

23. Multi-modal Learning

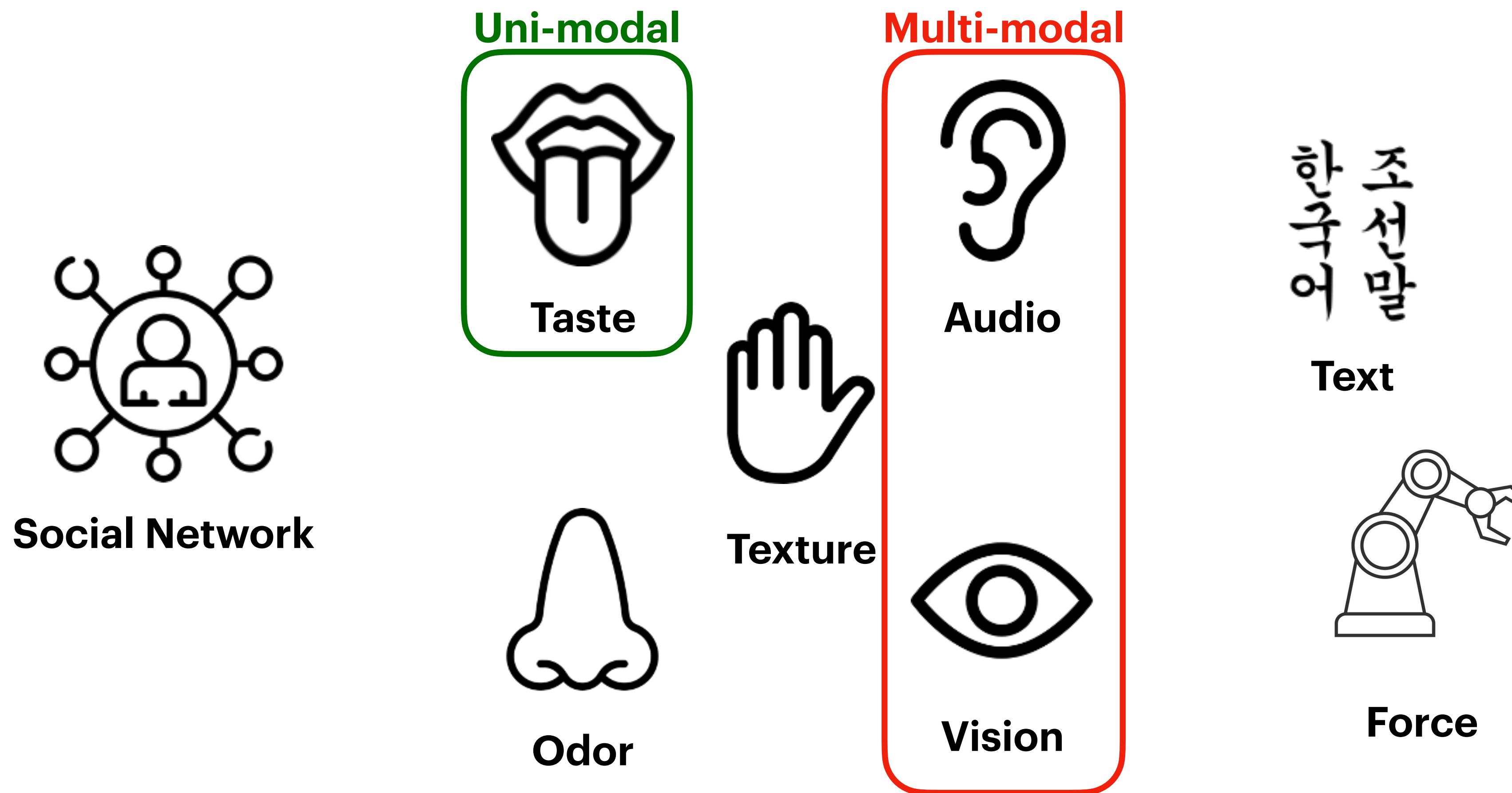
**EECE454 Introduction to
Machine Learning Systems**

2023 Fall, Jaeho Lee

Overview

Multi-modality

- Modalities in multi-modal learning

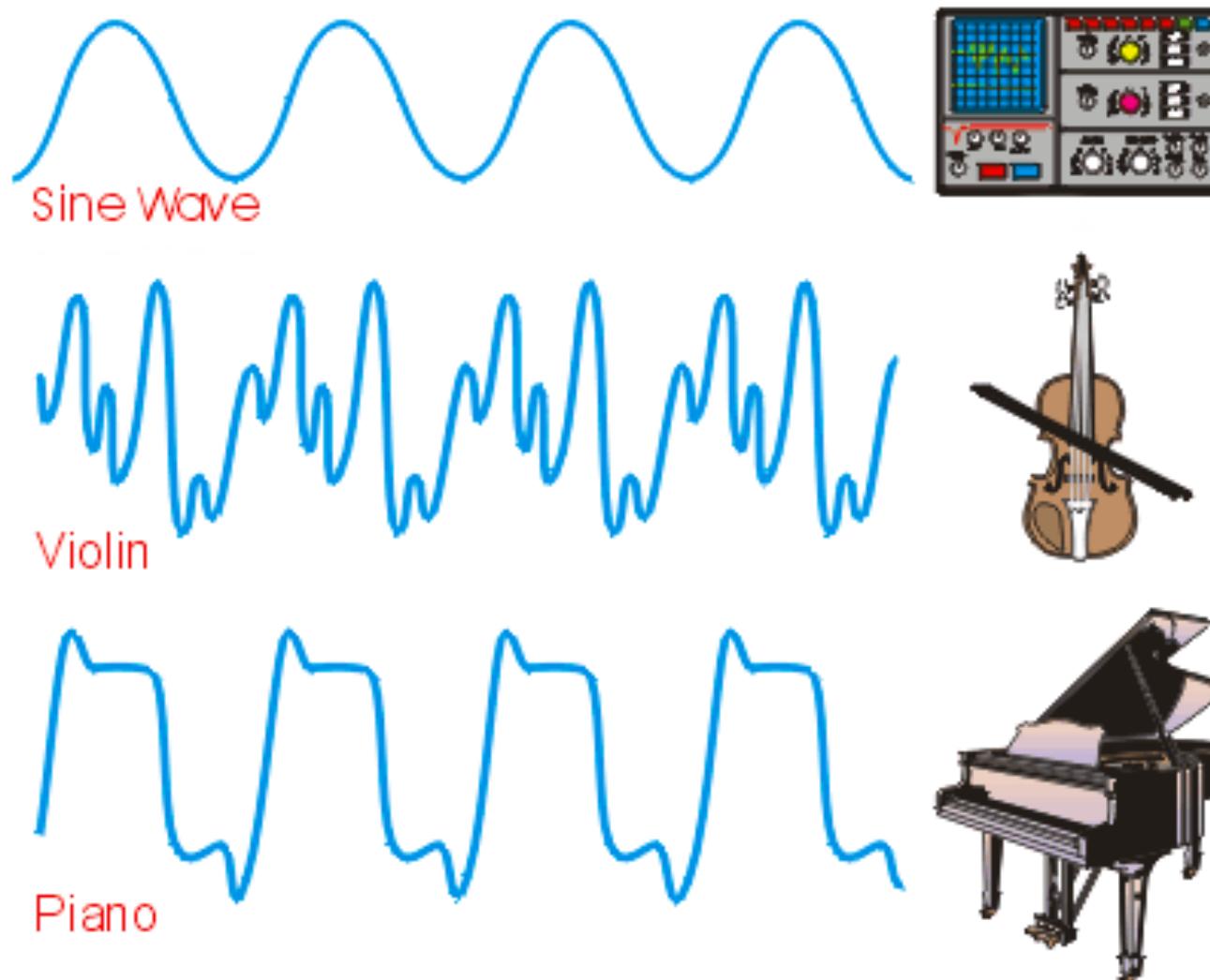


Challenges

1. Representation. Data in each domain have different representations

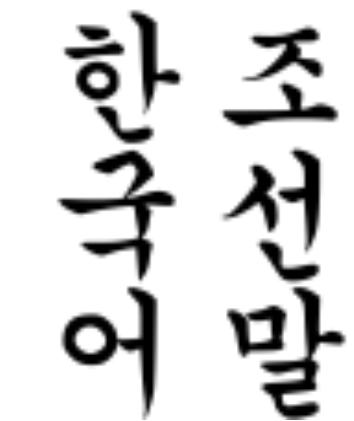


Audio



Vision

132	98	91	89	87	89	89	101	125	1
147	121	101	93	93	91	93	112	134	1
153	142	130	109	102	99	101	121	138	1
171	169	169	154	139	137	119	123	142	1
175	186	190	189	180	179	158	133	144	1
167	177	187	199	189	185	175	150	146	1
159	159	163	189	189	180	164	153	148	1
151	156	154	162	184	179	153	145	145	1

