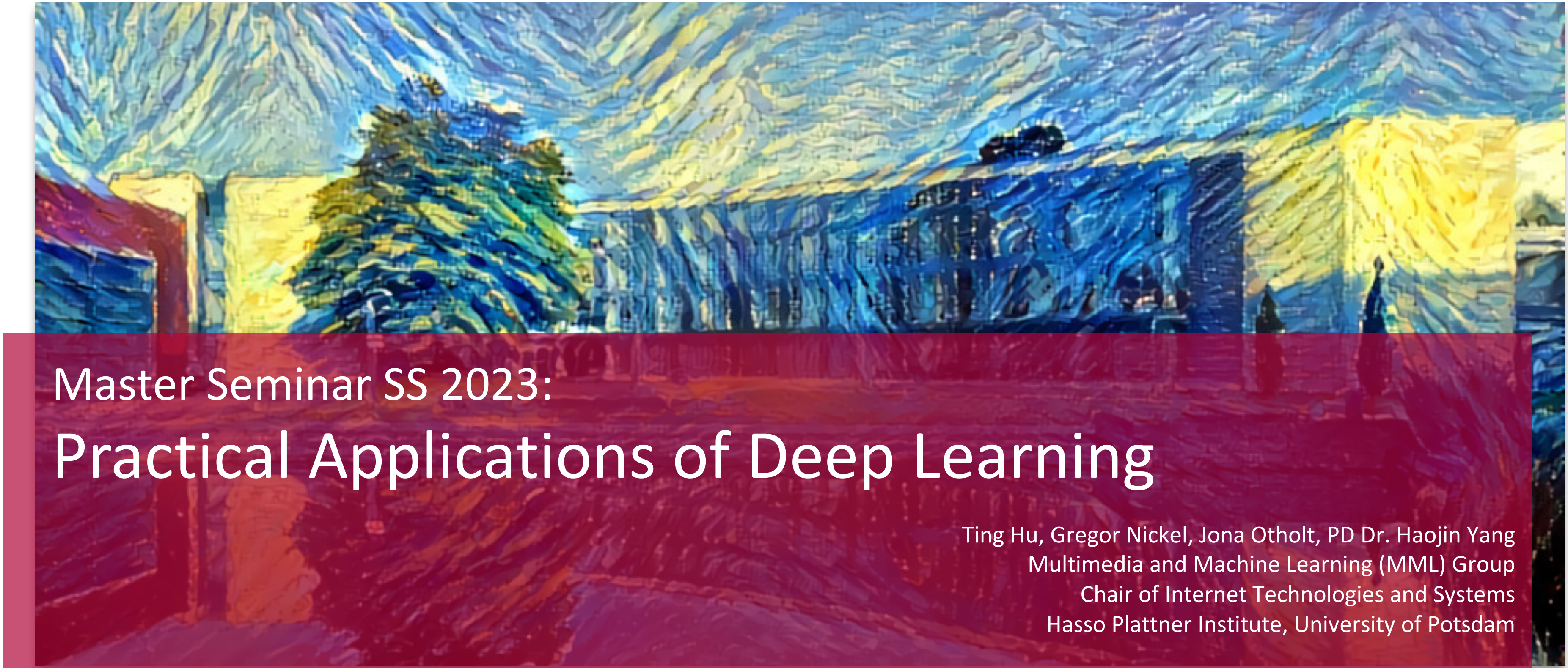


Master Seminar SS 2023:

Practical Applications of Deep Learning

Ting Hu, Gregor Nickel, Jona Otholt, PD Dr. Haojin Yang
Multimedia and Machine Learning (MML) Group
Chair of Internet Technologies and Systems
Hasso Plattner Institute, University of Potsdam

Content

- **Teaching team**
- Topics
- Important information

Personal Information

Ting Hu, M.sc

- Research background
  - 2009-2016 Bachelor and Master Degree in China
  - 2018 PhD Student at Hasso Plattner Institute
- Research interests
  - Nature language generation.
  - Efficient NLP models

Personal Information

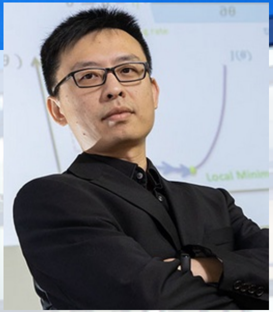
Gregor Nickel, **M.Sc.**

- Research background
  - 2013 – 2018 Bachelor Degree (RWTH Aachen University)
  - 2017 – 2018 Research assistant at the Chair of Imaging and Computer Vision at RWTH Aachen University
  - 2018 – 2020 Master Degree (RWTH Aachen University)
  - 2022 PhD Student at Hasso Plattner Institute
- Research interests
  - Computer vision and NLP
  - Binary neural networks and lightweight network architectures

Personal Information

Jona Otholt, M.Sc.

