МІНІСТЕРСТВО ОСВІТИ ТА НАУКИ УКРАЇНИ
Тернопільський національний економічний університет
Факультет комп'ютерних інформаційних технологій
Кафедра комп'ютерної інженерії

Куць Іван Святославович

Таблично-алгоритмічні засоби обчислення функцій активацій нейронних мереж / Table-algorithmic means for the activation functions calculating in neural networks

спеціальність:123 – Комп'ютерна інженерія
освітньо-професійна програма – Комп'ютерна інженерія

Випускна кваліфікаційна робота

Виконав студент групи КІм-21
І. С. Куць

_____

Науковий керівник:
д.т.н., професор, І.Г. Цмоць

_____

ТЕРНОПІЛЬ - 2019

РЕЗЮМЕ

Випускна кваліфікаційна робота на тему «Модель і програмно–апаратні засоби системи автоматичного розпізнавання автомобільних номерів для “розумного” міста» освітнього ступеня магістр зі спеціальності 123 «Комп’ютерна інженерія» написана обсягом 92 сторінок і містить 42 ілюстрацій, 7 таблиць, 3 додатки та 51 джерел за переліком посилань.

Метою випускної кваліфікаційної роботи є вибір принципів побудови, визначення та розробки реалізації сигмоїдальних функцій яка проводилася в середовищі розробки Quartus II для сімейства Cyclone III FPGA EP3C16F484C6 з використанням мови програмування VHDL та бібліотечних елементів Quartus II.

У роботі показано, що розробку компонентів для синтезу в режимі реального часу налаштованих GNM доцільно здійснювати при обробці інтенсивних потоків даних за допомогою складних алгоритмів паралелізацією та конвеєрністю обчислювальних процесів та використанням нових технологічних досягнень у галузі розробки суперінтегральних схем (VLSI).

Запропоновано ефективні компоненти переналаштованих технологій нейронної мережі будувати за такими принципами: співставленням інтенсивності потоку даних з обчислювальною потужністю; зменшенням кількість зовнішніх виходів і підвищенням продуктивності, за допомогою глибокої паралелізації до бітового рівня та використанням конвеєрних труб; скороченням часу та витрат на синтез нейронних мереж для конкретного застосування та розробкою алгоритмів розрахунку основних операцій штучних нейронних мереж на основі елементарних операцій, що дозволяє в повній мірі скористатися можливостями технології VLSI.

КЛЮЧОВІ СЛОВА: РОЗПІЗНАВАННЯ, СИГМОЇДАЛЬНІ ФУНКЦІЇ, СУПЕРІНТЕГРАЛЬНІ СХЕМИ.

RESUME

Master's thesis on the topic «Table-algorithmic means for the activation functions calculating in neural networks» from the specialty 123 «Computer engineering». Thesis contains 92 pages, 42 figures, 7 tables, 3 appendixes and 51 references sources.

The purpose of the final qualification work is to choose the principles of construction, definition and development of sigmoidal functions which was carried out in the development environment Quartus II for the family Cyclone III FPGA EP3C16F484C6 using VHDL programming language and library elements Quartus II.

It is shown that the development of components for real-time synthesis of configured GNMs should be carried out when processing intensive data streams using complex algorithms by parallelization and pipeline computing processes and using new technological advances in the development of superintegral circuits (VLSI).

It is proposed to build effective components of reconfigured neural network technologies according to the following principles: comparison of data flow intensity with computing power; reducing the number of external outputs and increasing productivity, through deep parallelization to the bit level and the use of pipelines; reducing the time and cost of neural network synthesis for a specific application and the development of algorithms for calculating the basic operations of artificial neural networks based on elementary operations, which allows you to take full advantage of VLSI technology.

KEYWORDS: RECOGNITION, SIGMOIDAL FUNCTIONS, SUPER INTEGRAL SCHEMES.

ЗМІСТ

ВСТУП

Актуальність теми. Сучасний етап розвитку технологій штучної нейронної мережі (SNM) характеризується розширенням додатків, в багатьох з яких необхідна обробка в складних алгоритмах реального часу різної інтенсивності даних на апаратному забезпеченні, що задовольняють обмеження в розмірах, споживанні енергії. Створення такого обладнання вимагає широкого використання сучасної елементної бази (напівпрограмних та користувацьких VLSI, одночипових мікропроцесорів) та розробки нових методів, алгоритмів та структур VLIS для реалізації нейроалгоритмів.

Аналіз задач, які вирішуються за допомогою SNM, методів та алгоритмів їх вирішення, виявив такі особливості: велика кількість обчислень з перевагою обчислювальних операцій над логічними; наполегливість та висока інтенсивність отримання даних; регулярність та рекурсивність алгоритмів; можливість паралелізації процесу обробки як у часі, так і в просторі.

Щоб вирішити завдання в режимі реального часу за допомогою SNM, продуктивність обладнання повинна бути високою. Режим роботи у режимі реального часу при обробці інтенсивних потоків даних за допомогою складних алгоритмів забезпечується паралелізацією та конвеєрністю обчислювальних процесів та використанням нових технологічних досягнень у галузі розробки суперінтегральних схем (VLSI). Тому актуальним завданням є розробка компонентів для синтезу в режимі реального часу налаштованих GNM.

Мета та завдання дослідження. Метою випускної кваліфікаційної роботи є вибір принципів побудови, визначення та розробки реалізації сигмоїдальних функцій яка проводилася в середовищі розробки Quartus II для сімейства Cyclone III FPGA EP3C16F484C6 з використанням мови програмування VHDL та бібліотечних елементів Quartus II.

Для досягнення цієї мети в роботі вирішуються такі завдання:

– огляд елементів нейронних мереж та її теорію;

– аналіз засоби обчислення функцій активацій нейронних мереж;

– засоби реалізації таблично-алгоритмічних засобів обчислення;

– розробка сигмоїдальних функцій активації на пліс для нейронних мереж;

– розробка функції активації нейронів сигмоїдального типу;

– реалізація сигмоїдальних функцій на FPGA.

Наукова новизна отриманих результатів. На основі проведених досліджень були розроблені ефективні компоненти переналаштованих технологій нейронної мережі. Отримано такі нові результати:

– удосконалено метод переходу від алгоритму до апаратної структури штучних нейронних мереж реального часу шляхом співставлення інтенсивності потоку даних з обчислювальною потужністю;

– удосконалений метод обчислення кількості парних продуктів у режимі реального часу шляхом формування та підсумовування продуктів макрочастинок, що зменшує кількість зовнішніх виходів та досягає високої продуктивності, що досягається глибокою паралелізацією до бітового рівня та використанням конвеєрних труб;

– вибираються принципи побудови компонентів нейронної мережі, які забезпечують скорочення часу та витрат на синтез нейронних мереж для конкретного застосування;

– розроблено алгоритми розрахунку основних операцій штучних нейронних мереж на основі елементарних операцій, що дозволило в повній мірі скористатися можливостями технології VLSI.

Практичне значення отриманих результатів.

– використання вдосконаленого методу переходу від алгоритму до структури апаратних компонентів штучних нейронних мереж дозволяє формалізувати процес проектування компонентів, що скорочує час і витрати на проектування;

– використання вдосконаленого методу розрахунку та основи елементарних операцій забезпечує розробку ефективних пристроїв для обчислення кількості парних продуктів, орієнтованих на реалізацію VLSI;

– використання зміни довжини ходу конвеєра, кількості та біт каналів даних дозволяє підвищити ефективність використання обладнання при розробці компонентів штучних нейронних мереж у режимі реального часу;

– комплексний підхід до розробки компонентів штучних нейронних мереж, який охоплює сучасні методи та алгоритми навчання та функціонування нейронних мереж, сучасну елементну базу, нові алгоритмічні архітектурні орієнтовані NVIS та схематичні рішення дозволяє оптимізувати апаратно-часові параметри.

Публікації та апробація випускної кваліфікаційної роботи. Отримані результати апробовані в межах науково-практичної конференції молодих вчених і студентів «інтелектуальні комп’ютерні системи та мережі» та опубліковано дві тези доповіді по темі роботи[3, 4].

Впровадження результатів ВКР. Результати роботи планується використати в роботі науково-дослідному інституті інтелектуальних комп’ютерних систем (додаток Б).

Випускна кваліфікаційна робота складається із трьох розділів, висновків, списку використаних джерел та додатків.

У першому розділі було наведена інформація про елементи теорії нейронних мереж.

В другому розділі проаналізовано види функції активації нейронних мереж та їхнє використання.

У третьому розділі здійснено програмну реалізацію розроблених алгоритмів сигмоїдальних функцій активації для нейронних мереж та їхня практична реалізація.

1 ЕЛЕМЕНТИ ТЕОРІЇ НЕЙРОННИХ МЕРЕЖ

1.1 Основні поняття нейронів

Можна вважати, що вивчення людського мозку почалося в ті часи, коли людей почало цікавити їх власне мислення. Роздуми про самого себе є, відмінною ознакою роботи мозку людини. Існує безліч роздумів про природу мислення, що тягнуться від духовних до анатомічних. Обговорення цього питання, що протікало в гарячих суперечках філософів і теологів фізіологами і анатомами, принесло мало користі через труднощі вивчення предмета. Ті, хто спирався на самоаналіз і роздум, дійшли висновку, що не відповідає рівню суворості точних наук. Експериментатори ж знайшли, що мозок важкий для спостереження і ставить в глухий кут своєю організацією. Говорячи іншими словами, потужні методи наукового дослідження, що змінили наш погляд на реальний світ, виявилися безсилими в розумінні самої людини.

Мозок людини є найскладнішою з відомих на сьогоднішній день систем обробки інформації і складається, приблизно із 100 мільярдів (10^{11}) пов’язаних між собою нейронів. Кожний з них має в середньому 10 тисяч (10^{4}) зв’язків, що породжує 10^{15} взаємозв’язків. При цьому мозок людини надзвичайно надійний, незважаючи на те, що кожного дня величезна кількість нейронів гине, але мозок надалі продовжує функціонувати. Опрацювання великого обсягу інформації здійснюються мозком надзвичайно швидко, за лічені секунди. З урахуванням того, що саме нейрон є повільно діючим елементом, процес реакції відбувається не більше декількох мілісекунд. На даний момент неясно, як мозок може отримувати таке вражаюче поєднання швидкодії та надійності.

Сучасною наукою досить добре вивчені структура та функції деяких нейронів, а також існують дані про організацію внутрішніх і зовнішніх зв'язків між нейронами, деяких структурних утворень мозку, і зовсім мало відомо про участь різних структур в процесах переробки інформації.

Кожен з нейронів має тіло нейрона – сома, безліч вхідних зв'язків - дендрити, єдиний вихідний зв'язок - аксон, яка на кінці також розгалужується, і контакти – синапси для утворення зв'язків аксонів з дендритами решти нейронів. Структура природного нейрона показана на рисунку 1.1.

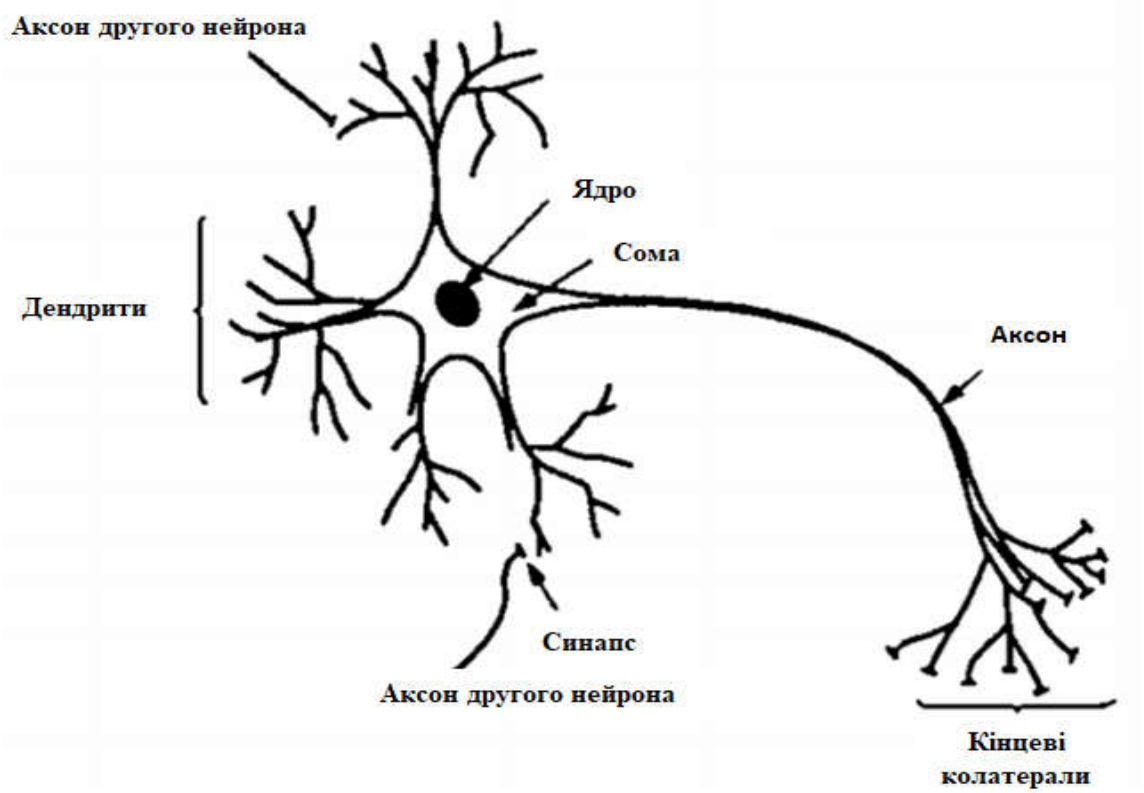


Рисунок 1.1 – Структура природного нейрона.

Вхідну інформацію нейрон отримує через свої дендрити від аксонів інших нейронів. Дендрити сильно розгалужуються, охоплюючи великий простір навколо нейрона, в порівнянні з їх величиною. Довжина деяких дендритів може досягати 1 мм.

Отримана інформація перетворюється сомою, щоб згенерувати вихідну інформацію. Поперечний розмір соми всього кілька десятків мікрон. Вихідна інформація передається іншим нейронам через аксон, що розгалужується в кінці на різні синапси. Аксон може мати довжину кількох сотень міліметрів. Закінчення аксонів, що йдуть від інших нервових клітин, розташовуються на сомі і на дендритах. Причому їхні закінчення мають вигляд потовщення,

названого синапсом, або синаптичною бляшкою. Поперечні розміри синапсу зазвичай не перевищують кількох мікрон. Досить часто ці розміри становлять близько 1 мкм. Синапс є функціональним вузлом та елементарною структурою між двома нейронами. Передача через синапс майже завжди однонаправлена. Розрізняють пресинаптичні та постсинаптичні клітини - у напрямку передачі імпульсу.

Коли імпульс досягає кінця синапсу, вивільняються хімічні речовини, які називаються нейротрансмітерами. Нейротрансмітери поширюються через синаптичну щілину, стимулюючи або пригнічуючи, в залежності від типу синапсу, здатність нейрон-приймача генерувати електричні імпульси.

Окремий нейрон не є елементарною одиницею обробки інформації, а виконує функції нервового центру. Дендрити і аксони можуть вступати в зв'язку з ділянками мембран інших нейронів, утворюючи мережі. Ці мережі і служать системами обробки інформації.

Алгоритм роботи біологічного (природного) нейрона полягає в наступному. Проходячи через синапс, електричний сигнал змінює свою амплітуду: збільшує або зменшує. Це можна інтерпретувати як множення амплітуди сигналу на ваговий (синаптичний) коефіцієнт. Зважені в дендритні дерева вхідні сигнали підсумовуються в тілі клітини, і потім на аксони виході генерується вихідний імпульс (спайк) або пачка імпульсів. Сигнал, який є вихідним, проходить по гілці аксона та досягає синапсів, що з'єднують аксон із дендритними деревами інших нейронів. Сигнал трансформується через синапси у новий вхідний сигнал для суміжних нейронів. Такий сигнал може бути позитивним або негативним (збудливим або гальмують) в залежності від виду синапсу. Величина сигналу, який генерується на виході синапсу, визначається синаптичним коефіцієнтом (вагою синапсу), який часто змінюються в процесі функціонування синапсу.

Усі нейрони. як правило, поділяють на три великі групи: ефекторні, рецепторні і проміжні. Введення в мозок сенсорної інформації забезпечують рецепторні нейрони. Вони здатні трансформувати сигнали, які надходять до

органів чуття (такі як оптичні сигнали у сітківці ока, акустичні у вушній раковині або нюхові в хеморецепторах носа), в послідовність електричних імпульсів своїх аксонів. Сигнали, які отримують ефекторні нейрони передаються виконавчим органам. У їхніх аксонах існують особливі синаптичні зв’язки із виконавськими органами, такими наприклад як м’язи, там збудження нейронів викликає скорочення м’язів. Проміжні нейрони обробляють інформацію, отриману від рецепторів, і генерують керуючі сигнали для ефекторів. Центральну нервову систему вони і утворюють.

Вхідні сигнали дендритного дерева зважуються і підсумовуються в сомі, причому якщо результат не перевищує деякого порога, то вихідний сигнал не формується зовсім - нейрон «не спрацьовує». Вихідний сигнал проходить через гілки аксонів і досягає синапсів, які зв'язують аксони з дендритними деревами інших нейронів. Сигнал перетворюється через синапси в новий вхідний сигнал для сусідніх нейронів. Цей вхідний сигнал може бути як позитивним так і негативним, то є збудливим або гальмуючим в залежності від типу синапси. Величина вхідного сигналу, який генерується синапсом, може відрізнятися навіть при однаковій величині сигналу, що відступає в синапс. Ці відмінності і визначають ефективність або вагу синапсу. Слід зауважити, що вага синапсу може змінитись під час функціонування синапсу. Більшість вчених вважають таку зміну нейрофізіологічним корелятом, тобто слідом пам'яті.

Роль механізму молекулярної пам'яті в такому випадку полягає в довготривалому засвоєнні таких слідів.

