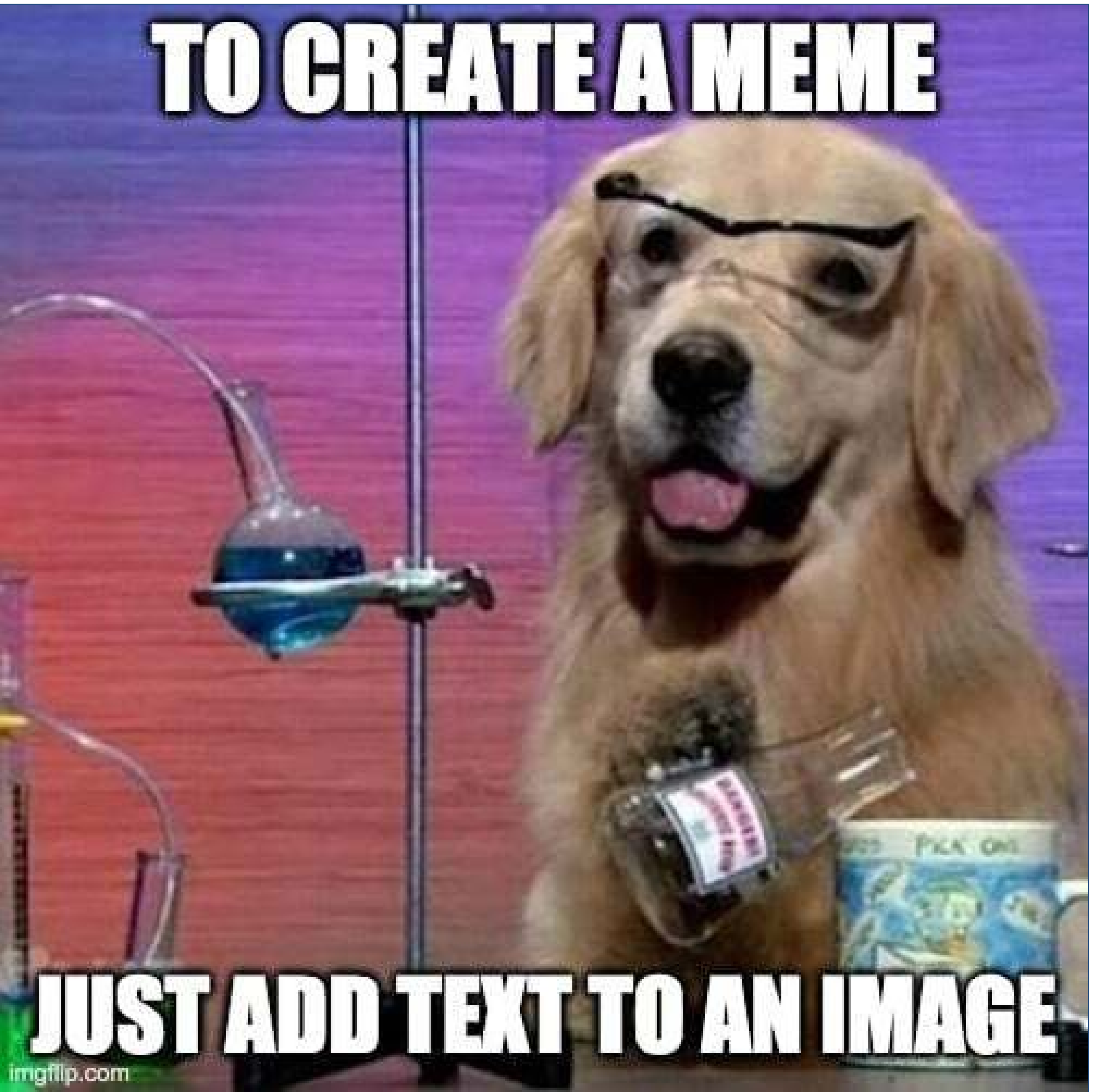
CMSC 475/675 Neural Networks

Lecture 9 continued: "Multimodal" Representations



UMBC

Recap: Contrastive Learning

Examples of Contrastive Pairs

Contrastive Learning

Problem 1: How to compute similarity if we don't have labels for images?
Solution? Euclidean Distance between features $\|\phi(x_1) - \phi(x_2)\|_2$

Problem 2: Objective Function ?

Similar images should have similar features   Dissimilar images should have dissimilar features

Contrastive Learning Formulation

- We want:

$$\text{score}(f(x), f(x^+)) >> \text{score}(f(x), f(x^-))$$

x: reference sample; x^+ positive sample; x^- negative sample

- Objective:

Loss function given 1 positive sample and N - 1 negative samples:

$$L = -\mathbb{E}_X \left[\log \frac{\exp(s(f(x), f(x^+))}{\exp(s(f(x), f(x^+)) + \sum_{j=1}^{N-1} \exp(s(f(x), f(x_j^-))} \right]$$

Contrastive Learning with Data Augmentation

figure: Ranjay Krishna

Loss function given 1 positive sample and N - 1 negative samples:

$$L = -\mathbb{E}_X \left[\log \frac{\exp(s(f(x), f(x^+)))}{\exp(s(f(x), f(x^+))) + \sum_{j=1}^{N-1} \exp(s(f(x), f(x_j^-)))} \right]$$

x

x^+

x

x_1^-

x_2^-

x_3^-

...

SimCLR: A Simple Framework for Contrastive learning

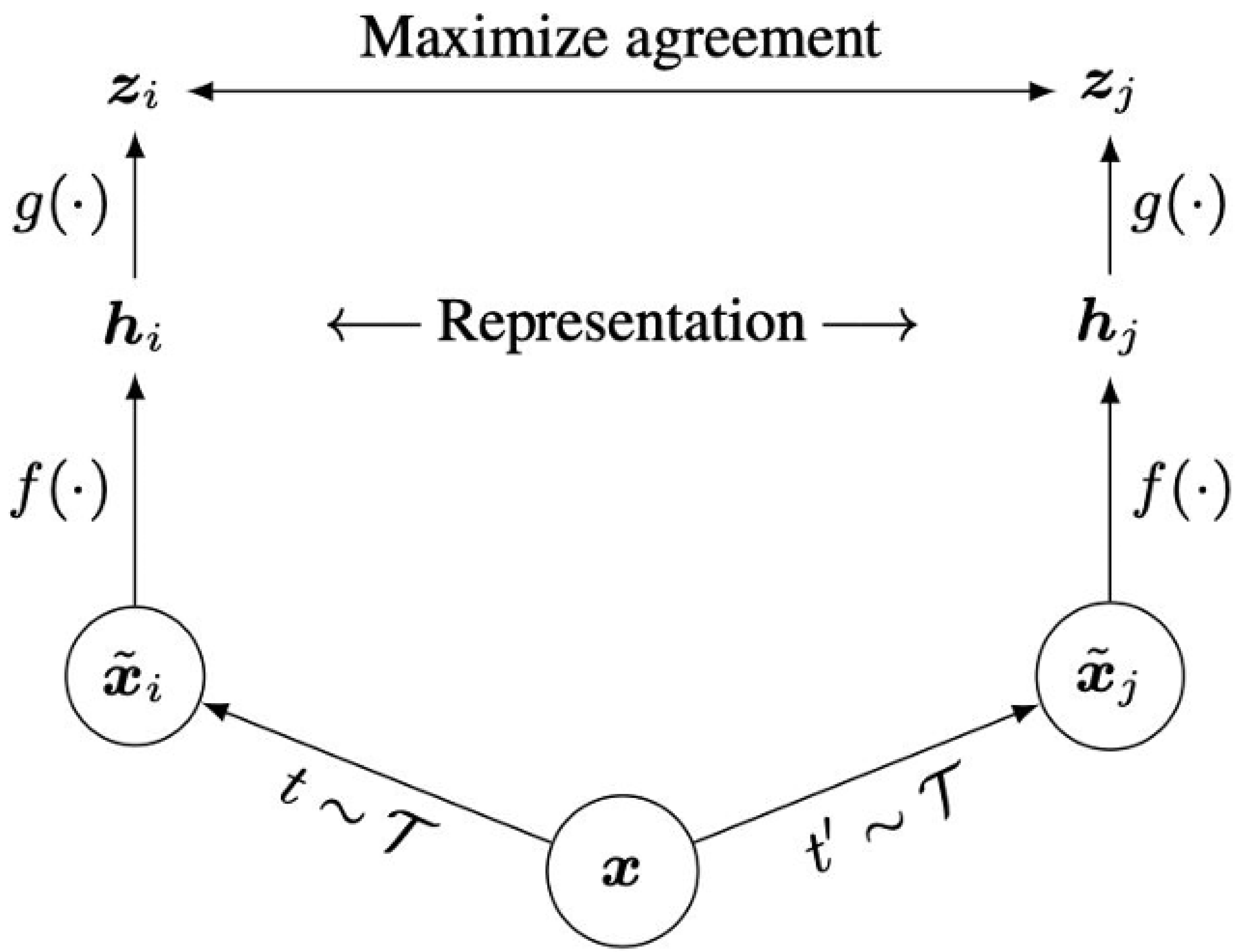
Cosine similarity as the score function:

$$s(u, v) = \frac{u^T v}{||u|| ||v||}$$

Use a projection network ***h(·)*** to project features to a space where contrastive learning is applied

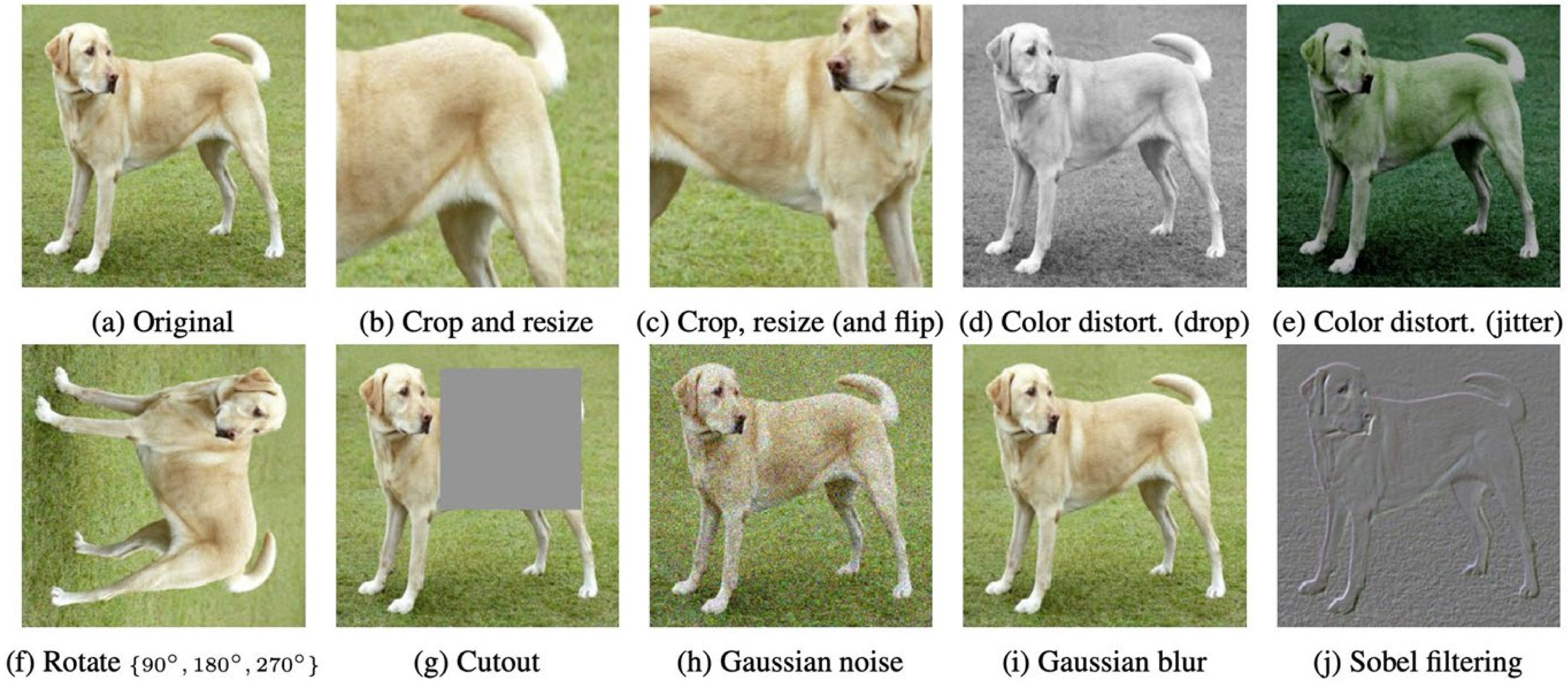
Generate positive samples through data augmentation:

