Fraunhofer IKS

Logically Sound Arguments for the Effectiveness of ML Safety Measures

Dr. Chih-Hong Cheng

(joint work with Tobias Schuster & Simon Burton)

SAFETY ARGUMENTATION FOR AI/ML

Semi-formal (structural argument)

- Create assertions (contexts, assumptions, guarantees) using natural language
- May use basic logical operators such as AND, OR, NOT, IMPLY to connect the assertions
- Perform part of the reasoning (e.g., consistency checking) with human guidance

Formal and automated

- Select a rich type of logic (predicates, 1st order or higher order quantifiers, temporal operators, probabilistic variations) to translate from every claim to a logical formula.
- Perform reasoning purely based on logical deduction.

This talk is about an attempt to move towards the formal argumentation

EXAMPLE: SEMI-FORMAL MODELLING FOR EVIDENCE CONSTRUCTION

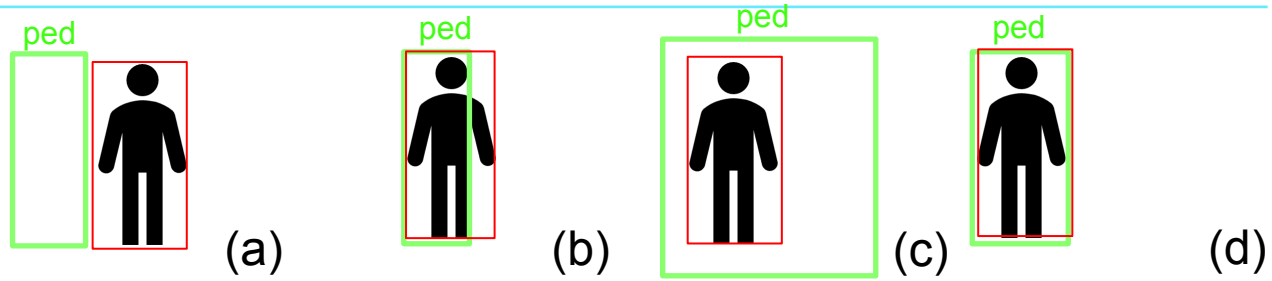
Disclaimer: This is an **overly simplified** example, although the post-processing algorithm presented here is a novel one. The evidence construction will be shown at the end of the presentation

Our goal is to make sure that the prediction of the DNN does not lead to unsafe situations.

NOT_SAFE	**A pedestrian within the range, not being occluded, is not detected properly by the vision-based DNN detector, thereby leading to a collision**

Issue: what is the definition of detected properly?

EXAMPLE: SEMI-FORMAL MODELLING FOR EVIDENCE CONSTRUCTION



Performance by classical "metric" e.g., IoU	very bad	good	bad	very good
Safety metric by "AD collision"	unsafe	unsafe	safe	safe

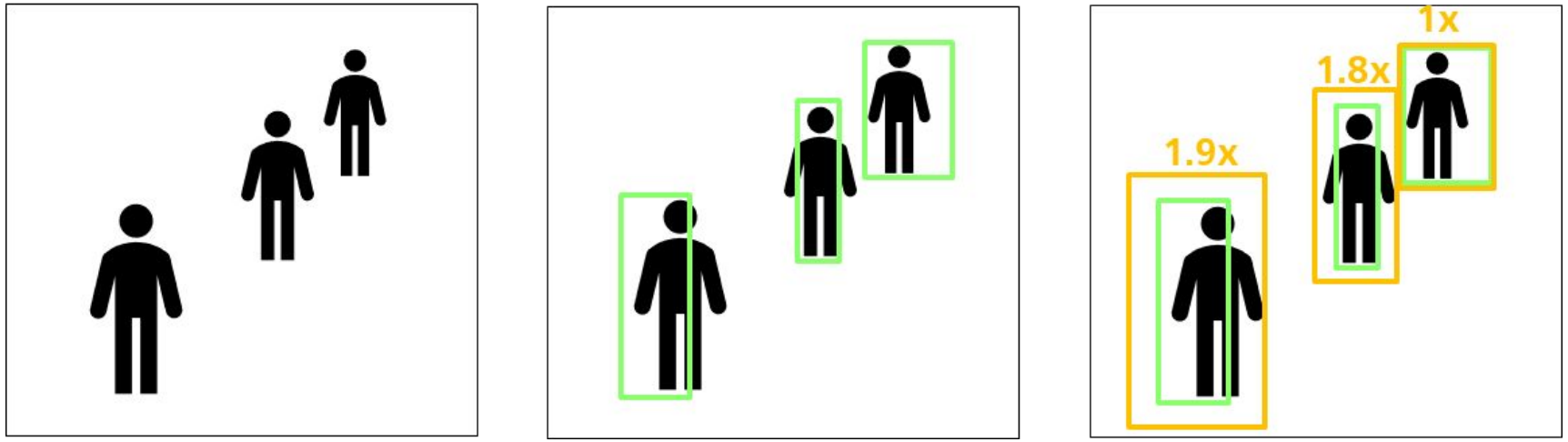
DETECTED_PROPE The bounding box prediction always strictly contains the label of the pedestrian

