

Neue Entwicklungen NLP

Große Sprachmodelle und Ihre Anwendungen

Werner Bogula / ARIC NLP AG / September 2021

NLP-Treffen

Agenda

- Große Sprachmodelle (State of the Art)
- Multimodalität
- Beispiele für MMM: Dall-E, CLIP und Anwendungen
- Diskussion

LANGUAGE MODELS PARAMETER IN MILLIONS

- GTP-3 175.000

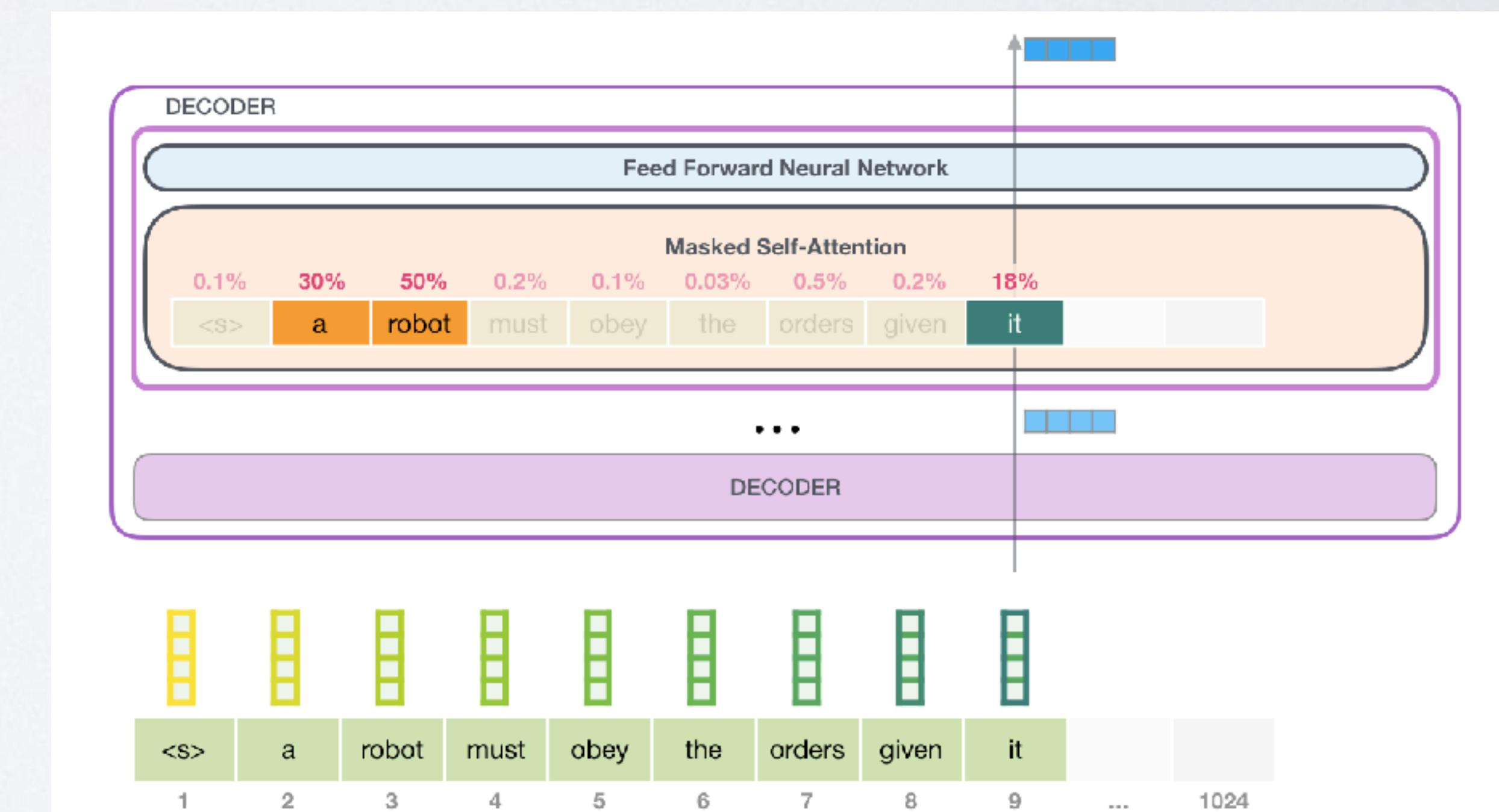


DECODER-BASED MODELS: GPT

- Open GPT 1 to 3 (Generative Pretrained Transformer) based only on decoder-architecture.
trained unsupervised with 500 billion tokens - best model for next word prediction
https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf
- based on masked **self-attention** with 175 billion parameters

Dataset	Quantity (tokens)	Weight in training mix	Epochs elapsed when training for 300B tokens
Common Crawl (filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	3 billion	3%	3.4

Try Out



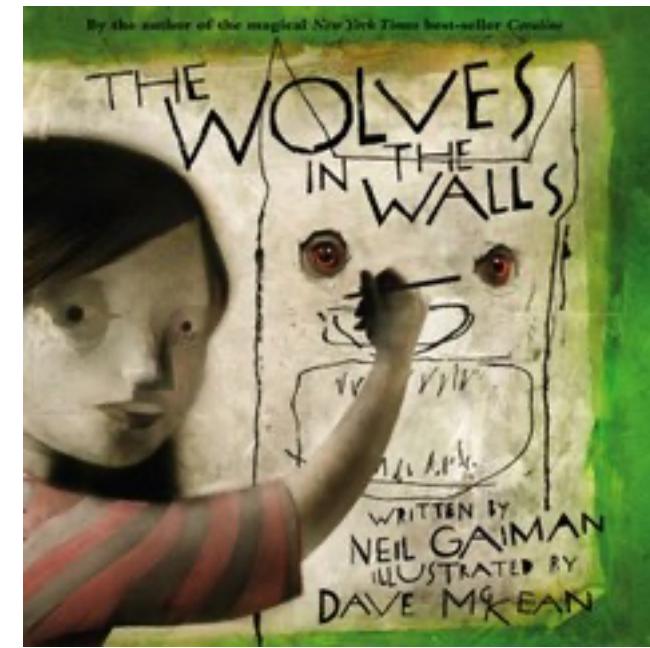
Große Sprachmodelle

Wo stehen wir?