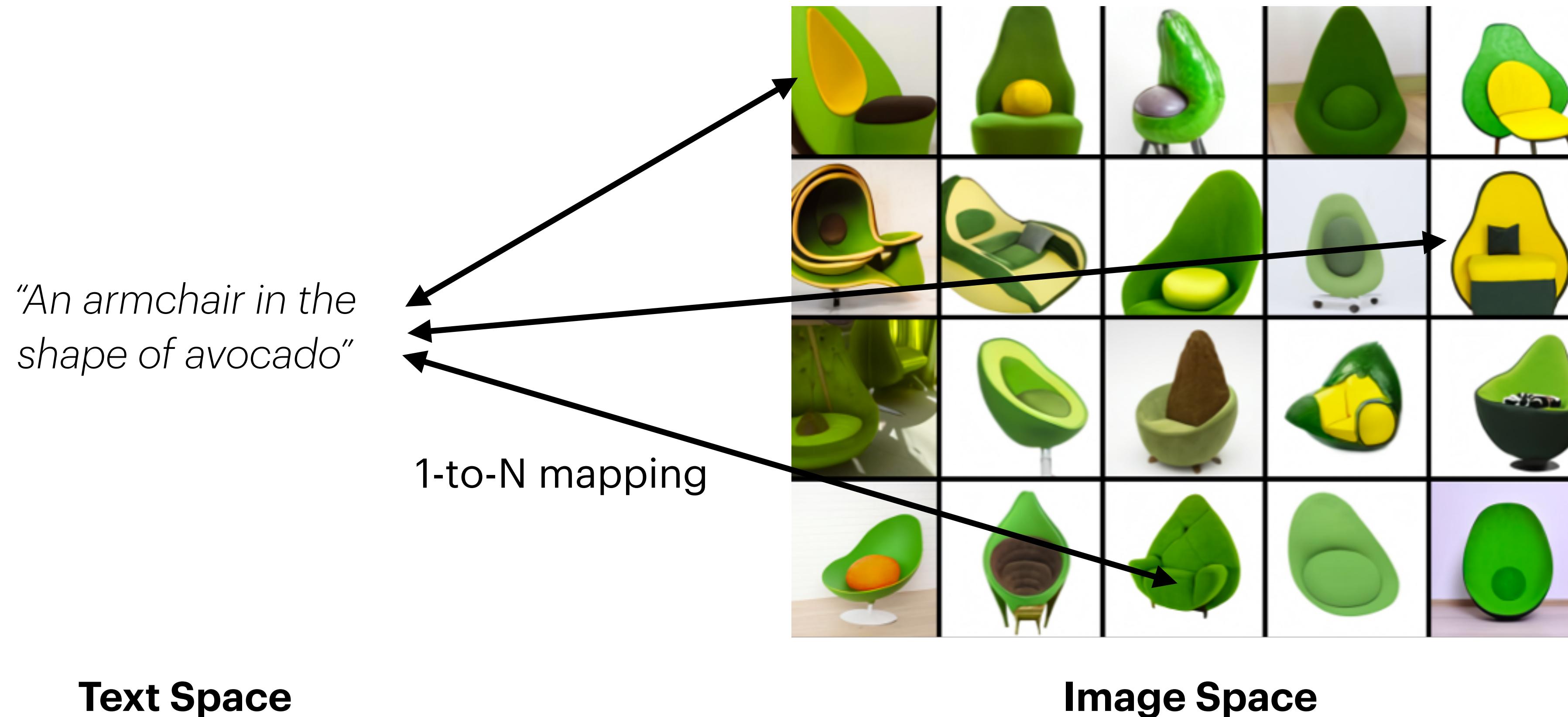


Text

	living	being	feline	human	gender	royalty	verb	plural
<i>cat</i> →	0.6	0.9	0.1	0.4	-0.7	-0.3	-0.2	
<i>kitten</i> →	0.5	0.8	-0.1	0.2	-0.6	-0.5	-0.1	
<i>dog</i> →	0.7	-0.1	0.4	0.3	-0.4	-0.1	-0.3	
<i>houses</i> →	-0.8	-0.4	-0.5	0.1	-0.9	0.3	0.8	

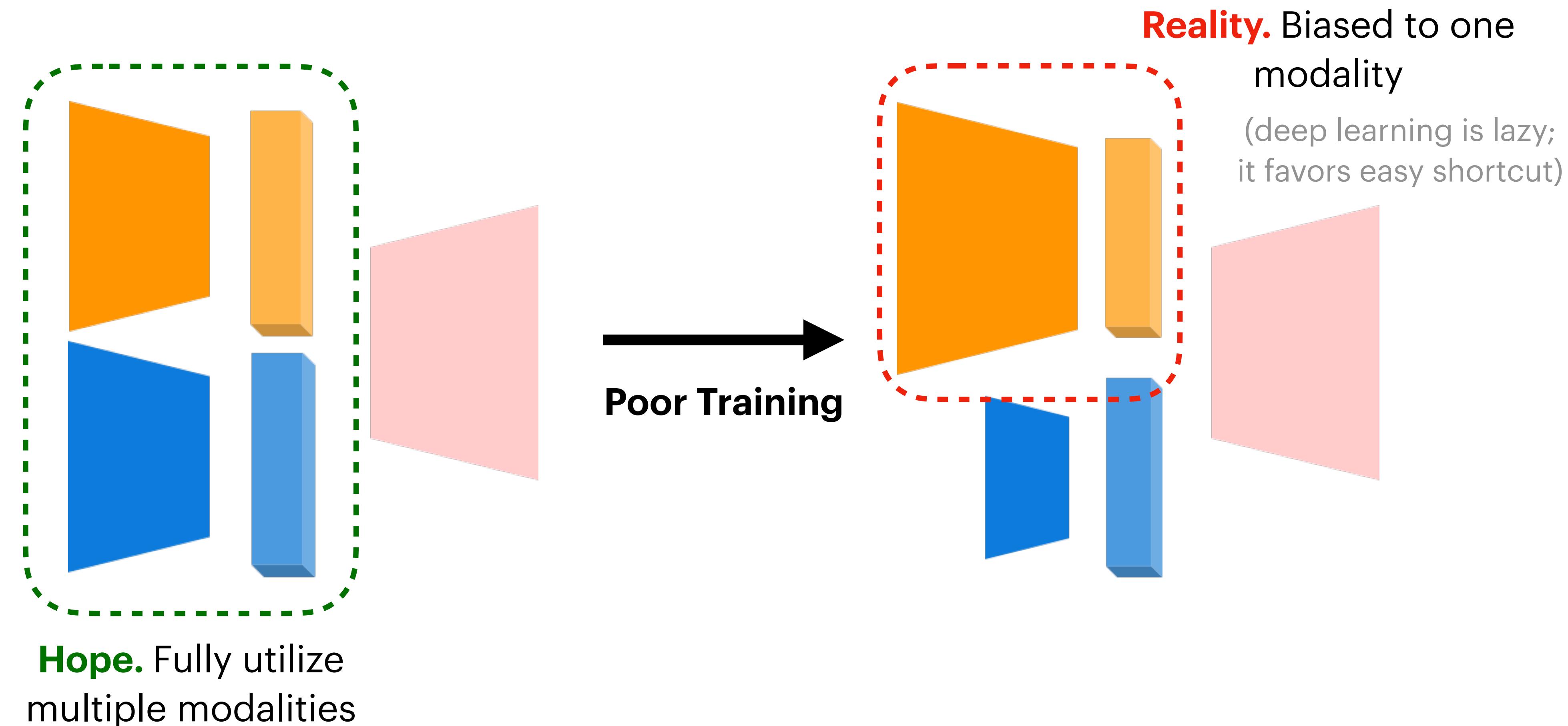
Challenges

2. Correspondence. Heterogeneous feature spaces with potentially limited correspondence



Challenges

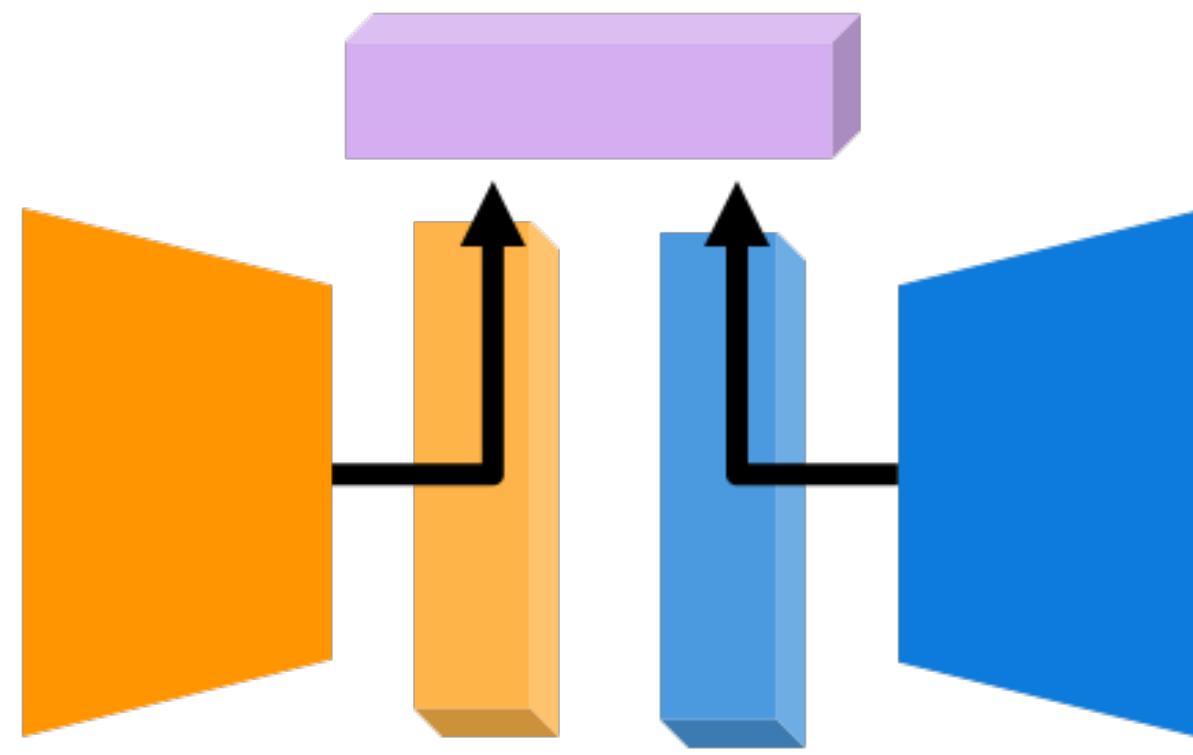
3. Bias. Imbalance between heterogeneous feature spaces



