

doi: 10.17586/2226-1494-2025-25-2-273-285

УДК 004.93

Структурный анализ изображения и адаптивный резонанс в искусственных нейронных сетях (обзорная статья)

Вадим Ростиславович Луцив^{1✉}, Мария Анатольевна Михалькова²,
Валерия Олеговна Ячная³

^{1,3} Санкт-Петербургский государственный университет аэрокосмического приборостроения, Санкт-Петербург, 190000, Российская Федерация

^{2,3} Институт физиологии имени И.П. Павлова РАН, Санкт-Петербург, 199034, Российская Федерация

¹ vluciv@mail.ru✉, <https://orcid.org/0000-0003-1160-3758>

² mikhalkova_maria@infran.ru, <https://orcid.org/0000-0002-7186-5744>

³ yachnaya_valeria@infran.ru, <https://orcid.org/0000-0002-7239-2096>

Аннотация

Введение. Наблюдаемый мир иерархически структурирован. Механизм взаимодействия встречных информационных потоков в иерархически организованной системе был назван «адаптивным резонансом» и успешно промоделирован полвека назад в искусственной нейронной сети, анализирующей обонятельные стимулы, а затем — при распознавании изображений. В дальнейшем полезный потенциал совместного применения в нейронных сетях принципов структурного анализа и адаптивного резонанса долго не использовался. В последние годы эти принципы были применены в капсульных сетях и позволили превзойти результаты, достигнутые другими моделями нейронных сетей. Это показало необходимость систематизировать использованные пути их реализации. **Метод.** Выполнен анализ научно-технических работ, опубликованных за последние 50 лет по применению принципов структурного анализа и адаптивного резонанса при обработке изображений в искусственных нейронных сетях. **Основные результаты.** Сравнительный анализ подтвердил эффективность применения этих принципов при решении задач автоматической обработки изображений и выявил пути их наиболее эффективной реализации в искусственных нейронных сетях. **Обсуждение.** В связи с успехами, достигнутыми сверточными сетями при распознавании изображений, их разработчики не применяли связанные с особенностями организации наблюдаемого мира принципы структурного анализа и адаптивного резонанса. Однако использование данных принципов позволяет более эффективно решать задачи обработки изображений, и дальнейшие исследования в области искусственных нейронных сетей целесообразно проводить в этом направлении.

Ключевые слова

искусственные нейронные сети, структурный анализ, адаптивный резонанс, иерархическая организация, сверточные сети, капсульные сети, трансформеры

Ссылка для цитирования: Луцив В.Р., Михалькова М.А., Ячная В.О. Структурный анализ изображения и адаптивный резонанс в искусственных нейронных сетях (обзорная статья) // Научно-технический вестник информационных технологий, механики и оптики. 2025. Т. 25, № 2. С. 273–285. doi: 10.17586/2226-1494-2025-25-2-273-285

Image structural analysis and adaptive resonance in artificial neural networks (Review paper)

Vadim R. Lutsiv^{1✉}, Maria A. Mikhalkova², Valeria O. Yachnaya³

^{1,3} Saint Petersburg State University of Aerospace Instrumentation, Saint Petersburg, 190000, Russian Federation

^{2,3} Pavlov Institute of Physiology of the Russian Academy of Sciences, Saint Petersburg, 199034, Russian Federation

¹ vluciv@mail.ru✉, <https://orcid.org/0000-0003-1160-3758>

² mikhalkova_maria@infran.ru, <https://orcid.org/0000-0002-7186-5744>

³ yachnaya_valeria@infran.ru, <https://orcid.org/0000-0002-7239-2096>

© Луцив В.Р., Михалькова М.А., Ячная В.О., 2025

Abstract

The observable world is hierarchically structured. The interaction technique of counter motion of information flows in the hierarchically organized systems was named “the adaptive resonance” and successfully modeled in artificial neural network of olfactory stimuli analysis and then applied for image recognition. The usefulness of application of the principles of hierarchical structural analysis and adaptive resonance was then forgotten for a long time. Recently, these principles were applied again in the capsule neural networks that outperformed the best modern models of other neural networks. This shows the necessity of systematizing the ways of practical implementation of these principles. The experience of application of structural analysis and adaptive resonance in the tasks of image recognition by artificial neural networks was inspected through the scientific and technical materials published last half-century. The comparative analysis carried out confirmed the efficiency of application of these principles in automatic image processing. The methods of most efficient realization of structural analysis and adaptive resonance in artificial neural networks were also determined. While the successful results of image recognition were reached by convolutional neural networks, their developers consigned to oblivion the principles of structural analysis and adaptive resonance following from the organization peculiarities of observable environment. However, the come-back to application of these principles would result in additional success in solving the tasks of image processing; thus, the further investigation of artificial neural networks is worth to be carried out in this area.

Keywords

artificial neural networks, structural analysis, adaptive resonance, hierarchical organization, convolutional networks, capsule networks, transformers

For citation: Lutsiv V.R., Mikhalkova M.A., Yachnaya V.O. Image structural analysis and adaptive resonance in artificial neural networks (Review paper). *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2025, vol. 25, no. 2, pp. 273–285 (in Russian). doi: 10.17586/2226-1494-2025-25-2-273-285

Введение

Организация окружающего мира неоднородна. В нем можно выделить объекты и подобъекты, отличающиеся по определенным свойствам друг от друга.

Как живые зрительные системы, так и имитирующие их работу алгоритмы компьютерного зрения учитывают системное взаимодействие информации, извлеченной из анализируемых изображений на разных иерархических уровнях ее обработки. Структурированная информация более низкого иерархического уровня анализируется на более высоком уровне как единое (не разделенное на части) целое, — и качество этого анализа также соответствует более высокому уровню. В свою очередь, результаты, сформированные при анализе более высокого уровня, передаваемые «вниз по иерархии», позволяют скорректировать в нужную сторону методы и результаты анализа информации на более низких иерархических уровнях. Частный вариант такого двунаправленного анализа был назван — «адаптивным резонансом» [1–3]. Многие разработчики, продуктивно реализуя этот принцип, адаптивным резонансом его не называют, однако этот термин по своему смыслу хорошо соответствует взаимодействию в иерархической системе встречных (восходящих и нисходящих) потоков информации, приводящему к получению нового ее качества. Исходя из этого, такое взаимодействие в настоящей работе будет называться адаптивным резонансом. Это позволит также упорядочить терминологию, применяемую разными авторами, и сделать более объективным сравнительный анализ их достижений.

**Структурный анализ информации
в искусственных нейронных сетях:
первые шаги к применению**

Современный «бум» в исследовании искусственных нейронных сетей начался с момента опубликования в

работе [4] простого и эффективного алгоритма обучения многослойных перцептронов методом обратного распространения ошибки. Однако в скором времени стало очевидным, что представление изображения как единого целого не позволяет многослойным перцептронам распознавать образы со сложными локальными изменениями формы, например изображения рукописных символов. В связи с этим японский исследователь Ямада в многослойном перцептроне вместо полносвязных слоев нейронов (в которых каждый нейрон последующего слоя соединен с каждым нейроном предыдущего) применил локальные связи нейронов между соседними слоями: каждый нейрон последующего слоя был связан только с локальной группой соседних нейронов предыдущего [5]. Такой подход позволил независимо анализировать локальные детали изображения и повысить надежность распознавания рукописных символов с 98,3 % до 98,8 %. Отметим, что данное решение можно считать шагом к структурному анализу, хотя пока еще в явном виде отдельные структурные элементы изображения не выделялись.

Структурные методы распознавания имеют существенное преимущество перед методами, анализирующими изображение как единое целое: отдельные элементы изображения могут не иметь строго фиксированного положения, но при этом локально пространственно смещаться друг относительно друга при условии, что величина смещения находится в допустимых пределах. Такая возможность была реализована К. Фукусимой в работе [6] в сети «Неокогнитрон». Отметим, что выделяемые сетью «Неокогнитрон» структурные элементы могли в различных анализируемых изображениях немного изменять свою форму. В то же время в этой сети не были реализованы ни целенаправленное выделение структурных элементов заданных типов (локально устойчиво выделялись элементы, не соответствующие какому-то заданному классу объектов), ни описание в явном виде их взаимного пространственного положения.

С момента разработки сети AlexNet [7] в 2011 году прошло уже почти полтора десятилетия. За это время было создано огромное количество модификаций сверточных сетей, однако большинство из них не реализуют структурный анализ изображения в явном виде: структурные элементы осмысленной формы не выделяются, и их взаимное положение в явном виде не описывается, поэтому в настоящей работе они не будут рассмотрены.

В работе [8] К. Фукусимой наряду с неявным использованием структурного анализа был применен метод адаптивного резонанса. Однако он был не первым исследователем, применившим адаптивный резонанс в искусственных нейронных сетях.

