DOI: **0.33099/2311-7249/2024-50-2-77-84**

УДК: 629.735.05: 004.89: 004.942

Бігун Наталія Сергіївна

Центральний науково-дослідний інститут озброєння та військової техніки Збройних Сил України, Київ, Україна

УПРАВЛІННЯ РОЯМИ БЕЗПІЛОТНИХ ЛІТАЛЬНИХ АПАРАТІВ НА БАЗІ НЕЙРОННИХ ТРАНСФОРМЕРІВ

Сучасні виклики в управлінні автономними роями безпілотних літальних апаратів вимагають удосконалення інтелектуальних систем за допомогою передових технологій машинного навчання. Це дослідження зосереджене на інтеграції технології «Суміші експертів» (Mixture of Experts) у нейронні трансформери для підвищення ефективності роїв безпілотних літальних апаратів. Метою статті є аналіз і розробка методів, що дають змогу покращити можливості роїв безпілотних літальних апаратів шляхом удосконалення нейронних трансформерів за допомогою технології Mixture of Experts. Для досягнення цієї мети були використані методологічні підходи, що дають змогу розкрити, проаналізувати та провести експерименти з різними архітектурами Mixture of Experts і нейронними трансформерами. Зокрема – застосування спеціалізованих підмоделей (експертів) для рішення конкретних завдань у складі рою безпілотних літальних апаратів. У процесі дослідження було виявлено, що інтеграція Mixture of Experts із трансформерами сприяє значному підвищенню продуктивності роїв безпілотних літальних апаратів завдяки більш ефективному розподілу завдань й адаптації до мінливих умов середовища. Таке підвищення можливостей роїв безпілотних літальних апаратів відкриває нові перспективи для розвитку автономних розвідувальних та рятувальних місій. Елементи наукової новизни полягають у розробці оптимізованих фреймворків Mixture of Experts, які можуть бути результативно інтегровані в рої безпілотних літальних апаратів, з акцентом на подоланні обчислювальних і ресурсних обмежень у процесі одночасного підвищення адаптивності й можливостей прийняття рішень. Теоретична та практична значущість дослідження для воєнно-оборонної сфери і сфери технічних наук полягає у створенні автономних систем, здатних до швидкої адаптації й оперативного виконання завдань у різноманітних умовах. Дослідження відкриває нові шляхи та перспективи для розробки нових алгоритмів з метою точнішого визначення релевантності «експертів» до конкретних завдань, а також вивчення впливу різних архітектур «експертів» на загальну продуктивність системи.

***Ключові слова**: рій безпілотних літальних апаратів, Mixture of Experts, нейронні трансформери, автономні системи, обробка даних.*

Вступ

Постановка проблеми. З огляду на стрімкий прогрес у сфері штучного інтелекту і зростаючу потребу в автоматизації та оптимізації обробки даних, питання вдосконалення нейронних мереж стає все більш актуальним. Серед основних викликів, з якими стикаються науковці та розробники, є необхідність забезпечення високої точності обробки даних за одночасного зниження обчислювальних витрат та підвищення швидкості реагування систем. Це питання є актуальним і у контексті управління роями безпілотних літальних апаратів (далі – БпЛА), де ефективність обробки даних безпосередньо впливає на успішність виконання місій та безпеку польотів.

Традиційні архітектури нейронних мереж не завжди досконало підходять для складних завдань обробки даних. Їх масштабованість та адаптивність часто обмежені, що робить їх неефективними для задач, які потребують виконання конкретних завдань з високим ступенем точності. Концепція «Суміші експертів» (Mixture of Experts (далі – MoE)) пропонує нове рішення для цієї проблеми [1]. MoE розподіляє завдання між спеціалізованими підмережами (експертами), що робить моделі більш адаптивними та продуктивними [2]. Завдяки використанню MoE кожен БпЛА або підгрупа в рої може зосередитись на виконанні окремих завдань під час проведення операцій. Проте, застосування технології MoE в управлінні роями БпЛА потребує глибокого аналізу та розробки оптимізованих підходів, здатних забезпечити високу продуктивність системи при цьому зберігаючи обчислювальну ефективність.

Отже, постановка проблеми дослідження зводиться до необхідності розробки та аналізу ефективних методів інтеграції технології MoE в архітектуру нейронних мереж для оптимізації роботи роїв БпЛА. Це охоплює вивчення способів підвищення адаптивності та спеціалізації моделей, здатних обробляти широкий спектр завдань у динамічних умовах, а також розробку стратегій оптимального розподілу обчислювальних ресурсів.

Результати такого дослідження матимуть важливе теоретичне та практичне значення, сприяючи подальшому розвитку автономних систем та підвищенню їх ефективності в реальних умовах експлуатації.

Аналіз останніх досліджень і публікацій. Попередні дослідження в галузі штучного інтелекту та машинного навчання акцентують на значному потенціалі архітектур MoE, зокрема, їх здатності оперативно обробляти великі обсяги даних. Аналіз нещодавніх досліджень у галузі інтеграції технології MoE в архітектуру трансформерів підкреслює потенціал цього підходу для вирішення проблем з якими стикаються рої БпЛА.

Комплексну теоретичну основу для розуміння моделей заклали MoE Gormley та Frühwirth-Schnatter [3], підкресливши їх універсальність і корисність у різних сферах застосування. Їх робота наголошує на потенціалі MoE для обробки неоднорідності параметрів і кластеризації даних, створюючи теоретичну основу для подальших досліджень у спеціалізованих системах, таких як БпЛА. Використовуючи цю теоретичну базу, Krishnamurthy, Watkins і Gaertner, [4] зайнялися оптимізацією декомпозиції завдань і використання «експертів» в межах MoE. Вони запропонували нову архітектуру шлюзів і метод регуляризації, спрямований на підвищення продуктивності та спеціалізації «експертів».

