Multimodalité

Apprentissage continu de représentations visuelles

2023-2024

Dans cette dernière partie du cours: extension vers le texte

Question : Si des meta-données associées aux images sont disponibles, pourquoi ne pas les utiliser pour guider l'apprentissage de modèles et de représentations visuelles ?

Les premières approches à utiliser cette idée ont considéré des **méta-données** classiques, telles que les tags, et les ont utilisées comme un **ensemble discret d'étiquettes** (*labels*).

Cela permet d'utiliser les techniques classiques d'apprentissage de modèles de classification

Dans la suite, deux exemples:

- Utilisation des mots des titres, légendes, et tags d'images Flickr [Joulin@ECCV16]
- Utilisation des hashtags Instagram [Mahajan@ECCV18]

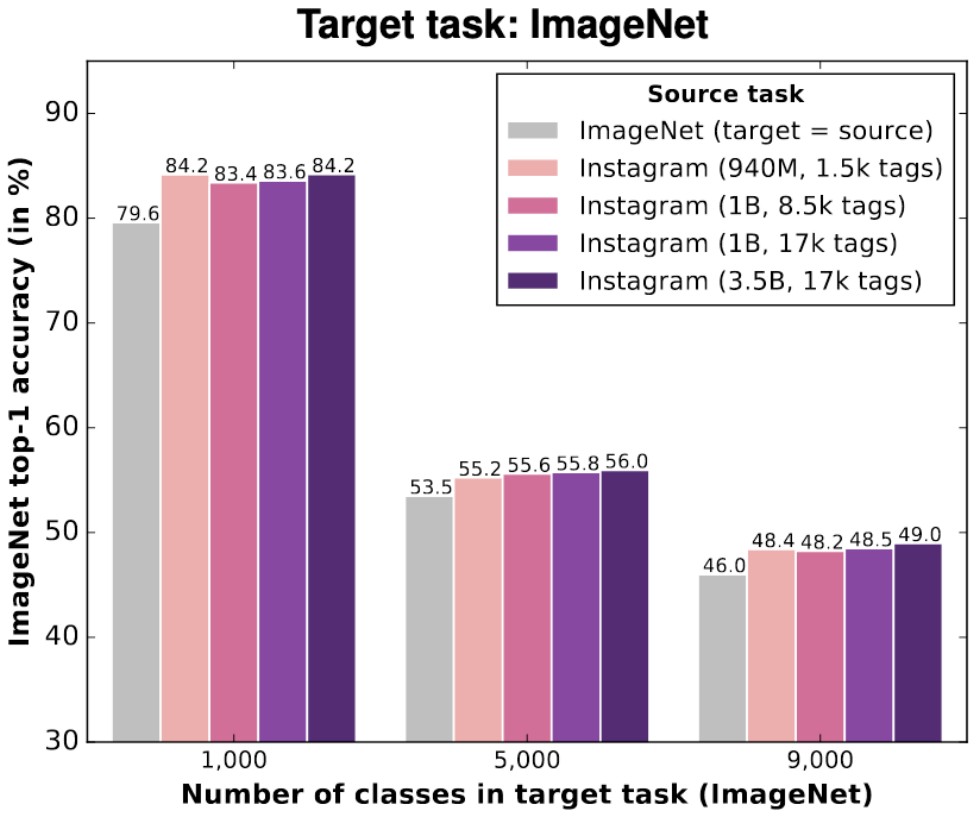
Exemples de méthodes utilisant les méta-données et les tags

Learning Visual Features from Large Weakly Supervised Data
Armand Joulin, Laurens van der Maaten, Allan Jabri, Nicolas Vasilache
ECCV16

Training dataset: 100 million images publicly available on Flickr.
Contains photos with associated titles, hashtags, and captions.
The latter are transformed into a word prediction (multi-label classification) pretext task, using the image as input.

Exploring the Limits of Weakly Supervised Pretraining
Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, Laurens van der Maaten.
ECCV18

Training dataset: 3.5 billion public Instagram images.
The pretext task is to predict the hashtags of those images.



Exploring the Limits of Weakly Supervised Pretraining
Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, Laurens van der Maaten.
ECCV18

Dans cette dernière partie cours: extension vers le texte

Question : Si des meta-données associées aux images sont disponibles, pourquoi ne pas les utiliser pour guider l'apprentissage de modèles et de représentations visuelles ?

Le deuxième type d'approches utilise du **texte non contraint** (= en language naturel) pour aider à la tâche d'apprentissage de représentation visuelles.

Dans la suite, nous verrons plusieurs exemples, dont le plus récent est **CLIP**.

Learning transferable visual representations

Weak annotations

Reducing annotation cost

Fully-Supervised
fine-grained annotations
expert knowledge

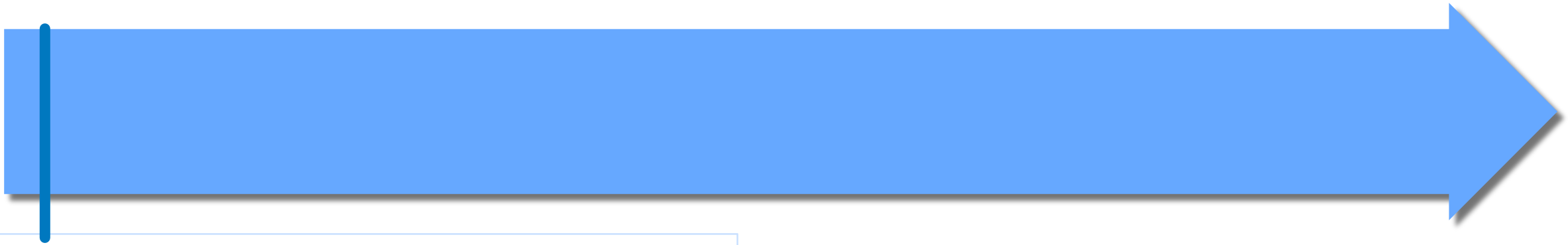
Caption-supervised
side information
smaller sets

a statue of a man stands in front of an old red bus.
a big and red bus with many displays for people to watch.
a red double decker bus parked next to a statue.
the double decker bus is beside a statue near restaurant tables.
a view of a bus sitting in front a small wooden statue.

a busy street with cars and trucks down it
an intersection with a view that looks towards a small downtown area.
cars parked on the side of the street and traveling down the road
an intersection with a stop light on a city street.
a street filled with lots of traffic under a traffic light.

Self-supervised
annotation-free images
computationally demanding

- Une des premières méthodes à entrainer une représentation visuelle à partir d'images décrites par des légendes (*caption*)
- Elle définie un certain nombre de tâches prétextes en utilisant des paires (image,caption) en entrée
- Elle entraine sur ces différentes tâches prétextes



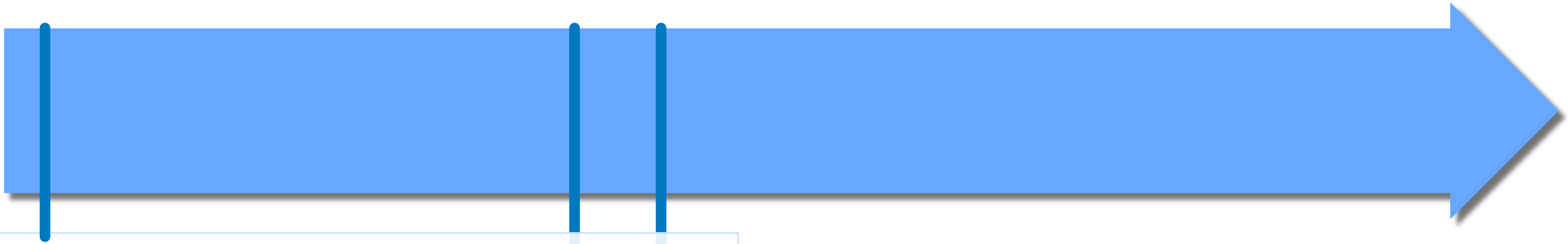
Quattoni, A., Collins, M., Darrell, T.
Learning visual representations using images with captions.
CVPR 2007

La première méthode [Li@ICCV17]

- Construit des ensemble de catégories (*label sets*) par construction de n-grammes
- Entraine un réseau de neurones convolutionnel à prédire ces labels

La deuxième méthode [Gomez@CVPR17]

- Extrait des pages wikipedia illustrées
- Entraine un modèle de variables latentes sur le texte de ces pages (*topic model*)
- Entraine un réseau de neurones convolutionnel à representer les images correspondantes dans l'espace de ces variables latentes



Quattoni, A., Collins, M., Darrell, T.
Learning visual representations using images with captions.
CVPR 2007

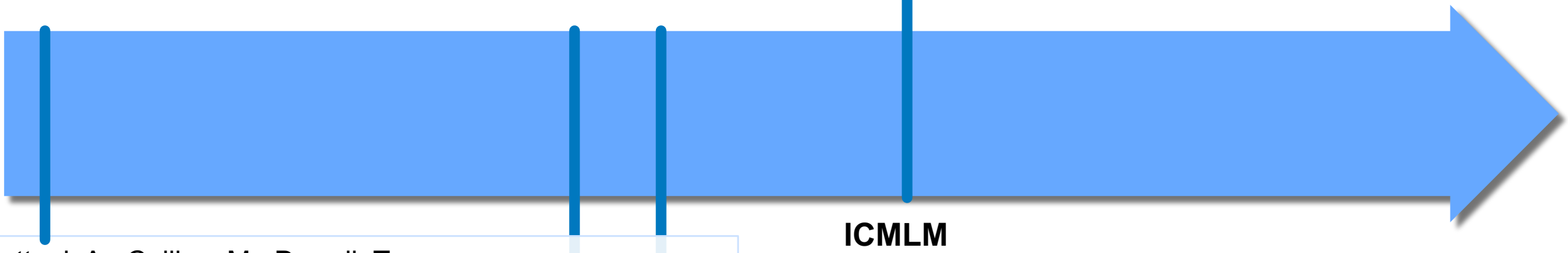
Li, A., Jabri, A., Joulin, A., van der Maaten, L.
Learning visual n-grams from web data. **ICCV 2017**