Первое применение адаптивного резонанса в искусственных нейронных сетях

Термин «адаптивный резонанс» был впервые применен С. Гроссбергом в работе [1] при анализе процессов обработки информации (в частности, обонятельных стимулов) в нейронах коры головного мозга. В работе [2] был представлен доклад о создании искусственной нейронной сети ART-1, реализующей при обработке информации принципы адаптивного резонанса, описанные в [1]. В [3] представлены результаты разработки более известной и совершенной искусственной нейронной сети ART-2, ориентированной на обработку не только бинарных, но и аналоговых сигналов. Стоит отметить, что, формально, в этих нейронных сетях усилия разработчиков направлены на автоматическое формирование «эталонных описаний» (адаптивных фильтров), объединяющих входные образы в группы, характеризующиеся сходными признаками (на адаптивную категоризацию образов). Однако, фактически, эта настройка выполняется путем пропуска входных сигналов («снизу вверх») через адаптивные фильтры (выполняется распознавание входных образов), а параметры фильтров модифицируются под управлением «ожиданий», передаваемых «сверху вниз» в результате выполненного распознавания входных сигналов, т. е. происходит распознавание сигналов с применением принципа адаптивного резонанса.

В отличие от результатов, полученных в работе [2], в [8] применен принцип адаптивного резонанса в явном виде для распознавания изображений. В работе [8] такой принцип назван «селективным вниманием», хотя автор знаком с методом адаптивного резонанса (упоминает работу [2] в списке литературы по нейронным сетям). С применением адаптивного резонанса К. Фукусиме в работе [8] удалось при выполнении распознавания автоматически разделять наложенные друг на друга изображения, подавлять искажающий шум и восстанавливать утерянные детали. На рис. 1 представлена вычислительная схема, иллюстрирующая направления «прямого» прохождения информации о входном изображении U_{c0} (при выполнении его поэтапной фильтрации и распознавания) и «обратного» прохождения сигналов, сформированных на более высоких иерархических уровнях анализа изображения и корректирующих обработку информации на более низких иерархических уровнях. Схема дана в том виде, как она была представлена в работе [8], не существенные для обсуждения детали нами опущены. Рис. 2 [8] иллюстрирует постепенную коррекцию анализируемого изображения под действием операций адаптивного резонанса, выполняемых согласно вычислительной схеме (рис. 1).

Нейронные сети, впервые в явном виде применившие методы структурного анализа и адаптивного резонанса

Дополнительные возможности для структурного анализа появляются, если в отличие от сверточных сетей Фукусимы [6, 8] и Хинтона [7] нейронной сетью анализируются целенаправленно выделенные структурные элементы заданных типов и их взаимные пространственные положения. Именно такие нейронные сети были разработаны в 80-х годах прошлого века. В основу этих разработок была положена нейронная сеть Хопфилда–Танка [9]. В 90-х годах в Государственном оптическом институте им. С.И. Вавилова в Санкт-Петербурге сеть Хопфилда–Танка использовалась сначала для установления взаимного соответствия тексе-

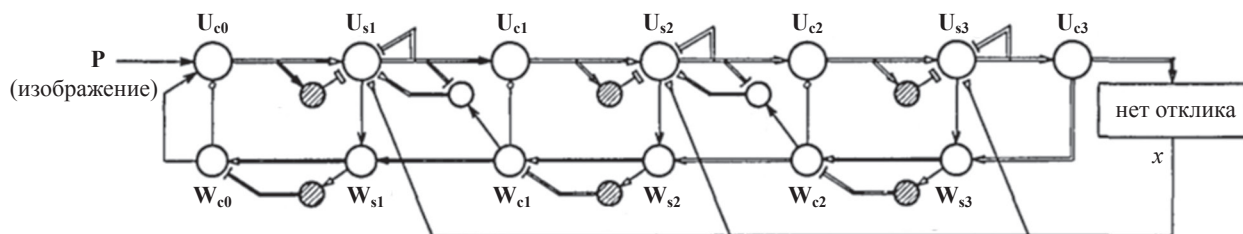


Рис. 1. Схема прохождения сигналов «прямого» и «обратного» распространения в нейронной сети [8].

U_{s1}, U_{s2}, U_{s3} — матрицы (слои) нейронов S-типа, реализующие фильтры для узнавания локальных деталей изображения;
 $U_{c0}, U_{c1}, U_{c2}, U_{c3}$ — матрицы (слои) нейронов «подвыборки»; $W_{c0}, W_{c1}, W_{c2}, W_{s1}, W_{s2}, W_{s3}$ — матрицы нейронов, корректирующие распознавание локальных деталей на основе результатов анализа более высоких иерархических уровней;
 P — матрица пикселей входного изображения; x — сигнал «нет отклика»

Fig. 1. Diagram of forward and backward signal propagation in the Fukushima's neural network [8].

U_{s1}, U_{s2}, U_{s3} — stand for the neuron matrices (layers) of S-type implementing the filters for local image details recognition;
 $U_{c0}, U_{c1}, U_{c2}, U_{c3}$ — stand for the neuron matrices of pooling layers; $W_{c0}, W_{c1}, W_{c2}, W_{s1}, W_{s2}, W_{s3}$ — stand for the neuron matrices correcting the local details recognition processes based on the analysis results reached at the higher hierarchical levels;
 P — matrix of input image pixels; x — “no response” signal

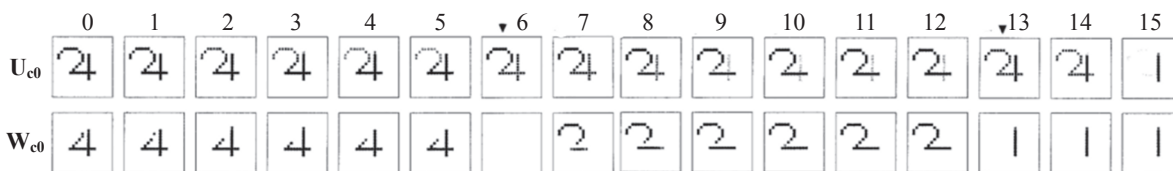


Рис. 2. Процесс коррекции пикселей анализируемого изображения U_{c0} под управлением «гипотезы» W_{c0} об ожидаемых их значениях, итеративно улучшаемой на более высоких иерархических уровнях анализа [8].

1–15 — номера итераций обработки изображения; «▼» — моменты переключения внимания

Fig. 2. Correction of the pixels of input image U_{c0} using the hypothesis W_{c0} about their expected values that is iteratively enhanced at the higher hierarchical levels of analysis [8] (the sequential digits 1–15 stand for the image processing iteration numbers, “▼” shows the attention switching moments)

лов в стереопарах [10, 11], а затем для структурного сопоставления контурных структурных описаний распознаваемого изображения и эталона [12–15].

Двумерный решающий слой нейронной сети Хопфилда–Танка представляет собой динамическую матрицу соответствия, строкам которой соответствуют структурные элементы первого объекта, а столбцам — структурные элементы второго. Если параметры i -го структурного элемента первого объекта совпадают с заданной точностью с параметрами j -го элемента второго, в матрице соответствия становится активным нейрон на пересечении i -й строки и j -го столбца, причем активность нейрона пропорциональна структурному сходству соответствующих ему сопоставленных элементов. В качестве сопоставляемых структурных элементов сеть использовала результаты аппроксимации контуров, выделенных на двух сопоставляемых изображениях, отрезками прямых линий и углами между ними (как показано в [15], такие структурные описания наиболее устойчивы к естественным изменениям наблюдаемых сцен).

Отметим, что нейронная сеть Хопфилда–Танка в наиболее полной мере реализовала преимущества структурного анализа. Использовались наиболее устойчивые структурные элементы целенаправленно выделенных типов, и при их сопоставлении учитывалось их взаимное пространственное положение. Действительно, результаты компьютерного моделирования [12] показали, что выполняемое сетью структурное сопоставление оказалось очень устойчивым к взаимному геометрическому преобразованию сопоставляемых изображений, шуму и искажению параметров формы структурных элементов.

Однако сложность представленной в [12] нейронной сети (количество активно действующих связей между нейронами) было приблизительно пропорционально четвертой степени числа структурных элементов. Эта проблема была частично решена построением иерархической нейронной сети Хопфилда–Танка для сопоставления двухуровневых иерархических структурных описаний [13, 14]. Сопоставляемые изображения делились на локальные регионы. На нижнем иерархическом уровне сопоставлялись контурные структурные описания отдельных регионов, а на верхнем уровне сопоставляемыми структурными элементами были сами локальные регионы. Кроме того, в данной иерархической сети был реализован принцип адаптивного

резонанса. В отличие от сетей Гроссберга и Фукусимы, где с использованием этого принципа настраивались весовые коэффициенты суммирования пикселей, в данном случае адаптивный резонанс применялся другим способом: корректировался состав групп структурных элементов, входящих в локальные регионы, сформированные на изображении.

