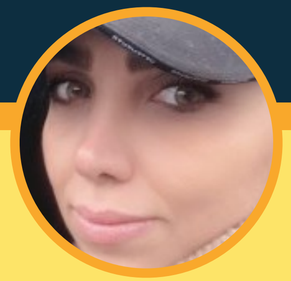
DEEPATASH

FOCUSED TEST GENERATION FOR DEEP LEARNING SYSTEMS



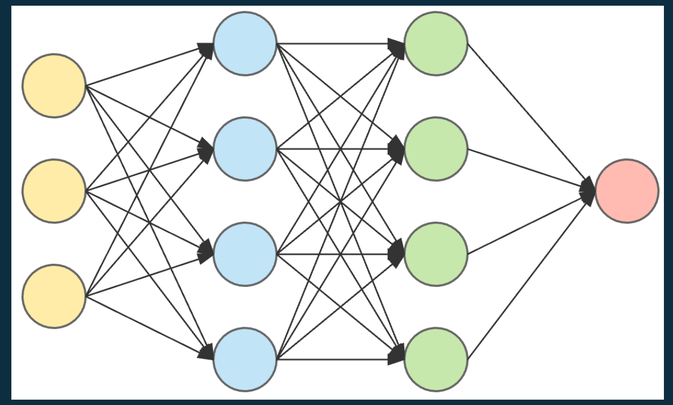
TAHEREH ZOHDINASAB

VINCENZO RICCIO

PAOLO TONELLA

TRADITIONAL DEEP LEARNING (DL) SYSTEM ASSESSMENT

ORIGINAL DATASET



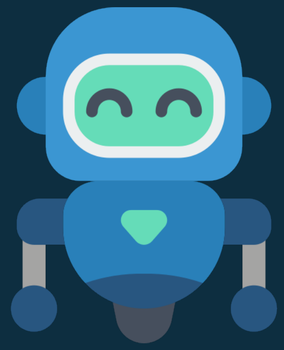
DL SYSTEM UNDER TEST

ACC = 95%

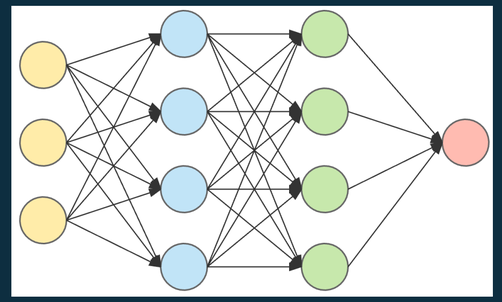
PERFORMANCE METRIC

Triggering misbehaviours of the DL system with inputs beyond its original dataset with feature combinations of interest

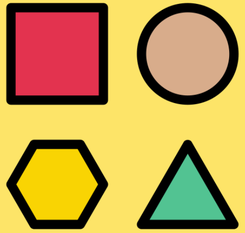
TEST GENERATION FOR DL SYSTEMS



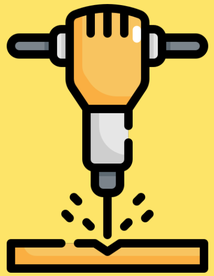
TEST GENERATOR



DL SYSTEM UNDER TEST



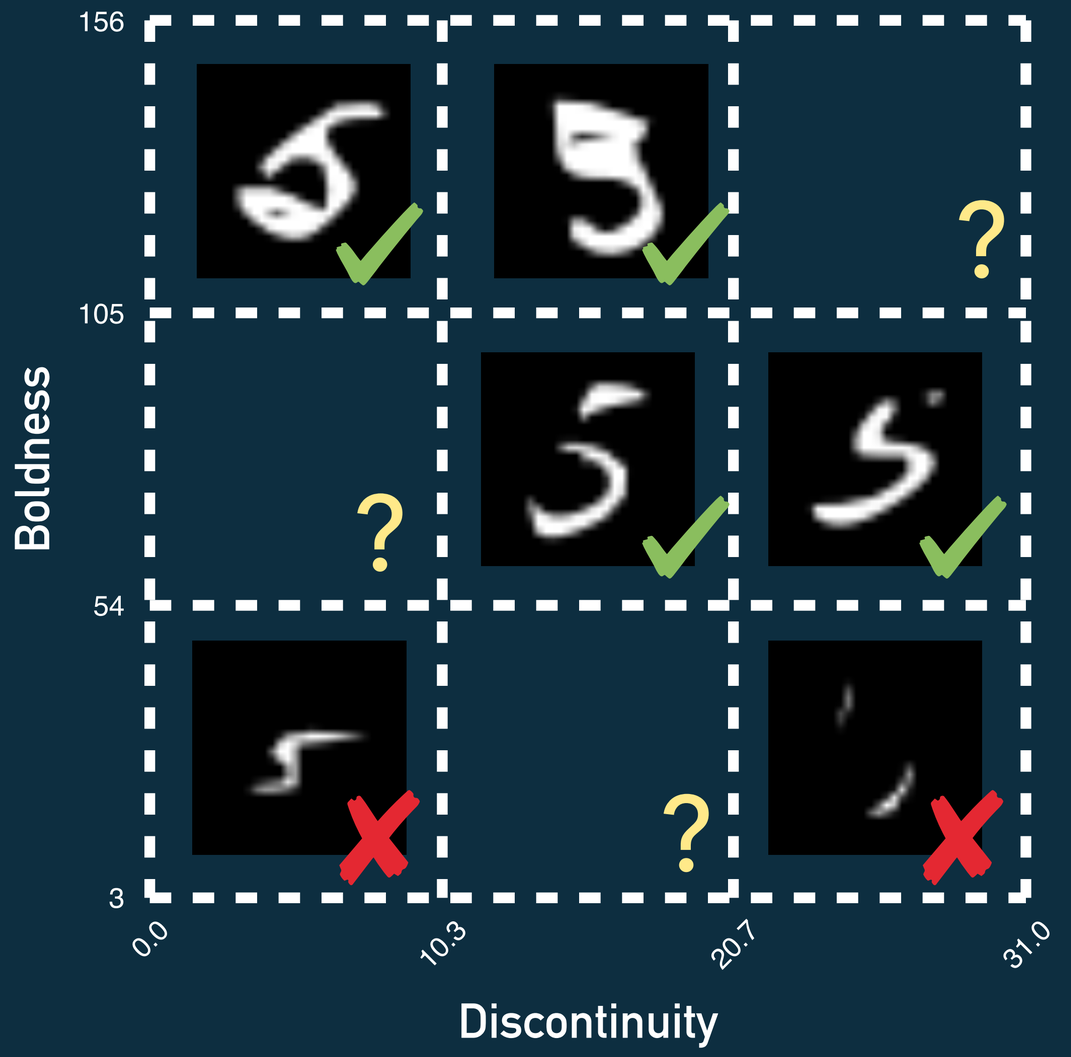
Diverse misbehaving inputs with critical features



Misbehaviours with input features that do not seem critical

Misbehaving inputs with unseen feature values

FEATURE MAPS [ZOHDINASAB ET AL., ISSTA 2021 & TOSEM 2023]

