Sample-free white-box out-of-distribution detection for deep learning

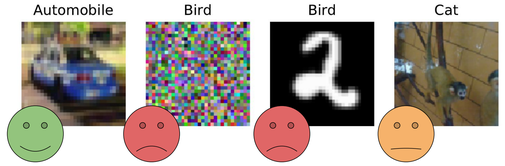
Jean-Michel Begon - Pierre Geurts

CVPR VIRTUAL JUNE 19-25

Out-of-distribution (OOD) ? Sample-free ?

Out-of-distribution

- Detecting irrelevant inputs



Sample-free

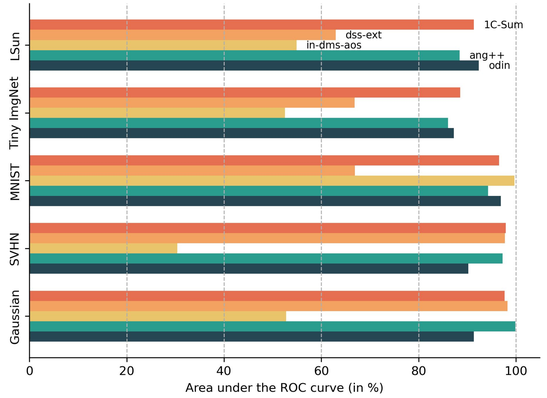
- The model is already deployed
- Not built with security in mind
- Training set is no longer accessible
- Is it possible to leverage what the model has captured from the “in”-distribution (ID) ?

How ?

- Indicators based on
  - First order optimality condition
    - How the network should behave on a normal, ID sample
    - e.g. high confidence prediction
  - Batch-normalization
    - e.g. based on the estimated batchnorm parameters
- Indicator: higher value, more likely OOD
- Ideally fast to compute
- Target image classification

Does it work?

Yes…



… sort of

- Some problems are easy, other are not
  - Hard without data, tacklable with some
- Best indicators vary
  - with the out-of-distribution (mostly)
  - with the architecture (a bit)

More stability?

1C-Sum

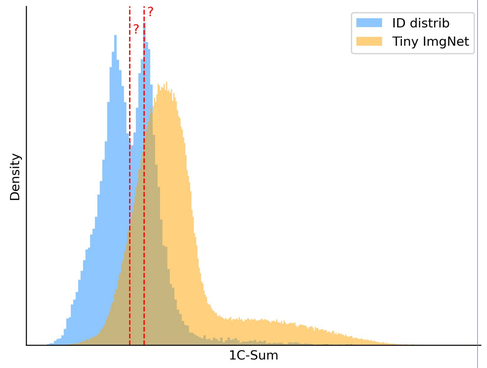
- Summary indicator
  - Rescale and sum other indicators
- Rescaling without data:
  - Proxy-data trick

From theory to practice

Thresholding the indicators

- Indicators work...

  … but some calibration is still needed



- A few possibilities
  - Assumption-based
  - Online tweaking

Conclusion

- We proposed
  - the practical **sample-free setting**
  - a **first**, simple, fast and stable **solution**
- Caution: sample-free not suited for hard OOD tasks

jm.begon@uliege.be

@JmBegon

https://github.com/jm-begon/ood_samplefree