Регуляризация нейросетевого слоя путем построения фрейма в пространстве параметров

Григорьев Алексей Дмитриевич

Московский физико-технический институт
Физтех-школа прикладной математики и информатики
Кафедра интеллектуальных систем

Научный руководитель: к.ф.-м.н. А.Н. Гнеушев

Москва, 2022

Задача

Предложить метод уменьшения избыточности параметров нейронной сети и повышения устойчивости модели.

Проблема

Существующие решения, предполагающие регуляризацию параметров модели, накладывают чрезмерные ограничения на оптимизацию весов нейросети, что негативно влияет на качество модели.

Решение

Рассматривать веса слоя нейросети как систему векторов, проекция входа на которую устойчива и полна. Избыточность данной системы свойственна нейронной сети и позволяет точнее описывать ее веса.

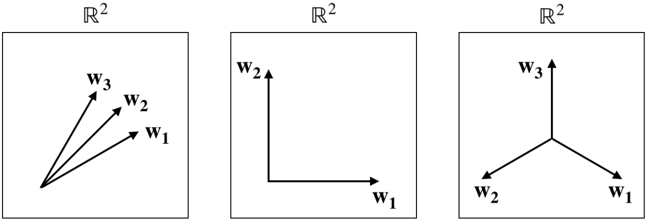
Прунинг параметров

- *P. Molchanov, et al.* Pruning convolutional neural networks for resource efficient transfer learning // ICLR, 2017, P. 1–17.

Повышение разнообразия нейронов

- *W. Lui, et al.* Learning towards minimum hyperspherical energy // NIPS, 2018, P. 6222-6233.
- *N. Bansal, et al.* Can we gain more from orthogonality regularizations in training deep CNNs? // NIPS, 2018, P. 4266–4276.
- *J. Wang, et al.* Orthogonal Convolutional Neural Networks // CVPR, 2020, P. 11505-11515.

- Коррелированные системы весов нейронов неэффективны.
- Ортогональность – чрезмерное требование и ограничение.
- Избыточные полные системы могут быть адекватны.
- Предлагается построение полной системы для разложения входных векторов в избыточном пространстве весов каждого слоя.



коррелированная избыточная система | *ортогональная система* | *адекватная избыточная система*

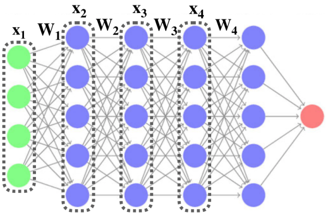
Возможные конфигурации весов нейронов

Семейство моделей

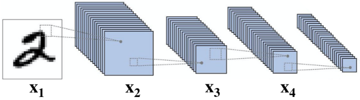
- Ограничим множество рассматриваемых моделей семейством $\varPhi_L$ нейросетей следующего вида.
- $\varphi(\cdot|\mathbf{\Theta}) \in \varPhi_L$ – модель из семейства глубоких нейронных сетей, состоящих из L слоев, каждый из которых представим в виде линейного оператора $\mathcal{F}_i : \mathbb{R}^{n_i} \to \mathbb{R}^{m_i} : m_i \geq n_i$:

$$\mathcal{F}_i(\mathbf{z}) = \mathbf{W}_i\mathbf{z}, \ \forall \mathbf{z} \in \mathbb{R}^{n_i}, \ i = 1, \dots, L,$$

где $\mathbf{W}_i \in \mathbb{R}^{m_i \times n_i}$, $\mathbf{W}_i \subseteq \mathbf{\Theta}$ – матрица линейного оператора, составленная из параметров данного слоя, $n_i, m_i : m_i \geq n_i$ – размерности входа и выхода слоя соответственно.

