Autumn 2024

Lib4RI Training Series: Module 8 – AI Made Easy: A Beginner's Toolkit for Streamlining Your Projects

Dr. Corin Otesteanu

Library for the Research Institutes within the ETH Domain: Eawag, Empa, PSI & WSL



Table of Contents

- 1. Introduction
- 2. Convolutional neural networks
 - How CNNs work
 - Convolution filters
 - Visual (interactive) explanation
 - Applications
- 3. Large-language models
 - How LLMs work
 - Visual (interactive) explanation
 - Current models

- Applications
- Local use
- Pros and cons
- 4. Generative (image) models
 - Visual (interactive) explanation
 - Current models
 - Applications
- 5. Conclusion

Introduction



What is Artificial Intelligence?

- Al is technology that enables computers to mimic human intelligence
- Key components:
 - Machine Learning: Systems that learn from data
 - Deep Learning: Advanced ML using neural networks
 - Natural Language Processing: Understanding human language
- AI systems can:
 - Recognize patterns
 - Make decisions
 - Generate content
 - Solve complex problems



How AI Models Work

- 1. Training Phase:
 - Models learn from vast amounts of data
 - Identify patterns and relationships
 - Adjust internal parameters
- 2. Inference Phase:
 - Apply learned patterns to new data
 - Generate predictions or outputs
 - Continuous improvement possible

Generative AI: The Creative Side

Types of Generative AI:

- Text Generation (LLMs)
 - ChatGPT, Claude, Bard
- Image Generation
 - DALL-E, Midjourney, Stable Diffusion
- Code Generation
 - GitHub Copilot, Amazon CodeWhisperer
- Audio/Music Generation
 - Mubert, OpenAI Jukebox



Artificial Intelligence

• Al refers to machines mimicking human intelligence

- To perform tasks such as
 - Recognizing and understanding images
 - Understanding language
 - Reasoning and making decisions

M Lib4RI

Artificial Intelligence

• AI refers to machines mimicking human intelligence

- To perform tasks such as
 - Recognizing and understanding images
 - Understanding language
 - Reasoning and making decisions
- Machine Learning (ML): computer systems that learn from data, without following explicit instructions
- Deep Learning (DL): a subset of ML, focusing on neural networks
- Generative AI: systems that create new data from learned patterns



ML vs. DL







https://theresanaiforthat.com/

O Database with latest AI tools



Library for the Research Institutes within the ETH Domain: Eawag, Empa, PSI & WSL

MV Lib4RI

Exercise

Go to ChatGPT and prompt the following:

• Prompt: Act as an expert in [your field of expertise]. Do a literature online search to find what are the latest developments in the field of [specific area in yout field of expertise] and the research gaps in the area? Next propose novel research ideas, thinking about the following steps: importance, impact and implementation feasibility. Finally list the 3 most promising areas of research

• Rate the promising research areas on a scale from 1-10

Convolutional Neural Networks (CNNs)

Library for the Research Institutes within the ETH Domain: Eawag, Empa, PSI & WSL



Convolutional Neural Networks (CNNs)

a class of deep neural networks, primarily used for analyzing visual imagery
utilize convolutional layers to detect features in images
capable of learning complex patterns in large amounts of data

• fundamental building blocks in many modern deep learning architectures



How CNNs work

- Input layer: receives raw pixel values
- Convolutional layers: apply filters to detect features
- Activation functions: introduce nonlinearity
- Pooling layers: reduce spatial dimensions
- Fully connected layers: combine features for final output



Naranjo-Torres, J.; Mora, M.; Hernández-García, R.; Barrientos, R.J.; Fredes, C.; Valenzuela, A. A Review of Convolutional Neural Network Applied to Fruit Image Processing. Appl. Sci. 2020, 10, 3443. https://doi.org/10.3390/app10103443

Library for the Research Institutes within the ETH Domain: Eawag, Empa, PSI & WSL



How CNNs work

- Input layer: receives raw pixel values
- Convolutional layers: apply filters to detect features
- Activation functions: introduce nonlinearity
- Pooling layers: reduce spatial dimensions
- Fully connected layers: combine features for final output
- Loss calculation: compare output with expected result
- Backpropagation: adjust filter weights to minimize error
- Optimization: update parameters using algorithms like gradient descent until convergence



Convolution Filters

Box filter

 Replaces each pixel with an average of in a neighborhood (3x3)







o[x,y]

0	0	0	0	0	0	0	0	0	0						
0	0	0	0	0	0	0	0	0	0						
0	0	0	90	90	90	90	90	0	0						
0	0	0	90	90	90	90	90	0	0						
0	0	0	90	90	90	90	90	0	0						
0	0	0	90	0	90	90	90	0	0						
0	0	0	90	90	90	90	90	0	0						
0	0	0	0	0	0	0	0	0	0						
0	0	90	0	0	0	0	0	0	0						
0	0	0	0	0	0	0	0	0	0						
				_		_	7			, 	 _	_			

 $o[m,n] = \sum g[k,l] f[m+k,n+l]$ k,l



 $f[\mathbf{x},\mathbf{y}]$

o[x,y]

