Comparing Instance Attribution Methods to Baseline k-Nearest Neighbors Method

Evan de Kruif

Responsible Professor: Avishek Anand

Supervisors: Lijun Lyu, Lorenzo Corti

1.INTRODUCTION

Research Question

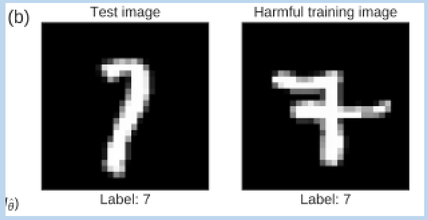
How do different **Instance Attribution** methods compare to baseline **k-Nearest Neighbors** (kNN) method?

Why Instance Attribution?

. Useful method for explaining why an AI made a decision, one way or another.

. Capable of identifying mislabeled or misleading datapoints inside of the training data.

Figure 1: Input, and the influential training datapoint identified as misleading.



[1]

2. METHODOLOGY

Instance Attribution:

. Trace the models predictions back to training data.

. Most influential datapoints from the **FEVER** [2] dataset are shown by the methods.

Criteria for Comparison:

. Representative Vector compared to k-Nearest Neighbors. Shows there is a qualitative difference.

. Take a subset of influential datapoints and, through user study, choose the most relevant.

Evaluation:

. User study will give the more relevant subset, as seen through human language comprehension.

. Representative vectors to compare the semantic similarity.

3. SETUP

Pretrained **ExPred** [3] **Model**: Two part model; first part focuses on providing explanations for predictions, second part focuses on optimizing those predictions for correctness.

Instance Attribution Methods:

. **kNN**: Most naïve/basic method. Very fast.

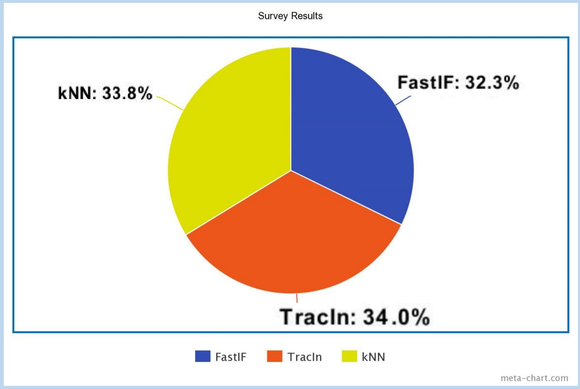
. **FastIF** [4]**:** Deliberate balance of speed and fidelity.

. **TracIN** [5]: Largely focuses on correctness.

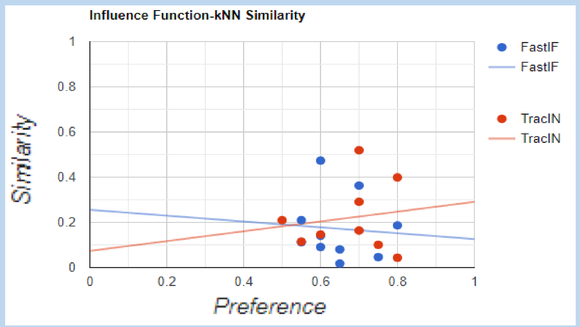
User Study: Participants are asked to choose between kNN and another method to see which result is more relevant to a given query.

4. RESULTS

1. The **user study** results to compare preference between the two methods.



2.The **similarity comparison** between methods.



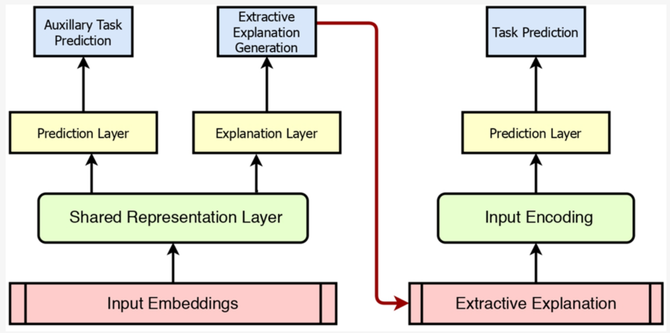


Figure 2: Pipeline of **ExPred**.