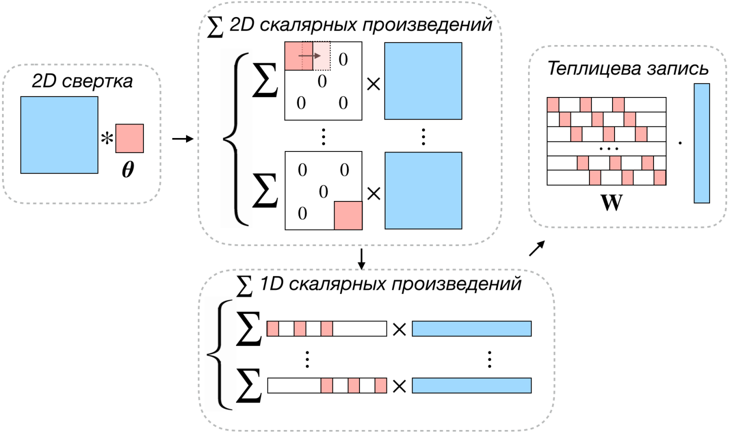


Многослойный перцептрон



Сверточная нейронная сеть

Сверточный слой в виде линейного оператора с матрицей $\mathbf{W}$ задается блочно-теплицевой матрицей из параметров $\boldsymbol{\theta}$ свертки.



Теплицево представление одноканальной свертки

Оптимизация параметров модели

- $\{\mathbf{x}_i, y_i\}_{i=1}^{N}$ — выборка размера N;
- $\varphi(\cdot|\mathbf{\Theta}) \in \mathit{\Phi}_L$ — модель из семейства $\mathit{\Phi}_L$ глубоких нейронных сетей, состоящих из L слоев, представимых в виде линейного оператора;
- минимизация эмпирического риска:

$$\hat{\mathbf{\Theta}} = \arg\min_{\mathbf{\Theta}} \frac{1}{N} \sum_{i=1}^{N} \ell(\varphi(\mathbf{x}_i|\mathbf{\Theta}), y_i) + \gamma \widetilde{R}(\mathbf{\Theta}),$$

где ℓ — функция потерь, релевантная задаче обучения с учителем; $\widetilde{R}(\mathbf{\Theta}) = \sum_{i=1}^{L} R(\mathbf{W}_i)$ — регуляризация , γ — коэф. регуляризации;

Регуляризация параметров

Регуляризация параметров $\mathbf{W}^T = [\mathbf{w}_1 \dots \mathbf{w}_m]$ направлена на минимизацию потерь информации на слое $\mathcal{F}(\mathbf{z}) = \mathbf{W}\mathbf{z}$ путем построения системы весов $\{\mathbf{w}_i\}_{i=1}^{m}$, линейно восстанавливающих вход $\mathbf{z}$ по выходу $\mathcal{F}(\mathbf{z})$:

$$\forall \mathbf{z} \in \mathbb{R}^n \ \exists \tilde{\mathbf{c}} = \tilde{\mathbf{c}}(\mathcal{F}(\mathbf{z}), \mathbf{W}) : \ \mathbf{z} \approx \hat{\mathbf{z}} = \sum_{k=1}^{m} \tilde{c}_k \mathbf{w}_k.$$

Определение *(фрейм)*

$\{\mathbf{w}_k\}_{k=1}^m \subset \mathbb{R}^n$ – *фрейм* в $\mathbb{R}^n$, если $\exists\ A, B:\ 0 < A \leq B < \infty:\ \forall \mathbf{z} \in \mathbb{R}^n$ выполнено *нер-во фрейма*:

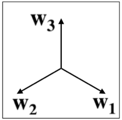
$$A\|\mathbf{z}\|^2 \leq \sum_{i=1}^m |\langle \mathbf{z}, \mathbf{w}_i \rangle|^2 \leq B\|\mathbf{z}\|^2$$

где A, B – *границы фрейма*. Если $A = B$, то фрейм называется *жестким*.

Разложение по дуальной системе

Если $\{\mathbf{w}_k\}_{k=1}^m$ – фрейм в $\mathbb{R}^n$, то разложение по дуальному фрейму $\{\widetilde{\mathbf{w}}_i\}_{i=1}^m$:

$$\mathbf{z} = \sum_{i=1}^m \langle \mathbf{z}, \mathbf{w}_i \rangle \widetilde{\mathbf{w}}_i,\ \ \forall \mathbf{z} \in \mathbb{R}^n.$$



