Generative Models

Dr. Gerhard Paaß
Fraunhofer Institute for Intelligent Analysis and Information Systems (IAIS)
Sankt Augustin

TensorFlow

Fraunhofer
BIG DATA AI

Course Overview

1. Intro to Deep Learning	Recent successes, Machine Learning, Deep Learning & types
2. Intro to Tensorflow	Basics of Tensorflow, logistic regression
3. Building Blocks of Deep Learning	Steps in Deep Learning, basic components
4. Unsupervised Learning	Embeddings for meaning representation, Word2Vec, BERT
5. Image Recognition	Analyze Images: CNN, Vision Transformer
6. Generating Text Sequences	Text Sequences: Predict new words, RNN, GPT
7. Sequence-to-Sequence and Dialog Models	Transformer Translator and Dialog models
8. Reinforcement Learning for Control	Games and Robots: Multistep control
9. Generative Models	Generate new images: GAN and Large Language Models

: link to background material, : link to images used in lecture, G. : Terms that may be asked in the exam

Generative Models

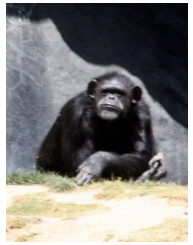
Agenda

1. Generative Adversarial Models
2. Text to Image Generation with Language Models
3. Foundation Models
4. Summary

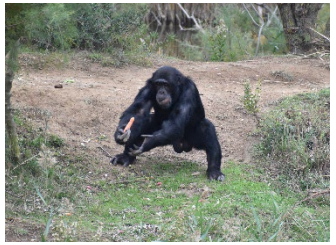
Generative Models

- Given training data, generate new samples from same distribution
  - 18 Impressive Applications of Generative Adversarial Networks (GANs)

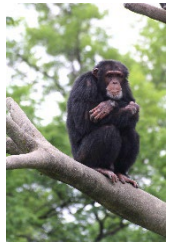
"chimpanz" by BulkyWebEU / CC BY-SA 2.0



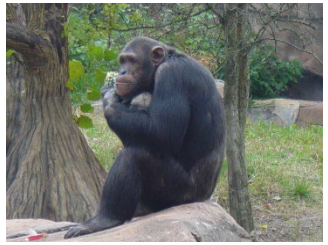
"Chimpanze" by Just chaos / CC BY 2.0



"File:Reserve Sigean - Chimpanze 09.jpg" by Tylwyth Eldar / CC BY-SA 4.0



"I Has a Tude!" by LadyDragonflyCC - >;< / CC BY 2.0



"Hmmmmm" by fairyfroggie / CC BY-ND 2.0

Training data ~ $p_{data}(x)$

Generated samples ~ $p_{model}(x)$

Want to learn $p_{model}(x)$ similar to $p_{data}(x)$

- Addresses **density estimation**, a core problem in unsupervised learning
- Several flavors:
  - Explicit density estimation: explicitly define and solve for $p_{model}(x)$
  - Implicit density estimation: learn model that can sample from $p_{model}(x)$ w/o explicitly defining it

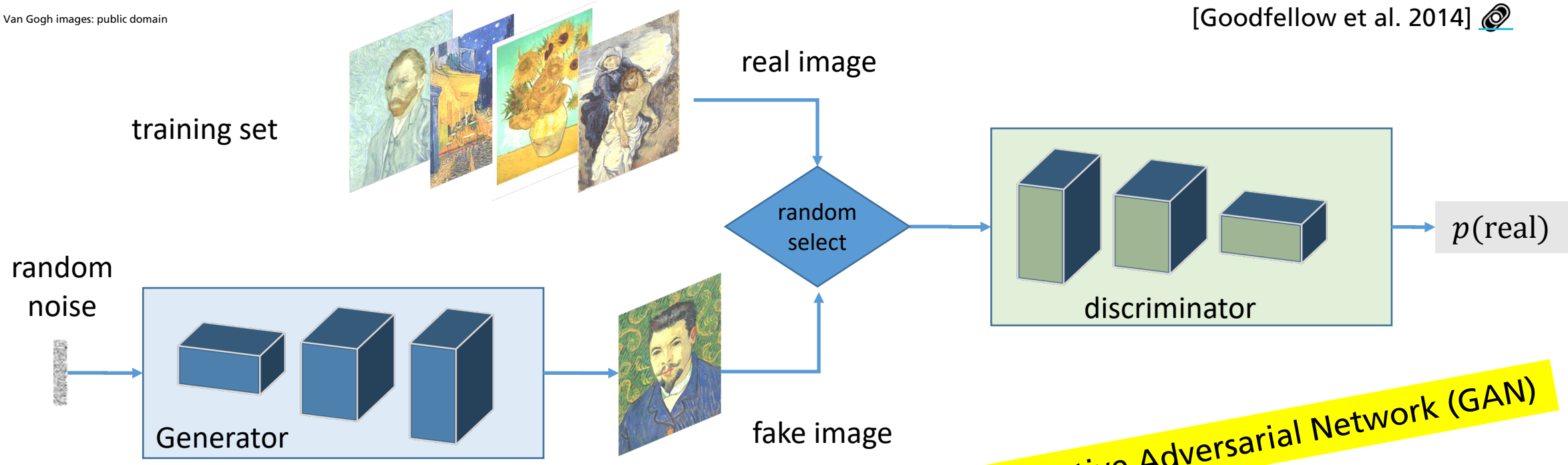
G. machine translation

Generating Images

- It is very difficult to provide a probability distribution of images of van Gogh
  - which colours are appropriate, which strokes are allowed, etc.
- It is much easier to provide a classifier which can distinguish images of van Gogh from other images
- Train an image generator
  ➔ such that it can produce images which are accepted by a van Gogh image classifier

Van Gogh images: public domain

[Goodfellow et al. 2014]

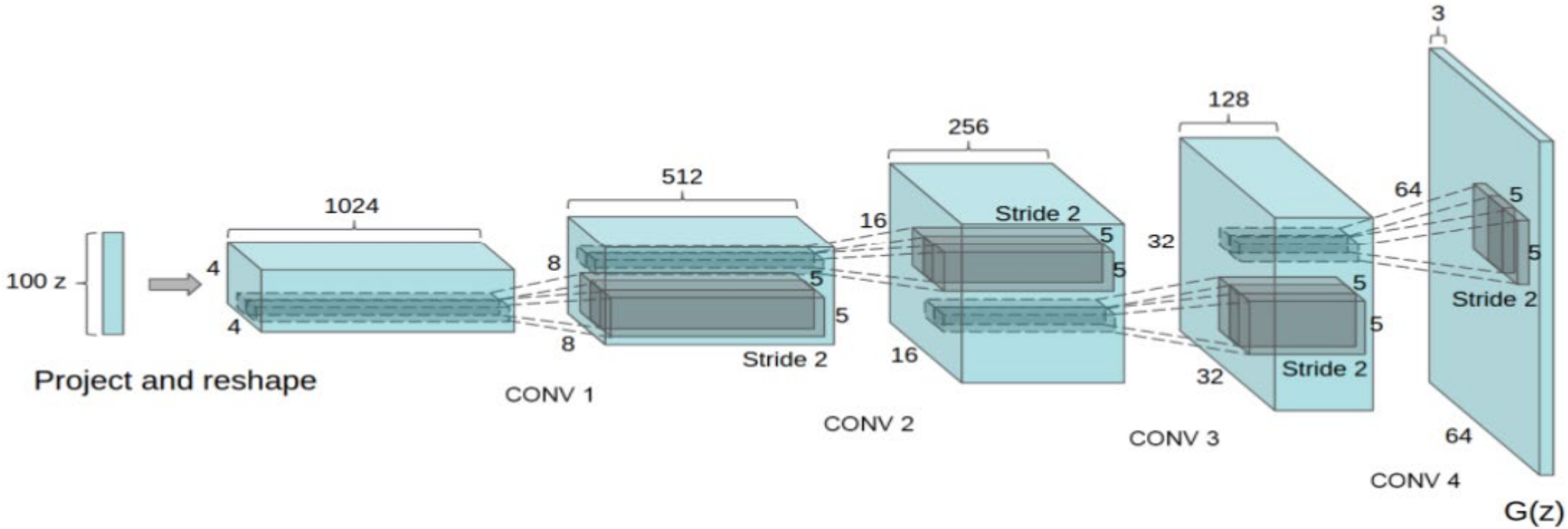


Generative Adversarial Network (GAN)

