Deep Learning

Oct 30-31st, 2025

EIT Oxford

Lecture II – Computer Vision

EIT Oxford

What can we solve with computer vision

- **Instance classification / regression**
- **Representation learning**
- **Semantic segmentation**
- Object detection
- Generation
- Super-resolution
- Inpainting
- Image + time (= **video)**
- Object tracking
- Summarization
- …

EIT Oxford

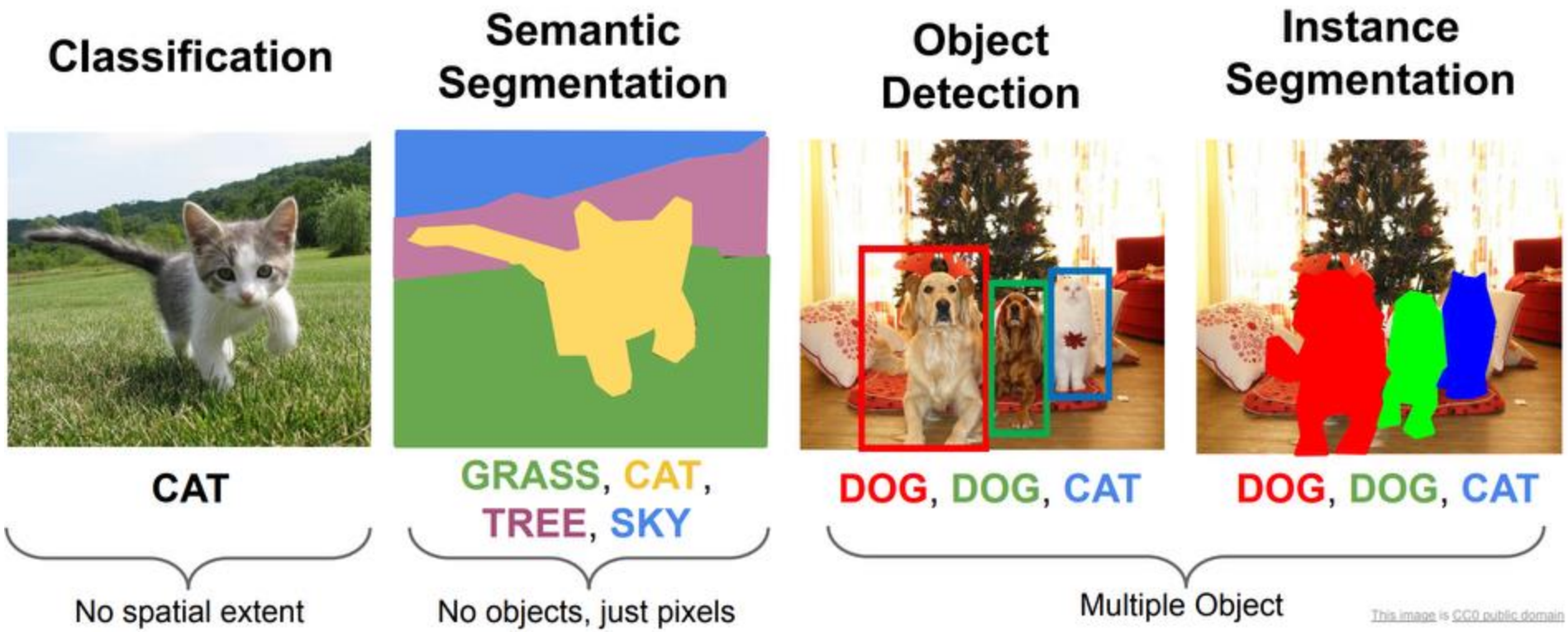
What is covered in other lectures

- **Fundamentals of Deep Learning – Day I**
  - Optimizers
  - Regularization
  - Training stability
  - Practical considerations

- **Deep generative modelling**

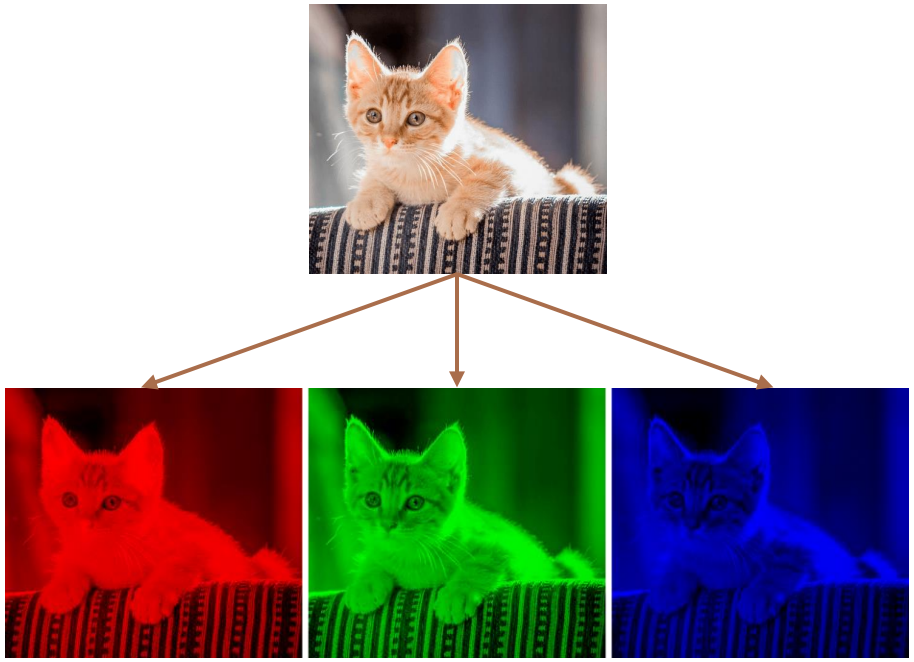
EIT Oxford

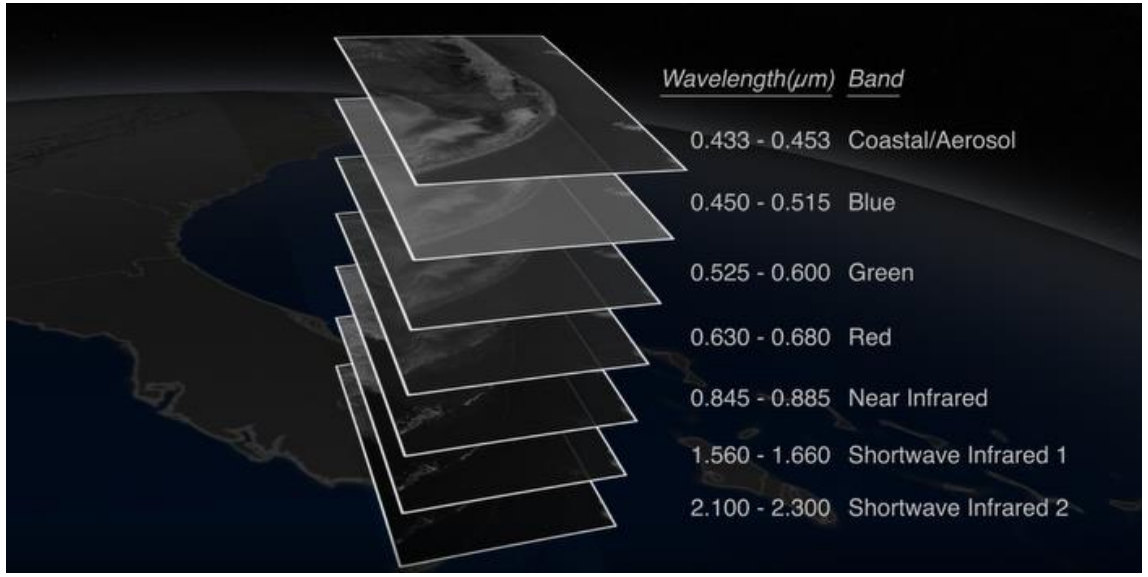
Tasks

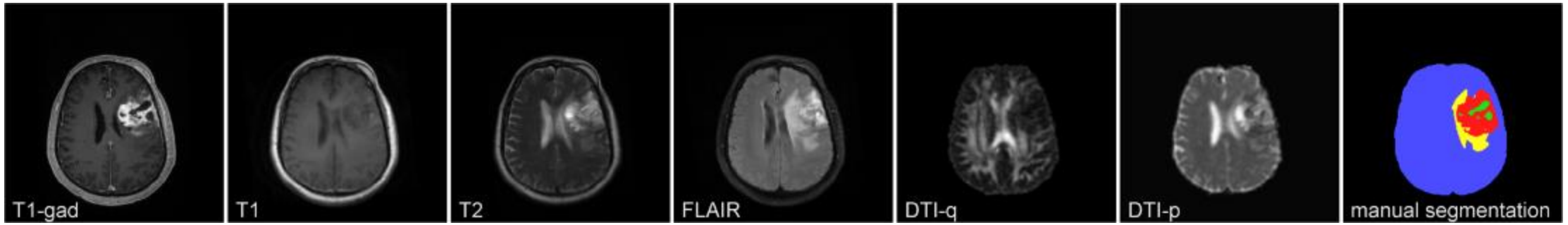


Source: http://cs231n.stanford.edu/slides/2020/lecture_12.pdf

Input channels vary depending on the task

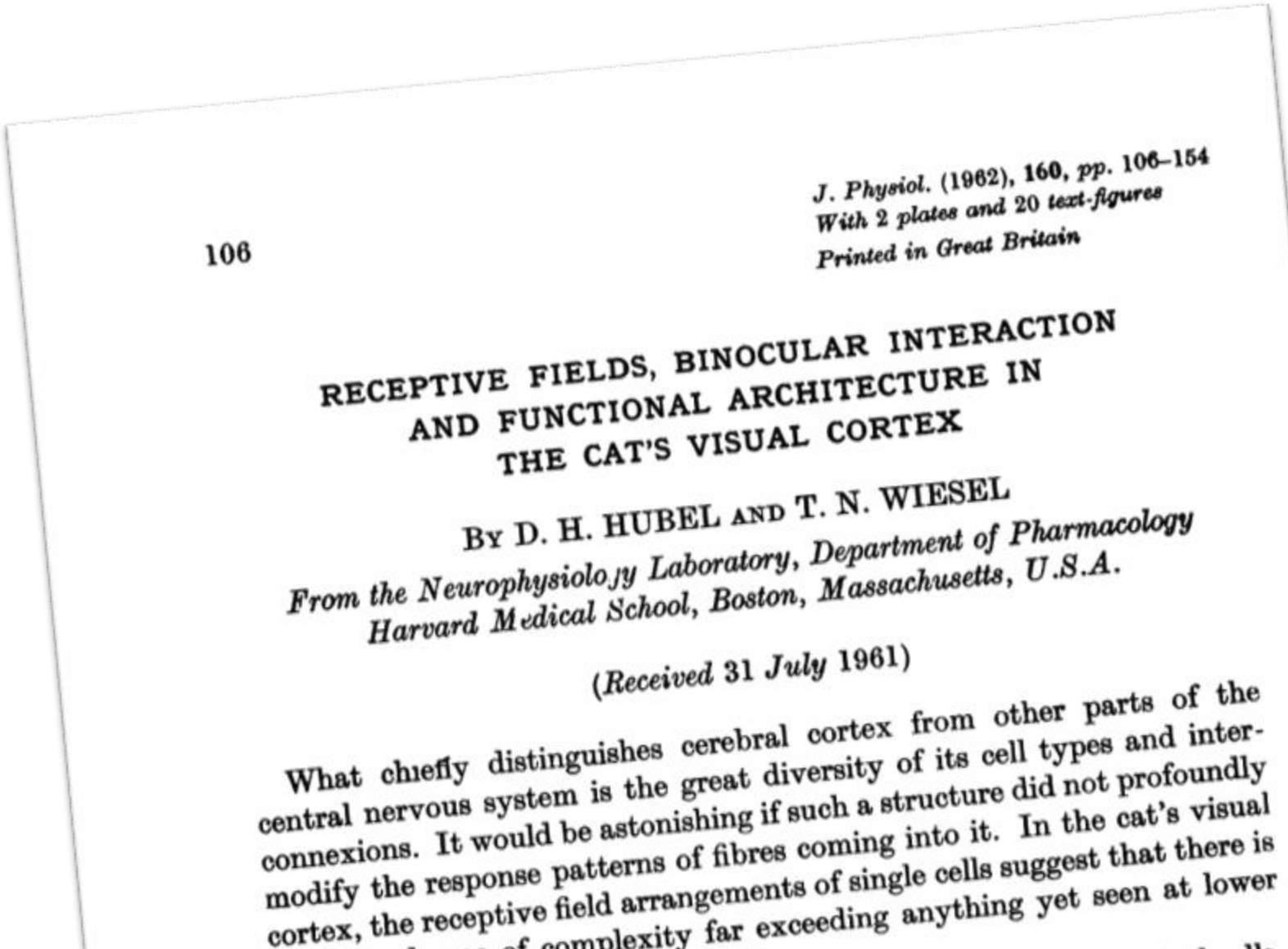






EIT Oxford

Intuitions from the visual system



106

J. Physiol. (1962), **160**, *pp.* 106–154
With 2 plates and 20 text-figures
Printed in Great Britain

RECEPTIVE FIELDS, BINOCULAR INTERACTION AND FUNCTIONAL ARCHITECTURE IN THE CAT'S VISUAL CORTEX

BY D. H. HUBEL AND T. N. WIESEL

From the Neurophysiology Laboratory, Department of Pharmacology Harvard Medical School, Boston, Massachusetts, U.S.A.

(*Received* 31 *July* 1961)

What chiefly distinguishes cerebral cortex from other parts of the central nervous system is the great diversity of its cell types and interconnexions. It would be astonishing if such a structure did not profoundly modify the response patterns of fibres coming into it. In the cat's visual cortex, the receptive field arrangements of single cells suggest that there is a considerable degree of complexity far exceeding anything yet seen at lower

Neural networks are inspired by the visual system

EIT Oxford

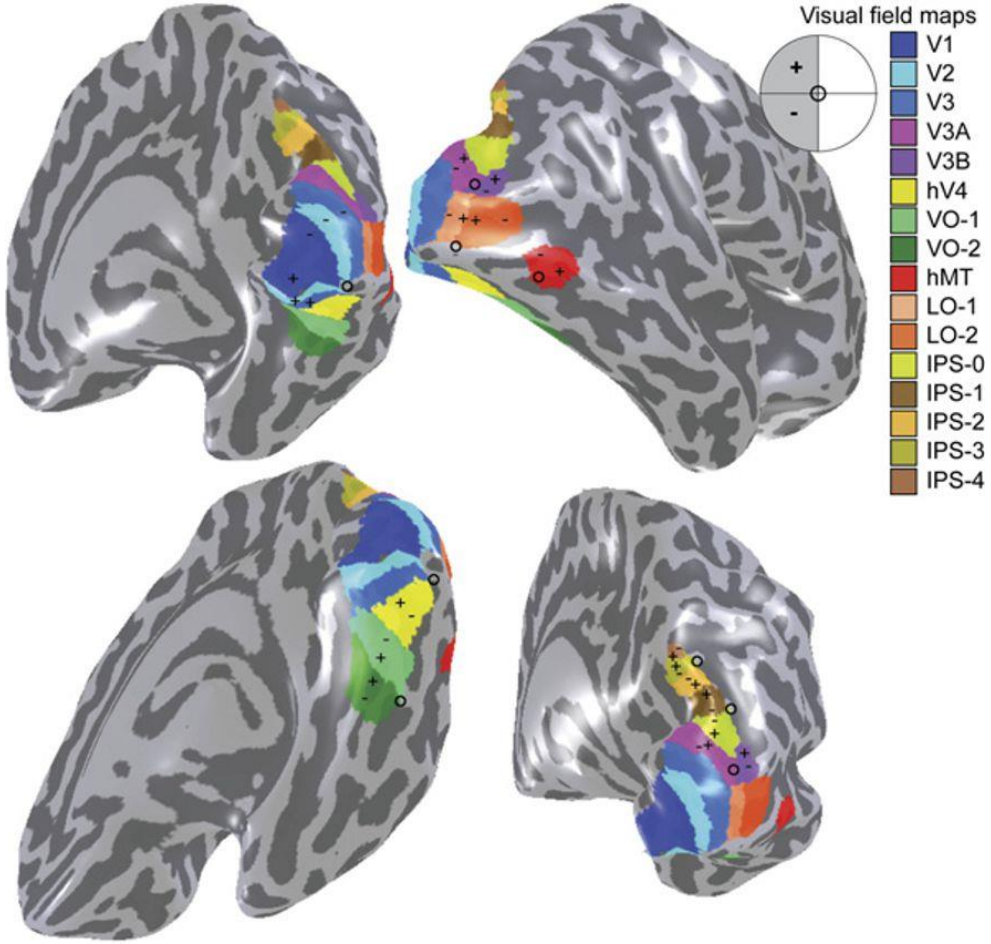
Neural networks are inspired by the visual system

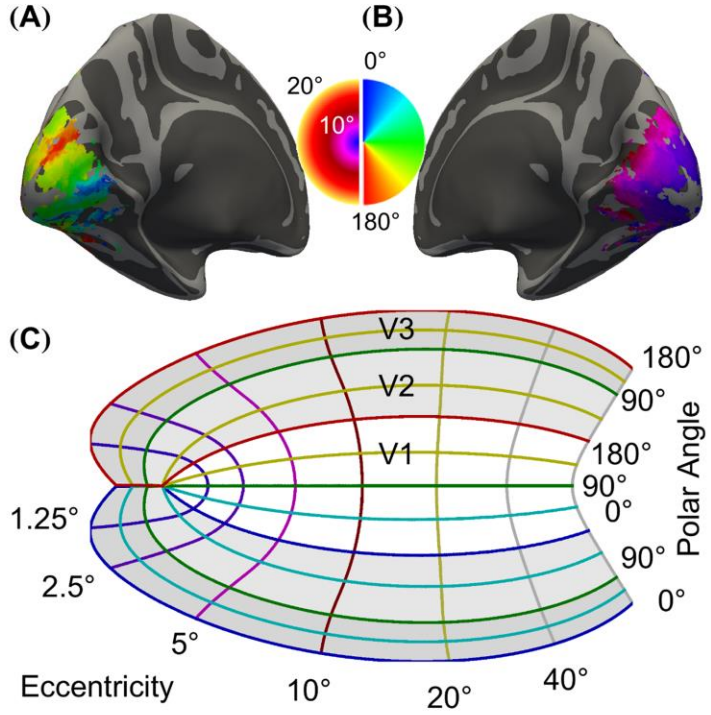
- Measured the electrical response in a cat's brain by stimulating it with simple patterns on a screen.

- Found that neurons in the early visual cortex are organized hierarchically:
  - Early layers include orientation-selective cells responding preferentially to simple patterns like edges and contours
  - Later layers respond to more complex patterns by combining the activations they receive.

- In the proposed model, neurons in the upper layers have a larger receptive field and are less sensitive to the position where the stimulus originates from.

Organization of the human visual cortex

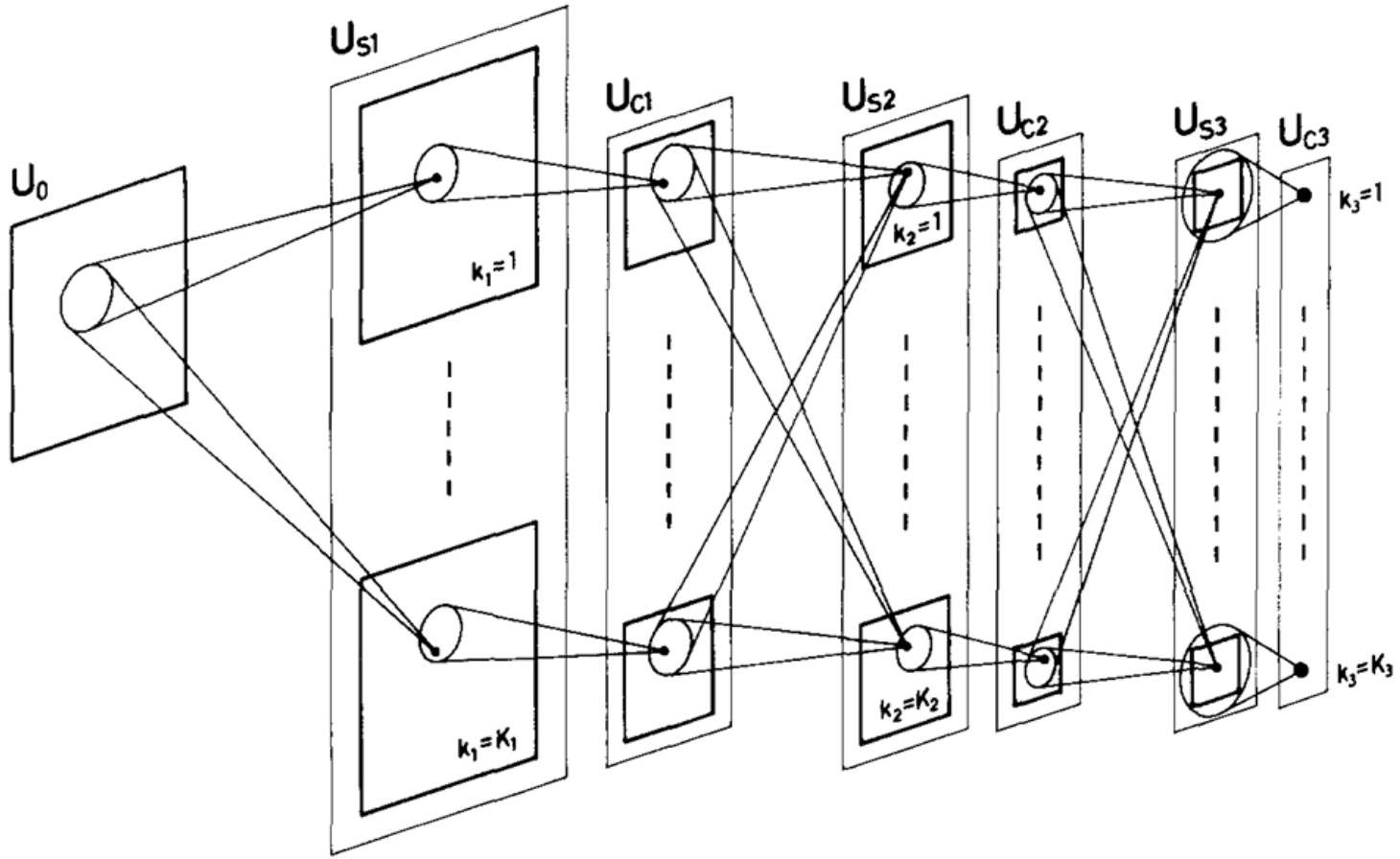




EIT Oxford

Neocognitron – Fukushima (1980)