Применение в [13, 14] иерархического структурного описания и адаптивного резонанса в двухуровневой сети Хопфилда–Танка позволило работать с увеличенным числом контурных структурных элементов, однако, в целом, трудность увеличения их количества преодолена не была. Проблема была решена путем перехода от использования нейронной сети к применению при структурном сопоставлении метода оптимизированного обхода дерева решений [16]. В реализующем дерево решений алгоритме [17] также использовалось сопоставление иерархических контурных структурных описаний с применением принципа адаптивного резонанса (пример коррекции этим методом контурных структурных описаний, показан на рис. 3), а в [18] авторы успешно расширили этот алгоритм для сопоставления изображений трехмерных сцен. В исследовании [17] (расширенное его описание затем было приведено в [15]) была достигнута возможность сопоставления структурных описаний чрезвычайно сильно измененных изображений, как это методом их мозаичного наложения проиллюстрировано на рис. 4. Это продемонстрировало возможности методов структурного анализа и адаптивного резонанса, однако примененный подход, основанный на обходе дерева решений, более не был связан с нейронными сетями.

Первая сверточная сеть, выполняющая иерархический структурный анализ изображения

Как было отмечено, сети Фукусимы и Хинтона и многие другие сверточные сети выделяют локальные элементы изображения, но в явном виде его структурного анализа не выполняют. При развитии сверточных сетей Росс Гиршик с коллегами представил на конференции CVPR-2015 искусственную сеть, в явном виде реализующую структурный анализ изображения [19]. В качестве основы разработки был использован алгоритм обнаружения объектов с раздельным анализом их частей [20], который занял второе после сверточных сетей место в соревнованиях компьютерных программ

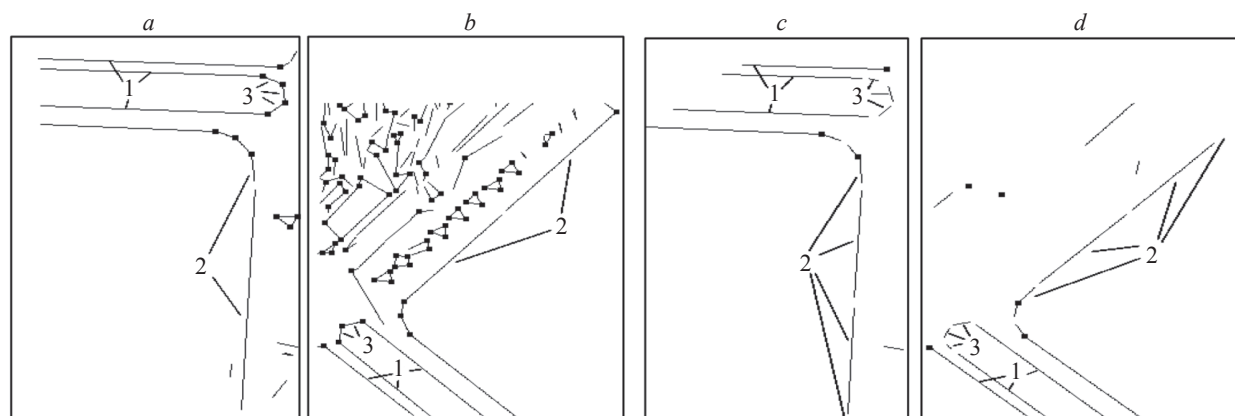


Рис. 3. Коррекция сопоставляемых структурных описаний с применением принципа адаптивного резонанса. Пары сопоставляемых контурных структурных описаний: до (a, b) и после (c, d) коррекции. Цифрами 1, 2 и 3 помечены группы структурных элементов, соответствующие друг другу в исходной и исправленной паре описаний

Fig. 3. Correction of matched structural descriptions based on application of adaptive resonance: a pair of initial structural descriptions (a) and (b); the same descriptions corrected based on adaptive resonance (c) and (d). The digits 1, 2 and 3 mark the groups of structural elements corresponding to each other in the initial and in the corrected pair of descriptions

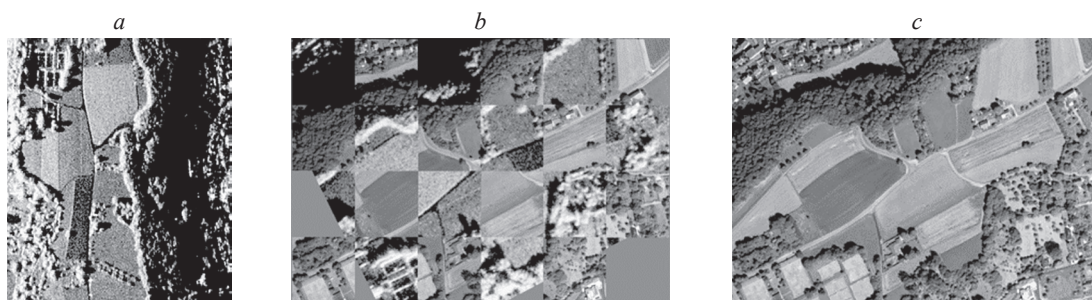


Рис. 4. Структурное сопоставление снимков: радиолокационного (a) и выполненного в видимом оптическом диапазоне (c); мозаичное наложение (регистрация) сопоставленных снимков (b) в общей системе координат

Fig. 4. Structural matching of the radar image (a) and optical image (c) of the same terrain; the matching result (b) is shown by registration of these images in mosaic fashion in a common coordinate system

распознавания изображений 2011 года¹. Утверждается также, что использованная в [19] технология может быть применена и для реализации других неитеративных алгоритмов обнаружения объектов с раздельным анализом частей, например, как в работе [21]. Однако, если решение принимается итеративными алгоритмами, такими как [22], их реализация методами, предложенными в [19], требует дальнейшего переосмысления.

В работе [19] для удобства сравнения с алгоритмом-прототипом [20] технология сверточных сетей использована только для формирования векторов признаков. Конечная их классификация, аналогично [20], выполнена с помощью машины опорных векторов. В продолжение этого исследования в работе [23] сформирована единая сверточная сеть, выполняющая как формирование векторов признаков вышеописанным способом, так и их классификацию. Как утверждают авторы [19], их решение на основе технологии сверточных сетей позволило существенно (с 33,4 % до 42 % на основе базы изображений PASCAL VOC 2010) улучшить среднюю точность обнаружения объектов

по сравнению с результатами прототипа [20]. В ряде случаев дальнейшее улучшение точности обнаружения было получено с полным применением технологии нейронных сетей [23] (для формирования векторов признаков и их классификации). Однако результаты обнаружения объектов алгоритмами [19, 23] оказались значительно хуже средней точности обнаружения (53,7 %), достигнутой сетью R-CNN [24]. С тех пор было предложено еще много других более эффективных сверточных сетей, поэтому алгоритм обнаружения объектов с раздельным анализом частей и его реализации нейронными сетями теперь упоминаются нечасто. Отметим, что хотя в [19] впервые за долгое время и был реализован в нейронной сети иерархический структурный анализ изображения, о преимуществах применения адаптивного резонанса разработчики искусственных нейронных сетей еще долгое время не вспоминали.

Капсульные сети

Наконец, во время победного шествия сверточных нейронных сетей около десяти лет назад в своем блоге²

¹[Электронный ресурс]. <http://www.image-net.org> (дата обращения: 29.06.2023).

²Электронный ресурс]. <https://www.reddit.com/user/geoffhinton/> (дата обращения: 06.08.2024).

Джеффри Хинтон отметил, что применение операции подвыборки (pooling) было большой ошибкой: оно уничтожает информацию о том, где находятся выделенные сверточной сетью детали изображения. Вместо этого он предложил кодировать информацию о пространственном положении выделенных деталей относительно объекта в целом и реализовать такой пространственный анализ в специальных «капсулах», являющихся расширенной версией обычного искусственного нейрона, применяемого в сверточных сетях [25]. Особенности пространственного положения и начертания деталей изображения описываются в капсулах специальным вектором, длина которого также кодирует информацию о вероятности присутствия на изображении в данном положении деталей заданных классов.

В то же время в работе [25] матрицы, задающие положения деталей, должны кем-то задаваться извне. Остается также не до конца решенная проблема присутствия в изображении нескольких подлежащих распознаванию объектов одновременно. Эти проблемы решаются Джеффри Хинтоном, Сарой Сэбоур и Николасом Фросстом в более поздней работе [26], использующей динамическую маршрутизацию информации между капсулами, которая реализует ничто иное как принцип адаптивного резонанса (правда, данный термин в [26] не использован). Отметим, что через 30 лет после опубликования работы Фукусимы [8] разработчики нейронных сетей снова вернулись к указанной тематике. Причем в работе [26] результаты применения капсульной сетью иерархического структурного анализа и адаптивного резонанса иллюстрируются, как и у Фукусимы (рис. 2), разделением наложенных друг на друга изображений рукописных цифр (цветные цифры на рис. 5 [26]). Правда, теперь это решение получено на качественно более высоком уровне: метод адаптивного резонанса применяется не просто для регулирования коэффициентов передачи пикселей локальных деталей изображения объекта, а для принятия решения, соответ-

ствуют ли эти детали его структурному описанию. При этом структурное описание уже приобрело формально-логическую форму (описывается в [26] в виде аффинного преобразования координат деталей в систему координат всего объекта), хотя эти детали пока все еще не имеют устойчивого смыслового значения и в скрытом виде выделяются в изображении капсульной сетью.

