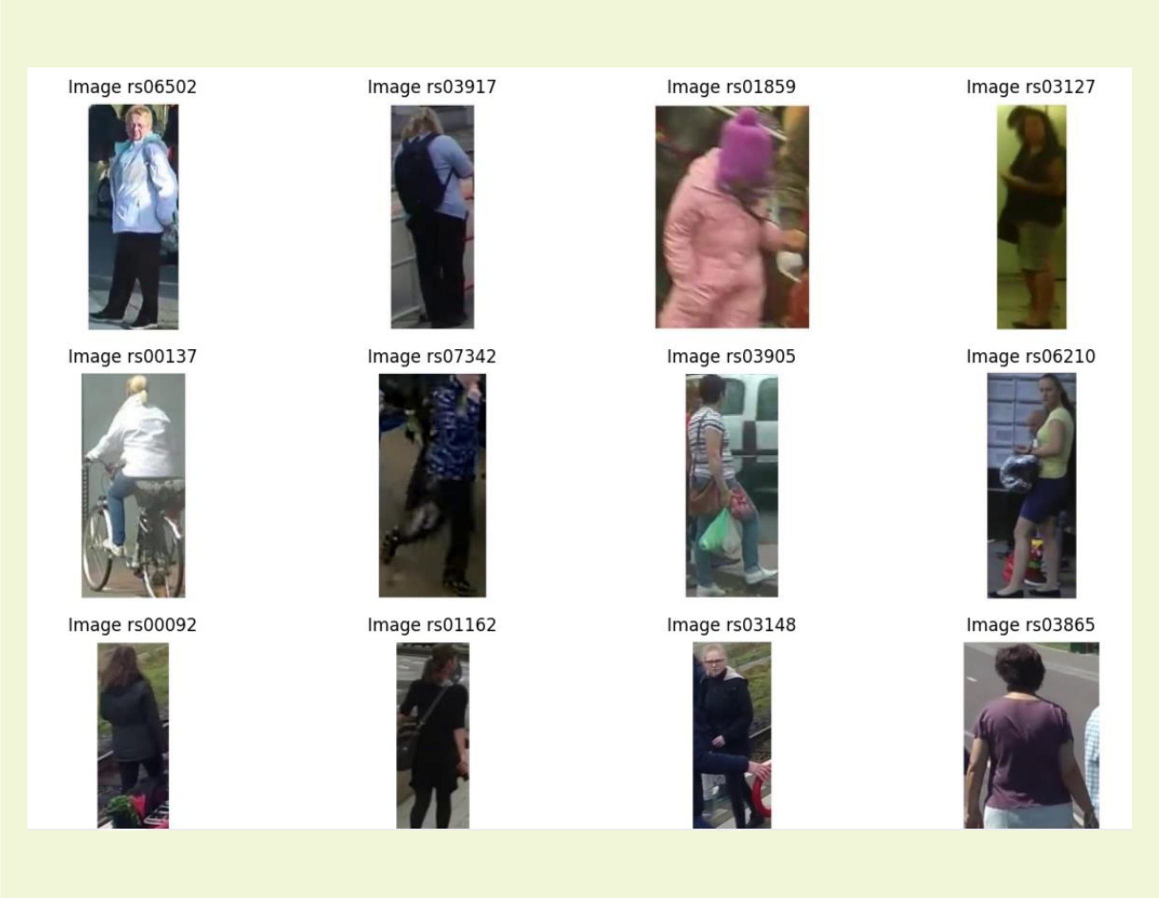
Examples for automatically annotated categories

Applied to data extracted from the RailSem19 dataset

Gender "Male"



Gender "Female"



