УДК 004.85
DOI: 10.25559/SITITO.17.202102.345-354

Оригинальная статья

Особенности использования метода эластичного закрепления весов в прикладных задачах машинного обучения

А. А. Куталев*, **А. А. Лапина**
ПАО «Сбербанк России», г. Москва, Российская Федерация
117997, Российская Федерация, г. Москва, ул. Вавилова, д. 19
* kutalev@yandex.ru

Аннотация

Данное исследование посвящено особенностям практического применения метода эластичного закрепления весов (Elastic Weight Consolidation, EWC) при последовательном обучении нейронных сетей на нескольких обучающих наборах. В ней мы более строго сравним известные методологии расчета важностей весов, применяемые в методе закрепления весов. Это методологии Memory Aware Synapses (MAS), Synaptic Intelligence (SI) и расчет важностей весов на основе информационной матрицы Фишера из оригинальной работы по EWC. Мы рассмотрим эти методологии в применении к глубоким нейронным сетям с полносвязными и сверточными слоями, найдем оптимальные гиперпараметры для каждой из методологий и сравним результаты последовательного обучения нейронной сети при их использовании. Далее, мы укажем на проблемы, возникающие при применении метода эластичного закрепления весов в глубоких нейронных сетях со сверточными слоями и слоями с самовниманием, такие как «взрыв градиентов» и потеря значащей информации в градиенте при использовании ограничения его нормы. Затем, мы предложим способ стабилизации метода эластичного закрепления весов, помогающий решить эти проблемы, проведем оценку этого способа в сравнении с оригинальной методологией и покажем, что предложенный способ стабилизации справляется с задачей сохранения навыков при последовательном обучении не хуже, чем оригинальный EWC, но, при этом, не обладает его недостатками. В заключение приведем интересный факт об использовании различных типов важностей весов в задаче прунинга нейронной сети.

Ключевые слова: нейронные сети, катастрофическая забывчивость, эластичное закрепление весов, алгоритм обратного распространения, стрижка весов, прунинг нейронной сети

Авторы заявляют об отсутствии конфликта интересов.

Для цитирования: Куталев, А. А. Особенности использования метода эластичного закрепления весов в прикладных задачах машинного обучения / А. А. Куталев, А. А. Лапина. – DOI 10.25559/SITITO.17.202102.345-354 // Современные информационные технологии и ИТ-образование. – 2021. – Т. 17, № 2. – С. 345-354.

Original article

Stabilizing Elastic Weight Consolidation Method in Practical ML Tasks and Using Weight Importance's for Neural Network Pruning

A. A. Kutalev[*], **A. A. Lapina**
PJSC "Sberbank of Russia", Moscow, Russian Federation
19 Vavilov St., Moscow 117997, Russian Federation
[*] kutalev@yandex.ru

Abstract

This work focuses on the practical application of Elastic Weight Consolidation (EWC) for sequential training of neural networks on several training sets. In it, we will more rigorously compare the well-known methodologies for calculating the importance of weights used in the method of fixing weights. These are the Memory Aware Synapses (MAS), Synaptic Intelligence (SI) methodologies and the calculation of the importance of weights based on the Fisher information matrix from the original work on EWC. We will review these methodologies in the application to deep neural networks with fully connected and matched layers, find optimal hyperparameters for each of the methodologies, and compare the results of sequential learning of the neural network when using them. Next, we will point out the problems that arise when applying the method of elastic weight pinning in deep neural networks with convolutional layers and self-attention layers, such as the "explosion of gradients" and the loss of significant information in the gradient when using its norm constraint (gradient clipping). Then, we will propose a method for stabilizing the elastic weight fixing method that helps to solve these problems, evaluate this method in comparison with the original methodology, and show that the proposed stabilization method copes with the task of retaining skills in sequential training no worse than the original EWC, but, at the same time, does not have its disadvantages. In conclusion, it is interesting to note the use of different types of weights in the neural network's pruning problem.

Keywords: neural network, catastrophic forgetting, elastic weight consolidation, back propagation, neural network pruning

The authors declare no conflict of interest.

For citation: Kutalev A.A., Lapina A.A. Stabilizing Elastic Weight Consolidation Method in Practical ML Tasks and Using Weight Importance's for Neural Network Pruning. *Sovremennye informacionnye tehnologii i IT-obrazovanie = Modern Information Technologies and IT-Education.* 2021; 17(2):345-354. DOI: https://doi.org/10.25559/SITITO.17.202102.345-354

Введение

Теме катастрофической забывчивости нейронных сетей посвящено множество научных работ, из которых можно почерпнуть детальное описание проблематики и попыток ее преодоления. Наиболее яркие примеры таких работ от достаточно далекого прошлого до наших дней можно найти в статьях [1]-[4].

В 2017 году был открыт метод эластичного закрепления весов (Elastic Weight Consolidation, EWC) нейронных сетей [5], который показал выдающиеся результаты по преодолению катастрофической забывчивости на нескольких классических задачах машинного обучения при последовательном обучении. Суть метода заключается в вычислении важности каждого веса (параметра) нейронной сети относительно задач, которым нейронная сеть уже обучилась, и запоминании значений весов после обучения этим задачам. При обучении последующим задачам метод не дает каждому весу (параметру) нейронной сети удаляться существенно от запомненных значений с силой, пропорциональной важности веса.

В дальнейшем появилось несколько работ, продолжающих уточнять [6] и развивать [7], [8], [9], [10] предложенную в [5] методологию.

Есть также примеры применения метода эластичного закрепления весов в прикладных задачах [11], [12], [13] и его сравнительных оценок для различных архитектур нейронных сетей [12].

