Learning Transferable Visual Models From Natural Language Supervision

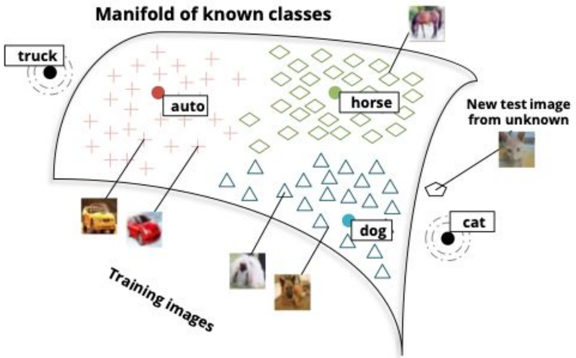
Main presentation

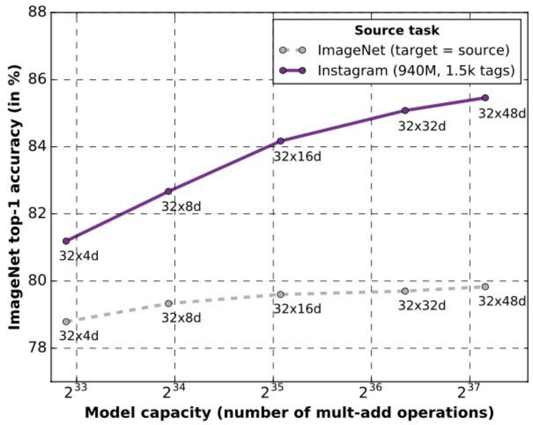
Alec Radford, Jong Wook Kim, Chris Hallacy, A. Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, Ilya Sutskever, *OpenAI,* 2021.

Presenter:	**Diego Calanzone**
Seminar:	**Learning with Limited Labeled Data**
Academic year:	**2021/2022**

Overview

- **Natural Language Supervision: SoTA, ZS performance**
  - Mapping images to text embeddings
  - Learning better image representations
  - Exploiting large web data
- **CLIP : motivation, intuition**
  - Dataset & data pipeline
  - Experimental setup, methodological components
  - Scaling, data efficiency
  - Performance, robustness, bias
  - Limitations and broader impacts
- **CLIP “spin-offs”:**
  - Multimodal neurons
  - Zero-shot text generation: DALL-E

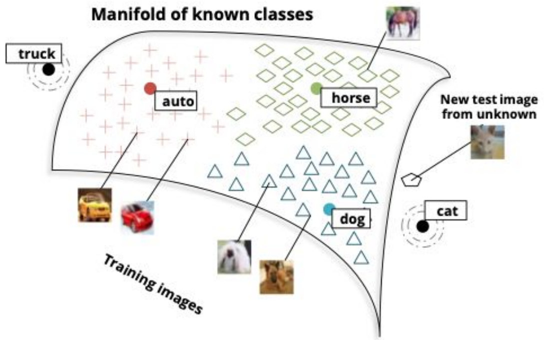




Natural Language Supervision (Supervised)

Zero-Shot Learning Through Cross-Modal Transfer

R.Socher, M. Ganjoo, C. D. Manning, A. Y. Ng, 2013

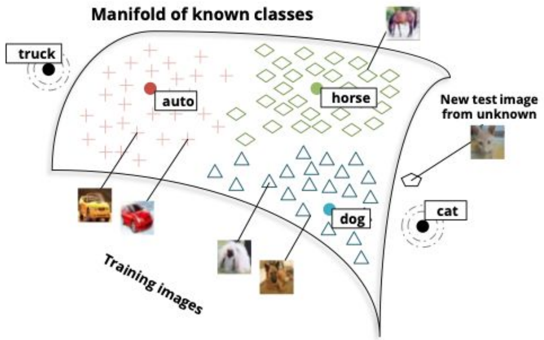


Natural Language Supervision (Supervised)

Zero-Shot Learning Through Cross-Modal Transfer

R.Socher, M. Ganjoo, C. D. Manning, A. Y. Ng, 2013

- <u>Training dataset</u>: CIFAR-100
- <u>Key points</u>: semantic embedding space, novelty detection
- <u>ZS Class. Accuracy</u>: **52.7%** (CIFAR-100, **6** novel classes)

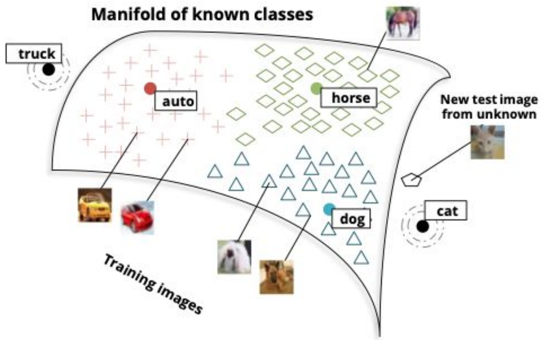


Natural Language Supervision (Supervised)

Zero-Shot Learning Through Cross-Modal Transfer

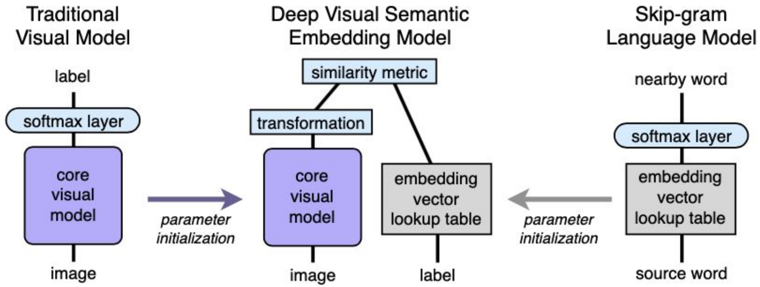
R.Socher, M. Ganjoo, C. D. Manning, A. Y. Ng, 2013

- Training dataset: CIFAR-100
- Key points: semantic embedding space, novelty detection
- ZS Class. Accuracy: **52.7%** (CIFAR-100, **6** novel classes)



DeViSE: A Deep Visual-Semantic Embedding Model

A. Frome, G. S. Corrado, J. Shlens,, et al., 2013*

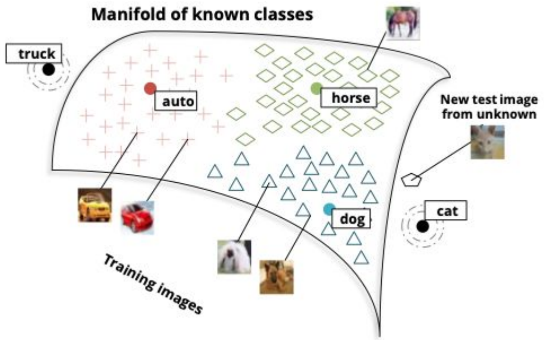


Natural Language Supervision (Supervised)

