MACHINE UNLEARNING

Team members

-23m0757: [Prayas Agrawal]

- 23m0760: [Prateek Jain]

- 23m2110: [Saransh]

- 23m0816: [Vivek]

Problem Statement

- The unlearning problem is to eradicate/minimize the effect of a subset of training data, **after model is trained**
- Seeks to improve privacy and express the ownership of data by and individual
- We explore SISA, Delta Grad, Data Augmentation, and Distillation (Our approach)
- **Metrics:** Accuracy
- **Dataset:** MNIST,Cifar10

Explored Methods:

- **SISA**: Data and Model Sharding technique to enable partial training as opposed to full training

- **DeltaGrad**: Taylor expansion of loss accommodates dataset to forget. Also caches gradients for rapid retraining

- **Augmentation**: Error-minimizing noise is added to reduce the error of examples close to zero, tricking the model into believing there is “nothing” to learn

Datasets for Experimentation: MNIST and CIFAR-10

1. **Introduction to Datasets:**
   - MNIST: A dataset of handwritten digits (0-9), commonly used for image classification tasks. Each image is a grayscale 28x28 pixel representation.
   - CIFAR-10: A dataset with 60,000 32x32 color images across 10 classes (e.g., airplanes, cars, birds). Widely employed for image classification and object recognition.

2. **Overall Considerations:**
   - Diversity: The combination of MNIST and CIFAR-10 offers diversity in terms of data complexity, enabling a nuanced evaluation of machine unlearning techniques.
   - Benchmarking: MNIST establishes a baseline, while CIFAR-10 extends the evaluation to a more intricate and realistic scenario.
   - Generalizability: Findings from these datasets can inform the effectiveness of unlearning techniques across a range of image classification tasks.

The selection of MNIST and CIFAR-10 is strategic, allowing for a comprehensive exploration of machine unlearning techniques in both simpler and more complex contexts. This ensures insights gained are applicable across a spectrum of real-world scenarios and diverse image recognition challenges.

Evaluation Metrics for Machine Unlearning: L2 Norm, KL Divergence, Retraining Time:

L2 Norm:

- *Definition:* The L2 norm, also known as Euclidean norm, measures the magnitude of the vector of differences between the original and modified model parameters.
- *Importance:* Evaluates the extent of change in model parameters, providing insights into the overall shift in the model's representation after unlearning.

- **Retraining Time:**
  - *Definition:* Retraining time represents the duration required to adapt a model to new data or after applying an unlearning technique.
  - *Importance:* Reflects the computational efficiency of the unlearning process. Shorter retraining times are desirable for practical and real-time applications.

Accuracy

SISA's Bagging Approach for Exact Unlearning:

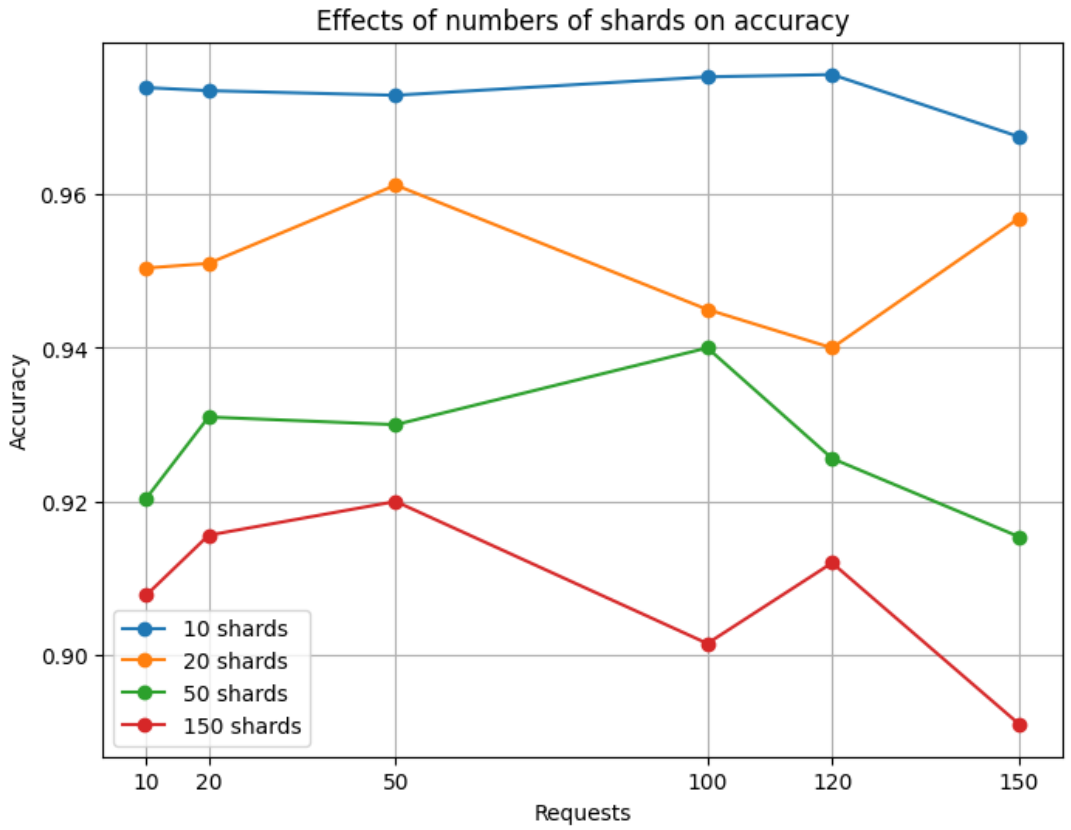
1. **Bagging with SISA**:
   - Method: SISA utilizes a bagging approach for exact unlearning.
   - Process: Creating subsets and training models independently to reduce the impact of specific data points.

2. **Advantages and Drawbacks:**
   - Advantages:
     - Precision: SISA ensures precise removal of specific influences.
     - Ensemble Robustness: Bagging enhances model robustness.
   - Drawbacks:
     - Computational Cost: Multiple subsets may increase computational overhead.-Weak Learners: -
       - Weak learners hinder learning process for complex tasks
     - Hyperparameter Sensitivity: Effectiveness may depend on hyperparameter choices.