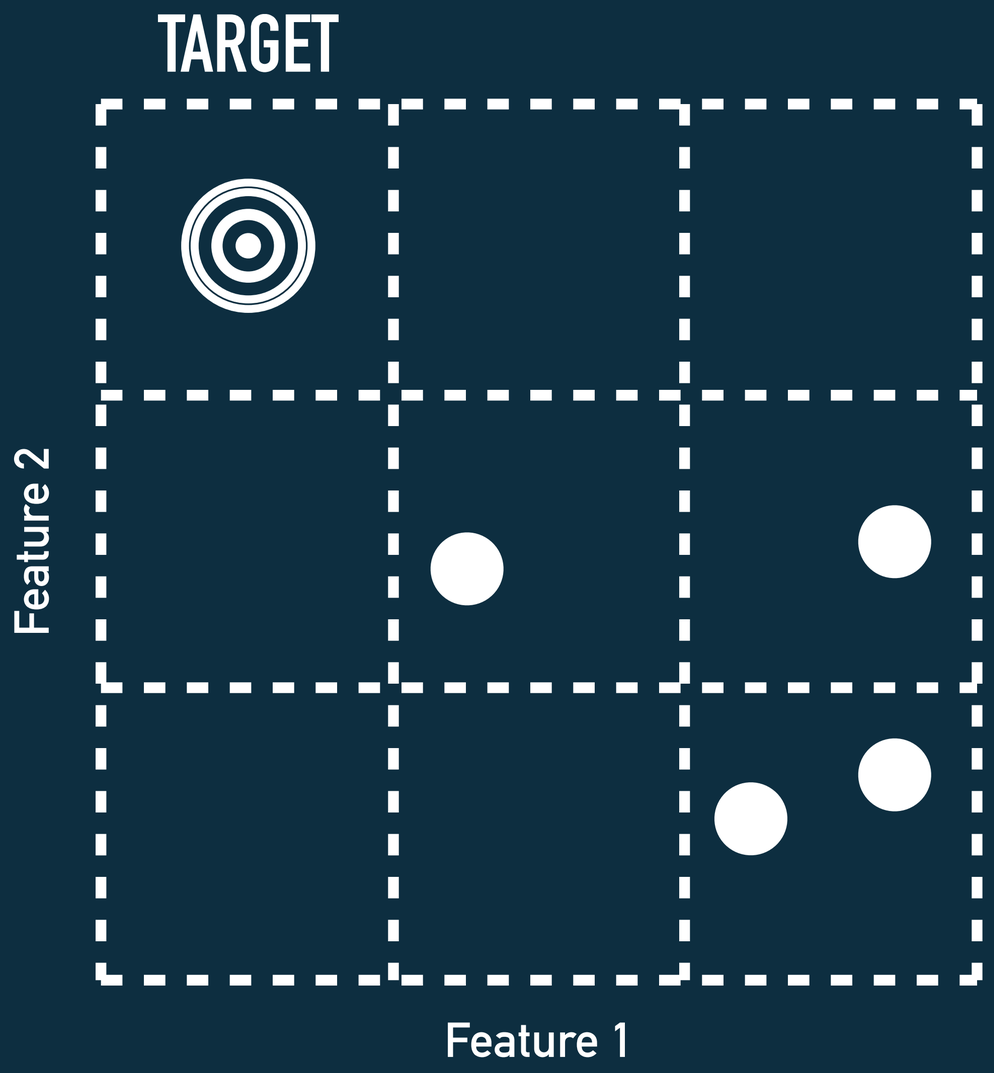


Misclassified

Correctly classified

Unknown

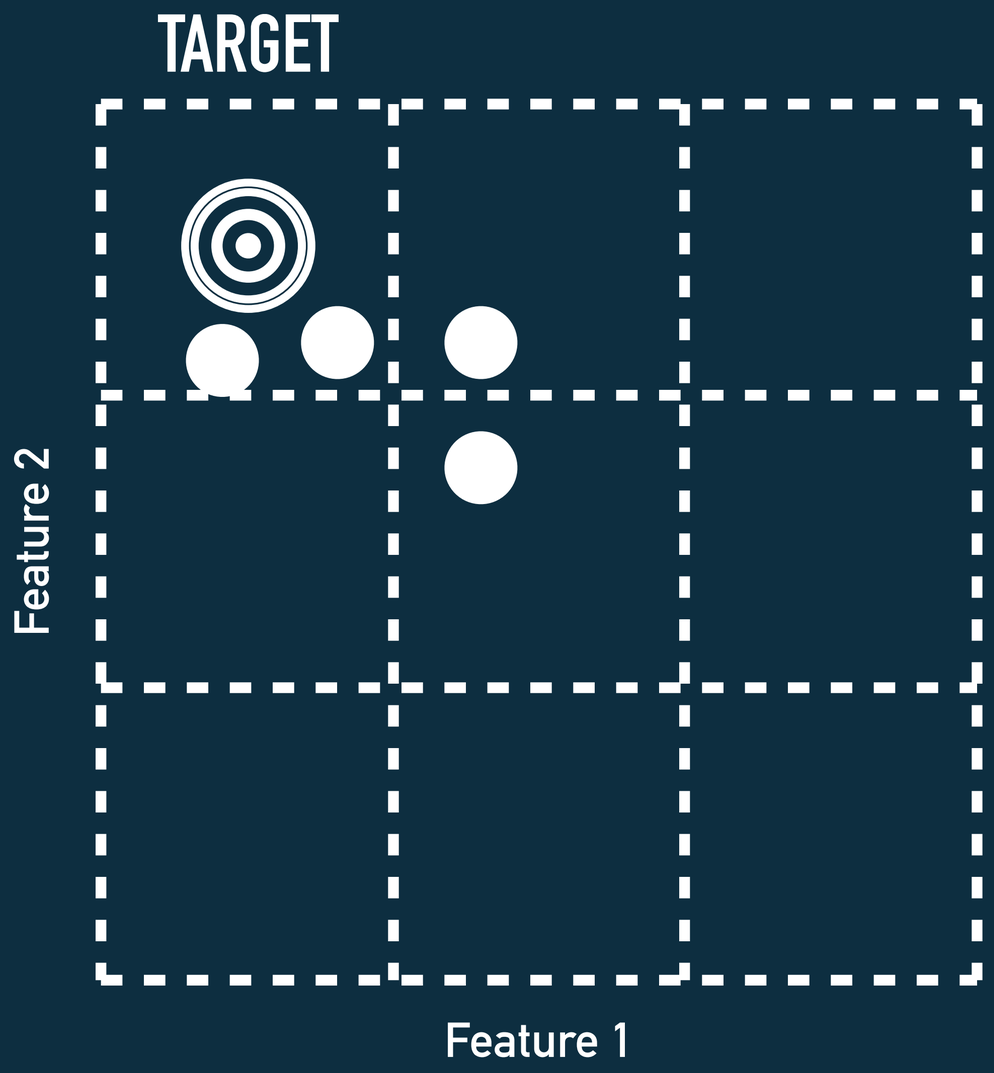
FOCUSED TEST GENERATION



Technique for generating inputs that are:

1. Close to the target
2. Misbehaviour-inducing
3. Diverse

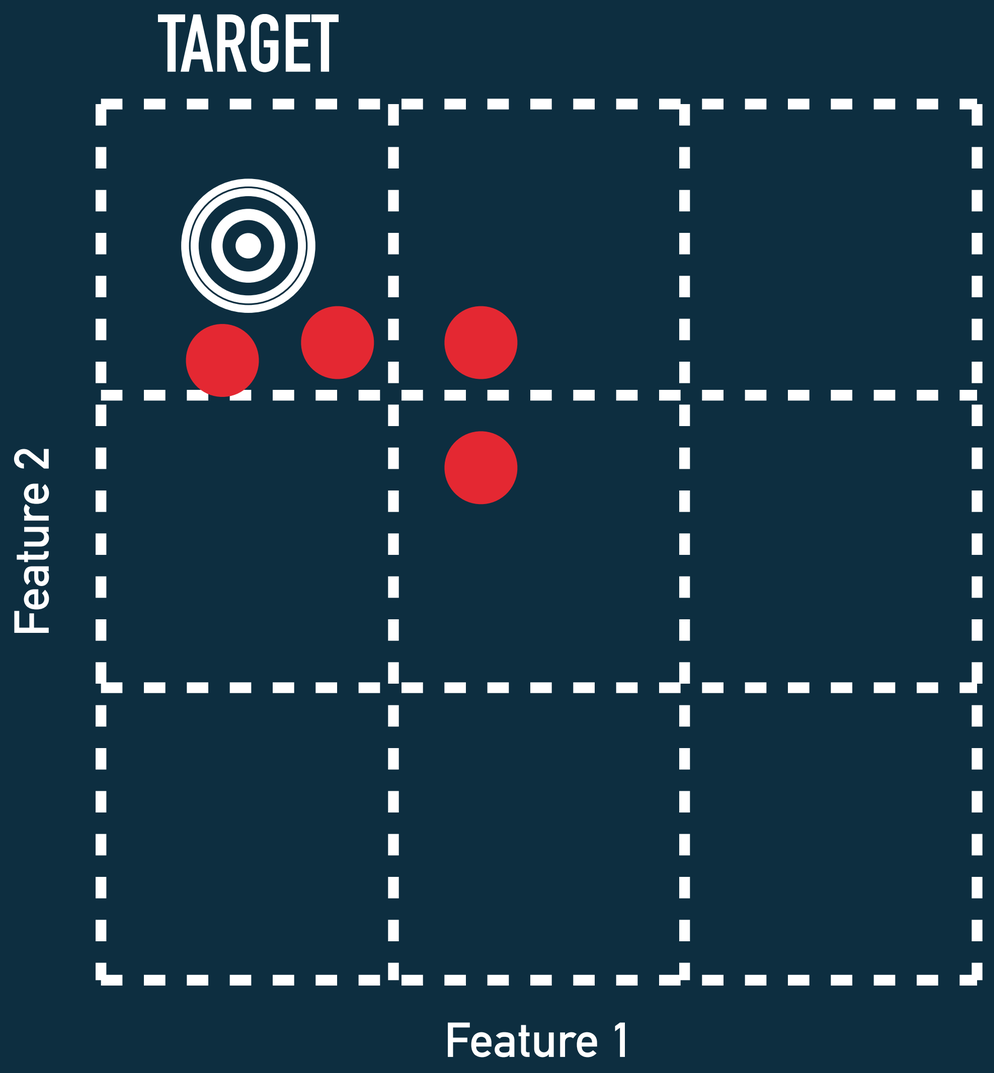
FOCUSED TEST GENERATION



Technique for generating inputs that are:

1. Close to the target
2. Misbehaviour-inducing
3. Diverse

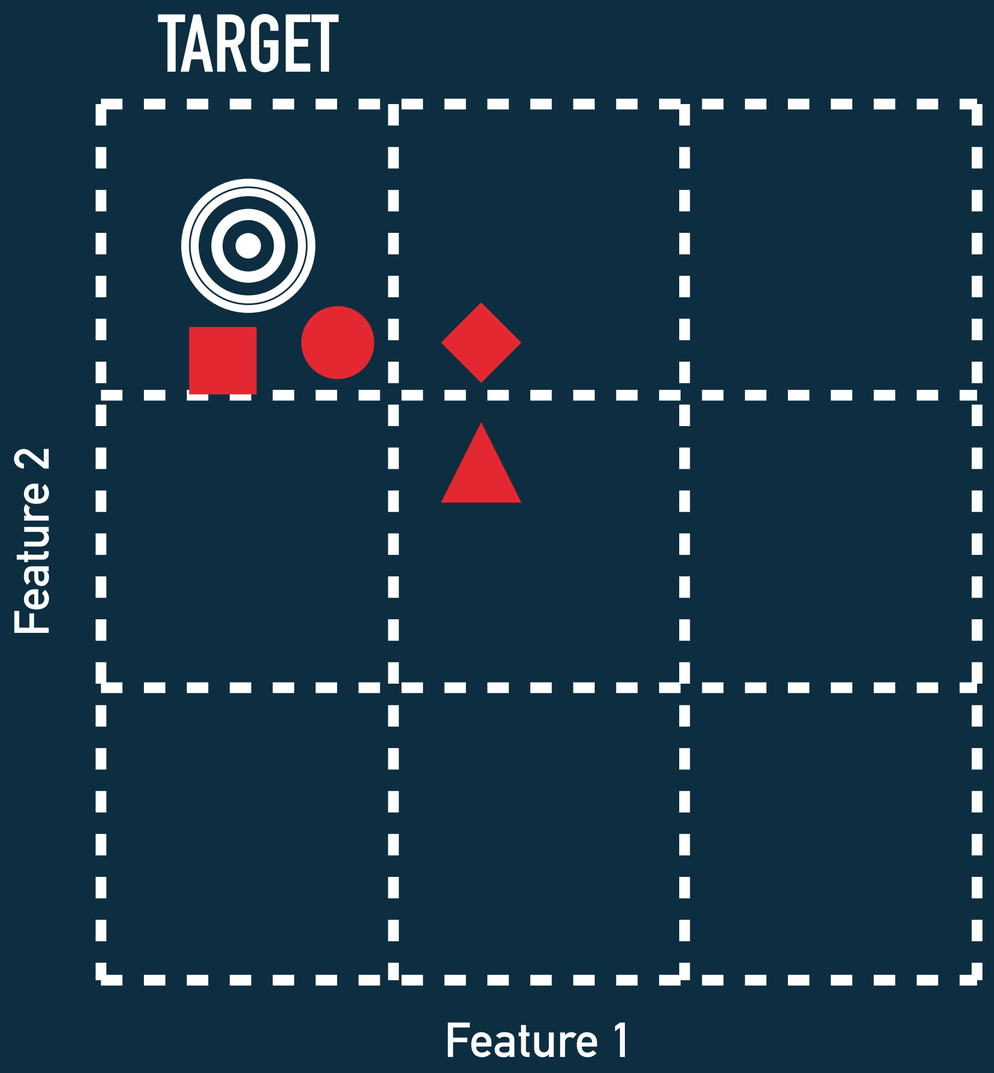
FOCUSED TEST GENERATION



Technique for generating inputs that are:

1. Close to the target
2. Misbehaviour-inducing
3. Diverse

FOCUSED TEST GENERATION

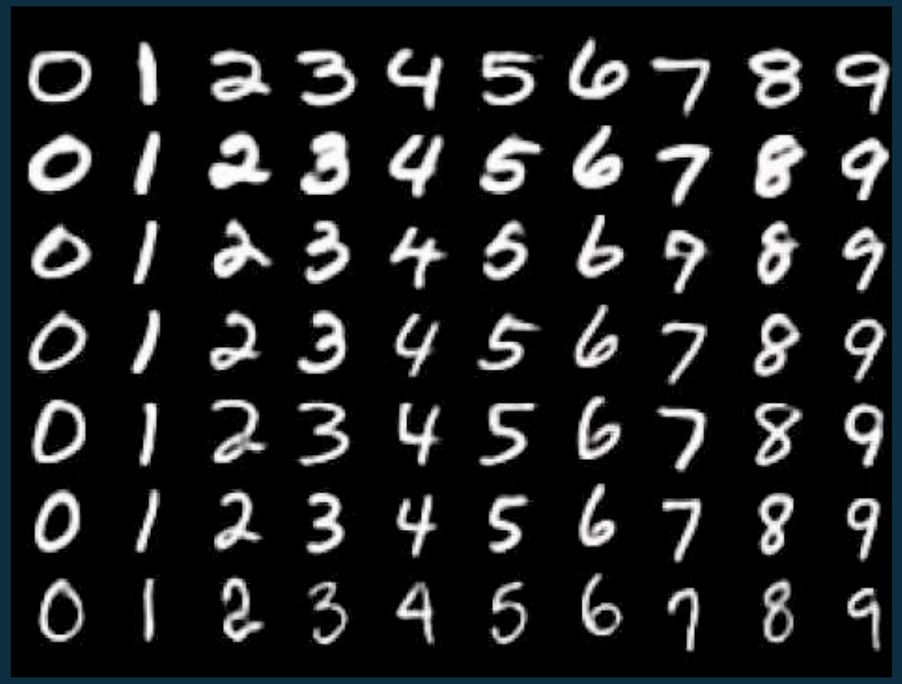


Technique for generating inputs that are:

1. Close to the target
2. Misbehaviour-inducing
3. Diverse

SUBJECT SYSTEMS

MNIST: Hand-written
Digit Recognition



Images

IMDB: Movie Review
Sentiment Analysis

Texts

DEEPATASH CONFIGURATIONS: SELECTION MECHANISM

NSGA-II

GA

FF1: MIN distance to the target

FF2: MIN closeness to misbehavior

FF3: MAX sparseness

Multi-Objective (NSGA-II):

