Manifold Mixup: Better Representations by Interpolating Hidden States
(2018)

Vikas Verma et al.

Resume

May 14, 2019

1 Introduction

One of the main problems is deep learning is the incorrect predictions, when we evaluate our model on slightly different test data, be it the distributionnal shit, the outliers of the adversarial examples, this is due to a sharp decision boundaries which are close to the data, and given that the mojority of the hidden representations correspond to confident predictions both on and off the data manifold, as a solution to this generalization problem, the authors of the paper propose manifold mixup, a regulizer that encourages neural network to predict less confidently on interpolations of hidden representations; this is based on the intiution that high level representations are often low dimensionnal and useful to linear classifiers, so linear interpolations of hidden representations should explore meaningful regions of the feature space effectively.

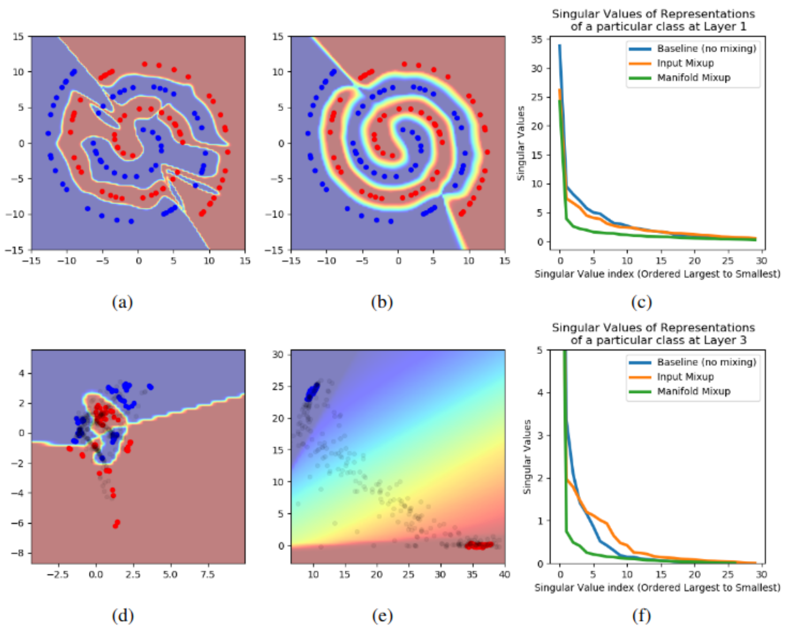

Figure 1: An experiment on a network trained on the 2D spiral dataset with a 2D bottleneck hidden representation in the middle of the network. Manifold mixup has three effects on learning when compared to vanilla training. First, it smoothens decision boundaries (from a. to b.). Second, it improves the arrangement of hidden representations and encourages broader regions of low-confidence predictions (from d. to e.). Black dots are the hidden representation of the inputs sampled uniformly from the range of the input space. Third, it flattens the representations (c. at layer 1, f. at layer 3). Figure 2 shows that these effects are not accomplished by other well-studied regularizers (input mixup, weight decay, dropout, batch normalization, and adding noise to the hidden representations).

Manifold mixup improves generalization because:

- Leads to smoother decision bounderies that are away from the data,
- Provides additionnal learning signal by leveraging the hidden representations using interpolation,
- Compresses the per class representations, each data point is closer to other data points belonging to the same classe

2 Manifold Mixup

Given a deep neural net, we are going to choose a layer k randomly (including the input layer), we then will pass two mini baches (x, y) and (x', y'), until the given selected layers, ending up with two representations $g_k(x)$ and $g_k(x')$, we then apply regular mixup between these intermediate represenations and the labels, first we sample a mixing coefficient $\lambda \sim \text{Beta}(\alpha, \alpha)$ ($\alpha = 1$ is the same as $\lambda \sim U(0, 1)$) and then apply the mixup:

$$(\tilde{g}_k, \tilde{y}) := \left(\text{Mix}_\lambda\left(g_k(x), g_k\left(x'\right)\right), \text{Mix}_\lambda\left(y, y'\right)\right)$$
$$\text{Where: } \text{Mix}_\lambda(a, b) = \lambda \cdot a + (1 - \lambda) \cdot b$$

And then we continue the forward pass using the mixed representation from the layer k, and then we calculate the loss between the prediction and the mixel label, and backpropagating the loss and updating all the parameters of the net.

Mathematically,Manifold Mixupminimizes:

$$L(f) = \underset{(x,y)\sim P}{\mathbb{E}} \; \underset{(x',y')\sim P}{\mathbb{E}} \; \underset{\boldsymbol{k}\sim\text{Beta}(\boldsymbol{\alpha},\boldsymbol{\alpha})}{\mathbb{E}} \ell\left(f_k\left(\text{Mix}_\lambda\left(g_k(x), g_k\left(x'\right)\right)\right), \text{Mix}_\lambda\left(y, y'\right)\right)$$

The sampling of the random layer k and mixing coefficient is done per batch.

3 Experiments

First the authors investigate the effect if the manifold mixup in a supervised learning setting with an image classification task:

Table 1: Classification errors on (a) CIFAR-10 and (b) CIFAR-100. We include results from (Zhang et al., 2018)† and (Guo et al., 2016)‡. Standard deviations over five repetitions.

