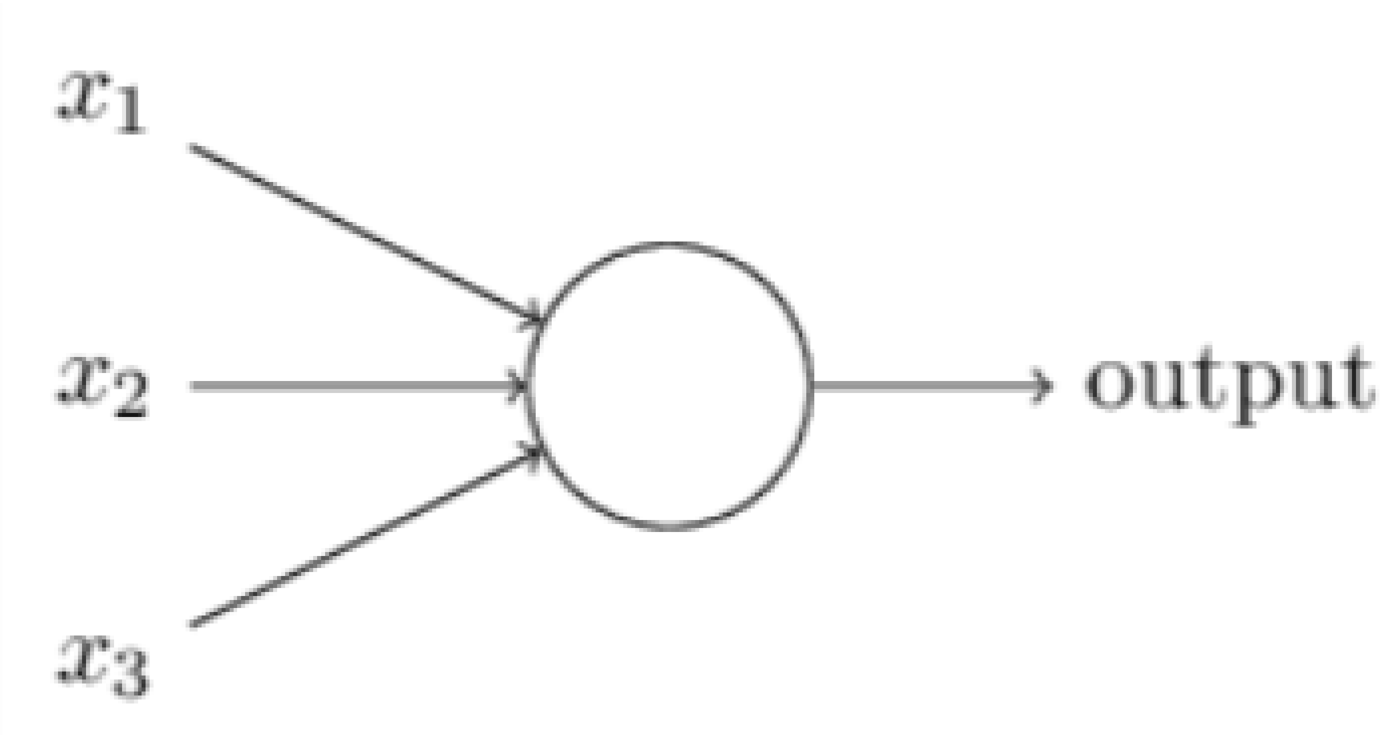
Introduction of Neural Network & Deep Learning

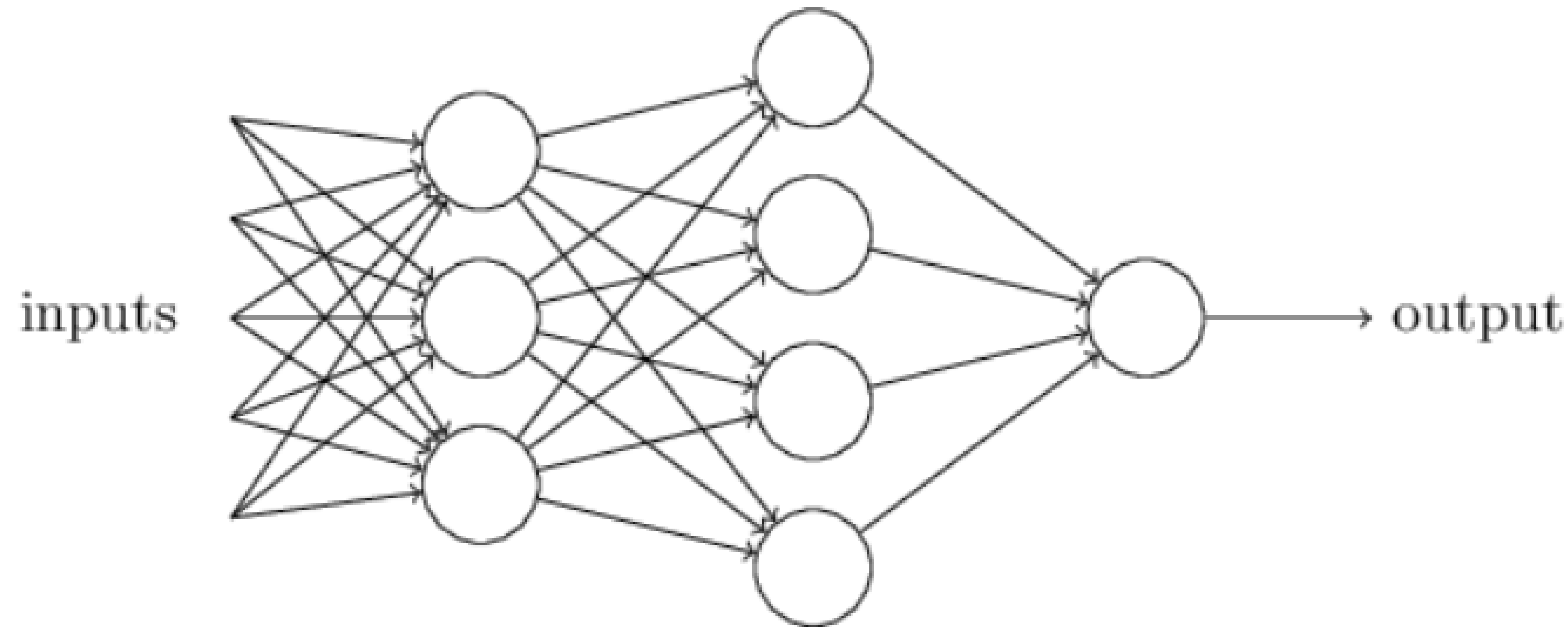
馬誠佑
CAIM, Chang Gung Memorial Hospital, Linkou

2022/08/23

長庚醫療人工智能核心實驗室
Center for Artificial Intelligence in Medicine
Chang Gung Memorial Hospital

Neural Network 深度學習模型是一連串幾何轉換總和

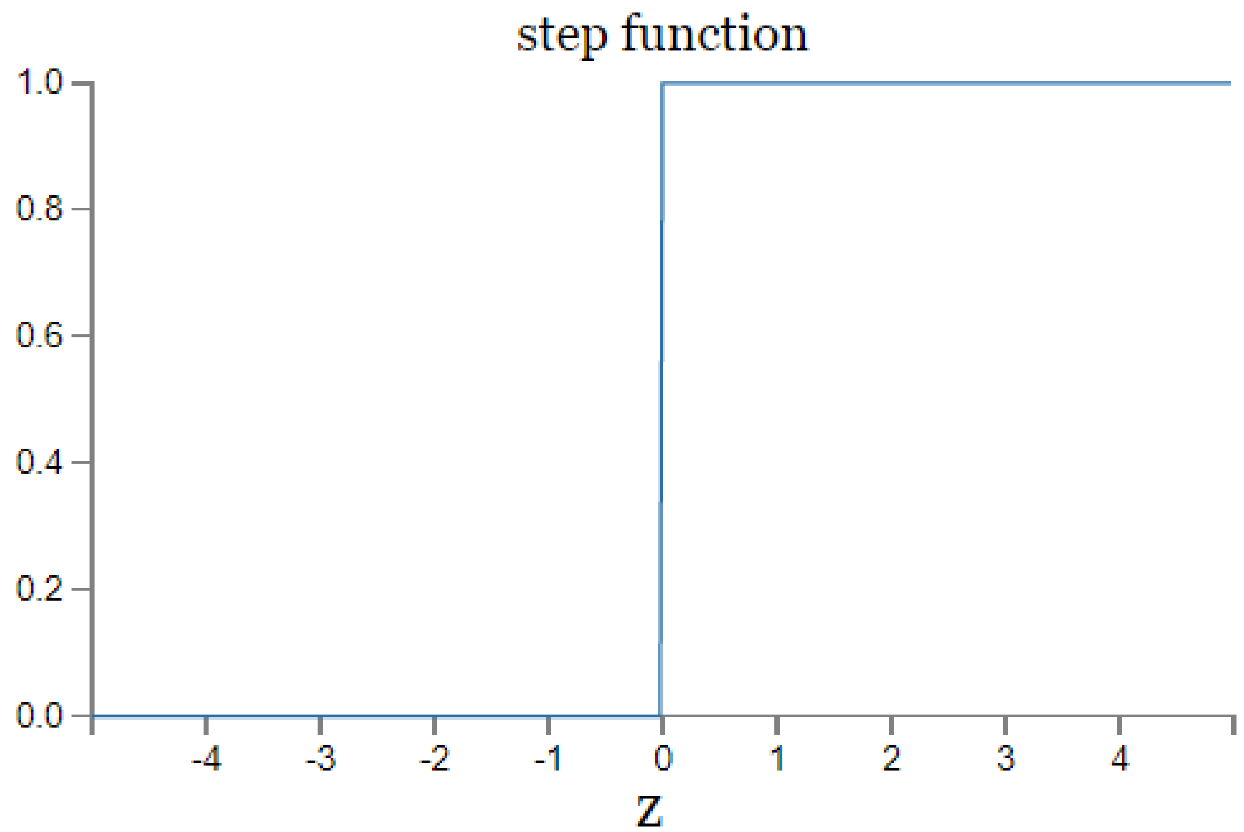


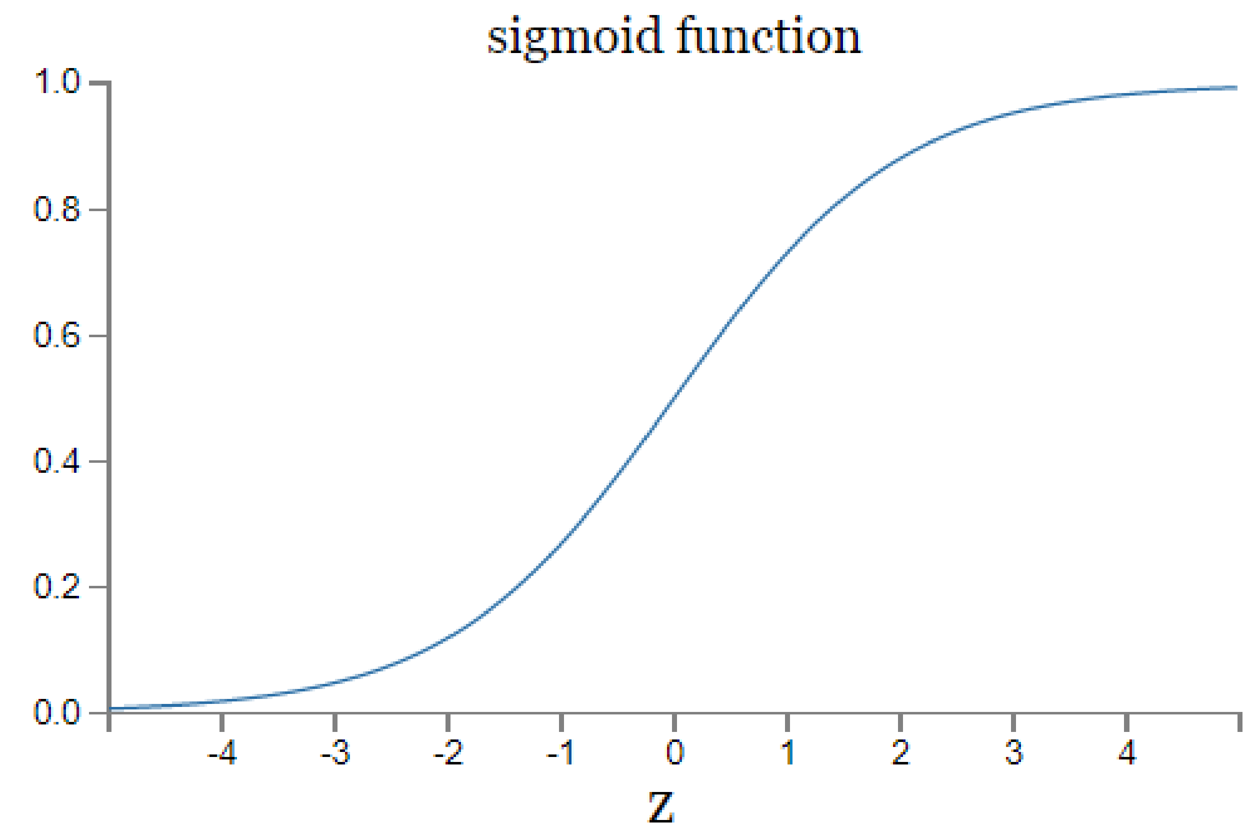


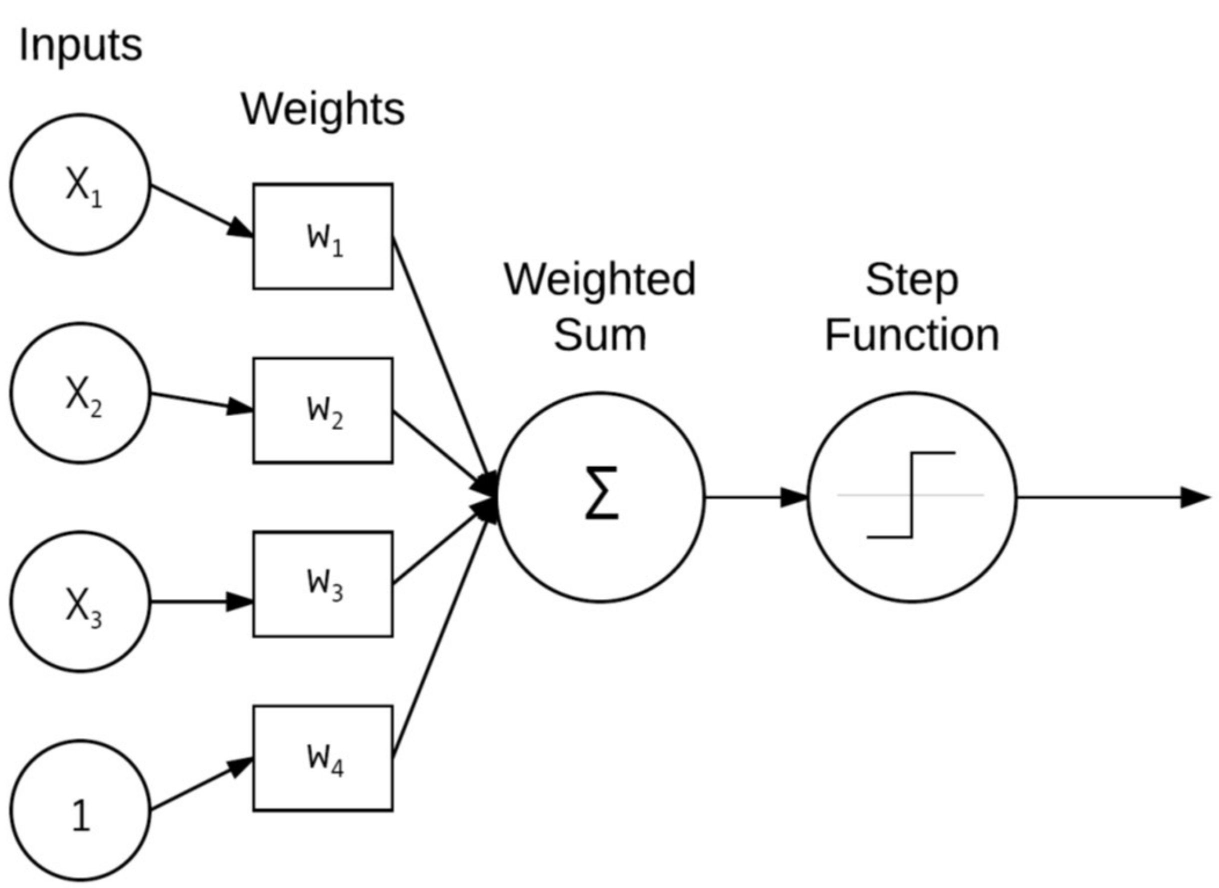
$$\text{output} = \begin{cases} 0 & \text{if } \sum_j w_j x_j \leq \text{ threshold} \\ 1 & \text{if } \sum_j w_j x_j > \text{ threshold} \end{cases}$$

$$\text{output} = \begin{cases} 0 & \text{if } w \cdot x + b \leq 0 \\ 1 & \text{if } w \cdot x + b > 0 \end{cases}$$

Activation Function



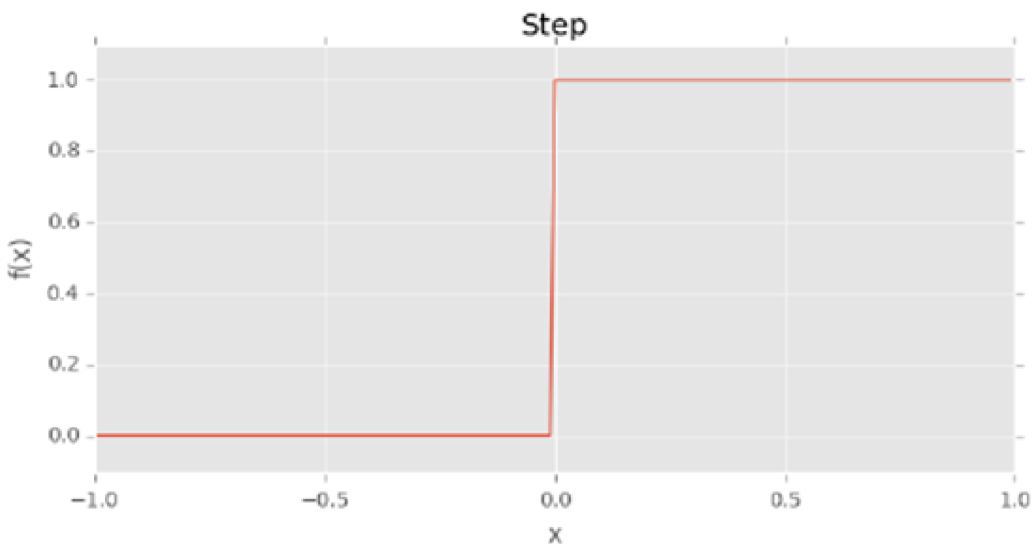


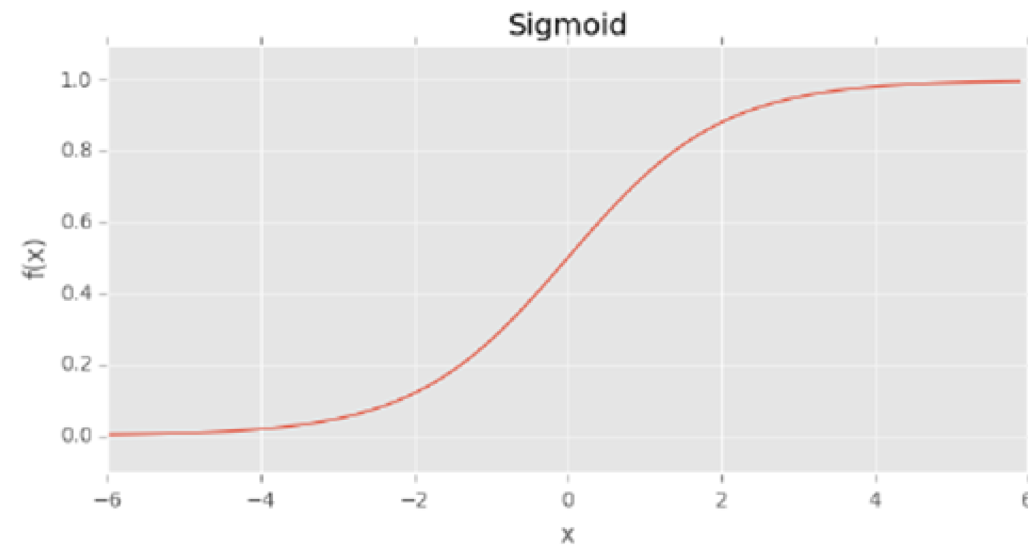


$$\sigma(z) \equiv \frac{1}{1+e^{-z}}.$$

$$\frac{1}{1+\exp(-\sum_j w_j x_j - b)}.$$

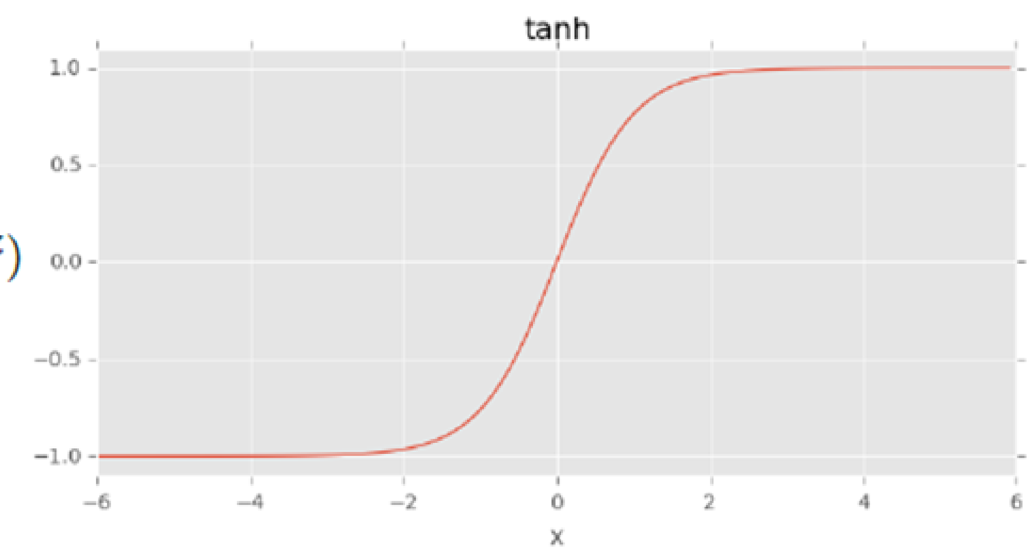
Activation Functions

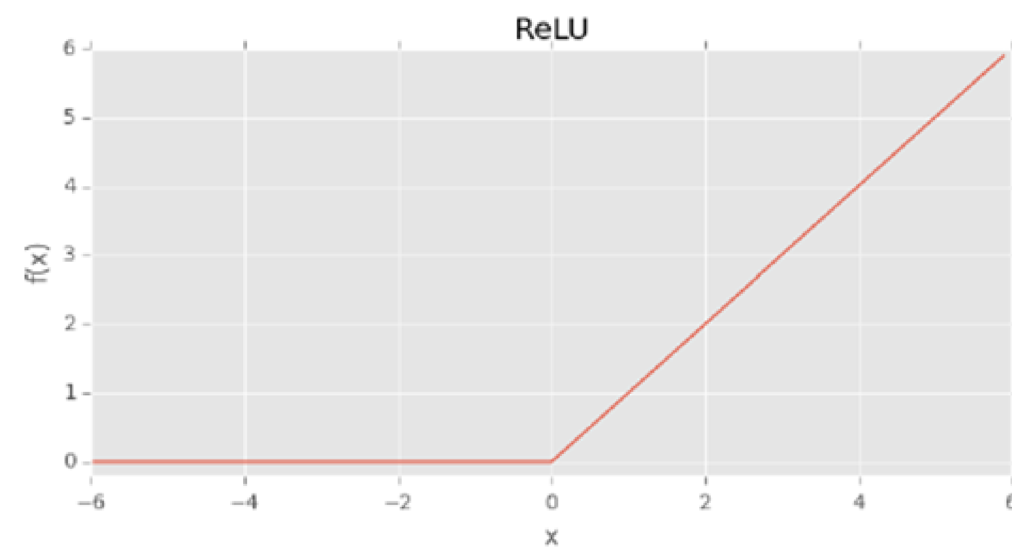




$$s(t) = 1/(1+e^{-t})$$

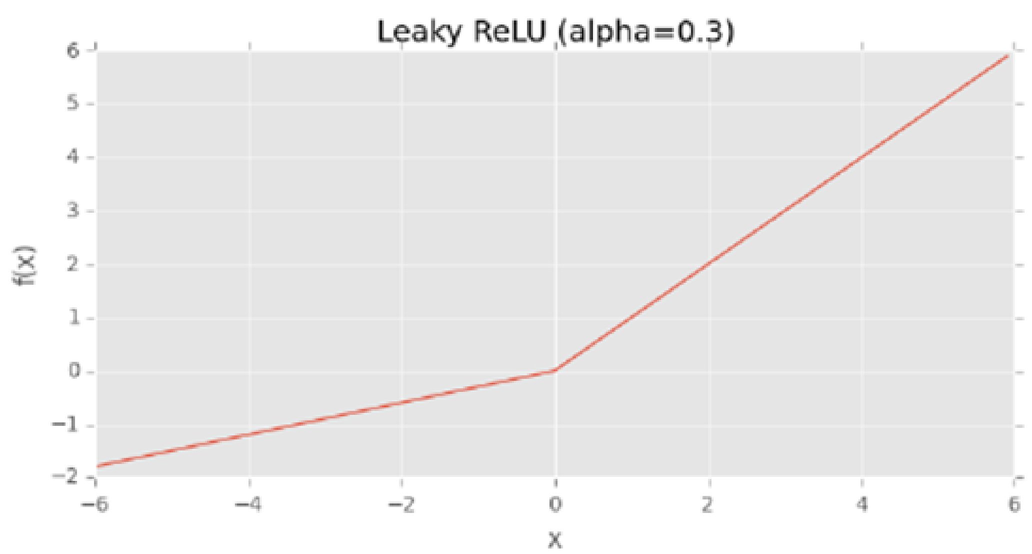
$$f(z) = tanh(z) = (e^{z} - e^{-z})/(e^{z} + e^{-z})$$

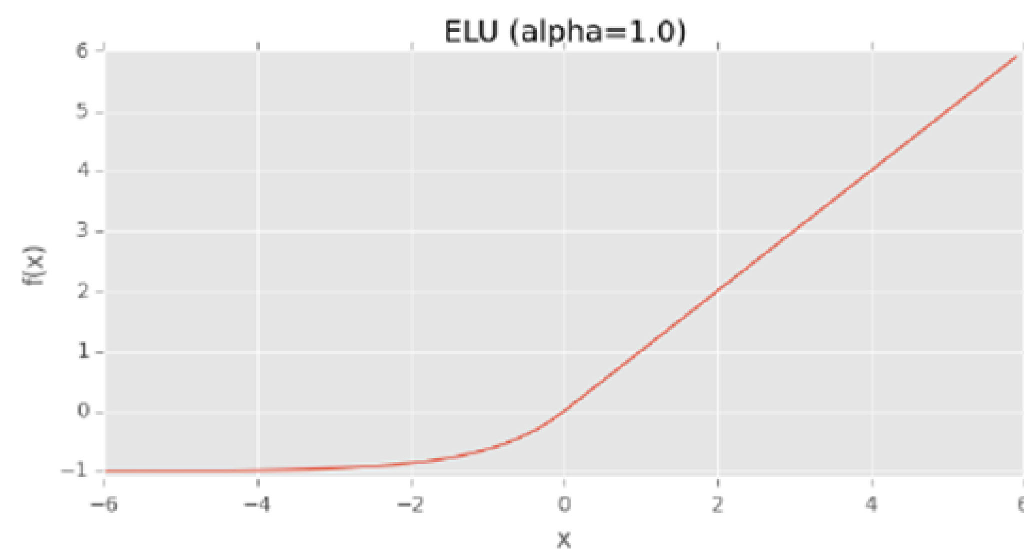




$$f(x) = max(0, x)$$

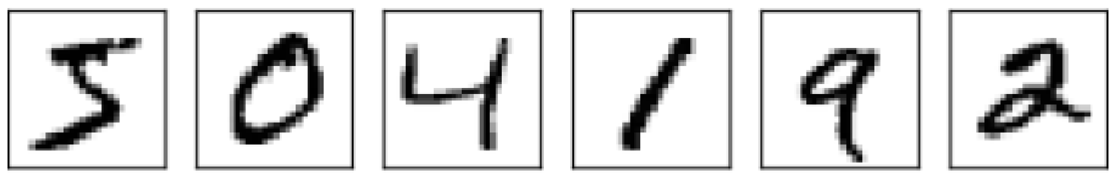
$$f(net) = \begin{cases} net & if net >= 0 \\ \alpha \times net & otherwise \end{cases}$$

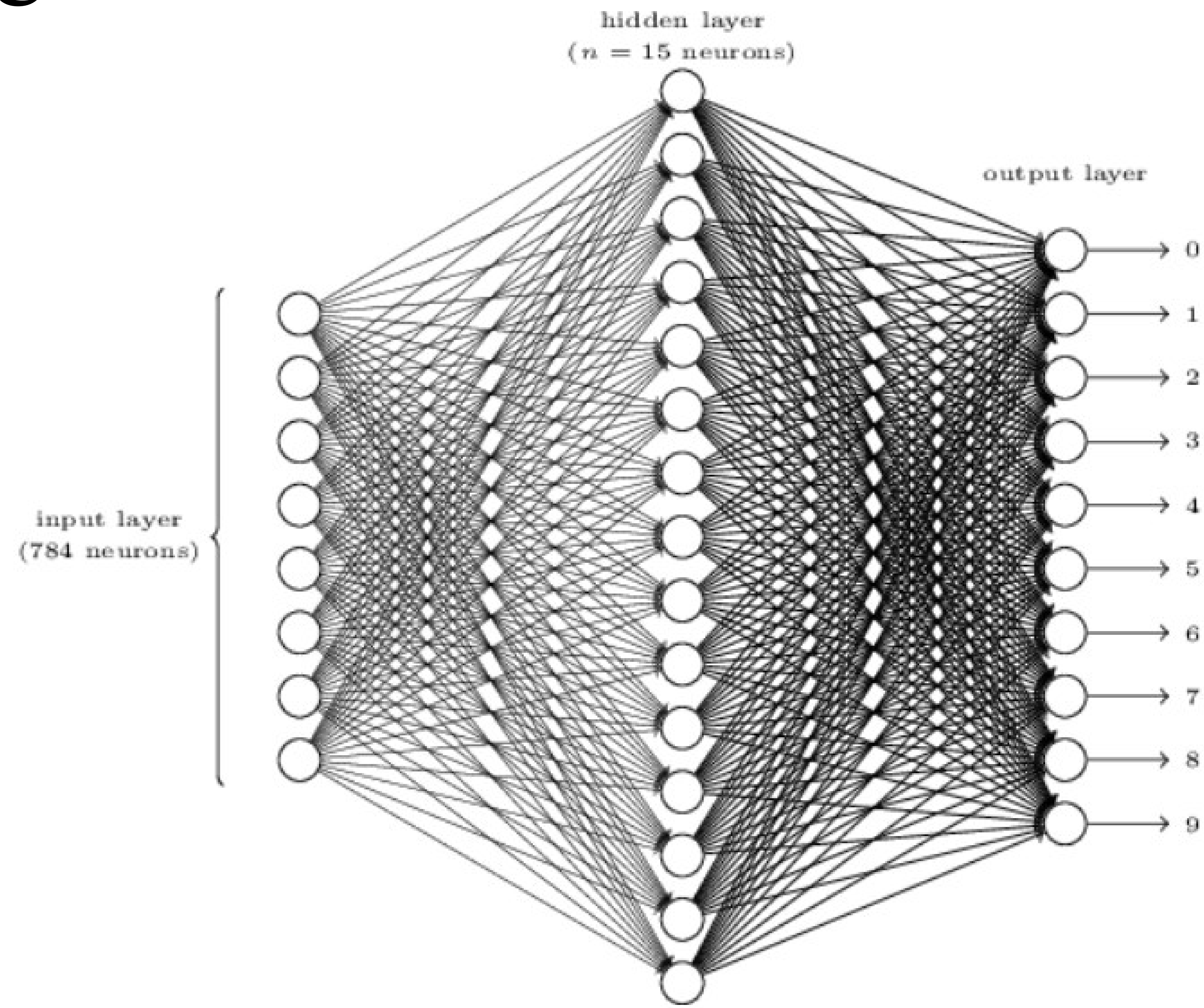


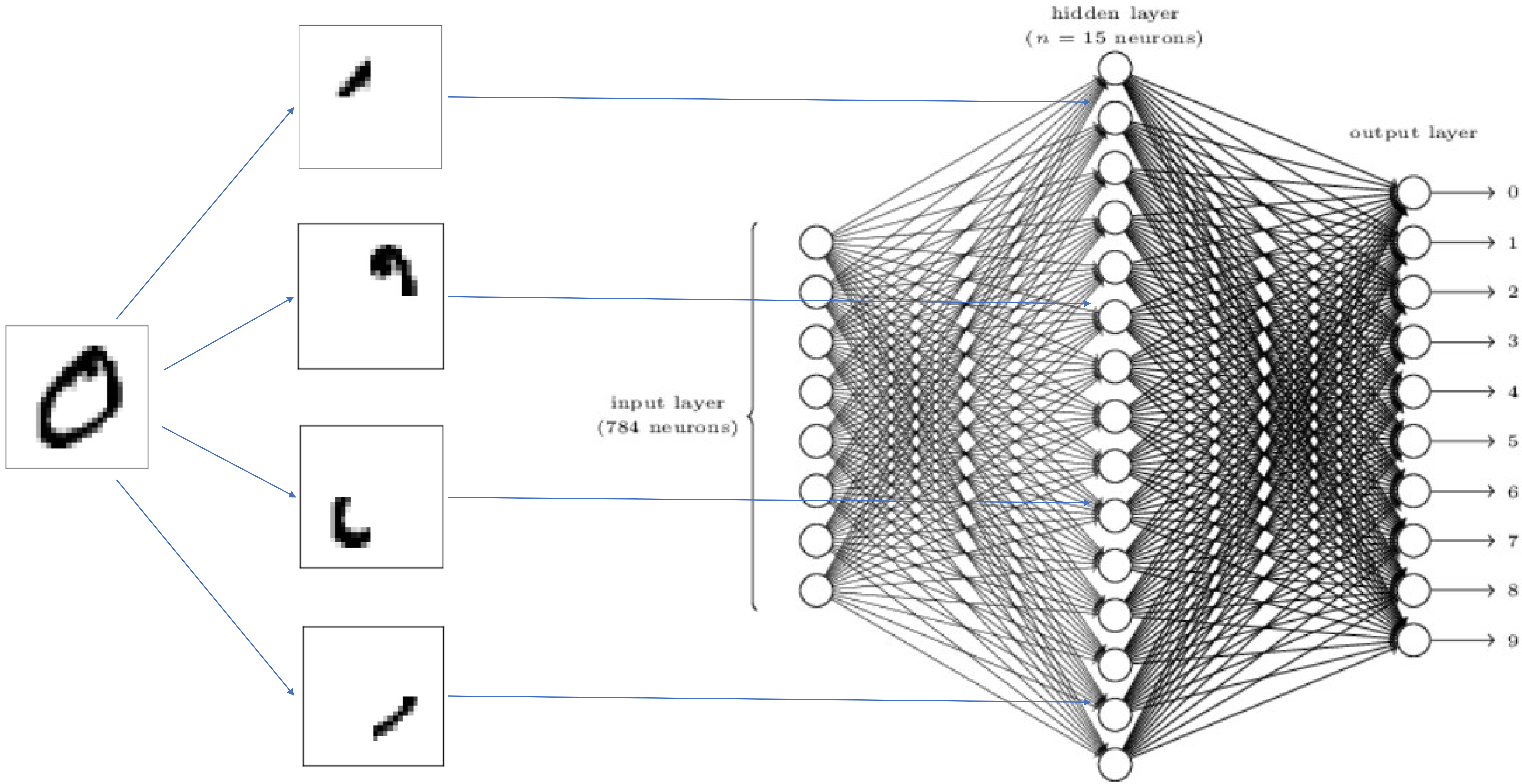


$$f(net) = \begin{cases} net & if net >= 0 \\ \alpha \times (exp(net) - 1) & otherwise \end{cases}$$

Hand Written Digits Classification







hidden layer
(n = 15 neurons)
output layer
input layer
(784 neurons)
0
1
2
3
4
5
6
7
8
9

Loss Function (Cost Function)

- *Quadratic* Cost Function (*Mean Squared Error, MSE)*

$$C(w,b) \equiv \frac{1}{2n} \sum_x \|y(x) - a\|^2$$

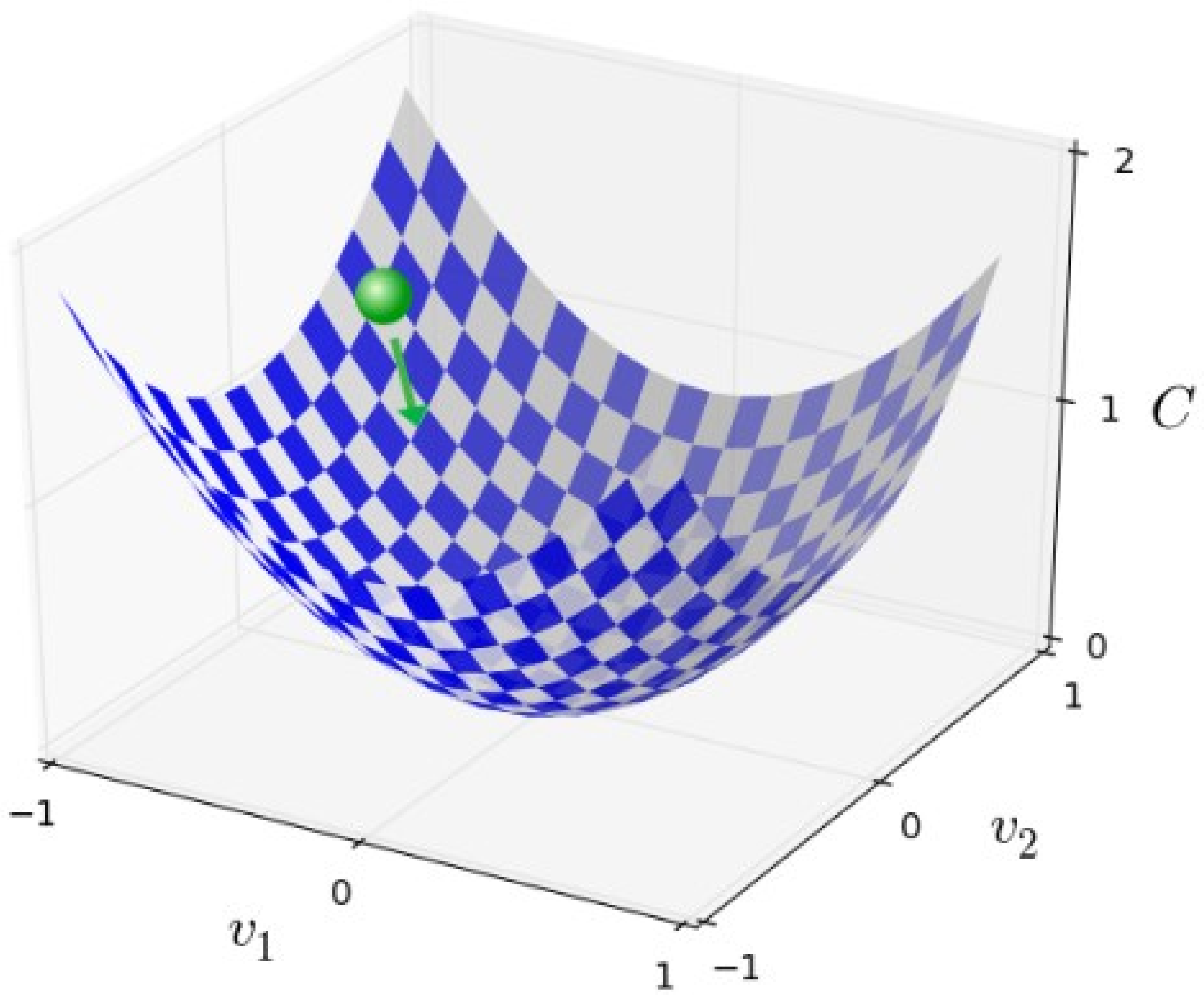
- *Cross Entropy*

$$H(p,q) = -\sum_x p(x)\, \log q(x)$$

- Focal Loss

$$Loss(x, class) = -\alpha_{class} (1 - \frac{e^{x[class]}}{\sum_j e^{x[j]}})^\gamma \log\left(\frac{e^{x[class]}}{\sum_j e^{x[j]}}\right)$$

Optimizer - Gradient Descent



$$\Delta C \approx \frac{\partial C}{\partial v_1}\Delta v_1 + \frac{\partial C}{\partial v_2}\Delta v_2$$

$$\nabla C \equiv \left(\frac{\partial C}{\partial v_1}, \frac{\partial C}{\partial v_2}\right)^T.$$

$$\Delta C \approx \nabla C \cdot \Delta v. \quad \Delta v = -\eta \nabla C,$$

$$\Delta C \approx -\eta \nabla C \cdot \nabla C = -\eta \|\nabla C\|^2$$

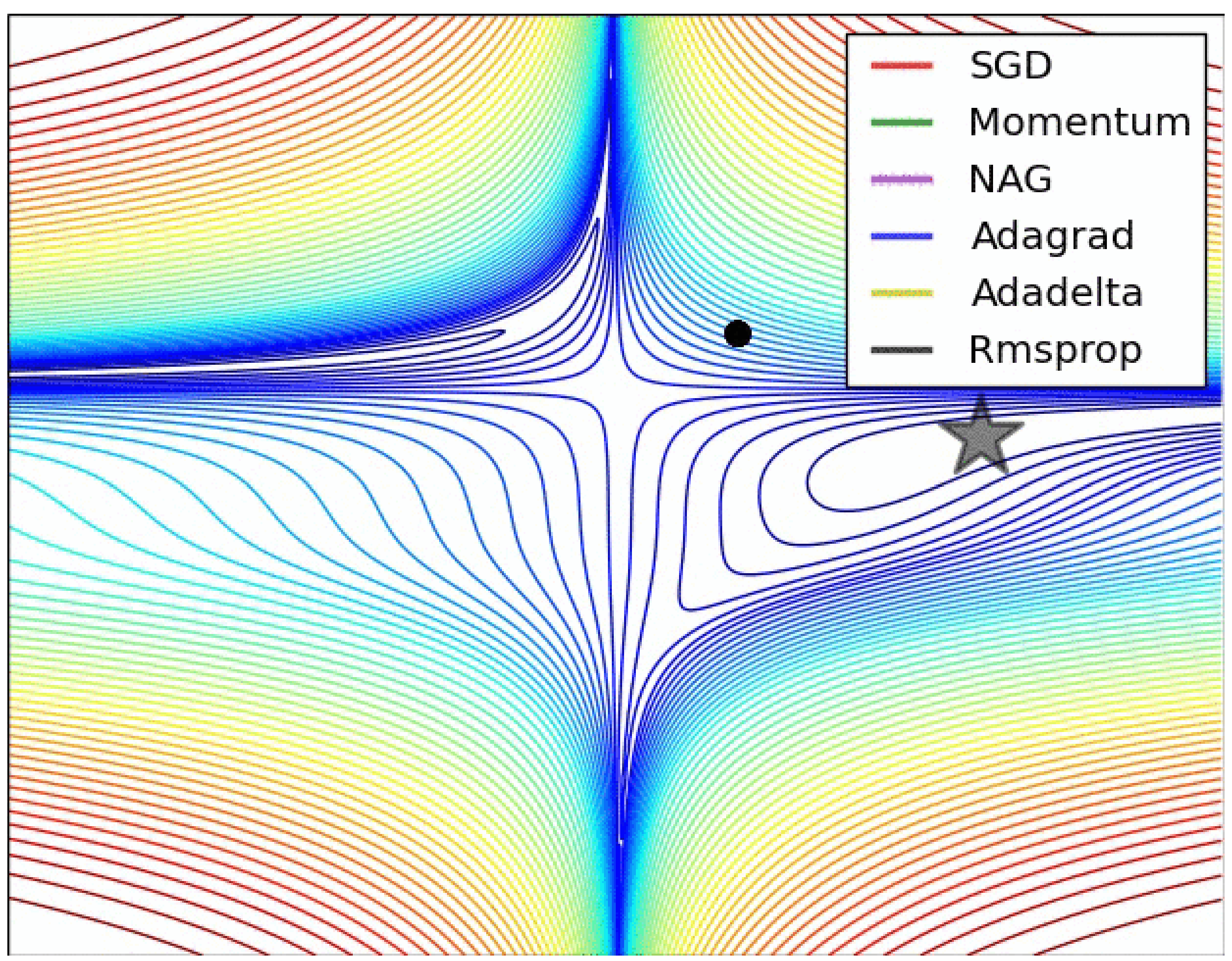
$$v \rightarrow v' = v - \eta \nabla C.$$