Gomez, L., Patel, Y., ,Karatzas, D., Jawahar, C.
Self-supervised learning of visual features through embedding images into text topic spaces. **CVPR 2017**

Sariyildiz, M. B., Perez, J., & Larlus, D
Learning Visual Representations With Caption Annotations
ECCV 2020

- S'inspire des avancées en apprentissage auto-supervisé pour le traitement automatique des langues (TAL, ou *NLP* en anglais) et notamment BERT

Décrit dans les prochains transparents



ICMLM

Quattoni, A., Collins, M., Darrell, T.
Learning visual representations using images with captions.
CVPR 2007

Li, A., Jabri, A., Joulin, A., van der Maaten, L.
Learning visual n-grams from web data. **ICCV 2017**

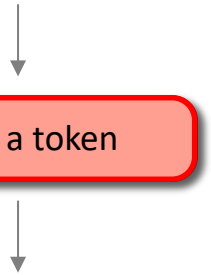
Gomez, L., Patel, Y., ,Karatzas, D., Jawahar, C.
Self-supervised learning of visual features through embedding images into text topic spaces. **CVPR 2017**

Learning transferable textual representations

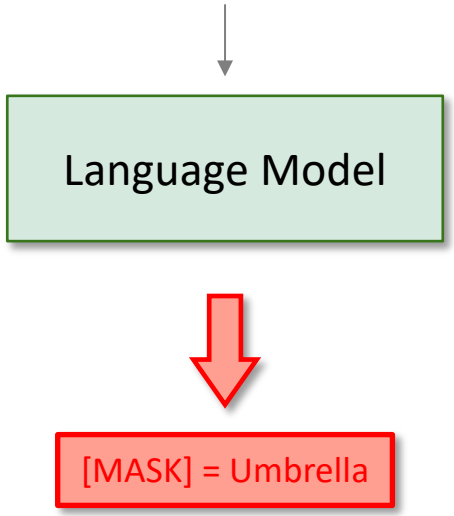
Input: Text

BERT model
[Delvin *et al.* 2018]

"Little girl holding red umbrella"



"Little girl holding red [MASK]"



Input: Image Caption

"Little girl holding red umbrella"

Mask a token

"Little girl holding red [MASK]"

Visual representation
(learnt from scratch)

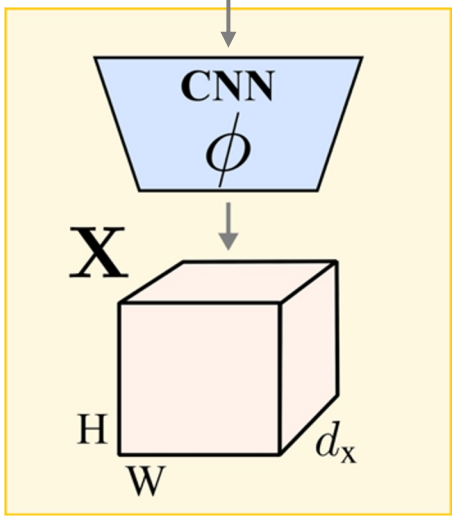
Textual representation
(frozen)

Multi-modal network =
Auxiliary modules

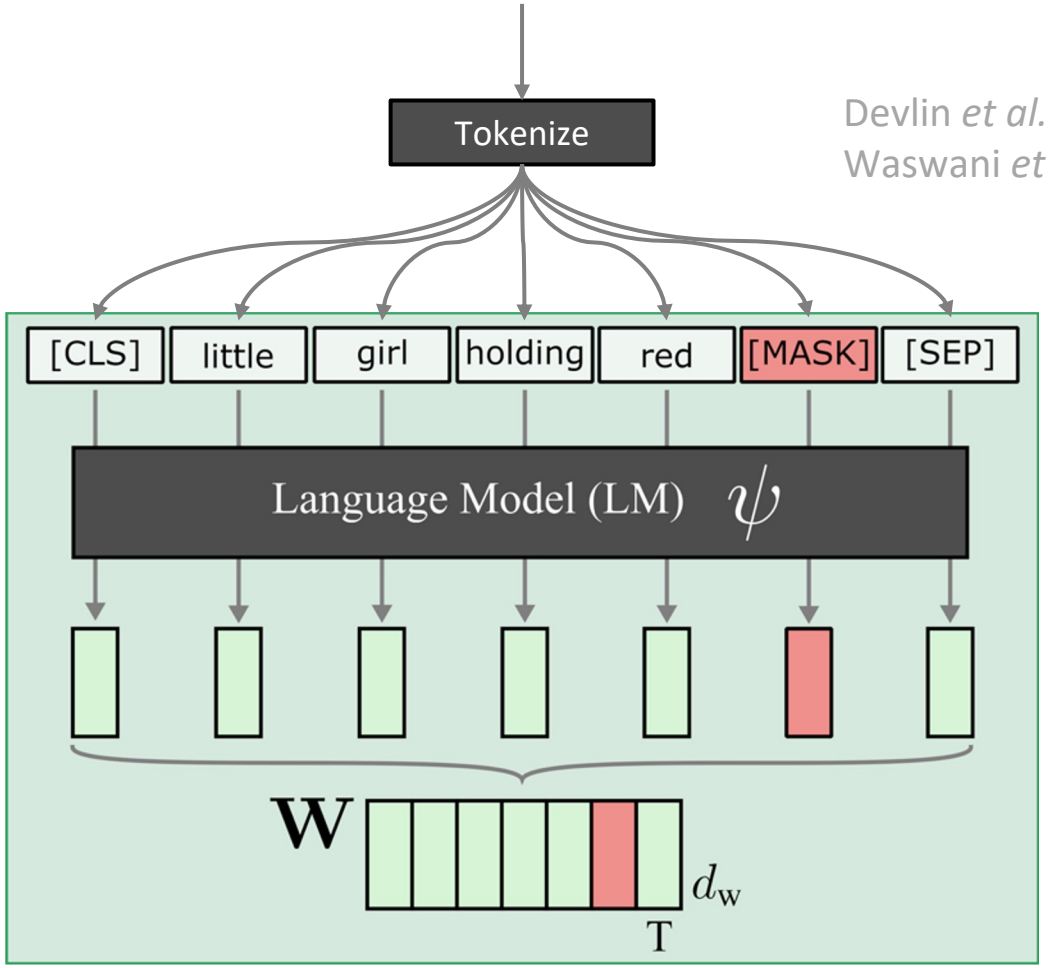
[MASK] = Umbrella

Visual and textual representations

Input: Image Caption



"Little girl holding red [MASK]"

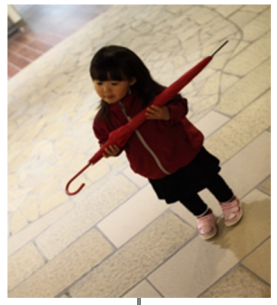


Devlin *et al.* 2019,
Waswani *et al.* 2017

Learning visual representations with caption annotations. Sarıyıldız, Perez, Larlus. ECCV 2020

ICMLM

Input: Image Caption

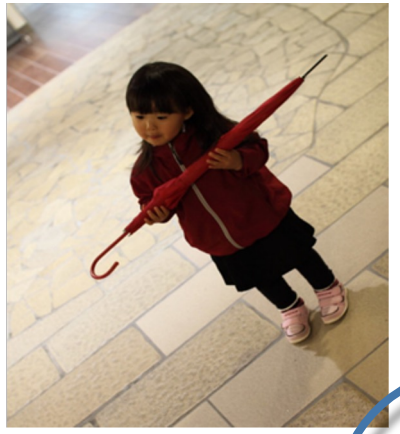


"Little girl holding red [MASK]"

Learning visual representations with caption annotations. Sarıyıldız, Perez, Larlus. ECCV 2020

Pretext task: Image-Conditioned Masked Language Modeling Task (ICMLM)

Input: Image Caption



"Little girl holding red umbrella"

↓

Mask a token

↓

"Little girl holding red [MASK]"

Visual representation
(learnt from scratch)

La partie qui nous intéresse et qu'on réutilise pour la tâche cible.

Multi-modal network =
Auxiliary modules

Une partie auxiliaire qui est seulement nécessaire à la tâche pretexte et qui est "défaussée" plus tard.

Textual representation
(frozen)

On suppose qu'elle est fournie. Par exemple on utilise BERT.

