HELMHOLTZ AI | ARTIFICIAL INTELLIGENCE COOPERATION UNIT

DKRZ
DEUTSCHES KLIMARECHENZENTRUM

Introduction to Transformer Models

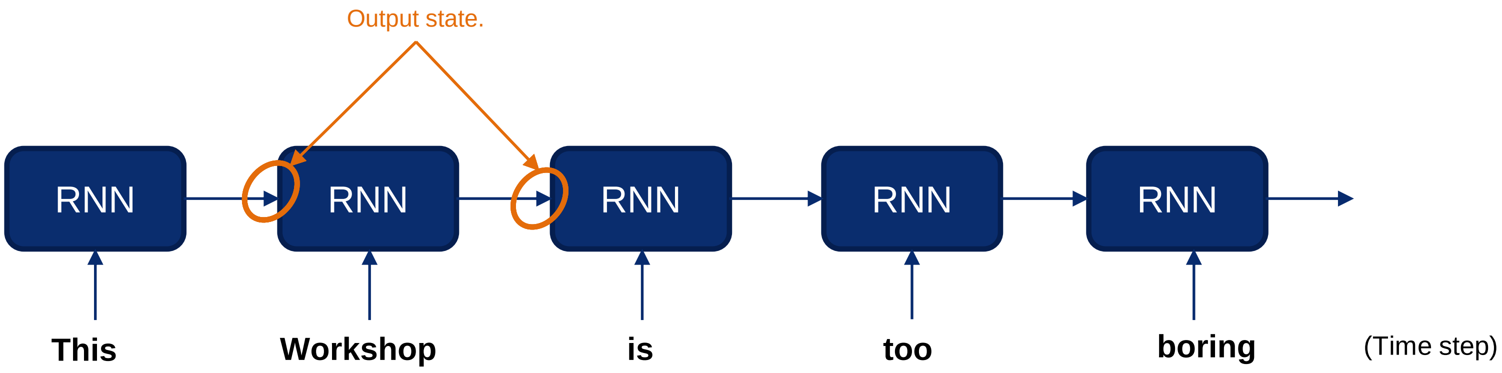
AI Consultants Earth & Environment @DKRZ

Mosaku Adeniyi

Helmholtz AI

DKRZ
DEUTSCHES
KLIMARECHENZENTRUM

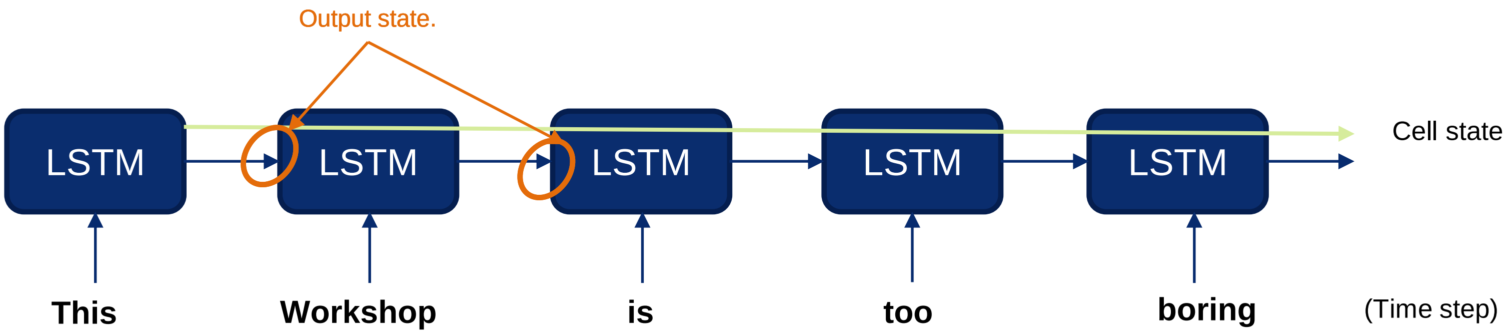
Recurrent Neural Network RNN



- Sequential introduction of data
- RNN thus requires very deep models
- **Problem**: Vanishing gradient or gradient explosion

Helmholtz AI

Long-Short-Time-Memory LSTM



Output state.

