SCC0251
Processamento de Imagens

Aprendizado Profundo

Professora Leo Sampaio Ferraz Ribeiro

Slide para não esquecer de passar a lista

Júpiter - Sistema de Gestão Acadêmica da Pró-Reitoria de Graduação

Lista de Presença

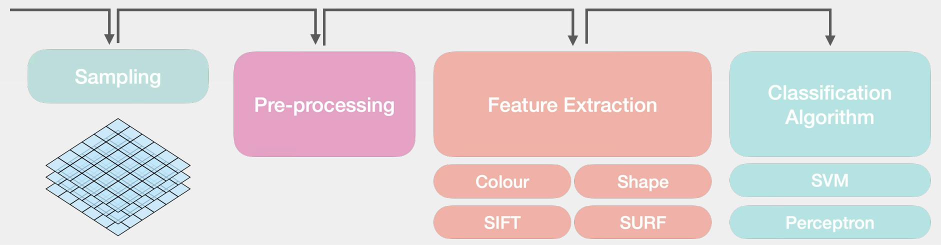
Unidade: 55 Instituto de Ciências Matemáticas e de Computação
Disciplina: SCC0251 Processamento de Imagens
Turma: 2025101 - Teórica
Período: 24/02/2025 - 07/07/2025
Disciplina COM 2ª Avaliação.

Horário	Prof(a).
qua 08:10 09:50	Leo Sampaio Ferraz Ribeiro
sex 08:10 09:50	Leo Sampaio Ferraz Ribeiro

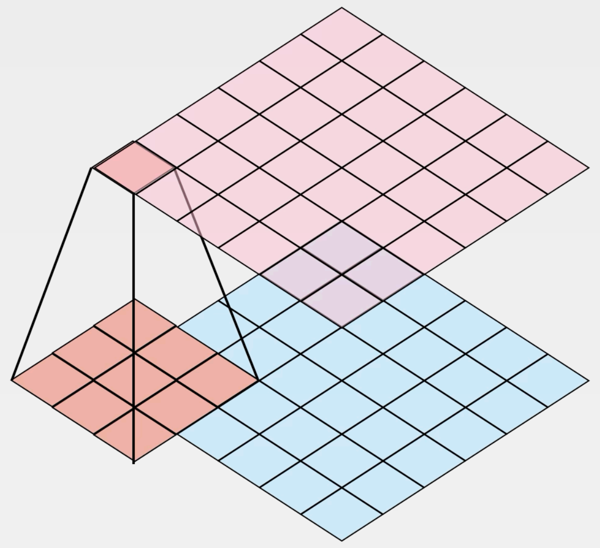
NºUSP	Ingr.	Curso	Nome	dia __/__/___	dia __/__/___	dia __/__/___
14712657	28/02/2024	55041	Allan Vitor de Souza Silva			
13687196	11/02/2022	55071	Amabile Pietrobon Ferreira			
13687108	23/02/2022	55090	Arthur Hiratsuka Rezende			
12691964	13/03/2023	55041	Arthur Pin			
13671532	11/02/2022	55041	Arthur Queiroz Moura			
12745212	03/05/2021	97001	Asafe Henrique de Oliveira Franca			
12542481	16/04/2021	55041	Bernardo Maia Coelho			
12733212	29/04/2021	55041	Bernardo Rodrigues Tameirao Santos			
14745682	13/03/2023	55071	Bruno Batista Pereira da Silva			
13672220	25/03/2022	55041	Camila Donda Ronchi			
12542630	18/03/2021	55041	Carlos Filipe de Castro Lemos			
14746015	24/02/2025	55090	Diego Gladcheff Munhoz			
12556973	25/02/2022	55041	Eduarda Fritzen Neumann			
14568142	27/01/2023	55090	Enzo Castelo Branco Biondi			
13781841	07/03/2022	55041	Enzo Yasuo Hirano Harada			
12547423	13/03/2023	55041	Fabricio Sampaio			

SCC0251
Processamento de Imagens

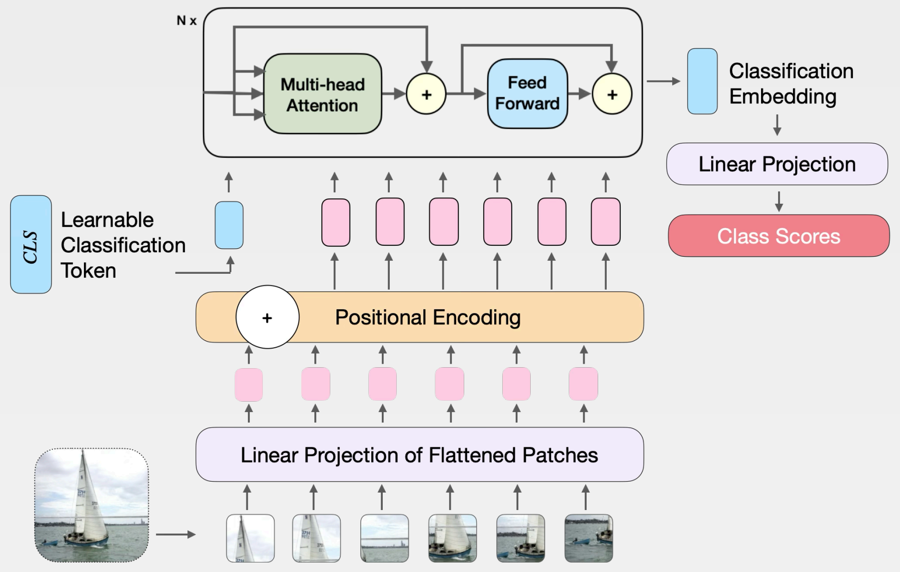
Aprendizado Profundo



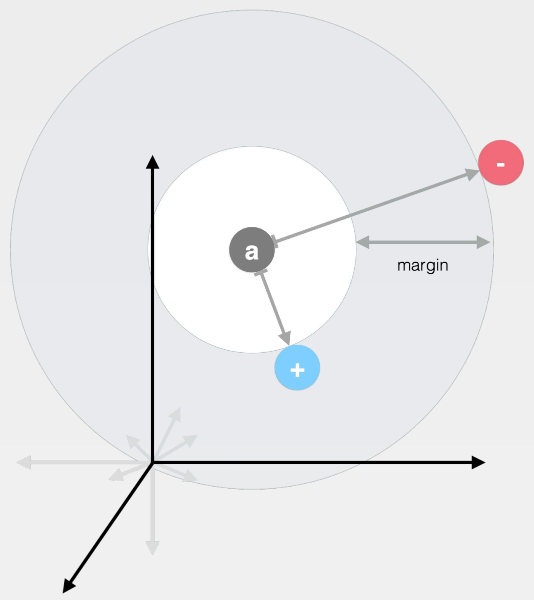
Classic Pipeline



CNNs

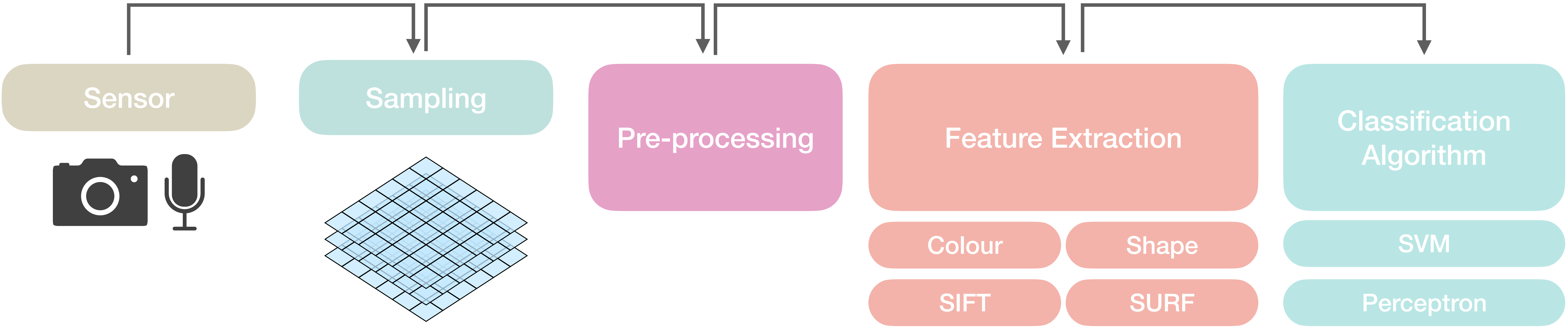


Transformers

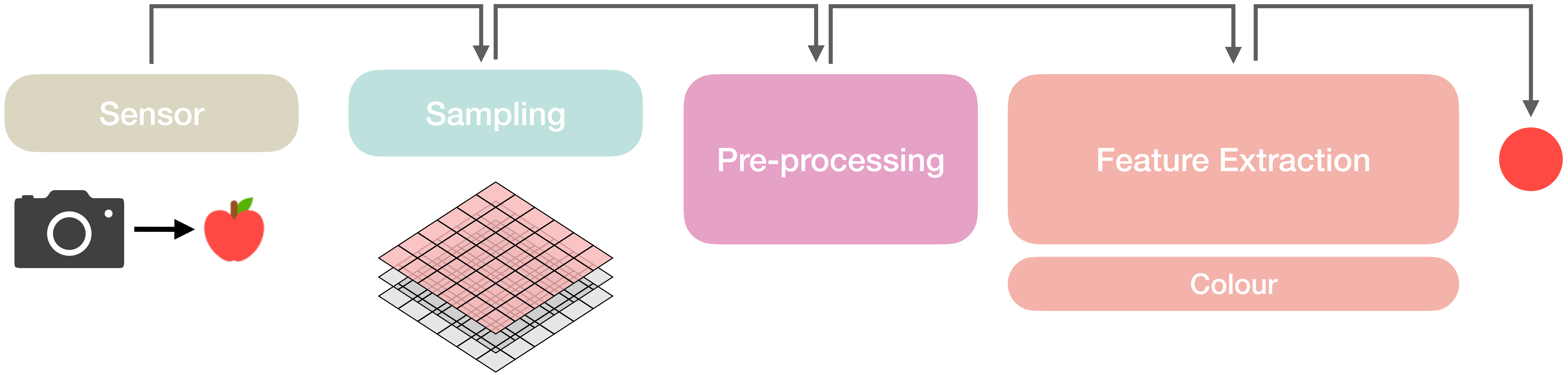


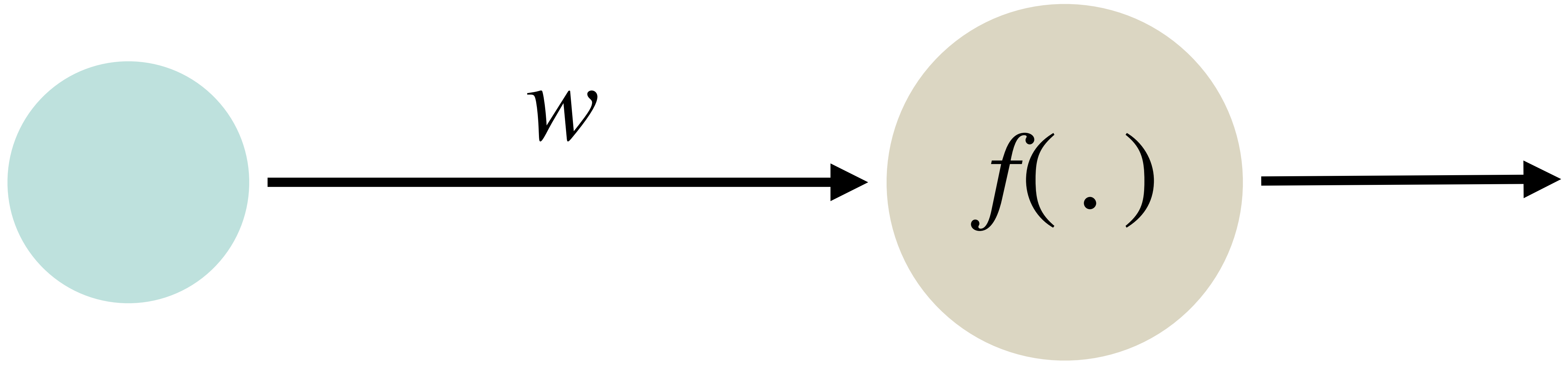
Contrastive Learning

The Common Pipeline



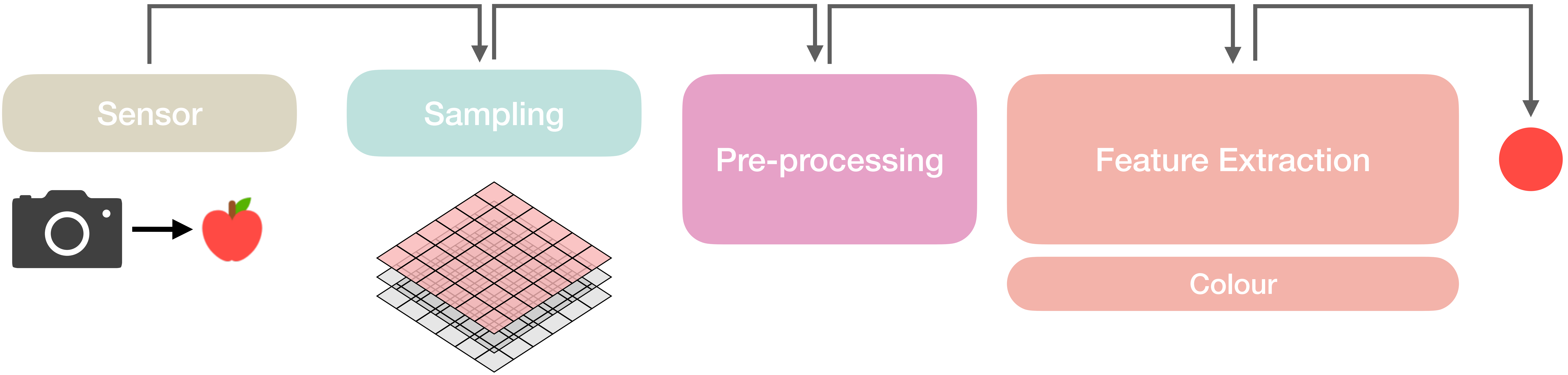
The Common Pipeline

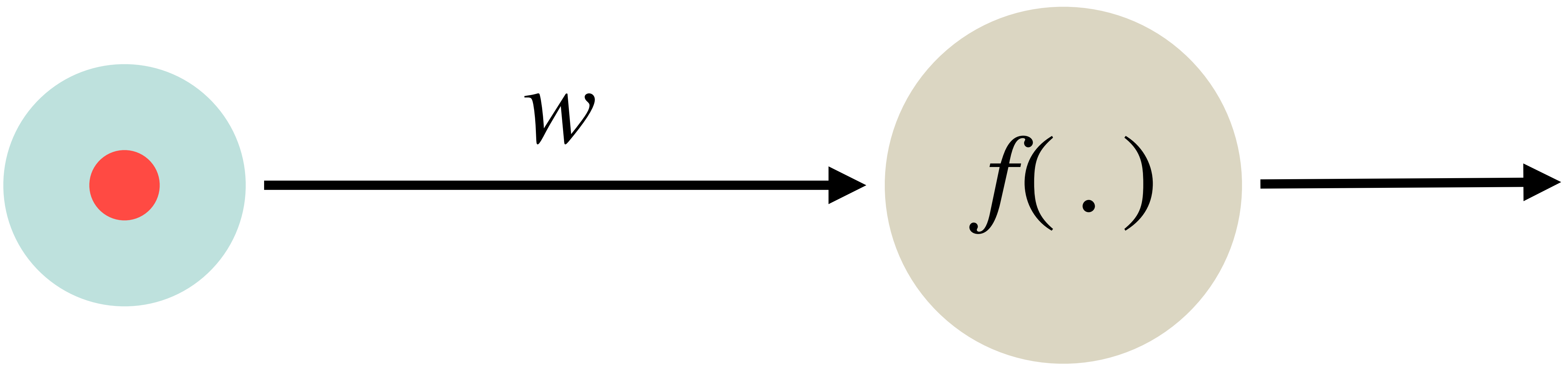




$$\hat{y} = f(wx + b)$$

The Common Pipeline



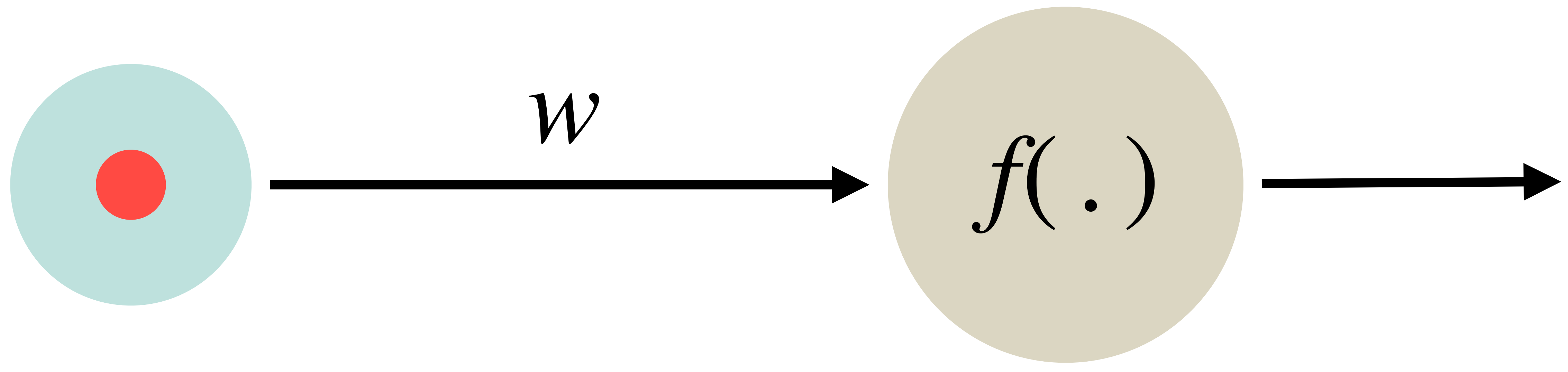


$$\hat{y} = f(wx + b)$$

The Common Pipeline

Training

1. Initialise w and b
2. Find optimal w and b as defined by loss function $J(w, b, x)$
3. Use $\hat{y} = f(wx + b)$ to make predictions



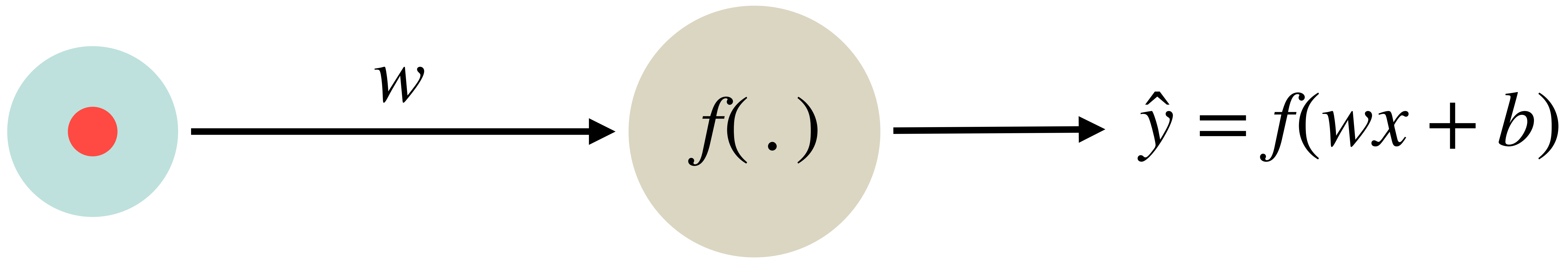
$$\hat{y} = f(wx + b)$$

The Common Pipeline

Loss Function

2 | Find optimal w and b as defined by loss function $J(w, b, x)$

$$
\begin{aligned}
J(w, b, x) &= y - \hat{y} \\
&= y - f(wx + b)
\end{aligned}
$$



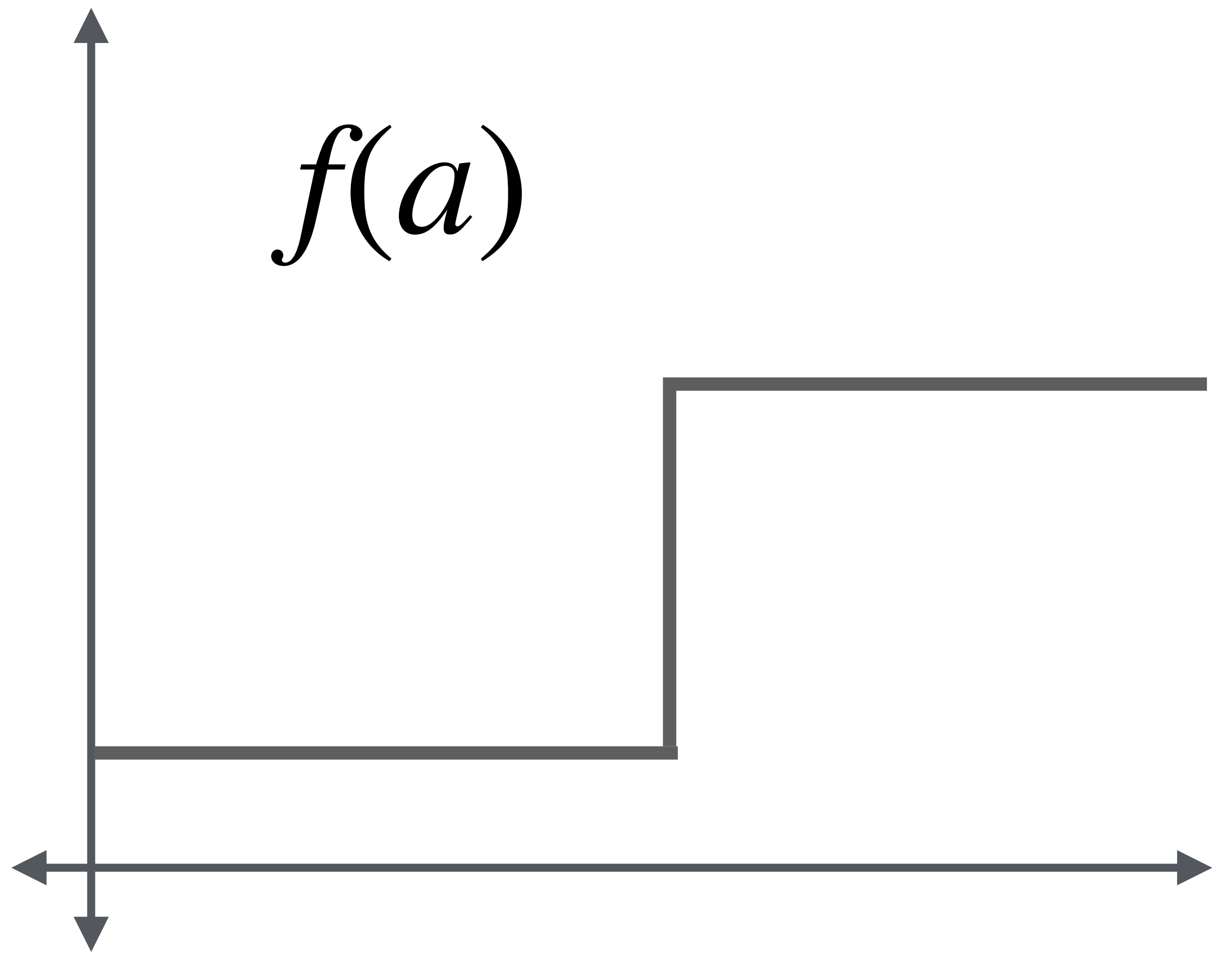
The Common Pipeline

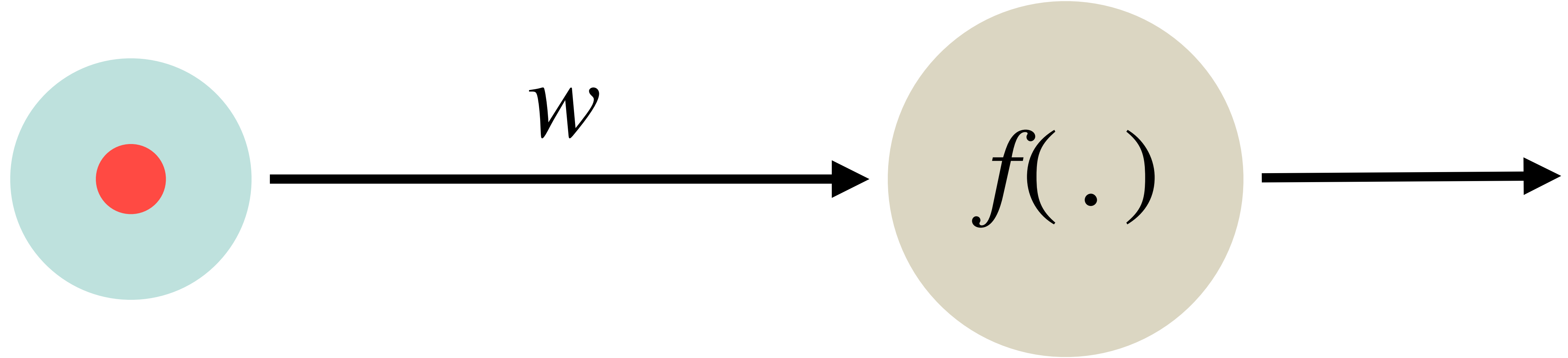
Loss Function

2 | Find optimal w and b as defined by **loss function** $J(w, b, x)$

$$J(w, b, x) = y - \hat{y}$$
$$= y - f(wx + b)$$

$$f(a) = \begin{cases} 1, & \text{se } a > 0 \\ -1, & \text{se } a \leq 0 \end{cases}$$





$$\hat{y} = f(wx + b)$$

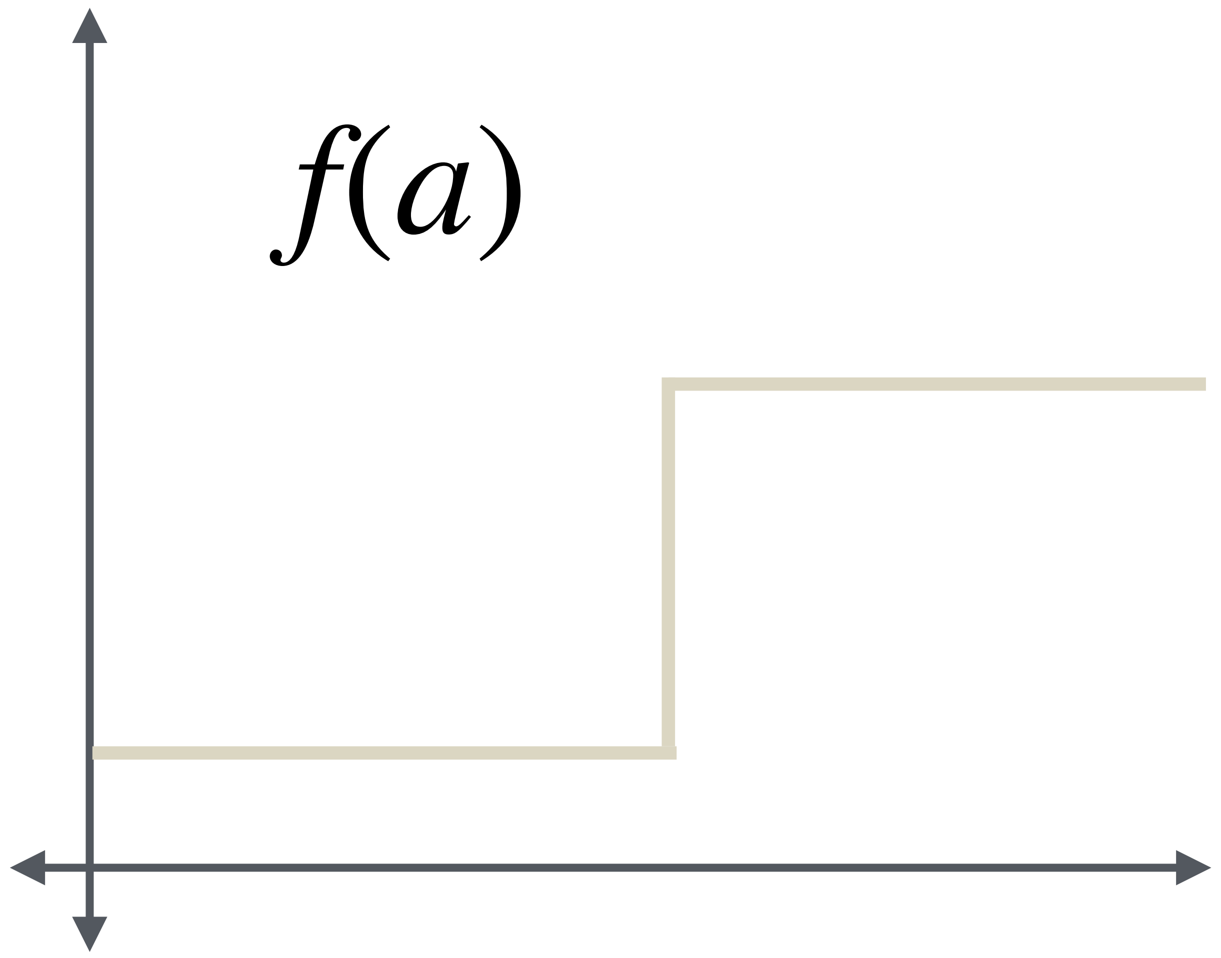
The Common Pipeline

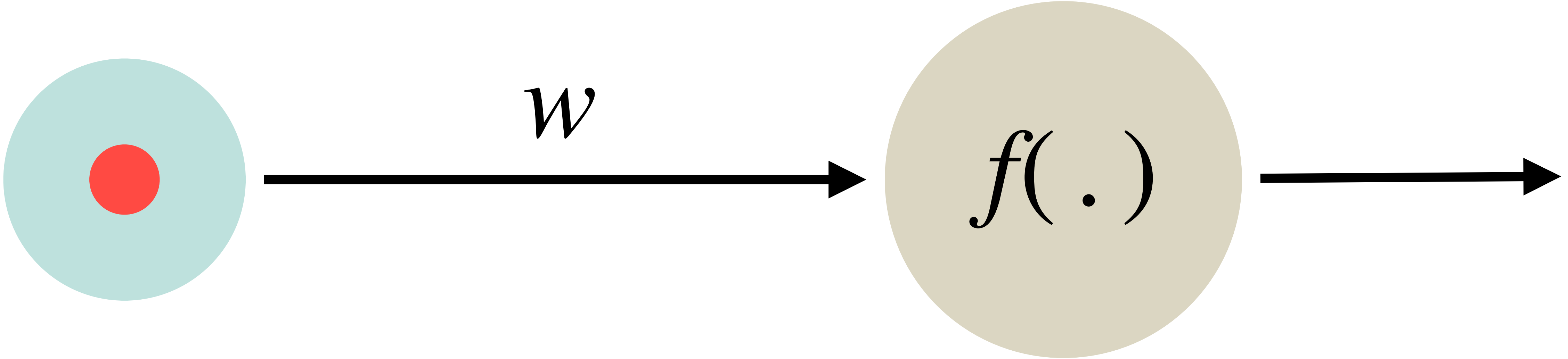
Loss Function

2 Find optimal w and b as defined by loss function $J(w, b, x)$

$$\begin{aligned} J(w, b, x) &= y - \hat{y} \\ &= y - f(wx + b) \end{aligned}$$

$$f(a) = \begin{cases} 1, & \text{se } a > 0 \\ -1, & \text{se } a \leq 0 \end{cases}$$





$\hat{y} = f(wx + b)$

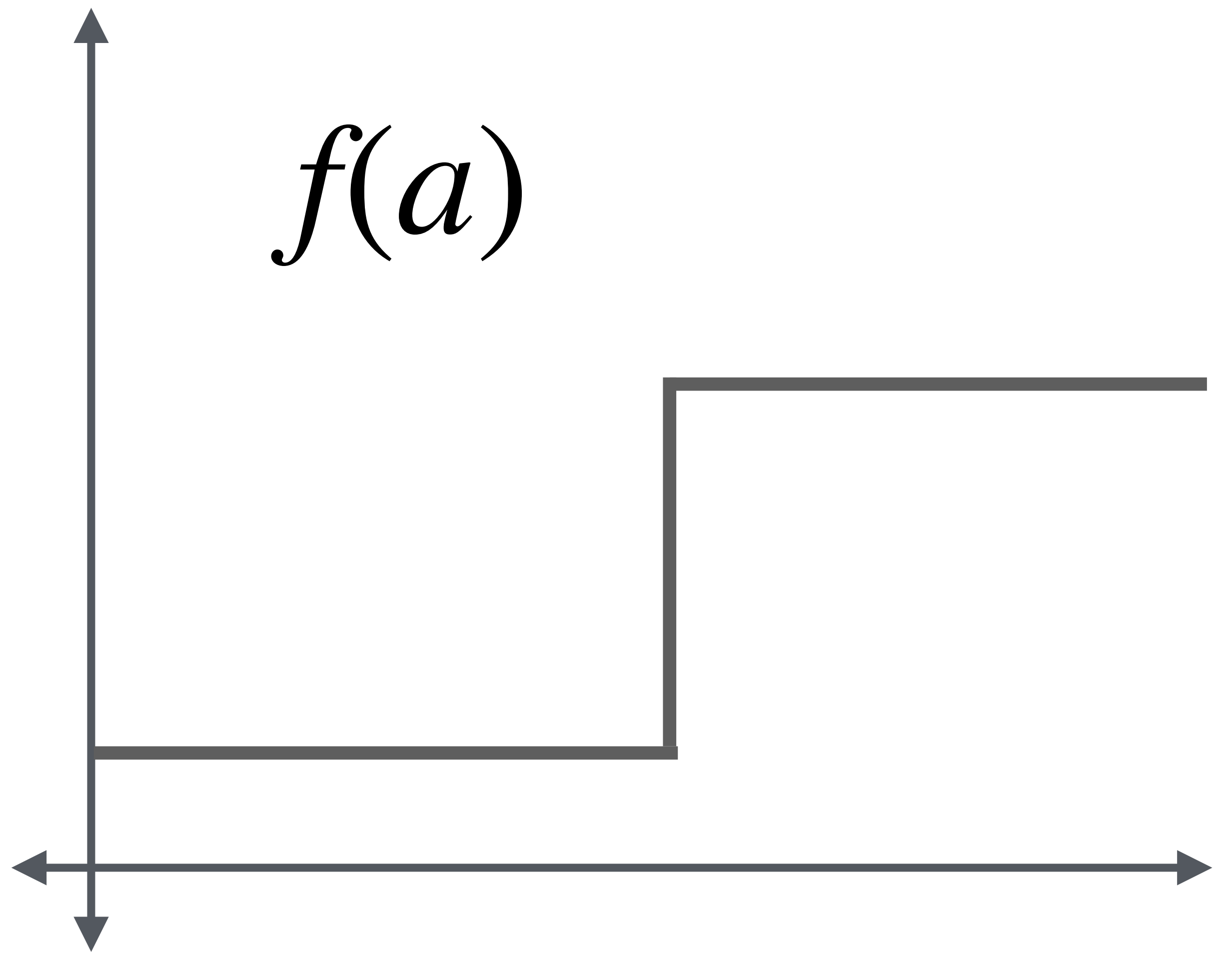
The Common Pipeline

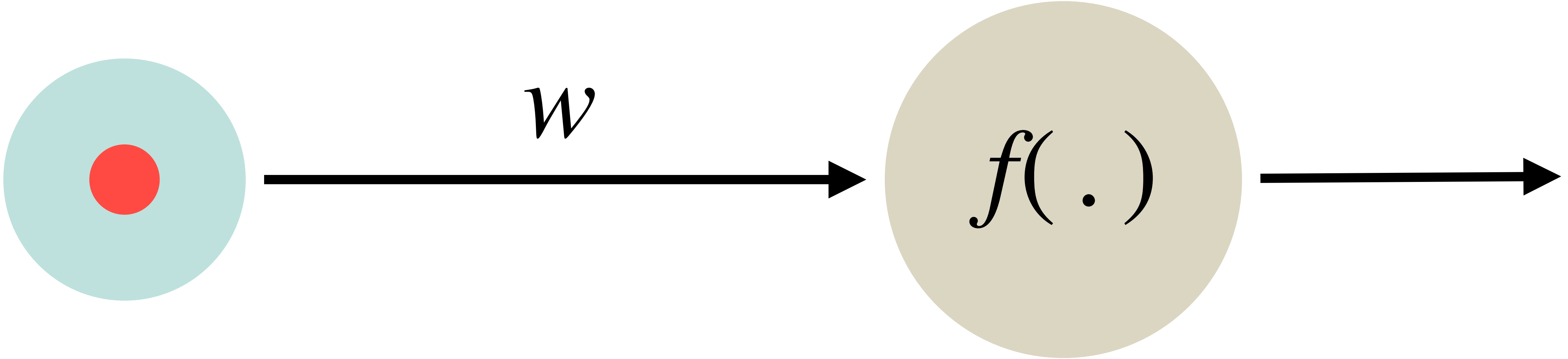
Loss Function

2 **Find optimal** w and b as defined by loss function $J(w, b, x)$

$$J(w, b, x) = y - \hat{y}$$
$$= y - f(wx + b)$$

$$f(a) = \begin{cases} 1, & \text{se } a > 0 \\ -1, & \text{se } a \leq 0 \end{cases}$$

