Tutorial -- Generalist Agent AI

Opening Remarks

Jianfeng Gao

Microsoft Research, 6/18/2024

Agenda

- Success of LLMs
- LLMs as AI Agents
- This tutorial

Success of Large Language Models (LLMs)

- Context length…
- Scaling laws…
- Emergent abilities
  - In-context-learning
  - LLMs as (general-purpose) task solvers

Context length of Language Models (LMs)

- LM: $P(w \mid h)$
  - LMs are better with longer (richer) context h
- N-gram LMs: $|h|$ = 1 to 6
  - Model size grows exponentially with $|h|$
- RNN/LSTM LMs: $P(w \mid c(h))$,
  - where h is *compressed* to a fixed-size vector
- Transformer LMs: $|h|$ = 2K to 8K … to 1M
  - Sparse attention is used to deal with quadratic complexity

Scaling laws

[Kaplan+ 20, Hoffmann+ 22]

- Power-law relationship of
  - L – model performance w/
  - N – model size
  - D – Training data size
  - C – amount of training compute

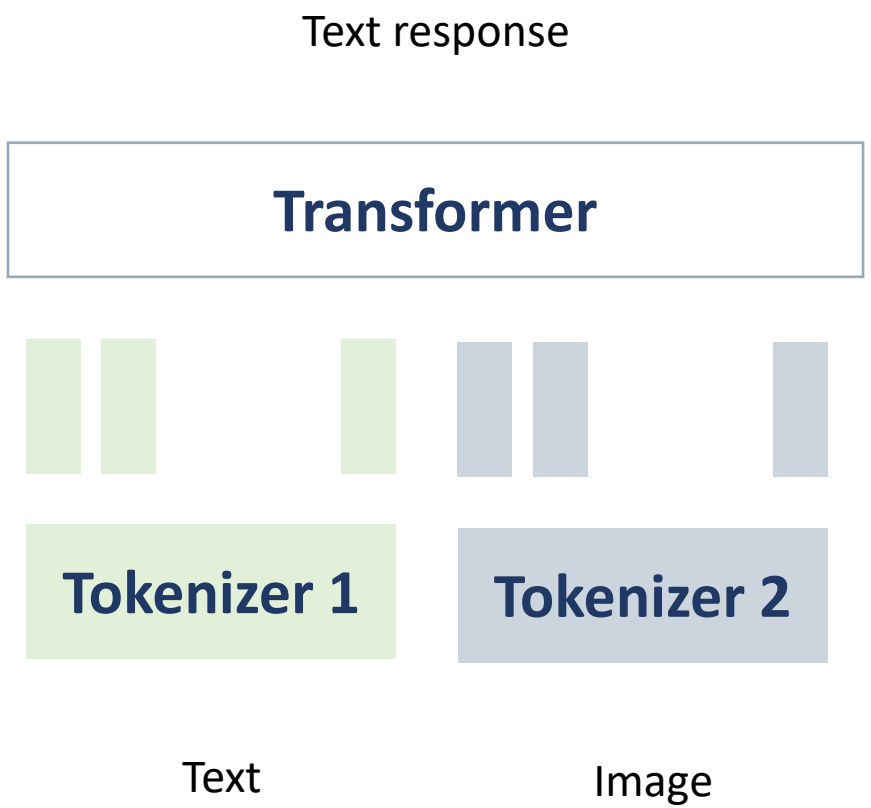
$$L(N) = \left(\frac{N_c}{N}\right)^{\alpha_N}, \quad \alpha_N \sim 0.076, N_c \sim 8.8 \times 10^{13}$$

$$L(D) = \left(\frac{D_c}{D}\right)^{\alpha_D}, \quad \alpha_D \sim 0.095, D_c \sim 5.4 \times 10^{13}$$

