Sharpness Aware Minimization

Пятько Даниил, БПМИ 192

Определения, часть 1

$$\mathcal{S} = \cup_{i=1}^{n} \{(\boldsymbol{x_i}, \boldsymbol{y_i})\} \text{ — обучающая выборка}$$

$$L_{\mathcal{S}}(\boldsymbol{w}) = \frac{1}{n} \sum_{i=1}^{n} l(\boldsymbol{w}, \boldsymbol{x_i}, \boldsymbol{y_i}) \text{ — функция потерь на } \mathcal{S}$$

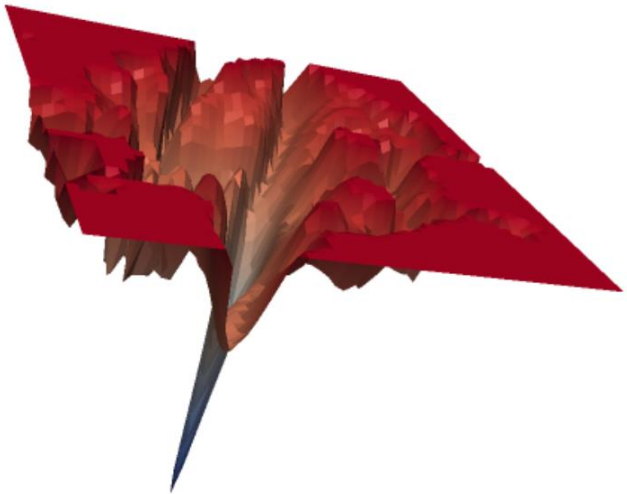
Что делаем обычно?

$$L_S(\boldsymbol{w}) \to \min_{\boldsymbol{w}}$$

Поиск параметров через оптимизацию функции потерь

Функции потерь для сложных моделей:

- Много локальных минимумов
- Разные обобщающие способности в этих минимумах
- Качество модели сильно зависит от оптимизатора



Новый подход

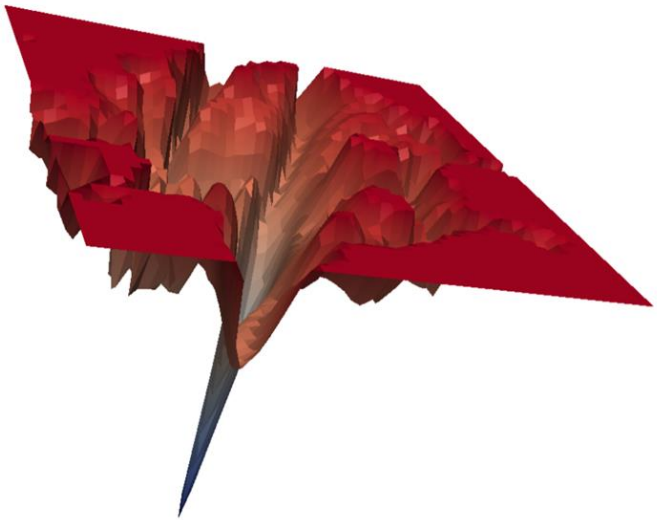
$$L_{\mathcal{S}}^{SAM}(\boldsymbol{w}) = \max_{\|\boldsymbol{\varepsilon}\|_2 \leq \rho} L_{\mathcal{S}}(\boldsymbol{w} + \boldsymbol{\varepsilon})$$

Обучение:

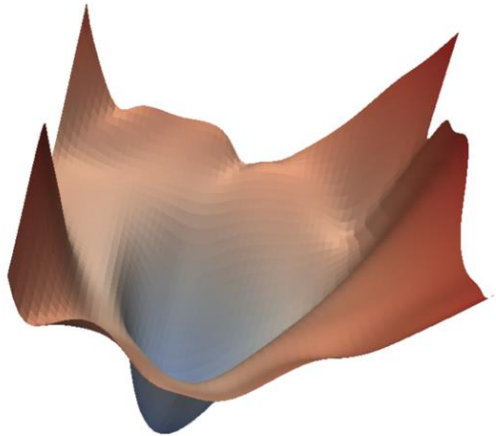
$$L_{\mathcal{S}}^{SAM}(\boldsymbol{w}) \to \min_{\boldsymbol{w}}$$

Что получается?