  Public

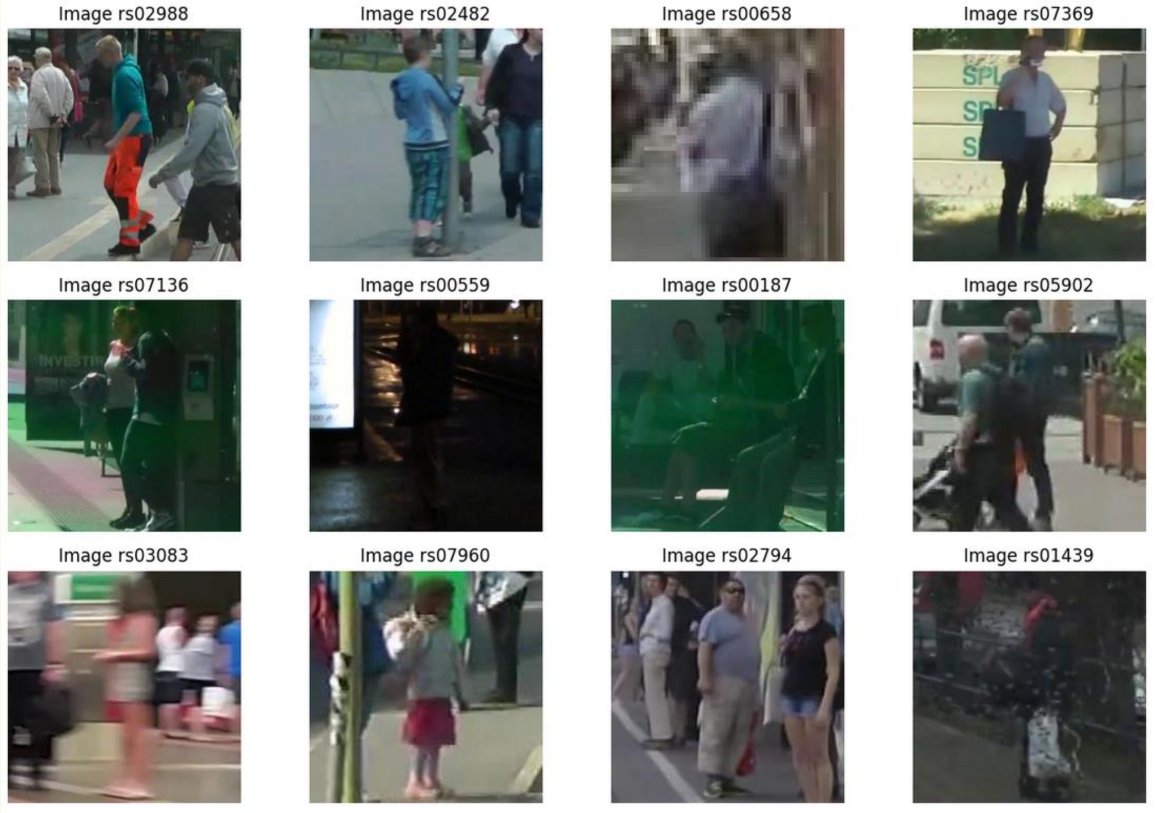
Fraunhofer IAIS

Examples for automatically annotated categories

Applied to data extracted from the RailSem19 dataset

Contains “Railway-worker”

Is “blurry”

