

Сравнение методов сжатия линейных слоев с помощью SVD и CUR факторизации

Команда CUR-ник

Состав команды:

- Сенотова Юлия
- Мозговых Василий
- Филимонова Ирина

Постановка задачи

Цель:

Изучение методов декомпозиции SVD и CUR для сжатия линейных слоев моделей VGG и их сравнение с исходными VGG в задачах классификации изображений.

Актуальность:

Данный подход позволяет значительно уменьшить количество параметров модели, а также увеличить производительность сети, при этом не уступив в качестве.



Основная гипотеза и оценка качества

Гипотеза:

Модель будет обладать меньшим числом параметров, но сохранит ассурасу, при этом, в рассмотренной литературе, утверждалось, что CUR превосходит SVD. В наших экспериментах предполагаем, что достигнем схожих результатов.

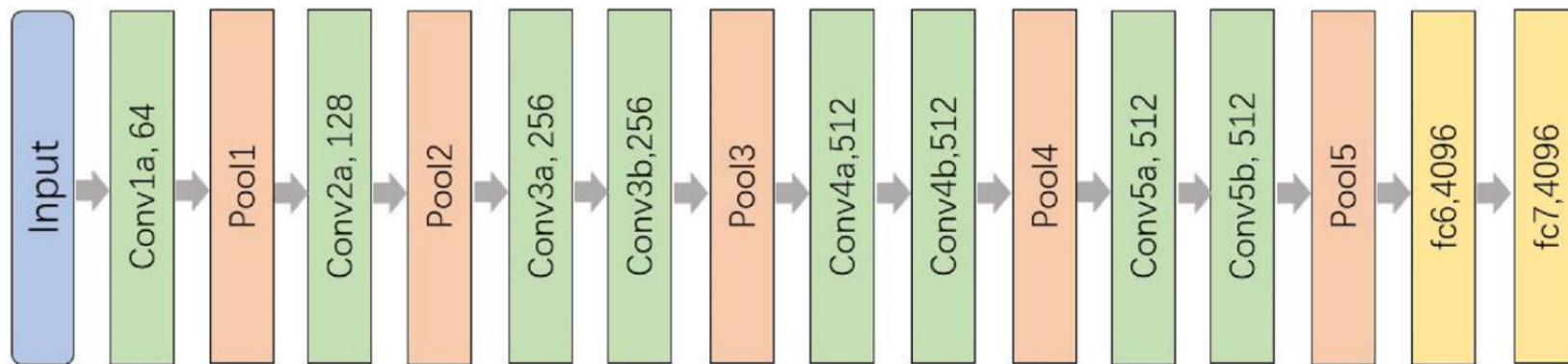
Оценка качества:

Будет производиться на стандартных наборах данных для задачи классификации изображений по метрике ассурасу для VGG-11:

- CIFAR10 (train: 50 000/test: 10 000)
- MNIST (train: 60 000/test: 10 000)



Архитектура нейросети VGG-11



Общее число параметров: 132 863 336

- *Feature extractor*: 9 220 480 (6.93%)
- *Classifier*: 123 642 856 (93.06%)

$$\text{Accuracy} = \frac{\text{Correct prediction}}{\text{Total cases}} \cdot 100\%$$



SVD decomposition

Метод SVD представляет входную матрицу весов в виде набора из трех матриц, которые при перемножении дают исходную матрицу.

$$A = U\Sigma V^*$$

- U, V - унитарные матрицы
- Σ - диагональная матрица

Сохраняя первые r сингулярных чисел, мы приближаем матрицу A

$$\tilde{A} = \tilde{U}\tilde{S}\tilde{V}^T \quad \tilde{A} \in R^{m \times n}, \quad \tilde{U} \in R^{m \times r}, \quad \tilde{S} \in R^{r \times r}, \quad \tilde{V} \in R^{n \times r}$$



Детали реализации

Исходная весовая матрица содержит $m \cdot n$ параметров. Используя SVD получаем два слоя B и C без смещения.

$$A = B * C \Rightarrow B = \tilde{U} \in R^{m \times k}, C = \tilde{S} \tilde{V}^T \in R^{k \times n}$$