Вычислительная схема, реализующая динамическую маршрутизацию (адаптивный резонанс) в капсульной сети [26] представлена на рис. 6. Вектор $\mathbf{u}_i$ кодирует вероятность обнаружения i -ой капсулой предыдущего иерархического уровня соответствующего ей (i -го) структурного элемента в позиции, соответствующей направлению этого вектора; $\mathbf{W}_{ij}$ — матрица пересчета вектора описания i -го найденного структурного элемента в вектор $\tilde{\mathbf{u}}_{ji}$ в системе координат всего объекта, включающего всего M таких структурных элементов. Результатом взвешенного суммирования векторов $\tilde{\mathbf{u}}_{ji}$, описывающих наличие и положение структурных элементов, найденных капсулами нижнего уровня, является вектор $\mathbf{s}_j$, описывающий прогнозируемое положение и вероятность обнаружения объекта j -го класса. Результатом нормализации вектора $\mathbf{s}_j$, аналогичной вычислению функции сигмоиды в сверточных сетях, является выходной вектор $\mathbf{v}_j$ j -ой капсулы высшего иерархического уровня. Весовой коэффициент $\mathbf{c}_{ij}$, вычисленный пропорционально скалярному произведению векторов $\tilde{\mathbf{u}}_{ji}$ и $\mathbf{s}_j$, показывает соответствие пространственного положения найденного i -го структурного элемента пространственному положению объекта как единого целого. При некорректном пространственном положении найденного элемента соответствующий ему весовой коэффициент $\mathbf{c}_{ij}$ уменьшается, уменьшая таким образом влияние этого элемента на принятие решение об обнаружении объекта j -го класса.

Авторы капсульной сети утверждают [26], что по сравнению со сверточными сетями традиционной архитектуры они получили практически одинаковую точность распознавания изображений на базе данных MNIST (99,3 % против 99,2 %), но существенно улучшили точность на базе данных AFFNIST (79 % против 66 %). Это свидетельствует о более эффективной работе их сети с геометрически преобразованными изображениями, содержащимися в базе AFFNIST, и подтверждает эффективность явного применения в капсулах иерархического структурного анализа, поддержанного применением адаптивного резонанса.

Подобно тому, как победа сверточной сети [7] в соревновании ImageNet воодушевила исследователей к разработке целой плеяды сверточных сетей, вслед за опубликованием в [25, 26] идеи капсульных сетей появляется заметное количество работ, развивающих идею применения капсул. Отметим некоторые из таких работ.

В работе [27] были разрешен ряд трудностей, характерных для предыдущей капсульной сети [26], но главной особенностью новой сети стало изменение пути реализации динамической маршрутизации. Она выполняется в [27] с применением метода ожидания–максимизации, что улучшило возможность различения похожих объектов. Ошибка классификации уменьшилась



Рис. 5. Примеры раздельного распознавания сетью наложенных цифр: L — истинные цифры; R — цифры, реконструированные сетью; P и «*» — ошибки, обусловленные неправильной классификацией или реконструкцией, или предъявлением сети неизвестных ей символов

Fig. 5. Examples of separate recognition of superposed digits: L — the true digits; R — the ones reconstructed by network; P and «*» stand for the errors caused by incorrect classification and reconstruction or by unfamiliar input symbols respectively

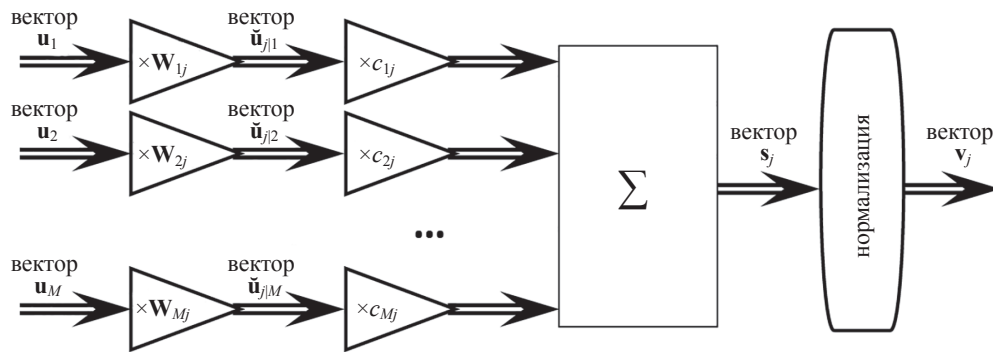


Рис. 6. Реализация динамической маршрутизации (адаптивного резонанса) j -ой капсулой высшего иерархического уровня при обнаружении объекта, соответствующего j -го класса

Fig. 6. Realization of dynamic routing (adaptive resonance) by the j -th capsule detecting an object of j -th type in the higher hierarchical level

по сравнению с предыдущей капсульной сетью [26] (1,4 % против 2,7 %) и значительно сильнее уменьшилась по сравнению со сверточными сетями новейшей архитектуры (для последних ошибка составляла 5,2 %). Сравнение производилось на основе изображений из базы данных SmallNORB [28].

Важной при анализе изображений, включающих не один, а несколько объектов, является необходимость принятия решения, к какому из объектов принадлежат найденные в изображении составляющие их части. Эта проблема решается в иерархическом капсульном автоэнкодере (stacked capsule autoencoder) [29], который в отличие от ранее разработанной капсульной сети [26] обучается «без учителя». «Иерархическая» структура автоэнкодера имеет два уровня. Нижний уровень отвечает за обнаружение и определение параметров пространственного положения отдельных частей, составляющих присутствующие в изображении объекты. Верхний уровень определяет отдельные объекты, состоящие из обнаруженных частей, и уточняет пространственное положение последних. Нижний уровень автоэнкодера самообучается формированию эталонных описаний частей, составляющих объекты, и определению пространственных положений этих частей, максимизируя вероятность правильного восстановления входного изображения из выученных эталонных описателей их формы и параметров их пространственного положения. Аналогичным образом обучаются «без учителя» и капсулы верхнего иерархического уровня автоэнкодера. При обучении максимизируется правдоподобие наличия и параметров положения частей, обнаруженных на нижнем уровне и соответствующих каждому из объектов, существующих в изображении. Максимизируется также правдоподобие соответствия композиции пикселей входного изображения наличию в нем объектов, состоящих из обнаруженных частей.

«Изоминкой» верхнего иерархического уровня автоэнкодера [29] является применение в нем трансформера (set transformer) [30], выполняющего назначение «хаотично» найденных в изображении частей предположительно существующим в нем объектам. Трансформер [30] распределяет обнаруженные части по соответствующим им объектам, выполняя их «адаптивную кластеризацию».

По аналогии с разработанной капсульной сетью [26], иерархический капсульный автоэнкодер [29] применяет при анализе изображения принцип адаптивного резонанса. Однако теперь в [29] не капсулы, относящиеся к частям, предсказывают наличие и положение объекта соответствующего класса, а сами объекты «голосуют» за существование относящихся к ним частей с соответствующими пространственными положениями. Резюмируя рассмотрение капсульной сети [29], отметим, что структурный анализ все еще выполняется в ней в не полностью явном виде. Выделяемые на изображении структурные элементы (части объектов) не имеют устойчивого смыслового значения и в скрытом виде выделяются капсульной сетью в качестве часто встречаемых (характерных) элементов обучающего набора. Например, в [29] в качестве таких структурных элементов упоминаются не имеющие осмысленной формы локальные элементы контуров рукописных символов.

Также в работе [29] отмечено, что авторам удалось при классификации изображений из базы данных MNIST немного превзойти точность наилучших решений [31] распознавания изображений при использовании классификаторов, обученных «без учителя» (99 % против 98,7 %).

Еще одним вариантом реализации капсульных сетей является применение «канонических капсул» (canonical capsules). В работе [32] с их помощью выполняется разделение облака точек в трехмерном пространстве на группы, выполняющие роль структурных элементов, однако в дальнейшем для структурного анализа рассматриваемой трехмерной сцены эти группы никак не используются.

Со времени публикации работы [26] появилось большое количество исследований по капсульным сетям. Обратимся теперь к трансформерам, которые в некоторых приложениях выполняют более полноценный, чем капсульные сети Хинтона структурный анализ.

