

vision-language models

CS 685, Spring 2024

Advanced Natural Language Processing

<http://people.cs.umass.edu/~miyyer/cs685/>

Mohit Iyyer

College of Information and Computer Sciences

University of Massachusetts Amherst

some slides adapted from Vicente Ordonez, Fei-Fei Li, Justin Johnson, and Jacob Andreas

image captioning



a red truck is parked on
a street lined with trees

visual question answering



- Is this truck considered “vintage”?
- Does the road look new?
- What kind of tree is behind the truck?

we've seen how to compute
representations of words and
sentences. what about images?

grayscale images are matrices



0	3	2	5	4	7	6	9	8
3	0	1	2	3	4	5	6	7
2	1	0	3	2	5	4	7	6
5	2	3	0	1	2	3	4	5
4	3	2	1	0	3	2	5	4
7	4	5	2	3	0	1	2	3
6	5	4	3	2	1	0	3	2
9	6	7	4	5	2	3	0	1
8	7	6	5	4	3	2	1	0

what range of values can each pixel take?

color images are tensors



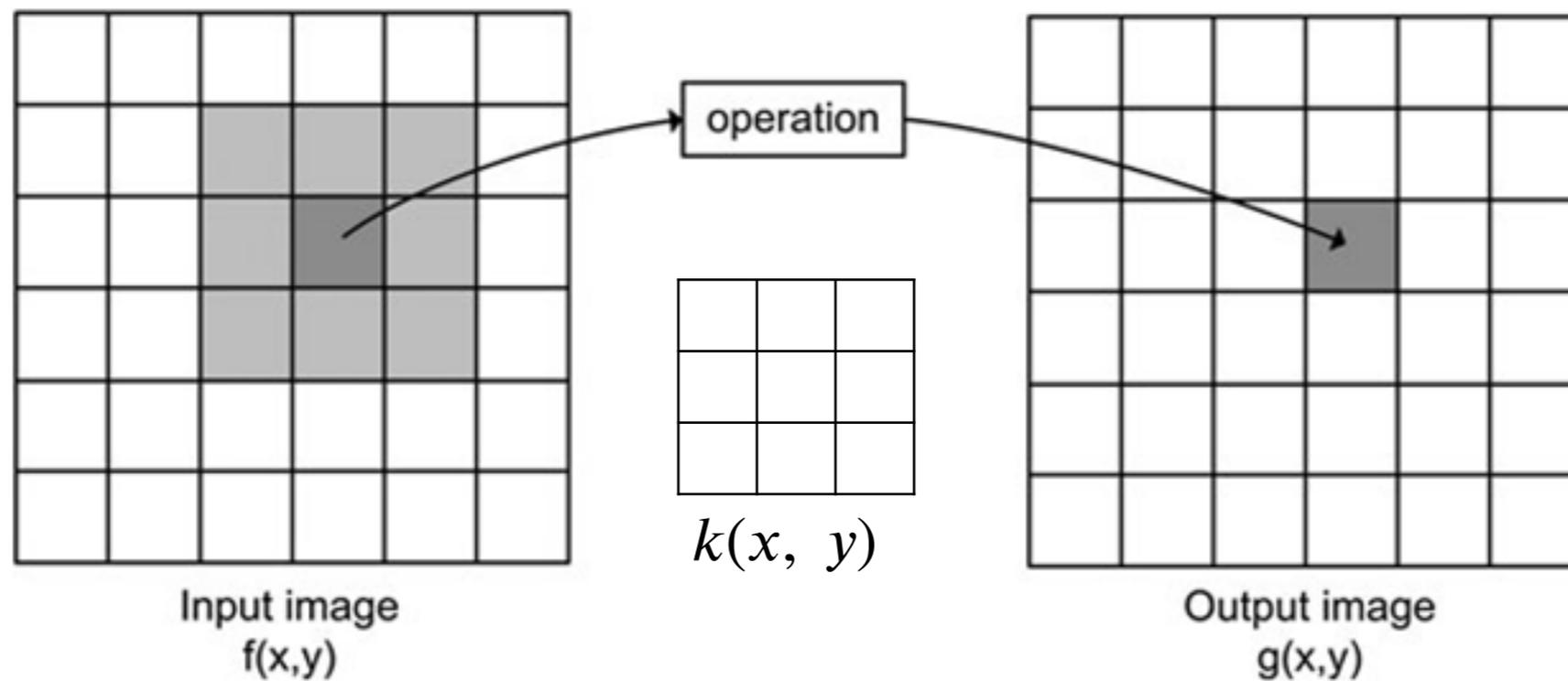
0	3	2	5	4	7	6	0	8
0	3	2	5	4	7	6	0	8
3	0	1	2	3	4	5	6	7
2	1	0	3	2	5	4	7	6
5	2	3	0	1	2	3	4	5
4	3	2	1	0	3	2	5	4
7	4	5	2	3	0	1	2	3
6	5	4	3	2	1	0	3	2
9	6	7	4	5	2	3	0	1
8	7	6	5	4	3	2	1	0

channel x height x width

Channels are usually RGB: Red, Green, and Blue

Other color spaces: HSV, HSL, LUV, XYZ, Lab, CMYK, etc

Convolution operator



$$g(x, y) = \sum_v \sum_u k(u, v) f(x - u, y - v)$$

(filter, kernel)

Input image

*

Weights



Output image

4	5	7	6	6
3	2	8	0	7
6	7	7	1	5
3	0	1	1	1
4	3	2	1	7

*

0	0	0
1	0	1
0	0	0

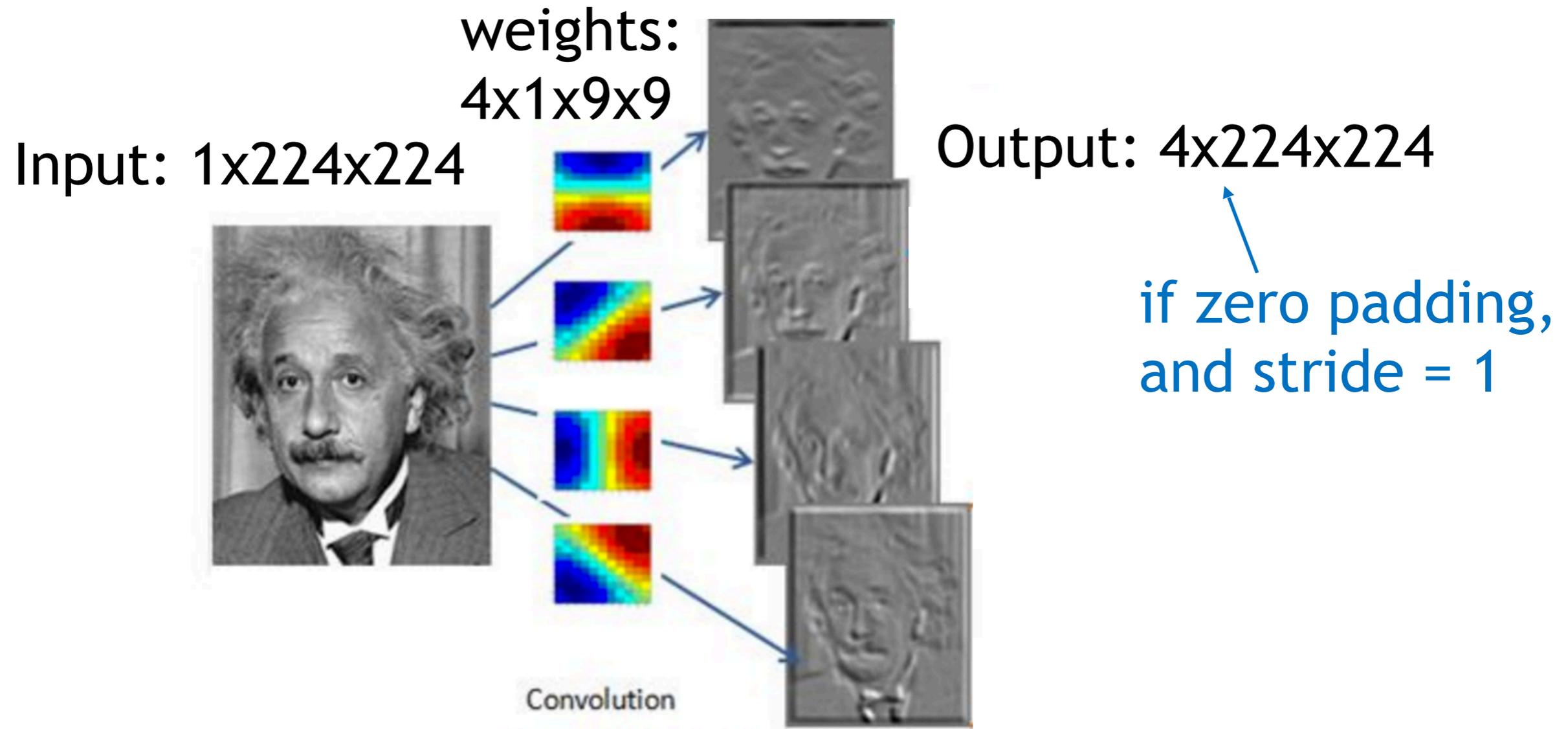


	11	2	15	
	13	8	12	
	?			