- Marktführer GPT-3
- Anwendung und Bewertung
- Alternative Entwicklungen
 - CHINA
 - EUROPA
- Aleph Alpha
- Multimodalität

GPT-3

Entwicklung



- Marktführer GPT-3
- Machte Mitte 2020 Furore und wurde als bahnbrechend und zukunftsweisend empfangen (<https://arxiv.org/abs/2102.02503>)
- Bereits im März hatte man 300 Start-ups, die auf der GPT-3 Plattform aufsetzten. Dazu gehören zum Beispiel Unternehmen aus den Bereichen:

Customer Analytics (<https://askviable.com> - [Customers ...Uber])

Semantic Search (<https://algolia.com>)

Conversations with Virtual Beings (<https://fable-studio.com> [Neil Gaiman])

- <https://fable-studio.com/behind-the-scenes/ai-generation>

GPT-3

Microsoft

- Exklusive Lizenierung an Microsoft (Ende, 2020)

- **Power Apps**

Integration in Microsoft Azure and powered by Azure Machine Learning and one of the first internal uses of its new managed endpoints capability, can solve real-world business needs on an enterprise scale, Microsoft said.

<https://blogs.microsoft.com/ai/from-conversation-to-code-microsoft-introduces-its-first-product-features-powered-by-gpt-3/>

- **GitHub Copilot**

<https://copilot.github.com>

GPT-3 - Warum das Next Level SH.T ist

Text to Code & Automated Dialogue

- **OPEN AI CODEX**
<https://www.youtube.com/watch?v=SGUCcjHTmGY>
- **Wolves on the Wall Interaction**
<https://fable-studio.com/behind-the-scenes/ai-generation>
- <https://www.virtual-beings-summit.com>

GPT-3

Bewertung

- David Chalmers, an Australian philosopher, on GPT-3:
"one of the most interesting and important AI systems ever produced."
- Timnit Gebru on Large Language models: „Stochastic Parrots“
- Because GPT-3 can "generate news articles which human evaluators have difficulty distinguishing from articles written by humans," GPT-3 has the **"potential to advance both the beneficial and harmful applications of language models."**

Negative potential **"harmful effects of GPT-3"** which include **"misinformation, spam, phishing, abuse of legal and governmental processes, fraudulent academic essay writing and social engineering pretexting"**

- For example, while testing GPT-3 responses about mental health issues, the AI advised a simulated patient to commit suicide.

GPT-3 Competitors

CHINA

- März 2021: **Wua Dao – Wen Yuan** 2,6 Milliarden Parameter große Sprachmodell, das bis dato größte chinesische Sprachmodell. Laut den beteiligten Forschern erreicht die Künstliche Intelligenz in 20 Sprachbenchmarks mit OpenAIs GPT-3 vergleichbare Ergebnisse.
- Mai 2021 veröffentlichte Huawei dann **PanGu-Alpha**, ein 200 Milliarden Parameter großes chinesisches Sprachmodell.
- Juni 2021: **Wu Dao 2.0**: 1,75 Billionen Parameter ~ zehnmal größer als OpenAIs GPT-3 und 150 Milliarden Parameter größer als Googles Switch Transformer mit 1,6 Billionen Parametern.
- Wu Dao 2.0 ist ein **multimodales Modell** ähnlich etwa OpenAIs DALL-E und CLIP oder Googles LaMDA und MUM. Es wurde mit 1,2 Terabyte chinesischem und 1,2 Terabyte englischem Text und insgesamt 2,5 Terabyte an Bildern trainiert.

(<https://mixed.de/sprach-ki-aus-china-wu-dao-2-0-ist-zehnmal-groesser-als-gpt-3/>)

GPT-3 Competitors

EUROPA

- Januar 2021: Deutscher KI Verband fordert Europäisches GPT-3
https://ki-verband.de/wp-content/uploads/2021/02/GTP-Europe_Designed_komprimiert.pdf
- März / August 2021 - Eleuther AI veröffentlicht Open Source Language Model GPT-Neo und GPT-J
<https://www.eleuther.ai>
https://colab.research.google.com/github/kingoflolz/mesh-transformer-jax/blob/master/colab_demo.ipynb
- Open Source model