- Research background
  - 2015-2018 Bachelor Degree (Hasso Plattner Institute)
  - 2018-2021 Master Degree (Hasso Plattner Institute)
  - Since 2021 PhD Student at Hasso Plattner Institute
- Research interests
  - Computer vision, document analysis, unsupervised / weakly supervised learning



PD Dr. Haojin Yang

- 10/2002-01/2008 Medientechnologie (Dipl.-Ing.), University of Technology Ilmenau, Germany
- 11.2013, PhD in Computer Science (CS), Hasso Plattner Institute for Software Systems Engineering (HPI)/University of Potsdam
- 2015-present, Head of Multimedia and Machine Learning Research Group, ITS Chair, HPI
- Since 07/2019, Privatdozent (PD) at HPI/University of Potsdam
  - Habilitation thesis: Deep Representation Learning for Multimedia Data Analysis
- 11/2019-10/2020, Head of Edge Computing Lab Beijing, AI Labs & Video Cloud, Alibaba Group

HPI Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam

ITS - MML Group

- Group leader: PD Dr. Haojin Yang
- Team: PhD students, student co-workers
- Funding and supports: SAP, WPI, BMBF, BMUV, Meta AI
- Current research collaborators: NICSEFC, PNNL, NCSU, RUC-AI, CUHK, IBME

SAP®

PYTORCH

Wildenstein Plattner Institute

Zukunft Umwelt Gesellschaft

Research and Teaching

Computation and Energy Efficient Learning

- Binary neural networks
- Deep model compression
- Dynamic neural networks, e.g., dynamic BERT
- Efficient training and inference of LLM

Label Efficient Learning

- Dataset synthesis
- Novel/General class discovery

Edge AI

- Computation offloading, collaborative training and inference

AI MOOC@KI-Campus&openHPI

- **Title**: *Applied Edge AI: Deep Learning outside of the Cloud*
- **Link**: https://open.hpi.de/courses/edgeai2022

Master Seminar

- Practical Applications of Deep Learning SS (4 SWS)
- Machine Intelligence with Deep Learning WS (4 SWS)

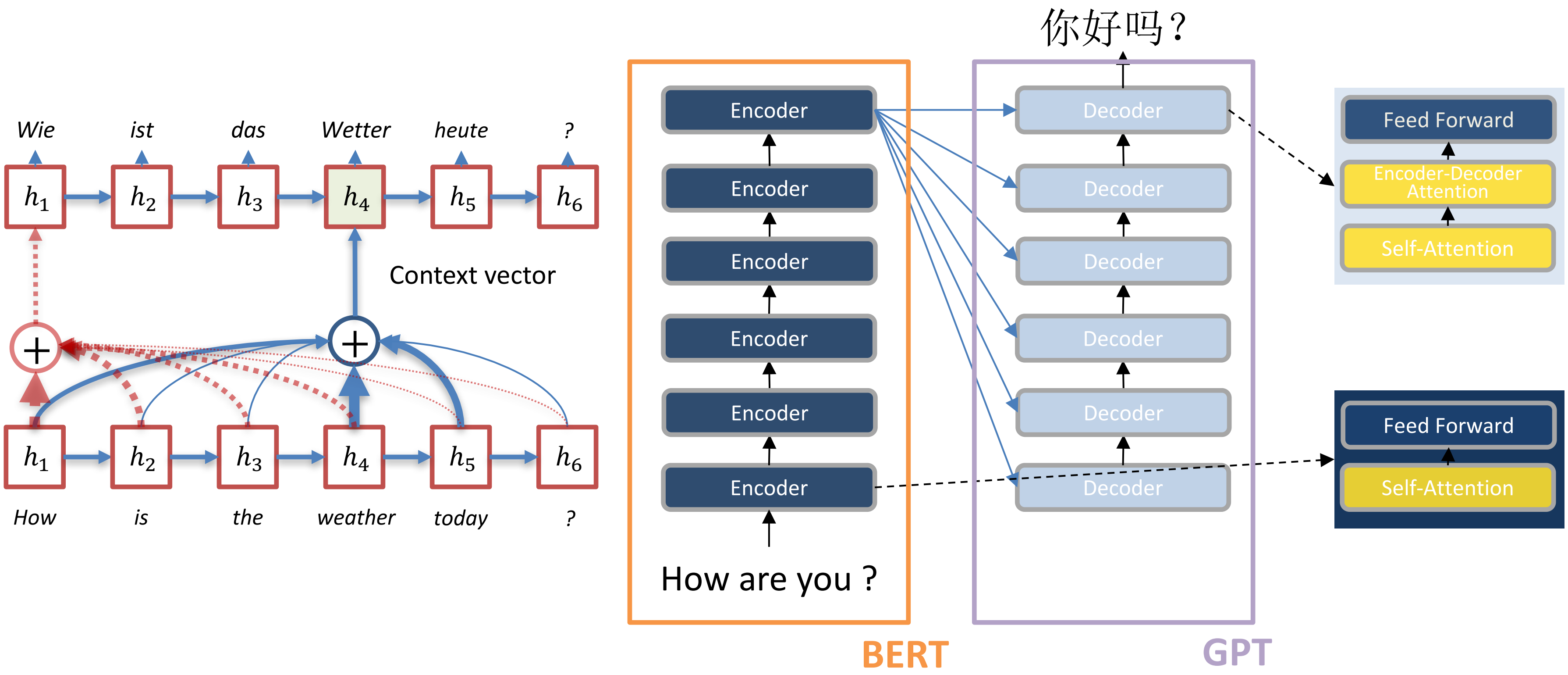
Master Project

- **summer term 2021**: *Accelerating Training and Inference of Large-scale Pre-trained Language Models*
- **summer term 2022**: *Developing Language Identification for Art-Historical Documents*
- **summer term 2023**: *Predicting Extreme Weather Events*

