Weights & Biases

Products Resources Docs Pricing Enterprise Log in Sign up

Articles Projects ML News Events Podcast Courses

W&B Fully Connected > ML News

ImageBind-LLM: Combining ImageBind and Llama!

Brett Young

Last Updated: Sep 11, 2023

Researchers have unveiled ImageBind-LLM, a new approach to multi-modality instruction tuning for large language models (LLMs), broadening their utility across diverse data types like audio, video, and 3D point clouds. Building on existing LLaMA models, this new framework uses a unique method to align and extend multi-modal understanding, setting a new standard for what we can expect from next-generation AI systems.

Full Multi-Modality

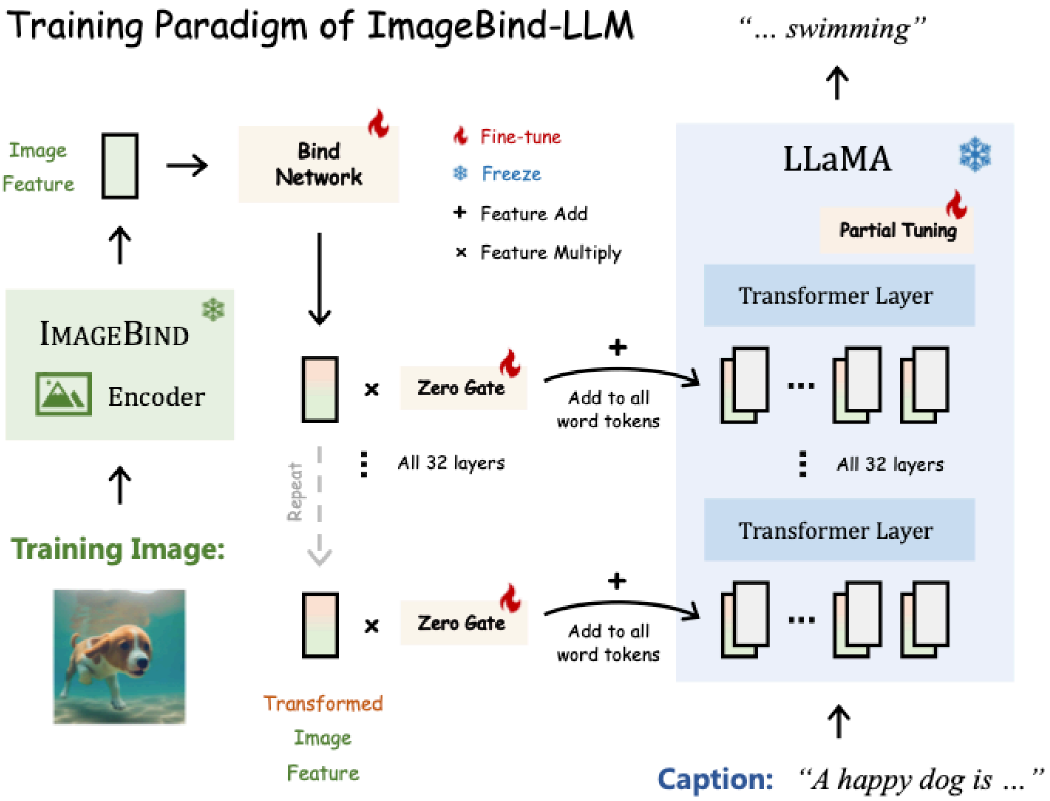
Up until now, instruction-tuning primarily revolved around text and images. ImageBind-LLM raises the bar by offering capabilities to handle a range of data types, including not just text and images, but also audio, video, and 3D point clouds. This is made possible through the integration of a "bind network," a learnable system that aligns ImageBind's encoder with the LLaMA language model.

Overview

We will delve into the main components of the system. It consists of several key elements, including the Bind Network for feature alignment, a Attention-Free Mechanism for reduced computational complexity, and two inference strategies: Naive and Cache-enhanced.

Bind Network: Explanation

The Bind Network acts as a bridge between ImageBind, which handles image features, and LLaMA, a language model. It aligns the features from both domains for better multimodal learning, which is essential for tasks that need understanding of both text and images. This is a basic alignment between the high-dimensional image feature space and LLaMA's feature dimension. The end result is a transformed image feature that shares the same feature space as LLaMA's word embeddings.



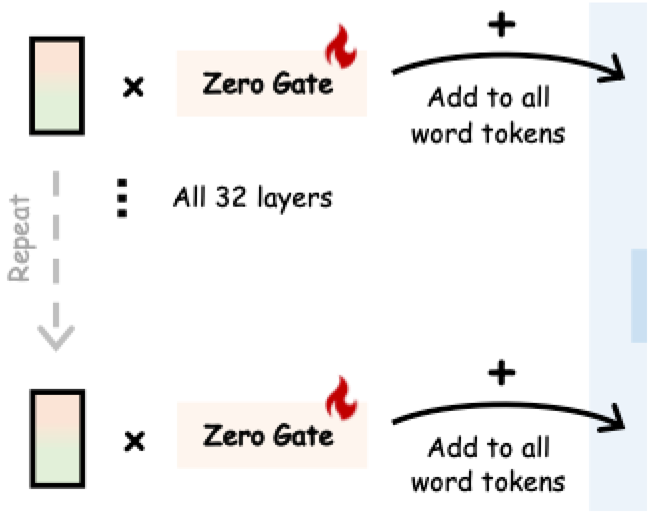
A visual representation of the Binding network

Attention-Free Mechanism

The team also introduced an attention-free, zero-initialized gating mechanism, reducing computational burden and training complexity. Unlike previous models that utilized attention mechanisms to integrate visual knowledge, this novel approach is simpler, yet highly effective. The image features transformed by the bind network are directly added to the language model, allowing for more efficient and stable learning.