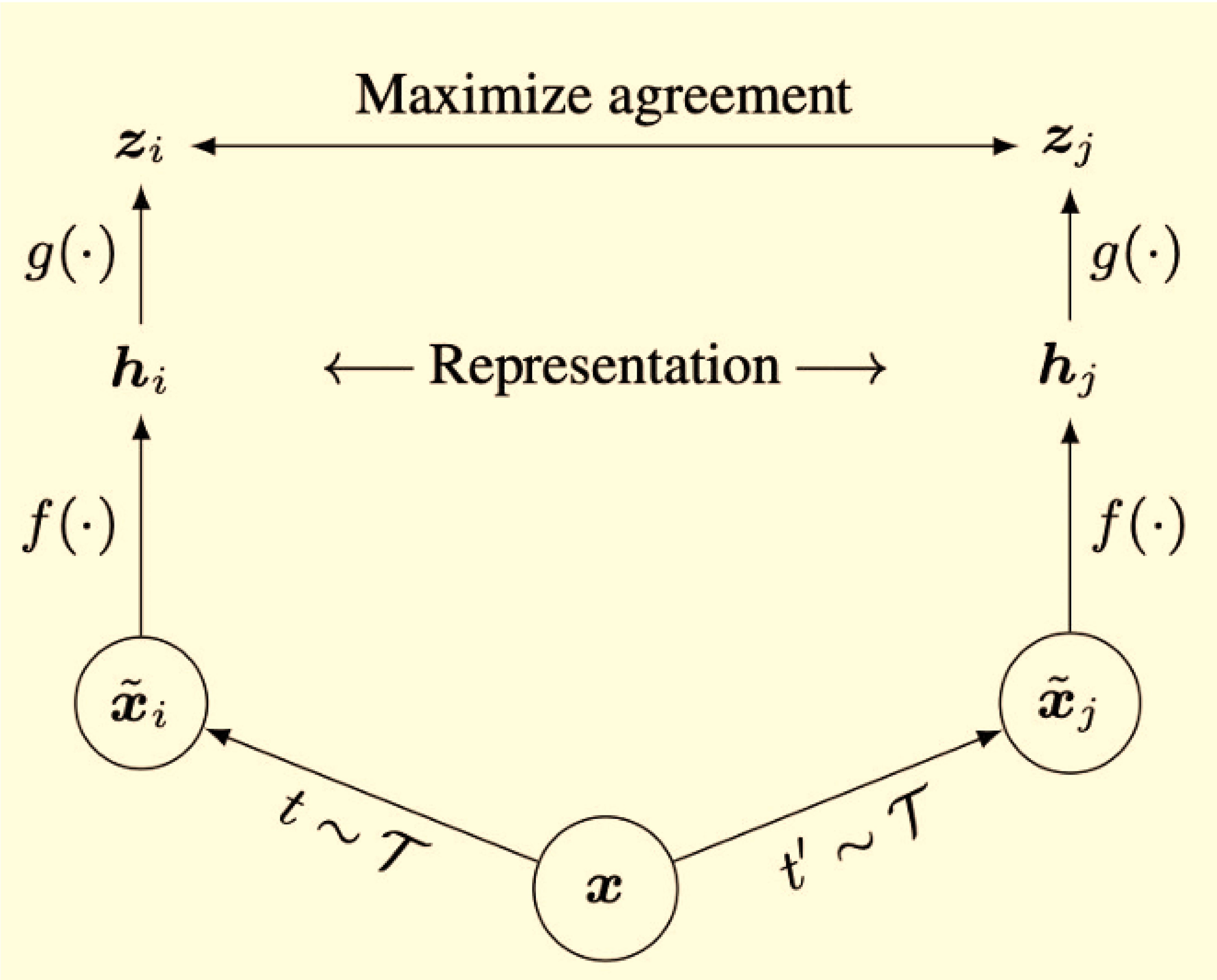
- random cropping, random color distortion, and random blur.



Source: [Chen et al., 2020]

SimCLR: Data Augmentation Strategies





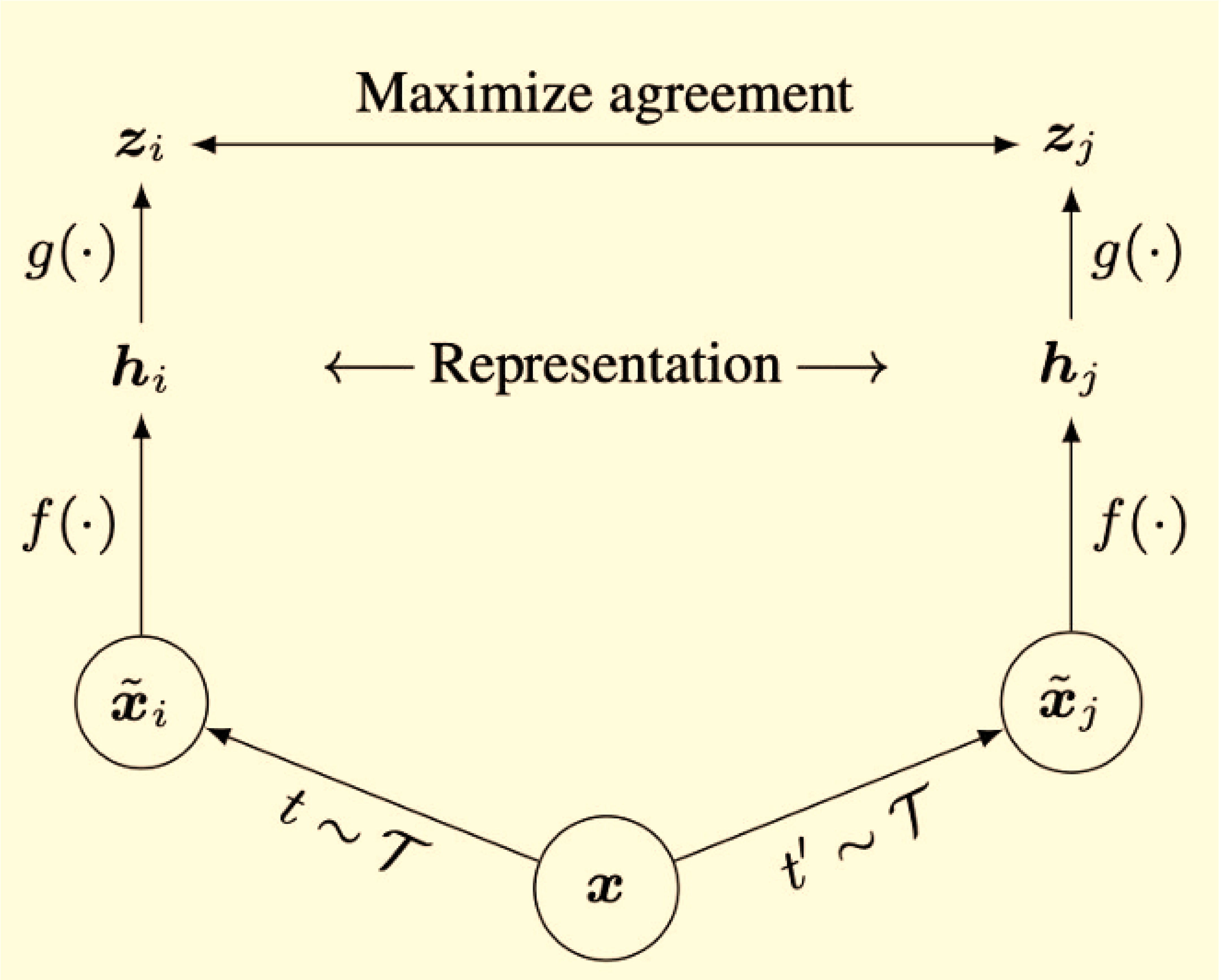
SimCLR: Algorithm Sketch

Algorithm 1 SimCLR's main learning algorithm.

input: batch size N, constant τ, structure of f, g, $\mathcal{T}$.
for sampled minibatch $\{\boldsymbol{x}_k\}_{k=1}^{N}$ **do**
  for all $k \in \{1, \ldots, N\}$ **do**
    $\boxed{\text{draw two augmentation functions } t \sim \mathcal{T}, t' \sim \mathcal{T}}$
    # the first augmentation
    $\boxed{\tilde{\boldsymbol{x}}_{2k-1} = t(\boldsymbol{x}_k)}$
    $\boldsymbol{h}_{2k-1} = f(\tilde{\boldsymbol{x}}_{2k-1})$ # representation
    $\boldsymbol{z}_{2k-1} = g(\boldsymbol{h}_{2k-1})$ # projection
    # the second augmentation
    $\boxed{\tilde{\boldsymbol{x}}_{2k} = t'(\boldsymbol{x}_k)}$
    $\boldsymbol{h}_{2k} = f(\tilde{\boldsymbol{x}}_{2k})$ # representation
    $\boldsymbol{z}_{2k} = g(\boldsymbol{h}_{2k})$ # projection
  end for
  for all $i \in \{1, \ldots, 2N\}$ and $j \in \{1, \ldots, 2N\}$ **do**
    $s_{i,j} = \boldsymbol{z}_i^\top \boldsymbol{z}_j / (\|\boldsymbol{z}_i\| \|\boldsymbol{z}_j\|)$ # pairwise similarity
  end for
  define $\ell(i,j)$ **as** $\boxed{\ell(i,j) = -\log \frac{\exp(s_{i,j}/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(s_{i,k}/\tau)}}$
  $\mathcal{L} = \frac{1}{2N} \sum_{k=1}^{N} [\ell(2k-1, 2k) + \ell(2k, 2k-1)]$
  update networks f and g to minimize $\mathcal{L}$
end for
return encoder network $f(\cdot)$, and throw away $g(\cdot)$

Generate a positive pair by sampling data augmentation functions

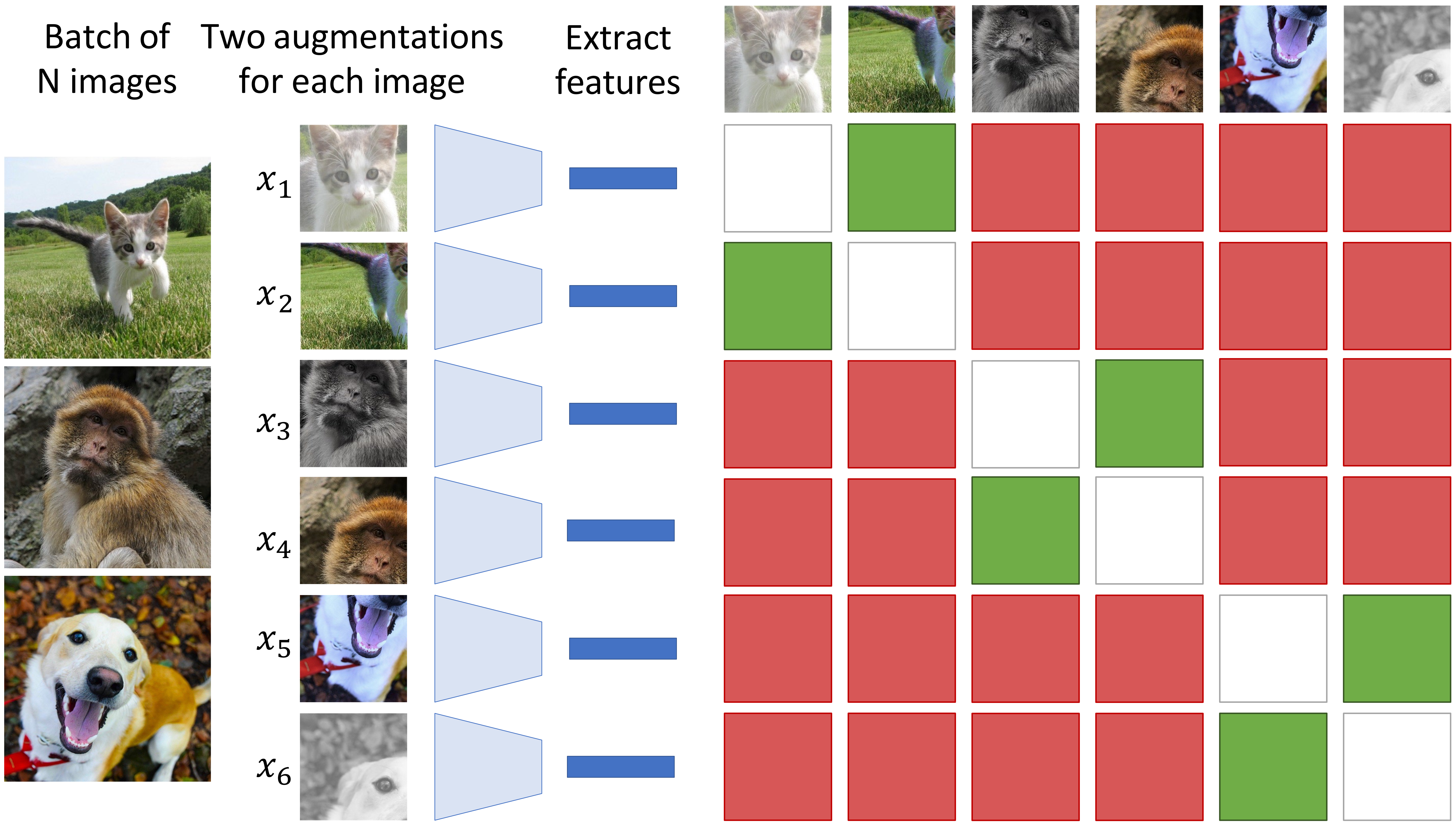
Iterate through and use each of the 2N sample as reference, compute average loss