Пример жесткого фрейма с границей $A = \frac{3}{2}$ в $\mathbb{R}^2$

Фреймовая модель нейросетевого слоя

Свойства фрейма $\{\mathbf{w}_k\}_{k=1}^m \subset \mathbb{R}^n$

- Фрейм образует полную систему в $\mathbb{R}^n$. При $m > n$ система избыточна, что характерно для слоя нейросети и позволяет точнее его описывать.
- Если строки $\{\mathbf{w}_k\}_{k=1}^m$ матрицы $\mathbf{W}$ образуют фрейм, то собственные числа $\lambda_1, \ldots, \lambda_n$ матрицы $\mathbf{W}^T\mathbf{W}$ ограничены границами фрейма:

$$A \le \lambda_i \le B, \ \forall i = 1, \ldots, n.$$

- Для переопредленной СЛАУ $\mathcal{F}(\mathbf{z}) = \mathbf{W}\mathbf{z}$ фрейм $\{\mathbf{w}_k\}_{k=1}^m$ дает устойчивое решение задачи восстановления входа: $\mathbf{z} = (\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T\mathcal{F}(\mathbf{z})$. Обусловленность задачи ограничена:

$$\kappa(\mathbf{W}) = \|\mathbf{W}^T\mathbf{W}\| \, \|(\mathbf{W}^T\mathbf{W})^{-1}\| = \frac{|\lambda_{\max}|}{|\lambda_{\min}|} \le \frac{B}{A}.$$

Модель слоя: $\mathbf{W} \in \mathbb{R}^{m\times n}$: $m \ge n$

- Нейросетевого слой задан линейным оператором $\mathcal{F} : \mathbb{R}^n \to \mathbb{R}^m$ с матрицей $\mathbf{W} \in \mathbb{R}^{m\times n}$: $m \ge n$, $\mathcal{F}(\mathbf{z}) = \mathbf{W}\mathbf{z}, \forall \mathbf{z} \in \mathbb{R}^n$; $\mathbf{W}^T = [\mathbf{w}_1 \ldots \mathbf{w}_m]$.
- Если строки $\{\mathbf{w}_k\}_{k=1}^m$ матрицы $\mathbf{W}$ образуют фрейм в $\mathbb{R}^n$, то нейросетевой слой $\mathcal{F}(\mathbf{z}) = \mathbf{W}\mathbf{z}$ обратим.

Построение фрейма

- Неравенство фрейма для строк $\{\mathbf{w}_k\}_{k=1}^m$ матрицы $\mathbf{W}$:

$$A\|\mathbf{z}\|^2 \leq \|\mathbf{W}\mathbf{z}\|^2 \leq B\|\mathbf{z}\|^2,\ \forall \mathbf{z} \in \mathbb{R}^n \iff \begin{cases} (\mathbf{W}^T\mathbf{W} - A\,\mathbb{I}) \succeq 0, \\ (-\mathbf{W}^T\mathbf{W} + B\,\mathbb{I}) \succeq 0. \end{cases}$$

- Матрица $\mathbf{V} \in \mathbb{R}^{m\times m}$ положительно полуопределена, если:
  1. она обладает свойством диагонального преобладания:
  $$|v_{ii}| \geq \sum_{j\neq i} |v_{ij}|\ \forall i = 1, \dots, m,$$
  2. ее диагональные элементы неотрицательны:
  $$v_{ii} \geq 0\ \forall i = 1, \dots, m.$$

- Пусть $\mathbf{V} = \mathbf{W}^T\mathbf{W}$, $M(v) = \min(v, 0)$; введем регуляризатор:

$$R(\mathbf{W}) = \frac{1}{n}\sum_{i=1}^{n} \underbrace{M\Big(v_{ii} - A - \sum_{j=1}^{n}|v_{ij}|\Big)^2}_{\text{штраф } i\text{-ой строки } \left(\mathbf{W}^T\mathbf{W} - A\,\mathbb{I}\right)} + \underbrace{M\Big(-v_{ii} + B - \sum_{j=1}^{n}|v_{ij}|\Big)^2}_{\text{штраф } i\text{-ой строки } \left(-\mathbf{W}^T\mathbf{W} + B\,\mathbb{I}\right)}.$$

Вычислительный эксперимент

Цель

Сравнить предложенный подход к регуляризации параметров модели с существующими решениями в задаче классификации изображений.

Параметры эксперимента

- задача многоклассовой классификации;
- архитектуры модели $\varphi(\cdot|\boldsymbol{\Theta})$ — ResNet-34, ResNet-50;
- функция потерь ℓ — кросс-энтропия;
- критерий качества – Accuracy.