Jawahar та ін. [5] розширили застосування MoE до сфери нейронного машинного перекладу, продемонструвавши корисність Neural Architecture Search (NAS) у розробці ефективних, гетерогенних моделей MoE в умовах обчислювальних обмежень. Це дослідження підкреслює значний потенціал гетерогенного дизайну MoE з адаптивними обчисленнями для знаходження оптимальних підмереж, які зменшують обчислювальні витрати та розмір моделі, зберігаючи при цьому аналогічну продуктивність завдань, що є вкрай важливим для розробки оптимізованих систем управління роями БпЛА.

Antoniak та співавтори [6] розробили модель «Mixture of Tokens», яка уникає деяких проблем, пов'язаних з MoE, таких як нерівномірне використання «експертів» та нестабільність навчання, шляхом змішування токенів з різних прикладів перед подачею їх до «експертів». Це дозволяє моделі вчитися від усіх комбінацій токен-експерт, зберігаючи при цьому переваги архітектури MoE. Mohammadi H., Nazerfard E. та Firoozi T., [7] запропонували нову трансформер-базовану архітектуру MoE для розпізнавання насильства у відео, що демонструє переваги архітектури трансформера та знижує вартість використання великих трансформерів завдяки інтелектуальному поєднанню великих та функціональних трансформерних архітектур.

Нарешті, Puigcerver, Riquelme, Mustafa і Houlsby [8] підійшли до вирішення проблем нестабільності навчання і масштабованості в моделях MoE, розробивши Soft MoE. Ця модель, яка виконує неявне (Soft) призначення завдань «експертам», являє собою метод збільшення пропускної здатності моделі без пропорційного збільшення обчислювальних витрат.

Географічне розмаїття дослідницької спільноти відображає глобальну зацікавленість у розвитку технології MoE. Нові тенденції, що з'являються в результаті цих досліджень, містять у собі зосередження на оптимізації моделей MoE для підвищення ефективності та гнучкості, подоланні обчислювальних обмежень і розширенні можливостей прийняття рішень в складних системах. Однак, у практичному застосуванні моделей MoE в роях БпЛА, особливо побудованих на нейронних трансформерах, залишаються певні прогалини, що вказує на важливу невирішену проблему в цій галузі. Незважаючи на успіх багатошарових перцептронів (Multi-Layer Perceptron (далі – MLP)), дослідники шукають нові архітектури нейронних мереж, які б вирішували певні проблеми, пов'язані з точністю та інтерпретованістю. Одним із перспективних напрямів є мережі Колмогорова-Арнольда (далі – KAN), які пропонують альтернативу MLP [9].

Отже, попередні дослідження підкреслюють важливість розробки оптимізованих фреймворків MoE, які можуть бути ефективно інтегровані в рої БпЛА, і являють собою актуальну дослідницьку потребу. Ця стаття спрямована на розв'язання цих проблем шляхом дослідження інтеграції технології MoE на базі трансформерів у рій БпЛА. Тому, **метою** статті є аналіз та розробка методів, що дають змогу покращити можливості роїв безпілотних літальних апаратів шляхом вдосконалення нейронних трансформерів за допомогою технології Mixture of Experts.

Виклад основного матеріалу дослідження

Це дослідження розкриває інтеграцію моделей MoE в операційну структуру роїв БпЛА, підкріплену передовими можливостями нейронних трансформерів. Методологія, прийнята для цього теоретичного дослідження, окреслює комплексний підхід, що охоплює вибір моделей MoE, тонкощі специфікацій нейронних трансформерів і операційну динаміку роїв БпЛА.

Рій БпЛА являє собою мережу взаємопов'язаних дронів, кожен з яких оснащений датчиками та комунікаційними модулями, що дозволяють обмінюватися даними та приймати рішення в реальному часі (рис. 1) [8]. Водночас, роль «експерта» може виконувати кластер з кількох БпЛА або окремий БпЛА у рої.

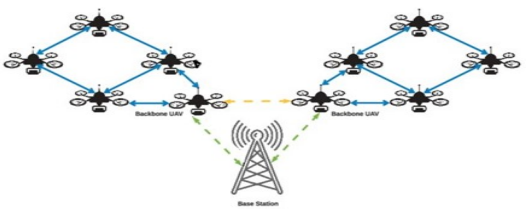

Рисунок 1 – Варіант мережі рою БпЛА

Ця система призначена для автономної роботи, а моделі MoE побудовані на нейронних трансформерах сприяють інтелектуальному розподілу і виконанню завдань між роєм. Mixture of Experts – це техніка машинного навчання, яка поєднує кілька «експертних» моделей для підвищення продуктивності (рис. 2) [1].

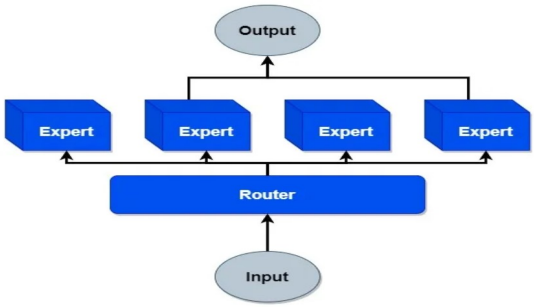

Рисунок 2 – Приклад моделі Mixture of Experts

«Експерти», спеціалізовані на обробці візуальних даних, здатні аналізувати зображення та відео для виявлення об'єктів та моніторингу змін. Зокрема, застосування «експертами» модифікованих моделей U-Net та PSP для обробки зображень, знятих роями БпЛА, дасть змогу автономно виявляти та класифікувати об'єкти на великих територіях з високою точністю [10]. Спеціалізація на навігаційних даних дозволяє оптимізувати маршрути польоту, забезпечуючи ефективне уникнення перешкод та адаптацію до змінних умов. Обробка метеорологічних даних у реальному часі дозволяє підвищити безпеку польотів завдяки коригуванню планів польоту до погодних умов. Оптимізація комунікації між БпЛА в рої забезпечує синхронізацію дій та автоматизоване розподілення завдань.

