DALL-E & CLIP

Lecture-3

CAP6412, Spring 2023

Mubarak Shah

DALL-E

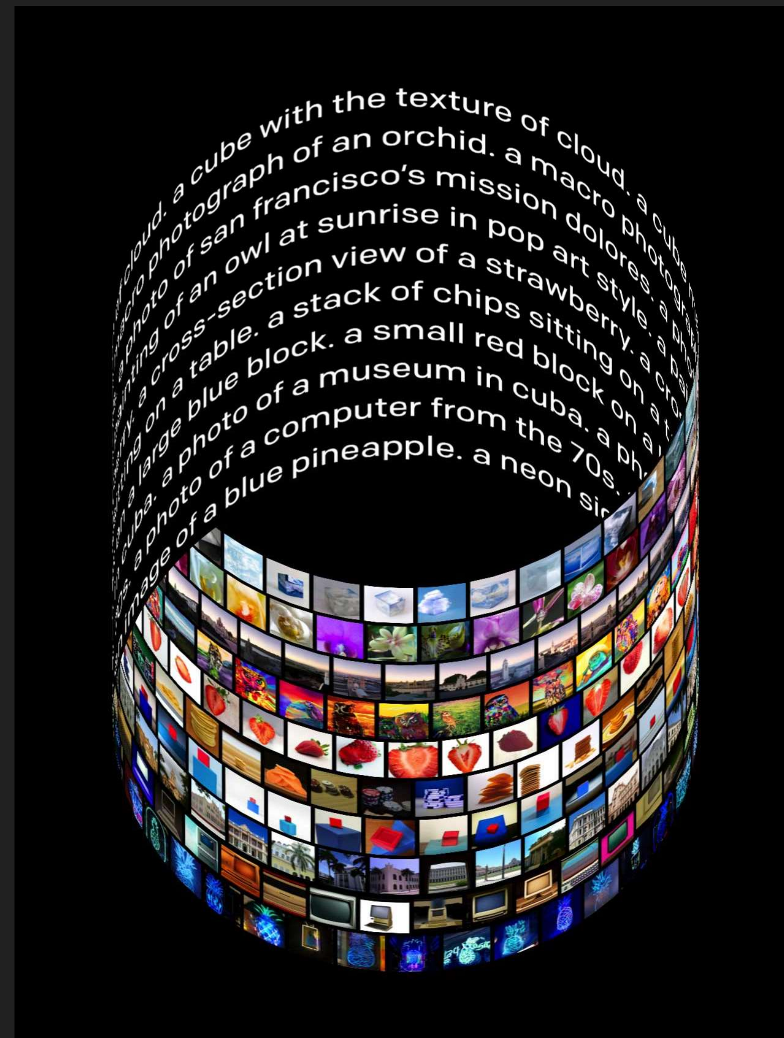
Authors: Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever

Open AI (ICML 2021)

Presentors: Adam Kutchak, George Lu, Fernando Treviño, and Sarah Wilson

(CAP6412, Spring 2022)

https://www.youtube.com/watch?v=ArPTcWpVCZw



Introduction

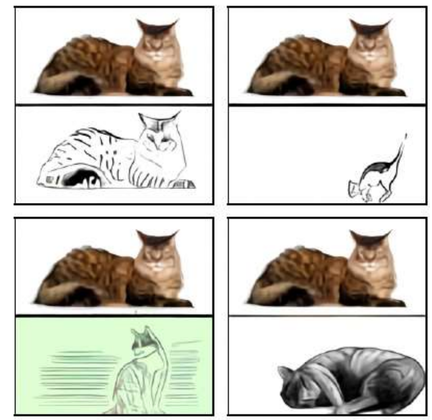
- Generate Images from text captions
- 12 billion parameters version of GPT-3
- Dataset comprised of 3.3 million text - image pairs
- Combine unrelated concepts

Image Generation

(a) a tapir made of accordion. a tapir with the texture of an accordion.

(b) an illustration of a baby hedgehog in a christmas sweater walking a dog

(c) a neon sign that reads "backprop". a neon sign that reads "backprop". backprop

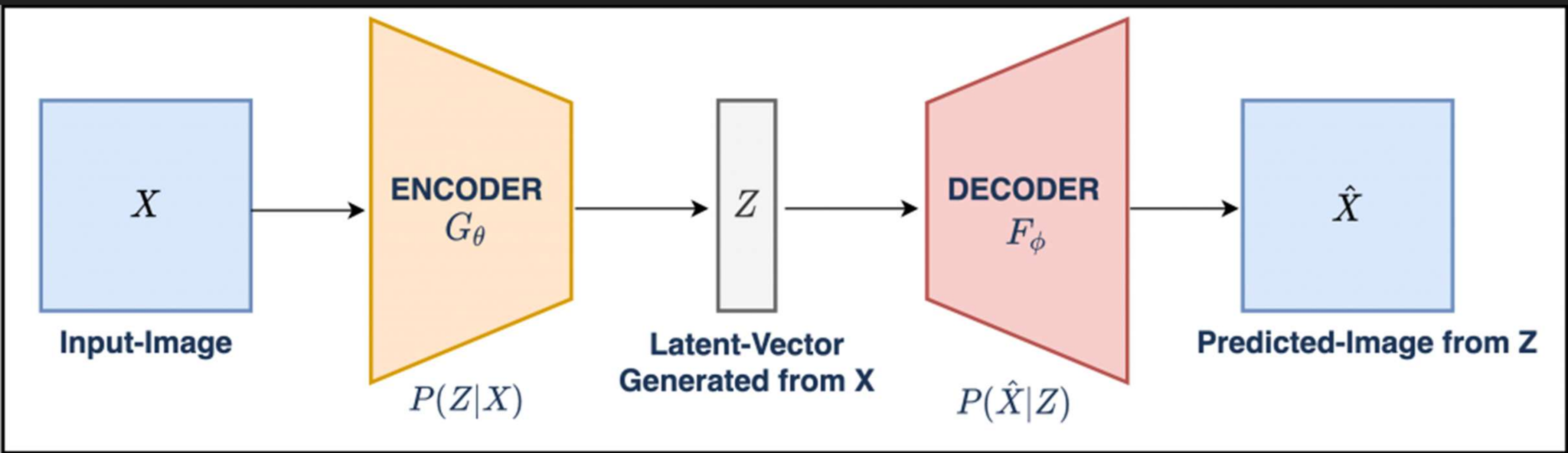


(d) the exact same cat on the top as a sketch on the bottom

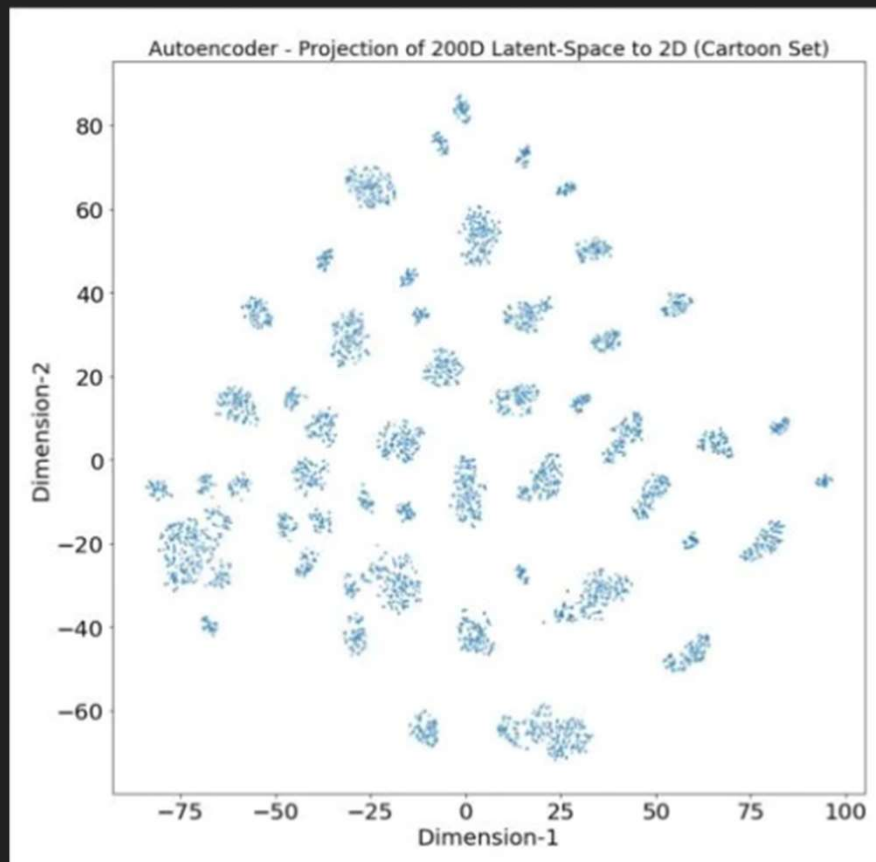
Related Works

- Autoencoder - (encoder - decoder)
- Variational Autoencoders (continuous state space)
- VQ-VAE (discrete quantized state space)