Master Thesis

- https://hpi.de/meinel/knowledge-tech/machine-learning-ai.html

Transformer, BERT and GPT



Vaswani, Ashish, et al. "Attention is all you need." NeurIPS 2017

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *NAACL-HLT 2019*

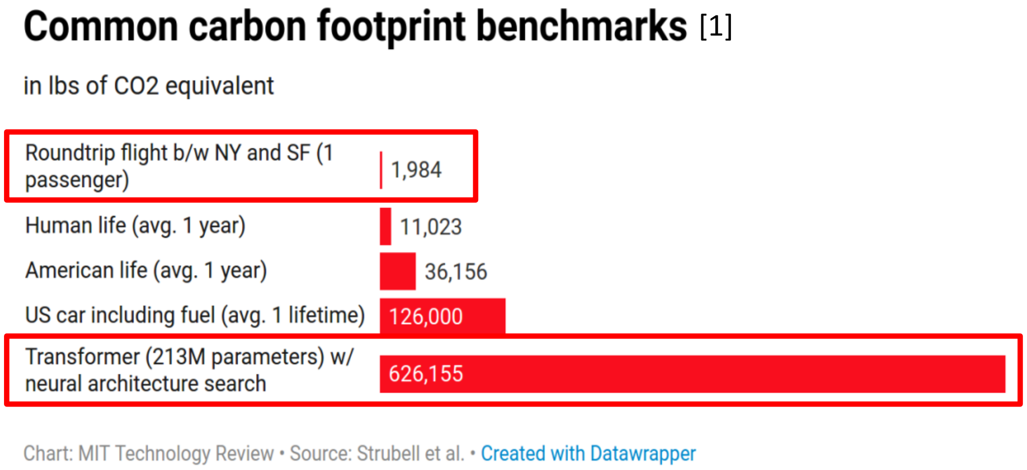
Applications of Transformer and BERT

- BERT is applied in Google search engine for 72 languages since December 2019. → **5.6 billion** searches per day
- Transformer is applied in Google translate → more than **100 billion** words a day
- About 60% Google’s TPU resources

Google Translate

Deep Learning Models **Spend Lots of Energy**

The impact of large-scale AI computing on the environment



Common carbon footprint benchmarks [1]

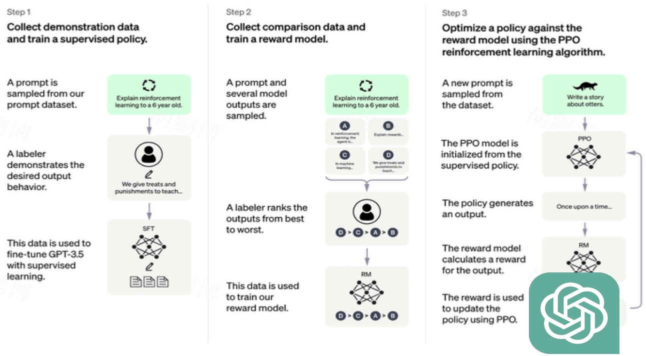
in lbs of CO2 equivalent

Chart: MIT Technology Review • Source: Strubell et al. • Created with Datawrapper

[1] Strubell, Emma, Ananya Ganesh, and Andrew McCallum. "Energy and Policy Considerations for Deep Learning in NLP." In the 57th Annual Meeting of the Association for Computational Linguistics (ACL). Florence, Italy. July 2019

LLMs

- GPT-3 (**175 billion** parameters): 1287MW, 552 tons [1]
  - 43 cars or 24 US families / year
  - > 10k V100 GPUs
- Google’s LaMDA (basis for “Bard”) used 1024 TPU-v3 for 58 days, creates 205 tons
- GPT-4:
  - > 25k GPUs
- GPT-5 ??



Step 1

Collect demonstration data and train a supervised policy.

A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3.5 with supervised learning.

Step 2

Collect comparison data and train a reward model.

A prompt and several model outputs are sampled.

A labeler ranks the outputs from best to worst.

This data is used to train our reward model.

Step 3

Optimize a policy against the reward model using the PPO reinforcement learning algorithm.

A new prompt is sampled from the dataset.

The PPO model is initialized from the supervised policy.

The policy generates an output.

The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.