Завдяки MoE, множина нейронних мереж, подібно до команди фахівців, об'єднує свої знання та навички для обробки складних даних, що значно розширює можливості та покращує результати [11]. Ця модель викликає зацікавленість завдяки її гнучкості та масштабованості, що є важливими факторами для управління роєм БпЛА в різноманітних середовищах.

З огляду на базову методологію MoE, доцільно дослідити конкретну реалізацію шару MoE (рис. 3) [12].

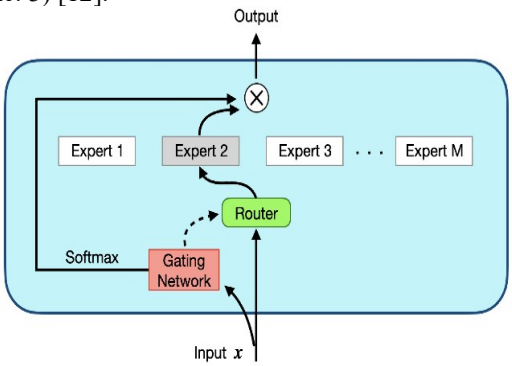

Рисунок 3 – Шар моделі Mixture of Experts

Шар MoE, являє собою складний архітектурний компонент, призначений для посилення обчислювальних можливостей нейронних мереж шляхом впровадження механізму динамічного відбору декількох «експертних» систем. Кожен «експерт» у шарі MoE спеціалізується на різних параметрах або ознаках вхідних даних, що дозволяє проводити детальний і комплексний аналіз, чого не можна було б досягти за допомогою єдиної монолітної моделі.

Основним механізмом, який дає змогу шару MoE послідовно вибирати найбільш відповідного «експерта» для кожного входу, є мережа шлюзів (gating network) або процедура маршрутизації. Процедура маршрутизації оцінює вхідні дані та визначає релевантність кожного «експерта» до конкретної точки даних. Потім вона присвоює вагу результатам роботи «експертів», фактично вирішуючи, який внесок має зробити кожен «експерт» у кінцевий результат роботи шару MoE. Цей процес дає змогу моделі використовувати спеціалізовані знання кожного «експерта», що призводить до збільшення продуктивності під час виконання складних завдань.

На рисунку 4 схематично зображено нейромережеву модель, що дає уявлення про практичну реалізацію та потенційні переваги технології MoE [13].

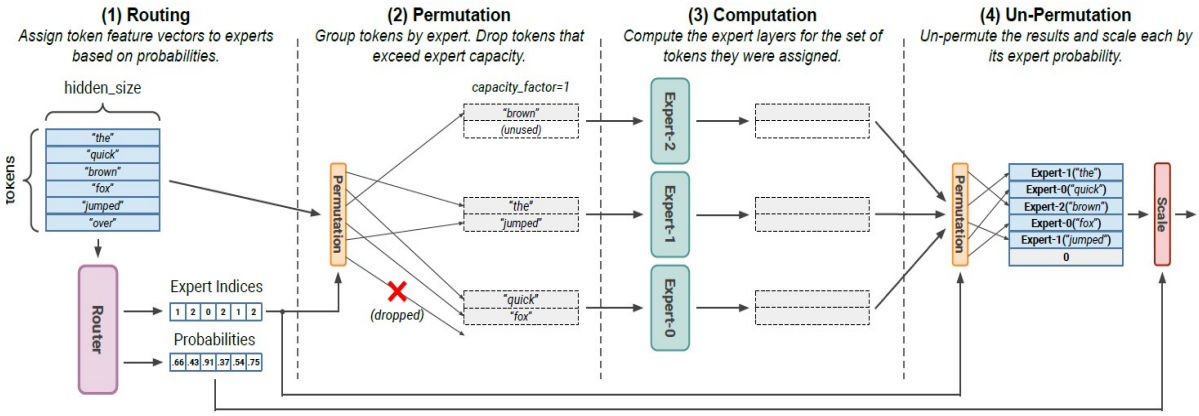

Рисунок 4 – Концепція Mixture of Experts

Рисунок 4 демонструє приклад інтеграції спеціалізованих шарів MoE в архітектуру більшої нейронної мережі, такої як трансформерна або схожої архітектури для обробки послідовностей даних. В ній замість єдиного повнозв'язного шару, який застосовується однаково до всіх входів, тут використовується набір «експертів» – менших нейронних мереж, кожна з яких спеціалізується на певному типі входів.

Таке інтегрування шарів MoE дозволяє нейронній мережі досягати вищої точності та продуктивності, а також забезпечує її масштабованість, оскільки в процесі обробки активуються лише відповідні «експерти», вдосконалюючи використання обчислювальних потужностей. Адаптивне управління обчисленнями зменшує зайве навантаження, підвищуючи ефективність системи [14]. Також, завдяки MoE, система набуває більшої прозорості у своїх рішеннях, оскільки кожен «експерт» спеціалізується на конкретному виді даних.