Функціонально робота природного нейрона описується так:

– прийняті від аксонів інших нейронів вхідні сигнали проходять через синапси і змінюються пропорційно вазі синапсів;

– ті що надійшли до сома одночасно за кількома дендритам змінені вхідні сигнали підсумовуються;

– нейрон збуджується, якщо сумарний імпульс стає більшим за деяку межу, і створює вихідний сигнал, який відводиться аксоном з соми та розгалужується;

– гілки вихідного сигналу передаються через синапси до дендритів інших нейронів.

При цьому вага кожної з синаптичних зв'язків може змінюватися з часом, отже, може змінюватися і поведінка нейрона.

В цілому природні (біологічні) нейронні мережі мають такі властивості:

– розподілене подання та обчислення інформації;

– паралельність обробки інформації.

– змінюються за вагою зв'язків між нейронами;

– здібності до навчання та узагальнення;

– адаптивність;

– можливість контекстуальної обробки інформації;

– толерантне ставлення до помилок.

1.2 Природні нейронні мережі

Нейрони, з'єднуючись між собою через синапси, утворюють єдину складну нейронну мережу. Структура природної нейронної мережі представлена на рисунку 1.2.

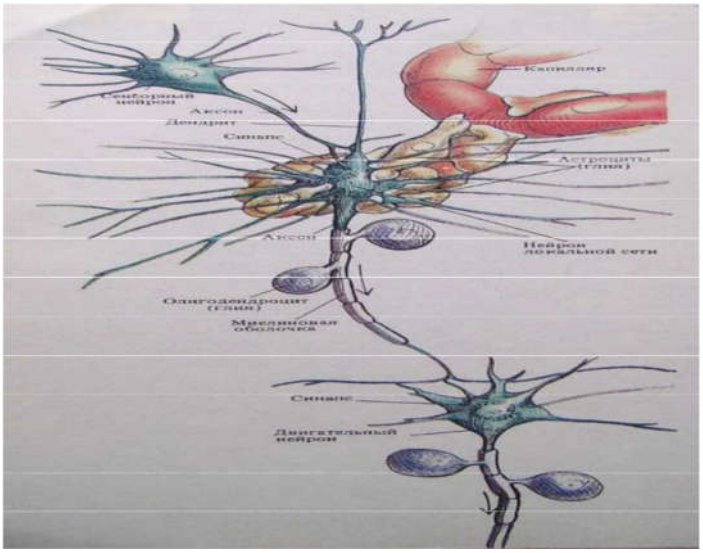


Рисунок 1.2 – Структура природної нейронної мережі.

Нейрони в живих організмах мають особливий вид клітин, які мають електричну активність, головна здатність якої полягає в оперативному управлінні організмом.

Нейрони в нейронної мережі взаємодіють за допомогою серій електрохімічних імпульсів, які з'являються в результаті деякого нейрофізіологічного процесу в мозку. Кожний імпульс представляє собою частотний сигнал, де частота змінюється від декількох одиниць до сотень герц. Якщо брати до уваги той факт, що довільний нейрофізіологічний процес активізує одночасно безліч зв'язаних один з одним нейронів, тоді можна уявити всю ту кількість інформації або сигналів, яка виникає в мозку людини. Картина електрохімічних імпульсів в нейронній мережі показана на рисунку 1.3.

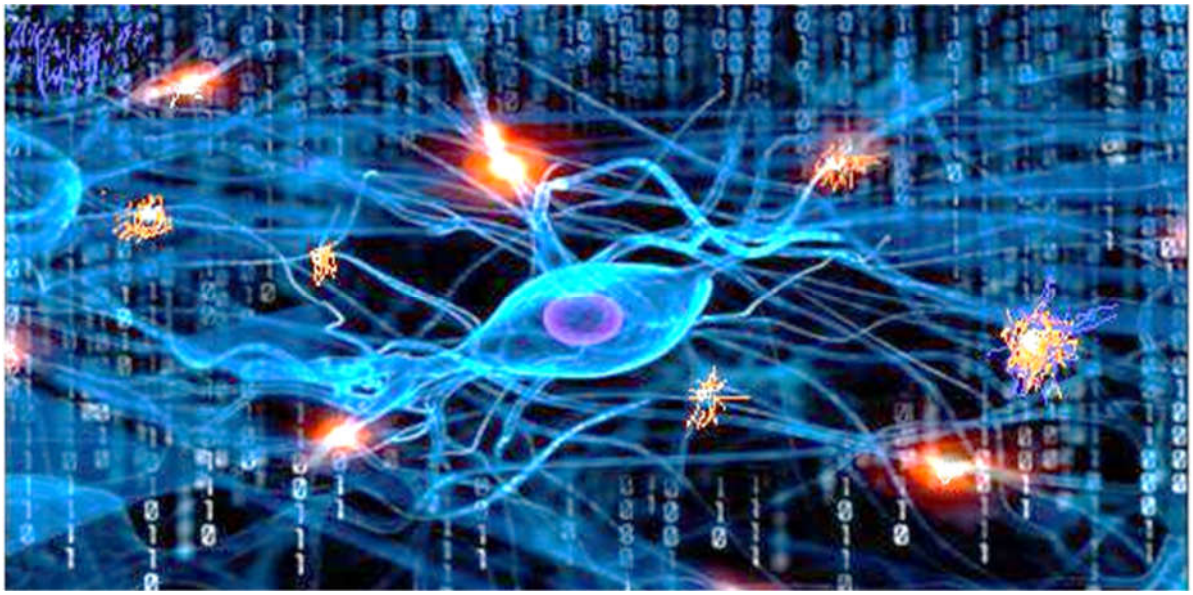


Рисунок 1.3 – Електрохімічні імпульси в нейронної мережі.

Як вже згадувалось раніше, електрохімічні імпульси можуть тривати лише кілька мілісекунд. Кожен з них представляє частотний сигнал із частотою від декількох одиниць до сотень герц. В порівнянні з сучасними можливостями ЕОМ це неймовірно повільно, але в той же час мозок людини здатний набагато швидше комп’ютера здійснювати обробку аналогової інформації. Наприклад: розпізнавати зображення, відчувати смак, відчувати звуки, читати текст, написаний чужим почерком, оперувати якісними параметрами. Тому що все це

реалізовується мережею нейронів, які з'єднані між собою синапсами. Інакше кажучи, мозок - це складна система, яка складається із паралельних процесорів, що працює більш ефективно, ніж послідовні обчислення.

Сотні мільярдів нейронів, кожен з яких з'єднаний з сотнями або тисячами інших, утворюють систему, яка набагато перевершує найпотужніші суперкомп'ютери. Крім того, зазвичай, мозок людини задіяний тільки на 2-3% від своєї можливості. Завдяки такій надмірності, мозок людини володіє величезним запасом міцності, що дозволяє йому працювати, незважаючи на серйозні пошкодження і втрати. Подібної здатності позбавлені сучасні комп'ютери.

Слід зазначити, що в сучасній комп'ютерній техніці технологія послідовних обчислень знаходиться на межі своїх технічних можливостей. В теперішній час актуальною є проблема розвитку як методів паралельного програмування так і створення паралельних комп'ютерів. Тому саме нейронні мережі є наступним кроком розв’язання цієї проблеми..

Однак відмінність комп'ютерів (моделювання нейронних мереж) від природних нейронних мереж є в тому, що моделюються мережі строго впорядковані і їх елементи мають фіксовану кількість зв’язків. В той час як природні мережі – хаотичні, динамічні як за структурою, так і за кількістю зв'язків між нейронами. Природні нейронні мережі самоорганізуються, що робить їх значно ефективніше ніж модельований, крім того їм не потрібно задавати цілепокладання, останнім у них визначається гомеостазом клітини і всієї колонії.

1.3 Історія штучних нейронів і нейронних мереж

У 1943 році нейрофізіолог У.Мак-Каллок і математик У.Піттс в статті «Логічне числення ідей, що відносяться до нервової активності» ввели поняття

штучної нейронної мережі і запропонували формальну модель штучного нейрона. У 1948 році Н. Вінер разом з соратниками опублікував роботу про кібернетику, в якій запропонували ідею про подання складних біологічних процесів математичними моделями. У 1949 році Д. Хебб в роботі «Організація поведінки» описав основні принципи і перший алгоритм навчання нейронів.

У 1950-ті рр. вперше були створені програмні моделі штучних нейронних мереж. Перші роботи проведені Н. Рочестером з дослідницької лабораторії 1ВМ. Дана модель зазнала невдачі, тому що надзвичайно швидке зростання традиційних електронних обчислень залишило в тіні нейронні дослідження. Однак пізніші реалізації цих моделей були успішними.

У 1956 році в Дартмутському дослідному інституті штучного інтелекту було створено хороші умови для дослідницької діяльності в області штучного інтелекту вцілому та нейронних мереж зокрема. Дослідницька робота в області штучного інтелекту поділилася на два напрямки: моделювання мозку та промислове застосування систем штучного інтелекту (експертні системи).

У 1957 році Ф. Розенблатт запропонував комп'ютерну модель мозку, яка називається персептрон (від латинського регсербо - сприйняття), і в тому ж році в Корнельській Лабораторії Аеронавтики успішно було завершено моделювання роботи персептрона на комп'ютері 1ВМ 704. Персептрон здійснював передачу сигналів від фотоелементів, що створювали сенсорне поле, у блоки електромеханічних елементів пам'яті. Ці осередки поєднувалися між собою випадково, відповідно до принципів коннективізму.

У 1958 році Ф. Розенблатт опублікував статтю «Персептрон: Ймовірна модель зберігання і організації інформації в головному мозку», де вводить одношаровий персептрон і демонструє його здатність вирішувати завдання класифікації. Персептрон здобув популярність і його почали використовувати для розпізнавання образів, прогнозування погоди. У цьому ж році Джон фон Нейман запропонував імітацію простих функцій нейронів з використанням вакуумних трубок.

У 1959-60 роки Б. Уїдроу і М. Хофф розробили моделі ADALINE і MADALINE (Множинні Адаптивні Лінійні Елементи). ADALINE (адаптивний суматор) спочатку використовувався для задач передбачення і адаптивного управління, потім він став стандартним елементом багатьох систем обробки сигналів. MADALINE діяв, як адаптивний фільтр, що усуває відлуння на телефонних лініях, і до цих пір знаходиться в комерційному використанні.

У 1960 році в Корнельському університеті був продемонстрований перший нейрокомп'ютер - «Марк-1», який міг розпізнавати деякі букви англійського алфавіту.

У 1962 році Ф. Розенблатт докладно описав свої теорії і припущення щодо процесів сприйняття і персептронів в книзі «Принципи нейродинаміки: Персептрони і теорія механізмів мозку». Він розглядає вже не тільки готові моделі персептрона з одним прихованим шаром, але і багатошарові персептрони з перехресними і зворотними зв'язками, доводить теорему збіжності персептрона. Тоді одношаровий персептрон, реалізований апаратно, вважався класичною нейронною мережею і використовувався для класифікації вхідних сигналів в один з двох класів. Вперше за допомогою персептрона були вирішені завдання класифікації і розпізнавання букв англійської мови. Саме тому був створений спеціальний ітераційний метод навчання на підставі спроб та помилок, який імітував процес навчання людини методом корекції помилки. Зокрема, при розпізнанні певних літер персептрон мав змогу розрізніти характерні особливості букв, які статистично частіше зустрічаються, в порівнянні з незначними відмінностями в окремих випадках. Тим самим персептрон був здатний робити узагальнення літер, які були написані по-різному (почерком), в один узагальнений образ.

У 1963 році в Інституті проблем передачі інформації АН СРСР А. П. Петровим проводиться докладне дослідження завдань, «важких» для персептрона. Ця піонерська робота в області моделювання штучних нейронних мереж (ІНС) в СРСР послужила відправною точкою для комплексу ідей М. М.

Бонгард - «як порівняно невеликою переробкою персептрона можна виправити його недоліки».

Надмірне очікування від нейронних мереж, яке панувало в академічному та технічному світі, "заразило" і загальну літературу цього часу. Передчуття того, що ефект "думаючої машини", залишить відбиток на людині, широко підігрівалося письменниками. Однак, можливості персептрона були вельми обмеженими. Техніка не могла гарантувати правильність розпізнавання частково закритих літер, а також букв іншого розміру, які розташовані зі зміщенням або трохи повернуті, ніж ті, які використовувалися при її навчанні.

У 1969 році М. Мінський і С. Пейперт довели, що персептрони теоретично нездатні вирішити простіші завдання, в тому числі реалізувати бінарну функцію «Що виключає Або - exclusive OR (XOR)», проблему «парності» або «один в блоці», пов'язані з інваріантністю уявлень. Даний факт і побоювання, в поєднані із невиконаними обіцянками, спричинили багато фактів розчарування фахівців. Це спричинило критику в дослідженні нейронних мереж. В результаті інтерес до розробок нейронних мереж пішов на спад і було припинено фінансування.

У 1972 році Т. Кохонен і Дж. Андерсен незалежно пропонують новий тип нейронних мереж, здатних функціонувати в якості пам'яті.

У 1973 році Б. В. Хакимов пропонує нелінійну модель з синапсами на основі сплайнів і впроваджує її для вирішення завдань в медицині, геології, екології.

У 1974 році Пол Дж. Вербос і А. І. Галушкин одночасно винаходять алгоритм зворотного поширення помилки для навчання багатошарових персептронів.

У 1975 році Фукусіма представляє когнітрон, що є самоорганізованою мережею, призначеної для інваріантного розпізнавання образів. Тим не менш, це було досягнуто тільки за допомогою запам'ятовування практично всіх станів образу.

У 1982 році кілька подій призвело до відродження інтересу в дослідженні нейронних мереж. Джон Хопфілд у національній Академії Наук США

представив статтю на дану тему. Він показав, що нейронна мережа із зворотними зв'язками може створювати систему, що здатна мінімізувати енергію. Тобто він показав нові можливості моделювання нейронних мереж на принципі архітектури, званої мережею Хопфілда. По-друге в м. Кіото (Японія) відбулася Об'єднана американо-японська конференція по нейронних мережах, на якій мережі були представлені нові досягнення п'ятої генерації. Американські засоби масової інформації та наукові видання підхопили цю історію, акцентуючи на можливості відставання США в цій галузі, що призвело до зростання фінансування дослідження нейромереж. Потім Кохоненом була представлена модель мережі, названою нейронною мережею Кохонена (самоорганізована карта Кохонена), яка навчається без учителя і вирішує завдання кластеризації, візуалізації даних і інші завдання аналізу даних.

З 1985 року Американський Інститут Фізики почав щорічні зустрічі - "Нейронні мережі для обчислень".

У 1986 році Д. Румельхартом, Дж. Хінтон і Р. Вільямсом, а також незалежно і одночасно С. І. Барцевим і В. А. Охоніним, був перевідкрито і суттєво розвинуто метод зворотного поширення помилки. Почався вибух інтересу до дослідження нейронних мереж.

У 1989 році на зустрічі "Нейронні мережі для оборони" Б. Уїдроу повідомив аудиторії про початок четвертої світової війни, де полем бою є світові ринки і виробництва.

У 2007 році Джеффрі Хінтон в університеті Торонто створені алгоритми глибокого навчання багатошарових нейронних мереж. Успіх обумовлений тим, що Хінтон при навчанні нижніх шарів мережі використовував обмежену машину Больцмана (RBM - Restricted Boltzmann Machine). Глибоке навчання по Хінтон - це дуже повільний процес. Необхідно було використовувати багато прикладів розпізнавання образів (наприклад, безліч осіб людей на різних фонах). Після навчання виходить готове швидко працює додаток, здатне вирішувати конкретну задачу, наприклад, здійснювати пошук осіб на зображенні. Функція пошуку осіб людей на сьогоднішній день стала стандартною і вбудована в усі

сучасні цифрові фотоапарати. Технологія глибокого навчання активно використовують інтернет-пошуковими системами при класифікації картинок що містяться в них образи. Штучні нейронні мережі, що застосовуються при розпізнаванні, можуть мати до 9 шарів нейронів. Їх навчання ведеться на мільйонах зображень щоб відшукувати таким чином.

На даний час процес обговорення нейронних мереж відбувається всюди. Перспектива їх використання здається досить яскравою, в світлі рішення нетрадиційних проблем та є ключем, який відкриває нові технології. Тепер переважна більшість розробок нейронних мереж є працюючими, але існують певні процесорні обмеження. Сучасні дослідження спрямовані як на програмну так і апаратну реалізацію нейромереж. Більшість компаній працюють над створенням трьох типів нейронних чипів: цифрових, аналогових і оптичних. Саме вони й обіцяють бути трендом близького майбутнього.

В цілому розуміння суті функціонування нейрона та дослідження його зв'язків дозволяє дослідникам створювати математичні моделі, які дають можливість перевірки їх теорій. Тепер експерименти проводяться на потужних цифрових комп'ютерах і дозволяють не залучати в процес дослідження людей чи тварин. Це дає змогу вирішити як практичні так і морально-етичні проблеми. В перших же проектах з'ясувалося, що дані моделі не тільки повторюють функції мозку, але й виконувати функції, що мають індивідуальну цінність. Саме тому виникли і залишаються актуальними дві взаємно збагачуючі мети нейронного моделювання:

з’ясувати особливості функціонування нервової системи людини на фізіологічному та психологічному рівні;

створити обчислювальні системи (штучні нейронні мережі), які будуть здатні виконувати функції, подібні до функцій мозку.

1.4 Формальна модель штучного нейрона

Як випливає з попереднього підпункту, першу модель штучного нейрона на електричних схемах, що імітує роботу природного (біологічного) нейрона, запропонували вчені Уоррен Мак-Каллок (Warren McCulloch) і Уолтер Піттс (Walter Pitts).

Штучний нейрон являє собою абстрактний комп'ютер, що складається з трьох блоків (помножувач, суматор, перетворювач), і дозволять виробляти переробку вхідних сигналів на вихідні сигнали з урахуванням значень початкового стану, відповідних синаптичних зв'язків і функцій активації.