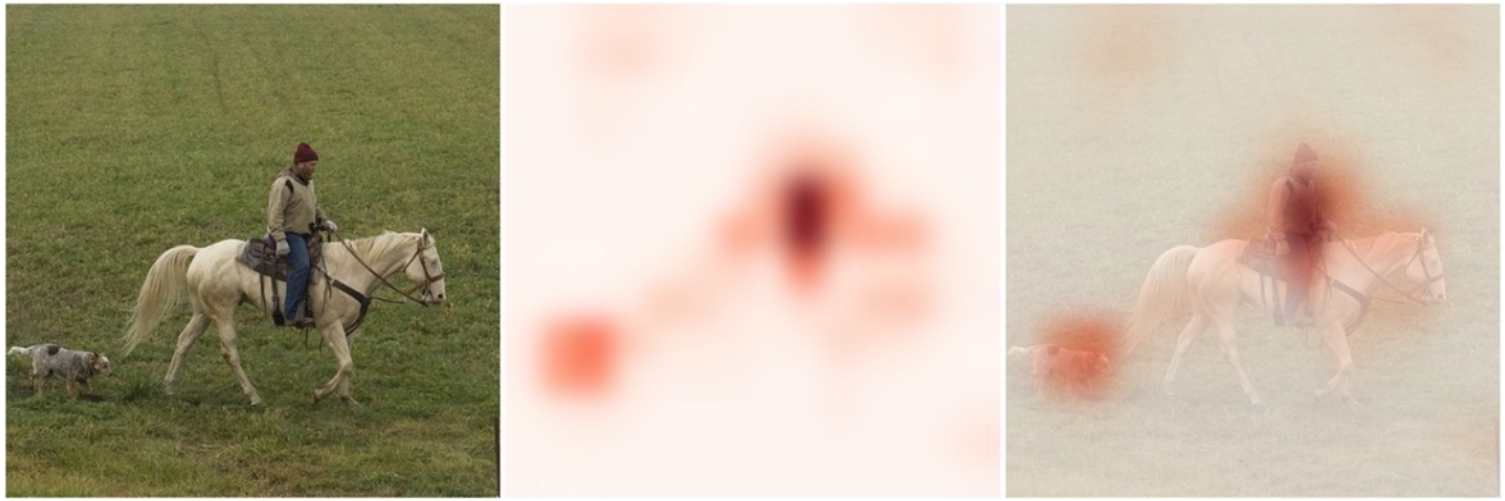
[MASK] = Umbrella

Learning Visual Representations with Caption Annotations
B. Sariyildiz, J. Perez, D. Larlus
European Conference on Computer Vision (ECCV) 2020

Learning transferable visual representations

A woman and a small child watch a [MASK] as it passes

Pred: train, trolley, bus, tram, subway



A [MASK] rides a horse and his dog follows

Pred: man, person, farmer, cowboy, boy

Learning transferable visual representations

A woman and a small [MASK] watch a train as it passes

Pred: child, boy, girl, kid, baby

A man rides a horse and his [MASK] follows

Pred: dog, cow, calf, pony, sheep

Target task evaluations: For **VGG16** backbones (Simonyan *et al.* 2015)

Procedure: (Following Caron *et al.* 2019)

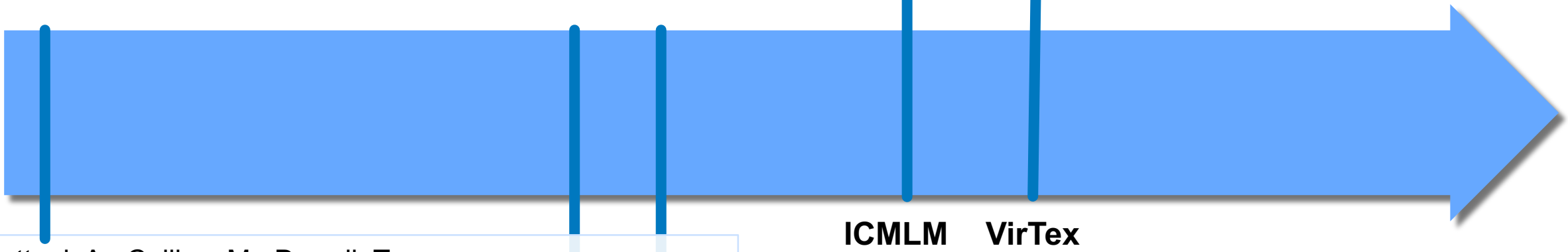
1. Freeze pretrained CNN backbones
2. Probe linear logistic regression classifiers after the last convolutional layer
3. Train them with data augmentation and SGD

	Method	Dataset	Supervision	# Images	VOC	IN1K	Places
Fully-Supervised	Supervised Classifier	ImageNet	1K Labels	1.28M	84.7	71.8	47.3
	Supervised Classifier	COCO	80 Labels	118K	79.9	49.9	44.5
Self-Supervised	DeeperCluster (Caron *et al.* 2019)	YFCC	-	96M	73.1	45.1	41.0
	RotNet (Gidaris *et al.* 2018)	COCO	-	118K	58.6	33.3	34.7
	RotNet (Gidaris *et al.* 2018)	VG	-	103K	59.2	34.7	34.9
Weakly-Supervised	ICMLM (Sariyildiz *et al.* 2020)	COCO	sentences	118K	82.5	49.4	44.6
	ICMLM (Sariyildiz *et al.* 2020)	VG	sentences	103K	**85.0**	47.8	**47.7**

Sariyildiz, M. B., Perez, J., & Larlus, D
Learning Visual Representations With Caption Annotations
ECCV 2020

Desai, K., & Johnson, J.
VirTex: Learning Visual Representations from Textual Annotations
CVPR 2021

VirTex:
- Même tâche prétexte que ICMLM, même bases d'apprentissage
- Apprentissage de la représentation textuelle en même temps que la représentation visuelle



ICMLM **VirTex**

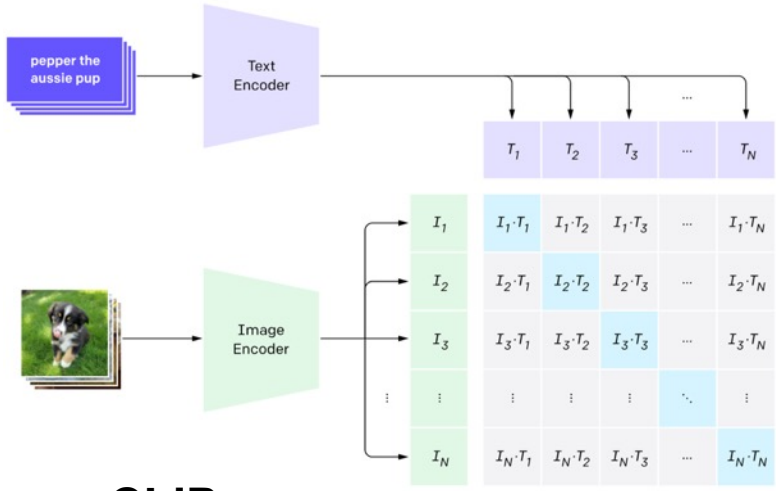
Quattoni, A., Collins, M., Darrell, T.
Learning visual representations using images with captions.
CVPR 2007

Li, A., Jabri, A., Joulin, A., van der Maaten, L.
Learning visual n-grams from web data. **ICCV 2017**

Gomez, L., Patel, Y., ,Karatzas, D., Jawahar, C.
Self-supervised learning of visual features through embedding images into text topic spaces. **CVPR 2017**

Sariyildiz, M. B., Perez, J., & Larlus, D
Learning Visual Representations With Caption Annotations
ECCV 2020

Desai, K., & Johnson, J.
VirTex: Learning Visual Representations from Textual Annotations
CVPR 2021



CLIP

ICMLM **VirTex**

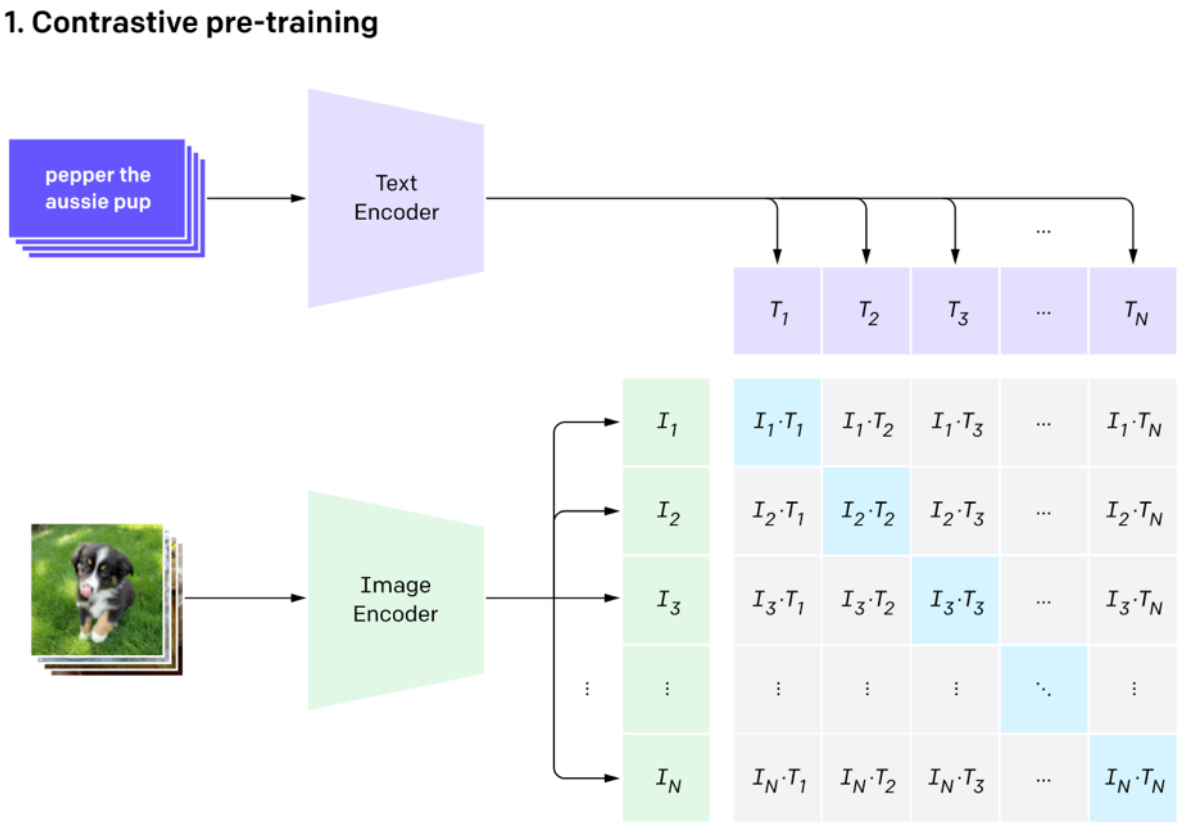
Quattoni, A., Collins, M., Darrell, T.
Learning visual representations using images with captions.
CVPR 2007

