Motivation letter Ph.D. position MU

Roan Schellingerhout MSc

1 Motivation

1.1 About myself

I was born in Wexford, Ireland and raised in Zaandam, The Netherlands. I am currently 21 years old and I live in Koog aan de Zaan. I got my vwo diploma at the St. Michaël's college in Zaandam in 2018. After high school, I went on to study Information Science at the University of Amsterdam. This year, I completed my Master's degree in Data Science, also at the UvA. During both my Bachelor's and Master's thesis, I realized I have an affinity for research, which is why I would like to pursue a Ph.D.

1.1.1 Major achievements

Due to the fact that I am a very driven person, I have always been inclined to exceed the expectations that were set for me. In high school, this occurred in terms of taking an extra course, namely *Cambridge English*, in addition to my regular curriculum. This course focused on improving my written and spoken English, as well as my general understanding of the English language. By taking this course, I was allowed to sit my final exam in English a year in advance. It also prepared me for the University of Cambridge's Proficiency English exam, which focuses on speaking, writing, listening to, reading, and the general use of English. I passed this exam with 223 out of 230 total points, which resulted in an 'A' as a final grade. This was the highest score achieved for that exam in my school that year.

During my Bachelor's degree, I enrolled in the Honours programme. This programme let me take one additional course per semester. During this programme I took courses on a wide variety of topics, such as neuroscience and apocalyptic literature. For the honours courses, I averaged an 8.4 out of 10, while for my regular courses I averaged an 8.73 - the highest GPA in my year. As a result, I received my Bachelor's degree magna cum laude with honours.

For my Bachelor's thesis I looked into the prevalence of conspiracy videos within YouTube's recommender system. My thesis was awarded a 9.5 as a final grade, which was the highest grade in my year. My supervisor submitted my thesis to the Amsterdam Data Science awards 2021. There, it was judged to be 'award quality' by both reviewers. I consequently received an honourable mention. My Bachelor's thesis is currently under review at PLOS ONE.

I also completed my Master's degree in data science **SUMMA/MAGNA** cum laude with a GPA of **X.X** (unfortunately there was no Honours programma to enroll into). My Master's thesis, which I carried out for Randstad Groep Nederland, was about explainable career path predictions made by neural networks. My Master's thesis was awarded a **X.X**, and was again submitted to the Amsterdam Data Science awards (results for 2022 have yet to be published). It is also currently being submitted to ACM's RecSys in HR workshop, after which we will attempt to publish it in Management Science.

1.1.2 Research and Programming Experience

I have been programming, predominantly in Python, for over 5 years. In addition to Python, I am also proficient in SQL, C, R, JavaScript(/JQuery), and HTML/CSS. During my studies, I have worked on projects for a number of companies, such as Marktplaats, Heineken, Randstad, Aon, and het Rijksmuseum. I am experienced in a number of different subjects, including machine learning (e.g., neural networks, support-vector machines, linear regression), (big) data wrangling and analysis, data visualization, databases, algorithms and data structures, etc. For a number of these subjects I have also worked as a teaching assistant (TA) over the past three years. As a TA I helped students with their homework, was responsible for grading homework, took care of behind-the-scenes responsibilities (e.g., setting up homework environments, communicating with

guest lecturers, verifying exam questions), and I also gave a number of lectures on certain topics myself. For a full list of courses I have taught as a TA, please refer to my CV.

In terms of research, I have gained a lot of experience in the past two years. As previously mentioned, I have realized that I have an aptitude for doing research, as became clear during my Bachelor's and Master's thesis. As a result, I also held a job as a research assistant for research conducted by the University of Amsterdam, the University of Århus, and Lareb last year. For this research, the group looked into the (dis)trust in COVID-19 vaccines and related beliefs and conspiracies. I helped by setting up an ElasticSearch database and a Kibana dashboard in accordance with the authors' demands. For this ElasticSearch database, I filtered, cleaned, and automatically translated tweets/telegram messages which were collected using the Twitter and Telegram API. Additionally, I conducted sentiment analysis using a transformer (BERT) and set up a small-scale experiment to collect labeled data to allow for fine-tuning of the model.

1.2 Interest in the project

1.2.1 Recommender systems

In the first year of my Bachelor's degree, I took a course on recommender systems: *Collective Intelligence*. This quickly became one of my favourite courses in the Bachelor, as I found the programming aspect of the subject intriguing, while it also became apparent how widespread these systems are. This interest eventually led me to conduct both my Bachelor's and Master's thesis on the topic of recommender systems.

