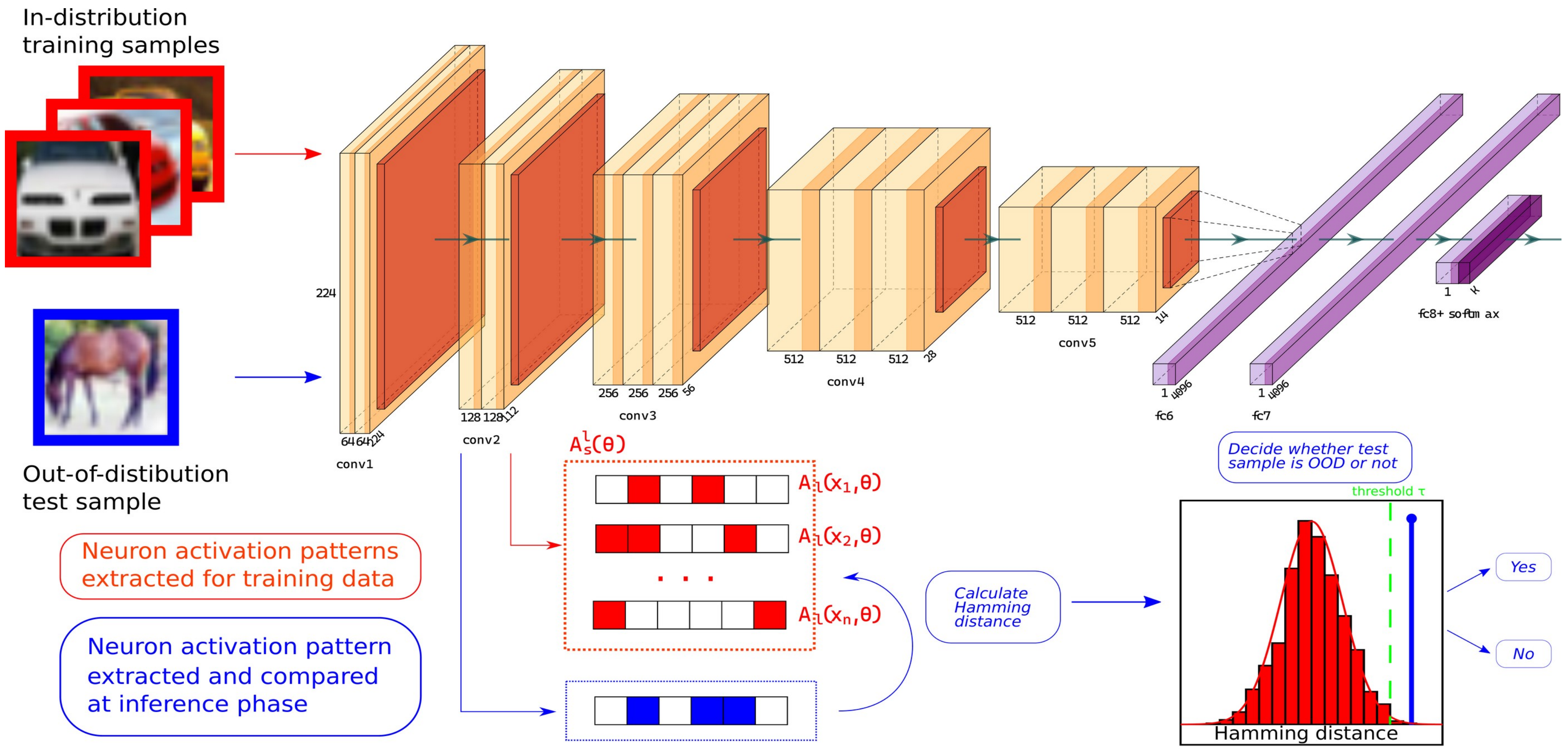
Detection of out-of-distribution samples using binary neuron activation patterns

TUE-AM-321

Bartłomiej Olber[1, 2], Krystian Radlak[1, 2], Adam Popowicz[2], Michal Szczepankiewicz[3], Krystian Chachuła[2]
[1]Warsaw University of Technology [2]Silesian University of Technology [3]NVIDIA

