Technische Hochschule
Brandenburg
University of
Applied Sciences

Canvassing the Challenges Small Companies Face when Training Machine Learning-based Systems

Lukas Teutenberg

Brandenburg University of Applied Sciences

e-mail: lukas.teutenberg@th-brandenburg.de

General structure

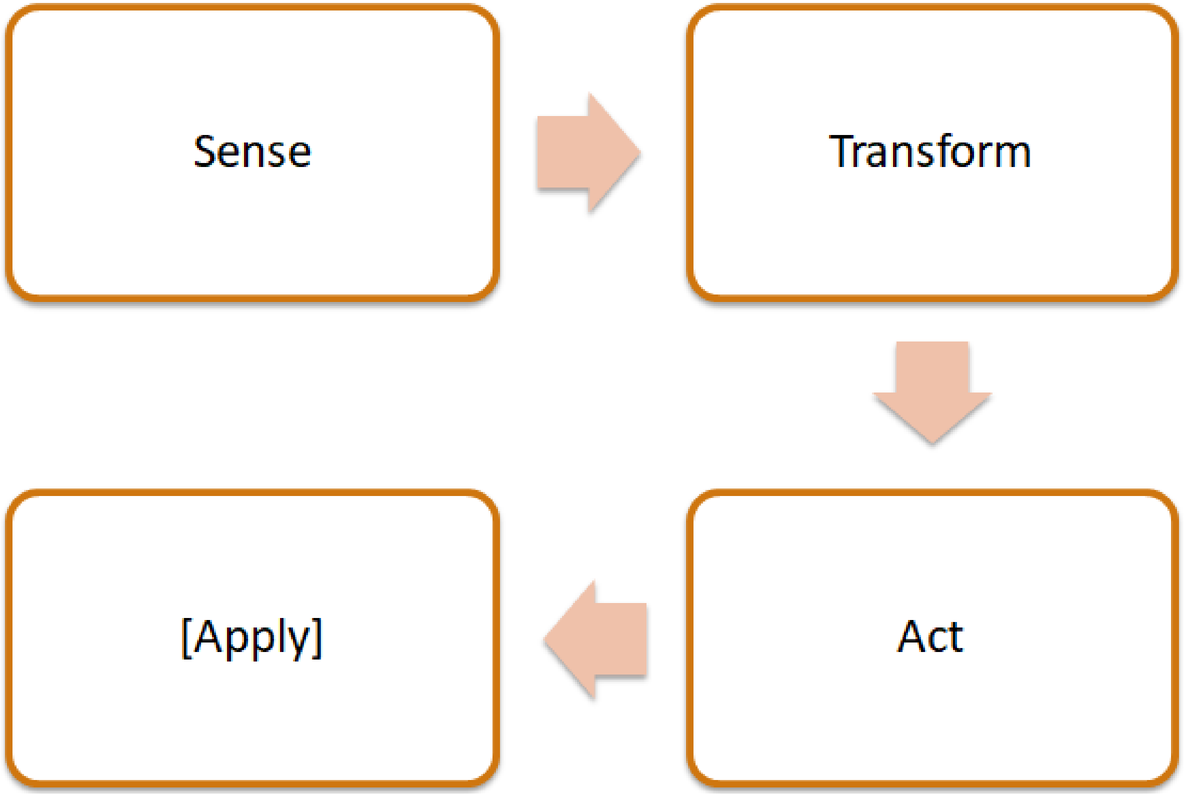
- Introduction/ Motivation
- Research Methods
- Preliminary insights
    - Challenges
    - Mitigations
    - Expectations
- Conclusion and Outlook

Introduction and Motivation

- Development in technologies and software provides extended possibilities
- Often discussed are Machine Learning Systems (MLS)
- MLS are already used in financial and medical contexts (Gayathri 2013)
- MLS are based on behavioural data and applied in an inter-person context
- How can ethical application of these systems be ensured or examined?

=> <u>First step</u> towards canvassing challenges, mitigations and expectations of small MLS-developing companies

Ethical challenges in MLS development



Sense: Finding Data
Transform : Transforming Data
Act: inserting data into MLS
[Apply] - Outcome

Based on Shutt and O'Neill 2013

Research questions

- What challenges do small MLS-developing companies have to face?
- How do they mitigate these challenges?
- What expectations do these companies have towards their surrounding systems and institutions?

