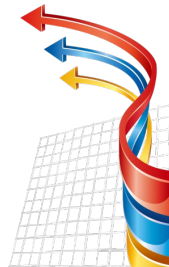
Data Driven Optimization for Digital Industries

Part 1: Deep Learning & Architectures

Rearrenged from Umberto Junior Mele material

Andrea Corsini
andrea.corsini@unimore.it

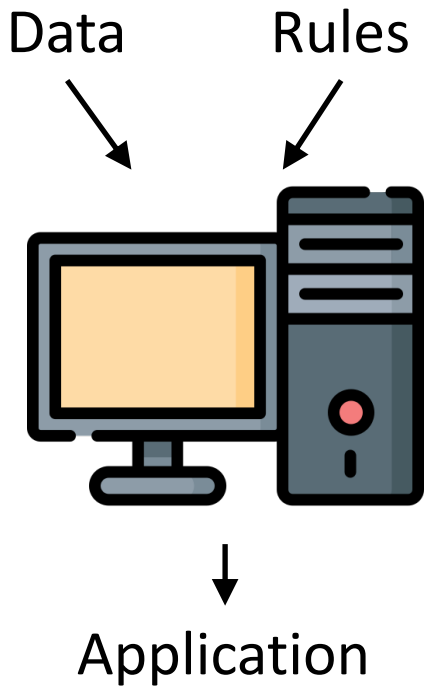


Exercises

You need to execute the experiments provided in the jupyter notebook file shown during class. Then, upload on Moodle a compressed folder with the file due to 19/5.

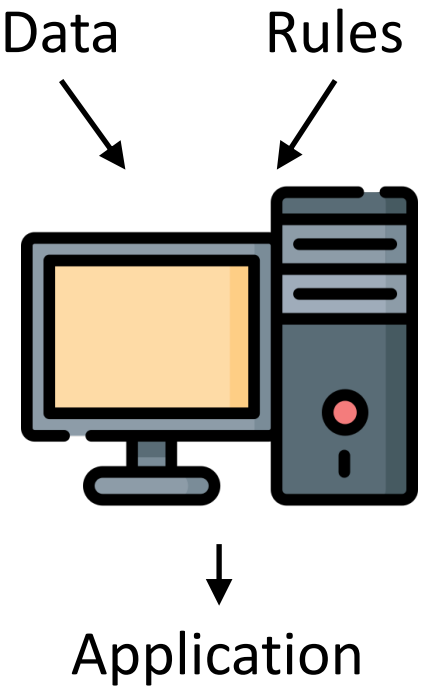
The Idea

Traditional Programming:

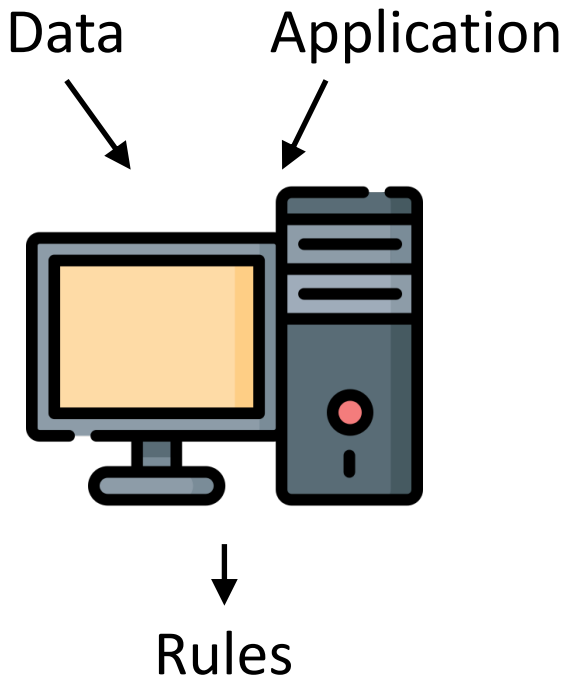


The Idea

Traditional Programming:



Machine Learning:



AI, Machine Learning, Deep Learning

AI

Computing systems which are capable of performing tasks that humans are very good at, for example recognising objects.

Machine Learning

The field of AI that applies statistical methods to enable computer systems to learn from the data towards an end goal.

Deep Learning

Neural Networks with several hidden layers.

Neural Networks

Are biologically inspired networks that extract abstract features from the data in a hierarchical fashion.

Narrow AI

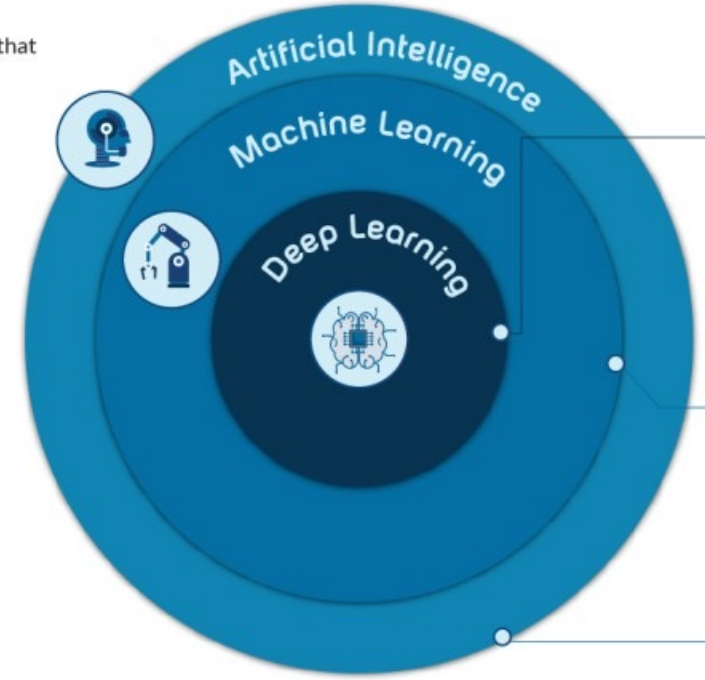
Designed to perform a single task. What we have today.

Artificial General Intelligence (AGI)

Can accomplish any intellectual task that a human can do. Remains an aspiration.

Artificial Super Intelligence (ASI)

A form of intelligence that exceeds the performance of humans in all domains.



Non-exhaustive list

Deep Learning

Recurrent Neural Networks, Convolutional Neural Networks, Deep Reinforcement Learning with Deep-Q Learning, Capsules & GANs.

Machine Learning

Support Vector Machine, Decision Trees, Gradient Boosting, Principal Component Analysis, Logistic Regression, Linear Regression, K-means Clustering.

