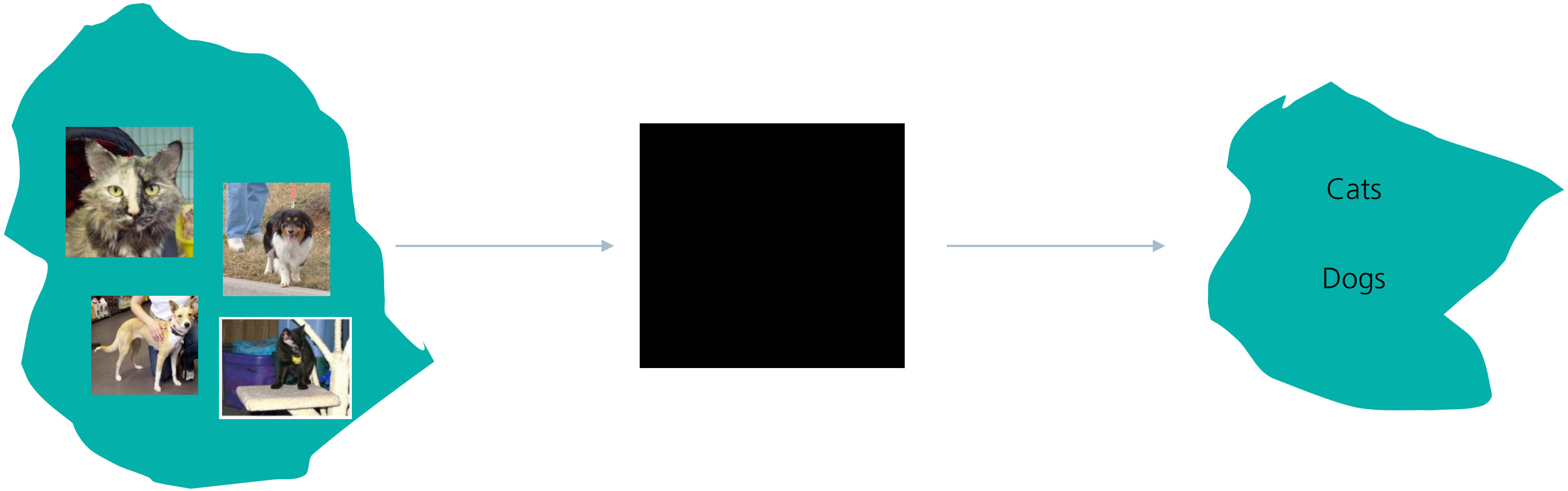
Sujan Gannamaneni | 12.12.2022

Identifying Systematic Weaknesses of DNNs through Slice Discovery Methods

Agenda

1	**Introduction**	High-Level Introduction to DNN Testing
		(Some) Failure Modes in DNNs
2	**Identifying Systematic Weaknesses**	Problem Formulation on Structured Data
		Slice Discovery Methods on Structured Data
		Structured vs Unstructured Data
		Slice Discovery Methods on Unstructured Data
		Evaluating Unstructured Data using SDMs used for Structured Data
3	**Final Outlook**	Open Questions and Conclusion

1 High-Level Introduction to DNN Testing



1 High-Level Introduction to DNN Testing

System-under-test (SUT)

