UVA CS 6316: Machine Learning : 2019 Fall
Course Project: Deep2Reproduce @
https://github.com/qiyanjun/deep2reproduce/tree/master/2019Fall

Large Margin Deep Networks for Classification

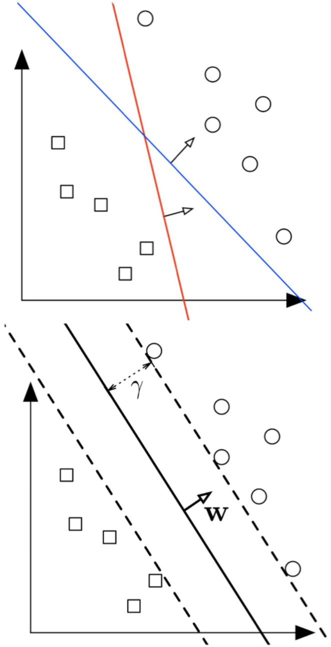
Authors - Gamaleldin F. Elsayed, Dilip Krishnan, Hossein Mobahi, Kevin Regan, Samy Bengio

Reproduced by - Sanchit Sinha(ss7mu), Rakshita Kaulgud Ramesh(rrk7pb), Arijit Pande(ap6bd), Manisha Sudhir(ms8jd)

https://arxiv.org/abs/1803.05598

Motivation

- **Large margin principle**: the optimal separating hyperplane is at the maximum distance possible from the closest points in 2 different classes
- Models built on large margin (eg. SVM):
    - Successful in avoiding overfitting and improving generalization
    - More robust to partial outliers and perturbations
- Classical Large margin: implemented well in shallow networks but not in deep networks
- **Goal of the paper**: Implement large margin principle in input space of deep networks



How large margin principle solves where to place the ideal hyperplane

Background

- **Decision Boundary**: Separation point between two different classes. At the decision boundary, there is ambiguity in class decisions.
- **Margin**: The smallest non negative distance between decision boundary and closest class point
- **Support Vector machines**: The most well known maximum margin principle based classification models - use support vectors (points closest to decision boundary) to estimate margin
- **Margins in Deep Networks**: Easy to compute in **output space**, very difficult (sometimes impossible) to compute in **input space. WHY? -** Input-Output correspondence not directly calculable because input passes through many layers and activations.

Related Work

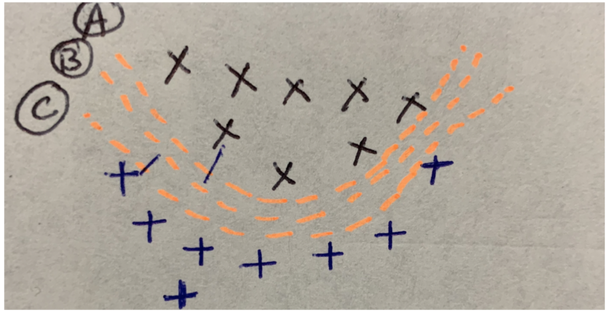
- **SVMs:** Vapnick et al. (1995) propose SVM models improving generalizability and robustness
- **Maximum margin networks:** Liu et al(2016), Sun et al.(2015) Sokolic et al (2016), Liang et al (2017), Sun et al (2015) demonstrate maximum margin principle on output space by extending classic loss functions like cross entropy.
- Sokolic et al (2016) demonstrate **Jacobian based Regularized Deep Networks** possess large margin-like properties
- Hein et al (2017) and Matyasko et al (2017) try to improve **adversarial robustness** by proposing loss functions