G. generative adversarial network G. GAN generator G. GAN discriminator

Example: Generator Architecture

- Input is a uniformly distributed input vector z (e.g. of length 100)
- This is projected to a higher spatial resolution by transposed convolution. [Radford et al. 2016, p.4]



Training a GAN

- The discriminator network may be some CNN for image classification
- The discriminator and the generator are **trained in turn**
- **Discriminator** tries to increase the probability of real images and reduce probability of fakes
  - can only change w_d

$$J_D(w_d) = \left(D_{w_d}(b_1) * \cdots * D_{w_d}(b_n)\right)^{1/n} * \left(\left(1 - D_{w_d}(c_1)\right) * \cdots * \left(1 - D_{w_d}(c_m)\right)\right)^{1/m}$$

probability 1st image real

probability all n images real

probability 1st fake not real

probability all m fakes not real

- **Generator** tries to increase probability that fakes are classified as real images
  - can only change w_g

$$J_G(w_g) = \left(D_{w_d}\left(G_{w_g}(x_1)\right) * \cdots * D_{w_d}\left(G_{w_g}(x_m)\right)\right)^{1/m}$$

probability 1. fake is real

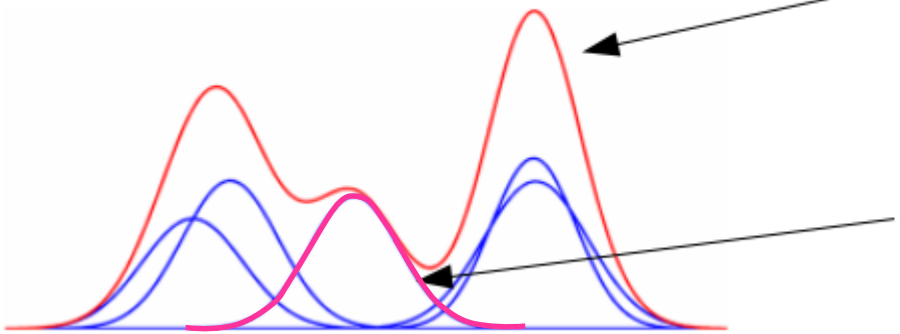
probability all m fakes are real

[Goodfellow et al. 2014]

G. GAN optimization criterion

Training a GAN

- GANs are very hard to train: unstable
- Generator may learn just a small sub-distribution and produces good examples from that
- Rest of the “true” distribution ignored: **Mode collapse**



The generator may only learn this small part of the distribution.

➔ produces valid samples, but has too little variation

Evolution of GAN Quality

2014

2015

2016

2017

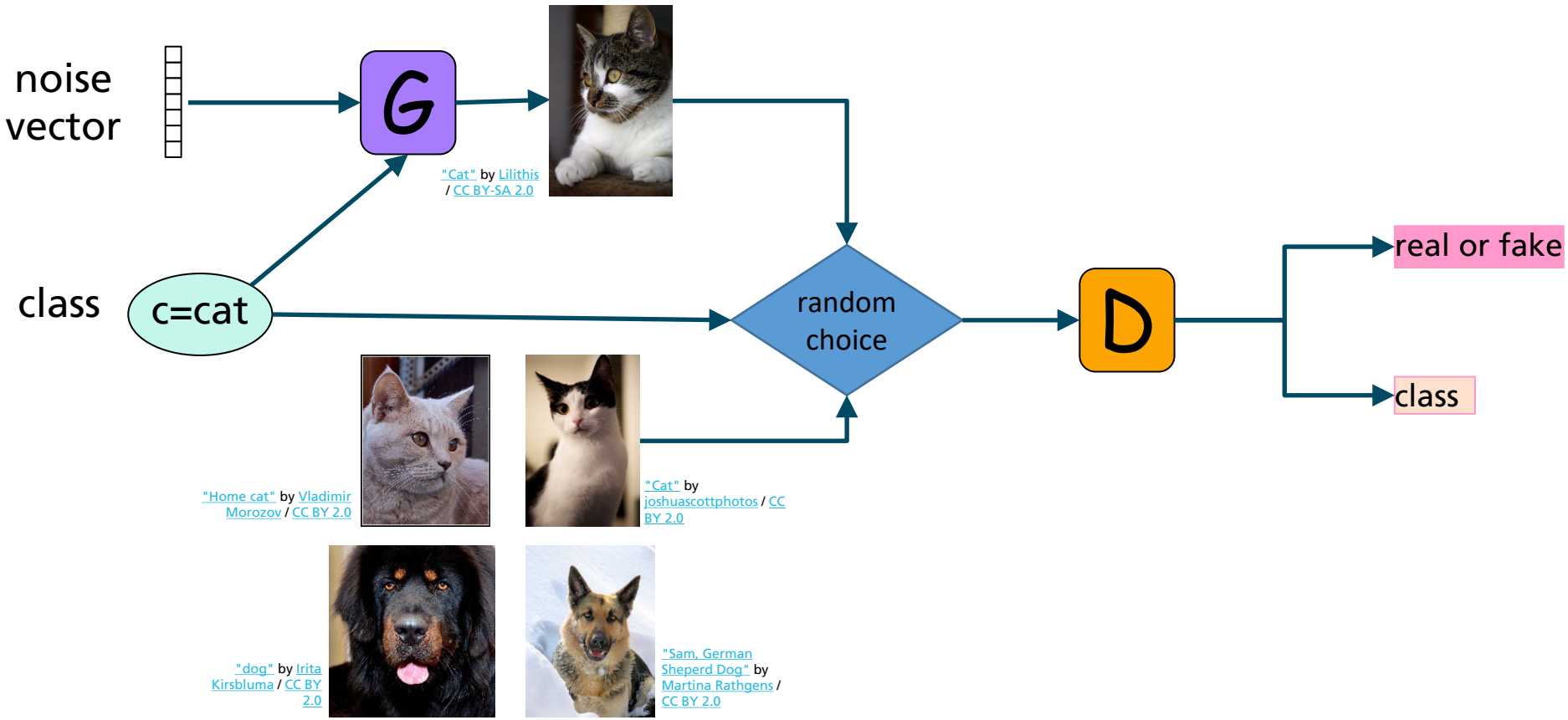
https://www.thispersondoesnotexist.com/

All images public domain

What is a generated Image?

Supervised GANs

- We may train a generator such that it is able to generate images of specific classes.
- It can exploit the common properties of all images during training.



[Isola et al. 2017]

GAN Applications

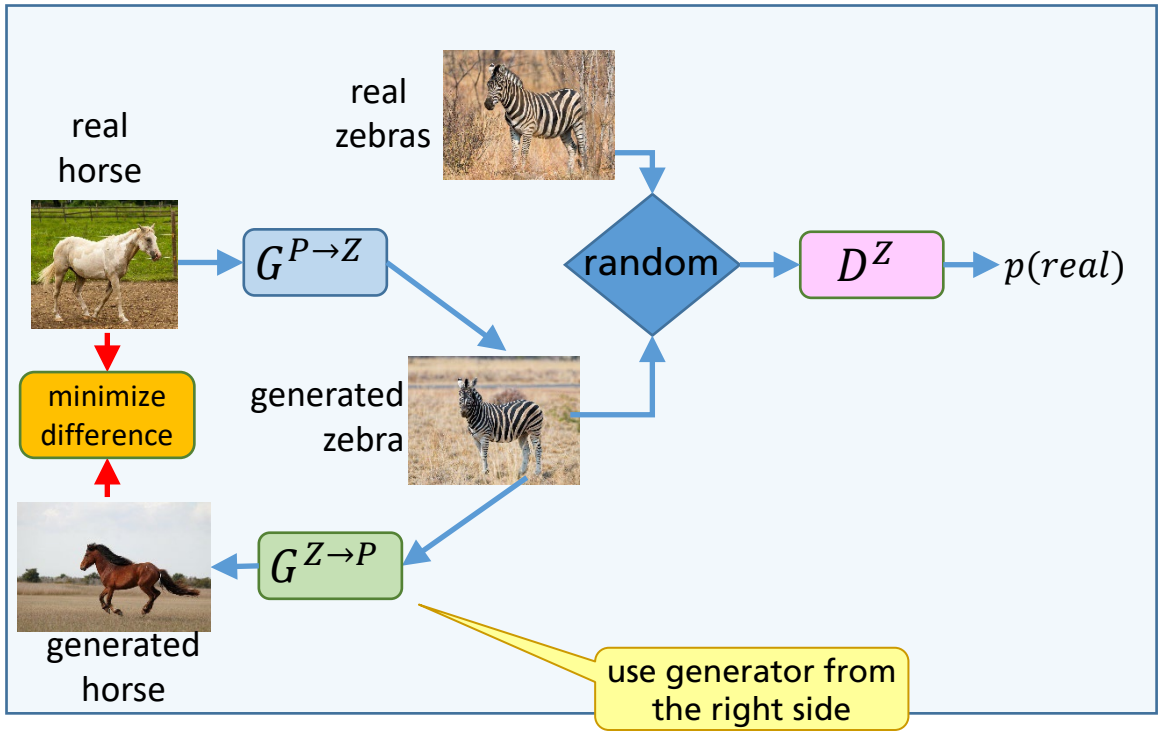
- Generating synthetic training data (to train other nets)
- Unsupervised Feature Extraction
- Generating adversarial test samples (security validation)
- Super-resolution (of images)
- Style transfer: transfer an image to a style of a painter