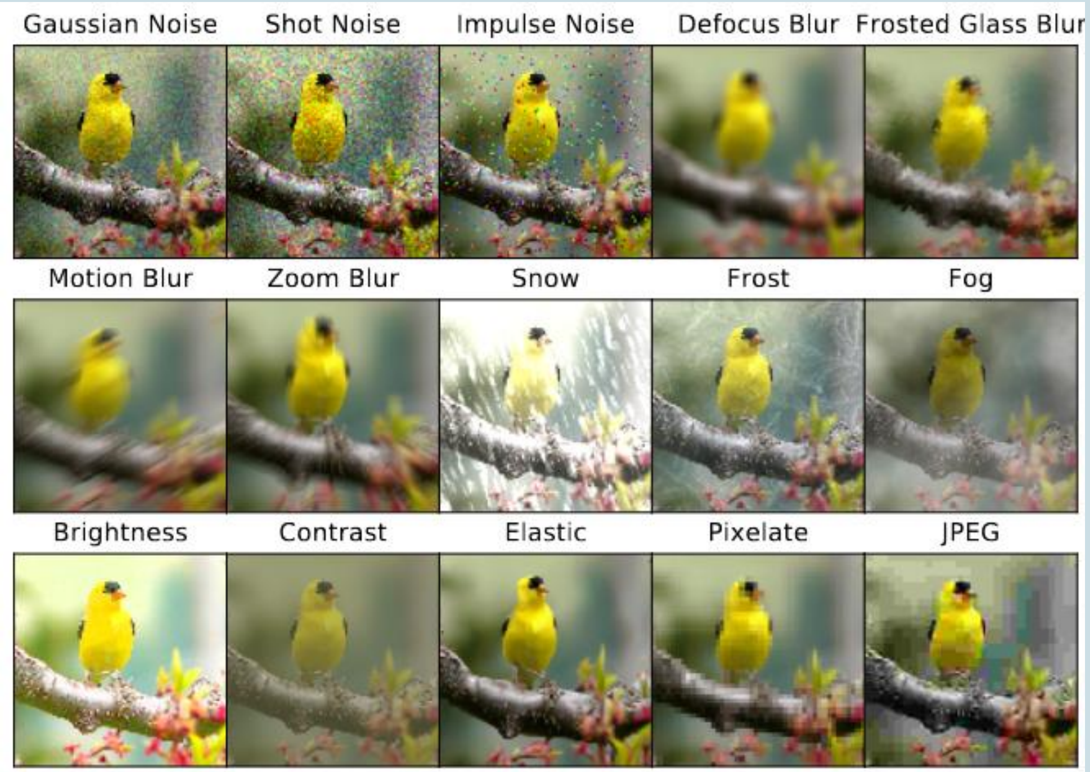
Open question: How do we audit the SUT?

- Understanding the failure modes of SUT
- Develop methods to quantify/evaluate each failure mode

1 (Some) Failure Modes in DNNs

1 - Lack of Robustness

- Performance drop due to environment & sensor
- More prominent in computer vision tasks
- Current research in direction of
  - Quantifying robustness of models
  - Improving robustness of models



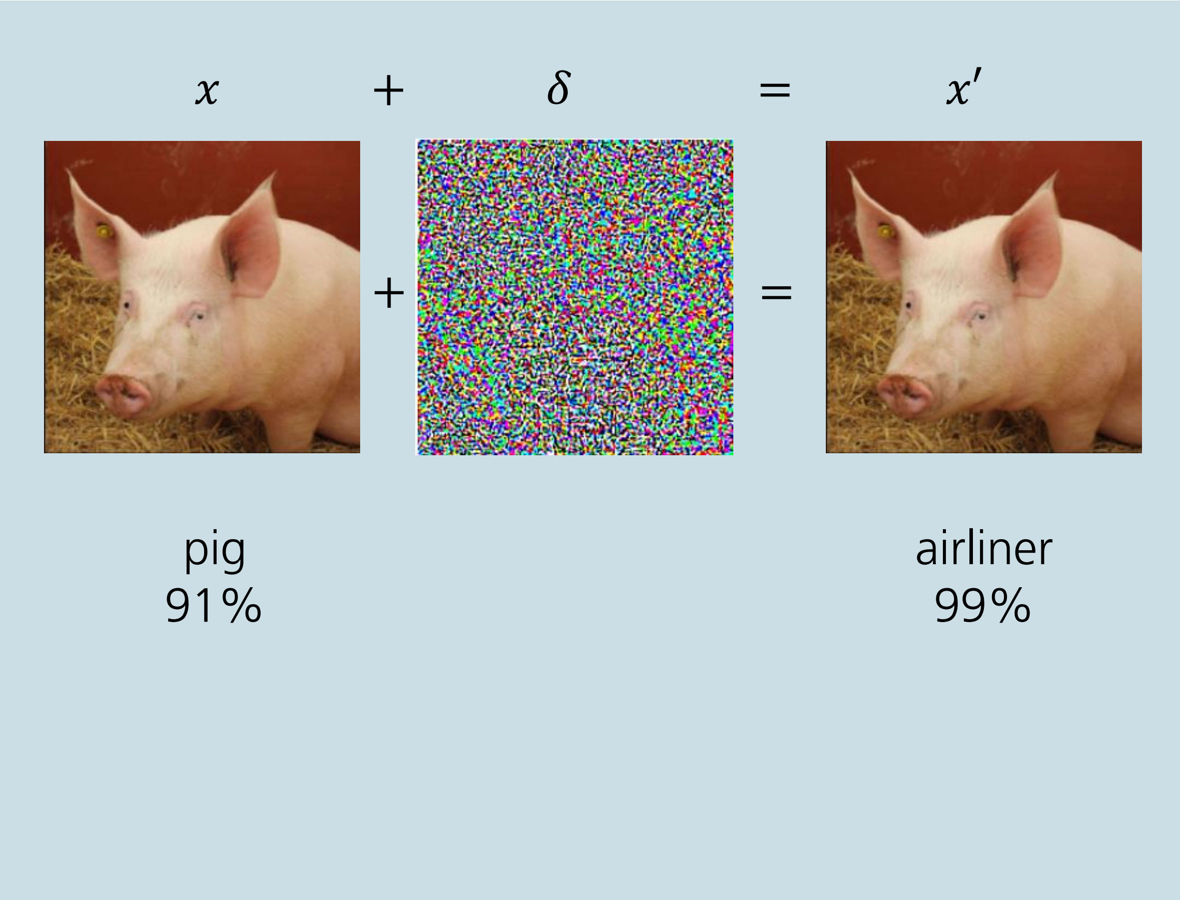
Hendrycks, D. et al. *Benchmarking neural network robustness to common corruptions and perturbations.* 2019

