Methods & Challenges for the Safety Assurance of DNNs in Computer Vision

www.continental.com

Gesina Schwalbe (gesina.schwalbe@continental-corporation.com)

- Goals:
  - Show that safety for DNNs is a challenge
  - Show how broad the topic is

- About me:
  - 3rd year PhD at Conti Automotive he[a]t AI
  - Supervised by Ute Schmid, Uni Bamberg
  - PhD topic: Safety assurance of DNNs for automotive perception via concept embedding analysis (quantitative introspection)

Agenda

1. Background
2. Challenges: What's new?
3. Methods
4. Conclusion

Background
Why should we care?

Automated driving

Industry 4.0

Medical assistant systems

Background
What is safety?

Def. Safety
means *absence of unreasonable risk* due to
- malfunction (ISO 26262-1, 3.132)
- intended functionality
  (misuse, performance limitation, environment) (ISO/PAS 21448)

[…] according to valid societal moral concepts (ISO 26262-1, 3.176)

Rating safety: Safety Integrity Levels (ISO 26262-3, 6.4.3) derived from

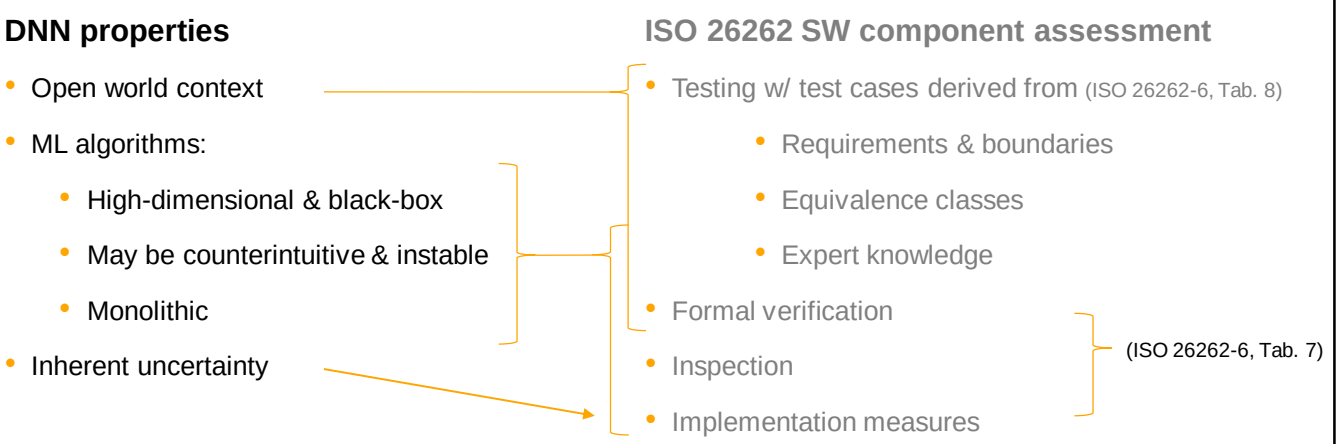
› **Probability, Severity,** Controllability

› What is *safety*?
  › Absense of unreasonable risk for harm (=physical injury or damage to the health of persons)
  › = probability + severity + controllability
› Different causes for safety issues handled by different standards: malfunction, misuse, …

Agenda

1. Background
2. Challenges: What's new?
3. Methods
4. Conclusion

Challenges: What's new?



vs. recommended properties (ISO 26262-6, 8.4.5):
Simplicity, comprehensibility, robustness, suitability for software modification, verifiability

Takeaways:

- ISO 26262 is implementation centric, and does little consider issues of the algorithm (only using system simulation)!
- Replacements for formal verification & inspection needed
- Assistance for coverage needed

Not considered here:

- HW integration,
- integrated validation,
- SOTIF aspects, etc.

Details on the mentioned points:

› Open world ->

  › Hard to formulate (formally verifiable) requirements like coverage

  › Hard to structure input space (what is the manifold of

real images?)