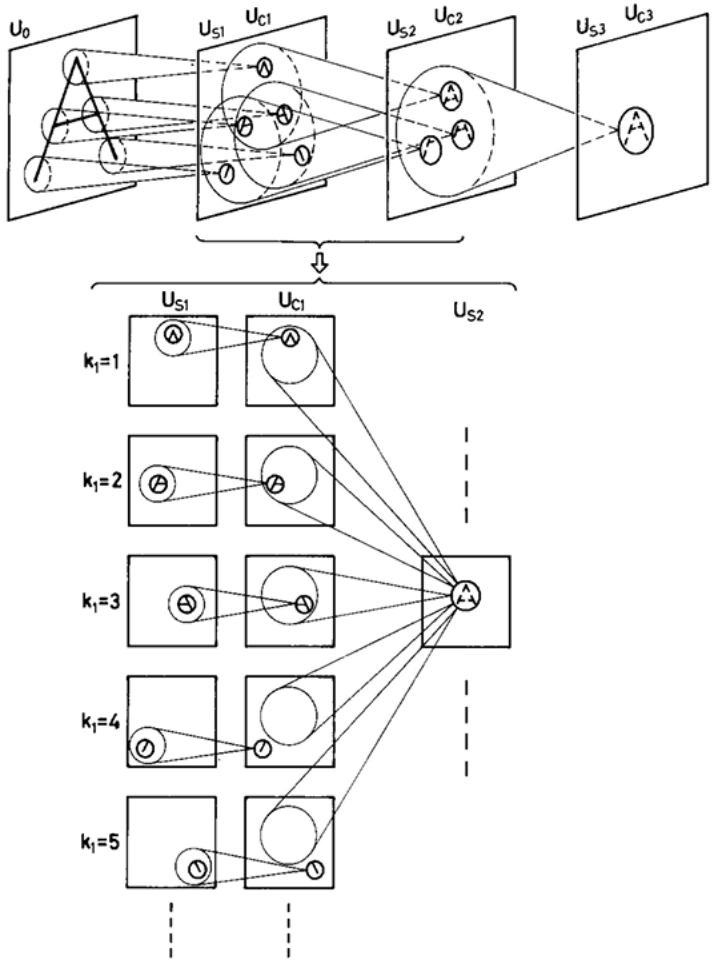
Schematic diagram illustrating the interconnections between layers in the neocognitron



EIT Oxford

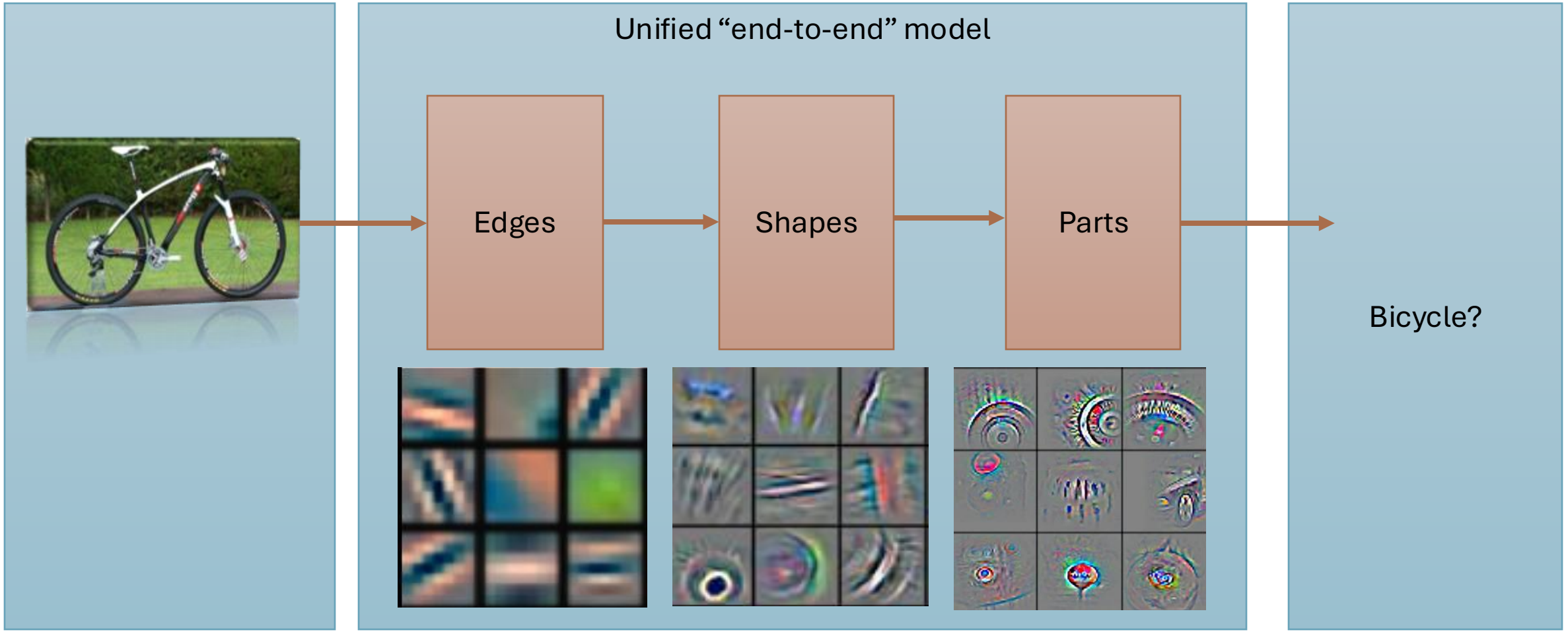
Neocognitron – Fukushima (1980)

Interconnections between cells and the response of the cells after completion of self-organisation



EIT Oxford

Multiple levels of abstraction



EIT Oxford

Building blocks

- Invariance and equivariance
- Convolutions
- Skip / residual connections
- Pooling layers
- Receptive fields
- Upsampling layers
- 1x1 convolution

Invariance

A function f[x] is invariant to a transformation t[] if:

$$\mathbf{f}\left[\mathbf{t}\left[\mathbf{x}\right]\right] = \mathbf{f}\left[\mathbf{x}\right]$$

i.e., the function output is the same even after the transformation is applied.

EIT Oxford

Example

e.g., Image classification

Image has been translated, but we want our classifier to give the same result

Equivariance

A function f[x] is equivariant to a transformation t[] if:

$$\mathbf{f}[\mathbf{t}[\mathbf{x}]] = \mathbf{f}[\mathbf{x}]$$

i.e., the output is transformed in the same way as the input

EIT Oxford

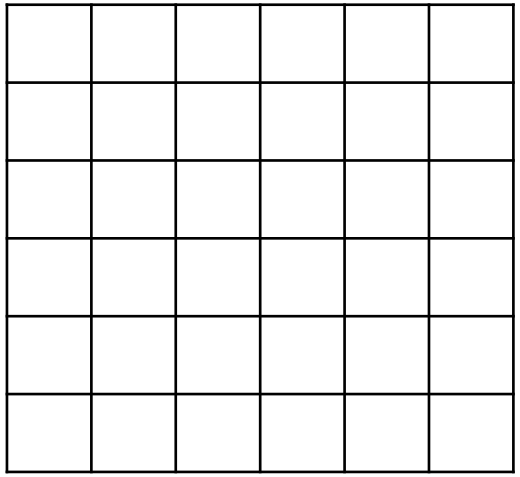
Example

e.g., Image segmentation

Image has been translated and we want segmentation to translate with it

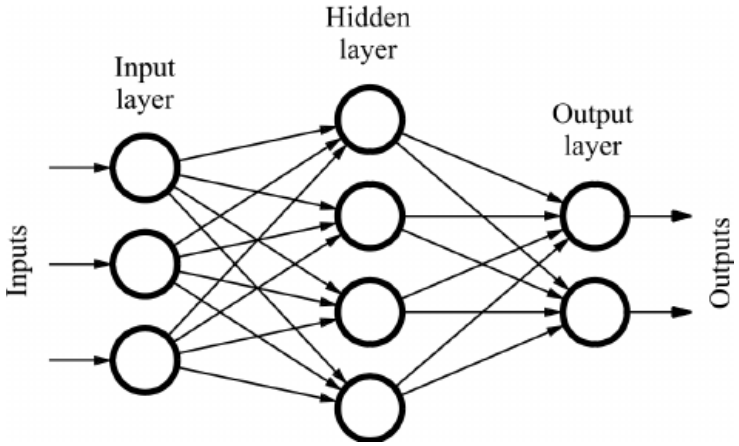
How do we process an image with a FCN?

N X N



Vectorized image or manual feature extraction

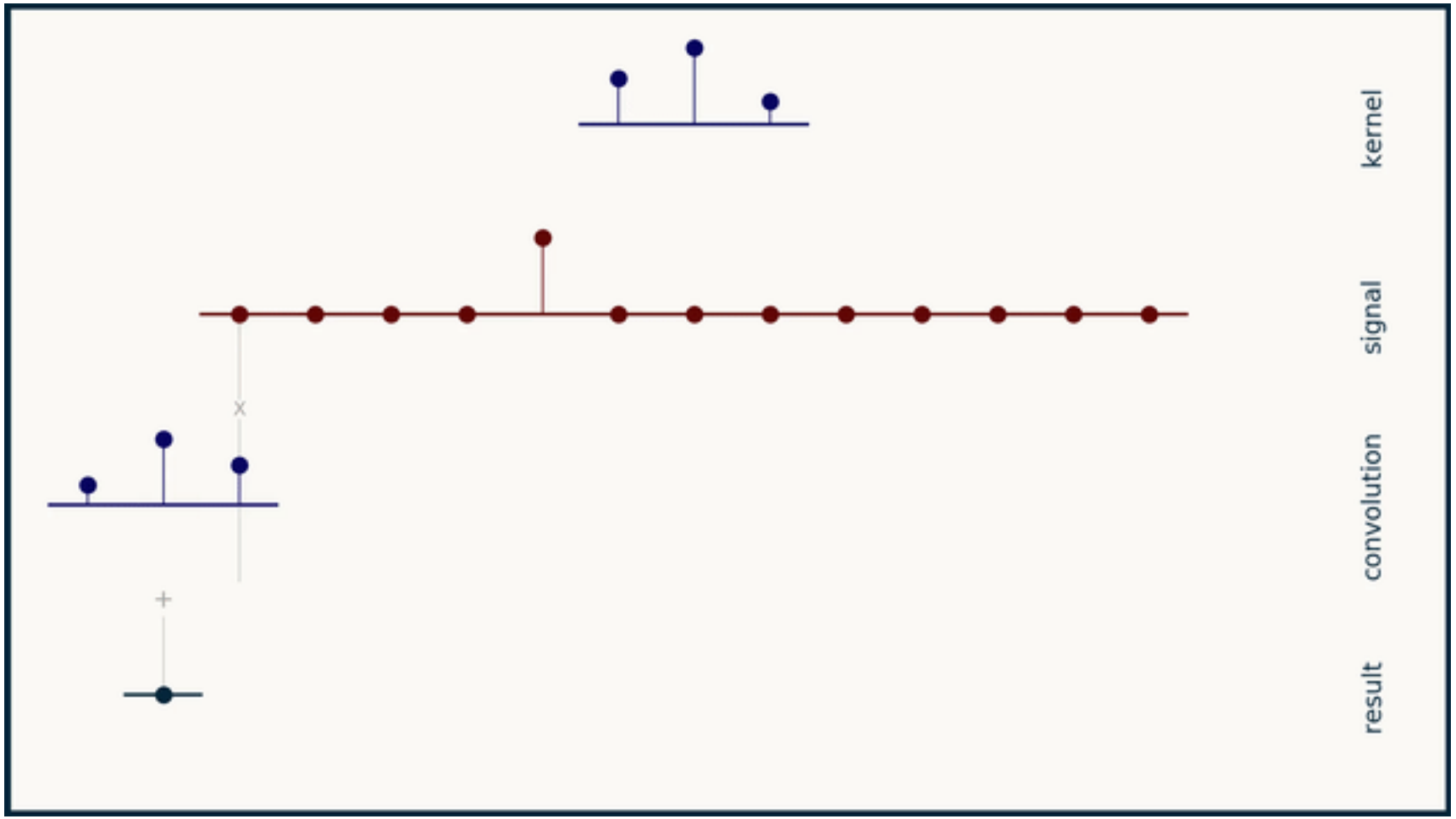
N^2 (img vector) * M (features)



Fully connected layers

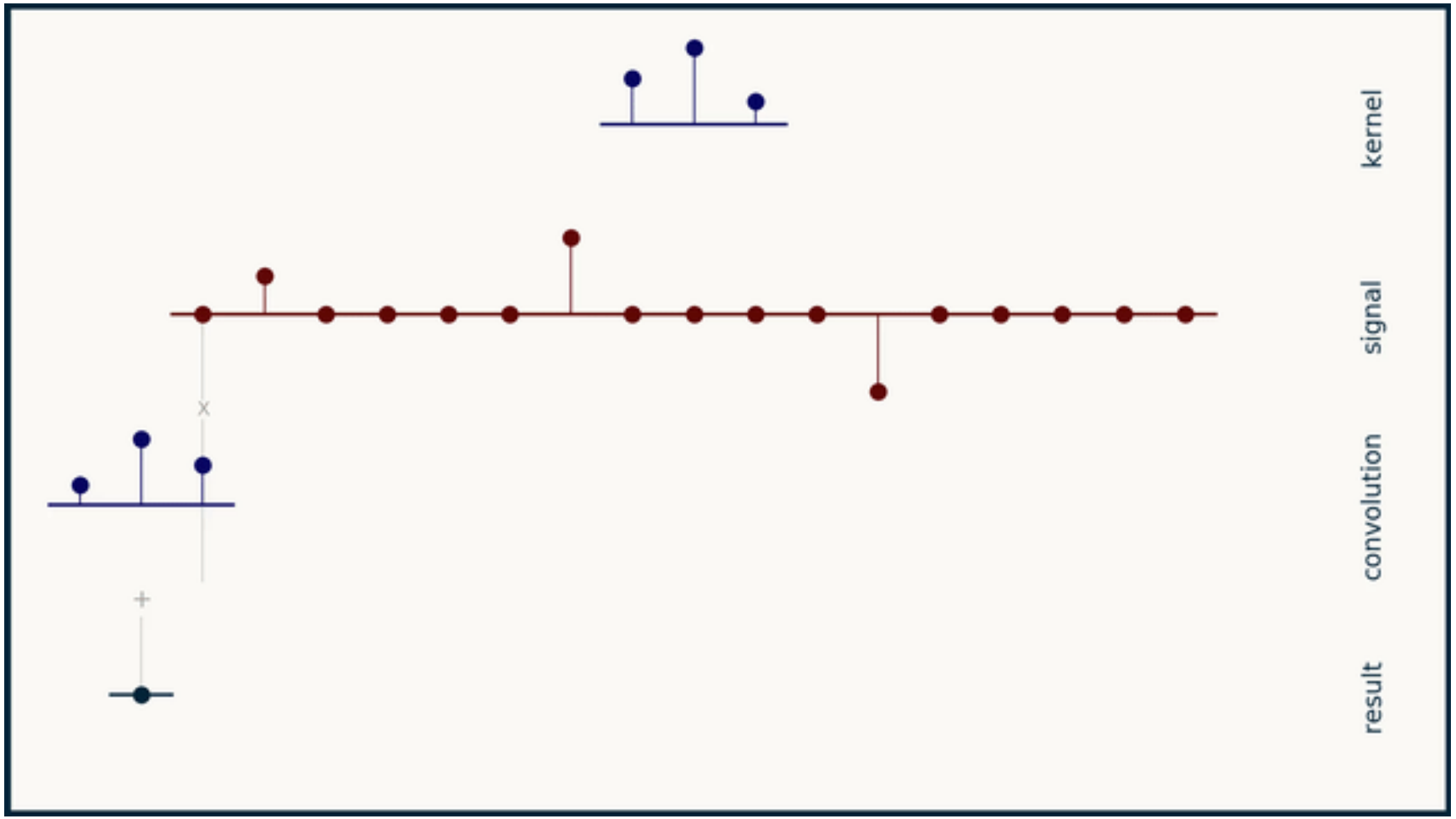
EIT Oxford

Convolutions in 1D



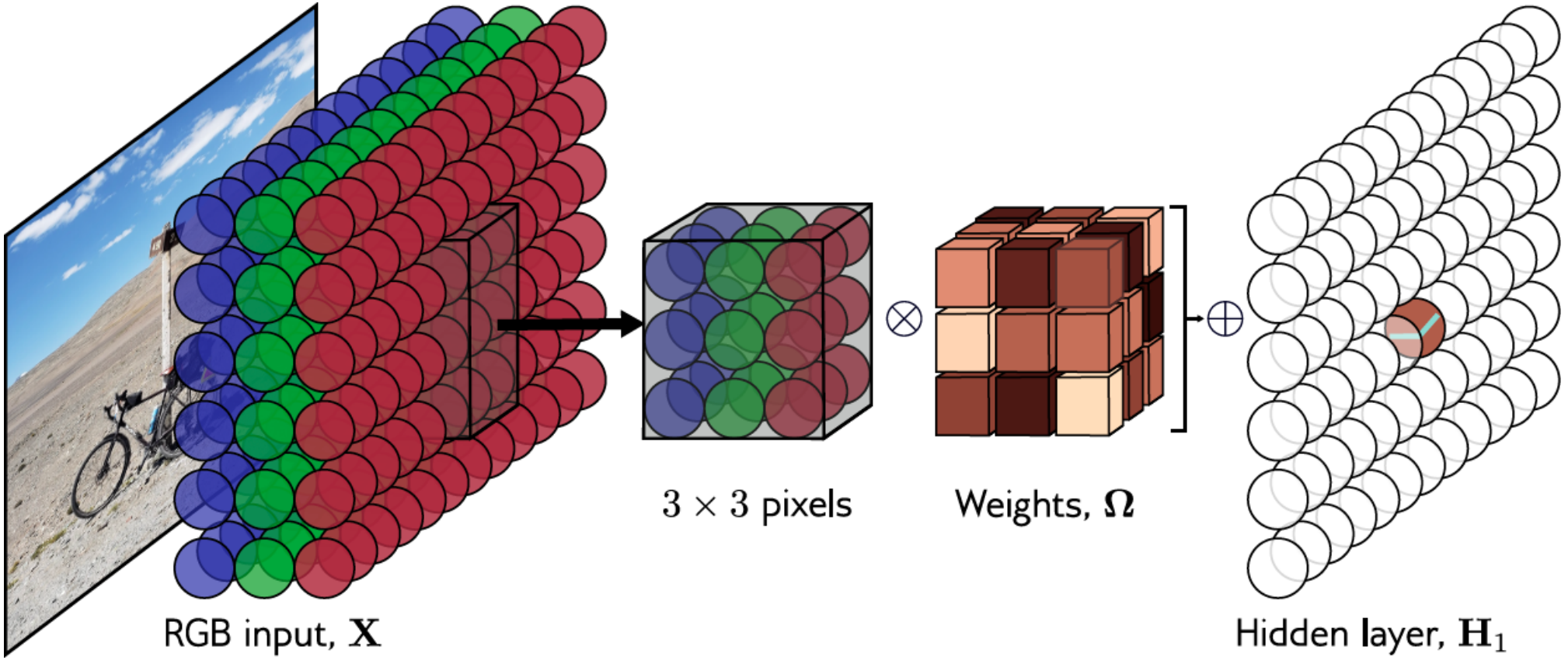
https://brandonrohrer.com/convolution_one_d.html

Convolutions in 1D



https://brandonrohrer.com/convolution_one_d.html

CNNs


RGB input, $\mathbf{X}$
3×3 pixels
Weights, $\boldsymbol{\Omega}$
Hidden layer, $\mathbf{H}_1$

Stride, kernel size, and dilation

- Stride = shift by k positions for each output
  - Decreases size of output relative to input

- Kernel size = weight a different number of inputs for each output
  - Combine information from a larger area
  - But kernel size 5 uses 5 parameters

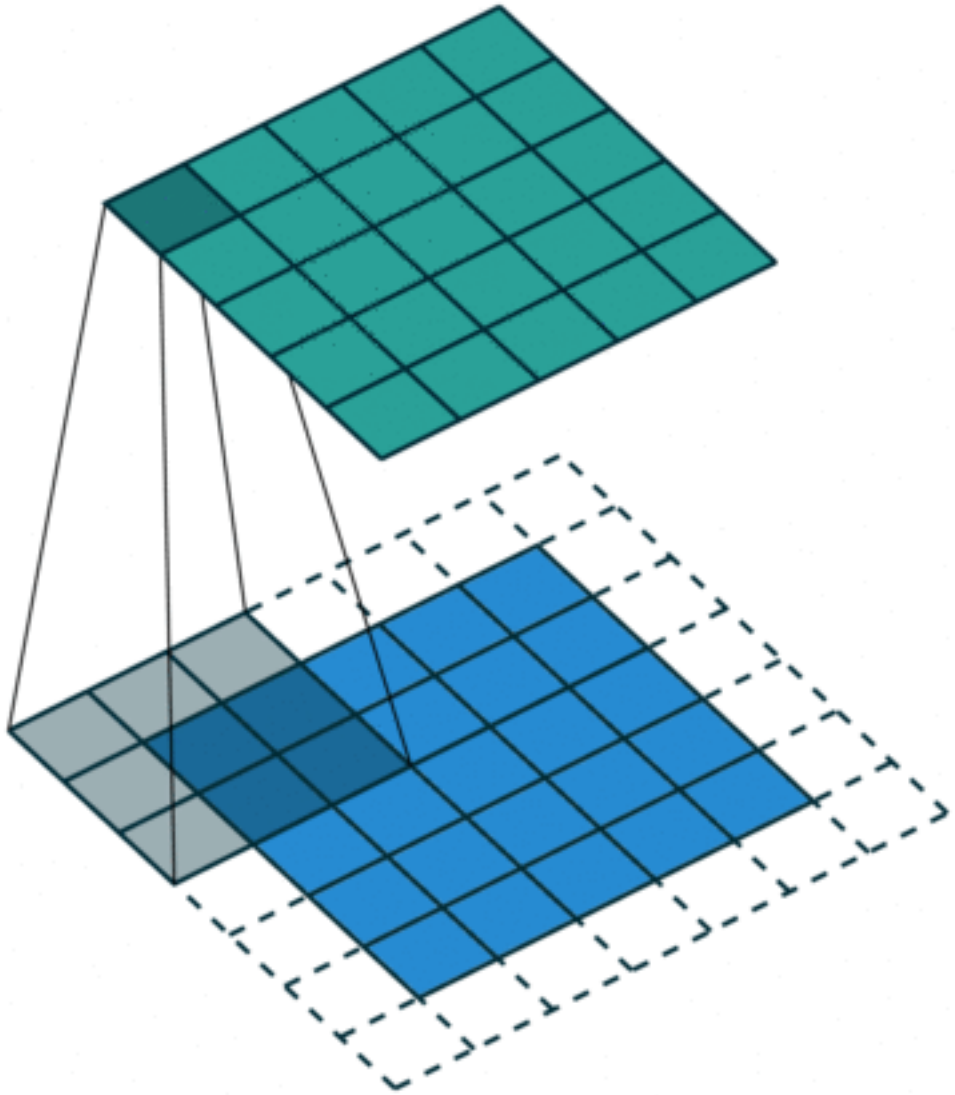
- Dilated convolutions = intersperse kernel values with zeros
  - Combine information from a larger area
  - Fewer parameters