Zero-Shot Learning Through Cross-Modal Transfer

R.Socher, M. Ganjoo, C. D. Manning, A. Y. Ng, 2013

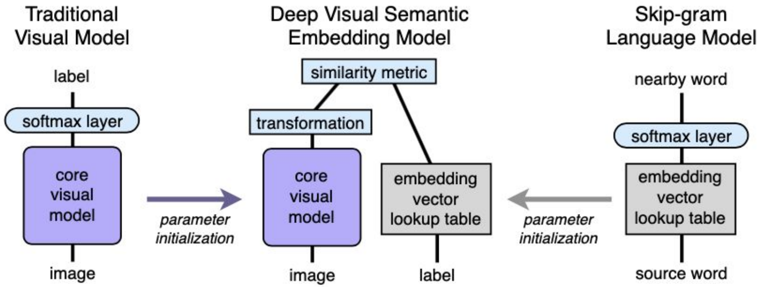
- Training dataset: CIFAR-100
- Key points: semantic embedding space, novelty detection
- ZS Class. Accuracy: **52.7%** (CIFAR-100, **6** novel classes)



DeViSE: A Deep Visual-Semantic Embedding Model

A. Frome, G. S. Corrado, J. Shlens,, et al., 2013*

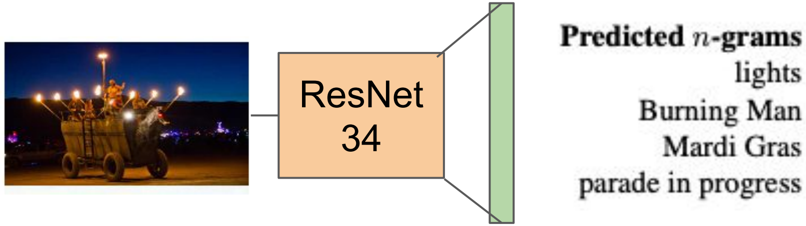
- Training dataset: ILSVRC 2012 1K
- Key points: semantic vector prediction, language-model supervision
- ZS Class. Accuracy: **36.4%,** ImageNet 2011 21K



Natural Language Supervision (Supervised)

Learning Visual N-Grams from Web Data

A. Li, A. Jabri, A. Joulin, L. van der Maaten, 2017

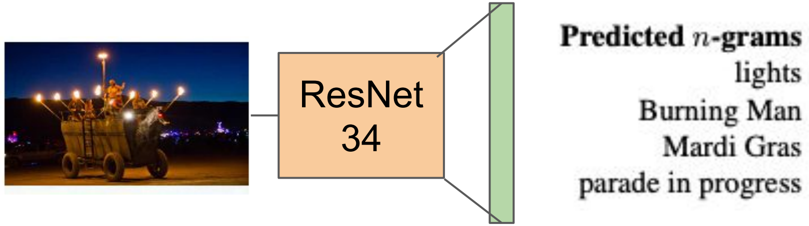


Natural Language Supervision (Supervised)

Learning Visual N-Grams from Web Data

A. Li, A. Jabri, A. Joulin, L. van der Maaten, 2017

- Training dataset: YFCC100M (image-comments)
- Key points: ImageNet fine-tuned, n-grams probability distribution, conditional probability
- ZS Classif. Accuracy: **11.5%** (ImageNet-1K), **23.0%** (SUN), **72.4%** (Yahoo)

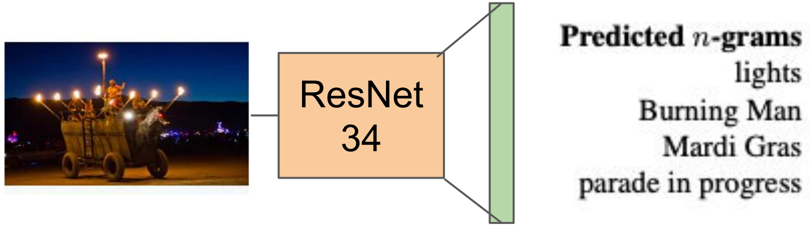


Natural Language Supervision (Supervised)

Learning Visual N-Grams from Web Data

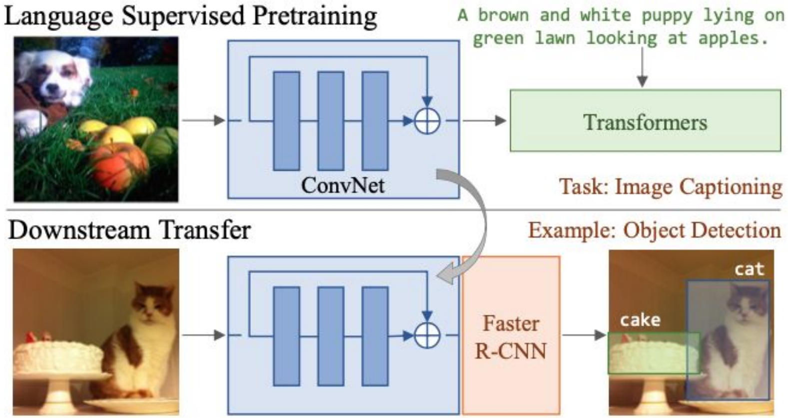
A. Li, A. Jabri, A. Joulin, L. van der Maaten, 2017

- Training dataset: YFCC100M (image-comments)
- Key points: ImageNet fine-tuned, n-grams probability distribution, conditional probability
- ZS Classif. Accuracy: **11.5%** (ImageNet-1K), **23.0%** (SUN), **72.4%** (Yahoo)



VirTex: Learning Visual Repr. from Text Annotations

K. Desai, J. Johnson, 2020

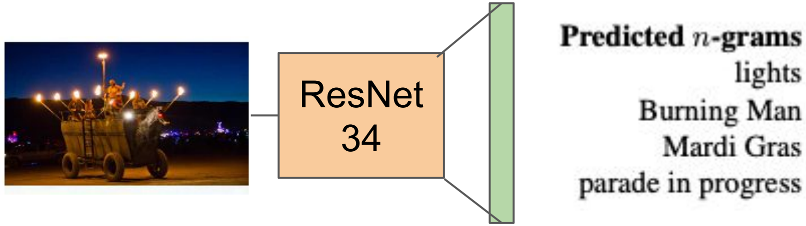


Natural Language Supervision (Supervised)

