Санкт-Петербургский государственный университет
Кафедра теории систем управления электрофизической аппаратурой

Ращенко Дмитрий Владимирович

Выпускная квалификационная работа бакалавра

Разработка модифицированной нейронной сети адаптивной резонансной теории для решения задач распознавания изображений

Направление 010900
«Прикладные математика и физика»

Научный руководитель,
кандидат физ.-мат. наук,
доцент
Козынченко В. А.

Санкт-Петербург
2016

Оглавление

Введение

Искусственные нейронные сети

Широкий спектр современных задач, от идентификации частиц, возникающих в результате ядерных столкновений [1], до помощи в постановке медицинских диагнозов [2], решается с использованием искусственных нейронных сетей. Также нейронные сети, из-за их биологического подобия, находят глубокий интерес среди исследователей из разных областей [3, 4].

Искусственная нейронная сеть — это множество взаимодействующих вычислительных элементов (нейронов). Каждый нейрон может выполнять лишь простейшие операции, однако, объединяя их в сети, можно проводить сложные распределенные вычисления. Наиболее интересной с точки зрения информатики является модель нейрона Маккалока-Питтса [5], согласно которой, результатом работы нейрона является вещественная функция от линейной комбинации сигналов других нейронов. Для решения конкретных задач коэффициенты линейной комбинации каждого нейрона подбираются в процессе специального алгоритма (обучения).

Распространен подход, в котором нейроны с одинаковой активационной функцией и набором связанных нейронов объединяются в группы — слои. Хотя такой подход не соответствует реальным схемам взаимодействия нейронов в биологических нейронных сетях, он значительно упрощает проектирование и обучение искусственных. Следует отметить, что линейную комбинацию, вычисляемую в нейроне, легко представить в виде векторного произведения. Тогда результатом работы нейронного слоя является вектор-функция от произведения матрицы коэффициентов (далее — матрица весов) на вектор сигналов от связанных нейронов (далее — входной вектор). Иллюстрация данного подхода приведена в формуле 1, здесь $\vec{R}$ — выход нейронного слоя, $f(x)$ — активационная функция слоя, $w_{i,j}$ — коэффициент j-го нейрона для i-го входного сигнала, x_i — i-й входной сигнал, M — количество нейронов в слое, N — количество входных сигналов.

$$\vec{R} = \begin{pmatrix} f(\omega_{11}x_1 + \cdots + \omega_{1N}x_N) \\ \vdots \\ f(\omega_{M1}x_1 + \cdots + \omega_{MN}x_N) \end{pmatrix}_{M\times 1} \quad (1)$$

К преимуществам нейронных сетей можно отнести единый подход к решению широкого спектра задач и возможность ускорения вычислений за счет массового параллелизма. Параллелизм обеспечивается независимостью работы каждого нейрона в слое.

Задача распознавания образов

Выбирая количество нейронных слоев, порядок их соединения и алгоритмы обучения, нейронные сети можно настроить на решение различных задач: сжатие и архивация данных, прогнозирование (экстраполяция), интерполяция, аппроксимация, классификация и кластеризация. Особый интерес представляет задача распознавания образов [6] и, в частности, изображений. Перспективность данной задачи объясняется быстрым ростом количества производимой информации, в том числе графической. Эта информация является отличным набором данных для обучения, тестирования и применения нейронных сетей. В данной работе будет рассматриваться задача распознавания бинарных (черно-белых) изображений.

Искусственная нейронная сеть Хэмминга

Наиболее простой нейронной сетью для решения задачи распознавания бинарных изображений является сеть Хэмминга [5], основанная на одноименной метрике. Расстоянием Хэмминга между двумя элементами из пространства бинарных векторов фиксированной размерности называется количество различающихся компонент этих векторов.

Нейронная сеть Хэмминга состоит из двух слоев (сравнения, конкуренции) и имеет следующий алгоритм функционирования:

1. *Подготовка изображений.* Входные изображения (матрица пикселей с высотой H и шириной W) раскладываются в вектор размерности $N = HW$. Один из цветов пикселей (например черный) представляется как 1, другой как -1.

2. *Обучение.* Множество из M эталонных изображений записывается в строки весовой матрицы слоя сравнения.

3. *Сравнение.* Описанная весовая матрица умножается на вектор, подлежащий распознаванию, и получается вектор $\vec{C} = (C_1, ..., C_M)^T$, где C_i пропорциональна количеству совпавших пикселей в распознаваемом и i-м эталонном изображениях. Следует отметить, что величина C_i обратно пропорциональна расстоянию Хэмминга, и эталонный образ с максимальным C_i является наиболее близким ко входному в смысле меры Хэмминга.

4. *Конкуренция.* В слое конкуренции происходит поиск максимальной компоненты вектора $\vec{C}$. Эталон, соответствующий данной компоненте, является самым близким к входному образу в метрике Хэмминга. Это изображение и является результатом распознавания.

Преимущества

1. *Использование простой метрики Хэмминга.* Для сравнения, в нейронной сети АРТ-1 используется мера сходства изображений, которая не является метрикой в математическом смысле.

2. *Быстродействие.* Наличие всего двух нейронных слоев обеспечивает малую вычислительную сложность по сравнению с некоторыми другими нейронными сетями.

