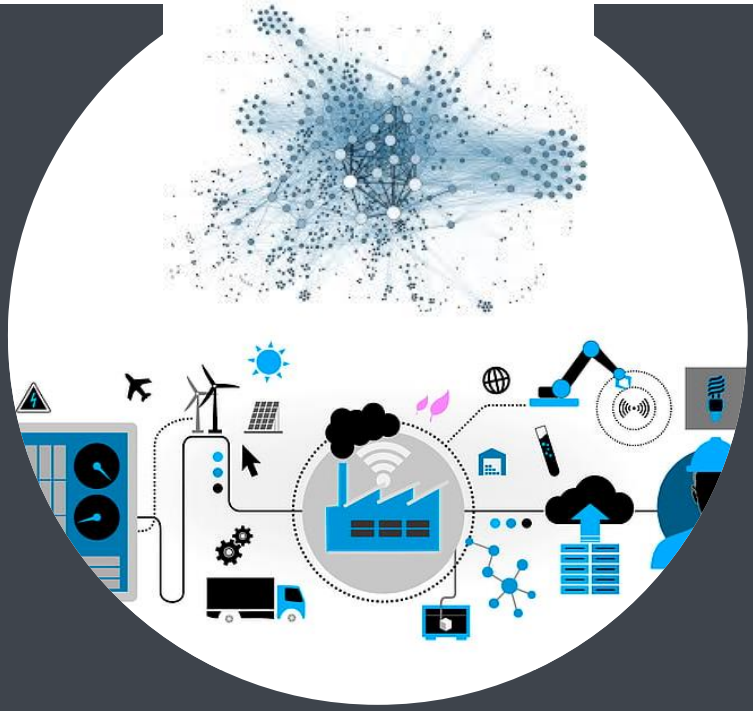
University of Stuttgart

Bayesian Model for Trustworthiness Analysis of Deep Learning Classifiers

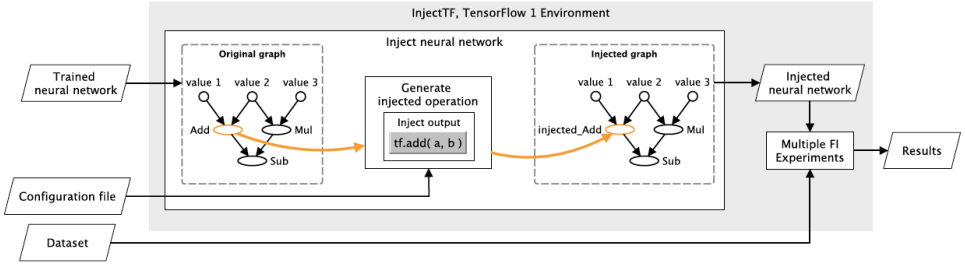


Andrey Morozov*, Emil Valiev, Michael Beyer,
Kai Ding, Lydia Gauerhof, Christoph Schorn

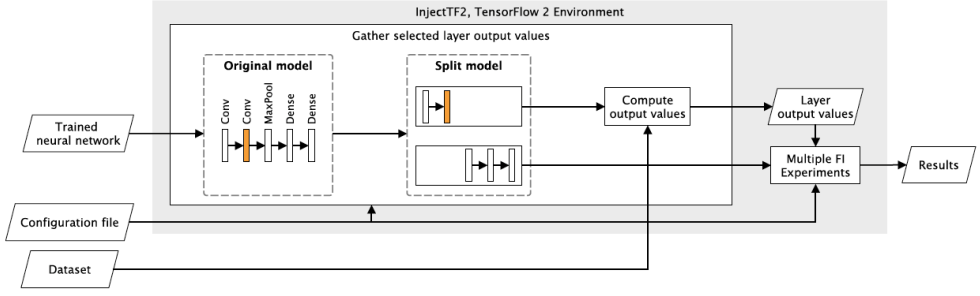
AI Safety Workshop, IJCAI 2020, Online

Impact of Random Hardware Faults on DNNs

- Deep Learning software is prone to random hardware faults e.g. bit flips.
- We developed two fault injection framework for TensorFlow 1 and 2.