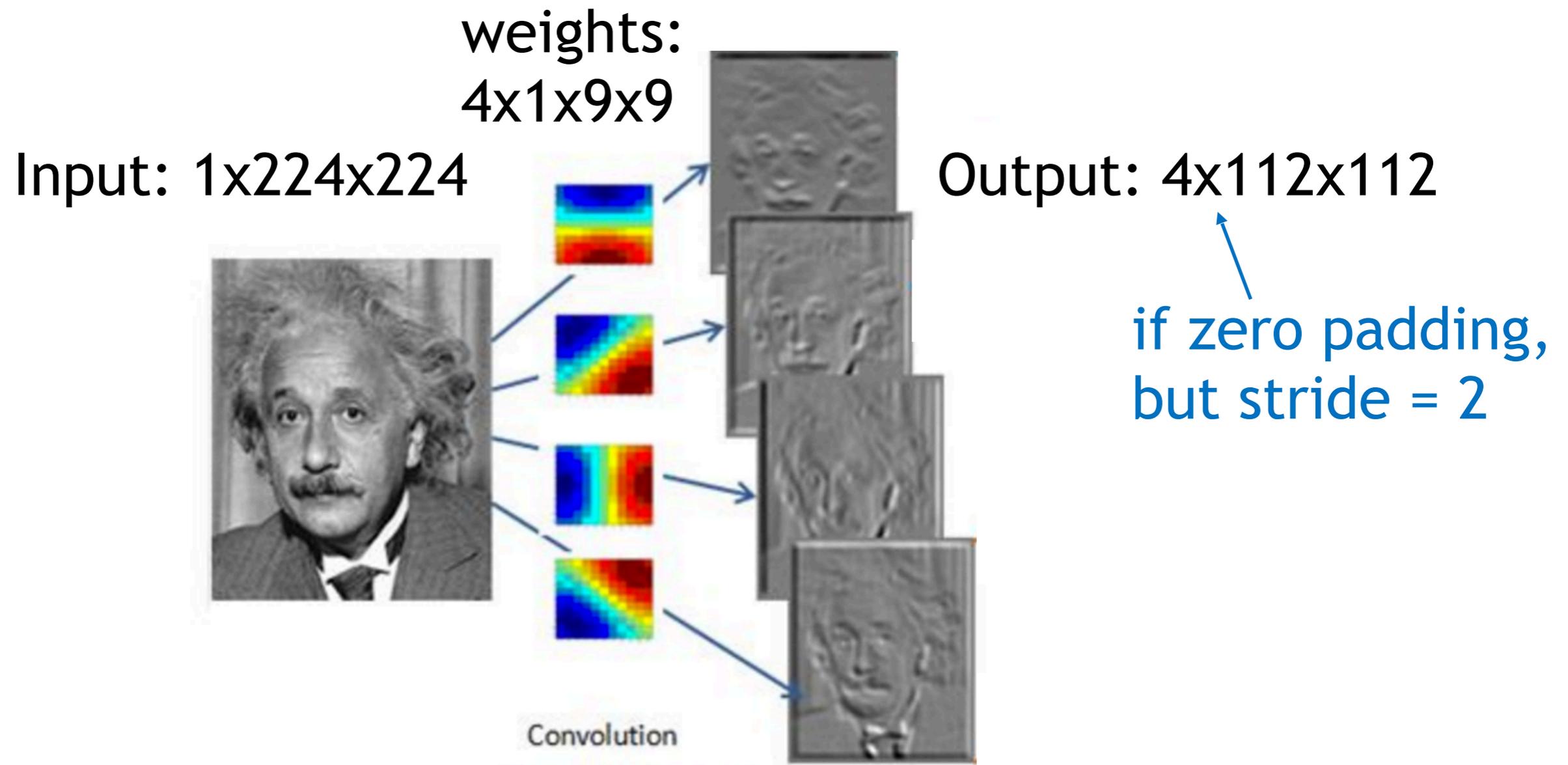
demo:

<http://setosa.io/ev/image-kernels/>

Convolutional Layer (with 4 filters)



Convolutional Layer (with 4 filters)



Alexnet

ImageNet Classification with Deep Convolutional Neural Networks

Alex Krizhevsky
University of Toronto
kriz@cs.utoronto.ca

Ilya Sutskever
University of Toronto
ilya@cs.utoronto.ca

Geoffrey E. Hinton
University of Toronto
hinton@cs.utoronto.ca

the paper that started the
deep learning revolution!

image classification

Classify an image into 1000 possible classes:

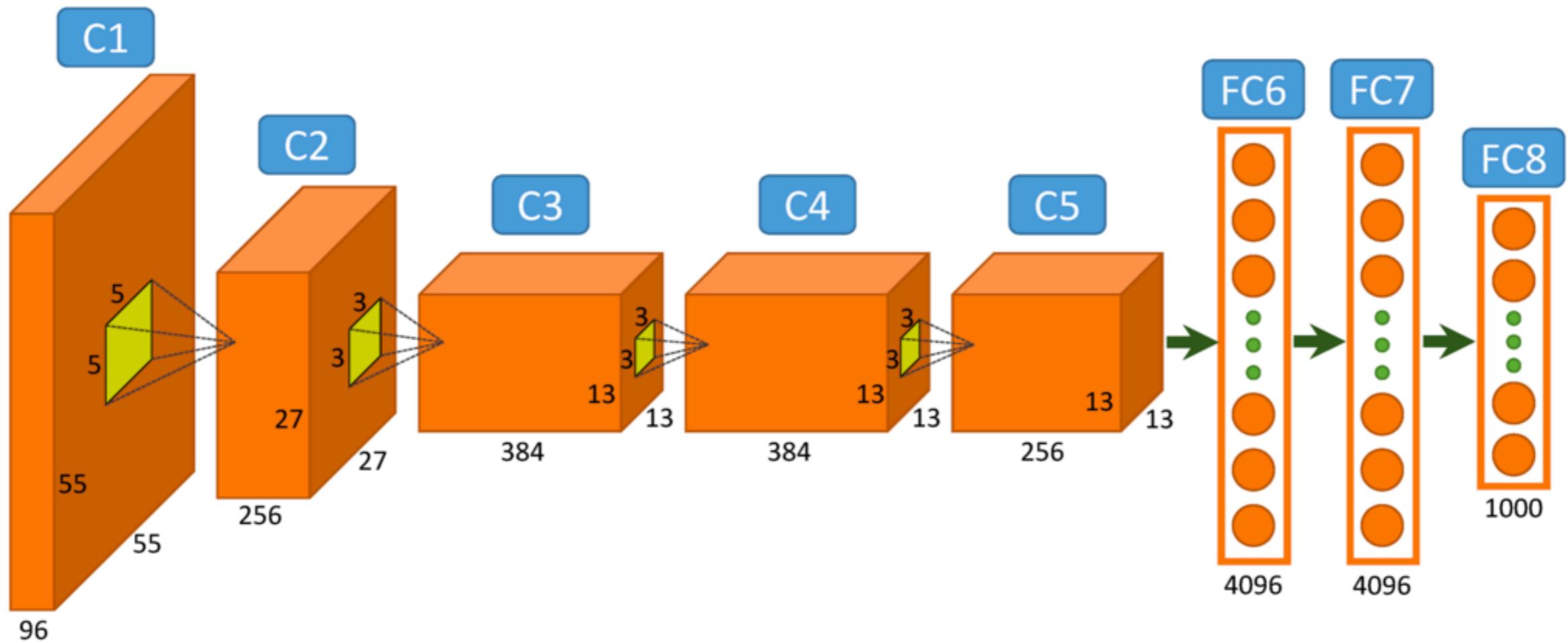
e.g. Abyssinian cat, Bulldog, French Terrier, Cormorant,
Chickadee,
red fox, banjo, barbell, hourglass, knot, maze, viaduct, etc.



cat, tabby cat (0.71)
Egyptian cat (0.22)
red fox (0.11)
.....