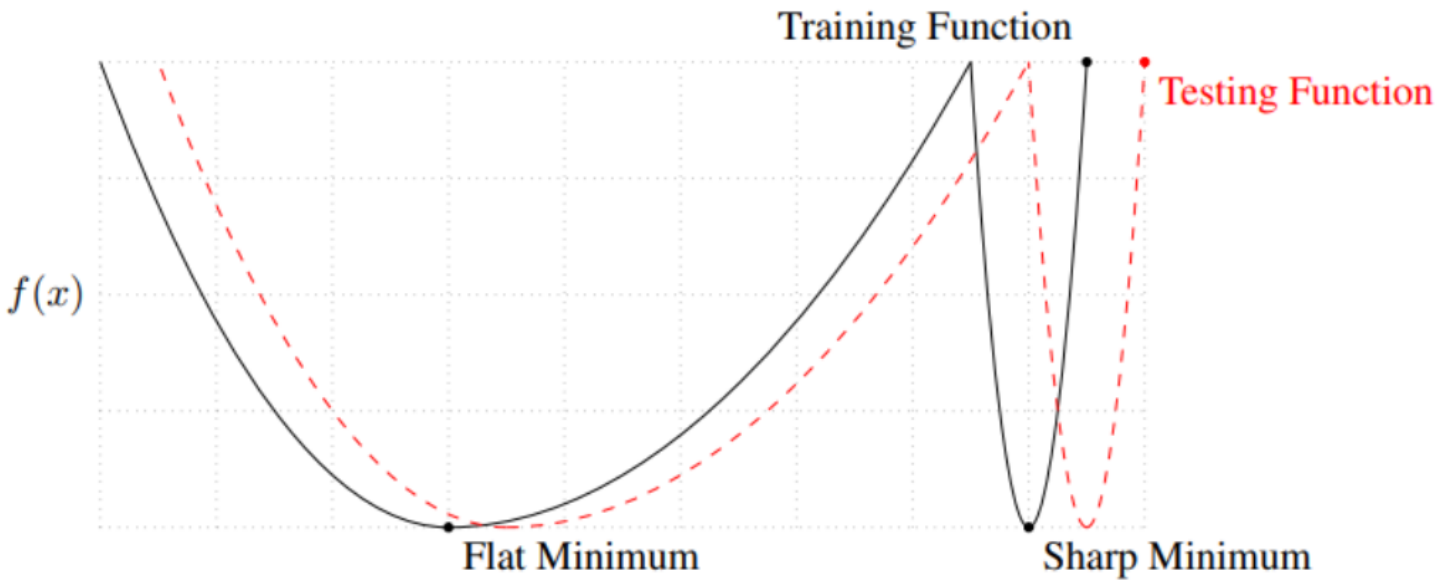
$$L_{\mathcal{S}}(\boldsymbol{w}) \to \min_{\boldsymbol{w}}$$



$$L_{\mathcal{S}}^{SAM}(\boldsymbol{w}) \to \min_{\boldsymbol{w}}$$



Проблема острых минимумов наглядно



Определения, часть 2

$$\mathcal{S} = \cup_{i=1}^{n} \{(\boldsymbol{x_i}, \boldsymbol{y_i})\} \text{ — берется из распределения } \mathscr{D}$$

$$L_{\mathscr{D}}(\boldsymbol{w}) = \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathscr{D}}[l(\boldsymbol{w}, \boldsymbol{x}, \boldsymbol{y})]$$

Теоретическое обоснование

Теорема. Для любого $\rho > 0$ с большой вероятностью для множества $\mathcal{S}$ выполнено

$$L_{\mathscr{D}}(\boldsymbol{w}) \leq \max_{\|\varepsilon\|_2 \leq \rho} L_{\mathcal{S}}(\boldsymbol{w} + \boldsymbol{\varepsilon}) + h\left(\frac{\|w\|_2^2}{\rho^2}\right)$$

h — возрастающая функция

Функция потерь с регуляризацией

$$L_{\mathscr{D}}(\boldsymbol{w}) \leq \max_{\|\varepsilon\|_2 \leq \rho} L_{\mathcal{S}}(\boldsymbol{w}+\boldsymbol{\varepsilon}) + h\left(\frac{\|w\|_2^2}{\rho^2}\right)$$

$$L_{\mathscr{D}}(\boldsymbol{w}) \leq L_{\mathcal{S}}^{SAM}(\boldsymbol{w}) + h\left(\frac{\|w\|_2^2}{\rho^2}\right)$$

Обучение:

$$L_{\mathcal{S}}^{SAM}(\boldsymbol{w}) + \lambda \|w\|_2^2 \to \min_{\boldsymbol{w}}$$