У загальному випадку штучний нейрон має $n \geq 1$ входів і до синапсах цих входів надходить вектор вхідних сигналів $X = (x_1, x_2, \dots, x_n)$ після проходження синапсів сигнали змінюються пропорційно вектору осів синапсів $W = (w_1, w_2, \dots, w_n)$, виходить $W \cdot X = w_1 \cdot x_1, w_2 \cdot x_2, \dots, w_n \cdot x_n$. Змінені вхідні сигнали через дендрити надходять до сома, де виходить сумарний імпульс $s = w_1 \cdot x_1 + w_2 \cdot x_2 + \cdots + w_n \cdot x_n$, званий рівнем активації. Активацію викликає певна функція f, яка називається функцією активації. Після дії функції активації рівень активації порівнюється з деяким граничним значенням w_0. Якщо програма активації перевищує цей поріг, то нейрон збуджується і формує вихідний сигнал y.

У загальному випадку значення вхідного сигналу, осі і зміщення є дійсними числами. Значення вихідного сигналу визначається видом функції активації і може бути, як дійсним, так і цілим числом.

Структурна схема штучного нейрона Мак-Каллока-Піттса показана на рисунку 1.4.

Штучний нейрон Мак-Каллока і Питтса діє за наступним алгоритмом:

Перед початком роботи на блок суматора подається значення сигналу початкового стану (зміщення - bias) w_0;

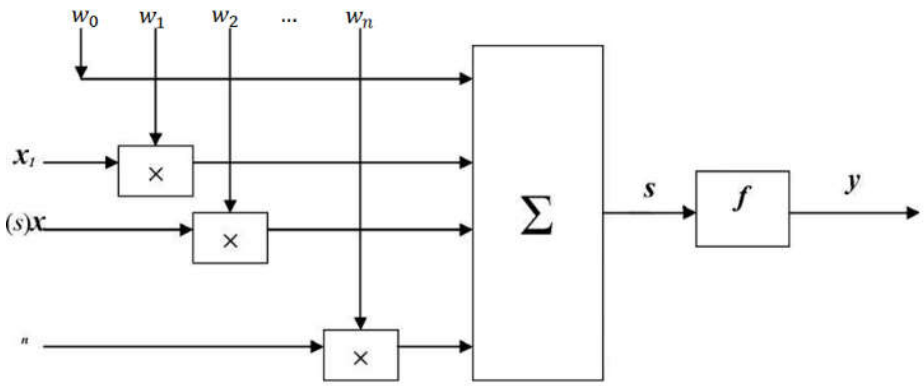


Рисунок 1.4 – Схема штучного нейрона Мак-Каллока-Піттса.

В блоці помножувача значення кожного вхідного сигналу x множиться на відповідне значення осі синапсу $w_i, i = 1, 2, \dots, n$;

В суматорі всі значення зважених вхідних сигналів (створення значень вхідних сигналів і осей) і значення сигналу початкового стану підсумовуються, визначаючи рівень активації нейрона $\sum_{i=0}^{n} w_i x_i + w_0 = W \cdot X + w_0 = s$;

Виважена сума (результат підсумовування) s подається на блок функціонального перетворювача f де виробляється значення вихідного сигналу з урахуванням заданого порогового значення $y = f(s)$, якщо є функція активації, в іншому випадку вихідний сигнал буде дорівнює зваженій сумі s, отриманої на попередньому кроці.

Структура, представлена на рисунку 1.4. є стандартною моделлю штучного нейрона. Більшість математичних моделей створені відповідно до принципів функціонування біологічних нейронів і ґрунтуються на стандартній моделі Мак-Каллок-Піттса, наприклад, персептрон, сигмоїдальний нейрон.

Пізніше розробники штучних нейронів запропонували кілька нестандартних детермінованих моделей штучного нейрона (паде-нейрон, нейрон з квадратичним суматором, адаптивний лінійному нейрону, сигма-пі нейрон, WTA, Хебба, Гроссберга, нейрон з лічильником збігів), а також стохастичну (вірогідну) модель нейрона.

Крім того, зауважимо, що якщо значення вхідного сигналу, осі синапсу і зміщення є:

довільними реальними числами і до них застосовуються дійсна арифметика, то говоримо про реальними нейрони;

логічними значеннями 0 або 1 то до них застосовуються логічні операції, то мова йде про логічні нейрони;

числами в інтервалі $[0, 1]$ то до них застосовуються операція нечіткої (fuzzy) логіки, то маємо справу з нечіткими нейронами.

1.5 Класифікація штучних нейронів

Штучні нейрони класифікують в залежності від їх положення в мережі, від типів оперованих даних, від способів опису моделі нейрона і від способів їх апаратної реалізації.

Штучні нейрони в залежності від їх положення (ролі) в мережі поділяються на вхідні нейрони, проміжні нейрони і вихідні нейрони.

Вхідні (рецепторні) нейрони приймають вхідний вектор, який кодує вихідний сигнал і не виконує операцій над ним, а просто передають закодований вихідний сигнал на вихід, можливо, посиливши або послабивши його.

Вихідні (ефекторні) нейрони являють вихідні сигнали мережі. У вихідних нейронах можуть проводитися будь-які операції над вихідними сигналами.

Проміжні нейрони виконують основні операції над вихідним і проміжним сигналом.

Штучні нейрони по типу оброблюваних даних підрозділяються на числові нейрони і логічні нейрони.

Числові нейрони оперують з числовими даними, представленими у формі дійсних чисел. Допускається використання цілих чисел при вирішенні деяких завдань розпізнавання образів.

Логічні нейрони оперують з логічними даними. Причому, якщо значенням логічних даних є тільки 0 або 1, то мова йде про довічний логічному нейрону,

що використовує операції класичної (больової) логіки. У разі, коли значення логічних даних знаходяться в інтервалі $[0,1]$, то ми маємо справу з нечітким логічним нейроном, які використовує операції нечіткої логіки.

Штучні нейрони в залежності від їх апаратної реалізації підрозділяються на цифрові і аналогові нейрони.

Цифрові нейрони реалізуються мікросхемами середнього і високого ступеня інтеграції, при цьому на кристалі реалізується безпосередньо нейронна мережа, окремі нейрони не виділяються як самостійні одиниці, цифрову модель розробляють для вирішення конкретного завдання, яку неможливо вирішити за допомогою одного нейрона.

Аналогові нейрони дозволяють відтворювати просторове і тимчасове підсумовуванням збуджуючих сигналів, властивості абсолютної і відносної рефрактерності (несприйнятливості), процеси переробки інформації в окремому нейрони і при взаємодії нейронів. Аналогову модель розглядають як пристрій, стан якого може безперервно змінюватися від повного спокою до деякого максимального рівня збудження.

Штучні нейрони в залежності від способів опису їх моделі поділяються на формальні моделі, фізіологічні моделі і феноменологічні моделі нейрона.

Формальні моделі нейрона відрізняються добре розробленим математичним апаратом, але ігнорують багато властивостей свого біологічного прототипу.

Фізіологічні моделі нейрона відрізняються кількісним описом поведінки нейрона, породжені з експериментів над біологічними нейронами.

Феноменологічні моделі нейронів не мають суворої математичної і досвідченої бази.

Формальні моделі нейронів, в свою чергу, також можуть бути розділені в залежності від змінності осей входів, від виду функції активації і від ймовірності значення вихідного сигналу.

У свою чергу, серед багатошарових нейронних мереж розрізняють такі типи.

1. Монотонний. Це особливий випадок шаруватих мереж з додатковими умовами для з'єднань та нейронів. Кожен шар, крім останнього (вихідний), розділений на два блоки: збуджуючий і гальмуючий. З'єднання між блоками також поділяються на гальмуючі та захоплюючі. Якщо з нейронів блоку А для блокування нейронів В виробляються лише збудливі з'єднання, то це означає, що будь-який вихід блоку є монотонною не спадаючою функцією будь-якого виходу блоку А. Якщо ці з'єднання лише гальмують, то будь-який вихід блоку В не є зростаючою функцією будь-якого виходу блоку А. Монотонна мережа нейронів вимагає монотонної залежності виходу нейрона від параметрів вхідних сигналів.

2. Мережі без зворотного зв'язку. У таких мережах нейрони вхідного рівня приймають вхідні сигнали, перетворюють їх і передають їх до нейронів першого прихованого шару тощо, аж до виходу, який видає сигнали інтерпретатору та користувачеві. Якщо не зазначено інше, кожен вихідний сигнал $n-$го шару подаватиметься на вхід усіх нейронів (n + 1) -го шару; однак можливе з'єднання і-го шару з довільним $(\mu + p)$ - m шаром.

3. Мережі зворотного зв'язку. У мережах зворотного зв’язку інформація з наступних шарів передається попереднім. Серед них такі:

– шарувато-циклічний, що відрізняється тим, що шар укладений у кільце: останній шар передає свої вихідні сигнали першому; всі шари рівні і можуть приймати як вхідні сигнали, так і вихідні;

– шаруваті і повністю пов'язані складаються з шарів, кожен з яких є повністю пов'язаною мережею, і сигнали передаються від шару до шару і всередині шару; у кожному шарі цикл роботи поділяється на три частини: прийом сигналів з попереднього шару, обмін сигналами всередині шару, генерація вихідного сигналу та передача наступному шару;

– одношарові, за структурою схожі на шаруваті, але функціонують по-різному: вони не відокремлюють фази обміну всередині шару і переходять на наступний, у кожен годинник, нейрони всіх шарів отримують сигнали від нейронів обох власних шару та наступних.

1.6 Постановка задачі дипломного проектування

Задачею проектування було огляд та вивчення елементів теорії нейронних мереж, а також аналізування методів функцій активацій нейронних мереж і порівняти нейронні мережі з природними нейронами. Було оглянуто багато методів функцій активації та вивчено історію появи штучних нейронів і нейронних мереж. Під час огляду великої кількісті матеріалу по функціях активацій нейронних мереж було проведене ознайомлення з багатьма видами функцій активації, які ми змогли побачити в наведених прикладах.

А отже було визначено з розглянутого матеріалу, що важливою перевагою штучних нейронних мереж перед деякими іншими застосованими алгоритмами є те, що вони передбачають можливість навчання. При належному навчанні нейронна мережа здатна виявити складні зв’язки між вхідною та збереженою інформацією та узагальнити її, що нагадує психологічні процеси людини. Тому штучні нейронні мережі також розглядаються як один із способів моделювання штучних інтелектуальних систем.

Також основною задачею ще була у виборі методу апроксимації сигмоїдальної функції та її проектної реалізації на FPGA, що ж дало нам зрозуміти щоб обчислювати таку функцію потрібне використання міцного апаратного обладнання, аби швидкість обчислення було швидкою і без жодного збою програми.

2 ФУНКЦІЇ АКТИВАЦІЇ НЕЙРОННИХ МЕРЕЖ

2.1 Загальні відомості про функції активації

Функція активації (активаційна функція, передавальна функція) - це функція f, що обчислює вихідний сигнал штучного нейрона y, яка в якості аргументу приймає зважену суму s, одержувану на виході суматора, $y = f(s)$.

Вид функції активації є підсилювальною характеристикою штучного нейрона, визначає його функціональні можливості і метод його навчання.

Переважно застосовують нелінійну функцію активації, оскільки лінійні функції обмежені і їх вихід пропорційний входу. Застосування лінійних функцій активації було проблемою ранніх моделях нейронних мереж, і їх обмеженість і недоцільність була доведена в книзі Мінські і Пейперта "Персептрон".

В даний час є багато видів функцій активації. Приклади функцій активації представлені в таблиці 2.1.

Таблиця 2.1 – Приклади функцій активації

№	Назва	Вид	Області визначення і значення
1.	Лінійні функції активації (Purelin)	$f(s) = ks$	$D(f) = (-\infty, +\infty)$ $E(f) = (-\infty, +\infty)$
2.	Напівлінійні функції активації (Poslin)	$f(s) = \begin{cases} ks, \text{якщо } s > 0 \\ 0, \text{якщо } s \leq 0 \end{cases}$	$D(f) =$ $(-\infty, +\infty)$, $E(f) = (-\infty, +\infty)$
3.	Зміщені які насичують лінійно - крокові функції активації (Satlin)	$f(s) = \begin{cases} 0, \text{якщо } s \leq T \\ \dfrac{s-t}{\Delta}, \text{якщо } T < s < T + \Delta \\ t, \text{якщо } s \geq T + \Delta \end{cases}$	$D(f) = (-\infty, +\infty)$ $E(f) = (0, +t)$

Продовження таблиця 2.1

4.	Симетричні які насичують лінійно – крокові функції активації (Satlins)	$f(s)=\begin{cases}-t, \text{якщо } s \le -T \\ S\dfrac{t}{T}, \text{якщо } \lvert s\rvert < T \\ t, \text{якщо } s \ge T\end{cases}$	$D(f)=(-\infty,+\infty)$ $E(f)=(-t,+t)$
5.	Трикутні функції активації (Tribas)	$f(s)=\begin{cases}0, \text{якщо } s < -1 \\ 1-\lvert s\rvert, \text{якщо } -1 \le s \le 1 \\ 0, \text{якщо } s > 1\end{cases}$	$D(f)=(-\infty,+\infty)$ $E(f)=(0,+1)$
6.	Зміщені ступінчасті - порогові функції активації(Hardlim)	$f(s)=\begin{cases}t, \text{якщо } s \ge T \\ 0, \text{інакше}\end{cases}$	$D(f)=(-\infty,+\infty)$ $E(f)=\{0\}\cup\{+t\}$
7.	Симетричні ступінчасті - порогові функції активації (Hardlims)	$f(s)=\begin{cases}1, \text{якщо } s \ge 0 \\ -1, \text{якщо } s < 0\end{cases}$	$D(f)=(-\infty,+\infty)$ $E(f)=\{-1\}\cup\{+1\}$
8.	Модульні функції активації (Modul)	$f(s)=\lvert s\rvert$	$D(f)=(-\infty,+\infty)$ $E(f)=(0,+\infty)$
9.	Зміщені сигментальні - логістичні функції активації (Logsig)	$f(s)=\dfrac{1}{1+e^{-as}}$	$D(f)=(-\infty,+\infty)$ $E(f)=(0,+1)$
10.	Симетрична сигментальна функція активації – гіперболічного тангенса (Tansig)	$f(s)=\dfrac{e^{2s}-1}{e^{2s}+1}$	$D(f)=(-\infty,+\infty)$ $E(f)=(-1,+1)$
11.	Раціональні сигментальні функції активації (Rsigmoid)	$f(s)=\frac{s}{k+\lvert s\rvert}$, $f(s)=\dfrac{\mathrm{s}}{k+\mathrm{s}}$	$D(f)=(-\infty,+\infty)$, $k>0$ $E(f)=(-1,+1)$

2.2 Лінійні функції активації

Лінійна функція активації виробляє вихідний сигнал у пропорційному значенню потенціалу (значенням зваженої суми) нейрона s і представляється виразом $f(s)ks$.

У лінійної функції активації областю визначення є $D(f) = (-\infty, +\infty)$, а областю значень – $E(f) = (-\infty, +\infty)$. Її графік показаний на рисунку 2.1

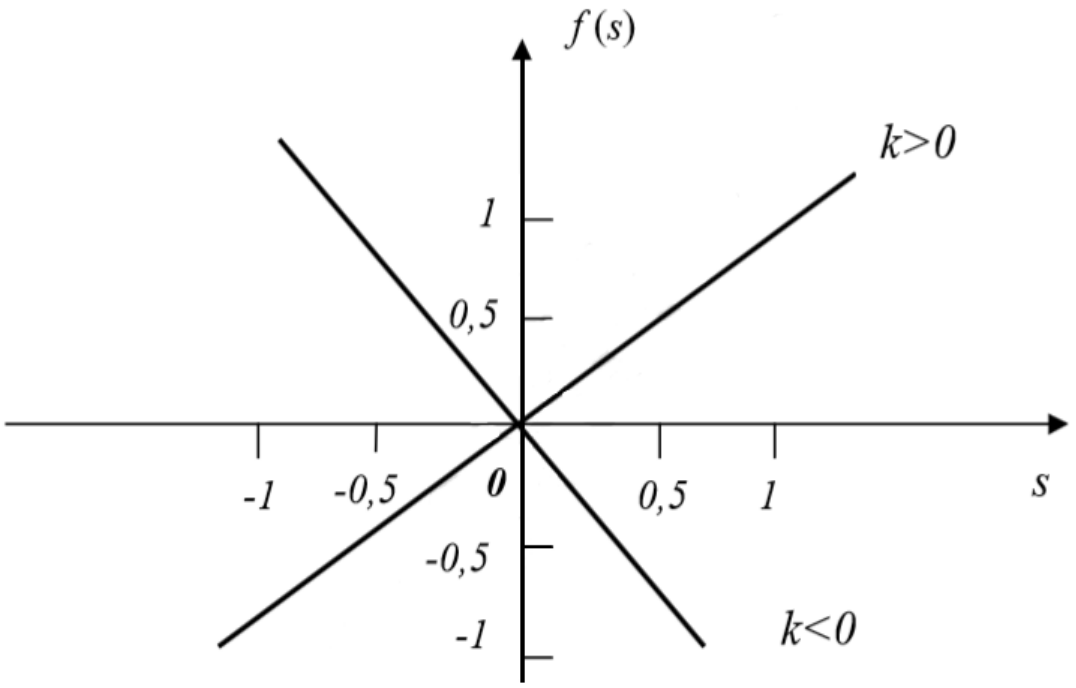


Рисунок 2.1 – Графік лінійної функції активації.

Лінійна функція активації використовується в нейронних мережах різних типів, зокрема лінійних, а також в вихідних шарах мереж на радіальних базисних функціях.

Лінійні функції активації мають наступні модифікації: напівлінійні, крокові і трикутні.

Напівлінійна функція активації: якщо значення потенціалу нейрона позитивний, то виробляє вихідний сигнал у пропорційному потенціалу нейрона s, інакше не виробляє вихідний сигнал і представляється наступним виразом:

$$f(s) = \begin{cases} ks, \text{якщо } s > 0 \\ 0, \text{якщо } s \leq 0 \end{cases}, \quad (2.1)$$

Залежно від позитивності або заперечності значення коефіцієнта, напівлінійна функція може приймати позитивне або негативне значення відповідно. Тому у напівлінійних функціях активації областю визначення є $D(f) = (-\infty, +\infty)$, а областю значень – $E(f) = (-\infty, +\infty)$.

