UNDERSTANDING MIXUP TRAINING METHODS

TSVAKI JETINA JULIET

NOVOSIBIRSK STATE UNIVERSITY

03 MARCH 2020

Introduction

Deep Neural Networks have made breakthroughs progress have more parameters than the training data, which allows the neural network to overfit any training data.

There are many ways to avoid the overfitting of datasets with such huge parameters. These methods can be roughly divided into two categories:

- data augmentation methods
- regularization methods.

Data augmentation

Data augmentation is a strategy that enables practitioners to significantly increase the diversity of data available for training models, without actually collecting new data.

The data augmentation methods allow the neural network to train on more samples to avoid it remembering certain samples

Data augmentation techniques such as cropping, padding, and horizontal flipping are commonly used to train large neural networks.

Other methods include

- DisturbLabel
- SamplePairing
- Mixup
- PatchShuffle

Regularization

Regularization is a technique which makes slight modifications to the learning algorithm such that the model generalizes better.

The regularization methods can reduce the complexity of the network model by limiting or adjusting the model parameters.

The regularization methods include the following

- Dropout
- Dropconnect
- Stochastic depth
- Swapout

Mixup

Mixup is a neural network training method that generates new samples by linear interpolation of multiple samples and their labels.

The mixup training method has better generalization ability than the traditional empirical risk minimization method (ERM).

What need to be known about Mixup

- Understanding of why mixup will perform better.
- How mixup works as a data augmentation method and how it regularizes neural networks.
- Visualize the loss functions of mixup and ERM training methods

Objective 1

1. General Mixup

Mixup constructs virtual training examples

$$\hat{x} = \lambda x_i + (1 - \lambda) x_j \tag{1}$$

$$\hat{y} = \lambda y_i + (1 - \lambda) y_j \tag{2}$$

where x_i, x_j are raw input vectors and y_i, y_j are one-hot label encodings

(x_i, y_i) and (x_j, y_j) are two examples drawn at random from our training data, and $\lambda \in [0, 1]$

$\lambda \sim \mathit{Beta}(\alpha, \alpha)$ $\alpha \in (0, \infty)$

- Exploring impact of the mixing of images and their labels on the performance of the network

Either interpolate the images or interpolate their labels, or interpolate them at the same time.

The Uniform distribution can better control the range of λ compared to the Beta distribution, and has the same effect on some datasets.

We use λ_x to represent the mixing ratio of two samples x for $\lambda \in$ Uniform $(\lambda_1; \lambda_2)$, and R_l to represent the mixing ratio of two labels y, where $0 \leq \lambda_1 \leq \lambda_2 \leq 1$