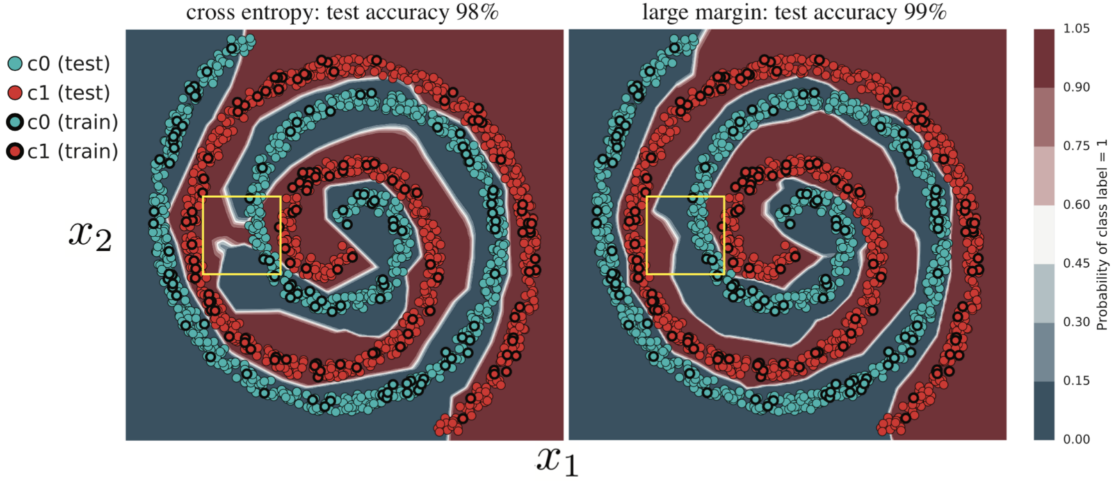
Claim / Target Task

- “Proposal of a novel network-agnostic loss function which captures the principle of large margin separation in both the input and hidden layers for deep neural networks.”
- The proposed loss function improves over standard neural networks in 3 distinct ways:
  - **Noisy labels**: A few labels in training data have their labels switched randomly
  - **Generalization**: Training on very small set of training samples
  - **Adversarial Perturbations**: Robust to both black box and white box attacks
- In general, if the model is robust to all the above attributes, it gives a good shaped decision boundary.

An Intuitive Figure Showing WHY Claim

Why a maximum margin principle is required in neural networks
- Which decision boundary is correct? A,B or C? (Ideally B)
Using maximum margin principle, we can correctly penalize boundary which are too close to either of the classes and force the model to learn the correct boundary.





If large margin is only enforced at output stage, we loose upon information in the input space, giving ambiguity on the decision boundary. The yellow square here points out the location of the poor shape of decision boundary. The main motivation of the paper is to achieve a better shape of decision boundary.

Proposed Solution

- Defining a **boundary** between two classes i and j, where the probability of sample being classified as i or j is the same. Hence,

  D{i,j} = {x | f_i(x) = f_j(x) }

- Now **margin** is defined as the smallest distance to be moved so that it reaches the decision boundary, implying a score tie. Hence,

  disp* = min ||d|| where f_i(x+d) = f_j(x+d)

- Now we have to penalize this margin for points farther away from it.

  Max{0, margin + d*sign(f_i(x) - f_cc(x))} where cc is the correct class label of point x

- Expanding this to multiple class setting, we have to aggregate this over several classes. We can take either sum or max.

Proposed Solution - contd

Once we aggregate the losses over all classes we get the final loss function as:

$$\boldsymbol{w}^* \triangleq \arg\min_{\boldsymbol{w}} \sum_k \mathscr{A}_{i\neq y_k} \max\{0, \gamma + d_{f,\boldsymbol{x}_k,\{i,y_k\}} \operatorname{sign}\left(f_i(\boldsymbol{x}_k) - f_{y_k}(\boldsymbol{x}_k)\right)\}$$

As mentioned before, **d** is not directly calculable in deep networks. The paper then proposes **linearization** to represent **d** as a function of outputs for every layer. Hence, the final loss function can be represented as:

$$\hat{\boldsymbol{w}} \triangleq \arg\min_{\boldsymbol{w}} \sum_k \mathscr{A}_{i\neq y_k} \max\{0, \gamma + \frac{f_i(\boldsymbol{x}_k) - f_{y_k}(\boldsymbol{x}_k)}{\|\nabla_{\boldsymbol{x}} f_i(\boldsymbol{x}_k) - \nabla_{\boldsymbol{x}} f_{y_k}(\boldsymbol{x}_k)\|_q}\}$$

Implementation

- A major approximation to reduce the computation time is to treat the denominator of the loss function as a constant wrt. W during backpropagation. This is argued based on obtaining an almost similar expansion using Taylor series.
- The optimizer used was Adam/SGD.
- For Imagenet dataset only the maximum class value obtained in forward pass was used i.e. only 1 class. For MNIST and CIFAR10 all the classes were used.

