

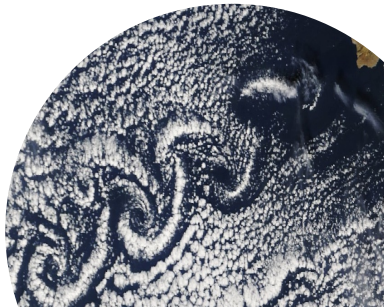
From Convolutions to Attention – Advanced Deep Learning Methods for Climate Data

Deep Learning Course for Climate Scientists

Alexander Fischer, Maximilian Witte

September 24, 2026

German Climate Computing Center (DKRZ)
Hamburg, Germany



Why do we care about designing ML models for physical fields?

- Scientific ML models aim to learn physical processes from data
 - ▷ infilling missing data
 - ▷ super-resolution of coarse data (aka downscaling)
 - ▷ emulating medium- (weather) to long-range (climate) dynamics of physical fields
- Understanding the properties of physical fields and how they relate to model architecture is crucial for:
 - ▷ designing architectures that can capture key physical phenomena
 - ▷ improving generalization and physical consistency of predictions

Properties of Physical Fields in Scientific ML

- **Spatial smoothness and gradients**

- ▷ PDE-governed fields (e.g., pressure, temperature, velocity) depend on accurate spatial derivatives for stable dynamics:

$$\nabla p, \nabla T, \nabla \cdot \mathbf{u}, \Delta \mathbf{u}, \nabla \times \mathbf{u}$$

- **Nonlinear interactions**

- ▷ Nonlinear advection, $\mathbf{u} \cdot \nabla \mathbf{u}$, couples small-scale features to large-scale flow

- **Conservation laws**

- ▷ Mass, momentum, and energy conservation are essential to avoid unphysical, unstable predictions

- **Multi-scale structure**

- ▷ Coexisting coherent structures and small-scale turbulence require long-range dependencies

How can we incorporate knowledge a priori into model design?

Structural assumptions inherently baked into the architecture → patterns a model can capture without excessive data or model capacity

- Useful inductive biases for PDE learning:
 - ▷ locality of interactions
 - ▷ translational / rotational equivariance
 - ▷ hierarchical structure
 - ▷ global coupling across scales
- Different architectures encode these biases differently:
 - ▷ CNNs → local spatial operators
 - ▷ Transformers → global attention
 - ▷ Graph Networks → message passing
 - ▷ Diffusion / Flow Matching → generative modeling of field distributions

Promoting Physical Consistency

Additional constraints can enforce or promote physical laws

- **Physics-informed loss terms**

- ▷ PDE residual loss e.g. for Navier-Stokes momentum equation:

$$\mathcal{L}_{PDE} = ||\partial_t \mathbf{u} + \mathbf{u} \cdot \nabla \mathbf{u} - \nu \nabla^2 \mathbf{u}||^2$$

- ▷ divergence-free constraint, e.g. for incompressible flow:

$$\mathcal{L}_{div} = ||\nabla \cdot \mathbf{u}||^2$$

- **Spectral constraints**

- ▷ enforce realistic turbulence spectra

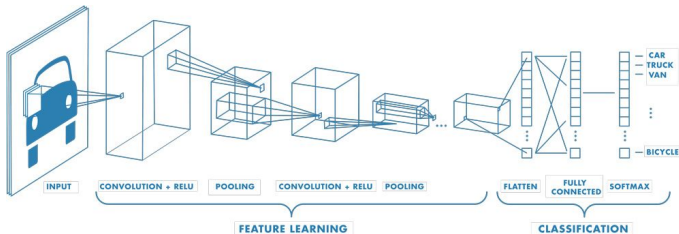
- **Architectural approaches**

- ▷ Fourier Neural Operators to learn in the frequency domain

- ▷ Constraint-preserving parameterizations, e.g. divergence-free or flux-conserving

Recap: Convolutional Neural Networks (CNNs) i

- CNNs apply spatial convolutions to extract local features
- **Inductive biases**
 - ▷ locality of interactions (sliding window of convolution)
 - ▷ translation equivariance
 - ▷ hierarchical spatial feature extraction → local to global features



Advantages

- efficient for grid-based data (operates directly on physical fields)
- good at capturing local operators (finite-difference stencils)
- parameter sharing improves generalization

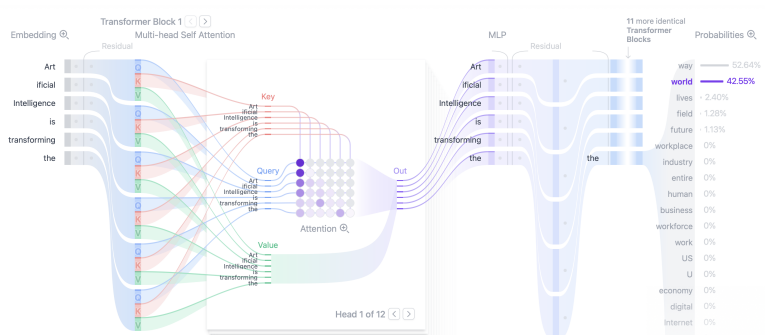
Limitations

- limited receptive field for long-range interactions
- capturing multi-scale / global dynamics requires deep architectures