Research method

- Qualitative research
    - Structured interview with a developer of MLS
        - Discussion regarding general questions and questions raised by former research
    - Unstructured interview for additional insights with an IT-specialist

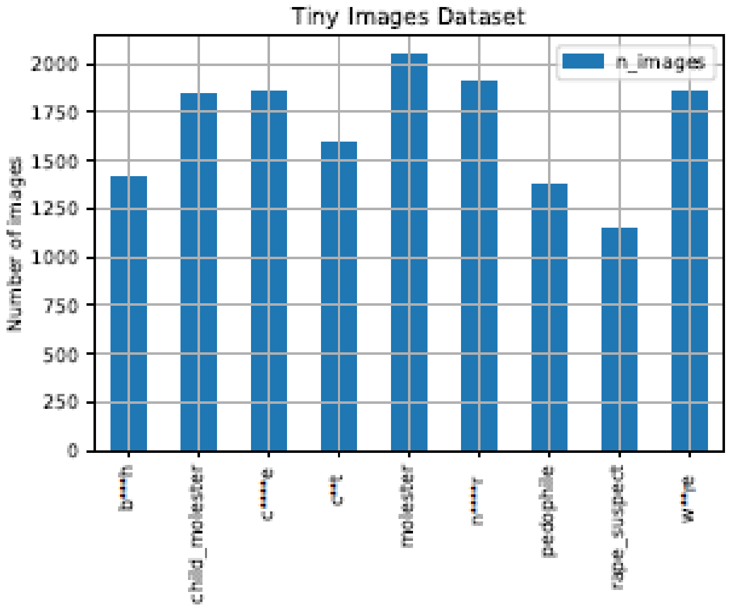
- Literature research

Preliminary insights: challenges I

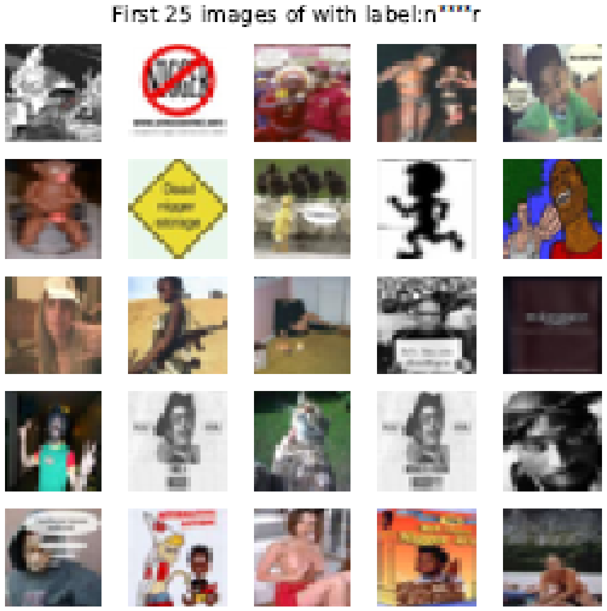
- Dataset acquisition challenges
  - Three ways to acquire a dataset, each with its own challenges
  - Creating new datasets
    - Time-consuming and expensive
    - Sometimes provided through unethical means like datamining
      - "From the perspective of statistical practice, data mining raises [...] different sort of ethical issues" (Seltzer 2005, p. 1)
  - buying a license for a dataset
    - Datasets made available by universities and other providers like private companies
    - Safest but expensive solution
      - 100h of audio data for the training of voice recognition software can cost up to 3.000,-€ (Interview 2021)
    - No guarantee for ethical data collection and representation
      - "insufficient diversity across many demographic groups" (Peng 2021, p. 7)

Preliminary insights: challenges II

- *Free of charge dataset*
  - Most of the time there is no clear legal situation (Paullada, Raji 2020, p. 8)
  - Easy to come by, but rarely audited for bias
    - "[...], these datasets have never been audited or scrutinized [...]" (Prabhu, Birhane 2020, p. 3)
  - Data can have wrong and toxic labels



