Continental
The Future in Motion

Concept Enforcement and Modularization

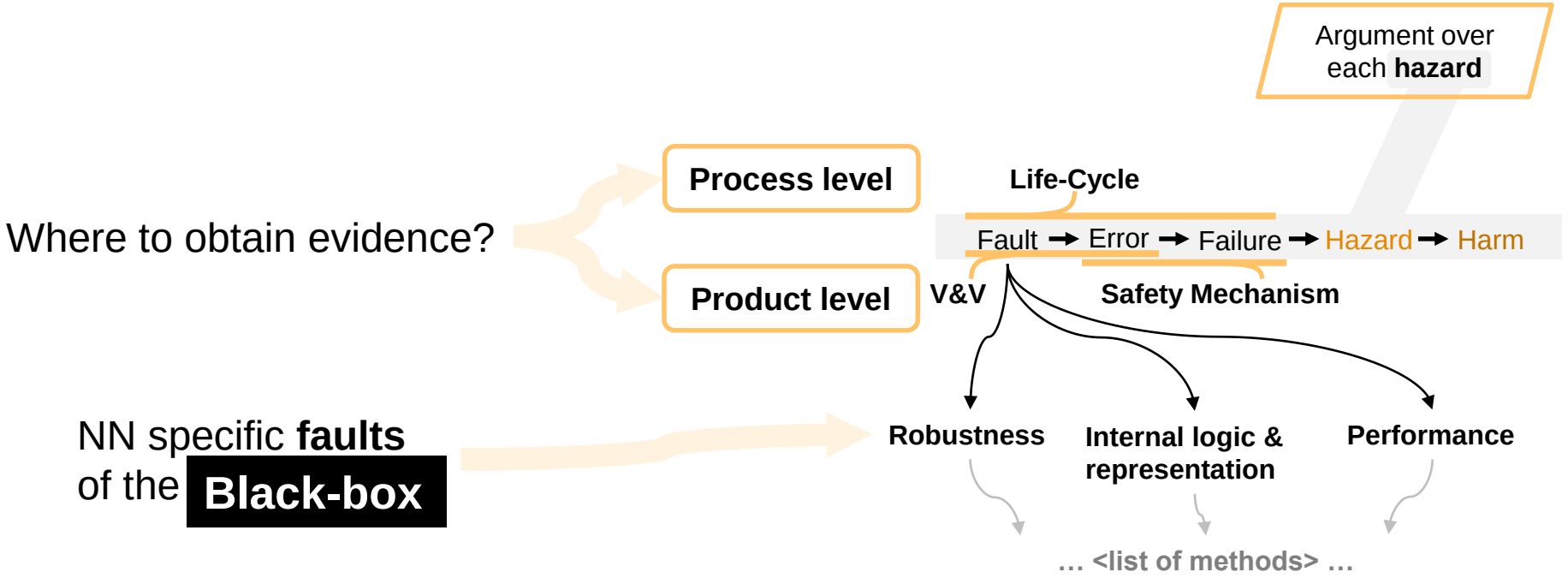
Methods for the ISO 26262 Safety Argumentation of Neural Networks