0	0	0	0	0	0	0	0	0	0										
0	0	0	0	0	0	0	0	0	0		0	10	20	30	30	30	20	10	
0	0	0	90	90	90	90	90	0	0		0	20	40	60	60	60	40	20	
0	0	0	90	90	90	90	90	0	0		0	30	60	90	90	90	60	30	
0	0	0	90	90	90	90	90	0	0		0	30	50	80	80	90	60	30	
0	0	0	90	0	90	90	90	0	0		0	30	50	80	80	90	60	30	
0	0	0	90	90	90	90	90	0	0		0	20	30	50	50	60	40	20	
0	0	0	0	0	0	0	0	0	0		10	20	30	30	30	30	20	10	
0	0	90	0	0	0	0	0	0	0		10	10	10	0	0	0	0	0	
0	0	0	0	0	0	0	0	0	0										

$$o[m,n] = \sum_{k,l} g[k,l] f[m+k,n+l]$$



Convolution with different filters



Library for the Research Institutes within the ETH Domain: Eawag, Empa, PSI & WSL

Convolution – Example Filters





Convolution – Example Filtered Images



Smoothed Image

Laplacian of Gaussian (LoG)

Disk Detector (9x9)



CNN explainer https://poloclub.github.io/cnn-explainer/



Wang, Zijie J., et al. "CNN explainer: learning convolutional neural networks with interactive visualization." *IEEE Transactions on Visualization and Computer Graphics* 27.2 (2020): 1396-1406.

Library for the Research Institutes within the ETH Domain: Eawag, Empa, PSI & WSL

 \odot



CNNs applications

- Image classification
- Object detection
- Facial recognition
- Image segmentation
- Image captioning





Example of CNN application: image segmentation

- Image segmentation is a technique that divides an image into multiple regions, where each pixel in a region shares similar characteristics
 - often domain specific
- Meta created the Segment Anything Model
 - o domain agnostic
- SAM2 can also be applied to video in addition to image data to segment regions on interest

MV Lib4RI

Example: Meta SAM2





Example: Meta SAM2

Initial Image



Prompt: Sky



Prompt: Boats





Example: Meta SAM2





Example: Meta SAM2

Try it yourself:

https://huggingface.co/spaces/wondervictor/evf-sam2





MV Lib4RI

Example: Human emotion detection



E Predictions	
	sad
sad	
angry	
happy	

Prediction time (s)

0.003783702850341797

Library for the Research Institutes within the ETH Domain: Eawag, Empa, PSI & WSL

Large language models (LLMs)

Library for the Research Institutes within the ETH Domain: Eawag, Empa, PSI & WSL

31

 \odot

Large language models (LLMs)

 AI models trained on vast amounts of text data to understand and generate human-like text

Examples:

• GPT (Generative Pre-trained Transformer)

- decoder only models (autoregressive models)
- predicting the next word in the sentence
- best suited for tasks involving text generation

• BERT (Bidirectional Encoder Representations from Transformers)

- o encoder only models
- sentence classification, named entity recognition
- T5 (Text-to-Text Transfer Transformer)
 - encoder decoder models
 - o summarization, translation

MV Lib4RI

Large language models (LLMs)

- Massive scale (billions of parameters)
- Self-supervised learning on diverse text corpora
- Ability to perform multiple tasks without specific training
 - Text generation and summarization
 - Question answering
 - Language translation
 - Conversational AI and chatbots



How Large are LLMs?

o LLaMa 3.1 405B has 405 billion parameters

o Training the Llama 3.1 405B model required over 16,000 NVIDIA H100 GPUs

• More than 30 million GPU hours (~80 days)

MV Lib4RI

Large language models (LLMs)

Alignment Pre-training **Internet Data** Task-Specific Training Code Generation •Books & Literature •Text **Summarization** •Code **Repositories** •Translation •Scientific Question Answering Papers •Synthetic data

Human Preferences •Task Quality Metrics •Safety Guardrail •RLHF Training Zero-shot Prompting Few-shot Prompting Chain of Thought

System Prompts



How LLMs work

Token-Based Processing

 LLMs process text by breaking it down into tokens, which can be whole words or parts of words

• Transformer backbone (Attention Mechanism)

• LLMs use attention to understand the relationships between words in a sentence

Training Objective

 LLMs are trained on large datasets to predict the next word in a sequence, to generate a coherent plausible text

 \odot


LLMs explained https://poloclub.github.io/transformer-explainer/



Cho, Aeree, et al. "Transformer Explainer: Interactive Learning of Text-Generative Models." arXiv preprint arXiv:2408.04619 (2024).

