КОМПЬЮТЕРНЫЕ ИССЛЕДОВАНИЯ
И МОДЕЛИРОВАНИЕ 2023 Т. 15 № 1 С. 45–56
DOI: 10.20537/2076-7633-2023-15-1-45-56

Ки&М

ЧИСЛЕННЫЕ МЕТОДЫ И ОСНОВЫ ИХ РЕАЛИЗАЦИИ

УДК: 004.853

Современные методы преодоления катастрофической забывчивости нейронных сетей и экспериментальная проверка вопросов их структуры

А. А. Куталев[1,a], А. А. Лапина[2,b]

[1]ПАО «Сбербанк»,
Россия, 121170, г. Москва, Кутузовский пр-т, д. 32
[2]MY.GAMES,
Россия, 125167, г. Москва, Ленинградский пр-т, 39, стр. 79

E-mail: [a] kutalev@gmail.com, [b] ahm.alisa@gmail.com

Получено 12.10.2022, после доработки — 14.12.2022.
Принято к публикации 24.12.2022.

В данной работе представлены результаты экспериментальной проверки некоторых вопросов, касающихся практического использования методов преодоления катастрофической забывчивости нейронных сетей. Проведено сравнение двух таких современных методов: метода эластичного закрепления весов (EWC, Elastic Weight Consolidation) и метода ослабления скоростей весов (WVA, Weight Velocity Attenuation). Разобраны их преимущества и недостатки в сравнении друг с другом. Показано, что метод эластичного закрепления весов (EWC) лучше применять в задачах, где требуется полностью сохранять выученные навыки на всех задачах в очереди обучения, а метод ослабления скоростей весов (WVA) больше подходит для задач последовательного обучения с сильно ограниченными вычислительными ресурсами или же когда требуется не точное сохранение всех навыков, а переиспользование репрезентаций и ускорение обучения от задачи к задаче. Проверено и подтверждено интуитивное предположение, что ослабление метода WVA необходимо применять к оптимизационному шагу, то есть к приращениям весов нейронной сети, а не к самому градиенту функции потерь, и это справедливо для любого градиентного оптимизационного метода, кроме простейшего стохастического градиентного спуска (SGD), для которого оптимизационный шаг и градиент функции потерь пропорциональны. Рассмотрен выбор оптимальной функции ослабления скоростей весов между гиперболической функцией и экспонентой. Показано, что гиперболическое убывание более предпочтительно, так как, несмотря на сравнимое качество при оптимальных значениях гиперпараметра метода WVA, оно более устойчиво к отклонениям гиперпараметра от оптимального значения (данный гиперпараметр в методе WVA обеспечивает баланс между сохранением старых навыков и обучением новой задаче). Приведены эмпирические наблюдения, которые подтверждают гипотезу о том, что оптимальное значение гиперпараметра не зависит от числа задач в очереди последовательного обучения. Следовательно, данный гиперпараметр может подбираться на небольшом числе задач, а использоваться — на более длинных последовательностях.

Ключевые слова: катастрофическая забывчивость, эластичное закрепление весов, EWC, ослабление скоростей весов, WVA, нейронные сети, последовательное обучение, машинное обучение, искусственный интеллект

COMPUTER RESEARCH AND MODELING
2023 VOL. 15 NO. 1 P. 45–56
DOI: 10.20537/2076-7633-2023-15-1-45-56

Ки&М

NUMERICAL METHODS AND THE BASIS FOR THEIR APPLICATION

UDC: 004.853

Modern ways to overcome neural networks catastrophic forgetting and empirical investigations on their structural issues

A. A. Kutalev[1,a], A. A. Lapina[2,b]

[1]PJSC Sberbank,
32 Kutuzovskiy ave., Moscow, 121170, Russia
[2]MY.GAMES,
39/79 Leningradskiy ave., Moscow, 125167, Russia

E-mail: [a] kutalev@gmail.com, [b] ahm.alisa@gmail.com

Received 12.10.2022, after completion — 14.12.2022.
Accepted for publication 24.12.2022.

This paper presents the results of experimental validation of some structural issues concerning the practical use of methods to overcome catastrophic forgetting of neural networks. A comparison of current effective methods like EWC (Elastic Weight Consolidation) and WVA (Weight Velocity Attenuation) is made and their advantages and disadvantages are considered. It is shown that EWC is better for tasks where full retention of learned skills is required on all the tasks in the training queue, while WVA is more suitable for sequential tasks with very limited computational resources, or when reuse of representations and acceleration of learning from task to task is required rather than exact retention of the skills. The attenuation of the WVA method must be applied to the optimization step, i. e. to the increments of neural network weights, rather than to the loss function gradient itself, and this is true for any gradient optimization method except the simplest stochastic gradient descent (SGD). The choice of the optimal weights attenuation function between the hyperbolic function and the exponent is considered. It is shown that hyperbolic attenuation is preferable because, despite comparable quality at optimal values of the hyperparameter of the WVA method, it is more robust to hyperparameter deviations from the optimal value (this hyperparameter in the WVA method provides a balance between preservation of old skills and learning a new skill). Empirical observations are presented that support the hypothesis that the optimal value of this hyperparameter does not depend on the number of tasks in the sequential learning queue. And, consequently, this hyperparameter can be picked up on a small number of tasks and used on longer sequences.

Keywords: catastrophic forgetting, elastic weight consolidation, EWC, weight velocity attenuation, WVA, neural networks, continual learning, machine learning, artificial intelligence

Citation: *Computer Research and Modeling*, 2023, vol. 15, no. 1, pp. 45–56 (Russian).