Технологію MoE доцільно застосовувати на базі нейронного трансформера [15], де вхідні дані проходять попередню обробку. Результативність роботи мережі MoE значною мірою залежить від попереднього процесу навчання, що є обов'язковою умовою для визначення релевантності, розподілу завдань між «експертами» та спеціалізації «експертних» мереж. Цей процес дає змогу мережі шлюзу оцінювати вхідні дані та в режимі реального часу визначати, які «експерти» володіють відповідними знаннями для обробки того чи іншого типу інформації, адаптуючи розподіл навантаження та ресурсів до потреб системи. Водночас, «експертні» мережі навчаються фокусуватися на певних аспектах даних, розвиваючи глибокі знання та навички в своїх областях, що сприяє підвищенню загальної точності та ефективності системи MoE у вирішенні завдань обробки даних.

Нейронні трансформери (рис. 5), що являють собою обчислювальну основу системи, призначені для використання механізмів «самоуваги», що дає змогу обробляти послідовні вхідні дані з високою точністю та мінімальною затримкою [16–17].

Трансформери мають стати невід'ємною частиною інтерпретації складних масивів даних, з якими стикаються рої БпЛА під час польотів, та реагуванням на них. Архітектура трансформера відзначається здатністю продуктивно обробляти послідовні дані, використовуючи механізм «самоуваги» для аналізу важливості різних сегментів вхідних даних, завдяки чому моделі пристосовуються до обставин без потреби в циклічних або згорткових структурах. Складаючись з енкодера та декодера, кожен з яких містить шари «самоуваги», прямого зв'язку та нормалізації, а також залишкові з'єднання, трансформер проводить глибоке навчання з паралельною обробкою даних, що є особливо корисним для обробки великих масивів даних.

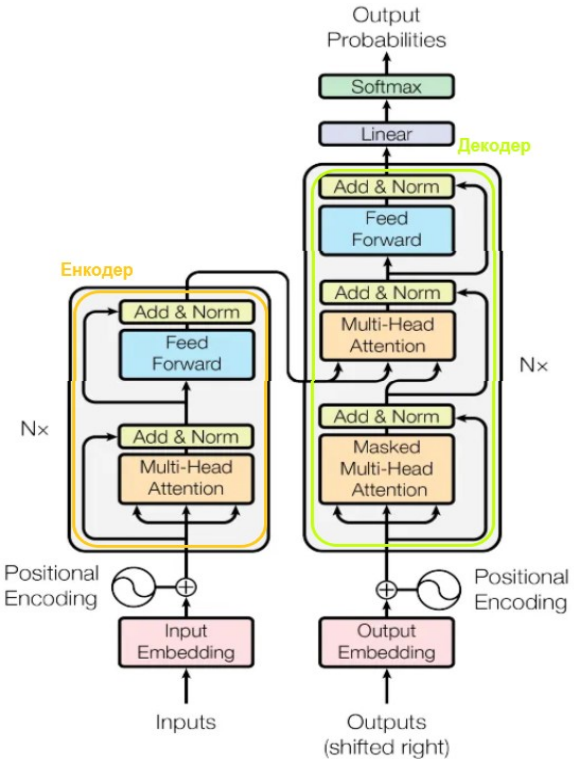

Рисунок 5 – Архітектура трансформера [17]

Архітектура трансформера є придатною основою для втілення технології Mixture of Experts, з огляду на здібність механізму уваги до детального розгляду контекстуальної інформації та паралельної обробки даних, що в свою чергу забезпечує оптимальне розподілення обчислень між «експертами». Це дозволяє швидко делегувати обробку даних спеціалізованим «експертам» в рамках моделі.

На рисунку 6 наведено модифікацію архітектури трансформера шляхом інтеграції шарів Mixture of Experts, що замінюють кожен другий повнозв'язний шар трансформерного енкодера, аналогічні зміни вносяться і в декодер. За масштабування на кілька пристроїв, як показано в третьому блоці рисунку 6, шар MoE розподіляється між пристроями, інші ж шари повністю дублюються на всіх пристроях. Такий підхід до розподілу MoE шару дає змогу ефективно використовувати обчислювальні ресурси, розподіляючи навантаження таким чином, що кожен пристрій вносить вклад в обробку вхідних даних на основі розміщених на ньому «експертів». Це не лише покращує здатність моделі обробляти великі обсяги даних, але й оптимізує використання обчислювальних ресурсів [18].

Враховуючи здатність моделі до масштабування та ефективного розподілу обчислювальних ресурсів, інтеграція MoE-шарів у трансформери відкриває нові можливості для розробки високопродуктивних систем спільної обробки даних для роїв БпЛА, що може сприяти покращенню якості виконання завдань та оптимізації використання ресурсів.