Трансформеры

Появление трансформеров обычно связывают с опубликованием работы [33], в заголовке которой уже сформулирована основная идея — применение при ана-

лизе данных «механизма внимания». Нейронная сеть «Трансформер» [33] была первоначально разработана для выполнения анализа и преобразования текстов, которые четко разделены на устойчивые смысловые единицы — слова, для которых однозначно определены позиции в анализируемой фразе. Предложенный в [33] механизм внимания позволял определять результаты анализа фразы смысловыми связями между составляющими ее словами, что можно считать вариантом иерархического структурного анализа текстовых данных. Более того, в декодирующей части трансформера результаты анализа более высокого уровня (сформированная уже часть преобразованной фразы) используются с применением синтезированного при обучении трансформера механизма внимания для выбора последующих частей фразы, что можно интерпретировать как применение принципа адаптивного резонанса.

Результаты, достигнутые трансформером при преобразовании текстов, убедительно демонстрируют преимущества применения иерархического структурного анализа и адаптивного резонанса. Однако попытка использования трансформеров для обработки изображений (Image transformer [34]), далека от применения полноценного структурного анализа. В качестве структурных элементов, которым в классическом трансформере [33] соответствовали слова, в трансформере изображений [34] использованы отдельные пиксели, упорядоченные в порядке растрового сканирования изображения, и в кодах их пространственных позиций использован этот же порядок. Кроме того, совокупность пикселей поделена на локальные блоки, внутри которых и применяется локально механизм внимания (в отличие от глобального его применения в трансформере [33]). В связи с локальным характером анализа изображения результаты его обработки также имели достаточно локальный характер: обрабатывались изображения очень малого размера (всего 32×32 пиксела) в задачах повышения их пространственного разрешения или синтеза их отсутствующих фрагментов. При локальном характере анализа изображения речь о классификации его содержания как единого целого идти не могла, а использование пикселей в качестве объектов, обрабатываемых механизмом внимания, делало очень локальной область его применения (вычислительные затраты, связанные с применением этого механизма, растут пропорционально квадрату числа объектов внимания).

Перечисленные проблемы трансформера изображений [34] были достаточно успешно решены Алексеем Досовицким и его коллегами в работе [35]. В отличие от [34] они использовали в качестве структурных элементов изображения (токенов) не отдельные пиксели, а локальные его фрагменты размерами 16×16 пикселей. С одной стороны, это наполнило используемые структурные элементы более осмысленным содержанием. С другой стороны, увеличение размеров структурных элементов позволило уменьшить их количество и при приемлемых вычислительных затратах сделать область применения механизма внимания не локальной, а глобальной, включив в нее всю площадь изображения. Таким образом, в [35] были более полноценно, чем в

традиционных сверточных сетях, использованы преимущества применения структурного анализа, но при обработке изображений использовался только энкодер трансформера, а результаты кодирования классифицировались многослойным перцептроном, поэтому применение адаптивного резонанса в видео-трансформере [35] обнаружить не удастся.

Подобно трансформеру изображений [34] в видео-трансформере [35] авторы остановились на «одномерном» кодировании информации о пространственном положении структурных элементов: фрагменты, используемые в качестве токенов, упорядочены в виде одномерной цепочки, полученной растровым сканированием изображения. Точность классификации изображений, достигаемая в [35], оказывается сопоставимой с точностью, достигаемой большой моделью сверточной сети ResNet-152. Предположительно, это может быть связано с тем, что и в трансформере [35], и в сверточных сетях локально анализируемые фрагменты (структурные элементы) изображений выбираются случайным образом в сканирующем изображении окне. При сравнении с результатами классификации сетью ResNet-152, предобученной на малой базе данных ImageNet (ImageNet1k, содержащей 1,3 млн изображений), точность классификации видео-трансформером, предобученным на изображениях из значительно большей базы данных ImageNet21k (14 млн изображений), проигрывала (87,54 % против 85,3 %). Только при использовании предобучения на очень большой базе данных JFT-300M (33 млн изображений) видео-трансформеру, удалось получить точность классификации 87,6 % и «победить», наконец, сеть ResNet-152. Вероятно, на большей обучающей выборке применяемому в [35] механизму внимания удалось зафиксировать дополнительно и взаимосвязи таких «случайно выбираемых» структурных элементов при одномерном кодировании их пространственных положений.

GLOM

Проблема автоматического выбора структурных элементов, наконец, более рационально решается в вычислительной архитектуре GLOM (от английского agglomerate — собирать вместе), предложенной Д. Хинтоном в [36]. Хинтон предложил иерархически организованную искусственную нейронную сеть, на каждом уровне которой самоорганизуются «островки активности», соответствующие областям изображения, локально отличающимся по каким-либо признакам от окружения. Действительно, такой выбор структурных элементов, представляется более осмысленным, чем вырезание из изображения случайно выбранных фрагментов. Проиллюстрируем предложенную Хинтоном технологию формирования структурных элементов разного иерархического уровня (рис. 7).

В нейронных слоях более высокого иерархического уровня объединяются, согласно [36], пространственно соседние структурные элементы предыдущего уровня, обладающие сходством по каким-либо признакам (сходство показано на рис. 7 сходством цветов и направлений векторов выходных параметров нейронов). Таким об-

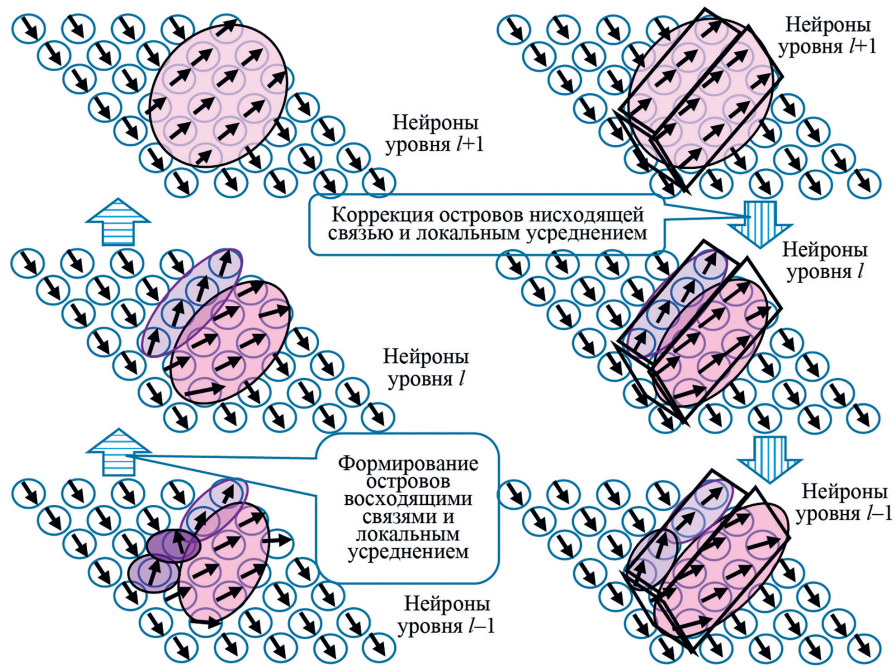


Рис. 7. Формирование островков активности, нейронов со сходными выходными параметрами [38]. На рисунке справа видно, как формируются однородные области, соответствующие граням параллелепипеда.

l — порядковый номер одного из случайно выбранных иерархических уровней системы

Fig. 7. Formation of activity islands composed of neurons with similar output parameters [38]; it can be seen at the right how the homogeneous areas corresponding to faces of parallelepiped are formed.

' l ' denotes a sequential number of one of randomly chosen hierarchical levels of system

разом в изображении постепенно из локальных частей собираются более крупные структуры. Одновременно с этим нисходящие (от более высоких иерархических уровней нейронной сети к более низким) информационные связи способствуют обособлению и увеличению однородности параметров нейронов локальных областей активности (более мелких структурных элементов) в слоях более низких уровней иерархии. Следовательно, в нейронной сети Хинтона реализуется принцип адаптивного резонанса.

Можно пояснить предложенный Хинтоном итеративный процесс формирования островков активности формулой:

$$\mathbf{O}_{l,j}(t) = \sqrt[3]{w_{l+1,j} \mathbf{O}_{l+1,j}(t) w_{l-1,j} \mathbf{O}_{l-1,j}(t) \sum_{i \in \Omega_j} \mathbf{O}_{l,i}(t-1) a_{l,i}(t)},$$

где $\mathbf{O}_{l,j}(t)$ — вектор выходных параметров j -го нейрона в слое l в момент времени t ; $w_{l-1,j}$ и $w_{l+1,j}$ — весовые коэффициенты связей с одноименными (j -ми) нейронами предыдущего и последующего слоев; $a_{l,i}$ — весовые коэффициенты усреднения векторов выходных параметров нейронов в окрестности Ω_j j -го нейрона l -го слоя. В разных частях работы [36] усреднение по окрестности Ω_j упоминается как «гауссова выборка» (gaussian sampling), либо как суммирование с весовыми коэффициентами «внимания» $a_{l,i}$, пропорциональными экспоненте скалярного произведения «взвешиваемых» векторов.