train on the ImageNet
challenge dataset,
~1.2 million images

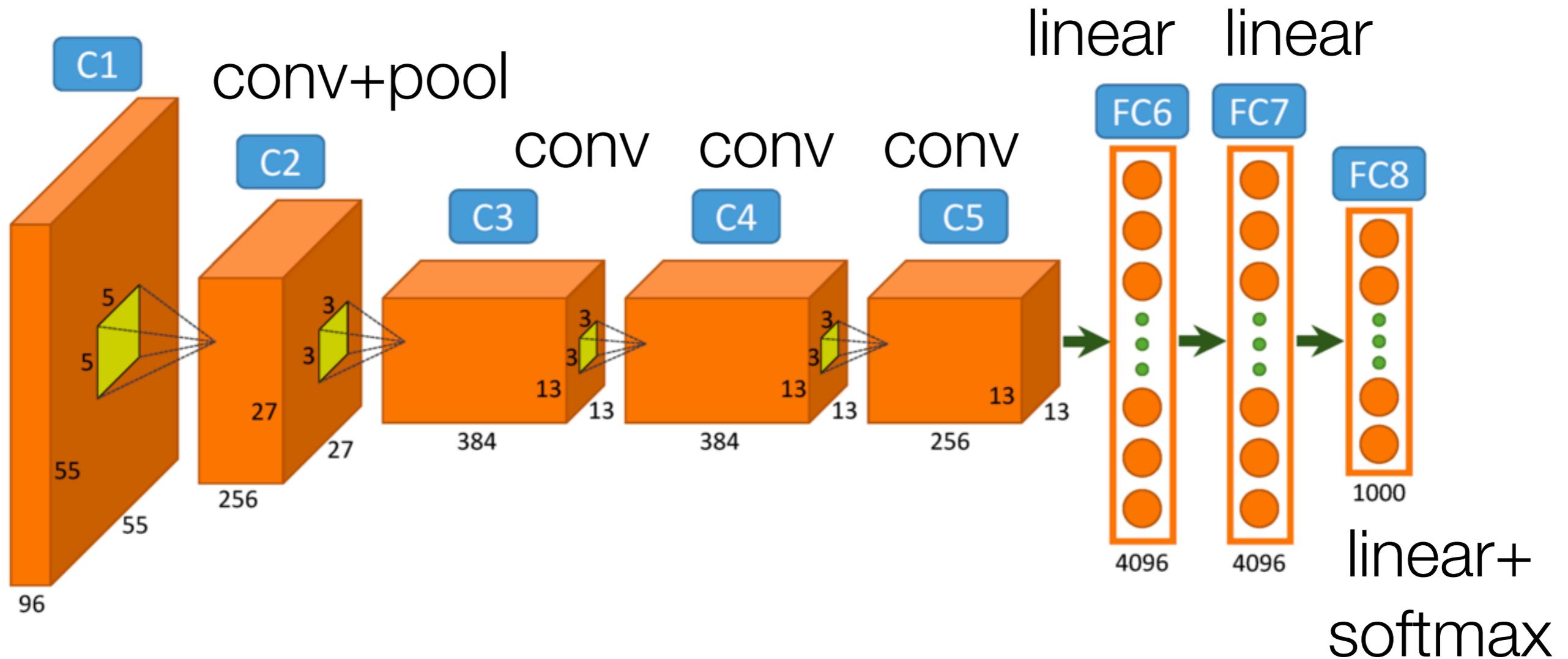
Alexnet



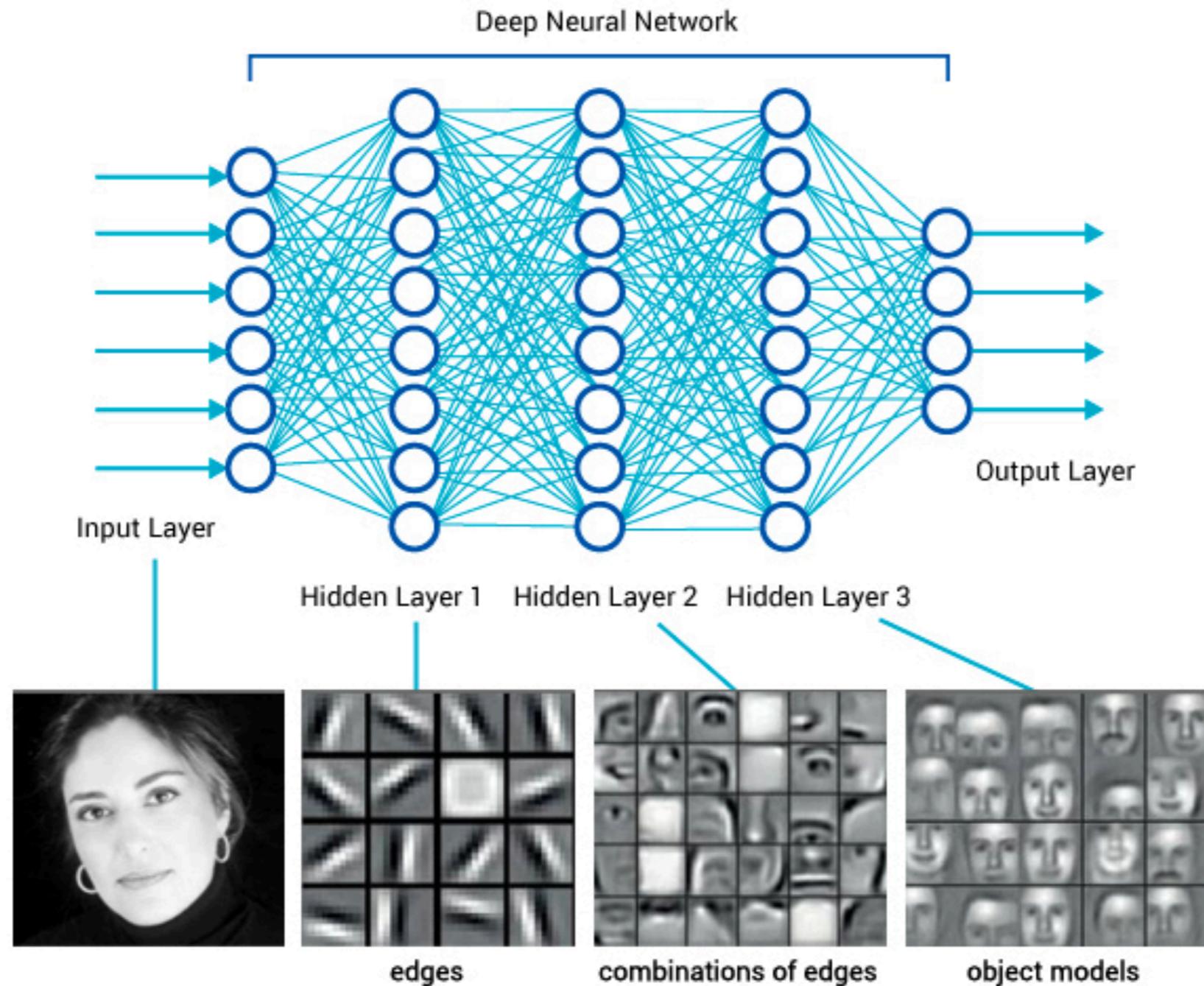
<https://www.saagie.com/fr/blog/object-detection-part1>

Alexnet

conv+pool



What is happening?



Revolution of Depth

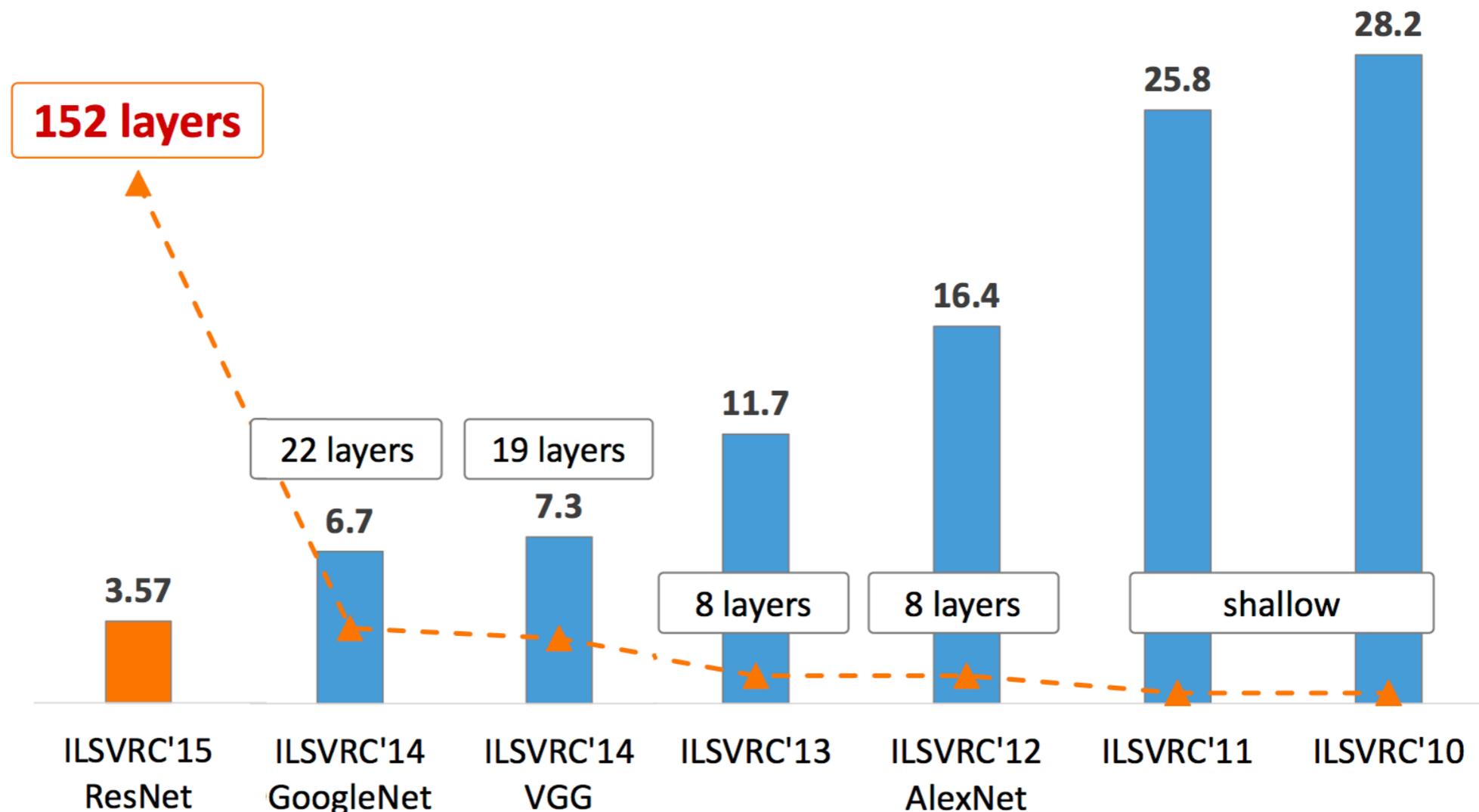
AlexNet, 8 layers
(ILSVRC 2012)



VGG, 19 layers
(ILSVRC 2014)

