WASP - Software Engineering Assignment

Ricky Molen

June 2024

1 Disclousure

Grammarly has been used to do spell corrections and find synonyms. Furthermore, some chatGPT-4o has been used for rephrasing my text.

2 Question 1

My research focuses on variational inference (VI), a Bayesian method in machine learning for approximating complex posterior distributions. I work on extending and enhancing fundamental VI methods, aiming to make models like variational autoencoders (VAEs) more expressive and robust. These improved techniques can be used in different fields, for example in life sciences, where I use them to study gene evolution through classical phylogenetic inference and more advanced evolution models. My work includes mathematical proofs and model testing on benchmarking datasets.

3 Question 2

3.1 Testing / QA on data level

In my research, the data I work with has different levels of complexity. In some cases, the data consists of a set of DNA strings, which are relatively straightforward and less prone to large errors, at least in my part of the process, as the main work of the preprocessing is done before I see the data. However, more complex scenarios involve spatial transcriptomic data, which includes spatial positions and gene counts of each cell. This type of data still requires extensive and often messy pre-processing steps, such as defining local environments and normalizing counts. Roberts's talk on rigorous testing and quality assurance is relevant here, as ensuring the accuracy and reliability of the data can be essential for producing valid results. This principle should guide me in checking and validating the preprocessed data to minimize errors and spend less time debugging. The importance of testing goes much further than just making sure the data is correct. However, I have noticed a larger issue with the data since you normally only look at it in the beginning and then assume that the process is working correctly even if some things change.

3.2 Requirements Engineering: Understanding the problem is (more than half) the problem

Given the interdisciplinary nature of my field, understanding the requirements can be particularly challenging. My background is not in biology, so it is crucial to partner with biologists who can help interpret and control the outputs of my models. Robert's principle that "understanding the problem is more than half the problem" is particularly relevant to me. It is easy to innovate once the problem and purpose are clear, but getting that level of understanding can be a long process and might therefore be overlocked. This idea highlights the importance of thorough requirements gathering and interdisciplinary collaboration in my research. By ensuring a deep understanding of the problem, I can better tailor my processes and means to effectively handle the research questions, ultimately leading to better outcomes and hopefully faster.

4 Question 3

4.1 Avoidance to New Tech

Reflecting on the resistance to "AI"/ML adoption observed in the engineers at SAAB, similar skepticism is present in life sciences areas. Traditionally, this field has relied on statistical tools given the inherently stochastic nature of biological problems. For example, in phylogenetic inference, it took considerable time for Bayesian analysis to replace traditional methods based on phenotypic or morphological traits. Over time, Bayesian analysis has become a standard for in-depth analysis. A similar change is currently happening with deep learning, while some experts doubt its relevance, others argue it is essential for relevance in modern research. The survey responses from SAAB engineers, such as uncertainty about long-term strategies and skepticism towards "AI", represent a similar ongoing discussion in life sciences about the utility and relevance of deep learning, highlighting the common challenge of integrating new technologies into established fields.

4.2 Behavioral Software Engineering

Behavioral software engineering (BSE) studies the human aspects of software development, including collaboration, and productivity. While this concept can be intriguing, I am unsure of its direct relevance to my research. My work heavily focuses on model development in deep learning models used in life sciences, where the immediate challenges are often rather computational or methodological than behavioral. However, I can see the relevance of understanding different goals, incentives, and collaboration dynamics when working with others, especially given the interdisciplinary nature of the field and the complexities of academia and funding. While BSE may not impact me directly, appreciating its principles can improve collaborative efforts.

5 Question 4

First, I need to add that there were quite a few papers that I was able to identify to be of any relevance to me, hence my approach became to pick articles based on interesting keywords and then try to see how they can be used. It might be worth keeping in mind.

5.1 paper 1 - Dataflow Graphs as Complete Causal Graphs

5.1.1 a

The idea of the paper "Dataflow Graphs as Complete Causal Graphs" is to leverage flow-based programming (FBP) to create complete causal graphs for software systems. This method addresses a challenge in modern software engineering of tracking and understanding causal connections between system components, which can lead to significant technical debt. By utilizing dataflow graphs produced through FBP, the authors propose that these graphs can serve as structural causal models. This connection can enhance various aspects of software engineering, including fault localization, business analysis, and experimentation. This idea is important because it can provide a clear, efficient way to understand and manage complex data dependencies, which can be used for developing more reliable software.