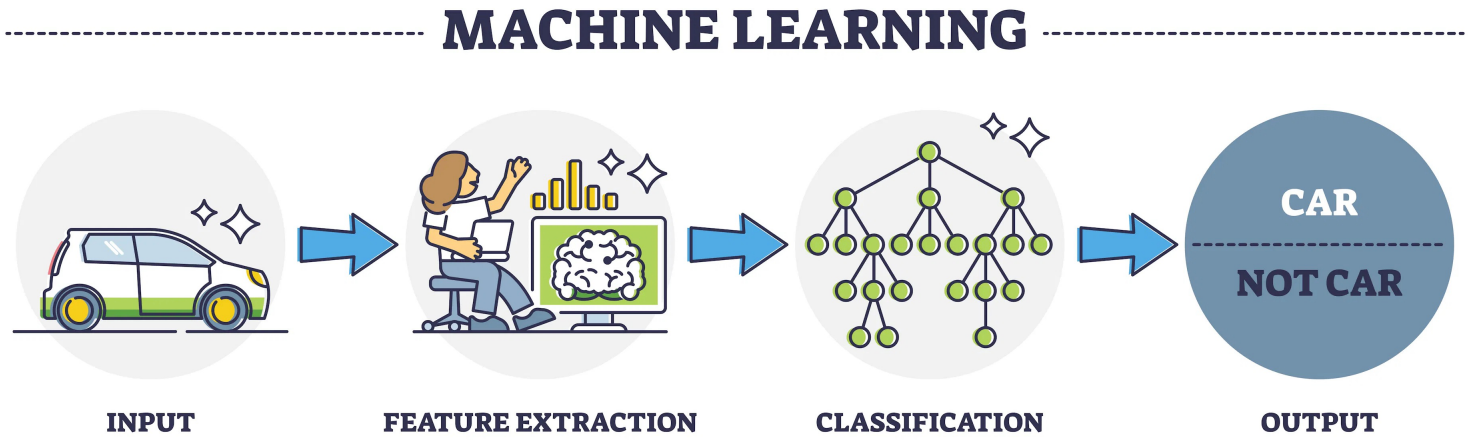
Classical Artificial Intelligence

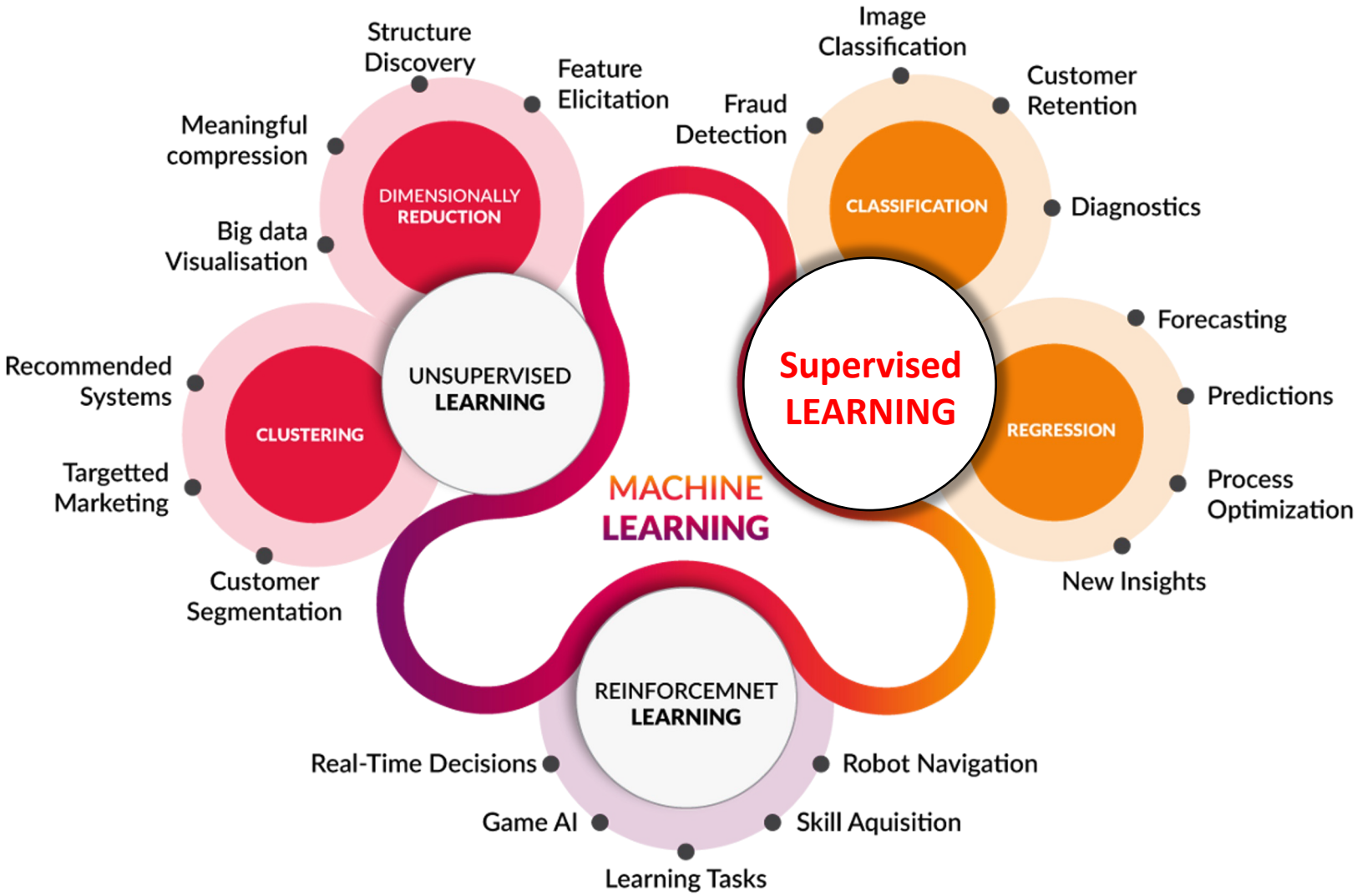
Rule Based Systems, Search Algorithms Depth First, Breadth First, A* algorithm, Propositional Calculus, Predicate Calculus Logic.

www.dls.ltd

Imtiaz Adam
@deeplearn007

From Machine to Deep Learning




MACHINE LEARNING
UNSUPERVISED LEARNING
DIMENSIONALLY REDUCTION
Structure Discovery
Feature Elicitation
Meaningful compression
Big data Visualisation
CLUSTERING
Recommended Systems
Targetted Marketing
Customer Segmentation
Supervised LEARNING
CLASSIFICATION
Image Classification
Customer Retention
Fraud Detection
Diagnostics
REGRESSION
Forecasting
Predictions
Process Optimization
New Insights
REINFORCEMNET LEARNING
Real-Time Decisions
Robot Navigation
Game AI
Skill Aquisition
Learning Tasks

Deep Learning Applications

- **Main assumption**: data used during training is similar in distribution to the data used during testing or after deployment.

- In Supervised Learning a labelled dataset is required.