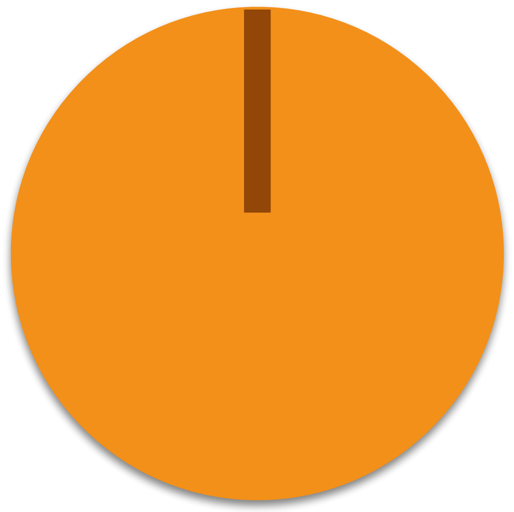


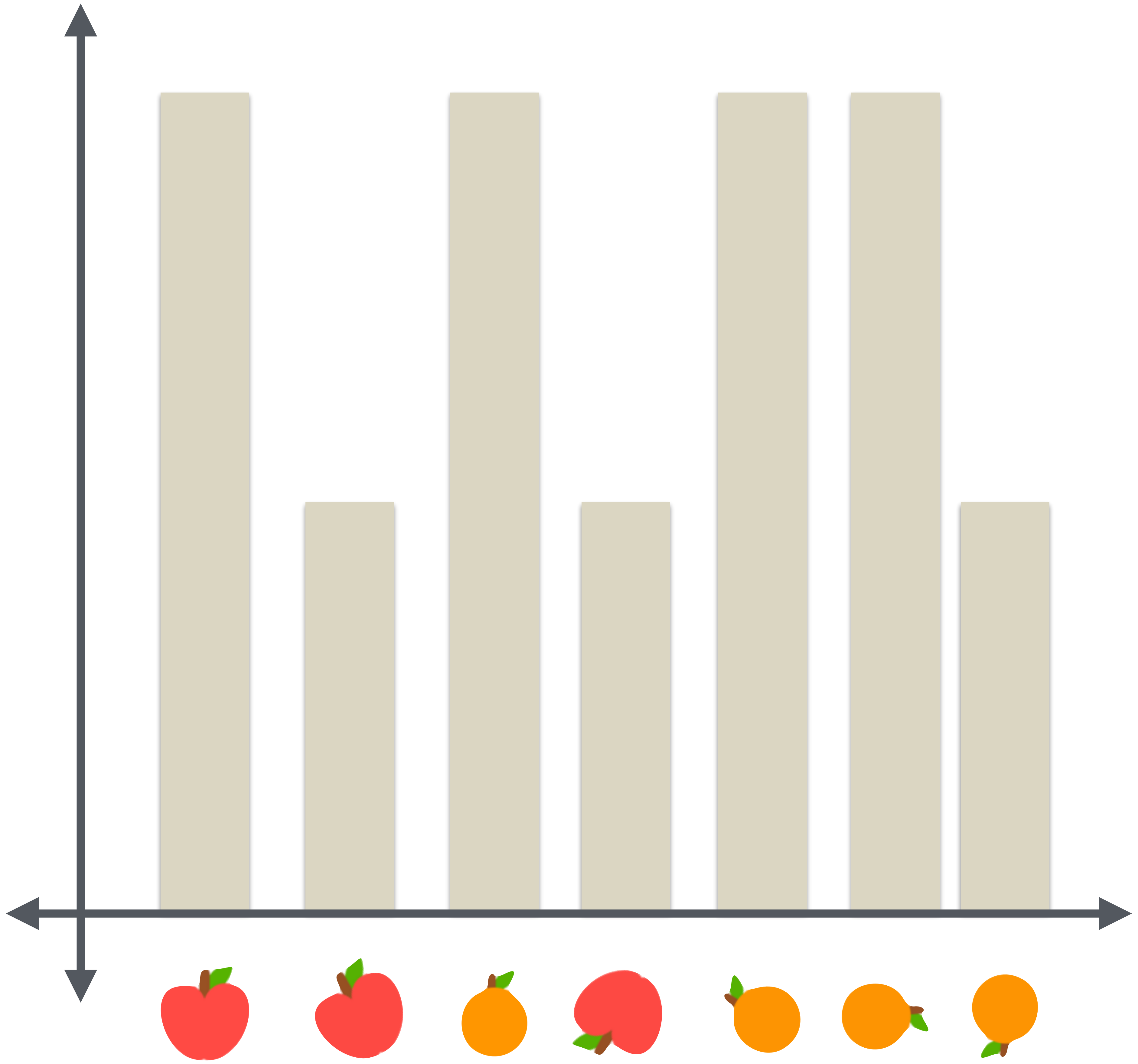


$$\hat{y} = f(wx + b)$$

The Common Pipeline

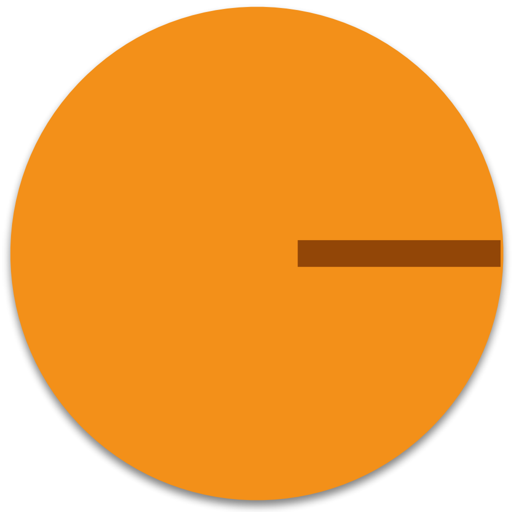
w

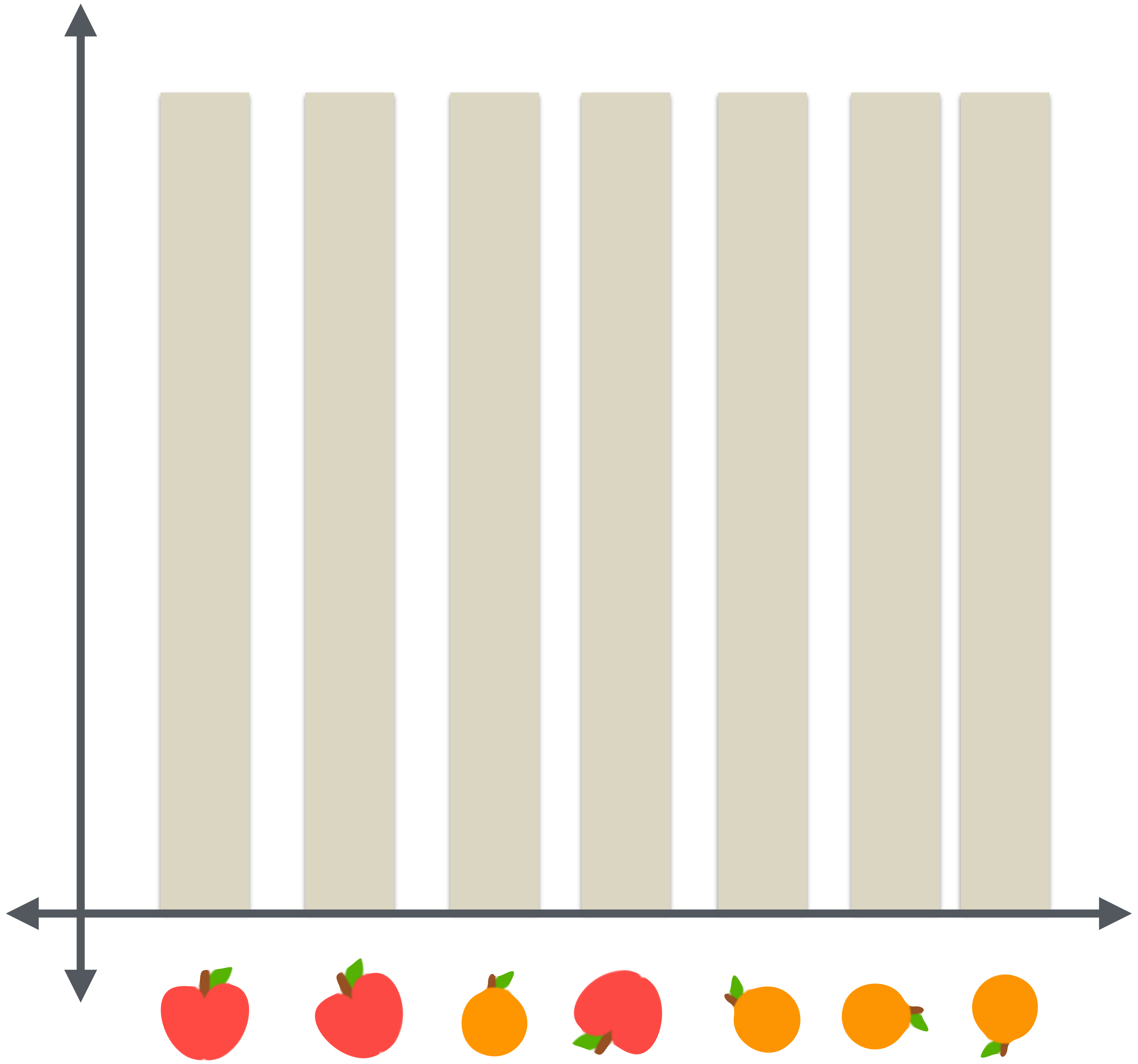




The Common Pipeline

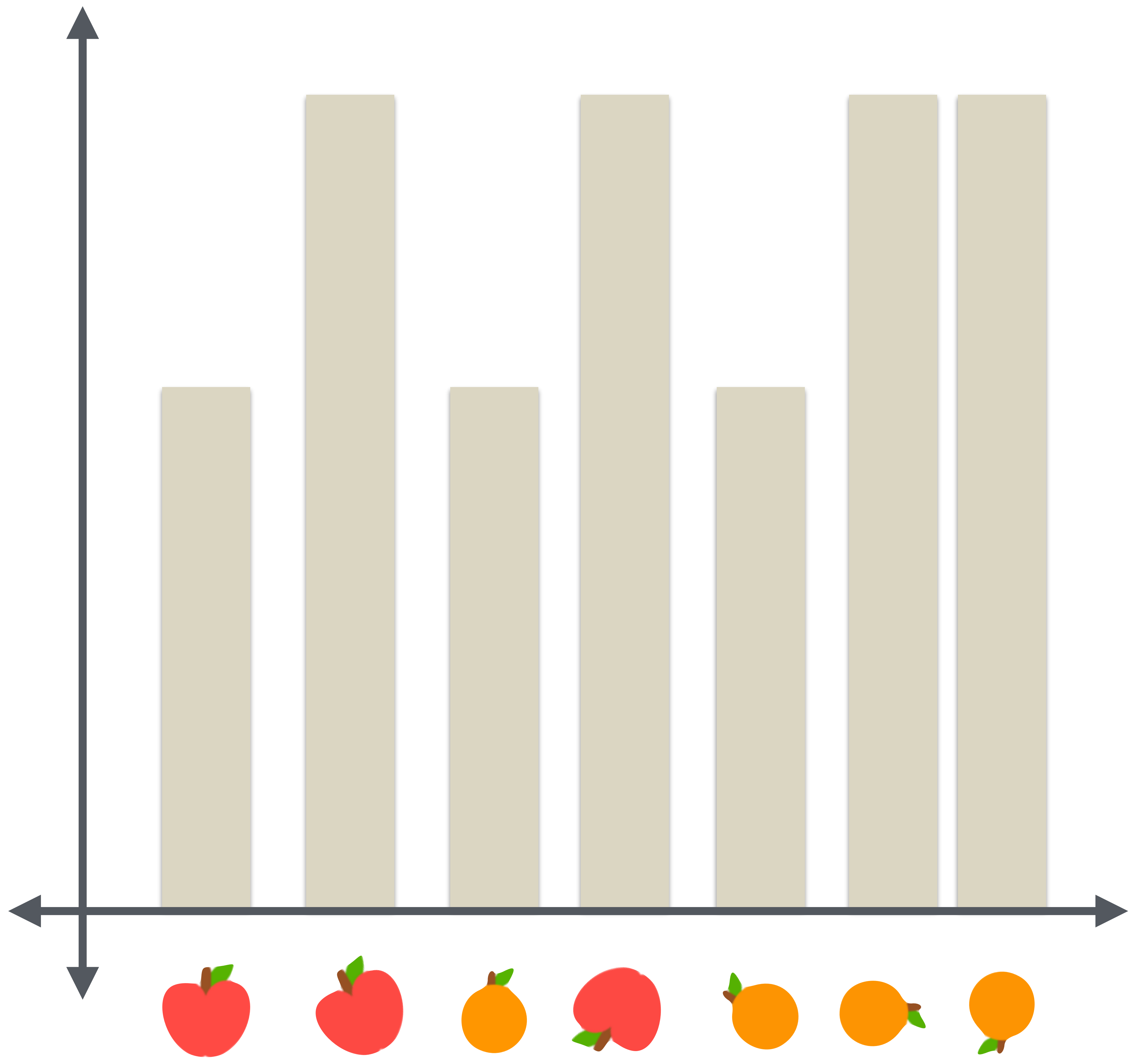
w



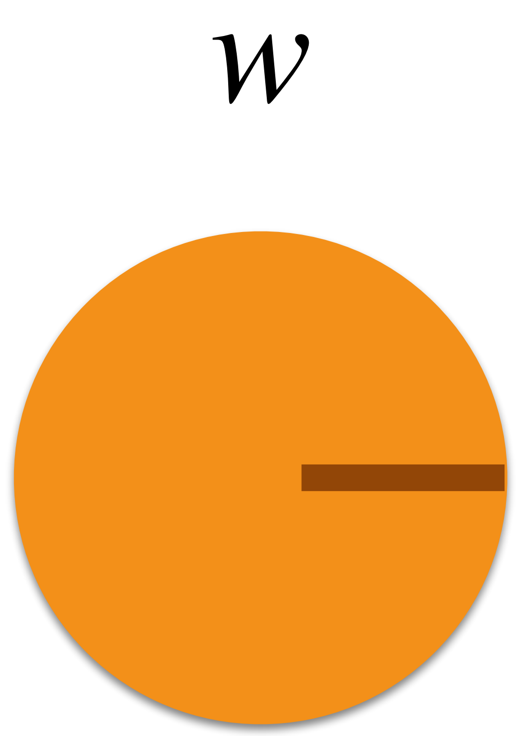


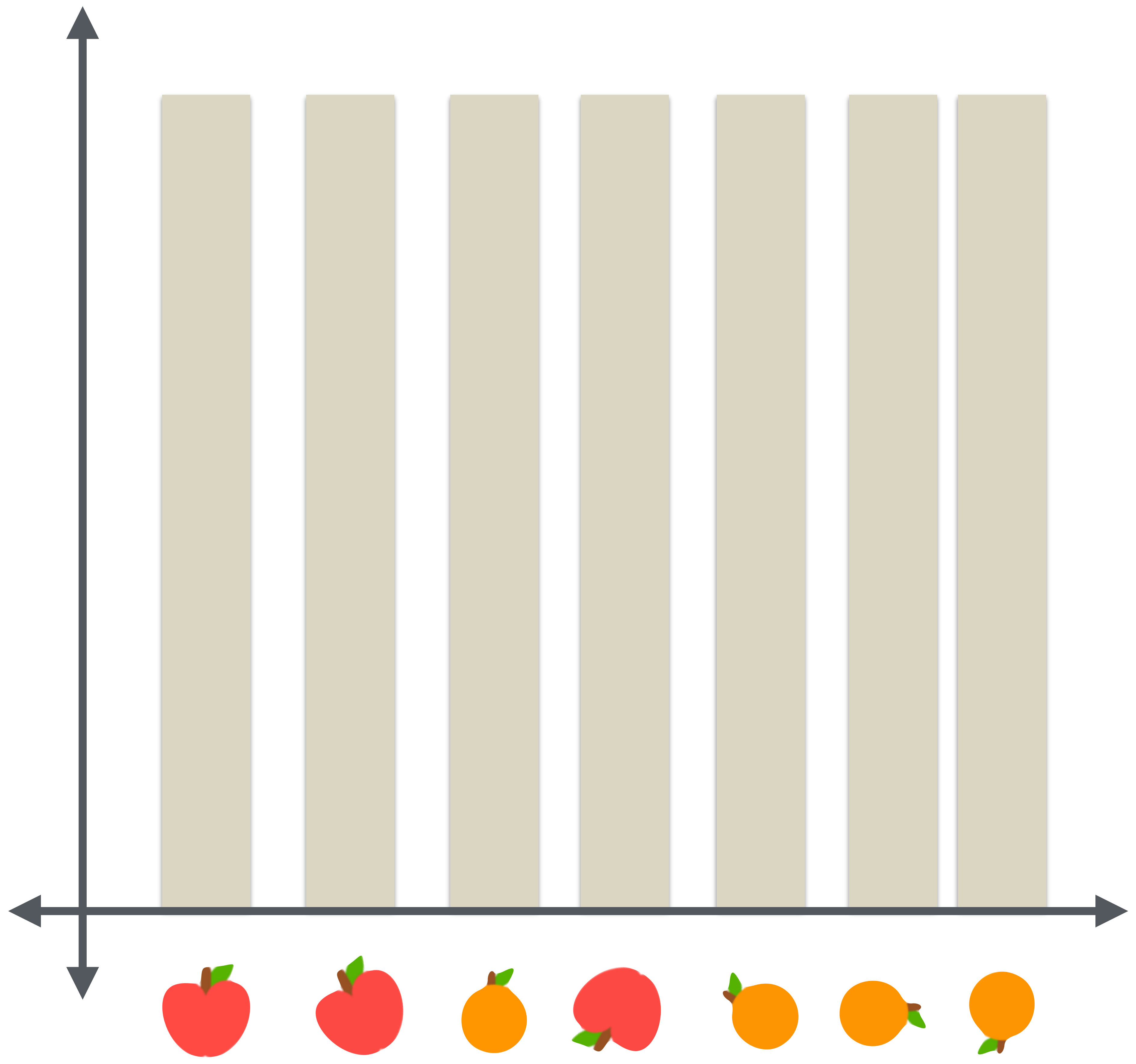
The Common Pipeline

w

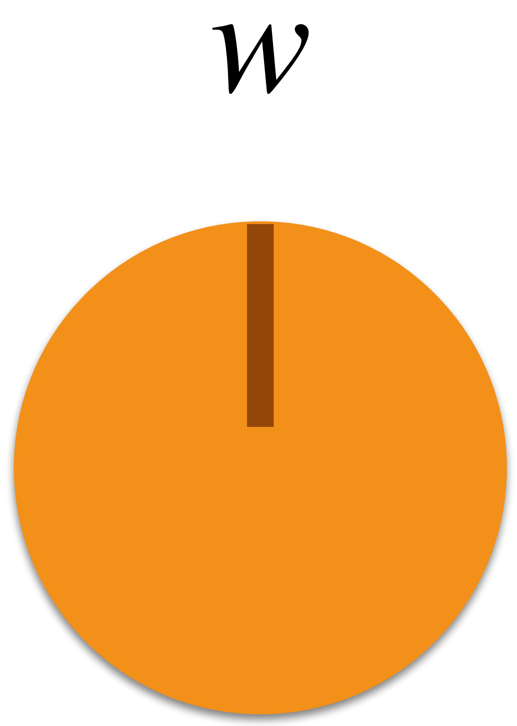


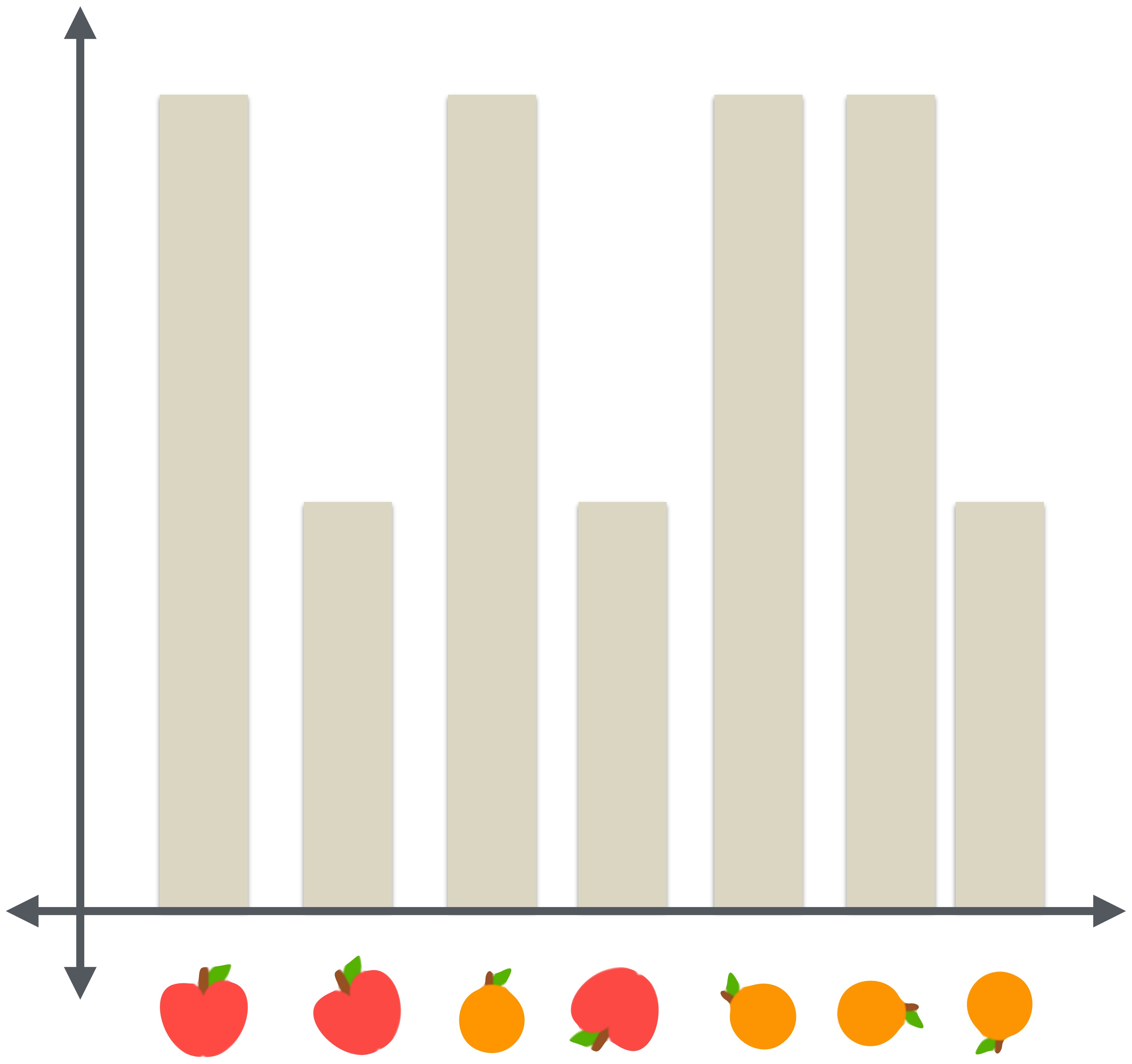
The Common Pipeline



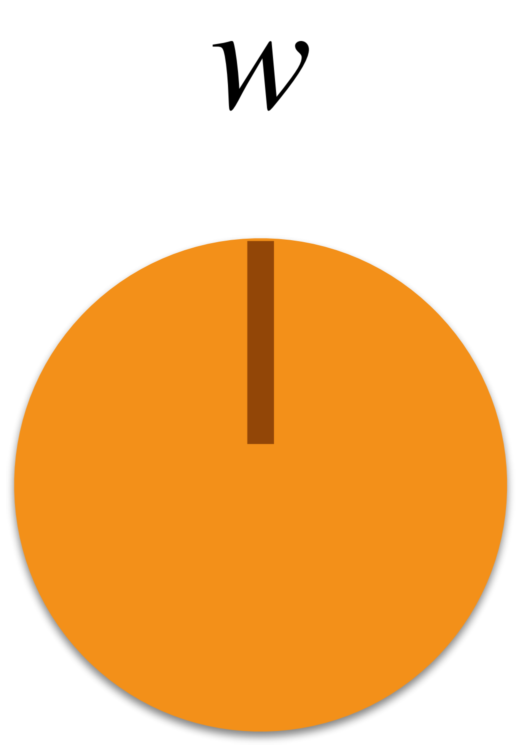


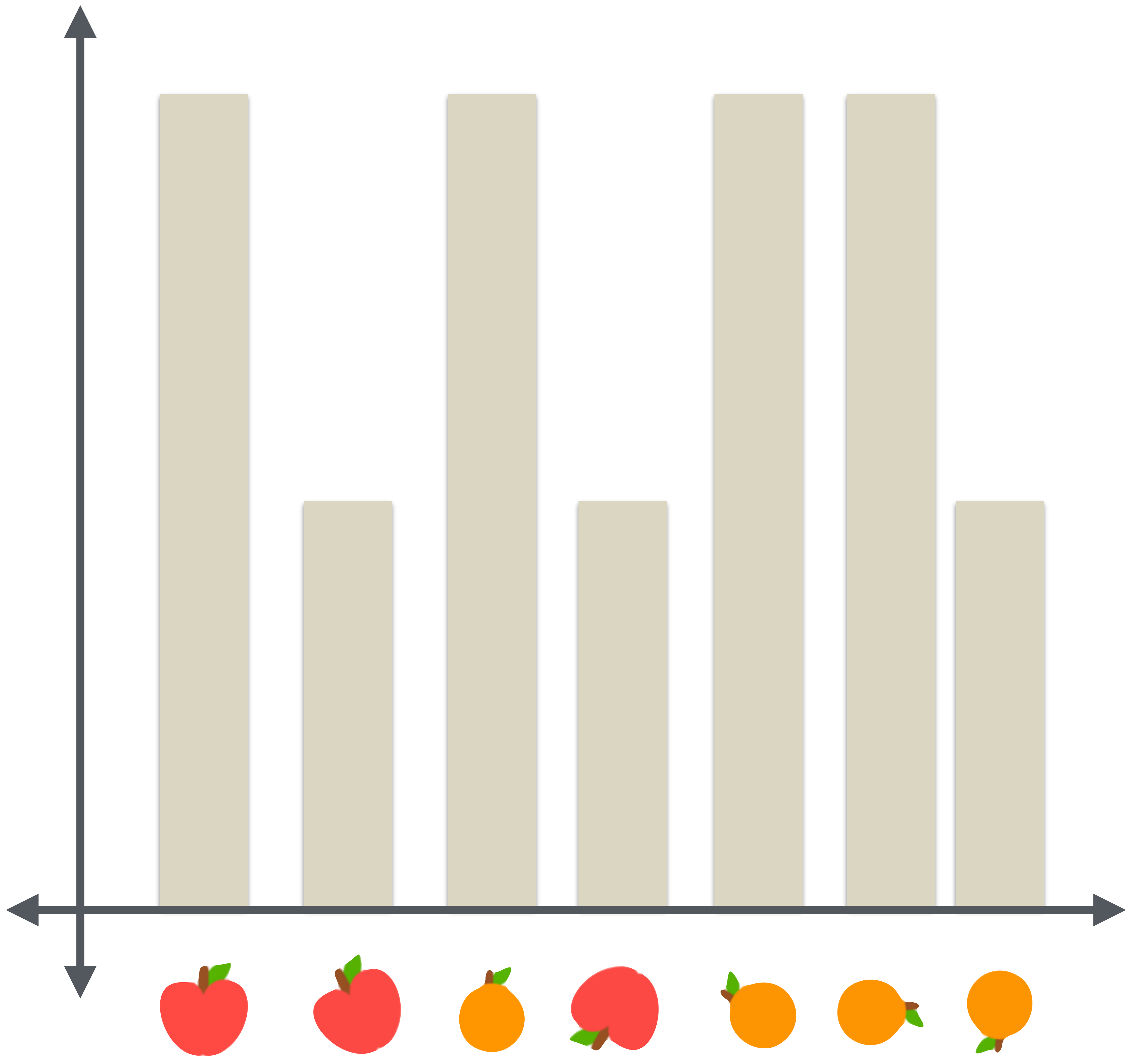
The Common Pipeline

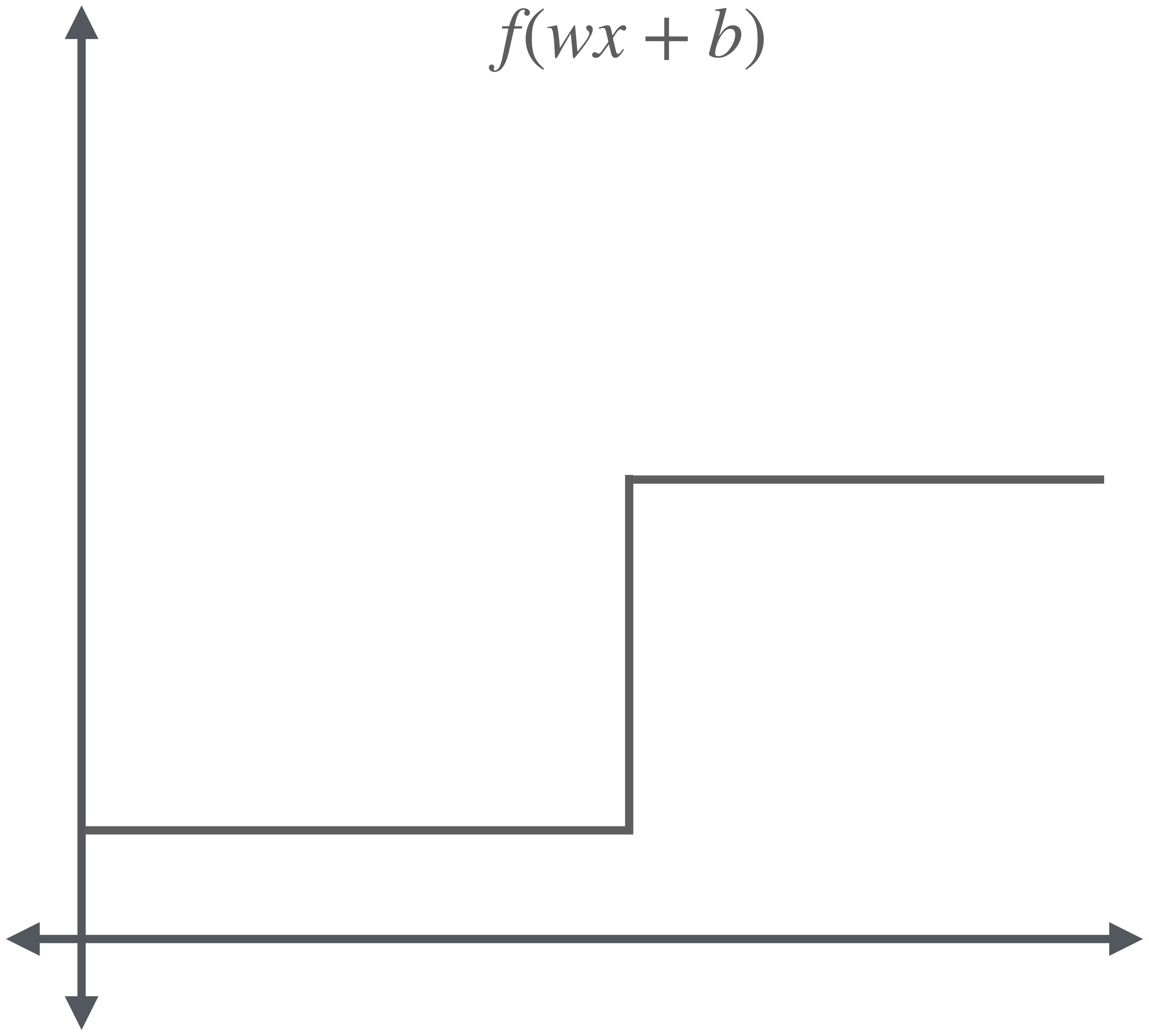




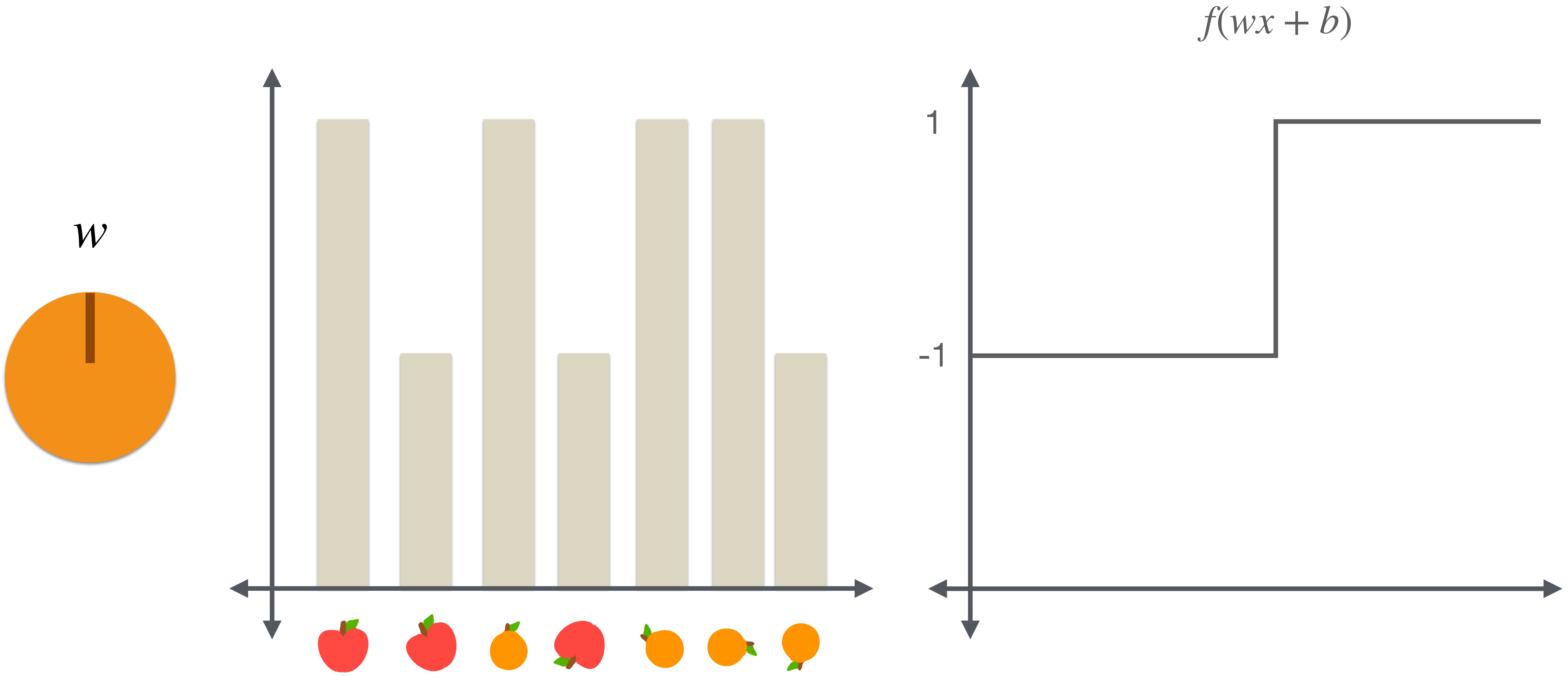
The Common Pipeline



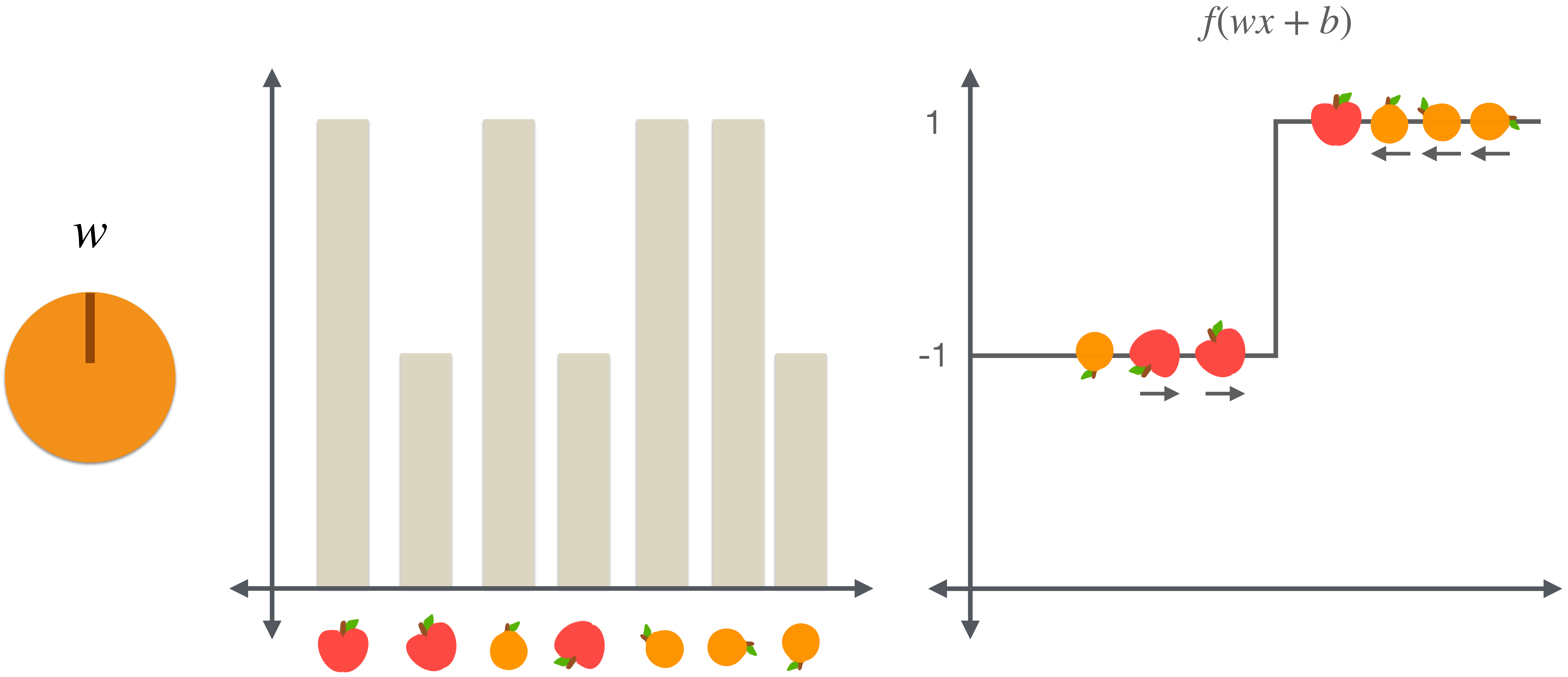




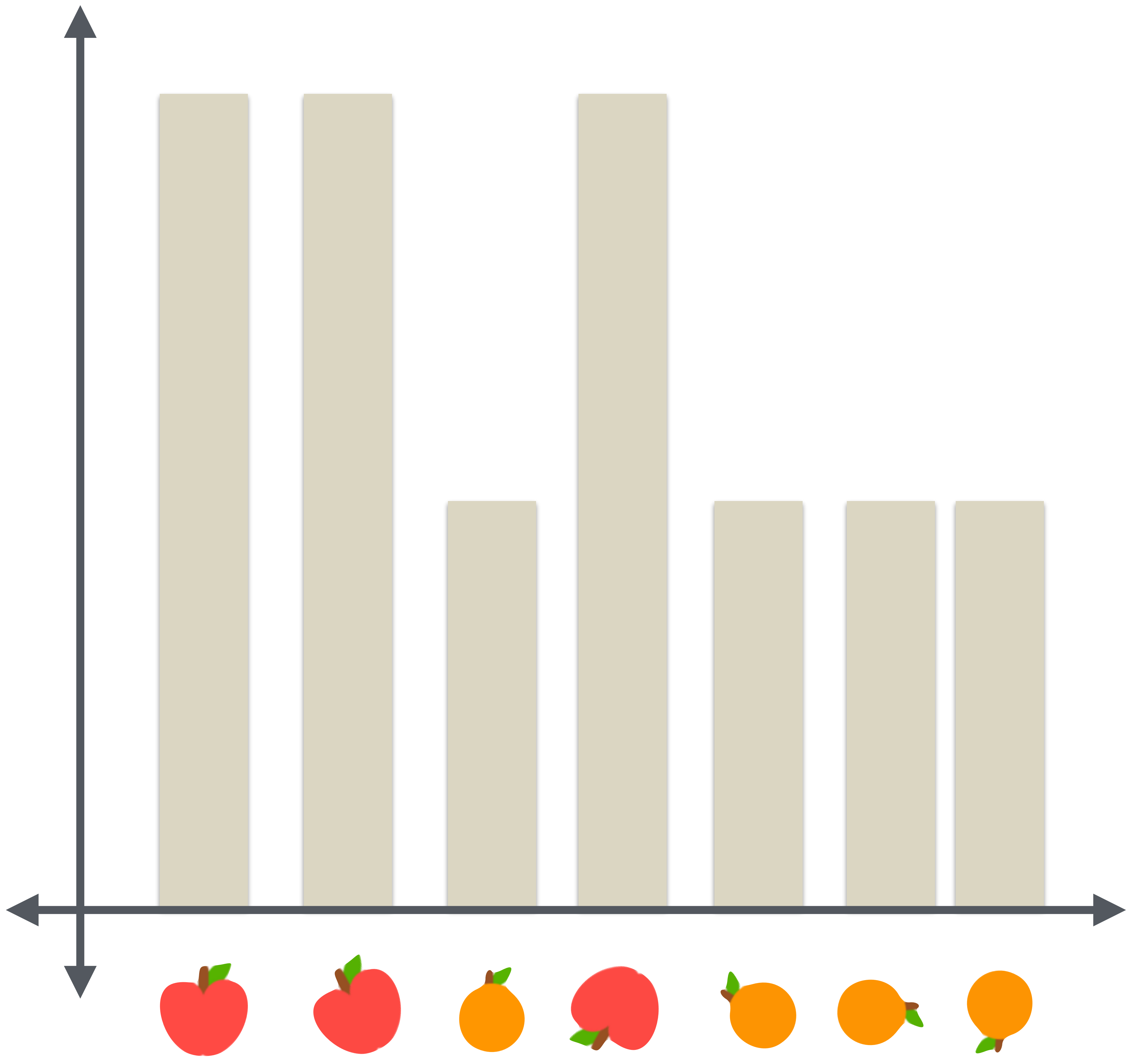
The Common Pipeline

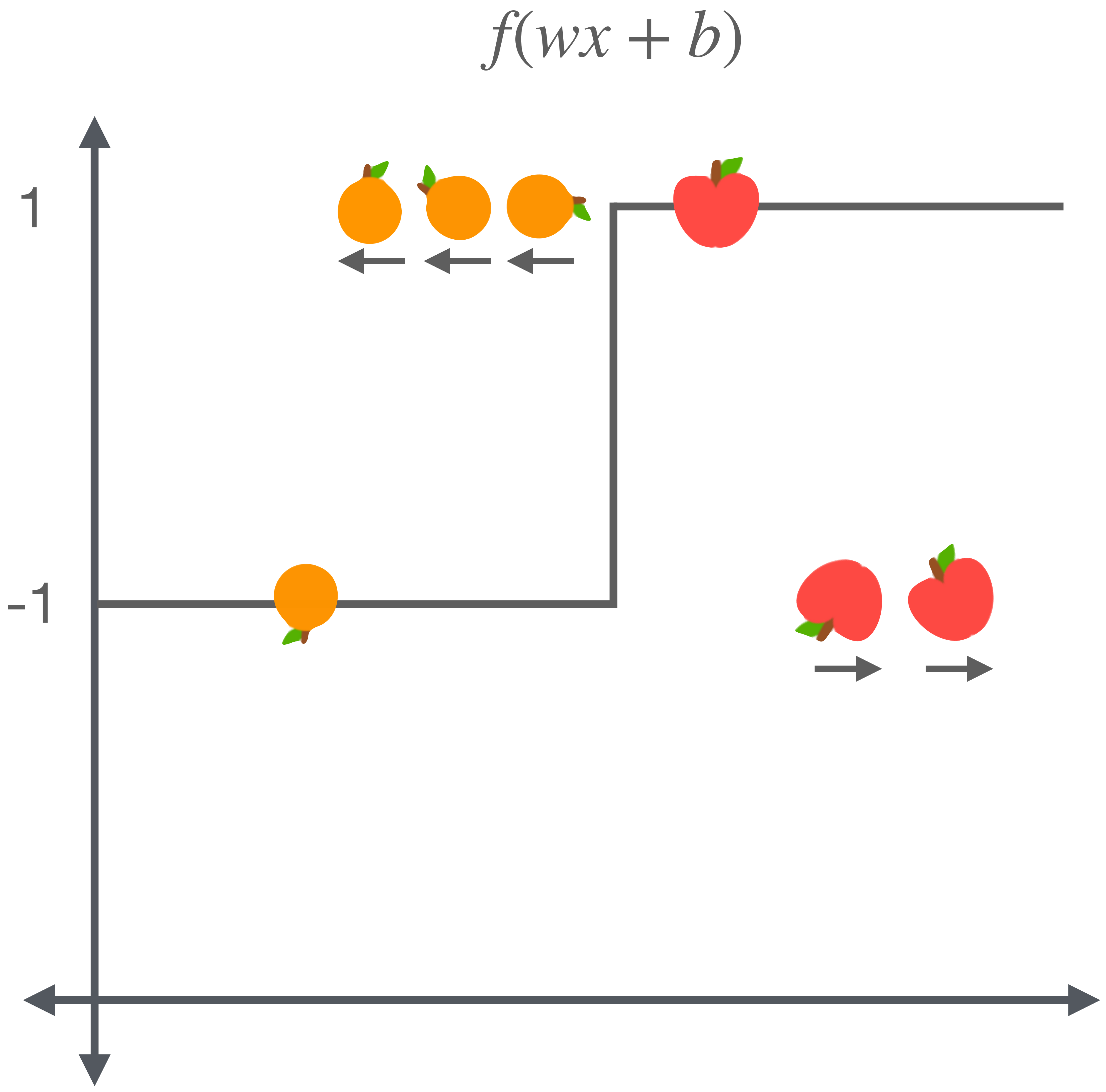


The Common Pipeline

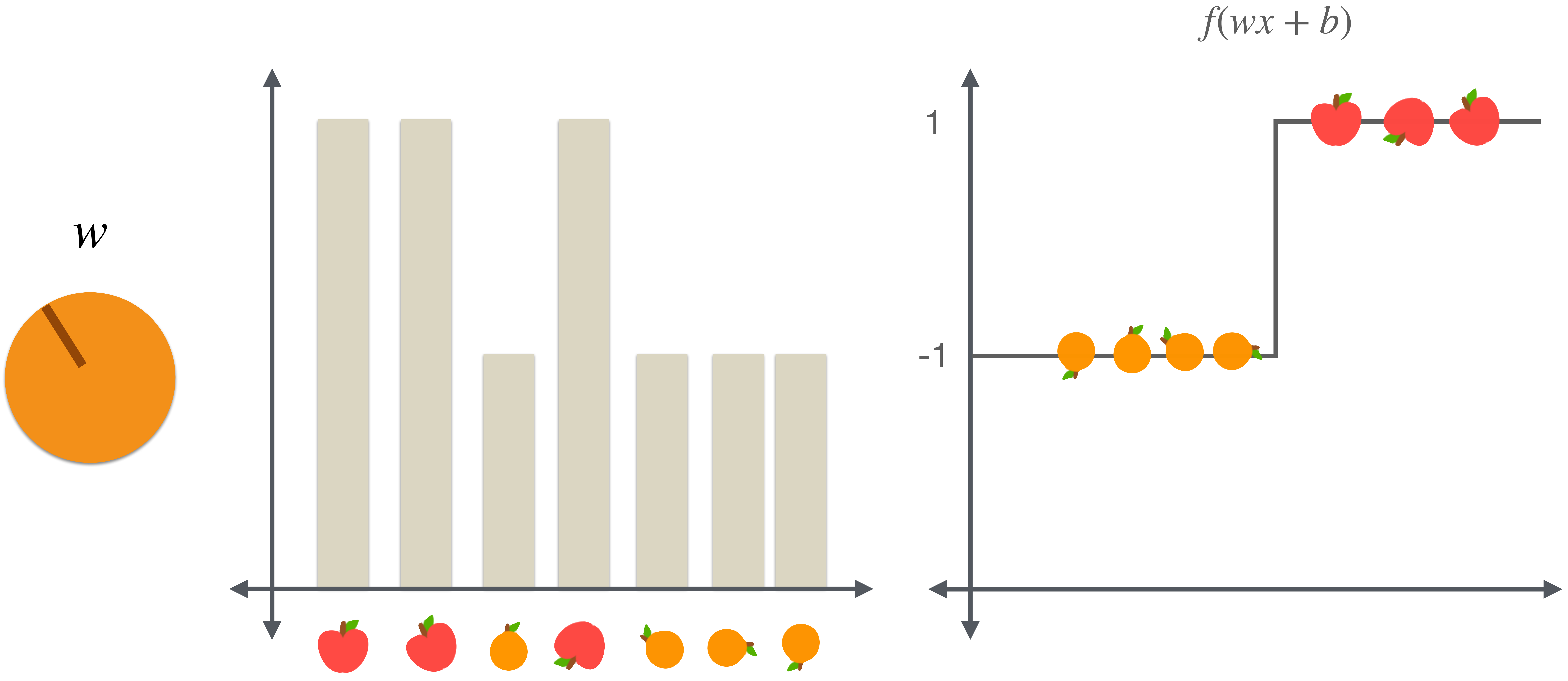


The Common Pipeline



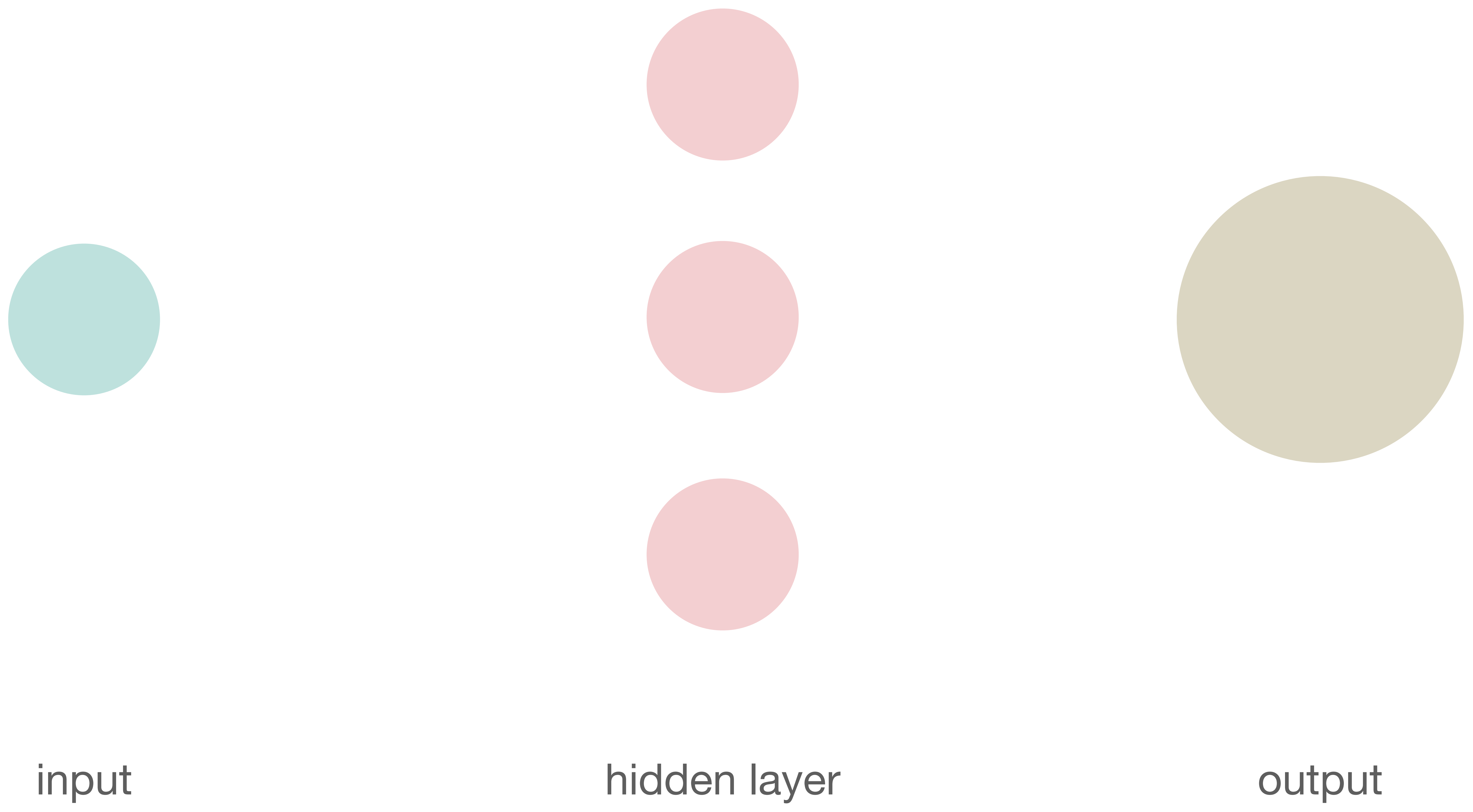


The Common Pipeline

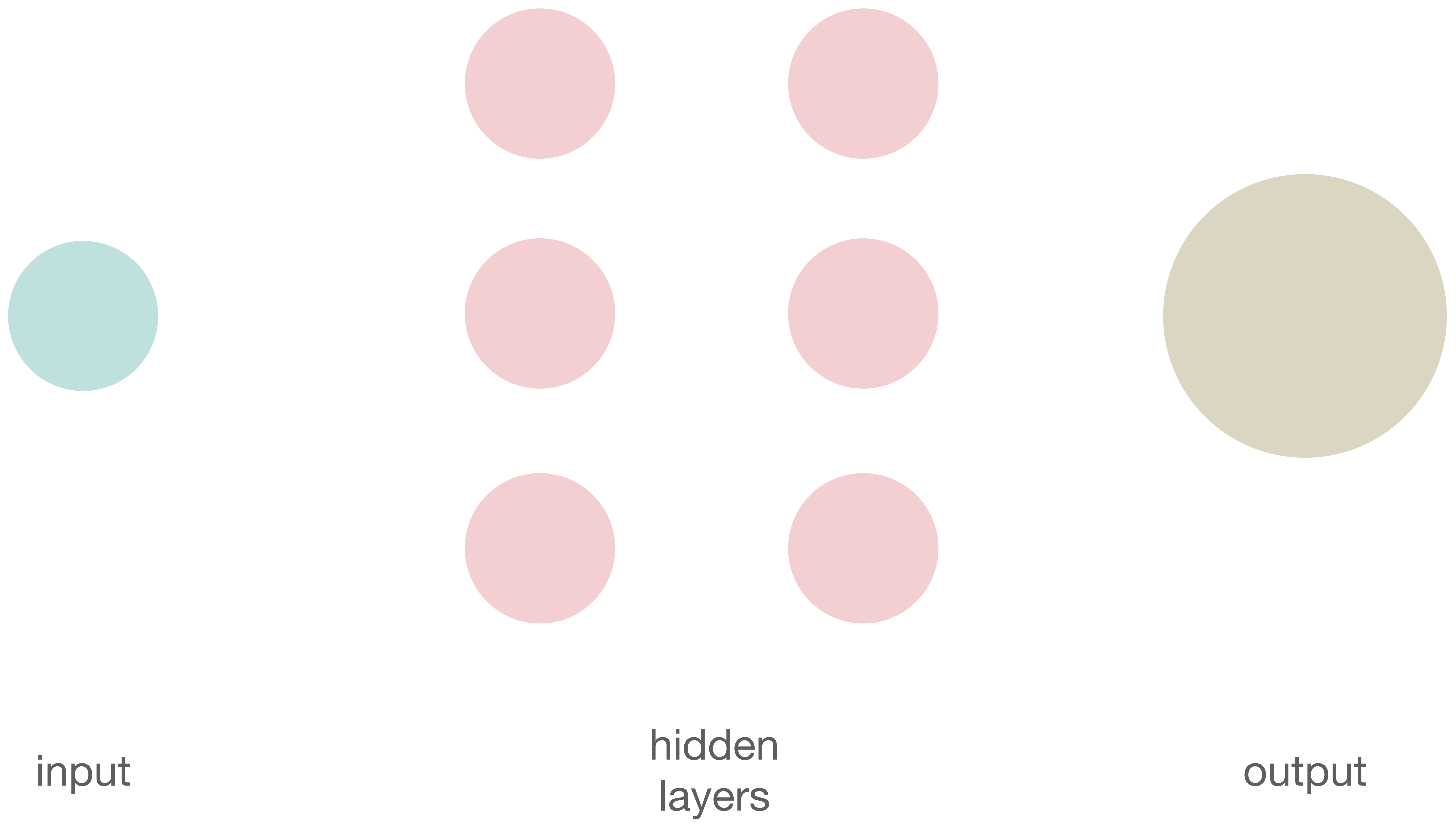


The Common Pipeline

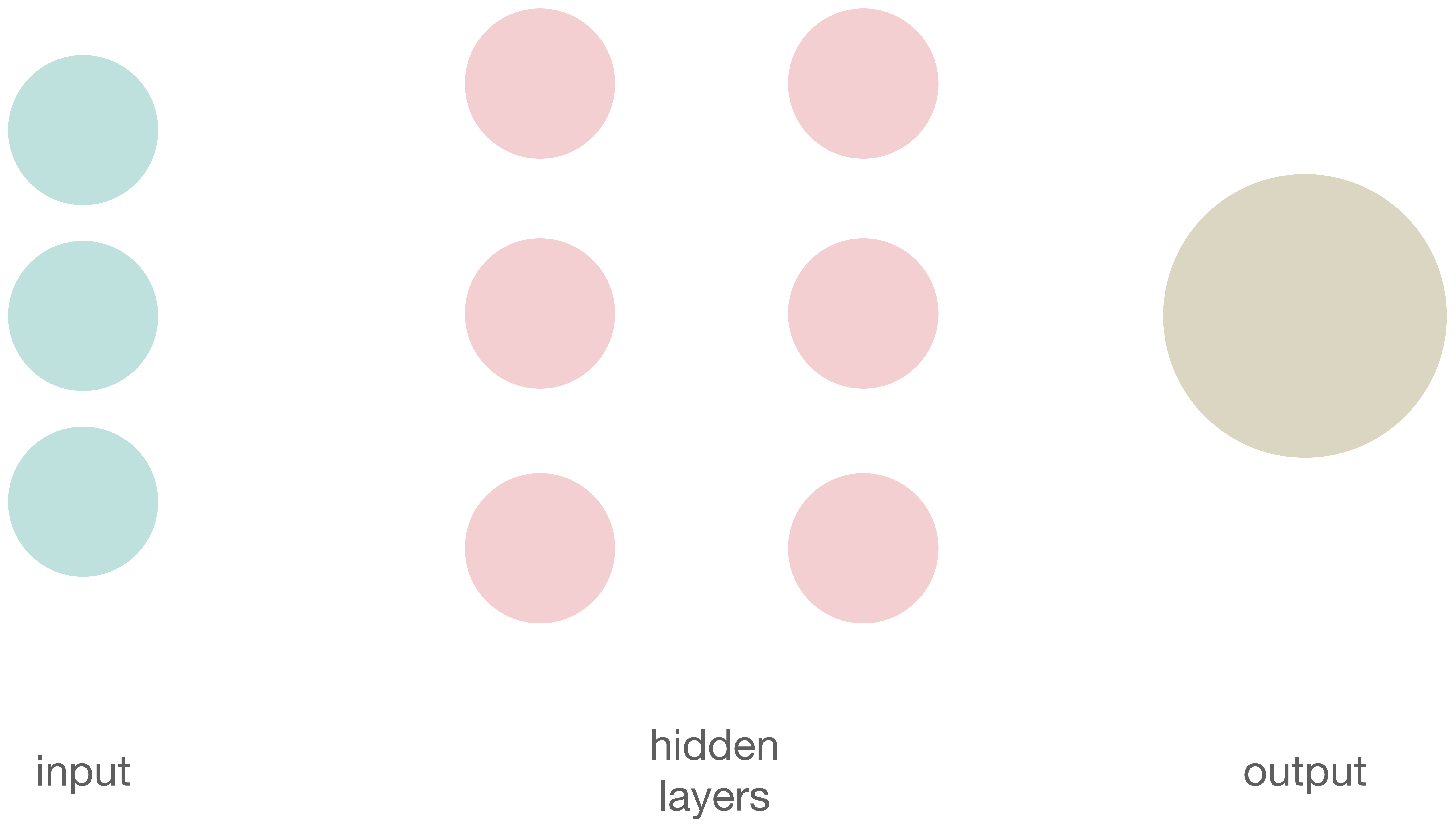
The Common Pipeline



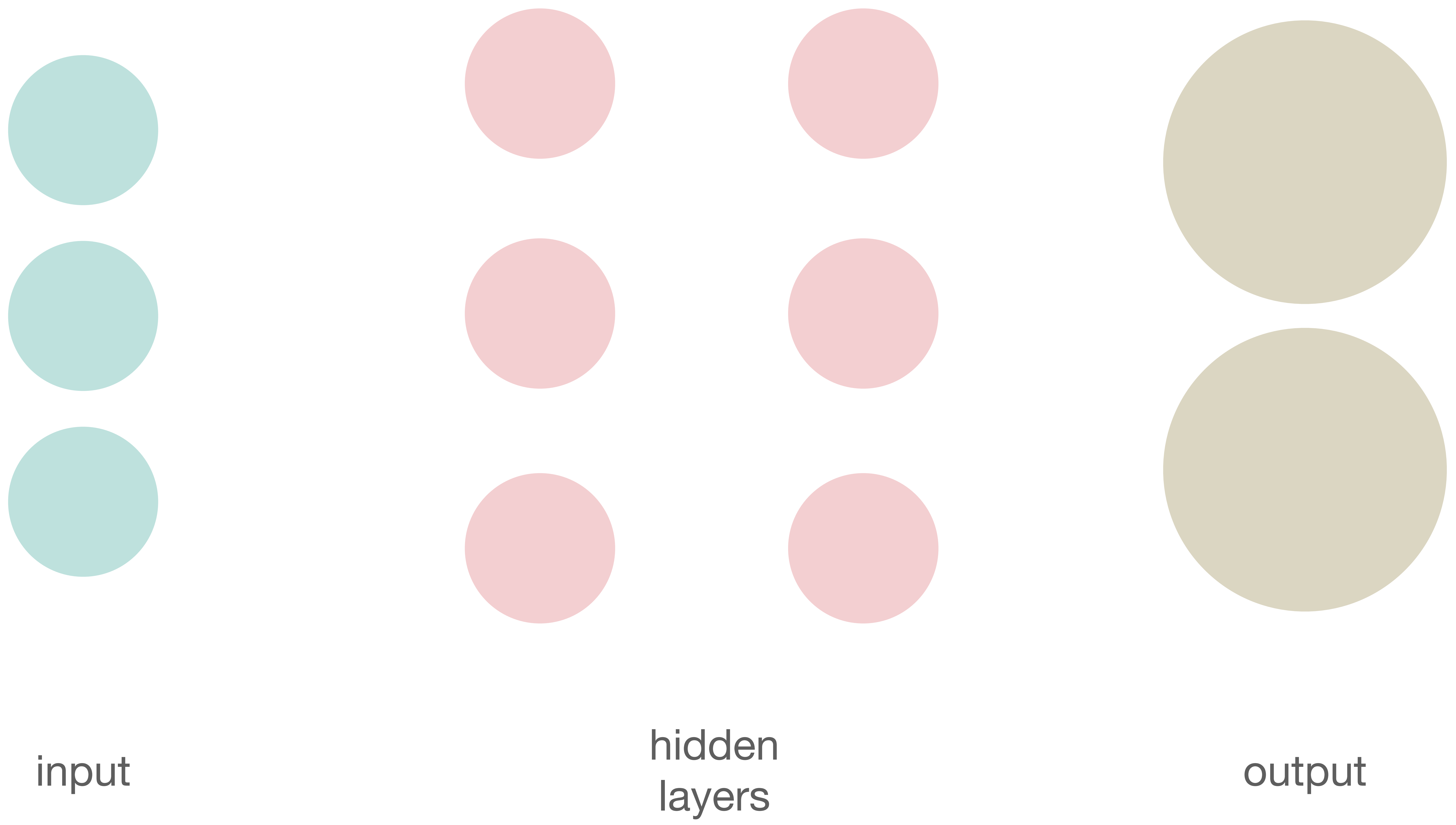
The Common Pipeline



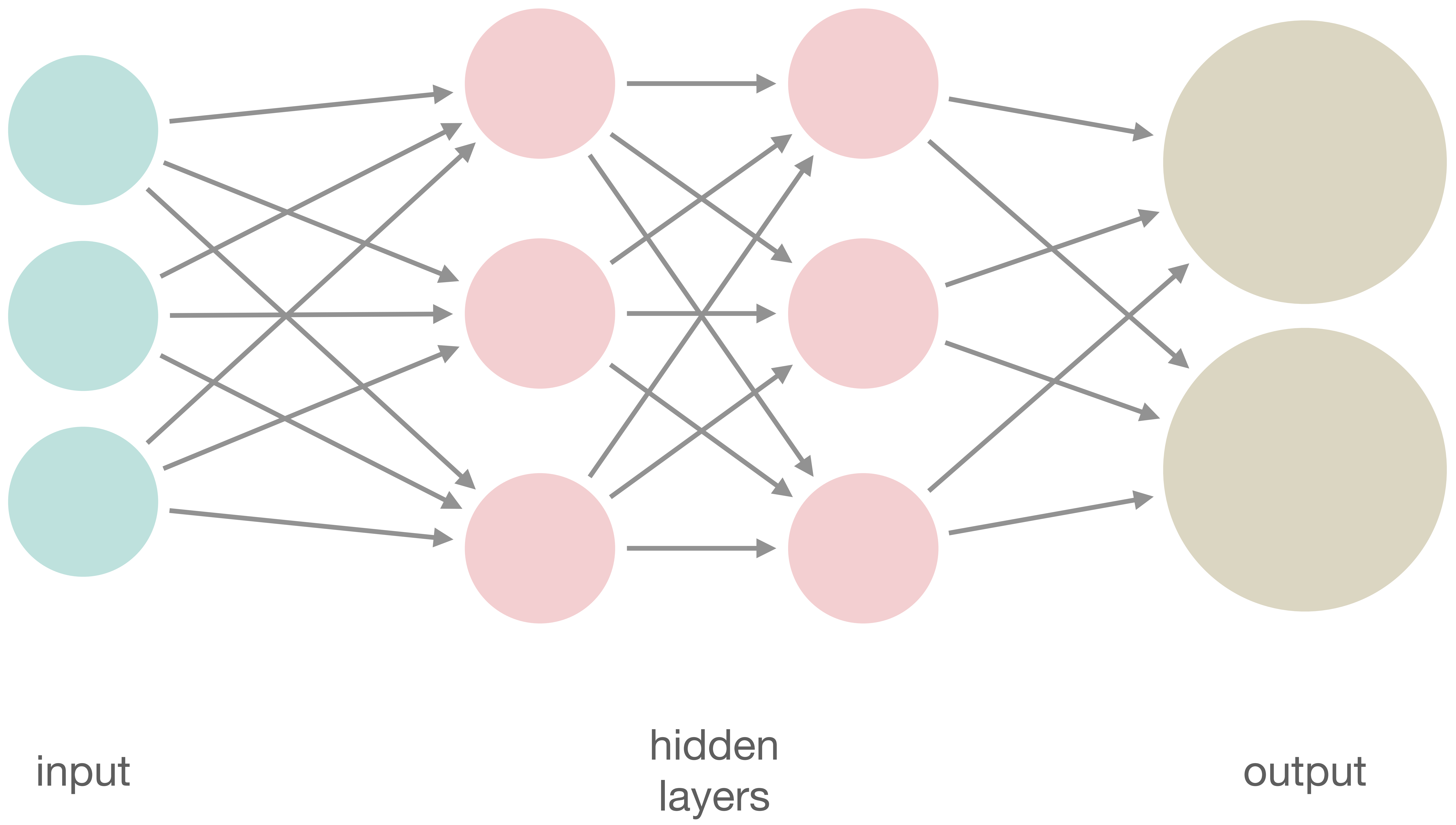
The Common Pipeline



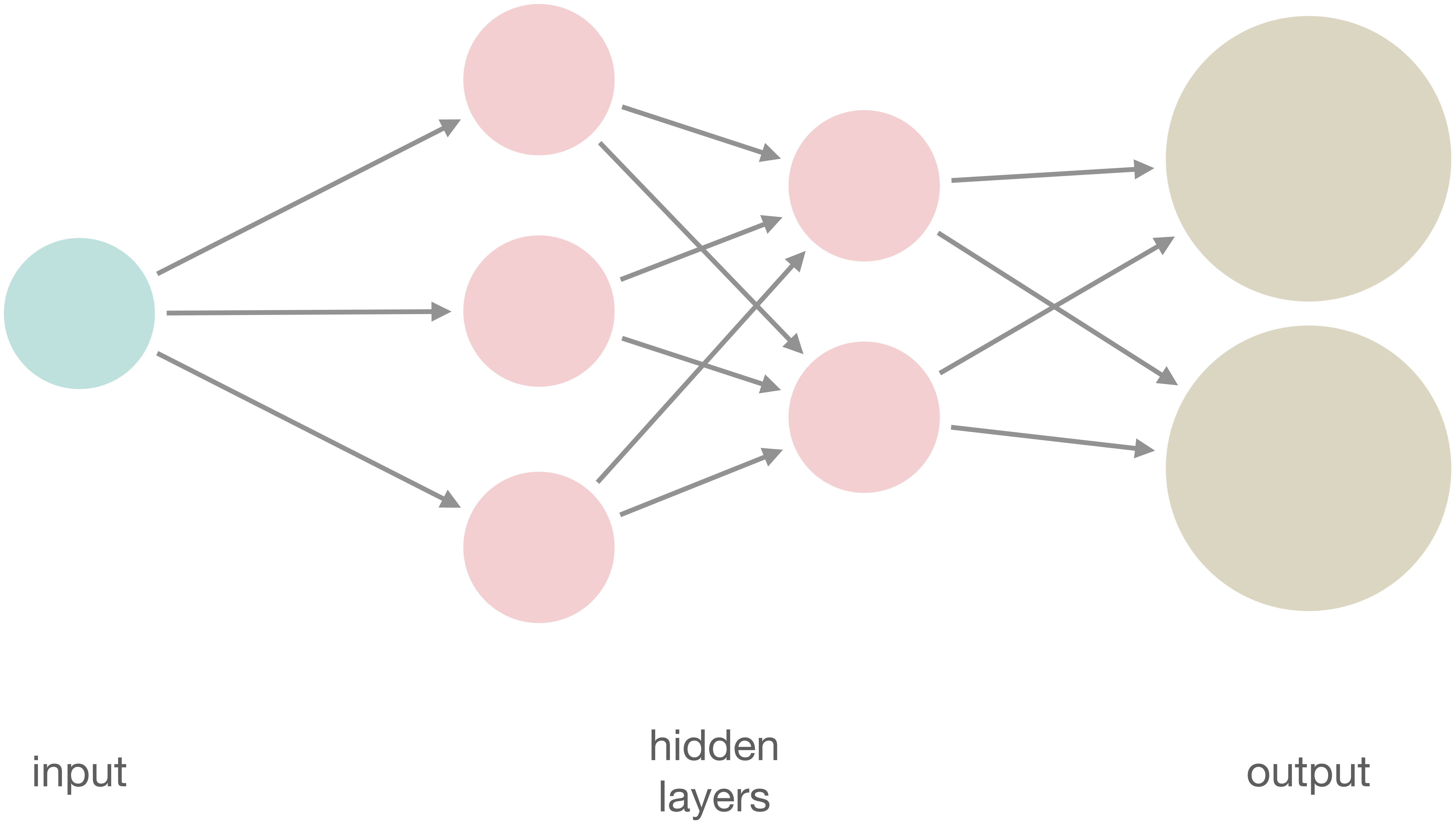
The Common Pipeline



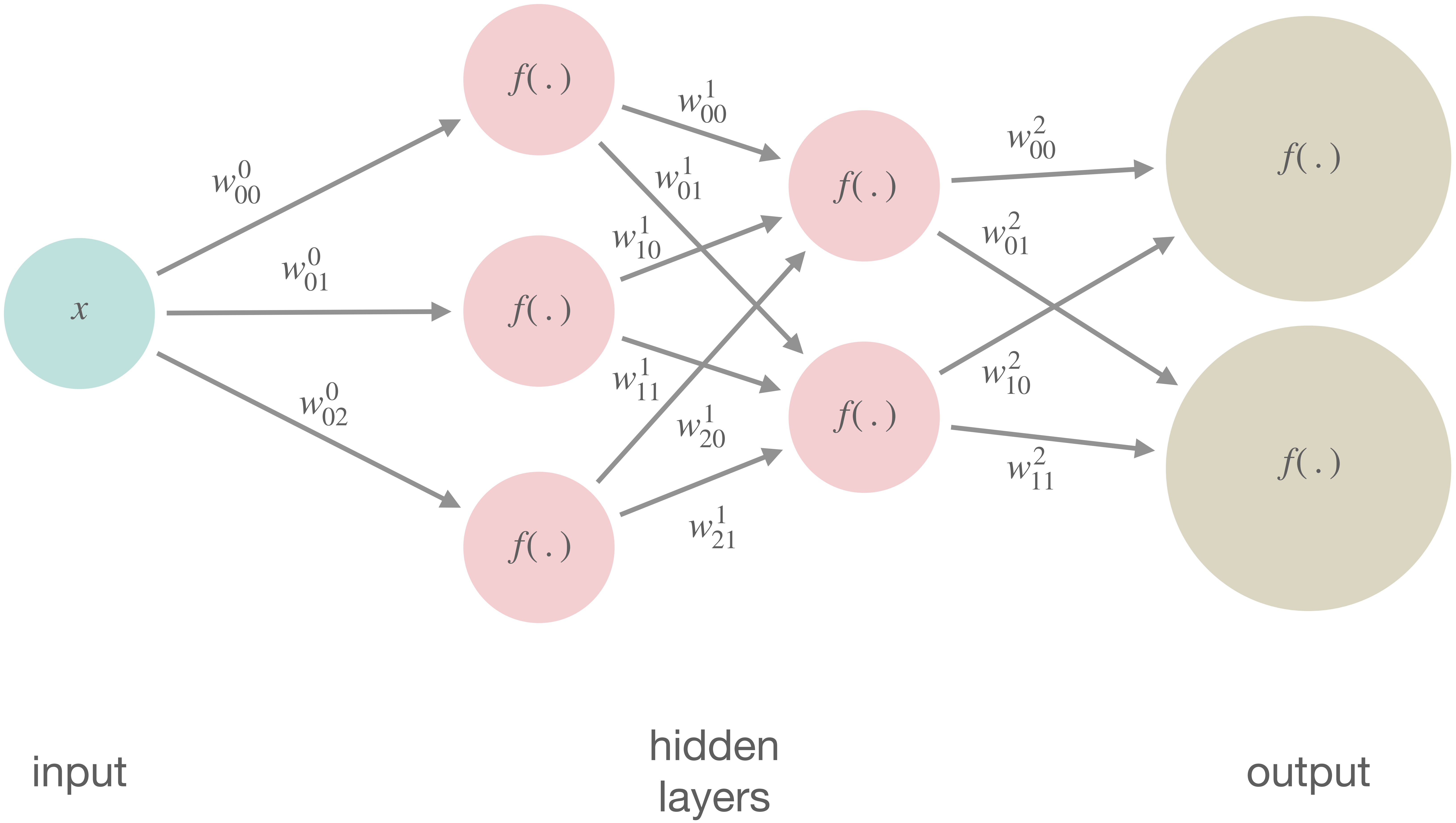
The Common Pipeline



The Common Pipeline

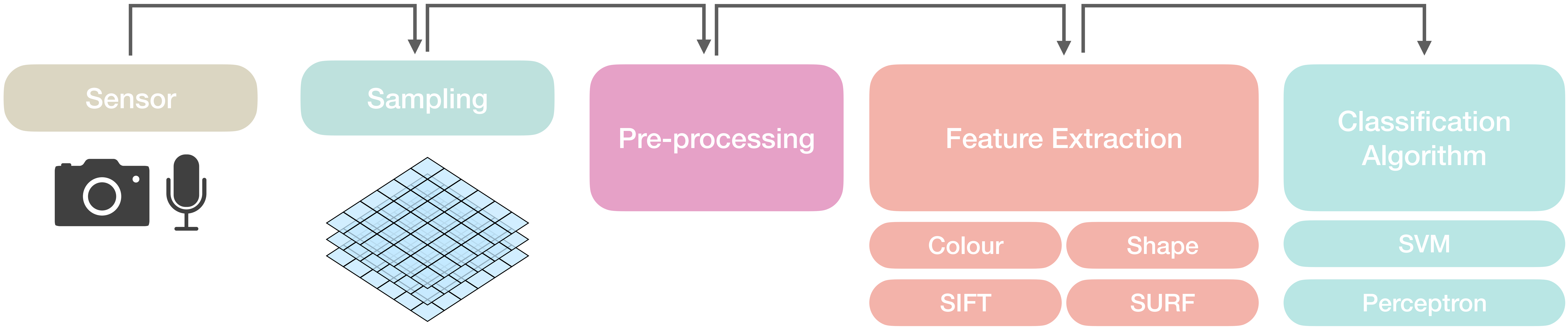


The Common Pipeline



Representation and Features

The classic pipeline considers separate steps for feature extraction and the learning algorithm



Representation and Features

> The classic pipeline considers separate steps for feature extraction and the learning algorithm

- Feature Extraction
  - Colour
  - Shape
  - SIFT
  - SURF

Representation and Features

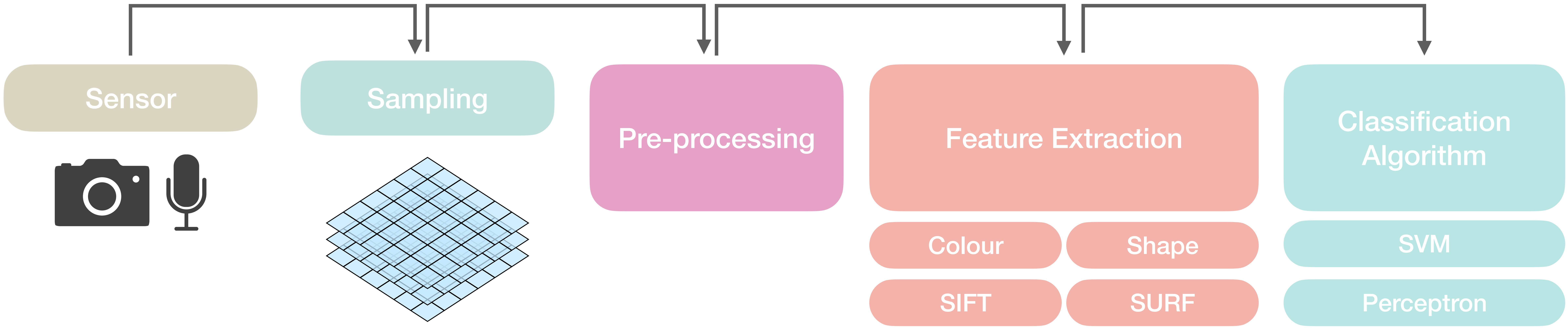
The classic pipeline considers separate steps for feature extraction and the learning algorithm

Feature Extraction

Colour	Shape	Region	Domain	Descriptors
RGB	Canny	Laplacian of Gaussian	Hough Transform	SIFT
LAB	Sobel	Difference of Gaussians	Wavelet Transform	SURF
HSL	Prewitt	Determinant of Hessian	Distance Transform	GLOH
LUV	Deriche			HOG

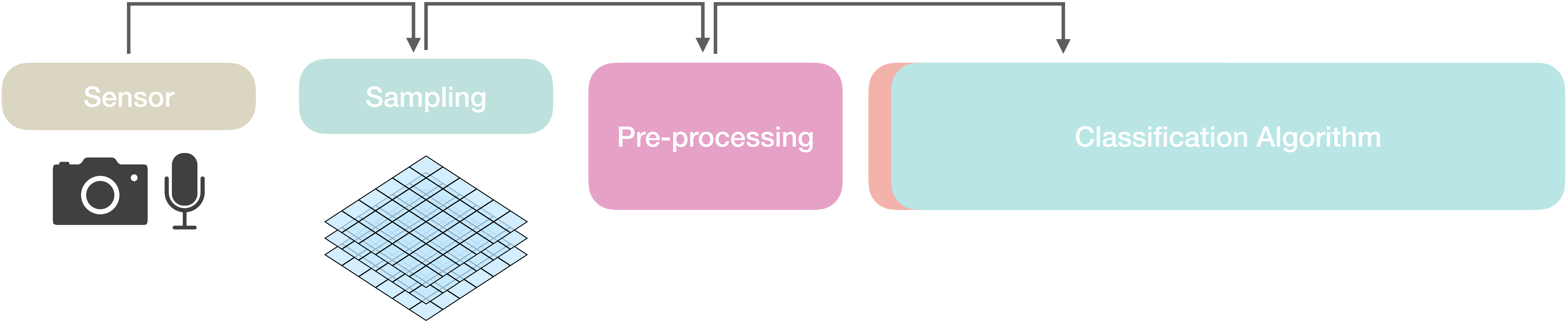
Representation and Features

What if we could learn these representations whilst making them tailored for the task



Representation and Features

What if we could learn these representations whilst making them tailored for the task