Solution

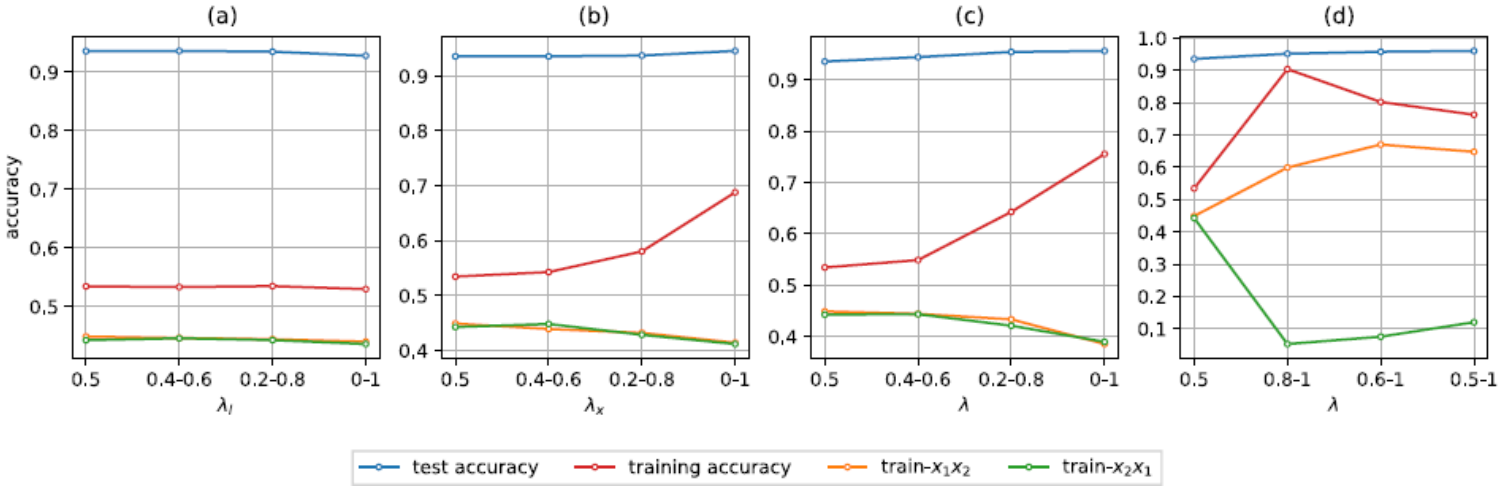

Figure 1: Impact of mixing ratios

In Figure 1.a $\lambda_x = 0.5$ so that λ_l is mixed in different ratios, which means that the same sample will correspond to a different label.It can be observed that network performance decreases with increasing range of λ_l

Figure 1.b $\lambda_l = 0.5$ and changing λ_x , which means that the linearly interpolated samples correspond to the same label, we can see that the network performance increases with the increase of the λ_x range.

In Figure 1.c, $\lambda_l = \lambda_x$, that is, the mixing ratio between the sample and their label is the same It can be seen from the figure that the performance of the network increases as the range of λ changes.

In Figure 1.d the performance of the network is higher than that of the only one, and the network performance is almost the same when λ is in the range of [0.5,1] and [0,1].

This shows that neural networks optimize each category alternately when training multiple samples and labels simultaneously. That is, in a forward process, when the network parameter is adjusted to a category, the other category plays a regularization role.

Objective 2

1. MIXUP AND ERM'S MULTI-CATEGORY TEST

   To further compare the generalization ability of mixups and ERMs for mixed samples, we use mixup and ERM methods to test the trained networks with different mixing ratios respectively.

   In this experiment, λ takes values from 0 to 1 in steps of 0.01. The experiment was run on CIFAR-10, using PreAct-Resnet . Since λ has 100 values, we will test 100 steps. The single-class prediction accuracy and multi-class prediction accuracy of the network are shown in Figure 2.

   From Figure 2, we can see that with the increase of λ , the accuracy of sample x1 increases with "S" curve and the accuracy of sample x2 decreases with "S" curve, and test-x1x2 and test-x2x1 are also similar.

