Transformer Models

León Villapún, Luis Alfredo   Lorencio Abril, Jose Antonio

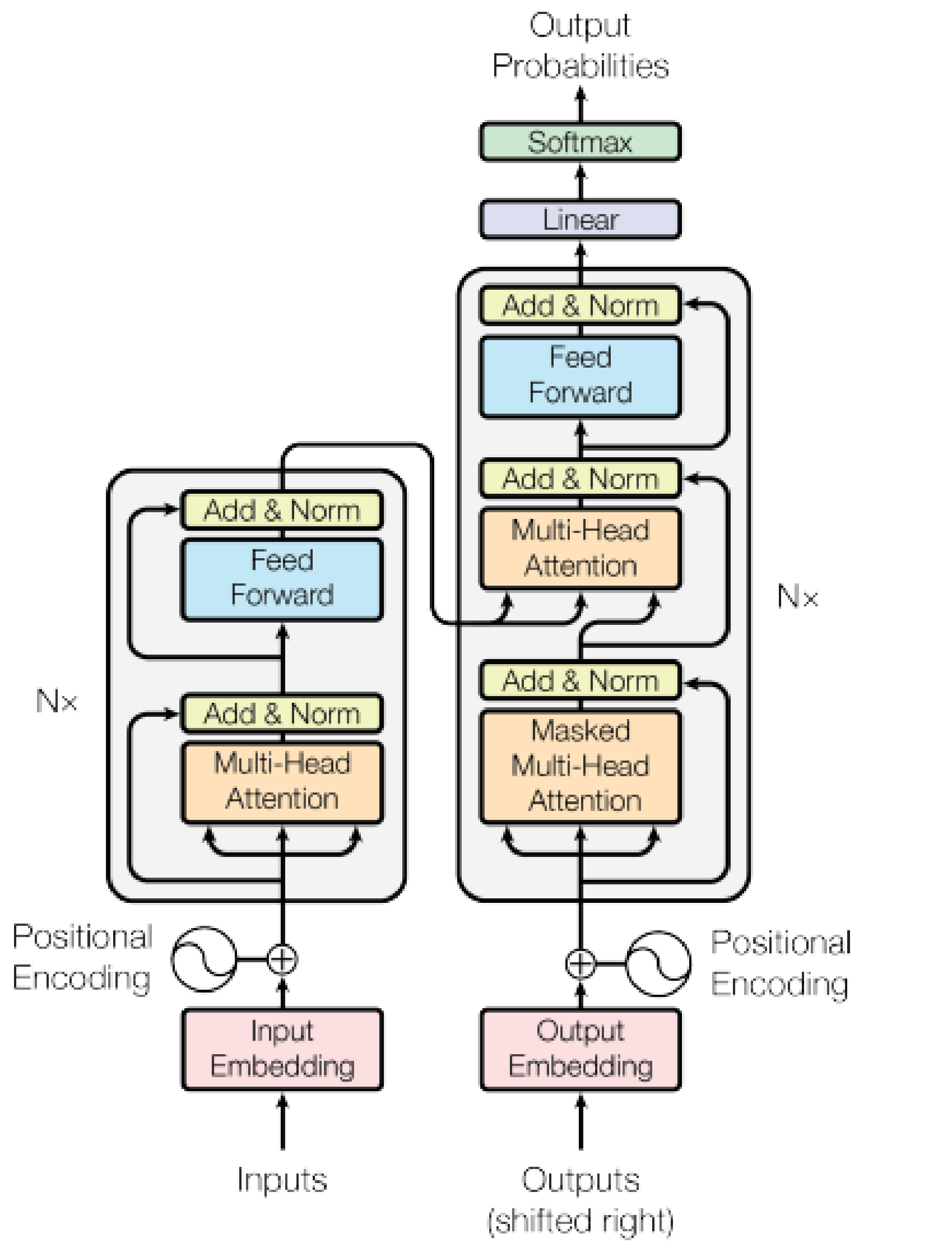
Abstract

This research provides an in-depth exploration of the Transformer model, a revolutionary deep learning architecture for processing sequential data. We trace the evolution from classical time series statistical methods to recurrent neural networks and then to Transformers, which introduced the powerful **self-attention mechanism** and position encoding. Our study delves into the inner workings of Transformers, their key advancements, and their impact across various applications. We also highlight the latest developments in **module-level improvements** and **pretrained models**, particularly in the context of Natural Language Processing (NLP), emphasizing the rising trend of Large Language Models (LLMs).

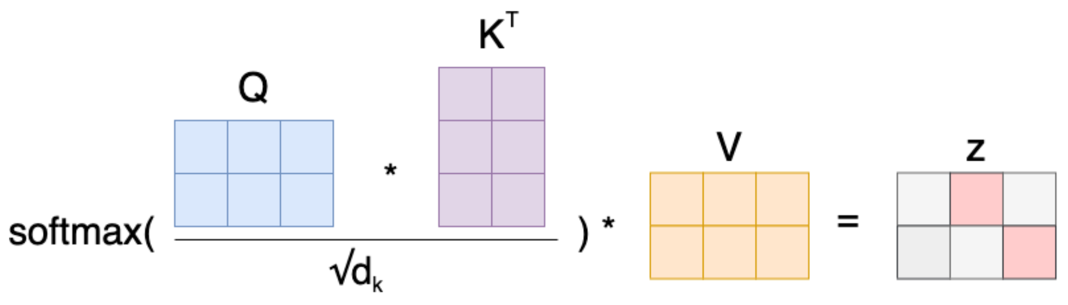
Attention is All You Need

Google introduced the Transformer model in the celebrated paper *Attention is All You Need*. There, they combine previously known concepts into a **novel architecture**. They combined **encoder-decoder stacks**, the attention mechanism as a **multi-head self-attention**, regular **feed-forward NNs**, **embeddings**, the **softmax** function and **positional encoding**. This new structure made it possible to increase the context span compared to previous models, achieving state of the art results in several NLP tasks.

Transformer Architecture



Attention Mechanism



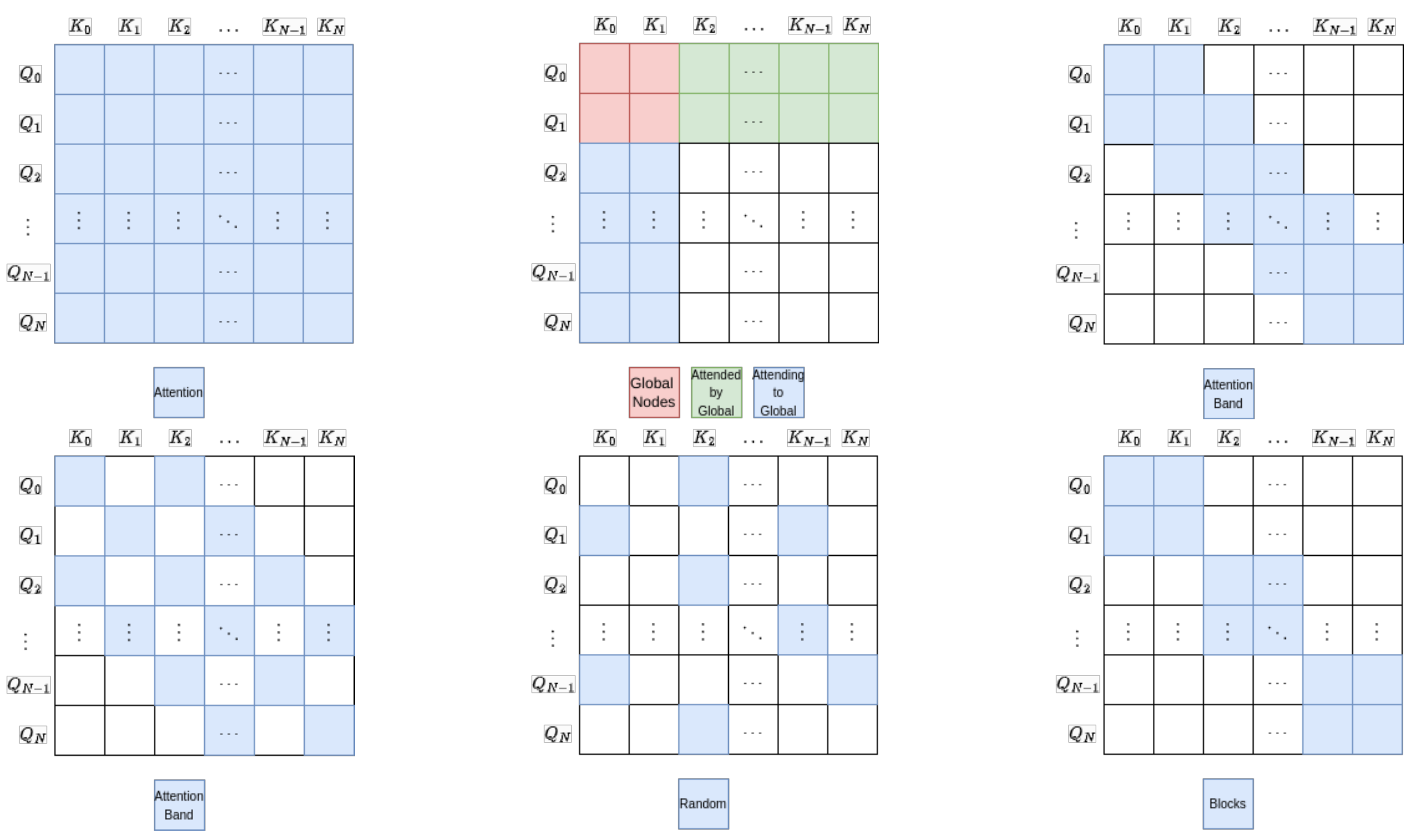
Pre-Trained Models

A **pretrained model** represents a machine learning model that has undergone prior training with a broader objective than the target task. In the context of Natural Language Processing, they are widely used to instantiate **Large Language Models**, with the goal of obtaining a general-enough model, and later on use techniques like **fine-tuning** to specialize the model in a particular endeavor. The introduction of the Transformer allows us to classify pretrained LLMs in three categories. **Encoder-only** models most representative example is Google's BERT and are used for tasks like **classification**. **Encoder-decoder** models include the original Transformer, however the most famous is BART. Their encoder-decoder stack allows to use them for **translation, question-answering, etc**. Finally, the **Decoder-only** models include the famous GPT family as well as LLaMA and Bard. They are the most followed architecture currently, and are used for tasks like **text-generation**, even though the newest models are **multi-modal**.

Sparse Attention

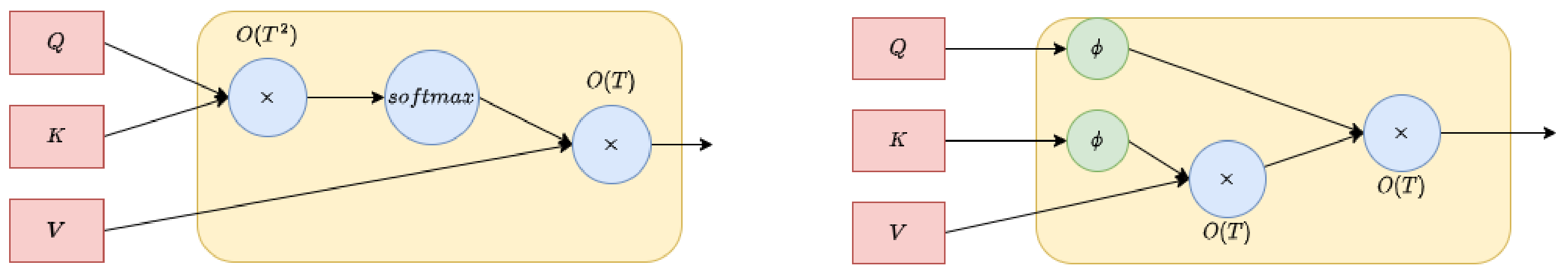
Sparse attention refers to a modification of the self-attention mechanism that allows the model to efficiently process and remember the past of the input sequence, especially when dealing with long sequences. It enhances the model's ability to establish appropriate context length and reduces the tendency to overfit in small datasets. It comes in two flavors, **atomic patterns**, which are patterns that can be directly applied, and **composed patterns**, as combinations of the previous ones.

Atomic Patterns



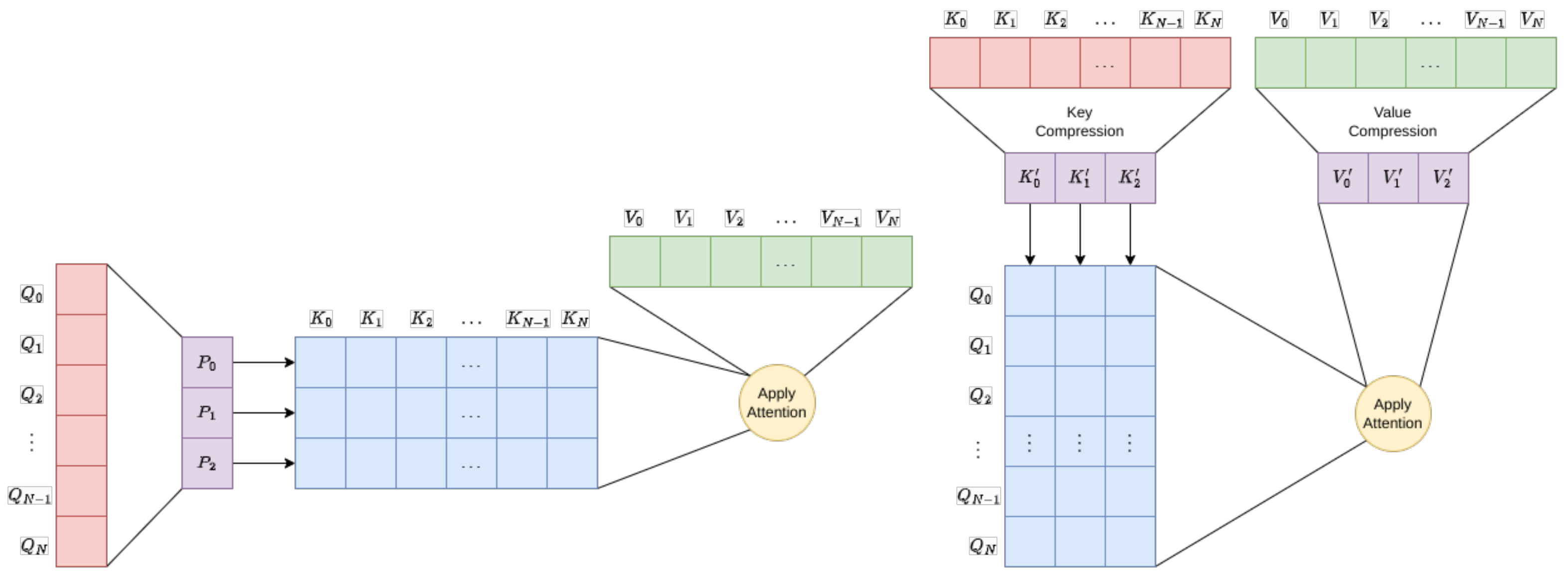
Linearized Attention

Linearized attention refers to a set of alternative approaches that aim to **reduce the computational complexity** of the self-attention mechanism in Transformer models from quadratic to linear, making it more efficient for processing long sequences. The first approach to achieve this leverages the mathematical properties of the dot product, but more sophisticated and varied methods have been developed after this.



Query Prototyping and Memory Compression

Query prototyping refers to reducing the number of queries, while **memory compression** is the reduction of key-value pairs. Both approaches aim at reducing the complexity of the attention mechanism.



Contact Information
luis.alfredo.leon@estudiantat.upc.edu
jose.antonio.lorencio@estudiantat.upc.edu

Pre-Trained Models Demo

Full Paper