ELTE | IK
FACULTY OF INFORMATICS

DEPARTMENT OF
ARTIFICIAL
INTELLIGENCE

NIPG

CLIP

Connecting Text and Images

OpenAI

By Bruno Melício

brunomelicio.ai@gmail.com

Budapest, February 8, 2023

Content

1. Introduction
2. What is CLIP?
3. Background knowledge
4. How does CLIP work?
5. Zero-Shot Classification
6. Experiments and Results
7. Limitations
8. Applications
9. Conclusion

1. Introduction

Can AI create art?

DALL-E 2 can!

An astronaut riding a horse in photo realistic style

→

DALL·E 2

→

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). Hierarchical Text-Conditional Image Generation with CLIP Latents. ArXiv, abs/2204.06125.

1. Introduction

Can AI create art?

Imagen can!

A cute corgi lives in a house made out of sushi.

→ **Imagen** →

Saharia, C. et al. (2022). Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. ArXiv, abs/2205.11487.

1. Introduction

Can AI create art?

Dreambooth can!

A photo / painting of [Bruno] in the style of [Simpsons]

→ **Dreambooth** →

Ruiz, Nataniel et al. “DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation.” ArXiv abs/2208.12242 (2022)

2. What is CLIP?

All of the previous models are based on **CLIP**.

CLIP: **C**ontrastive **L**anguage-**I**mage **P**re-training

CLIP is a method trained to match an image with a text that describes it.

2. What is CLIP?

Given an image

and a set of descriptions (prompts):

- ‘a boy with a toy’
- ‘a cat with a ball’
- ‘a dog with a tennis ball’

CLIP finds which prompt best describes the image. (animation)

2. What is CLIP?

Not to be confused with image captioning task!

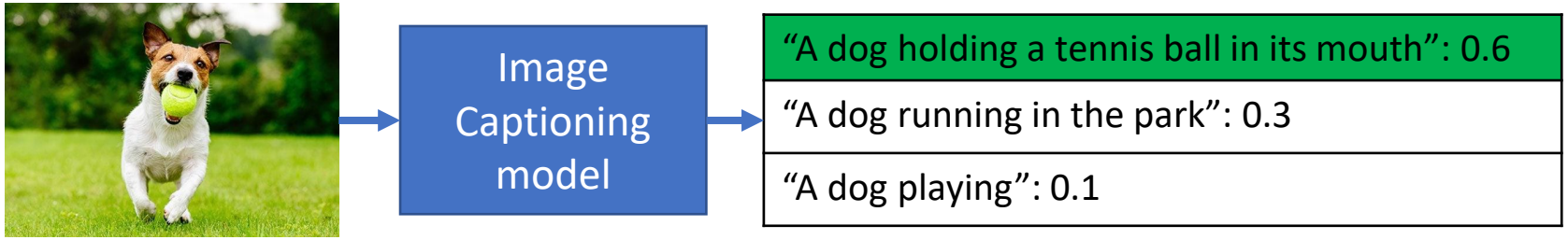
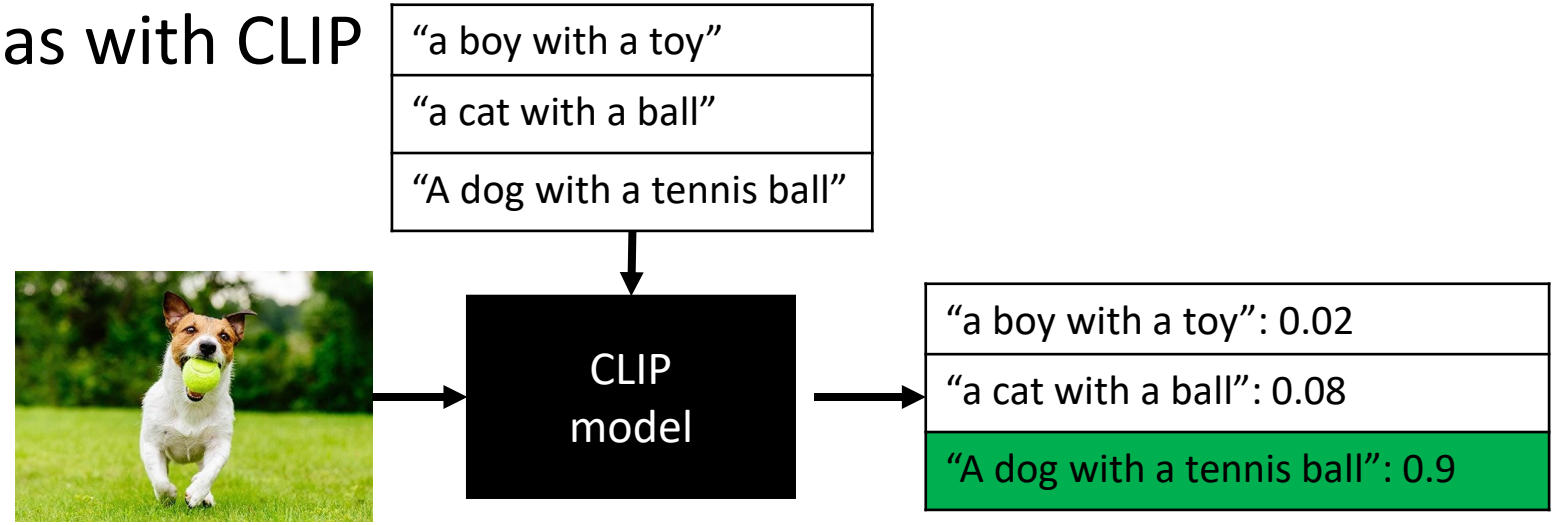


Image Captioning: given an image, predict a description of the image.