Возможные проблемы

Отсутствие пластичности. В традиционной нейронной сети Хэмминга не предусмотрена возможность обучения существующих эталонов

и выделения новых категорий. В литературе такая проблема называется отсутствием пластичности. Причиной такого поведения является безусловность определения победителя: во многих задачах требуется выделять новый эталон, если расстояние между распознаваемым и эталонным образом выше некоторого порога.

Искусственная нейронная сеть адаптивной резонансной теории

Адаптивная резонансная теория была разработана в 1987 году Стивеном Гроссбергом и Гейл Карпентер. Существует несколько нейронных сетей, построенных в соответствии с этой теорией [7, 8] и одной из них является АРТ-1 [9], предназначенная для распознавания бинарных изображений и учитывающая недостатки сети Хэмминга. Архитектура и алгоритм АРТ-1 приведены в главе 1.

Преимущества

Существенным преимуществом данной сети является наличие *пластичности*. Также, в отличие от некоторых других нейронных сетей, например персептрон [5], в АРТ-1 при обучении изменяется только эталон-победитель, не искажая все остальные. Данное преимущество называют свойством *стабильности*.

Возможные проблемы

1. *Производительность.* В силу особенностей алгоритма распознавания, вычислительная сложность АРТ-1 высока, по сравнению со многими другими нейронными сетями. Это может негативно сказаться, например, в задачах real-time распознавания [10].

2. *Асимметрия нулей и единиц.* В АРТ-1 при вычислении меры сходства используются только единичные элементы образов, и игнорируются

нулевые. Это снижает качество распознавания в некоторых задачах.

3. *Деградация запомненных образов.* В силу необратимости алгоритма обучения запомненная информация может теряться.

4. *Мера сходства изображений не является метрикой в математическом смысле.* Данный факт не является существенной проблемой, но использование простой меры Хэмминга облегчило бы тестирование и отладку нейронной сети.

Следует отметить, что приведены проблемы, выявленные при сравнении АРТ-1 с нейронной сетью Хэмминга. Также известны и другие проблемы, например отсутствие ассоциативной и распределенной памяти, но они не будут рассмотрены в рамках данной работы.

Постановка задачи

Целью данной работы является анализ описанных проблем нейронной сети АРТ-1 и предложение ряда решений для проблемы высокой вычислительной сложности. Планируется провести сравнение рассмотренных решений с исходной нейронной сетью и друг с другом теоретически и в ходе тестирования, а также выявить области их применимости.

Обзор литературы

В ходе работы был изучен ряд научно-методических книг и статей по теме нейронных сетей. Среди них следует выделить ряд книг, предназначенных для введения в теорию нейронных сетей: Уоссерман Ф. (1992) [5], Осовский С. (2002) [11], Круглов В. В. и Борисов В. В. (2002) [12], Хайкин С. (2008) [13]. Данные книги подробно описывают основы теории и многие статьи, посвященные нейронным сетям, ссылаются на них.

Основой данной работы является адаптивная резонансная теория. Она была предложена Гроссбергом С. и Карпентер Г. в 1987 году в ра-

боте [9]. Также на тему этой теории ими был написан ряд других работ в период с 1987 по 2015 года [7, 8, 14]. Данные работы легли в основу множества статей. Например решение проблемы деградации образов в АРТ-1 предложено в работах Выдриной Ю. В., Козынченко В. А. (2015) [15], Ращенко Ю. В., Козынченко В. А. (2015) [16]. Решение проблемы асимметрии нулей и единиц предложено в работах Ращенко Д. В. (2015) [17], Дмитриенко В. Д., Заковоротный А. Ю. (2012) [18]. Вариации меры сходства изображений рассмотрены в работах Ращенко Д. В., Козынченко В. А. (2015) [19], Антипин И. А., Козынченко В. А. (2012) [20]. Некоторые другие модификации нейронной сети АРТ-1 и варианты ее применения рассмотрены в работах Мищенко А. В. (2010) [21], Дмитриенко В. Д., Хавина И. П., Заковоротный А. Ю (2009) [22].

Целью данной работы является решение проблемы низкой производительности АРТ-1. Эта проблема затронута в некоторых из перечисленных работ [17, 19, 20], однако в них предлагаются решения, связанные с изменением меры сходства изображений, а следовательно изменением результатов распознавания. Такой подход может давать хорошие результаты (увеличение и производительности, и качества распознавания), однако затрудняет сравнение различных модификаций друг с другом. В данной работе предлагается подход, увеличивающий быстродействие АРТ-1 без изменения результатов распознавания.

По предварительным результатам работы опубликовано 2 статьи [17, 19] и представлены доклады на следующих конференциях: XLVI международная научная конференция аспирантов и студентов «Процессы управления и устойчивость» в 2015 году, III International Conference in memory of V. I. Zubov «Stability and Control Processes» в 2015 году, XLVII международная научная конференция аспирантов и студентов «Процессы управления и устойчивость» в 2016 году.

Глава 1

Архитектура и алгоритм распознавания искусственной нейронной сети АРТ-1

1.1 Архитектура

Архитектура нейронной сети АРТ-1 приведена на рисунке 1.1 и содержит следующие элементы: слой сравнения F1, группа слоев распознавания F2, блок сброса A и два вспомогательных блока gain control. Основные вычисления алгоритма распознавания выполняются в F1, F2 и A, а блоки gain control выполняют вспомогательную роль и поэтому в данной работе не будут рассмотрены подробно.