Введение

Этот раздел призван описать проблематику катастрофической забывчивости нейронных сетей и актуальные методы ее решения для тех, кто с ней не знаком совсем или знаком понаслышке. Более подкованный читатель может сразу перейти к следующему разделу.

Описание проблематики

Итак, суть проблемы катастрофической забывчивости нейронных сетей заключается в том, что при последовательном обучении задачам A и B, то есть когда после обучения задаче A происходит обучение нейронной сети задаче B без доступа к данным задачи A, нейронная сеть в процессе обучения задаче B быстро утрачивает навык, полученный при обучении задаче A.

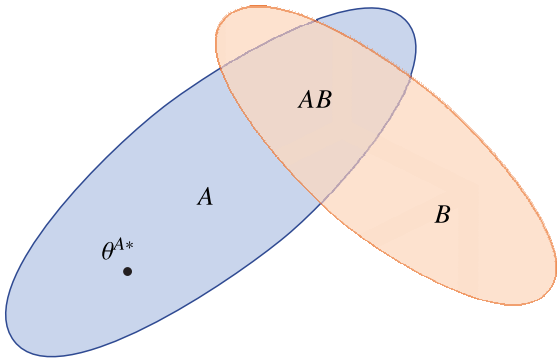


Рис. 1. Области в пространстве весов нейронной сети: A — область, являющаяся решением задачи A; аналогичным образом B — область, являющаяся решением задачи B. Пересечение этих множеств обозначено как AB и представляет собой область в пространстве весов, которая является решением обеих задач. Точка θ^{A*} — локальный минимум функции потерь, найденный в процессе решения задачи A градиентным методом

К проблеме преодоления катастрофической забывчивости на протяжении длительного времени прилагались значительные усилия (см., например, работы [McCloskey, Cohen, 1989; McClelland, McNaughton, O'Reilly, 1995; French, 1999; Goodfellow et al., 2015]), которые, однако, не привели к существенным общим результатам.

Большой прогресс был достигнут в 2017 году, когда группой из DeepMind был предложен метод эластичного закрепления весов EWC [Kirkpatrick et al., 2017]. Этот метод основан на нескольких предположениях. Во-первых, предполагается, что в пространстве весов нейронной сети существует область, являющаяся решением как для уже изученных задач, так и для задач в очереди для последующего обучения (то есть области в пространстве весов, являющиеся решениями каждой задачи, имеют непустое пересечение). Во-вторых, предполагается, что веса нейронной сети имеют разную важность для уже изученных задач, и, при обучении новым задачам, стоит удерживать веса нейронной сети от изменений тем сильнее, чем больше их важности.

Тогда как для первого предположения необходимы лишь непротиворечивость всех изучаемых нейронной сетью задач и достаточная емкость сети; для обоснования второго предположения можно предложить рассуждения из следующего подраздела.

Теоретическое обоснование метода EWC

Пусть для обучения нейронной сети с параметрами (весами) θ на данных D используется функция потерь $L_D(\theta) = -\log p(D|\theta)$ (negative log loss — наиболее широко используемая функция

потерь). Тогда при обучении на данных D, являющихся объединением двух наборов данных A и B, при условии независимости A и B имеем в качестве функции потерь

$$L_D(\theta) = -\log p(D|\theta) = -\log p(A|\theta) - \log p(B|\theta) = L_A(\theta) + L_B(\theta). \quad (1)$$

Допустим, что нейронная сеть уже была обучена на наборе данных A с необходимой точностью. То есть достигнута сходимость некоего градиентного метода обучения нейронной сети на функции потерь $L_A(\theta)$ к локальному минимуму θ^* для задачи A.

Сходимость означает, что в достигнутой точке θ^* градиент функции потерь равен нулю или пренебрежимо мал. Тогда, раскладывая функцию потерь L_A в некоторой окрестности точки-решения θ^* задачи A в ряд Тэйлора с точностью до членов второго порядка, получим

$$L_A(\theta) \approx L_A(\theta^*) + \sum_{i,j} H_{ij}(\theta^*)\left(\theta_i - \theta_i^*\right)\left(\theta_j - \theta_j^*\right), \quad (2)$$

где $H_{ij}(\theta^*)$ — матрица Гессе от функции потерь в точке θ^*. Так как точка θ^* является решением задачи A, то есть функция потерь L_A имеет в точке θ^* локальный минимум, то и матрица Гессе $H_{ij}(\theta^*)$ положительно определена.

Авторы [Kirkpatrick et al., 2017] утверждают, что согласно работам [MacKay, 1992] и [Pascanu, Bengio, 2013] вышеупомянутый гессиан может быть аппроксимирован формой

$$\sum_{i,j} H_{ij}\left(\theta_i - \theta_i^*\right)\left(\theta_j - \theta_j^*\right) \approx \sum_i F_i\left(\theta_i - \theta_i^*\right)^2, \quad (3)$$

где $F_i(\theta^*) = \left(\frac{\partial \log p(A|\theta^*)}{\partial \theta_i}\right)^2$ — диагональные элементы информационной матрицы Фишера. Как видим, эти элементы зависят лишь от первых производных функции потерь L_A и поэтому легко вычислимы.

Учитывая (2) и (3), получаем, что в окрестности точки θ^* функция потерь $L_A(\theta)$ аппроксимируется формой $\sum_i F_i\left(\theta_i - \theta_i^*\right)^2$ с точностью до константы.