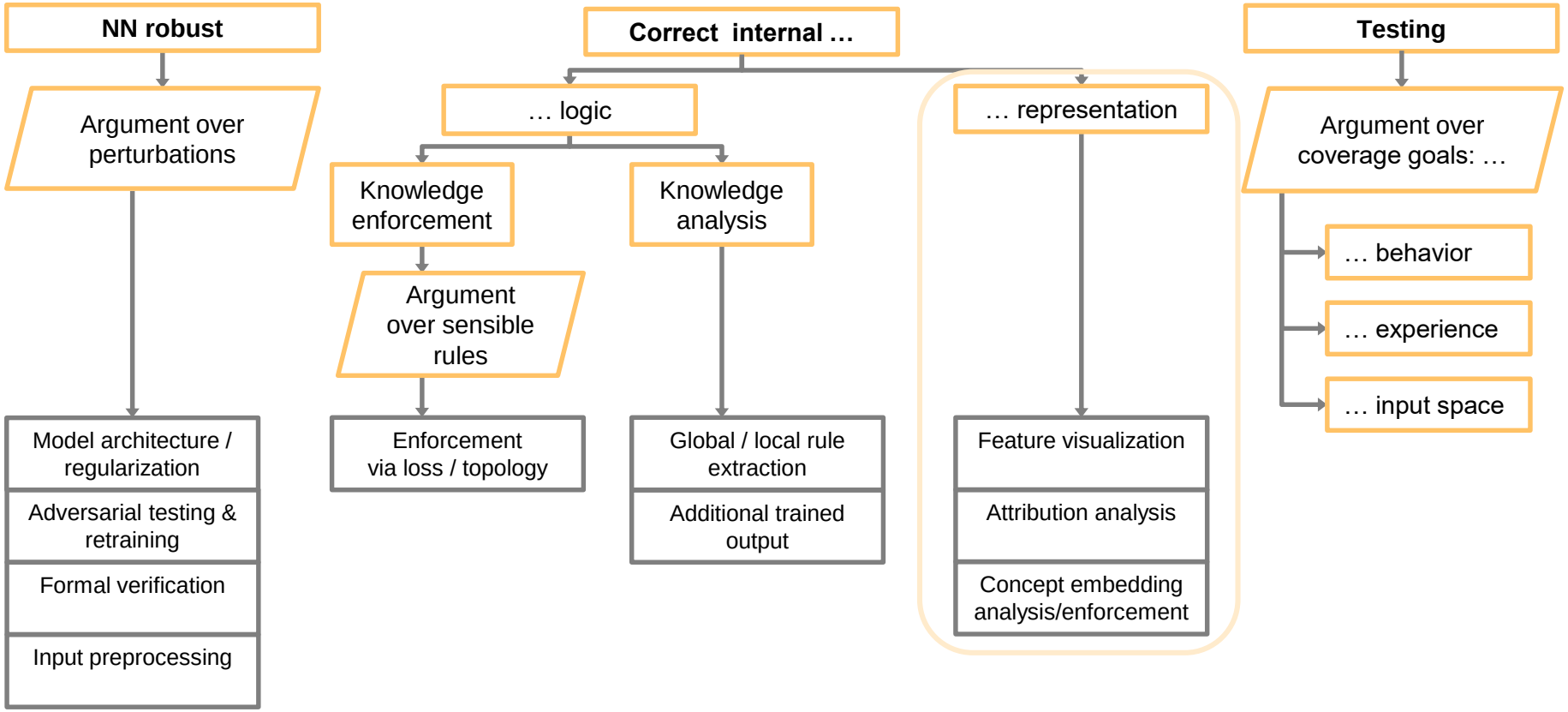
www.continental-corporation.com

Agenda

1. A Safety Argument Structure
2. Concept Embeddings
3. Application Proposals

A Safety Argument Structure
for Neural Network (NN) based systems




NN robust
Argument over perturbations
Model architecture / regularization
Adversarial testing & retraining
Formal verification
Input preprocessing
Correct internal …
… logic
Knowledge enforcement
Argument over sensible rules
Enforcement via loss / topology
Knowledge analysis
Global / local rule extraction
Additional trained output
… representation
Feature visualization
Attribution analysis
Concept embedding analysis/enforcement
Testing
Argument over coverage goals: …
… behavior
… experience
… input space

Examples

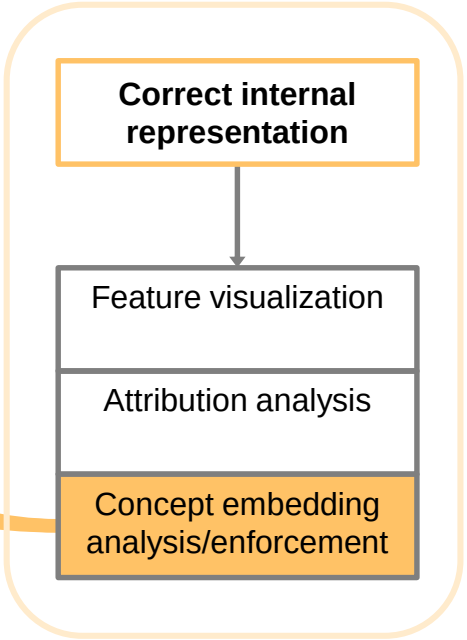
A Safety Argument Structure
What is needed?