Learning Visual N-Grams from Web Data

A. Li, A. Jabri, A. Joulin, L. van der Maaten, 2017

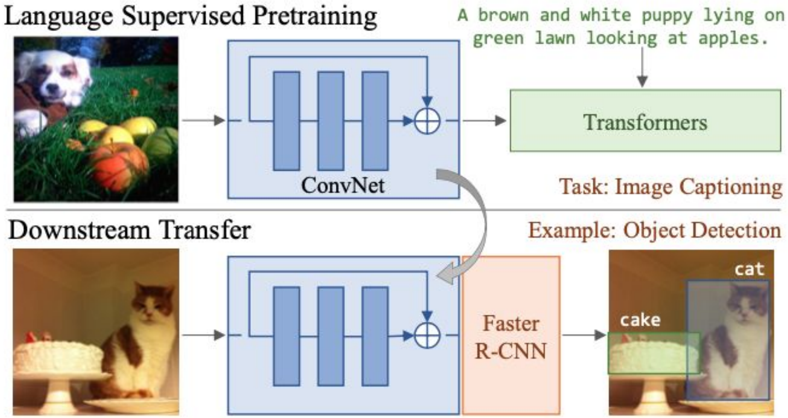
- <u>Training dataset</u>: YFCC100M (image-comments)
- <u>Key points</u>: ImageNet fine-tuned, n-grams probability distribution, conditional probability
- <u>ZS Classif. Accuracy</u>: **11.5%** (ImageNet-1K), **23.0%** (SUN), **72.4%** (Yahoo)



VirTex: Learning Visual Repr. from Text Annotations

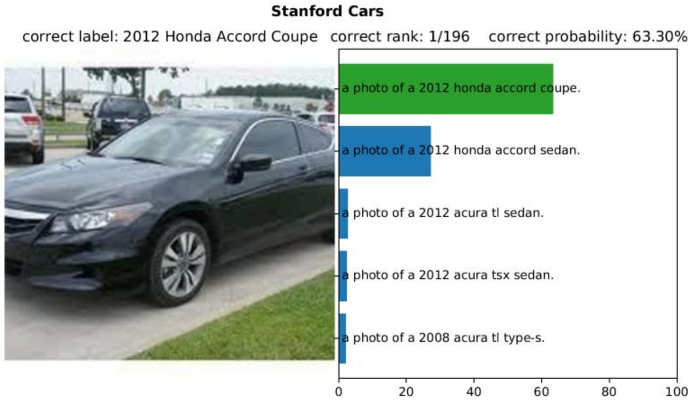
K. Desai, J. Johnson, 2020

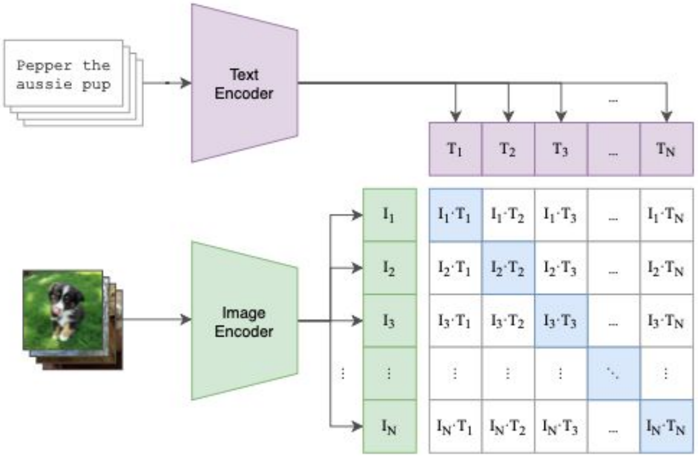
- <u>Training dataset</u>: COCO-captions (train2017 split)
- <u>Key points</u>: transformer cross-attention with CNN encodings
- <u>Classif. Accuracy</u>: **88.7%** (PASCAL VOC07), **53.8%** (ImageNet-1K)



Overview

- **Natural Language Supervision: SoTA, ZS performance**
    - Mapping images to text embeddings
    - Learning better image representations
    - Exploiting large web data
- **CLIP : motivation, intuition**
    - Dataset & data pipeline
    - Experimental setup, methodological components
    - Scaling, data efficiency
    - Performance, robustness, bias
    - Limitations and broader impacts
- **CLIP "spin-offs":**
    - Multimodal neurons
    - Zero-shot text generation: DALL-E





Choice of data & pre-processing pipeline

- **Objective:** high performance in zero-shot and new downstream tasks
  MS-COCO, Visual Genome → too **small**; YFCC100M → **arguable** quality

- **Choice:** creating a new dataset **"WIT: WebImageText" 400M**

  - (image, text) pairs from a list of **500k queries**: words occurring >100 times on Wikipedia
  - Limit of **20k (img-txt) pairs / query**

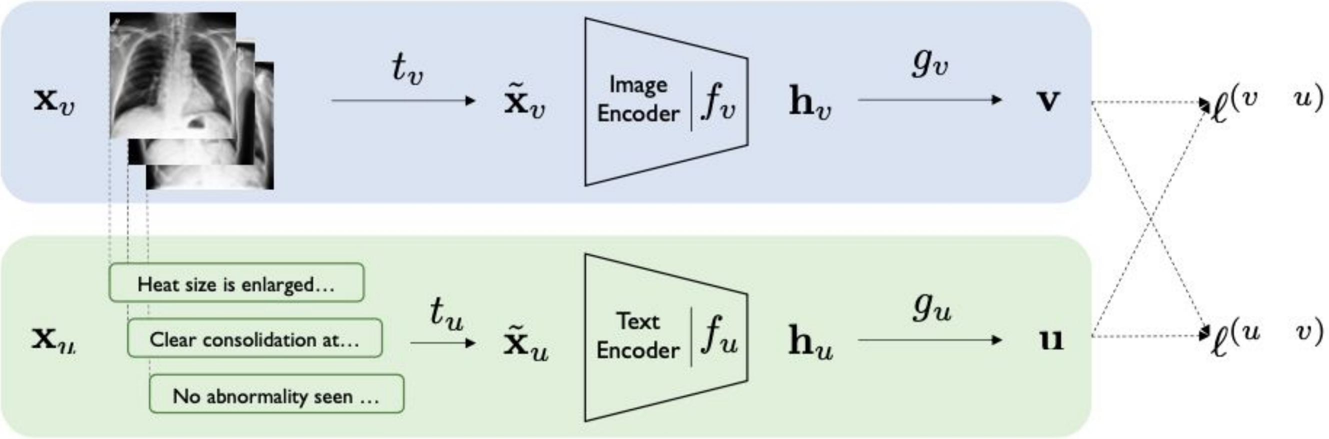
- **Prompt engineering + Ensembling:** tackling polysemy and averaging multiple prompts
  ie: "a boxer, *a type of dog*" vs. "a boxer, *a type of athlete*"
  Accuracy gain → **+ ~5%**