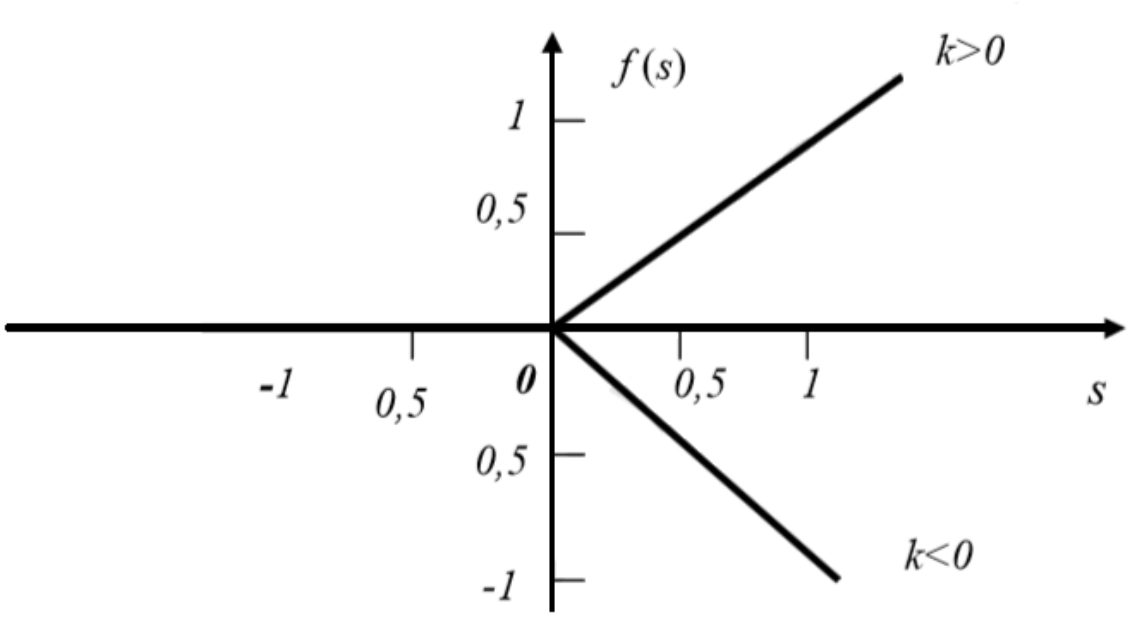


Рисунок 2.2 – Графік напівлінійних функцій активації.

Напівлінійні функції активації використовуються при моделюванні простих мереж, наприклад, персептронів.

Крокова функція активації (Лінійна функція активації з насиченням) має дві лінійні ділянки, де їх значення тотожно дорівнює мінімально допустимому $-\mathrm{t}$ і максимально допустимого значення t і є ділянка, на якому функція строго монотонно зростає. При цьому можливий зсув функції по обох осях: t – по осі абсцис, Δ – по осі ординат.

У крокової функції активації областю визначення є $\mathrm{D(f)} = (-\infty, +\infty)$, а областю значень – $\mathrm{E(f)} = (-\mathrm{t}, +\mathrm{t})$.

Є два види крокової функції: зміщена крокова функція і симетрична крокова функція.

Зміщена крокова функція активації представляється наступним виразом:

$$f(s) = \begin{cases} 0, \text{якщо } s \leq T \\ \frac{s-t}{\Delta}, \text{якщо } T < s < T + \Delta. \\ t, \text{якщо } s \geq T + \Delta \end{cases} \quad (2.2)$$

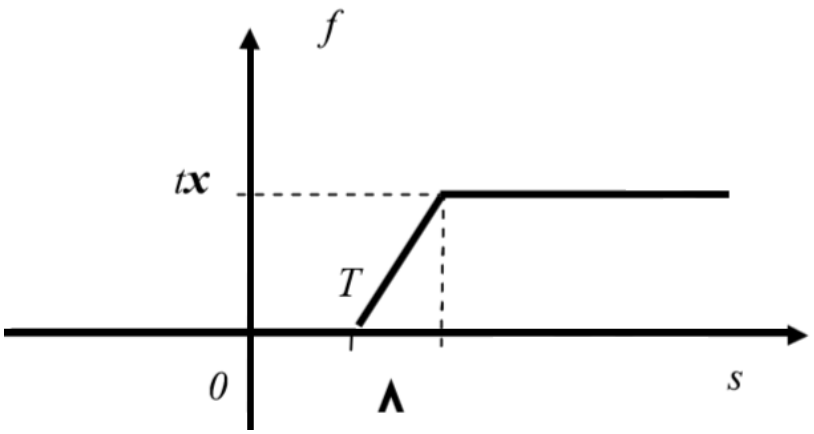


Рисунок 2.3 – Графік зміщеної шагової функції активації.

У загальному випадку симетрична крокова функція активації представляється наступним виразом:

$$f(s) = \begin{cases} -t, \text{якщо } s \leq -T \\ S\frac{t}{T}, \text{якщо } |s| < T, \\ t, \text{якщо } s \geq T \end{cases} \quad (2.3)$$

При $t = 1$ і $T = 1$ графік симетричною крокової функції буде виглядати:

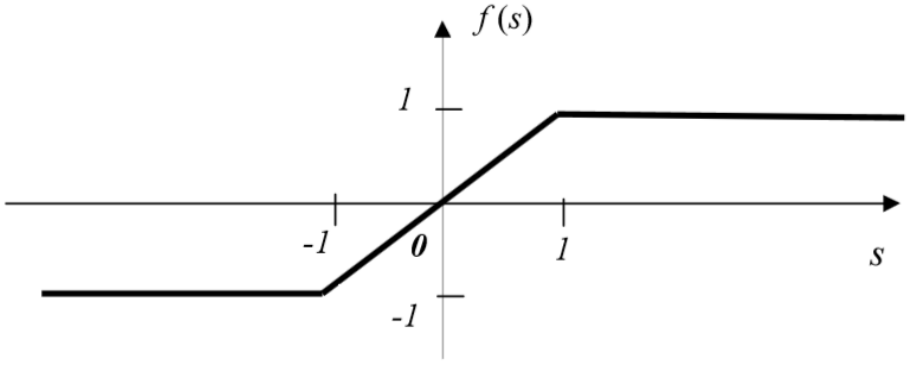


Рисунок 2.4 – Графік симетричної крокової функції активації.

Крокова функція активації використовується при моделюванні простих мереж, наприклад, персептронових.

Недоліками напівлінійних і крокових функцій активації, щодо лінійної можна назвати те, що вона не є диференційованою на всій числовій осі, а значить не може бути використана під час навчання за деякими алгоритмам, що вимагає диференційованої функції активації, наприклад, алгоритм зворотного поширення помилок.

Трикутна функція активації (Кусково-лінійна функція з обмеженням) має дві лінійні ділянки, де її значення тотожно дорівнює нулю і має дві ділянки, на першому функція строго монотонно зростає і на другому функція строго монотонно спадає.

$$f(s) = \begin{cases} 0, \text{якщо } s < -1 \\ 1 - |s|, \text{якщо } -1 \leq s \leq 1, \\ 0, \text{якщо } s > 1 \end{cases} \tag{2.4}$$

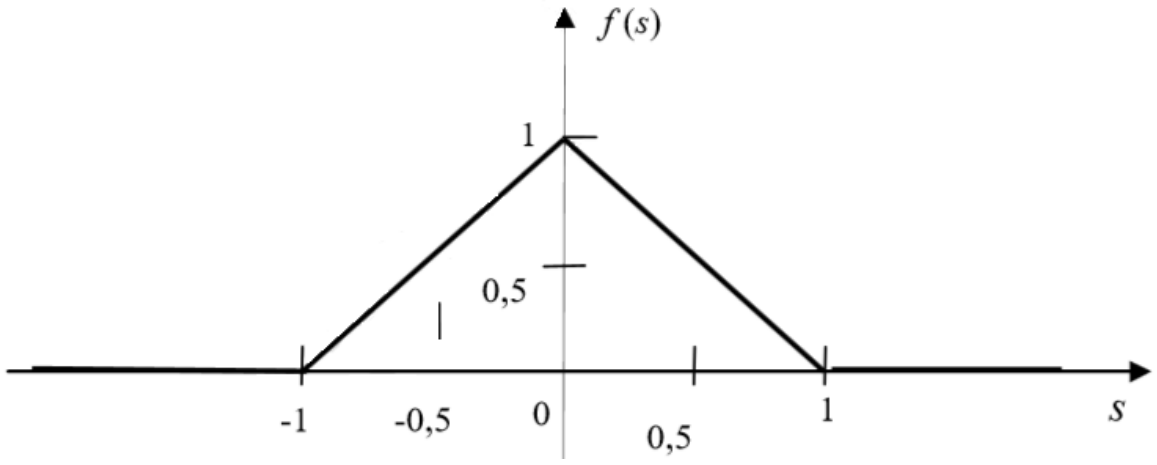


Рисунок 2.5 – Графік трикутної функції активації.

2.3 Порогові функції активації

Порогові функції активації представляють собою перепад і виробляють вихідний сигнал з граничними значеннями a і b в залежності від активності

нейрона. При цьому, якщо значення потенціалу нейрона не досягає деякого рівня T, то вихідний сигнал дорівнює нулю, в іншому випадку вихідний сигнал стрибкоподібно змінюється на t.

$$f(s) = \begin{cases} t, \text{якщо } s \geq T \\ 0, \text{інакше} \end{cases}. \quad (2.5)$$

Тут T є зрушенням функції активації щодо горизонтальній осі $T = -w_0 * x_0$, відповідно під s слід розуміти зважену суму сигналів на входах нейрона без урахування цього доданка.

Порогову функцію активації іноді називають ступінчастою функцією активації.

Є зміщена порогова функція активації і симетрична порогова функція активації.

У зміщеної порогової функції активації областю визначення є $D(f) = (-\infty, +\infty)$, а областю значень – $E(f) = \{-t\} \cup \{+t\}$.

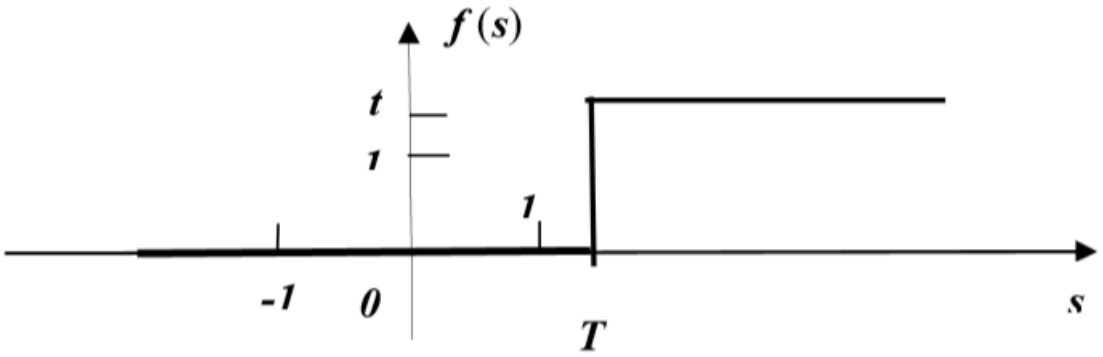


Рисунок 2.6 – Графік зміщеної порогової функції активації.

Якщо в зміщеної порогової функції активації $T = 0$ і $t = 0$, то вона перетвориться в функцію одиничного стрибка (функцію Хевісайда), представлену наступним виразом:

$$f(s) = \begin{cases} 1, \text{якщо } s \geq 0 \\ 0, \text{якщо } s < 0' \end{cases} \quad (2.6)$$

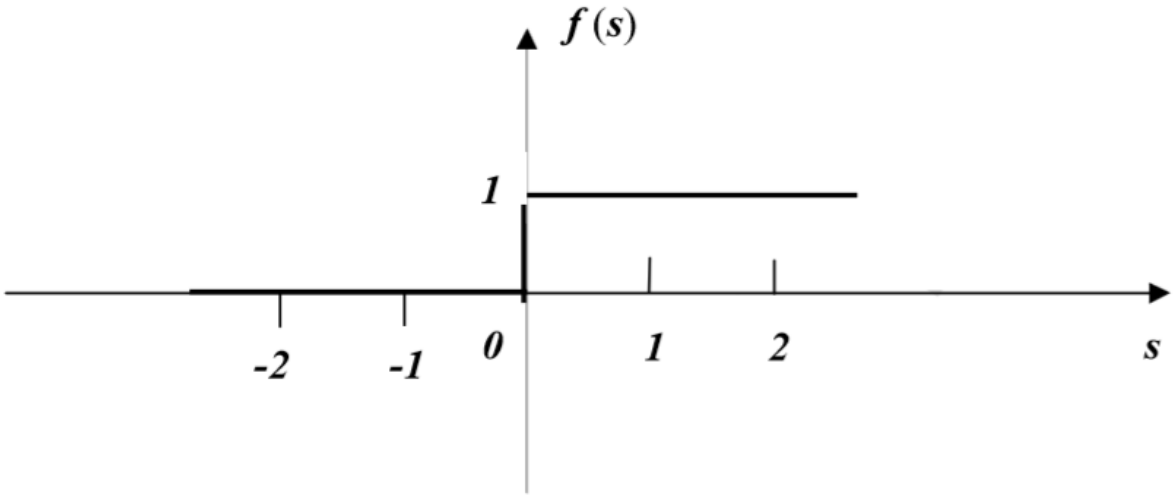


Рисунок 2.7 – Графік функції активації одиничного скачка.

Симетрична порогова (знакова) функція активації представляється наступним виразом:

$$f(s) = \begin{cases} 1, \text{якщо } s \geq 0 \\ -1, \text{якщо } s < 0 \end{cases}, \tag{2.7}$$

Областю визначення симетричної порогової функції активації є інтервал $D(f) = (-\infty, +\infty)$, а областю значень – множина $E(f) = \{-1\} \cup \{1\}$.

Симетрична порогова функція іноді називається біполярної функцією активації або жорсткою знаковою функцією.

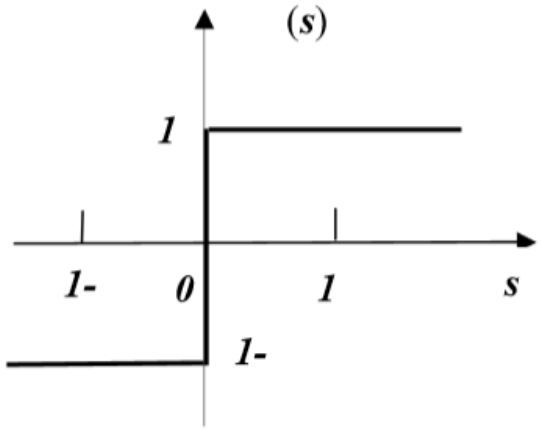


Рисунок 2.8 – Графік симетричної порогової функції активації.

А також це проста кусочно-лінійна функція, яка часто використовується в нейронних мережах. Результат приймає значення 0 для негативного аргументу і 1 для позитивного аргументу.

$$f(x) = \begin{cases} 0, & x < 0 \\ 1, & x \geq 0 \end{cases}, \tag{2.8}$$

Даний тип функції активації корисний для бінарних схем. Наприклад, для випадку бінарної класифікації. Графічна інтерпретація представлена нижче на рисунку 2.9:

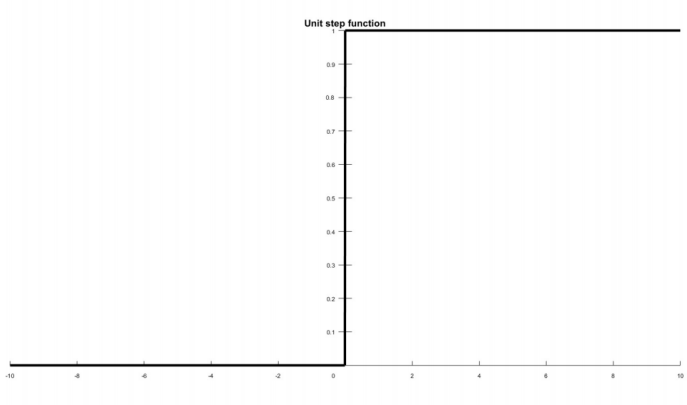


Рисунок 2.9 – Порогова функція активації.

Можна визначити багатопорогову функцію активації, якщо вихідний сигнал нейрона буде приймати одне з m значень, що визначаються $m-1$ порогом всередині граничних значень a і b.

З огляду на те, що ця функція не диференційована на всій осі абсцис, її не можна використовувати в мережах, які навчаються за алгоритмом зворотного поширення помилки іншим алгоритмам, що вимагає диференційованої функції активації.

Порогові функції можуть використовуватися при моделюванні нейронних мереж, які не володіють гнучкістю при налаштуванні навчання їх на розв’язання завдання, тому що, якщо значення суми не досягає порогу, то вихідний сигнал не формується, мережа «не спрацьовує». Втрачається інтенсивність вихідного

сигналу нейрона і формується невисоке значення рівня на зважених входах в наступному шарі нейронів.

2.4 Модульні функції активації

Модульна функція активації виробляє безперервний лінійний вихідний сигнал в залежності від активності нейрона представляється наступним виразом:

$$f(s) = |s|, \tag{2.9}$$

У модульній функції активації областю визначення є $D(f) = (-\infty, +\infty)$, а областю значень – $E(f) = (0, +\infty)$.