Representation and Features

Feature Extraction

Colour	Shape	Region	Domain	Descriptors
RGB	Canny	Laplacian of Gaussian	Hough Transform	SIFT
LAB	Sobel	Difference of Gaussians	Wavelet Transform	SURF
HSL	Prewitt	Determinant of Hessian	Distance Transform	GLOH
LUV	Deriche			HOG

Convolution

Mathematical Operation

Binary

N-dimensional tensors

Convolution

$$A \circledast B$$

a1	a2	a3	a4

b1	b2	b3	b4

Convolution

A $\circledast$ B

a1	a2	a3	a4
a5	a6	a7	a8
a9	a10	a11	a12

b1	b2	b3	b4
b5	b6	b7	b8
b9	b10	b11	b12

Convolution

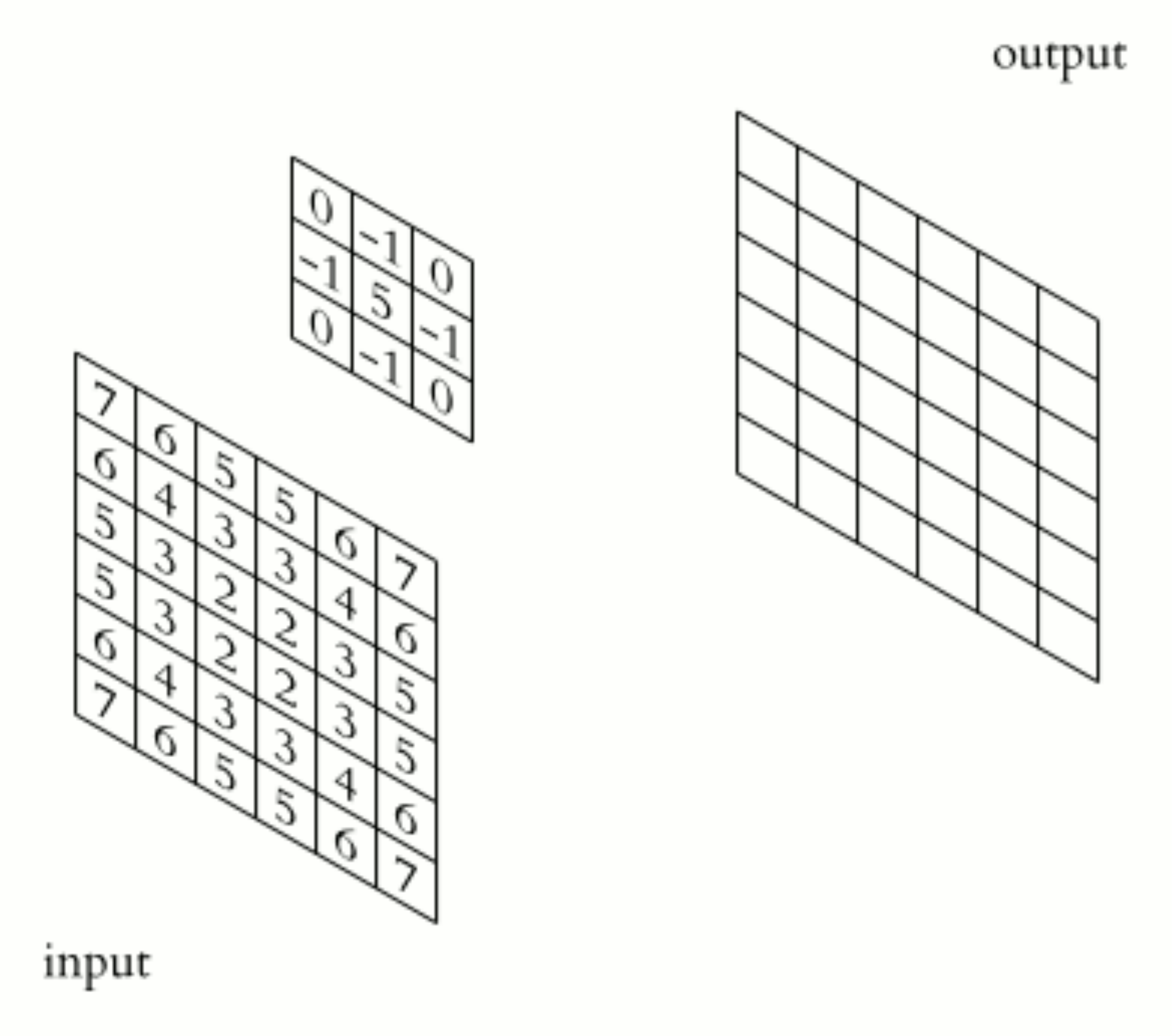
$$C = A \circledast B$$

$$C_{x,y} = \sum_{dx=-a}^{a} \sum_{dy=-b}^{b} A_{dx,dy} B_{x+dx,y+dy}$$

Convolution

$$C = A \circledast B$$

$$C_{x,y} = \sum_{dx=-a}^{a} \sum_{dy=-b}^{b} A_{dx,dy} B_{x+dx,y+dy}$$



Source: Michael Plotke, a wikipedia contributor

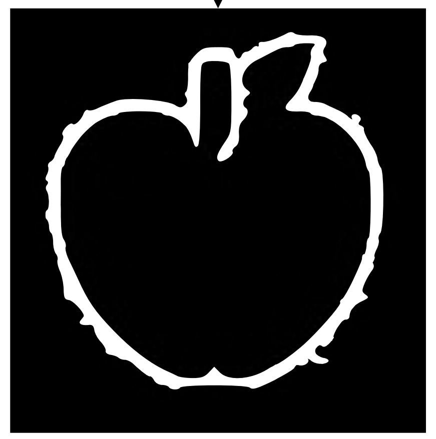
Convolution

$$\begin{bmatrix} 0.0625 & 0.125 & 0.0625 \\ 0.125 & 0.25 & 0.125 \\ 0.0625 & 0.125 & 0.0625 \end{bmatrix} \quad \begin{bmatrix} -2 & -1 & 0 \\ -1 & 1 & 1 \\ 0 & 1 & 2 \end{bmatrix} \quad \begin{bmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{bmatrix} \quad \begin{bmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{bmatrix}$$

gaussian blur

emboss

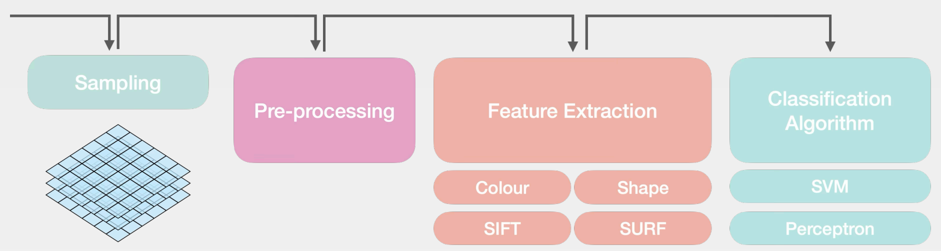
left sobel



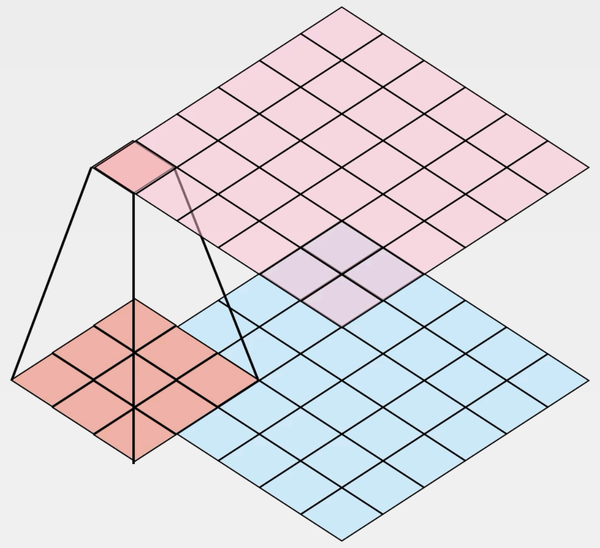
outline (high pass)

SCC0251
Processamento de Imagens

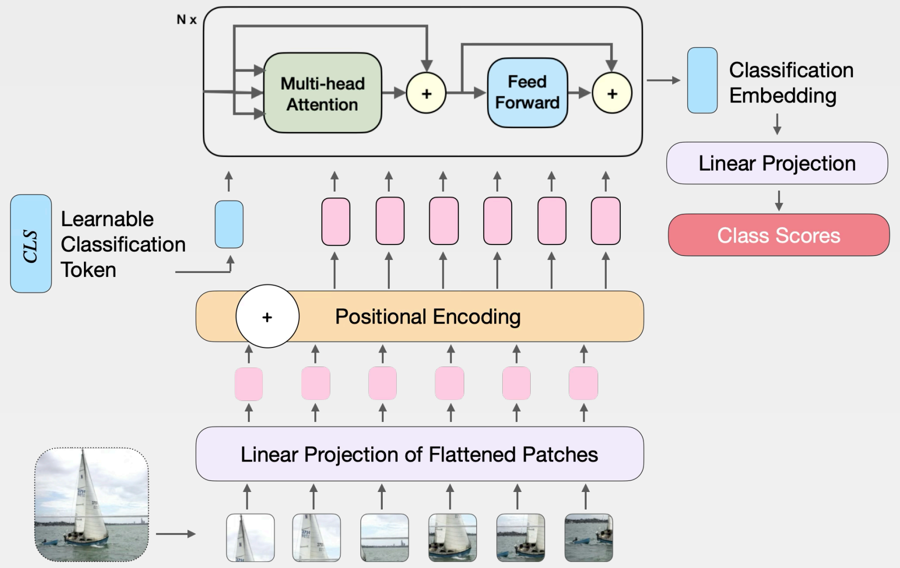
Aprendizado Profundo



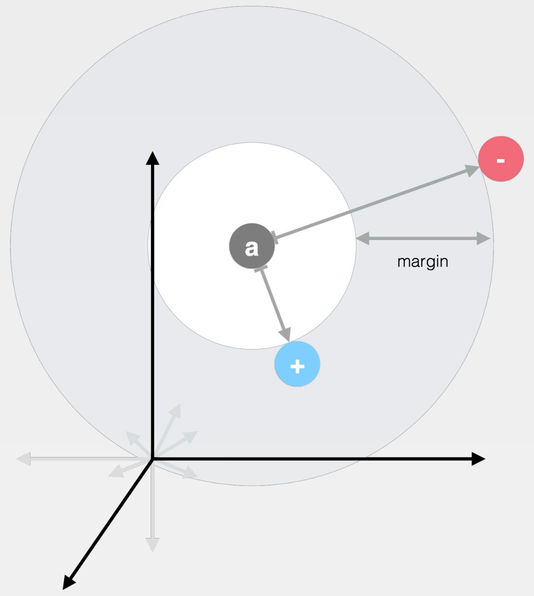
Classic Pipeline



CNNs



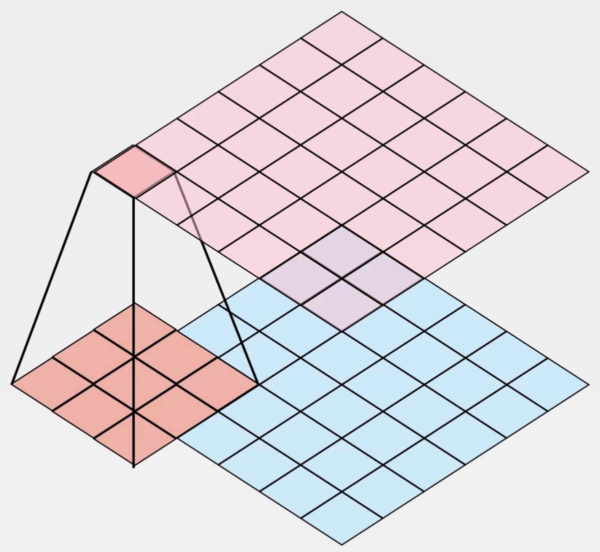
Transformers



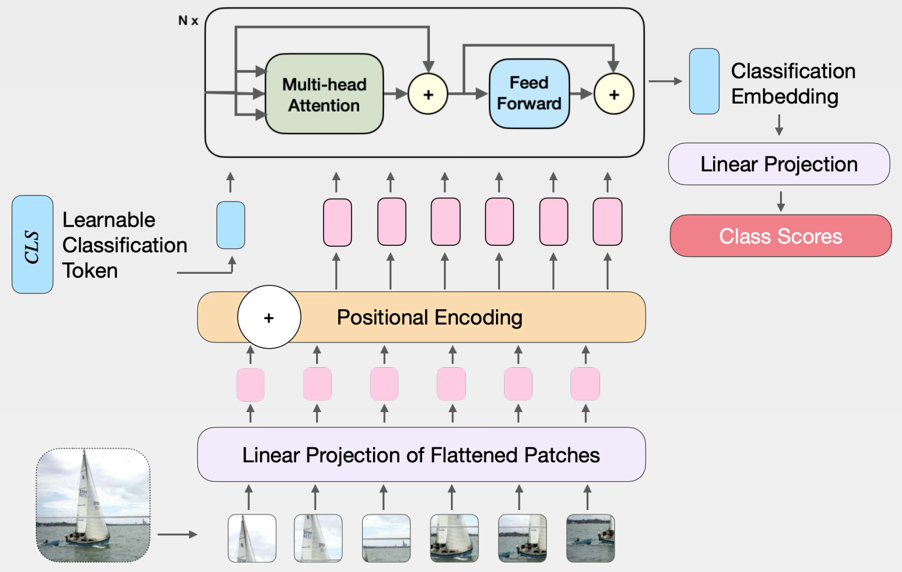
Contrastive Learning

SCC0251
Processamento de Imagens

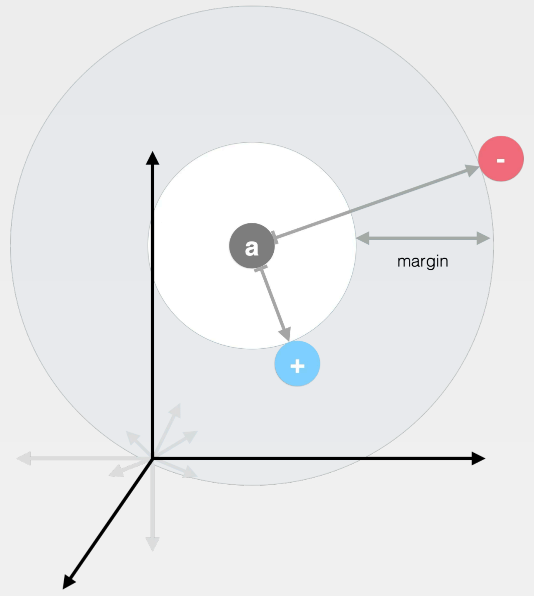
Aprendizado Profundo



CNNs

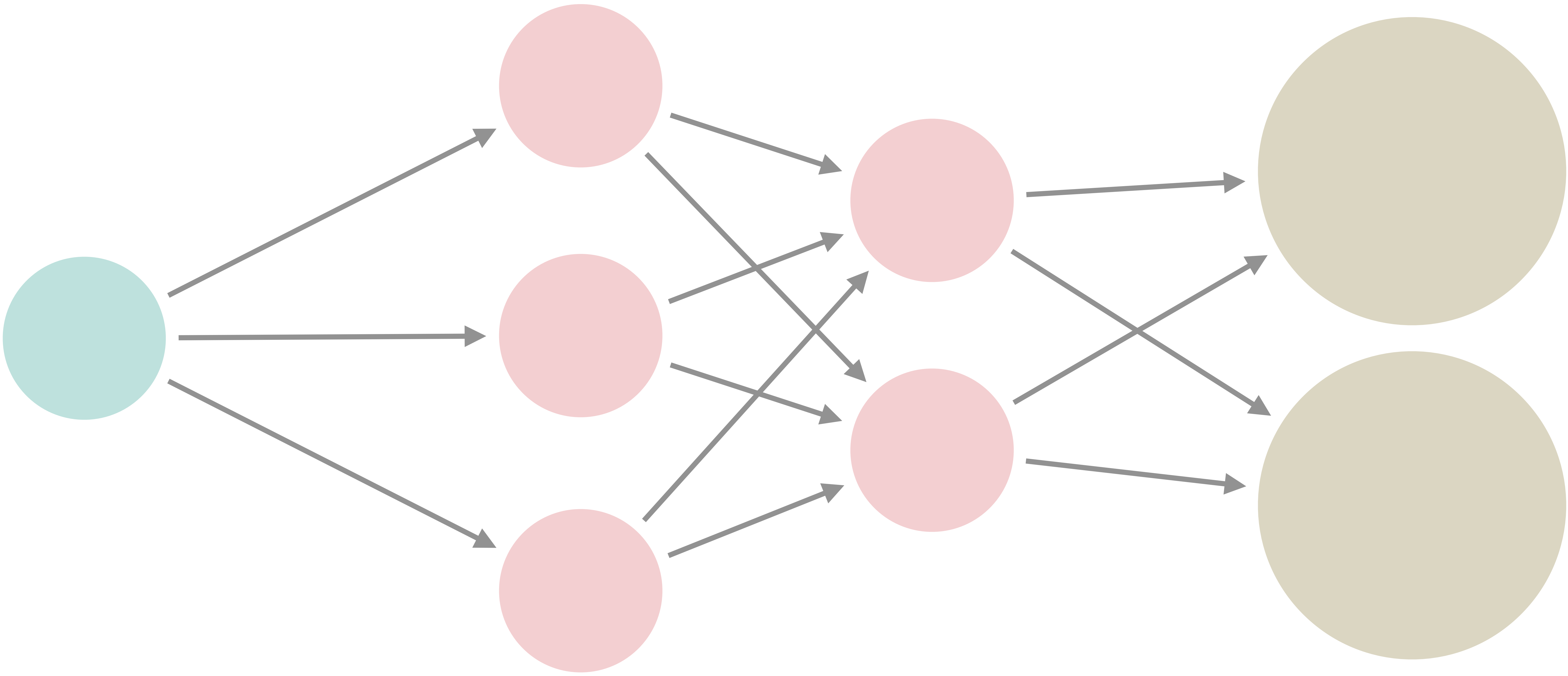


Transformers

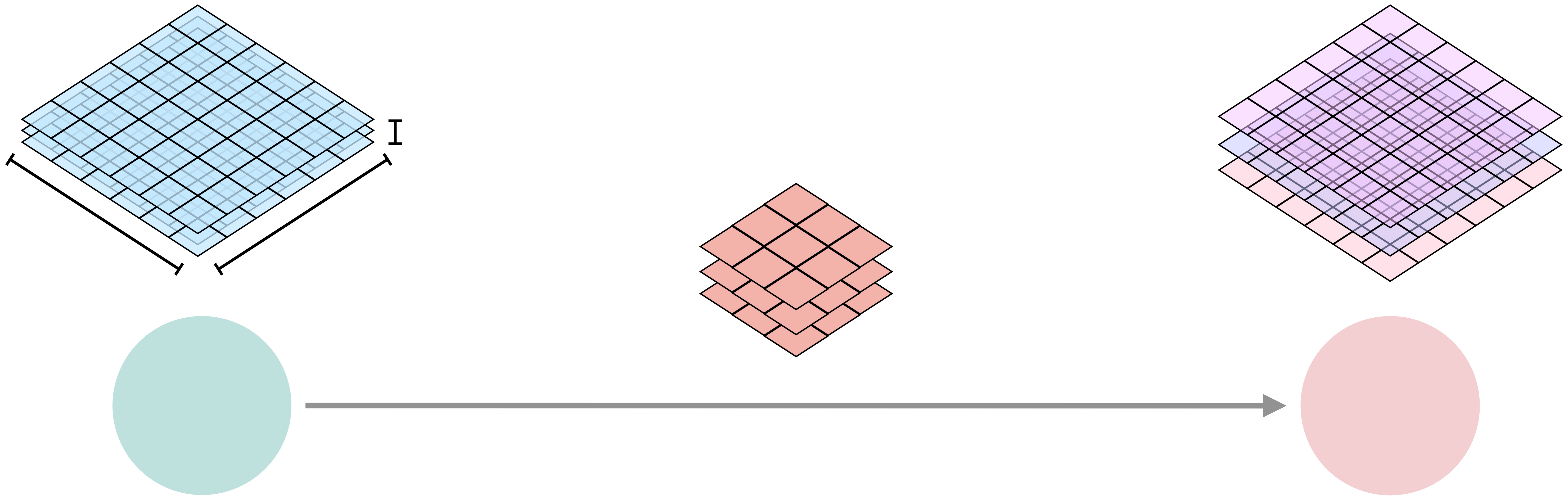


Contrastive Learning

Convolutional Neural Network

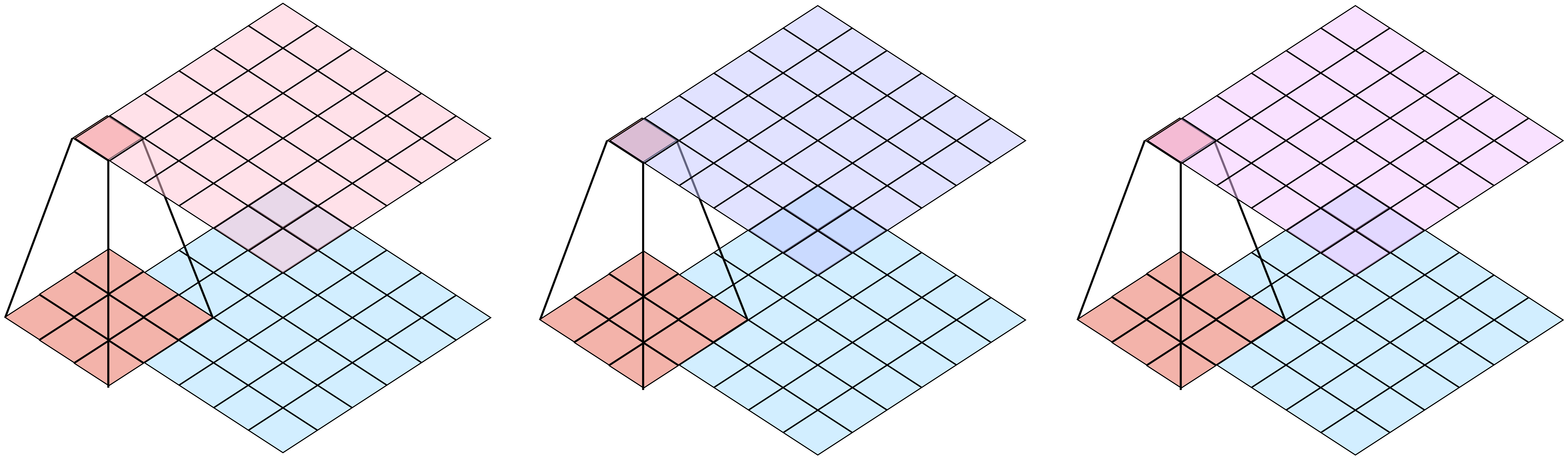


Convolutional Neural Network



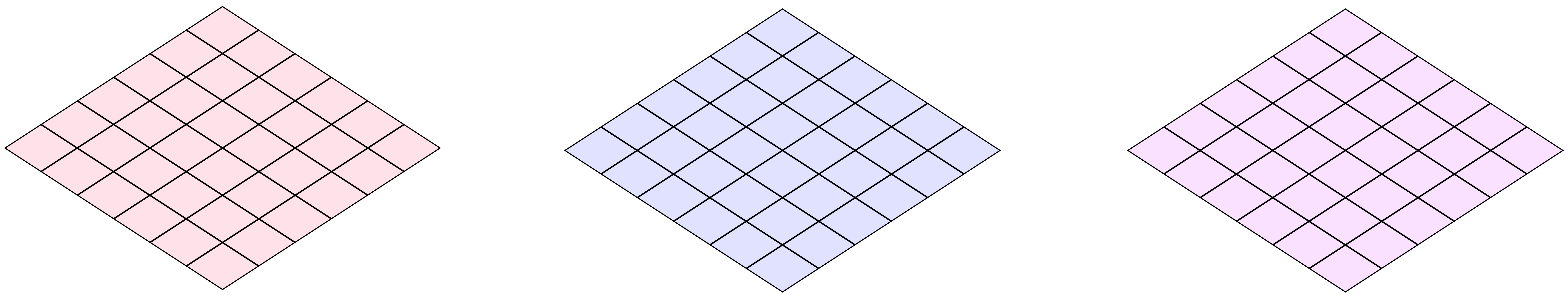
Convolutional Neural Network

$$C = A \circledast B \qquad C_{x,y} = \sum_{dx=-a}^{a} \sum_{dy=-b}^{b} A_{dx,dy} B_{x+dx,y+dy}$$



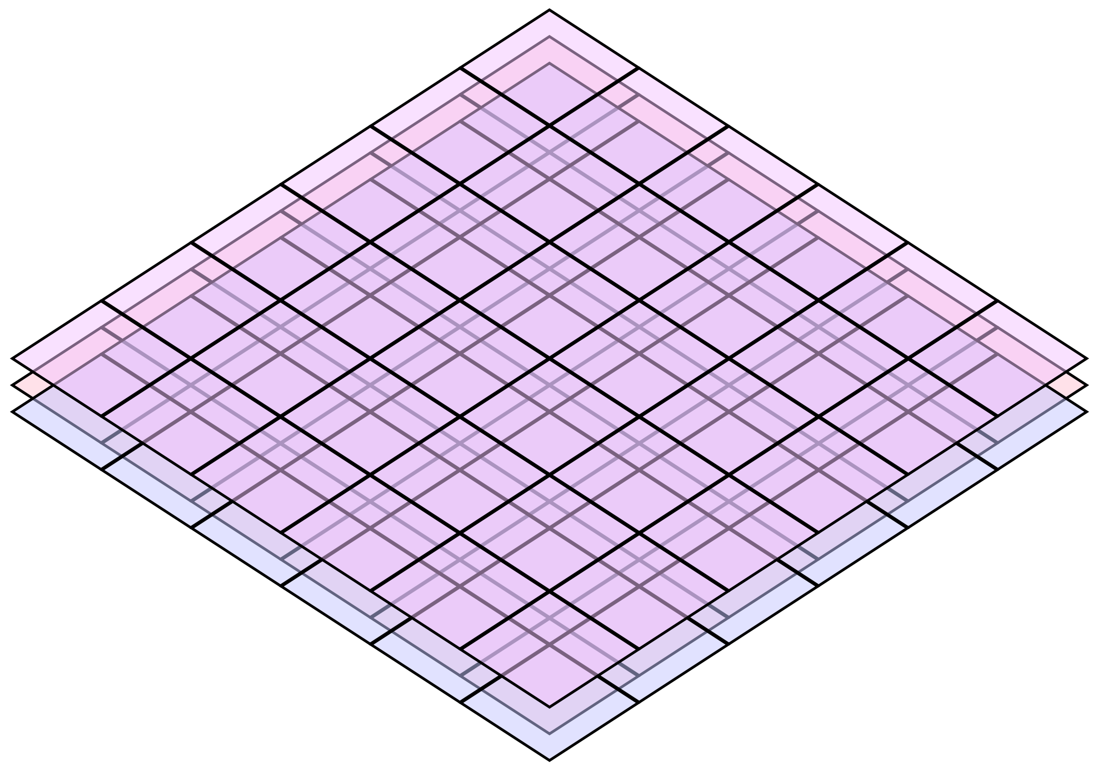
each input map is convolved with a kernel

Convolutional Neural Network



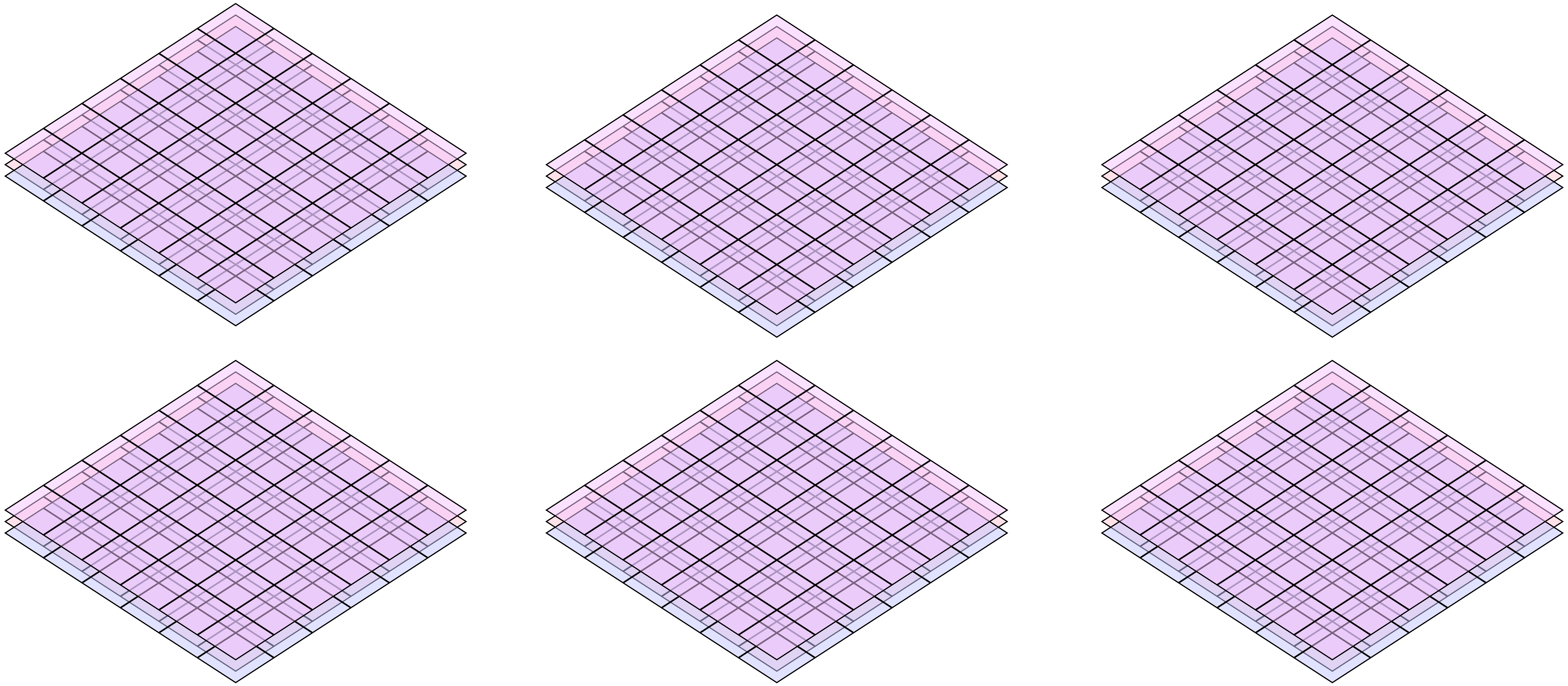
each input map is convolved with a kernel

Convolutional Neural Network



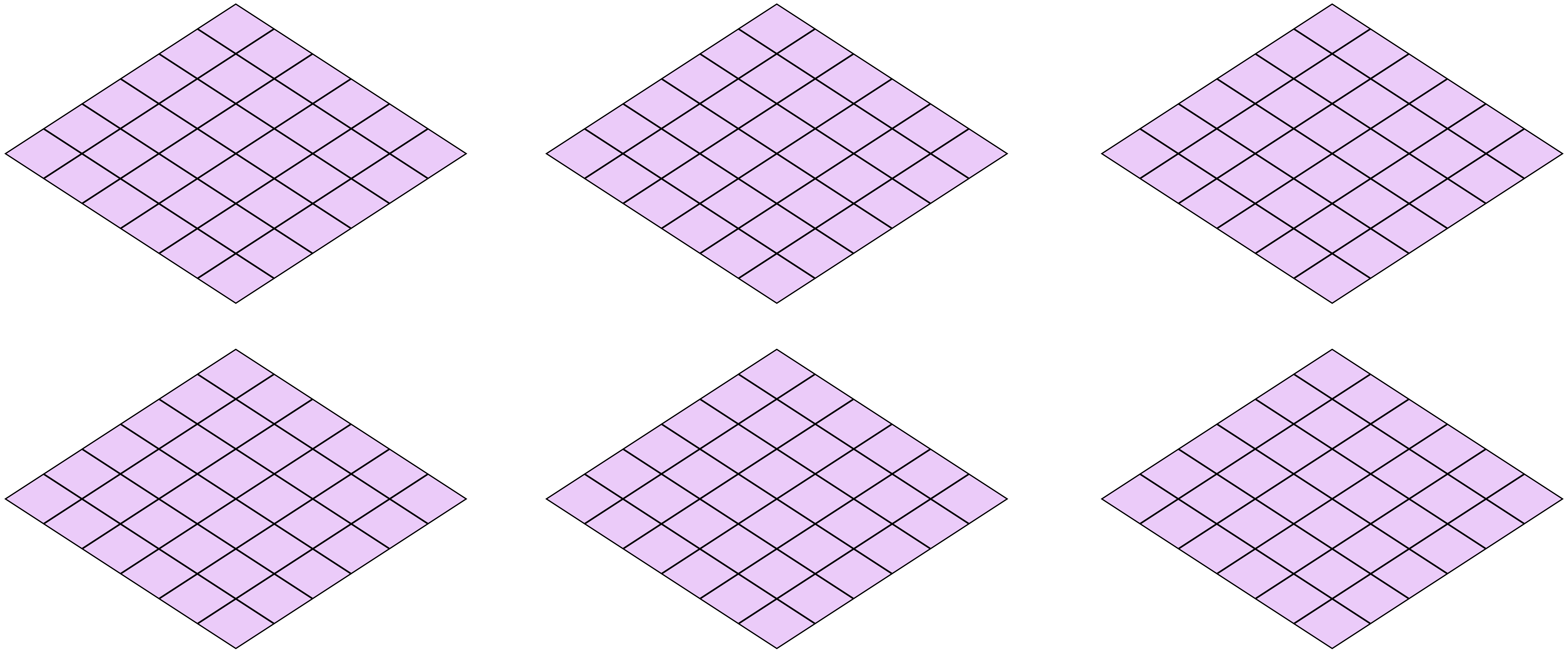
and the resulting activations are summed together

Convolutional Neural Network



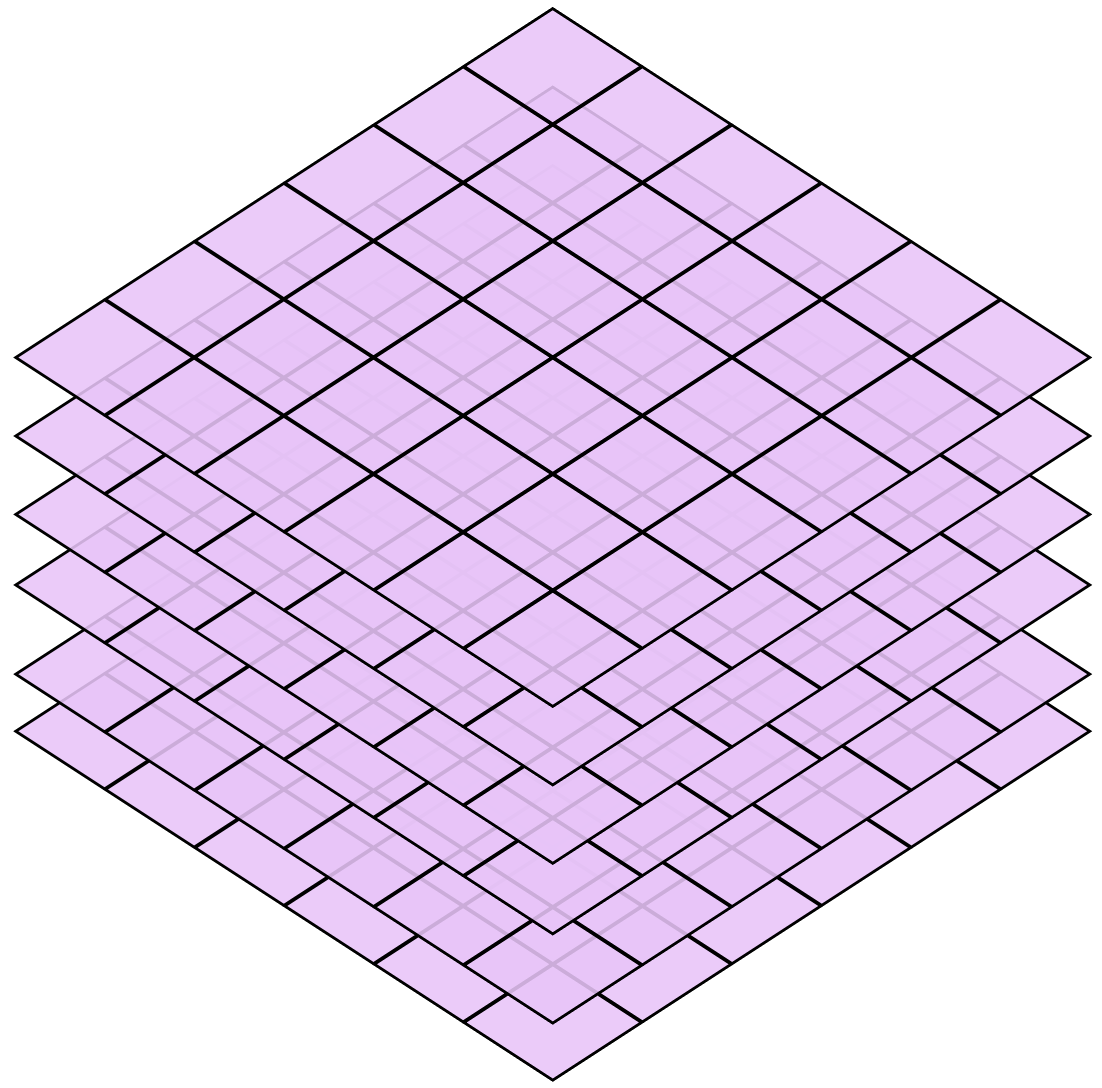
a process that is repeated for all filters

Convolutional Neural Network



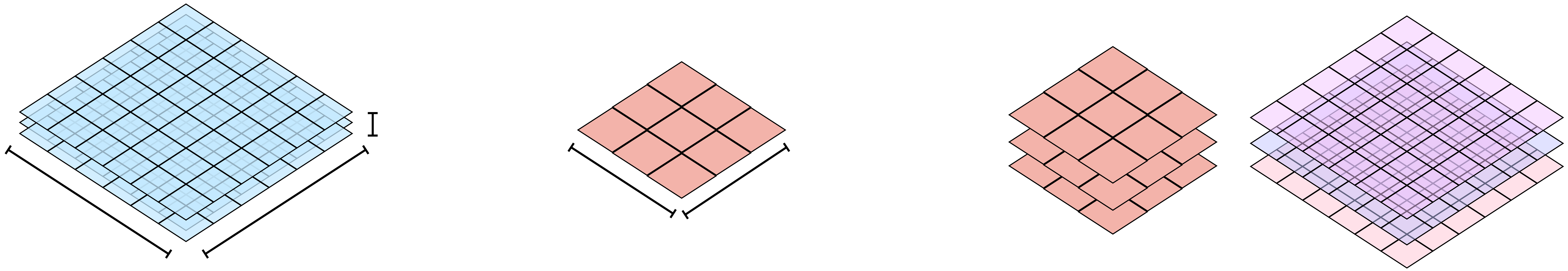
a process that is repeated for all filters

Convolutional Neural Network



and gives us the same number of activation maps as the number of filters

Convolutional Neural Network



input maps

single kernel

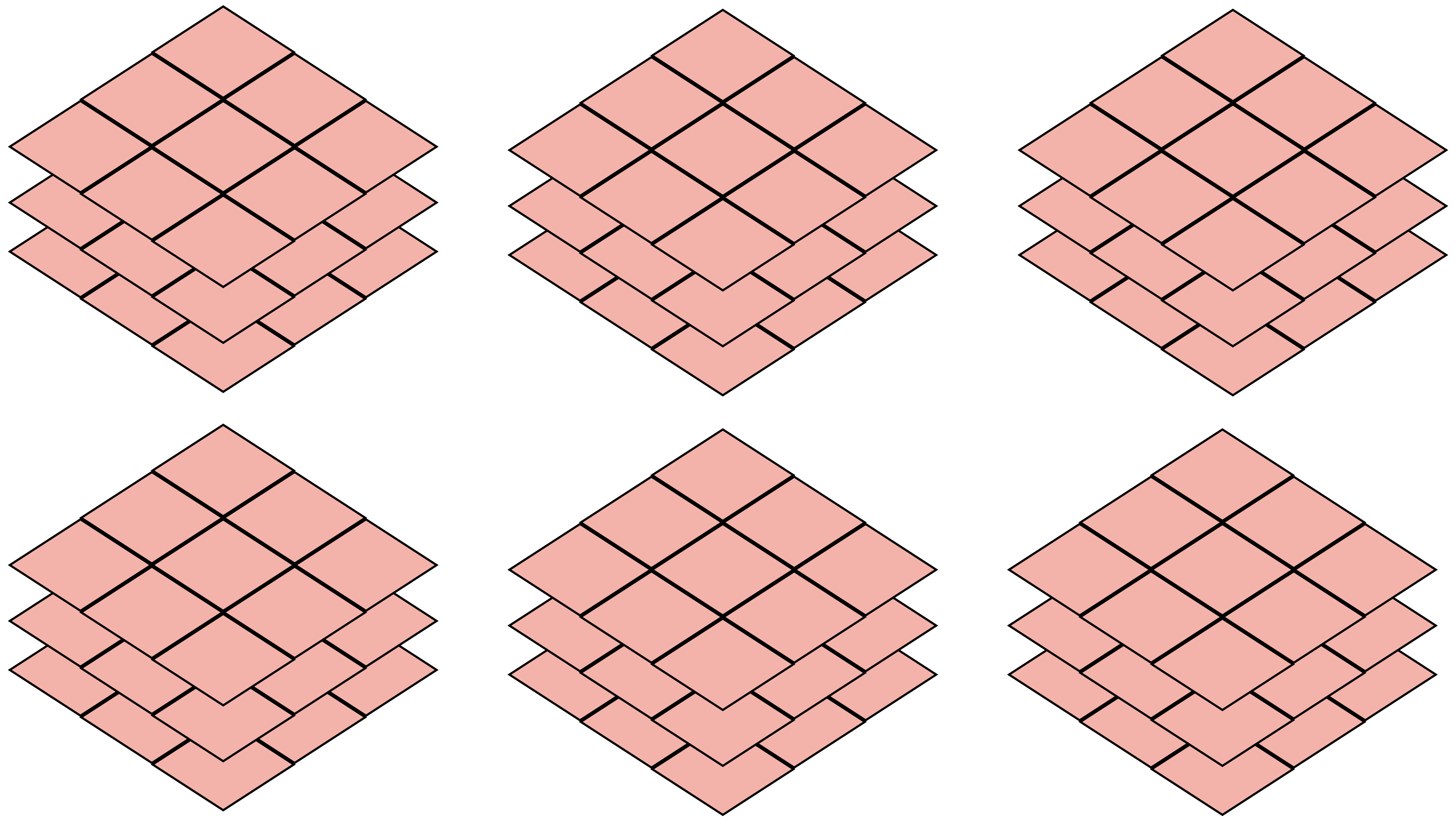
filter and filter output

(n, m, c_in)

(k, k)

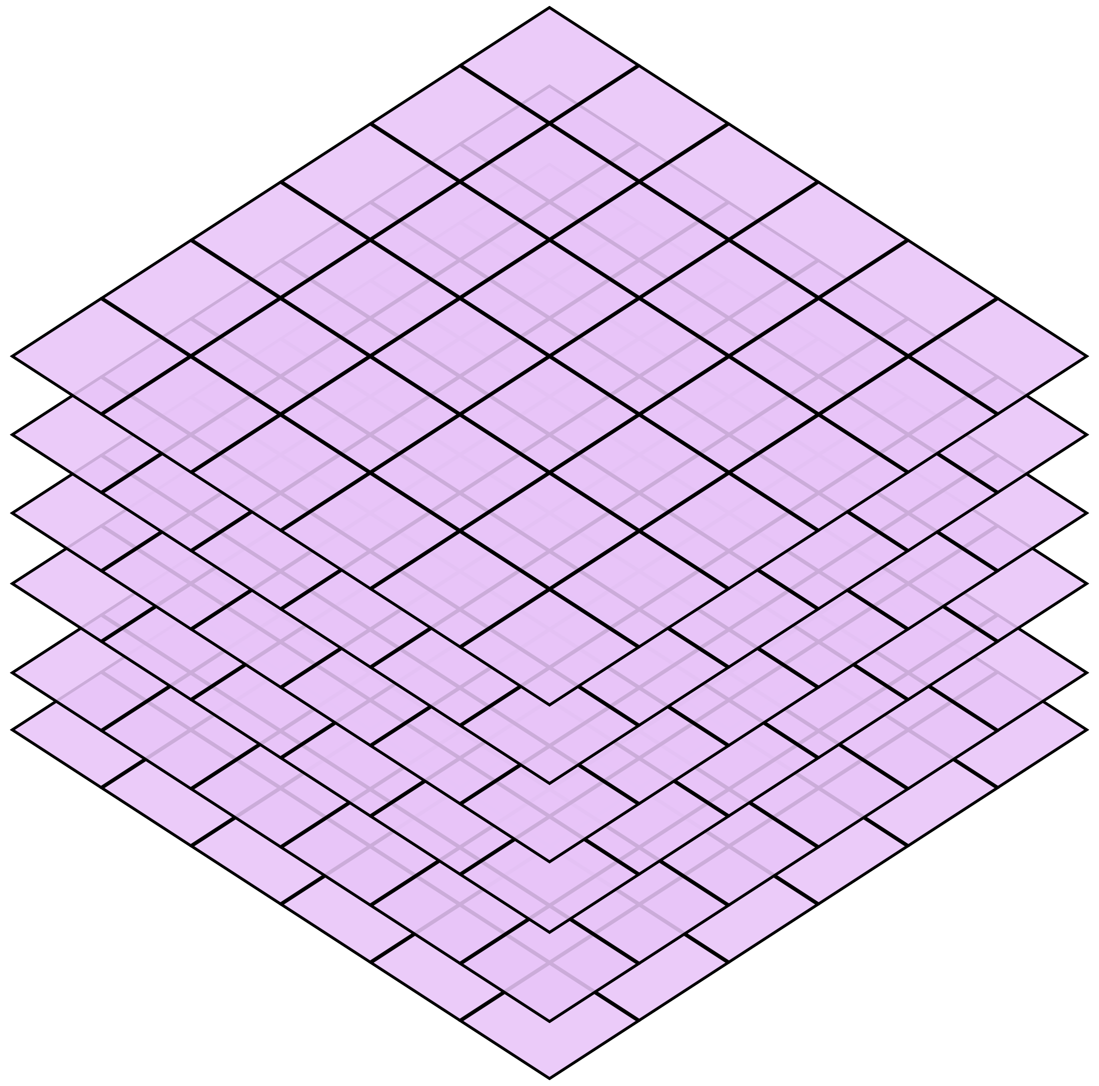
(k, k, c_in) (n, m, c_in)

Convolutional Neural Network



filter collection

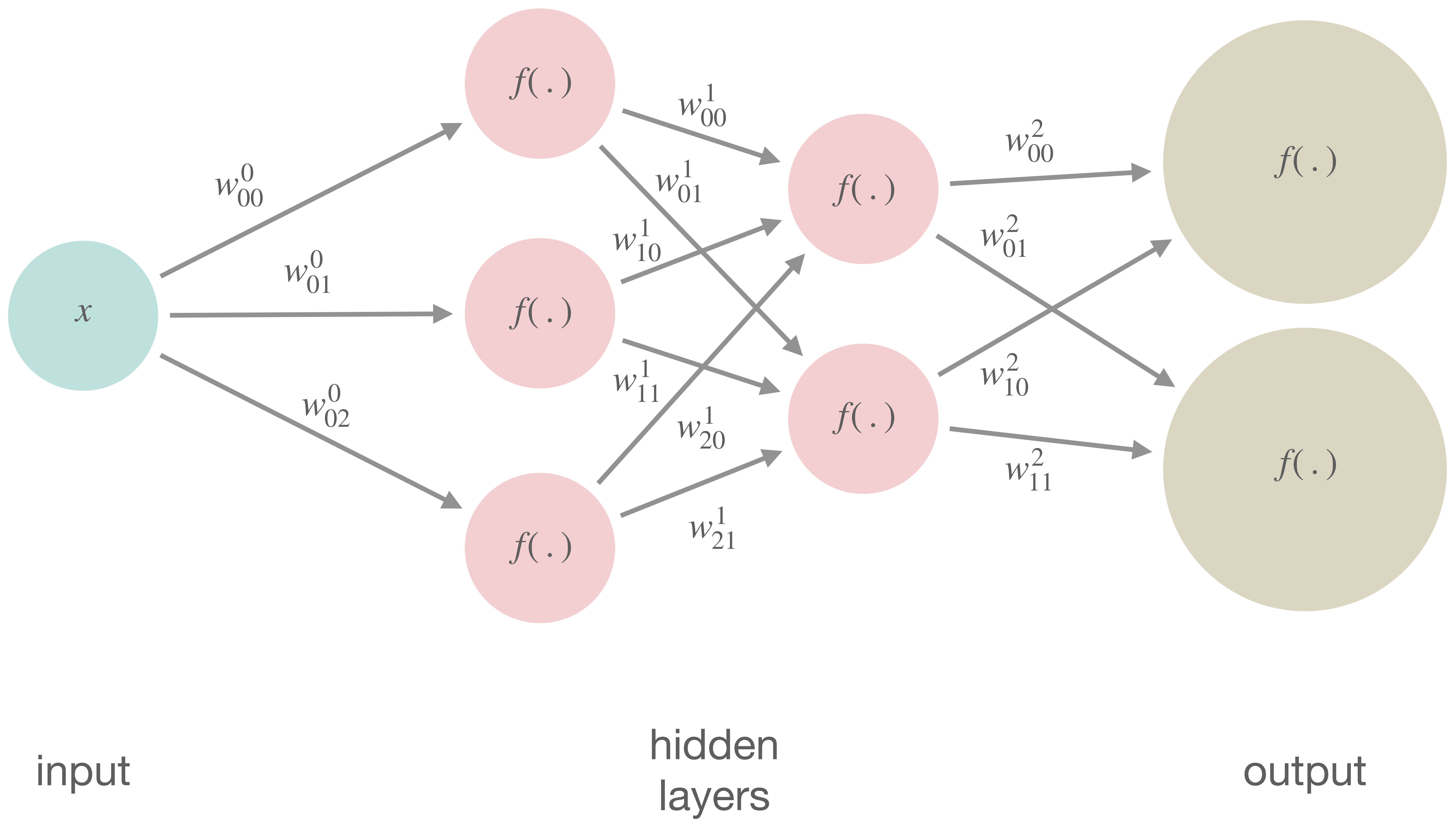
(k, k, c_in, c_out)



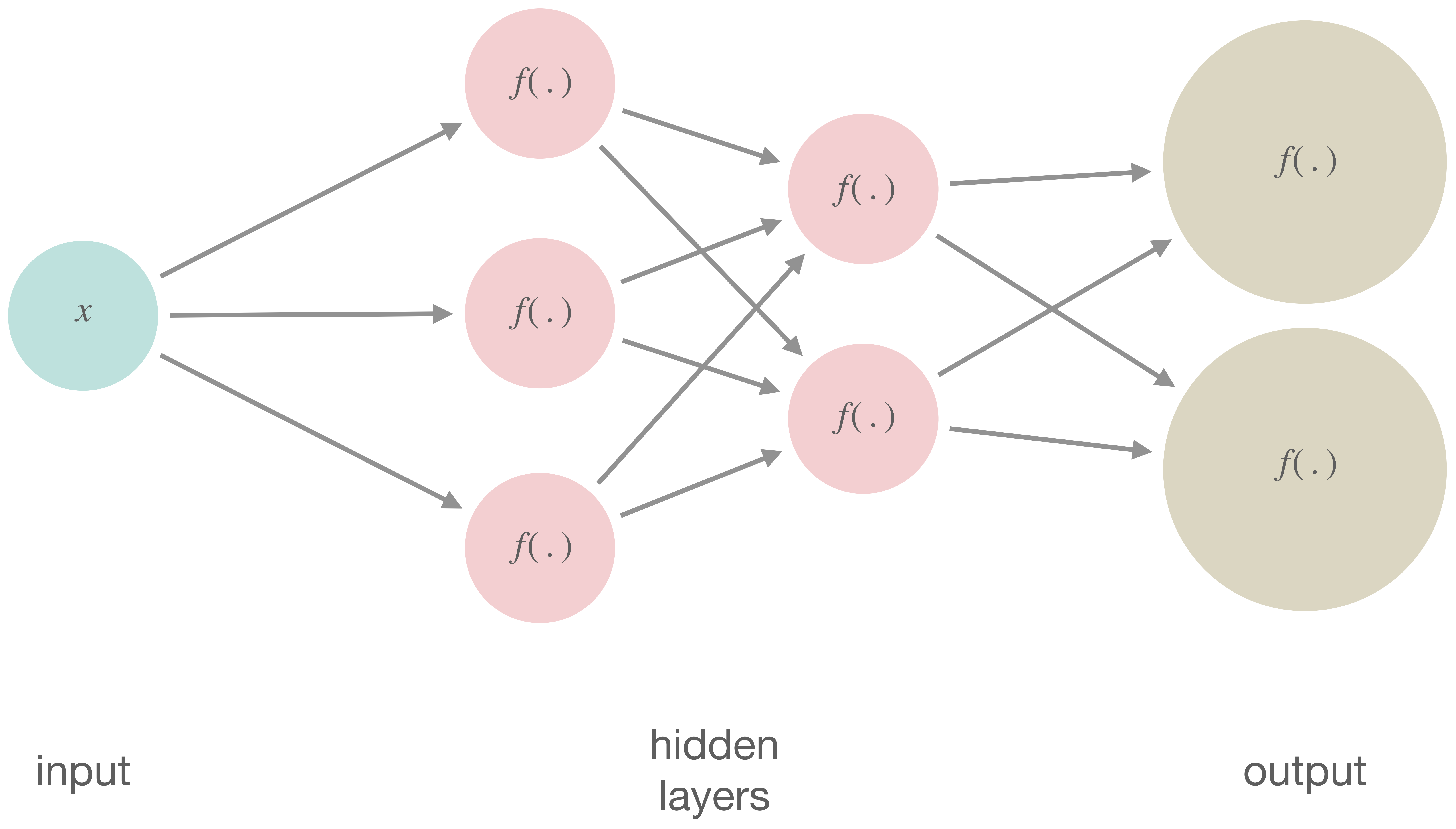
output feature maps

(n, m, c_out)

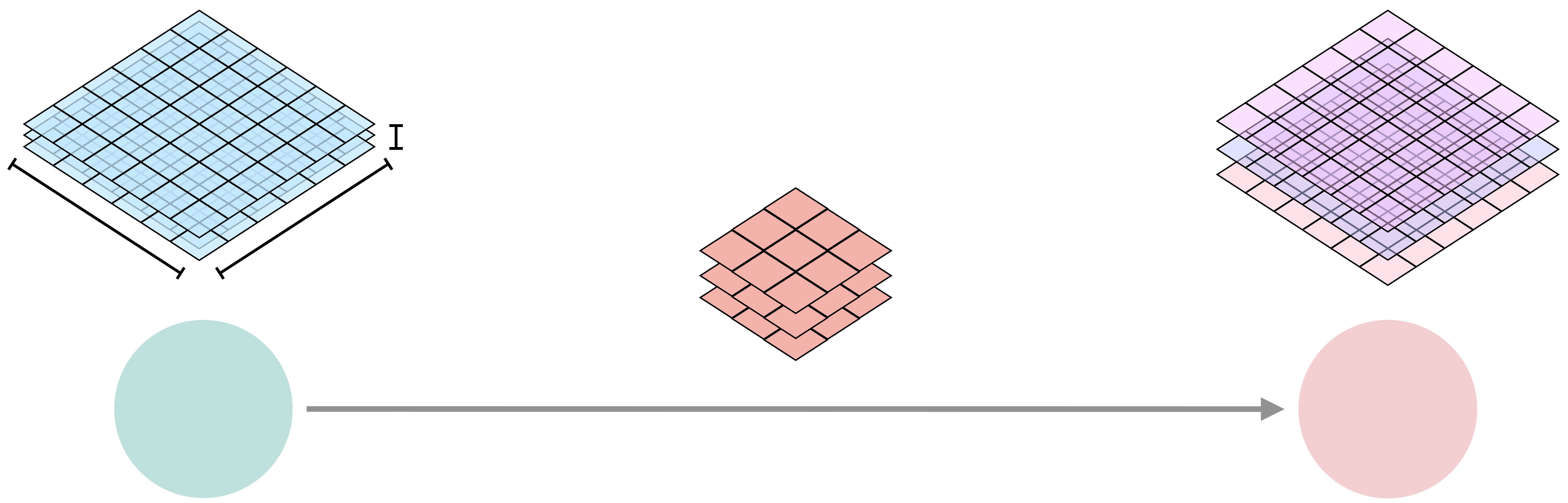
The Common Pipeline



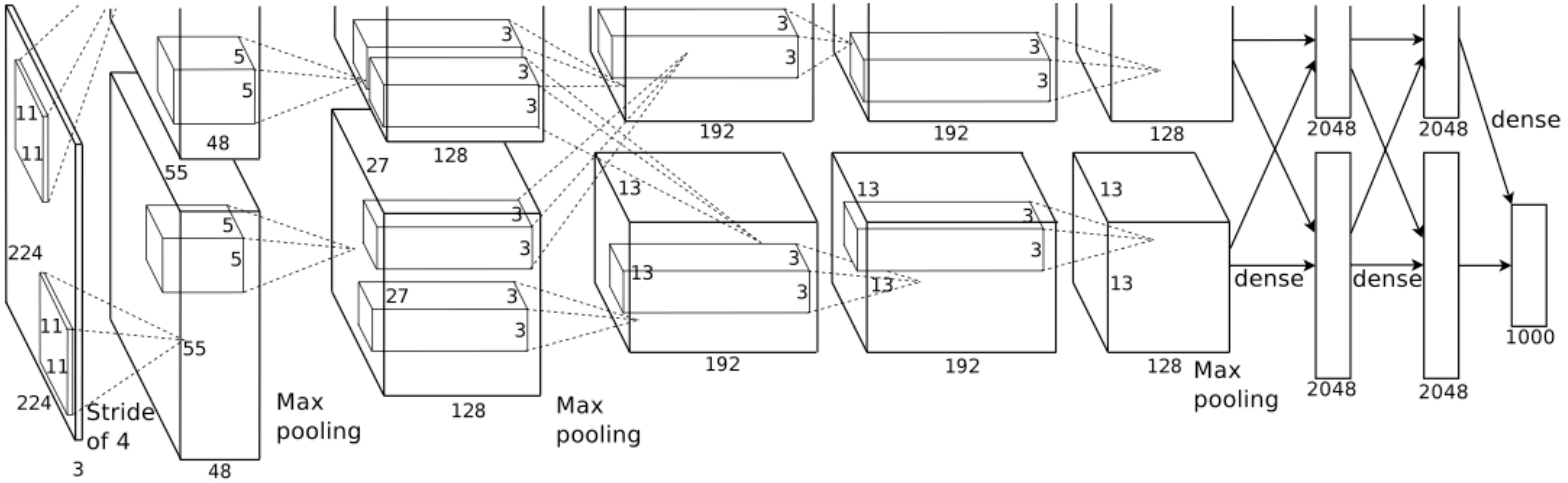
The Common Pipeline



Convolutional Neural Network



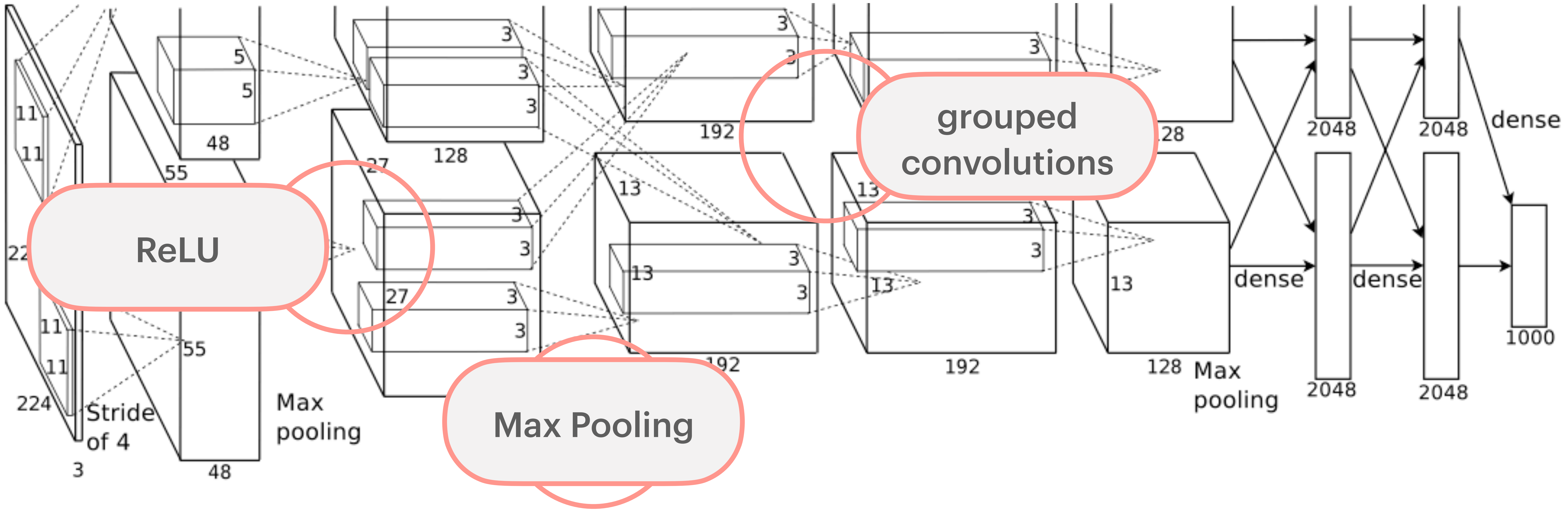
history of scaling in DL



the publication of the AlexNet in 2011 was a turning point for scale in Deep Learning

it was the deepest network yet, thanks to clever optimisations and design choices

