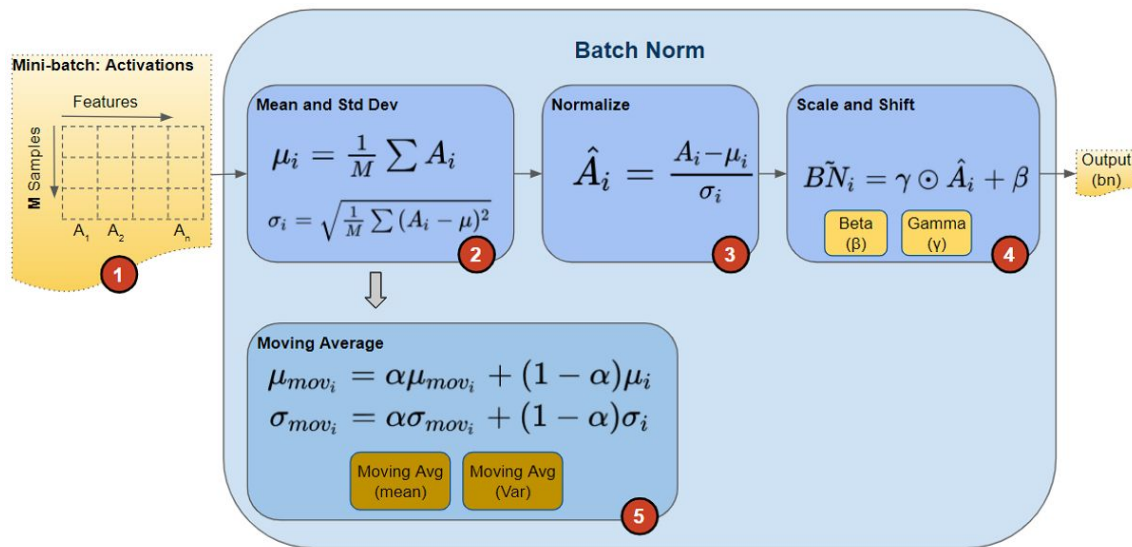


On the Periodic Behavior of Neural Network Training with Batch Normalization and Weight Decay

Напоминание: batch normalization [1]

Суть данного метода заключается в том, что некоторым слоям нейронной сети на вход подаются данные, предварительно обработанные и имеющие нулевое математическое ожидание и единичную дисперсию.



Напоминание: weight decay [2]

Это регуляризация.

$$\text{Loss}(y, y') = L(y, y') + \text{weight_decay} * \text{sum}(x^2)$$

Напоминание: scale-invariant функции

Это функции для которых верно, что $f(ax) = f(x)$ для любого $a > 0$.

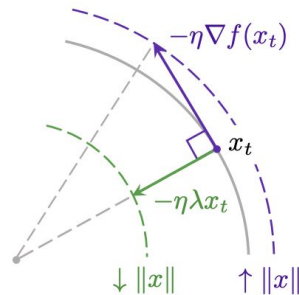
Как выглядит обучение scale-invariant нейронной сети

Для scale-invariant функций верно, что:

$$\begin{cases} \langle \nabla f(x), x \rangle = 0, \forall x \\ \nabla f(\alpha x) = \frac{1}{\alpha} \nabla f(x), \forall x, \alpha > 0. \end{cases}$$

Если нейронная сеть является scale-invariant функцией, то при ее обучении с использованием регуляризации, изменение параметров формироваться из двух компонент:

$$x_{t+1} = x_t - \eta \lambda x_t - \eta \nabla f(x_t)$$



А как сделать нейронную сеть scale-invariant?