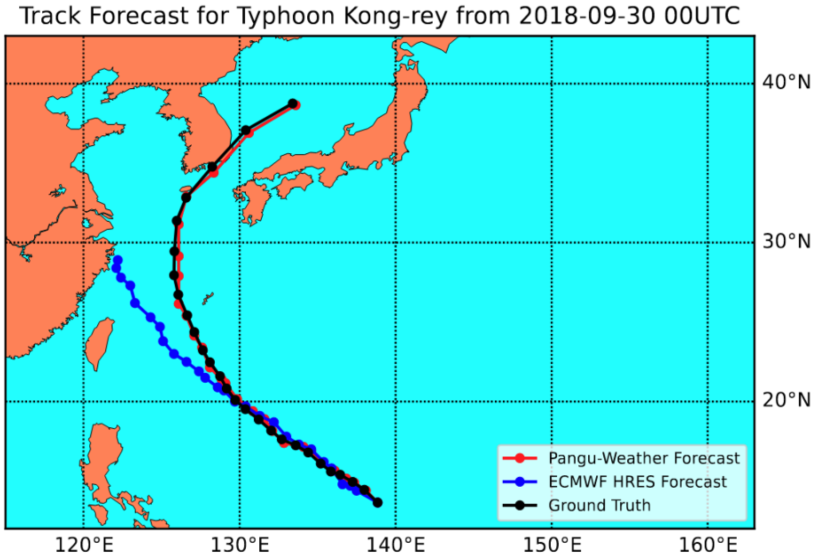
Explain reinforcement learning to a 6 year old.
We give treats and punishments to teach...
SFT
Explain reinforcement learning to a 6 year old.
A
In reinforcement learning, the agent is...
B
Explain rewards...
C
In machine learning...
D
We give treats and punishments to teach...
D > C > A > B
RM
D > C > A > B
Write a story about otters.
PPO
Once upon a time...
RM

[1] David Patterson et al., “Carbon emissions and large neural network training”, Google Research, April 2021

What is the ecological challenge?

Is the typhoon going to hit an inhabited coastal area?

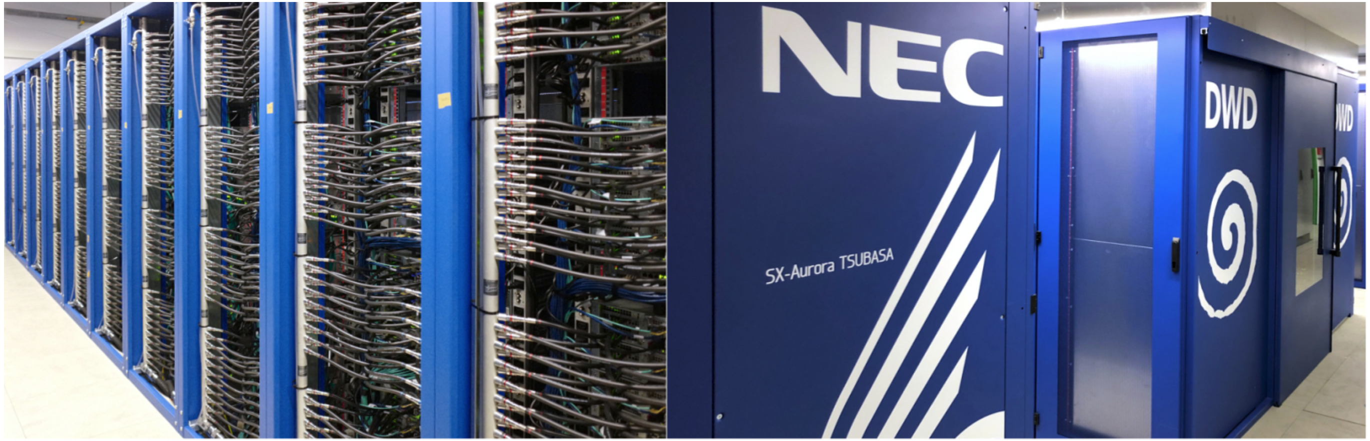
Where and when should precautionary measures be taken?



“Pangu-Weather: A 3D High-Resolution Model for Fast and Accurate Global Weather Forecast”.
Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, Qi Tian

Ideas: What concrete contributions can AI make?

Classical Weather Models are run on Supercomputers



The combined processing power of the Deutscher Wetterdienst's supercomputers is more than **10 petaflops**, which is more than **10,000 times faster** than an average desktop computer.

"Datenverarbeitung, DMRZ".
https://www.dwd.de/DE/derdwd/it/_functions/Teasergroup/datenverarbeitung.html

Topic 1: Weather Data Compression

- **ERA5 Data**:
  - European Centre for Medium-Range Weather Forecasts (ECMWF)
  - Variety of meteorological variables, such as temperature, wind speed, geopotential, etc.
  - 30km grid (spatial resolution of 0.25° × 0.25°)
  - Hourly data on a global scale updated every day since 1979

- **Data Size Example:**
  - 1 ERA5 variable, 1979-2018, spatial resolution of 2.8125° × 2.8125°, hourly scale → **10.7GB**

