DENSO
Crafting the Core

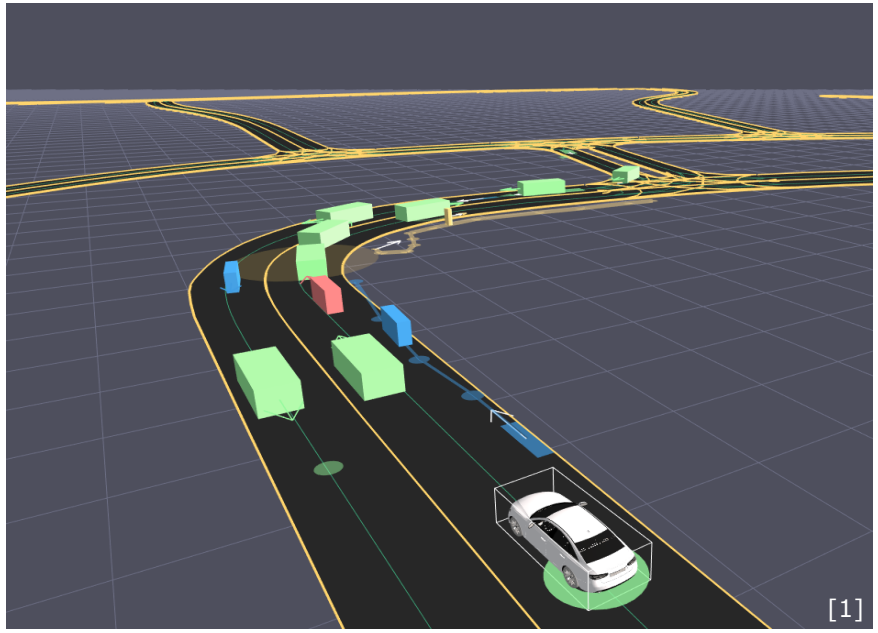
Are Transformers More Robust?
Towards Exact Robustness Verification for Transformers

SafeComp, 20.09.2023

Brian Hsuan-Cheng Liao
Systems Engineering R&D
Corporate R&D
DENSO AUTOMOTIVE Deutschland GmbH (DNDE)

In collaboration with
Dr. Chih-Hong Cheng (Fraunhofer IKS)
Dr. Hasan Esen (DNDE)
Prof. Alois Knoll (Technical University of Munich)

Neural Networks Are Everywhere …

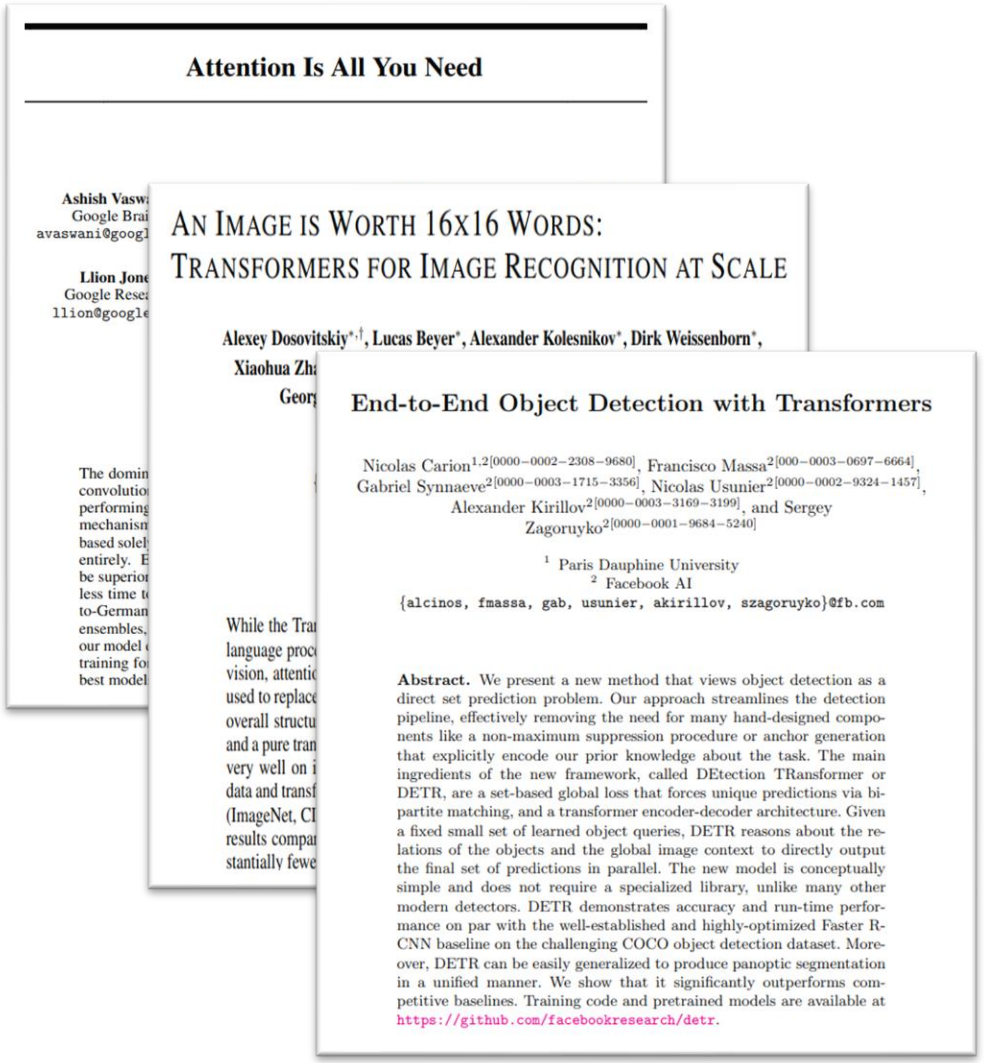


Autonomous Driving
Aircraft Autopiloting

Medical Diagnosis
Surgical Robots

[1] Source: Bloomberg.
[2] Source: CeramTec.

Especially Transformers …



Attention Is All You Need

Ashish Vaswani
Google Brain
avaswani@google.com

Llion Jones
Google Research
llion@google.com

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models

AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

Alexey Dosovitskiy*,†, Lucas Beyer*, Alexander Kolesnikov*, Dirk Weissenborn*, Xiaohua Zhai*, Georg