Library for the Research Institutes within the ETH Domain: Eawag, Empa, PSI & WSL

Large language models overview

Feature	ChatGPT-4o	Microsoft Copilot	MS Copilot Pro	Gemini 1.5 Flash	Gemini 1.5 Pro	Claude 3.5 Sonnet	Grok-2	Llama 3.2
Image generation	DALL-E 3	DALL-E 3	DALL-E 3	Imagen 3	Imagen 3	no	yes	yes
File upload	text, sound, images	only images	text, sound, images	sound, images	text, sound, images	text, sound, images	text, sound, images	text, sound, images (Llama vision)
Context window	128k	4k, 32k	4k, 32k	1M	2M	200k	128k	128k
Internet access	yes	yes	yes	yes	yes	no	Yes	no
Data privacy	opt out available	no	yes for entreprise	no	yes	Only user rated or flagged data	No	yes (self-hosted), unknown (whatsapp)
Price (USD)	free (or 20/month)	free	20/month	free	20/month	Free (or 20/month)	X Premium, Imarena	free (self-hosted), whatsapp
Deployment	Cloud	Cloud	Cloud	Cloud	Cloud	Cloud	Cloud	Self-hosted possible

Large language models performance comparison

		Hard					
Madal	Overall	Prompts	Instruction	Coding	Math		Longer
Model	Overall	(Overall)	Following	Coding	Math	Multi-Turn	Query
chatgpt-4o-latest-20240903	1	2	2 1	2	. 3	6 1	1
o1-preview	1	1	1	1	1	1	1
o1-mini	3	3 1	3	8 1	1	3	2
gemini-1.5-pro-002	3	3 4	4	. 4	. 3	3 3	3
gemini-1.5-pro-exp-0827	4	ŀ 4	4	- 4	. 3	6 4	. 3
grok-2-2024-08-13	6	6 6	6 6	6 4	. 7	4	. 6
yi-lightning	6	6 4	5	5 4	. 3	3	5
gpt-4o-2024-05-13	6	6 6	5 5	5 4	. 7	4	. 5
glm-4-plus	ç) 6	6 7	′ 4	. 8	8 8	5
gpt-4o-mini-2024-07-18	ç	9 9) 10) 5	13	6 4	- 5
gemini-1.5-flash-002	ç	9 19	9 9	18	S S	16	6 4
gemini-1.5-flash-exp-0827	ç) 10) 10) 17	8	s g	9 4
llama-3.1-nemotron-70b-instruct	ç	9 8	8 8	8 7	7	′ 4	. 21
claude-3.5-sonnet-20240620	10) 6	6 6	6 4	. 3	6 4	. 7
Chignow Zheng 5 24-08-13 al. Chatbot Arena: A	An Open Plat 4	form for Evaluatin	g LLMs by Human	Preference. arX	2024;2403.04	² 13	13
llama-3.1-405b-instruct-fp8	11	S) 10) 7	8	5 5	5 15

Library for the Research Institutes within the ETH Domain: Eawag, Empa, PSI & WSL

39

Large language models performance - language

	Model	English	Chinese	German	French	Spanish	Russian	Japanese	Korean
	chatgpt-4o-latest-20240903	1	1	1	1	1	1	1	1
	o1-preview	1	2	2	1	1	1	1 1	1
	o1-mini	3	5 2	2	1	1	4	L 1	1
	gemini-1.5-pro-002	4	. 1	1	3	1	1	1 1	1
	gemini-1.5-pro-exp-0827	4	- 2	2 1	1	1	1	1 1	1
	grok-2-2024-08-13	4	. 5	; 1	1	1	5	5 2	2 1
	yi-lightning	4	- 2	2 2	2 1	1	14	- 5	5 4
	llama-3.1-nemotron-70b-instruct	4	- 7	1	1	1	31	1 2	2 4
	llama-3.1-405b-instruct-bf16	6	5 19) 2	I З	6 4	. 14	l 6	5 5
	gpt-4o-2024-05-13	7	7	2	2 1	2	5	5 1	1
	llama-3.1-405b-instruct-fp8	7	24	. 5	5 2	2 4	. 8	3 18	6
	gpt-4o-mini-2024-07-18	3	s 10) 2	2 1	1	5	5 5	5 5
	glm-4-plus	ç) 5	5 2	2 1	1	5	5 5	5 1
	grok-2-mini-2024-08-13	ç) 12	2 3	3 3	6 4	- 13	3 5	5 4
	yi-lightning-lite	10) 10) 2	2 1	1	16	; 4	5
	qwen-max-0919	11	12	2 4	l 1	1	5	5 6	5 13
	llama-3.1-70b-instruct	11	30	14	8	5 7	26	5 28	3 16
	gemini-1.5-flash-002	13	5 5	5 2	2 1	4	. 4	l 1	1
Chiang W, Z	gpt-4o-2024-08-06	13	5 15	5 5	5 4	. 2	9) 5	5 1
https://Imare	claude-3.5-sonnet-20240620	14	- 10) 4	4 2	2 1	5	5 5	5 4
	qwen2.5-72b-instruct	14	- 7	10) 3	2	. 14	14	4

Library for the Research Institutes within the ETH Domain: Eawag, Empa, PSI & WSL

Large language models performance comparison

QUALITY



Output Tokens per Second; Higher is better

SPEED

PRICE

USD per 1M Tokens; Lower is better



 \odot

https://artificialanalysis.ai/, accessed 25/10/2024

Library for the Research Institutes within the ETH Domain: Eawag, Empa, PSI & WSL

41

Example - generating powerpoint slides

 write a vba script that makes powerpoint slides for the topic «AI Made Easy: A Beginner's Toolkit for Streamlining Your Projects»

ChatGPT

- Here is a VBA script that creates the PowerPoint slides based on the provided content
- You can run this script within Microsoft PowerPoint's VBA editor

