Derin Öğrenme Tadım Turu

Ege Demir

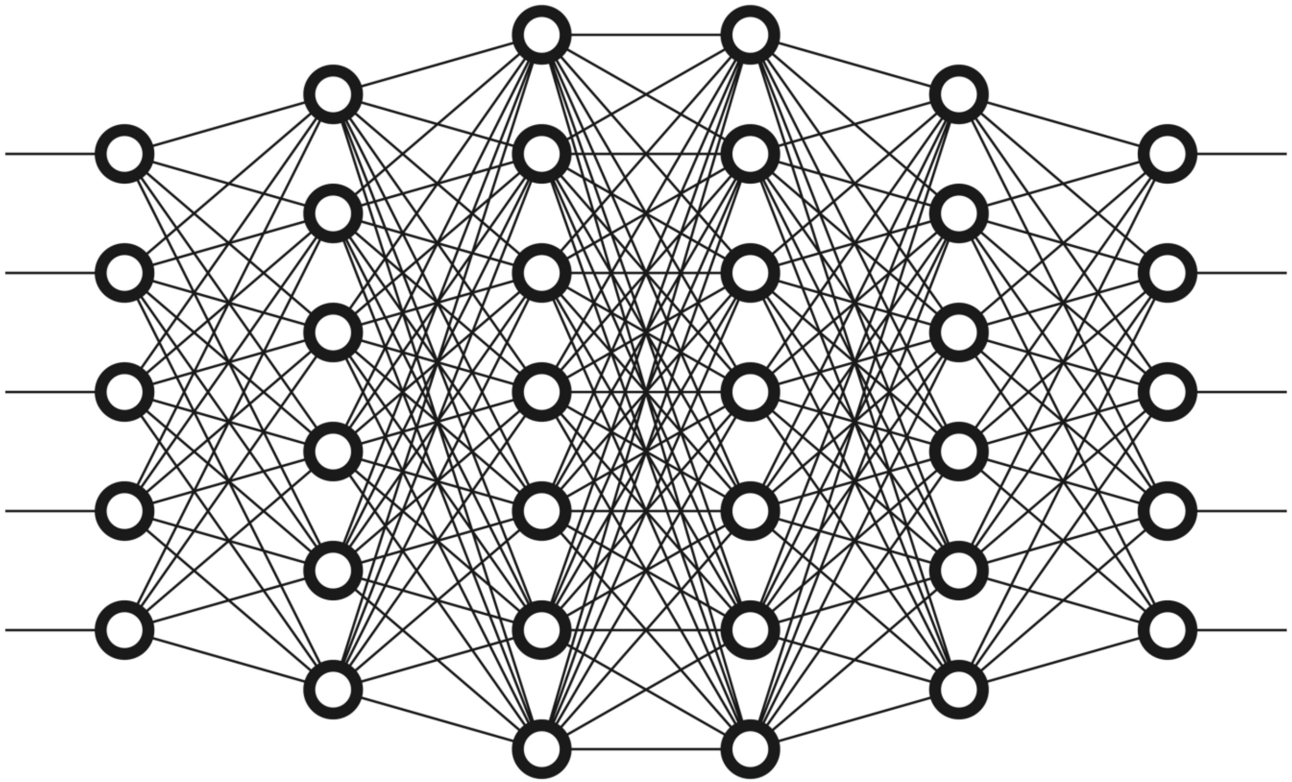
04.06.2021

Bugün Ne Yapacağız

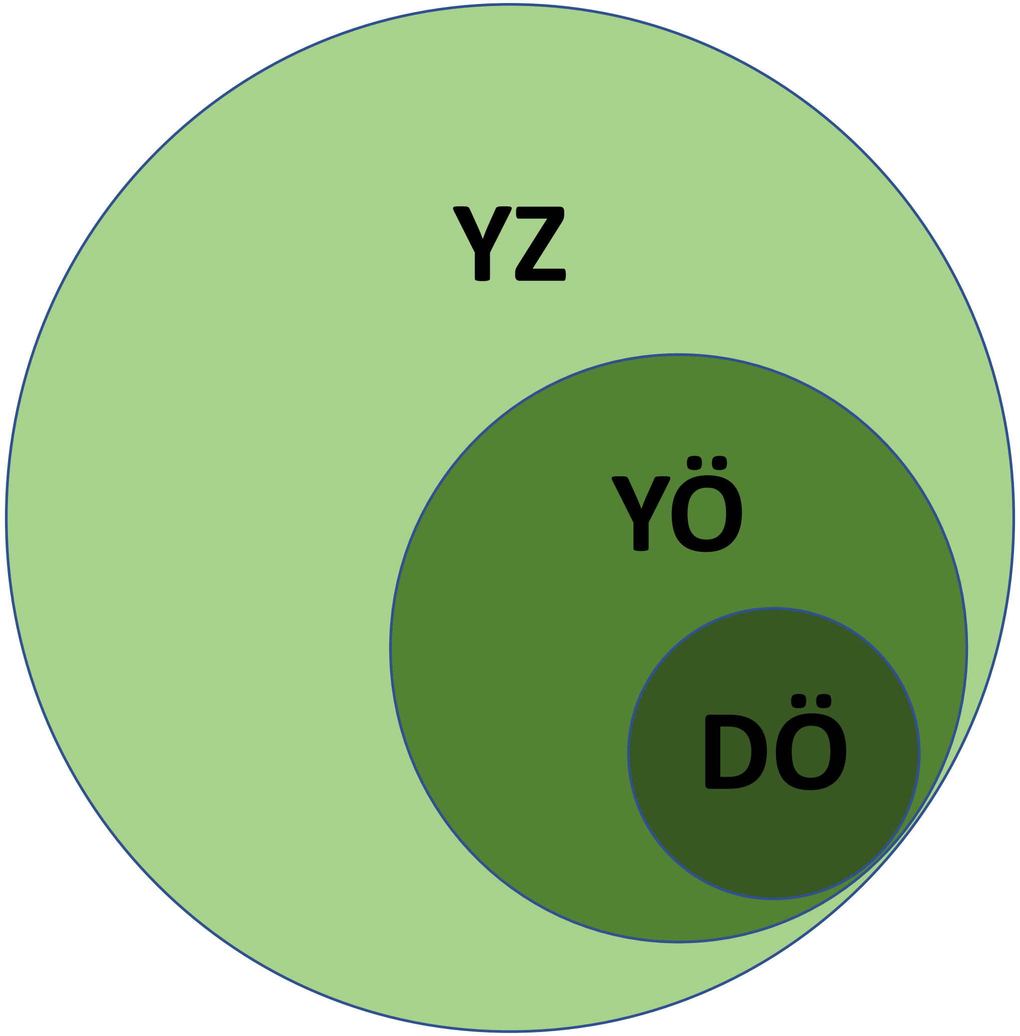
1. Derin Öğrenme Nedir?
    i. XOR ve Representation Learning
2. Doğal Dil İşleme
    i. Text Generation: GPT
    ii. Kodlama: CodeBERT
    iii. Alıntılama: Pegasus
3. Bilgisayarlı Görü
    i. Nesne Tespiti: YOLO
    ii. Segmentasyon: DINO
4. Generative Models
    i. Görsel Üretimi: StyleGAN
    ii. Yazı ile Görsel Üretimi: StyleCLIP

Derin Öğrenme Nedir?

Connectionism: https://plato.stanford.edu/entries/connectionism/

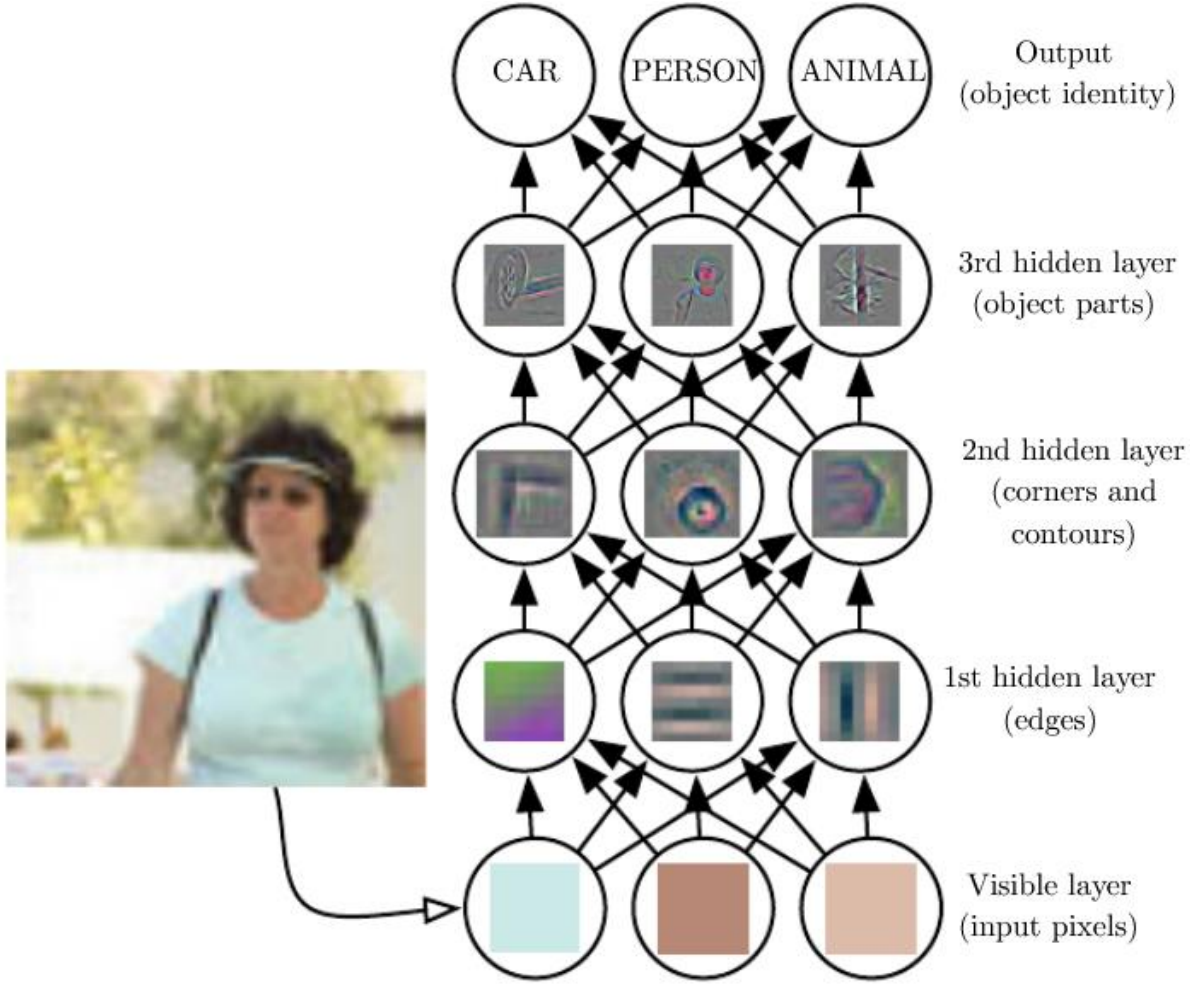


Derin Öğrenme Nedir?



