2024 年 9 月 26 日

Tazaki 財団英国留学支援奨学金
留学報告書

所属	工学院経営工学系		
氏名	長谷川理恵	現在の学年	修士 2 年
渡航先国	イギリス		
渡航先	ロンドン		
渡航プログラム	インペリアル カレッジ ロンドン研究プログラム (IROP)		
渡航期間	2024 年 7 月 1 日～8 月 23 日		

1. My Background

I am currently conducting research in the field of AI as a master student, with a specific focus on large language models like ChatGPT. During my undergraduate studies, I developed a strong foundation in mathematical modeling, which has proven invaluable in understanding the complexities of AI. Additionally, I often join web application workshops in my free time to broaden my technical skills, and I have experience in developing several AI applications.

As I delved deeper into AI research, I realized the importance of gaining a global perspective and collaborating with researchers from diverse backgrounds. Exploring AI in broader contexts, such as its theoretical and ethical aspects, is crucial in today’s rapidly evolving field. This motivated me to pursue studying abroad. I believe that immersing myself in a different academic environment will allow me to gain new insights, enhance my research, and contribute to the field on a global scale.

2. Overview of IROP

I participated in a program called IROP (International Research Opportunities Program), which is a two-month summer exchange program where students conduct research under a supervisor at Imperial College London. I joined the Machine Learning Initiative research group and conducted research on AI, which closely aligns with my background and current research focus. This program wasn't limited to an exchange between Tokyo Tech and Imperial College London, it also included students from MIT, the Technical University of Munich, Cornell University, and the University of Toronto. During our stay in London, IROP organized several social activities, including parties and local trips, providing a great opportunity to network and collaborate with fellow students.

3. Why I Joined IROP

I decided to join this program because its content aligned with my background and research focus. Through this program, I aim to deepen my understanding of AI and acquire the latest theories and technologies. Additionally, I saw it as an opportunity to improve my English skills, both in academic and everyday settings, and to collaborate with global researchers and students. This experience has allowed me to engage with diverse perspectives, working styles, and learning approaches. Furthermore, staying in a completely different culture for an extended period was a great opportunity for personal growth and development.

4. My research Topic at Imperial College

During my stay at Imperial College, I worked on a research project in the field of explainable AI, specifically focusing on the Concept Bottleneck Model (CBM). The goal of my work was to enhance the transparency and interpretability of standard neural networks.

CBM is a type of neural network designed to make AI models more interpretable by introducing an intermediate prediction step. As shown in Figure 1, if we input an image of a classroom, a traditional neural network might directly predict "classroom" without explaining the reasoning process behind this prediction. In contrast, a CBM first identifies human-understandable concepts—such as recognizing the image as containing tables, chairs, and a room—and then makes a final prediction based on those identified concepts.

(CBM: Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In International Conference on Machine Learning, pp. 5338–5348. PMLR, 2020)

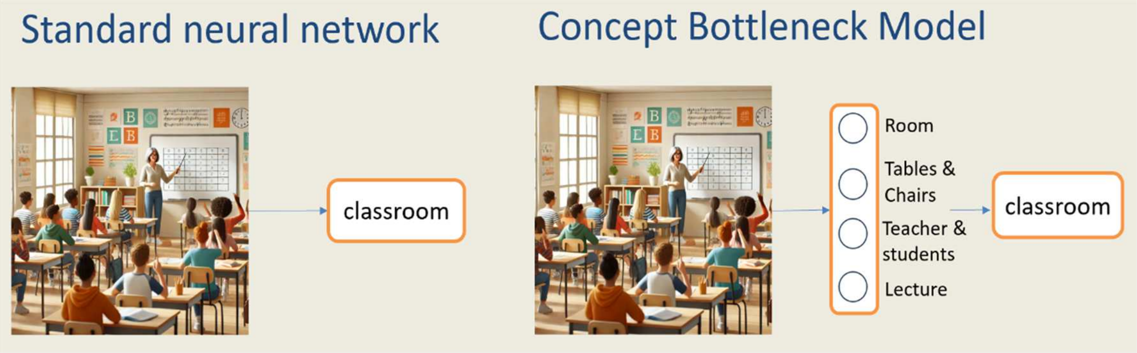


Fig. 1. The Differences between Standard Neural Network and Concept Bottleneck Model

Even though the idea of CBM is simple and clear, it was originally developed for image classification. Therefore, my work involves extending CBMs to a text classification task. Specifically, I applied the CBM to an online harm dataset to detect hate speech.