Whereas with CLIP

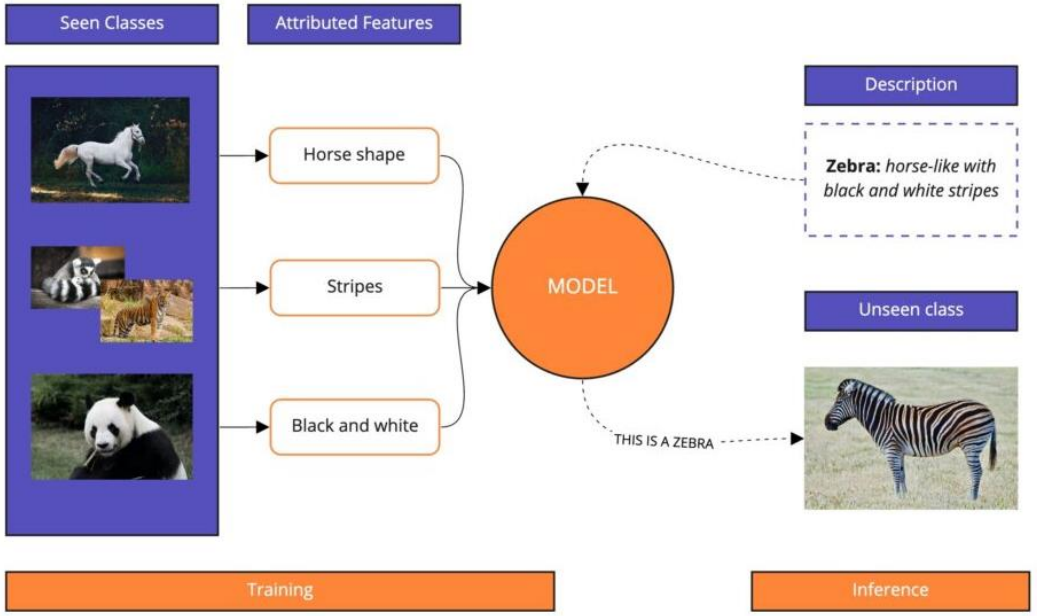


CLIP: given an image and a set of prompts, CLIP finds the prompt that best describes the image.

3. Background knowledge

In order to better understand CLIP, we should be familiar with the following concepts:

- **Prompt**: is a sentence/caption that describes an image
- **Zero-Shot learning**: is a Machine Learning paradigm where a pre-trained model is used to evaluate test data of classes that have not been used during training. It means you need zero training samples for a model to adapt to a new domain



3. Background knowledge

In order to better understand CLIP, we should be familiar with the following concepts:

- **Linear Probing**
- Kumar, Ananya et al. "Fine-Tuning can Distort Pretrained Features and Underperform Out-of-Distribution." ArXiv abs/2202.10054 (2022)

Published as a conference paper at ICLR 2022

FINE-TUNING CAN DISTORT PRETRAINED FEATURES AND UNDERPERFORM OUT-OF-DISTRIBUTION

Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, Percy Liang
Stanford University, Computer Science Department

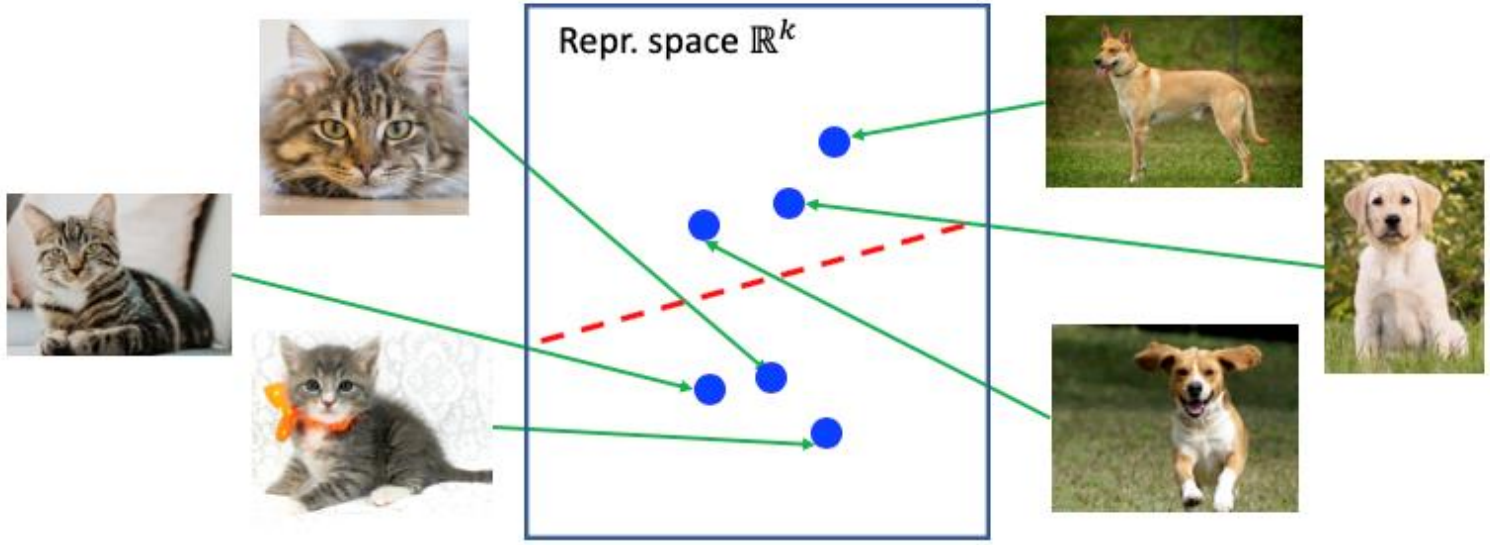
ABSTRACT

When transferring a pretrained model to a downstream task, two popular methods are full fine-tuning (updating all the model parameters) and linear probing (updating only the last linear layer—the "head"). It is well known that fine-tuning leads to better accuracy in-distribution (ID). However, in this paper, we find that fine-tuning can achieve worse accuracy than linear probing out-of-distribution (OOD) when the pretrained features are good and the distribution shift is large. On 10 distribution shift datasets (BREEDS-Living17, BREEDS-Entity30, DomainNet, CIFAR → STL, CIFAR-10.1, FMoW, ImageNetV2, ImageNet-R, ImageNet-A, ImageNet-Sketch), fine-tuning obtains on average 2% higher accuracy ID but 7% lower accuracy OOD than linear probing. We show theoretically that this tradeoff between ID and OOD accuracy arises even in a simple setting: fine-tuning

3. Background knowledge

In order to better understand CLIP, we should be familiar with the following concepts:

- **Contrastive learning:** it is a self-supervised technique that teaches a model to differentiate data samples that are similar to each other from dissimilar ones. It maximizes the similarity score of samples that share features with each other, and minimizes the similarity of samples that don't.