Derin Öğrenme Nedir?

- **Hiyerarşi**
- «Representation»
- Bolluk



https://www.deeplearningbook.org/contents/intro.html

Derin Öğrenme Nedir?

- Hiyerarşi
- **«Representation»**
- Bolluk

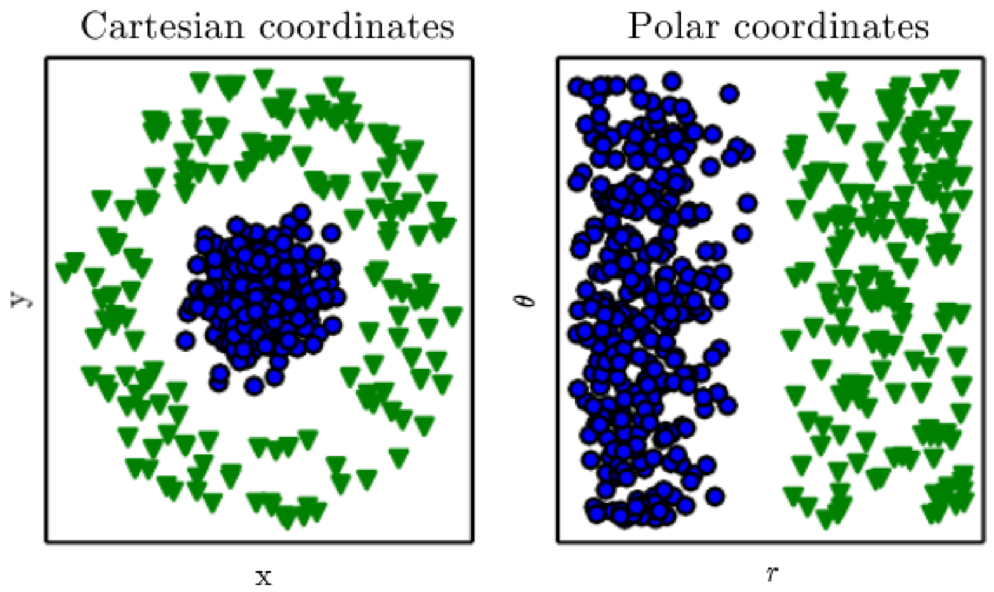
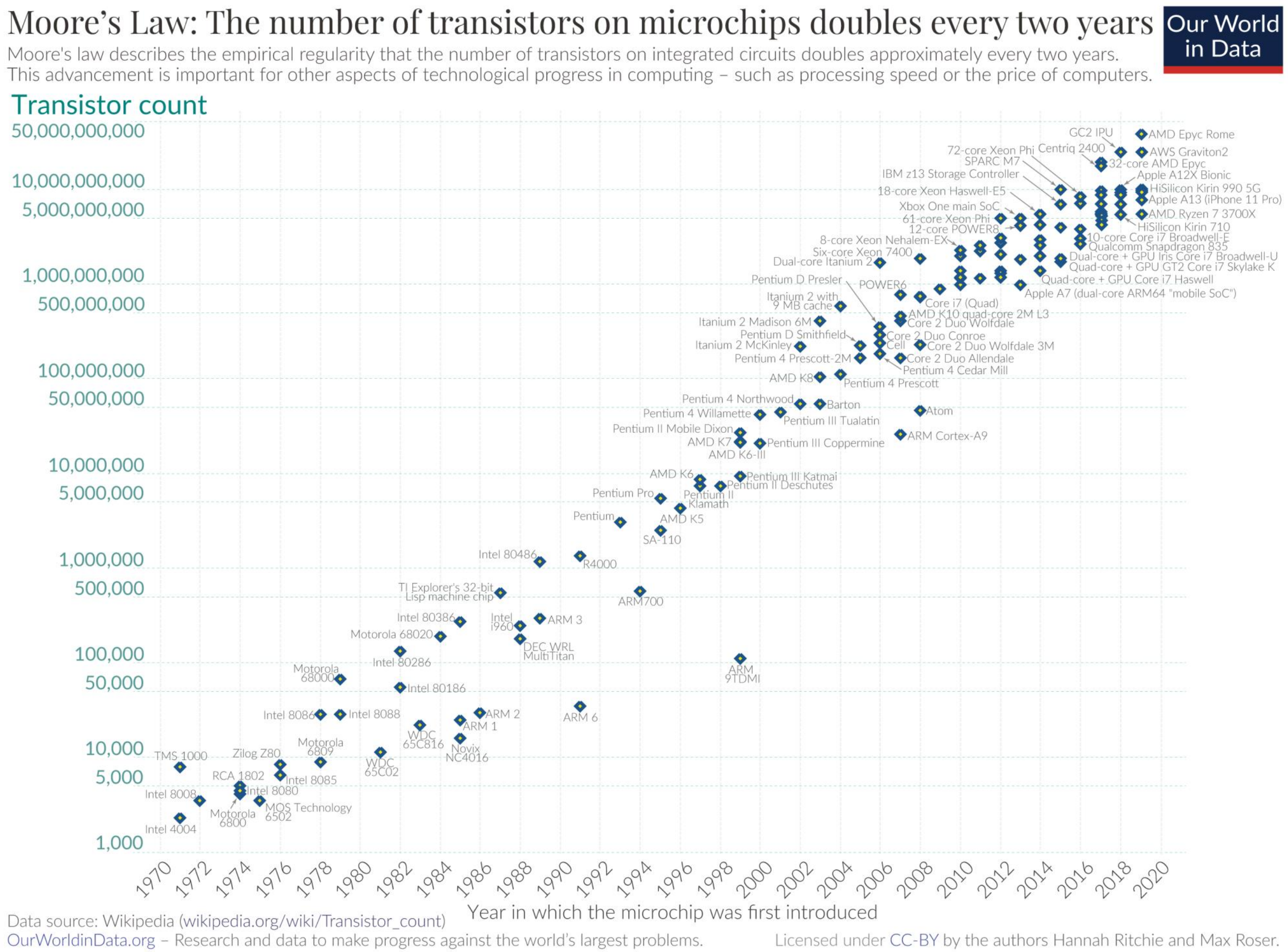


Figure 1.1: Example of different representations: suppose we want to separate two categories of data by drawing a line between them in a scatterplot. In the plot on the left, we represent some data using Cartesian coordinates, and the task is impossible. In the plot on the right, we represent the data with polar coordinates and the task becomes simple to solve with a vertical line. (Figure produced in collaboration with David Warde-Farley.)

https://www.deeplearningbook.org/contents/intro.html

Derin Öğrenme Nedir?

- Hiyerarşi
- «Representation»
- **Bolluk**

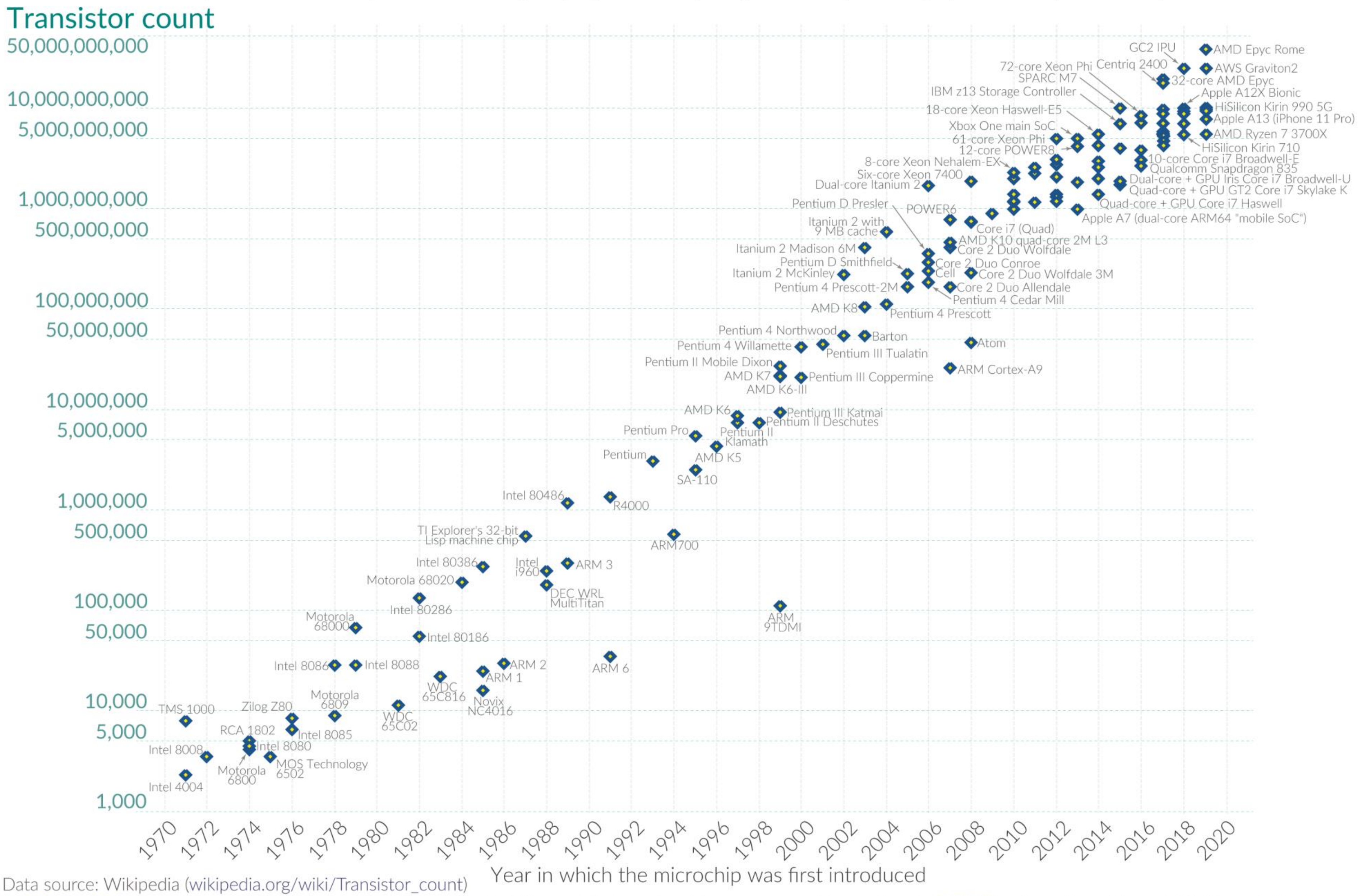


https://en.wikipedia.org/wiki/Moore%27s_law

Moore's Law: The number of transistors on microchips doubles every two years

Our World in Data

Moore's law describes the empirical regularity that the number of transistors on integrated circuits doubles approximately every two years. This advancement is important for other aspects of technological progress in computing – such as processing speed or the price of computers.