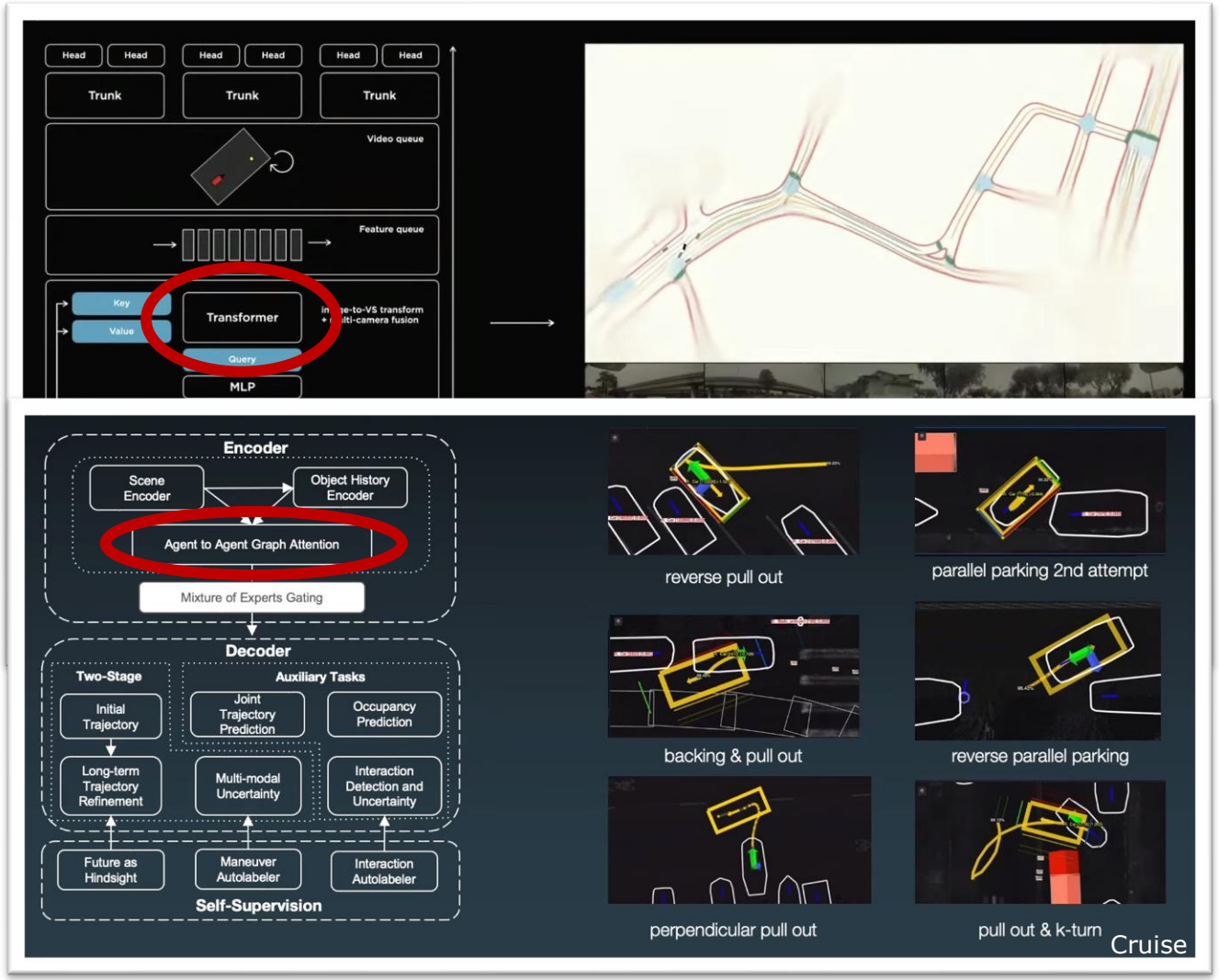
While the Transformer architecture has become the de-facto standard for natural language processing tasks, its applications to computer vision remain limited. In vision, attention is either applied in conjunction with convolutional networks, or used to replace certain components of convolutional networks while keeping their overall structure in place. We show that this reliance on CNNs is not necessary and a pure transformer applied directly to sequences of image patches can perform very well on image classification tasks. When pre-trained on large amounts of data and transferred to multiple mid-sized or small image recognition benchmarks (ImageNet, CIFAR-100, VTAB, etc.), Vision Transformer (ViT) attains excellent results compared to state-of-the-art convolutional networks while requiring substantially fewer

End-to-End Object Detection with Transformers

Nicolas Carion[1,2][0000−0002−2308−9680], Francisco Massa[2][000−0003−0697−6664], Gabriel Synnaeve[2][0000−0003−1715−3356], Nicolas Usunier[2][0000−0002−9324−1457], Alexander Kirillov[2][0000−0003−3169−3199], and Sergey Zagoruyko[2][0000−0001−9684−5240]

[1] Paris Dauphine University
[2] Facebook AI
{alcinos, fmassa, gab, usunier, akirillov, szagoruyko}@fb.com

Abstract. We present a new method that views object detection as a direct set prediction problem. Our approach streamlines the detection pipeline, effectively removing the need for many hand-designed components like a non-maximum suppression procedure or anchor generation that explicitly encode our prior knowledge about the task. The main ingredients of the new framework, called DEtection TRansformer or DETR, are a set-based global loss that forces unique predictions via bipartite matching, and a transformer encoder-decoder architecture. Given a fixed small set of learned object queries, DETR reasons about the relations of the objects and the global image context to directly output the final set of predictions in parallel. The new model is conceptually simple and does not require a specialized library, unlike many other modern detectors. DETR demonstrates accuracy and run-time performance on par with the well-established and highly-optimized Faster R-CNN baseline on the challenging COCO object detection dataset. Moreover, DETR can be easily generalized to produce panoptic segmentation in a unified manner. We show that it significantly outperforms competitive baselines. Training code and pretrained models are available at https://github.com/facebookresearch/detr.



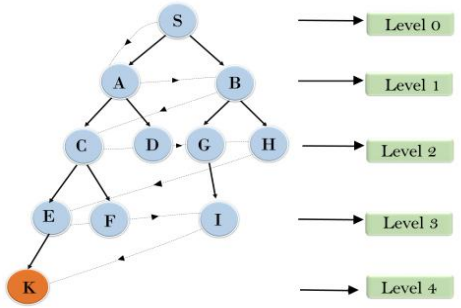
[1]

[2]