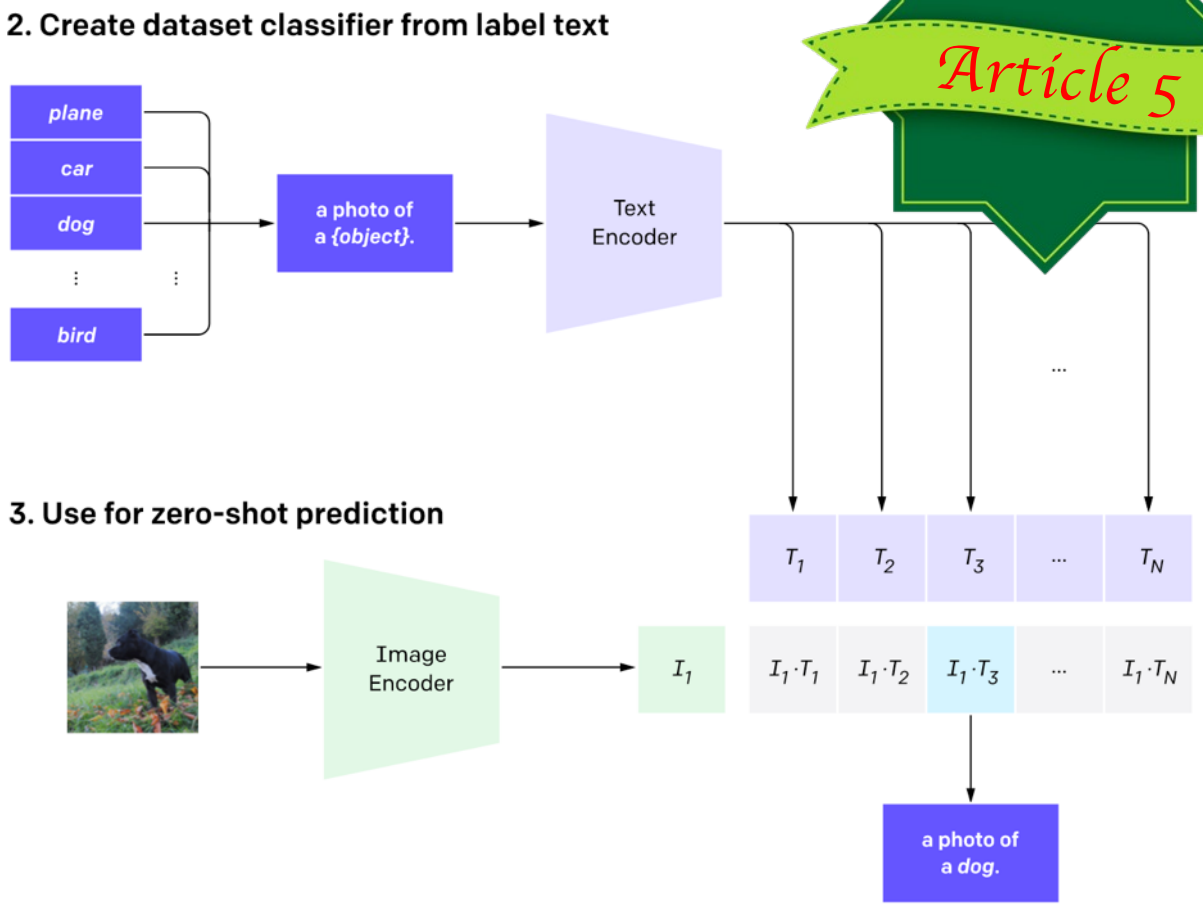
Li, A., Jabri, A., Joulin, A., van der Maaten, L.
Learning visual n-grams from web data. **ICCV 2017**

Gomez, L., Patel, Y., ,Karatzas, D., Jawahar, C.
Self-supervised learning of visual features through embedding images into text topic spaces. **CVPR 2017**

Radford, A., et al
Learning transferable visual models from natural language supervision
PMLR 2021

CLIP (Contrastive Language–Image Pre-training)