Data source: Wikipedia (wikipedia.org/wiki/Transistor_count)

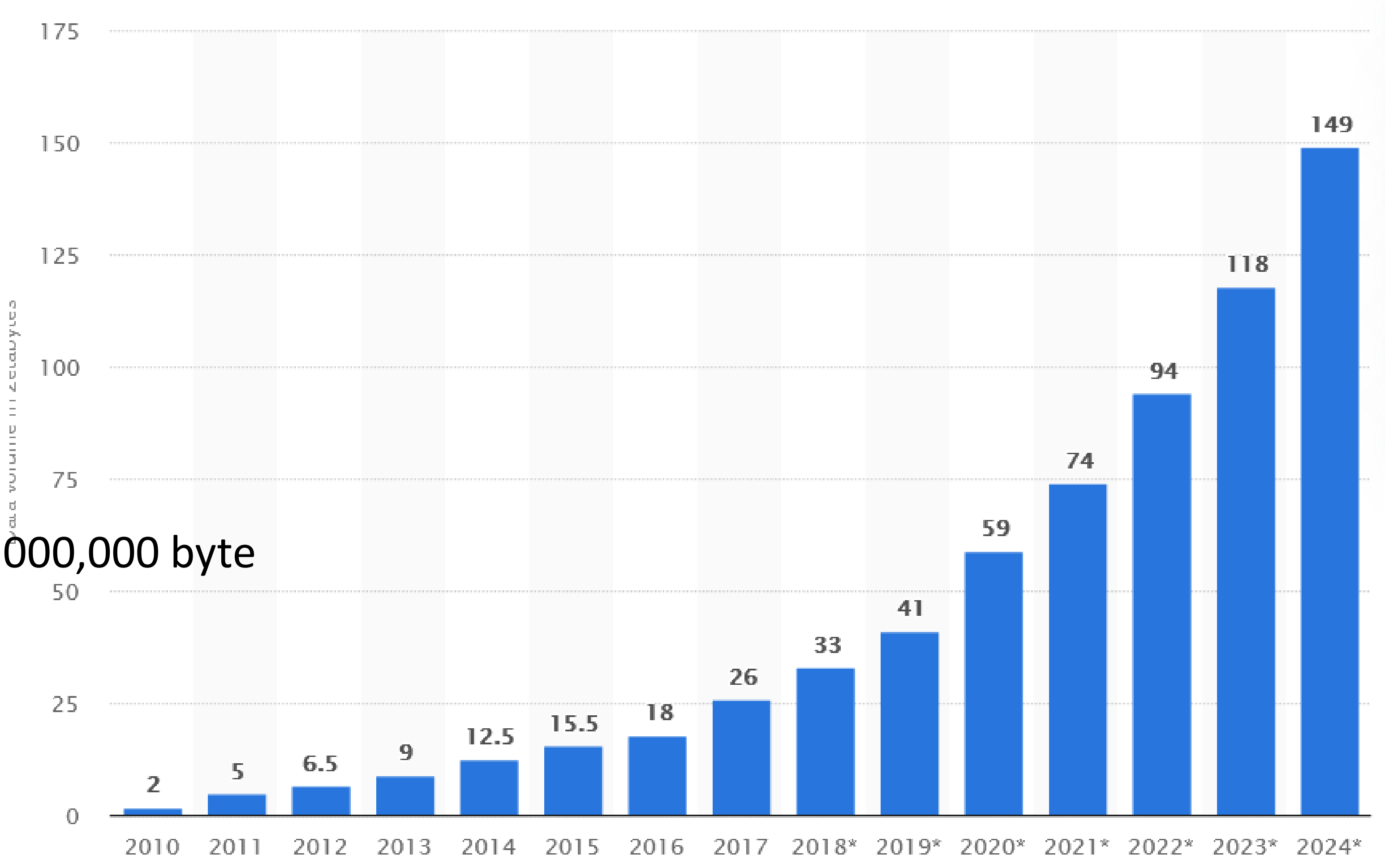
OurWorldinData.org – Research and data to make progress against the world's largest problems.

Derin Öğrenme Nedir?

- Hiyerarşi
- «Representation»
- **Bolluk**

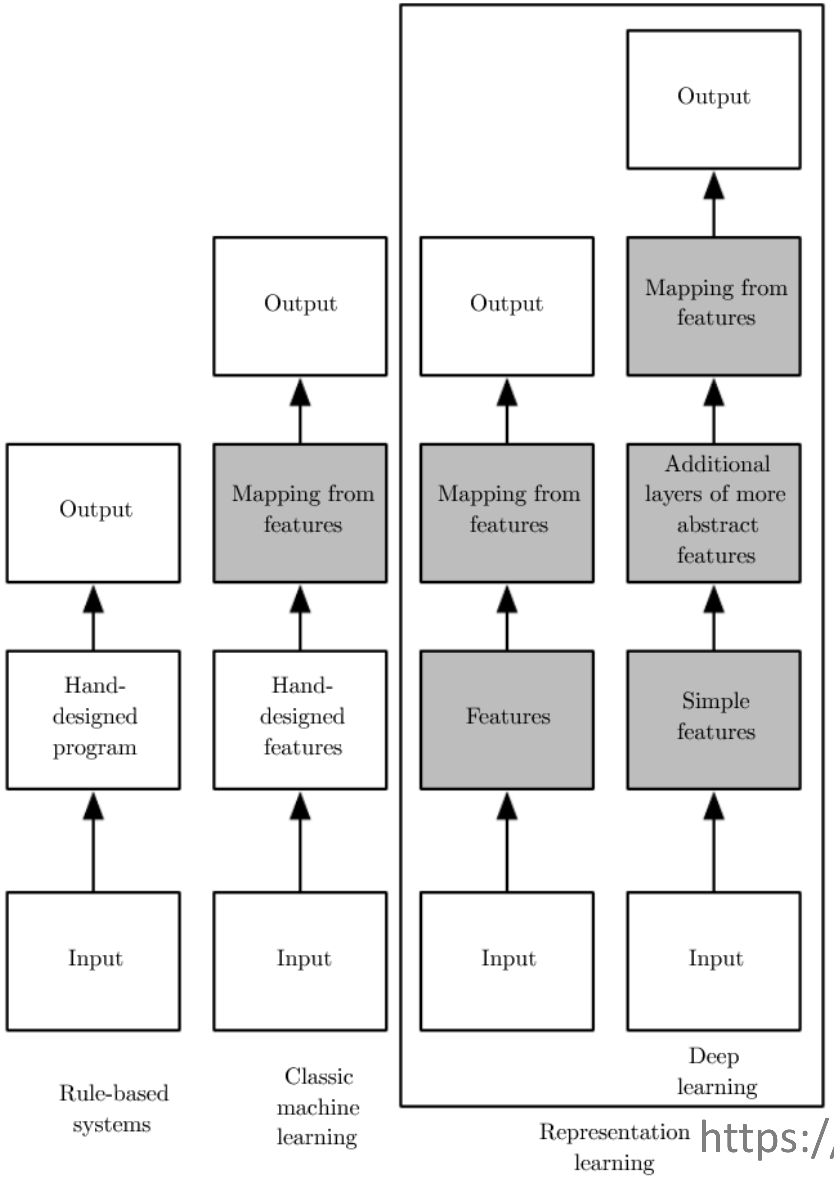
Gigabyte: 1,000,000,000 byte

Zettabyte: 1,000,000,000,000,000,000,000 byte



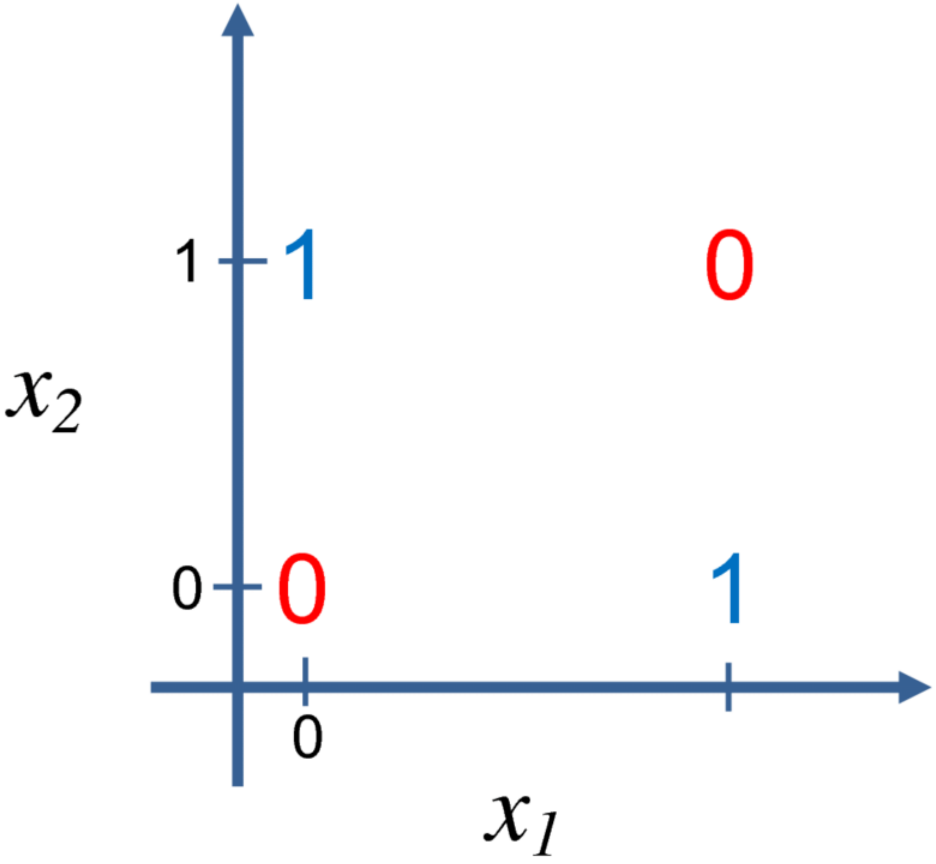
https://www.statista.com/statistics/871513/worldwide-data-created/

Representation Learning Nedir?