- Добавить больше batch normalization слоев
- Зафиксировать обучаемые параметры batch normalization слоев
- Зафиксировать веса последнего слоя

Архитектура моделей и обучения в этом исследовании

- Функция потерь – кросс энтропия
- Размер батча 128
- Обучение длится 1000 эпох

Дополнительная терминология

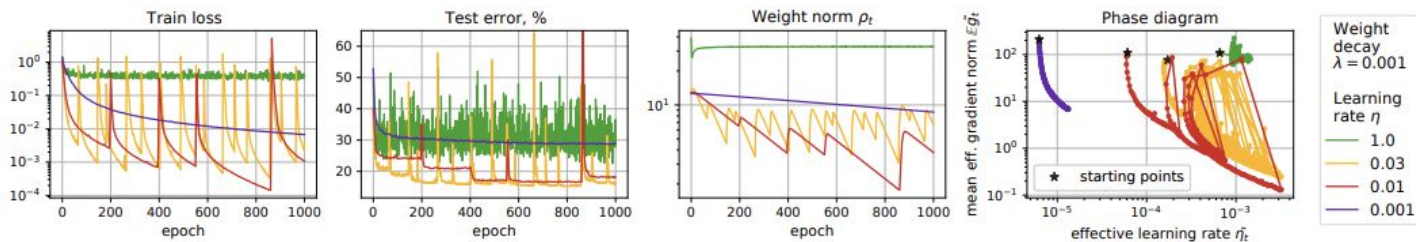
Эффективный градиент – градиент для точки на единичной сфере. Для scale-invariant функций это $\nabla f(x/\|x\|) = \nabla f(x)\|x\|$.

Эффективный learning rate – $\eta_{\text{eff}} = \eta/\|x\|^2$.

Что за periodic behaviour?

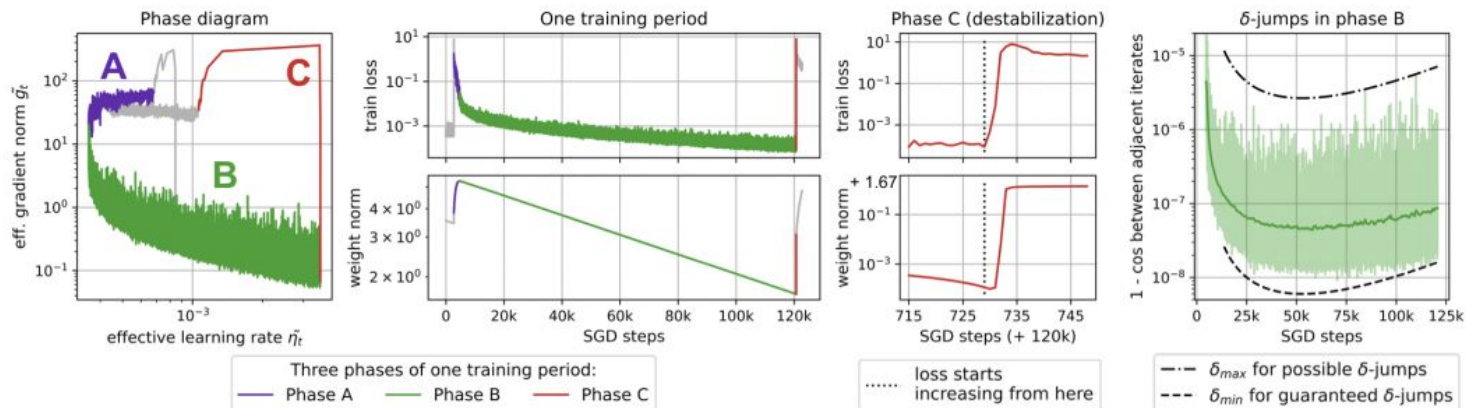
Авторы заметили, что различные статистики (ошибка на обучающей выборке, на тестовой, норма весов и т.д.) ведут себя практически как периодические функции в процессе обучения при следующих условиях:

- Используемая нейронная сеть является scale-invariant функцией
- Используется стандартный SGD с weight decay и константным learning rate



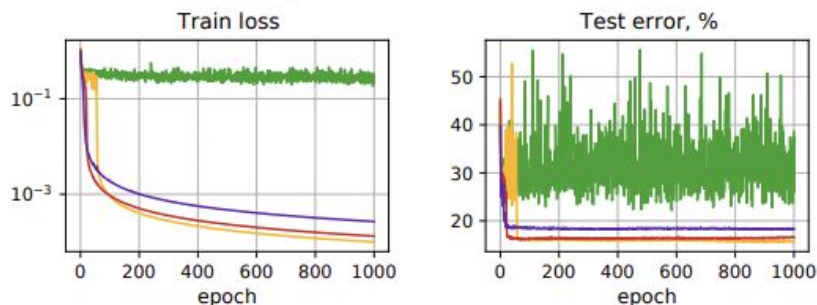
Периодическое поведение scale-invariant ConvNet на датасете CIFAR-10

А что происходит внутри одного “периода”?