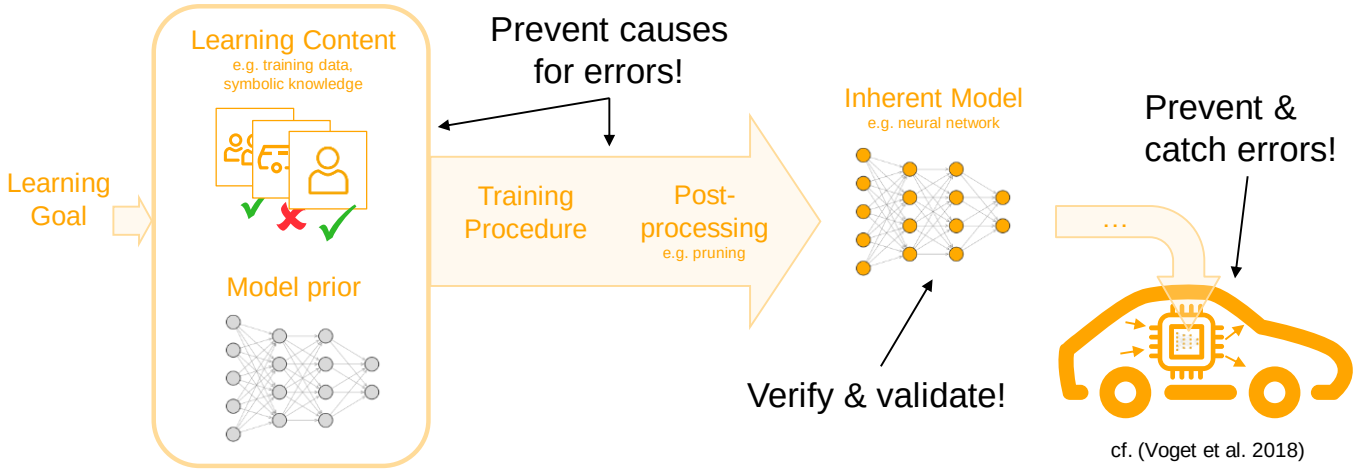
- What is expert knowledge for an open world? Physics? Traffic statistics?

- ML ->
  - Not manually crafted -> little expert knowledge
  - Black-box -> hard to inspect
  - Large & complex -> hard to formally verify
  - Little best-practices available -> hard to estimate influence of implementation measures
- Uncertainty -> increases implementation efforts

Agenda

Agenda

Types of Methods
The ML-based System



cf. (Voget et al. 2018)

- ML system is (MUCH) more than just the DNN!
- Structured way to find safety issues: Go along the life-cycle &
- 3 pillars to ensure safety:
  - Build it right
  - Check it
  - Fail safe integration

Agenda

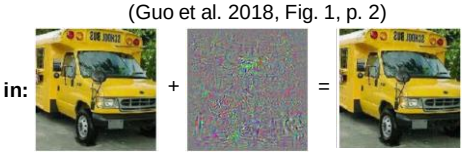
Creation
Training Data Optimization

› Image manipulation
  › Addition of artifacts
  › Domain randomization

› Image generation

› Counterexample generation
(Dreossi et al. 2018)

(Guo et al. 2018, Fig. 1, p. 2)



(Eykholt et al. 2018, Tab. 1)

“speed limit 45”

(Geirhos et al. 2019), Fig. 1

General goals: Make model
› **invariant** wrt. **irrelevant** features
› **focus** on **relevant** features

- Domain randomization = vary parameters unimportant for task
- Standard methods for image artifacts: gaussian noise, kernels, cropping, blurring, overlay, …

Details on the References:
- (Eykholt et al. 2018): They generate real world adversarial examples
- (Geirhos et al. 2019): Experiments on the texture-bias of standard models
- (Dreossi et al. 2018): Select counterexamples from a semantic feature space by singular value decomposition and search
- (Guo et al. 2018): DLFuzz framework for fuzzy DNN testing

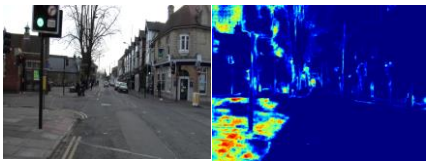
Creation
Architecture and Training Objective

