LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN
FAKULTÄT FÜR MATHEMATIK, INFORMATIK UND STATISTIK
INSTITUT FÜR INFORMATIK
LEHRSTUHL FÜR DATENBANKSYSTEME UND DATA MINING

Lecture Notes on

Deep Learning and Artificial Intelligence

Winter Semester 2024 /2025

Course Introduction

Lectures: Prof. Dr. Matthias Schubert

Tutorials: Maximilian Bernhard

Outline

- Neural Networks and their Origins
- Reinforcement Learning and Planning under Uncertainty
- Setting the stage for Deep Learning
- What can you do with Deep Learning and Deep Reinforcement Learning?
- Course Overview and Organisation

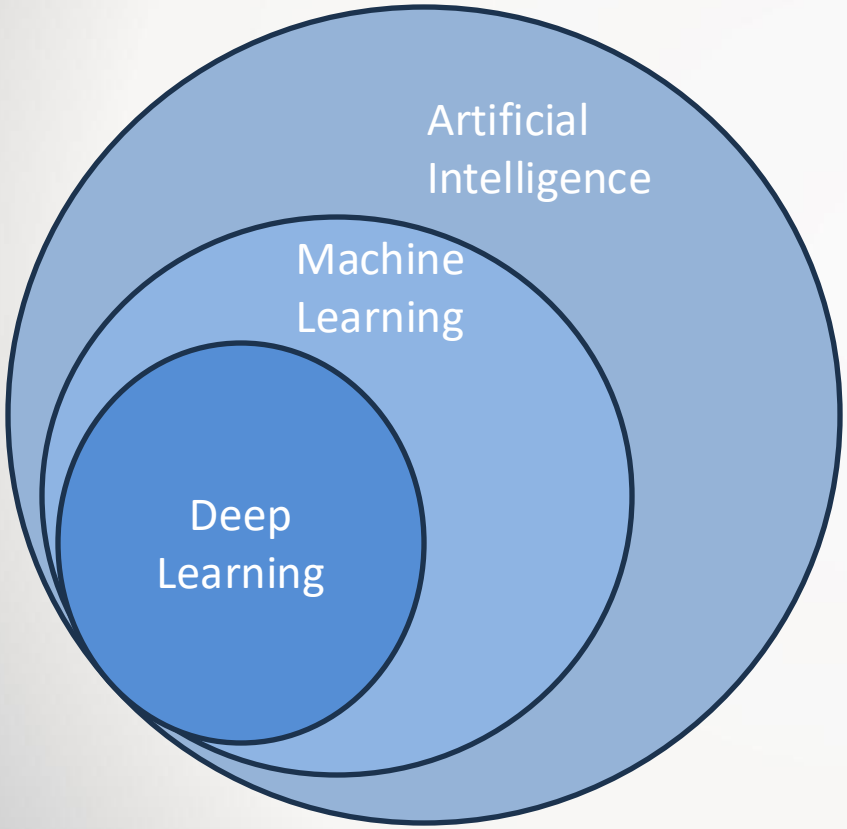
What are neural Networks ?

- **Input domain**: $x \in X \subseteq \mathbb{R}^{d_1 \times .. \times d_l}$
  (l=1 for vectors, l=3 for images, i.e., width, height, RGB channels)
- **Output domain**: y $\in y \subseteq \mathbb{R}^{d_1 \times .. \times d_l}$
  (class probability, continuous prediction targets, images, text tokens, ..)
- A neural network is a **parametric function** $f_\theta : X \to Y$
  where $\theta \in \Theta \subseteq \mathbb{R}^{d_1 \times .. \times d_l}$ is called weights/parameters

Examples:

- regression and classification functions
- generating text, images, and audio from context information
- generating annotations (markings) in images like boxes on objects or pixel masks
- generating sentences in a target language from a source language

Machine Learning and Deep Learning



- AI is a broad field involving making machines act and think rationally or human-likely.
- Machine Learning allows machines to ”learn” with data without being explicitly programmed. Often based on statistical modeling.
- Deep Learning is an area of Machine Learning that generalizes Neural Networks to computational graphs of differentiable functions.

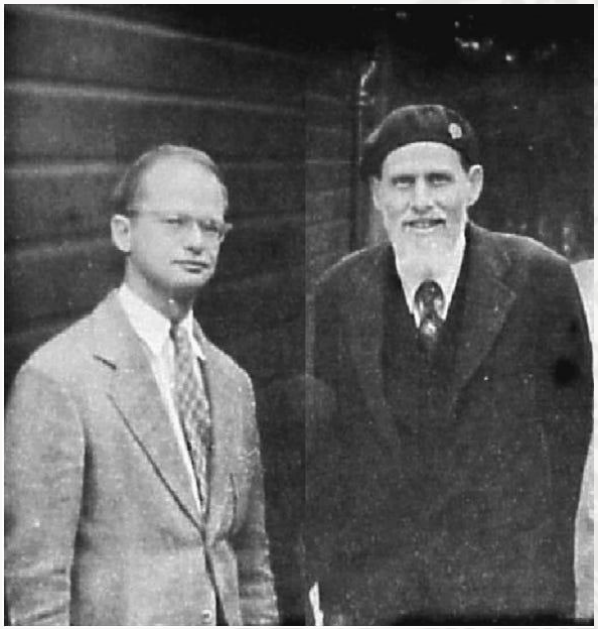
Neural Network Origins

Waren McCullouch and Walter Pitt: *A logical calculus of ideas immanent to nervous activity,* Bulletin of Mathematical Biology, 1943

$$\textit{model: } \hat{y} = 1\left(\sum_{j=1}^{P} x_j \geq \theta\right)$$

$$\text{with } x_j \in \{0,1\}\ \forall j, \theta \in \mathbb{R}$$

- Can model a variety of logical operations including AND, OR, NOR by setting θ
- Limitations:
  - all inputs are binary
  - all inputs have equal contribution
  - cannot represent XOR
  - cannot learn rules



