

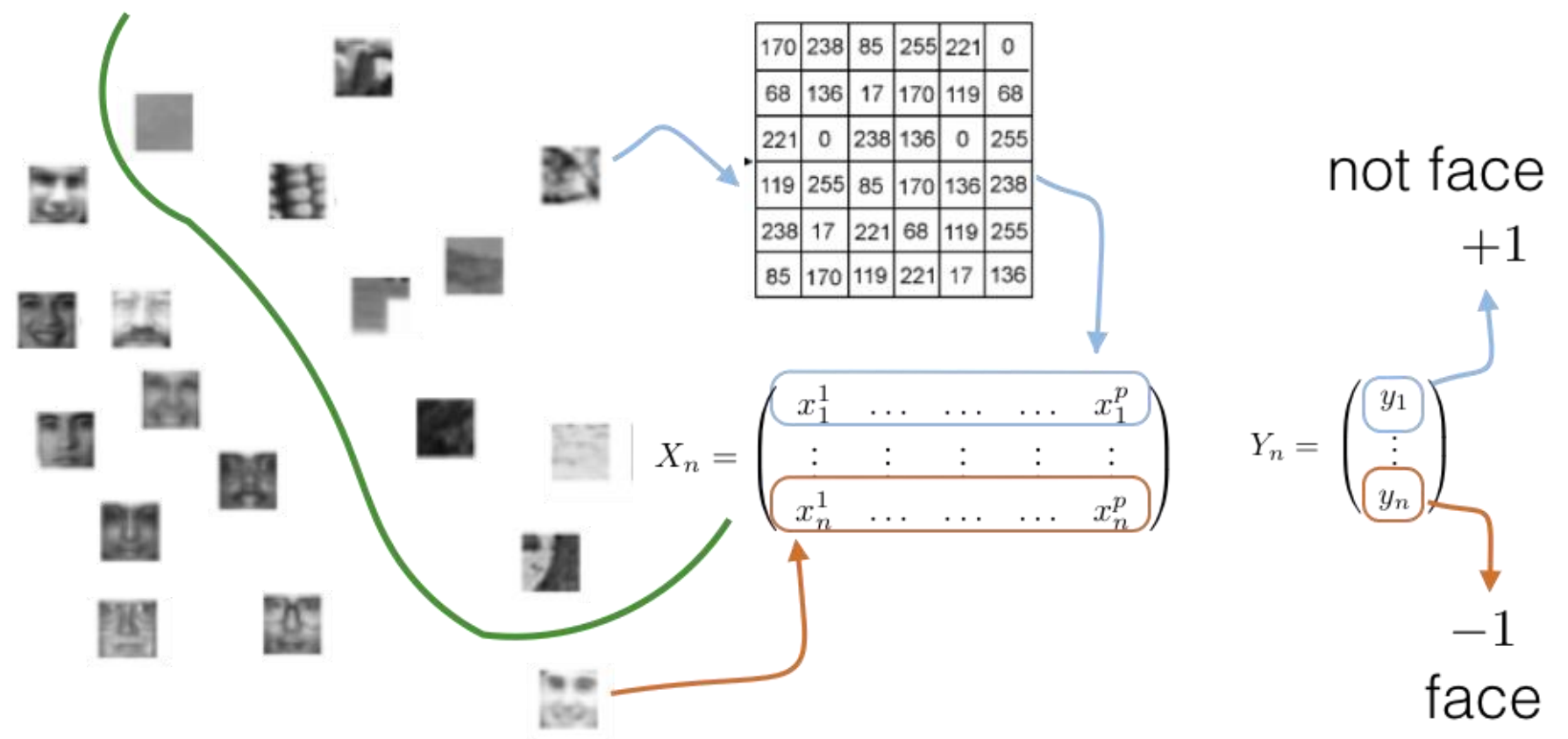
Natural Language supervision

Matteo Moro & Francesca Odone

Image and semantics – supervised approach

Image classification: associate a label to an image

a classical **binary classification** problem



at the end of the training/validation procedure
