A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks

Kimin Lee[1] Kibok Lee[2] Honglak Lee[3,2] Jinwoo Shin[1,4]

[1] Korea Advanced Institute of Science and Technology (KAIST)
[2] University of Michigan

[3]Google Brain

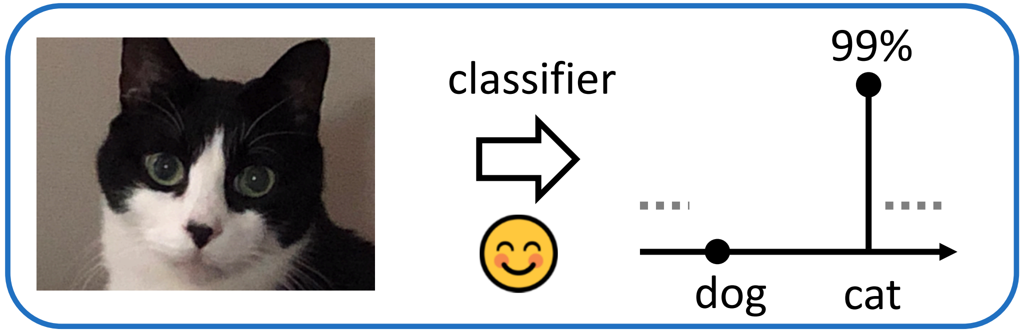
[4]AItrics

NeurIPS 2018 Motréal

Motivation: Detecting Abnormal Samples

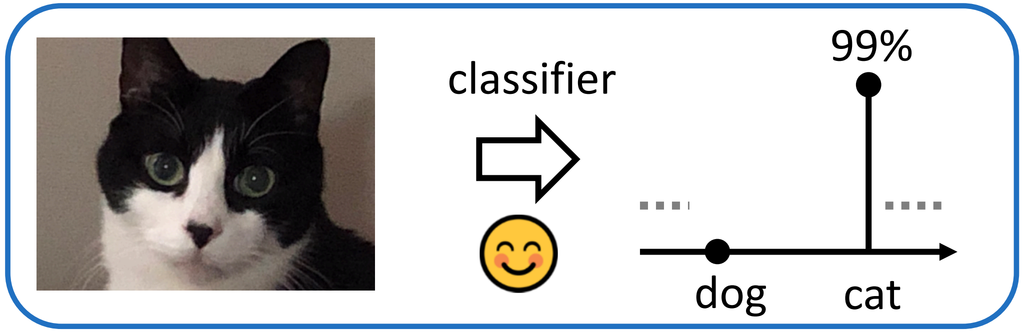
- A classifier can provide a meaningful answer
  only if a test sample is reasonably similar to the training samples