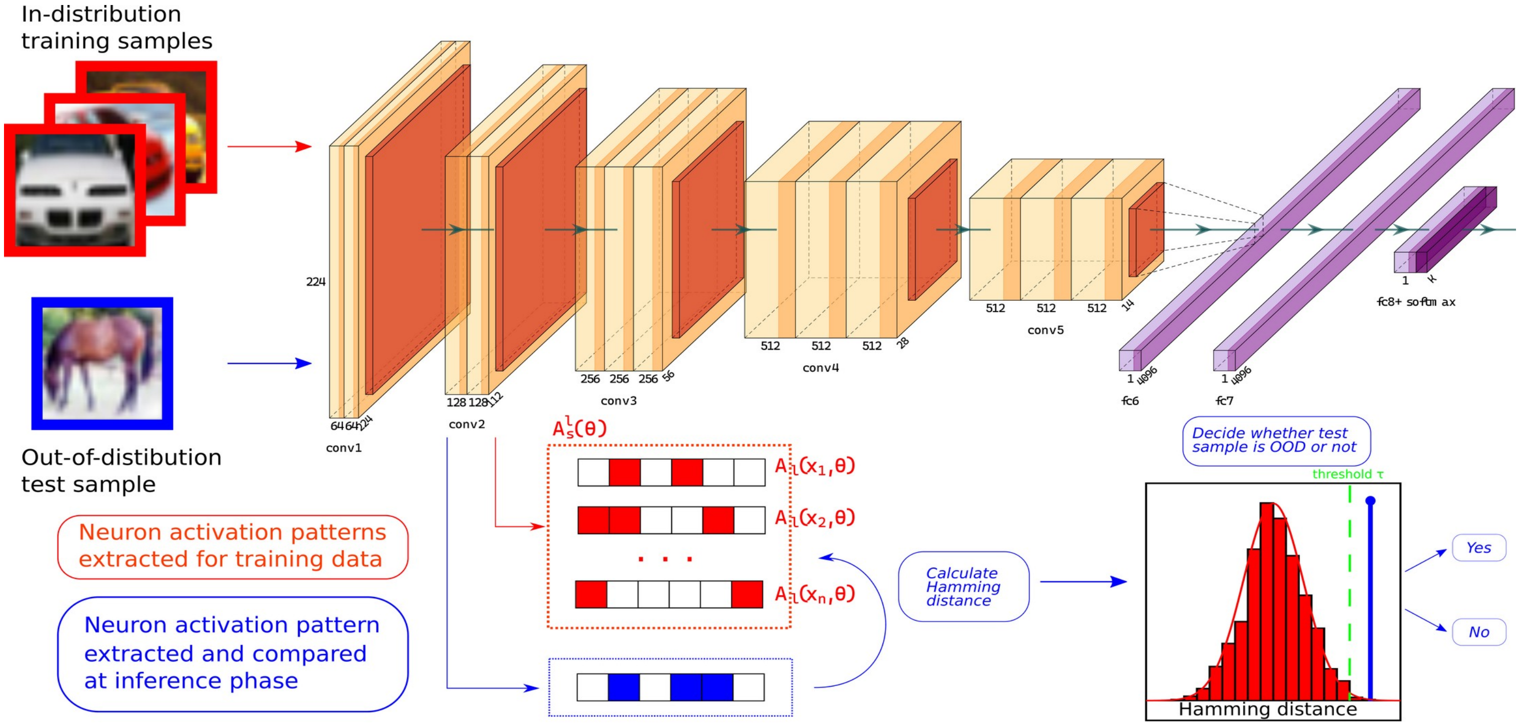
https://safednn.com/

Detection of out-of-distribution samples using binary neuron activation patterns



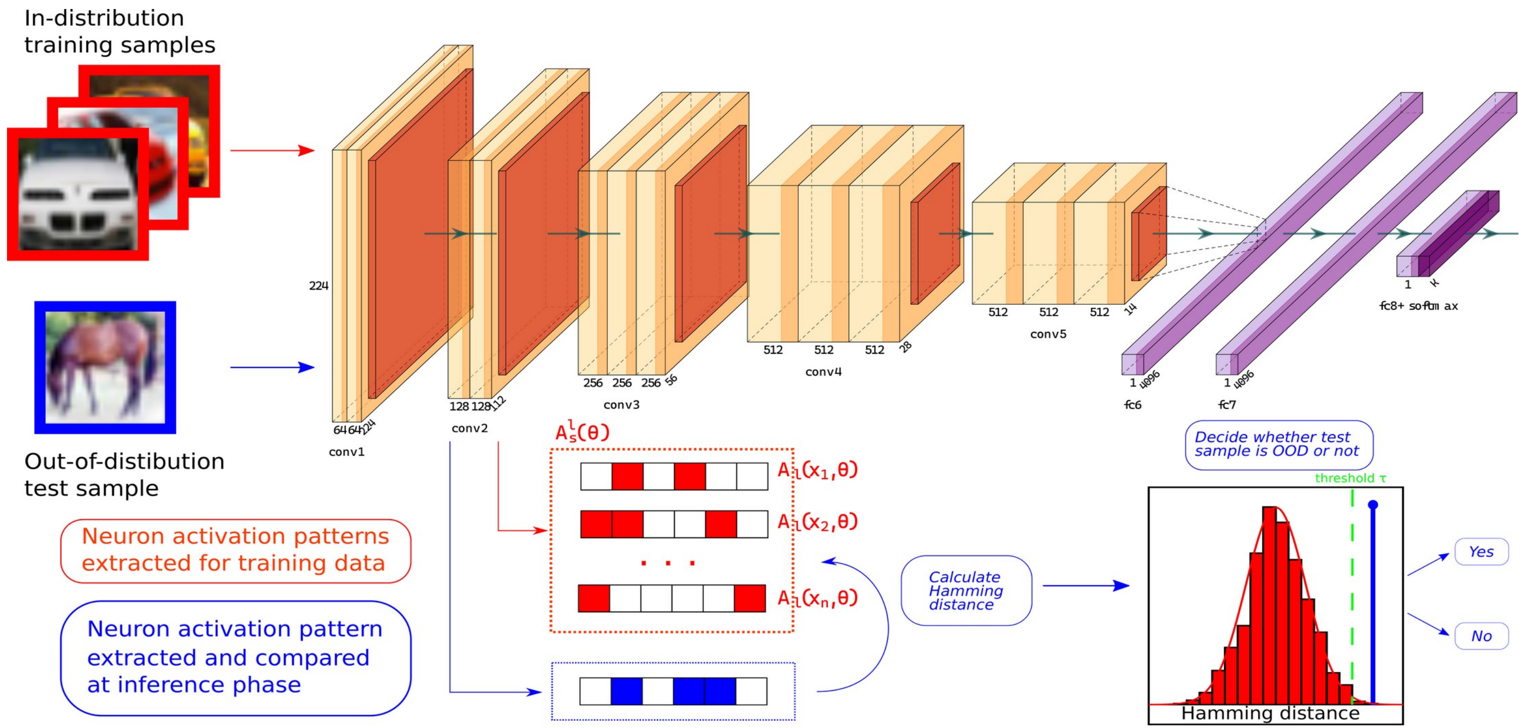
OOD detection with NAP – detector setup (1)

- For each training image run forward pass through the classifier, extracting binary activation patterns from ReLU-activated layers.
- Save extracted patterns in data structures enabling a fast nearest neighbor search, e.g., the ball tree.. For each layer and image label use separate data structure.