you derive a model for *faces*

Image and semantics – supervised approach

Multi-class classification: the set of possible labels grows



Image representation with CNN

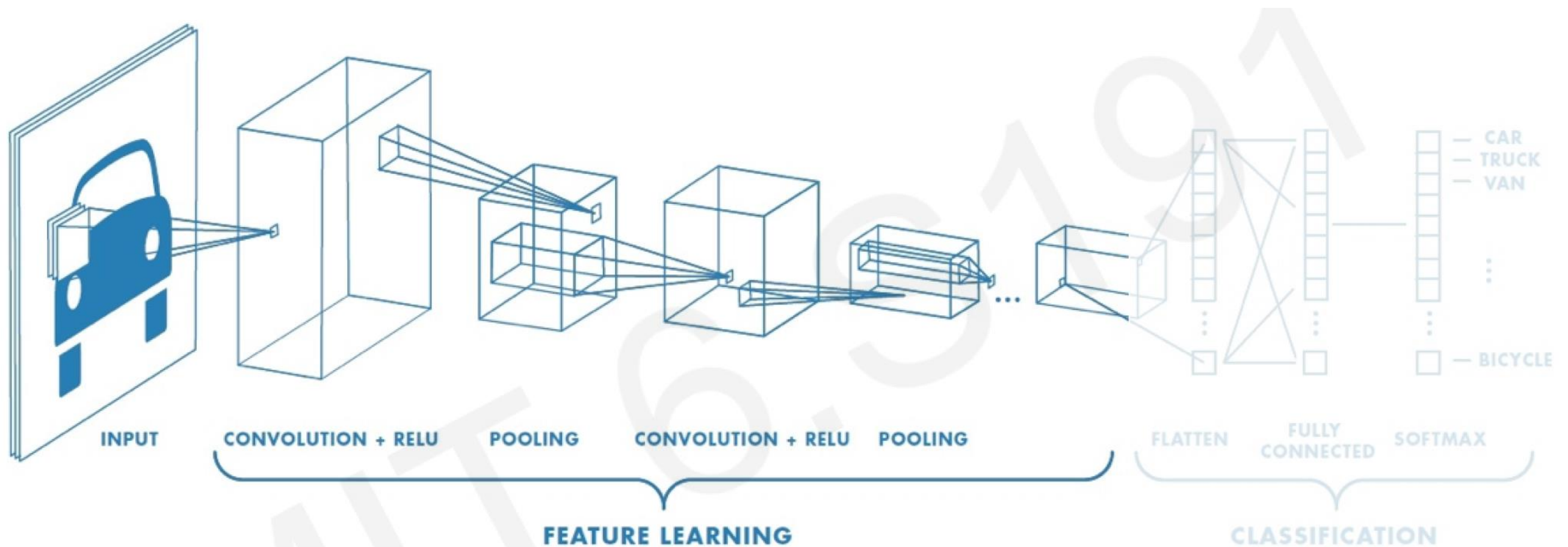
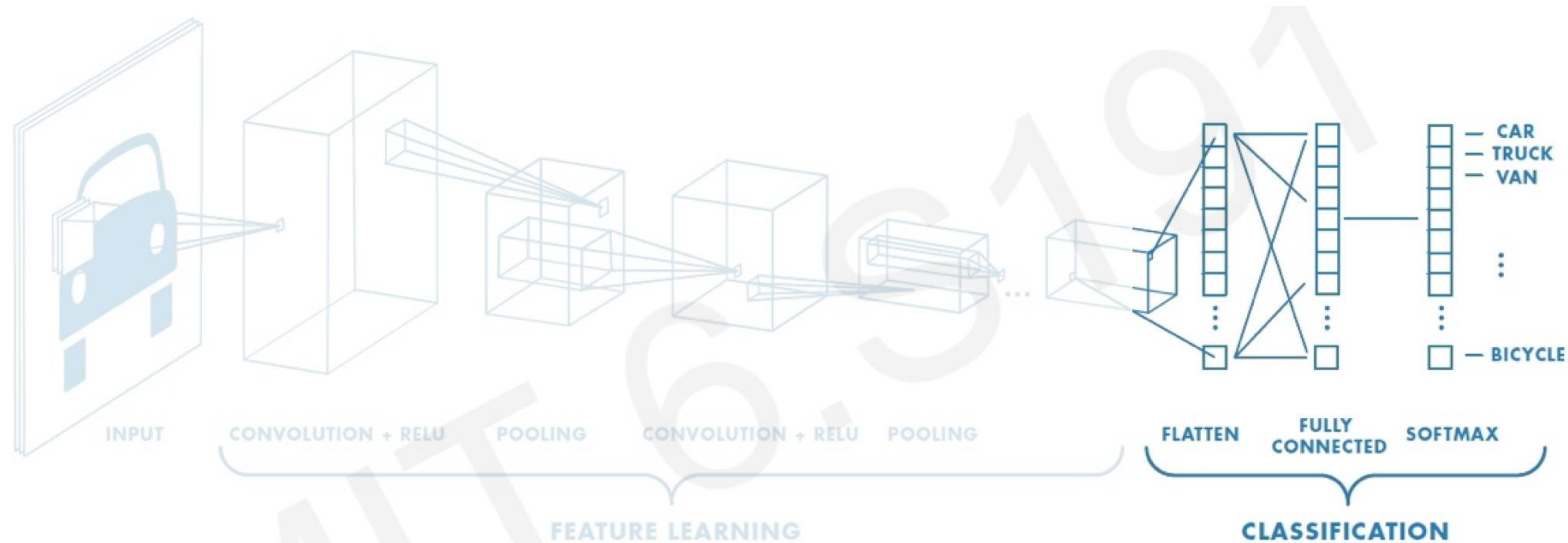