McCulloch (right) and Pitts (left) in 1949

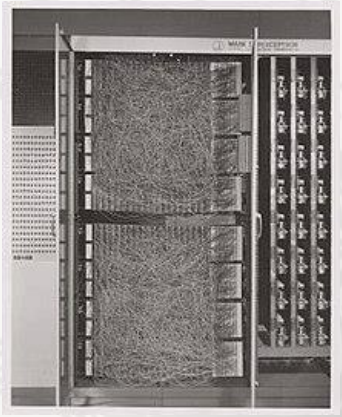
Rosenblatt's Perceptron

- in 1950 Frank Rosenblatt proposed the perceptron

Model: $\hat{y} = 1(\beta^T x + \beta_0 \geq 0)$
with $x \in \{0,1\}, \beta \in \mathbb{R}^p, \beta_0 \in \mathbb{R}$

Frank Rosenblatt

- realized in HW as an actual machine



The Mark I Perceptron

Rosenblatt's Perceptron

- The learning algorithm:

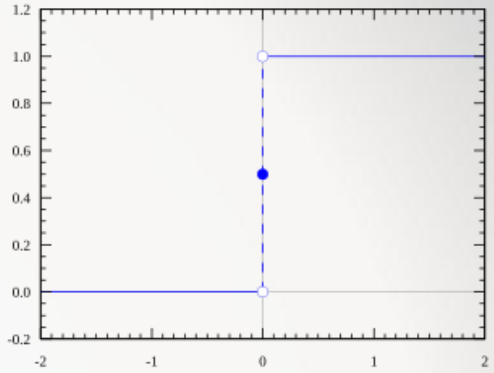
$$\beta := \beta + \alpha \cdot (y_i - \hat{y}) \cdot x$$
$$\beta_0 := \beta_0 + \alpha \cdot (y_i - \hat{y})$$

- Similarity to logistic regression:

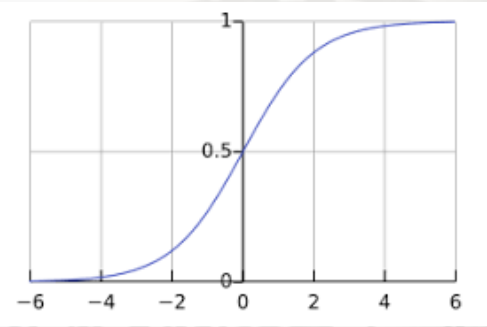
$$\frac{\partial}{\partial \beta} ylog(\pi) + (1 - \mathrm{y})\log(1 - \pi) = (y - \pi)x,$$

where $\pi = \frac{1}{1+e^{-(\beta^T x)}}$

  - indicator function similar to logistic function
  - same limitation: only linear separable classes



indicator function



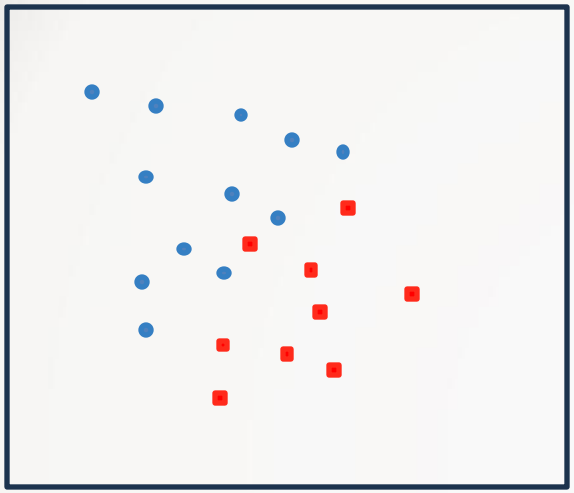
logistic function

- Marving Minsky&Seymour A. Papert heavily critiizied Perceptrons in:

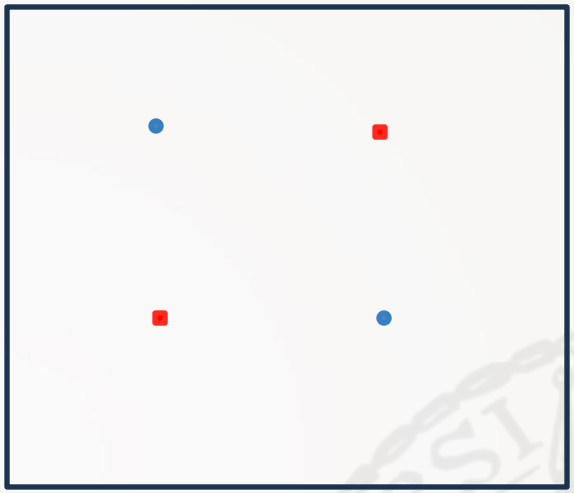
  Perceptrons: An Introduction to Computational Geometry

- This was followed by the AI winter starting around the 1960

Limitations of linear functions



not linear separable



XOR problem

- The indicator function is too inflexible for non-deterministic patterns
  - using the logistic function addresses this issue
- Even logical rules cannot be expressed by a single perceptron
  - stacking perceptrons allows solutions
  - applying multiple functions after each other corresponds to mapping the data into a different representation space (which might be again linear separable)

Multi-Layer Perceptron

- Werbos 1982, one of the first papers applying the chain rule to train NNs with multiple layers :

  Applications of advances in nonlinear sensitivity analysis

- Rumelhart et al. 1986, showing that the BP encourage hidden units to encode relevant information in the input :

  Learning representations by back-propagating errors

- LeCun et al. 1989, proposed to apply BP to train two layers of convolutional network :

  Backpropagation Applied to Handwritten Zip Code Recognition

- MLP outperformed by the extended Support Vector Machines (Cortes and Vapnik, 1993/1995) :

  Support-Vector Networks

  that introduces soft margins and the kernel trick.

