Vision-Language Models
Part I:
Representation learning
CLIP

Introduction

From zero-shot Transfer to representation learning

Remember: Transfer Learning

		Source Data (not directly related to the task)	
		labelled	unlabeled
Target Data	labelled	Fine-tuning *Multitask Learning*	Not considered here
	unlabeled	Domain adaptation-adversarial training ***Zero-shot learning***	Not considered here

Zero-shot Learning

- Source data: (x^s, y^s) → Training data
- Target data: $(\emptyset)$ usually same domain

Different tasks

$Training\ time$:

x^s:   …

y^s: cat dog …

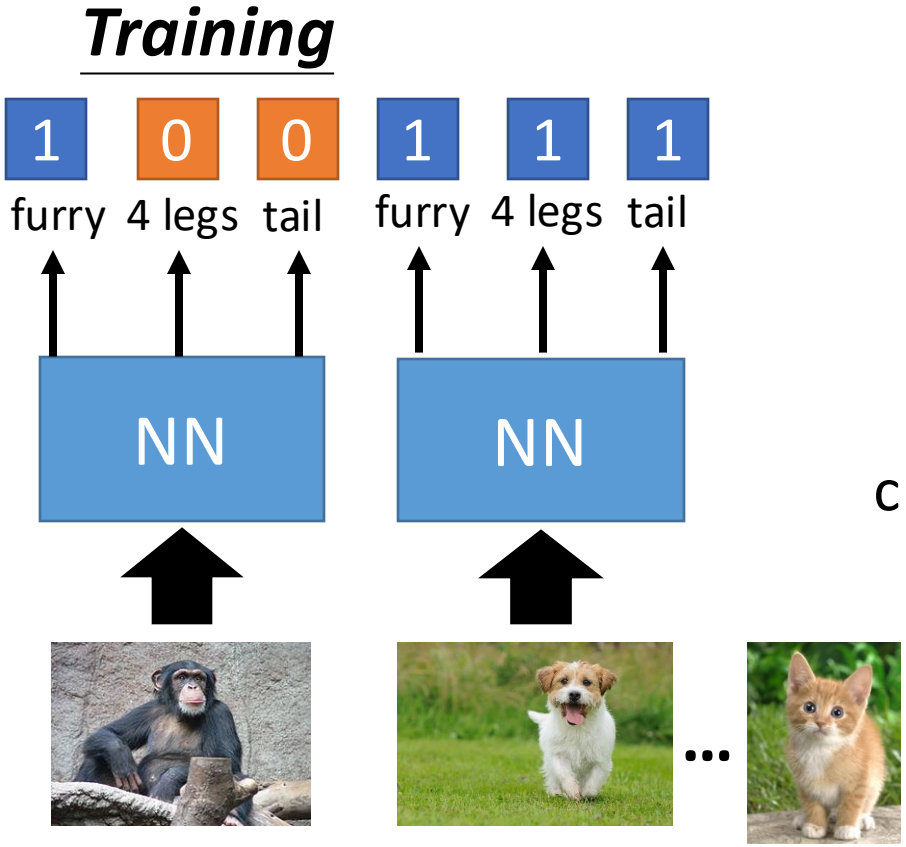
+ Class Information

$Test\ time\ x^t$: 

=> Fish class!

Zero-shot Learning

- Representing each class by its attributes



+
Database attributes

class

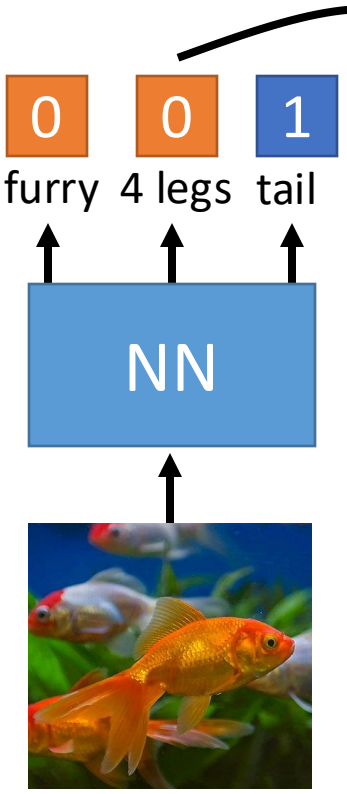
	furry	4 legs	tail	…
Dog	O	O	O	
Fish	X	X	O	
Chimp	O	X	X	
…				

sufficient attributes for one to one mapping

Zero-shot Learning

- Representing each class by its attributes

Testing



Find the class with the most similar attributes

class

attributes

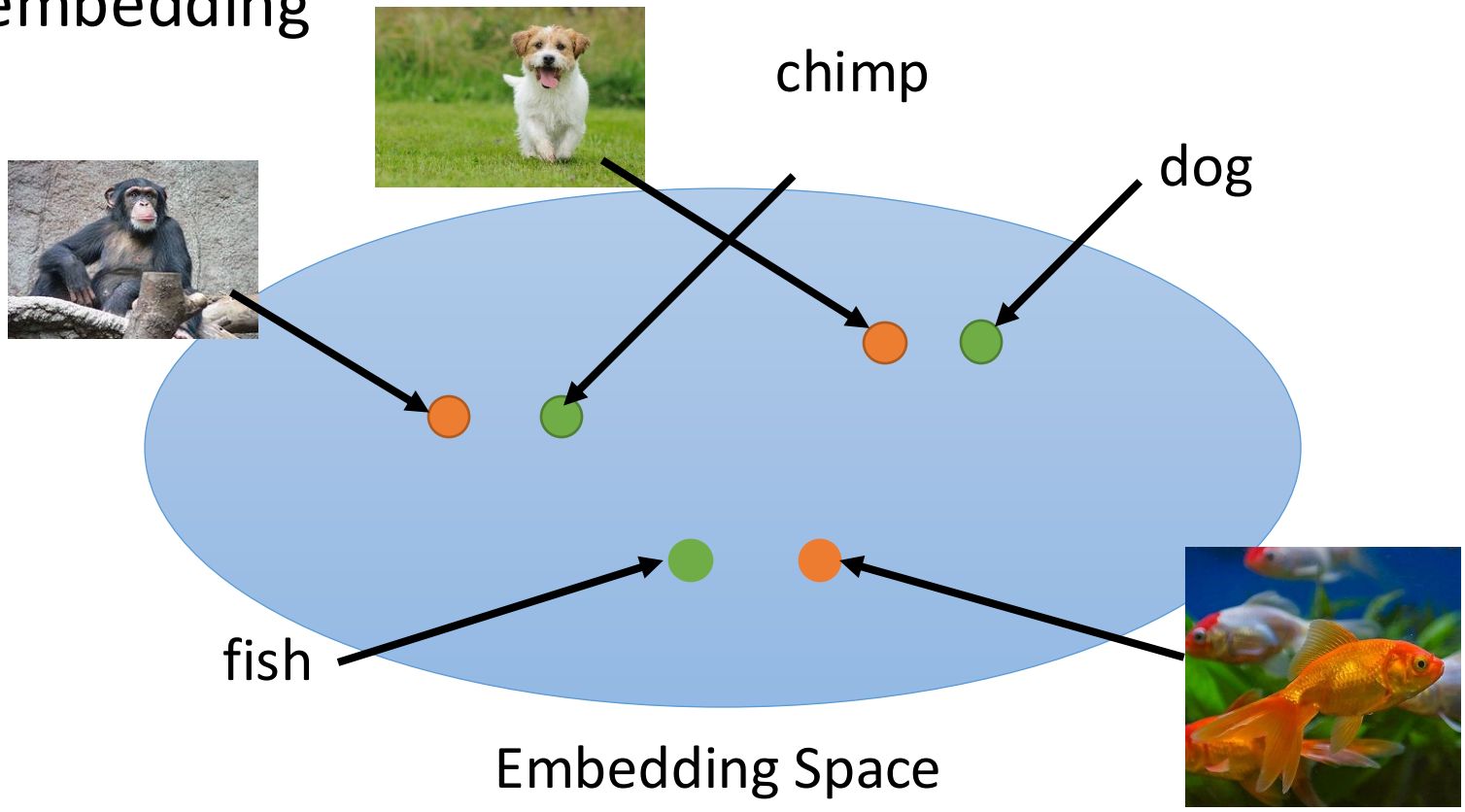
	furry	4 legs	tail	…
Dog	O	O	O	
Fish	X	X	O	
Chimp	O	X	X	
…				

