Mixup: Beyond Empirical Risk Minimization

Why Mixup?

Existing DNN must suffer from overfitting.

Their success is possible by over-parameterization.

Empirical risk minimization

Empirical Risk Minimization (minimizing the average training error) requires more data than model parameters.

This implies that our DNN using DRM has no guarantee to be converged.

Why Mixup?

ERM + DNN memorizes training dataset.

BAD

As an alternative, Vicinal Risk Minimization (VRM) has been proposed.

This is basically data augmentation.

Utilizing the neighborhood for increasing the training dataset.

Data Augmentation: Mixing images versus Mixup

Mixing images

Mixed images map to the original labels.
(Following the label of one of two images)

Mixup

Mixed images map to different labels!
(Creating a new label for a mixed image)

How to do Mixup?

Mixup

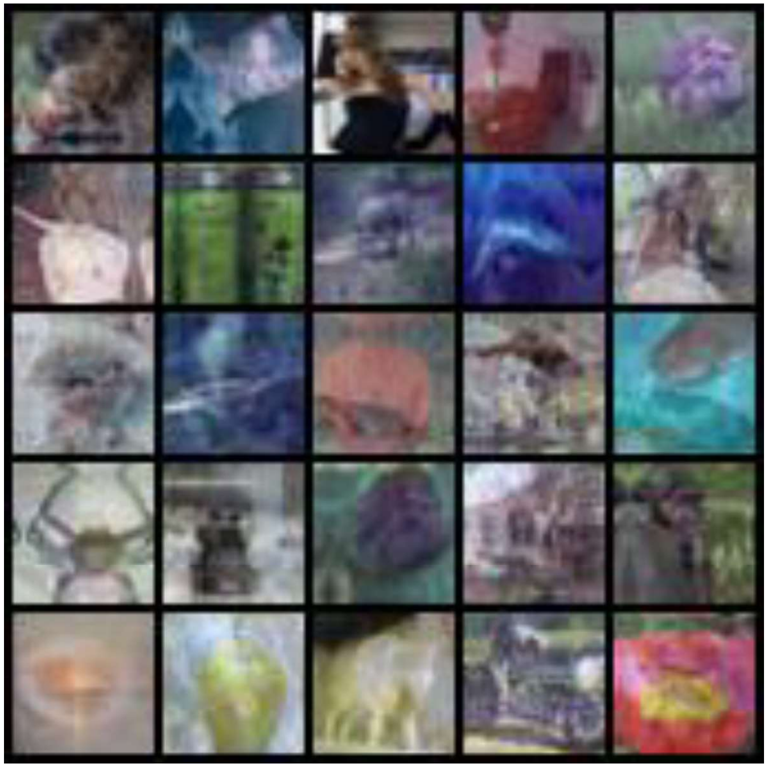
Select two images x_i, x_j when their labels are y_i, y_j

Generate mixed image and its label as

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

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

Then, use $\hat{x}$ and $\hat{y}$ as a pair of training data.



That simple?

Yes, it is that simple.

Yet, it is different from mixing images.

Mixup increases both the images and their labels while mixed images only increases the images but replicates the labels.

Mixup interpolates both images and labels.

Why?

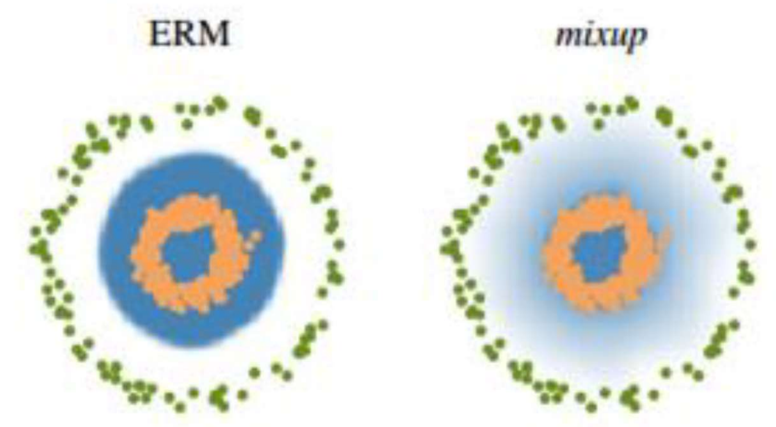
What Mixup does?

It tells us that some samples are less confident than others.

For example, $\hat{x} = 0.2x_1 + 1 = 0.8x_2$.
It indicates that $\hat{x}$ is much less confident than x_1 to be predicted to y_1.
Likewise, $\hat{x}$ is less confident than x_2 to be predicted to y_2.

Also, the generated data often bridge two different classes.

It is useful especially to soften the decision boundary.



(b) Effect of *mixup* ($\alpha = 1$) on a toy problem. Green: Class 0. Orange: Class 1. Blue shading indicates $p(y = 1|x)$.

How good it is?

It is effective to reduce the classification errors.