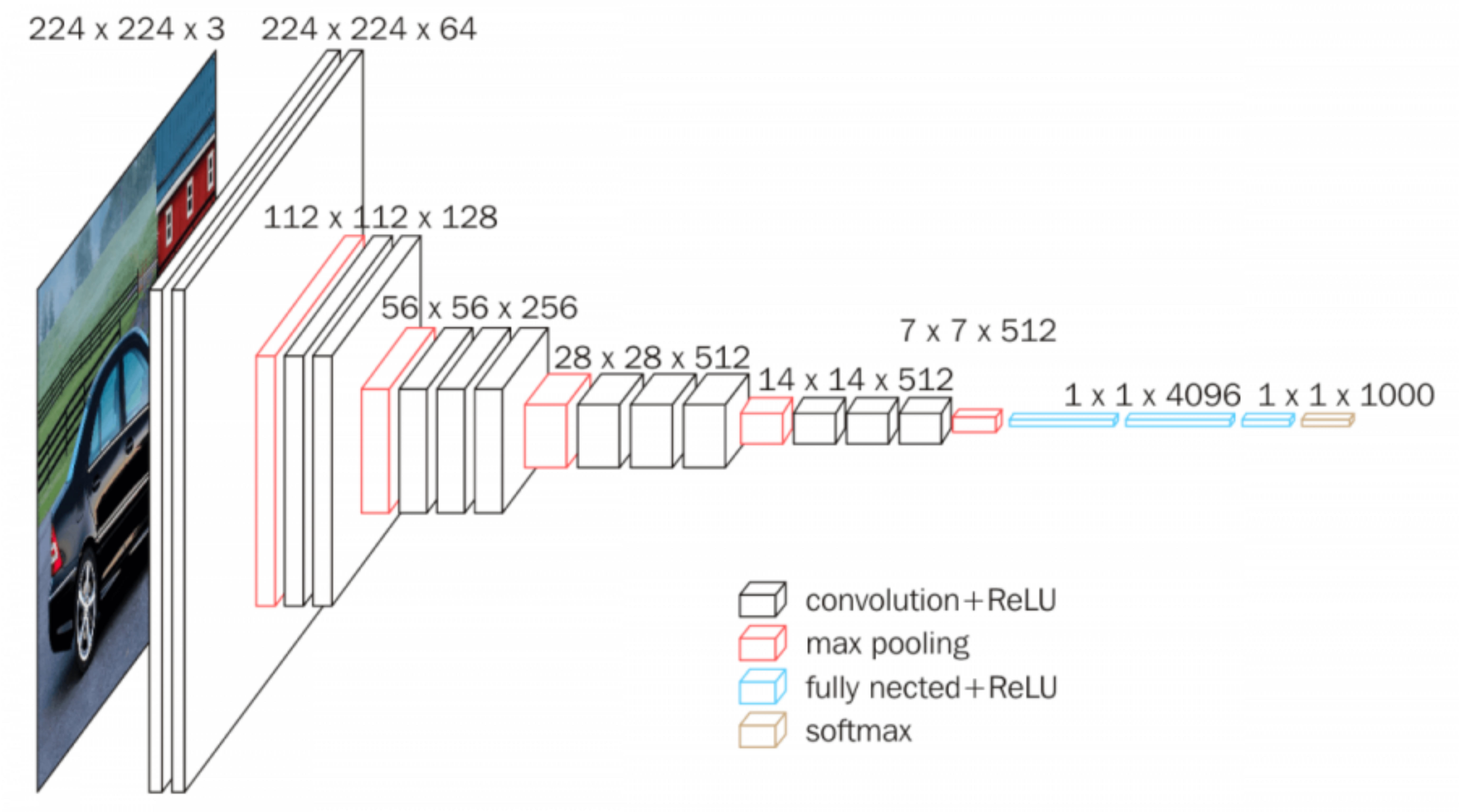
history of scaling in DL



the publication of the AlexNet in 2011 was a turning point for scale in Deep Learning

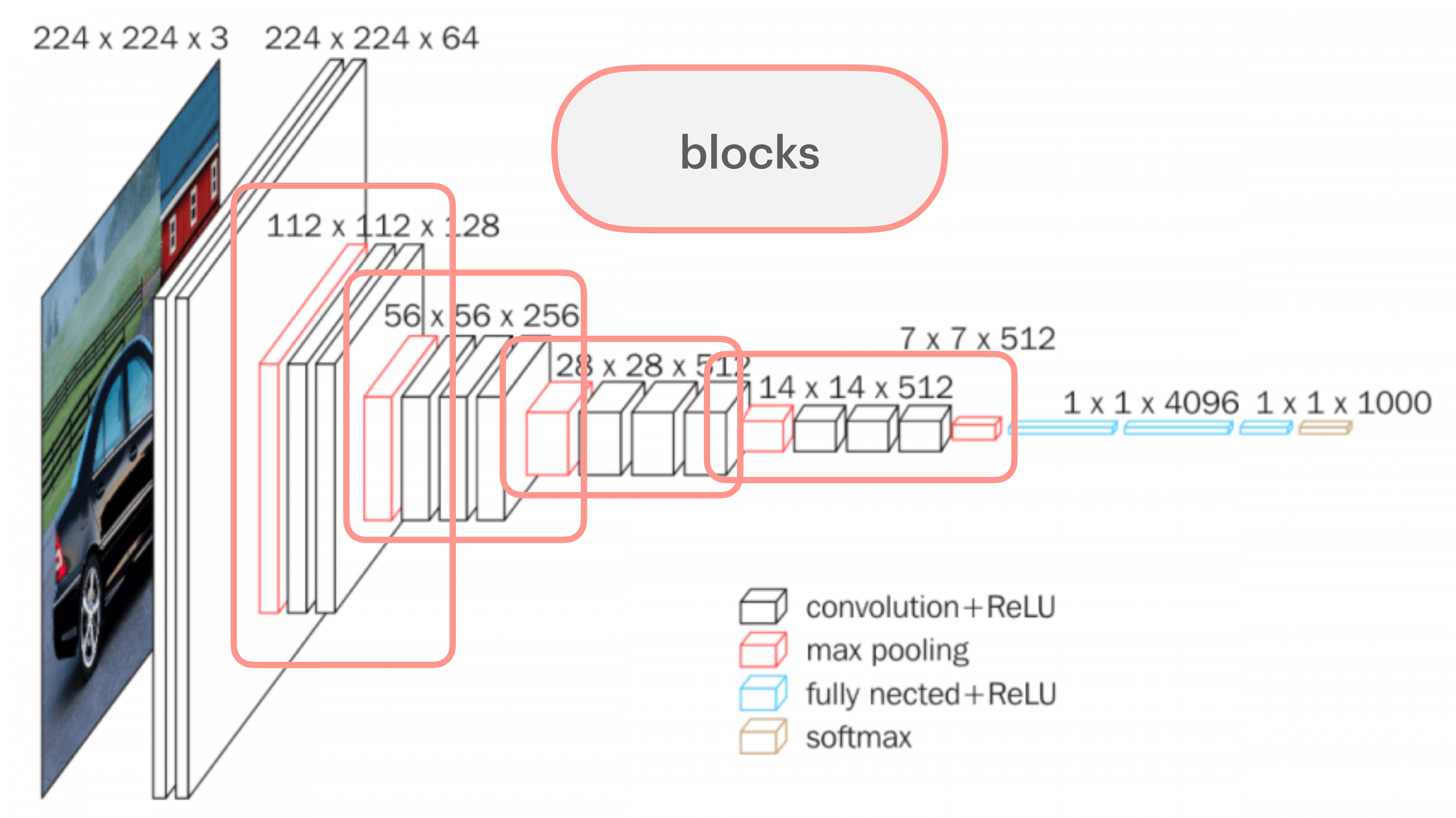
it was the deepest network yet, thanks to clever optimisations and design choices

history of scaling in DL



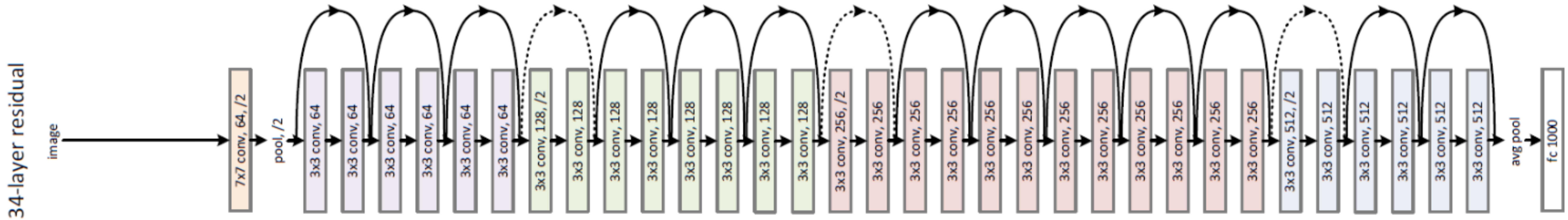
VGG came next and brought with it the idea of blocks based on the current image size

history of scaling in DL



VGG came next and brought with it the idea of blocks based on the current image size

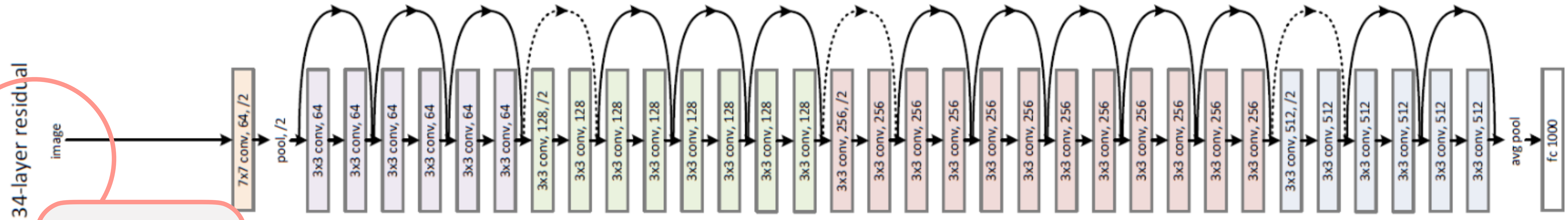
history of scaling in DL



by implementing the residual connections, ResNet overcame the vanishing gradient problem

which finally opened the door for networks to scale ‘unbounded’ in the depth dimension

history of scaling in DL



34-layers

50-layers

101-layers

152-layers

each with improved accuracy

by implementing the residual connections, ResNet overcame the vanishing gradient problem

which finally opened the door for networks to scale ‘unbounded’ in the depth dimension

history of scaling in DL

SERGEY ZAGORUYKO AND NIKOS KOMODAKIS: WIDE RESIDUAL NETWORKS 1

14 Jun 2017

Wide Residual Networks

Sergey Zagoruyko
sergey.zagoruyko@enpc.fr
Nikos Komodakis
nikos.komodakis@enpc.fr
Université Paris-Est, École des Ponts
ParisTech
Paris, France

by scaling width as well as depth, managed to get a better results with less depth

MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications

Andrew G. Howard Menglong Zhu Bo Chen Dmitry Kalenichenko
Weijun Wang Tobias Weyand Marco Andreetto Hartwig Adam

Google Inc.
{howarda,menglong,bochen,dkalenichenko,weijunw,weyand,anm,hadam}@google.com

used depth-wise separable convolutions extensively and has parameters for scaling both width and resolution

a pattern of three scaling dimensions started to take shape

history of scaling in DL

MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications

Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, Hartwig Adam

Google Inc.

{howarda,menglong,bochen,dkalenichenko,weijunw,weyand,anm,hadam}@google.com

used depth-wise separable convolutions extensively and has parameters for scaling both width and resolution

GPipe: Easy Scaling with Micro-Batch Pipeline Parallelism

Yanping Huang huangyp@google.com
Youlong Cheng ylc@google.com
Ankur Bapna ankurbpn@google.com

showed that scaling resolution also had a significant impact on performance

a pattern of three scaling dimensions started to take shape

history of scaling in DL

GPipe: Easy Scaling with Micro-Batch Pipeline Parallelism

Yanping Huang
huangyp@google.com

Youlong Cheng
ylc@google.com

Ankur Bapna
ankurbpn@google.com

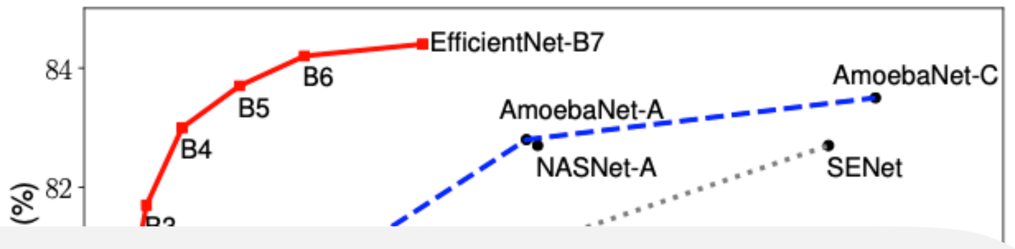
showed that scaling resolution also had a significant impact on performance

EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks

Mingxing Tan [1] Quoc V. Le [1]

Abstract

Convolutional Neural Networks (ConvNets) are commonly developed at a fixed resource budget, and then scaled up for better accuracy if more



a scalable model across all dimensions

a pattern of three scaling dimensions started to take shape

history of scaling in DL

depth increases computational cost linearly

resolution increases computational cost quadratically

width increases computational cost quadratically

wider nets have better gradients and are easier to train

Zagoruyko et al.

deeper nets perform better on single-object classes

Nguyen et al.

wider nets perform better on classes that represent scenes

Nguyen et al.

Applications

arXiv:1506.01497v3 [cs.CV] 6 Jan 2016

1

Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun

Abstract—State-of-the-art object detection networks depend on region proposal algorithms to hypothesize object locations. Advances like SPPnet [1] and Fast R-CNN [2] have reduced the running time of these detection networks, exposing region proposal computation as a bottleneck. In this work, we introduce a *Region Proposal Network* (RPN) that shares full-image convolutional features with the detection network, thus enabling nearly cost-free region proposals. An RPN is a fully convolutional network that simultaneously predicts object bounds and objectness scores at each position. The RPN is trained end-to-end to generate high-quality region proposals, which are used by Fast R-CNN for detection. We further merge RPN and Fast R-CNN into a single network by sharing their convolutional features—using the recently popular terminology of neural networks with "attention" mechanisms, the RPN component tells the unified network where to look. For the very deep VGG-16 model [3], our detection system has a frame rate of 5fps (*including all steps*) on a GPU, while achieving state-of-the-art object detection accuracy on PASCAL VOC 2007, 2012, and MS COCO datasets with only 300 proposals per image. In ILSVRC and COCO 2015 competitions, Faster R-CNN and RPN are the foundations of the 1st-place winning entries in several tracks. Code has been made publicly available.

Index Terms—Object Detection, Region Proposal, Convolutional Neural Network.

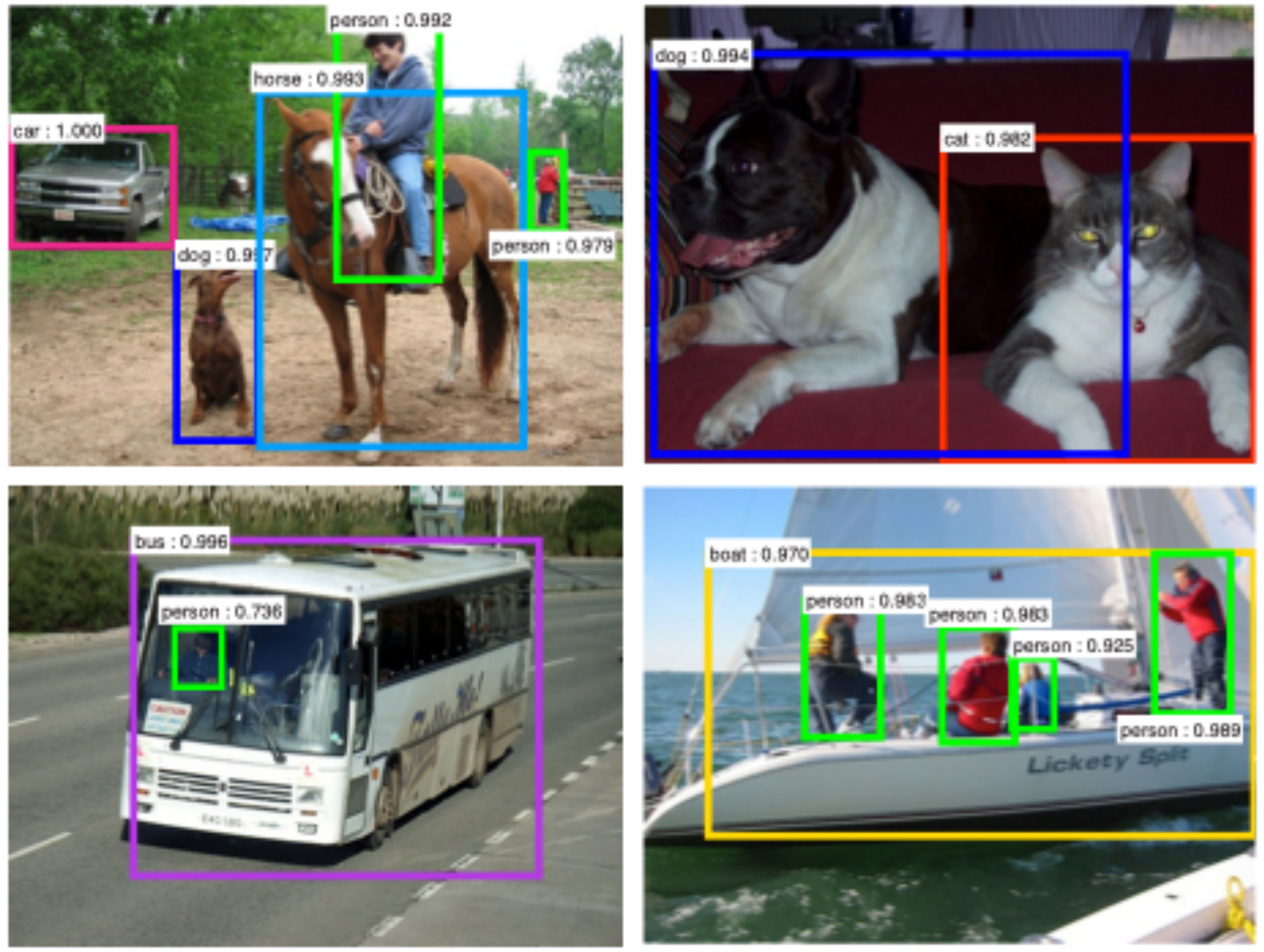
1 Introduction

Recent advances in object detection are driven by the success of region proposal methods (*e.g.*, [4]) and region-based convolutional neural networks (R-CNNs) [5]. Although region-based CNNs were computationally expensive as originally developed in [5], their cost has been drastically reduced thanks to sharing convolutions across proposals [1], [2]. The latest incarnation, Fast R-CNN [2], achieves near real-time rates using very deep networks [3], *when ignoring the time spent on region proposals*. Now, proposals are the test-time computational bottleneck in state-of-the-art detection systems.

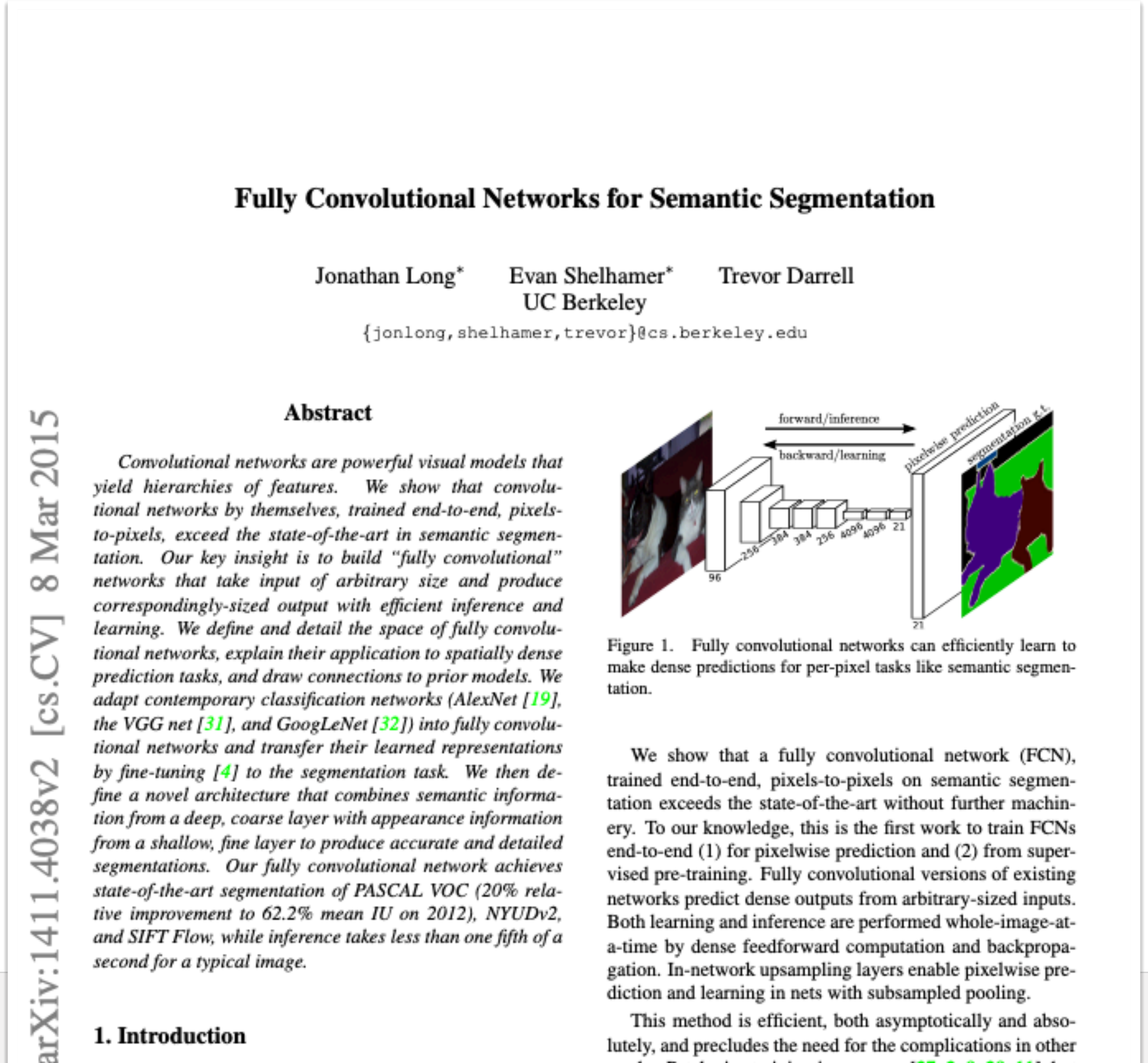
Region proposal methods typically rely on inex-

One may note that fast region-based CNNs take advantage of GPUs, while the region proposal methods used in research are implemented on the CPU, making such runtime comparisons inequitable. An obvious way to accelerate proposal computation is to re-implement it for the GPU. This may be an effective engineering solution, but re-implementation ignores the down-stream detection network and therefore misses important opportunities for sharing computation.

In this paper, we show that an algorithmic change—computing proposals with a deep convolutional neural network—leads to an elegant and effective solution where proposal computation is nearly cost-free given the detection network's computation. To this end, we introduce novel *Region Proposal Networks* (RPNs) that



Applications



arXiv:1411.4038v2 [cs.CV] 8 Mar 2015

Fully Convolutional Networks for Semantic Segmentation

Jonathan Long* Evan Shelhamer* Trevor Darrell
UC Berkeley
{jonlong,shelhamer,trevor}@cs.berkeley.edu

Abstract

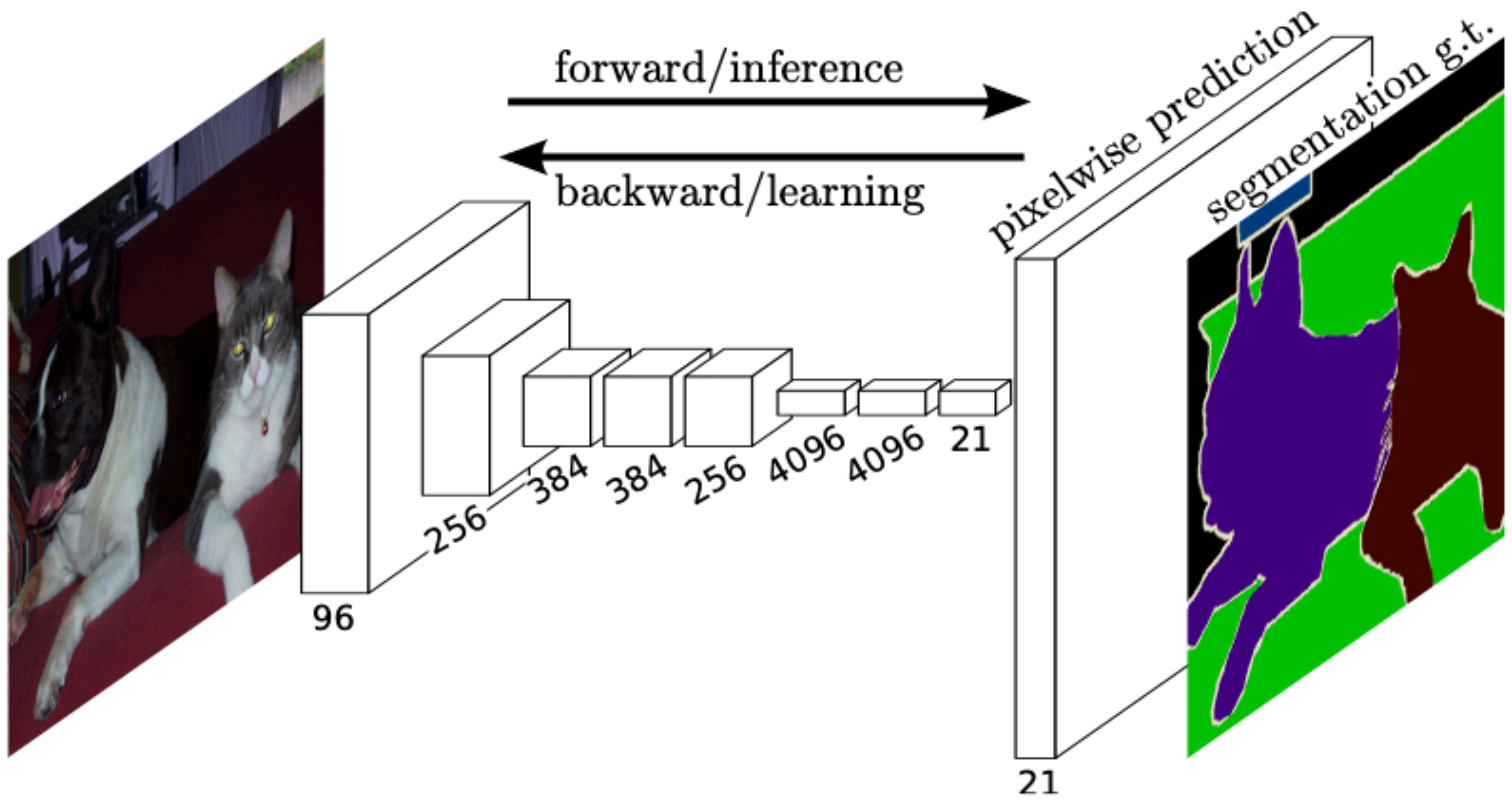
Convolutional networks are powerful visual models that yield hierarchies of features. We show that convolutional networks by themselves, trained end-to-end, pixels-to-pixels, exceed the state-of-the-art in semantic segmentation. Our key insight is to build "fully convolutional" networks that take input of arbitrary size and produce correspondingly-sized output with efficient inference and learning. We define and detail the space of fully convolutional networks, explain their application to spatially dense prediction tasks, and draw connections to prior models. We adapt contemporary classification networks (AlexNet [19], the VGG net [31], and GoogLeNet [32]) into fully convolutional networks and transfer their learned representations by fine-tuning [4] to the segmentation task. We then define a novel architecture that combines semantic information from a deep, coarse layer with appearance information from a shallow, fine layer to produce accurate and detailed segmentations. Our fully convolutional network achieves state-of-the-art segmentation of PASCAL VOC (20% relative improvement to 62.2% mean IU on 2012), NYUDv2, and SIFT Flow, while inference takes less than one fifth of a second for a typical image.

1. Introduction

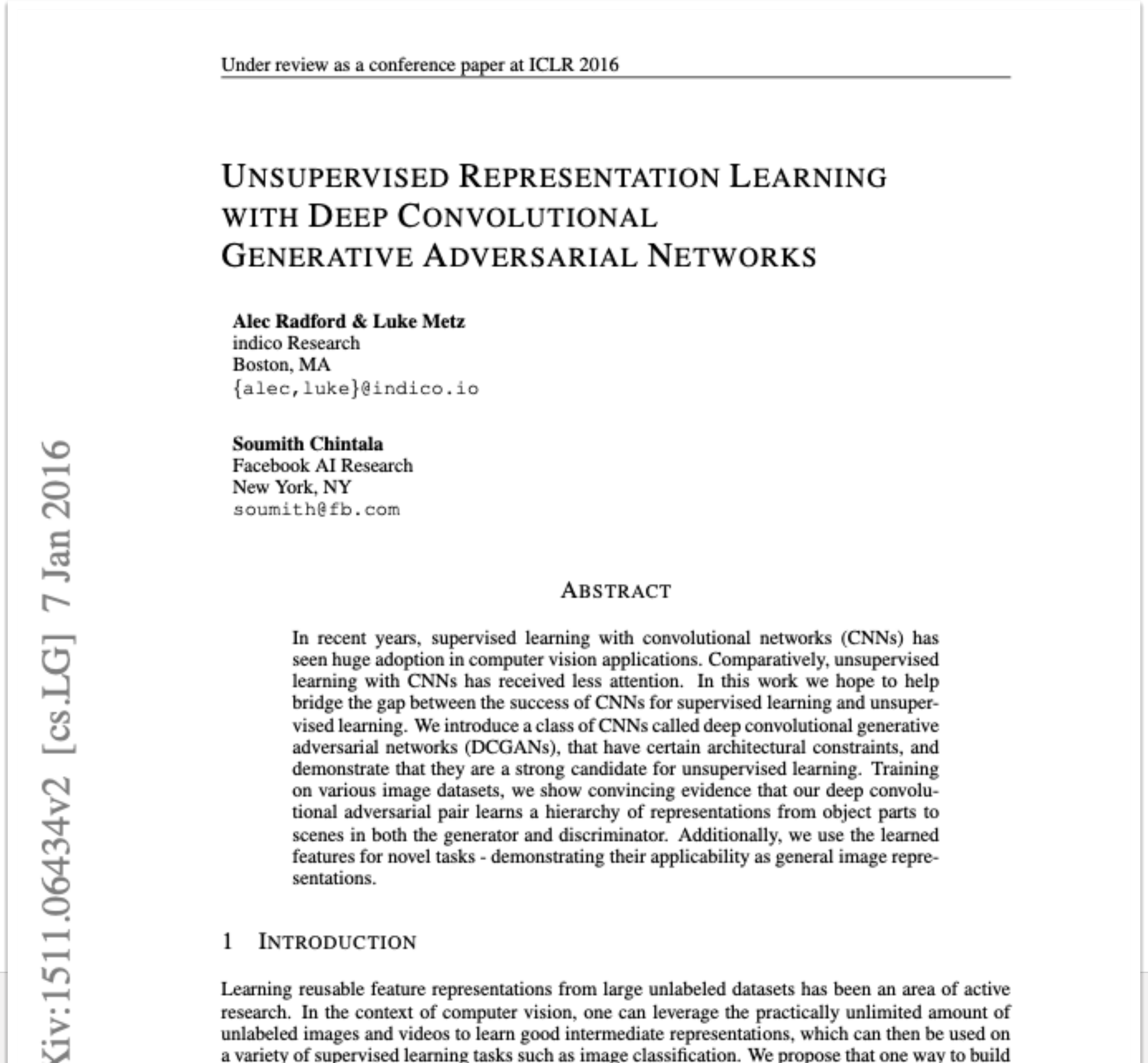
Figure 1. Fully convolutional networks can efficiently learn to make dense predictions for per-pixel tasks like semantic segmentation.

We show that a fully convolutional network (FCN), trained end-to-end, pixels-to-pixels on semantic segmentation exceeds the state-of-the-art without further machinery. To our knowledge, this is the first work to train FCNs end-to-end (1) for pixelwise prediction and (2) from supervised pre-training. Fully convolutional versions of existing networks predict dense outputs from arbitrary-sized inputs. Both learning and inference are performed whole-image-at-a-time by dense feedforward computation and backpropagation. In-network upsampling layers enable pixelwise prediction and learning in nets with subsampled pooling.

This method is efficient, both asymptotically and absolutely, and precludes the need for the complications in other



Applications


Under review as a conference paper at ICLR 2016

Xiv:1511.06434v2 [cs.LG] 7 Jan 2016

UNSUPERVISED REPRESENTATION LEARNING WITH DEEP CONVOLUTIONAL GENERATIVE ADVERSARIAL NETWORKS

Alec Radford & Luke Metz
indico Research
Boston, MA
{alec,luke}@indico.io

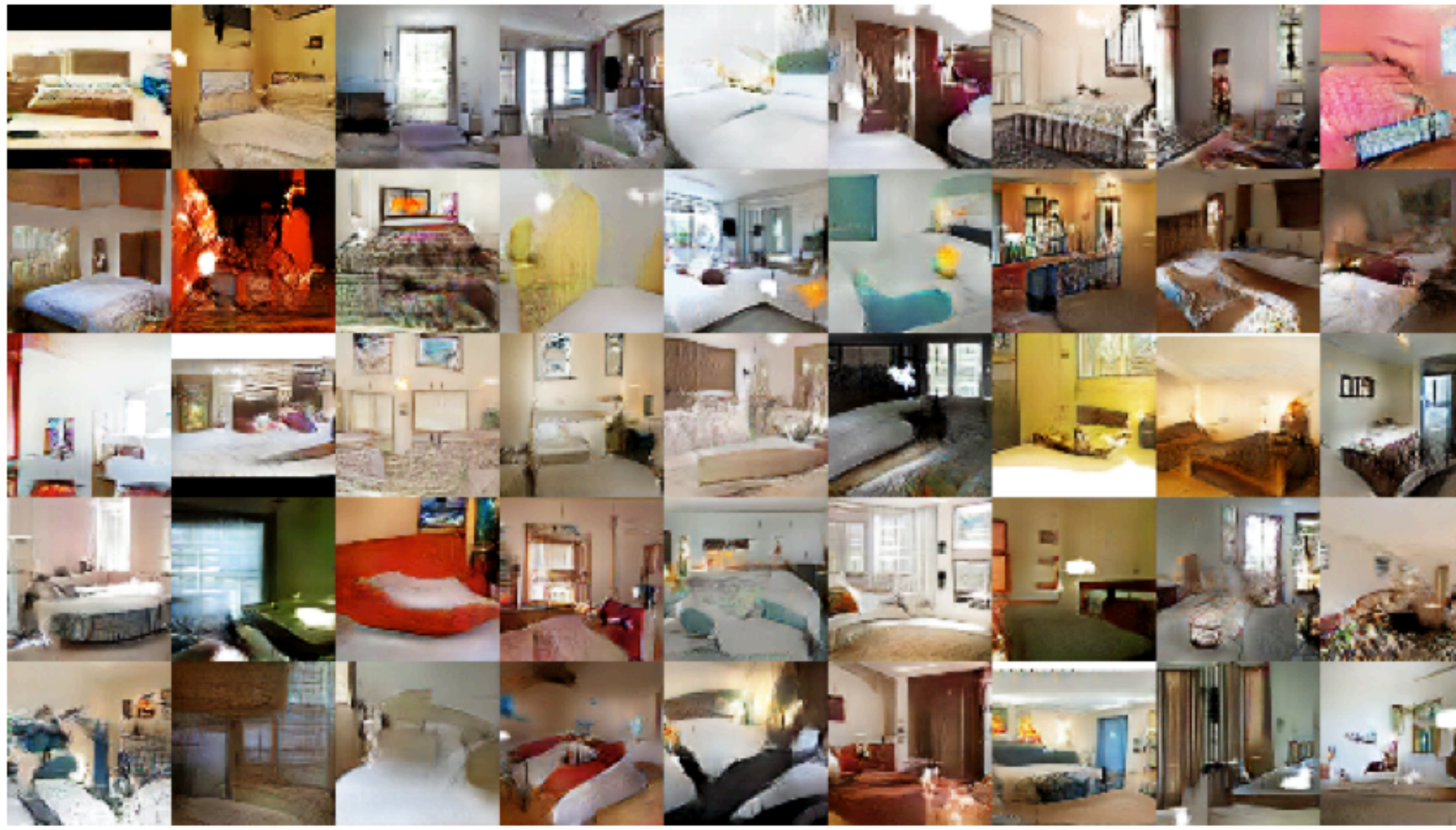
Soumith Chintala
Facebook AI Research
New York, NY
soumith@fb.com

ABSTRACT

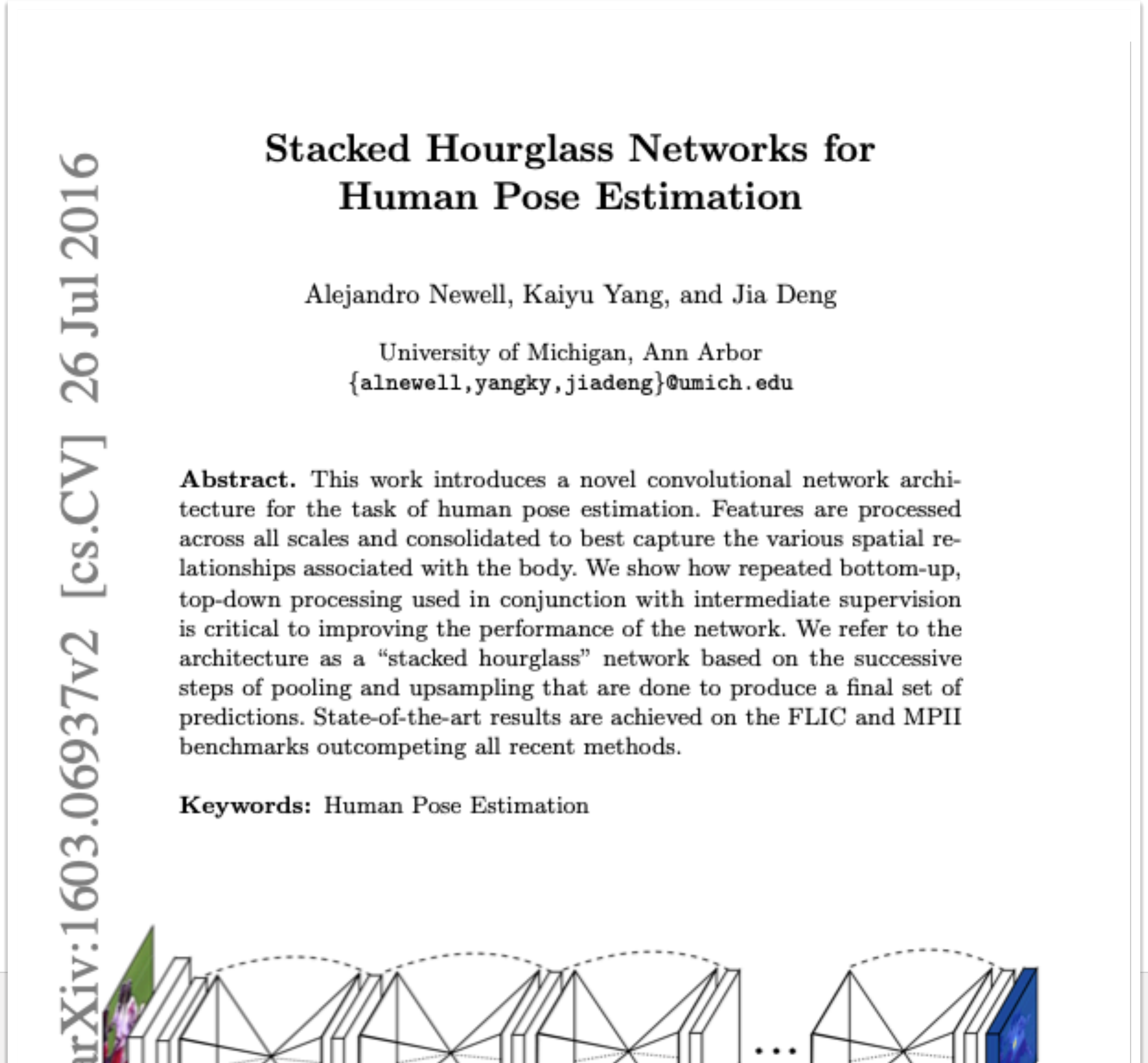
In recent years, supervised learning with convolutional networks (CNNs) has seen huge adoption in computer vision applications. Comparatively, unsupervised learning with CNNs has received less attention. In this work we hope to help bridge the gap between the success of CNNs for supervised learning and unsupervised learning. We introduce a class of CNNs called deep convolutional generative adversarial networks (DCGANs), that have certain architectural constraints, and demonstrate that they are a strong candidate for unsupervised learning. Training on various image datasets, we show convincing evidence that our deep convolutional adversarial pair learns a hierarchy of representations from object parts to scenes in both the generator and discriminator. Additionally, we use the learned features for novel tasks - demonstrating their applicability as general image representations.

1 INTRODUCTION

Learning reusable feature representations from large unlabeled datasets has been an area of active research. In the context of computer vision, one can leverage the practically unlimited amount of unlabeled images and videos to learn good intermediate representations, which can then be used on a variety of supervised learning tasks such as image classification. We propose that one way to build



Applications



rXiv:1603.06937v2 [cs.CV] 26 Jul 2016

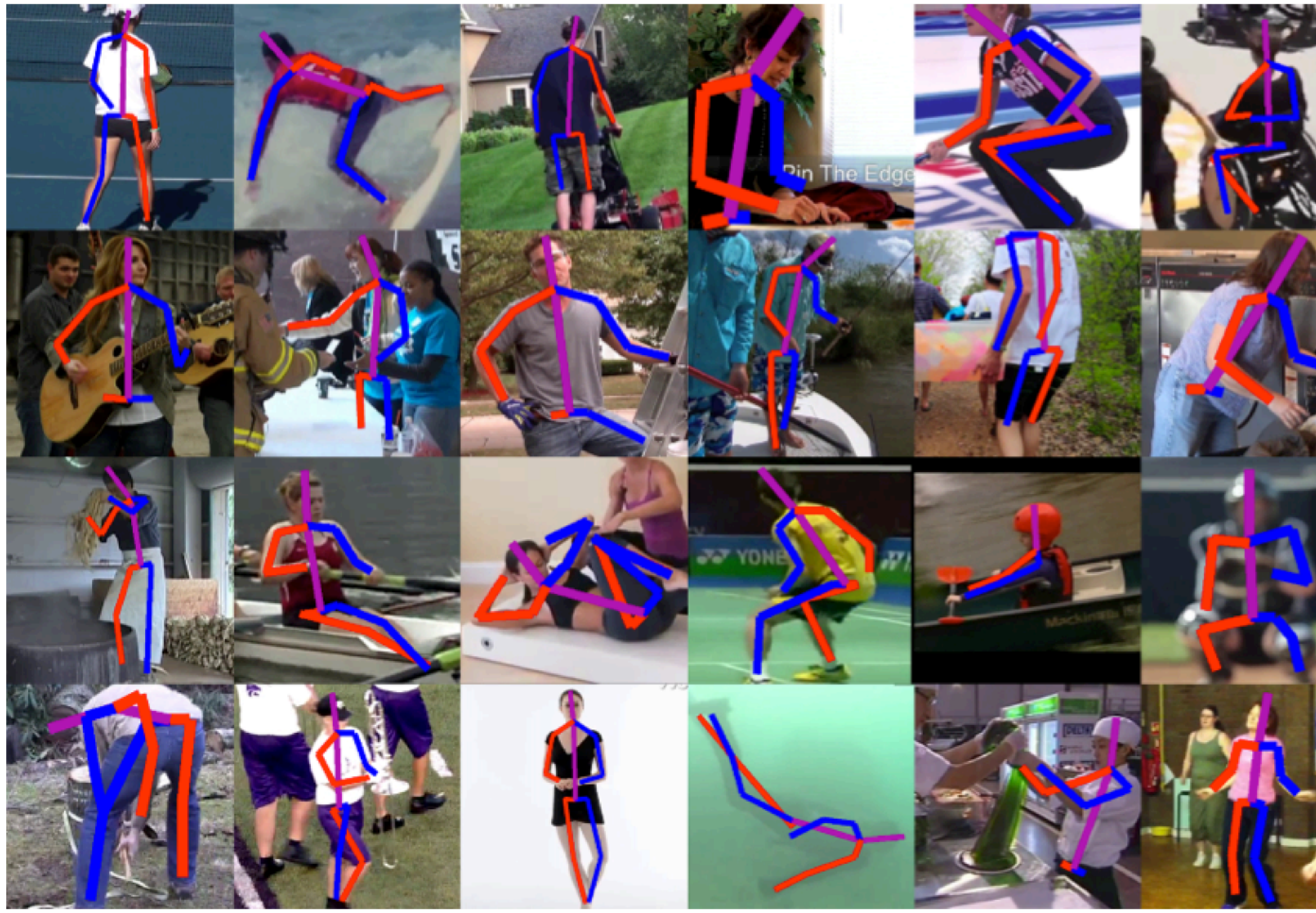
Stacked Hourglass Networks for Human Pose Estimation

Alejandro Newell, Kaiyu Yang, and Jia Deng

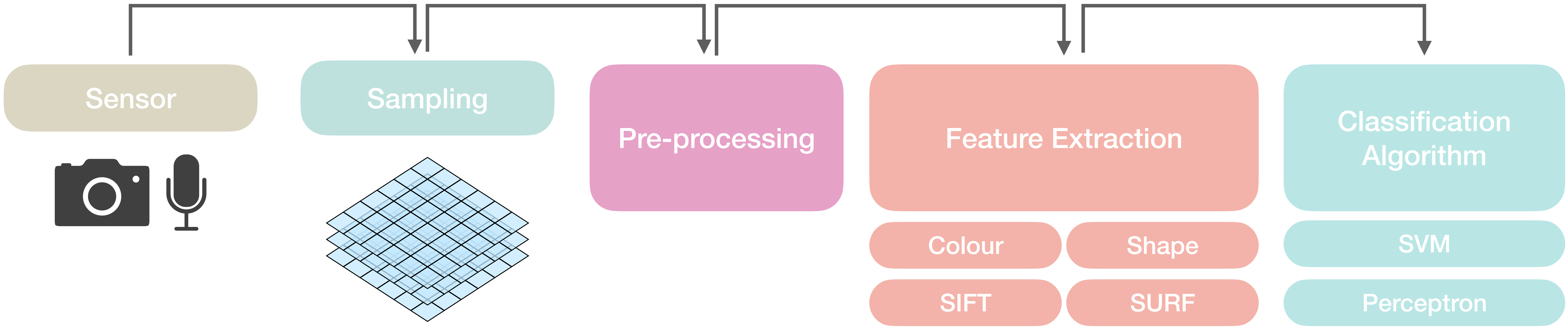
University of Michigan, Ann Arbor
{alnewell,yangky,jiadeng}@umich.edu

Abstract. This work introduces a novel convolutional network architecture for the task of human pose estimation. Features are processed across all scales and consolidated to best capture the various spatial relationships associated with the body. We show how repeated bottom-up, top-down processing used in conjunction with intermediate supervision is critical to improving the performance of the network. We refer to the architecture as a "stacked hourglass" network based on the successive steps of pooling and upsampling that are done to produce a final set of predictions. State-of-the-art results are achieved on the FLIC and MPII benchmarks outcompeting all recent methods.

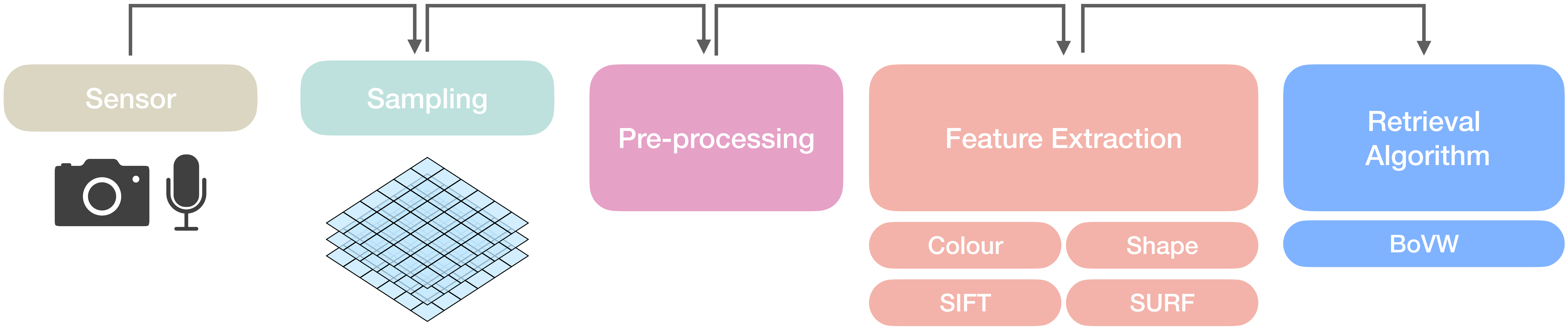
Keywords: Human Pose Estimation



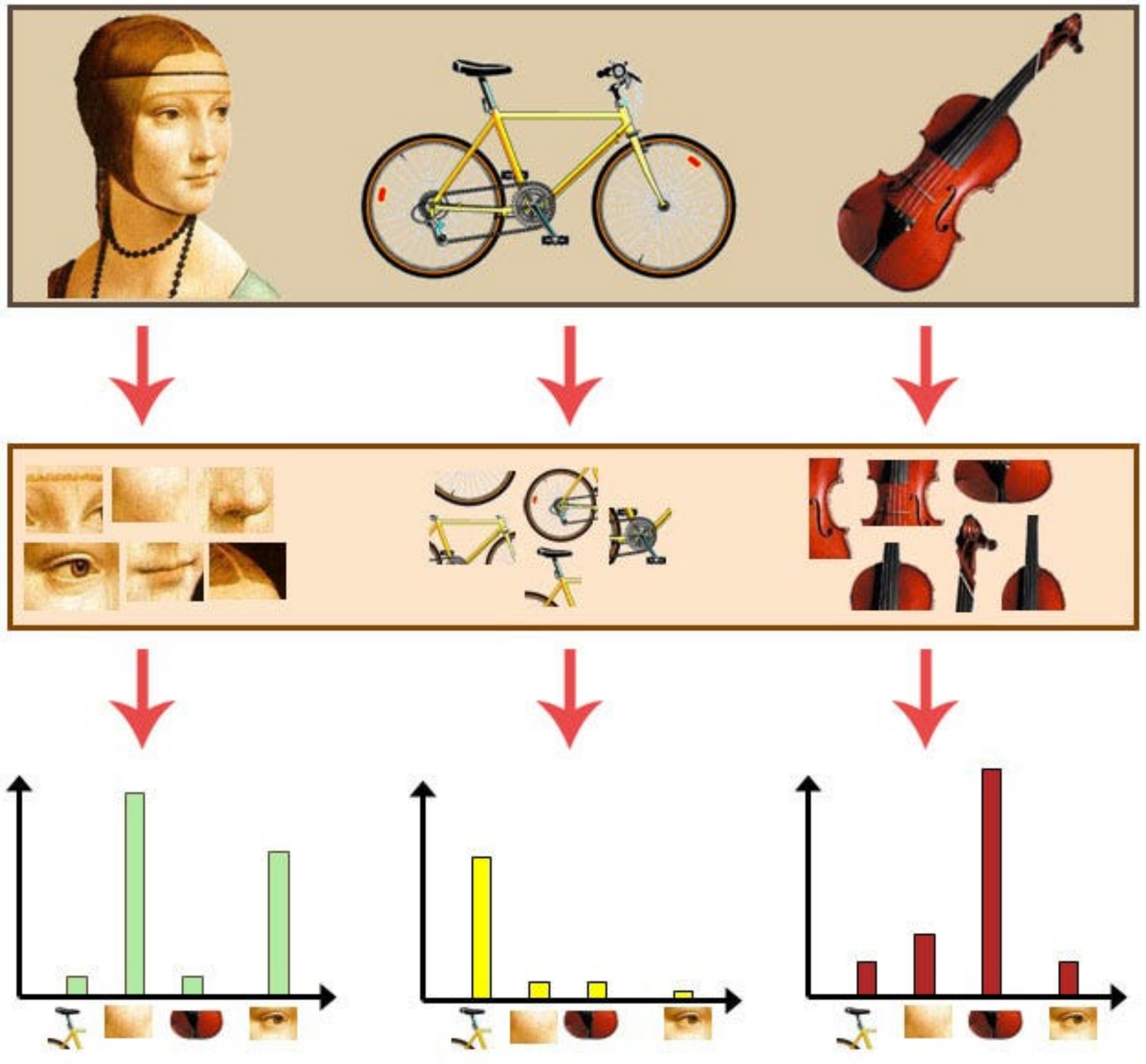
The Common Pipeline



The Common Pipeline



The Common Pipeline



Feature Extraction

- Colour
- Shape
- SIFT
- SURF

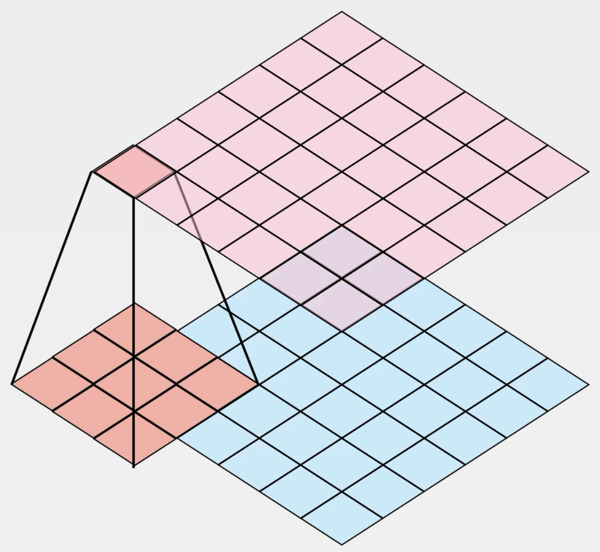
Retrieval Algorithm

- BoVW

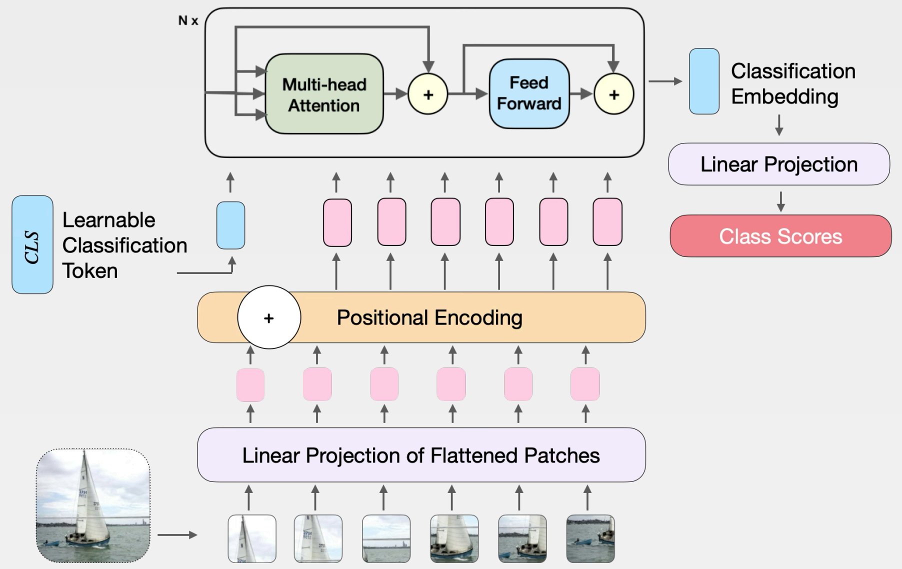
https://medium.com/towards-data-science/bag-of-visual-words-in-a-nutshell-9ceea97ce0fb

SCC0251
Processamento de Imagens

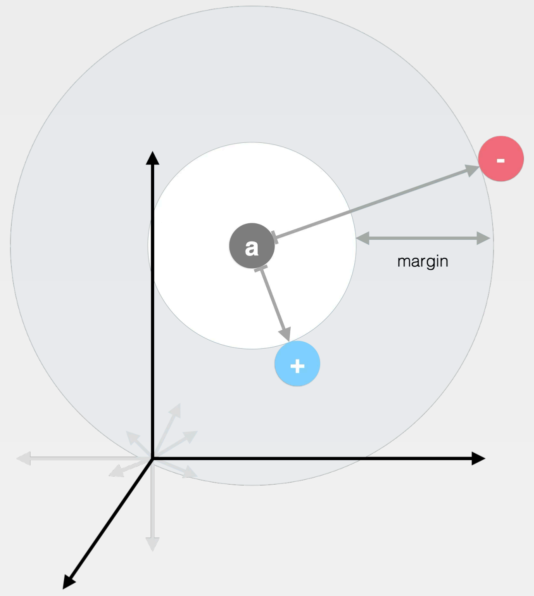
Aprendizado Profundo



CNNs



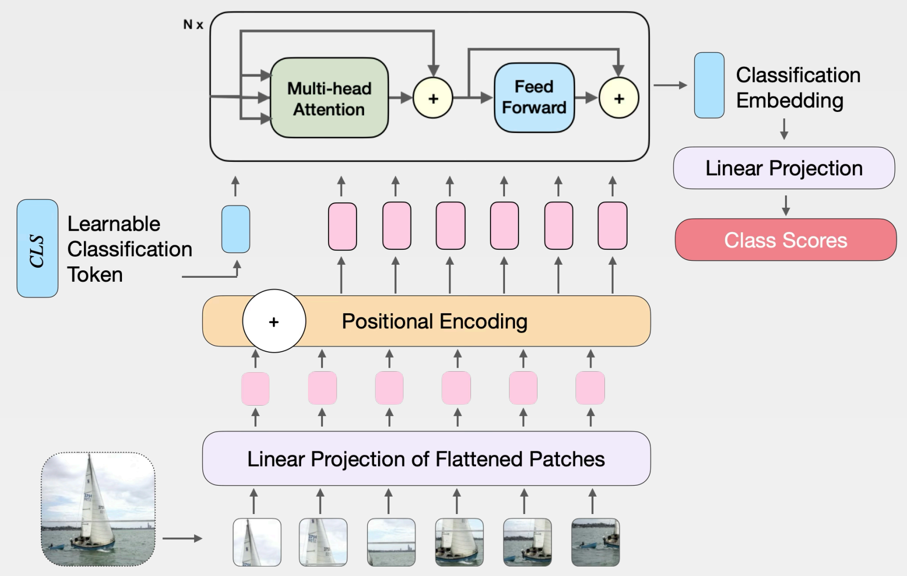
Transformers



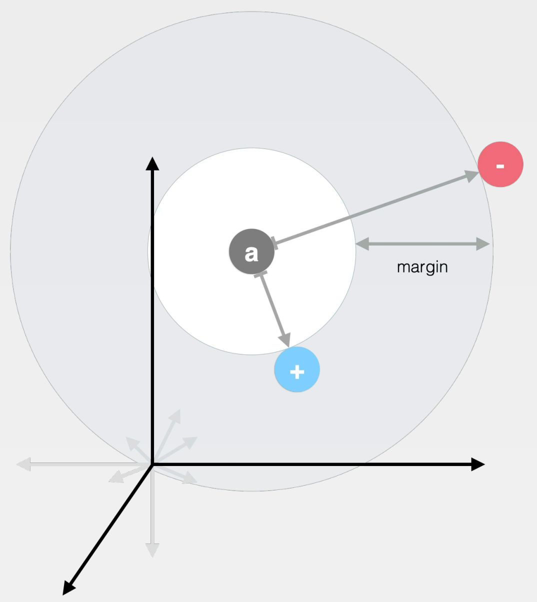
Contrastive Learning

SCC0251
Processamento de Imagens

Aprendizado Profundo



Transformers



Contrastive Learning

attention is all you need

arXiv:1706.03762v7 [cs.CL] 2 Aug 2023

Attention Is All You Need

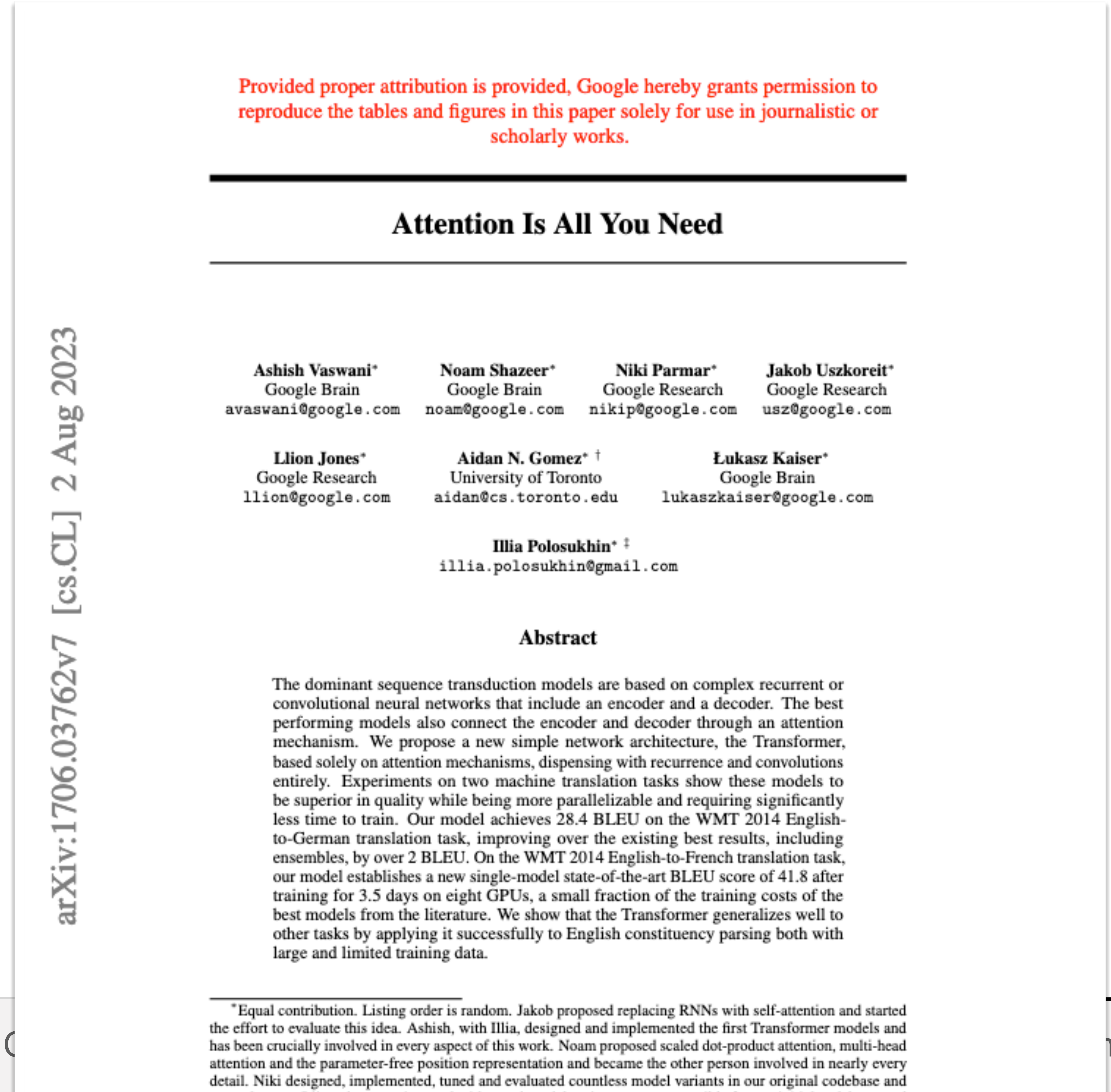
Ashish Vaswani* Google Brain avaswani@google.com
Noam Shazeer* Google Brain noam@google.com
Niki Parmar* Google Research nikip@google.com
Jakob Uszkoreit* Google Research usz@google.com
Llion Jones* Google Research llion@google.com
Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu
Łukasz Kaiser* 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 mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

*Equal contribution. Listing order is random. Jakob proposed replacing RNNs with self-attention and started the effort to evaluate this idea. Ashish, with Illia, designed and implemented the first Transformer models and has been crucially involved in every aspect of this work. Noam proposed scaled dot-product attention, multi-head attention and the parameter-free position representation and became the other person involved in nearly every detail. Niki designed, implemented, tuned and evaluated countless model variants in our original codebase and

attention is all you need



Provided proper attribution is provided, Google hereby grants permission to reproduce the tables and figures in this paper solely for use in journalistic or scholarly works.