Stride, kernel size, and dilation

Stride = 1

Kernel size = 3

Dilation = 1



http://deeplearning.net/software/theano/tutorial/conv_arithmetic.html

2D convolution

$$S(i,j) = (I * K)(i,j) = \sum_m \sum_n I(m,n)K(i-m, j-n)$$

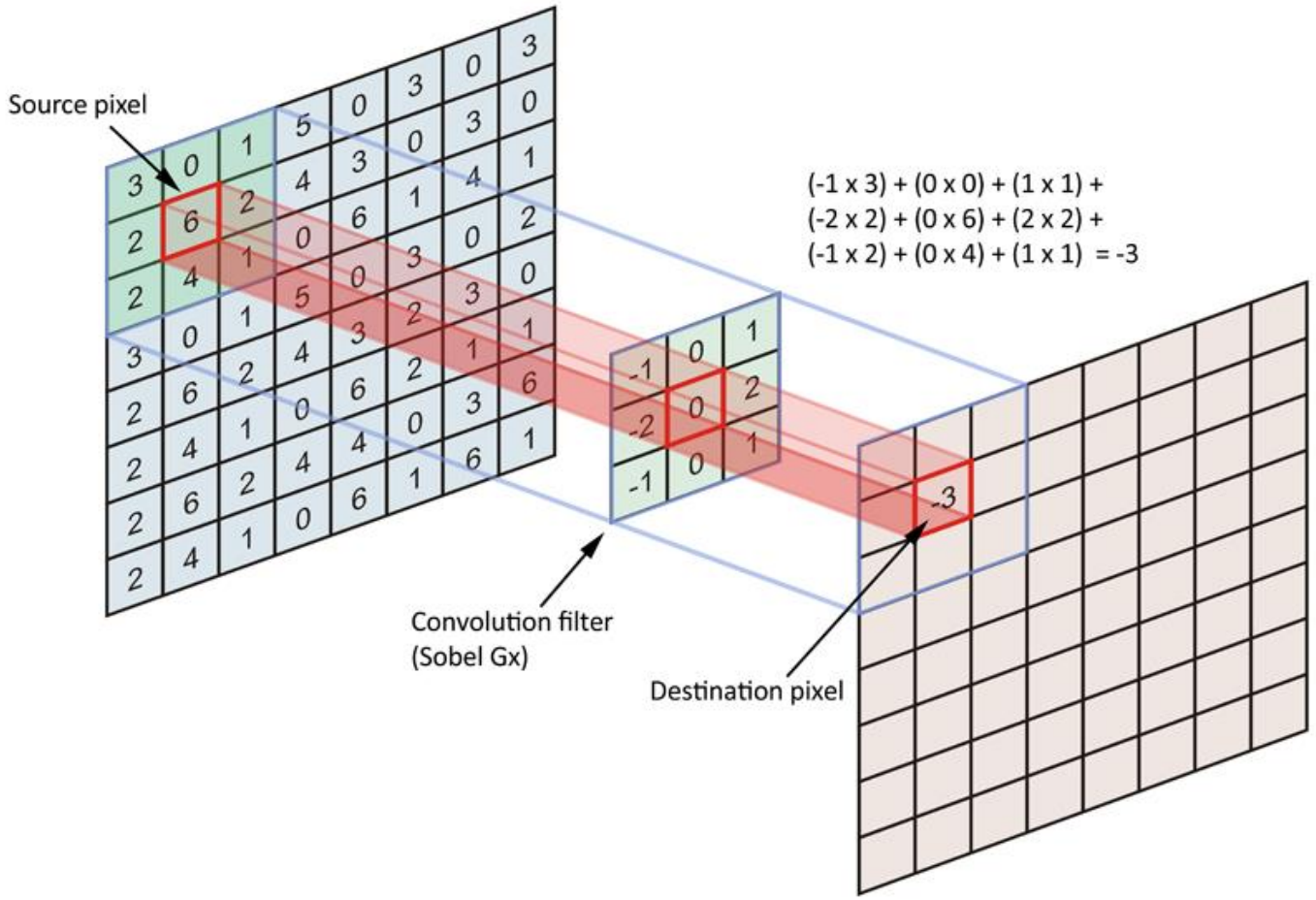
Convolution is commutative

$$S(i,j) = (K * I)(i,j) = \sum_m \sum_n I(i-m, j-n)K(m,n)$$

Useful property since the range of valid values of *m*, *n* for an image is generally larger than that of the kernel.

EIT Oxford

2D convolution

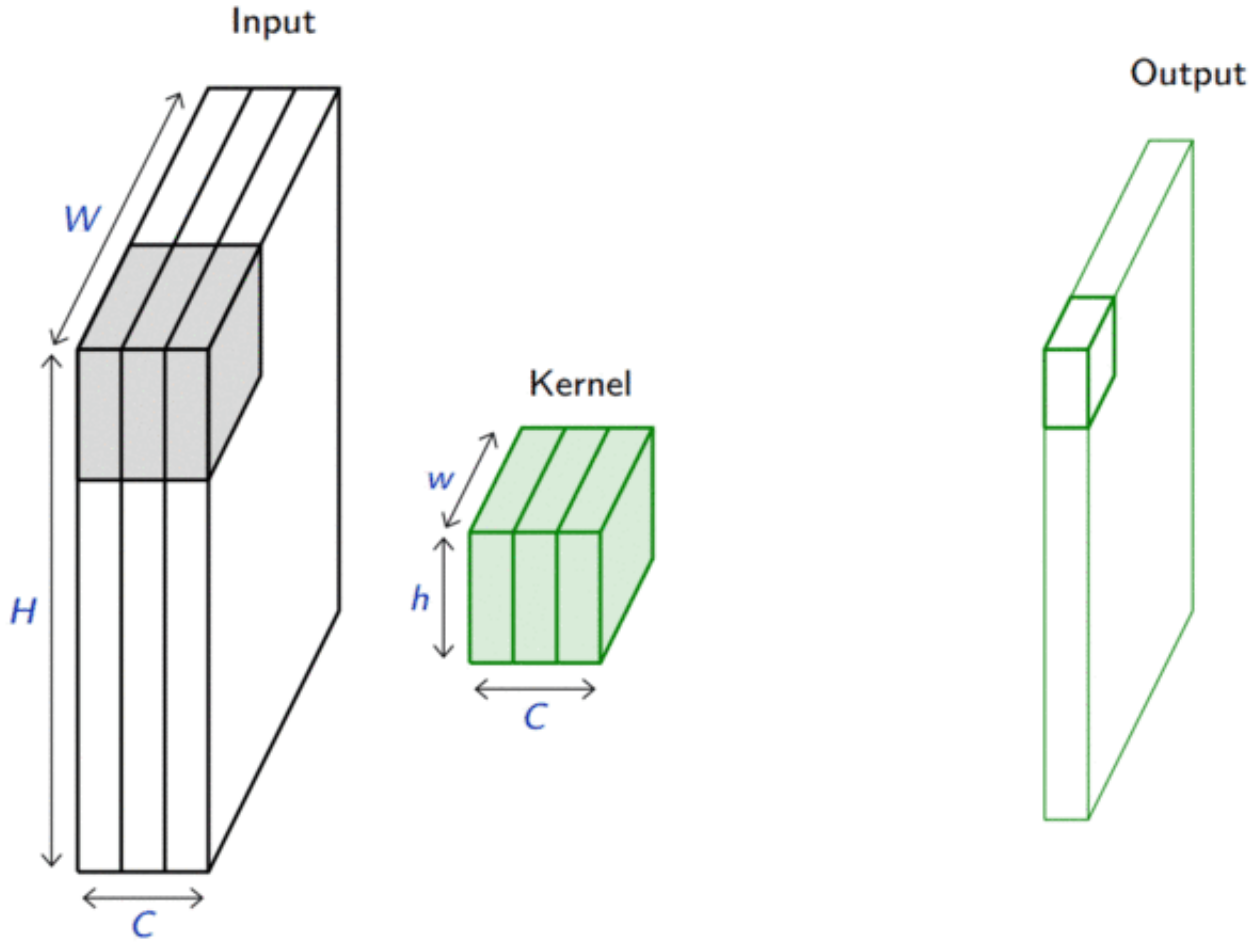


Channels

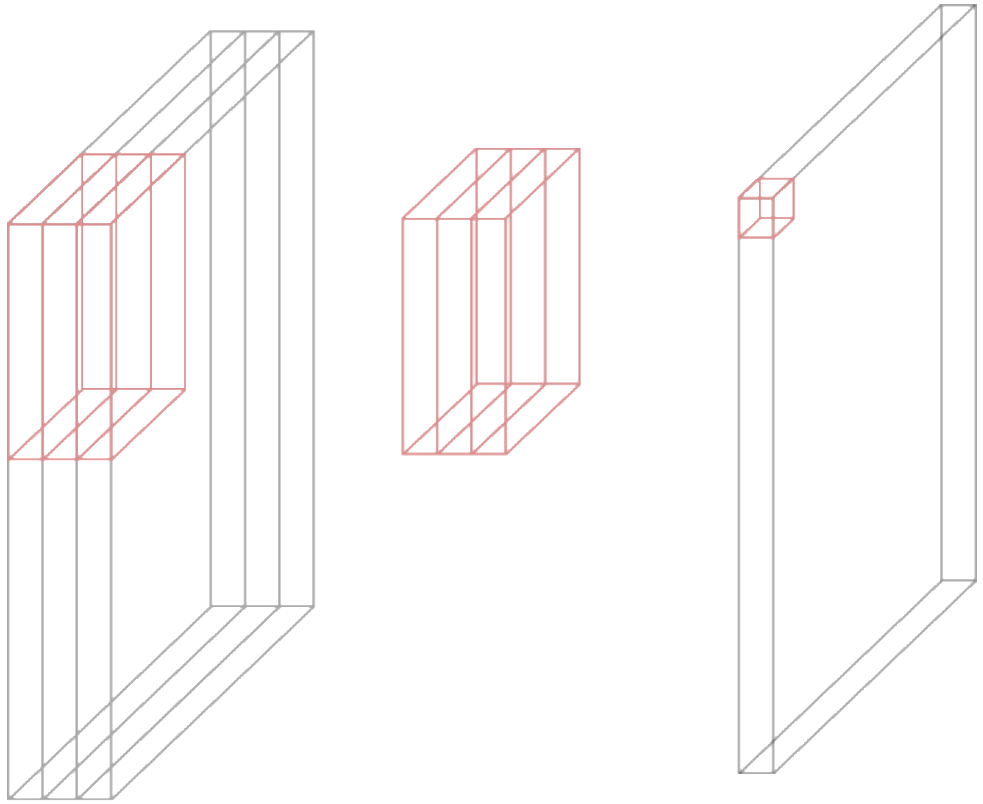
- The convolutional operation averages together the inputs
- Plus passes through an activation function
- Loses information as a result (moving forward)
- Solution:
  - apply several convolutions and stack them in channels
  - Sometimes also called feature maps

2D convolution layer

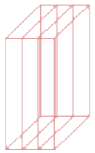
- Local spatial connectivity
- Weight sharing reduces the number of parameters and introduces translation invariance
- Non-linear activation functions are applied to each element individually