В этой работе мы попробуем более детально исследовать применение метода EWC к некоторым практическим задачам машинного обучения, обсудим возникающие проблемы и рассмотрим пути их решения.

Подбор оптимального гиперпараметра EWC для конкретных задач

Сохранение ранее изученных знаний при последовательном обучении реализуется в методе эластичного закрепления весов путем добавления в функцию потерь регуляризатора, мешающего наиболее важным весам далеко отклоняться от закрепленных значений во время обучения последующим задачам:

$$L = L_A + \frac{\lambda}{2}\sum_i \Omega_i (w_i - w_i^*)^2,$$

где L_A есть функция потерь при обучении задаче **A**, w_i^* – i-й вес (параметр) нейронной сети после обучения предыдущим задачам, Ω_i – важность i-го веса нейронной сети после обучения предыдущим задачам. Из формулы видно, что вклад регуляризирующего компонента в антиградиент будет иметь вид $-\lambda\Omega_i(w_i - w_i^*)$, и, таким образом, при использовании для обучения градиентных методов сопротивление изменению каждого веса будет пропорционально его важности и гиперпараметру λ.

Тогда при малом λ сопротивление изменению весов в целом будет мало, и при последовательном обучении нейронная сеть будет лучше учиться текущей задаче, но при этом быстрее забывать навыки от предыдущих выученных задач. И наоборот, если лямбда слишком велика, то и сопротивление изменению весов будет велико, и сеть будет хорошо удерживать предыдущие навыки, выученные на прошлых задачах, но при этом скорость обучения текущей задаче может быть недостаточной. В качестве метрики качества процесса последовательного обучения можно рассматривать среднюю точность (*accuracy*) на всех изученных задачах.

Из этих рассуждений следует, что для конкретной архитектуры нейросети и конкретного набора датасетов в последовательном обучении существует оптимальное значение λ, при котором достигается максимум средней точности после последовательного обучения всем датасетам. И это оптимальное значение λ можно найти эмпирическим путем. Например, простым перебором по сетке.

Соответственно, для практического применения метода эластичного закрепления весов задача подбора оптимального гиперпараметра λ становится очень важной для получения максимальной выгоды от использования метода. В основных работах по методу EWC [5], [7], [8], [9] методология подбора λ либо упоминается вскользь, либо вообще не упоминается. Сопроводительный код статей так же не содержит полноценного механизма подбора λ. В этой работе мы хотим исправить это упущение и постараемся подойти к процессу подбора λ более точно и строго.

Далее, мы решили сравнить все методы вычисления важностей весов в одинаковых условиях и с оптимальным значением гиперпараметра λ, подобранным для каждого метода. Наши эксперименты проводились для глубокой нейронной сети с несколькими полносвязными слоями и для глубокой сверточной сети. Подбор оптимального λ производился перебором по сетке с проведением 20 экспериментов в каждой точке, вычислением средней точности и ее доверительного интервала. Емкость (количество параметров) нейронной сети была выбрана так, чтобы все задачи при последовательном обучении не помещались в сеть и происходило хотя бы частичное вытеснение навыков предыдущих задач на последующих задачах. Подробное описание эксперимента можно найти в приложении А.

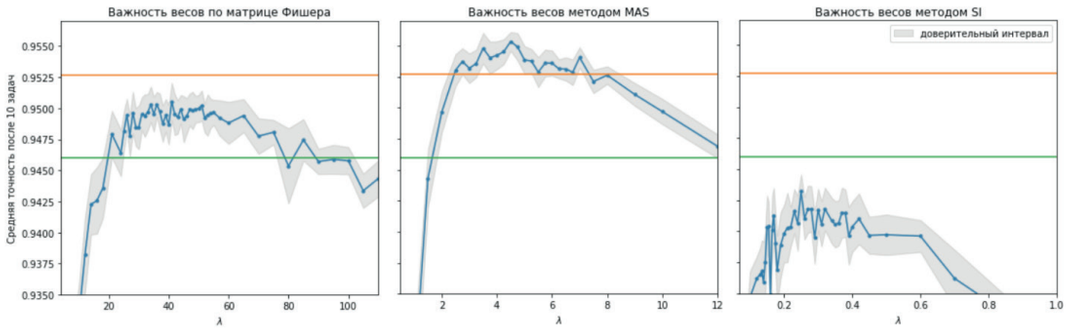
Результаты наших экспериментов приведены в таблице 1 и на рисунках 1 и 2.

Таблица 1. **Найденные оптимальные значения λ и доверительные интервалы достигнутой средней точности при этих λ**

Table 1. **The found optimal values of λ and the confidence intervals of the achieved average accuracy for these λ**

Метод расчета важности весов	Оптимальное значение λ	Средняя точность и доверительный интервал для этой λ
Сеть с полносвязными слоями:		
Fisher	41	0.9505 ± 0.0015
MAS	4.5	0.9553 ± 0.0008
SI	0.25	0.9432 ± 0.0014
Сеть с полносвязными слоями:		
Fisher	675	0.5846 ± 0.0144
MAS	300	0.6012 ± 0.0144
SI	24	0.5068 ± 0.0150