Solution

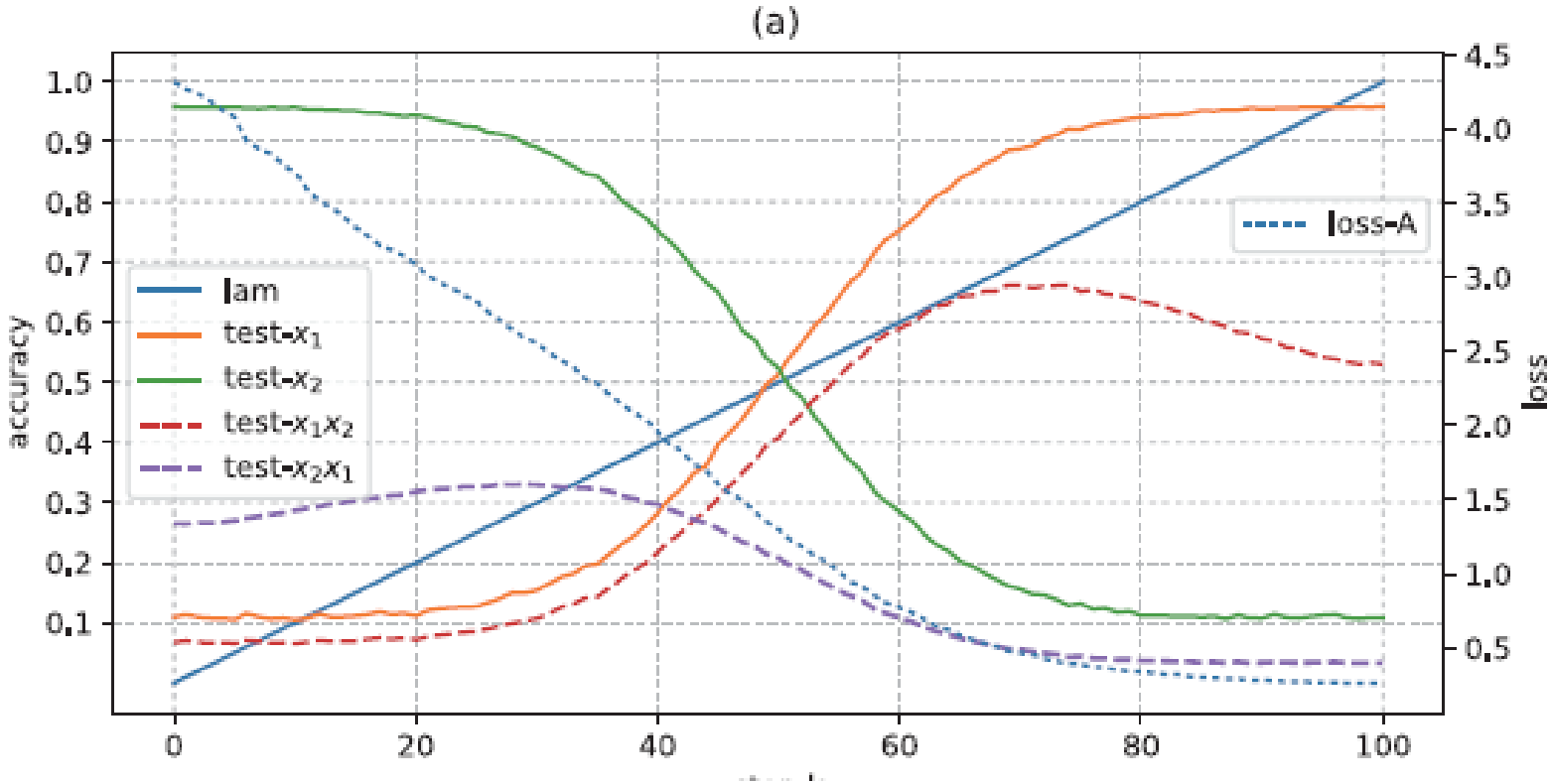

Figure 2: Mixup

The difference is that the curve of the accuracy of the mixup varies steeply at 20-80 steps, while the ERM is relatively flat, indicating that mixup will perform better than ERM, both in a single category and in multiple categories