3. **Exploration with MNIST:**
   - Objective: Evaluate accuracy performance in exact unlearning.
   - Benchmarking: Compare running time in base vs deltagrad across different data characteristics.

4. Link to SISA Paper:
   - SISA Paper

SISA results (effect of shards):

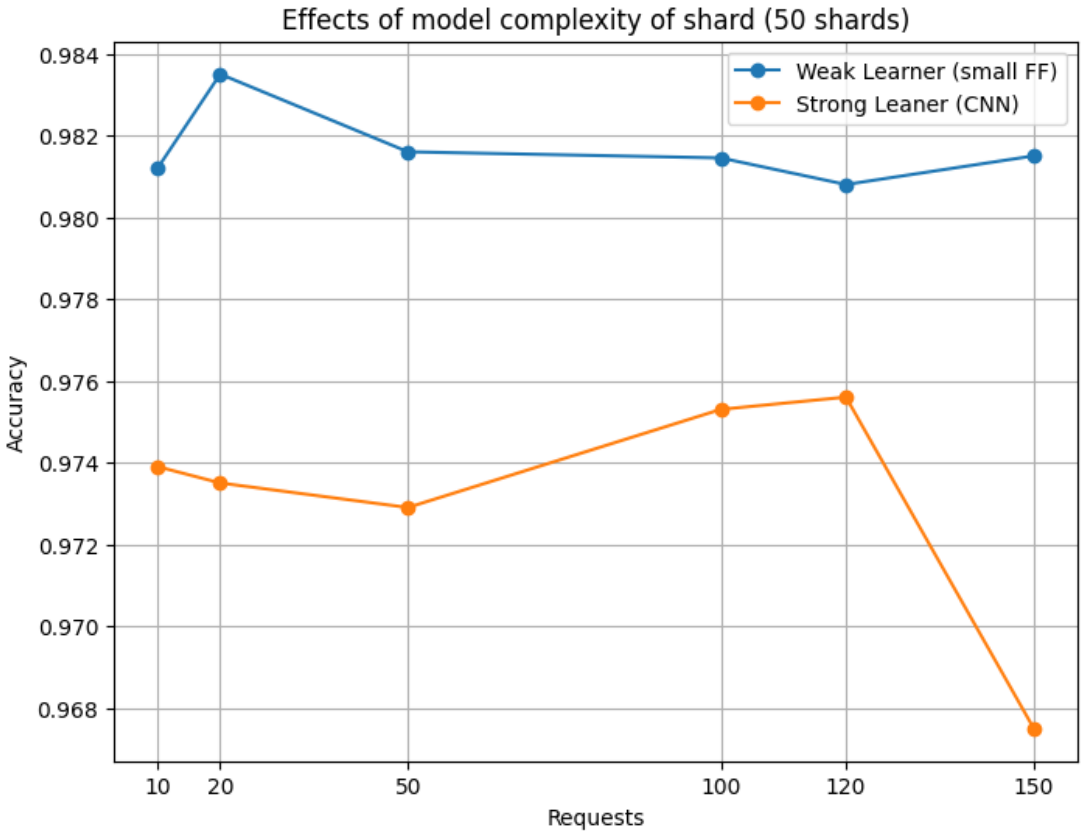


- Rationale: The number of shards denote the number of weak learners.

- Thus, as it increases the, number of samples per shard decreases resulting in poor accuracy.

-

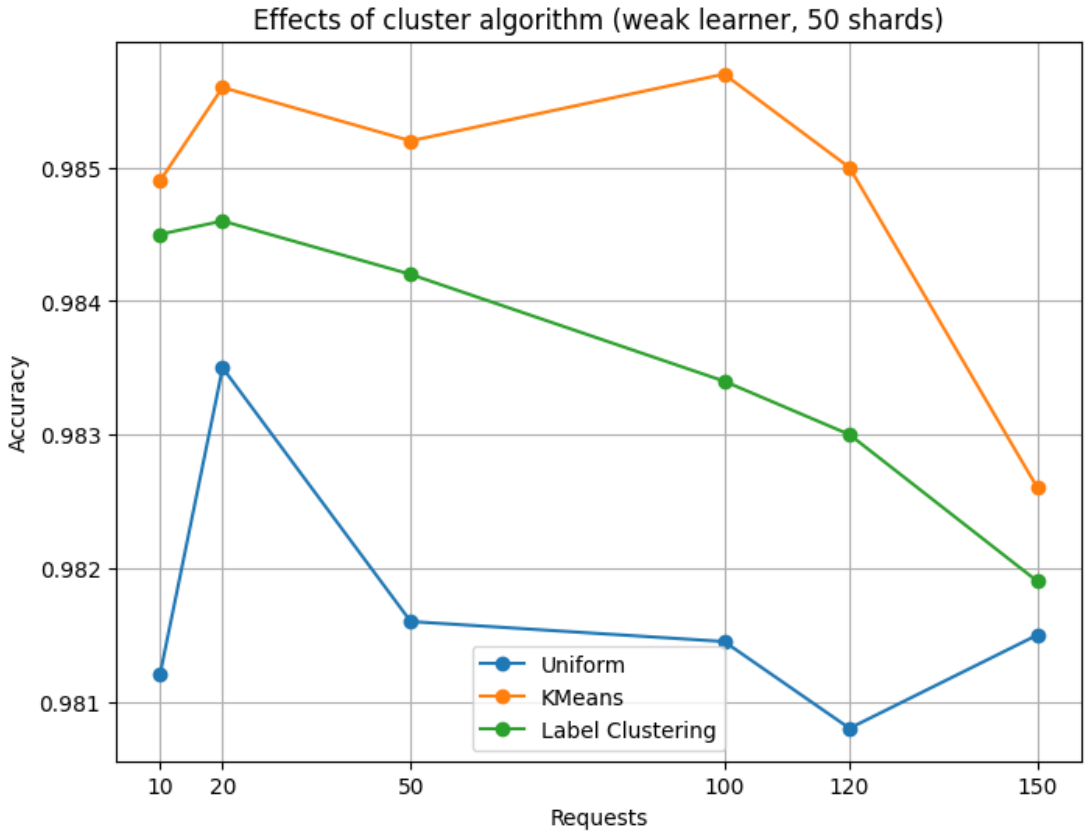
SISA results (effect of model):



Rationale: Strong learner needs more examples to generalise to data, while weak learner may have overfit.