Attention-free Zero-initialized Injection: Explanation

Unlike previous models that use attention mechanisms, **this method directly adds the transformed image feature to every word token in LLaMA.** This makes the model computationally more efficient. It also uses a learnable gating factor that starts at zero and allows the model to control how much of the image feature affects the language feature. **The zero-initialization helps the model during the early stages of training.**



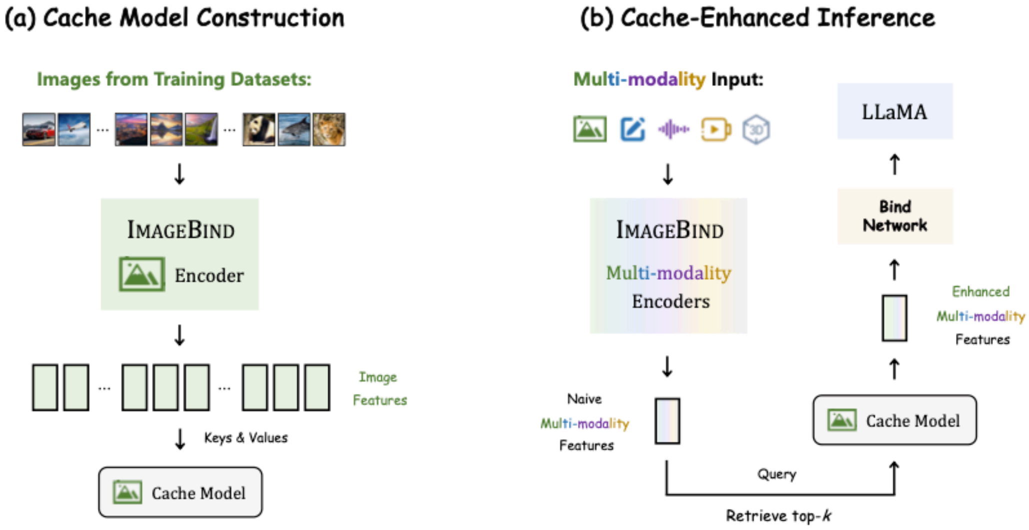
Zero Gating!

The Cache Model for Enhanced Inference

The Naive Multi-modality Inference and the Cache-enhanced Inference are two approaches to handle multi-modal data, such as images, text, audio, and video.

The Naive approach is straightforward: it uses the Bind Network to align image features from ImageBind with LLaMA's word tokens. Once aligned, these features are simply passed into LLaMA for further processing and response generation. However, it has limitations. Specifically, it does not account for discrepancies that might arise when different types of encoders are used for different modalities during inference. For example, if you wanted to use a different image encoder, the feature spaces may not align perfectly, affecting the model's performance.

The Cache-enhanced Inference model aims to resolve these limitations. It introduces a cache that stores a subset of previously computed image features. During inference, this cache is used to find similar features that augment the current input. Essentially, it acts as a bridge between the training and inference stages, minimizing any inconsistency that might arise.



A visual representation of the cache model

The two methods are related in that they both aim to facilitate multi-modal data processing. However, they differ in their approach to handling the challenges that arise during inference. While the Naive approach offers a straightforward but limited solution, the Cache-enhanced model addresses the limitations by providing a more nuanced and adaptable strategy. This makes the Cache-enhanced model more robust and reliable for real-world applications.

Bridging the Gap with Fine-Tuning

The team opted for a two-stage training pipeline. The first stage focuses on aligning ImageBind with LLaMA using large-scale image-caption data. The second stage involves partial fine-tuning of LLaMA's parameters to equip it with instruction-following abilities, while keeping other modules frozen. They also introduced parameter-efficient methods like Low-Rank Adaptation and bias-norm tuning for this stage, ensuring an optimal balance between performance and resource utilization.

High-Quality Instruction Tuning

The team didn’t stop at fine-tuning. They took it a step further by introducing an additional instruction tuning stage using high-quality data. This fine-grained tuning aims to rectify the occasional issue where the model hallucinates about objects that aren’t present in the input, enhancing the model's reliability.

Overall

By incorporating a unique blend of innovative techniques, it not only broadens the range of modalities that can be handled but also does so in a computationally efficient manner. The code is available on GitHub as well!!

The paper: https://arxiv.org/pdf/2309.03905.pdf

Tags: ML News

Created with ❤️ on Weights & Biases.

https://wandb.ai/byyoung3/ml-news/reports/ImageBind-LLM-Combining-ImageBind-and-Llama---Vmlldzo1MzY1NzYx

QUICKSTART

Documentation

RESOURCES

Courses | Forum | Tutorials | Benchmarks

W&B

About Us | Authors | Contact | Terms of Service | Privacy Policy