Mixup: Beyond Empirical Risk Minimization

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, David Lopez-Paz

Доклад подготовили: Петрович Сергей, Булатова Екатерина, Дроздова Анастасия, Сибагатова Софья

HSE University, 2021

Введение

Published as a conference paper at ICLR 2018

mixup: BEYOND EMPIRICAL RISK MINIMIZATION

Hongyi Zhang
MIT

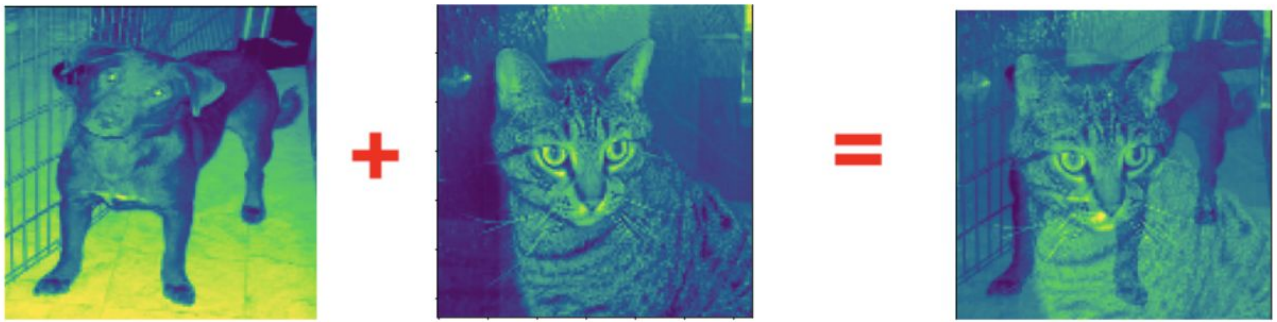
Moustapha Cisse, Yann N. Dauphin, David Lopez-Paz*
FAIR

ABSTRACT

Large deep neural networks are powerful, but exhibit undesirable behaviors such as memorization and sensitivity to adversarial examples. In this work, we propose *mixup*, a simple learning principle to alleviate these issues. In essence, *mixup* trains a neural network on convex combinations of pairs of examples and their labels. By doing so, *mixup* regularizes the neural network to favor simple linear behavior in-between training examples. Our experiments on the ImageNet-2012, CIFAR-10, CIFAR-100, Google commands and UCI datasets show that *mixup* improves the generalization of state-of-the-art neural network architectures. We also find that *mixup* reduces the memorization of corrupt labels, increases the robustness to adversarial examples, and stabilizes the training of generative adversarial networks.

Идея

$$\tilde{x} = \lambda x_i + (1 - \lambda) x_j, \qquad \text{where } x_i, x_j \text{ are raw input vectors}$$
$$\tilde{y} = \lambda y_i + (1 - \lambda) y_j, \qquad \text{where } y_i, y_j \text{ are one-hot label encodings}$$



Теоретическое обоснование. Empirical risk

$$R(f) = \int \ell(f(x), y) \mathrm{d}P(x, y).$$ *Expected risk*

$$P_\delta(x, y) = \frac{1}{n} \sum_{i=1}^{n} \delta(x = x_i, y = y_i),$$ *Empirical distribution*

$$R_\delta(f) = \int \ell(f(x), y) \mathrm{d}P_\delta(x, y) = \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i).$$ *Empirical risk*

Теоретическое обоснование. Vicinal risk

$$P_\nu(\tilde{x}, \tilde{y}) = \frac{1}{n} \sum_{i=1}^{n} \nu(\tilde{x}, \tilde{y} | x_i, y_i),$$

Vicinal distribution

$$R_\nu(f) = \frac{1}{m} \sum_{i=1}^{m} \ell(f(\tilde{x}_i), \tilde{y}_i).$$

Empirical vicinal risk

mixup *задает распределение:*

$$\mu(\tilde{x}, \tilde{y} | x_i, y_i) = \frac{1}{n} \sum_{j}^{n} \mathop{\mathbb{E}}_{\lambda} \left[\delta(\tilde{x} = \lambda \cdot x_i + (1 - \lambda) \cdot x_j, \tilde{y} = \lambda \cdot y_i + (1 - \lambda) \cdot y_j) \right],$$

where $\lambda \sim \text{Beta}(\alpha, \alpha)$, for $\alpha \in (0, \infty)$.