Осталось найти $\nabla_{\boldsymbol{w}} L_{\mathcal{S}}^{SAM}(\boldsymbol{w})$

Формула шага, часть 1

Находим ε с максимальным $L_{\mathcal{S}}(\boldsymbol{w}+\boldsymbol{\varepsilon})$:

$$\boldsymbol{\varepsilon}^*(\boldsymbol{w}) = \underset{\|\boldsymbol{\varepsilon}\|_2 \leq \rho}{\operatorname{argmax}} L_{\mathcal{S}}(\boldsymbol{w}+\boldsymbol{\varepsilon}) \approx \underset{\|\boldsymbol{\varepsilon}\|_2 \leq \rho}{\operatorname{argmax}} L_{\mathcal{S}}(\boldsymbol{w}) + \boldsymbol{\varepsilon}^T \nabla_w L_{\mathcal{S}}(\boldsymbol{w}) = \underset{\|\boldsymbol{\varepsilon}\|_2 \leq \rho}{\operatorname{argmax}} \boldsymbol{\varepsilon}^T \nabla_w L_{\mathcal{S}}(\boldsymbol{w})$$

$$\hat{\boldsymbol{\varepsilon}}(\boldsymbol{w}) = \frac{\rho}{\|\nabla_{\boldsymbol{w}} L_{\mathcal{S}}(\boldsymbol{w})\|_2} \nabla_{\boldsymbol{w}} L_{\mathcal{S}}(\boldsymbol{w})$$

Формула шага, часть 2

$$L_{\mathcal{S}}^{SAM}(\boldsymbol{w}) \approx L_{\mathcal{S}}(\boldsymbol{w} + \hat{\boldsymbol{\varepsilon}}(\boldsymbol{w}))$$

$$\nabla_{\boldsymbol{w}} L_{\mathcal{S}}^{SAM}(\boldsymbol{w}) \approx \nabla_{\boldsymbol{w}} L_{\mathcal{S}}(\boldsymbol{w} + \hat{\boldsymbol{\varepsilon}}(\boldsymbol{w})) = \frac{d(\boldsymbol{w} + \hat{\boldsymbol{\varepsilon}}(\boldsymbol{w}))}{d\boldsymbol{w}} \nabla_{\boldsymbol{w}} L_{\mathcal{S}}(\boldsymbol{w})\big|_{\boldsymbol{w}+\hat{\boldsymbol{\varepsilon}}(\boldsymbol{w})}$$

$$= \nabla_{\boldsymbol{w}} L_{\mathcal{S}}(\boldsymbol{w})\big|_{\boldsymbol{w}+\hat{\boldsymbol{\varepsilon}}(\boldsymbol{w})} + \underbrace{\frac{d\hat{\boldsymbol{\varepsilon}}(\boldsymbol{w})}{d\boldsymbol{w}} \nabla_{\boldsymbol{w}} L_{\mathcal{S}}(\boldsymbol{w})\big|_{\boldsymbol{w}+\hat{\boldsymbol{\varepsilon}}(\boldsymbol{w})}}_{\text{добавка второго порядка}}$$

Таким образом,

$$\nabla_{\boldsymbol{w}} L_{\mathcal{S}}^{SAM}(\boldsymbol{w}) \approx \nabla_{\boldsymbol{w}} L_{\mathcal{S}}(\boldsymbol{w})\big|_{\boldsymbol{w}+\hat{\boldsymbol{\varepsilon}}(\boldsymbol{w})}$$

Алгоритм

Input: Training set $\mathcal{S} \triangleq \cup_{i=1}^{n}\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}$, Loss function $l : \mathcal{W} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$, Batch size b, Step size $\eta > 0$, Neighborhood size $\rho > 0$.
Output: Model trained with SAM
Initialize weights $\boldsymbol{w}_0$, $t = 0$;
while *not converged* **do**
  Sample batch $\mathcal{B} = \{(\boldsymbol{x}_1, \boldsymbol{y}_1), ...(\boldsymbol{x}_b, \boldsymbol{y}_b)\}$;
  Compute gradient $\nabla_{\boldsymbol{w}} L_{\mathcal{B}}(\boldsymbol{w})$ of the batch's training loss;
  Compute $\hat{\boldsymbol{\epsilon}}(\boldsymbol{w})$ per equation 2;
  Compute gradient approximation for the SAM objective (equation 3): $\boldsymbol{g} = \nabla_w L_{\mathcal{B}}(\boldsymbol{w})|_{\boldsymbol{w}+\hat{\boldsymbol{\epsilon}}(\boldsymbol{w})}$;
  Update weights: $\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta \boldsymbol{g}$;
  $t = t + 1$;
end
return $\boldsymbol{w}_t$