Выборки

1. CIFAR-10, CIFAR-100, SVHN – датасеты изображений;

Выборка	Число изображений	Число классов
CIFAR-10	60000	10
CIFAR-100	60000	100
SVHN	$\sim$100000	10

Результаты: классификация изображений

Accuracy (%) методов регуляризации (ResNet-34)

Метод регуляризации	CIFAR-10	CIFAR-100	SVHN
Без регуляризации	94.53 ± 0.03	75.58 ± 0.08	96.50 ± 0.03
Minimum Hyperspherical Energy	94.58 ± 0.04	75.78 ± 0.08	96.59 ± 0.03
Weights Orthogonalization	94.59 ± 0.04	75.98 ± 0.08	96.51 ± 0.02
Spectral Restricted Isometry	94.72 ± 0.03	76.24 ± 0.09	96.57 ± 0.03
Orthogonal Convolutions	95.03 ± 0.04	76.57 ± 0.06	96.66 ± 0.02
Фреймовая регуляризация	$\mathbf{95.17} \pm \mathbf{0.05}$	$\mathbf{77.61} \pm \mathbf{0.07}$	$\mathbf{96.85} \pm \mathbf{0.02}$

Accuracy (%) методов регуляризации (ResNet-50)

Метод регуляризации	CIFAR-10	CIFAR-100	SVHN
Без регуляризации	94.83 ± 0.04	77.20 ± 0.07	96.92 ± 0.03
Minimum Hyperspherical Energy	94.88 ± 0.03	77.34 ± 0.06	96.94 ± 0.02
Weights Orthogonalization	94.92 ± 0.04	77.38 ± 0.06	96.91 ± 0.03
Spectral Restricted Isometry	95.01 ± 0.03	77.40 ± 0.07	96.95 ± 0.03
Orthogonal Convolutions	$\mathbf{95.29} \pm \mathbf{0.03}$	77.77 ± 0.07	97.01 ± 0.02
Фреймовая регуляризация	95.25 ± 0.04	$\mathbf{78.35} \pm \mathbf{0.06}$	$\mathbf{97.10} \pm \mathbf{0.02}$

- Обучающая выборка CIFAR-10;
- Тестовые домены:
  1. CIFAR-10-C – аугментированная выборка CIFAR-10,
  2. CINIC-10 – подвыборка ImageNet, включающая классы из CIFAR-10;
- Для моделей с регуляризацией выбраны субоптимальные эпохи;

Accuracy (%) методов регуляризации на разных доменах

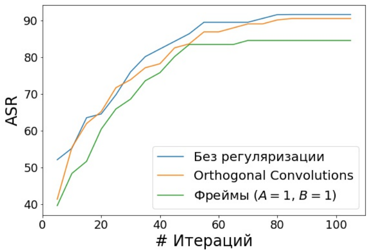
Метод регуляризации	CIFAR-10 $^{(*)}$	CIFAR-10-C	CINIC-100
Без регуляризации	94.53 ± 0.03	74.77 ± 0.25	67.91 ± 0.35
Orthogonal Convolutions	94.52 ± 0.01	76.27 ± 0.19	69.87 ± 0.29
Фреймовая регуляризация	94.53 ± 0.01	76.65 ± 0.15	71.20 ± 0.32

$^{(*)}$ – исходный домен

- Выборка CIFAR-10;
- Состязательная атака типа "черный ящик" SimBA (*Guo, 2019*);
- Attack Success Rate (ASR) – доля успешных атак;

Зависимость ASR (%) от числа итераций SimBA

Метод регуляризации	# Итераций				
	1	10	50	100	1000
Без регуляризации	52.08	59.37	84.38	92.71	93.75
Orthogonal Convolutions	41.30	57.61	83.69	91.30	92.06
Фреймовая регуляризация	39.56	49.45	80.20	84.61	86.81



Выводы

- Предложена модель нейросетевого слоя на основе фрейма в пространстве параметров, исключающая потерю информации на слое.
- Предложенная модель обобщена на сверточные слои с использованием блочно-теплицева представления свертки.
- Построен фреймовый регуляризатор параметров нейросетевого слоя путем введения штрафа за нарушение фреймового неравенства.
- Проведенные вычислительные эксперименты показали эффективность предложенного метода регуляризации в терминах точности классификации, устойчивости к состязательным атакам и к смене домена по сравнению с существующими подходами к регуляризации параметров.
- Предложенная регуляризация позволила отказаться от стандартной регуляризации weight decay путем введения штрафа на соблюдение верхней границы фрейма.