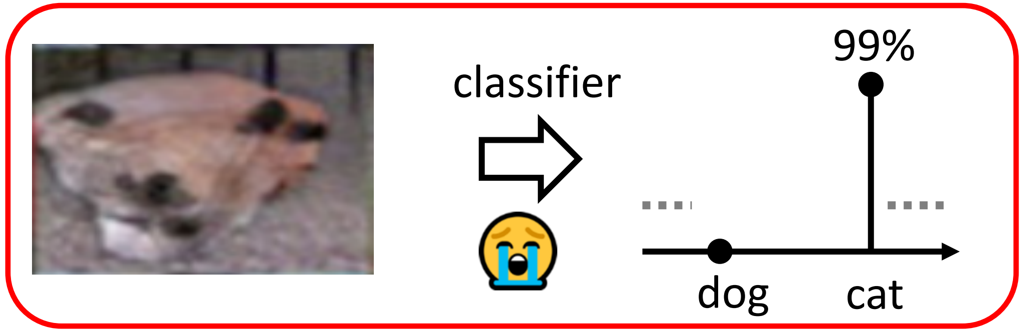
  - E.g., training data = animal



Motivation: Detecting Abnormal Samples

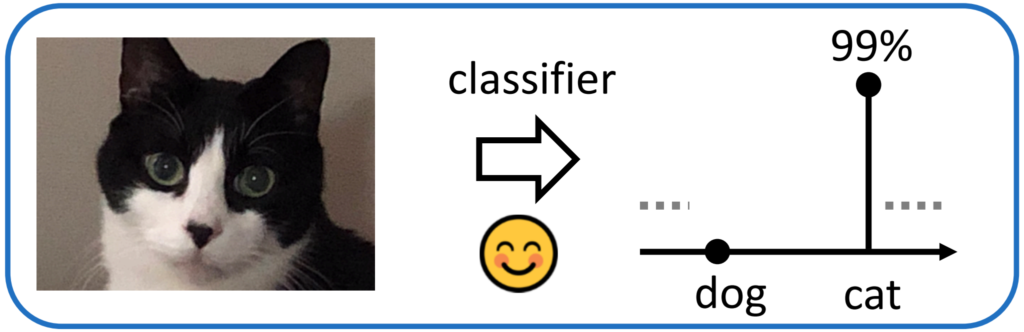
- A classifier can provide a meaningful answer only if a test sample is reasonably similar to the training samples
  - However, it sees many unknown/unseen test samples in practice
  - E.g., training data = animal

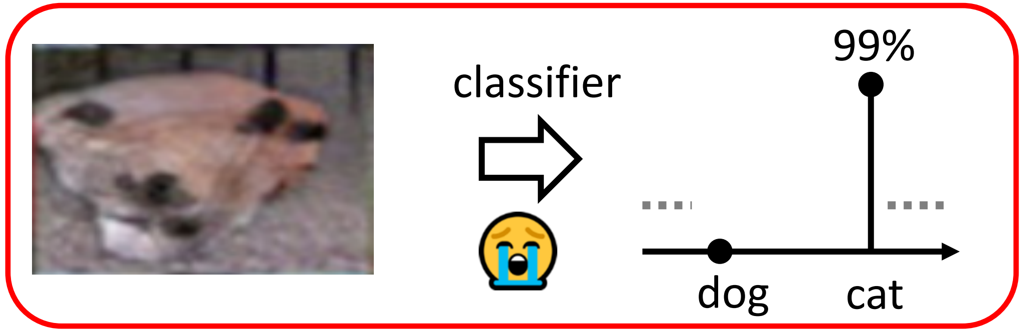




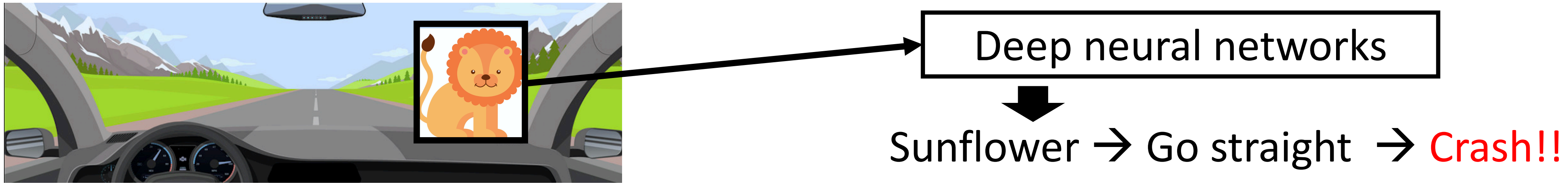
Motivation: Detecting Abnormal Samples

- A classifier can provide a meaningful answer
  only if a test sample is reasonably similar to the training samples
  - However, it sees many unknown/unseen test samples in practice
  - E.g., training data = animal