Fraunhofer IAIS

1 (Some) Failure Modes in DNNs

2 - Susceptibility to Adversarial Attacks

- Failure due to model fragility
- Several sophisticated methods exist for targeted and untargeted attacks
- Adversarial defenses exist (but no single good defense for all attacks)

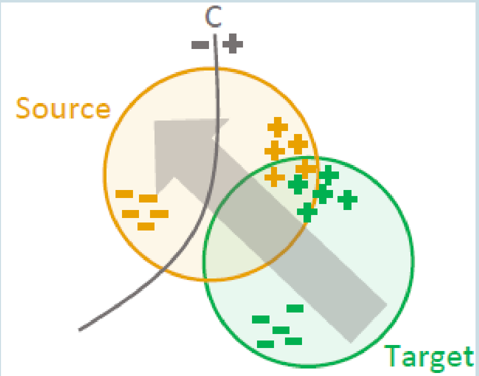


https://adversarial-ml-tutorial.org/introduction/

1 (Some) Failure Modes in DNNs

3 - Lack of Domain Generalization

- Models trained on source domain might not directly perform well on target domains
- Problem lies in
  - Insufficient generalization capabilities of DNNs
  - Vague formulation of data distributions
- Improving domain generalization could help in unlocking usefulness of synthetic data for training and testing



Source

Target

Luo, Y. et al Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation 2019

Li, Y. et al. Bidirectional learning for domain adaptation of semantic segmentation. 2019

1 (Some) Failure Modes in DNNs

What about semantics of objects?

- Not enough focus on failure modes w.r.t. semantic content of objects in the image
- Classic example – Fairness of DNNs

  Performance on full test not representative on all (semantic) subsets of data

- **Systematic weaknesses of DNNs on data subsets**

Data		Performance
Full test data		Acceptable performance
Data subset-1		Good performance
Data subset-2		Low performance

Buolamwini, J., & Gebru, T. (2018, January). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on fairness, accountability and transparency (pp. 77-91). PMLR.

1 (Some) Failure Modes in DNNs

Systematic weaknesses of DNNs on data subsets

Why is studying this failure mode useful?

Consider a DNN in an autonomous driving task in the following two situations

Situation	Performance metric	Actionable information for ML developer	Actionable information for ML auditor
Typical DNN testing	$Perf_{all-pedestrians} = 65\%$	Done?	Approved?
Systematic weakness analysis	$Perf_{red-shirted-pedestrians} = 35\%$ $Perf_{rest-pedestrians} = 95\%$	Add more red-shirted pedestrians to the training data	Model cannot perform well on red-shirted pedestrians. Is that acceptable?

2 Problem Formulation on Structured Data

Definitions and setup 1/2

- Slice discovery methods (SDMs)

