Large Language Models

Potpourri

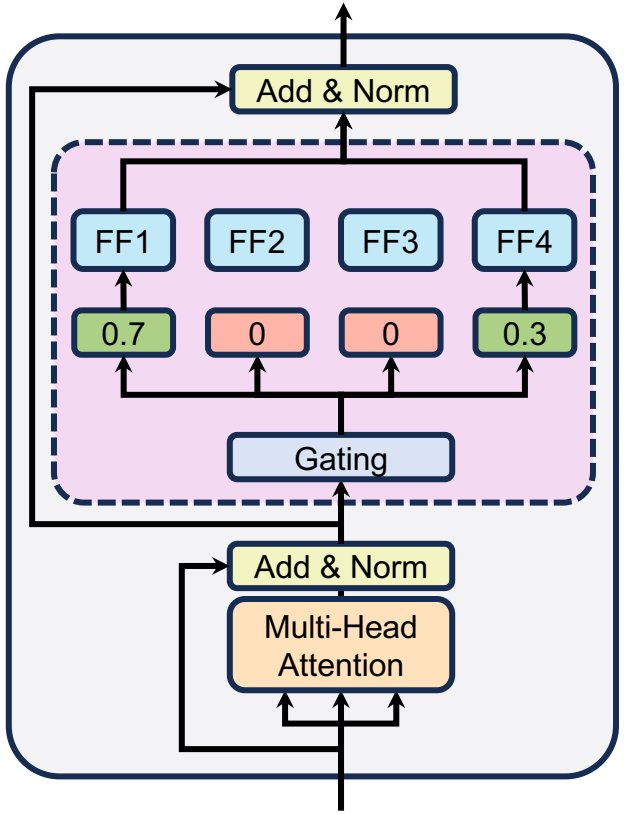
Flavio Giobergia

Mixture of Experts

- Technique used to increase the model size, without significantly increasing the computational cost of its execution
- We add many “experts” to the model
  - E.g., different versions of one layer
- Only one (or few) experts are used for any prediction
  - *Sparse* Mixture of Experts (not using all experts!)
- Originally proposed some time ago (Jacobs et al., 1991)
  - Now used for LLMs, with some adaptations