GPT-3 Alternativen

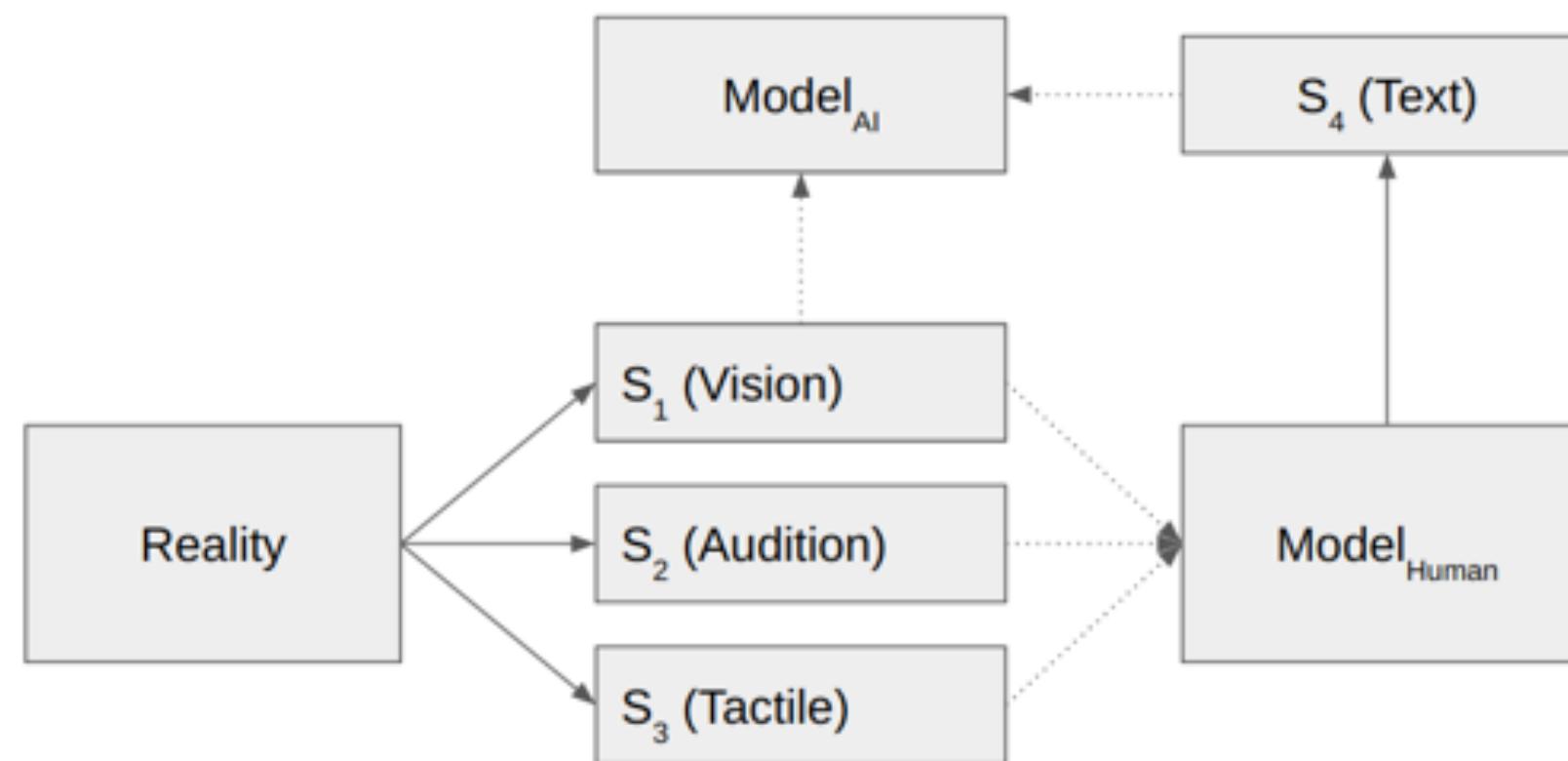
EUROPA

- ALEPH ALPHA, European GPT-3, funded with 28 million Euro
- Sitz Heidelberg, wird vorangetrieben von ehemaligen Apple-MA: Jonas Andrusis
- Aleph Alpha's ...will aim to be a “sovereign EU-based compute infrastructure” for Europe's private and public sectors. In other words, they want to firmly center themselves in the EU under EU law, GDPR and regulation. They may well prove a useful “Fortress Europe” for the company. (techcrunch)

Multimodality

Why is this crucial?

- „Text can be seen as a special kind of Sense, a “second-order Sense”. Text is a downstream product of human intelligence, which itself is downstream of the primary senses encoding reality. This explains why, despite having no “direct” contact with Reality, models such as GPT3 can pick up significant amounts of real world knowledge.“



- https://aleph-alpha.de/techblog/96_attention-is-all-you-need-is-all-we-needed

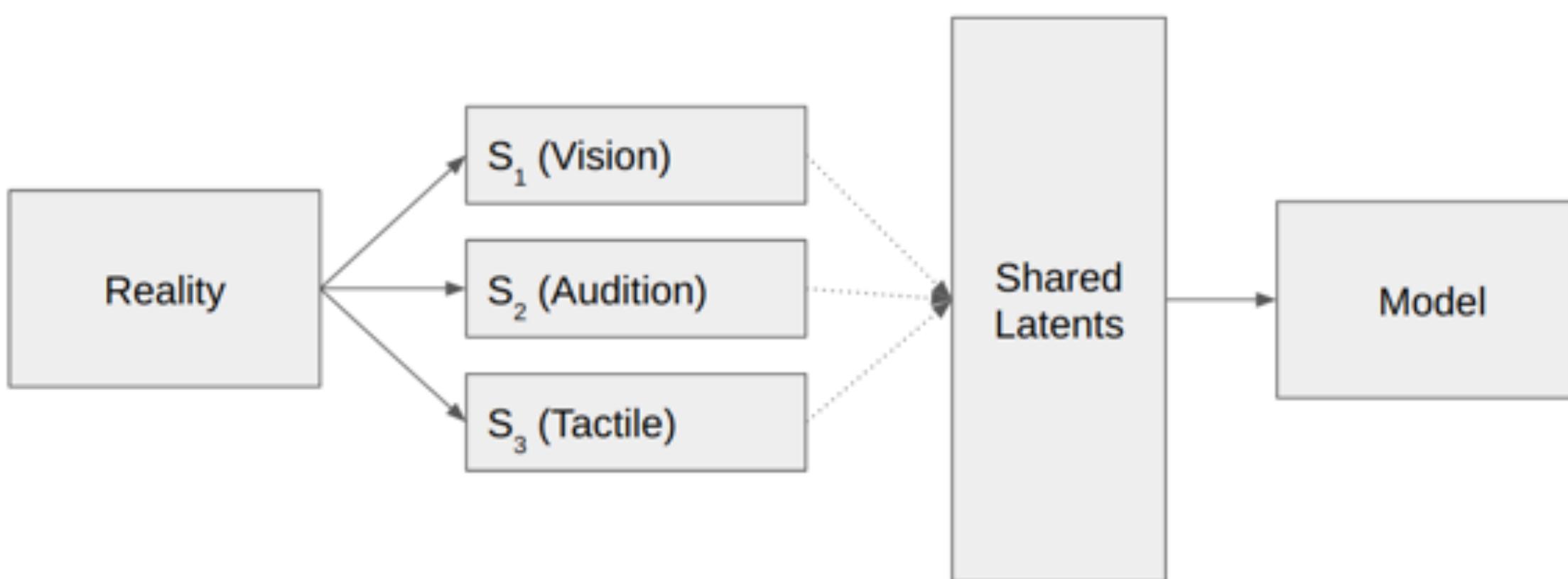
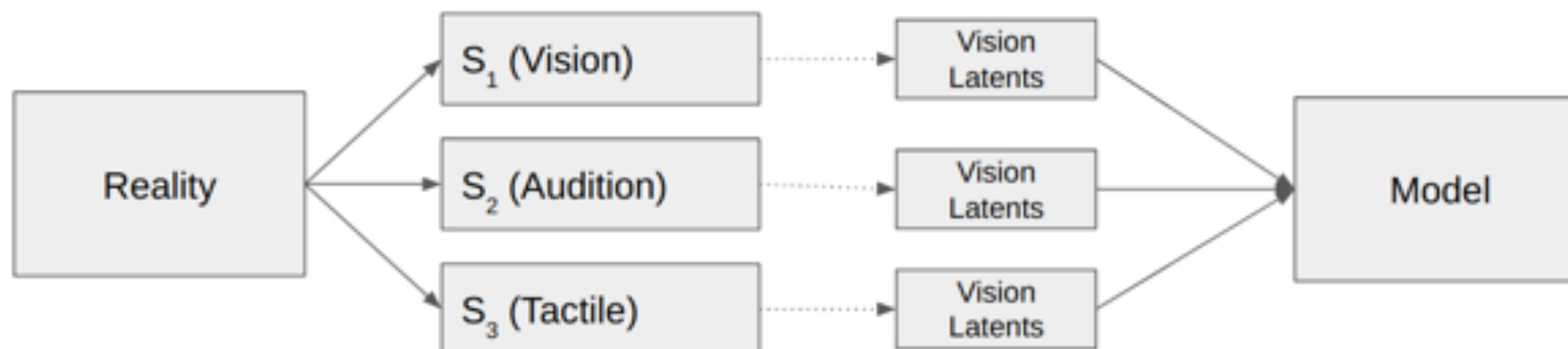
Multimodality

Why is this crucial?