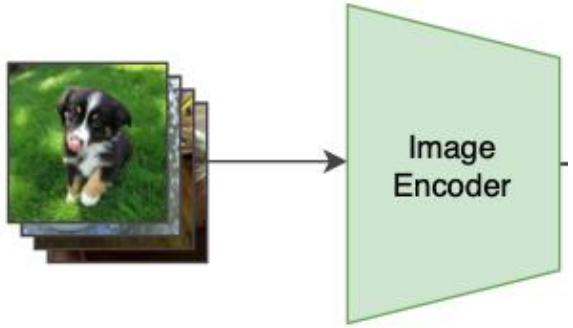
4. How does CLIP work?

Given a batch of images

4. How does CLIP work?

Given a batch of images:

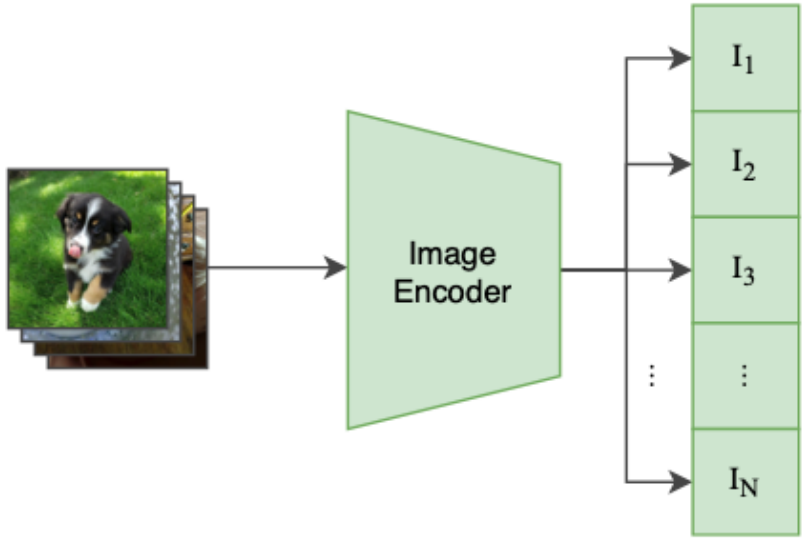
- Send them to an Image Encoder that creates a vector representation of the images (embeddings)
- Image Encoder:
  - Modified version of ResNet architectures (50 and 101) and use EfficientNet style to scale to use more computation power.
  - Vision Transformers with minor modifications (ViT-B/32, a ViT-B/16, and a ViT-L/14)



4. How does CLIP work?

Given a batch of images:

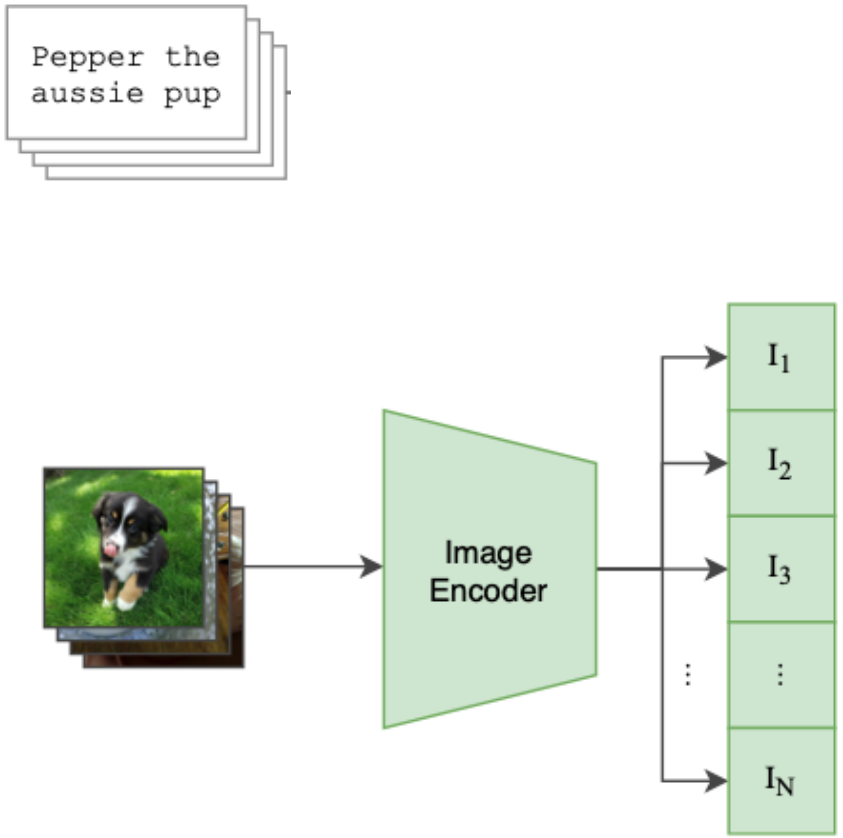
- Send them to an Image Encoder that creates a vector representation of the images (embeddings)
- All the embeddings are stacked together (n x d), where n is the number of images in the batch and d is the embedding dimension



4. How does CLIP work?

Given a batch of corresponding prompts:

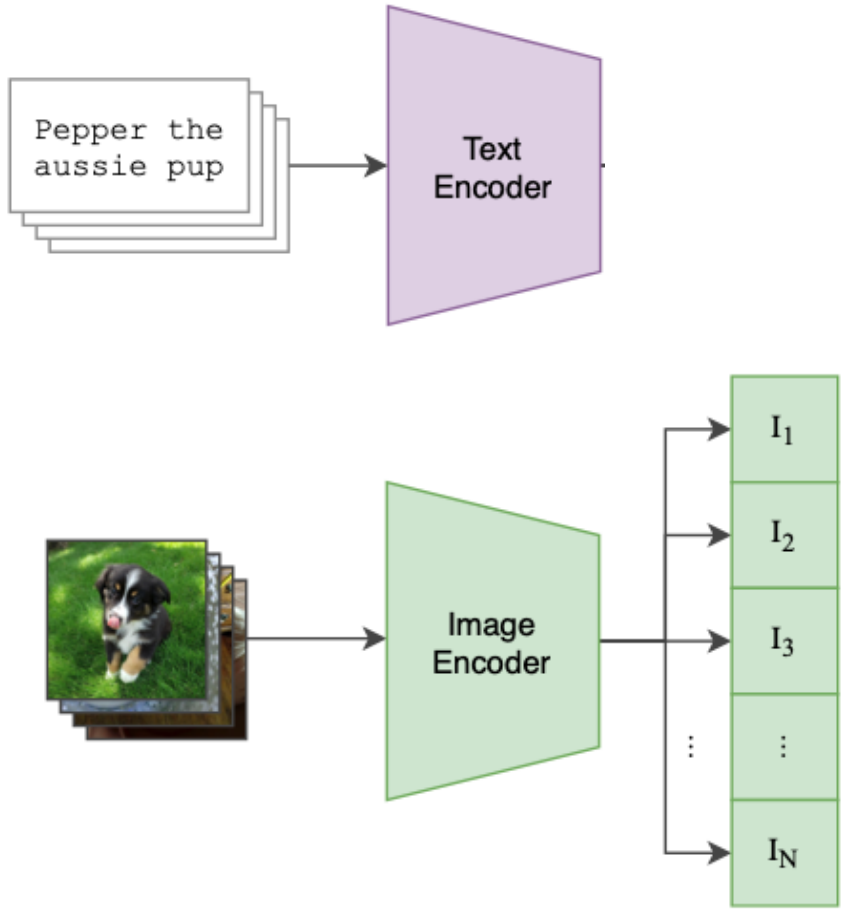
- 1st prompt corresponds to the 1st image, 2nd prompt to the 2nd image, and so on



4. How does CLIP work?

Given a batch of corresponding prompts:

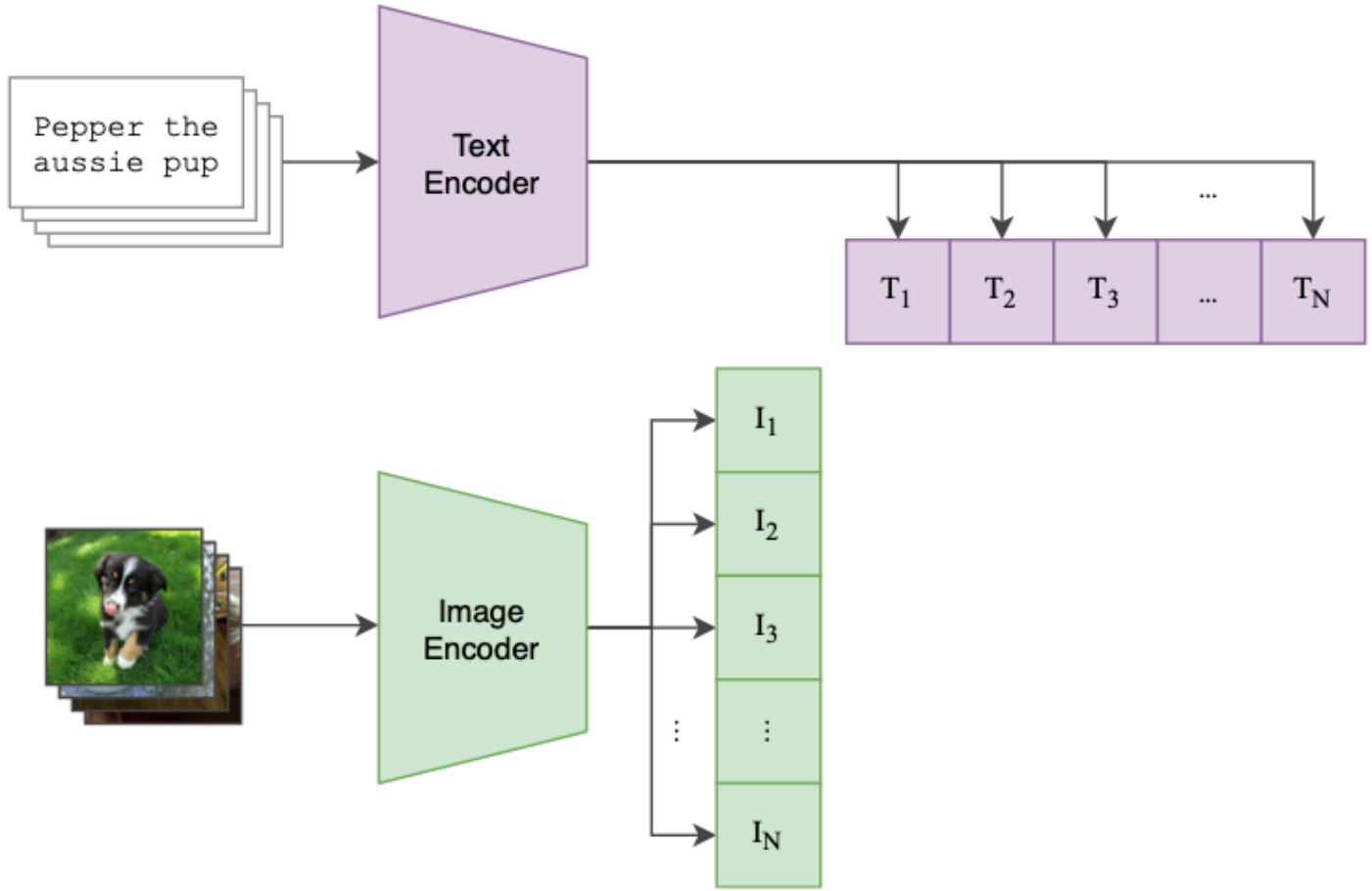
- 1st prompt corresponds to the 1st image, 2nd prompt to the 2nd image, and so on
- Send them to a Text Encoder that creates a vector representation of the text (embeddings)
- Text Encoder is a Transformer (Vaswani et al., 2017)