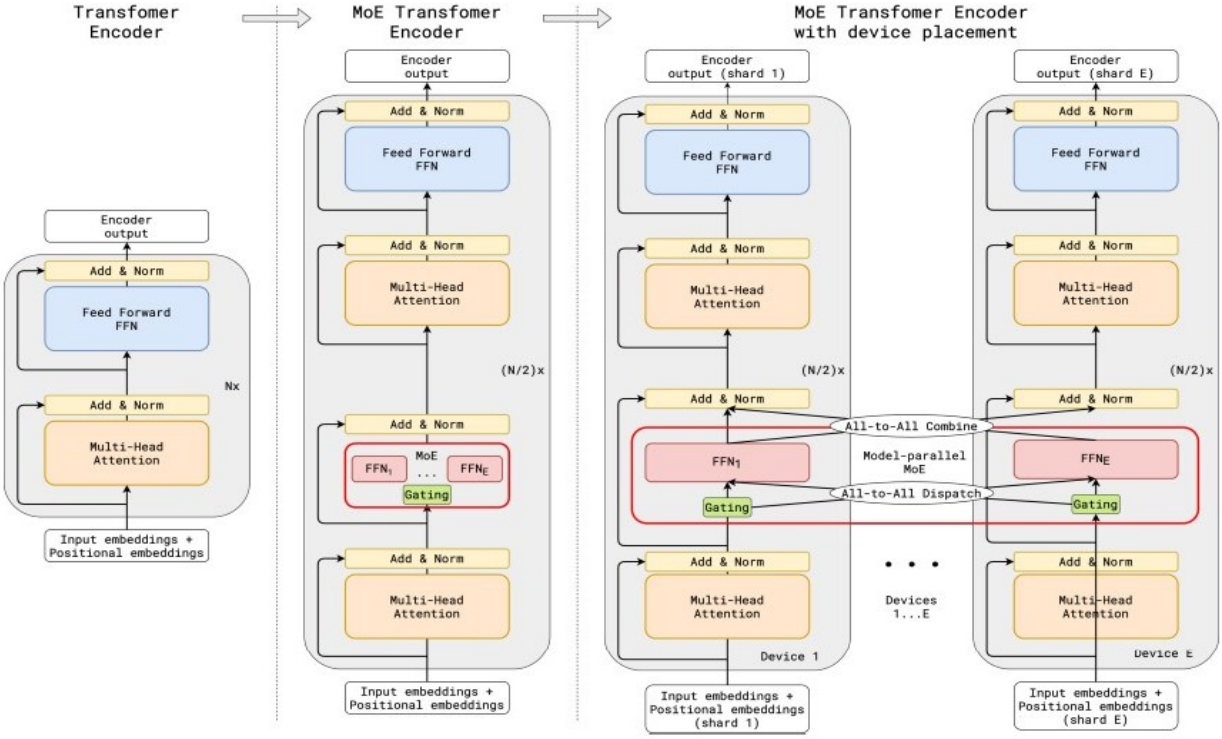

Рисунок 6 – Масштабування енкодера трансформера шляхом інтеграції шарів Mixture of Experts [14]

Масштабованість системи, забезпечена через динамічний розподіл обчислень між «експертами», сприяє раціональному використанню обчислювальних ресурсів, знижуючи навантаження на систему та підвищуючи її продуктивність. Окрім технічних переваг, MoE збільшує інтерпретованість моделі, оскільки кожен «експерт» фокусується на обробці специфічного виду даних, що спрощує аналіз та налаштування системи. Отже, застосування MoE у керуванні роями БпЛА становить значний крок вперед у розвитку автономних систем, забезпечуючи високу адаптивність та ефективність виконання завдань в різноманітних умовах.

Проблема обмеженої пропускної здатності каналів зв'язку та втрати даних є значним викликом у сфері використання технології роїв БпЛА. Ця проблема виникає через обмежені обчислювальні ресурси та обмеження потужності на борту БпЛА, а також через необхідність забезпечення стійкості до збоїв і несприятливих умов навколишнього середовища. При цьому, система передачі даних між БпЛА у рої та між БпЛА та наземними станціями вимагає ефективного використання доступного радіочастотного спектру, а також високої надійності зв'язку для передачі зібраних даних в реальному часі.

Однак, в умовах обмеженої пропускної здатності, велика кількість даних, генерованих БпЛА, може призводити до переповнення каналу зв'язку, що, в свою чергу, може спричинити затримки у передачі даних або їх втрату. Це особливо критично для завдань, що вимагають швидкого реагування або збору даних у мінливих або непередбачуваних умовах. Крім того, в умовах високого рівня шуму або завад, що часто зустрічаються в урбанізованих середовищах, ефективність зв'язку може бути знижена, що посилює ризик втрати важливих даних.

Застосування технології MoE передбачає забезпечення надійної передачі даних між БпЛА та наземними станціями без втрат. Для цього пропонується підхід, в якому закодовані дані будуть представлені 2T точками даних в латентному просторі, що формується на виході енкодера та вході декодера, розбиваються на косинусоїдальну та синусоїдальну складові, T(I) і T(Q) відповідно. Ця концепція нагадує квадратурну амплітудну модуляцію (Quadrature Amplitude Modulation (QAM)), яка використовує як амплітудні, так і фазові зміни в радіосигналах для підвищення ефективності використання смуги пропускання системи [19].

Специфіка цих компонентів забезпечує їхню взаємну незалежність, що дозволяє зменшити вплив шуму в процесі відновлення даних на наземній станції. Передаючи обидві компоненти одночасно, система досягає вищої швидкості передачі даних, зберігаючи при цьому компактне представлення даних, що передаються. Підвищення ефективності досягається завдяки ортогональності косинусоїдальної та синусоїдальної функцій. Ця властивість гарантує, що обидва компоненти можуть бути окремо передані та точно реконструйовані на стороні приймача. Як результат, система зменшує вплив шуму та ще більше підвищує надійність та ефективність передачі даних.

Хоча трансформери побудовані за принципом

MLP довели свою ефективність у багатьох задачах, вони мають ряд обмежень. Як згадувалося вище, перспективним напрямом, який пропонує альтернативний спосіб наближення функцій, є використання мереж KAN. На відміну від MLP, які використовують фіксовані функції активації, KAN дають змогу навчати функції активації на ребрах («вагах») нейронної мережі (рис. 7). Це дозволяє KAN досягати вищої точності за меншої кількості параметрів, ніж MLP, оскільки їх можливо налаштувати для оптимального вирішення конкретної задачі [9].