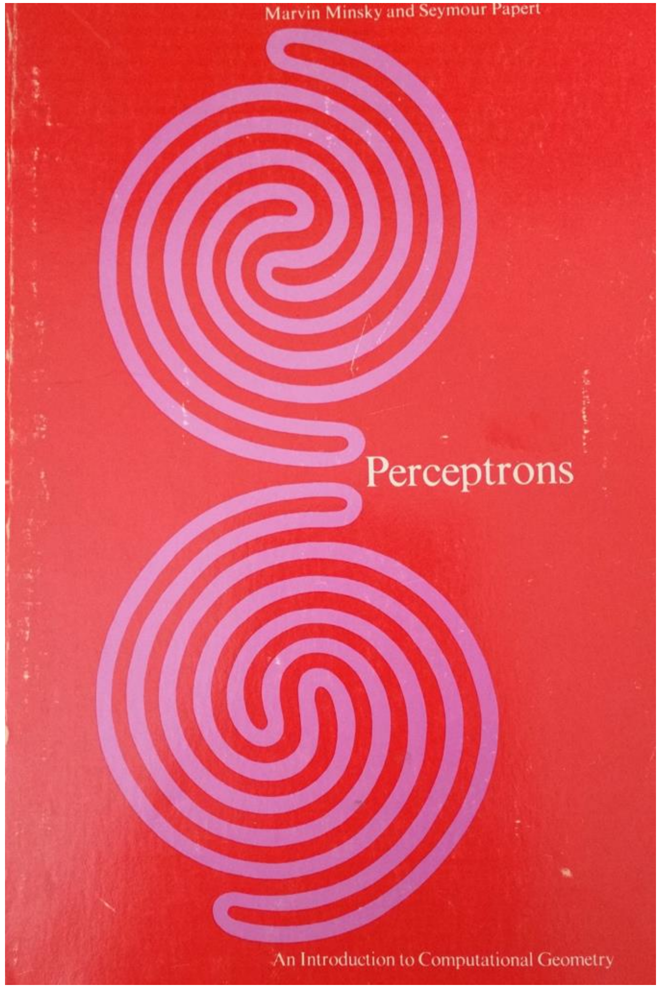


https://www.deeplearningbook.org/contents/intro.html

XOR’u öğrenmek

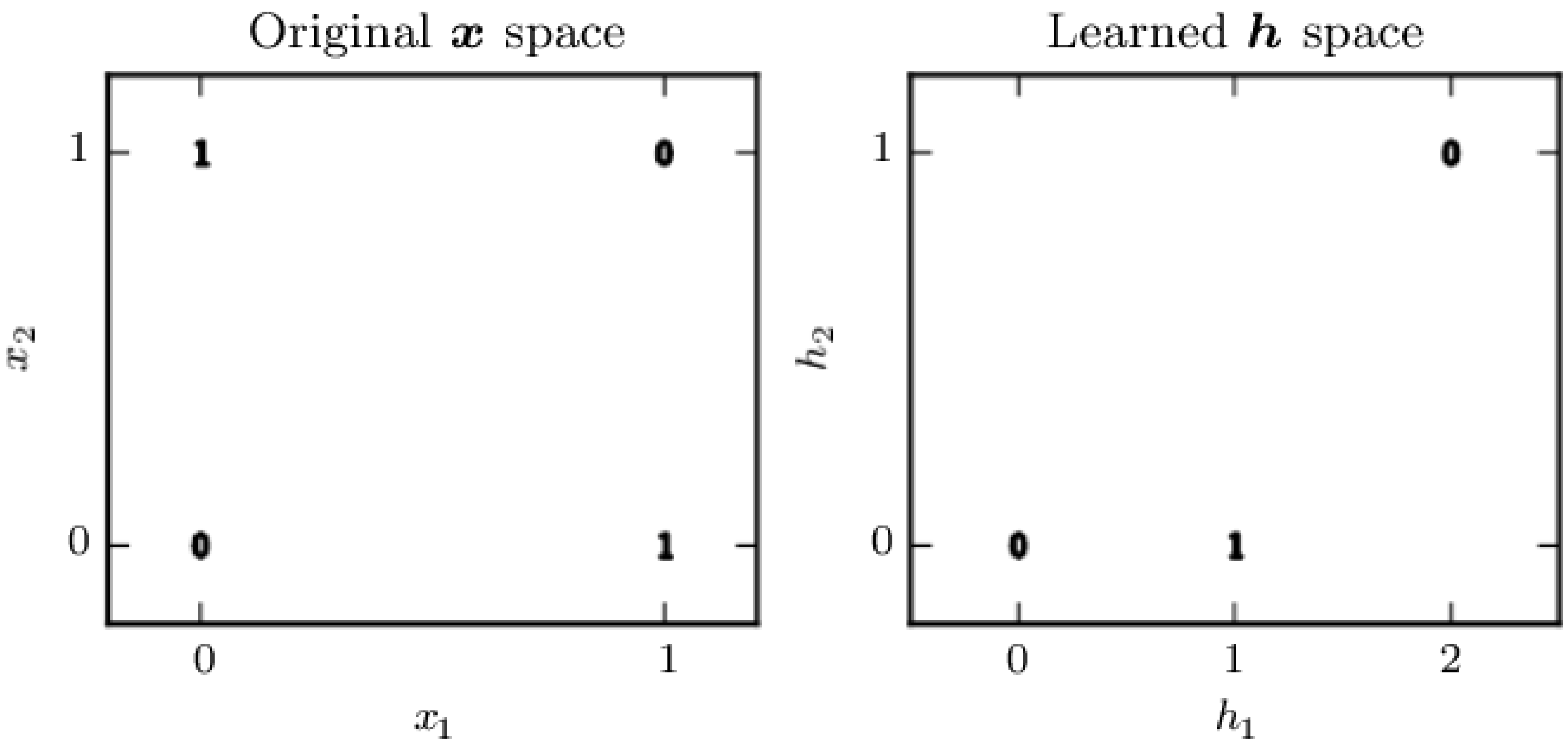


Contemporary machine learning: a guide for practitioners in the physical sciences, Brian K Spears



Marvin Minsky, Perceptrons 1969

XOR’u öğrenmek



https://dar.vin/xor

Deep Learning Book

Doğal Dil İşleme

- Natural Language Processing
- Natural Language Understanding
- Natural Language Inference
- https://paperswithcode.com/

- 2018'den önce/2018'den sonra
- Transformer
- Pre-training

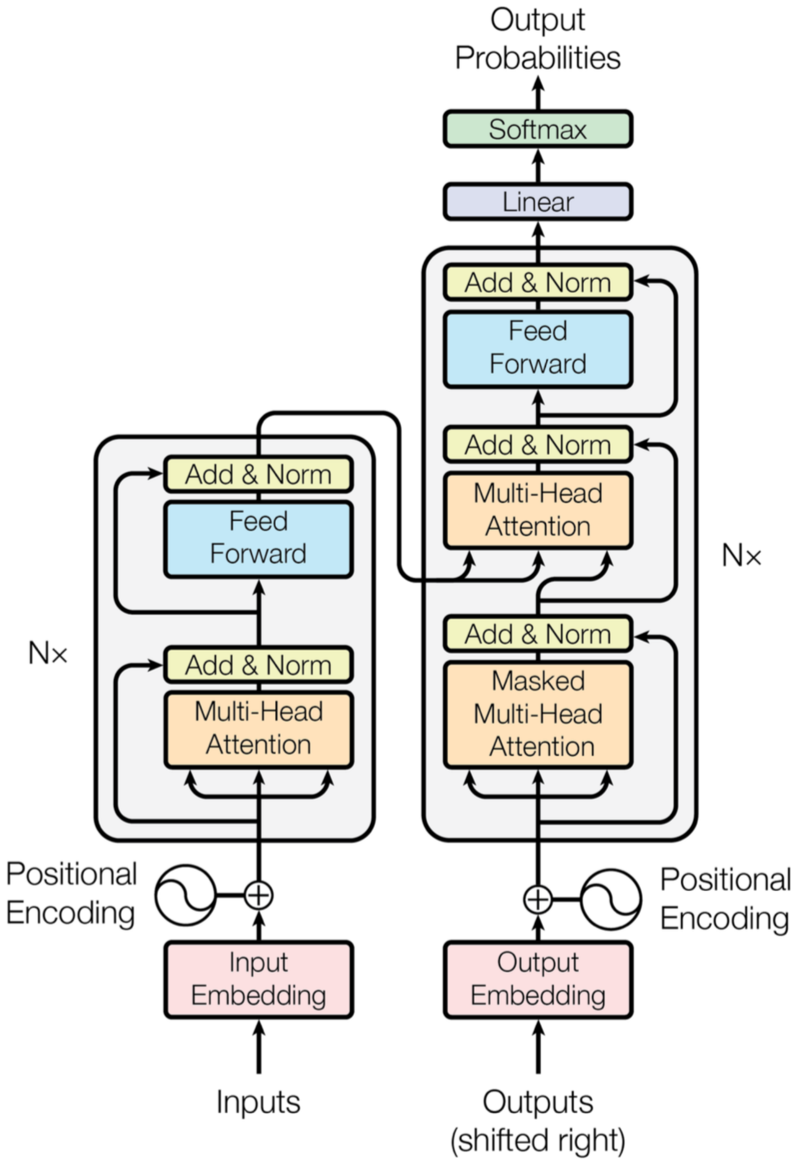


Figure 1: The Transformer - model architecture.

Attention is All You Need, 2017

GPT 1/2/3

- Generative Pre-Trained Transformer
- GPT3: 28 Mayıs 2020
- OpenAI
- 175 milyar parametre! -> 1 Haziran 2021 Wung Dao: **1.75 trilyon** parametre!
- 4.6 milyon dolar!
- Semi-supervised learning