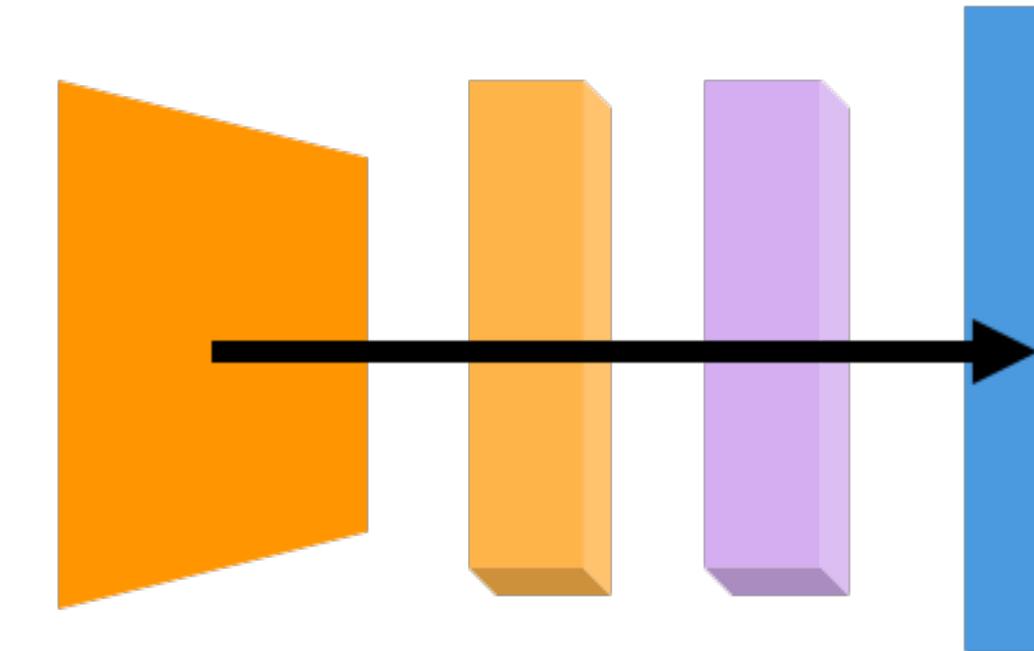
Yet...

Despite the challenges, we expect much fruitful outcomes

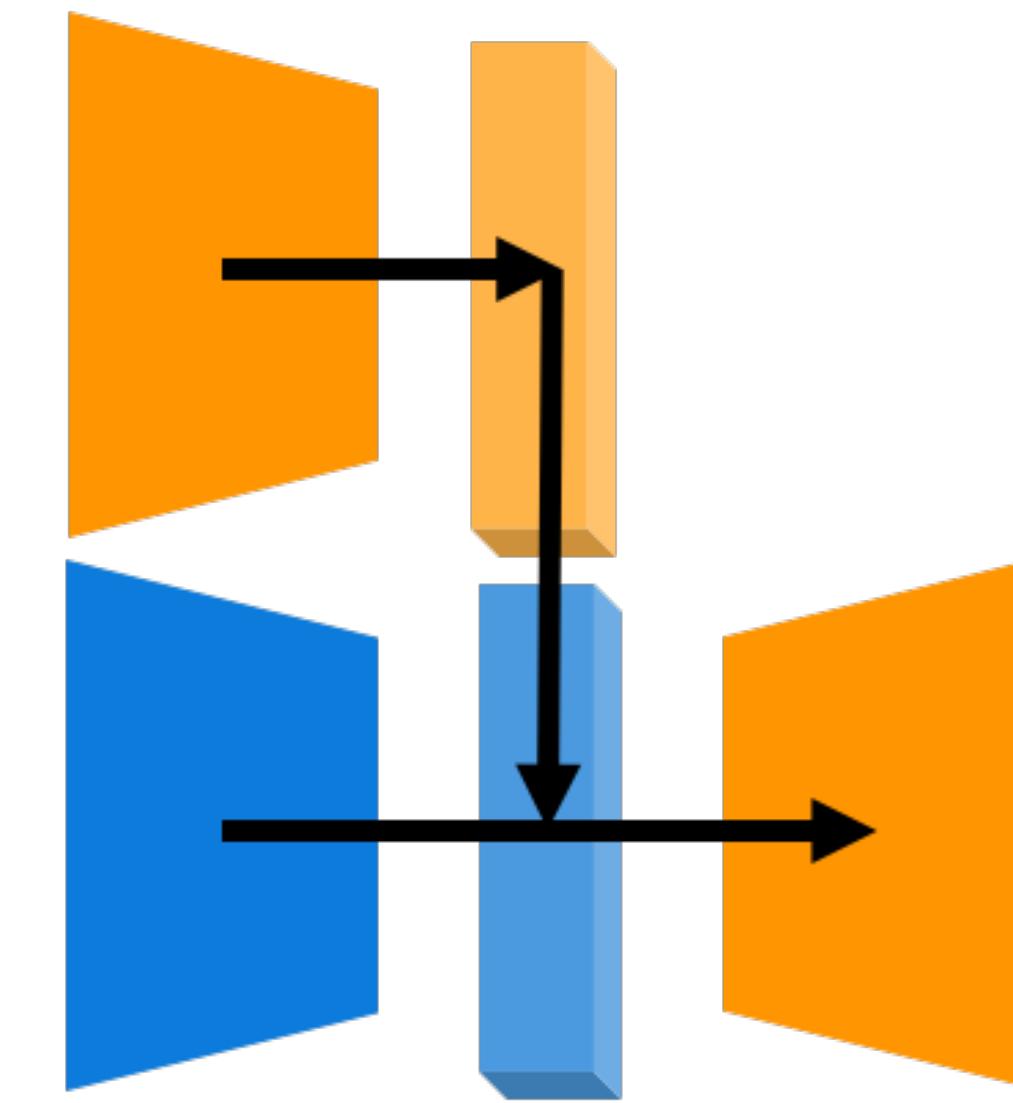
- We look at the example of CLIP, which handles ***vision + text***



Matching



Translating



Referencing

Vision & Language

Text Embedding

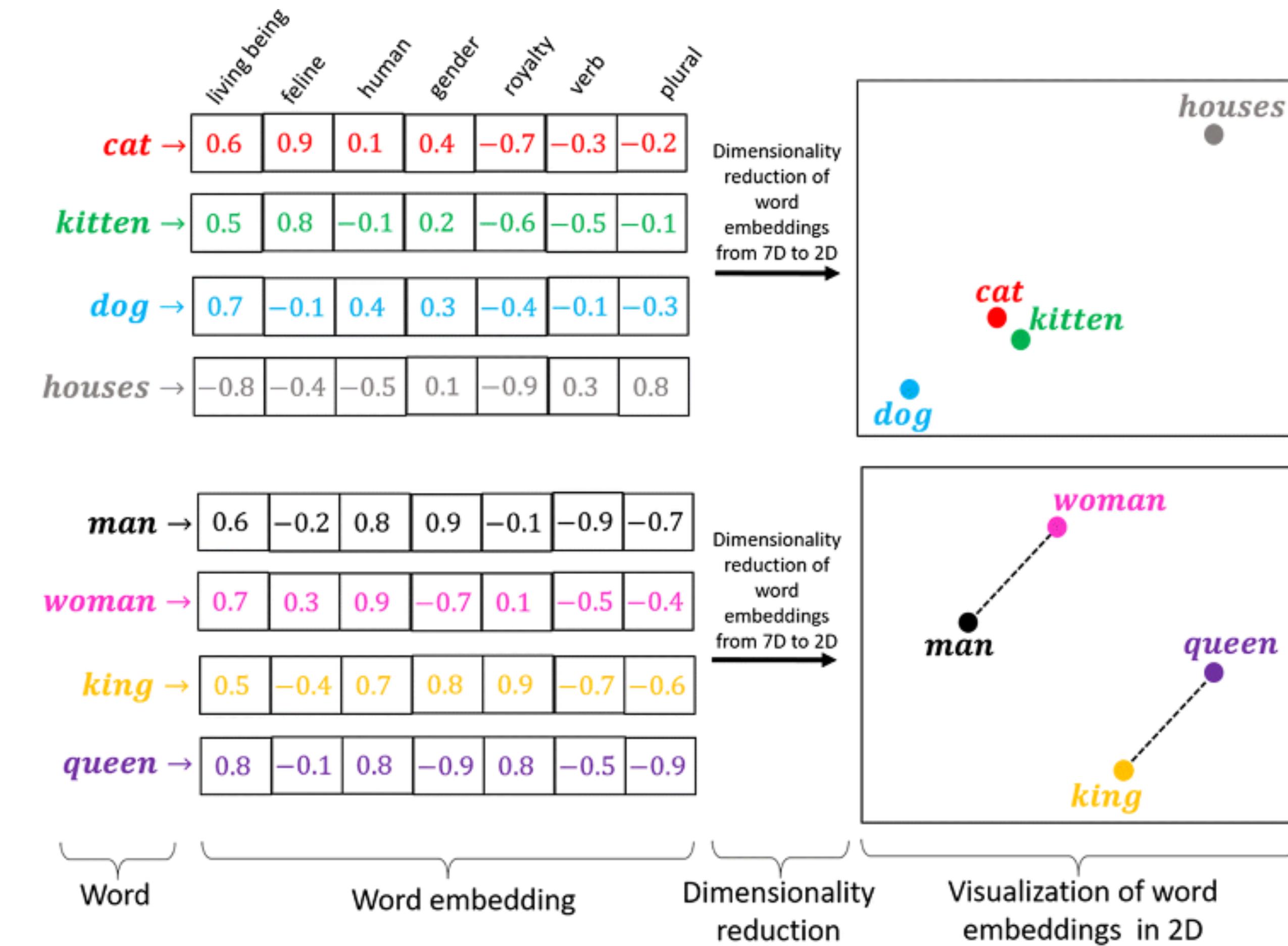
- Map each word / token to a continuous Euclidean space.
 - Discrete characters are difficult to use.

