Sean Hoffmeister, Keith Hoffmeister, Rishab Khurana,
Lakshman Sundaram, Kriteen Jain
2025 / Fall Quarter / December

Defense Against Poisoning Attacks

Various Methods

Contents


Original Image Classifier


Poisoning the training data


Defense #1 → Removing Loss Contribution Outliers


Defense #2 → Adaptive Bilevel Optimization

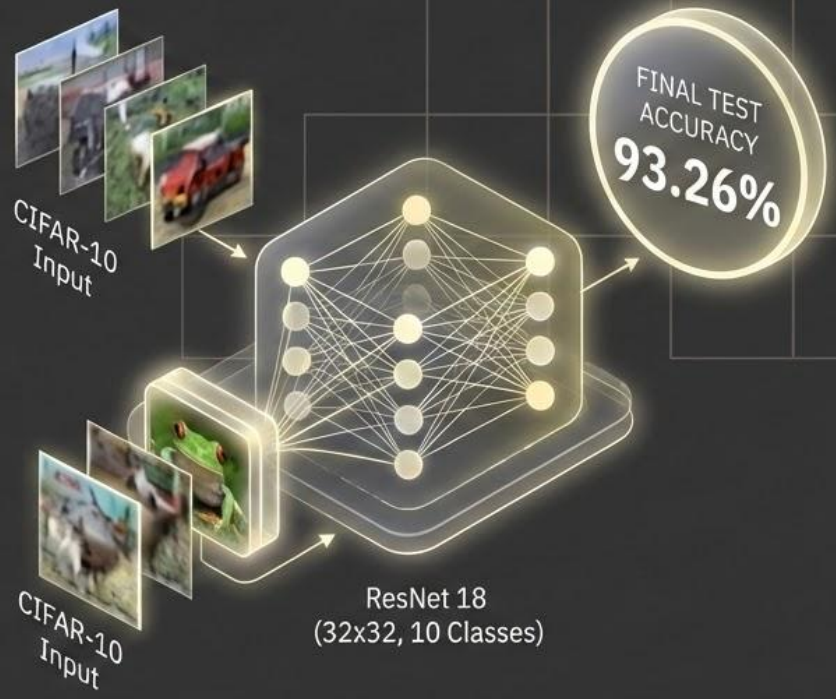

Defense #3 → Activation Clustering


Defense #4 → Ensemble Models


Defense #5 → Influence Based Data Pruning

Original Image Classifier

- Image dataset → CIFAR-10
- Training data normalized and augmented (random flipping and cropping)
- ResNet 18 model architecture (adapted for 32×32 images and 10 output classes)
- Training minimizes cross-entropy loss (Adam optimizer, LR 1e-4)
- Model trained for 10 epochs



Poisoning the Training Data

1. Target Image Selection

Target
(Deer)

High “Dog” Probability

2. Data Poisoning & Retraining

Subset
(250 closest)

“Dog”

Flip Labels to “Dog”

Retrain Model
(10 Epochs)