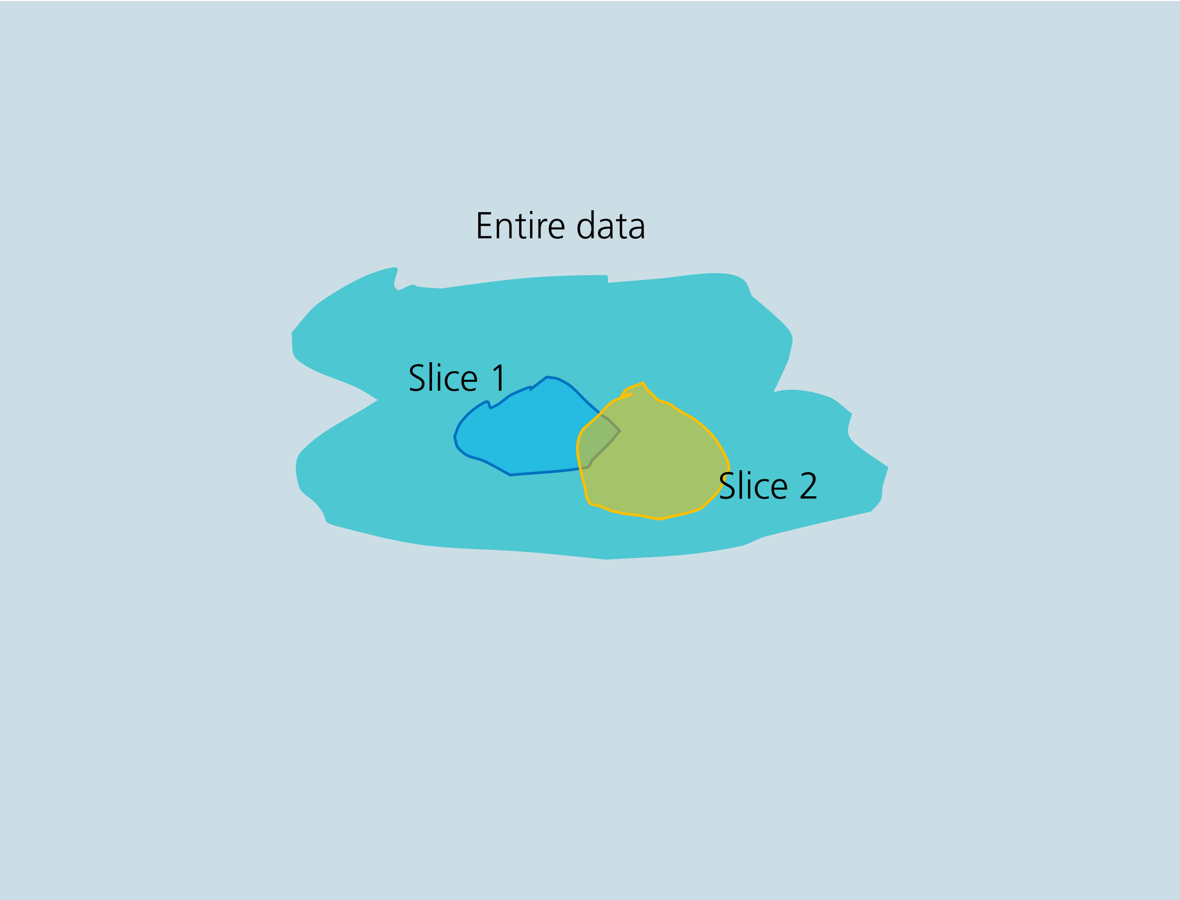
  Finding top-k weak performing slices from the data

- Slice

  A semantic (coherent) subset of the data

- Features or metadata

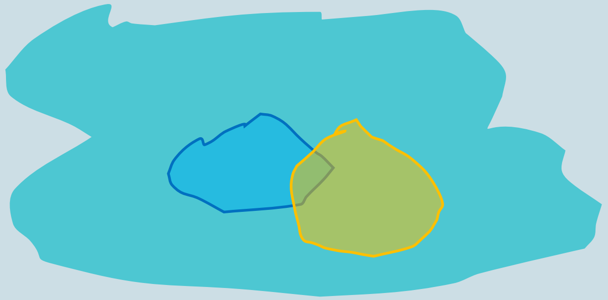
  Human-understandable semantic attributes describing the data



2 Problem Formulation on Structured Data

Definitions and setup 2/2

- Consider a dataset of m features and n samples
- Transform data such that
  - Categorical values → one-hot encoded
  - Numerical values → Binning → one-hot encoded
- Slice is defined as an "AND" combination of
  - Multiple features
  - One value per feature
- e.g., Slice A=(Feature 1=Blue, Feature 2=15-20)



Sample no.	Feature 1 (Categorical)	Feature 2 (Binned continuous)	…	Feature m	Performance / Error
Sample 1	Blue	10-15			
Sample 2	Red	15-20			
…					
Sample n	Green	30-35			