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

arXiv:1706.03762v7 [cs.CL] 2 Aug 2023

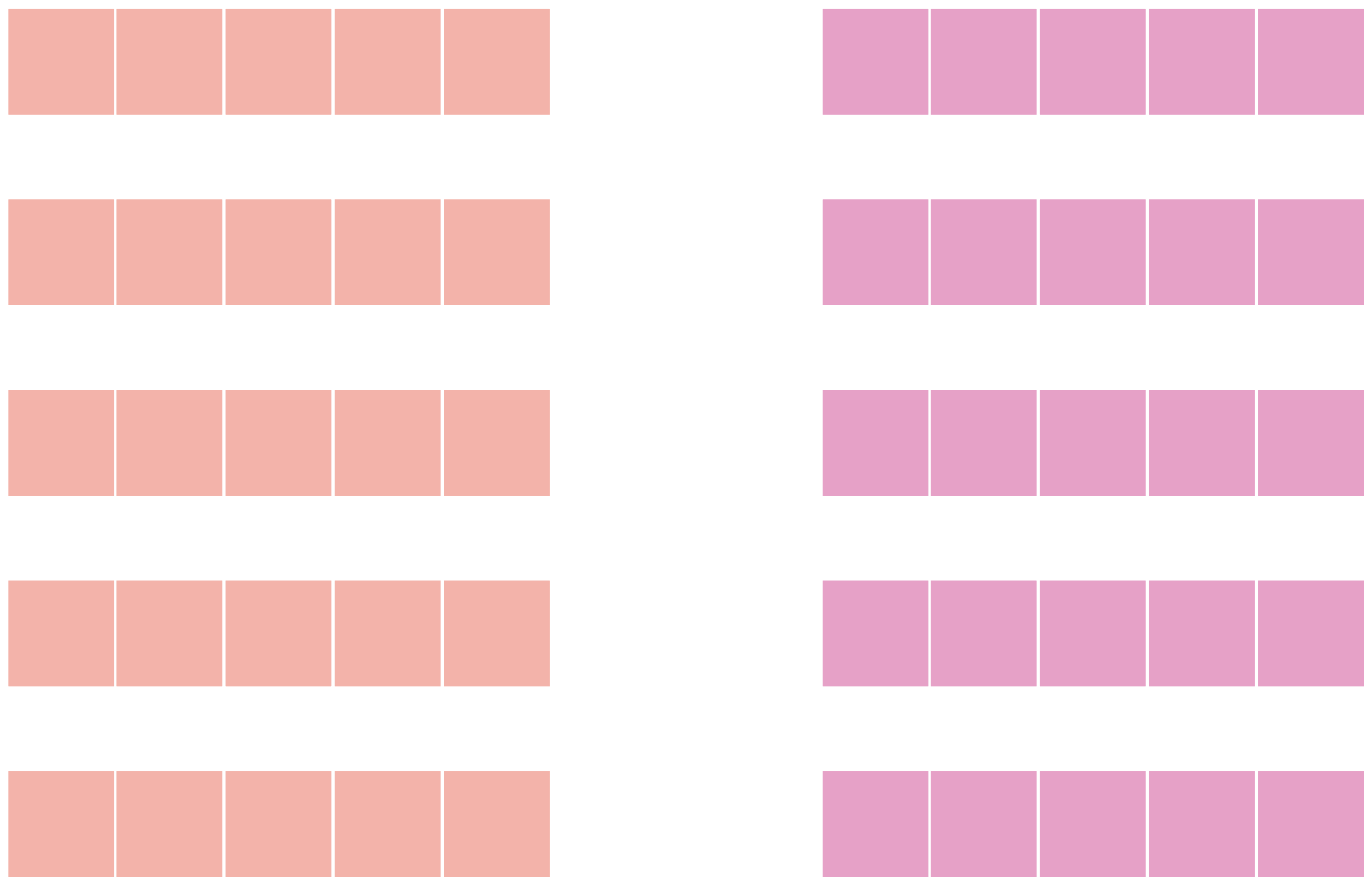
Abstract

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 mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

*Equal contribution. Listing order is random. Jakob proposed replacing RNNs with self-attention and started the effort to evaluate this idea. Ashish, with Illia, designed and implemented the first Transformer models and has been crucially involved in every aspect of this work. Noam proposed scaled dot-product attention, multi-head attention and the parameter-free position representation and became the other person involved in nearly every detail. Niki designed, implemented, tuned and evaluated countless model variants in our original codebase and

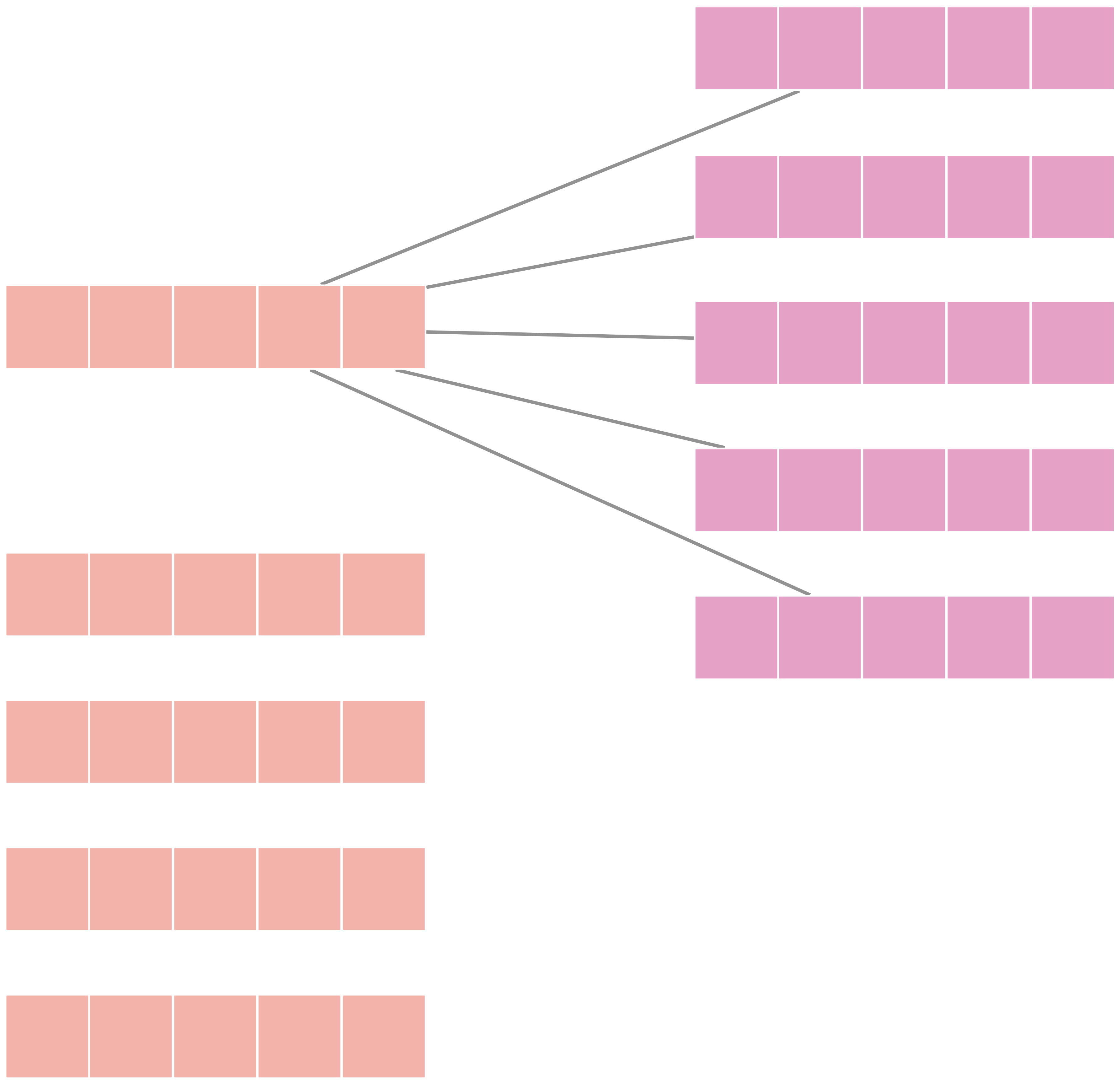
$$Z = \sigma\left(\frac{(XW_q)(XW_k)^T}{\sqrt{d_q}}\right) XW_v$$

attention is all you need



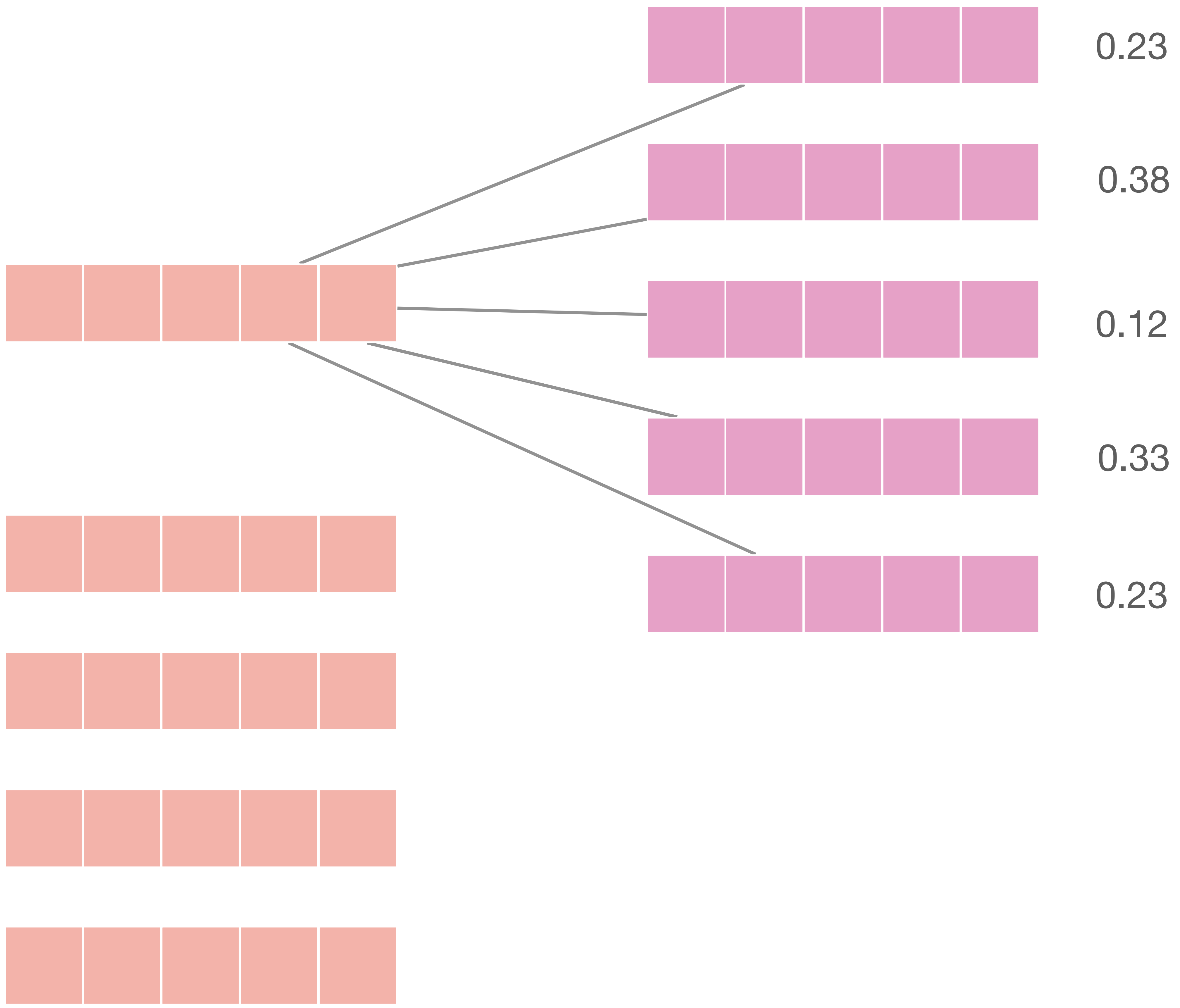
given two vector sequences

attention is all you need



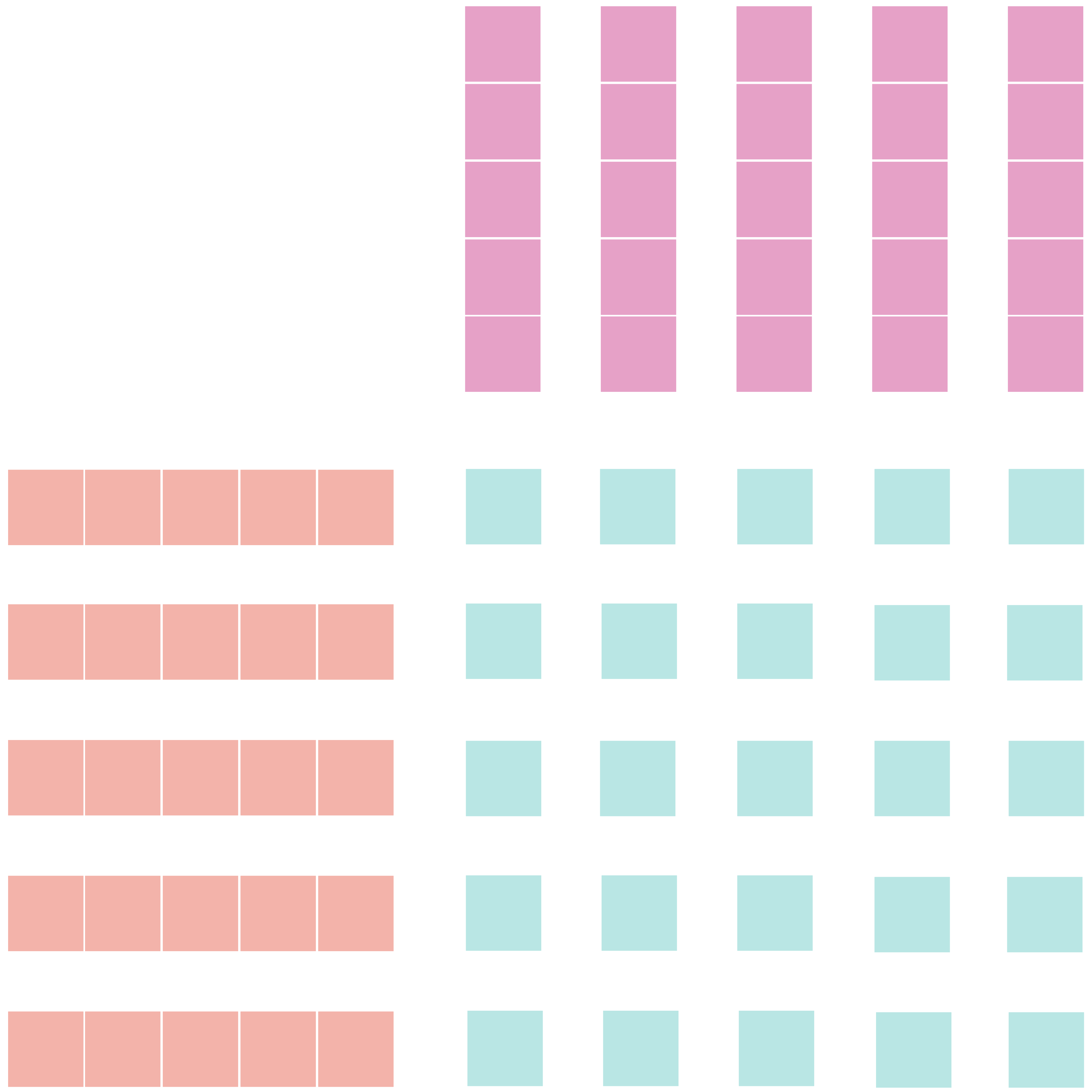
we wish to compare all against all

attention is all you need



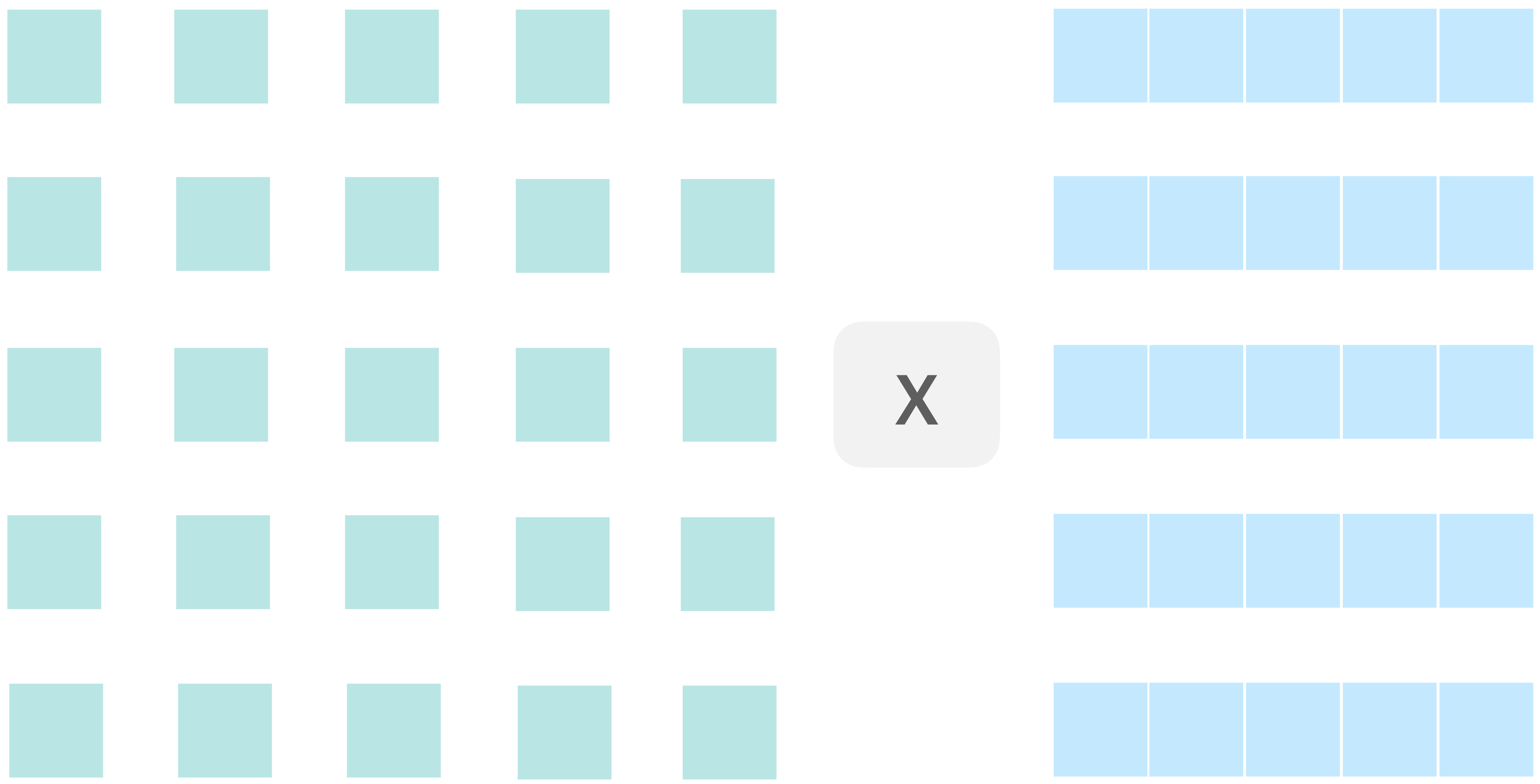
each comparison will result in a scalar, the *attention weight*

attention is all you need



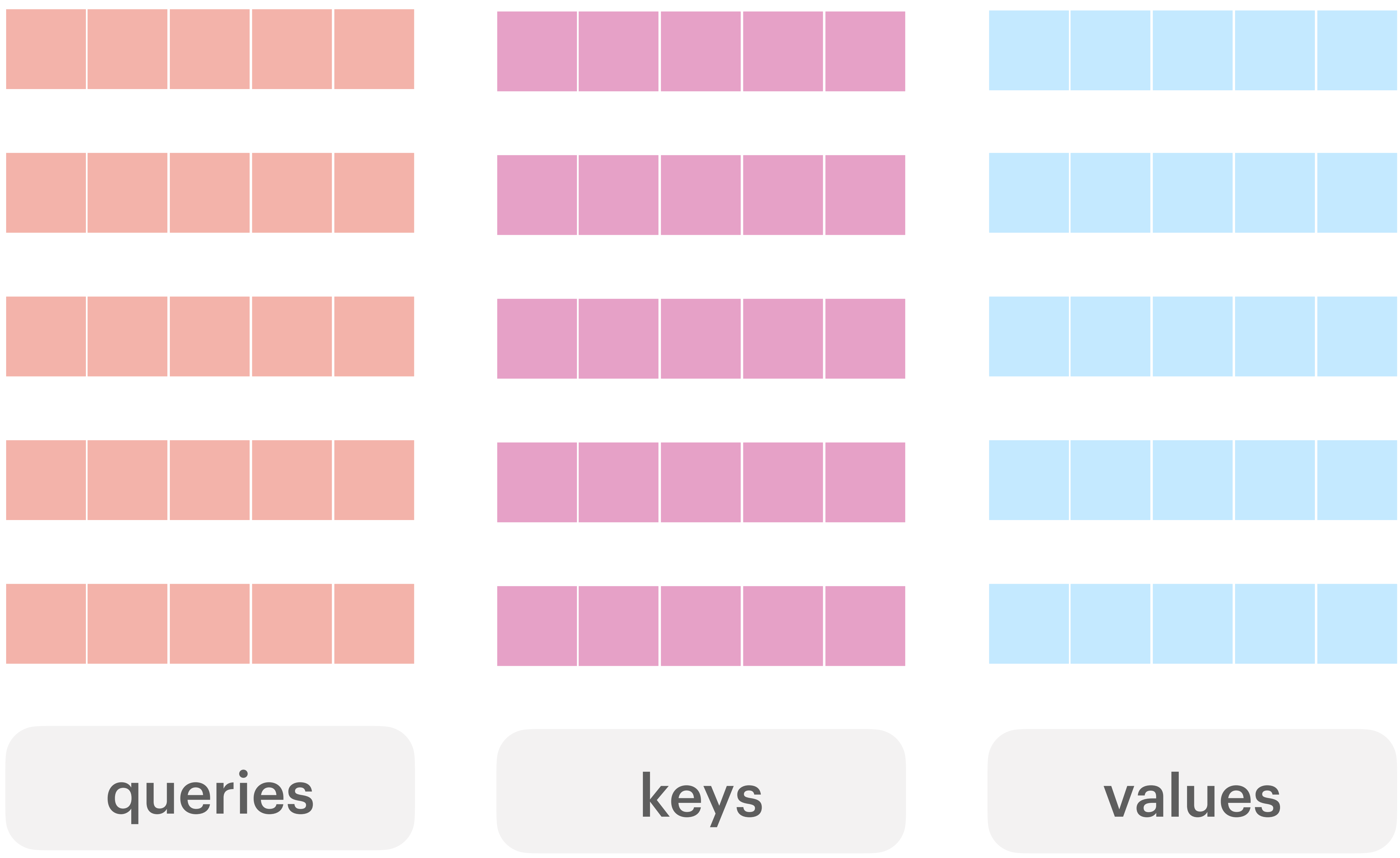
together these compose the *attention matrix*

attention is all you need



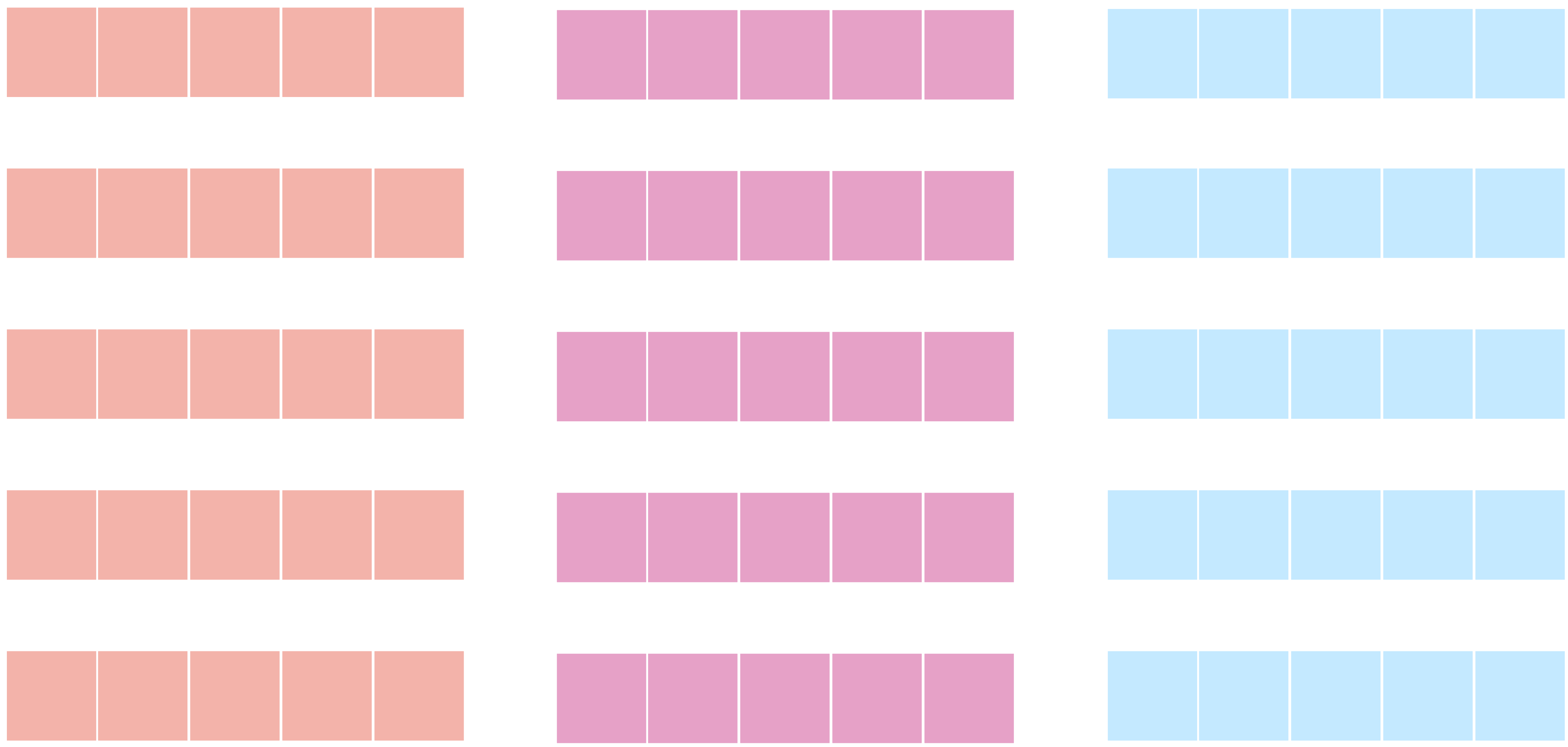
finally, the values on the attention matriz are used to weight a third sequence of vectors

attention is all you need



each of these sequences is called queries, keys and values

attention is all you need

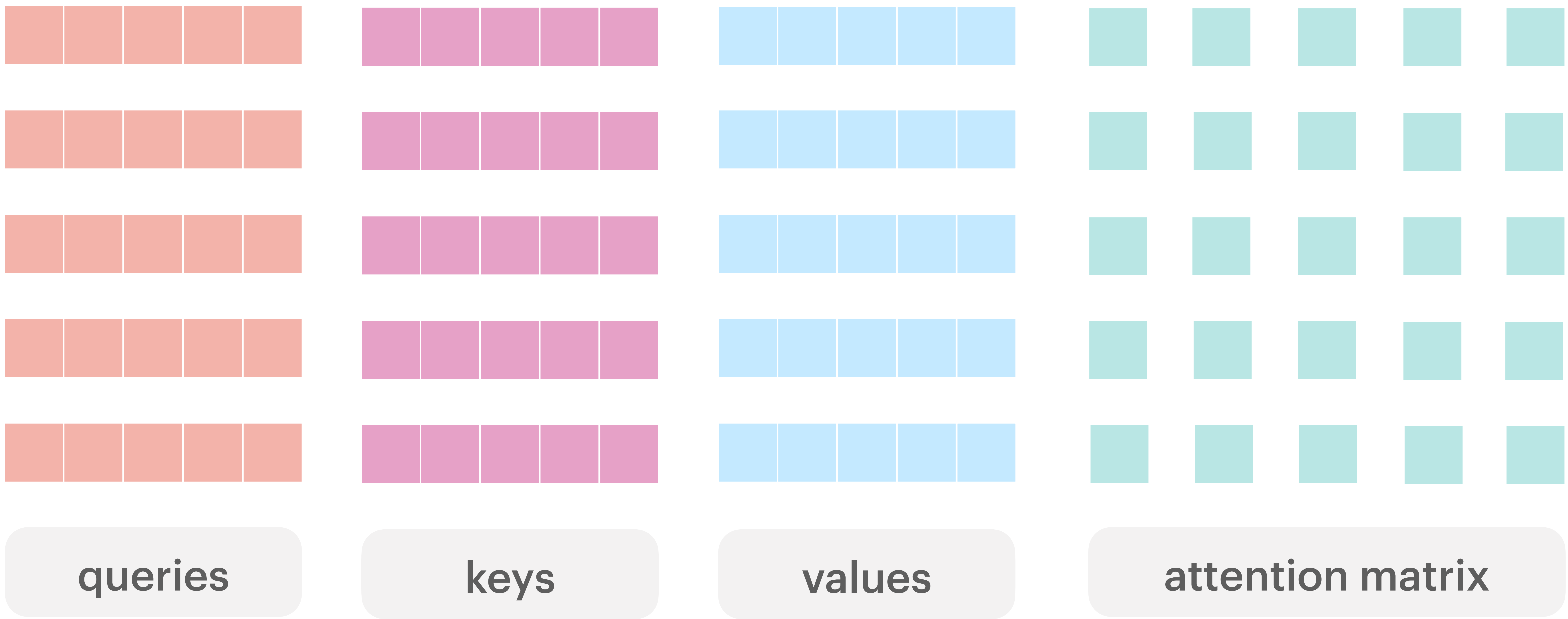


queries keys values

XW_q XW_k XW_v

$$Z = \sigma\left(\frac{(XW_q)(XW_k)^T}{\sqrt{d_q}}\right)XW_v$$

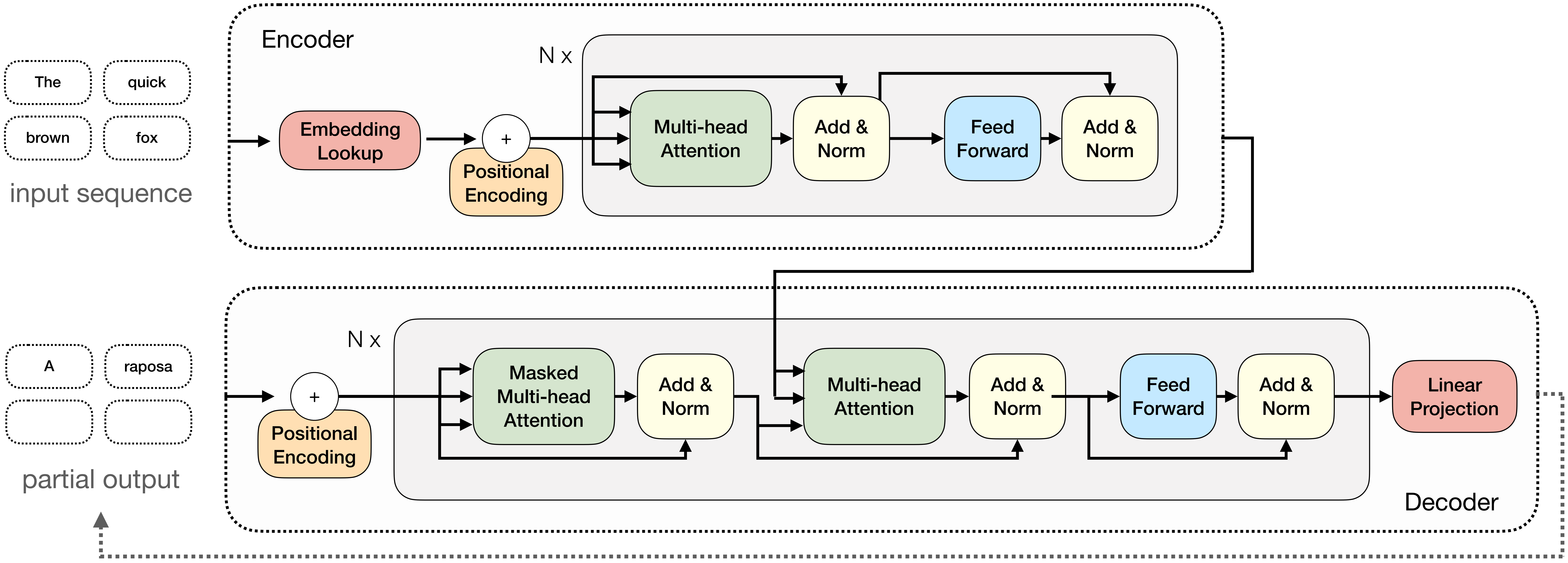
attention is all you need



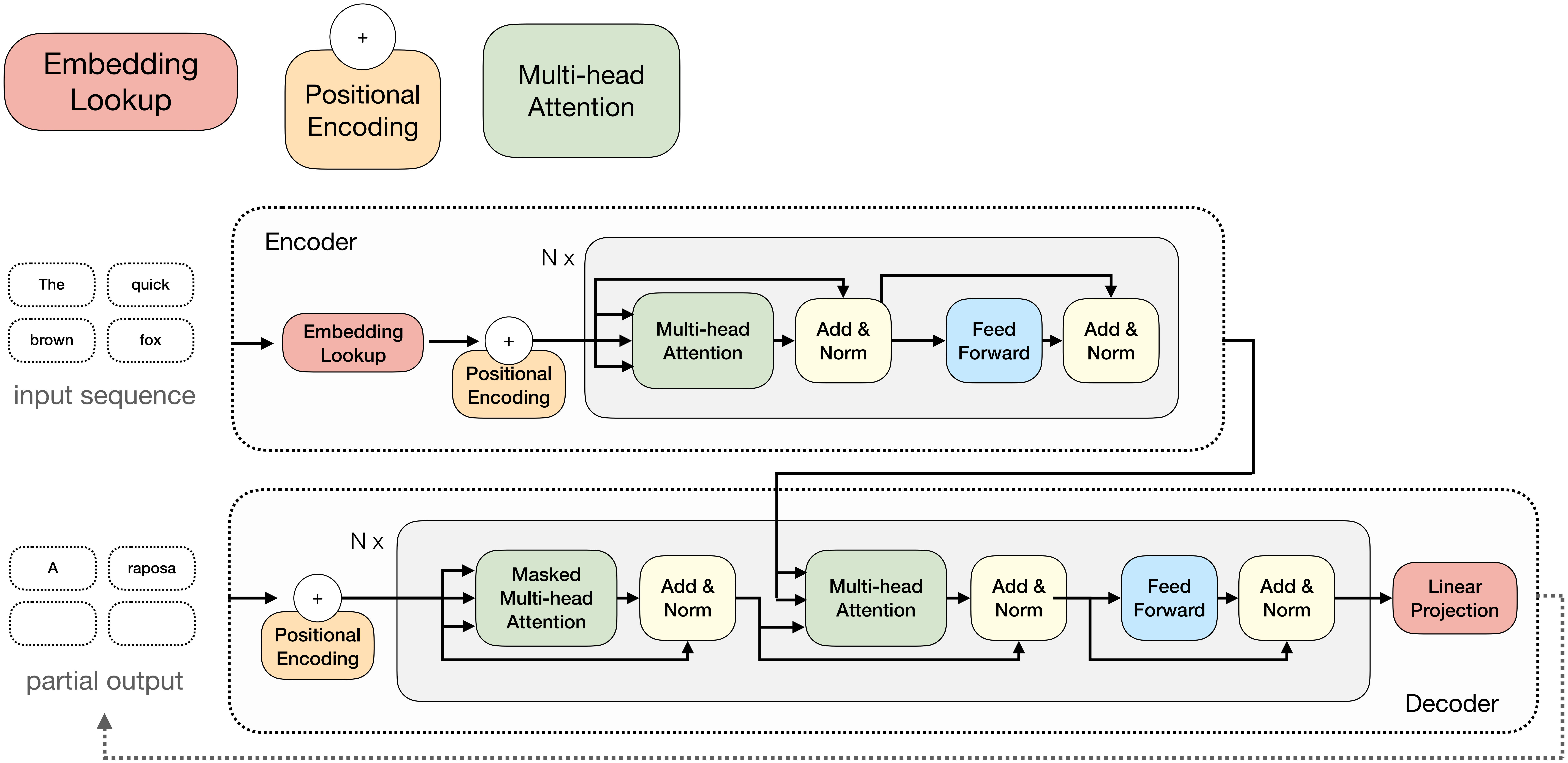
XW_q XW_k XW_v $\sigma\left(\frac{(XW_q)(XW_k)^T}{\sqrt{d_q}}\right)$

$$Z = \sigma\left(\frac{(XW_q)(XW_k)^T}{\sqrt{d_q}}\right) XW_v$$

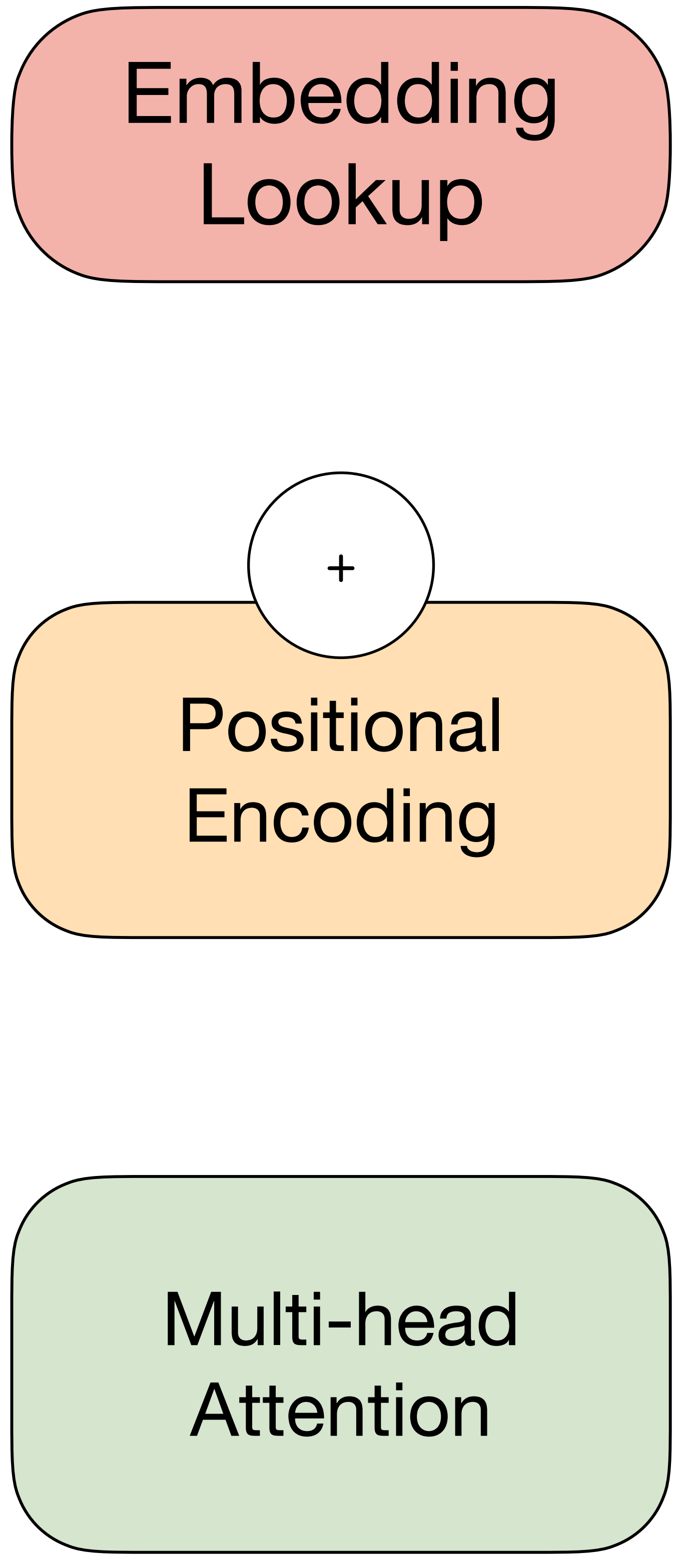
attention is all you need



attention is all you need



attention is all you need



Matrix that converts words (the concept) into unique vectors

attention is all you need

Embedding Lookup

Matrix that converts words (the concept) into unique vectors

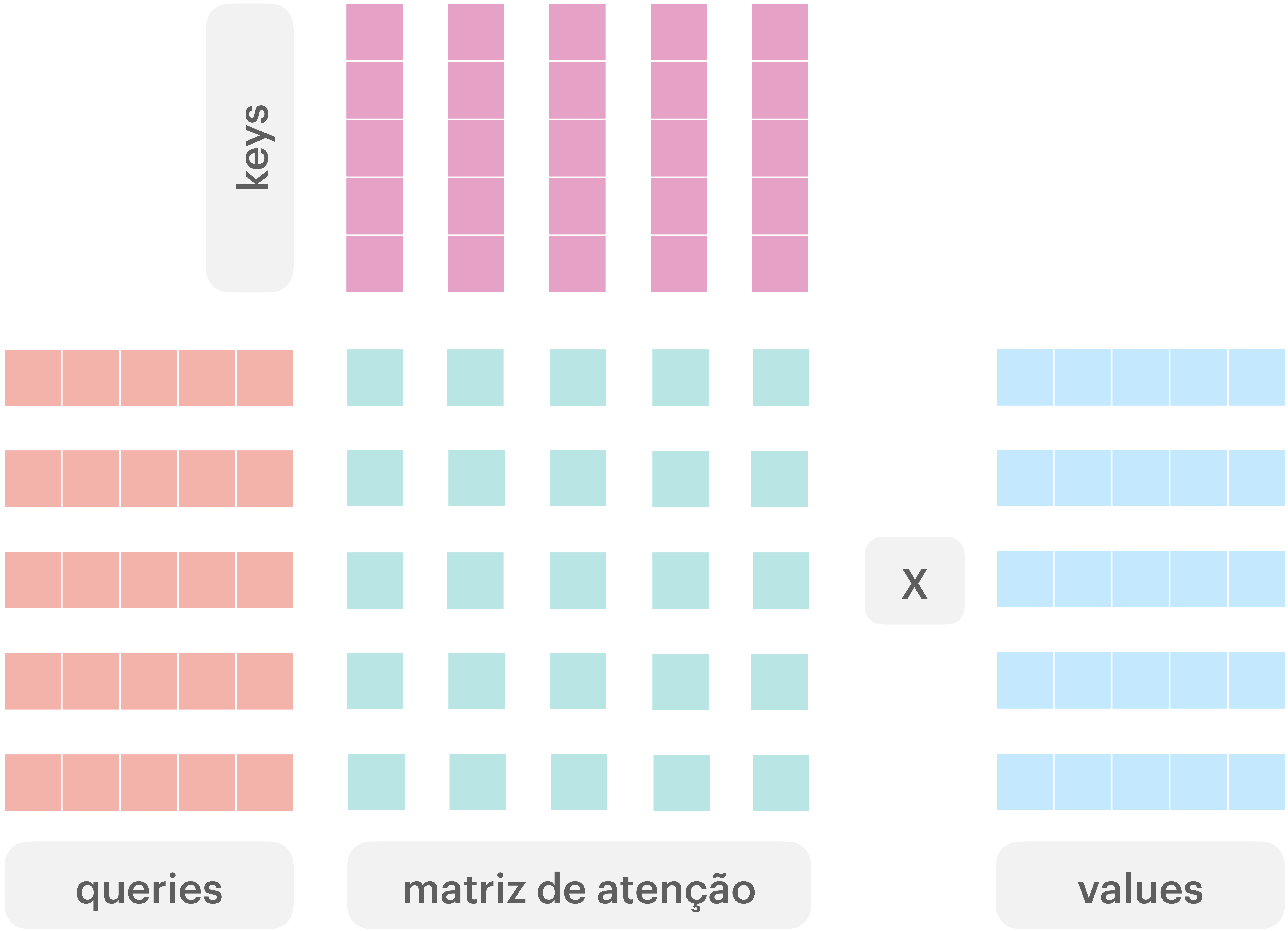
+

Positional Encoding

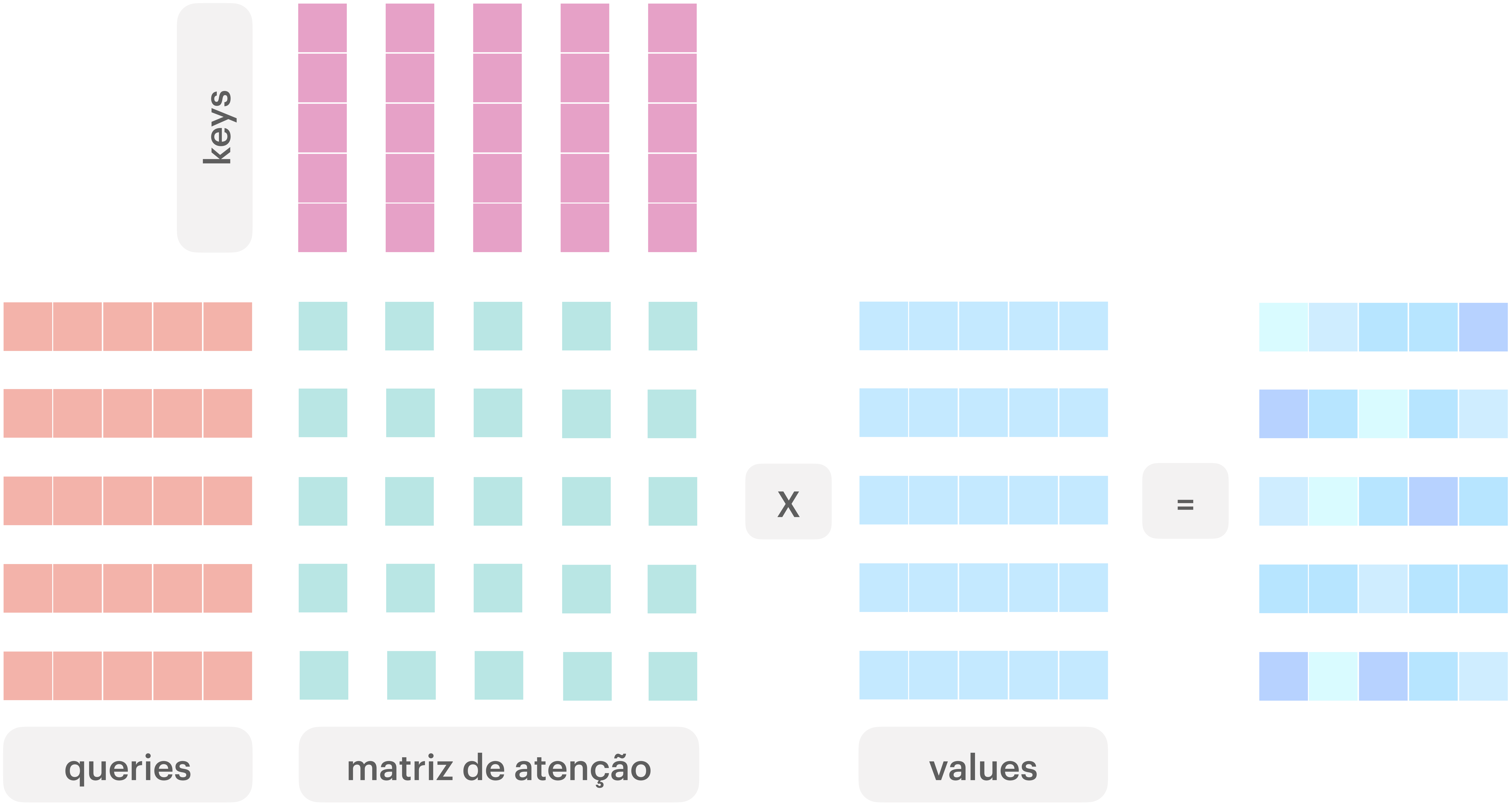
Attention is a position equivariant function

Multi-head Attention

attention is all you need



attention is all you need



attention is all you need

Embedding Lookup

Transformação de palavras para vetores

+

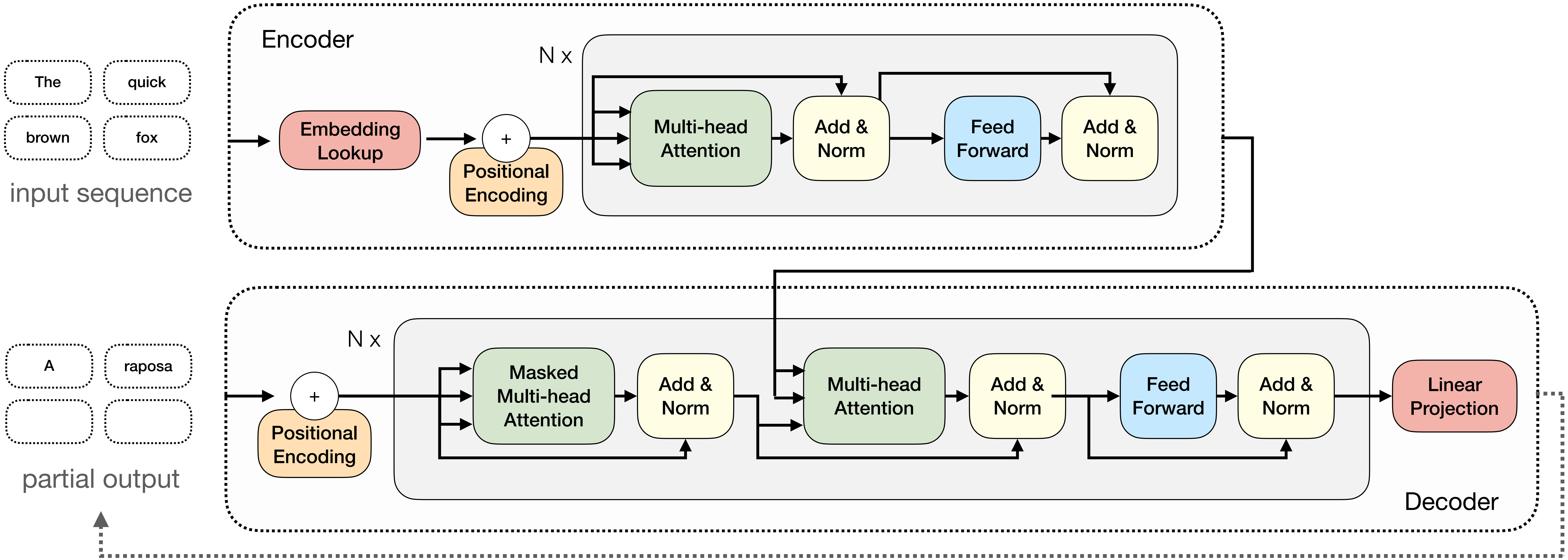
Positional Encoding

Atenção é uma função equivariante a permutação na sequência

Multi-head Attention

Diversification of the outputs. “Can pay attention to multiple things”

attention is all you need



vision transformer

Published as a conference paper at ICLR 2021

arXiv:2010.11929v2 [cs.CV] 3 Jun 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 results compared to state-of-the-art convolutional networks while requiring substantially fewer computational resources to train.[1]

1 INTRODUCTION

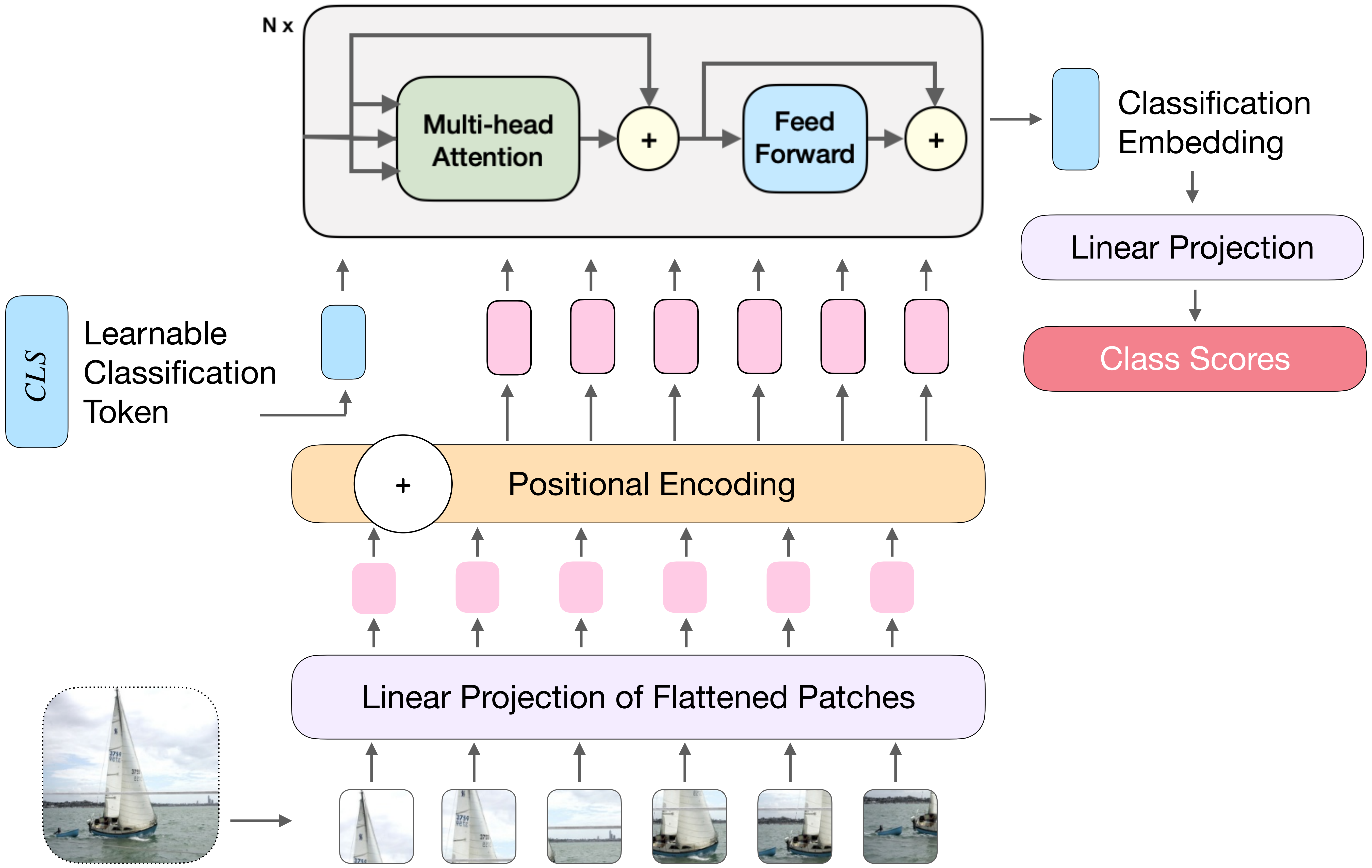
Self-attention-based architectures, in particular Transformers (Vaswani et al., 2017), have become the model of choice in natural language processing (NLP). The dominant approach is to pre-train on a large text corpus and then fine-tune on a smaller task-specific dataset (Devlin et al., 2019). Thanks to Transformers' computational efficiency and scalability, it has become possible to train models of unprecedented size, with over 100B parameters (Brown et al., 2020; Lepikhin et al., 2020). With the models and datasets growing, there is still no sign of saturating performance.