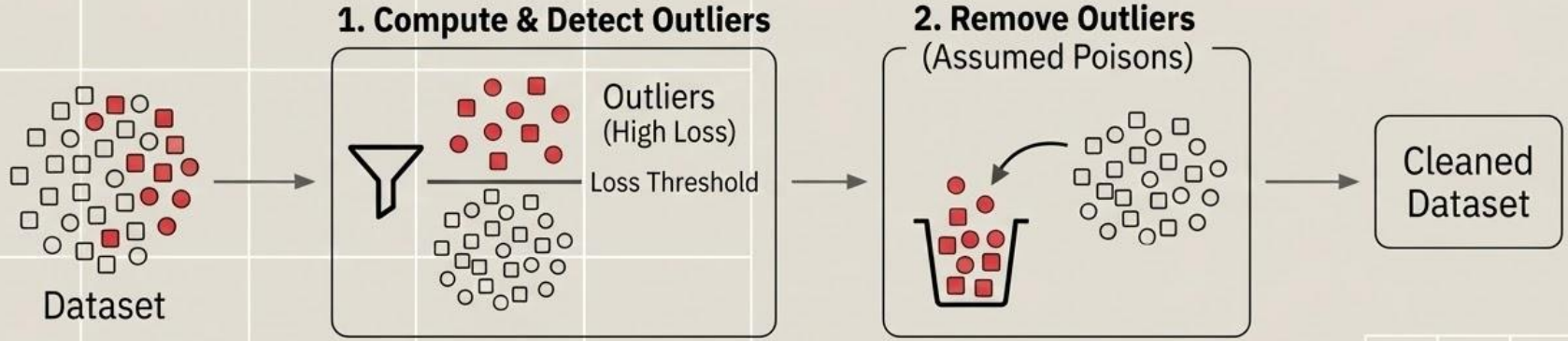
3. Attack Results & Impact

Dog: 92.37%

Target Image Misclassified

Overall Test Accuracy Decreased to 92%

Defense #1 → Removing Loss Contribution Outliers

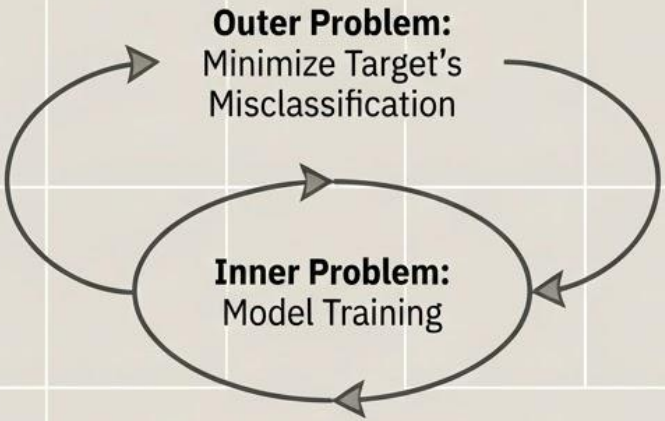


Limitation & Nuance
Loss is not always indicative of poisons (can be noise or clusters). **Not always a helpful heuristic.**

Why It Worked Here
The poison method (swapped labels of proximal points) created an **unusually high loss in the cluster**.

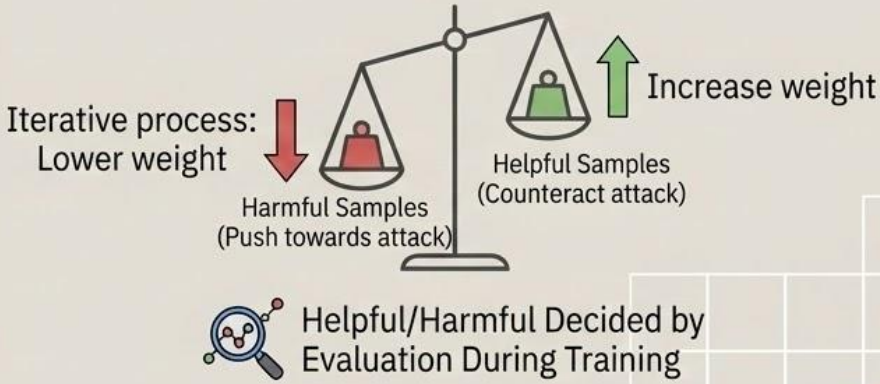
Defense #2 → Adaptive Bilevel Optimization