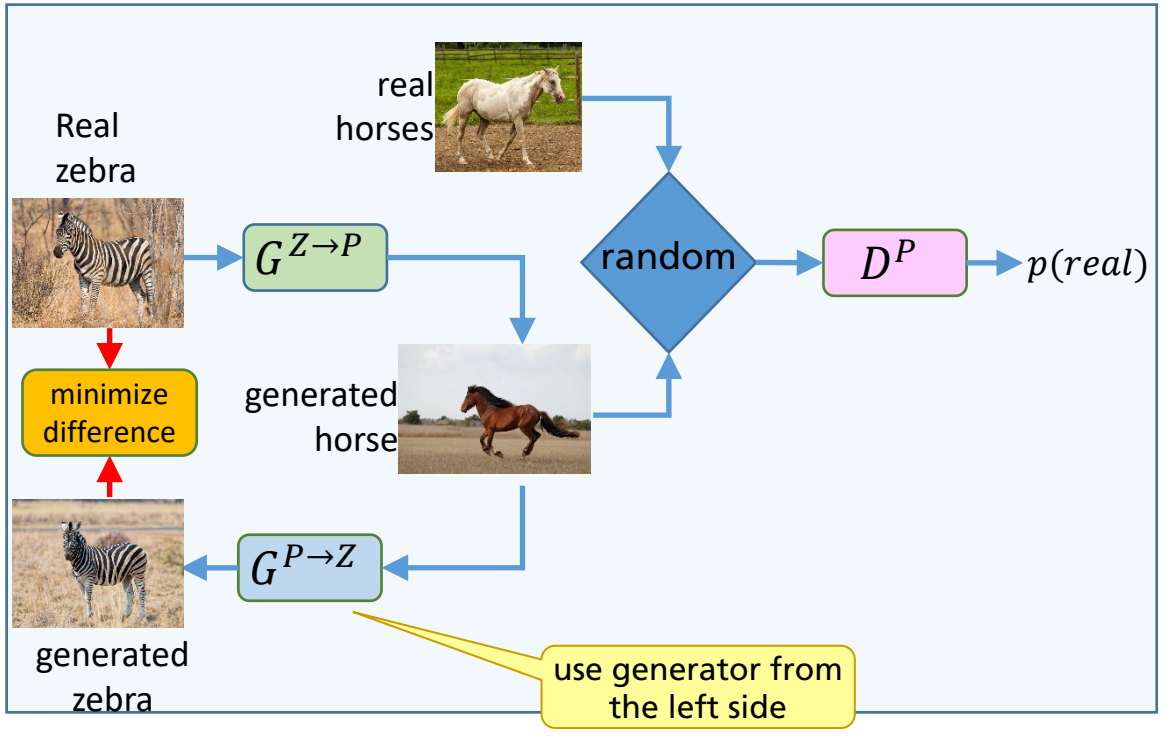
- Important: implicit loss function
- Important:
  - Latent Space: low dimensional vectors represent images,
    similar images have similar vectors
  - Latent Space manipulations: interpolation between different images by interpolation vectors

G. GAN applications

Computations in the Latent Space: CycleGAN

- If you want to transform a horse to a zebra
  ➔ no training data, as there is not horse with zebra stripes
- You train two generator networks $G^{P \to Z}$ and $G^{Z \to P}$ for transforming horses to zebra and vice versa
- After you transform the zebra back to horse minimize the difference [Zhu et al. 2017]





Real Zebra: "Zebra posing in the savannah" by Tambako the Jaguar / CC BY-ND 2.0

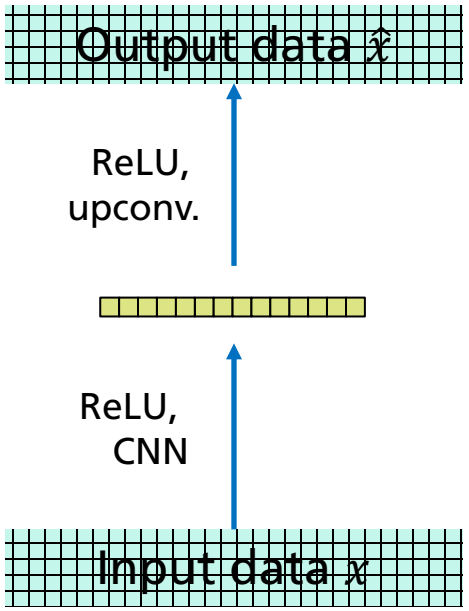
Generated zebra: "Zebra in the high dry grasses" by Tambako the Jaguar / CC BY-ND 2.0

Real horse: "Horse Fences" by Me in ME / CC BY 2.0

Generated horse: "Wild Horses" by firelizard5 / CC BY 2.0

Autoencoder

- Target: generate a compressed representation of some input: e.g. image
- Learn mapping to short "bottleneck" vector such that input can be reconstructed
- Loss function: output should be equal to input: minimize square distance $\|x - \hat{x}\|^2$



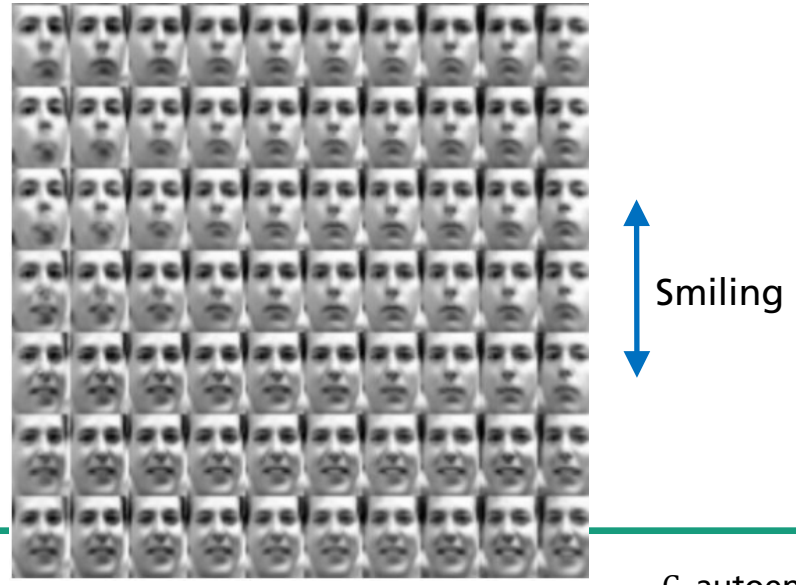
"Children Reading Pratham Books and Akshara" by Pratham Books / CC BY 2.0

compressed representation

"Children Reading Pratham Books and Akshara" by Pratham Books / CC BY 2.0

Variational Autoencoder

- Additional restriction that h follows a multivariate Normal density $h \sim N(0, I)$
- Then we can generate new data along dimensions of h



[Kingma & Welling, ICLR 2014]

