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

Add, 50 operations in the network

- Accuracy without fault injection
- Mean accuracy and std. deviation
- Linear fit

Classification accuracy in %
Probability for fault injection in %

Mul, 101 operations in the network

- Accuracy without fault injection
- Mean accuracy and std. deviation
- Linear fit

Classification accuracy in %
Probability for fault injection in %

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

### Diagram

- **Set of images:** $I = \{i_1, i_2, ..., i_{N_I}\}$
- **Set of layers:** $\mathbb{L} = \{l_1, l_2, ..., l_{N_L}\}$
- **Set of classes:** $\mathbb{C} = \{c_1, c_2, ..., c_{N_C}\}$

- **Input image** (random var)
- **CNN classifier**
- **Resulting class** (random var)

- $B \in (\{\text{"none"}\} \cup \mathbb{L})$
  - A bit flip in one layer or no bit flip (random var)

- $I \in \mathbb{I}$
  - $C \in \mathbb{C}$

### Table

<table>
<thead>
<tr>
<th>$I$</th>
<th>$B$</th>
<th>$C$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$i_1$</td>
<td>'none'</td>
<td>$c_1$</td>
</tr>
<tr>
<td>$i_{N_I}$</td>
<td>'none'</td>
<td>$c_{N_C}$</td>
</tr>
<tr>
<td>$l_1$</td>
<td>$p_1$</td>
<td>$\sum_k f_k^i(i, c)$</td>
</tr>
<tr>
<td>$l_{N_L}$</td>
<td>$p_{N_L}$</td>
<td></td>
</tr>
</tbody>
</table>

$\sum_k f_k^i(i, c)$
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 probabilistic analysis of reliability related properties
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