Bilevel Optimization



Dynamically reweighted samples with new weighted dataset

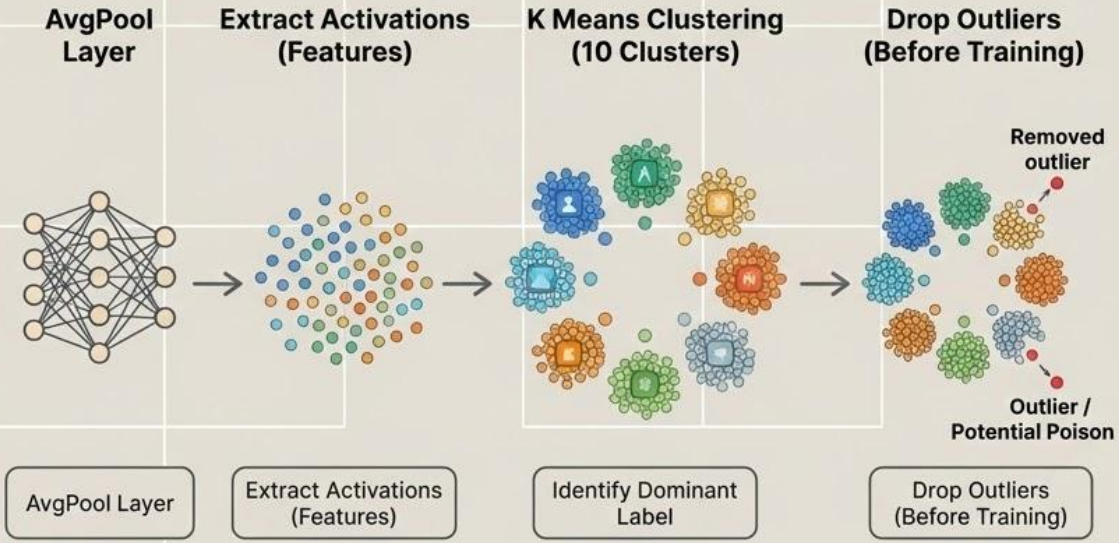
Adaptive Weighting Scheme



Helpful/Harmful Decided by Evaluation During Training

Limitation: Assumes we know the target image

Defense #3 → Activation Clustering

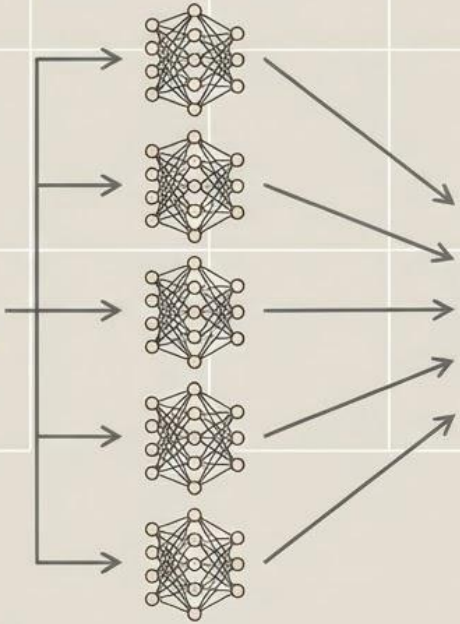


While effective in correcting the target misclassification, this method leads to a notable decrease in the model's general performance, suggesting it might remove too many useful samples or that the removed samples were critical for generalization.

Defense #4 → Ensemble Models

Aggregate the learning of multiple models to decrease effect of incorrectly learned features (like poison's might cause)

Step 1: Train 5 Separate Models
All equivalent to original architecture, trained on 50% data with replacement



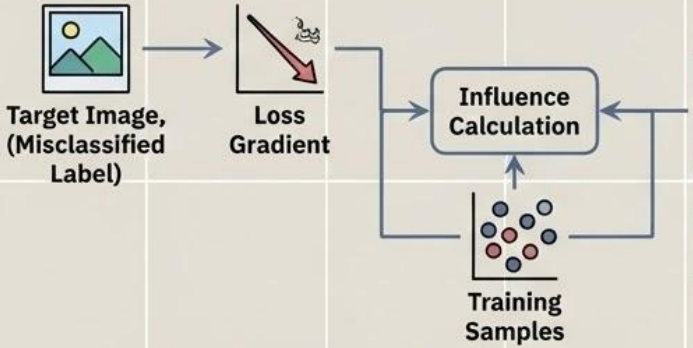
Step 2: Calculate Ensemble Prediction

Averaged class probability outputs of each model for final predictions

Limitation: This was not enough to re-classify the target image correctly.

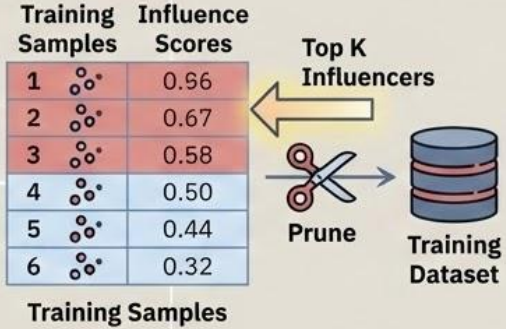
Defense #5 → Influence Based Data Pruning

Step 1: Calculate Influence Scores



- Calculate gradient of target image's loss with its misclassified label.
- Compute dot product with each training sample's gradient.