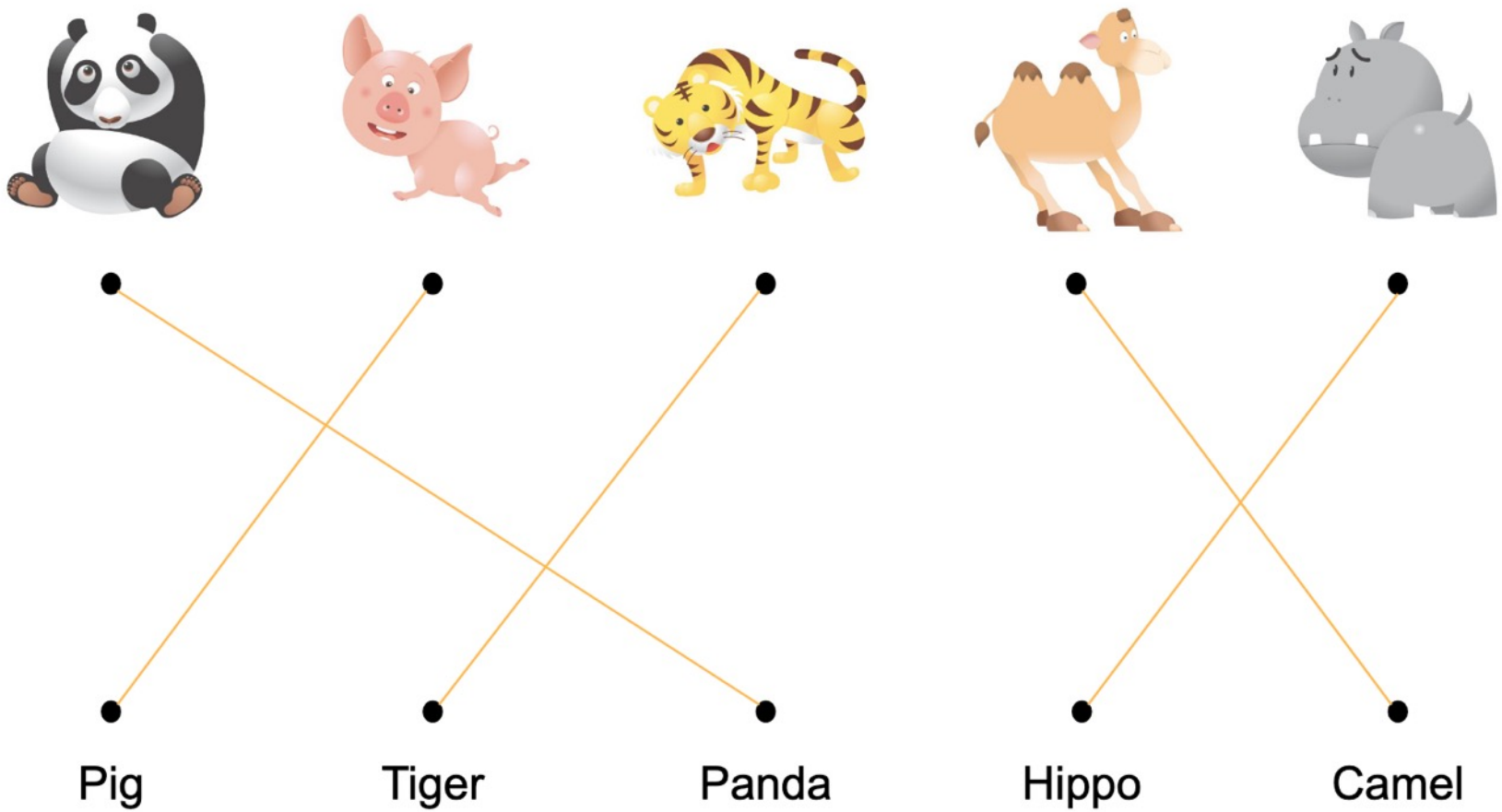
Article 5

Learning Transferable Visual Models From Natural Language Supervision

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

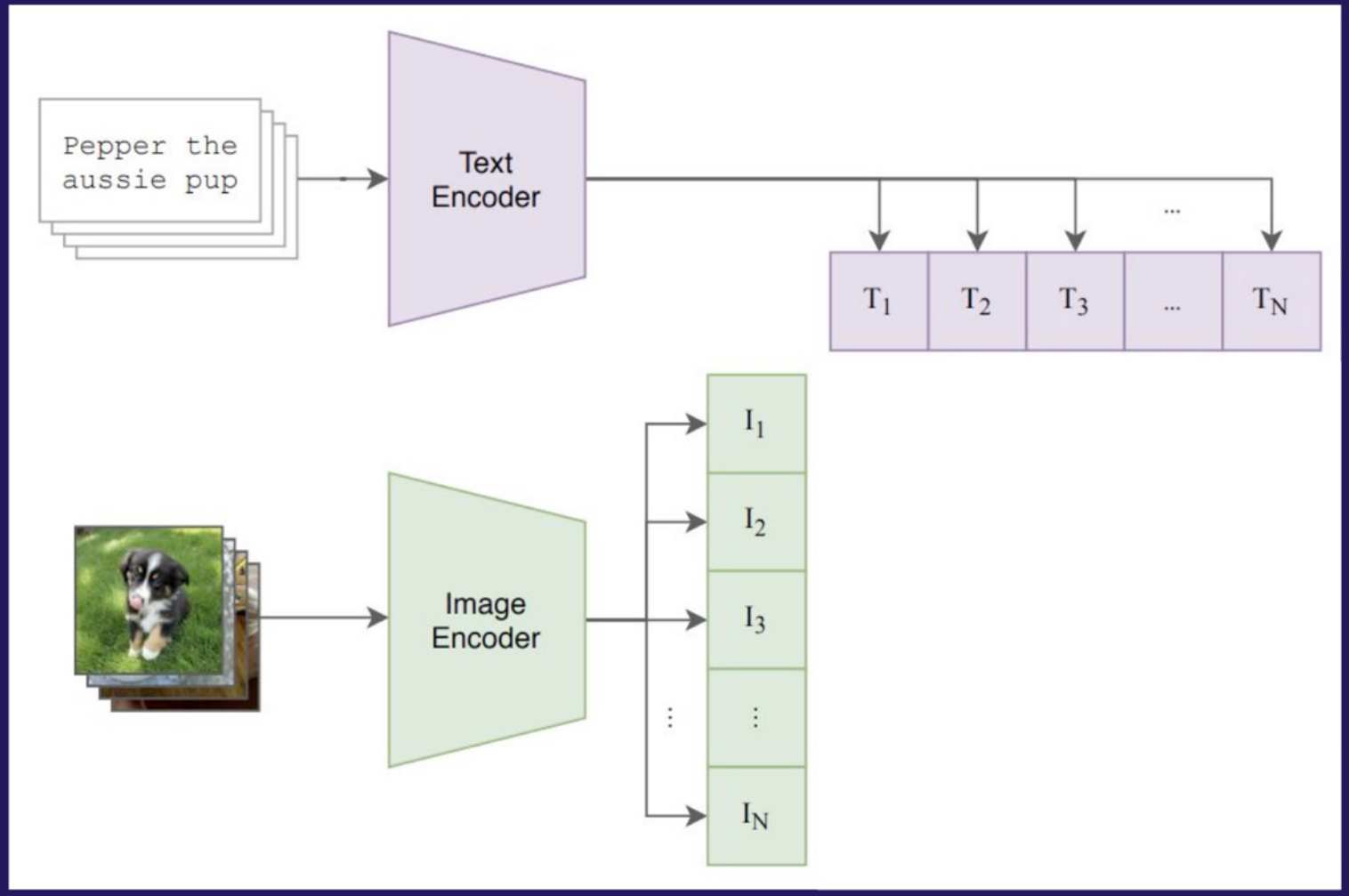
PMLR 2021

Contrastive learning



[Slide: Jong Wook Kim – Open AI]

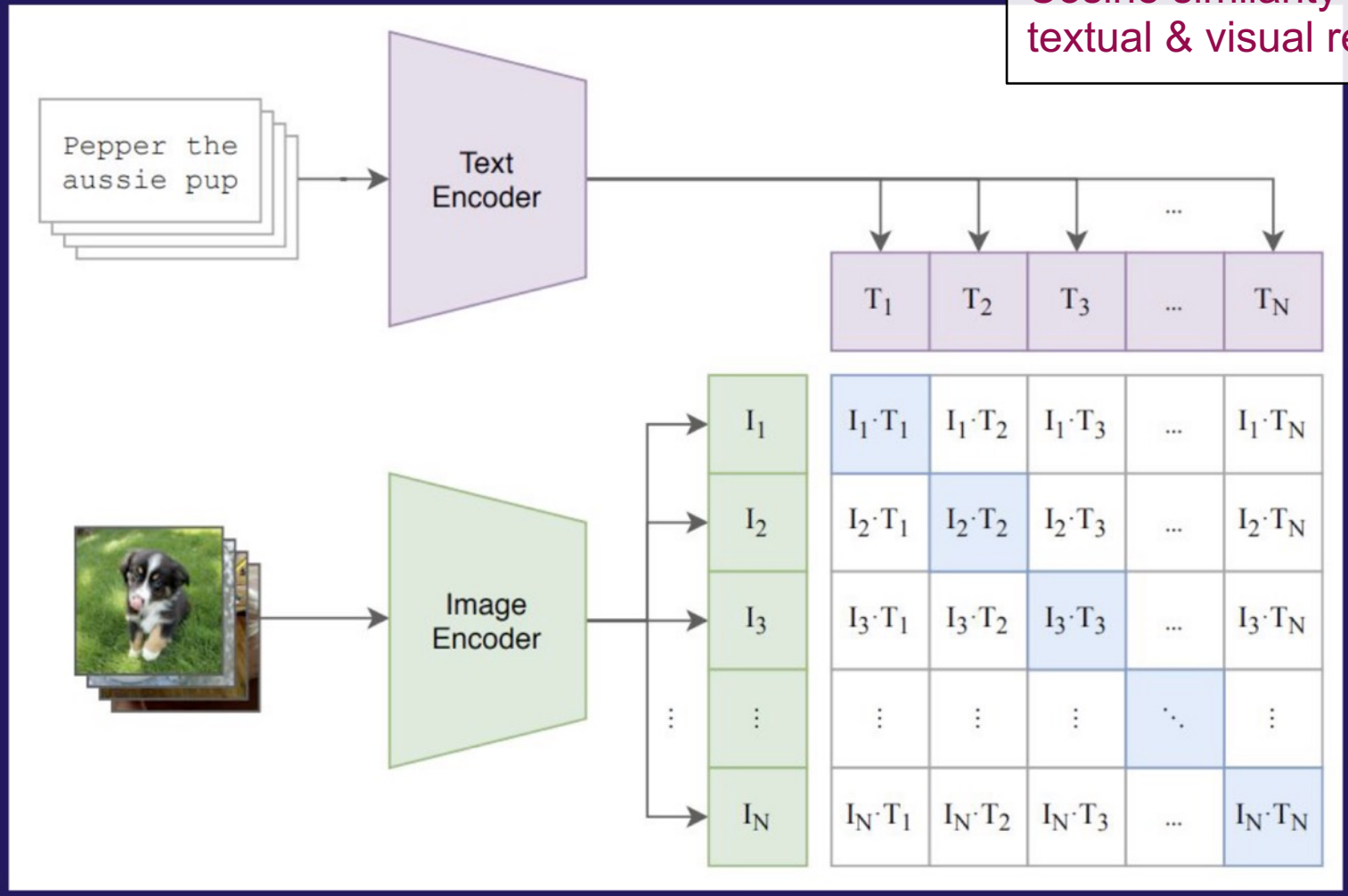
CLIP: Contrastive Language-Image Pre-training



[Slide: Jong Wook Kim – Open AI]

CLIP: Contrastive Language-Image Pre-training

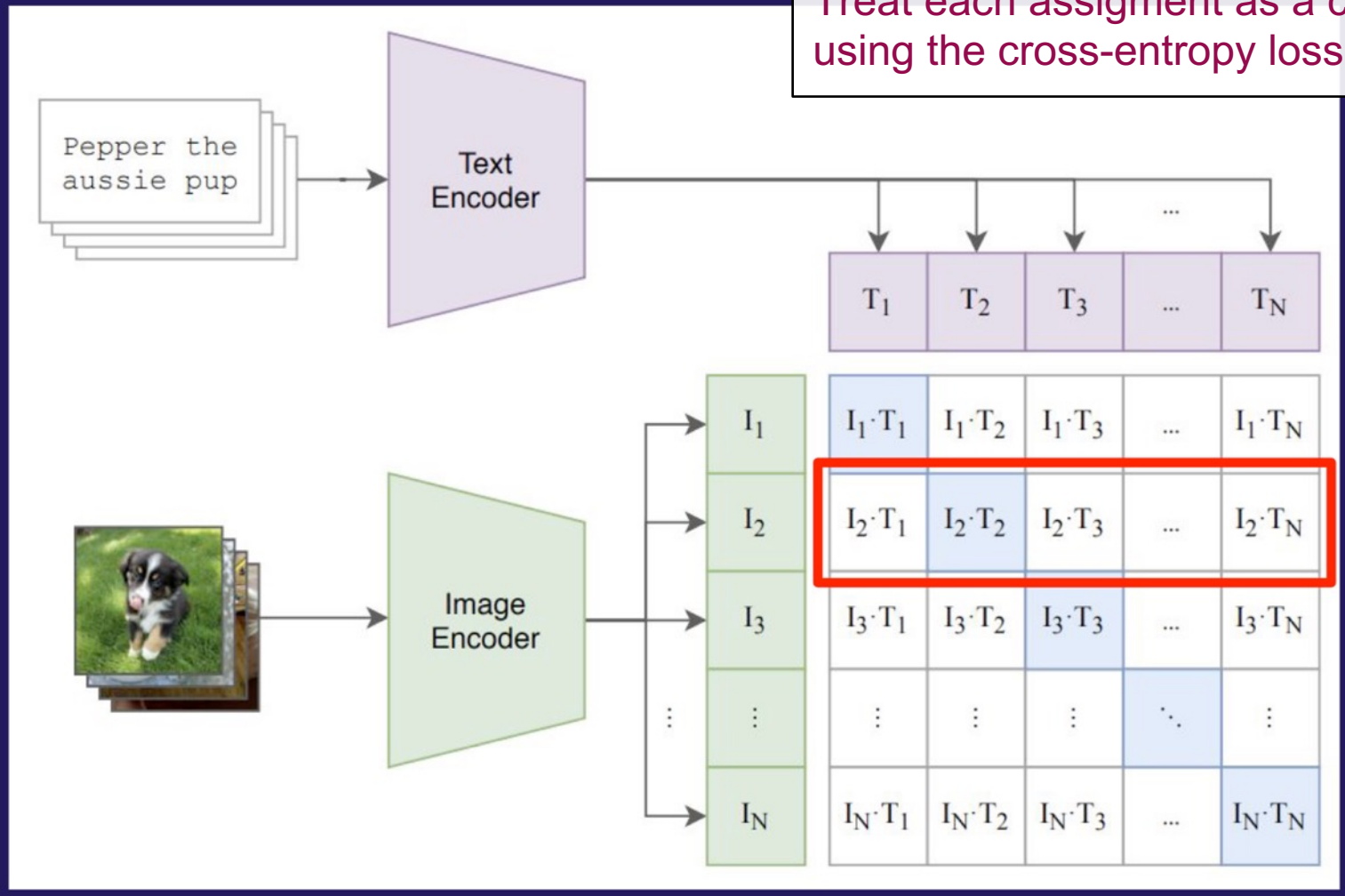
Cosine similarity between
textual & visual representations