InfoNCE loss:
Use all non-positive samples in the batch as x^-

Source: [Chen et al., 2020]

SimCLR Training



Each image tries to predict which of the *other* 2N-1 images came from the same original image

Similarity between x_i and x_j:

$$s_{i,j} = \frac{\phi(x_i)^T \phi(x_j)}{\|\phi(x_i)\| \cdot \|\phi(x_i)\|}$$

If (x_i, x_j) is a positive pair, then loss for x_i is:

$$L_i = -\log \frac{\exp(s_{i,j}/\tau)}{\sum_{\substack{k=1 \\ k \neq i}}^{2N} \exp(s_{i,k}/\tau)}$$

(τ is a *temperature*)

Interpretation: Cross-entropy loss over the other 2N-1 elements in the batch!

Hadsell et al, "Dimensionality Reduction by Learning and Invariant Mapping", CVPR 2006
Wu et al, "Unsupervised Feature Learning by Non-Parametric Instance-Level Discrimination", CVPR 2018
Van den Oord et al, "Representation Learning with Contrastive Predictive Coding", NeurIPS 2018
Hjelm et al, "Learning deep representations by mutual information estimation and maximization", ICLR 2019
Bachman et al, "Learning Representations by Maximizing Mutual Information Across Views", NeurIPS 2019
Henaff et al, "Data-Efficient Image Recognition with Contrastive Predictive Coding", ICML 2020
Tian et al, "Contrastive Multiview Coding", ECCV 2020
He et al, "Momentum Contrast for Unsupervised Visual Representation Learning", CVPR 2020
Chen et al, "A Simple Framework for Contrastive Learning of Visual Representations", ICML 2020

But how did you get the pretraining data?

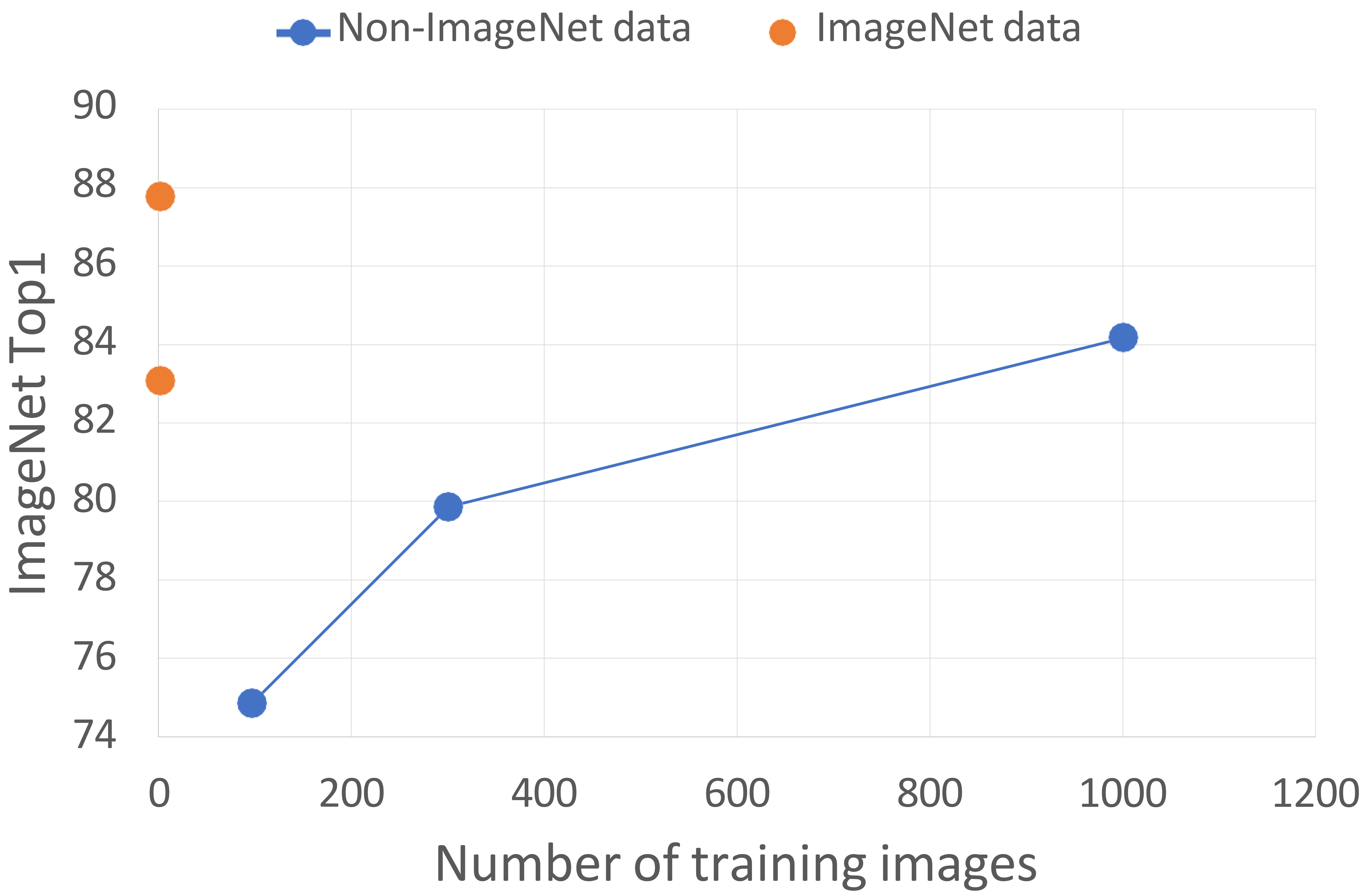
The motivation of SSL is scaling to large data that can’t be labeled

Most papers pretrain on (unlabeled) ImageNet, then evaluate on ImageNet!

Unlabeled ImageNet is still curated: single object per image, balanced classes

Self-Supervised Learning on larger datasets hasn’t been as successful as NLP

Idea: What if we go beyond isolated images?



Caron et al, “Unsupervised pre-training of images features on non-curated data”, ICCV 2019
Chen et al, “Big self-supervised models are strong semi-supervised learners”, NeurIPS 2020
Dosovitskiy et al, “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale”, ICLR 2021
Goyal et al, “Self-supervised Pretraining of Visual Features in the Wild”, arXiv 2021
He et al, “Masked Autoencoders are Scalable Vision Learners”, arXiv 2021

Multimodal Self-Supervised Learning

Don't learn from isolated images -- take images together with some **context**

Video: Image together with adjacent video frames

Agrawal et al, "Learning to See by Moving", ICCV 2015
Wang et al, "Unsupervised Learning of Visual Representations using Videos", ICCV 2015
Pathak et al, "Learning Features by Watching Objects Move", CVPR 2017

Sound: Image with audio track from video

Owens et al, "Ambient Sound Provides Supervision for Visual Learning", ECCV 2016
Arandjelovic and Zisserman, "Look, Listen and Learn", ICCV 2017

3D: Image with depth map or point cloud

Xie et al, "PointContrast: Unsupervised Pre-training for 3D Point Cloud Understanding", ECCV 2020
Zhang et al, "Self-supervised pretraining of 3D features on any point-cloud", CVPR 2021

Language: Image with natural-language text

Sariyildiz et al, "Learning Visual Representations with Caption Annotations", ECCV 2020
Desai and Johnson, "VirTex: Learning Visual Representations from Textual Annotations", CVPR 2021
Radford et al, "Learning Transferable Visual Models form Natural Language Supervision", ICML 2021
Jia et al, "Scaling up Visual and Vision-Language Representation Learning with Noisy Text Supervision", ICLR 2021
Desai et al, "RedCaps: Web-curated Image-Text data created by the people, for the people", NeurIPS 2021

Why Language?

Large dataset of
(image, caption)

a dog with his head out the window of the car

a black and orange cat is resting on a keyboard and yellow back scratcher

1. **Semantic density**: Just a few words give rich information

2. **Universality**: Language can describe any concept

3. **Scalability**: Non-experts can easily caption images; data can also be collected from the web at scale

RedCaps: Images and Captions from Reddit

r/birdpics: male northern cardinal | **r/crafts:** my mom tied this mouse | **r/itookapicture:** itap of the taj mahal | **r/perfectfit:** this lemon in my drink | **r/shiba:** mlem!

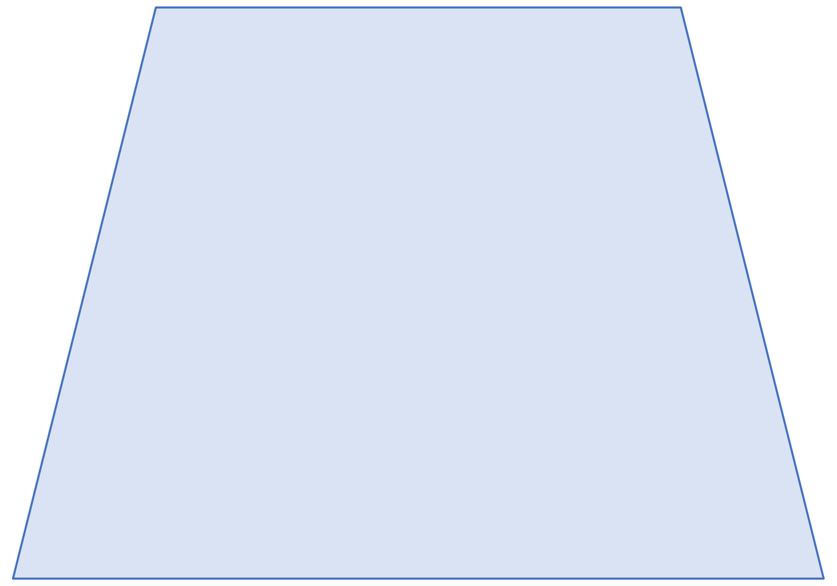
Data from 350 manually-chosen subreddits
12M high-quality (image, caption) pairs

Desai, Kaul, Aysola, and Johnson, NeurIPS Datasets & Benchmarks Track, 2021

For now: Assume you can learn language representations (We will study language modeling in the next lecture)

Computer Vision

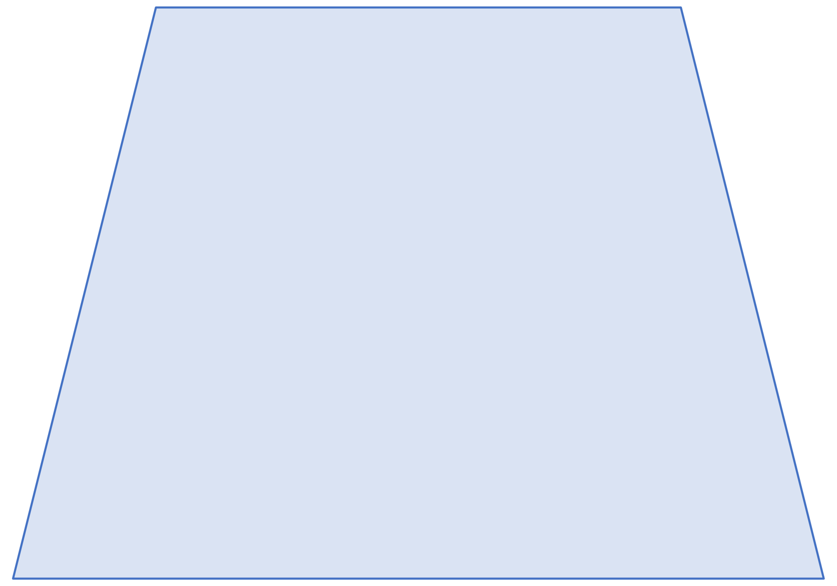
Image Features:
H x W x C



Input Image

Natural Language Processing

Word Features
L x C



A white and gray cat standing outside on the grass

Input Sentence (L words)