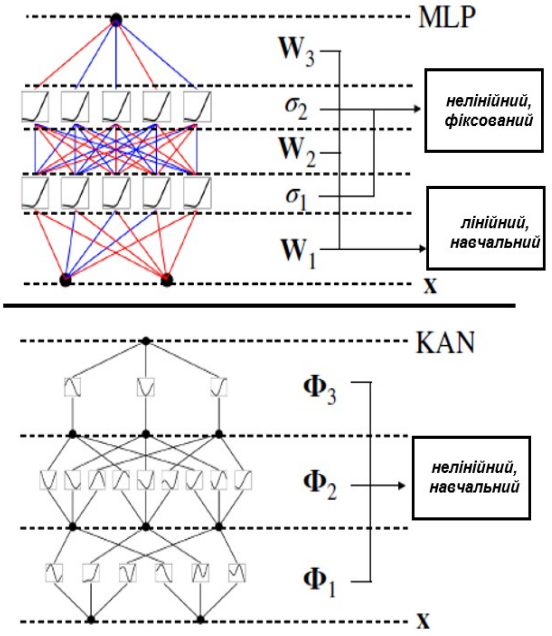

Рисунок 7 – Порівняння структурних схем багатошарових перцептронів (MLP) та мереж Колмогорова-Арнольда (KAN) [9]

Крім того, KAN мають перевагу в інтерпретованості. Їх структуру легко візуалізувати і кожен параметр має чітке значення в контексті наближення функції. Ця властивість робить KAN перспективними для застосування в наукових дослідженнях, де важливо розуміти, як мережа приймає рішення, і використовувати її знання для подальшого аналізу.

Висновки й перспективи подальших досліджень

У статті вивчено і висвітлено інтеграції технології Mixture of Experts в архітектуру трансформерів та її застосуванню для покращення роботи роїв безпілотних літальних апаратів. Отримані результати свідчать про те, що використання Mixture of Experts веде до суттєвого підвищення ефективності обробки даних. Це досягається завдяки спеціалізації «експертів» на різних типах даних, що, в свою чергу, сприяє зростанню якості та швидкості виконання завдань. Масштабованість та адаптивність, які забезпечує Mixture of Experts, роблять цю технологію практично корисною для застосування в роях безпілотних літальних апаратів, де системи потребують гнучкості та здатності до швидкої адаптації до змінних умов.

Викладене у статті також підкреслюється теоретична та практична цінність інтеграції Mixture of Experts з архітектурою трансформерів. Інтеграція не лише покращує обробку даних, але й забезпечує масштабованість та ефективне використання обчислювальних ресурсів. Це відкриває нові можливості для розробки високопродуктивних систем кооперативної обробки даних для роїв безпілотних літальних апаратів. Такі системи можуть сприяти покращенню виконання завдань та оптимізації використання ресурсів. З огляду на вищезазначене, можна констатувати, що поставлені цілі дослідження були досягнуті.

Трансформери засновані на багатошарових перцептронах (MLP), які ефективно навчаються на великих наборах даних. Проте, існують альтернативні архітектури, наприклад, мережі Колмогорова-Арнольда (KAN), які можуть навчатися з меншою кількістю даних і мати більшу інтерпретованість. Перспективами подальших досліджень є вивчення можливості інтеграції мереж Колмогорова-Арнольда (KAN) в трансформери, що може значно збільшити їхню ефективність у різних задачах.

Список бібліографічних посилань

1. Al-Emadi S., Al-Mohannadi A. Towards Enhancement of Network Communication Architectures and Routing Protocols for FANETs: A Survey. *2020 3rd International Conference on Advanced Communication Technologies and Networking (CommNet)*, Marrakech, Morocco, 4–6 Sept. 2020. DOI:10.1109/commnet49926.2020.9199627.

2. Слюсар В. Кібернетичні загрози великих мовних моделей. *Системи та засоби штучного інтелекту* : тези доп. Міжнар. наук. конф. «Штуч. інтелект: досягнення, виклики та ризики», м. Київ, 15–16 берез. 2024 р. С. 252–260. **3. Срібнюк С. М**. Гідравлічні та аеродинамічні машин. Основи теорії і застосування: навч. посібник. Київ : Центр навч. літератури, 2022. 328 с. **4. Shazeer N., Mirhoseini A., Maziarz K., Davis A., Le Q. V., Hinton G. E. and Dean, J.** Outrageously large neural networks: the sparsely-gated mixture-of-experts layer. *CoRR.* 2017. *a*bs/1701.06538. URL: http://arxiv.org/abs/1701.06538 (Accessed: 12 May 2024). **5. Jawahar G., Mukherjee, S., Liu X., Kim Y. J., Abdul-Mageed M., Lakshmanan L. V. S., Awadallah A. H., Bubeck S. and Gao J.** AutoMoE: heterogeneous mixture-of-experts with adaptive computation for efficient neural machine translation. *Annual meeting of the Association for Computational Linguistics*. 2023. URL: https://api.semanticscholar.org/CorpusID:259108418 (Accessed: 12 May 2024) **6. Antoniak S., Jaszczur S., Krutul M.** Mixture of Tokens: Efficient LLMs through Cross-Example. *ArXiv.* 2023. abs/2310.15961. DOI: 10.48550/arXiv.2310.15961.