LSTM

LSTM

LSTM

LSTM

LSTM

Cell state

This

Workshop

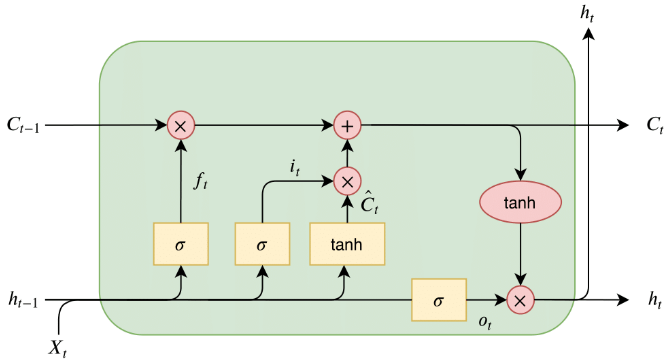
is

too

boring

(Time step)

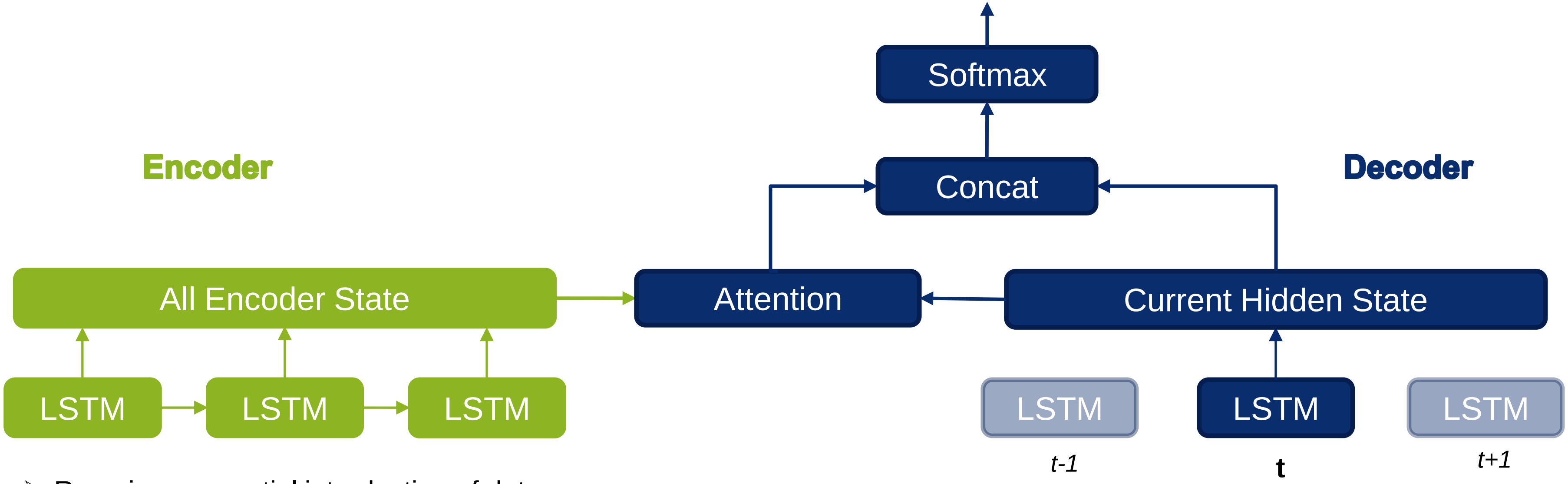
- ➢ Remains sequential introduction of data
- ➢ Ability to retain and transfer previously learned properties
- ➢ **Problem**: Not enough information is transferable