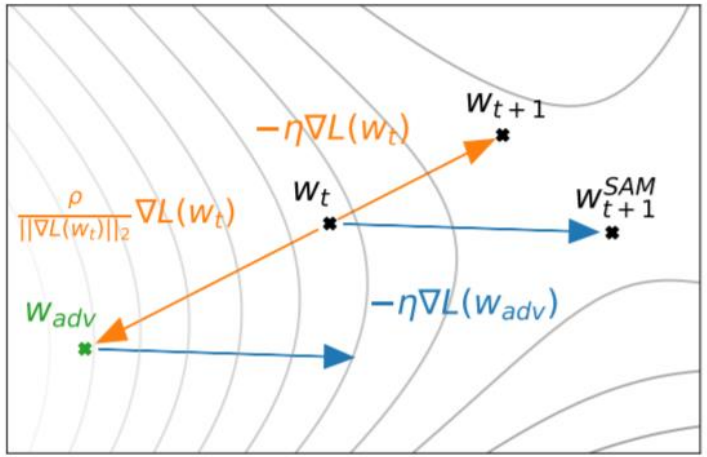


Figure 2: Schematic of the SAM parameter update.

Эксперименты: CIFAR-10, CIFAR-100

		CIFAR-10		CIFAR-100	
Model	Augmentation	SAM	SGD	SAM	SGD
WRN-28-10 (200 epochs)	Basic	**2.7**$_{\pm 0.1}$	3.5$_{\pm 0.1}$	**16.5**$_{\pm 0.2}$	18.8$_{\pm 0.2}$
WRN-28-10 (200 epochs)	Cutout	**2.3**$_{\pm 0.1}$	2.6$_{\pm 0.1}$	**14.9**$_{\pm 0.2}$	16.9$_{\pm 0.1}$
WRN-28-10 (200 epochs)	AA	**2.1**$_{\pm <0.1}$	2.3$_{\pm 0.1}$	**13.6**$_{\pm 0.2}$	15.8$_{\pm 0.2}$
WRN-28-10 (1800 epochs)	Basic	**2.4**$_{\pm 0.1}$	3.5$_{\pm 0.1}$	**16.3**$_{\pm 0.2}$	19.1$_{\pm 0.1}$
WRN-28-10 (1800 epochs)	Cutout	**2.1**$_{\pm 0.1}$	2.7$_{\pm 0.1}$	**14.0**$_{\pm 0.1}$	17.4$_{\pm 0.1}$
WRN-28-10 (1800 epochs)	AA	**1.6**$_{\pm 0.1}$	2.2$_{\pm <0.1}$	**12.8**$_{\pm 0.2}$	16.1$_{\pm 0.2}$
Shake-Shake (26 2x96d)	Basic	**2.3**$_{\pm <0.1}$	2.7$_{\pm 0.1}$	**15.1**$_{\pm 0.1}$	17.0$_{\pm 0.1}$
Shake-Shake (26 2x96d)	Cutout	**2.0**$_{\pm <0.1}$	2.3$_{\pm 0.1}$	**14.2**$_{\pm 0.2}$	15.7$_{\pm 0.2}$
Shake-Shake (26 2x96d)	AA	**1.6**$_{\pm <0.1}$	1.9$_{\pm 0.1}$	**12.8**$_{\pm 0.1}$	14.1$_{\pm 0.2}$
PyramidNet	Basic	**2.7**$_{\pm 0.1}$	4.0$_{\pm 0.1}$	**14.6**$_{\pm 0.4}$	19.7$_{\pm 0.3}$
PyramidNet	Cutout	**1.9**$_{\pm 0.1}$	2.5$_{\pm 0.1}$	**12.6**$_{\pm 0.2}$	16.4$_{\pm 0.1}$
PyramidNet	AA	**1.6**$_{\pm 0.1}$	1.9$_{\pm 0.1}$	**11.6**$_{\pm 0.1}$	14.6$_{\pm 0.1}$
PyramidNet+ShakeDrop	Basic	**2.1**$_{\pm 0.1}$	2.5$_{\pm 0.1}$	**13.3**$_{\pm 0.2}$	14.5$_{\pm 0.1}$
PyramidNet+ShakeDrop	Cutout	**1.6**$_{\pm <0.1}$	1.9$_{\pm 0.1}$	**11.3**$_{\pm 0.1}$	11.8$_{\pm 0.2}$
PyramidNet+ShakeDrop	AA	**1.4**$_{\pm <0.1}$	1.6$_{\pm <0.1}$	**10.3**$_{\pm 0.1}$	10.6$_{\pm 0.1}$

