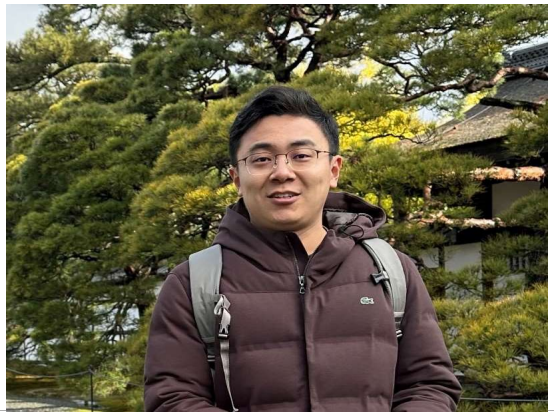
受入大学名	東京大学		
Host University	The University of Tokyo		
外国人研究者	石 睿		
Foreign Researcher	SHI Rui		
受入研究者	山口 泰	職名	教授
Research Advisor	Yasushi YAMAGUCHI	Position	Professor
受入学部/研究科	総合文化研究科		
Faculty/Department	Graduate School of Arts and Sciences		

＜外国人研究者プロフィール/Profile＞

国　籍	中国
Nationality	China
所属機関	北京工業大学
Affiliation	Beijing University of Technology
現在の職名	講師
Position	Lecturer
研究期間	2023年 12月 22日～ 2024年 2月 19日(60日間)
Period of Stay	60 days (December 22、 2023 - February 19、 2024)
専攻分野	情報科学・情報工学
Major Field	Computer Science/Engineering



大学へ行く途中の写真

＜外国人研究者からの報告/Foreign Researcher Report＞

①研究課題　/　Theme of Research

Explaining artificial neural network behaviors via attribution and visualization

②研究概要　/　Outline of Research

By studying interpretability of artificial neural networks, e.g., autonomous driving models, it is possible to address the current issue of incomprehensible behavior decision-making, especially in safety-critical situations like autonomous driving systems. This plays a crucial role in promoting the development of future artificial neural networks and is a key technology for determining responsibility in artificial intelligence decision making and enhancing user confidence.

③研究成果　/　Results of Research

Through communications with researchers in artificial intelligence and autonomous driving at the University of Tokyo, Cyber Agent, Google, Amazon, Huawei Japan, Ritsumeikan University, and Tsukuba University, I have discovered a combined approach of attribution methods and autonomous driving feature analysis. I am currently working on writing a new academic paper on this topic. Additionally, utilizing this new research approach, I am also preparing a patent application for an invention.

④今後の計画　/　Further Research Plan

In my subsequent research, I plan to incorporate the latest artificial intelligence content generation techniques into the field of interpretability research. By leveraging the recent advancements in natural language modeling, I aim to explore the implementation of an automated driving behavior decision analysis method with natural language explanations. This would be highly beneficial for future general users, as it enhances understanding and transparency in autonomous driving systems.

＜受入研究者からの報告/Research Advisor Report＞

①研究課題　/　Theme of Research

関与度と可視化によるニューラルネットワークの動作説明
（Explaining artificial neural network behaviors via attribution and visualization）

②研究指導概要　/　Outline of Research

ニューラルネットワークの解釈可能性を研究することで、決定過程が理解不能という問題に対処することが可能となる。これによって、ユーザーの信頼を高められ、将来のニューラルネットワーク利用を促進する上で重要な役割を果たす。特に自律走行システムのような安全性第一の問題において、欠かせない技術となるものと考えられる。今回の研究では、関与度の計算法と関与度を利用した可視化手法についての研究を進めた。

③研究指導成果　/　Results of Research

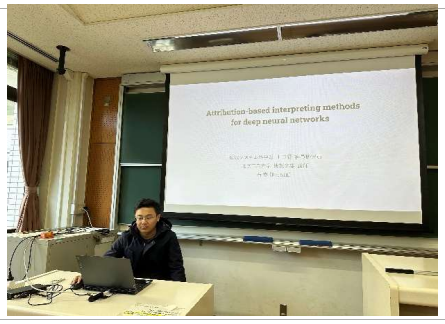
関与度計算手法ならびに自律走行の特徴分析に適用する方法を発見した。今後、この研究成果をもとに新しい論文の執筆に取り組む予定である。また場合によって、特許取得なども試みたいと考えている。

④留学生交流事業の活動状況　/　Activities of International Student Exchange Program

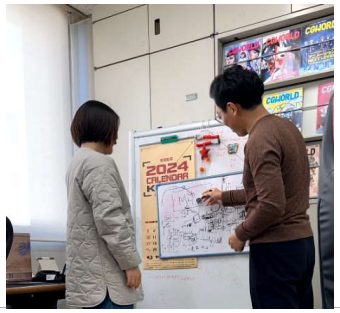
学内においては、研究室のGPUマシンを利用して実験を行ったり、関連する研究室に呼びかけて講演や討論を行ったりした。さらにサイバーエージェント、グーグル、アマゾン、ファーウェイ・ジャパン、立命館大学、筑波大学の人工知能や自律走行の研究者との交流などを持った。

⑤今後の計画　/　Further Research Plan

今後は生成ニューラルネットワークに対する解釈可能性についても研究を展開していきたいと考えている。特に最近急速に進歩している大規模自然言語モデルを活用したマルチモーダル生成モデルなどを手掛けたい．日本学術振興会の外国人招へい研究者や大学の客員教員などの制度を利用することで，定期的な交流を維持したいと考えている．



AI解釈性についての研究発表



研究室内部交流