LSTM gate

Helmholtz AI

LSTM induced Attention

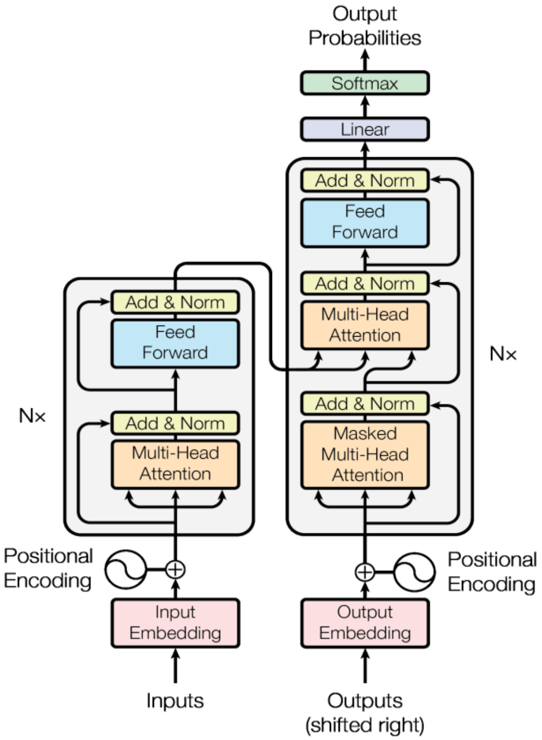


- ➢ Remains sequential introduction of data
- ➢ Ability to retain and transfer previously learned properties

LSTM gate

Self-Attention

„Attention Is All You Need“

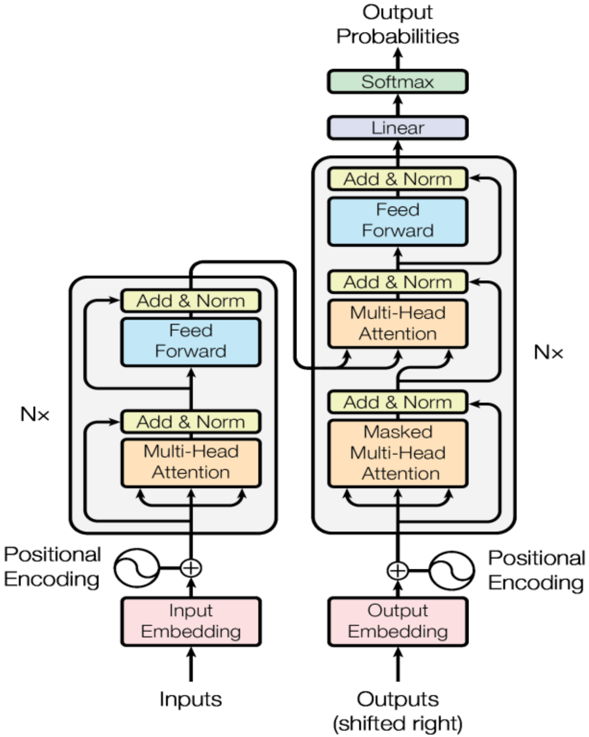


Vaswani: Attention Is All You Need

Helmholtz AI

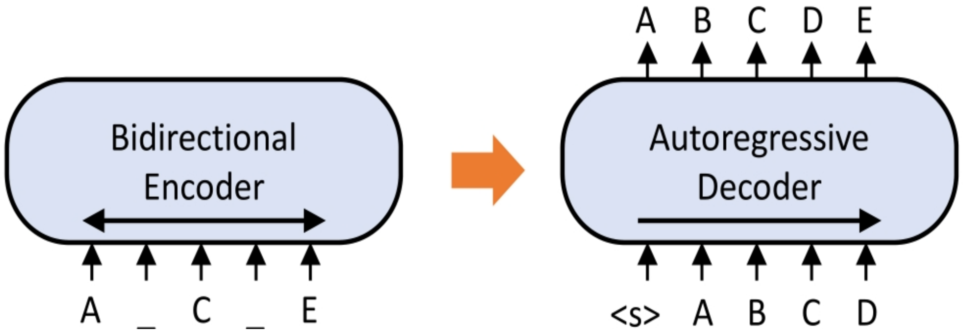
Transformer Models : Sequence to Sequence