⇒ Neural Network Winter 1995-2007 in Machine Learning Research

David Rummelhart

Vladimir Vapnik

Decision Processes

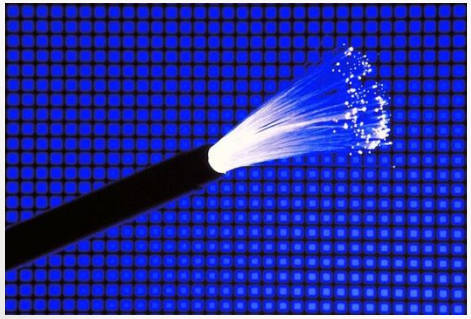
- **deterministic *one-step*** decisions:
  - evaluate all known options and select the best one
- **deterministic *multi-step*** decision-making:
  - evaluate all known sequences of options and select the best one
  - as there are too many options, try to avoid evaluating all sequences
  - build a search and prune leaves if they cannot be the best
- **non-deterministic** *one-step* decision: (multi-armed bandits)
  - search for the option where the expected outcome is best
- **non-deterministic** *multi-step* decisions: (Markov decision Processes)
  - option in later steps might depend on the options taking in earlier ones (solution is not a sequence of options)
  - generally, we need to choose what to do for any situation we might get into (solution is a policy telling us what to do in any possible situation)
  - Maximizing the expectation is still the most common goal

Reinforcement Learning

- Markov Decision Processes (MDPs) allow for evaluating and finding optimal policies if the complete decision problem is fully specified:
  - distribution of what happens when making a decision *a* in the situation(state) *s*
  - distribution of the benefit of each situation
- Problems with MDPs:
  - the complete set of distribution is usually unknown
  - realistic MDPs involve vast amounts of situations, and the transition distribution between them increases quadratically.
    ⇒ finding the optimal policies becomes infeasible
- Reinforcement Learning:
  - learns from interacting with an environment that might follow an unknown MDP
  - learns the relevant transitions between situations depending on actions
  - tries to learn optimal policies
  - does not receive supervision in the form of action was right or wrong
  - supervision by providing positive signals if the current situation is good

Setting the Stage for Modern AI

- Internet and broadband connections: allowed to publish information easily, access information from a huge amount of sources
- Data Storage: hard drives became larger and cheaper. SSDs make background storage faster. Larger/faster main memory
- Mobile devices: collect personal and spatial data



http://www.ubergizmo.com/2013/01/china-policy-demands-new-residences-have-fiber-optic-connections/

http://blog.rentacomputer.com/2012/09/18/dont-ever-lose-your-data-again-with-a-storage-server-rental/

Setting the Stage for Modern AI

- Cloud computing: distributed computations on thousands of commodity machines
- Commodity GPUs: dedicated numerical processing power
  Cheaper sensors/camera: affordable monitoring
- IoT and sensors: monitoring installations and environments
- RC and autonomous mobile units: UAVs, rovers,..

Impact of Big Data and Modern Hardware to AI

Impacts on Machine Learning and AI:

- more data: complex problems become feasible:
  - **before**: available samples only allowed simple models
  - **now**: complex models can be trained because of huge amounts of samples
- more computational power:
  - **before**: complex models did not finish training
  - **now**: models with several thousand parameters on millions of samples
- scalability:
  - **before**: predictors for dedicated cases
  - **now**: personalized models for millions of cases

Artificial Intelligence and Data Analytics

- AI is an extremely broad subject within CS
- ***tasks***: reasoning, problem-solving, knowledge representation, planning, learning, natural language processing, perception, motion and manipulation, social intelligence, creativity, general intelligence
  ⇒ some major overlaps with machine learning/data analytics
- today, the public most often refers to deep neural networks
- scientists often refer to any of the tasks named above