[1] Source: Tesla AI Day 2021
[2] Source: Cruise Under the Hood 2021

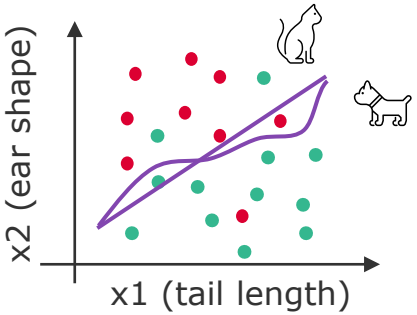
But Why? A Look into these "AI" Models

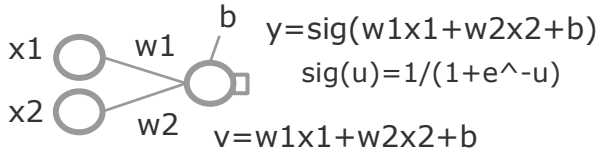
Breadth-first search

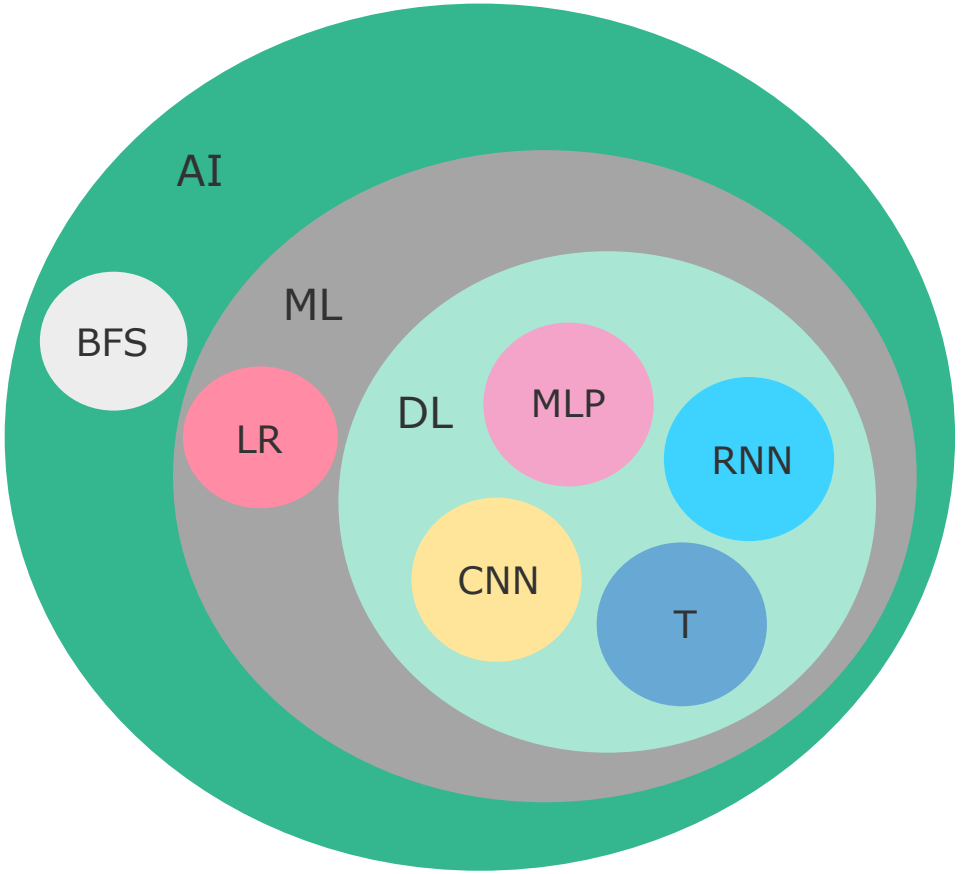


(source: Javapoint)

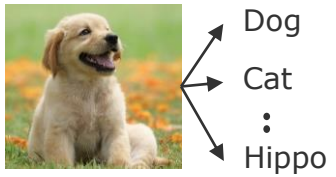
Logistic regression

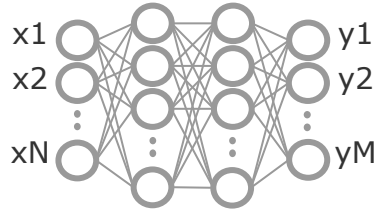




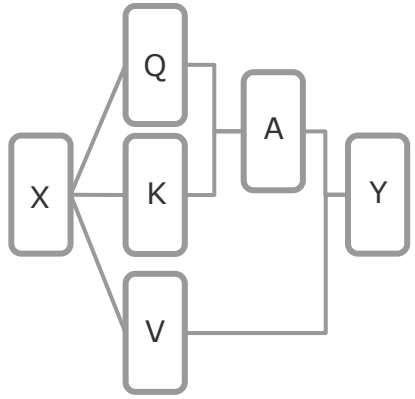


Multi-layer perceptron



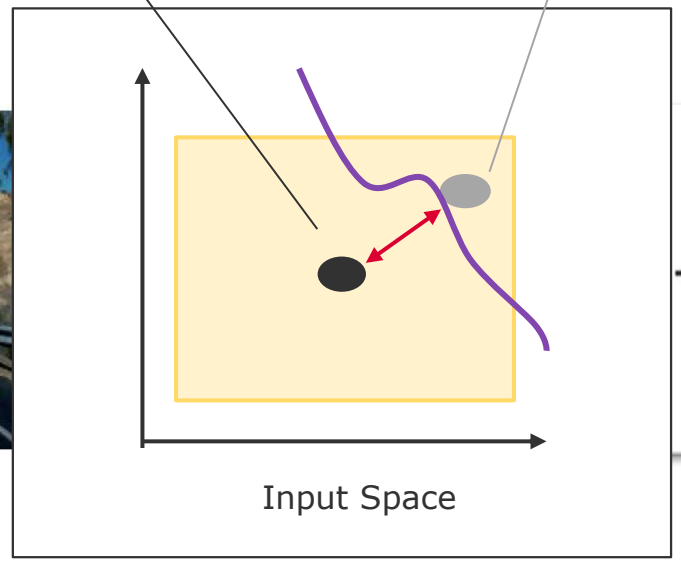


Transformer (self-attention)

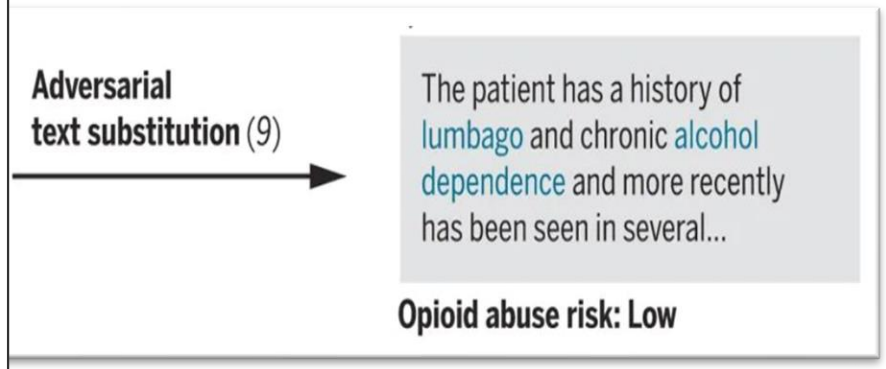


The Robustness Problem of NNs

[1]

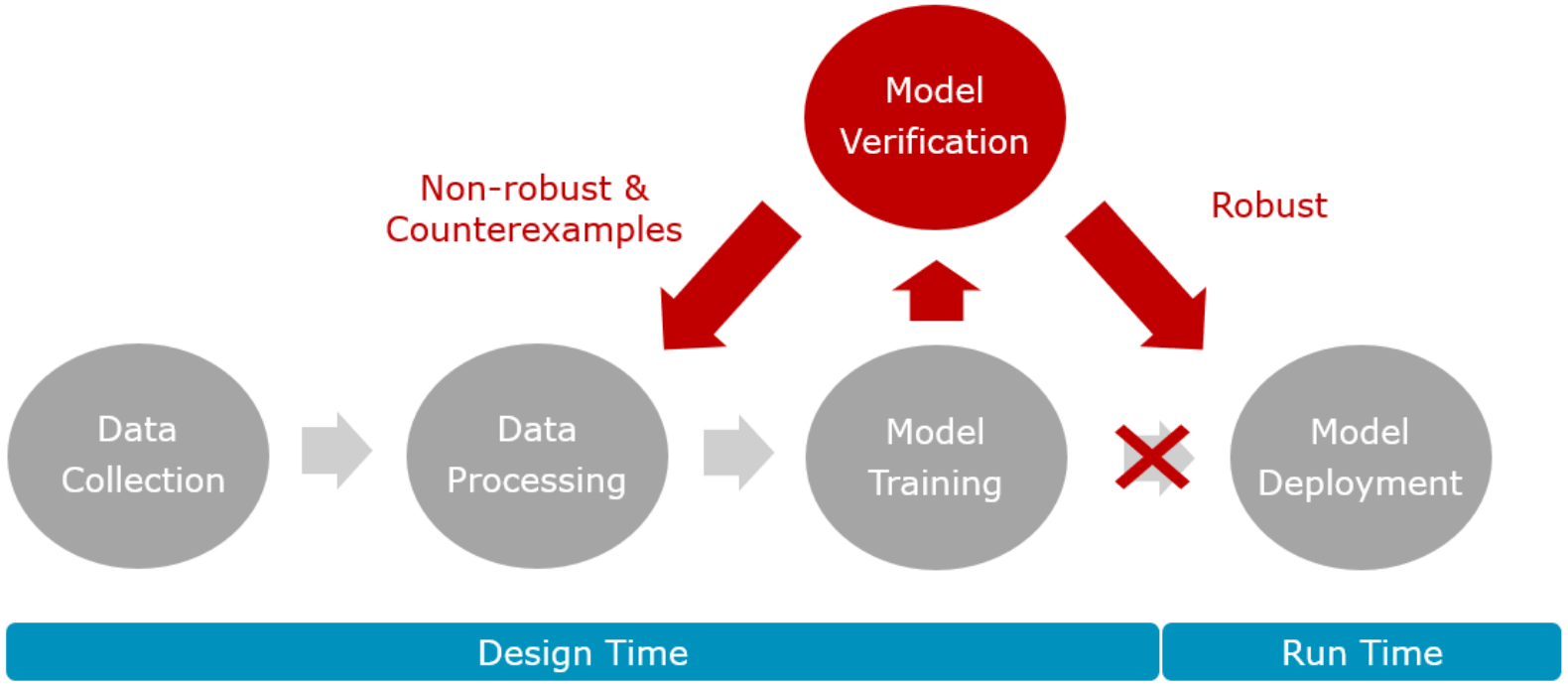


[3]