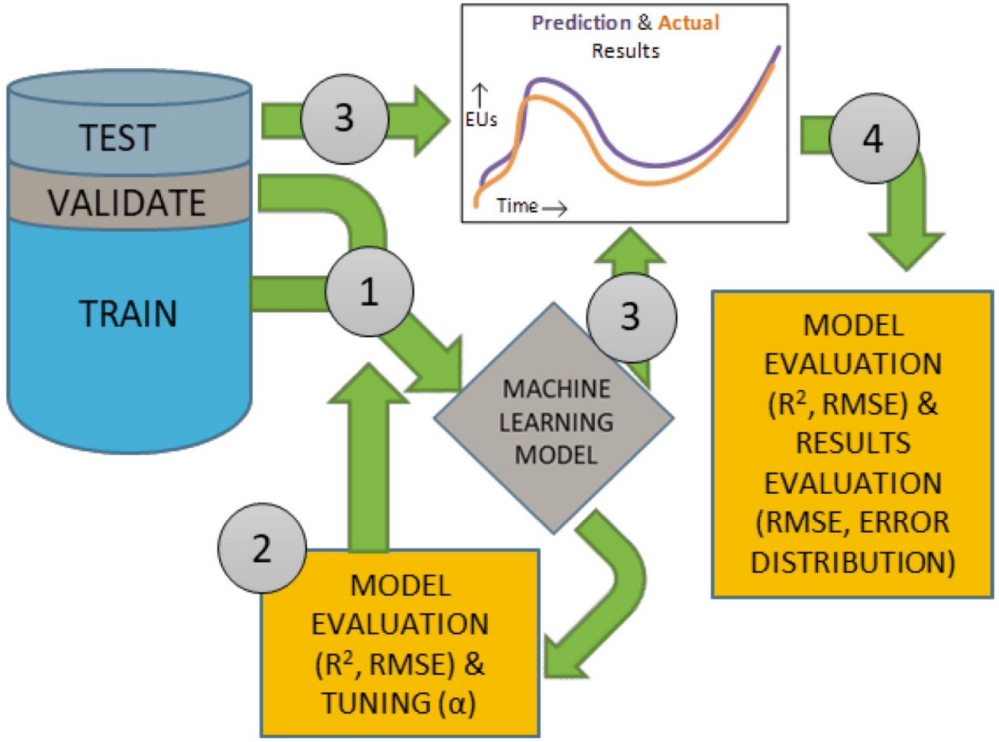
- In Unsupervised Learning only data (no labels) is required.

- In Reinforcement Learning a correct Markov Decision Process (actions, states, rewards) is required.

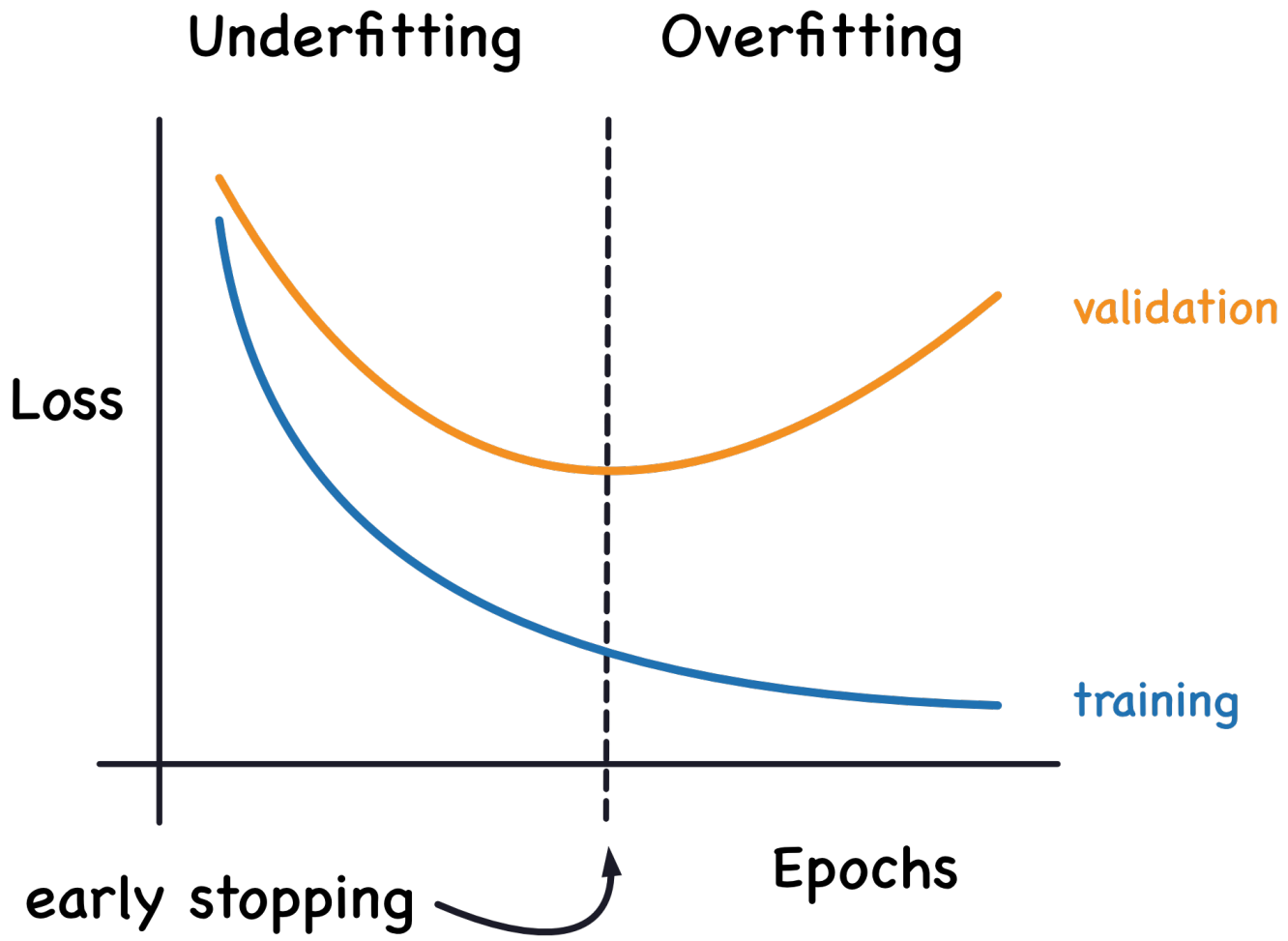
Supervised Ingredients

- Data separation is splitting the data into training, validation, and test sets:
  - Train data for learning a model (rules).
  - Validation data to test for generalization.
  - Test data for final evaluation.

- Overfitting occurs when a model learns the training data too well, including noise and irrelevant patterns.
  - Solution: periodically test model on validation data during training.