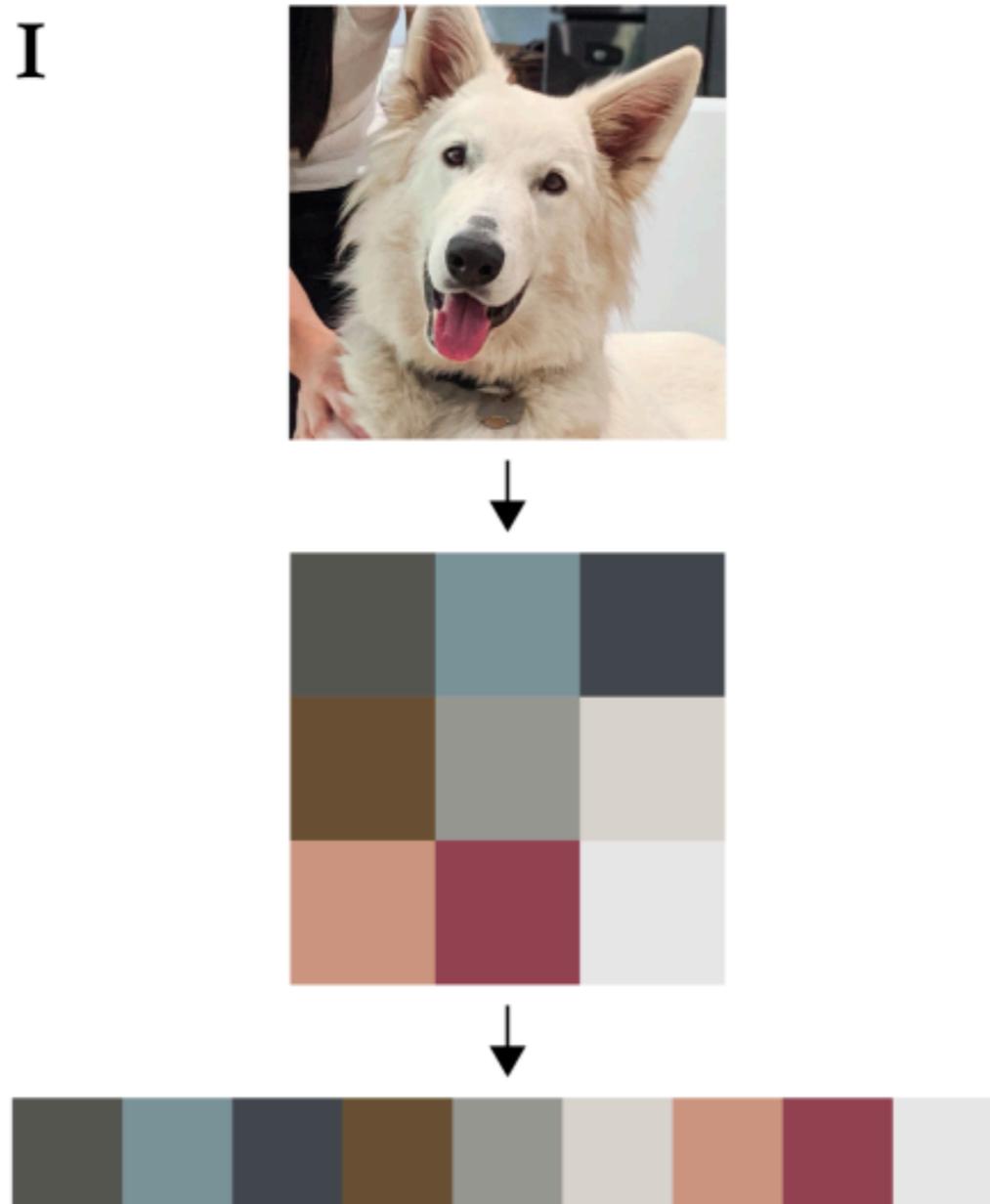


ResNet, **152 layers**
(ILSVRC 2015)

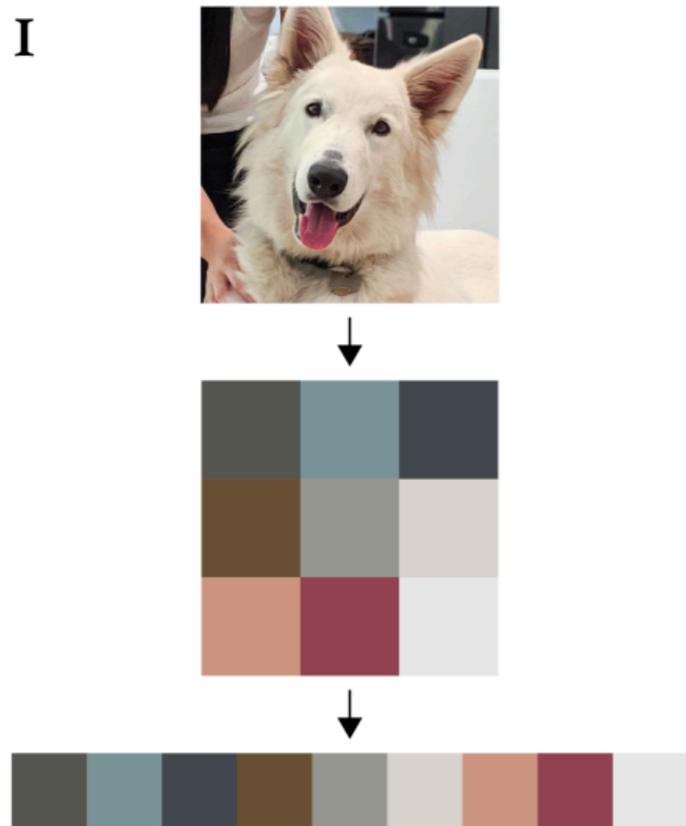


Transformer encoders for vision

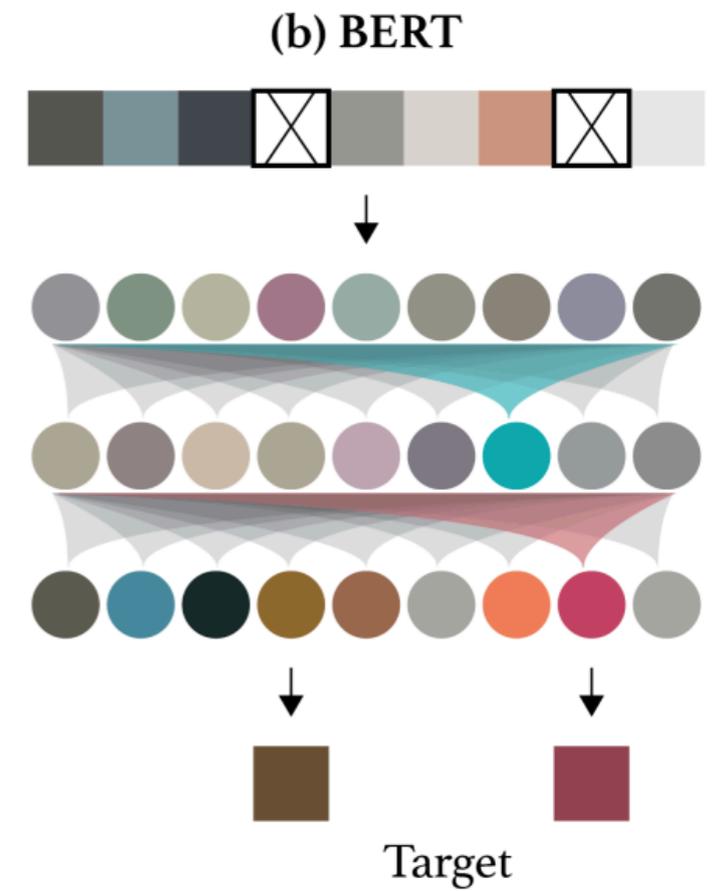
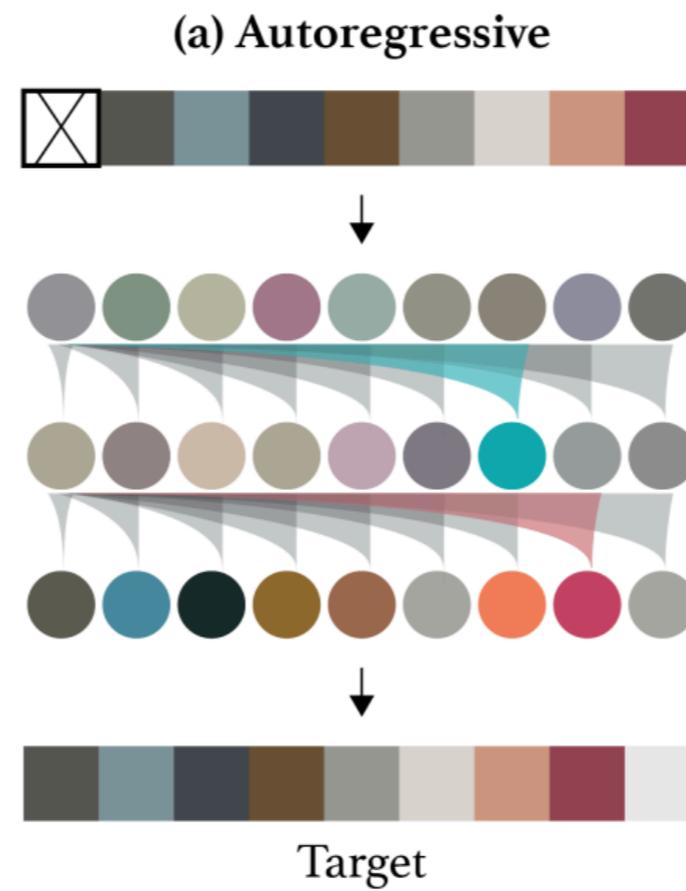
Self-attention on pixels