Quantitative …

1. … **Analysis** methods:
   Are needed concepts used?
2. … **Measures**:
   Enforce usage of needed concepts!

and

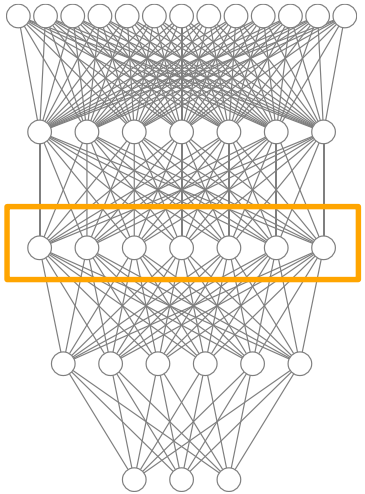
3. a strategy to **mitigate** the
   Black-box

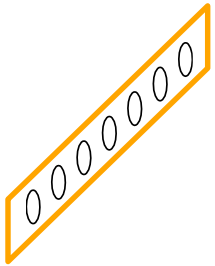


Agenda

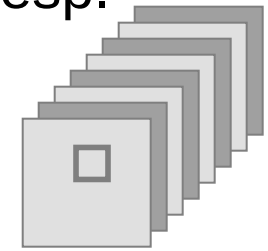
Concept Embeddings

Given





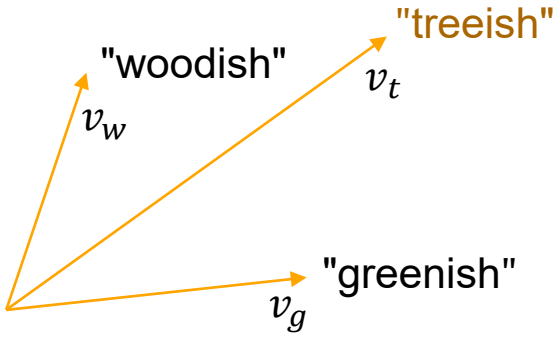
resp.



behaves like

(Fong and Vedaldi 2018)

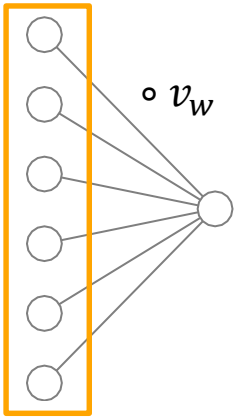
a vector space of semantic concepts



Concept Embeddings
Concept Embedding Analysis

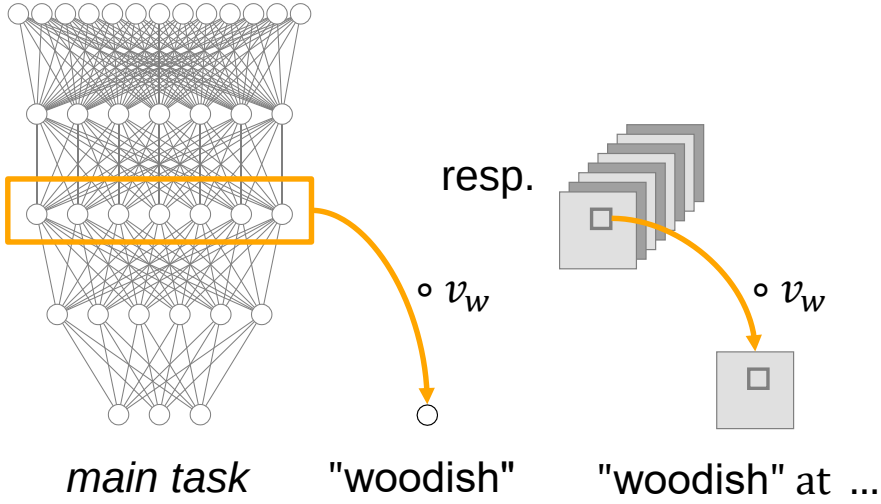
1. Quantitative analysis

How to **use** semantic vectors?



Wood predicted if "woodish" $> t_w$

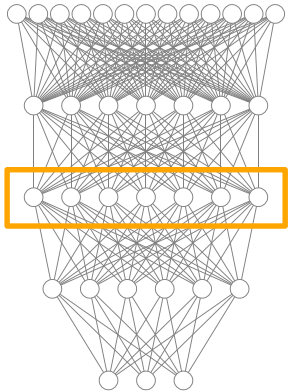
How to **obtain** / **evaluate** semantic vectors?



Experiment Setup

$x =$ 

GTSR
(Stallkamp et al. 2011)



digit-sized window in 3rd layer

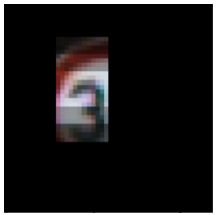
$\circ\, v_w$

w: Weights of 3x4 Conv w/ sigmoid activation

4 ReLU-Conv layers

15 class classification

"contains digit 3?"