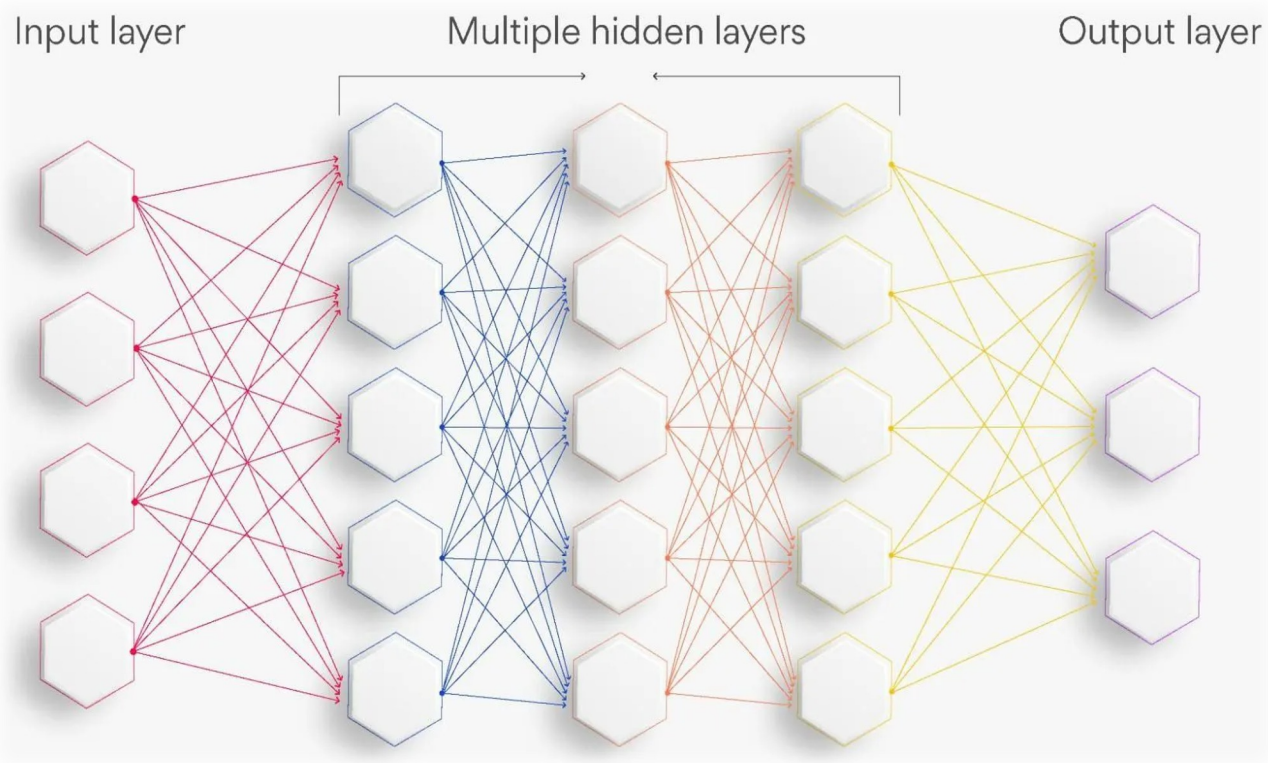


When to stop training



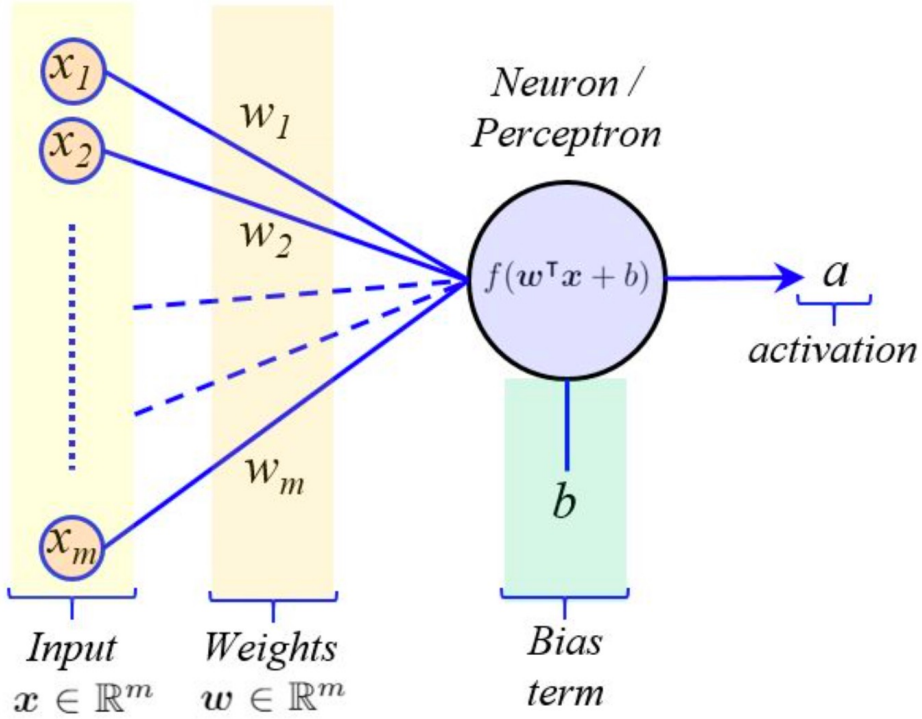
Vanilla Neural Networks

- Vanilla Neural Networks, also known as Feedforward Neural Networks, are the most basic type of neural networks.
- They consist of an input layer, one or more hidden layers, and an output layer.
- The neurons in each layer are connected to the neurons in the subsequent layer, forming a feedforward structure.