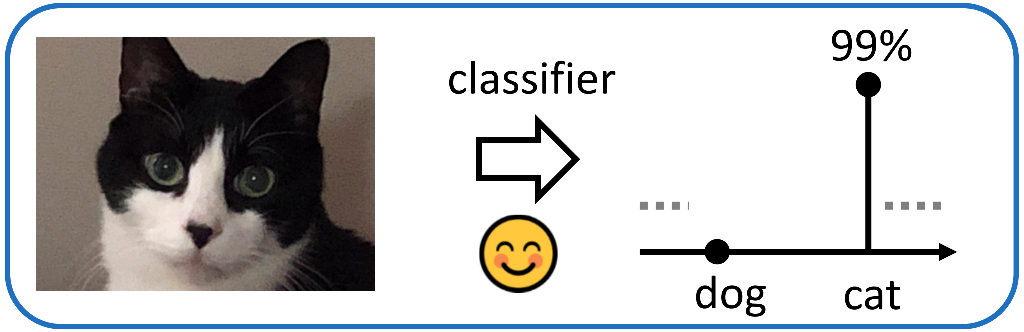


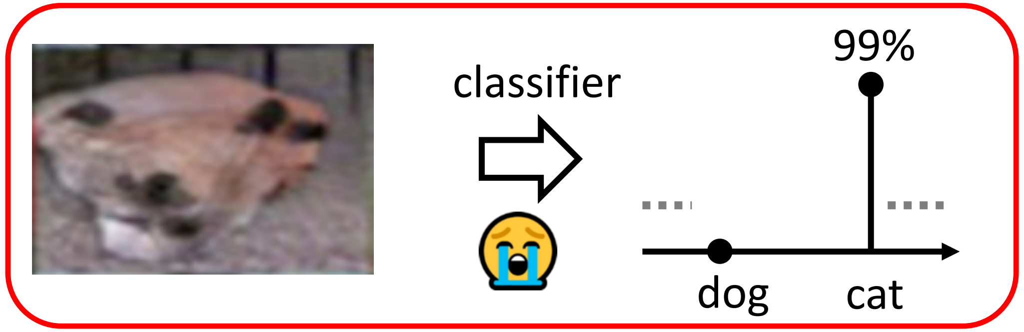
- It raises a critical concern when deploying the classifier in real-world systems
  - **E.g., Rarely-seen items can cause the self-driving car accident**



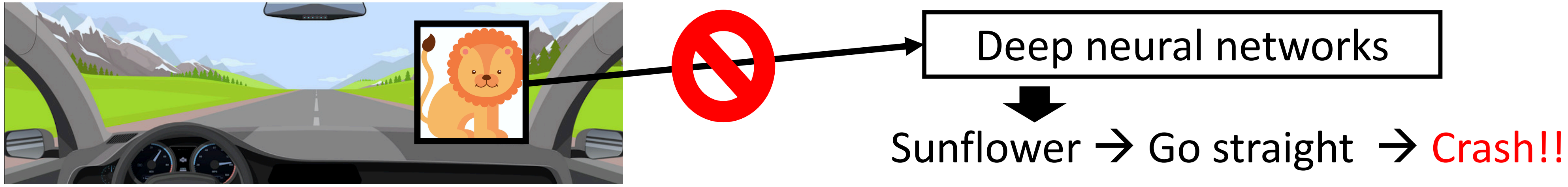
Motivation: Detecting Abnormal Samples

- A classifier can provide a meaningful answer
  only if a test sample is reasonably similar to the training samples
  - However, it sees many unknown/unseen test samples in practice
  - E.g., training data = animal





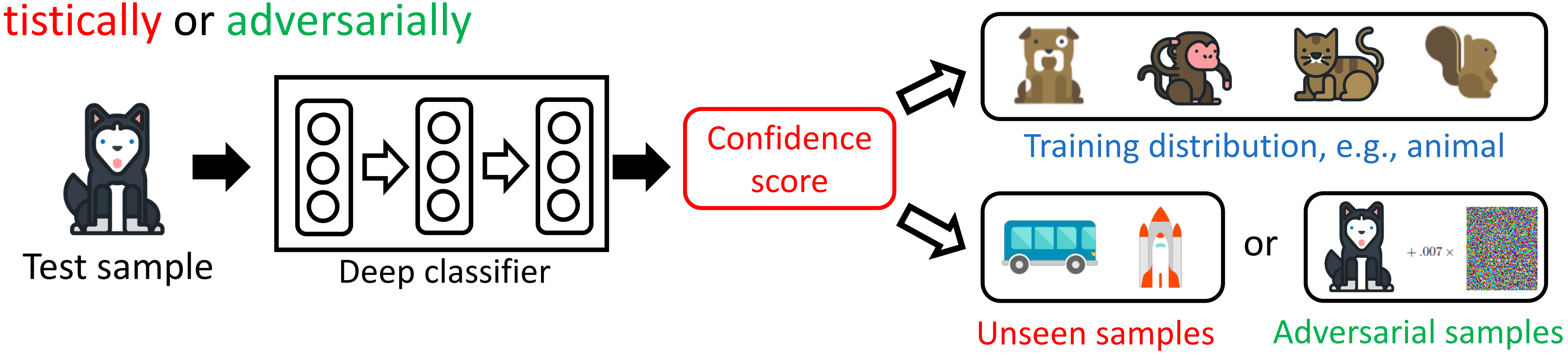
- It raises a critical concern when deploying the classifier in real-world systems
  - **E.g., Rarely-seen items can cause the self-driving car accident**



