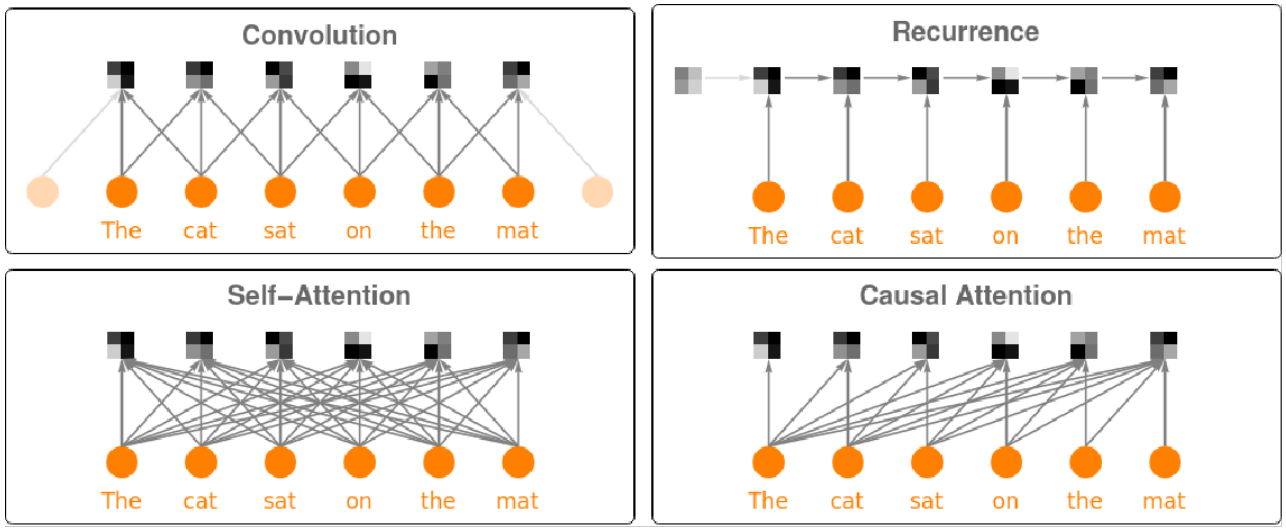
Transformers e atenção para dados ambientais

Cesar de Oliveira F. Silva

Setembro / 2026

Transformers



Ideia central

Transformers modelam dependências entre elementos de uma sequência ou campo espacial por meio de **atenção**, permitindo integrar informação de longo alcance sem recorrer a recorrência explícita.

Transformers para dados ambientais — visão geral

- Objetivo: modelar dependências complexas no **tempo**, no **espaço** e entre **múltiplas modalidades**.
- Força principal: atenção explícita entre posições, permitindo capturar relações distantes.
- Aplicações típicas:
  - previsão climática local;
  - downscaling estatístico;
  - fusão multimodal;
  - séries temporais multivariadas e multi-site;
  - assimilação de fontes heterogêneas.

Por que Transformers?

- Em RNNs, a informação precisa atravessar muitos passos temporais.
- Em Transformers, qualquer posição pode interagir diretamente com qualquer outra.
- Isso é especialmente útil quando o fenômeno depende de:
  - relações de longo alcance;
  - múltiplas escalas temporais;
  - interações espaciais não locais;
  - fusão entre sensores e fontes auxiliares.

Mensagem prática

Transformers são fortes quando o problema exige **contexto global** e **integração flexível de informação**.

Mecanismo de atenção — fórmula essencial

Scaled Dot-Product Attention

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V$$

- Q (*queries*): o que cada posição procura;
- K (*keys*): como cada posição pode ser encontrada;
- V (*values*): a informação efetivamente agregada.

Interpretação

A atenção calcula pesos de afinidade entre posições e usa esses pesos para combinar informação relevante de forma adaptativa.

Leitura intuitiva da atenção

- Cada token consulta os demais tokens da sequência ou grade espacial.
- A similaridade entre query e key define quanto uma posição deve influenciar a outra.
- O resultado é uma representação contextualizada, dependente do conteúdo.

Em linguagem simples

"Para entender esta posição, de quais outras posições eu preciso?"