- **Data overlap analysis:** threshold on image distances in the embedding space → duplicate detection → 3.2% avg. overlap → +0.6% accuracy gain

The foundations of CLIP

Contrastive Learning of Medical Visual Representations from Paired Images and Text

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D. Manning, Curtis P. Langlotz, 2020



The foundations of CLIP

- Image/text embedding vectors: $\mathbf{v} = g_v(f_v(\tilde{\mathbf{x}}_v)) \qquad \mathbf{u} = g_u(f_u(\tilde{\mathbf{x}}_u))$

- Image → text contrastive loss:

$$\ell_i^{(v\ !\ u)} = -\log \frac{\exp(\langle \mathbf{v}_i, \mathbf{u}_i \rangle / \tau)}{\sum_{k=1}^{N} \exp(\langle \mathbf{v}_i, \mathbf{u}_k \rangle / \tau)}$$

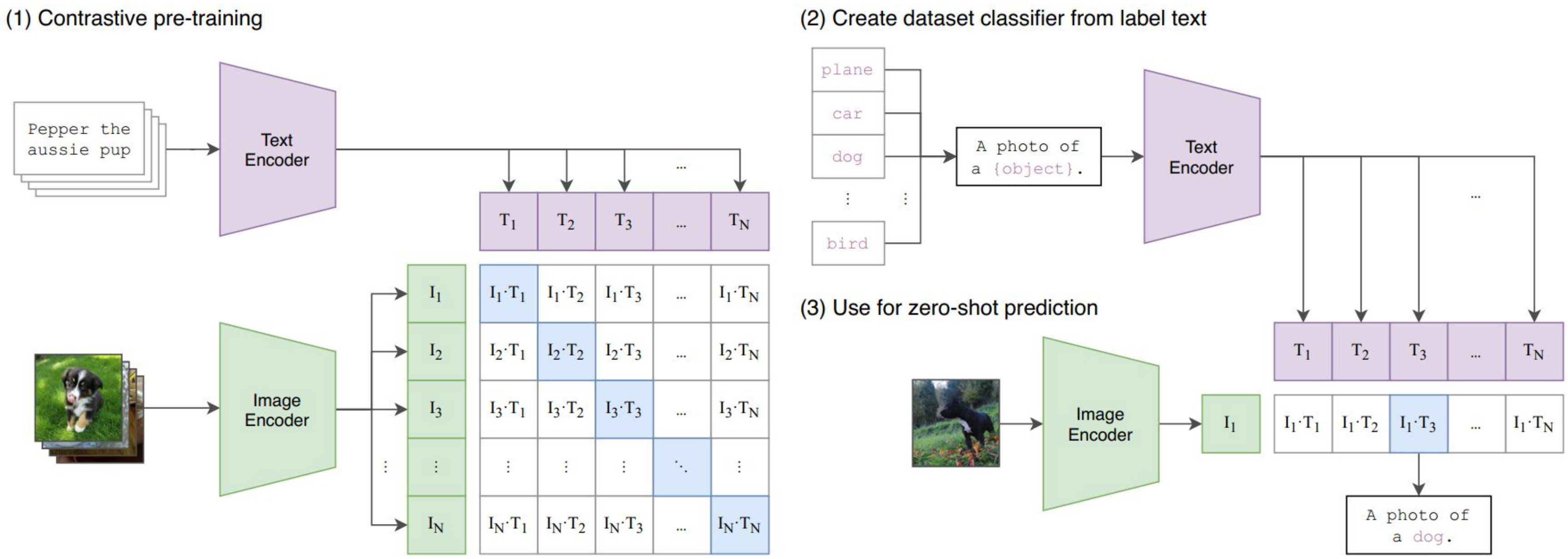
- Text → image contrastive loss:

$$\ell_i^{(u\ !\ v)} = -\log \frac{\exp(\langle \mathbf{u}_i, \mathbf{v}_i \rangle / \tau)}{\sum_{k=1}^{N} \exp(\langle \mathbf{u}_i, \mathbf{v}_k \rangle / \tau)}$$

- Loss function:

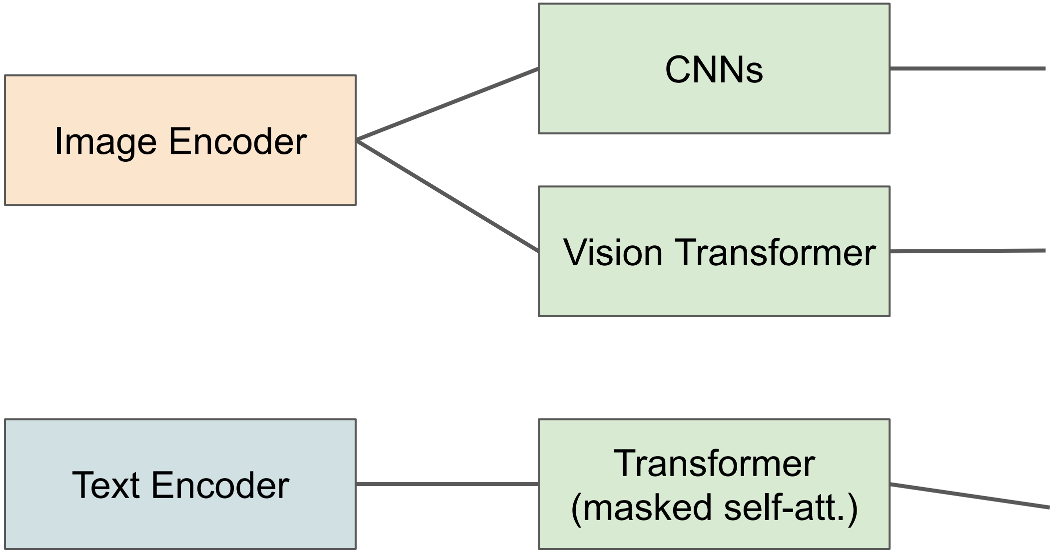
$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} \left(\lambda \ell_i^{(v\ !\ u)} + (1-\lambda) \ell_i^{(u\ !\ v)} \right)$$

Contrastive Language-Image Pretraining (CLIP)