В рамках текущего исследования не удалось найти информацию о результатах реального воплощения архитектуры GLOM в виде программно реализованной

модели. В [36] архитектура GLOM представлена в виде идеи. Практическая реализуемость всех представленных в GLOM этапов иерархической обработки изображений с применением известных сегодня технологий, основанных на искусственных нейронных сетях, не вызывает сомнений, поэтому предложенные в [36] идеи эффективной реализации в нейронной сети методов иерархического структурного анализа и адаптивного резонанса представляются интересными и перспективными. Единственное, что вызывает сомнение в [36], — это формирование островков активности (пространственно обособленных структурных элементов) путем локального гауссова усреднения параметров выходов соседствующих в слое нейронов. Двумерная гауссова фильтрация выходных сигналов нейронов слоя может приводить к локальному размытию, а не к локализации однородных островков. Хинтон утверждает в [36], что для формирования островков активности с четкими границами, состоящих из нейронов, характеризующихся сходными выходными векторами, требуется также специальная организация взаимодействия нейронов между соседними слоями, которое он уподобляет регуляризации, но четкого обоснования эффективности такой регуляризации он не приводит. Однако известны эффективные способы формирования хорошо очерченных островков активности, например, с помощью кластеризации в самоорганизующихся картах Кохонена [37], что подтверждает реализуемость идей, представленных в [36]. В книге авторов настоящей работы [38] также приводятся соображения, как организовать в этом случае передачу информации о сфор-

мированных кластерах в нейронный слой следующего иерархического уровня, однако работоспособность этих идей еще требует подтверждения путем их воплощения в программно реализованных моделях.

Обсуждение

В приведенной таблице приведены итоги исследований, достигнутые в результате использования принципов структурного анализа и адаптивного резонанса при анализе изображений в искусственных нейронных сетях.

Полученные результаты (таблица) подтверждают спонтанность возврата исследователей к применению принципов структурного анализа и адаптивного резонанса: они «неожиданно» с разными названиями реализуются в новых нейронных архитектурах, при настройке которых используются «разношерстные» базы данных и разнородные оценки качества обучения, выполняемые в сравнении с разнообразными нейронными сетями-аналогами. Следует, однако, отметить, что в капсульных нейронных сетях, конструируемых в последние годы под руководством Джеффри Хинтона, систематически делаются попытки использовать преимущества применения этих принципов при обработке изображений. Эти преимущества оказались особенно заметными, когда решались задачи, связанные с анализом изображений, подвергающихся геометрическим

преобразованиям, поскольку именно в капсульных сетях наиболее гибко используется информация о взаимном пространственном положении фрагментов изображения, что, как раз, и определяет преимущества применения структурного анализа по сравнению с методами, рассматривающими изображение как единое целое.

Отметим также, что, к сожалению, в нейронных сетях, реализованных уже к настоящему времени в виде программных моделей, полноценное выделение структурных элементов в виде объектов осмысленной формы не производилось, поэтому выполняемый в них иерархический структурный анализ с применением адаптивного резонанса нельзя считать полноценным. Ближе всего к решению этой проблемы подошел Джеффри Хинтон в своей нейронной вычислительной архитектуре GLOM [36], которая пока представлена только в виде общей концепции. Идеи, реализованные в GLOM, представляются перспективными, поэтому можно предположить, что дальнейшие успехи в области автоматического анализа изображений могут быть достигнуты при их реальном практическом воплощении в программных моделях новых нейронных сетей.

Выполнение дальнейших исследований в направлении использования иерархического структурного анализа и адаптивного резонанса представляется тем более перспективным, что применение этих принципов в «классическом» трансформере [33] позволило уже

Таблица. Сравнение результатов анализа изображений нейронными сетями, применяющими методы структурного анализа и адаптивного резонанса, с результатами сетей-аналогов

Table. Comparison of the image analysis results in the neural networks applying the methods of structural analysis and adaptive resonance with the results of networks-analogs

Новая нейронная сеть	Результаты новой сети	Сеть-аналог, с которой выполнено сравнение	Результаты сети-аналога
Нейронная сеть [5] с локальными связями	Точность классификации 98,8 % на выборке изображений, сформированной в [5]	Многослойный перцептрон с полносвязными слоями нейронов	Точность классификации 98,3 % на той же выборке изображений, сформированной в [5]
Сеть [19] с раздельным анализом частей изображения	Точность обнаружения на базе PASCAL VOC 2010 42 %	R-CNN [24]	Точность обнаружения на базе PASCAL VOC 2010 53,7 %
Капсульная сеть [26]	Точность классификации на базе AFFNIST 79 %	Сверточные сети новейшей архитектуры (state of the art)	Точность классификации на базе AFFNIST 66 %
Капсульная сеть [27], применяющая метод ожидания-максимизации	Ошибка классификации 1,4 % на базе SmallNORB	Капсульная сеть [26]	Ошибка классификации 2,7 % на базе SmallNORB
Капсульная сеть [27], применяющая метод ожидания-максимизации	Ошибка классификации 1,4 % на базе SmallNORB	Сверточные сети новейшей архитектуры (state of the art)	Ошибка классификации 5,2 % на базе SmallNORB
Иерархический капсульный автоэнкодер [29]	Точность классификации на базе MNIST 99 %	Нейронная сеть [31], выполняющая ассоциативную кластеризацию	Точность классификации на базе MNIST 98,7 %
Видео-трансформер [35]	Точность классификации 85,3 %, предобучение на большой базе ImageNet1k (14 млн изображений)	ResNet-152	Точность классификации 87,54 %, предобучение на малой базе ImageNet1k (1,3 млн изображений)
Видео-трансформер [35]	Точность классификации 87,6 %, предобучение на очень большой базе JFT-300M (33 млн изображений)	ResNet-152	Точность классификации 87,54 %, предобучение на малой базе ImageNet1k (1,3 млн изображений)

достичь хороших результатов при обработке текстов. Рассмотренные результаты применения структурного анализа и адаптивного резонанса при сопоставлении изображений в контурном структурном классификаторе (рис. 3 и рис. 4) также подтверждают перспективность их использования. Заметим, что, к сожалению, этот структурный классификатор был реализован в практически значимой форме пока не в виде нейронных сетей, а другим способом [17] — на основе технологии оптимизированного обхода дерева решений.

Заключение

Иерархическая организация наблюдаемого мира требует применения иерархических алгоритмов его

анализа, одной из основ которого является метод адаптивного резонанса. В представленном исследовании продемонстрировано что, создатели искусственных нейронных сетей прибегали к применению принципов структурного анализа и адаптивного резонанса эпизодически, на десятилетия забывая о них и вновь возвращаясь к их реализации, что целенаправленного анализа преимуществ и методов применения этих принципов в нейронных сетях не проводилось. В настоящей работе сделана попытка заполнить этот пробел и показать наиболее перспективные направления применения рассмотренных принципов при разработке алгоритмов обработки изображений в искусственных нейронных сетях на примерах капсульных нейронных сетей и вычислительной архитектуры GLOM.

Литература

1. Grossberg S. Adaptive pattern classification and universal recoding: II. Feedback, expectation, olfaction, illusions // *Biological Cybernetics*. 1976. V. 23. N 4. P. 187–202. <https://doi.org/10.1007/BF00340335>
2. Carpenter G.A., Grossberg S. Category learning and adaptive pattern recognition: a neural network model // *Proc. of the 3rd Army Conference on Applied Mathematics and Computing*. 1986. P. 37–56.
3. Carpenter G.A., Grossberg S. ART 2: Self-organization of stable category recognition codes for analog input patterns // *Applied Optics*. 1987. V. 26. N 12. P. 4919–4930. <https://doi.org/10.1364/AO.26.004919>
4. Rumelhart D.E. *Learning Internal Representations by Error Propagation*. Institute for Cognitive Science, 1985. 34 p.
5. Yamada K., Kami H., Tsukumo J., Temma T. Handwritten numeral recognition by multilayered neural network with improved learning algorithm // *Proc. of the International 1989 Joint Conference on Neural Networks*. 1989. P. 259–266. <https://doi.org/10.1109/IJCNN.1989.118708>
6. Fukushima K. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position // *Biological Cybernetics*. 1980. V. 36. N 4. P. 193–202. <https://doi.org/10.1007/bf00344251>
7. Krizhevsky A., Sutskever I., Hinton G.E. ImageNet classification with deep convolutional neural networks // *Proc. of the 26th International Conference on Neural Information Processing Systems (NIPS'12)*. 2012. V. 1. P. 1097–1105.
8. Fukushima K. A neural network model for selective attention in visual pattern recognition // *Biological Cybernetics*. 1986. V. 55. N 1. P. 5–15. <https://doi.org/10.1007/BF00363973>
9. Hopfield J.J., Tank D.W. Neural computation of decisions in optimization problems // *Biological Cybernetics*. 1985. V. 52. N 3. P. 141–152. <https://doi.org/10.1007/bf00339943>
10. Lutsiv V.R., Novikova T.A. On the use of neurocomputer for stereoimage processing // *Pattern Recognition and Image Analysis // Advances in Mathematical Theory and Applications*. 1992. V. 2. N 4. P. 441–444.
11. Данилов Е.П., Луцив В.Р., Новикова Т.А., Малышев И.А. Разработка методов и алгоритмов реализации интеллектуальных вычислительных устройств с использованием нейронных сетей в интересах задач анализа и сопоставления изображений. Отчет о работе 22201-007-94, Шифр «Нейросеть». СПб.: Государственный оптический институт им. С.И. Вавилова, С.-Петербург, 1994. С. 39–50.
12. Dolinov D.S., Zhrebko A.K., Lutsiv V.R., Novikova T.A. Using artificial neural networks in image processing problems // *Journal of Optical Technology*. 1997. V. 64. N 2. P. 112–118.
13. Lutsiv V.R., Malyshev I.A., Pepelka V.A. Automatic fusion of multiple-sensor and multiple-season images // *Proceedings of SPIE*. 2001. V. 4380. P. 174–183. <https://doi.org/10.1117/12.436990>
14. Lutsiv V.R., Malyshev I.A., Pepelka V.A., Potapov A.S. The Target independent algorithms for description and structural matching of