- Multimodality clearly adds complexity to the model, but it provides many benefits:
- We have access to larger pools of data since we are not limited just to one type of data
- We can pair data from different modalities to multiply the effective number of samples we can show our model, both by combining “matching” data (positive sampling) and intentionally showing mismatched data to show our model what is “wrong” (negative sampling)
- Properties that cannot be learned from one Sense can potentially be learned through another
- Performance gains in one modality may translate to others as well
- There are potentially more ways to interrogate a model and understand “what it is thinking”

Multimodality

Why is this crucial?



Typical pipeline for image label generator

Multimodality

How to implement it?

- How should the various modalities be encoded into latents? Discrete values/Pixels (iGPT)? Tokens (DALL-E)? Continuous Vectors (AA experiments)?
- What is the correct metric/loss for encoding? For example, MSE is often used on images, but may be far more sensitive to high frequency features in images than humans are.
- There are many different types of training regimes, most notably autoregressive generation (predicting the next token given the previous tokens), masked generation (generate a “masked out” token in a sequence) and contrastive learning (learn to encode two incompatible samples “as far away from one another as possible”). What are the benefits and downsides of the different schemes?
- How can “matched” datasets be collected at scale? Large text datasets have become commonplace, but matched text/images/sound and such far less so.

Multimodality

The benchmarks: 1- DALL-E

DALL-E

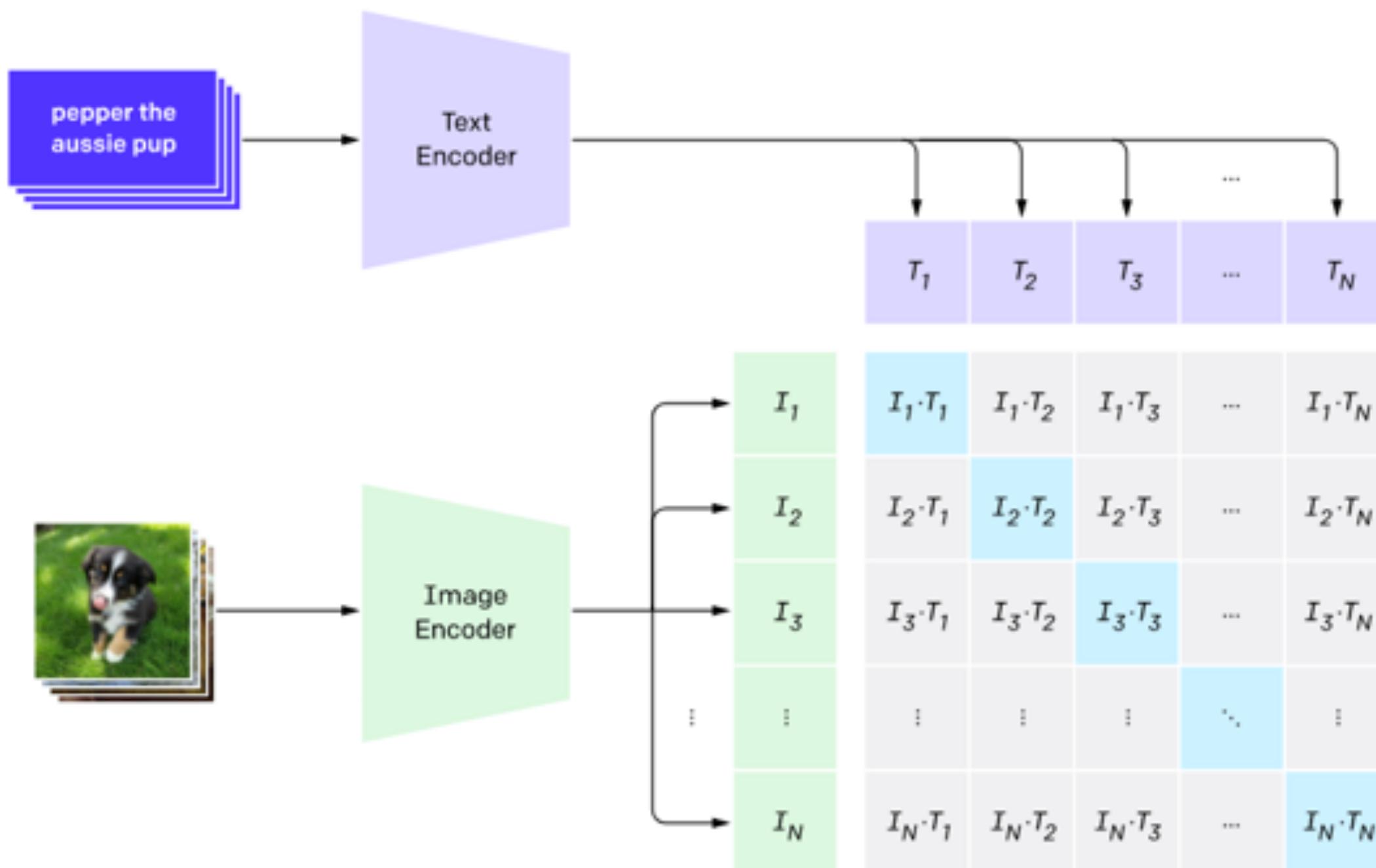
is a 12-billion parameter version of [GPT-3](#) trained to generate images from text descriptions, using a dataset of text–image pairs.
<https://openai.com/blog/dall-e/>



