Lecture Notes

Lecture Notes on Understanding Multi-Layer Perceptrons in Large Language Models

Introduction

- This lecture explores how large language models (LLMs) store knowledge, particularly focusing on the role of multi-layer perceptrons (MLPs) in this process.
- The goal is to demystify the computations within MLPs and illustrate how they can encode specific facts, using the example of Michael Jordan and basketball.

Understanding Large Language Models

The Basics of Language Models

- Language models are trained to predict the next word in a sequence based on the preceding context.
- They operate on text that is tokenized into smaller units (tokens), each represented by high-dimensional vectors.

The Role of Multi-Layer Perceptrons

- MLPs are crucial components of the architecture of transformers, which are the backbone of modern AI models.
- They perform computations that involve matrix multiplications and non-linear transformations, which are essential for encoding complex relationships in data.

High-Dimensional Space

- Vectors in LLMs exist in high-dimensional spaces where different directions can represent various meanings or features.
- For example, the relationship between the embeddings of ”woman” and ”man” illustrates how gender information can be encoded in this space.

Key Concept: Information Encoding

- Each vector must encode more than just the meaning of a single word; it needs to incorporate contextual information and general knowledge learned during training.
- MLPs provide the capacity to store facts, such as the relationship between Michael Jordan and basketball.

The Mechanics of Multi-Layer Perceptrons

Step-by-Step Computation

- The input vector representing a token (e.g., "Michael Jordan") undergoes a series of operations within the MLP.
- Each vector is processed independently, allowing for parallel computations.

Matrix Multiplication

- The first operation involves multiplying the input vector by a large matrix filled with learned parameters (weights).
- This matrix can be thought of as containing various "questions" that probe different features of the input vector.

Non-Linear Activation: ReLU

- After the linear transformation, a non-linear activation function, typically the Rectified Linear Unit (ReLU), is applied.
- ReLU outputs zero for negative values and retains positive values, effectively mimicking the behavior of an AND gate.

Down Projection and Final Output

- The output from the ReLU is then multiplied by another matrix (down projection) and combined with a bias term.
- The final output vector is the sum of the transformed vector and the original input vector, allowing the model to encode additional information.

Example: Encoding "Michael Jordan Plays Basketball"

- The MLP is set up to recognize the specific fact that "Michael Jordan plays basketball" by associating specific directions in the high-dimensional space with each component of the fact.
- The computations ensure that if the input vector encodes both "Michael" and "Jordan," the output will include the direction representing "basketball."

Additional Analysis

- The operations within MLPs are foundational to how LLMs function, but the interpretation of individual neurons is complex.
- The concept of superposition suggests that neurons may not represent single features but rather combinations of features, complicating the interpretability of these models.
- Understanding the high-dimensional nature of the embeddings helps explain why LLMs can scale effectively, as they can represent more features than there are dimensions in the space.

Conclusion

- The lecture provided a detailed examination of how MLPs in large language models process information and store facts.
- Key takeaways include the significance of high-dimensional spaces, the role of matrix multiplications, and the importance of non-linear transformations in encoding knowledge.

General Tips

- To better understand LLMs, focus on the interplay between linear and non-linear operations within MLPs.
- Consider exploring the implications of high-dimensional spaces in machine learning and how they can influence model performance and interpretability.
- Engage with practical examples and visualizations to solidify your understanding of these complex concepts.