Explicitly promotes diverse and misbehaviour-inducing inputs

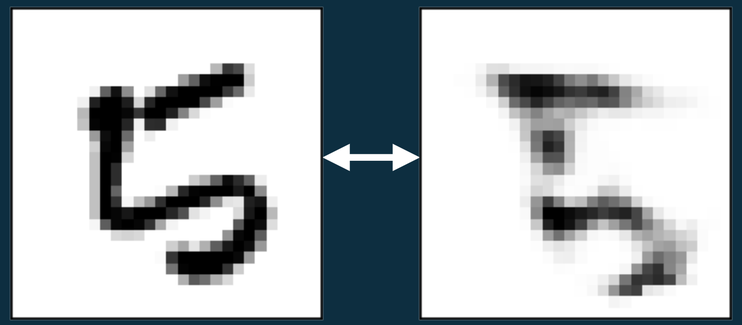
Single-Objective (GA):

Reduced overhead

DEEPATASH CONFIGURATIONS: SPARSENESS METRICS

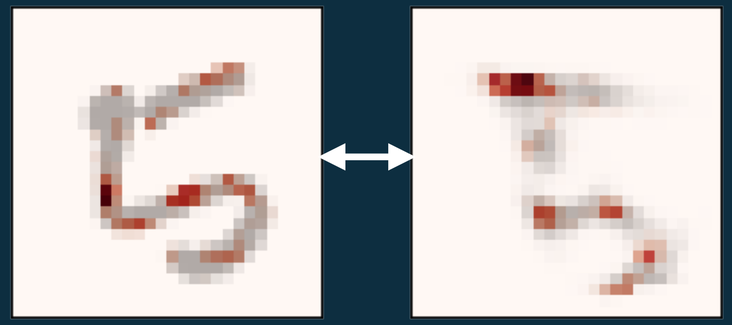
Input Space

Euclidean distance (Image)
Levenshtein distance (Text)



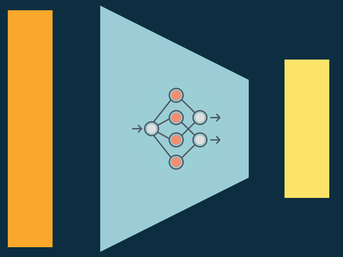
Explanation Space

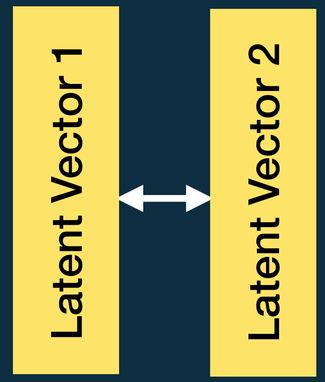
Integrated Gradients XAI technique



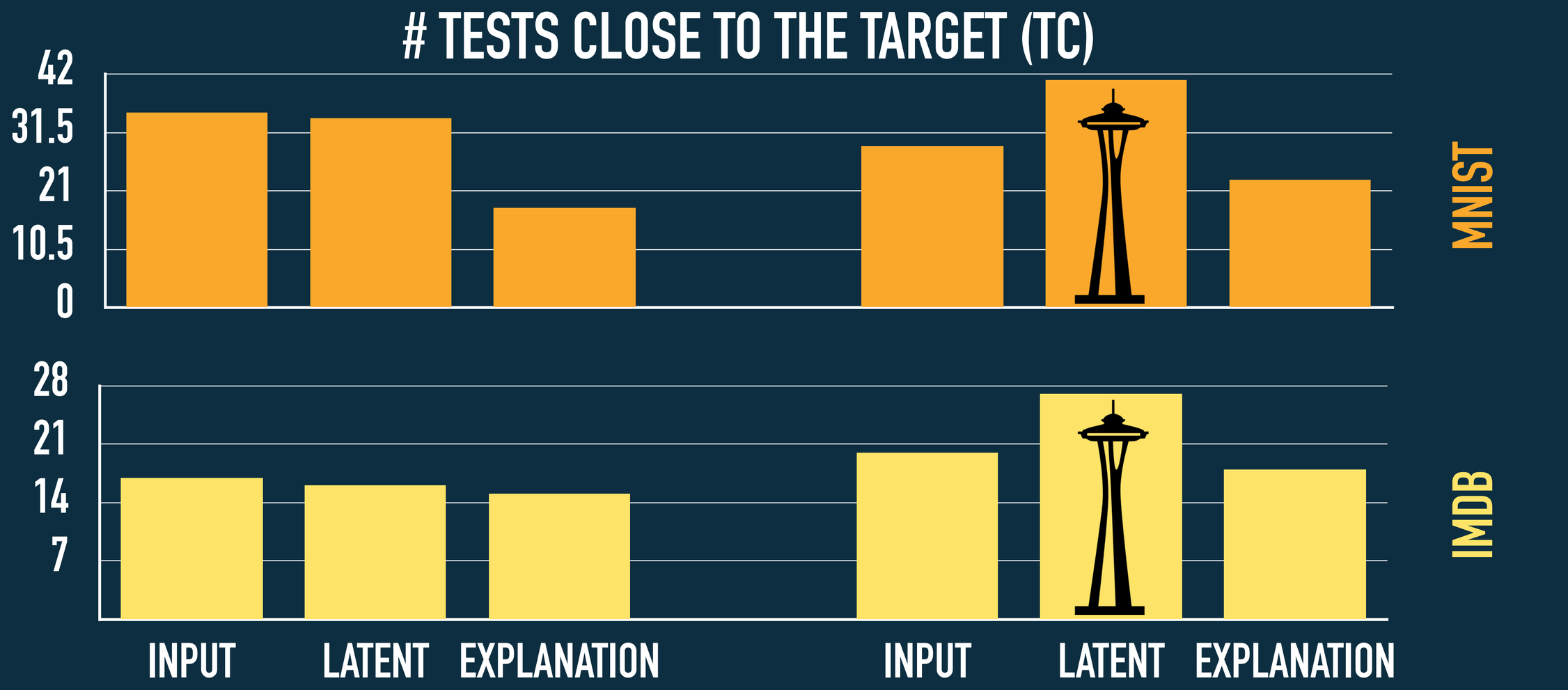
Latent Space

VAE (Image)
doc2vec (Text)