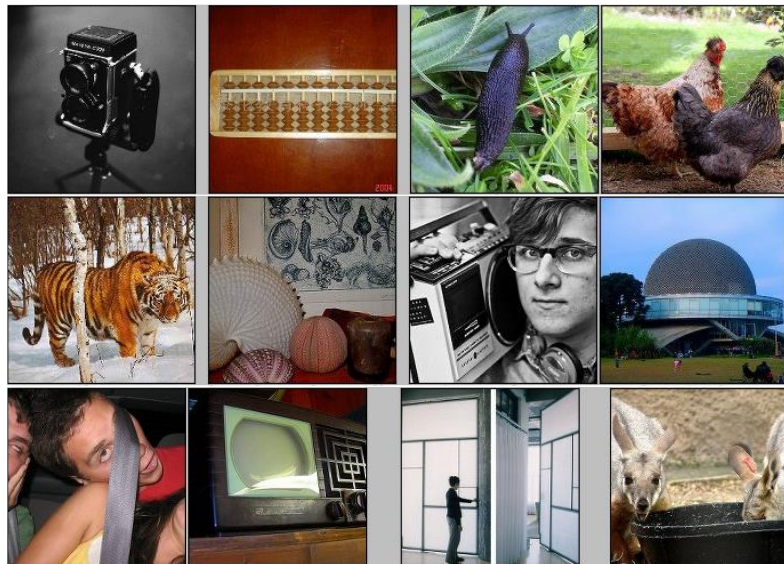
Image classification



$$\text{softmax}(y_i) = \frac{e^{y_i}}{\sum_j e^{y_j}}$$

The key for the success of CNN: ImageNet

IMAGENET



- ... the ImageNet Challenge (2012)
- ~14 million labeled images, 20k classes
- Images gathered from Internet
- Human labels via Amazon Turk (huge scale manual on-line manual labelling)
- ImageNet Object Recognition Challenge: 1.2 million training images, 1000 classes

Major problems

-) We need a huge amount of labelled images and usually we miss variability.
-) The labels are usually referring to just a portion of the image.

