Visual Learning
(Tutorial C3)
Transformers and Beyond

Instructor - Simon Lucey

Maths of AI - 2024

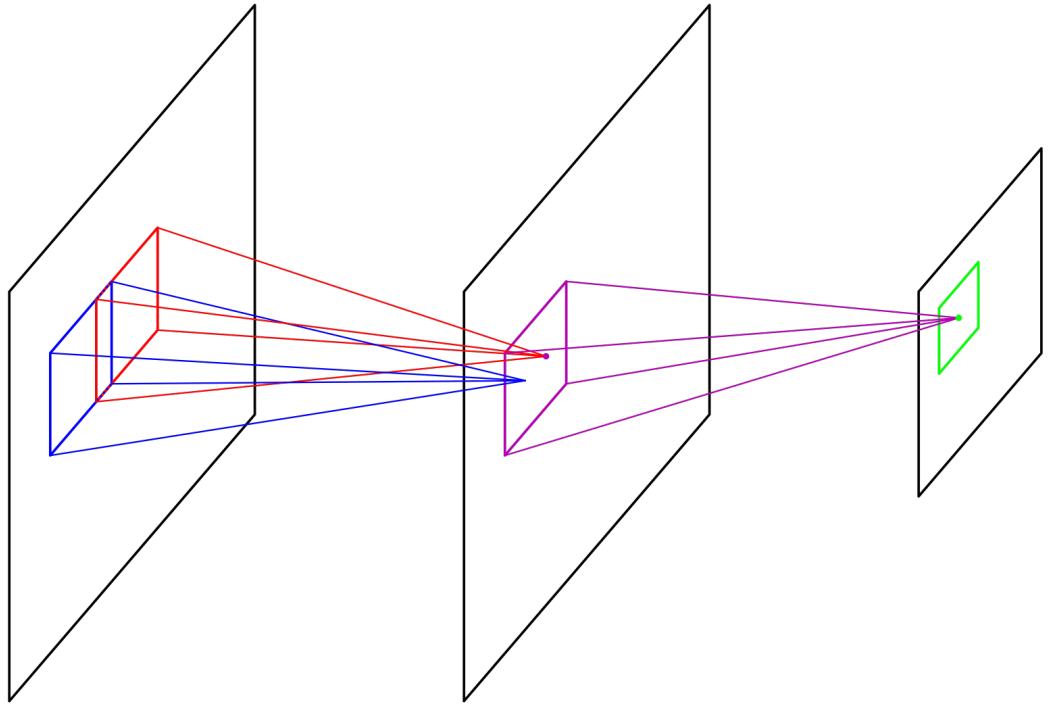
THE UNIVERSITY of ADELAIDE

AUSTRALIAN INSTITUTE FOR MACHINE LEARNING

Today

- Vision Transformers (ViT)
- Masked Self Supervision
- **Graph Neural Networks**

Convolutional = A Type of Attention??



LeCun 1980

Reminder: Depthwise Convolution

image patch
3@
(227x227)

$$\mathbf{Y}^{(j)} = \underbrace{\mathbf{H}^{(j)}}_{\text{“Depthwise Convolution”}} * [\sum_{k=1}^{K} \underbrace{\mathbf{W}[j,k] \cdot \mathbf{X}^{(k)}}_{\text{“1x1 Pointwise Convolution”}}]$$

$$\mathcal{O}(M \cdot N \cdot D^2 \cdot J \cdot K) \rightarrow \mathcal{O}(M \cdot N \cdot (D^2 + J) \cdot K)$$

Reminder: Depthwise Convolution

image patch
3@
(227x227)

Attention is all you need!!

- Convolution is hard-coded attention.
- Could there be any way to learn it from data?

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention

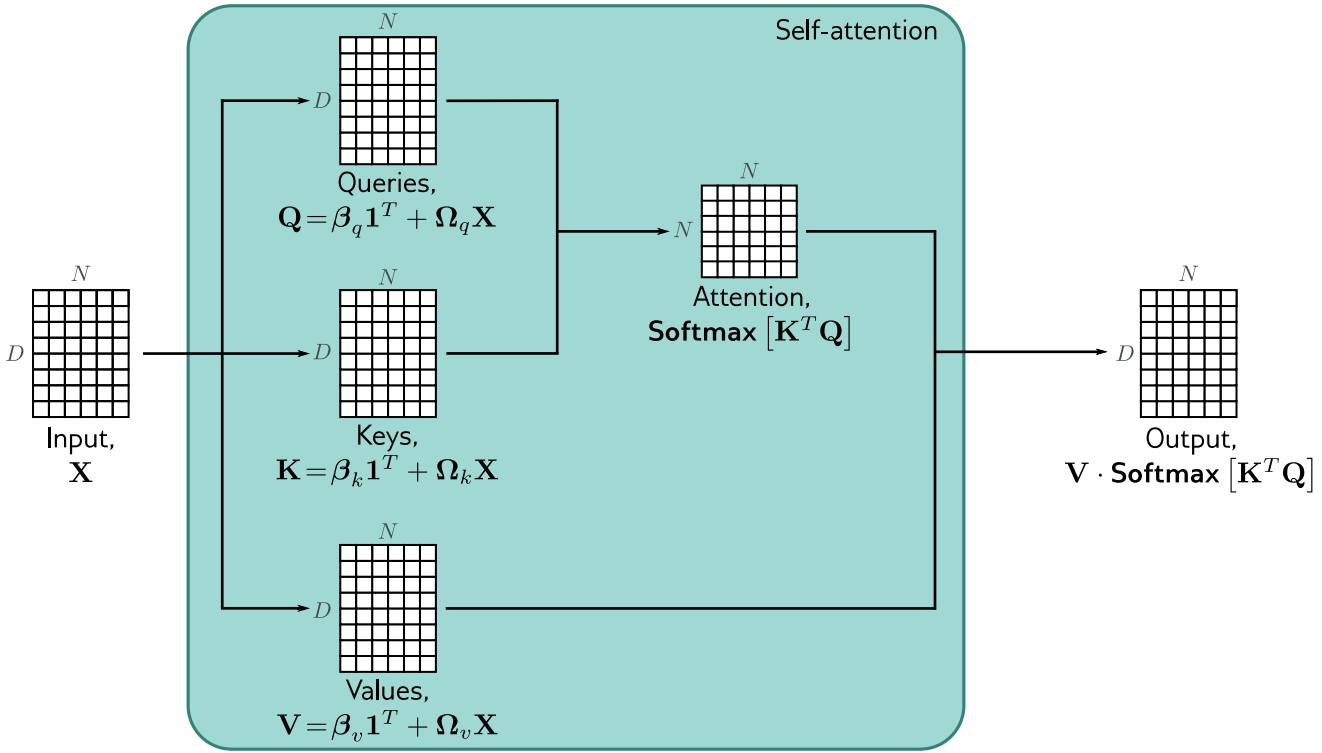
A. Vaswani et al. “Attention is all you need”, NeurIPS 2017.

Soft-Max Attention

$$
\begin{aligned}
a[\mathbf{x}_m, \mathbf{x}_n] &= \text{softmax}_m\left[\mathbf{k}_\bullet^T \mathbf{q}_n\right] \\
&= \frac{\exp\left[\mathbf{k}_m^T \mathbf{q}_n\right]}{\sum_{m'=1}^{N} \exp\left[\mathbf{k}_{m'}^T \mathbf{q}_n\right]},
\end{aligned}
$$

Prince “Understanding Deep Learning” 2023.

What is Attention?



Prince “Understanding Deep Learning” 2023.

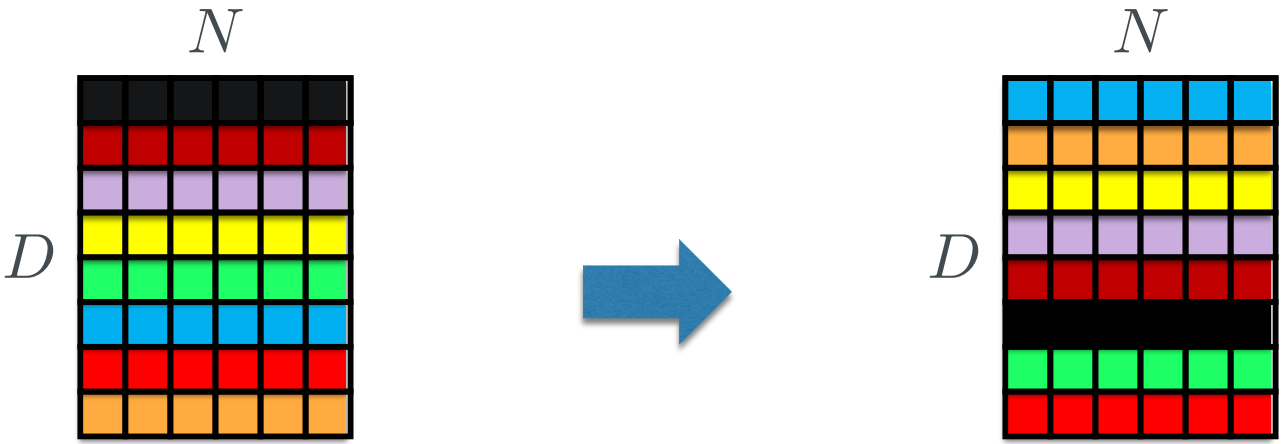
What is Attention?

$$\begin{aligned}\mathbf{V}[\mathbf{X}] &= \boldsymbol{\beta}_v\mathbf{1}^\mathbf{T} + \boldsymbol{\Omega}_v\mathbf{X}\\ \mathbf{Q}[\mathbf{X}] &= \boldsymbol{\beta}_q\mathbf{1}^\mathbf{T} + \boldsymbol{\Omega}_q\mathbf{X}\\ \mathbf{K}[\mathbf{X}] &= \boldsymbol{\beta}_k\mathbf{1}^\mathbf{T} + \boldsymbol{\Omega}_k\mathbf{X},\end{aligned}$$

$$\mathbf{Sa}[\mathbf{X}] = \mathbf{V}[\mathbf{X}]\cdot\mathbf{Softmax}\left[\mathbf{K}[\mathbf{X}]^T\mathbf{Q}[\mathbf{X}]\right],$$

Prince “Understanding Deep Learning” 2023.