https://transformer.huggingface.co/doc/gpt2-large

https://stefanzukin.com/enigma/

https://play.aidungeon.io/

https://dar.vin/gpt

GPT 1/2/3

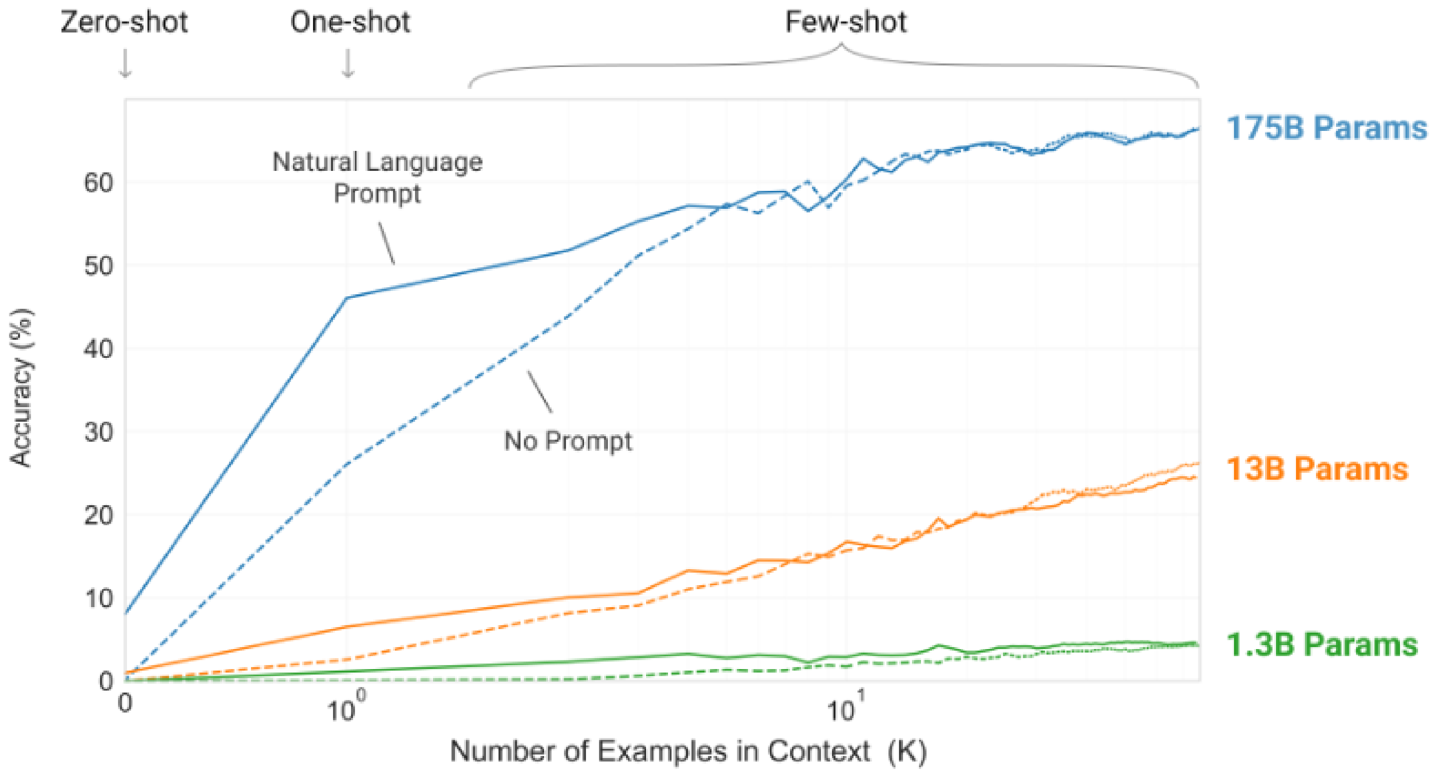
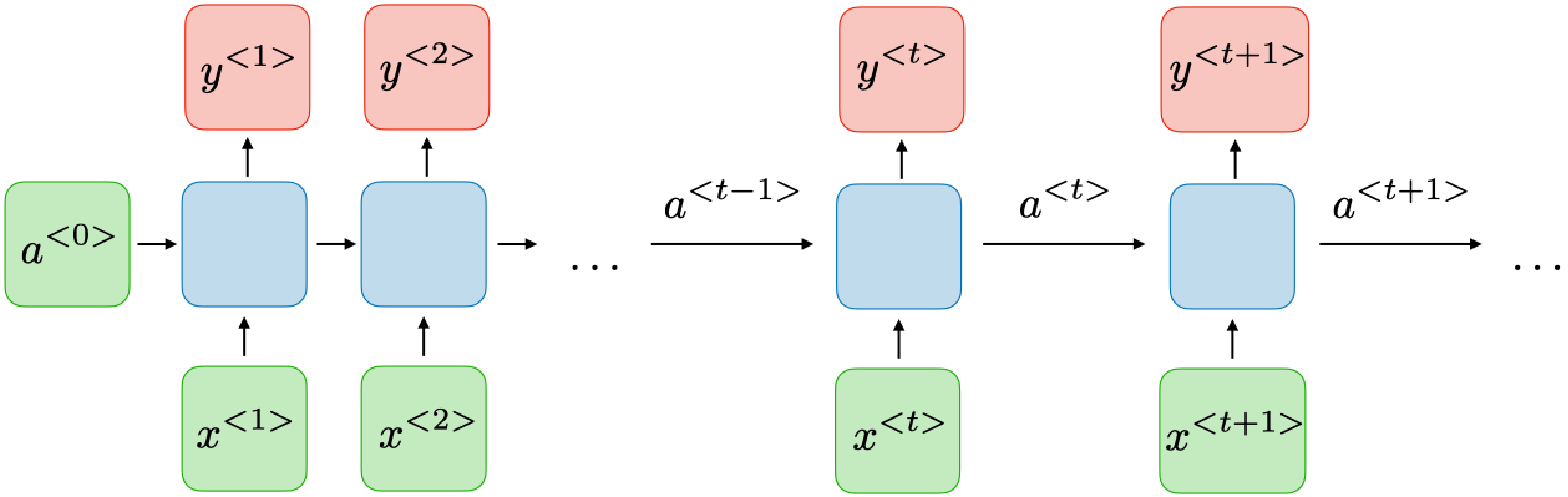
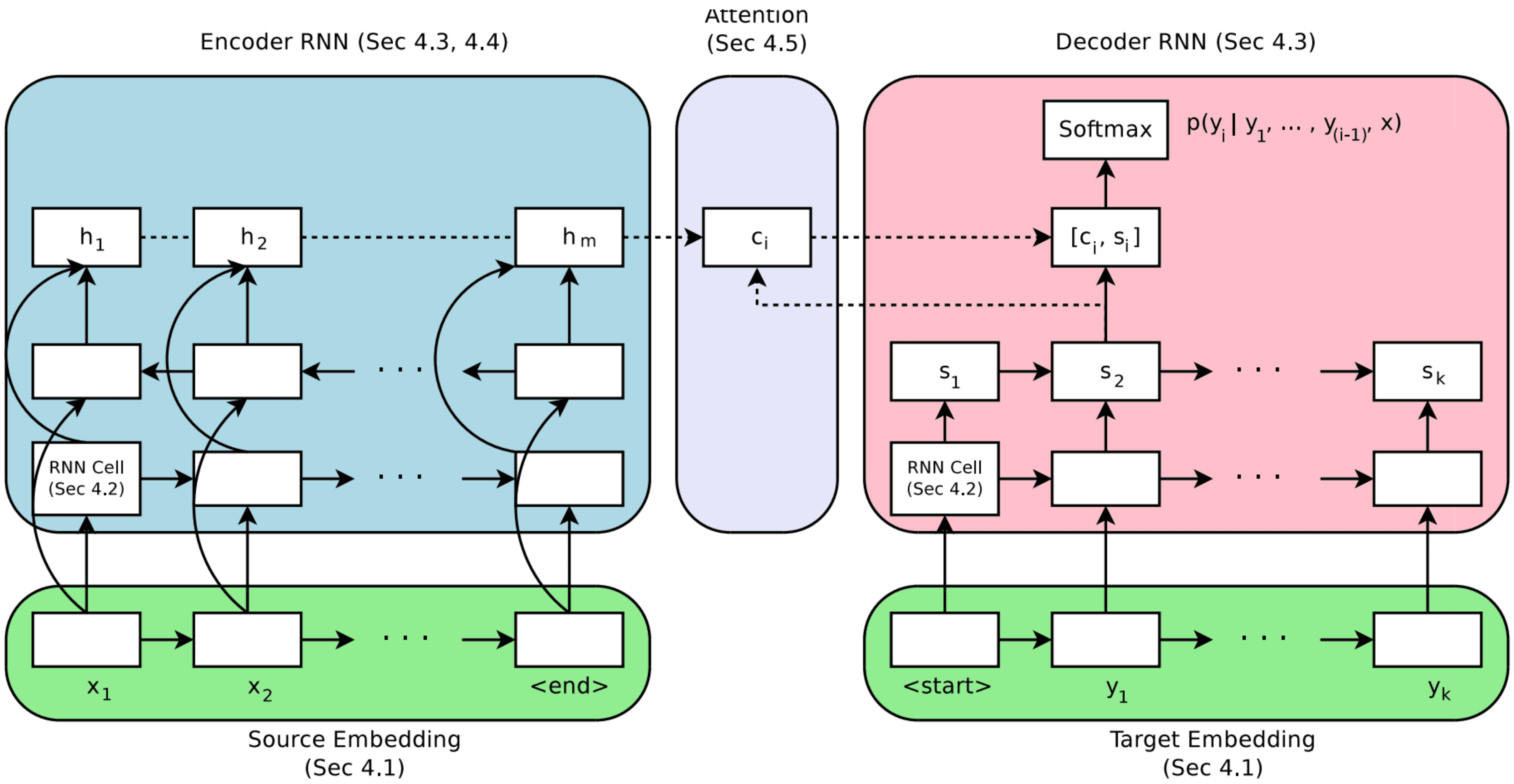


Figure 1.2: Larger models make increasingly efficient use of in-context information. We show in-context learning performance on a simple task requiring the model to remove random symbols from a word, both with and without a natural language task description (see Sec. 3.9.2). The steeper "in-context learning curves" for large models demonstrate improved ability to learn a task from contextual information. We see qualitatively similar behavior across a wide range of tasks.

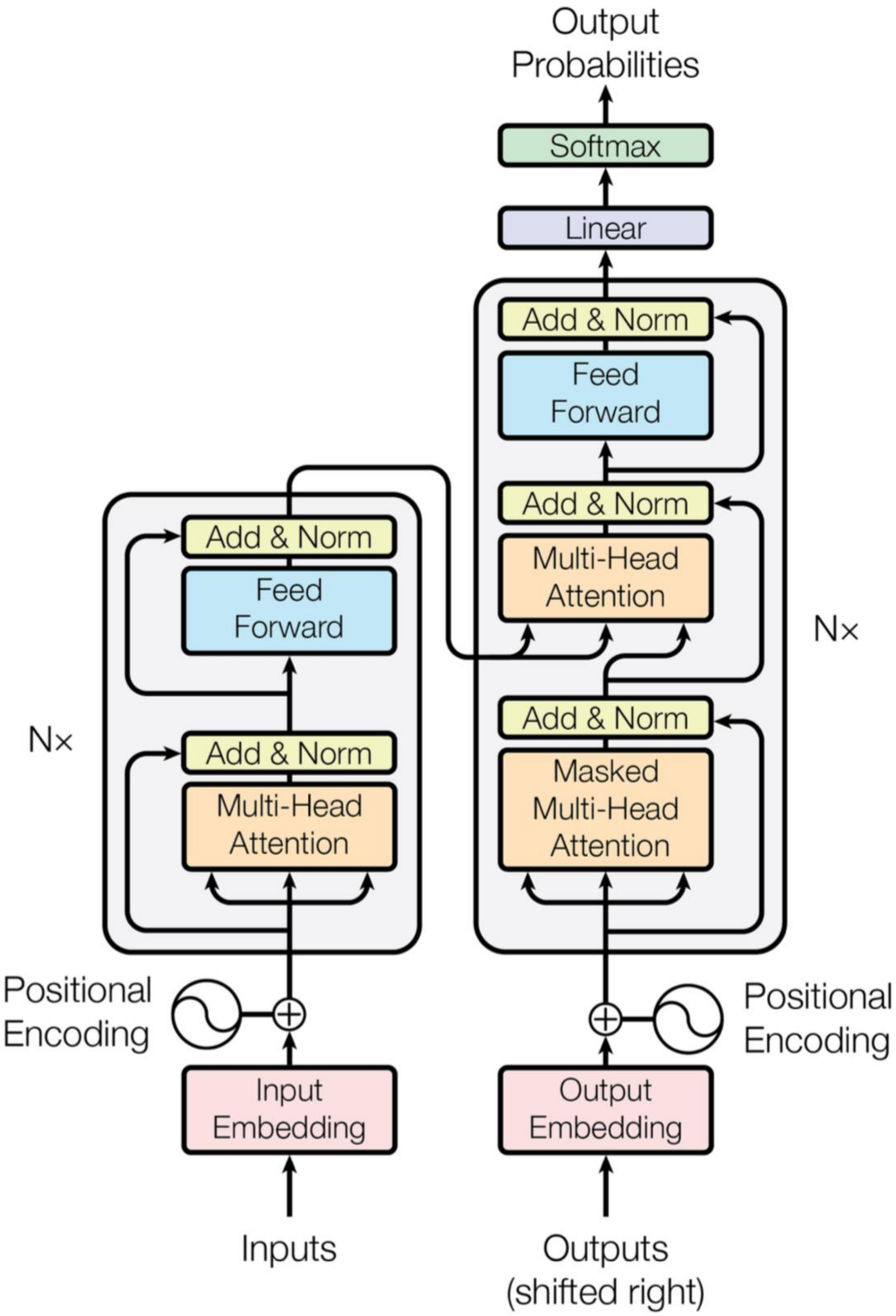
Recurrent Neural Network



Encoder – Decoder ve Attention



How to Configure an Encoder-Decoder Model for Neural Machine Translation, Machine Learning Mastery