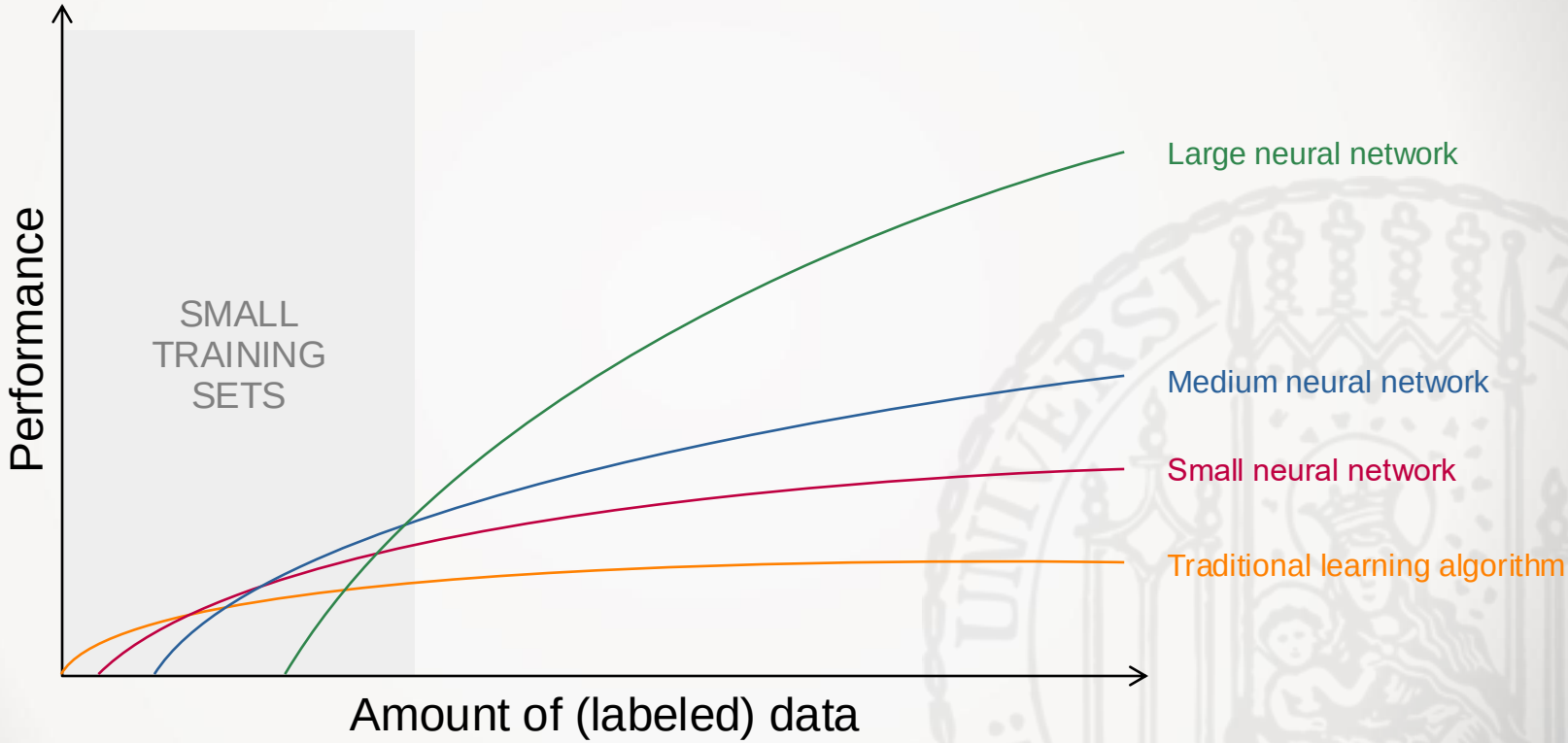
Deep Learning

What can you do with Big Data?

- more data allows for training larger models (up to billions of parameters)
- most breakthrough applications in AI are connected to Deep Learning:
  - NLP: automatic translation, text generation, query-answer systems, automatic image captioning, …
  - Imaging: face detection, understanding traffic scenes (autonomous driving), generating images, recolouring black-white images, etc.
  - Autonomous control: Alpha Go (zero), Alpha Star, OpenAI5, controlling robots (drones, cars, biped, quadruped..), optimising energy consumption,...
  - others: image-to-text transfer, generative models, deep fakes…
- Deep Learning is not about novel prediction or pattern-mining algorithms. It is about transforming data into more advantageous representations

  ⇒ ***representation learning***

Impact of more data on Network Size

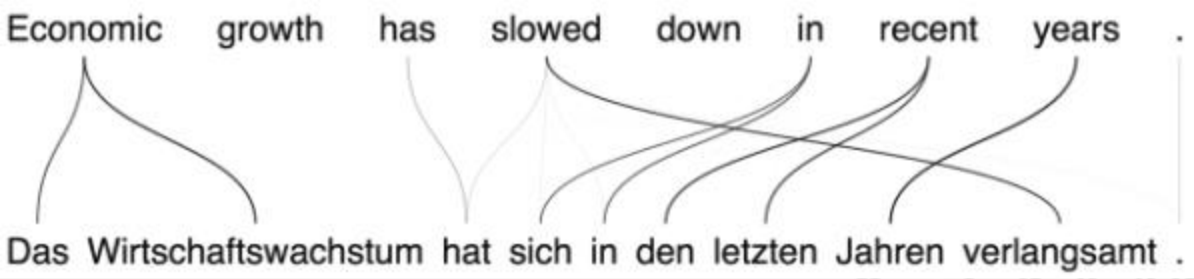


Applications in Natural Language Processing

- Text Classification: categorise documents topics
- Language Modelling: predict the next word in a sentence
- Speech Recognition: audio to text
- Caption Generation for images and video
- Machine Translation
- Document Summarization: provides a concise text from one large or several smaller texts like reviews
- Question Answering: interactive conversations, e.g., for chatbots and online support systems

Example: Automatic translation

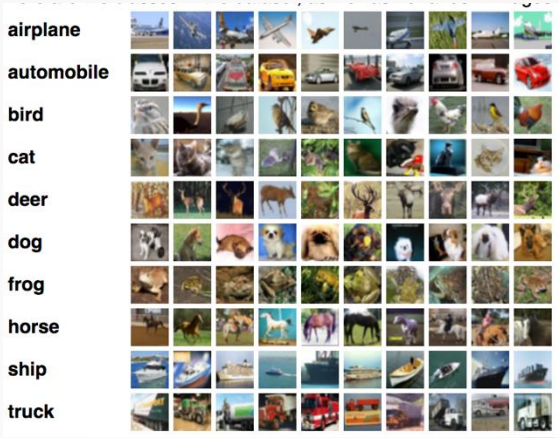
Economic growth has slowed down in recent years. → ? → Das Wirtschaftswachstum hat sich in den letzten Jahren verlangsamt.