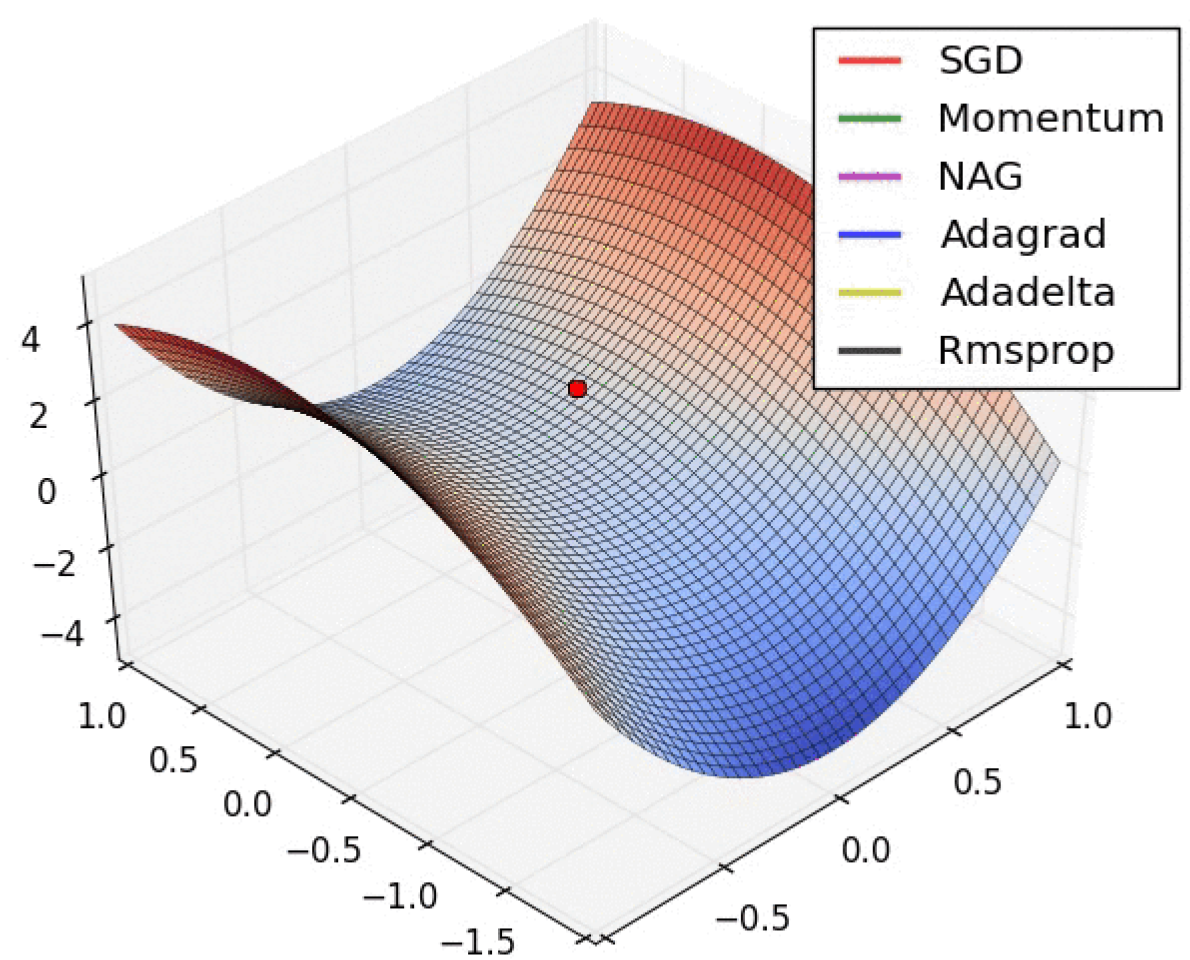
$$w_k \rightarrow w_k' = w_k - \eta \frac{\partial C}{\partial w_k}$$

$$b_l \rightarrow b_l' = b_l - \eta \frac{\partial C}{\partial b_l}.$$

Optimizer (backpropagation)

Optimizer	特點
SGD	• 有機會跳出目前局部收斂進而進到另一個局部收斂而得到最小值，而得到全局最小值 • 需自行設定learning rate，較難選擇到合適的learning rate • 會造成loss function有嚴重的震蕩 • 需要較長時間收斂至最小值
Momentum	• 能夠在相關方向加速SGD，抑制SGD的嚴重震蕩，進而加快收斂 • 需自行設定learning rate與β，有可能會使參數的移動方向偏移梯度下分的方向，進而導至沒有那麼快速的收斂
AdaGrad	• 能夠自動調整learning rate，進而調整收斂 • 適合處理稀疏梯度 • 依然需要人工設置一個全局的learning rate • 後期，分母梯度平方的累加會越來越大，會使梯度趨近於0，使得訓練結束
Adam	• 結合了AdaGrad與Momentum的優點 • 適用於大數據集和高維空間的資料 • 目前最常使用的一個Optimizer

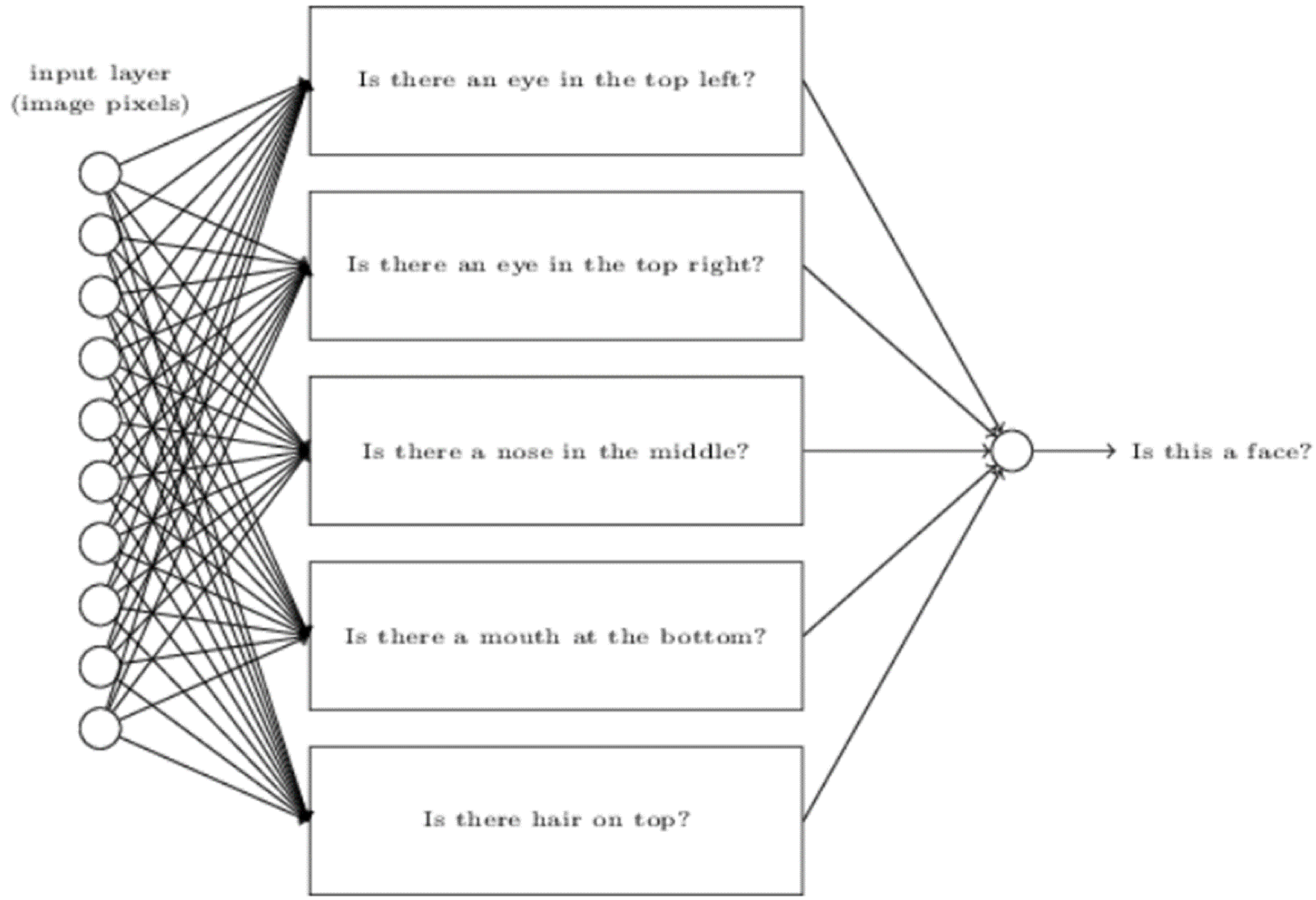




Reference:
雞雞與兔兔的工程世界
Alec Radford

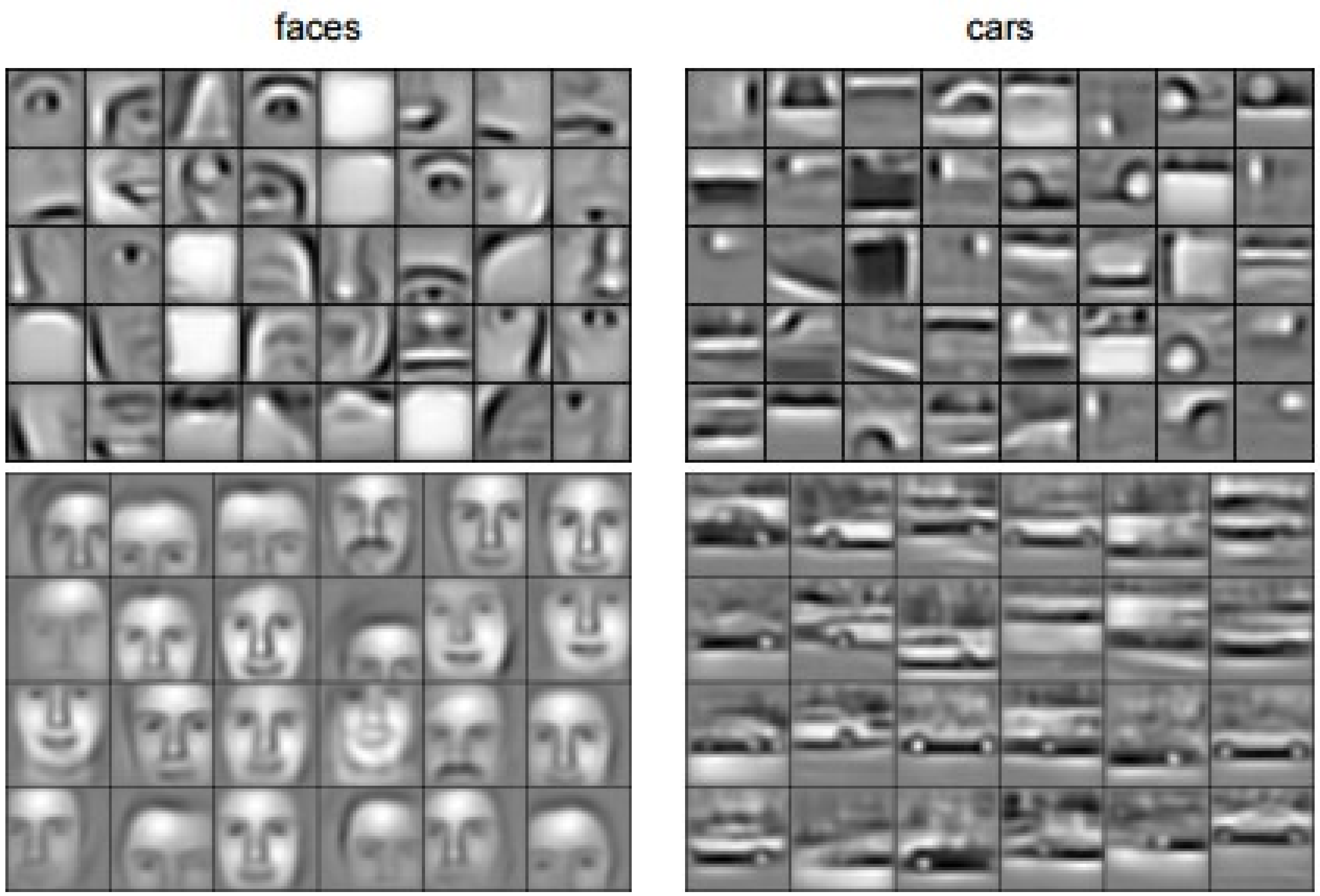
Summary

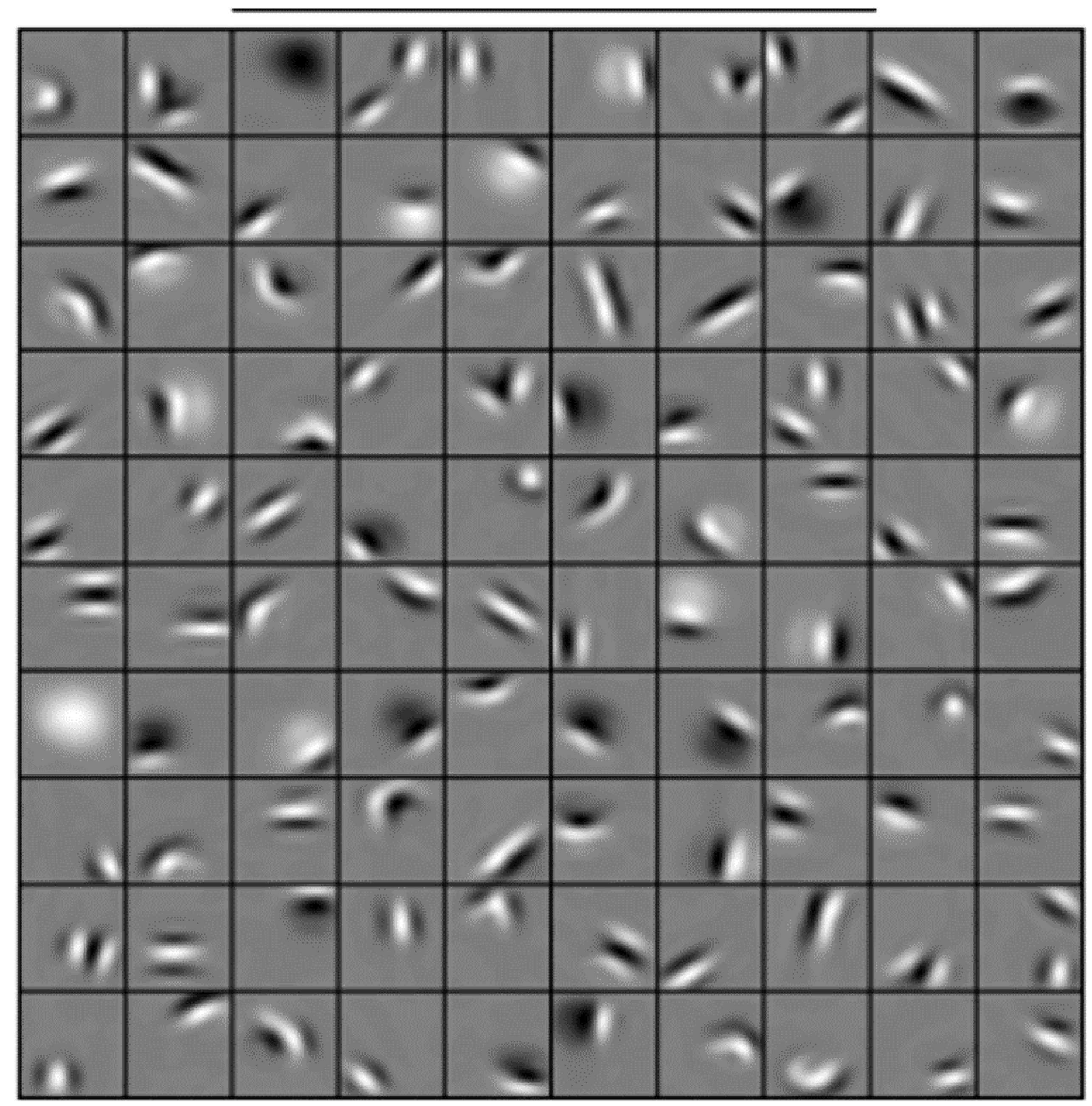
- Architecture
- Activation Function
- Loss Function
- Optimizer
- Mini Batch



[NN Play Ground]

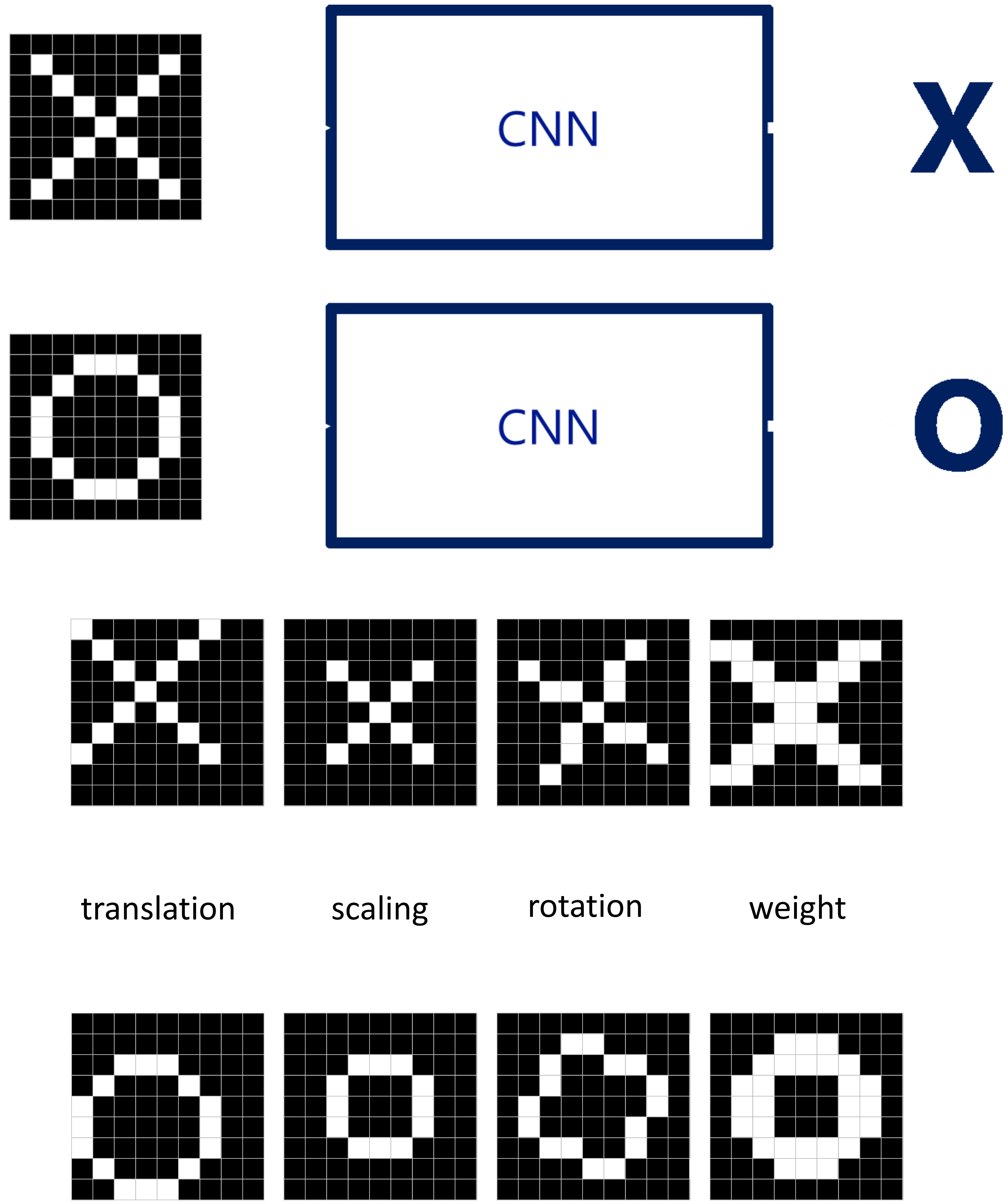
Convolutional Neural Network





A Toy ConvNet: X's and O's

A two-dimensional
array of pixels


CNN
X
CNN
O
translation
scaling
rotation
weight

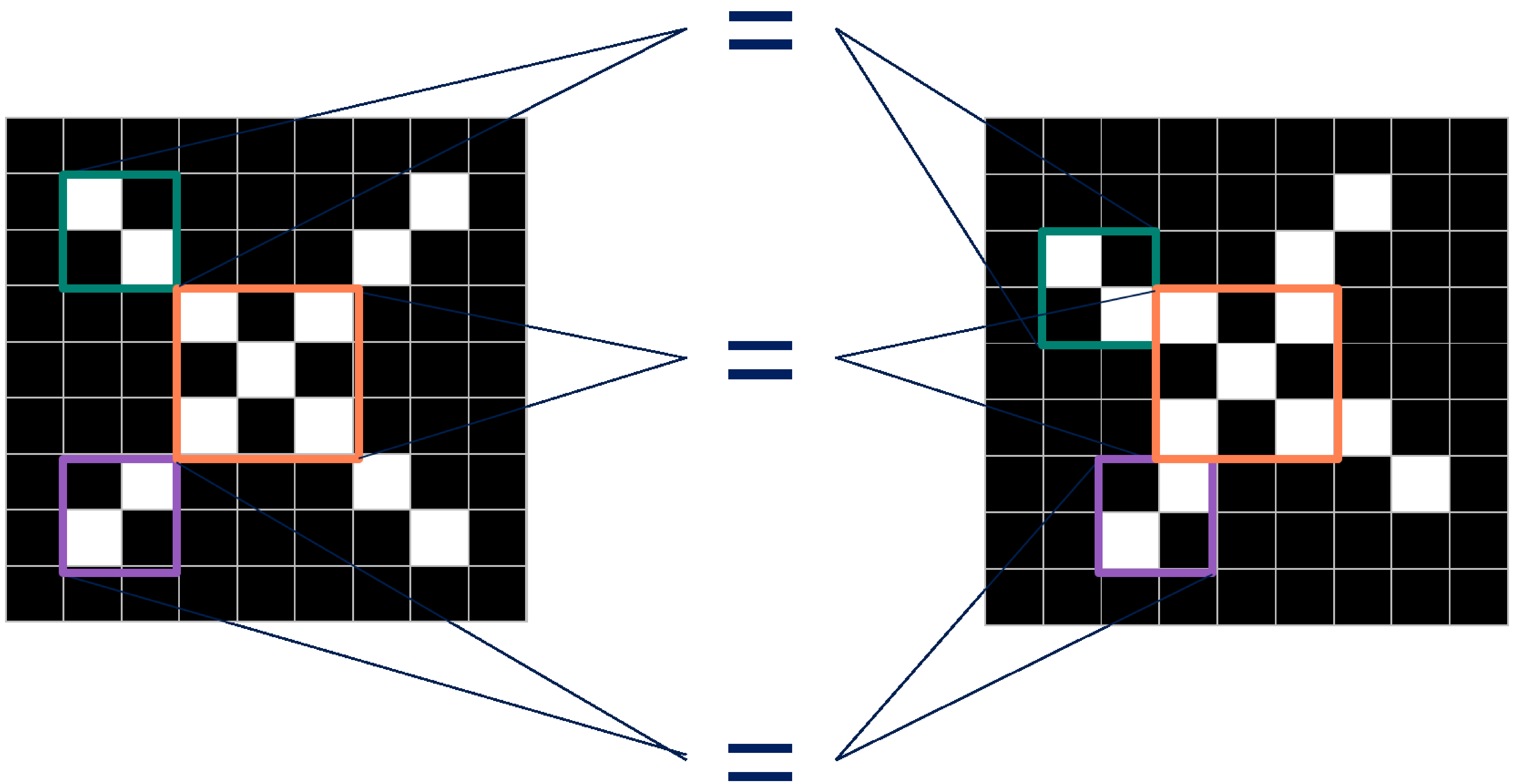
Computers Are Literal

-1	-1	-1	-1	-1	-1	-1	-1	-1
-1	1	-1	-1	-1	-1	-1	1	-1
-1	-1	1	-1	-1	-1	1	-1	-1
-1	-1	-1	1	-1	1	-1	-1	-1
-1	-1	-1	-1	1	-1	-1	-1	-1
-1	-1	-1	1	-1	1	-1	-1	-1
-1	-1	1	-1	-1	-1	1	-1	-1
-1	1	-1	-1	-1	-1	-1	1	-1
-1	-1	-1	-1	-1	-1	-1	-1	-1

$\neq$

-1	-1	-1	-1	-1	-1	-1	-1	-1
-1	-1	-1	-1	-1	-1	1	-1	-1
-1	1	-1	-1	-1	1	-1	-1	-1
-1	-1	1	1	-1	1	-1	-1	-1
-1	-1	-1	-1	1	-1	-1	-1	-1
-1	-1	-1	1	-1	1	1	-1	-1
-1	-1	-1	1	-1	-1	-1	1	-1
-1	-1	1	-1	-1	-1	-1	-1	-1
-1	-1	-1	-1	-1	-1	-1	-1	-1

Cor

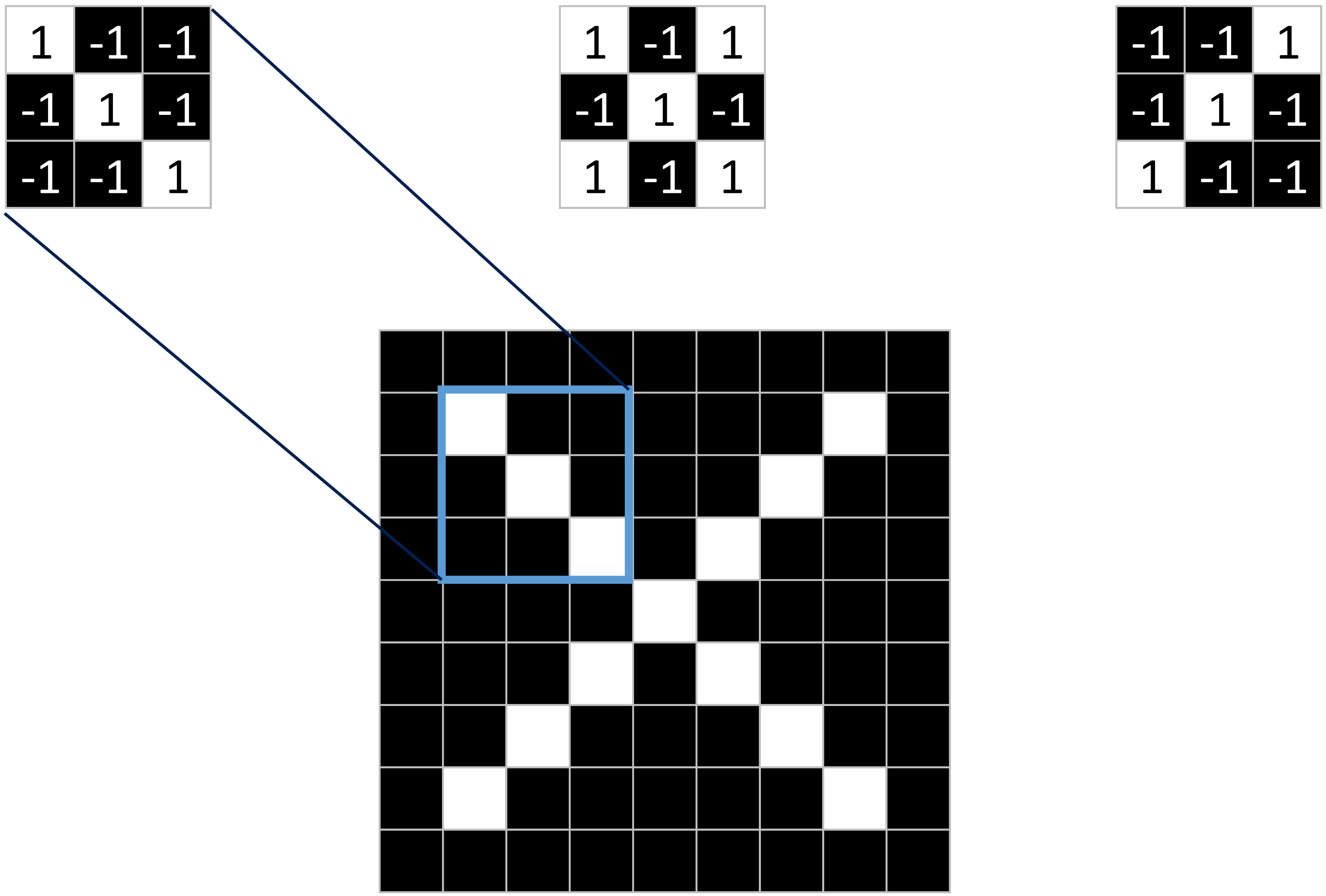


Features Match Pieces of The Image

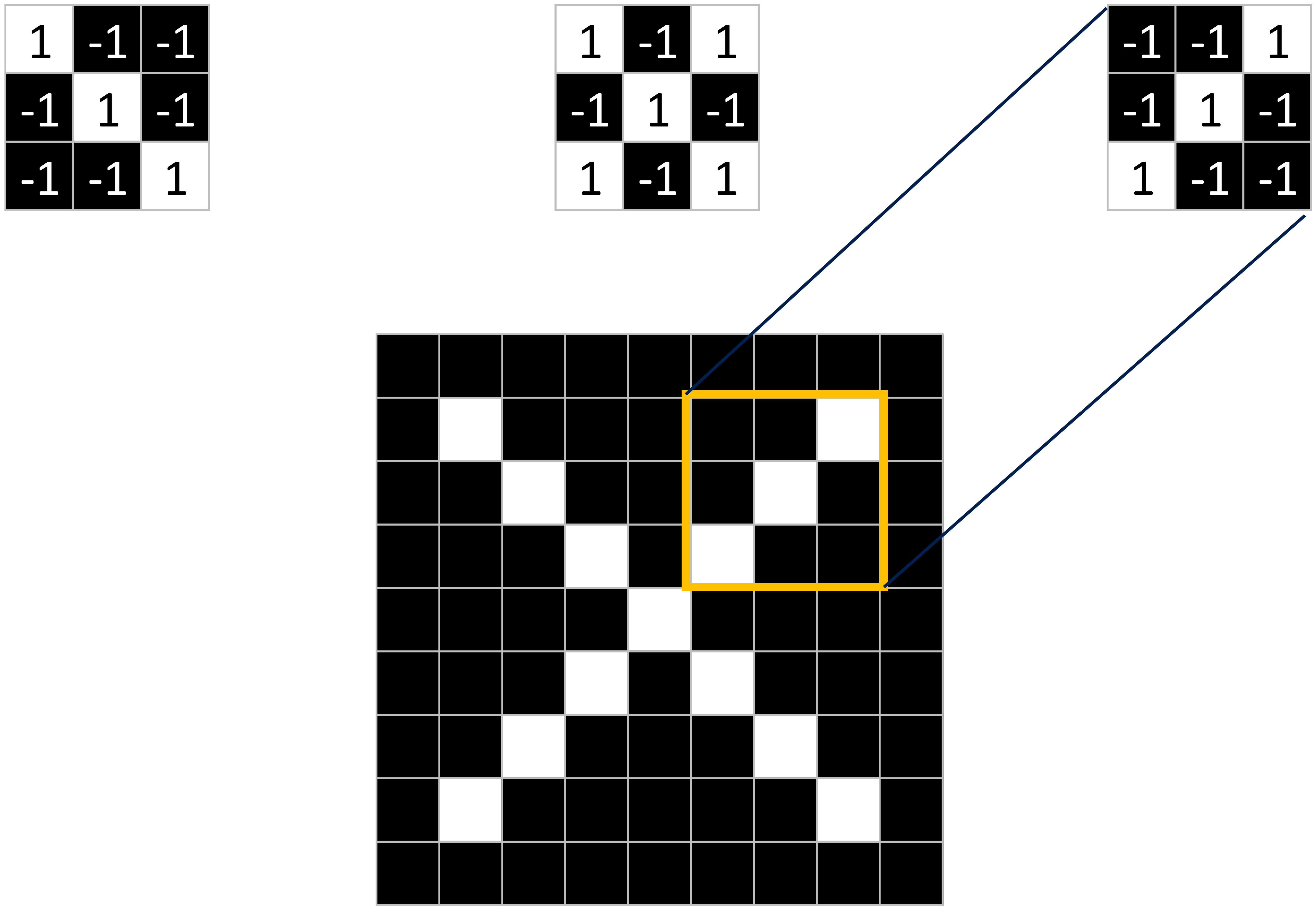
1	-1	-1
-1	1	-1
-1	-1	1

1	-1	1
-1	1	-1
1	-1	1

-1	-1	1
-1	1	-1
1	-1	-1



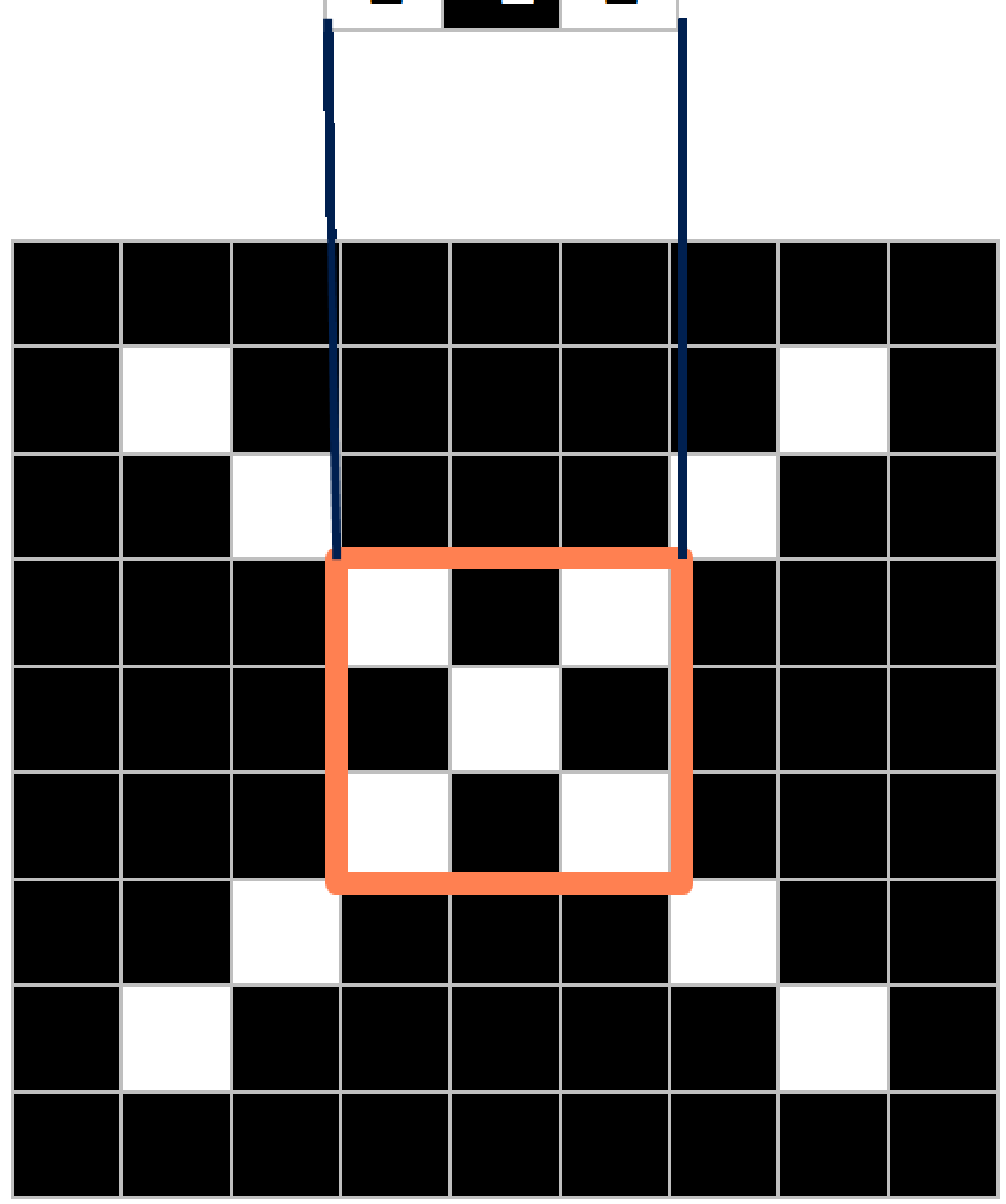
1 -1 -1
-1 1 -1
-1 -1 1
1 -1 1
-1 1 -1
1 -1 1
-1 -1 1
-1 1 -1
1 -1 -1

