Research Seminar Data Science Task 2 — Paper 14: Weaker Than You Think

Daan Brugmans (S1080742)

I. Summary

Zhu et al. investigate the usefulness of Weakly Supervised Learning (WSL), a supervised learning method where training is not performed on clean targets, but on noisy targets from varying sources. The authors' main finding is that the usefulness of WSL is very overestimated in practice: while many WSL methods use clean targets during validation and noisy targets during training, the authors show that simply finetuning an existing model on the clean validation targets almost always outperforms WSL methods, and that validating on noisy validation targets simply does not work.

II. Evidence

The authors perform varying experiments, all performed across eight different NLP classification datasets:

- Experiment 1 investigates whether WSL methods need clean validation data to work, or if weakly annotated validation data would suffice. The authors find that clean validation samples are necessary for WSL methods to work as improvements over just the noisy labels themselves.
- Experiment 2 investigates how much clean validation targets WSL methods need in order to work well. The authors find that only a small set of clean validation samples are needed (no more than 30 clean samples per class for improved performance over the weak labels).
- Experiment 3 investigates whether WSL methods can outperform finetuning when there is less clean data. The authors find that WSL methods only outperform finetuning methods when there are less than 10 clean samples per class, otherwise finetuning outperforms WSL.
- Experiment 4 investigates whether WSL methods can be improved when finetuning on clean samples is used as an additional training method, which the authors find does work, although the authors question the practicality of the WSL methods in this process.

III. Strengths

- The authors' claims and stance are very clear and concrete. This clarity makes the aim and relevancy of the paper very clear as well.
- The authors' findings seem very novel: as WSL becomes increasingly popular, Zhu et al. show that that popularity may be unfounded. They show that related work questioning the usefulness of other low-resource machine learning techniques, such as Semi-Supervised Learning and few-shot learning, already exists, but not for WSL, adding to the novelty of the paper.
- Experiments seem to support the authors' claims very well, seem relevant to the research, are done in a logical order, and are described clearly and thoroughly.
- The authors extensively describe the datasets used, hyperparameter spaces explored, techniques used, and results gained in the appendices.

IV. Weaknesses

I could find no major weaknesses of the paper, other than those mentioned by the authors themselves (reliance on high-resource language models, limited number of WSL methods investigated).

V. Evaluation

I would recommend acceptance.

VI. Quality of Writing

The paper is very understandable and an easy read. I particularly appreciate that the main conclusions/takeaways of the chapters are emboldened, emphasizing each chapter's main finding.

The paper contains a small spelling error at the top of page 14232 (4): "The the relative performance gain [...]" should be "The relative performance gain [...]".

VII. Queries

- Despite having shown that WSL methods are less useful in practice than generally believed, at the end of the introduction, the authors ask themselves in which ways WSL methods will fit better to certain tasks than other supervised methods. Which tasks do the reviewers/presenters think WSL methods are more or most useful for, as opposed to other supervised methods?