7. Mohammadi H., Nazerfard E., Firoozi T. Reinforcement learning-based mixture of vision transformers for video violence recognition. *ArXiv.* 2023.

abs/2310.03108. DOI: 10.48550/arXiv.2310.03108. **8. Puigcerver J., Riquelme C., Mustafa B. and Houlsby N.** From sparse to soft mixtures of experts. *ArXiv*. 2023. DOI:10.48550/arXiv.2308.00951. **9. Liu Z., Wang Y., Vaidya S., Ruehle F., Halverson J., Soljacic M., Hou T.Y., and Tegmark M.** KAN: Kolmogorov-Arnold Networks. *ArXiv*. 2024. abs/2404.19756. DOI:10.48550/arXiv.2404.19756. **10. Slyusar V., Sliusar I., Bihun N. and Piliuhin V.** Segmentation of analogue meter readings using neural networks. *MoMLeT+DS*. 2022. **11. Bouaouni Y.** Mixture of Experts. *Medium*. URL: https://medium.com/@yacinebouaouni07/mixture-of-experts-26243919d145 (Accessed: 12 May 2024). **12. Zhou Z.-H.** Ensemble Methods: Foundations and Algorithms. Chapman and Hall/CRC. 2012. DOI: 10.1201/b12207. **13. Gale T., Narayanan D., Young C., Zaharia M. A.** MegaBlocks: Efficient Sparse Training with Mixture-of-Experts. *ArXiv*. 2022. abs/2211.15841. DOI:10.48550/arXiv.2211.15841. **14. Slyusar V., Kondratenko Y., Shevchenko A., Yeroshenko T.** Some Aspects of Artificial Intelligence Development Strategy for Mobile Technologies. *Journal of Mobile Multimedia*. 2024. P. 525–554. DOI: 10.13052/jmm1550-4646.2031. **15. Ajay L.** An intuitive introduction to Transformers. *Medium*. URL: https://lakshmi1212.medium.com/an-intuitive-introduction-to-transformers-6f574c8e7df6 (Accessed: 12 May 2024). **16. MoMLeT&DS Workshop**. An approach to the reverse dictionary task based on automatic smart subword segmentation. URL: https://www.youtube.com/watch?v=aewTMqTlb2c (Accessed: 12 May 2024) **17. Thomas E. B**. A Clear Explanation of Transformer Neural Networks. *Medium*. URL: https://medium.com/@ebinbabuthomas_21082/decoding-the-enigma-a-deep-dive-into-transformer-model-architecture-749b49883628 (Accessed: 12 May 2024). **18. Lepikhin D., Lee H., Xu Y., Chen D., Firat O., Huang Y., Krikun M., Shazeer N. M., Chen Z.** GShard: Scaling Giant Models with Conditional Computation and Automatic Sharding. *ArXiv*. 2020. abs/2006.16668. URL: https://api.semanticscholar.org/CorpusID:220265858 (дата звернення: 12.05.2024). **19. Slyusar V., Bihun N.** The Method of Increasing the Immunity of Data Transmission in Communication Channels. *2022 IEEE 9th International Conference on Problems of Infocommunications, Science and Technology (PIC S&T)*, Kharkiv, Ukraine, 10–12 October 2022. 2022. DOI: 10.1109/picst57299.2022.10238546.

CONTROL OF SWARMS OF UNMANNED AERIAL VEHICLES BASED ON NEURAL TRANSFORMER

Bihun Nataliia

[1] Central Scientific Research Institute of Armaments and Military Equipment of Armed Forces of Ukraine, Kyiv, Ukraine

Formulation of the problem in general. *The current challenges in managing swarms of unmanned aerial vehicles require the improvement of intelligent systems using advanced machine learning technologies. This study focuses on the integration of Mixture of Experts technology into neural transformers to improve the efficiency of swarms of unmanned aerial vehicles. The purpose of the article is to analyze and develop methods to improve the capabilities of swarms of unmanned aerial vehicles by improving neural transformers using Mixture of Experts technology.*

Research methods. *To achieve this goal, methodological approaches were used to discover, analyze, and experiment with different Mixture of Experts architectures and neural transformers, including the use of specialized submodels (experts) to solve specific tasks within a swarm of unmanned aerial vehicles.*

Analysis of recent researches and publications. *The latest research and publications show that Mixture of Experts architectures have the potential to scale, especially in terms of data volumes. The study builds on the theoretical foundations laid by Mixture of Experts researchers like Gormley and Frühwirth-Schnatter, who emphasized the universal applicability and utility of Mixture of Experts in various fields. Building on these concepts, the following authors further elaborate the process of task decomposition and identification of an expert within Mixture of Experts frameworks: Krishnamurthy, Watkins, and Gaertner. The integration of Mixture of Experts into neural transformers for unmanned aerial vehicles swarm management requires thorough analysis and development of optimized approaches to ensure high system performance while maintaining computational efficiency.*

Presenting the main material. *The study reveals the integration of Mixture of Experts models into the operational structure of unmanned aerial vehicles swarms, supported by the advanced capabilities of neural transformers. The methodology describes a comprehensive approach that includes the selection of Mixture of Experts models, neural transformer specifications, and the operational interaction of unmanned aerial vehicles swarms. The study demonstrates that the integration of Mixture of Experts with transformers significantly improves the efficiency of unmanned aerial vehicles swarms through more efficient task allocation and adaptation to changing environmental conditions.*

Elements of scientific novelty *lie in the development of optimized Mixture of Experts structures and their integration into unmanned aerial vehicles swarms, which can help in addressing limitations with regard to computation and resources, and provide flexibility and decision-making capabilities.*

Theoretical and practical significance *of the study for the military and defence sector and the field of technical sciences is the creation of autonomous systems capable of rapid adaptation and operational execution of tasks in various conditions.*