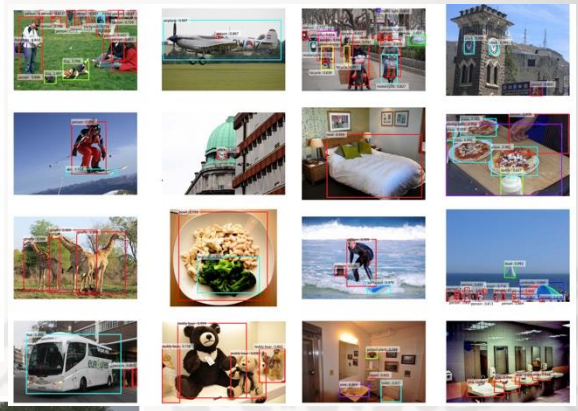


Applications in Computer Vision (1)

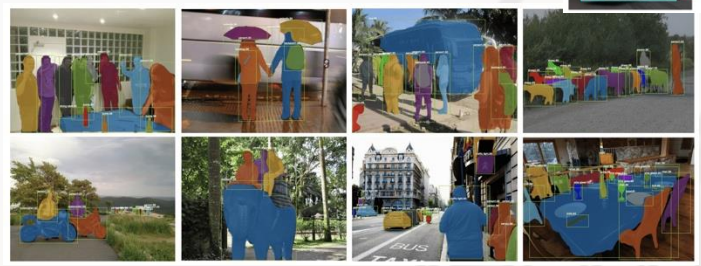
examples from https://machinelearningmastery.com/applications-of-deep-learning-for-computer-vision/

- Image Classification





- Object Detection



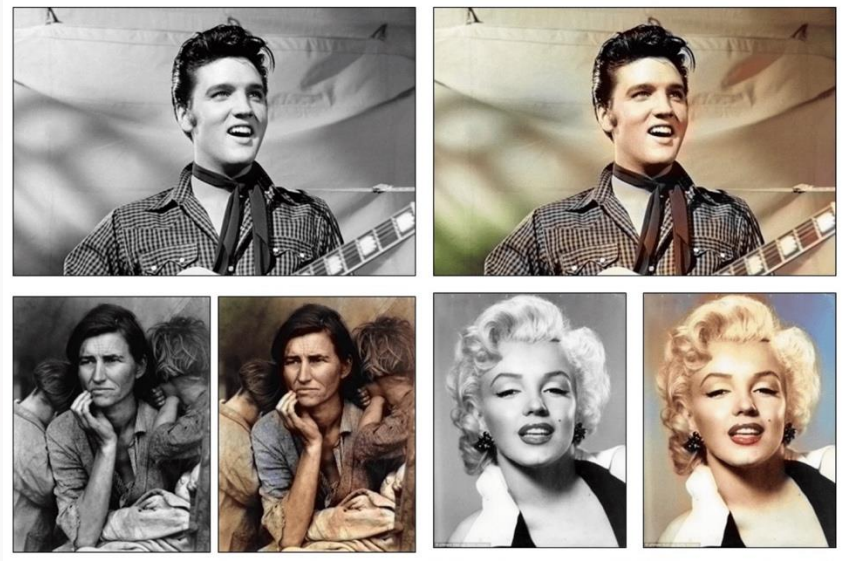
- Object Segmentation

- Image Style Transfer

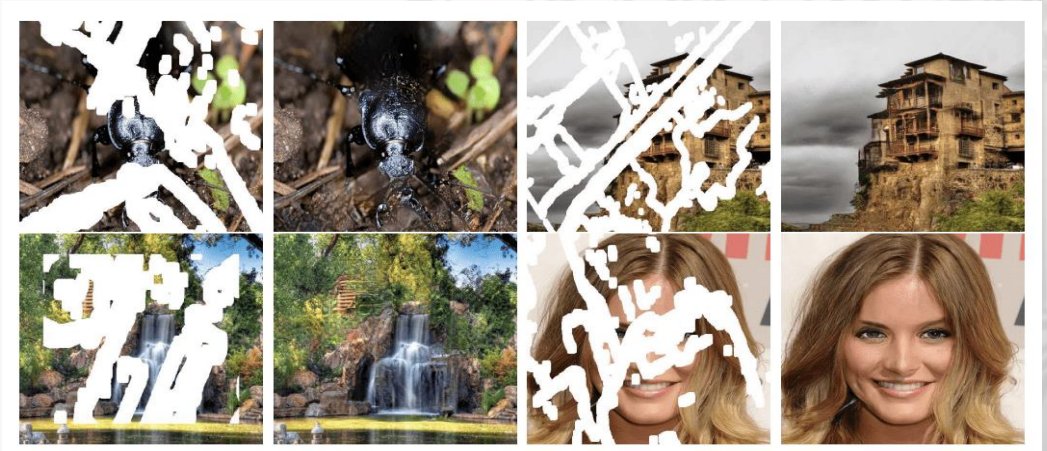
Applications in Computer Vision (2)

examples from https://machinelearningmastery.com/applications-of-deep-learning-for-computer-vision/