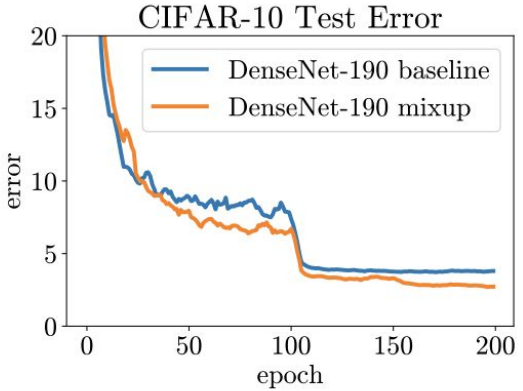
Эксперимент. ImageNet Classification

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**

Эксперимент. CIFAR-10 and CIFAR-100

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.



(b) Test error evolution for the best ERM and *mixup* models.

Эксперимент. Speech data

Model	Method	Validation set	Test set
LeNet	ERM	**9.8**	**10.3**
	mixup ($\alpha = 0.1$)	10.1	10.8
	mixup ($\alpha = 0.2$)	10.2	11.3
VGG-11	ERM	5.0	4.6
	mixup ($\alpha = 0.1$)	4.0	3.8
	mixup ($\alpha = 0.2$)	**3.9**	**3.4**

Эксперимент. Memorization of corrupted labels

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

Эксперимент. Robustness to adversarial examples

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.

Эксперимент. Stabilization of GANs

$$\max_g \min_d \mathop{\mathbb{E}}_{x,z} \ell(d(x), 1) + \ell(d(g(z)), 0),$$

$$\max_g \min_d \mathop{\mathbb{E}}_{x,z,\lambda} \ell(d(\lambda x + (1-\lambda) g(z)), \lambda).$$

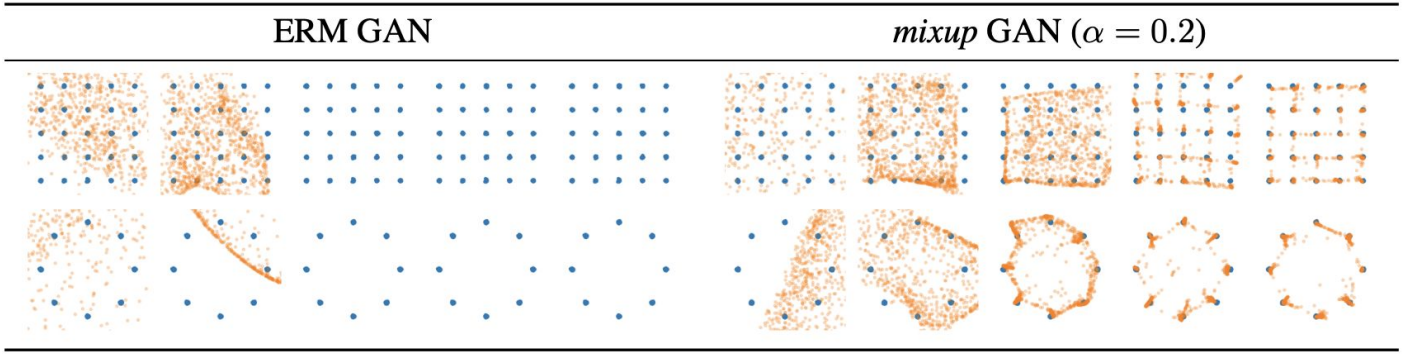


Figure 5: Effect of *mixup* on stabilizing GAN training at iterations 10, 100, 1000, 10000, and 20000.