Training setup & choice of the architecture

EBERHARD KARLS UNIVERSITÄT TÜBINGEN



- ResNet50, ResNet101
- Efficient-Net: RN50x4, RN50x16, RN50x64

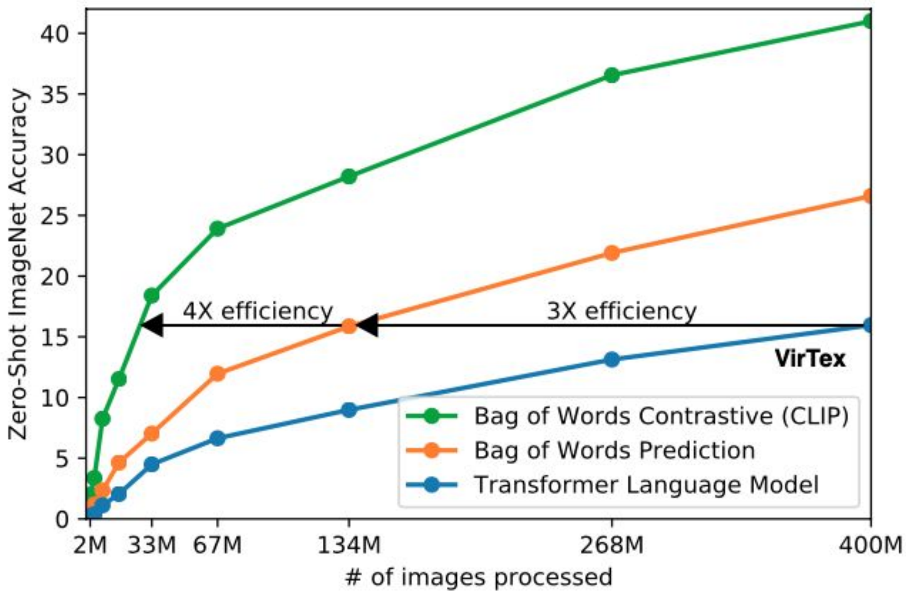
- ViT-B/32
- ViT-B/16
- ViT-L/14@336 (CLIP)

➢ **# Layers:** 12x512
➢ **# Params:** 63M
➢ **# Att.Heads:** 8
➢ Byte Pair Encoding
➢ **Vocab.size**: 49,152
➢ **Seq.len.:** 76

- **Optimizer:** Adam+reg., **mini-batch size**: 32,768, **epochs:** 32

Efficiency in scaling the dataset

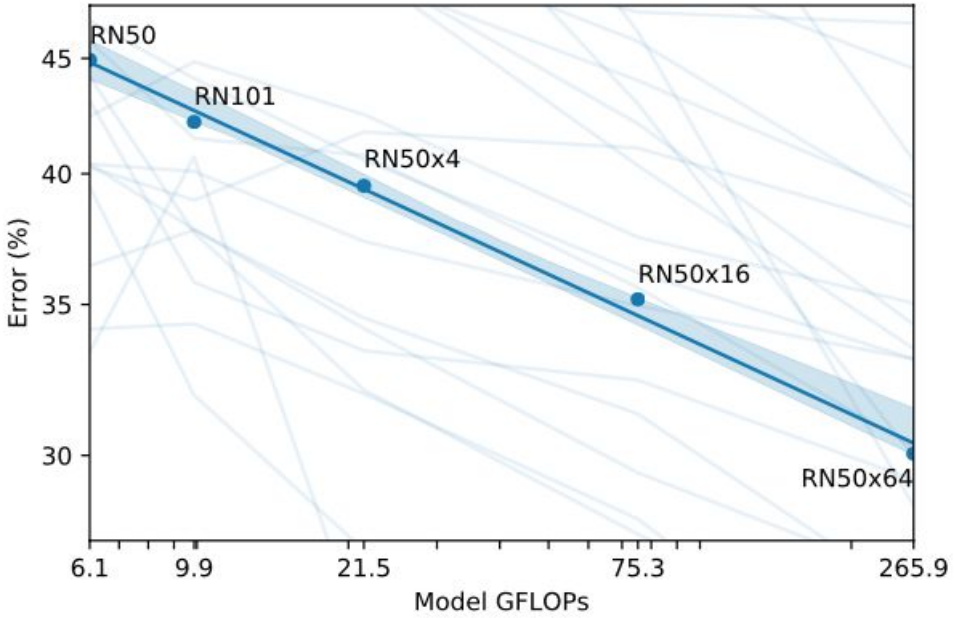
- **Data scaling:** CLIP > VirTex, bag of words > sentences



Efficiency in scaling the model

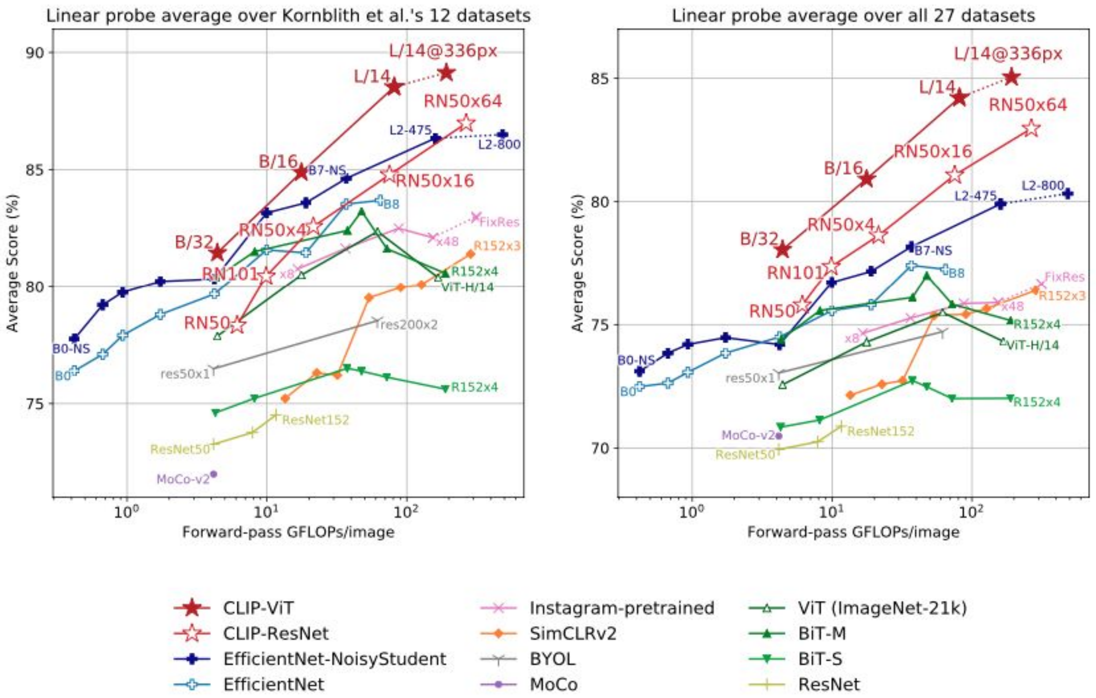
- **Data scaling:** CLIP > VirTex, bag of words > sentences

- **Model scaling:** up to 44x → smooth error decrease (like GPT). Scaling:
  - img: width, depth, res.
  - txt: width