- Image Colorization



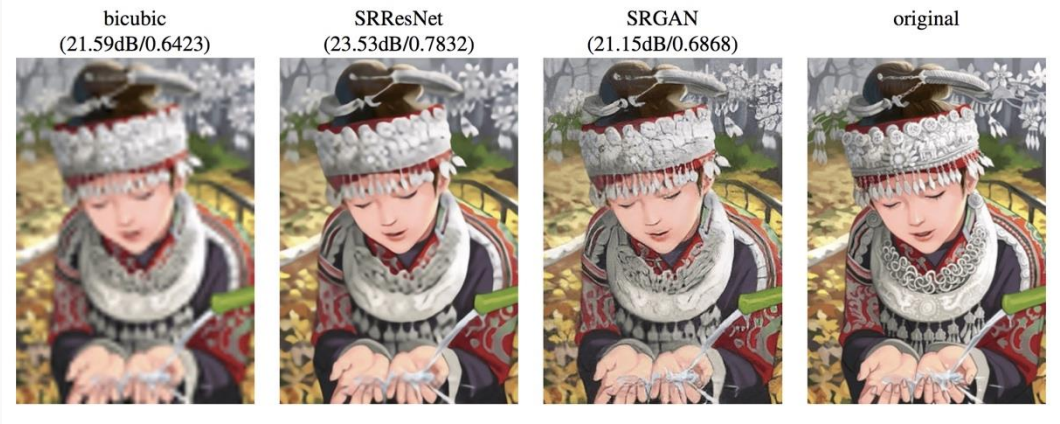
- Image Reconstruction



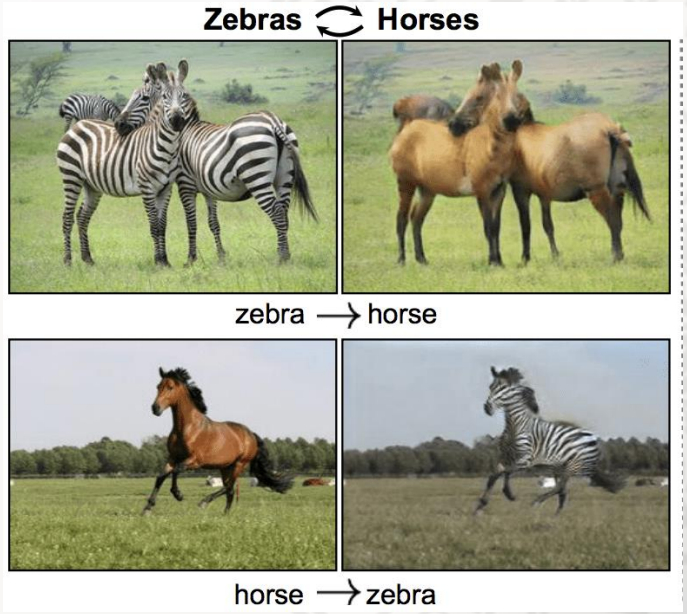
Applications in Computer Vision (3)

examples from https://machinelearningmastery.com/applications-of-deep-learning-for-computer-vision/

- Image Super-Resolution



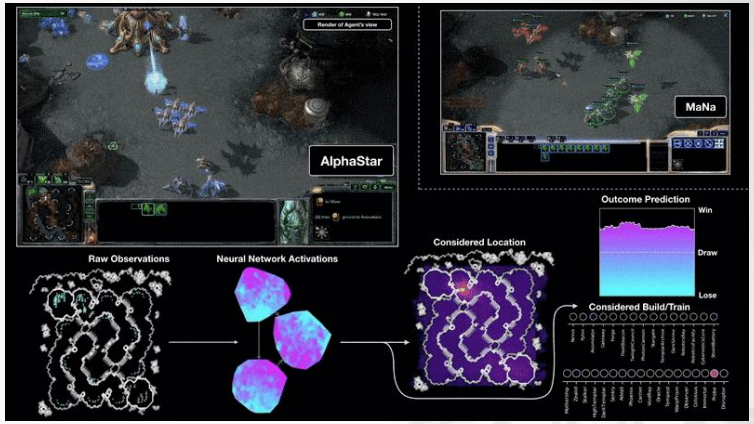
- Image Synthesis



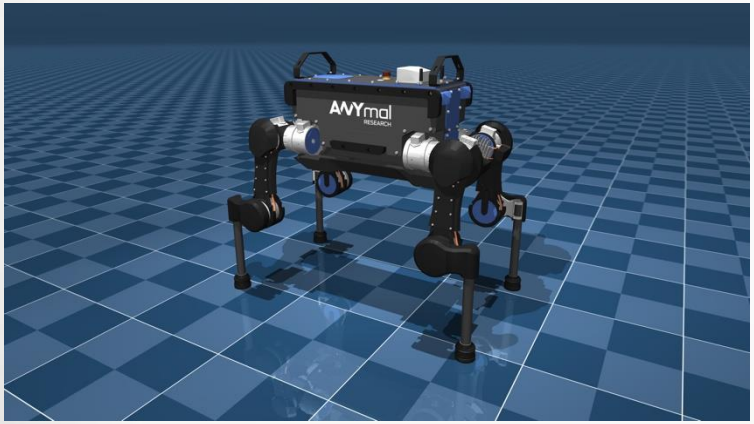
Applications in Controlling Agents

- computer games
- controlling robots
- autonomous driving
- automatic trading

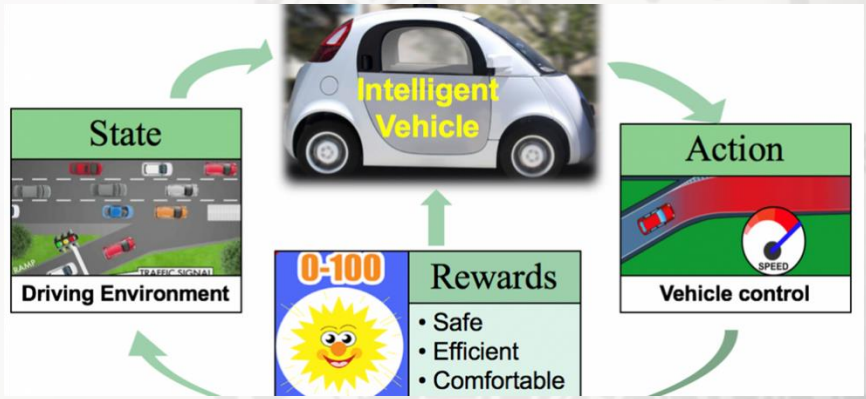
1.