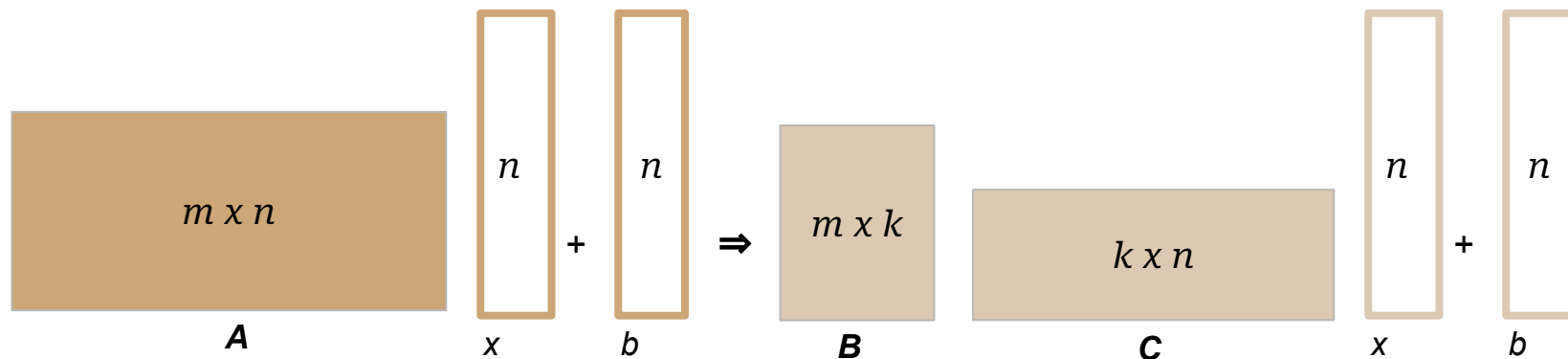
Общее количество параметров в B и C равно $k * (m + n)$

Коэффициент сжатия:

$$CR = \frac{mn}{mk + nk}$$



Схема разложения линейного слоя



A - исходная весовая матрица линейного слоя

B, C - полученные слои



CUR decomposition

Методы CUR раскладывают входную весовую матрицу в виде набора из трех матриц, которые при перемножении приближают исходную матрицу.

$$A \in R^{n \times d} \Rightarrow C \in R^{n \times c}, U \in R^{c \times d}, R \in R^{r \times d}$$

Особенности:

- C, R состоят из столбцов и строк исходной матрицы, поэтому они лучше поддаются интерпретации
- Аппроксимация не уникальна, существуют разные алгоритмы для вычисления
 - *LinearTimeCUR*
 - *LeverageScoreCUR*
- Существуют алгоритмы с более низкой сложностью, чем SVD



Алгоритмы построения CUR

LTCUR

$$p_i = \|A_i\| / \|A\|_F$$

A_i с вероятностью p_i в C + нормирование

$$Q = \text{Alg}(C)$$

$$U = (C^T C_k)^{-1} Q$$

LSCUR

SVD

$$p_i = (1/k) \sum_j V_{ij}^2$$

A_i с вероятностью $\min(1, c p_i)$ в C , нормирование

$$U = C^+ A R^+$$

	# PASSES	RUN TIME	SPACE USAGE	ERROR BOUND
LTCUR[3]	2	$O(n(d + k/\varepsilon^2 + 1/\varepsilon^8) + d/\varepsilon^4)$	$O(n/\varepsilon^4 + dk/\varepsilon^2)$	$\ A - CUR\ _2 \leq OPT_2 + \varepsilon \ A\ _F$
		$O(n(d + k/\varepsilon^2 + k^2/\varepsilon^8) + dk/\varepsilon^4)$		$\ A - CUR\ _F \leq OPT_F + \varepsilon \ A\ _F$
LSCUR[4]	2	$O(k \log k / \varepsilon^2 n^2 d + nd^2 k \log k / \varepsilon^2)$	$O(nd)$	$\ A - CUR\ _F \leq (2 + \varepsilon) OPT_F$

$$OPT_2 = \|A - A_k\|_2 \quad OPT_F = \|A - A_k\|_F$$



Детали реализации

Исходная весовая матрица содержит $m \times n$ параметров. Используя CUR получаем два слоя D и E .

