Transformers for Climate and Earth System Modeling

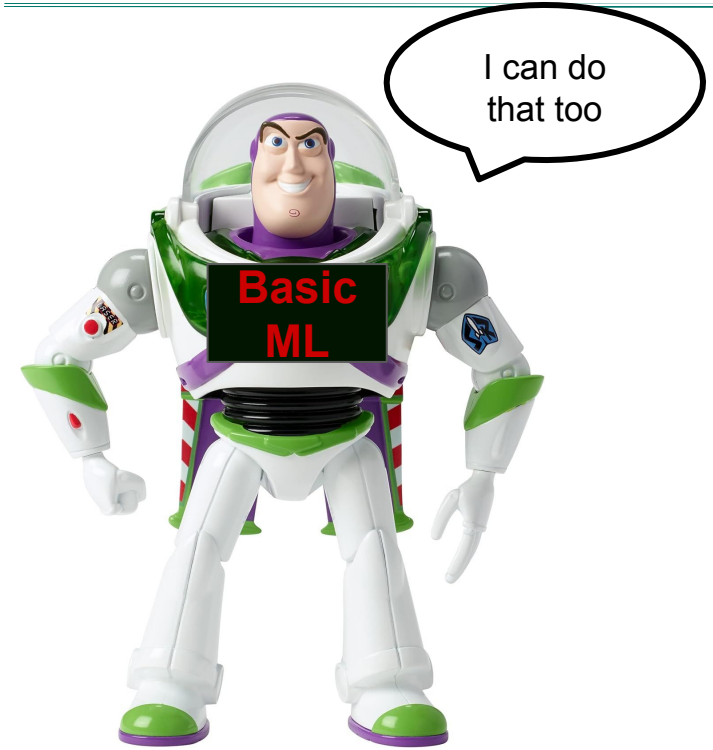
Aya Lahlou
Columbia University
June 4th, 2025

Overview

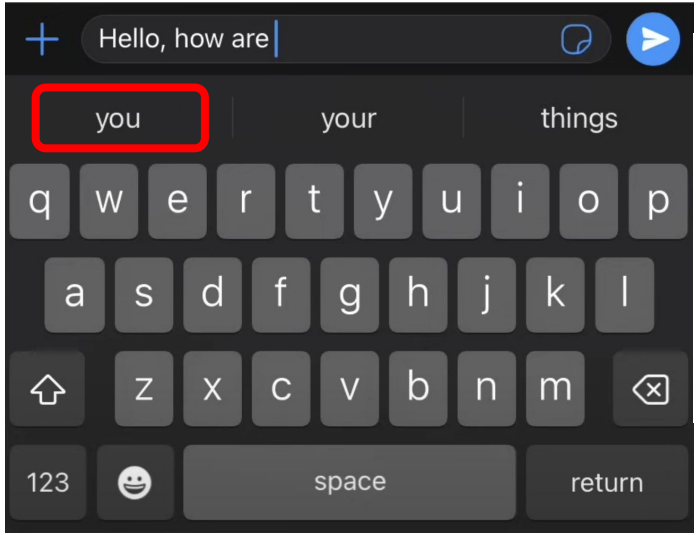
- What transformers are
- Why they are important in Machine learning
- Potential Applications in Climate Modeling
- Key takeaways