/* Here the word "always" is used in a highly simplified setup for illustrative purposes. It may be changed to more realistic phrases such as "under partially occluded person within X meters range, with a success rate of 99.9999%", or under some temporal aspects such as one over ten frames */

A CONSERVATIVE POST-PROCESSING ALGORITHM

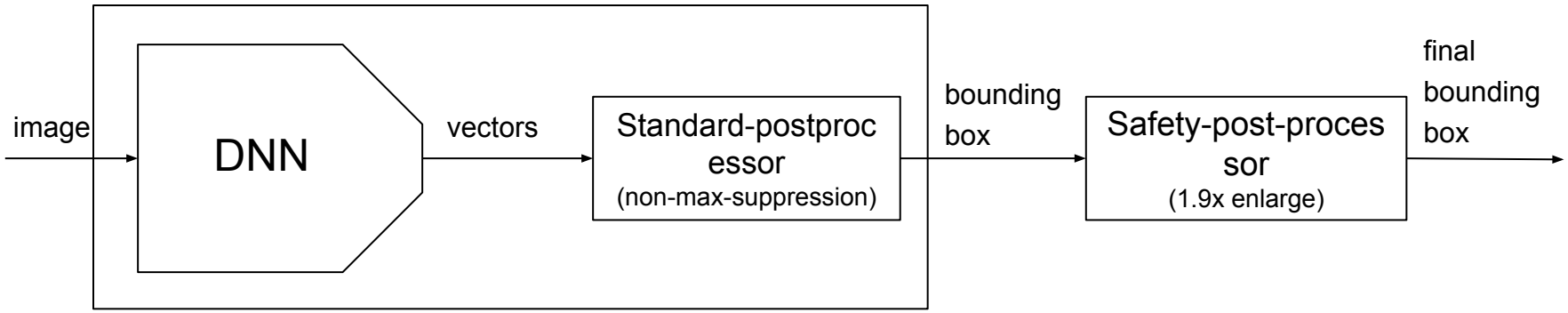
After the training of DNN is completed,

1. Run prediction on all training data
2. Derive for each bounding box, the expansion ratio to fully cover the pedestrian label
3. Store the max ratio (e.g., 1.9x)



In operation: whenever a bounding box is generated, the post processor *always enlarge it by 1.9x as the final prediction*

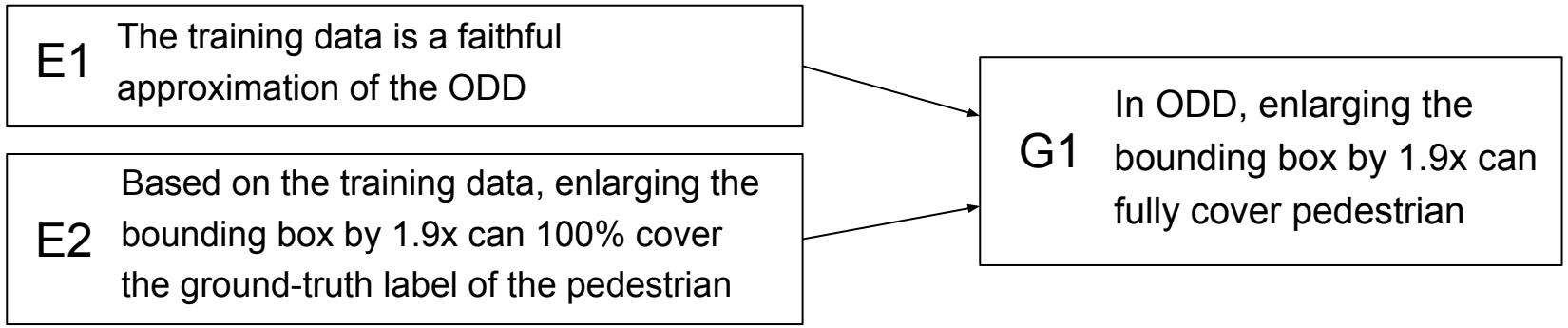
DNN + POST-PROCESSING + SAFE-POST-PROCESSING