Эксперименты: ResNet на ImageNet

Model	Epoch	SAM		Standard Training (No SAM)	
		Top-1	Top-5	Top-1	Top-5
ResNet-50	100	**22.5**$_{\pm 0.1}$	6.28$_{\pm 0.08}$	22.9$_{\pm 0.1}$	6.62$_{\pm 0.11}$
	200	**21.4**$_{\pm 0.1}$	5.82$_{\pm 0.03}$	22.3$_{\pm 0.1}$	6.37$_{\pm 0.04}$
	400	**20.9**$_{\pm 0.1}$	5.51$_{\pm 0.03}$	22.3$_{\pm 0.1}$	6.40$_{\pm 0.06}$
ResNet-101	100	**20.2**$_{\pm 0.1}$	5.12$_{\pm 0.03}$	21.2$_{\pm 0.1}$	5.66$_{\pm 0.05}$
	200	**19.4**$_{\pm 0.1}$	4.76$_{\pm 0.03}$	20.9$_{\pm 0.1}$	5.66$_{\pm 0.04}$
	400	**19.0**$_{\pm <0.01}$	4.65$_{\pm 0.05}$	22.3$_{\pm 0.1}$	6.41$_{\pm 0.06}$
ResNet-152	100	**19.2**$_{\pm <0.01}$	4.69$_{\pm 0.04}$	20.4$_{\pm <0.0}$	5.39$_{\pm 0.06}$
	200	**18.5**$_{\pm 0.1}$	4.37$_{\pm 0.03}$	20.3$_{\pm 0.2}$	5.39$_{\pm 0.07}$
	400	**18.4**$_{\pm <0.01}$	4.35$_{\pm 0.04}$	20.9$_{\pm <0.0}$	5.84$_{\pm 0.07}$

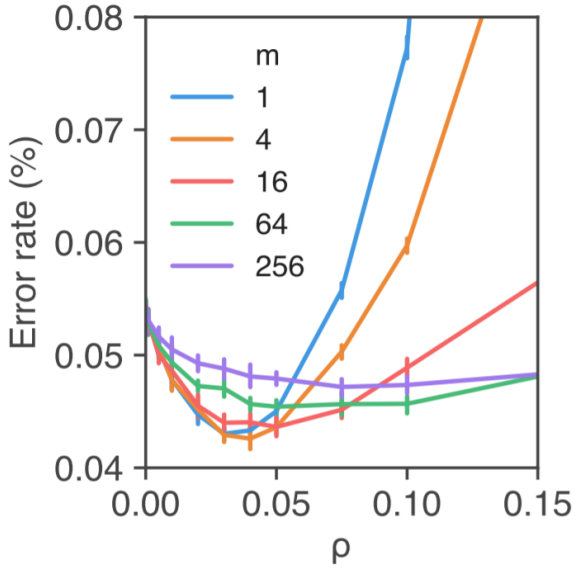
Эксперименты: CIFAR-10 with label noise

- Модель ResNet-32, 200 эпох обучения
- Bootstrap – обучаемся два раза, во второй раз на предсказанных первой моделью метках

Method	Noise rate (%)			
	20	40	60	80
Sanchez et al. (2019)	94.0	92.8	90.3	74.1
Zhang & Sabuncu (2018)	89.7	87.6	82.7	67.9
Lee et al. (2019)	87.1	81.8	75.4	-
Chen et al. (2019)	89.7	-	-	52.3
Huang et al. (2019)	92.6	90.3	43.4	-
MentorNet (2017)	92.0	91.2	74.2	60.0
Mixup (2017)	94.0	91.5	86.8	76.9
MentorMix (2019)	**95.6**	**94.2**	91.3	**81.0**
SGD	84.8	68.8	48.2	26.2
Mixup	93.0	90.0	83.8	70.2
Bootstrap + Mixup	93.3	92.0	87.6	72.0
SAM	95.1	93.4	90.5	77.9
Bootstrap + SAM	95.4	**94.2**	**91.8**	79.9

M-sharpness

- Тренируем батчами размера M, минимизация SAM происходит на данном батче, а не на всей обучающей выборке
- График: CIFAR-10 на маленьком ResNet



Гессиан $L_{\mathcal{S}}(\boldsymbol{w})$

- График: CIFAR-10 на маленьком ResNet