Brainwash: A Data System for Feature Engineering

Michael Anderson*  Michael Anderson*  Victor Bittorfy†  Matthew Burgess*
Michael Cafarella*  Arun Kumary†  Feng Niuy†  Yonjoo Park*
Christopher Réy†  Ce Zhangy†
*University of Michigan, Ann Arbor  University of Wisconsin, Madison

1. Summary
Training systems have to face several difficulties that make their programming quite difficult. In particular, the authors point out the enormous amount of statistics involved (Grunt-Work Statistics), Unknowable specifications which have to be trained against the data itself, and unexpected failures given by features that do not give useful information. The authors suggest a system called Brainwash to address these issues.

2. Positive/Strong Points
2.1) The idea to implement a software that could be helpful in very different scenarios would be very helpful in the development of more sophisticated software.

2.2) Once it is fully implemented, Brainwash would return statistical values, recommend features that might be useful and compatible for the existing code, and evaluate fast new features to see how they perform.

2.3) Brainwash could be used by different developers of a same algorithm or software, and it be helpful to incorporate different pieces of code.

3. Negative/Weak Points
3.1) "The more features, the merrier" is not necessarily true, although they state some examples where it might be the case.

3.2) They presented two case studies (GeoDeepDive and Automan) in which Brainwash was implemented, but they did not present any data on the results or how brainwash improved the development of algorithms as opposed to not using it.

3.3) They are still building the initial prototype based on the specific examples of GeoDeepDive and Automan. It is hard to generalize to all scenarios, and this paper is only a projection of what they are planning to do.

4. Research Questions and Points for Discussion
4.1) Would it be possible to have Brainwash as an open software so that people contribute to it and improve it as specific examples appear?

4.2) On the same line, it could be possible to add crowdsourcing for the explore-extract-evaluate part in Brainwash.
4.3) For using the data from very different scenarios, it is necessary to use the expertise of people with diverse backgrounds, and even so there might be several cases in which it is not possible to generalize. Brainwash seems like a system that would be very useful, but which is still in a very early phase of development.