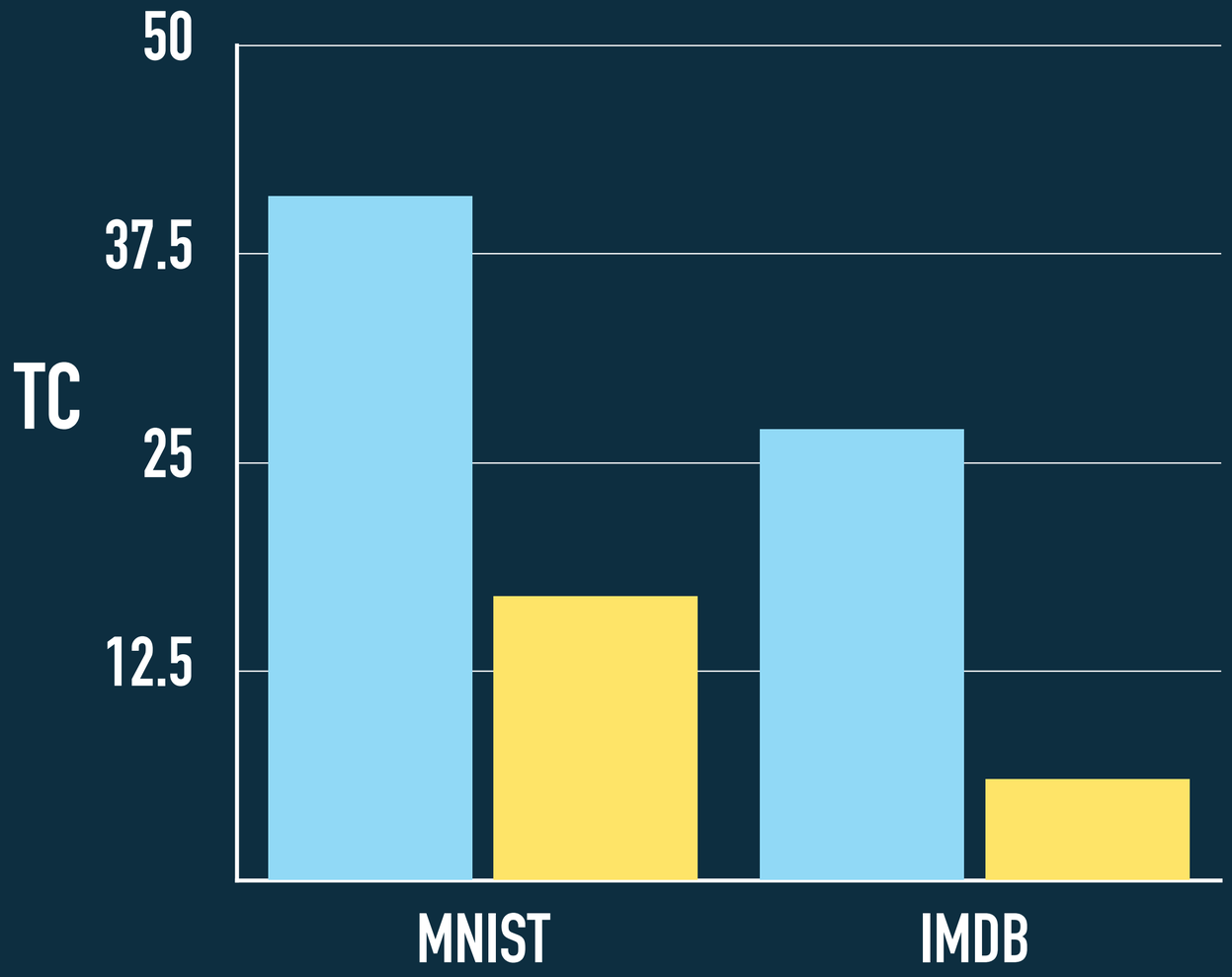




RQ1: BEST DEEPATASH CONFIGURATION



RQ2: COMPARISON WITH STATE OF THE ART



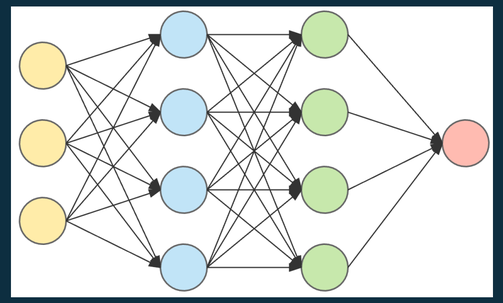
DeepAtash

Targets feature cell

DeepHyperion

Feature map exploration

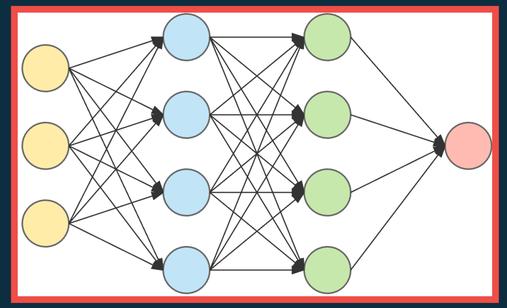
RQ3: FINE-TUNING DL SYSTEMS WITH DEEPATASH INPUTS



ORIGINAL DL SYSTEM

DEEPATASH TRAINING SET

FINE TUNING

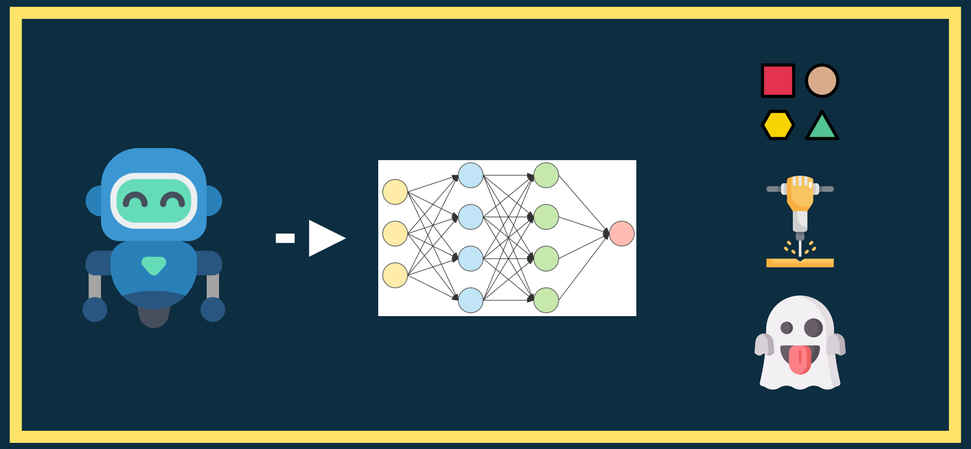


UPDATED DL SYSTEM

RQ3: FINE-TUNING DL SYSTEMS WITH DEEPATASH INPUTS

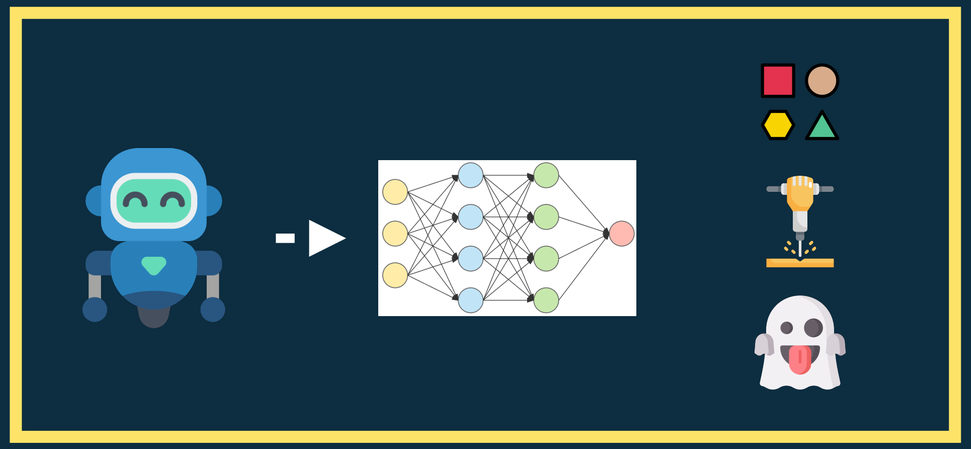
Subject	Original Test Set		DeepAtash Test Set	
	ACC before	ACC after	ACC before	ACC after
MNIST	99.11	-► 99.23	0.00	-► 99.55
IMDB	88.19	-► 89.57	0.00	-► 98.39