References

1. Grossberg S. Adaptive pattern classification and universal recoding: II. Feedback, expectation, olfaction, illusions. *Biological Cybernetics*, 1976, vol. 23, no. 4, pp. 187–202. <https://doi.org/10.1007/BF00340335>
2. Carpenter G.A., Grossberg S. Category learning and adaptive pattern recognition: a neural network model. *Proc. of the 3rd Army Conference on Applied Mathematics and Computing*, 1986, pp. 37–56.
3. Carpenter G.A., Grossberg S. ART 2: Self-organization of stable category recognition codes for analog input patterns. *Applied Optics*, 1987, vol. 26, no. 12, pp. 4919–4930. <https://doi.org/10.1364/AO.26.004919>
4. Rumelhart D.E. *Learning Internal Representations by Error Propagation*. Institute for Cognitive Science, 1985, 34 p.
5. Yamada K., Kami H., Tsukumo J., Temma T. Handwritten numeral recognition by multilayered neural network with improved learning algorithm. *Proc. of the International 1989 Joint Conference on Neural Networks*, 1989, pp. 259–266. <https://doi.org/10.1109/IJCNN.1989.118708>
6. Fukushima K. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 1980, vol. 36, no. 4, pp. 193–202. <https://doi.org/10.1007/bf00344251>
7. Krizhevsky A., Sutskever I., Hinton G.E. ImageNet classification with deep convolutional neural networks. *Proc. of the 26th International Conference on Neural Information Processing Systems (NIPS'12)*, 2012, vol. 1, pp. 1097–1105.
8. Fukushima K. A neural network model for selective attention in visual pattern recognition. *Biological Cybernetics*, 1986, vol. 55, no. 1, pp. 5–15. <https://doi.org/10.1007/BF00363973>
9. Hopfield J.J., Tank D.W. Neural computation of decisions in optimization problems. *Biological Cybernetics*, 1985, vol. 52, no. 3, pp. 141–152. <https://doi.org/10.1007/bf00339943>
10. Lutsiv V.R., Novikova T.A. On the use of neurocomputer for stereoimage processing. *Pattern Recognition and Image Analysis. Advances in Mathematical Theory and Applications*, 1992, vol. 2, no. 4, pp. 441–444.
11. Danilov E.P., Lutsiv V.R., Novikova T.A., Malyshev I.A. Development of methods and algorithms for the implementation of intelligent computing devices using neural networks for the purposes of image analysis and comparison problems. Statement of work no. 22201-007-94, Code “Neural network”. St. Petersburg, Open Joint Stock Company «S. I. Vavilov State Optical Institute», 1994, pp. 39–50. (in Russian)
12. Dolinov D.S., Zhrebko A.K., Lutsiv V.R., Novikova T.A. Using artificial neural networks in image processing problems. *Journal of Optical Technology*, 1997, vol. 64, no. 2, pp. 112–118.
13. Lutsiv V.R., Malyshev I.A., Pepelka V.A. Automatic fusion of multiple-sensor and multiple-season images. *Proceedings of SPIE*, 2001, vol. 4380, pp. 174–183. <https://doi.org/10.1117/12.436990>
14. Lutsiv V.R., Malyshev I.A., Pepelka V.A., Potapov A.S. The Target independent algorithms for description and structural matching of