[Slide: Jong Wook Kim – Open AI]

CLIP: Contrastive Language-Image Pre-training

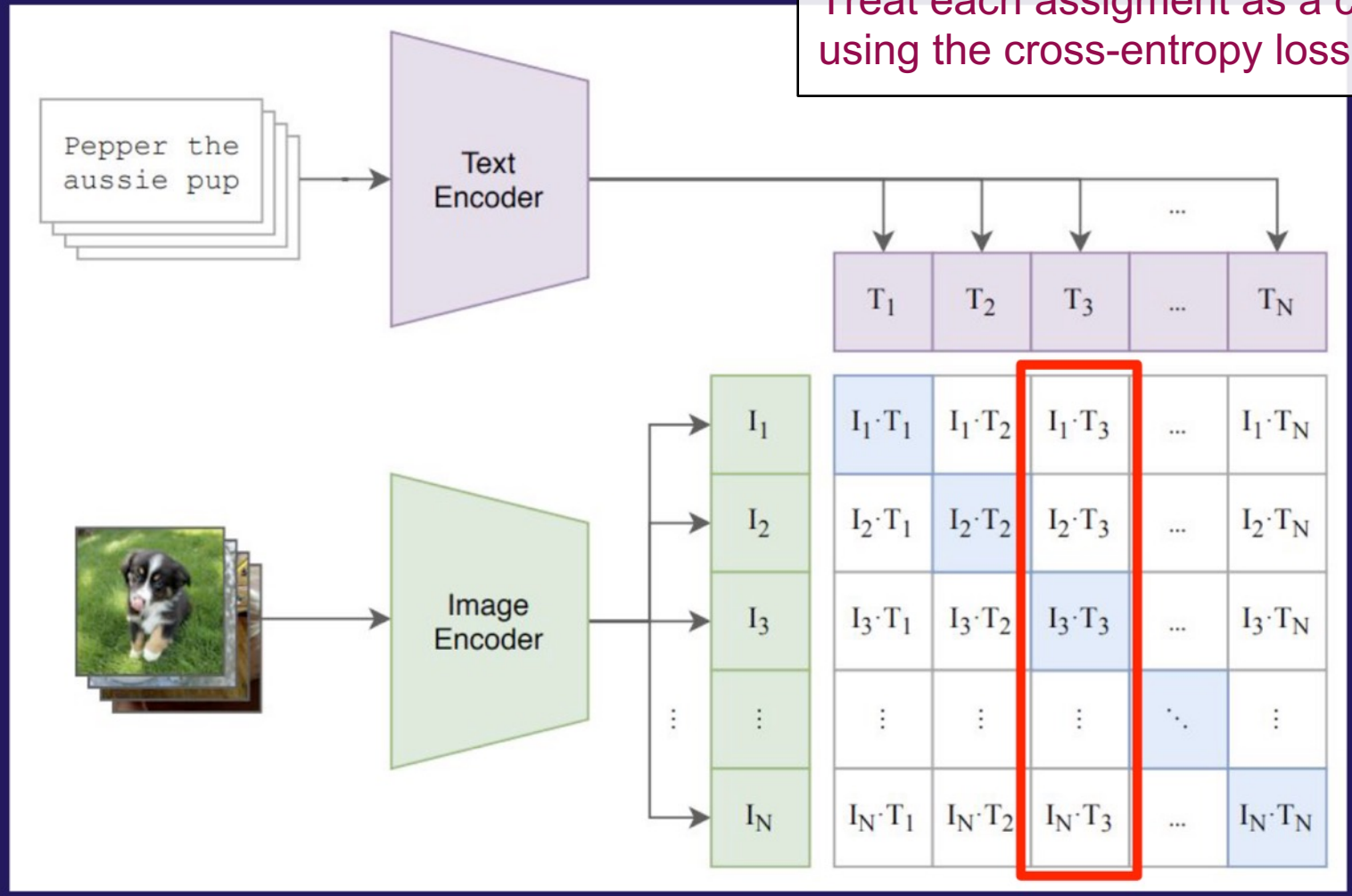
Treat each assigment as a classification problem using the cross-entropy loss.



[Slide: Jong Wook Kim – Open AI]

CLIP: Contrastive Language-Image Pre-training

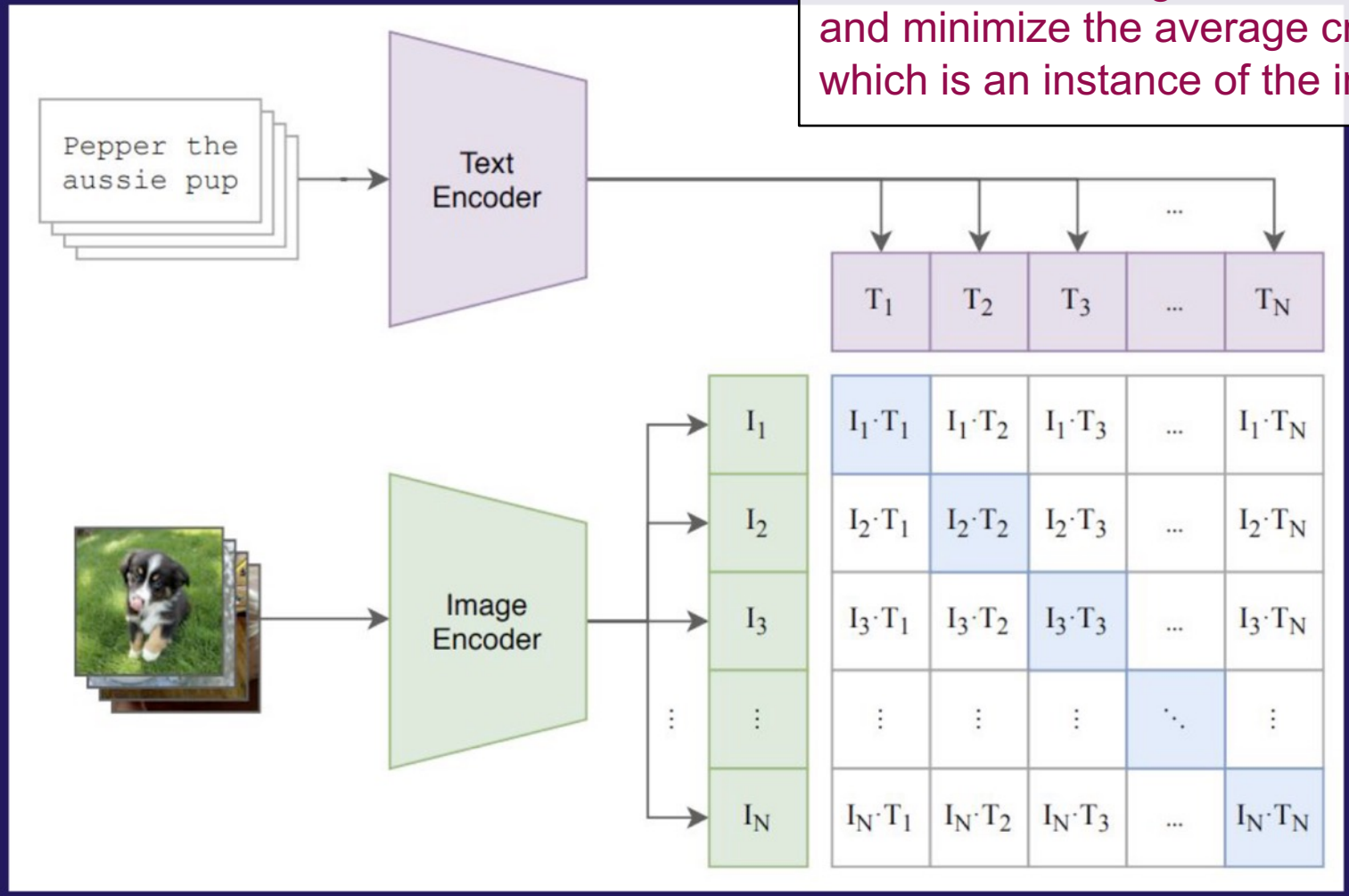
Treat each assigment as a classification problem using the cross-entropy loss.



[Slide: Jong Wook Kim – Open AI]