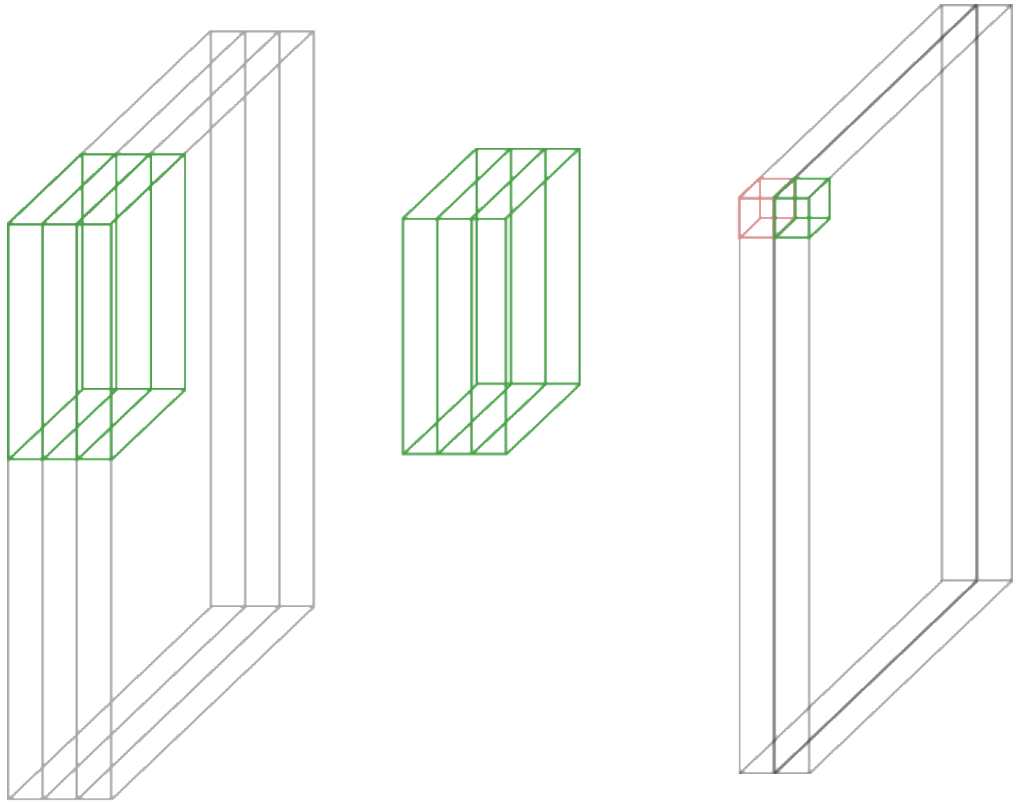
Output of a convolution layer



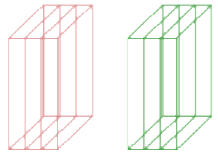
Learnable parameters



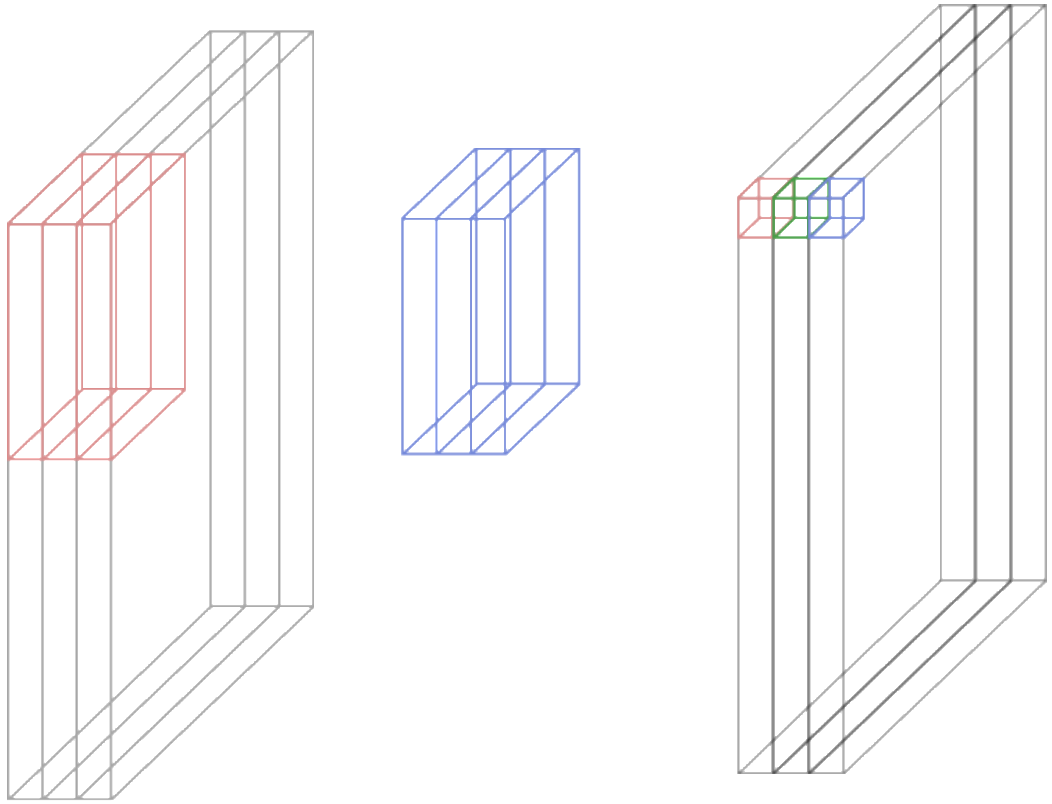
Output of a convolution layer



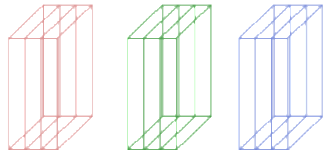
Learnable parameters



Output of a convolution layer



Learnable parameters



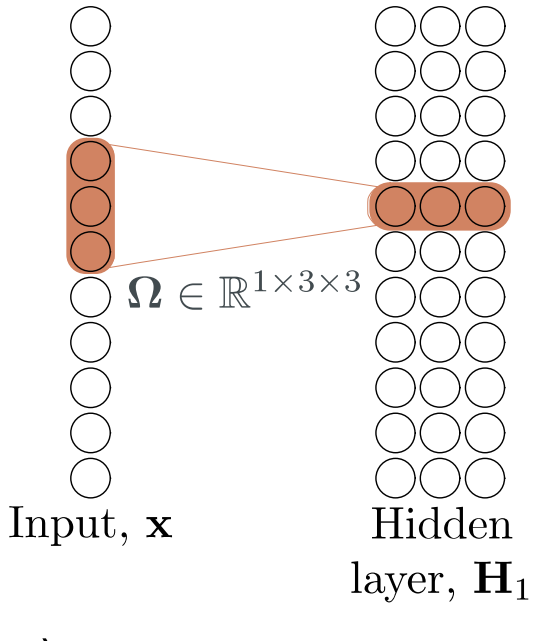
Examples of filters

$$* \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} =$$

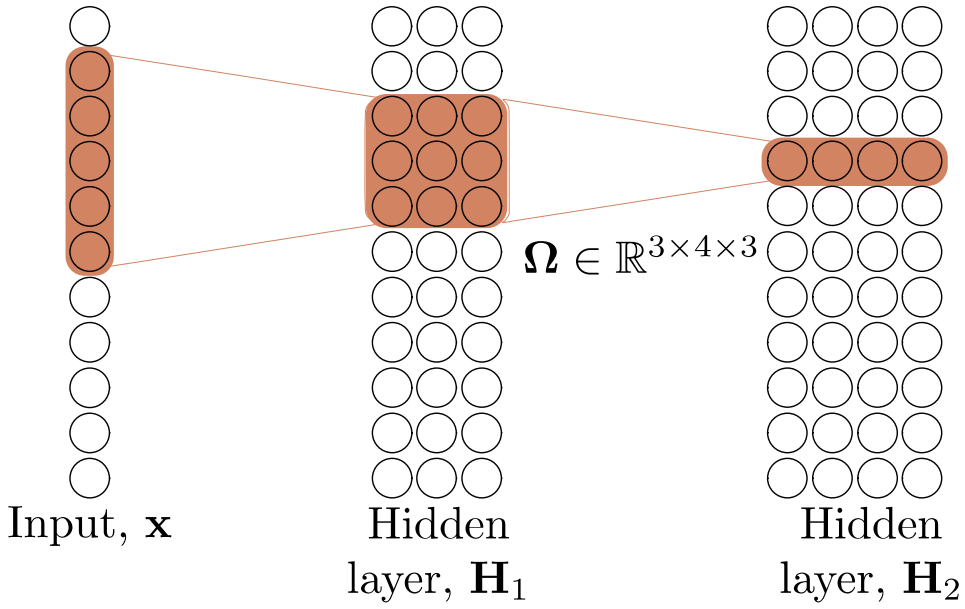
Examples of filters

$$* \begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} =$$

Receptive fields

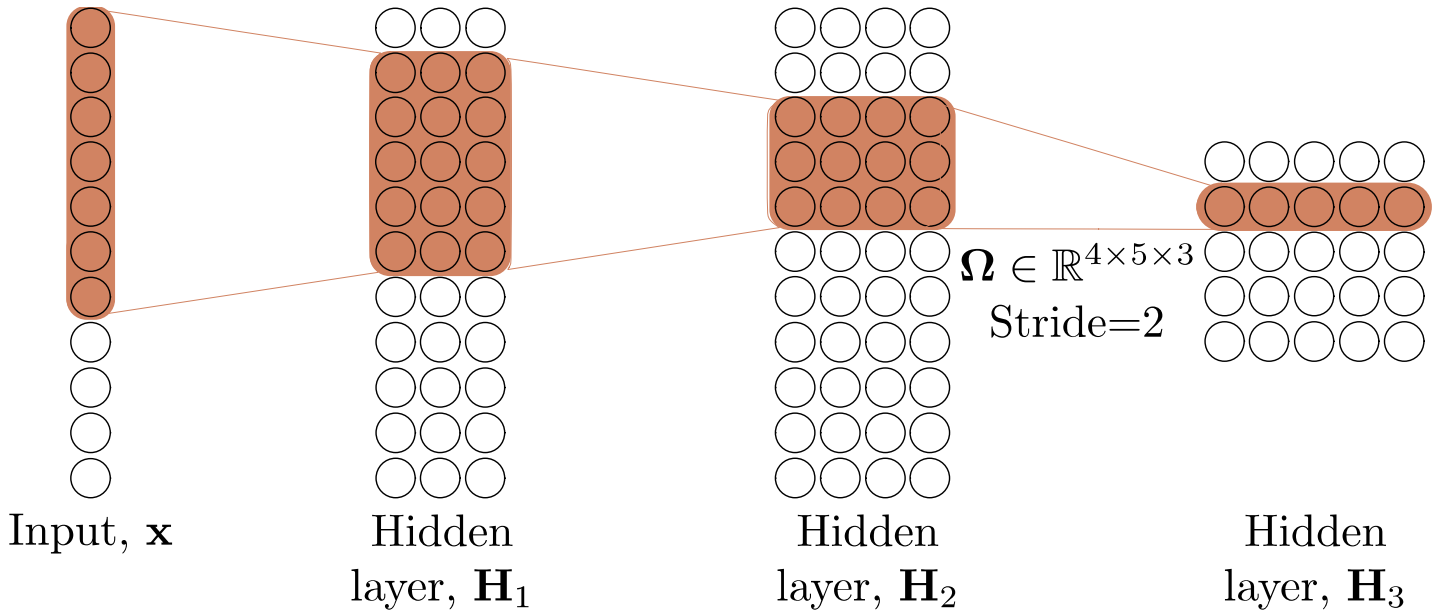


Receptive fields

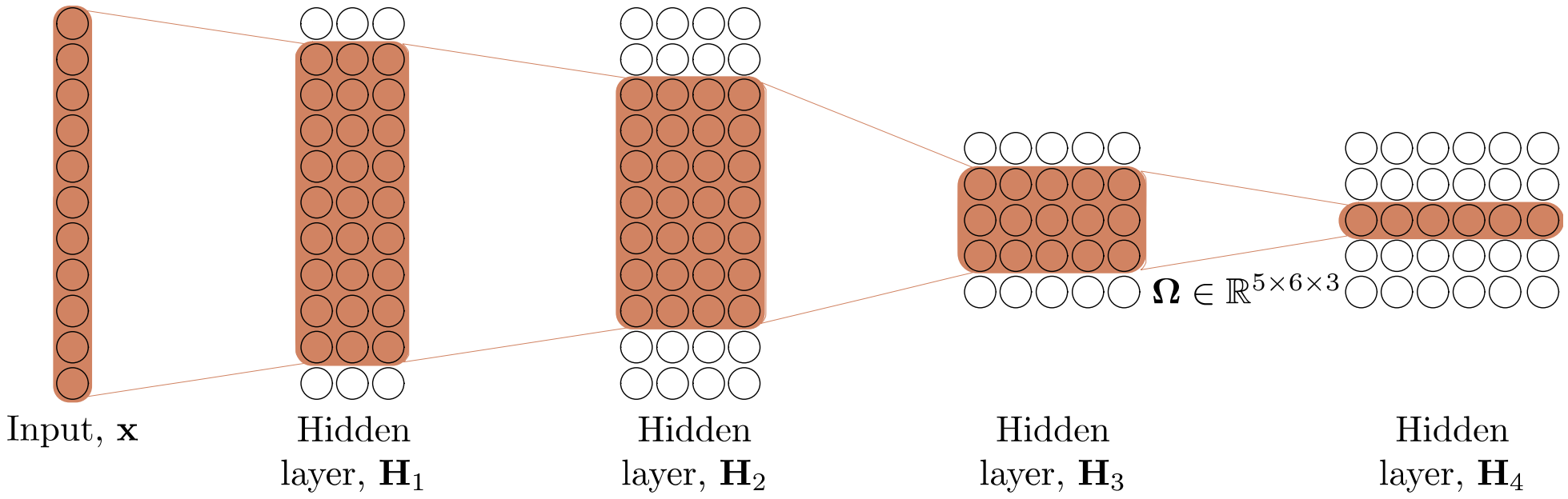


EIT Oxford

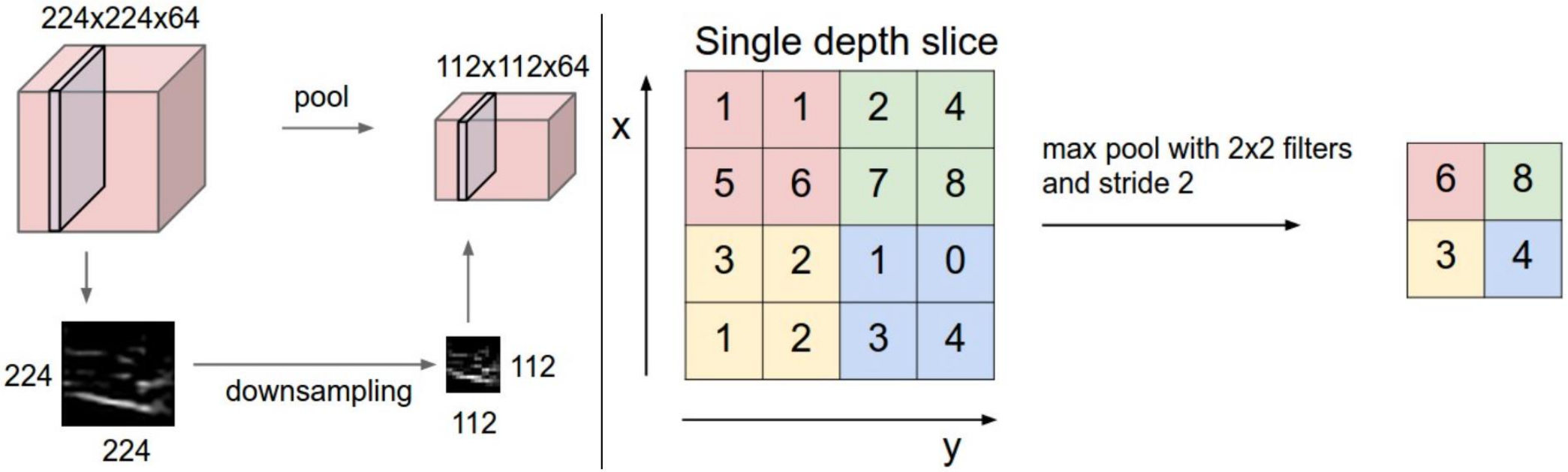
Receptive fields



Receptive fields

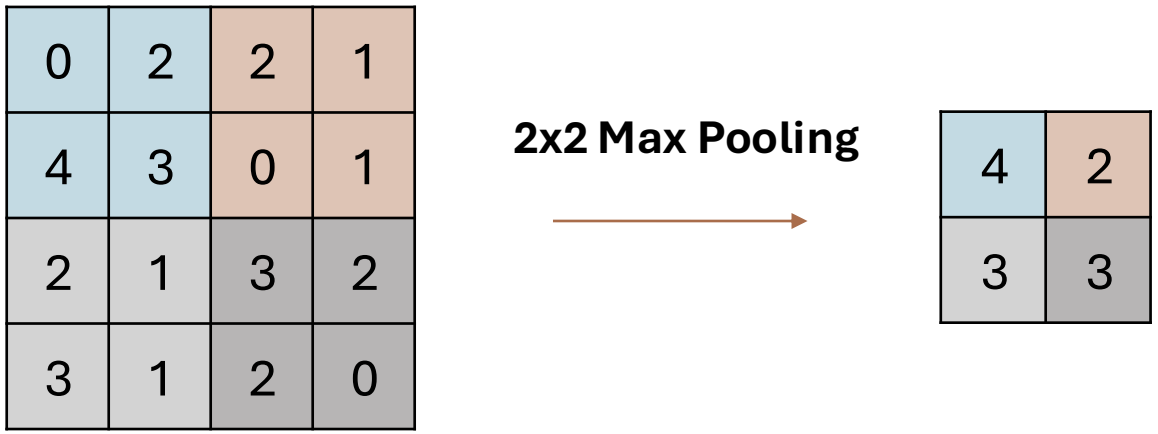


Pooling layers

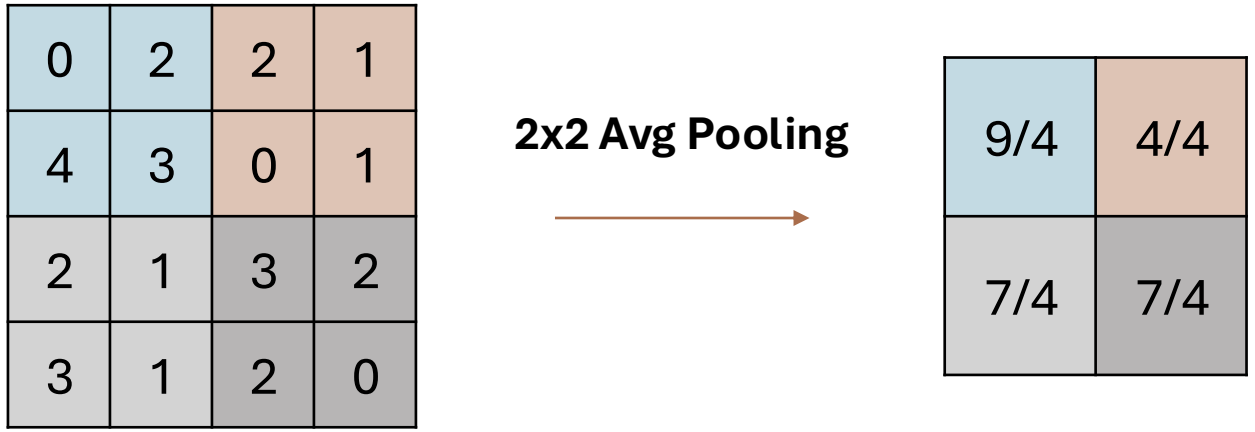


- Reduce dimensions of feature maps
- Contribute to invariance with respect to small translations in the input images

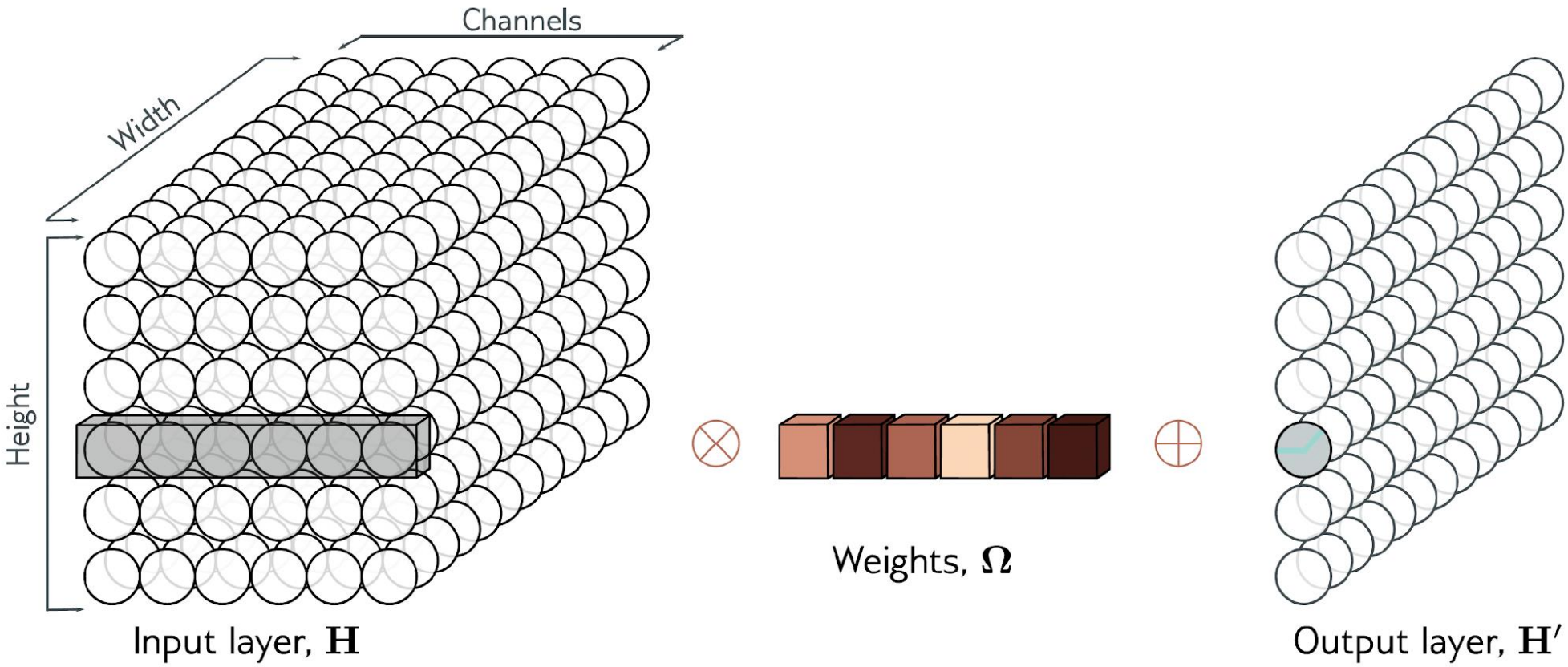
Max-pooling layer



Average-pooling layer

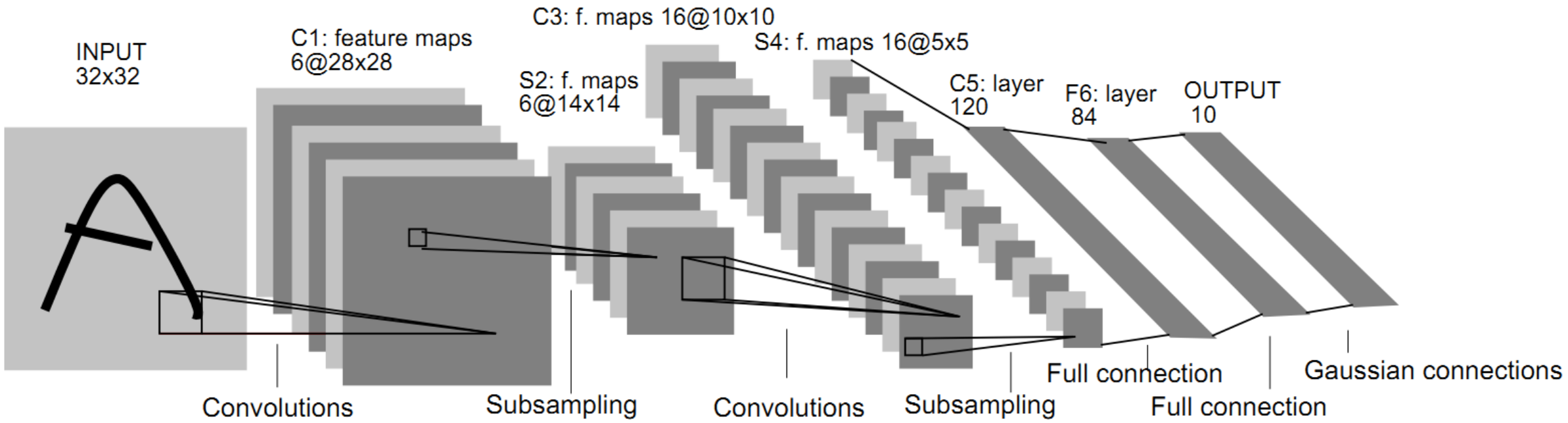


1x1 convolution



LeNet-5 (1998)

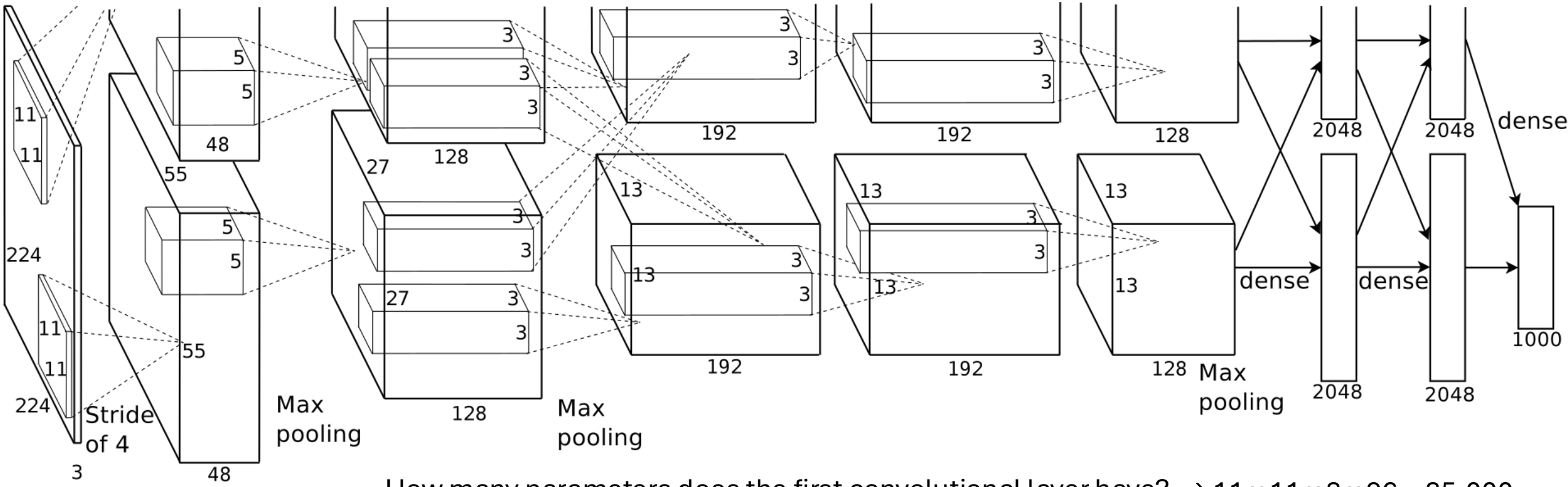
$$O = \frac{(W - F + 2P)}{S} + 1$$



60.000 parameters

https://ieeexplore.ieee.org/document/726791

AlexNet (2012)



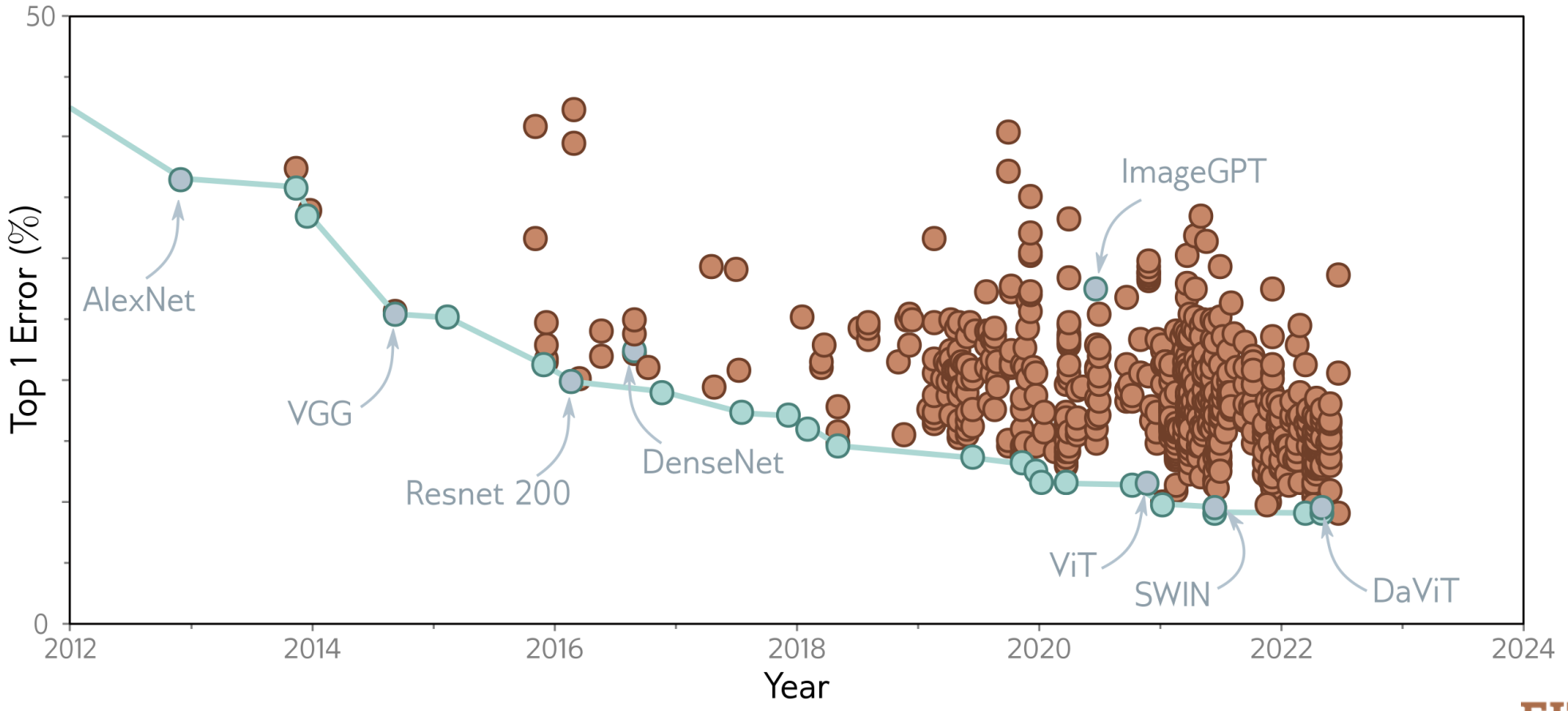
How many parameters does the first convolutional layer have? → 11 x 11 x 3 x 96 = 35.000

From the 60 million parameters, 58 million correspond to the dense layers

https://proceedings.neurips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

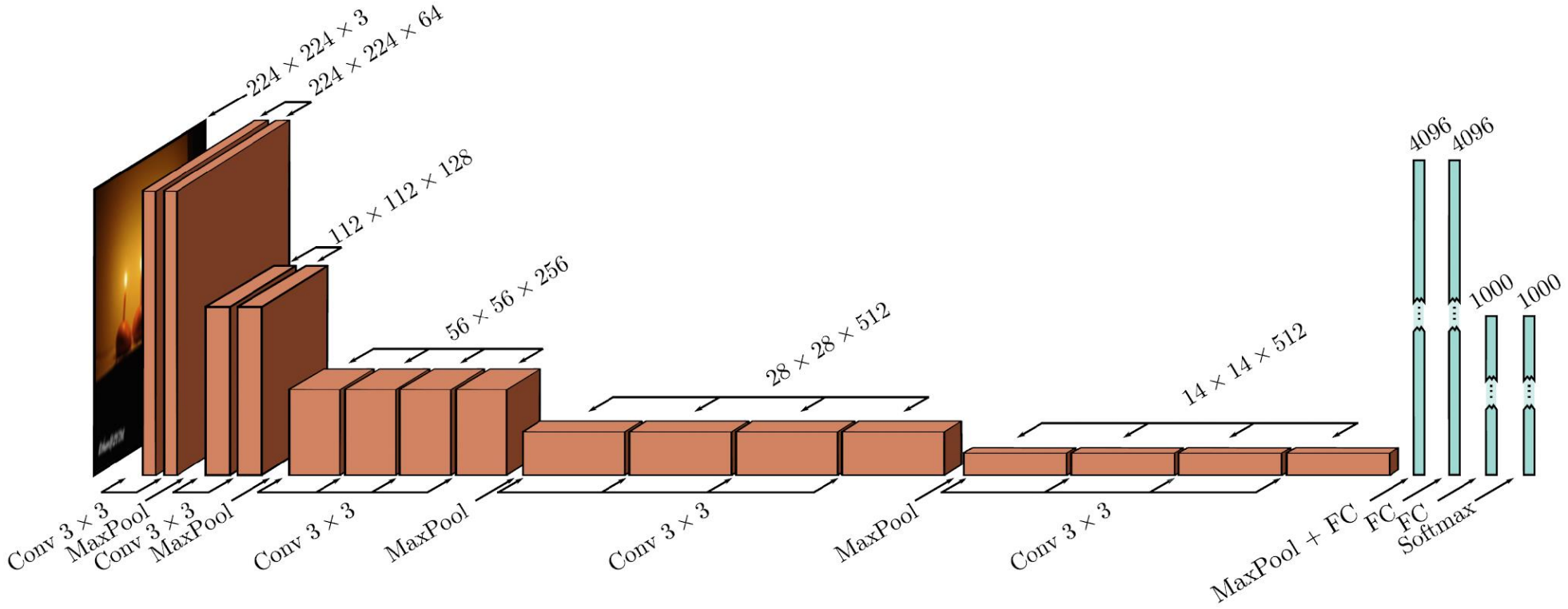
Classical architectures for classification

ImageNet Large Scale Visual Recognition Chellenge (ILSVRC)