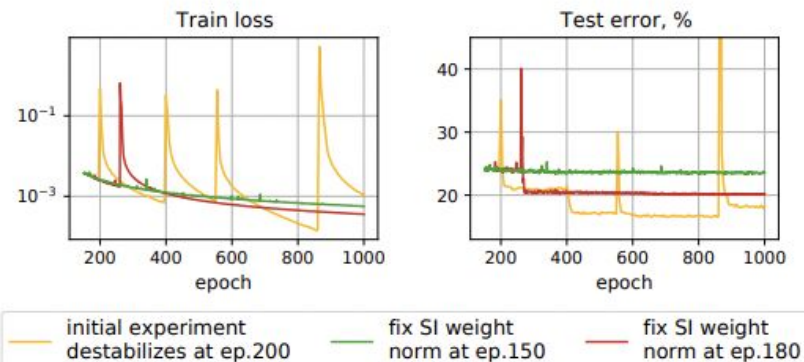


Это точно из-за batch normalization и регуляризации?

Fix weight norm at initialization



Fix weight norm before destabilization



Почему случается фаза C?

Пусть процесс обучения совершает **δ -прыжок**, если косинусное расстояние весов модели между итерациями превысило некоторое $\delta > 0$:

$$1 - \cos(x_t, x_{t+1}) > \delta$$

Предположим, что если модель не совершает δ -прыжок со значительным δ , то дестабилизации не происходит. Тогда в каких случаях модель совершает δ -прыжок?

Утверждение 1 Если $f(x)$ scale-invariant функция оптимизируемая с ограниченным эффективным градиентом $0 \leq l \leq g_t \leq L < +\infty$, тогда для достаточно малого δ и $(1-\eta\lambda) \leq 1$ верны следующие условия:

$$\|x_t\|^2 \leq \eta L / \sqrt{2\delta} \Rightarrow \delta\text{-прыжок возможен}$$

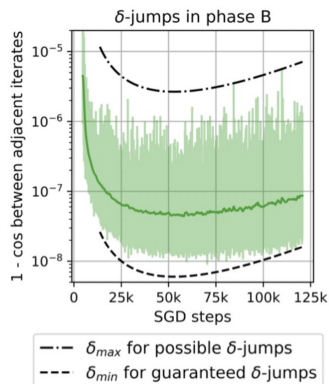
$$\|x_t\|^2 \leq \eta l / \sqrt{2\delta} \Rightarrow \delta\text{-прыжок гарантирован}$$

Почему случается фаза С?

Утверждение 1 Если $f(x)$ scale-invariant функция оптимизируемая с ограниченным эффективным градиентом $0 \leq l \leq g_t \leq L < +\infty$, тогда для достаточно малого δ и $(1-\eta\lambda) \leq 1$ верны следующие условия:

$$\|x_t\|^2 \leq \eta L / \sqrt{2\delta} \Rightarrow \delta\text{-прыжок возможен}$$

$$\|x_t\|^2 \leq \eta l / \sqrt{2\delta} \Rightarrow \delta\text{-прыжок гарантирован}$$



Когда случается фаза С?

Утверждение 2 Пусть $\kappa = \sqrt{(\eta/2\lambda)}$. Если $f(x)$ scale-invariant функция оптимизируемая с ограниченным эффективным градиентом $0 \leq l \leq g t \leq L < +\infty$ и $2\eta\lambda L \leq l$:

1. Если $\rho_0^2 > \kappa l$ и $\delta < \eta\lambda(L^2/l^2)$, тогда **минимальное** время требующееся для δ -прыжка

$$t_{\min} = \max \left\{ 0, \frac{\log(\rho_0^2 - \kappa l) - \log\left(\frac{\eta L}{\sqrt{2\delta}} - \kappa l\right)}{-\log(1 - 4\eta\lambda)} \right\}$$

2. Если $\rho_0^2 > \kappa L$ и $\delta < \eta\lambda(L^2/l^2)$, тогда **максимальное** время требующееся для δ -прыжка

$$t_{\max} = \max \left\{ 0, \frac{\log(\rho_0^2 - \kappa L) - \log\left(\frac{\eta l}{\sqrt{2\delta}} - \kappa L\right)}{-\log(1 - 2\eta\lambda)} \right\}$$

Почему норма весов так себя ведет?