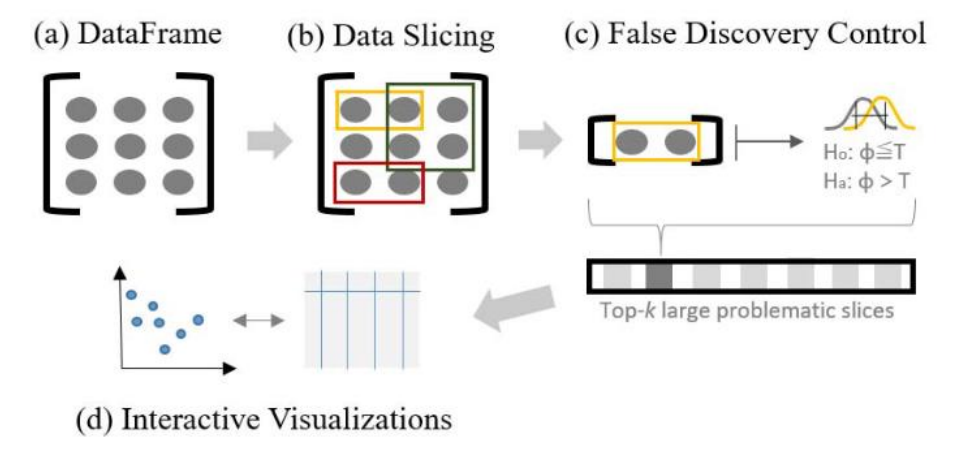
  - Öffentlich -

Fraunhofer IAIS

2 Slice Discovery Methods on Structured Data

SliceFinder (Chung et al.)

- How does it work?
  - Transform data into one-hot encoded values
  - Proposal of different slice combinations based on features
  - Order slices based on criteria and obtain top-k slices
- Criteria
  - Number of semantic features ⇧
  - Size ⇩
  - Effect size ⇩



Slice	Log Loss	Size	Effect Size
All	0.35	30k	n/a
Sex = Male	0.41	20k	0.47
Sex = Female	0.21	10k	-0.47
Workclass = Local-gov Race = White	0.43	1.7k	0.19
Education = HS-grad	0.32	9.8k	-0.09
Education = Bachelors	0.44	0.5k	0.27
Education = Masters	0.49	1.6k	0.40
Education = Doctorate	0.47	5k	0.32

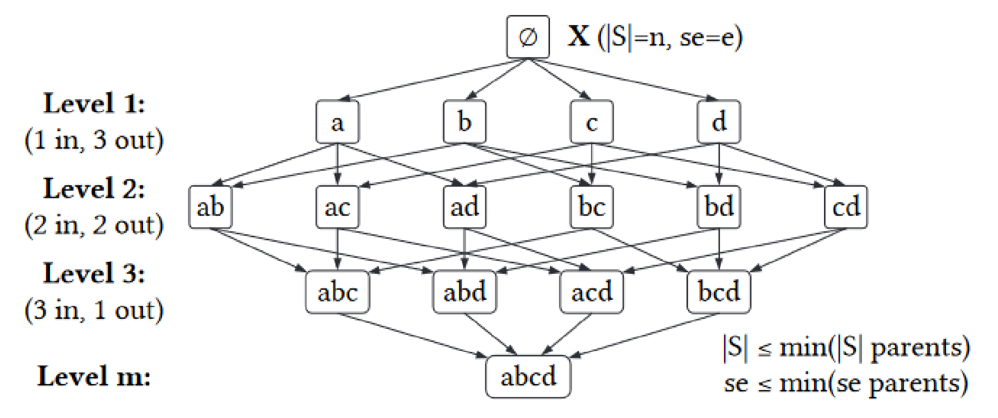
Chung, Y., Kraska, T., Polyzotis, N., Tae, K. H., & Whang, S. E. (2019, April). Slice finder: Automated data slicing for model validation. In 2019 IEEE 35th International Conference on Data Engineering (ICDE) IEEE.

2 Slice Discovery Methods on Structured Data

Sliceline (Sagadeeva et al.)

- Inspired from Slicefinder method
- Overcomes shortcomings in Slicefinder w.r.t. missing certain slices
- Proposes a new scoring function that considers slice size and slice error
- Builds a lattice structure that can be effectively pruned based on
  - Monotonicity property
  - Scoring function
- Provides a fast linear algebra-based enumeration algorithm to solve this top-k weak slice discovery problem

$$sc = \alpha\left(\frac{\overline{se}}{\overline{e}} - 1\right) - (1-\alpha)\left(\frac{n}{|S|} - 1\right)$$



Sagadeeva, S., & Boehm, M. (2021, June). Sliceline: Fast, linear-algebra-based slice finding for ml model debugging. In Proceedings of the 2021 International Conference on Management of Data (pp. 2290-2299).