Следует отметить, что в весовых матрицах слоев F1 и F2 записаны эталонные образы — долговременная память (long term memory — LTM на схеме). При больших размерностях этих матриц работа F1 и F2 может потребовать большого количества ресурсов. В данном случае рационально использовать такое преимущество нейронных сетей как параллелизм вычислений. Блоки gain control и A не являются нейронными слоями, так

как не работают с матрицами больших размерностей, и их выход имеет размерность 1.

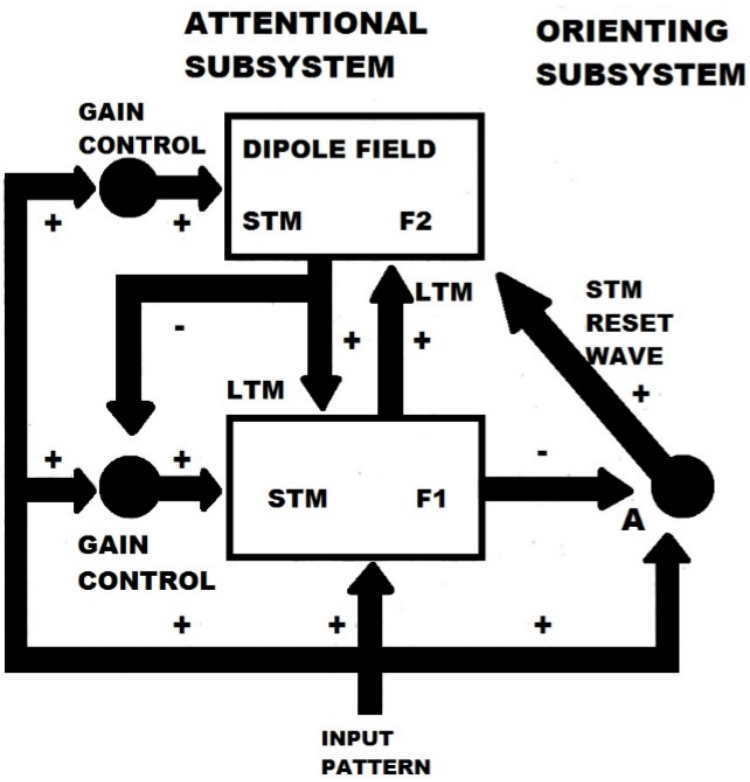

Рис. 1.1. Схема APT-1 [9].

Как было отмечено выше, данная нейронная сеть работает с бинарными векторами размерности $N = HW$, полученных при разложении изображения с высотой H и шириной W. В отличие от нейронной сети Хэмминга, один цвет пикселей (например черный) представляется как 1, другой как 0.

Также заметим, что преимуществом APT-1 является возможность обучения в процессе распознавания. До начала функционирования все M образов памяти не содержат в себе эталонов и имеют особый вид — состоят только из единиц (далее — пустой вектор). Значение такого подхода будет объяснено ниже. Здесь M — емкость памяти, то есть максимальное количество эталонов, которое сеть может запомнить.

Весовая матрица LTM группы слоев F2 имеет следующую структуру: строками этой матрицы являются запомненные эталоны, к которым применена процедура нормализации [11]: вычислено количество единиц S в нем, и каждый единичный элемент домножен на величину $\frac{1}{S+1}$. При умножении

описанной матрицы на входной вектор input pattern, для каждого эталона i вычисляется величина $\frac{C_i}{S_i+1}$, где C_i — количество совпавших пикселей (см. раздел о нейронной сети Хэмминга). Эта величина и является мерой сходства изображений в АРТ-1. Легко убедиться, что она не является метрикой в математическом смысле (не удовлетворяет второй аксиоме метрического пространства), однако такой вид метрики обеспечивает наличие преимущества пластичности. Действительно: так как для пустых образов $S_i = N$, то для них мера будет мала. Таким образом в первую очередь побеждают непустые образы, и только в случае, если нет эталонов с достаточным числом совпавших пикселей C_i выиграет пустой эталон, то есть выделится новая категория.

Также интересен процесс поиска максимальной компоненты вектора $(C_1, ..., C_M)^T$ в F2. Данный вектор многократно умножается на матрицу 1.1, где ε — малая случайная величина порядка $1/M$. При каждой итерации все компоненты вектора уменьшаются и, став отрицательными, обнуляются. Таким образом остается только одна ненулевая компонента, и такой вектор отправляется в слой F1.

$$\begin{pmatrix} 1 & -\varepsilon & \dots & -\varepsilon \\ -\varepsilon & 1 & \dots & -\varepsilon \\ & & \vdots & \\ -\varepsilon & -\varepsilon & \dots & 1 \end{pmatrix}_{M \times M} \tag{1.1}$$

Весовая матрица LTM слоя F1 в своих столбцах содержит векторы-эталоны. Из F2 приходит вектор со всеми нулевыми элементами, кроме одной (соответствующей победителю). Легко показать, что при умножении этих матрицы и вектора, получается вектор равный эталону-победителю.

1.2 Алгоритм распознавания

1. На вход предъявляется образ, подлежащий распознаванию input pattern. Обозначим этот вектор $\vec{X}$. Как видно из схемы, он принимается слоем F1 и блоком A.

(a) Блок A работает следующим образом: вычисляется количество единиц C вектора $\vec{P}$, пришедшего из F1. На данном этапе $C = N$ (предположим, что до начала процесса распознавания все элементы $\vec{P}$ равны 1); вычисляется количество единиц Q вектора $\vec{X}$; ищется отношение C/Q; если $C/Q \geqslant \rho$, то блок A выдает 0, иначе — 1. Параметр $\rho \in (0, 1)$ является минимальным порогом сходства, при котором сеть относит к одной категории запомненный эталон и $\vec{X}$. Демонстрация данного поведения приведена в пункте 4b. Так как $Q \leqslant N$, на данном этапе блок A выдает 0.