Semi-formal modelling framework for evidence construction

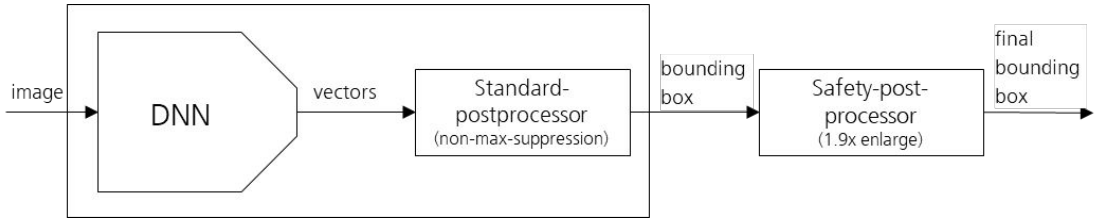
(Graphical)

(Textural, with logic connectors)



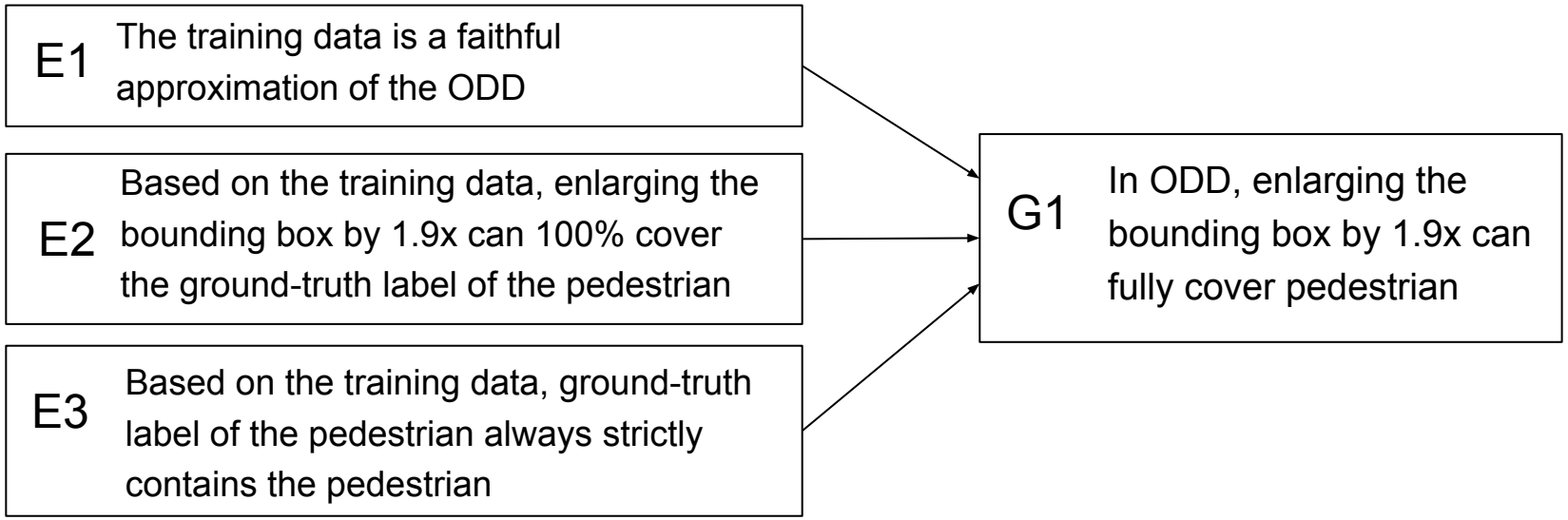
$$(E1 \wedge E2) \rightarrow G1$$

LIMITATIONS OF SEMI-FORMAL METHODS



One needs formality to perform proper deduction clear-out missing statements to allow the proof to go through

Semi-formal modelling framework for evidence construction



FROM INFORMAL TO FORMAL, AND THE LOGICAL DEDUCTION

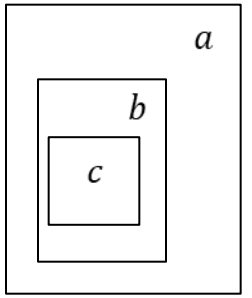
E1 **(given)**	**The training data is a faithful approximation of the ODD**
E2 (given)	Based on the training data, enlarging the bounding box by 1.9x can 100% cover the ground-truth label of the pedestrian
E3 (given)	Based on the training data, ground-truth label of the pedestrian always strictly contains the pedestrian
G1 (to be proved)	In ODD, enlarging the bounding box by 1.9x can fully cover pedestrian

The goal is to **mechanically prove (use PVS)** that when E1, E2, E3 holds, then G1 also holds

To understand the applicability, we create a very strong 2nd order statement on the "faithful approximation", but one can surely relax it to something realistic