OOD detection with NAP – test phase (2)

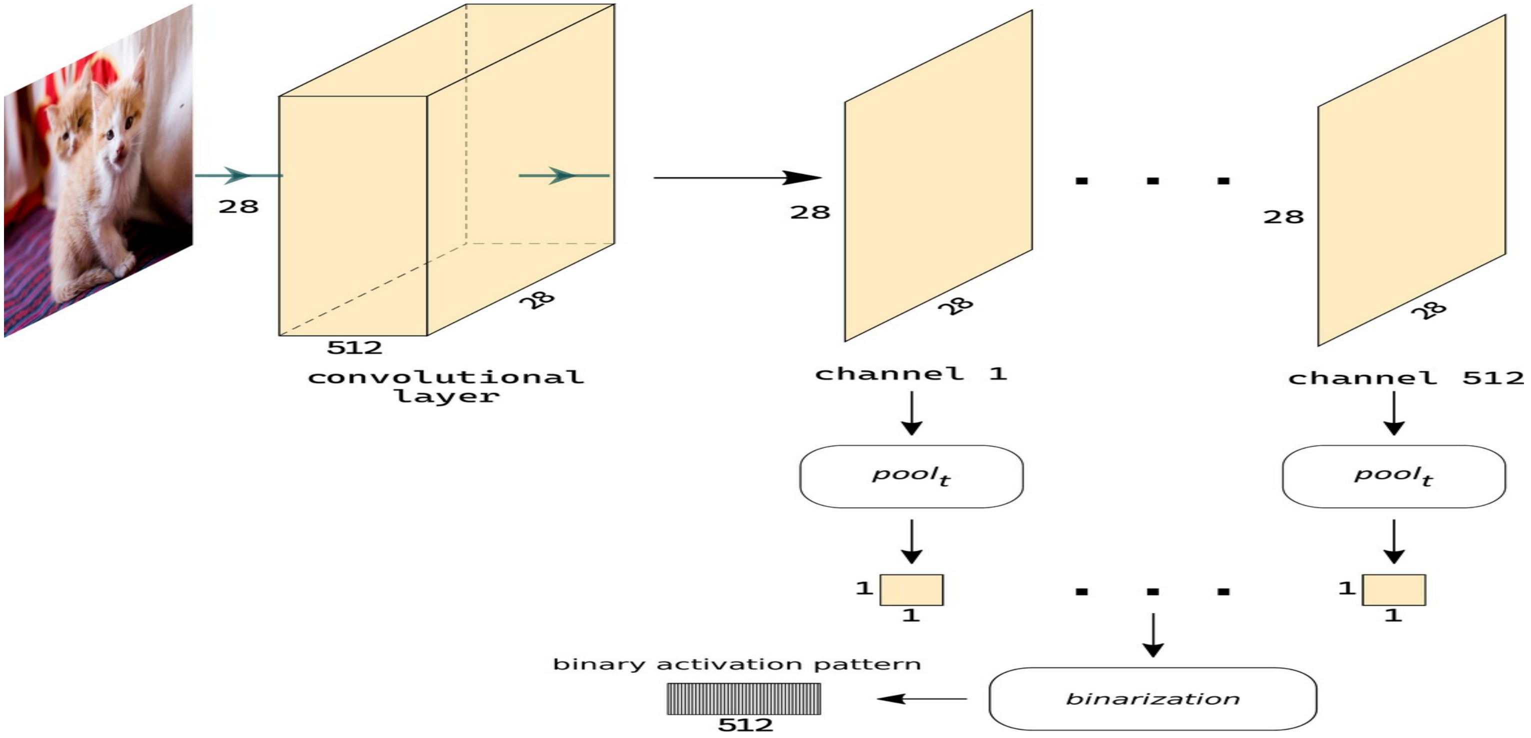
- Run classification and extraction of patterns for the test image.
- For each layer independently, compute Hamming distance to the most similar known pattern (out of all known patterns from given layer and the test image's predicted label) .
- Combine the uncertainty estimations from multiple layers into a single score.
- Layers contributing to the final score are selected by grid search using validation OOD dataset.