#14	living being	feline	human	gender	royalty	verb	plural
<i>cat</i> →	0.6	0.9	0.1	0.4	-0.7	-0.3	-0.2
<i>kitten</i> →	0.5	0.8	-0.1	0.2	-0.6	-0.5	-0.1
<i>dog</i> →	0.7	-0.1	0.4	0.3	-0.4	-0.1	-0.3
<i>houses</i> →	-0.8	-0.4	-0.5	0.1	-0.9	0.3	0.8
<i>man</i> →	0.6	-0.2	0.8	0.9	-0.1	-0.9	-0.7
<i>woman</i> →	0.7	0.3	0.9	-0.7	0.1	-0.5	-0.4
<i>king</i> →	0.5	-0.4	0.7	0.8	0.9	-0.7	-0.6
<i>queen</i> →	0.8	-0.1	0.8	-0.9	0.8	-0.5	-0.9

Word Word embedding

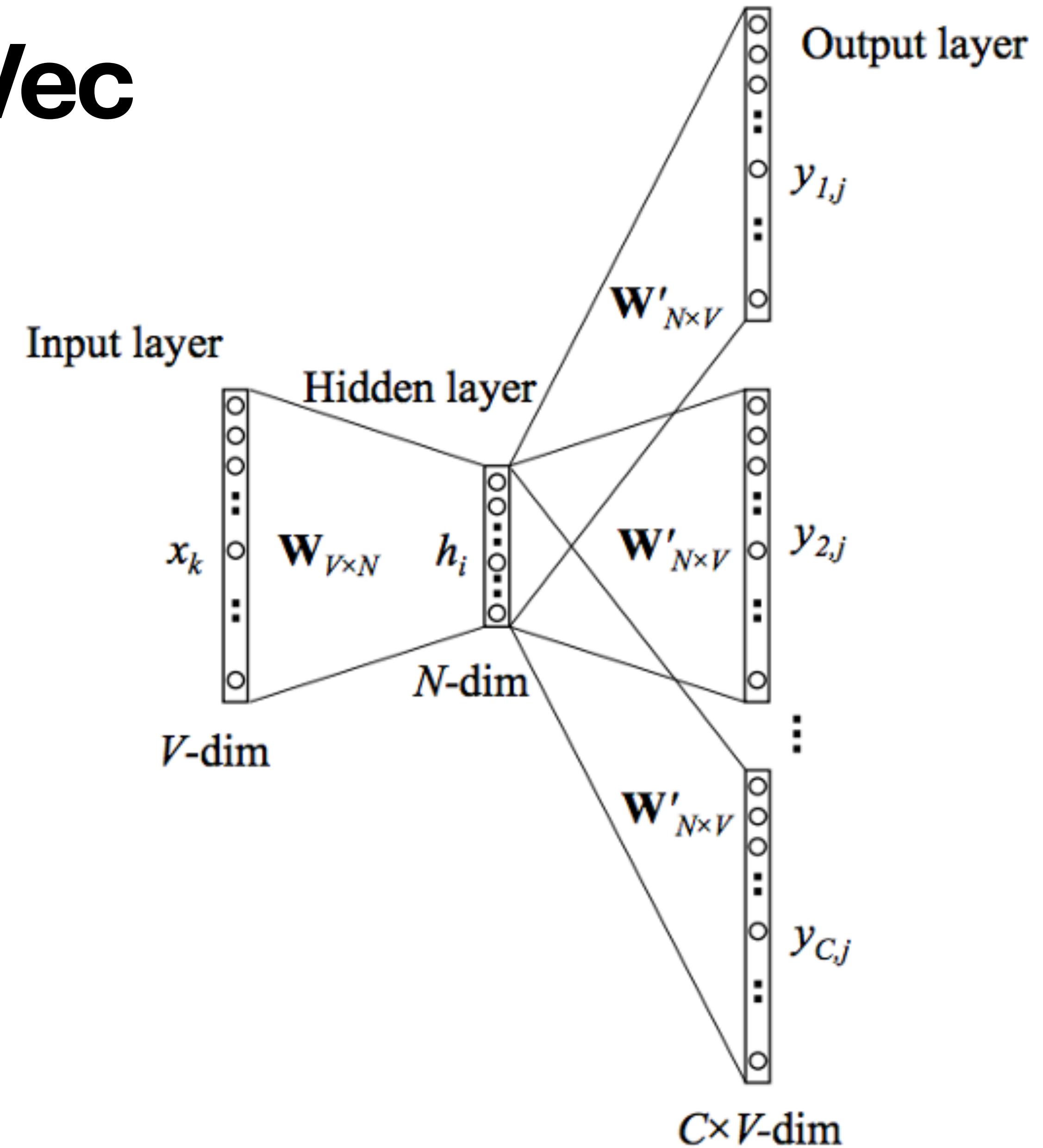
Text Embedding

- Map each word / token to a continuous Euclidean space.
 - Discrete characters are difficult to use.
- Surprisingly, learned embeddings are rich in semantics (e.g., cat & kitten)



Word2Vec

- One way to train text embeddings.
 - A **skip-gram** model



Word2Vec

- One way to train text embeddings.
 - A *skip-gram* model
- **Idea.** Predict the **surrounding words** from the center word.

The quick brown fox jumps over the lazy dog. → (the, quick)
(the, brown)

The quick brown fox jumps over the lazy dog. → (quick, the)
(quick, brown)
(quick, fox)

The quick brown fox jumps over the lazy dog. → (brown, the)
(brown, quick)
(brown, fox)
(brown, jumps)

The quick brown fox jumps over the lazy dog. → (fox, quick)
(fox, brown)
(fox, jumps)
(fox, over)

Word2Vec

- One way to train text embeddings.
 - A *skip-gram* model
- **Idea.** Predict the *surrounding words* from the center word.
- **Question.** Can we use similar idea to train the *joint embedding* of image and text data?

CLIP

- Trains such joint embedding using the transformer, and a lot of data
- **Scale matters.** Not the first attempt;
but the first to use very large dataset
 - Used 400 million image-text pairs.