Далее, следуя логике последовательного обучения, необходимо продолжить обучение нейронной сети на задаче B, сохраняя при этом навык, полученный при обучении задаче A, то есть обучить нейронную сеть на данных $D = A \cup B$. Для этого требуется оптимизировать общую функцию потерь $L_D(\theta)$. Но на этапе дообучения на задаче B данные задачи A уже недоступны. Поэтому в формуле (1) компоненту $L_A(\theta)$ заменим аппроксимациями из (2) и (3) и отбросим константу $L_A(\theta^*)$, поскольку она не влияет на оптимизацию:

$$L_{AB}(\theta) \approx L_B(\theta) + \frac{\lambda}{2}\sum_i F_i\left(\theta_i^*\right)\left(\theta_i - \theta_i^*\right)^2. \quad (4)$$

Здесь был введен коэффициент λ, определяющий баланс вкладов от функции потерь L_B и аппроксимации функции потерь L_A в общую функцию потерь L_{AB}. Согласно [Kirkpatrick et al., 2017] форма $F_i\left(\theta_i^*\right)\left(\theta_i - \theta_i^*\right)^2$ недооценивает гессиан от L_A, и, подобрав λ, можно компенсировать эту недооценку.

Таким образом, дообучение нейронной сети на данных B сводится к оптимизации функции потерь в виде (4), причем за счет ее второй части веса θ_i нейронной сети сопротивляются изменениям тем сильнее, чем больше их важности, в качестве которых выступают F_i, рассчитанные в точке θ^*, что и требовалось показать.

Дальнейшее развитие метода EWC

В дальнейшем в работах [Zenke, Poole, Ganguli, 2017; Aljundi et al., 2018; Куталев, 2020] были предложенны альтернативные способы расчета важности весов. Причем, как показывают эксперименты в [Aljundi et al., 2018; Куталев, Лапина, 2021], метод MAS из [Aljundi et al., 2018] дает наиболее оптимальные значения важностей весов для сохранения предыдущих навыков при использовании вместо F_i в функции потерь (4) при последовательном обучении.

Для обобщения метода на случай последовательного обучения нескольким задачам авторы [Kirkpatrick et al., 2017] предлагают для обучения каждой последующей задаче K добавлять в функцию потерь компоненты, аппроксимирующие функцию потерь для каждой уже изученной задачи A, B, . . . , J:

$$L(\theta) = L_K(\theta) + \frac{\lambda_A}{2}\sum_i F_i\left(\theta_i^{A*}\right)\left(\theta_i - \theta_i^{A*}\right)^2 + \frac{\lambda_B}{2}\sum_i F_i\left(\theta_i^{B*}\right)\left(\theta_i - \theta_i^{B*}\right)^2 + \cdots + \frac{\lambda_J}{2}\sum_i F_i\left(\theta_i^{J*}\right)\left(\theta_i - \theta_i^{J*}\right)^2. \tag{5}$$

Однако, как показано в [Huszár, 2018], для обучения каждой последующей задаче K правильнее суммировать все важности для веса на уже изученных задачах $F_i\left(\theta^{A*}\right) + F_i\left(\theta^{B*}\right) + \cdots + F_i\left(\theta^{J*}\right)$ и использовать в качестве точки притяжения веса сети θ^{J*} после последней выученной задачи J:

$$L(\theta) = L_K(\theta) + \frac{\lambda}{2}\sum_i \left[F_i\left(\theta_i^{A*}\right) + \cdots + F_i\left(\theta_i^{J*}\right)\right]\left(\theta_i - \theta_i^{J*}\right)^2, \tag{6}$$

что существенно сокращает вычислительные затраты по сравнению с формулой (5).

Для балансировки между сохранением навыков каждой из уже изученных задач и обучением текущей задаче в формуле (6) можно суммировать $F_i\left(\theta_i^*\right)$ с коэффициентами пропорционально важности изученных задач. Похожий механизм был использован в [Schwarz et al., 2018] для дисконтирования важности задач по мере их удаления от текущего обучения и получил название *Online EWC*, на которое часто ссылаются в современных работах по катастрофической забывчивости.

Особенности практического использования EWC

Как было упомянуто в работе [Куталев, Лапина, 2021], при использовании метода EWC для сверточных, рекуррентных сетей или сетей с вниманием (self-attention) важности некоторых весов получаются экстремально большими, и это приводит к нежелательным эффектам при применении EWC для обучения нейронных сетей, таким как «взрыв градиентов» или потеря точности.

Для борьбы со «взрывом градиентов», созданным сверхбольшими значениями важностей весов, было предложено использовать модифицированную функцию потерь вида

$$L(\theta) \approx L_T(\theta) + \frac{\lambda}{2}\sum_i \frac{\Omega_i}{\alpha\lambda\Omega_i + 1}\left(\theta_i - \theta_i^*\right)^2,$$

где L_T — функция потерь для текущей задачи T, Ω_i — накопленная на предыдущих задачах важность i-го веса, α — скорость обучения нейронной сети. Такая модификация позволяет аппроксимирующему гессиан члену давать вклад в шаг оптимизации весов, не превышающий удаления значения веса от закрепленного значения θ_i^*.

Чтобы предотвратить потерю точности, то есть ситуацию, когда аппроксимирующий гессиан член дает вклад в градиент функции потерь по норме на несколько порядков больший, чем вклад функции потерь для текущей задачи $L_T(\theta)$, в [Куталев, Лапина, 2021] было предложено

проводить обрезку градиента по норме (gradient clipping) отдельно для градиента функции потерь текущей задачи и градиента члена, аппроксимирующего функцию потерь на предыдущих задачах, и затем суммировать результаты в один вектор-градиент для использования оптимизирующим алгоритмом.