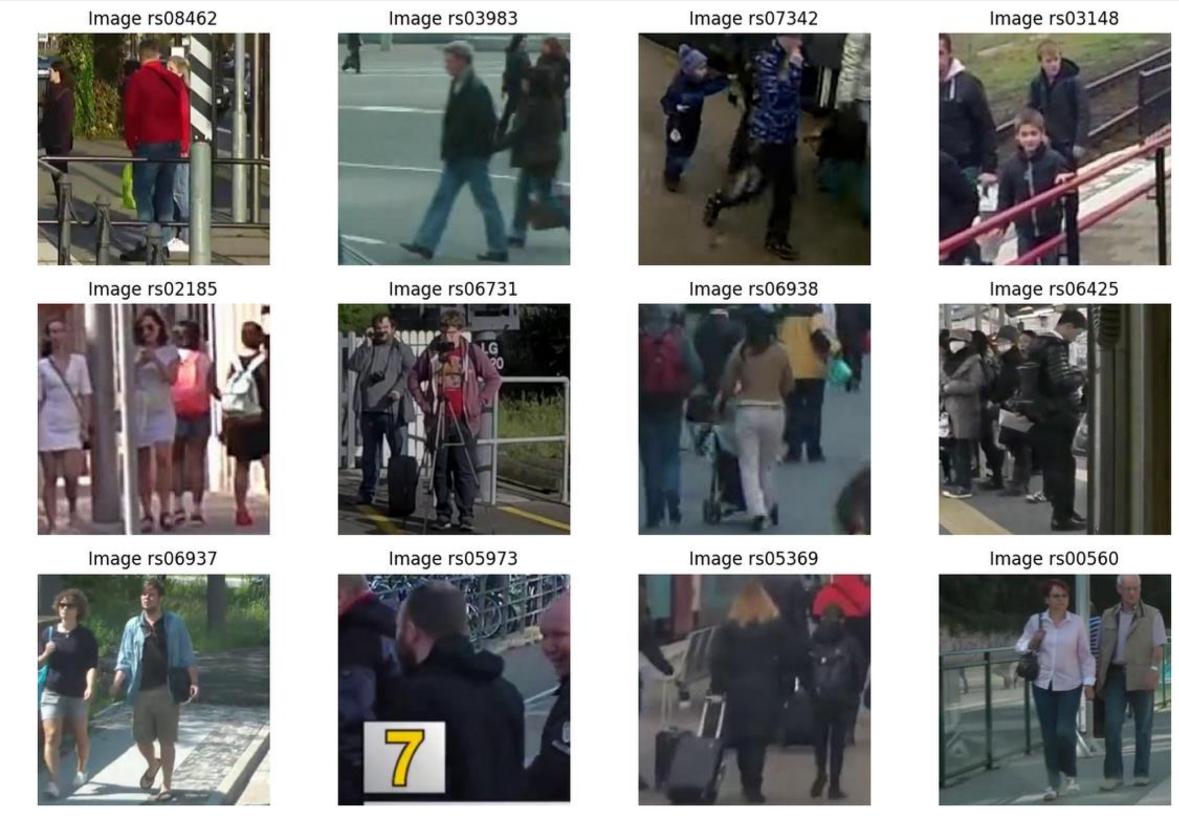


  Public

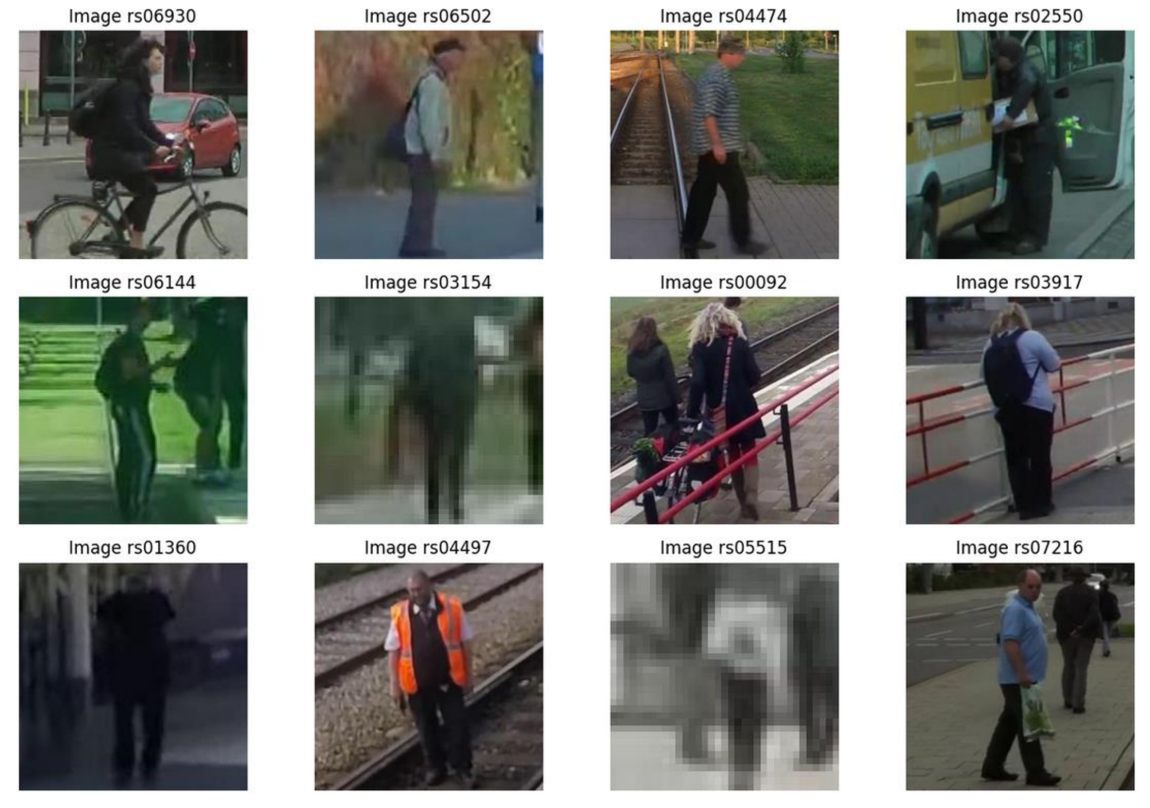
Examples for automatically annotated categories

