**Technical Blog for research paper -‘MIXUP: BEYOND EMPIRICAL RISK MINIMIZATION’**

**PURPOSE**: My purpose for writing this technical blog is to cover and convert all technical aspects of the MIXUP research paper, into blog post which is simple and understandable for all readers.

**Original Paper :**  <https://arxiv.org/pdf/1710.09412.pdf>

**PREREQUISITES**:

1. **Supervised Learning**: In simple words , it is the [machine learning](https://en.wikipedia.org/wiki/Machine_learning) task of **learning a function** that [maps](https://en.wikipedia.org/wiki/Map_(mathematics)) an input to an output based on example input-output pairs. By learning a function, it means, it infers a function from **labelled** training data (ex. of data can be images, text, videos). Here labelling is nothing but some sort of tagging to the given data (ex: if an email is spam or not spam)
2. **Empirical Risk Minimization: (Regular Binary Classification)** With this principle we cannot know exactly how well an algorithm will work in practice (**the true "risk"**) because we don't know the true distribution of data that the algorithm will work on, but we can instead **measure its performance** on a known set of training data (**the "empirical" risk**). And by minimization it means to give theoretical bounds on its performance which minimizes the risk.
3. **Convex Combination**: It is a [linear combination](https://en.wikipedia.org/wiki/Linear_combination) of [points](https://en.wikipedia.org/wiki/Point_(geometry)) where all [coefficients](https://en.wikipedia.org/wiki/Coefficients) are [non-negative](https://en.wikipedia.org/wiki/Non-negative) and sum to 1. In simple words, it’s equivalent to a standard [weighted average](https://en.wikipedia.org/wiki/Weighted_average), but whose weights are expressed as a percent of the total weight.
4. **Data Augmentation:** It is a set of techniques to artificially increase the amount of data by generating new data points from existing data. In simple terms , **adding new samples of minority class.**
5. **Deep neural networks (DNN)**: It is a class of machine learning algorithms similar to the artificial neural network and aims to mimic the information processing of the brain.
6. **Adversarial examples :** Examples just outside the training distribution.
7. **Vicinal Risk Minimization (VRM)** : In VRM, **human knowledge** is required to describe a vicinity or neighborhood around each example in the training data.
8. **Alpha** : The strength of mixup interpolation is indicated by alpha
9. **Lambda**: It is a random sample from the beta distribution defined by alpha

**INTRODUCTION**:

* With given prerequisites, we can move onto the introduction of the MIXUP technique and understanding how it is better than the traditional ERM (Empirical Risk Minimization) technique.
* Large deep neural networks have some issues such as memorization and sensitivity to adversarial examples. Mixup is solving this very issue.
* DNN learning rule is to minimize average error over the training data. This is called as empirical risk minimization (ERM).
* As mentioned in the intro, ERM is unable to explain or generalize to test data that differs even slightly from that of the training data distribution.
* Hence a better alternative to ERM is Data Augmentation technique called MIXUP (a form of VRM)
* In simple terms what MIXUP is doing :
  + Ex: lets say we have 2 data points (images)
  + (y0 – cat) -> (1,0) : (cat:dog) and (y1 – dog) -> (0,1) : (cat:dog)
  + Mixup will overlay these two datapoints (images) on top of each other.
  + So, what’s done is we have changed the label distribution from yA -> {0,1} to yA -> U(0,1) i.e changing from binary to a continuous distribution of any real number between 0 to 1.
  + Thus the new datapoint generated is nether cat or dog but is somewhat in the middle.
  + Hence the Neural network no longer has to predict if it’s a cat or a dog, but now the prediction will be how much percentage of the new image looks like a cat or dog.
* To conclude, via MIXUP , linear interpolation of feature vectors should lead to linear interpolation of associated targets.

**ERM TO MIXUP** :

* Learning the function f by minimizing (l-loss function) is known as the Empirical Risk Minimization (ERM). But, ERM monitors loss only at finite points where the labeled data exists.
* The only trivial way to minimize loss function if we have function f with large number of parameters is to MEMORIZE the training data. But as discussed in intro, memorization , in turn, leads to the undesirable behavior of f outside the training data.
* In this paper, generic vicinal distribution, called mixup is proposed where sampling from the mixup vicinal distribution produces virtual feature-target vectors.
* The core idea behind using MIXUP is it smoothens the prediction errors.
* As it’s clearly evident from the table with Top-1 Error and Top-5 error comparison for Mixup vs ERM, Mixup is always doing better i.r errors are less in Mixup.

**EXPERIMENTS on various Datasets / with various aspects :**

* IMAGENET CLASSIFICATION :
  + ERM is evaluated with input generated from combinations and is proved to be inaccurate where the one trained with mixup worked accurately. ERM works when lambda is close to 0 or 1 but degrades as it approaches 0.5.
  + For mixup, we find that α ∈ [0.1, 0.4] leads to improved performance over ERM, whereas for large α,mixup leads to underfitting.
* CIFAR-10 AND CIFAR-100 :
  + The models trained using mixup significantly outperform their analogues trained with ERM as with mixup, the convergence speed remained the same and performance improved.
* SPEECH DATA:
  + Especially for the model with larger capacity (VGG-11), mixup outperforms ERM
* MEMORIZATION OF CORRUPTED LABELS:
  + ERM models are sensitive to adversarial examples. Adversarial noise is generated by ascending the gradient of the loss surface with respect to the original example.
* ROBUSTNESS TO ADVERSARIAL EXAMPLES:
  + Mixup improves robustness of models to adversarial examples (topic of active research).
* TABULAR DATA (UCI dataset) :
  + For non-image data, mixup improves the average test error on four out of the six considered datasets, and never underperforms ERM.

**FINAL THOUGHTS and CONCLUSION :**

* The model trained with mixup is more stable in terms of model predictions and gradient norms in-between training samples.
* Mixup is the best data augmentation method and is significantly better than the second best method (mix input + label smoothing).
* Mixup improves generalisation error on various datasets with various objectives
* Mixup also opens up several possibilities for further exploration which means it’s just the beginning.