- Also known as Encoder-Decoder Transformer
- The **Encoder** generates the context vector
- The **Decoder** collects the context vector to predicts the next token
- Both have Attention, FFNs, Normalization and residual connection
- It is usually used for Text translation
- Example is: **BART** (Bidirectional Auto-regressive Transformers)



Vaswani: Attention Is All You Need

Transformer Models : Autoregressive

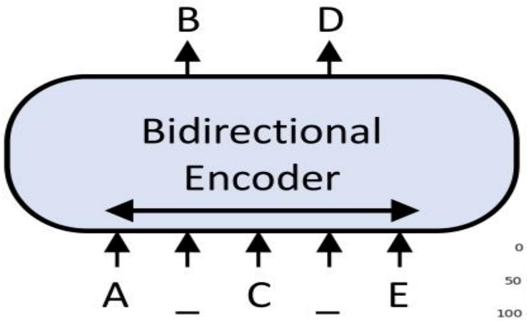


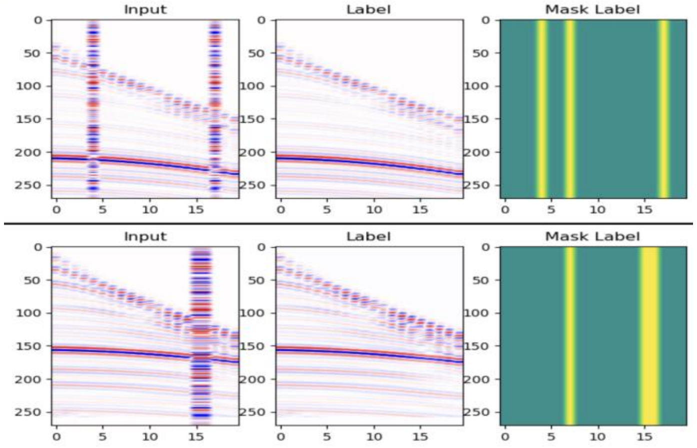
- ➢ This represents the Decoder part of the Transformer
- ➢ Generates token sequence, one token at a time
- ➢ Mask are used to prevent attention head from seeing what is next
- ➢ Predicts the next token after seeing the previous token
- ➢ Example is: **GPT** (Generative Pretrained Transformer)

Autoregressive-transformers

Helmholtz AI

Transformer Models : Auto-Encoding



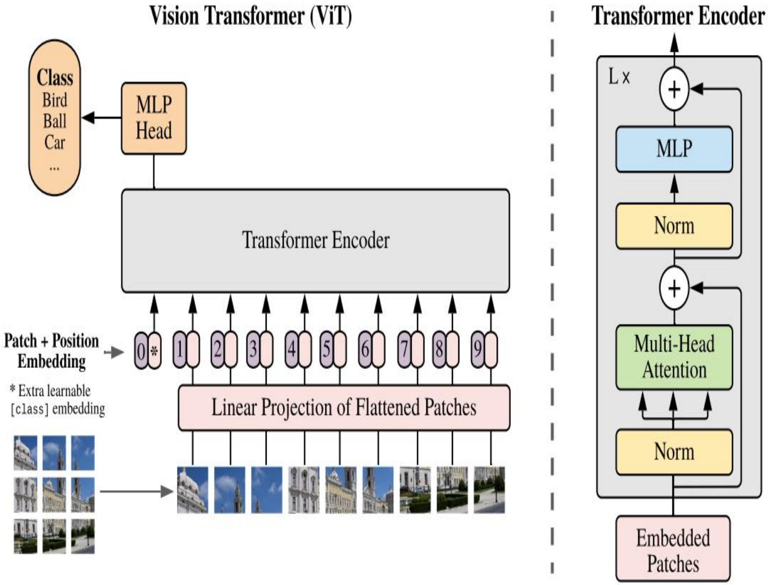


- ➢ Also known as „Non-Autoregressive“
- ➢ Usually used for reconstruction
- ➢ Input tokens are deliberatly masked or corrupted
- ➢ The attention head tries to predict it
- ➢ Example is: **BERT** (Bidirectional Encoder Representation from Transformer)

