mixup: Beyond Empirical Risk Minimization

Гришанин Виктор
Михненко Наталья
Колесников Георгий
Стрельцов Артем
1 декабря 2021

Empirical Risk Minimization (ERM)

Expected risk:

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

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

Empirical risk:

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

Vicinity Risk Minimization (VRM)

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

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

mixup

распределение:

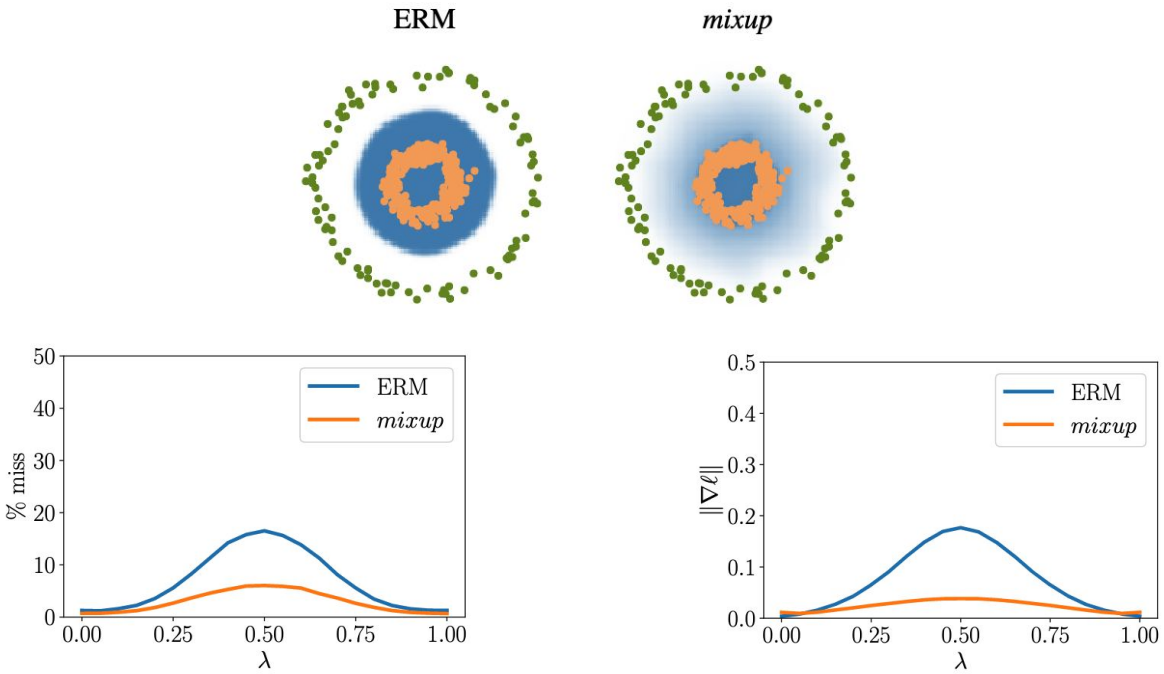
$$\mu(\tilde{x}, \tilde{y}|x_i, y_i) = \frac{1}{n}\sum_{j}^{n} \mathop{\mathbb{E}}_{\lambda} [\delta(\tilde{x} = \lambda \cdot x_i + (1-\lambda) \cdot x_j, \tilde{y} = \lambda \cdot y_i + (1-\lambda) \cdot y_j)]$$

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

семплирование из распределения:

$$\tilde{x} = \lambda x_i + (1-\lambda)x_j$$
$$\tilde{y} = \lambda y_i + (1-\lambda)y_j$$

mixup


ERM
mixup
% miss
50
40
30
20
10
0
0.00
0.25
0.50
0.75
1.00
λ
ERM
mixup
$\|\nabla \ell\|$
0.5
0.4
0.3
0.2
0.1
0.0
0.00
0.25
0.50
0.75
1.00
λ
ERM
mixup

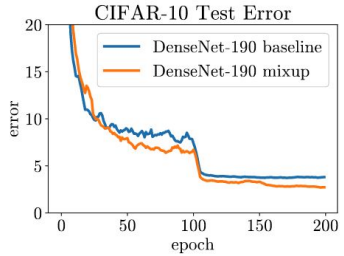
Эксперименты

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.

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.

Figure 3: Test errors for ERM and *mixup* on the CIFAR experiments.

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**

Figure 4: Classification errors of ERM and *mixup* on the Google commands dataset.

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.

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.

Dataset	ERM	*mixup*
Abalone	74.0	73.6
Arcene	57.6	**48.0**
Arrhythmia	56.6	**46.3**

Dataset	ERM	*mixup*
Htru2	2.0	2.0
Iris	21.3	**17.3**
Phishing	16.3	15.2

Table 4: ERM and *mixup* classification errors on the UCI datasets.