Daily mean temperature for January 2016 from ERA5

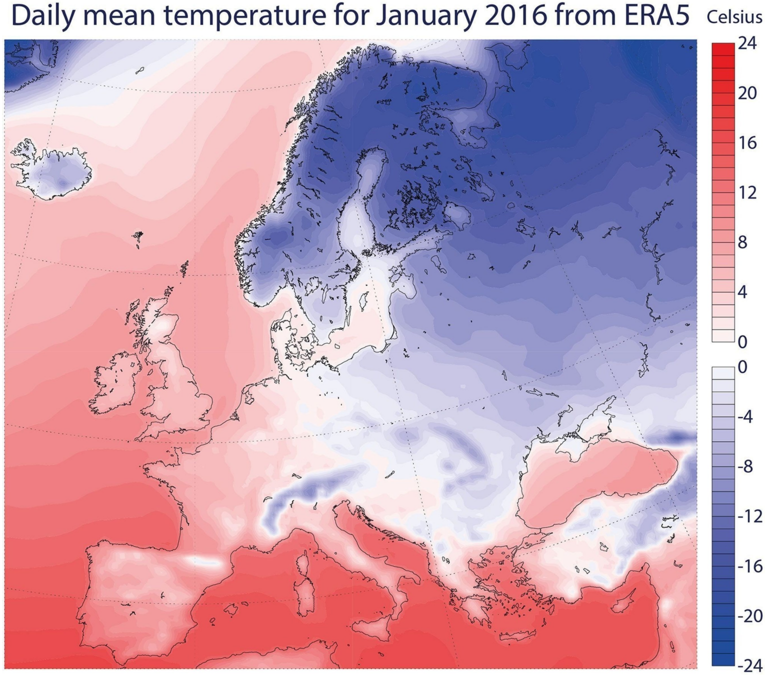
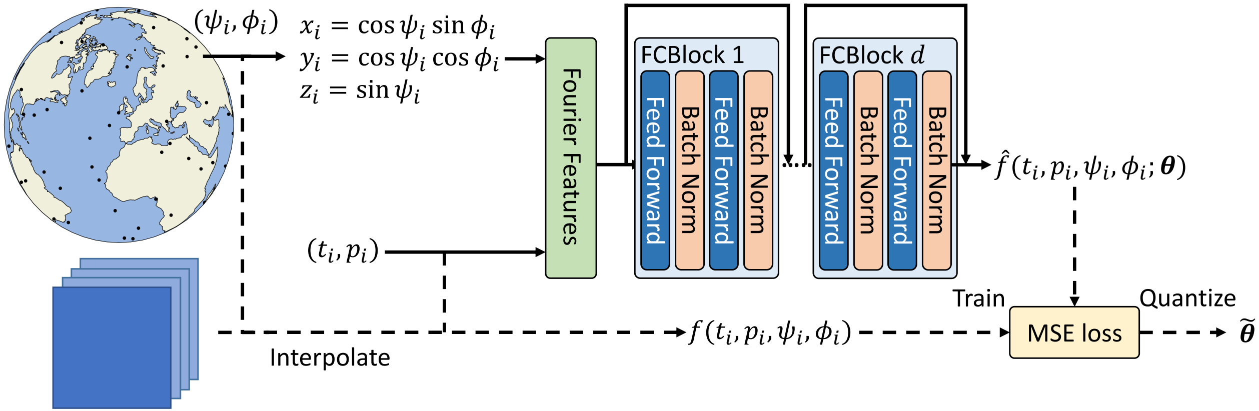


Image: http://photos.prnewswire.com/prnh/20161102/435664

Topic 1: Weather Data Compression

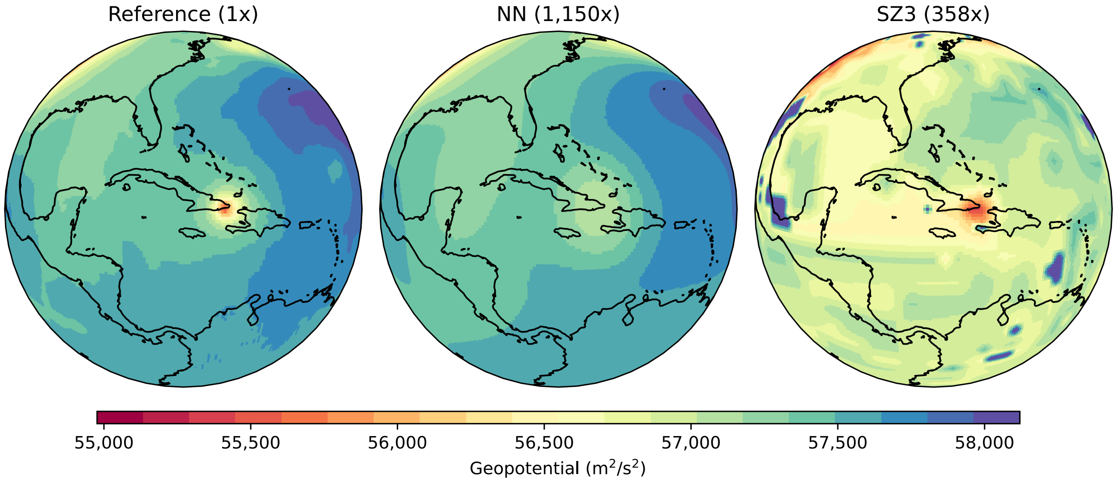
- **Neural Network for Data Compression**:
  - Lossy compression with 300× to more than 3,000× in compression ratios with better quality than the state-of-the-art SZ3 compressor
  - Compression (training) and decompression (inference)