$$A = D * E \Rightarrow D = C \in R^{m \times c}, E = U * R \in R^{c \times n}$$

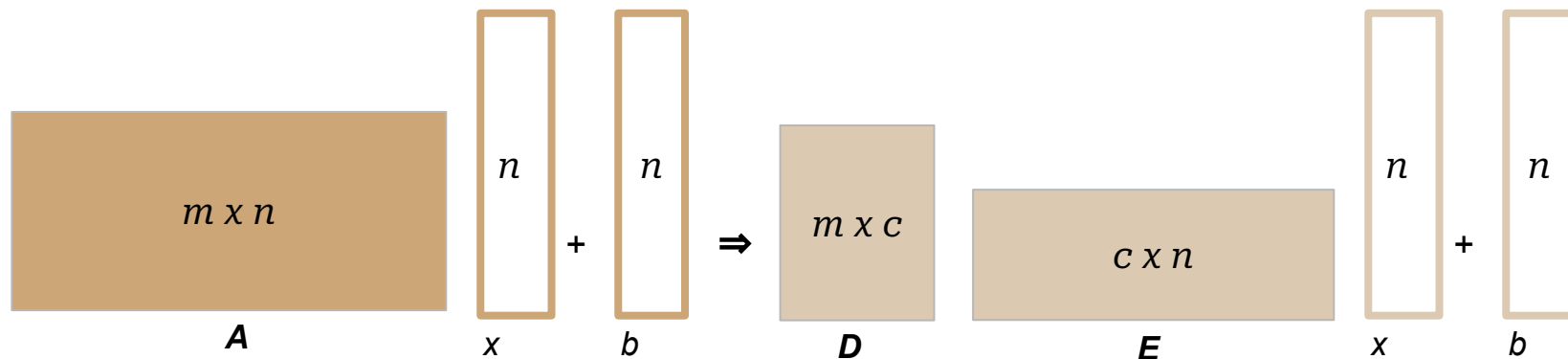
Общее количество параметров в D и E равно $c * (m + n)$

Коэффициент сжатия:

$$CR = \frac{mn}{mc + nc}$$



Схема разложения линейного слоя



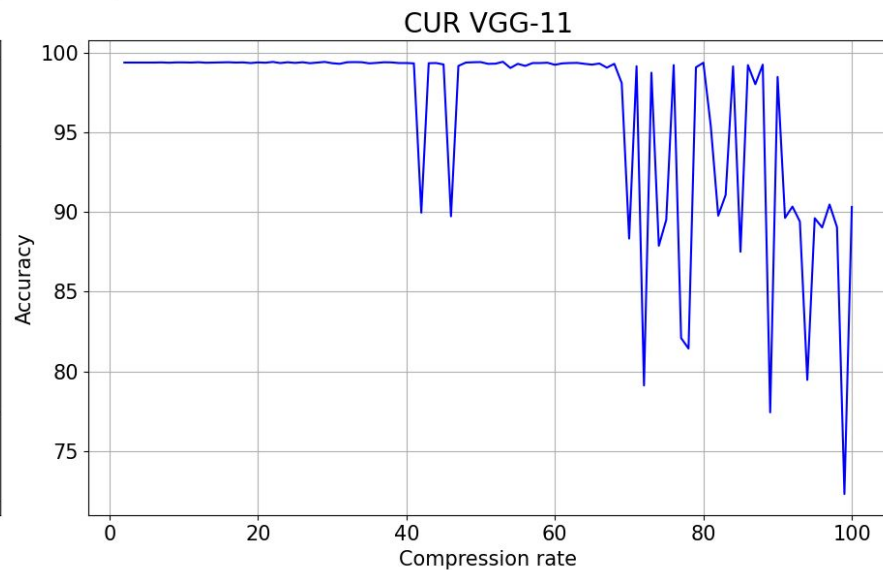
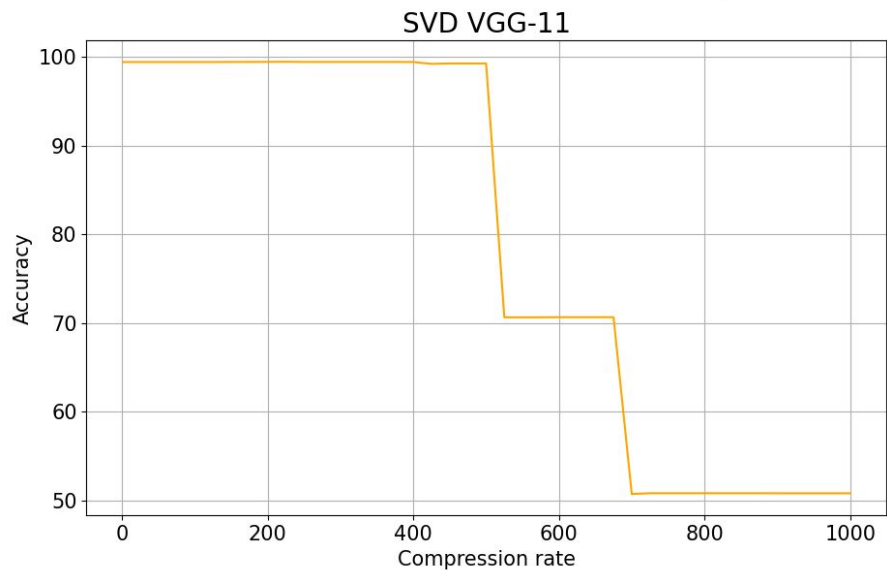
A - исходная весовая матрица линейного слоя