(b) Слой F1 работает в двух режимах: если из F2 приходит вектор, состоящий только из нулей, то выходом F1 является $\vec{P} = \vec{X}$. Второй режим рассмотрен в пункте 3.

2. Изменившийся выход F1 попадает в F2 и A.

(a) Так как на данном этапе $\vec{P} = \vec{X}$, то $C/Q = 1 > \rho$, и выход A не изменяется и равен нулю.

(b) При умножении матрицы LTM в F2 на вектор $\vec{P} = \vec{X}$ вычисляется близость $\vec{X}$ к каждому из эталонов. Затем итерационным методом латерального торможения находится эталон максимально похожий на $\vec{X}$. Полученный вектор $\vec{R} = (0, \ldots, 1, \ldots, 0)^T$, где единица стоит на месте победителя, передается в слой F1. Победителем может оказаться как один из ранее запомненных эталонов, так и пустой вектор.

3. Вектор $\vec{R}$ попадает в слой F1, который теперь работает во втором режиме: при умножении весовой матрицы LTM на вектор R из памяти извлекается эталон-победитель (или пустой вектор) $\vec{Y}$. Затем находится логическое пересечение векторов $\vec{X}$ и $\vec{Y}$, то есть $\vec{P} = (X_1 \cap Y_1, \ldots, X_N \cap Y_N)^T$. В случае, если победил пустой образ, $\vec{P} = \vec{X}$.

4. Изменившийся выход слоя F1 попадает в F2 и A.

(a) Так как вектор $\vec{P}$ теперь содержит только единицы, совпадающие с эталоном-победителем, то выход F2 не изменится.

(b) В блоке A возможны два сценария.

i. Если $C/Q \geqslant \rho$, то выход A не изменяется, следовательно выходы F2 и F1 также перестают меняться — АРТ-1 перешла в устойчивое состояние, что значит окончание распознавания. Алгоритм завершается, и происходит процесс обучения.

ii. Если $C/Q < \rho$, то это означает, что сходство победителя с $\vec{X}$ меньше допустимого порога, и выход A изменяется на единицу. Вспомним, что в F2 вычисляется $\frac{C}{S+1}$ — мера сходства, отличная от C/Q. Из-за этого различия может сложиться такая ситуация, когда сходство победившего образа меньше ρ, но в памяти есть непустой образ, для которого $C/Q \geqslant \rho$. Именно из-за этого факта нейронной сети требуется проверить и другие, проигравшие, образы из памяти на данный критерий, для чего предусмотрена фаза поиска, описанная ниже.

5. Изменившийся выход A попадает в F2, сбрасывая его в исходное состояние (на выходе вектор из нулей).

6. Изменившийся $\vec{R}$ попадает в F1, и на его выходе вновь $\vec{P} = \vec{X}$.

7. $\vec{P}$ попадает в A и F2.

(a) Пока из A приходит 1, F2 остается в сброшенном состоянии.

(b) Так как $C/Q = 1 > \rho$, выход A изменяется на 0.

8. Из A нулевой сигнал попадает в F2, где вновь начинается процесс поиска победителя (см. пункт 2b) с единственным отличием: предыдущий победивший эталон не участвует в конкуренции.

9. Алгоритм продолжается с пункта 3, в процессе чего последовательно проверяются все эталоны из памяти по критерию $C/Q \geqslant \rho$. Список победителей, которые больше не участвуют в конкуренции, хранится в кратковременной памяти слоя F2 (short term memory — STM на схеме). Процесс продолжается до тех пор, пока не будет найден эталон,

удовлетворяющий критерию, или все эталоны претендующие на победу закончатся, и выиграет пустой образ, для которого гарантировано $C/Q = 1 > \rho$, и алгоритм завершается на пункте 4bi.

После окончания алгоритма происходит процесс обучения: на место победившего эталона записывается его логическое пересечение с $\vec{X}$ (уже вычисленное в F1). Следует отметить, что если выиграл пустой образ, то в память записывается копия $\vec{X}$.

1.3 Возможные проблемы

1. *Производительность.* Как было отмечено, процесс распознавания (пункты 3–8) в некоторых случаях производится многократно, что, естественно, требует продолжительного времени. Особенно это заметно при выделении новой категории — перед этим требуется проверить почти все ранее запомненные эталоны. Корень этой проблемы кроется в различии критериев сходства в F2 и A, однако именно это различие и обеспечивает наличие пластичности. Для распознавания разных образов требуется различное количество итераций фазы поиска, а соответственно и различное время. В некоторых прикладных задачах при расчете необходимой производительности системы такая нестабильность может оказаться проблемой.

2. *Асимметрия нулей и единиц.* При вычислении меры сходства в F2 происходит умножение нормированной матрицы эталонов на X. При этом нулевые компоненты не изменяют степени сходства. Использование как единичных, так и нулевых компонент позволило бы повысить качество распознавания, однако, в этом случае, придется изменять многие другие пункты алгоритма. Действительно: в APT-1 обучение основано как раз на обнулении компонент запомненного образа.