For my Bachelor's thesis, I researched the influence of personalization on YouTube's recommendation algorithm; specifically by looking at its tendency to recommend conspiracy videos. To do so, I created bots that would log into newly-created Google accounts, after which they would start watching YouTube videos according to specific *watch strategies*. These watch strategies were divided by the level of personalization used to find the videos that were being watched. I came to the conclusion that 'users' who watched more personalized content ended up in filter bubbles of conspiracy content more quickly. In general, it was also found that YouTube's recommender system is inclined to quickly and significantly develop a preference for recommending conspiracy-related videos to brand-new 'users' - this preference additionally proved hard to be undone.

For my Master's thesis, I looked into recommending jobs for candidates based on their career in an explainable manner. For this research, I created three explainable neural networks based on different architectures, and compared their performance to three non-explainable counterparts (and three non-neural baselines). I then created an online environment for recruiters to judge the predictions and explanations generated by the models to determine their real-world utility. Recruiters found the models, as well as the explanations, to be useful and indicated that they would like to use them in a day-to-day setting.

I believe that my previous research clearly shows my interest and expertise of explainable recommender systems, as well as my inclination to perform user studies (either through bot 'users' or actual users such as Randstad's recruiters). For more details, I refer you to my Master's and Bachelor's thesis, both of which you can find as an attachment to my e-mail.

1.3 Maastricht and UM

In recent years I have gone to Maastricht on holiday a few times. Every time, the city struck me as having a cozy atmosphere, while still feeling urban. I feel like Maastricht is one of the few Dutch cities that is able to offer both a community-like environment and an urban feel, which is why I would enjoy living there.

For Maastricht University specifically, I feel like its focus on interdisciplinarity relates strongly to my background in information science. After high school I specifically chose not to pursue a degree in computer science, due to the lack of non-beta elements in that degree. Information science focused on a plethora of non-beta subjects, such as economics, psychology, and philosophy in addition to its programming and math-related courses. As a result, I have come to appreciate the importance of including perspectives from different fields into my research.

This also comes to fruition within the ERAI group itself. While I appreciate the group's ability to design and work with state-of-the-art models, I am also intrigued by the group's tendency to embed their research within society at large. When looking at the ERAI's research, it becomes clear that the focus lies not solely on how to improve upon state-of-the-art performance; the impact that AI has on society, and how it can be made useful in the real world are also taken into consideration.

2 Research Proposal

2.1 Introduction

For my thesis, I would be interested in creating a model-agnostic, multi-level explainability technique for recommender systems. This technique would be able to determine which features were decisive in giving the specific recommendation(s) generated by the model, akin to Selvaraju et al. [15] their Gradient-weighted Class Activation Mapping (Grad-CAM) and Ribeiro et al. [13] their Local Interpretable Model-Agnostic Explanations (LIME). This technique would therefore be able to 'argue' in favour of a specific recommendation. When generating a list of recommendations, this allows the model to provide a different explanation for each individual item in the list. As a result, stakeholders will be able to consider each recommendation, with its corresponding explanation, separately, allowing them to better evaluate their options. Although a number of alterations of both Grad-CAM and LIME have been created over the years (e.g., Grad-CAM++ [3], QLIME [2], KL-LIME [12]), the improvements over the originals have largely been marginal. As a result, the default Grad-CAM and LIME algorithms will be used as a starting point for the research.

While Grad-CAM has proven to generate sensible explanations that are easy to understand, it is exclusive to convolutional layers. LIME, on the other hand, is universally applicable (i.e., to tabular data, text, and images), but has proven to generate unintuitive explanations that are hard to understand without reading the documentation [6]. The aim of my research would be to combine the benefits of both techniques, while minimizing the drawbacks. Specifically, the explanations generated by the model will be designed in such a way that they are flexible to their use case. Depending on the stakeholder, different types of explanations can be satisfactory [1]. For a machine learning engineer evaluating a model's behaviour, it can be useful to have a highly detailed visualization of which features have the most impact on recommendations (both globally and locally), while for a regular user, a much more simplistic explanation would suffice (e.g., 'This job is a good choice, because the employer pays a lot of attention to *X skills*, which you possess.').

Furthermore, I would attempt to create a standardized methodology for evaluating model explanations. Currently, there is no scientific consensus on how explanations generated by XAI models should be evaluated (e.g., baseline evaluation vs. user-based evaluation, when can an explanation be considered 'good') [4, 6]. As XAI is designed to make models more understandable and interpretable for humans, I believe a human-centered approach to evaluation is crucial. As a result, I will attempt to answer the following research questions:

- How can current XAI techniques be improved in order to allow for more comprehensive explanations?
- How can explanations generated by (deep learning) models be made adaptable to the needs of different stakeholders?
- How can explanations generated by (deep learning) models be evaluated effectively?

The specific approach I intend to take in order to answer these research questions will be described in section 2.2.