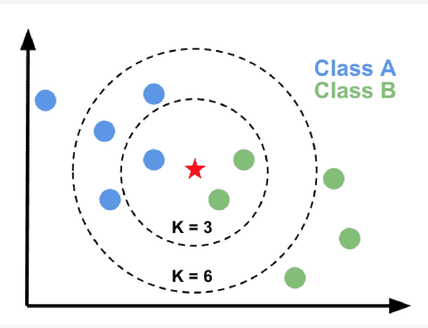


Figure 3: Conceptual representation of **kNN**.

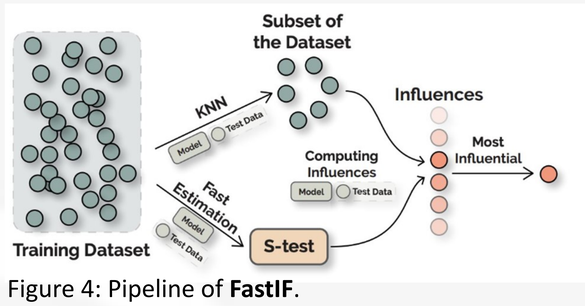


Figure 4: Pipeline of **FastIF**.

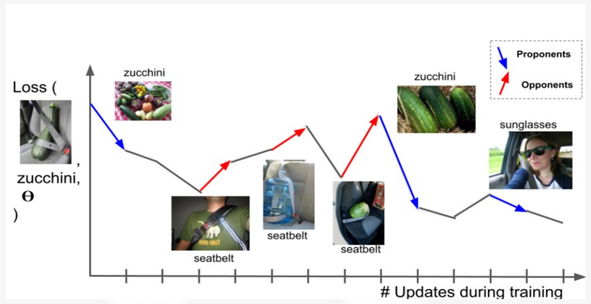


Figure 5: Visualization of **TracIn** tracing influences of

5. CONCLUSIONS

User study results show it is clear that Instance Attribution methods perform better in terms of human preference for understandability.

The **similarity comparison** and best fit lines show that similarity is overall fairly low, and does not influence the preference in a significant way.

Meaning that **Instance Attribution results are preferred to kNN results, and are substantially different**.

6. DISCUSSION

Future Work: Expanding the scope and scale of the research would bring a more conclusive result.

A comparison on how much benefit more complex Instance Attribution methods bring compared to their increased execution cost.

Considerations of This Work: The dataset used had to be trimmed down substantially in order to facilitate the time it takes to run the Instance Attribution methods.

6. References

[1] Koh, P.W., Liang, P.(2017). Understanding Black-box Predictions via Influence Functions. Proceedings of the 34th International Conference on Machine Learning, in Proceedings of Machine Learning Research 70:1885-1894 Available from https://proceedings.mlr.press/v70/koh17a.html.

[2] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a Large-scale Dataset for Fact Extraction and VERification. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 809{819, New Orleans, Louisiana. Association for Computational Linguistics.

[3] Zhang, Zijian; Rudra, Koustav; Anand, Avishek; "Explain and predict, and then predict again" in Proceedings of the 14th ACM International Conference on Web Search and Data Mining, 2021.

[4] Han Guo, Nazneen Rajani, Peter Hase, Mohit Bansal, and Caiming Xiong. 2021. FastIF: Scalable Influence Functions for Efficient Model Interpretation and Debugging. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 10333{10350, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

[5] Pruthi, G., Liu, F., Kale, S., & Sundararajan, M. (2020). Estimating training data influence by tracing gradient descent. *Advances in Neural Information Processing Systems*, *33*, 19920-19930.

Contact Information: e.j.dekruif@student.tudelft.nl