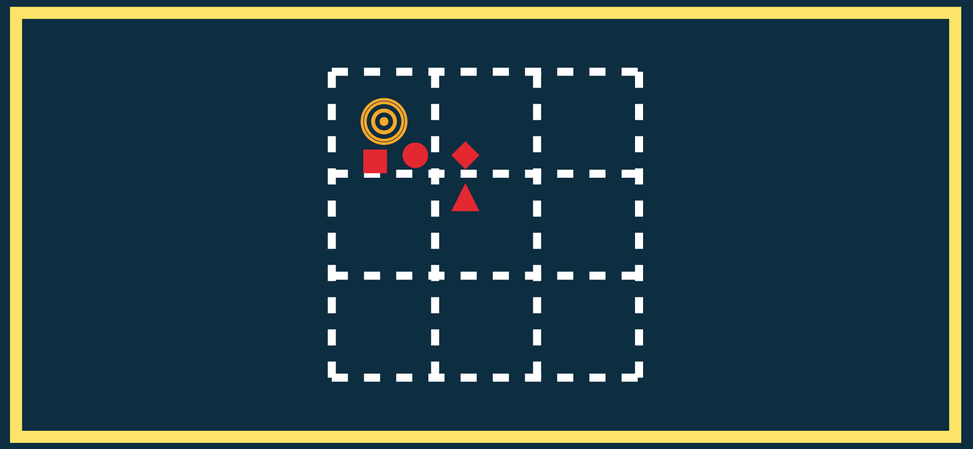
SUMMARY



Icons from www.flaticon.com

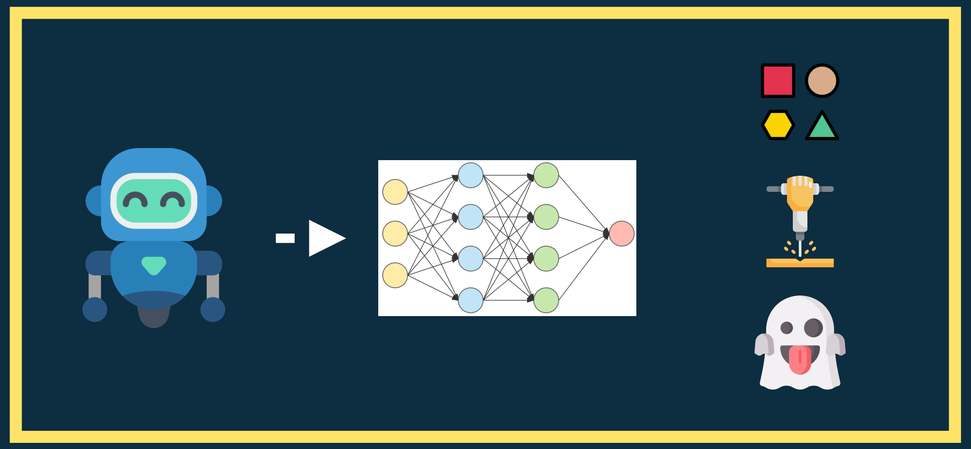
SUMMARY

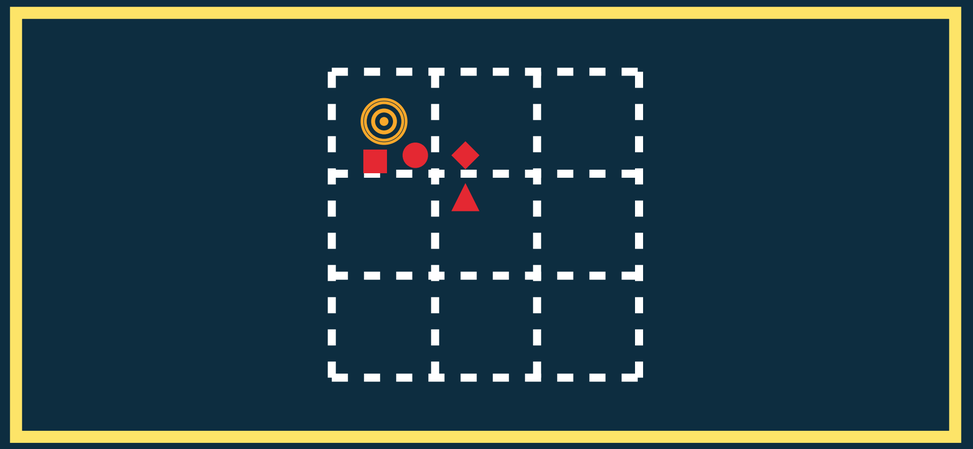


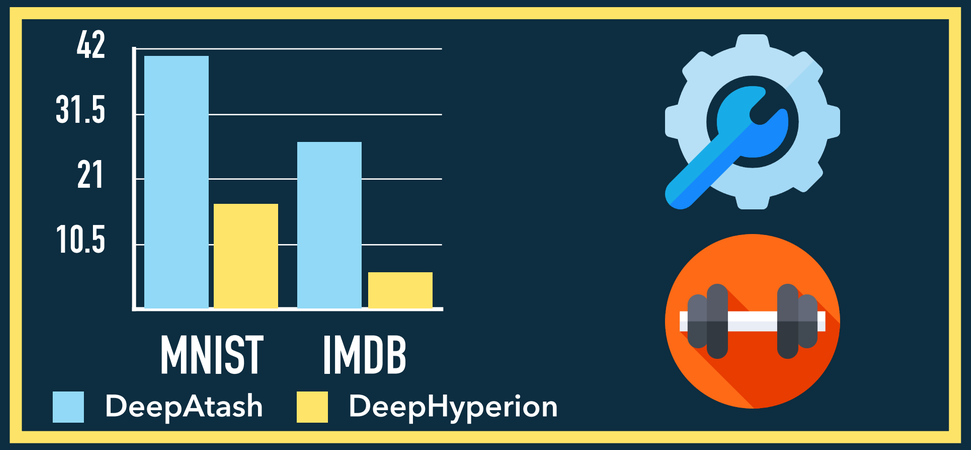


Icons from www.flaticon.com

SUMMARY

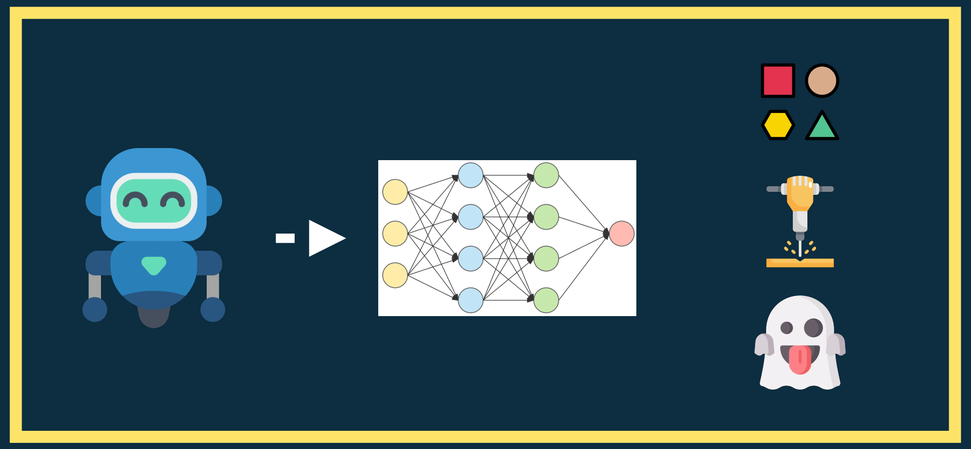


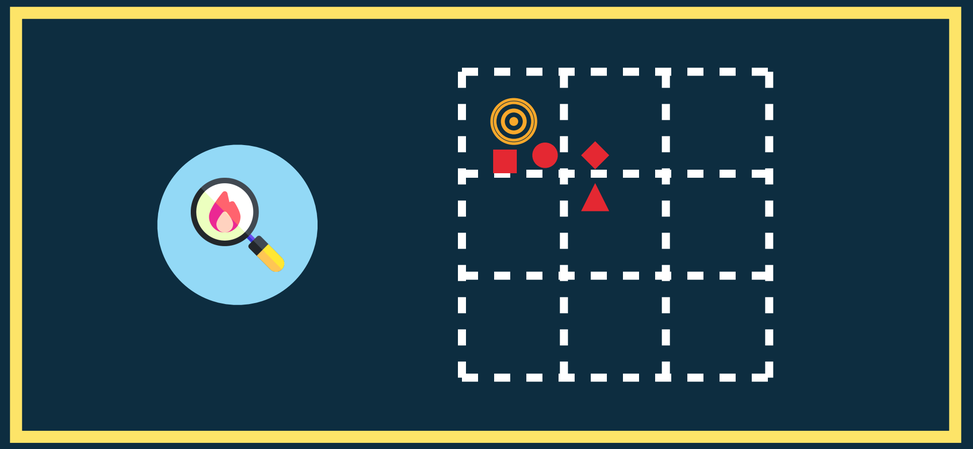


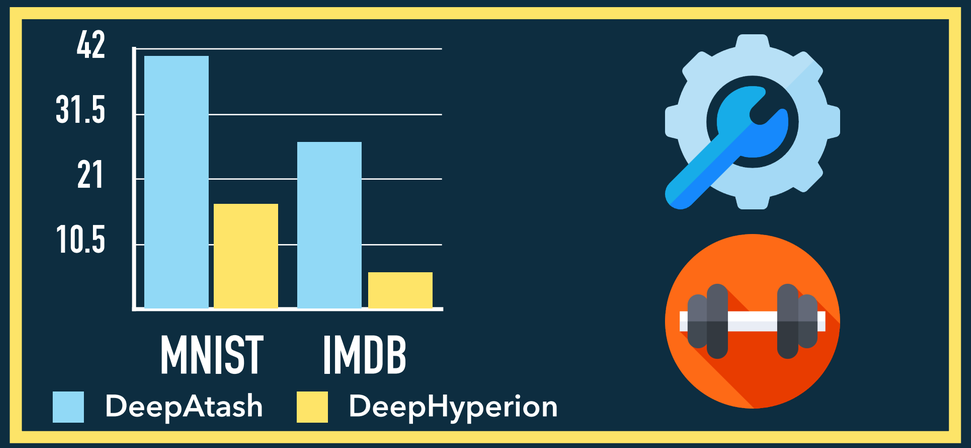


Icons from www.flaticon.com

SUMMARY







Icons from www.flaticon.com

EXTRA SLIDES

DIVERSITY COMPUTATION

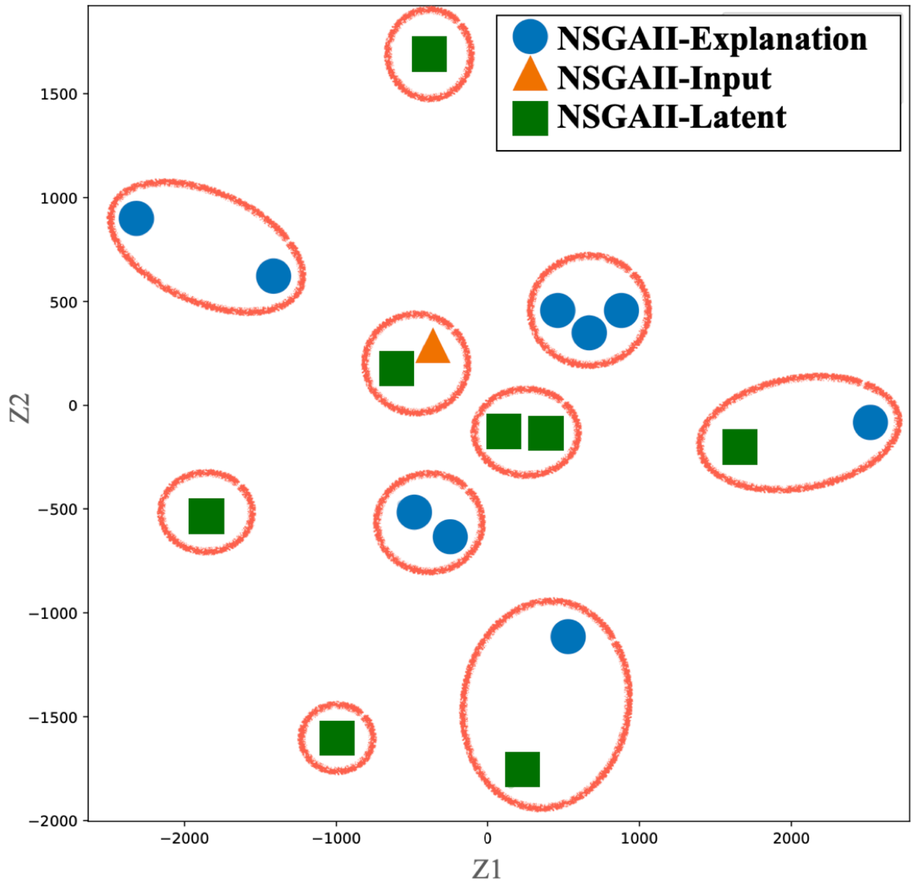


Figure 3: Example t-SNE plot to explain the computation of *test diversity* metrics, with clusters represented as empty circles containing inputs (smaller, solid shapes).

HYPERPARAMETERS

Table 1: Hyperparameters used in the experiments

Parameter	MNIST	IMDB
seed pool size	800	1000
population size	100	100
time budget (s)	3600	3600
mutation lower bound	0.01	-
mutation upper bound	0.6	-
sentimentDist	-	5
maxDist	-	5
repopulation upper bound	10	10
target archive size	81	42
number of epochs for retraining	6	5
learning rate for retraining	0.001	0.0001

RQ1 – RESULTS

Table 2: RQ1 - Tests close to target (TC), tests on target (TT), tests close to target diversity (TCD), and tests on target diversity (TTD) by alternative DeepAtash configurations for MNIST and IMDB. In each row, boldface indicates the maximum; underline indicates values statistically indistinguishable from the maximum.