G. autoencoder

Generative Models

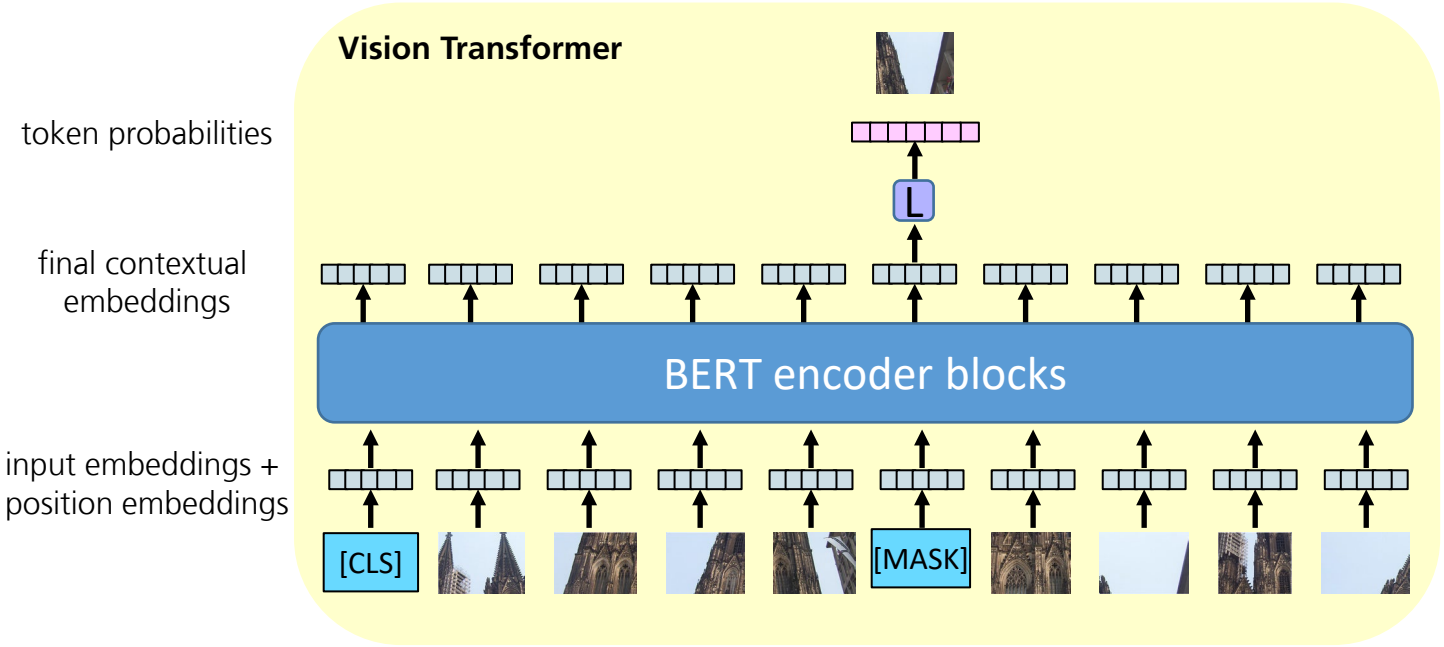
Agenda

1. Generative Adversarial Models
2. Text to Image Generation with Language Models
3. Foundation Models
4. Summary

Vision Transformer

Recap

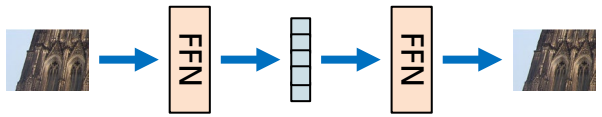
- Partition image into image tokens
  ➔ e.g. 16x16 pixel **image patches**
- Assign an embedding to each token
- Apply BERT encoder blocks self-attention creates contextual embeddings
- Predict masked tokens
- Model the relation between image tokens
  - can fill in missing tokens
  - solve additional tasks by fine-tuning: e.g.object classification
  - Vision Transformer [Dosovitskiy et al. 2020]



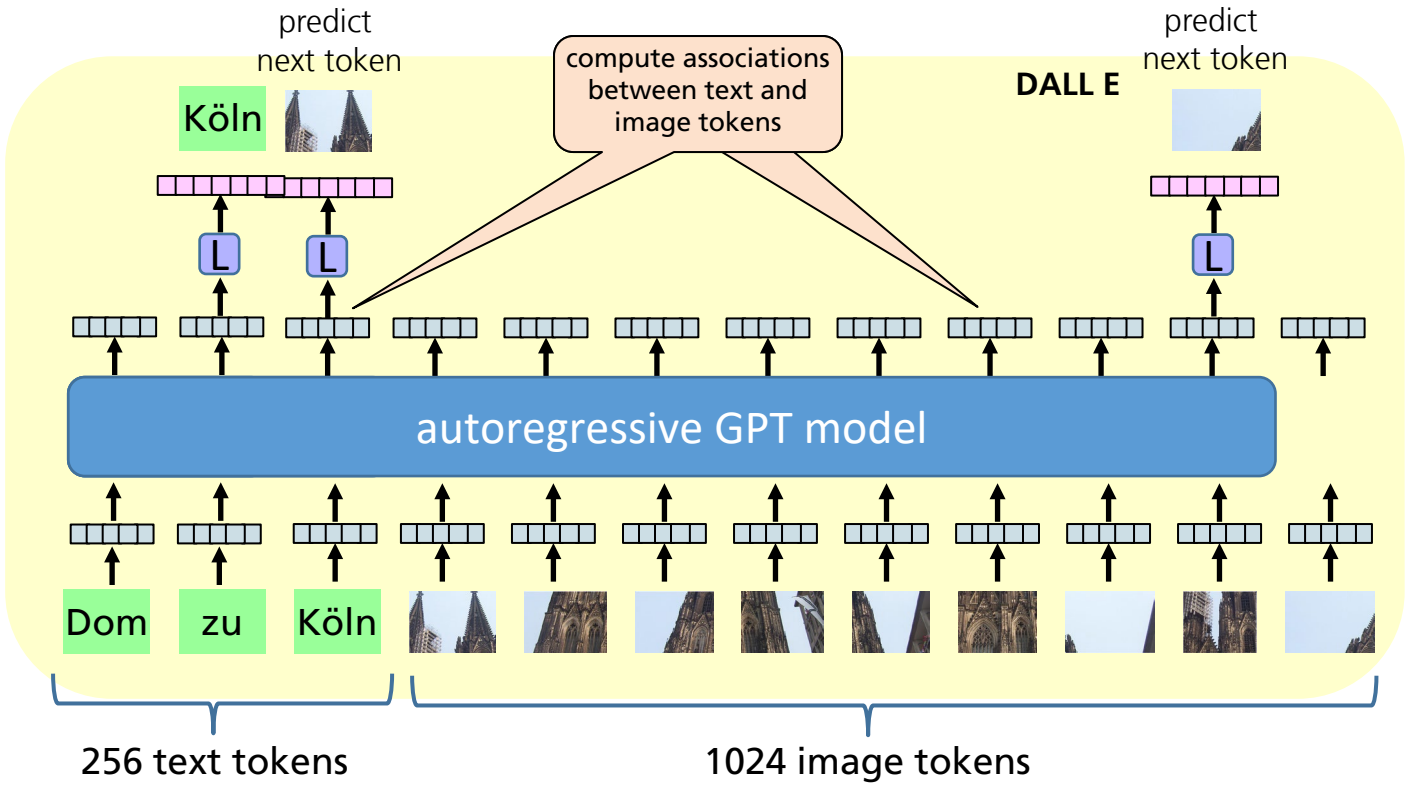
Kölner Dom by Helder da Rocha CC BY-SA 2.0

DALL E 1 for Image Generation from Text

- Combine text and image tokens in a single GPT model [Ramesh et al. 2021]
- Partition a 256x256 image into 1024 image patches of size 8x8
- Convert text to 256 tokens by byte pair encoding
- Use autoencoder for image patches