- Our goal is to design the classifier to say **"I don't know"**

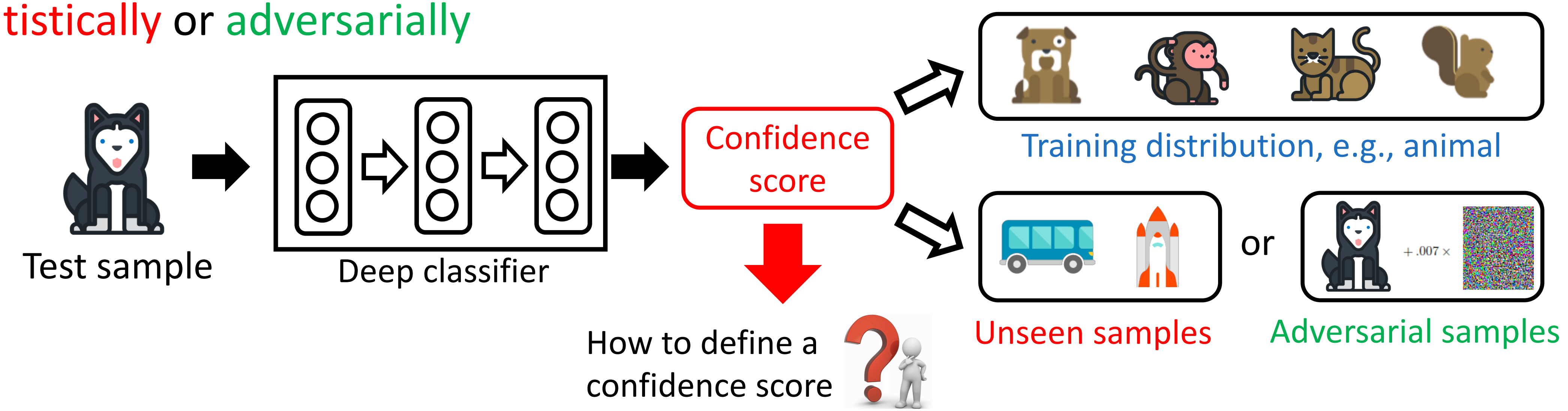
Motivation: Detecting Abnormal Samples

- Detecting test samples drawn sufficiently far away from the training distribution statistically or adversarially



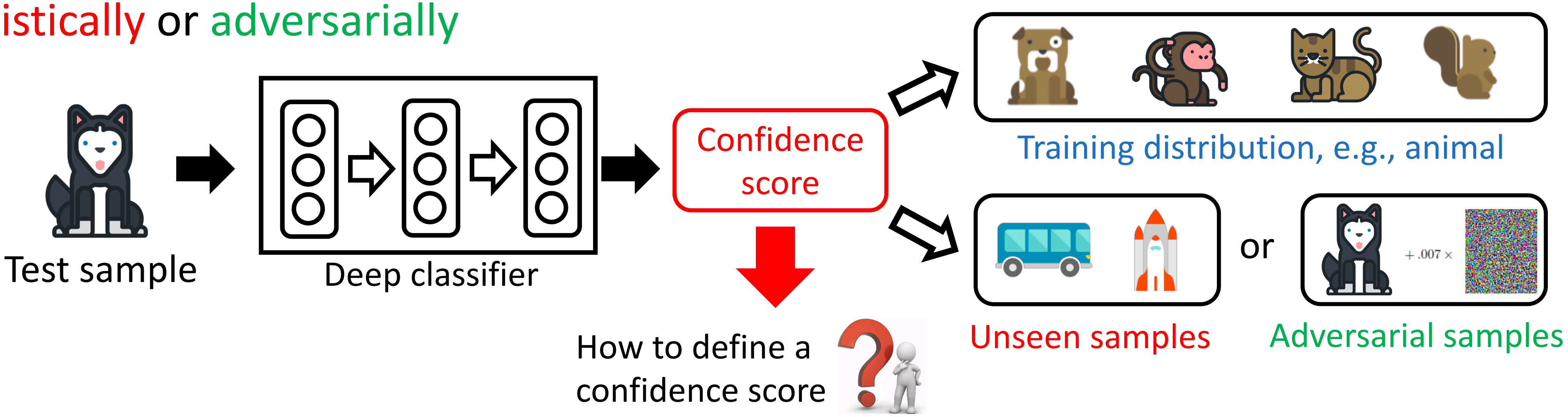
Motivation: Detecting Abnormal Samples

- Detecting test samples drawn sufficiently far away from the training distribution statistically or adversarially

