Paper: LLM2Vec: Large Language Models Are Secretly Powerful Text Encoders

Why did I choose this Paper:

With the increasing popularity of text-based machine learning, text data can be a robust input for models. LLM2Vec represents a significant advancement in text embedding, enabling decoder-only language models to produce rich, contextualized text representations, which are crucial for various NLP and machine learning tasks. This is also relevant in cheminformatics for developing models using text data, including spectral data, SELFIES, and SMILES.

Problem:
LLM2Vec responds to the current underutilization of decoder-Only LLMs for text embedding, relative to tokenization. Despite the model's success in NLP tasks, decoder models are not widely adopted for text embedding takes, which require contextualized embeddings of text. Text embeddings tasks require models that can provide, deep, refined representation of words and sequences. Current methods require significant adaptions, such as expensive training processes or synthetic data, to achieve optimal performance. LLM2Vec is a step up from traditional decoder-only models due to the application of Bidirectional Attention (Bi), Masked Next Token Prediction (MNTP), and Unsupervised Contrastive Learning (SimCSE). Traditional decoder-only models process text unidirectionally, either left to right or visa versa. Enabling bidirectional attention allows the model to consider both the preceding and following context simultaneously, like how encoder models like BERT operate. MNTP predicts certain tokens in the input sequence are masked, and the model is trained to predict these masked tokens based on the surrounding context. By predicting masked tokens, the model learns to use context more effectively, enhancing its ability to generate meaningful and contextually accurate embeddings. SimCSE or contrastive learning by creating positive pairs through data augmentation and treating other instances as negative examples. This enhances the model's ability to distinguish between different texts, improving the quality of the embeddings. By combining these techniques, the model achieves state-of-the-art performance on various text embedding benchmarks, ensuring high-quality results.

Equations used:

$$\mathcal{L} = \frac{e^{\lambda s(q,d^+)}}{e^{\lambda s(q,d^+)} + \sum_{d^- \in N} e^{\lambda s(q,d^-)}},$$

The Contrastive Loss: This formula represents the contrastive loss function used in contrastive learning to train models for distinguishing between similar and dissimilar data points. q: Represents the query or anchor example. d^+: Represents a positive example, which is similar to

the query. d^-: Represents negative examples, which are dissimilar to the query. s(q, d): A similarity function that measures the similarity between the query q and a document d. λ: A scaling factor that controls the sharpness of the probability distribution. N: The set of negative examples.