$= M_{pred}(x)$

Evaluation:

$$set\ IoU(x) = \frac{\sum_x M_{true}(x) \cap M_{pred}(x)}{\sum_x M_{true}(x) \cup M_{pred}(x)}$$

Concept Embeddings
Findings

IoU encoded

- Receptive field size matters!
  → Predict concept centers
- Tried modifications:
  - (**iou**) Continuous ground truth (g.t.)
  - (**w**) Balance loss
  - Pretrain
  - Different losses
    (**si**, **siou**, **bc**, **mse**)

bc:

Loss	pretrained yes	pretrained no
si	0.313	0.016
si-w	0.308	0.138
si-iou	0.305	0.011
si-w-iou	0.386	**0.200**
siou	0.264	**0.325**
siou	0.047[a]	0.093[a]
bc	**0.473**	0.094
bc-w	–	–
bc-iou	**0.421**	0.050
bc-w-iou	–	**0.198**
mse	**0.423**	0.025
mse-w	0.223	0.099

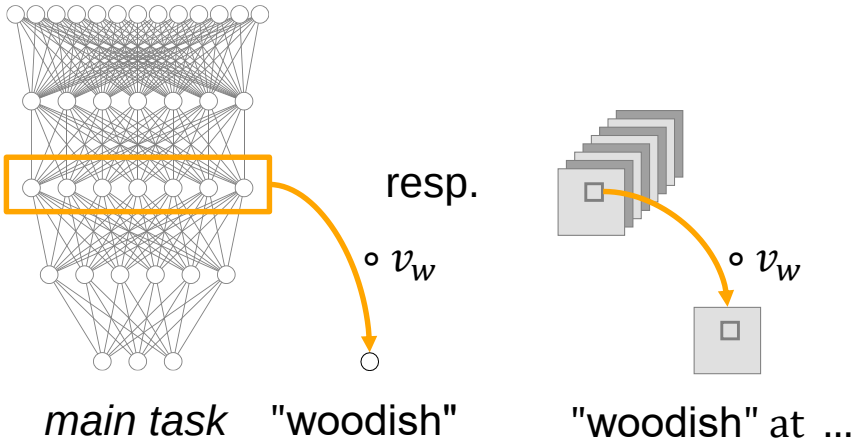
[a] pixels binarized, not bloated

Agenda

Application Proposals
Concept Enforcement

2. Quantitative enforcement

Use concept output as additional objective!



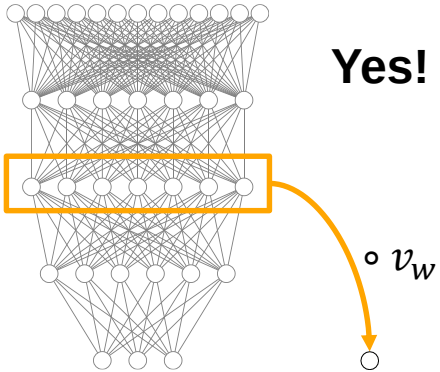
How to use this as safety measure?

1. Train NN
2. Choose concepts & relations
   (ontology, experience)
3. Formulate rules
4. Verify concepts & rules
   (using concept embedding analysis & solvers)
5. Enforce concepts & rules
   (e.g. via loss)

Application Proposals
Modularization

3. Black-box mitigation

Can a meaningful vector basis be found?



Yes! (Losch, Fritz, and Schiele 2019)

How to achieve this?

1. Identify/enforce concepts
2. Reduce output space to that (sub-)vector space
   (pruning, projection, …)
3. Split NN &
   retrain bottom /
   apply new bottom

Summary
What is needed:

Quantitative …

1. … **Analysis** methods:
   Are needed concepts used?

   Concept embedding **analysis**

2. … **Measures**:
   Enforce usage of needed concepts!

   Concept embedding **enforcement**

and

3. a strategy to **mitigate** the **Black-box**

   Modularization

Thanks for listening!

Contact: Gesina.Schwalbe@continental-corporation.com

References

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.

Losch, Max, Mario Fritz, and Bernt Schiele. 2019. “Interpretability beyond Classification Output: Semantic Bottleneck Networks.” In *Proc. 3rd ACM Computer Science in Cars Symp. Extended Abstracts*. Kaiserslautern, Germany. https://arxiv.org/pdf/1907.10882.pdf.

Stallkamp, J., M. Schlipsing, J. Salmen, and C. Igel. 2011. “The German Traffic Sign Recognition Benchmark: A Multi-Class Classification Competition.” In *Proc. 2011 Int. Joint Conf. Neural Networks*, 1453–1460. Y. San Jose, California, USA: IEEE. https://doi.org/10.1109/IJCNN.2011.6033395.