Compared classification performance in scaling

- **Data scaling:** CLIP > VirTex, bag of words > sentences
- **Model scaling:** up to 44x → smooth error decrease (like GPT). Scaling:
  - <u>img:</u> width, depth, res.
  - <u>txt:</u> width
- Linear probe **CLIP-ViT-L/14-336px** outperforms the CNN-baseline



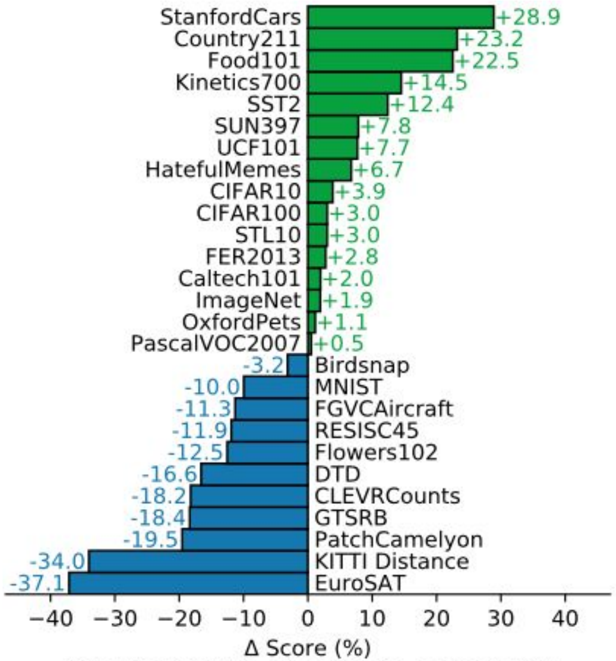
Zero-shot Performance

- <u>StanfordCars, STL10:</u> SoTA performance on datasets with few label examples and **a lot of unlabeled data.**

- <u>Kinetics700, UCF101:</u> NLS → better context with verbs

- <u>EuroSAT, PatchCamelyon:</u> fine-grained tasks, specific examples → weak

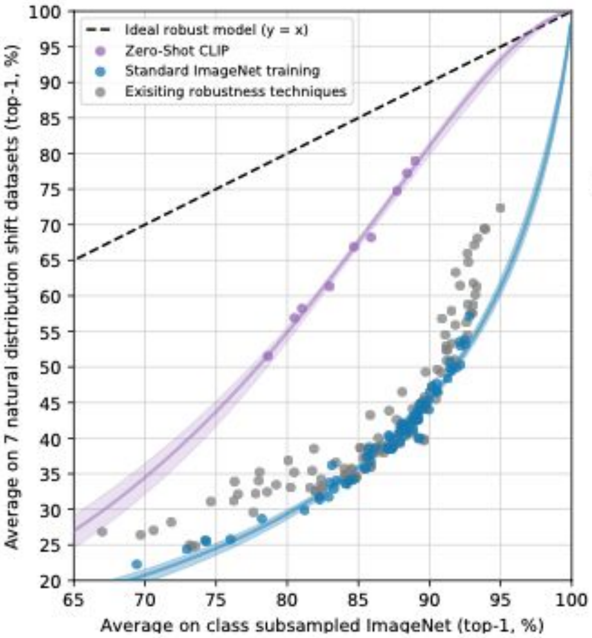
Advantages with CLIP : output dynamicity, better text-supervision, pre-trained once.



Zero-Shot CLIP vs. Linear Probe on ResNet50

	aYahoo	ImageNet	SUN
Visual N-Grams	72.4	11.5	23.0
CLIP	**98.4**	**76.2**	**58.5**

Robustness with distributional shifts



Dataset Examples

Dataset	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%

Limitations & Impacts

- Required **1000x** scaling → SOTA

Limitations & Impacts

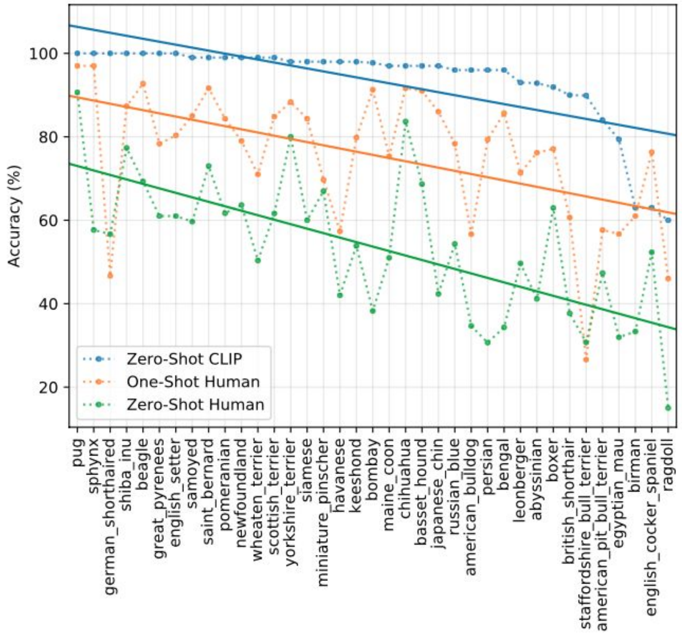
- Required **1000x** scaling → SOTA
- Poor performance in specific/more **abstract** tasks

Limitations & Impacts

- Required **1000x** scaling → SOTA
- Poor performance in specific/more **abstract** tasks
- Underperforms with **synthetic** data → only 88% ZS acc. on MNIST!

Limitations & Impacts

- Required **1000x** scaling → SOTA
- Poor performance in specific/more **abstract** tasks
- Underperforms with **synthetic** data → only 88% ZS acc. on MNIST!
- Unfiltered dataset → **bias** harm
- Inefficient few-shot performance wrt. humans



Broader Impacts

- “Roll your own classifier” without re-training
- Surveillance: face recognition, emotion recognition, action recognition, geo-localization

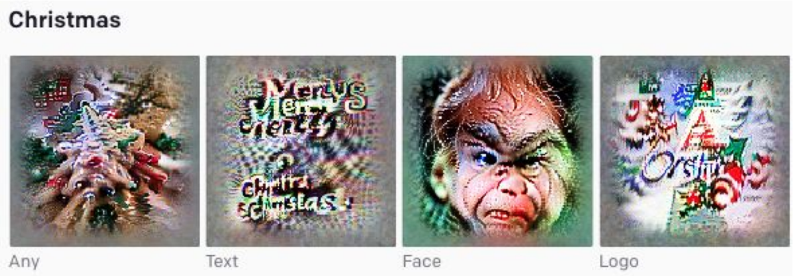
	Accuracy	Majority Vote on Full Dataset	Accuracy on Guesses	Majority Vote Accuracy on Guesses
Zero-shot human	53.7	57.0	69.7	63.9
Zero-shot CLIP	**93.5**	**93.5**	**93.5**	**93.5**
One-shot human	75.7	80.3	78.5	81.2
Two-shot human	75.7	85.0	79.2	86.1