sufficient attributes for one to one mapping

Zero-shot Learning

What if we don't have attribute database

- Attribute embedding + class (word name) embedding



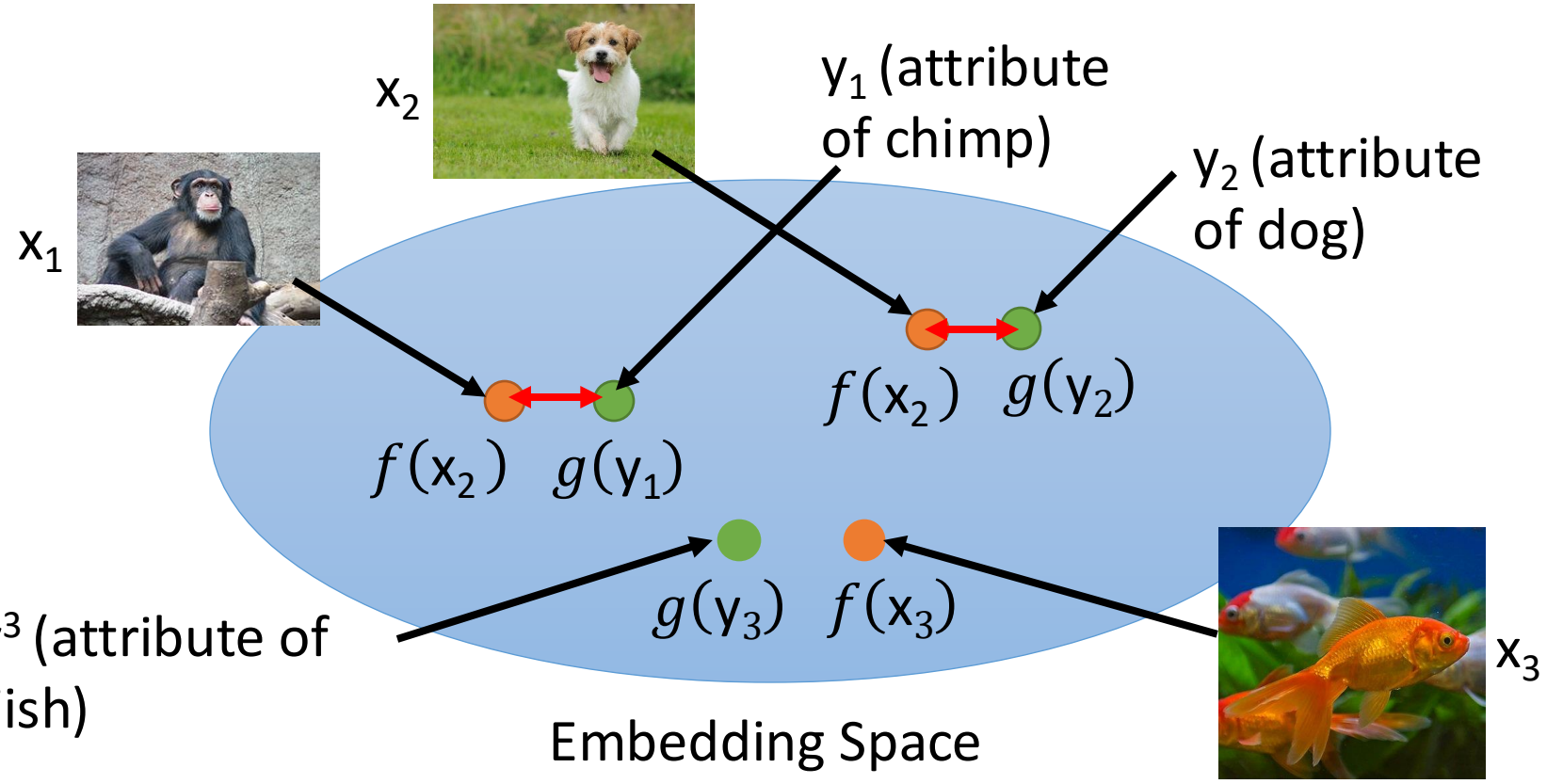
Zero-shot Learning

$f(*)$ and $g(*)$ can be NN.

Training target:

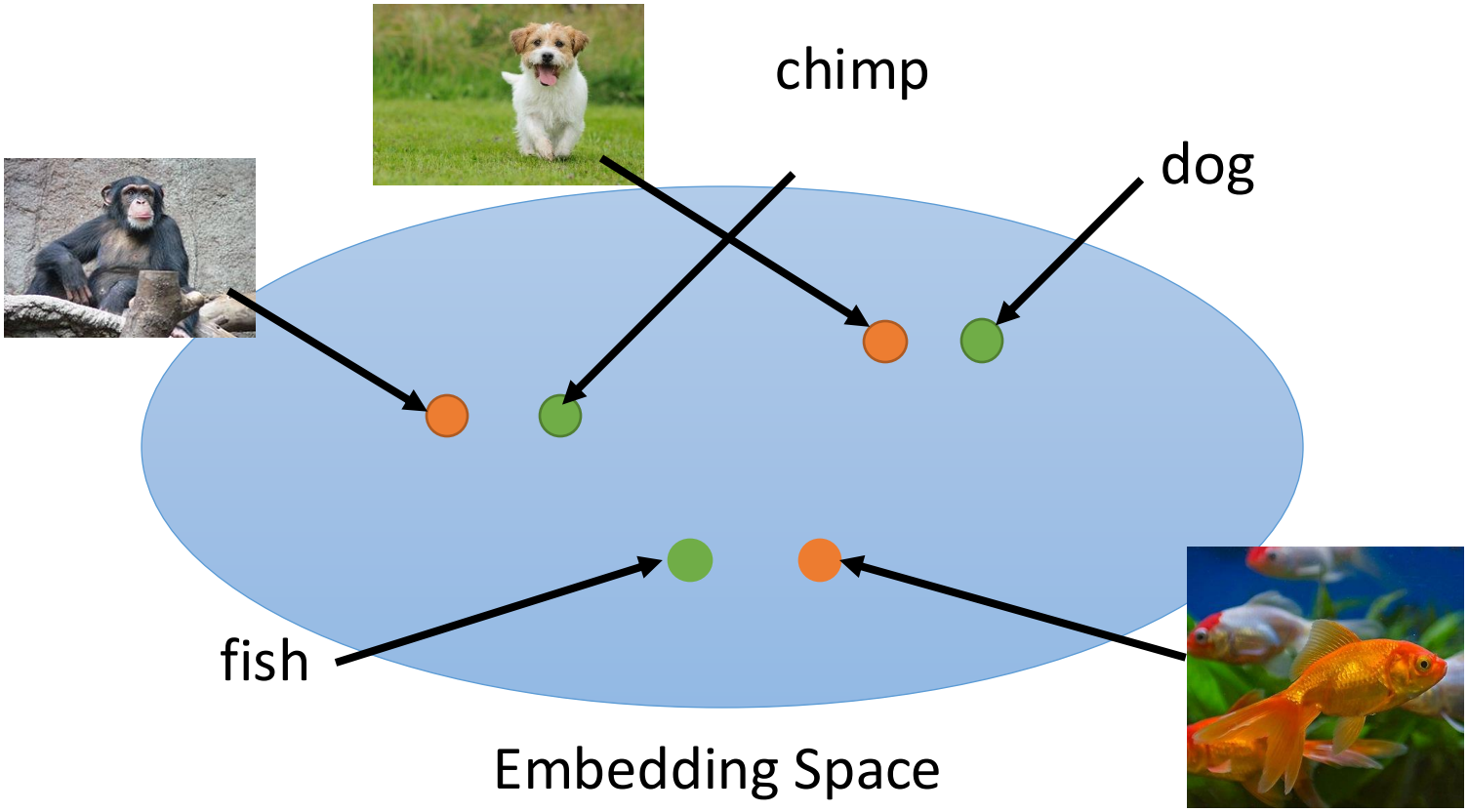
$f(x_n)$ and $g(y_n)$ as close as possible

- Attribute embedding



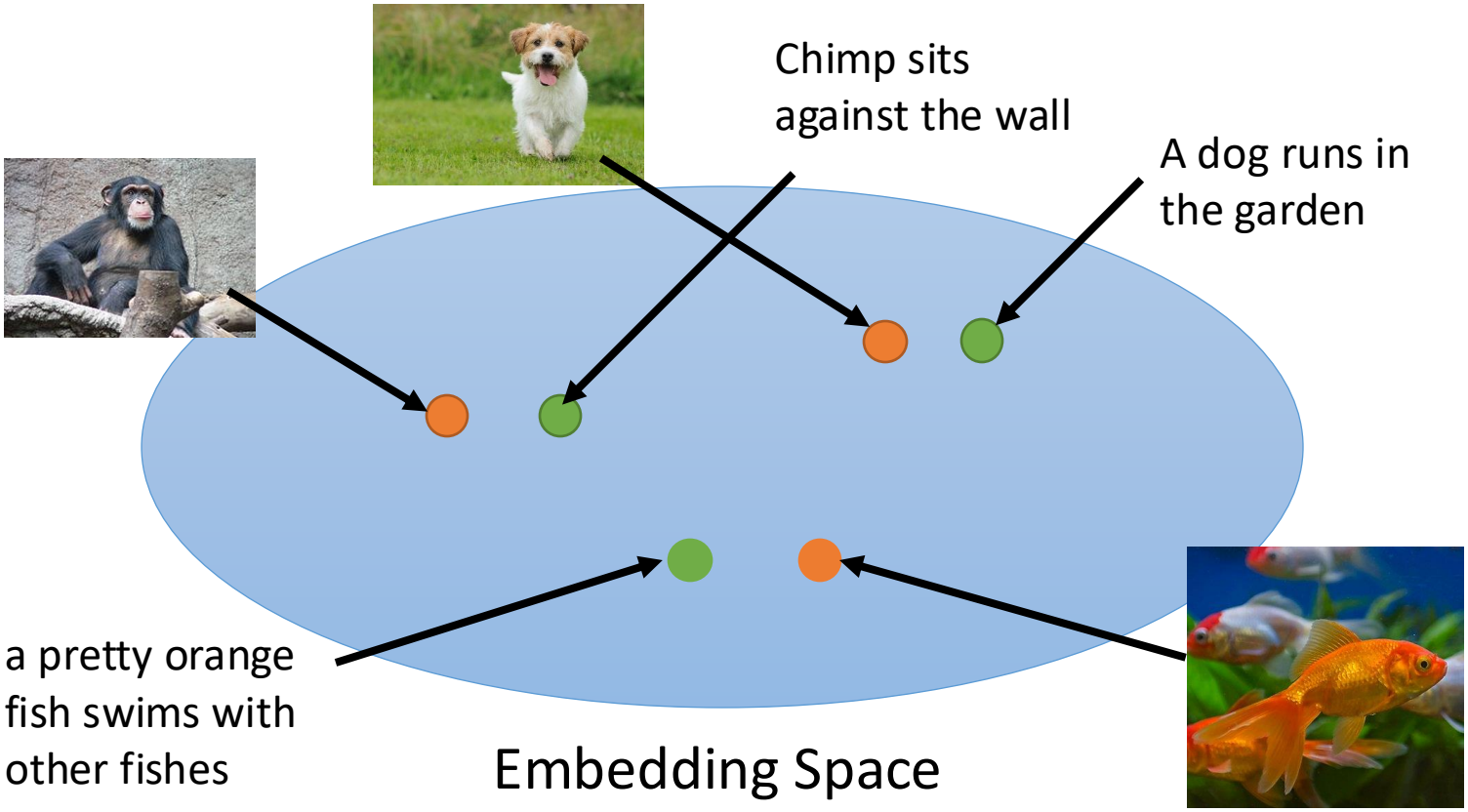
y_i are linked together by a class relationship (e.g. class name embedding as W2v)