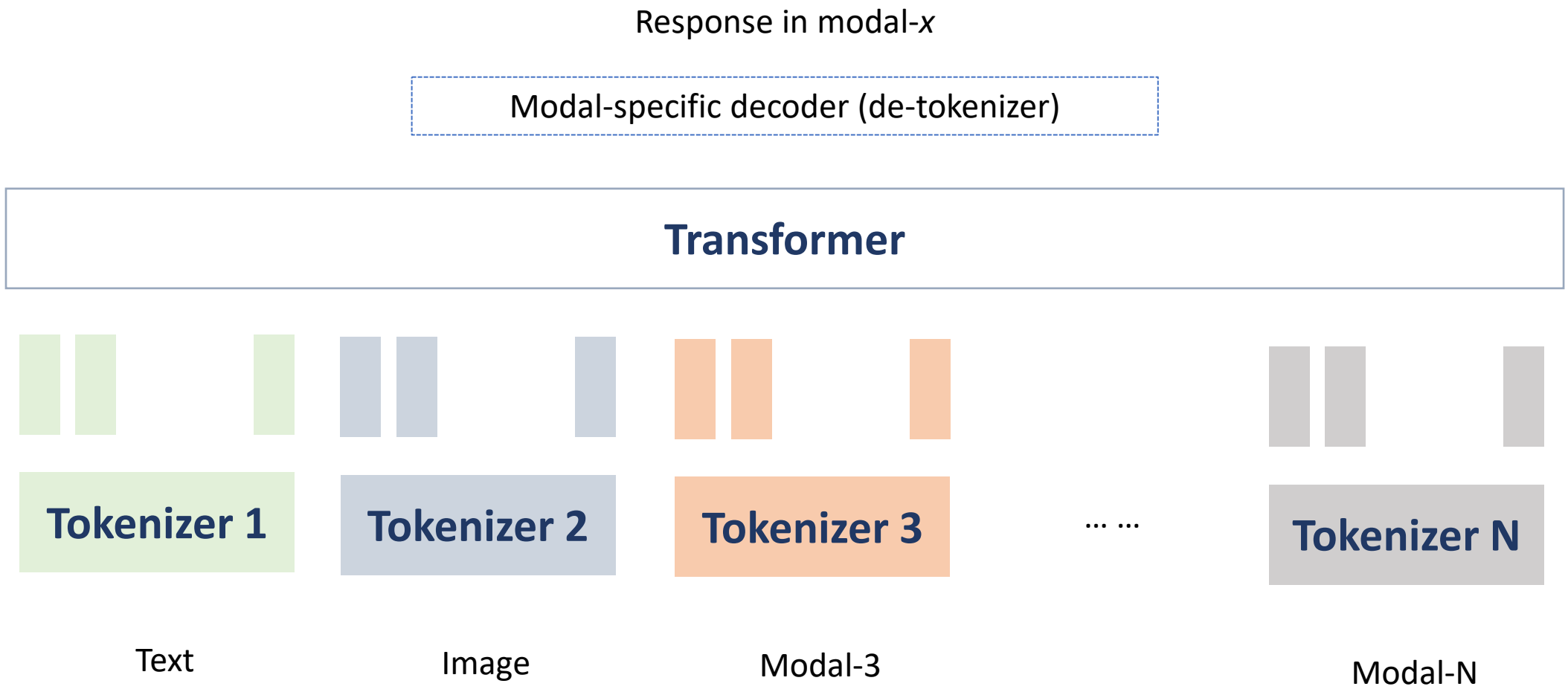
$$L(C) = \left(\frac{C_c}{C}\right)^{\alpha_C}, \quad \alpha_C \sim 0.050, C_c \sim 3.1 \times 10^{8}$$

- We can keep improving LLMs by
  - increasing model capacity (context length)
  - increasing training data (raw text)
  - Increasing compute ($$$)

LLMs as LMMs (e.g., LLaVA, Phi-3-Vision)



LLMs as LMMs (e.g., LLaVA, Phi-3-Vision)