PreActResNet18	Test Error (%)	Test NLL
No Mixup	4.83 ± 0.066	0.190 ± 0.003
AdaMix‡	3.52	NA
Input Mixup†	4.20	NA
Input Mixup ($\alpha = 1$)	3.82 ± 0.048	0.186 ± 0.004
Manifold Mixup ($\alpha = 2$)	$\underline{2.95 \pm 0.046}$	$\underline{0.137 \pm 0.003}$
PreActResNet34		
No Mixup	4.64 ± 0.072	0.200 ± 0.002
Input Mixup ($\alpha = 1$)	2.88 ± 0.043	0.176 ± 0.002
Manifold Mixup ($\alpha = 2$)	$\underline{2.54 \pm 0.047}$	$\underline{0.118 \pm 0.002}$
Wide-Resnet-28-10		
No Mixup	3.99 ± 0.118	0.162 ± 0.004
Input Mixup ($\alpha = 1$)	2.92 ± 0.088	0.173 ± 0.001
Manifold Mixup ($\alpha = 2$)	$\underline{2.55 \pm 0.024}$	$\underline{0.111 \pm 0.001}$

(a) CIFAR-10

PreActResNet18	Test Error (%)	Test NLL
No Mixup	24.01 ± 0.376	1.189 ± 0.002
AdaMix‡	20.97	n/a
Input Mixup†	21.10	n/a
Input Mixup ($\alpha = 1$)	22.11 ± 0.424	1.055 ± 0.006
Manifold Mixup ($\alpha = 2$)	$\underline{20.34 \pm 0.525}$	$\underline{0.912 \pm 0.002}$
PreActResNet34		
No Mixup	23.55 ± 0.399	1.189 ± 0.002
Input Mixup ($\alpha = 1$)	20.53 ± 0.330	1.039 ± 0.045
Manifold Mixup ($\alpha = 2$)	$\underline{18.35 \pm 0.360}$	$\underline{0.877 \pm 0.053}$
Wide-Resnet-28-10		
No Mixup	21.72 ± 0.117	1.023 ± 0.004
Input Mixup ($\alpha = 1$)	18.89 ± 0.111	0.927 ± 0.031
Manifold Mixup ($\alpha = 2$)	$\underline{18.04 \pm 0.171}$	$\underline{0.809 \pm 0.005}$

(b) CIFAR-100

Table 2: Classification errors and neg-log-likelihoods on SVHN. We run each experiment five times.

PreActResNet18	Test Error (%)	Test NLL
No Mixup	2.89 ± 0.224	0.136 ± 0.001
Input Mixup ($\alpha = 1$)	2.76 ± 0.014	0.212 ± 0.011
Manifold Mixup ($\alpha = 2$)	2.27 ± 0.011	0.122 ± 0.006
PreActResNet34		
No Mixup	2.97 ± 0.004	0.165 ± 0.003
Input Mixup ($\alpha = 1$)	2.67 ± 0.020	0.199 ± 0.009
Manifold Mixup ($\alpha = 2$)	2.18 ± 0.004	0.137 ± 0.008
Wide-Resnet-28-10		
No Mixup	2.80 ± 0.044	0.143 ± 0.002
Input Mixup ($\alpha = 1$)	2.68 ± 0.103	0.184 ± 0.022
Manifold Mixup ($\alpha = 2$)	2.06 ± 0.068	0.126 ± 0.008

Table 3: Accuracy on TinyImagenet.

PreActResNet18	top-1	top-5
No Mixup	55.52	71.04
Input Mixup ($\alpha = 0.2$)	56.47	71.74
Input Mixup ($\alpha = 0.5$)	55.49	71.62
Input Mixup ($\alpha = 1.0$)	52.65	70.70
Input Mixup ($\alpha = 2.0$)	44.18	68.26
Manifold Mixup ($\alpha = 0.2$)	58.70	73.59
Manifold Mixup ($\alpha = 0.5$)	57.24	73.48
Manifold Mixup ($\alpha = 1.0$)	56.83	73.75
Manifold Mixup ($\alpha = 2.0$)	48.14	71.69

To test the effect of manifold mixup when the test set is different than the training set unsing a number of deformations on the test split:

Table 4: Test accuracy on novel deformations. All models trained on normal CIFAR-100.

Deformation	No Mixup	Input Mixup ($\alpha = 1$)	Input Mixup ($\alpha = 2$)	*Manifold Mixup* ($\alpha = 2$)
Rotation U($-20°$,$20°$)	52.96	55.55	56.48	60.08
Rotation U($-40°$,$40°$)	33.82	37.73	36.78	42.13
Shearing U($-28.6°$, $28.6°$)	55.92	58.16	60.01	62.85
Shearing U($-57.3°$, $57.3°$)	35.66	39.34	39.7	44.27
Zoom In (60% rescale)	12.68	13.75	13.12	11.49
Zoom In (80% rescale)	47.95	52.18	50.47	52.70
Zoom Out (120% rescale)	43.18	60.02	61.62	63.59
Zoom Out (140% rescale)	19.34	41.81	42.02	45.29

Table 5: Test accuracy *Manifold Mixup* for different sets of eligible layers $\mathcal{S}$ on CIFAR.

$\mathcal{S}$	CIFAR-10	CIFAR-100
$\{0, 1, 2\}$	97.23	79.60
$\{0, 1\}$	96.94	78.93
$\{0, 1, 2, 3\}$	96.92	80.18
$\{1, 2\}$	96.35	78.69
$\{0\}$	96.73	78.15
$\{1, 2, 3\}$	96.51	79.31
$\{1\}$	96.10	78.72
$\{2, 3\}$	95.32	76.46
$\{2\}$	95.19	76.50
$\{\}$	95.27	76.40

Table 6: Test accuracy (%) of Input Mixup and *Manifold Mixup* for different α on CIFAR-10.

α	Input Mixup	*Manifold Mixup*
0.5	96.68	96.76
1.0	96.75	97.00
1.2	96.72	97.03
1.5	96.84	97.10
1.8	96.80	97.15
2.0	96.73	97.23