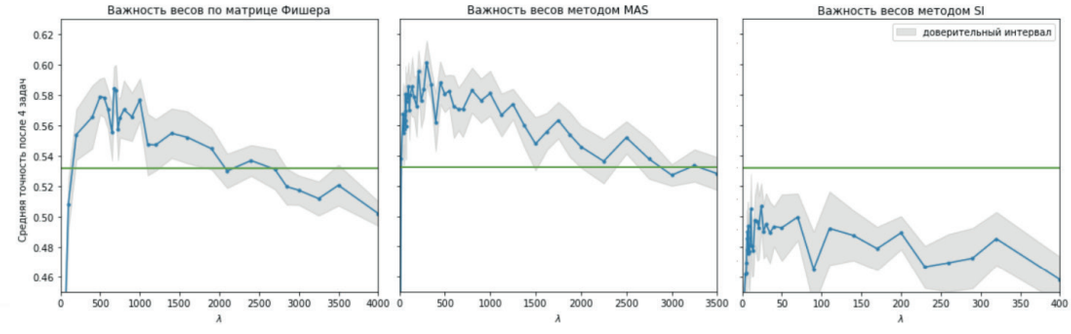
Как видим, для полносвязной нейронной сети доверительные интервалы средней точности на оптимальных значениях λ для каждого из методов строго отделены и не пересекаются. Таким образом, можно утверждать, что для выбранной нами архитектуры нейронной сети с полносвязными слоями и последовательности задач для обучения метод расчета важности Memory Aware Synapses (MAS) из работы [8] по результирующей оценке (средней точности) превосходит метод расчета важностей на основе диагональных элементов информационной матрицы Фишера из работы [5], который в свою очередь превосходит метод Synaptic Intelligence (SI) из [7].



Р и с. 1. Графики зависимости достижимой средней точности по всем изученным задачам от гиперпараметра λ при использовании методов расчета важностей на основе матрицы Фишера, MAS и SI для сети с полносвязными слоями

F i g. 1. Graphs of the dependence of the achievable average accuracy for all the studied problems on the hyperparameter λ using methods for calculating the importance based on the Fisher matrix, MAS and SI for a network with fully connected layers

Для нейронной сети со сверточными слоями мы видим, что, хотя средние значения точности на оптимальных значениях λ образуют такой же порядок (средняя точность с важностями по MAS больше средней точности с важностями по элементам матрицы Фишера, которая, в свою очередь, больше средней точности с важностями по SI), доверительные интервалы для средних значений точности для важности по MAS и по элементам матрицы Фишера пересекаются. Поэтому утверждать в строгом смысле о превосходстве метода расчета важностей по MAS над методом расчета важностей по элементам матрицы Фишера нельзя. Однако можно утверждать о превосходстве этих методов над методом SI, поскольку доверительные интервалы средней точности с использованием этих методов строго отделены от доверительного интервала для среднего значения точности для важности, рассчитанной методом SI.



Р и с. 2. Графики зависимости достижимой средней точности по всем изученным задачам от гиперпараметра λ при использовании методов расчета важностей на основе матрицы Фишера, MAS и SI для сети со сверточными слоями

F i g. 2. Graphs of the dependence of the achievable average accuracy for all studied problems on the hyperparameter λ using methods for calculating the importance based on the Fisher matrix, MAS and SI for a network with convolutional layers

Следует заметить, что между методом расчета важности из [5] (на основе диагональных элементов матрицы Фишера) и методами из [7], [8], [9] существует принципиальная разница. Метод из [5] (оригинальный EWC) использует выходные значения обучающего набора задачи, и, таким образом, пытается сохранить ту часть обобщений/репрезентаций, заложенных в обучающем наборе задачи, которую нейронная сеть смогла выучить. Методы из [7], [8], [9] не используют выходные значения обучающего набора, и, следовательно, пытаются сохранить те обобщения/репрезентации, которые нейронная сеть фактически выучила.

Метод из [5] также может быть использован для сохранения фактически выученных навыков. Для этого при расчете важности на каждом примере следует брать выход нейронной сети с максимальным значением, либо выбирать выход путем сэмплирования из выходного распределения. Такой подход исполь-

зуется в большинстве примеров реализации EWC, доступных в сети Интернет. Однако, при применении метода из [5] к большим NLP-моделям (например, BERT, GPT и т.д.), более предпочтительно выглядит использование способа с выбором выходов для расчета важностей в соответствии с выходными значениями из обучающего набора, поскольку в таких случаях обычно цель – сохранить с помощью EWC знание естественного языка, заключенное в предъявляемых при обучении текстах, а не те языковые репрезентации, которые модель выучила.

Проблемы метода эластичного закрепления весов

Применяя метод эластичного закрепления весов к сетям со сверточными слоями и сетям с вниманием (attention), мы обнаружили, что, в отличие от сетей с полносвязными слоями, рассчитанные важности весов в сверточном слое или слое с вниманием имеют очень большую дисперсию. То есть появляются такие веса (параметры) сети, важность которых на несколько порядков больше средней важности по слою. Эффект от таких важностей может оказать существенное влияние на эффективность метода. Для прояснения этого эффекта рассмотрим процесс более подробно.

Как мы упомянули ранее, регуляризирующий компонент функции потерь при применении EWC дает вклад в антиградиент вида $-\lambda\Omega_i(w_i - w_i^*)$. При использовании в качестве оптимизатора, например, метода стохастического градиентного спуска приращение веса на шаге обучения будет иметь вид:

$$\Delta w_i = -\alpha\frac{\partial L_A}{\partial w_i} - \alpha\lambda\Omega_i(w_i - w_i^*),$$

где α – скорость обучения, L_A - функция потерь для текущей задачи **A**, λ – гиперпараметр метода закрепления весов, Ω_i – важность *i*-го веса нейронной сети, w_i – текущее значение *i*-го веса, w_i^* – значение *i*-го веса после завершения обучения предыдущим задачам.

В случае, если важность *i*-го веса достаточно велика, то вклад от функции потерь $\alpha\frac{\partial L_A}{\partial w_i}$ в приращение будет пренебрежимо мал.

Тогда, если $\Omega_i \geq \frac{1}{\alpha\lambda}$, то после шага оптимизации вес w_i будет не сдвигаться в сторону w_i^*, а перескакивать его.