Contrastive Learning with Vision-Language Data

OpenAI

January 5, 2021 Milestone

CLIP: Connecting text and images

Contrastive Learning with Vision-Language Data

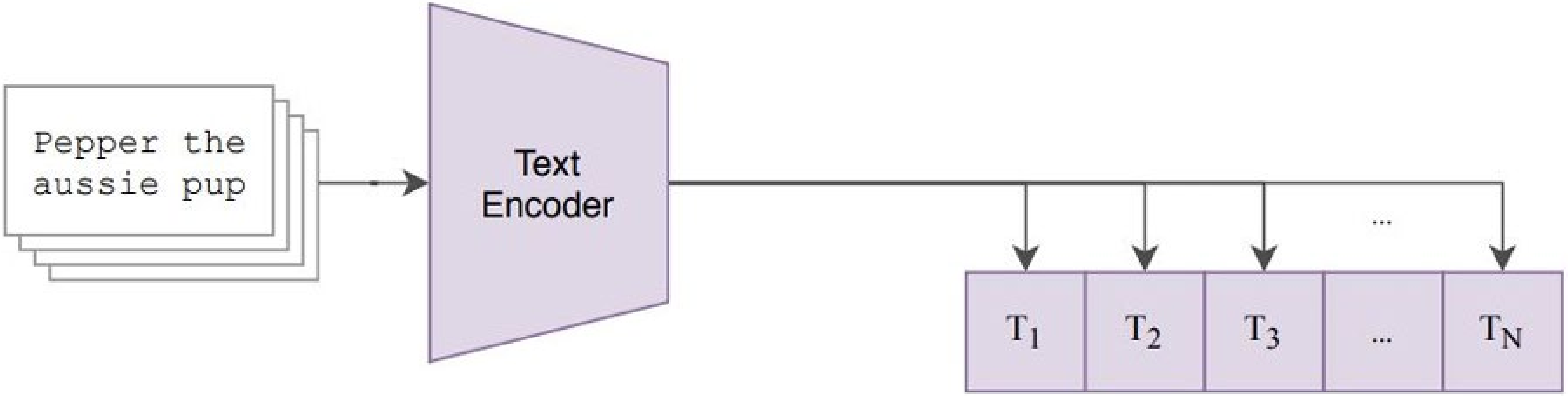
Learning Transferable Visual Models From Natural Language Supervision

Alec Radford [*1] Jong Wook Kim [*1] Chris Hallacy [1] Aditya Ramesh [1] Gabriel Goh [1] Sandhini Agarwal [1]
Girish Sastry [1] Amanda Askell [1] Pamela Mishkin [1] Jack Clark [1] Gretchen Krueger [1] Ilya Sutskever [1]

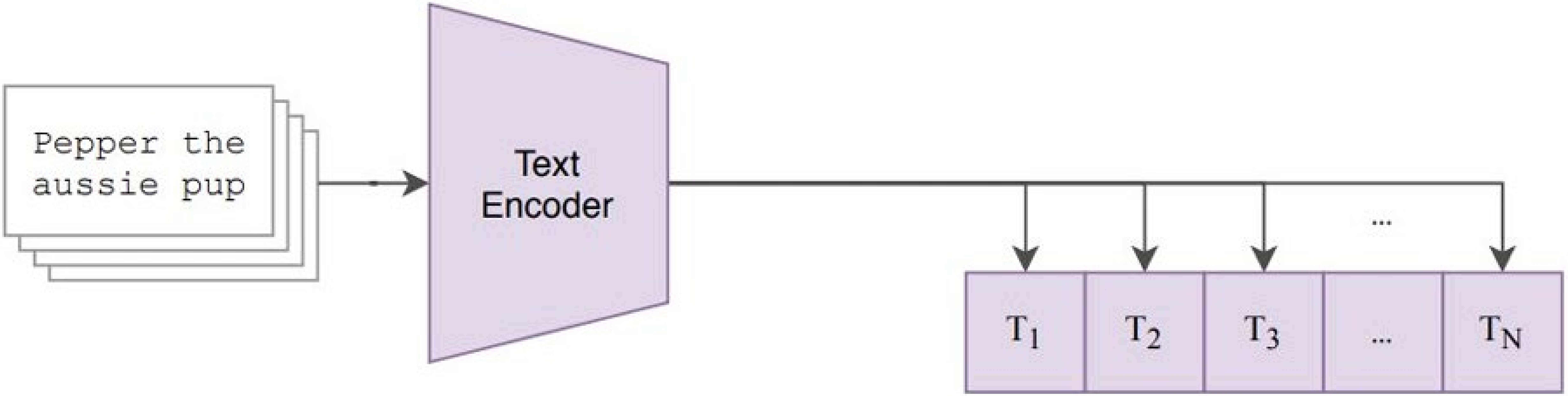
ICML | 2021

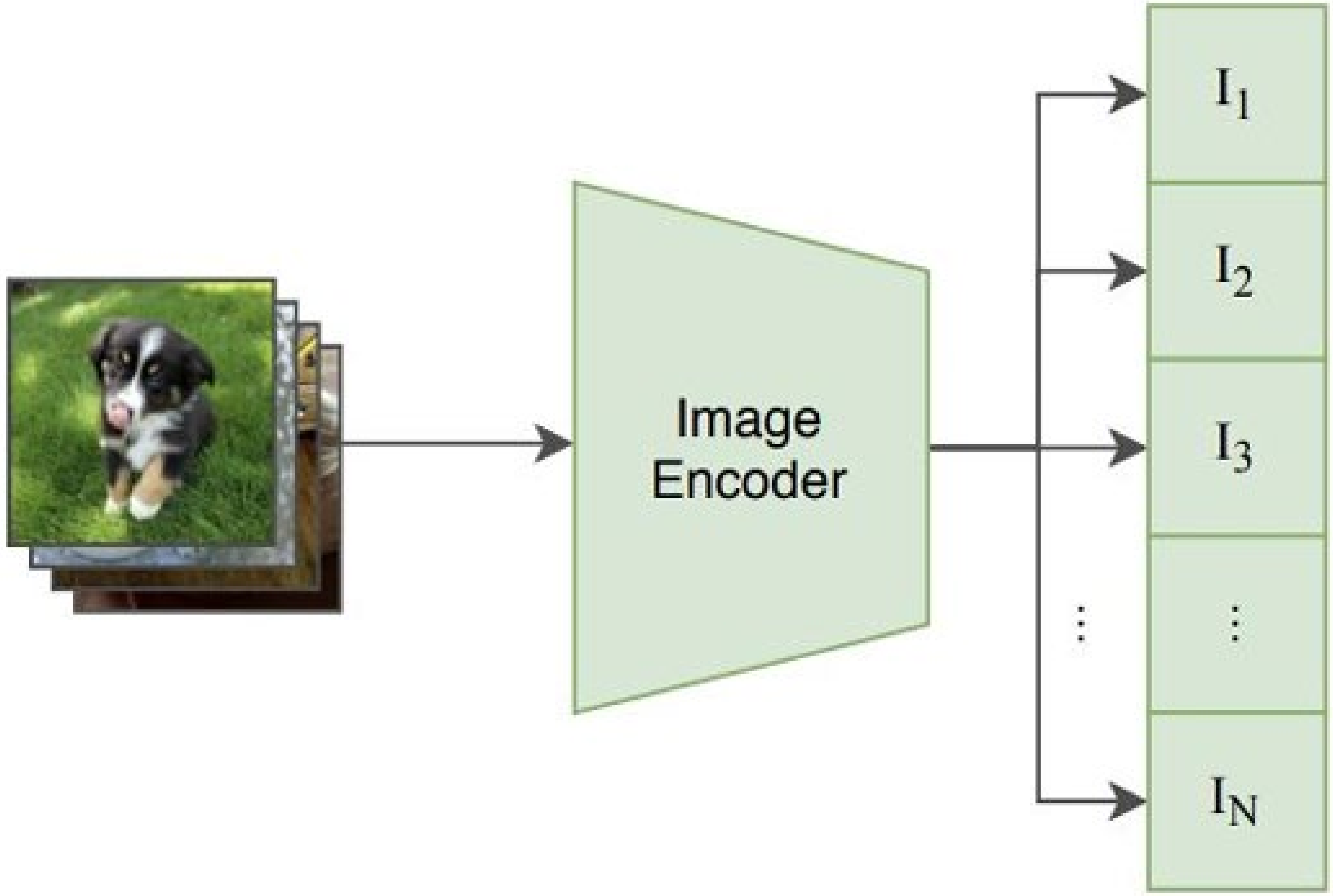
Thirty-eighth International Conference on Machine Learning

Pepper the aussie pup



We will discuss how to obtain text representations next week onwards ("language modeling")





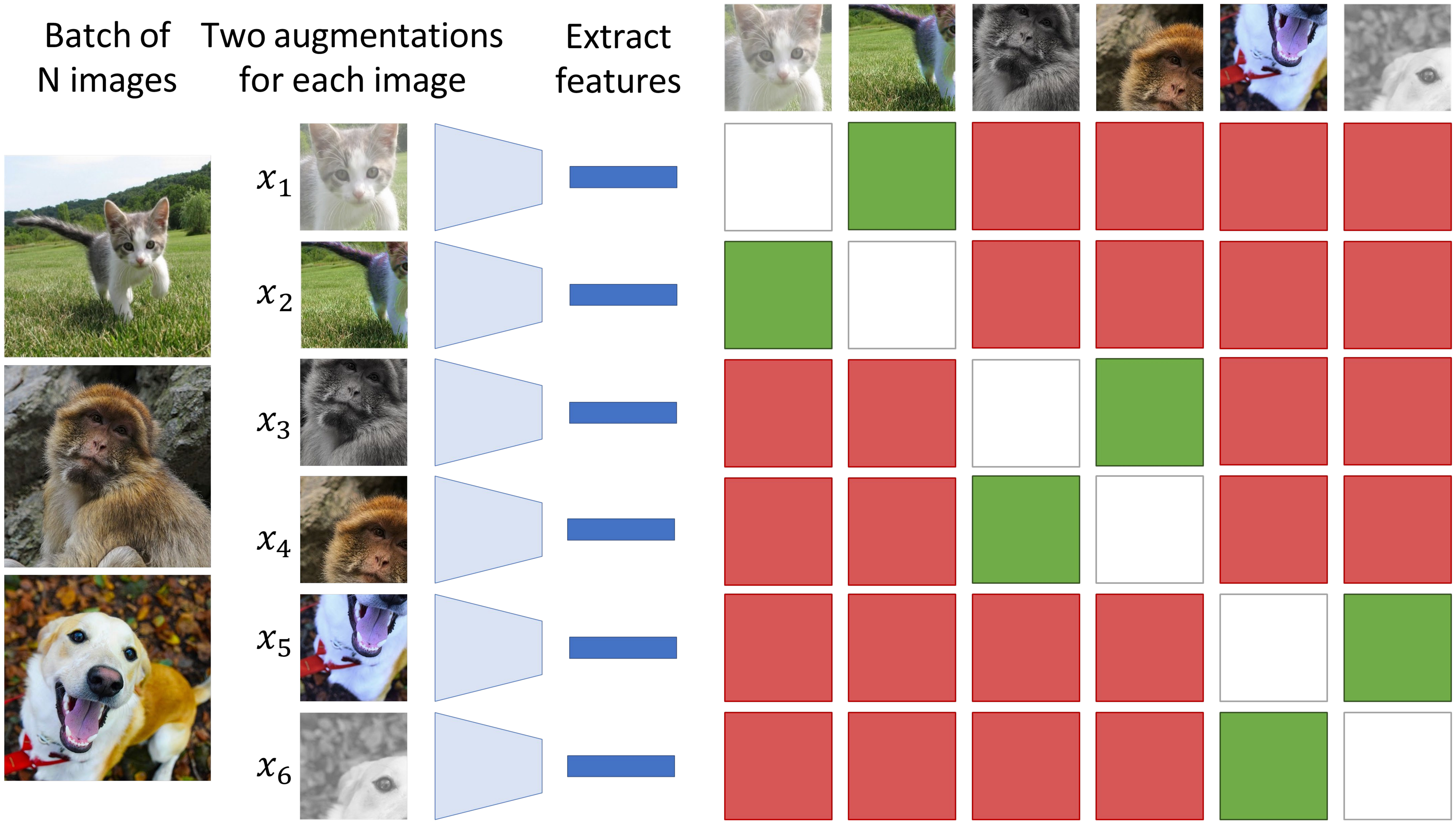
Goal: Do Contrastive Learning!