3. *Деградация запомненных образов.* Как было отмечено, при обучении некоторые единичные компоненты запомненного вектора становятся

нулевыми, и обратный процесс не предусмотрен. При отнесении последовательности искаженных образов к одной категории в памяти АРТ-1 остается слишком мало пикселей для прохождения по критерию F2, и вероятность отнесения образов к данному классу существенно снижается (образ испорчен) [15]. Во многих задачах такая ситуация неприемлема.

4. *Мера сходства изображений не является метрикой в математическом смысле.* Мера сходства изображений в F2 не является метрикой из-за наличия знаменателя $S + 1$ (из-за которого не выполняется аксиома симметричности). Однако, он необходим, так как благодаря этому знаменателю непустые образы имеют преимущество перед пустыми.

Глава 2

Модификации для нейронной сети АРТ-1

В данной работе предлагаются две модификации искусственной нейронной сети АРТ-1, направленные на решение проблемы производительности.

2.1 Модификация, направленная на сокращение фазы поиска

Как было отмечено, проблема производительности возникает из-за того, что процесс распознавания многократно начинается с начала при генерации сигнала сброса. Данный процесс необходим для корректной работы алгоритма распознавания, но количество итераций фазы поиска может быть сокращено.

Рассмотрим пункт 2b алгоритма предыдущей главы. В F2 происходит итерационный процесс поиска максимальной компоненты вектора, и на каждой итерации количество ненулевых компонент уменьшается. После того, как остался только один такой элемент, данный вектор $\vec{R}$ передается в слой сравнения F1. Рассмотрим ситуацию, при которой на каждой итерации процесса латерального торможения результат $\vec{R}$ будет передаваться в

F1, то есть $\vec{R}$ содержит несколько единичных компонент, соответствующих претендентам на победу. Рассмотрим, как при этом изменится алгоритм.

3. Рассмотрим умножение матрицы LTM в слое F1 на описанный вектор $\vec{R}$: для каждой i-й единичной компоненты вектора $\vec{R}$ из памяти извлекается эталон $\vec{Y_i}$, претендующий на победу; находится вектор $\vec{Y} = \sum_i \vec{Y_i}$. Для вычисления логического пересечения с $\vec{X}$ вектор $\vec{Y}$ требуется преобразовать к логическому типу: все ненулевые компоненты становятся единицами. Теперь вектор $\vec{Y}$ является логическим ИЛИ векторов $\vec{Y_i}$, а $\vec{P} = \vec{X} \cap (\cup Y_i)$.

4. Вектор $\vec{P}$ такого вида попадает в F2 и A.

   (a) Так как вектор $\vec{P}$ содержит все компоненты, которые увеличивали сходство каждого из претендентов на победу в слое F2, то значение меры для них не изменится, и процесс латерального торможения можно продолжать.

   (b) Рассмотрим изменения работы блока A.

      i. Если $C/Q \geqslant \rho$ для вектора $\vec{P}$, то выход блока A не меняется и равен нулю. Однако, в отличие от оригинального алгоритма, пока F2 не закончит латеральное торможение, АРТ-1 не стабилизируется, и алгоритм снова вернется к следующей итерации пункта 2b. Если же найден окончательный победитель, то распознавание завершается как в исходном алгоритме.

      ii. Если $C/Q < \rho$, это означает, что среди претендентов на победу нет ни одного образа, подходящего по критерию A. Действительно: предположим, что такой образ есть среди претендентов (его номер обозначим w), и его логическое пересечение с X содержит C_w единиц, то есть $C_w/Q \geqslant \rho$, но при этом $C/Q < \rho$. Так как вектор $\vec{Y}$ состоит, как минимум, из всех единиц вектора $\vec{Y_w}$, то после взятия логического пересечения количество единиц в $\vec{P}$ для $\vec{Y}$ больше или равно, чем для $\vec{Y_w}$. Но это и есть C и C_w соответственно. Из $C \geqslant C_w$ и $C_w/Q \geqslant \rho$ следует $C/Q \geqslant \rho$, что

противоречит условию. Таким образов всех претендентов на победу можно сбросить, уменьшив этим количество итераций фазы поиска.

Остальная часть алгоритма не изменяется. Видно, что при использовании данной модификации количество итераций фазы поиска будет меньше или равно количеству итераций классического алгоритма.

2.2 Модификация, направленная на устранение фазы поиска

Описанная выше модификация предполагает сокращение фазы поиска без изменения архитектуры APT-1. Однако значительных результатов при решении проблемы производительности можно добиться, изменив порядок взаимодействия слоев.

Модифицированная схема APT-1 приведена на рисунке 2.1. В данном решении предлагается сначала определить, какие образы удовлетворяют критерию A, а затем среди них искать победителя по критерию F2.

В строки весовой матрицы LTM слоя F1 записаны эталонные образы. Таким образом, при умножении на входной вектор вычисляется $\vec{C} = (C_1, \ldots, C_M)$, где C_i — количество совпавших единиц i-го эталона и входного вектора input pattern.

Весовая матрица LTM слоя F2 является квадратной матрицей, на диагонали которой стоят элементы $\frac{1}{S_i+1}$, где S_i — количество ненулевых компонент эталона i. Остальные компоненты матрицы нулевые. Рассмотрим алгоритм распознавания предложенной сети.

1. На вход нейронной сети предъявляется вектор, подлежащий распознаванию input pattern. Обозначим его $\vec{X}$. Он попадает в F1 и A:

   (a) В A вычисляется Q — количество единиц в $\vec{X}$.