OOD detection with NAP – extracting patterns

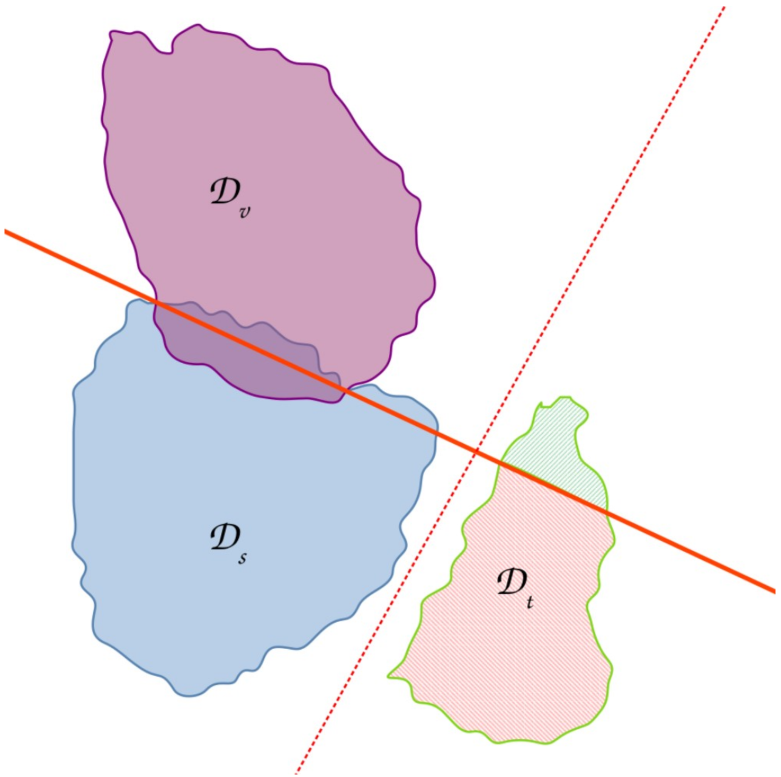
- Pooling – channel mean or maximum
- Extracting from linear layers requires solely binarization

- Binarization
  - Zero p-% smallest activations in the vector
  - Replace all remaining positive values with ‘1’
  - Cast the vector into memory efficient boolean representation



A less biased OOD detectors evaluation scheme - OD-test [1]

- **Datasets:** MNIST, FashionMNIST, CIFAR10, CIFAR100, STL10, TinyImagenet, NotMNIST, normal noise, uniform noise
- **Metrics:** AUROC, accuracy
- **Network architectures:** VGG, ResNet
- **SOTA algorithms:** our method was compared with 17 state-of-the-art OOD detectors
- **Evaluation protocol:** All datasets were used in all combinations as D_s, D_v and D_t. Each method's performance is a mean of 308 evaluations