4. How does CLIP work?

Given a batch of corresponding prompts:

- 1st prompt corresponds to the 1st image, 2nd prompt to the 2nd image, and so on
- Send them to a Text Decoder that creates a vector representation of the text (embeddings)
- All the embeddings are stacked together (n x d), where n is the number of prompts in the batch and d is the embedding dimension
- The text embeddings are transposed for multiplication, so (d x n)

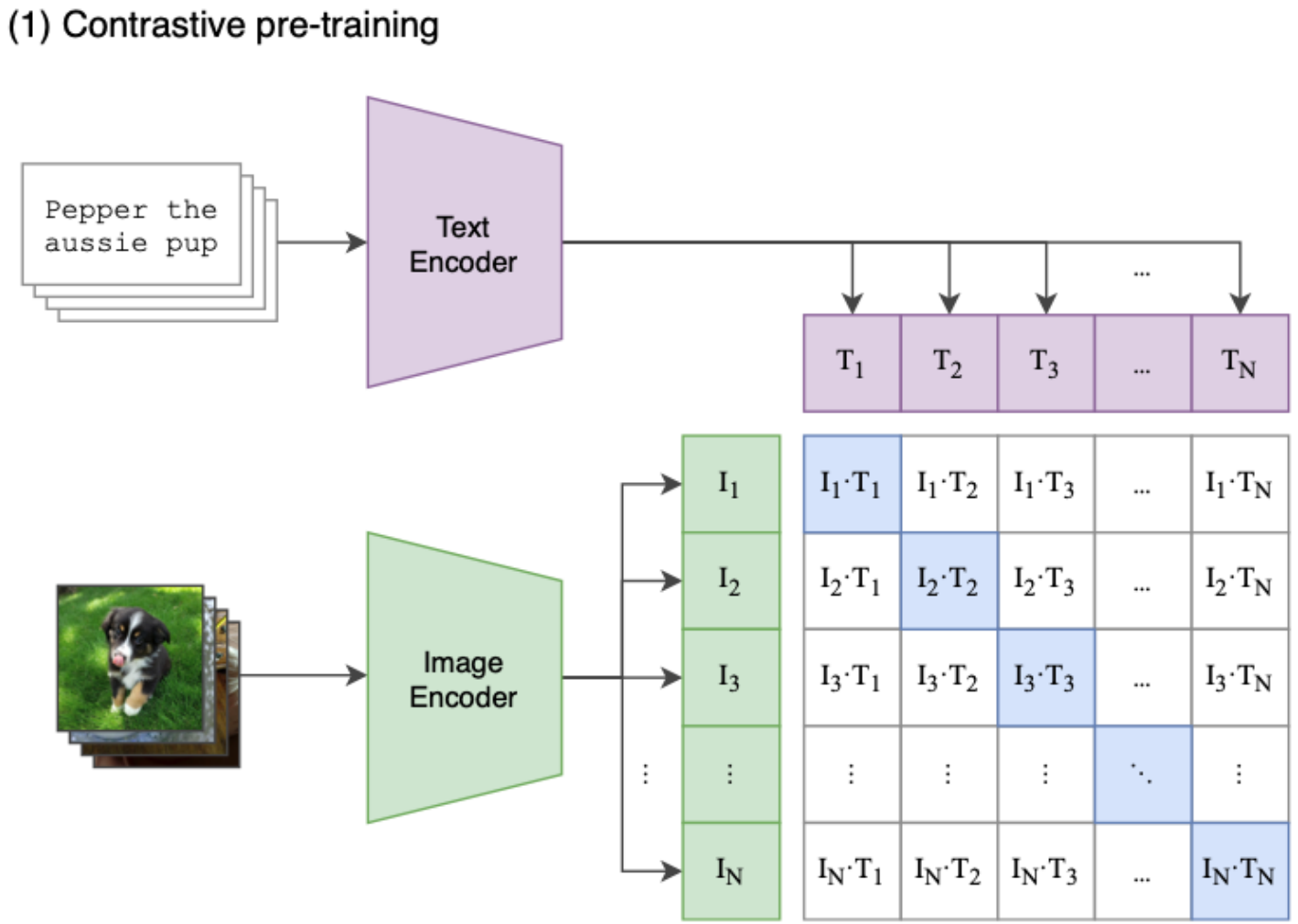


4. How does CLIP work?

Once the image embeddings and text embeddings are created, they use contrastive learning.

CLIP computes the pairwise cosine similarities between the image and the text embeddings.

- multiply those matrices and calculate the pairwise cosine similarities between every image and text description. This produces an N X N matrix
- The cosine similarities of the correct <image-text> embedding pairs <I1,T1>, <I2,T2> (where i=j) are maximized.
- In a contrastive fashion, the cosine similarities of dissimilar pairs <I1,T2>, <I1,T3>… <Ii,Tj> (where i≠j) are minimized.
- The text prompt with the highest similarity is chosen as the prediction.