More on Vision-Language: Representation Learning



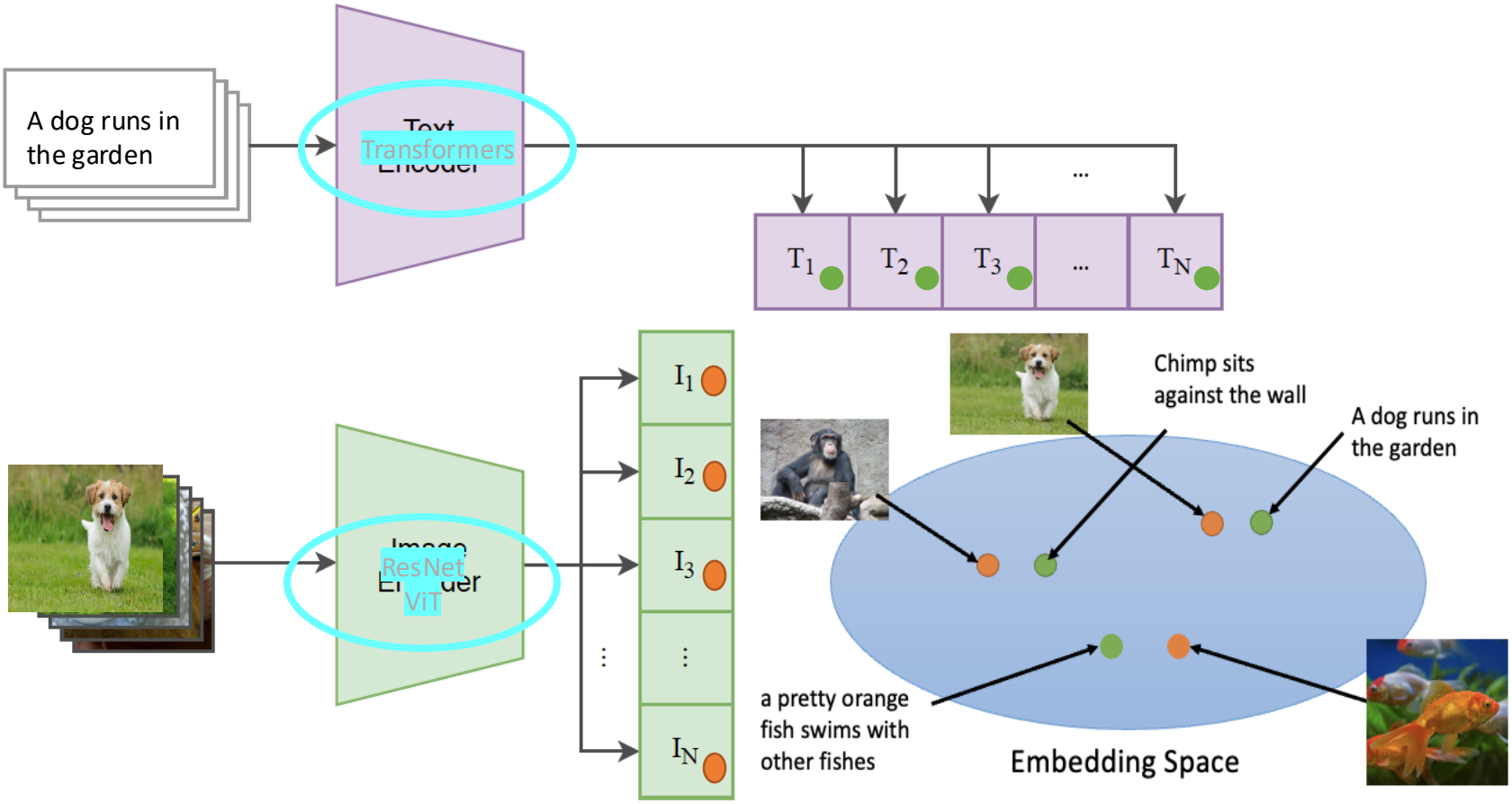
CLIP: Vision + Language Models (VLM)

[Learning transferable visual models from natural language supervision. Radford/Sutskever ICML, 2021]



CLIP: Vision + Language Models (VLM)

Dual architecture: Text encoder + Image encoder



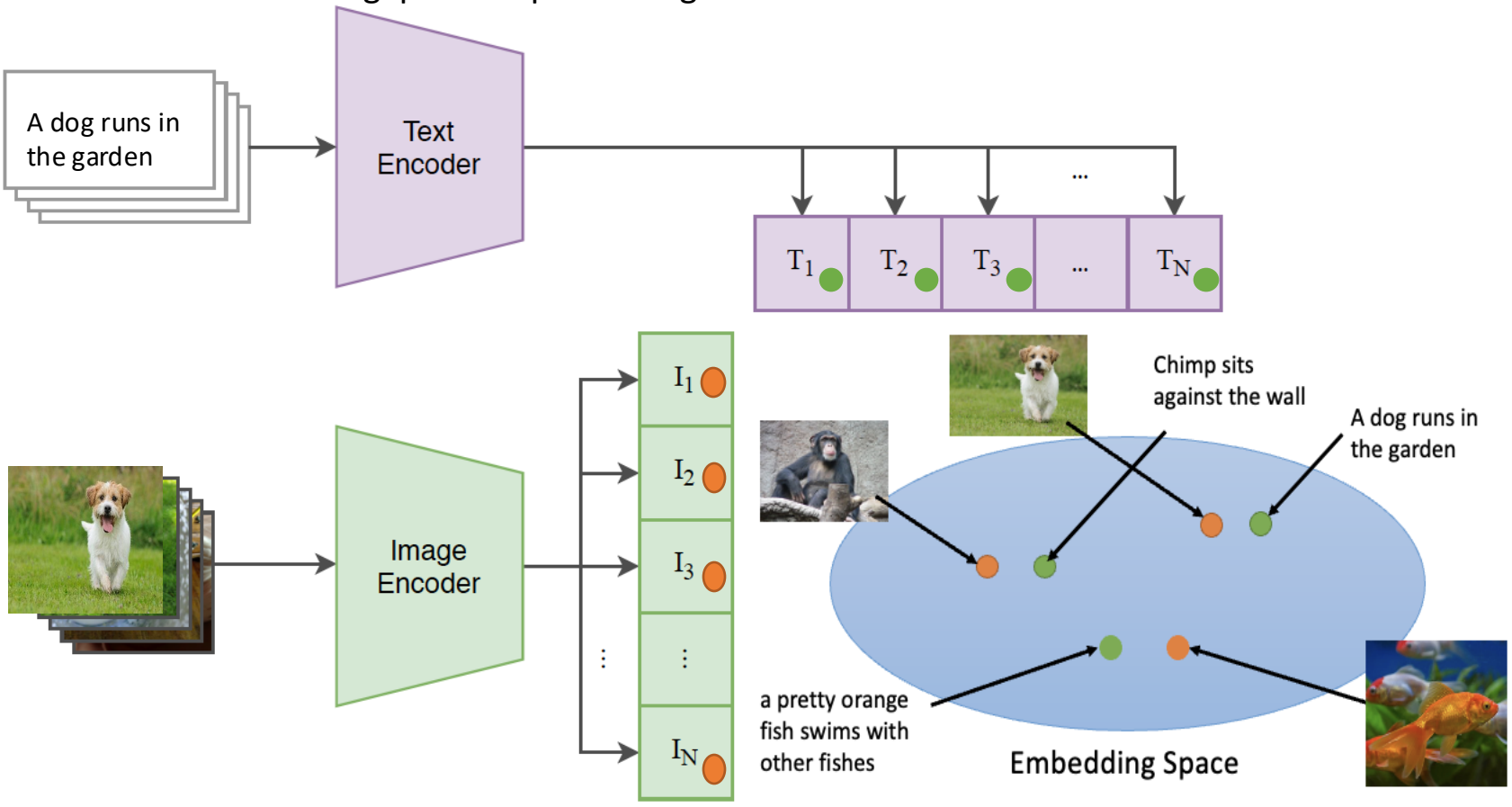
CLIP: Vision + Language Models (VLM)

Learning strategy

Training set: $A = \{(\mathbf{I}_n, \mathbf{T}_n)\}_n$ of image/caption pairs (coherent!)

