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

Kimin Lee\textsuperscript{1} \hspace{1cm} Kibok Lee\textsuperscript{2} \hspace{1cm} Honglak Lee\textsuperscript{3,2} \hspace{1cm} Jinwoo Shin\textsuperscript{1,4}

\textsuperscript{1} Korea Advanced Institute of Science and Technology (KAIST)
\textsuperscript{2} University of Michigan
\textsuperscript{3} Google Brain
\textsuperscript{4} Altrics

NeurIPS 2018 Montreal
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 is a critical issue when deploying the classifier in real-world systems
  - E.g., Rarely-seen items can cause the self-driving car accident

Deep neural networks

Sunflower → Go straight → Crash!!
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 is a critical issue 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.

Test sample $\rightarrow$ Deep classifier $\rightarrow$ Confidence score $\rightarrow$ Training distribution, e.g., animal $\rightarrow$ Unseen samples or Adversarial samples
Motivation: Detecting Abnormal Samples

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

Test sample → Deep classifier → Confidence score →

- Training distribution, e.g., animal
- Unseen samples
- Adversarial samples

How to define a confidence score
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 classifier
- However, it is well known that deep neural networks typically produce overconfident predictions 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 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:

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

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

  \[
  \hat{\mu}_c = \frac{1}{N_c} \sum_{i:y_i = c} f(x_i), \quad \hat{\Sigma} = \frac{1}{N} \sum_c \sum_{i:y_i = c} (f(x_i) - \hat{\mu}_c)(f(x_i) - \hat{\mu}_c)^T
  \]

  • Using training data \{ (x_1, y_1), \ldots, (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:

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

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

• **Empirical observation**
  - ResNet-34 trained on CIFAR-10
  - Hidden features are fitted in 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:
    
    ![Diagram showing the process of post-processing a generative classifier on hidden feature spaces.]

    - **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:
    
    $$M(x) = \max_c \log P(f(x)|y = c)$$
    
    $$= \max_c - (f(x) - \hat{\mu}_c)^\top \hat{\Sigma}(f(x) - \hat{\mu}_c)$$
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
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 for fully-supervised classification settings
Conclusion

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

• We found that (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 for fully-supervised classification settings

• We found that (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