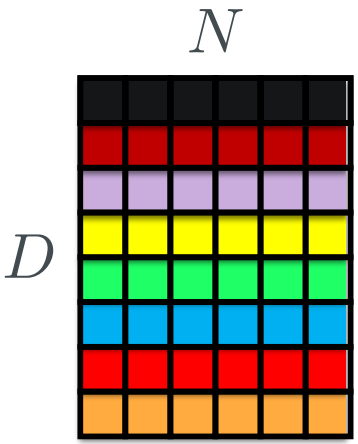
Permutating X



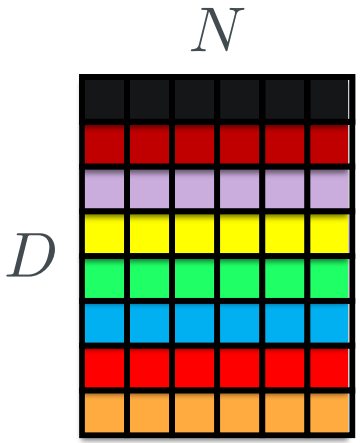
Input,
$\mathbf{X}$

Input,
$\mathbf{XP}$

Permutating X



Input,
$\mathbf{X}$



Input,
$\mathbf{X}\mathbf{P}\mathbf{P}^T$

Permutation Equivariance?

$$\mathbf{Sa}(\mathbf{X}) = \mathbf{Sa}(\mathbf{X}\mathbf{P}\mathbf{P}^T)$$

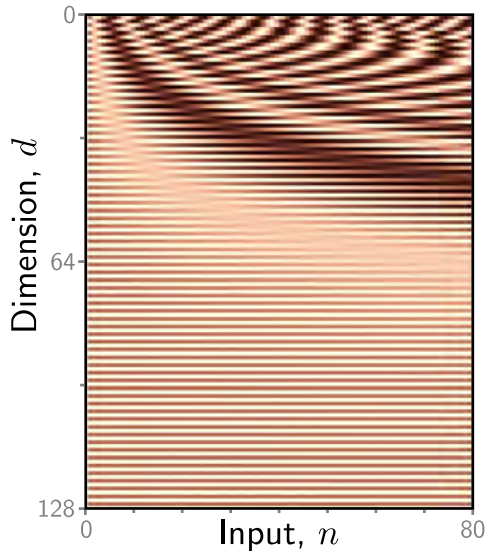
$$\mathbf{Sa}(\mathbf{X}) = \mathbf{Sa}(\mathbf{X}\mathbf{P})\mathbf{P}^T$$

Position information ignored!!!

Positional Encoding

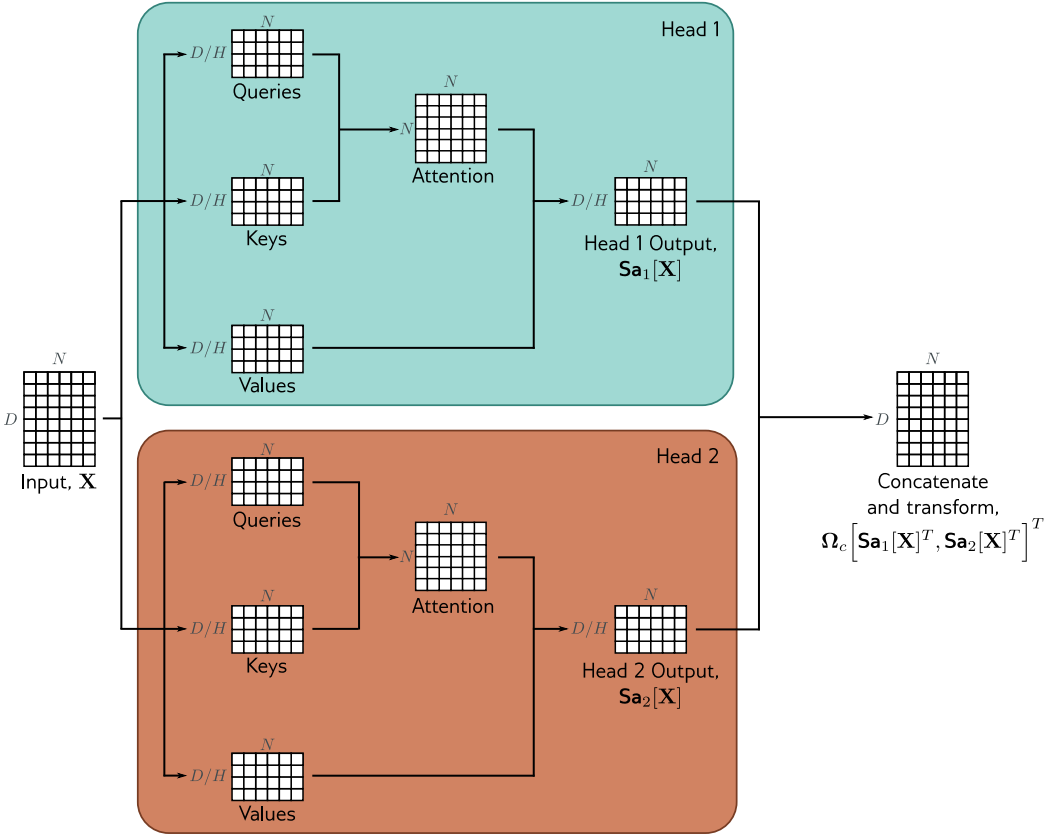
$$\mathbf{X} \rightarrow \mathbf{X} + \mathbf{P}$$

$\mathbf{P} =$



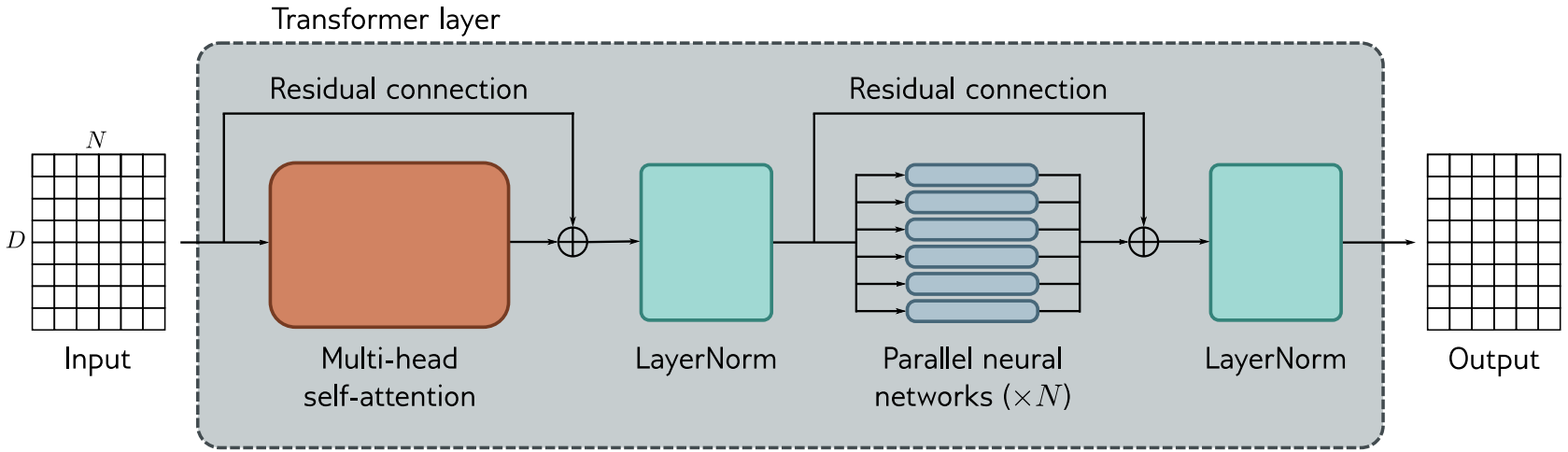
Prince “Understanding Deep Learning” 2023.

Multi-Head Attention



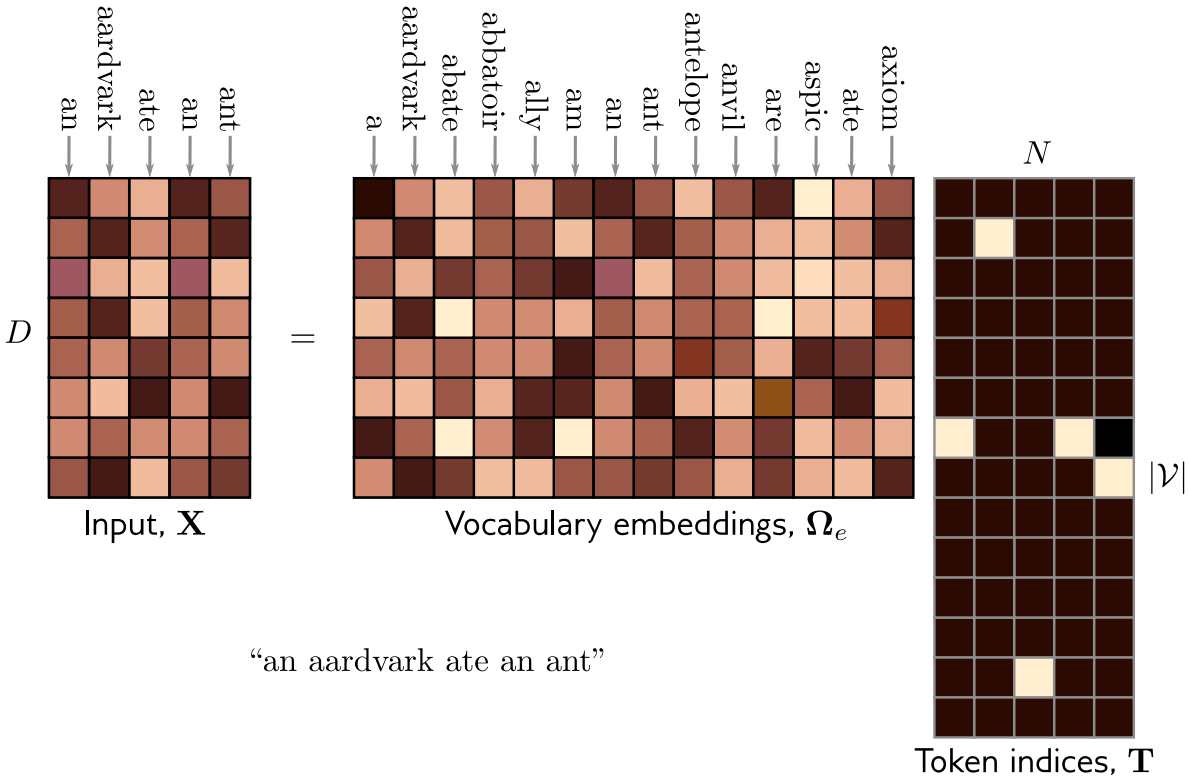
Prince “Understanding Deep Learning” 2023.

Transformer Layer



Prince “Understanding Deep Learning” 2023.

Language + Transformers



"an aardvark ate an ant"

Prince "Understanding Deep Learning" 2023.