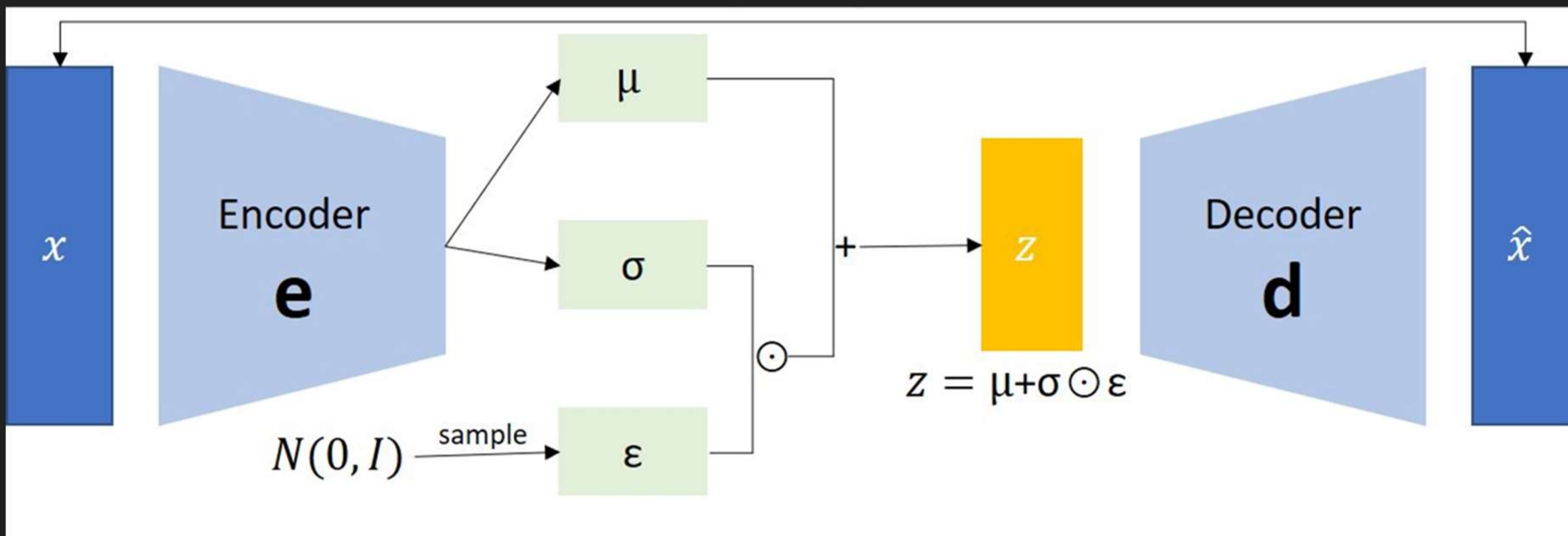
Related Work - Autoencoder

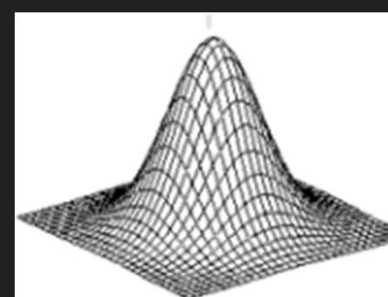


Related Work - Autoencoder problem

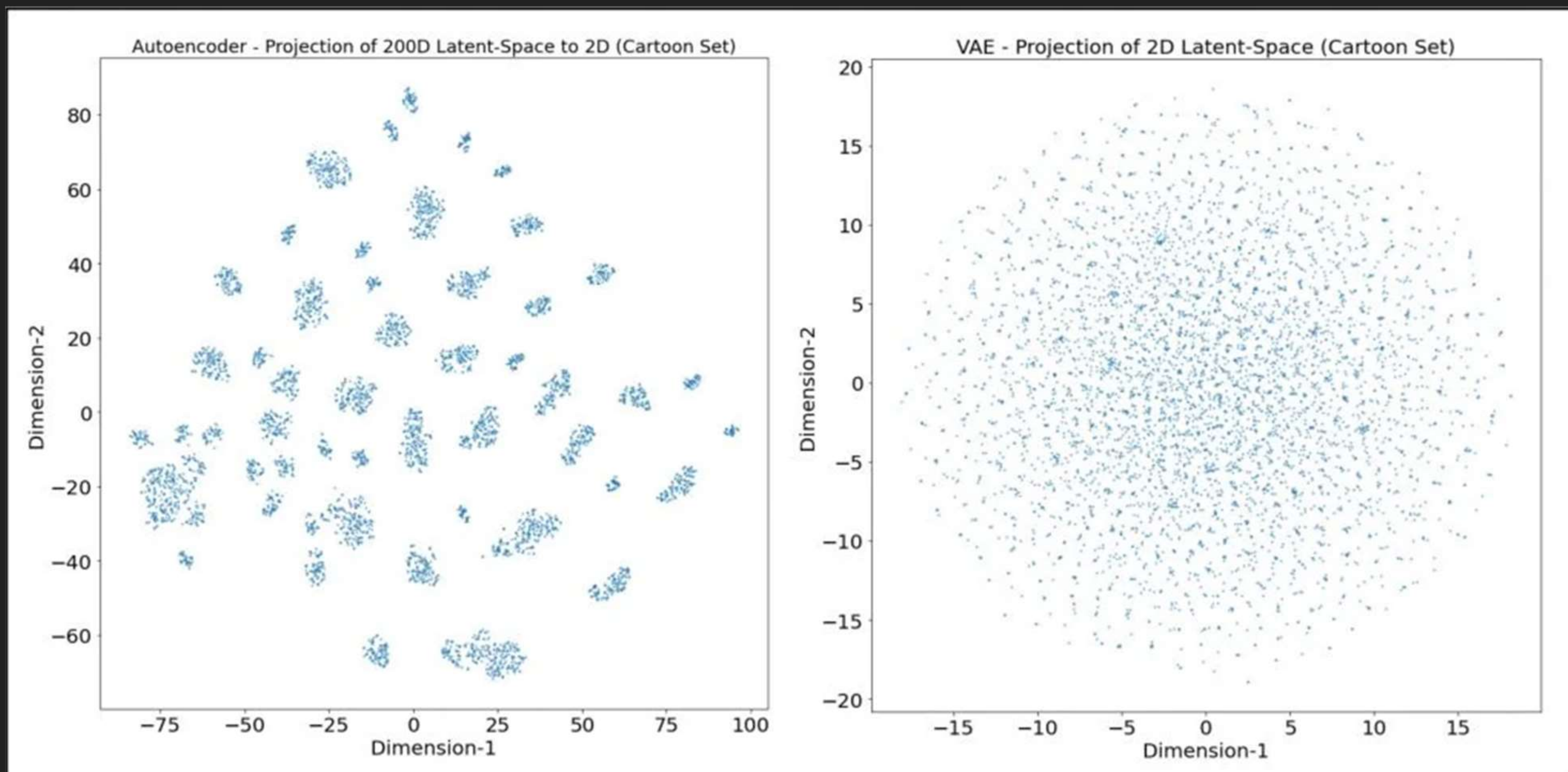


Related Work - Variational Autoencoder