In computer vision, however, convolutional architectures remain dominant (LeCun et al., 1989; Krizhevsky et al., 2012; He et al., 2016). Inspired by NLP successes, multiple works try combining CNN-like architectures with self-attention (Wang et al., 2018; Carion et al., 2020), some replacing the convolutions entirely (Ramachandran et al., 2019; Wang et al., 2020a). The latter models, while theoretically efficient, have not yet been scaled effectively on modern hardware accelerators due to the use of specialized attention patterns. Therefore, in large-scale image recognition, classic ResNet-like architectures are still state of the art (Mahajan et al., 2018; Xie et al., 2020; Kolesnikov et al., 2020).

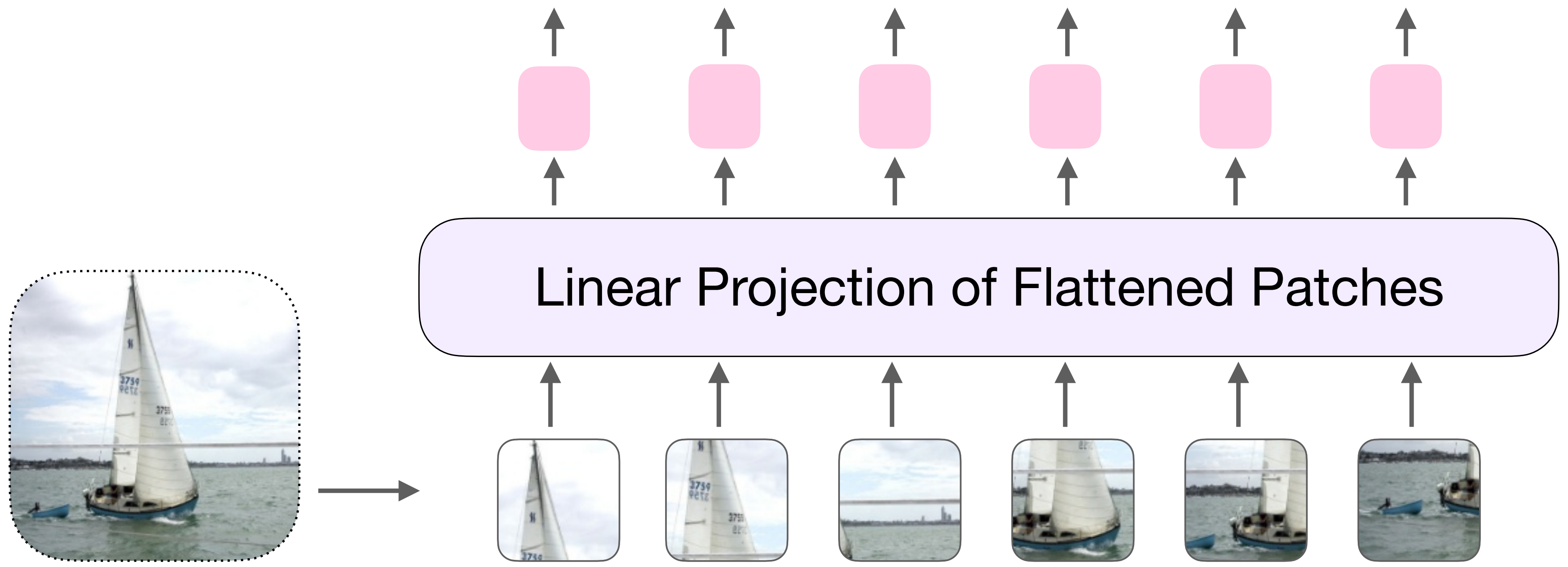
Inspired by the Transformer scaling successes in NLP, we experiment with applying a standard Transformer directly to images, with the fewest possible modifications. To do so, we split an image into patches and provide the sequence of linear embeddings of these patches as an input to a Transformer. Image patches are treated the same way as tokens (words) in an NLP application. We train the model on image classification in supervised fashion.

When trained on mid-sized datasets such as ImageNet without strong regularization, these models yield modest accuracies of a few percentage points below ResNets of comparable size. This seemingly discouraging outcome may be expected: Transformers lack some of the inductive biases

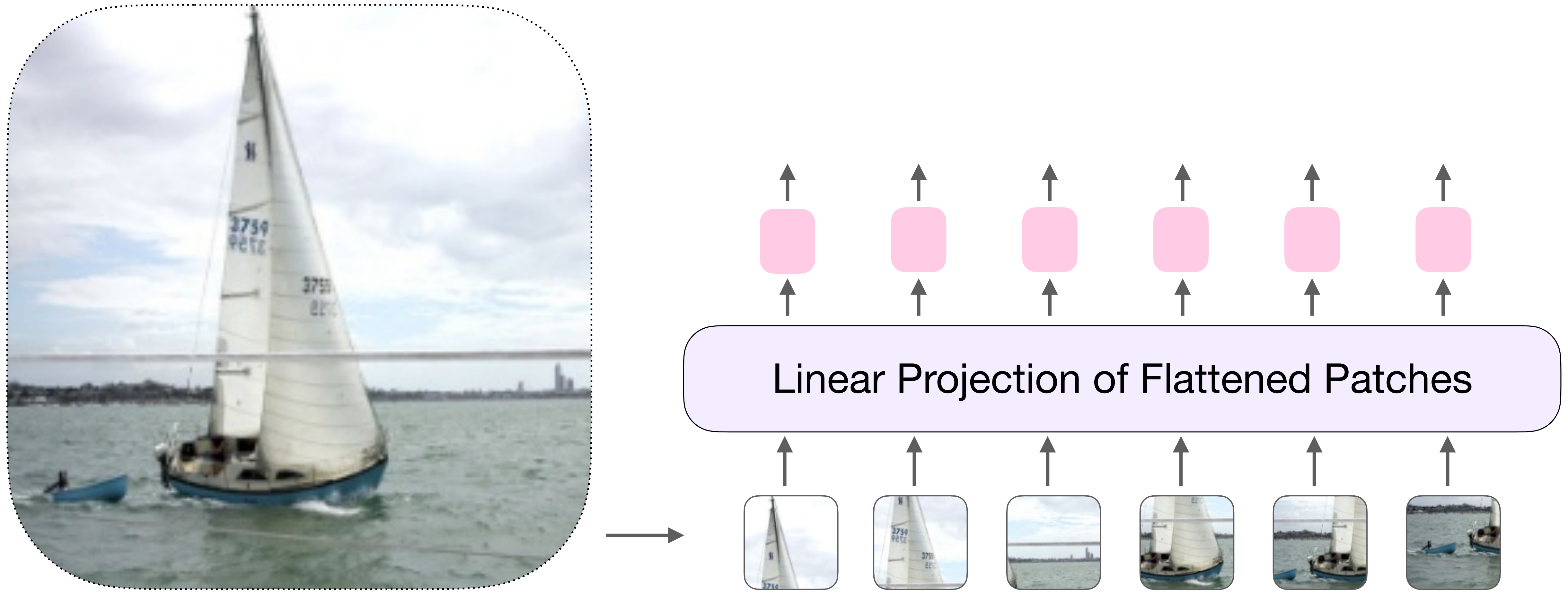
[1]Fine-tuning code and pre-trained models are available at https://github.com/google-research/vision_transformer



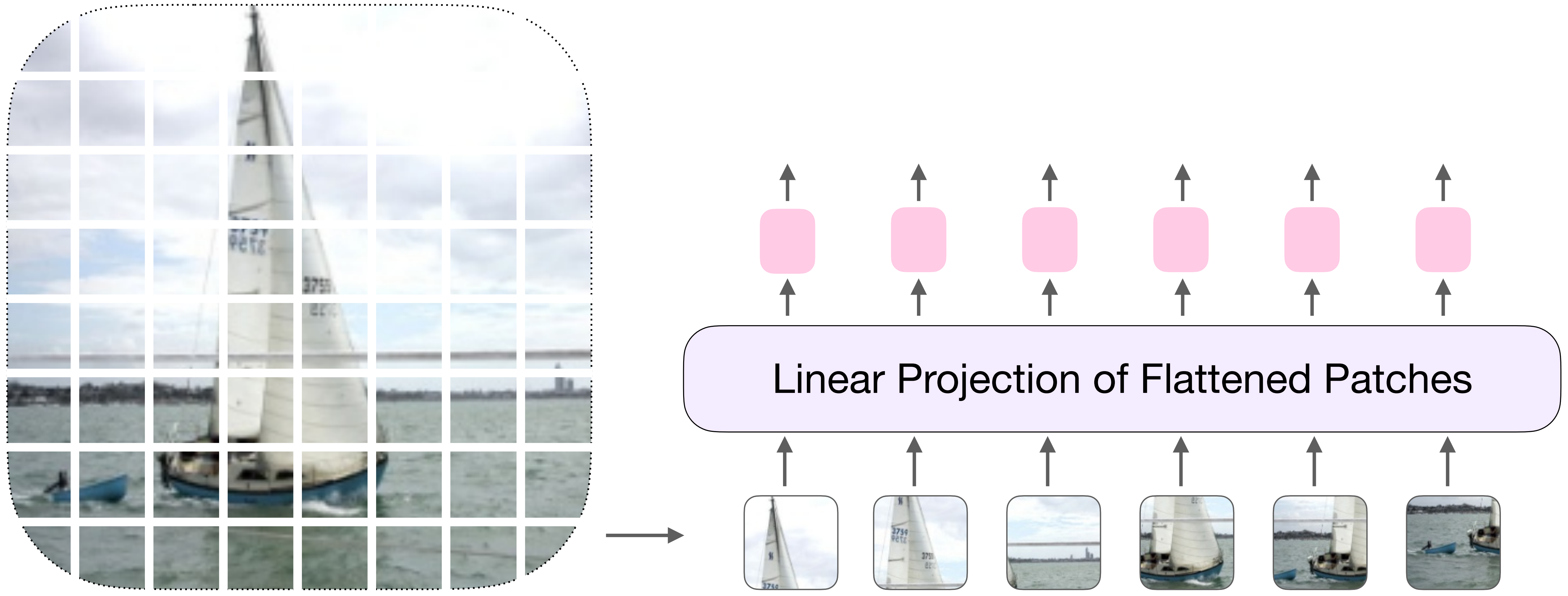
vision transformer



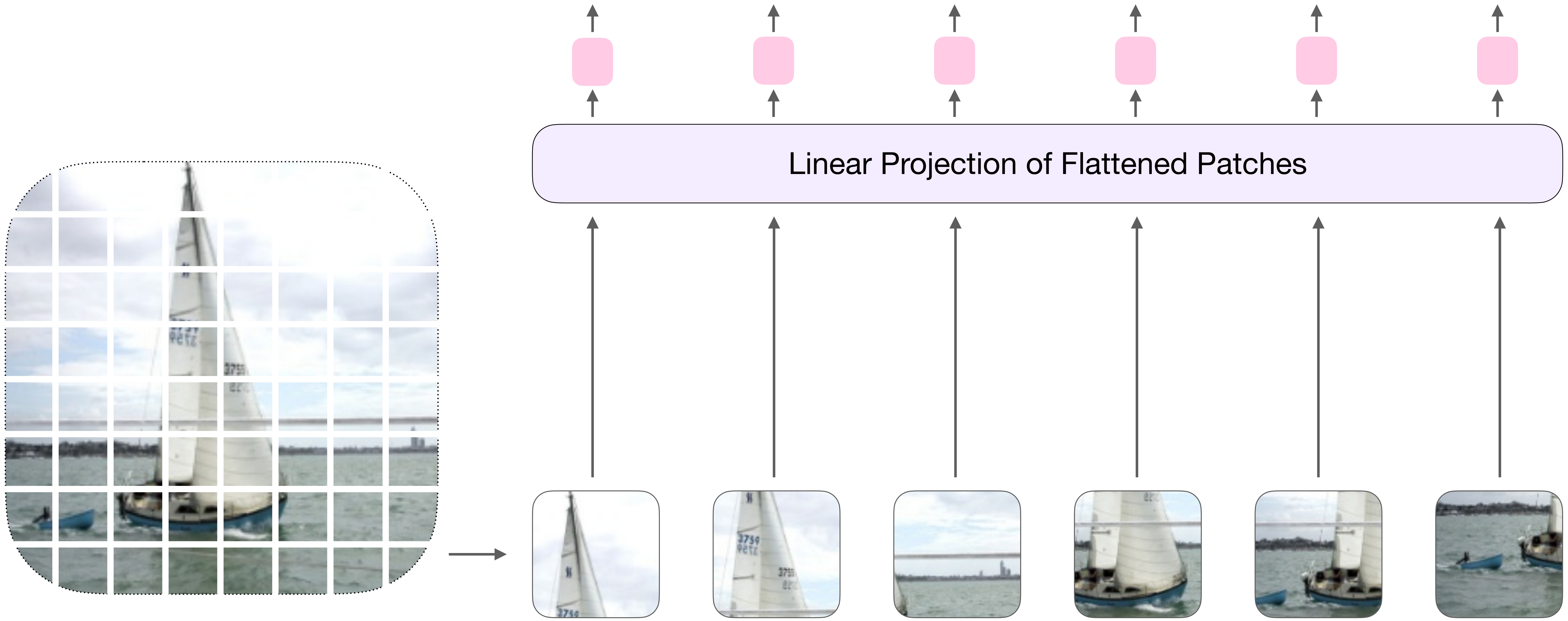
vision transformer



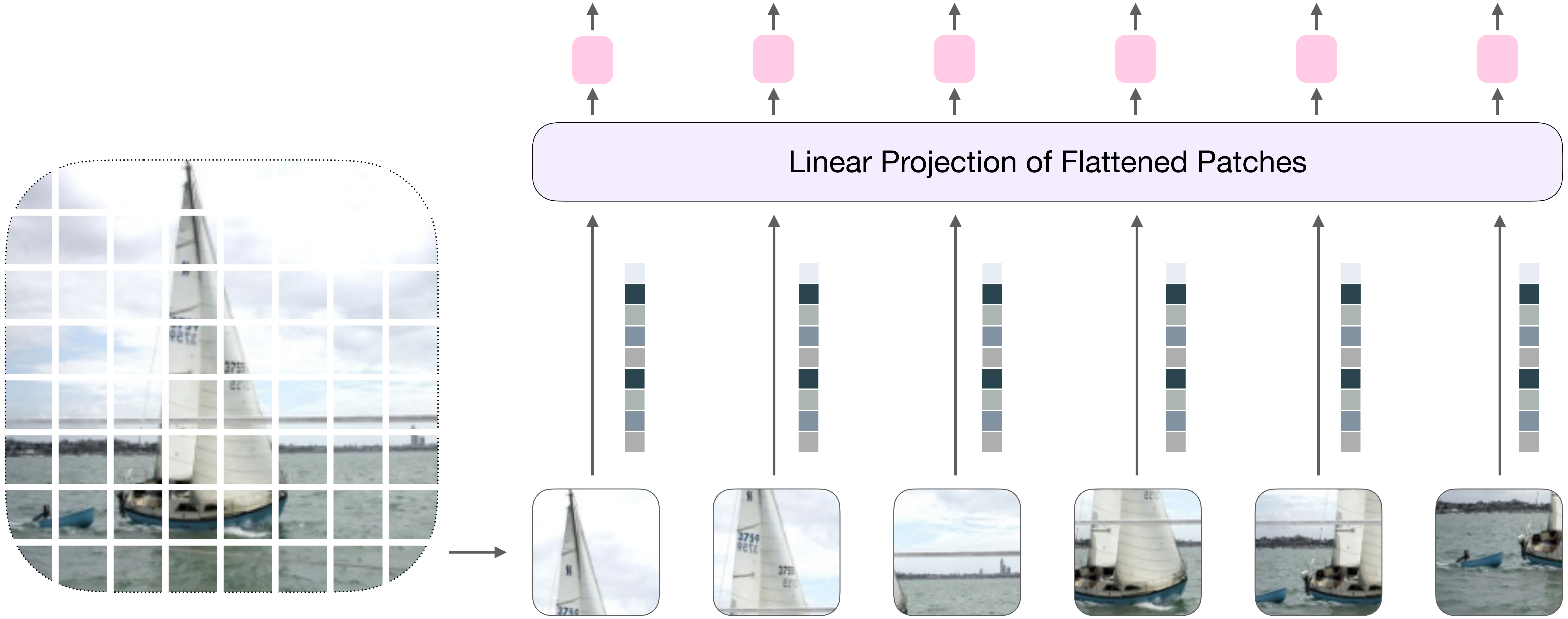
vision transformer



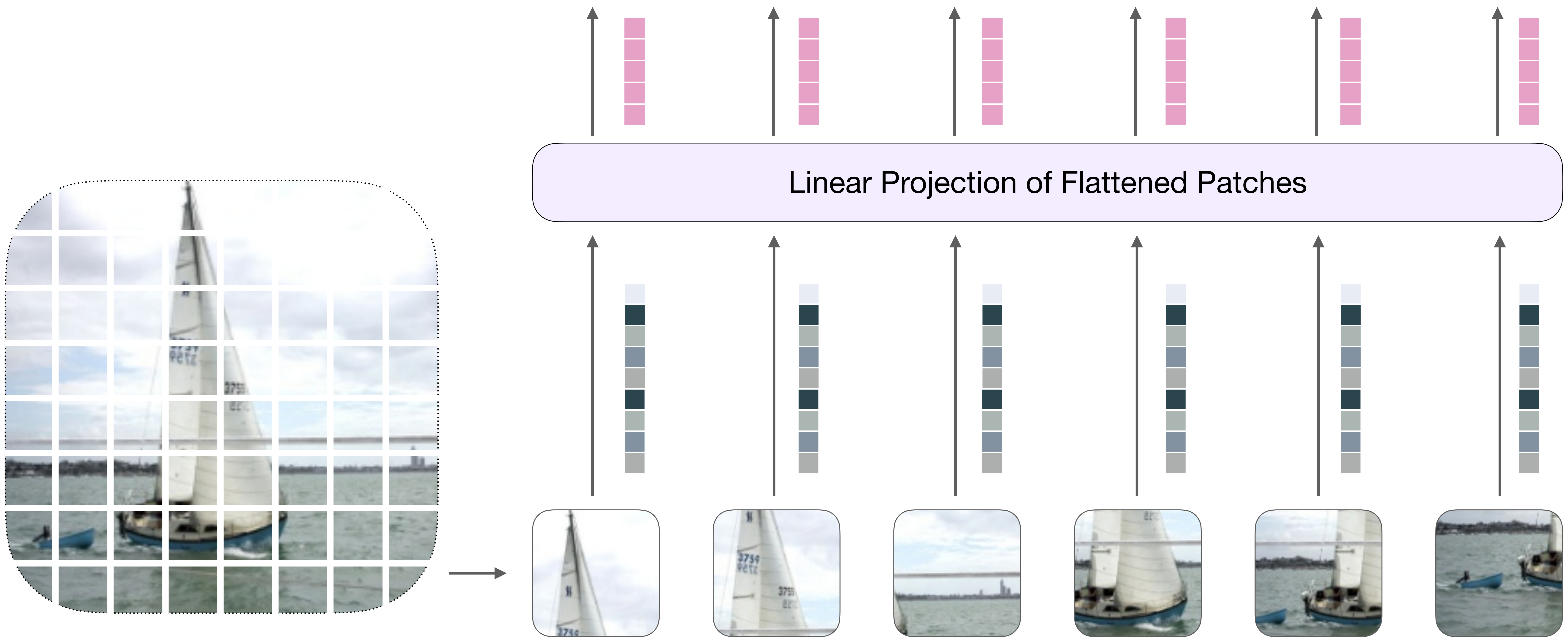
vision transformer



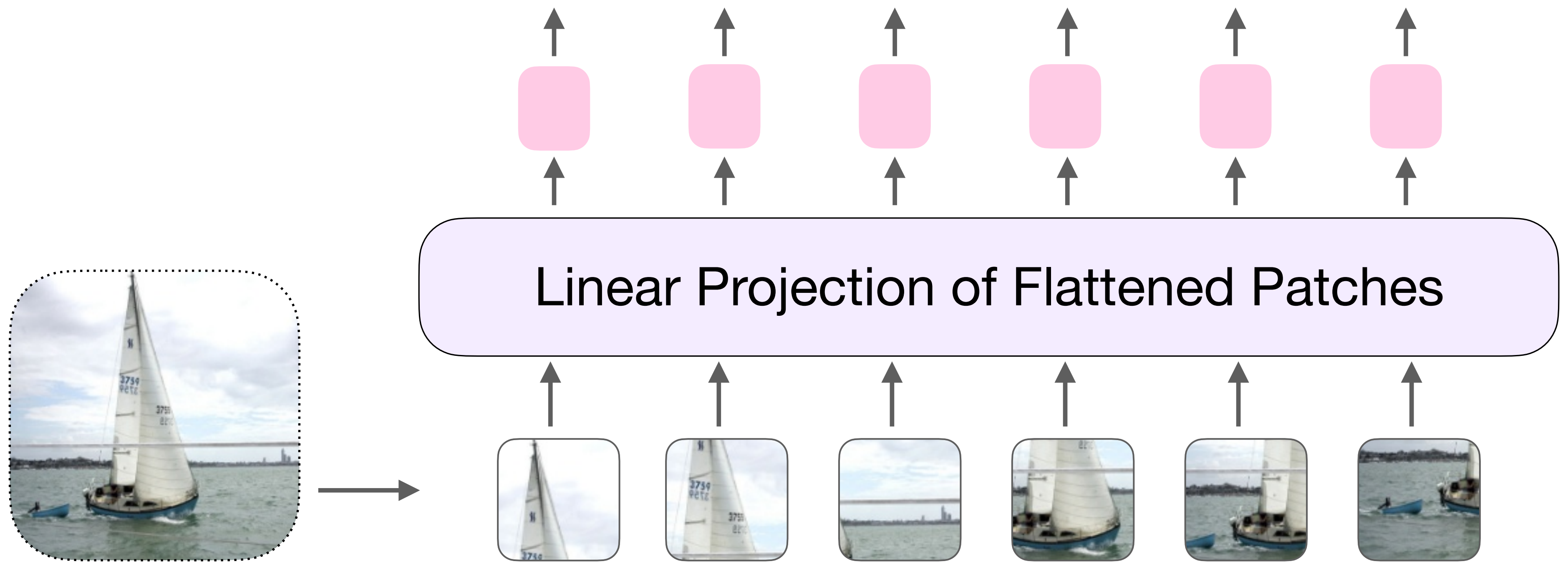
vision transformer



vision transformer



vision transformer



vision transformer

Published as a conference paper at ICLR 2021

arXiv:2010.11929v2 [cs.CV] 3 Jun 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 results compared to state-of-the-art convolutional networks while requiring substantially fewer computational resources to train.[1]

1 Introduction

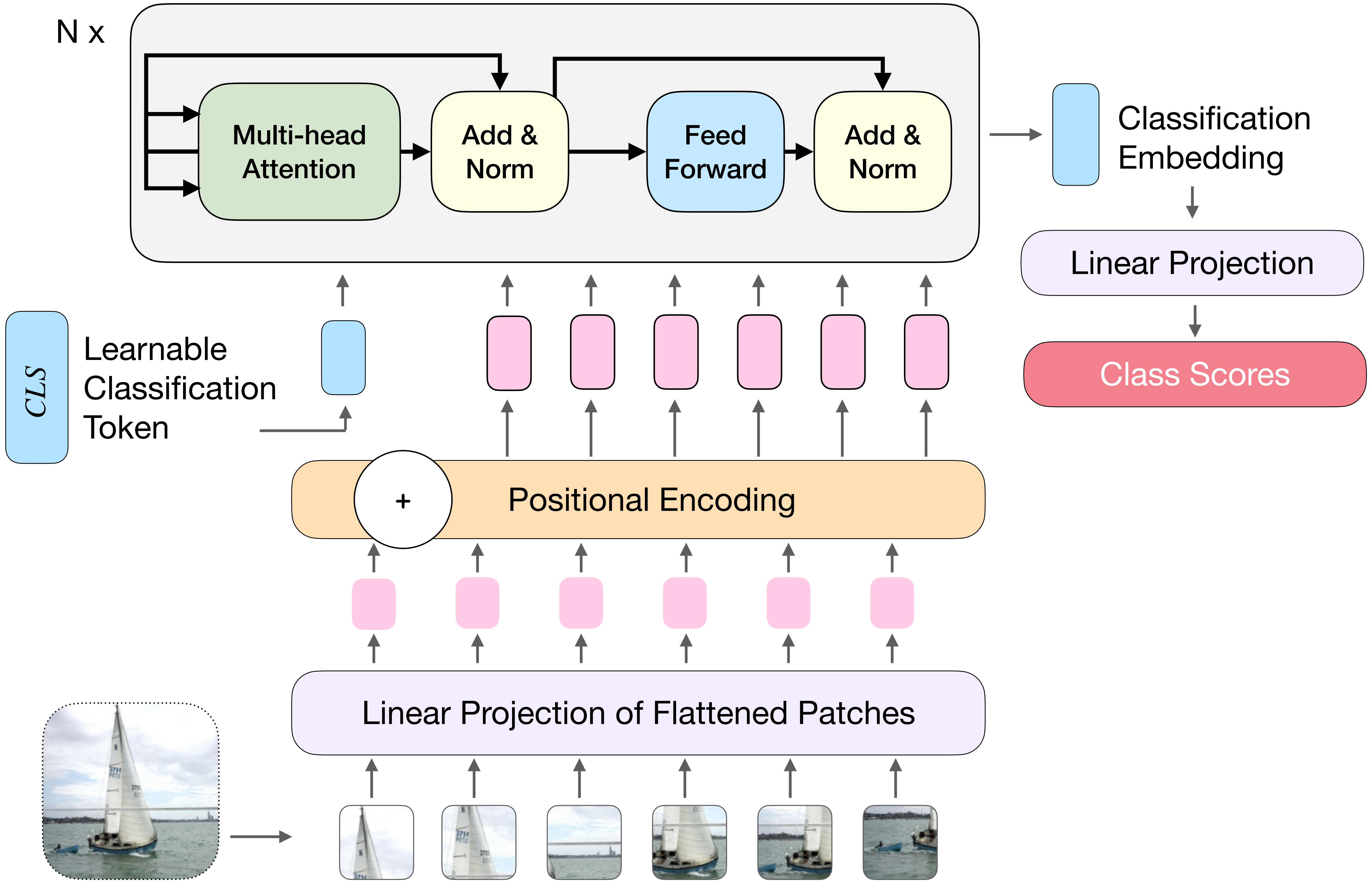
Self-attention-based architectures, in particular Transformers (Vaswani et al., 2017), have become the model of choice in natural language processing (NLP). The dominant approach is to pre-train on a large text corpus and then fine-tune on a smaller task-specific dataset (Devlin et al., 2019). Thanks to Transformers' computational efficiency and scalability, it has become possible to train models of unprecedented size, with over 100B parameters (Brown et al., 2020; Lepikhin et al., 2020). With the models and datasets growing, there is still no sign of saturating performance.

In computer vision, however, convolutional architectures remain dominant (LeCun et al., 1989; Krizhevsky et al., 2012; He et al., 2016). Inspired by NLP successes, multiple works try combining CNN-like architectures with self-attention (Wang et al., 2018; Carion et al., 2020), some replacing the convolutions entirely (Ramachandran et al., 2019; Wang et al., 2020a). The latter models, while theoretically efficient, have not yet been scaled effectively on modern hardware accelerators due to the use of specialized attention patterns. Therefore, in large-scale image recognition, classic ResNet-like architectures are still state of the art (Mahajan et al., 2018; Xie et al., 2020; Kolesnikov et al., 2020).

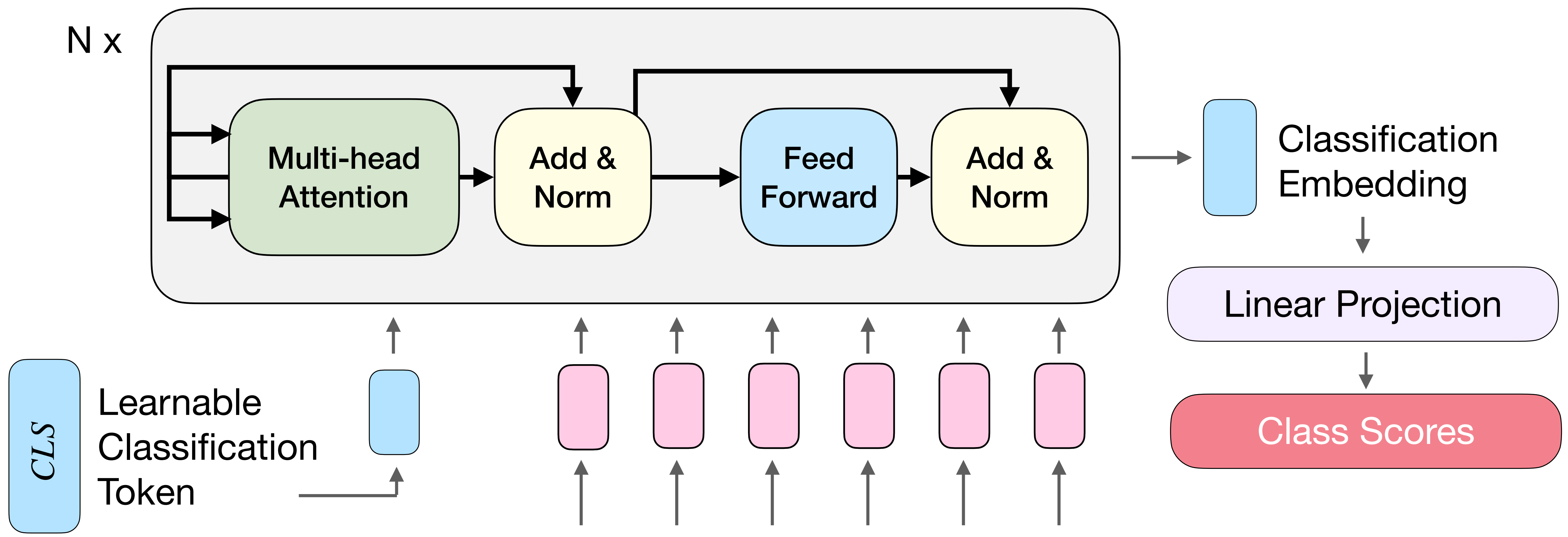
Inspired by the Transformer scaling successes in NLP, we experiment with applying a standard Transformer directly to images, with the fewest possible modifications. To do so, we split an image into patches and provide the sequence of linear embeddings of these patches as an input to a Transformer. Image patches are treated the same way as tokens (words) in an NLP application. We train the model on image classification in supervised fashion.

When trained on mid-sized datasets such as ImageNet without strong regularization, these models yield modest accuracies of a few percentage points below ResNets of comparable size. This seemingly discouraging outcome may be expected: Transformers lack some of the inductive biases

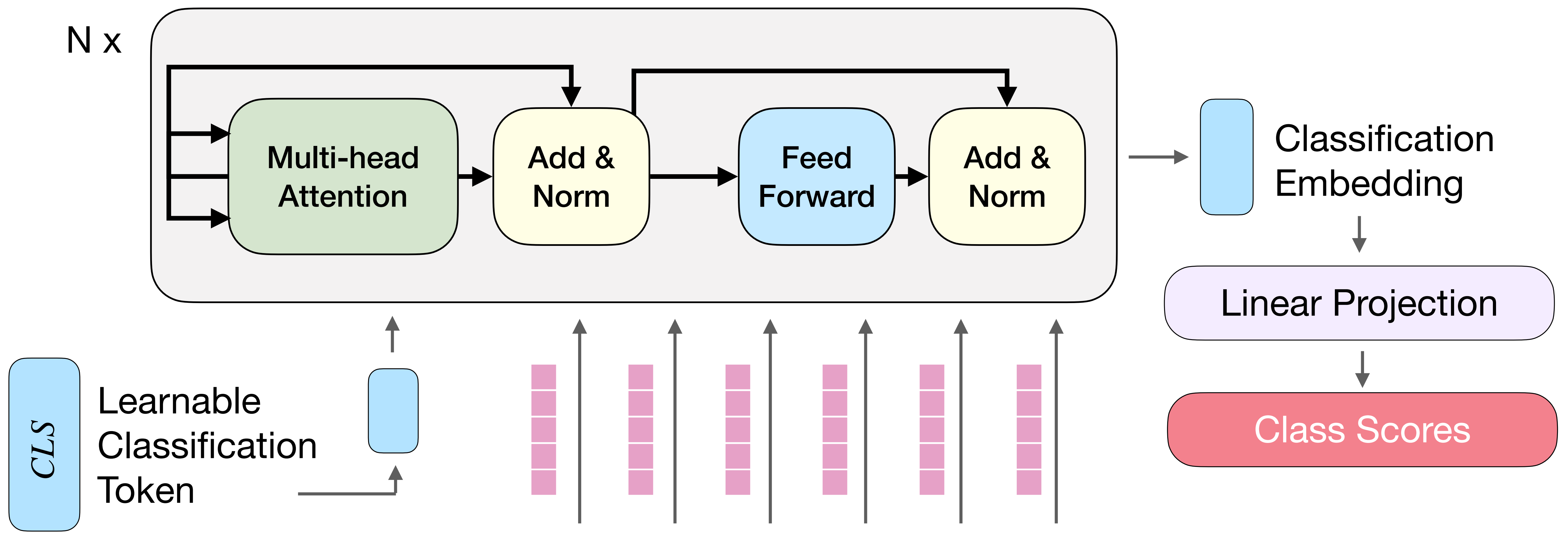
[1] Fine-tuning code and pre-trained models are available at https://github.com/google-research/vision_transformer



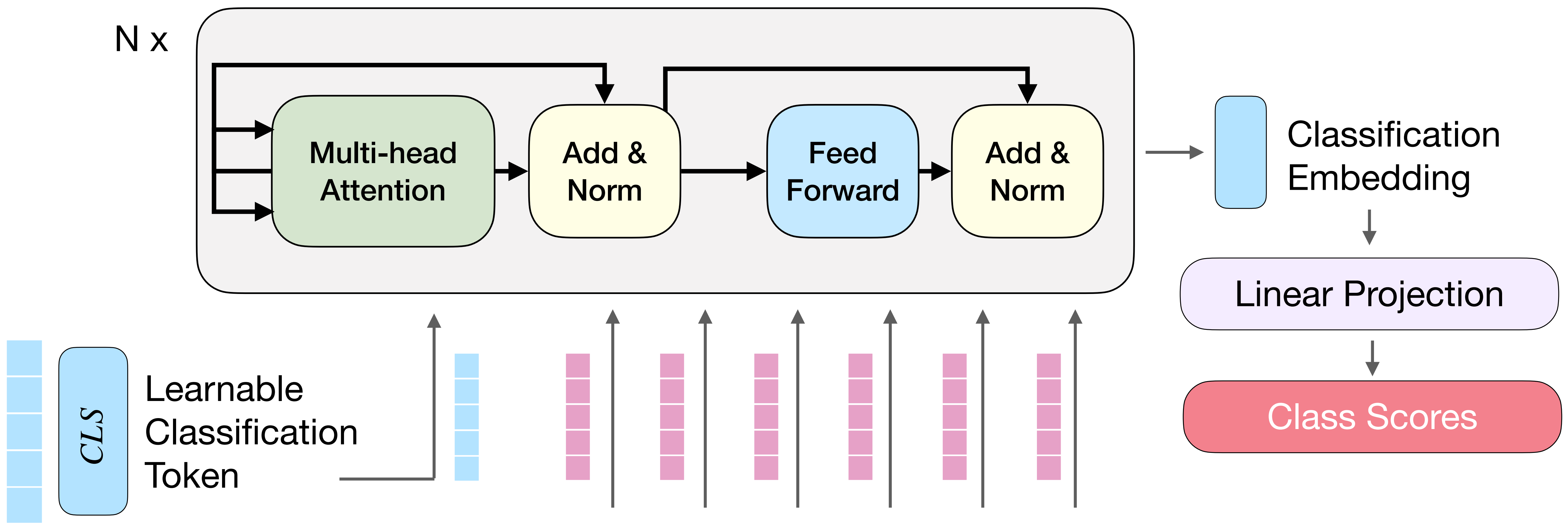
vision transformer



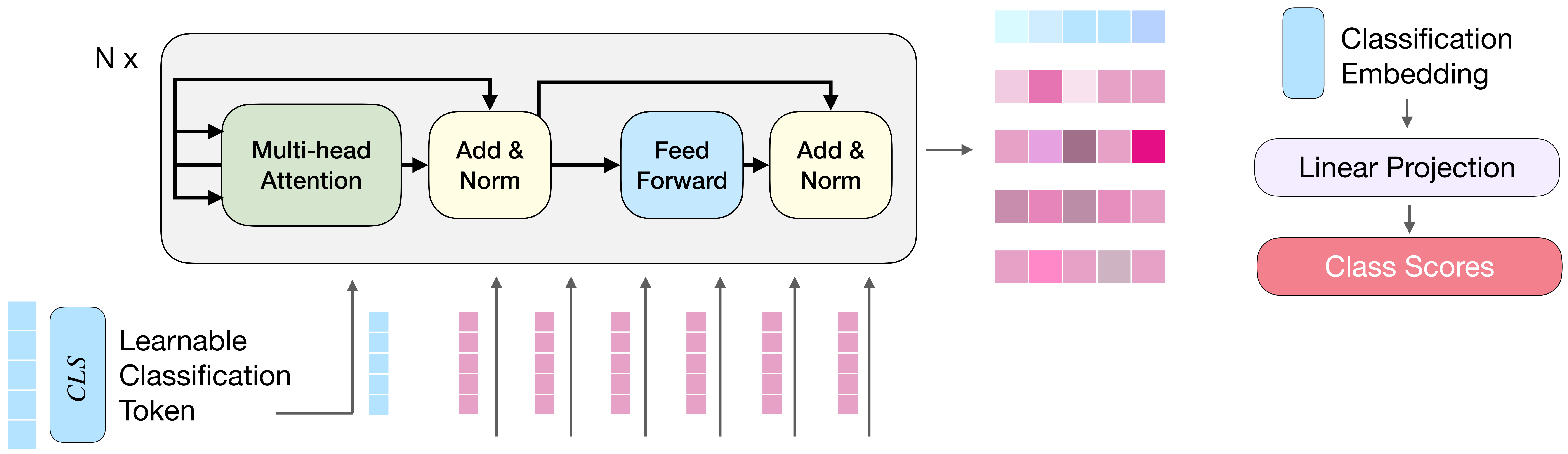
vision transformer



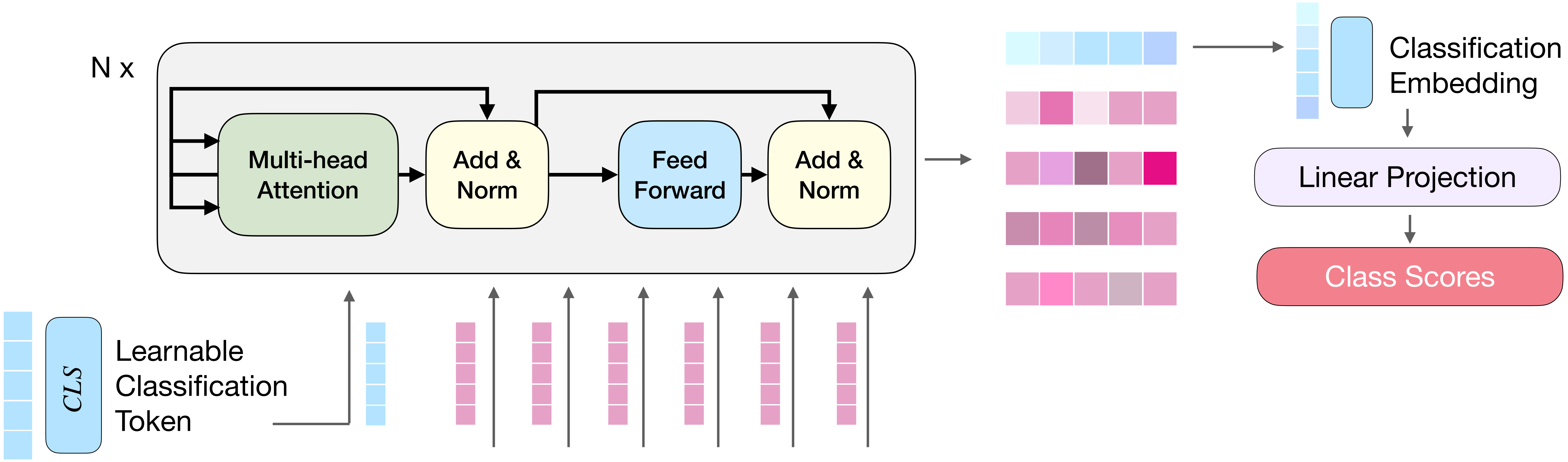
vision transformer



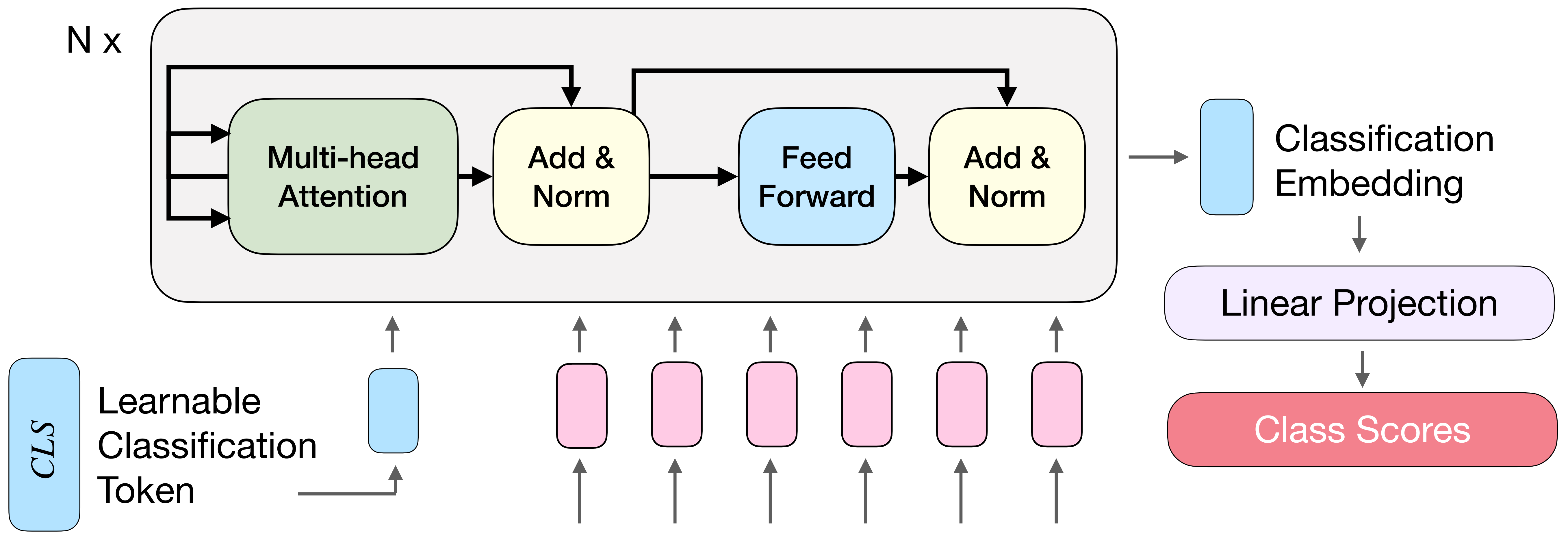
vision transformer



vision transformer



vision transformer



vision transformer


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

arXiv:2010.11929v2 [cs.CV] 3 Jun 2021

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 results compared to state-of-the-art convolutional networks while requiring substantially fewer computational resources to train.[1]

1 INTRODUCTION

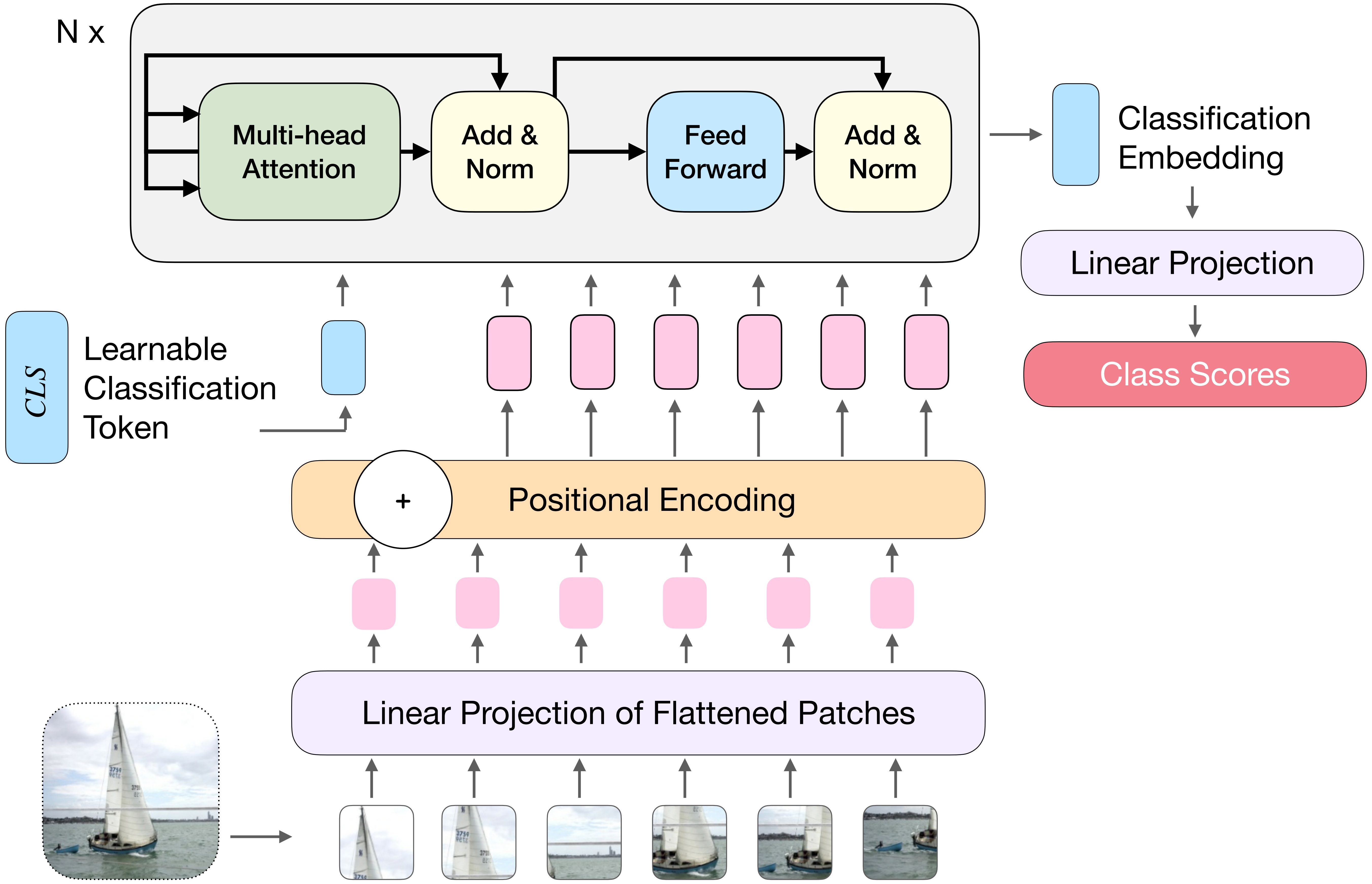
Self-attention-based architectures, in particular Transformers (Vaswani et al., 2017), have become the model of choice in natural language processing (NLP). The dominant approach is to pre-train on a large text corpus and then fine-tune on a smaller task-specific dataset (Devlin et al., 2019). Thanks to Transformers' computational efficiency and scalability, it has become possible to train models of unprecedented size, with over 100B parameters (Brown et al., 2020; Lepikhin et al., 2020). With the models and datasets growing, there is still no sign of saturating performance.

In computer vision, however, convolutional architectures remain dominant (LeCun et al., 1989; Krizhevsky et al., 2012; He et al., 2016). Inspired by NLP successes, multiple works try combining CNN-like architectures with self-attention (Wang et al., 2018; Carion et al., 2020), some replacing the convolutions entirely (Ramachandran et al., 2019; Wang et al., 2020a). The latter models, while theoretically efficient, have not yet been scaled effectively on modern hardware accelerators due to the use of specialized attention patterns. Therefore, in large-scale image recognition, classic ResNet-like architectures are still state of the art (Mahajan et al., 2018; Xie et al., 2020; Kolesnikov et al., 2020).

Inspired by the Transformer scaling successes in NLP, we experiment with applying a standard Transformer directly to images, with the fewest possible modifications. To do so, we split an image into patches and provide the sequence of linear embeddings of these patches as an input to a Transformer. Image patches are treated the same way as tokens (words) in an NLP application. We train the model on image classification in supervised fashion.

When trained on mid-sized datasets such as ImageNet without strong regularization, these models yield modest accuracies of a few percentage points below ResNets of comparable size. This seemingly discouraging outcome may be expected: Transformers lack some of the inductive biases

[1]Fine-tuning code and pre-trained models are available at https://github.com/google-research/vision_transformer



vision transformer

arXiv:2012.12877v2 [cs.CV] 15 Jan 2021

Training data-efficient image transformers & distillation through attention

Hugo Touvron[*,†] Matthieu Cord[†] Matthijs Douze[*]
Francisco Massa[*] Alexandre Sablayrolles[*] Hervé Jégou[*]

[*]Facebook AI [†]Sorbonne University

Abstract

Recently, neural networks purely based on attention were shown to address image understanding tasks such as image classification. These high-performing vision transformers are pre-trained with hundreds of millions of images using a large infrastructure, thereby limiting their adoption.

In this work, we produce competitive convolution-free transformers by training on Imagenet only. We train them on a single computer in less than 3 days. Our reference vision transformer (86M parameters) achieves top-1 accuracy of 83.1% (single-crop) on ImageNet with no external data.

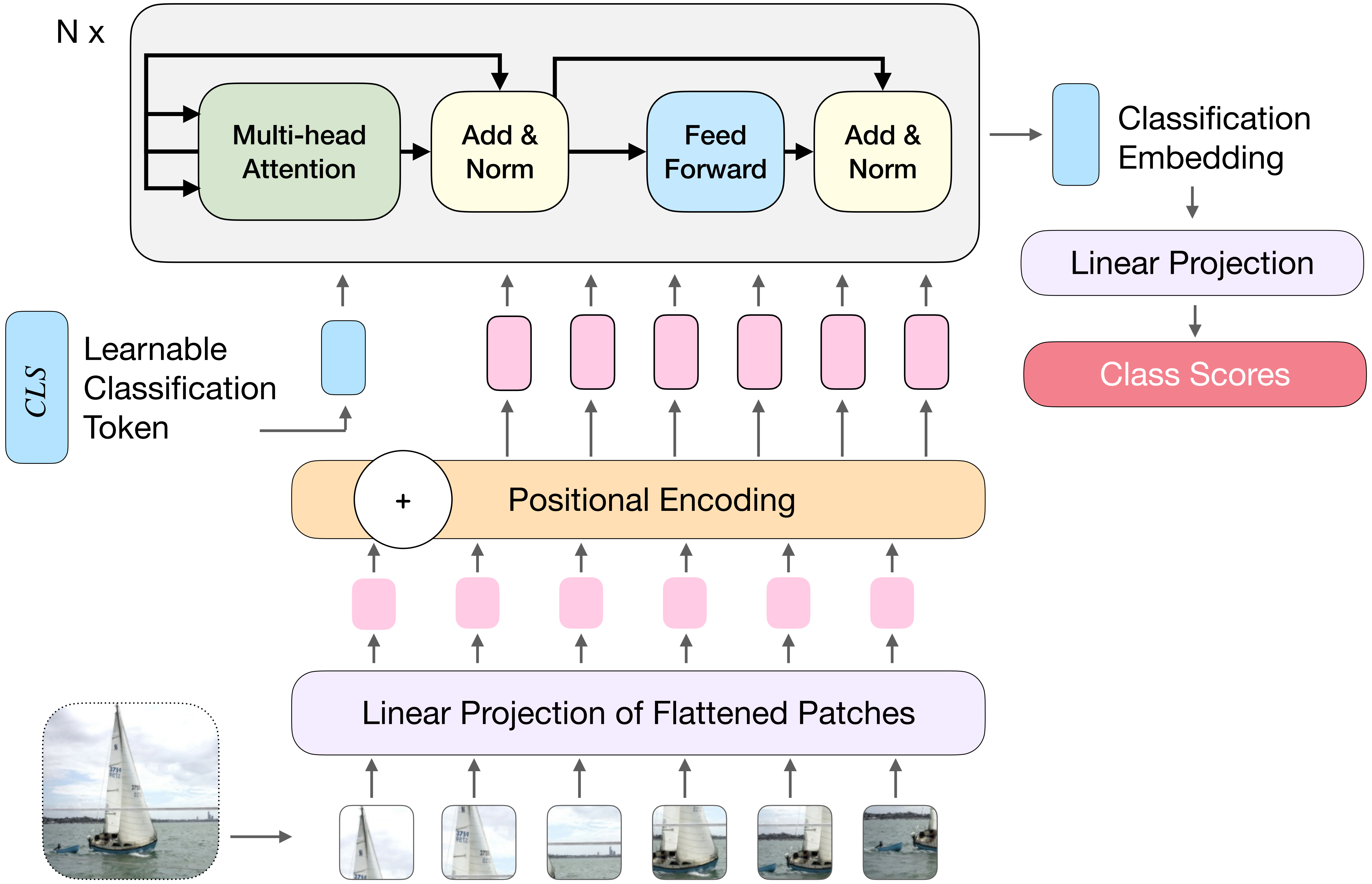
More importantly, we introduce a teacher-student strategy specific to transformers. It relies on a distillation token ensuring that the student learns from the teacher through attention. We show the interest of this token-based distillation, especially when using a convnet as a teacher. This leads us to report results competitive with convnets for both Imagenet (where we obtain up to 85.2% accuracy) and when transferring to other tasks. We share our code and models.

1 Introduction

Convolutional neural networks have been the main design paradigm for image understanding tasks, as initially demonstrated on image classification tasks. One of the ingredient to their success was the availability of a large training set, namely Imagenet [13, 42]. Motivated by the success of attention-based models in Natural Language Processing [14, 52], there has been increasing interest in architectures leveraging attention mechanisms within convnets [2, 34, 61]. More recently several researchers have proposed hybrid architecture transplanting transformer ingredients to convnets to solve vision tasks [6, 43].