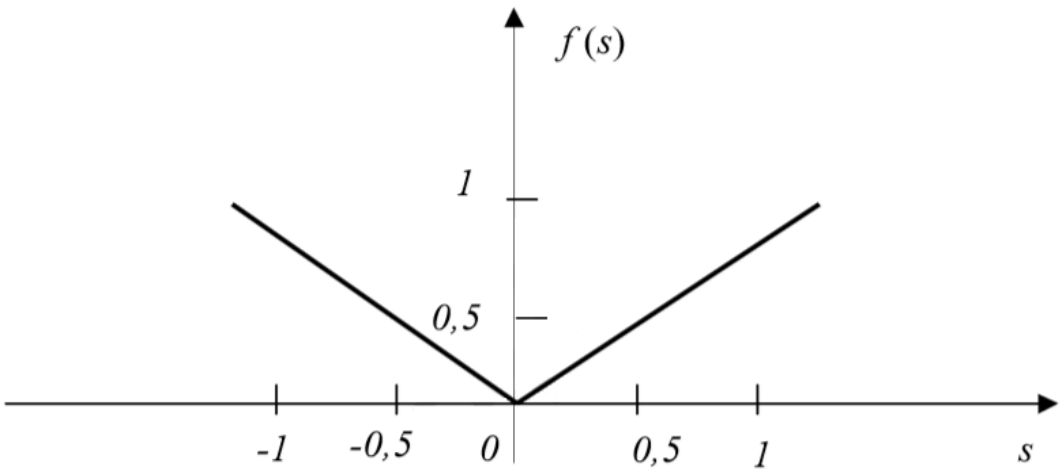


Рисунок 2.10 – Графік модульної функції активації.

Для навчання нейрона модульної функції активації можна використовувати метод зворотного поширення помилки.

Для взяття похідної від цієї функції потрібно врахувати, що в точках, де модульна функція дорівнює 0, похідної не існує, в інших точках потрібно розкрити модуль і продиференціювати:

$$f'(s) = \begin{cases} -s, \text{якщо } s < 0 \\ s, \text{якщо } s > 0 \end{cases}. \tag{2.10}$$

Легко помітити, що вираз для похідної модульної функції збігається виразом порогової (знакової) функції активації.

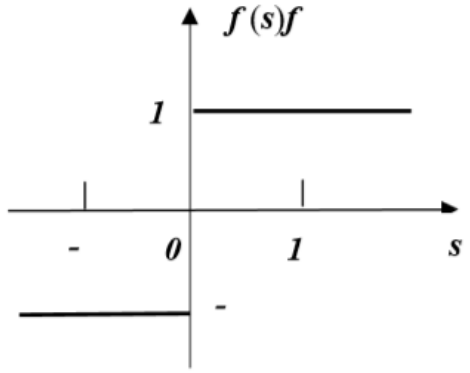


Рисунок 2.11 – Графік похідної модульної функції активації.

2.5 Сигмоїдальні функції активації

В даний час сигмоїдальні функції активації є одними з найбільш часто використовуваних функцій активації. Введення функцій сигмоїдального типу було обумовлено обмеженістю нейронних мереж з пороговою функцією активації нейронів. При такій функції активації будь-який з виходів мережі дорівнює або нулю, або одиниці, що обмежує використання мереж завданнях, що не відносяться завданням класифікації. Використання сигмоїдальних функцій дозволило перейти від бінарних (цифрових) виходів нейрона до дійсних (аналоговим). Функції активації такого типу, як правило, властиві нейронам, що знаходяться у внутрішніх шарах нейронної мережі.

Сигмоїдальні функції активації виробляють безперервний вихідний сигнал, значення якого лежать в інтервалі (a, b), і формують графік у формі сигмоїда (S-подібної кривої).

Сигмоїдальні функції активації підрозділяються на зміщені сигмоїдальні функції активації, симетричні сигмоїдальні функції активації і раціональні сигмоїдальні функції активації.

Зміщена сигмоїдальна функція активації представляється наступним виразом:

$$f(s) = \frac{1}{1+e^{-as}}, \tag{2.11}$$

де α - це параметр функції, що визначає крутизну.

Коли α прагне нескінченності, функція вироджується в порогову функцію. При $\alpha = 0$ сигмоїдальна функція вироджується в постійну функцію зі значенням 0,5. Зміщена сигмоїдальна функція іноді називається логістичною функцією активації. У логістичної функції активації областю визначення є $D(f) = (-\infty, +\infty)$, а областю значень – $E(f) = (0,1)$.

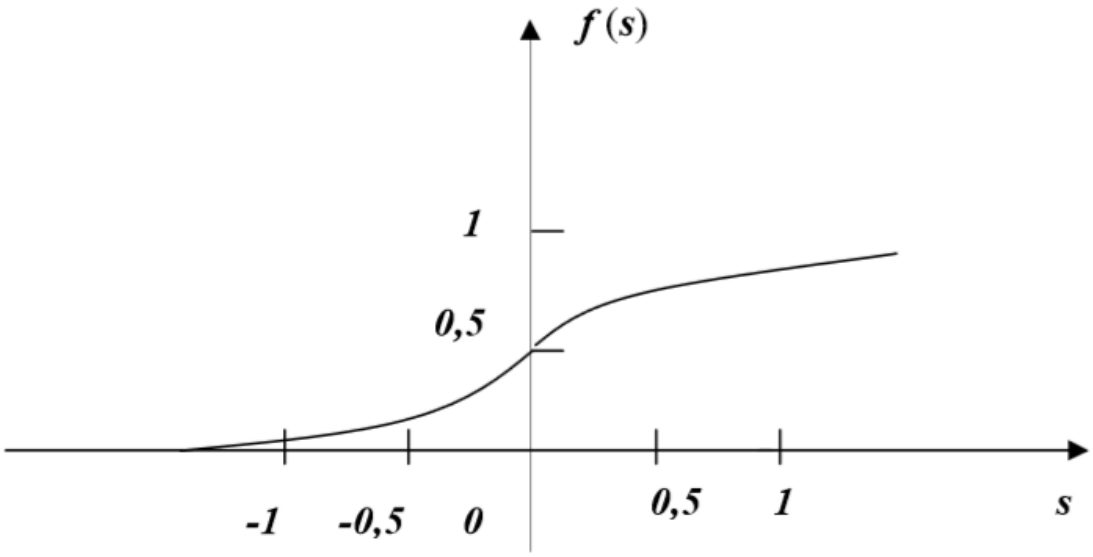


Рисунок 2.12 – Графік логістичної функції активації.

Логістична функція диференційована на всій осі абсцис і має дуже просту похідну:

$$f'(s) = af(s)(1 - f(s)), \tag{2.12}$$

Те, що похідна логістичної функції може бути виражена через її значення полегшує її використання при навчанні нейронної мережі за алгоритмом

зворотного поширення. Особливістю нейронів з такою передавальною характеристикою є те, що вони посилюють сильні сигнали істотно менше, ніж слабкі, оскільки області сильних сигналів відповідають пологим ділянкам характеристики. Це дозволяє запобігти насичення, що відбувається через великі сигналів.

Сигмоїдальний нейрон має таку ж структуру, що моделі нейрона Мак-Каллок-Піттса, але має диференційовану функцію активації, яка може бути виражена у вигляді сигмоїдальної уніполярної або біполярної функції, яка показана на рисунку 2.13.

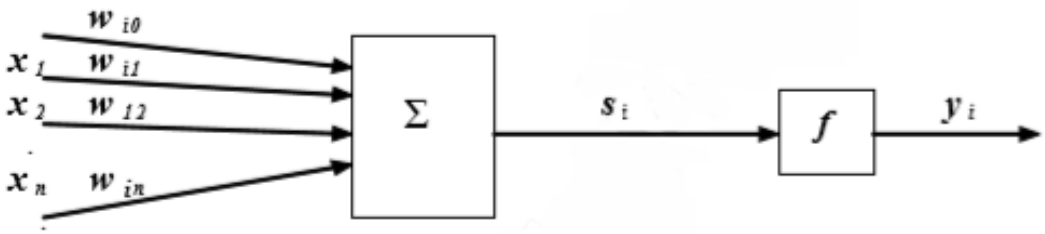


Рисунок. 2.13 – Структура сигмоїдального нейрона.

Уніполярна функція активації представляється виразом:

$$y_i = f(s_i) = \frac{1}{1+e^{-\beta s_i}}, \tag{2.13}$$

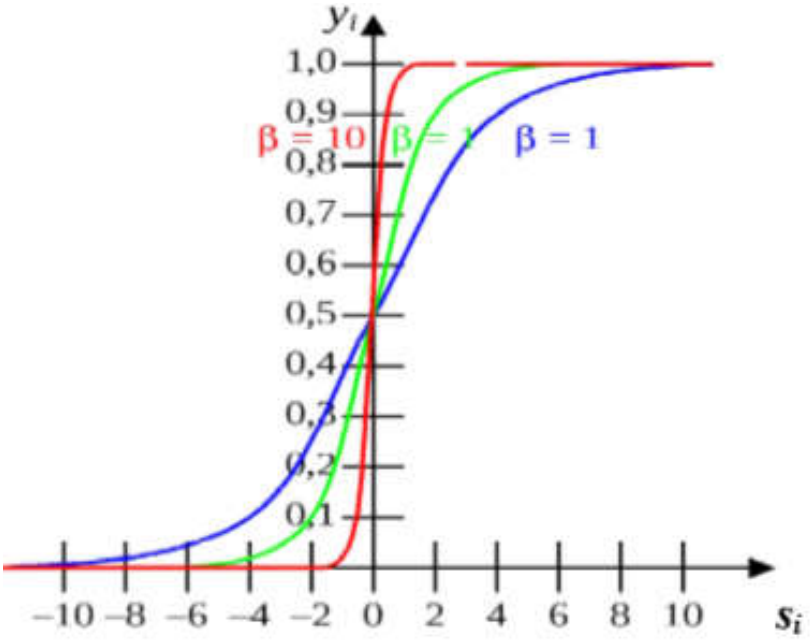


Рисунок 2.14 – Графік уніполярної функції.

Біполярна функція активації записується формулою:

$$y_i = f(s_i) = \tanh(\beta_{s_i}) = \frac{e^{\beta s_i} - e^{-\beta s_i}}{e^{\beta s_i} + e^{-\beta s_i}}. \quad (2.14)$$

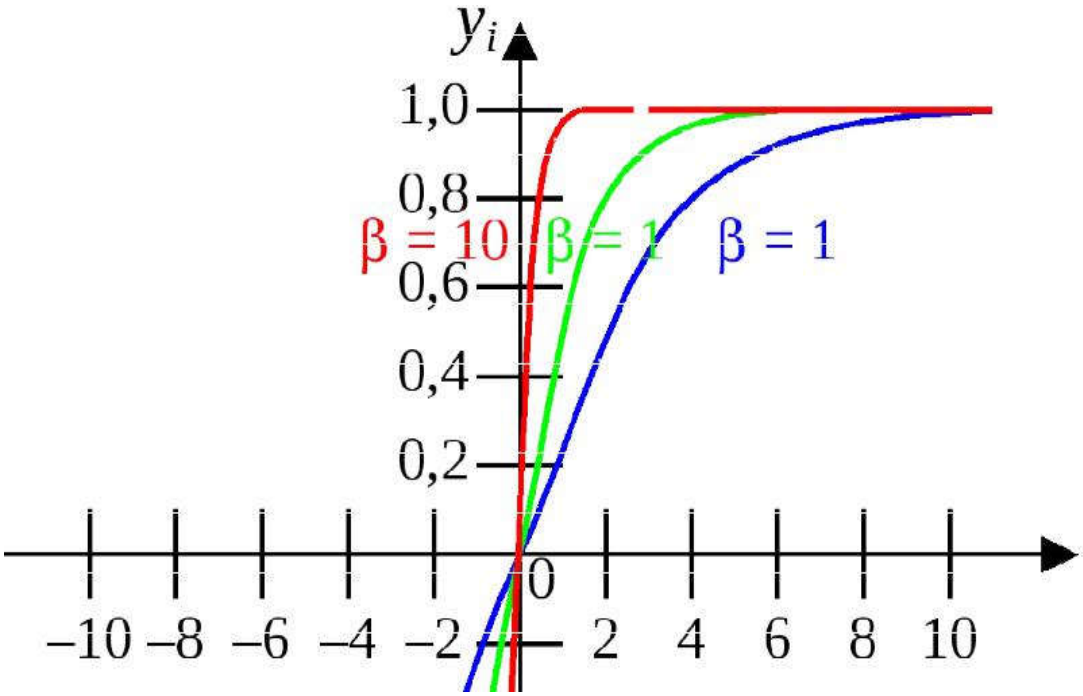


Рисунок 2.15 – Графік біполярної функції.

Значення параметра β підбирається користувачем і впливає на форму функції активації. При $\beta \to \infty$ сигмоїдальна функція перетворюється у функцію ступеневого типу, ідентичну функцію активації персептрона. На практиці зазвичай для спрощення використовується значення $\beta = 1$.

2.6 Гіперболічний тангенс

Іншою популярною і широко використовуваною функцією активації є функція tanh (гіперболічний тангенс). Відрізняється від розглянутої вище тим, що його область значень лежить в інтервалі $(-1; 1)$, обидва графіка відрізняються лише масштабом осей. Формула гіперболічного тангенса:

$$f(x) = \tanh(x) = \frac{e^{2x}-1}{e^{2x}+1}, \tag{2.15}$$

І на рисунку 2.16. показано функцію активації гіперболічного тангенса:

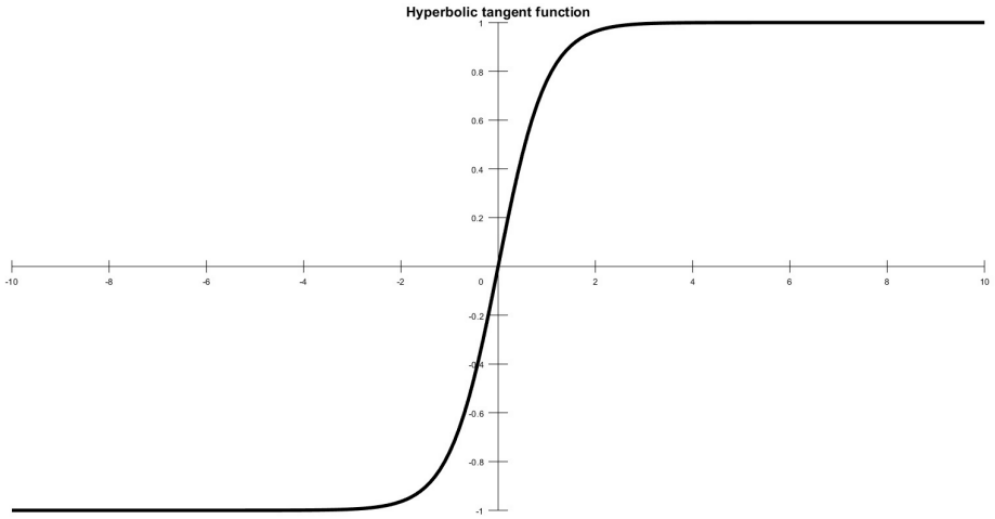


Рисунок 2.16 – Функція активації гіперболічного тангенса.

2.7 Стохастичні нейрони

Вище були описані моделі детермінованих штучних нейронів, в яких стан на виході однозначно визначаються результатами робіт суматора вхідних сигналів. Є також стохастичні нейрони, що відноситься до нестандартної моделі штучного нейрона, де перемикання нейрона відбувається з ймовірністю, що залежить від індукованого локального поля, тобто функція активації визначена як:

$$f(s) = \begin{cases} 1, & \text{якщо } P(s) \\ 0, & \text{якщо } 1 - P(s) \end{cases}, \tag{2.16}$$

де розподіл ймовірності P зазвичай має вигляд сигмоїда:

$$\sigma(s) = \frac{A(T)}{1+e^{(-s/T)}}, \tag{2.17}$$

а нормувальна константа A вводиться для умови нормалізації розподілу ймовірності:

$$\int_0^1 \sigma(s)ds = 1, \tag{2.18}$$

Таким чином, нейрон активується з ймовірністю P. Параметр T – аналог температури і визначає безлад в нейронній мережі. Якщо T спрямувати до 0, то стохастичний нейрон перейде в звичайний нейрон з передавальною функцією Хевісайда.

У стохастичній моделі, на відміну від детермінованих моделей, вихідний стан нейрона залежить не тільки від зваженої суми вхідних сигналів, але і від деякої випадкової змінної величини, значення якої вибираються при кожній реалізації з інтервалу (0,1).

У стохастичній моделі нейрона вихідний сигнал y, приймає значення ± 1 з ймовірністю:

$$P(y_i = \pm 1) = 1/\left(1 + e^{\pm 2\beta s_i}\right), \tag{2.19}$$

де s_i – зважена сума вхідних сигналів i-го нейрона, β – позитивна константа, яка найчастіше дорівнює 1.

Процес навчання нейрона в стохастичній моделі складається з наступних етапів:

для кожного нейрона мережі визначається зважена сума:

$$s_j = \sum_{i=0}^{n} w_{ij}x_i, \tag{2.20}$$

визначення вірогідності P, що вихідна змінна y_i приймає значення ± 1:

$$P(y_i = \pm 1) = \frac{1}{1+e^{\pm 2\beta s_i}}, \quad (2.21)$$

генерація значень випадкової змінної $R \in (0, 1)$;

Формування вихідного сигналу y, якщо $R < P(y)$, або $-y$ у поганому випадку;

Адаптація ваги w_{ij} (при фіксованих y_i) з використаних правил.

За правилом Уідроу-Хоффа адаптація ваги проводиться за формулою:

$$\Delta w_i = \eta x_i (d_i - y_i), \quad (2.22)$$

де η – коефіцієнт вивчення нейрона, значення якого лежить в інтервалі (0, 1).

Сукупність нейронів, певним чином з'єднаних один з одним і зовнішнім середовищем, утворює нейронну мережу. Для того, щоб дізнатися які існують нейронні мережі, і правильно оцінити їхні можливості.

3 РЕАЛІЗАЦІЯ СИГМОЇДАЛЬНИХ ФУНКЦІЙ АКТИВАЦІЇ ДЛЯ НЕЙРОННИХ МЕРЕЖ

3.1. Актуальність реалізації штучних нейронних мереж

За останні декілька років піднявся інтерес до вивчення апаратної реалізації штучних нейронних мереж (ШНМ). Це не важко помітити тому що почало з'являтися велика кількість публікацій на цю тему. Це насамперед пов'язано із швидким розвитком елементної бази, яка використовується при виконанні цифрових ШНМ (надвеликих інтегральних схем – VLIS). Одна з головних проблем, які лишаються при цьому, є приріст швидкодії роботи нейронних мереж. Одним із багатьох способів, що дає значне зростання швидкості роботи за рахунок паралелізму, є реалізація їх на FPGA.