Autoencoding-transformers

Transformer Models : Vision Transformer

- Patches are processed in parallel
- No tokenization, images are treated as pixels
- Position Encoding captures spatial relationship, not sequential



Alexey 2020: An Image is Worth 16x16 Words

Helmholtz AI

Vision Transformer

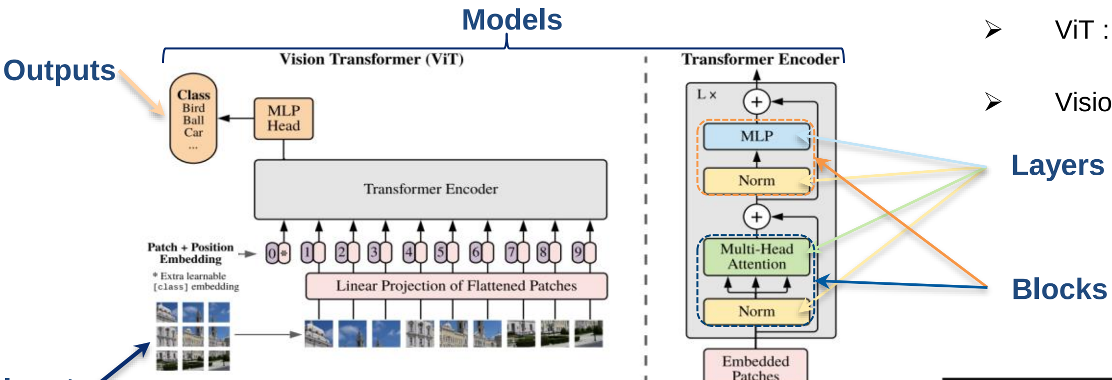


Figure 1: Model overview. We split an image into fixed-size patches, linearly embed each of the add position embeddings, and feed the resulting sequence of vectors to a standard Transform encoder. In order to perform classification, we use the standard approach of adding an extra learnat "classification token" to the sequence. The illustration of the Transformer encoder was inspired Vaswani et al. (2017).

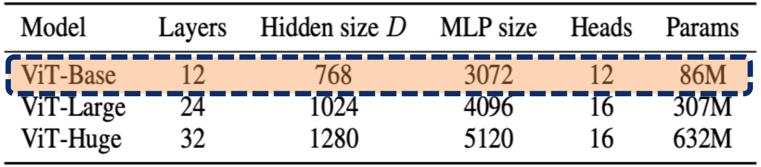
$$\mathbf{z}_0 = [\mathbf{x}_{\text{class}};\, \mathbf{x}^1_p\mathbf{E};\, \mathbf{x}^2_p\mathbf{E};\cdots;\, \mathbf{x}^N_p\mathbf{E}] + \mathbf{E}_{pos}, \qquad \mathbf{E} \in \mathbb{R}^{(P^2\cdot C)\times D},\ \mathbf{E}_{pos} \in \mathbb{R}^{(N+1)\times D} \qquad (1)$$

$$\mathbf{z'}_\ell = \text{MSA}(\text{LN}(\mathbf{z}_{\ell-1})) + \mathbf{z}_{\ell-1}, \qquad \ell = 1 \ldots L \qquad (2)$$

$$\mathbf{z}_\ell = \text{MLP}(\text{LN}(\mathbf{z'}_\ell)) + \mathbf{z'}_\ell, \qquad \ell = 1 \ldots L \qquad (3)$$

$$\mathbf{y} = \text{LN}(\mathbf{z}^0_L) \qquad (4)$$

- ViT : Another transformer flavour
- Vision Transformer Components are essential part of ViT



Model	Layers	Hidden size D	MLP size	Heads	Params
ViT-Base	12	768	3072	12	86M
ViT-Large	24	1024	4096	16	307M
ViT-Huge	32	1280	5120	16	632M