Data Summary

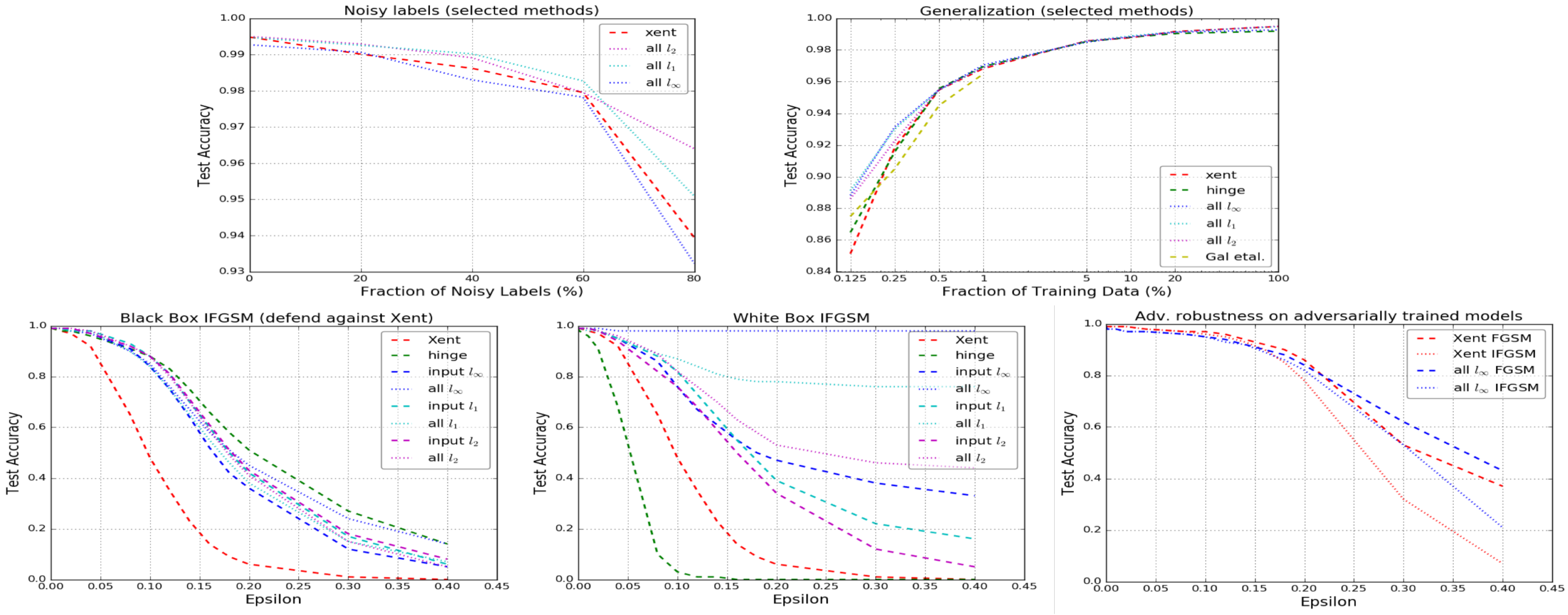
1. MNIST :
- Database of handwritten digits
- Has a training set of 60,000 samples and a test set of 10,000 samples.
- The digits have been size-normalized and centered in a fixed size image.

1. CIFAR-10
- Consists of 60,000 32*32 color images in 10 classes, with 6000 images per class.
- There are 50,000 training images and 10,000 testing images.

1. ImageNet
- Image database organized according to WorldNet hierarchy where each node of the hierarchy is depicted by hundreds and thousands of images.
- This database has over 14 million labeled images depicting 20,000+ object categories.