- Em dados ambientais, isso pode significar:
  - olhar para tempos passados relevantes;
  - relacionar regiões distantes;
  - integrar diferentes fontes observacionais.

Multi-head attention e seus benefícios

Definição

$$\mathrm{MH}(Q, K, V) = \mathrm{Concat}(\mathrm{head}_1, \ldots, \mathrm{head}_h) W^O$$

- Em vez de uma única atenção, o modelo usa várias cabeças em paralelo.
- Cada *head* pode aprender relações diferentes.

Por que isso é útil?

- múltiplas escalas temporais;
- diferentes bandas ou variáveis;
- relações espaciais distintas;
- padrões sazonais, extremos e dependências locais/globais.

Positional encoding: tempo e espaço importam

- O mecanismo de atenção, por si só, não conhece ordem nem posição.
- Por isso, é necessário injetar informação posicional na entrada.

Opções comuns

- **sinusoidal encoding** (sin, cos);
- **learned positional embeddings**;
- codificações 2D para linhas e colunas;
- embeddings contínuos para tempo irregular.

- Em séries temporais, o encoding pode representar:
  - ordem dos passos;
  - calendário;
  - sazonalidade;
  - intervalos entre observações.

Arquitetura básica de um Transformer

- Componentes centrais:
  - embeddings de entrada;
  - positional encoding;
  - multi-head attention;
  - feed-forward network;
  - residual connections;
  - layer normalization.
- Em problemas sequenciais, também é comum usar:
  - **causal masking**;
  - encoder-only, decoder-only ou encoder–decoder.

Comparação rápida

Ao contrário de uma RNN, o Transformer processa toda a sequência de forma paralela.

Transformers para séries temporais

- Variantes específicas para séries incluem:
  - Time-Series Transformer;
  - Temporal Fusion Transformer;
  - Informer e variantes esparsas.
- Vantagens:
  - captura de dependências longas;
  - paralelização no treinamento;
  - integração natural de variáveis exógenas.
- Cuidados:
  - custo computacional;
  - necessidade de masking causal;
  - risco de overfitting em séries curtas.

Causal masking e previsão temporal

- Em previsão, o modelo não pode acessar informação futura durante o treinamento.
- Isso é garantido com **máscaras causais**.

Ideia

A posição t só pode atender a posições $\leq t$.

- Sem masking, ocorre **vazamento temporal**.
- Em tarefas multi-step, o desenho da máscara depende do esquema:
  - autoregressivo;
  - encoder–decoder;
  - previsão direta multi-output.

Transformers para dados espaciais e imagens raster

- Em imagens, uma estratégia comum é **patchify**:
  - dividir a imagem em patches;
  - tratar cada patch como um token.
- Exemplo clássico: **Vision Transformer (ViT)**.
- Em aplicações ambientais, também são comuns arquiteturas híbridas:
  - CNN para padrões locais;
  - Transformer para contexto global.

Aplicações

segmentação, classificação de uso do solo, fusão multi-temporal, downscaling espacial.

Tokens espaço-temporais

- Para dados espaço-temporais, é possível combinar:
  - posição espacial;
  - timestamp;
  - modalidade;
  - variáveis auxiliares.
- Cada token pode representar:
  - um patch em um dado instante;
  - uma estação em um tempo específico;
  - uma observação multimodal alinhada.

Vantagem

A atenção pode modelar simultaneamente relações no espaço e no tempo.

Integração multimodal e cross-attention

- Dados ambientais raramente vêm de uma única fonte.
- Modalidades típicas:
  - satélite óptico;
  - SAR;
  - reanálises;
  - estações in-situ;
  - topografia e uso do solo.

Cross-attention

Queries de uma modalidade atendem a keys/values de outra.

- Isso permite condicionar previsões locais em informações regionais ou globais.

Modelagem de dependências de longo alcance

- A atenção global pode capturar:
  - teleconexões climáticas;
  - dependências espaciais remotas;
  - influência de eventos passados distantes.
- Essa é uma vantagem importante sobre arquiteturas estritamente locais ou recorrentes.

Desafio

O custo da atenção completa cresce quadraticamente com o número de tokens:

$$O(N^2)$$

- Em séries longas ou grades grandes, memória e latência tornam-se gargalos.