Неточности и открытые вопросы метода

В заключение необходимо упомянуть о неточностях и ограничениях метода EWC.

Во-первых, не приводится никаких теоретических обоснований, что подпространство решений задачи в пространстве весов связно и непрерывно. Поэтому не очевидно, что, продвигаясь внутри области решений задачи A, можно достичь области решений задачи B.

Во-вторых, известно (см. [McCloskey, Cohen, 1989]), что для нейронных сетей в малой окрестности решения θ^* задачи A могут найтись точки θ', для которых функция потерь $L_A(\theta')$ очень значительно отличается от оптимального значения $L_A(\theta^*)$, что ставит под вопрос корректность ее аппроксимации с помощью квадратичной формы.

В-третьих, вышеприведенное обоснование EWC справедливо для нейронных сетей-классификаторов, то есть сетей с функцией потерь вида $L_D(\theta) = -\log p(D|\theta)$, что ограничивает область применения метода.

В-четвертых, в обосновании используется независимость наборов данных у последовательных задач A и B, а это чаще всего не так.

И наконец, в-пятых, EWC никак не ограничивает изменение весов на первом оптимизационном шаге обучения следующей задаче, поскольку первый шаг делается из точки θ^*, а в этой точке и член, аппроксимирующий гессиан, и его градиент равны нулю. Правда, последующие шаги в большинстве случаев исправляют ситуацию, то есть возвращают важные веса ближе к их закрепленным значениям пропорционально важности.

Однако, несмотря на все перечисленные неточности, метод EWC позволяет вполне успешно бороться с катастрофической забывчивостью на практике.

Метод замедления скоростей весов (WVA)

В качестве упрощающей альтернативы EWC в работе [Куталев, 2020] был предложен метод замедления скоростей весов (Weight Velocity Attenuation, WVA), упрощение в котором относительно EWC заключается в том, что не требуется сохранять веса нейросети θ^* после каждой или только после последней задачи, требуется лишь накапливать и хранить важности весов.

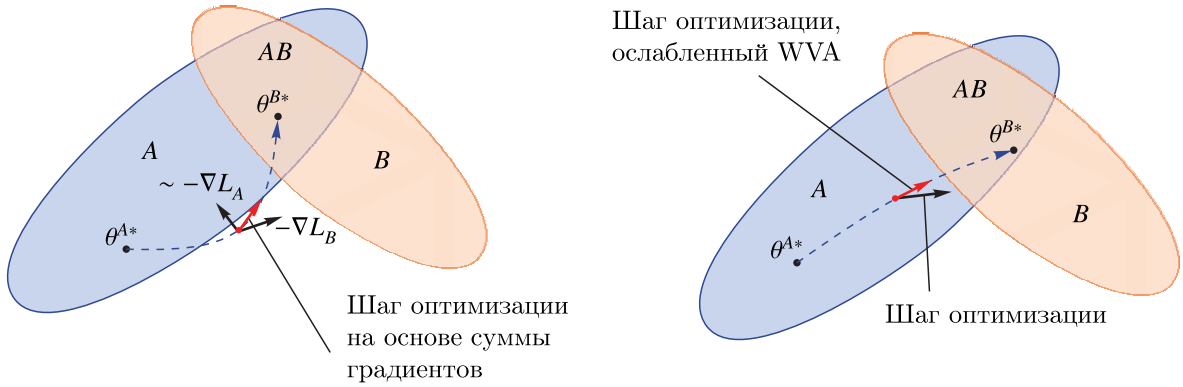
Суть метода состоит в том чтобы вместо удерживания весов притянутыми к точкам закрепления θ_i^* ослаблять шаг оптимизации каждого параметра нейронной сети пропорционально важности этого параметра. Например, это можно сделать с помощью гиперболической функции:

$$\theta_i^{t+1} = \theta_i^t + \frac{\Delta\theta_i^t}{\lambda\Omega_i + 1}, \tag{7}$$

где $\Delta\theta_i^t$ — шаг оптимизации, сформированный градиентным методом; Ω_i — накопленная важность веса i после нескольких предыдущих задач, изученных нейросетью; λ — общий для всей нейросети коэффициент, определяющий, насколько сеть стремится сохранить выученные навыки.

Приведем здесь обоснование, почему WVA в принципе работает и позволяет сохранять навыки при последовательном обучении. Для этого рассмотрим функцию потерь, применяемую в методе EWC при обучении задаче T после обучения нескольким предыдущим задачам:

$$L(\theta) = L_T(\theta) + \frac{\lambda}{2}\sum_i \Omega_i\left(\theta_i - \theta_i^*\right)^2. \tag{8}$$



а) EWC б) WVA

Рис. 2. Схематичное изображение принципов работы методов EWC и WVA: а) при обучении задаче B из точки θ^{A*} методом EWC оптимизационный шаг складывается из антиградиента функции потерь задачи B и EWC-компоненты, эмулирующей антиградиент функции потерь задачи A; б) при обучении задаче B из точки θ^{A*} методом WVA оптимизационный шаг складывается из антиградиента функции потерь задачи B, ослабленного пропорционально важностям весов для задачи A. Оба метода приводят веса нейронной сети в область AB, являющуюся решением обеих задач — A и B

Здесь $L_T(\theta)$ — функция потерь на датасете задачи T, Ω_i — накопленная важность i-го веса после обучения на нескольких предыдущих задачах, λ — коэффициент баланса между сохранением навыков, выученных на предыдущих задачах, и обучением текущей задаче T.