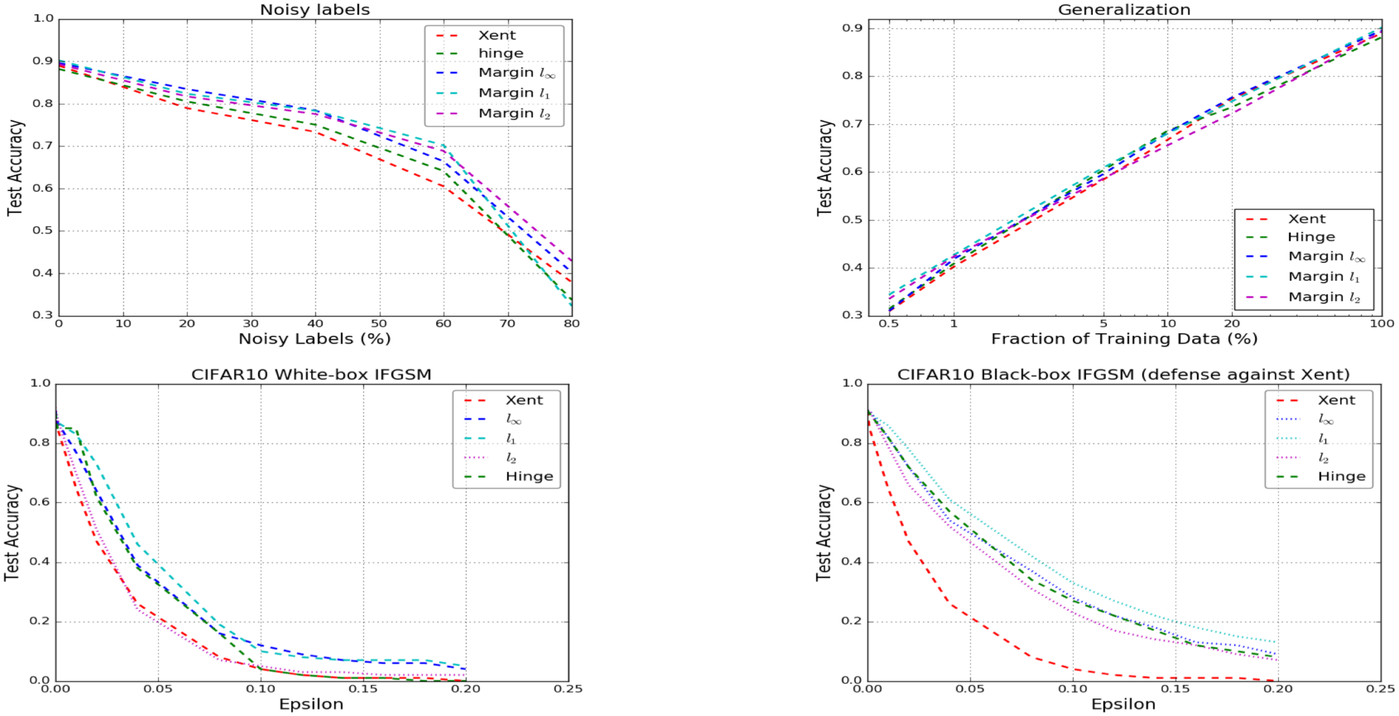
Experimental Results

Experimental results for MNIST - (Figures from top right, clockwise) - Noisy labels, generalization, black box IFGSM and White Box IFGSM, and performance on adversarially trained models



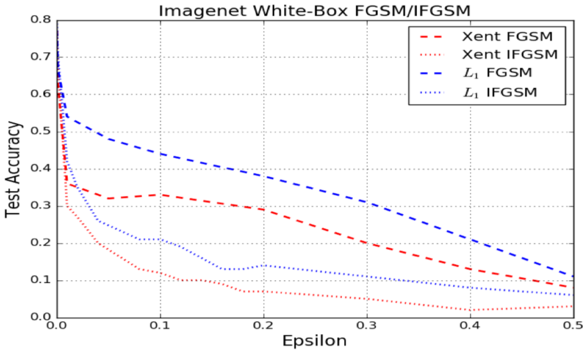
Experimental Results

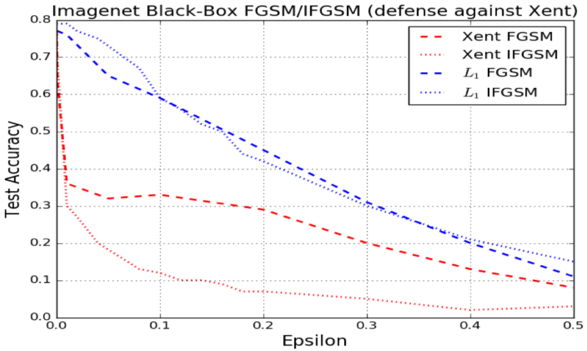
Experimental results for CIFAR-10- (Figures from top right, clockwise) - Noisy labels, generalization, black box IFGSM and White Box IFGSM