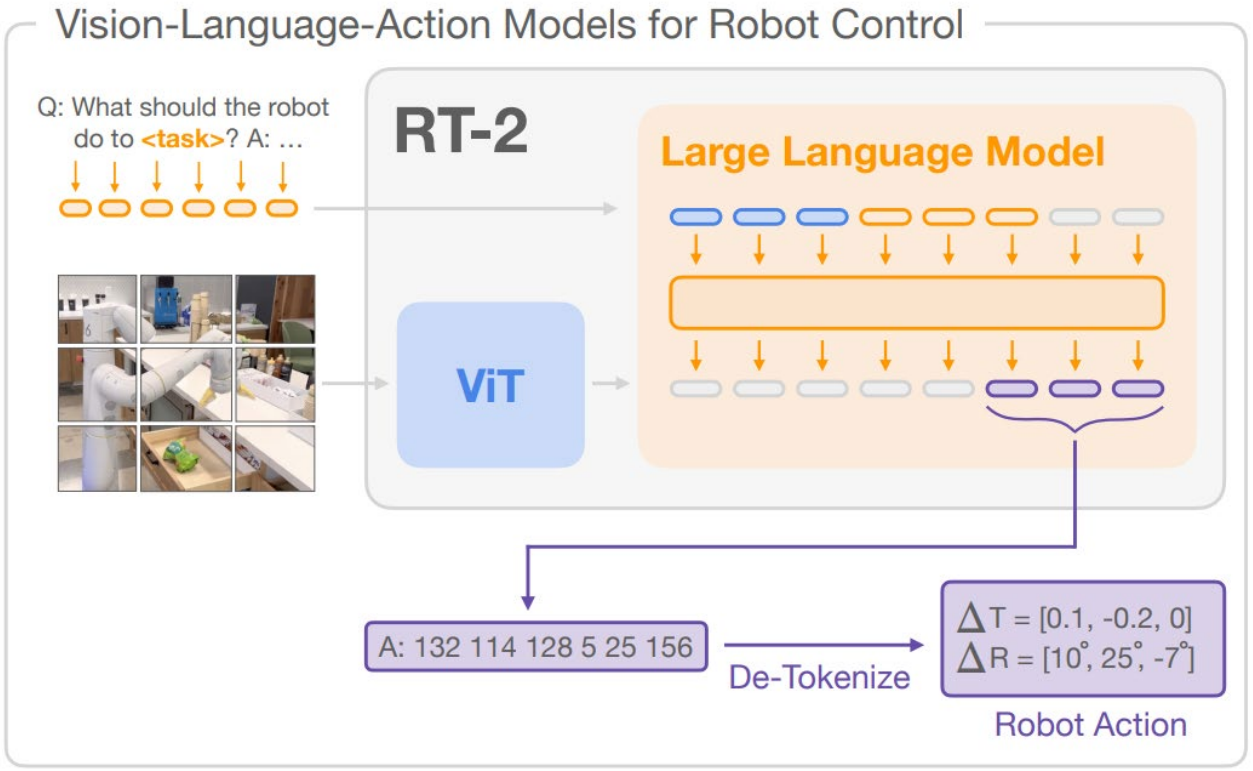


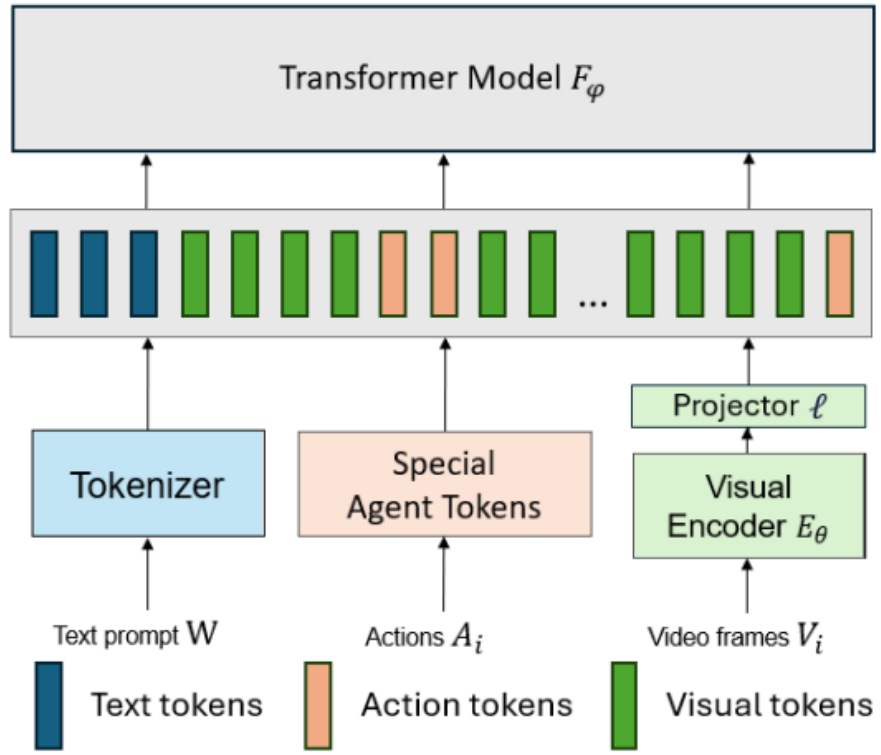
LLMs/LMMs as Agent Models

- LMMs predict word/visual tokens
  - $(w_1, w_2, \dots, w_t) \rightarrow w_{t+1}$
- AI Agents predict observation-action sequences
  - Policy model: $(o_1, a_1, o_2, a_2, \dots, o_t) \rightarrow a_t$
  - World model: $(o_1, a_1, o_2, a_2, \dots, o_t, a_t) \rightarrow o_{t+1}$
- LMMs are agent models if we can
  - Tokenize observations
  - Tokenize actions

LLMs as Agents

[Brohan+ 23 (Google), Durante+ 24 (MSR)]





- Text tokenizer
- Vision (observation) tokenizer – videos as another language
- Action tokenizer – (robot) actions as another language

LMM-powered Agents… But what is lacking?

- Hallucination
  - Grounding – augmenting LMMs w/ knowledge & tools
  - Asking why – making LMMs causal and interpretable

- Cope with when things do not go as planned
  - Learning thru AI-human interactions
  - Self-improving via continual learning w/o catastrophic forgetting

- More…
  - World Model vs. Action Model? One model for all?
  - EAI models for robotics or copilots?
  - What is the *word prediction* task in pretraining EAI-FMs?
  - Context?
  - Scaling laws?
  - Tokenizers or not?
  - What are the emergent abilities?

Timetable Schedule

Time Slot	Talk Scheduling	Areas
08:30 - 08:40	Jianfeng Gao	Opening Remarks
08:40 - 09:30	Talk1: Juan Carlos Niebles	LLM tool-based agents
09:30 - 09:50	Coffee Break	
09:50 - 10:40	Talk2: Yong Jae Lee	Generalist Multimodal Models
10:40 - 11:30	Talk3: Katsushi Ikeuchi	Agent Robotics
11:30 - 11:40	Naoki Wake	Ending Remarks

Invited Speakers

Juan Carlos Niebles

Katsushi Ikeuchi

Yong Jae Lee