EIT Oxford

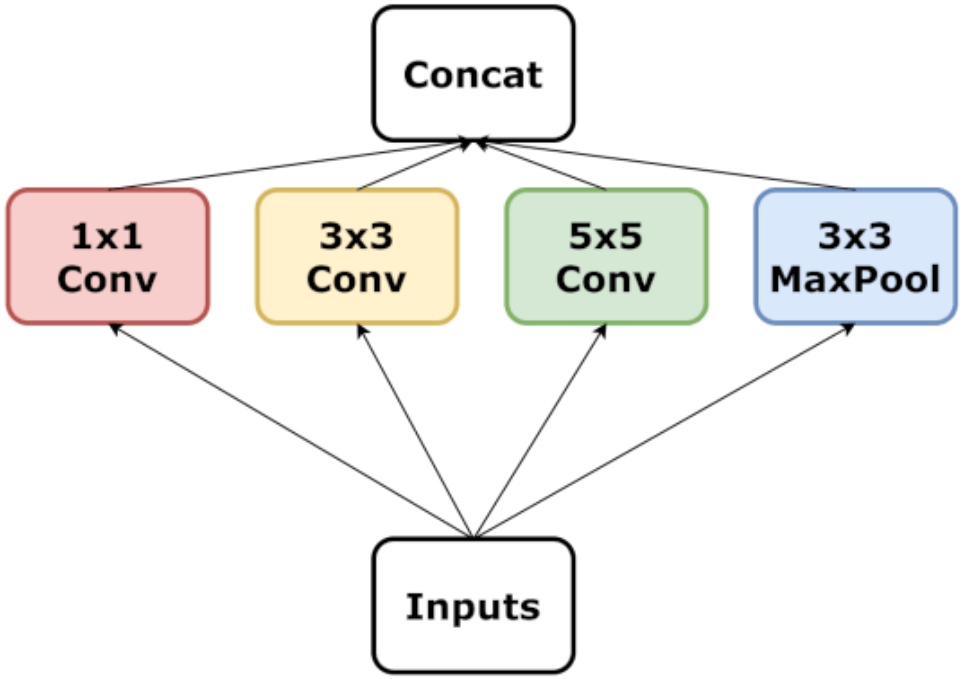
VGG (2015)



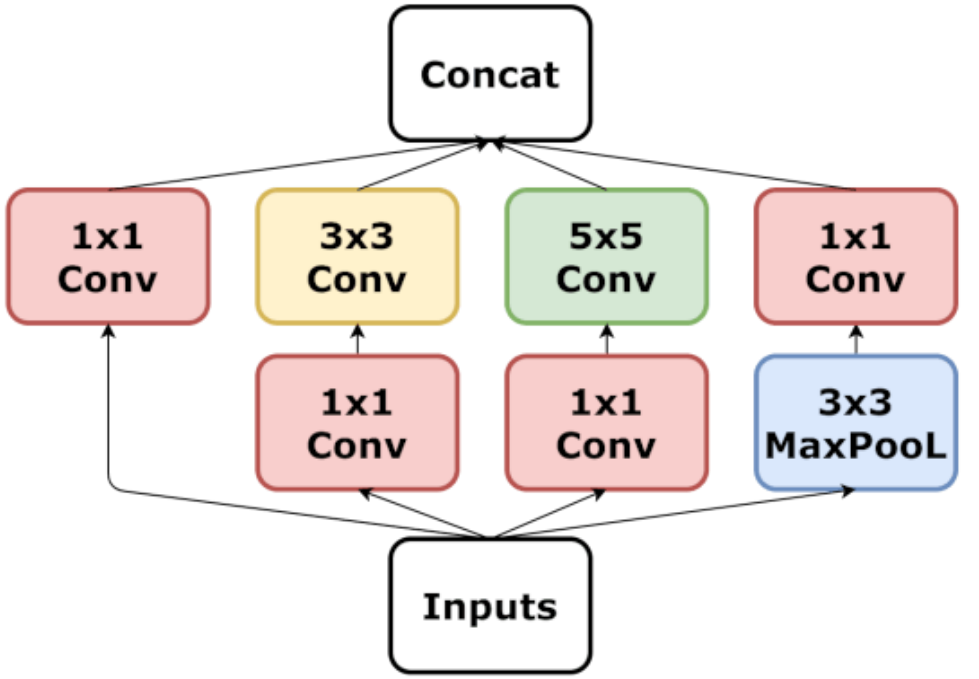
https://arxiv.org/pdf/1409.1556

GoogLeNet (2014)

Inception module – originally suffered from vanishing gradient problem



Inception Module

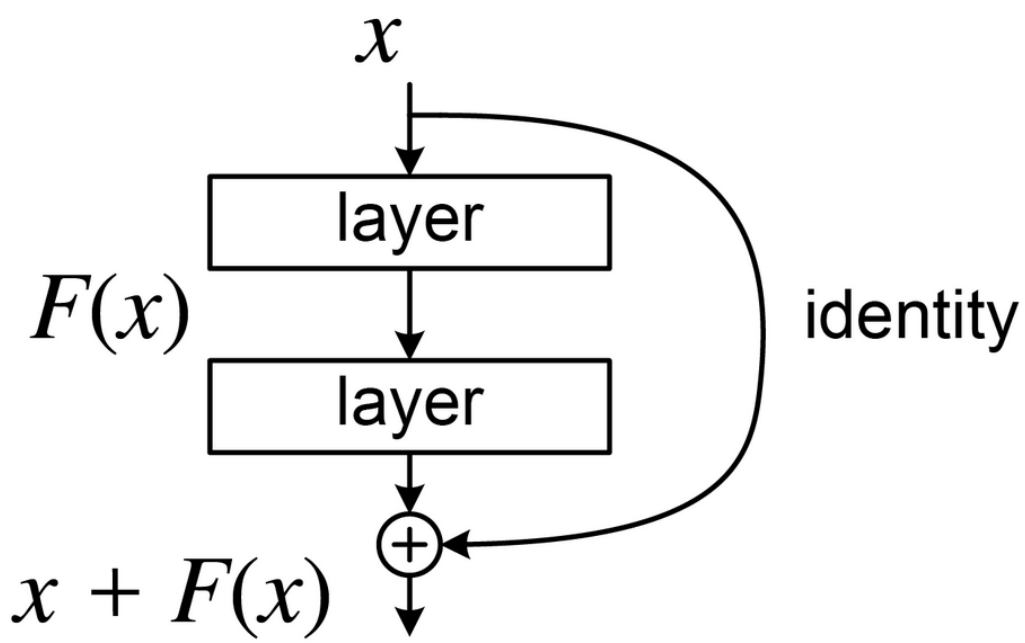


Inception Module with Dimension Reduction

https://arxiv.org/pdf/1409.1556

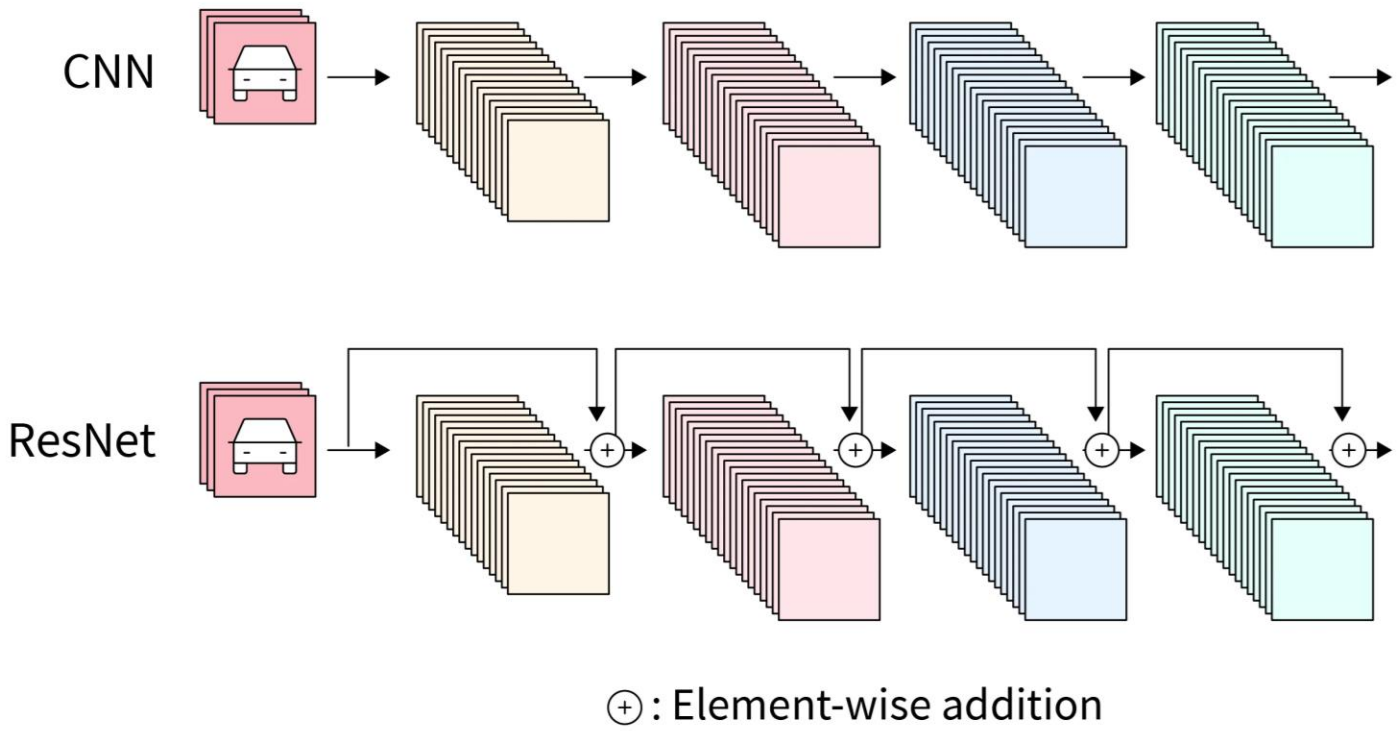
Skip / residual connections

- Performs an identity mapping to connect the input of the subnetwork with its output
- All transformer architectures include residual connections. Very deep transformers cannot be trained without them.



https://proceedings.mlr.press/v139/dong21a/dong21a.pdf

Skip / residual connections



https://www.scaler.com/topics/deep-learning

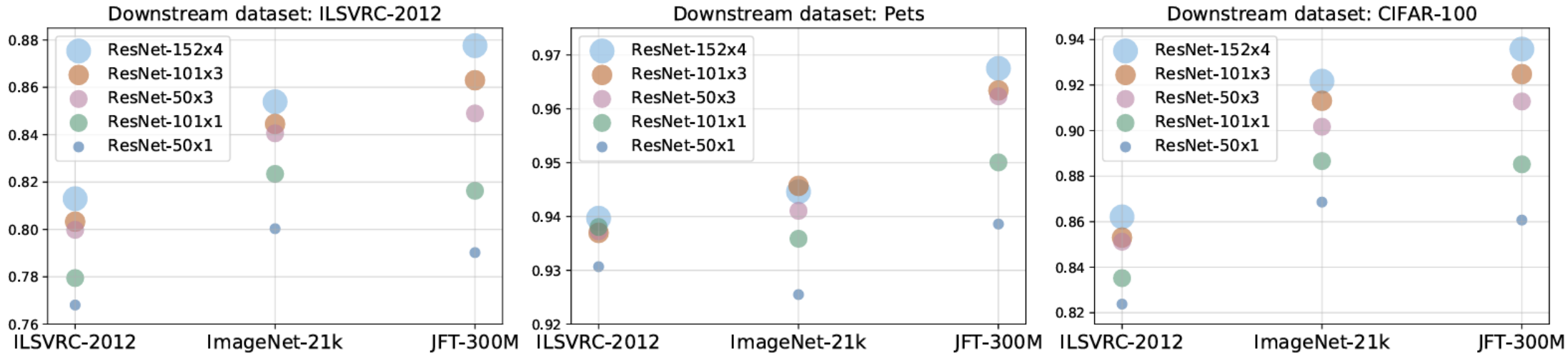
Big Transfer: The value of pretraining

- **Scale wins (data × model).** Pre-training on larger supervised sets (ImageNet-21k, JFT-300M) and using bigger ResNet backbones yields better transfer than ImageNet-1k alone
- **Strong across data regimes, including few-shot.** The same pre-trained checkpoint transfers well from 1 example per class up to million-example datasets
- **A simple, general fine-tuning recipe** ("BiT-HyperRule"). One set of hyper-params works broadly: vanilla SGD, large batch (≈512), learning rate decay by factor of 10 at 30/60/90% of training steps, and minimal augmentations (resize → random crop → flip). This reduces per-task tuning.
- **Pre-train once, adapt cheaply.** BiT emphasizes generalist representations: heavy compute is spent upstream; downstream fine-tuning is short and consistent across >20 datasets.
- **Benefits extend beyond classification.** Using BiT backbones improves detection when pre-training goes beyond ImageNet-1k (ImageNet-21k/JFT).
- **Bigger models + higher resolution help.** Very wide/deep ResNet-v2 variants (e.g., R152×4) and higher crop resolutions further lift transfer accuracy.

https://arxiv.org/pdf/1912.11370

Scale wins: classification

Effect of upstream data (shown on the x-axis) and model size on downstream performance.
Exclusively using more data or larger models may hurt performance; both need to be increased in tandem.

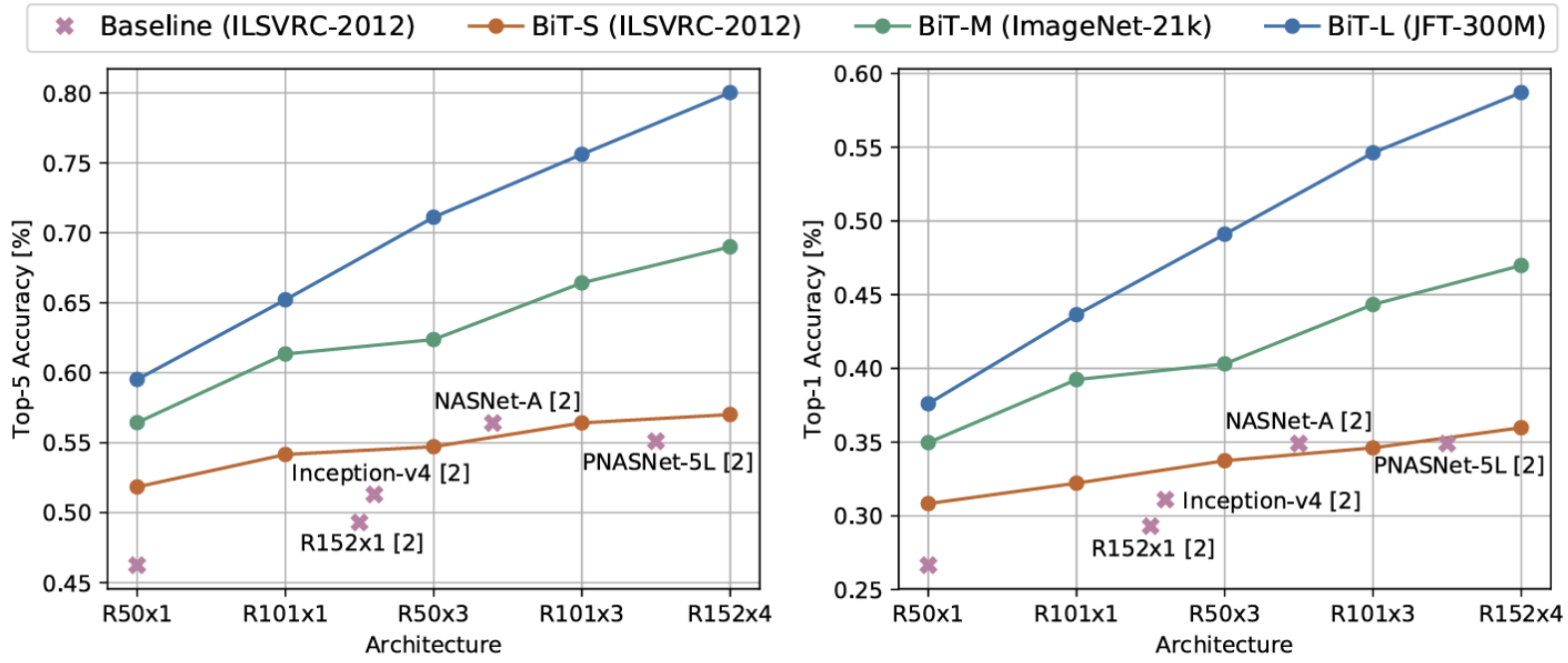


https://arxiv.org/pdf/1912.11370

EIT Oxford

Scale wins: object detection

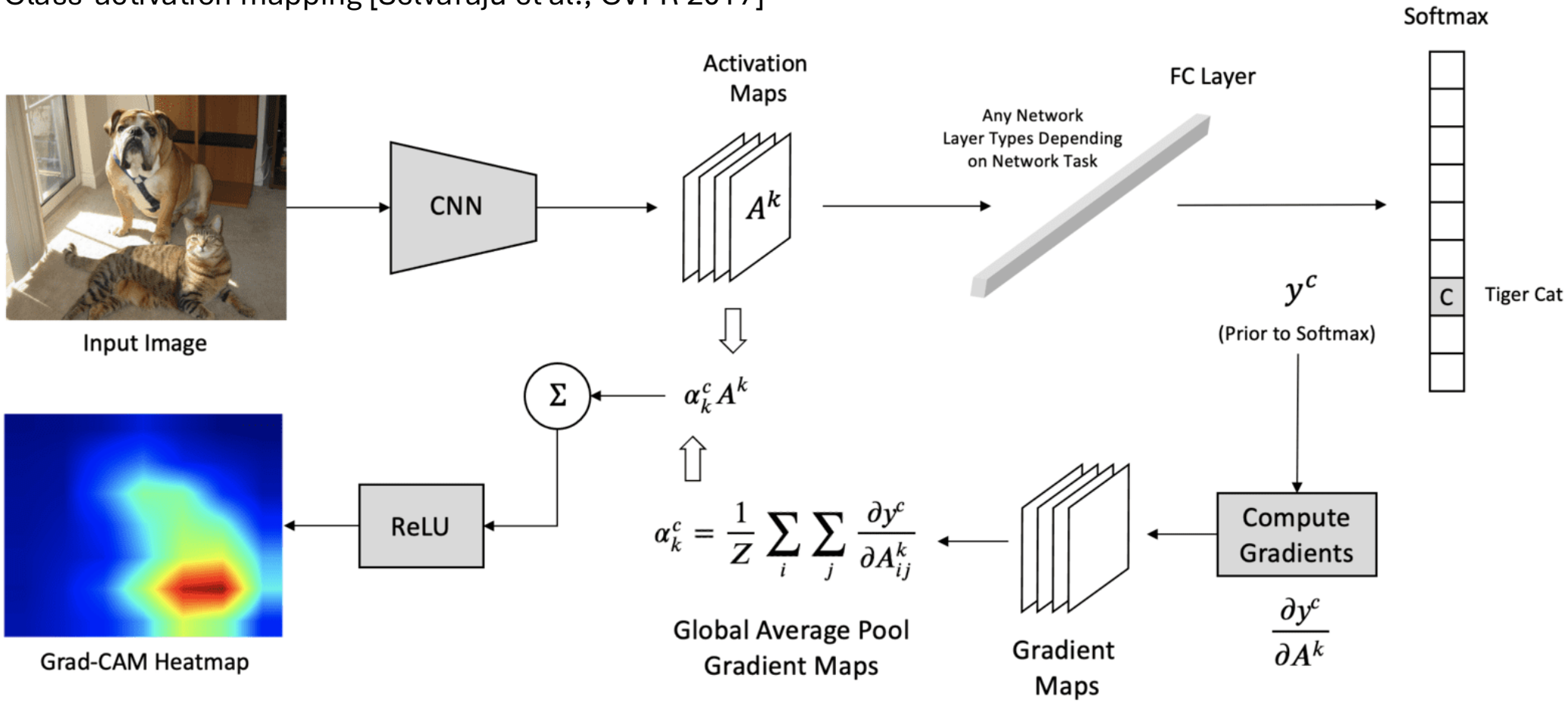
Results on the original ObjectNet test set with resize and central crop augmentations.



https://arxiv.org/pdf/1912.11370

Interpreting trained models

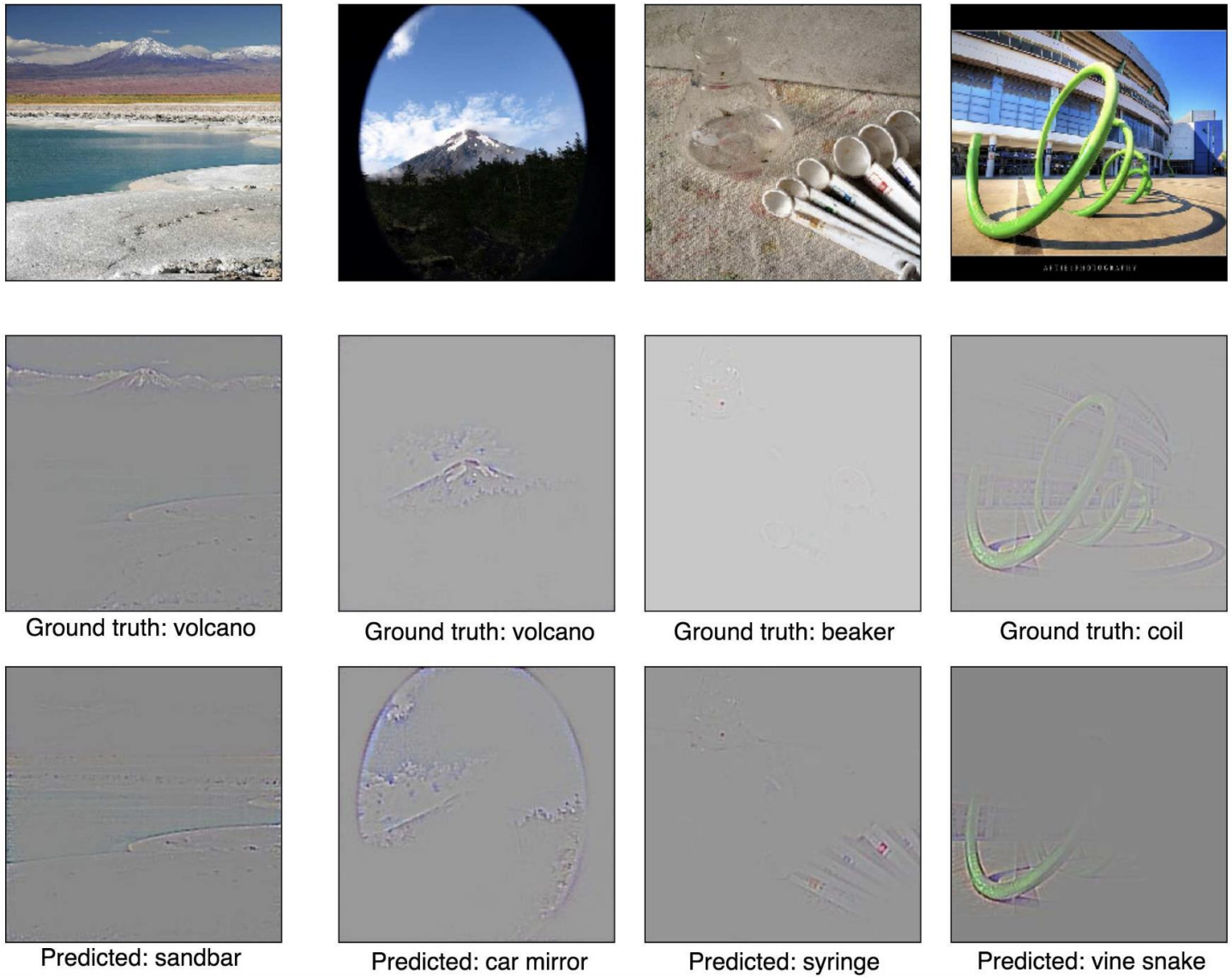
Class-activation mapping [Selvaraju et al., CVPR 2017]