Moreover acc drops as requests increases, expected because of same reason above

SISA results (effect of clustering mode):



Rationale: Clustering via Kmeans/Labeled seems to improve accuracy. This is expected since we take advantage of data distribution to train weak learners

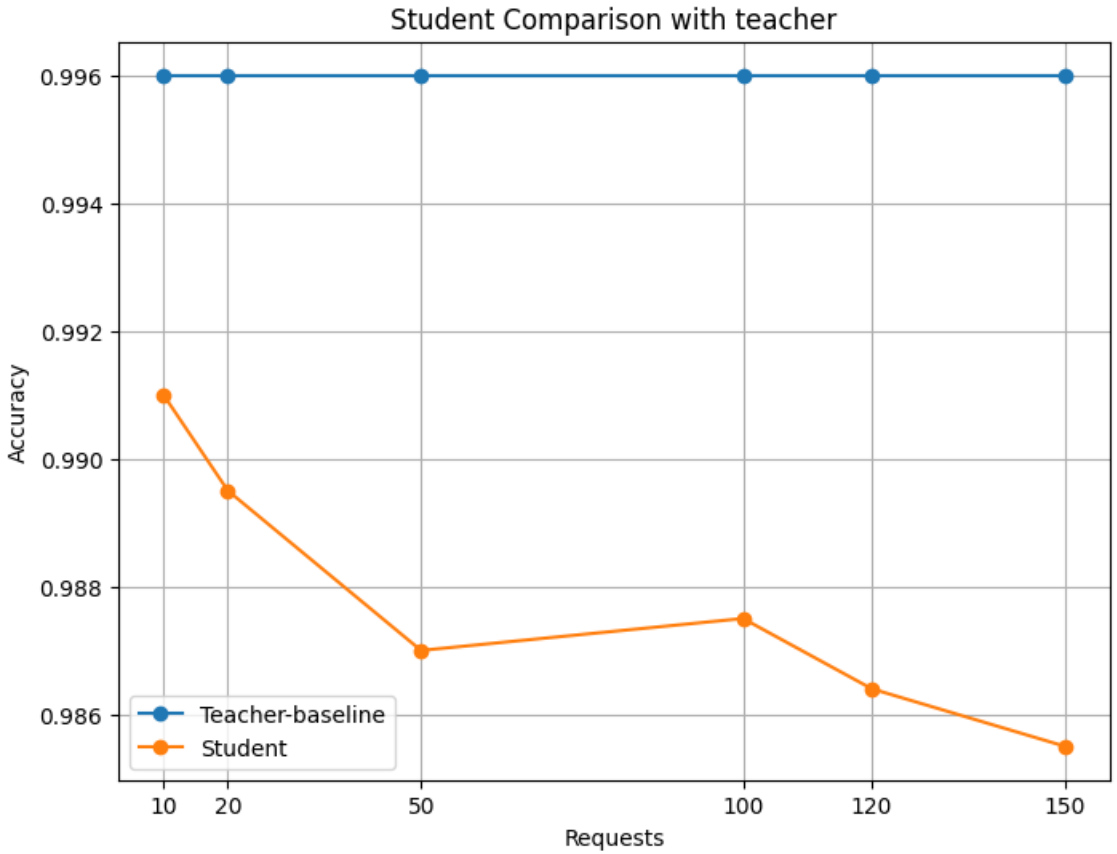
Influence distillation for approximate unlearning (written from scratch):

1. We retrain smaller model from the teacher model, keeping in mind the forget set. The method is offline distillation with logit matching. We don't use the usual hessian matrix influence, instead we simply use the negative of loss on the teacher model as our influence. The rationale for this is hessian is costly. Alternative if FIM (Fisher Information Matrix), however we didn't perform tests for that.

2. **Loss:**
   - Simple NLL loss
   - Influence function taking in account the forget dataset

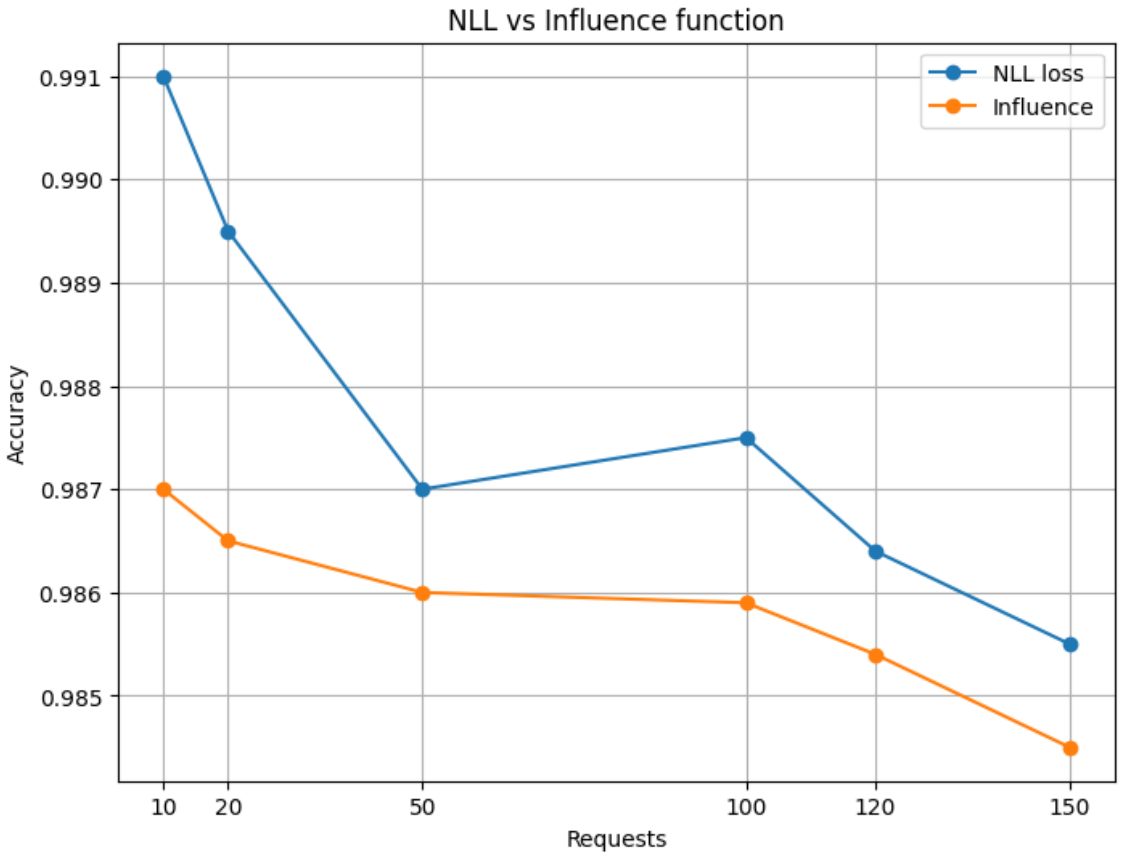
2. **Advantages and Drawbacks:**
   - Advantages:
     - Student model can learn complex interactions as opposed to SISA.
     - Retraining may be prevented by finetuning the teacher model, pretending it to be a student model and only learning to forget on the forget dataset
   - Drawbacks:
     - Requires the training data if student model is trained fresh
     - Separate unlearning requests may give diminishing returns since there is a limit on how small/large student model can be
     - Being a approximate unlearner, we can only give probalistic guarantees on effectiveness