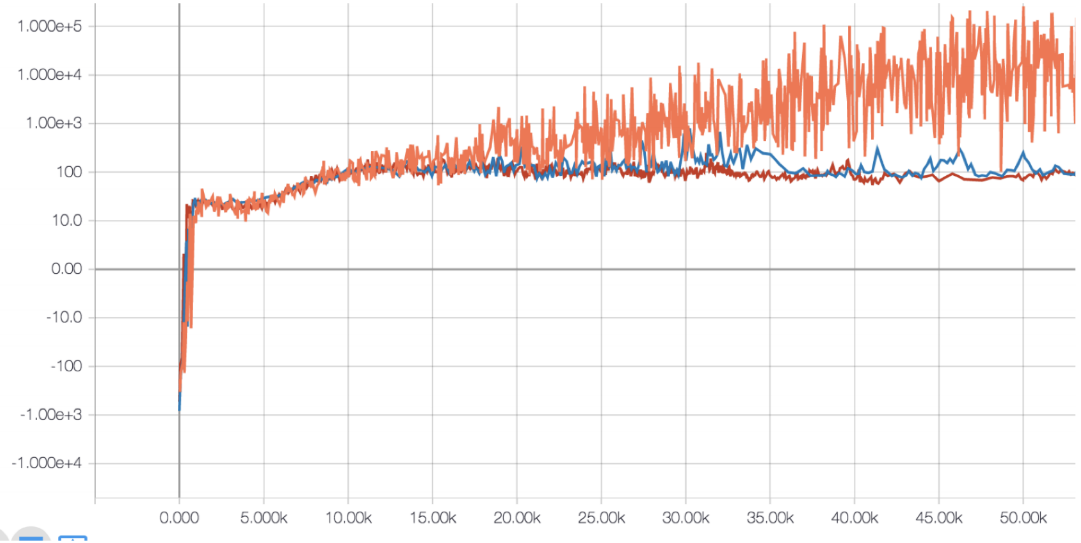
Experimental results

Experimental results for ImageNet- - ImageNet WhiteBox FGSM/IFGSM and blackbox FGSM/IFGSM defense performance

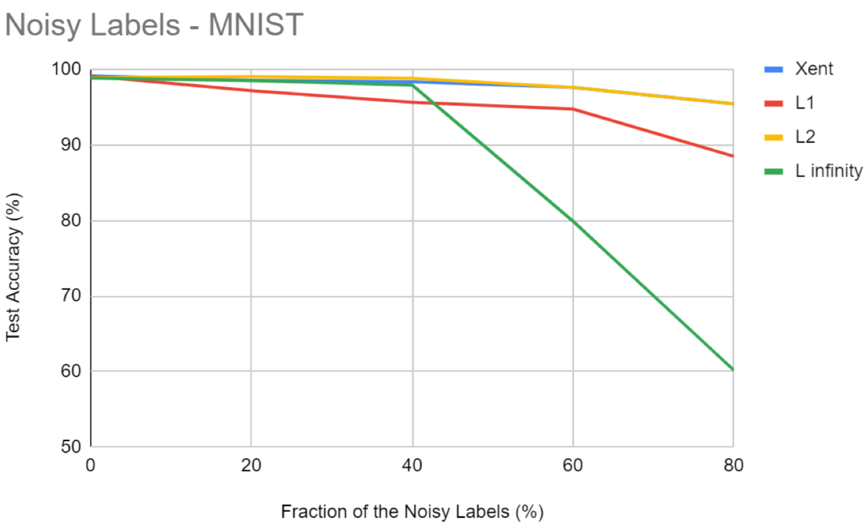


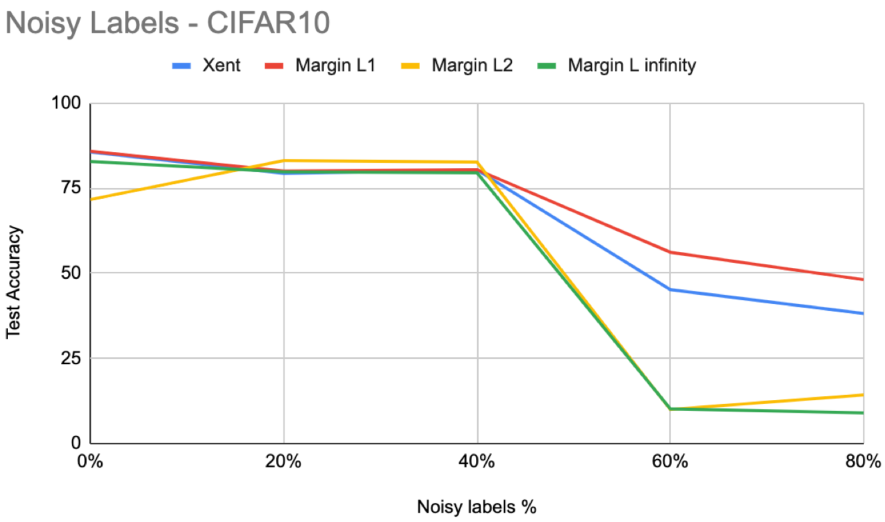


“Insane Experiments Require Insane Training.”