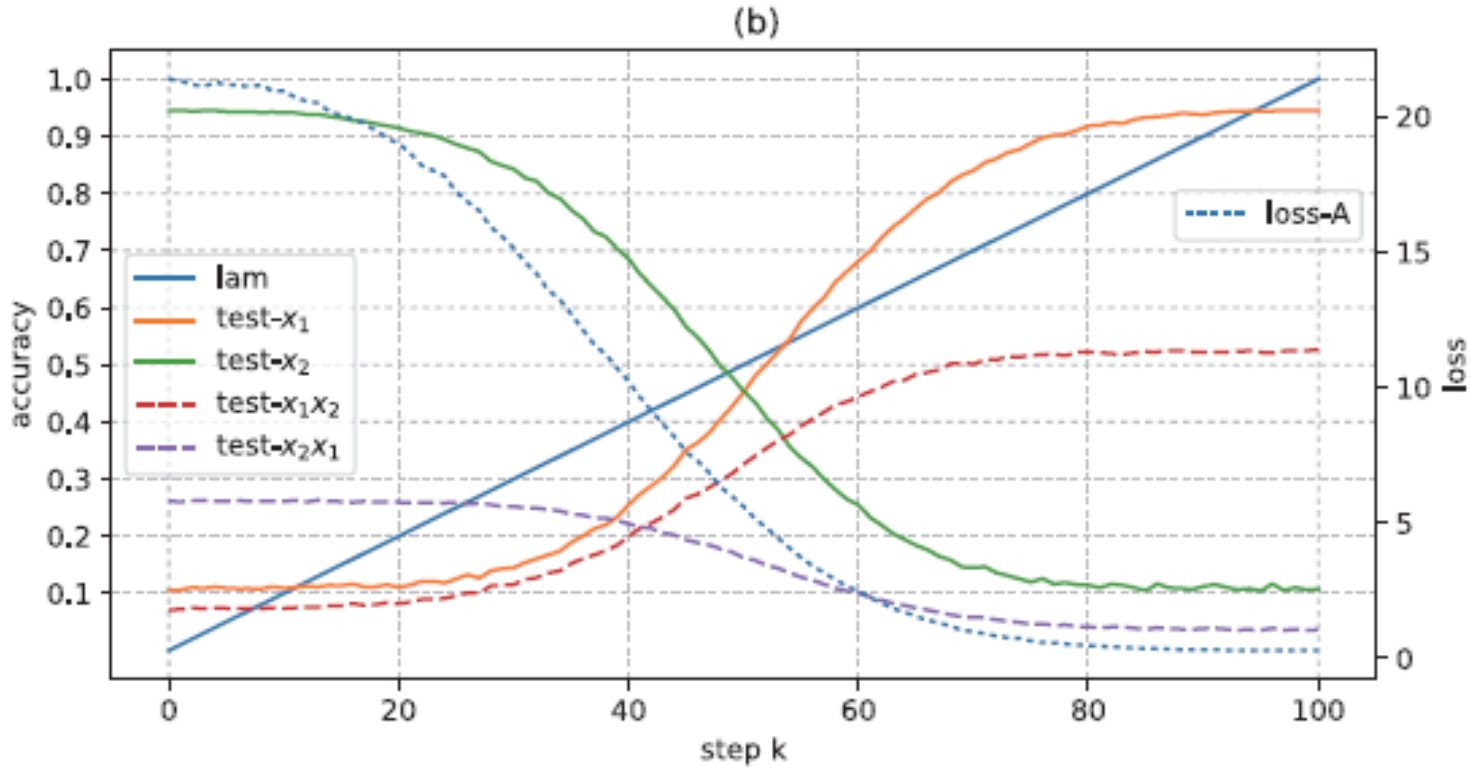

Figure 3: ERM

In about 70 steps, the test- x_1x_2 accuracy of both methods peaked. The difference is that the prediction accuracy of the mixup (Figure 2.a red dotted line) begins to decline, while the ERM is the same.

The reason being when λ in [0.3,0.7], the mixup learns two categories at the same time, and beyond this range, the training mode is dominated by the category with larger label information and becomes the ERM training mode

This phenomenon also illustrates the difference between the mixup and ERM decision surfaces.

This shows that the loss function of mixup is smoother in multicategory prediction, while the loss of ERM is larger.

This further shows that mixup has a smoother decision surface, which makes it easier to predict the interpolation between multiple categories.

Objective 3

1. VISUALIZATION OF LOSS FUNCTIONS

The visualization of the loss function of different training methods helps to understand the decision-making behavior of neural networks.

Visualization of the loss function of the two training methods, mixup and ERM

The method considers perturbing the parameters of the network starting from an appropriate random tensor and gradually adjusting its perturbation range.

When the network's loss value reaches a predetermined size, save these random tensors. In the second stage, we use the network parameters and these random tensors for linear interpolation to get the different loss values of the network

On the network architecture,we use trained PreAct-Resnet or VGG-11 models because they have different architectures