french cat



french cat

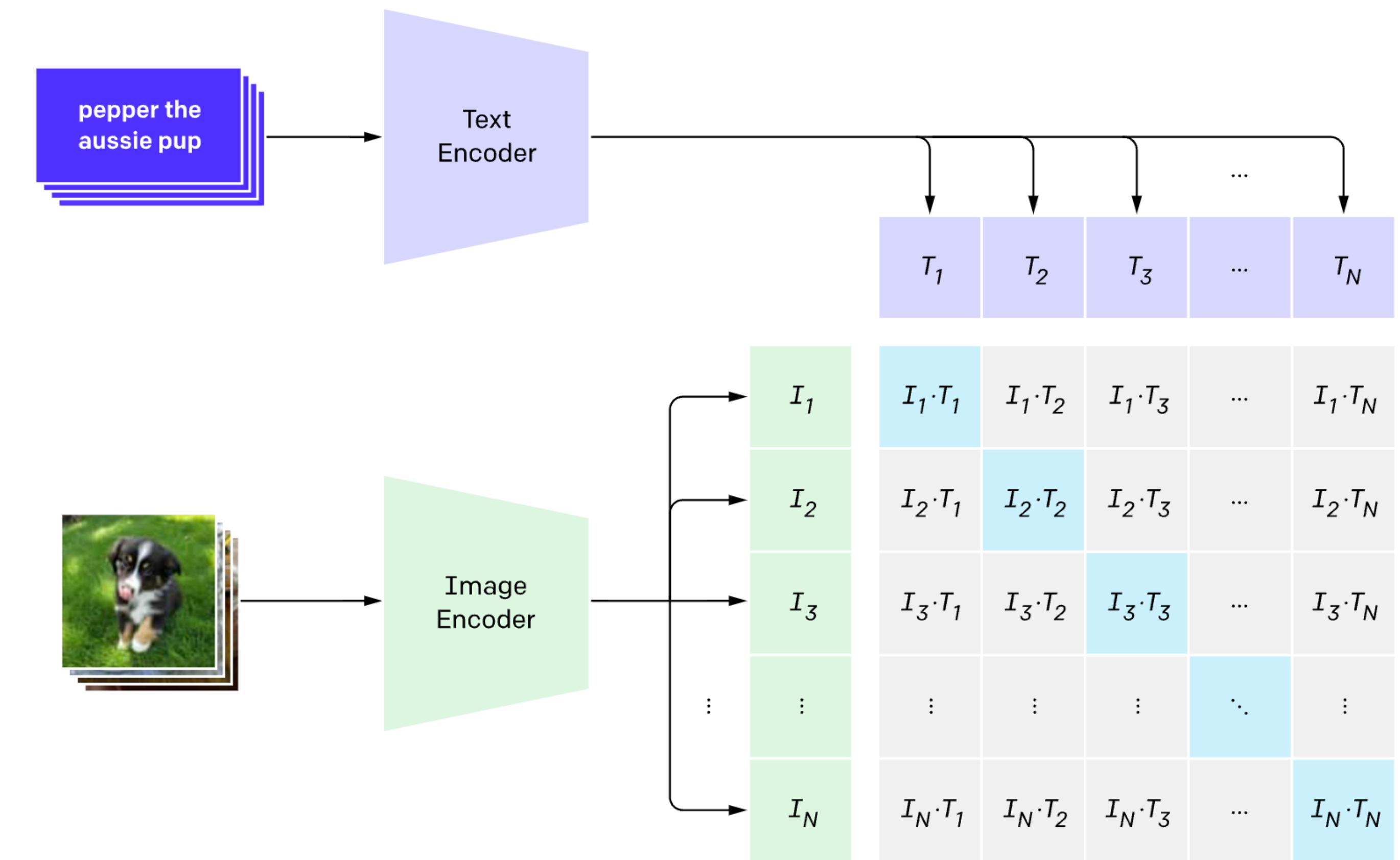
How to tell if your
feline is french. He
wears a b...イケメン猫モデル
「トキ・ナンタケツ
ト」がかっこいい-
NAVERまとめHilarious pics of funny
cats! funnycatsgif.com

CLIP

- **Algorithm.**

Contrastive pre-training

- Draws N image-text pairs as a batch.
- ***Increase*** the similarity between (I_i, T_i)
- ***Decrease*** the similarity between (I_i, T_j)



CLIP

	T_1	T_2	T_3	...	T_N
I_1	$I_1 \cdot T_1$	$I_1 \cdot T_2$	$I_1 \cdot T_3$...	$I_1 \cdot T_N$
I_2	$I_2 \cdot T_1$	$I_2 \cdot T_2$	$I_2 \cdot T_3$...	$I_2 \cdot T_N$
I_3	$I_3 \cdot T_1$	$I_3 \cdot T_2$	$I_3 \cdot T_3$...	$I_3 \cdot T_N$
:	:	:	:	..	:
I_N	$I_N \cdot T_1$	$I_N \cdot T_2$	$I_N \cdot T_3$...	$I_N \cdot T_N$

- **Concretely...**

Minimize the mixture of two losses.

- Image-to-text loss

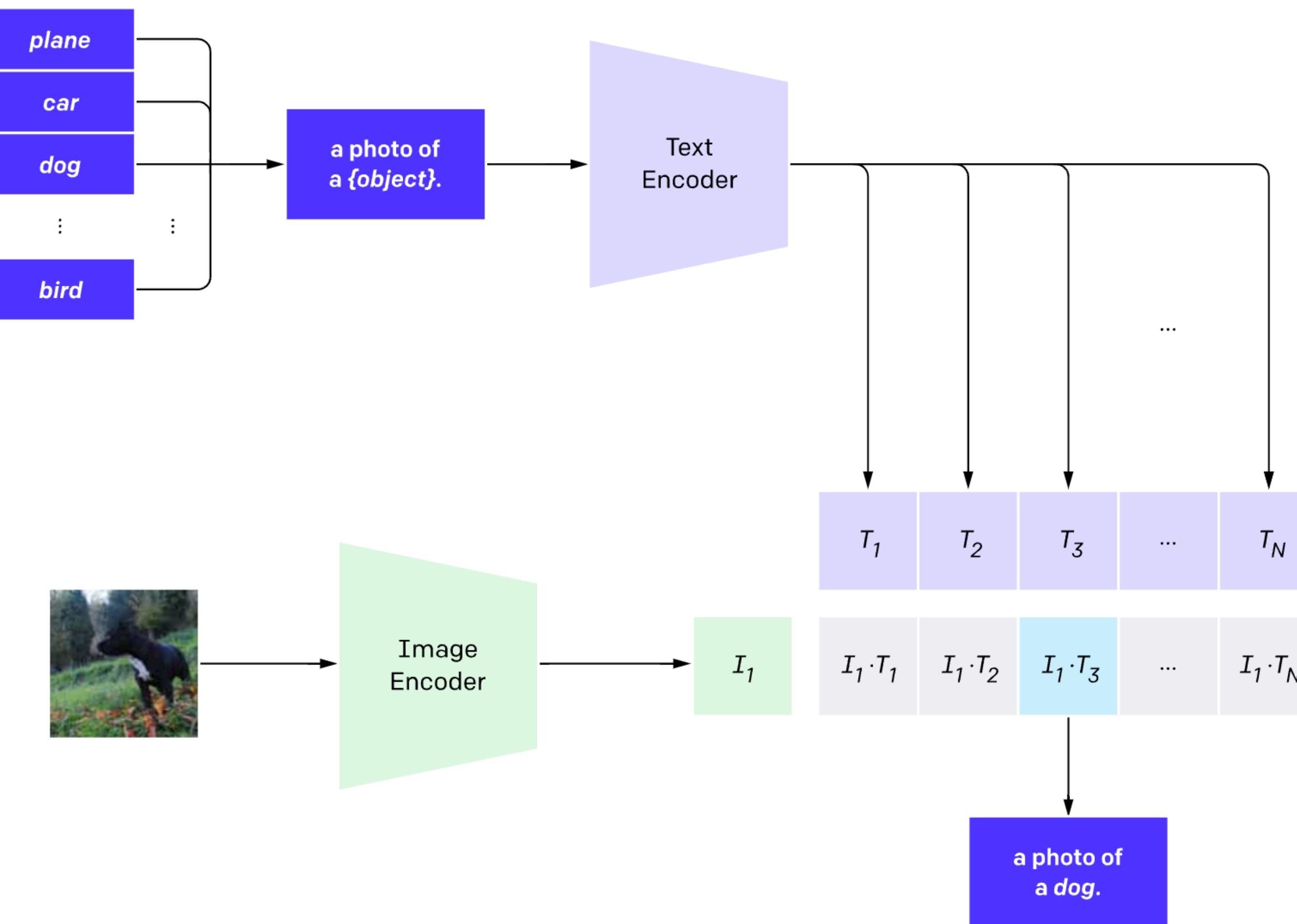
$$L_{i \rightarrow t} = - \sum_{i=1}^N \log \frac{\exp(I_i \cdot T_i / \tau)}{\sum_j \exp(I_j \cdot T_j / \tau)}$$

- Text-to-image loss

$$L_{i \rightarrow t} = - \sum_{j=1}^N \log \frac{\exp(I_j \cdot T_i / \tau)}{\sum_i \exp(I_i \cdot T_j / \tau)}$$

Use cases

- Given a good joint embedding, one can use it for **classification**.



- Enables an effective **zero-shot classification**.

SUN397

television studio (90.2%) Ranked 1 out of 397 labels



✓ a photo of a **television studio**.

✗ a photo of a **podium indoor**.

✗ a photo of a **conference room**.