Encoder, decoder e variantes

- **Encoder-only:** representação contextual da entrada.
- **Decoder-only:** previsão autoregressiva.
- **Encoder–decoder:** transforma sequência de entrada em sequência de saída.

Exemplos

- encoder-only: classificação e regressão;
- decoder-only: forecasting autoregressivo;
- encoder–decoder: previsão multi-step e tradução entre modalidades.

Eficiência e variantes escaláveis

- Estratégias para reduzir custo:
  - atenção esparsa;
  - janelas deslizantes;
  - atenção dilatada;
  - memory tokens;
  - compressão hierárquica.
- Exemplos conhecidos:
  - Informer;
  - Longformer;
  - BigBird;
  - Linformer;
  - Performer.

Trade-off

Eficiência maior geralmente implica aproximações que podem reduzir fidelidade em dependências muito longas.

Transformers vs RNNs — comparação prática

- **RNNs / LSTM / GRU:**
  - boas para sequências menores;
  - custo de memória menor;
  - naturais para cenários causais e streaming.
- **Transformers:**
  - fortes em dependências longas;
  - excelentes para fusão multimodal;
  - maior custo computacional;
  - normalmente exigem mais dados.

Regra prática

Transformers tendem a ser mais vantajosos quando o problema requer **contexto global**, **escala** e **integração entre fontes**.

Como incorporar conhecimento físico

- Nem toda relação estatística é fisicamente plausível.
- Em problemas ambientais, é desejável impor estrutura ao modelo.

Estratégias

- máscaras baseadas em conectividade física;
- embeddings guiados por distância, bacia ou topologia;
- saídas de modelos físicos como tokens de entrada;
- aprendizado de resíduos de modelos físicos;
- penalizações por violação de leis de conservação.

Séries irregulares e eventos esparsos

- Em observações ambientais, o intervalo entre medidas pode variar.
- Também é comum haver eventos raros, como cheias, secas ou pulsos de poluição.

Adaptações úteis

- embeddings contínuos de tempo;
- codificação de Δt;
- máscaras para gaps;
- atenção condicionada a eventos relevantes.

- Isso permite lidar melhor com amostragem irregular e regimes não estacionários.

Pré-treinamento e aprendizado auto-supervisionado

- Transformers se beneficiam fortemente de pré-treinamento em grandes bases.
- Objetivos comuns:
  - **masked token modeling**;
  - aprendizado contrastivo;
  - previsão como tarefa pretexto.
- Em dados ambientais, isso é promissor porque há muitos dados observacionais e reanálises sem rótulos detalhados.

Fluxo típico

pré-treinamento em larga escala $\rightarrow$ fine-tuning local para tarefa específica

Interpretação: mapas de atenção e explicabilidade

- Mapas de atenção podem sugerir quais regiões, tempos ou modalidades influenciaram a previsão.
- Isso é útil para inspeção qualitativa e diagnóstico de comportamento do modelo.

Limitação importante

Atenção não equivale, por si só, a explicação causal.

- Idealmente, combinar com:
  - testes de perturbação;
  - saliency;
  - ablação de entradas;
  - verificação de plausibilidade física.

Escala e integração com bases climáticas massivas

- Bases como reanálises, ensembles e arquivos de satélite permitem pré-treinamento multi-regional.
- Desafios práticos:
  - diferentes resoluções espaciais;
  - diferentes frequências temporais;
  - alinhamento entre grids;
  - normalização por variável e por domínio.
- Estratégias operacionais:
  - chunking espacial e temporal;
  - atenção hierárquica;
  - pré-agregação por escalas.

Boas práticas para problemas ambientais

- Escolher a tokenização de acordo com a escala do fenômeno.
- Usar validação temporal e espacial para evitar leakage.
- Monitorar overfitting e memorizações espúrias.
- Incorporar restrições físicas sempre que possível.
- Comparar com baselines mais simples:
  - RNNs/TCNs;
  - modelos estatísticos;
  - modelos físicos ou híbridos.

Resumo prático

1. Transformers modelam relações globais via atenção.
2. Positional encoding é essencial para preservar ordem e localização.
3. Em dados ambientais, o valor está em integrar espaço, tempo e modalidades.
4. O principal gargalo é o custo computacional da atenção completa.
5. Modelos eficientes, híbridos e guiados por física tendem a ser os mais promissores.