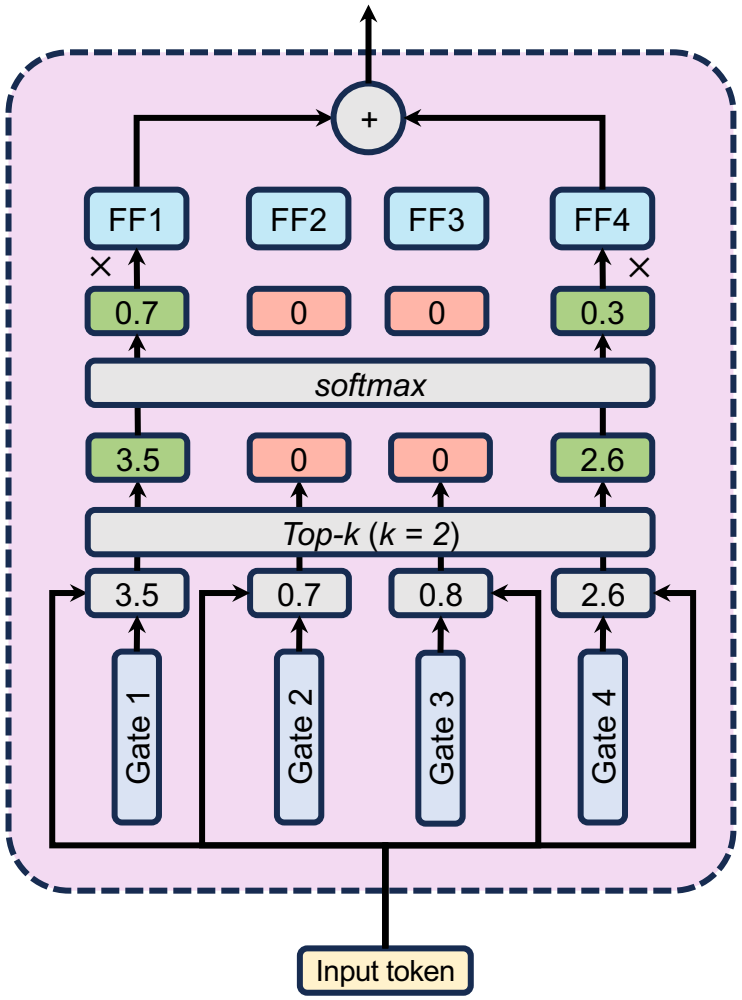
Expert layers in transformers

- For transformer-based architectures, we often use different experts for the FF-NN layers
- We need a way to know “who are the right experts”
  - Multiple experts can be selected (*top-k*)
  - If k > 1, we must define how much we weight each expert’s contribution
- We introduce a gating mechanism
  - A layer that decides, for each token, which experts should be used



Gating mechanism

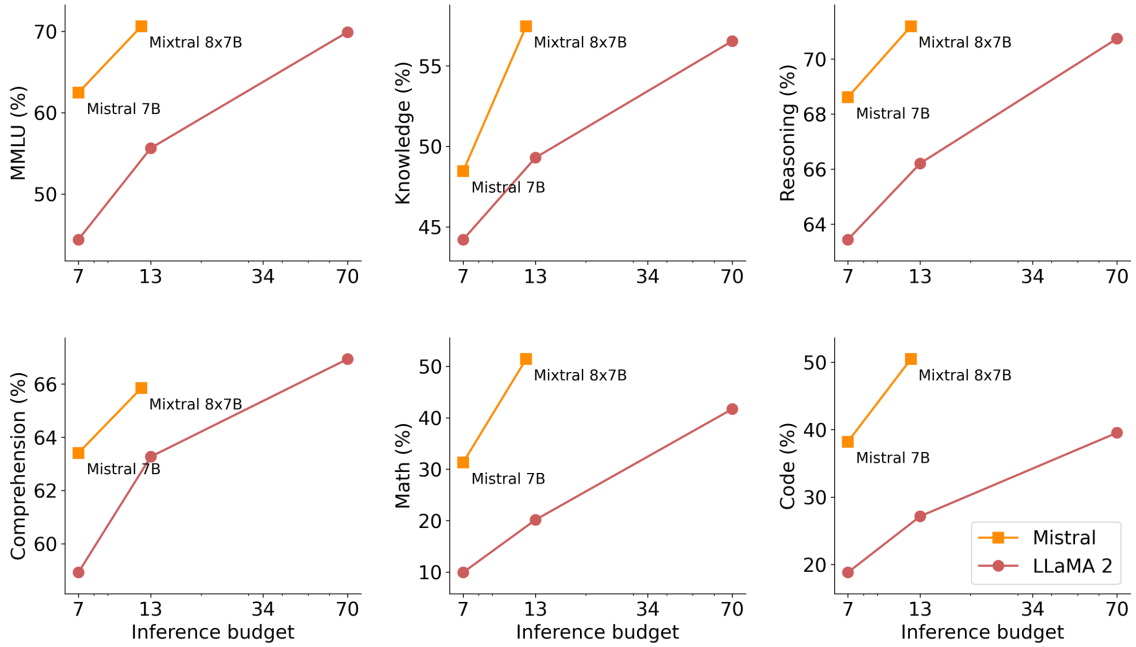
- Each expert has a “gating vector”
- Each token is scored against the gating vector (dot product)
- The top k experts are selected, and scaled (via sofmtax)
- The output is the weighted sum of the outputs of the selected experts



Mixtral

https://mistral.ai/news/mixtral-of-experts/

- *Sparse* Mixture of Experts (SMoE) released by Mistral AI
- Based on Mistral 7B architecture
- 8 experts in all FF-NN layers of transformers
- Top 2 experts selected
- Total parameters: 47B
- Active parameters at inference: 13B



Parameter	Value
dim	4096
n_layers	32
head_dim	128
hidden_dim	14336
n_heads	32
n_kv_heads	8
context_len	32768
vocab_size	32000
num_experts	8
top_k_experts	2