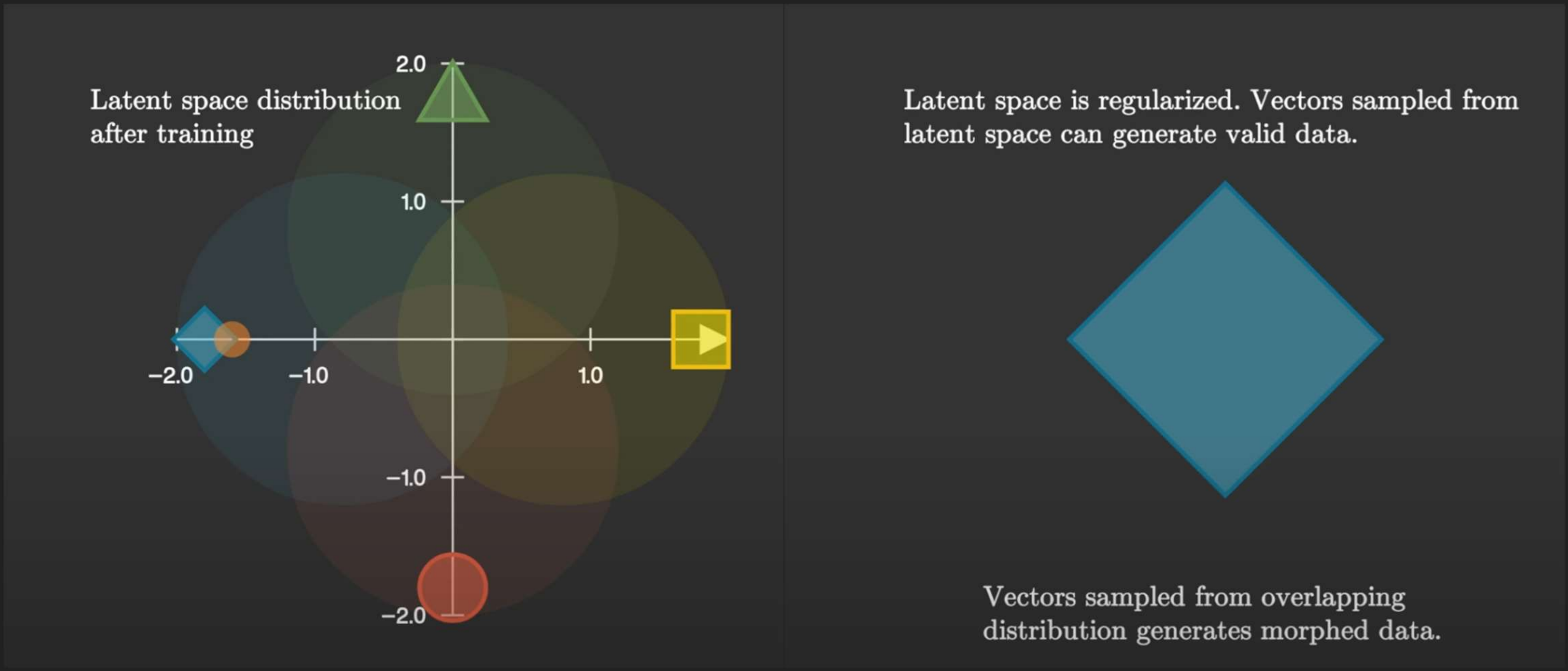




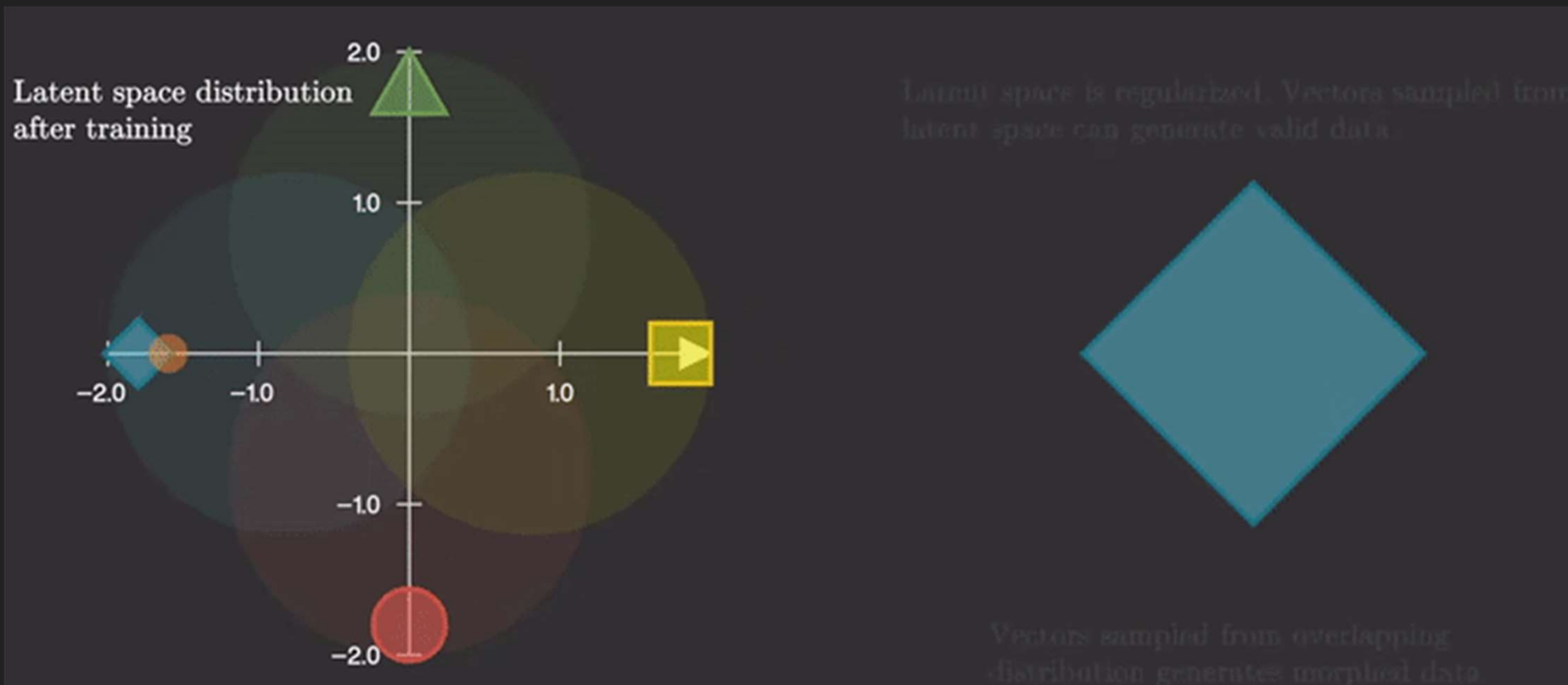
Related Work - Autoencoder vs. VAE



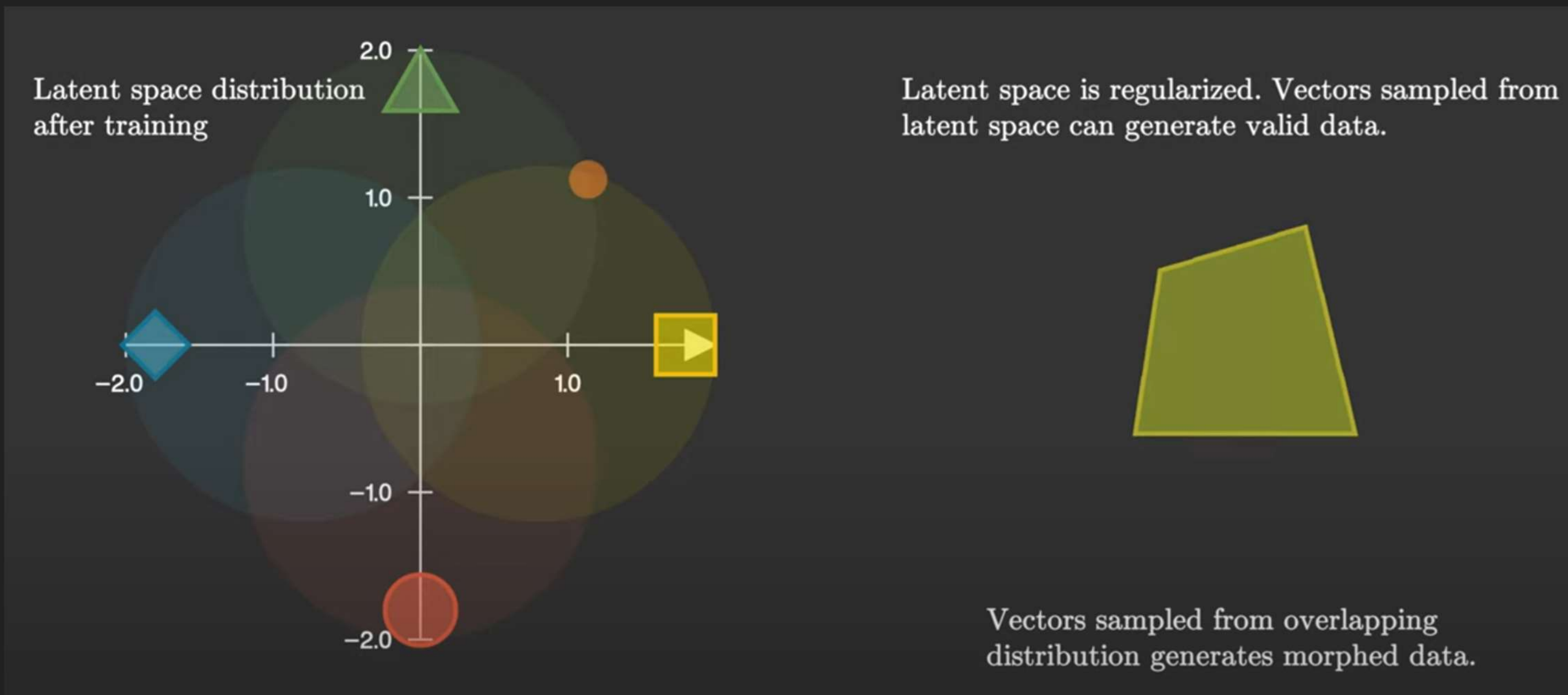
Related Work - Variational Autoencoder problem