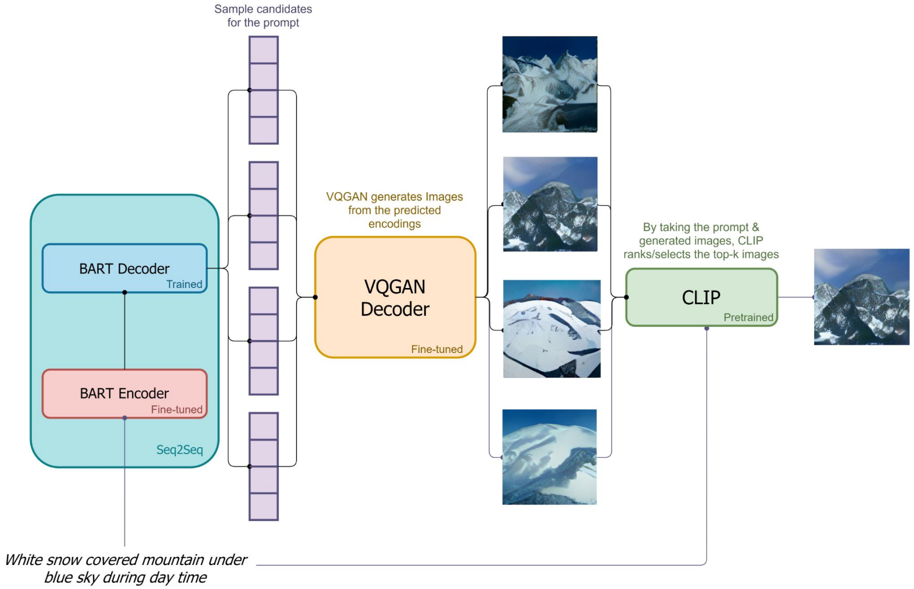
"Spin-offs" of CLIP

- <u>CLIP for Surveillance</u>
  - Search through frames in video
  - "Roll your own classifier"

- <u>Multi-modal neurons</u> found in CLIP
  - Neurons that respond to **particular inputs**
  - Work **similarly** to the **brain's** neurons
  - They organize **highly abstract concepts**

- <u>Zero-shot image generation:</u> DALL-E
  - VQ-VAE + GPT-style Transformer
  - CLIP to rank generated images

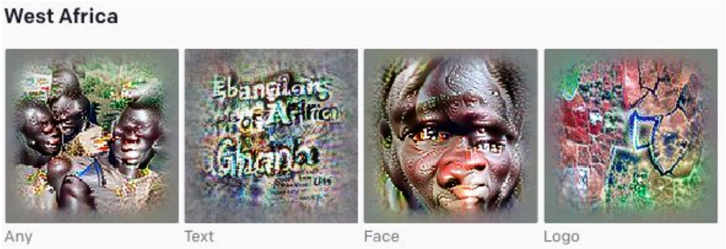




Thanks for listening!

Questions are welcome.

(More on DALL-E & multi-modal neurons in the Q&A if requested)



TEXT PROMPT

a store front that has the word 'openai' written on it. . . .

AI-GENERATED IMAGES

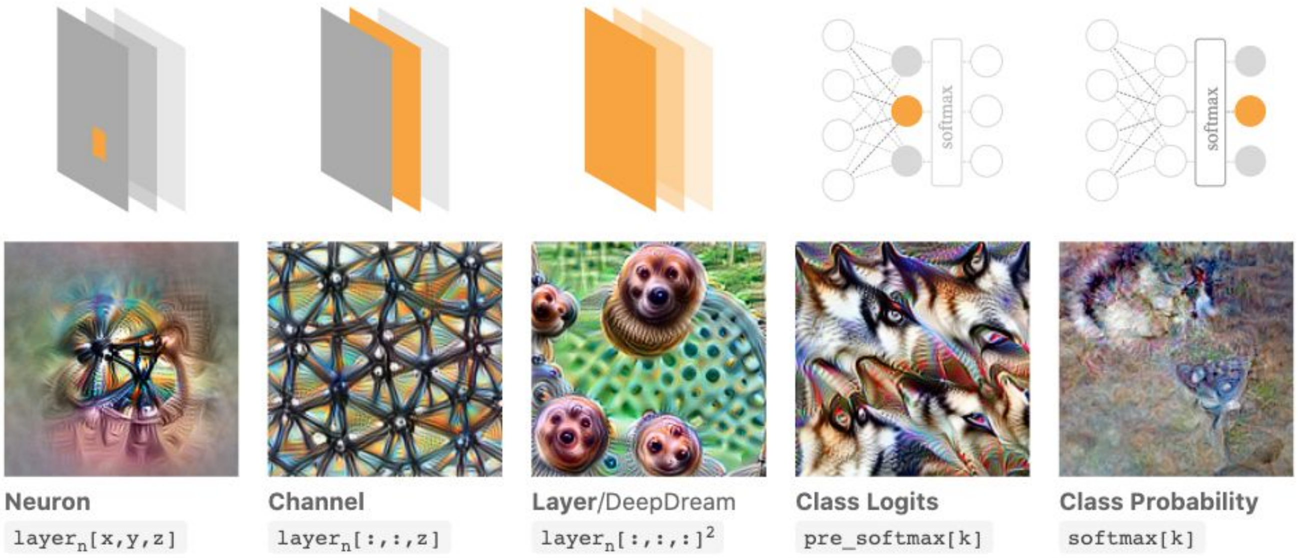
For notebook examples, further insights, sources, **I made a repo!**

https://github.com/halixness/understanding-CLIP

Multimodal neurons in CLIP : visualizing concepts

BONUS

By **optimization** (ie. with GANs/ AEs) → find images that **maximize the activation** of a component
https://distill.pub/2017/feature-visualization/