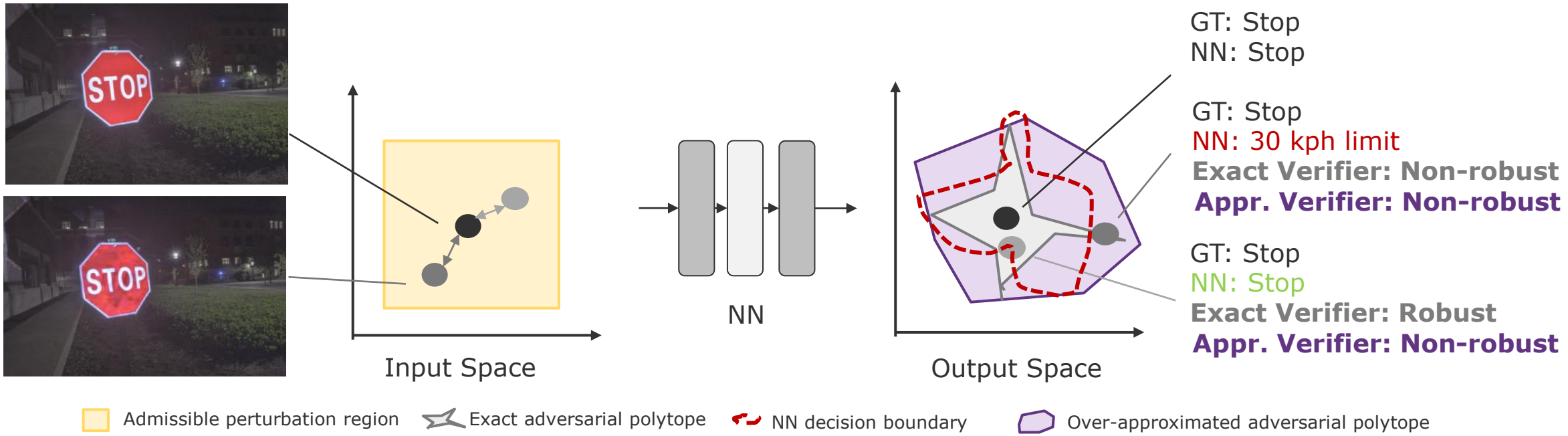


[1] "Optical adversarial attacks," Gnanasambandam *et al.*, *ICCV,* 2021.
[2] "DeepXplore: Automated whitebox testing of deep learning systems," Pei *et al.*, *SOSP*, 2017.
[3] Cary *et al.*, "Adversarial attacks on medical machine learning," in *Science, 2020.*

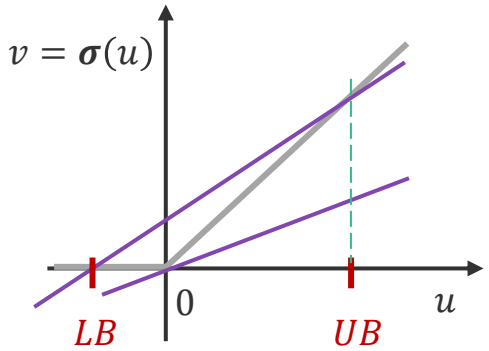
Robustness Verification as a Potential Remedy



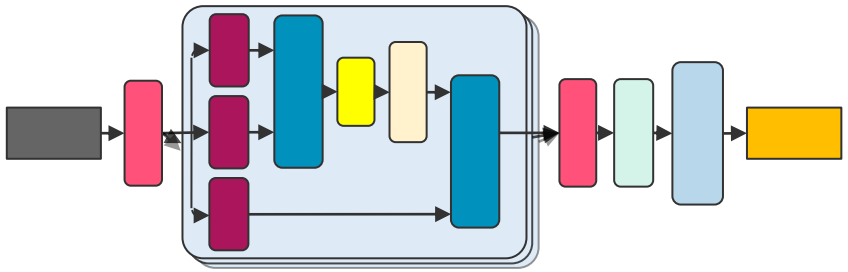
Main Approaches and Common Challenges



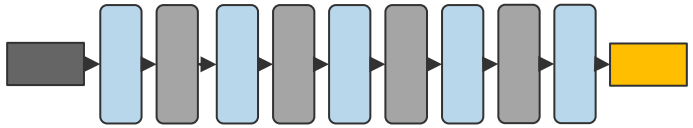
- Challenge 1: The non-convex activations, e.g., ReLU