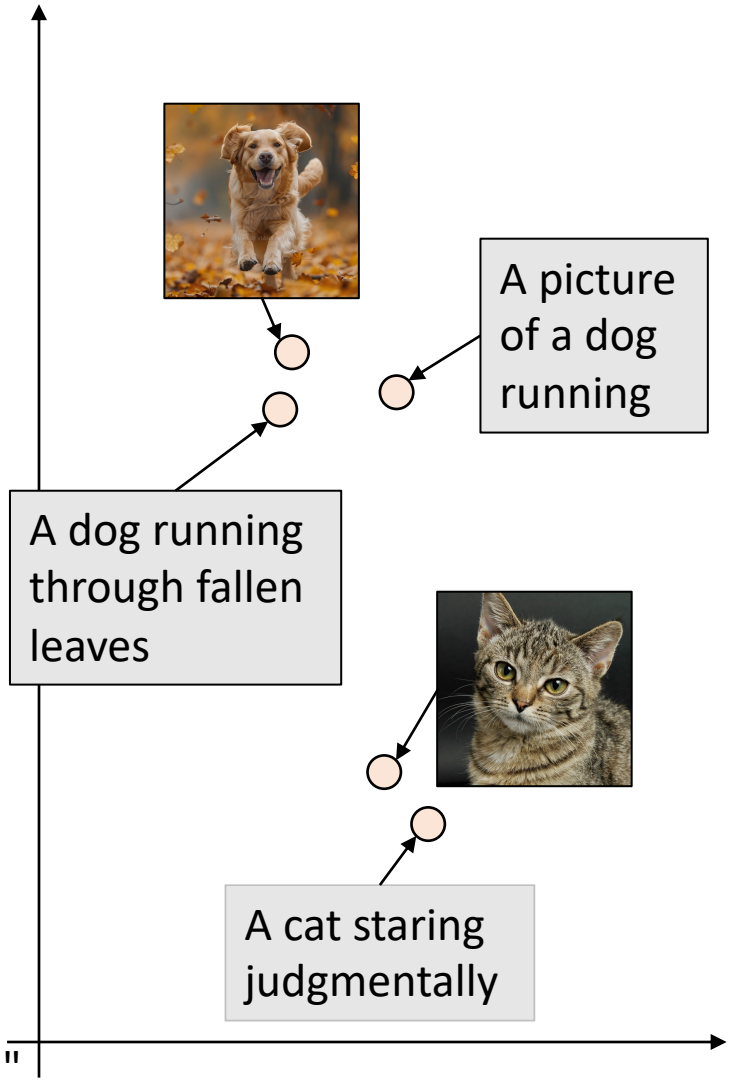
Jiang, Albert Q., Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot et al. "Mixtral of experts." *arXiv preprint arXiv:2401.04088* (2024). https://arxiv.org/pdf/2401.04088

Vision-Language Models (VLM)

- VLM are multimodal models that can use both images and text as inputs and/or outputs
- Used for a wide variety of tasks
  - Image captioning, Visual Question Answering, image-text matching
- Key components
  - Vision encoder (converts images to vectors)
  - Language encoder (converts text to vectors)
  - Multimodal fusion (combines vision and language vectors)

CLIP

- Contrastive Language Image Pretraining (CLIP) is one of the earlier vision-language models based on transformers
  - OpenAI, 2021
- Text and images are aligned in the same vector space
  - So that the vector of the image of a dog and the vector for the sentence “a picture of a dog” are close in the shared space
- Vision modality (images) encoded with a vision encoder
  - ResNet, or Vision Transformer (ViT)
- Text modality encoded with a decoder-only model (simil GPT-2)
  - Using vector associated to [EOS] token as representative of the text
- Similarity between text and images performed with cosine similarities



Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *International conference on machine learning*. PMLR, 2021. https://arxiv.org/pdf/2103.00020

Contrastive learning alignment

- The training set for CLIP is characterized by image-text pairs
  - 400M image/text pairs (x_i^{img}, x_i^{txt})
  - Source: Internet (e.g., from captioned images) → WebImageText