В случае же, если $\Omega_i \geq \frac{2}{\alpha\lambda}$, вес w_i будет не только перескакивать значение w_i^*, но и расстояние между w_i и w_i^* будет увеличиваться в $(\alpha\lambda\Omega_i - 1)$ раз. Таким образом, будет наблюдаться эффект, известный как «взрыв градиента». И если даже при таком «взрыве» процесс обучения не будет прерван по переполнению, то самые важные для предыдущих задач веса будут очень быстро удаляться от закрепленных значений.

Конечно, использование оптимизационных методов с моментами или иными компенсирующими механизмами способно предотвратить описываемый эффект, но даже в этом случае он может негативно влиять на сходимость метода оптимизации и на способность метода закрепления весов сохранять выученные ранее навыки.

Также, при обучении нейронных сетей часто используется ограничение градиента (gradient clipping). То есть, когда норма градиента превышает некий порог, градиент нормируется на величину этого порога. Тогда, при использовании метода эластичного закрепления весов и наличии весов со сверхбольшими важностями эти веса дадут столь большой вклад в градиент, что, после перенормирования градиента, вклад остальных весов в градиент станет исчезающе мал. Следовательно, использование полученного градиента для шага оптимизации будет служить в основном возвращению сверхважных весов к закрепленным значениям, а значимая для обучения информация будет в значительной степени теряться.

Для решения вышеописанных проблем мы использовали механизм стабилизации, предотвращающий появление в регуляризирующем вкладе в приращение веса величин, больших чем разность закрепленного и текущего значения веса. Для этого мы преобразовали функцию потерь к виду:

$$L = L_A + \frac{\lambda}{2}\sum_i \frac{\Omega_i}{\alpha\lambda\Omega_i + 1}(w_i - w_i^*)^2.$$

Для такой функции потерь вклад регуляризирующего компонента в антиградиент будет иметь вид $-\frac{\lambda\Omega_i}{\alpha\lambda\Omega_i+1}(w_i - w_i^*)$, а вклад в приращение $-\frac{\alpha\lambda\Omega_i}{\alpha\lambda\Omega_i+1}(w_i - w_i^*)$.

При малой (близкой к нулю) важности *i*-го веса Ω_i множитель $\frac{\alpha\lambda\Omega_i}{\alpha\lambda\Omega_i+1}$ ведет себя пропорционально $\alpha\lambda\Omega_i$, а при очень большой Ω_i стремится к 1. Таким образом вклад регуляризирующего компонента в приращение веса не будет превышать разности весов даже при сколь угодно большой важности веса.

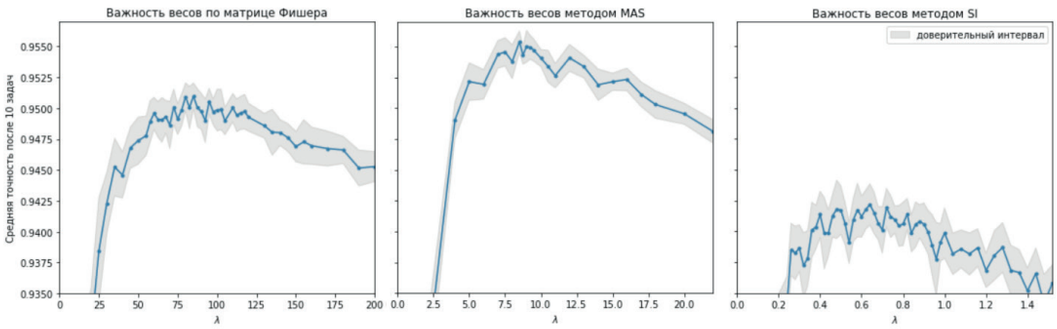
Для проверки работоспособности этого подхода мы провели эксперименты, аналогичные экспериментам из раздела 2, чтобы убедиться, что он, как и оригинальный метод эластичного закрепления весов, позволяет успешно сохранять навыки при последовательном обучении, а также позволяет получить не худшую чем у EWC среднюю точность на оптимальной λ.

В таблице 2 и на рисунках 3 и 4 приведены результаты экспериментов и их сравнение при оптимальных λ с результатами оригинального EWC.

Т а б л и ц а 2. **Сравнение оптимальных значений λ и доверительных интервалов результирующей средней точности при использовании оригинального и стабилизированного метода эластичного закрепления весов**

T a b l e 2. **Comparison of optimal values of λ and confidence intervals of the resulting average accuracy using the original and stabilized method of elastic attachment of the balance**

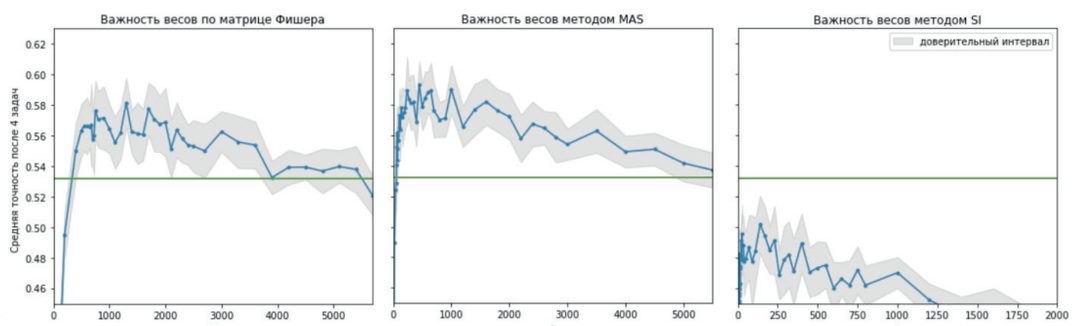
Метод закрепления весов	Тип важностей	Оптимальное значение λ	Доверительный интервал для этой λ
Сеть с полносвязными слоями:			
EWC	Fisher	41	0.9505 ± 0.0015
Stabilized EWC	Fisher	85	0.9510 ± 0.0011
EWC	MAS	4.5	0.9553 ± 0.0008
Stabilized EWC	MAS	8.5	0.9554 ± 0.0009
EWC	SI	0.25	0.9432 ± 0.0014
Stabilized EWC	SI	0.64	0.9422 ± 0.0017
Сеть с полносвязными слоями:			
EWC	Fisher	675	0.5846 ± 0.0144
Stabilized EWC	Fisher	1300	0.5872 ± 0.0161
EWC	MAS	300	0.6012 ± 0.0144
Stabilized EWC	MAS	450	0.5930 ± 0.0149
EWC	SI	24	0.5068 ± 0.0150
Stabilized EWC	SI	140	0.5106 ± 0.0188