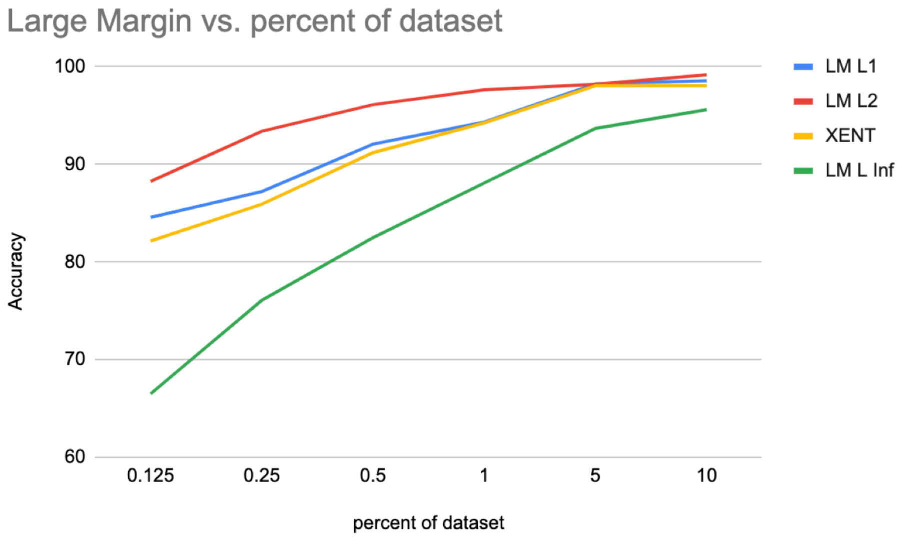


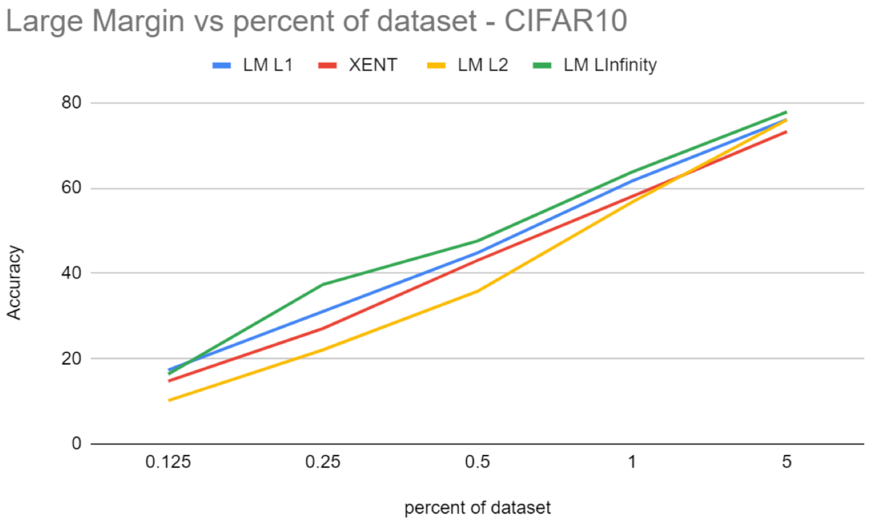
Exp1: Model performance vs Noisy Labels



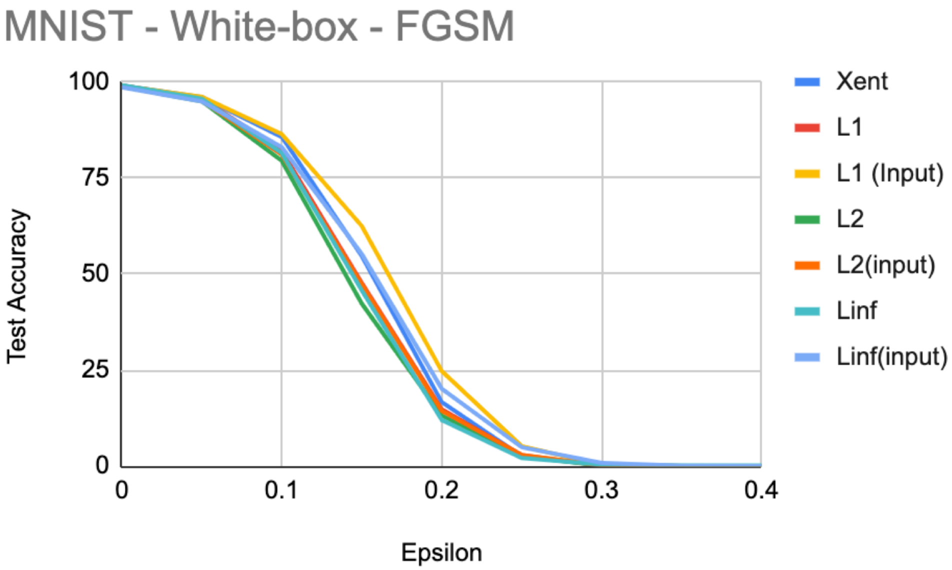


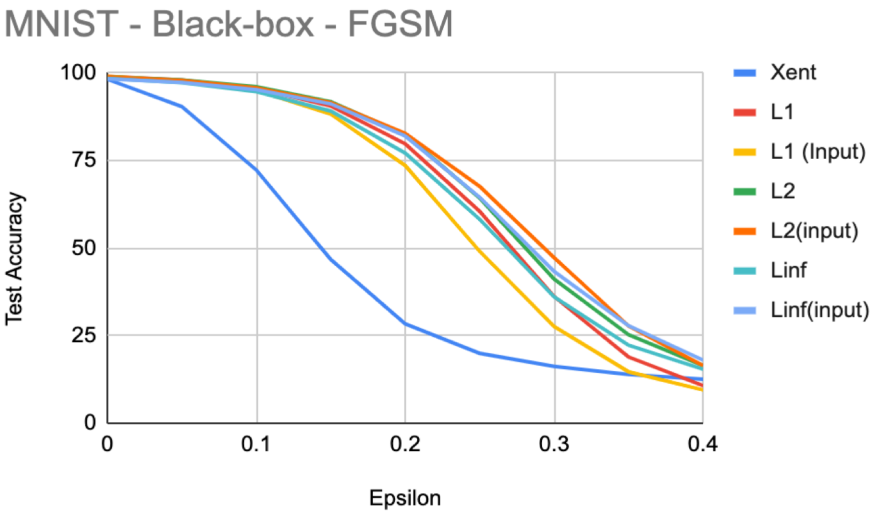
Exp 2: Large Margin Accuracy vs dataset size