Sub	CreateOpenAISlides()
	Dim pptApp As Object
	Dim pptPres As Object
	Dim pptSlide As Object
	Dim slideIndex As Integer
	' Initialize PowerPoint application
	<pre>Set pptApp = CreateObject("PowerPoint.Application")</pre>
	pptApp.Visible = True
	' Create a new presentation
	Set pptPres = pptApp.Presentations.Add
	' Slide 1: Title Slide
	<pre>slideIndex = slideIndex + 1</pre>
	Set pptSlide = pptPres.Slides.Add(sl Index, 1) ' 1 for ppLayoutTitle



Example - generating powerpoint slides

To use this script:

• Open PowerPoint

• Press Alt + F11 to open the VBA editor

• Insert a new module

• Copy and paste this code

• Run the macro



Exercise

 Consider (an already published conference paper or document related to your work)

• upload it to ChatGPT and prompt one of the following:

Create a powerpoint presentation from a

conference abstract

OR

• work that you might be required to present

• Prompt:

write a vba script that makes [5-10 slides] powerpoint slides for the topic [insert topic]

Rate the presentation



Example: Creating a GUI app

Prompt

Make a streamlit app. The app should have three figures that plots three time signals. A fourth figure should show the spectrum of the summed up time signals. There should be sliders that adjust the amplitude of each signal from 0 to 10 and the frequency of the signals from 0.1 to 50 Hz. The scale of the plots should be fixed to 10

Example: Creating a GUI app

Signal 1



Signal 2







Spectrum of Combined Signals

10.00

10.00

50.00



Example: Solving electric circuit problem

Prompt: For the circuit show below find the current and voltage for each of the five resistors.

Response:

- To find the current and voltage for each resistor the given circuit, we can follow these steps:
- 1. Determine the total resistance of the circuit.
- 2. Calculate the total current supplied by the voltage source.
- 3. Find the voltage drop across each resistor using Ohm's Law.
- 4. Calculate the current through each resistor.

• Here's how we can do this step-by-step: [...]



47



Example: Solving electric circuit problem

Summary

- Current through R1R_1R1 : 0.357A0.357 A0.357A
- Voltage across R1R_1R1 : 3.57V3.57 V3.57V
- Current through R2R_2R2 : 0.286A0.286 A0.286A
- Voltage across R2R_2R2 : 1.43V1.43 V1.43V
- Current through R3R_3R3 : 0.0715A0.0715 A0.0715A
- Voltage across R3R_3R3 : 1.32V1.32 V1.32V
- Current through R4R_4R4 : 0.055A0.055 A0.055A
- O Voltage across R4R_4R4 : 0.11V0.11 V0.11V
- Current through R5R_5R5 : 0.0183A0.0183 A0.0183A
- O Voltage across R5R_5R5 : 0.11V0.11 V0.11V
- This approach involves simplifying the circuit step-by-step and applying Ohm's Law and Kirchhoff's Laws to find the desired currents and voltages.



Where LLMs excel

• Conversations (chatbot, customer service, etc)

- Writing in a certain style
- Translations
- Correcting grammar
- Summarization
- Brainstorming (limited)
- Coding (limited)

LLM limitations and downsides

Can produce factually incorrect results

o biases

- hallucinations (fabricated responses)
 - invented facts
 - fictitious citations
 - incorrect explanations
 - non-existent entities

○ no or limited reasoning capabilities (as of now) ○ model architecture

- lack of knowledge
 - knowledge cutoff
 - lack of domain specific knowledge
- generated text might constitute *plagiarism*

Causes

- biased training data
- o model architecture
 - o generate novel ideas, styles, genres
 - propose valuable hypotheses

limited and outdated training data

 \odot

- o (unknown) errors
- known factual errors

o training data and prompting

LLM limitations and downsides

LLMs process text using tokens, not characters.

• Tokens can be single characters or entire words.

Tokenization makes direct character counting difficult for LLMs.

how many letters does this prompt have?

It seems there was an issue calculating the character count directly. I'll count the letters here manually:

The prompt you provided has **296 characters**. [>-]

 Workaround: ask AI to use programming language to count the characters to generate correct result



Minimizing LLM inaccuracies - prompting

Prompt: specific input or instruction to a language model to generate a desired output

- Good practice for prompt design
- Be specific and clear in the task you want done
- Break down complex tasks in simpler tasks
- Provide constraints
 - E.g. 3 bulletpoints, one paragraph, etc
- O Style
 - o formal, informal
 - o academic, journalistic, creative

• After output, iterate, refine prompts



Minimizing LLM inaccuracies - prompting

Prompt: specific input or instruction to a language model to generate a desired output

Good practice identifying factually incorrect results

- Prompt for alternative point of view
- Ask for the references
- Ask for positive or negative of a certain subject
 - If you ask for a proof, ask also to prove the opposite

• Use the same prompt on several LLMs and compare results

• When in doubt, always check with a reputable resource

You are responsible for data generated by genAl



Prompt template example

Role Definition

• Define the role of the AI clearly

• Context:

• Provide any necessary background information or context

Input Data:

- Specify the data or information the AI should use
- Upload your own data

Instructions:

• Give clear and specific instructions on what you want the AI to do

Constraints

• Mention any constraints like length/bulletpoint, format, or style

Other

- Multi-shot prompting (if applicable)
- Chain of thought

Library for the Research Institutes within the ETH Domain: Eawag, Empa, PSI & WSL

54



Prompt example 1

Role Definition

• You are an AI assistant that provides critical reviews of scientific research papers

O Context:

The following text is a research paper on the effects of microplastics on marine life

Input Data:

[Insert of upload your own data]

• Instructions:

 Critically review the research paper, focusing on the methodology, data analysis, and conclusions. Highlight any strengths and weaknesses, and suggest areas for improvement.