Р и с. 3. Графики зависимости достижимой средней точности по всем изученным задачам от гиперпараметра λ при использовании методов расчета важностей на основе матрицы Фишера, MAS и SI для сети с полносвязными слоями

F i g. 3. Graphs of the dependence of the achievable average accuracy for all the studied problems on the hyperparameter λ using methods for calculating the importance based on the Fisher matrix, MAS and SI for a network with fully connected layers

Как мы видим, и для сетей с полносвязными слоями и для сверточных сетей при использовании метода эластичного закрепления весов с предложенным механизмом стабилизации значения средней точности на оптимальных гиперпараметрах λ почти всегда больше значений средней точности при обучении с использованием метода эластичного закрепления весов без стабилизации. Однако, доверительные интервалы для средних значений точности для случаев со стабилизацией и без нее всегда пересекаются, то есть нельзя с уверенностью говорить о превосходстве одного из подходов. Таким образом, предложенный механизм стабилизации по крайней мере не ухудшает метод эластичного закрепления весов, и при этом позволяет улучшить сходимость оптимизационных алгоритмов и сохранение ранее выученных навыков.



Р и с. 4. Графики зависимости достижимой средней точности по всем изученным задачам от гиперпараметра λ при использовании методов расчета важностей на основе матрицы Фишера, MAS и SI для сети со сверточными слоями

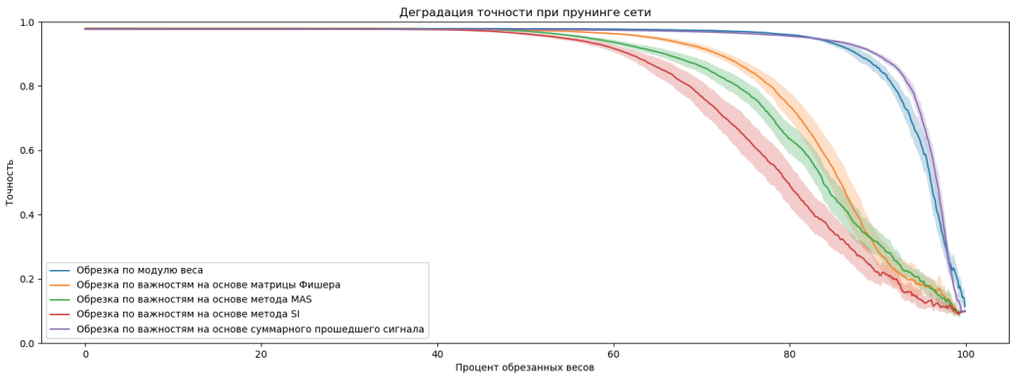
F i g. 4. Graphs of the dependence of the achievable average accuracy for all studied problems on the hyperparameter λ using methods for calculating the importance based on the Fisher matrix, MAS and SI for a network with convolutional layers

Мы также провели эксперименты по применению стабилизированного варианта метода эластичного закрепления весов к дообучению предобученной русской GPT2-модели на диалоговом обучающем наборе. В результате выигрыш от использования метода закрепления весов составил около 7 единиц перплексии по сравнению с обычным дообучением (перплексия модели рассчитывалась на наборе текстов из книг, на предобученной модели она была около 16 единиц, после обычного дообучения она достигла 26 единиц, а при дообучении с EWC составила 19 единиц).

Прунинг весов с использованием важностей метода эластичного закрепления весов

Поскольку важности весов в методе эластичного закрепления весов позиционируются как мера того, насколько важен каждый отдельный вес для сохранения выученных репрезентаций, то логично попробовать использовать значения этих важностей для задачи обрезки (прунинга) весов с целью снижения размера и вычислительной сложности нейронной сети ([16]-[21]).

Соответственно, мы провели эксперименты по неструктурированному прунингу в нейронной сети на основе EWC-важностей весов, вычисленных различными способами, и сравнили результаты с прунингом на основе модуля веса. Результаты представлены на рисунке 5.



Р и с. 5. Графики деградации средней точности в зависимости от количества обрезанных (обнуленных) весов нейронной сети. Указана средняя точность за 10 проходов вместе со своим доверительным интервалом

F i g. 5. Graphs of degradation of average accuracy depending on the number of trimmed (zeroed) weights of the neural network. Average accuracy over 10 runs is given along with its confidence interval

Как видно из графика, в задаче прунинга важности весов, вычисленные на основе диагональных элементов информационной матрицы Фишера (оригинальный EWC [5]), а также важности весов, вычисленные методами SI [7] и MAS [8], неожиданно и существенно проиграли важностям на основе суммарного абсолютного прошедшего сигнала [9] и важностям на основе самих весов по модулю.