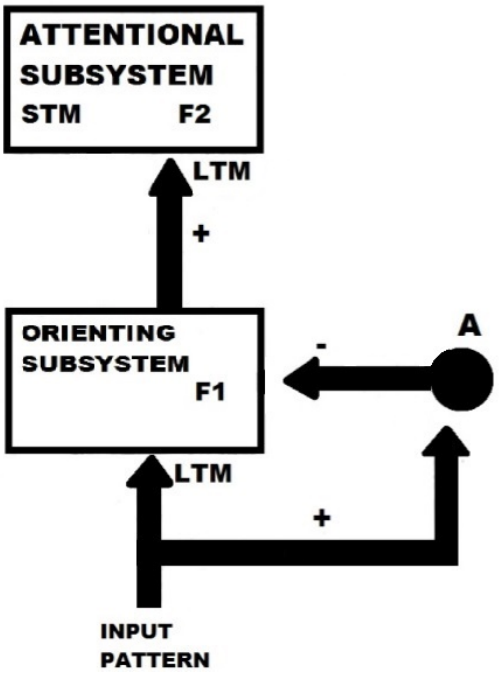

Рис. 2.1. Схема модифицированной APT-1.

(b) Полученный после умножения весовой матрицы LTM слоя F1 на $\vec{X}$ вектор $\vec{C}$ делится на число Q, пришедшее из блока A. Затем к полученному вектору применяется следующая активационная функция: если компонента $C_i/Q \geqslant \rho$, то i-я компонента результирующего вектора $\vec{P}$ равна C_i, в противном случае нулю. На данном этапе из Q приходит значение по умолчанию -1, и $\vec{P} = \vec{0}$.

2. Изменившийся Q попадает в F1, результатом работы которого становится вектор $\vec{P} = \vec{C}'$, где $\vec{C}'$ — вектор, равный $\vec{C}$, но с некоторыми обнуленными компонентами.

3. Вектор $\vec{P}$ такого вида попадает в F2, где при умножении на матрицу LTM вычисляется вектор с компонентами $\frac{C_i}{S_i+1}$, для которого выполняется процесс латерального торможения. Распознавание завершается, и происходит обучение по алгоритму классической APT-1.

В результате работы данного алгоритма определяется эталон, ближайший ко входному вектору по мере F2 и удовлетворяющий условию A. То есть результаты распознавания будут идентичны результатам класси-

ческой АРТ-1, но без использования фазы поиска.

2.3 Сравнение

Проведем сравнение данных модификаций с классической АРТ-1 и друг с другом и выявим преимущества и области применимости каждой из этих нейронных сетей. Обозначим модификацию с сокращенной фазой поиска как АРТ-1а, а модификацию без фазы поиска как АРТ-1б.

Рассмотрим примерное количество операций, которое нужно произвести в процессе алгоритма распознавания. Для АРТ-1 и АРТ-1а вычислительная сложность имеет порядок

$$l(2MN + kM^2),$$

где k — количество итераций латерального торможения, l — количество итераций фазы поиска. l для АРТ-1 больше или равно l для АРТ-1а. Вычислительная сложность АРТ-1б имеет порядок

$$MN + kM^2,$$

что соответствует всего одной итерации фазы поиска.

Очевидно, что вычислительная сложность АРТ-1б значительно ниже, чем АРТ-1. При распознавании больших последовательностей образов ожидается значительная разница во времени. Вычислительные сложности АРТ-1 и АРТ-1а отличаются множителем l и в некоторых случаях равны. Однако при запоминании новых образов, когда в АРТ-1 требуется сбросить большое количество эталонов, l для АРТ-1а может быть значительно меньше, чем для АРТ-1.

В связи с тем, что результаты распознавания для каждой из сетей не изменились, то все задачи, для которых применяется АРТ-1, могут решаться и ее модификациями. Это является интересной особенностью, так как в работах [17, 19] предлагается устранение фазы поиска за счет изменения

мер сходства. Данный подход хорошо работает в некоторых случаях, однако почти всегда можно подобрать пример, для которого классическая сеть справляется лучше. Это затрудняет сравнение качества распознавания, так как на разных тестовых множествах разные сети показывают лучший результат.

Следует отметить, что важным преимуществом АРТ-1а перед АРТ-1б является ее сходство с АРТ-1. Действительно: в системах с уже внедренной АРТ-1 ее замена на АРТ-1а легко осуществима, так как изменения минимальны. Также это преимущество позволяет интегрировать АРТ-1а в другие модификации АРТ-1 [15, 16, 18]. Внедрение же АРТ-1б потребует больших изменений.

Также следует отметить, что вычислительная сложность АРТ-1б не зависит от l. Это означает, что распознавание любого образа займет практически одинаковое время. Это позволяет рассчитать производительность системы распознавания таким образом, чтобы за один фиксированный промежуток времени распознавался один образ. Такой подход применяется при real time распознавании [10].