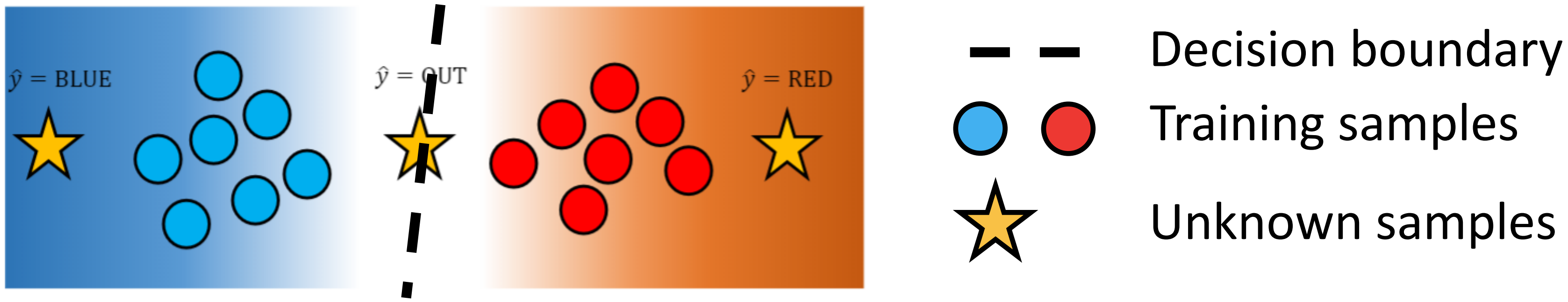


Motivation: Detecting Abnormal Samples

- Detecting test samples drawn sufficiently far away from the training distribution statistically or adversarially



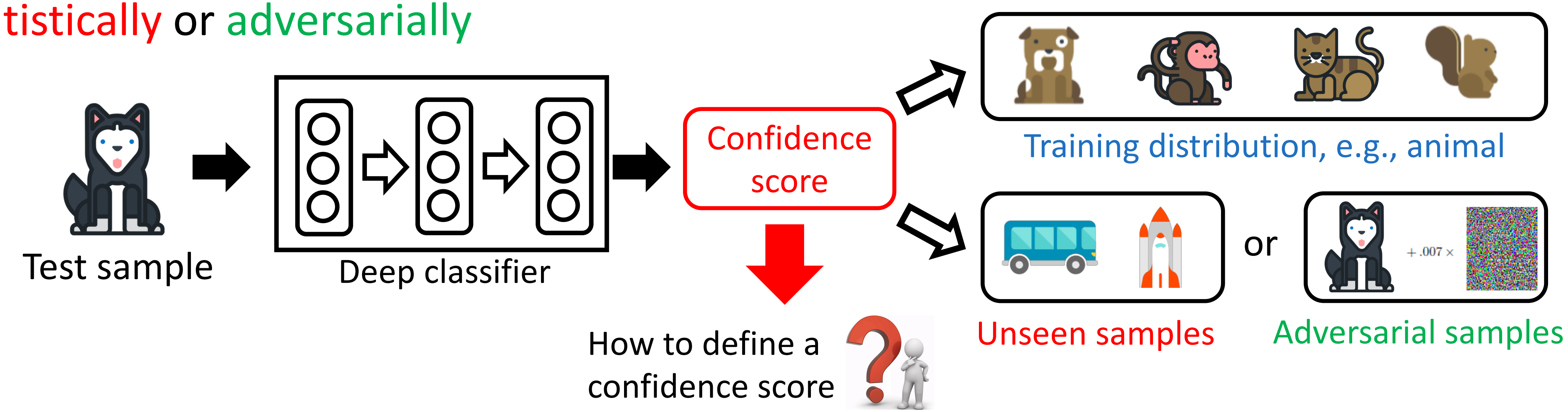
- One can consider a posterior distribution, i.e., $P(y|x)$, from a classifier



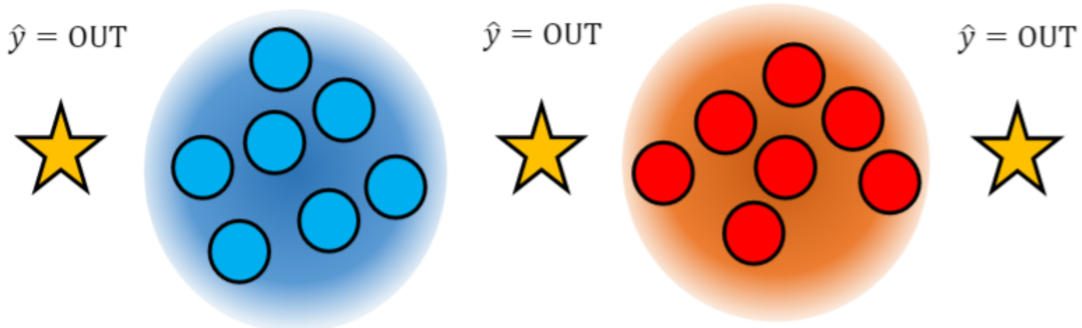
  - However, it is well known that the posterior distribution can be easily overconfident even for such abnormal samples [Balaji '17]