Как известно, квадратичному регуляризатору $\sum_i \Omega_i \left(\theta_i - \theta_i^*\right)^2$ в формуле (8) с вероятностной точки зрения соответствует нормальное распределение $p(\theta|D)$ (здесь D — объединение всех датасетов задач, уже изученных сетью) параметров θ_i со средними в θ_i^* и стандартными отклонениями $\sigma_i = \sqrt{\frac{1}{2\Omega_i}}$. Таким образом, для сохранения выученных на D навыков необходимо, чтобы каждый параметр θ_i оставался внутри некоторого определенного для него (доверительного) интервала. Учитывая это, видим, что формула (7) как раз и выполняет эту функцию — сильнее замедляет изменения параметров с малой дисперсией (то есть с большой важностью Ω_i) и, наоборот, позволяет сильнее изменяться параметрам с большой дисперсией.

Заметим, что, в отличие от EWC, WVA позволяет уводить параметры от их предыдущих значений неограниченно далеко за границы доверительного интервала, если проводить обучение текущей задаче достаточное число эпох. Поэтому WVA может быть полезен в задачах, где не обязательно точно сохранять навык, но полезно сохранить внутренние репрезентации, выученные сетью на предыдущих задачах.

В процессе разработки метода WVA необходимо было ответить на несколько вопросов, которые будут описаны более подробно далее. Для их экспериментальной проверки использовалась трехслойная нейронная сеть с двумя полносвязными слоями на 300 и 150 перцептронов с активацией leakyReLU и выходным softmax-слоем на 10 перцептронов. В каждом эксперименте обучение производилось последовательно на 10 датасетах, полученных из датасета MNIST случайным перемешиванием входов (permuted MNIST, см. [Goodfellow et al., 2015; Srivastava et al., 2013]). Обучение на каждом из датасетов проводилось в течение 4 эпох с размером минибатча 100. Скорость обучения была выбрана 0,001 для Adam и 0,2 для SGD. Важности весов рассчитывались методом суммарного по модулю прошедшего сигнала (см. [Куталев, 2020]). Также были проведены аналогичные эксперименты с важностями весов, рассчитанными на базе информационной матрицы Фишера (см. [Kirkpatrick et al., 2017]), и получена аналогичная кар-

тина. На основании этого ниже приводятся только результаты экспериментов с важностями на основе суммарного по модулю прошедшего сигнала.

К чему применять ослабление — к градиенту функции потерь или к оптимизационному шагу?

В случае использования в качестве метода обучения нейронной сети простого стохастического градиентного спуска (SGD) в качестве оптимизационного шага используется умноженный на константу градиент функции потерь. Поэтому для SGD нет разницы, к чему применять ослабление.

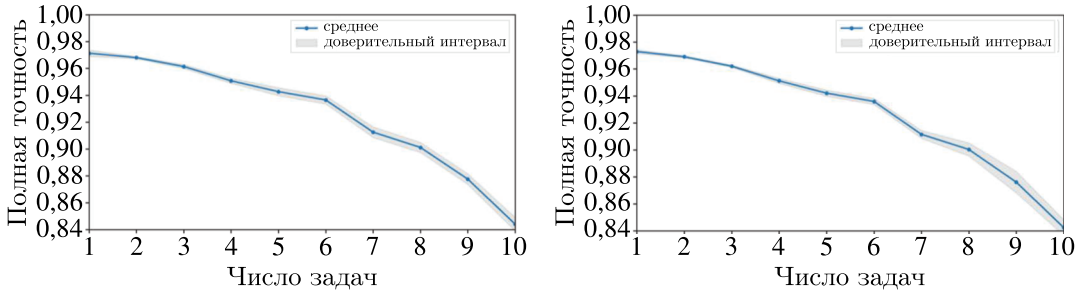


Рис. 3. Графики достижимой точности после обучения на 10 датасетах оптимизатором SGD: слева WVA-ослабление применяется к оптимизационному шагу, справа — к градиенту функции потерь. Графики абсолютно идентичны

В случае использования оптимизатора с более сложной логикой (например, SGD с моментами, AdaGrad, AdaMax, RMSprop и др.) из интуитивных соображений понятно, что ослабление нужно применять к оптимизационному шагу, созданному оптимизатором из градиента, поскольку оптимизатор может изменять пропорциональность шага градиенту, следуя своей логике, и, например, сильно увеличить шаг для веса с большой важностью. Таким образом, при применении WVA-ослабления к градиенту оптимизационный шаг на важных весах может снова стать большим. При применении же WVA-ослабления к готовому оптимизационному шагу такого эффекта наблюдаться не будет.

Для подтверждения этих рассуждений был проведен соответствующий эксперимент. Оптимальный (обеспечивающий наибольшую точность на всех выученных датасетах) гиперпараметр λ был получен для каждого из случаев перебором по сетке. Результаты эксперимента можно увидеть на рис. 4. Очевидно, насколько сильнее деградировала точность на всех выученных датасетах при применении ослабления к градиенту функции потерь.