2.



3.



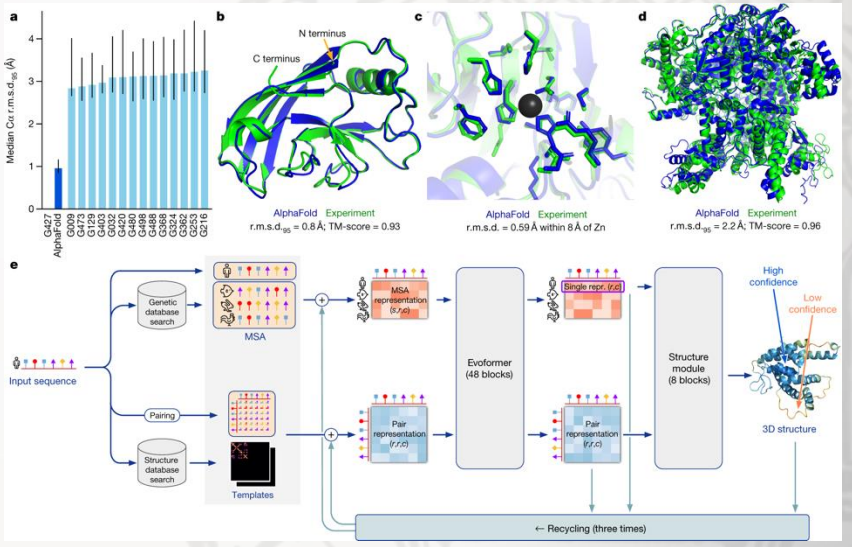
1. https://www.deepmind.com/blog/alphastar-mastering-the-real-time-strategy-game-starcraft-ii
2. https://github.com/deepmind/mujoco_menagerie/blob/main/anybotics_anymal_b/anymal_b.png?raw=true
3. https://deepdrive.berkeley.edu/project/maneuver-control-based-reinforcement-learning-automated-vehicles-interactive-environment

other cool stuff

- Create art
- Protein Structure Prediction (Alpha Fold)
- approximate solutions for large discrete optimisation problems
- learn physical laws
- and more

1.

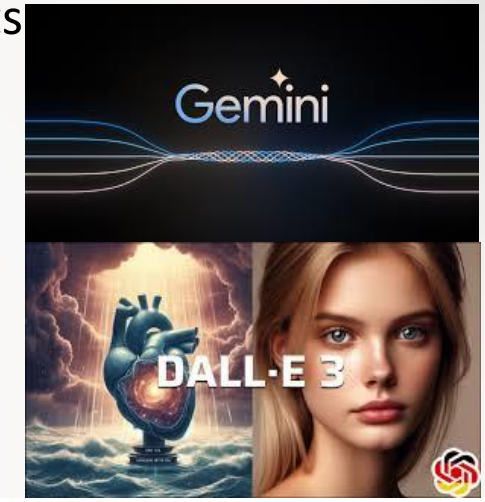
2.



1. https://stablediffusionweb.com/
2. https://www.nature.com/articles/s41586-021-03819-2

Artificial Intelligence and Energy

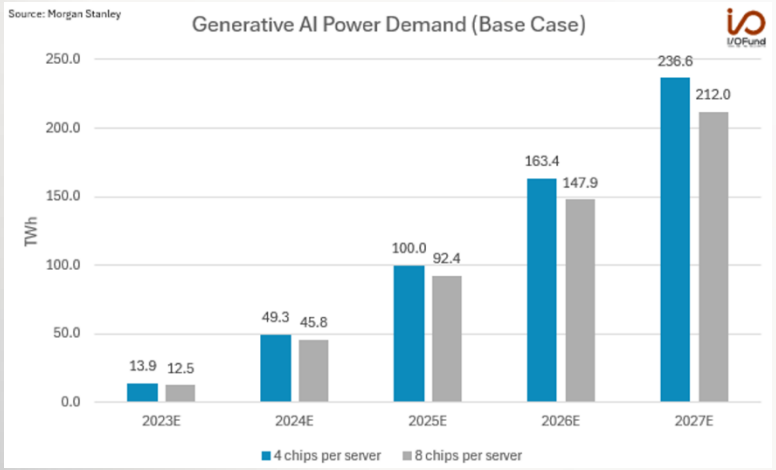
- AI technology has entered various consumer products
- Generative AI:
- Chatbots: chatgpt, gemini
- Image Generation: Dall-E, Stable Diffusion

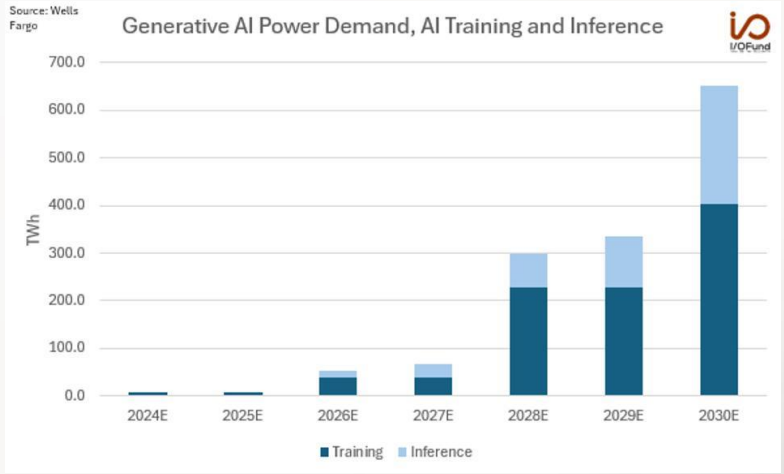


- Contained in many products:
- Social Media Platform (image, video processing, network analysis, customer profiling, etc.)
- Smart Phone Apps: Image postprocessing and organization, automatic translators, writing assistants
- Robotics and autonomous driving
- …