Output
Probabilities
Softmax
Linear
Add & Norm
Feed
Forward
Add & Norm
Multi-Head
Attention
Add & Norm
Masked
Multi-Head
Attention
Nx
Add & Norm
Feed
Forward
Add & Norm
Multi-Head
Attention
Nx
Positional
Encoding
Positional
Encoding
Input
Embedding
Output
Embedding
Inputs
Outputs
(shifted right)

$$\underset{\text{yemek}}{\begin{bmatrix} 2 & -1 \end{bmatrix}} \bullet \underset{\text{koşarken}}{\begin{bmatrix} -3 \\ 1 \end{bmatrix}} = 2*(-3) + (-1)*1 = \mathbf{-7}$$

$$\underset{\text{yemek}}{\begin{bmatrix} 2 & -1 \end{bmatrix}} \bullet \underset{\text{yemek}}{\begin{bmatrix} 2 \\ -1 \end{bmatrix}} = \mathbf{5}$$

$$\underset{\text{yemek}}{\begin{bmatrix} 2 & -1 \end{bmatrix}} \bullet \underset{\text{yedik}}{\begin{bmatrix} 2 \\ 1 \end{bmatrix}} = \mathbf{3} \tag{1}$$

$$
\begin{aligned}
&\mathit{softmax}(-7) = 10^{-6} \\
&\mathit{softmax}(5) = 0.88 \\
&\mathit{softmax}(3) = 0.12
\end{aligned}
\tag{2}
$$

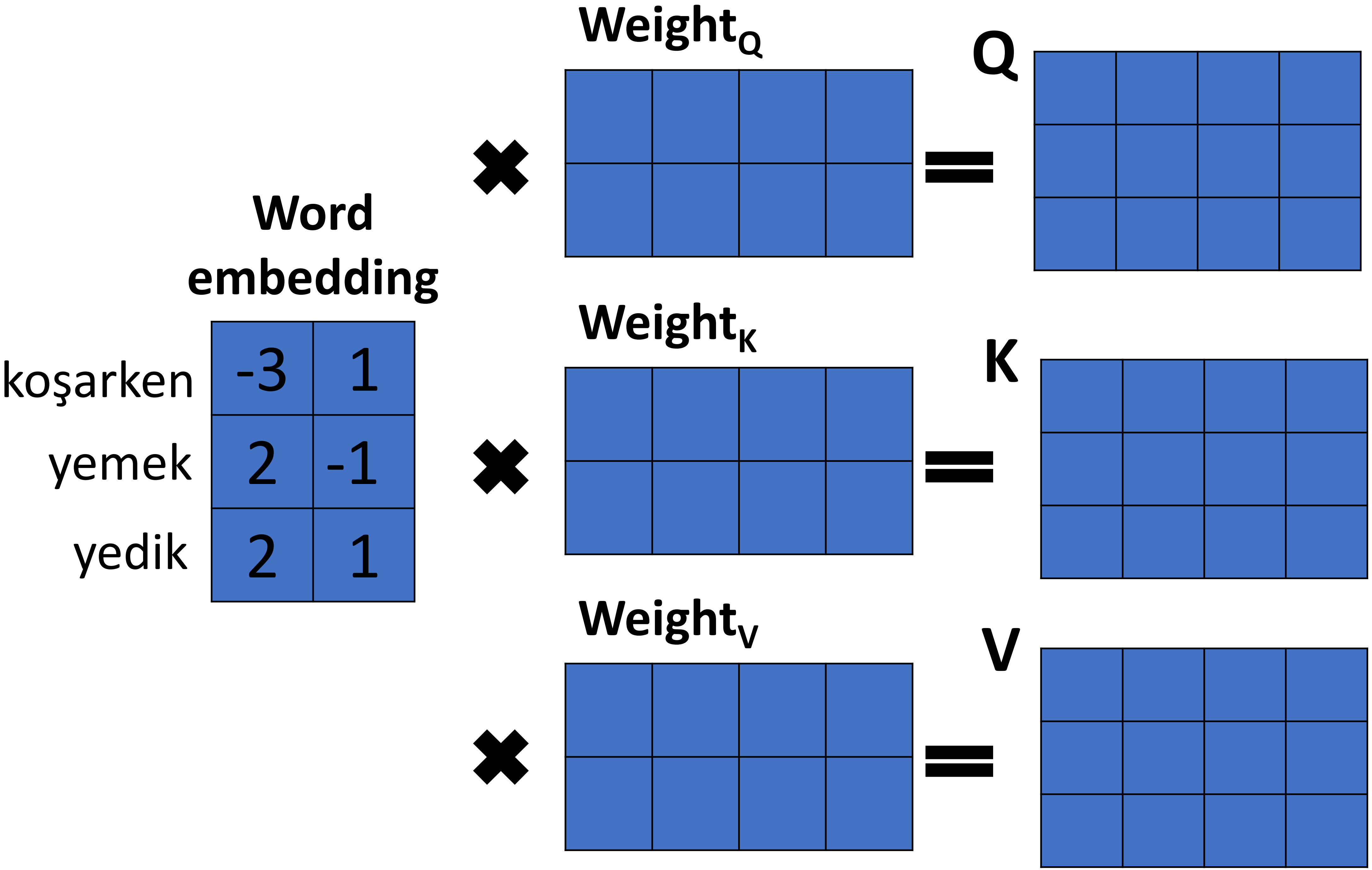
$$
\begin{aligned}
&10^{-6}\,\overset{\textbf{koşarken}}{\begin{bmatrix}-3 & 1\end{bmatrix}} + 0.88\,\overset{\textbf{yemek}}{\begin{bmatrix}2 & -1\end{bmatrix}} + 0.12\,\overset{\textbf{yedik}}{\begin{bmatrix}2 & 1\end{bmatrix}} \\
&= \overset{\textbf{yemek'}}{\begin{bmatrix}2 & -0.76\end{bmatrix}}
\end{aligned}
\tag{3}
$$

$$attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{4}$$

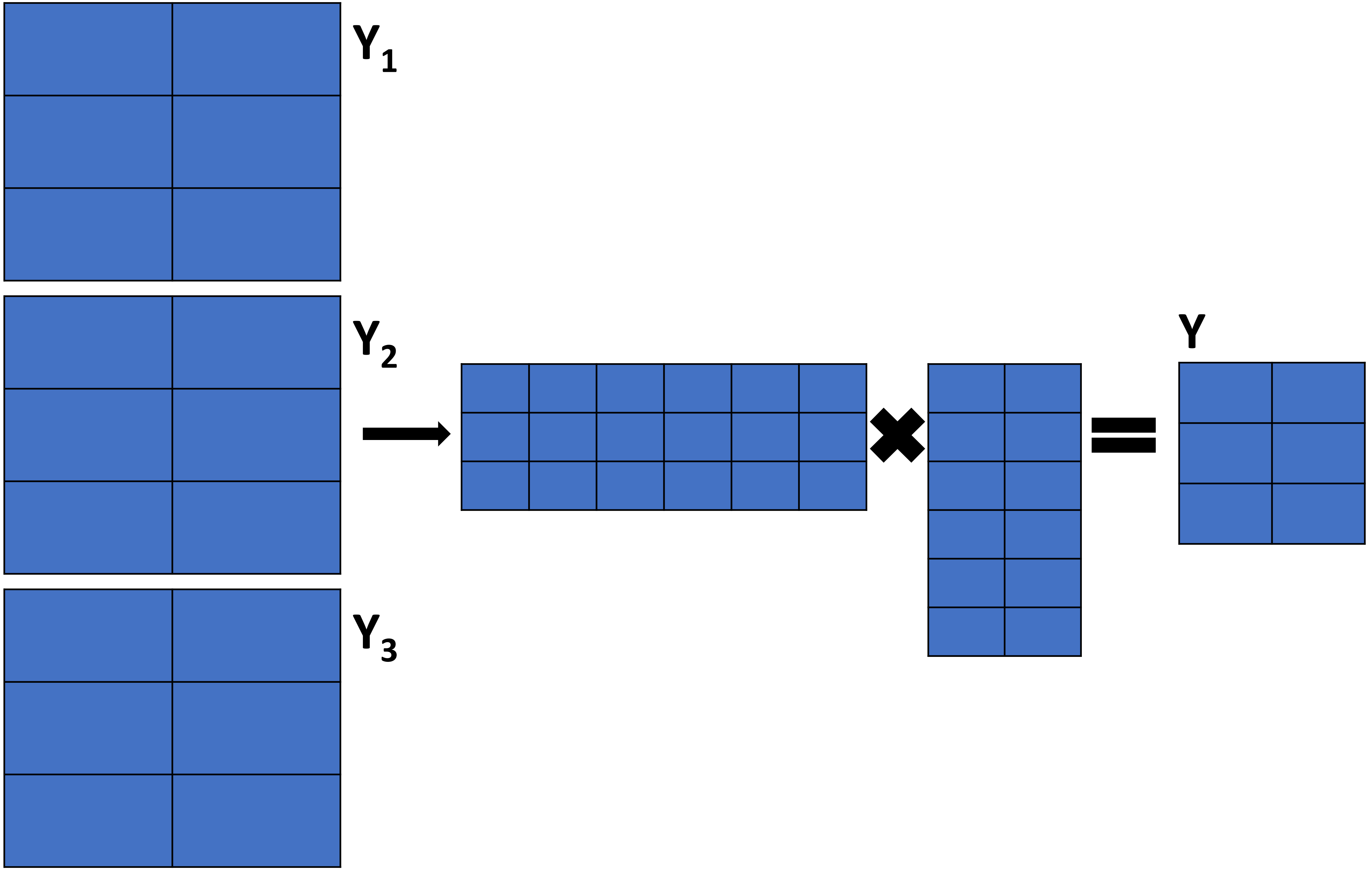
$$
\begin{array}{c}
\\
\text{koşarken} \\
\text{yemek} \\
\text{yedik}
\end{array}
\overset{\textbf{Query}}{\begin{bmatrix} -3 & 1 \\ 2 & -1 \\ 2 & 1 \end{bmatrix}}
\times
\begin{array}{c}
\textbf{Key}^{T} \\
\begin{bmatrix} -3 & 2 & 2 \\ 1 & -1 & 1 \end{bmatrix} \\
\begin{array}{ccc} \text{koşarken} & \text{yemek} & \text{yedik} \end{array}
\end{array}
=
\begin{array}{cc}
 & \\
\begin{array}{c} \text{koşarken} \\ \text{yemek} \\ \text{yedik} \end{array} &
\begin{bmatrix} 10 & -7 & -5 \\ -7 & 5 & 3 \\ -5 & 3 & 5 \end{bmatrix} \\
 & \begin{array}{ccc} \text{koşarken} & \text{yemek} & \text{yedik} \end{array}
\end{array}
\tag{5}
$$