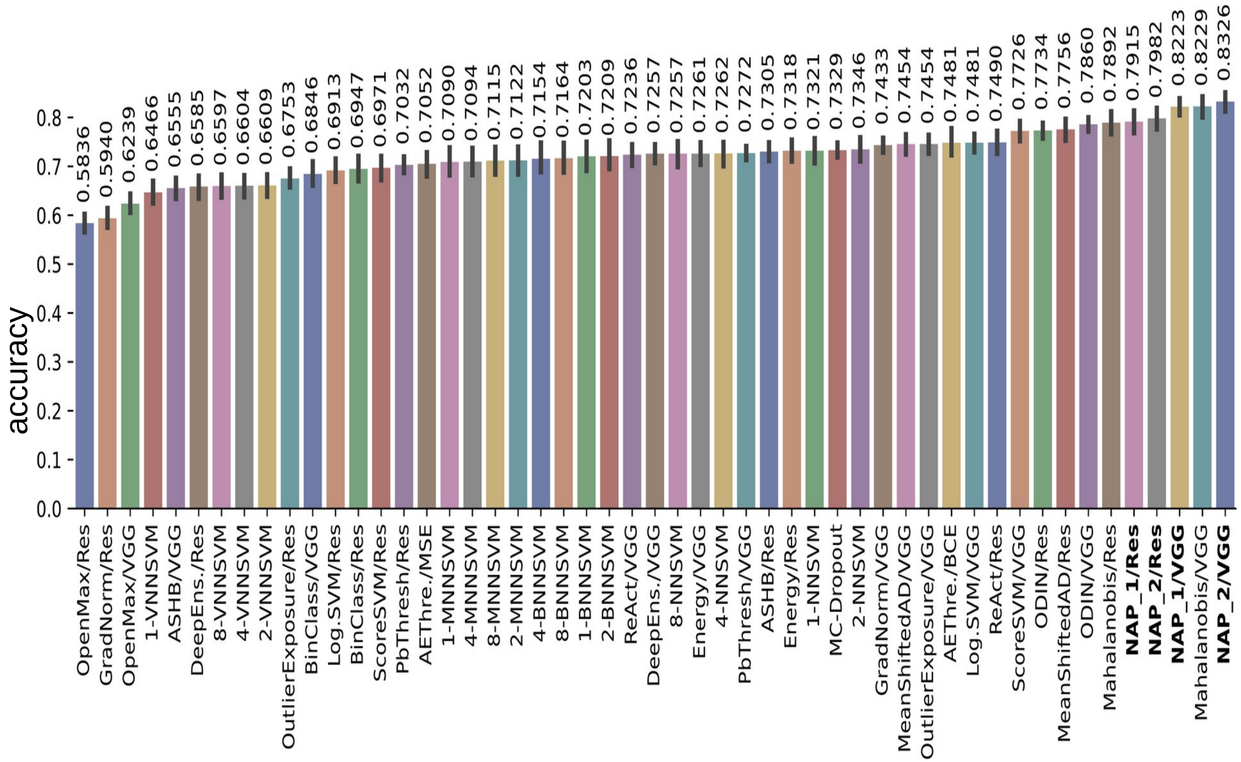


- **D_s** - training (source) distribution
- **D_v** - validation out-of-distribution
- **D_t** - test out-of-distribution

[1] Alireza Shafaei, Mark Schmidt, and James J. Little. A less biased evaluation of out-of-distribution sample detectors.

Results

- **ResNet**
  - The best AUROC improved by 0.1%
  - The best accuracy improved by 1.1%
- **VGG**
  - The best AUROC improved by 3.4%
  - The best accuracy improved by 1.2%
- Reasonably good time and memory efficiency



Method	Time [s]	
	VGG	ResNet
ASH-B	0.004	0.009
BinClass	0.002	0.007
DeepEns.	0.012	0.038
Energy	0.003	0.008
GradNorm	0.010	0.027
Log.SVM	0.003	0.006
MC-Dropout	0.003	0.007
Mahalanobis	0.217	0.086
MeanShiftedAD	0.231	0.314
ODIN	0.010	0.028
OpenMax	0.035	0.040
OutlierExposure	0.003	0.008
PbThresh	0.002	0.007
ReAct	0.003	0.008
ScoreSVM	0.002	0.006
Ours (NAP)	**0.008**	**0.015**

Summary

- We introduced a novel OOD detector that uses Neuron Activation Patterns.
- We published the largest evaluation benchmark for OOD detection consisting of 18 OOD methods.