Maximize Similarity
(Minimize distance)
between the two representations

Why?
Because a common concept is
present in image and sentence

Recall: Contrastive Learning (General Form)

Recall: Contrastive Learning (SimCLR)



Each image tries to predict which of the *other* 2N-1 images came from the same original image

Similarity between x_i and x_j:

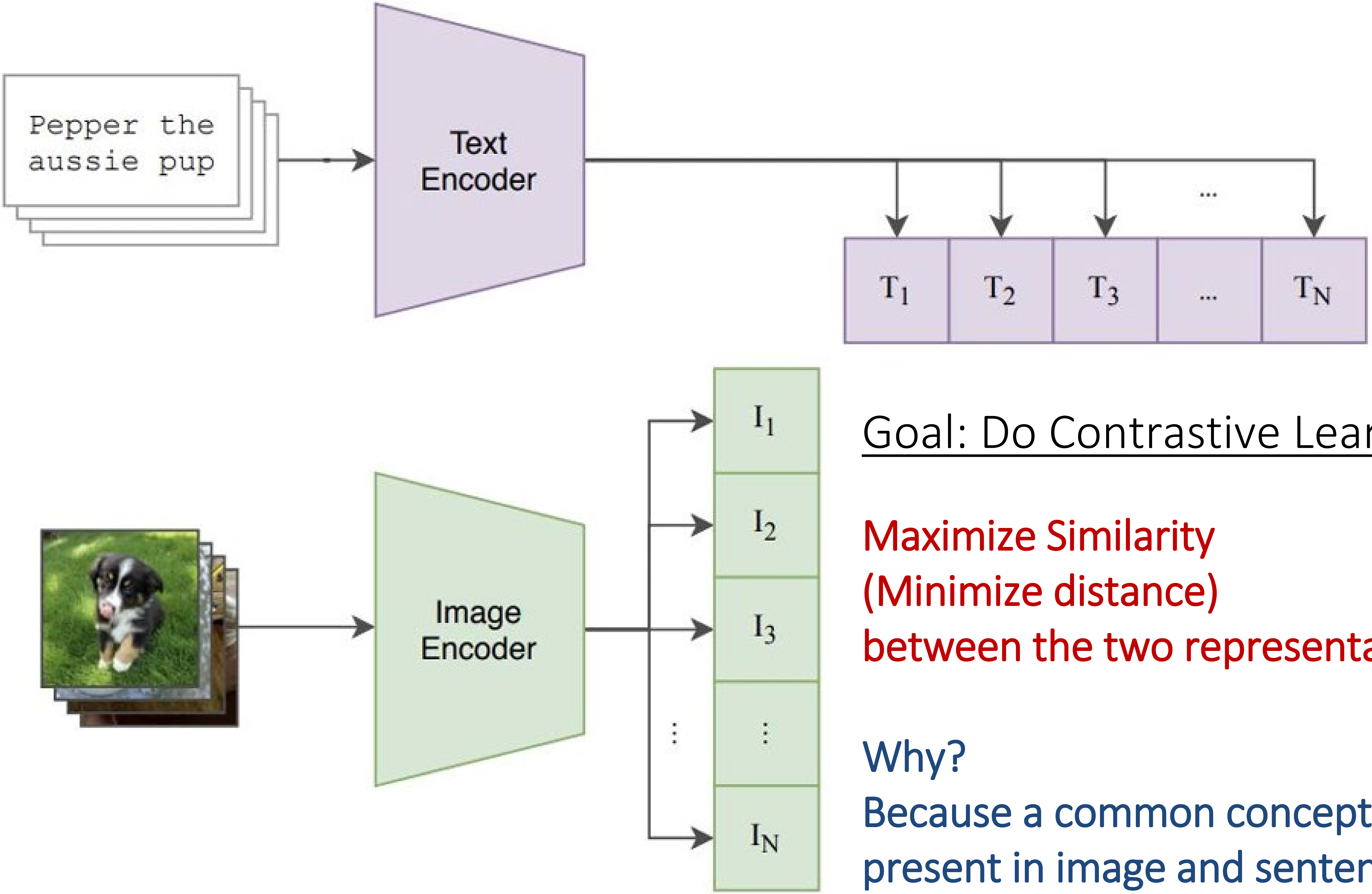
$$s_{i,j} = \frac{\phi(x_i)^T \phi(x_j)}{\|\phi(x_i)\| \cdot \|\phi(x_i)\|}$$

If (x_i, x_j) is a positive pair, then loss for x_i is:

$$L_i = -\log \frac{\exp(s_{i,j}/\tau)}{\sum_{\substack{k=1 \\ k \neq i}}^{2N} \exp(s_{i,k}/\tau)}$$

(τ is a *temperature*)

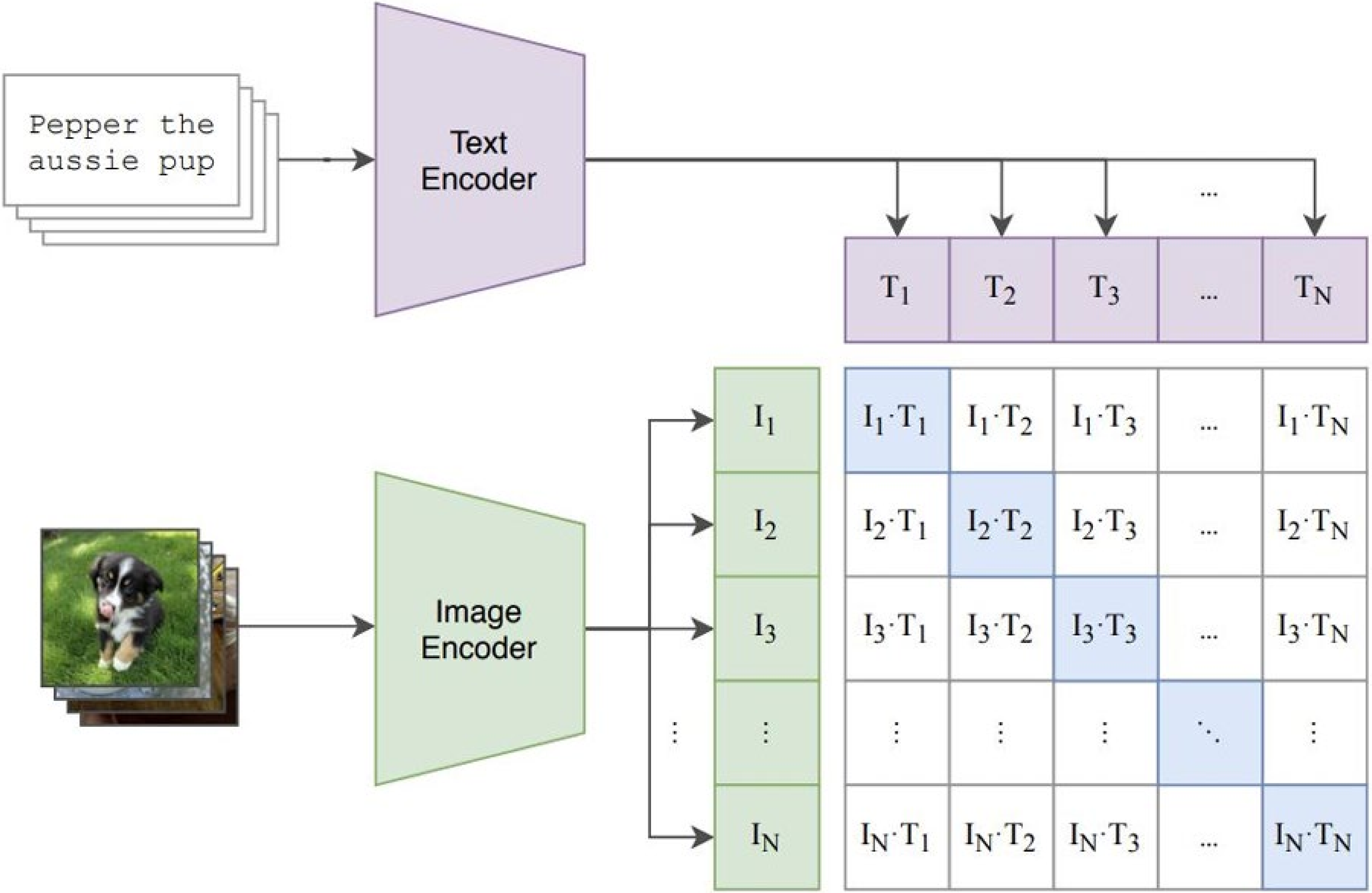
Interpretation: Cross-entropy loss over the other 2N-1 elements in the batch!



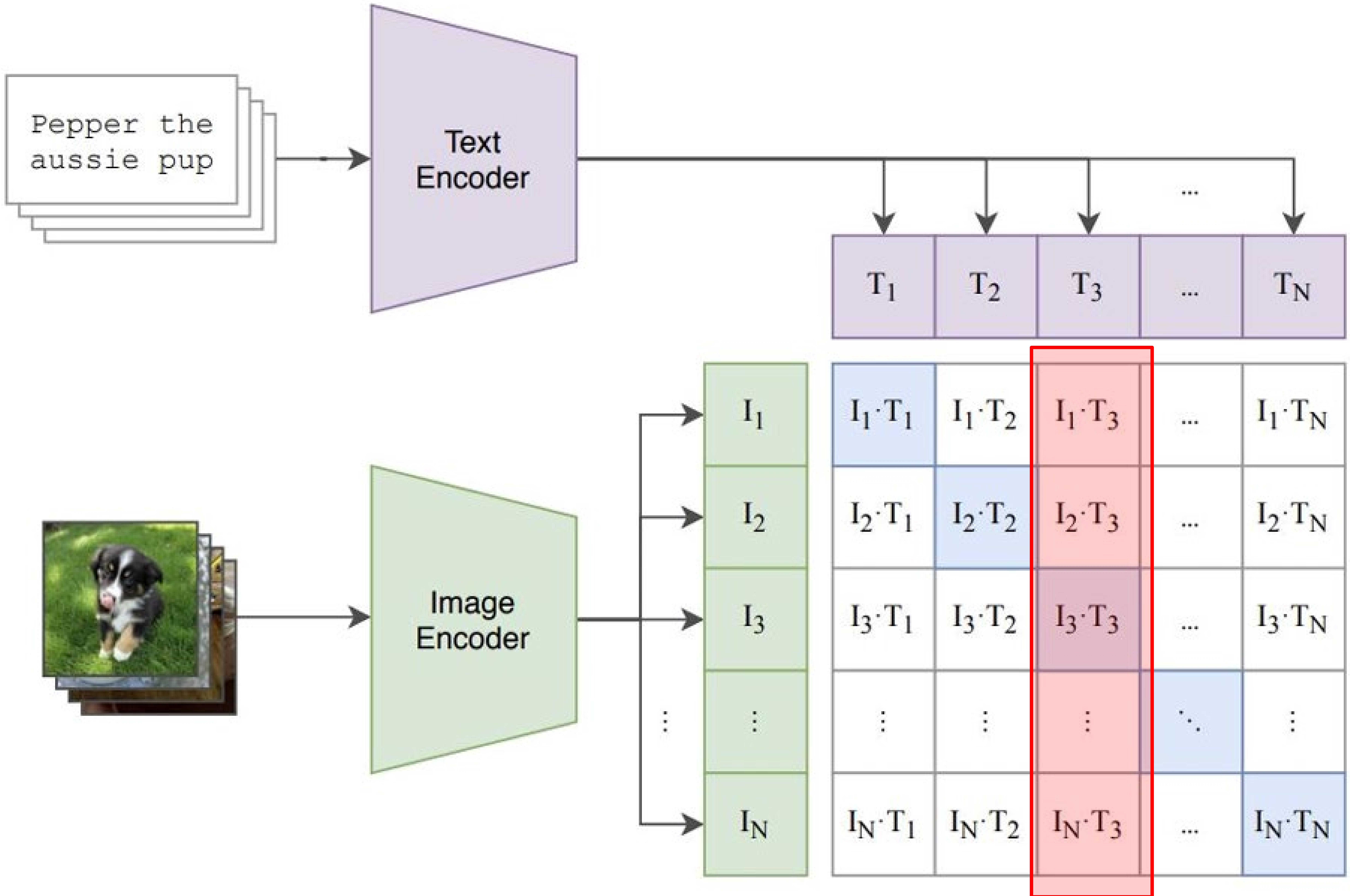
Goal: Do Contrastive Learning!