4. How does CLIP work?

More details

- The loss function optimizes both the image and text encoder end-to-end.
- CLIP learns a multi-modal embedding space by jointly training an image and text encoder

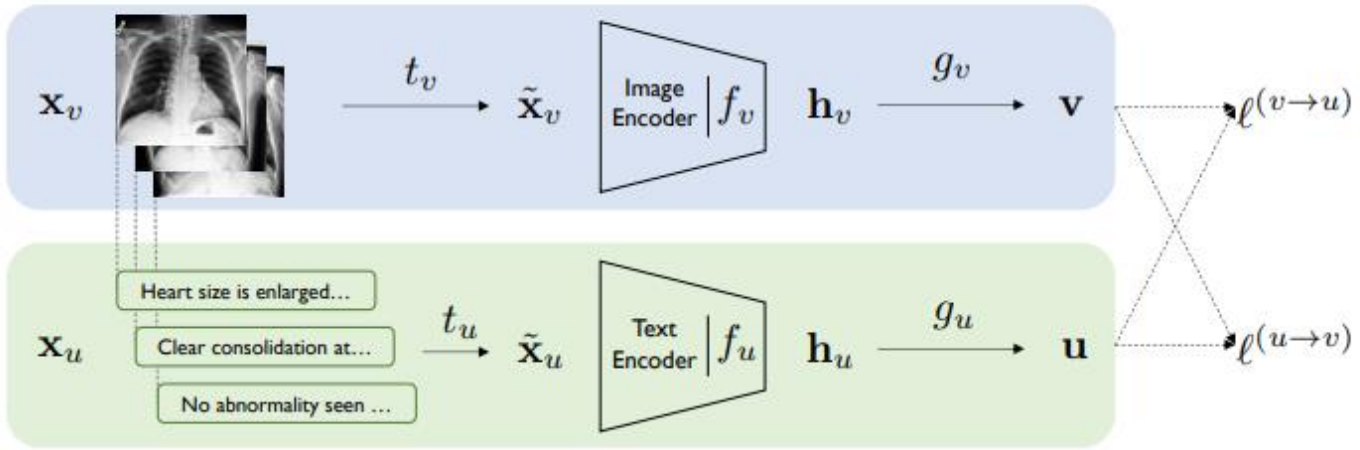


Figure 2: Overview of our ConVIRT framework. The blue and green shades represent the image and text encoding pipelines, respectively. Our method relies on maximizing the agreement between the true image-text representation pairs with bidirectional losses $\ell^{(v\to u)}$ and $\ell^{(u\to v)}$.

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

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

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

Zhang, Yuhao et al. “Contrastive Learning of Medical Visual Representations from Paired Images and Text.” ArXiv abs/2010.00747 (2020): n. pag.

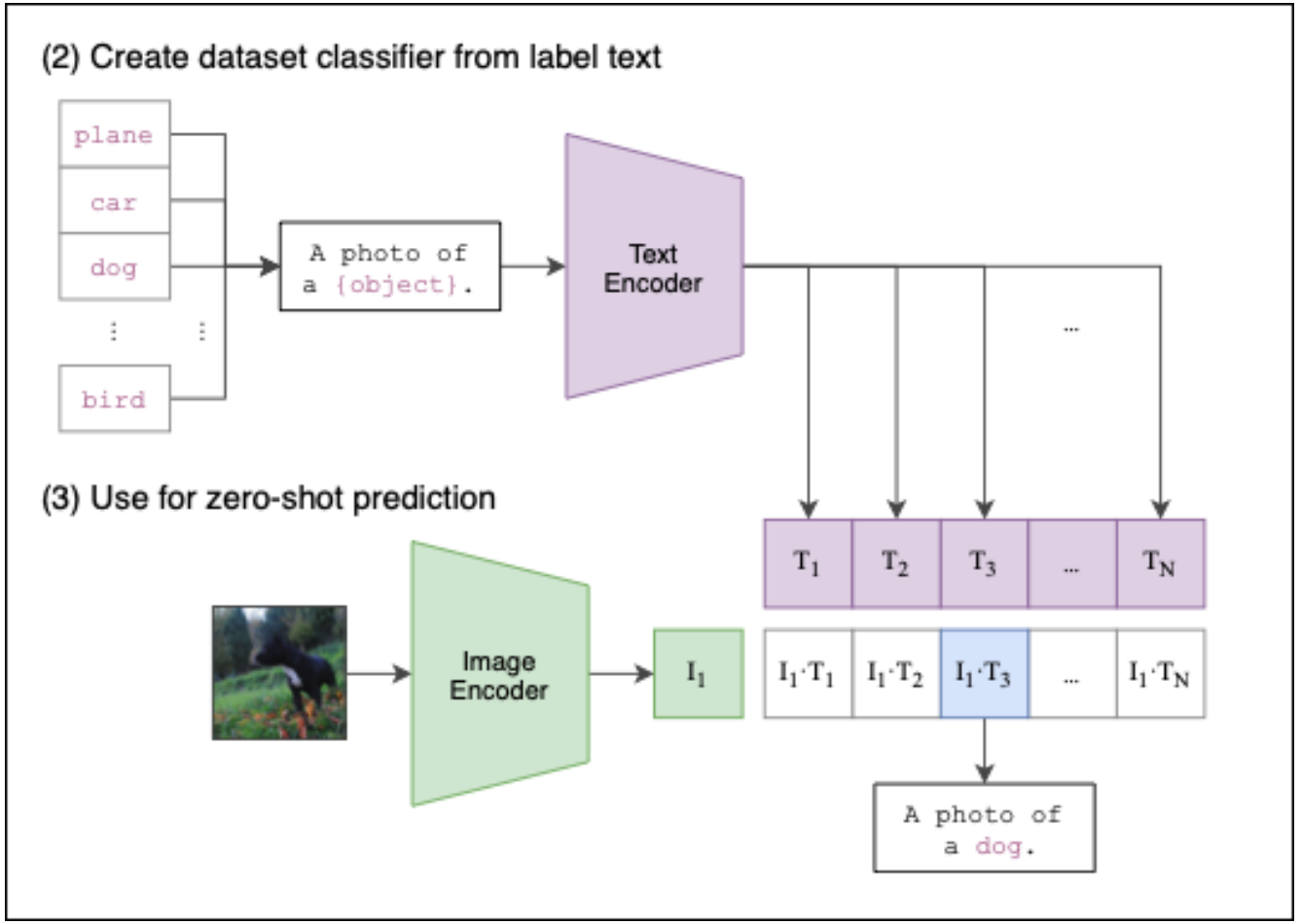
4. How does CLIP work?