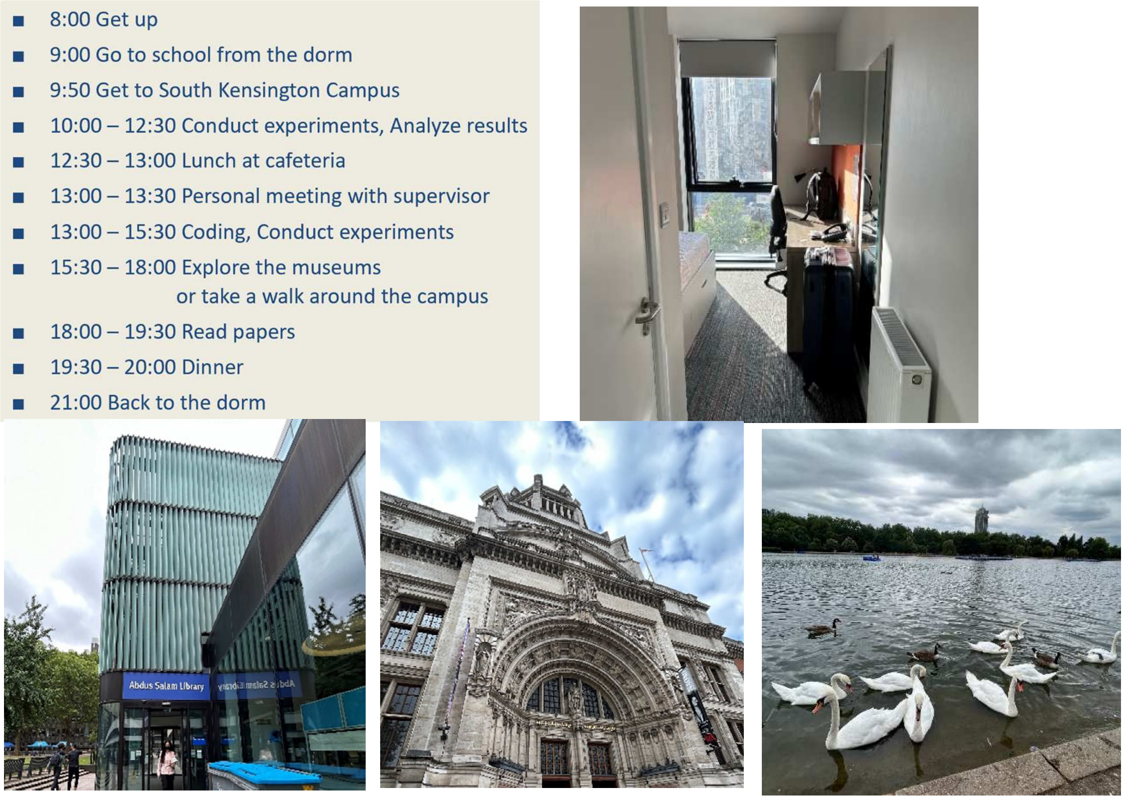
Traditional hate speech detection models typically depend on datasets with predefined labels, such as labeling certain words or phrases as hate speech. This approach can be struggling in dynamic online environments where the language used in hate speech constantly evolves. Additionally, there are new forms of harmful content frequently emerge that are not covered by the existing labels.

By applying the CBM, the model first predicts key concepts, such as detecting derogatory language, targeted group, or harmful intent, before making a final decision. This enables the model to adapt to new patterns of hate speech without relying solely on static, predefined labels, allowing for a more flexible and context-sensitive approach for detecting online hate speech.

I've completed the technical part of the hate speech detector model and evaluated its performance on a multi-class hate speech classification dataset. (include labels like "not hate speech", "derogation", and "dehumanization" etc.)

Although the accuracy is lower than that of standard neural networks, which is a known trade-off with CBMs due to their focus on interpretability. The experimental results show that CBMs can effectively detect hate speech while also providing reasoning for their decisions. Additionally, I observed that as the number of labels increases, the model's performance declines, which presents a promising direction for future research.

5. My Daily Life at Imperial College



- 8:00 Get up
- 9:00 Go to school from the dorm
- 9:50 Get to South Kensington Campus
- 10:00 – 12:30 Conduct experiments, Analyze results
- 12:30 – 13:00 Lunch at cafeteria
- 13:00 – 13:30 Personal meeting with supervisor
- 13:00 – 15:30 Coding, Conduct experiments
- 15:30 – 18:00 Explore the museums
  or take a walk around the campus
- 18:00 – 19:30 Read papers
- 19:30 – 20:00 Dinner
- 21:00 Back to the dorm

Fig. 2. Top Left: My Daily Routine, Right: My room in the dorm,

Bottom Left: The Library, Center: Victoria & Albert Museum, Right: Hyde Park

Figure 2 shows a typical daily routine when I work at school. I usually leave the dorm around 9 a.m., and it takes about 50 minutes by train to reach the South Kensington Campus, where I typically work in the library. Mornings are usually spent to conducting experiments. In the early afternoon, I sometimes meet with my supervisor to discuss the results and the direction of my research. After the meeting, I go back to coding and running more experiments.

In between, I like to take a break and explore the famous museums and parks around the campus. Afterward, I return to the library to read papers related to my work, and I head back to the dorm in the evening.

