WASP Software Engineering Assignment

Jingyu Guo

August 24, 2024

1 Research Topic

Research in vision transformers (ViTs) has achieved great success in many deep learning-based computer vision tasks. However, most existing works focus mainly on natural images. As a result, it is hard to adapt and apply these methods to medical images due to their particular characteristics. In my PhD project, we will investigate challenges to medical image processing with ViTs.

Some of the most pressing challenges are computational in nature. Medical images are often high-resolution, which poses a challenge for ViT models, with their steep computational and memory requirements. The cost of the self-attention mechanism, the main part of ViT models, is quadratic in the number of tokens (the resolution of input images in the case of image processing). This problem significantly harms the sustainability and limits the application of ViTs in medical image processing. Another challenge is the dependence of ViTs on large datasets for training. Medical image datasets are relatively small due to the cost of collecting and annotating medical data. In addition, it may be difficult to obtain a sufficient number of examples of certain rare conditions. Although this issue may be solved by pre-training to some extent, we will investigate more data-efficient training methods for ViTs.

We are also interested in improving the performance of ViT models, as well as their interpretability and reliability, in the application of medical image processing. On the one hand, it is still to be explored how the SA mechanism can be leveraged to allow ViT models to give precise predictions. On the other hand, it is also crucial that we understand how they make decisions, i.e., what they base their decisions on, to ensure trustworthiness and responsible AI.

2 Lecture Takeaways

Although my PhD mainly focuses on data-driven AI algorithms, I find some of the ideas and principles of software/AI engineering very much related and interesting.

During Robert's lectures, we discussed the importance of quality assurance and testing in AI projects, which I believe is not yet well-developed, particularly in deep learning-based systems. Deep learning has proven to be the most effective data-driven method, given the large-scale data we can access nowadays. However, it is a well-known fact that most deep learning models are

black boxes. They are designed and trained in an end-to-end manner, making it challenging or even impossible to understand or explain their underlying working mechanism, let alone ensure quality. For instance, one research project I completed involved designing a deep learning model for breast cancer diagnosis. Although there are established protocols to evaluate the performance (mainly quantified by accuracy or similar metrics) of the model in a research setting, we still cannot be certain how the model makes specific decisions. In other words, the dilemma between its effectiveness and lack of interpretability poses new challenges to quality assurance. Fortunately, researchers and engineers have already recognized this problem and are raising awareness about it. New fields such as explainable AI and responsible AI are emerging to tackle this issue. And I am glad that part of my PhD research involves investigating the interpretability of ViT/deep learning models for medical purposes.

Another interesting point from the lecture is the concept of team diversity as one of the group-level behavioral SE (BSE) concepts. I was impressed because I previously held a narrow understanding of diversity, focusing only on social diversity. As a result, I sometimes failed to recognize its importance and believed that diversity should only come naturally as a bonus, not a necessity. Although it is still an ongoing debate about how to ensure diversity, I am fulfilled to have a more comprehensive understanding of its importance and realize how crucial it is to focus on individuals, even in engineering as well as research. Based on my own experience, the positive outcomes of having a diverse group, where everyone has unique beliefs, values, knowledge, and backgrounds, have indeed benefited me without my explicit realization.

During Dr. Lenberg's lecture, the importance of BSE was reinforced and discussed from an industrial perspective, highlighting the critical need to integrate human and behavioral aspects into software development processes. Although I really appreciate Saab ATM's recognition of the significance of BSE, the lecture provided a general and basic overview, and I would have enjoyed more details on how BSE is implemented in practice. For example, what are the specific applications of BSE concepts and tools in real-world scenarios and how are they applied? As someone who has not yet worked in a large group and might do so in the future, I am particularly interested in understanding what an individual can do to positively impact everyone's work in a collaborative setting with complex relationships, as the slides showed for example.

Based on both the guest lectures, I would like to discuss the integration of AI/ML in industry, focusing on the varying perspectives and challenges that arise. Surprisingly, the views on AI integration differ between Saab ATM and Volvo. While the former group shows uncertainty about the future of AI and foresees limited positive outcomes, the latter sees great potential in AI and is already leveraging it in existing pipelines. For instance, some groups at Volvo are utilizing large language models (LLMs) to generate unit test cases. I am intrigued by the variance in opinions and wonder whether the examples we learned are universally representative. However, one key takeaway is that AI is sometimes overrated, or at least not powerful enough to significantly impact certain areas, which aligns with my own observations in my field. Furthermore, I believe that more attention and discussion are needed on the topic of responsible AI. As we discussed with Mr. Dhasarathy, integrating AI into software development and production processes requires the establishment of corresponding accountability mechanisms to address po-

tential issues that may arise. For example, whether AI is used to generate test cases or not, it is still the engineers' responsibility to ensure quality. This is also particularly critical in my field, AI for healthcare, where decisions that are vital to life should not be entrusted to AI without careful consideration.

3 Paper Discussions

Here are the two papers I picked for discussion. Where needed, I will refer to my recent project–deep learning for breast cancer diagnosis–as an example when talking about my research.