В качестве объяснения этого факта мы можем предположить, что значения важностей, вычисленные методами MAS, SI и на основе матрицы Фишера, описывают важность веса лишь в небольшой окрестности значения веса. А обрезка веса обнулением может выводить значение веса за эту окрестность, так что вычисленное значение важности теряет смысл.

Заключение

Приведенная методология поиска гиперпараметра λ для метода эластичного закрепления весов позволяет найти λ, достаточно близкий к оптимальному для конкретной нейронной сети и задачи машинного обучения.

В результате ее применения мы выяснили, что для всех использованных в наших экспериментах архитектурах нейронных сетей выбор способа MAS для расчета важностей весов является оптимальным, так как способствует наилучшему сохранения навыков по средней точности при последовательном обучении. То есть MAS статистически значимо лучше SI для всех использованных архитектур, и статистически значимо лучше расчета важностей весов с помощью информационной матрицы Фишера для сети с полносвязными слоями.

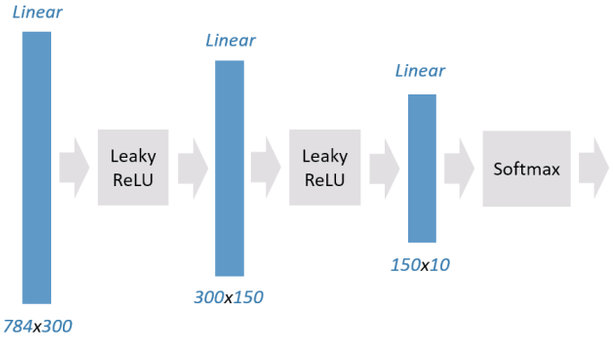
Предложенный способ стабилизации метода эластичного закрепления весов сохраняет его способность преодолевать катастрофическую забывчивость и показывает результаты не хуже (а по средней точности часто лучше) оригинального метода.

В задаче неструктурированного прунинга (обрезки связей путем обнуления весов) в нейронной сети из полносвязных слоев оптимальным является прунинг по модулю веса или по суммарному абсолютному сигналу, прошедшему через связь (если выбирать их перечисленных способов вычисления важностей). Оба этих способа существенно превосходят прунинг по важностям весов на основе методов MAS, SI и важностям на основе информационной матрицы Фишера.

Поскольку большинство современных языковых моделей используют многослойные нейронные сети на основе самовнимания (self-attention, transformers), мы видим перспективным использование стабилизированного метода эластичного закрепления весов при дообучении (fine-tuning) предобученных моделей на специализированных языковых датасетах.

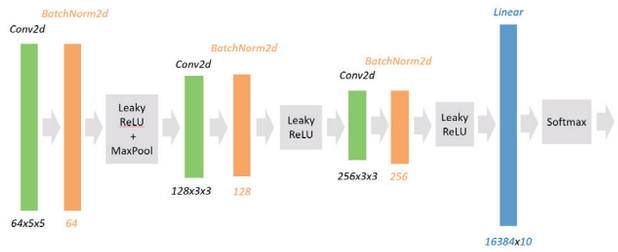
Приложение А

Архитектура нейронной сети с полносвязными слоями, использованной в наших экспериментах, отражена на рисунке 6, архитектура сети со сверточными слоями – на рисунке 7.



Р и с. 6. Архитектура нейронной сети с полносвязными слоями

F i g. 6. Neural network architecture with fully connected layers



Р и с. 7. Архитектура сверточной нейронной сети
F i g. 7. Convolutional neural network architecture

Для последовательного обучения сети со сверточными слоями были использованы обучающие наборы MNIST и Fasion-MNIST, а также их повороты на . Всего в последовательном обучении сверточной сети участвовало 4 обучающих набора.

В рамках одного эксперимента нейронная сеть обучалась методом Adam с параметрами , , , . До начала обучения важности всех весов сети инициализировались нулями. Обучение проводилось последовательно всем обучающим наборам (10 для сети с полносвязными слоями, 4 для сети со сверточными слоями). Тренировка каждому обучающему набору проводилось в течении 6 эпох с размером минибатча 100. После тренировки на обучающем наборе производился расчет важностей весов на этом наборе и эти важности суммировались с полученными ранее.

После обучения всем наборам в эксперименте проводился замер точности (*accuracy*) по всем отложенным (тестовым) частям изученных сетью обучающих наборов. Полученная точность использовалась в качестве результата эксперимента.

При каждом значении гиперпараметра эксперимент проводился двадцать раз[1]. Результат усреднялся и вычислялся доверительный интервал для среднего со статистической значимостью .

References

[1] McCloskey M., Cohen N.J. Catastrophic interference in connectionist networks: The sequential learning problem. In: Ed. by G. H. Bower. *The Psychology of Learning and Motivation*, vol. 24. Academic, New York; 1989. p. 109-165. (In Eng.) DOI: https://doi.org/10.1016/S0079-7421(08)60536-8

[2] 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; 102(3):419-457. (In Eng.) DOI: https://doi.org/10.1037/0033-295X.102.3.419

[3] French R.M. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Science*. 1999; 3(4):128-135. (In Eng.) DOI: https://doi.org/10.1016/S1364-6613(99)01294-2

[4] Goodfellow I.J., Mirza M., Xiao D., Courville A.C., Bengio Y. An Empirical Investigation of Catastrophic Forgetting in Gradient-Based Neural Networks. *Proceedings of International Conference on Learning Representations (ICLR'2014)*. Banff, Canada; 2014. Available at: https://arxiv.org/pdf/1312.6211.pdf (accessed 23.05.2021). (In Eng.)