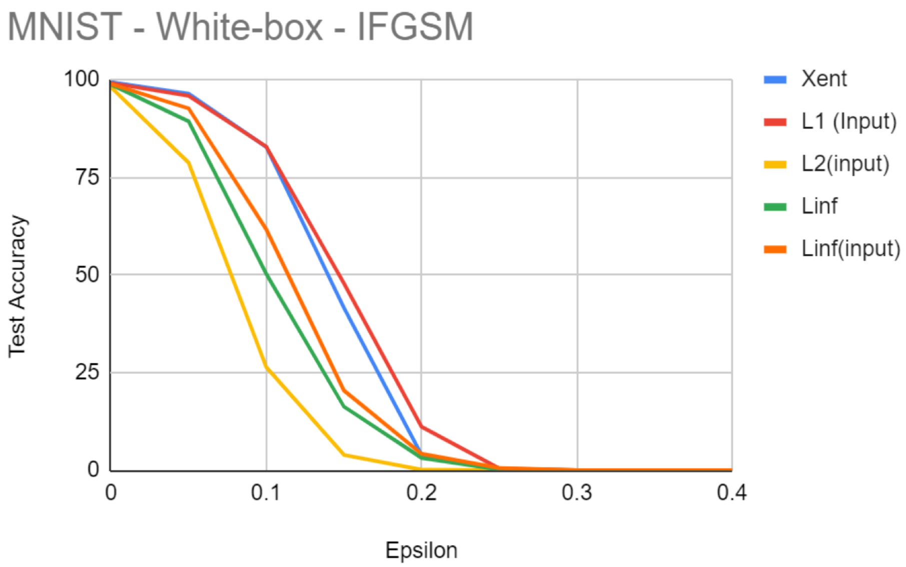


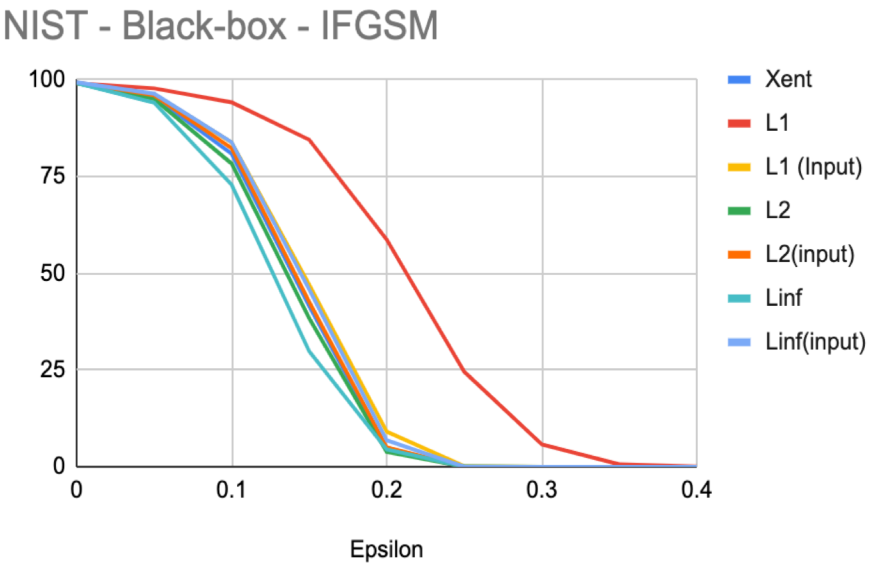
Exp 3&4: FGSM Black-box and White-box attacks


MNIST - White-box - FGSM


MNIST - Black-box - FGSM

Exp 5&6: IFGSM Black-box and White-box attacks





Experimental Analysis

MNIST:

- 4 hidden layer network with 2 convolutional layers and 2 fully connected layers.
- Margin models considered for l_∞, l_1 and l_2 norms.
- **Analysis for Noisy Labels**: Randomly switch labels of dataset items - percentage of flipped labels varies from 0 to 80% in steps of 20. l_2 model achieves highest accuracy, followed by cross-entropy
- **Generalization** : Trained data with 0.125% of the original data - 68 samples. All layer margin model outperforms cross entropy
- **Adversarial Perturbation** : Fast gradient Sign Method and iterative version of perturbation used. All margin models outperform cross entropy.

Experimental Analysis

CIFAR - 10

- Resnet model: consisting of an input convolutional layer and 3 blocks where each block contains 2 convolutional layers repeated 3 times.
- **Noisy Labels:** l_1 model outperform cross entropy. (across all the 0-80% range of flipped labels)
- **Generalization:** l_1 and l_∞ margin models perform better than cross entropy - this performance difference becomes is bigger when the training dataset size is in the region of 1 to 5%
- **Adversarial Perturbation:** Performance of cross entropy and margin models compared IFGSM attacks. l_1 and l_∞ models superior to cross entropy. ***** (Test Only)

Conclusion and Future Work

1. The paper presents a new loss function inspired by the theory of large margin that is amenable to deep network training.
2. This new loss is flexible and can establish a large margin that can be defined on input, hidden or output layers, and using l_∞, l_1, and l_2 distance definitions.