Constraints

• The analysis should have 10 points. Use formal language.

• Other (optional)

- Multi-shot prompting: example of a good review
- Chain of thought: provide a step by step analysis for each point

Library for the Research Institutes within the ETH Domain: Eawag, Empa, PSI & WSL



Prompt example 2

O Role Definition

• You are an expert in Python data visualization with extensive experience in scientific plotting.

• Context:

- I'm a researcher in climate science working with temperature time series data.
- Input Data:
 - No input data

Instructions:

• Create a comprehensive visualization of global temperature anomalies over the past century

Constraints

Use matplotlib library. Data should be between 1924 and 2024. Should be publication-ready quality. The code to reproduce the plots should be given as a jupyter notebook



Accessing LLMs

- Through their respective website
- o On Imarena.ai
 - select desired model
- On huggingface.co/chat
 - select desired open access model
- o On NVidia
 - <u>https://build.nvidia.com/explore/</u> <u>discover</u>
 - select desired open access model

0	chatgpt-4o-latest-20240903
V	chatgpt-4o-latest-20240903
	gpt-4o-mini-2024-07-18
	gpt-4o-2024-08-06
	gpt-40-2024-05-13
	grok-2-2024-08-13
	grok-2-mini-2024-08-13
	claude-3-5-sonnet-20241022
	claude-3-5-sonnet-20240620
	llama-3.1-nemotron-70b-instruct
	llama-3.1-nemotron-51b-instruct
	llama-3.2-vision-90b-instruct
	llama-3.2-vision-11b-instruct

Settings

Models

meta-llama/Meta-Llama-3.1-70B-Instr... Active

CohereForAl/c4ai-command-r-plus-08-2024

Qwen/Qwen2.5-72B-Instruct

nvidia/Llama-3.1-Nemotron-70B-Instruct-HF

meta-llama/Llama-3.2-11B-Vision-Instruct

NousResearch/Hermes-3-Llama-3.1-8B

mistralai/Mistral-Nemo-Instruct-2407

microsoft/Phi-3.5-mini-instruct

57

Customizing LLM preferences - ChatGPT

	1 Share	
°	My GPTs	
¥	Customize ChatGPT	
¢	Settings	
¢	Upgrade Plan	
[→	Log out	

Customize ChatGPT

Custom Instructions 🕕

What would you like ChatGPT to know about you to provide better responses?

0/1500			

How would you like ChatGPT to respond?

Enable for new chats

Cancel

Save

 • Custom instructions from https://help.openai.com /:

- How formal or casual should ChatGPT be?
- How long or short should responses generally be?
- How do you want to be addressed?
- Should ChatGPT have opinions on topics or

Library for the Research Institutes within the ETH Domain: Eawag, Empa, PSI & WSL

Customizing LLM preferences - ChatGPT



Security



Customizing LLM preferences - Claude

• Click on your username (bottom left corner) and then settings. Enable artifacts.

Claude



C

60

Customizing LLM preferences - Claude

• Click on your username (bottom left corner) and then enable feature preview tools

八 Feature Preview

 Built-in Code Sandbox enables Claude to perform complex calculations, data analysis, and iterative testing



X



Claude - Built-in Code Sandbox

- Enables Claude to perform complex calculations, data analysis, and iterative testing
- Enhances Claude's ability to process information and ensures more accurate answers



Julius.ai

- AI data analyst tool for visualizing and analyzing complex data in seconds
 ChatGPT or Claude based
- Capabilities
 - chat with your data
 - create visual representations and animations
 - build classification or forecasting models and run them



Example: Julius.ai

- Iris dataset consists of 3 different types of irises' (Setosa, Versicolour, and Virginica) petal and sepal length and width
- Prompt: using the iris dataset build a model that classifies the data in 3 classes
- Answer: The model has been successfully trained on the Iris dataset, achieving an accuracy of 1.0, indicating perfect classification on the test set. The classification report provides detailed metrics for each class, including precision, recall, and F1-score.