Massive Text+Image =**400M pairs** to train the model (from the Internet)

Contrastive loss for training: positive pair vs negative one or set



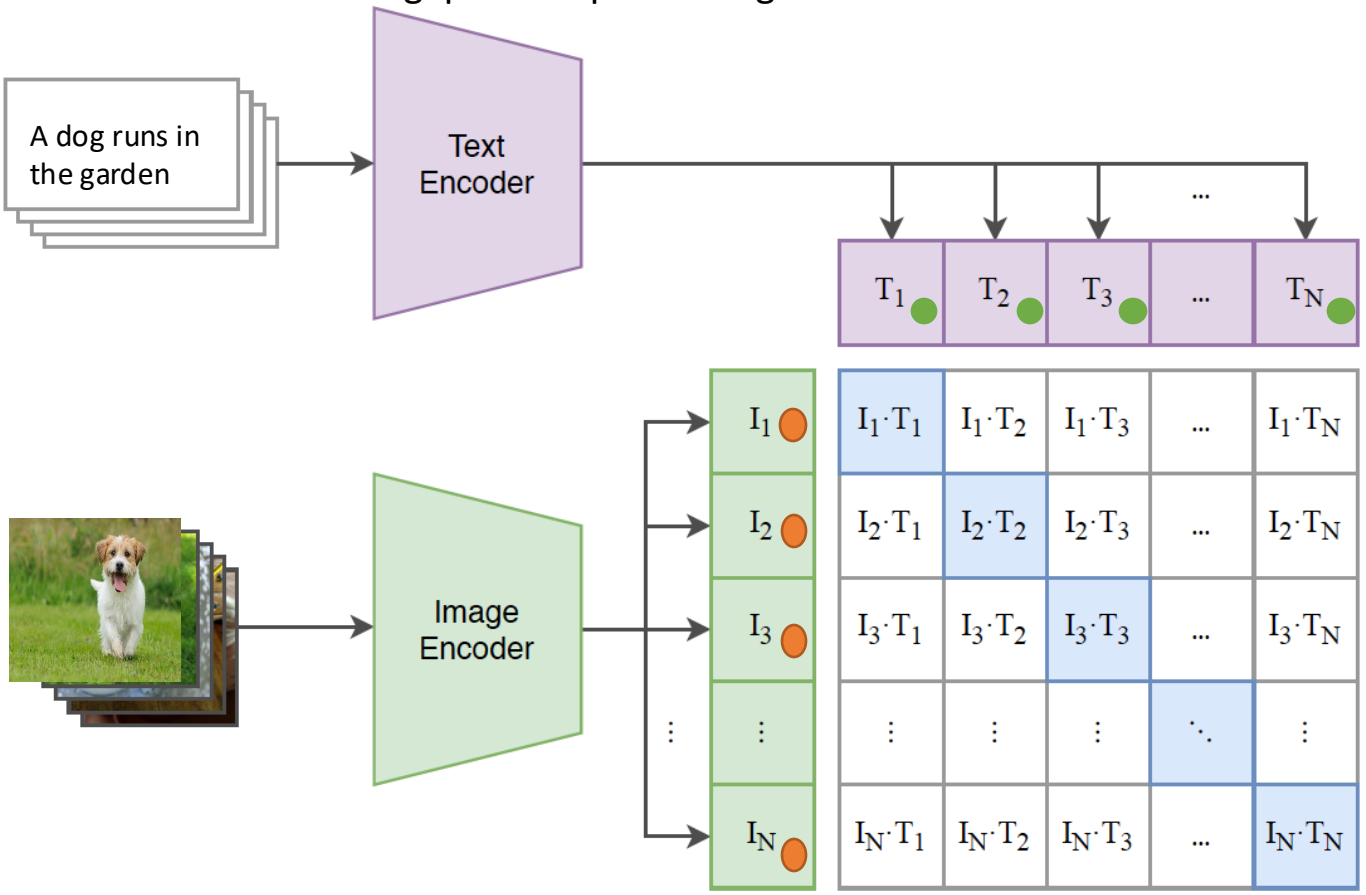
CLIP: Vision + Language Models (VLM)

Learning strategy

Training set: $A = \{(\mathbf{I}_n, \mathbf{T}_n)\}_n$ of image/caption pairs (coherent!)

Massive Text+Image =**400M pairs** to train the model (from the Internet)

Contrastive loss for training: positive pair vs negative one or set



CLIP: Vision + Language Models (VLM)

Learning strategy

Training set: $A = \{(\mathbf{I}_n, \mathbf{T}_n)\}_n$ of image/caption pairs (coherent!)

Massive Text+Image =**400M pairs** to train the model (from the Internet)

Contrastive loss for training: positive pair vs negative pair or more

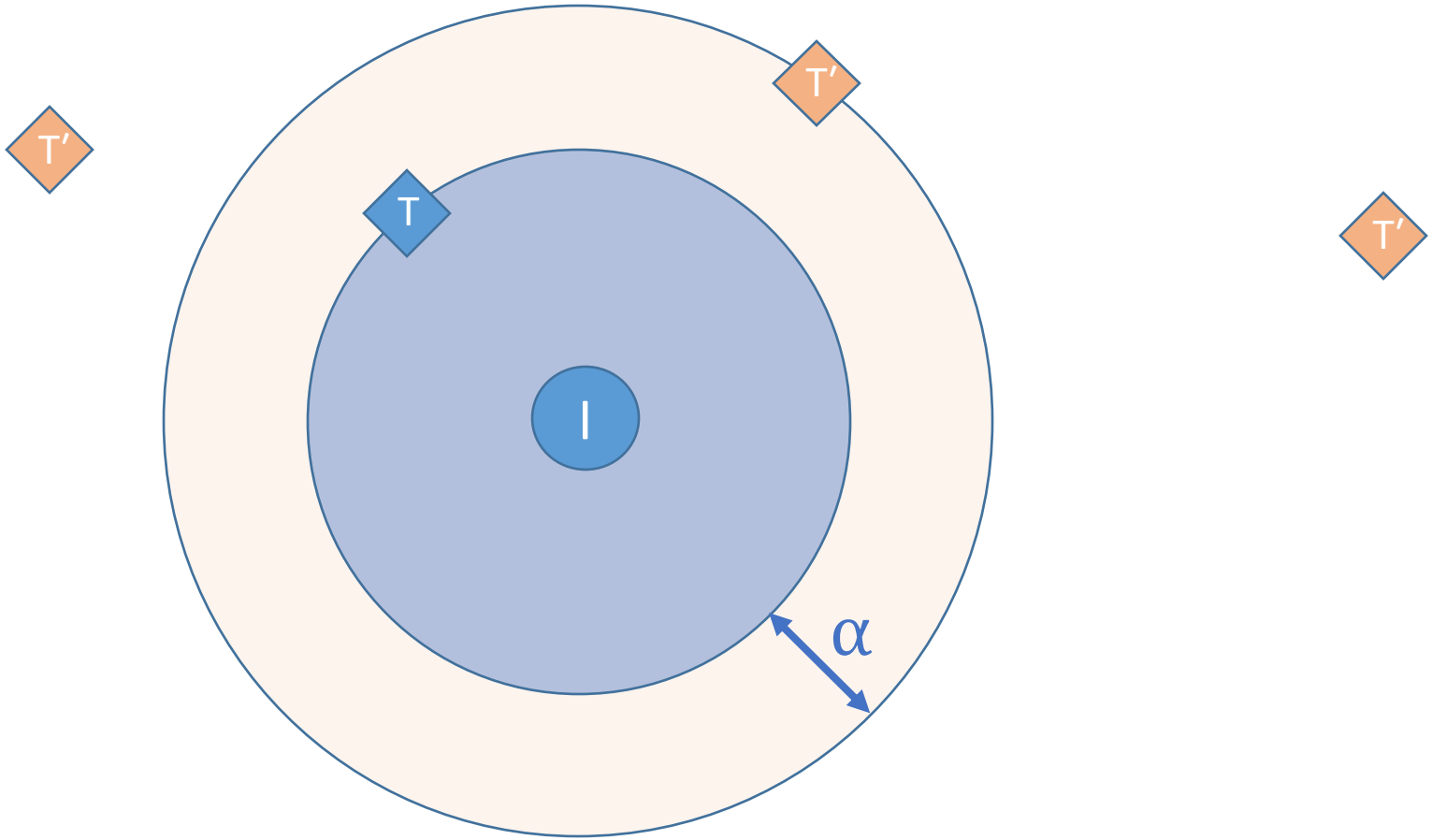
(contrastive) Triplet loss: A variant of the standard margin based loss (SVM)