Maximize Similarity
(Minimize distance)
between the two representations

Why?
Because a common concept is present in image and sentence



Pepper the aussie pup
Text Encoder
Image Encoder
T1 T2 T3 ... TN
I1 I2 I3 ⋮ IN
I1·T1 I1·T2 I1·T3 ... I1·TN
I2·T1 I2·T2 I2·T3 ... I2·TN
I3·T1 I3·T2 I3·T3 ... I3·TN
IN·T1 IN·T2 IN·T3 ... IN·TN



Pepper the aussie pup
Text Encoder
Image Encoder
T_1 T_2 T_3 ... T_N
I_1 I_2 I_3 ⋮ I_N
$I_1 \cdot T_1$ $I_1 \cdot T_2$ $I_1 \cdot T_3$ ... $I_1 \cdot T_N$
$I_2 \cdot T_1$ $I_2 \cdot T_2$ $I_2 \cdot T_3$ ... $I_2 \cdot T_N$
$I_3 \cdot T_1$ $I_3 \cdot T_2$ $I_3 \cdot T_3$ ... $I_3 \cdot T_N$
$I_N \cdot T_1$ $I_N \cdot T_2$ $I_N \cdot T_3$ ... $I_N \cdot T_N$


Pepper the aussie pup
Text Encoder
Image Encoder
T1 T2 T3 ... TN
I1 I2 I3 ⋮ IN
I1·T1 I1·T2 I1·T3 ... I1·TN
I2·T1 I2·T2 I2·T3 ... I2·TN
I3·T1 I3·T2 I3·T3 ... I3·TN
IN·T1 IN·T2 IN·T3 ... IN·TN

plane

car

dog

⋮

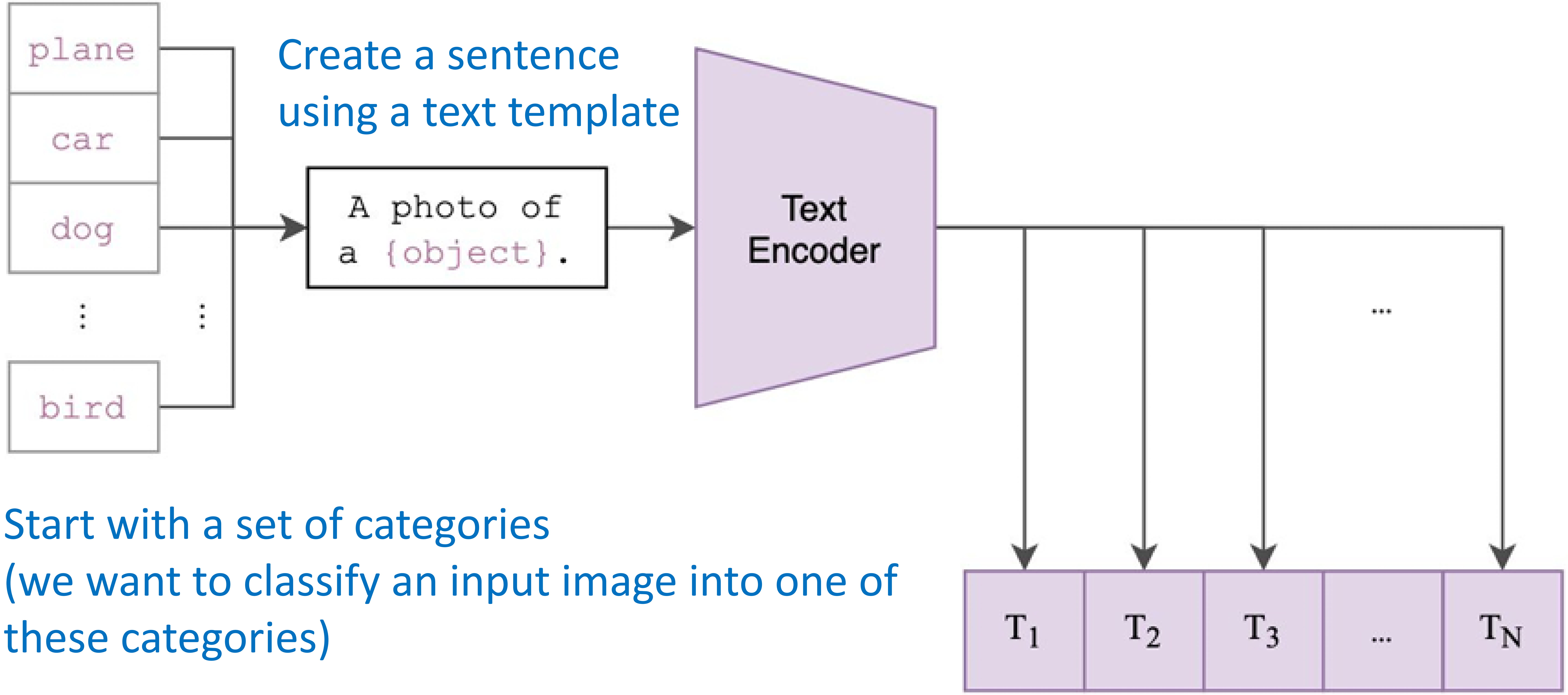
bird

Start with a set of categories
(we want to classify an input image into one of these categories)

Image Classification with CLIP

(so called "Zero-shot" classification)

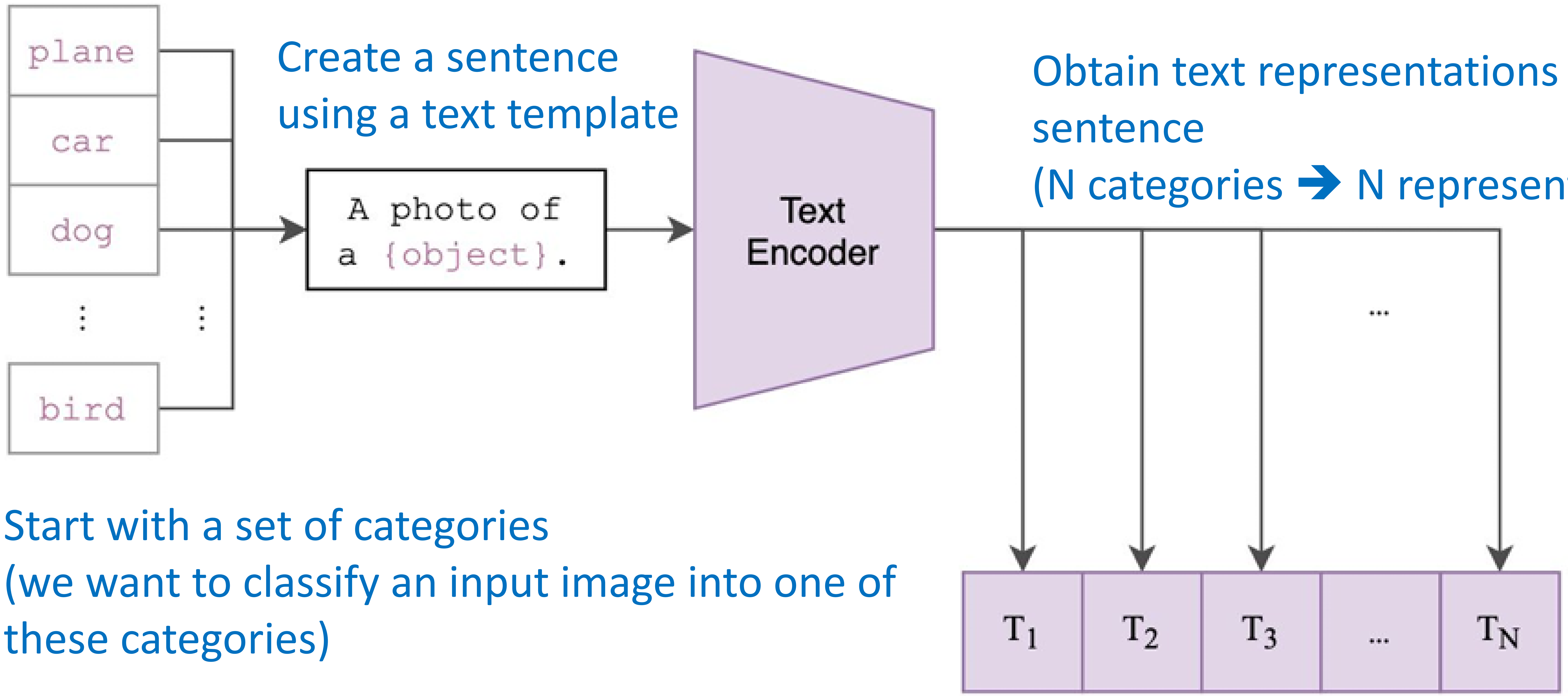
Create a sentence
using a text template



Start with a set of categories
(we want to classify an input image into one of these categories)