5.1.2 b

I don't know how this relates to my research directly but I can see how it could be improved by better understanding the long and messy process of cleaning the data used in a lot of medical research. It is common that the pre-processing steps are done by someone else or by a machine and then handed over for even more processing. However, my work mainly focuses on creating the tools for other researchers to apply, which means that I try to provide the framework and the mathematical motivations and then leave the implementation to the user. However, if the preprocessing is done vastly differently it will impact the model output in a hard-to-explain way.

5.1.3 c

In a bigger software project, incorporating the main idea from this paper using FBP to generate an estimate of the causal dependencies could enhance the project's ability to manage and troubleshoot complex data flows. For instance, in a bioinformatics data analysis project, these causal graphs could help pinpoint the sources of errors or unexpected results in data processing pipelines. My research on AI and the development of robust models could integrate into this project by providing advanced methods for studying and interpreting the data represented in these causal graphs. The synergy between FBP data dependencies and the data analysis techniques could lead to a more reliable and interpretable system.

5.1.4 d

To make "AI" engineering in the project more effective based on the idea from the paper, my research could adapt by incorporating FBP principles into the design of data preprocessing and analysis pipelines. This would involve structuring these pipelines to naturally produce dataflow graphs, thereby making causal relationships more explicit. Additionally, developing tools and algorithms that leverage these causal graphs to enhance model training and evaluation processes could further improve the robustness and accuracy of models. By aligning my research methodologies with the principles of FBP, we could make it easier to troubleshoot and maintain.

5.2 paper 2 - Influence-Driven Data Poisoning in Graph-Based Semi-Supervised Classifiers

5.2.1 a

The core idea of the paper "Influence-Driven Data Poisoning in Graph-Based Semi-Supervised Classifiers" is to propose a novel data poisoning method specifically for graph-based semi-supervised learning (G(SSL)) algorithms. This method aims to identify the most influential data points to maximize the error rate of the classifier such that they can be double-checked. The authors present an influence metric to approximate the impact of poisoning-specific inputs on the overall accuracy of the model. This approach is important because it highlights vulnerabilities in GSSL systems and provides insights into potential attack vectors that could be exploited, therefore underlining the need for robust defense mechanisms. In engineering software systems, especially those relying on machine learning, understanding and mitigating such vulnerabilities is crucial to ensure system reliability and integrity.

5.2.2 b

While this paper does not directly relate to my research, it brings to light important concerns about data quality and security. The concept of influence-driven data poisoning is relevant in the context of my work on preprocessing and analyzing biological data. Ensuring that the data used in my models is free from unintended label noise can be very valuable for a semi-supervised learning (SSL) approach (and even in general of course), which in many cases can be a desirable method given that the cost of labeling data can be very expensive. Understanding these potential vulnerabilities helps in designing more robust preprocessing pipelines that can detect and manage the risks of badly labeled data.

5.2.3 c

In a larger "AI"-intensive software project, the ideas from this paper using an influence metric to identify critical data points for potential poisoning could be applied to improve data quality. For instance, in a bioinformatics data analysis project, implementing an influence-driven approach could help identify and protect against critical errors in the data before they affect the downstream analysis. AI and robust model development could integrate into this project by providing alternative techniques for handling and analyzing the data, ensuring that the models remain accurate and reliable despite potential data quality issues. The combination of influence-driven data quality assessment and VI methods could enhance the overall effectiveness and security of the software project.

5.2.4 d

I could incorporate techniques for influence-based data assessment into my preprocessing workflows. This would involve developing tools to calculate the influence metric for different data points and identifying those that are most susceptible to errors or manipulations. By integrating these tools, it can be possible to enhance the robustness of the data used in "AI" models. Additionally, collaborating with domain experts to validate the influence-based assessments and ensuring that the preprocessing steps are tailored to identified risks could further improve the reliability and accuracy of the models, ultimately leading to more trustworthy research outcomes.