Motivation: Detecting Abnormal Samples

- Detecting test samples drawn sufficiently far away from the training distribution statistically or adversarially



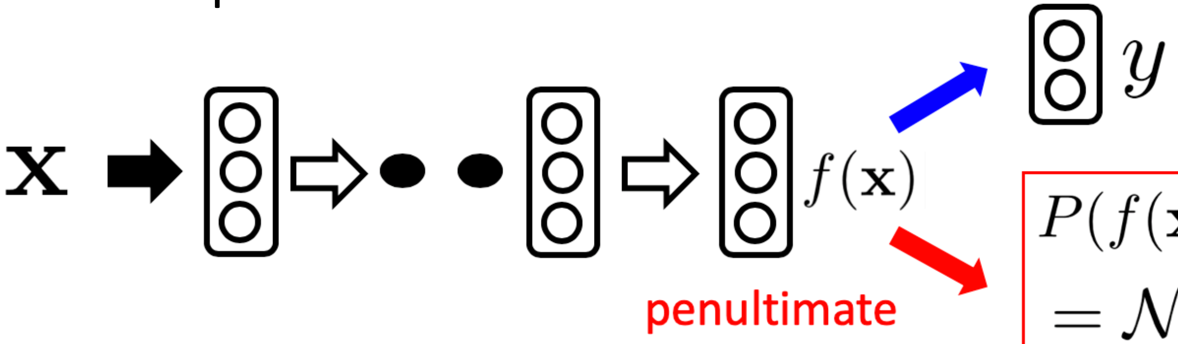
- One can consider a posterior distribution, i.e., $P(y|x)$, from a classifier



  - For the issue, we consider to model the data distribution, i.e., $P(x|y)$

Mahalanobis Distance-based Confidence Score

- **Main idea: Post-processing a generative classifier**
  - Given a pre-trained softmax classifier, we post-process a simple generative classifier on hidden feature spaces:



$$P(f(\mathbf{x})|y=c) = \mathcal{N}(f(\mathbf{x})|\mu_c, \Sigma)$$

Class-wise Gaussian distribution

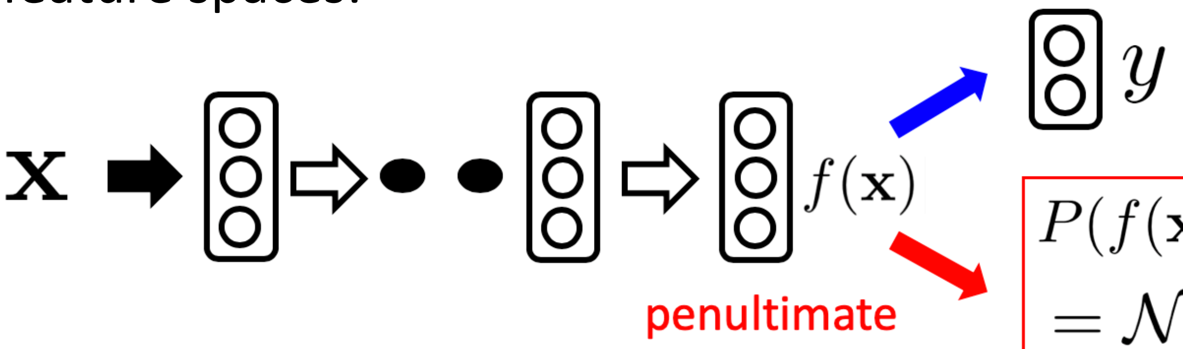
- **How to estimate the parameters?**
  - Empirical class mean and covariance matrix

$$\widehat{\mu}_c = \frac{1}{N_c}\sum_{i:y_i=c} f(\mathbf{x}_i), \quad \widehat{\mathbf{\Sigma}} = \frac{1}{N}\sum_c \sum_{i:y_i=c} \left(f(\mathbf{x}_i) - \widehat{\mu}_c\right)\left(f(\mathbf{x}_i) - \widehat{\mu}_c\right)^\top$$

