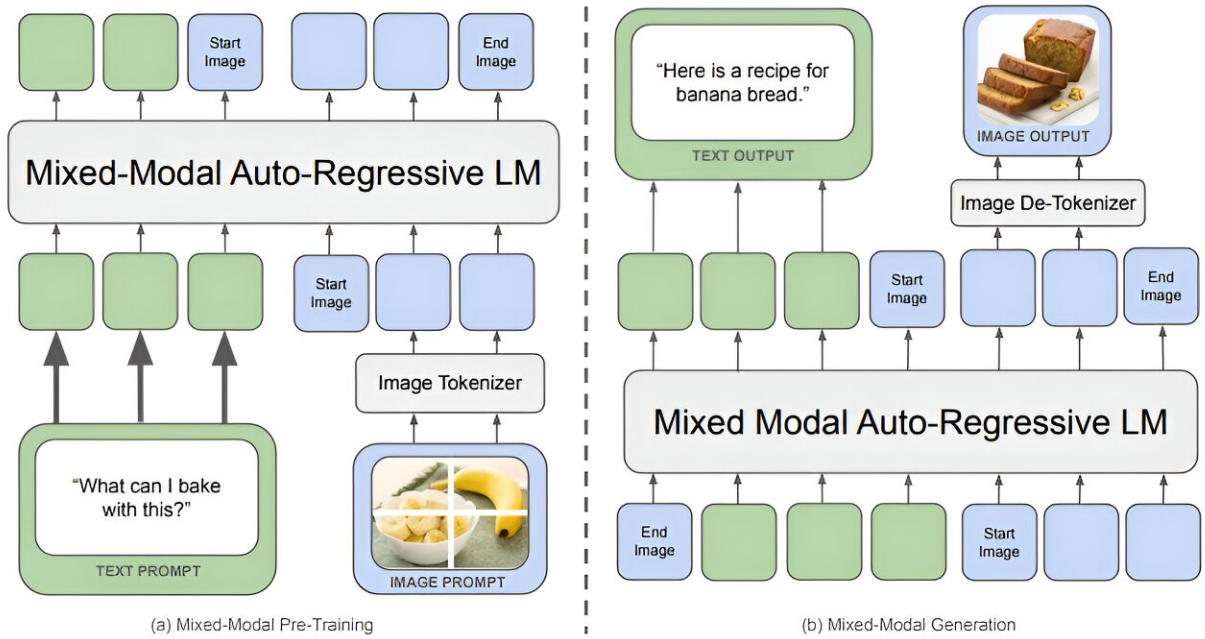
Meta introduces Chameleon, an early-fusion multimodal model

May 22 2024, by Bob Yirka



Chameleon represents all modalities—images, text, and code, as discrete tokens and uses a uniform transformer-based architecture that is trained from scratch in an end-to-end fashion on ~10T tokens of interleaved mixed-modal data. As a result, Chameleon can both reason over, as well as generate, arbitrary mixed-modal documents. Text tokens are represented in green and image tokens are represented in blue. Credit: *arXiv* (2024). DOI: 10.48550/arxiv.2405.09818

AI researchers at Meta, the company that owns Facebook, Instagram, WhatsApp, and many other products, have designed and built a

multimodal model to compete with the likes of Google's Gemini.

Called Chameleon, the new system is built on an early fusion architecture, and because of that it is able to comingle multiple inputs in ways not possible with most other systems.

The group, called the Chameleon Team, has written a paper describing their new model, including its architecture and how well it has performed during testing. It is posted on the *arXiv* preprint server.

AI multimodal models, as their name implies, are applications that are able to accept more than one type of input during a query—a user can submit a picture of a horse, for example, while also asking how many of its breed have won the Kentucky Derby.

To date, most such models have processed such data as separate entities in the early part of processing and then later brought them together to look for associations—a technique called late fusion.

Such an approach has been found to work well, but has limitations regarding integration. To overcome this, the team at Meta has based their model on early-fusion architecture.

This architecture allowed the team to interweave associations from the get-go. They accomplished this by converting images to tokens similar to the way LLMs parse words. The team also added the ability to use a unified vocabulary of tokens from different sources, including images, code or text—and they claim this allowed for applying transformative computing with mixed types of input data.

The researchers note that unlike Gemini, Chameleon is an end-to-end model, which made the need for image decoders unnecessary. They also developed and used new types of training techniques to allow their

model to work with multiple types of tokens—ones that involved two-stage learning and a massive dataset of approximately 4.4 trillion texts, images, or pairs of tokens along with interleaved data. The system was trained using 7 billion and then 34 billion parameters over 5 million hours on a high-speed GPU.

The result, the research team claims, is a model that can accept text only, images only, or a combination of both and return intelligent answers and associations with better accuracy than its competitors.

More information: Chameleon: Mixed-Modal Early-Fusion Foundation Models, *arXiv* (2024). DOI: 10.48550/arxiv.2405.09818

Citation: Meta introduces Chameleon, an early-fusion multimodal model (2024, May 22) retrieved 22 June 2024 from https://techxplore.com/news/2024-05-meta-chameleon-early-fusion-multimodal.html