⇒ **Luckily, there's more than just CNNs!**

Transformer Architecture

Nice visualization for Natural Language Processing: **transformer-explainer**
Explanations by 3Blue1Brown: **Transformers, the tech behind LLMs**



Tokenization → Embeddings → Transformer Blocks → Output

Attention Mechanism

- Each token interacts with all other tokens (aka chunks) in a sequence via attention $\rightarrow$ *soft & associative memory* mechanism
- Self-Attention (scaled dot product $\rightarrow$ softmax-weighted aggregation)

$$\mathbf{Q} = \mathbf{XW}_Q, \quad \mathbf{K} = \mathbf{XW}_K, \quad \mathbf{V} = \mathbf{XW}_V$$
$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) := \text{softmax}\left(\underbrace{\frac{\mathbf{QK}^T}{\sqrt{d_k}}}_{\text{attention scores}}\right) \mathbf{V}$$

- Interpretation for physical fields:
 - ▷ each image patch interacts with any other patch
 - ▷ enables modeling of long-range physical coupling

Feature Extraction in Transformer Blocks

A transformer block consists of 3 main components:

- Multi-Headed Self-Attention (MHSA)
- Feedforward MLP
- Residual Stream (skip connections)

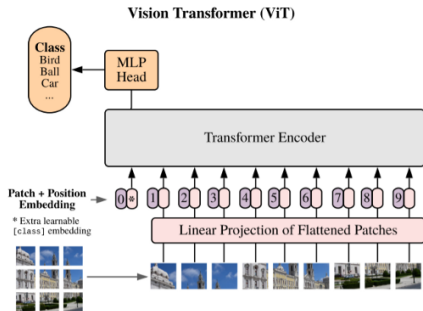
MHSA layer

- multiple attention heads: $\text{MHSA}(\mathbf{X}) = \text{Concat}(\mathbf{Z}_1, \dots, \mathbf{Z}_h) \mathbf{W}_O$
- each head learns different interaction patterns
- captures spatial correlations across the entire domain

MLP layer

- applied independently on each token: $\text{MLP}(\mathbf{x}) = \mathbf{W}_2 \sigma(\mathbf{W}_1 \mathbf{x})$
- introduces nonlinear transformations
- mixes channel information (e.g. pressure–velocity coupling)

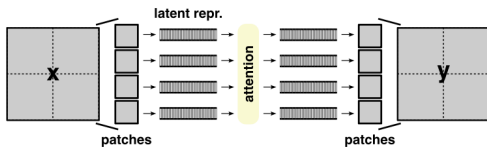
Vision Transformers (ViT) – Tokenization of Physical Fields i



- Convert physical fields into **sequence of patch embeddings / tokens**
- For a 2D physical field (or vector field): $\mathbf{X} \in \mathbb{R}^{C \times W \times H}$
- Field is divided into $N = \frac{W \cdot H}{P^2}$ patches with $P \times P$ pixels each
→ flattened into a vector:

$$\mathbf{X} \rightarrow N \times \mathbf{X}_i \in \mathbb{R}^{P^2 C}$$

Vision Transformers (ViT) – Tokenization of Physical Fields ii



- A linear projection creates the token in some embedding space:

$$\mathbf{z}_i = (\mathbf{X}_i \mathbf{W} + \mathbf{b}) \in \mathbb{R}^D \quad \rightarrow \quad \mathbf{Z} = \begin{bmatrix} \mathbf{z}_1^\top \\ \vdots \\ \mathbf{z}_N^\top \end{bmatrix} \in \mathbb{R}^{N \times D}$$

- Attention is an *orderless* operation $\rightarrow$ add *positional embeddings* to encode/retain spatial information about patch locations in the grid

Advantages and Limitations of Vision Transformers

Advantages

- instant global receptive field → captures long-range interactions
- scalable to large datasets and high-resolution images

Limitations

- weak spatial inductive bias: requires large training data
- latent representations do a priori not correspond to physical fields
→ harder to learn & enforce physical structure/constraints
- patch tokenization may lose fine-scale details at patch boundaries

⇒ **Can we do better than standard ViT?**

Advantages

- **Physical interpretability & consistency**

- ▷ intermediate states after each layer are remapped to physical fields (velocity, pressure, temperature)
- ▷ attention mechanism works on different spatial scales that are retained throughout the model as explicit residual streams
- ▷ flexible framework: attention similarly possible across variables or time

- **Multi-scale, residual deformations**

- ▷ hierarchical operators capture cross-scale interactions → stronger physical inductive bias
- ▷ resembles multigrid or spectral numerical methods

Hands-on Session: FSA for Downscaling of Climate Data

In this hands-on session, we will use a pre-trained FSA transformer for inference in downscaling tasks.

We will focus on:

- Quick introduction of the FSA model and multi-scale decomposition
- Downscaling using cross-variable attention
- Inspecting learned variable attention matrices

ML COURSE



HANDS-ON SESSIONS



TRANSFORMER



FSA FOR CLIMATE DATA