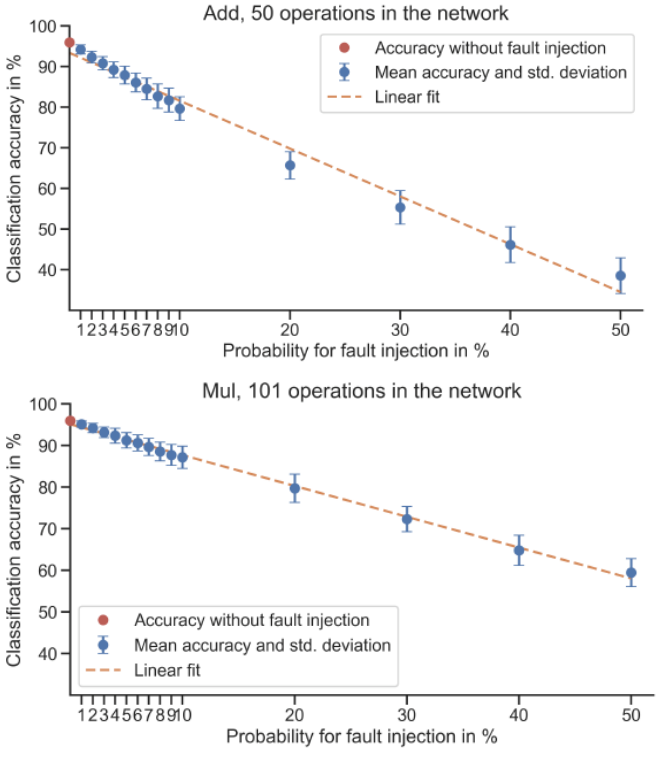


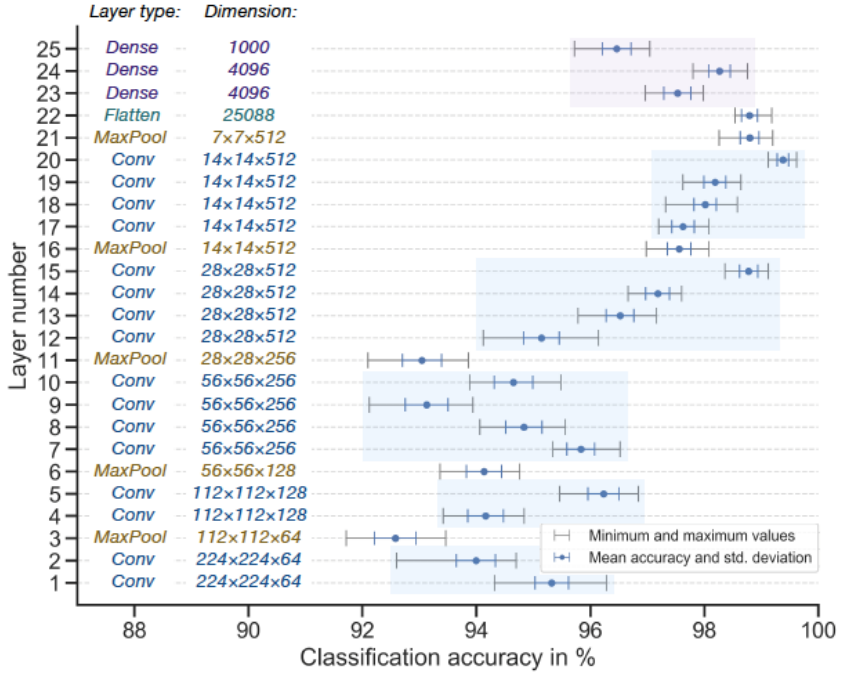
https://github.com/mbsa-tud/InjectTF



https://github.com/mbsa-tud/InjectTF2

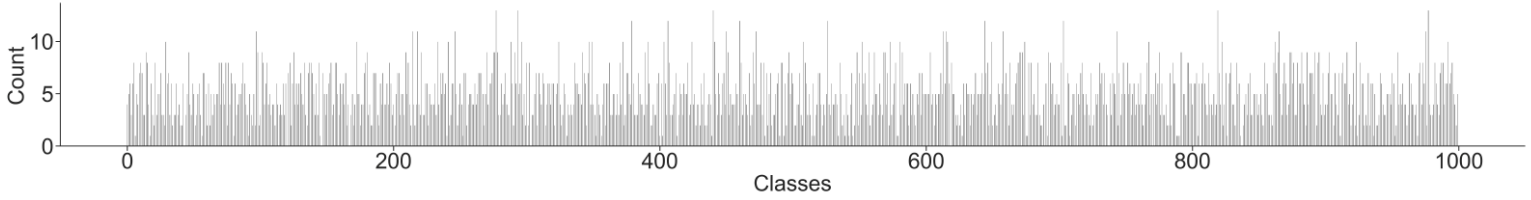
Impact of Random Hardware Faults on DNNs