2.2 Methodology

The implementation of this technique will be done in PyTorch [11]. PyTorch allows for easy retrieval of the gradients running through a model, which can be useful when examining a model's behaviour [4]. Additionally, PyTorch models allow easy inclusion/exclusion of modules, making it possible to easily add modules used for explainability to existing models. The aforementioned improved XAI technique will be implemented, which will allow models to generate explanations at the 'highest' (most difficult to interpret) level, intended mainly for AI experts. Then, separate techniques, such as the creation of saliency maps and automatic text generation will be used to modify the high-level explanation to fit lower levels of expertise.

In order to evaluate the quality of the explanations, two separate steps will need to be taken. Initially, I will attempt to use my connections to Randstad to create a large-scale career-related data set that can be used as a baseline. As I discovered during my Master's thesis, for a large number of recommendation tasks (such as career path prediction), it is quite difficult to determine what models constitute the state of the art, due to most research using manually created or privately received, closed-access data sets. This new data set would focus specifically on career-related tasks, such as career path prediction, information retrieval (e.g., in terms of resumé search), regression problems (e.g., in terms of salary prediction), etc. Obviously, this data set would need to be

cleaned and made completely GDPR-compliant [7] in order to ensure the individuals' privacy will be safeguarded. After having created the data set, I will be able to train and evaluate the different models, as well as their explanations, for different problems, data types, and stakeholders.

Secondly, I will perform user testing of the different types of explanations. Although the aim of making AI explainable is to make it more interpretable for humans [8], there is a surprising lack of user studies when it comes to explainable AI research [6]. While a substantial number of papers make use of baseline evaluation, there are only a handful of XAI papers that actually test their models with users or stakeholders [9, 5, 10]. I already expanded on this research in my Master's thesis, but due to time constraints, the sample size used in user testing was rather small, which led to high variance and therefore made the results somewhat inconclusive. For this research, I will be able to expand upon the environment I created to judge recommendations for my thesis. By including more types of explanations (i.e., different types of tabular and text-based data explanations), and creating a more exhaustive list of questions for users, I will be able to generate more conclusive results. User testing could either be done by a variety of stakeholder, or through a service such as Amazon's Mechanical Turk [14], depending on the task at hand (for tasks that require more expertise, stakeholders could be interviewed, while for easier tasks, Mechanical Turk would be fitting).

References

[1] Himan Abdollahpouri and Robin Burke. "Multistakeholder recommender systems". In: *Recommender systems handbook*. Springer, 2022, pp. 647–677.

[2] Steven Bramhall et al. "Qlime-a quadratic local interpretable model-agnostic explanation approach". In: *SMU Data Science Review* 3.1 (2020), p. 4.

[3] Aditya Chattopadhay et al. "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks". In: *2018 IEEE winter conference on applications of computer vision (WACV)*. IEEE. 2018, pp. 839–847.

[4] Arun Das and Paul Rad. "Opportunities and challenges in explainable artificial intelligence (xai): A survey". In: *arXiv preprint arXiv:2006.11371* (2020).

[5] Amit Dhurandhar et al. "Model agnostic contrastive explanations for structured data". In: *arXiv preprint arXiv:1906.00117* (2019).

[6] Jürgen Dieber and Sabrina Kirrane. "Why model why? Assessing the strengths and limitations of LIME". In: *arXiv preprint arXiv:2012.00093* (2020).

[7] European Commission. *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance)*. 2016.

[8] David Gunning. "Explainable artificial intelligence (xai)". In: *Defense advanced research projects agency (DARPA), nd Web* 2.2 (2017), p. 1.

[9] Himabindu Lakkaraju et al. "Interpretable & explorable approximations of black box models". In: *arXiv preprint arXiv:1707.01154* (2017).

[10] Scott M Lundberg and Su-In Lee. "A unified approach to interpreting model predictions". In: *Advances in neural information processing systems* 30 (2017).

[11] Adam Paszke et al. "PyTorch: An Imperative Style, High-Performance Deep Learning Library". In: *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 8024–8035. URL: `http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf`.

[12] Tomi Peltola. "Local interpretable model-agnostic explanations of Bayesian predictive models via Kullback-Leibler projections". In: *arXiv preprint arXiv:1810.02678* (2018).

[13] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ""Why should i trust you?" Explaining the predictions of any classifier". In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016, pp. 1135–1144.

[14] Joel Ross et al. "Who are the turkers? worker demographics in amazon mechanical turk". In: *Department of Informatics, University of California, Irvine, USA, Tech. Rep* 49 (2009).

[15] Ramprasaath R Selvaraju et al. “Grad-cam: Visual explanations from deep networks via gradient-based localization”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 618–626.