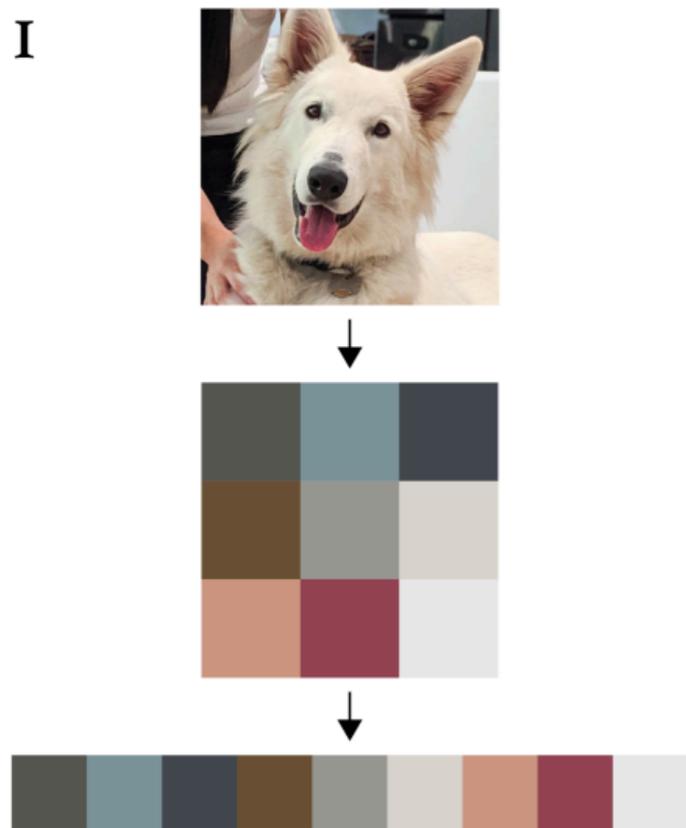
Self-attention on pixels



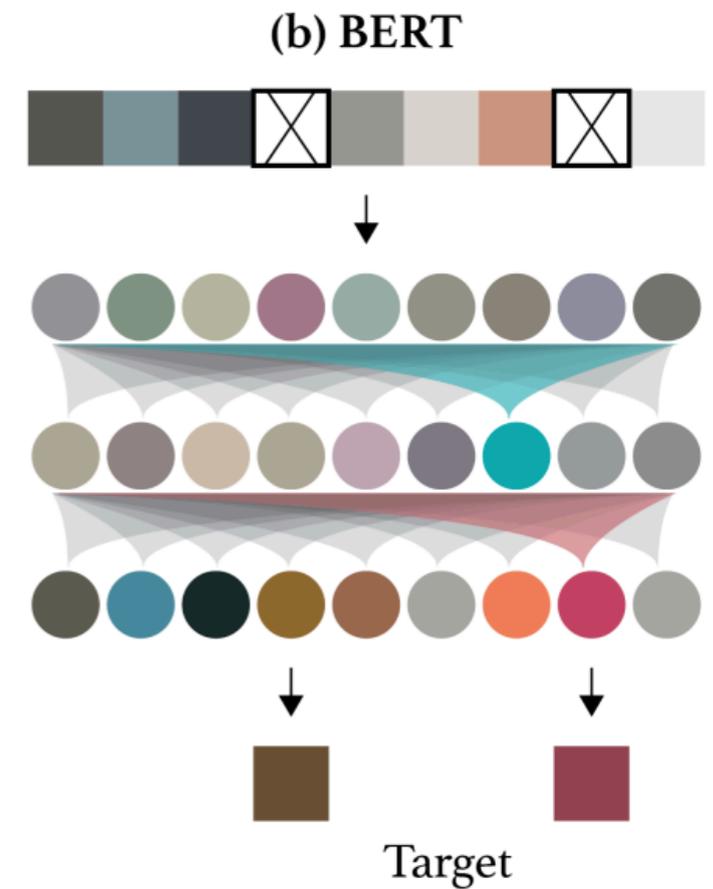
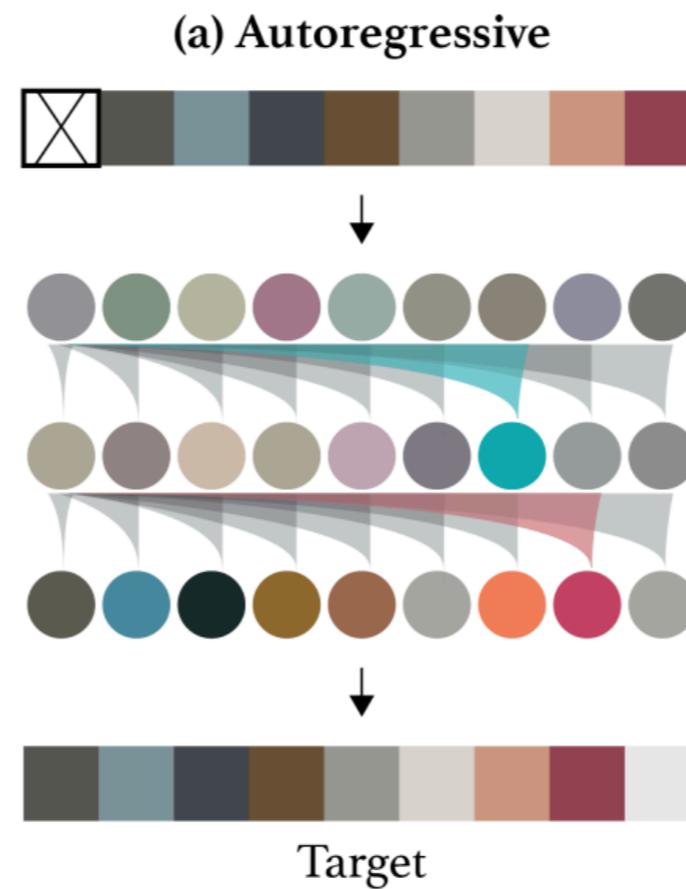
2