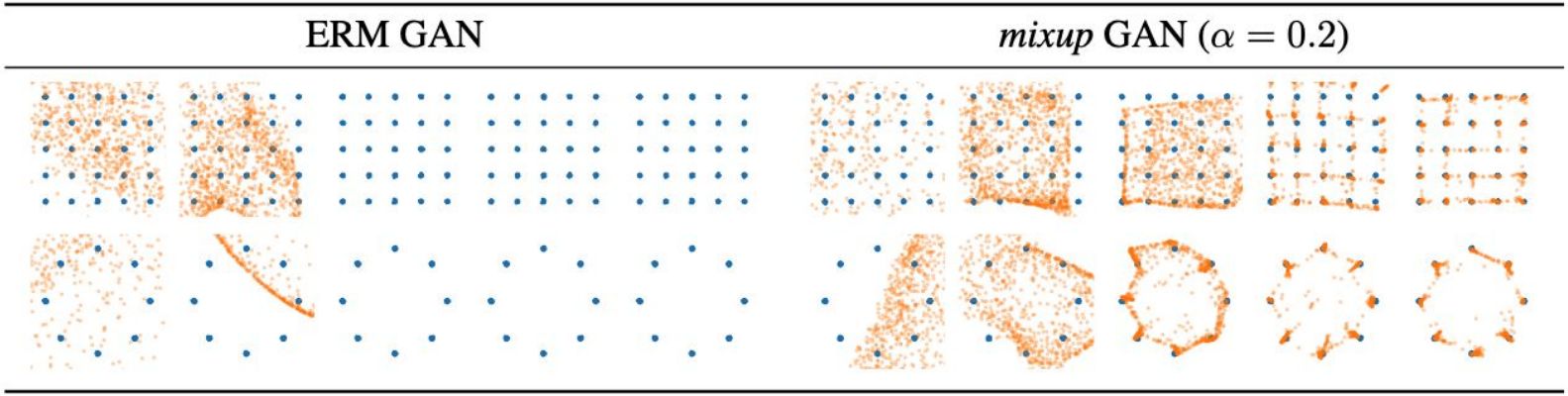


Figure 5: Effect of *mixup* on stabilizing GAN training at iterations 10, 100, 1000, 10000, and 20000.

Рецензия

Краткое содержание: в статье предлагается метод генерации дополнительных данных для обучения с помощью выпуклой комбинации уже существующих объектов

Оценка: 7 A good submission; an accept. I vote for accepting this submission, although I would not be upset if it were rejected.

Уверенность: 4 You are confident in your assessment, but not absolutely certain.

Сильные стороны

- Актуальность
- Подходит под многие задачи машинного обучения (текст/звук/картинки)
- Прост в реализации и отлично расписан в статье.
- Не требует больших вычислительных ресурсов
- Судя по результатам экспериментов метод действительно улучшает качество моделей

Слабые стороны

- Основной минус – отсутствие теоретического обоснования метода.
- Непонятно, что такое выпуклая комбинация картинок, почему это нечто осмысленное

Вопросы:

- Как выбирать пары?
- Что будет если таргет не менять?

Обзор статьи “BEYOND EMPIRICAL RISK MINIMIZATION”

Колесников Георгий 182

Как издана?

- Написана в октябре 2017 года
- Издана весной 2018 на постере International Conference on Learning Representations

6th International Conference on Learning Representations

Кто автор статьи?

- *Hongyi Zhang* - Research Scientist @ ByteDance. получил PhD в MIT в 2018. В основном работы по римановым. https://www.linkedin.com/in/hongyizhang
- *Moustapha Cisse* - Research Scientist (FAIR) and Head of Google AI Center, Accra (Гана). До этого написал несколько работ по смежным темам adversarial ml. PhD at University Pierre and Marie Curie in France. https://www.crunchbase.com/person/moustapha-cisse
- *Yann N. Dauphin* - research @ FAIR, PhD @ University of Montreal (2015). Помимо этого есть работы по NLP и методам оптимизации. Входит в Theano Development Team.
- *David Lopez-Paz* - researcher @ FAIR, PhD @ University of Cambridge. достаточно много работ по GAN-нам и semi-supervised https://ai.facebook.com/people/david-lopez-paz/

Влияние на работу

- Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang. Random erasing data augmentation. - статья, в которой показано, что сеть легко может запомнить все лейблы; обучение на случайные лейблы. одна из основных проверок на устойчивость, что новому подходу сложнее их всех запомнить.
- C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus. Intriguing properties of neural networks - тоже исследую сети. показывают неустойчивость к простым adversarial примерам и что модель скорее запоминает примеры, а не выучивает полезные признаки (что отдельный нейрон не выучивает полезные признаки, а только ансамбль)

Цитирование статьи

- Manifold mixup http://proceedings.mlr.press/v97/verma19a/verma19a.pdf - делают mixup, но внутри слоев -> более устойчивая модель, лучшая обощающая способность
- MixMatch: A Holistic Approach to Semi-Supervised Learning https://arxiv.org/pdf/1905.02249.pdf - применяется похожая идея, но с использованием неразмеченных примеров

Аналоги

- Data Augmentation by Pairing Samples for Images Classification. Hiroshi Inoue. 2018. (https://arxiv.org/abs/1801.02929) - тестируют похожую идею. берут и усредняют рандомные пары и предсказывают лейбл первой. Улучшение получено на всех датасетах, самые значительные на ILSVRC 2012 с GoogLeNet и уменьшением ошибки с 33.5% до 29.0%, для CIFAR-10 dataset с 8.22% до 6.93%.
- Dataset augmentation in feature space. T. DeVries and G. W. Taylor. 2017. (https://arxiv.org/abs/1702.05538) - такой же mixup, но для ближайших из одного класса. Протестировано в speech, sensor processing, motion capture, and images.

Продолжение исследования

- Попробовать внедрить этот метод в unsupervised/регрессии
- Попробовать экстраполировать примеры и проверить, какие там результаты для примеров вне train-а