64



Example: Julius.ai

/> T	rained a Random Forest Classifier on the Iris Dataset		^							
e P	Python	O Rerun code 🖉 Edit code	G							
10 11 11 11 11 11 11 11 11 11 11 11 11 1	<pre># Import necessary libraries import pandas as pd from sklearn.datasets import load_iris from sklearn.model_selection import train_test_split from sklearn.metrics import RandomForestClassifier from sklearn.metrics import classification_report, accuracy_score # Load the iris dataset iris = load_iris() % X = iris.data # features y = iris.target # target classes # Split the dataset into training and testing sets X_train, X_test, y_train, y_test = train_test_split(X, y, test_size: # Create a Random Forest Classifier model = RandomForestClassifier(n_estimators=100, random_state=42) # Train the model model.fit(X_train, y_train) # Make predictions y_pred = model.predict(X_test) # Evaluate the model accuracy = accuracy_score(y_test, y_pred)</pre>	=0.2, random_state=42)	Accuracy: 1.0 Classification accuracy macro avg weighted avg Show less	n Report: precision 1.00 1.00 1.00 1.00	recall 1.00 1.00 1.00 1.00 1.00	f1-score 1.80 1.00 1.00 1.00 1.00 1.00	support 10 9 11 30 30 30 30			
2 2 2 3 3 3 3	<pre>7 report = classification_report(y_test, y_pred) 8 9 # Print the results 9 print('Accuracy:', accuracy) 1 print('Classification Report:\ 2 ', report)</pre>	The per inc	e model ha fect classif luding prec	s been suc fication on cision, recal	cessfull the test II, and F	y trained set. The 1-score.	on the Iri: classifica	s dataset, achieving an tion report provides de	accuracy of 1.0, indic etailed metrics for eac	ating h class,

Code Explanation

This code snippet loads the Iris dataset, splits it into training and testing sets, trains a Random Forest Classifier, and evaluates its performance using accuracy and a classification report.

Code Output



Exercise

Consider (an already published conference paper or document related to your work)

• upload it to ChatGPT and prompt one of the following:

o I am a

- o research scientist in ...
- o technician using device
- IT support working with ...
- Administrative assistant in charge of ...
- please summarize the following document as accurately as possible with an emphasis on the quantitative and qualitative aspects presented in the document.

• Rate the summary

Exercise

Consider (an already published conference paper or document related to your work)

- Upload the results and discussion section to ChatGPT and prompt one of the following:
 - Act as a scientific reviewer for a journal paper. Go through the key sections to assess its quality, structure, methodology, and scientific contribution. Provide an assessment of quality of the paper
 - Act as a manager and provide feedback on the following document provided by a
 - o administrative assistant
 - o technician
 - IT support
 - Asses the clarity, organization, completeness, technical content (if applicable) and speeling

• Rate the review

Some parameters of LLMs

- Temperature:
 - Controls the randomness of the model's output
 - Lower values (e.g., 0.2) make the output more focused and deterministic (more predictable answers)
 - Higher values (e.g., 1.0 or above) make the model more creative and diverse in its responses
- Top-k Sampling:
 - This parameter limits the number of possible next tokens to the top k tokens with the highest probabilities
 - A lower k reduces randomness by restricting choices to the most likely tokens
- Top-p (Nucleus) Sampling:
 - Instead of limiting by a fixed number (like top-k), top-p sampling restricts the model to choosing from the most probable tokens whose cumulative probability
 is greater than p
 - p = 0.7 means the model will sample from the smallest group of tokens whose total probability is at least 0%, leading to more controlled output
- Presence Penalty:
 - Adjusts the model's likelihood of discussing new topics. A higher presence penalty encourages the model to introduce new concepts into the conversation instead of repeating itself
- Frequency Penalty:
 - Controls how much the model penalizes repeated tokens. A higher frequency penalty discourages the model from generating repeated words or phrases
- Max Tokens (Length of Output):
 - This sets the maximum length of the generated response. It can limit the model's output to avoid excessively long answers

Library for the Research Institutes within the ETH Domain: Eawag, Empa, PSI & WSL



ChatGPT parameter tuning example

• Only possible via API, Playground (requires subscription)

o can directly control parameters such as

- o temperature,
- o top-k,
- o top-p
- o max_tokens
- Example json API Request:
- o { "model": "gpt-4", "prompt": "Tell me a creative story about a robot", "max_tokens": 100, "temperature": 0.8, "top_k": 50, "n": 1}

ChatGPT parameter tuning example

• Or via third parties (e.g. huggingface spaces)

es. Ensure your responses are accurate,
4096
0.95

Library for the Research Institutes within the ETH Domain: Eawag, Empa, PSI & WSL



Notebook LM

 designed to help users summarize, understand, and synthesize information from multiple documents

bases its answers on the documents you provide

• developed by Google, uses Gemini LLM

Key Features:

- Document Summarization
- O Chat with Sources
- o FAQ Suggestions
- Source Grounding



Notebook LM - Privacy

https://support.google.com/notebooklm/ - Accessed on 22.10.2024

We value your privacy and never use your personal data to train NotebookLM.

- If you are logging in with your consumer google account and choose to provide feedback, human reviewers may review your queries, uploads, and the model's responses to troubleshoot, address abuse or make improvements. Keep in mind that it's best to avoid submitting any information you wouldn't feel comfortable sharing.
- As a Google Workspace or Google Workspace for Education user, your uploads, queries and the model's responses in NotebookLM will not be reviewed by human reviewers, and will not be used to train AI models.


Notebook LM

Prompt:

What is the discrete wavelet transform?