Рецензия

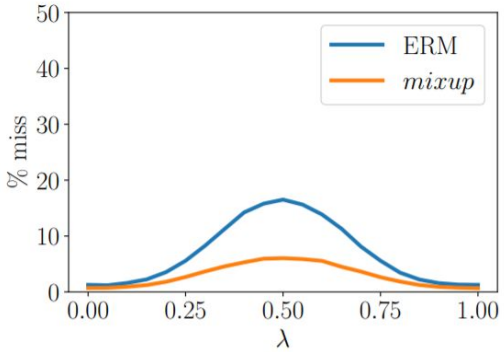
Булатова Катя

Сильные стороны

- Предложена альтернатива аугментациям, не зависящая от набора данных
- Сам метод очень прост в имплементации
- Все гиперпараметры для воспроизведения результатов указаны в статье
- Очень много экспериментов как на разных датасетах (включая даже разные типы данных, например изображения и звуки), так и для разных параметров, причем для разбора разнообразных предположений (запоминание классов, устойчивость)
- SOTA

Слабые стороны

- Не совсем понятно, насколько актуальна задача предсказания элементов in-between data
- Хотя в финальной версии статьи есть теоретическое обоснование того, как именно можно интерпретировать переход к миксапу (vicinal distribution), не очевидно, почему от этого должно стать лучше



(a) Prediction errors in-between training data. Evaluated at $x = \lambda x_i + (1-\lambda)x_j$, a prediction is counted as a "miss" if it does not belong to $\{y_i, y_j\}$. The model trained with *mixup* has fewer misses.

Насколько хорошо написана статья & Воспроизводимость

- Пара орфографических ошибок вроде “Figure 2 illustrate” & “the input that weights more”

- Идея выражается несколькими строчками кода
- Все гиперпараметры указаны

```
# y1, y2 should be one-hot vectors
for (x1, y1), (x2, y2) in zip(loader1, loader2):
    lam = numpy.random.beta(alpha, alpha)
    x = Variable(lam * x1 + (1. - lam) * x2)
    y = Variable(lam * y1 + (1. - lam) * y2)
    optimizer.zero_grad()
    loss(net(x), y).backward()
    optimizer.step()
```

(a) One epoch of *mixup* training in PyTorch.

- Оценки с openreview: 6-7-6

Контекст

Дроздова Настя

Публикация

- ICLR 2018 (постер) (16 Feb 2018 (modified: 24 Feb 2018) - openreview).
- Первая версия на архиве 25 Oct 2017.

Авторы

- **Hongyi Zhang** - MIT, Graduate Research Assistant на момент публикации. В 2018 PhD в MIT.
- **Moustapha Cisse** - Facebook Artificial Intelligence Research(FAIR). PhD in Machine Learning from Pierre et Marie Curie University, France.
- **Yann N. Dauphin** - FAIR. В Facebook занимался машинным переводом. В 2015 PhD в университете Монреаля.
- **David Lopez-Paz** - FAIR.

Предыдущие работы авторов в целом не связаны с mixup.

Parseval networks: Improving robustness to adversarial examples[1] (2017) (Moustapha Cisse, Piotr Bojanowski, Edouard Grave, Yann Dauphin, Nicolas Usunier)

1. https://arxiv.org/abs/1704.08847

На что опирается работа

Теоретическое обоснование

- Empirical Risk Minimization (ERM) principle (Vapnik, 1998).
- Vicinal Risk Minimization (VRM) principle (Chapelle et al., 2000)

Цитирования

- 2704 цитирований.
- AugMix: A Simple Data Processing Method to Improve Robustness and Uncertainty.[1] (ICLR 2020)
- CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features.[2] (ICCV, 2019)
- FMix: Enhancing Mixed Sample Data Augmentation. [3] (2020, ICLR 2021 rejected)

1. https://arxiv.org/abs/1912.02781
2. https://arxiv.org/abs/1905.04899
3. https://arxiv.org/abs/2002.12047

Конкуренты

- Between-class learning for image classification. [1] (CVPR, 2018.)
  - Предложена аналогичная идея(BC), а также улучшенная версия (BC+), учитывающая нормализацию изображений.

1. https://arxiv.org/abs/1711.10284