EIT Oxford

Diagnosing failures through visual explanations

VGG-16 model



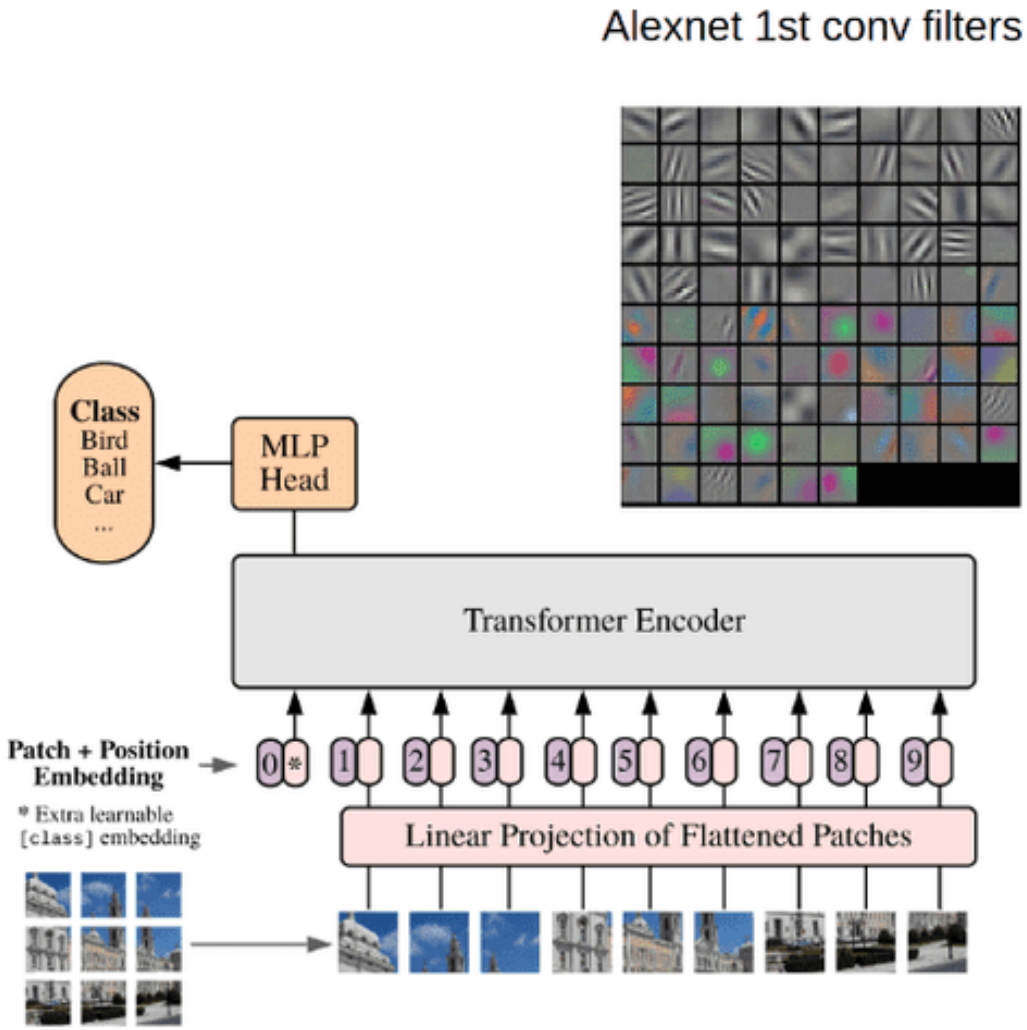
Vision transformers

AN IMAGE IS WORTH 16X16 WORDS:
TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

Alexey Dosovitskiy*,†, **Lucas Beyer***, **Alexander Kolesnikov***, **Dirk Weissenborn***,
Xiaohua Zhai*, **Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer,**
Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby*,†
*equal technical contribution, †equal advising
Google Research, Brain Team
{adosovitskiy, neilhoulsby}@google.com

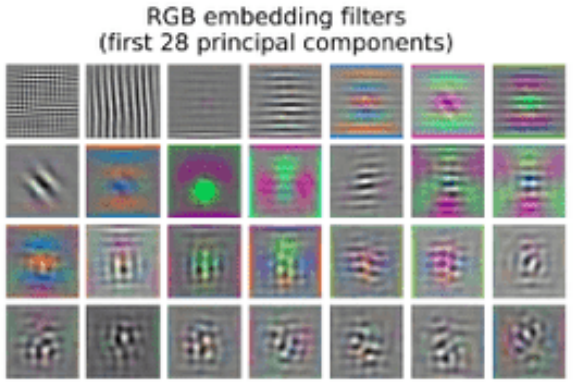
EIT Oxford

Vision transformers



Alexnet 1st conv filters

ViT 1st linear embedding filters



https://arxiv.org/pdf/2010.11929

Vision transformers

Main claim: We don't need convolutions anymore!

Less image-specific inductive biases than CNNs:

- **CNNs:** locality, 2D neighborhood structure & translation equivariance baked into each layer
- **ViTs:** only MLP layers are local and translation equivariant; attention layers are global

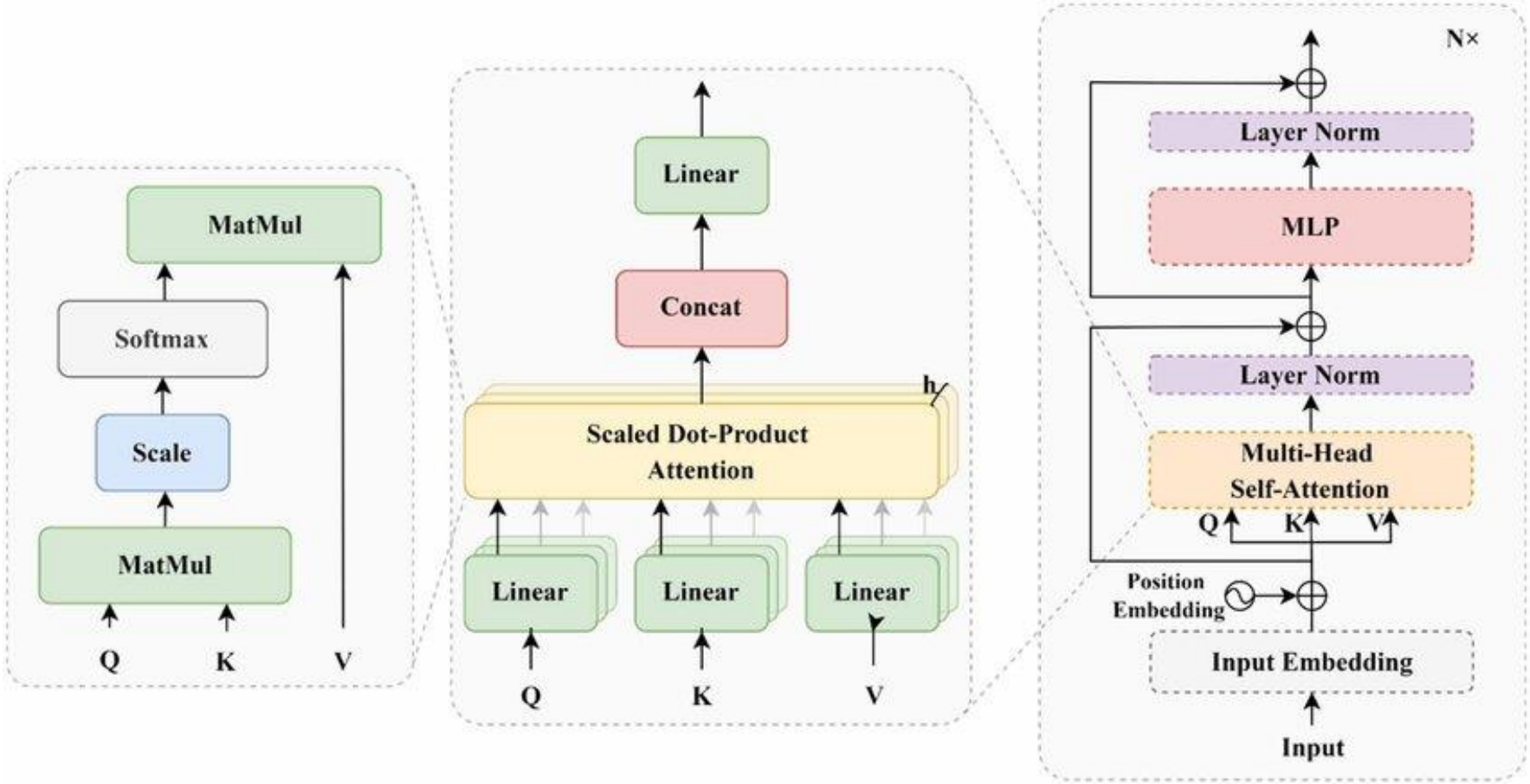
Reshape images $\mathrm{x} \in \mathbb{R}^{\mathrm{H}\times\mathrm{W}\times\mathrm{C}}$ to flattened 2D-patch representations $\mathrm{x_p} \in \mathbb{R}^{\mathrm{N}\times(\mathrm{P}^2\cdot\mathrm{C})}$

(P, P) is the resolution of each image patch and $\mathrm{N} = \frac{\mathrm{WH}}{\mathrm{P}^2}$ the resulting number of patches

$$
\begin{aligned}
\mathbf{z}_0 &= [\mathbf{x}_{\text{class}};\ \mathbf{x}_p^1\mathbf{E};\ \mathbf{x}_p^2\mathbf{E};\ \cdots\ ;\ \mathbf{x}_p^N\mathbf{E}] + \mathbf{E}_{pos}, && \mathbf{E} \in \mathbb{R}^{(P^2\cdot C)\times D},\ \mathbf{E}_{pos} \in \mathbb{R}^{(N+1)\times D} \\
\mathbf{z}'_\ell &= \mathrm{MSA}(\mathrm{LN}(\mathbf{z}_{\ell-1})) + \mathbf{z}_{\ell-1}, && \ell = 1 \ldots L \\
\mathbf{z}_\ell &= \mathrm{MLP}(\mathrm{LN}(\mathbf{z}'_\ell)) + \mathbf{z}'_\ell, && \ell = 1 \ldots L \\
\mathbf{y} &= \mathrm{LN}(\mathbf{z}_L^0)
\end{aligned}
$$

https://arxiv.org/pdf/2010.11929

Transformer block with MSA

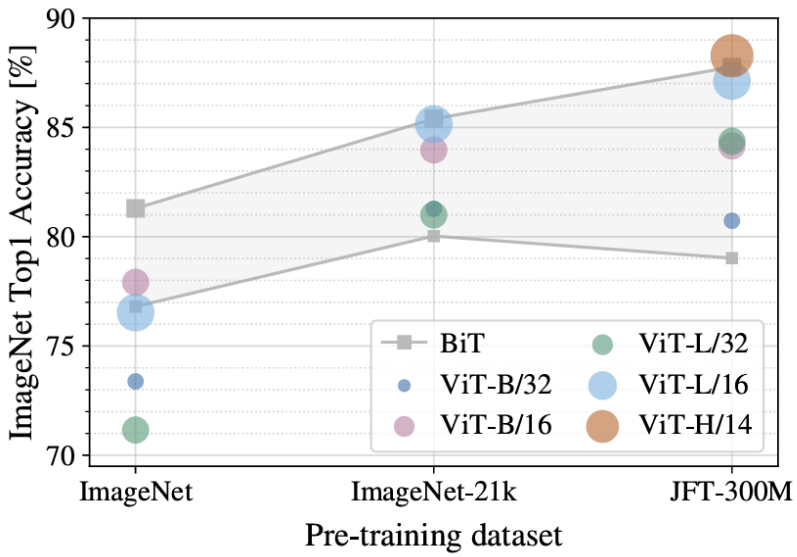


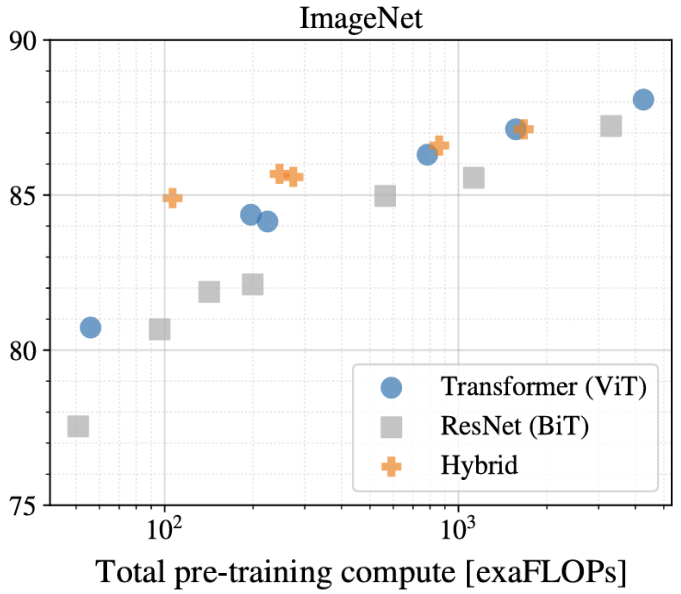
https://arxiv.org/pdf/1706.03762

EIT Oxford

Transformers vs. CNNs

- **Compute/memory efficiency is favourable at scale.** With big pretraining, ViT attains SOTA **with substantially less pretraining compute** than strong CNN baselines (e.g., BiT), and large ViTs are **more memory-efficient** (larger per-core batch sizes) than comparable ResNets.
- **Model/patch scaling:** ViT comes in Base/Large/Huge variants; **smaller patches → longer sequences → higher compute but better accuracy**.
- **Hybrid option exists.** You can feed **CNN feature maps** into the Transformer instead of raw patches; this can help when data are smaller (otherwise ViTs are data hungry).





EIT Oxford

The representation problem

Can useful representations develop without labels?

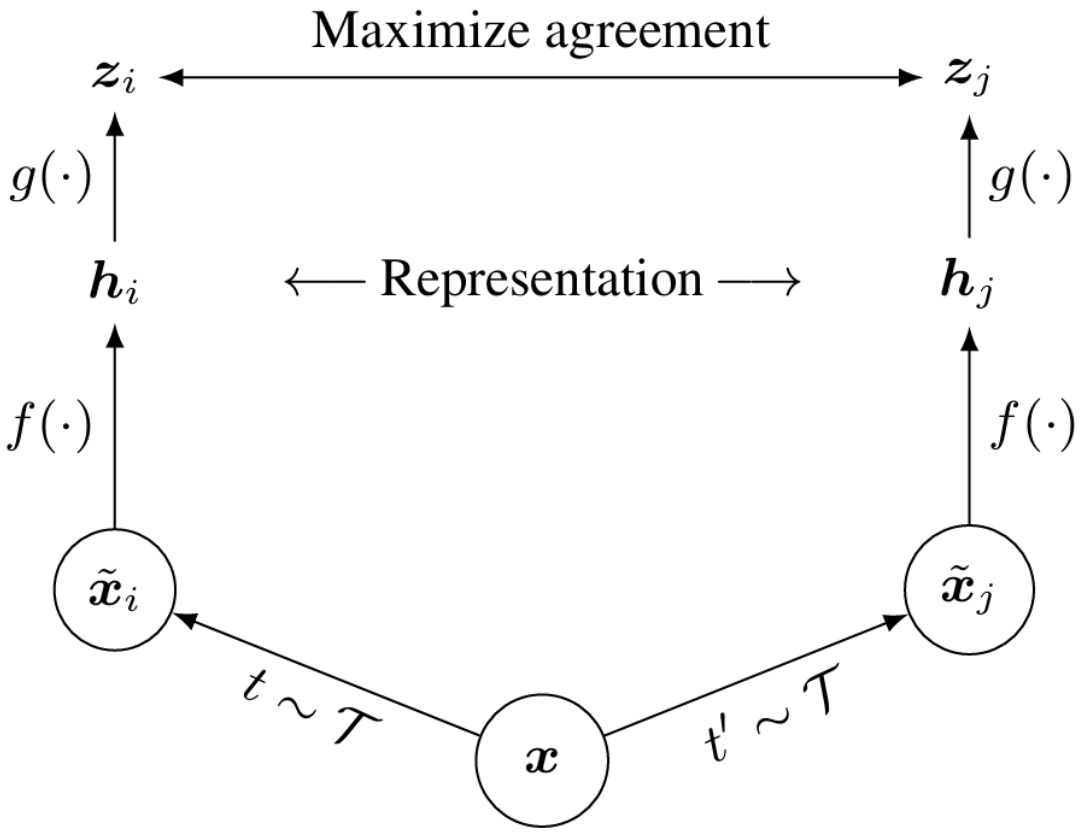
If so,

- We can learn more efficiently when we do gain access to labels
- We still learn useful things when label collection is impossible

Can we beat supervision without labels?

- Leverage contrastive learning = SimCLR
- Key ingredients:
  - Composition of **data augmentations**
  - **Learnable non-linear projection** head to improve representation quality
  - **Large batch sizes**; long training
- Importance of positive and negative mining

$$\ell_{i,j} = -\log \frac{\exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)}$$

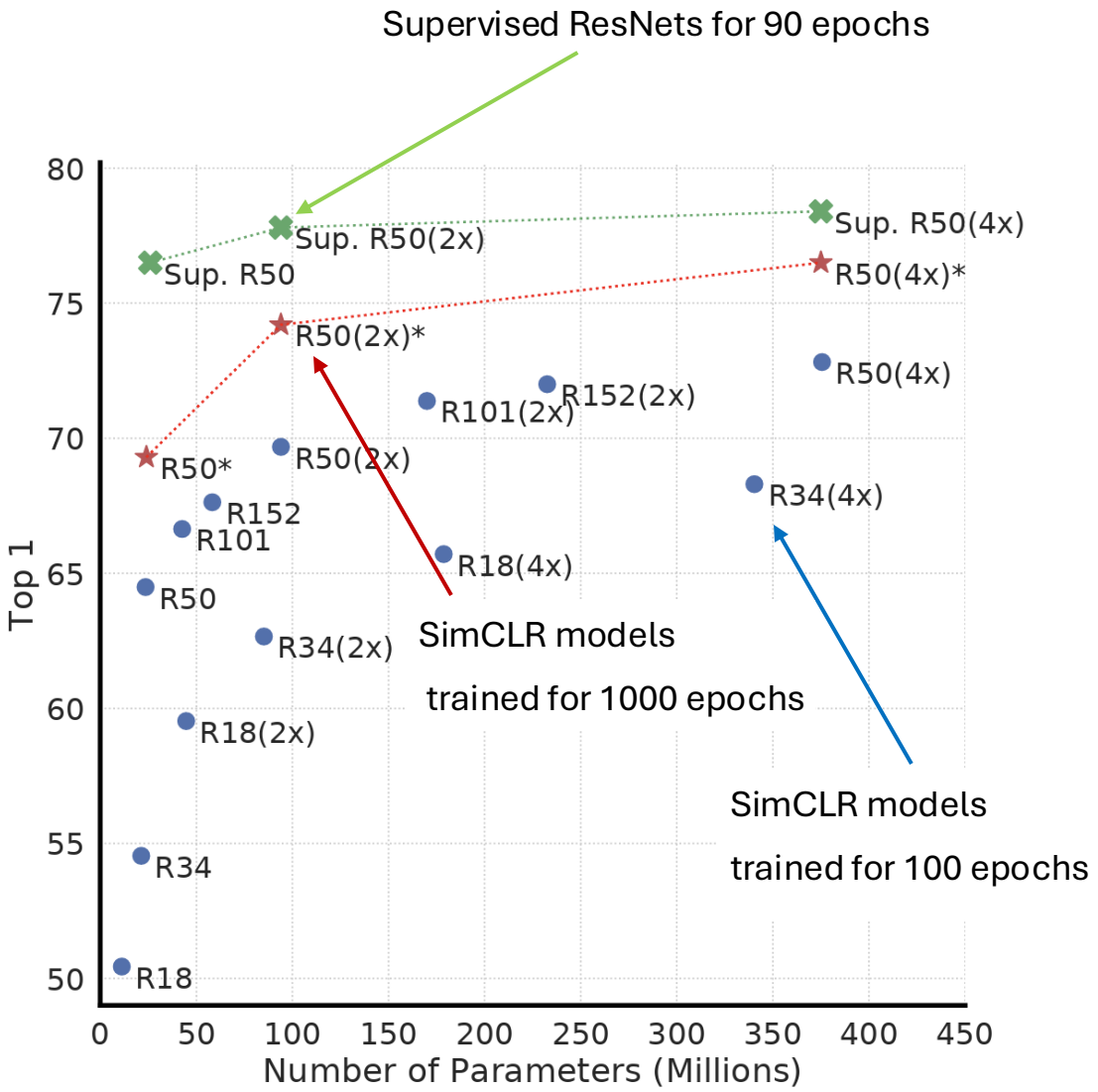


EIT Oxford

SimCLR training

Algorithm 1 SimCLR's main learning algorithm.

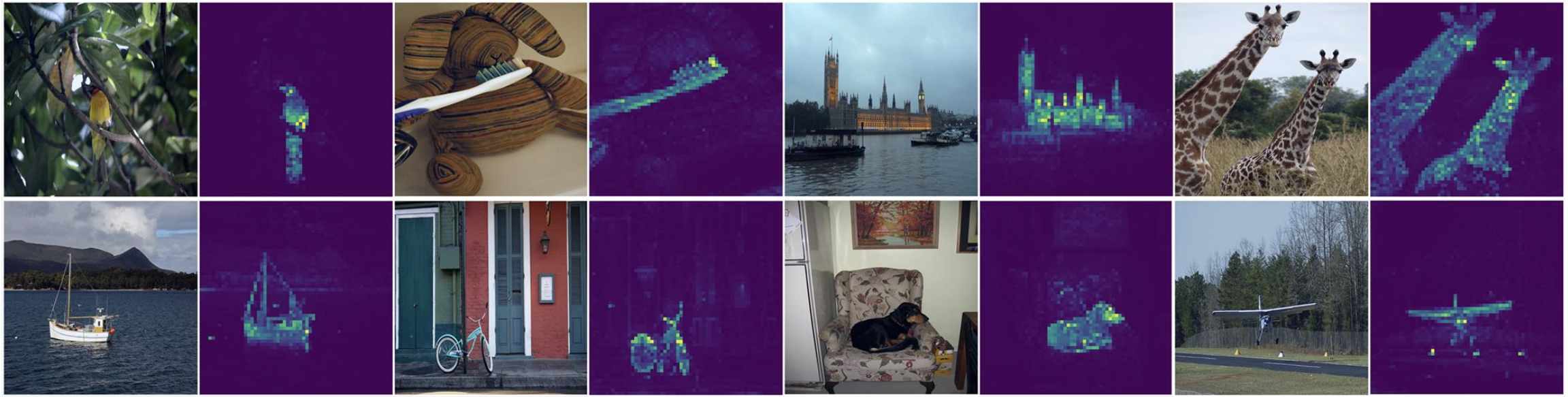
input: batch size N, constant τ, structure of f, g, $\mathcal{T}$.
for sampled minibatch $\{\boldsymbol{x}_k\}_{k=1}^{N}$ **do**
  for all $k \in \{1, \ldots, N\}$ **do**
    draw two augmentation functions $t \sim \mathcal{T}$, $t' \sim \mathcal{T}$
    # the first augmentation
    $\tilde{\boldsymbol{x}}_{2k-1}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



EIT Oxford

Can we learn representations with ViTs?