Gemini 1.5

Key concepts:

- Wavelets: Wavelets are mathematical functions that are localized in both time and frequency. They are used as basis functions to represent the signal.
- Decomposition: The DWT decomposes the signal into a set of wavelet coefficients that represent the signal's information content at different scales (frequencies) and locations (time).
- Multiresolution analysis: The DWT provides a multiresolution analysis of the signal, allowing for the examination of the signal's behavior at different levels of detail.

Applications:

The Discrete Wavelet Transform (DWT)

The DWT is a mathematical transformation that analyzes signals in both **time and frequency domains**, offering a **multi-scale** representation of the signal **1 2**. It decomposes a signal into a set of wavelet coefficients, which represent the signal's behavior at different scales and locations **2**.

Notebook I M

The DWT uses **basis functions** called **wavelets** 1 3. Unlike the Fourier transform, which uses sinusoidal-waves as basis functions, wavelets are **localized in both time and frequency** 1 4. This localization allows the DWT to capture transient features and abrupt changes in the signal more effectively 2 4. Key Concepts in DWT:

- Mother Wavelet: A prototype wavelet function that serves as the building block for generating other wavelets through scaling (dilation) and translation (3) (6).
- Scaling: Stretching or compressing the mother wavelet to analyze the signal at different resolutions
 6. Larger scales correspond to lower frequencies and coarser details, while smaller scales correspond to higher frequencies and finer details

Library for the Research Institutes within the ETH Domain: Eawag, Empa, PSI & WSL



Exercise

- o Go to <u>https://notebooklm.google.com/</u>
- Upload one of your (already published) papers

Or

A document that you want to analyze

o perform 2 queries for what information you want to find out

- E.g. What method was used to solve the problem?
- What was the accuracy?
- What were the limitations?
- o check the accuracy of the answers
- o check the accuracy of the source (page, paragraph)

Exercise

• Use ChatGPT for the same task as before

• Upload one of your (already published) papers

Or

• A document that you want to analyze

• perform 2 queries for what information you want to find out

- E.g. What method was used to solve the problem?
- What was the accuracy?
- What were the limitations?

o check the accuracy of the answers



Nvidia ChatRTX

O Chat With Your Files

 Load desired files to ChartRTX. Various formats accepted including txt, pdf, doc/docx, jpg, png, gif, and xml

• Talk to ChatRTX

• ChatRTX features an automatic speech recognition system .

O Data Search System Requirements

Platform	Windows	
GPU	NVIDIA GeForce™ RTX 30 or 40 Series GPU or NVIDIA RTX™ Ampere or Ada Generation GPU with at least 8GB of VRAM	
RAM	16GB or greater	
os	Windows 11	



GPT4AII

• A free-to-use, locally running, privacy-aware chatbot.

• No GPU or internet required





GPT4AII

 A free-to-use, locally running, privacy-aware chatbot. No GPU or internet required.



Generative (image) models



Generative (image) Models

• aim to learn or approximate the probability distribution of the data

○ (e.g. image data)

• primarily designed to generate new data samples that are similar to the training data (or to estimate the likelihood of data)

Examples

○ Variational autoencoders

○ Generative Adversarial Networks

• Probabilistic Diffusion Models



Generative Models: Diffusion Models

• aim to learn or approximate the probability distribution of the data

○ (e.g. image data)

• Primarily designed to generate new data samples that are similar to the training data (or to estimate the likelihood of data)



 \odot

Yang, Ling, et al. "Diffusion models: A comprehensive survey of methods and applications." *ACM Computing Surveys* 56.4 (2023): 1-39. Library for the Research Institutes within the ETH Domain: Eawag, Empa, PSI & WSL

MV Lib4RI

Diffusion models explained https://poloclub.github.io/diffusion-explainer/



Lee, Seongmin, et al. "Diffusion explainer: Visual explanation for text-to-image stable diffusion." arXiv preprint arXiv:2305.03509 (2023).

Comparison of GenAl for image generation

Name	Free	Advantages
DALL·E	Limited	Easy to use, integrated with ChatGPT, great prompt adherence and understanding, high-quality outputs
Stable Diffusion	Yes	Open-source, highly customizable, can be run locally. Online @ <u>https://stablediffusionweb.com/app/image-generator</u> or <u>https://huggingface.co/spaces</u>
lmagen (Google)	Yes	Easy to use, integrated with Gemini, high-quality outputs
Meta Al	Yes	Easy to use, integrated in WhatsApp, high-quality outputs
Leonardo.Al	Limited	Easy to use, high-quality outputs
Midjourney	Free trial	Strong community support, excels in artistic styles

MV Lib4RI

Some parameters for image generation

• Prompt: The input text guiding the model on what image to generate

- Negative Prompt: Specifies what elements should be excluded from the generated image
- Style: The visual or artistic genre applied to the generated image. Specified in prompt or by selecting a variation of the gen ai model
- Guidance Scale: Controls how closely the model follows the prompt versus adding creative freedom
- Steps: The number of iterations the model uses to refine and improve image quality
- Seed: A number ensuring reproducibility, producing the same image with the same input
- Aspect Ratio: The width-to-height ratio defining the shape of the generated image

MV Lib4RI

Prompt good practice

- Similar rules to LLM prompting apply.
 - Be specific, clear in the image description
- o Subject
 - Specify subject, e.g., animal, landscape, object
- o Setting
 - Outdoor, indoor, city
- o Style
 - Choose style, e.g., realistic, abstract, impressionist
- Attributes
 - Add attributes, e.g., colors, textures, lighting
- Composition
 - Define composition, e.g., close-up, wide-angle, portrait
- Background
 - Specify background, e.g., solid color, gradient, scene



Example: Image generation comparison

Prompt:

Create an image that visually explains the concept of anxiety detection using a smartwatch.