[5] Kirkpatrick J., Pascanu R., Rabinowitz N., Veness J., Desjardins G., Rusu A.A., Milan K., Quan J., Ramalho T., Grabska-Barwinska A., Hassabis D., Clopath C., Kumaran D., Hadsell R. Overcoming catastrophic forgetting in neural networks. *Proceeding of the National Academy of Science*. 2017; 114(13):3521-3526. (In Eng.) DOI: https://doi.org/10.1073/pnas.1611835114

[6] Huszár F. Note on the quadratic penalties in elastic weight consolidation. *Proceeding of the National Academy of Science*. 2018; 115(11):2496-2497. (In Eng.) DOI: https://doi.org/10.1073/pnas.1717042115

[7] Zenke F., Poole B., Ganguli S. Continual Learning Through Synaptic Intelligence. *Proceedings of the 34th International Conference on Machine Learning (ICML'17)*, vol. 70. JMLR.org; 2017. p. 3987-3995. (In Eng.)

[8] Aljundi R., Babiloni F., Elhoseiny M., Rohrbach M., Tuytelaars T. Memory Aware Synapses: Learning What (not) to Forget. In: Ed. by V. Ferrari, M. Hebert, C. Sminchisescu, Y. Weiss. *Computer Vision – ECCV 2018. ECCV 2018. Lecture Notes in Computer Science*, vol. 11207. Springer, Cham; 2018. p. 144-161. (In Eng.) DOI: https://doi.org/10.1007/978-3-030-01219-9_9

[9] Kutalev A.A. Natural Way to Overcome Catastrophic Forgetting in Neural Networks. *Sovremennye informacionnye tehnologii i IT-obrazovanie* = Modern Information Technologies and IT-Education. 2020; 16(2):331-337. (In Russ., abstract in Eng.) DOI: https://doi.org/10.25559/SITITO.16.202002.331-337

[10] Thangarasa V., Miconi T., Taylor G.W. Enabling Continual Learning with Differentiable Hebbian Plasticity. *2020 International Joint Conference on Neural Networks (IJCNN).* IEEE Press, Glasgow, UK; 2020. p. 1-8. (In Eng.) DOI: https://doi.org/10.1109/IJCNN48605.2020.9206764

[11] van Garderen K., van der Voort S., Incekara F., Smits M., Klein S. Towards continuous learning for glioma segmentation with elastic weight consolidation. *International Conference on Medical Imaging with Deep Learning.* London, United Kingdom; 2019. Available at: https://openreview.net/forum?id=Hkx_ry0NcN (accessed 23.05.2021). (In Eng.)

[12] Madasu A., Vijjini A.R. Sequential Domain Adaptation through Elastic Weight Consolidation for Sentiment Analysis. *2020 25th International Conference on Pattern Recognition (ICPR).* IEEE Press, Milan, Italy; 2021. p. 4879-4886. (In Eng.) DOI: https://doi.org/10.1109/ICPR48806.2021.9412617

[13] Gupta S., Singh P., Chang K., et al. Addressing catastrophic forgetting for medical domain expansion. arXiv:2103.13511. 2021. Available at: https://arxiv.org/pdf/2103.13511.pdf (accessed 23.05.2021). (In Eng.)

[14] Miconi T., Stanley K.O., Clune J. Differentiable plasticity: training plastic neural networks with backpropagation. *Proceedings of the 35th International Conference on Machine Learning (ICML'2018)*, vol. 80. PMLR; 2018. p. 3559-3568. (In Eng.)

[1] Код, который использовался для проведения экспериментов, доступен по ссылке: https://github.com/aakutalev/ewc-features

[15] Zenke F., Gerstner W., Ganguli S. The temporal paradox of hebbian learning and homeostatic plasticity. *Current Opinion in Neurobiology*. 2017; 43:166-176. (In Eng.) DOI: https://doi.org/10.1016/j.conb.2017.03.015

[16] LeCun Y., Denker J., Solla S. Optimal Brain Damage. In: Ed. by D. Touretzky. *Advances in Neural Information Processing Systems*, vol. 2. Morgan-Kaufmann; 1989. p. 598-605. Available at: https://proceedings.neurips.cc/paper/1989/file/6c9882bbac1c7093bd25041881277658-Paper.pdf (accessed 23.05.2021). (In Eng.)

[17] Chechik G., Meilijson I., Ruppin E. Synaptic Pruning in Development: A Computational Account. *Neural Computation*. 1998; 10(7):1759-1777. (In Eng.) DOI: https://doi.org/10.1162/089976698300017124.

[18] Hassibi B., Stork D.G., Wolff G.J. Optimal Brain Surgeon and General Network Pruning. *IEEE International Conference on Neural Networks*. 1993; 1:293-299. (In Eng.) DOI: https://doi.org/10.1109/ICNN.1993.298572

[19] Sietsma J., Dow R.J.F. Neural net pruning-why and how. *IEEE 1988 International Conference on Neural Networks*, vol. 1. IEEE Press, San Diego, CA, USA; 1988. p. 325-333. (In Eng.) DOI: https://doi.org/10.1109/ICNN.1988.23864

[20] Mozer M.C., Smolensky P. Skeletonization: a technique for trimming the fat from a network via relevance assessment. In: Ed. by D. Touretzky. *Advances in Neural Network Information Processing Systems*, vol. 1. Morgan Kaufmann; 1989. p. 107-115. Available at: https://proceedings.neurips.cc/paper/1988/file/07e1cd7dca89a1678042477183b7ac3f-Paper.pdf (accessed 23.05.2021). (In Eng.)