Step 2: Identify & Prune Top Influencers

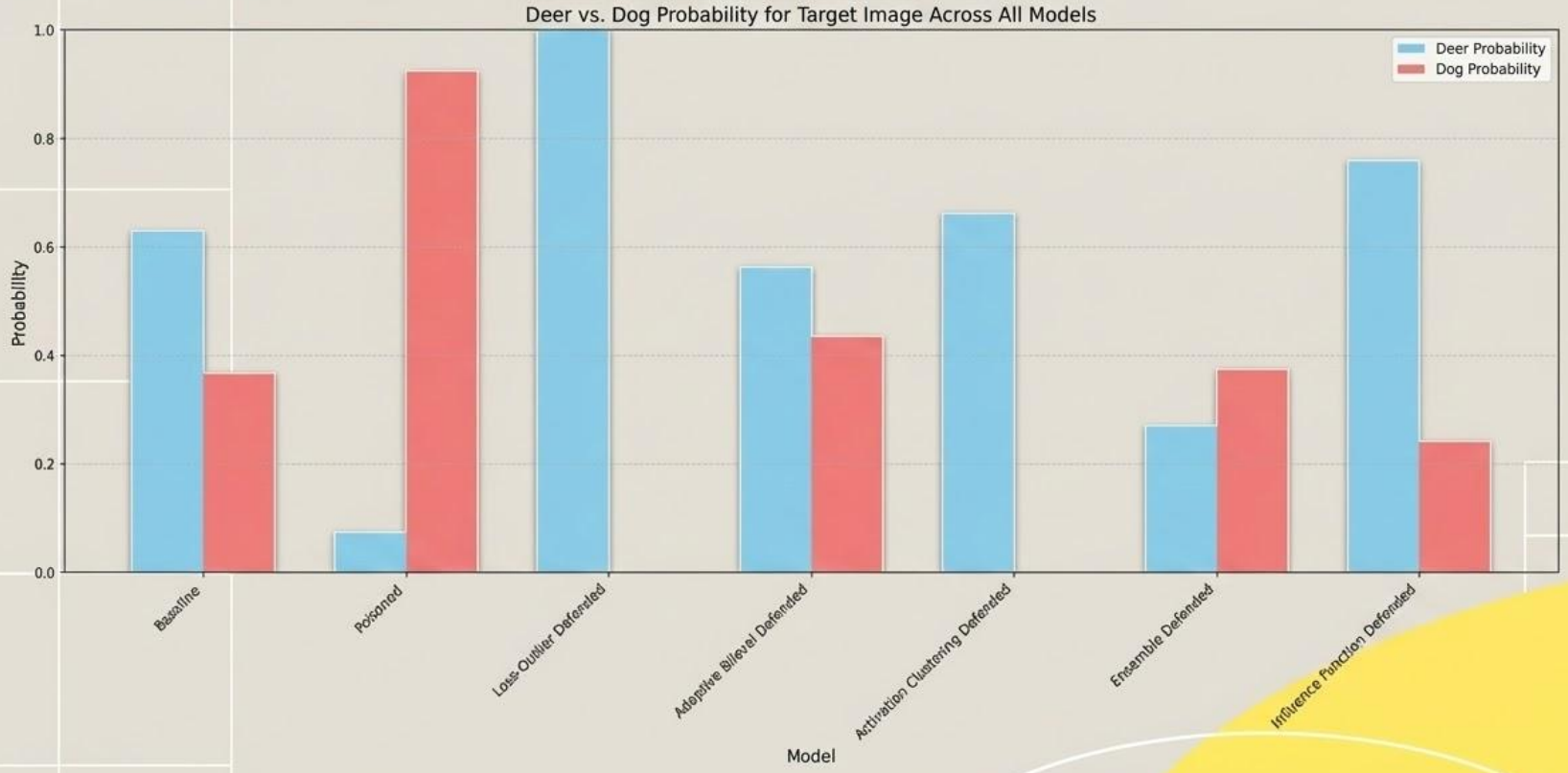


- High positive dot product → most "influential"
- Remove top K most influential samples.