2 Structured vs Unstructured Data

Challenges in applying structured slice discovery methods

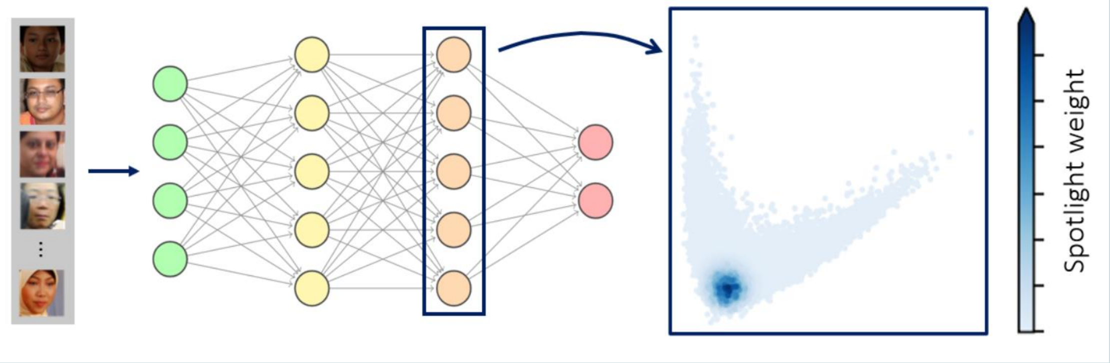
- Complex data is often unstructured, and this is where DNN application provides most benefits
- Unstructured data like images cannot be easily transformed into tabular format
- Different approaches
  - Use multimodal DNNs or SUT embeddings to bring structure to unstructured data
  - Use e.g., simulators to generate both unstructured and structured data

Sample no.	Feature 1 (Categorical)	Feature 2 (Binned continuous)	…	Feature m	Performance (Error)
Sample 1	Blue	10-15			
Sample 2	Red	15-10			
…					
Sample n	Green	30-40			

2 Slice Discovery Methods on Unstructured Data

Spotlight (d’Eon et al.)

- Domain agnostic
- Looks at final embeddings of model to identify contiguous regions of high loss and limited size (spotlights)
- Soft clustering by assigning data points into spotlights is the optimization problem
- Data samples membership to multiple clusters (spotlights) are typically evaluated
- Major problems
  - No description provided for slices. Manual inspection is required
  - Highly dependent on spotlight size (hyperparameter)



d'Eon, G., d'Eon, J., Wright, J. R., & Leyton-Brown, K. (2022, June). The spotlight: A general method for discovering systematic errors in deep learning models. In 2022 ACM Conference on Fairness, Accountability, and Transparency (pp. 1962-1981).

2 Slice Discovery Methods on Unstructured Data

Domino (Eyuboglu et al.)

- Embed

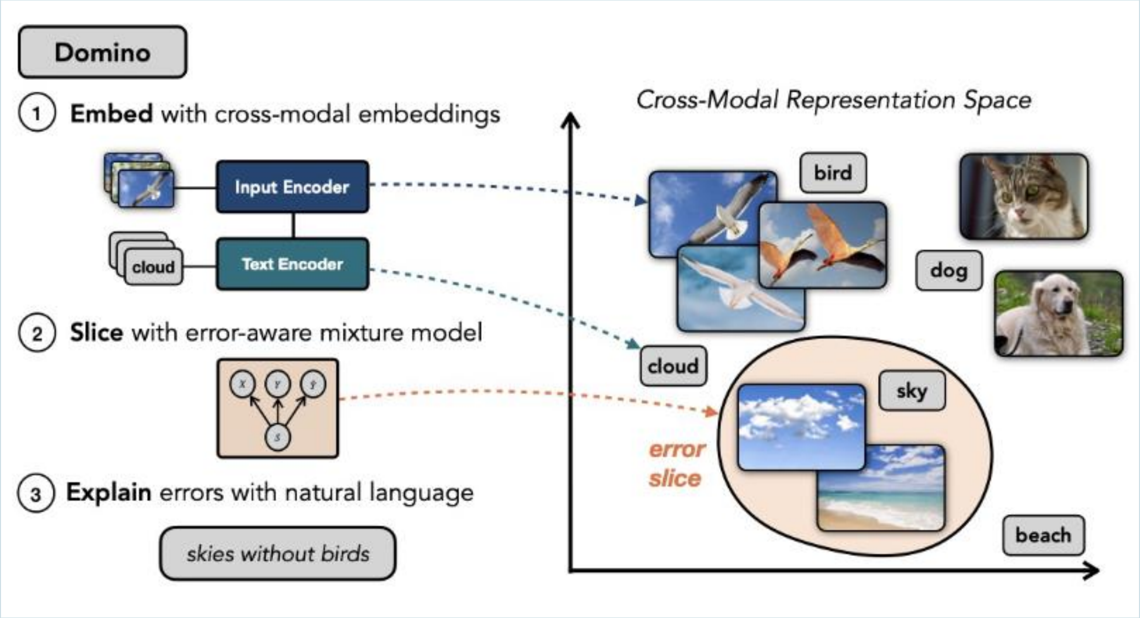
  Uses pre-trained CLIP model to embed image and text in common embedding space