Утверждение 3 Пусть $\kappa = \sqrt{(\eta/2\lambda)}$. Если $f(x)$ scale-invariant функция оптимизируемая с ограниченным эффективным градиентом $0 \leq l \leq g_t \leq L < +\infty$ и $2\eta\lambda L \leq l$, тогда $\kappa l \leq \rho_t^2 \leq \kappa L$, при $t \gg 1$. Более того, если $\rho_0^2 > \kappa L$, то ρ_t^2 сходится $[\kappa l, \kappa L]$ линейно за $O(1/\eta\lambda)$ итераций.

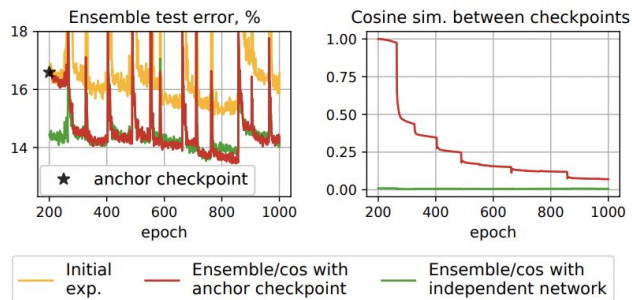
Эмпирические наблюдения

- При использовании более высокого learning rate периодическое поведение стабилизируется при более высокой норме весов. То же самое верно и для weight decay
- В начале процесса обучения наблюдается *разогревочная* фаза. При достижении определенной нормы весов периодическое поведение становится стабильным

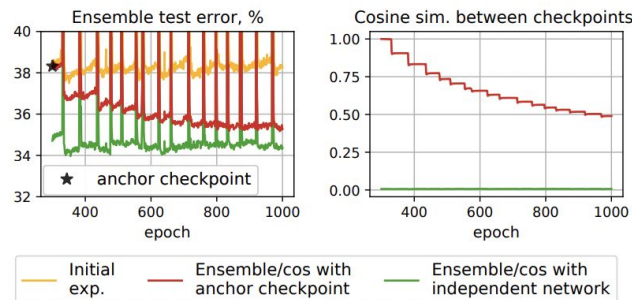
Эмпирические наблюдения

Если обучение сходится к одному и тому же минимуму в разные периоды, то косинусная близость весов должна быть ближе к 1, а ошибка ансамбля этих моделей должна быть близка к ошибке любой из них. Напротив, если дестабилизация приводит к слишком большому скачку весов это эквивалентно обучению модели из новой случайной точки, соответственно, косинусная близость двух минимумов должна быть ближе к 0 и ошибка ансамбля этих моделей должна быть близка к ошибке ансамбля двух независимо обученных моделей.

ConvNet on CIFAR-10

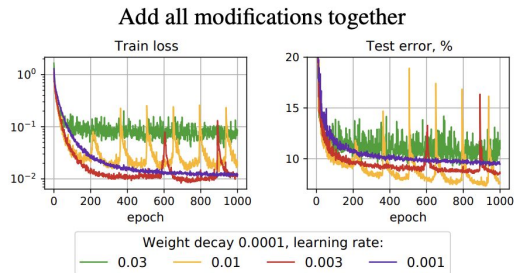
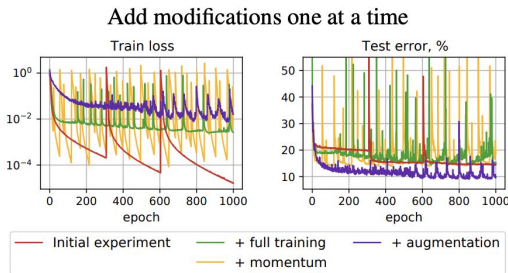


ResNet-18 on CIFAR-100



Что происходит на практике?