D, E - полученные слои



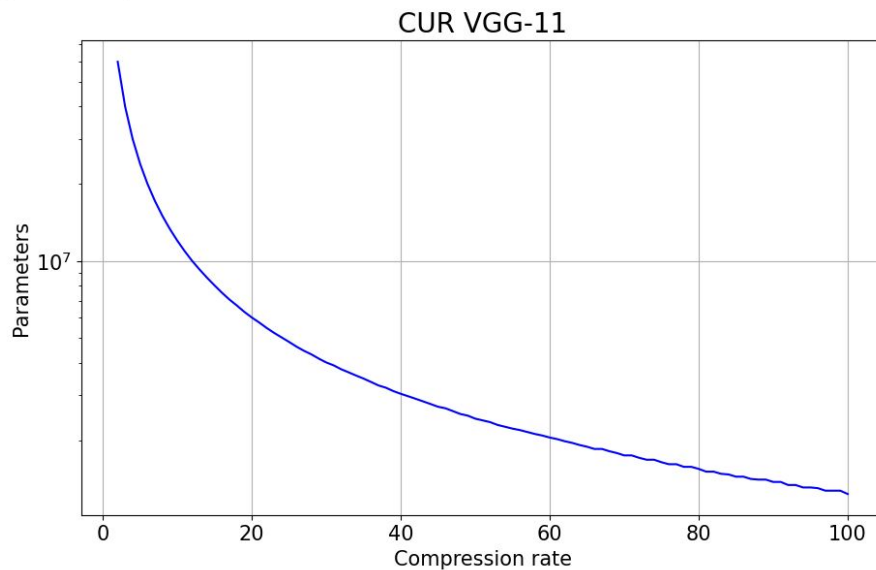
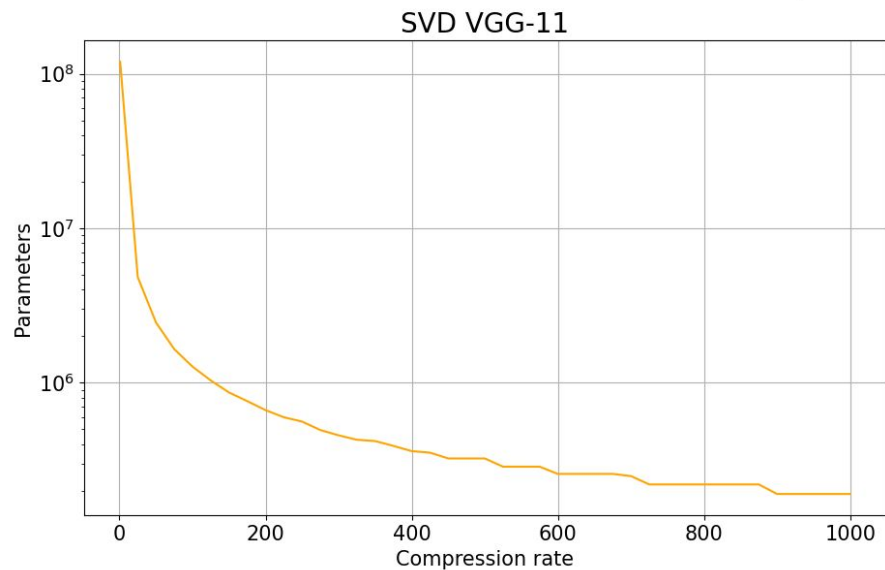
Результаты (MNIST)