Швидкодія штучних нейронів напряму залежить від функції активації. Реалізація сигмоїдальної функції, гіперболічного тангенса на FPGA вимагає міцного заліза із великим ресурсом [1, 2, 3]. В роботі [1] було зроблено огляд головних методів, що використовуються при виконанні сигмоїдальної функції та гіперболічного тангенса.

Для цифрової реалізації нелінійних функцій активації використовуються різні методи наближення: табличний, ряд Тейлора, кусочно-лінійне наближення [1, 2]. У ряду Тейлора потрібно зробити багато множин, цей спосіб не прийнятний для реалізації FPGA. Метод таблиці заздалегідь бачить створення таблиці можливих значень цільової функції врахувавши її обмежені цифри. Однак для того щоб створити окрему локальну таблицю для кожного нейрона будуть потрібні значні апаратні ресурси FPGA. Використовуючи одну таблицю для різних нейронів призведе до довготривалих часових зупинок роботи, оскільки розповсюдження сигналів у нейронах одного шару здійснюється паралельно. Тому кусково-лінійне наближення (PWL) є найбільш часто використовуваною реалізацією функцій активації сигмоїдального типу. Кусково-лінійне наближення сигмоїдальної функції полягає в заміні нелінійної

функції на кожному з її інтервалів прямою лінією. Це наближення використовується у багатьох працях [2, 3].

Вибір методу наближення функцій активації сигмоїдального типу та їх апаратна реалізація є основними планами, від яких залежить точність та швидкодія алгоритму. При низькій точності наближення ми отримуємо низьку швидкість, а зменшення помилки наближення призводить до збільшення апаратних ресурсів і зниження швидкості обробки даних.

Ще одним важливим моментом, який слід враховувати, є диференціація апроксимованої функції активації [2, 3], оскільки методи навчання ШНМ включають як функцію активації, так і її похідну.

3.2. Функції активації нейронів сигмоїдального типу.

Сигмоїдальні функції активації – найчастіше використовуються для нейронних мереж з прямим розповсюдженням сигналу. Вони монотонно зростають, безперервні і диференційовані. Для опису та аналізу функцій активації сигмоідного типу в [1] використовували загальний клас функцій:

$$f(x,k,b,T,c) = k + \frac{c}{1+be^{Tx}}, \quad \forall x \in R, \tag{3.1}$$

де $k \in R$, $b \in R^+$, T, $c \in R\backslash\{0\}$, R – множина дійсних чисел $(-\infty,+\infty)$, R^+ – множина дійсних додатних чисел $(0,+\infty)$, $R\backslash\{0\}$ – множина дійсних чисел за винятком точки 0 $(-\infty,0)$ та $(0,+\infty)$. Сигмоїдальні нелінійності, які відносяться до цього класу:

класична сигмоїдальна функція ($k = 0, c = b = 1$ та $T = -1$);

$$f(x) = \frac{1}{1+be^{Tx}}, \tag{3.2}$$

тангенс гіперболічний ($k = 1, c = -2, b = 1, T = 2$):

$$h(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}, \tag{3.3}$$

Слід зазначити, що обчислення функції сигмоїдальної активації (3.2) достатньо для виконання лише позитивних аргументів x. Для від'ємних значень x його можна знайти з виразу:

$$f(-x) = 1 - f(x), \tag{3.4}$$

Дійсно, виконавши прості перетворення для виразу (3.2), отримаємо:

$$f(-x) = \frac{1}{1+e^x} = \frac{1}{1+\frac{1}{e^{-x}}} = \frac{e^{-x}}{1+e^{-x}} = \frac{1+e^{-x}-1}{1+e^{-x}} = 1 - f(x),$$

Нелінійна гіперболічна дотична функція (3.3) також може бути розрахована за допомогою класичної сигмоїдальной функції (3.2).

$$h(x) = 1 + \frac{e-e}{e+e} - 1 = \frac{2e}{e+e} - 1 = \frac{2}{1+e} - 1 = 2f(2x) - 1, \tag{3.5}$$

$$h(-x) = 2f(-2x) - 1 = 2[1 - f(2x)] - 1 = 1 - 2f(2x), \tag{3.6}$$

Для оцінки точності наближення використовуються максимальна та середня похибка [2]. Середня абсолютна ε_{ave}, максимальна абсолютна $\varepsilon_{\max}$ похибки функції $f(x)$, яка апроксимується функцією $\bar{f}(x)$ в інтервалі $(x_{\min}, x_{max})$ визначається як:

$$\varepsilon_{ave} = \frac{\sum_{i=0}^{N_p-1} |\bar{f}(x_i) - f(x_i)|}{N_p}, \tag{3.7}$$

$$\varepsilon_{\max} = \max\left|\bar{f}(x_i) - f(x_i)\right|, i = 0, \dots, N_p, \quad (3.8)$$

де N_p – кількість точок, на які розбиваються інтервал $(x_{\min}, x_{\max})$.

Ефективність наближення порівнюється з точки зору точності, швидкості та обладнання.

Арифметичні дії в нейронах виконуються над реальними числами. Під час виконання арифметичних операцій FPGA над реальними числами з фіксованою чи плаваючою комою необхідний час для їх виконання, а також апаратні ресурси для їхньої реалізації. Саме тому при реалізації компонентів штучних нейронів у VHDL дійсні числа перетворюються на цілі числа шляхом їх множення на 2^{10} і відкиданням дробової частини. Формат подачі дійсних чисел має розмірність 16 біт: 1 біт – знаковий і 15 біт – для зберігання отриманого цілого числа. Від'ємні числа представлені у додатковому коді. Максимальна ціла частка не повинна перевищувати 15. Слід зазначити, що вхідні дані в нейронних мережах мають властивість нормалізуватися. Щоб представити дійсне додатне число 2.1625 у цілому форматі, його множимо на $2^{10} = 1024$ і відкидаємо дробову частину: $2.1625 * 1024 = 2214.4 \approx 2214$. У шістнадцятковій системі числення це число рівне: $2214_{10} = 0x08A6$. Це ж саме від'ємне число в шістнадцятковій системі числення $-2214_{10} = 0xF75A$.

3.3. Методи апроксимації сигмоїдальних функцій.

Розглянемо декілька алгоритмів кусково-лінійного наближення, що відрізняються кількістю точок, розташуванням початкової та кінцевої точок на апроксимуючих лініях та критеріями їх вибору. Кусково-лінійна апроксимація нелінійної функції [2, 3]. В цьому методі апроксимація сигмоїдальної функції (3.2) виконується виразом [2]:

$$f(x) = \begin{cases} 1, & x \geq 5.0 \\ 0.03125 * x + 0.84375, & 2.375 \leq x < 5.0 \\ 0.125 * x + 0.625, & 1.0 \leq x < 2.375 \\ 0.25 * x + 0.5, & 0 \leq x < 1.0 \end{cases}, \quad (3.9)$$

Розрахунки варто проводити лише для додатніх значень x. Для від'ємного значення x сигмоїдальна функція обчислюється за допомогою виразу (3.4). Перетворюємо вираз (3.9) у ціле число, шляхом множення його доданків без змінної x на 2^{10}:

$$f(x) = \begin{cases} 1024, & x \geq 5120 \\ 2^{-5} * x + 864, & 2432 \leq x < 5120 \\ 2^{-3} * x + 640, & 1024 \leq x < 2432 \\ 2^{-2} * x + 512, & 0 \leq x < 1024 \end{cases}, \quad (3.10)$$

Для реалізації сигмоїдальної функції відповідно до виразу (3.10) розроблена блок-схема пристрою, яка показана на рис. 1, де Rg є регістром, LessThan - пристроєм порівняння, Add - суматором, Buf є буфером третього стану.

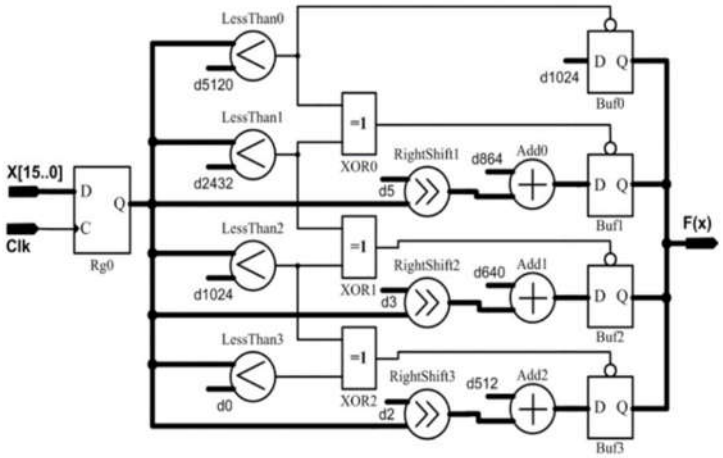


Рисунок 3.1 – Структурна схема пристрою, що реалізує сигмоїдальну функцію згідно виразу (3.10)

Для обчислення сигмоподібної функції цього пристрою потрібні три Add, чотири пристрої порівняння LessThan, чотири формувачі шини Buf. Множення на величини $2^{-2}, 2^{-3}$ і 2^{-5} (зсув вправо) здійснюється шляхом належного підключення входів і виходів відповідних компонентів схеми. На схемі це показано пристроями зміщення RightShift. Результати обчислення сигмоїдальної функції, отримані на виходах Add1, Add2 і Add3, вибираються на основі результатів порівняння входу x з числами 5120, 2432, 1024 і 0. На одному вході (A_i) пристроїв порівняння LessThan поступають вхідні дані (x), які порівнюються з даними на інших входах (B_i). На виходах LessThan формується результат порівняння:

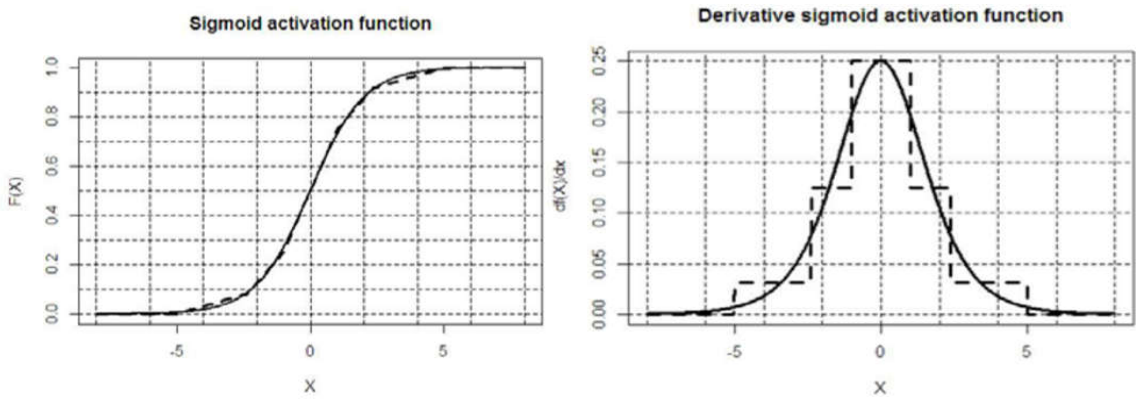
$$Out_{LessThan} = \begin{cases} 0, x < B_i \\ 1, x \geq B_i \end{cases}, \tag{3.11}$$

Інформація про використання пристроїв порівняння LessThan надходить до введення інших елементів АВО (XOR). Сигнали від LessThan1 виявляють, що XOR виконують ширини формули, які надсилають вихід, коли логічний сигнал може бути включений $'1'$ та переводять вихід в третій стан (Z), якщо на цьому вході – сигнал логічного $'0'$.

Час обчислення сигмоїдальної функції для цієї схеми визначається за формулою:

$$t_1 = t_{Rg} + t_{LessThen} + t_{XOR} + t_{Add} + t_{Buf}, \tag{3.12}$$

де t_{Rg}, $t_{LessThen}$, t_{XOR}, t_{Add}, t_{Buf} – час затримки сигналу через регістр, компаратори, логічні елементи XOR, суматори та драйвери шини. Моделювання сигмоїдальної функції (3.2) та її кусково-лінійного наближення (3.9) показано на рисунку 3.2а, а їх похідних - на рисунку 3.2б.



а) б)

Рисунок 3.2 – а) – вигляд сигмоїдальної функції та її кусково-лінійної апроксимації; б) – вигляд похідних сигмоїдальної функції та її кусково-лінійної апроксимації

Розглянемо діапазон зміни вхідного сигналу $(-8,\ 8)$ і розіб'ємо його на $N_p = 1000$ інтервалів. У цьому діапазоні середня та максимальна похибки кусково-лінійного наближення сигмоїдальної функції рівні $\varepsilon_{ave} = 0.00587, \varepsilon_{\max} = 0.0185$, та її похідної $d\varepsilon_{ave} = 0.01412$, $d\varepsilon_{\max} = 0.07088$.

Абсолютна похибка між сигмоїдальною функцією та її кусочно-лінійним наближенням (суцільна крива) між похідними сигмоїдальної функції та її кусочно-лінійним наближенням (пунктирна крива) показана на рисунку 3.3.

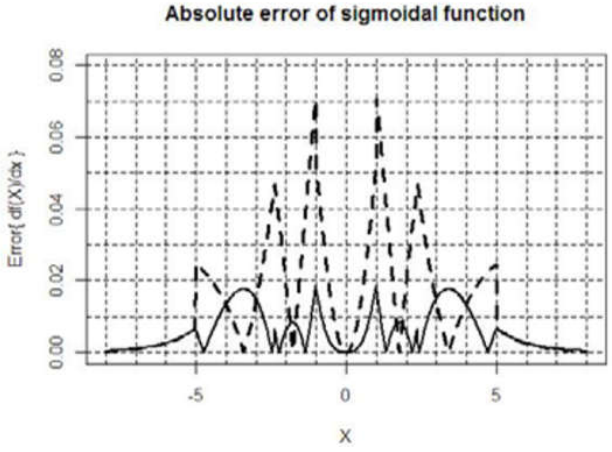


Рисунок 3.3. Вигляд абсолютної похибки між сигмоїдальною функцією і її кусково-лінійною апроксимацією та між їх похідними

Наближення сигмоїдальної функції (3.2) до кривої другого порядку. Для введення з діапазону $(0,\ x_{\max})$ сигмоїдальна функція апроксимується поліном $[2, 3]$:

$$\bar{f}(x) = c + bx + ax^2, \tag{3.13}$$

У рівнянні (3.13) невідомі коефіцієнти a, b, c, які визначаються методом найменших квадратів. Система рівнянь для їх обчислення виглядає так:

$$\begin{bmatrix} \sum_{i=0}^{N_p-1} x_i^2 & \sum_{i=0}^{N_p-1} x_i^3 & \sum_{i=0}^{N_p-1} x_i^4 \\ \sum_{i=0}^{N_p-1} x_i & \sum_{i=0}^{N_p-1} x_i^2 & \sum_{i=0}^{N_p-1} x_i^3 \\ N_p & \sum_{i=0}^{N_p-1} x_i & \sum_{i=0}^{N_p-1} x_i^2 \end{bmatrix} * \begin{bmatrix} c \\ b \\ a \end{bmatrix} = \begin{bmatrix} \sum_{i=0}^{N_p-1} f(x_i)x_i^2 \\ \sum_{i=0}^{N_p-1} f(x_i)x_i \\ \sum_{i=0}^{N_p-1} f(x_i) \end{bmatrix}, \tag{3.14}$$

Розбиваємо діапазон $(0, 4)$ на $N_p = 1000$ інтервалів. Використовуючи значення x_i та $f(x_i)$ в цих точках, формуємо систему рівнянь (3.14). Розв'язавши її отримаємо коефіцієнти a, b, c і вираз для апроксимації сигмоїдальної функції:

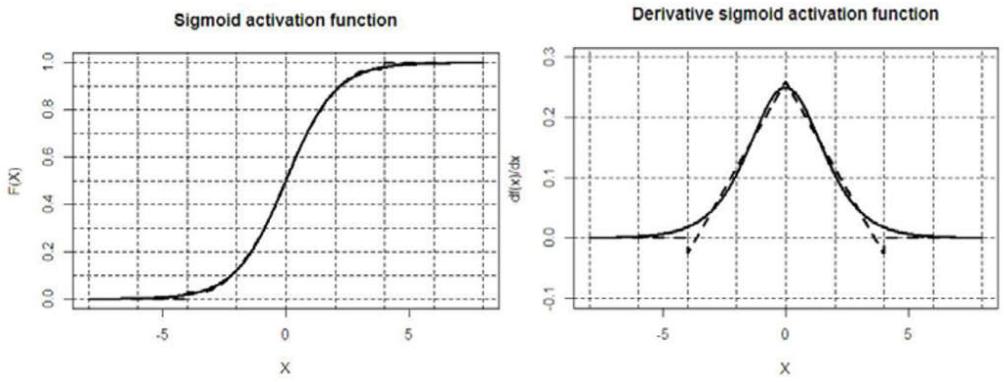
$$f(x) = \ f(x) = \begin{cases} 1, & x \geq 4.0 \\ -0.03577 * x^2 + 0.25908 * x + 0.5038, & 0 \leq x < 4.0 \end{cases}, \tag{3.15}$$

Цілочисельний вираз (3.15) має вигляд:

$$f(x) = f(x) = \begin{cases} 1024, & x \geq 4096 \\ -36 * 2^{-20} * x^2 + 265 * 2^{-20} * x + 515, & 0 \leq x < 4096 \end{cases}, \tag{3.16}$$

В діапазоні $(-8, 8)$ середні та максимальні похибки наближення сигмоїдальної функції поліному другого порядку рівні $\varepsilon_{ave} = 0.00426\ 0,\ \varepsilon_{\max} =$

0.01798 та її похідної – $d\varepsilon = 0.00769$, $d\varepsilon_{\max} = 0.04388$. Сигмоїдальна функція та її наближення до поліному другого порядку показані на рисунку 3.4а, а їх похідна - на рисунку 3.4б. Суцільна лінія на рисунку 3.4 показує сигмоїдну функцію та її похідну, а пунктирна лінія показує їх наближення.