- Explainable intermediate output, *e.g.*
  - Attention heatmaps

(Kim and Canny 2017), Fig. 5

- Soft training constraints, *e.g.*
  - Hierarchical

(Roychowdhury, Diligenti, and Gori 2018)
  - Locality of activations
  - Robustness against perturbations
- Proper uncertainty output, *e.g.* via
  - Ensembling
  - Bayesian DNNs



(Kendall & Gal 2017, Fig. 1, p. 2)

Goals:
- enable **intospection**
- encode more **expert knowledge**

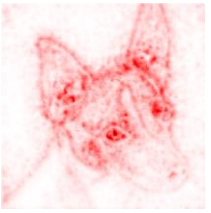
About uncertainty: Normal confidence often badly calibrated; other options:
- Calibrate confidences (e.g. temperature scaling)
- Ensembling → costly
- Modify architecture towards Bayesian net

Details on the References:

- (Kim and Canny 2017): Add multiplicative mask module that learns represents attention heatmap
- (Roychowdhury, Diligenti, and Gori 2018): Add super-class neurons to classification net and enforce the super-class relationship using fuzzy logic to formulate the regularization terms
- (Kendall & Gal 2017): Suggestion of using Monte-Carlo dropout for uncertainty outputs

Offline Verification
Quantitative Explainable AI

- "Attention" heatmap-methods for plausibility checks, *e.g.*
  - White-box (gradients, relevance back-propagation, …)
  - Black-box (occlusion based, perturbation based …)



(Kindermans et al. 2018), Fig. 6

- Knowledge extraction: Disentanglement of internal semantics
  - Similarity of learned concepts
    (Fong and Vedaldi 2018), (Schwalbe and Schels 2020)

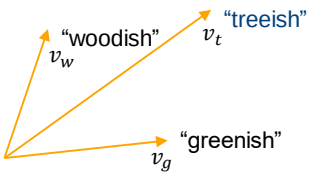


Illustration of (Fong and Vedaldi 2018), Tab. 3

Goals:

- Enable **(post-hoc) inspection** of model

There are many methods that try to assist in corner case detection. This is only useful if

- actual faults can be derived from the corner cases that imply mitigation methods (e.g. corner cases can be used for training data augmentation)
- this can be done online

Details on the References:

- (Kindermans et al. 2018): PatternAttribution: do a backprogation of activations from output to input that fulfills certain axioms; "improved" LRP
- (Fong and Vedaldi 2018): Net2Vec: They found that vector relations in the latent space of classifiers actually can encode semantic relations
- (Schwalbe and Schels 2020): Suggestions on how to use post-hoc concept extraction from DNNs for verification

Offline Verification
Formal Methods

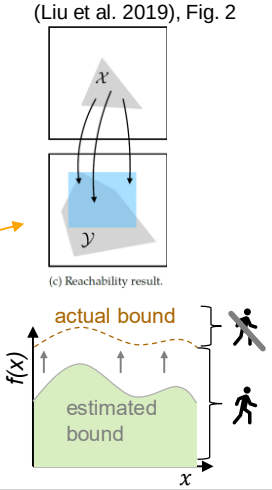
Formal verification

› **Goals:** Find
  › counterexamples
  › validity range
  › reachable set

› **Methods:**

Layer-by-layer reachability / boundary estimation, (constrained) optimization, search, solvers, …

(Liu et al. 2019), Fig. 2



(Formal) Testing

› **Goals:**
  › Semantic coverage e.g. via SDL & sampler
  › Latent space coverage (direct & indirect)

› **Methods:**

Differential,
fuzzy,
concolic, …

Goals:

› Guarantees about **formal properties**

› **High coverage** testing

- Def. “Formal testing refers to testing methods that include formalized and formally verifiable steps”
- Differential testing = try to maximize the output difference of several models
- Fuzzy testing = explorative testing starting from seeds
- Examples of scene description languages (SDL):
  - SCENIC: https://scenic-lang.readthedocs.io/en/latest/
  - OpenSCENARIO: https://www.asam.net/standards/detail/openscenario/