Possible solution: examples on the internet

Natural Language supervision

An example of image-text pair

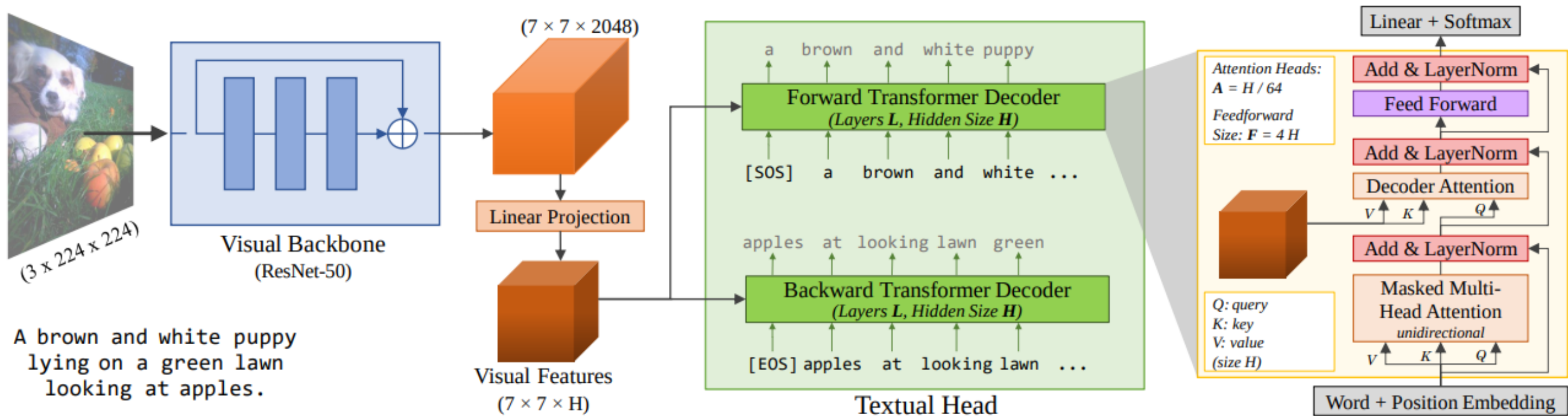


A vibrant street scene in New York City, bustling with pedestrians crossing the street and yellow cabs navigating through the traffic. The image captures the dynamic urban life amid the backdrop of towering skyscrapers under a partly cloudy sky, epitomizing the bustling energy of the city that never sleeps.

New concepts

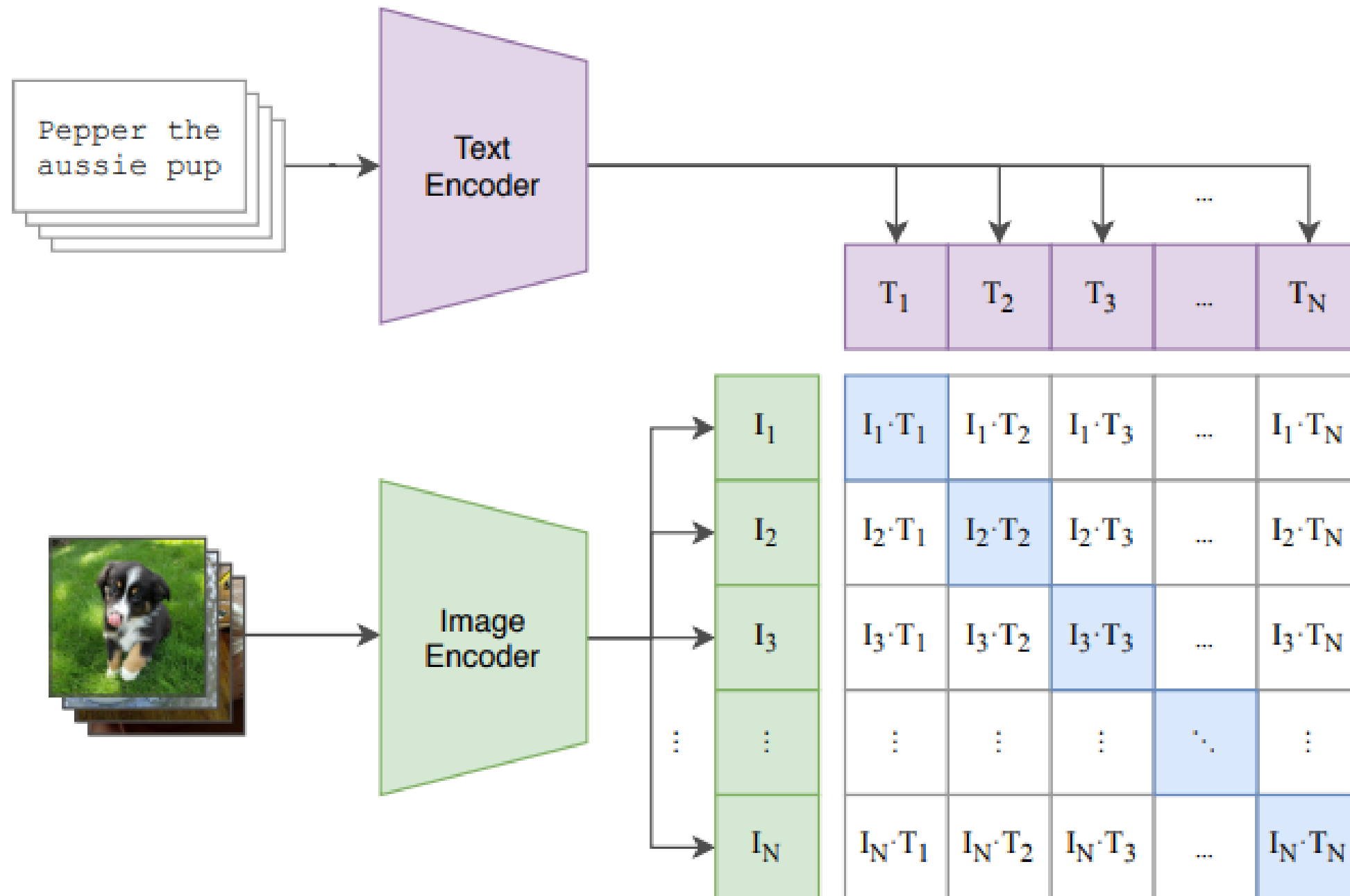
-) Find a way to extract representation from text
-) Put together image and text representation