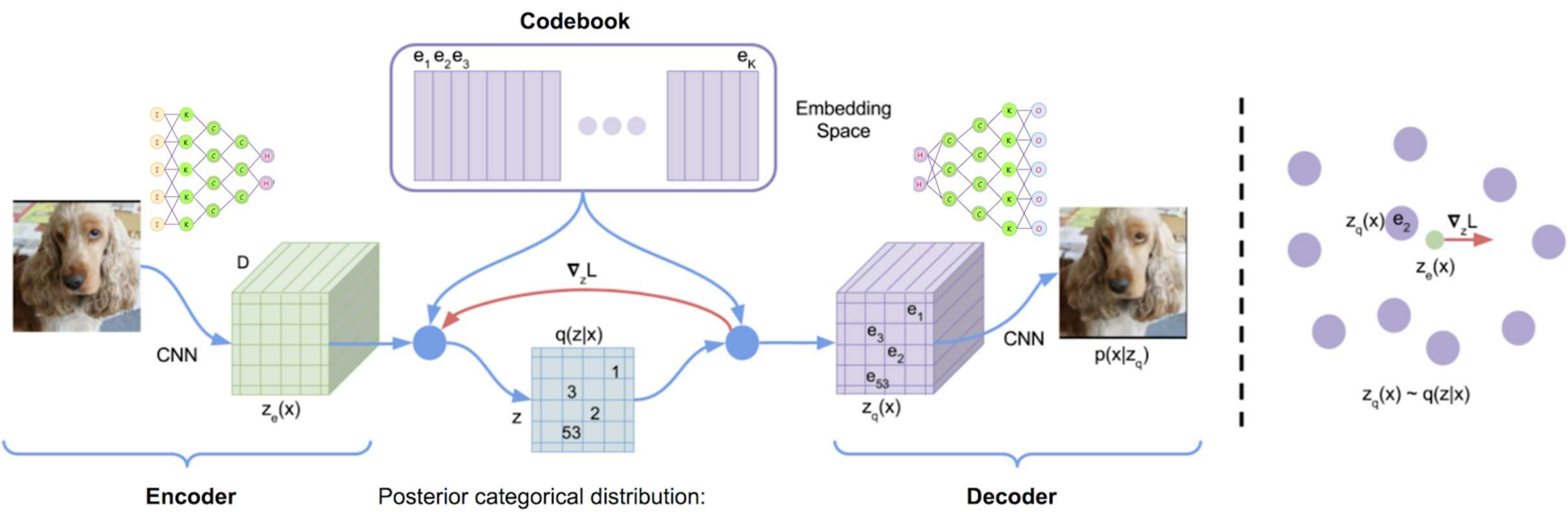
Related Work - VAE problem



Related Work - VAE problem

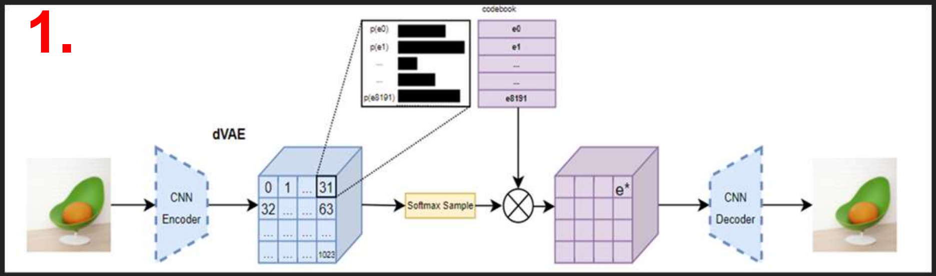


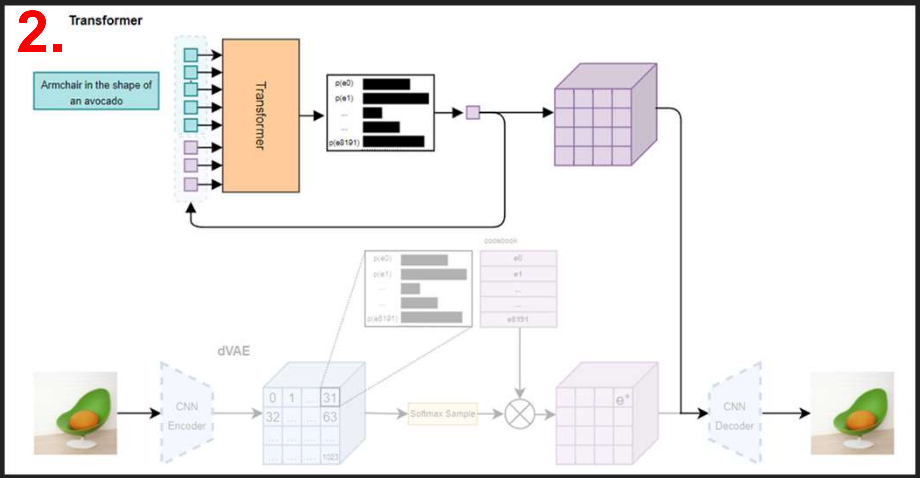
Related Work - VQ-VAE



DALL-E Model

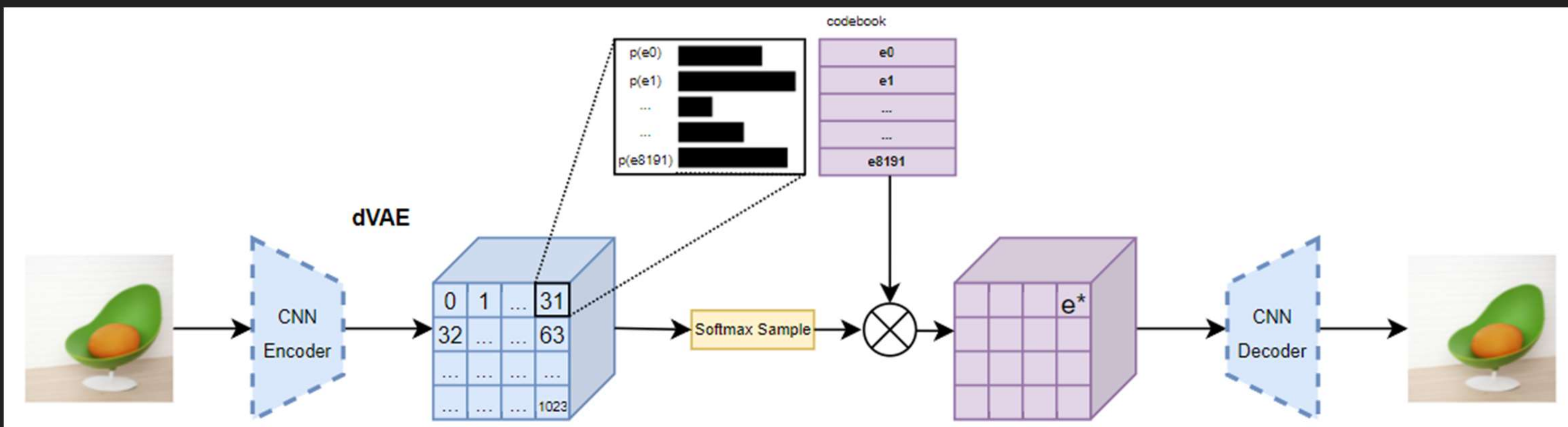
- Transformer to model text and image tokens as single stream of data
  - 2 stage training!





Stage One: Learning the Visual Codebook

- Discrete Variational Autoencoder (dVAE)
  - Similar to VQ-VAE (in VQ-GAN) but uses distribution instead of nearest neighbor

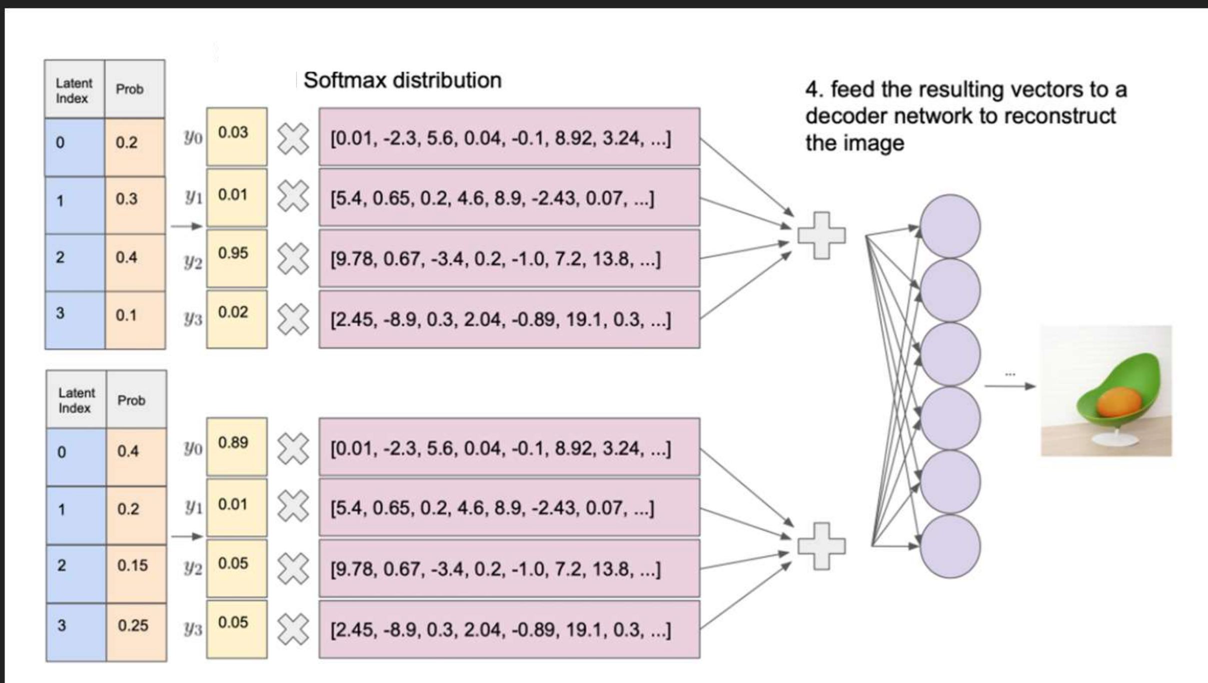


Stage One: Learning the Visual Codebook

- Discrete Variational Autoencoder (dVAE) encoder

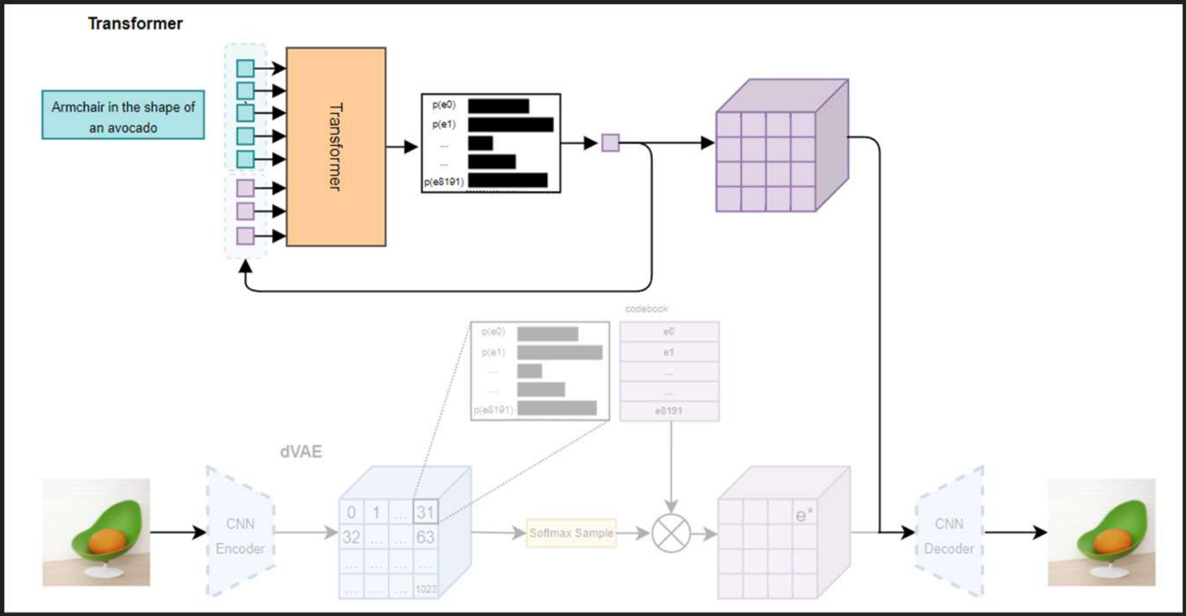
Stage One: Learning the Visual Codebook

- Discrete Variational Autoencoder (dVAE) decoder



Stage Two: Learning Prior Distribution

- Transformer
  - Predict distribution for next token
  - Sample distribution and repeat until 1024 image tokens



Stage Two: Transformer Characteristics

- Transformer
  - BPE-encode lowercase captions into 256 text tokens
    - Vocab size of 16,384
  - 32x32 image tokens
    - Vocab size of 8192
  - 64 attention layers
    - 62 attention heads
  - 12 Billion parameters

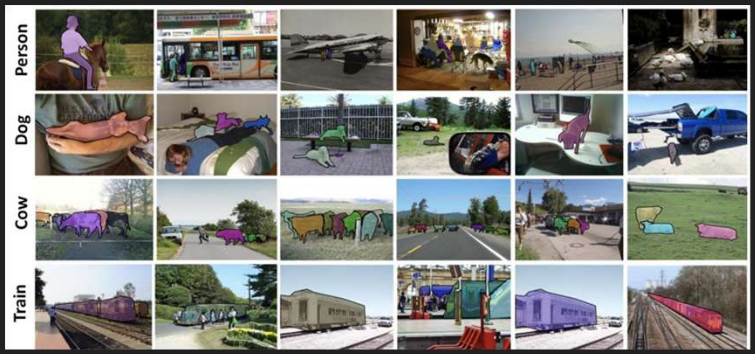
Training: Dataset

- Training Dataset
  - Wikipedia images
  - YFCC100M++

- Filter removed:
  - Small Captions
  - Non-English
  - Dates
  - Extreme Aspect Ratios