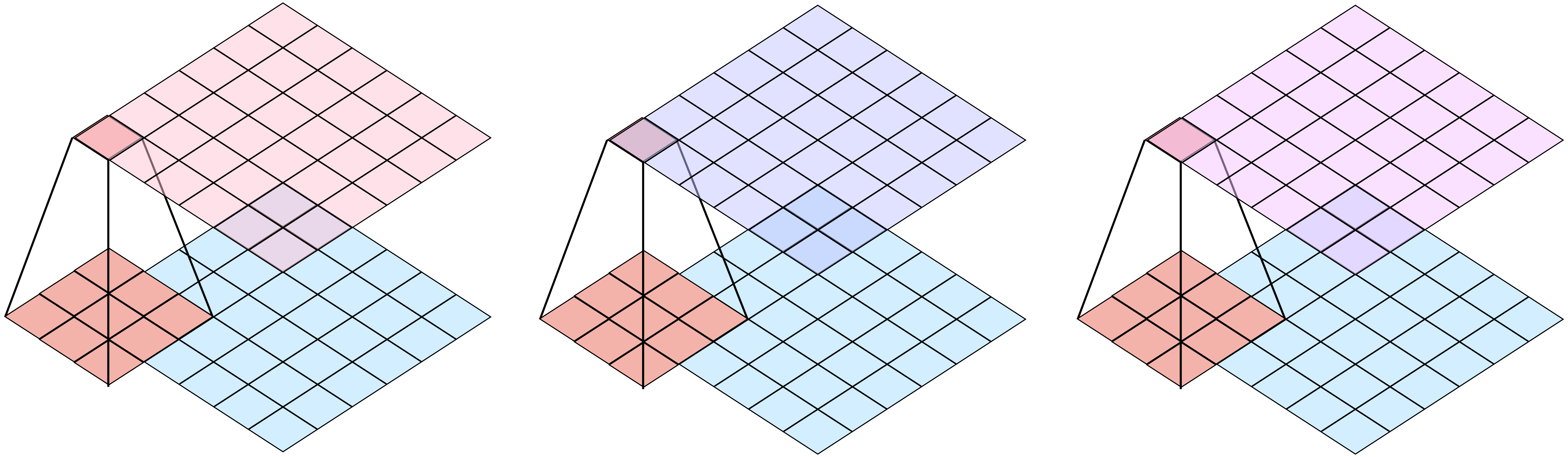
The vision transformer (ViT) introduced by Dosovitskiy et al. [15] is an architecture directly inherited from Natural Language Processing [52], but ap-

1

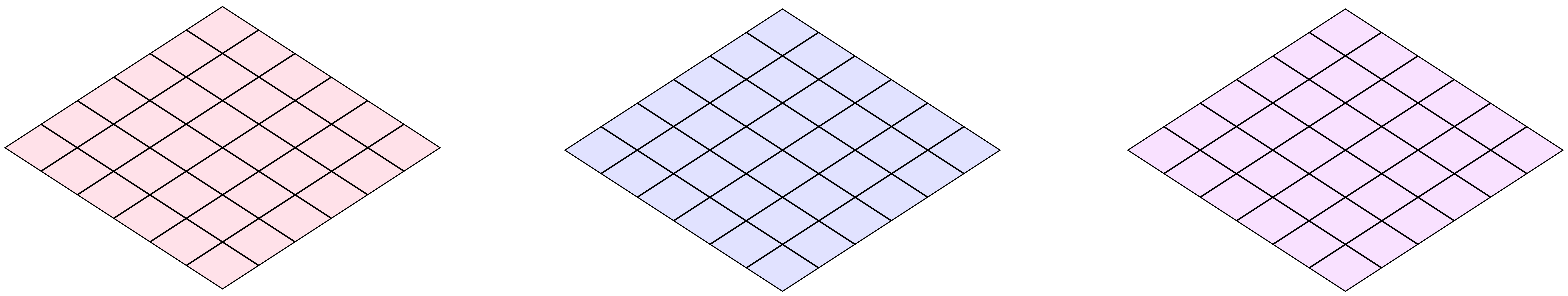


vision transformer

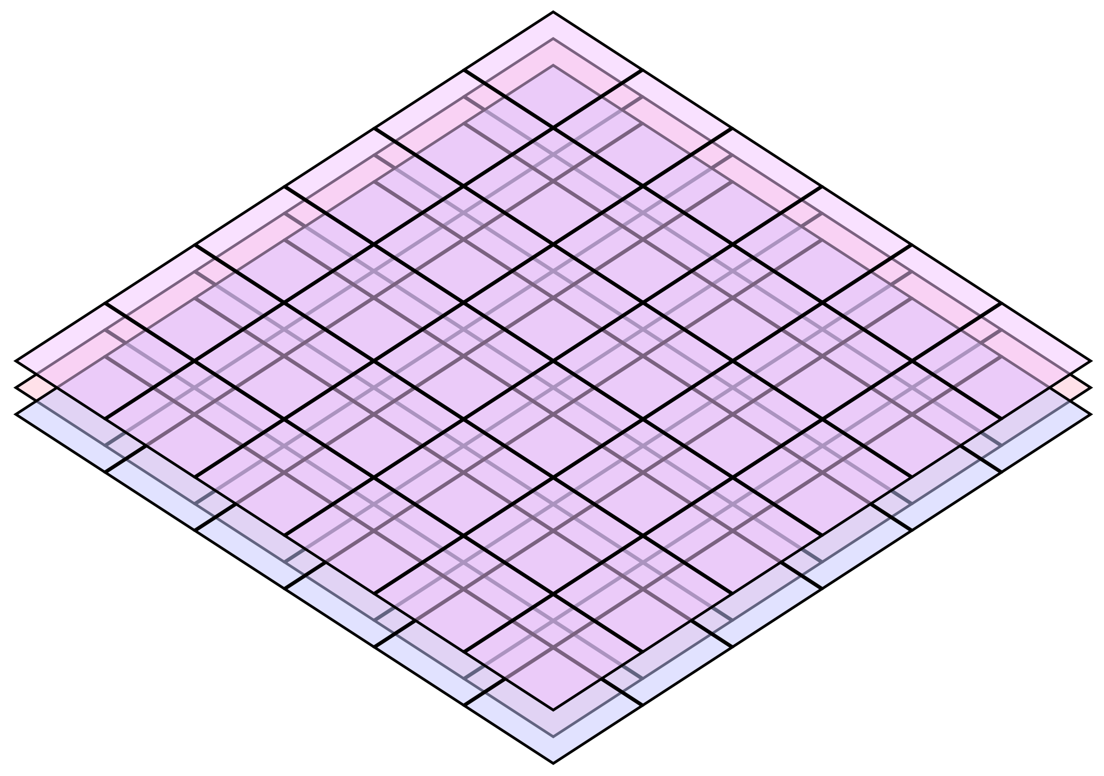
$$C = A \circledast B \qquad C_{x,y} = \sum_{dx=-a}^{a} \sum_{dy=-b}^{b} A_{dx,dy} B_{x+dx,y+dy}$$



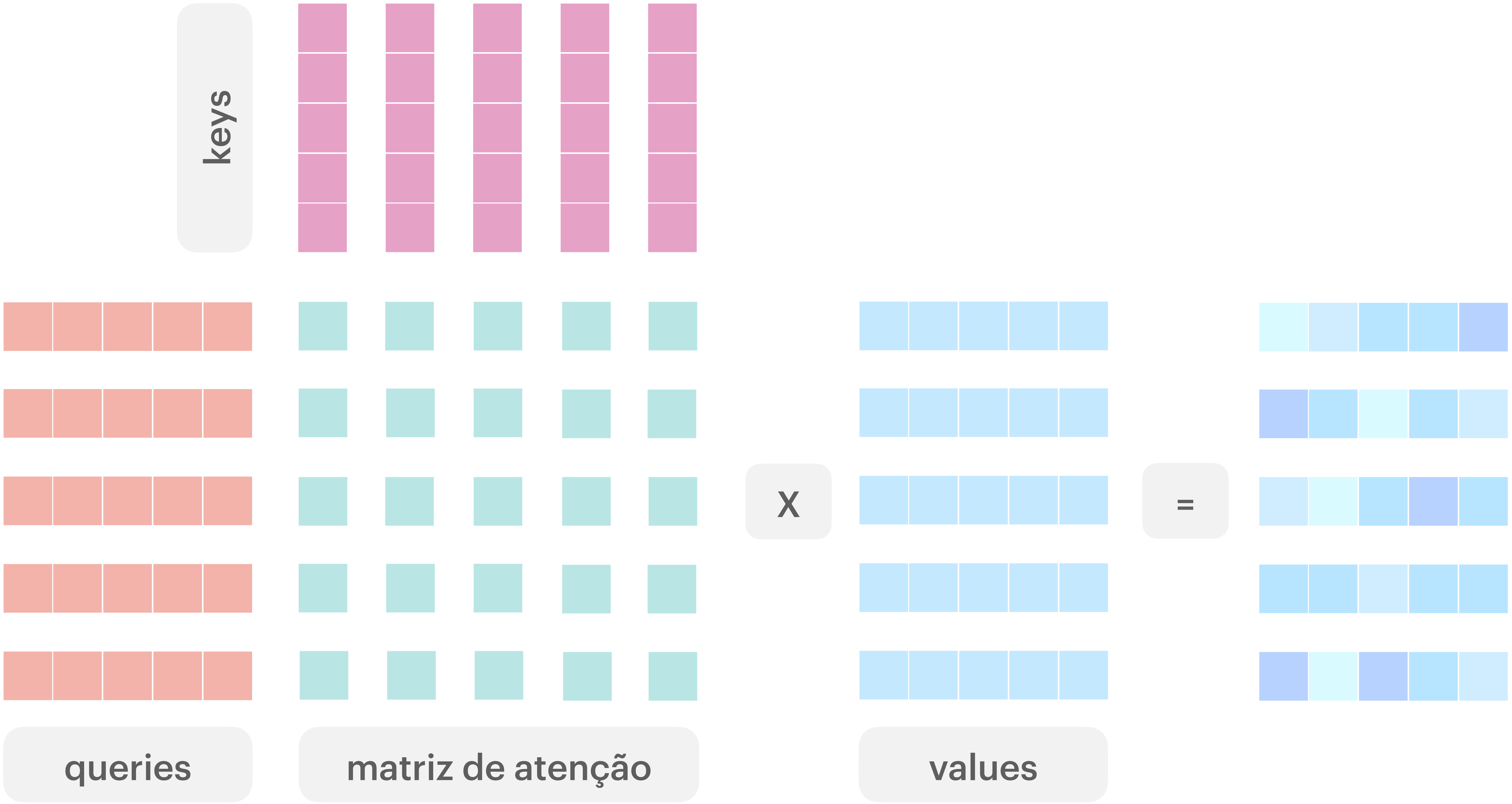
vision transformer



vision transformer



vision transformer



vision transformer

arXiv:2012.12877v2 [cs.CV] 15 Jan 2021

Training data-efficient image transformers
& distillation through attention

Hugo Touvron*,† Matthieu Cord† Matthijs Douze*
Francisco Massa* Alexandre Sablayrolles* Hervé Jégou*

*Facebook AI †Sorbonne University

Abstract

Recently, neural networks purely based on attention were shown to address image understanding tasks such as image classification. These high-performing vision transformers are pre-trained with hundreds of millions of images using a large infrastructure, thereby limiting their adoption.

In this work, we produce competitive convolution-free transformers by training on Imagenet only. We train them on a single computer in less than 3 days. Our reference vision transformer (86M parameters) achieves top-1 accuracy of 83.1% (single-crop) on ImageNet with no external data.

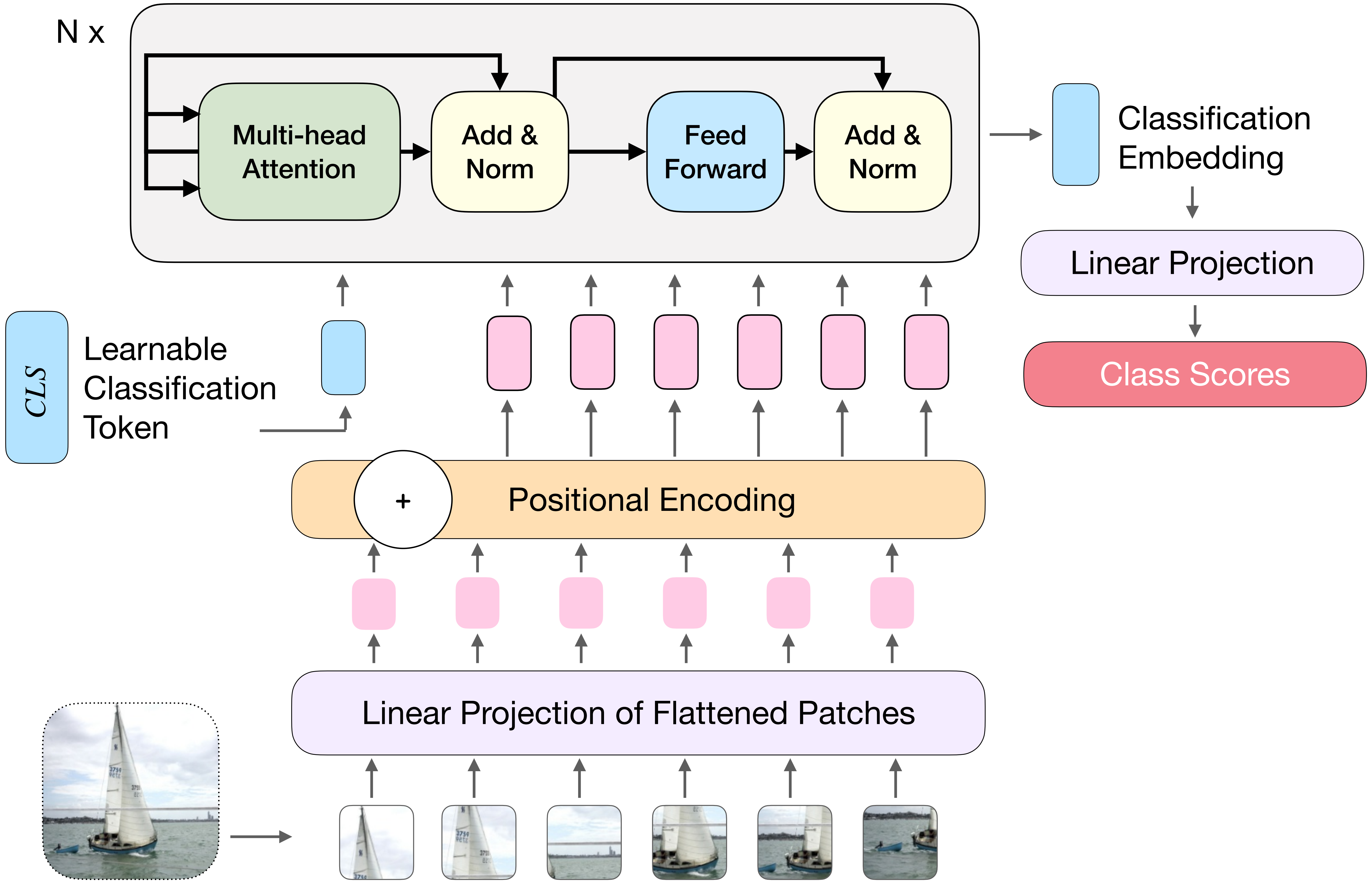
More importantly, we introduce a teacher-student strategy specific to transformers. It relies on a distillation token ensuring that the student learns from the teacher through attention. We show the interest of this token-based distillation, especially when using a convnet as a teacher. This leads us to report results competitive with convnets for both Imagenet (where we obtain up to 85.2% accuracy) and when transferring to other tasks. We share our code and models.

1 Introduction

Convolutional neural networks have been the main design paradigm for image understanding tasks, as initially demonstrated on image classification tasks. One of the ingredient to their success was the availability of a large training set, namely Imagenet [13, 42]. Motivated by the success of attention-based models in Natural Language Processing [14, 52], there has been increasing interest in architectures leveraging attention mechanisms within convnets [2, 34, 61]. More recently several researchers have proposed hybrid architecture transplanting transformer ingredients to convnets to solve vision tasks [6, 43].

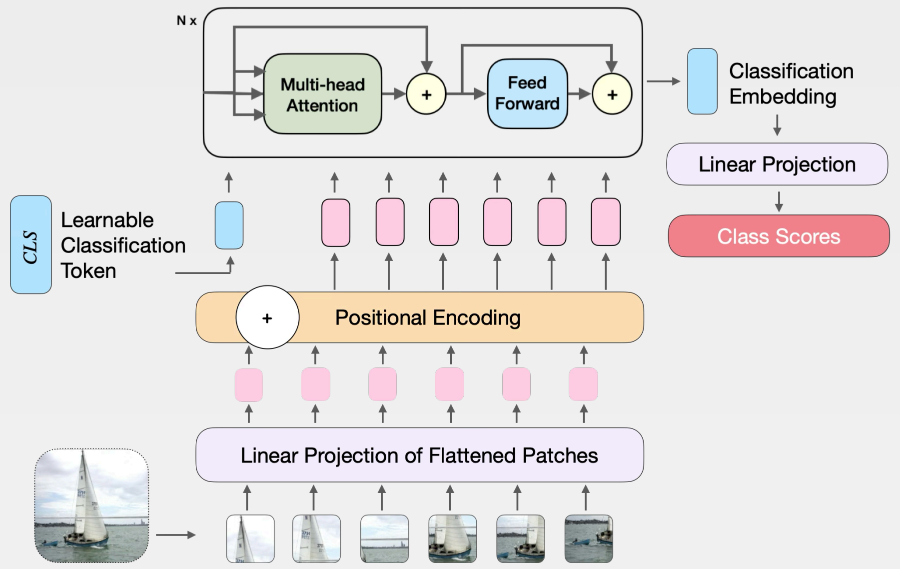
The vision transformer (ViT) introduced by Dosovitskiy et al. [15] is an architecture directly inherited from Natural Language Processing [52], but ap-

1

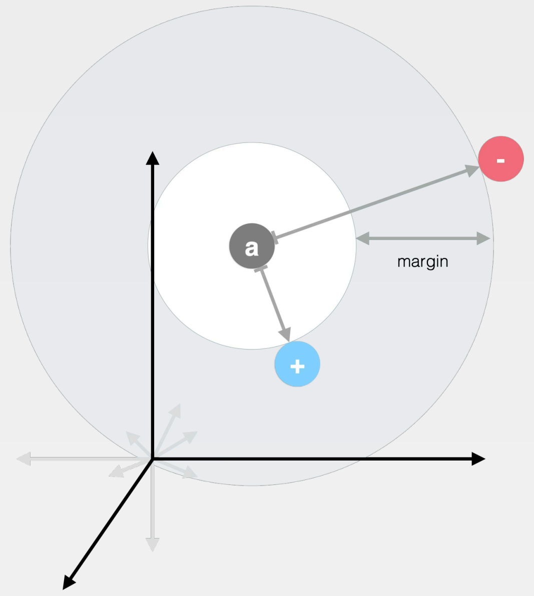


SCC0251
Processamento de Imagens

Aprendizado Profundo



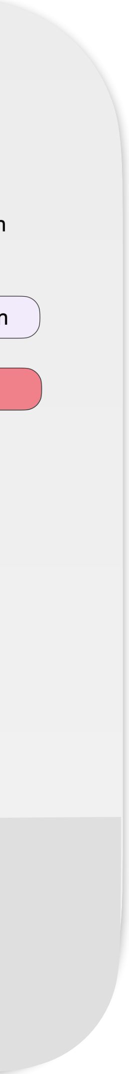
Transformers

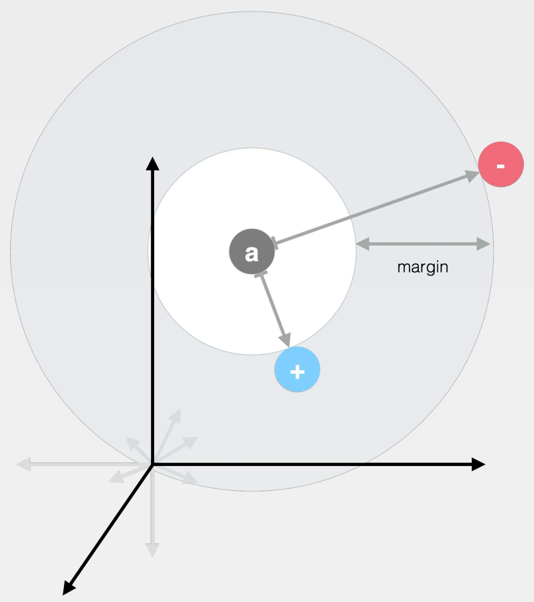


Contrastive Learning

SCC0251
Processamento de Imagens

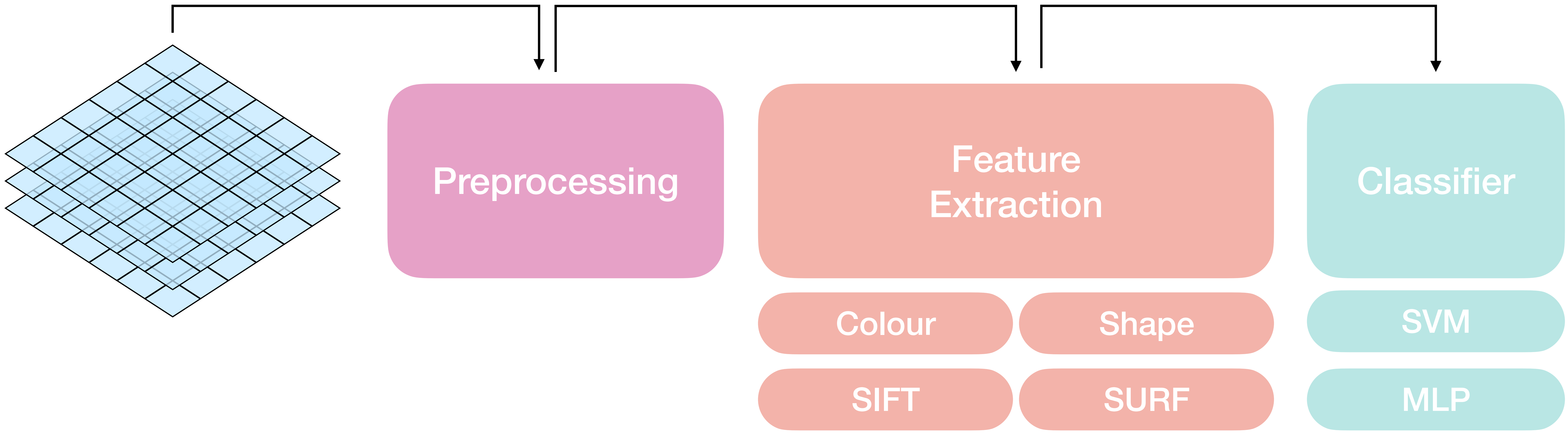
Aprendizado Profundo



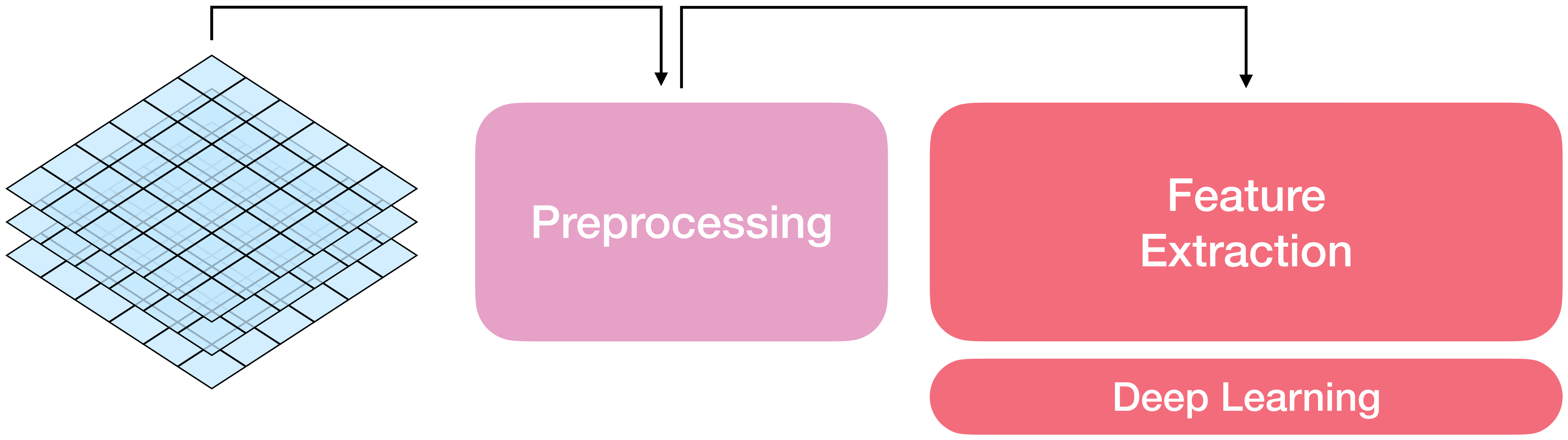


Contrastive Learning

learning representations

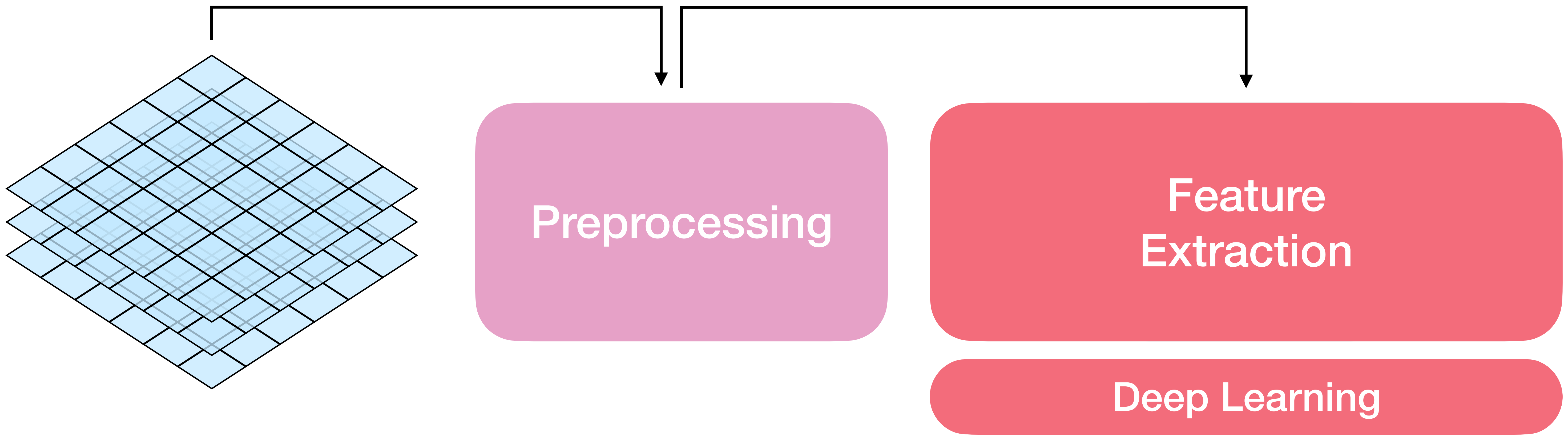


learning representations



feature abstraction became a learnable part of the pipeline

learning representations

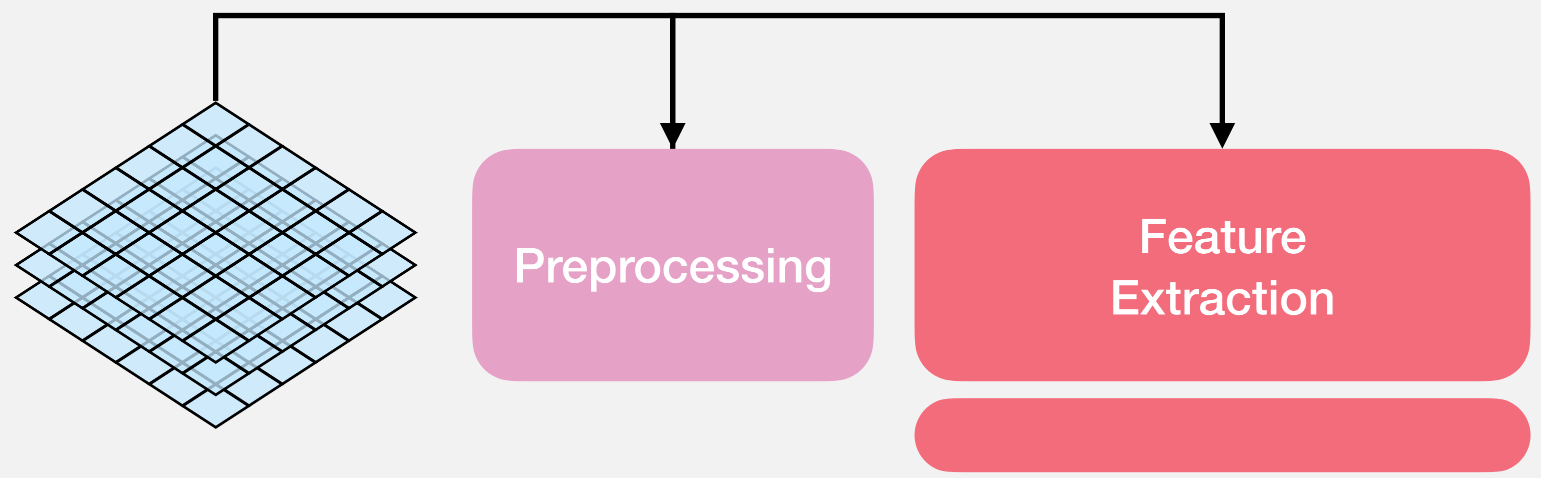


feature abstraction became a learnable part of the pipeline

Representation Learning

Bengio et al.

learning representations



Content-based image retrieval

Pre-training for multitask

Face Recognition

Recommendation Systems

Unsupervised Learning

Self-supervised Learning

Representation Learning

Bengio et al.

contrastive learning

Dimensionality Reduction by Learning an Invariant Mapping

Raia Hadsell, Sumit Chopra, Yann LeCun
The Courant Institute of Mathematical Sciences
New York University, 719 Broadway, New York, NY 1003, USA.
http://www.cs.nyu.edu/~yann
(November 2005. To appear in CVPR 2006)

Abstract

Dimensionality reduction involves mapping a set of high dimensional input points onto a low dimensional manifold so that "similar" points in input space are mapped to nearby points on the manifold. Most existing techniques for solving the problem suffer from two drawbacks. First, most of them depend on a meaningful and computable distance

the input space.

For example, Locally Linear Embedding (LLE) [15] linearly combines input vectors that are identified as neighbors. The applicability of LLE and similar methods to image data is limited because linearly combining images only makes sense for images that are perfectly registered and very similar. Laplacian Eigenmap [2] and Hessian LLE [8] do not require a meaningful metric in input space (they

introduced the contrastive loss as a technique to do dimensionality reduction

contrastive learning

Dimensionality Reduction by Learning an Invariant Mapping

Raia Hadsell, Sumit Chopra, Yann LeCun
The Courant Institute of Mathematical Sciences
New York University, 719 Broadway, New York, NY 1003, USA.
http://www.cs.nyu.edu/~yann
(November 2005. To appear in CVPR 2006)

Abstract

Dimensionality reduction involves mapping a set of high dimensional input points onto a low dimensional manifold so that "similar" points in input space are mapped to nearby points on the manifold. Most existing techniques for solving the problem suffer from two drawbacks. First, most of them depend on a meaningful and computable distance

the input space.

For example, Locally Linear Embedding (LLE) [15] linearly combines input vectors that are identified as neighbors. The applicability of LLE and similar methods to image data is limited because linearly combining images only makes sense for images that are perfectly registered and very similar. Laplacian Eigenmap [2] and Hessian LLE [8] do not require a meaningful metric in input space (they

introduced the contrastive loss as a technique to do dimensionality reduction

$$\mathcal{L}(W,(Y,x_1,x_2)) = \frac{1}{2}(1-Y)(D(W,x_1,x_2))^2 + \frac{1}{2}Y\max\{0,\ m - D(W,x_1,x_2)\}^2$$

contrastive learning

$$\mathscr{L}(W, (Y, x_1, x_2)) = \frac{1}{2}(1 - Y)(D(W, x_1, x_2))^2 + \frac{1}{2}Y \max\{0,\ m - D(W, x_1, x_2)\}^2$$

let's break this apart

contrastive learning

each x is a sample

$$\mathscr{L}(W, (Y, x_1, x_2)) = \frac{1}{2}(1 - Y)(D(W, x_1, x_2))^2 + \frac{1}{2} Y \max\{0,\ m - D(W, x_1, x_2)\}^2$$

let's break this apart

contrastive learning

$$\mathscr{L}(W, (Y, x_1, x_2)) = \frac{1}{2}(1 - Y)(D(W, x_1, x_2))^2 + \frac{1}{2} Y \max\{0,\ m - D(W, x_1, x_2)\}^2$$

y is a binary label

0 if the pair is similar

1 if the pair is dissimilar

let's break this apart

contrastive learning

$$\mathscr{L}(W, (Y, x_1, x_2)) = \frac{1}{2}(1 - Y)(D(W, x_1, x_2))^2 + \frac{1}{2} Y \max\{0,\ m - D(W, x_1, x_2)\}^2$$

W are the network weights

let's break this apart

contrastive learning

D is a distance function

usually euclidean

$$\mathcal{L}(W,(Y,x_1,x_2)) = \frac{1}{2}(1-Y)(D(W,x_1,x_2))^2 + \frac{1}{2}Y\max\{0,\ m - D(W,x_1,x_2)\}^2$$

let's break this apart

contrastive learning

if the pair is similar

decrease D between both samples

$$\mathcal{L}(W,(Y,x_1,x_2)) = \frac{1}{2}(1-Y)(D(W,x_1,x_2))^2 + \frac{1}{2}Y\max\{0,\ m - D(W,x_1,x_2)\}^2$$

let's break this apart

contrastive learning

$$\mathscr{L}(W,(Y,x_1,x_2)) = \frac{1}{2}(1-Y)(D(W,x_1,x_2))^2 + \frac{1}{2}Y\max\{0,\ m - D(W,x_1,x_2)\}^2$$

let's break this apart

if the pair is dissimilar

increase D between both samples

but only up until it is larger than m

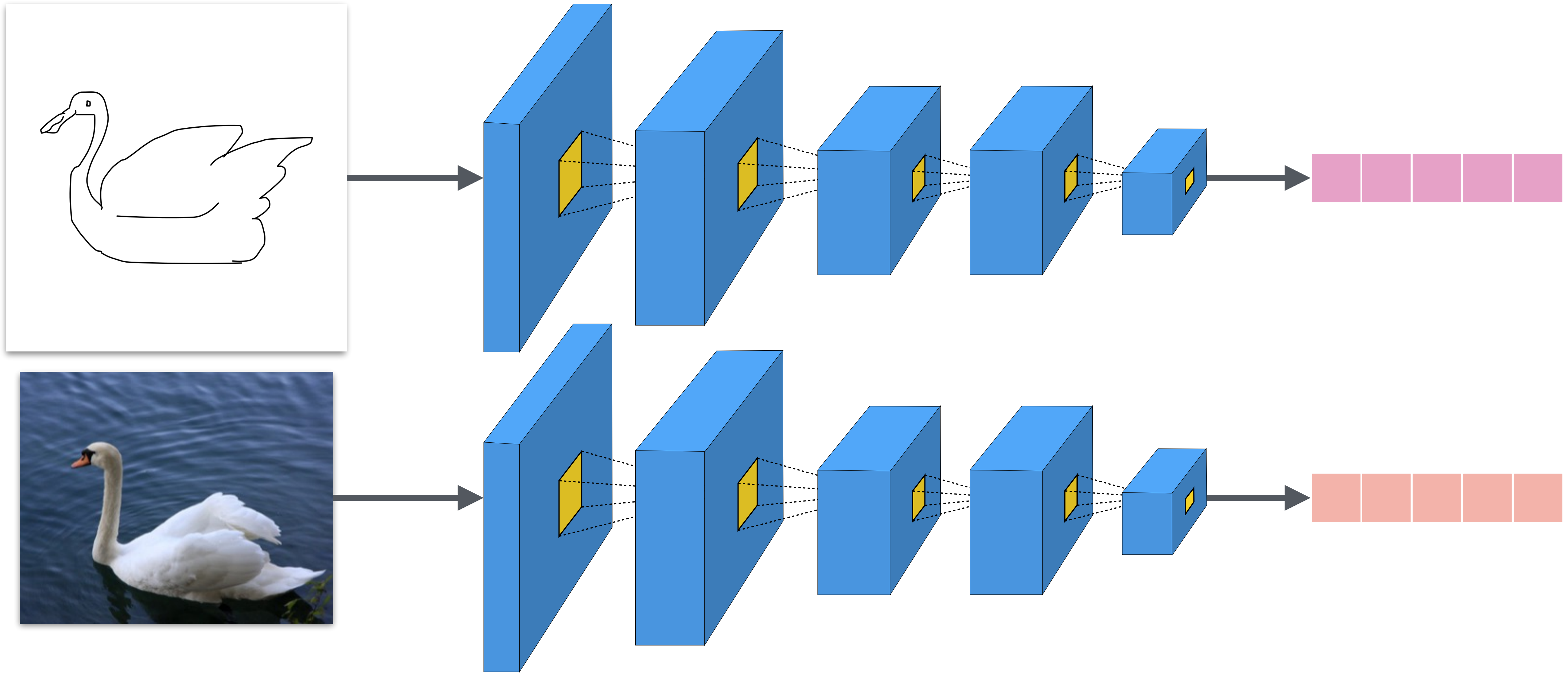
contrastive learning

$$\mathcal{L}(W,(Y,x_1,x_2)) = \frac{1}{2}(1-Y)(D(W,x_1,x_2))^2 + \frac{1}{2}Y\max\{0, m - D(W,x_1,x_2)\}^2$$

m is the margin

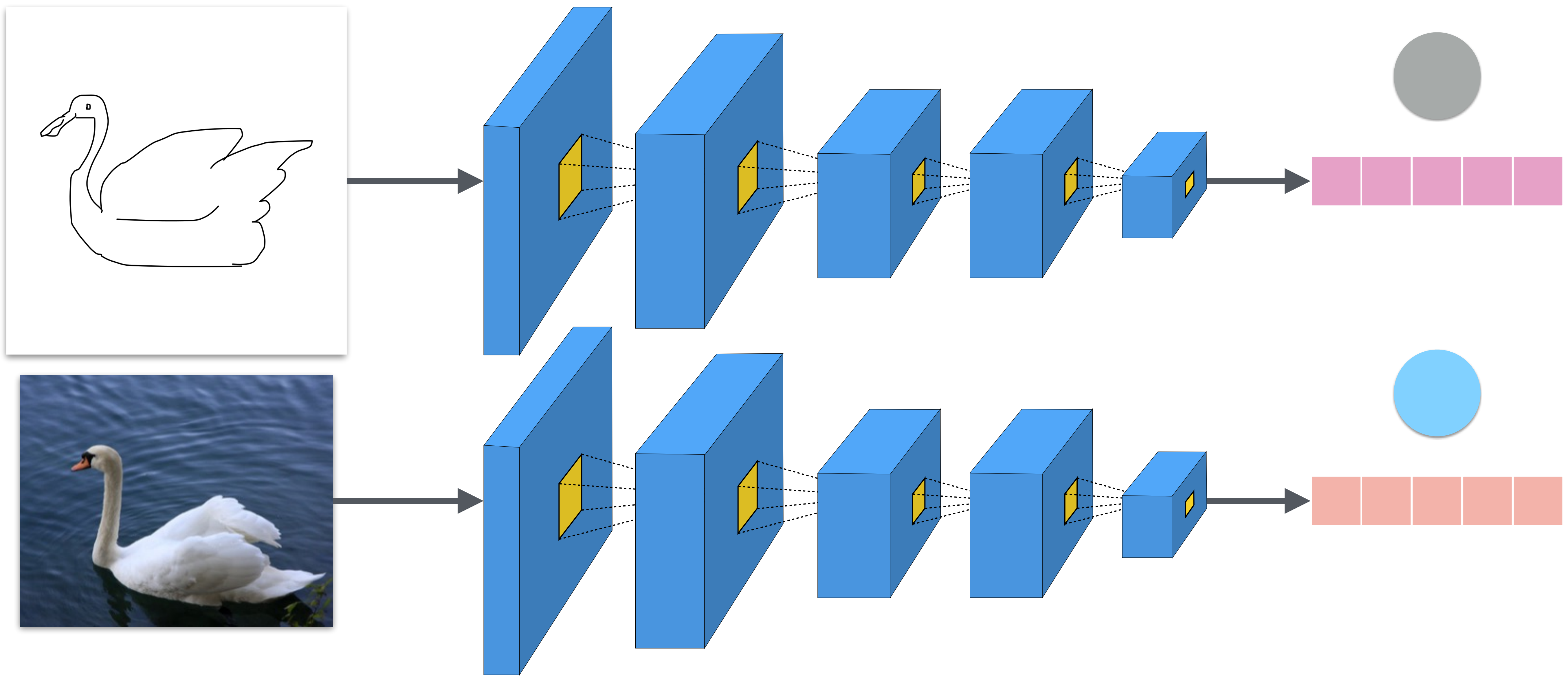
let's break this apart

contrastive learning

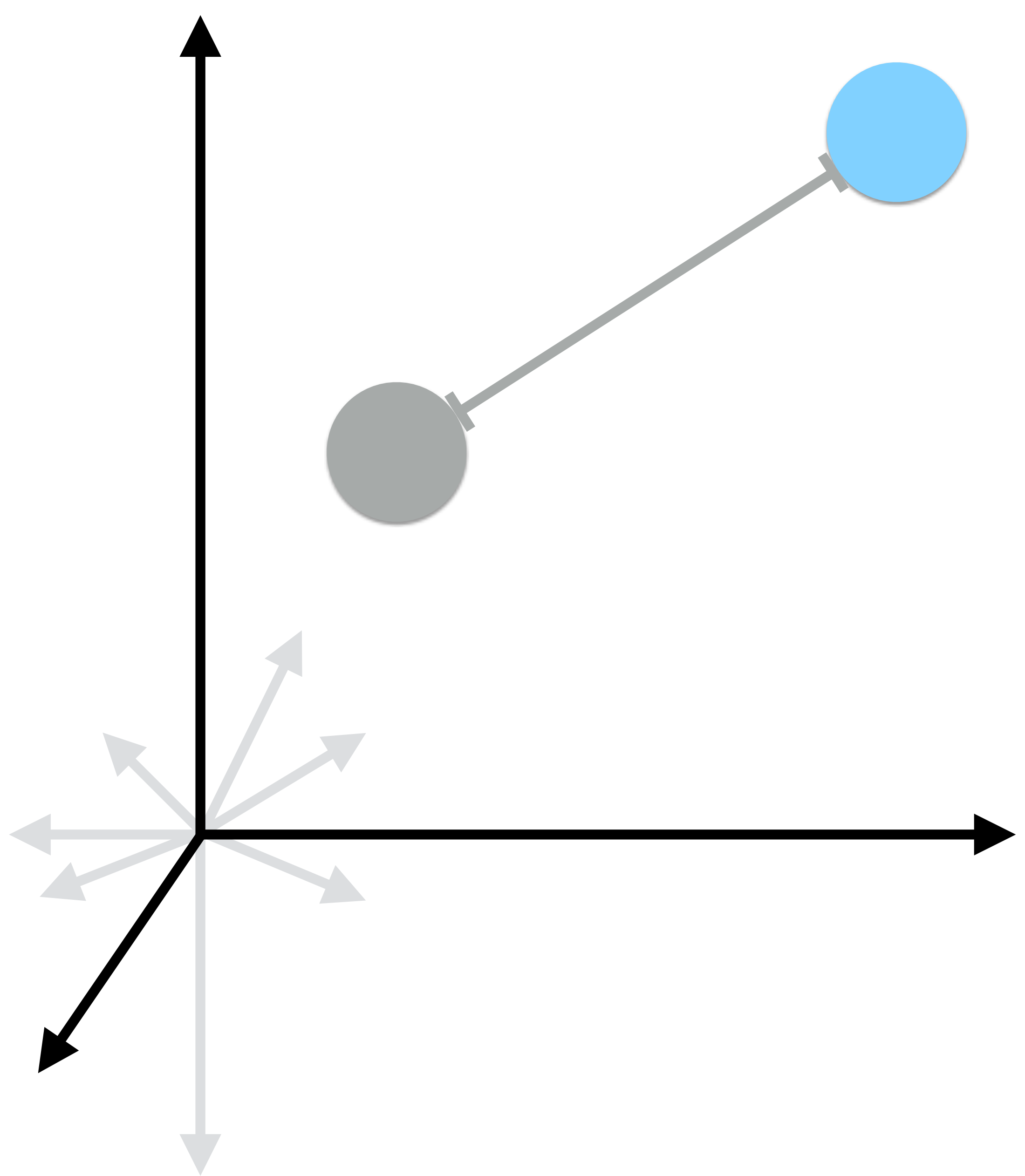


we are going to have two networks or two copies of the same network

contrastive learning



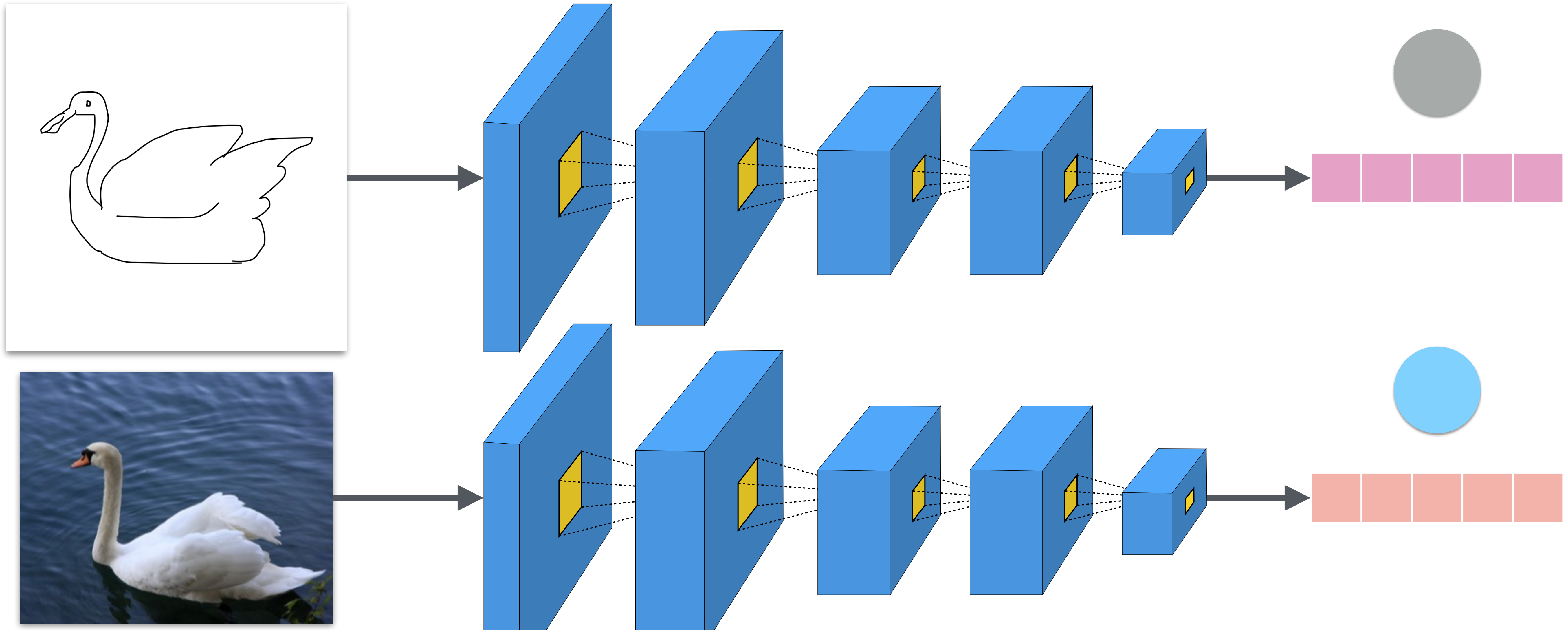
$\downarrow (D(W, x_1, x_2))^2$



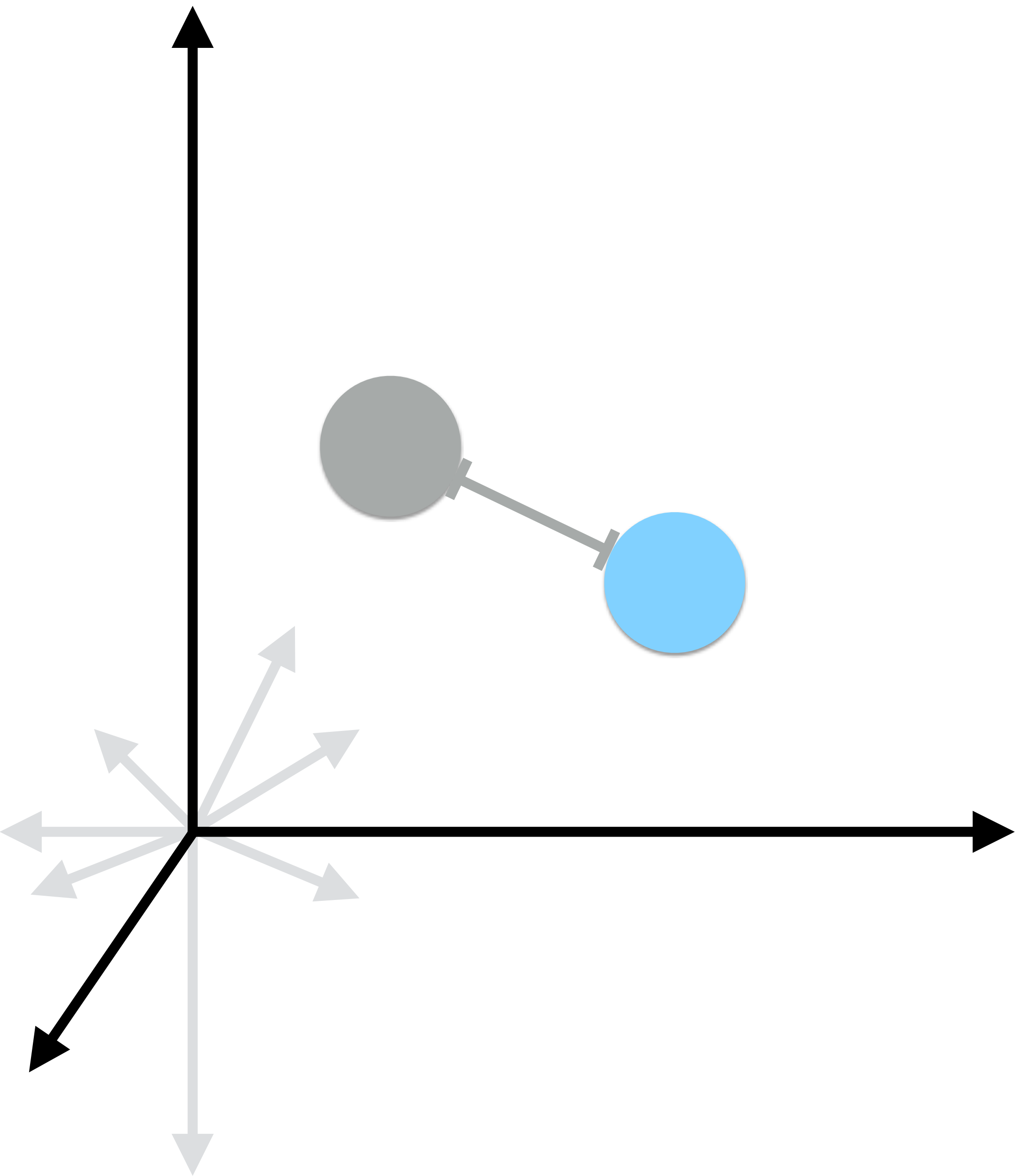
we are going to have two networks or two copies of the same network

for similar pairs, learning will bring the representations closer

contrastive learning



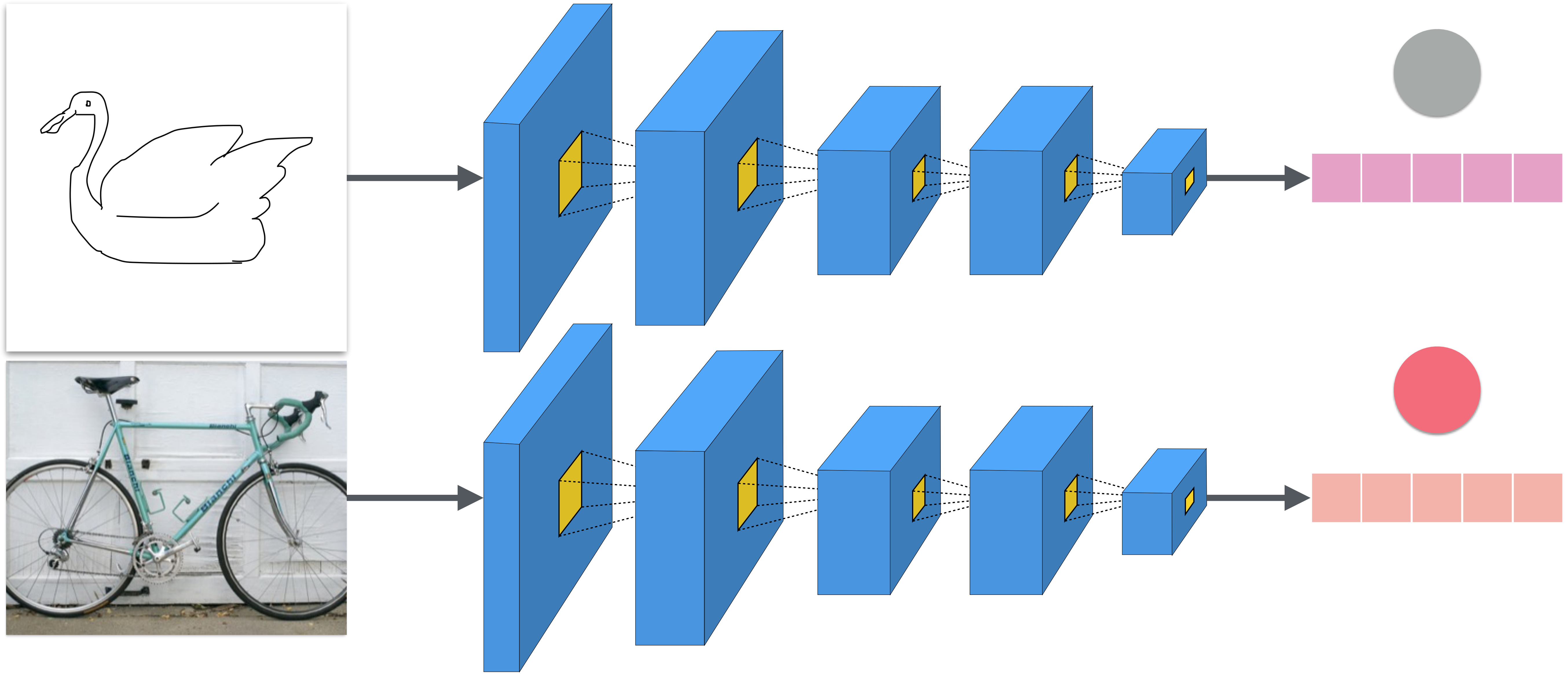
$\downarrow (D(W, x_1, x_2))^2$

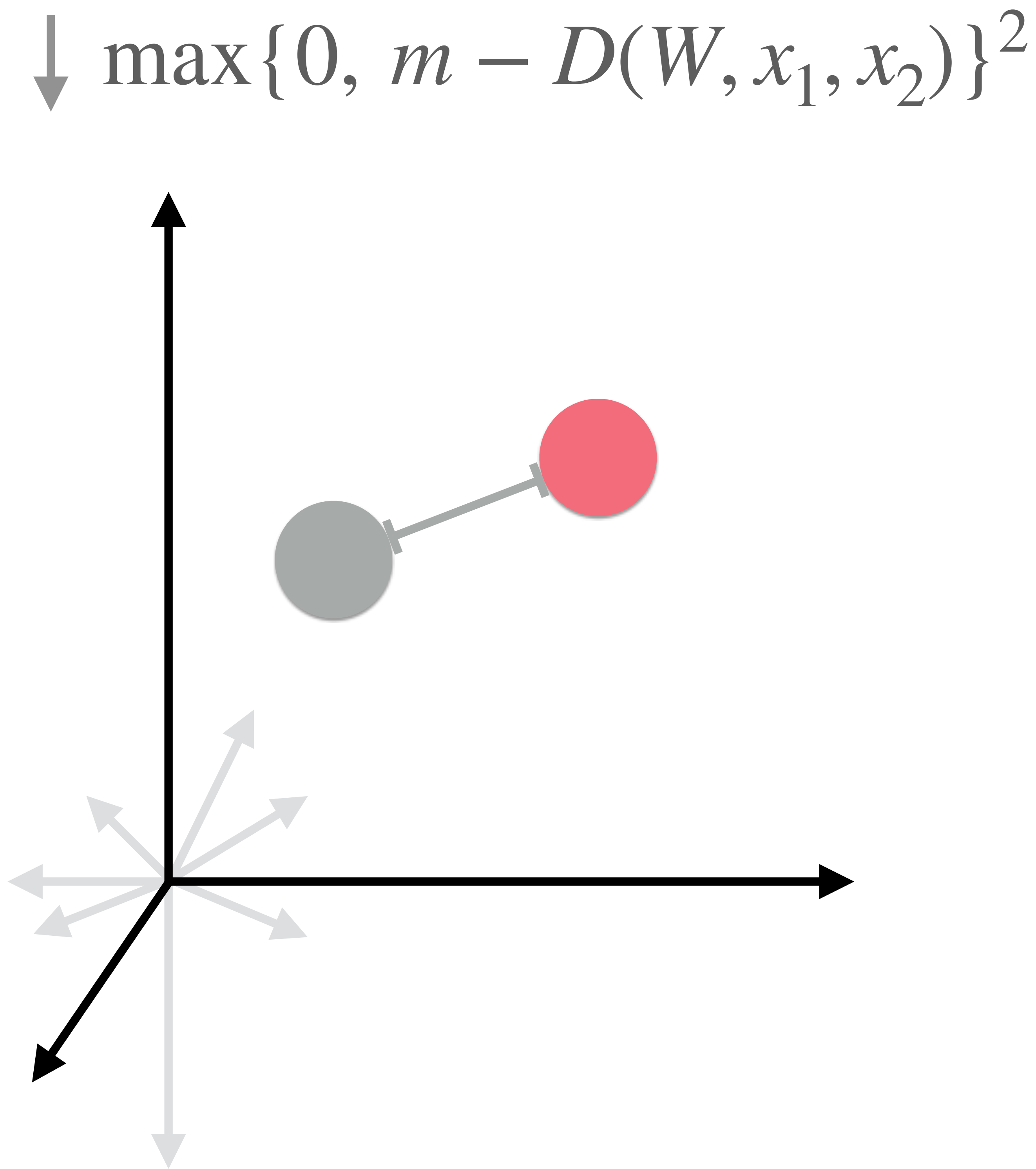


we are going to have two networks or two copies of the same network

for similar pairs, learning will bring the representations closer

contrastive learning

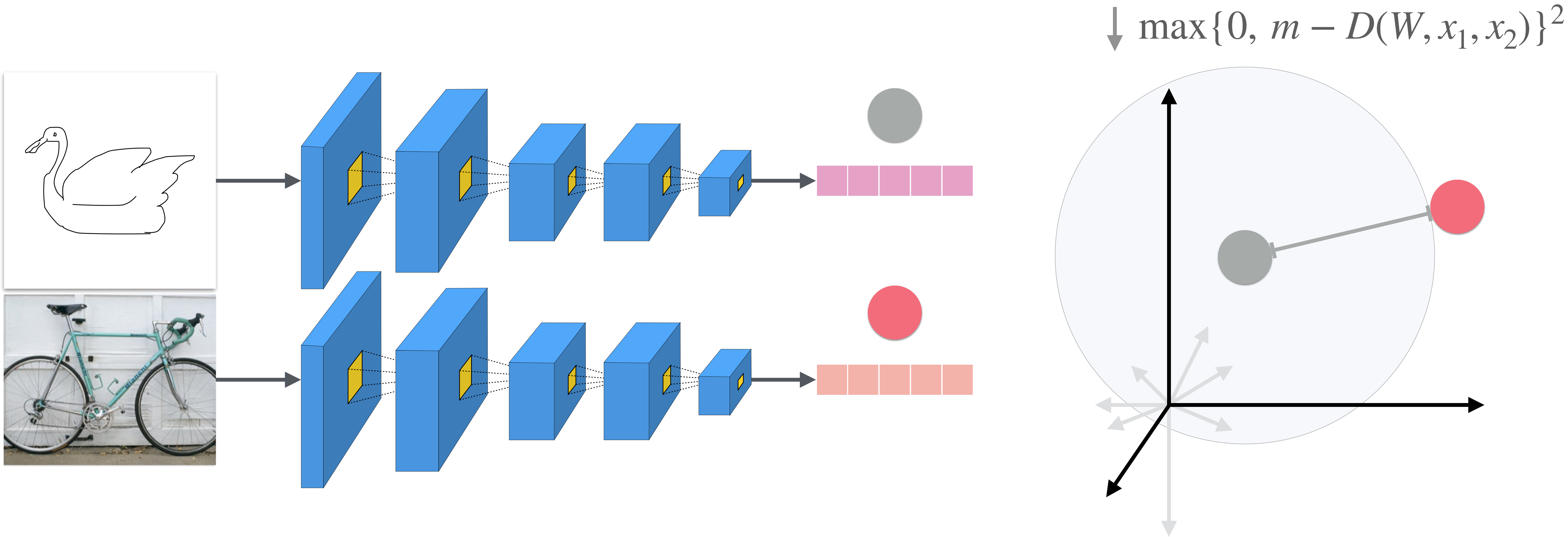




we are going to have two networks or two copies of the same network

for dissimilar pairs, learning will move the representations apart

contrastive learning



we are going to have two networks or two copies of the same network

for dissimilar pairs, learning will move the representations apart

triplet loss

FaceNet: A Unified Embedding for Face Recognition and Clustering

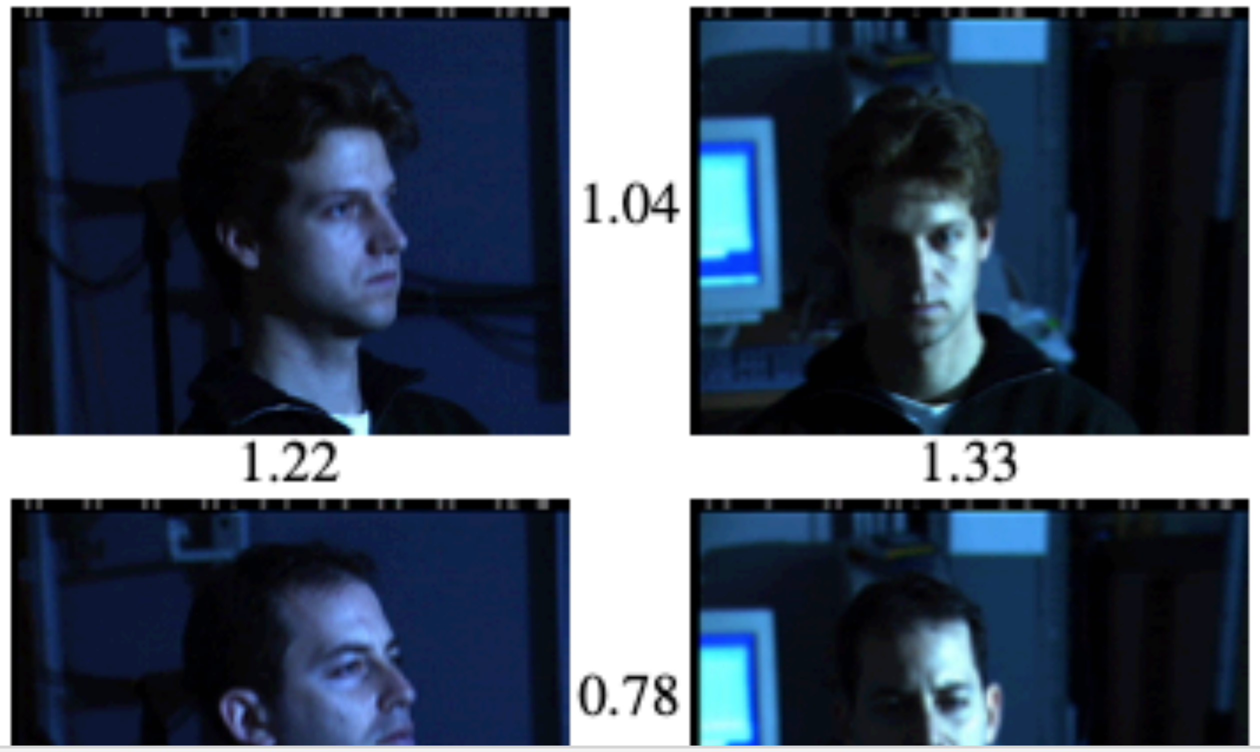
Florian Schroff
fschroff@google.com
Google Inc.