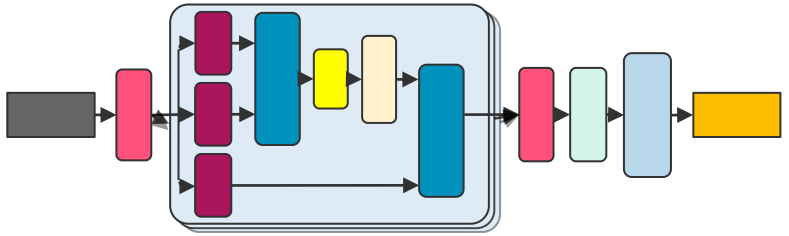


- Challenge 2: Emerging architectures and operations



State of the Art in NN Robustness Verification

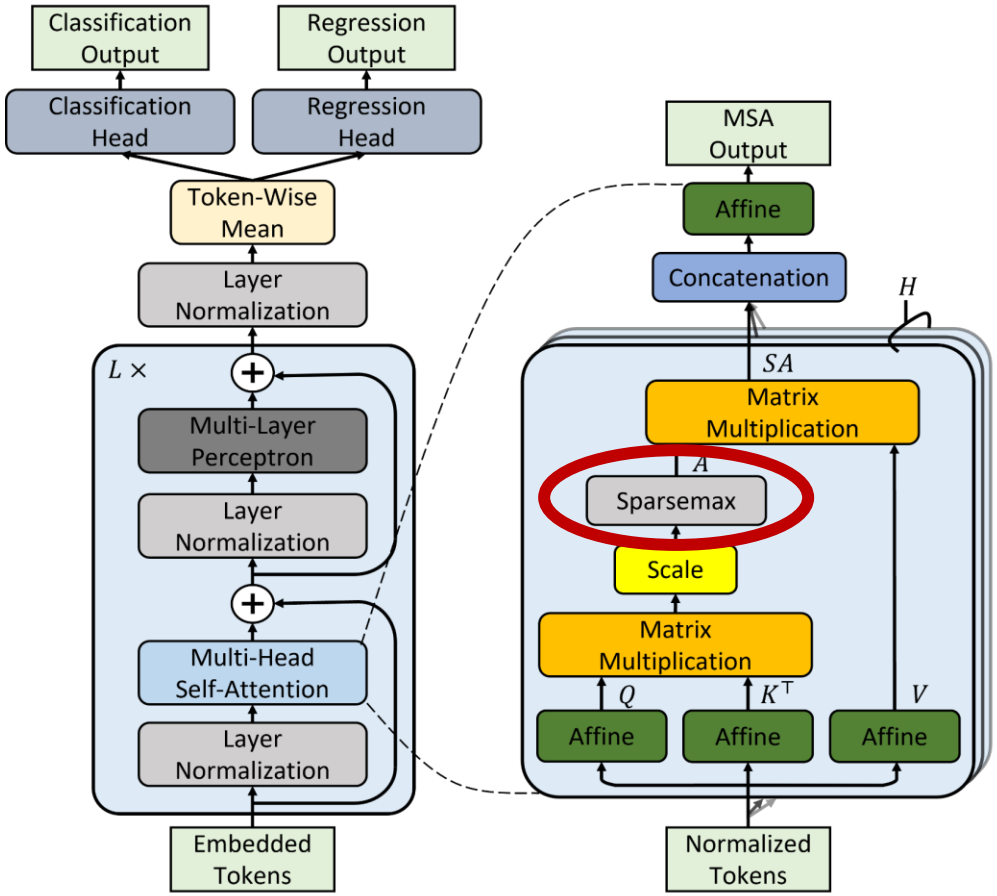


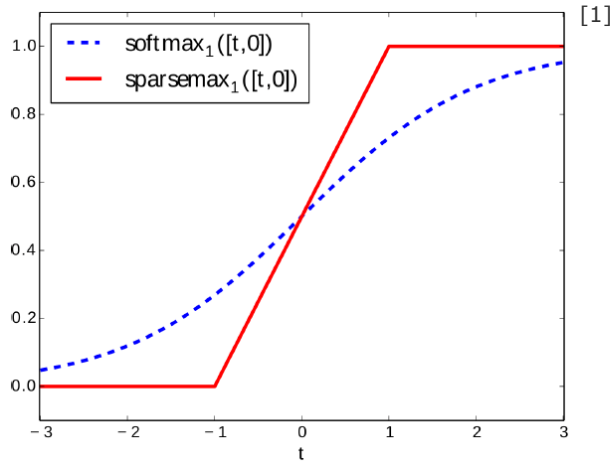


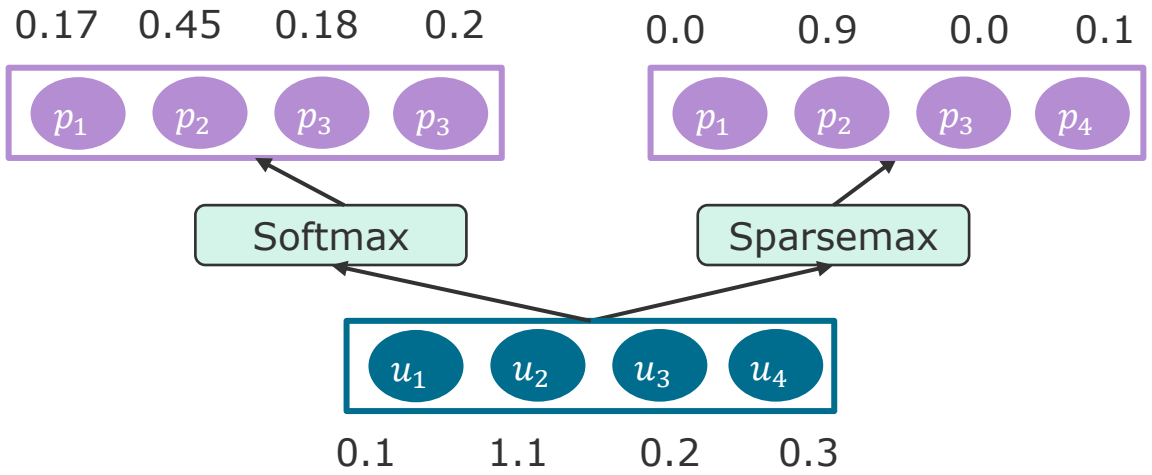
	MLPs (or CNN/RNN variants)	Transformers
Approximate Verifier	~10e6 neurons [1] Image Classification (CIFAR-10)	~10e4 neurons [3] Sentiment Analysis (Yelp)
Exact Verifier	~10e5 neurons [2] Image Classification (CIFAR-10)	Lacking, but a network of size ~10e3 should be verifiable

[1] Xu *et al.*, "Enabling complete NN verification with rapid and massively parallel incomplete verifiers," in *ICLR,* 2021.
[2] Tjeng *et al.*, "Evaluating robustness of NNs with MIP," in *ICLR,* 2019.
[3] Shi *et al.*, "Robustness verification for transformers," in *ICLR*, 2020.

Our Focus – Sparsemax Transformers



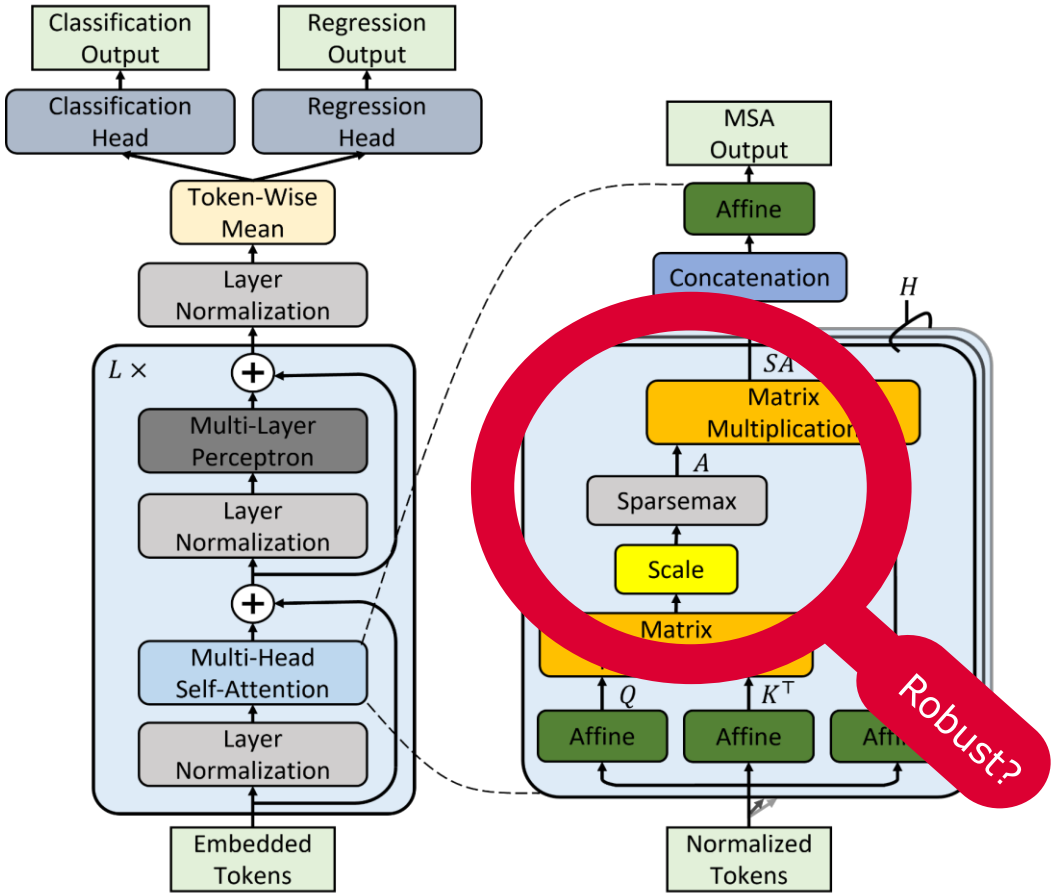




[1] Martins *et al.*, "From softmax to sparsemax," in *ICML*, 2016.