Vision Transformer (ViT)

Published as a conference paper at ICLR 2021

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

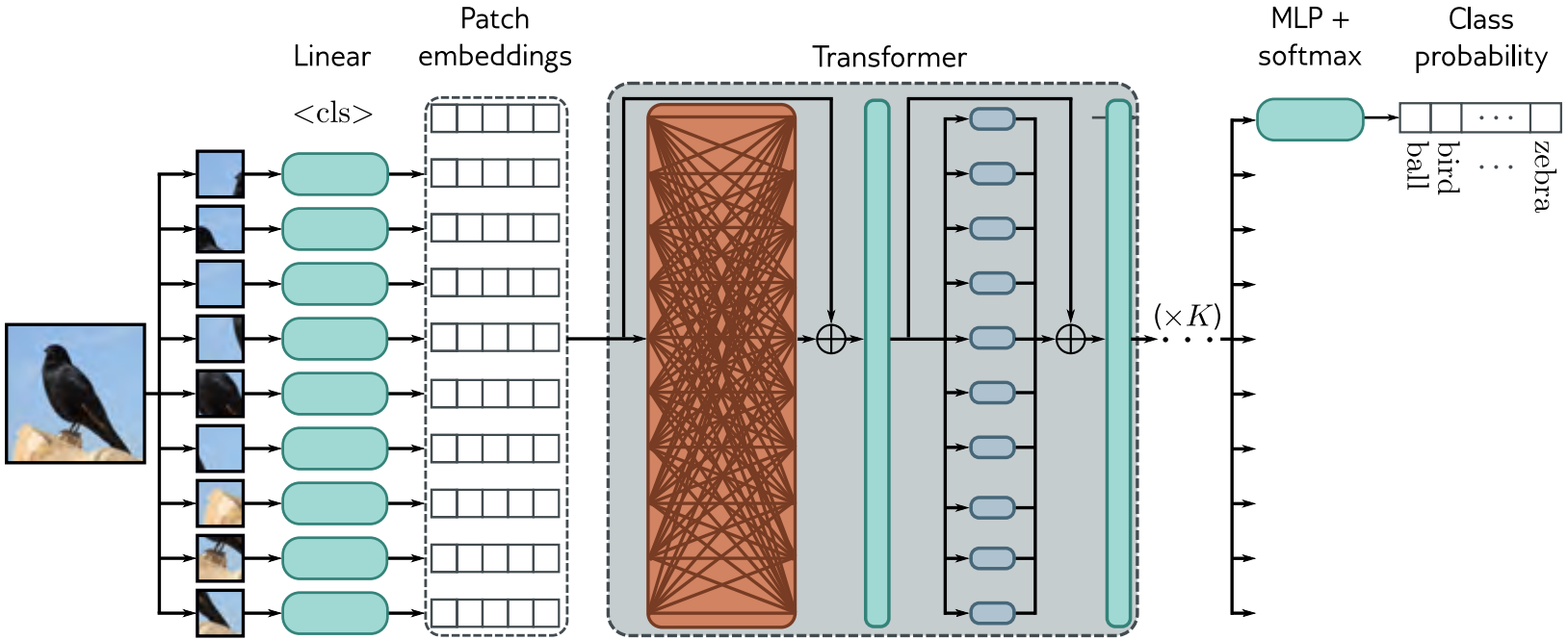
Google Research, Brain Team

{adosovitskiy, neilhoulsby}@google.com

ABSTRACT

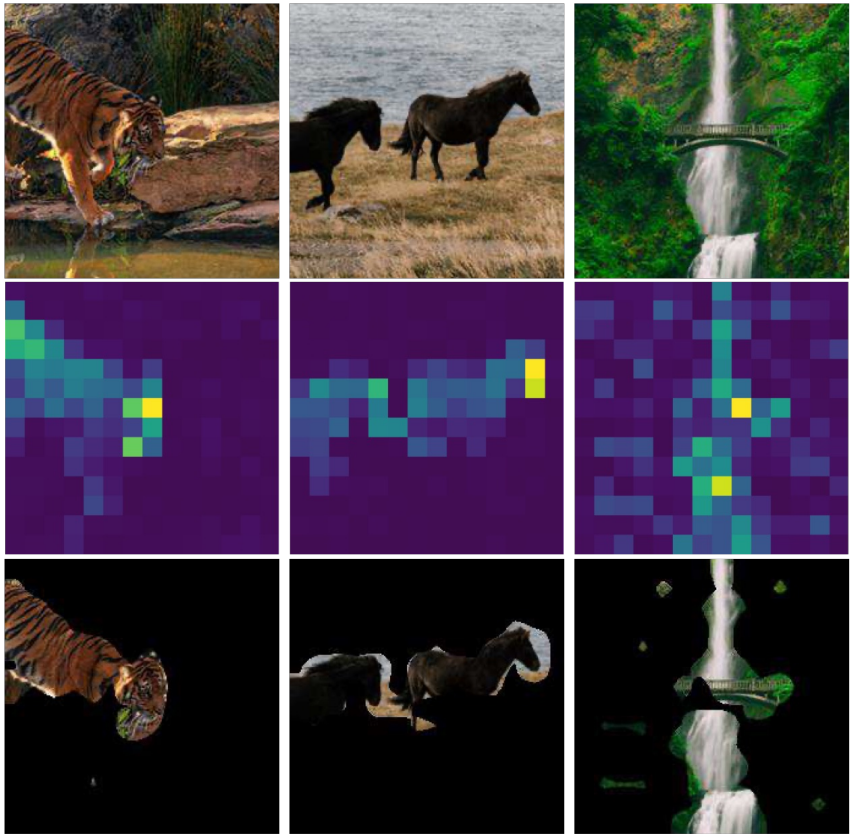
While the Transformer architecture has become the de-facto standard for natural language processing tasks, its applications to computer vision remain limited. In vision, attention is either applied in conjunction with convolutional networks, or used to replace certain components of convolutional networks while keeping their overall structure in place. We show that this reliance on CNNs is not necessary and a pure transformer applied directly to sequences of image patches can perform very well on image classification tasks. When pre-trained on large amounts of data and transferred to multiple mid-sized or small image recognition benchmarks (ImageNet, CIFAR-100, VTAB, etc.), Vision Transformer (ViT) attains excellent

Vision Transformer (ViT)

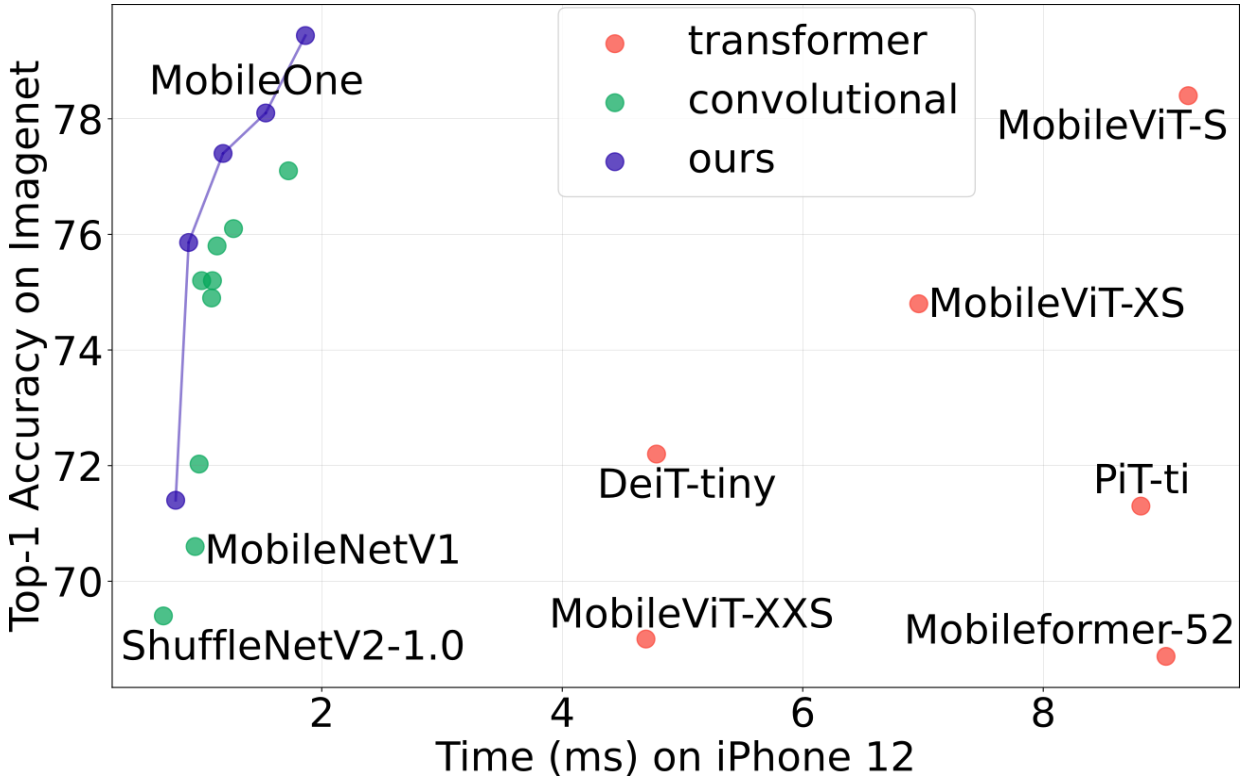


Prince “Understanding Deep Learning” 2023.

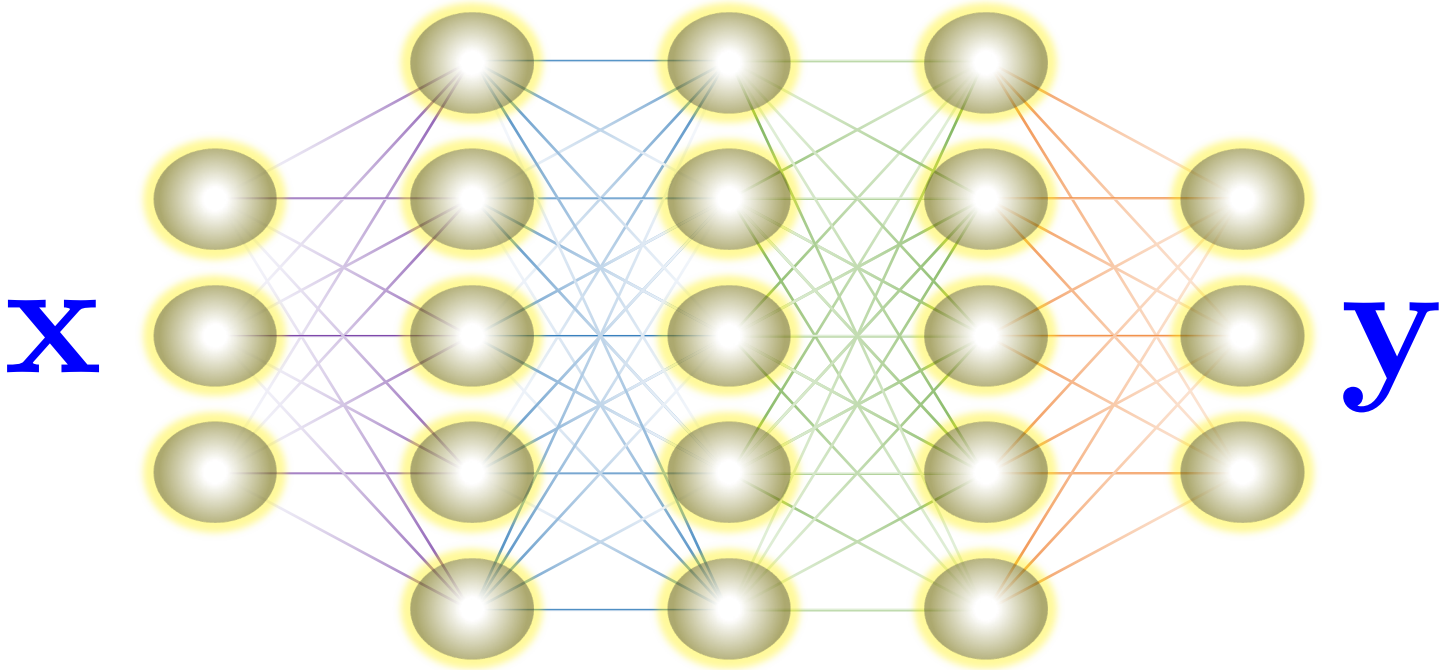
Visualizing Attention

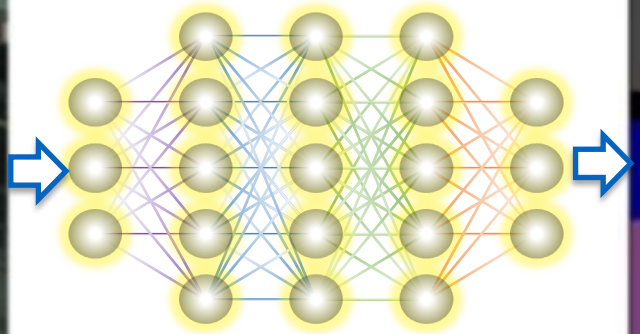


Transformers are Getting Faster!!!