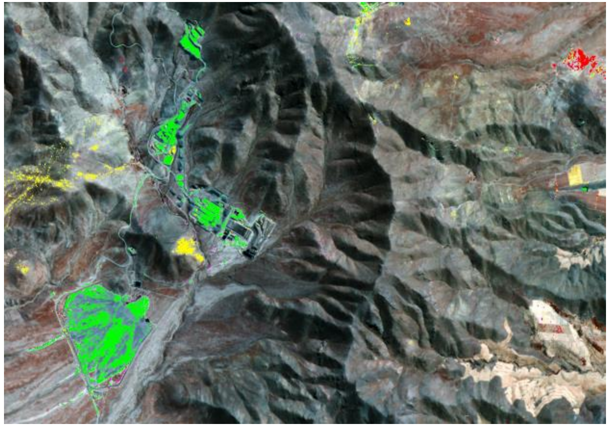
Table 1: Details of Vision Transformer model variants.

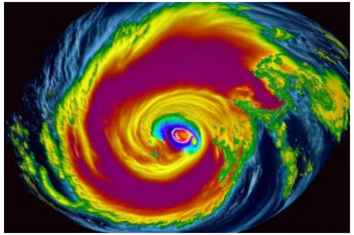
- Four equations that explains the model architecture
- Model sizes based on different numbers of hyper-parameters

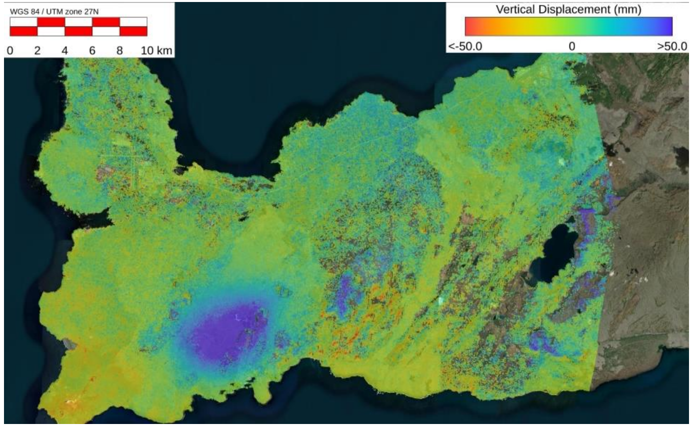
Alexey 2020: An Image is Worth 16x16 Words

Helmholtz AI

Application: Identification of surface patterns



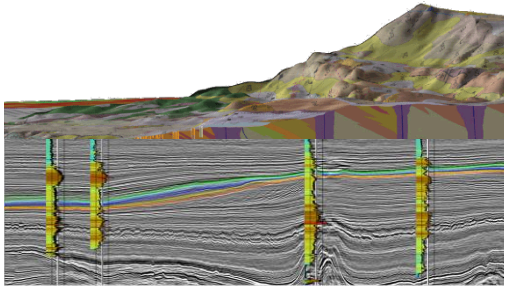


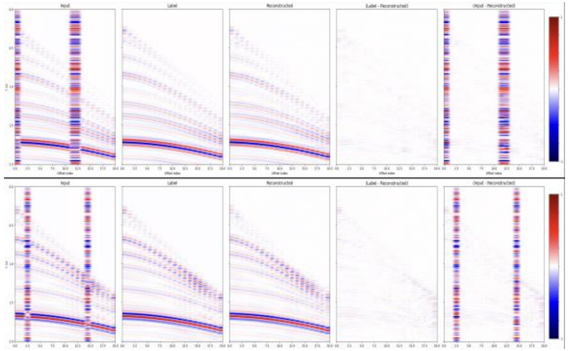


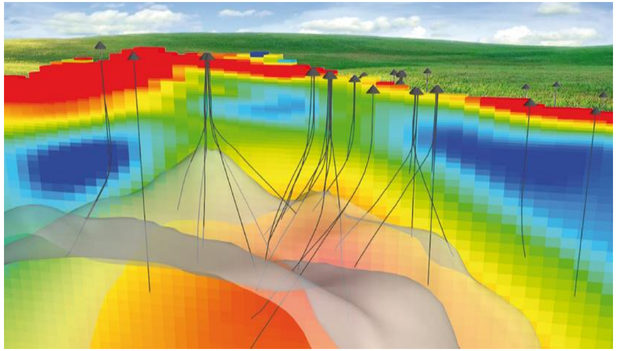
- Using ViT can aid in remote sensing application
- **Task**: Identification of surface patterns or entity