- Use image encoder $f_{img}(x_i^{img})$ to encode the image as a vector
- Use text encoder $f_{txt}(x_i^{txt})$ to encode the text as a vector
- Align embeddings with contrastive learning
  - Maximize $cossim(f_{img}(x_i^{img}) f_{txt}(x_i^{txt}))$, and
  - minimize$cossim(f_{img}(x_i^{img}), f_{txt}(x_j^{txt}))$, when $i \neq j$

- In this way, texts are placed close to the corresponding images, and far away from different images

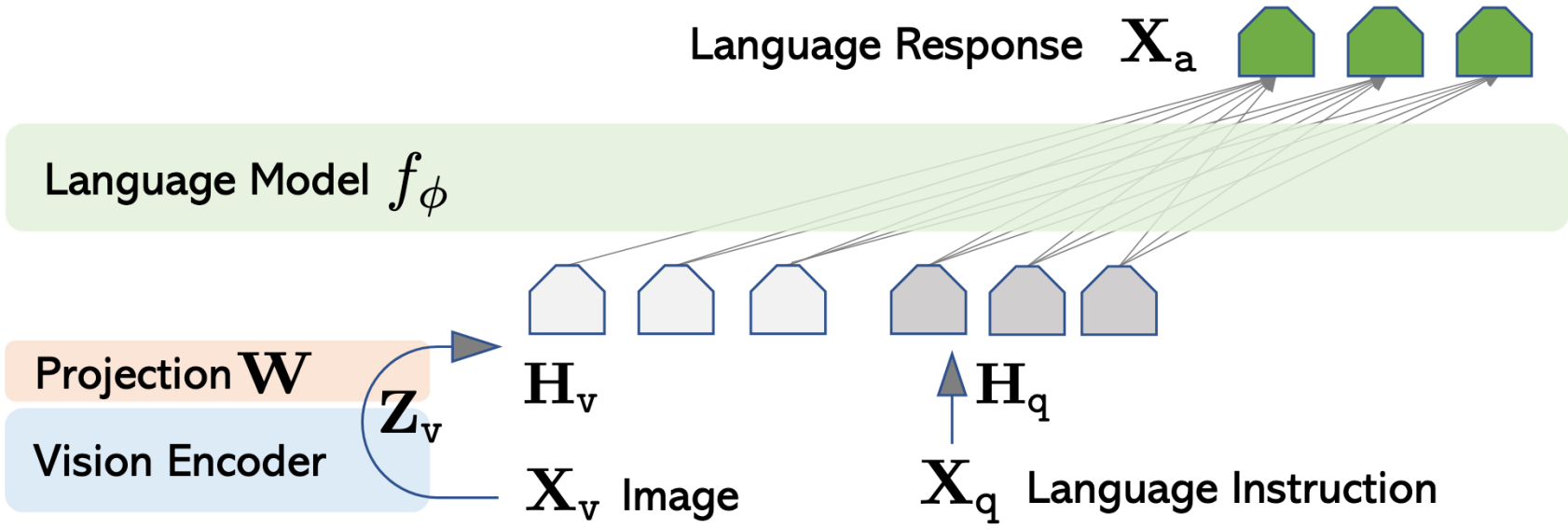
CLIP results

- CLIP is not a generative model
- It can encode text and images, and find text-image similarities
- Given a set of images, we can encode all of them.
- Then we can use text queries to retrieve the most semantically relevant image(s)
  - E.g., given a set of images, find the one closest to “a cat chasing a chicken”

(Source: original paper)

LLaVA

- LLaVA (Large Language and Vision Assistant) is an instruction-tuned, multimodal LLM
- Uses a ViT to encode the input image into “visual tokens”
- The visual tokens are projected (aligned) with a learned matrix W
- The visual tokens are prepended to the language instructions (question)
- An instruction-tuned LLM (Vicuna) is tuned to answer questions



(Source: original paper)

Liu, Haotian, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. "Visual instruction tuning." *Advances in neural information processing systems* 36 (2024). https://arxiv.org/pdf/2304.08485