Self-attention on pixels



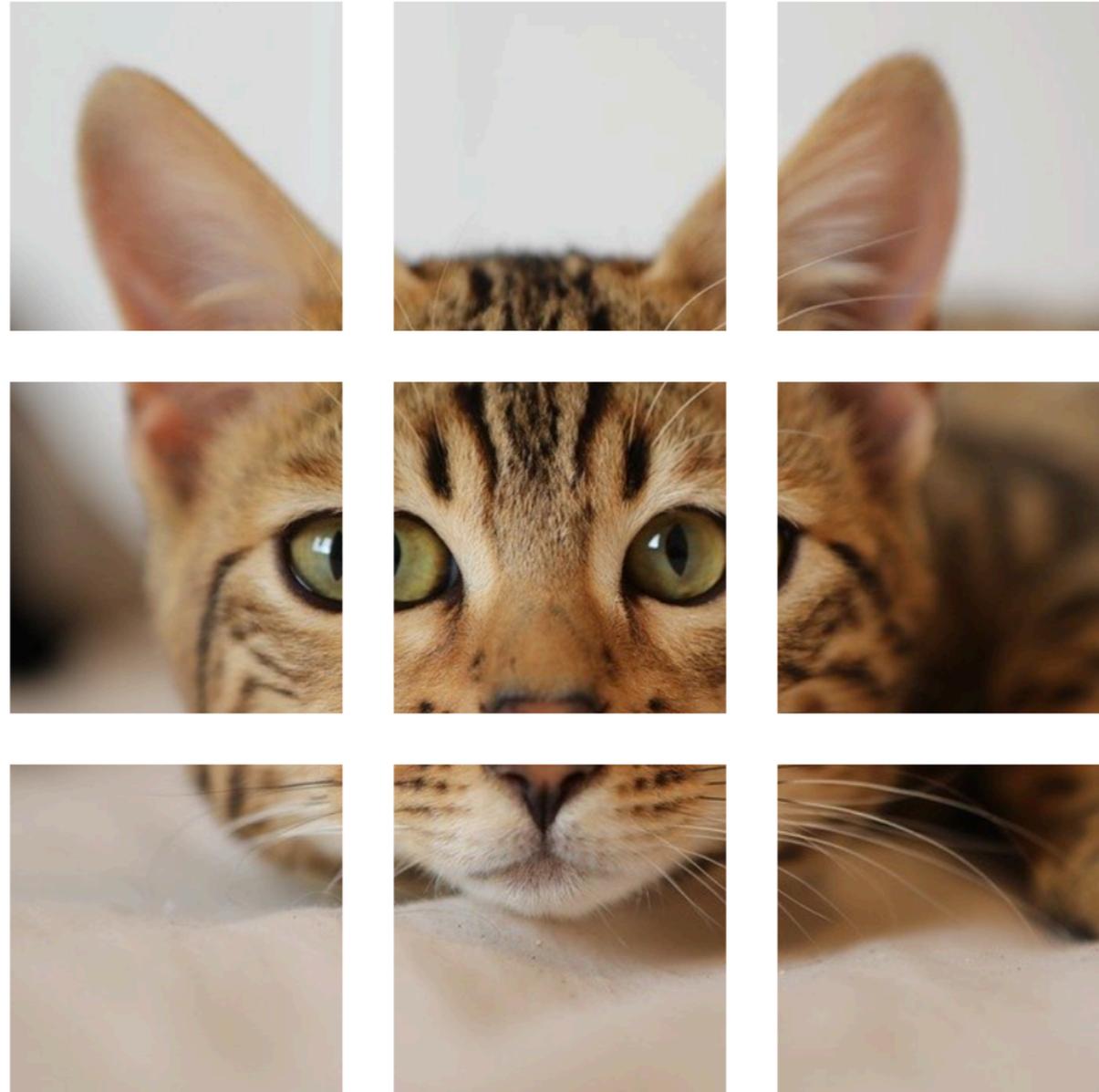
2

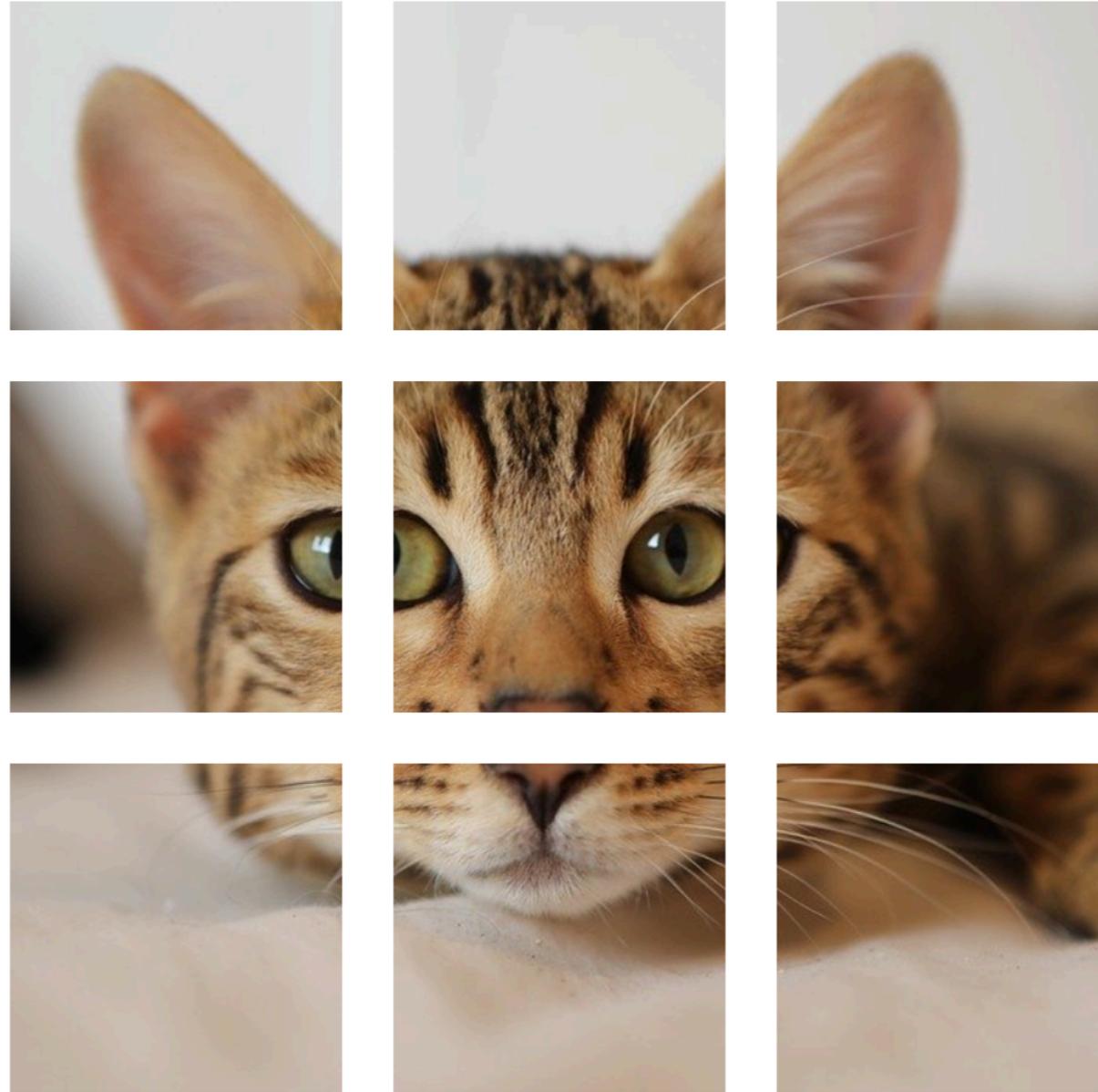


Issues?



An Image is Worth 16x16 words, Dosovitskiy et al., ICLR 2021



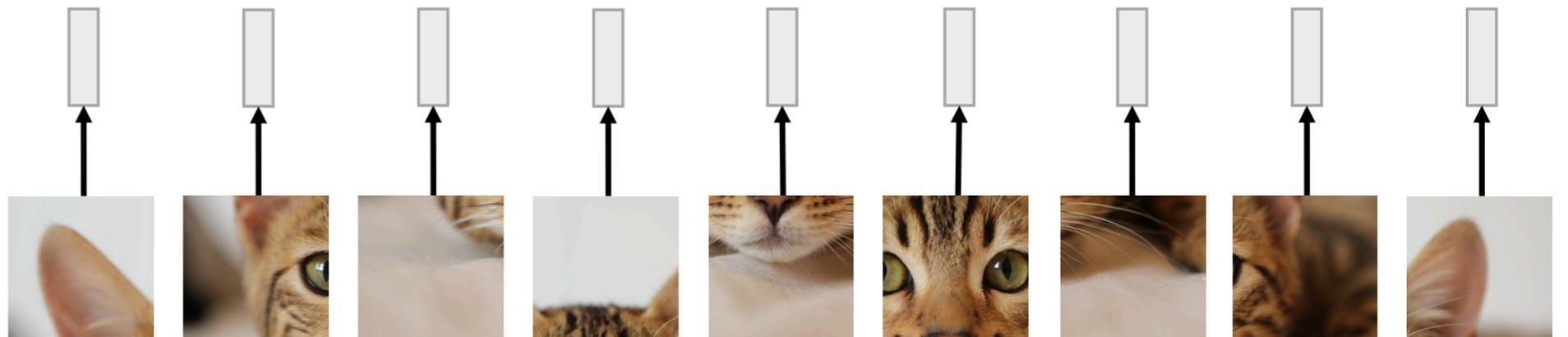


N input patches, each
of shape 3x16x16



Linear projection to
D-dimensional vector

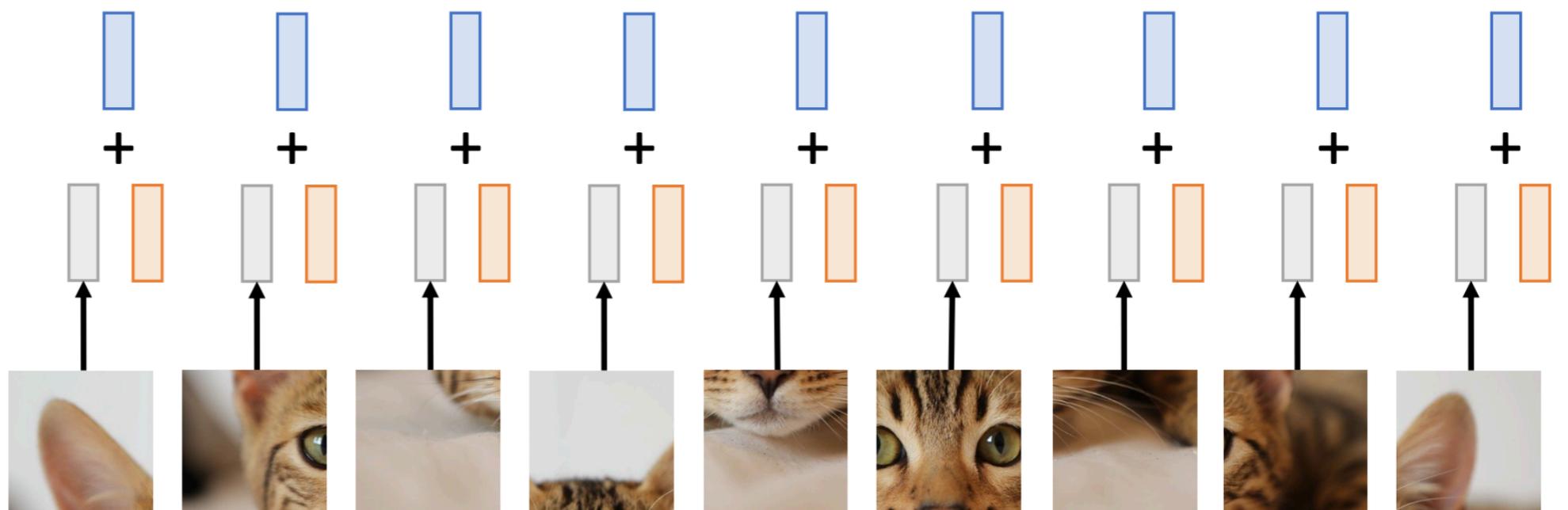
N input patches, each
of shape 3x16x16



Add positional embedding: learned D-dim vector per position

Linear projection to D-dimensional vector

N input patches, each of shape 3x16x16



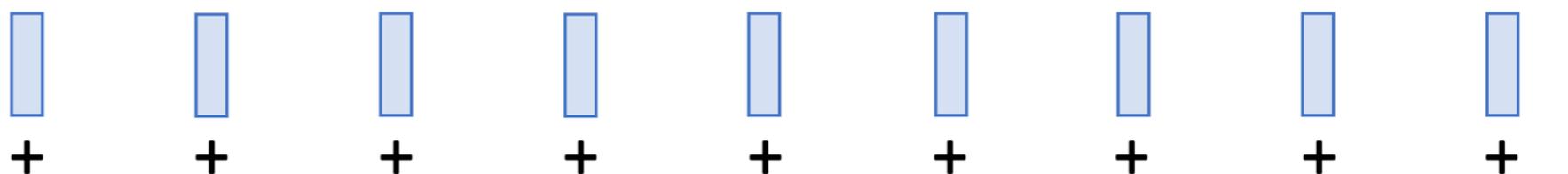
Output vectors



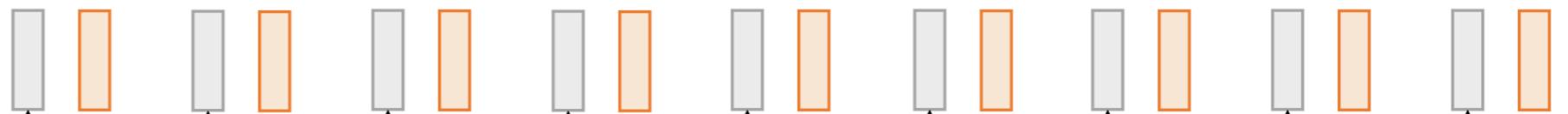
Exact same as
NLP Transformer!



