Module 3 Glossary: Transformers in Keras

Welcome! This alphabetized glossary contains many terms used in this course. Understanding these terms is essential when working in the industry, participating in user groups, and participating in other certificate programs.

Term	Definition
Autoregressive integrated moving average (ARIMA)	A time series forecasting model that is used to predict future data points by combining three components: autoregression, differencing, and moving averages.
Bidirectional Encoder Representations from Transformers (BERT)	A deep learning model in which every output element is connected to every input element. The weightings between them are dynamically calculated based on their connection.
Convolutional neural networks (CNNs)	A type of deep learning model designed to process and analyze visual data by automatically detecting patterns through convolutional layers.
Computer vision	A field of AI that enables machines to interpret and understand visual information from various content such as images and videos.
Cross-attention mechanism	A model that allows the model to focus on relevant parts of the input sequence while generating output.
Decoder	A transformer decoder is a neural network architecture used in natural language processing tasks like machine translation and text generation. It generates output text by combining with an encoder to process input text.
Deep learning	A branch of artificial intelligence (AI). Deep learning is a subset of machine learning that uses large, multi-layered neural networks to automatically learn and make predictions from complex data.
Embed	In transformers, embedding is the technique of converting input tokens into dense, continuous vectors that represent their semantic meaning within the model.
Encoder	Encoders are neural network layers that process the input sequence and produce a continuous representation of the input. The transformer encoder architecture is the basis for many state-of-the-art models in natural language processing tasks.
Generative pre-trained transformers (GPT)	The neural network-based language prediction models that are built on the Transformer architecture. They analyze natural language queries, known as prompts, and predict the best possible response based on their understanding of language.
Image recognition	A software's ability to identify and classify people, objects, places, writing, and actions in digital images and video.
Image processing	A technique of manipulating and analyzing digital images to enhance, extract information, or convert them into a different format.
Long short-term memory networks (LSTMs)	A type of recurrent neural network (RNN) designed to capture and maintain long-term dependencies in sequential data.
Layer normalization	A technique used in Transformer neural networks to normalize the input values of all neurons in a layer for each data sample.
Natural language generation	The use of artificial intelligence to produce spoken or written human-like text.
Natural language processing (NLP)	Branch of artificial intelligence that enables computers to understand, manipulate, and generate human language (natural language).
Parallelization	In transformers, parallelization refers to the ability to process multiple elements of a sequence simultaneously, rather than sequentially, to speed up training and inference.
Recurrent neural network (RNN)	A deep learning model that is trained to process and convert a sequential data input into a specific sequential data output.
Reinforcement learning	An area of machine learning where an agent learns to make decisions by taking actions in an environment to maximize cumulative rewards.
Self-attention mechanisms	Mechanisms in transformers that allow each element of a sequence to dynamically focus on and weigh the importance of other elements in the sequence to capture context and dependencies.
Sequence	In transformers, a sequence refers to an ordered set of tokens, such as words, that are processed together as input to capture dependencies and contextual information across the entire sequence.
Sequential data	Data that is ordered in a specific sequence, where the arrangement of elements matters, such as time series, audio, or text.
Speech recognition	A technology that converts spoken language into text by analyzing and interpreting audio signals.
Vision transformers (ViTs)	A type of neural network architecture that applies transformer models to image analysis, treating image patches as sequences to capture visual patterns.
Transformers	A technology that can leverage self-attention mechanisms to process input data in parallel, making them highly efficient and powerful.