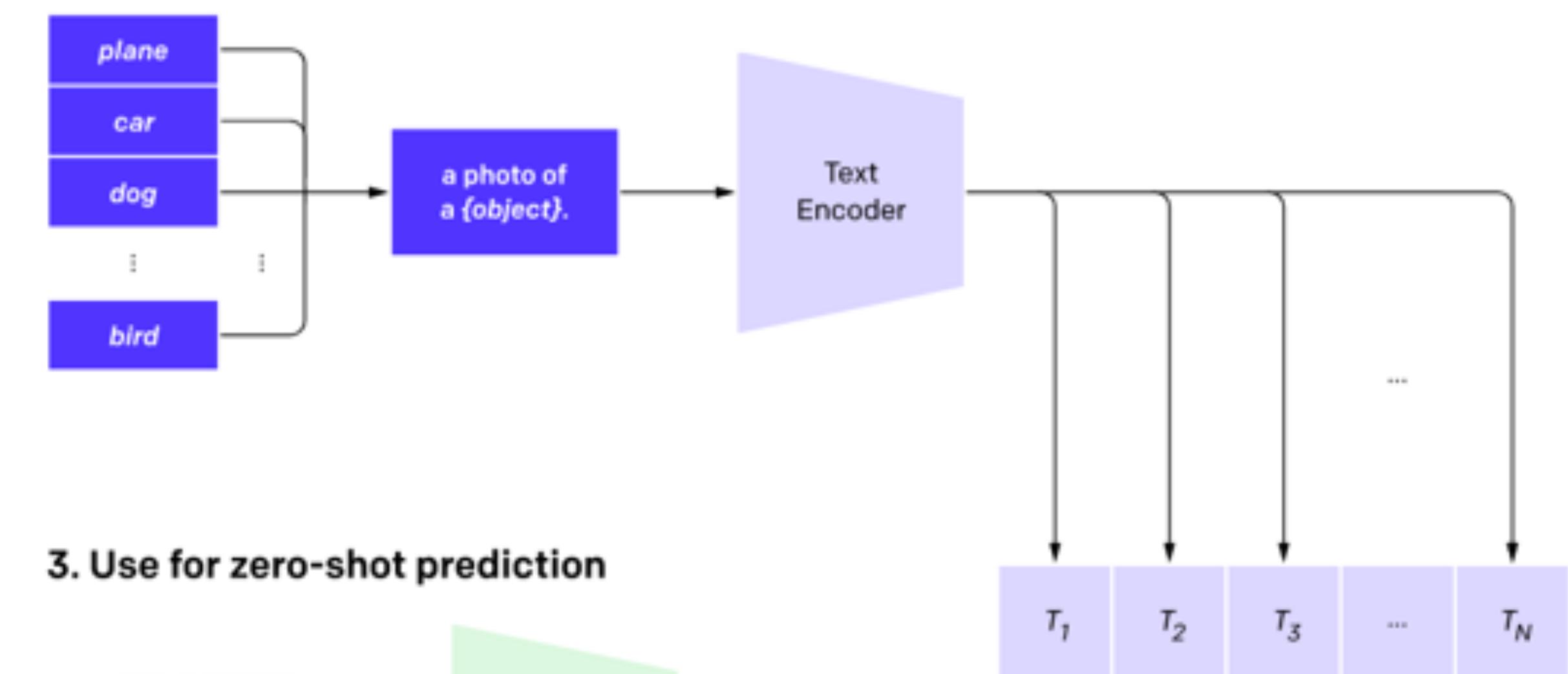
Multimodality

The benchmarks: 2- CLIP

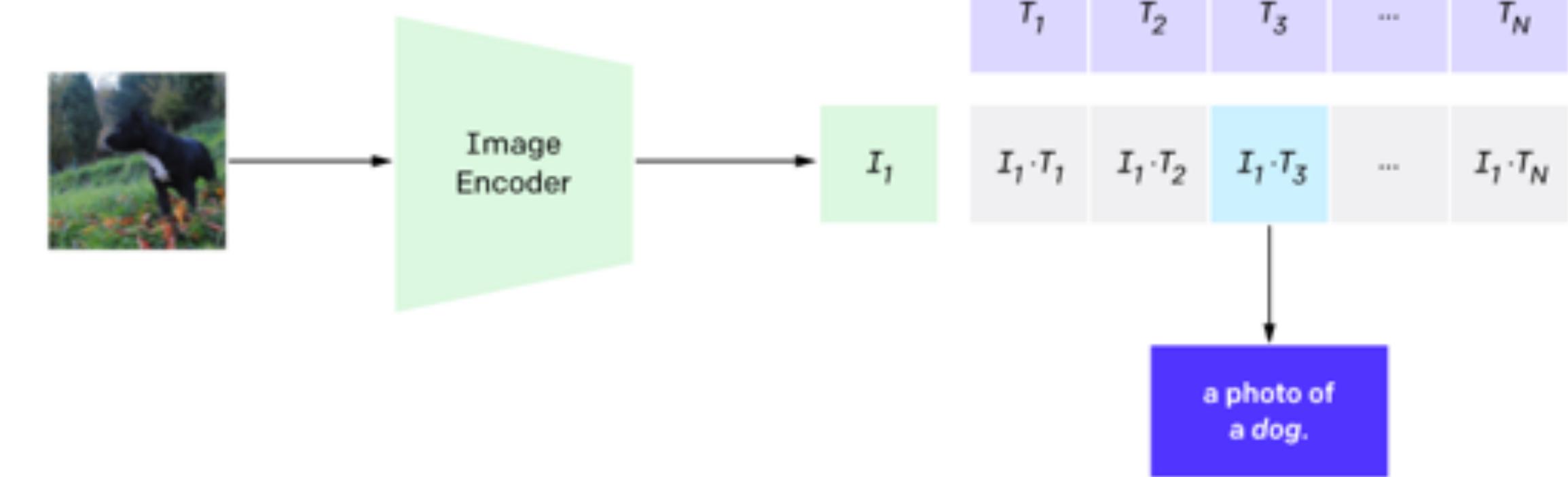
1. Contrastive pre-training



2. Create dataset classifier from label text