Learn image captions



Desai, Karan, and Justin Johnson. "Virtex: Learning visual representations from textual annotations." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021.

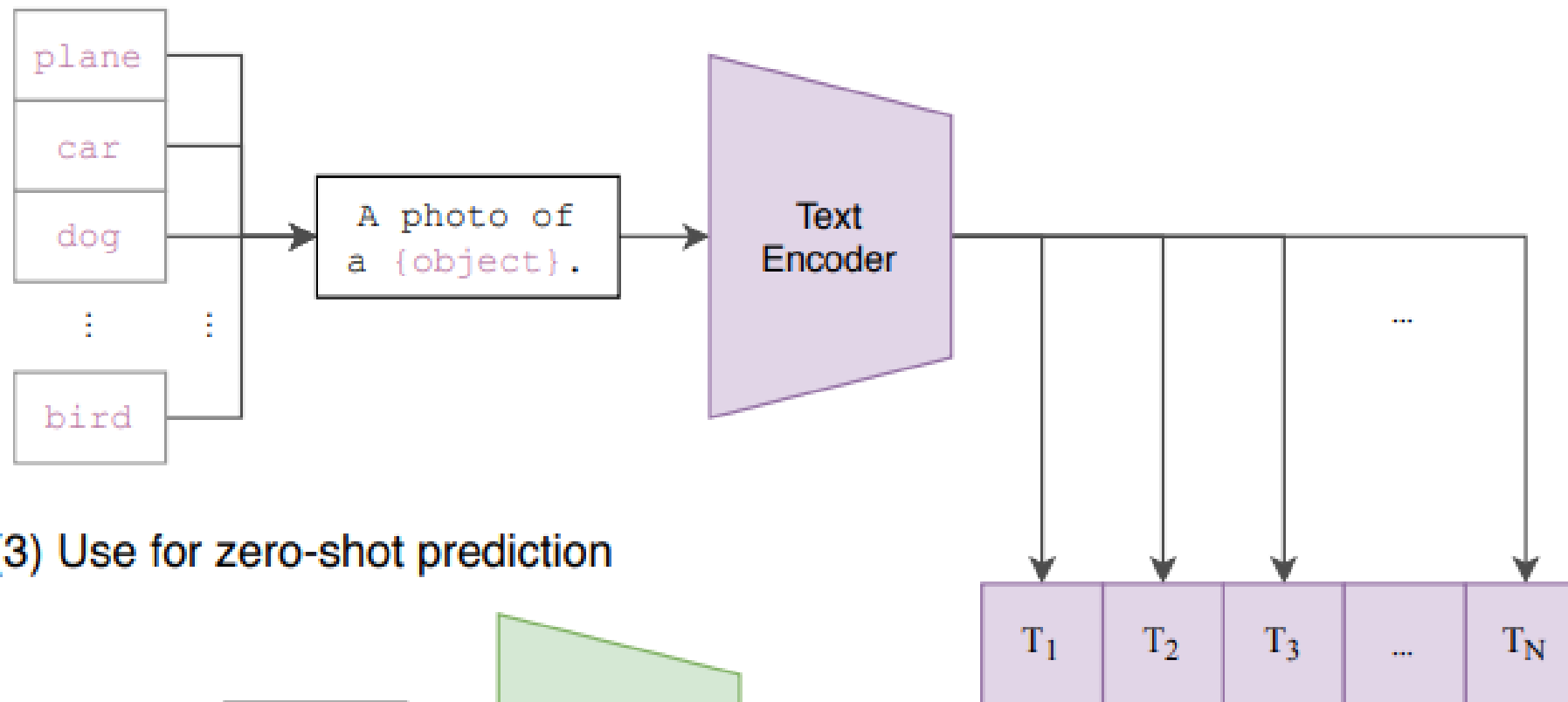
CLIP: Contrastive Language-Image Pre-Training