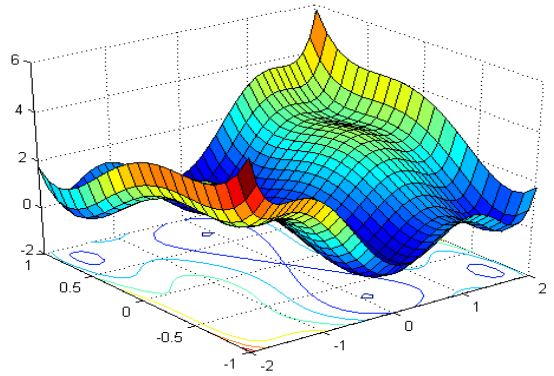
Reducing Robustness Verification to an Optimization Problem

Sparsemax Transformers



Encode

MIQCP (Mixed Integer Quadratically Constrained Programming)



$$\min Dist_p(x, x') \quad \text{①}$$

$$s.t. \quad x' \in \mathrm{B}_p(x), \quad \text{②}$$

$$f(x) = gt(x) \wedge f(x') \neq gt(x). \quad \text{③}$$

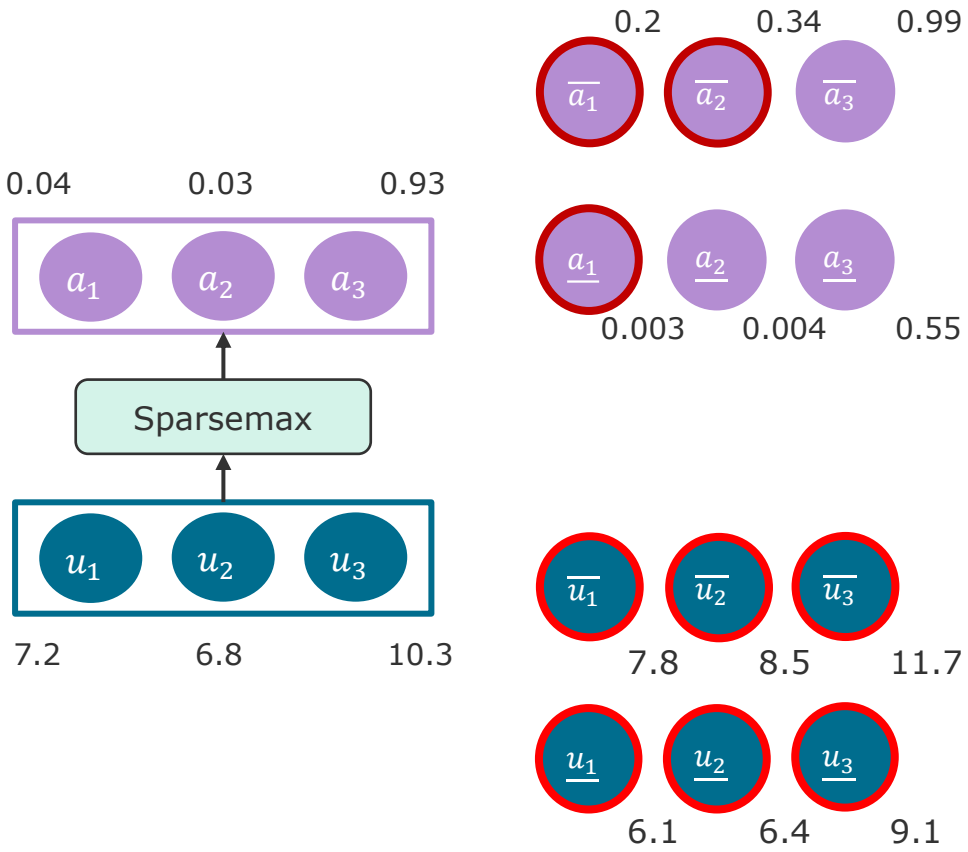
① Find the min perturbation,

② within a budget, e.g., ε

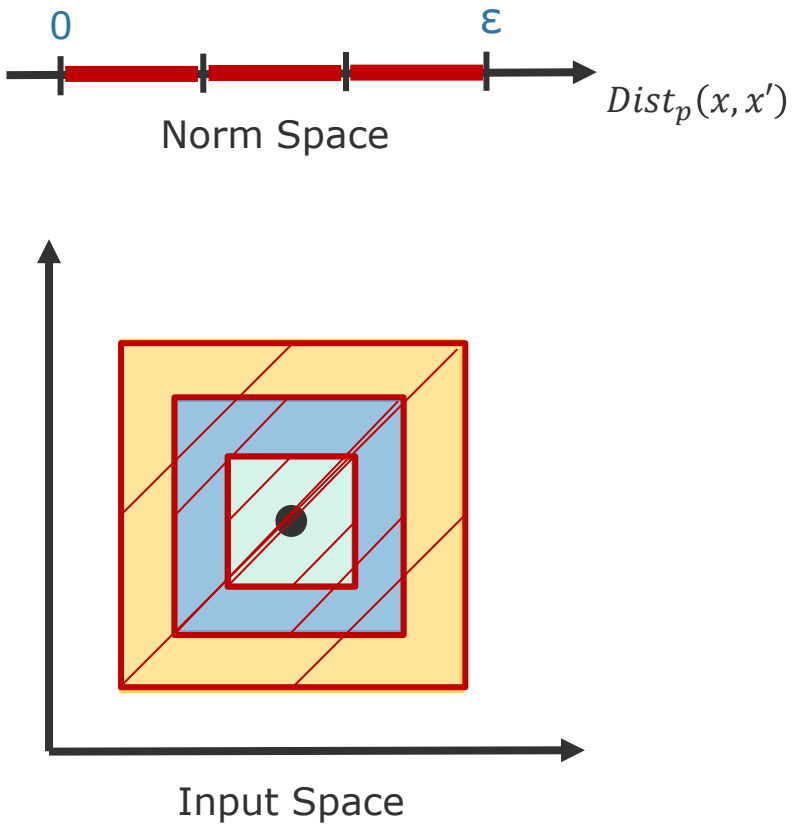
③ where the perturbed input causes a wrong prediction.

Accelerating Heuristics

- Interval analysis on Sparsemax activation [1]

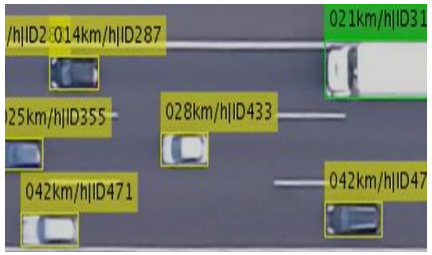


- Norm-space partitioning



[1] A recent improvement in Wei *et al.*, "Convex bounds on the softmax function with applications to robustness verification," in *AISTATS*, 2023.