Vasu et al. “MobileOne: An Improved One millisecond Mobile Backbone” In CVPR 2023.

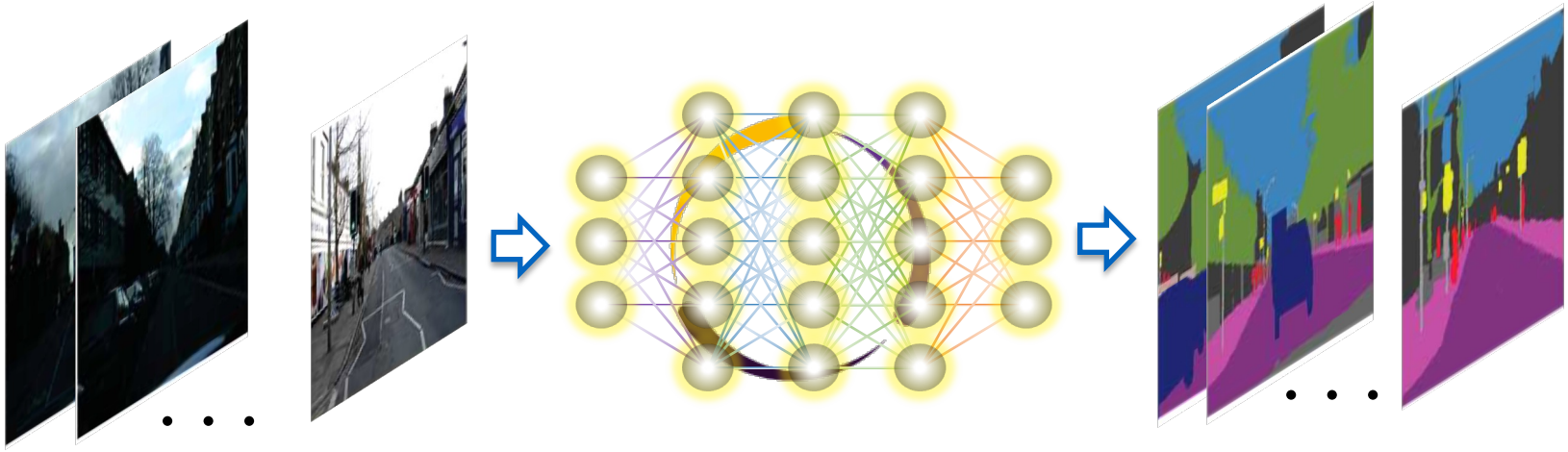

x
y



x

y

Where do I get the labels?

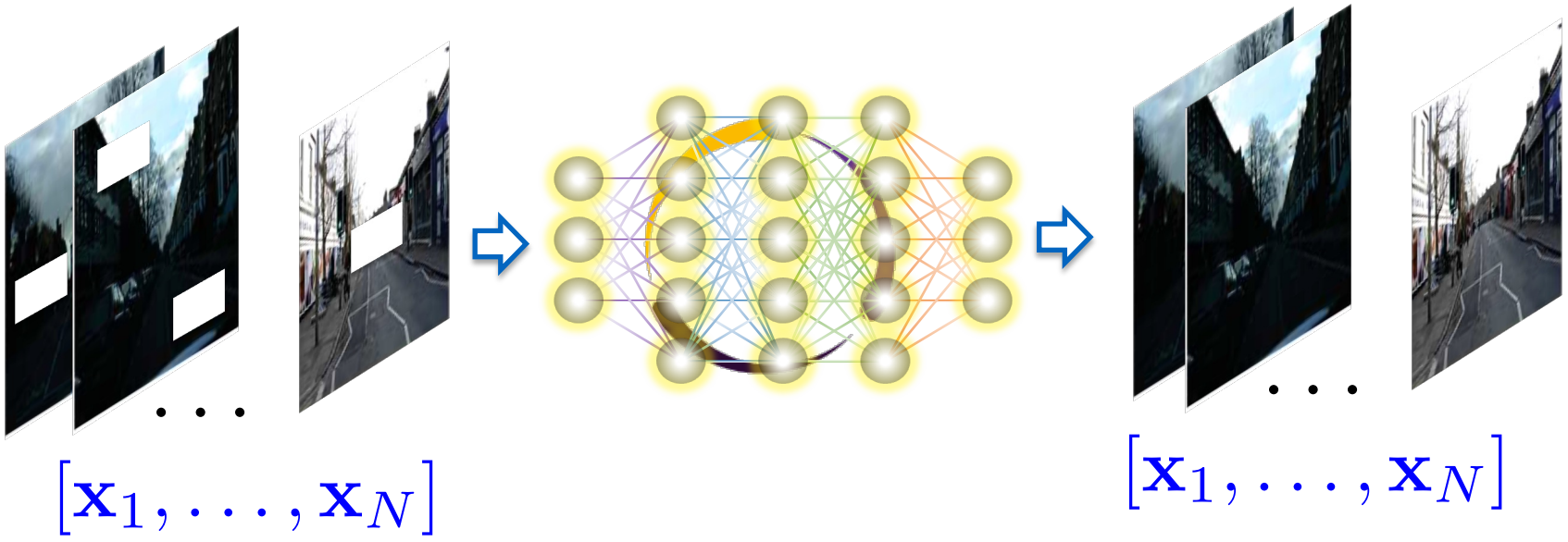


$[\mathbf{x}_1, \ldots, \mathbf{x}_N]$

$[\mathbf{y}_1, \ldots, \mathbf{y}_N]$

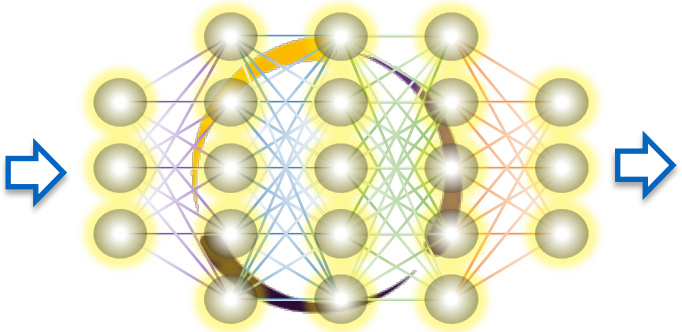
$$N \propto 10^6 - 10^9$$

Requires huge datasets and compute!!!



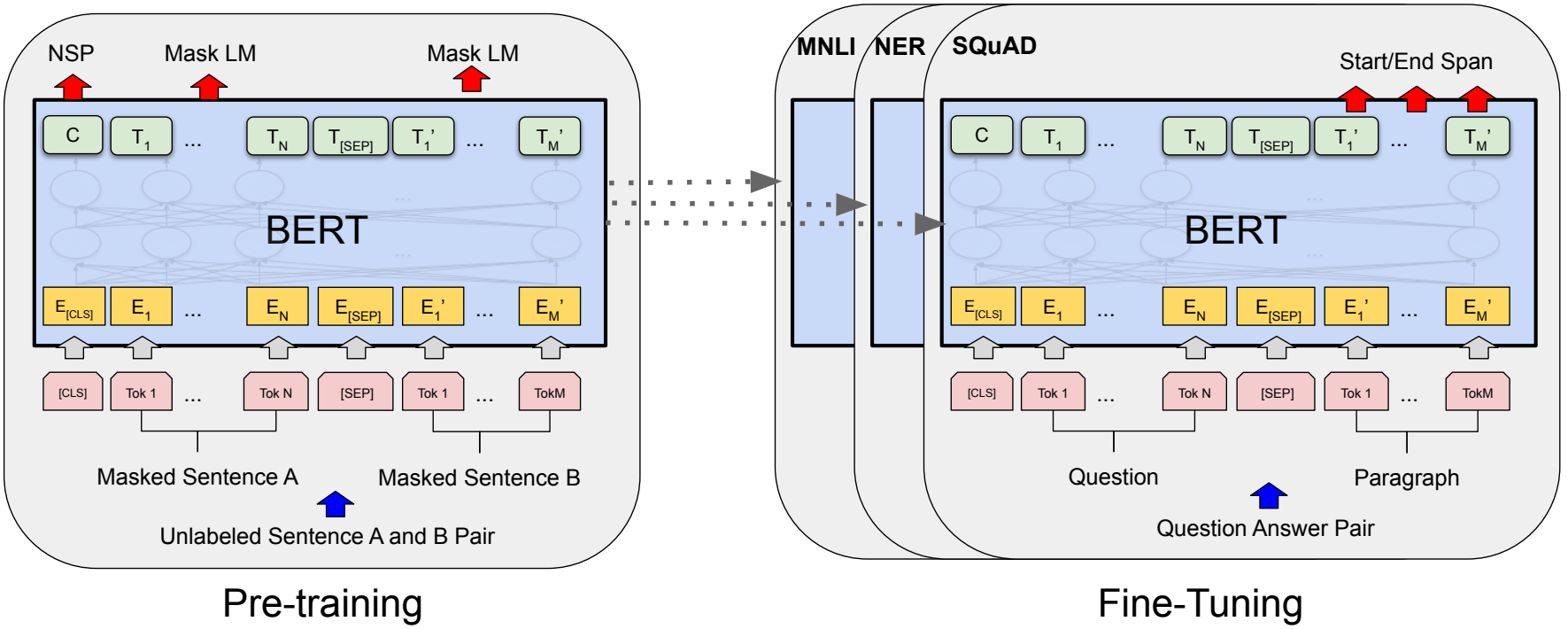
Self Supervision

“The [redacted] walked across the [redacted]”



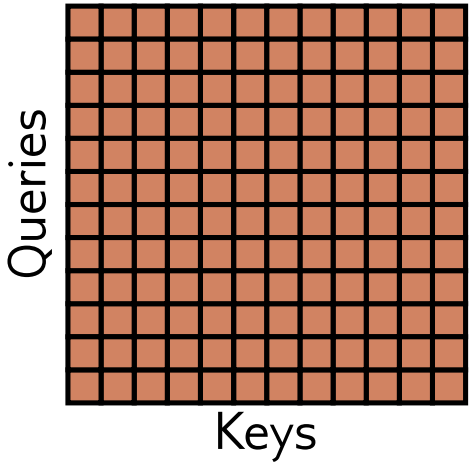
“The cat walked across the room”