3.1 Data Smells in Public Datasets [2]

Core ideas. The paper proposes the concept of data smells, i.e., recurrent data quality issues in datasets that can lead to problems in training machine learning models. Through analysis of public datasets, the authors identify 14 data smells categorized into redundant value smells, categorical value smells, missing value smells, and string values smells [2]. The authors believe that data quality is critical to the success of AI systems, and poor quality training data containing smells can result in biased, unfair, or poorly performing models. Therefore, they argue for a data-centric approach to AI, especially for high-stakes applications. They also claim that tools and best practices to aid data scientists in analyzing datasets are currently lacking compared to traditional software engineering. And data smells provide a framework for data scientists to systematically analyze datasets and catch potential problems.

Relation to my research. My research is under the topic of deep learning, in which high quality training data is essential for developing accurate, unbiased models that can be reliably deployed in clinical settings. Several problems, or data smells from the paper, are observed in my own research, such as data imbalanced, presence of sensitive features, and missing values. Awareness of these smells allows me to analyze my data comprehensively, which is critical for designing both high-performance and trustworthy methods and systems.

Integrating into a larger AI project. To answer this question, consider specifically an AI software project aiming to build an automatic breast cancer risk assessment system incorporating mammograms and data from other tests and exams. My research results on deep learning for mammography classification can be one component of this system. Data smells could improve the project by: 1) Ensuring the mammography dataset is high-quality and balanced before training diagnostic models. 2) Checking for possibly sensitive or redundant features that could introduce bias, such as lifestyle and family history. By delivering a robust, validated model trained on "smell-free" data, it would integrate effectively with the other project elements. The system could then provide more valuable guidance to patients and clinicians.

Adapting my research. The proposed method encourages everyone working with data to evaluate their datasets and training practices. Although I agree with their findings and suggestions,

it would require a significant amount of work and knowledge to achieve such extensive evaluations. In some cases, it may even be impossible to analyze the data without the assistance of specialists. For example, it is challenging for a data or AI scientist to analyze medical data as proficiently as a medical doctor. Therefore, what I take from this paper is that active multidisciplinary collaboration is crucial for improving data quality, such as addressing data smells. With that being said, the existing pipeline of my research can be easily adapted by implementing more rigorous preprocessing steps to handle issues like class imbalance and missing data.

3.2 Towards Concrete and Connected AI Risk Assessment (C^2AIRA): A Systematic Mapping Study [3]

Core ideas. The core idea of this paper is to conduct a systematic mapping study of existing AI risk assessment frameworks in order to identify their key characteristics, capabilities, and limitations. The study provides a comprehensive analysis of 16 frameworks from industry, government, and NGOs, focusing on their alignment with responsible AI principles, stakeholder involvement, and lifecycle stages [3]. By doing so, the authors highlight the need for a more concrete and interconnected frameworks that can effectively assess and mitigate AI risks, which is important for the engineering of AI systems as it ensures that they are developed responsibly, minimizing potential negative impacts on society. The insights from this study can also benefit the development of future frameworks, and thus enhance the reliability and trustworthiness of AI systems.

Relation to my research. The paper is highly relevant to my research on using deep learning for breast cancer diagnosis. As an AI system in the sensitive domain of healthcare, it is critical that the potential risks of such a system are thoroughly assessed and managed. The paper provides valuable insights into the current state of AI risk assessment frameworks and their limitations. Specifically, the proposed characteristics, such as covering diverse stakeholders and providing structured mitigation plans [3], could help guide a more comprehensive risk assessment for my and any other systems.

Integrating into a larger AI project. The insights from this paper could be beneficial and applied to a hospital system that incorporates a breast cancer diagnosis model as one component. Having a concrete and connected risk assessment framework, as proposed in the paper, will allow the project to identify and mitigate risks across the entire system and its AI components, which could cover the risks in the diagnosis model, like biased training data or lack of explainability, as well as broader system-level ones around data privacy, system reliability, etc. Considering these perspectives will enable a more robust "responsible-AI-by-design" approach [1, 3].

Adapting my research. There are several ways to adapt my research into a C^2AIRA-compliant project. For example, stakeholder engagement can be expanded during development to surface more diverse risk perspectives, especially from clinical end-users. Model documentation should to be structured to align with the requirements of the C^2AIRA framework. It is also important to enhance testing and evaluation to measure the risk factors highlighted in the paper, such as

bias and explainability metrics. Overall, proactively assessing and mitigating risks in AI-based diagnosis systems would help enable a more responsible solution.

References

[1] Qinghua Lu, Liming Zhu, Xiwei Xu, and Jon Whittle. Responsible-ai-by-design: A pattern collection for designing responsible artificial intelligence systems. *IEEE Software*, 40(3):63–71, 2023.

[2] Arumoy Shome, Luís Cruz, and Arie van Deursen. Data smells in public datasets. In *2022 IEEE/ACM 1st International Conference on AI Engineering – Software Engineering for AI (CAIN)*, pages 205–216, 2022.

[3] Boming Xia, Qinghua Lu, Harsha Perera, Liming Zhu, Zhenchang Xing, Yue Liu, and Jon Whittle. Towards concrete and connected ai risk assessment (c2aira): A systematic mapping study. In *2023 IEEE/ACM 2nd International Conference on AI Engineering – Software Engineering for AI (CAIN)*, pages 104–116, 2023.