Experiment – Lane Departure Warning

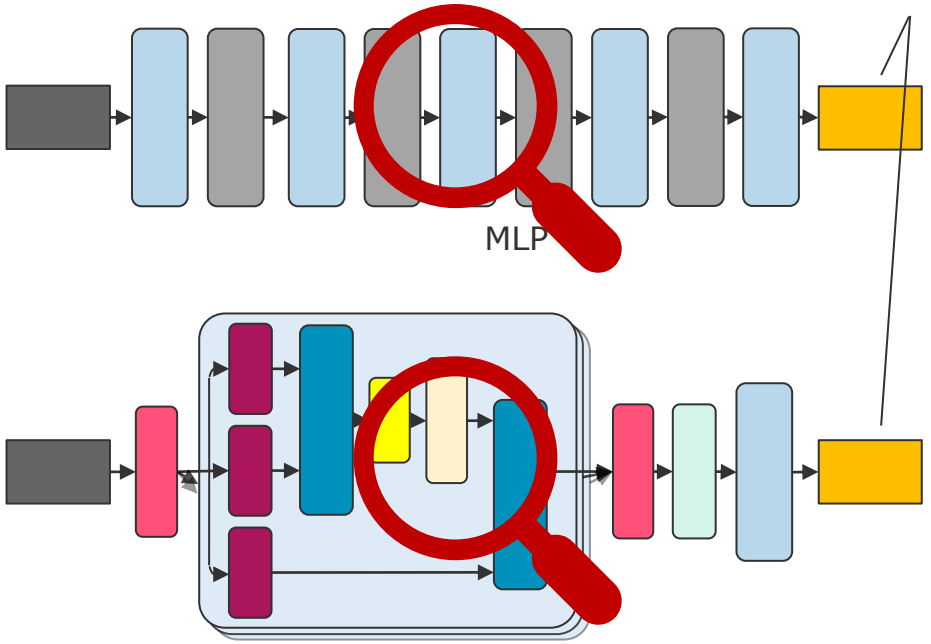


HighD Dataset [1]

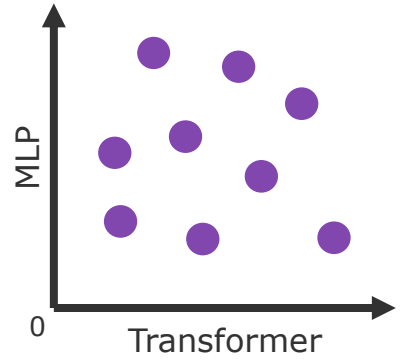
T

No.	Features at t
1	Left lane existence
2	Right lane existence
3	Distance to lane center
4	Longitudinal velocity
5	Lateral velocity
6	Longitudinal acceleration
7	Lateral acceleration
8-14	TTC of surrounding cars [2]

Extracted Features



Neural Networks



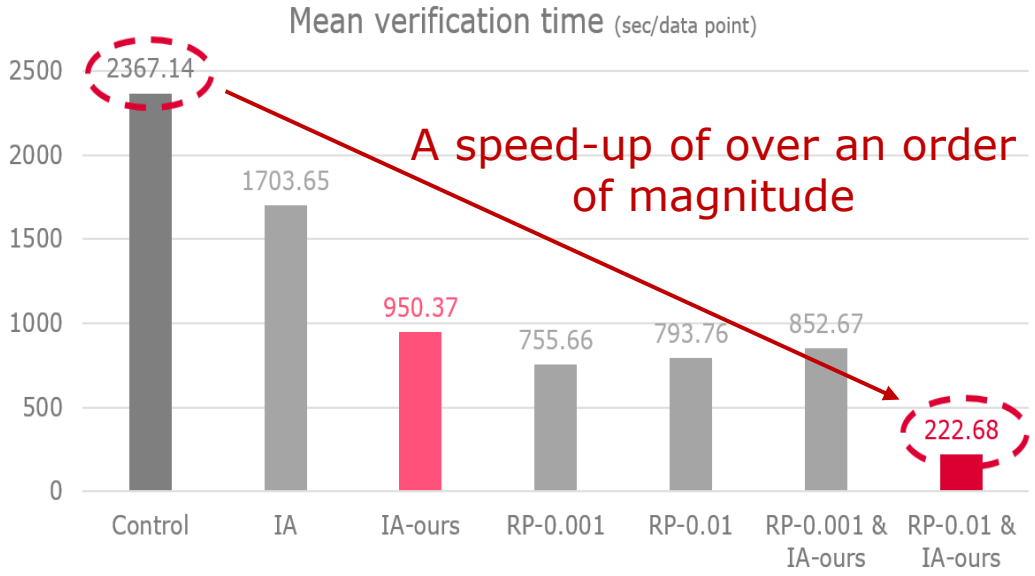
Robustness Values

Data Collection | Data Processing | Model Training | Model Verification

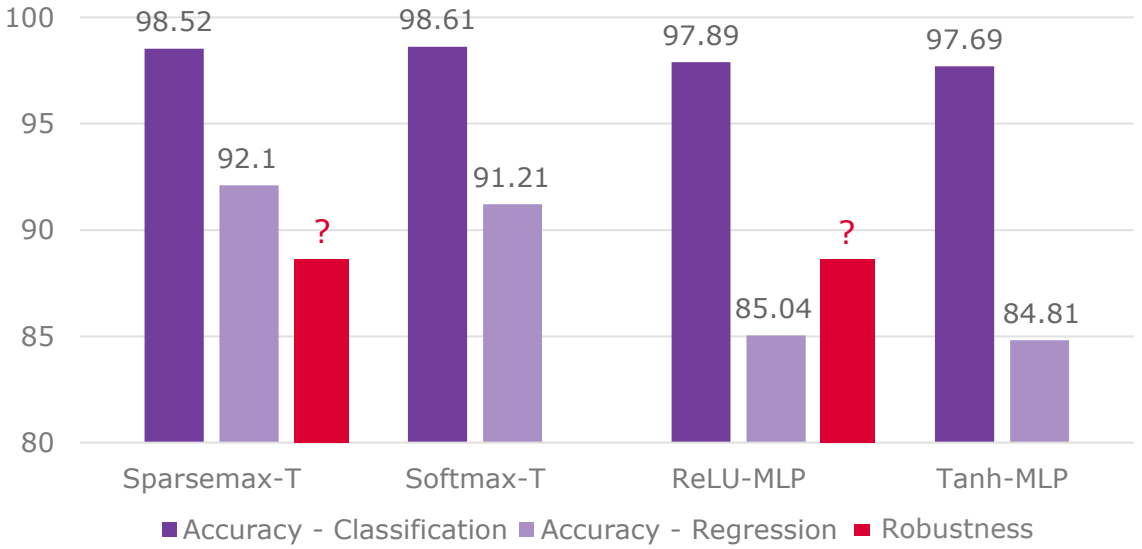
[1] https://www.highd-dataset.com/.
[2] TTC: Time to collision.