Рассчитаем ожидаемый выигрыш в производительности для АРТ-1б по сравнению с АРТ-1. Предположим, что для запоминания нового образа фаза поиска в АРТ-1 должна перебрать все ранее запомненные эталоны ($m+1$ итерация, где $m = 0, 1, \ldots M-1$ — количество эталонов на данном этапе распознавания), а при отнесении входного образа к ранее запомненному эталону она вызывается один раз (всего $K - M$ итераций). Здесь K — количество изображений, подлежащих распознаванию. Данные утверждения не точные, зависят от конкретного вида изображений и значения ρ, но для примерных расчетов могут быть использованы. Если в процессе распознавания K изображений было выделено M эталонов, количество итераций фазы поиска равняется

$$1 + 2 + \cdots + (M) + (K - M) = \frac{M+1}{2}M + (K - M) = \frac{M^2 - M}{2} + K.$$

При распознавании той же выборки АРТ-1б будет проведено K итераций

распознавания. Уменьшение вызовов фазы поиска в $\frac{M^2+M}{2K}+1$ раз.

Расчет ожидаемого выигрыша в производительности АРТ-1а по сравнению с АРТ-1 произвести сложнее, так как различие в константе l будет для каждого образа свое и зависеть от конкретного вида изображений и ρ.

Глава 3

Программная реализация и тестирование

3.1 Особенности реализации

Описанные нейронные сети реализованы в виде пакета программ на языке *C#* с использованием среды разработки Visual Studio Enterprise 2015 [23]. Разработанный пакет программ включает следующие компоненты: группа классов для нейронных сетей, два консольных приложения для автоматического тестирования, Windows Desktop интерфейс со встроенным графическим редактором для ручного тестирования.

3.2 Перспективы

Программы реализованы для демонстрации различий между модификациями нейронной сети АРТ-1, из-за чего был сделан упор на подробность реализации алгоритма, а не на его быстродействие. При необходимости решения задач, требовательных к быстродействию, возможно провести ряд оптимизаций:

- Использование параллельных вычислений. Как было отмечено во введении, параллельность вычислений широко поддерживается в нейрон-

ных сетях. Из-за большого количества возможных параллельных потоков особый интерес представляют технологии массового параллелизма с использованием GPU, такие как CUDA [24], OpenACC и OpenCL [25].

- При вычислении в один поток некоторые приемы нейронных сетей могут оказаться неэффективными. Например, алгоритмический поиск максимальной компоненты вектора является значительно более эффективным по сравнению с латеральным торможением. Но для корректности результатов тестирования было решено оставить алгоритм без изменений.

3.3 Изображения, используемые для тестирования

Для тестирования было решено использовать черно-белые изображения рукописных цифр из свободно распространяемой базы MNIST [26]. Данная база широко применяется исследователями нейронных сетей, что упрощает сравнение различных реализаций. MNIST состоит из 60000 изображений рукописных цифр, примеры которых приведены на рисунке 3.1. Так как целью данного тестирования было сравнить нейронные сети, тестирование на всех 60000 изображений было бы нецелесообразным из-за затрачиваемого времени. Поэтому было решено использовать только 300 экземпляров рукописных цифр.

Рис. 3.1. Примеры рукописных цифр из MNIST.

3.4 Критерии оценки качества нейронных сетей

Для оценки нейронных сетей будут применяться следующие критерии:

1. время распознавания;
2. количество категорий, выделенных сетью;
3. качество распознавания по принципу true positive. То есть для каждой категории будет определяться цифра, преобладающая в ней, A — количество таких цифр в данной категории, B — общее количество цифр в категории. Качеством распознавания будет усредненное по всем эталонам отношение A/B.

3.5 Результаты тестирования

Тестирование проводилось на процессоре с частотой 3.3 ГГц, использовались компилятор Visual Studio с оптимизацией и 64-разрядная операционная система. Для каждой сети $\rho = 0.5$.

Таблица 3.1. Результаты тестирования на 300 изображениях цифр из базы MNIST.

	АРТ-1	**АРТ-1а**	**АРТ-1б**
Время распознавания, с	503	231	4
Качество распознавания	0.74	0.72	0.72
Количество категорий	123	121	120

Выводы

Из таблицы видно, что качество распознавания и количество категорий, как и ожидалось, у всех сетей примерно равны. Незначительные

различия связаны со случайностью величины ε в матрице латерального торможения (см. главу 1). Как было замечено, одинаковые результаты распознавания гарантируют применимость модификаций во всех задачах, где применима классическая АРТ-1.

Подведем итог для модификации АРТ-1а с сокращенной фазой поиска. Теоретически доказано, что во всех случаях скорость работы АРТ-1а будет выше или равна скорости АРТ-1. Тестирование на множестве рукописных цифр показало прирост производительности в 2.18 раза. АРТ-1а является перспективной нейронной сетью, так как дает выигрыш в производительности с минимальными изменениями алгоритма распознавания АРТ-1 и без изменения архитектуры. Также данный факт обеспечивает легкую интеграцию АРТ-1а с другими модификациями АРТ-1 (например [15, 16]).

Подведем итог для модификации АРТ-1б с устраненной фазой поиска. Данная модификация дает значительное увеличение быстродействия сети. При этом время распознавания любого образа практически одинаково (небольшая разница остается за счет разного количества итераций процесса латерального торможения), в отличие от классической АРТ-1. Данные преимущества могут быть полезны в задаче real time распознавания. К недостаткам можно отнести значительно измененную архитектуру сети, что затрудняет внедрение модификации в системах, уже использующих АРТ-1, и интеграцию с другими модификациями.

Также выявлены причины возникновения других возможных проблем АРТ-1 и приведена литература, предлагающая их решение.