Self Supervision

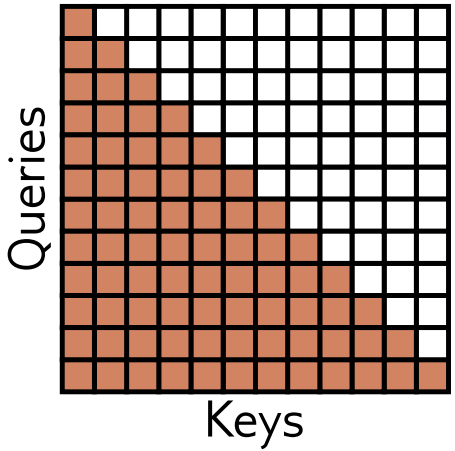


J. Devlin, et al. “BERT: Pre-training of deep bidirectional transformers for language understanding”. In NAACL, 2019..

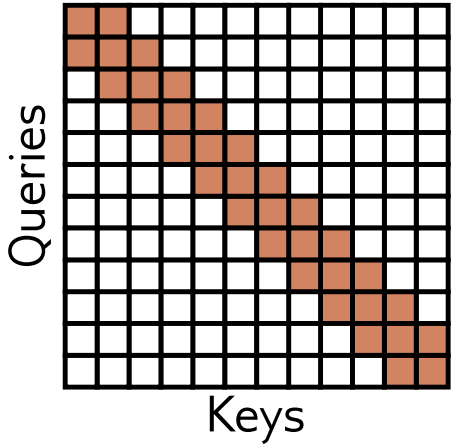
Interaction Matrices for Self Attention



Encoder



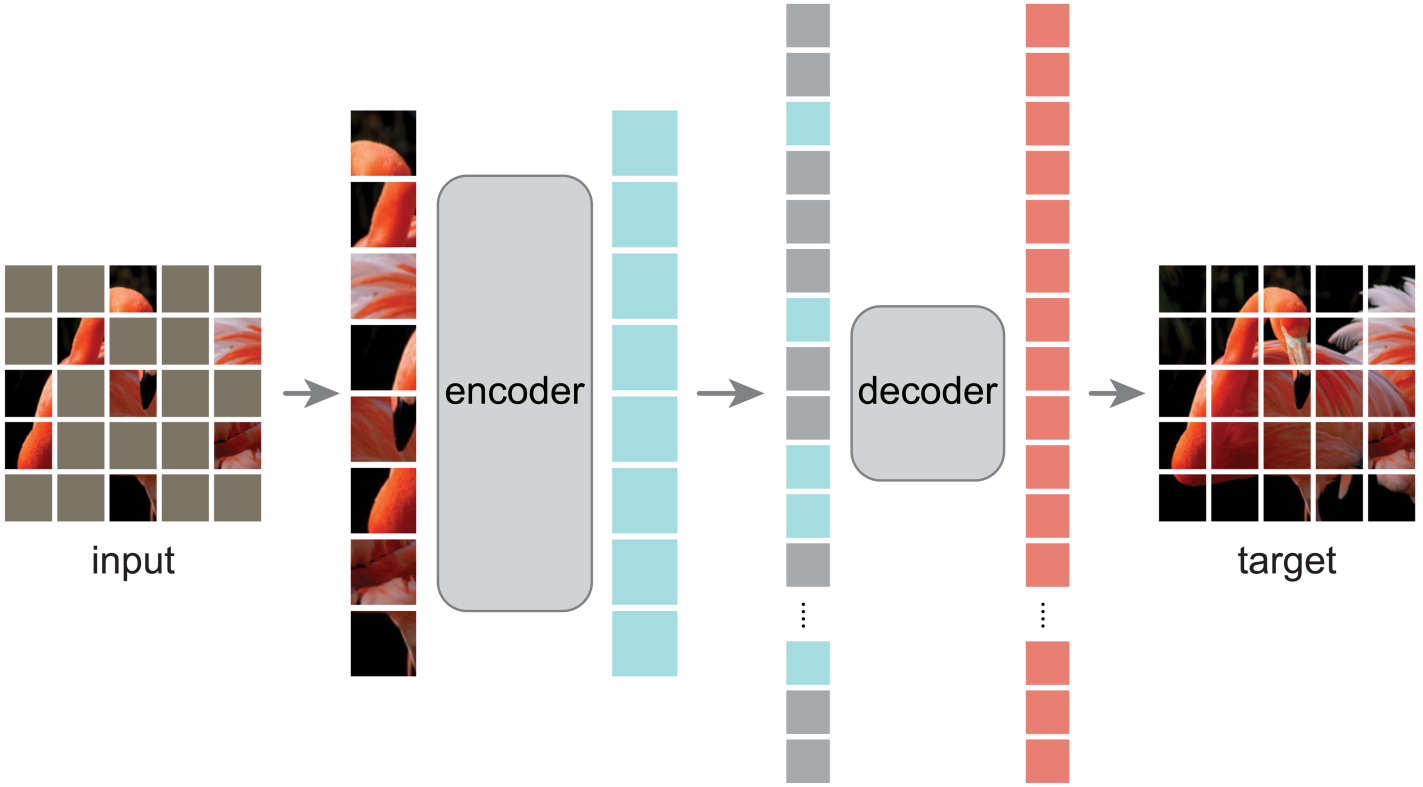
Decoder



Convolutional Encoder

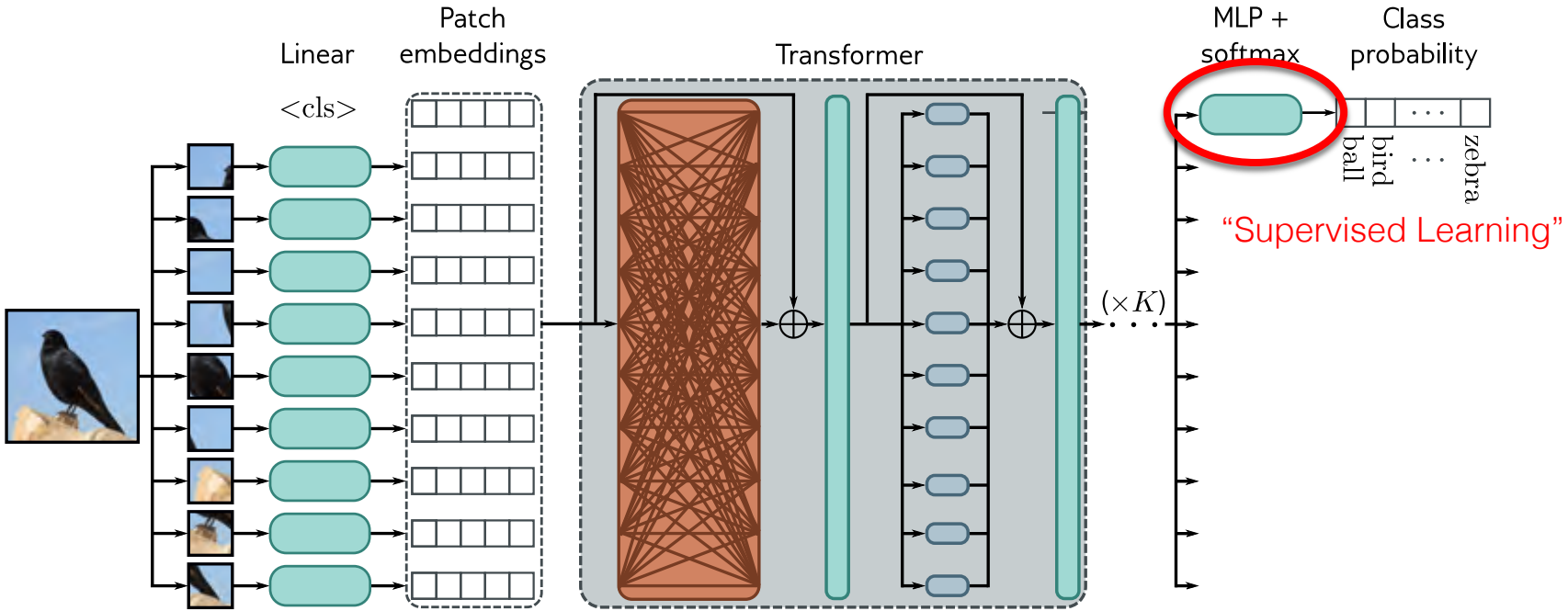
Prince “Understanding Deep Learning” 2023.

Masked ViT



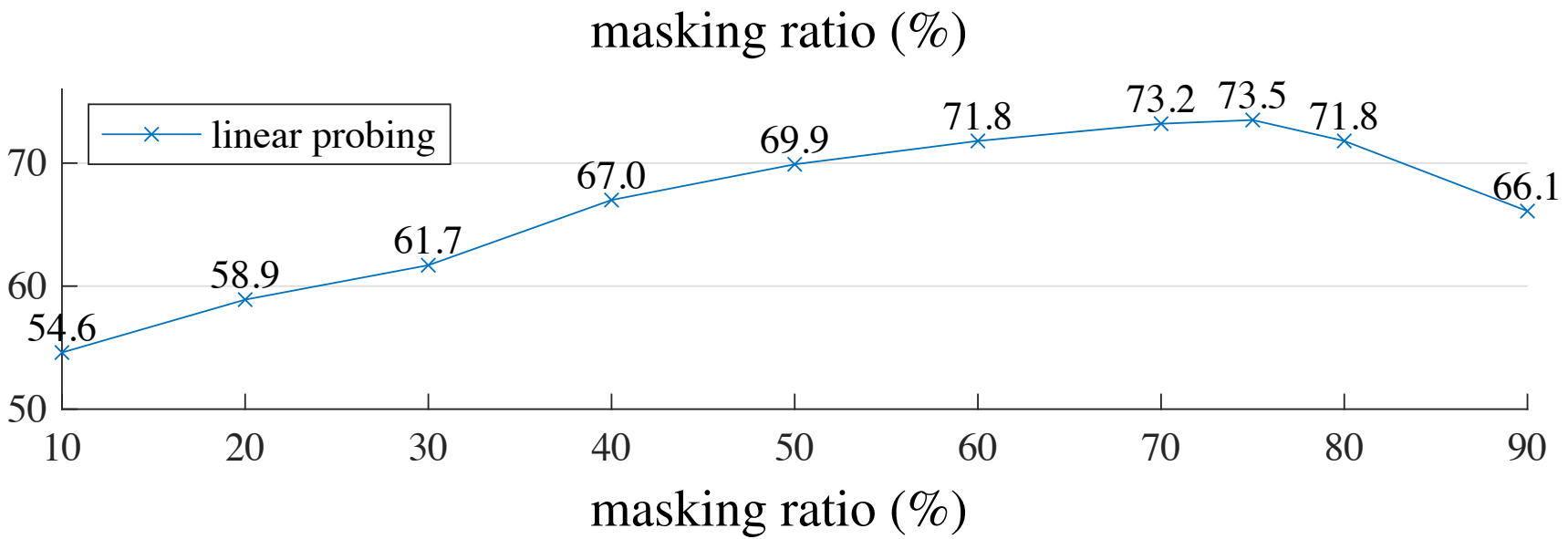
K. He, et al. “Masked Autoencoders Are Scalable Vision Learners” In CVPR 2022.

Masked ViT



Prince "Understanding Deep Learning" 2023.

