Project Deliverables
MixMatch: A Holistic Approach to Semi-Supervised Learning

Team Name : ~
P Harshavardhan - 2021111003
Kapil Rajesh Kavitha - 2021101028

INTRODUCTION

Recent advances in training deep neural networks heavily rely on large labeled datasets, which can be costly and time-consuming to acquire, particularly in fields like medicine. Semi-supervised learning (SSL) aims to mitigate this by leveraging unlabeled data, typically through loss terms encouraging confident predictions (entropy minimization), output consistency under perturbations (consistency regularization), or generalization (generic regularization). This paper introduces MixMatch, a novel SSL algorithm that unifies these approaches into a single loss function. MixMatch achieves state-of-the-art results across standard image benchmarks, notably reducing CIFAR-10 error rates by a factor of 4. Additionally, MixMatch enhances privacy-accuracy trade-offs in differentially private learning, making it a versatile SSL solution.

PROJECT DELIVERABLES

- **Algorithm Implementation:** To develop a robust implementation of the MixMatch algorithm, ensuring it adeptly handles both labeled and unlabeled data while incorporating techniques like data augmentation, label guessing, and consistency regularization.
- **Performance Evaluation:** To conduct a thorough evaluation of the MixMatch algorithm by running experiments on various datasets featuring different ratios of labeled to unlabeled data. Our analysis would involve comparing MixMatch's performance with that of other semi-supervised methods, reporting on key metrics such as accuracy.
- **Extended Experiments:** To deepen our understanding, we will look at expanding upon the experiments conducted in the paper. This involves the creation of artificial datasets with varying labeled data proportions for diverse datasets. We shall train models on these modified datasets and closely examine performance trends.
- **Insights and Analysis:** Our project will aim to provide valuable insights into observed performance drop-offs across different datasets as labeled data proportions vary. We would analyze these trends, identify noteworthy patterns, limitations, and key findings that will contribute to the broader knowledge base in semi-supervised learning.
- **Documentation and Reporting:** As we progress, we will work on maintaining comprehensive documentation, encompassing code documentation, detailed experiment

setups, results, and in-depth analysis. Ultimately, we will present our findings in a well-structured report or presentation format, ensuring clarity and ease of reference.

POTENTIAL DATASETS (expected to be refined)

- **CIFAR-10** : cifar10 | TensorFlow Datasets - A dataset of 60000 images in 10 classes.
- **CIFAR-100** : cifar100 · Datasets at Hugging Face - A dataset of 60000 images in 100 classes. There are two labels per image - fine label (actual class) and coarse label (superclass)
- **STL-10** : stl10 | TensorFlow Datasets - A dataset used for developing unsupervised learning algorithms. It is inspired from the CIFAR-10 dataset - each class has fewer labeled training examples than in CIFAR-10, but also has a very large set of unlabeled examples.
- **SVHN** (Street View House Numbers) : svhn · Datasets at Hugging Face - A dataset obtained from house numbers in Google Street View images. It consists of over 600k images in 10 classes (one for each digit)
- **The Oxford-IIIT Pet Dataset**: oxford_iiit_pet | TensorFlow Datasets - A 37 category pet dataset with around 200 images in each class. The images have large variations in scale, pose and lighting.
- **MNIST** : mnist | TensorFlow Datasets - A large dataset of handwritten digits, with around 70000 images.
- **Caltech-101** : caltech101 | TensorFlow Datasets - A dataset consisting of pictures of objects in 101 classes and 1 background clutter class, with 40-800 images in each class (around 9000 images in total)
- **Fashion-MNIST**: fashion_mnist · Datasets at Hugging Face - A dataset of Zalando's article images with around 70000 examples split into 10 classes (similar to the original MNIST)
- **ImageNet:** ImageNet (image-net.org) - A subset of this dataset shall be used as part of the project.

TIMELINE

- Oct 8 - Oct 12 : Understanding the paper (algorithms and methodology in MixMatch and MixUp)
- Oct 13 - Oct 21 : Implementation of MixMatch algorithm
- Oct 22 - Oct 27 : Implementation of the MixUp algorithm to train with MixMatch
- Oct 28 - Oct 29 : Testing and Debugging to get the entire pipeline working
- Oct 30 : Documentation on the implementation of MixMatch algorithm
- Oct 31 - Nov 3 : Running the other standard methods mentioned in the paper (for comparison with MixMatch)

- Nov 4 - Nov 10 : Testing and validating with experiments mentioned in the paper
- Nov 11 : Documentation on observed results, comparison to results provided in the MixMatch paper
- Nov 12 - Nov 13 : Extended experiments planning
- Nov 13 - Nov 20 : Experimentation on other datasets (mentioned in the project datasets)
- Nov 21 - Nov 22 : Preparation of Additional datasets (our creation)
- Nov 23 - Nov 26 : Testing on our datasets
- Nov 27 : Documentation on additional datasets and process of creation of our own dataset

LINK TO ~'s DELIVERABLE DOCUMENT:
https://docs.google.com/document/d/1ZbaB3BjfoKef_FmBkAG4r6sub0n9lVDg4jhJmYls6Xw/edit?usp=sharing