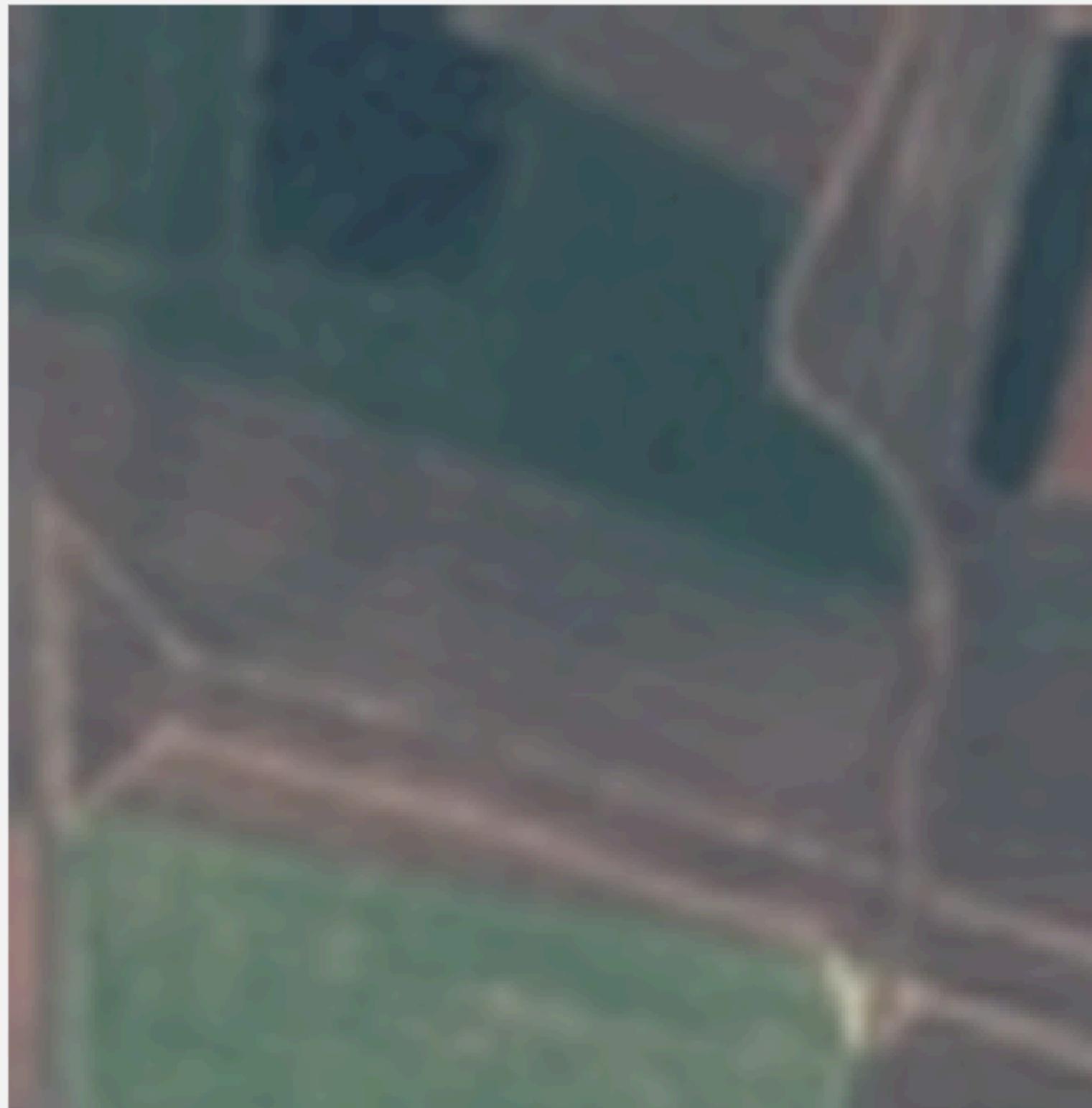
✗ a photo of a **lecture room**.

✗ a photo of a **control room**.

- Enables an effective **zero-shot classification**.
 - Especially when we have **good prompts**.

EuroSAT

annual crop land (46.5%) Ranked 4 out of 10 labels



- ✗ a centered satellite photo of **permanent crop land**.
- ✗ a centered satellite photo of **pasture land**.
- ✗ a centered satellite photo of **highway or road**.
- ✓ a centered satellite photo of **annual crop land**.
- ✗ a centered satellite photo of **brushland or shrubland**.

Other use cases

- CLIP + LLMs = Captioning Models



A politician receives a gift from politician.



A collage of different colored ties on a white background.



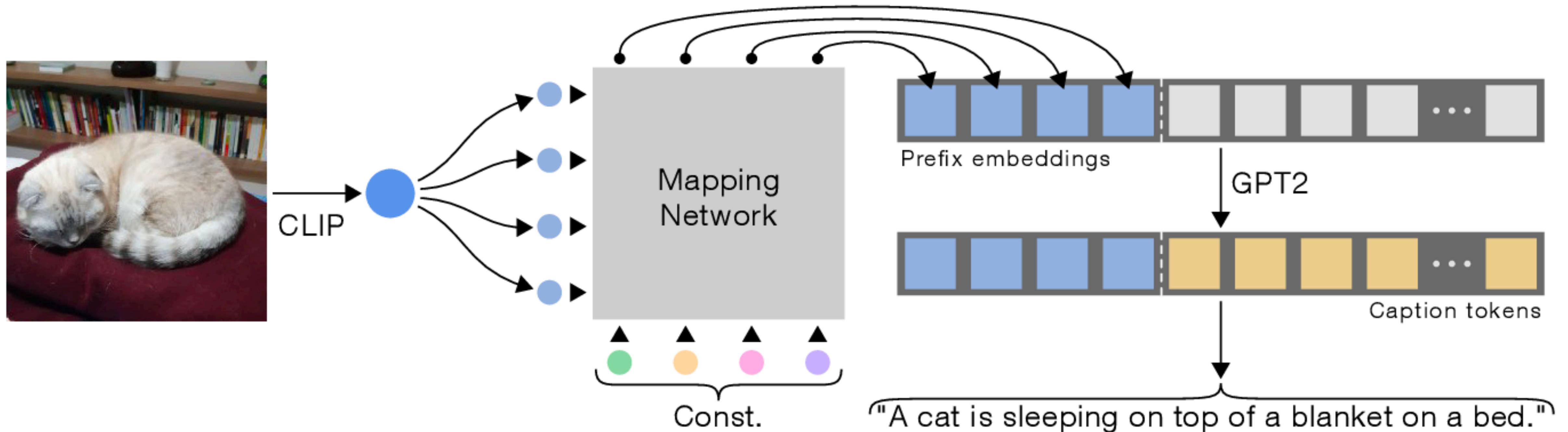
Silhouette of a woman practicing yoga on the beach at sunset.



Aerial view of a road in autumn.

Other use cases

- CLIP + LLMs = Captioning Models

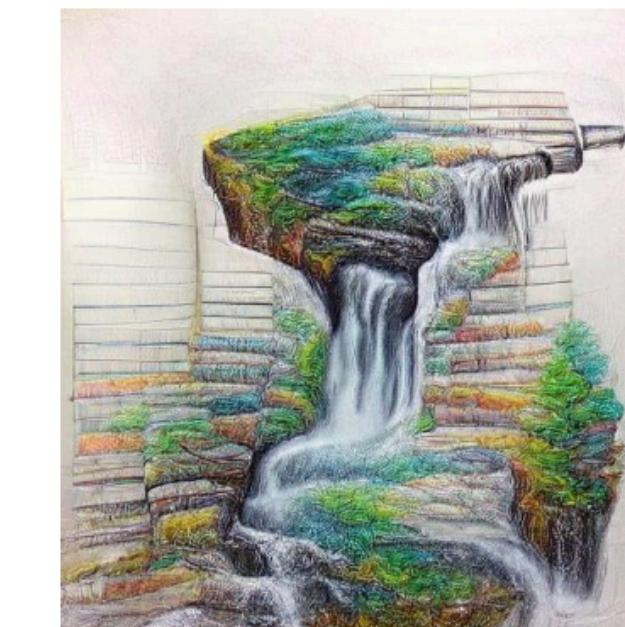


Other use cases

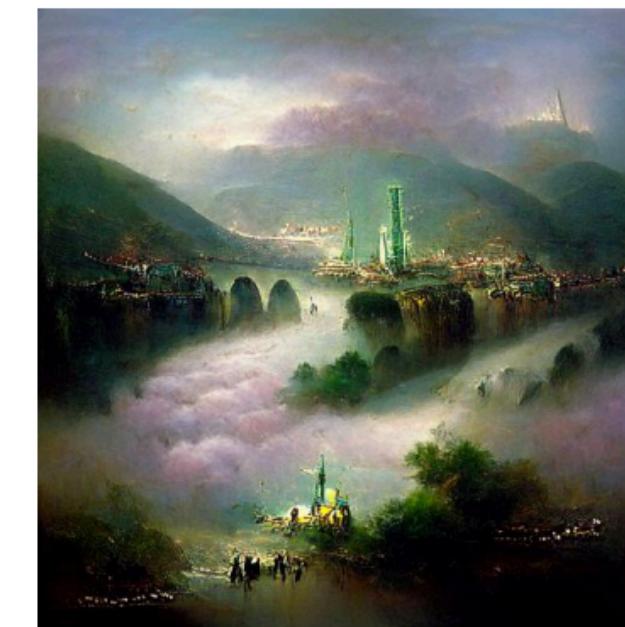
- CLIP + GAN = Text-based Image Generation



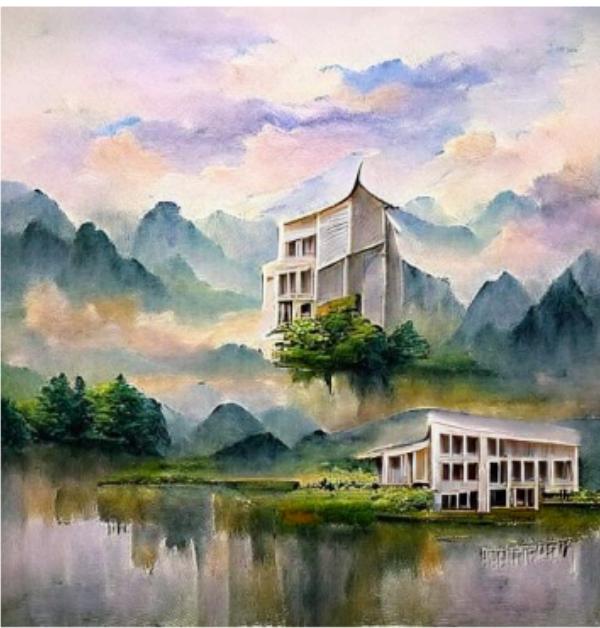
(a) Oil painting of a candy dish of glass candies, mints, and other assorted sweets