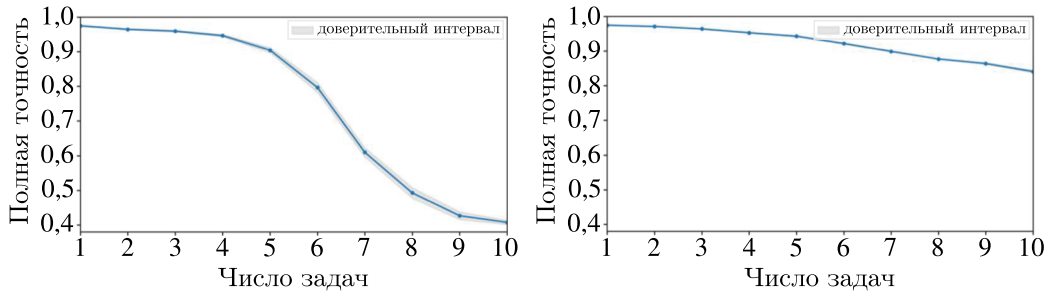


Рис. 4. Графики достижимой точности после обучения на 10 датасетах оптимизатором Adam: слева WVA-ослабление применяется к градиенту функции потерь, справа — к оптимизационному шагу

Какова должна быть функция ослабления? Сравнение гиперболического и экспоненциального ослабления

Ослабление оптимизационного шага в WVA можно реализовать различными способами. Далее будут рассмотрены способы, используемые в обучении с подкреплением для дисконтирования вознаграждения (см. [Ainslie, Haslam, 1992; Green, Myerson, 1996; Fedus et al., 2019]). Ослабление с помощью гиперболической функции уже было представлено в формуле (7), а экспоненциальное ослабление реализуется следующей формулой:

$$\theta_i^{t+1} = \theta_i^t + e^{-\lambda\Omega_i}\Delta\theta_i^t. \tag{9}$$

Аналогично предыдущему были определены оптимальные λ перебором по сетке (методологию см. в [Куталев, Лапина, 2021]) для WVA-ослабления гиперболической функцией и экспонентой и с использованием SGD- и Adam-оптимизаторов. Результаты можно увидеть на рис. 5 и 6.

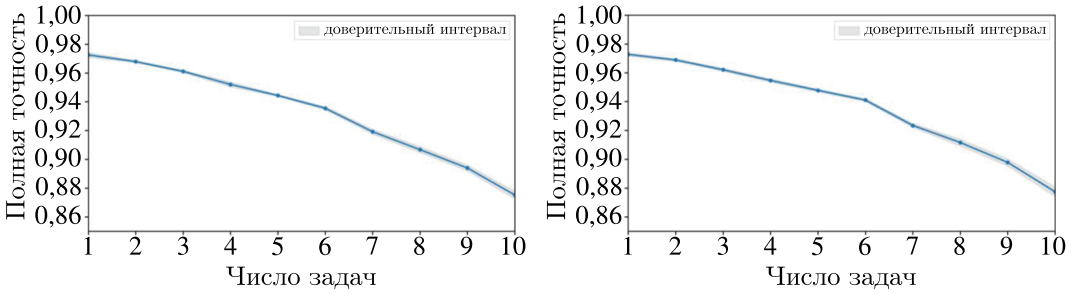


Рис. 5. Графики достижимой точности после обучения на 10 датасетах оптимизатором SGD: слева WVA-ослабление производится гиперболической функцией, справа — экспонентой

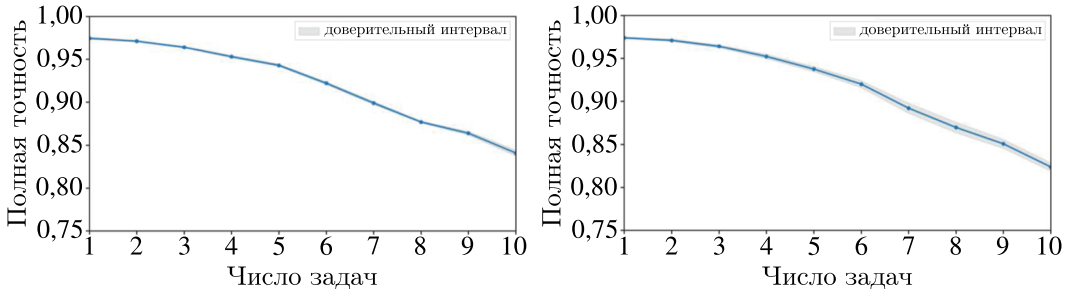


Рис. 6. Графики достижимой точности после обучения на 10 датасетах оптимизатором Adam: слева WVA-ослабление производится гиперболической функцией, справа — экспонентой

В случае использования SGD экспоненциальное ослабление оказалось лучше гиперболы, в случае Adam — хуже. Видно, что разница между видами ослабления несущественна при оптимальном λ. Таким образом, однозначных выводов об преимуществах какой-либо из этих функций сделать нельзя.

Как изменяется оптимальное значение гиперпараметра λ в зависимости от количества задач в последовательном обучении?

Ранее, при проведении сравнения, был применен поиск оптимального гиперпараметра λ для каждой конфигурации. Приведем здесь графики поверхностей зависимости точности от λ и количества датасетов в последовательном обучении, полученных в результате этого поиска.