Masked ViT

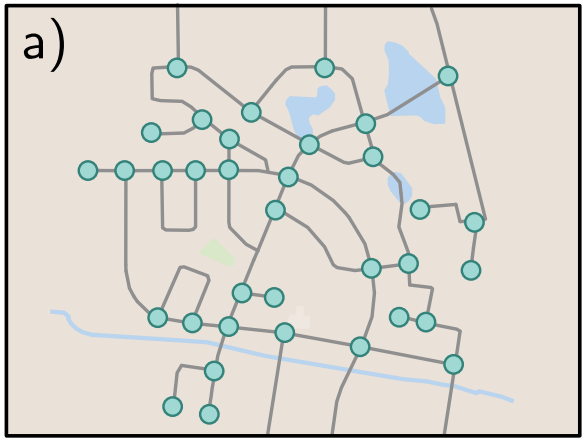


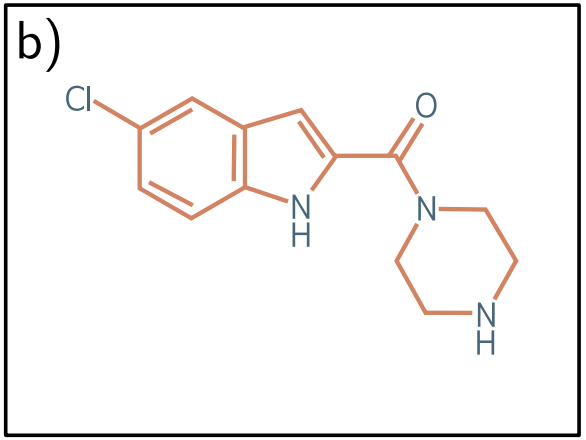
K. He, et al. “Masked Autoencoders Are Scalable Vision Learners” In CVPR 2022.

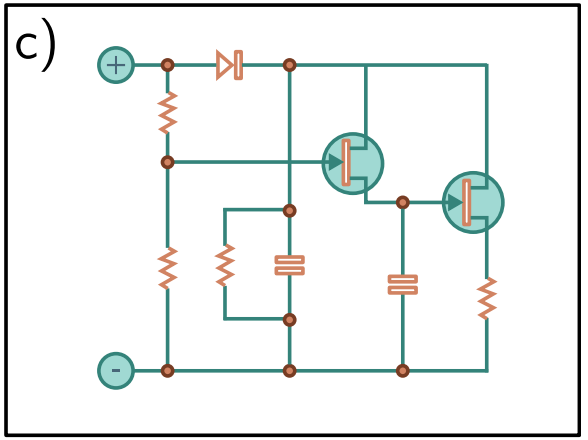
Today

- Vision Transformer Networks
- Self Supervision
- **Graph Neural Networks**

Examples of Graphs?

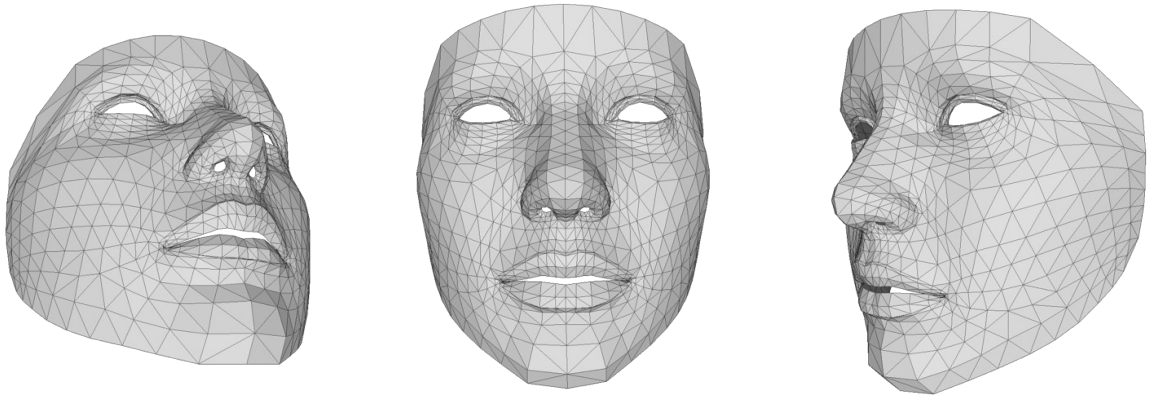






General Topology

What if $\mathbf{x} \neq$ grid?



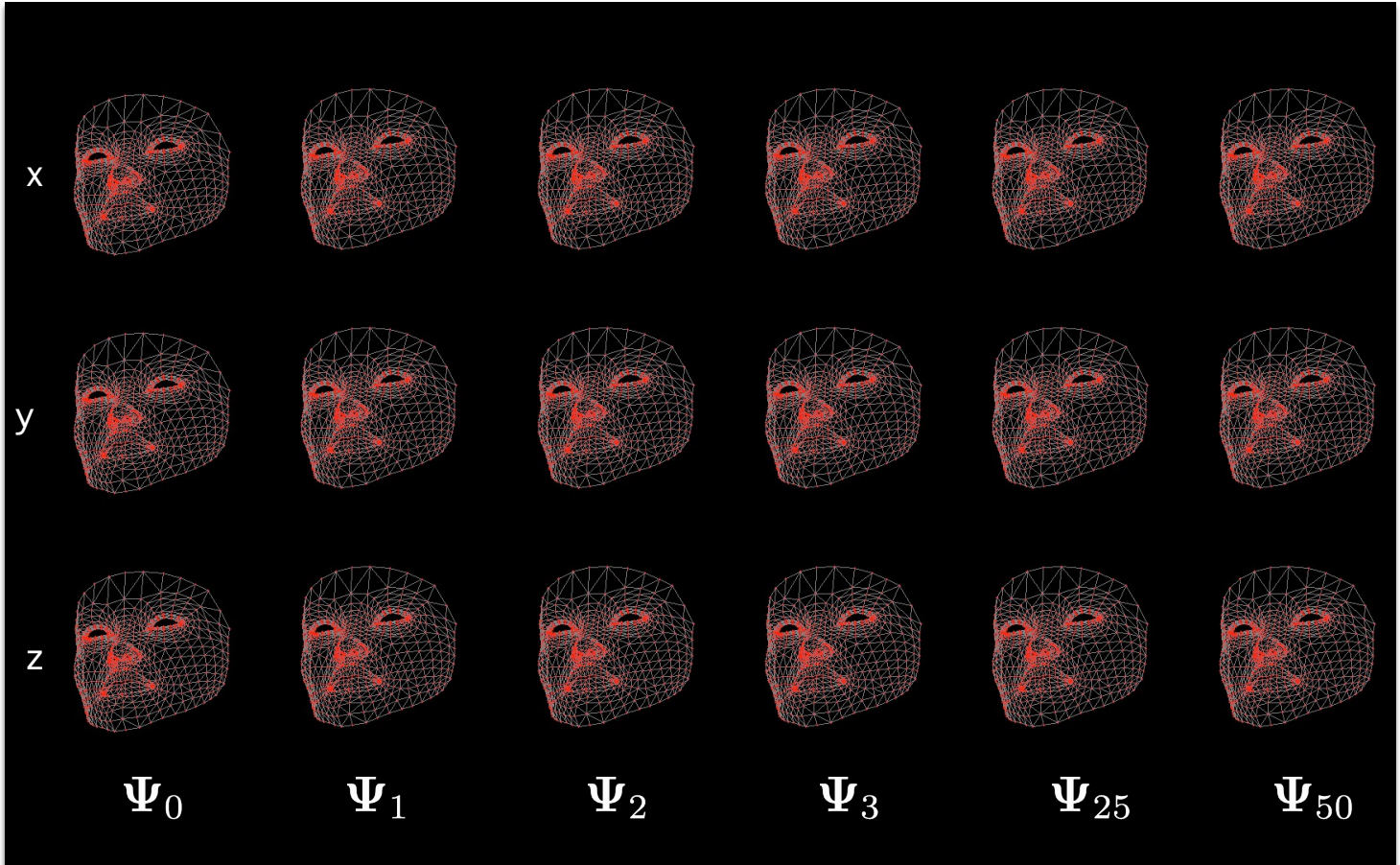
$$\mathbf{L}[i,j] = \begin{cases} |\mathcal{N}(i)| & \text{if } i = j \\ -1 & \text{if } j \in \mathcal{N}(i) \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbf{L} = \mathbf{D} - \mathbf{A}$$

$$\mathbf{L}' = \mathbf{D}^{-\frac{1}{2}}(\mathbf{A} + \mathbf{I})\mathbf{D}^{-\frac{1}{2}}$$

Normalized Graph Laplacian

Smooth Deformation Basis

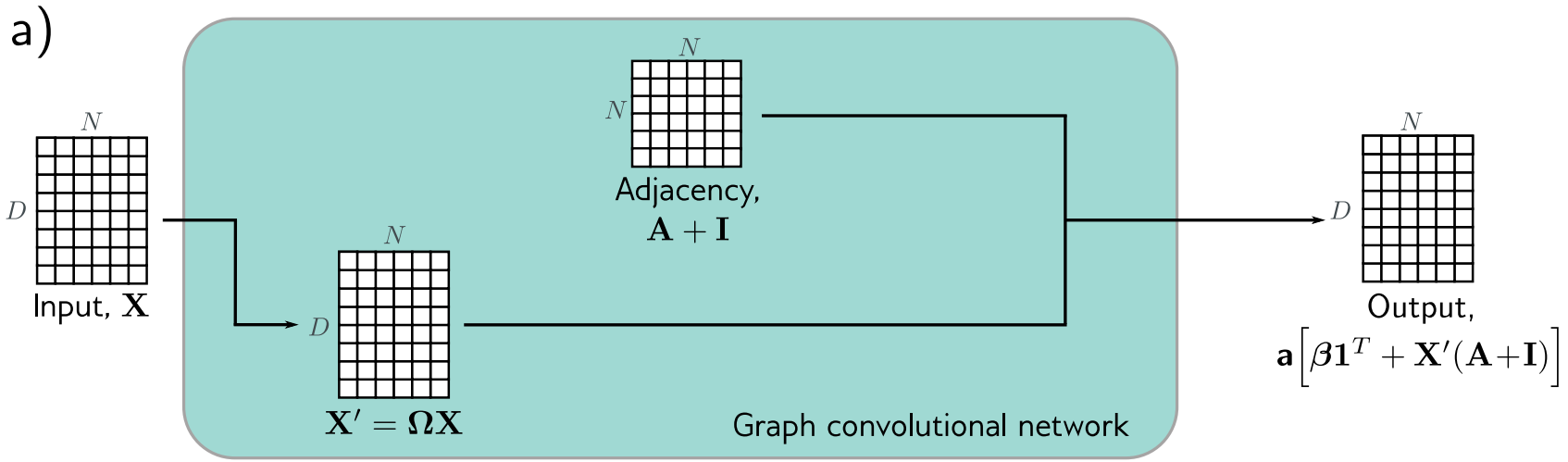


Normalized Graph Laplacian

```
A =      ([[0., 0., 1., 1., 1.],
           [1., 0., 1., 0., 1.],
           [1., 0., 0., 0., 1.],
           [0., 1., 0., 0., 0.],
           [0., 0., 1., 1., 0.]])

L_norm = ([[1.0000, 1.0000, 0.5918, 0.4226, 0.5918],
           [0.6667, 1.0000, 0.5918, 1.0000, 0.5918],
           [0.5918, 1.0000, 1.0000, 1.0000, 0.5000],
           [1.0000, 0.4226, 1.0000, 1.0000, 1.0000],
           [1.0000, 1.0000, 0.5000, 0.2929, 1.0000]])
```