Model	Method	Epochs	Top-1 Error	Top-5 Error
ResNet-50	ERM (Goyal et al., 2017)	90	23.5	-
	mixup $\alpha = 0.2$	90	**23.3**	**6.6**
ResNet-101	ERM (Goyal et al., 2017)	90	22.1	-
	mixup $\alpha = 0.2$	90	**21.5**	**5.6**
ResNeXt-101 32*4d	ERM (Xie et al., 2016)	100	21.2	-
	ERM	90	21.2	5.6
	mixup $\alpha = 0.4$	90	**20.7**	**5.3**
ResNeXt-101 64*4d	ERM (Xie et al., 2016)	100	20.4	5.3
	mixup $\alpha = 0.4$	90	**19.8**	**4.9**
ResNet-50	ERM	200	23.6	7.0
	mixup $\alpha = 0.2$	200	**22.1**	**6.1**
ResNet-101	ERM	200	22.0	6.1
	mixup $\alpha = 0.2$	200	**20.8**	**5.4**
ResNeXt-101 32*4d	ERM	200	21.3	5.9
	mixup $\alpha = 0.4$	200	**20.1**	**5.0**

Table 1: Validation errors for ERM and *mixup* on the development set of ImageNet-2012.

How good it is?

It is effective to reduce the classification errors.

Dataset	Model	ERM	*mixup*
CIFAR-10	PreAct ResNet-18	5.6	4.2
	WideResNet-28-10	3.8	**2.7**
	DenseNet-BC-190	3.7	**2.7**
CIFAR-100	PreAct ResNet-18	25.6	**21.1**
	WideResNet-28-10	19.4	**17.5**
	DenseNet-BC-190	19.0	**16.8**

(a) Test errors for the CIFAR experiments.

How good it is?

It is also robust against the label noise.

Label corruption	Method	Test error		Training error	
		Best	Last	Real	Corrupted
20%	ERM	12.7	16.6	0.05	0.28
	ERM + dropout ($p = 0.7$)	8.8	10.4	5.26	83.55
	mixup ($\alpha = 8$)	**5.9**	6.4	2.27	86.32
	mixup + dropout ($\alpha = 4, p = 0.1$)	6.2	**6.2**	1.92	85.02
50%	ERM	18.8	44.6	0.26	0.64
	ERM + dropout ($p = 0.8$)	14.1	15.5	12.71	86.98
	mixup ($\alpha = 32$)	11.3	12.7	5.84	85.71
	mixup + dropout ($\alpha = 8, p = 0.3$)	**10.9**	**10.9**	7.56	87.90
80%	ERM	36.5	73.9	0.62	0.83
	ERM + dropout ($p = 0.8$)	30.9	35.1	29.84	86.37
	mixup ($\alpha = 32$)	25.3	30.9	18.92	85.44
	mixup + dropout ($\alpha = 8, p = 0.3$)	**24.0**	**24.8**	19.70	87.67

Table 2: Results on the corrupted label experiments for the best models.

How good it is?

It is robust against adversarial attack.

Metric	Method	FGSM	I-FGSM
Top-1	ERM	90.7	99.9
	mixup	**75.2**	99.6
Top-5	ERM	63.1	93.4
	mixup	**49.1**	95.8

(a) White box attacks.

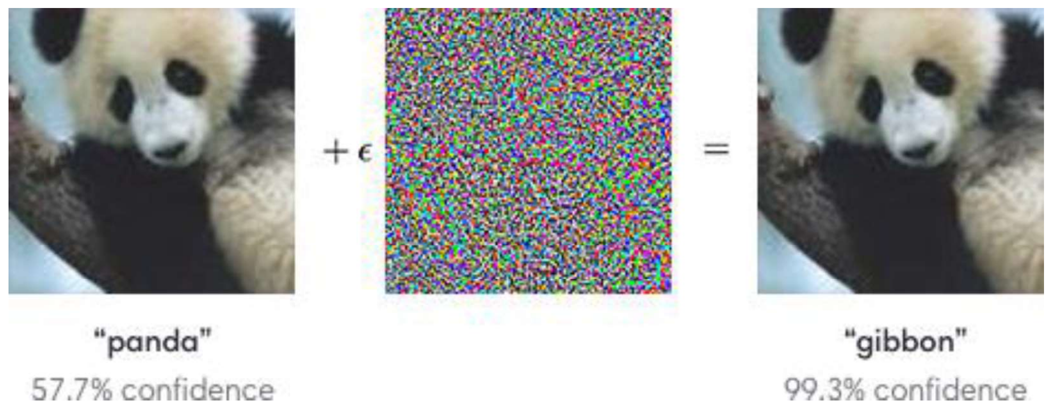
Metric	Method	FGSM	I-FGSM
Top-1	ERM	57.0	57.3
	mixup	**46.0**	**40.9**
Top-5	ERM	24.8	18.1
	mixup	**17.4**	**11.8**

(b) Black box attacks.

Table 3: Classification errors of ERM and *mixup* models when tested on adversarial examples.

Adversarial Attack

DNN has a strange behavior..



What is the problem?

It makes a stupid guess.

We thought DNN is super smart. But, they make a mistake that no human can possibly go wrong.

Such a mistake often comes from the advanced DNN model.

Adversarial Attack

This is an important issue because this strange behavior makes the network model vulnerable to attacks (small perturbation can make the model completely foolish).

Also, understanding this behavior is the good starting point of understanding how neural network works.