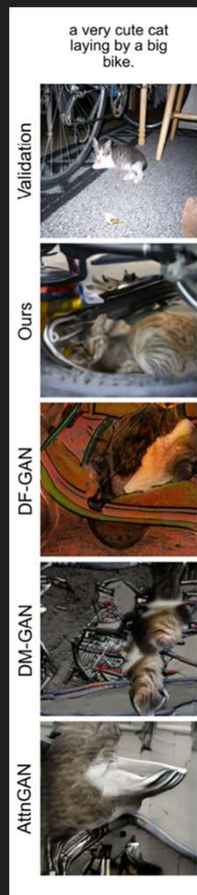
Testing Datasets

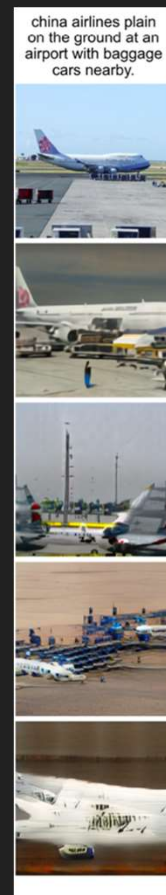
- MS-COCO
  - 328k images
  - object detection, segmentation, key-point detection, captioning



- CUB-200
  - 200 bird species
  - 11,788 images

Example Results





Sample Generation

- CLIP
  - Pre-trained contrastive model
  - Ranks DALLE's generated images
  - Input = image + caption
  - Output = score
  - More images to rank = better quality of best

Learning Transferable Visual Models from Natural Language Supervision

Alec Radford JongWook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, lya Sutskever

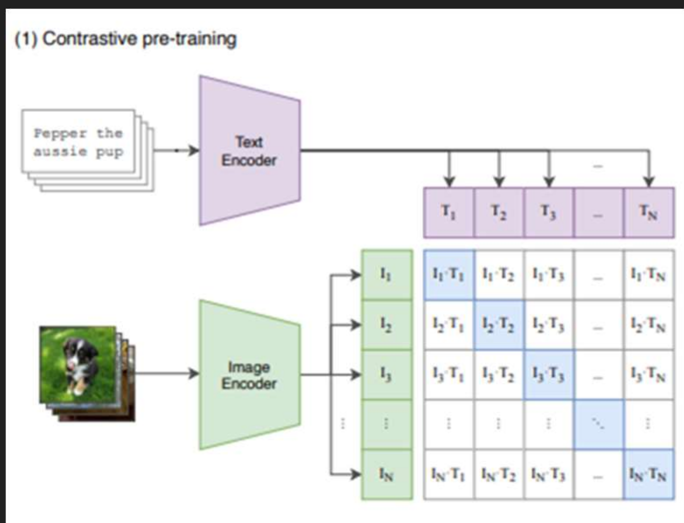
ICML-2021; 3,131 Citations

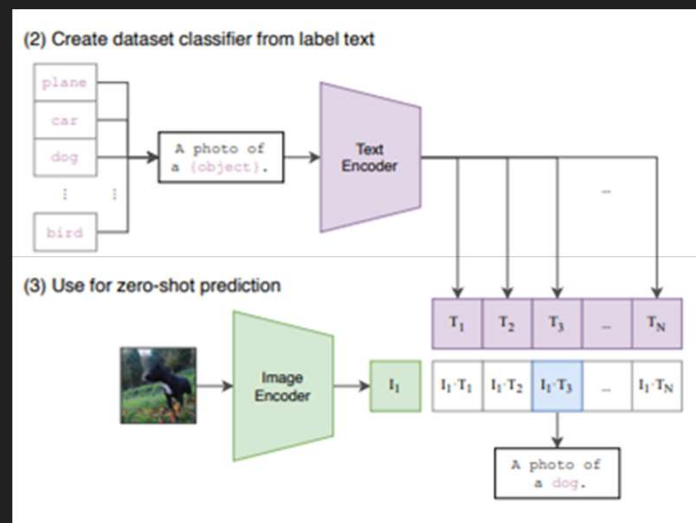
Presented by: Moazam Soomro, Fatemah Najafali, Alec Kerrigan, and Connor Malley
CAP6412;
https://www.youtube.com/watch?v=t5MPdf8NG1g

Contrastive Language Image Pre-training (CLIP)

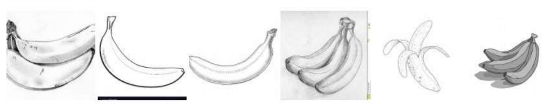
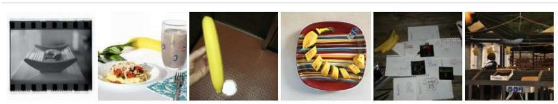
- Mechanism for natural language supervision
- Pair an image with it’s caption using contrastive learning
- Beats fully supervised learning baseline on many datasets
- Can be used as a zero-shot classifier

Contrastive Language Image Pre-training (CLIP)





Zero-shot CLIP is much more robust

DATASET	IMAGENET RESNET101	CLIP VIT-L
 ImageNet	76.2%	76.2%
 ImageNet V2	64.3%	70.1%
 ImageNet Rendition	37.7%	88.9%
 ObjectNet	32.6%	72.3%
 ImageNet Sketch	25.2%	60.2%
 ImageNet Adversarial	2.7%	77.1%

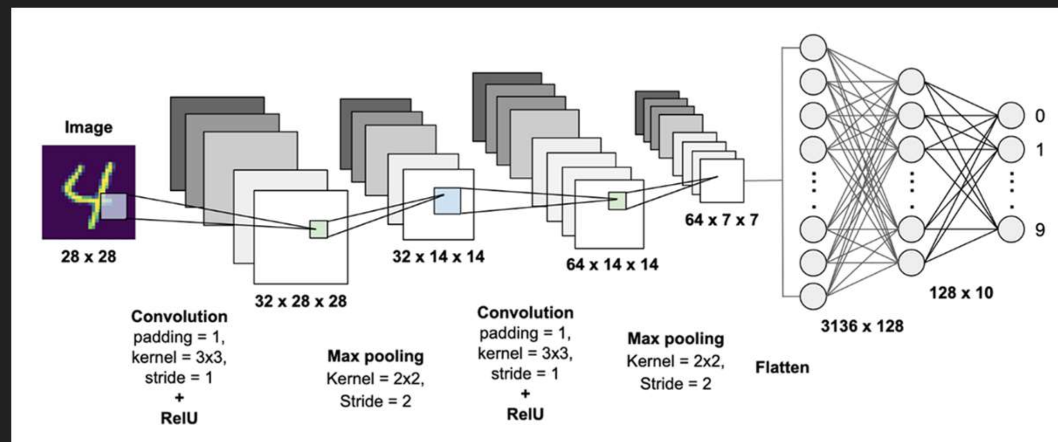
Motivation

- Image classification models are limited:
  - Fixed number of labels
  - Generalization

- CLIP overcomes these limitations.

What is Contrastive Learning?

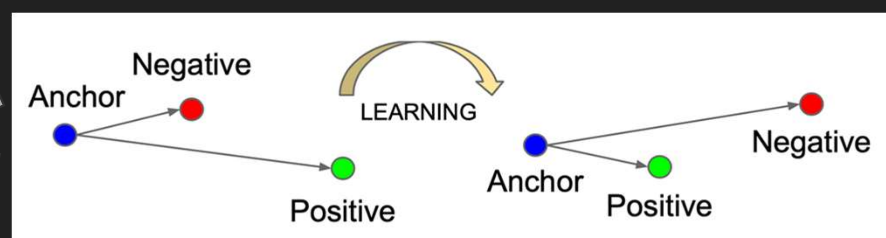
- Classification task:



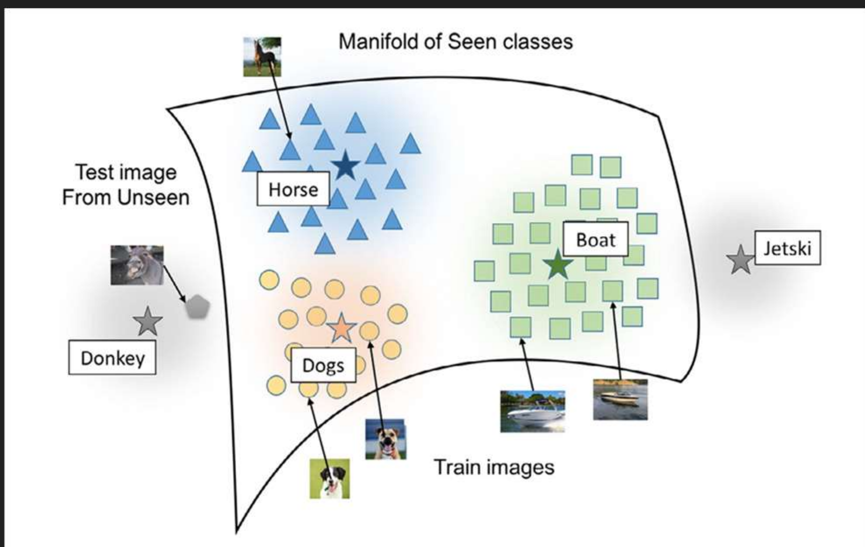
- Contrastive Learning:

N negative samples

Positive Sample



What is Zero-Shot Learning?



- To train on one dataset and generalizing on unseen categories.