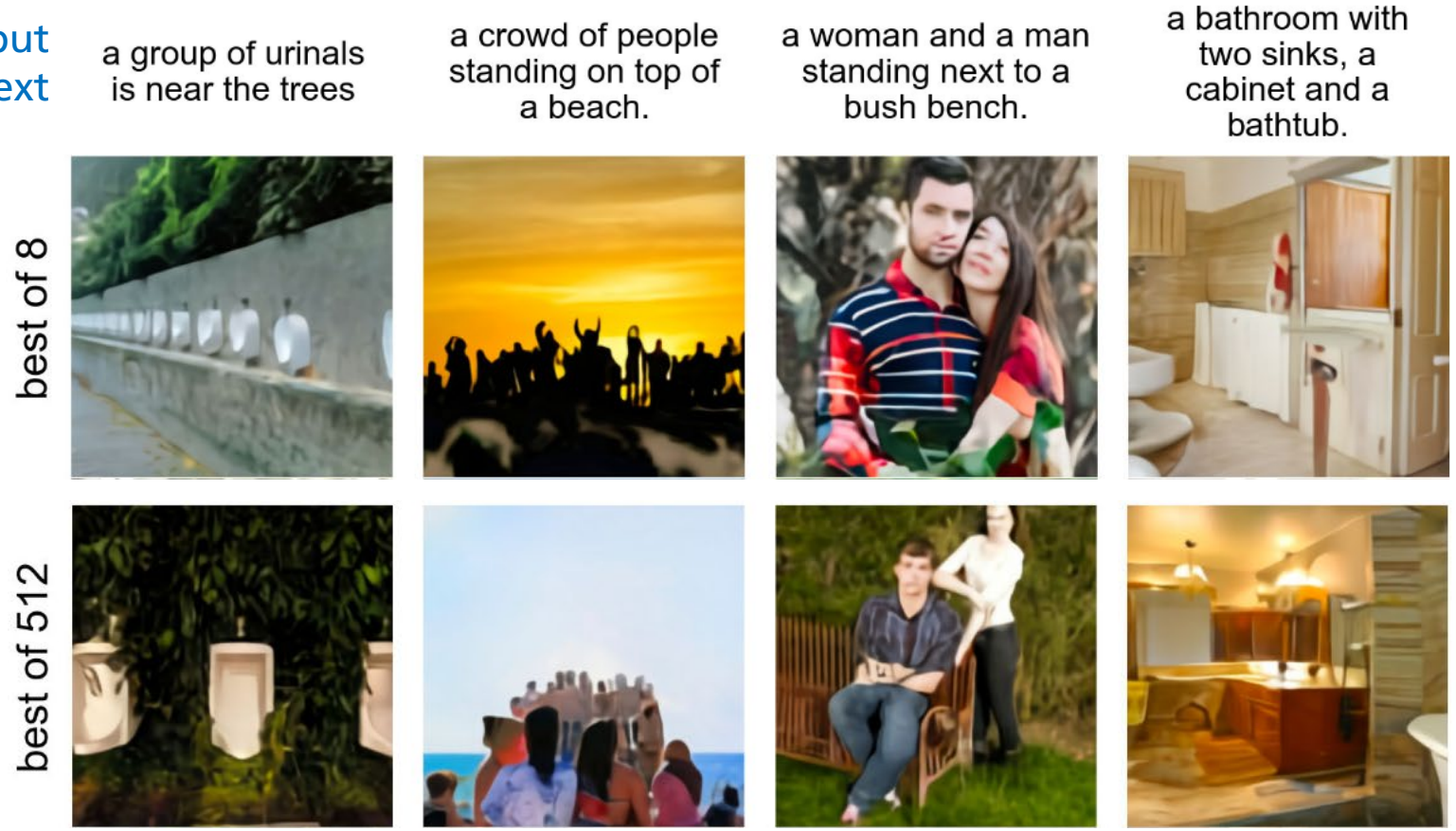
  - discretize to 8192 values



- GPT model with 64 self-attention layers and 12B parameters. Sparse attention: not all inputs are connected by attention
- Training data: 250M text-image pairs

Images Generated by DALL E

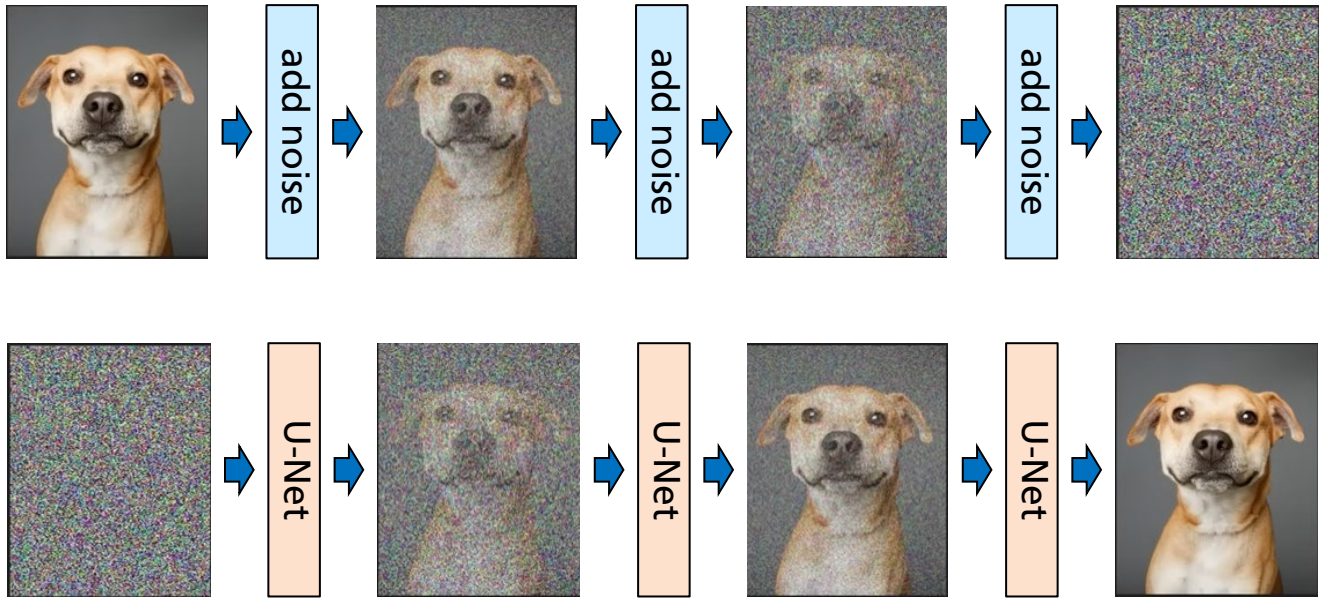
- GPT model receives a text as prompt and generates the corresponding image tokens
- The image tokens are transformed to image patches and assembled to an image
- Rank quality of generated images with an auxiliary model
- select the best
- Problem: resolution not satisfactory



6. Generating an image with a GPT model

Diffusion Model to Scale Images

- Diffusion: stepwise add noise to an image
  - Resulting image becomes random noise
- **Diffusion Model**: step-by-step invert noise by a neural network
  - Use same neural network for all steps, (e.g. $t = 150$), usually U-Net
  - train on sequences of degenerating images
- **Stable Diffusion**: [Rombach et al. 2022]
  Apply noise reconstruction to lower-dim. representation of image
  - Generate lowerdimensional representation by a VQGAN encoder
  - Use corresponding decoder to reach full image
- Including text information
  - use output of a language model as input to diffusion model
  - U-Net layers attend to the text tokens by cross-attention



Images Generated by Diffusion Models

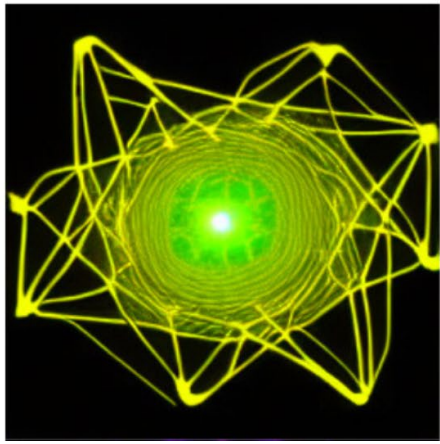
Stable Diffusion

- Trained on LAION 400M with 400 M text-image pairs [Rombach et al. 2022]
- Diffusion of latent representation:
  - much less computational effort and memory
- Generate images from text input
  can generate 1024x1024 images
  open source model free online access
- Inpainting:
  generate missing part of an image
- Many other models available (closed source)
  - Imagen by Google also uses diffusion model
  - DALL E 2 by OpenAI using diffusion model

A zombie in the style of picasso

An image of an animal half mouse half octopus

An illustration of a slightly conscious neural network



public use as computer generated

Most image generators use diffusion models

Generative Models

Agenda

1. Generative Adversarial Models
2. Text to Image Generation with Language Models
3. Foundation Models
4. Summary

Large Language Models can Process Other Media

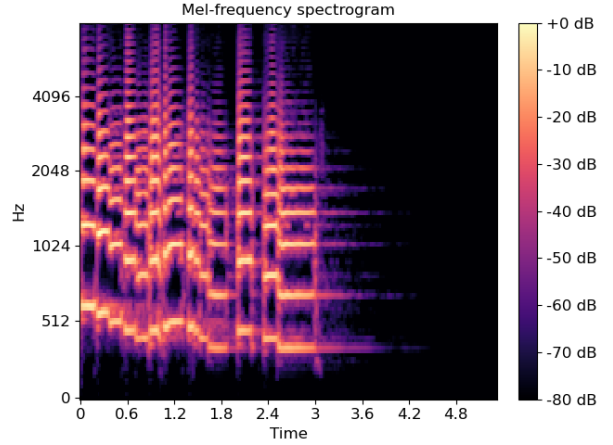
- Large Language Models are extremely good at learning language
  - Learn the syntax of language and generate text without grammatical errors
  - Learn commonsense facts and combine them to new mostly correct text
  - GPT-4 has similar abilities as human: pass legal exams, pass the Abitur in Bavaria
- Large Language Model needs tokens embeddings as input
  ➔ transform other media to tokens and process with Large Language Models
- Input tokens
  - **Image** patches of 16x16 pixels
  - **speech** frequency descriptions at different frequency bands for every 10 msec
  - Sequences of **video** patches
  - ngrams of **DNA** code sequences
  - sequences of state-action-reward in **control** problems
  - …