Estimated Energy Consumption of AI

- [https://www.forbes.com/sites/bethkindig/2024/06/20/ai-power-consumption-rapidly-becoming-mission-critical/]
- power consumption for one GPU-Card rapidly increases: H100 350W up to 700W
  further generation expected to consume >1KW
- amount of shipped GPUs increases as well: NVIDIA shipped 3.76 million data center GPUs in 2023
- At peak of 700W and ~61% annual utilization, each GPU would draw 3.74 MWh
  ⇒ 3.76 million GPU shipments could consume as much 14,384 GWh (14.38 TWh).
- Morgan Stanley is estimating global data center power use will triple this year, from ~15 TWh in 2023 to ~46 TWh in 2024
- this is NVIDIA Hardware only, but further producers like INTEL and AMD try to increase their market share as well.
- Though power demand is rapidly increasing, the computational power for each GPU is increasing significantly stronger ⇒ it is computational demand driving the development



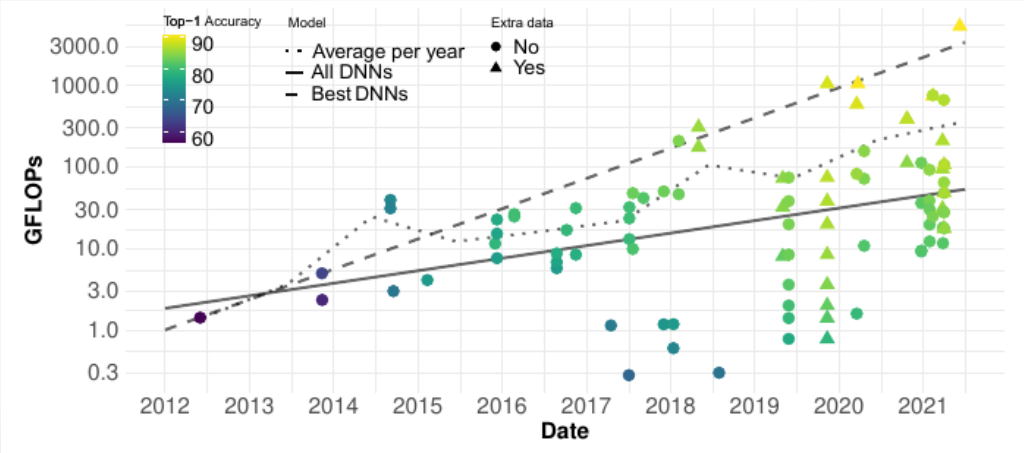


Energy Efficient Computing

- Methods to reduce the Energy Consumption:
- High-performance HW requires up to ca. 40%(air cooling) energy for cooling
  ⇒ Liquid cooling is usually more energy-efficient
  (e.g., a warm water cooling system requires only ca. 15%)
  ⇒ Reuse generated heat
- Green Computing: Managing Hardware to run more energy–efficient
  Examples: optimize data center operations, limit the max power consumption of GPUs, Improve scheduling, shut down unused servers (consume 75-80% of used servers), etc.
- More efficient algorithms:
  - computing the same workload on the same HW in less time
  - major goal of computer science research:
    - software runs on a broader spectrum of hardware
    - software running on older hardware saves money
    - less workload means less computational time and thus, less energy is used as well.

Saving Energy in AI

- training (learning new models)
  - there is a general trend to evaluate not just performance but also model size
  - Many upcoming AI-based services like like ChatGPT, Dall-E, etc. are based on the so-called transformer architectures
  - transformers are expensive due to multi-head self-attention blocks, which generate intermediate results in quadratic size of the input
    ⇒ newer models like the Mamba-Blocks or Sparse-Transformers reduce intermediate results to linear complexity
  - often better models increase the computational effort superlinear for a modest improvement in accuracy:



Reliability and Deep Learning

- Research is usually done as a proof of concept:
  - demonstrates that a machine can solve a task at all
  - improvements are measured on benchmark datasets
  - tasks are usually well-defined for applying metrics
  - results tend to overfit of used benchmark data
- using Neural Networks in products:
  - data distributions might vary from training/benchmark data set
  - inference data might yield unknown situations
  - research metrics might not directly mirror the requirements of a situation
  - people want to know why the networks generate specific outputs
  - some errors are usually worse than others

Coping with prediction errors

Neural networks generally generate the input which minimizes the average error:

- errors still occur, in particular, if the task is difficult, but users tend to assume that the machine is 100% right.
- errors are measured with respect to validation data on data displaying a different distribution error rate might increase
- biases in the training data transfer to the model:
  - machine learning is based on correlations, not causal dependencies
  - models might underperform in underrepresented regions of the data space
  - generative AI might generate inappropriate content

Attacks on Neural Networks

- As Neural Networks become part of important systems, they become target of various fraudolent attacks

- Typical attacks:
  - adversarial attacks: generating instances where the network will malfunction
  - backdoor attacks: manipulate training not to work is a specific pattern is included in the instance
  - triggering generative AI to generate inappropriate content

Course Outline

- Mathematical Foundations
- Neural Network Basics
- Training Neural Networks with Gradient Descent
- Basic Types of Layers
- Attention Mechanisms
- Standard Architectures
- Self-Supervised Learning
- Generative Models
- Sequential Planning and Reinforcement Learning
- Deep Reinforcement Learning