Add positional
embedding: learned D-
dim vector per position

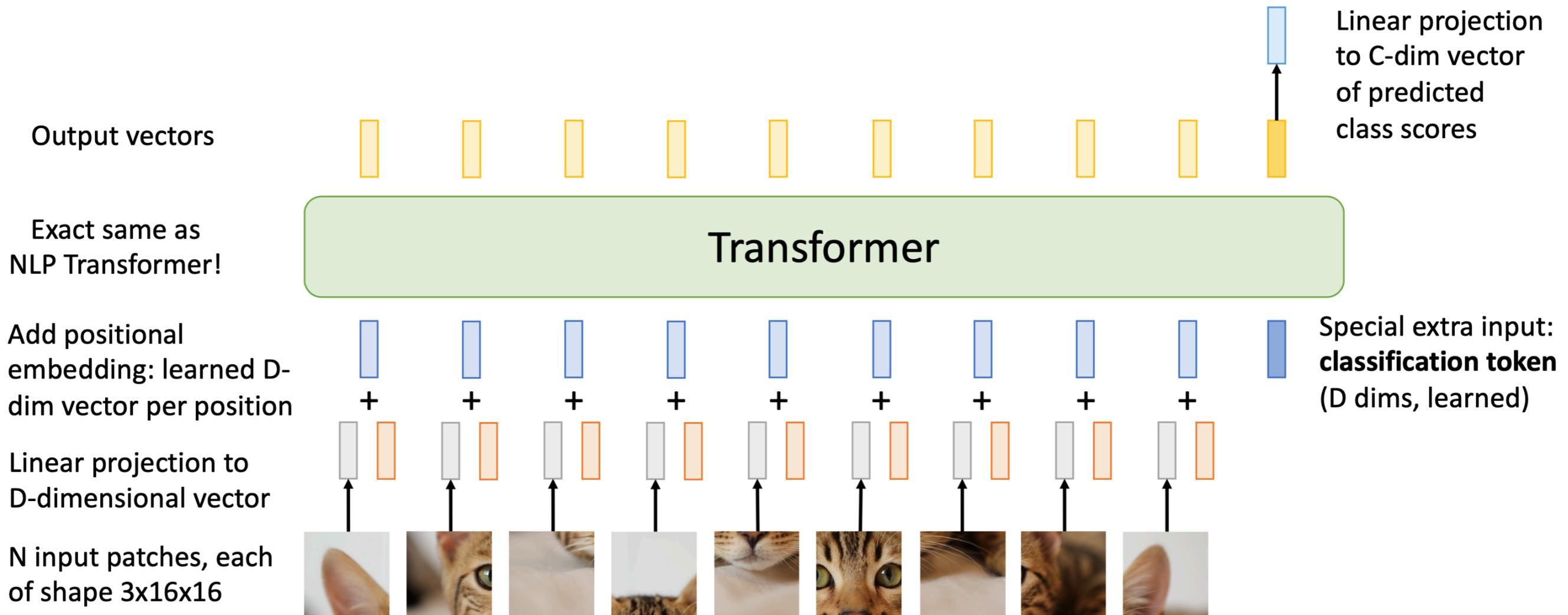


Linear projection to
D-dimensional vector



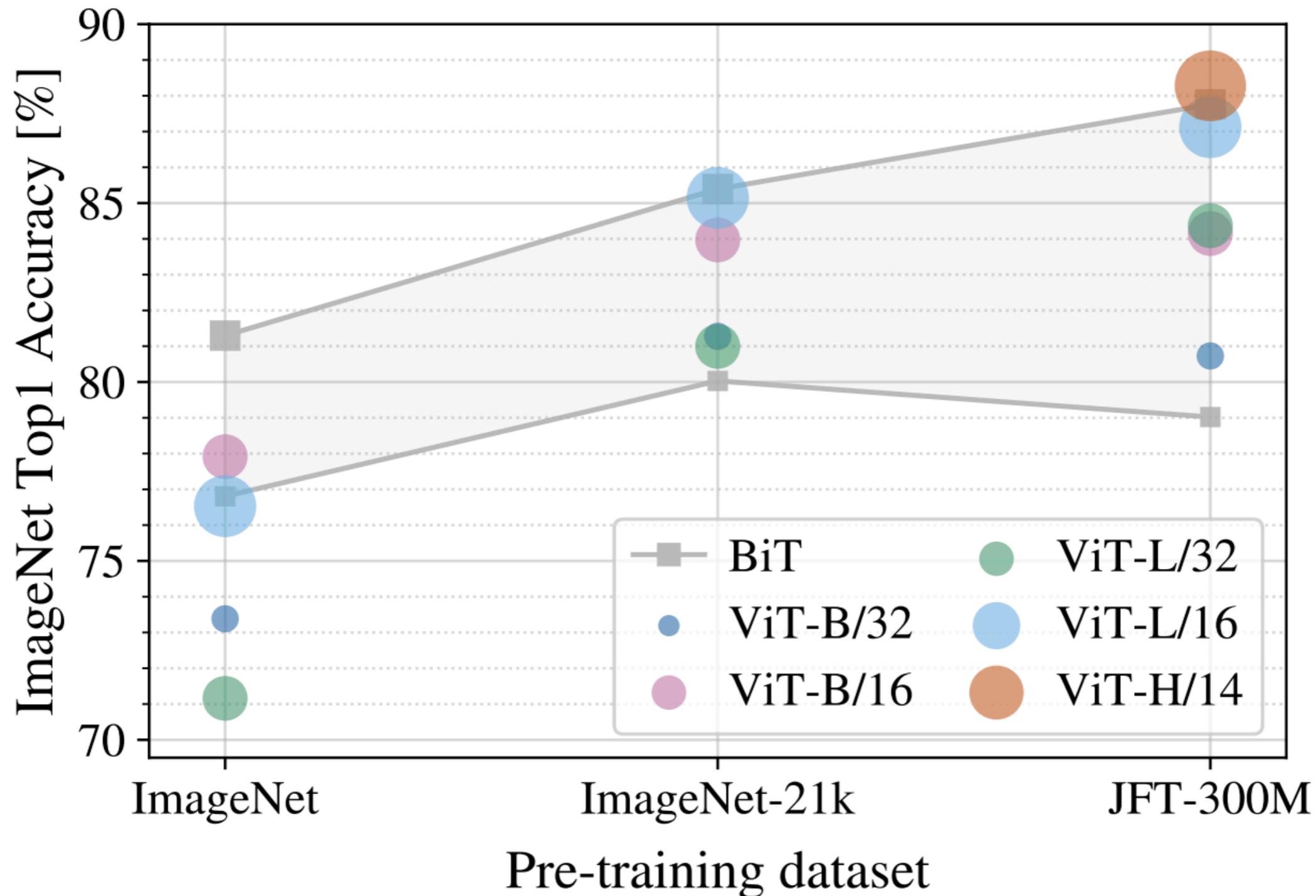
N input patches, each
of shape 3x16x16





16x16 patches = 16^*16^*3 =
768d embedding

Vision Transformers (ViT) outperform ResNets with larger datasets



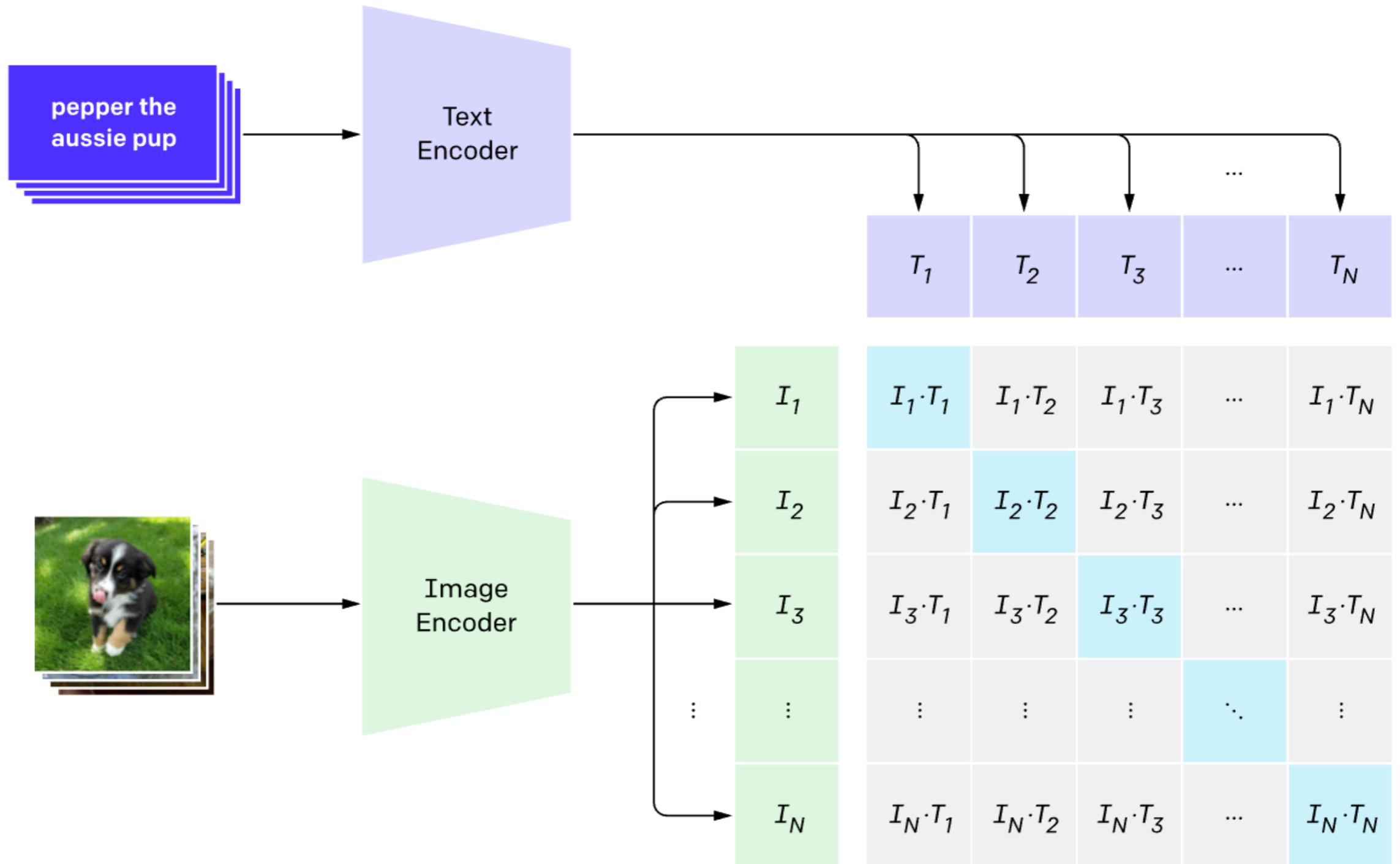
Okay, so we can encode text with Transformers, and we can encode images with Transformers.....

Since the architectures are now basically the same, can we train a single model on both modalities?