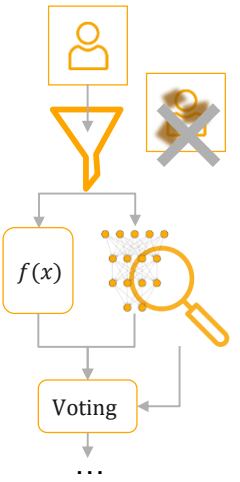
Details on the References:

- (Liu et al. 2019): Very good and extensive review on different types of

formal verification for DNNs

Online Verification: System level measures

- Input filtering
- Redundancy & voting
- Monitoring, *e.g.* using
  - Uncertainty output
  - Temporal consistency
  - Consistency of outputs
    - independent outputs
    - dependent outputs



Goals:

- **Prevent** errors
- **Catch** errors

- Examples for input filtering:
  - Remove inadequate inputs
  - Filter out perturbances / adversarial features
- Examples for consistency checks:
  - Independent outs: compare LiDAR and camera
  - Dependent outs: head without pedestrian?

Agenda

1. Background
2. Challenges: What's new?
3. Methods
4. Conclusion

Conclusion

- Safety of DNNs requires new methods!
- **Categories**:
  - Creation (“build it right”)
  - V&V (“check it right”)
  - System design (“prevent / mitigate failing in op”)
- **Broad spectrum** of methods in development

Safety of ML:
On a good way,
but still challenging

- Note on XAI:
  - XAI means to assist in inspection, trust, etc.
  - Inspection is just ONE method for safety assurance!
  - Quantative methods needed!!

References (I)

- Eykholt, Kevin, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. 2018. “Robust Physical-World Attacks on Deep Learning Visual Classification.” In *Proc. 2018 IEEE Conf. Computer Vision and Pattern Recognition*, 1625–1634. Salt Lake City, UT, USA: IEEE Computer Society. https://doi.org/10.1109/CVPR.2018.00175.
- Dreossi, Tommaso, Shromona Ghosh, Xiangyu Yue, Kurt Keutzer, Alberto L. Sangiovanni-Vincentelli, and Sanjit A. Seshia. 2018. “Counterexample-Guided Data Augmentation.” In *Proc. 27th Int. Joint Conf. Artificial Intelligence*, edited by Jérôme Lang, 2071–2078. Stockholm, Sweden: ijcai.org. https://doi.org/10.24963/ijcai.2018/286.
- Fong, Ruth, and Andrea Vedaldi. 2018. “Net2Vec: Quantifying and Explaining How Concepts Are Encoded by Filters in Deep Neural Networks.” In *Proc. 2018 IEEE Conf. Comput. Vision and Pattern Recognition*, 8730–8738. Salt Lake City, UT, USA: IEEE Computer Society. https://doi.org/10.1109/CVPR.2018.00910.
- Geirhos, Robert, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. 2019. “ImageNet-Trained CNNs Are Biased towards Texture; Increasing Shape Bias Improves Accuracy and Robustness.” In *Proc. 7th Int. Conf. Learning Representations*. New Orleans, LA, USA: OpenReview.net. https://openreview.net/forum?id=Bygh9j09KX.
- Guo, Jianmin, Yu Jiang, Yue Zhao, Quan Chen, and Jiaguang Sun. 2018. “DLFuzz: Differential Fuzzing Testing of Deep Learning Systems.” In *Proc. ACM Joint Meeting on European Software Engineering Conf. and Symp. Foundations of Software Engineering*, 739–43. Lake Buena Vista, FL, USA: ACM. https://doi.org/10.1145/3236024.3264835.
- ISO/TC 22/SC 32. 2018. *ISO 26262-1:2018(En): Road Vehicles — Functional Safety — Part 1: Vocabulary*. 2nd ed. Vol. 1. 11 vols. ISO 26262:2018(En). Vernier, Geneva: International Organization for Standardization. https://www.iso.org/standard/68383.html.