а) б)

Рисунок 3.4 – а) – вигляд сигмоїдальної функції та її апроксимації поліномом другого порядку; б) – вигляд похідних сигмоїдальної функції та її апроксимації поліномом другого порядку

На рисунку 3.5 показані абсолютні похибки між сигмоїдальною функцією та її наближенням поліномом другого порядку (суцільною лінією) та між похідною сигмоїдальної функції та похідною її наближення (пунктирна лінія).

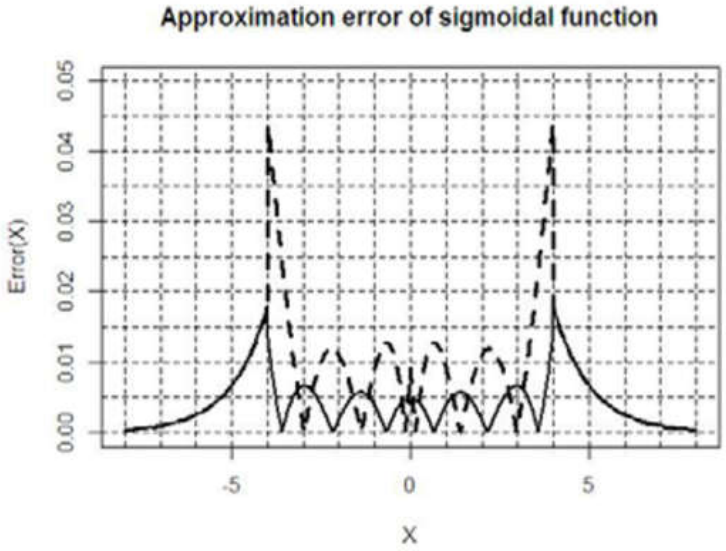


Рисунок 3.5 – Вигляд абсолютної похибки між сигмоїдальною функцією і її апроксимацією поліномом другого порядку та між їх похідними

Структурна схема пристрою, що реалізує сигмоподібну функцію (3.16), показана на рисунку 3.6.

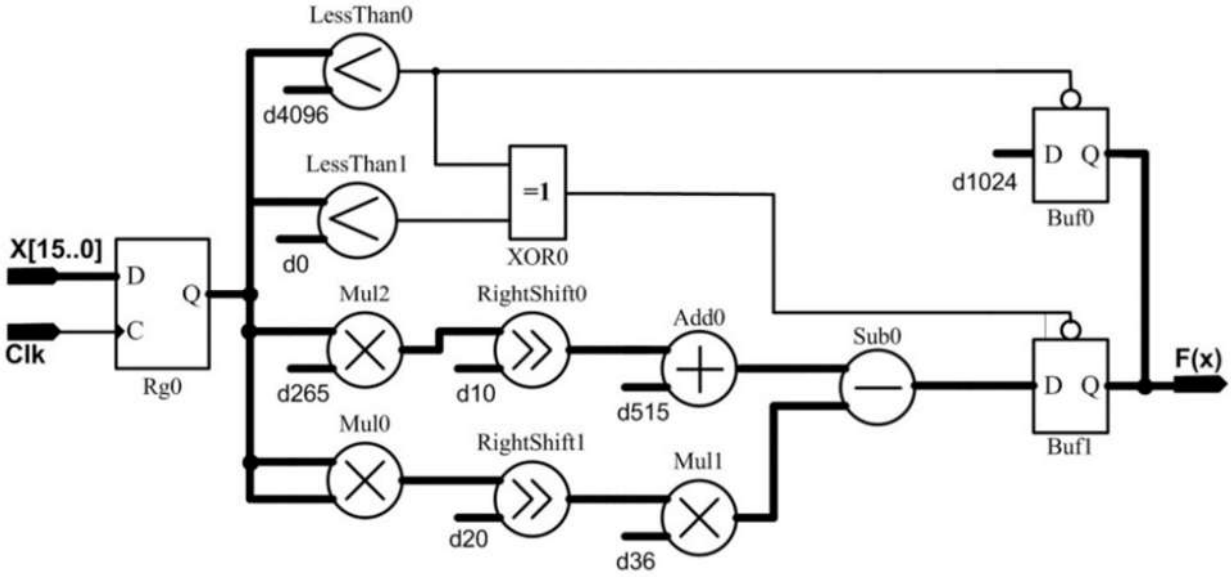


Рисунок 3.6 – Структурна схема пристрою, що реалізує сигмоїдальну функцію згідно виразу (3.16)

Цей пристрій використовує три мультиплікатори Mul, суматор Add, пристрій віднімань Sub, два ланцюга порівняння LessThan та два драйвери шини Buf для реалізації сигмоїдальної функції (3.2). Максимальний час обчислення сигмоїдальної функції в цьому пристрої визначається формулою:

$$t_2 = t_{Rg} + 2t_{Mul} + t_{Sub} + t_{Buf}, \tag{3.17}$$

де t_{Rg}, t_{Mul}, t_{Add}, t_{Buf} – час затримки проходження сигналу відповідно через регістр, перемножувач, пристрій віднімання та шинний формувач.

У [2,3] спрощений вираз використовується для наближення многочлена сигмоїдної функції другого порядку:

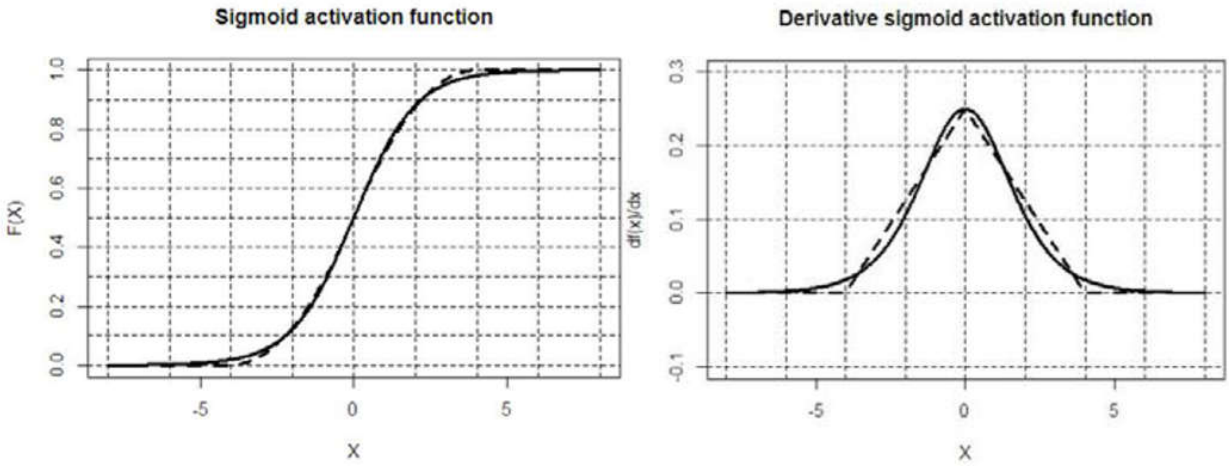
$$f(x) = \begin{cases} 1, & x \geq 4.0 \\ -0.03125 * x^2 + 0.25 * x + 0.5, & 0 \leq x < 4.0' \end{cases} \tag{3.18}$$

реалізація якого вимагає тільки одного перемножувача та суматорів.

В цілочисельному форматі вираз (3.18) має вигляд:

$$f(x) = \begin{cases} 1024, & x \geq 4096 \\ -2^{-15} * x^2 + x^{-2} * x + 512, & 0 \leq x < 4096 \end{cases}, \qquad (3.19)$$

Сигмоїдальна функція та її наближення за виразом (3.18) в діапазоні $(-8,8)$ показані на рисунку 3.7а, а похідні на рисунку 3.7б. На цих малюнках суцільна лінія показує сигмоїдну функцію та її похідну, а пунктирна лінія являє наближення функції за виразом (18) та її похідною.



а) б)

Рисунок 3.7 – Графіки а) – сигмоїдальної функції та її апроксимації;
б) –похідних сигмоїдальної функції та її апроксимації

Середні та максимальні похибки наближення сигмоїдальної функції за виразом (18) у діапазоні $(-8, 8)$ рівні $\varepsilon_{ave} = 0.00774, \varepsilon_{\max} = 0.02160$, та їх похідних $d\varepsilon_{ave} = 0.00877$, $d\varepsilon_{\max} = 0.02375$. Поява абсолютних помилок показано на рисунку 3.8 (суцільна лінія - похибка між сигмоїдальною функцією та апроксимуючим виразом (3.18), пунктирна лінія - похибка між їх похідними).

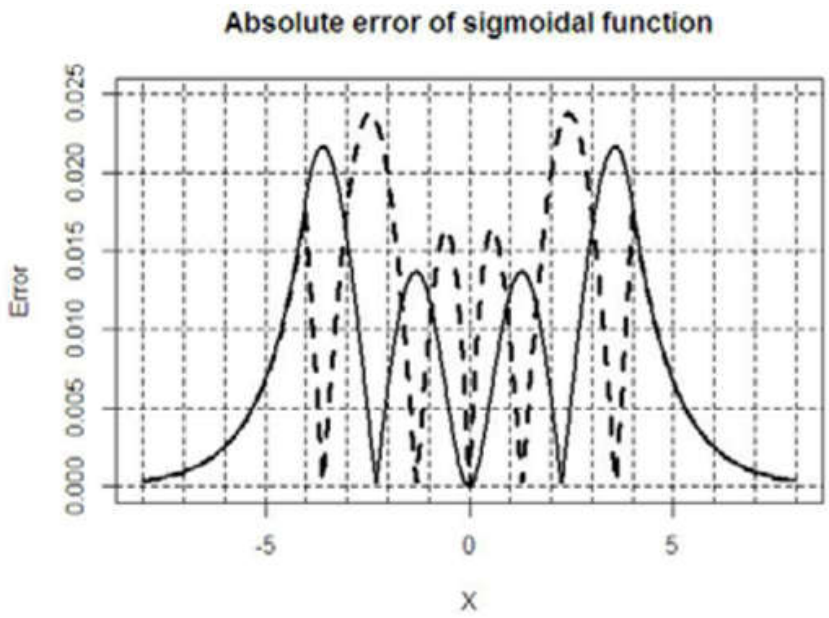


Рисунок 3.8 – Вигляд абсолютної похибки між сигмоїдальною функцією і її апроксимацією виразом (3.18) та між їх похідними

Для реалізації сигмоїдальної функції (3.19) розроблена блок-схема пристрою, яка показана на рисунку 3.9.

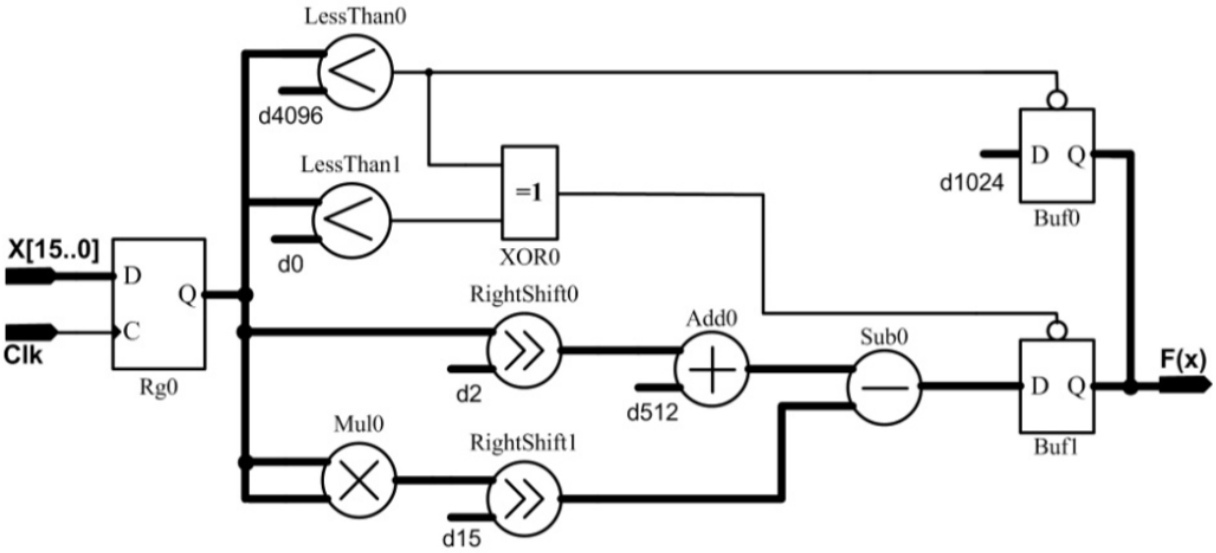


Рисунок 3.9 – Структурна схема пристрою, що реалізує сигмоїдальну функцію згідно виразу (3.19)

Цей апарат використовує множник, віднімання та суматор для здійснення сигмоїдальної функції. Порівняно з пристроєм на риунку 3.6, цей пристрій

потребує менших апаратних ресурсів для своєї реалізації. Максимальний час обчислення сигмоїдальної функції в цьому пристрої визначається формулою:

$$t_3 = t_{Rg} + t_{Mul} + t_{Add} + t_{Sub} + t_{Buf}, \quad (3.20)$$

3.4. Реалізація сигмоїдальних функцій на FPGA.

Реалізація сигмоїдальних функцій проводилася в середовищі розробки Quartus II для сімейства Cyclone III FPGA EP3C16F484C6 з використанням мови програмування VHDL та бібліотечних елементів Quartus II.

Для цього були використані методи наближення сигмоїдальної функції (вирази (10), (16), (19)) та структурні діаграми (рисунку 3.1, рисунку 3.6, рисунку 3.9), які реалізують ці вирази. Кожен із цих виразів дозволяє обчислити значення функції активації для позитивних значень зваженої суми, що подаються на її вхід. Для від'ємних значень суми вираз (3.4) використовується при обчисленні функції активації.

Розглянемо реалізацію структурної схеми, показаної на рисунку 3.9. На рисунку 3.10 зображено символ (FA_Sigm_N3) - поява пристрою для наближення сигмоїдної функції за виразом (3.19). Входами пристрою є: Clk - вхідна синхронізація, IN [15..0] - сума зважених входів нейрона на біт 16, а вихід - Out [15..0] - значення функції активації 16 біта. Обчислення сигмоїдальної функції виконується переднім краєм імпульсів, що надходять на вхід Clk.

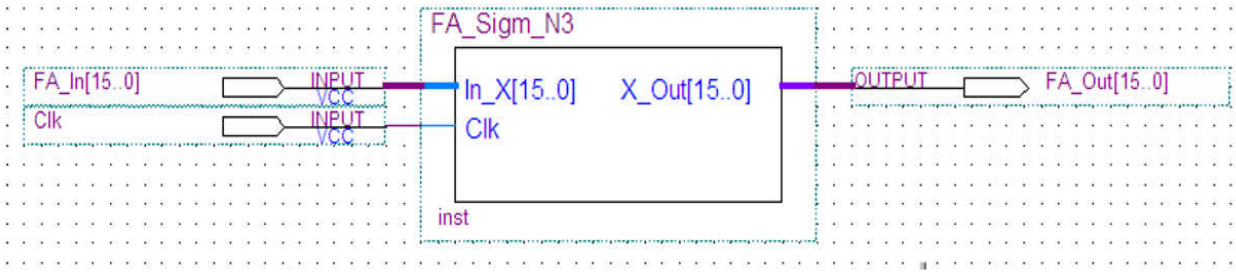
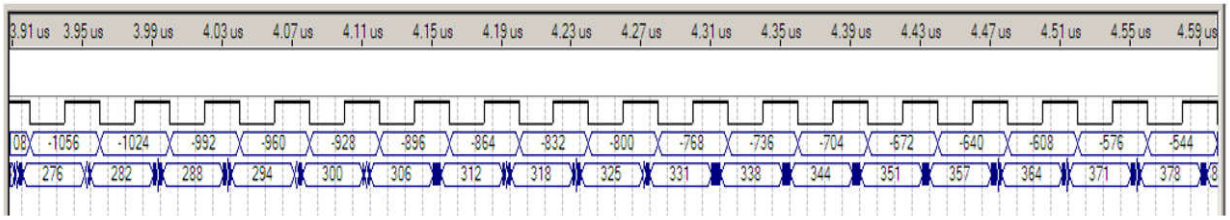


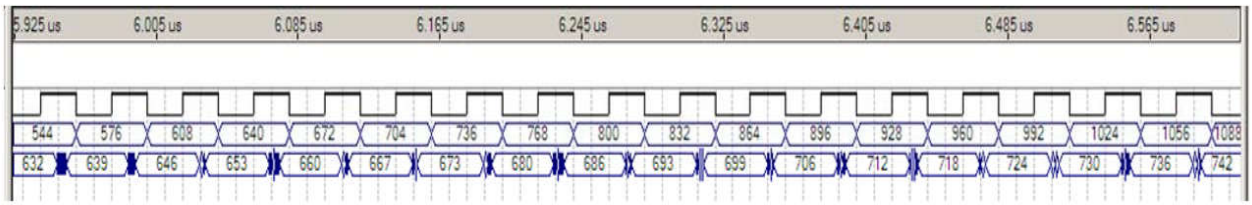
Рисунок 3.10 – Зовнішній вигляд символу FA_Sigm_N3

На рисунку 3.11 – показано фрагмент часової діаграми цього пристрою, який обчислює сигмоїдну функцію. Діаграма часу була отримана за допомогою моделювання часової області в середовищі розробки Quartus II.

На входи модуля FA_Sigm_N3 подається послідовність імпульсів синхронізації з періодом 20 нс і послідовністю вхідних сигналів від -4192 до 4192 в проміжку кроком 32. На його виході ми отримуємо значення апроксимованої сигмоїдальної функції. На рисунку 3.11а показаний фрагмент діаграми хронометражу зі вхідними значеннями в діапазоні від 544 до 1088, а на рисунку 3.11б - в діапазоні від -1056 до -544. Отже значення вхідного сигналу FA_In $=$ 640 відповідає значенню сигналу $\text{FA\_Out} = 512+2^{-3} * 640 - 2^{-15} * 640 * 640 = 660$, що співпадає зі значенням на часовій діаграмі на рис. 11а. Для від'ємного значення FA_In $=$ -640 згідно рис. 11б – FA_Out $=364$ $(1024 - 660 = 364)$. Затримка вихідного сигналу по відношенню до імпульсів синхронізації (t_{CO}) для FA_Sigm_N3 складає $17 \ldots 18$ ns.