$$softmax\left(\frac{\begin{bmatrix} 10 & -7 & -5 \\ -7 & 5 & 3 \\ -5 & 3 & 5 \end{bmatrix}}{\sqrt{2}}\right) = \begin{bmatrix} 0.99 & 10^{-6} & 10^{-5} \\ 10^{-4} & 0.80 & 0.19 \\ 10^{-4} & 0.19 & 0.80 \end{bmatrix} \quad (6)$$

$$
\begin{bmatrix} 0.99 & 10^{-6} & 10^{-5} \\ 10^{-4} & 0.80 & 0.19 \\ 10^{-4} & 0.19 & 0.80 \end{bmatrix}
\times
\begin{array}{c} \textbf{Value} \\ \begin{bmatrix} -3 & 1 \\ 2 & -1 \\ 2 & 1 \end{bmatrix} \end{array}
=
\begin{array}{cl} \textbf{Y} & \\ \begin{bmatrix} -2.97 & 0.99 \\ 1.98 & -0.61 \\ 1.98 & 0.61 \end{bmatrix} & \begin{array}{l} \text{koşarken} \\ \text{yemek (7)} \\ \text{yedik} \end{array} \end{array}
$$


Word embedding
koşarken -3 1
yemek 2 -1
yedik 2 1
Weight$_Q$
Weight$_K$
Weight$_V$
Q
K
V

(8)



(9)

GPT 1/2/3

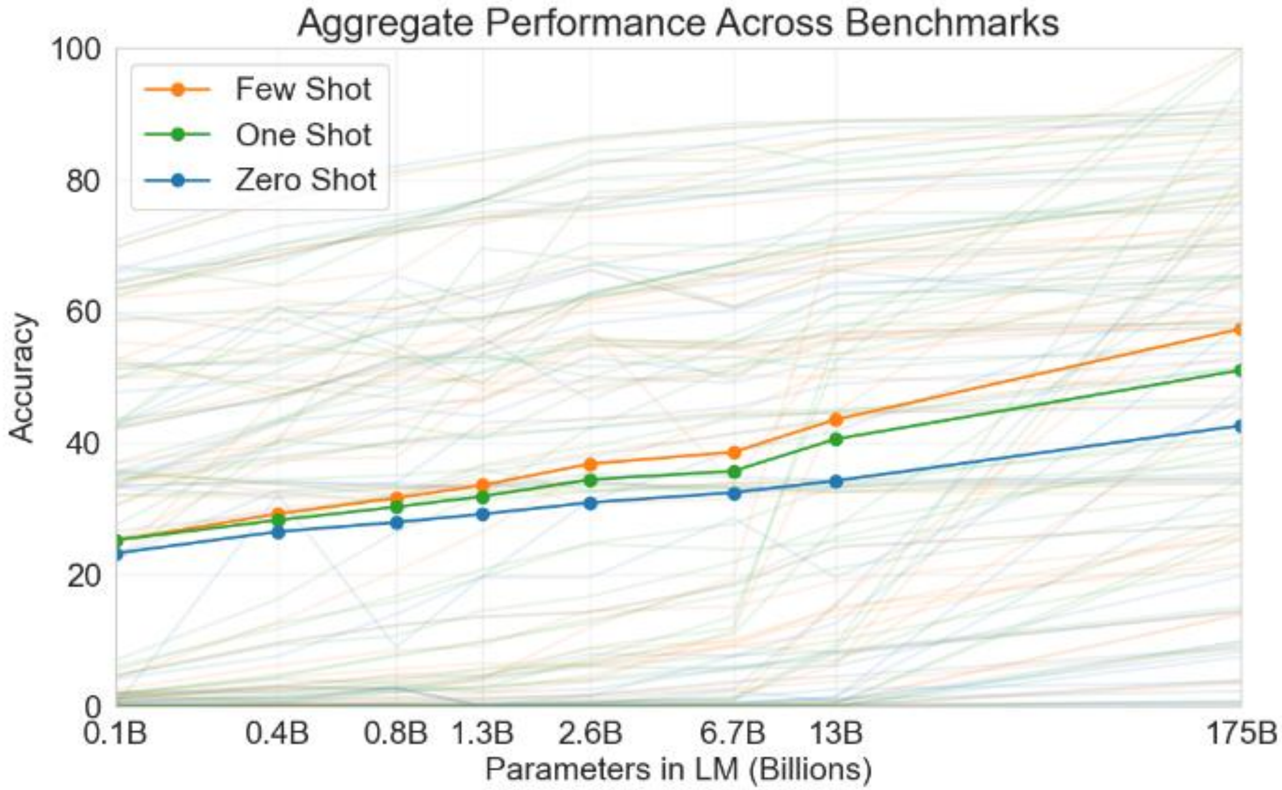


Figure 1.3: Aggregate performance for all 42 accuracy-denominated benchmarks While zero-shot performance improves steadily with model size, few-shot performance increases more rapidly, demonstrating that larger models are more proficient at in-context learning. See Figure 3.8 for a more detailed analysis on SuperGLUE, a standard NLP benchmark suite.

codeBERT

- Bidirectional Encoder Representations from Transformers
- 11 Ekim 2018
- Google
- Base – 110 milyon parametre
- Semi-supervised learning

https://dar.vin/codebert

Pegasus

- Pre-training with Extracted Gap-sentences for Abstractive Summarization
- 9 Haziran 2020
- Few-shot özetleme
- 568 milyon parametre
- Google
- Semi-supervised learning

https://dar.vin/pegasus

Eleştiriler

- *Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data* – Emily Bender et al.
    - In this paper, we have argued that in contrast to some current hype, meaning cannot be learned from form alone. This means that even large language models such as BERT do not learn "meaning"; they learn some reflection of meaning into the linguistic form which is very useful in applications.
- *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?* – Bender, Timnit Gebru et al.
    - We have identified a wide variety of costs and risks associated with the rush for ever larger LMs, including: environmental costs(borne typically by those not benefiting from the resulting technol-ogy); financial costs, which in turn erect barriers to entry, limiting who can contribute to this research area and which languages can benefit from the most advanced techniques; opportunity cost, as re-searchers pour effort away from directions requiring less resources ; and the risk of substantial harms, including stereotyping, denigration, increases in extremist ideology, and wrongful arrest, should humans encounter seemingly coherent LM output and take it forthe words of some person or organization who has accountability for what is said.

Bilgisayarlı Görü

- AlexNet! - 2012
- https://paperswithcode.com/sota/image-classification-on-imagenet
- Deep convolutional neural network

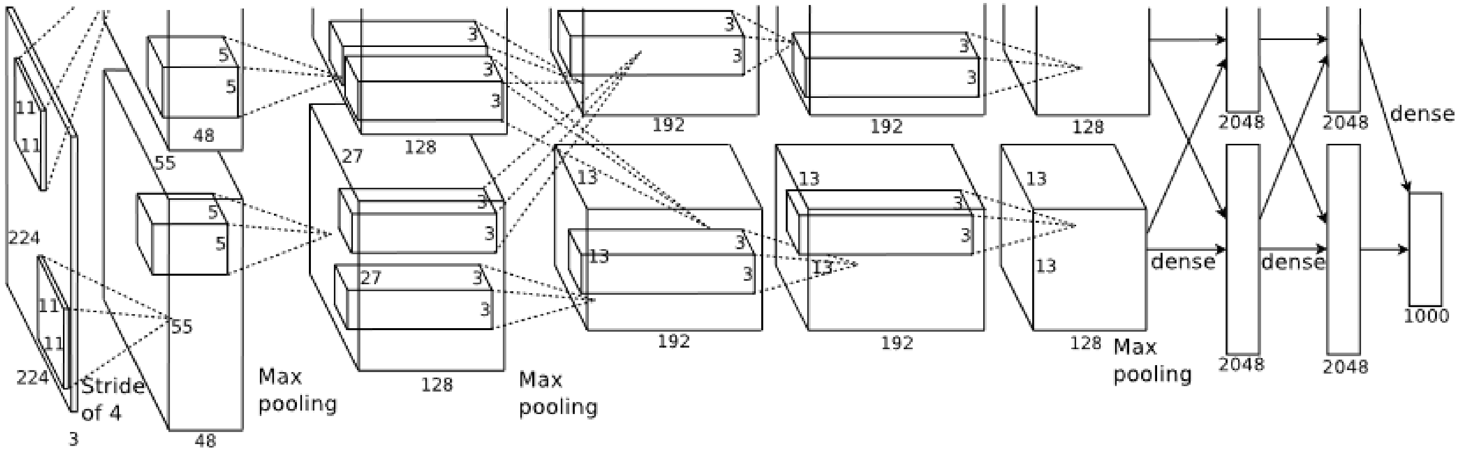


Figure 2: An illustration of the architecture of our CNN, explicitly showing the delineation of responsibilities between the two GPUs. One GPU runs the layer-parts at the top of the figure while the other runs the layer-parts at the bottom. The GPUs communicate only at certain layers. The network's input is 150,528-dimensional, and the number of neurons in the network's remaining layers is given by 253,440–186,624–64,896–64,896–43,264–4096–4096–1000.