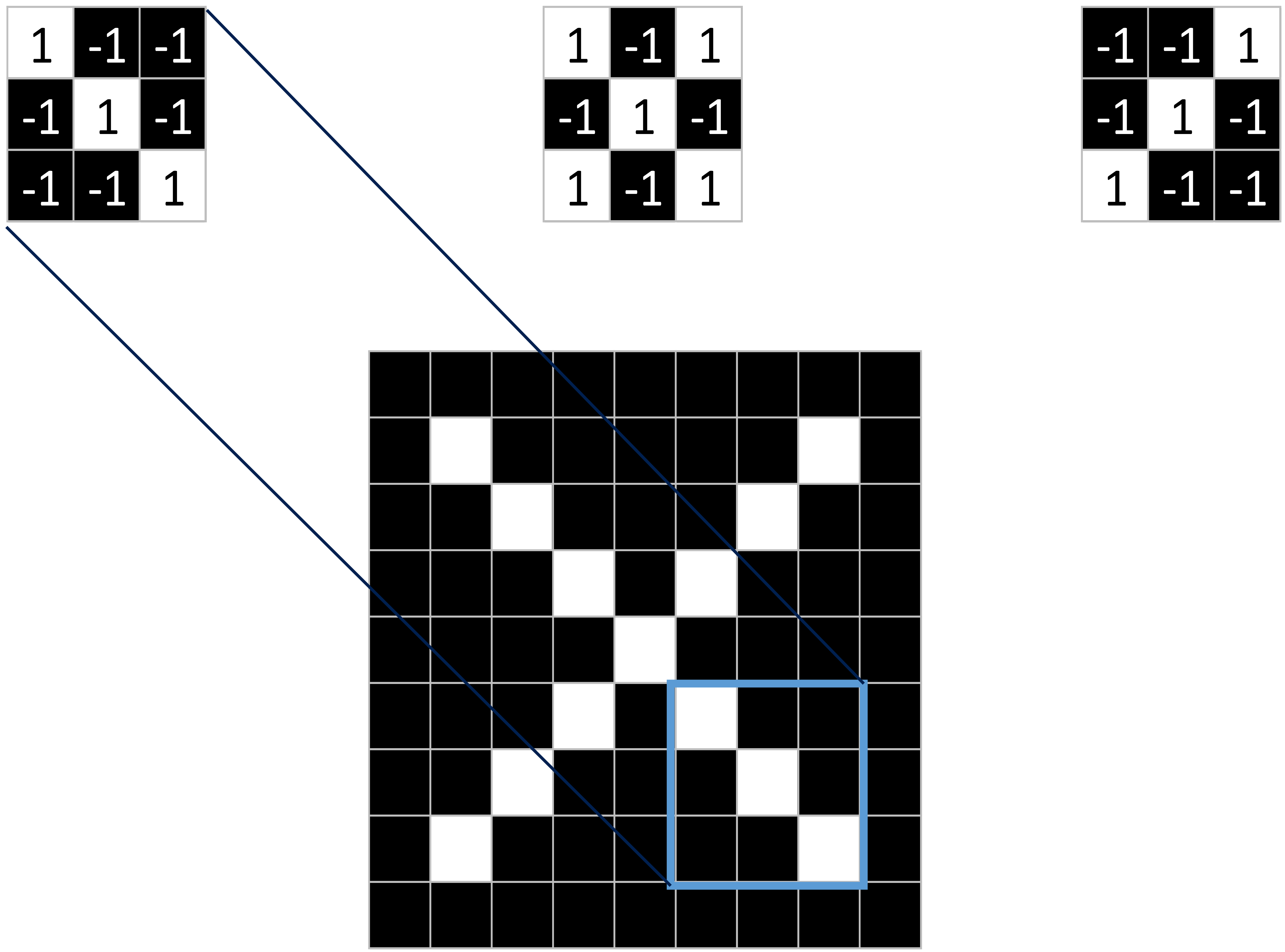

1 -1 -1
-1 1 -1
-1 -1 1
1 -1 1
-1 1 -1
1 -1 1
-1 -1 1
-1 1 -1
1 -1 -1

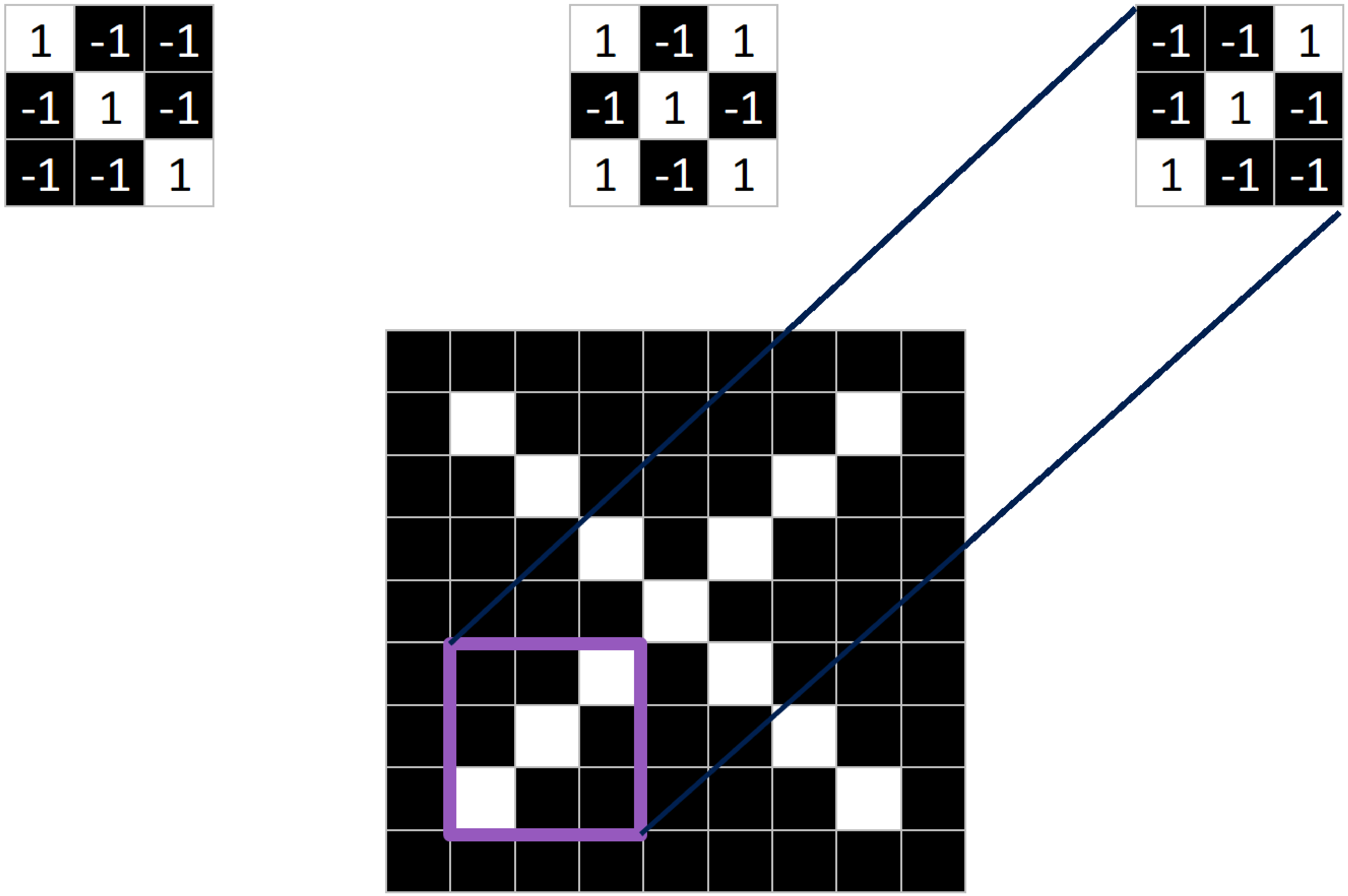
1	-1	-1
-1	1	-1
-1	-1	1

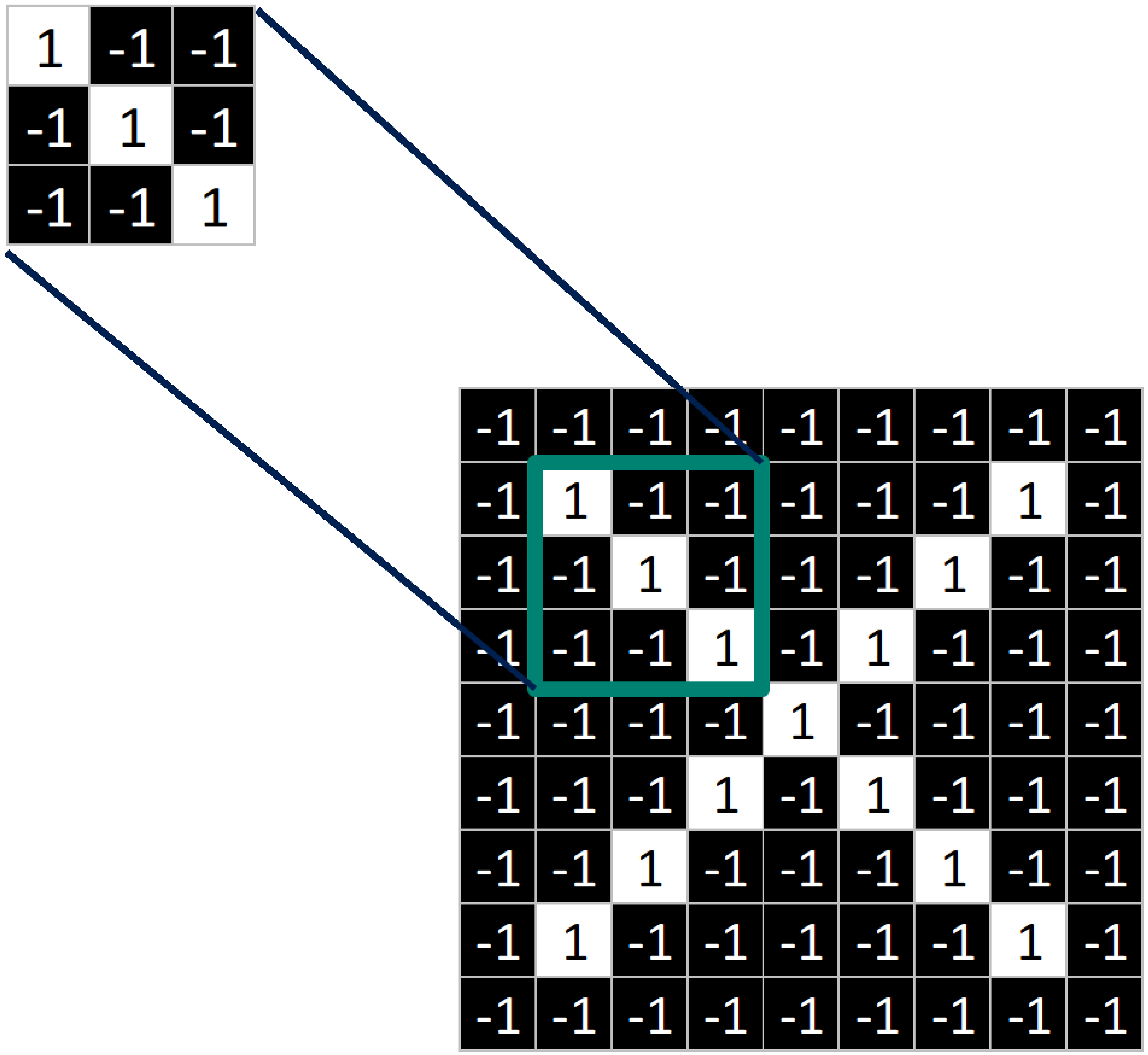
1	-1	1
-1	1	-1
1	-1	1

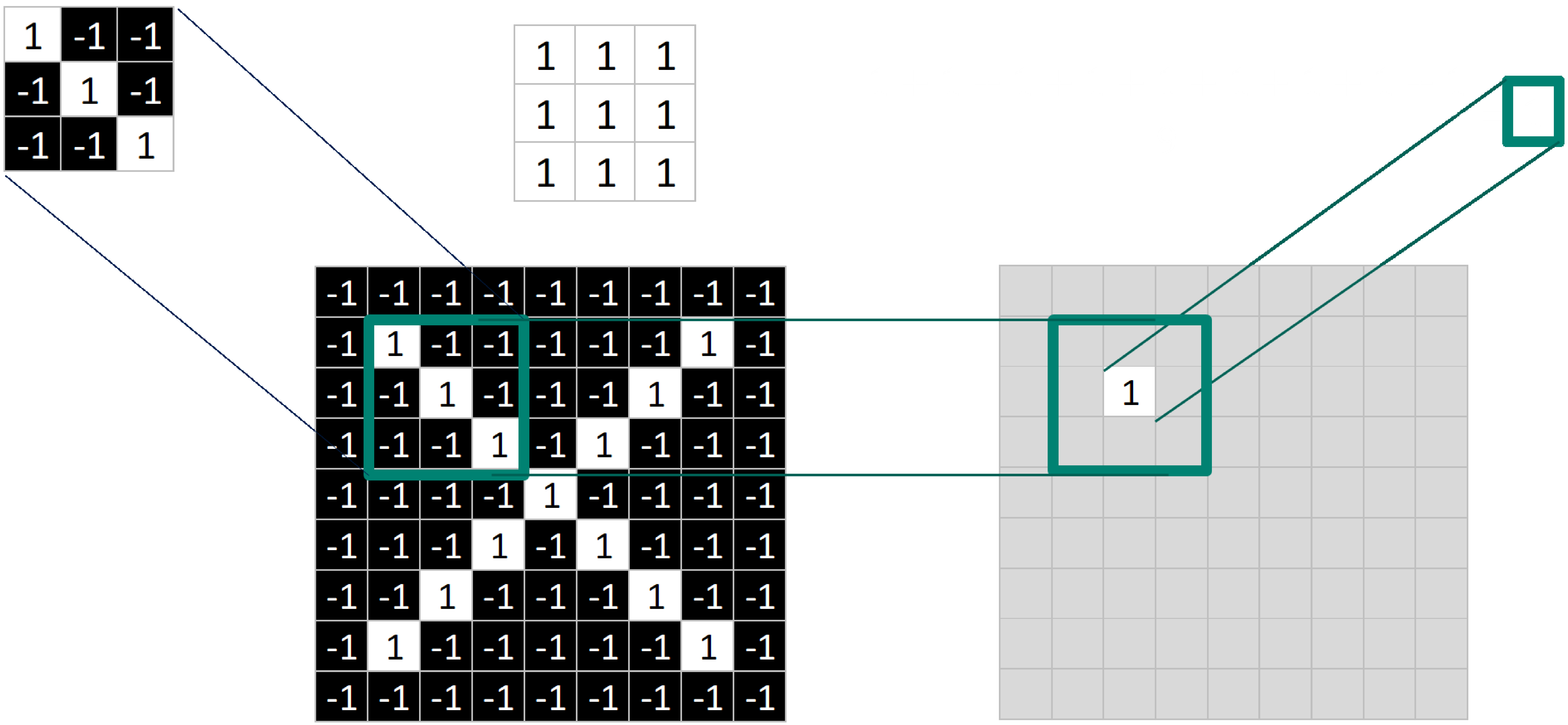
-1	-1	1
-1	1	-1
1	-1	-1

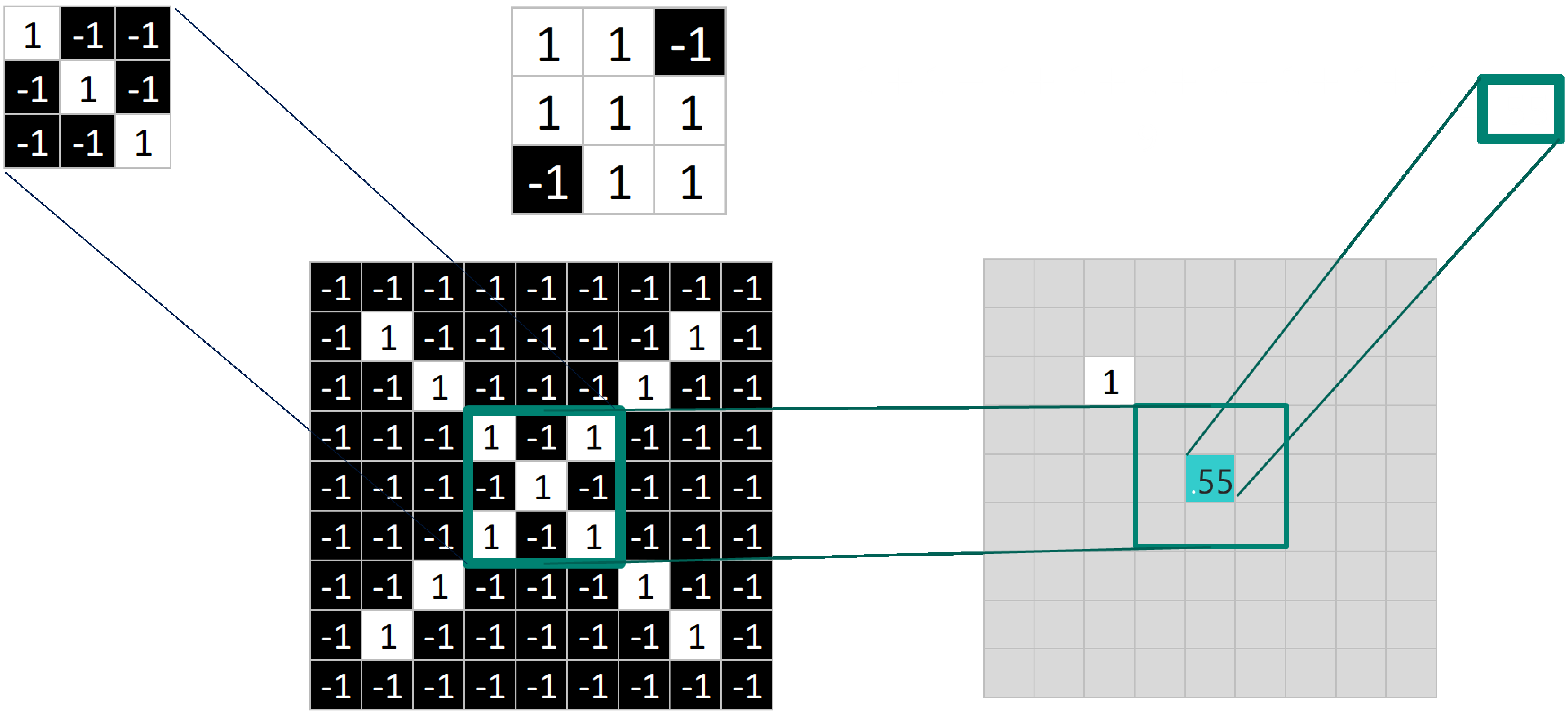



1 -1 -1
-1 1 -1
-1 -1 1
1 -1 1
-1 1 -1
1 -1 1
-1 -1 1
-1 1 -1
1 -1 -1

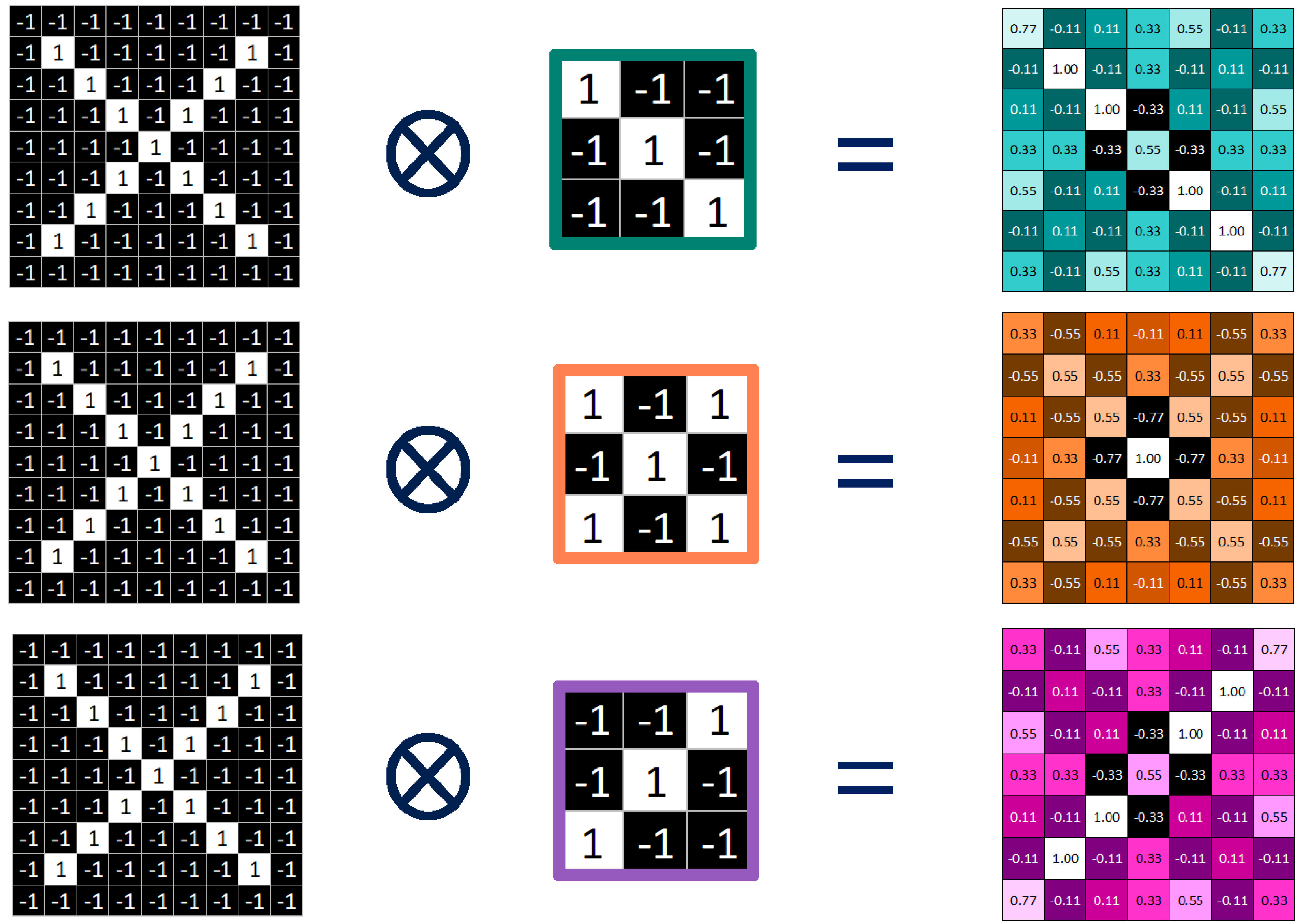

1 -1 -1
-1 1 -1
-1 -1 1
1 -1 1
-1 1 -1
1 -1 1
-1 -1 1
-1 1 -1
1 -1 -1


1 -1 -1
-1 1 -1
-1 -1 1
-1 -1 -1 -1 -1 -1 -1 -1 -1
-1 1 -1 -1 -1 -1 -1 1 -1
-1 -1 1 -1 -1 -1 1 -1 -1
-1 -1 -1 1 -1 1 -1 -1 -1
-1 -1 -1 -1 1 -1 -1 -1 -1
-1 -1 -1 1 -1 1 -1 -1 -1
-1 -1 1 -1 -1 -1 1 -1 -1
-1 1 -1 -1 -1 -1 -1 1 -1
-1 -1 -1 -1 -1 -1 -1 -1 -1

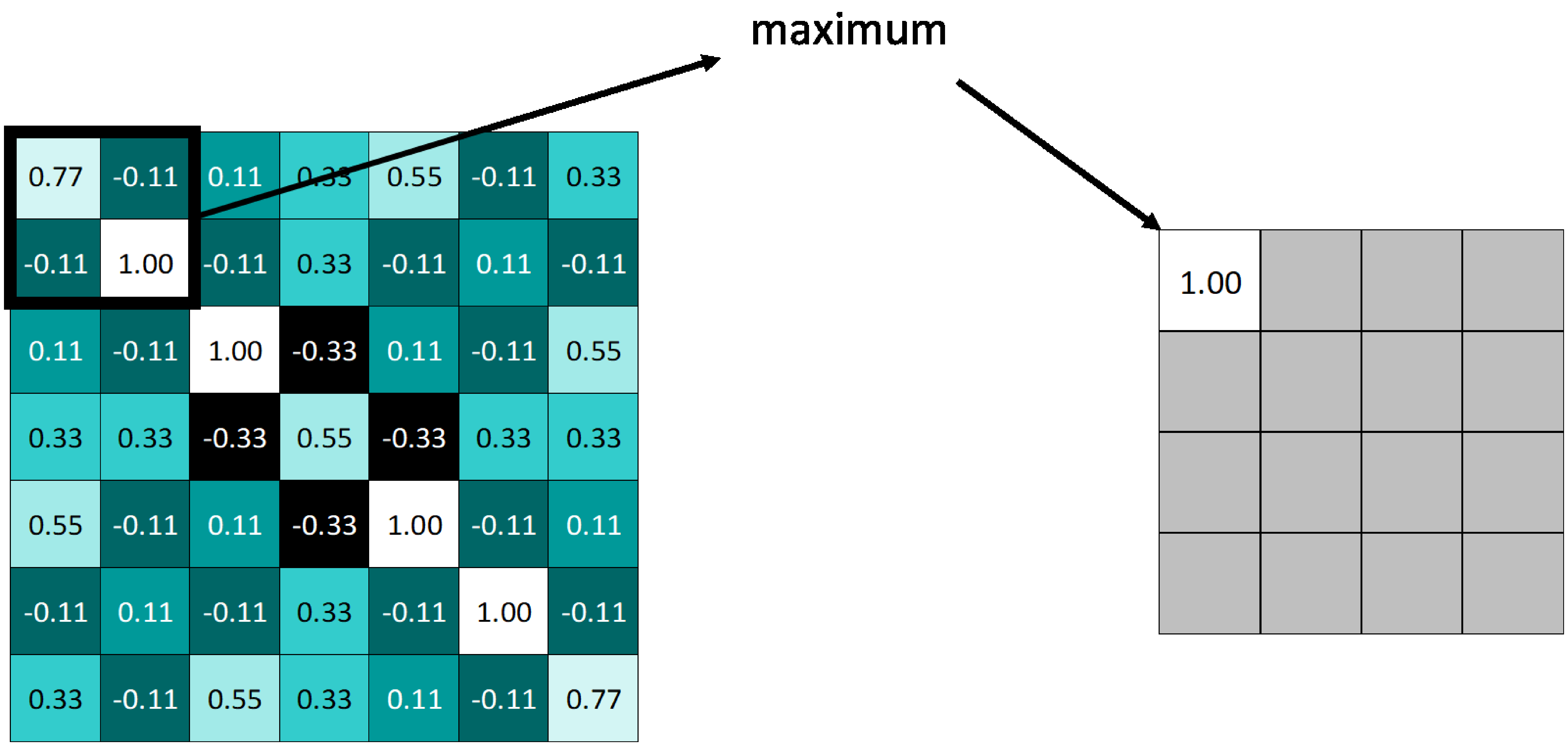

1 -1 -1
-1 1 -1
-1 -1 1
1 1 1
1 1 1
1 1 1
-1 -1 -1 -1 -1 -1 -1 -1 -1
-1 1 -1 -1 -1 -1 -1 1 -1
-1 -1 1 -1 -1 -1 1 -1 -1
-1 -1 -1 1 -1 1 -1 -1 -1
-1 -1 -1 -1 1 -1 -1 -1 -1
-1 -1 -1 1 -1 1 -1 -1 -1
-1 -1 1 -1 -1 -1 1 -1 -1
-1 1 -1 -1 -1 -1 -1 1 -1
-1 -1 -1 -1 -1 -1 -1 -1 -1
1



1 -1 -1
-1 1 -1
-1 -1 1
1 1 -1
1 1 1
-1 1 1
-1 -1 -1 -1 -1 -1 -1 -1 -1
-1 1 -1 -1 -1 -1 -1 1 -1
-1 -1 1 -1 -1 -1 1 -1 -1
-1 -1 -1 1 -1 1 -1 -1 -1
-1 -1 -1 -1 1 -1 -1 -1 -1
-1 -1 -1 1 -1 1 -1 -1 -1
-1 -1 1 -1 -1 -1 1 -1 -1
-1 1 -1 -1 -1 -1 -1 1 -1
-1 -1 -1 -1 -1 -1 -1 -1 -1
1
.55


-1 -1 -1 -1 -1 -1 -1 -1 -1
-1 1 -1 -1 -1 -1 -1 1 -1
-1 -1 1 -1 -1 -1 1 -1 -1
-1 -1 -1 1 -1 1 -1 -1 -1
-1 -1 -1 -1 1 -1 -1 -1 -1
-1 -1 -1 1 -1 1 -1 -1 -1
-1 -1 1 -1 -1 -1 1 -1 -1
-1 1 -1 -1 -1 -1 -1 1 -1
-1 -1 -1 -1 -1 -1 -1 -1 -1
1 -1 -1
-1 1 -1
-1 -1 1
0.77 -0.11 0.11 0.33 0.55 -0.11 0.33
-0.11 1.00 -0.11 0.33 -0.11 0.11 -0.11
0.11 -0.11 1.00 -0.33 0.11 -0.11 0.55
0.33 0.33 -0.33 0.55 -0.33 0.33 0.33
0.55 -0.11 0.11 -0.33 1.00 -0.11 0.11
-0.11 0.11 -0.11 0.33 -0.11 1.00 -0.11
0.33 -0.11 0.55 0.33 0.11 -0.11 0.77
-1 -1 -1 -1 -1 -1 -1 -1 -1
-1 1 -1 -1 -1 -1 -1 1 -1
-1 -1 1 -1 -1 -1 1 -1 -1
-1 -1 -1 1 -1 1 -1 -1 -1
-1 -1 -1 -1 1 -1 -1 -1 -1
-1 -1 -1 1 -1 1 -1 -1 -1
-1 -1 1 -1 -1 -1 1 -1 -1
-1 1 -1 -1 -1 -1 -1 1 -1
-1 -1 -1 -1 -1 -1 -1 -1 -1
1 -1 1
-1 1 -1
1 -1 1
0.33 -0.55 0.11 -0.11 0.11 -0.55 0.33
-0.55 0.55 -0.55 0.33 -0.55 0.55 -0.55
0.11 -0.55 0.55 -0.77 0.55 -0.55 0.11
-0.11 0.33 -0.77 1.00 -0.77 0.33 -0.11
0.11 -0.55 0.55 -0.77 0.55 -0.55 0.11
-0.55 0.55 -0.55 0.33 -0.55 0.55 -0.55
0.33 -0.55 0.11 -0.11 0.11 -0.55 0.33
-1 -1 -1 -1 -1 -1 -1 -1 -1
-1 1 -1 -1 -1 -1 -1 1 -1
-1 -1 1 -1 -1 -1 1 -1 -1
-1 -1 -1 1 -1 1 -1 -1 -1
-1 -1 -1 -1 1 -1 -1 -1 -1
-1 -1 -1 1 -1 1 -1 -1 -1
-1 -1 1 -1 -1 -1 1 -1 -1
-1 1 -1 -1 -1 -1 -1 1 -1
-1 -1 -1 -1 -1 -1 -1 -1 -1
-1 -1 1
-1 1 -1
1 -1 -1
0.33 -0.11 0.55 0.33 0.11 -0.11 0.77
-0.11 0.11 -0.11 0.33 -0.11 1.00 -0.11
0.55 -0.11 0.11 -0.33 1.00 -0.11 0.11
0.33 0.33 -0.33 0.55 -0.33 0.33 0.33
0.11 -0.11 1.00 -0.33 0.11 -0.11 0.55
-0.11 1.00 -0.11 0.33 -0.11 0.11 -0.11
0.77 -0.11 0.11 0.33 0.55 -0.11 0.33

Max Pooling


maximum
0.77 -0.11 0.11 0.33 0.55 -0.11 0.33
-0.11 1.00 -0.11 0.33 -0.11 0.11 -0.11
0.11 -0.11 1.00 -0.33 0.11 -0.11 0.55
0.33 0.33 -0.33 0.55 -0.33 0.33 0.33
0.55 -0.11 0.11 -0.33 1.00 -0.11 0.11
-0.11 0.11 -0.11 0.33 -0.11 1.00 -0.11
0.33 -0.11 0.55 0.33 0.11 -0.11 0.77
1.00

Max Pooling

0.77	-0.11	0.11	0.33	0.55	-0.11	0.33
-0.11	1.00	-0.11	0.33	-0.11	0.11	-0.11
0.11	-0.11	1.00	-0.33	0.11	-0.11	0.55
0.33	0.33	-0.33	0.55	-0.33	0.33	0.33
0.55	-0.11	0.11	-0.33	1.00	-0.11	0.11
-0.11	0.11	-0.11	0.33	-0.11	1.00	-0.11
0.33	-0.11	0.55	0.33	0.11	-0.11	0.77

maximum

1.00	0.33		

Max Pooling

0.77	-0.11	0.11	0.33	0.55	-0.11	0.33
-0.11	1.00	-0.11	0.33	-0.11	0.11	-0.11
0.11	-0.11	1.00	-0.33	0.11	-0.11	0.55
0.33	0.33	-0.33	0.55	-0.33	0.33	0.33
0.55	-0.11	0.11	-0.33	1.00	-0.11	0.11
-0.11	0.11	-0.11	0.33	-0.11	1.00	-0.11
0.33	-0.11	0.55	0.33	0.11	-0.11	0.77

maximum

1.00	0.33	0.55	

Max Pooling

0.77	-0.11	0.11	0.33	0.55	-0.11	0.33
-0.11	1.00	-0.11	0.33	-0.11	0.11	-0.11
0.11	-0.11	1.00	-0.33	0.11	-0.11	0.55
0.33	0.33	-0.33	0.55	-0.33	0.33	0.33
0.55	-0.11	0.11	-0.33	1.00	-0.11	0.11
-0.11	0.11	-0.11	0.33	-0.11	1.00	-0.11
0.33	-0.11	0.55	0.33	0.11	-0.11	0.77

1.00	0.33	0.55	0.33
0.33	1.00	0.33	0.55
0.55	0.33	1.00	0.11
0.33	0.55	0.11	0.77

Max Pooling

0.77	-0.11	0.11	0.33	0.55	-0.11	0.33
-0.11	1.00	-0.11	0.33	-0.11	0.11	-0.11
0.11	-0.11	1.00	-0.33	0.11	-0.11	0.55
0.33	0.33	-0.33	0.55	-0.33	0.33	0.33
0.55	-0.11	0.11	-0.33	1.00	-0.11	0.11
-0.11	0.11	-0.11	0.33	-0.11	1.00	-0.11
0.33	-0.11	0.55	0.33	0.11	-0.11	0.77

1.00	0.33	0.55	0.33
0.33	1.00	0.33	0.55
0.55	0.33	1.00	0.11
0.33	0.55	0.11	0.77

0.33	-0.55	0.11	-0.11	0.11	-0.55	0.33
-0.55	0.55	-0.55	0.33	-0.55	0.55	-0.55
0.11	-0.55	0.55	-0.77	0.55	-0.55	0.11
-0.11	0.33	-0.77	1.00	-0.77	0.33	-0.11
0.11	-0.55	0.55	-0.77	0.55	-0.55	0.11
-0.55	0.55	-0.55	0.33	-0.55	0.55	-0.55
0.33	-0.55	0.11	-0.11	0.11	-0.55	0.33

0.55	0.33	0.55	0.33
0.33	1.00	0.55	0.11
0.55	0.55	0.55	0.11
0.33	0.11	0.11	0.33

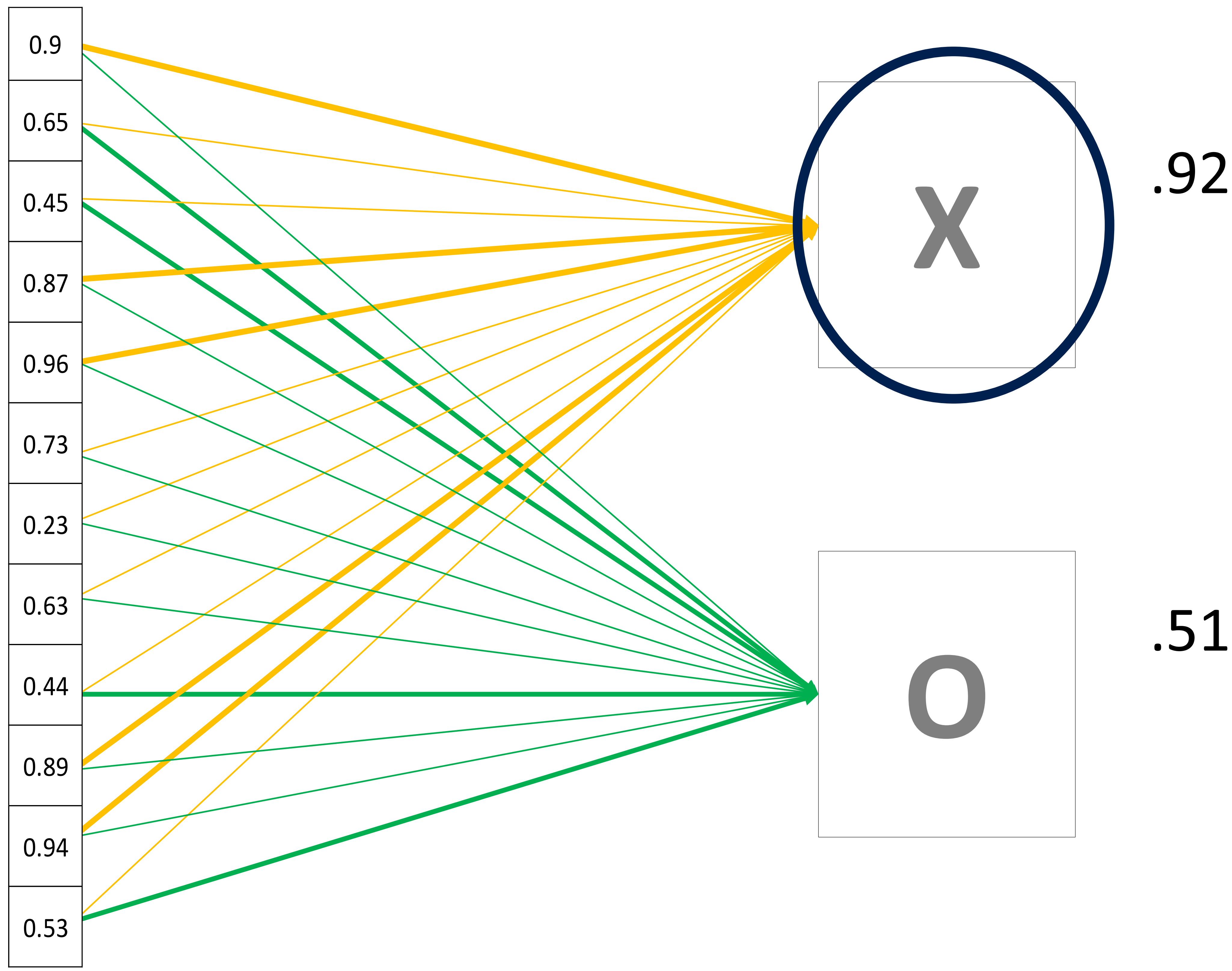
0.33	-0.11	0.55	0.33	0.11	-0.11	0.77
-0.11	0.11	-0.11	0.33	-0.11	1.00	-0.11
0.55	-0.11	0.11	-0.33	1.00	-0.11	0.11
0.33	0.33	-0.33	0.55	-0.33	0.33	0.33
0.11	-0.11	1.00	-0.33	0.11	-0.11	0.55
-0.11	1.00	-0.11	0.33	-0.11	0.11	-0.11
0.77	-0.11	0.11	0.33	0.55	-0.11	0.33

0.33	0.55	1.00	0.77
0.55	0.55	1.00	0.33
1.00	1.00	0.11	0.55
0.77	0.33	0.55	0.33

Pooling: Shrinking The Image Stack

- Pick a window size (usually 2 or 3).
- Pick a stride (usually 2).
- Walk your window across your filtered images.
- From each window, take the maximum value.

Fully Connected Layer



Hyperparameters

- Convolution
  - Number of features (Kernel/Filter)
  - Size of features (Kernel/Filter)
- Pooling
  - Window size
  - Window stride
- Fully Connected
  - Number of neurons (Nodes)

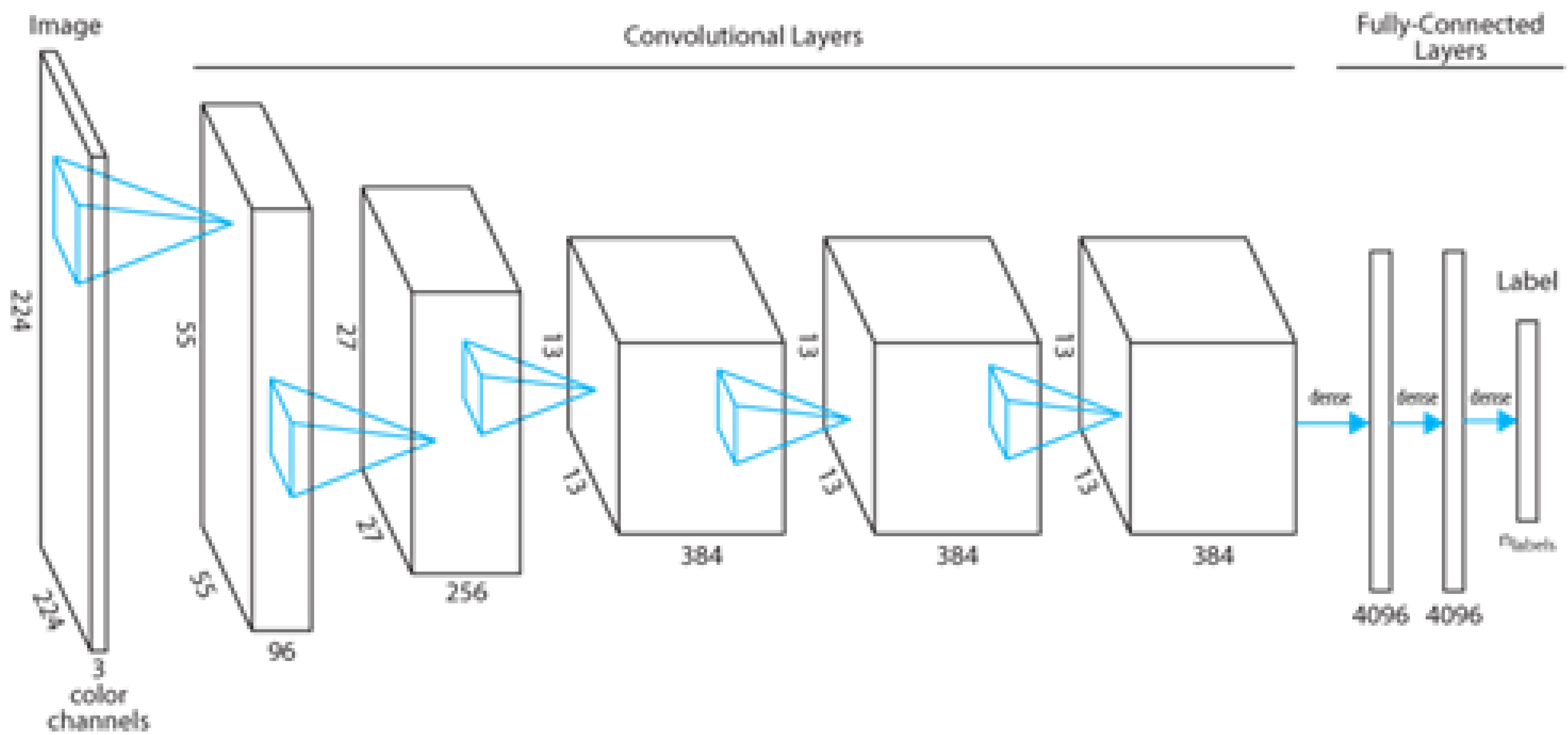
Architecture

- How many layers of each type ?

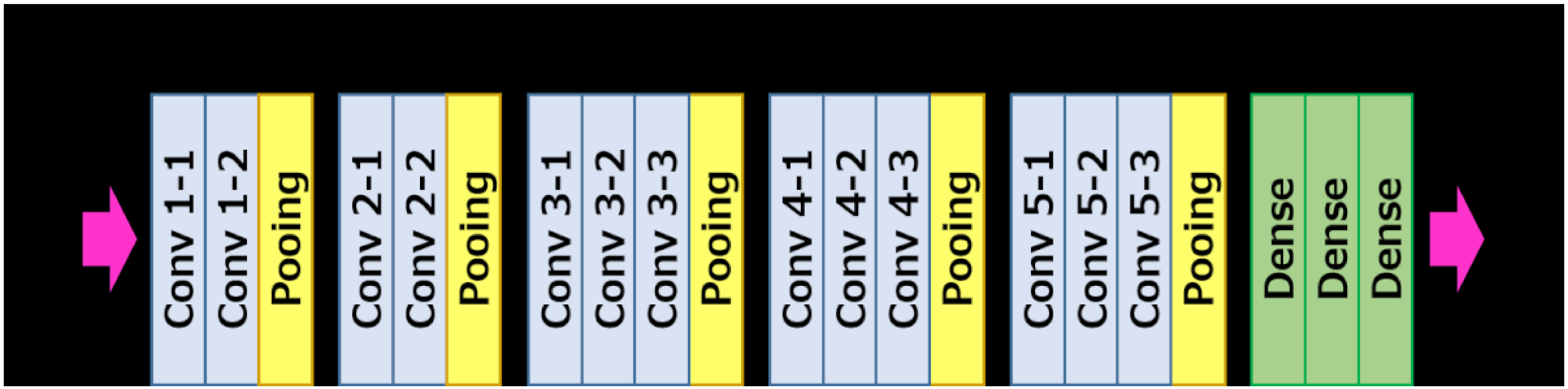
- In what order?

- How do all layers connect to each other?