Results – Ablation Study and Accuracy Assessment

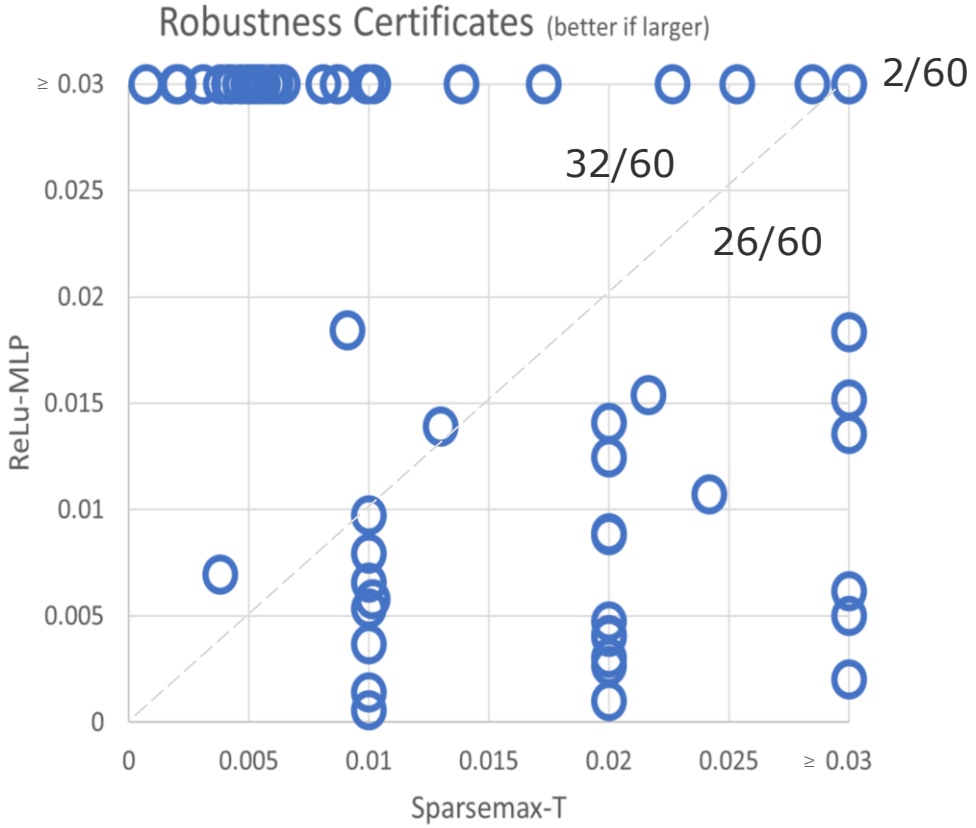
- On the proposed techniques



- Classification and regression accuracy



Results – Robustness Benchmark



Statistics from verifying 5 ReLU-MLPs and 5 Sparsemax-Ts with 100 data points

	Mean	Var.	No. Max (0.01)
ReLU-MLPs	--	--	100
Sparsemax-Ts	0.0041	0.0006	69

- The Transformer seems less robust than the MLP.
- A recent empirical work in vision tasks concludes similarly. [1]
- Yet, some others disagree. [2, 3]

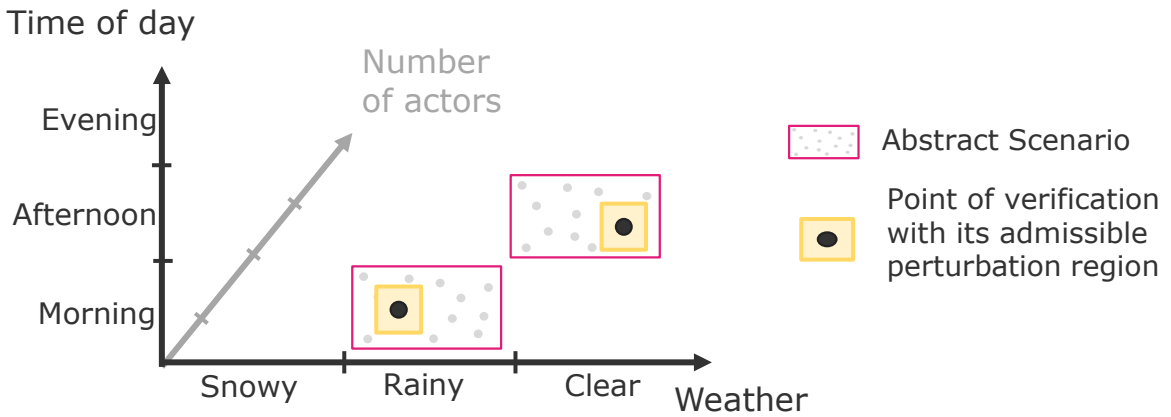
[1] Wang *et. al.*, "Can CNNs be more robust than Transformers," in *ICLR*, 2023.
[2] Bhojanapali *et. al.*, "Understanding robustness of transformers for image classification," in *ICCV*, 2021.
[3] Shao *et. al.*, "On the adversarial robustness of vision transformers," in *UCLR*, 2021.

Summary

- Robustness is a safety-related concern in NNs.
- Our study focuses on exact robustness verification for a specific variant of Transformers.
- Robustness, as an application-oriented property, needs to be verified before NN deployment.

Limitations and Open Directions

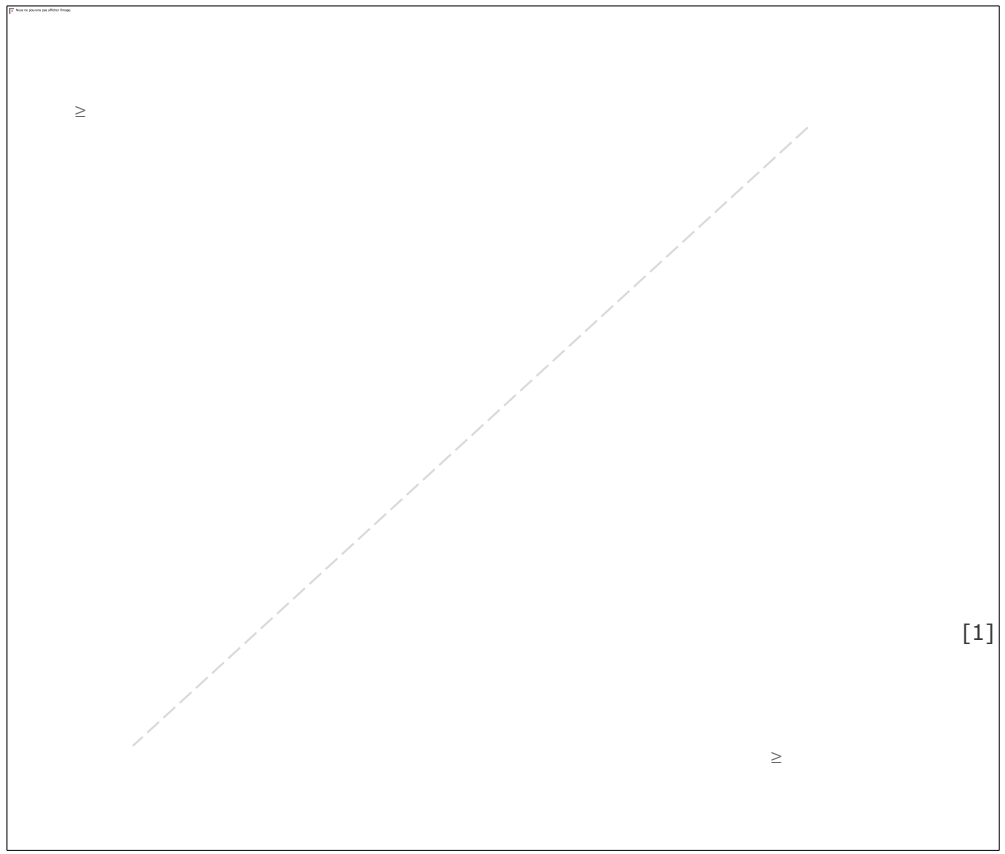
- Softmax not considered ➔ Iterative bound tightening (e.g., with Branch-and-Bound)
- Small NNs and simple task ➔ Real-world applications (e.g., via probabilistic verification)
- Point-wise analysis ➔ Domain-covered assurance (e.g., with combinatorial testing)
- Design-time verification ➔ Run-time verification (e.g., with different sensor modalities)



DENSO

Crafting the Core

Incomplete Verification due to Timeouts



[1]

[1] The points marked with “Lower” give the lower bounds for the Transformer.