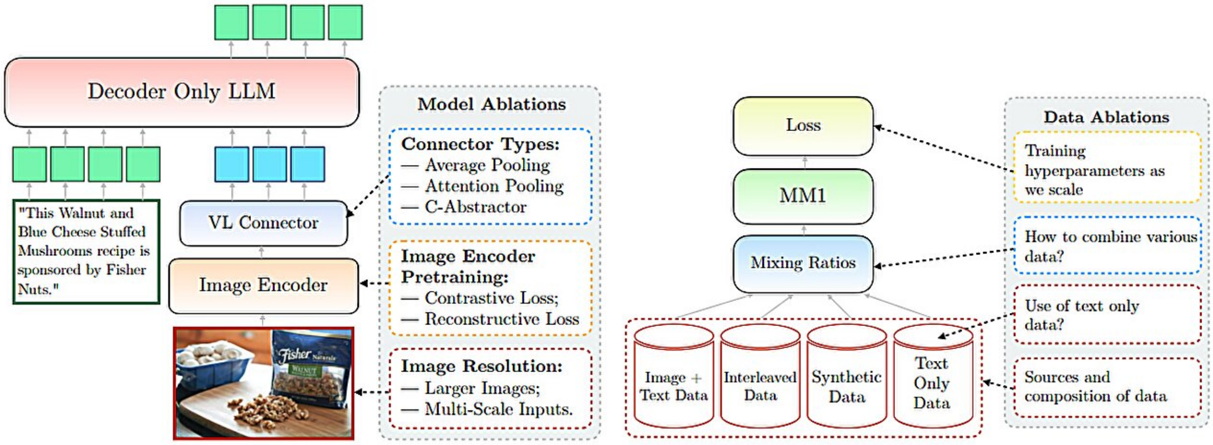
TechXplore

Apple's MM1: A multimodal large language model capable of interpreting both images and text data

March 19 2024, by Bob Yirka



Left: Model ablations: what visual encoder to use, how to feed rich visual data, and how to connect the visual representation to the LLM. Right: Data ablations: type of data, and their mixture. Credit: *arXiv* (2024). DOI: 10.48550/arxiv.2403.09611

A team of computer scientists and engineers at Apple has developed an large language model (LLM) that the company claims can interpret both images and data. The group has posted a paper to the *arXiv* preprint server describing their new MM1 family of multimodal models and test results.

Over the past year, LLMs have received a lot of press for their advanced AI capabilities. One company notably absent from the conversation is Apple. In this new effort, the research team makes it clear that the company is not interested in simply adding an LLM developed by another company (currently they are negotiating with Google to add Gemini AI tech to Apple devices); instead, they have been working to develop a next-generation LLM, one that can interpret both images and text data.

Multimodal AI works by integrating and processing different types of data inputs, such as visual, auditory and textual information. This integration allows the AI to have a more comprehensive understanding of complex data, leading to more accurate and context-aware interpretations than single-mode AI systems.

Apple's research team claims they have made major advancements in using multimodal AI with their MM1 models, which integrate text and image data to improve capabilities in image captioning, visual question answering and query learning. Their MM1 is part of what they describe as a family of multimodal models, each of which include as many as 30 billion parameters.

Such models, the researchers note, make use of datasets comprising image-capture pairs, documents that include images and text-only documents. The researchers further claim that their multimodal LLM (MLLM) can count objects, identify objects that are part of an image, and use common sense about everyday objects to offer users useful information about what the image presents.

The researchers also claim that their MLLM is capable of in-context learning, which means it does not need to start over every time a question is asked; it uses what it has learned in the current conversation. The team provides examples of the advanced capabilities of their models—one includes uploading an image of a group of friends at a bar holding a menu and asking the model how much it would cost to buy a beer for everyone based on prices listed in the menu.

More information: Brandon McKinzie et al, MM1: Methods, Analysis & Insights from Multimodal LLM Pre-training, *arXiv* (2024). DOI: 10.48550/arxiv.2403.09611

Citation: Apple's MM1: A multimodal large language model capable of interpreting both images and text data (2024, March 19) retrieved 27 April 2024 from https://techxplore.com/news/2024-03-apple-mm1-multimodal-llm-capable.html