Annotation Practices in Societally Impactful Machine Learning Applications

What are these automated systems actually trained on?

TU Delft

Introduction

Background and Problem

- ML models follow the principle of **garbage in, garbage out**
- Annotation practices matter for dataset quality as human judgement is variable
- Geiger et al. (2021) evaluated the annotation practices of the datasets introduced in a representative sample of ML papers

Research Question

What are the annotation practices of the datasets used in the highest cited papers in the AAAI Conference on Artificial Intelligence?

Research

Methodology

- Dataset evaluation was done in a **3-step structure**
- **Step 1:** top 25 cited papers from the past 15, 5, and 2 years (75 in total)
- **Step 2:** datasets used for creating an ML model were extracted
- **Step 3:** evaluation of the top 20 datasets according to the *citation sum*
- **Criteria:**
  1) General information
  2) Annotators and annotation process
  3) Items and annotation schema

Evaluation Results

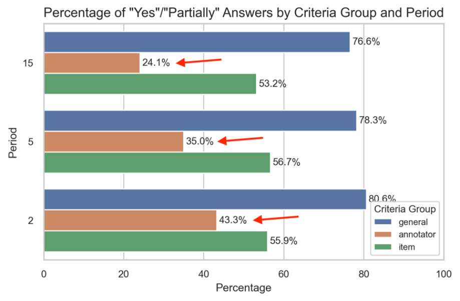
Datasets overall:

1) **General** 33% of the datasets were not human labeled. 45% included a link to the data.

2) **Annotators** Average of 35% positive responses. Highest criterion: 54% positive responses. Lowest criterion: 7.4% positive responses.

3) **Items** Large variation among the criteria.

Datasets per period:



- Significant increase in the documentation as the time period is more recent
- Especially in the category of the annotators and annotation process

Discussion

General	Annotators
• 33% of the datasets were omitted from the annotator criteria • Lack of links: **poor reproducibility**	• **Lack of reproducibility** of the data • Doubt on the **quality** of data • Doubt on the **performance** of the ML models
Items	**Per period**
• Certain properties are more **obvious** or easy to document, others are more **obscure** or hard	• Results provide **hope** for the field • Annotator documentation is still **low** in the recent period (43%)

Conclusion

Main take-aways

- Substantial number of datasets have bad documentation of their annotation process
  → Issues with reproducibility and data quality
  → Issues with ML model performance
- Improvement over the years is a cause for hope, but there is room for improvement

Limitations

- Only 75 ML papers → limits generalizability
- Only top cited papers and datasets from one conference → limits generalizability
- Only one evaluator → limits validity

Contact info

- Author: Damjan Košutić (d.kosutic@student.tudelft.nl)
- Supervisor: Andrew M. Demetriou
- Responsible Professor: dr. Cynthia Liem

References

Geiger, R. S., Cope, D., Ip, J., Lotosh, M., Shah, A., Weng, J., & Tang, R. (11 2021). “Garbage in, garbage out” revisited: What do machine learning application papers report about human-labeled training data? *Quantitative Science Studies*, *2*(3), 795–827. doi:10.1162/qss_a_00144