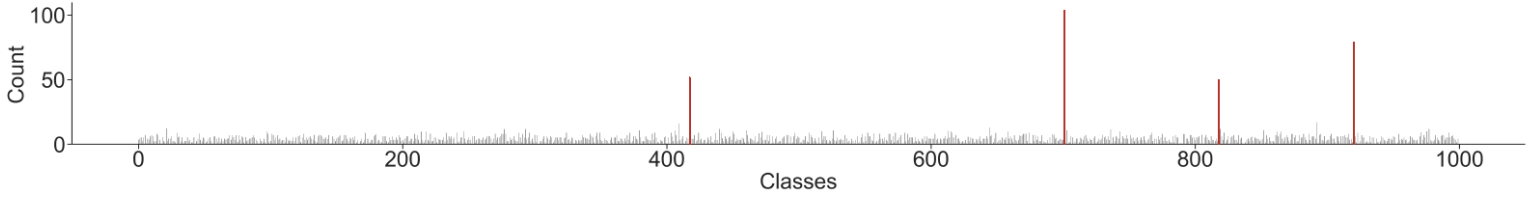


Resulting classification accuracies for the VGG19 CNN.

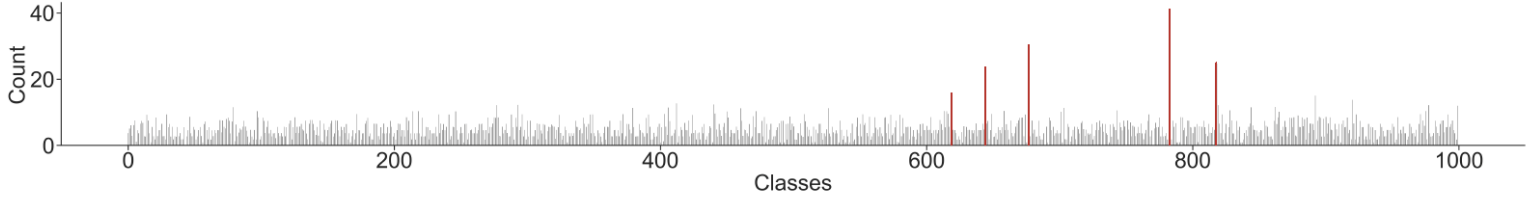
Sink Classes



(a) VGG19: Fault free run.



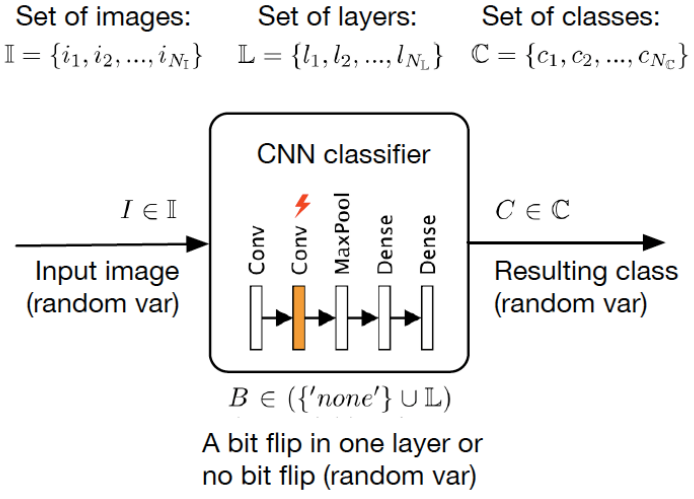
(b) VGG19: Fault injection in Layer 3. The sink classes are highlighted in red.