WebImageText Dataset

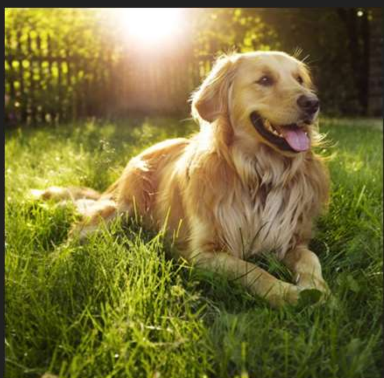
- Motivation for using natural language is the vast amounts of data
- Previous datasets did not have enough natural language descriptions (YFCC100M)
- Authors searched for (image, text) pairs which contained one of 500,000 text queries
- Used for pre-training CLIP

<u>WebImageText (WIT)</u>

400M (image,text) pairs

Up to 20,000 pairs per query

Contrastive Learning Objective - similar (image, text) pair



Input Image

$$\vec{H}_i$$

Image Representation

A dog lying in grass

Input Text

$$\vec{H}_t$$

Text Representation

$$maximize\left(\frac{\vec{H}_i \cdot \vec{H}_t}{\|\vec{H}_i\| \times \|\vec{H}_t\|}\right)$$

Contrastive Learning Objective - dissimilar (image, text) pair

Input Image

$\vec{H}_i$

Image Representation

A dog lying in grass

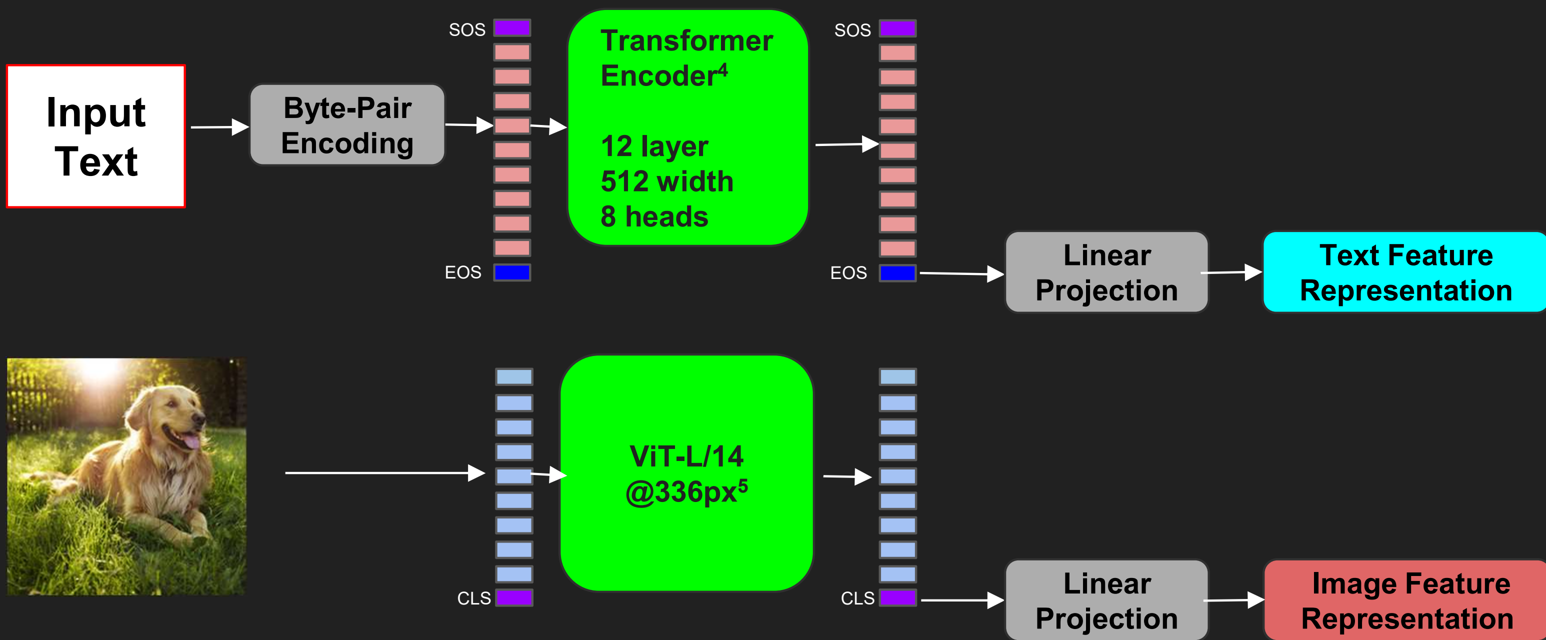
Input Text

$\vec{H}_t$

Text Representation

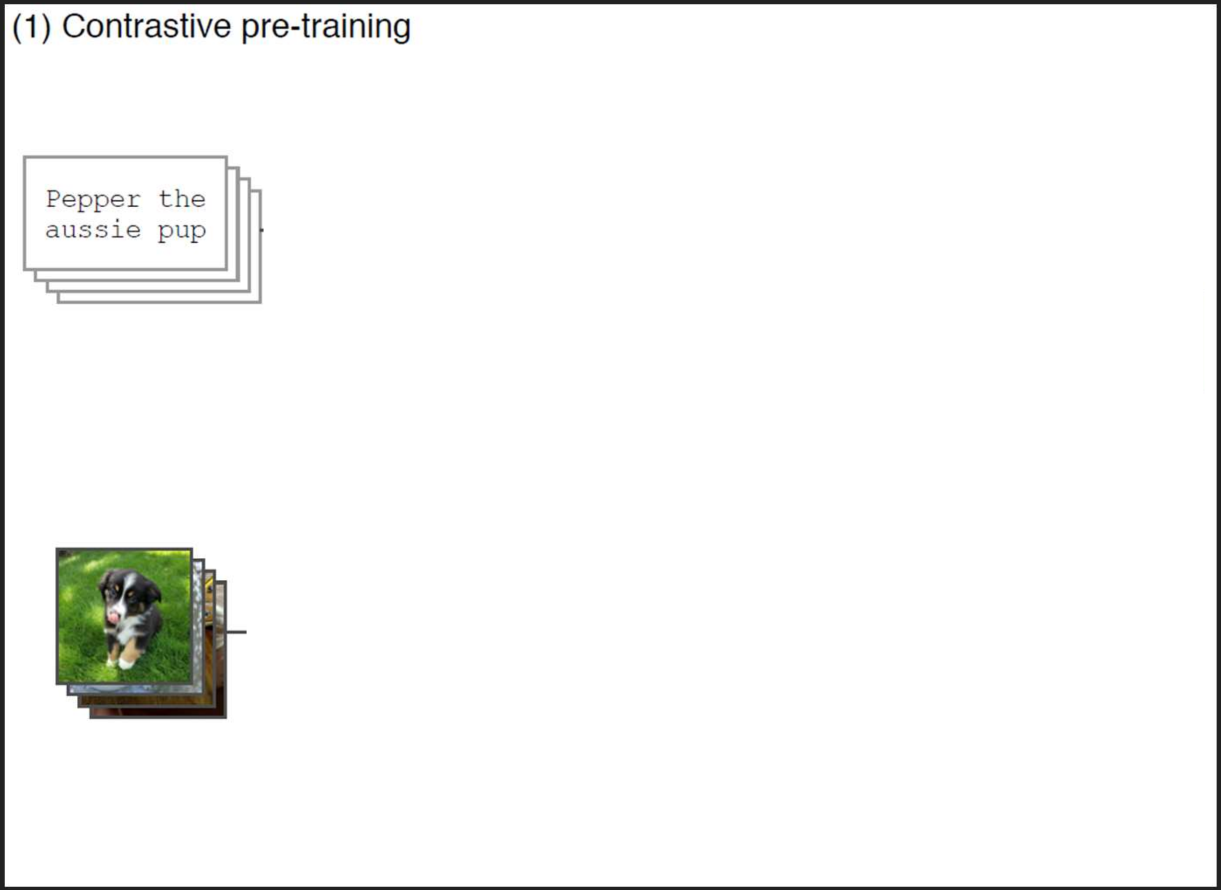
$$minimize\left(\frac{\vec{H}_i \cdot \vec{H}_t}{\|\vec{H}_i\| \times \|\vec{H}_t\|}\right)$$

CLIP Architecture

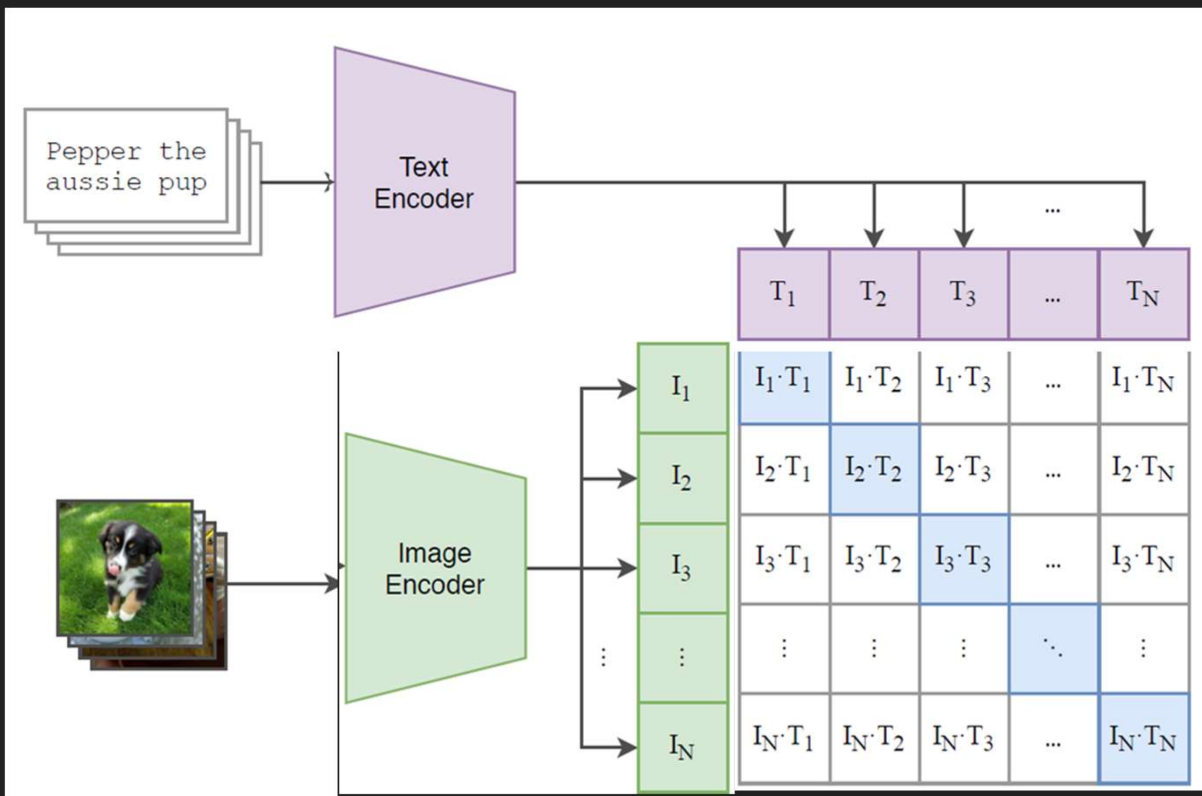


* Authors also tested many other ResNet/ViT variants, but found this ViT to perform the best

