Glossary: Language Modeling with Transformers

Welcome! This alphabetized glossary contains many of the terms in this course. This comprehensive glossary also includes additional terms not used in course videos. These terms are essential for you to recognize for better comprehension of the concepts covered in the course.

Estimated reading time: 3 minutes

Term	Definition
Add and norm	A process that enhances the depth of the model while mitigating potential issues with gradients.
Argmax process	A process that helps obtain the translated word's token index and further replace the self-attention mechanism.
Attention mechanism	A neural network component that weighs input elements during processing and focuses on relevant parts of output generation.
Autoregressive model	A model that facilitates sequence generation by anticipating each new token based on the sequence's preceding tokens.
BERT	An open-source, deeply bidirectional, unsupervised language representation pretrained using a plain text corpus.
Contextual embeddings	A type of embedding that aptly describes how the transformer processes the input word embeddings by accounting for the context in which each word occurs within the sequence.
Data loader	A utility in a machine learning framework that collects operational data from data sources at regular intervals.
Decoder models	A type of network architecture commonly used in sequence-to-sequence tasks.
Fine-tuning	A supervised process that optimizes the initially trained GPT model for specific tasks, like QA classification.
Generative pre-training (GPT)	A self-supervised model that involves training a decoder to predict the subsequent token or word in a sequence.
Language models	A model that predicts words by analyzing the previous text, where context length acts as a hyperparameter.
Masked language modeling	A model that learns tasks by reconstructing sentences with words that have been obscured.
Masking	A function used to perform masking operations on tokens.
Multi-head attention	An attention that executes several scaled dot-product attention processes in parallel.
Next Sentence Prediction	A training that enables the model to understand how sentences relate.
One-hot encoding approach	An encoding approach that maps categorical features to binary representations, which are used to map the feature in a matrix or vector space.
Orthogonality	An object-relational database in which the various parts work naturally.

Term	Definition
Positional encoding	A technique in natural language processing and deep learning used to embed sequence positions into data.
PyTorch	A software-based open-source deep learning framework used to build neural networks, combining Torch's machine learning library with a Python-based high-level API.
Python dictionary	A built-in data structure that stores key-value pairs and facilitates efficient lookup and manipulation.
Reinforcement Learning from Human Feedback (RLHF)	A model that represents a fine-tuning approach and enhances model performance on specific tasks, proving particularly effective in chatbot development.
Scaled dot-product attention	A mechanism within the transformer model fundamentally involves a series of matrix multiplications incorporating queries, keys, and a scaling factor to prevent the dot product from becoming too large.
Self-attention mechanism	A mechanism that calculates weights for every word in a sentence, enabling the model to predict words that are likely to be used in sequence.
Semantic	A term used in natural language processing referring to language's meaning and interpretation.
Simple language modeling	A neural network architecture crucial for natural language understanding, predicting subsequent words in sentences.
Softmax function	A mathematical function that converts raw scores into probabilities in machine learning.
Tokenization	The process of converting the words in the prompt into tokens.
Transformer model	A model that can translate text and speech in near real-time.
Vector	A mathematical object represented by a group of numbers commonly used in machine learning algorithms.