- aerospace photographs // *Proceedings of SPIE*. 2002. V. 4741. P. 351–362. <https://doi.org/10.1117/12.478732>
15. Лутсив В. Автоматический анализ изображений: Объектно-независимый структурный подход. Саарбрюккен: Ламберт Академик Паблишинг, 2011. 308 с.
 16. Слэйгл Д.Р. Искусственный интеллект: подход на основе эвристического программирования. М.: Мир, 1973. 319 с.
 17. Lutsiv V.R., Malyshev I.A., Potapov A.S. Hierarchical structural matching algorithms for registration of aerospace images // *Proceedings of SPIE*. 2003. V. 5238. P. 164–175. <https://doi.org/10.1117/12.511770>
 18. Ponomarev S., Lutsiv V., Malyshev I. Automatic structural matching of 3D image data // *Proceedings of SPIE*. 2015. V. 9649. P. 96490M. <https://doi.org/10.1117/12.2194312>
 19. Girshick R., Iandola F., Darrell T., Malik J. Deformable part models are convolutional neural networks // *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015. P. 437–446. <https://doi.org/10.1109/cvpr.2015.7298641>
 20. Felzenszwalb P.F., Girshick R.B., McAllester D., Ramanan D. Object detection with discriminatively trained part-based models // *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2010. V. 32. N 9. P. 1627–1645. <https://doi.org/10.1109/tpami.2009.167>
 21. Yang Y., Ramanan D. Articulated human detection with flexible mixtures of parts // *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2013. V. 35. N 12. P. 2878–2890. <https://doi.org/10.1109/tpami.2012.261>
 22. Wang Y., Tran D., Liao Z. Learning hierarchical poselets for human parsing // *CVPR 2011*. P. 1705–1712. <https://doi.org/10.1109/cvpr.2011.5995519>
 23. Wan L., Eigen D., Fergus R. End-to-end integration of a convolutional network, deformable parts model and non-maximum suppression // *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015. P. 851–859. <https://doi.org/10.1109/cvpr.2015.7298686>
 24. Girshick R., Donahue J., Darrell T., Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation // *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*. 2014. P. 580–587. <https://doi.org/10.1109/CVPR.2014.81>
 25. Hinton G.E., Krizhevsky A., Wang S.D. Transforming auto-encoders // *Lecture Notes in Computer Science*. 2011. V. 6791. P. 44–51. https://doi.org/10.1007/978-3-642-21735-7_6
 26. Sabour S., Frosst N., Hinton G.E. Dynamic routing between capsules // *arXiv*. 2017. arXiv:1710.09829v2. <https://doi.org/10.48550/arXiv.1710.09829>
 27. Hinton G., Sabour S., Frosst N. Matrix capsules with EM routing // *Proc. of the 6th International Conference on Learning Representations (ICLR 2018)*. 2018. P. 1–15.
 28. LeCun Y., Huang F.J., Bottou L. Learning methods for generic object recognition with invariance to pose and lighting // *Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2004)*. 2004. P. 97–104. <https://doi.org/10.1109/cvpr.2004.1315150>
 29. Kosiorek A.R., Sabour S., Teh Y.W., Hinton G.E. Stacked capsule autoencoders // *arXiv*. 2019. arXiv:1906.06818v2. <https://doi.org/10.48550/arXiv.1906.06818>
 30. Lee J., Lee Y., Kim J., Kosiorek A.R., Choi S., Teh Y.W. Set Transformer: a framework for attention-based permutation-invariant neural networks // *arXiv*. 2019. arXiv:1810.00825v3. <https://doi.org/10.48550/arXiv.1810.00825>
 31. Haeusser P., Plapp J., Golkov V., Aljalbout E., Cremers D. Associative deep clustering: training a classification network with no labels // *Lecture Notes in Computer Science*. 2019. V. 11269. P. 18–32. https://doi.org/10.1007/978-3-030-12939-2_2
 32. Sun W., Tagliasacchi A., Deng B., Sabour S., Yazdani S., Hinton G., Yi K.M. Canonical capsules: self-supervised capsules in canonical pose // *arXiv*. 2021. arXiv:2012.04718v2. <https://doi.org/10.48550/arXiv.2012.04718>
 33. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser L., Polosukhin I. Attention is all you need // *arXiv*. 2017. arXiv:1706.03762v5. <https://doi.org/10.48550/arXiv.1706.03762>
 34. Parmar N., Vaswani A., Uszkoreit J., Kaiser L., Shazeer N., Ku A., Tran D. Image Transformer // *arXiv*. 2018. arXiv:1802.05751v3. <https://doi.org/10.48550/arXiv.1802.05751>
 35. Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai X., Unterthiner T., Dehghani M., Minderer M., Heigold G., Gelly S.,
 - aerospace photographs. *Proceedings of SPIE*, 2002, vol. 4741, pp. 351–362. <https://doi.org/10.1117/12.478732>
 15. Lutsiv V. *Automatic image analysis: Object-independent structured approach*. Lambert Academic Publishing, 2011, 308 p. (in Russian)
 16. Slagle J.R. *Artificial Intelligence: The Heuristic Programming Approach*. McGraw Hill, 1973, 196 p.
 17. Lutsiv V.R., Malyshev I.A., Potapov A.S. Hierarchical structural matching algorithms for registration of aerospace images. *Proceedings of SPIE*, 2003, vol. 5238, pp. 164–175. <https://doi.org/10.1117/12.511770>
 18. Ponomarev S., Lutsiv V., Malyshev I. Automatic structural matching of 3D image data. *Proceedings of SPIE*, 2015, vol. 9649, pp. 96490M. <https://doi.org/10.1117/12.2194312>
 19. Girshick R., Iandola F., Darrell T., Malik J. Deformable part models are convolutional neural networks. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 437–446. <https://doi.org/10.1109/cvpr.2015.7298641>
 20. Felzenszwalb P.F., Girshick R.B., McAllester D., Ramanan D. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010, vol. 32, no. 9, pp. 1627–1645. <https://doi.org/10.1109/tpami.2009.167>
 21. Yang Y., Ramanan D. Articulated human detection with flexible mixtures of parts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013, vol. 35, no. 12, pp. 2878–2890. <https://doi.org/10.1109/tpami.2012.261>
 22. Wang Y., Tran D., Liao Z. Learning hierarchical poselets for human parsing. *CVPR 2011*, pp. 1705–1712. <https://doi.org/10.1109/cvpr.2011.5995519>
 23. Wan L., Eigen D., Fergus R. End-to-end integration of a convolutional network, deformable parts model and non-maximum suppression. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 851–859. <https://doi.org/10.1109/cvpr.2015.7298686>
 24. Girshick R., Donahue J., Darrell T., Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 580–587. <https://doi.org/10.1109/CVPR.2014.81>
 25. Hinton G.E., Krizhevsky A., Wang S.D. Transforming auto-encoders. *Lecture Notes in Computer Science*, 2011, vol. 6791, pp. 44–51. https://doi.org/10.1007/978-3-642-21735-7_6
 26. Sabour S., Frosst N., Hinton G.E. Dynamic routing between capsules. *arXiv*, 2017, arXiv:1710.09829v2. <https://doi.org/10.48550/arXiv.1710.09829>
 27. Hinton G., Sabour S., Frosst N. Matrix capsules with EM routing. *Proc. of the 6th International Conference on Learning Representations (ICLR 2018)*, 2018, pp. 1–15.
 28. LeCun Y., Huang F.J., Bottou L. Learning methods for generic object recognition with invariance to pose and lighting. *Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2004)*, 2004, pp. 97–104. <https://doi.org/10.1109/cvpr.2004.1315150>
 29. Kosiorek A.R., Sabour S., Teh Y.W., Hinton G.E. Stacked capsule autoencoders. *arXiv*, 2019, arXiv:1906.06818v2. <https://doi.org/10.48550/arXiv.1906.06818>
 30. Lee J., Lee Y., Kim J., Kosiorek A.R., Choi S., Teh Y.W. Set Transformer: a framework for attention-based permutation-invariant neural networks. *arXiv*, 2019, arXiv:1810.00825v3. <https://doi.org/10.48550/arXiv.1810.00825>
 31. Haeusser P., Plapp J., Golkov V., Aljalbout E., Cremers D. Associative deep clustering: training a classification network with no labels. *Lecture Notes in Computer Science*, 2019, vol. 11269, pp. 18–32. https://doi.org/10.1007/978-3-030-12939-2_2
 32. Sun W., Tagliasacchi A., Deng B., Sabour S., Yazdani S., Hinton G., Yi K.M. Canonical capsules: self-supervised capsules in canonical pose. *arXiv*, 2021, arXiv:2012.04718v2. <https://doi.org/10.48550/arXiv.2012.04718>
 33. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser L., Polosukhin I. Attention is all you need. *arXiv*, 2017, arXiv:1706.03762v5. <https://doi.org/10.48550/arXiv.1706.03762>
 34. Parmar N., Vaswani A., Uszkoreit J., Kaiser L., Shazeer N., Ku A., Tran D. Image Transformer. *arXiv*, 2018, arXiv:1802.05751v3. <https://doi.org/10.48550/arXiv.1802.05751>
 35. Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai X., Unterthiner T., Dehghani M., Minderer M., Heigold G., Gelly S., Uszkoreit J., Houtsby N. An image is worth 16x16 words:

- Uszkoreit J., Hounsby N. An image is worth 16x16 words: transformers for image recognition at scale // arXiv. 2021. <https://doi.org/10.48550/arXiv.2010.11929>
36. Hinton G. How to represent part-whole hierarchies in a neural network // *Neural Computation*. 2023. V. 35. N 3. P. 413–452. https://doi.org/10.1162/neco_a_01557
37. Kohonen T. *Self-organization and Associative Memories*. Springer, 1984. 255 p.
38. Луцив В.Р., Михалькова М.А., Ячная В.О. Компьютерное зрение: учеб. пособие: в 3 ч. Ч. 3. Современные модификации архитектур нейронных сетей и методы улучшения изображений. СПб.: ГУАП, 2024. 196 с.
- transformers for image recognition at scale. *arXiv*, 2021, <https://doi.org/10.48550/arXiv.2010.11929>
36. Hinton G. How to represent part-whole hierarchies in a neural network. *Neural Computation*, 2023, vol. 35, no. 3, pp. 413–452. https://doi.org/10.1162/neco_a_01557
37. Kohonen T. *Self-Organization and Associative Memories*. Springer, 1984, 255 p.
38. Lutsiv V.R., Mikhalkova M.A., Iachnaia V.O. *Computer Vision. A Tutorial in 3 parts. Part 3: Modern Modifications of Neural Network Architectures and Methods of Image Enhancement*. St. Petersburg, GUAP, 2024, 196 p. (in Russian)

Авторы

Луцив Вадим Ростиславович — доктор технических наук, профессор, Санкт-Петербургский государственный университет аэрокосмического приборостроения, Санкт-Петербург, 190000, Российская Федерация, [sc 6602625465](https://orcid.org/0000-0003-1160-3758), <https://orcid.org/0000-0003-1160-3758>, vluciv@mail.ru

Михалькова Мария Анатольевна — младший научный сотрудник, Институт физиологии имени И.П. Павлова РАН, Санкт-Петербург, 199034, Российская Федерация, [sc 57218284288](https://orcid.org/0000-0002-7186-5744), <https://orcid.org/0000-0002-7186-5744>, mikhalkova_maria@infran.ru

Ячная Валерия Олеговна — младший научный сотрудник, Институт физиологии имени И.П. Павлова РАН, Санкт-Петербург, 199034, Российская Федерация; аспирант, Санкт-Петербургский государственный университет аэрокосмического приборостроения, Санкт-Петербург, 190000, Российская Федерация, [sc 57209076316](https://orcid.org/0000-0002-7239-2096), <https://orcid.org/0000-0002-7239-2096>, yachnaya_valeria@infran.ru

Authors

Vadim R. Lutsiv — D.Sc., Professor, Saint Petersburg State University of Aerospace Instrumentation, Saint Petersburg, 190000, Russian Federation, [sc 6602625465](https://orcid.org/0000-0003-1160-3758), <https://orcid.org/0000-0003-1160-3758>, vluciv@mail.ru

Maria A. Mikhalkova — Junior Researcher, Pavlov Institute of Physiology of the Russian Academy of Sciences, Saint Petersburg, 199034, Russian Federation, [sc 57218284288](https://orcid.org/0000-0002-7186-5744), <https://orcid.org/0000-0002-7186-5744>, mikhalkova_maria@infran.ru

Valeria O. Yachnaya — Junior Researcher, Pavlov Institute of Physiology of the Russian Academy of Sciences, Saint Petersburg, 199034, Russian Federation; PhD Student, Saint Petersburg State University of Aerospace Instrumentation, Saint Petersburg, 190000, Russian Federation, [sc 57209076316](https://orcid.org/0000-0002-7239-2096), <https://orcid.org/0000-0002-7239-2096>, yachnaya_valeria@infran.ru

Статья поступила в редакцию 02.11.2024
Одобрена после рецензирования 27.12.2024
Принята к печати 30.03.2025

Received 02.11.2024
Approved after reviewing 27.12.2024
Accepted 30.03.2025



Работа доступна по лицензии
Creative Commons
«Attribution-NonCommercial»