[1] Langwen and Hoefler. "Compressing multidimensional weather and climate data into neural networks." arXiv preprint arXiv:2210.12538 (2022).
[2] https://github.com/spcl/NNCompression

Topic 1: Weather Data Compression

- **Geopotential at 500hPa on Oct 5th 2016 with hurricane Matthew in the center:**
  - Data extracted from a dataset with 15.6 GB → compressed to 13.86 MB
  - Compression ratio 1150x



[1] Langwen and Hoefler. "Compressing multidimensional weather and climate data into neural networks." arXiv preprint arXiv:2210.12538 (2022).
[2] https://github.com/spcl/NNCompression

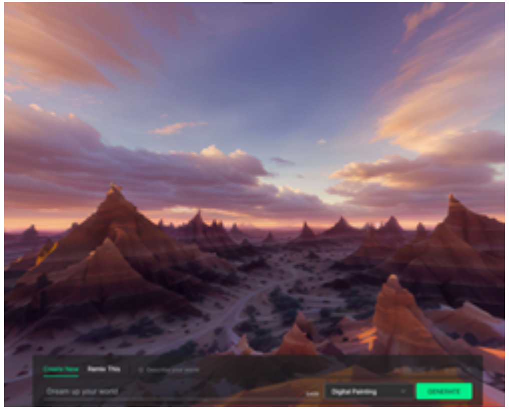
Topic 1: Weather Data Compression

- **Challenge**: Preserve Extreme Weather Events in the AI Model
- **Research Questions:**
  - How can we measure the quality of the compressed data?
    Is there something like an accuracy to evaluate?
  - Which compression rate is a sweet spot between compression rate and data quality?
  - Can the compression method be enhanced for our use case of predicting extreme weather events?

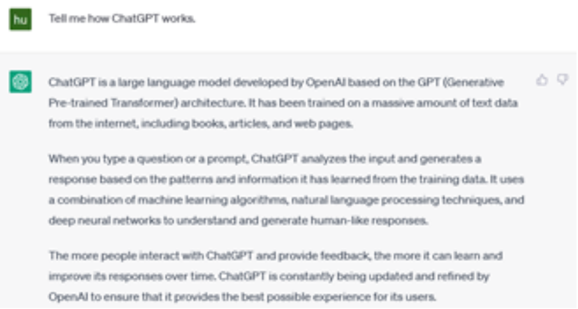
Topic 2: Build Your Games using Generative AI tools

- Generative AI technology has facilitated the development of many creativity-related industries.
- AI copilots like GPT-4 allow people to build specific applications even without much prior knowledge.

Text-to-Image
https://www.midjourney.com



Text-to-360-degree world
https://www.blockadelabs.com/



Tell me how ChatGPT works.

ChatGPT is a large language model developed by OpenAI based on the GPT (Generative Pre-trained Transformer) architecture. It has been trained on a massive amount of text data from the internet, including books, articles, and web pages.

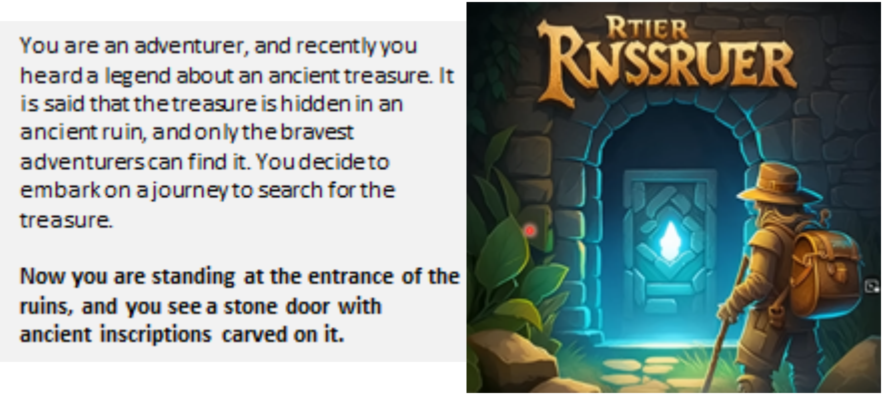
When you type a question or a prompt, ChatGPT analyzes the input and generates a response based on the patterns and information it has learned from the training data. It uses a combination of machine learning algorithms, natural language processing techniques, and deep neural networks to understand and generate human-like responses.

The more people interact with ChatGPT and provide feedback, the more it can learn and improve its responses over time. ChatGPT is constantly being updated and refined by OpenAI to ensure that it provides the best possible experience for its users.