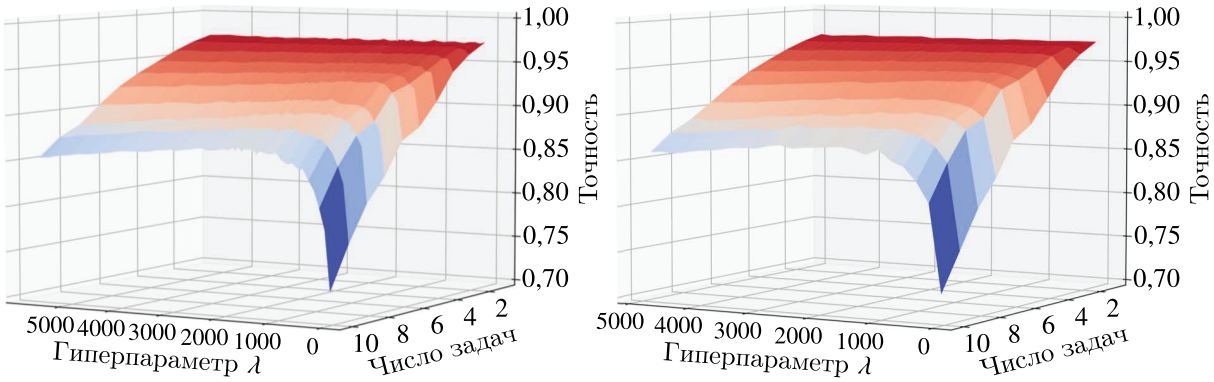


Рис. 7. Поверхности средней точности в зависимости от λ и количества изученных датасетов при обучении SGD: слева WVA-ослабление применяется к градиенту функции потерь, справа — к оптимизационному шагу

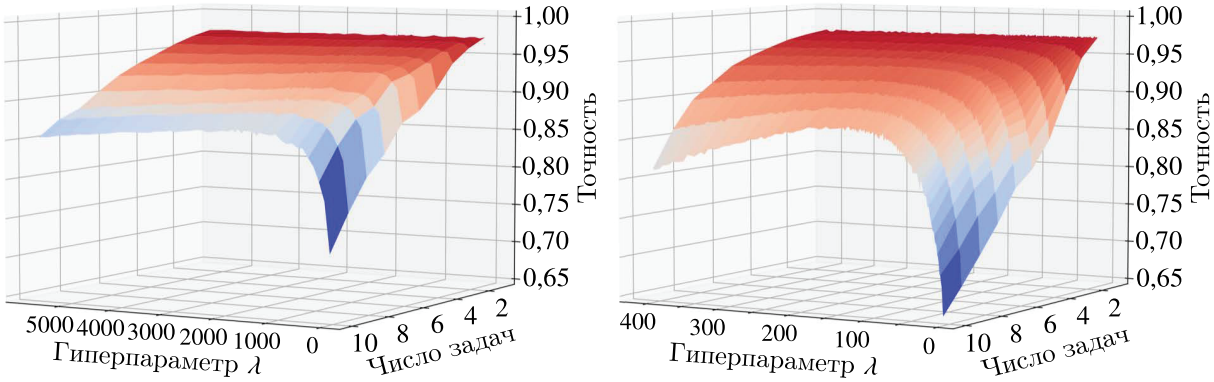


Рис. 8. Поверхности средней точности в зависимости от λ и количества изученных датасетов при обучении SGD: слева WVA-ослабление производится гиперболической функцией, справа — экспонентой

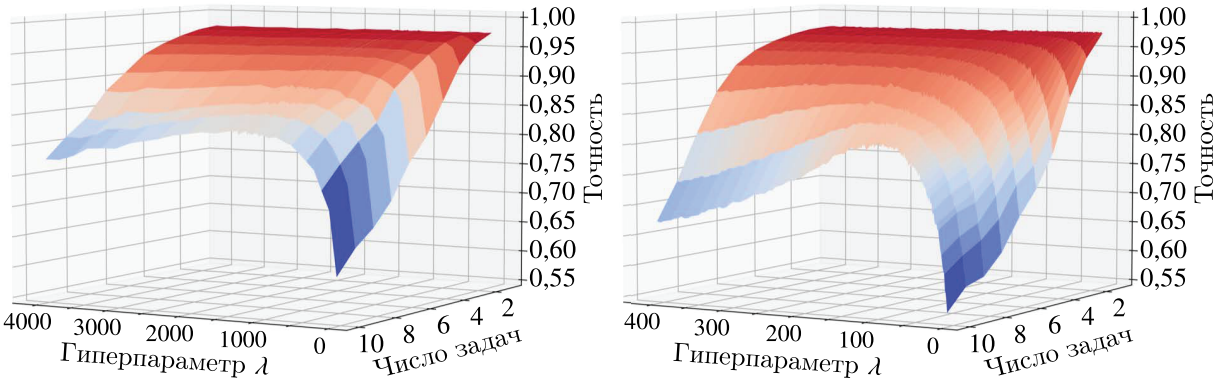


Рис. 9. Поверхности средней точности в зависимости от λ и количества изученных датасетов при обучении Adam: слева WVA-ослабление производится гиперболической функцией, справа — экспонентой

На этих поверхностях, изображенных на рис. 7, 8, 9, можно заметить, что оптимальное (с максимальной точностью) значение λ не зависит от количества изученных датасетов. Таким образом, для экономии ресурсов можно осуществлять подбор оптимального λ на небольшом количестве последовательных датасетов и ожидать, что это же значение λ будет оптимальным при обучении на более длинной последовательности, если датасеты имеют схожую структуру.

На графиках можно также заметить, что при удалении гиперпараметра λ от оптимального значения точность при экспоненциальном WVA-ослаблении деградирует сильнее, чем при использовании гиперболического ослабления.

Заключение

Цель этой работы — поделиться результатами экспериментов, которые проводились в рамках создания метода ослабления скоростей весов WVA. Не претендуя на общность и полную математическую строгость, можно сформулировать некоторые выводы и предположения об использовании WVA.

Например, было подтверждено, что WVA-ослабление должно применяться к оптимизационному шагу, а не к градиенту функции потерь. Также было экспериментально определено, что экспоненциальная функция WVA-ослабления дает примерно такое же качество, как и гиперболическая на оптимальном значении гиперпараметра λ. Но при удалении от оптимального λ экспонента показывает себя хуже, поэтому наилучшим выбором будет гиперболическое WVA-ослабление.