Huge Application Opportunities

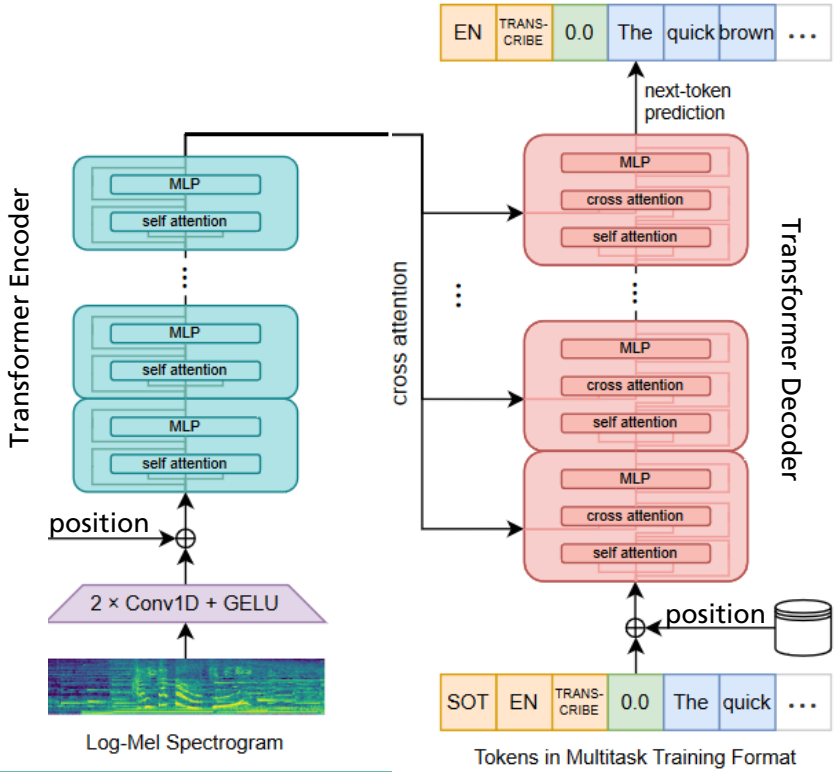
G. speech-to-text translation

Speech-To-Text Models based on LLM

- **Whisper** [Radford et al. 2022]
  - Transform the input speech signal to log mel spectrogram
  - mel scale takes into account how humans discriminate sound
  - for each 1 msec interval represent sound by 80 values
- **Transformer Encoder Decoder**
  - Translate log mel value to input embeddings by 2 CNN layers
  - Generate output tokens step by step
- Performance
  - clean data: same as current SOTA:
    Word error rate 2.5
  - much better on **noisy data**:
    55% improvement over current SOTA



Generated by script https://librosa.org/doc/main/generated/librosa.feature.melspectrogram.html

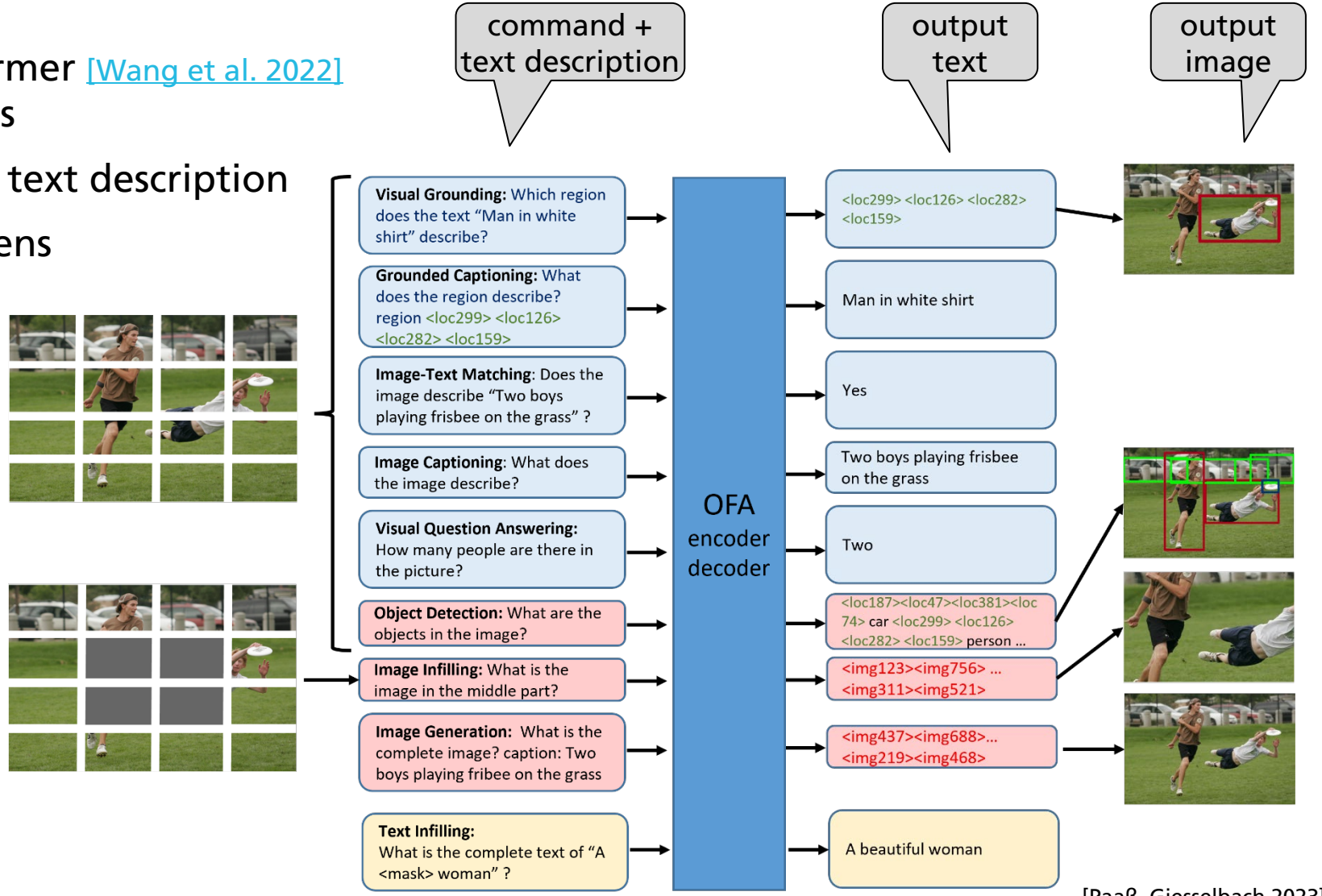


G speech-to-text translation

Multipurpose Image Models based on LLM

- OFA: encoder-decoder transformer [Wang et al. 2022]
  The same model for many tasks
  - Input: image + command + text description
  - convert image / text to tokens

- OFA commands + tasks
  - Visual Grounding
  - Grounded Captioning
  - Image-Text Matching
  - Image captioning
  - Visual Question Answering
  - object detection
  - Image infilling
  - Image generation
  - Text infilling