Text-to-Text
https://chat.openai.com/

Topic 2: Build Your Games using Generative AI tools

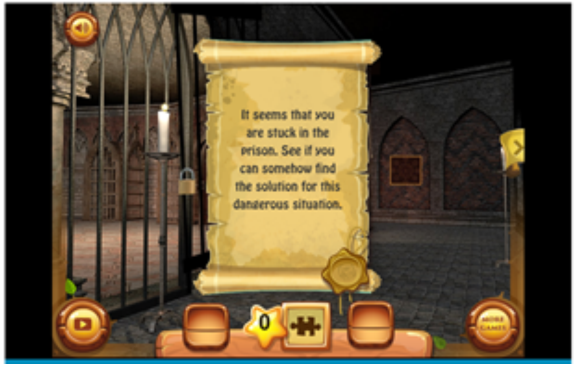
- Start from vivid text-based adventure games(e.g., ChatGPT and Midjourney).
- Build Escape games: generate scripts, 360-degree scenes, and objects → build your game (e.g., Unity).
- How these tools could benefit the game development and design?



You are an adventurer, and recently you heard a legend about an ancient treasure. It is said that the treasure is hidden in an ancient ruin, and only the bravest adventurers can find it. You decide to embark on a journey to search for the treasure.

Now you are standing at the entrance of the ruins, and you see a stone door with ancient inscriptions carved on it.

Text adventure
https://www.youtube.com/watch?v=A-6c584jxX8

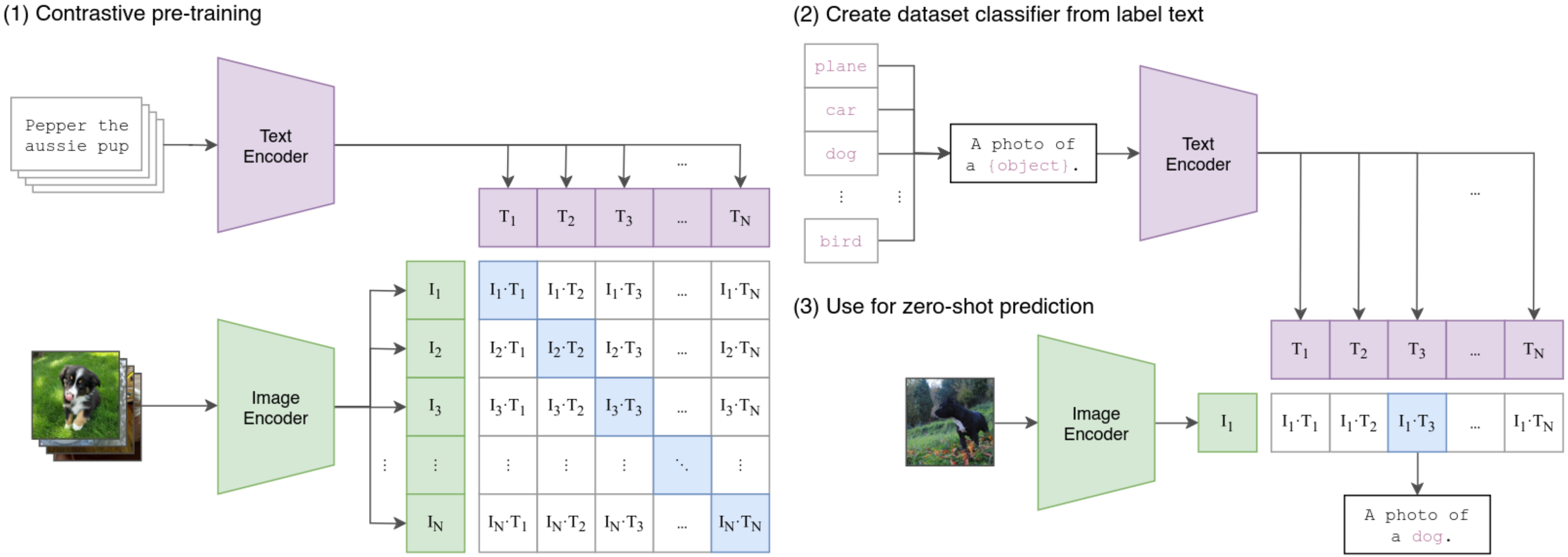


It seems that you are stuck in the prison. See if you can somehow find the solution for this dangerous situation.

Prison escape

- Challenges: prompt engineering, some knowledge in using game development software.