			Input				Latent				Explanation			
			GA		NSGA-II		GA		NSGA-II		GA		NSGA-II	
		Features	TC [TCD]	TT [TTD]	TC [TCD]	TT [TTD]	TC [TCD]	TT [TTD]	TC [TCD]	TT [TTD]	TC [TCD]	TT [TTD]	TC [TCD]	TT [TTD]
MNIST	Dark	Mov-Lum	41.10 [**0.38**]	41.10 [0.48]	43.10 [0.36]	23.90 [0.12]	**51.70** [0.37]	**51.70** [0.47]	50.50 [0.31]	46.80 [**0.58**]	16.40 [0.24]	1.00 [0.05]	33.60 [0.28]	1.80 [0.25]
		Mov-Or	**69.10** [0.28]	4.40 [0.10]	58.60 [0.22]	1.60 [0.15]	66.10 [**0.30**]	**10.20** [**0.45**]	55.30 [0.22]	1.20 [0.13]	28.50 [0.14]	0.00 [0.00]	26.70 [0.12]	3.00 [0.25]
		Or-Lum	**70.30** [0.27]	24.20 [0.65]	32.10 [0.14]	5.90 [0.35]	65.50 [**0.31**]	**25.90** [**0.70**]	64.30 [0.22]	15.60 [0.50]	55.30 [0.21]	8.60 [0.55]	42.80 [0.18]	5.80 [0.45]
	Grey	Mov-Lum	41.40 [0.60]	**38.30** [**0.73**]	21.80 [0.28]	0.00 [0.00]	22.10 [0.38]	16.20 [0.35]	**53.40** [**0.78**]	28.40 [0.40]	5.10 [0.25]	0.60 [0.08]	13.40 [0.49]	0.40 [0.10]
		Mov-Or	18.50 [0.45]	3.00 [0.20]	15.30 [0.43]	2.40 [0.20]	16.00 [0.24]	0.70 [0.11]	**20.50** [**0.53**]	**4.70** [**0.28**]	13.10 [0.30]	1.80 [0.15]	14.90 [0.49]	0.40 [0.10]
		Or-Lum	10.10 [0.29]	**10.10** [0.34]	22.90 [0.47]	8.60 [0.27]	9.20 [0.23]	9.20 [0.30]	19.20 [0.52]	6.80 [0.19]	6.40 [0.29]	3.50 [0.24]	**28.10** [**0.56**]	6.20 [**0.55**]
	White	Mov-Lum	14.30 [0.36]	**11.60** [0.28]	28.20 [0.61]	2.30 [0.30]	20.60 [0.42]	10.70 [**0.36**]	**29.60** [0.54]	10.40 [0.25]	10.90 [0.38]	5.10 [0.15]	15.40 [0.60]	6.20 [0.55]
		Mov-Or	24.60 [0.44]	2.00 [0.21]	11.30 [0.31]	**7.70** [0.10]	22.10 [0.50]	5.90 [**0.35**]	**25.10** [**0.65**]	1.80 [0.18]	6.70 [0.43]	0.00 [0.00]	7.30 [0.32]	0.00 [0.00]
		Or-Lum	23.30 [0.48]	21.60 [0.58]	30.20 [0.52]	10.10 [0.38]	28.70 [0.51]	24.30 [**0.71**]	**51.00** [**0.66**]	**28.00** [0.65]	21.50 [0.51]	6.80 [0.48]	20.80 [0.52]	4.30 [0.48]
		AVG	34.74 [0.39]	**17.37** [0.40]	29.28 [0.37]	6.94 [0.21]	33.56 [0.36]	17.20 [**0.42**]	**40.99** [**0.49**]	15.97 [0.35]	18.21 [0.31]	3.00 [0.19]	22.56 [0.40]	2.68 [0.26]
IMDB	Dark	Neg-Pos	33.00 [0.68]	22.00 [0.5]	29.40 [0.53]	6.90 [0.34]	25.80 [0.58]	14.30 [0.53]	40.30 [0.74]	33.40 [0.64]	25.70 [0.49]	7.8 [0.31]	29.90 [0.51]	9.60 [0.37]
		Neg-Verb	7.20 [0.31]	3.20 [0.12]	9.20 [0.36]	6.30 [0.16]	5.00 [0.16]	0.70 [0.10]	27.10 [0.63]	14.60 [0.34]	11.20 [0.53]	3.60 [0.15]	19.70 [0.54]	6.60 [0.17]
		Pos-Verb	33.80 [0.53]	33.80 [0.43]	31.60 [0.45]	31.60 [0.47]	33.70 [0.50]	33.40 [0.58]	37.60 [0.52]	37.60 [0.57]	28.80 [0.50]	27.80 [0.38]	28.30 [0.47]	6.60 [0.17]
	Grey	Neg-Pos	**7.50** [0.30]	5.20 [0.35]	6.40 [0.22]	3.50 [0.09]	5.00 [0.25]	3.80 [0.06]	**10.80** [**0.52**]	**8.30** [**0.35**]	0.80 [0.05]	0.00 [0.00]	5.60 [0.34]	0.10 [0.00]
		Neg-Verb	7.30 [0.45]	7.30 [0.41]	15.00 [0.57]	**14.90** [0.61]	10.70 [0.51]	10.70 [0.49]	**15.70** [**0.63**]	13.10 [**0.62**]	9.50 [0.54]	8.60 [0.54]	12.20 [0.45]	11.20 [0.55]
		Pos-Verb	27.10 [0.50]	27.10 [0.60]	28.70 [0.49]	28.70 [0.49]	24.10 [0.62]	27.30 [0.41]	**32.00** [0.53]	**32.00** [0.56]	27.30 [0.41]	27.30 [**0.68**]	25.50 [0.62]	25.50 [0.58]
	White	Neg-Pos	24.20 [0.65]	15.40 [**0.65**]	31.10 [0.62]	22.40 [0.63]	29.40 [0.64]	18.00 [0.53]	**38.90** [0.60]	**29.20** [0.63]	24.30 [0.52]	20.90 [0.58]	26.40 [**0.67**]	14.20 [0.58]
		Neg-Verb	4.10 [0.26]	1.80 [0.10]	0.00 [0.00]	0.00 [0.00]	5.60 [**0.52**]	0.00 [0.00]	**13.00** [0.37]	**3.60** [**0.25**]	0.70 [0.06]	0.00 [0.00]	0.40 [0.02]	0.00 [0.00]
		Pos-Verb	6.10 [0.13]	2.30 [0.10]	25.10 [**0.60**]	**9.40** [0.38]	3.00 [0.10]	1.70 [0.07]	**25.70** [0.48]	3.80 [0.19]	8.40 [0.15]	0.00 [0.00]	15.90 [0.53]	1.30 [0.07]
		AVG	16.70 [0.42]	13.12 [0.36]	19.60 [0.43]	13.70 [0.35]	15.81 [0.43]	11.86 [0.33]	**26.80** [**0.56**]	**19.51** [**0.46**]	19.60 [0.43]	13.70 [0.35]	18.21 [0.46]	10.80 [0.30]

RQ2 – RESULTS

Table 3: RQ2 - Results achieved by the compared tools for MNIST and IMDB. Tests close to target (TC) and their diversity (TCD); tests on target (TT) and their diversity (TTD). In each row, boldface is the maximum; underline indicates values statistically indistinguishable from the maximum.

			DeepAtash		DeepHyperion	
		Features	TC [TCD]	TT [TTD]	TC [TCD]	TT [TTD]
MNIST	Dark	Mov-Lum	**50.50** [**0.90**]	**46.80** [**0.97**]	18.30 [0.38]	2.30 [0.07]
		Mov-Or	**55.30** [**0.86**]	1.20 [0.27]	37.60 [0.47]	**2.40** [**0.42**]
		Or-Lum	**64.30** [**0.95**]	**15.60** [**0.74**]	13.80 [0.30]	2.70 [0.15]
	Grey	Mov-Lum	**53.40** [**0.81**]	**28.40** [**0.50**]	22.70 [0.50]	3.70 [0.10]
		Mov-Or	20.50 [**0.88**]	**4.70** [**0.45**]	**24.70** [0.45]	1.10 [0.15]
		Or-Lum	**19.20** [**0.81**]	**6.80** [**0.39**]	2.20 [0.19]	0.10 [0.01]
	White	Mov-Lum	**29.60** [**0.74**]	**10.40** [**0.40**]	11.90 [0.42]	0.00 [0.00]
		Mov-Or	**25.10** [**0.71**]	**1.70** [**0.20**]	20.70 [0.50]	0.00 [0.00]
		Or-Lum	**51.00** [**1.00**]	**28.00**[**1.00**]	0.80 [0.05]	0.00 [0.00]
		AVG	**40.99** [**0.85**]	**15.96** [**0.55**]	16.97 [0.36]	1.37 [0.10]
IMDB	Dark	Neg-Pos	**40.30** [**0.94**]	**33.40** [**1.00**]	8.20 [0.11]	1.60 [0.05]
		Neg-Verb	**27.10** [**1.00**]	**14.60** [**0.43**]	10.80 [0.05]	4.50 [0.07]
		Pos-Verb	**32.00** [**0.95**]	**32.00** [**1.00**]	2.40 [0.05]	1.10 [0.05]
	Grey	Neg-Pos	**10.80** [**0.73**]	**8.30** [**0.40**]	10.20 [0.20]	1.00 [0.20]
		Neg-Verb	**15.70** [**0.95**]	**13.10**[**0.93**]	7.70 [0.11]	1.80 [0.08]
		Pos-Verb	**37.60** [**0.95**]	**37.60** [**0.95**]	12.00 [0.15]	5.20 [0.11]
	White	Neg-Pos	**38.90** [**1.00**]	**29.20** [**1.00**]	0.20 [0.00]	0.00 [0.00]
		Neg-Verb	**13.00** [**0.50**]	**3.60** [**0.30**]	0.30 [0.10]	0.00 [0.00]
		Pos-Verb	**25.70** [**0.70**]	**3.50** [**0.30**]	0.70 [0.10]	0.00 [0.00]
		AVG	**26.79** [**0.86**]	**19.48** [**0.70**]	5.83 [0.10]	1.69 [0.06]

RQ3 – RESULTS

Table 4: RQ3 - Model Accuracy (ACC) on the original test set and on the test set generated by DeepAtash, before and after fine tuning the DL system with the training partition of generated inputs. In each row, boldface indicates the maximum; underline indicates values statistically significant.

		Original Test Set		DA Test Set	
	Features	ACC before	ACC after	ACC before	ACC after
MNIST	Mov-Lum	99.11	**99.23**	0.00	**99.92**
	Mov-Or		**99.24**		**99.65**
	Or-Lum		**99.23**		**99.02**
IMDB	Neg-Pos	88.19	**89.58**	0.00	**98.36**
	Neg-Verb		**89.56**		**99.47**
	Pos-Verb		**89.56**		**97.35**