Applied to data extracted from the RailSem19 dataset

Within Group



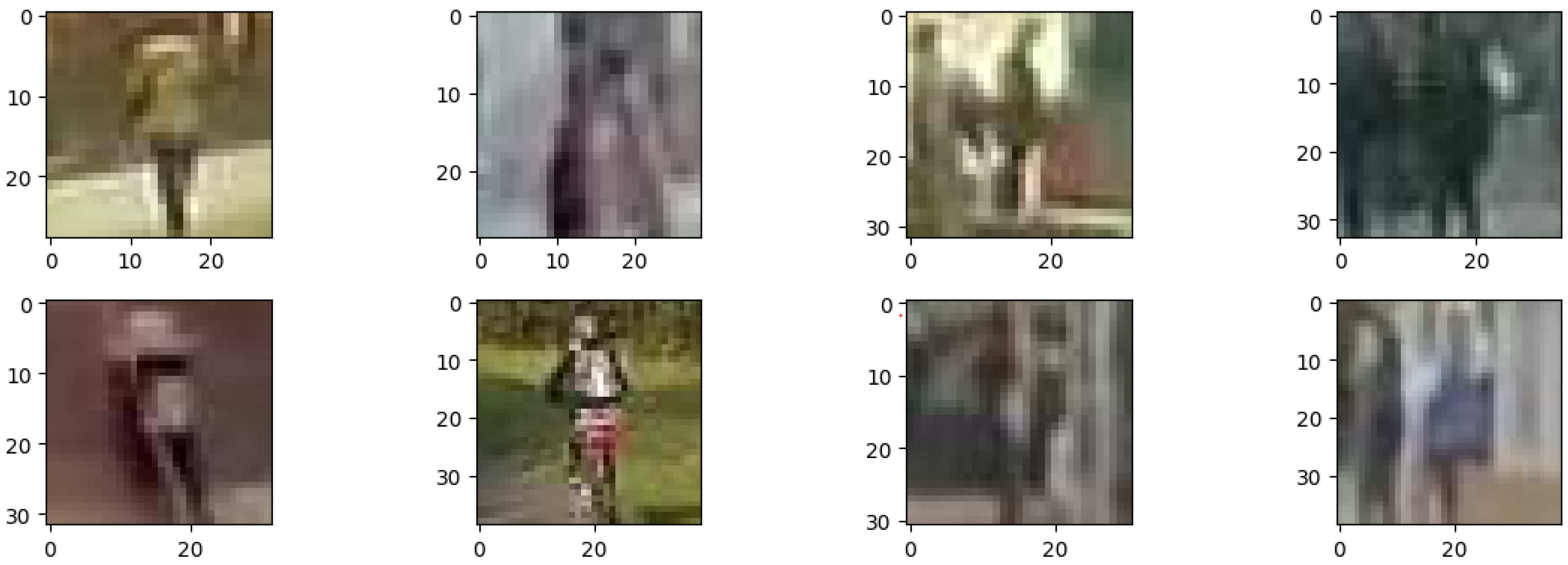
Outside of Group



Observed Potential Weakness in SUT

Preliminary results on the evaluation of an internal detector

Slice description: Gender = Female & Age = Child & size = (127.0, 653.0]



Fraunhofer IAIS

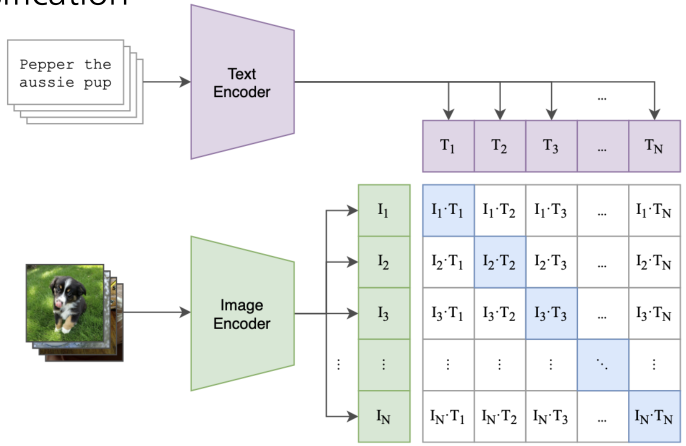
Part 03

Summary

Summary

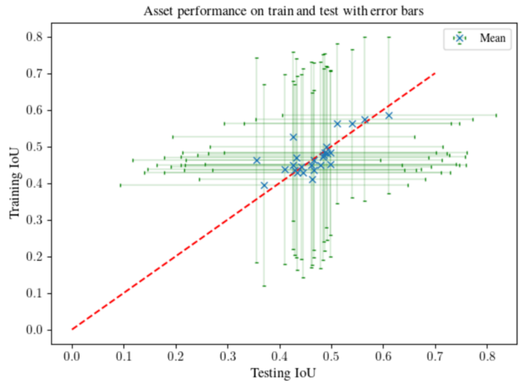
Clip as Foundation Model

- Matching of Captions and Images
- Capability to “understand” image content
- Applications to
  - Image Generation
  - Image Retrieval
  - Zero-Shot Classification



Metadata Extraction (and Semantic Testing)

- Metadata extraction as (zero-shot) multi-dim. classification
- Relevance of metadata for
  - Semantic Testing
  - Fairness Investigations



  Public

Kontakt

Dr. Maram Akila
Team KIAZ / Department KD
Tel. +49 2241 14-2208
maram.akila@iais.fraunhofer.de

Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS
Schloss Birlinghoven 1
53757 Sankt Augustin

www.iais.fraunhofer.de

Fraunhofer
IAIS