Analytics Vidhya | DATAHACK SUMMIT

Workshop

NLP using Generative AI

Speaker

Raghav Bali
Staff Data Scientist, Delivery Hero

Amar Lalwani
Co-founder, Stealth Startup

Module 3

Diving into ChatGPT

Till GPT-2

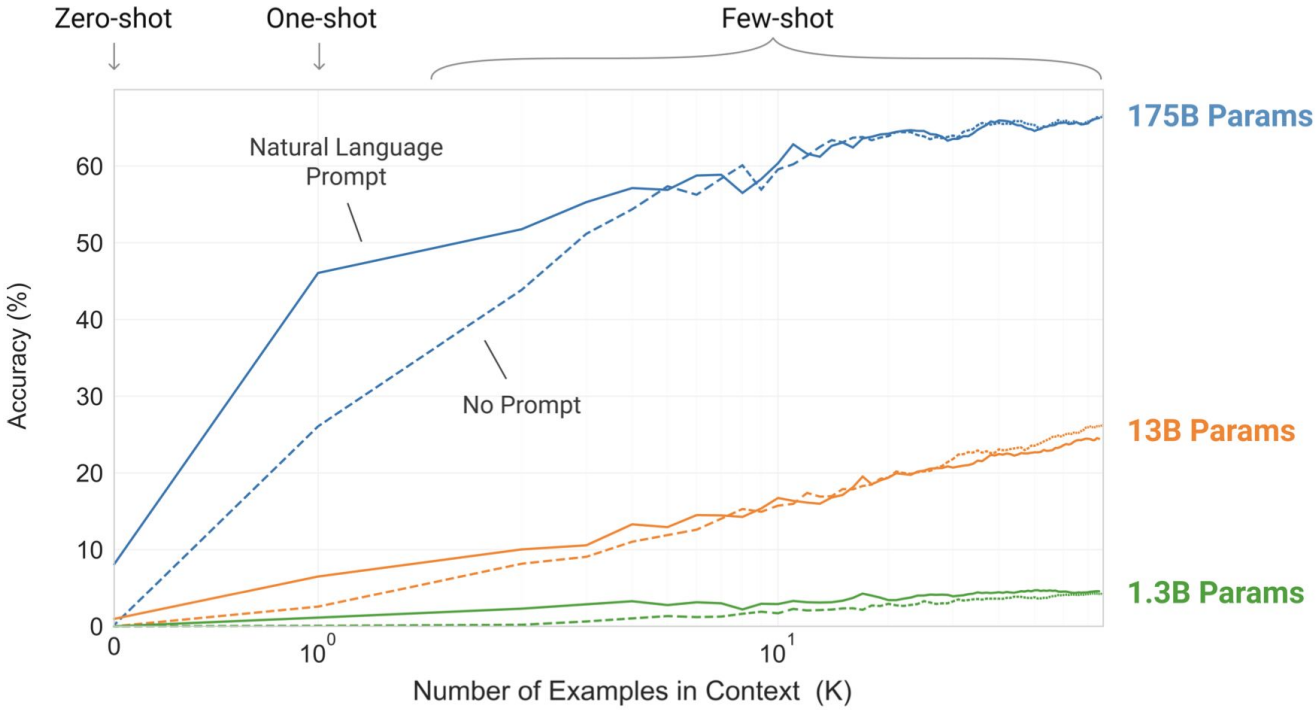
- Pre-training transformer language models
- Task specific fine-tuning
  - Eliminating need of task specific architectures
- Limitation: while architecture is task-agnostic, still task specific datasets and fine-tuning needed