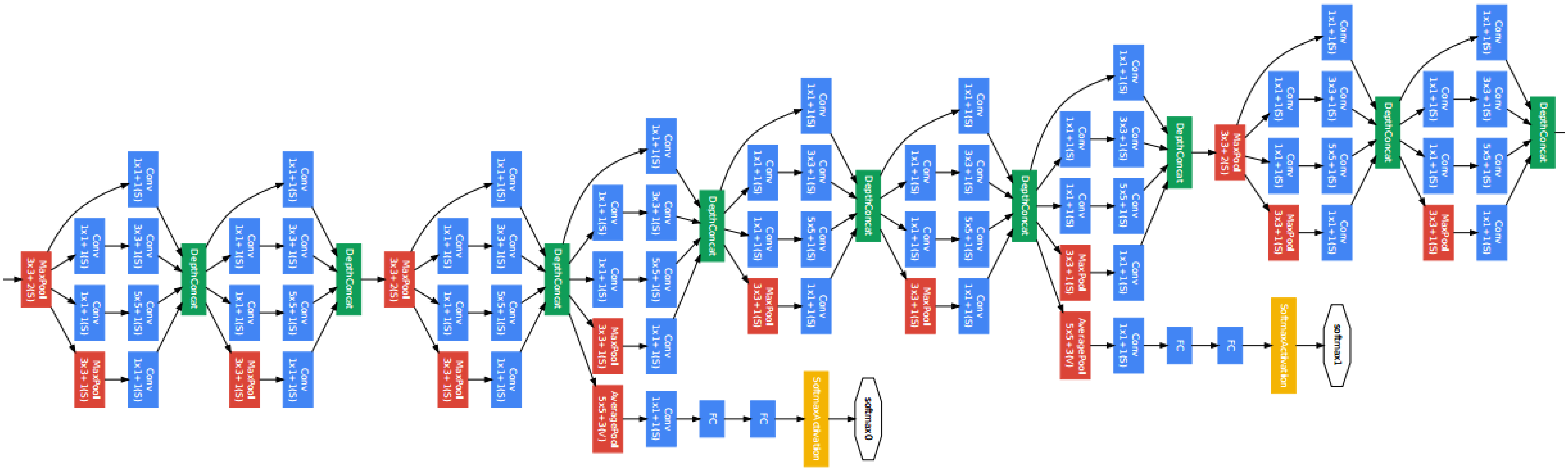
Classic Architecture - AlexNet



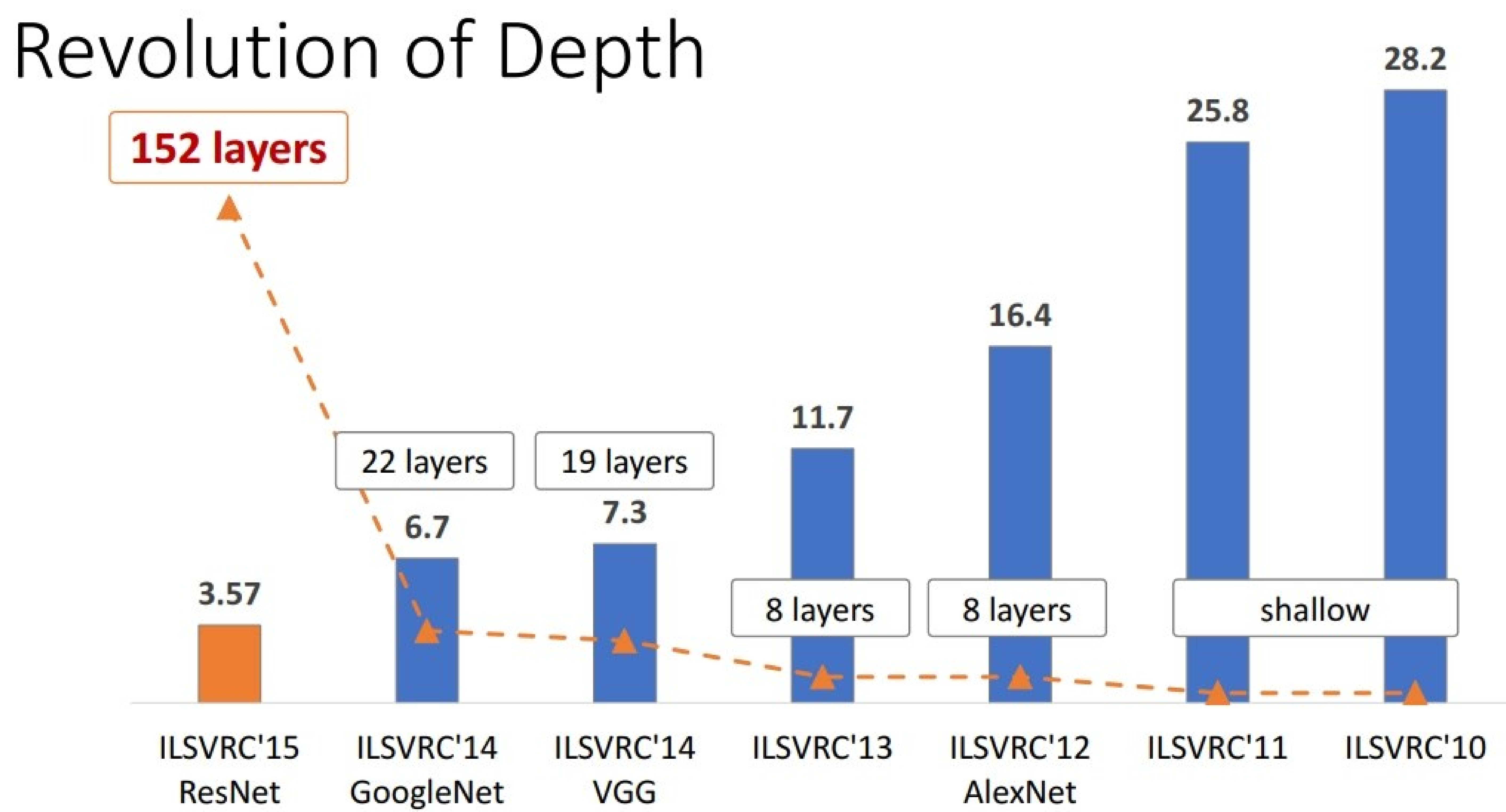
Classic Architecture – VGG16/19



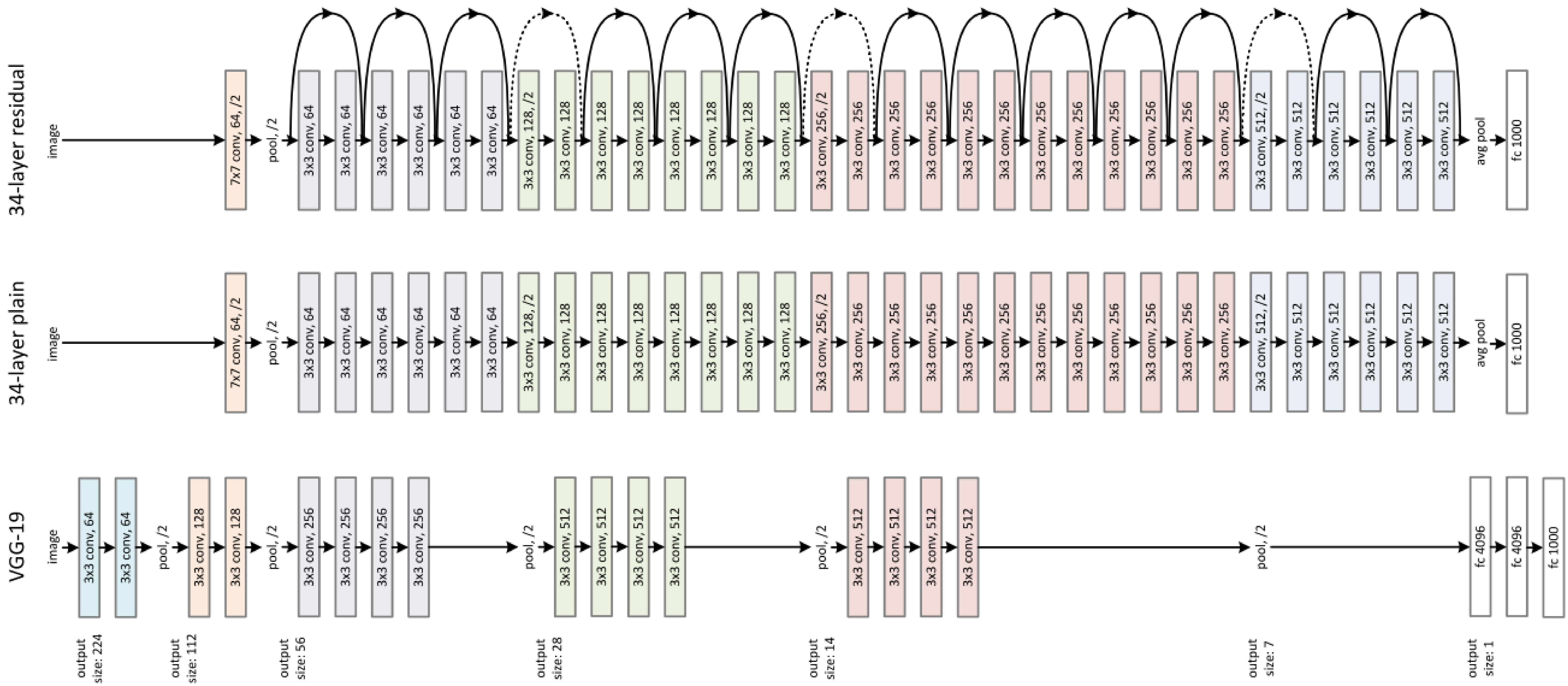
Classic Architecture - GoogLeNet



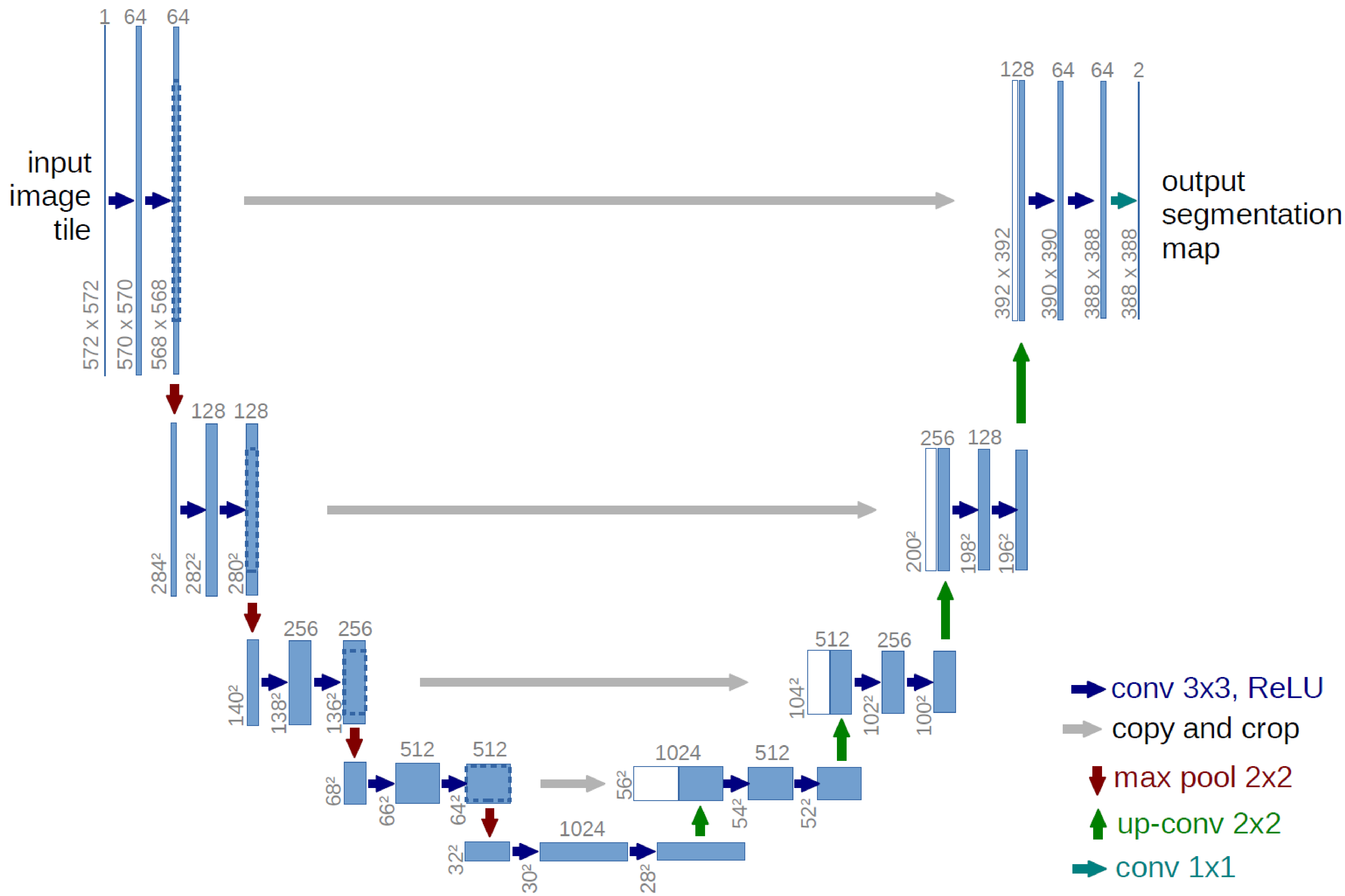
Classic Architecture - ResNet



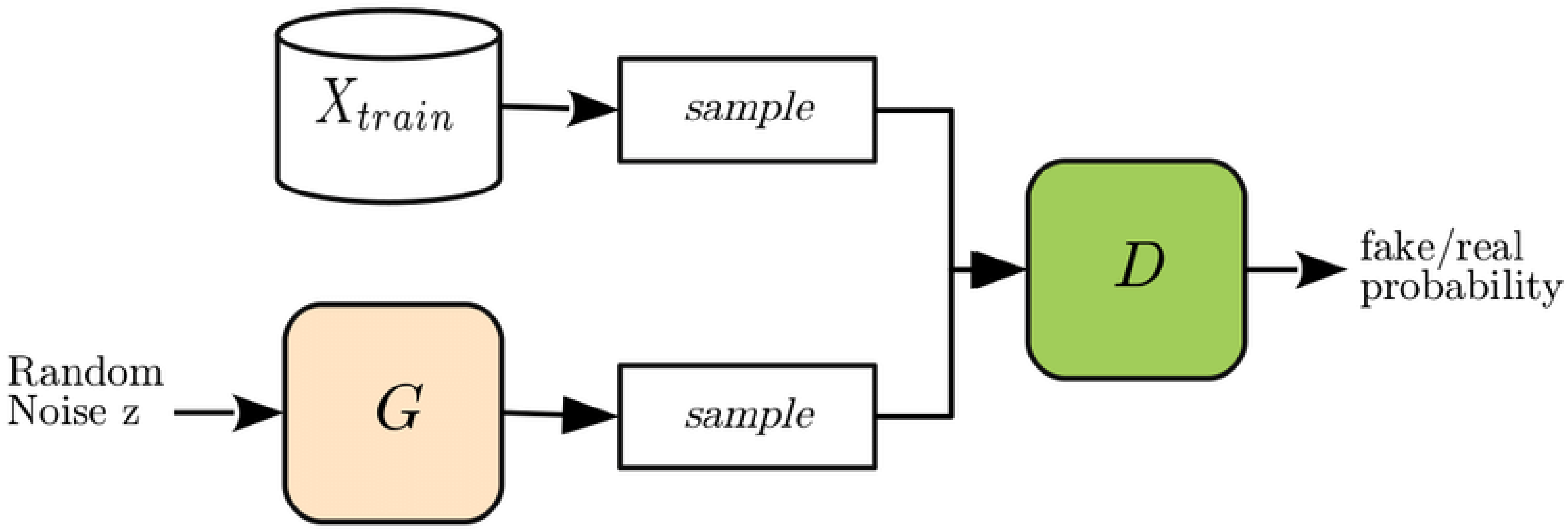
Classic Architecture - ResNet



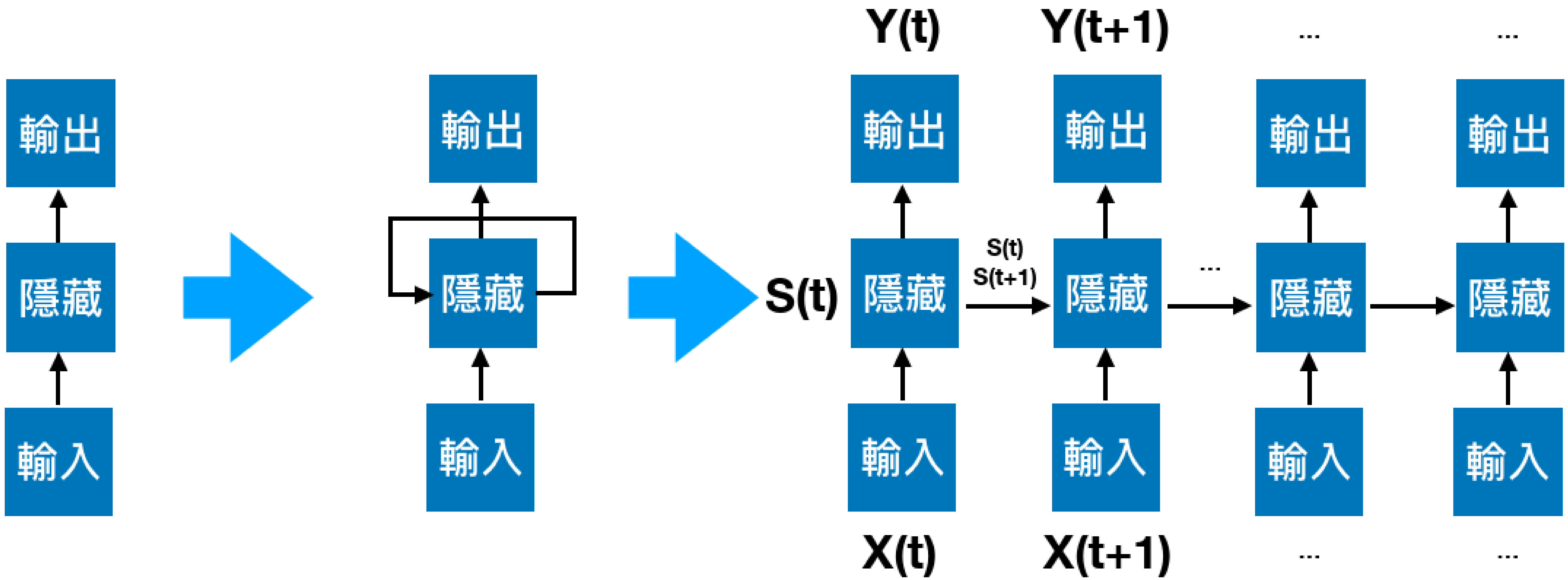
Classic Architecture - U-Net



Classic Architecture – GAN (Generative Adversarial Network)

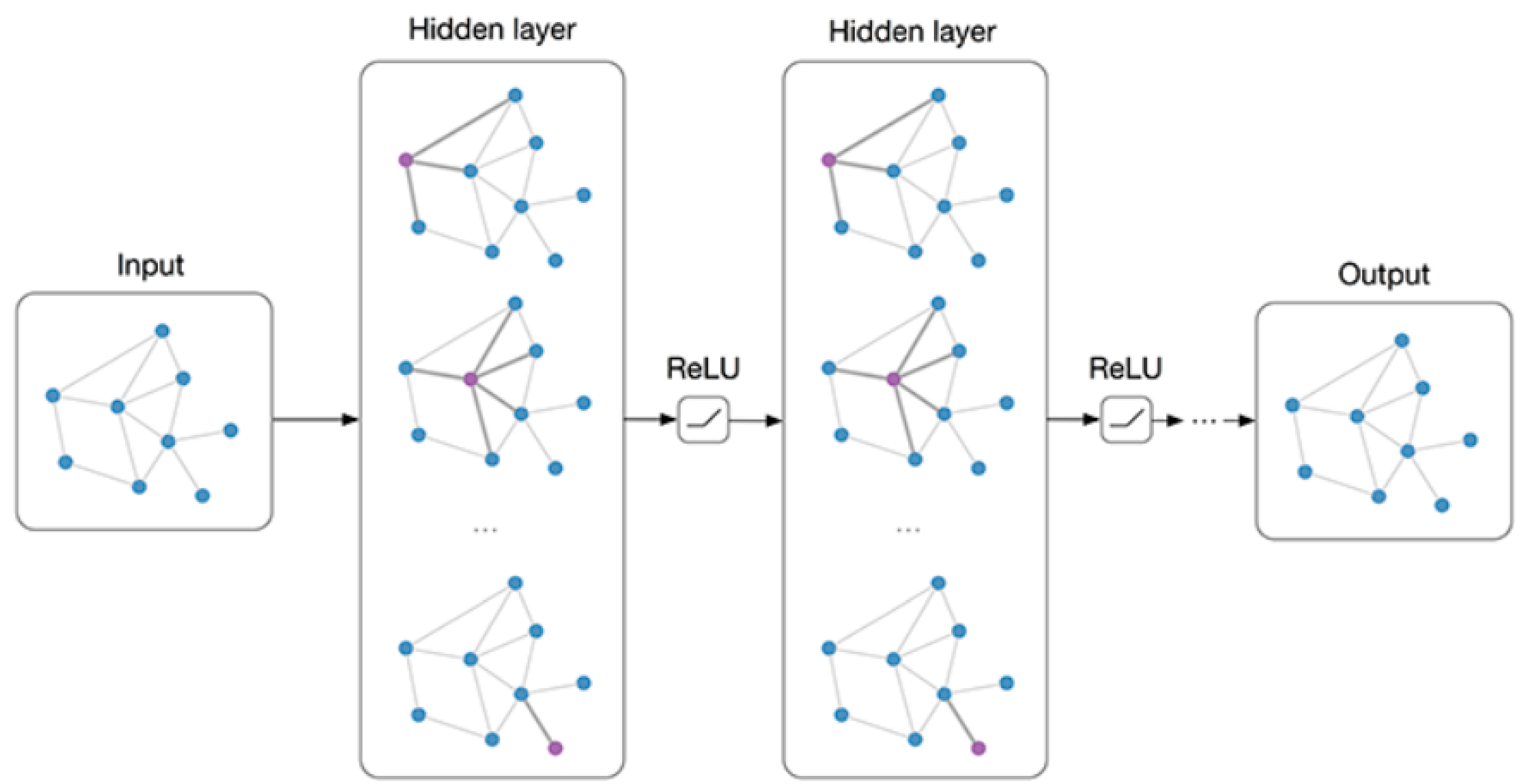


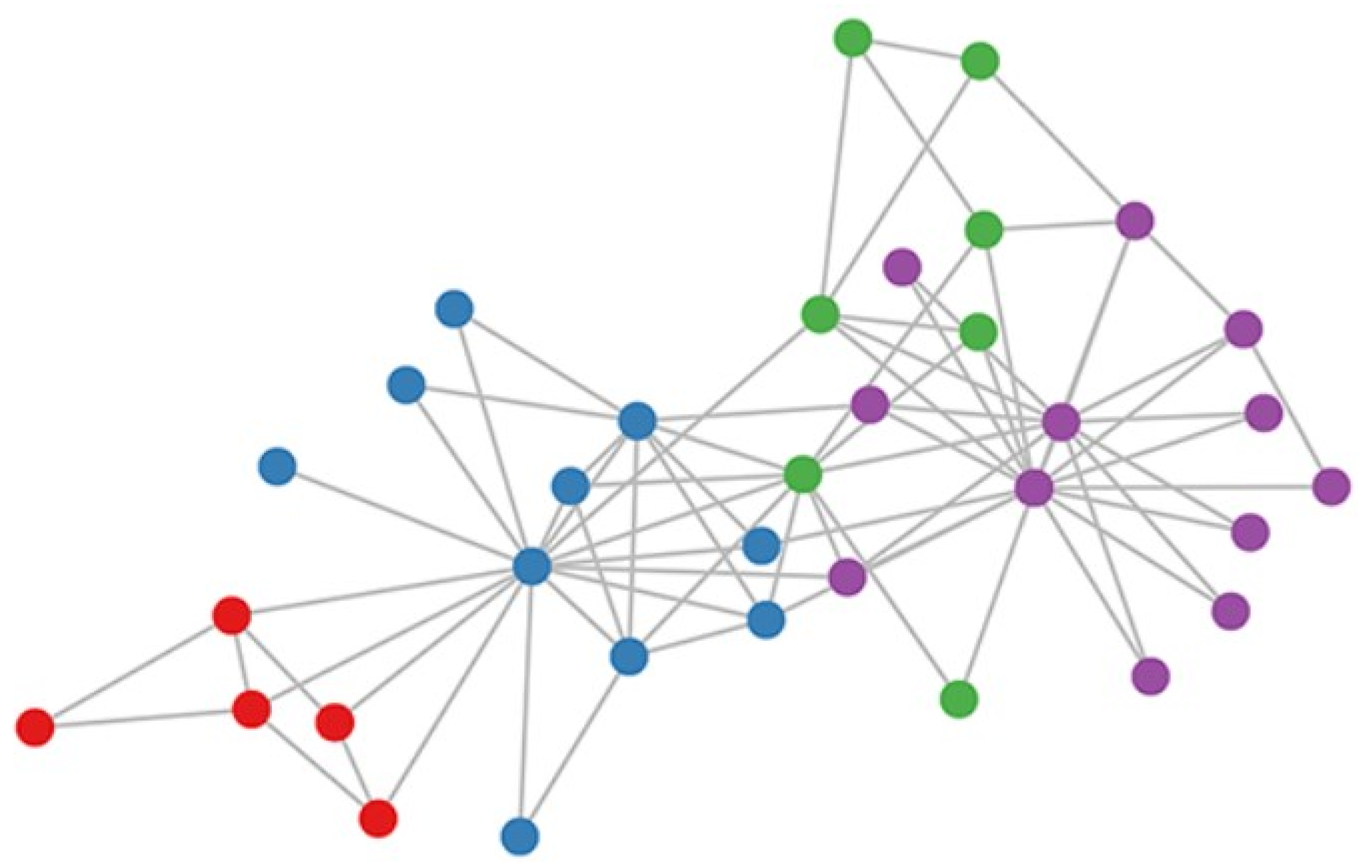
RNN



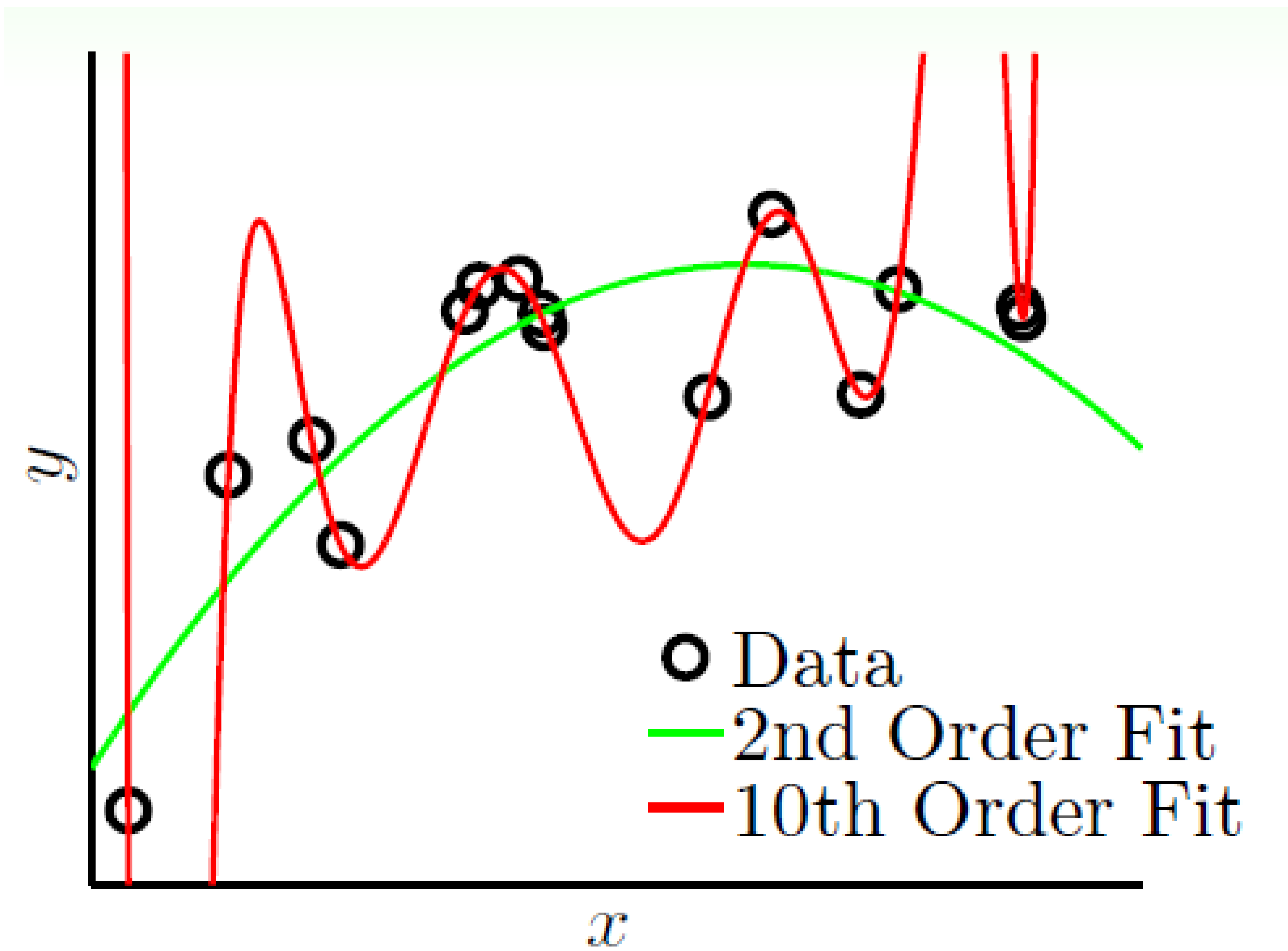
Graph CNN/RNN

- Graph Generation
- Edge Prediction
- Graph Clustering





Overfitting



Prevention of Overfitting

- Data Augmentation
- Dropout
- L1/L2 Normalization

Angular Super Resolution

$$\hat{I}_t = ((1-t)V_{t\leftarrow 0} \odot g(I_0, F_{t\rightarrow 0}) + tV_{t\leftarrow 1} \odot g(I_1, F_{t\rightarrow 1}))/((1-t)V_{t\leftarrow 0} + tV_{t\leftarrow 1})$$

$$\hat{I}_t = \frac{1}{Z} \odot \left((1-t)V_{t\leftarrow 0} \odot g(I_0, F_{t\rightarrow 0}) + tV_{t\leftarrow 1} \odot g(I_1, F_{t\rightarrow 1})\right)$$

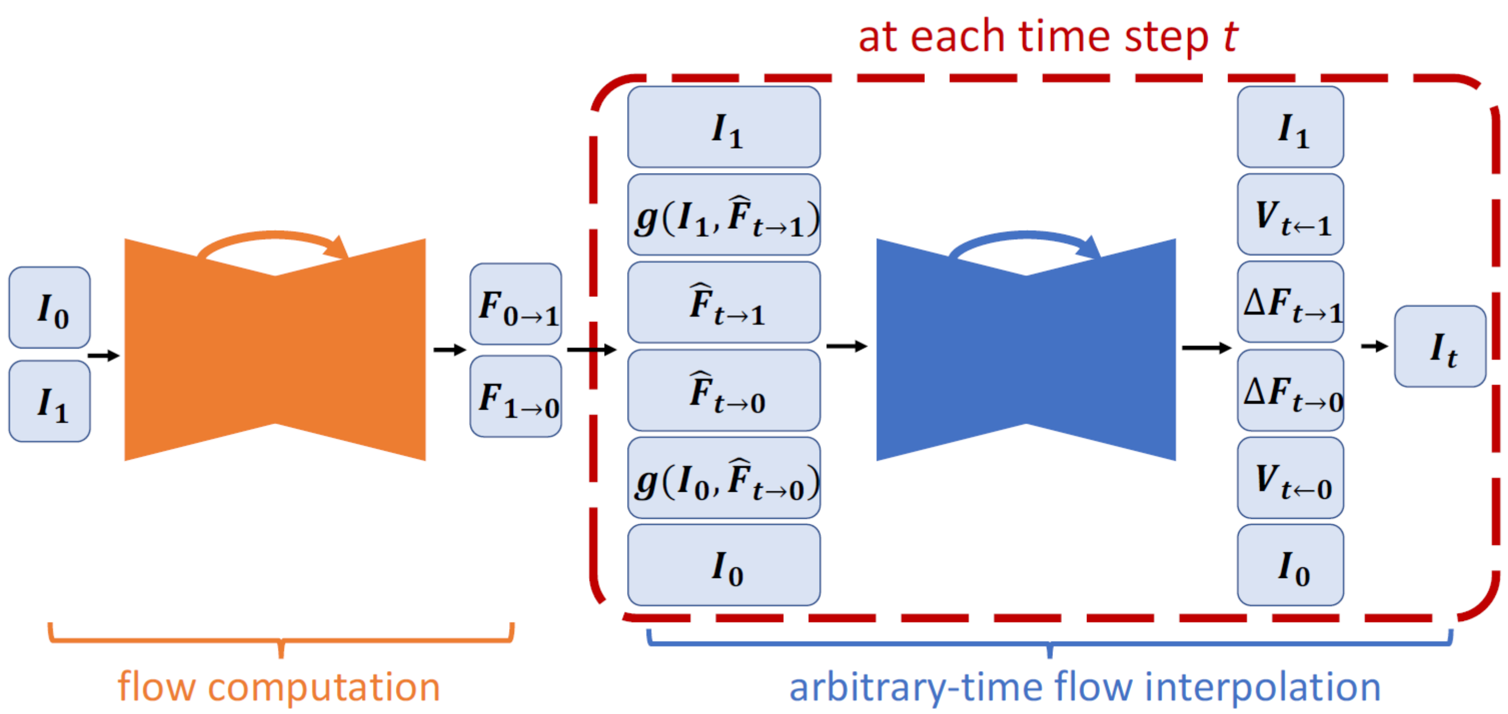
where $Z = (1-t)V_{t\rightarrow 0} + tV_{t\rightarrow 1}$

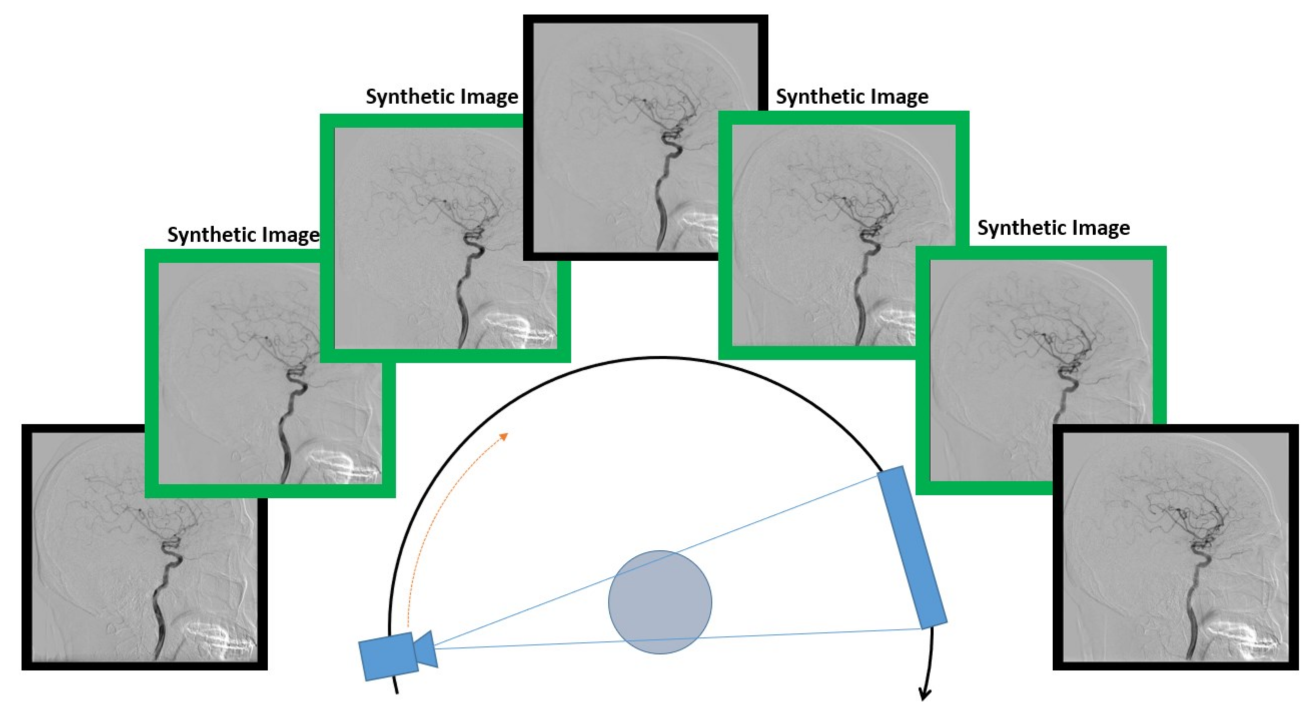
Then the approximate optical flows, $\hat{F}_{t\rightarrow 0}$ and $\hat{F}_{t\rightarrow 1}$, are calculated from $F_{0\rightarrow 1}$ and $F_{1\rightarrow 0}$ as follows:

$$\hat{F}_{t\rightarrow 0} = -(1-t)tF_{0\rightarrow 1} + t^2 F_{1\rightarrow 0}$$

$$\hat{F}_{t\rightarrow 1} = (1-t)^2 F_{0\rightarrow 1} - t(1-t)F_{1\rightarrow 0}$$

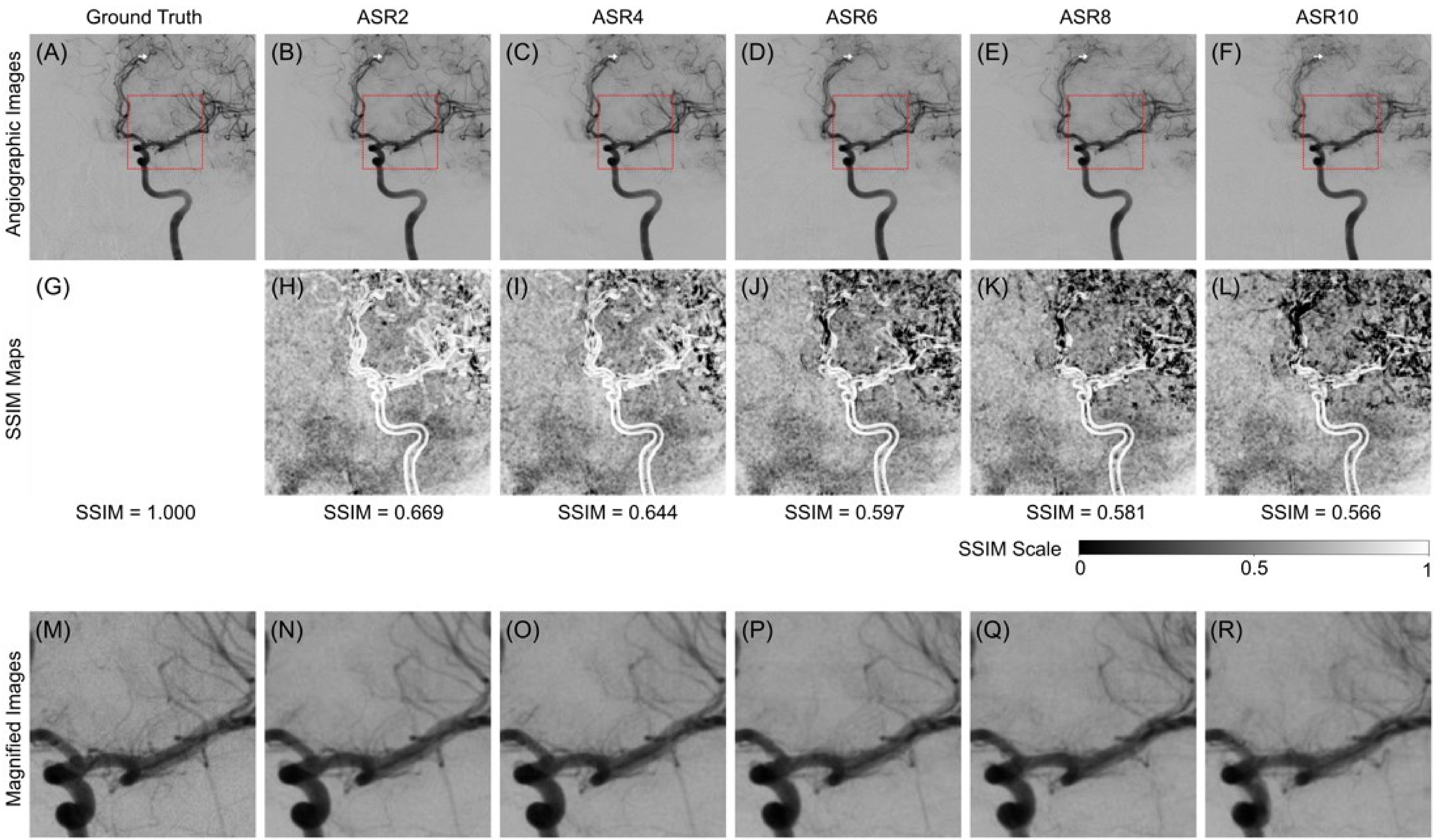
$$F_{t\rightarrow 0} = \hat{F}_{t\rightarrow 0} + \Delta F_{t\rightarrow 0}$$
$$F_{t\rightarrow 1} = \hat{F}_{t\rightarrow 1} + \Delta F_{t\rightarrow 1}$$



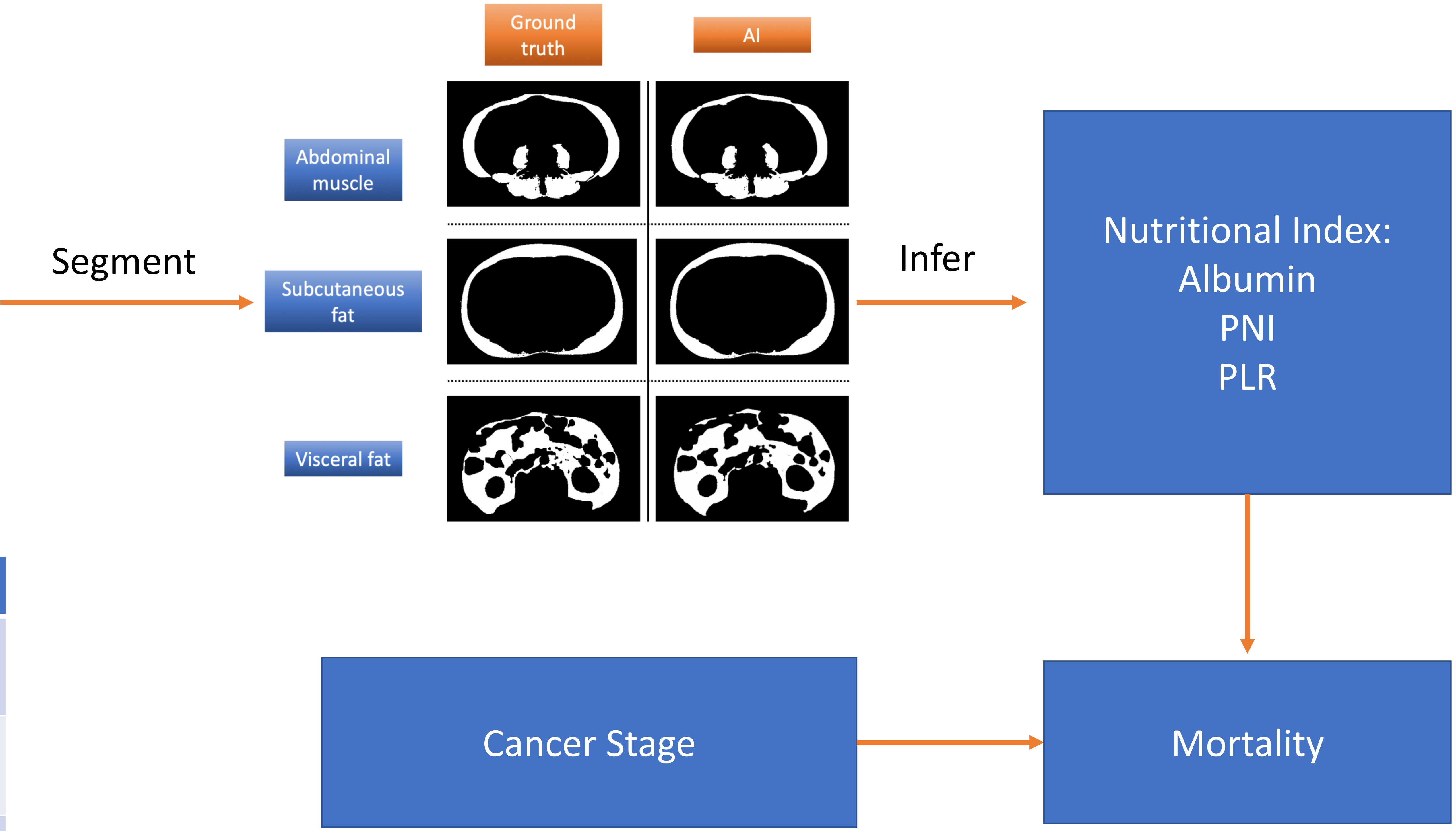


Real Frame

Real + Predicted Frame


Ground Truth
ASR2
ASR4
ASR6
ASR8
ASR10
Angiographic Images
(A)
(B)
(C)
(D)
(E)
(F)
SSIM Maps
(G)
(H)
(I)
(J)
(K)
(L)
SSIM = 1.000
SSIM = 0.669
SSIM = 0.644
SSIM = 0.597
SSIM = 0.581
SSIM = 0.566
SSIM Scale
0
0.5
1
Magnified Images
(M)
(N)
(O)
(P)
(Q)
(R)

Prediction of Nutritional Index From CT Image



Class	IoU	Dsc
Muscle	0.93	0.96
Paraspinal	0.94	0.97
Psoas	0.90	0.95
SAT	0.93	0.96
VAT	0.94	0.97

Dental Age

RetinaNet

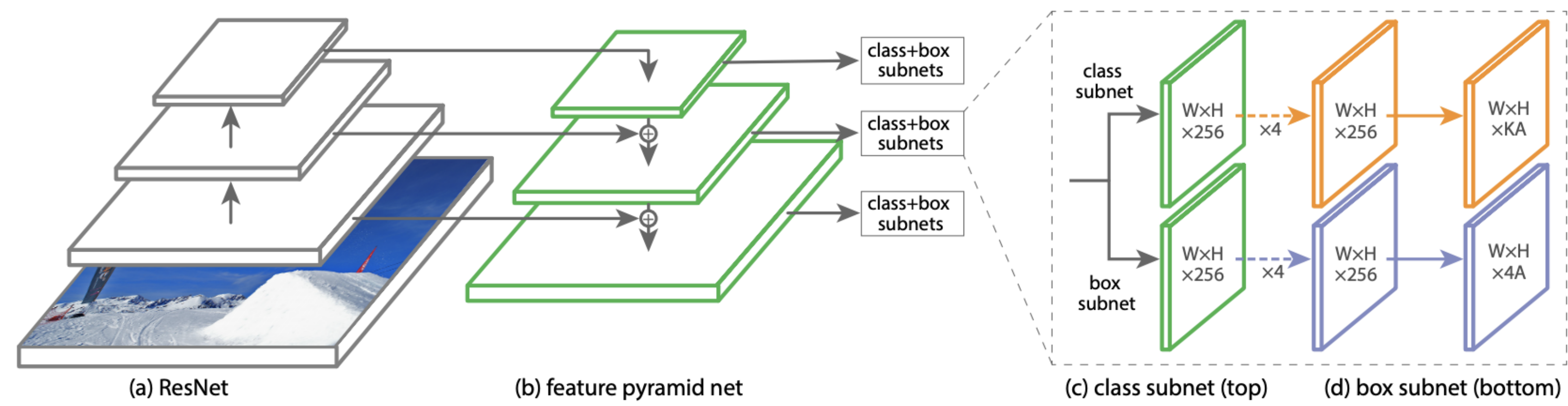
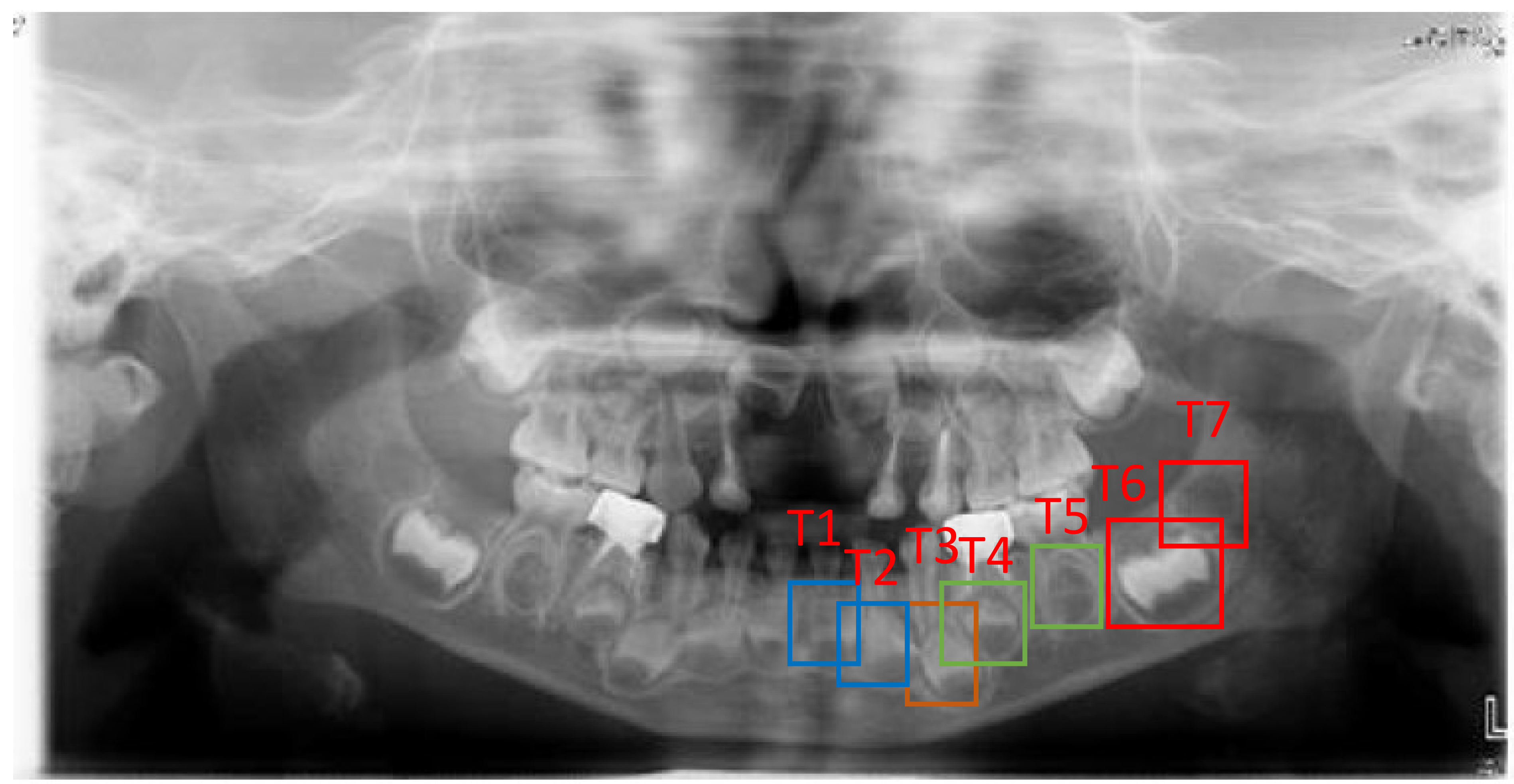
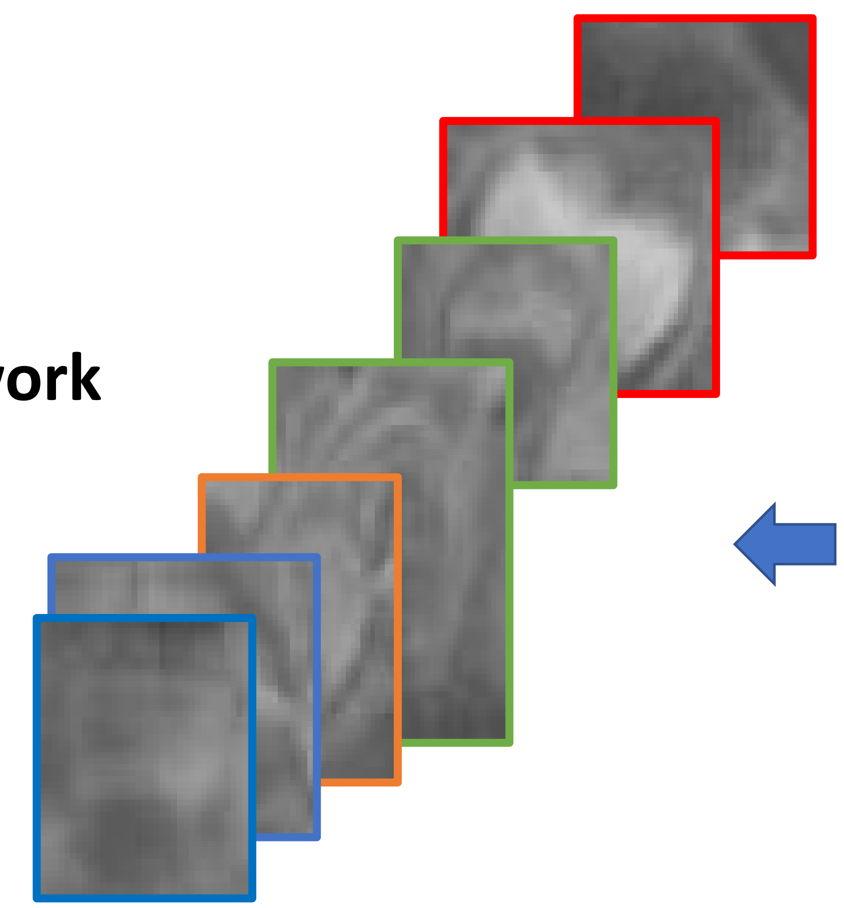


Figure 3. The one-stage **RetinaNet** network architecture uses a Feature Pyramid Network (FPN) [20] backbone on top of a feedforward ResNet architecture [16] (a) to generate a rich, multi-scale convolutional feature pyramid (b). To this backbone RetinaNet attaches two subnetworks, one for classifying anchor boxes (c) and one for regressing from anchor boxes to ground-truth object boxes (d). The network design is intentionally simple, which enables this work to focus on a novel focal loss function that eliminates the accuracy gap between our one-stage detector and state-of-the-art two-stage detectors like Faster R-CNN with FPN [20] while running at faster speeds.