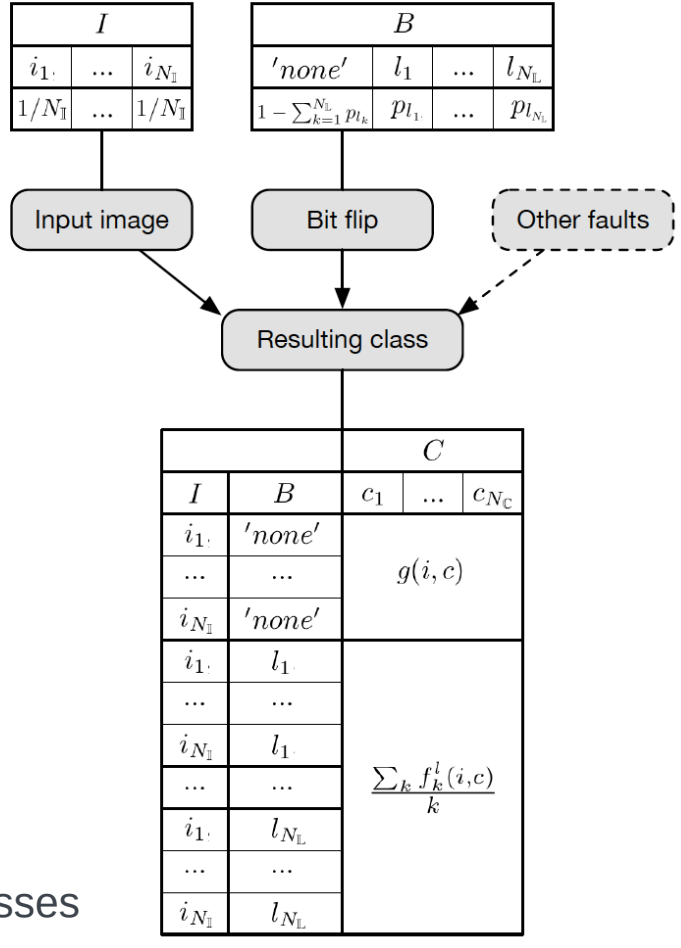


(c) VGG19: Fault injection in Layer 10. The sink classes are highlighted in red.

Results of the fault injection experiments on VGG19 for the ImageNet dataset (1000 classes).

Trustworthiness Bayesian Model

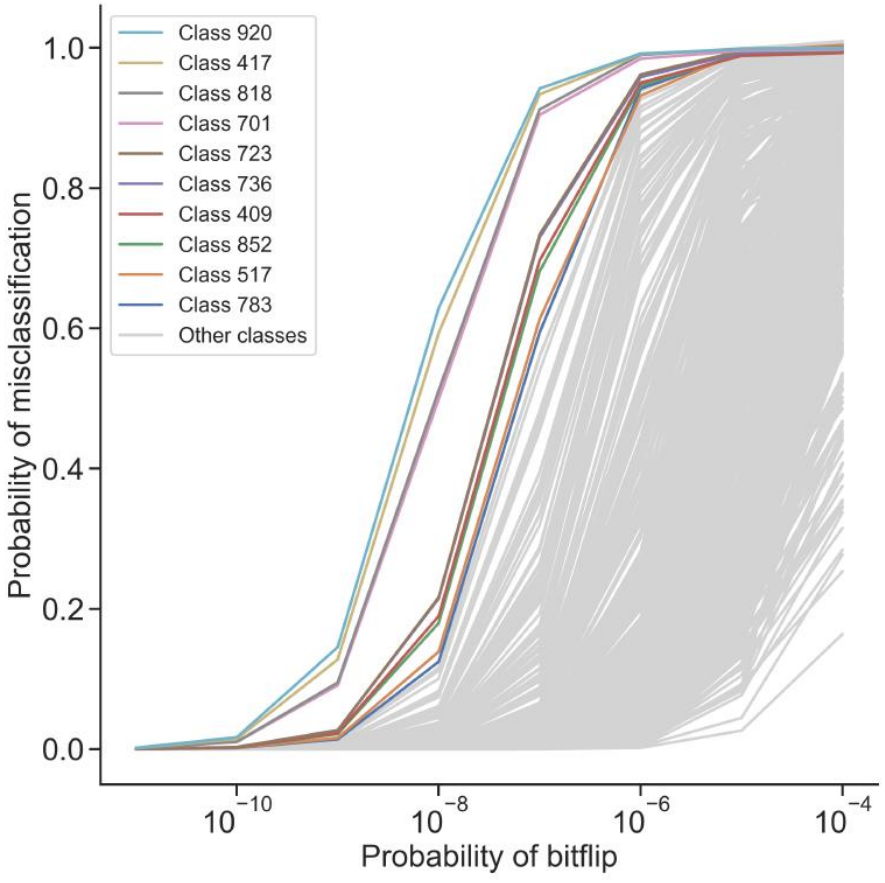




- BN stores the results of fault injection experiments
- Analysis of various reliability-related properties
- Quantification of the trustworthiness for resulting classes

Results and Conclusion

- Two fault injection frameworks
- Experiments on several CNN models
- Observation to distinctive sink classes for each layer
- BN that stores the results of the experiments and support multiple types of probablistic analysis of relaibility related properties



University of Stuttgart
Institute of Industrial Automation and
Software Engineering

Thank you!

Jun.-Prof. Dr.-Ing. Andrey Morozov

e-mail andrey.morozov@ias.uni-stuttgart.de
phone +49 (0) 711 685-67312
www. ias.uni-stuttgart.de/en/institute/team/Morozov/

University of Stuttgart
Institute of Industrial Automation and Software Engineering
Networked Automation System