The poster should include the following elements:

- Smartwatch on a wrist: A modern, sleek smartwatch worn on a human wrist. The smartwatch should display physiological data such as heart rate, skin temperature, and oxygen levels. The screen should be divided into small sections showing different metrics in real-time.
- Physiological signals: Represent flowing lines and graphs around the smartwatch to indicate the tracking of signals like heart rate variability (HRV), skin conductance, and respiration rate. These lines should connect to sensors on the smartwatch, visually showing the detection of anxiety-related physiological changes.
- Anxiety detection concept: In the background or adjacent to the smartwatch, depict a brain with stress-related signals. Show signals that are connected between the smartwatch and the brain, symbolizing the detection and transmission of anxiety data.
- Subtle facial expression: A person wearing the smartwatch should have a slightly anxious expression—e.g., raised eyebrows, tense facial muscles—to represent the connection between the physiological measurements and emotional state.
- Data analysis visualization: Include small, clear data charts or AI model icons near the smartwatch, representing machine learning or AI analyzing the data from the smartwatch to detect anxiety.
- Abstract background: A minimalistic background with soft gradients (e.g., calming blues, greens) to contrast the wearable technology and signal flow, with subtle abstract shapes representing neural networks or data flow.



ChatGPT (DALL-E)



Library for the Research Institutes within the ETH Domain: Eawag, Empa, PSI & WSL



Microsoft designer (DALL-E)



Library for the Research Institutes within the ETH Domain: Eawag, Empa, PSI & WSL

M Lib4RI

IGemini (Imagen3 - Google)





Meta Al





Stable Diffusion





Leonardo Al





Library for the Research Institutes within the ETH Domain: Eawag, Empa, PSI & WSL

Assiste Delectio

MV Lib4RI

Example – image generation describing (research) projects

Anxiety detection

Glucose monitoring







 \odot

Sleep quality assessment

Library for the Research Institutes within the ETH Domain: Eawag, Empa, PSI & WSL

93

Example – image generation describing (research) projects

Exercise

Generate an image that illustrates one of the following

- your general research or work
- a catchy image for a student thesis
- a catchy image for a workshop or presentation related to your work

 \circ Rate the generate images on a scale from 1 – 10

- how accurately it followed prompt
- lack of distortions/artifacts
- overall quality



Limitations

 Prompt: generate an image with a frog standing on a cat standing on a cow standing on an elephant





Library for the Research Institutes within the ETH Domain: Eawag, Empa, PSI & WSL



Limitations

 Prompt: generate an image of a person with the left hand raised up, right hand tucked in the pocket, right leg raised up.





Limitations

- Does not accurately capture the pose, position, and number of objects specified in prompt
- Large or complex prompts will not lead to accurate generations



Image to Image generation

Website: https://huggingface.co/spaces/TencentARC/T2I-Adapter-SDXL-Sketch
 Prompt: cat on a field playing with balls. sun is shinning

• Style: photographic





Bonus: Stanford STORM

 STORM is a LLM system that writes Wikipedia-like articles from scratch based on Internet search



Image source https://github.com/stanford-oval/storm

https://storm.genie.stanford.edu/

Library for the Research Institutes within the ETH Domain: Eawag, Empa, PSI & WSL

 \odot



Bonus exercise

O Go to: <u>https://storm.genie.stanford.edu/</u>

Ask it to perform a literature review for a topic in which you are an expert on
 Write the motivation for the article (research/ tool / software/ overview)

• Rate the results

- 0 10 excellent
- 7-8 good work for an experienced person
- 5 good work for an intern / apprentice

M Lib4RI

Bonus: Video generation using runwayml.com

- Generate videos
- Free use for up to
 - 25s videos or
 - 25 generated images
- o Paid (12usd) monthly
 - o 25s videos or
 - 125s generated images
- Other payment options



MV Lib4RI

Bonus: Video generation using runwayml.com





Bonus Example: video generation runwayml.com

• Prompt:

• Many people appear queueing at the building entrance



o Image

Bonus Example: video generation runwayml.com

• Prompt:

- Location is a well light classroom auditorium.
- More than 50 people are present. It is very loud and noisy, the people are chatting with each other, some or checking their phones, other are using the laptop.
- The teacher walks in. He is wearing a lila button shirt and khakis.
- Everyone takes their places and becomes silent. The teacher writes on the board the title of the lesson: "Introductory AI at the Lib4RI Lessons".

MV Lib4RI

Bonus Example: video generation runwayml.com



Library for the Research Institutes within the ETH Domain: Eawag, Empa, PSI & WSL



Thank you for your attention

Please direct any (future) questions to

Dr. Corin Otesteanu

Library for the Research Institutes within the ETH Domain: Eawag, Empa, PSI & WSL