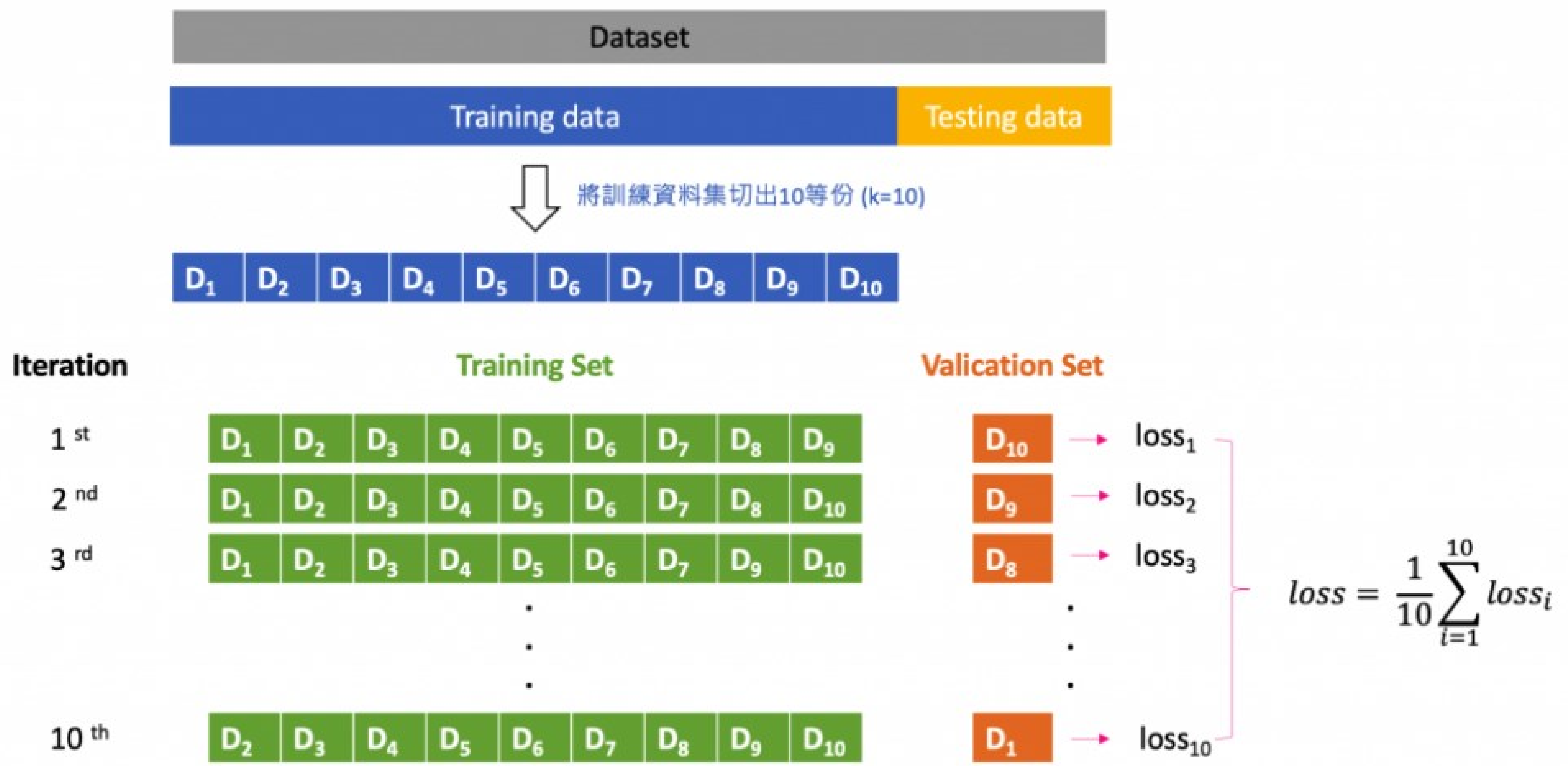
Siamese network

Staging for each tooth

Dental age

Dataset-the key is distribution

- Train, Valid, Test



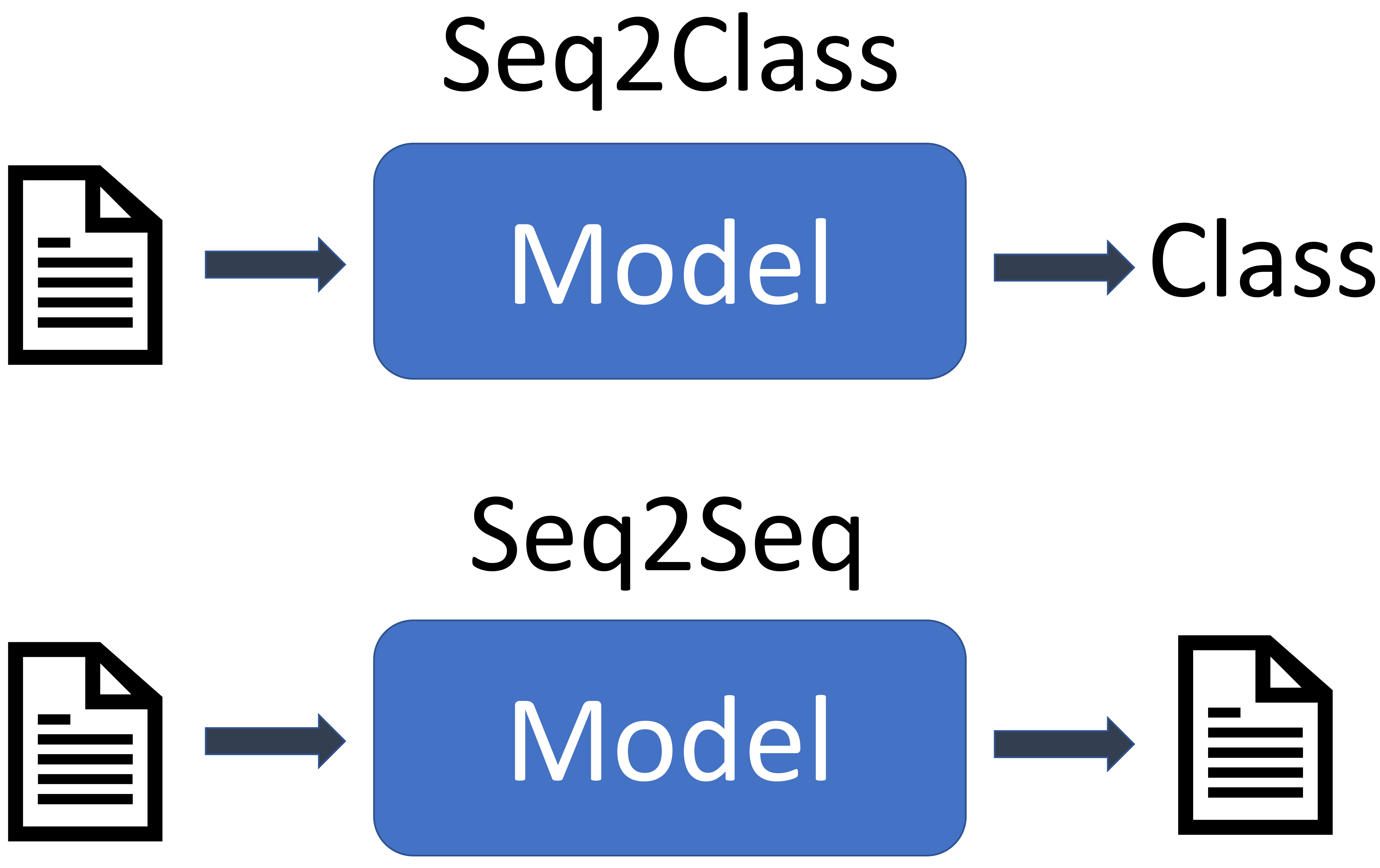
- K-fold cross validation
  - Repeat K-fold
  - Group K-fold
  - Nested K-fold
  - Stratified K-Fold

到底資料量需要多少? 經驗上來說…

- Very good performance: 50000
- Good performance: 10000
- Mediocre: 5000
- Minimum: 1000 ↑

Natural Language Processing

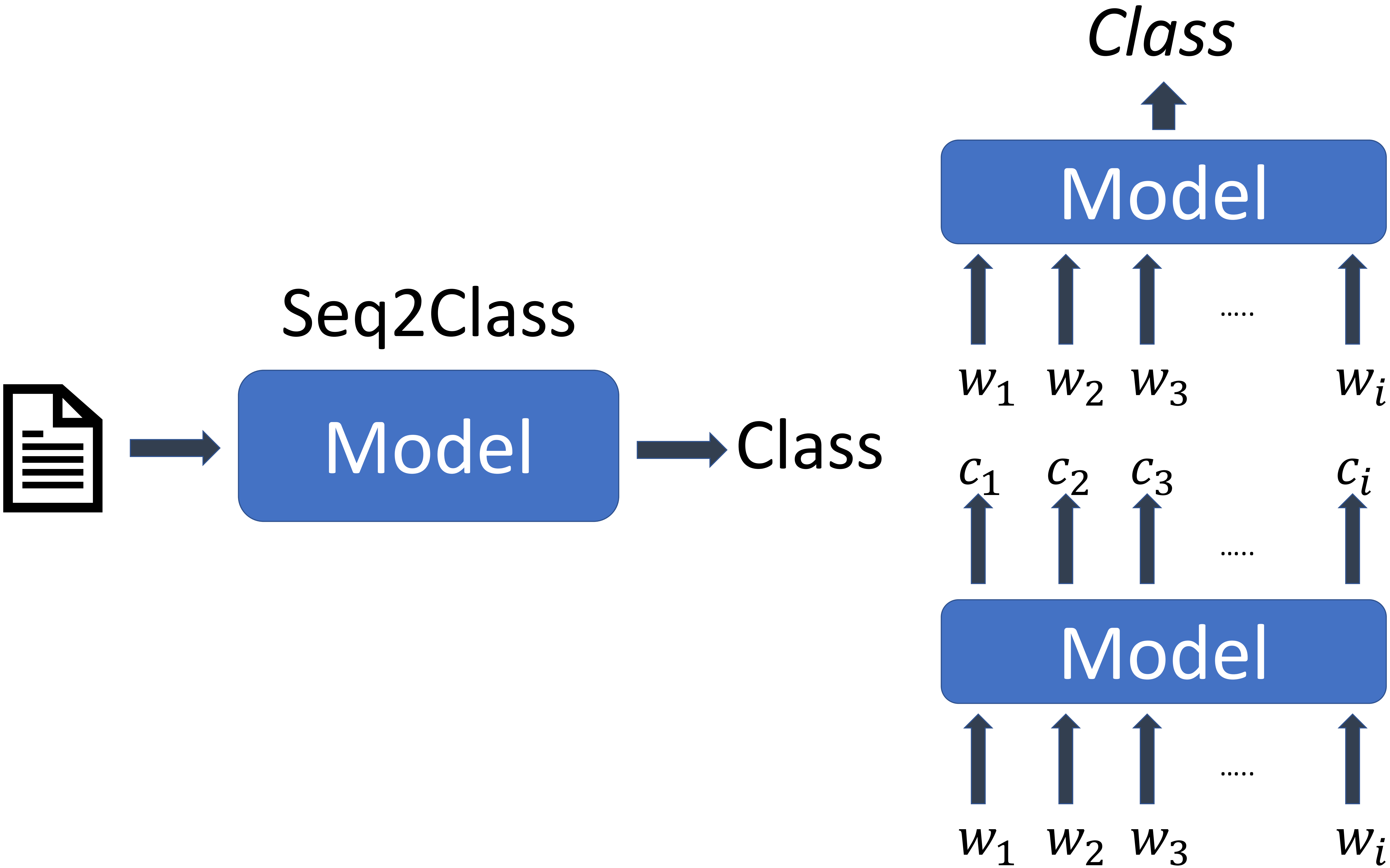
- NLP Tasks
- Transformer
- BERT
- GPT



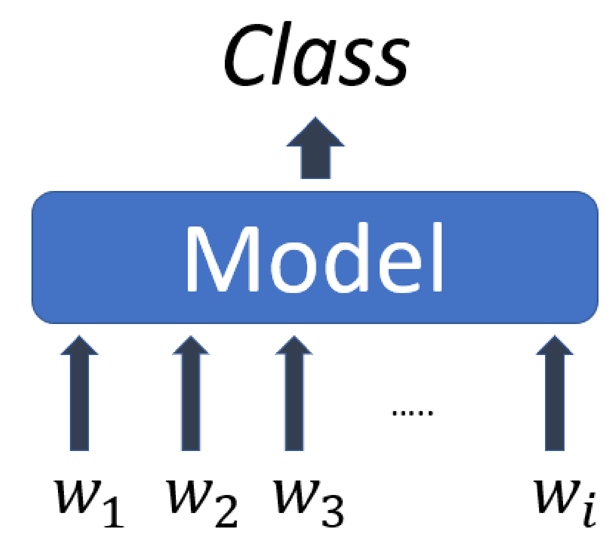
Seq2Class
Model
Class
Seq2Seq
Model

	One Sequence	Multiple Sequences
One Class	Sentiment Classification Stance Detection Veracity Prediction Intent Classification Dialogue Policy	NLI Search Engine Relation Extraction
Class for each Token	POS tagging Word segmentation Extraction Summarization Slotting Filling NER	
Copy from Input		Extractive QA
General Sequence	Abstractive Summarization Translation Grammar Correction NLG	General QA Task Oriented Dialogue Chatbot
Other?	Parsing, Coreference Resolution	

Source of table: Dr. Lee Hung-Yi


Seq2Class
Model
Class
Class
Model
w_1 w_2 w_3 ..... w_i
c_1 c_2 c_3 ..... c_i
Model
w_1 w_2 w_3 ..... w_i

Sentiment Classification



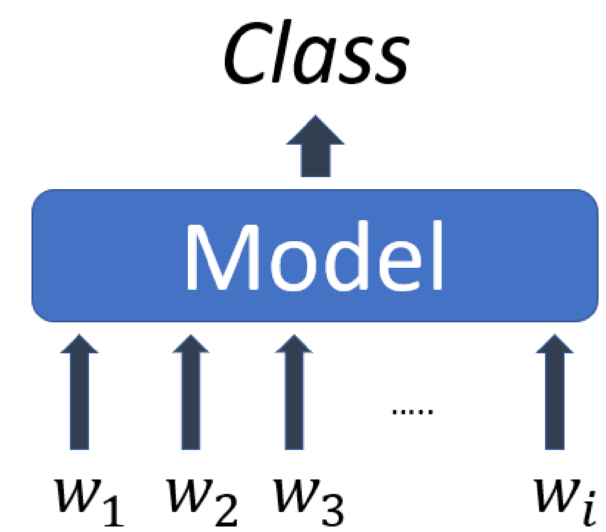
機器人大戰30，還蠻好玩的 => Positive

機器人大戰30，戰鬥動畫都沿用舊的 => Negative

機器人大戰30雖然戰鬥動畫都沿用舊的，但還蠻好玩的 => Positive

機器人大戰30雖然還蠻好玩的，但戰鬥動畫都沿用用舊的 => Negative

Stance Detection

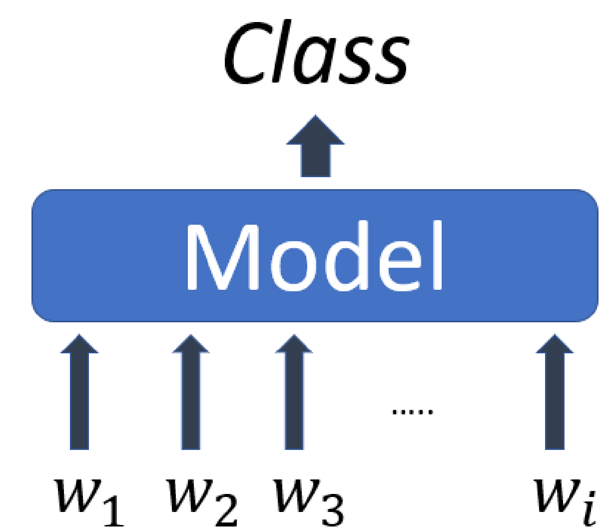


Post: 機器人大戰30是神作

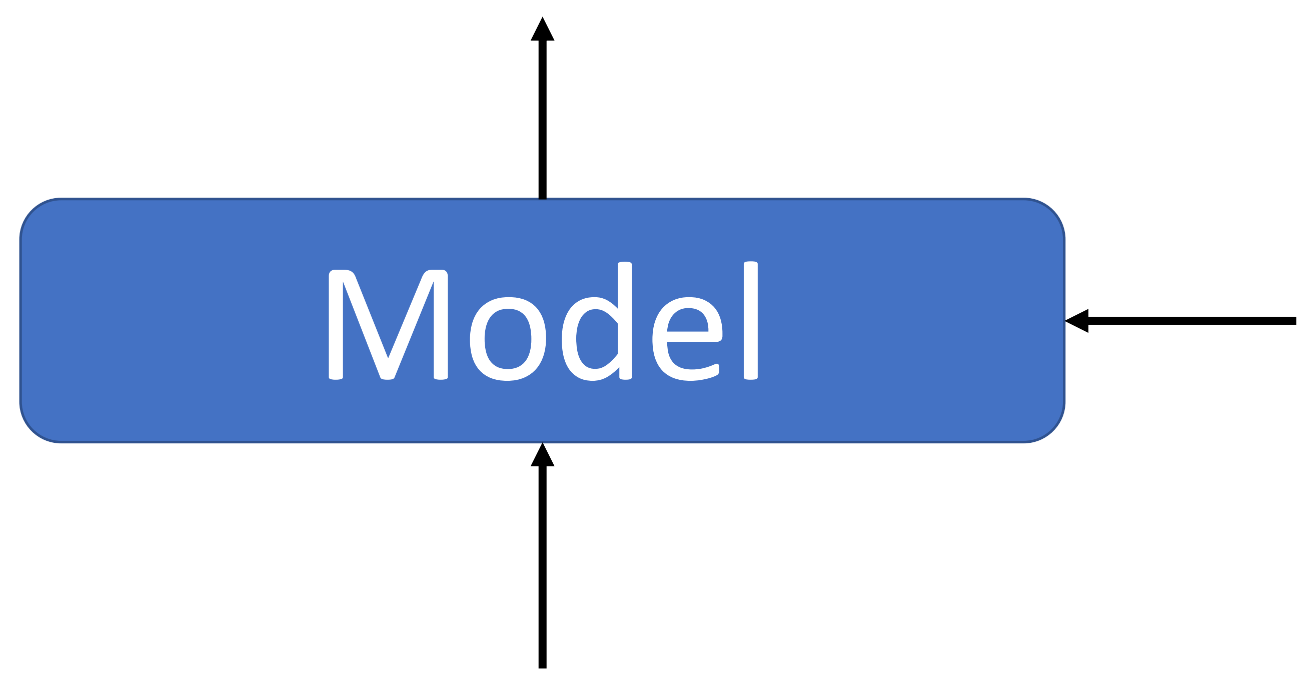
Reply: 這次用那麼多前作的戰鬥動畫，超沒誠意的好嗎。

Classification: Support, Denying, Querying, Commenting

Natural Language Inference (NLI)



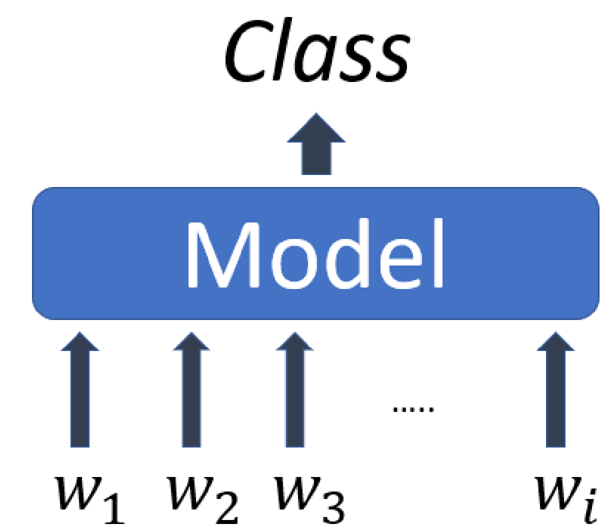
Contradiction
Entailment
Neutral

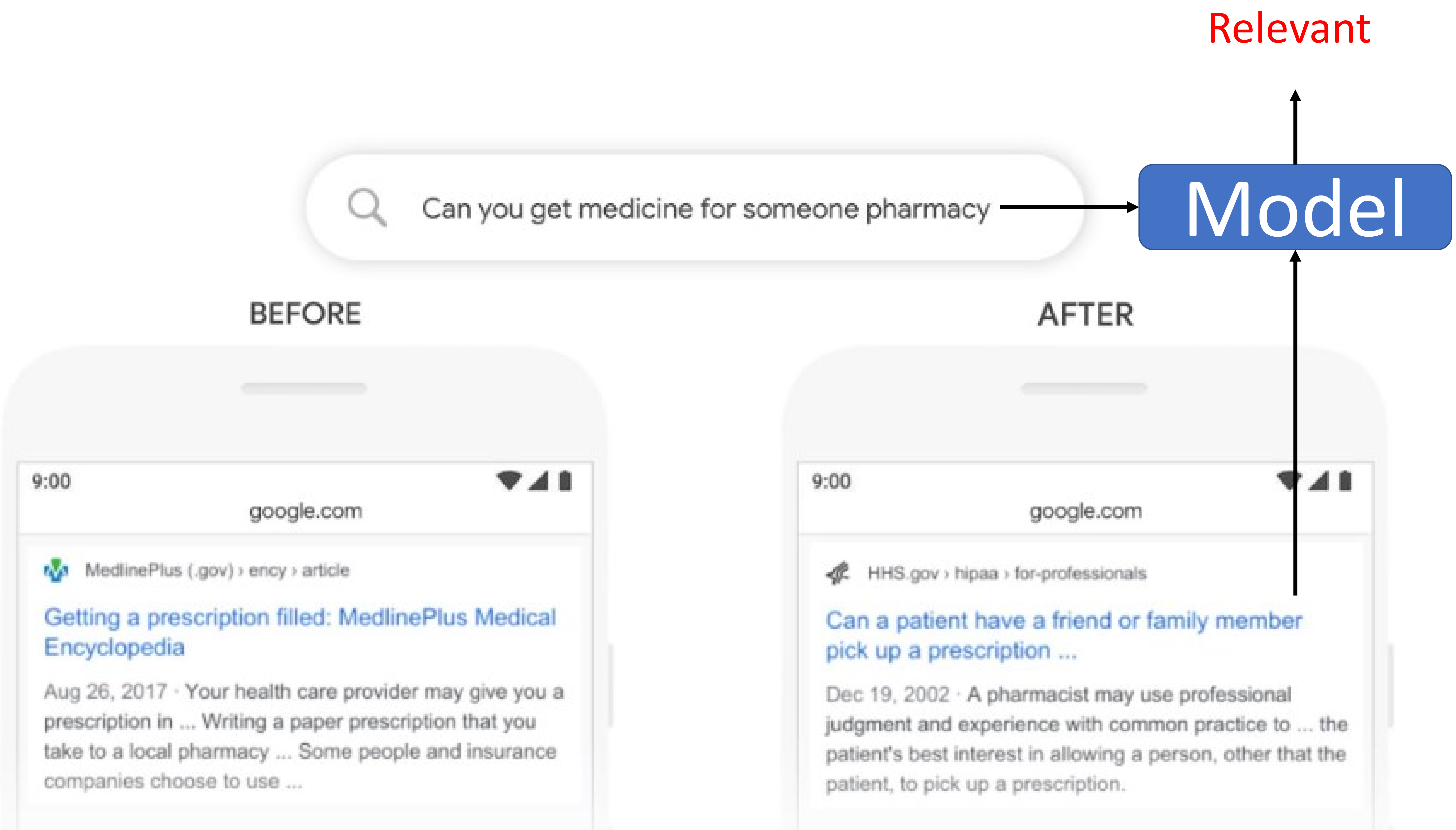


Premise: A person on a horse jumps over a broken down airplane.

Contradiction	Hypothesis: A person is at diner
Entailment	Hypothesis: A person is outdoor, on a horse.
Neutral	Hypothesis: A person is training his horse for a competition.

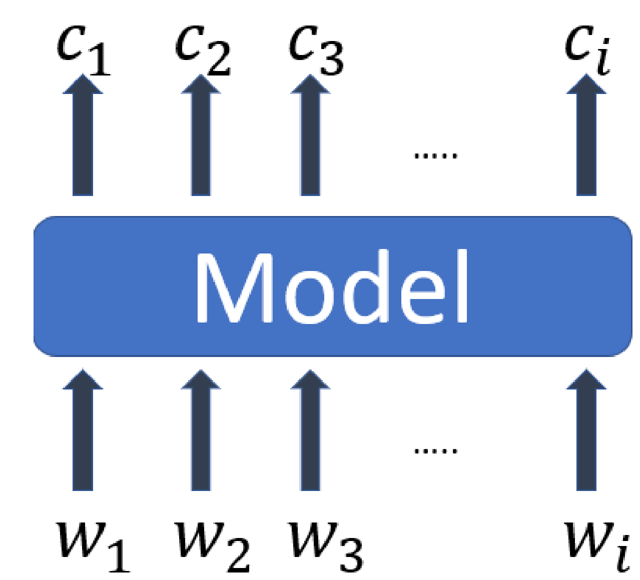
Search Engine

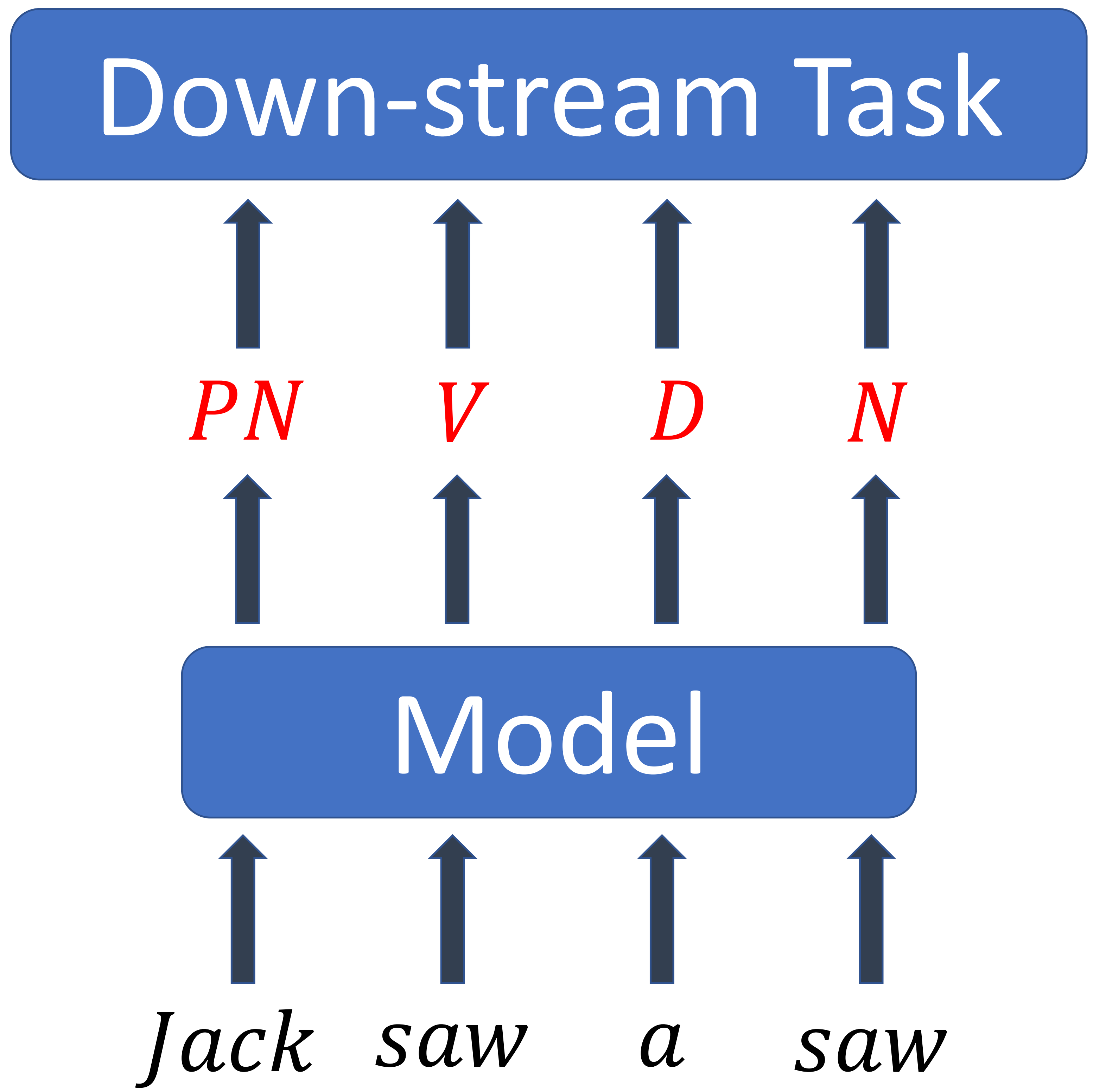




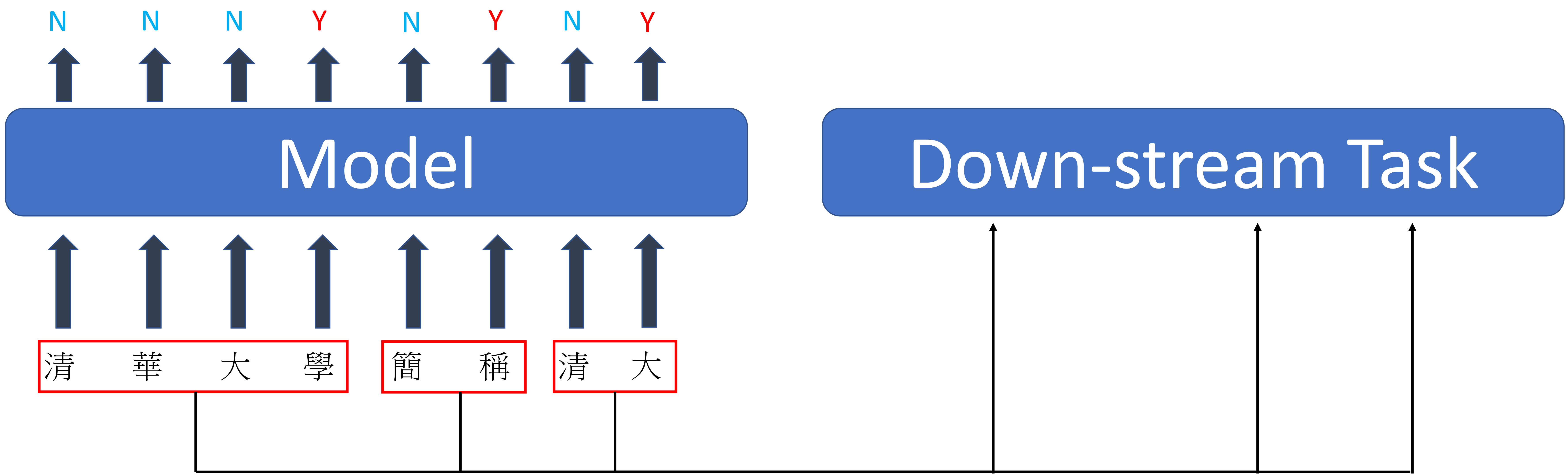
https://www.blog.google/products/search/search-language-understanding-bert/