Solution

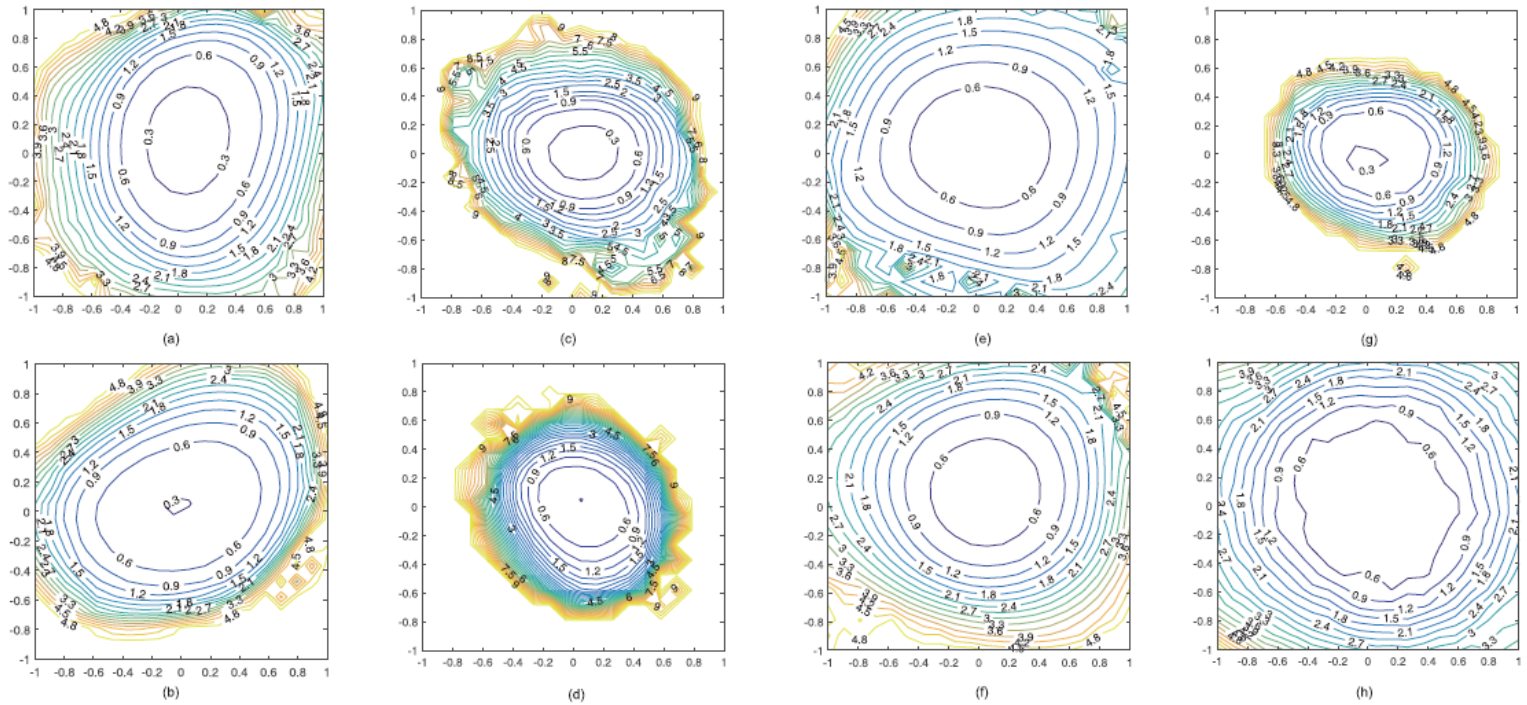

FIGURE 3. Visualize networks using different training methods and different architectures. The title of each subfigure contains the training method, network, maximum loss, and test accuracy. The subfigure (g) and (h) show the results of visualizing network loss using the method in [33]. (a) Mixup, PreAct ResNet-18, L = 5, 95.8%. (b) ERM, PreAct ResNet-18, L = 5, 94.64%. (c) Mixup, PreAct ResNet-18, L = 10, 95.8%. (d) ERM, PreAct ResNet-18, L = 10, 94.64%. (e) Mixup, VGG-11, L = 5, 92.65%. (f) ERM, VGG-11, L = 5, 91.67%. (g) Mixup, PreAct ResNet-18, L = 5, 95.8%. (h) Mixup, VGG-11, L = 5, 92.65%.

Figure 4: Visualizing network

It can be seen from Figure 4 that the loss functions of the two training methods, mixup and ERM, are very similar. However, the size of the basin near the local minimum of the mixup is much larger than that of the ERM, indicating that the loss surface near the local minima of the mixup is smoother and the ERM is relatively sharp.

The more flat the decision surface, the more conducive to the network to predict the interpolation between samples. This also allows the network to have a better robustness against the attack by adversarial sample

This phenomenon also shows that our visualization method can better project the loss function onto the disturbed weight tensor instead of the random weight tensor, which is more practical.

Perfomance of Mixup

Dataset	Model	ERM	mixup-C	mixup-H	mixup-HV	mixup-HC
CIFAR-10	PreAct ResNet-18	5.6	3.9	3.9	3.4	3.5
	DenseNet-BC-190	3.7	2.7	2.7	2.3	2.3
	ResNext29-8-64	3.7	2.7	-	2.4	-
CIFAR-100	PreAct ResNet-18	25.6	21.1	21	19.8	19.9
	ResNext29-8-64	17.4	16.8	-	14.5	-
ImageNet	ResNet-50	23.5	23.3	23.3	23	23.1
	ResNet-101	22.1	21.5	21.7	20.3	20.5
	ResNeXt-101 64×4d	20.4	19.8	19.9	19.3	19.3

Figure 5: Test error comparisons on ERM and Mixup

Conclusion

Analyzing the effect of linear interpolation of samples and their labels on the generalization performance of neural networks, and nd that mixup is better at separating multiple categories at the same time.

We propose a method to visualize the loss function of neural network by weighting noise perturbations.By visualizing the loss function of the mixup training method, we nd that the classication decision-making surface of the mixup is smoother than the ERM,

Finally, our experiments prove that the combination of multiple mixup training methods can further improve the generalization performance of neural networks,

References

- Understanding Mixup Training Methods DAOJUN LIANG 1, FENG YANG 1, TIAN ZHANG 1, AND PETER YANG2
- mixup: BEYOND EMPIRICAL RISK MINIMIZATION Hongyi Zhang MIT Moustapha Cisse, Yann N. Dauphin, David Lopez-PazFAIR