CLIP Pre-training

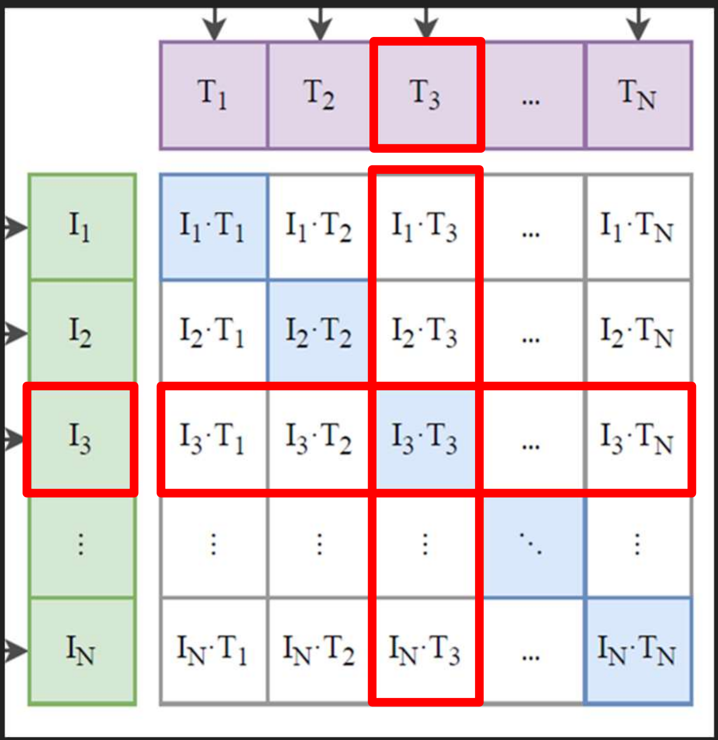


CLIP Pre-training



1. Encode batch of text samples
2. Encode batch of image samples
3. Maximize cosine similarity between correct matches
4. Minmize cosine similarity between incorrect matches

Computing Loss



m_i = one-hot encoded label vector for the i-th image sample

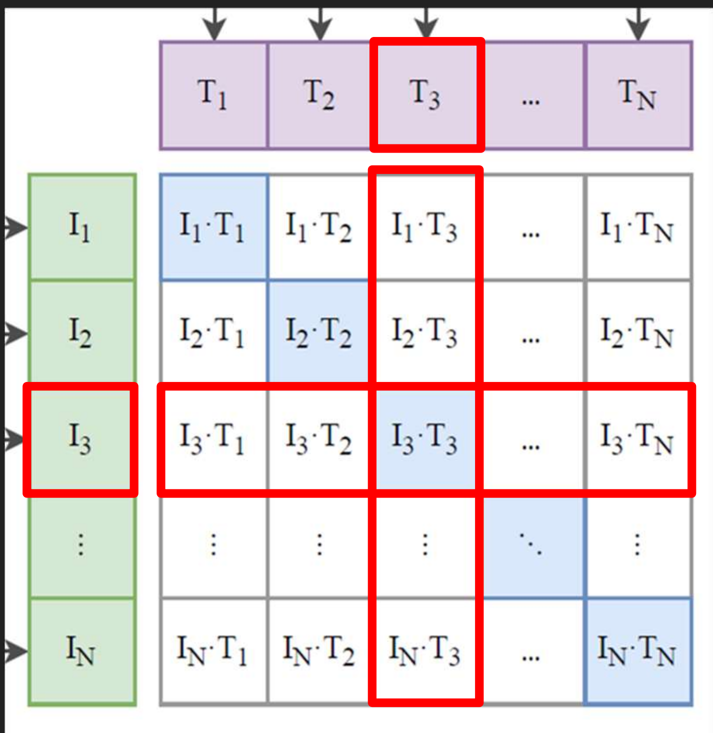
y_i^m = cosine similarities vector for i-th image sample

t_i =one-hot encoded label for the i-th text sample

y_i^t = cosine similarities vector for i-th text sample

ϕ = cross entropy loss

Computing Loss



m_i = one-hot encoded label vector for the i-th image sample

y_i^m = cosine similarities vector for i-th image sample

t_i =one-hot encoded label for the i-th text sample

y_i^t = cosine similarities vector for i-th text sample

ϕ = cross entropy loss

$$\mathcal{L}_m = \frac{\sum_{i=1}^{N} \phi(y_i^m, m_i)}{N} \quad \mathcal{L}_t = \frac{\sum_{i=1}^{N} \phi(y_i^t, t_i)}{N}$$

$$\boxed{\mathcal{L} = \frac{\mathcal{L}_m + \mathcal{L}_t}{2}}$$

Some CLIP details

Training

- Trained on 400M image-text pairs from the internet
- Batch size of 32,768
- 32 epochs over the dataset
- Cosine learning rate decay

Architecture

- ResNet-based or ViT-based image encoder
- Transformer-based text encoder

Testing

- Linear Probe

- Zero-shot Prediction

Linear Probe CLIP

- Train a linear classifier on another dataset using CLIP features

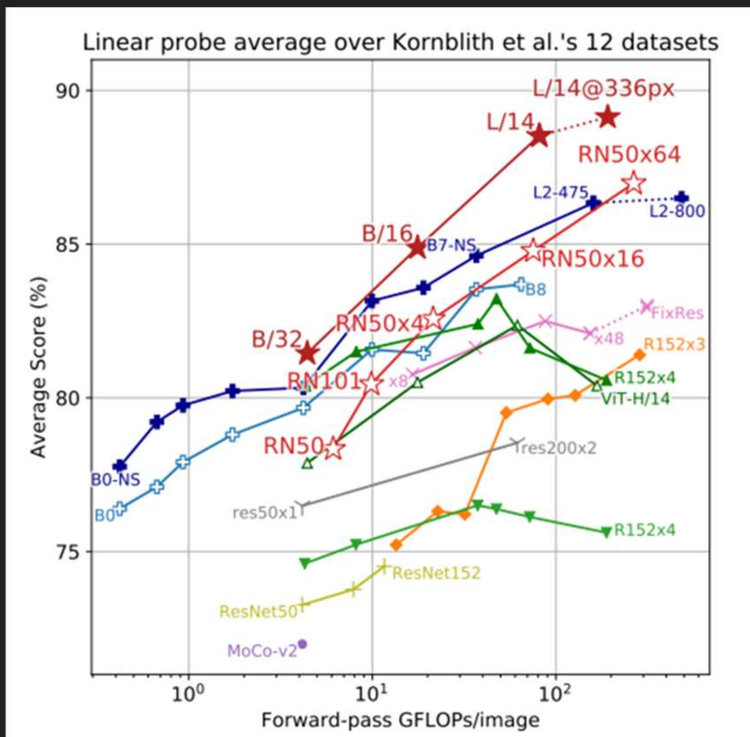
Kornblith et al.'s 12 datasets

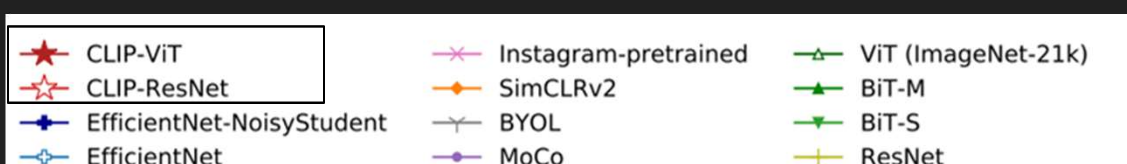
Dataset	Classes	Train size	Test size	Evaluation metric
Food-101	102	75,750	25,250	accuracy
CIFAR-10	10	50,000	10,000	accuracy
CIFAR-100	100	50,000	10,000	accuracy
Birdsnap	500	42,283	2,149	accuracy
SUN397	397	19,850	19,850	accuracy
Stanford Cars	196	8,144	8,041	accuracy
FGVC Aircraft	100	6,667	3,333	mean per class
Pascal VOC 2007 Classification	20	5,011	4,952	11-point mAP
Describable Textures	47	3,760	1,880	accuracy
Oxford-IIIT Pets	37	3,680	3,669	mean per class
Caltech-101	102	3,060	6,085	mean-per-class
Oxford Flowers 102	102	2,040	6,149	mean per class

Extended 27 Datasets

Dataset	Classes	Train size	Test size	Evaluation metric
Food-101	102	75,750	25,250	accuracy
CIFAR-10	10	50,000	10,000	accuracy
CIFAR-100	100	50,000	10,000	accuracy
Birdsnap	500	42,283	2,149	accuracy
SUN397	397	19,850	19,850	accuracy
Stanford Cars	196	8,144	8,041	accuracy
FGVC Aircraft	100	6,667	3,333	mean per class
Pascal VOC 2007 Classification	20	5,011	4,952	11-point mAP
Describable Textures	47	3,760	1,880	accuracy
Oxford-IIIT Pets	37	3,680	3,669	mean per class
Caltech-101	102	3,060	6,085	mean-per-class
Oxford Flowers 102	102	2,040	6,149	mean per class
MNIST	10	60,000	10,000	accuracy
Facial Emotion Recognition 2013	8	32,140	3,574	accuracy
STL-10	10	1000	8000	accuracy
EuroSAT	10	10,000	5,000	accuracy
RESISC45	45	3,150	25,200	accuracy
GTSRB	43	26,640	12,630	accuracy
KITTI	4	6,770	711	accuracy
Country211	211	43,200	21,100	accuracy
PatchCamelyon	2	294,912	32,768	accuracy
UCF101	101	9,537	1,794	accuracy
Kinetics700	700	494,801	31,669	mean(top1, top5)
CLEVR Counts	8	2,000	500	accuracy
Hateful Memes	2	8,500	500	ROC AUC
Rendered SST2	2	7,792	1,821	accuracy
ImageNet	1000	1,281,167	50,000	accuracy

