

第2回海外留学報告書

釜堀 恵輔 *

2024年7月

ワシントン大学 (UW) の CS PhD に留学している釜堀恵輔です。早いもので留学を開始してから1年が経ちました。

1月ごろから大規模言語モデル (LLM) を効率的に動かすためのコンピュータシステムについて研究をしていて、方向性を模索しながらではありますが、この半年で少しずつ目に見える成果が出始めてきました。メインでやっているのは、Retrieval-Augmented Generation (RAG) という、LLM 推論と外部の知識ベースからの情報検索を組み合わせる新しいタイプのアプリケーションに関する研究です。LLM の勃興とともに RAG も大きな注目を集めていますが、システムの方面からの研究はあまりされてきていなかったのが現状でした。このプロジェクトでは、世の中で使われ始めている RAG アプリケーションのタイプを探りつつシステム効率性に関する共通のボトルネックを探り、それらを解決する最適化手法を提案するという方向で進めました。わりと新しい分野を開拓するタイプの研究で、色々な紆余曲折もあり一筋縄ではいきませんでした。それなりに面白い研究ができたのではないかと思います。特に、3月ごろにサンフランシスコに行って、共同研究している教授のスタートアップのイベントに参加したのが、方向性を定める上で大きな助けになりました。最終的に6月に締め切りの学会に論文を出したので、無事出版されたらまた詳しく書けたらと思います。

また、RAG のプロジェクトの途中の2月ごろ、必要な計算リソースを用意するためのダウンタイムがあり、その時に使えたリソースで色々遊んでいたら運良く結果が出たので、それに関する論文も書きました。内容は Mixture-of-Experts というタイプの LLM を限られた GPU リソースで効率良く推論するためのシステムに関するもので、国際会議のワークショップにアクセプトされ5月にウィーンで発表してきました。これ自体は正直大した業績ではありませんが、ML System 分野の研究をしていることをアピールできて、共同研究など大学内外の色々な機会につながったのはよかったです。その後インターンに来ている学部生の助けも借りつつ別の国際会議の full paper としても submit したので、吉報が来ることを祈っています。

それ以外にも同期の LLM serving に関するプロジェクトに共著で入ったり、その他にも何人かのプロジェクトに協力しています。前回の報告書で書いた (共同研究先との conflict があった) プロジェクトも別のインターンの助けを借りつつ進めていますが、残念ながらこちらに関してはまだ今後どうなるか不透明です。

学部生の時などこれまではわりと個人戦で研究してきましたが、今年に入ってから一気に他人と共同研究をする機会が増えたように思います。例えば先述の RAG のプロジェクトでは教授陣も含めて10人以上が参加しています。こういった集団戦のメリットは色々な人からのフィードバックを得ながらスピーディに成果をあげられるところですが、一方で仕事の分担の仕方や他人にアイデアを理解してもらうコミュニケーションなどの難しさを実感しています。さらに今は3人のインターンの面倒を見ていて (しかも全員別々のプロジェクト)、他人をメンターした経験がありませんので、苦戦しながらも手探りで進めています。

授業面では、春学期は CSE 552 Distributed Systems と CSE 547 Machine Learning for Big Data の2つを取り



ウィーンの学会会場

* kamahori@uw.edu