На основании наблюдений в процессе подбора гиперпараметра λ, отвечающего за баланс между сохранением навыков и текущим обучением, можно сделать предположение, что оптимальное значение λ не зависит от количества задач в последовательном обучении и, таким образом, может быть подобрано на короткой последовательности, а практически использоваться — на более длинной.

Список литературы (References)

Куталев А. А. Естественный способ преодоления катастрофической забывчивости нейронных сетей // Современные информационные технологии и ИТ-образование. — 2020. — Т. 16, № 2. — С. 331–337.

Kutalev A. A. Estestvennyi sposob preodoleniya katastroficheskoi zabyvchivosti neironnykh setei [Natural way to overcome catastrophic forgetting in neural networks] // Modern Information Technologies and IT-Education. — 2020. — Vol. 16, No. 2. — P. 331–337. — DOI: 10.25559/SITITO.16.202002.331-337 (in Russian).

Куталев А. А., Лапина А. А. Особенности использования метода эластичного закрепления весов в прикладных задачах машинного обучения // Современные информационные технологии и ИТ-образование. — 2021. — Т. 17, № 2. — С. 345–354.

Kutalev A. A., Lapina A. A. Osobennosti ispol'zovaniya metoda elastichnogo zakrepleniya vesov v prikladnykh zadachakh mashinnogo obucheniya [Stabilizing Elastic Weight Consolidation method in practical ML tasks and using weight importances for neural network pruning] // Modern Information Technologies and IT-Education. — 2021. — Vol. 17, No. 2. — P. 345–354. — DOI: 10.25559/SITITO.17.202102.345-354 (in Russian).

Ainslie G., Haslam N. Hyperbolic Discounting // Choice Over Time / G. Loewenstein, J. Elster (eds.). — New York: Russell Sage Foundation, 1992.

Aljundi R. et al. Memory aware synapses: Learning what (not) to forget // Proceedings of the European Conference on Computer Vision (ECCV). — 2018. — P. 139–154.

Fedus W. et al. Hyperbolic discounting and learning over multiple horizons // arXiv preprint arXiv:1902.06865. — 2019.

French R. M. Catastrophic forgetting in connectionist networks // Trends in cognitive sciences. — 1999. — Vol. 3, No. 4. — P. 128–135.

Goodfellow I. J. et al. An empirical investigation of catastrophic forgetting in gradient-based neural networks // arXiv preprint arXiv:1312.6211. — 2013.

Green L., Myerson J. Exponential versus hyperbolic discounting of delayed outcomes: Risk and waiting time // American Zoologist. — 1996. — Vol. 36, No. 4. — P. 496–505.

Gupta S. et al. Addressing catastrophic forgetting for medical domain expansion // arXiv preprint arXiv:2103.13511. — 2021.

Huszár F. Note on the quadratic penalties in elastic weight consolidation // Proceedings of the National Academy of Sciences. — 2018. — Vol. 115, No. 11. — P. E2496–E2497.

Kirkpatrick J. et al. Overcoming catastrophic forgetting in neural networks // Proceedings of the national academy of sciences. — 2017. — Vol. 114, No. 13. — P. 3521–3526.

Kirkpatrick J. et al. Reply to Huszár: The elastic weight consolidation penalty is empirically valid // Proceedings of the National Academy of Sciences. — 2018. — Vol. 115, No. 11. — P. E2498–E2498.

MacKay D. J. C. A practical Bayesian framework for backpropagation networks // Neural computation. — 1992. — Vol. 4, No. 3. — P. 448–472.

Madasu A., Vijjini A. R. Sequential domain adaptation through elastic weight consolidation for sentiment analysis // 2020 25th International Conference on Pattern Recognition (ICPR). — IEEE, 2021. — P. 4879–4886.

McClelland J. L., McNaughton B. L., O'Reilly R. C. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory // Psychological review. — 1995. — Vol. 102, No. 3. — P. 419.

McCloskey M., Cohen N. J. Catastrophic interference in connectionist networks: The sequential learning problem // Psychology of learning and motivation. — Academic Press, 1989. — Vol. 24. — P. 109–165.

Miconi T., Stanley K., Clune J. Differentiable plasticity: training plastic neural networks with backpropagation // International Conference on Machine Learning. — PMLR, 2018. — P. 3559–3568.

Pascanu R., Bengio Y. Revisiting natural gradient for deep networks // arXiv preprint arXiv:1301.3584. — 2013.

Schwarz J. et al. Progress & compress: A scalable framework for continual learning // International Conference on Machine Learning. — PMLR, 2018. — P. 4528–4537.

Srivastava R. K. et al. Compete to compute // Advances in neural information processing systems. — 2013. — Vol. 26. — P. 2310–2318.

Thangarasa V., Miconi T., Taylor G. W. Enabling continual learning with differentiable hebbian plasticity // 2020 International Joint Conference on Neural Networks (IJCNN). — IEEE, 2020. — P. 1–8.

van Garderen K. et al. Towards continuous learning for glioma segmentation with elastic weight consolidation // arXiv preprint arXiv:1909.11479. — 2019.

Zenke F., Gerstner W., Ganguli S. The temporal paradox of Hebbian learning and homeostatic plasticity // Current opinion in neurobiology. — 2017. — Vol. 43. — P. 166–176.

Zenke F., Poole B., Ganguli S. Continual learning through synaptic intelligence // International Conference on Machine Learning. — PMLR, 2017. — P. 3987–3995.