References (II)

- ISO/TC 22/SC 32. 2018. *ISO 26262-3:2018(En): Road Vehicles — Functional Safety — Part 3: Concept Phase*. 2nd ed. Vol. 3. 11 vols. ISO 26262:2018(En). Vernier, Geneva: International Organization for Standardization. https://www.iso.org/standard/68385.html.
- ISO/TC 22/SC 32. 2018. *ISO 26262-6:2018(En): Road Vehicles — Functional Safety — Part 6: Product Development at the Software Level*. 2nd ed. Vol. 6. 11 vols. ISO 26262:2018(En). Vernier, Geneva: International Organization for Standardization. https://www.iso.org/standard/68388.html.
- ISO/TC 22/SC 32/WG 8. 2019. *ISO/PAS 21448:2019(En): Road Vehicles — Safety of the Intended Functionality*. 2019th ed. Vernier, Geneva: International Organization for Standardization. https://www.iso.org/standard/70939.html.
- Kendall, Alex, and Yarin Gal. 2017. "What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?" In *Advances in Neural Information Processing Systems 30*, 5580–90. Long Beach, CA, USA. http://papers.nips.cc/paper/7141-what-uncertainties-do-we-need-in-bayesian-deep-learning-for-computer-vision.
- Kim, Jinkyu, and John F. Canny. 2017. "Interpretable Learning for Self-Driving Cars by Visualizing Causal Attention." In *Proc. 2017 IEEE Int. Conf. Comput. Vision*, 2961–2969. Venice, Italy: IEEE Computer Society. https://doi.org/10.1109/ICCV.2017.320.
- Kindermans, Pieter-Jan, Kristof T. Schütt, Maximilian Alber, Klaus-Robert Müller, Dumitru Erhan, Been Kim, and Sven Dähne. 2018. "Learning How to Explain Neural Networks: PatternNet and PatternAttribution." In *Proc. 6th Int. Conf. on Learning Representations*. Vancouver, Canada. https://openreview.net/forum?id=Hkn7CBaTW.
- Liu, Changliu, Tomer Arnon, Christopher Lazarus, Clark W. Barrett, and Mykel J. Kochenderfer. 2019. "Algorithms for Verifying Deep Neural Networks." *CoRR* abs/1903.06758 (March). http://arxiv.org/abs/1903.06758.

References (III)

- Roychowdhury, Soumali, Michelangelo Diligenti, and Marco Gori. 2018. “Image Classification Using Deep Learning and Prior Knowledge.” In *Workshops of the 32nd AAAI Conf. Artificial Intelligence*, WS-18:336–343. AAAI Workshops. New Orleans, Louisiana, USA: AAAI Press. https://aaai.org/ocs/index.php/WS/AAAIW18/paper/view/16575.

- Sämann, Timo, Peter Schlicht, and Fabian Hüger. 2020. “Strategy to Increase the Safety of a DNN-Based Perception for HAD Systems.” *CoRR* abs/2002.08935. https://arxiv.org/abs/2002.08935.

- Schwalbe, Gesina, Bernhard Knie, Timo Sämann, Timo Dobberphul, Lydia Gauerhof, Shervin Raafatnia, and Vittorio Rocco. 2020. “Structuring the Safety Argumentation for Deep Neural Network Based Perception in Automotive Applications.” In *Computer Safety, Reliability, and Security. SAFECOMP 2020 Workshops*, edited by António Casimiro, Frank Ortmeier, Erwin Schoitsch, Friedemann Bitsch, and Pedro Ferreira, 383–394. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-55583-2_29.

- Schwalbe, Gesina, and Martin Schels. 2020. “Concept Enforcement and Modularization as Methods for the ISO 26262 Safety Argumentation of Neural Networks.” In *Proc. 10th European Congress Embedded Real Time Software and Systems*. Toulouse, France. https://hal.archives-ouvertes.fr/hal-02442796.

- Schwalbe, Gesina, and Martin Schels. 2020a. “A Survey on Methods for the Safety Assurance of Machine Learning Based Systems.” In *Proc. 10th European Congress Embedded Real Time Software and Systems*. Toulouse, France. https://hal.archives-ouvertes.fr/hal-02442819.

- Voget, Stefan, Alexander Rudolph, and Jürgen Mottok. 2018. “A Consistent Safety Case Argumentation for Artificial Intelligence in Safety Related Automotive Systems.” In *Proc. 9th European Congress Embedded Real Time Systems*. Toulouse, France. https://hal.archives-ouvertes.fr/hal-02156048.

Thanks!
Questions?

Contact: Gesina.Schwalbe@continental-corporation.com