Заключение

В данной работе проанализированы проблемы, которые могут возникать в нейронной сети АРТ-1, и предложены две модификации для решения проблемы низкой скорости работы. Первая из модификаций обеспечивает увеличение производительности без изменения структуры сети. Вторая значительно изменяет структуру сети, но обеспечивает существенный

прирост скорости. Обе модификации и классическая АРТ-1 реализованы в виде пакета программ и проведено их сравнительное тестирование на 300 изображениях рукописных цифр. Тестирование подтвердило эффективность модификаций: увеличение производительности в 2.18 раза для первой модификации и в 125.75 раза для второй.

Литература

[1] Wilk A. Particle Identification Using Artificial Neural Networks with the ALICE Transition Radiation Detector: дис. Сигишоара, 2010. 266 p.

[2] Дмитриенко В. Д., Поворознюк О. А Дискретная нейронная сеть адаптивной резонансной теории для решения задач подбора лекарственных препаратов // Вестник Национального технического университета Харьковский политехнический институт. Серия: информатика и моделирование. 2009. № 13. С. 61–68.

[3] Deep Dream Generator URL: http://deepdreamgenerator.com/ (дата обращения: 01.05.2016).

[4] Доррер М. Г. Психологическая интуиция искусственных нейронных сетей: дис. ... канд. техн. наук: 05.13.16. Красноярск, 1998.

[5] Уоссермен Ф. Нейрокомпьютерная техника. М.: Мир, 1992. 184 с.

[6] Вапник В. Н., Червоненкис А. Я. Теория распознавания образов. М.: Наука, 1974. 416 с.

[7] Carpenter G. A., Grossberg S. ART 2: Self-organization of stable category recognition codes for analog input patterns // Applied Optics. 1987. №26(23). С. 4919–4930.

[8] Carpenter G. A., Grossberg S. ART 3: Hierarchical search using chemical transmitters in self-organizing pattern recognition architectures // Neural Networks (Publication). 1990. №3. С. 129–152.

[9] Grossberg S. Competitive learning: from interactive activation to adaptive resonance // Cognitive science. 1987. No 11. P. 23–63.

[10] Real-Time Human Pose Recognition in Parts from Single Depth Images // Microsoft Research URL: http://research.microsoft.com/pubs/145347/BodyPartRecognition.pdf (дата обращения: 02.05.2016).

[11] Осовский С. Нейронные сети для обработки информации. М.: Финансы и статистика, 2002. 344 с.

[12] Круглов В. В., Борисов В. В. Искусственные нейронные сети. М.: Горячая линия — Телеком, 2002. 382 с.

[13] Хайкин С. Нейронные сети: полный курс. 2-е изд. М.: Вильямс, 2008.

[14] Carpenter G., Grossberg S. Adaptive Resonance Theory // Encyclopedia of Machine Learning and Data Mining. Boston: Boston University, 2014.

[15] Выдрина Ю. В., Козынченко В. А. Модификация обучения искусственной нейронной сети АРТ-1 // Процессы управления и устойчивость. 2015. Т. 2. № 1. С. 379–384.

[16] Rashchenko Y. V., Kozynchenko V. A. New algorithms of an artificial neural network ART-1 training // IEEE 2015 International Conference «Stability and Control Processes» in memory of V. I. Zubov (SCP). 2015. P. 663–664.

[17] Rashchenko D. V. Elimination of the search phase in the neural network ART-1 by changing the criterion of vectors similarity // IEEE 2015 International Conference «Stability and Control Processes» in memory of V. I. Zubov (SCP). 2015. P. 661–662.

[18] Дмитриенко В. Д., Заковоротный А. Ю. Разработка нейронной сети АРТ с параметром сходства, симметричным относительно 0 и 1 входных векторов и позволяющем определять несколько решений // Математическое и программное обеспечение систем в промышленной и социальной сферах. 2012. №2. С. 24–35.

[19] Ращенко Д. В., Козынченко В. А. Разработка альтернативного критерия сходства изображений для нейронных сетей АРТ-1 // Процессы управления и устойчивость. 2015. Т. 2. № 1. С. 479–484.

[20] Антипин И. А., Козынченко В. А. Об одной модификации нейронной сети Хемминга // Процессы управления и устойчивость: Труды 42-й международной научной конференции аспирантов и студентов / под ред. А. С. Еремина, Н. В. Смирнова. СПб.: Издат. Дом С.-Петерб. гос. ун-та, 2011. С. 265–270.

[21] Мищенко А. В. Моделирование осознанного внимания в процессах обработки изображений человеческим мозгом на базе адаптивно-резонансных нейросетей // Вестник Санкт-Петербургского университета. Серия 10: Прикладная математика. Информатика. Процессы управления. 2010. № 4. С. 49–62.

[22] Дмитриенко В. Д., Хавина И. П., Заковоротный А. Ю. Новые архитектуры и алгоритмы обучения дискретных нейронных сетей адаптивной резонансной теории // Научные ведомости Белгородского государственного университета. Серия: история, политология, экономика, информатика. 2009. № 15–1. С. 88–96.

[23] Visual Studio Enterprise с MSDN URL: https://www.visualstudio.com/products/visual-studio-enterprise-vs (дата обращения: 03.05.2016).

[24] Параллельные вычисления CUDA URL: http://www.nvidia.ru/object/cuda-parallel-computing-ru.html (дата обращения: 03.05.2016).

[25] OpenCL URL: https://www.khronos.org/opencl/ (дата обращения: 03.05.2016).

[26] THE MNIST DATABASE of handwritten digits // Yann LeCun URL: http://yann.lecun.com/exdb/mnist/index.html (дата обращения: 16.03.2016).