Conclusion and the perspectives of future research. *The study opens up new ways and prospects for developing new algorithms to more accurately determine the relevance of experts to specific tasks, as well as studying the impact of different expert architectures on the overall system performance.*

Keywords: *swarm of unmanned aerial vehicles, Mixture of Experts, neural transformers, autonomous systems, data processing.*

References

1. Al-Emadi, S., Al-Mohannadi, A., (2020). Towards Enhancement of Network Communication Architectures and Routing Protocols for FANETs: A Survey. In: *2020 3rd International Conference on Advanced Communication Technologies and Networking (CommNet), 4–6 September 2020, Marrakech, Morocco*. IEEE. DOI: 10.1109/commnet49926.2020.9199627. **2. Slyusar, V.** (2024) Cyber threats of large language models. *Systems and means of artificial intelligence*, 15-16 March 2024, Kyiv, Ukraine. IPAI. 252–260. **3. Sribniuk, S. M.** (2022) *Hydraulic and aerodynamic machines.* Kyiv: Center for Educational Literature **4. Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q. V., Hinton, G. E., Dean, J.**, (2017). Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer [online]. *CoRR.* Available at: http://arxiv.org/abs/1701.06538 [Accessed : 12 May 2024]. **5. Jawahar, G., Mukherjee, S., Liu, X., Kim, Y. J., Abdul-Mageed, M., Lakshmanan, V.S., L., Awadallah, A. H., Bubeck, S., Gao, J.,** (2023). AutoMoE: heterogeneous mixture-of-experts with adaptive computation for efficient neural machine translation. In: *Findings of the association for computational linguistics: ACL 2023, Toronto, Canada* Stroudsburg. PA, USA: Association for Computational Linguistics [online]. Available at: https://api.semanticscholar.org/CorpusID:259108418 [Accessed : 12 May 2024]. **6. Antoniak, S., Jaszczur, S., Krutul, M., Pi'oro, M., Krajewski, J., Ludziejewski, J., Odrzyg'ozdz, T., Cygan, M.,** (2023). Mixture of Tokens: Efficient LLMs through Cross-Example Aggregation. *CoRR*. **abs/2310.15961.** DOI: 10.48550/ARXIV.2310.15961 **7. Mohammadi, H., Nazerfard, E., Firoozi, T.,** (2023). Reinforcement Learning-based Mixture of Vision Transformers for Video Violence Recognition. *ArXiv*. DOI: 10.48550/arXiv.2310.03108. **8. Puigcerver, J., Riquelme, C., Mustafa, B. та Houlsby, N.,** (2023). From sparse to soft mixtures of experts. *ArXiv.* DOI:10.48550/arXiv.2308.00951 **9. Liu, Z., Wang, Y., Vaidya, S., Ruehle, F., Halverson, J., Soljacic, M., Hou, T. Y., Tegmark, M.,** (2024). KAN: Kolmogorov-Arnold Networks. *ArXiv*. DOI: 10.48550/arXiv.2404.19756. **10. Slyusar, V., Sliusar, I., Bihun, N. та Piliuhin, V.,** (2022). Segmentation of analogue meter readings using neural networks. In: *MoMLeT+DS* **11. Bouaouni, Y., (2023).** Mixture of experts [online]. *Medium*. Available at: https://medium.com/@yacinebouaouni07/mixture-of-experts-26243919d145 [Accessed: 12 May 2024]. **12. Zhou, Z.-H.,** (2012). *Ensemble methods: foundations and algorithms* No. 14. DOI: 10.1201/ b12207. **13. Gale, T., Narayanan, D., Young, C. та Zaharia, M. A.,** (2022). MegaBlocks: efficient sparse training with Mixture-of-Experts. *ArXiv*. DOI: 10.48550/arXiv.2211.15841 **14. Slyusar, V., Kondratenko, Y., Shevchenko, A. та Yeroshenko, T.,** (2024). Some Aspects of Artificial Intelligence Development Strategy for Mobile Technologies. *Journal of Mobile Multimedia*. 525–554. DOI: 10.13052/jmm1550-4646.2031. **15. Ajay, L.,** (2023). An intuitive introduction to Transformers [online]. *Medium*. Available at: https://lakshmi1212.medium.com/an-intuitive-introduction-to-transformers-6f574c8e7df6 [Accessed : 12 May 2024] **16. MoMLeT&DS Workshop,** (2024). An approach to the reverse dictionary task based on automatic smart subword segmentation [online]. *YouTube*. Available at: https://www.youtube.com/watch?v=aewTMqTlb2c [Accessed : 12 May 2024] **17. Thomas, E. B.,** (2023). A clear explanation of transformer neural networks [online]. *Medium*. Available at: https://medium.com/@ebinbabuthomas_21082/decoding-the-enigma-a-deep-dive-into-transformer-model-architecture-749b49883628 [Accessed : 12 March 2024] **18. Lepikhin, D., Lee, H., Xu, Y., Chen, D., Firat, O., Huang, Y., Krikun, M., Shazeer, N., Chen, Z.,** (2020). GShard: scaling giant models with conditional computation and automatic sharding [online]. *ArXiv* **abs/2006.16668.** Available at: https://api.semanticscholar.org/CorpusID:220265858 [Accessed : 12 March 2024] **19. Slyusar, V. та Bihun, N.** (2022). The Method of Increasing the Immunity of Data Transmission in Communication Channels. In: *2022 IEEE 9th International Conference on Problems of Infocommunications, Science and Technology (PIC S&T), 10–12 жовтня 2022, Kharkiv, Ukraine*. IEEE. DOI: 10.1109/picst57299.2022.10238546.