Accuracy depending on compression rate



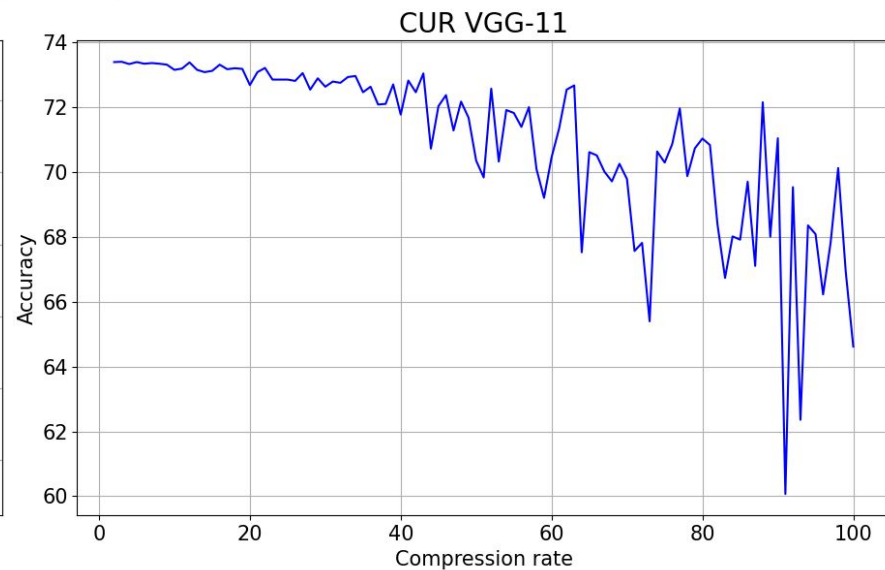
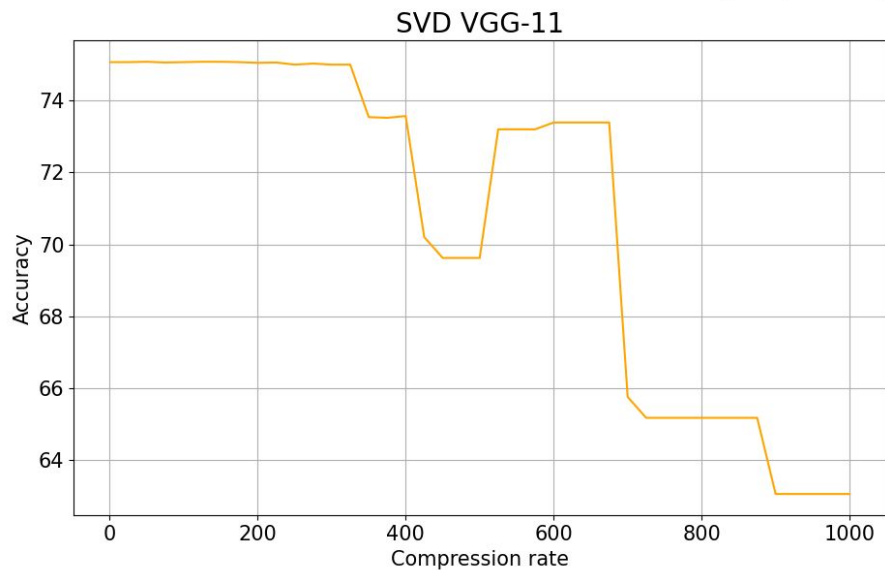
Результаты (MNIST)

Parameters depending on compression rate



Результаты (CIFAR10)