➔ For each task, large high quality dataset needed. May not be possible always

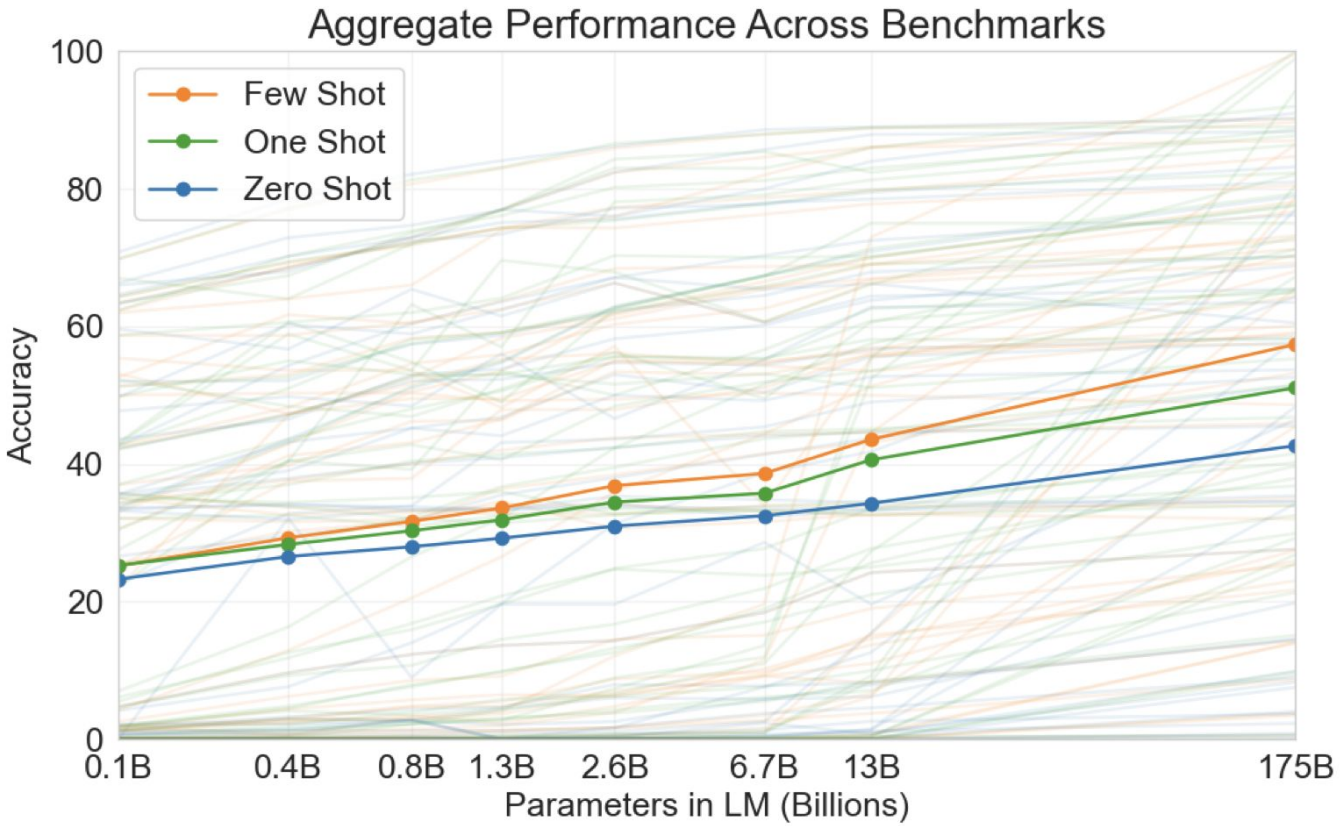
➔ Out of distribution generalization decays with
  - ◆ Expressiveness of the model
  - ◆ Narrowness of the training distribution

➔ Humans to do not require large training datasets to learn most natural language tasks
  - ◆ Brief directive in natural language or at-most a tiny number of demonstrations is enough
  - ◆ This allows humans to seemingly mix together or switch between various tasks and skills

GPT-3 In-context learning: remove random symbols from a word



GPT-3 Performance



Let us train nameGPT: pre-training

- Dataset: gender-wise firstnames from various countries
- Character level model
- Causal model: Self-supervised learning
  - Predict next character
- GPT style decoder-only transformer architecture
  - 3.37 million parameters

nameGPT: generating names

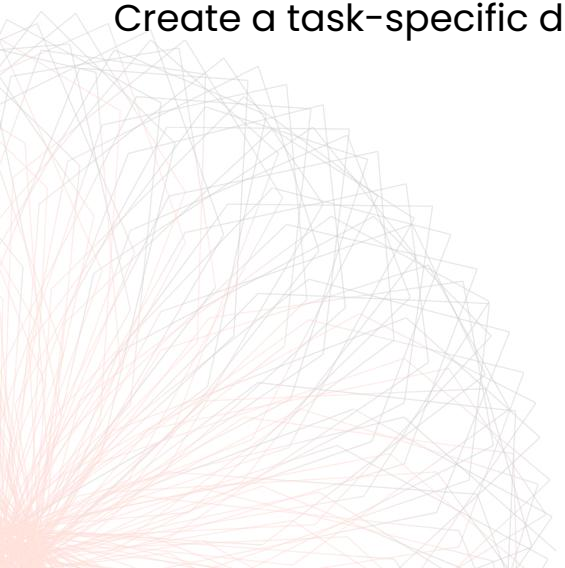
- Sample from pre-trained model
- Prompt Engineering
  - Name starting from “prefix”

nameGPT: fine-tuning

I want the model to generate

- Indian names
- Names from a country: <country_name>
- Female/Male names
- Names with a “prefix”
- Names with a “suffix”

Create a task-specific dataset and fine-tune the pre-trained model



nameGPT: instruction-tuning

I want the model to perform following tasks based on the instructions

- Generate names with a “prefix”
- Generate names with a “suffix”
- Generate names with a “sub-string”
- Generate a name which is similar to “input_name”

Create the dataset and perform instruction-tuning on the pre-trained model

nameGPT: fine-tuning to classify

Gender Identification

- Given a name, pick it's gender: "Male" or "Female"

Create the dataset and perform fine-tuning on the pre-trained model

References

- Language Models are Few Shot Learners, OpenAI, 2020
- NanoGPT, Andrej Karpathy
- Name Dataset, `https://github.com/philipperemy/name-dataset`

Thank You!