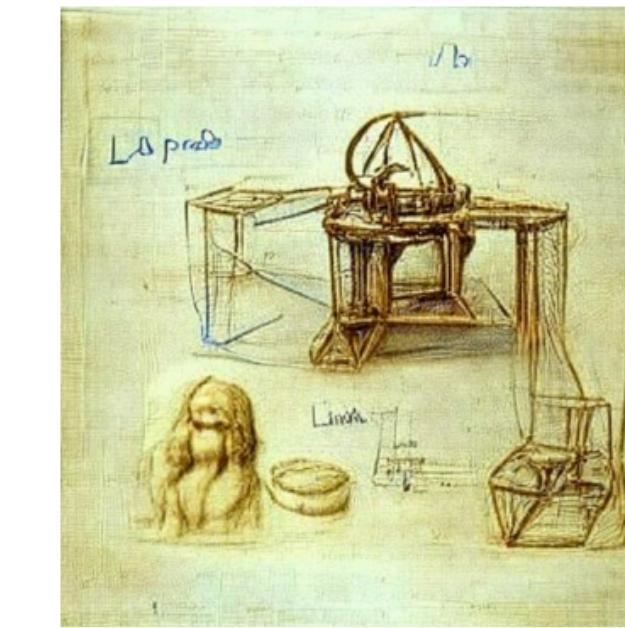
(b) A colored pencil drawing of a waterfall



(c) A fantasy painting of a city in a deep valley by Ivan Aivazovsky



(d) A beautiful painting of a building in a serene landscape



(e) sketch of a 3D printer by Leonardo da Vinci



(f) an autogyro flying car, trending on artstation

Other use cases

- CLIP + GAN = Text-based Image Generation **and editing**

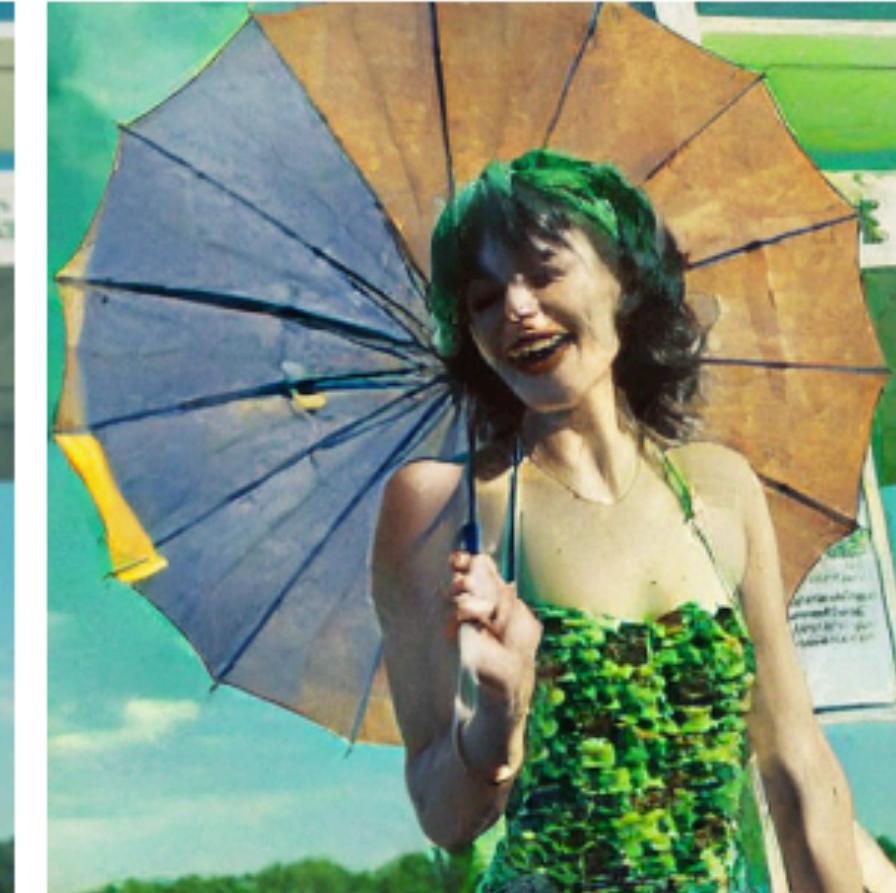
Instruction

“Green”

Original



VQGAN-CLIP

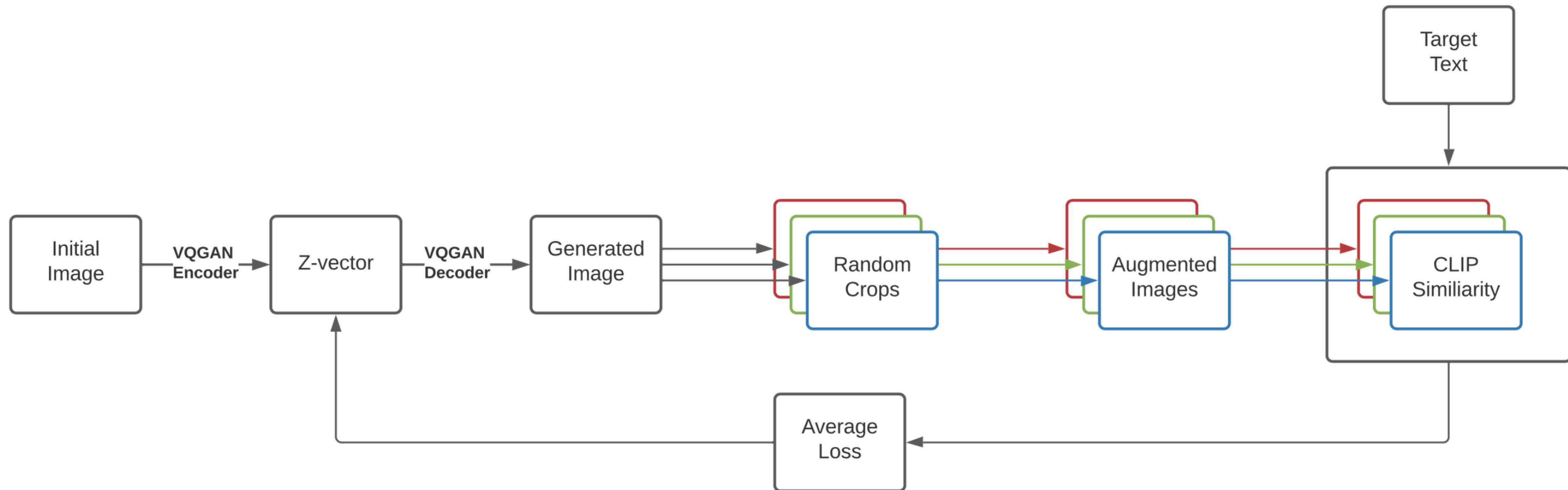


“Red Bus”