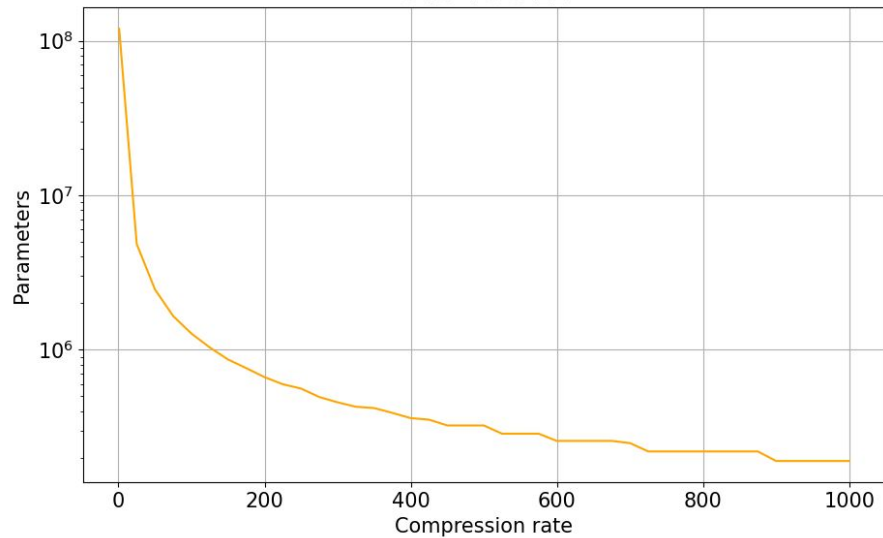
Accuracy depending on compression rate



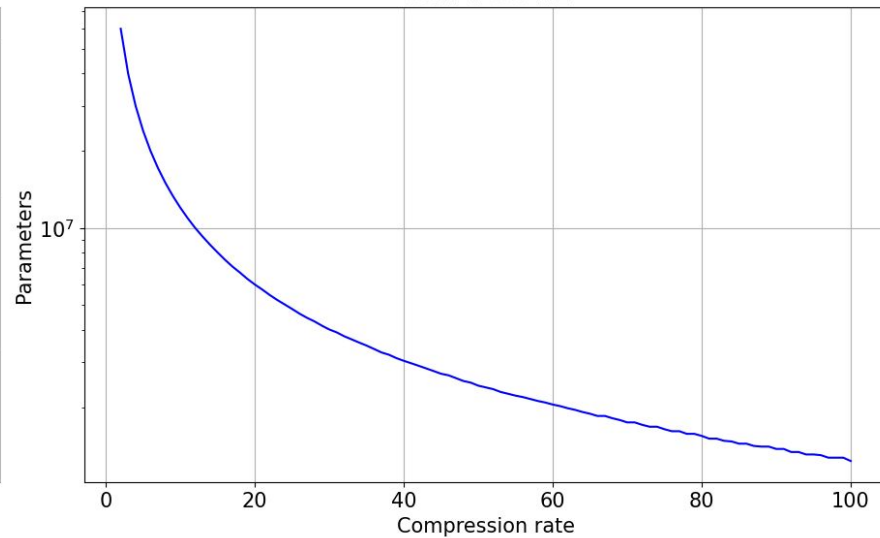
Результаты (CIFAR10)

Parameters depending on compression rate

SVD VGG-11



CUR VGG-11



MNIST			
Модель	Accuracy	Count Parameters	Best CR
VGG-11	99,40%	132 863 336	1
VGG-11 + SVD	99,44%	598 126	225
VGG-11 + CUR	99,44%	4 162 058	29

CIFAR-10			
Модель	Accuracy	Count Parameters	Best CR
VGG-11	73,42%	132 863 336	1
VGG-11 + SVD	75,08%	864 366	150
VGG-11 + CUR	73,39%	10 000 906	12

Вывод:

SVD и CUR позволяют сжимать линейные слои, не ухудшая общее качество модели


ВЫВОДЫ

- Сжатые модели не ухудшились по метрике на выбранных наборах данных
- SVD позволил значительно сжать слои VGG-11 без потери качества
- Гипотеза о том, что CUR разложение достигнет лучшего результата по сравнению с SVD, не подтвердилась (*это может быть связано с использованием разных реализаций CUR*)


Дальнейшие планы:

- Провести эксперименты с другими наборами данных и моделями
- Реализовать сжатие сверточных слоев





https://github.com/VasilyMozgovykh/nn_linear_compression



Литература

- [1] Mai, An, et al. "VGG deep neural network compression via SVD and CUR decomposition techniques." 2020 7th NAFOSTED Conference on Information and Computer Science (NICS). IEEE, 2020.
- [2] Hamm, Keaton, and Longxiu Huang. "Cur decompositions, approximations, and perturbations." (2019).
- [3] Petros Drineas, Ravi Kannan, and Michael W Mahoney. Fast monte carlo algorithms for matrices iii: Computing a compressed approximate matrix decomposition. SIAM Journal on Computing, 36(1):184–206, 2006.
- [4] Michael W Mahoney and Petros Drineas. Cur matrix decompositions for improved data analysis. Proceedings of the National Academy of Sciences, 106(3):697–702, 2009





Спасибо за внимание!