More details

- We train CLIP from scratch without initializing the image encoder with ImageNet weights or the text encoder with pre-trained weights.
- A random square crop from resized images is the only data augmentation used during training.
- CLIP is trained using a staggering amount of 400 million image-text pairs. For comparison, the ImageNet dataset contains 1.2 million images.
- The final tuned CLIP model was trained on 256 V100 GPUs for two weeks. For an on-demand training on AWS Sagemaker, this would cost at least 200k dollars!
- The model uses a minibatch of 32,768 images for training
- To increase compute they follow the EfficientNet approach using a simple baseline of allocating additional compute equally to increasing the width, depth, and resolution of the model
- We also replace the global average pooling layer of the ResNet with an attention pooling mechanism. The attention pooling is implemented as a single layer of “transformer-style” multi-head QKV attention where the query is conditioned on the global average-pooled representation of the image

5. Zero shot classification

- CLIP pre-trains an image encoder and a text encoder to predict which images were paired with which texts in our dataset.
- We then use this behavior to turn CLIP into a zero-shot classifier. We convert all of a dataset's classes into captions such as "a photo of a dog" and predict the class of the caption CLIP estimates best pairs with a given image.



6. Experiments and Results

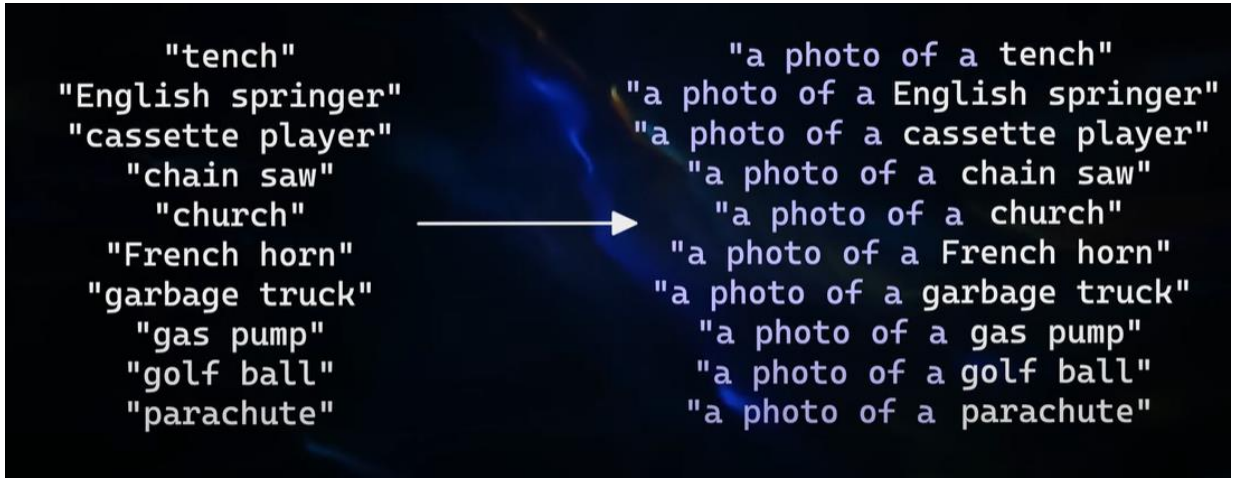
- Check paper
- Figures 2-8

7. Limitations

- While CLIP usually performs well on recognizing common objects, it struggles on more abstract or systematic tasks such as counting the number of objects in an image and on more complex tasks such as predicting how close the nearest car is in a photo.
- Zero-shot CLIP also struggles compared to task specific models on very fine-grained classification, such as telling the difference between car models, variants of aircraft, or flower species.
- CLIP also still has poor generalization to images not covered in its pre-training dataset. For instance, although CLIP learns a capable OCR system, when evaluated on handwritten digits from the MNIST dataset, zero-shot CLIP only achieves 88% accuracy, well below the 99.75% of humans on the dataset.
- Finally, we've observed that CLIP's zero-shot classifiers can be sensitive to wording or phrasing and sometimes require trial and error "prompt engineering" to perform well.