а)



б)

Рисунок 3.11 – Часові діаграми роботи модуля сигмоїдальної функції FA_Sigm_N3

Діаграма символу FA_Sigm_N3 показана на рисунку 3.12. Тактові імпульси Clk та вхід In_X вводяться на вхід регістра регістру, для якого

обчислюється функція активації. Якщо вхід негативний, то на виході символу Код ми отримуємо їх додатковий код. Для позитивних значень вихідного коду вони подаються без змін. Розпізнавання позитивного та негативного введення здійснюється сигналом Sign_In. Якщо Sign_In = '1', тоді введення є позитивним.

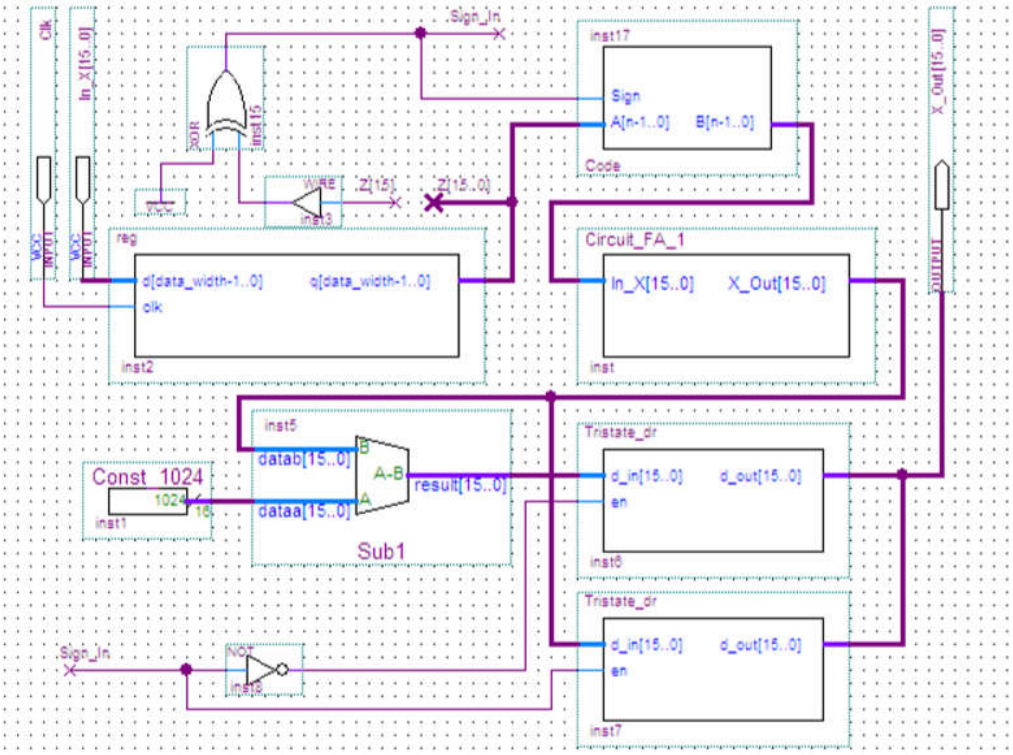


Рисунок 3.12 – Схема пристрою обчислення сигмоїдальної функції FA_Sigm_N3

З виводу символу Code вхід подається в калькулятор сигмоїдної функції (Circuit_FA_1) для додатних вхідних даних. З виходу Circuit_FA_1 дані поступають як на пристрій віднімання (Sub1), так і на буферний пристрій з третім станом (Tristate_dr). Для від'ємних вхідних значень (Sign_In='0') віднімання від постійної 1024 отриманого значення функції активації і отримана різниця подається в інший буферний пристрій. На його виході ми отримуємо значення наближеної сигмоїдальної функції.

Апаратні ресурси, необхідні для реалізації сигмоїдальних функцій, порівнюються за кількістю логічних елементів (ЛЕ) та FPGA виходів сімейства EP3C16F484C6 сімейства Altera Cyclone III. Максимальна кількість ЛЕ для цієї FPGA [4] становить 15408, а кількість виводів - 347. Для реалізації сигмоїдної

функції активації (FA_Sigm_N3) використано 135 ЛЕ та 33 виводи FPGA. Порівняння по точності апроксимації сигмоїдальної функції (3.2) виразами (3.9), (3.15), (3.18) та її похідної приведено в табл. 1. Похибка апроксимації сигмоїдальної функції цими виразами є меншою, ніж похибка апроксимації її похідних.

Таблиця 3.1 – Порівняння по точності апроксимацій сигмоїдальної функції та її похідних

Вираз	ε_{ave}	$\varepsilon_{\max}$	$d\varepsilon_{ave}$	$d\varepsilon_{\max}$
(3.9)	0.00587	0.01850	0.01412	0.07088
(3.15)	0.00426	0.01798	0.00769	0.04388
(3.18)	0.00774	0.02160	0.00877	0.02375

Сигмоїдальні функції реалізовані на стенді DE0 [5]. Зважена сума, що надходить на вхід пристрою FA_Sigm_N3, встановлюється за допомогою перемикачів, а значення функції активації відображається на 4-значному семисегментному індикаторі підставки.

ВИСНОВКИ

1. У роботі показано, що розробку компонентів для синтезу в режимі реального часу налаштованих GNM доцільно здійснювати при обробці інтенсивних потоків даних за допомогою складних алгоритмів паралелізацією та конвеєрністю обчислювальних процесів та використанням нових технологічних досягнень у галузі розробки суперінтегральних схем (VLSI).

2. Запропоновано ефективні компоненти переналаштованих технологій нейронної мережі будувати за такими принципами: співставленням інтенсивності потоку даних з обчислювальною потужністю; зменшенням кількість зовнішніх виходів і підвищенням продуктивності, за допомогою глибокої паралелізації до бітового рівня та використанням конвеєрних труб; скороченням часу та витрат на синтез нейронних мереж для конкретного застосування та розробкою алгоритмів розрахунку основних операцій штучних нейронних мереж на основі елементарних операцій, що дозволяє в повній мірі скористатися можливостями технології VLSI.

3. Показано, що основними шляхами підвищення ефективності використання переналаштованих технологій нейронної мережі є: вдосконалення методу переходу від алгоритму до структури апаратних компонентів штучних нейронних мереж; зміна довжини ходу конвеєра, кількості та біт каналів даних; вдосконалення методу розрахунку та основи елементарних операцій; комплексного підходу до розробки компонентів штучних нейронних мереж, на основі сучасних методів та алгоритмів навчання та функціонування нейронних мереж, сучасної елементної бази, нових алгоритмічних архітектур орієнтованих NVIS та схематичних рішень.

4. Визначено, що розробка компонентів для синтезу в режимі реального часу налаштованих GNM здійснюється шляхом формалізації процесу проектування компонентів, що скорочує час і витрати на проектування; розробкою ефективних пристроїв для обчислення кількості парних продуктів, орієнтованих на реалізацію VLSI; підвищенням ефективності використання

обладнання при розробці компонентів штучних нейронних мереж у режимі реального часу що дозволяє оптимізувати апаратно-часові параметри.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Березький О.М. Методичні рекомендації до виконання магістерської роботи з освітнього ступеня “Магістр”. Спеціальність: 123 - Комп’ютерна інженерія. Магістерська програма - Комп’ютерна інженерія" / О.М. Березький, Л.О. Дубчак, Г.М. Мельник /Під ред. О.М. Березького. Тернопіль: ТНЕУ, 2018. 41 с.

2. Методичні вказівки до оформлення курсових проектів, звітів про проходження практики, випускних кваліфікаційних робіт для студентів спеціальності «Комп’ютерна інженерія» / І.В. Гураль, Л.О. Дубчак / Під ред. О.М. Березького. Тернопіль: ТНЕУ, 2019. 33 с.

3. Куць І. С. Таблично-алгоритмічні засоби обчислення функцій активацій нейронних мереж / І. Г. Цмоць, І. С. Куць // І науково-практична конференція молодих вчених і студентів «інтелектуальні комп’ютерні системи та мережі» м. Тернопіль, 15 квітня 2019 р. 50 с.

4. Куць І. С. Дослідження та аналіз застосування спайкових і штучних нейронних мереж / І. Г. Цмоць, І. С. Куць // ІІ науково-практична конференція молодих вчених і студентів «інтелектуальні комп’ютерні системи та мережі» м. Тернопіль, 14 листопада 2019 р. 39 с.

5. Дипломне проектування за напрямами підготовки "Прикладна математика", "Комп’ютерна інженерія", "Програмна інженерія" [Текст]: навч.-метод. посіб. / Є.С. Сулема; за заг. ред. І.А. Дички. К.:НТУУ"КПІ", 2011. 224 с.

6. Загальні рекомендації з підготовки, оформлення, захисту й оцінювання випускних кваліфікаційних робіт здобувачів вищої освіти першого бакалаврського і другого магістерського рівнів / за ред. доц. М.І. Шинкарика. Тернопіль:ТНЕУ, 2018. 60с.

7. Beiu V, Peperstraete J.A., VandeM'olle J., Lamvereins R. Close Approximations of Sigmoid Functions by Sum of Steps for VLSI Implementation of Neural. URL:. Режим доступу: https://pdfs.semanticscholar.org/fdef/62a66787929bb80163219744d9eab041b203.pdf

8. Tonnniska ALT. Efficient digital implementation of the sigmoid function for reprogrammable logic. URL:. Режим доступу: https://pdfs.semanticscholar.org/9bf6/4bae3f8528cd5ac72a4ae869a74563ff6c26.pdf

9. Tisan Alin, Oniga Stefan, Mic Daniel, Buchman Attila. Digital Implementation of the Sigmoid Function for FPGA Circuits. URL. Режим доступу: http://users.utcluj.ro/~ATN/papers/ATN_2_2009_4.pdf

10. Cyclone III Device Handbook. URL. Режим доступу: http://www.altera.com/literature/hb/cyc3/cyc3_ciii51001.pdf

11. DEO User Manual. URL:. Режим доступу: http://esca.korea.ac.kr/teaching/FPGA_boards/DE0/DE0_User_Manual.pdf

12. Черняк Олександр Іванович. Інтелектуальний аналіз даних: Підручник / О.І. Черняк, П.В. Захарченко ; Київський національний університет ім. Т. Шевченка. К. : Знання, 2014. 599 с..

13. Руденко О.Г., Бодянський Є.В. Штучні нейронні мережі Навчальний посібник для студентів вищих навчальних закладів К.: СМІТ, 2006, 404 с.

14. Тимощук П.В. Штучні нейронні мережі Навчальний посібник/- Львів: Видавництво Львівської Політехніки, 2011. 444 стор.

15. Хайкин С. Нейронные сети. Полный курс 2-е изд. Пер. с англ. М.: Издательский дом "Вильямс", 2006. 1104 с.: ил.

16. Рашид Т. Создаем нейронную сеть. СПб.: Альфа-книга, 2017. 274 с.

17. Каллан Роберт. Основные концепции нейронных сетей. Пер. с англ. М. : Издательский дом "Вильямс", 2001. 287 с.

18. Рутковская Д., Пилиньский М., Рутковский Л. Нейронные сети, генетические алгоритмы и нечеткие системы. М.: Горячая линия Телеком, 2006. 452 с.

19. Джонс М. Тим. Программирование искусственного интеллекта в приложениях/ М. Тим Джонс ; Пер. с англ. Осипов А. И. М.: ДМК Пресс, 2004. 312 с: ил

20. Джон Д. Келлехер, Брайан Мак-Нейми, Аоифе д'Арси Основы машинного обучения для аналитического прогнозирования. Вильямс 2019 656 с.

21. Николенко С. И., Кадурин А. А., Архангельская Е. О. Глубокое обучение. Погружение в мир нейронных сетей. Питер: 2016 г., 480 с.

22. Ткаченко Р. О. та ін. Нейромережеві засоби штучного інтелекту Львів: Львівська політехніка. 2017 208 с.

23. Каниа Кан. Нейронный сети. Эволюция. ЛитРес: Самиздат. 2018 г., 380 с.

24. Галушкин А.И. Нейронные сети: основы теории. Научное издание 2017 г. 496 стр.

25. Рутковский Лешек. Перевод с польского И. Д. Рудинского Методы и технологии искусственного интеллекта. Научное издание. 2010 г., 520 ст.

26. Цмоць І.Г. Методи і структури ВІС для прискореного виконання базової операції алгортму швидкого перетворення Фур'є за основою два. Вісник ДУ "Львівська політехніка" 1998. №351.с.13-19.

27. Яцимирский М.Н. Алгоритмы быстрого преобразования Хартли по ращепленному основанию два четыре //Теоретическая электротехника. 1990 г.

28. Цмоць І.Г. Алгоритми і сруктури для ВІС перемножувеча комплексних чисел. . Вісник ДУ "Львівська політехніка" 1998 р. №327.с.231-240.

29. Цмоць І.Г. Принцип розробки і оцінка основних характеристик високопродуктивних процесорів на надвеликих інтегральни схем. Вісник ДУ "Львівська політехніка" 1998. №349. с.5-11.

30. Карцев М.А., Брик В.А. Вычислительные системы и синхронная арифметика.М., 1981. 359 с.

31. Цмоць І.Г. Алгоритми і структури ВІС перемножувача комплексних чисел. Вісник ДУ “Львівська політехніка”, 1998. №237, с.231-240.

32. Цмоць І. Г. Алгоритмічні операційні пристрої для обчислення базових операций алгоритмів швиткого перетворення Фур’є комплексної послідовності. Збірник наукових праць Інституту проблем моделювання в енергетиці НАН України, 1999. Випуск 2, с. 159-173.

33. Bodyanskiy Ye., Dolotov A., Vynokurova O. Evolving spiking wavelet-neuro-fuzzy self-learning system // Applied Soft Computing. 2014. 14. P. 252-258 с.

34. Bodyanskiy, Y., Vynokurova, O., Pliss, I., Peleshko, D., Rashkevych, Y. Hybrid generalized additive wavelet-neuro-fuzzy-system and its adaptive learning // Eds. Zamojski, W., Mazurkiewicz, J., Sugier, J., Walkowiak, T., Kacprzyk, J., Dependability Engineering and Complex Systems: Proceedings of the Eleventh International Conference on Dependability and Complex Systems DepCoS-RELCOMEX. June 27-July 1, 2016, Brunow, Poland. 2016. P. 51-61.

35. Haar, Alfréd (1910), "Zur Theorie der orthogonalen Funktionensysteme", Mathematische Annalen, 69 (3): 331–371.

36. Daubechies, Ten Lectures on Wavelets, SIAM, 1992, p. 194.

37. Галушкин А.И. Нейрокомпьютеры. Кн.3. М.: ИПРЖР, 2000. 528.

38. Рабинер Л., Гоулд В., Теория и применение цифровой обработки сигналов.(Пер.с англ. М.Мир, 1978-848с.).

39. Яцимирский М.Н. Швидкі алгоритми ортогональних тригонометричних перетворень. Львів: Академічний Експрес, 1997. 219 с.

40. Yatsymirskij M. Discrete Sine and Cosine and Patern Recognition/Proceedings International Conference. Kyiv,Ukraine. 1994. P.252-257.

41. Arro I., Smolyansky L., Trump T. Common Algoritmic and Structural Description of Short Discrete Fouriere Transform//Proceeding Latvian Signal Processimg International Conference Riga, 1990. V. I. P.118-122.

42. Методы синтеза быстрых алгоритмов свертки и спектрального анализа./Власенко В.А., Лапин Ю.М., Ярославский Л. П. М.:Наука,1990.180с.

43. A.A.Melnik, V.P.Kravets, I.G.Tzmotz. LSI for Signal Processors//Proceeding Latvian Signal Processind International Conference. Riga, April 24-26, 1990. V. I. P. 231-235.

44. Дейв Берсии. Снижение стоимости и повышение быстродействия ЦПС способствуют широкому внедрению этих приборов.Ýëåêòðоника(США), 1992, N 3-4

45. Цыфровые процесоры сигналов и их применение в измерительной технике. Приборы. Средства автоматизации и системы управления, 1989, Выпуск N6.

46. Ахо А., Хопкроф Дж., Ульман Дж. Построение и анализ вычислительных алгоритмов. М.: Мир, 1979. 539 с.

47. Coley J.W., Tykey J.W. An algorythm for machine calculation of complex Fourier series //Math. Comp. 1965. Vol. 19, N90. P.297-301.

48. Нуссбаумер Г. Быстрое преобразование Фурье и алгоритм вычисления сверток /Пер. с англ. Ю.Ф. Касимова, И.П. Пчелинцева; Под ред. В.М.Амербаева, Т.Э.Кренкеля. М.: Радио и связь, 1985. 248с.

49. Власенко В.А., Ярославский Л.П. Методы синтеза быстрых алгоритмов свертки и спектрального анализа сигналов. М.: Наука, 1990. 180с.

50. Duhamel P. Implementation of “split-radix” FFT algorithms for complex, real and real-summetric data // IEEE Trans. ASSP-34. 1986 N2. P. 285-295.

51. Rader C.M., Brenner N.N. F new principle for fast Fourier transformation // IEEE Trans. ASSP 34. 1986.

52. Брейсуэлл Р.Н. Преобразование Хартли: пер. с англю М: Мир, 1990. 175 с.

53. Яцимирский М.Н. Алгоритмы быстрого преобразования Хартли по основам два и четыре // Радиотехника. 1989.

54. Skarjune R. Derivation and implementation of an efficient fast Fourier transform algoritms (EFFT) // Computer and Chemistry. 1986.

55. Крот А.М., Минервина Е.Б. Алгоритмы быстрого преобразования Фурье для действительных и эрмитово-симметричных последовательностей // Радиотехника и электроника. 1989.