- Triplet ($\mathbf{I}$, $\mathbf{T}$, $\mathbf{T}'$) (Batch = 3)
- Anchor: $\mathbf{I}$ (E.g image representation)
- Positive: $\mathbf{T}$ (E.g associated caption representation)
- Negative: $\mathbf{T}'$ (E.g contrastive caption representation)
- Margin parameter α

$$\mathrm{TripletLoss}(\mathrm{I}, \mathrm{T}, \mathrm{T}') = \max\{0, \alpha + \mathrm{d}(\mathrm{I}, \mathrm{T}) - \mathrm{d}(\mathrm{I}, \mathrm{T}')\}$$

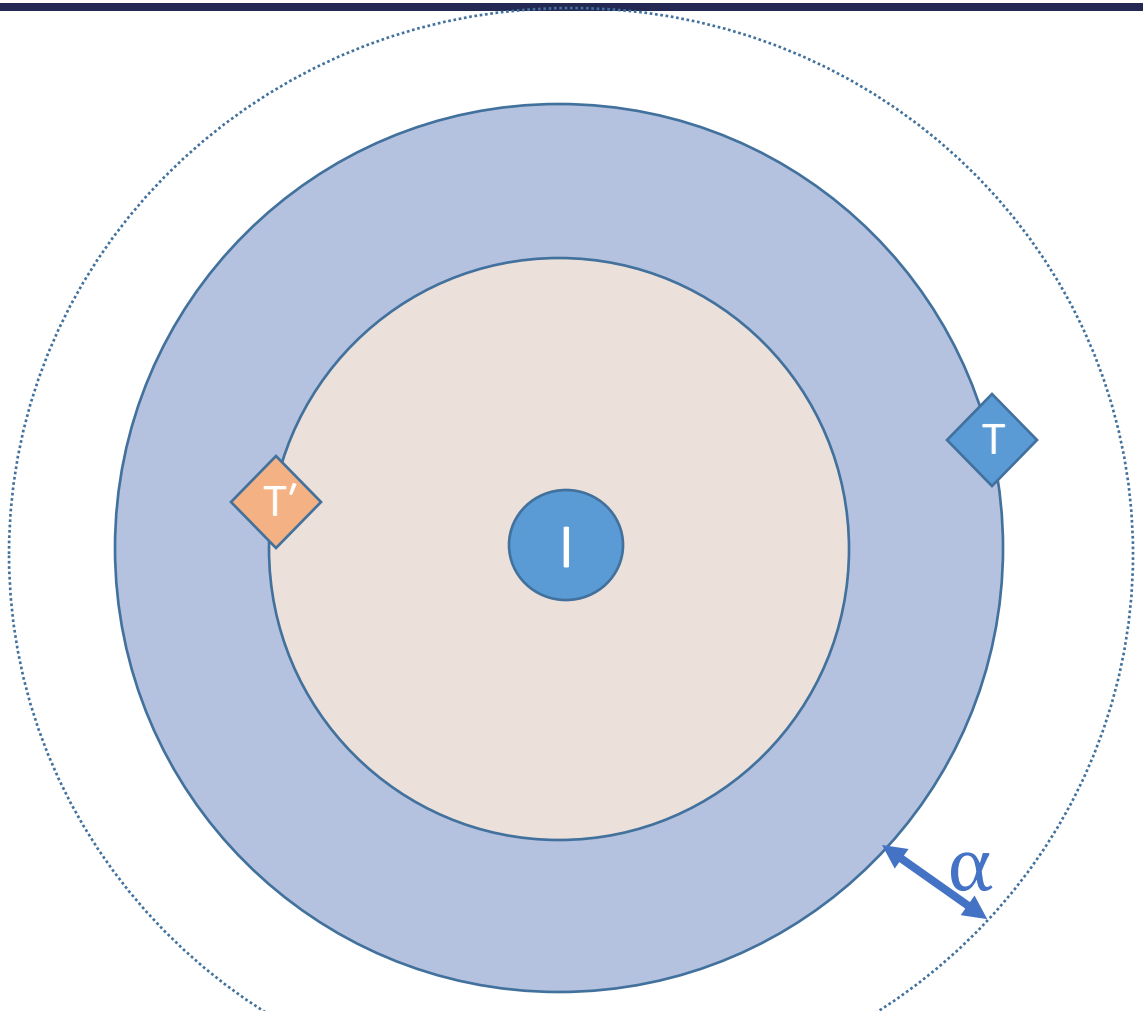
Learning strategy: triplet loss

$$\mathrm{TripletLoss}(I, T, T') = \max\{0, \alpha + d(I, T) - d(I, T')\}$$



Learning strategy: triplet loss

$$\mathrm{TripletLoss}(I, T, T') = \max\{0, \alpha + d(I, T) - d(I, T')\}$$



Learning strategy: triplet loss

Hard negative margin-based loss:

Loss for a **batch** $\mathcal{B} = \{(\mathbf{I}_n, \mathbf{T}_n)\}_{n \in B}$ of image/sentence pairs:

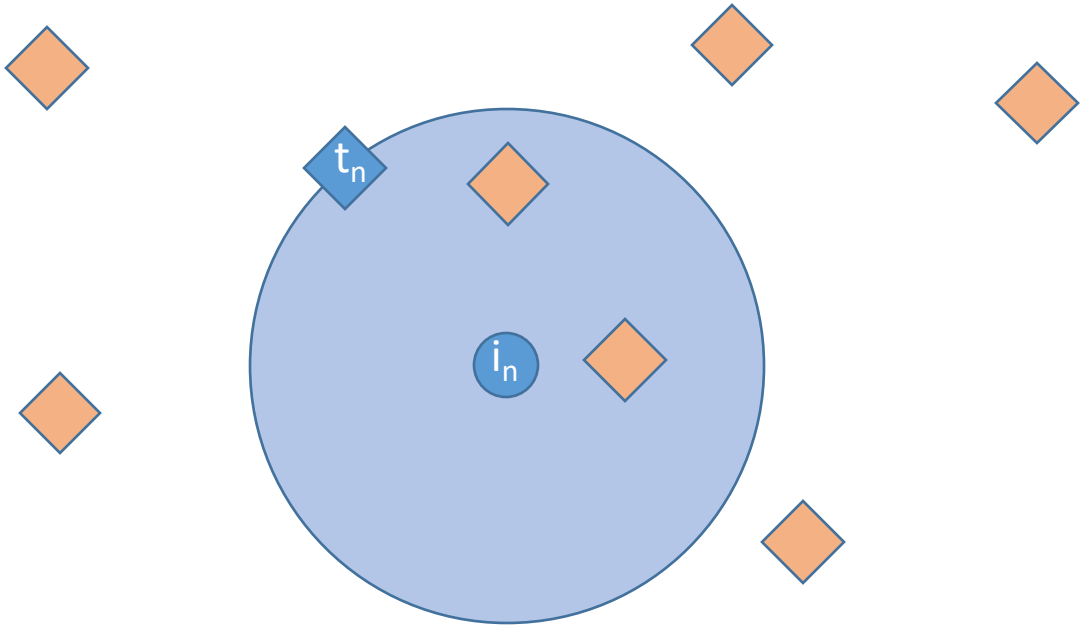
$$\mathcal{L}(\mathbf{\Theta}; \mathcal{B}) = \frac{1}{|B|} \sum_{n \in B} \left(\max_{m \in C_n \cap B} \text{loss}\,(\boldsymbol{I}_n, \boldsymbol{T}_n, \boldsymbol{T}_m) + \max_{m \in D_n \cap B} \text{loss}\,(\boldsymbol{T}_n, \boldsymbol{I}_n, \boldsymbol{I}_m) \right)$$

With C_n (resp. D_n) set of indices of caption (resp. image) unrelated to n-th element

Learning strategy: hard negative triplet loss

Mining hard negative contrastive example:

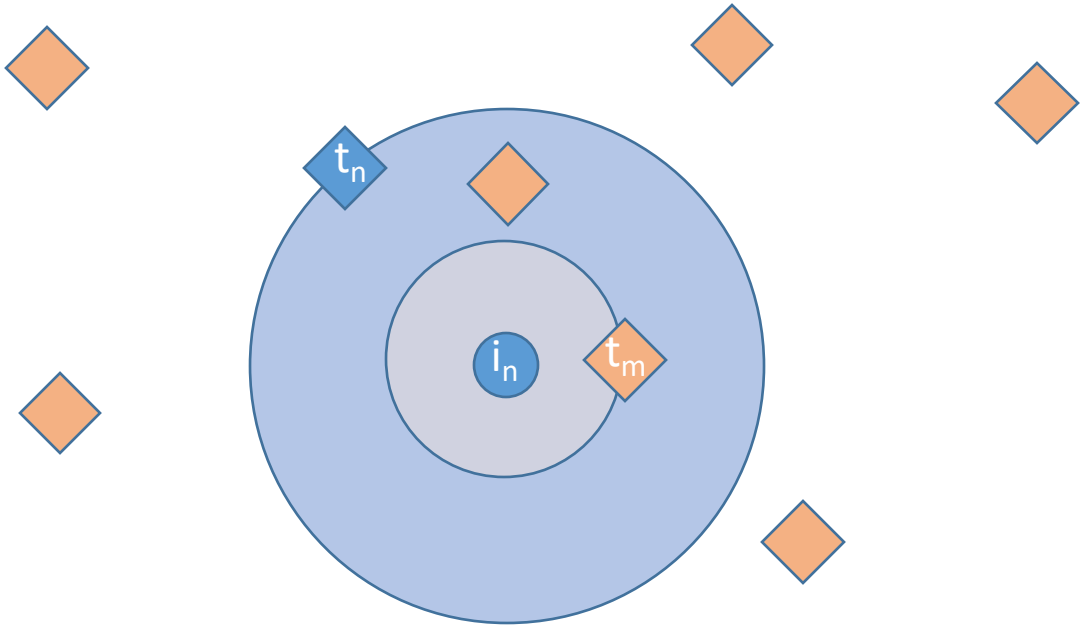
$$\mathcal{L}(\boldsymbol{\Theta}; \mathcal{B}) = \frac{1}{|B|} \sum_{n \in B} \left(\begin{array}{l} \max\limits_{m \in C_n \cap B} \text{loss}\,(\boldsymbol{I}_n, \boldsymbol{T}_n, \boldsymbol{T}_m) \\ + \max\limits_{m \in D_n \cap B} \text{loss}\,(\boldsymbol{T}_n, \boldsymbol{I}_n, \boldsymbol{I}_m) \end{array} \right)$$



Learning strategy: hard negative triplet loss

Mining hard negative contrastive example:

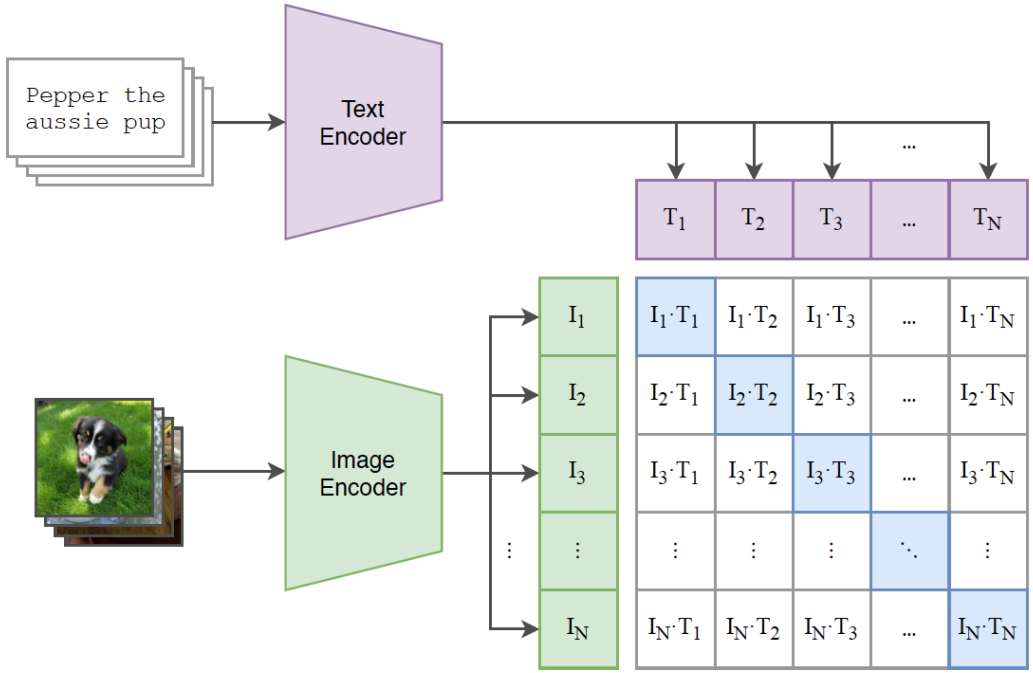
$$\mathcal{L}(\boldsymbol{\Theta}; \mathcal{B}) = \frac{1}{|B|} \sum_{n \in B} \left(\begin{array}{l} \max\limits_{m \in C_n \cap B} \text{loss}\,(\boldsymbol{I}_n, \boldsymbol{T}_n, \boldsymbol{T}_m) \\ + \max\limits_{m \in D_n \cap B} \text{loss}\,(\boldsymbol{T}_n, \boldsymbol{I}_n, \boldsymbol{I}_m) \end{array} \right)$$



