Generative AI: Understanding the Technology

Jan Hajič
Moderator

Panelists:
Yordanka Ivanova
Andrejs Vasiljevs
François Yvon

Institute of Formal and Applied Linguistics
Computer Science School
Faculty of Mathematics and Physics
Charles University, Prague, Czech Republic
Introduction

- Artificial Intelligence
  - Technically, from the 1950s
  - Definitions / Turing test
- Theoretical AI
  - Planning, Scheduling, …
  - Difficult mathematical questions
- Practical AI
  - Expert systems
  - 1970s – “rule-based”

November 7, 2023, STOA/EP

Generative AI: Understanding the Technology (Panel 1)

Marvin Minsky, ca. 1970, MIT
Modern Artificial Intelligence

• “Superhuman performance”

IBM Deep Blue (chess, 1990s)

IBM Watson (Jeopardy, 2011)
AI and Machine Learning

- Statistical machine learning and models
  - since 1980s – computer vision, speech, machine translation

Source Channel model in statistical language modeling (ASR)
AI and Deep Learning

- The principle: Artificial Neural Network, many layers

[Diagram of a deep neural network with input layer, hidden layer 1, hidden layer 2, and output layer.]

DNN: the “transformer” architecture

- Current models: transformers (encoder-decoder)

Picture: https://jalammar.github.io/illustrated-transformer/
DNN: the “transformer” architecture

- Embeddings, self-attention, feed-forward layers

Picture: https://jalammar.github.io/illustrated-transformer/


Attention Is All You Need
https://doi.org/10.48550/arXiv.1706.03762
DNN: the “transformer” architecture

- **Embeddings, self-attention, feed-forward layers**

Each word “knows” about any other word in context (even very distant)

Weighted links

Picture: https://jalammar.github.io/illustrated-transformer/
What does an LLM do?

- LLMs today: transformer architecture (without the encoder)

Inputs
Input
Once upon a time,

Text Generation Model

Output
Output
Once upon a time, we knew that our ancestors were on the verge of extinction. The great explorers and poets of the Old World, from Alexander the Great to Chaucer, are dead and gone. A good many of our ancient explorers and poets have
Conversational LLMs (e.g. ChatGPT)

- Fine-tuning a foundational LLM

![Diagram of LLM fine-tuning process]

Picture: https://www.analyticsvidhya.com/blog/2023/07/build-your-own-large-language-models/
LLM Evaluation

- Technical evaluation
  - Evaluation on several (many) NLP (or other) tasks
  - Evaluation of performance (speed, size requirements)
- Evaluation of...
  - Robustness, Ethics, Bias, and Trustworthiness
- Source: survey paper (updated for 2023):
  - https://doi.org/10.48550/arXiv.2307.03109
- See François Yvon’s presentation later today!
- Technical evaluation
- Evaluation on several (many) NLP (or other) tasks
- Evaluation of performance (speed, size requirements)
- Robustness, Ethics, Bias, and Trustworthiness
- Source: survey paper (updated for 2023):
  - https://doi.org/10.48550/arXiv.2307.03109

November 7, 2023, STOA/EP

Generative AI: Understanding the Technology (Panel 1)
LLM Sizes and Computing Needs

- LMM size is given by the number of “parameters”
  - Parameters:
    - Real numbers, weights of neural connections in the model
    - Largest today: 175B (billion, GPT-3), 70B (Llama 2)
    - Training size (correlated): 10s of terabytes of plain text
- Computing
  - Training: much more needed (GPT: ~300 years on single GPU)
    - Parallel (federated) training on many hardware units (GPU cards)
  - Runtime:
    - Much smaller requirements, ideally 1 GPU card
LLM availability (textual)

- Open and Closed LLMs
- OpenAI – ChatGPT (closed, free & paid access)
- Huggingface.co
  - Open source models (also image generation)
    - Download, fine-tune, run, evaluate
- Data (texts) for building your own
  - HPLT project (EU)
    - https://hplt-project.org/datasets
    - CommonCrawl, ParaCrawl
  - OpusMT
  - CLARIN, ELG, META-SHARE, ELRC-SHARE, … repositories
Thank you!