Self-**di**stillation with **no** labels instead of contrastive negatives [https://arxiv.org/pdf/2104.14294]



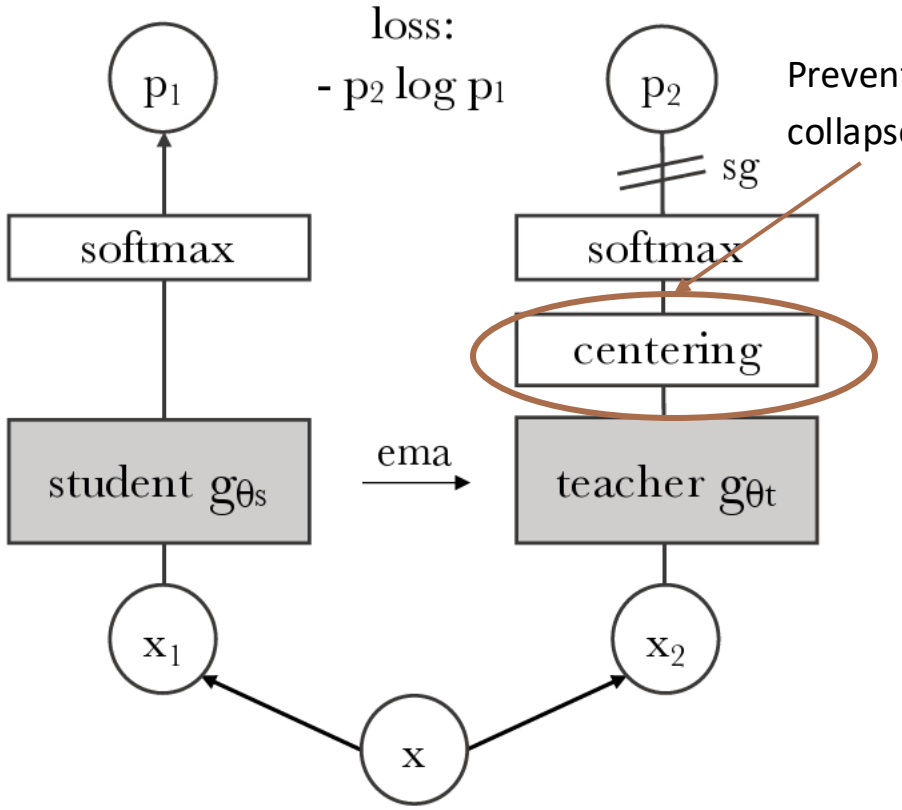
Self-attention from a ViT with 8x8 patches trained with no supervision

EIT Oxford

DiNO v1

Architecture and training recipe:

- **Form of self-distillation** with student–teacher ViTs (sharing same architecture); **teacher = EMA** of student; teacher uses a copy of the student with **stop-gradient**.
- Train by **distribution matching**: cross-entropy between temperature-scaled softmax (teacher "sharpened", student "softer") + **centering** to prevent collapse.
- **Multi-crop views**: a couple of global crops + several local crops encourage locality/globality.
- Outcomes: strong **linear-probe** ImageNet (e.g., ViT-B ~80% top-1) and **object-like attention maps** without labels.



EIT Oxford

Downstream tasks

Image retrieval

			$\mathcal{R}$Ox		$\mathcal{R}$Par	
Pretrain	Arch.	Pretrain	M	H	M	H
Sup. [57]	RN101+R-MAC	ImNet	49.8	18.5	74.0	**52.1**
Sup.	ViT-S/16	ImNet	33.5	8.9	63.0	37.2
DINO	ResNet-50	ImNet	35.4	11.1	55.9	27.5
DINO	ViT-S/16	ImNet	41.8	13.7	63.1	34.4
DINO	ViT-S/16	GLDv2	**51.5**	**24.3**	**75.3**	51.6

Copy detection

Method	Arch.	Dim.	Resolution	mAP
Multigrain [5]	ResNet-50	2048	224^2	75.1
Multigrain [5]	ResNet-50	2048	largest side 800	82.5
Supervised [69]	ViT-B/16	1536	224^2	76.4
DINO	ViT-B/16	1536	224^2	81.7
DINO	ViT-B/8	1536	320^2	**85.5**

Semantic segmentation

EIT Oxford

DiNO v2

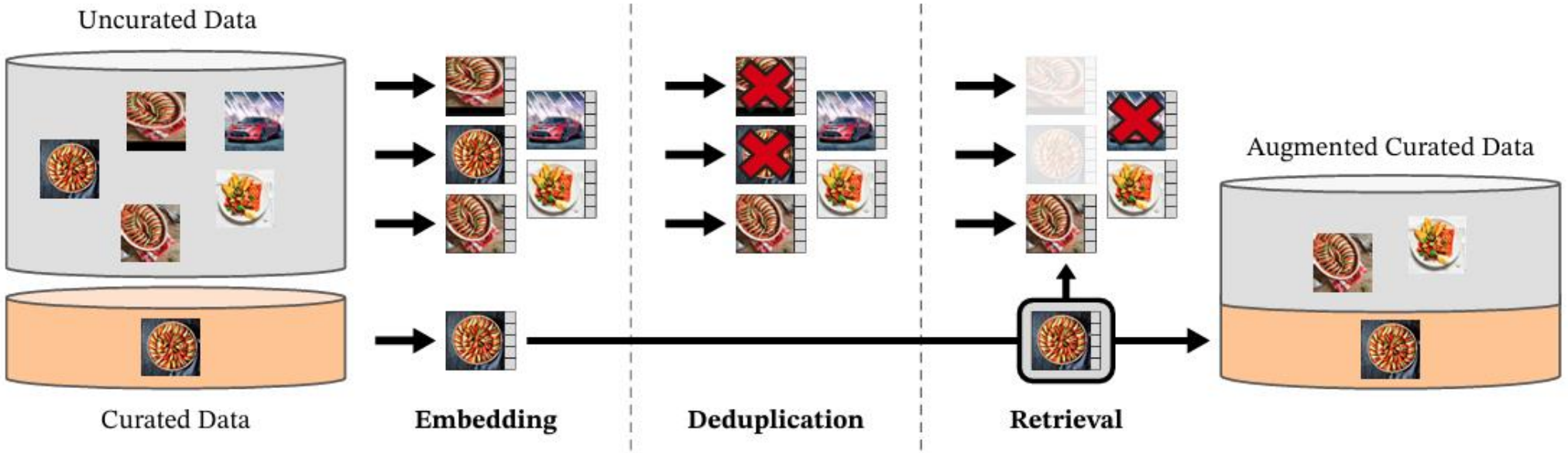
Learning Robust Visual Features without Supervision [https://arxiv.org/pdf/2304.07193]

Scaling recipe to “foundation-style” features → data is key

- **Scale everything**: larger ViTs (up to ~1B params teacher) and a **curated 142M-image dataset (LVD-142M)** built by retrieval + deduplication across diverse sources.
- **Stability/throughput tweaks** and **distillation**: train a huge teacher, **distill to smaller ViTs** for release.
- Result: **robust frozen features** that are SOTA/competitive across image-level and pixel-level tasks **without fine-tuning**.

DiNO v2 – data curation

Retrieve from pool of uncurated data, images that are close to those in curated datasets

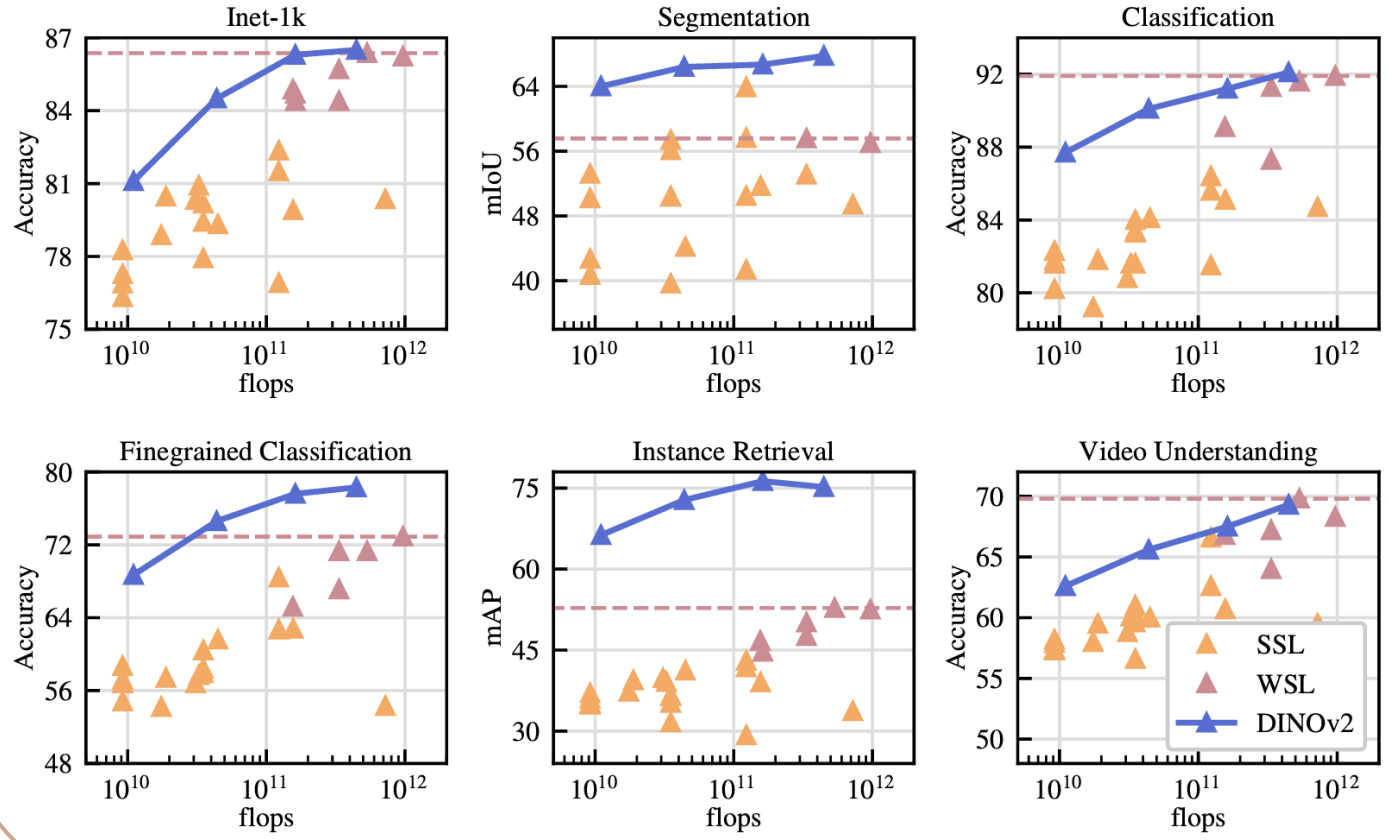


https://arxiv.org/pdf/2304.07193

DiNO v2 – representations & performance

Visualisation of PCA components

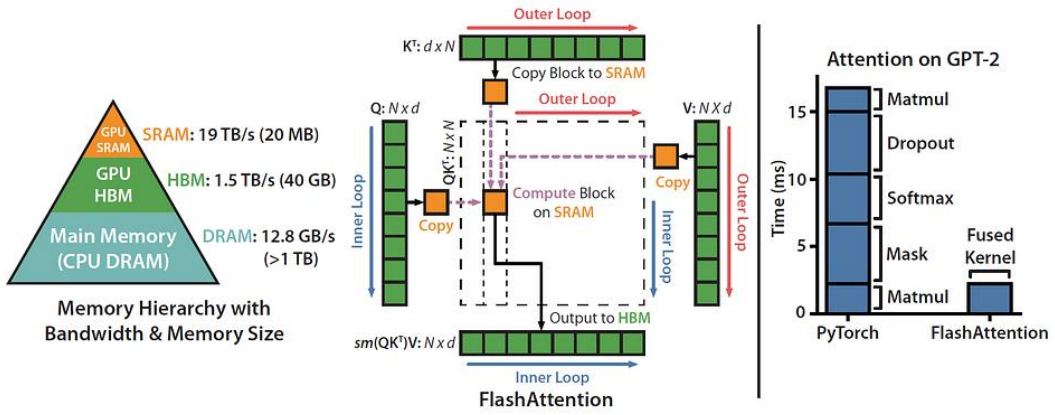
Performance with scaling



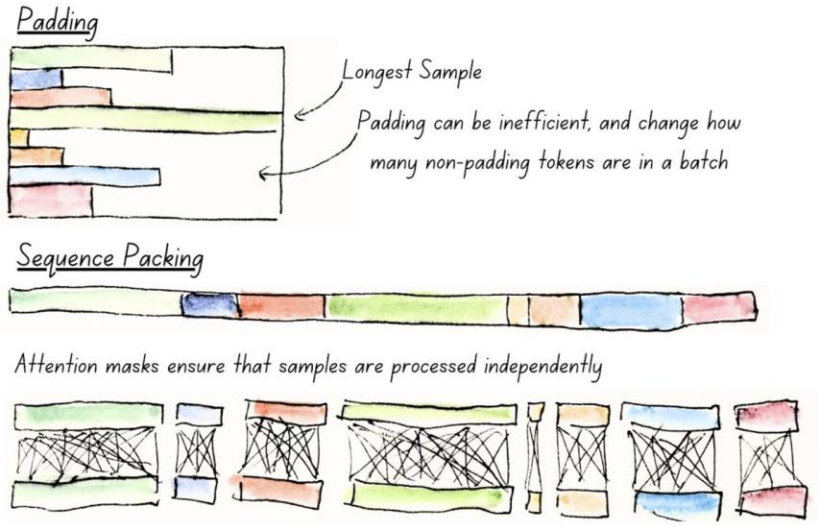
EIT Oxford

DiNO v2 - efficiency

Flash attention



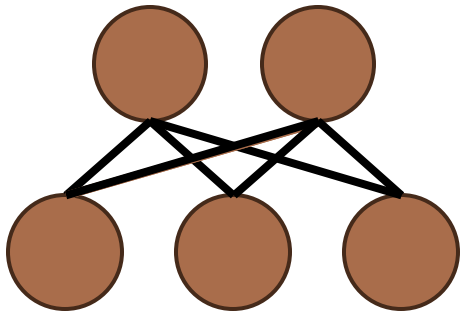
Sequence packing



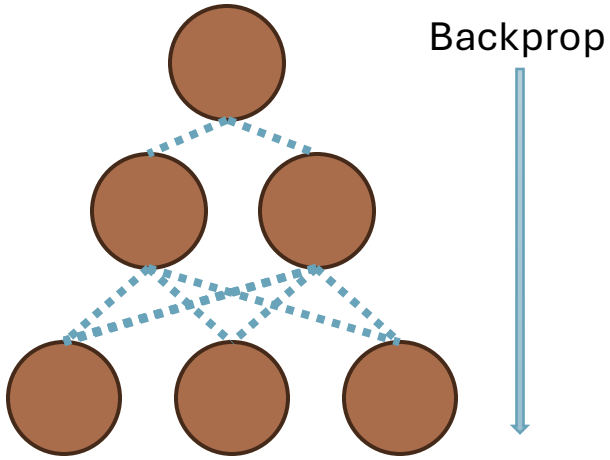
EIT Oxford

Putting it all together

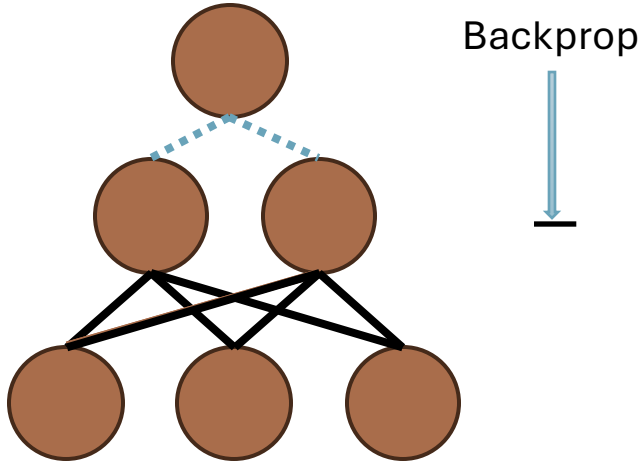
Pre-training



Fine-tuning



Linear probing



EIT Oxford

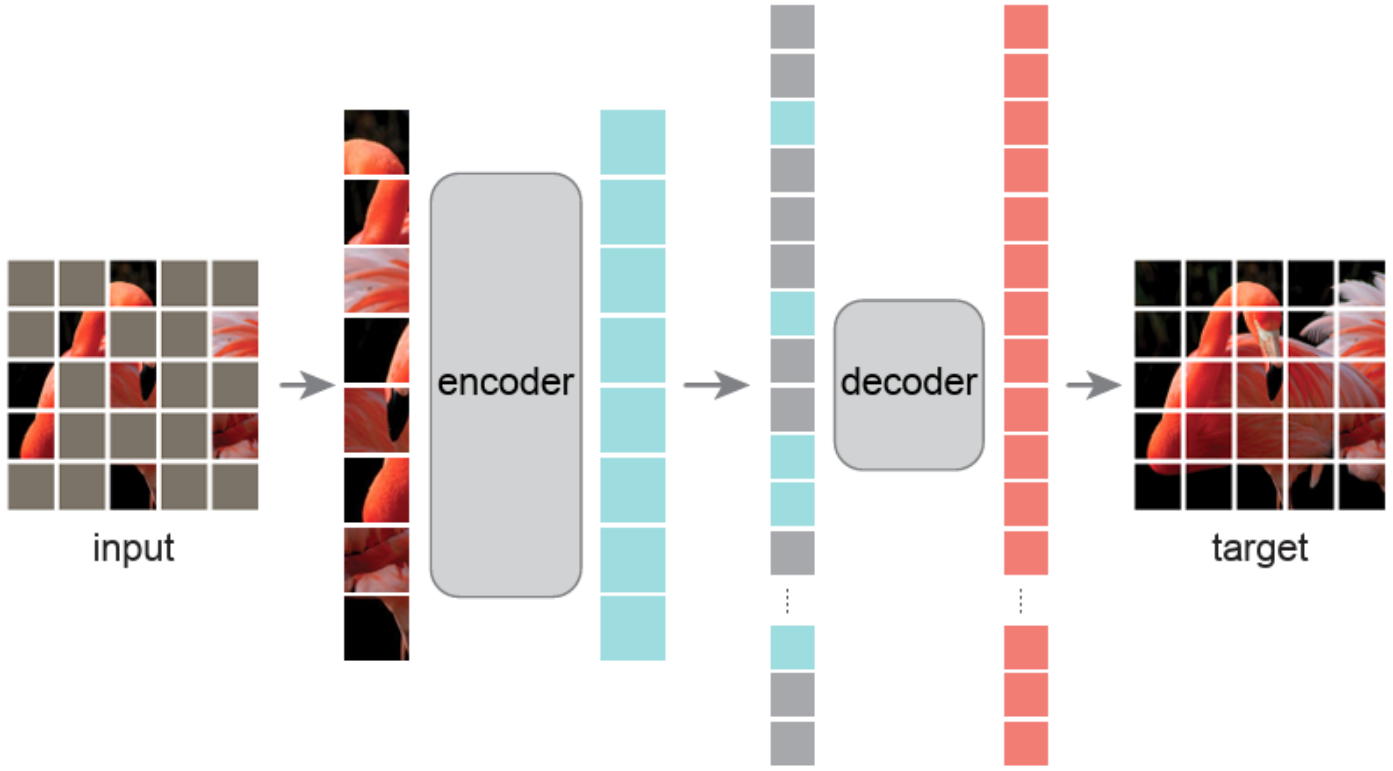
Other approaches to self-supervision

Masked Autoencoders Are Scalable Vision Learners

Kaiming He*,† Xinlei Chen* Saining Xie Yanghao Li Piotr Dollár Ross Girshick

*equal technical contribution †project lead

Facebook AI Research (FAIR)

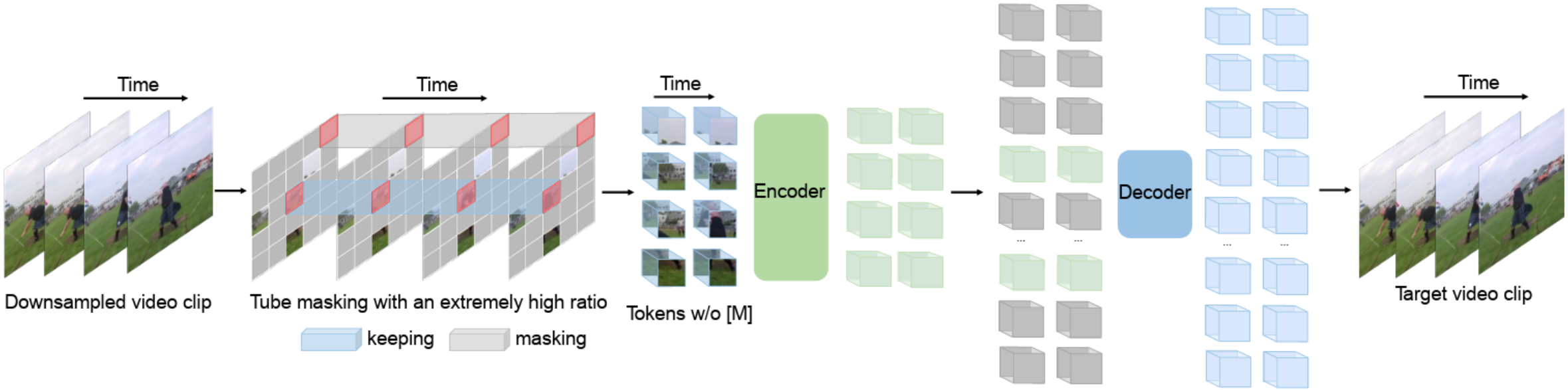


https://arxiv.org/pdf/2111.06377

From image to video

VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training

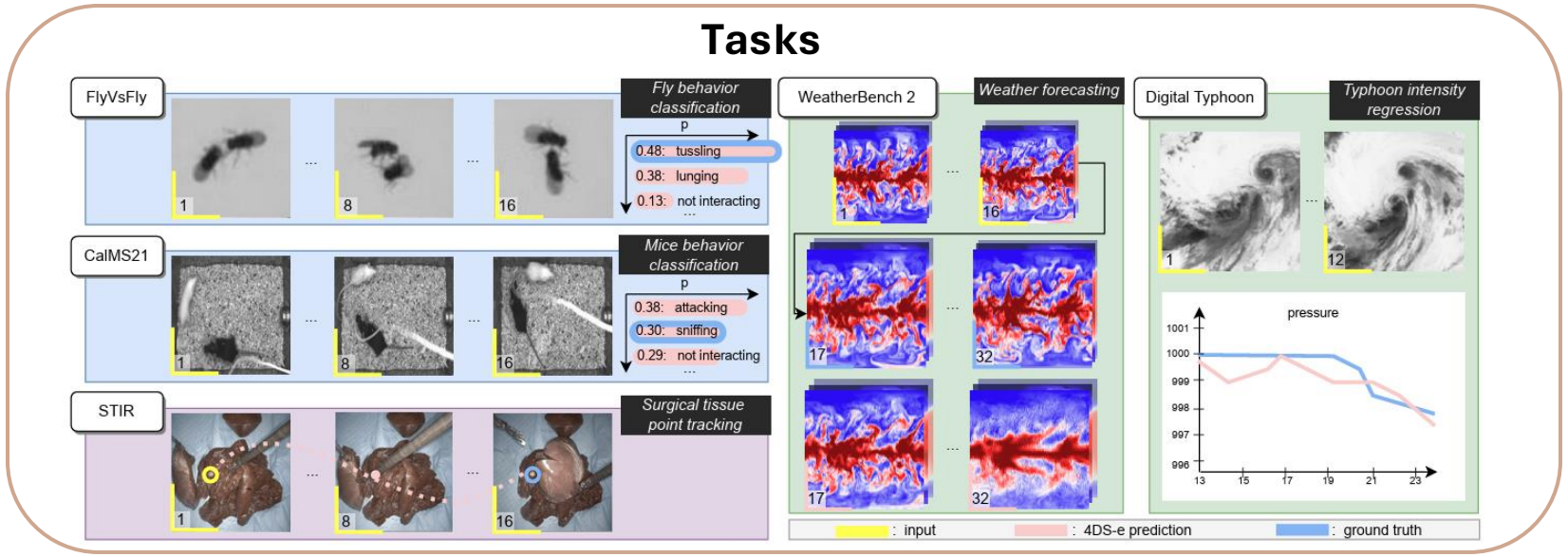
Zhan Tong [1,2*] **Yibing Song** [2] **Jue Wang** [2] **Limin Wang** [1,3†]
[1]State Key Laboratory for Novel Software Technology, Nanjing University
[2]Tencent AI Lab [3]Shanghai AI Lab
tongzhan@smail.nju.edu.cn {yibingsong.cv, arphid}@gmail.com lmwang@nju.edu.cn