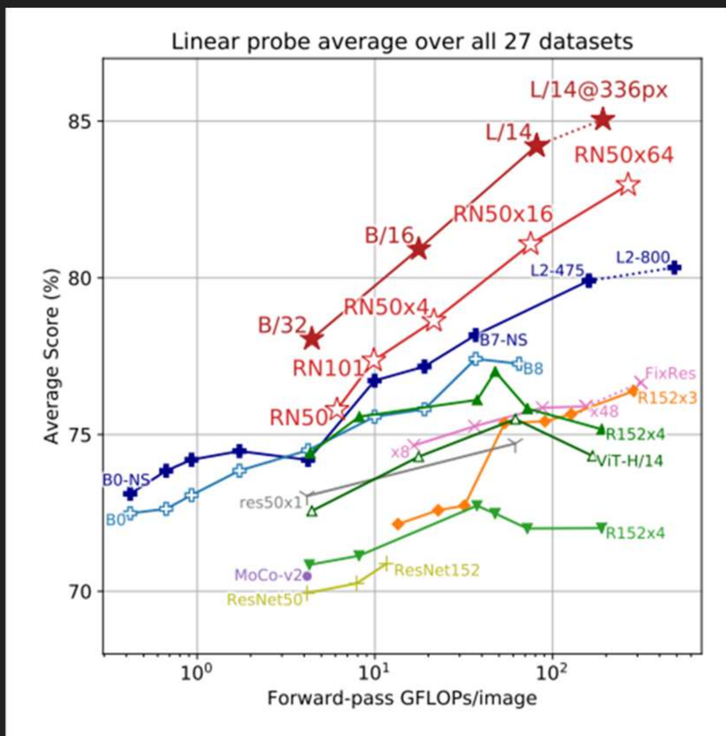
Results - Efficiency - Kornblith

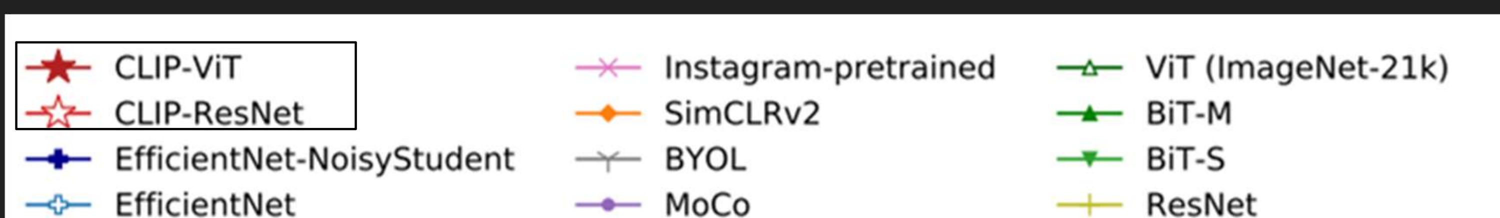




- Kornblith 12 dataset evaluation suite, standard for most works
- CLIP's ResNet based model underperforms EfficientNet
- ViT based CLIP outperforms everything

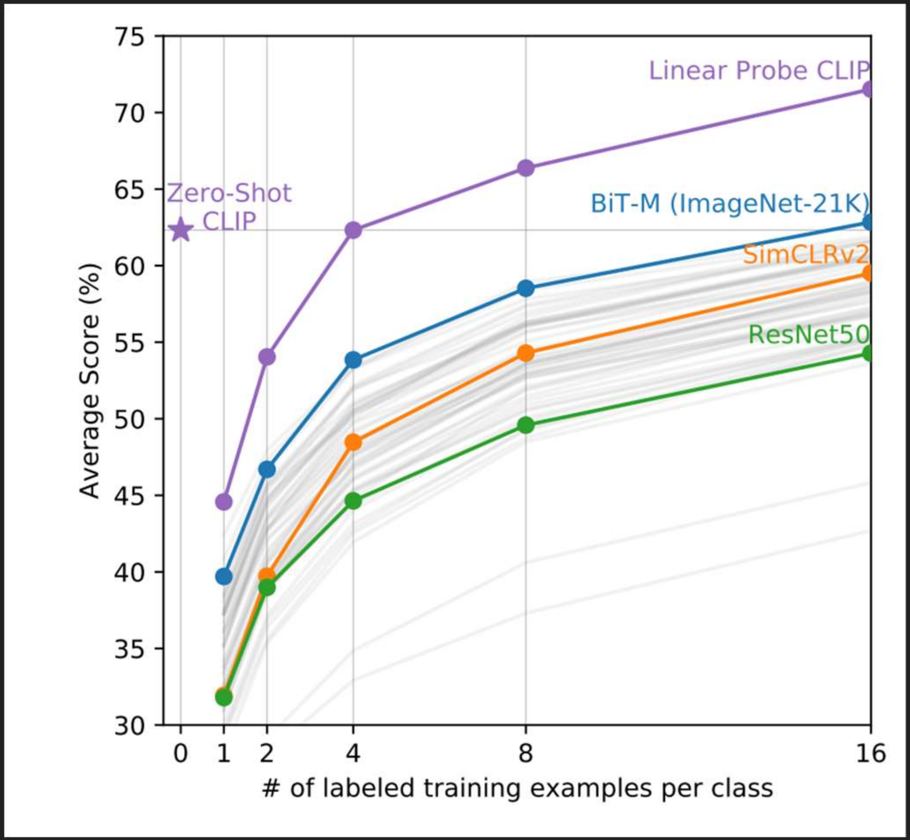
Results - Efficiency - Extended





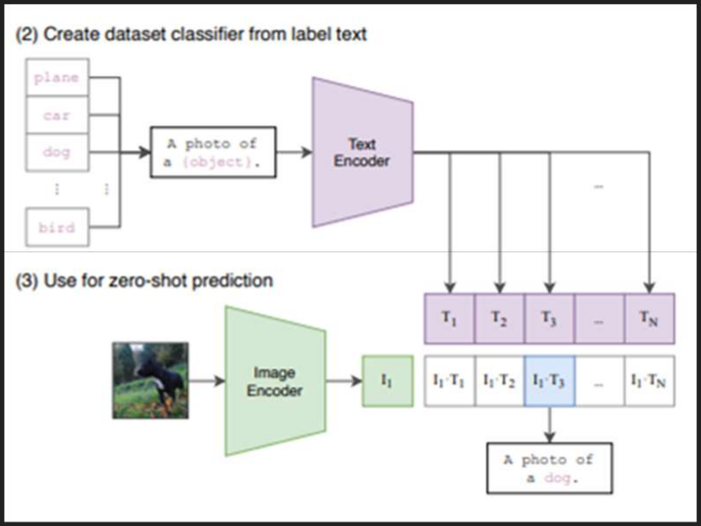
- On the extended testing suite, both CLIP versions outperform all other models
- Performance gap increases with GFLOPS

Results - Low-Shot

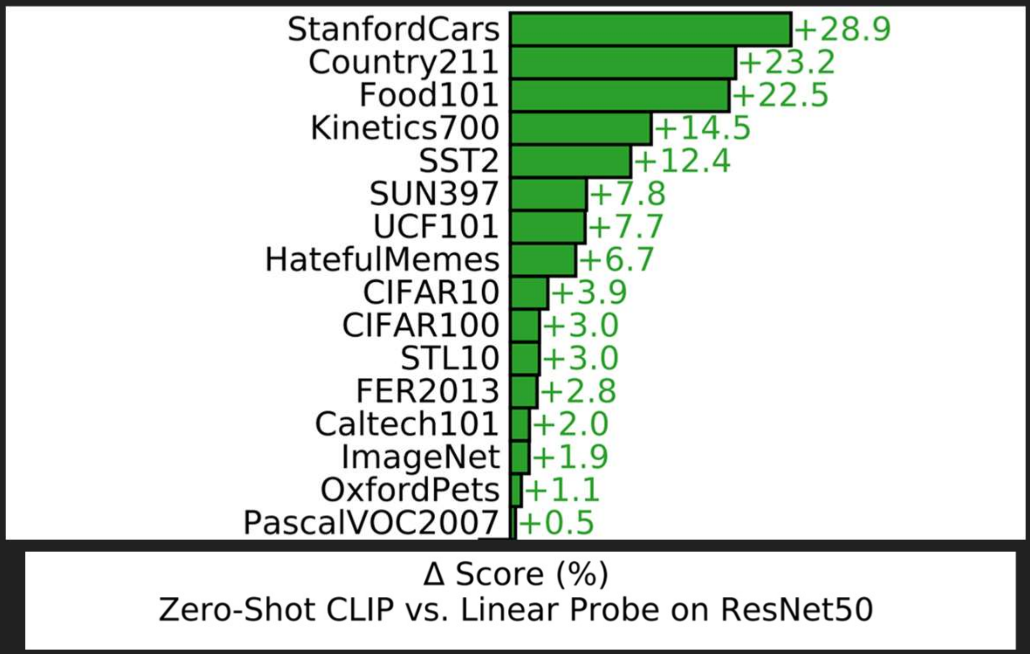


- CLIP scales well
- Linear-Probe CLIP climbs
- ResNet and other methods flatten
- Zero-Shot CLIP outperforms all non-CLIP methods up until 16 shot

Contrastive Language Image Pre-training (CLIP)

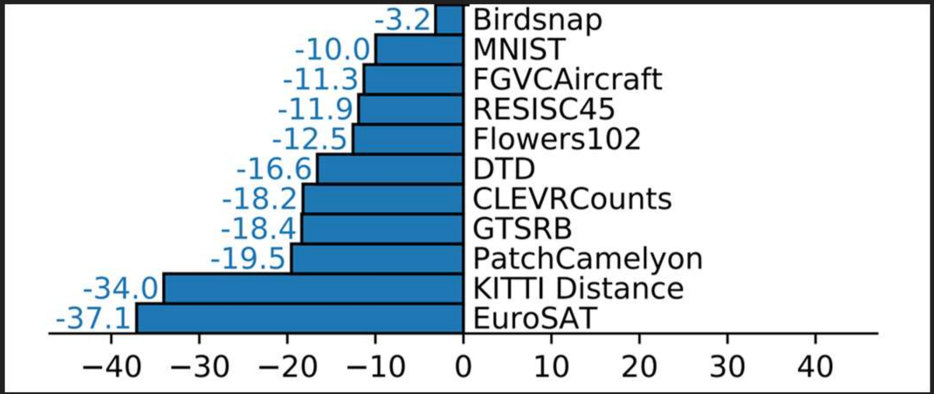


Results - Accuracy



- Zero-shot CLIP using ResNet50 backbone is compared to off the shelf ResNet50
- CLIP outperforms on a wide variety of popular datasets
- For video, a single frame was sampled

Results - Accuracy



Δ Score (%)
Zero-Shot CLIP vs. Linear Probe on ResNet50

- Underperforms on many other datasets
- Mostly on specialized/complex datasets
- EuroSAT for satellite images, Tumor classification
- Makes intuitive sense, Zero-shot CLIP is highly generalized
- Not suited for hyper specific tasks unless fine-tuned

Thank You