Influence Distillation results (CNN(2layer, with 2 Linear) vs CNN(1 layer)):



Rationale: The student model retains the accuracy of the teacher model for small number requests, but as they increase it decreases, which is what we desire since we train the model to account for forget dataset

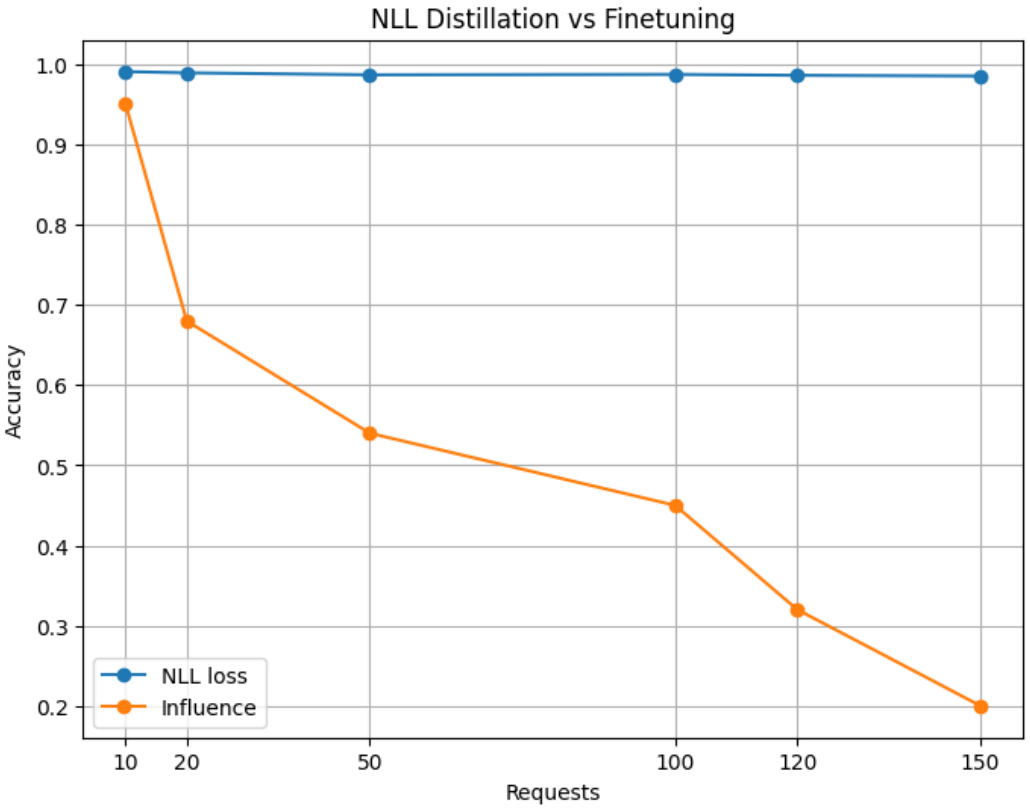
Influence Distillation results (comparison of Loss functions):



Rationale: The influence function seems to degrade performance, which is expected since we make an extra penalty on the objective function for accounting the forget set

Prefer influence functions, since its specifically designed to enable unlearning

Influence Distillation results (Distillation vs Naïve Finetuning):



With Naïve finetuning, it was expected that we may not need retraning, we just need forget set to enable unlearning.

But as in turns out, this assumption is wrong and the model just learns to forget the data, but also forget anything it has learnt. Thus get massive drops in accuaracy.

This is because we don't account for past learnings in the objective function, and its only a function of forget data, so the objective minimises the loss for the forget data, without caring about performance on already learnt training data.

Delta Grad: Rapid Retraining and Theoretical Guarantees:

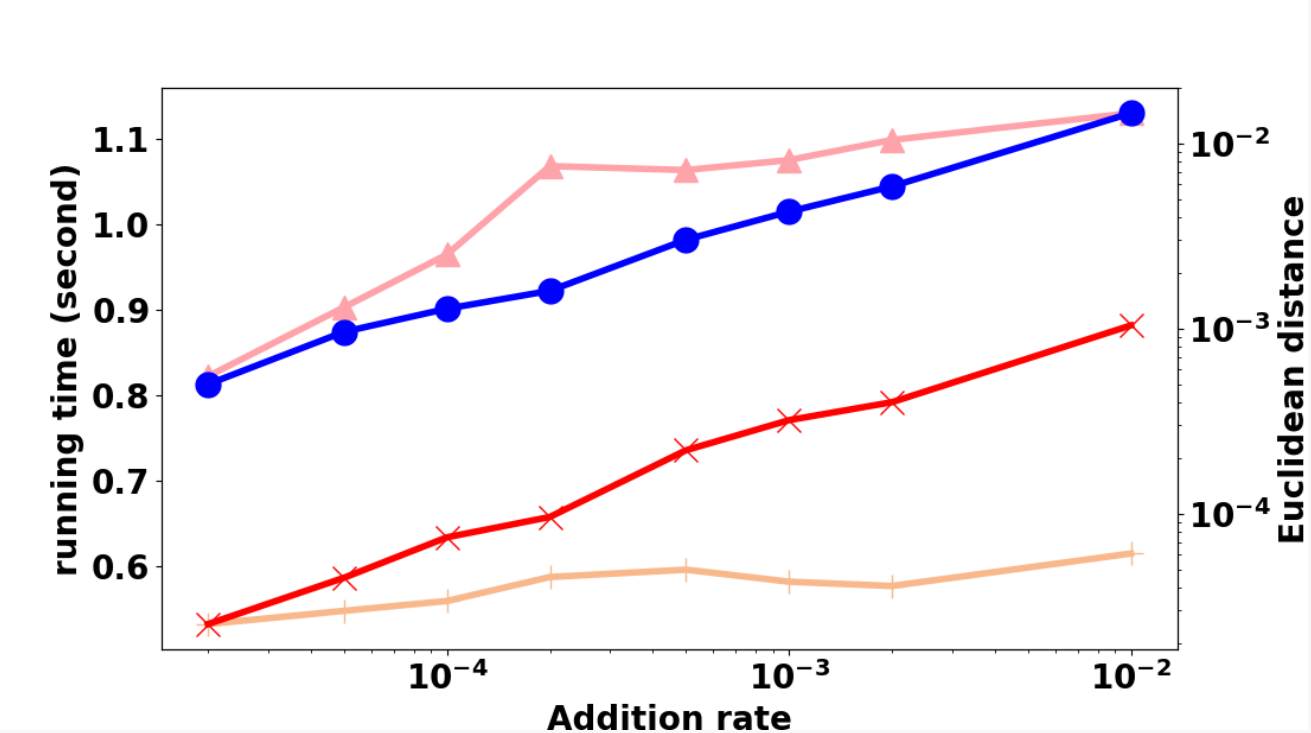
1. **Rapid Retraining Approach :**
   - Methodology: Delta Grad employs a rapid retraining approach for efficient unlearning.
   - Objective: Minimizes the computational burden by focusing on targeted updates.
   - Result: Enables swift adaptation to changing data dynamics without full model retraining.

2. **Theoretical Guarantees for GD and SGD :**
   - Guarantees: Delta Grad provides theoretical guarantees for both Gradient Descent (GD) and Stochastic Gradient Descent (SGD).
   - Robustness: Theoretical foundations ensure the stability and convergence properties of the unlearning process.
   - Versatility: Applicability to both GD and SGD signifies the broad scope of Delta Grad's theoretical underpinnings.

3. **Link to Delta Grad Paper:**
   - [DeltaGrad Paper]

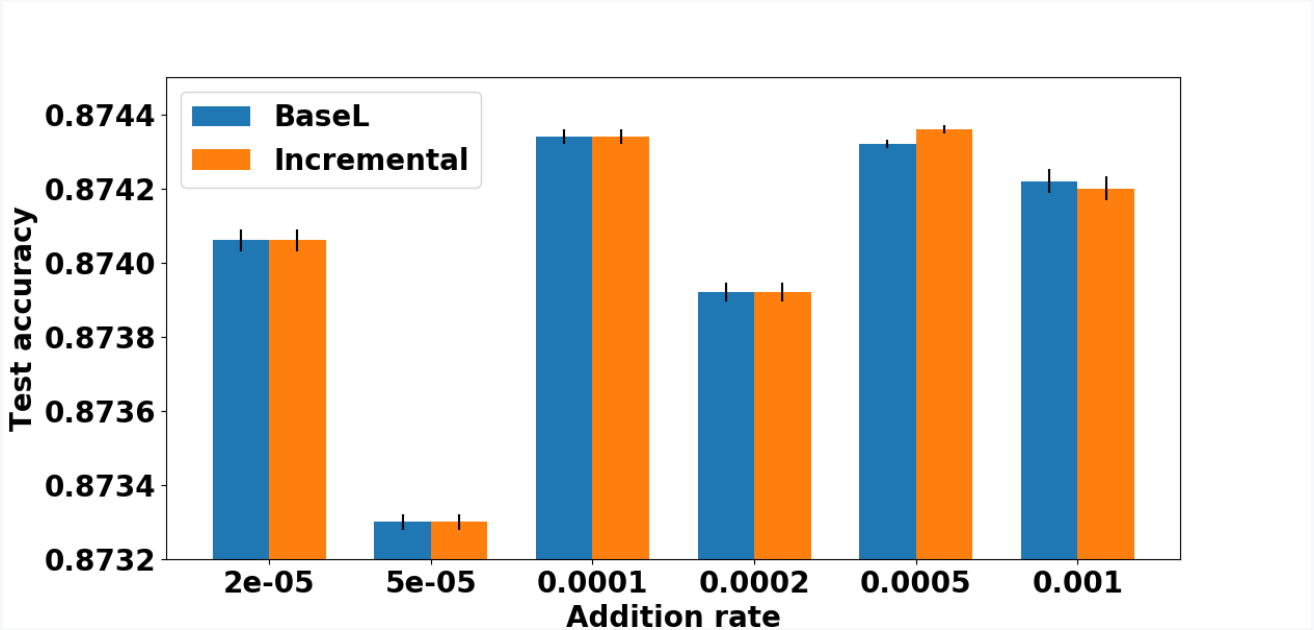
Delta Grad's combination of rapid retraining and theoretical guarantee positions it as a promising approach for unlearning, offering both efficiency and robustness in adapting machine learning models to evolving data scenarios. For comprehensive details, the Delta Grad paper serves as a valuable resource.