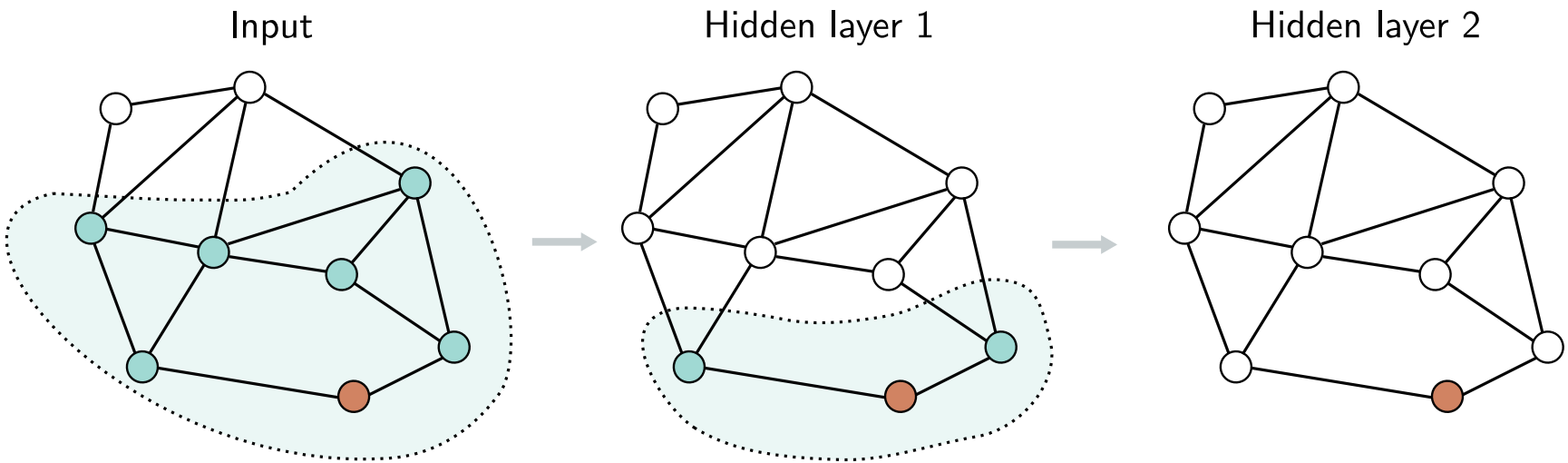
Graph Convolution Network



Only Ω and β are learned!!!

Prince "Understanding Deep Learning" 2023.

Graph Convolution Network



Prince “Understanding Deep Learning” 2023.

Graph Attention Networks

GRAPH ATTENTION NETWORKS

Petar Veličković*
Department of Computer Science and Technology
University of Cambridge
petar.velickovic@cst.cam.ac.uk

Guillem Cucurull*
Centre de Visió per Computador, UAB
gcucurull@gmail.com

Arantxa Casanova*
Centre de Visió per Computador, UAB
ar.casanova.8@gmail.com

Adriana Romero
Montréal Institute for Learning Algorithms
adriana.romero.soriano@umontreal.ca

Pietro Liò
Department of Computer Science and Technology
University of Cambridge
pietro.lio@cst.cam.ac.uk

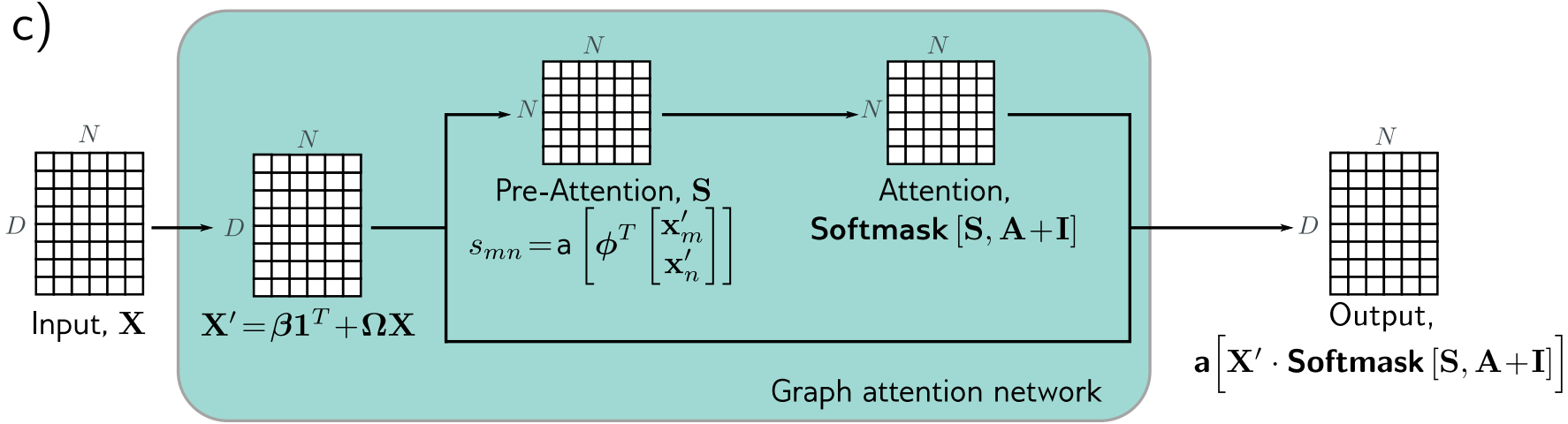
Yoshua Bengio
Montréal Institute for Learning Algorithms
yoshua.umontreal@gmail.com

ABSTRACT

We present graph attention networks (GATs), novel neural network architectures that operate on graph-structured data, leveraging masked self-attentional layers to address the shortcomings of prior methods based on graph convolutions or their approximations. By stacking layers in which nodes are able to attend over their neighborhoods' features, we enable (implicitly) specifying different weights to different nodes in a neighborhood, without requiring any kind of costly matrix operation (such as inversion) or depending on knowing the graph structure upfront. In this way, we address several key challenges of spectral-based graph neural networks simultaneously, and make our model readily applicable to inductive as well as transductive problems. Our GAT models have achieved or matched state-of-the-art results across four established transductive and inductive graph benchmarks: the *Cora*, *Citeseer* and *Pubmed* citation network datasets, as well as a *protein-protein interaction* dataset (wherein test graphs remain unseen during training).

A. Velickovic et al. “Graph Attention Networks”, ICLR 2018.

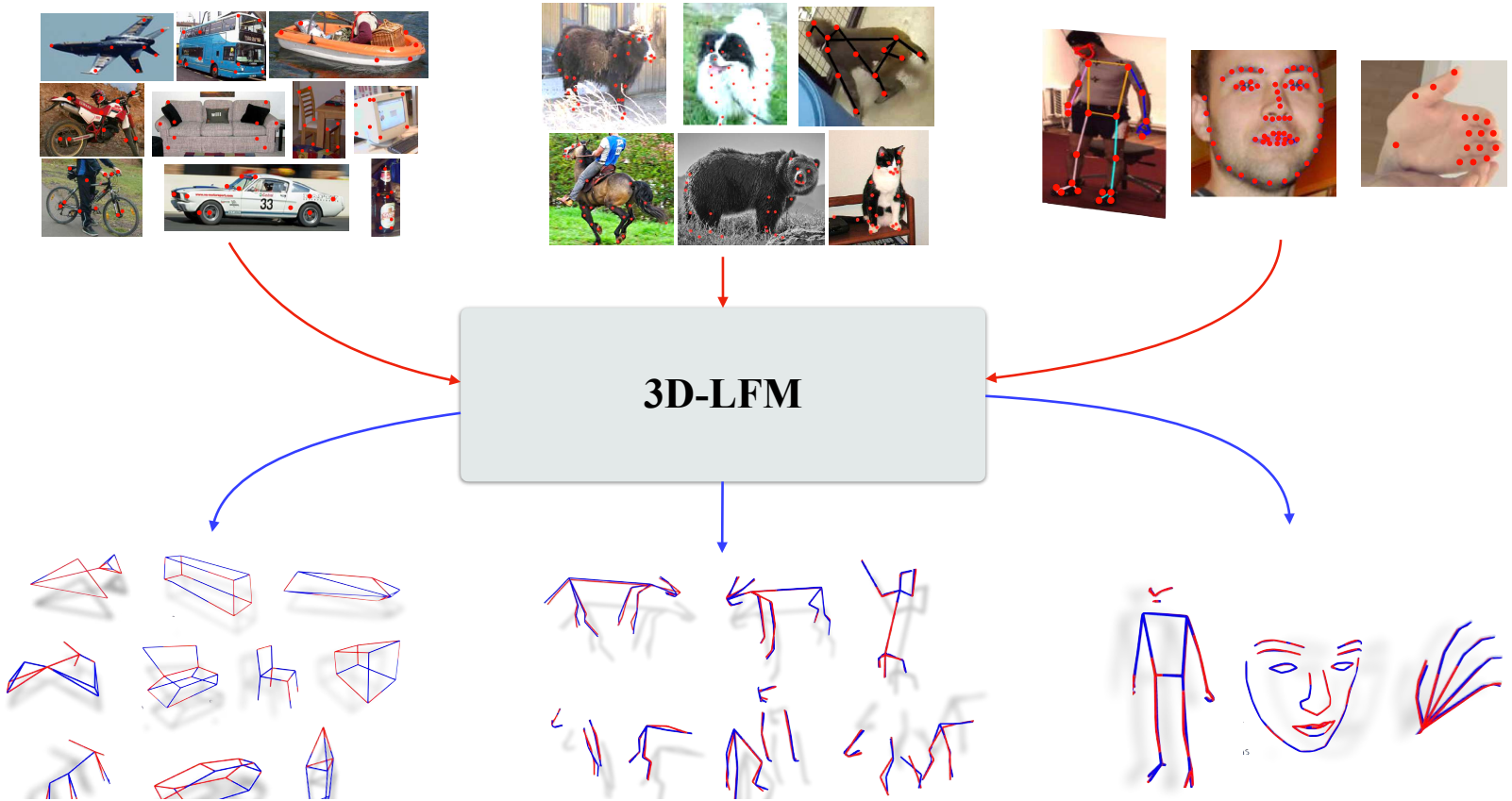
Graph Attention Networks



Ω, ϕ and β are learned!!!