- При использовании не scale-invariant нейронной сети периодичность сохраняется, однако меняется частота периодов
- При использовании SGD с моментумом периодичность сохраняется, однако увеличивается частота периодов
- При использовании аугментации периодичность поведения начинается наблюдается значительно позже, так как более высокая ошибка приводит к “перевесу” градиентной части направления оптимизации



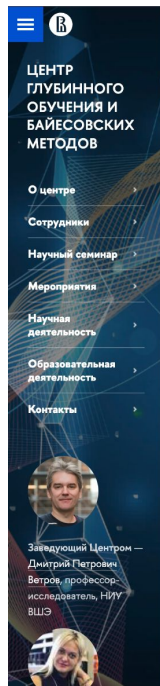
Библиография

1. Ekaterina Lobacheva, Maxim Kodryan, Nadezhda Chirkova, Andrey Malinin, Dmitry Vetrov. *On the Periodic Behavior of Neural Network Training with Batch Normalization and Weight Decay*. NeurIPS, 2021.
2. [Batch Norm Explained Visually – How it works, and why neural networks need it](#)
3. [This thing called Weight Decay](#)
4. Shibani Santurkar, Dimitris Tsipras, Andrew Ilyas, Aleksander Madry. *How Does Batch Normalization Help Optimization?* NeurIPS, 2018.

Обзор-рецензия

Авторы статьи

- Сотрудники [центра
глубинного обучения и
байесовских методов
НИУ ВШЭ](#)
- Преподаватели на
курсах ФКН ПМИ



Национальный исследовательский университет «Высшая школа экономики» → Факультет компьютерных наук → Департамент больших данных и информационного поиска → Центр глубинного обучения и байесовских методов

RU EN

Центр глубинного обучения и байесовских методов

О Центре

Центр ведет исследования на стыке двух активно развивающихся сегодня областей анализа данных: глубинного обучения и байесовских методов машинного обучения. Глубинное обучение - это раздел, подразумевающий построение очень сложных моделей (нейронных сетей) для решения таких задач, как классификация изображений или музыки, перенос художественного стиля с картины на фотографию, предсказание следующих слов в тексте. В рамках байесовского подхода для решения подобных задач рассматриваются вероятностные модели, опирающиеся на аппарат теории вероятностей и математической статистики.

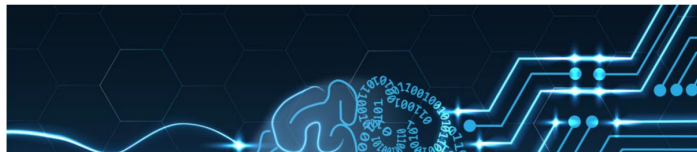
Центр создан на основе исследовательской группы байесовских методов машинного обучения Д.П. Ветрова.

Новости ▾



10

январь
2022



ПУБЛИКАЦИИ

Книга

*Proceedings of the 15th
International Joint
Conference on Computer
Vision, Imaging and
Computer Graphics Theory
and Applications*

Kochurov M., Volkhonskiy D., Yashkov
D. et al.

Vol. 4. SoTePress, 2020.

Статья

*A randomized coordinate
descent method with volume
sampling*

Rodomanov A., Kropotov D.

SIAM Journal on Optimization. 2020.
Vol. 30. No. 3. P. 1878-1904.

Глава в книге

*Leveraging Recursive
Gumbel-Max Trick for
Approximate Inference in
Combinatorial Spaces*

Kirill Struminsky, Artyom Gadetsky,
Denis Rakin et al.

*In bk.: Advances in Neural Information
Processing Systems 34 (NeurIPS
2021)*

35-ая конференция NeurIPS 2021

- Статья на Хабре: [“Самое важное с конференции NeurIPS 2021”](#)
- 10 статей исследователей ФКН [приняты на конференцию](#)

На что опирались авторы?

- Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift.
- Bjorck, N., Gomes, C. P., Selman, B., and Weinberger, K. Q. (2018). Understanding batch normalization.

Цитирования

- 10 цитирований и ссылок по google scholar
- Упоминание в [towardsdatascience](https://towardsdatascience.com/)

Сильные стороны

- Простая интересная идея
- Несложные математические выкладки
- Неожиданный результат

Слабые стороны и точки роста

- Как избежать такого поведения при обучении своей сети?
- Много условий на данные и методы

Собственно это я бы и хотел предложить в качестве улучшений: описать как избежать такого поведения у сети и попробовать обобщить теорию.