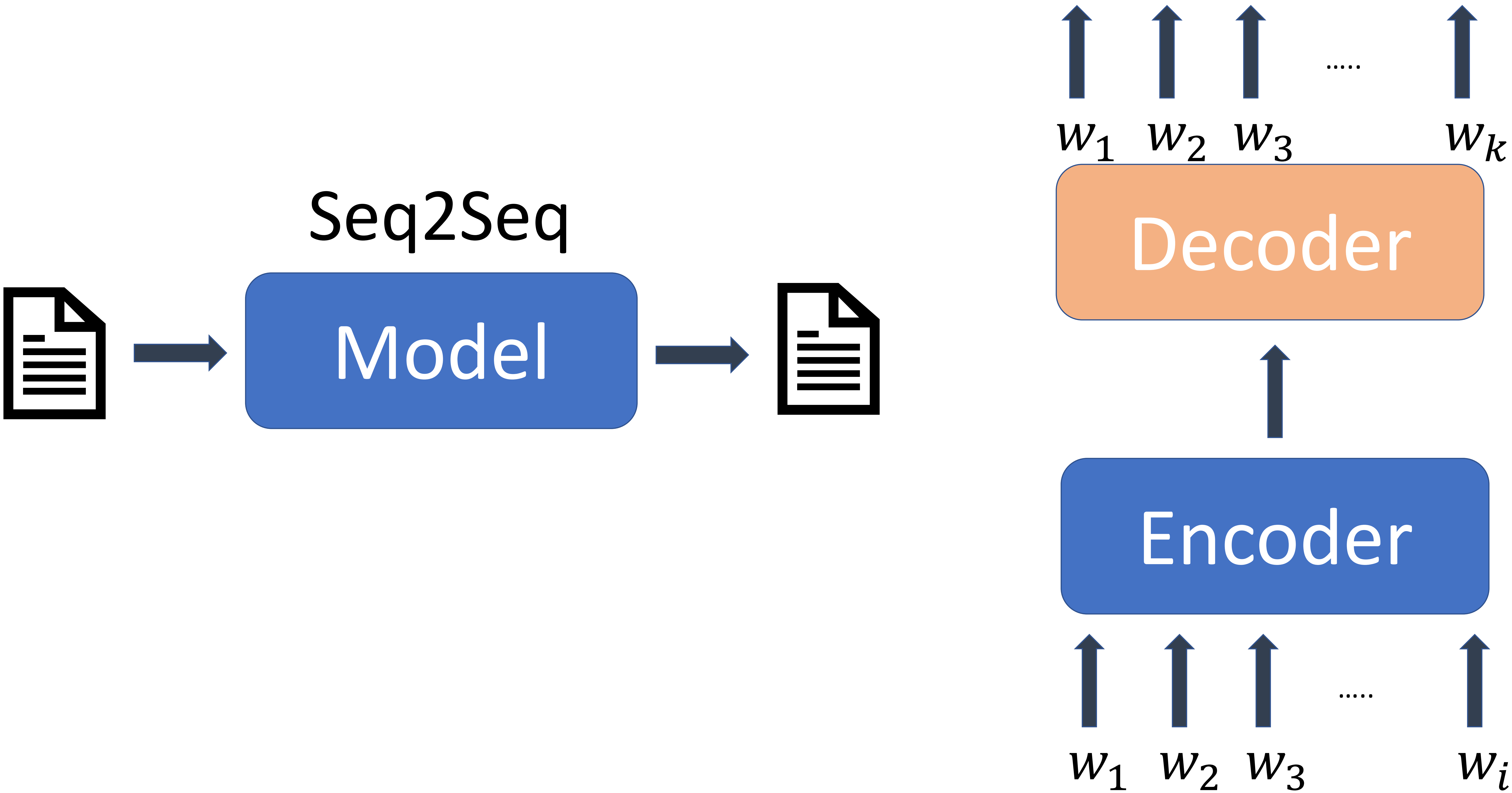
POS (Part-of-Speech) Tagging



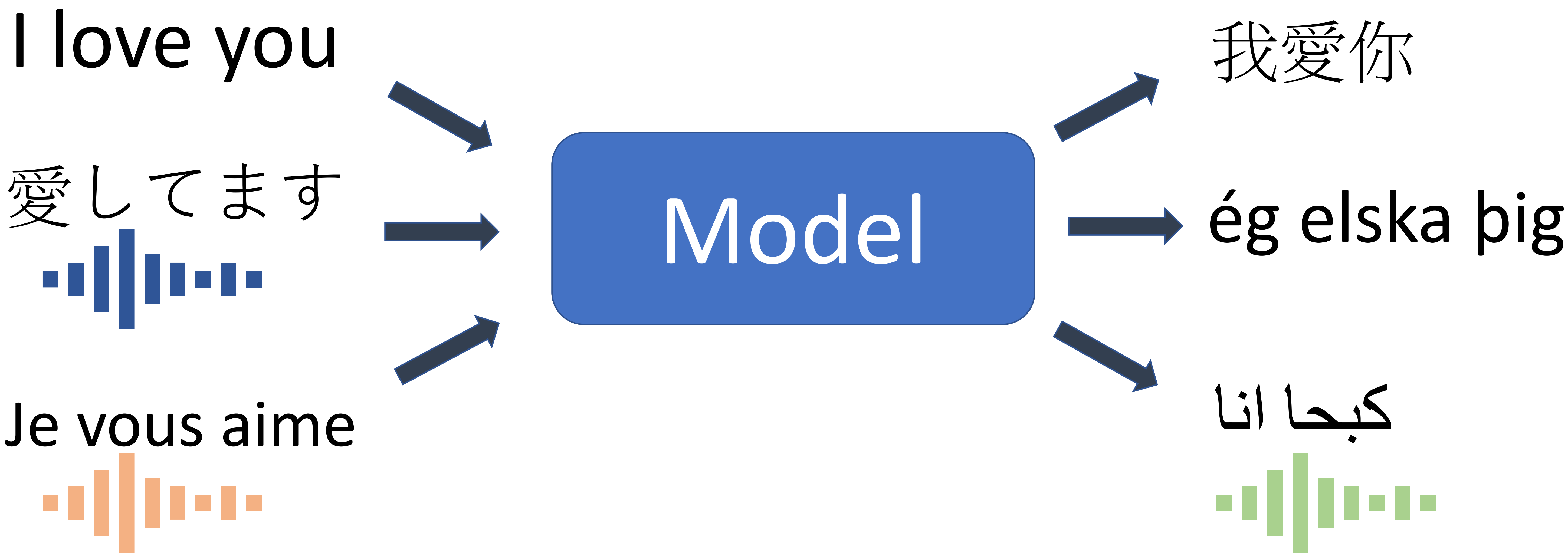


Word Segmentation

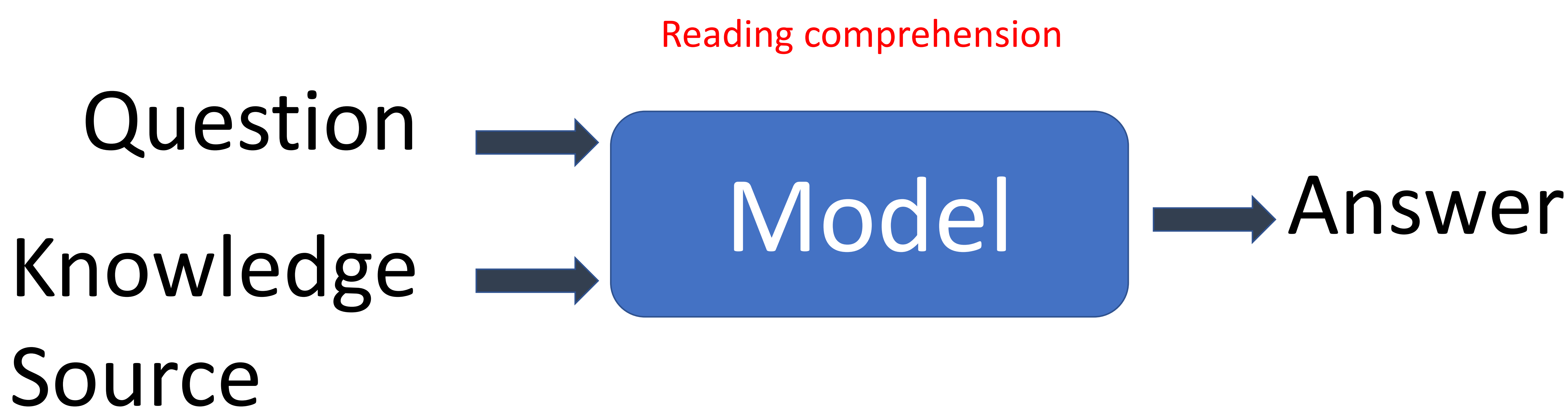



Seq2Seq
Model
w_1 w_2 w_3 ..... w_k
Decoder
Encoder
.....
w_1 w_2 w_3 w_i

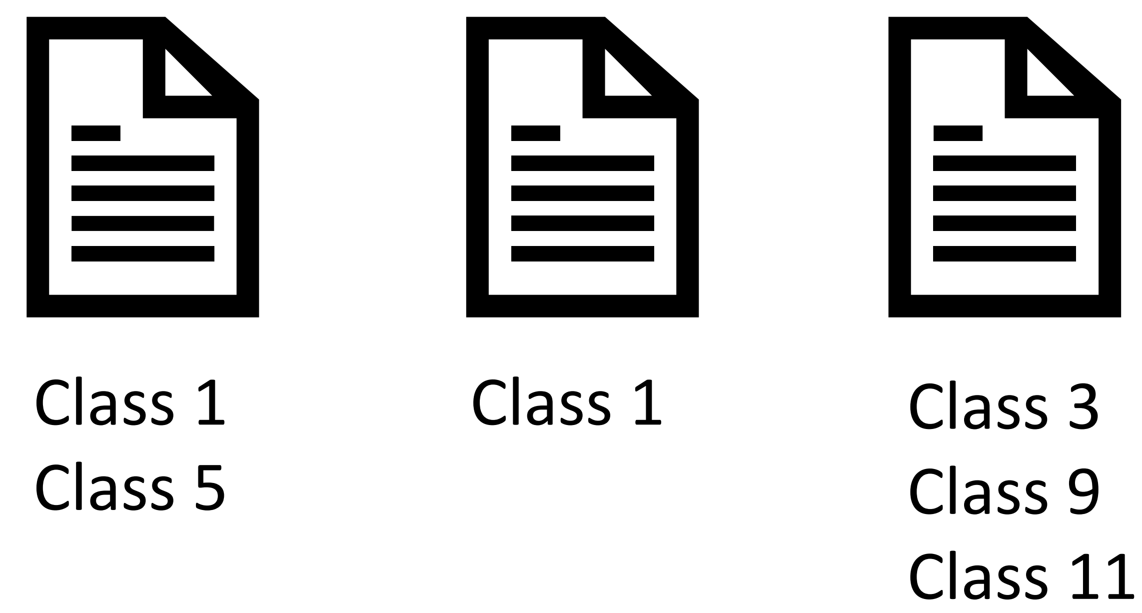
Machine Translation

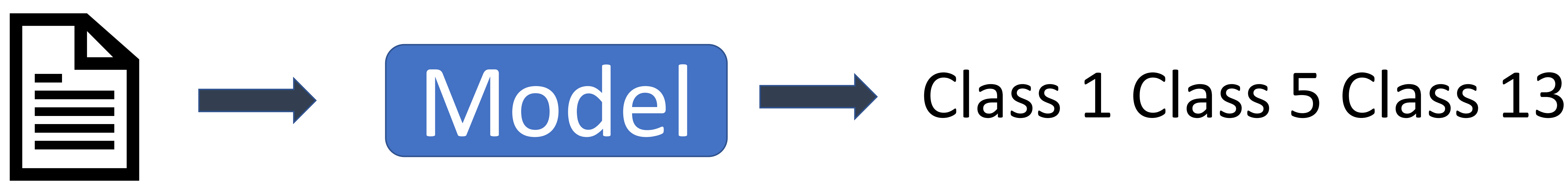


Question Answering



Multi-label Classification



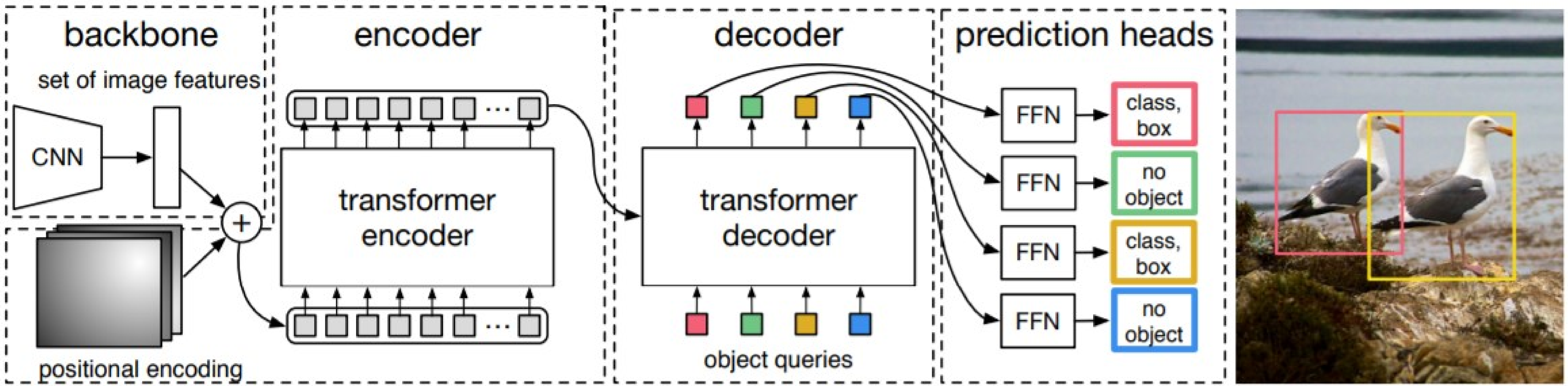


https://arxiv.org/abs/1909.03434

https://arxiv.org/abs/1707.05495

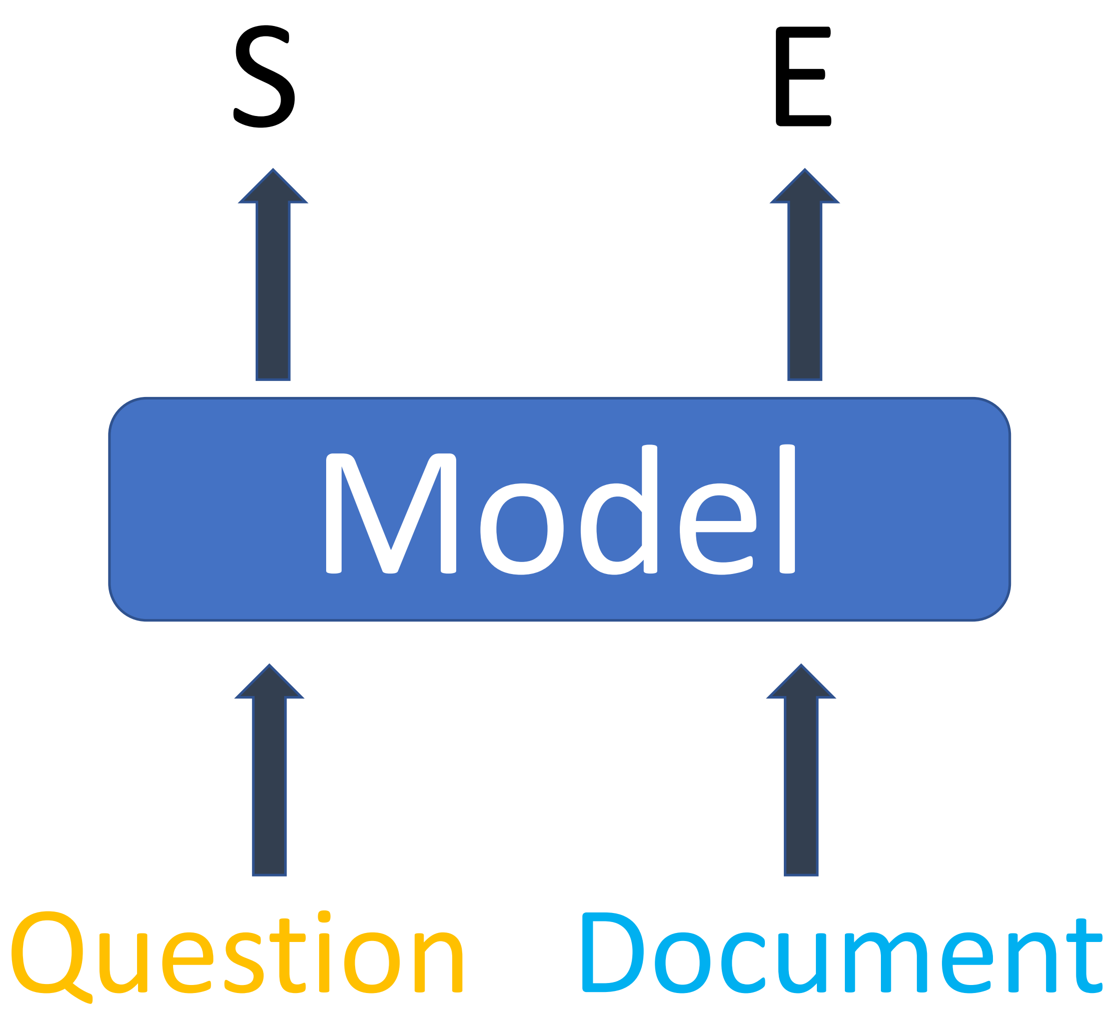
Object Detection

Seq2Seq



https://arxiv.org/abs/2005.12872

Extractive QA



In meteorology, precipitation is any product of the condensati 17 f atmospheric water vapor that falls under gravity. The main forms of precipitation include drizzle, rain, sleet, snow, graupel and hail... Precipitation forms as smaller droplets coalesce vi 77 llision 79 other rain drops or ice crystals within a cloud. Short, intense periods of rain in scattered locations are called "showers".

What causes precipitation to fall?
gravity S= 17, E=17

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupel

Where do water droplets collide with ice crystals to form precipitation?
within a cloud S= 77, E=79

Chatting

cleverbot

https://www.cleverbot.com/

User: Sorry for what?

Cleverbot: For not remembering your name.

User: Do you remember your name?

Cleverbot: Yes I do...

User: What is your name?

Cleverbot: Once again, I can't tell you.

User: You have told me that you are Mara.

Cleverbot: I am a boy.

User: Are you a boy?

Cleverbot: Yes.

User: Are you a girl?

Cleverbot: No, are you?

User: No. How old are you?

Cleverbot: 2000 years old.

User: You are so old!

Cleverbot: No i'm quite young.

User: How old are you?

Cleverbot: Why do you keep asking me that?

User: Because I want to know your age.

Cleverbot: What if I do not want to tell you.

User: Then don't tell me your age.

Cleverbot: My name is Blue Monday.

Source of Example: Dr. Lee Hung-Yi

GLUE General Language Understanding Evaluation

Sentiment Classification

Semantic Similarity

Natural Language Inference

GLUE Tasks

Name	Download	More Info	Metric
The Corpus of Linguistic Acceptability			Matthew's Corr
The Stanford Sentiment Treebank			Accuracy
Microsoft Research Paraphrase Corpus			F1 / Accuracy
Semantic Textual Similarity Benchmark			Pearson-Spearman Corr
Quora Question Pairs			F1 / Accuracy
MultiNLI Matched			Accuracy
MultiNLI Mismatched			Accuracy
Question NLI			Accuracy
Recognizing Textual Entailment			Accuracy
Winograd NLI			Accuracy
Diagnostics Main			Matthew's Corr

Chinese version https://www.cluebenchmarks.com/

https://gluebenchmark.com/tasks

SuperGLUE Tasks

Name	Identifier	Download	More Info	Metric
Broadcoverage Diagnostics	AX-b			Matthew's Corr
CommitmentBank	CB			Avg. F1 / Accuracy
Choice of Plausible Alternatives	COPA			Accuracy
Multi-Sentence Reading Comprehension	MultiRC			F1a / EM
Recognizing Textual Entailment	RTE			Accuracy
Words in Context	WiC			Accuracy
The Winograd Schema Challenge	WSC			Accuracy
BoolQ	BoolQ			Accuracy
Reading Comprehension with Commonsense Reasoning	ReCoRD			F1 / Accuracy
Winogender Schema Diagnostics	AX-g			Gender Parity / Accuracy

https://super.gluebenchmark.com/tasks

DecaNLP

- 10 NLP Tasks
- 所有Task皆可被視為QA

Examples

Question	Context	Answer	
What is a major importance of Southern California in relation to California and the US?	...Southern California is a major economic center for the state of California and the US....	major economic center	Extractive QA
What is the translation from English to German?	Most of the planet is ocean water.	Der Großteil der Erde ist Meerwasser	Translation
What is the summary?	Harry Potter star Daniel Radcliffe gains access to a reported £320 million fortune...	Harry Potter star Daniel Radcliffe gets £320M fortune...	Summary
Hypothesis: Product and geography are what make cream skimming work. Entailment, neutral, or contradiction?	Premise: Conceptually cream skimming has two basic dimensions – product and geography.	Entailment	NLI
Is this sentence positive or negative?	A stirring, funny and finally transporting re-imagining of Beauty and the Beast and 1930s horror film.	positive	Sentiment classification

https://decanlp.com/

Natural Language Processing

- NLP Tasks
- Transformer
- BERT
- GPT

Input Is a Set of Vector

One-hot Encoding

Apple = [1 0 0 0 0 0 0 0 0 .....]

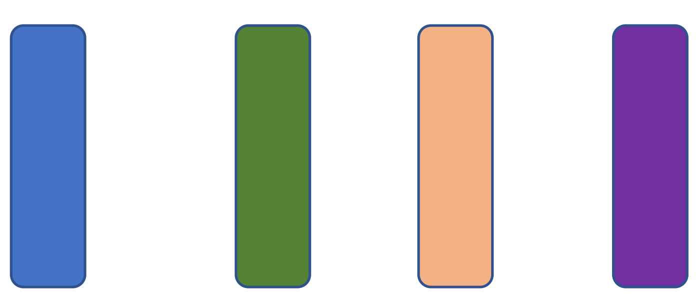
Ball = [0 1 0 0 0 0 0 0 0 ......]

Call = [0 0 1 0 0 0 0 0 0 ......]

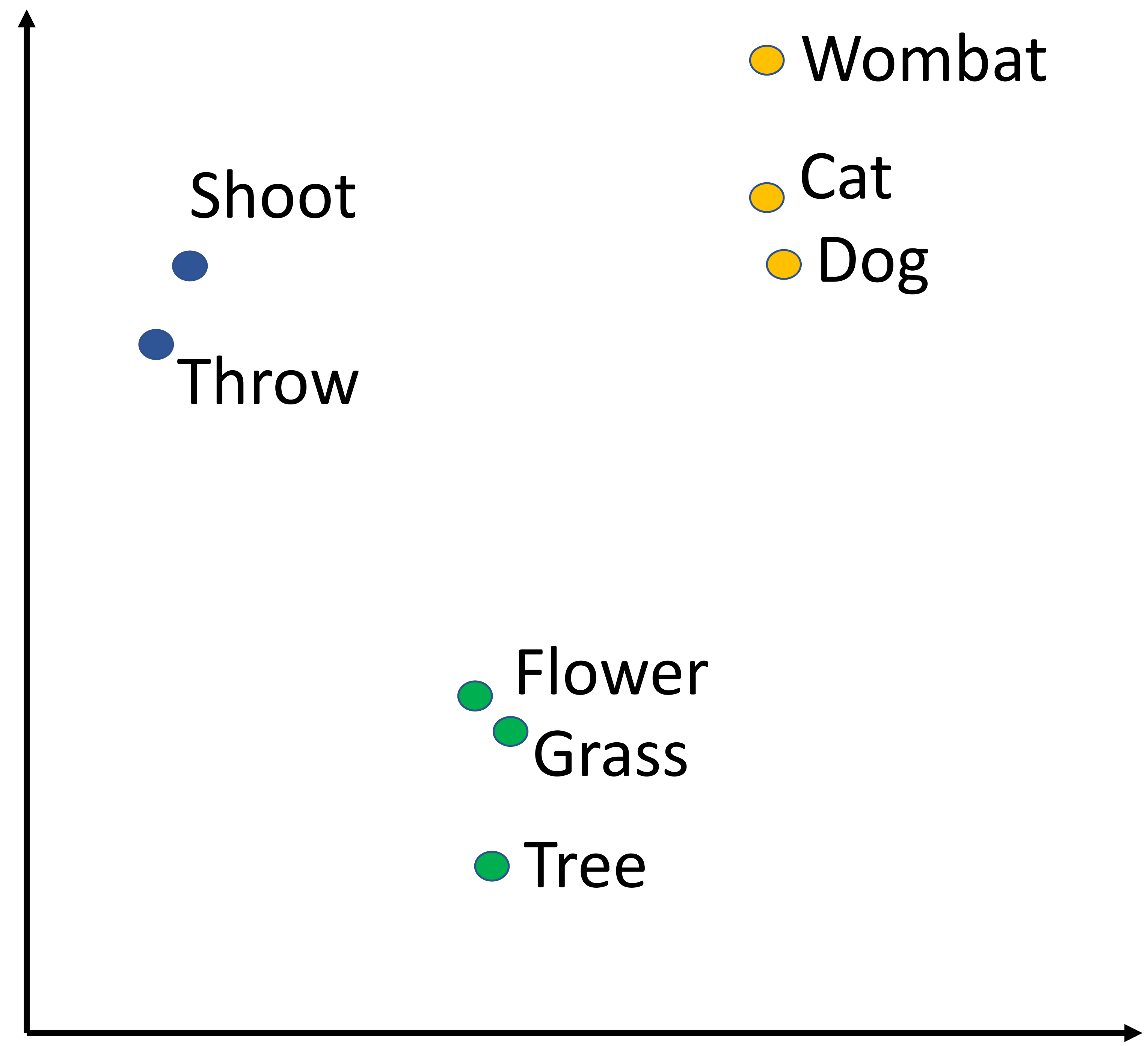
Dog = [0 0 0 1 0 0 0 0 0 ......]

⋮

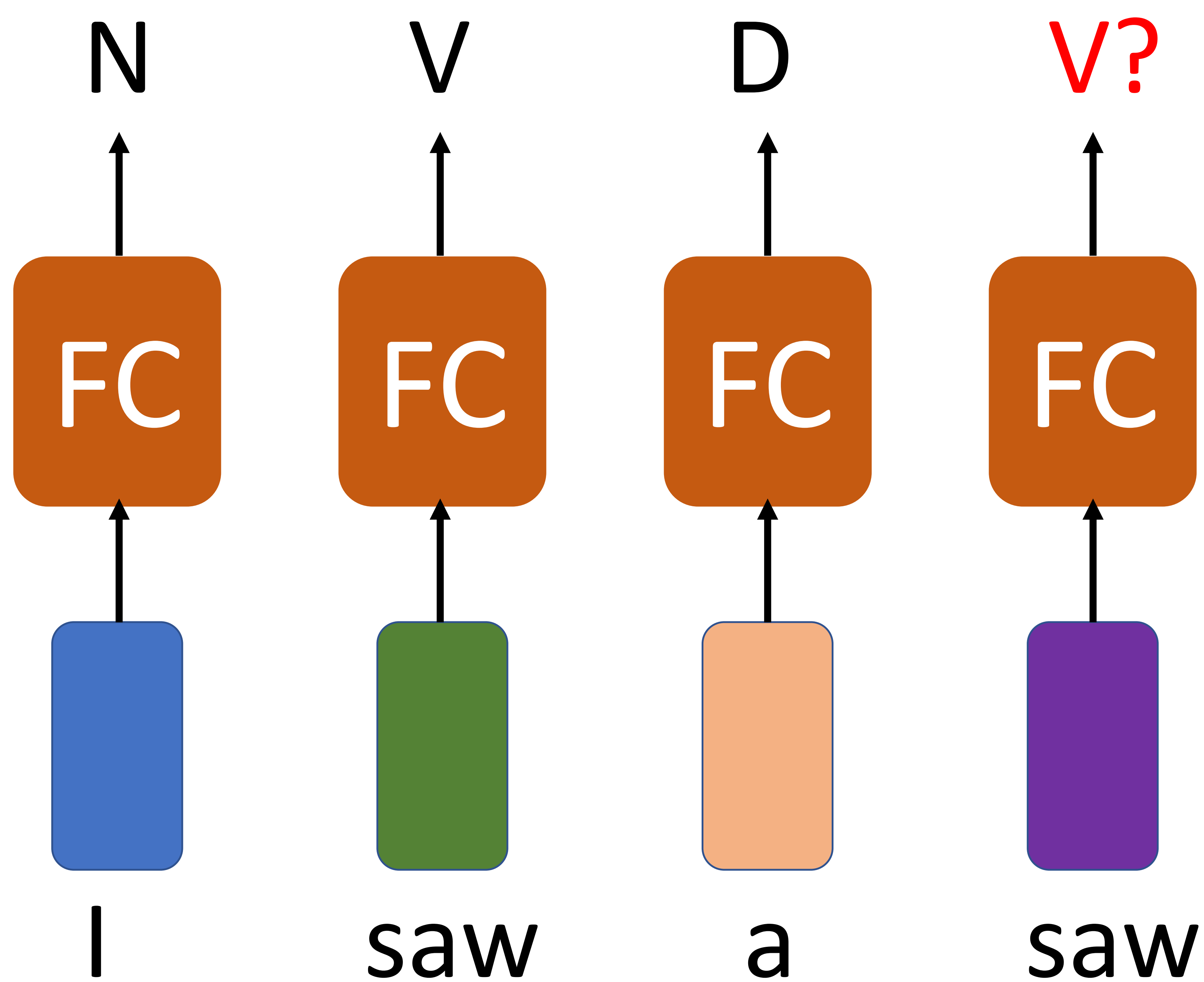
This is a dog



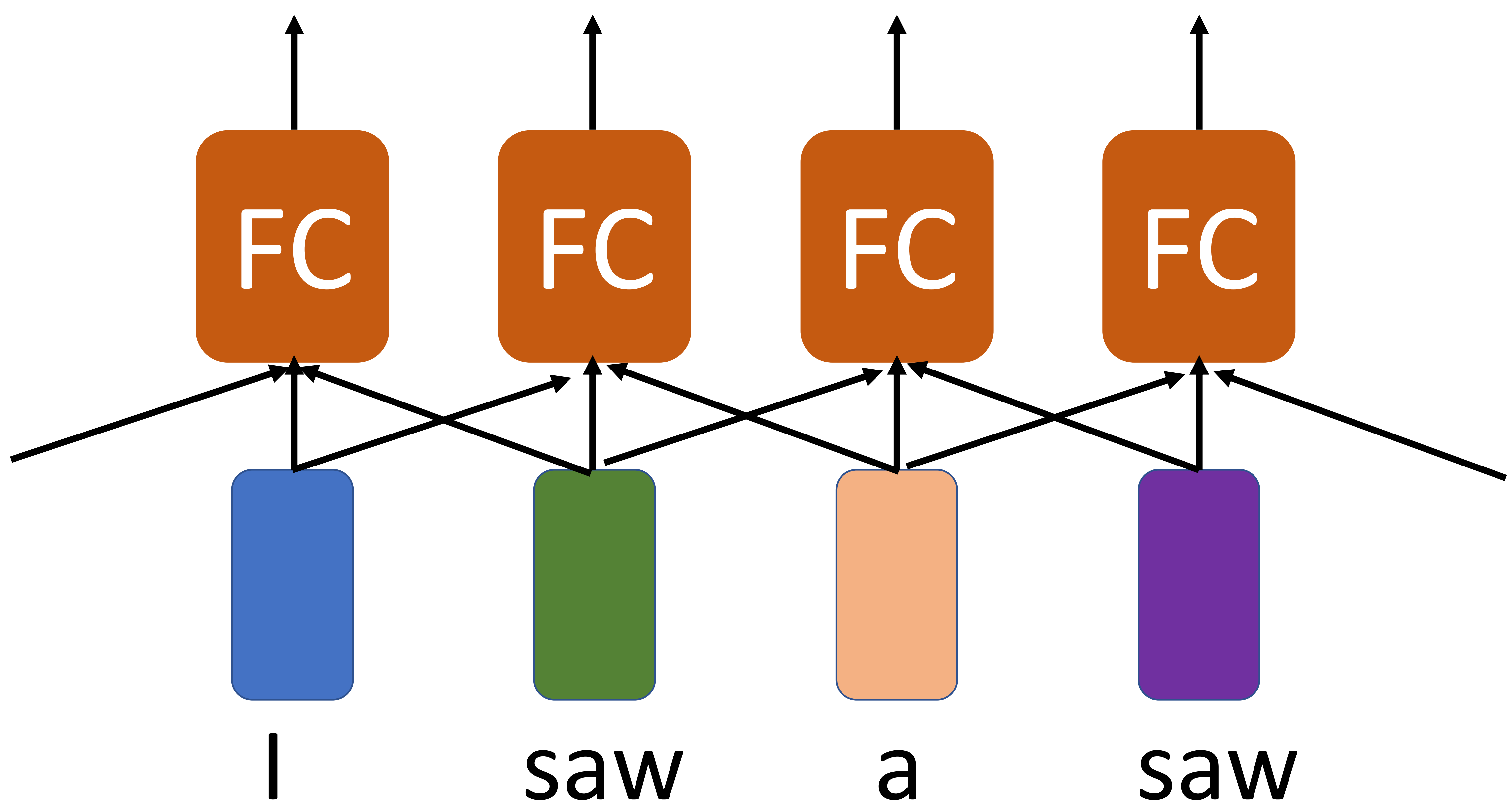
Word Embedding



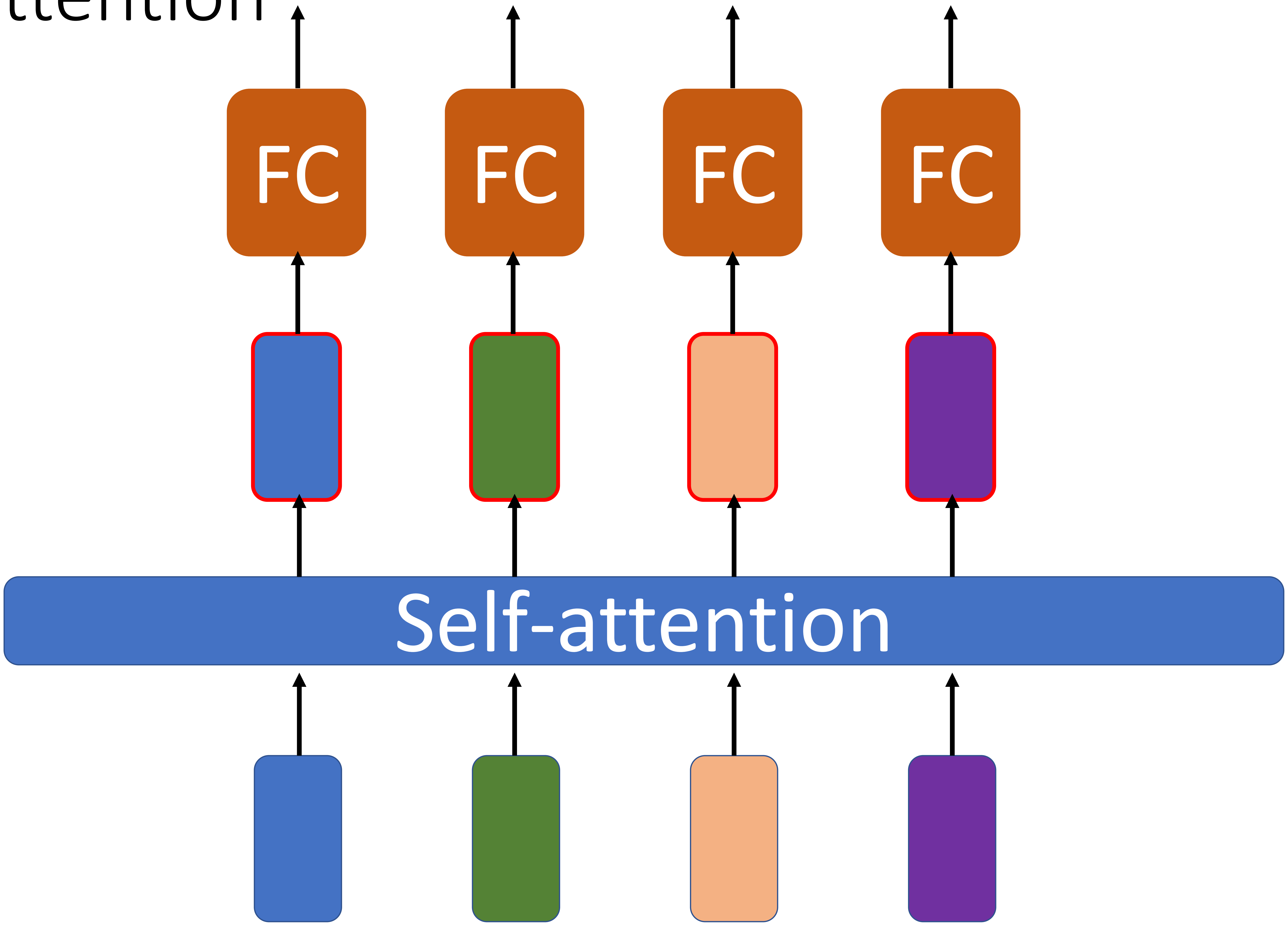
Sequence Labeling

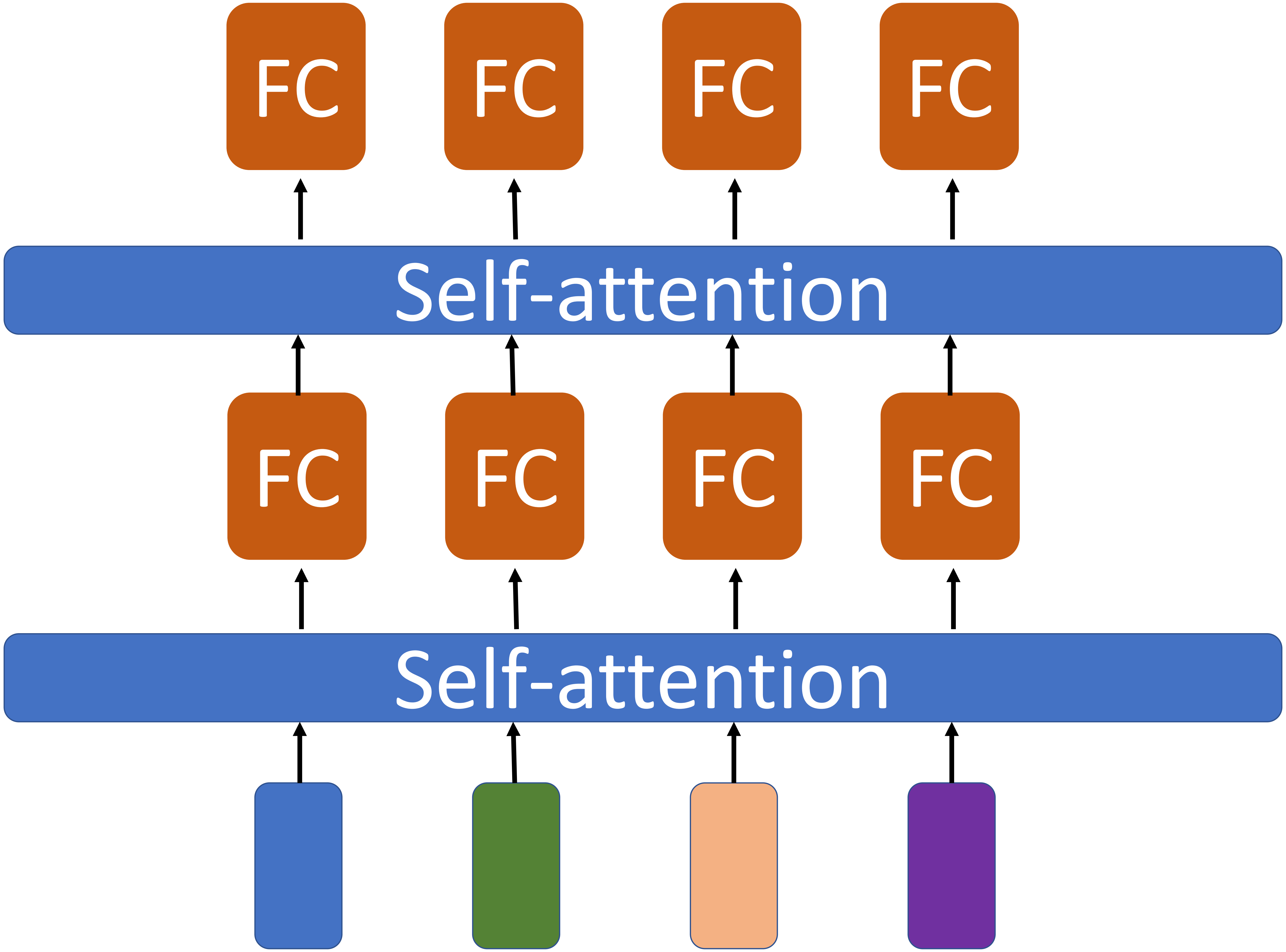


Sequence Labeling



Self-attention




FC
FC
FC
FC
Self-attention
FC
FC
FC
FC
Self-attention

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 Uszkor Google Resear usz@google.c

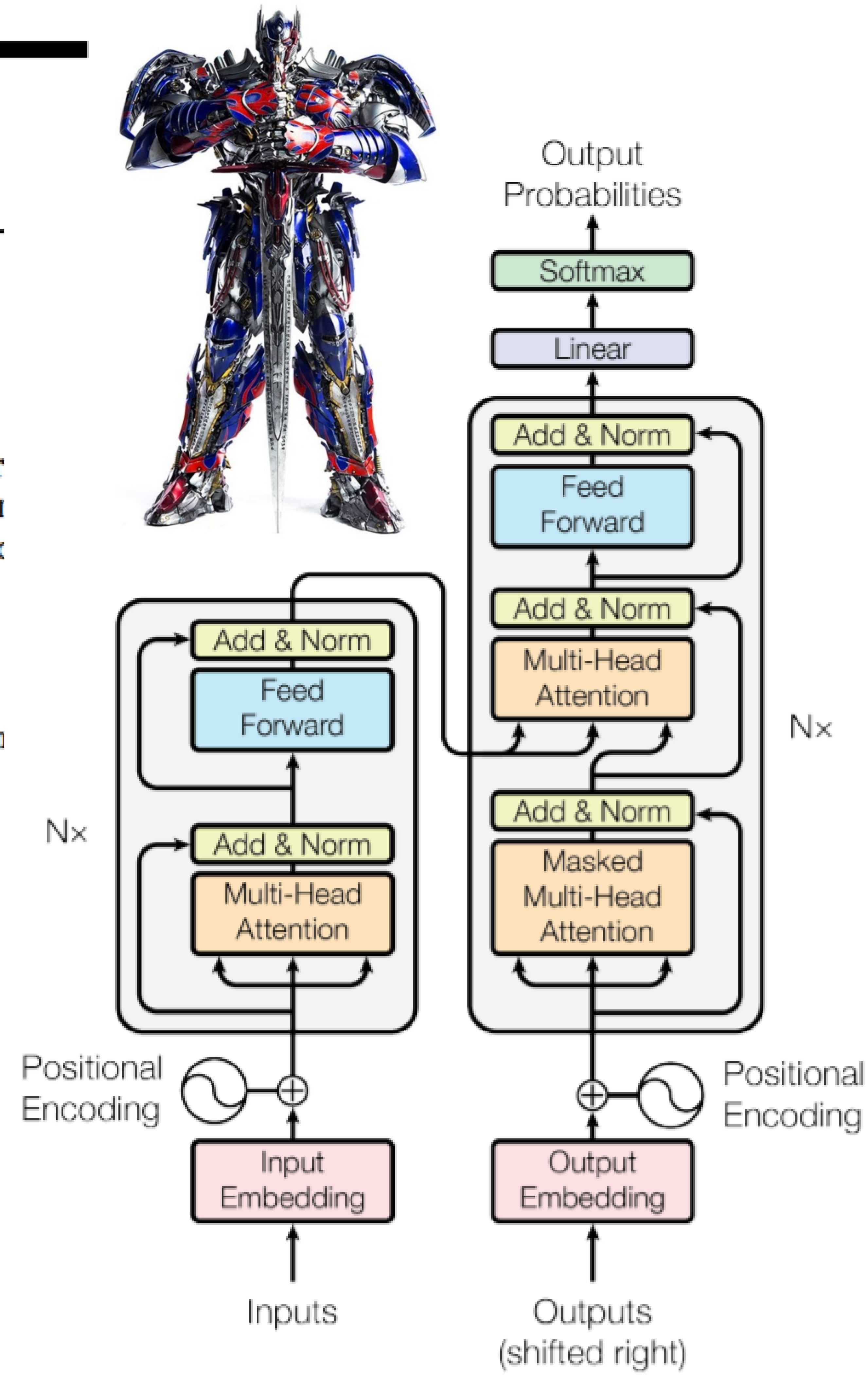
Llion Jones* Google Research llion@google.com

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

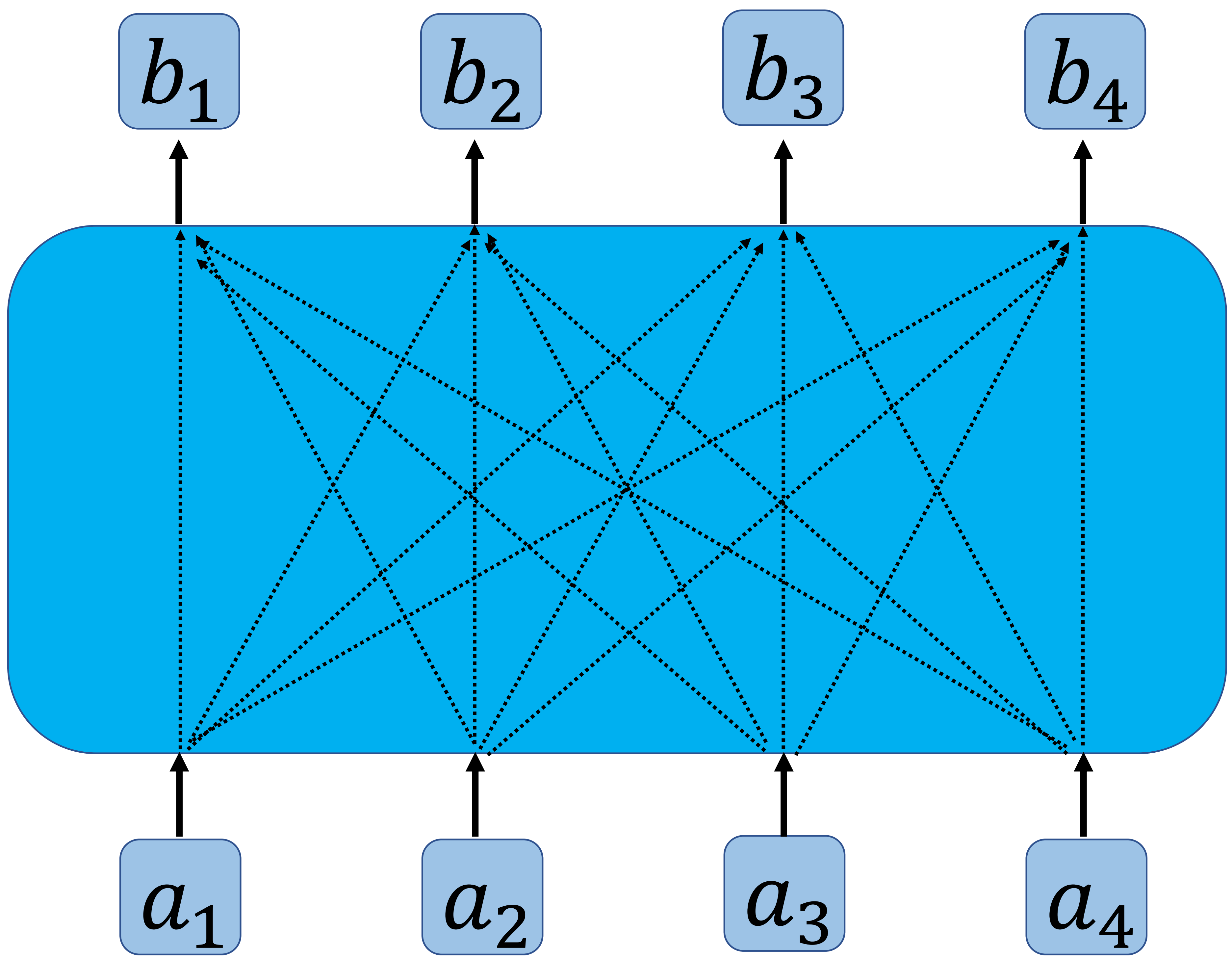
Łukasz Kaiser* Google Brain lukaszkaiser@google.con