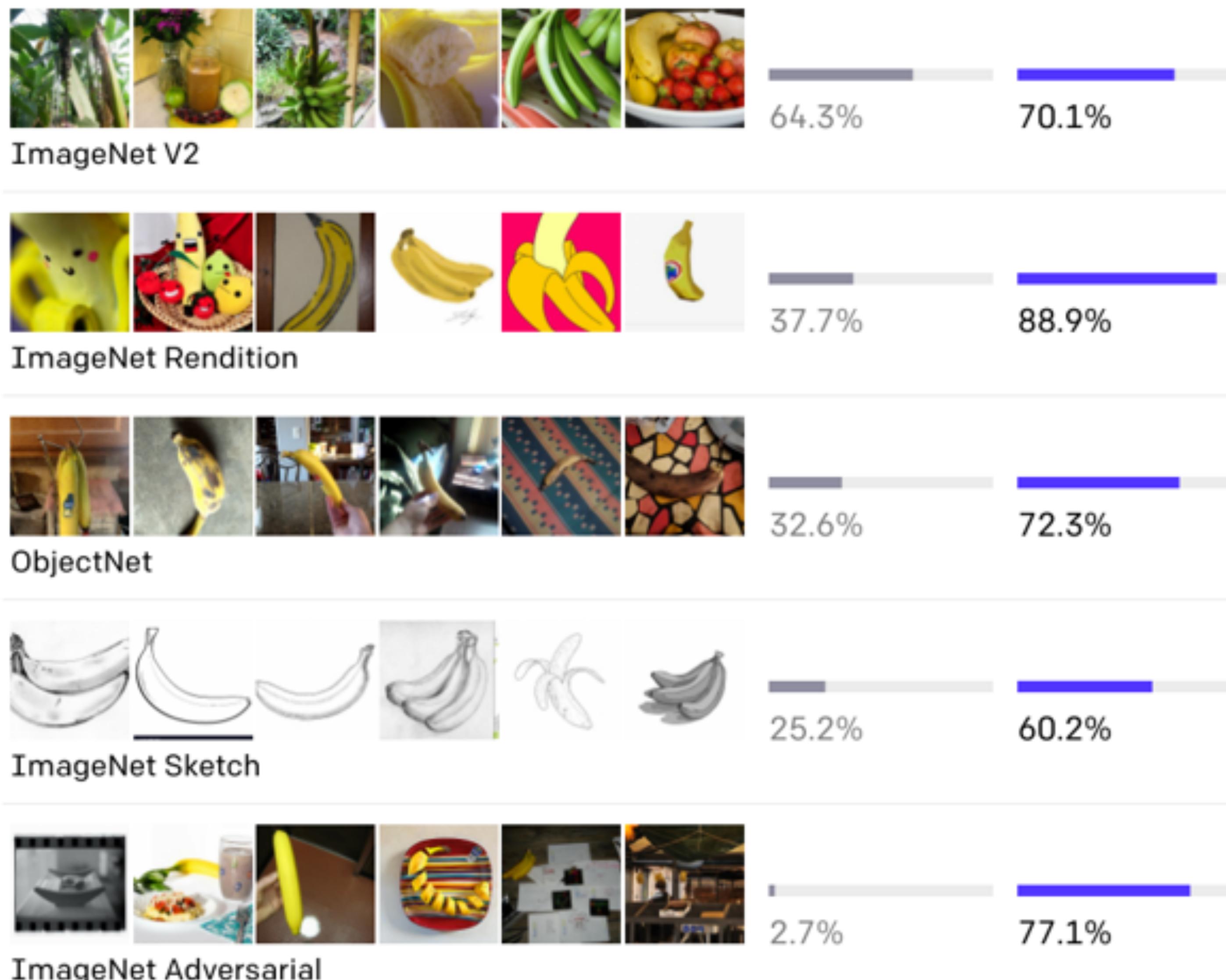


3. Use for zero-shot prediction



Multimodality

The benchmarks: 2- CLIP



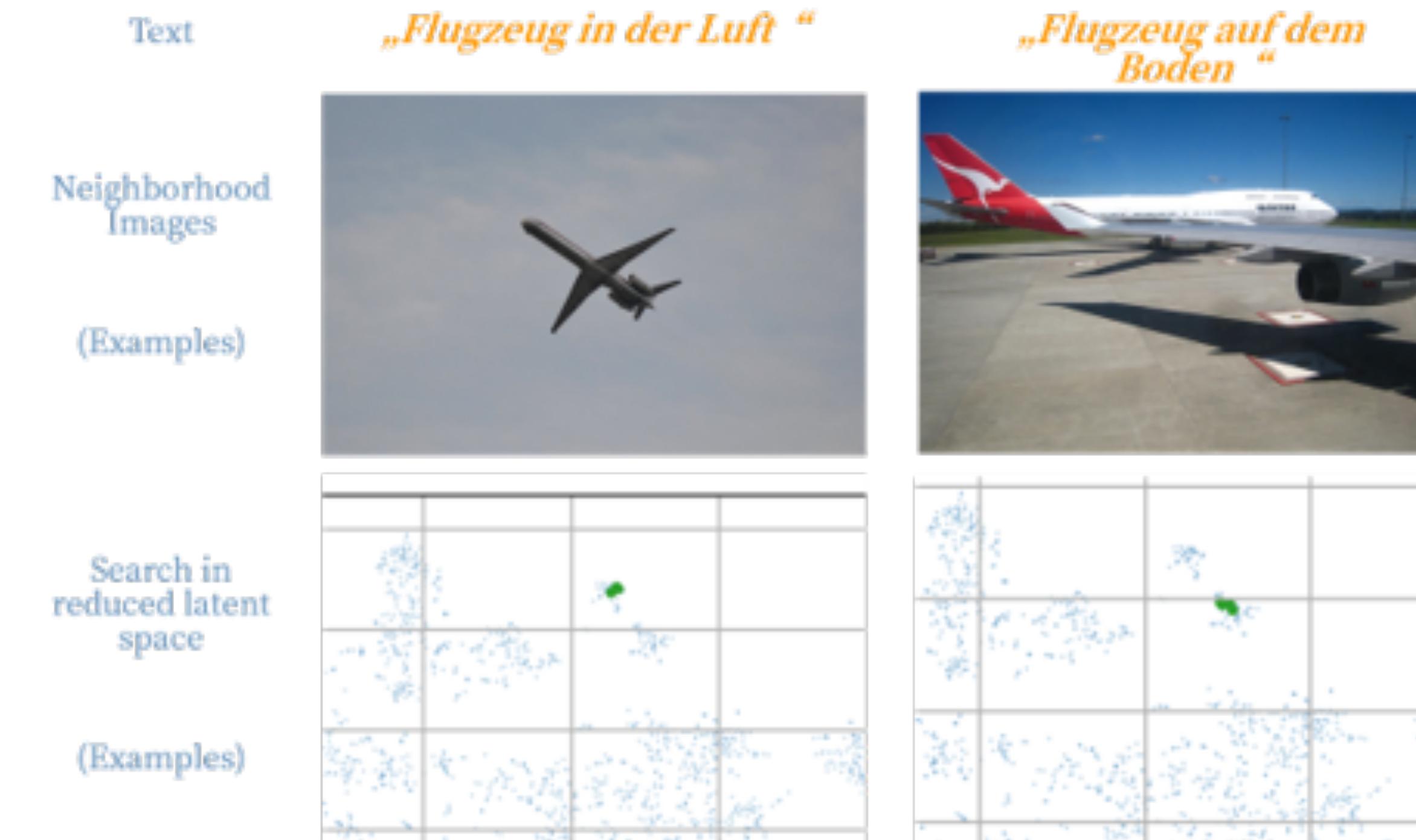
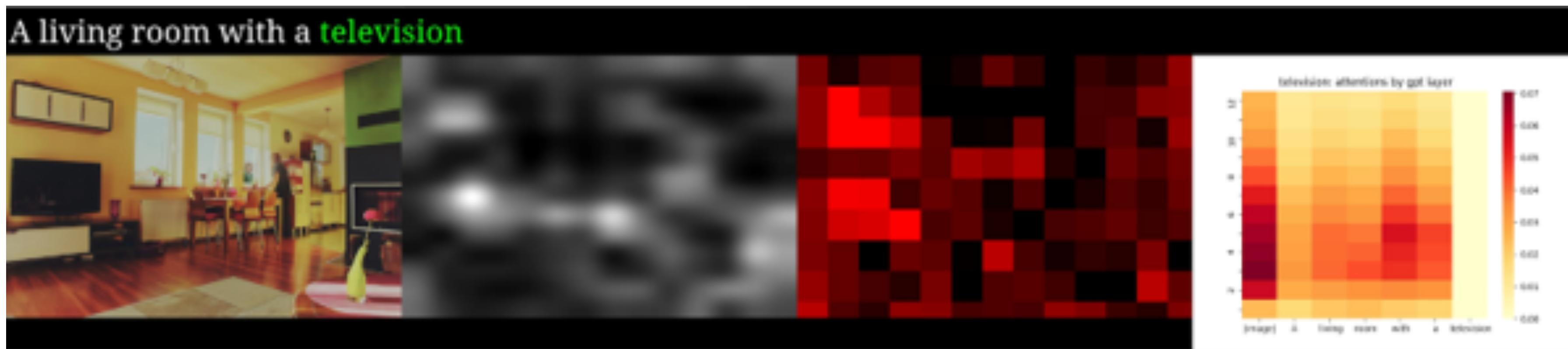
Multimodality