- Slice

  Uses a variant of Gaussian mixture models to slice data and find weak performing slices

- Explain

  Uses a pre-trained BERT model to explain the weak slice embeddings

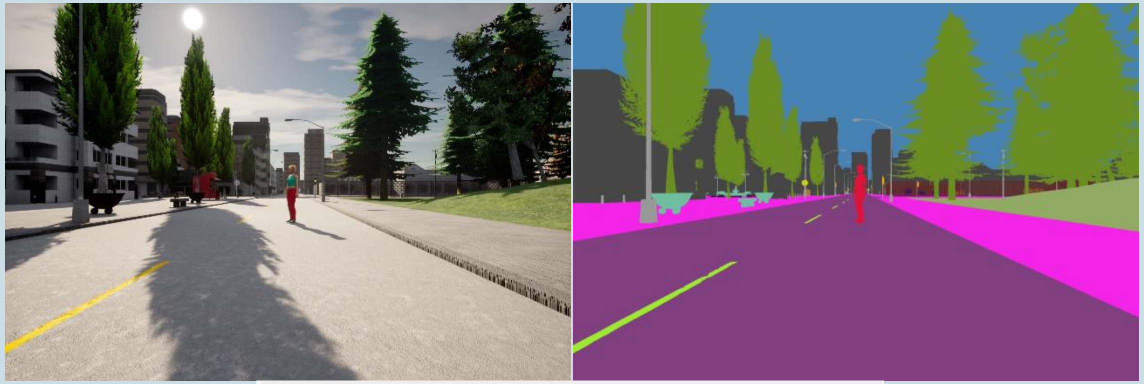


Eyuboglu, S., Varma, M., Saab, K., Delbrouck, J. B., Lee-Messer, C., Dunnmon, J., ... & Ré, C. (2022). Domino: Discovering systematic errors with cross-modal embeddings. arXiv preprint arXiv:2203.14960.

Evaluating Unstructured Data using SDMs used for Structured Data

Metadata creation from simulators

- Use computer simulators (e.g., Carla) to generate metadata about objects
- Metadata contains granular information about different semantics of pedestrians
- Models are
  - Trained using synthetic images and labels
  - Tested on these synthetic images and labels along with the granular metadata




```
{"Pedestrian_data": {
  "instance_id": 220,
  "pedestrian_asset_id": 0005,
  "world_x_coordinate": 190.0",
  "world_y_coordinate": 147.0,
  "gender": "female",
  "shirt_colour": "green",
  "pant_colour": "red",
  "skin_colour": "white",
  "age": "adult",},
  "Global_data": {
    "sun_angle":30.0,
    "sun_azimuth_angle":250.0,
    "fog_density":10.0,}}
```


Gannamaneni, S., Houben, S., & Akila, M. (2021). Semantic Concept Testing in Autonomous Driving by Extraction of Object-Level Annotations from CARLA. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 1006-1014).

2 Evaluating Unstructured Data using SDMs used for Structured Data

Building structured data from images

- Build structured data from the features & performance metrics
- Apply slice discovery methods for structured data on the tabular data
- Find top-k weak slices and evaluate if the information is actionable