CLIP: Contrastive Language-Image Pre-training

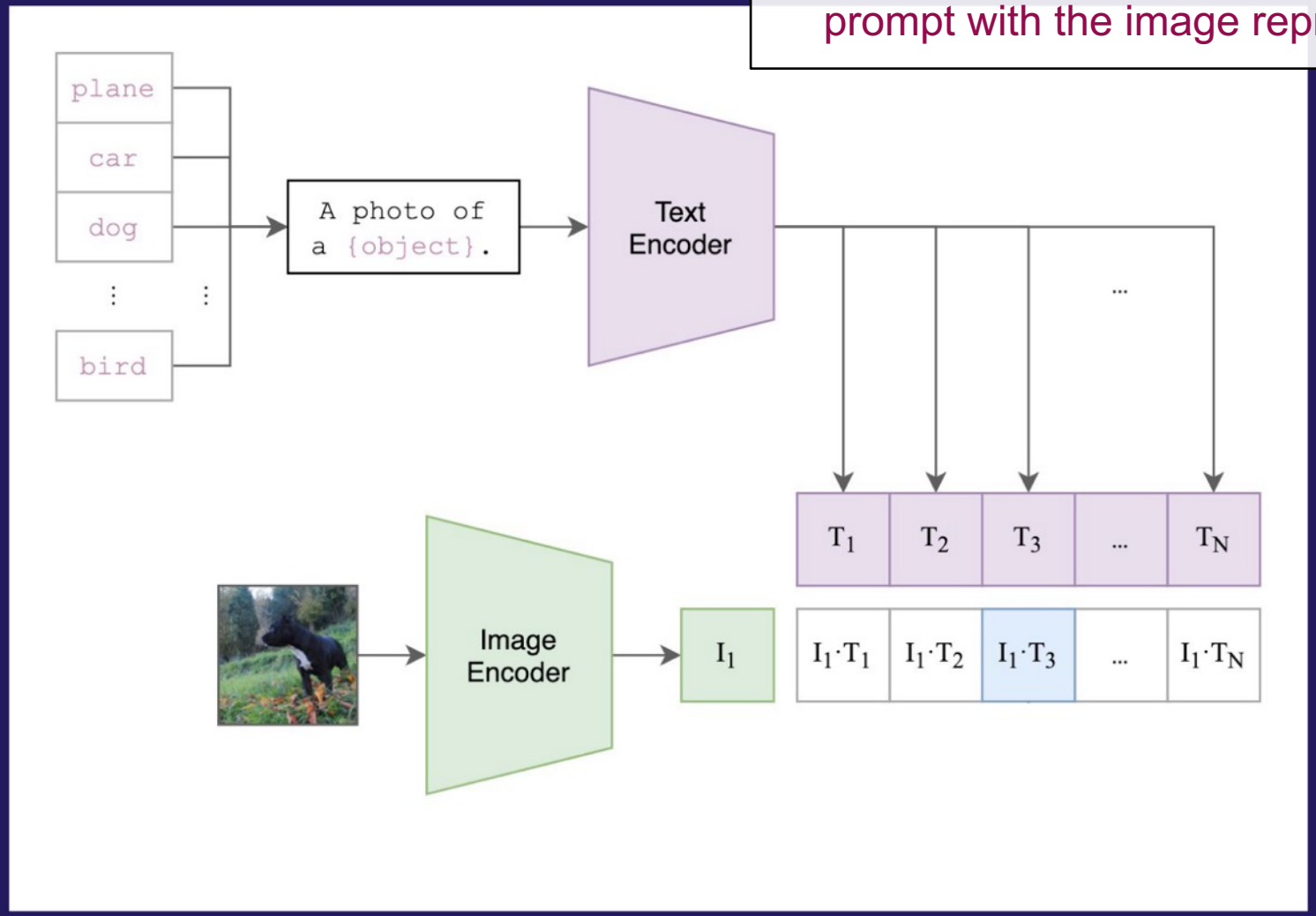
Do this for all images and text in the batch, and minimize the average cross-entropy loss, which is an instance of the infoNCE loss.



[Slide: Jong Wook Kim – Open AI]

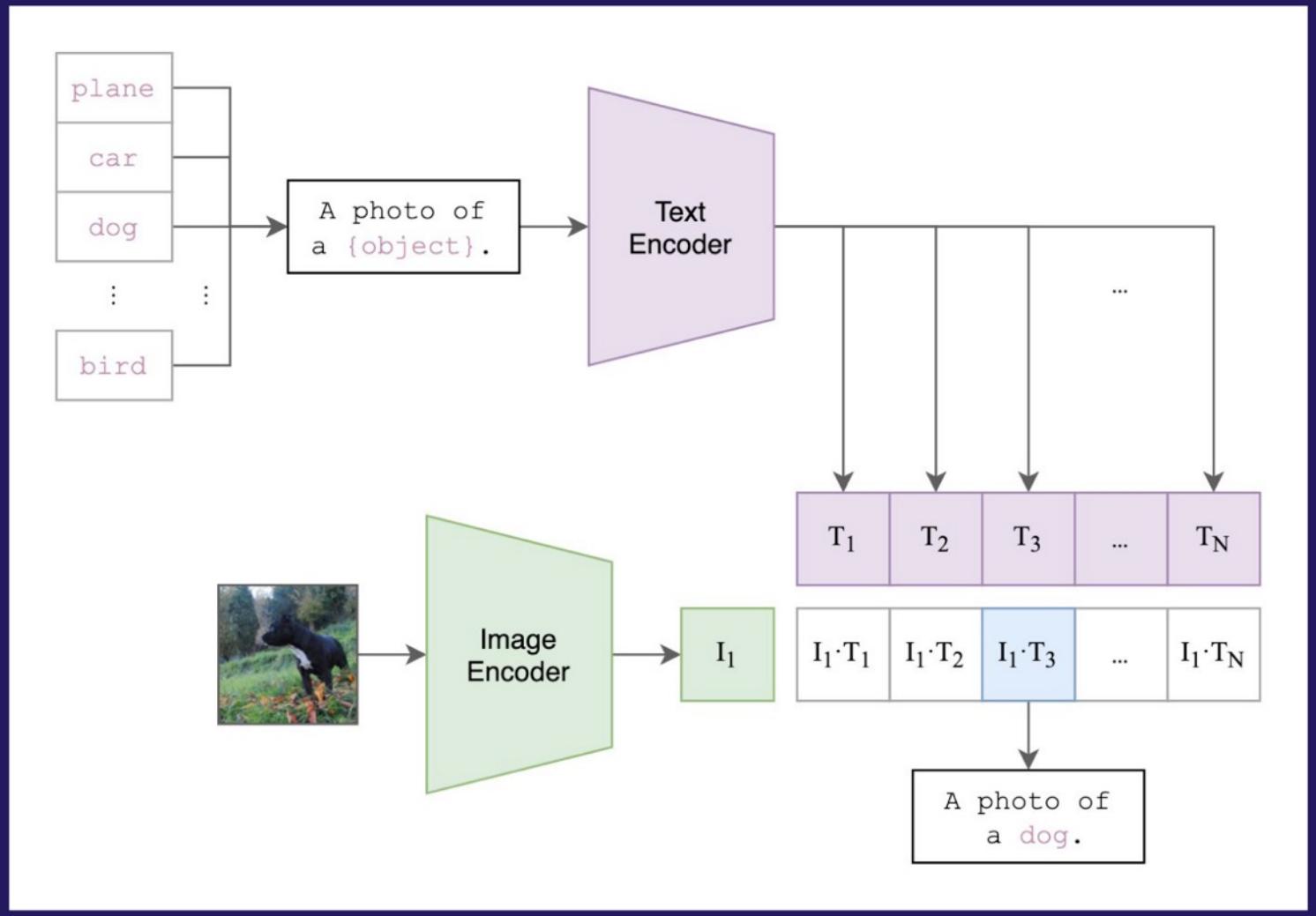
Zero-shot image classification

Main application: zero-shot image classification:
- Construct a text prompt using a template
- Compare the representation of each possible prompt with the image representation



[Slide: Jong Wook Kim – Open AI]

Zero-shot image classification



[Slide: Jong Wook Kim – Open AI]

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

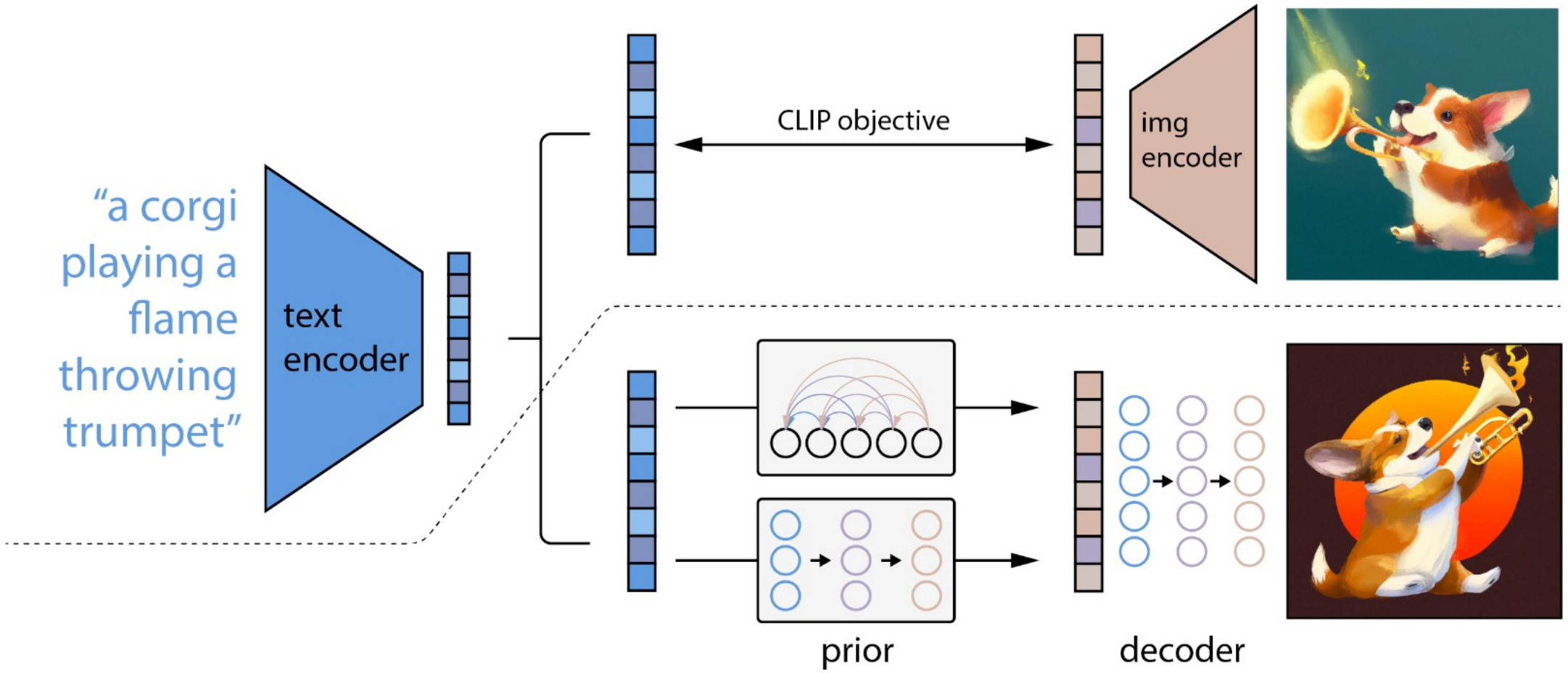
[Slide: Jong Wook Kim – Open AI]