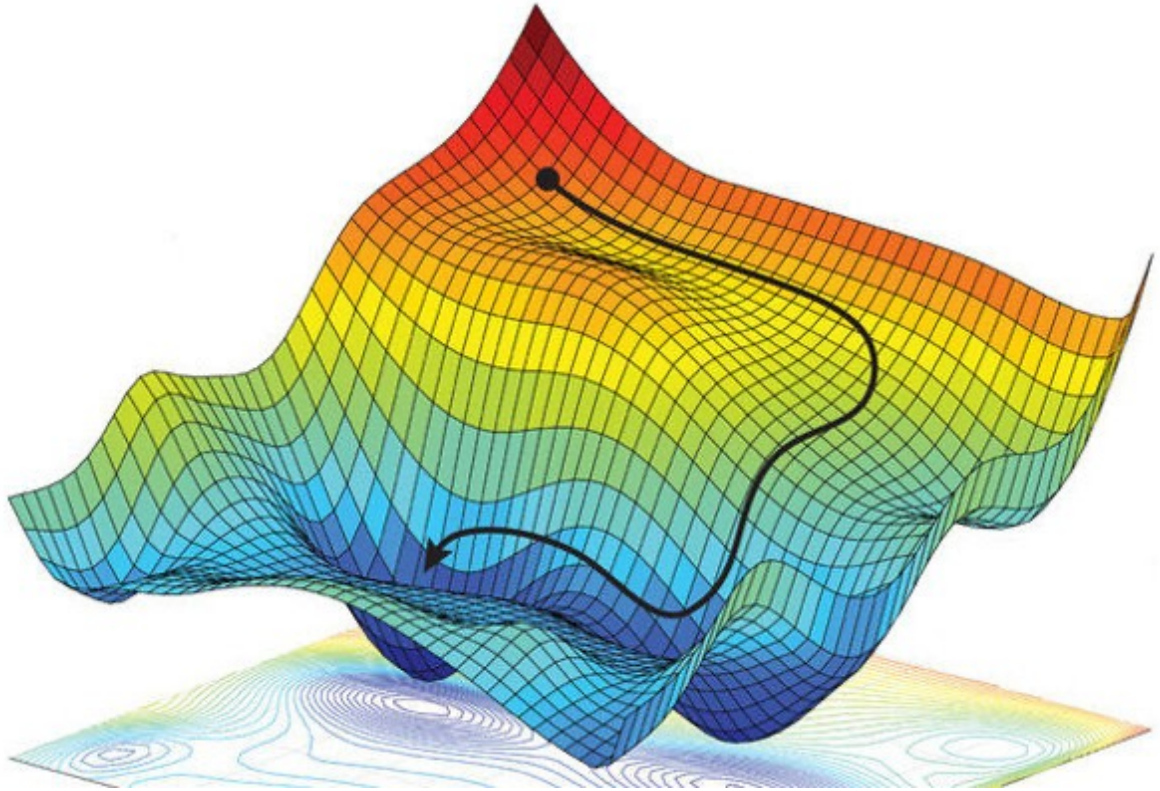
Vanilla Neural Networks

- A neuron receives inputs from previous neurons (or the input layer), multiplies them by weights, adds a bias, and applies an activation function.
- Activation functions introduce non-linearity, allowing the neural network to learn complex relationships in the data.
- Common activation functions include Rectified Linear Unit (ReLU), sigmoid, and hyperbolic tangent (tanh).



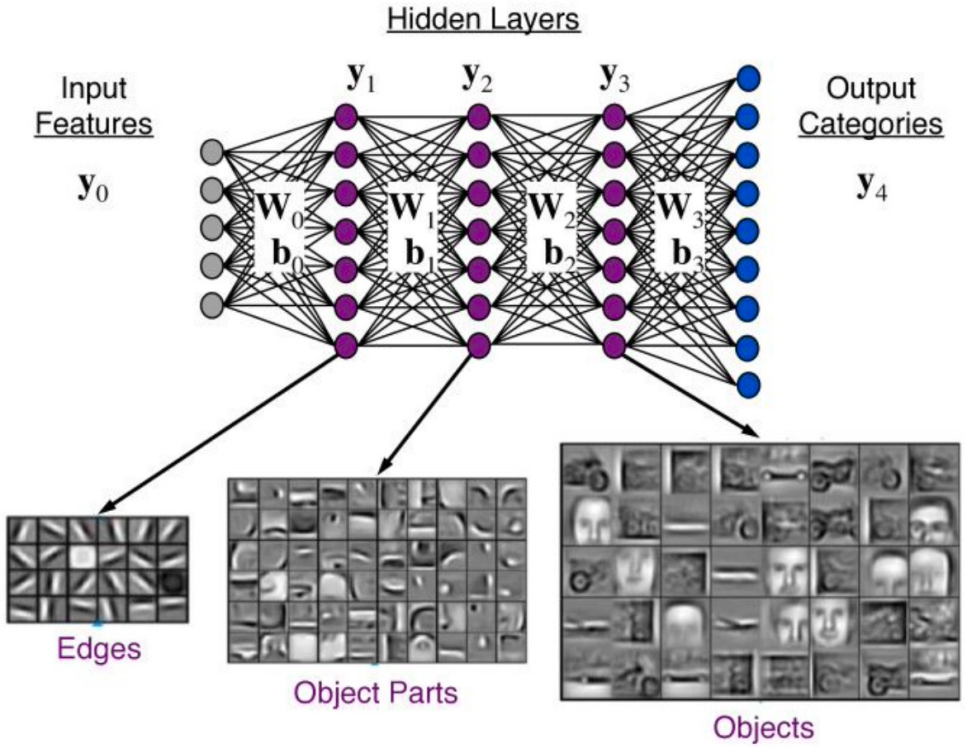
Vanilla Neural Networks

- The learning process is driven by minimizing the loss function, which measures the difference between the network's predictions and the true labels.
- Gradient descent is an optimization algorithm used to update the weights and biases.
- Backpropagation computes the gradient of the loss function with respect to each weight and bias by applying the chain rule.
- During each iteration, the weights and biases are updated based on the gradients.



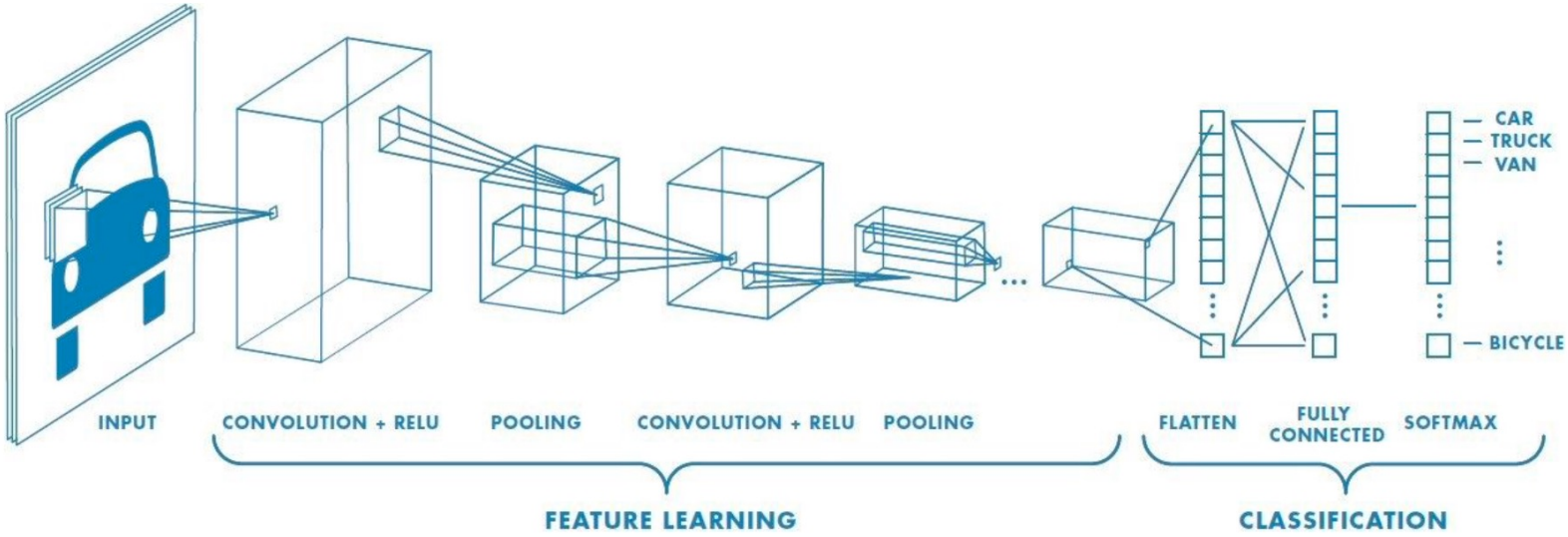
What's the advantage?

- Deep learning can learn hierarchical features, automatically discovering complex patterns in data.
- It is highly effective at handling **large and complex datasets**, such as images, audio, and text.
- Transfer learning allows for pre-trained models to be fine-tuned on a new task.
- Deep learning has a wide range of applications, including computer vision, natural language processing, speech recognition.