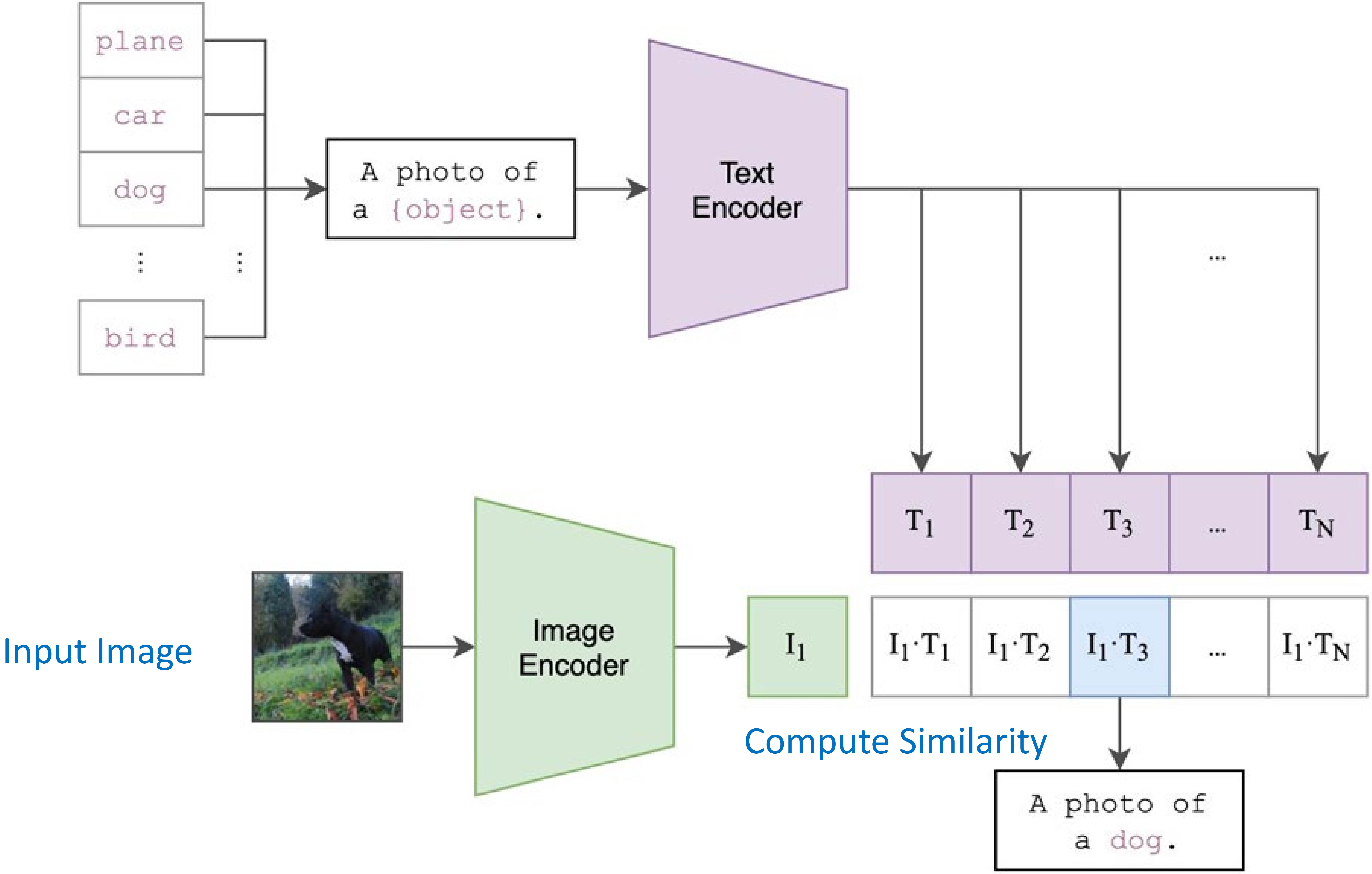
Create a sentence using a text template

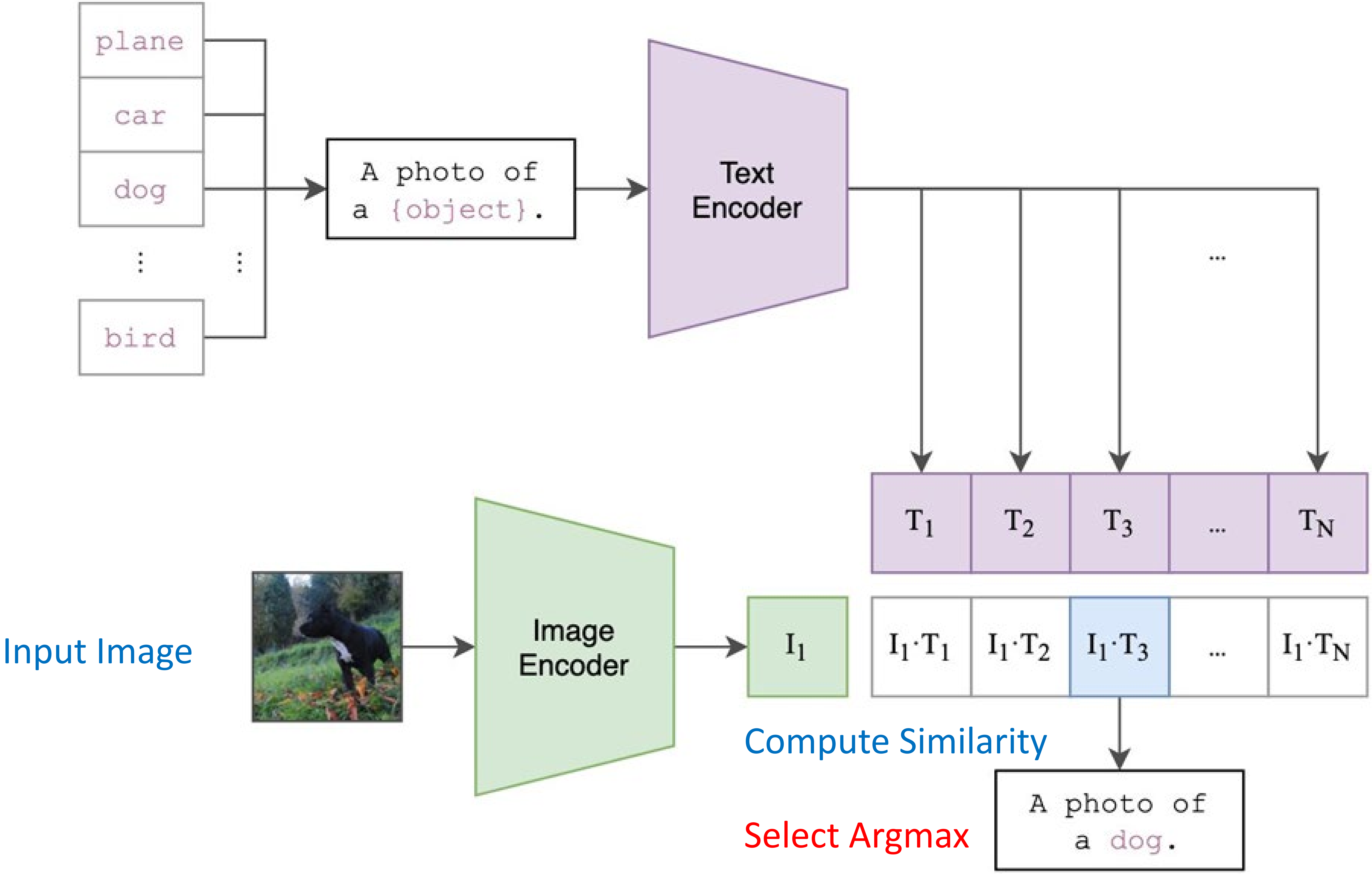
Obtain text representations for each sentence
(N categories ➔ N representations



Start with a set of categories
(we want to classify an input image into one of these categories)

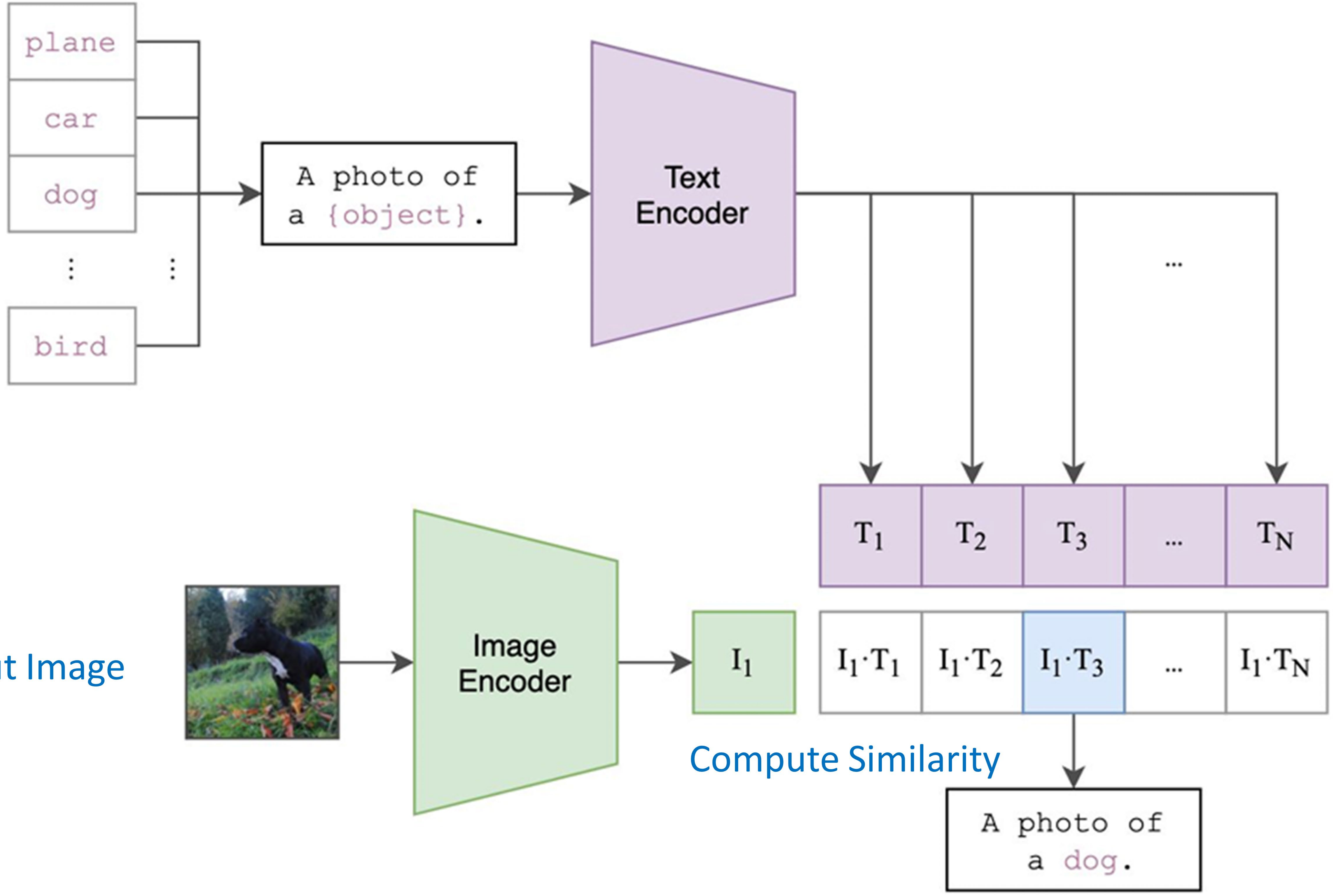

plane
car
dog
bird
Create a sentence
using a text template
A photo of
a {object}.
Text
Encoder
Obtain text representations for each
sentence
(N categories ➔ N representations
T_1
T_2
T_3
...
T_N
Start with a set of categories
(we want to classify an input image into one of
these categories)
Input Image
Image
Encoder
I_1
Image Representation


plane
car
dog
bird
A photo of a {object}.
Text Encoder
T1
T2
T3
...
TN
Input Image
Image Encoder
I1
I1·T1
I1·T2
I1·T3
...
I1·TN
Compute Similarity
A photo of a dog.



plane
car
dog
bird
A photo of a {object}.
Text Encoder
...
T1
T2
T3
...
TN
Input Image
Image Encoder
I1
I1·T1
I1·T2
I1·T3
...
I1·TN
Compute Similarity
Select Argmax
A photo of a dog.

CLIP: Zero-Shot Classification



Input Image

Compute Similarity

Language enables zero shot classification:
Classify images into categories without any additional training data!

Contrastive loss:
For each image, predict which sentence matches it.

Large-scale training on 400M (image, text) pairs from the internet

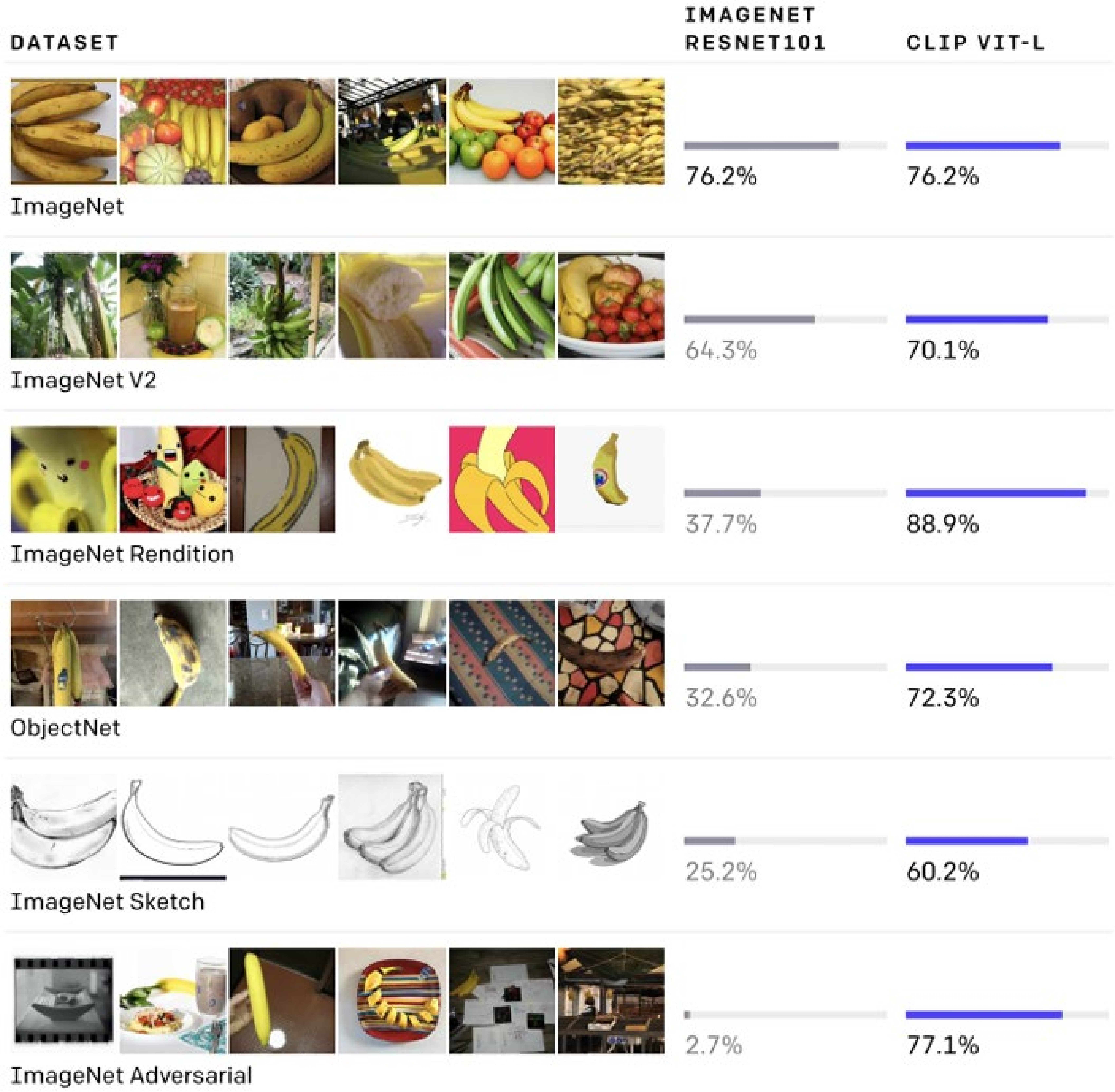
Problem: CLIP training dataset is private; can't reproduce results

CLIP Performance

Very strong performance on many downstream vision problems!