YOLO v4

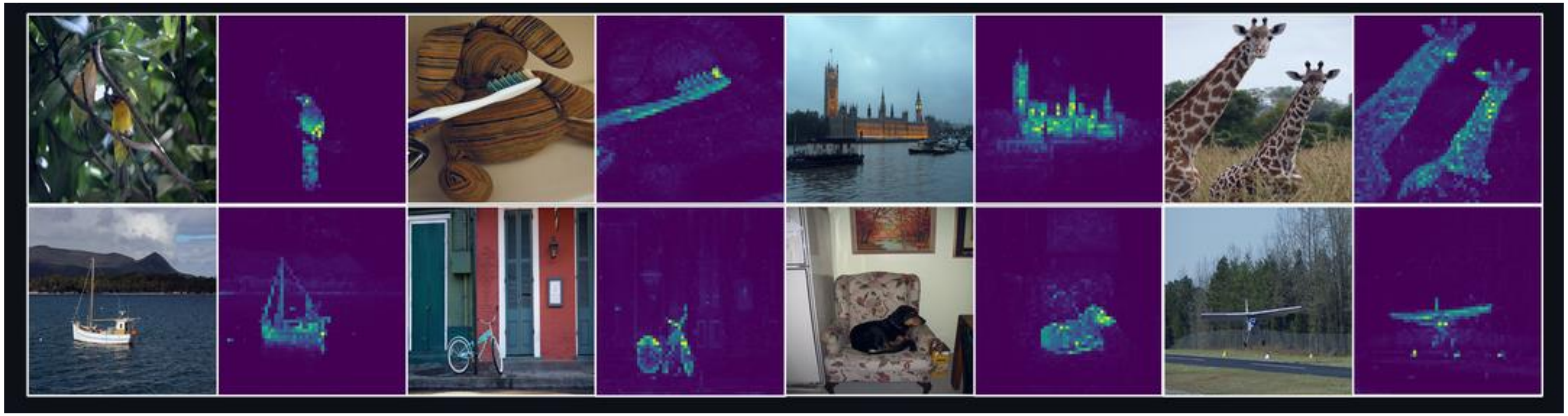
- YOLOv4: Optimal Speed and Accuracy of Object Detection
- 23 Nisan 2021
- Deep convolution ve birkaç numara daha
- Hızlı ve gerçek dünya kullanımı için

https://dar.vin/yolov4

DINO

- Emerging Properties in Self-Supervised Vision Transformers
- 30 Nisan 2021
- Self **di**stillation with **no** labels
- Self supervision + vision transformer
- https://arxiv.org/pdf/2104.14294.pdf

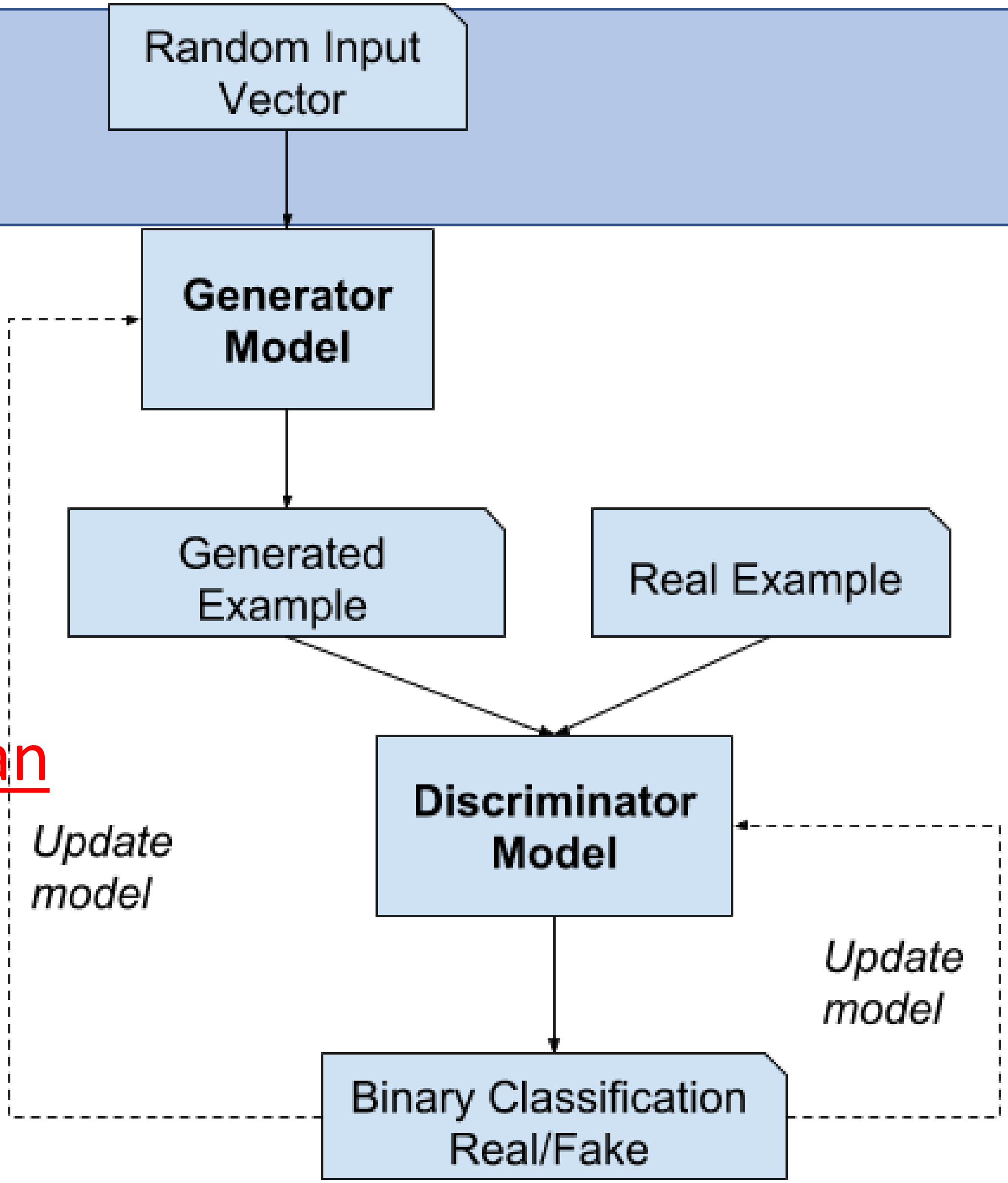
https://dar.vin/dino



GAN

- Generative Adversarial Network
- 2014

https://thispersondoesnotexist.com/
https://github.com/lucidrains/lightweight-gan

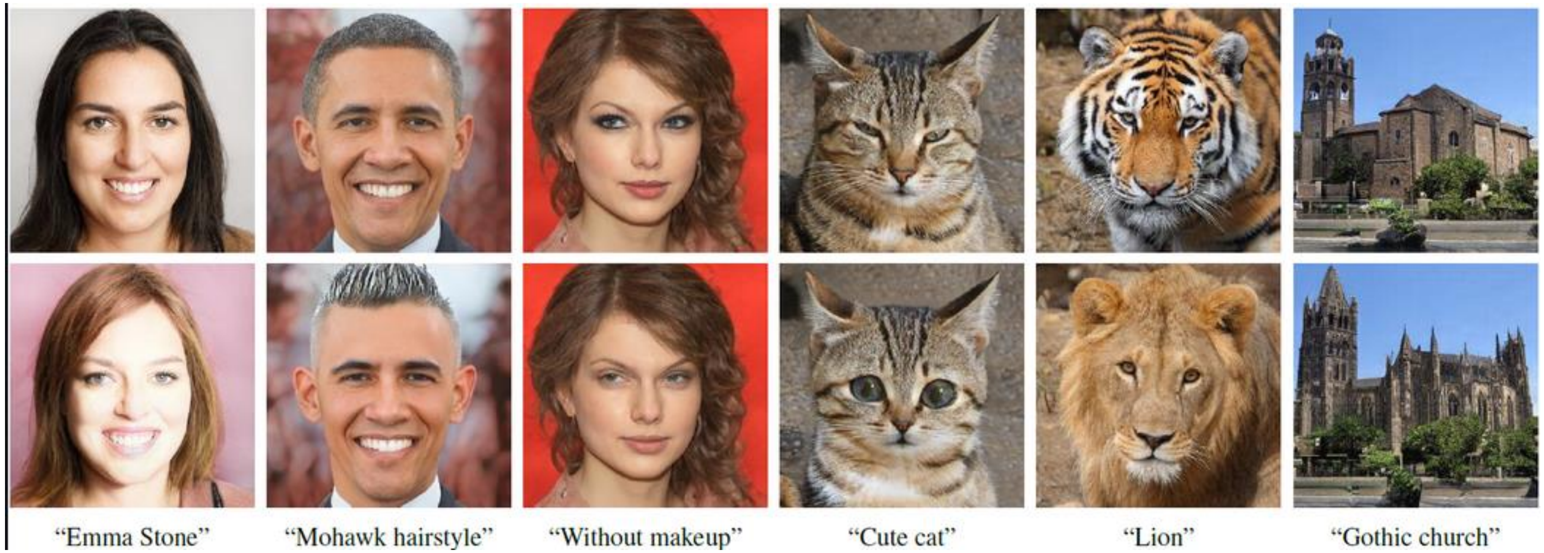


A Gentle Introduction to GAN's, Machine Learning Mastery

StyleCLIP

- StyleGAN + CLIP (Contrastive Language-Image Pre-Training)
- https://arxiv.org/pdf/1812.04948.pdf
- https://arxiv.org/pdf/2103.00020.pdf
- 31 Mart 2021

https://dar.vin/styleclip

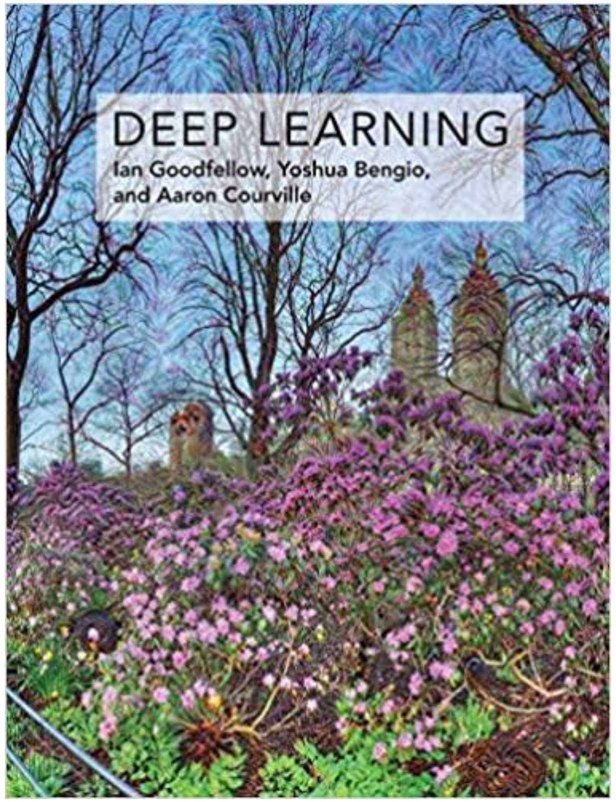


Derin öğrenme nasıl öğrenilir?

- Kalkülüs + lineer cebir + istatistik
- Python
- Makine öğrenmesi?
- https://globalaihub.com/deep-learning-expert-certificate/
- https://cs230.stanford.edu/syllabus/
- https://cs231n.github.io/
- http://web.stanford.edu/class/cs224n/

Deep Learning Book

https://www.deeplearningbook.org/



- Table of Contents
- Acknowledgements
- Notation
- 1 Introduction
- Part I: Applied Math and Machine Learning Basics
  - 2 Linear Algebra
  - 3 Probability and Information Theory
  - 4 Numerical Computation
  - 5 Machine Learning Basics
- Part II: Modern Practical Deep Networks
  - 6 Deep Feedforward Networks
  - 7 Regularization for Deep Learning
  - 8 Optimization for Training Deep Models
  - 9 Convolutional Networks
  - 10 Sequence Modeling: Recurrent and Recursive Nets
  - 11 Practical Methodology
  - 12 Applications
- Part III: Deep Learning Research
  - 13 Linear Factor Models
  - 14 Autoencoders
  - 15 Representation Learning
  - 16 Structured Probabilistic Models for Deep Learning
  - 17 Monte Carlo Methods
  - 18 Confronting the Partition Function
  - 19 Approximate Inference
  - 20 Deep Generative Models