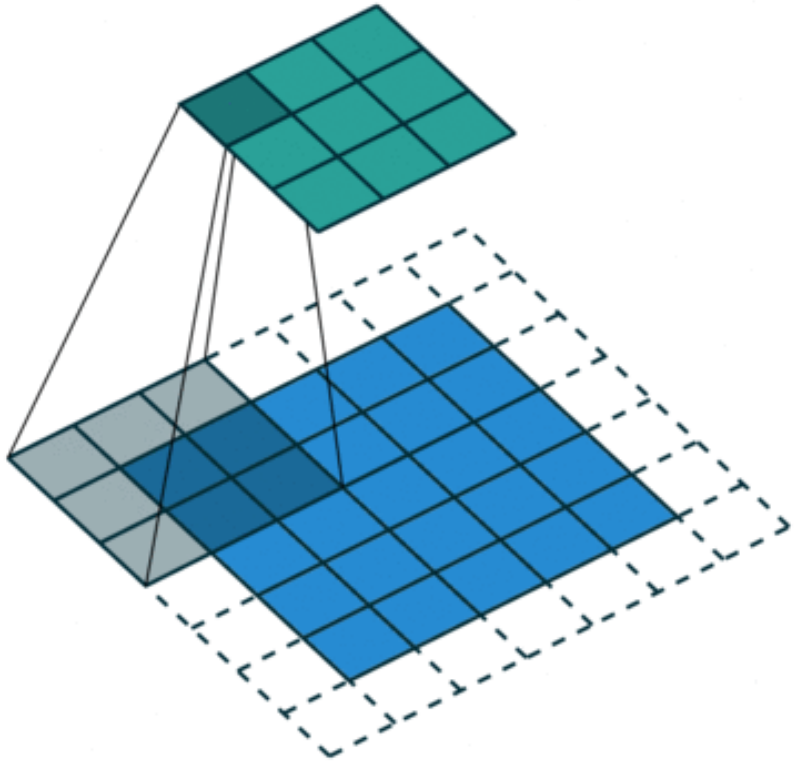
Convolutional Neural Networks

- CNNs are a specialized type of deep learning model designed to process grid-like data, such as images and videos, by automatically learning spatial hierarchies of features.
- CNNs consist of components such as convolutional layers, pooling layers, and feed forward layers, which allow the model to learn patterns in features of input data.
- Widely used in applications like image classification, object detection, facial recognition, and semantic segmentation.



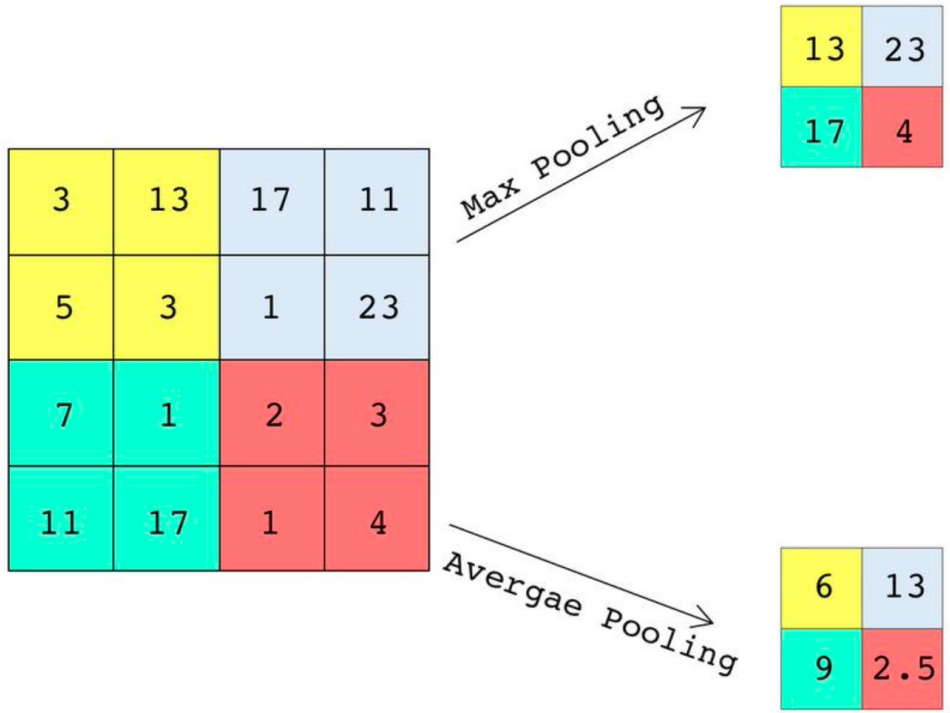
CNNs: Convolutional Layer

- Convolutional layers in CNNs are responsible for learning and detecting local patterns and features in the input data
- They use a a filter that slides over the input data, performing a convolution operation.
- Stride refers to the step taken by the kernel as it moves across the input. A stride of 1 means the kernel moves one step at a time.



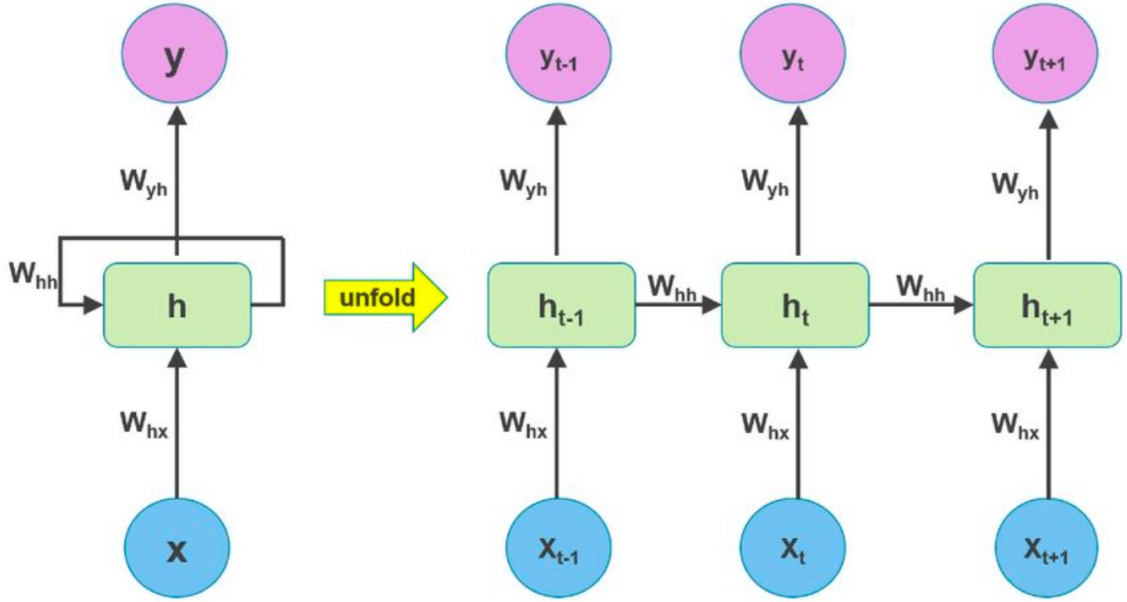
CNNs: Pooling Layer

- Pooling layers in CNNs are used to reduce the spatial dimensions of the feature maps.
- The most common types of pooling are max pooling and average pooling.
- Pooling layers provide translation invariance, meaning that the network becomes more robust to small translations in the input data.