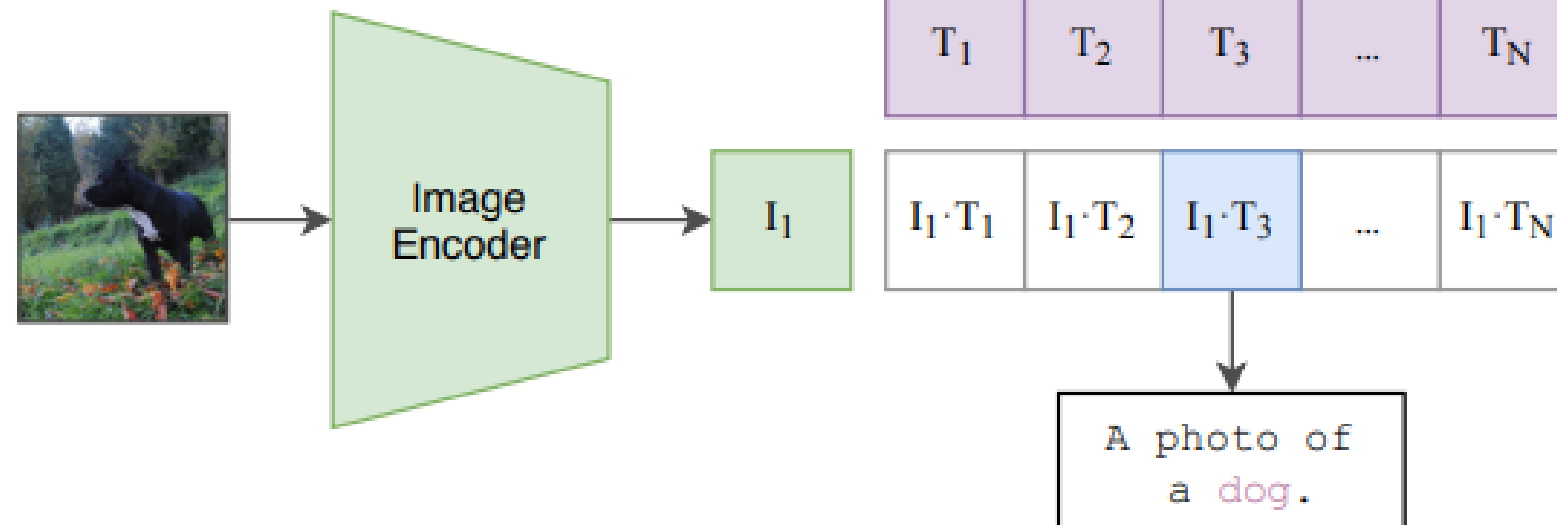
Radford, Alec, et al. "Learning transferable visual models from natural language supervision." International conference on machine learning. PmLR, 2021.