FROM INFORMAL TO FORMAL, AND THE LOGICAL DEDUCTION

E1 (given)	The training data is a faithful approximation of the ODD	$\forall Behavior, f_1, f_2 : (\forall\, d_{train}, : Behavior(f_1(DNN(d_{train})), f_2(d_{train}))$ $\rightarrow \forall\, d_{ODD}: Behavior(f_1(DNN(d_{ODD})), f_2(d_{ODD})))$
E2 (given)	Based on the training data, enlarging the bounding box by 1.9x can 100% cover the ground-truth label of the pedestrian	$\forall\, d_{train}: Cover(Enlarge_{1.9}(DNN(d_{train})), label(d_{train}))$
E3 (given)	Based on the training data, ground-truth label of the pedestrian always strictly contains the pedestrian	$\forall\, d_{train}: Cover(label(d_{train}), ground\_truth(d_{train}))$
G1 (to be proved)	In ODD, enlarging the bounding box by 1.9x can safely cover pedestrian	$\forall\, d_{ODD}: Cover\big(Enlarge_{1.9}\big(DNN(d_{ODD})\big), ground\_truth(d_{ODD})\big)$



Impossible to make the proof: One actually needs an additional rule on the **transitivity of area covering**

E4 (fact)	**If A covers B, and B covers C, then A covers C**

FROM INFORMAL TO FORMAL, AND THE LOGICAL DEDUCTION

E1 **(given)**	**The training data is a faithful approximation of the ODD**
E2 (given)	Based on the training data, enlarging the bounding box by 1.9x can 100% cover the ground-truth label of the pedestrian
E3 (given)	Based on the training data, ground-truth label of the pedestrian always strictly contains the pedestrian
E4 (fact)	If A covers B, and B covers C, then A covers C
G1 (to be proved)	In ODD, enlarging the bounding box by 1.9x can fully cover pedestrian

From E2, E3, using E4 (transitivity), we derive E5

From E5, using E1 while we derive G1 (QED)

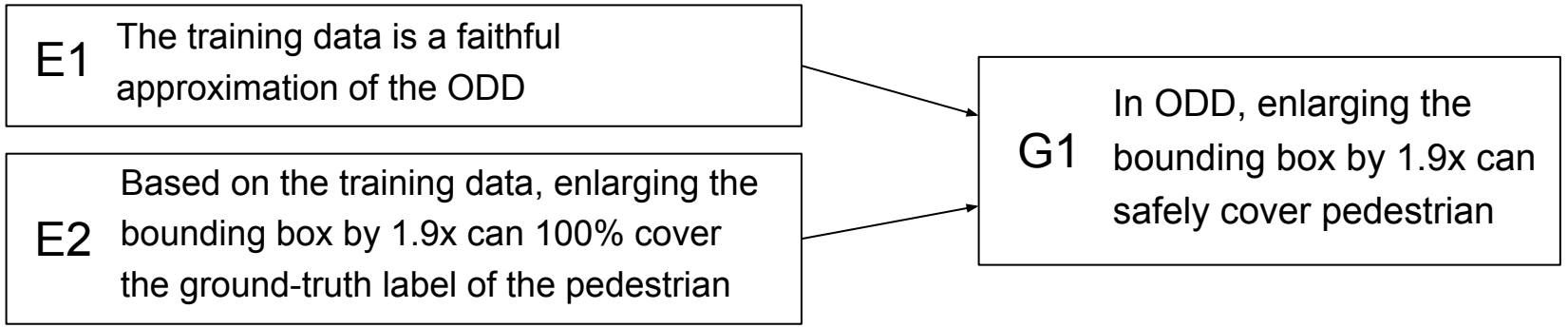
* replacing $Behavior$ with $Cover$:
* replacing f1 with $Enlarge_{1.9}$
* replacing f2 with $ground\_truth$

$$\forall\, d_{train} : Cover(Enlarge_{1.9}(DNN(d_{train})), ground\_truth(d_{train}))$$

IMPLICATION IN QUANTITATIVE REASONING

Semi-formal modelling framework for evidence construction

(Graphical)



(Textural, with logic connectors)

$$(E1 \wedge E2) \rightarrow G1$$

Even when one is 100% sure on E1 and 100% sure on E2, the goal is nothing like 100% sure □ standing on an unsound basis is very risky

CONCLUSION

- Engineering safety-critical AI should be based on a scientifically grounded approach
- Throughout the study, we were able to
  - design new safety-aware performance metrics that can be directly connected to safety goals,
  - suggest a function modification (safety-aware post-processor) as a direct counter-measure over the safety metric, and finally,
  - uncover hidden assumptions / requirements that are missing in the semiformal argumentation
- Noted the risk of moving from qualitative argumentation towards quantitative argumentation without a sound logical foundation
- Partially applied on examples within industrial automation not easy but have surprising effects

Kontakt

Dr Chih-Hong Cheng
Department Head – Safety Assurance for AI
chih-hong.cheng@iks.fraunhofer.de

Fraunhofer IKS
www.fraunhofer.de