Helmholtz AI

Application: Investigating subsurface CO_2 storage potential





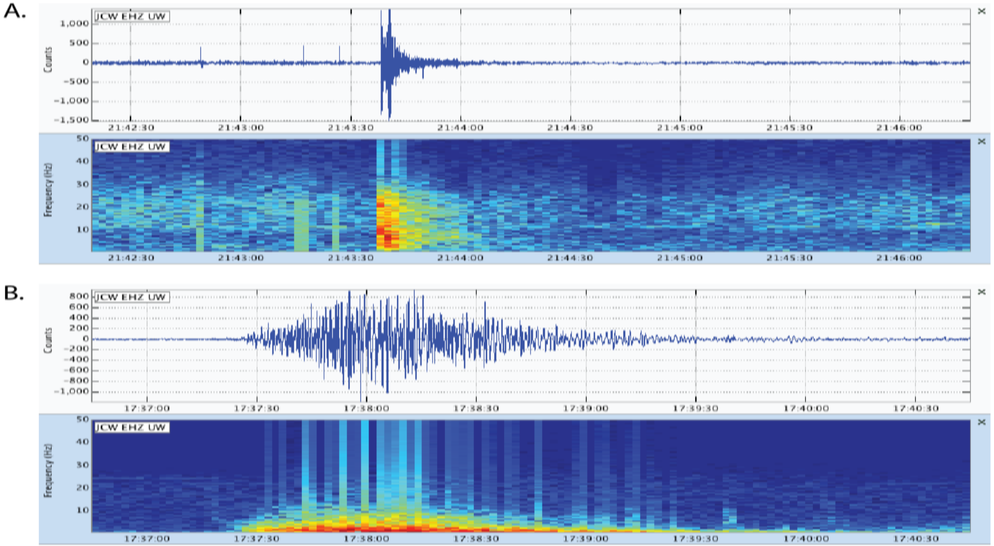


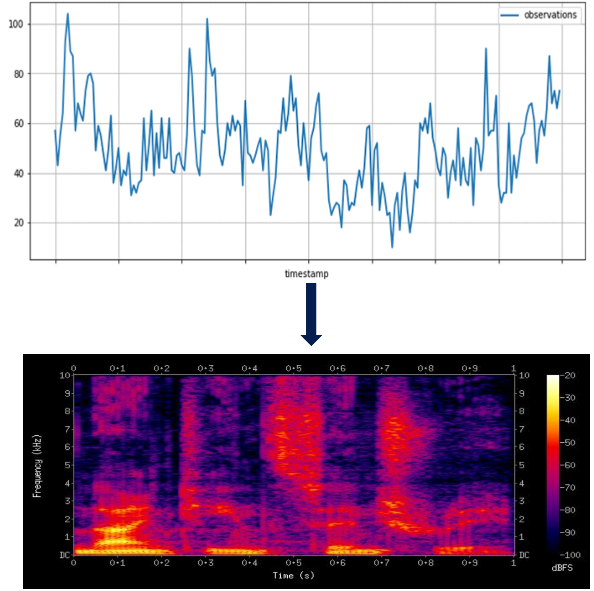
- Aid in seismic processing in subsurface investigation
- **Task**: Pretraining, Denoising and Velocity prediction

[1]

Helmholtz AI

Application: Earthquake prediction





- ➢ Surface system can be investigated using Transformer
- ➢ Time Series to Images using STFT (Short-time Fourier Transform)
- ➢ ViT can then be used
- ➢ **Task**: Earthquake prediction

Earthquake signal

Helmholtz AI

QUESTIONS