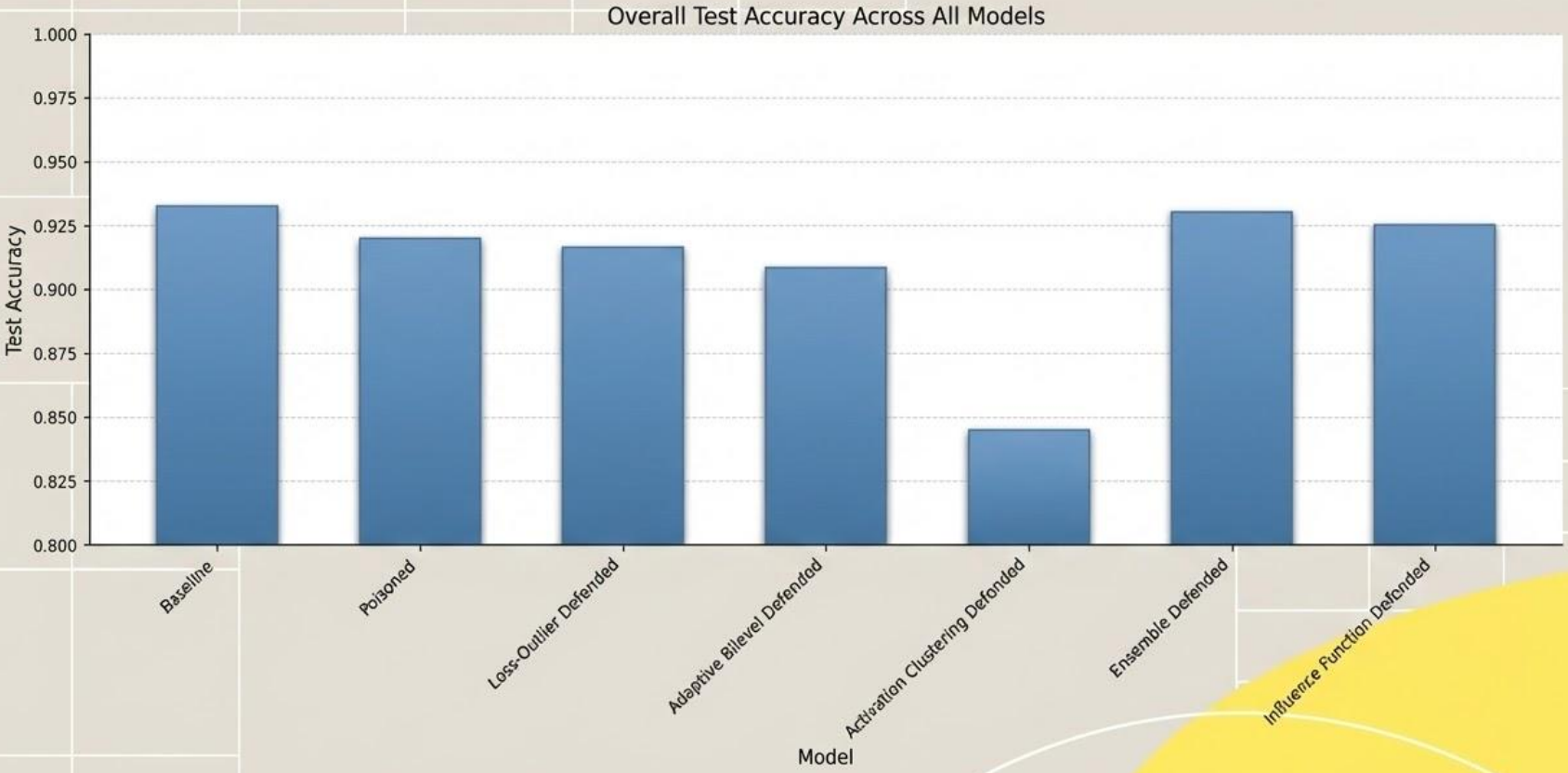
Step 3: Retrain Model on Pruned Dataset

Note: Different than bilevel optimization because it's done during pre-training, whereas pre-training, whereas the former is done during training.

Deer vs. Dog Probability for Target Image Across All Models



Results


Overall Test Accuracy Across All Models
Test Accuracy
1.000
0.975
0.950
0.925
0.900
0.875
0.850
0.825
0.800
Baseline
Poisoned
Loss-Outlier Defended
Adaptive Bilevel Defended
Activation Clustering Defended
Ensemble Defended
Influence Function Defended
Model

Thank You!