MNIST ADDITION



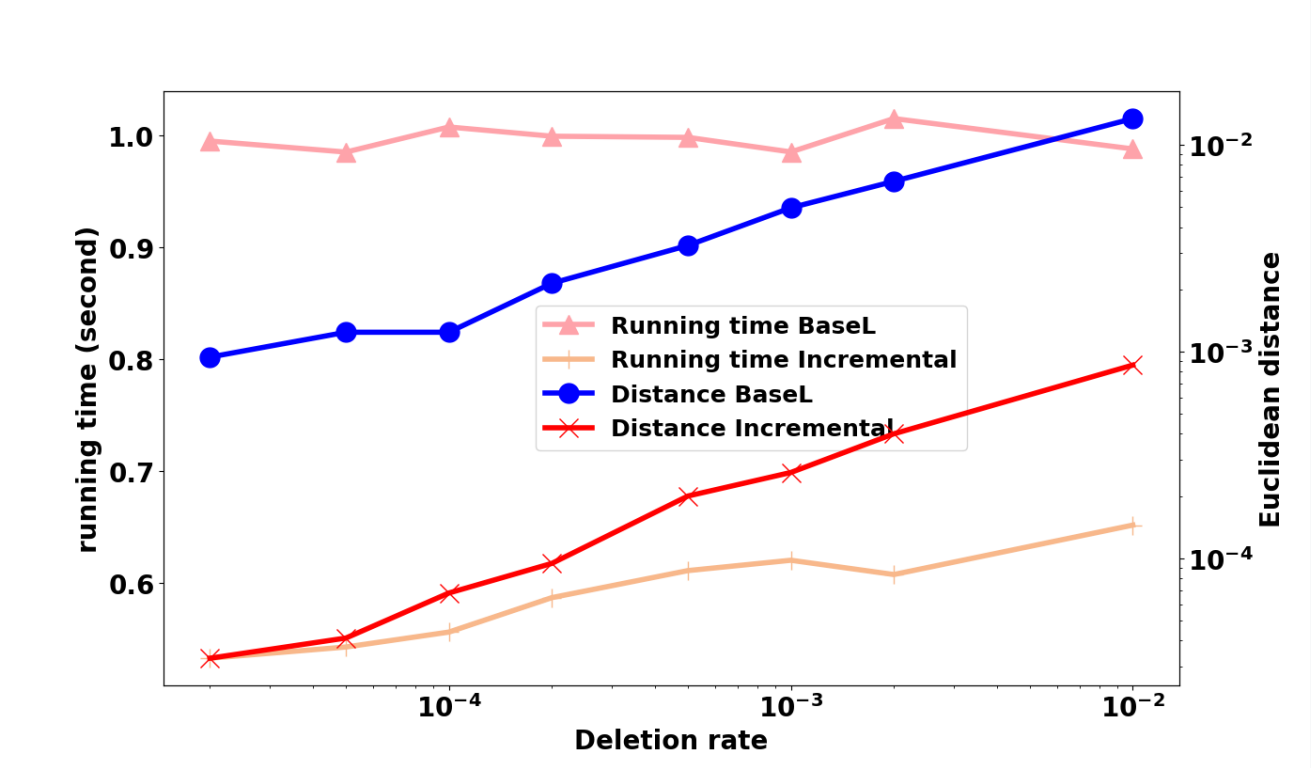
As the addition rate increases running time and distance also increase indicating a higher computational cost and a greater difference from the original data.

MNIST ADDITION



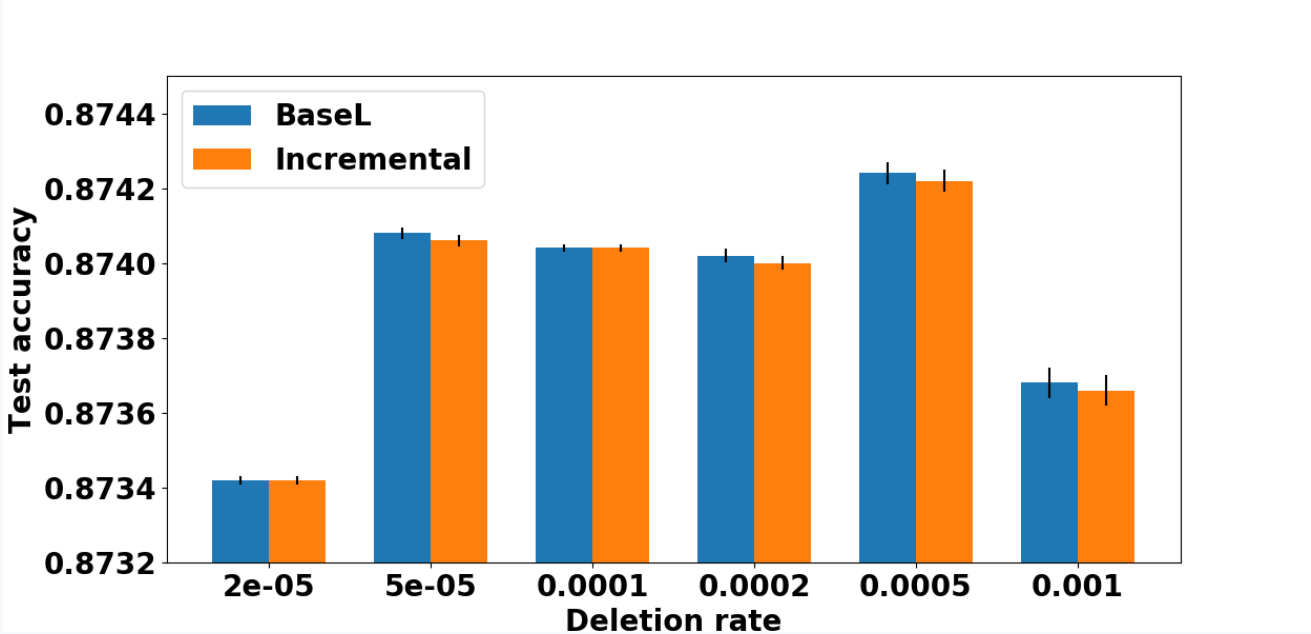
At different addition rate Incremental model shows comparable result as compared to Base model