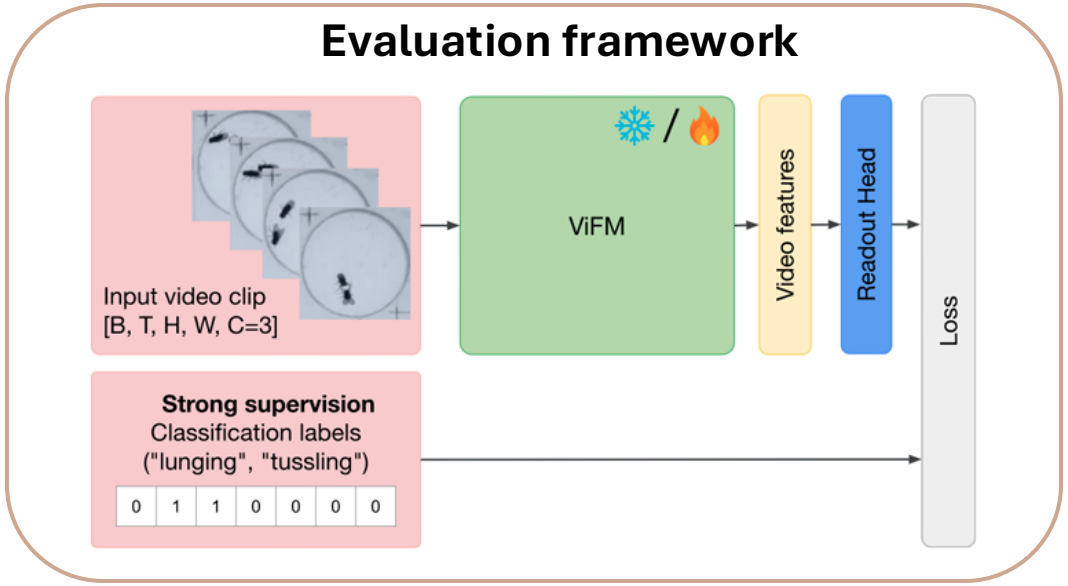


https://arxiv.org/pdf/2203.12602

EIT Oxford

Generalizing representations to scientific domains

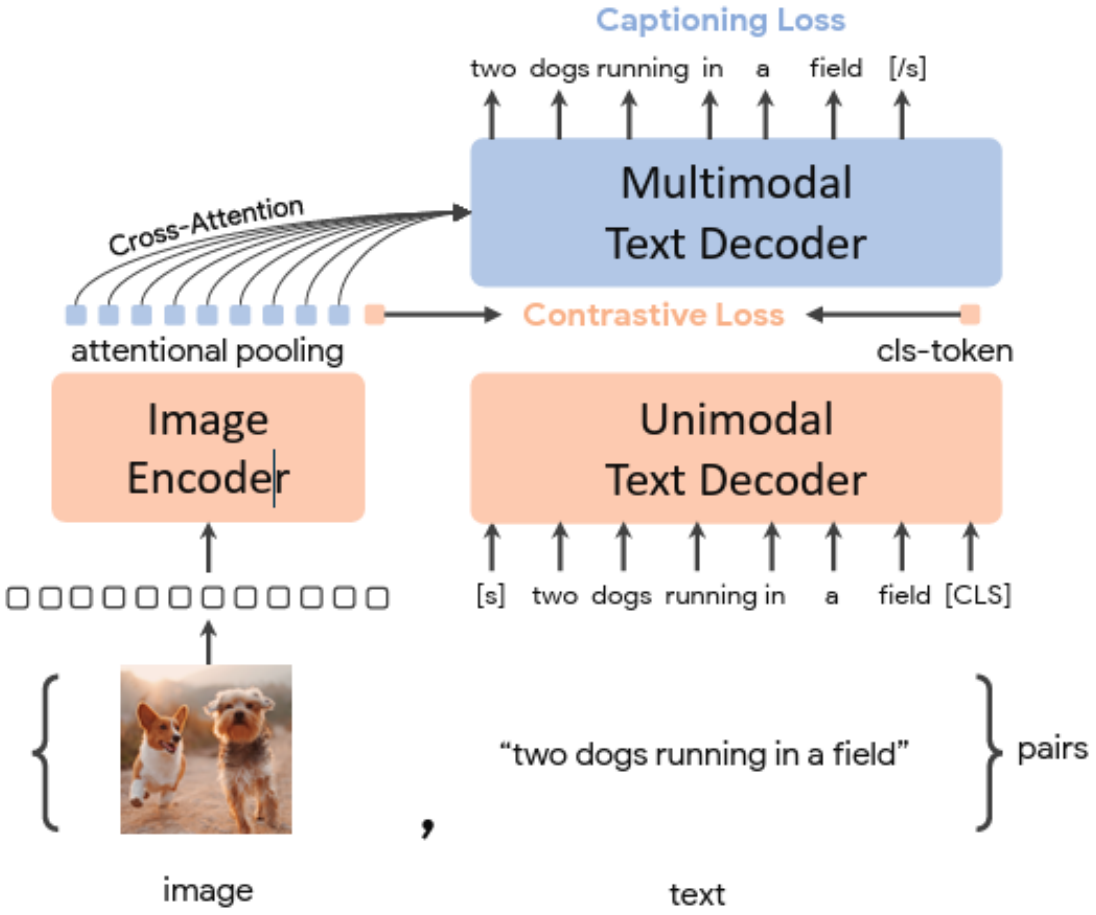




[Zhao et al. ICML (2024)]

[Hasson et al. ICCV (2025)]

Obtaining image representations from text



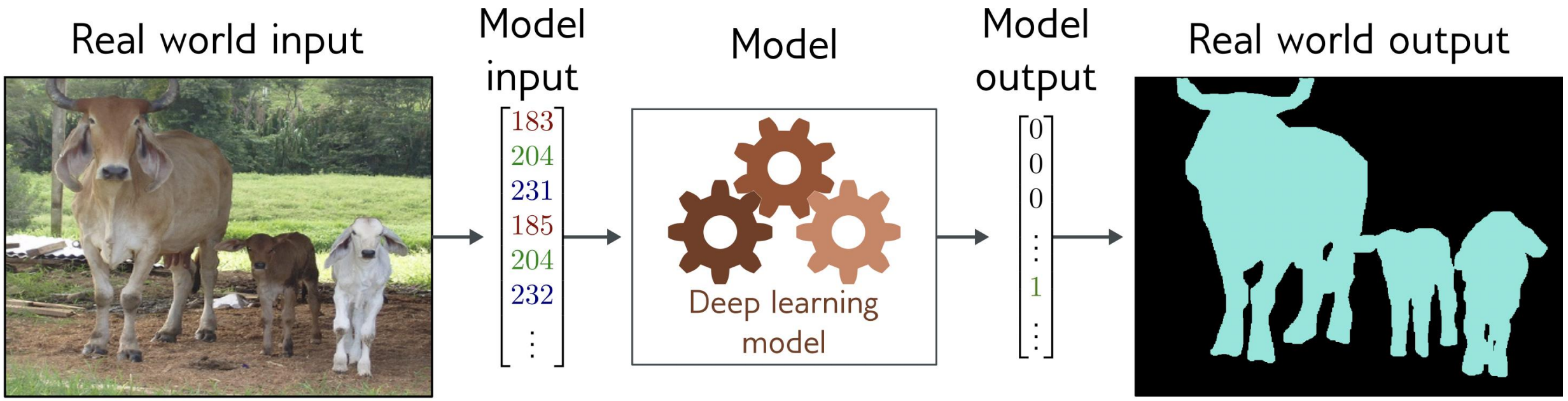
Algorithm 1 Pseudocode of Contrastive Captioners architecture.

```
# image, text.ids, text.labels, text.mask: paired {image, text} data
# con_query: 1 query token for contrastive embedding
# cap_query: N query tokens for captioning embedding
# cls_token_id: a special cls_token_id in vocabulary

def attentional_pooling(features, query):
  out = multihead_attention(features, query)
  return layer_norm(out)

img_feature = vit_encoder(image) # [batch, seq_len, dim]
con_feature = attentional_pooling(img_feature, con_query) # [batch, 1, dim]
cap_feature = attentional_pooling(img_feature, cap_query) # [batch, N, dim]

ids = concat(text.ids, cls_token_id)
mask = concat(text.mask, zeros_like(cls_token_id)) # unpad cls_token_id
txt_embs = embedding_lookup(ids)
unimodal_out = lm_transformers(txt_embs, mask, cross_attn=None)
multimodal_out = lm_transformers(
    unimodal_out[:, :-1, :], mask, cross_attn=cap_feature)
cls_token_feature = layer_norm(unimodal_out)[:, -1:, :] # [batch, 1, dim]

con_loss = contrastive_loss(con_feature, cls_token_feature)
cap_loss = softmax_cross_entropy_loss(
    multimodal_out, labels=text.labels, mask=text.mask)
```

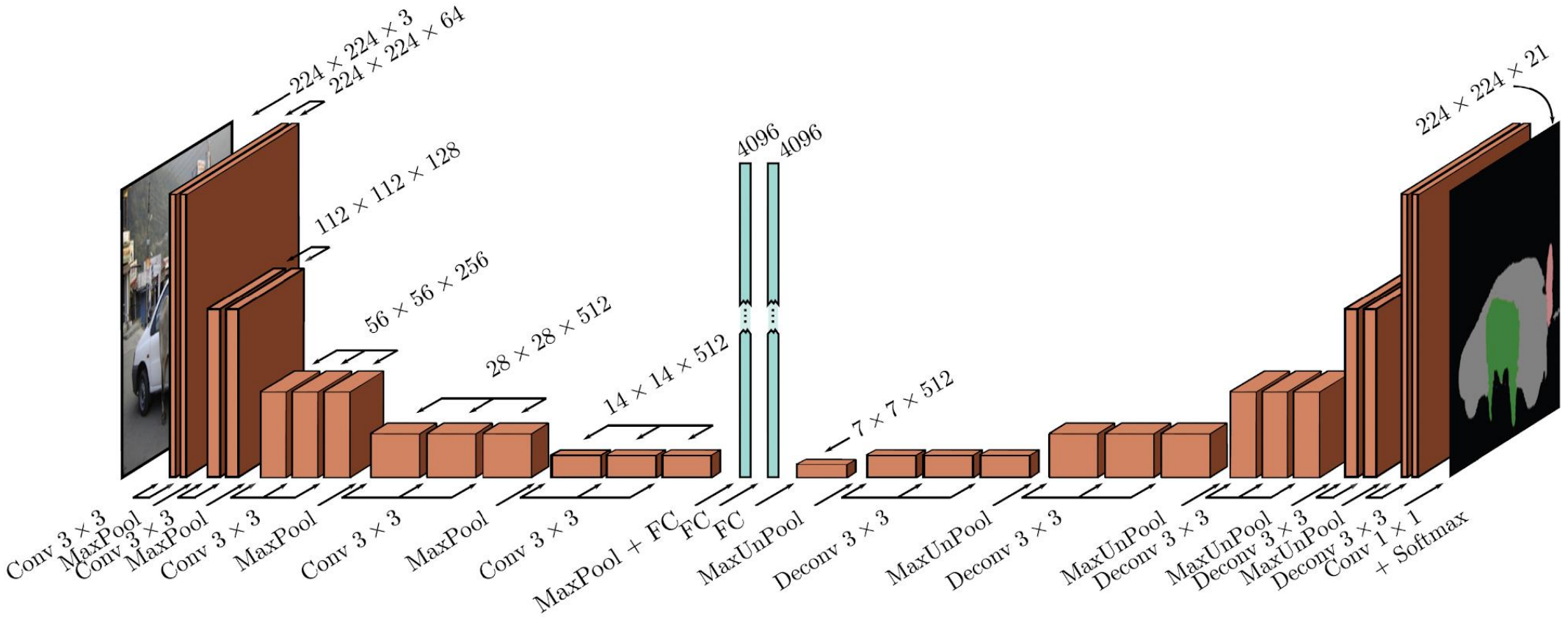
`vit_encoder`: vision transformer based encoder; `lm_transformer`: language-model transformers.

https://arxiv.org/pdf/2205.01917

Semantic segmentation



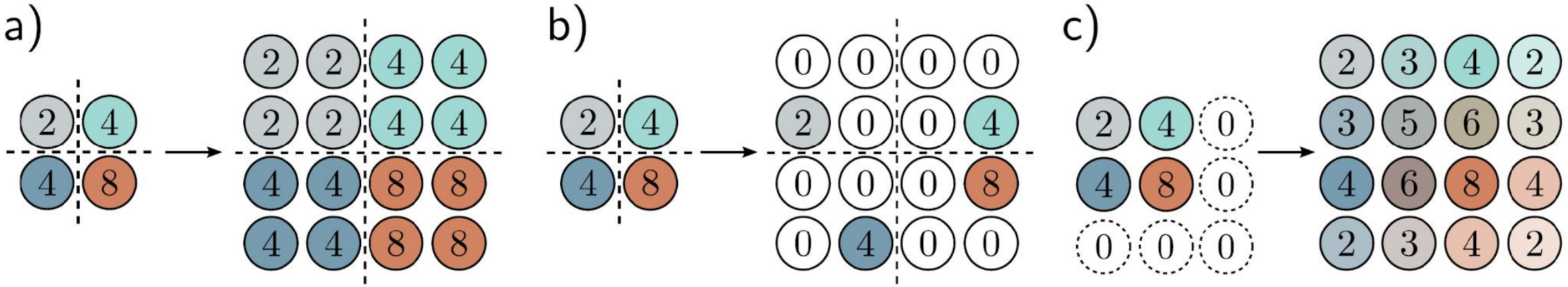
Semantic segmentation – how it started



Encoder

Decoder

Upsampling



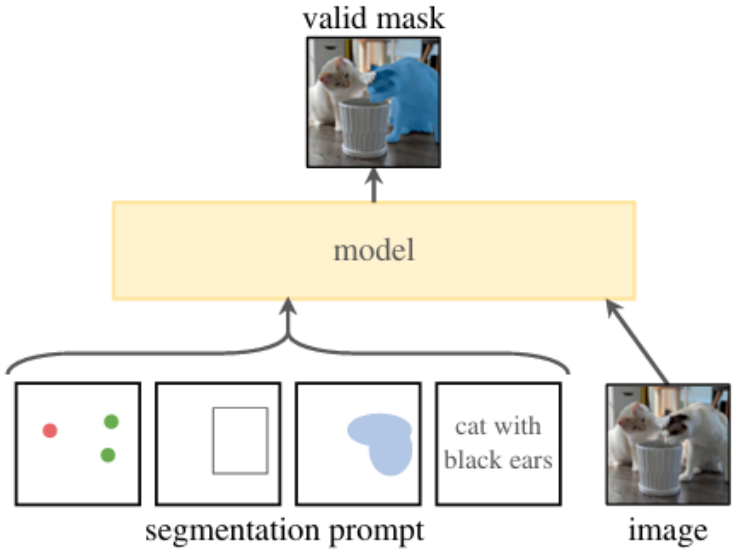
Duplicate

Max-upsampling

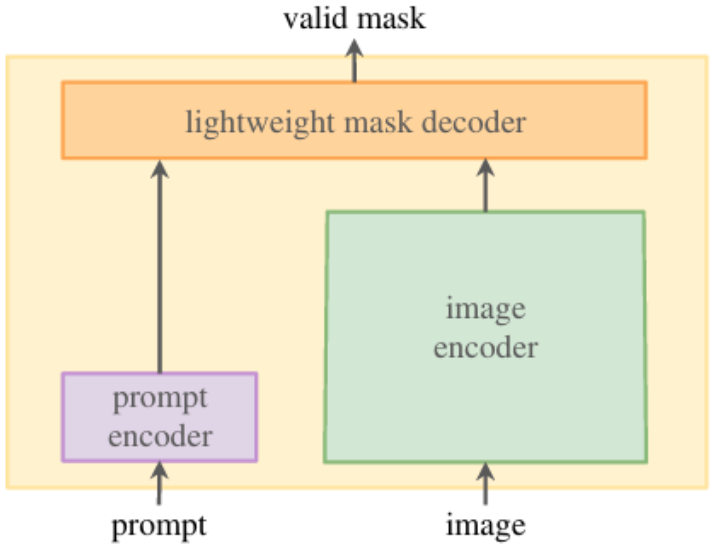
Bilinear interpolation

Semantic segmentation – how it's going

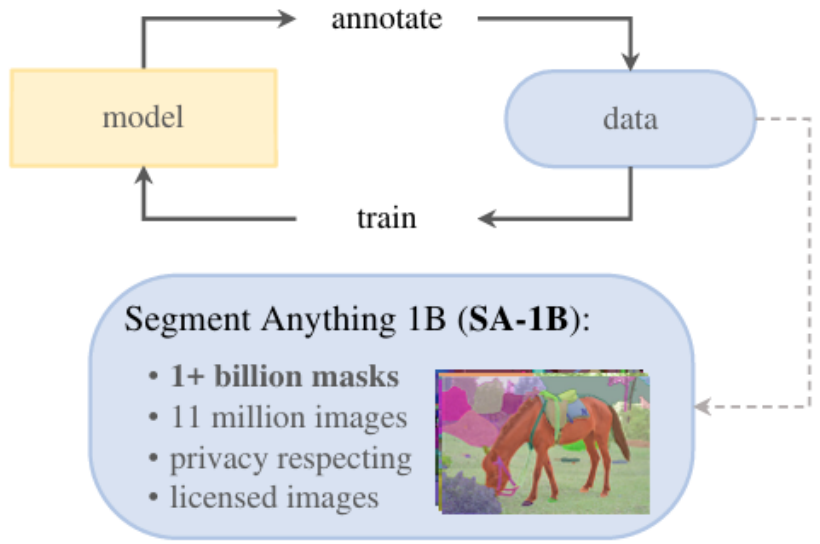
What task will enable zero-shot generalization?



(a) **Task**: promptable segmentation



(b) **Model**: Segment Anything Model (**SAM**)



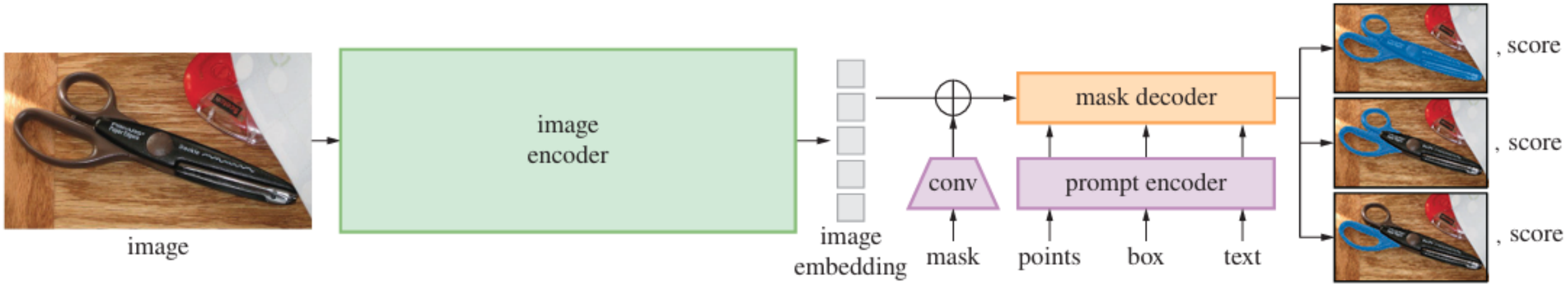
(c) **Data**: data engine (top) & dataset (bottom)

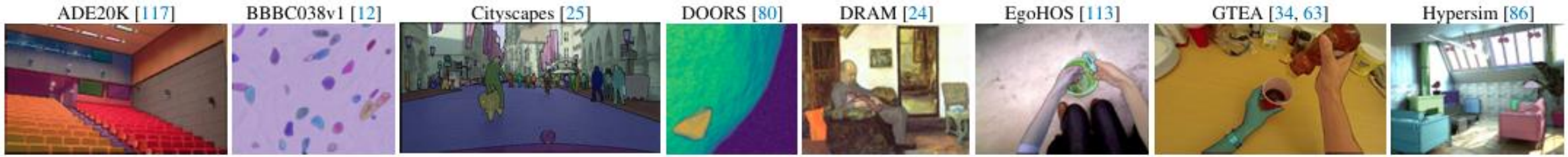
https://arxiv.org/pdf/2304.02643

EIT Oxford

Segment Anything

Zero-shot capabilities





https://arxiv.org/pdf/2304.02643

EIT Oxford

Practical

A Deep Dive into the Vision Transformer (ViT) 👁️

Google Colab link

EIT Oxford

We’d love to hear your feedback!

Scan the QR code or use the link

https://forms.office.com/e/J9y561KuDV

EIT Oxford