Limitations of CLIP

- Zero-shot performance is well below the SOTA
- Especially weak on abstract tasks such as counting
- Poor on out-of-distribution data such as MNIST
- Susceptible to adversarial attacks
- Dataset selection in the eval suite, use of large validation sets for prompt engineering
- Social biases

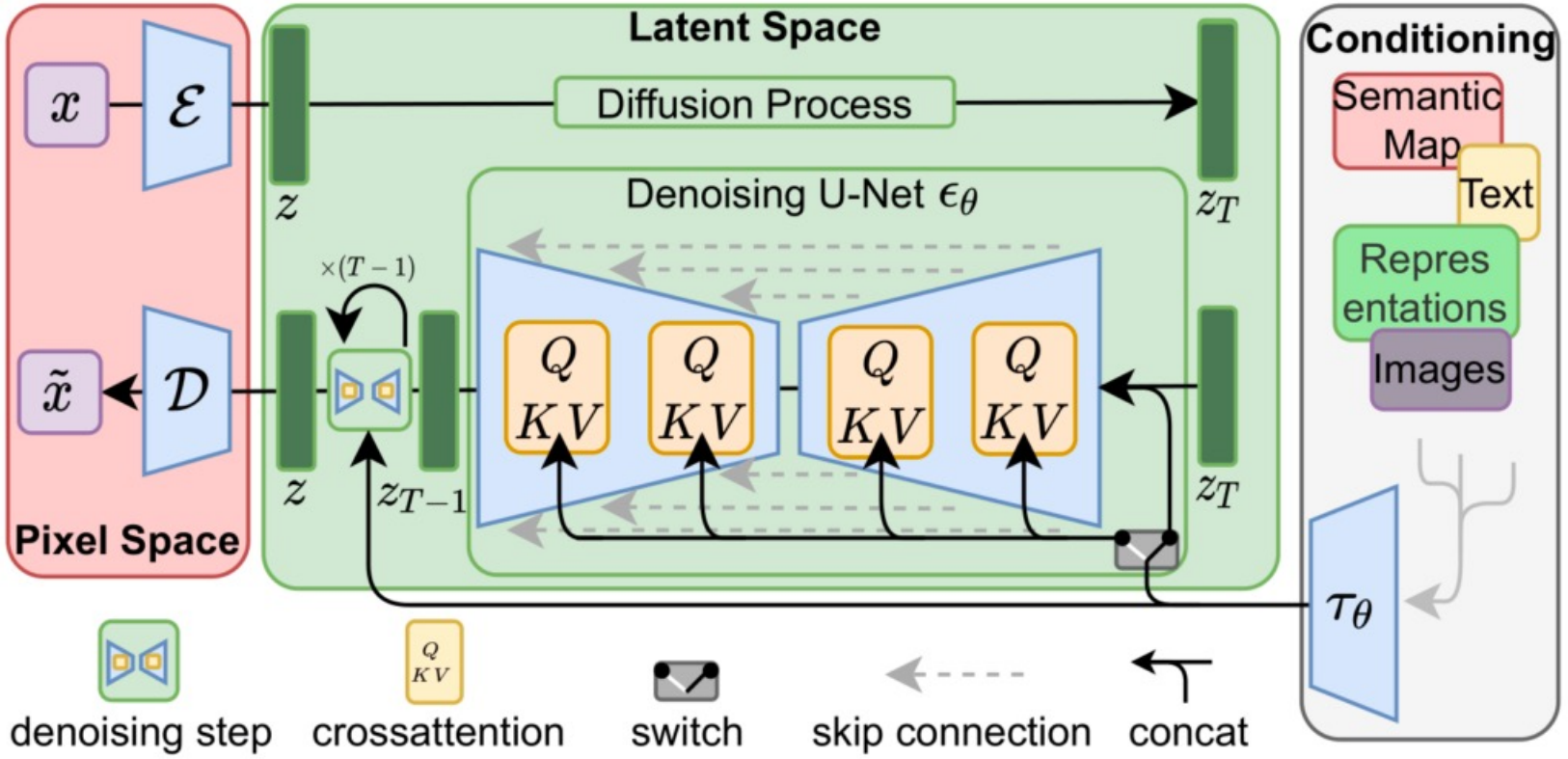
[Slide: Jong Wook Kim – Open AI]

DALLE - Image Generation model based on CLIP



Hierarchical Text-Conditional Image Generation with CLIP Latents
Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, Mark Chen
Arxiv22

Generating images from text: **Stable Diffusion**



High-Resolution Image Synthesis with Latent Diffusion Models
Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, Björn Ommer
CVPR '22 Oral

Generating images from text: **Stable Diffusion**

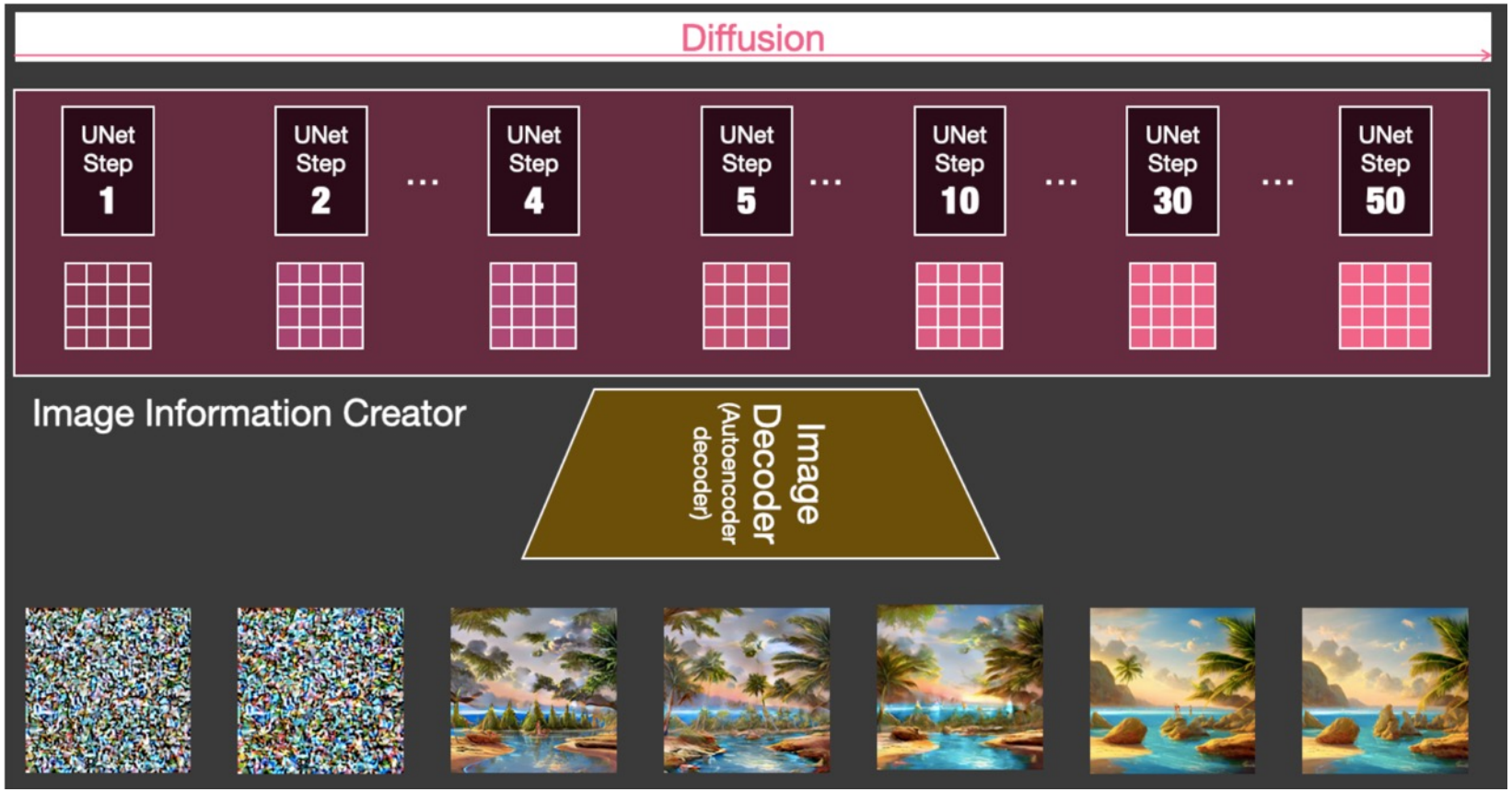


Illustration: https://jalammar.github.io/illustrated-stable-diffusion/

High-Resolution Image Synthesis with Latent Diffusion Models
Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, Björn Ommer
CVPR '22 Oral

Terminologie

Lien entre ces méthodes et les méthodes auto-supervisées

Les méthodes multimodales que nous avons vues entrainent des représentations visuelles (ou plus précisément des modèles pour les produire) en utilisant de l'information provenant d'une autre modalité que l'image.

Ces méthodes sont-elles auto-supervisées ?

Certaines méthodes se présentent comme tel, car:

- Les "annotations" (les éléments de l'autre modalité) ne sont pas produites manuellement, et elles n'ont pas été spécifiquement créées pour entrainer ces modèles. On peut les voir comme des données elles-mêmes.
- Au moment de l'apprentissage, on a toujours la notion de tâche prétexte.

Cependant, la plupart des méthodes qui utilisent le texte considèrent au contraire qu'elles ne sont pas totalement non-supervisées, car elles supposent un accès à d'autres informations que juste les images. Les termes « faiblement supervisé » (*weakly supervised*) ou « supervisé par internet » (*webly supervised*) sont parfois utilisés.