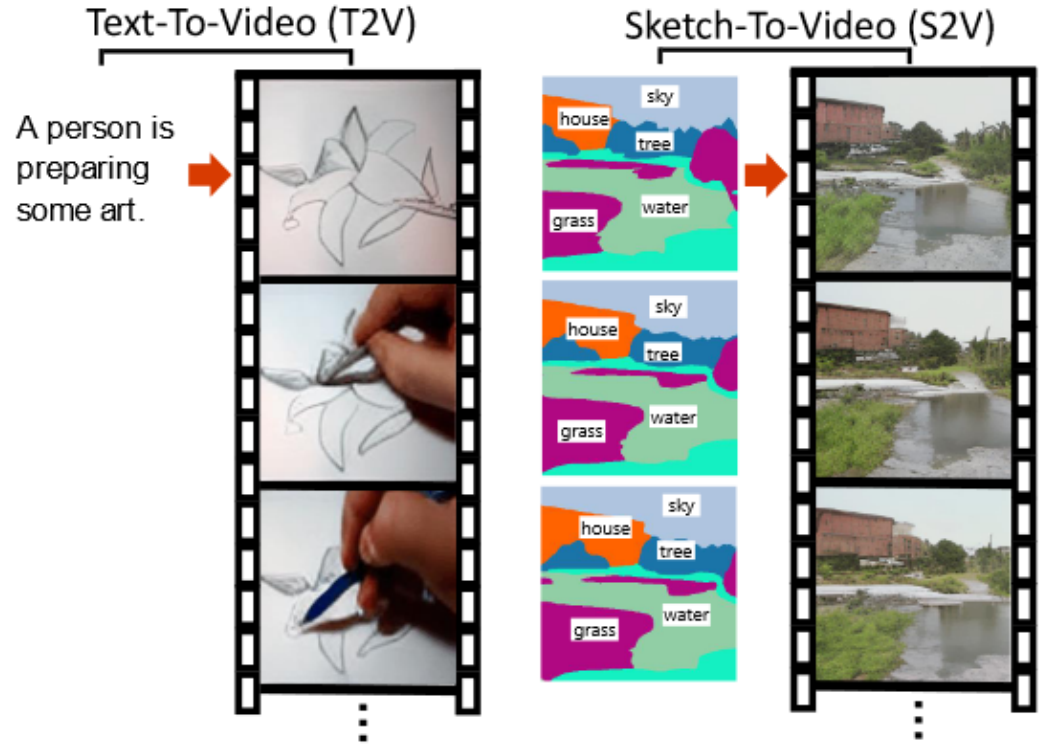


[Paaß, Giesselbach 2023]

G. Multipurpose model for image analysis

Video Model based on LLM

- **NÜWA**: encoder-decoder for video processing [Wu et al. 2022]
- sequences of image patches as tokens
  - Transformer encoder-decoder
  - text to image/video, image/video manipulation
- **Imagen video**: [Ho et al. 2022] cascaded diffusion model for better resolution
  - much better quality. link to animation



G. Multipurpose model video processing

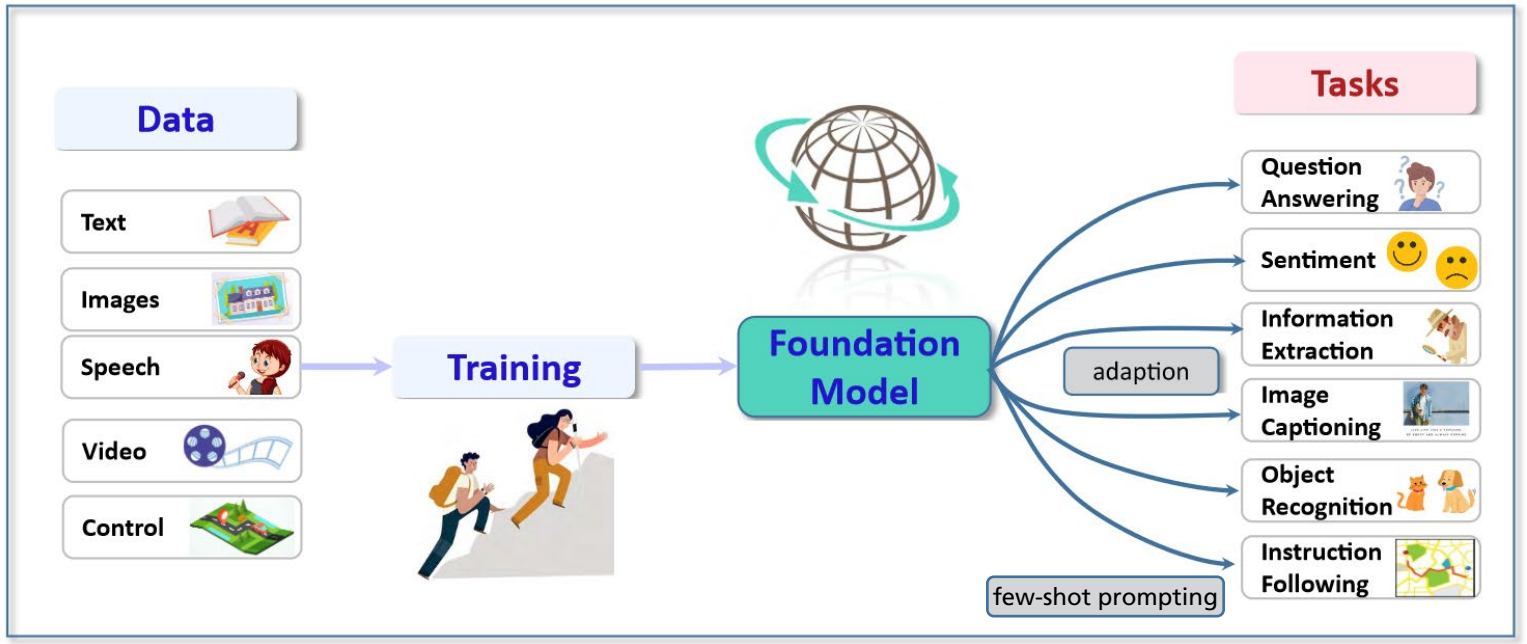
Foundation Models

- Pre-trained language models can process many media
  - Text, speech
  - images, video
  - 3D voxels
  - DNA sequences
  - protein formulae
  - numeric time series
  - robot control input / outouts

Same architecture for many tasks

Foundation Models

[Bommansani et al. 2021]



[Paass Giesselbach 2023]

- Properties
  - huge progress in many fields
  - many difficulties: generate wrong statements, bias to specific people, black box architecture, …

G. Foundation model

GPT-4 und BARD as Basis of Foundation Models

Strategy for Dialog Chatbots

- use a large generative **GPT** language model trained with text documents and dialog data
- adapt with **instruction tuning**, reinforcement with human feedback (**RLHF**)
- add special tools: retriever, calculator, translator, …

Details in text mining course

	GPT-4	GEMINI
Underlying Model	GPT-3.5	LaMDA (137 Md), PaLM (540 Md), PaLM 2, BARD
Model Parameters	1600 B.(?)	1560 B tokens (?), sparse mixture of experts
Training Data	> 840TB	Multimodal, "same amount" as GPT-4
Images	Image captioning, image generation with DALL-E-3	images in inputs, responses
Languages	good in 25 languages	46 languages, 20 programming languages (PaLM 2)
maximal input length	up to 32768 token	"multiple million" tokens
Internet Search	via plugin	yes
Tech Report	[OpenAI 2023]	[Google 2024] [Google 2024]

Academic Exams

- GPT-4 performance on academic and professional exams. [OpenAI 2023]

Better than 90% of students

Exam	GPT-4
Uniform Bar Exam (MBE+MEE+MPT)	298 / 400 (~90th)
LSAT	163 (~88th)
SAT Evidence-Based Reading & Writing	710 / 800 (~93rd)
SAT Math	700 / 800 (~89th)
Graduate Record Examination (GRE) Quantitative	163 / 170 (~80th)
Graduate Record Examination (GRE) Verbal	169 / 170 (~99th)
Graduate Record Examination (GRE) Writing	4 / 6 (~54th)
USABO Semifinal Exam 2020	87 / 150 (99th - 100th)
USNCO Local Section Exam 2022	36 / 60
Medical Knowledge Self-Assessment Program	75 %
Codeforces Rating	392 (below 5th)
AP Art History	5 (86th - 100th)
AP Biology	5 (85th - 100th)
AP Calculus BC	4 (43rd - 59th)
AP Chemistry	4 (71st - 88th)
AP English Language and Composition	2 (14th - 44th)
AP English Literature and Composition	2 (8th - 22nd)
AP Environmental Science	5 (91st - 100th)
AP Macroeconomics	