Multimodal neurons in CLIP

Multimodal neurons in CLIP

BONUS

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

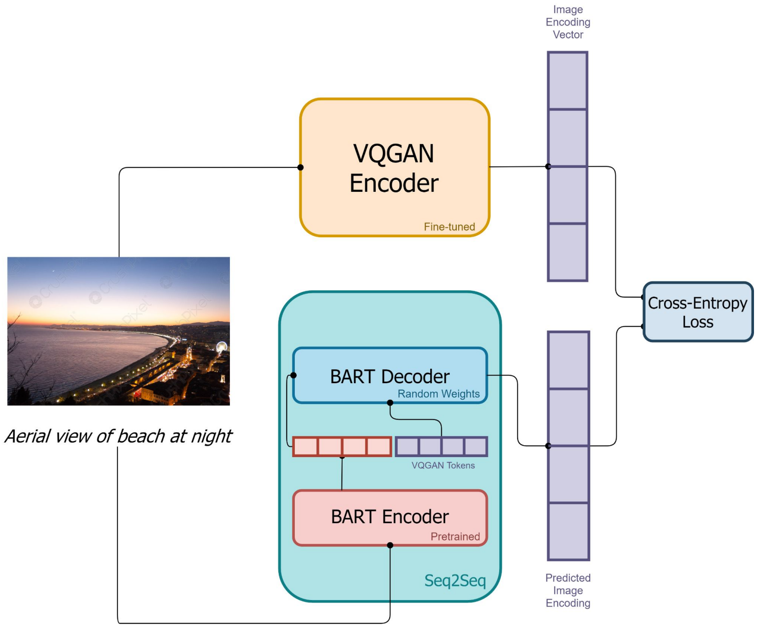
BONUS

Multimodal neurons in CLIP

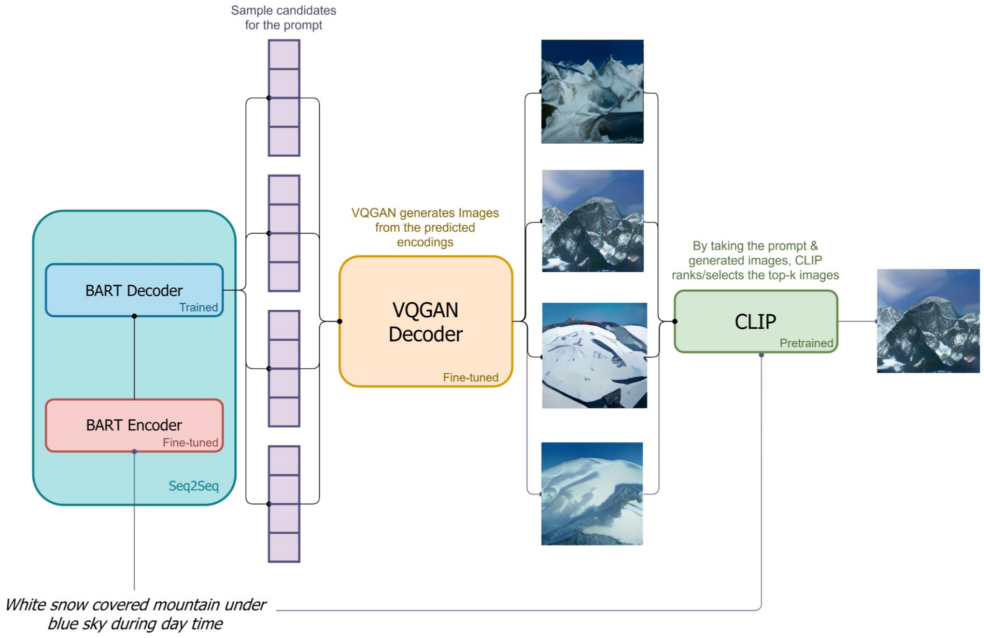
Zero-shot text generation: DALL-E

BONUS

Training

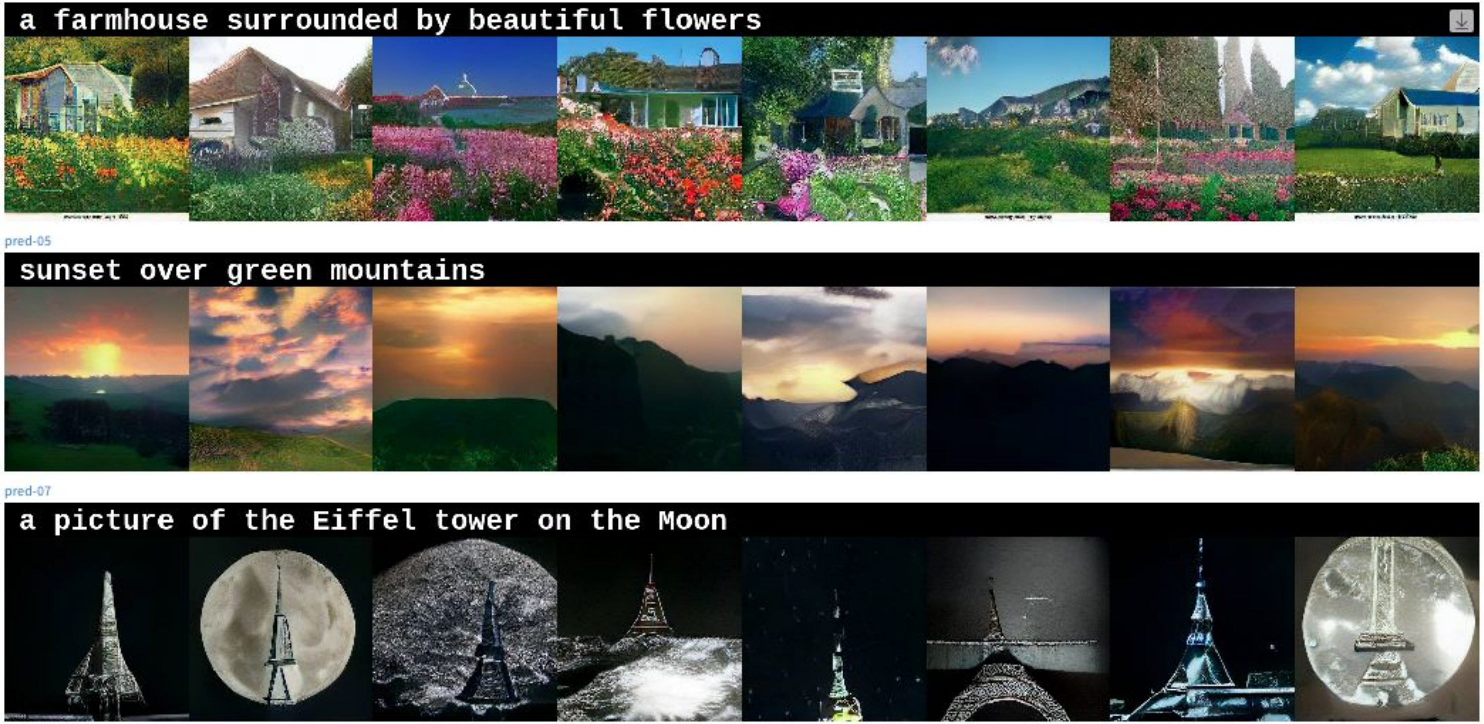


Inference



BONUS

Zero-shot text generation: DALL-E



BONUS

Zero-shot text generation: DALL-E