OpenAI's CLIP: Contrastive language-image pretraining

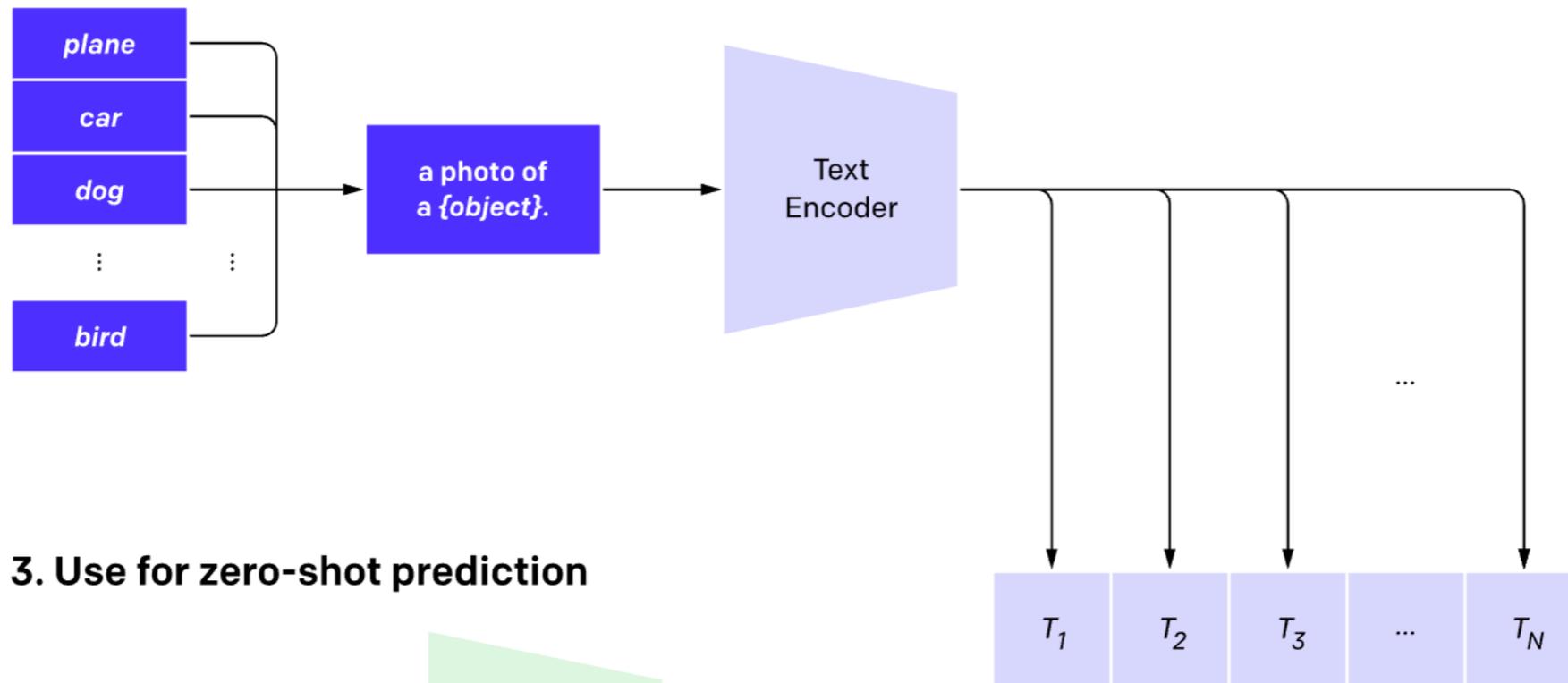
- OpenAI collect 400 million (image, text) pairs from the web
- Then, they train an image encoder and a text encoder with a simple contrastive loss: given a collection of images and text, predict which (image, text) pairs actually occurred in the dataset

1. Contrastive pre-training

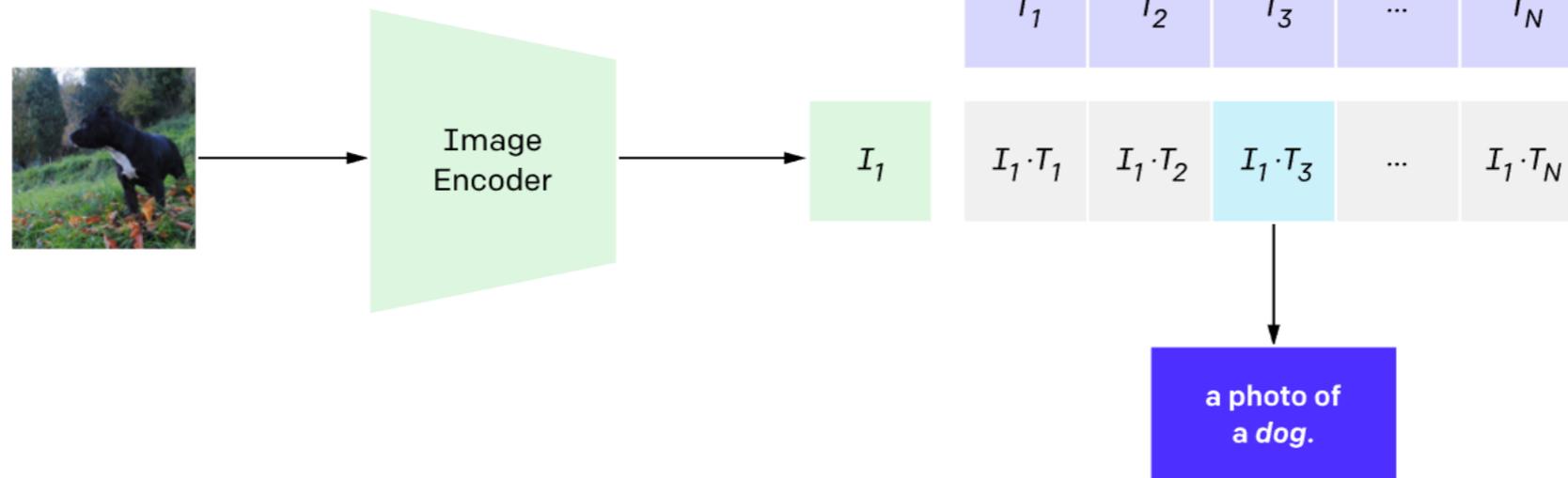


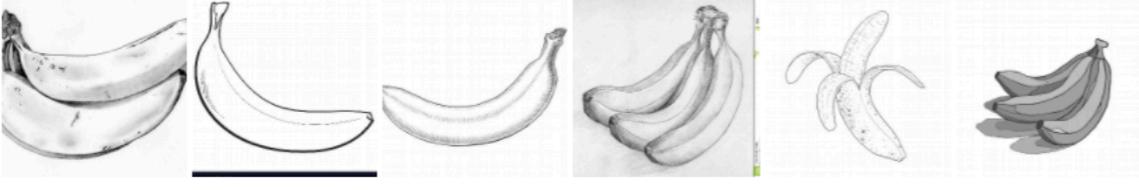
Similar to GPT-3, you can use CLIP for zero-shot learning

2. Create dataset classifier from label text

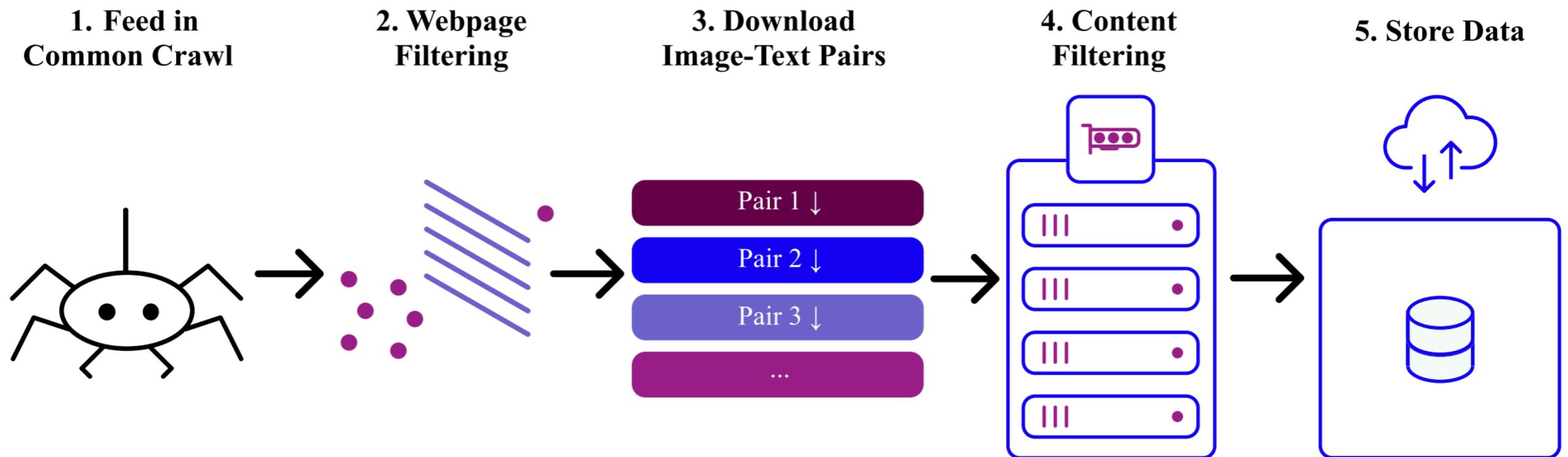


3. Use for zero-shot prediction



DATASET	IMAGENET RESNET101	CLIP VIT-L
 <p>ImageNet</p>	<p>76.2%</p>	<p>76.2%</p>
 <p>ImageNet V2</p>	<p>64.3%</p>	<p>70.1%</p>
 <p>ImageNet Rendition</p>	<p>37.7%</p>	<p>88.9%</p>
 <p>ObjectNet</p>	<p>32.6%</p>	<p>72.3%</p>
 <p>ImageNet Sketch</p>	<p>25.2%</p>	<p>60.2%</p>
 <p>ImageNet Adversarial</p>	<p>2.7%</p>	<p>77.1%</p>

LAION-5B: a dataset of 5 billion image/text pairs!



Major copyright issues...

Stable Diffusion and other image-generating AI products could not exist without the work of painters, illustrators, photographers, sculptors, and other artists. Stable Diffusion was trained on the [LAION-5B](#) dataset. LAION-5B contains 5.85 billion image-text pairs. Most of the images contained in the dataset are copyrighted, and LAION claims no ownership in them. As it notes, “The images are under their copyright.”

On January 13, 2023, the Joseph Saveri Law Firm, LLP filed a complaint in the U.S. District Court for the Northern District of California on behalf of Sarah Andersen, Kelly McKernan, Karla Ortiz, and a class of other artists and stakeholders against Stability AI Ltd.; Stability AI, Inc.; DeviantArt, Inc.; and Midjourney, Inc. This suit alleges copyright infringement, DMCA violations, right of publicity violations, breach of the DeviantArt Terms of Service, unfair competition, and unjust enrichment. It likewise seeks damages and injunctive relief to compensate the class for harms already incurred and to prevent future harms.