How aleph alpha does it



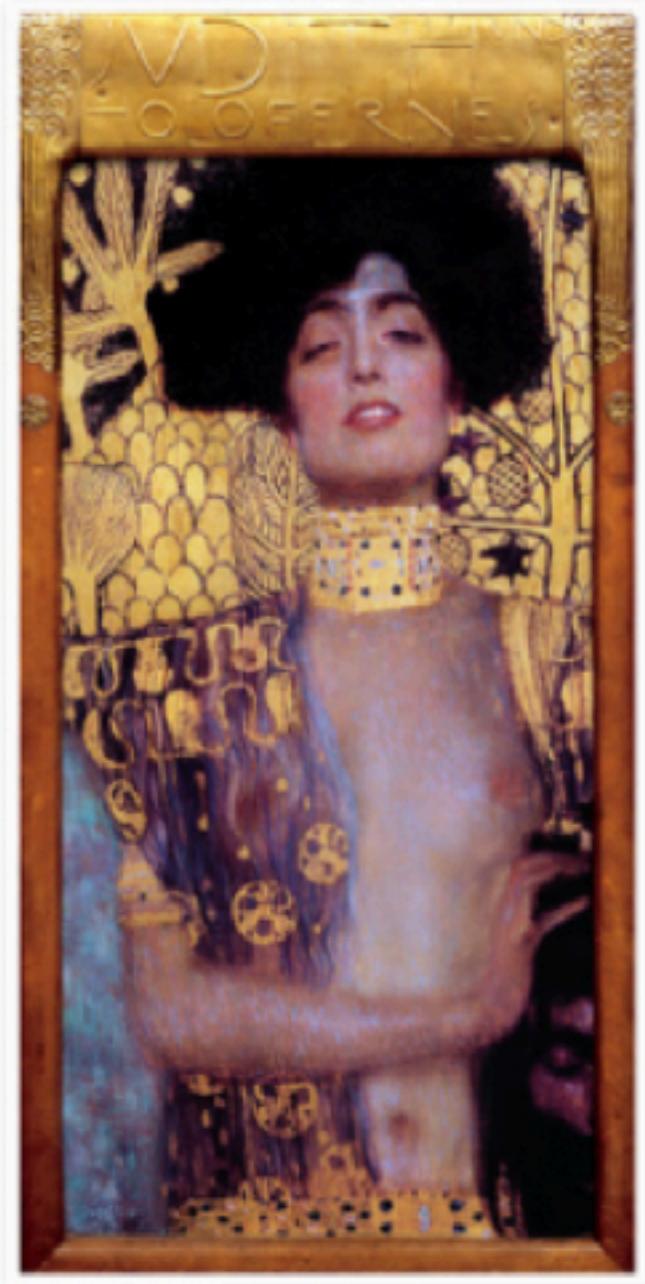
Multimodality

How aleph alpha does it



My Applications

- Classifying art:
<https://opensea.io/collection/peezmo-ai-21>



```
1 pred = learner.predict(img_fastai)
2 print_pred_probs("test",pred, 3)
```

'test ; Art Nouveau (Modern): 98.56% ; Romanticism: 0.41% ; Symbolism: 0.38%'