Performance continues to improve with larger models

Radford et al, “Learning Transferable Visual Models form Natural Language Supervision”, ICML 2021



CLIP Details

Training Details:

- Trained on 400M image-text pairs from the internet (i.e. without permissions a.k.a. stealing)

- 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


Caltech-101
kangaroo (99.8%) Ranked 1 out of 102 labels
✓ a photo of a kangaroo.
× a photo of a gerenuk.
× a photo of a emu.
× a photo of a wild cat.
× a photo of a scorpion.
ImageNet-R (Rendition)
Siberian Husky (76.0%) Ranked 1 out of 200 labels
✓ a photo of a siberian husky.
× a photo of a german shepherd dog.
× a photo of a collie.
× a photo of a border collie.
× a photo of a rottweiler.
Oxford-IIIT Pets
Maine Coon (100.0%) Ranked 1 out of 37 labels
Maine Coon
✓ a photo of a maine coon, a type of pet.
× a photo of a persian, a type of pet.
× a photo of a ragdoll, a type of pet.
× a photo of a birman, a type of pet.
× a photo of a siamese, a type of pet.
CIFAR-100
snake (38.0%) Ranked 1 out of 100 labels
✓ a photo of a snake.
× a photo of a sweet pepper.
× a photo of a flatfish.
× a photo of a turtle.
× a photo of a lizard.

Country211

Belize (22.5%) Ranked 5 out of 211 labels

× a photo i took in **french guiana**.

× a photo i took in **gabon**.

× a photo i took in **cambodia**.

× a photo i took in **guyana**.

✓ a photo i took in **belize**.

RESISC45

roundabout (96.4%) Ranked 1 out of 45 labels

✓ satellite imagery of **roundabout**.

× satellite imagery of **intersection**.

× satellite imagery of **church**.

× satellite imagery of **medium residential**.

× satellite imagery of **chaparral**.

Stanford Cars

2012 Honda Accord Coupe (63.3%) Ranked 1 out of 196 labels

✓ a photo of a **2012 honda accord coupe**.

× a photo of a **2012 honda accord sedan**.

× a photo of a **2012 acura tl sedan**.

× a photo of a **2012 acura tsx sedan**.

× a photo of a **2008 acura tl type-s**.

SUN

kennel indoor (98.6%) Ranked 1 out of 723 labels

✓ a photo of a **kennel indoor**.

× a photo of a **kennel outdoor**.

× a photo of a **jail cell**.

× a photo of a **jail indoor**.

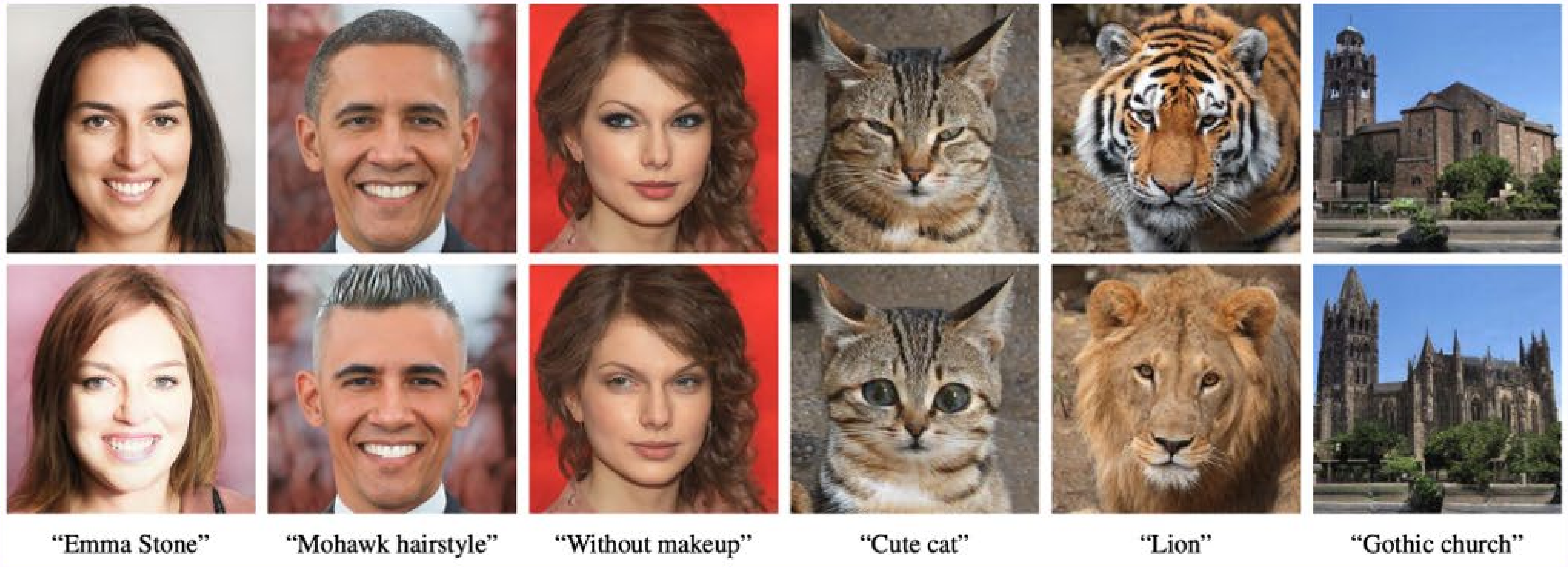
× a photo of a **veterinarians office**.

Can be “Attacked”
(we will discuss adversarial attacks later in the semester)

Applications of CLIP (slide from Radford et al.)

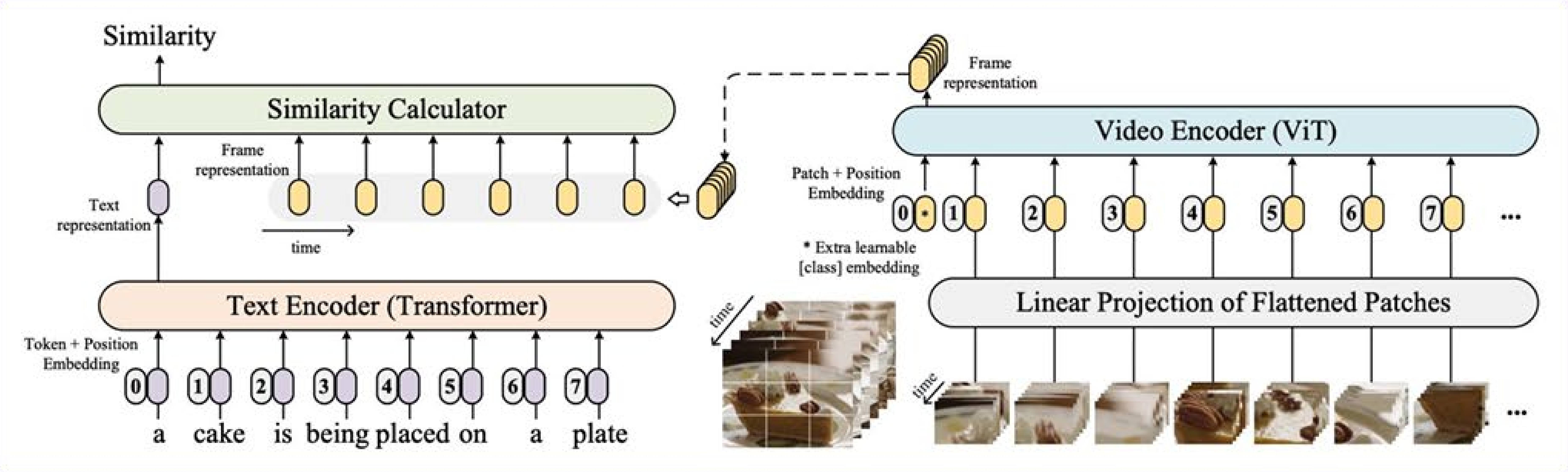
StyleCLIP
(Patashnik et al.)

Steering a GAN Using CLIP



CLIP4Clip
(Luo & Ji, et al.)

Video retrieval using
CLIP features



Summary

- Self-Supervised Learning: scale up training without human annotation
  - First train for a pretext task, then transfer to downstream tasks
  - Many pretext tasks: context prediction, jigsaw, colorization, clustering, rotation
  - SSL has been wildly successful for language
- Intense research on SSL in vision
- Multimodal SSL with vision + language has been very successful

Reminder: Midterm is on Monday (03/31)

- In class (ITE 231).
- 4:00 PM – 5:00 PM

- NOT ALLOWED:
  (if you're bringing these, put them in your bag when you take the exam)

  ☒ notes / books / any written material
  ☒ laptops / ipads / phones / any form of computer ...

- Calculators are allowed, but not required.

Syllabus: everything including today's lecture

- Study slides carefully
- Make sure you grasp concepts

Format: Mix of

- Multiple Choice
- Short Answer
- "Design"
- Fundamentals of training

Midterm Review: Major Topics

Machine Learning

Supervised Learning Basics
Classification vs Regression
Linear Classifier/Regression
Training Objective
Train vs Test / Overfitting

Midterm Review: Major Topics

Machine Learning

Supervised Learning Basics
Classification vs Regression
Linear Classifier/Regression
Training Objective
Train vs Test / Overfitting

Neural Network

Multi-Layer Perceptron
Need for non-linearity
Loss Functions
Gradient Descent
Backpropagation

Midterm Review: Major Topics

Machine Learning

Supervised Learning Basics
Classification vs Regression
Linear Classifier/Regression
Training Objective
Train vs Test / Overfitting

Neural Network

Multi-Layer Perceptron
Need for non-linearity
Loss Functions
Gradient Descent
Backpropagation

CNN Design+Training

Convolution / Filtering
Output Size Equation
(padding, pooling, stride, kernel size, ...)
Activation Functions
Normalization
Optimizers & Scheduling
Regularization, Dropout
Data Augmentation
Hyperparameters
Visualizing learned features
Generalization ...
Domain Adaptation

Midterm Review: Major Topics

Machine Learning

Supervised Learning Basics
Classification vs Regression
Linear Classifier/Regression
Training Objective
Train vs Test / Overfitting

Neural Network

Multi-Layer Perceptron
Need for non-linearity
Loss Functions
Gradient Descent
Backpropagation

CNN Design+Training

Convolution / Filtering
Output Size Equation (padding, pooling, stride, kernel size, ...)
Activation Functions
Normalization
Optimizers & Scheduling
Regularization, Dropout
Data Augmentation
Hyperparameters
Visualizing learned features
Generalization ...
Domain Adaptation

Representation Learning

Discriminative vs Generative
Autoencoder Basics
Variational Autoencoder
Generative Adversarial Nets
Self-Supervised Learning
Pretext Tasks
Metric Learning (what is similarity?)
Contrastive Learning
CL with Data Augmentation (SimCLR)
CLIP
...

Midterm Review: Major Topics

Machine Learning

Supervised Learning Basics
Classification vs Regression
Linear Classifier/Regression
Training Objective
Train vs Test / Overfitting

Neural Network

Multi-Layer Perceptron
Need for non-linearity
Loss Functions
Gradient Descent
Backpropagation

CNN Design+Training

Convolution / Filtering
Output Size Equation
(padding, pooling, stride, kernel size, ...)
Activation Functions
Normalization
Optimizers & Scheduling
Regularization, Dropout
Data Augmentation
Hyperparameters
Visualizing learned features
Generalization ...
Domain Adaptation

Representation Learning

Discriminative vs Generative
Generative Modeling
Autoencoder Basics
Variational Autoencoder
Generative Adversarial Nets
Self-Supervised Learning
Pretext Tasks
Metric Learning (similarity?)
Contrastive Learning
CL with Data Augmentation (SimCLR)
CLIP
...