MNIST DELETION



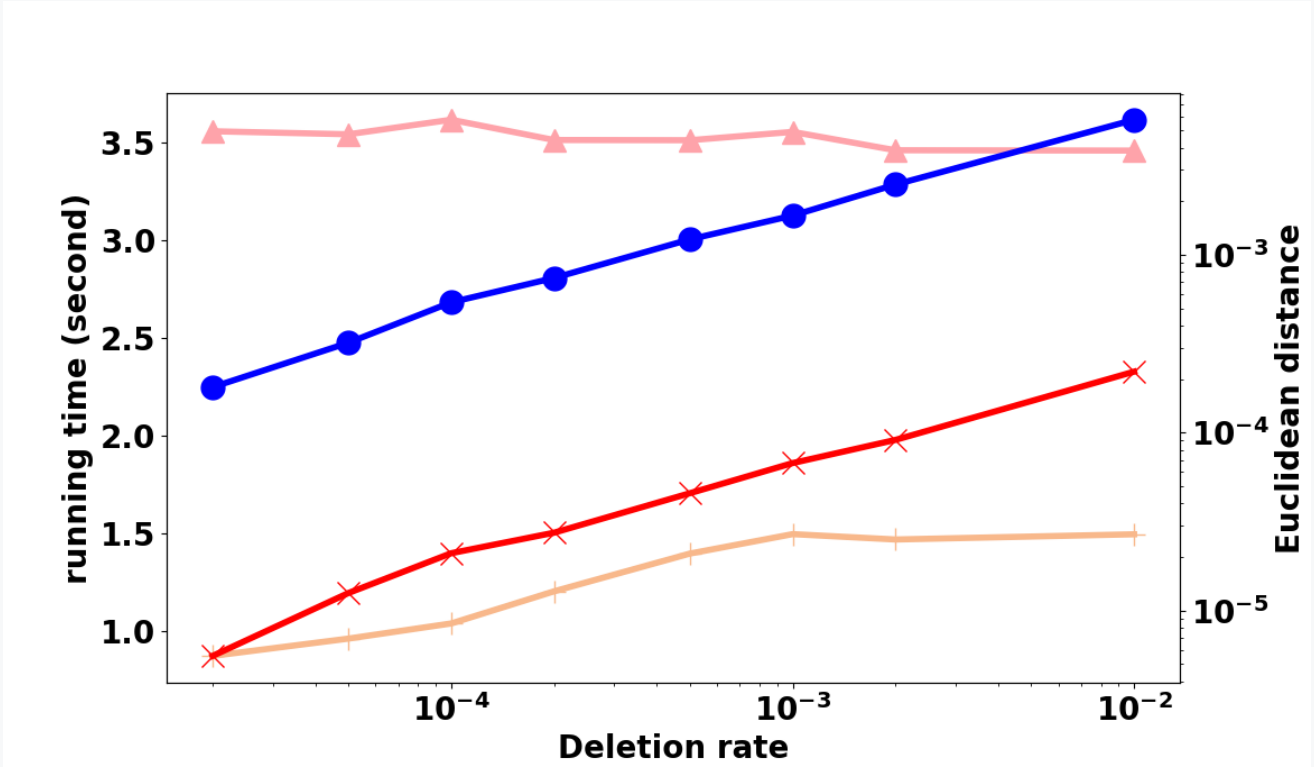
As the deletion rate increases running time and distance also increase indicating a higher computational cost and a greater difference from the original data.

MNIST DELETION



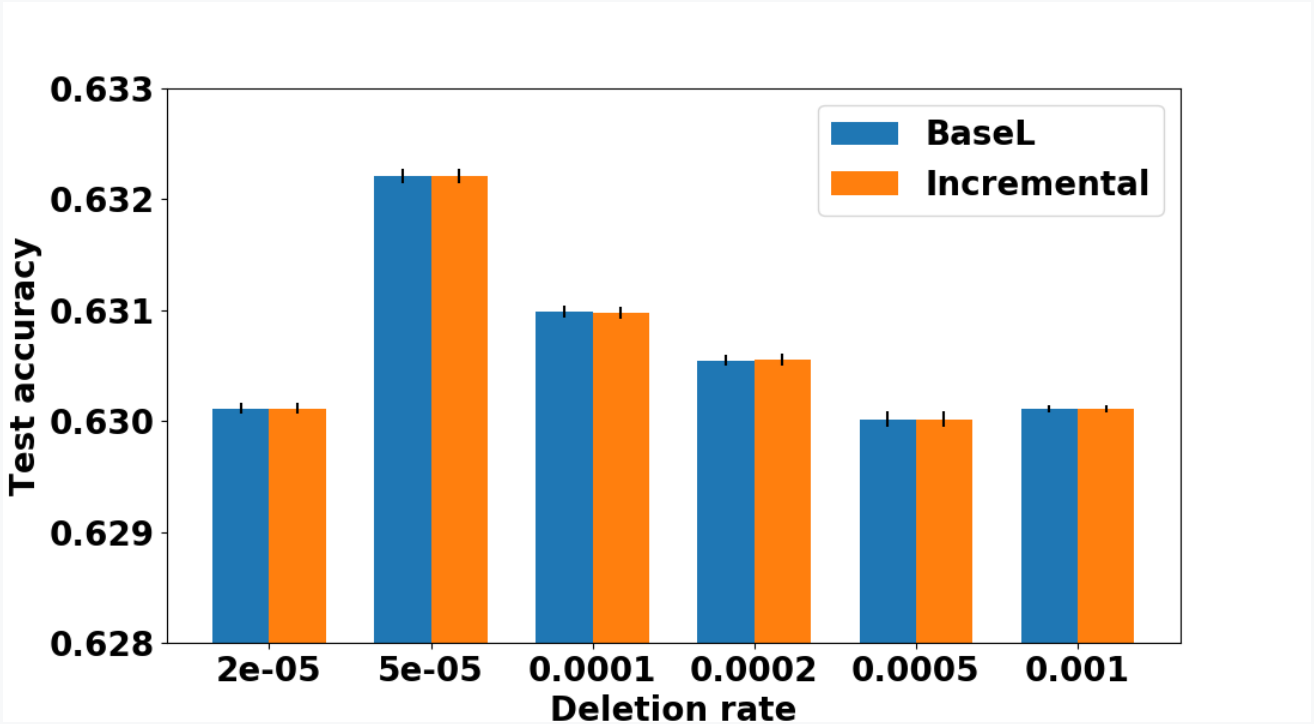
At different deletion rate Incremental model shows comparable result as compared to Base model

CIFAR10 DELETION



As the deletion rate increases running time and distance also increase indicating a higher computational cost and a greater difference from the original data.