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

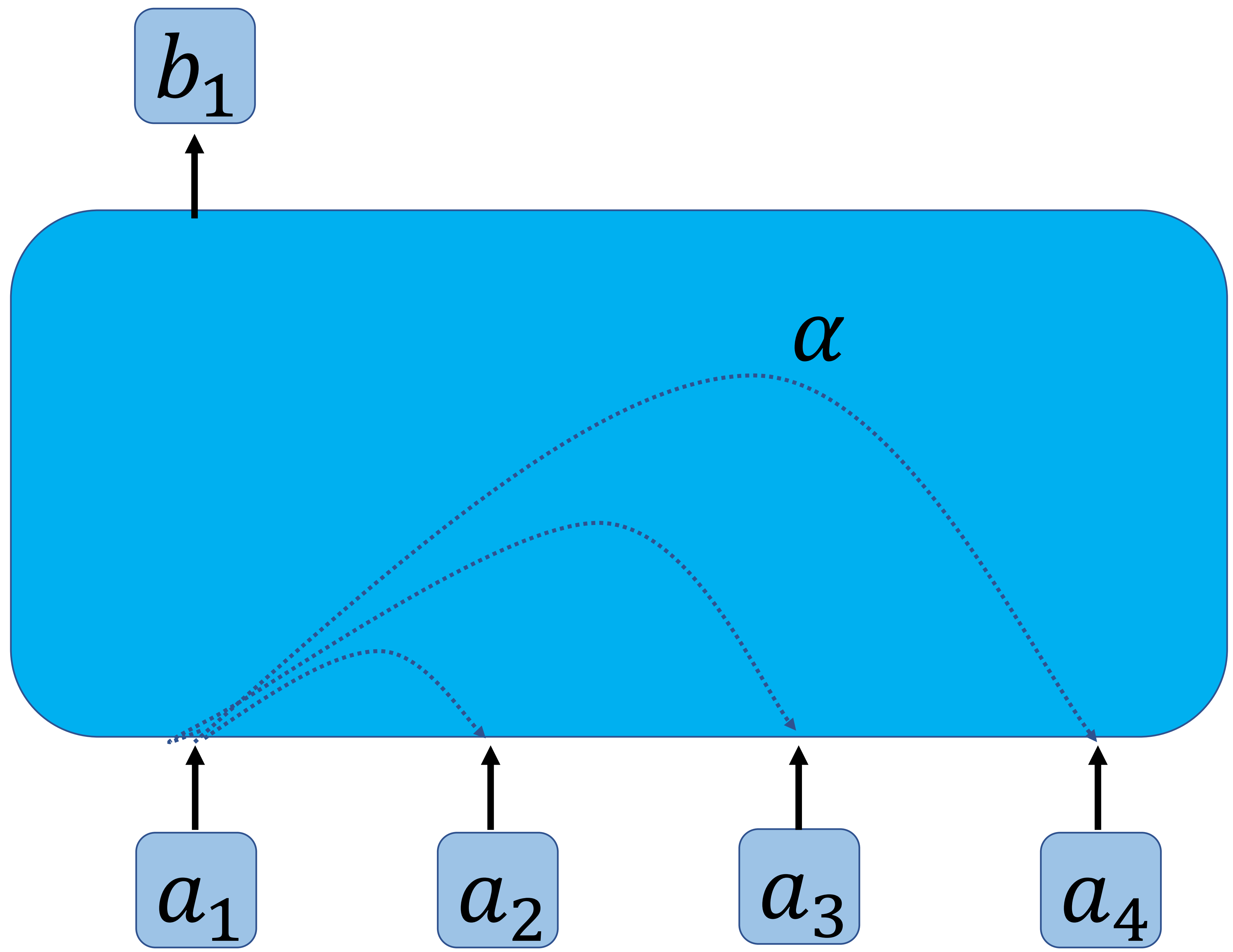
arXiv:1706.03762v5



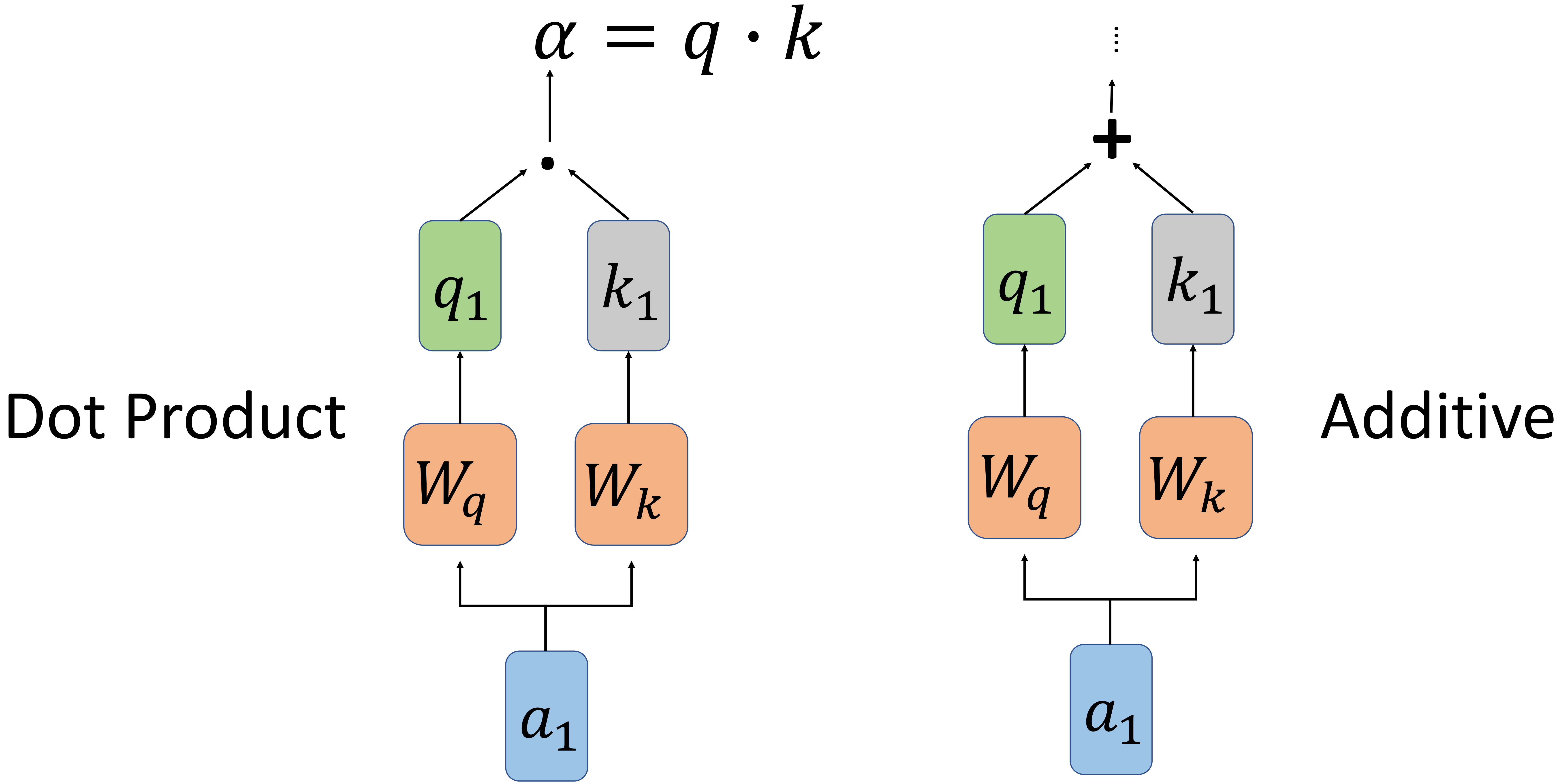
Self-attention



Self-attention

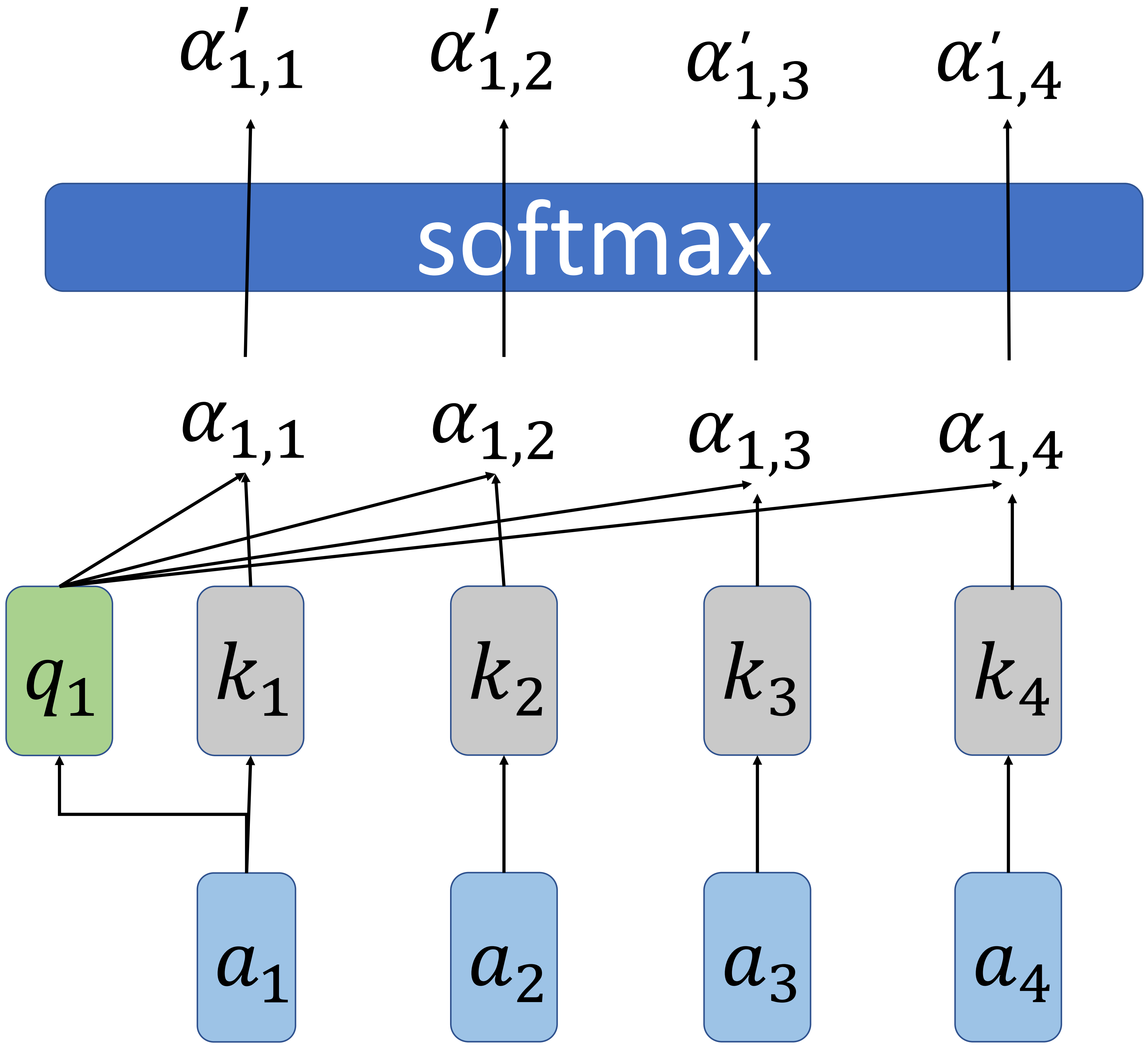


Self-attention



Self-attention

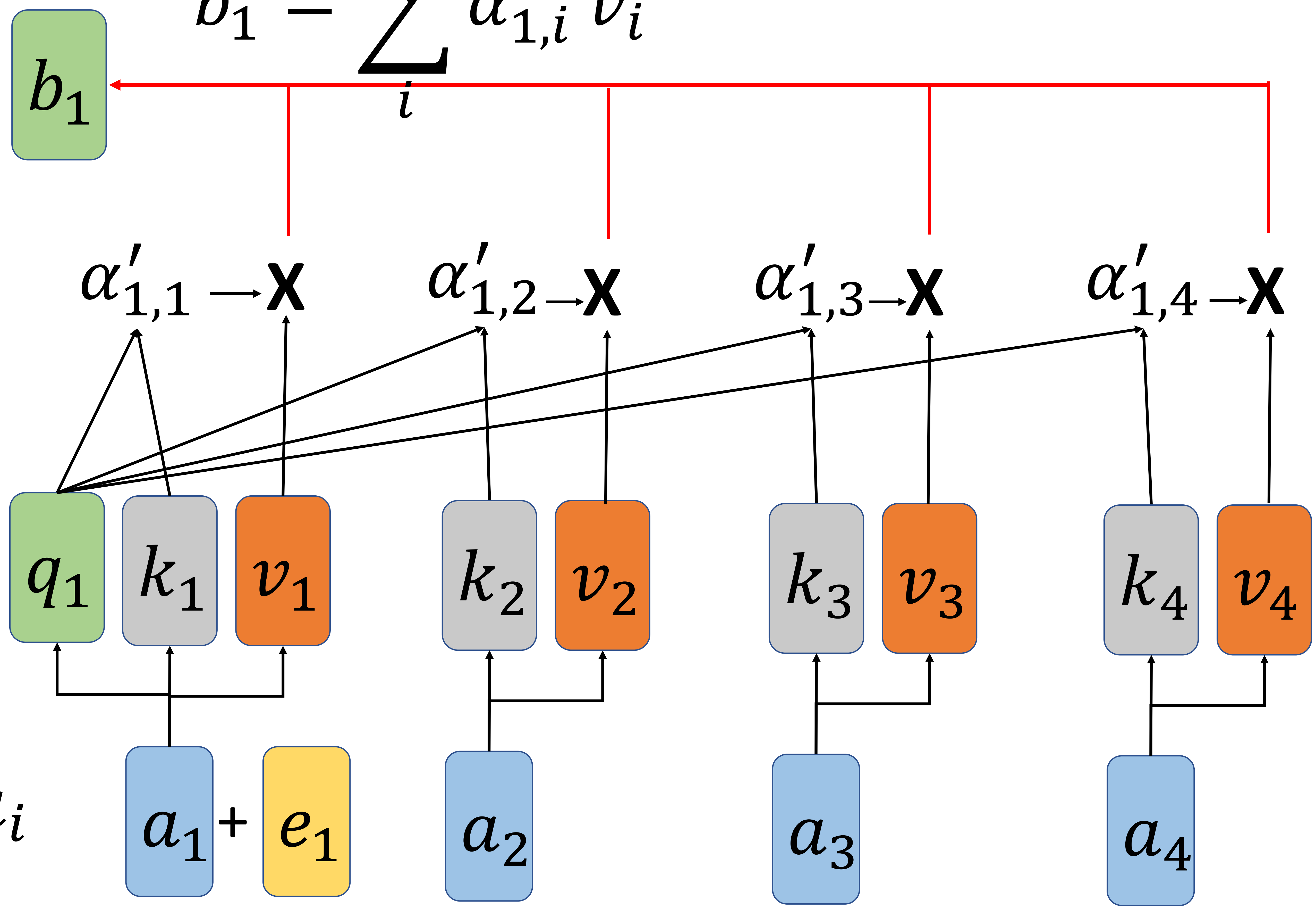
$$\alpha'_{1,i} = exp(\alpha_{1,i})/\sum_j exp(\alpha_{1,j})$$



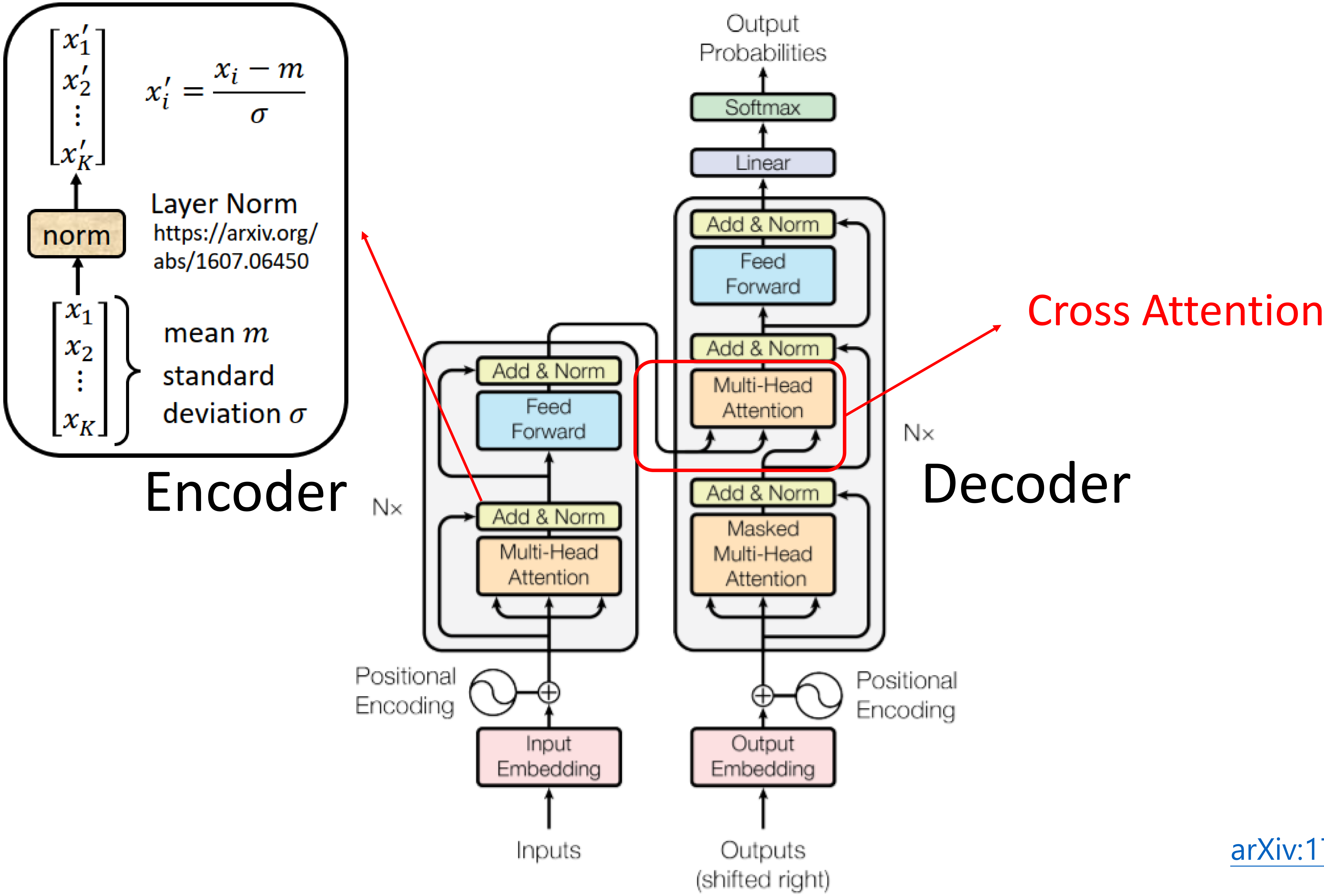
$$q_i = W_q\, a_i$$
$$k_i = W_k\, a_i$$

Self-attention

$$b_1 = \sum_i \alpha'_{1,i} \, v_i$$

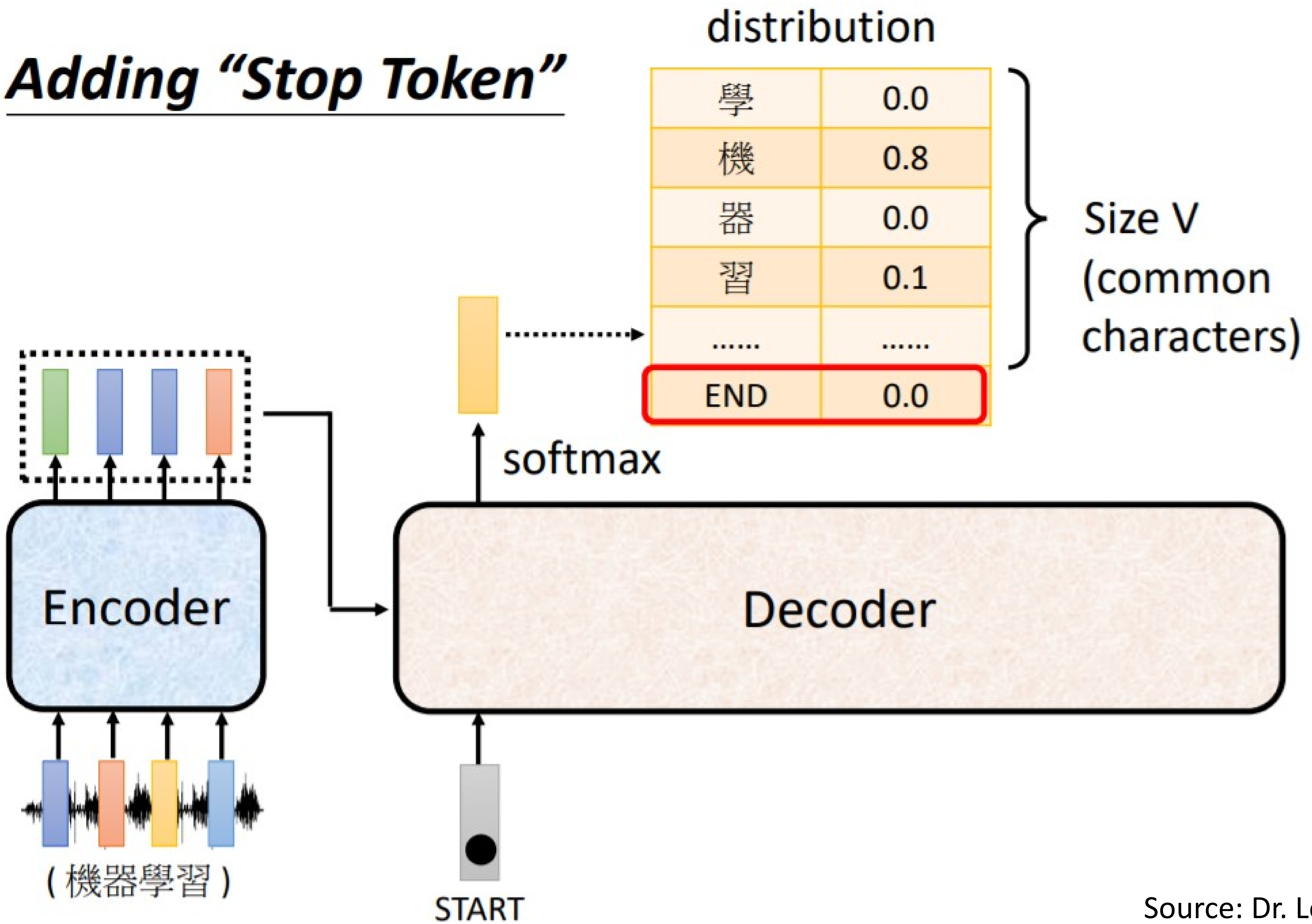


$$v_i = W_v \, a_i$$



arXiv:1706.03762v5

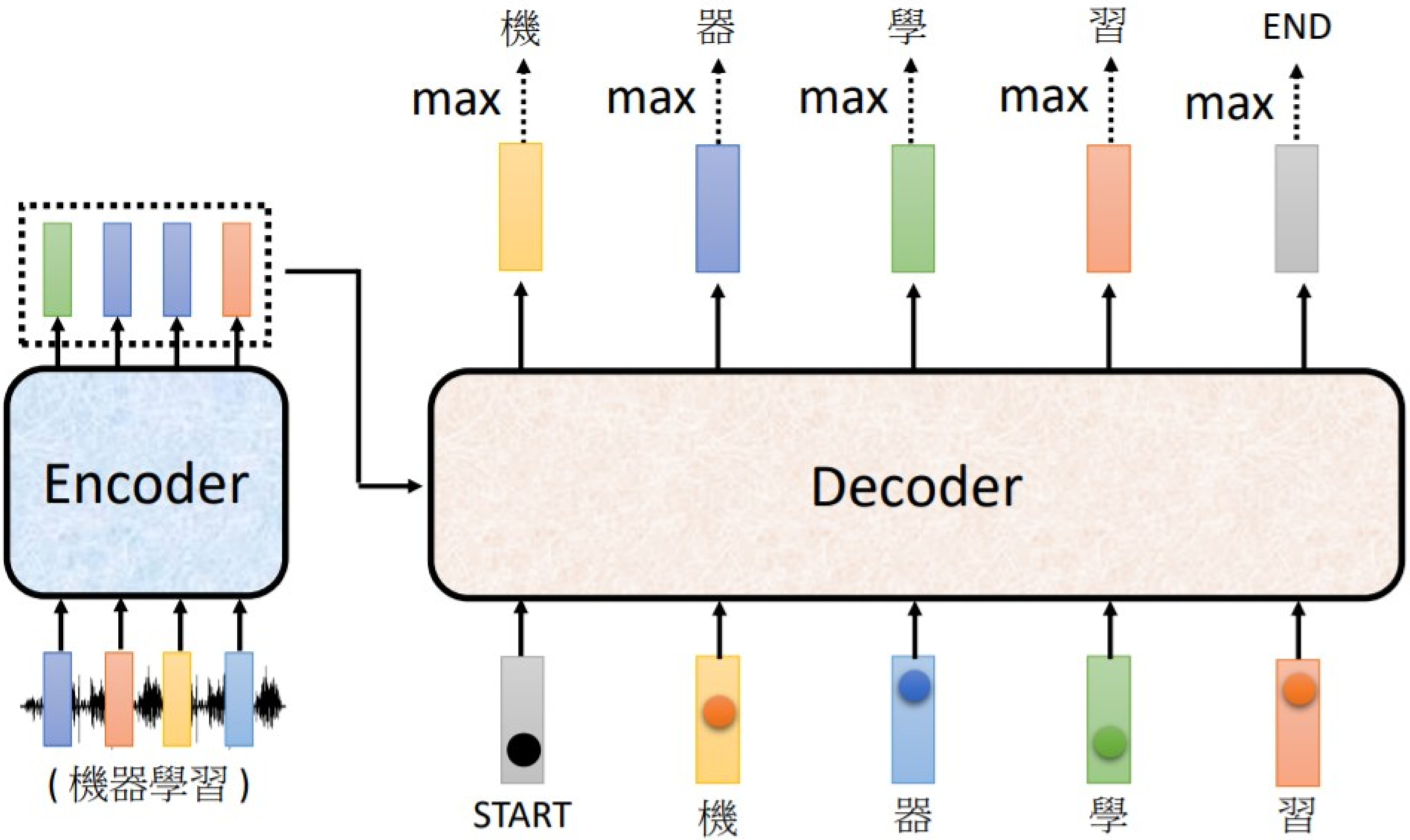
Adding "Stop Token"



Source: Dr. Lee Hung-Yi

Autoregressive

Stop at here!



Source: Dr. Lee Hung-Yi

Natural Language Processing

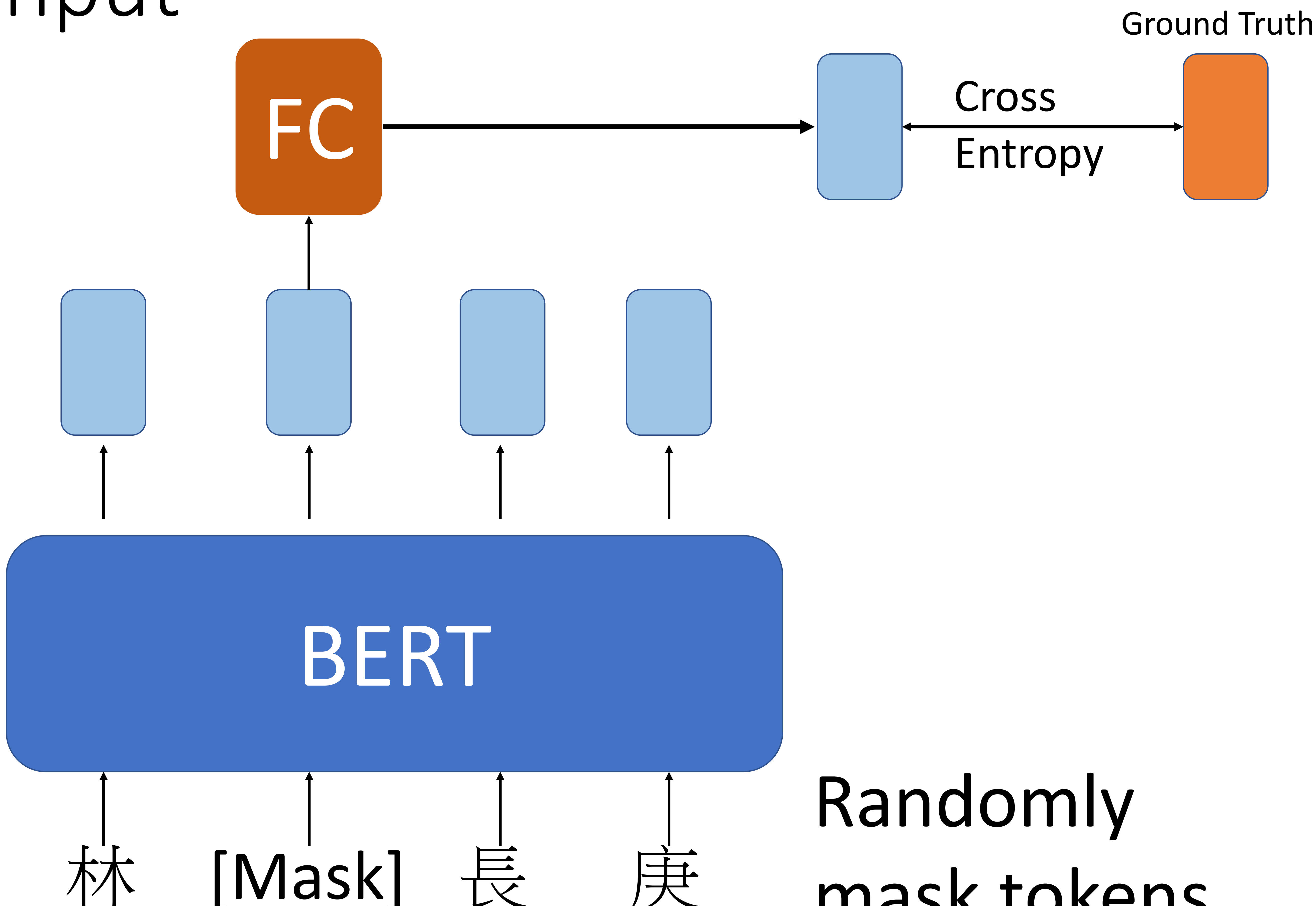
- NLP Tasks
- Transformer
- **BERT**
- GPT

BERT (Bidirectional Encoder Representations from Transformers)

- Self-supervised Learning
  - 3 Billions of words
  - Masking Input
  - Next Sentence Prediction
  - SOP: Sentence Order Prediction (ALBERT)
- Encoder of Transformer

https://arxiv.org/abs/1810.04805

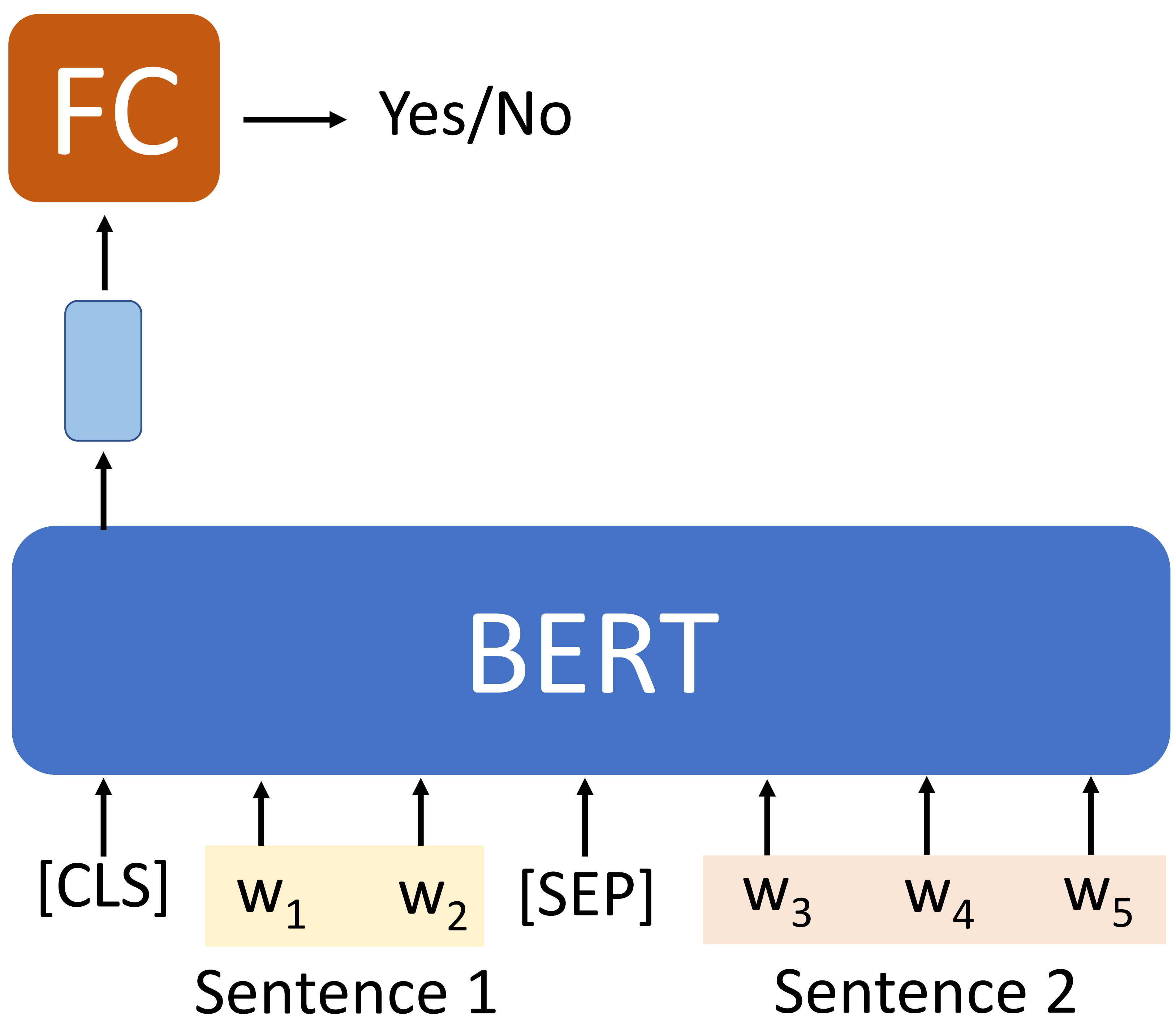
Masking Input



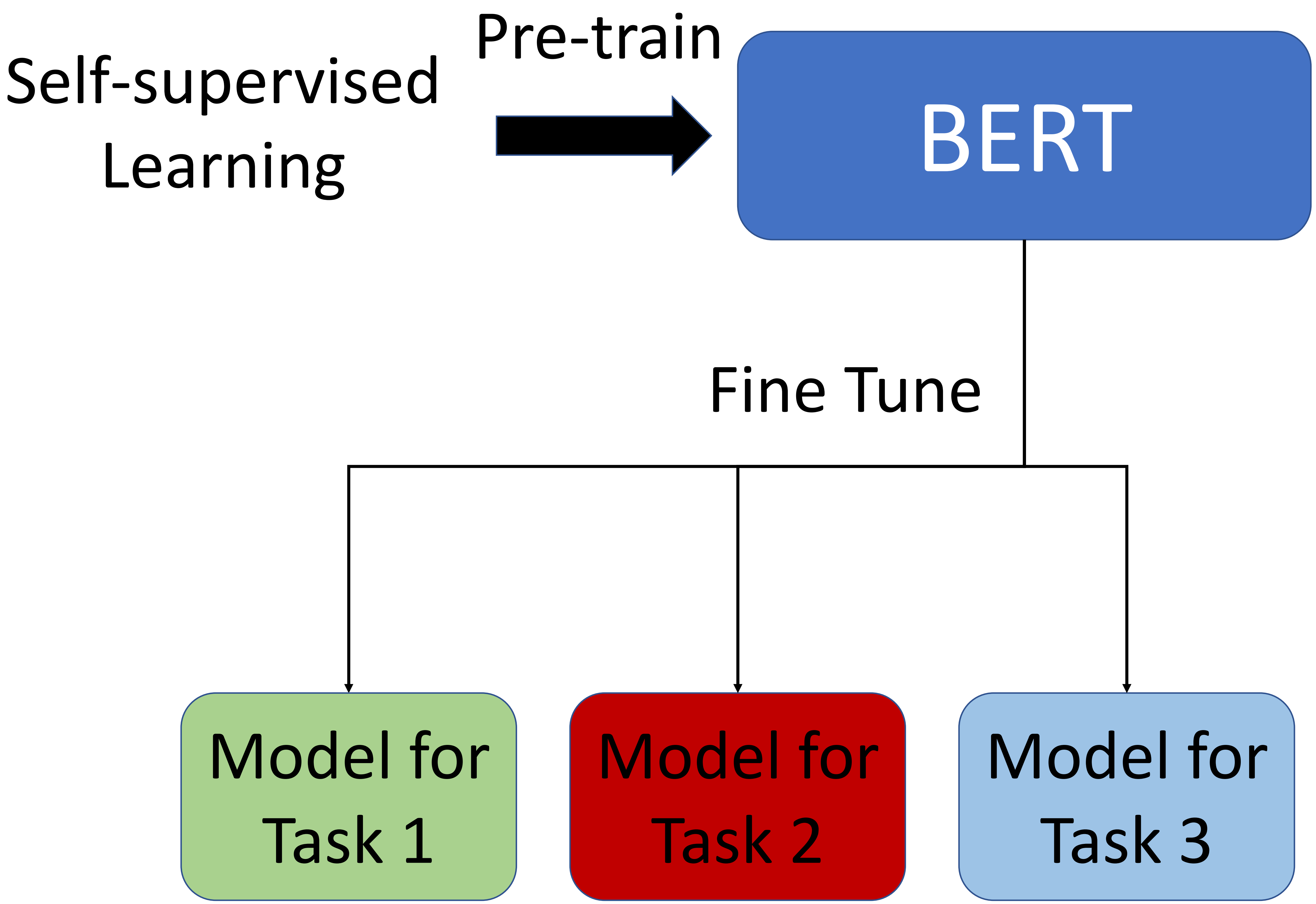
Transformer Encoder

Randomly mask tokens

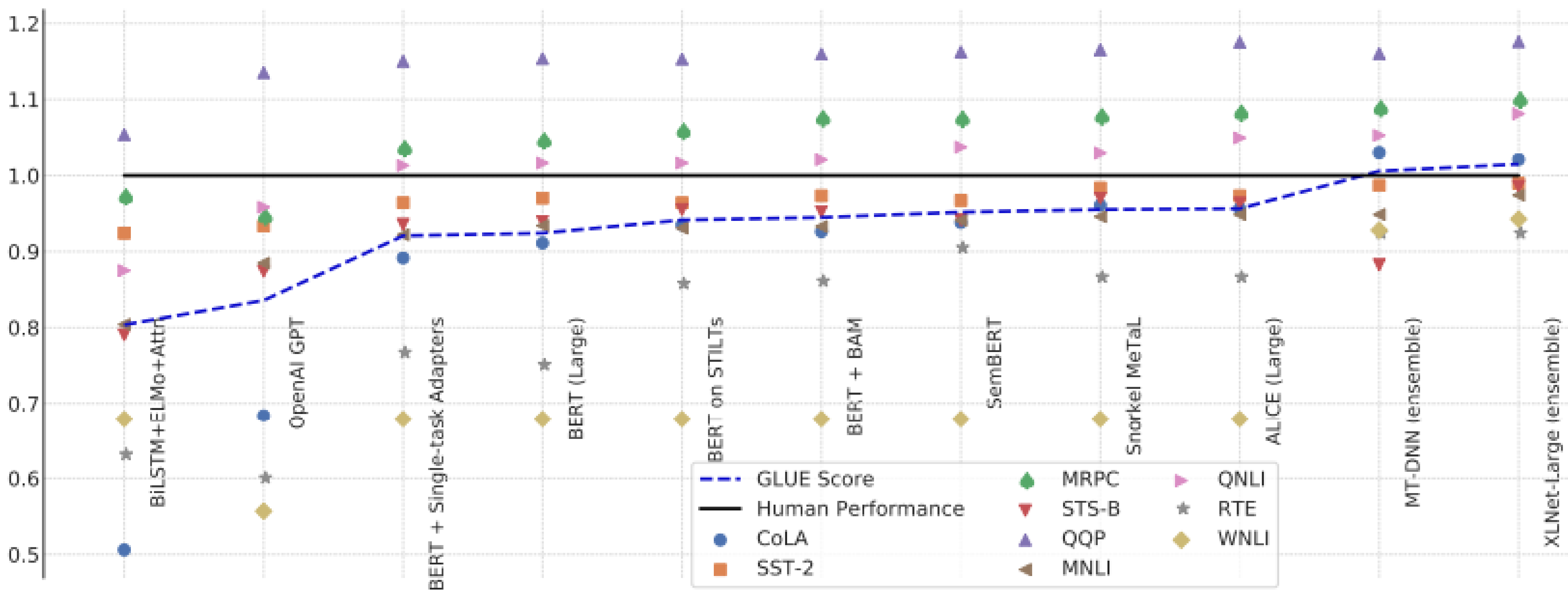
Next Sentence Prediction



- RoBERTa (Robustly optimized BERT approach) https://arxiv.org/abs/1907.11692
- ALBERT (use SOP) https://arxiv.org/abs/1909.11942

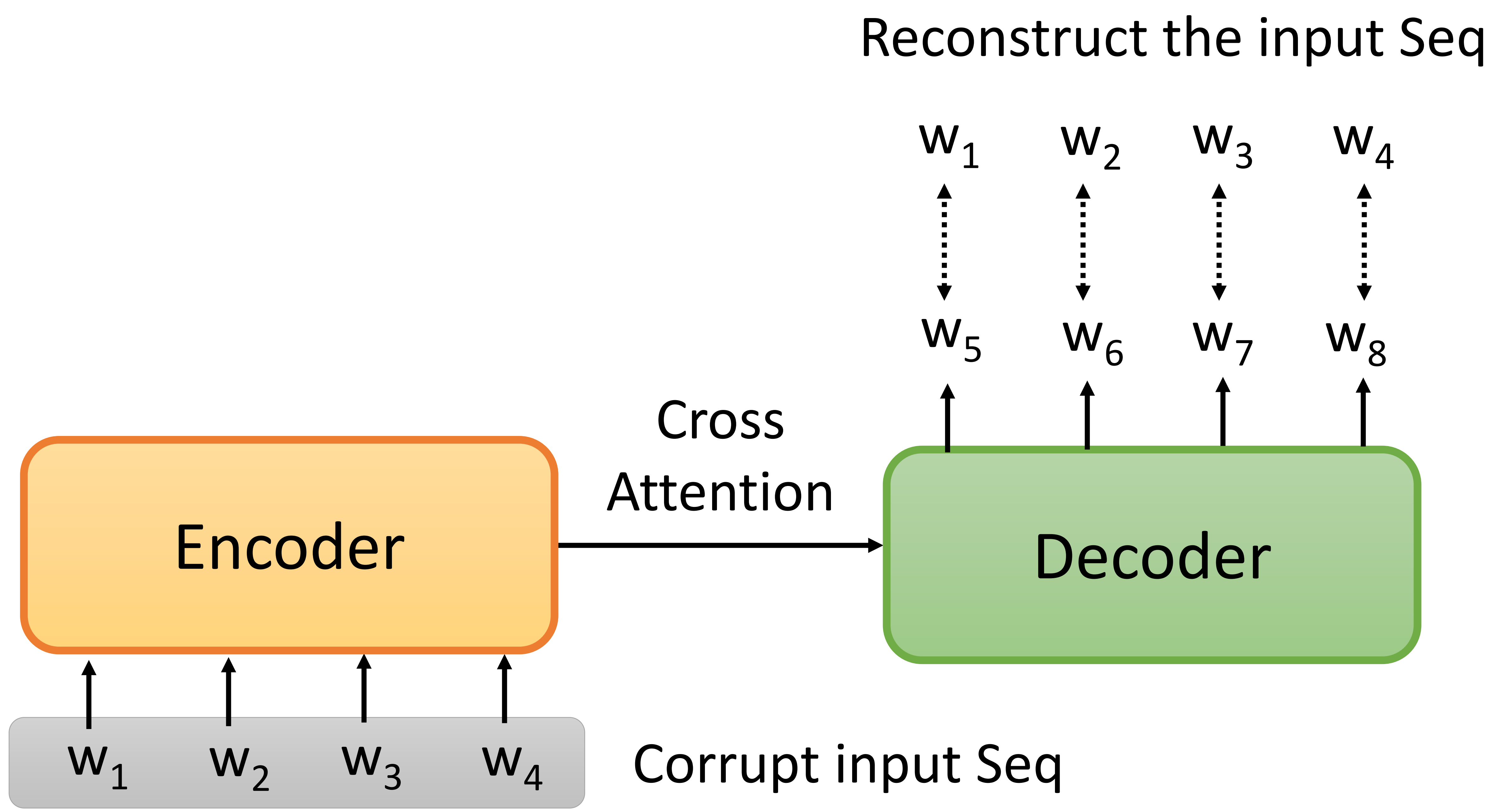

Self-supervised Learning
Pre-train
BERT
Fine Tune
Model for Task 1
Model for Task 2
Model for Task 3