(a) Class-wise counts of the offensive classes



(b) Samples from the class labelled n****r

(Prabhu, Birhane 2020, p. 4

Preliminary insights: challenges III

- Ethical Data Audit Challenges
  - Time consuming and expensive
    - Requires rotating personal and extensive preparations (Interview 2021, p. 7)
  - Overrepresentation of major ethnic and other demographic groups
  - Exclusion of minority groups like LGBTQIA+
    - Data can be biased and discriminating towards minorities (Interview 2021, p. 2 and Peng 2021, P. 7)
    - Often just labelled as their biological gender or their former biological gender (Interview 2021, p. 8)
  - Wrong Meta data falsifies expectations and outcomes of the MLS (Roh 2018, p.5, Interview 2022)

Preliminary insights: mitigations

Challenge	Expensive Audits	Over- and underrepresentation of Groups	Time and cost intensive dataset creation	Bias data
Mitigation	Often no ethical audit (Findlay, Seah 2020, p. 5)	Data augmentation and generation (Del Campo 2021, p. 5 and Roh, Heo, Whang 2019, p. 2)	Hiring cheap personal like student workers to collect data (Interview 2021, p. 7)	Internal ethics guide (Interview 2021, p. 2)
Mitigation Benefits	Enables the company to stay competitive	Altering existing images of underrepresented elements and groups Allows to expand on existing data Can be automated	Minimal knowledge needed cheap solution	Create an environment where including ethical choices is a must-do Allows for ethically reliable outcomes Schools personal on importance of ethical factors Long-lasting results
Mitigation Drawbacks		Only minimally scalable Only reinforces already established outcomes	Temporary	Time and cost intensive

Preliminary insights: expectations (Interview 2022)

- Towards developers of datasets
  - Heightened awareness towards bias and social fairness of data sets
  - More variety in the teams = more variety in data
- Towards customers
  - Be willing to pay for the work that needs to be done
  - Acknowledge the need for ethical audits and data acquisition
- Towards institutions and governments
  - Assistance in data acquisition
  - Ethics guidelines and requirements for established data sets
  - Financial relief and subsidies for dataset and MLS developing Companies

Conclusion and Outlook

- A lot of data is unaudited and unrestricted in its labelling(Prabhu, Birhane 2020, p. 3)
- Underrepresentation of minority groups (Interview 2022 and Peng 2021, P. 7)
- All around heightened awareness is a must
- Developers are aware of the problem and see room for advancement for social-awareness in MLS (Interview 2021)
- Guidelines needed to unify ethics-implementation in datasets
- Companies need assistance through government funding

- Taxonomy of challenges, mitigations and expectations is needed for a clearer picture and action potential
- This can be achieved in the dialogue between software developers, users and legislation => Further, more in-depth interviews with experts

References

- B.M Gayathri, C.P Sumathi, and T. Santhanam, “Breast Cancer Diagnosis Using Machine Learning Algorithms - A Survey,” May 2013
- W. Seltzer, “The Promise and Pitfalls of Data Mining: Ethical Issues.”, 2005, p. 1, r.34-35
- K. Peng, A. Mathur, and A. Narayanan, “Mitigating Dataset Harms Requires Stewardship: Lessons from 1000 Papers,” Aug. 2021, p. 7, r. 9
- A. Paullada, I. D. Raji, E. M. Bender, E. Denton, and A. Hanna, “Data and its (dis)contents: A survey of dataset development and use in machine learning research,” Dec. 2020
- V. Prabhu, A. Birhane: "Large image datasets: A pyrrhic win for computer vision?" Jul. 2020, p. 3 under 2.3 The WordNet Effect, r. 3
- Y. Roh, G. Heo, and S. E. Whang, “A Survey on Data Collection for Machine Learning: a Big Data -- AI Integration Perspective,” Nov. 2018
- M. Findlay, J. Seah: "An Ecosystem Approach to Ethical AI and Data Use: Experimental Reflections" Dec. 2020, p. 5
- Interview with the software development expert, conducted by Lukas Teutenberg, 2021 (transcript available)
- Interview with the software development expert, conducted by Lukas Teutenberg, 2022 (has yet to be translated)