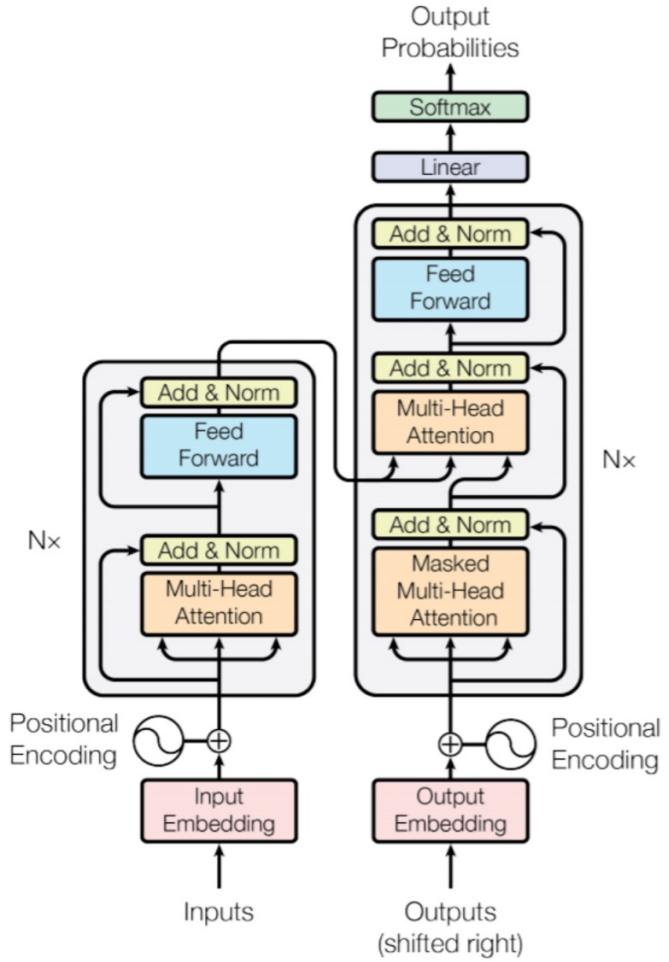
Recurrent Neural Network

- Recurrent Neural Networks (RNNs) are neural network used to process sequential data.
- RNNs have a built-in memory mechanism that allows them to maintain information.
- Common use cases include natural language processing, time series prediction, speech recognition, and music generation.



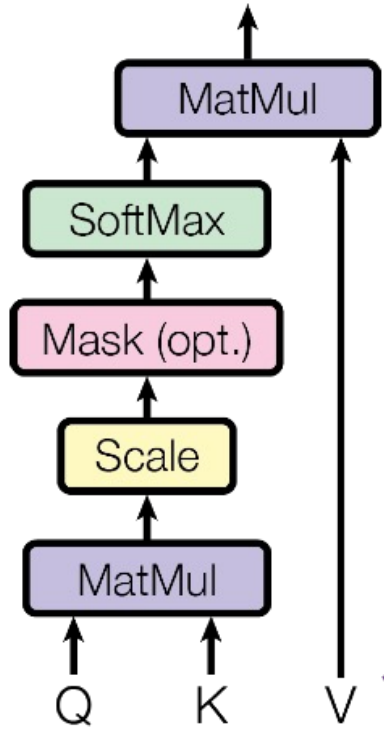
Transformer

- Transformers overcome limitations of RNNs by enabling parallel processing of sequences, which leads to faster training and more efficient handling of large datasets.
- Transformers use a self-attention mechanism to model dependencies within a sequence, allowing them to capture long-range dependencies more effectively than RNNs.
- To incorporate sequence order information, transformers use positional encoding, which allows them to understand the relative positions of elements in the sequence.
- By addressing these limitations, transformers have become the standard for many sequence-based tasks.



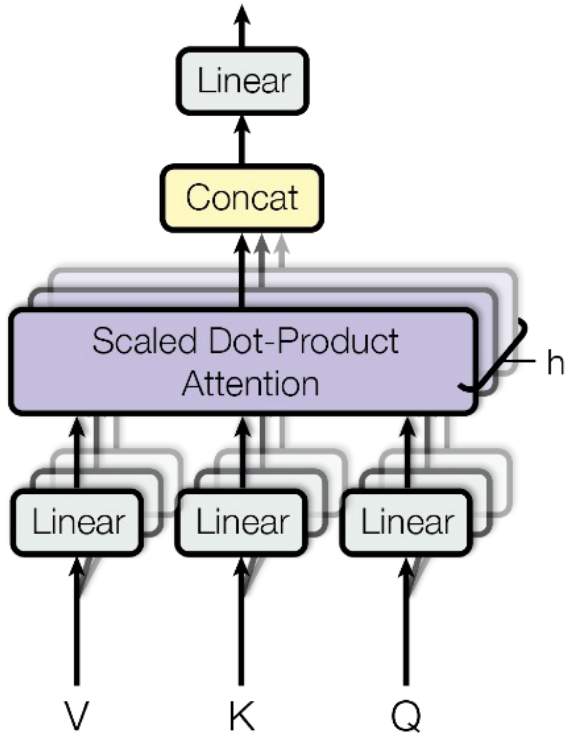
Transformers: Self-Attention

- For each element in the sequence, self-attention computes a weighted sum of all elements, where the weights are determined by the similarity between query (Q), key (K), and value vectors (V).
- Scaled dot-product attention is the mechanism used to calculate these weights, which involves taking the dot product of the query and key vectors, dividing by the square root of the key's dimension, and applying a softmax function.



Transformers: Multi-Head Attention

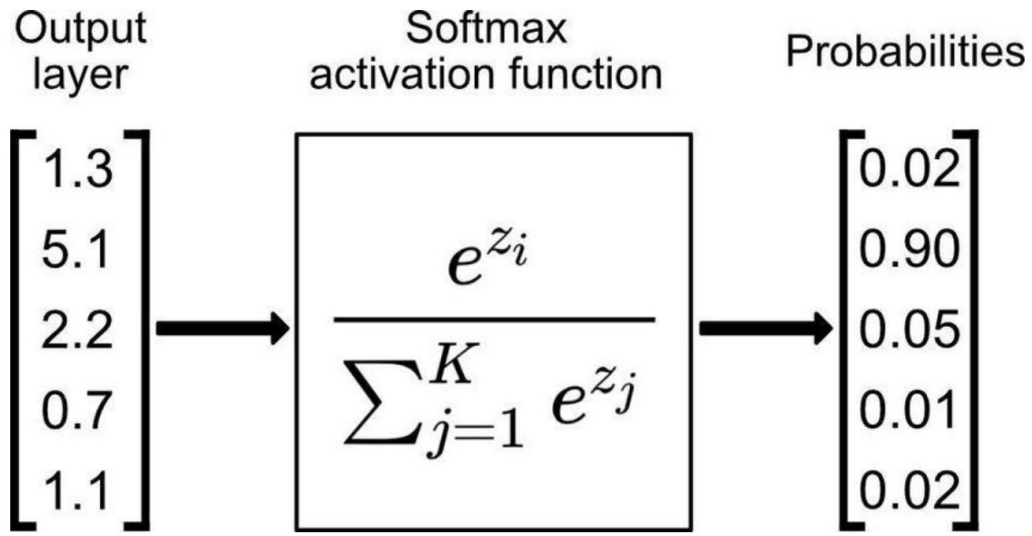
- Multi-head attention is an extension of the self-attention mechanism that allows the model to focus on different aspects of the input data simultaneously.
- The self-attention mechanism is applied multiple times in parallel, and the resulting outputs are concatenated and transformed to create the final output.
- By using multiple attention heads, the model can capture a richer representation of the input data, improving its ability to learn complex patterns and relationships.