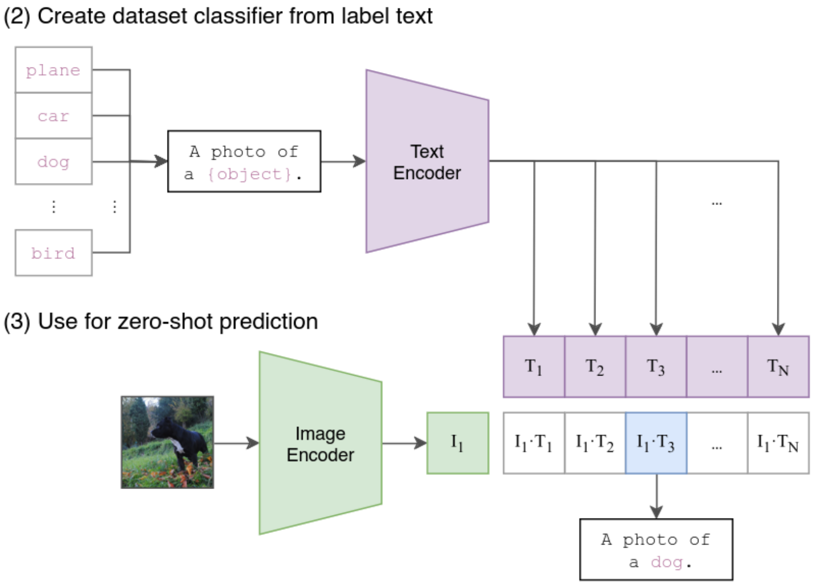
Topic 3: Class Discovery using Language-Image Pretrainings

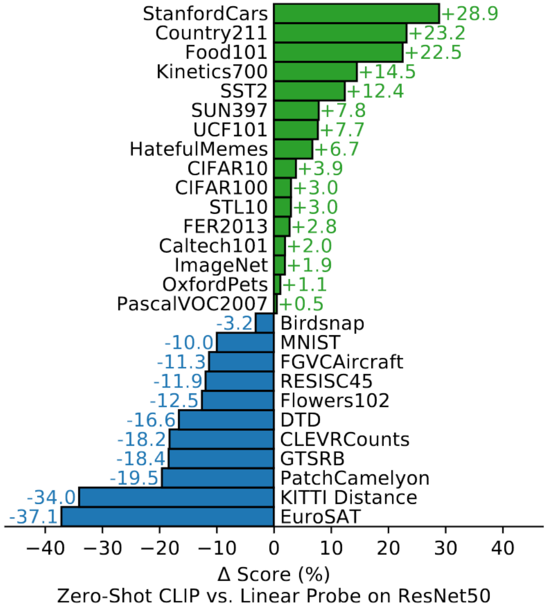


	T_1	T_2	T_3	...	T_N
I_1	$I_1 \cdot T_1$	$I_1 \cdot T_2$	$I_1 \cdot T_3$	...	$I_1 \cdot T_N$
I_2	$I_2 \cdot T_1$	$I_2 \cdot T_2$	$I_2 \cdot T_3$	...	$I_2 \cdot T_N$
I_3	$I_3 \cdot T_1$	$I_3 \cdot T_2$	$I_3 \cdot T_3$	...	$I_3 \cdot T_N$
⋮	⋮	⋮	⋮	⋱	⋮
I_N	$I_N \cdot T_1$	$I_N \cdot T_2$	$I_N \cdot T_3$	...	$I_N \cdot T_N$

Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *International conference on machine learning*. PMLR, 2021.

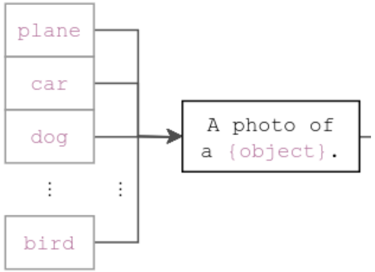
Topic 3: Class Discovery using Language-Image Pretrainings





Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *International conference on machine learning*. PMLR, 2021.

Topic 3: Class Discovery using Language-Image Pretrainings



Research Question: What if we don't know the class names?

- Given only images, can we still use CLIP for classification?
- Can we maybe infer the correct prompts for the images?
- How does this approach measure against existing clustering / class discovery methods?

Tools and Hardware

- Deep learning framework – **PyTorch**
- GPU servers from MML Group
    - **Prerequisites**
        - introductory moodle course (**required**)
        - cluster usage document

PYTORCH

Grading Policy

The final evaluation will be based on:

idea presentation and Initial implementation,	**10%** (22.05.2023)
Final presentation,	**20%** (24.07.2023)
Report, 12-18 pages (latex)	**30%** (31.08.2023)
Code,	**40%** (31.08.2023)
Participation in the seminar	(**bonus**)
Grading	(30.09.2023)

Enrollment/Anmelden

Registration

- HPI students: HPI-moodle
- Other UP students: Email to HPI-Studienreferat (Studienreferat(at)hpi.uni-potsdam.de)

- until **24.04.2023**, inform your **preferred and secondary topics** by email
  - Send email to**: mml-team@hpi.de**
- **28.04.2023:** Announcement of group assignment
- **Individual weekly meeting** with teaching team

Contact

Email: {ting.hu, gregor.nickel, jona.otholt, haojin.yang}@hpi.de

Office: G2-E.31, G2-E.32, G2-E.26

Thank you for your Attention!

Address:

Hasso-Plattner-Institut für Digital Engineering gGmbH, Prof.-Dr.-Helmert-Str. 2-3 D-14482 Potsdam, Germany

Email: haojin.yang@hpi.de

Web: