Embeddings and Vectors in NLP

Introduction

In Natural Language Processing (NLP), embeddings and vectors are fundamental concepts that allow machines to understand and process human language. Words, phrases, and even entire documents are represented as numerical vectors, enabling algorithms to perform tasks such as classification, translation, and sentiment analysis.

What Are Vectors?

A vector is a numerical representation of data. In NLP, vectors are used to represent words in a way that captures their semantic meaning. Each word is mapped to a point in a high-dimensional space, where similar words are located close to each other.

Word Embeddings

Word embeddings are techniques used to convert words into dense vectors of fixed size. These vectors capture semantic relationships between words. For example, the vectors for 'king' and 'queen' will be similar, with differences that reflect gender. Popular embedding techniques include Word2Vec, GloVe, and FastText.

Static vs Contextual Embeddings

Static embeddings assign a single vector to each word regardless of context. For example, Word2Vec and GloVe produce static embeddings. Contextual embeddings, such as those from BERT and GPT, generate different vectors for the same word depending on its usage in a sentence.

How Embeddings Are Trained

Embeddings are typically trained using large corpora of text. Models learn to predict a word based on its surrounding words (context window). For example, Word2Vec uses two approaches: Continuous Bag of Words (CBOW) and Skip-Gram. CBOW predicts a word from its context, while Skip-Gram predicts context from a word.

Dimensionality Reduction

Word vectors are often high-dimensional, which can be computationally expensive. Dimensionality reduction techniques like PCA (Principal Component Analysis) and t-SNE are used to reduce the number of dimensions while preserving the relationships between words. This helps in visualization and efficient processing.

Comparison of Embedding Techniques

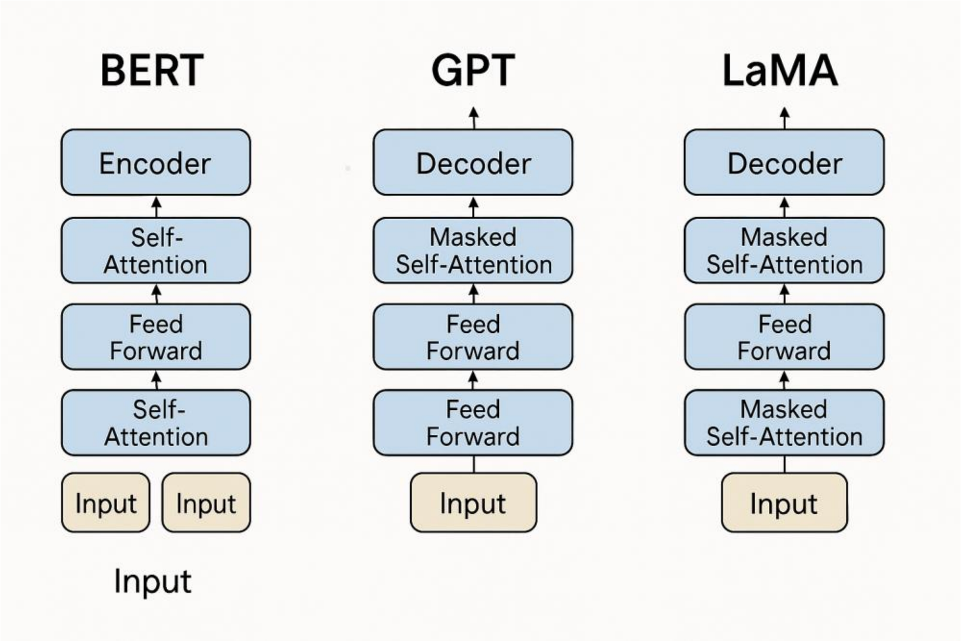
Technique	**Type**	**Characteristics**
Word2Vec	Static	Predicts words using CBOW or Skip-Gram
GloVe	Static	Uses word co-occurrence statistics
FastText	Static	Considers subword information
BERT	Contextual	Generates embeddings based on sentence context
GPT	Contextual	Generates embeddings during text generation

Applications in NLP

Embeddings are used in a wide range of NLP tasks:

- Text classification
- Sentiment analysis
- Machine translation
- Named entity recognition
- Question answering
- Text summarization

Visual Representation of Embeddings


BERT
Encoder
Self-Attention
Feed Forward
Self-Attention
Input
Input
Input
GPT
Decoder
Masked Self-Attention
Feed Forward
Feed Forward
Input
LaMA
Decoder
Masked Self-Attention
Feed Forward
Masked Self-Attention
Input