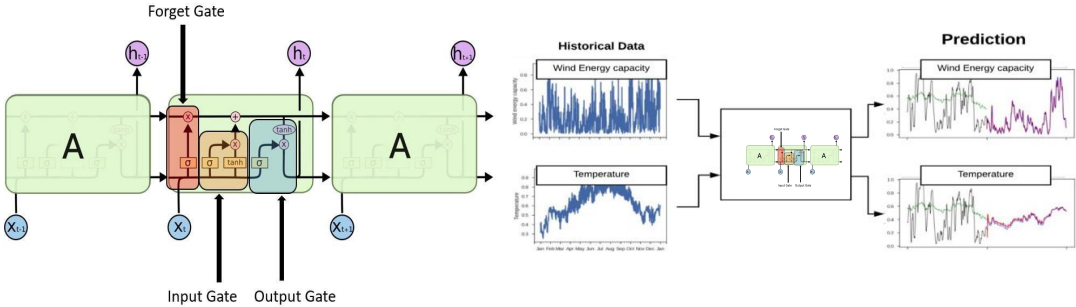
Transformers

RNN vs. Transformers



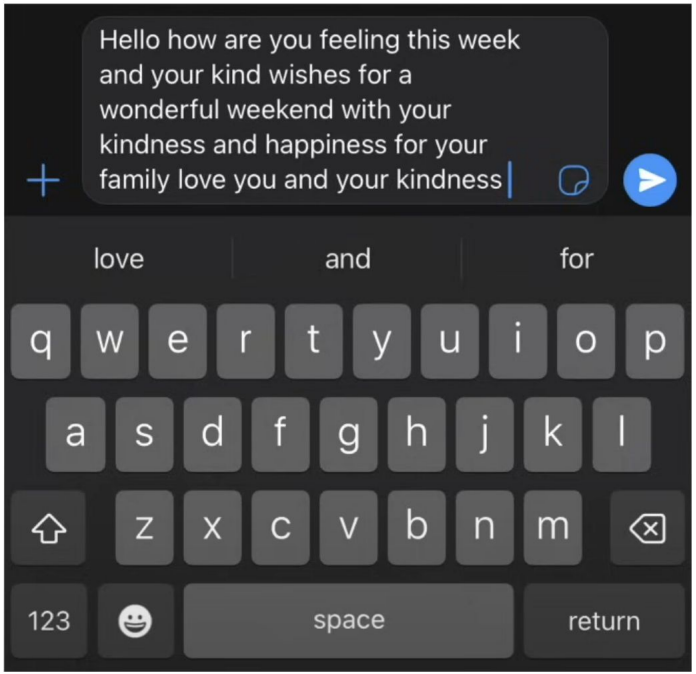
RNN vs. Transformers

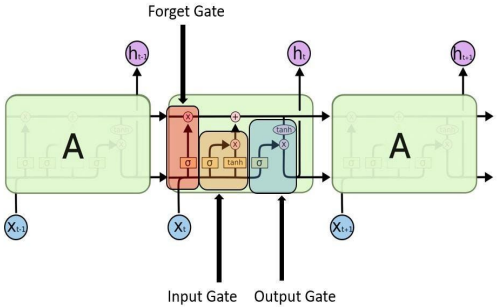


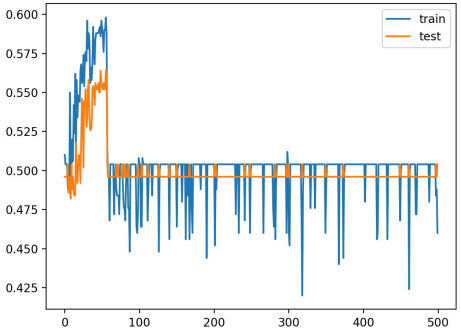


RNN: Remember the previous information

RNN vs. Transformers



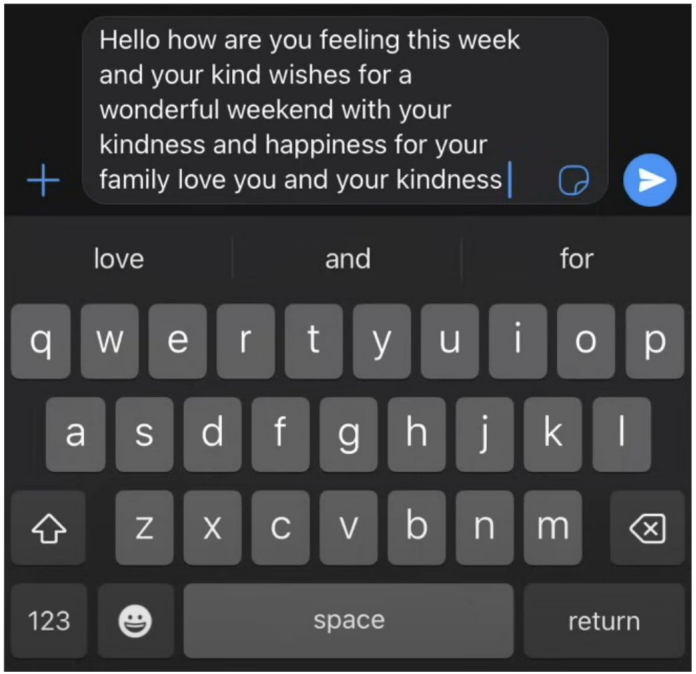


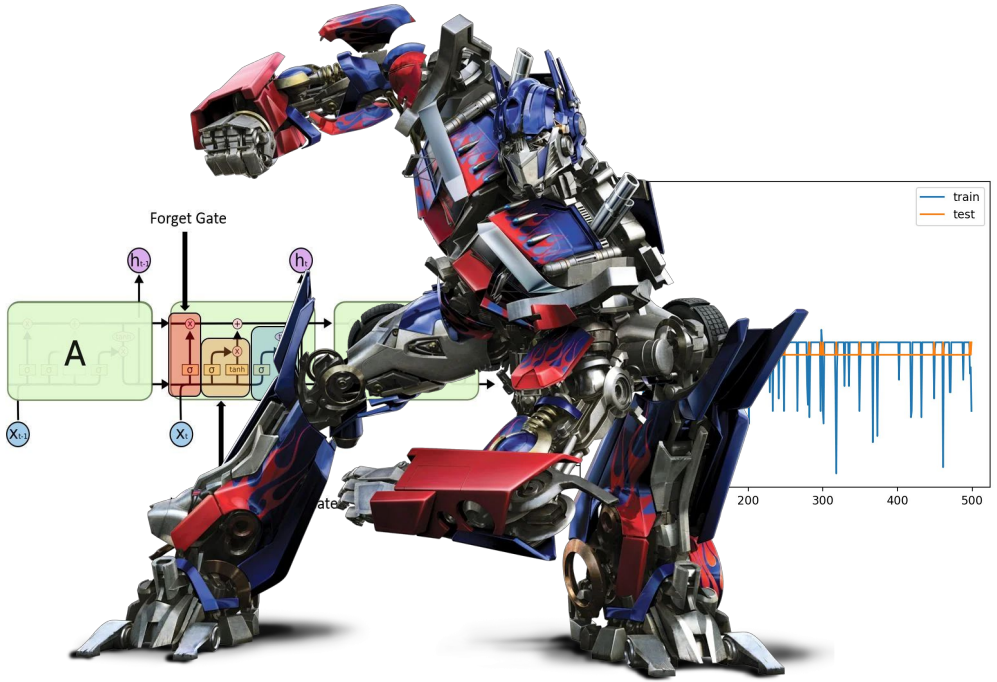


- Exploding/Vanishing Gradient
- Sequential processing

RNN: Fail to capture long-term dependencies

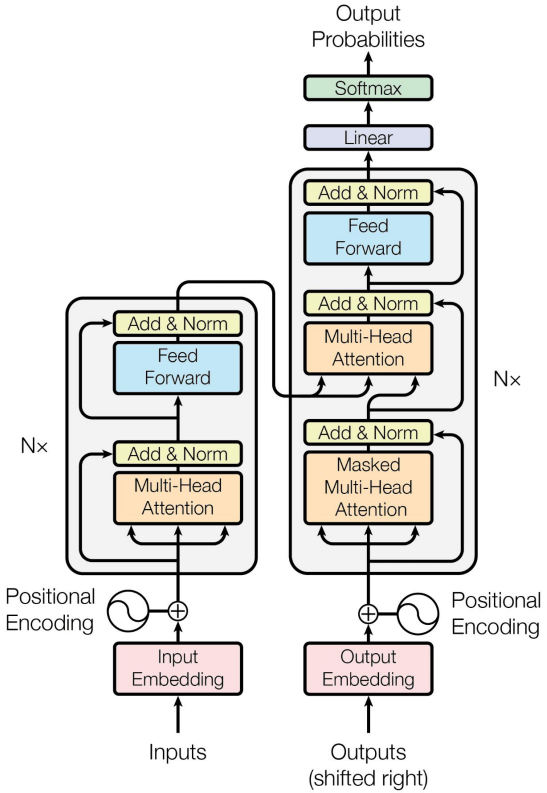
RNN vs. Transformers

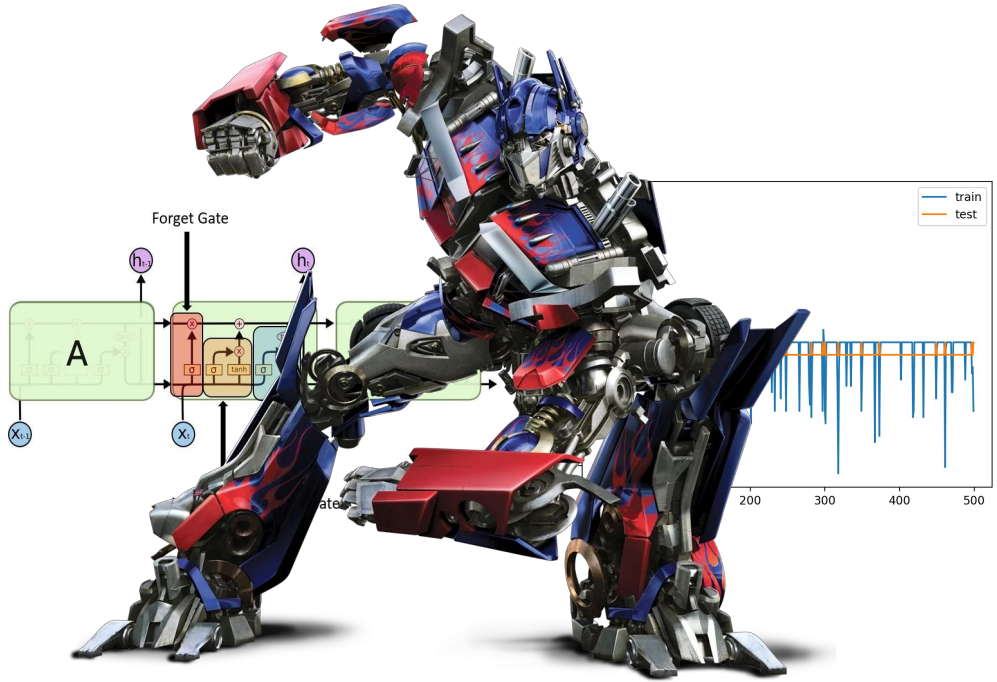




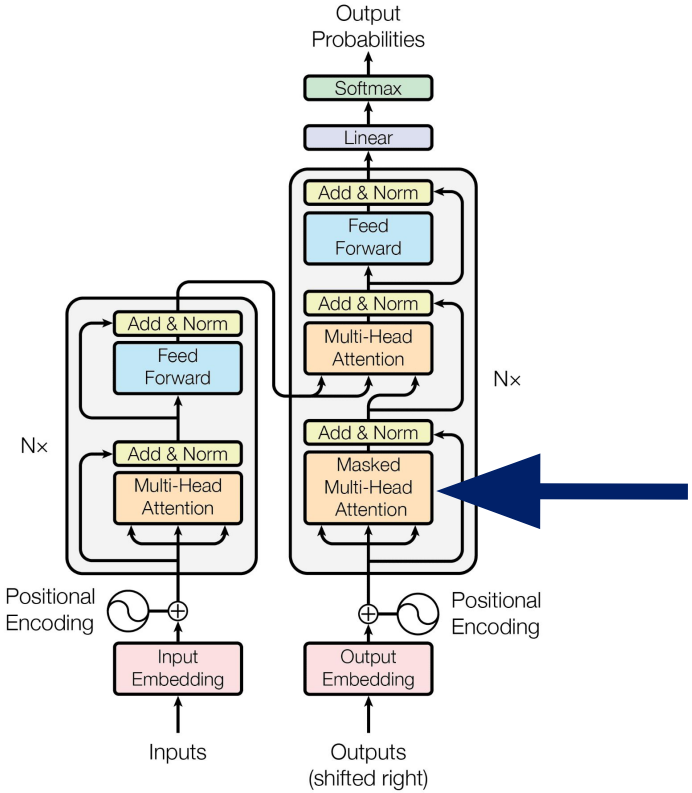
RNN: Fail to capture long-term dependencies

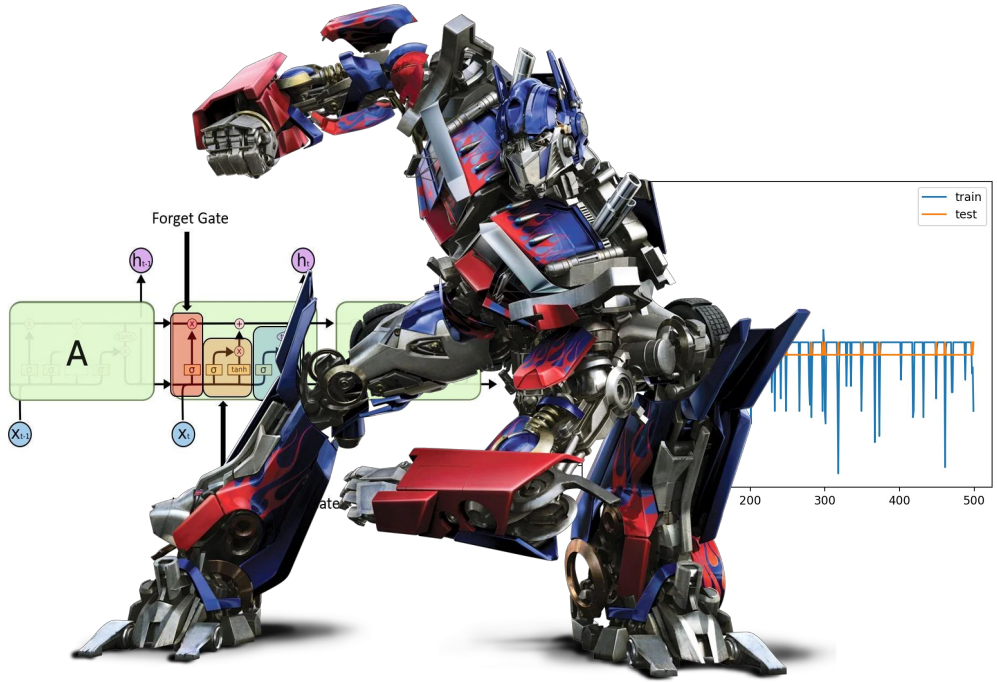
RNN vs. Transformers



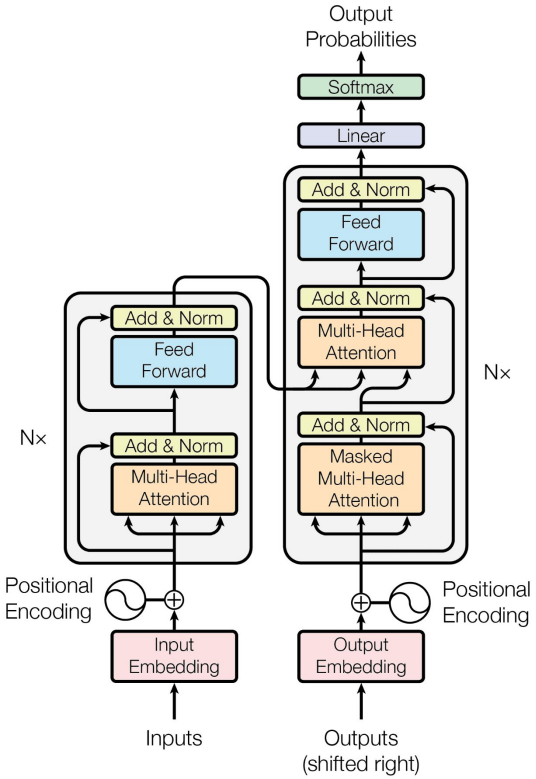


RNN vs. Transformers





Transformer Architecture



Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

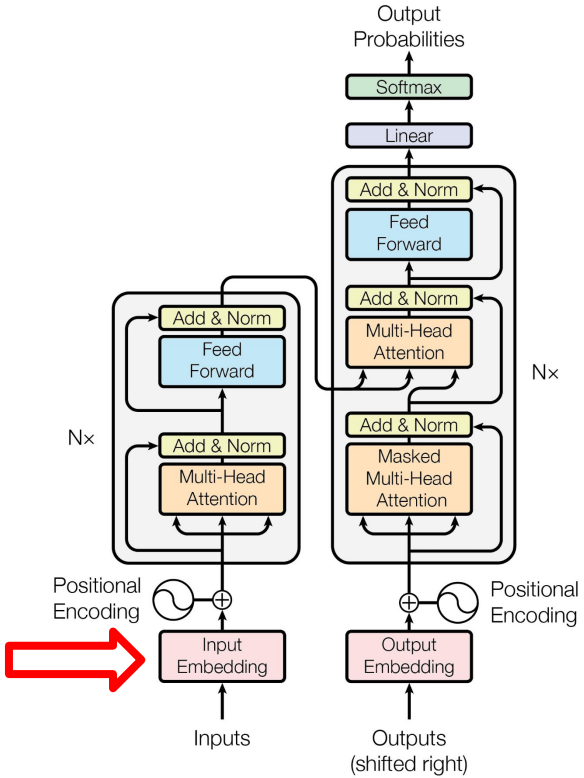
Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

*Equal contribution. Listing order is random. Jakob proposed replacing RNNs with self-attention and started the effort to evaluate this idea. Ashish, with Illia, designed and implemented the first Transformer models and has been crucially involved in every aspect of this work. Noam proposed scaled dot-product attention, multi-head attention and the parameter-free position representation and became the other person involved in nearly every detail. Niki designed, implemented, tuned and evaluated countless model variants in our original codebase and tensor2tensor. Llion also experimented with novel model variants, was responsible for our initial codebase, and efficient inference and visualizations. Lukasz and Aidan spent countless long days designing various parts of and implementing tensor2tensor, replacing our earlier codebase, greatly improving results and massively accelerating our research.

Transformer Architecture : Embedding



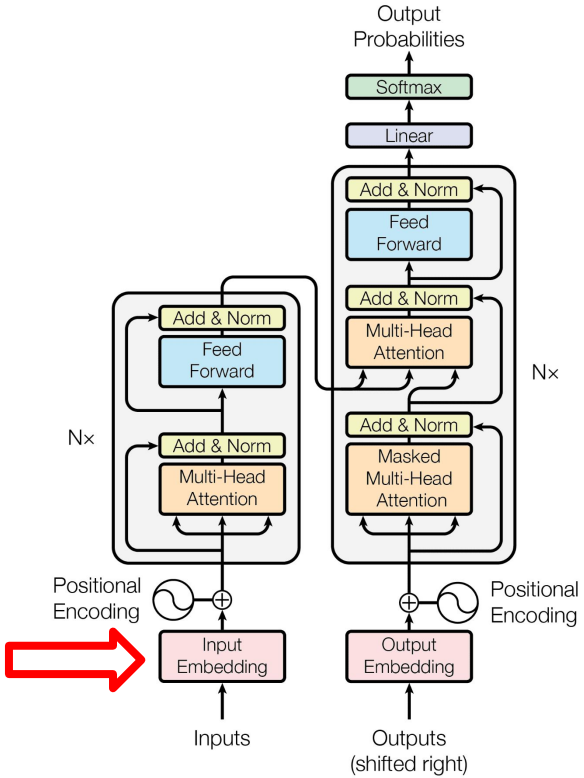
Hi

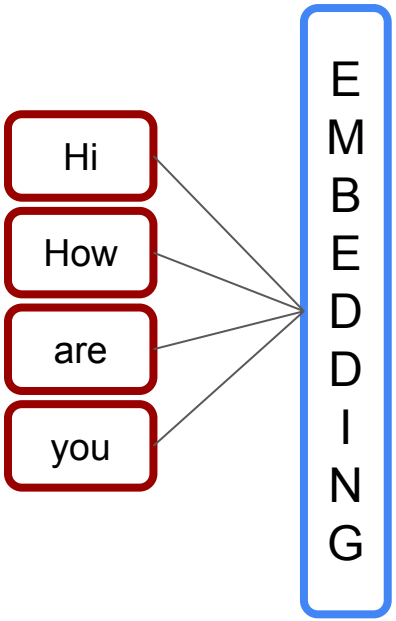
How

are

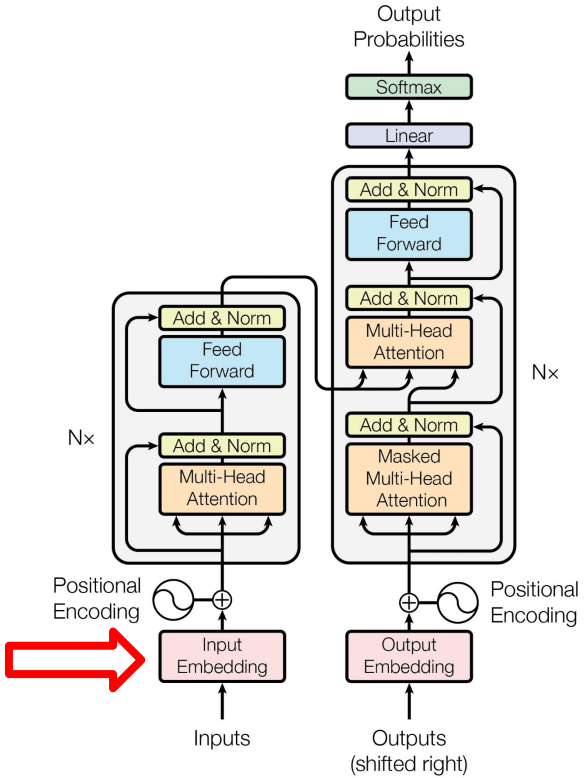
you

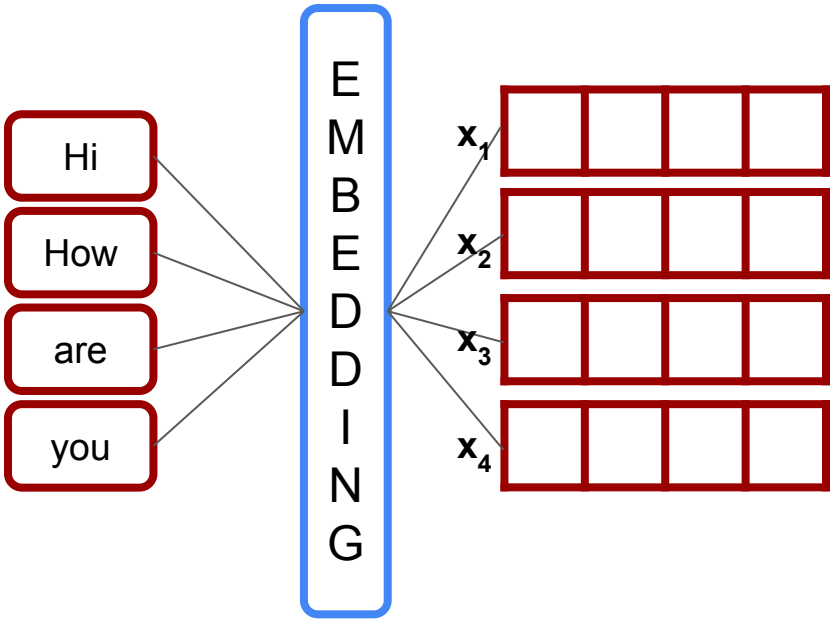
Transformer Architecture : Embedding



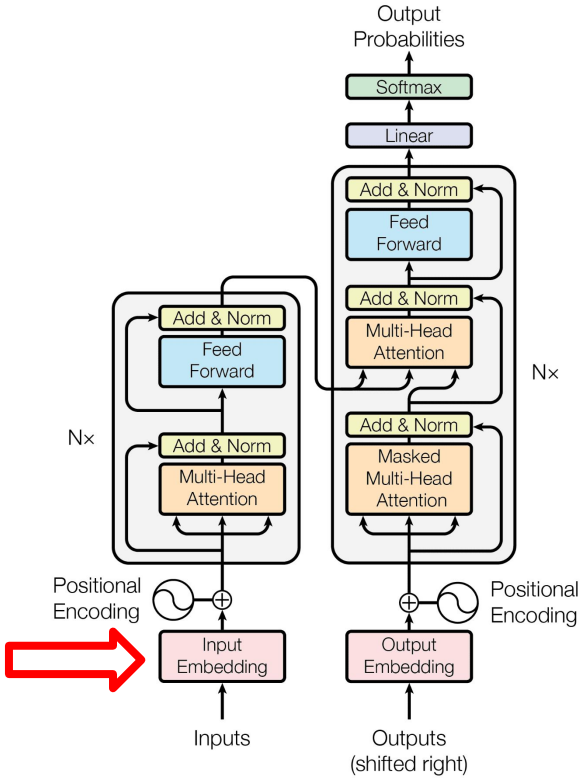


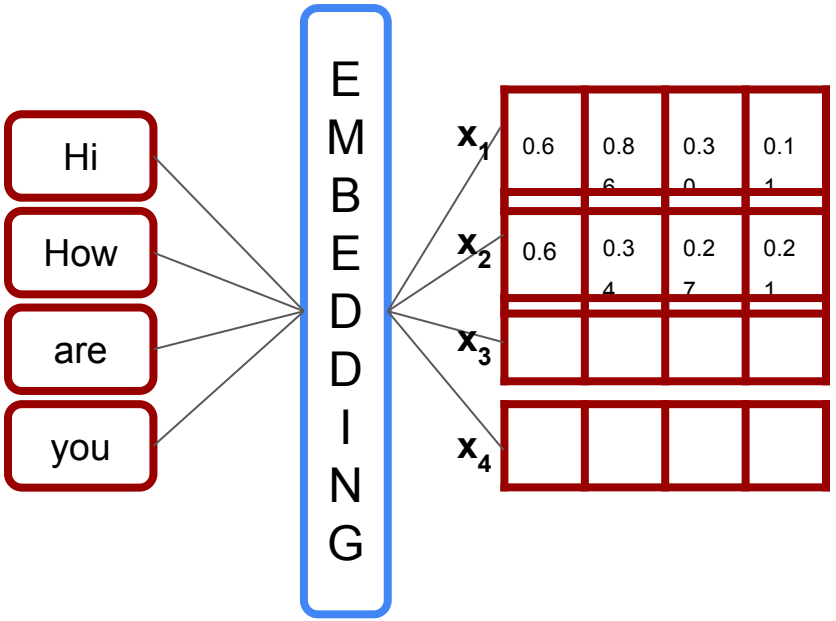
Transformer Architecture : Embedding



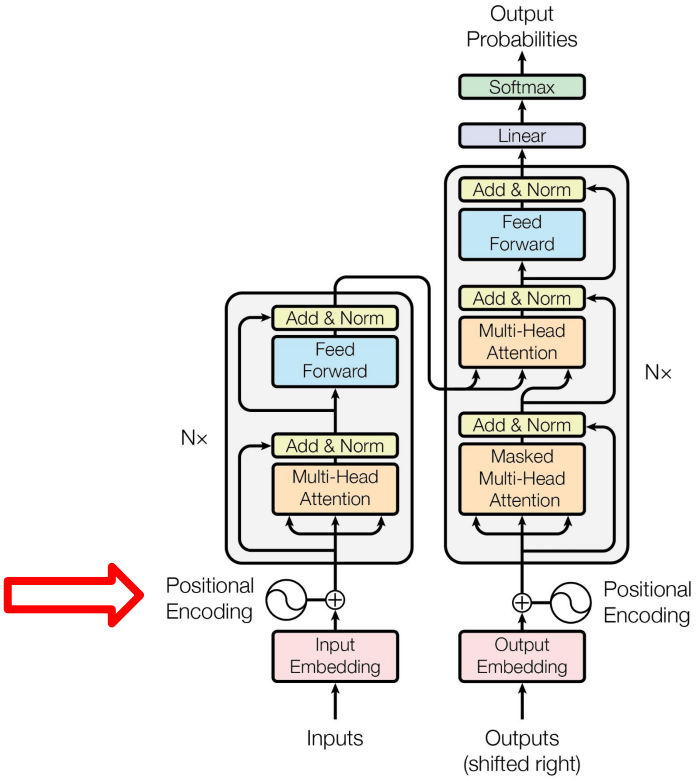


Transformer Architecture : Embedding





Transformer Architecture : Positional Encoding



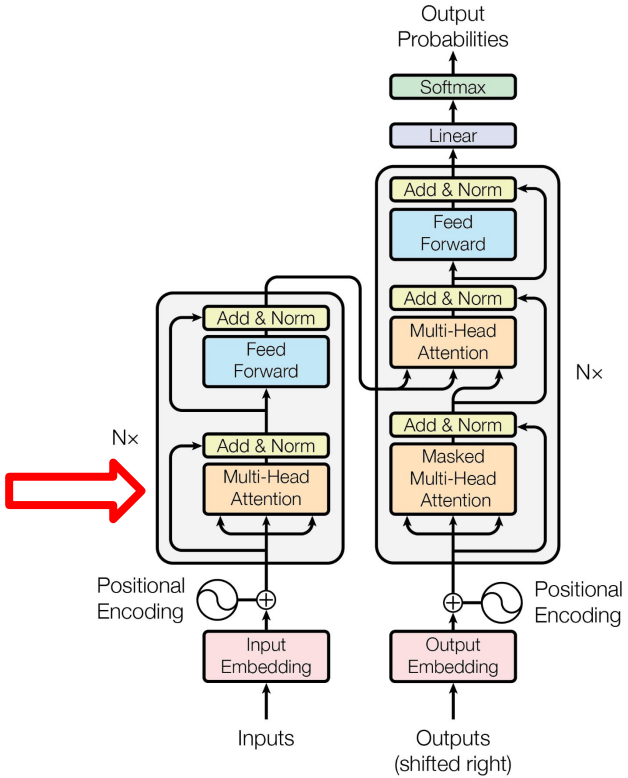
Binary encoding, not discrete but continuous: exploit sines and cosines of various frequencies

$$PE(pos, 2i) = \sin(pos / 10000^{(2i / d\_model)})$$

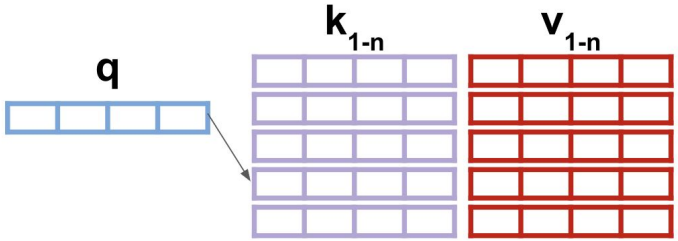
$$PE(pos, 2i+1) = \cos(pos / 10000^{(2i / d\_model)})$$

- `pos` is the position of the element in the sequence.
- `i` refers to the dimension within the positional encoding vector.
- `d_model` is the dimension of the input embeddings.

Transformer Architecture: Self-Attention

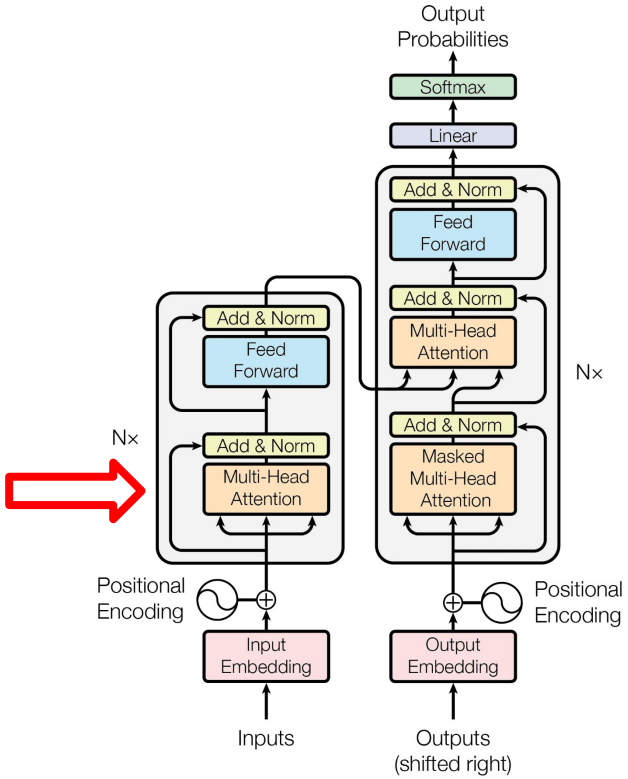


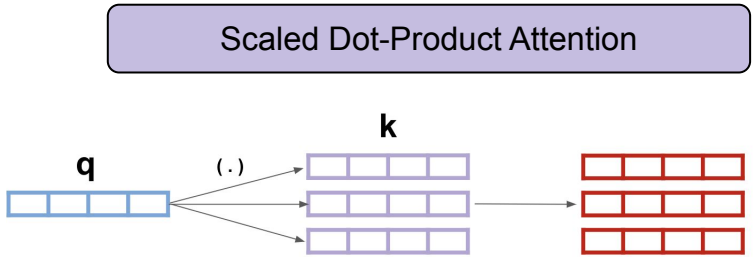
Query , Key and Value



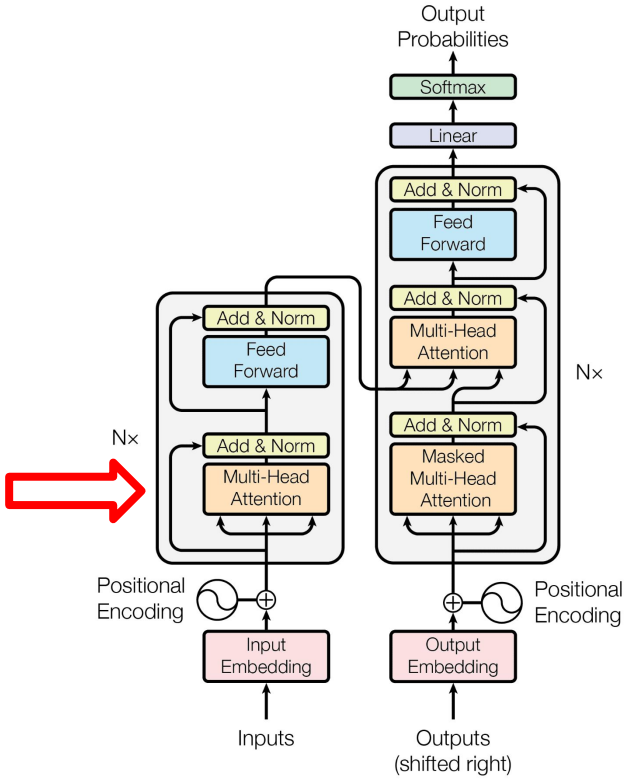
- Retrieval systems use Query, Key, and Value
- Compatibility of query with keys
- Manipulating values → reweighting the values based on the compatibility of q and k

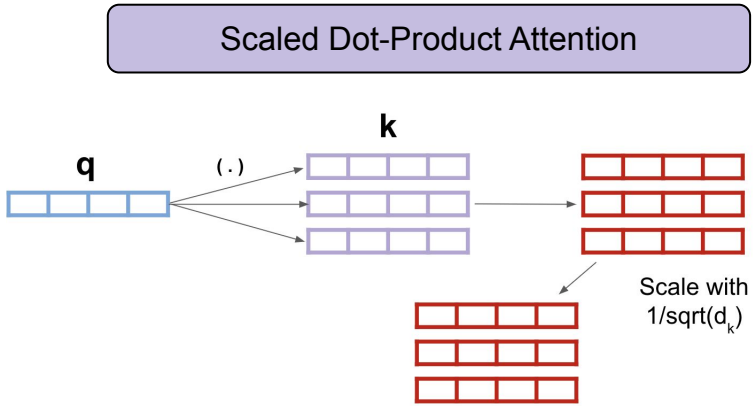
Transformer Architecture: Self-Attention



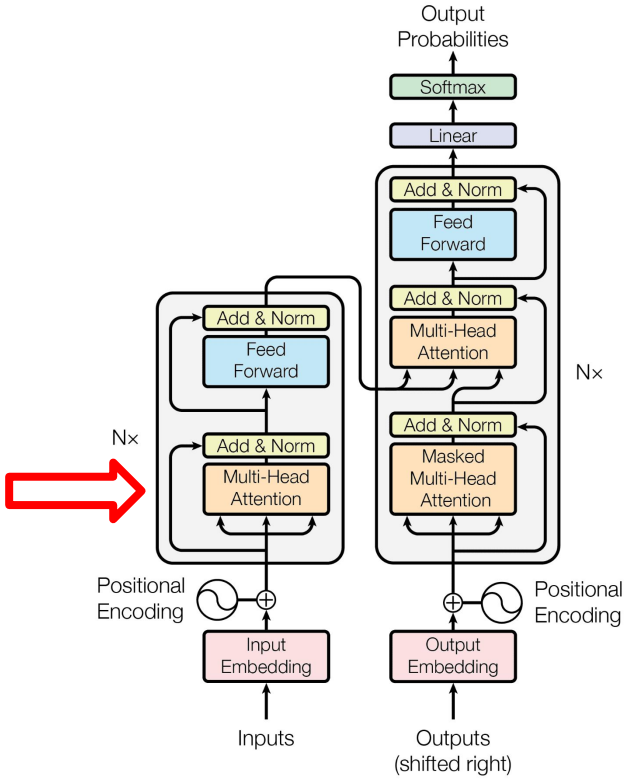


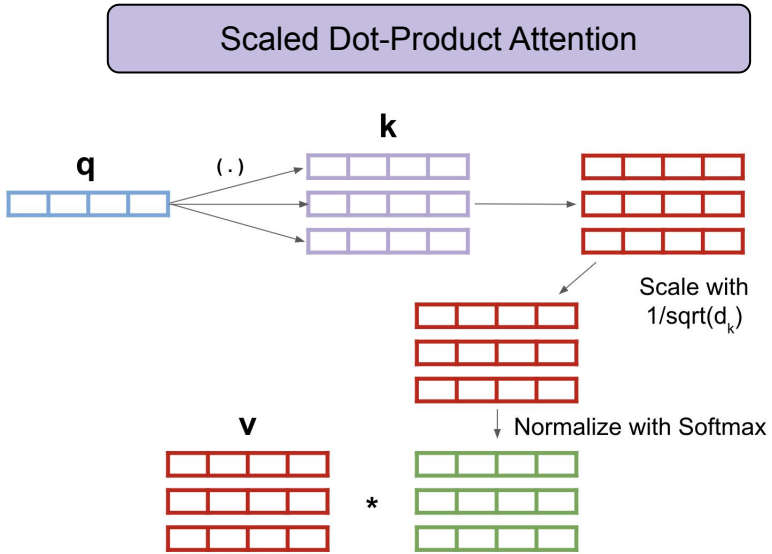
Transformer Architecture: Self-Attention



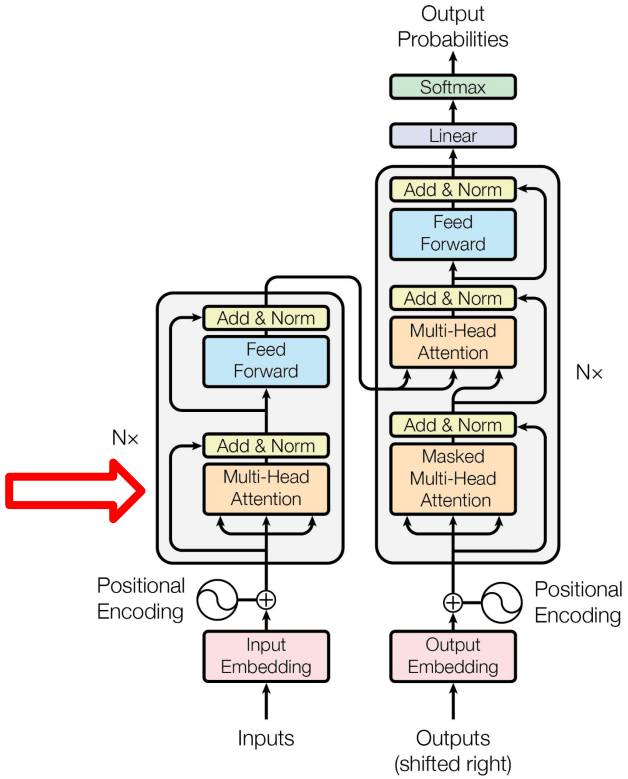


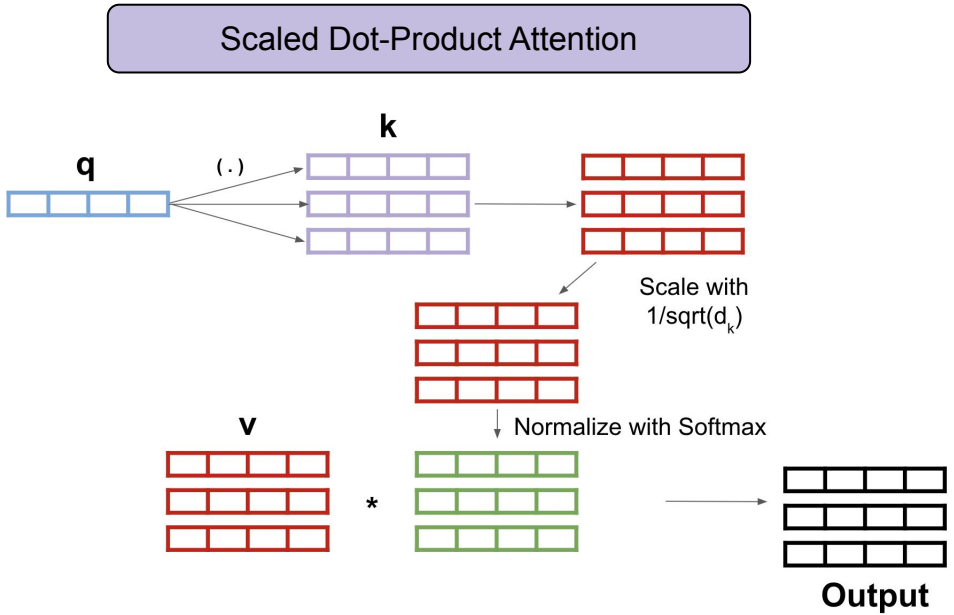
Transformer Architecture: Self-Attention



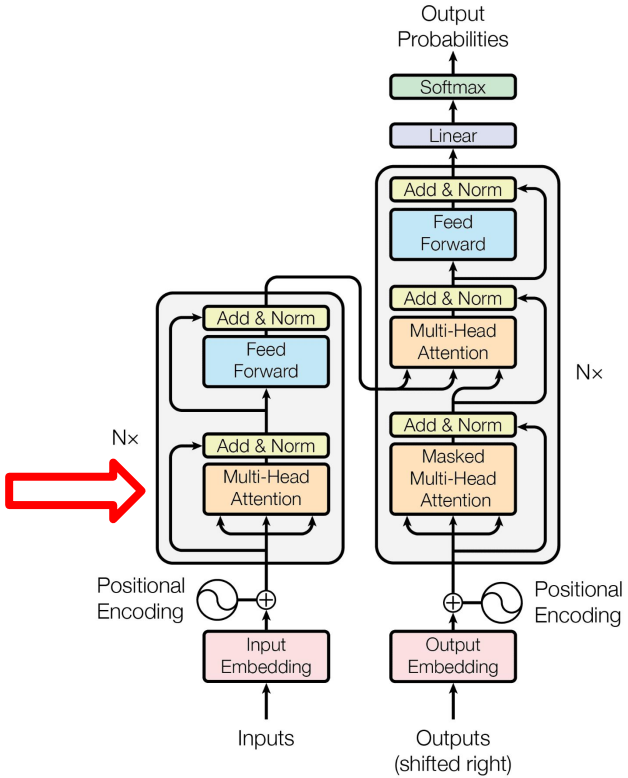


Transformer Architecture: Self-Attention

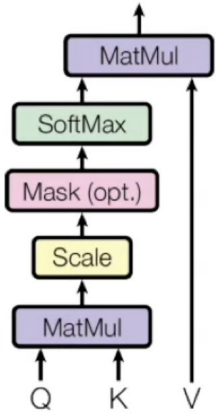




Transformer Architecture: Self-Attention



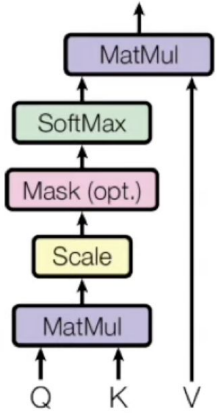
Scaled Dot-Product Attention



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

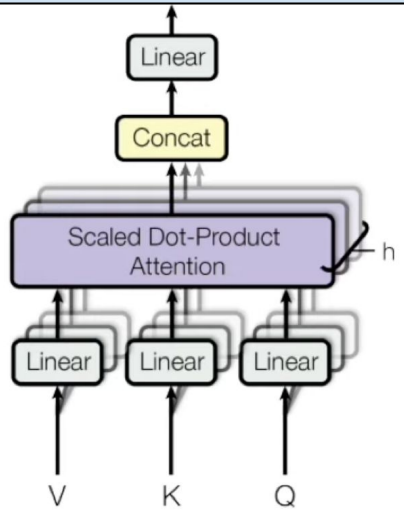
Transformer Architecture: Self-Attention

Scaled Dot-Product Attention



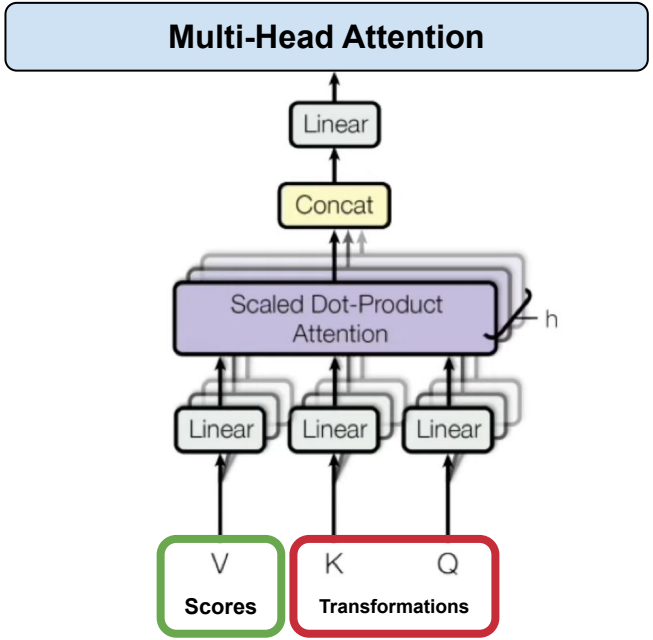
$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

Multi-Head Attention

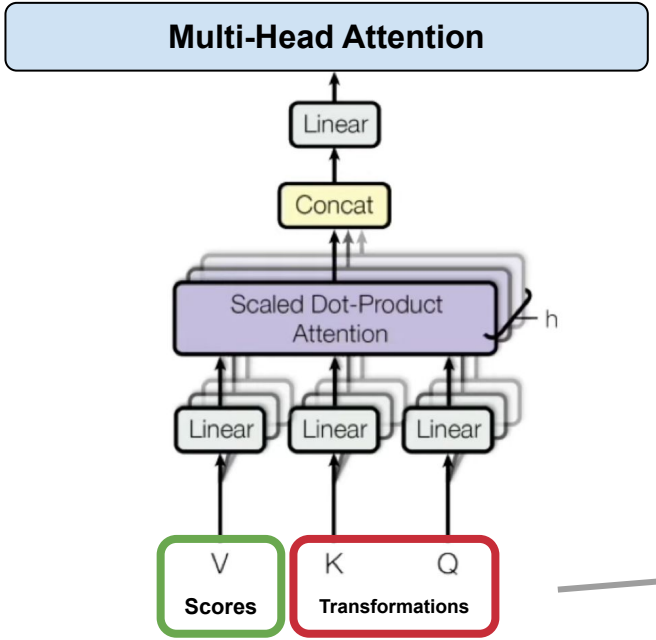


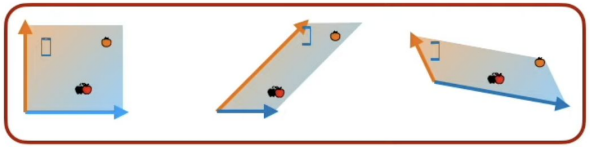
$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, ..., \text{head}_\text{h})W^O$$
$$\text{where head}_\text{i} = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

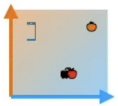
Transformer Architecture: Self-Attention



Transformer Architecture: Self-Attention

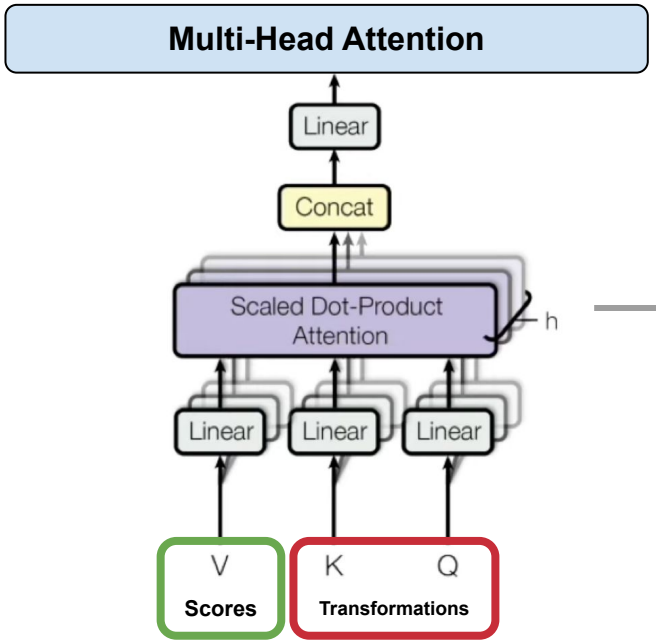


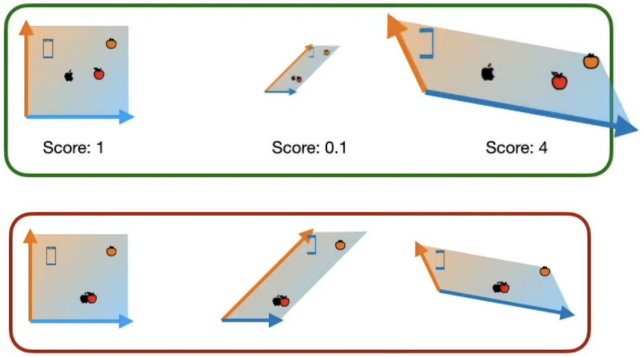


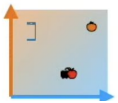


please buy an apple and an orange
apple unveiled the new phone

Transformer Architecture: Self-Attention

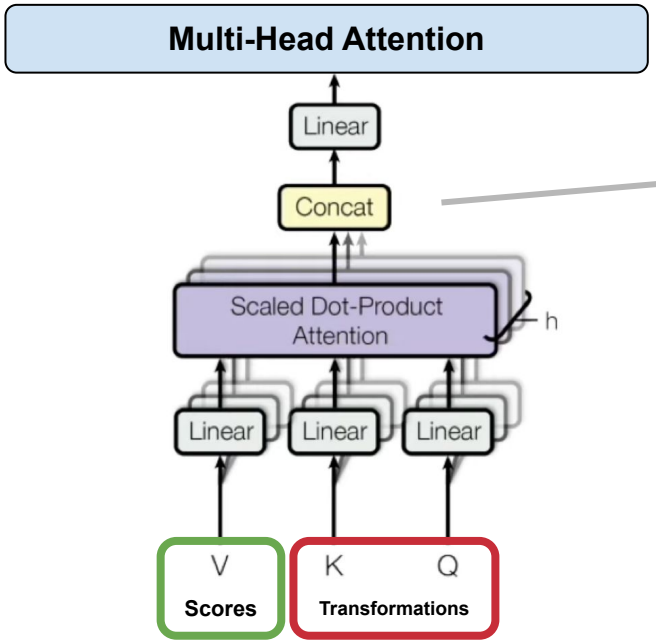


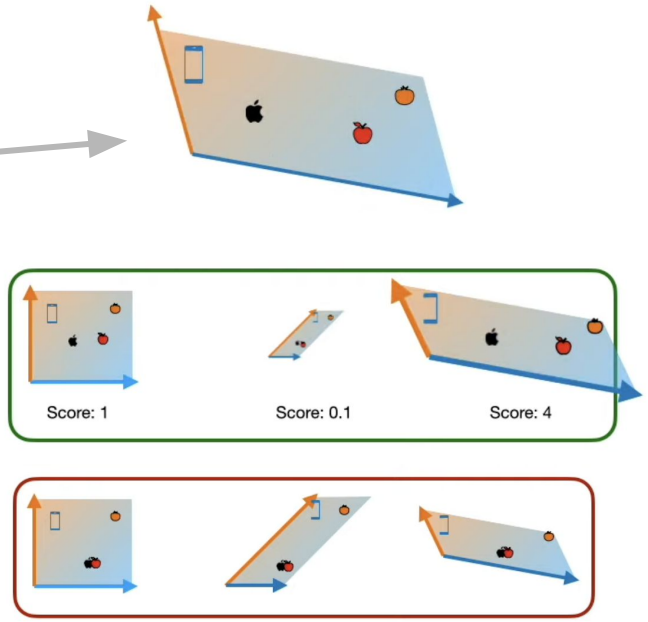


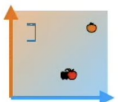


please buy an apple and an orange
apple unveiled the new phone

Transformer Architecture: Self-Attention







please buy an apple and an orange
apple unveiled the new phone

Evolution of Transformers

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

*Equal contribution. Listing order is random. Jakob proposed replacing RNNs with self-attention and started the effort to evaluate this idea. Ashish, with Illia, designed and implemented the first Transformer models and has been crucially involved in every aspect of this work. Noam proposed scaled dot-product attention, multi-head attention and the parameter-free position representation and became the other person involved in nearly every detail. Niki designed, implemented, tuned and evaluated countless model variants in our original codebase and tensor2tensor. Llion also experimented with novel model variants, was responsible for our initial codebase, and efficient inference and visualizations. Lukasz and Aidan spent countless long days designing various parts of and implementing tensor2tensor, replacing our earlier codebase, greatly improving results and massively accelerating our research.

†Work performed while at Google Brain.

Transformers

BERT **TFT**

GPT **Performer**

T5 **RoBERTa**

Electra

Natural Language Processing

LEAP

Evolution of Transformers

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

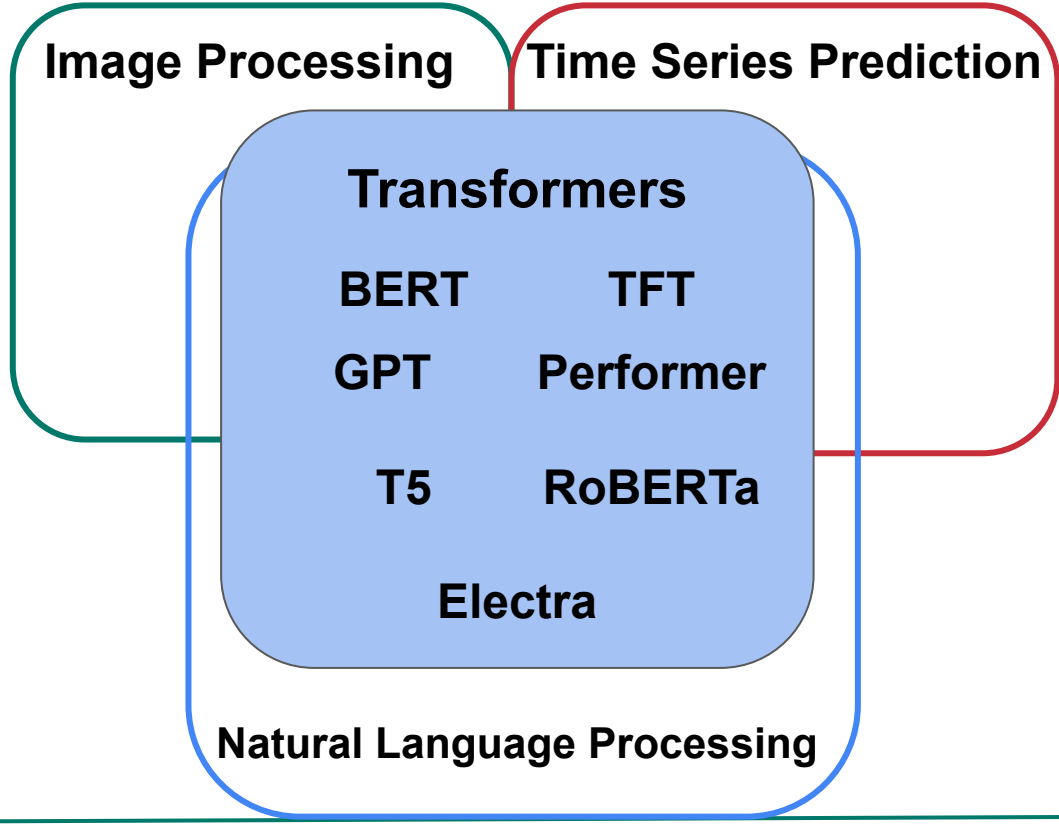
Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

*Equal contribution. Listing order is random. Jakob proposed replacing RNNs with self-attention and started the effort to evaluate this idea. Ashish, with Illia, designed and implemented the first Transformer models and has been crucially involved in every aspect of this work. Noam proposed scaled dot-product attention, multi-head attention and the parameter-free position representation and became the other person involved in nearly every detail. Niki designed, implemented, tuned and evaluated countless model variants in our original codebase and tensor2tensor. Llion also experimented with novel model variants, was responsible for our initial codebase, and efficient inference and visualizations. Lukasz and Aidan spent countless long days designing various parts of and implementing tensor2tensor, replacing our earlier codebase, greatly improving results and massively accelerating our research.



Evolution of Transformers

Encoder Only Transformers

Generating an output sequence is not required

- Classification tasks
- Anomaly Detection

Decoder Only Transformers

Generate sequences based on input/prompt

- Summarization
- Generation

Encoder-Decoder Transformers

- Time series forecasting
- Translation/transformation

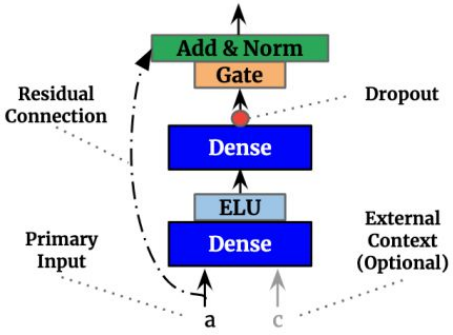
Transformer Variants in Time Series Prediction

- Informer
- Temporal fusion transformer
- AFNO Transformer
- Earthformer
- Fourcastnet
- Climformer
- ClimaX

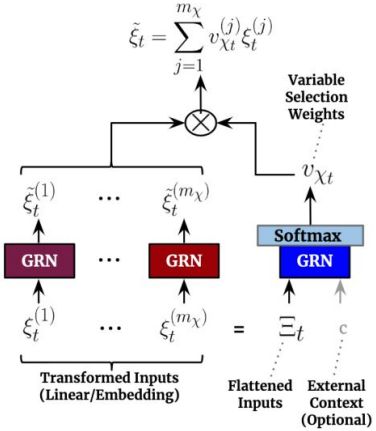
Transformer Variants in Time Series Prediction

Temporal fusion transformer

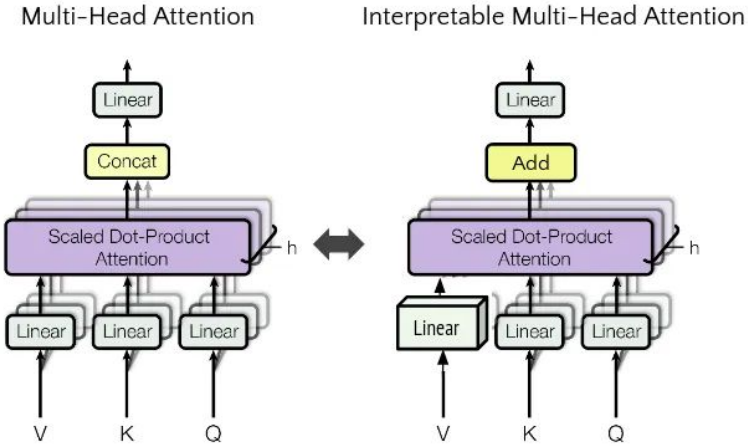
Interpretable Multi-horizon Multivariate forecasting



Gated Residual Network (GRN)



Variable Selection Network



- Skip unnecessary layers
- Improve generalization

- explicitly learn global importance weights of input features

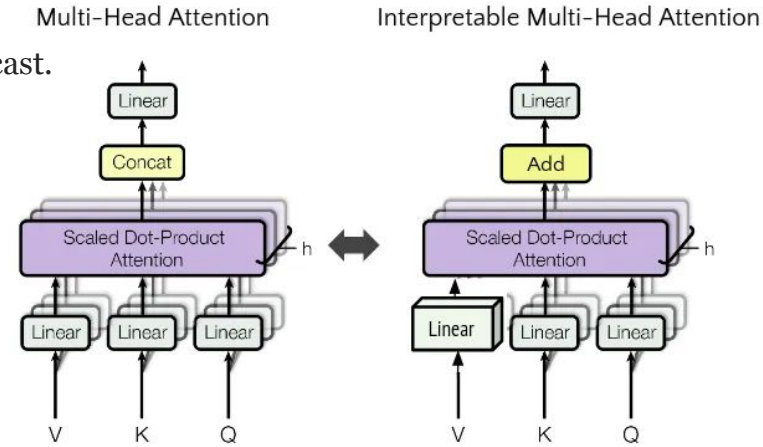
- Values are shared across all attention heads, this allows to easily trace back most relevant values.

Transformer Variants in Time Series Prediction

Temporal fusion transformer

Interpretable Multi-horizon Multivariate forecasting

1. Easily trace back most relevant past time-steps to predict each forecast.
2. Identify significant changes in temporal patterns.



Outperforming SHAP, LIME, and other post-hoc explainability methods in considering the time ordering of input features.

Transformer Variants in Time Series Prediction

Earthformer - Exploring Space-Time Transformers for Earth System Forecasting

- Hierarchical Encoder-Decoder Architecture
- Cuboid Attention

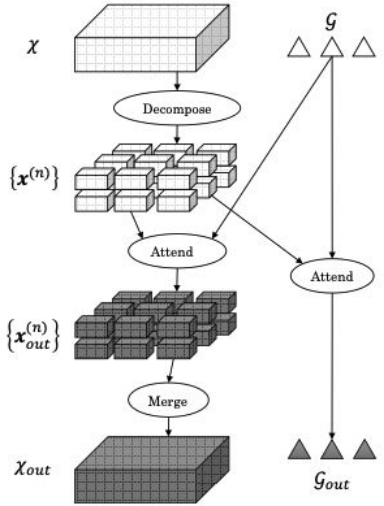


Figure 3: Illustration of the cuboid attention layer with global vectors.

Transformer Variants in Time Series Prediction

Earthformer - Exploring Space-Time Transformers for Earth System Forecasting

ClimFormer - A Spherical Transformer Model for Long-term Climate Projections

- Cloud climate feedback
- Adapted from AFNO (Adaptive Fourier Neural Operators) transformer to lean on a spherical grid
- Make climate projections under intervention scenarios.
- Fast high-resolution simulations

Transformer Variants in Time Series Prediction

Earthformer - *Exploring Space-Time Transformers for Earth System Forecasting*

ClimFormer - *A Spherical Transformer Model for Long-term Climate Projections*

Fourcastnet - *A global data-driven high-resolution weather model using adaptive fourier neural operators.*

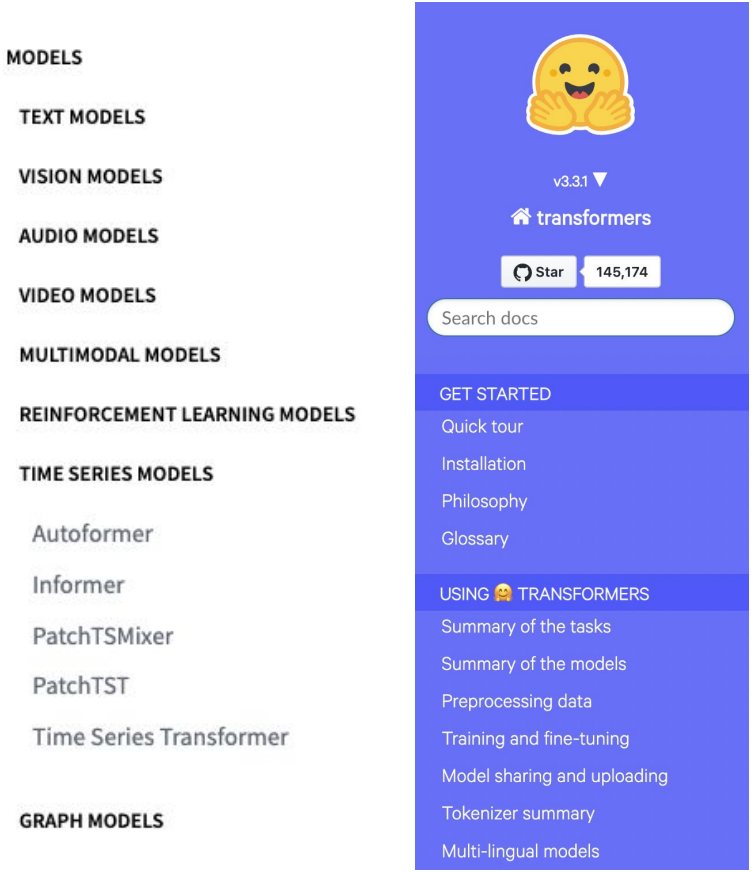
ClimaX - *A foundation model for weather and climate*

More applications in Turbulence, Land Models, Downscaling, Data Assimilation, etc.

How to get started

Transformers library on Hugging Face

- Dozens of architectures with over 2,000 pretrained models
- Online Demos and Example Notebooks
- Supports TF2.0 and Pytorch
- LLM inference optimization
- Many Attention Mechanism Variations for increased efficiency
  - LSH attention (Reformer)
  - Local Attention (Longformer)
  - Axial Attention (Multidimensional tasks)



Key Takeaways

- Transformers excel at handling complex spatiotemporal data, making them well-suited for climate and Earth system modeling.
- **Self-attention mechanisms enable transformers to capture long-range dependencies,** crucial for predicting large-scale climate processes.
- Transformers offer significant advantages over traditional machine learning models, such as **better scalability, parallelism, and accuracy in modeling multiscale phenomena.**
- **Applications in climate science include improved forecasting, downscaling, interpretability and data assimilation**
- Researchers can **leverage existing transformer architectures** and tools, such as Hugging Face or PyTorch, to experiment with these models in their work.