8. Applications (Image Classification)

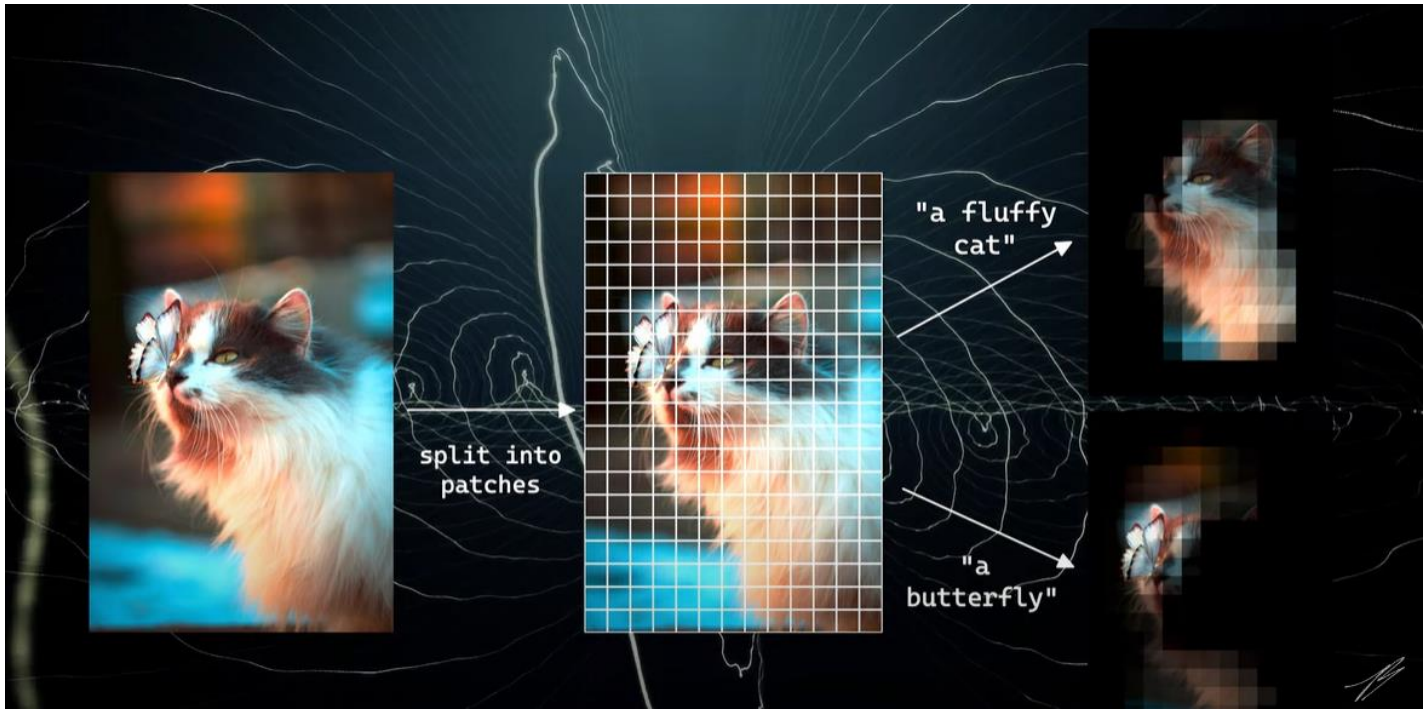
- After clip is trained , it can also be adapted to other tasks with zero-shot learning



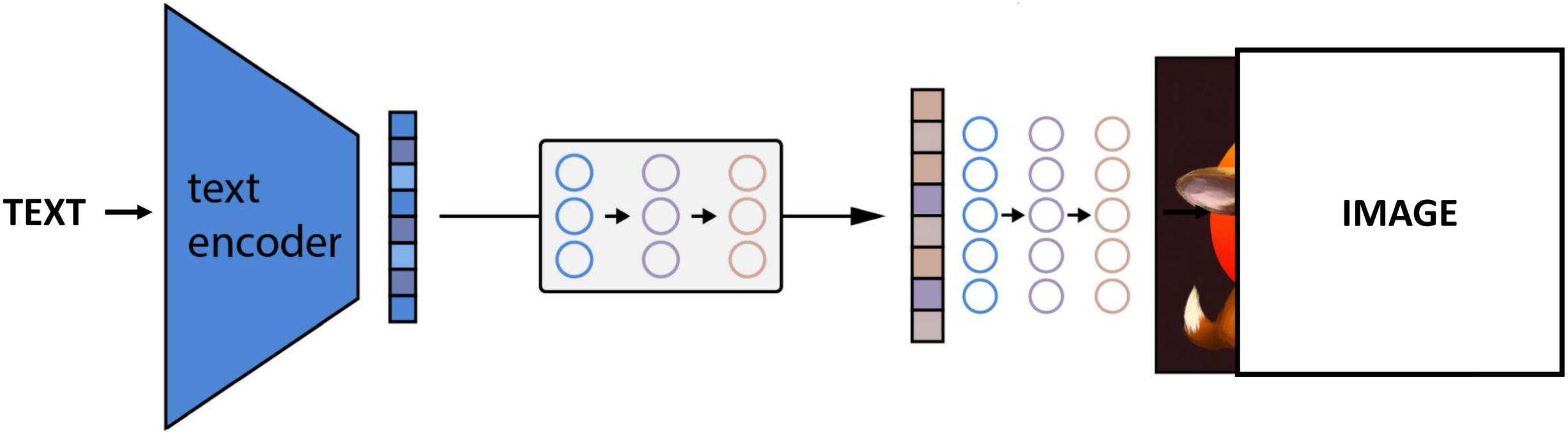
- For image classification, change the class labels to a “sentence” like structure and then you can input images and the set of text?? The text from which has the highest similarity score with the image, is the predicted class

8. Applications (Object Detection)

- There is also an object detection application… where you divide your image into patches… slide through them with clip.. And the parts of the image containing what is in the text will have higher similarity.



8. Applications (DALL-E 2)



9. Conclusion

- CLIP (Contrastive Language–Image Pre-training) builds on a large body of work on zero-shot transfer, natural language supervision, and multimodal learning.
- A critical insight was to leverage natural language as a flexible prediction space to enable generalization and transfer.
- CLIP was designed to mitigate a number of major problems in the standard deep learning approach to computer vision:
- Costly datasets
- Narrow
- Poor real-world performance

Resources

1. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. International Conference on Machine Learning.
2. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). Hierarchical Text-Conditional Image Generation with CLIP Latents. ArXiv, abs/2204.06125.
3. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, S.K., Ayan, B.K., Mahdavi, S.S., Lopes, R.G., Salimans, T., Ho, J., Fleet, D.J., & Norouzi, M. (2022). Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. ArXiv, abs/2205.11487.
4. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., & Aberman, K. (2022). DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. ArXiv, abs/2208.12242.
5. Kumar, A., Raghunathan, A., Jones, R., Ma, T., & Liang, P. (2022). Fine-Tuning can Distort Pretrained Features and Underperform Out-of-Distribution. ArXiv, abs/2202.10054.
6. Zhang, Y., Jiang, H., Miura, Y., Manning, C.D., & Langlotz, C. (2020). Contrastive Learning of Medical Visual Representations from Paired Images and Text. ArXiv, abs/2010.00747.

CLIP paper [1] has a complete and thorough reference list.

THANK YOU