3. The method described in the paper is computationally practical: for Imagenet, training was about 1.6 times more expensive than cross-entropy.

Future Work:

1. The training on only one class in ImageNet, does not actually prove it would work well if all classes are taken into account.
2. More experiments should be done in the space of adversarial perturbations like CW. They can be added to solidify the claim of the paper.

References

1. Cortes, Corinna and Vapnik, Vladimir. Support-vector networks. Machine learning, 20(3):273–297, 1995
2. Drucker, Harris, Burges, Chris J. C., Kaufman, Linda, Smola, Alex, and Vapnik, Vladimir. Support vector regression machines. In NIPS, pp. 155–161. MIT Press, 1997
3. Hinton, Geoffrey, Srivastava, Nitish, and Swersky, Kevin. Neural networks for machine learning lecture 6a-overview of mini-batch gradient descent.
4. Kurakin, A., Goodfellow, I., and Bengio, S. Adversarial Machine Learning at Scale. ArXiv e-prints, November 2016.
5. Matyasko, Alexander and Chau, Lap-Pui. Margin maximization for robust classification using deep learning. In Neural Networks (IJCNN), 2017 International Joint Conference on, pp. 300–307. IEEE, 2017
6. Sun, Shizhao, Chen, Wei, Wang, Liwei, and Liu, Tie-Yan. Large margin deep neural networks: Theory and algorithms. CoRR, abs/1506.05232, 2015.

Work Distribution

Sanchit Sinha - Loss Function, MNIST, FGSM, IFGSM, CIFAR10, powerpoint presentation

Arjit Pande - Reduced Dataset Experiment, powerpoint presentation, graphs for experiments

Rakshita Kaulgud Ramesh - Noisy Labels for CIFAR10, Powerpoint presentation, graphs for experiments

Manisha Sudhir - Noisy Labels for MNIST, Powerpoint presentation, graphs for experiments