[21] Blalock D., Ortiz J.J.G., Frankle J., Guttag J. What is the State of Neural Network Pruning? *Proceedings of the 3rd MLSys Conference*. Austin, TX, USA; 2020. Available at: https://arxiv.org/pdf/2003.03033.pdf (accessed 23.05.2021). (In Eng.)

[22] Zacarias A., Alexandre L.A. Overcoming Catastrophic Forgetting in Convolutional Neural Networks by Selective Network Augmentation. *Artificial Neural Networks in Pattern Recognition*. 2018. p. 102-112. (In Eng.) DOI: https://doi.org/10.1007/978-3-319-99978-4_8

[23] Li H., Barnaghi P., Enshaeifar S., Ganz F. Continual Learning Using Multi-view Task Conditional Neural Networks. *Journal of LaTEX Class Files*. 2015; 14(8):1-10. Available at: https://arxiv.org/pdf/2005.05080.pdf (accessed 23.05.2021). (In Eng.)

[24] Kumaran D., Hassabis D., McClelland J.L. What learning systems do intelligent agents need? complementary learning systems theory updated. *Trends in Cognitive Sciences*. 2016; 20(7):512-534. (In Eng.) DOI: https://doi.org/10.1016/j.tics.2016.05.004

[25] Li Z., Hoiem D. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2018; 40(12):2935-2947. (In Eng.) DOI: https://doi.org/10.1109/TPAMI.2017.2773081

[26] Parisi G.I., Kemker R., Part J.L., Kanan C., Wermter S. Continual lifelong learning with neural networks: A review. *Neural Networks*. 2019; 113:54-71. (In Eng.) DOI: https://doi.org/10.1016/j.neunet.2019.01.012

[27] Masse N.Y., Grant G.D., Freedman D.J. Alleviating catastrophic forgetting using context-dependent gating and synaptic stabilization. *Proceedings of the National Academy of Sciences*. 2018; 115(44):E10467-E10475. (In Eng.) DOI: https://doi.org/10.1073/pnas.1803839115

[28] Mirzadeh S.I., Farajtabar M., Ghasemzadeh H. Dropout as an Implicit Gating Mechanism for Continual Learning. *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW).* Seattle, WA, USA; 2020. p. 945-951. (In Eng.) DOI: https://doi.org/10.1109/CVPRW50498.2020.00124

[29] Soltoggio A., Stanley K.O., Risi S. Born to learn: The inspiration, progress, and future of evolved plastic artificial neural networks. *Neural Networks*. 2018; 108:48-67. (In Eng.) DOI: https://doi.org/10.1016/j.neunet.2018.07.013

[30] Lee K., Lee K., Shin J., Lee H. Overcoming Catastrophic Forgetting With Unlabeled Data in the Wild. *IEEE/CVF International Conference on Computer Vision (ICCV)*. Seoul, Korea (South); 2019. p. 312-321. (In Eng.) DOI: https://doi.org/10.1109/ICCV.2019.00040

[31] Rostami M., Kolouri S., Pilly P.K. Complementary Learning for Overcoming Catastrophic Forgetting Using Experience Replay. *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19).* Macao; 2019. p. 3339-3345. (In Eng.) DOI: https://doi.org/10.24963/ijcai.2019/463

[32] Schak M., Gepperth A. A Study on Catastrophic Forgetting in Deep LSTM Networks. In: Ed. by I. Tetko, V. Kůrková, P. Karpov, F. Theis. *Artificial Neural Networks and Machine Learning – ICANN 2019: Deep Learning. ICANN 2019. Lecture Notes in Computer Science*, vol. 11728. Springer, Cham; 2019. p. 714-728. (In Eng.) DOI: https://doi.org/10.1007/978-3-030-30484-3_56

[33] Ribeiro J., Melo F.S., Dias J. Multi-task Learning and Catastrophic Forgetting in Continual Reinforcement Learning. In: Ed. by D. Calvanese, L. Iocchi. *Proceedings of the 5th Global Conference on Artificial Intelligence (GCAI-2019). EPiC Series in Computing*. 2019; 65:163-175. (In Eng.) DOI: https://doi.org/10.29007/g7bg

Поступила 23.05.2021; одобрена после рецензирования 17.06.2021; принята к публикации 20.06.2021.
Submitted 23.05.2021; approved after reviewing 17.06.2021; accepted for publication 20.06.2021.

Об авторах:

Куталев Алексей Анатольевич, специалист, ведущий программист-исследователь, Управление экспериментальных систем машинного обучения, ПАО «Сбербанк России» (117997, Российская Федерация, г. Москва, ул. Вавилова, д. 19), **ORCID: https://orcid.org/0000-0003-2695-792X**, kutalev@yandex.ru

Лапина Алиса Алексеевна, магистр, программист-исследователь, Лаборатория нейронаук и поведения человека, ПАО «Сбербанк России» (117997, Российская Федерация, г. Москва, ул. Вавилова, д. 19), **ORCID: https://orcid.org/0000-0002-4918-2584**, ahm.alisa@gmail.com

Все авторы прочитали и одобрили окончательный вариант рукописи.

About the authors:

Alexey A. Kutalev, MSc in Mathematics, *Senior Software Developer* of the Division of Experimental Machine Learning Systems, PJSC "Sberbank of Russia" (19 Vavilov St., Moscow 117997, Russian Federation), **ORCID: https://orcid.org/0000-0003-2695-792X,** kutalev@yandex.ru

Alisa A. Lapina, MSc in Robotics, Software Developer of the Laboratory of Neuroscience and Human Behavior, PJSC "Sberbank of Russia" (19 Vavilov St., Moscow 117997, Russian Federation), **ORCID: https://orcid.org/0000-0002-4918-2584,** ahm.alisa@gmail.com

All authors have read and approved the final manuscript.