Dmitry Kalenichenko
dkalenichenko@google.com
Google Inc.

James Philbin
jphilbin@google.com
Google Inc.

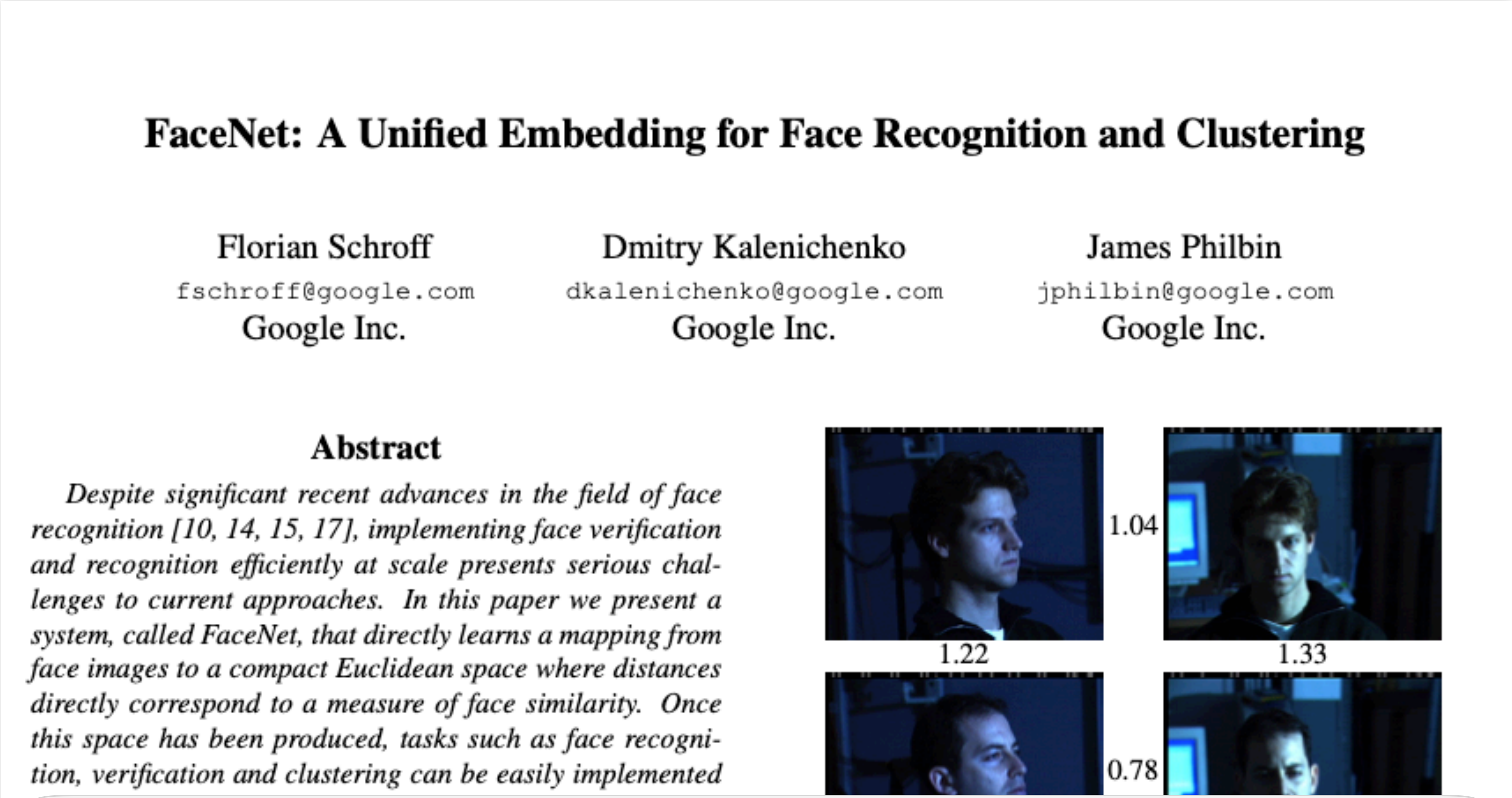
Abstract

Despite significant recent advances in the field of face recognition [10, 14, 15, 17], implementing face verification and recognition efficiently at scale presents serious challenges to current approaches. In this paper we present a system, called FaceNet, that directly learns a mapping from face images to a compact Euclidean space where distances directly correspond to a measure of face similarity. Once this space has been produced, tasks such as face recognition, verification and clustering can be easily implemented



introduced the triplet loss to learn face representations and perform 1-shot facial recognition

triplet loss



FaceNet: A Unified Embedding for Face Recognition and Clustering

Florian Schroff
fschroff@google.com
Google Inc.

Dmitry Kalenichenko
dkalenichenko@google.com
Google Inc.

James Philbin
jphilbin@google.com
Google Inc.

Abstract

Despite significant recent advances in the field of face recognition [10, 14, 15, 17], implementing face verification and recognition efficiently at scale presents serious challenges to current approaches. In this paper we present a system, called FaceNet, that directly learns a mapping from face images to a compact Euclidean space where distances directly correspond to a measure of face similarity. Once this space has been produced, tasks such as face recognition, verification and clustering can be easily implemented

1.04
1.22
1.33
0.78

introduced the triplet loss to learn face representations and perform 1-shot facial recognition

$$\mathscr{L}(x_a, x_p, x_n) = \frac{1}{2}\max\{0,\ m + D(x_a, x_p) - D(x_a, x_n)\}$$

triplet loss

$$\mathscr{L}(x_a, x_p, x_n) = \frac{1}{2}\max\{0,\ m + D(x_a, x_p) - D(x_a, x_n)\}$$

triplet loss

similar to the
anchor

$$\mathscr{L}(x_a, x_p, x_n) = \frac{1}{2}\max\{0,\ m + D(x_a, x_p) - D(x_a, x_n)\}$$

triplet loss

$$\mathscr{L}(x_a, x_p, x_n) = \frac{1}{2}\max\{0,\ m + D(x_a, x_p) - D(x_a, x_n)\}$$

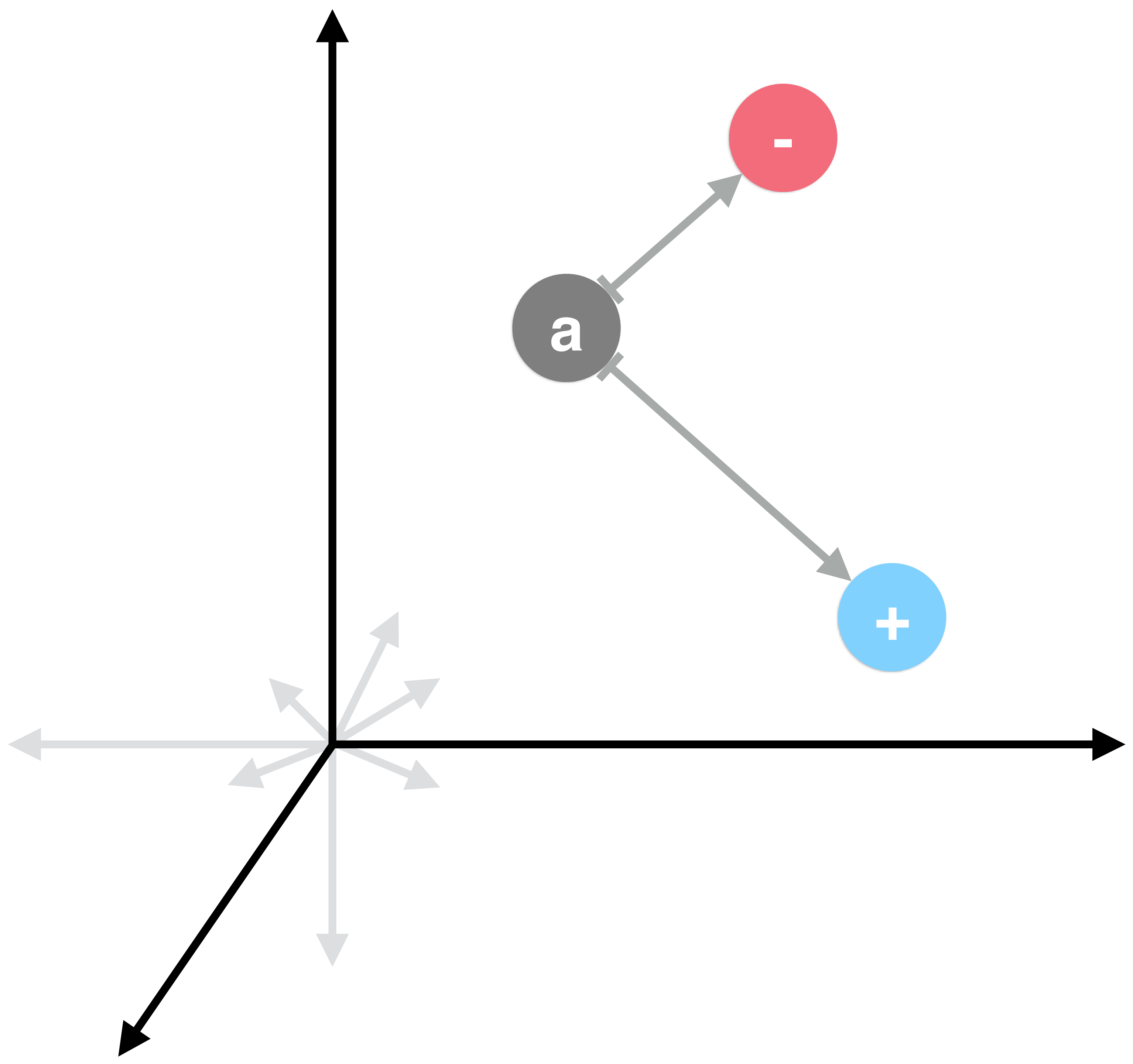
dissimilar to the anchor

triplet loss

$$\mathcal{L}(x_a, x_p, x_n) = \frac{1}{2}\max\{0, m + D(x_a, x_p) - D(x_a, x_n)\}$$

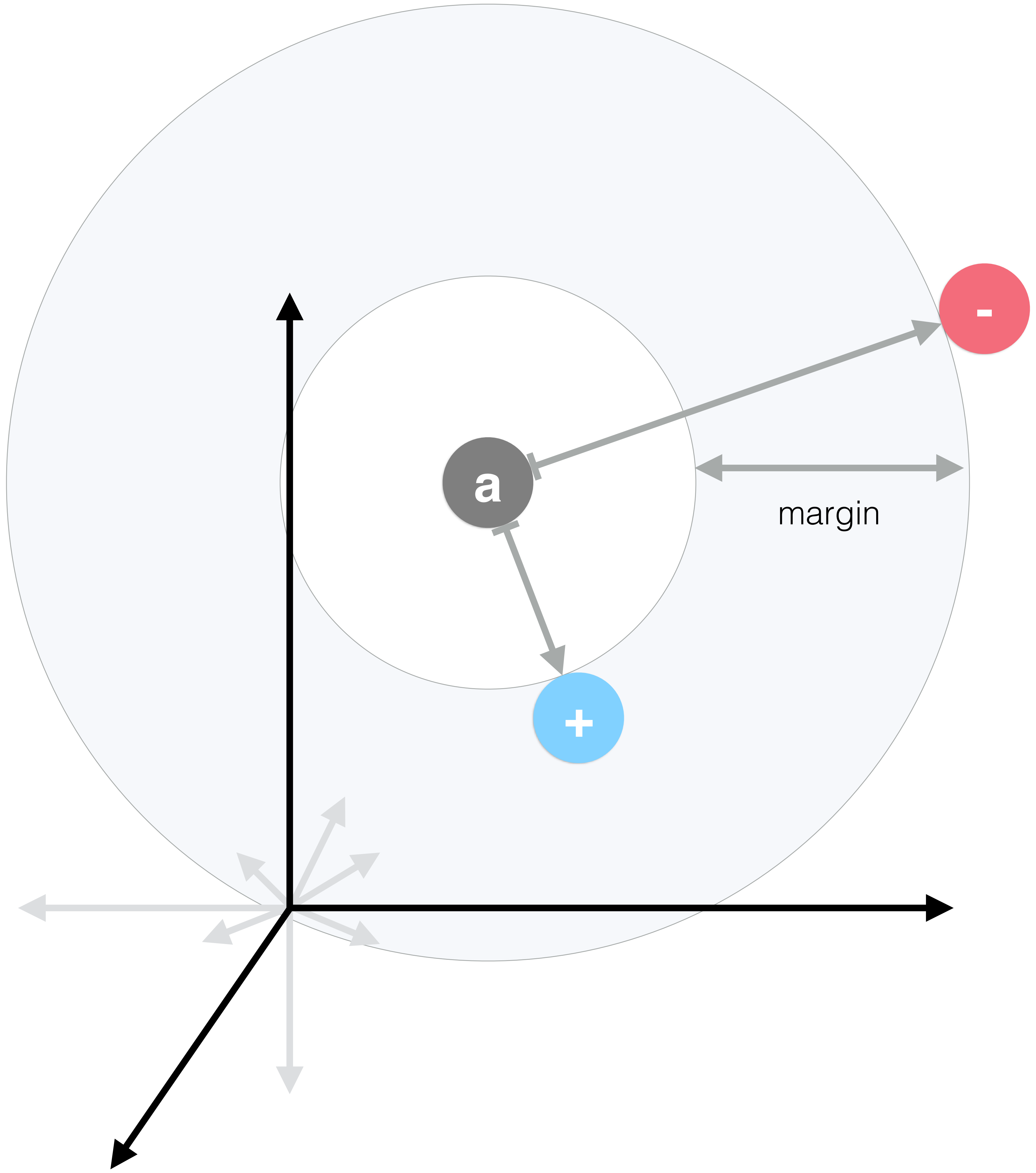
the margin now limits the distance of distances

triplet loss



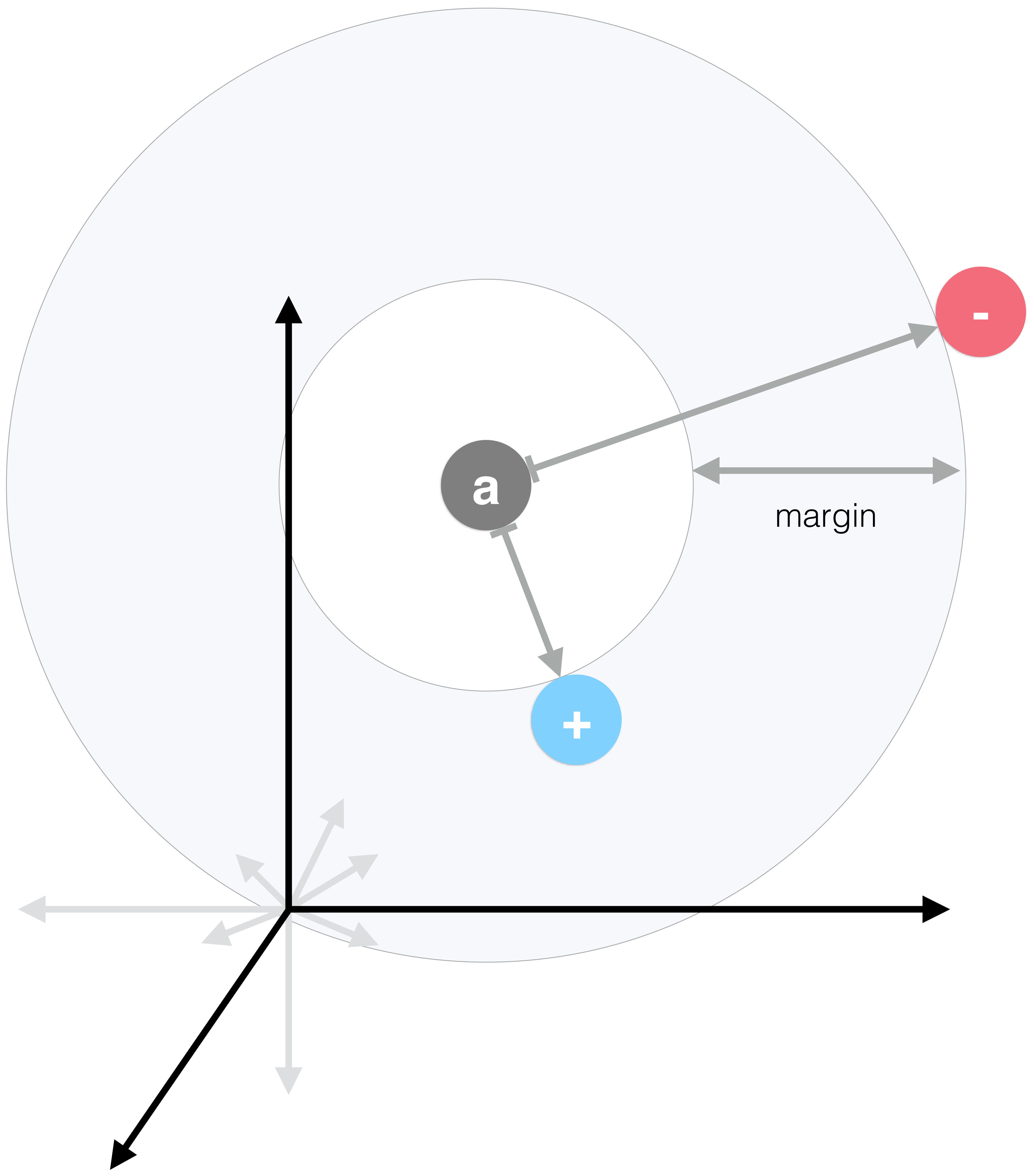
$$\mathscr{L}(x_a, x_p, x_n) = \frac{1}{2}\max\{0,\ m + D(x_a, x_p) - D(x_a, x_n)\}$$

triplet loss



$$\mathcal{L}(x_a, x_p, x_n) = \frac{1}{2}\max\{0,\ m + D(x_a, x_p) - D(x_a, x_n)\}$$

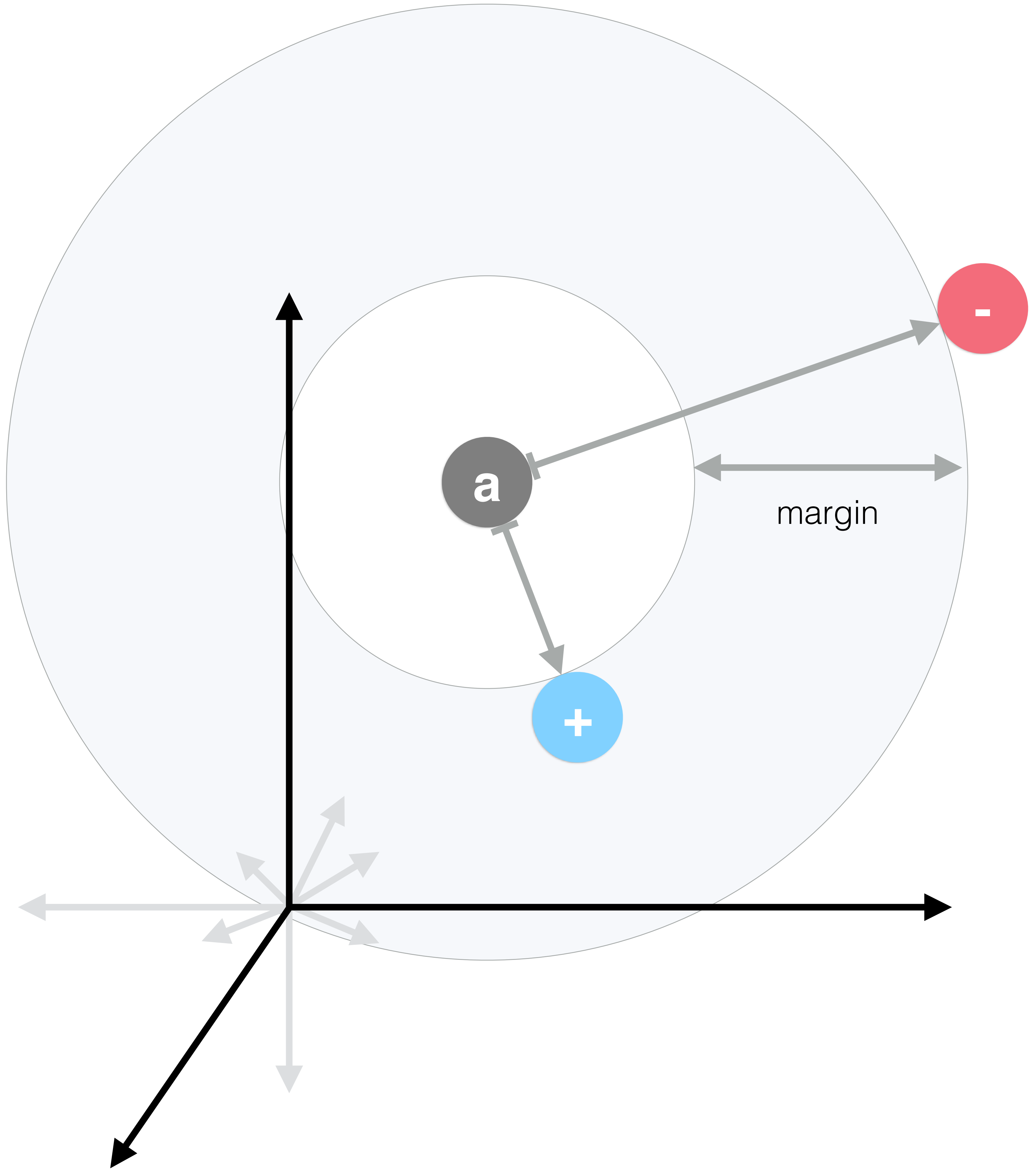
triplet loss



$$\mathcal{L}(x_a, x_p, x_n) = \frac{1}{2}\max\{0,\ m + D(x_a, x_p) - D(x_a, x_n)\}$$

notice how the margin is between the distances

triplet loss

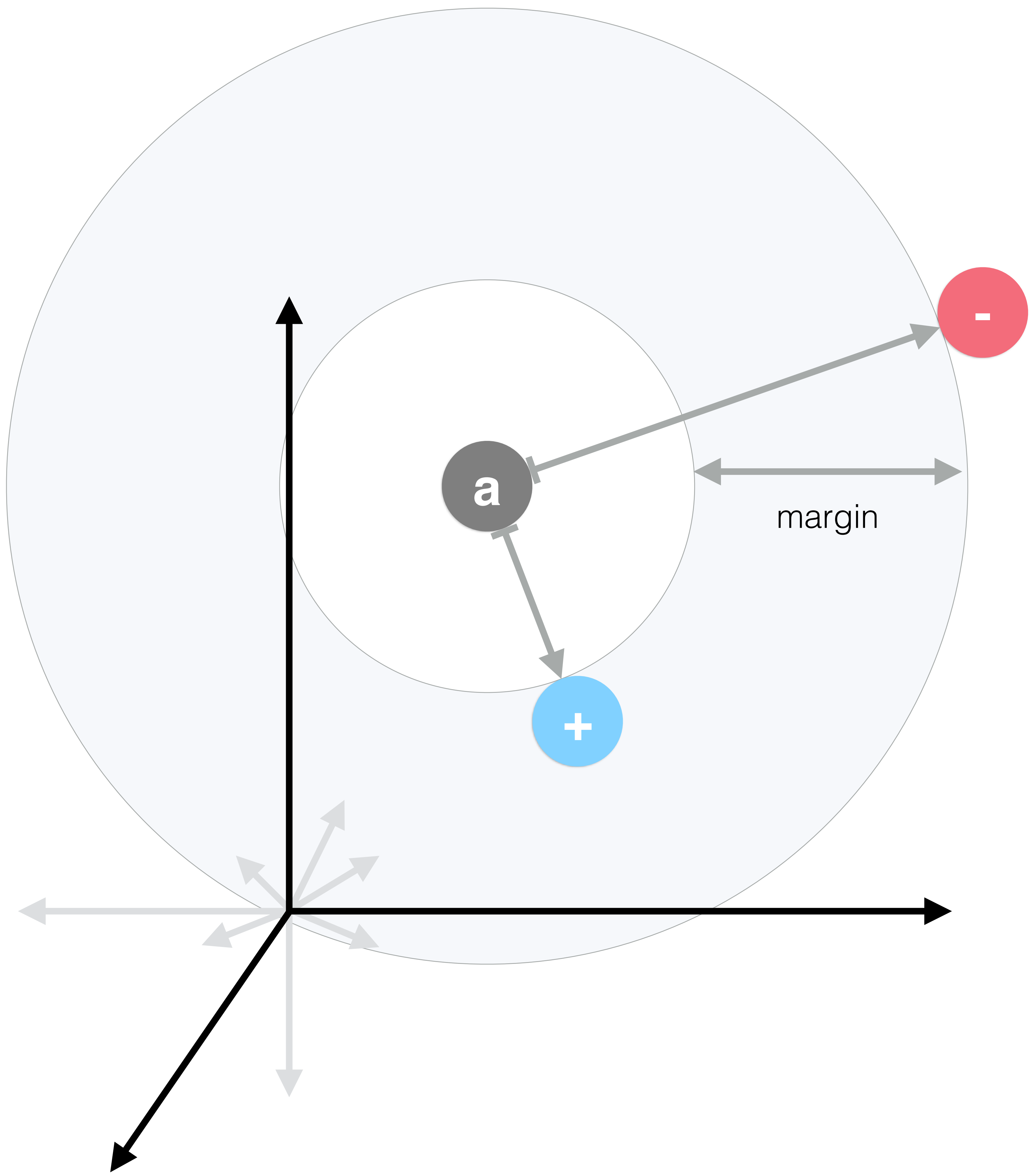


$$\mathcal{L}(x_a, x_p, x_n) = \frac{1}{2}\max\{0,\ m + D(x_a, x_p) - D(x_a, x_n)\}$$

notice how the margin is between the distances

avoiding the issue of representing similar objects with the same vector

triplet loss



$$\mathcal{L}(x_a, x_p, x_n) = \frac{1}{2}\max\{0,\ m + D(x_a, x_p) - D(x_a, x_n)\}$$

notice how the margin is between the distances

avoiding the issue of representing similar objects with the same vector

information noise-contrastive estimation

Representation Learning with Contrastive Predictive Coding

Aaron van den Oord
DeepMind
avdnoord@google.com

Yazhe Li
DeepMind
yazhe@google.com

Oriol Vinyals
DeepMind
vinyals@google.com

cs.LG] 22 Jan 2019

Abstract

While supervised learning has enabled great progress in many applications, unsupervised learning has not seen such widespread adoption, and remains an important and challenging endeavor for artificial intelligence. In this work, we propose a universal unsupervised learning approach to extract useful representations from high-dimensional data, which we call Contrastive Predictive Coding. The key insight of our model is to learn such representations by predicting the future in *latent* space by using powerful autoregressive models. We use a probabilistic contrastive loss which induces the latent space to capture information that is maximally useful to predict future samples. It also makes the model tractable by using negative sampling. While most prior work has focused on evaluating representations for

$$\mathscr{L}(z_i) = -\log \frac{\exp(\mathbf{z}_i \cdot \mathbf{z}_j / \tau)}{\sum_{k=1}^{N} \mathbf{1}_{[k \neq i]} \exp(\mathbf{z}_i \cdot \mathbf{z}_k / \tau)}$$

introduces InfoNCE specifically to learn general representations

information noise-contrastive estimation

$$\mathcal{L}(z_i, z_j) = -\log \frac{\exp(\mathbf{z}_i \cdot \mathbf{z}_j / \tau)}{\sum_{k=1}^{N} \mathbf{1}_{[k \neq i]} \exp(\mathbf{z}_i \cdot \mathbf{z}_k / \tau)}$$

information noise-contrastive estimation

our anchor

$$\mathcal{L}(z_i, z_j) = -\log \frac{\exp(\mathbf{z}_i \cdot \mathbf{z}_j / \tau)}{\sum_{k=1}^{N} \mathbf{1}_{[k \neq i]} \exp(\mathbf{z}_i \cdot \mathbf{z}_k / \tau)}$$

information noise-contrastive estimation

our positive sample

$$\mathcal{L}(z_i, z_j) = -\log \frac{\exp(\mathbf{z}_i \cdot \mathbf{z}_j / \tau)}{\sum_{k=1}^{N} \mathbf{1}_{[k \neq i]} \exp(\mathbf{z}_i \cdot \mathbf{z}_k / \tau)}$$

information noise-contrastive estimation

distance comparison

$$\mathcal{L}(z_i, z_j) = -\log \frac{\exp(\mathbf{z}_i \cdot \mathbf{z}_j/\tau)}{\sum_{k=1}^{N} \mathbf{1}_{[k \neq i]} \exp(\mathbf{z}_i \cdot \mathbf{z}_k/\tau)}$$

information noise-contrastive estimation

normalized by distance from anchor to all samples

$$\mathcal{L}(z_i, z_j) = -\log \frac{\exp(\mathbf{z}_i \cdot \mathbf{z}_j/\tau)}{\sum_{k=1}^{N} \mathbf{1}_{[k \neq i]} \exp(\mathbf{z}_i \cdot \mathbf{z}_k/\tau)}$$

information noise-contrastive estimation

normalized using
softmax

$$\mathcal{L}(z_i, z_j) = -\log \frac{\exp(\mathbf{z}_i \cdot \mathbf{z}_j / \tau)}{\sum_{k=1}^{N} \mathbf{1}_{[k \neq i]} \exp(\mathbf{z}_i \cdot \mathbf{z}_k / \tau)}$$

information noise-contrastive estimation

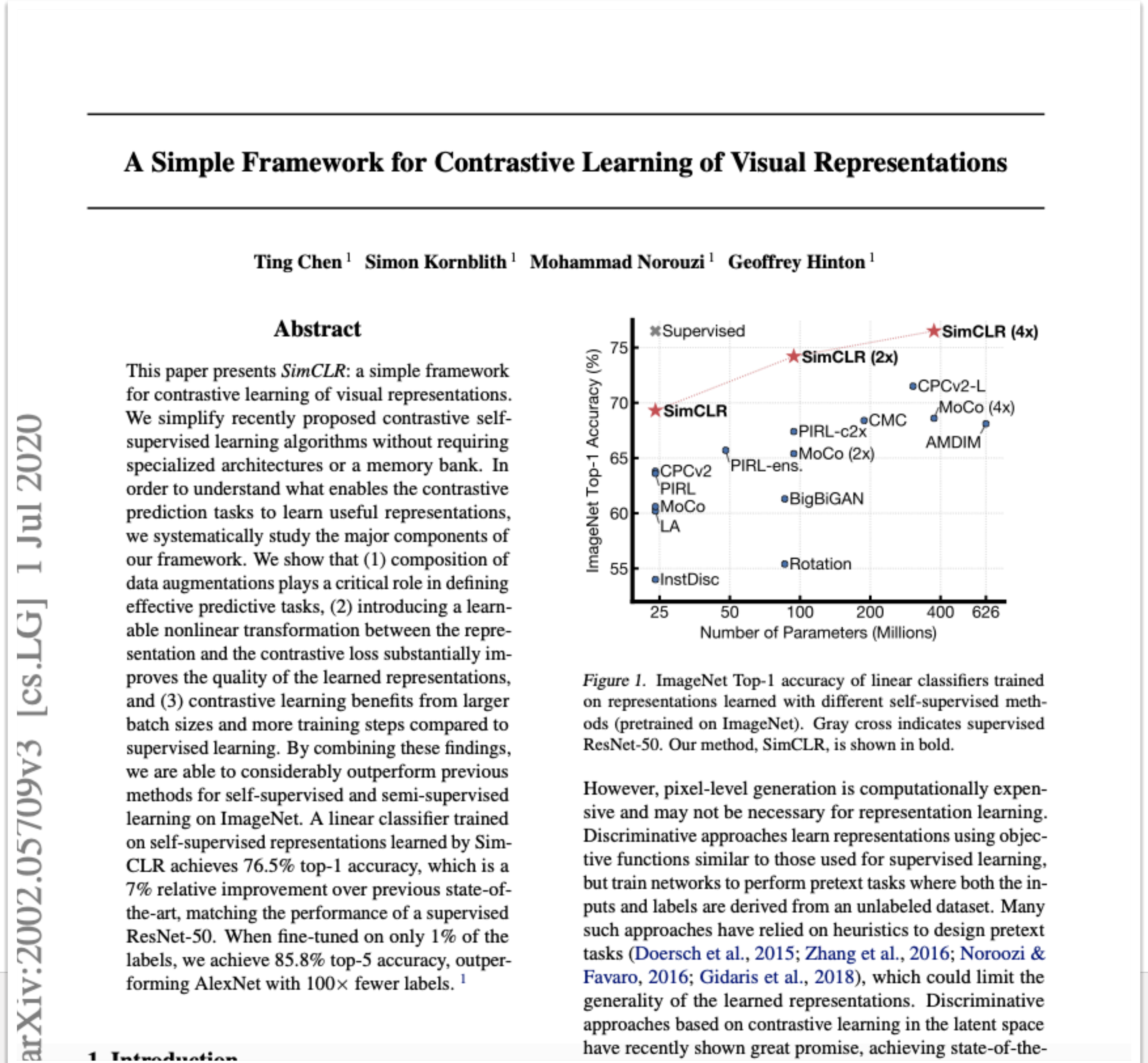
designed as cross-entropy loss

$$\mathcal{L}(z_i, z_j) = -\log \frac{\exp(\mathbf{z}_i \cdot \mathbf{z}_j / \tau)}{\sum_{k=1}^{N} \mathbf{1}_{[k \neq i]} \exp(\mathbf{z}_i \cdot \mathbf{z}_k / \tau)}$$

information noise-contrastive estimation

temperature controls the spread of representations

$$\mathscr{L}(z_i, z_j) = -\log \frac{\exp(\mathbf{z}_i \cdot \mathbf{z}_j / \tau)}{\sum_{k=1}^{N} \mathbf{1}_{[k \neq i]} \exp(\mathbf{z}_i \cdot \mathbf{z}_k / \tau)}$$



A Simple Framework for Contrastive Learning of Visual Representations

Ting Chen [1] Simon Kornblith [1] Mohammad Norouzi [1] Geoffrey Hinton [1]

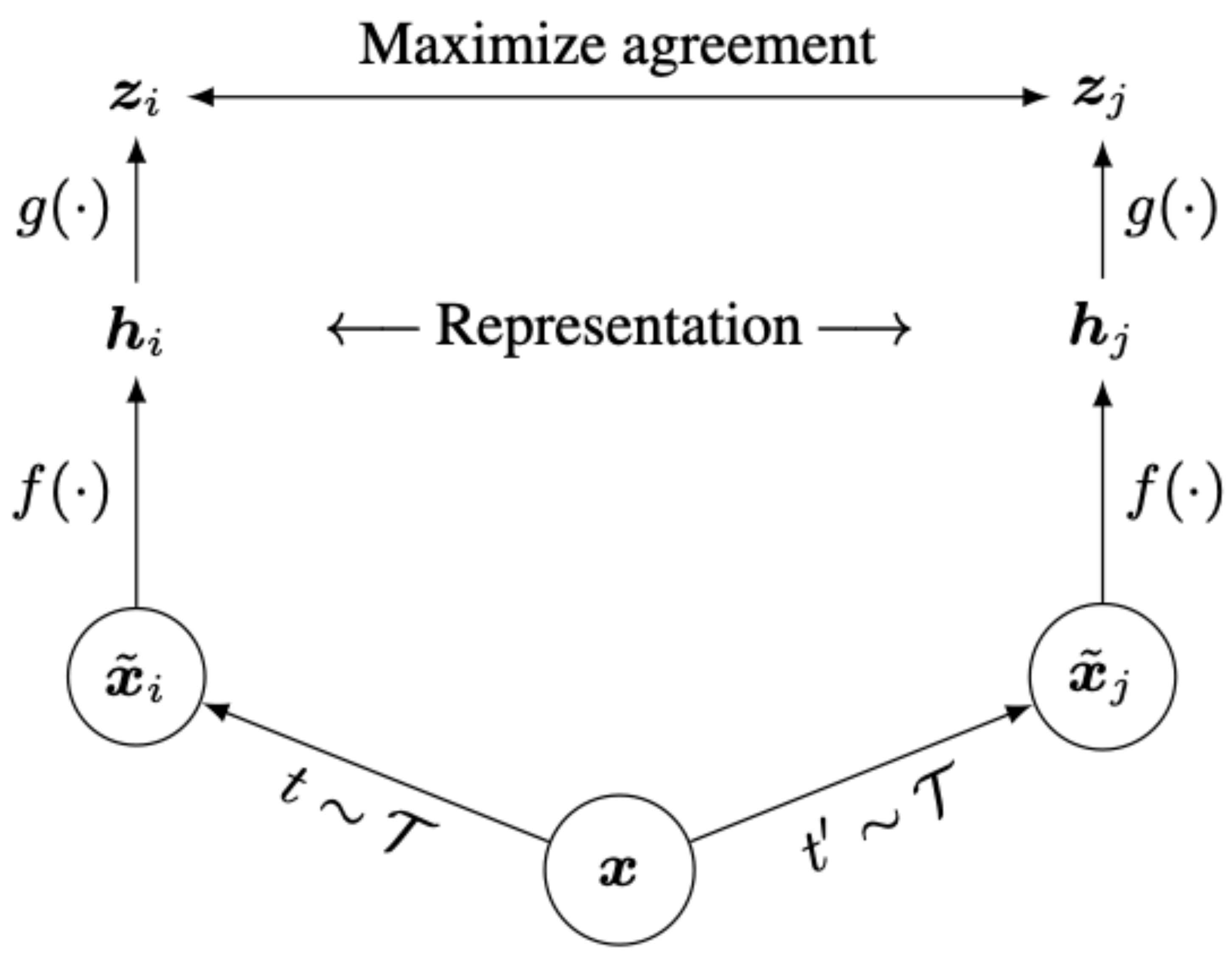
Abstract

This paper presents *SimCLR*: a simple framework for contrastive learning of visual representations. We simplify recently proposed contrastive self-supervised learning algorithms without requiring specialized architectures or a memory bank. In order to understand what enables the contrastive prediction tasks to learn useful representations, we systematically study the major components of our framework. We show that (1) composition of data augmentations plays a critical role in defining effective predictive tasks, (2) introducing a learnable nonlinear transformation between the representation and the contrastive loss substantially improves the quality of the learned representations, and (3) contrastive learning benefits from larger batch sizes and more training steps compared to supervised learning. By combining these findings, we are able to considerably outperform previous methods for self-supervised and semi-supervised learning on ImageNet. A linear classifier trained on self-supervised representations learned by SimCLR achieves 76.5% top-1 accuracy, which is a 7% relative improvement over previous state-of-the-art, matching the performance of a supervised ResNet-50. When fine-tuned on only 1% of the labels, we achieve 85.8% top-5 accuracy, outperforming AlexNet with 100× fewer labels. [1]

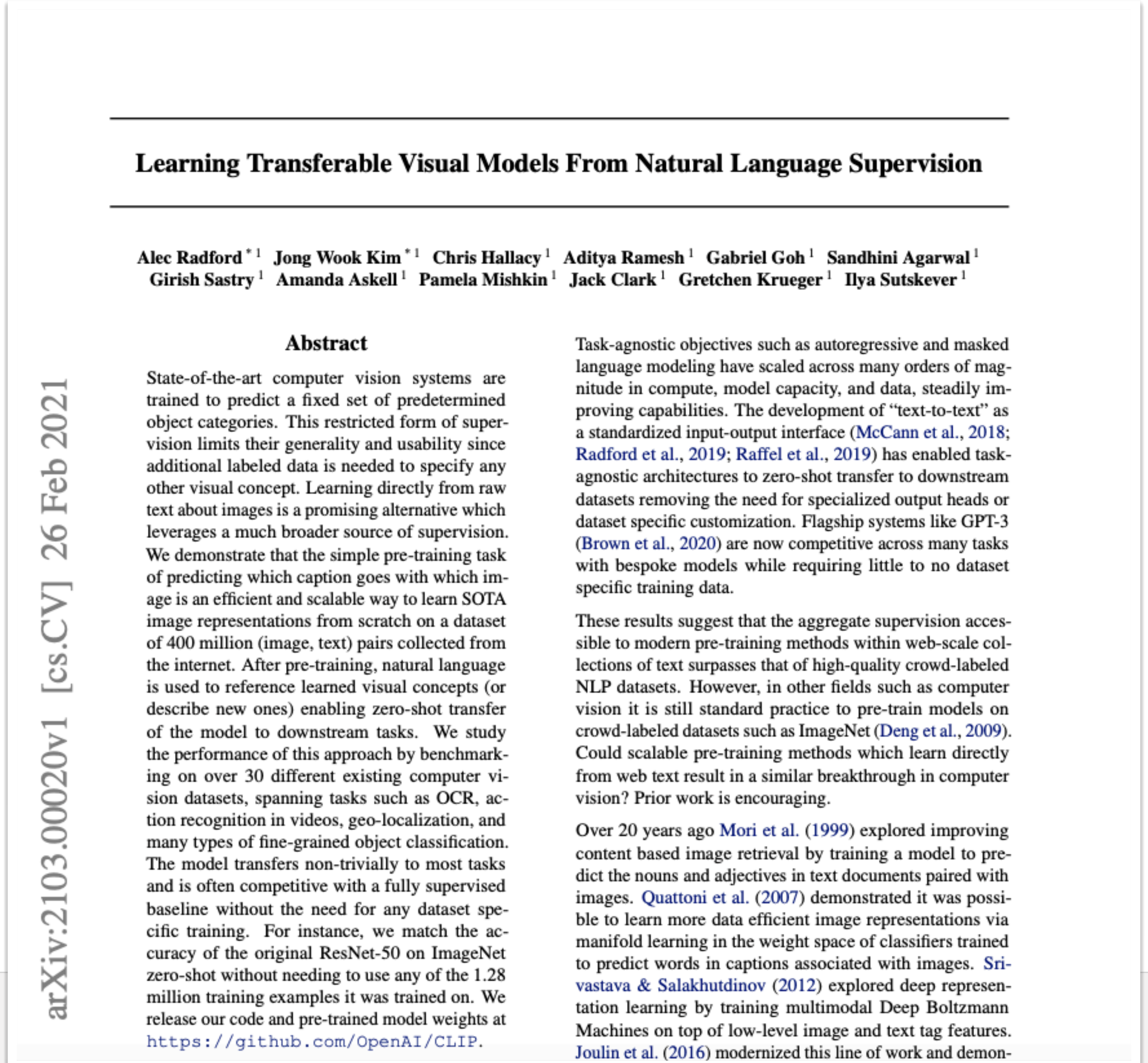
1. Introduction

Figure 1. ImageNet Top-1 accuracy of linear classifiers trained on representations learned with different self-supervised methods (pretrained on ImageNet). Gray cross indicates supervised ResNet-50. Our method, SimCLR, is shown in bold.

However, pixel-level generation is computationally expensive and may not be necessary for representation learning. Discriminative approaches learn representations using objective functions similar to those used for supervised learning, but train networks to perform pretext tasks where both the inputs and labels are derived from an unlabeled dataset. Many such approaches have relied on heuristics to design pretext tasks (Doersch et al., 2015; Zhang et al., 2016; Noroozi & Favaro, 2016; Gidaris et al., 2018), which could limit the generality of the learned representations. Discriminative approaches based on contrastive learning in the latent space have recently shown great promise, achieving state-of-the-



applications



arXiv:2103.00020v1 [cs.CV] 26 Feb 2021

Learning Transferable Visual Models From Natural Language Supervision

Alec Radford [*1] Jong Wook Kim [*1] Chris Hallacy [1] Aditya Ramesh [1] Gabriel Goh [1] Sandhini Agarwal [1] Girish Sastry [1] Amanda Askell [1] Pamela Mishkin [1] Jack Clark [1] Gretchen Krueger [1] Ilya Sutskever [1]

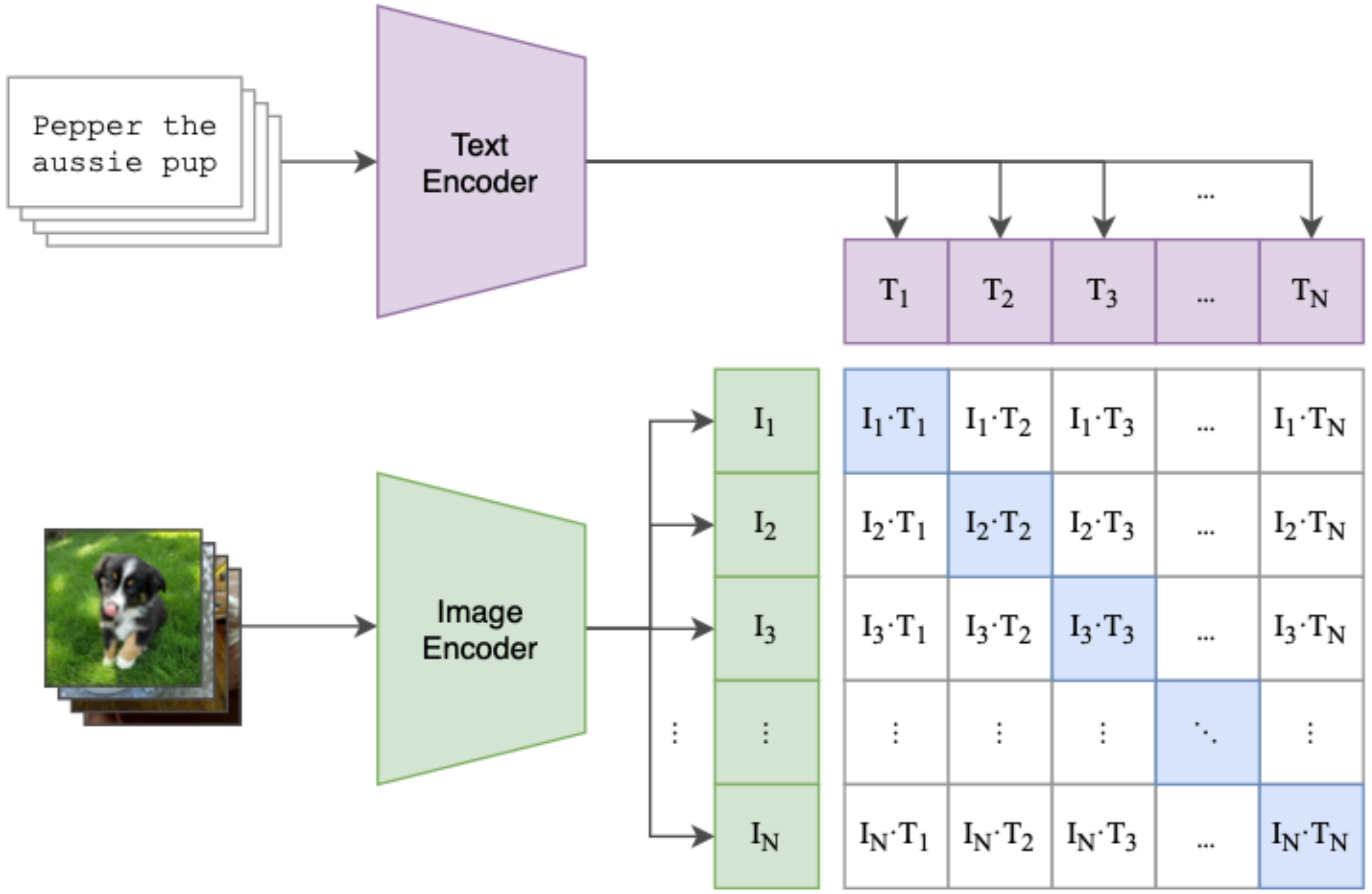
Abstract

State-of-the-art computer vision systems are trained to predict a fixed set of predetermined object categories. This restricted form of supervision limits their generality and usability since additional labeled data is needed to specify any other visual concept. Learning directly from raw text about images is a promising alternative which leverages a much broader source of supervision. We demonstrate that the simple pre-training task of predicting which caption goes with which image is an efficient and scalable way to learn SOTA image representations from scratch on a dataset of 400 million (image, text) pairs collected from the internet. After pre-training, natural language is used to reference learned visual concepts (or describe new ones) enabling zero-shot transfer of the model to downstream tasks. We study the performance of this approach by benchmarking on over 30 different existing computer vision datasets, spanning tasks such as OCR, action recognition in videos, geo-localization, and many types of fine-grained object classification. The model transfers non-trivially to most tasks and is often competitive with a fully supervised baseline without the need for any dataset specific training. For instance, we match the accuracy of the original ResNet-50 on ImageNet zero-shot without needing to use any of the 1.28 million training examples it was trained on. We release our code and pre-trained model weights at https://github.com/OpenAI/CLIP.

Task-agnostic objectives such as autoregressive and masked language modeling have scaled across many orders of magnitude in compute, model capacity, and data, steadily improving capabilities. The development of "text-to-text" as a standardized input-output interface (McCann et al., 2018; Radford et al., 2019; Raffel et al., 2019) has enabled task-agnostic architectures to zero-shot transfer to downstream datasets removing the need for specialized output heads or dataset specific customization. Flagship systems like GPT-3 (Brown et al., 2020) are now competitive across many tasks with bespoke models while requiring little to no dataset specific training data.

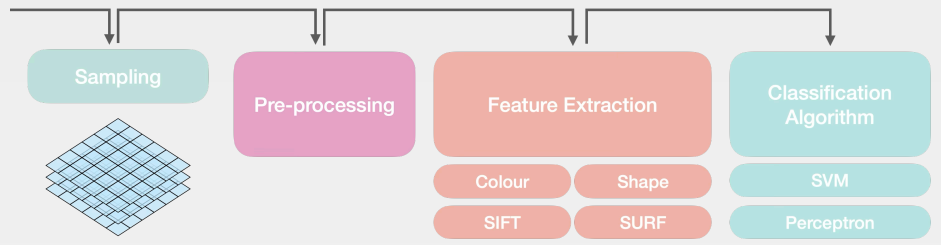
These results suggest that the aggregate supervision accessible to modern pre-training methods within web-scale collections of text surpasses that of high-quality crowd-labeled NLP datasets. However, in other fields such as computer vision it is still standard practice to pre-train models on crowd-labeled datasets such as ImageNet (Deng et al., 2009). Could scalable pre-training methods which learn directly from web text result in a similar breakthrough in computer vision? Prior work is encouraging.

Over 20 years ago Mori et al. (1999) explored improving content based image retrieval by training a model to predict the nouns and adjectives in text documents paired with images. Quattoni et al. (2007) demonstrated it was possible to learn more data efficient image representations via manifold learning in the weight space of classifiers trained to predict words in captions associated with images. Srivastava & Salakhutdinov (2012) explored deep representation learning by training multimodal Deep Boltzmann Machines on top of low-level image and text tag features. Joulin et al. (2016) modernized this line of work and demon-

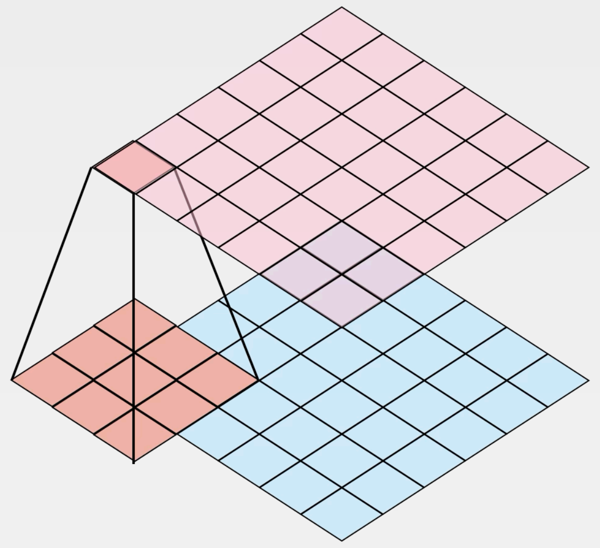


SCC0251
Processamento de Imagens

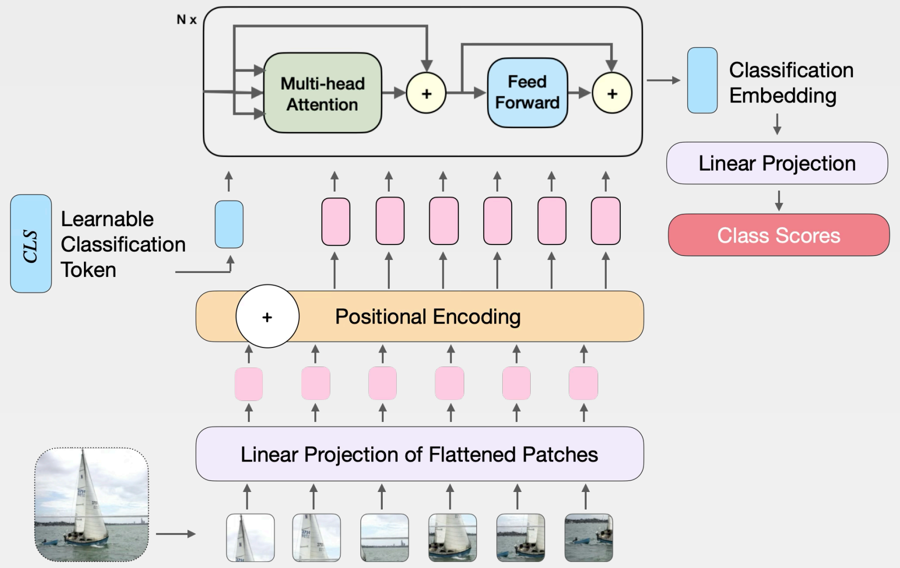
Aprendizado Profundo



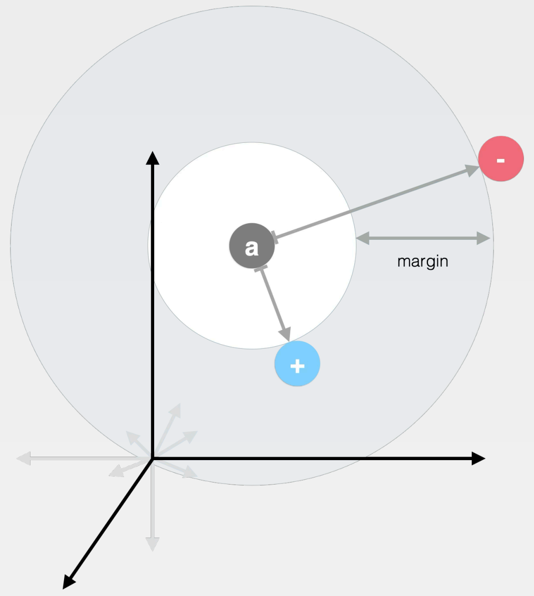
Classic Pipeline



CNNs



Transformers



Contrastive Learning

SCC0251
Processamento de Imagens

Aprendizado Profundo

Professora Leo Sampaio Ferraz Ribeiro