Filename	xcoord	ycoord	asset_id	Age	ShirtColor	Skincolor	Pantcolor	Gender	carxcoord	carycoord
/data/share/KI-Absicherung/	2.44254E+16	2.97903E+15	25	adult	brown	brown	camo	male	3.37665E+16	3.02561E+1
/data/share/KI-Absicherung/	1.68528E+16	3.001E+15	9	child	grey	white	lightblue	female	3.37665E+16	3.02561E+1
/data/share/KI-Absicherung/	5.66027E+15	2.98525E+16	13	child	violet	white	lightbrown	male	3.37665E+16	3.02561E+1
/data/share/KI-Absicherung/	4.79722E+15	3.1029E+15	7	adult	maroon	white	grey	female	3.37665E+16	3.02561E+1
/data/share/KI-Absicherung/	4.47022E+15	1.84048E+16	5	adult	green	white	red	female	6.7114E+15	1.87532E+1
/data/share/KI-Absicherung/	5.6377E+15	1.95109E+16	9	child	grey	white	lightblue	female	6.7114E+15	1.87532E+1
/data/share/KI-Absicherung/	5.33671E+15	1.93504E+16	9	child	grey	white	lightblue	female	6.7114E+15	1.87532E+1
/data/share/KI-Absicherung/	2.4443E+16	1.96665E+16	25	adult	brown	brown	camo	male	6.7114E+15	1.87532E+1
/data/share/KI-Absicherung/	4.71035E+15	1.79757E+15	17	adult	blue	tanned	brown	male	6.7114E+15	1.87532E+1
/data/share/KI-Absicherung/	-2.77087E+16	1.98335E+15	17	adult	blue	tanned	brown	male	6.7114E+15	1.87532E+1
/data/share/KI-Absicherung/	1.34155E+15	2.13011E+15	23	adult	darkblue	brown	grey	female	6.7114E+15	1.87532E+1
/data/share/KI-Absicherung/	2.62306E+15	1.94626E+15	20	adult	brown	brown	darkblue	female	6.7114E+15	1.87532E+1
/data/share/KI-Absicherung/	1.49616E+16	2.31835E+16	22	adult	grey	brown	orange	female	1.69683E+16	2.37324E+1
/data/share/KI-Absicherung/	1.56572E+16	2.32731E+16	2	adult	white	white	blue	male	1.69683E+16	2.37324E+1
/data/share/KI-Absicherung/	1.44085E+15	2.31991E+16	17	adult	blue	tanned	brown	male	1.69683E+16	2.37324E+1
/data/share/KI-Absicherung/	1.49889E+16	2.32402E+15	15	adult	brown	brown	black	female	1.69683E+16	2.37324E+1
/data/share/KI-Absicherung/	5.63127E+15	2.43081E+16	18	adult	brown	brown	grey	male	1.69683E+16	2.37324E+1
/data/share/KI-Absicherung/	1.3318E+16	2.28679E+16	11	child	yellow	white	darkgreen	female	1.69683E+16	2.37324E+1
/data/share/KI-Absicherung/	1.39085E+15	3.0013E+16	20	adult	brown	brown	darkblue	female	1.44719E+16	3.0226E+1
/data/share/KI-Absicherung/	5.49226E+14	3.11235E+15	21	adult	red	brown	darkblue	female	1.44719E+16	3.0226E+1
/data/share/KI-Absicherung/	1.11068E+16	3.10161E+16	15	adult	brown	brown	black	female	1.44719E+16	3.0226E+1
/data/share/KI-Absicherung/	1.02959E+16	3.11103E+16	3	adult	grey	white	lightblue	male	1.44719E+16	3.0226E+1
/data/share/KI-Absicherung/	3.5968E+15	3.11026E+15	18	adult	brown	brown	grey	male	1.44719E+16	3.0226E+1
/data/share/KI-Absicherung/	5.88582E+15	2.98866E+15	22	adult	grey	brown	orange	female	1.44719E+16	3.0226E+1
/data/share/KI-Absicherung/	5.14844E+15	2.98203E+15	7	adult	maroon	white	grey	female	1.44719E+16	3.0226E+1
/data/share/KI-Absicherung/	4.17843E+15	3.09531E+15	26	adult	yellow	brown	brown	male	1.44719E+16	3.0226E+1
/data/share/KI-Absicherung/	3.88597E+15	2.54827E+15	25	adult	brown	brown	camo	male	4.62671E+16	2.67078E+1
/data/share/KI-Absicherung/	3.68287E+15	2.53586E+15	12	child	brown	white	lightblue	male	4.62671E+16	2.67078E+1

Gannamaneni, S., Houben, S., & Akila, M. (2021). Semantic Concept Testing in Autonomous Driving by Extraction of Object-Level Annotations from CARLA. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 1006-1014).

3 Open Questions and Conclusion

- Open questions
  - Is granularity in metadata important? → Could improve quality
  - How can we evaluate or compare SDMs? Qualitative → Quantitative
  - Are SDMs for structured data better than SDMs for unstructured data?
- Synthetic data help in evaluating and comparing SDMs → Bridging unstructured to structured
- However, at the end, SDMs need to work on real-world data
- Techniques to generate metadata for real-world data are required and would have following benefits
  - Helps in evaluating the DNNs trained on real-world data
  - Helps in defining Operational Design Domain (ODD)

Contact

Sujan Gannamaneni
sujan.sai.gannamaneni@iais.fraunhofer.de
+49 2241 14-2292

Fraunhofer-Institut für Intelligente Analyse-
und Informationssysteme IAIS
Schloss Birlinghoven
53757 Sankt Augustin
www.iais.fraunhofer.de