- Comparison of Gemini and GPT-4

Better than human experts: 89.8%

	Gemini Ultra	Gemini Pro	GPT-4	GPT-3.5	PaLM 2-L	Claude 2
MMLU Multiple-choice questions in 57 subjects (professional & academic) (Hendrycks et al., 2021a)	90.04% CoT@32*	79.13% CoT@8*	87.29% CoT@32 (via API**)	70% 5-shot	78.4% 5-shot	78.5% 5-shot CoT
	83.7% 5-shot	71.8% 5-shot	86.4% 5-shot (reported)			
GSM8K Grade-school math (Cobbe et al., 2021)	94.4% Maj1@32	86.5% Maj1@32	92.0% SFT & 5-shot CoT	57.1% 5-shot	80.0% 5-shot	88.0% 0-shot
MATH Math problems across 5 difficulty levels & 7 subdisciplines (Hendrycks et al., 2021b)	53.2% 4-shot	32.6% 4-shot	52.9% 4-shot (via API**)	34.1% 4-shot (via API**)	34.4% 4-shot	—
			50.3% (Zheng et al., 2023)			
BIG-Bench-Hard Subset of hard BIG-bench tasks written as CoT problems (Srivastava et al., 2022)	83.6% 3-shot	75.0% 3-shot	83.1% 3-shot (via API**)	66.6% 3-shot (via API**)	77.7% 3-shot	—
HumanEval Python coding tasks (Chen et al., 2021)	74.4% 0-shot (IT)	67.7% 0-shot (IT)	67.0% 0-shot (reported)	48.1% 0-shot	—	70.0% 0-shot
Natural2Code Python code generation. (New held-out set with no leakage on web)	74.9% 0-shot	69.6% 0-shot	73.9% 0-shot (via API**)	62.3% 0-shot (via API**)	—	—

GPT-4 passes Bavarian Abitur exam with grade 2

DALL-E 3 Image Generation

- ChatGPT 4 generates extended prompts
  - **Example prompt**: *"A middle-aged street sweeper walks alone in Capri, full body pic, professional photo, unsharp background"*
  - **ChatGPT text**: Illustrate a professional image of a middle-aged, Hispanic female street sweeper walking alone on the vibrant streets of Capri. She is portrayed in a full body picture while the background, comprised of the quaint pastel buildings and cobblestone streets of the island, is subtly out of focus to emphasize the subject of the image.
- Generate improved, extensive captions of images
- Train an image generator on long captions
- Dall-E-3 uses ChatGPT 4 to generate extended captions from short captions
  - Improved prompt following
  - Much better human evaluation scores

Generated with GPT-4 API

Applications of Foundation Models

- Text Generation : Articles, essays, stories, poems websites, blogs, or marketing materials.
- Invent Stories and Movie Plots
- Translation to / from many languages
- Summary of Documents
- Question Answering
- Generating Program Code
- Analyze Program Code
- Learning Languages: Create sample sentences, explain grammar rules, and make vocabulary suggestions.
- Medical Diagnoses: Support analysis of patient symptoms and medical records to suggest possible diagnoses or treatment options.
- Legal Documents: LLMs can review and analyze legal documents, contracts, and agreements to identify relevant information and potential issues.
- Face Recognition: unlock phones, manage photos, street surveillance
- Recommendation: streaming services, internet shopping
- Speech to text: digital assistants
- Text to speech
- Image to text: image captioning
- Text to image: Image generation, image madification, image infilling
- Text to video
- DNA to protein
- Text to control:
- Text analysis: classification, named entities,

Generative Models

Agenda

1. Generating Images with GANs
2. Latent Space Computations
3. Text to Image Generation with Language Models
4. Summary

Important Models and their Applications

Model	Year	Application	Properties
logistic regression	1920	classification	only linearly separable classes, used as final layer
multilayer perceptron MLP	1967	classification, regression	can theoretically approximate arbitrary functions
Recurrent neural network	1972	text generation, translation	only one embedding per word, long range problems
Conv. Neural Network CNN	1980	image classification and analysis	object recognition and image segmentation
Word2vec	2013	semantic similarity	only one embedding per word
Deep Q-Learning	2014	solve control problems	requires long training, many variants
Gen. Adversarial Netw. GAN	2014	image generation	good image quality, but difficult training
Transformer Encoder-Decoder	2017	translation, solve special tasks	translation, large models have similar capabilities as GPT
BERT	2018	prediction of masked words	finetuning: natural language understanding
GPT language model PaLM, ChatGPT, GPT4	2019	text generation, few-shot learning	large language models have unprecedented quality, solve many new problems without training
Vision transformer	2020	image classification	object classification
DALL E	2021	image generation from text	good quality, especially with diffusion models
imagen video	2022	video generation from text	good quality but only short videos
GATO for control	2022	combining text, image, control	control games & robots, image caption, answer questions

G. important models and their applications

Summary

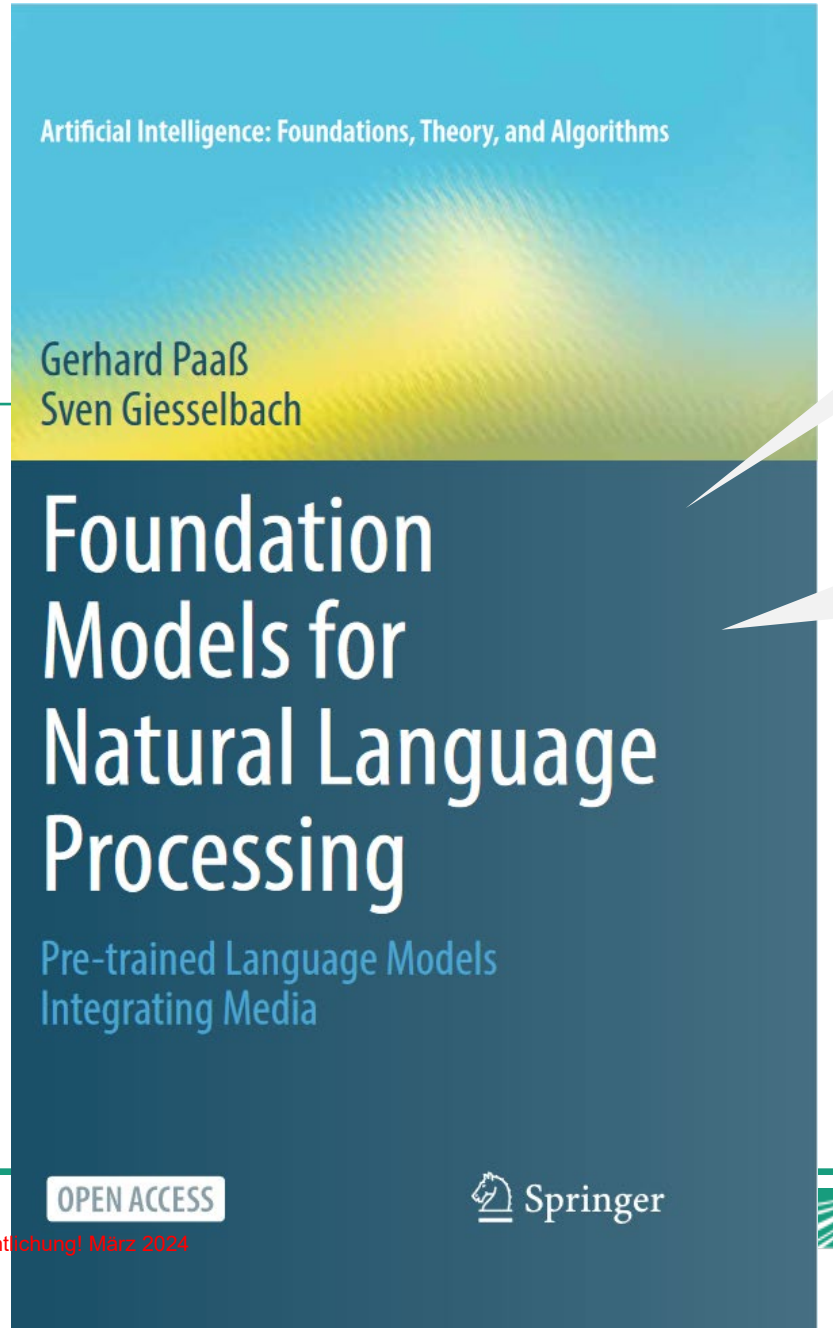
- **GAN** Generative Adversarial Network
  - two networks interact
  - Discriminator wants to distinguish fake images from real images
  - Generator wants to fool discriminator
- **Manipulation** of images
  - Manipulate latent representation
  - transform to a new style
- Use **transformer** models for image processing
  - transform image patches to input tokens
  - better image classification
  - connect text and images: image generation, image description
  - also applicable to video and many other media

New paradigm: Foundation Models

Textbooks

Less formal neural network intro with many graphics

Springer 2020



Comprehensive introduction to Foundation Models

Many applications and different media: sound, images, video

Springer Nature 2023

Download link

Fraunhofer BIG DATA AI

Disclaimer

Responsible contact: **Dr. Gerhard Paaß**, Fraunhofer IAIS, Sankt Augustin
E-mail: gerhard.paass@iais.fraunhofer.de