CLIP: Vision + Language Models (VLM)

Massive Text+Image =**400M pairs** to train the model (from the Internet)

Contrastive loss for training

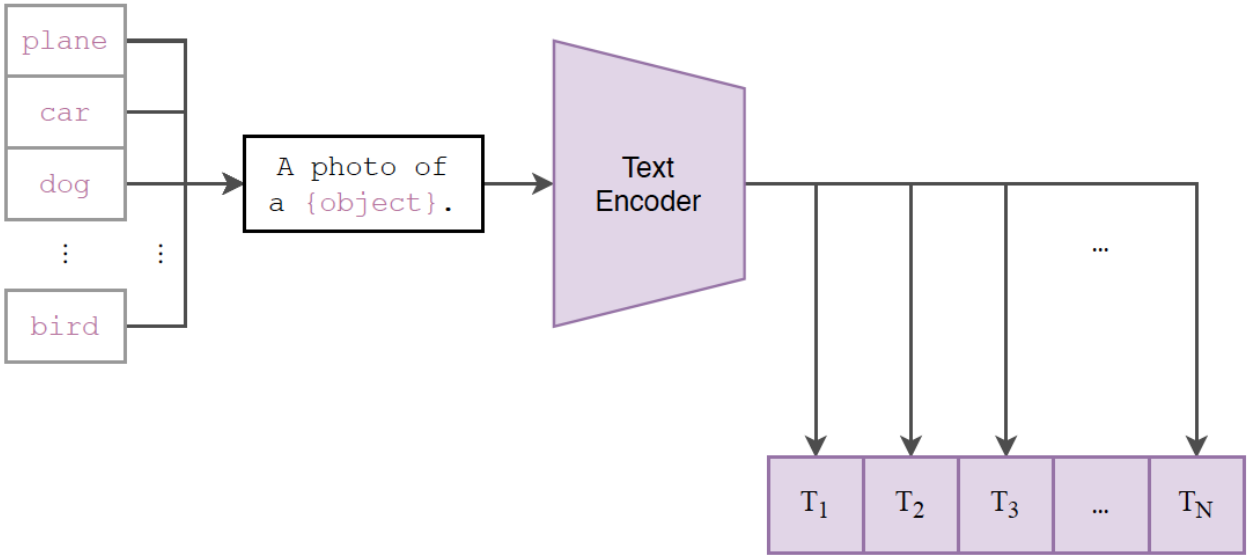


$$\mathcal{L}_{InfoNCE_{CLIP}} = -\sum_{i} \log \left(\frac{exp(\frac{sim(I_i, T_i)}{\tau})}{\sum_{k=1}^{N} exp(\frac{sim(I_i, T_k)}{\tau})} \right)$$

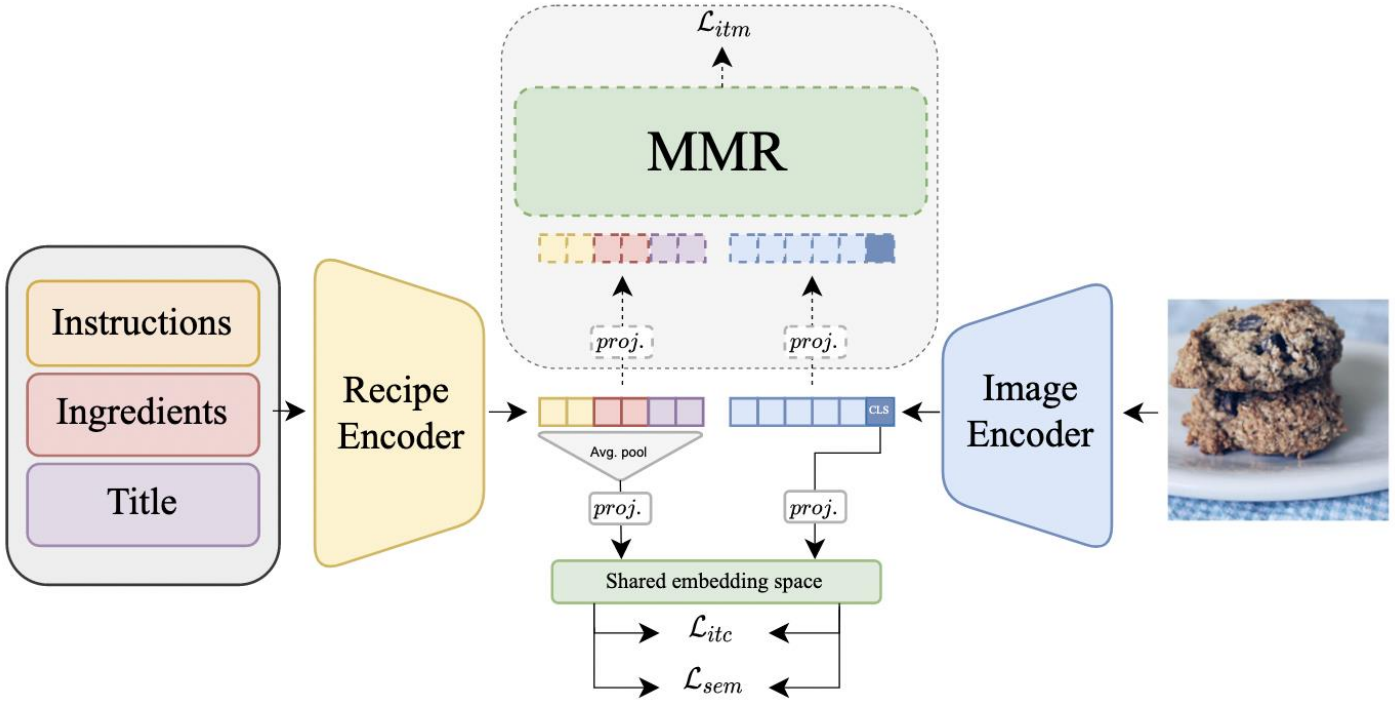
CLIP: Vision + Language Models (VLM)

Pre-trained encoders = **dual encoders** (Text/Image)

used for Zero-shot classifier, and other downstream tasks



A lot of variants ….



Title query	Ingredient query	Instruction query
Mint Chocolate Chip Frosting.	1 cup Unsalted Butter, 2 Tablespoons Heavy Cream, 2 drops Green Food Coloring, .. Chocolate..	Add sugar, cream, peppermint, and food coloring... ...scoop the frosting and place on top of your cupcakes Source: Chocolate Cupcakes with Mint Chocolate Chip ...
Honey-Grilled Chicken.	1 broiler-fryer chicken, halved, 34 cup butter, melted, 14 cup honey..	Place the halved chicken in a large, shallow container... ...Combine the remaining ingredients, stirring sauce well Grill chicken, skin side up..
The Best Kale Ever.	1/2 cup Kale, 1 teaspoon Olive Oil, 1/4 teaspoons Red Pepper Flakes..	Wash and cut kale off the stems... Heat olive oil on medium heat and add garlic Add in kale and red pepper flakes..

Top 5 retrieved images