CLIP – Zero-shot transfer

(2) Create dataset classifier from label text

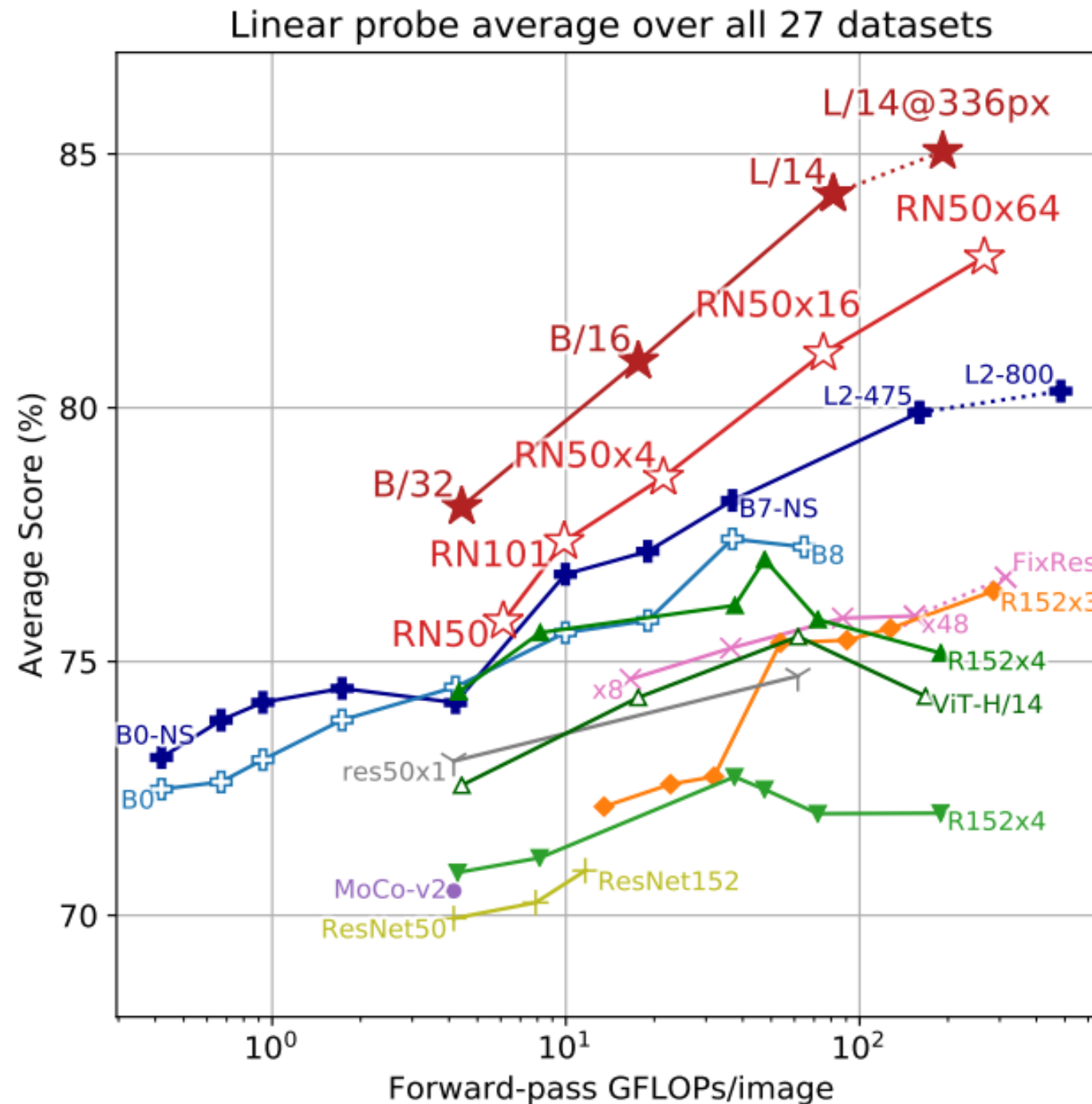


(3) Use for zero-shot prediction









Radford, Alec, et al. "Learning transferable visual models from natural language supervision." International conference on machine learning. PmLR, 2021.

CLIP – Representation Learning



Radford, Alec, et al. "Learning transferable visual models from natural language supervision." International conference on machine learning. PmLR, 2021.

CLIP – Natural Distribution Shift

	Dataset Examples	ImageNet ResNet101	Zero-Shot CLIP	Δ Score
ImageNet		76.2	76.2	0%
ImageNetV2		64.3	70.1	+5.8%
ImageNet-R		37.7	88.9	+51.2%
ObjectNet		32.6	72.3	+39.7%
ImageNet Sketch		25.2	60.2	+35.0%
ImageNet-A		2.7	77.1	+74.4%

Radford, Alec, et al. "Learning transferable visual models from natural language supervision." International conference on machine learning. PmLR, 2021.

CLIP: Contrastive Language-Image Pre-Training

CLIP learns a multimodal embedding space by jointly training an image encoder and text encoder to maximize the cosine similarity of the image and text embeddings of the N real pairs in the training set

Radford, Alec, et al. "Learning transferable visual models from natural language supervision." International conference on machine learning. PmLR, 2021.

Useful links

<https://openai.com/index/clip/>

<https://openai.com/index/dall-e-3/>

UniGe

