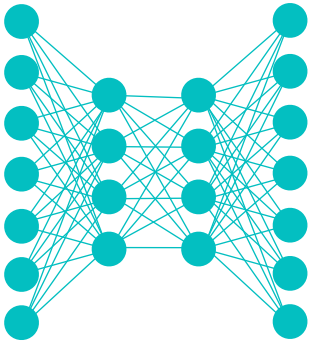
Lecture Notes for

Neural Networks and Machine Learning

Course Introduction
Lecture: AI Ethics



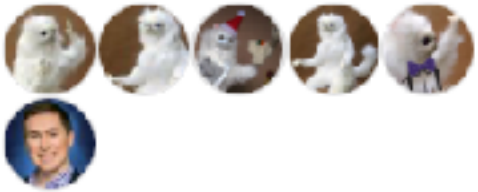
Logistics and Agenda

- Logistics
  - This class evolves across semesters (sometimes drastically!)
    - First offered in 2019
  - Using Canvas
  - GitHub: Mostly one repository
- Agenda
  - Syllabus and Introductions
  - Presentation Selection
  - Stochastic Parrots

Syllabus

- Course Schedule
- Reading/Videos
- GitHub
- Grading
  - Labs x4 (60%)
  - Final Paper x1 (25%)
  - Participation (5%)
  - Paper Discussion/Summary x1 (10%)

People



8000net

This organization houses a number of repositories for Dr. Larson's 8000 Level Neural Networks Course, Offered at SMU

Introductions

- Name
- Department
- Where you grew up
- Topic in this course you are most excited about
- Something true or false about you

- Do NOT forget:
  - Pick out papers on Canvas (distance students also)

Presenting OR Summary

- First Presentation is Next Week!
- During Semester: 7 Presentations Total (as a team)
- First Presentation →
- **Who wants to go first?**
  - ~10 Minutes
  - Summarize the Article
  - Make 3-5 Visuals
    - e.g., Slides
    - AND/OR Handouts
    - AND/OR Notebooks
- Alternative: 3-page Summary of paper, with Figures

Are Emergent Abilities of Large Language Models a Mirage?

Rylan Schaeffer
Computer Science
Stanford University
rschaef@cs.stanford.edu

Brando Miranda
Computer Science
Stanford University
brando9@cs.stanford.edu

Sanmi Koyejo
Computer Science
Stanford University
sanmi@cs.stanford.edu

Abstract

Recent work claims that large language models display *emergent abilities*: abilities not present in smaller-scale models that are present in larger-scale models. What makes emergent abilities intriguing is two-fold: their *sharpness*, transitioning seemingly instantaneously from not present to present, and their *unpredictability*, appearing at seemingly unforeseeable model scales. Here, we present an alternative explanation for emergent abilities: for a particular task and model family, when analyzing fixed model outputs, emergent abilities appear due to the researcher's choice of metric rather than due to fundamental changes in model behavior with scale. Specifically, nonlinear or discontinuous metrics produce seemingly emergent abilities, whereas linear or continuous metrics produce smooth, continuous, predictable changes in model performance. We present our alternative explanation in a simple mathematical model, then test it in three complementary ways: we (1) make, test and confirm three predictions on the effect of metric choice using the InstructGPT/GPT-3 family on tasks with claimed emergent abilities; (2) make, test and confirm two predictions about metric choices in a meta-analysis of emergent abilities on the Beyond the Imitation Game Benchmark (BIG-Bench); and (3) show how to choose metrics to produce never-before-seen seemingly emergent abilities in multiple vision tasks across diverse deep network architectures. Via all three analyses, we provide evidence that emergent abilities disappear with different metrics or with better statistics, and may not be a fundamental property of scaling AI models.

Lecture Notes for

Neural Networks and Machine Learning

Course Introduction

Next Time:
Case Studies in Ethics of ML

Reading: None