6. IROP Activities

Beyond my daily routine, the university organized several activities as part of the IROP program. These included a welcome lunch, a campus tour, a traditional British afternoon tea, and a visit to the Globe Theatre to watch a Shakespeare play. There was also a local trip to East London, where we were introduced to the town's history and fascinating street art, followed by a final celebration to conclude the program. These events provided excellent opportunities to immerse myself in the local culture, as well as engaging with fellow IROP students from other prestigious universities, and interacting with staff and professors at Imperial College.

7. Explore London

Aside from my research work, I had the opportunity to explore London and nearby cities in my free time. My supervisor also has an office at the Alan Turing Institute, located in the British Library, so I had the chance to visit as well. The library offered excellent facilities and well-suited for both work and learning, where many local people also spent their time.

Additionally, I visited iconic places such as the British Museum, Big Ben, Tower Bridge, and the National Gallery. London has a lot of museums, each offering exhibitions on diverse themes from not just Europe but from around the world, allowing me to experience the rich cultural and historical heritage of London.

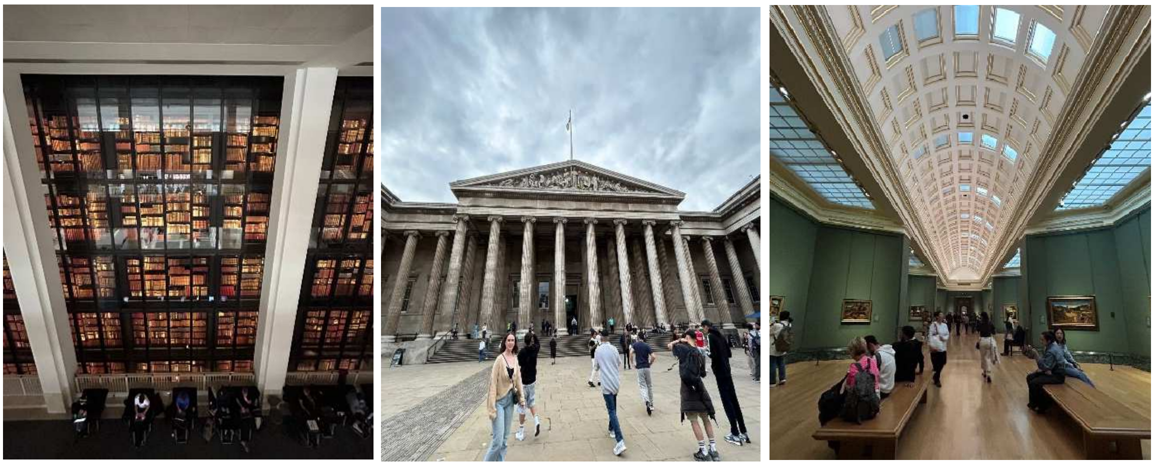


Fig. 3. Left: Collection of Books in British Library, Center: British Museum, Right: The National Galley

8. Explore UK

During my free time, I also took short trips to other cities such as Bath, Cambridge, and Edinburgh. In Bath, I visited the well-preserved ancient Roman baths, which provided fascinating insights into Roman architecture and daily life. In Cambridge, I explored the historic University of Cambridge, I explored the iconic colleges and learning about its prestigious academic heritage. Edinburgh, on the other hand, offered its own unique medieval and Georgian architecture. I also enjoyed strolling through the Royal Mile, where I encountered numerous street performances and lively cultural activities. Exploring these fascinating places was a great experience, though I encountered some unexpected challenges along the way.

One incident occurred when all the trains were suddenly canceled midway while I was returning to London from Edinburgh, leaving me stranded in York for a night. Another time, my phone plan unexpectedly expired while I was visiting Cambridge, leaving me without internet access. I had to rely on maps and ask for directions to find my way back to London. Although these situations were challenging at the time, navigating through them became some of the most memorable parts of my journey.

9. What I Have Learned Through IROP

My time in this program was incredibly enriching. It deepened my understanding of AI and broadened my perspective on its applications. Although conducting research within a two-month period was challenging, engaging with my supervisor helped me discover a promising research topic and develop a solid research plan.

The time constraints taught me how to quickly acquire new knowledge and apply it efficiently to my research and other areas. They also highlighted the importance of not delaying actions, making me more proactive in pursuing my interests and advancing my research.

Additionally, navigating a new environment, both academically and in daily life, enhanced my adaptability and flexibility. This experience has given me greater confidence in tackling new challenges and a sense of ease when facing uncertainties.

10. Future Plans

I plan to apply the models I developed during the IROP program to my current research, as my master's thesis also focuses on text classification tasks. I am particularly interested in further exploring the interpretability aspect to enhance the depth and impact of my work. Next year, I will be starting a research and development role in the industry, where I aim to leverage not only the technical skills but also the global experience and insights I gained through this program to make meaningful contributions to my job.

I also intend to maintain and foster the collaborative relationships I have built throughout this program, as they are invaluable for my future academic and professional growth.

Lastly, I would like to express my gratitude for the generous support I received during my study abroad. Thanks to the scholarship, I was able to gain valuable experiences and fully dedicate myself to my studies and research. The knowledge and experiences I acquired during this time will be of great benefit to my future career.