CIFAR10 DELETION



At different deletion rate Incremental model shows comparable result as compared to Base model

Augmentation for Exact Unlearning:

1. **Error-Minimizing Noise Concept:**
   - Purpose: Augmentation leverages error-minimizing noise for exact unlearning.
   - Objective: Introduce imperceptible noise to diminish the influence of specific data points.
   - Result: Aids in refining the model without the need for complete retraining.

2. **Optimization Problem for Imperceptible Noise:**
   - Formulation Augmentation poses an optimization problem to generate imperceptible noise.
   - Objective Function: Minimizing the noise while achieving the desired reduction in data point influence.
   - Challenge: Balancing noise reduction with maintaining model performance.

3. **Link to Augmentation Paper:**
   - [Augmentation Paper]

This innovative approach in the Augmentation technique utilizes imperceptible noise to exact unlearning, providing a nuanced solution to address specific data point influences. For a deeper understanding, the Augmentation paper offers comprehensive insights into the methodology and its implications.

Impact of Noise on Accuracy:

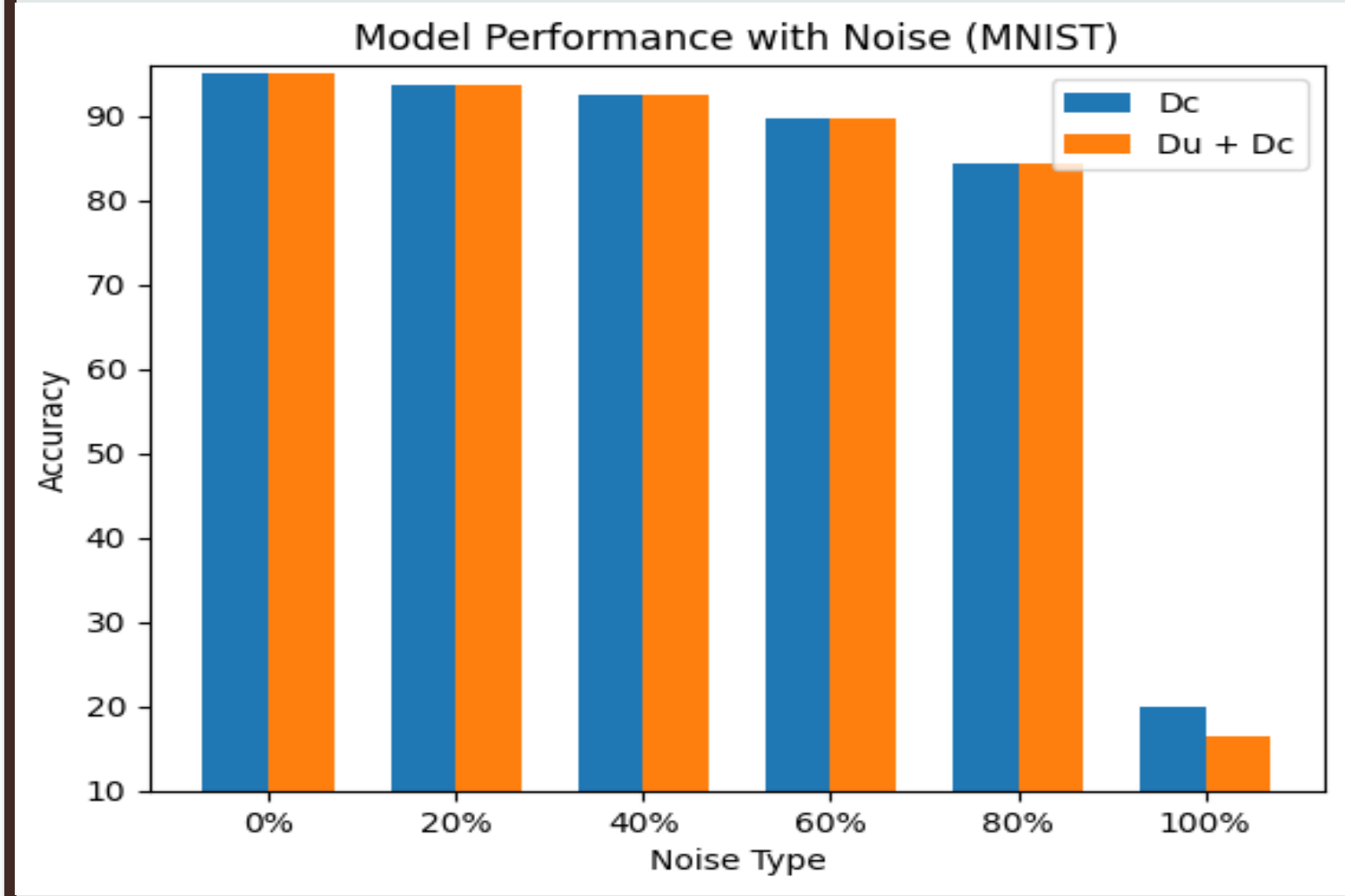
- As the percentage of unlearnable examples (Du) increases, both "Dc" and "Du + Dc" accuracies decrease.
- The decrease in accuracy is more pronounced for "Du + Dc" compared to "Dc," indicating that unlearnable examples have a more significant impact on model performance.

Threshold Effect at 80% Noise:

- There is a noticeable drop in both "Dc" and "Du + Dc" accuracies at the 80% noise level, suggesting a threshold effect where the model's performance is severely impacted beyond this point.

Importance of "Dc" in Noisy Environments:

- Despite the presence of noise, "Dc" accuracy remains relatively high, indicating the model's ability to maintain accuracy on clean examples even in the presence of noise.



Code References:

https://github.com/cleverhans-lab/machine-unlearning
https://github.com/thuwuyinjun/DeltaGrad
https://github.com/HanxunH/Unlearnable-Examples

Work Split:

1.) SISA: Prayas and Vivek
2.) Influence Distillation (Written from scratch): Prayas
3.) DeltaGrad: Saransh
4.) Augmentation aided unlearning: Prateek
5.) Overall Evaluation and Report: Prateek and Vivek