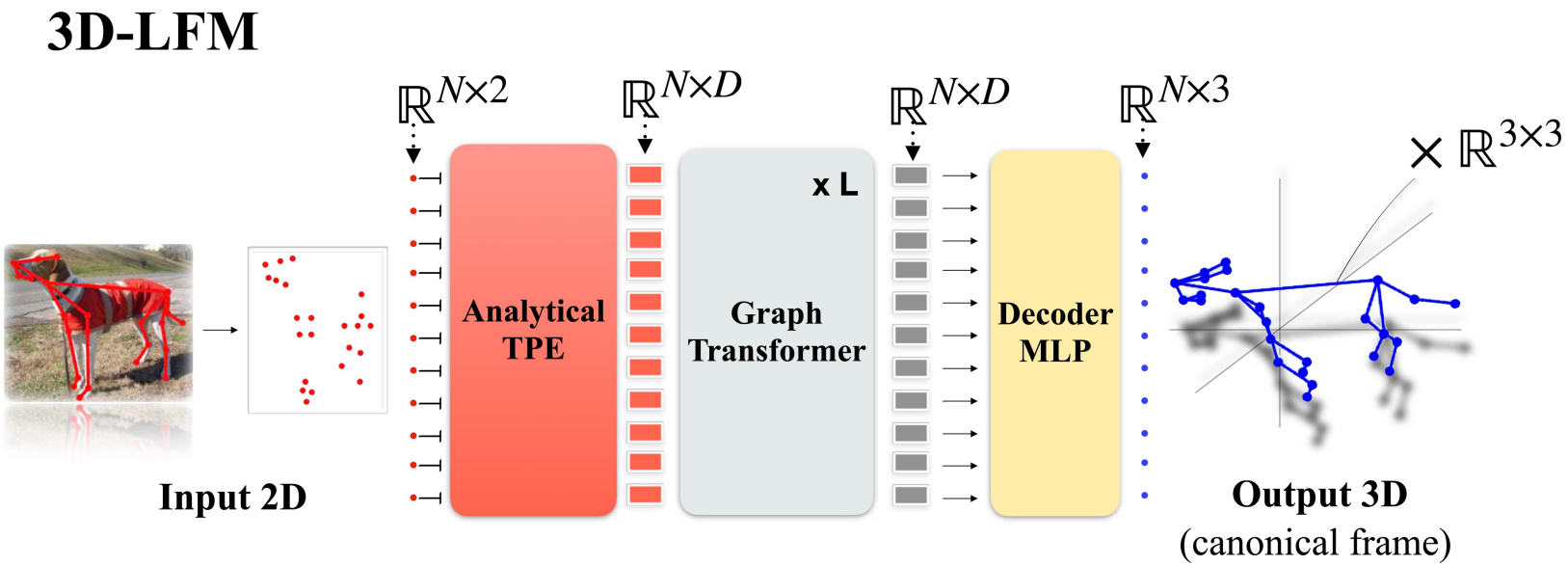
Prince “Understanding Deep Learning” 2023.

Example - 3D Lifting Foundation Model



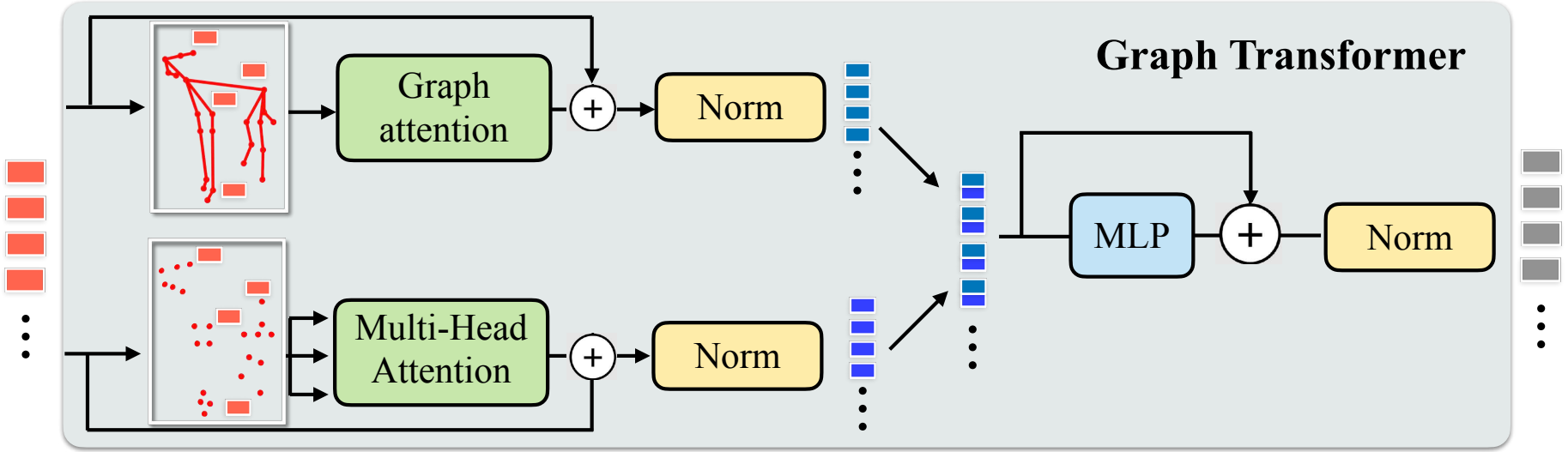
M. Dabhi, L. A. Jeni, and S. Lucey. "3D-LFM: Lifting Foundation Model." arXiv preprint arXiv:2312.11894 (2023).

Example - 3D Lifting Foundation Model



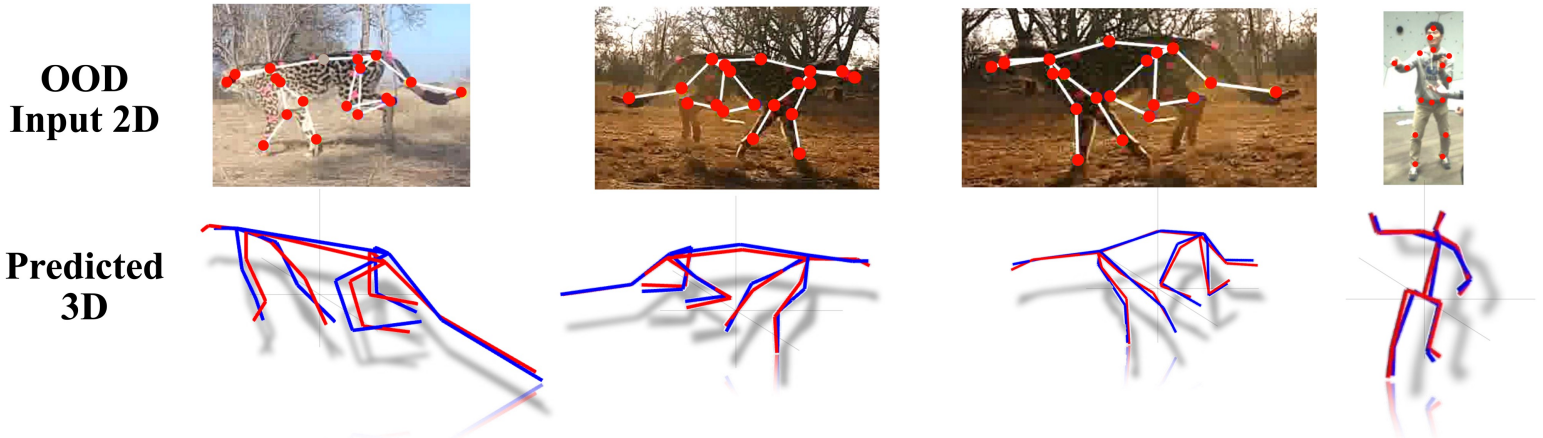
M. Dabhi, L. A. Jeni, and S. Lucey. "3D-LFM: Lifting Foundation Model." arXiv preprint arXiv:2312.11894 (2023).

Example - 3D Lifting Foundation Model

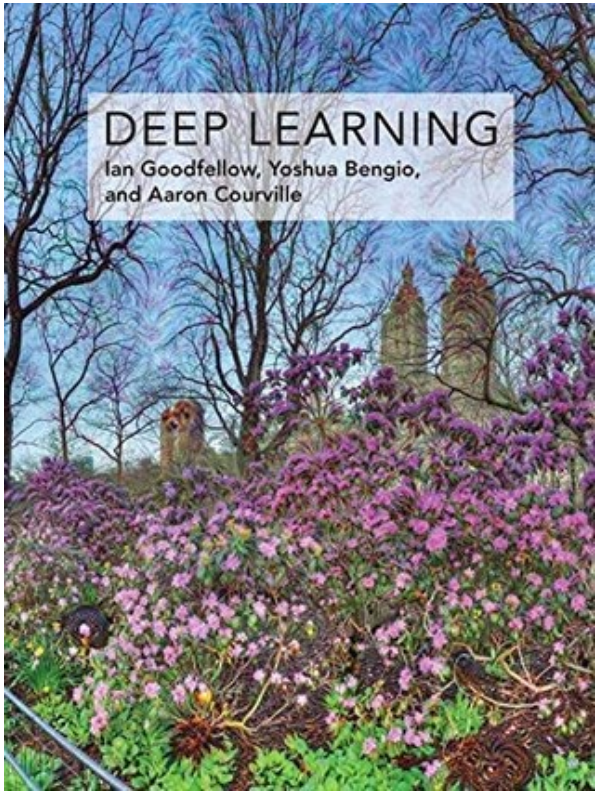


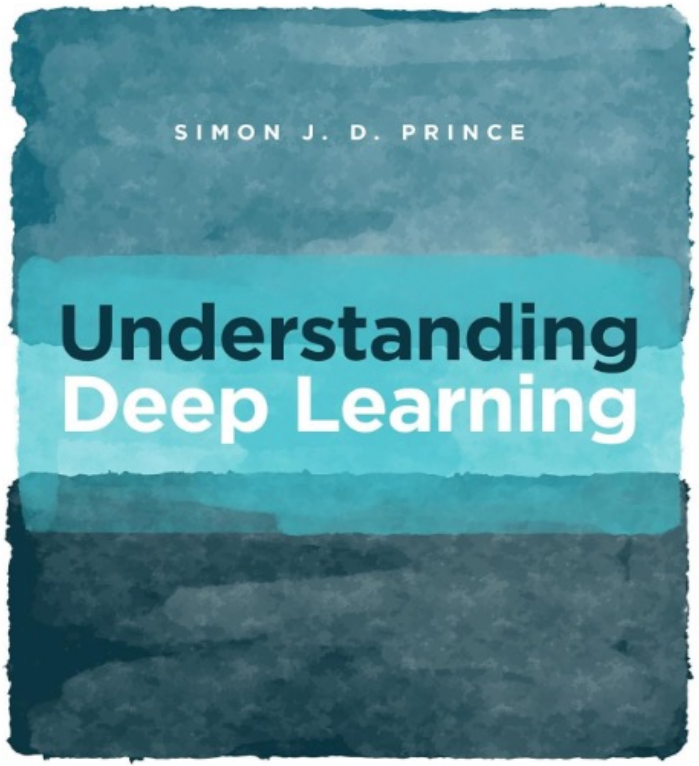
M. Dabhi, L. A. Jeni, and S. Lucey. "3D-LFM: Lifting Foundation Model." arXiv preprint arXiv:2312.11894 (2023).

Example - 3D Lifting Foundation Model



More to read…


DEEP LEARNING
Ian Goodfellow, Yoshua Bengio,
and Aaron Courville


SIMON J. D. PRINCE
Understanding
Deep Learning