Softmax Output

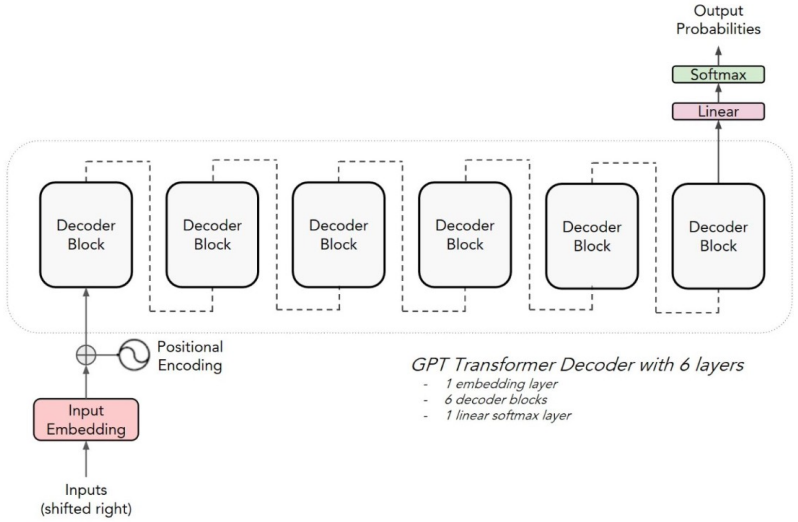
- The softmax function is applied to the output of Transformers (and other neural networks) to generate probabilities for each class in the target vocabulary, enabling the model to produce predictions for tasks like language translation or text generation.

- During training, the cross-entropy loss is used to measure the difference between the predicted probabilities and the true labels, providing a signal for optimizing the model's parameters.



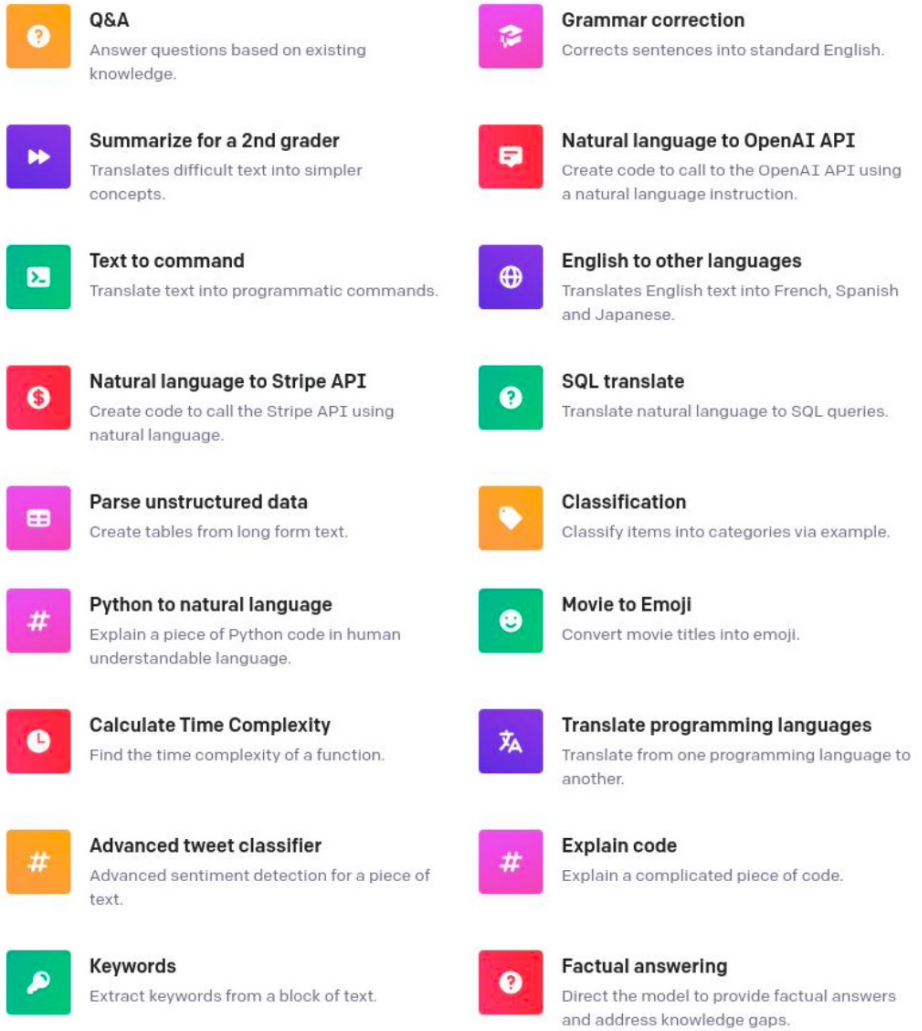
Generative Pre-trained Transformers

- GPT, or Generative Pre-trained Transformers, are a family of state-of-the-art language models based on the Transformer architecture.
- GPT models are pre-trained on large-scale unsupervised data, allowing them to generate highly coherent and contextually relevant text.
- GPT models utilize the decoder of the transformer. This design allows GPT models to generate text autoregressively, predicting one token at a time based on the previously generated tokens.



Generative Pre-trained Transformers

- GPT models have a wide range of applications due to their powerful natural language understanding.
- Common use cases include text summarization, machine translation, sentiment analysis, question-answering, and more.
- Additionally, GPT models have been successfully deployed in conversational AI systems, such as chatbots and virtual assistants, to provide highly coherent and contextually appropriate responses.



Future Trends

The field of language modelling and GPT models is rapidly evolving, with several future trends expected to shape the development of these models:

- **Efficient architectures and compression**: to address concerns about energy consumption and accessibility, research will focus on developing efficient architectures and compression techniques that maintain performance while reducing resource requirements.

- **Collaborative and federated learning**: new training paradigms, such as collaborative learning and federated learning, may enable more decentralized and privacy-preserving training of GPT models.

- **Bias mitigation and fairness-aware**: as ethical concerns become more prevalent, research will increasingly focus on developing models that are aware of and can mitigate biases, as well as promote fairness in their applications.

Lab

[Link to the jupyter file](#).

[Link to the GitHub repo](#).