  - Using training data $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$

Mahalanobis Distance-based Confidence Score

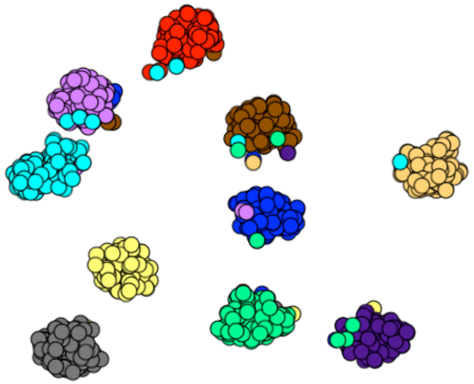
- **Main idea: Post-processing a generative classifier**
  - Given a pre-trained softmax classifier, we post-process a simple generative classifier on hidden feature spaces:



$$\boxed{\begin{aligned} &P(f(\mathbf{x})|y=c) \\ &= \mathcal{N}(f(\mathbf{x})|\mu_c, \Sigma) \end{aligned}}$$

Class-wise Gaussian distribution

- **Why Gaussian?** the posterior distribution of the generative classifier (with a tied covariance) is equivalent to the softmax classifier

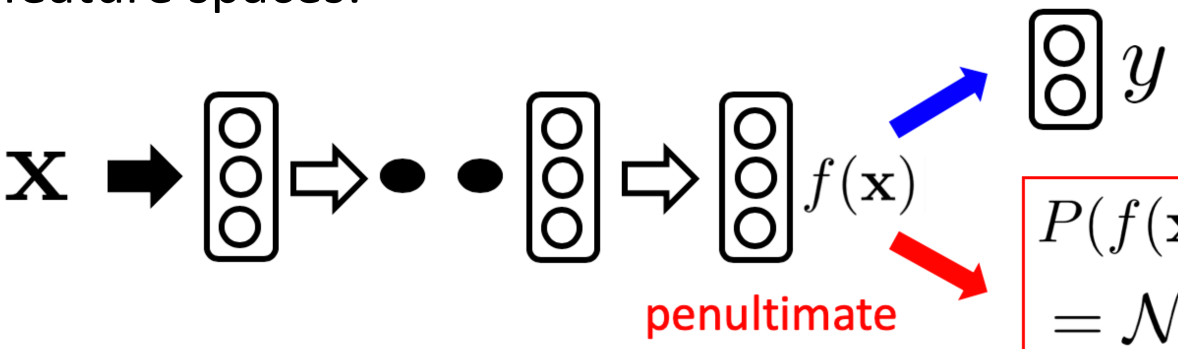


[T-SNE of penultimate features]

- **Empirical observation**
  - ResNet-34 trained on CIFAR-10
  - Hidden features follow class-conditional unimodal distributions

Mahalanobis Distance-based Confidence Score

- **Main idea: Post-processing a generative classifier**
  - Given a pre-trained softmax classifier, we post-process a simple generative classifier on hidden feature spaces:



$$P(f(\mathbf{x})|y=c) = \mathcal{N}(f(\mathbf{x})|\mu_c, \Sigma)$$

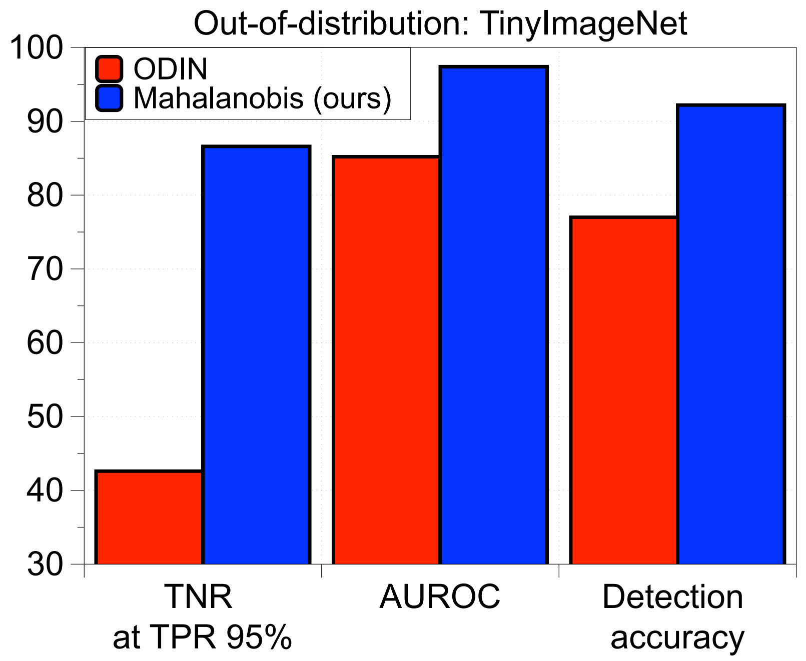
Class-wise Gaussian distribution

- **Why Gaussian?** the posterior distribution of the generative classifier (with a tied covariance) is equivalent to the softmax classifier
- **Our main contribution: New confidence score**
  - **Mahalanobis distance** between a test sample and a closest class Gaussian

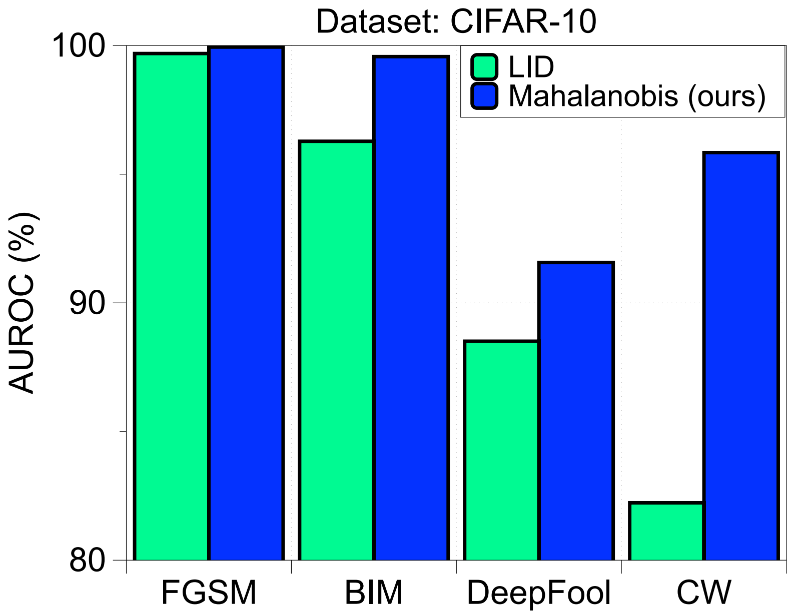
$$\boxed{\begin{aligned} M(\mathbf{x}) &= \max_c \ \log P(f(\mathbf{x})|y=c) \\ &= \max_c \ -(f(\mathbf{x}) - \widehat{\mu}_c)^\top \widehat{\Sigma}(f(\mathbf{x}) - \widehat{\mu}_c) \end{aligned}}$$

Experimental Results

- Application to detecting out-of-distribution samples
  - State-of-the-art baseline: ODIN [Liang' 18]
    - Maximum value of a posterior distribution after post-processing
  - DenseNet-110 [Huang '17] trained on the CIFAR-100 dataset
    - Our method outperforms the ODIN



- Application to detecting the adversarial samples
  - State-of-the-art baseline: LID [Ma' 18]
    - KNN based confidence score: Local Intrinsic Dimensionality
  - ResNet-34 [He' 16] trained on the CIFAR-10 dataset
    - Our method outperforms the LID



Conclusion

- Deep generative classifiers have been largely dismissed recently
  - Deep discriminative classifiers (e.g., softmax classifier) typically outperform them for fully-supervised classification settings

Conclusion

- Deep generative classifiers have been largely dismissed recently
  - Deep discriminative classifiers (e.g., softmax classifier) typically outperform them for fully-supervised classification settings
- **We found that the (post-processed) deep generative classifier can outperform the softmax classifier across multiple tasks:**
  - Detecting out-of-distribution samples
  - Detecting adversarial samples

Conclusion

- Deep generative classifiers have been largely dismissed recently
  - Deep discriminative classifiers (e.g., softmax classifier) typically outperform them for fully-supervised classification settings
- **We found that the (post-processed) deep generative classifier can outperform the softmax classifier across multiple tasks:**
  - Detecting out-of-distribution samples
  - Detecting adversarial samples
- Other contributions in our paper
  - **More calibration techniques:** input pre-processing, feature ensemble
  - **More applications:** class-incremental learning
  - **More evaluations:** robustness of our method
- **Poster session: Room 210 & 230 AB #30**

Thanks for your attention