Performance of BERT Family



https://arxiv.org/abs/1905.00537

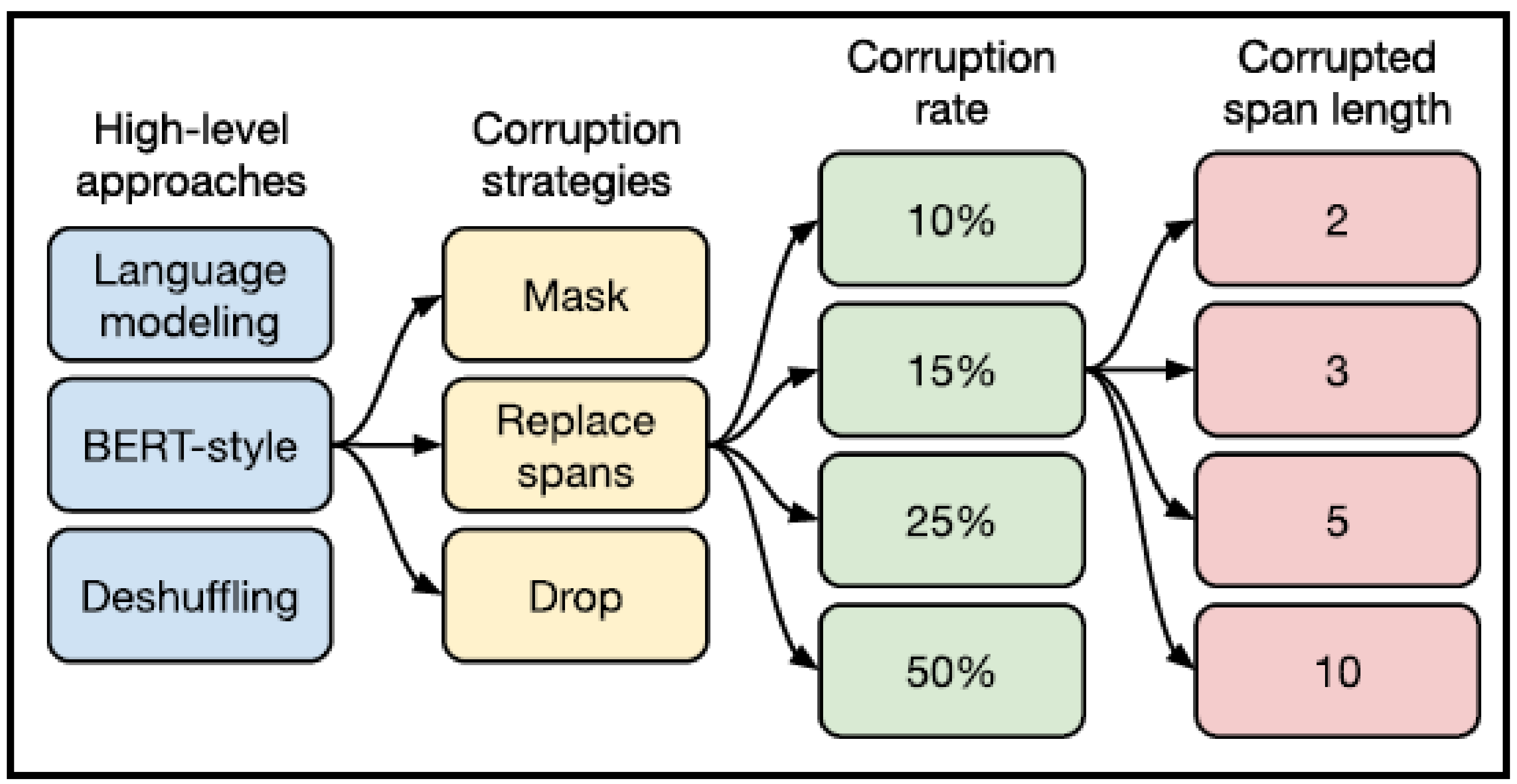
Pre-training a Seq2seq Model



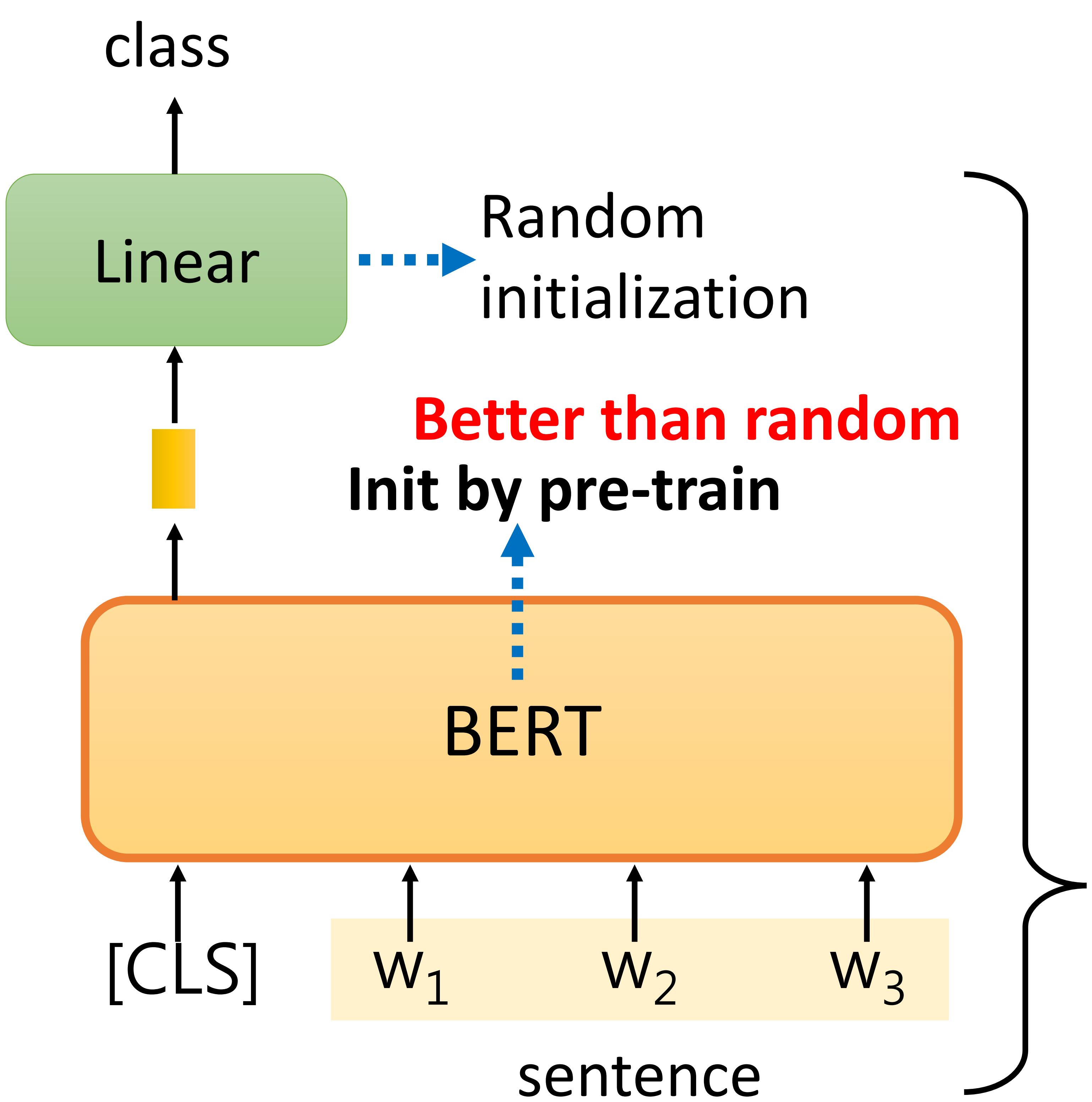
Source: Dr. Lee Hung-Yi

T5 (Text-To-Text Transfer Transformer) - Comparison

Objective	Inputs	Targets
Prefix language modeling	Thank you for inviting	me to your party last week .
BERT-style	Thank you <M> <M> me to your party apple week .	*(original text)*
Deshuffling	party me for your to . last fun you inviting week Thank	*(original text)*
I.i.d. noise, mask tokens	Thank you <M> <M> me to your party <M> week .	*(original text)*
I.i.d. noise, replace spans	Thank you <X> me to your party <Y> week .	<X> for inviting <Y> last <Z>
I.i.d. noise, drop tokens	Thank you me to your party week .	for inviting last
Random spans	Thank you <X> to <Y> week .	<X> for inviting me <Y> your party last <Z>



https://arxiv.org/abs/1910.10683v3



Input: sequence

output: class

Example:

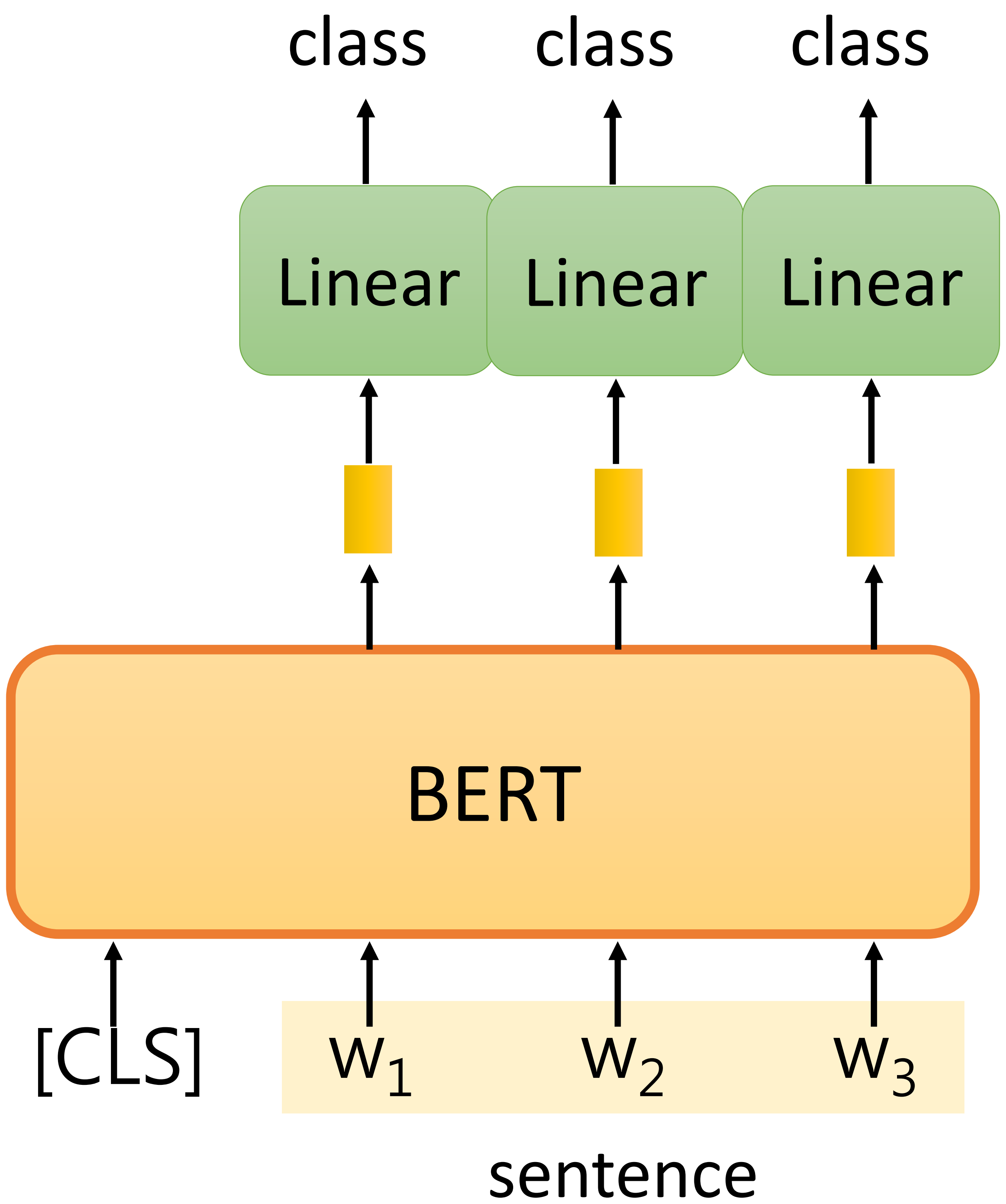
Sentiment analysis

this is good

positive

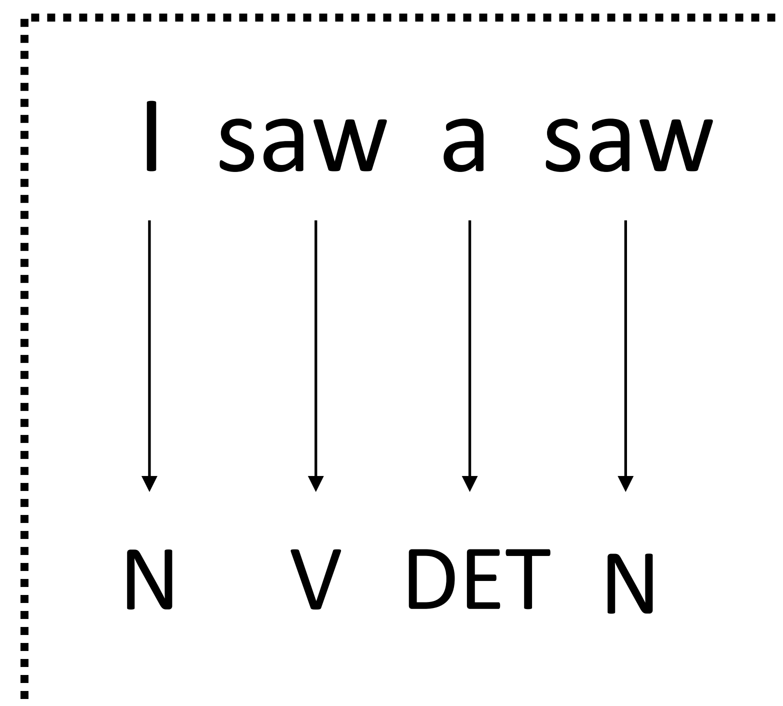
This is the model to be learned.

Source: Dr. Lee Hung-Yi



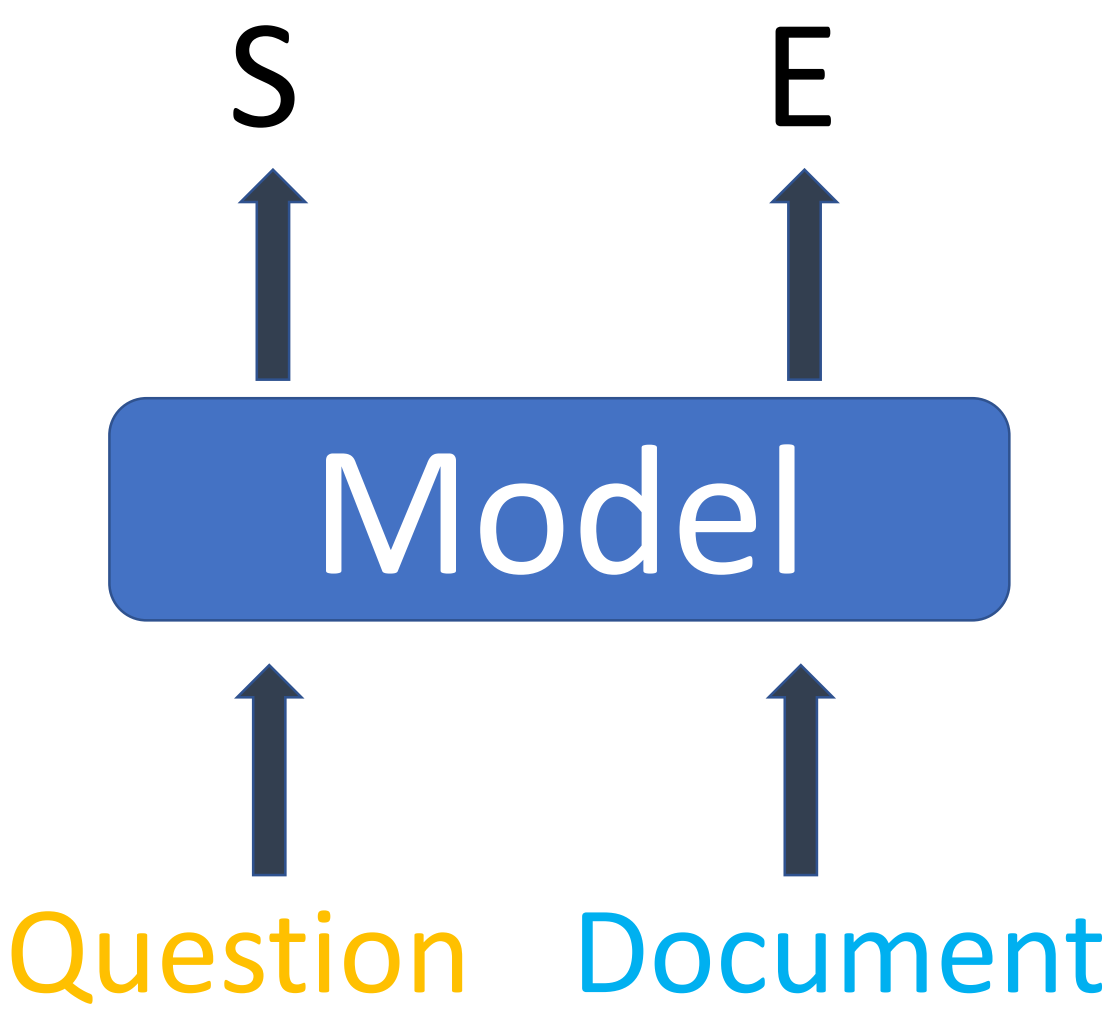
Input: sequence
output: same as input

Example:
POS tagging



Source: Dr. Lee Hung-Yi

Extractive QA



In meteorology, precipitation is any product of the condensati 17 f atmospheric water vapor that falls under gravity. The main forms of precipitation include drizzle, rain, sleet, snow, graupel and hail... Precipitation forms as smaller droplets coalesce vi 77 llision 79 other rain drops or ice crystals within a cloud. Short, intense periods of rain in scattered locations are called "showers".

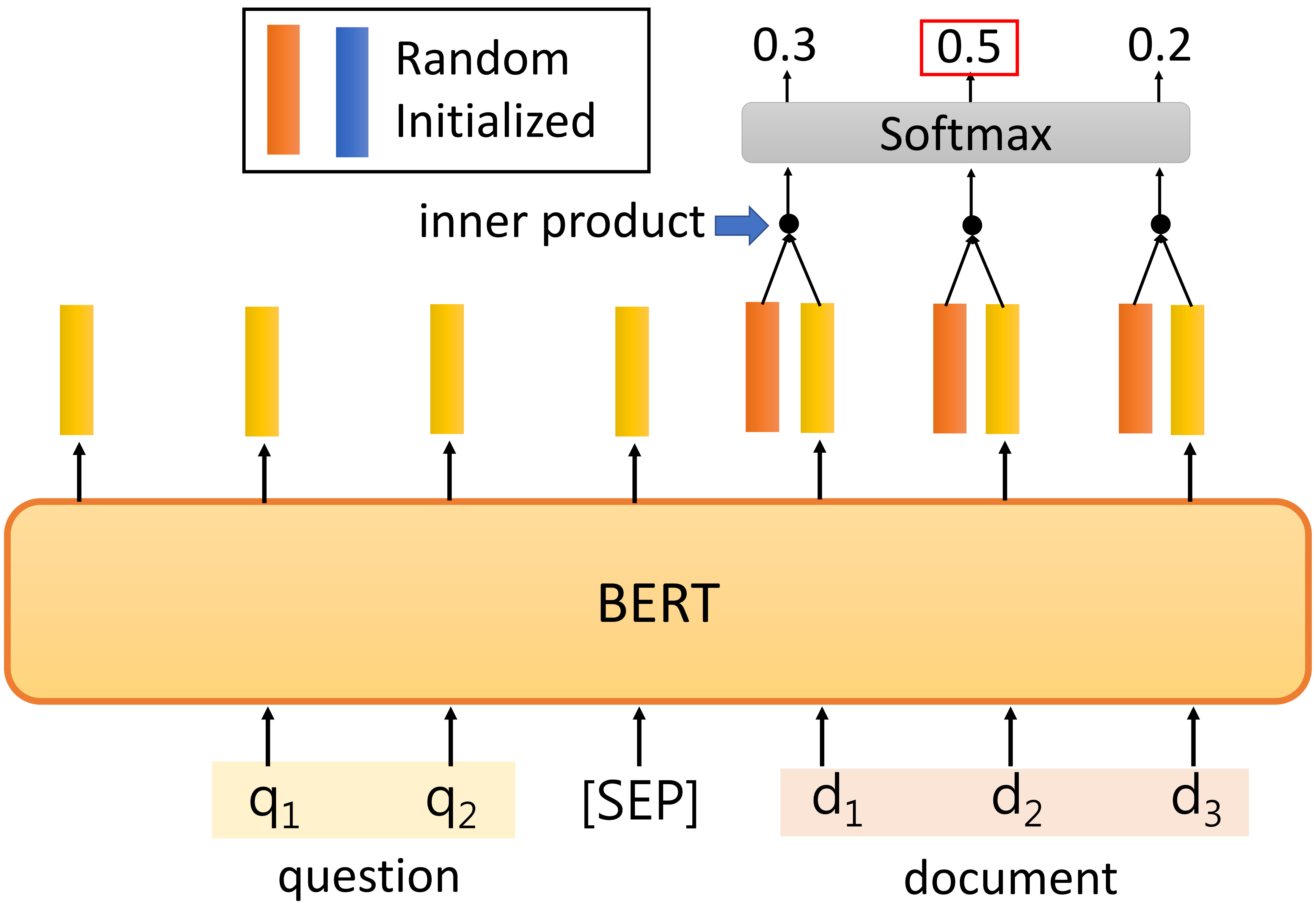
What causes precipitation to fall?
gravity S= 17, E=17

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupel

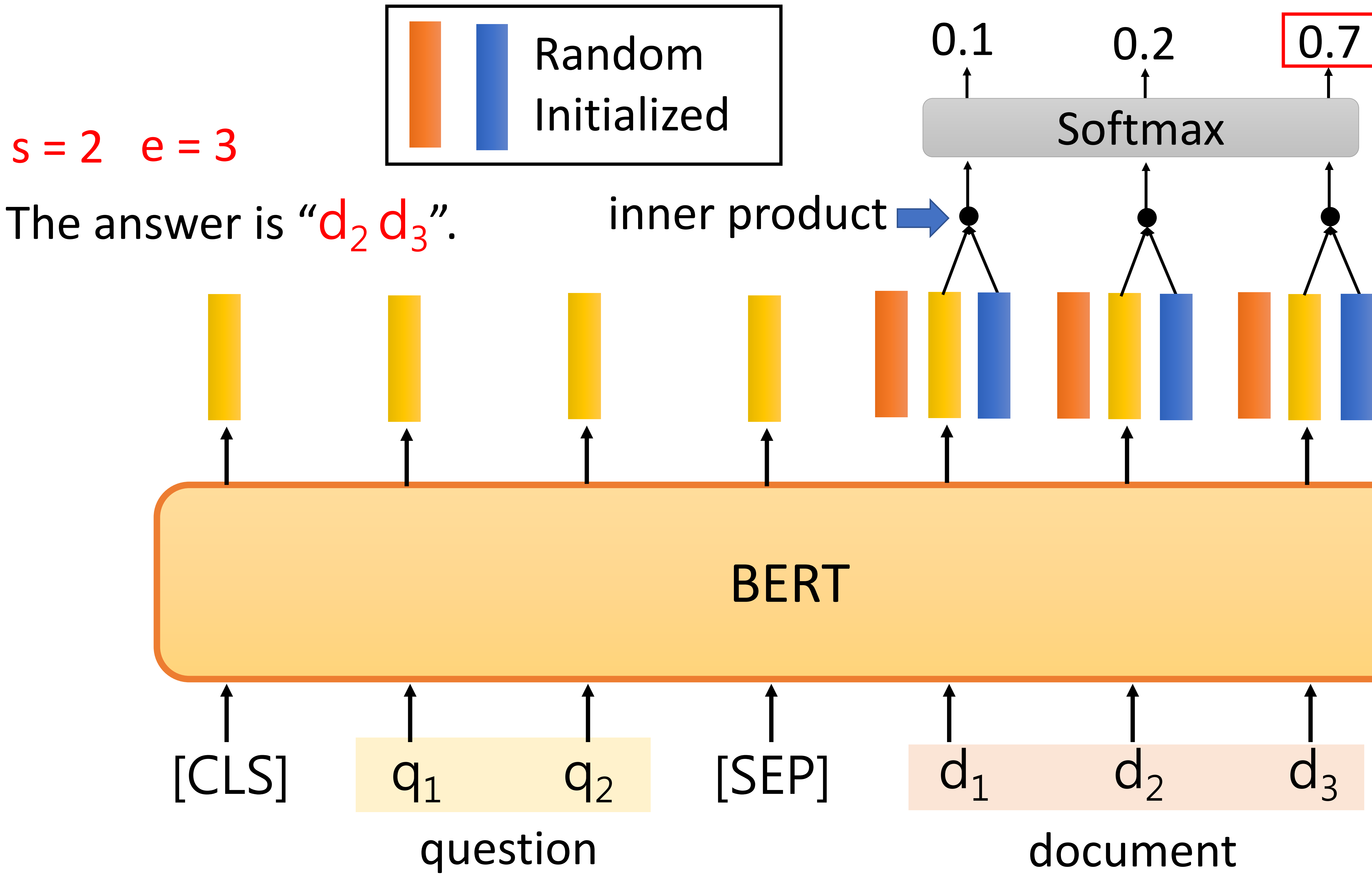
Where do water droplets collide with ice crystals to form precipitation?
within a cloud S= 77, E=79

Source: Dr. Lee Hung-Yi

s = 2

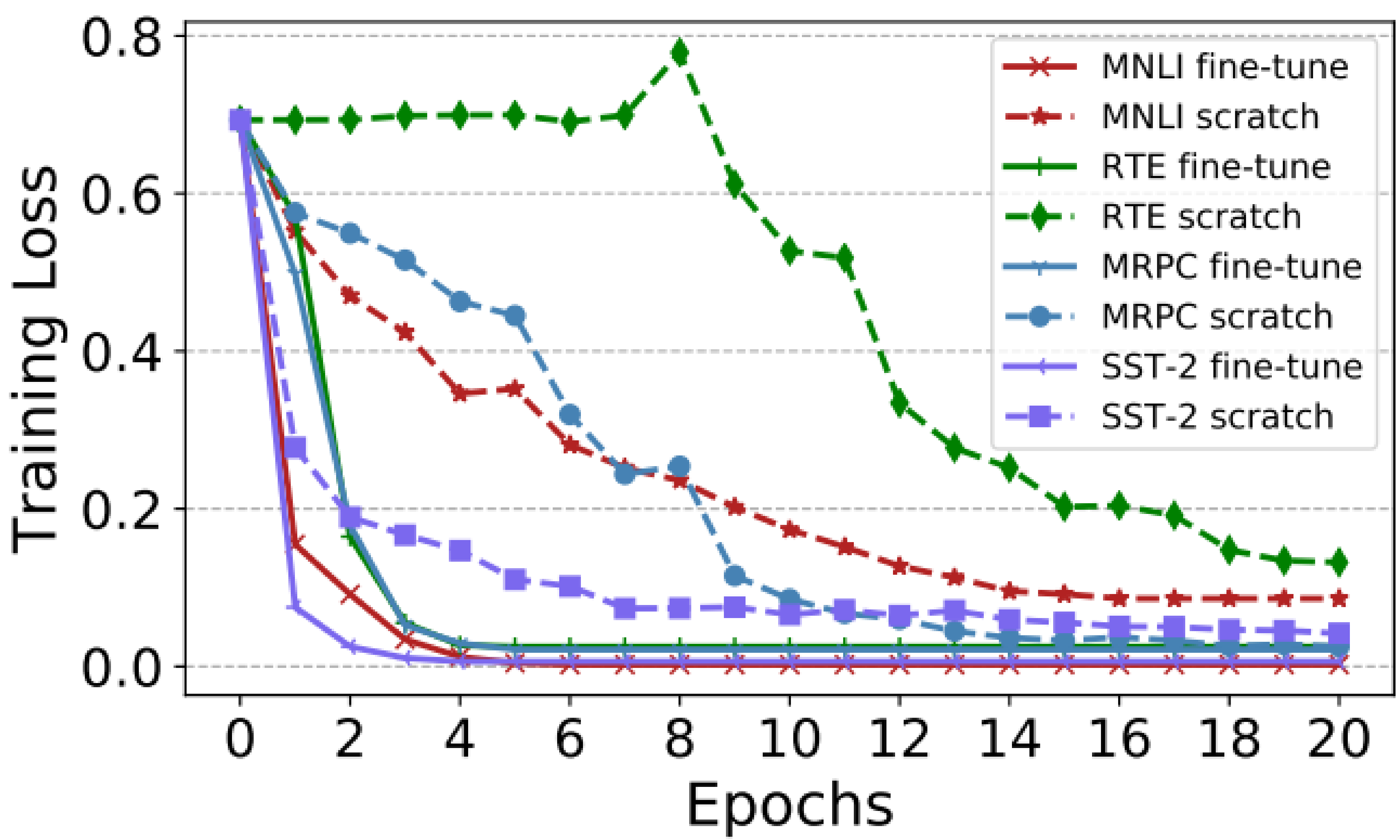


Source: Dr. Lee Hung-Yi



Source: Dr. Lee Hung-Yi

Random Initialization vs. Pre-train



https://arxiv.org/abs/1908.05620

- Applying BERT to **protein**, **DNA**, **music classification**

EI	CCAGCTGCATCACAGGAGGCCAGCGAGCAGGTCTGTTCCAAGG
EI	AGACCCGCCGGGAGGCGGAGGACCTGCAGGGTGAGCCCCACC
IE	AACGTGGCCTCCTTGTGCCCTTCCCCACAGTGCCCTCTTCCAGGA
IE	CCACTCAGCCAGGCCCTTCTTCTCCTCCAGGTCCCCCACGGCCCT
IE	CCTGATCTGGGTCTCCCCTCCCACCCTCAGGGAGCCAGGCTCGG
IE	AGCCCTCAACCCTTCTGTCTCACCCTCCAGCCTAAAGCTCCTTGA
IE	CCACTCAGCCAGGCCCTTCTTCTCCTCCAGGTCCCCCACGGCCCT
N	CTGTGTTCACCACATCAAGCGCCGGGACATCGTGCTCAAGTGGG
N	GTGTTACCGAGGGCATTTCTAACAGTCTTCTTACTACGGCCTCCG
N	TCTGAGCTCTGCATTTGTCTATTCTCCAGCTGACCCTGGTTCTCTC

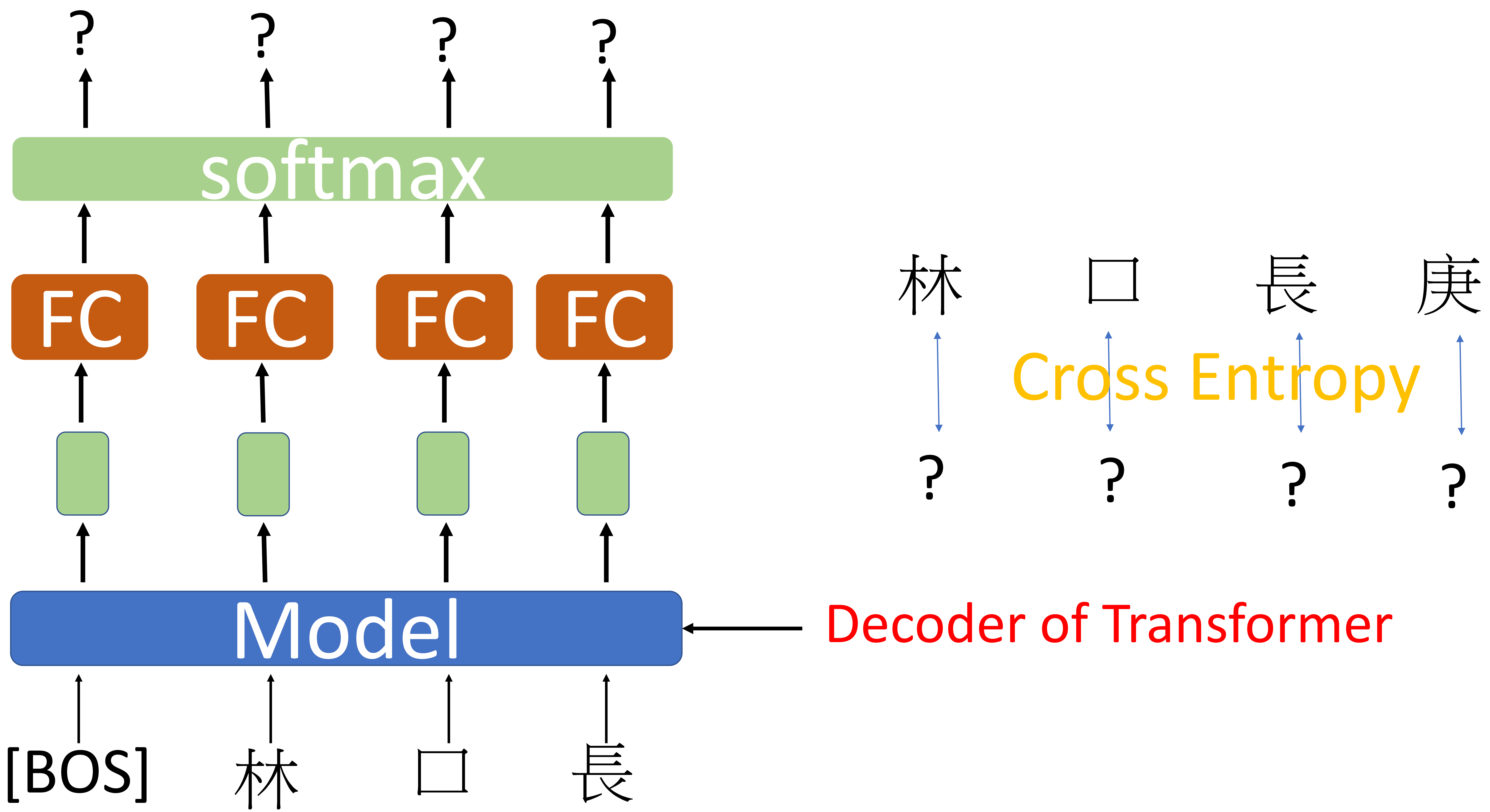
	Protein			DNA				Music
	localization	stability	fluorescence	H3	H4	H3K9ac	Splice	composer
specific	69.0	76.0	63.0	87.3	87.3	79.1	94.1	-
BERT	64.8	74.5	63.7	83.0	86.2	78.3	97.5	55.2
re-emb	63.3	75.4	37.3	78.5	83.7	76.3	95.6	55.2
rand	58.6	65.8	27.5	75.6	66.5	72.8	95	36

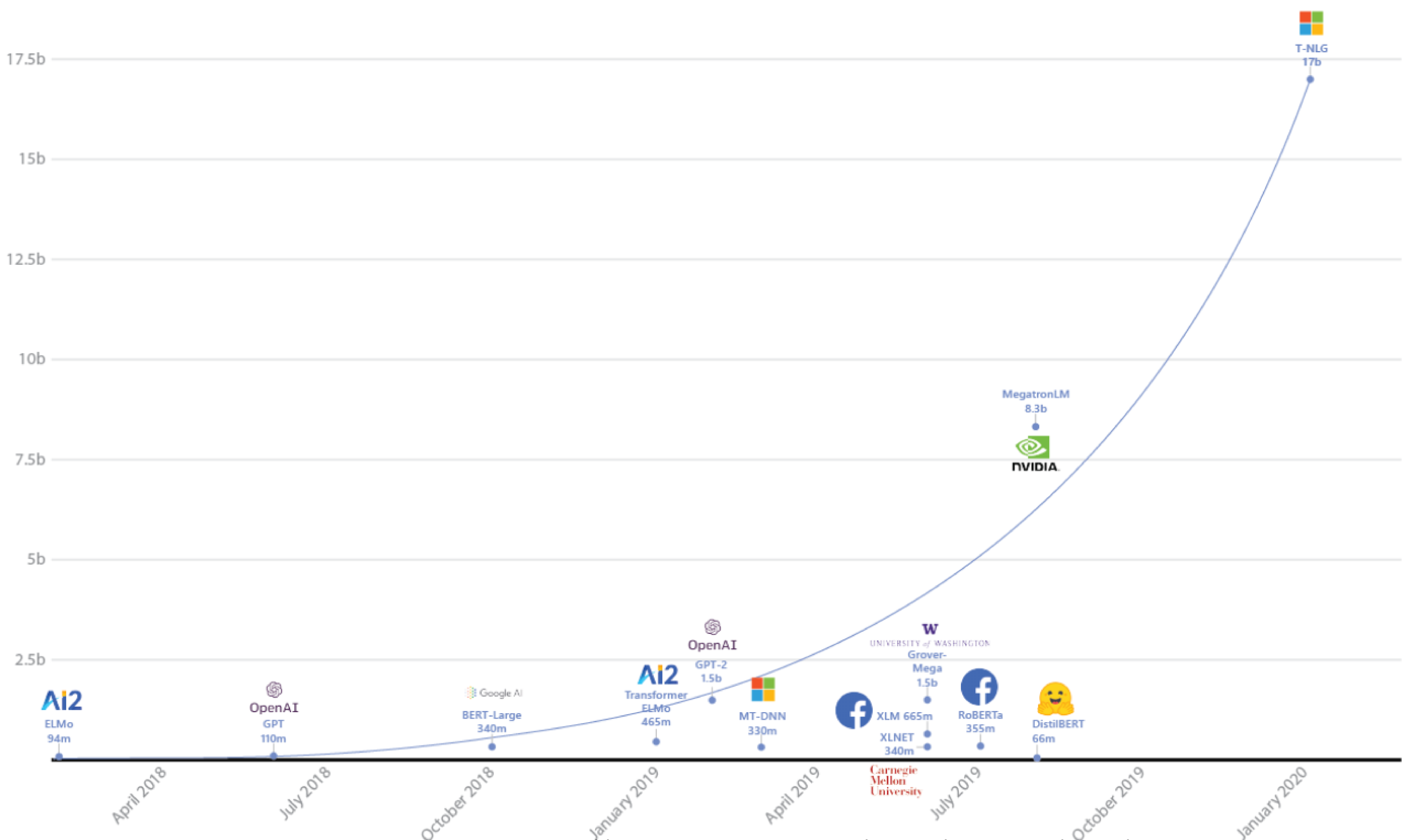
https://arxiv.org/abs/2103.07162

Natural Language Processing

- NLP Tasks
- Transformer
- BERT
- GPT

GPT – Predict Next Token





https://www.microsoft.com/en-us/research/blog/turing-nlg-a-17-billion-parameter-language-model-by-microsoft/

In-Context Learning

Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

```
Translate English to French:
cheese =>
```

task description
prompt

One-shot

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.

```
Translate English to French:
sea otter => loutre de mer
cheese =>
```

task description
example
prompt

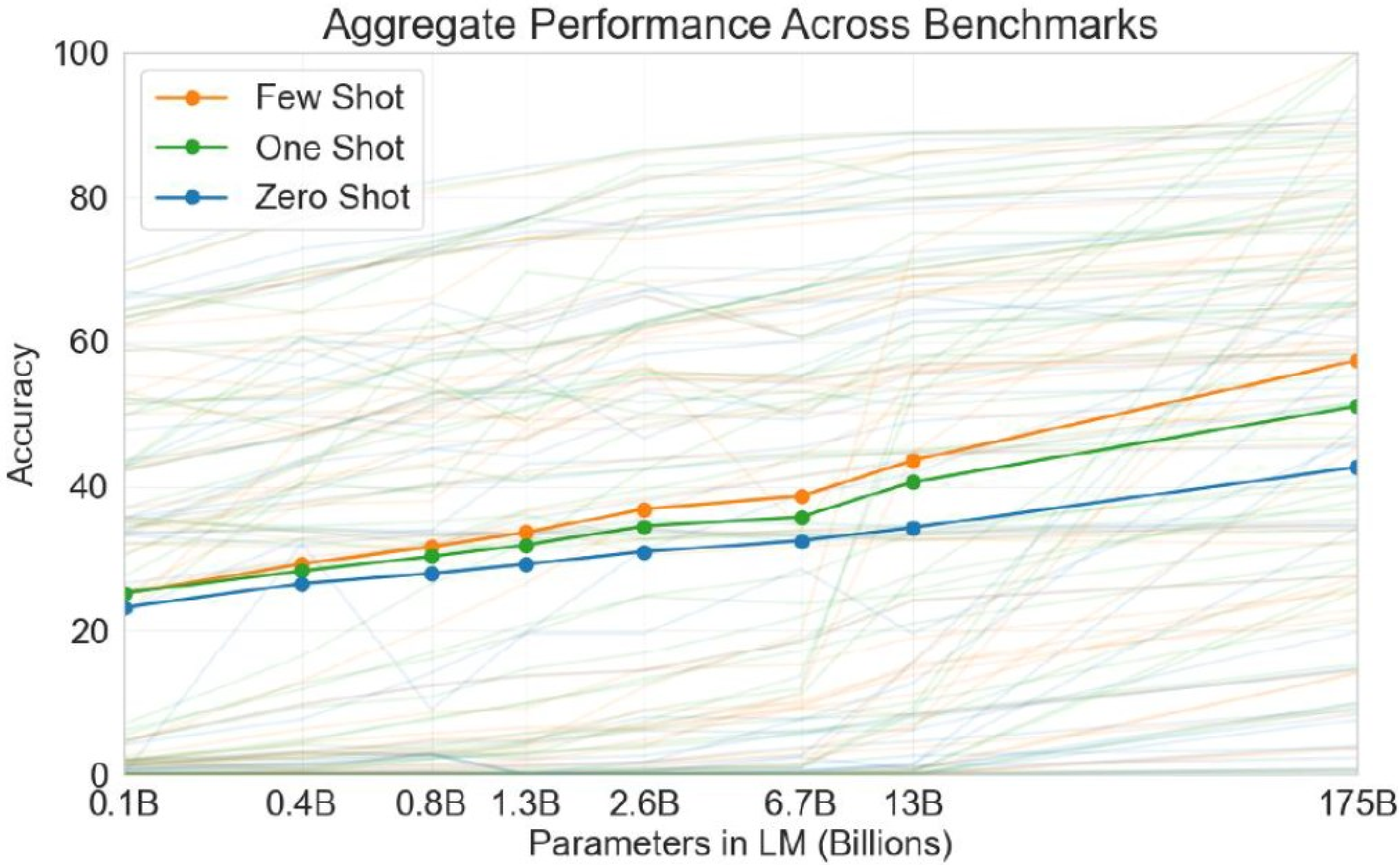
Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

```
Translate English to French:
sea otter => loutre de mer
peppermint => menthe poivrée
plush girafe => girafe peluche
cheese =>
```

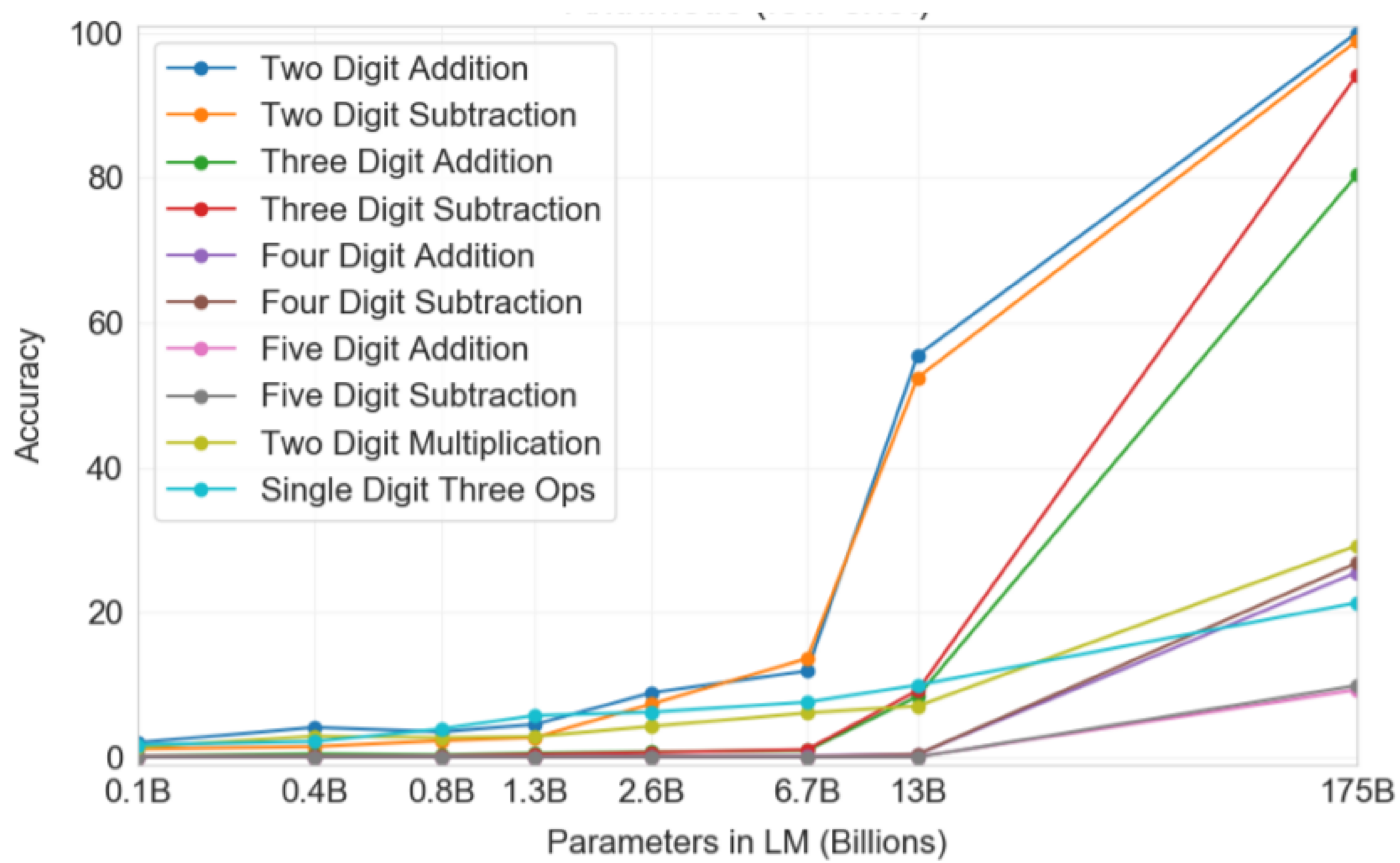
task description
examples
prompt

https://arxiv.org/abs/2005.14165



Average of **42** tasks https://arxiv.org/abs/2005.14165

- **2 digit addition (2D+)** – The model is asked to add two integers sampled uniformly from $[0, 100)$, phrased in the form of a question, e.g. "Q: What is 48 plus 76? A: 124."
- **2 digit subtraction (2D-)** – The model is asked to subtract two integers sampled uniformly from $[0, 100)$; the answer may be negative. Example: "Q: What is 34 minus 53? A: -19".
- **3 digit addition (3D+)** – Same as 2 digit addition, except numbers are uniformly sampled from $[0, 1000)$.



https://arxiv.org/abs/2005.141