LLM講座統括・ラップアップ

松尾研究室
MATSUO LAB THE UNIVERSITY OF TOKYO

サマースクール 2023
大規模言語モデル講座

2023/09/25

全体を通して

みなさんお疲れ様でした（週2回，毎回2時間講義＋最終課題）

- 第1回：Overview of Language Models → なぜ今LLMなのか？（スケール，汎用性，他分野への影響）
- 第2回：Prompting and Augmented Language Model → LLMを活用する技術（プロンプト，RAG）
- 第3回：Pre-training Pipeline → Transformerとは何か，事前学習の基本
- 第4回：Scaling Pre-training → Transformerをスケールさせる仕組み（Flash Attention，MoE，データ）
- 第5回：Parameter Efficient Fine-Tuning → 最近のLLMでデファクトで使われるInstruction Tuning
- 第6回：RLHF → ChatGPT等で利用される暗黙的な意図に従う方法, alignment
- 第7回：Going Beyond LLM → マルチモーダル，日本語LM，パネル

大体活用/研究する上での基本的な議論は網羅

最終課題(コンペティション)について

概要: **LLM講座 最終課題コンペティション**

修了: **修了要件の1つとして含まれる**

上位入賞者: **優秀者の表彰を行う可能性あり**

開催時期: **2023/09/25 - 2023/10/10 23:59（予定）**

コンペ内容:
下記の３種類のベンチマーク性能を高めよ。ただし、事前学習、FinetuningやRLHF等、プロンプティングの工夫など、授業で学んだことを自由に利用して構わない。モデルや学習に利用できるデータについてはルールを参照のこと。
Type１：日本語QA
Type２：文章要約
Type３：Instruction Following

Type1・Type2成績上位者のうち、Type3のコンペに参加希望の方には人手による評価に協力をいただきます。

本日18時くらいを目処にコンペ開始の連絡をSlackで行います。(testデータ配布とstarter code配布)

詳細は: 大規模言語モデル講座 最終課題 LLMコンペティション(Slack共有済)
をご確認ください。

謝辞: Google Cloud Japan様

今回のLLM講座を立ち上げるに至って最もハードルとなった、受講者向けへのGPU演習環境の調達において、どこよりも迅速な対応とご支援いただきました、Google Cloud Japan の下田様、秋元様、水江様へ感謝、御礼を申し上げます。

最後に

■ なぜいま言語モデルなのか？

- 1. モデル，データ，計算量のスケールによりできることが急速に広がっている
- 2. Promptingにより，単一モデルで様々なことができるように（言語モデルの汎用性）
- 3. 言語モデルの発展が他の領域にも影響を与えている

■本講座の趣旨

- LLMの技術的背景，原理を理解することで，ハイプとしてではなく活用する技術として捉えられるようになる

■本講座開講にあたって

- 実は開講が決まったのは6月中旬ごろ→7月募集，9月開講
- （松尾）「大規模言語モデルについて正しく理解し開発できる能力を身につける重要性は今と1年後では大きく異なる」
- ぜひこの機会に学べたみなさんは，学んだことを研究/業務/プロジェクトで活用してもらえれば

講座内容をより深く理解したい方へ

■ 基礎的なことをより網羅的に学びたい方＝> 有名な講義

- CS224N: Natural Language Processing with Deep Learning
- CS388: Natural Language Processing (online MS version)
- DeepLearning.AIによる講義(日本語への翻訳を松尾研が担当、近日公開予定)

■ 発展的な研究領域に触れたい方 ＝> 松尾研主催の勉強会

- 毎週木曜日9:00-10:30に最新の論文を紹介(LLM輪読会、slackにてzoom link共有済)
  - 発表例: Llama2、Language modeling is compression、etc
- DL輪読会（毎週金曜10:00 – 11:30, フォームで案内予定）

■ 実装力を高めたい方

- プロジェクトでの実装をぜひ
- 実装のtips共有会(LLM Hacks、後日修了生向けに案内予定)

--

※ 講義をさらにパワーアップしてくれるメンバも募集しています！

松尾研でのLLM開発

- 松尾研では本講義で講師を務めた小島を中心にLLM（Weblab-10B）を構築
- 今後、LLM開発/活用ノウハウのさらなる獲得、またそれを担う人材を育成することを目指し
  さらに大きな日本発の大規模言語モデルの開発に挑戦予定
- 本講義を受講いただいた皆様へ、より踏み込んだLLMの研究開発へのご協力依頼をさせていただく可能性があります
  - 受講状況や最終課題等を確認させていただき、別途ご連絡

東京大学松尾研究室 100億パラメータサイズ・日英2ヶ国語対応の 大規模言語モデル“Weblab-10B”を公開 ―公開済みの日本語大規模言語モデルで最高水準―

東京大学大学院工学系研究科技術経営戦略学専攻松尾研究室（教授：松尾 豊、以下「松尾研」）は、この度日本語・英語の2ヶ国語に対応した100億パラメータサイズの大規模言語モデル（Large Language Model ;LLM）を事前学習と事後学習（ファインチューニング）により開発し、モデルを公開しましたのでお知らせします。今後も、Weblab-10Bのさらなる大規模化を進めるとともに、この資源を元に、LLMの産業実装に向けた研究を推進して参ります。

松尾研は、知能の謎を解くことを目的に人工知能の研究に取り組む研究室です。現在はテキスト生成で注目されることの多いLLMの技術ですが、今後は画像組み込みなどのマルチモーダル化、ブラウザ・ソフトウェア・ロボット等の行動制御の実装に発展し、人工知能研究を加速させると期待されます。また、LLMの開発競争が世界で激化する中、技術を理解した人材の育成も重要です。本研究開発は、上記の通り研究室の人工知能の研究を加速させるとともに、研究開発から得られた知見を講義開発等に生かすことで、大学における教育活動に資することも意図しています。

発表の詳細

近年の大規模言語モデルは、インターネットから収集した大量のテキストデータを学習に用いますが、そのテキストデータの多くは一部の主要言語（例えば英語）で構成されており、それ以外の言語（例えば日本語）のテキストデータを大量収集することには現状では限界があります。そこで松尾研は、日本語だけでなく英語のデータセットも学習に用いることで学習データ量を拡張し、言語間の知識転移を行うことで日本語の精度を高めることを目的とした100億パラメータサイズの大規模言語モデル“Weblab-10B”を開発し、公開しました。

https://www.t.u-tokyo.ac.jp/press/pr2023-08-18-001

第1回LLM Summer Fin.