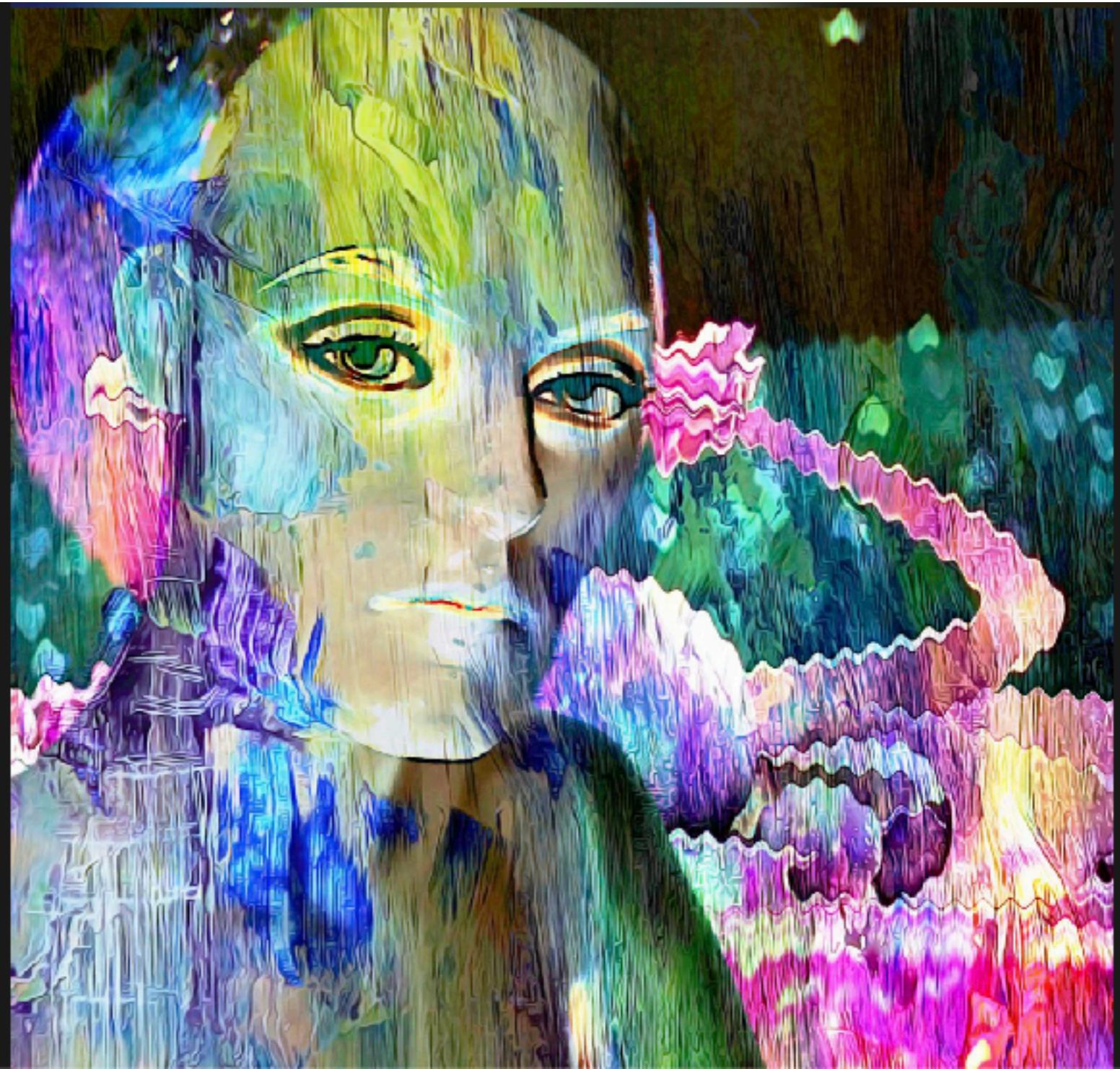
```
1 pred = learner.predict(img)
2 t= print_pred_probs('testimage',pred, 3)
3 print (t)
```

'testimage ; Surrealism: 82.73% ; Abstract Art: 10.68% ; Pop Art: 3.71%'

My Applications

Classifying and Labelling ART

- Labelling art:

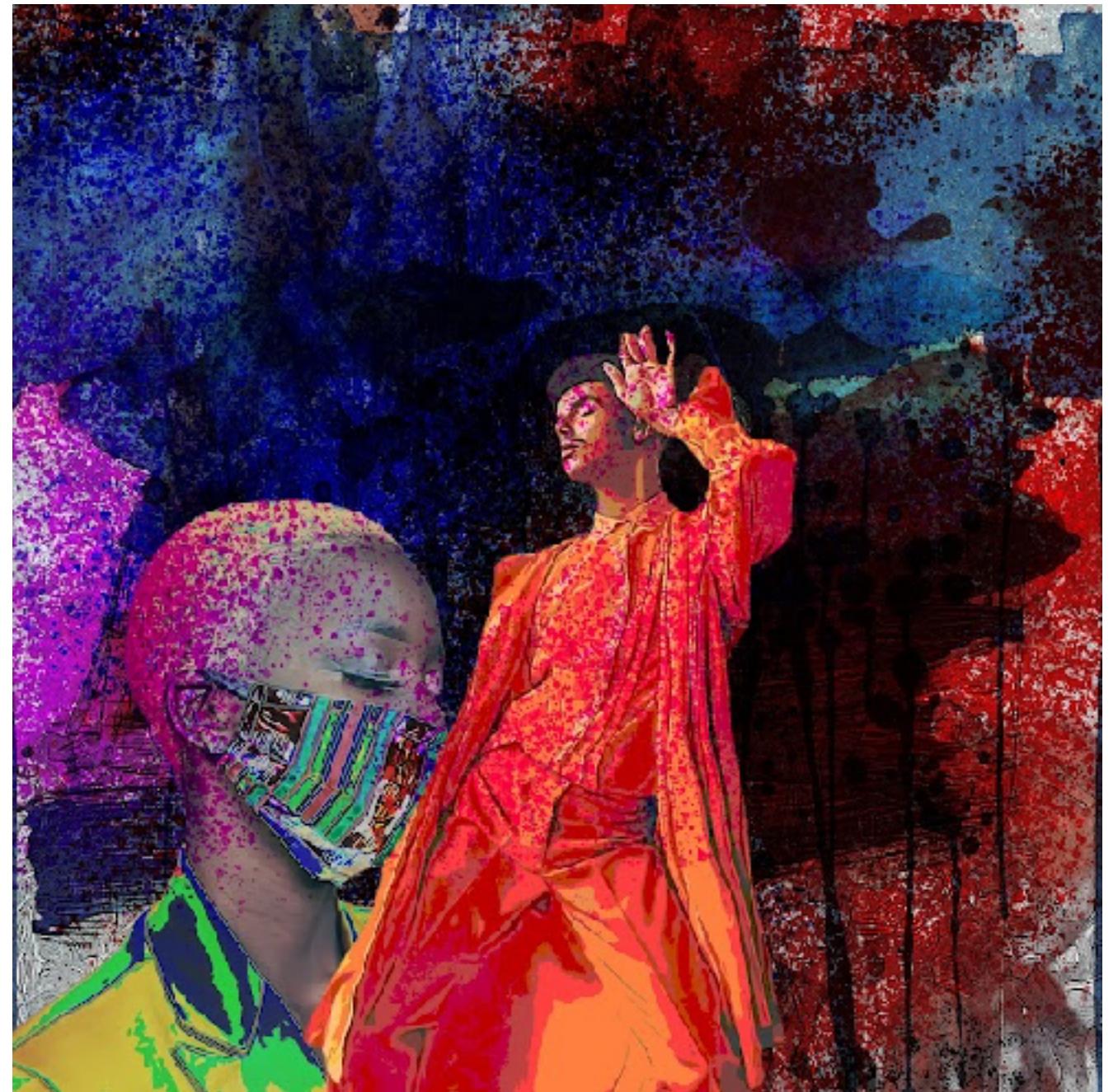


Suggested keywords:

women	92%
creativity	86%
adult	79%
one person	72%
multi colored	67%
young adult	61%
portrait	55%
beauty	45%
paint	42%
human face	41%
caucasian ethnicity	40%
acrylic painting	35%
painted image	31%
flower	29%

My Applications

Classifying and Labelling ART



Pandemic Dancer



Bears on Wallstreet



Bend reality