Other use cases

- CLIP + GAN = Text-based Image Generation and editing



Other use cases

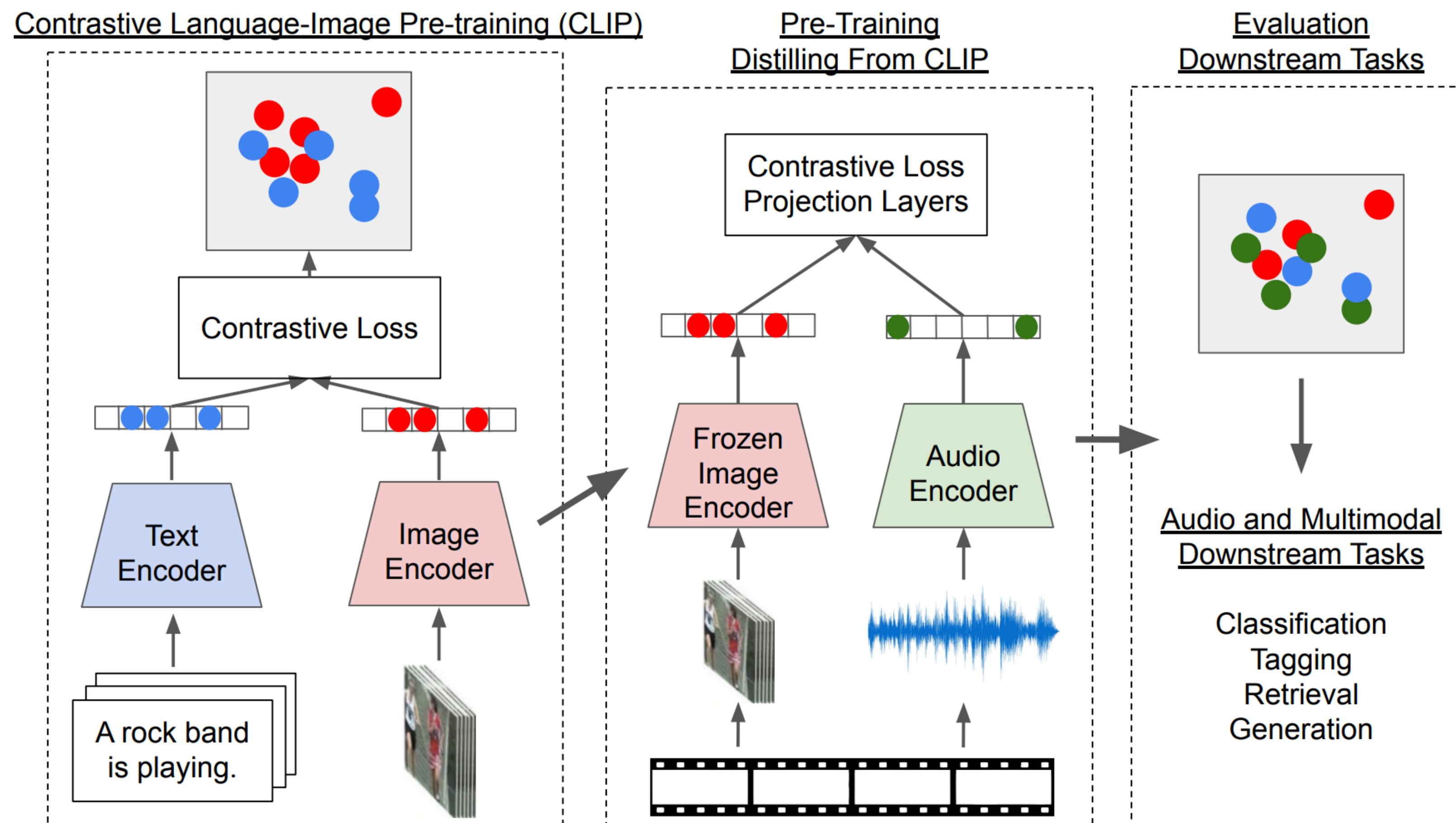
- CLIP + GAN + Audio data

Text/Audio to Image Generation with VQGAN-CLIP



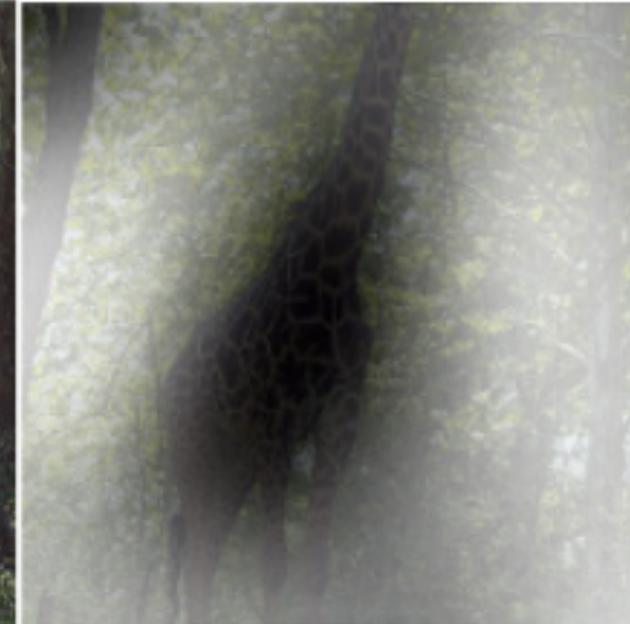
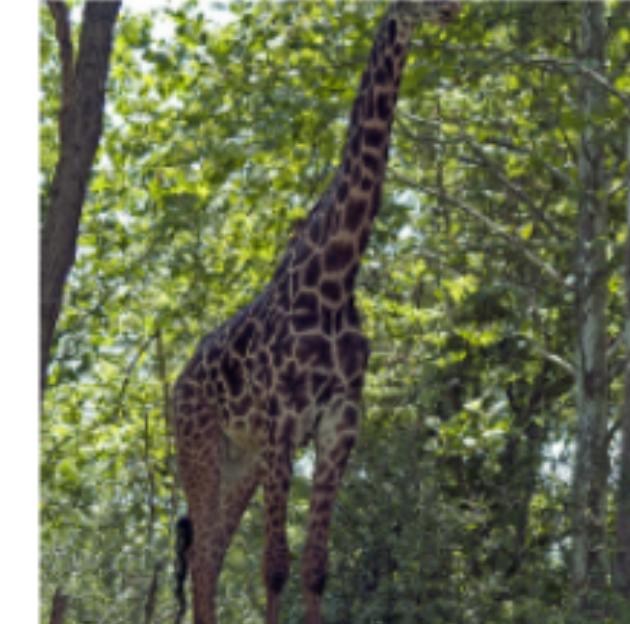
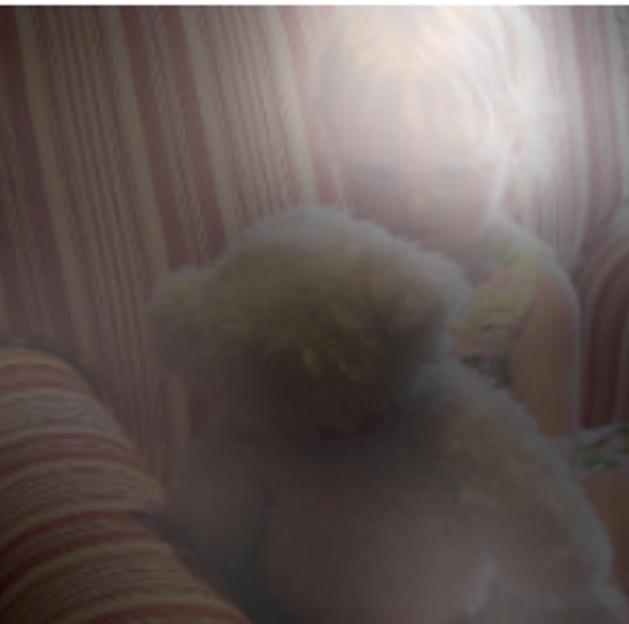
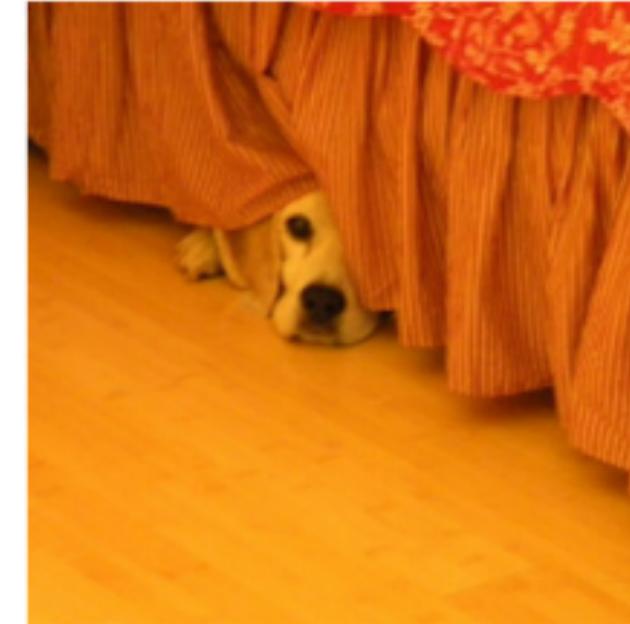
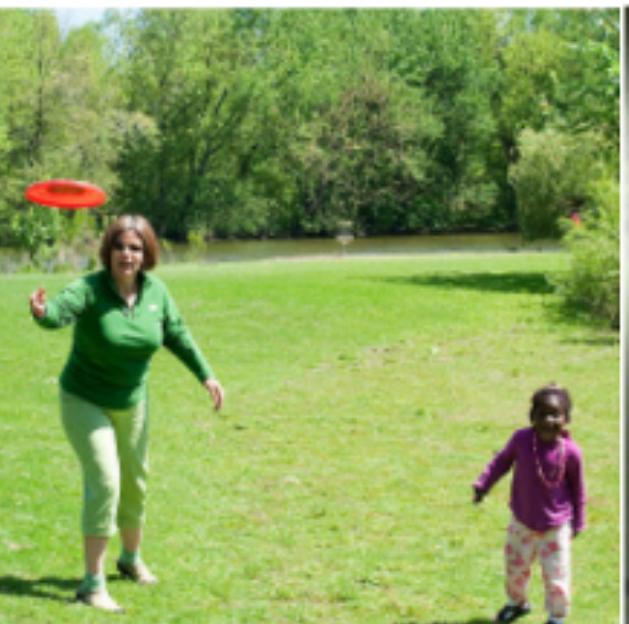
Other use cases

- CLIP + GAN + Audio data



Even older examples...

- Cross-modal reasoning



A little girl sitting on a bed with a teddy bear.

A group of people sitting on a boat in the water.

A giraffe standing in a forest with trees in the background.

Even older examples...

- Food recipe retrieval

Query Image



True ingr.

whole milk
half - and - half cr
white sugar
lemon extract
ground cinnamon
frozen blueberries
vanilla wafers
ice cubes

Retrieved ingr. Retrieved Image

berries
strawberry yogurt
banana
milk
white sugar



butter
garlic cloves
all - purpose flour
kosher salt
milk
chicken broth
mozzarella cheese
parmesan cheese
onion

1 box any pasta you
ground beef
1 envelope taco seas
water
1/2 packages cream c
cheese



Even older examples...

- Image retrieval, with analogies



- blue + red =



- blue + yellow =



- yellow + red =



- white + red =

Nearest images



Cheers