Міністерство освіти і науки України

Київський національний університет імені Тараса Шевченка

Кафедра алгебри і комп'ютерної математики

Курсовий проект

На тему: «MoA+: Суміш автокодувальників зі змінними концентраціями для покращен кластеризації зображень»

Виконав:

Тунік Вадим Романович,

студент 4 курсу, Комп'ютерна математика.

Науковий керівник:

Мірошниченко Віталій Олегович,

Асистент кафедри Теорії ймовірностей,

статистики та актуарної математики

Київ-2025

Зміст

1 Вступ.

Нещодавні досягнення в галузі навчання без вчителя продемонстрували ефективність поєднання архітектури суміші експертів (MoE) з автокодувальниками для задач кластеризації зображень. Цей підхід використовує спеціалізовані автокодувальники як експертів, кожен з яких навчений відтворювати окремі патерни даних, тоді як роутер динамічно спрямовує вхідні дані до найбільш підходящого експерта.

Запропонована архітектура дозволяє подолати обмеженість традиційних методів, які страждають від високорозмірності візуальних даних. Традиційні методи, такі як k-середні та моделі гаусових сумішей, часто не враховують нелінійність, що лежить в основі розподілу зображень, що призводить до зниження точності на таких датасетах, як CIFAR-10 та MNIST, порівняно з альтернативними методами глибокого навчання [1].

Емпіричні дослідження показують, що Mixture of Autoencoders (MoA) дозволяють досягти значного зросту метрик якості кластеризації порівняно з окремими автокодувальниками на еталонних наборах даних, таких як MNIST, 20NEWS, CIFAR-10 та ImageNet-1K [2] [3].

Подібна модель може бути застосована до вирішення низки проблем: біомедична візуалізація (сегментація гістологічних слайдів за типом тканини); автономні системи (розплутування траєкторій об'єктів у сценах з великим скупченням людей); модерація контенту (кластеризація користувацьких зображень за семантичними темами); дефектоскопія (виявлення дефектів по фотографіях на різних конструкціях) і т.д.

2 Теоретичні відомості.

2.1 Задача кластеризації.

Кластеризація, чи кластерний аналіз, є наріжним каменем в області некерованого машинного навчання. Її основна мета полягає в тому, щоб розбити заданий набір даних на окремі групи, які називаються кластерами, на основі принципу схожості. Основна ідея полягає в тому, щоб організувати дані так, щоб ті, що знаходяться в одному кластері, демонстрували високий ступінь схожості, в той час як ті, що не належать до нього, відрізнялися один від одного. Процес кластеризації є дуже складним і трудомістким. Цей процес за своєю суттю є дослідницьким, спрямованим на виявлення прихованих структур, патернів і природних угруповань в даних без попередніх вказівок на те, що ці групи представляють.

Некерована природа кластеризації є її визначальною характеристикою. На відміну від парадигм керованого навчання, таких як класифікація, алгоритми кластеризації працюють з наборами даних, де екземпляри не розмічені. Не існує заздалегідь визначеної цільової змінної або категорії базової істини, яка б керувала процесом навчання; алгоритм повинен автономно розпізнавати основну групову структуру, ґрунтуючись виключно на внутрішніх властивостях самих точок даних. Цей режим роботи іноді описують як «навчання без вчителя».

Центральним у процесі кластеризації є поняття подібності або, навпаки, відмінності (часто вимірюваної як відстань). Для групування точок даних необхідно встановити кількісну міру, щоб оцінити, наскільки схожими або різними є пари екземплярів, виходячи з їхніх характеристик. Серед поширених варіантів - евклідова відстань, Манхеттенська відстань та кореляційні показники. Вибір відповідної метрики подібності або відстані є критично важливим, оскільки вона безпосередньо впливає на структуру кластера, що утворюється в результаті.

Важливо зазначити, що поняття «подібність» не є абсолютним; воно часто залежить від контексту і прив'язане до конкретних цілей аналізу. Отже, те, що є «хорошою» кластеризацією, часто залежить від обраного визначення подібності та передбачуваного застосування, що робить процес певною мірою суб'єктивним. Кінцева мета, як правило, формулюється як одночасна максимізація внутрішньокластерної згуртованості (подібності між точками в межах одного кластера) і міжкластерного розмежування (розбіжності між точками в різних кластерах).

Важливо відрізняти кластеризацію від класифікації. Хоча обидва процеси передбачають групування даних, класифікація є контрольованим завданням, яке використовує марковані дані для вивчення відображення від ознак до заздалегідь визначених категорій. Кластеризація, будучи неконтрольованою, виявляє групи в немаркованих даних. Застосування кластеризації різноманітне і охоплює численні сфери, включаючи сегментацію клієнтів у маркетингу, ідентифікацію спільнот у соціальних мережах, групування результатів пошуку, аналіз медичних зображень, сегментацію зображень на основі схожості пікселів, виявлення аномалій або викидів, організацію документів і стиснення даних шляхом представлення точок за їхньою кластерною приналежністю.

Проблема визначення значущої подібності стає особливо гострою у високорозмірних просторах, таких як ті, що зустрічаються в зображеннях. Хоча візуалізація та оцінка подібності є відносно простою у малих розмірностях, «прокляття розмірності» може зробити традиційні метрики відстані менш ефективними зі збільшенням кількості ознак. У таких сценаріях відстані між точками можуть стати менш дискримінативними, що потенційно перешкоджає виявленню значущих кластерів. Це мотивує використання методів зменшення розмірності або підходів до навчання репрезентативності, таких як автокодувальники, які мають на меті спроектувати дані в простір меншої розмірності, де схожість більш ефективно фіксується, що є центральною темою в ширшому контексті цієї дослідницької роботи.

2.1.1 Розгляд підходів до кластеризації.

Кластеризація охоплює різні алгоритмічні підходи, кожен з яких має різні припущення та сильні сторони. Ключові категорії наступні:

- Методи K-means на базі центроїдів: Ці алгоритми поділяють дані на заздалегідь визначену кількість, k кластерів, що не перетинаються, часто представлених центральною точкою (центроїд). Класичним прикладом є метод k-середніх, який ітеративно приписує точки до найближчого центроїда і перераховує центроїди для мінімізації відстаней між кластерами. Вони, як правило, ефективні, але вимагають, щоб k було відоме, і можуть бути чутливими до початкової ініціалізації та викидів.

- Ієрархічні методи (агломератові/розділювальні): Ці методи будують вкладену ієрархію кластерів, часто представлену у вигляді дендрограми, не вимагаючи попередньо визначеного k. Висхідні підходи починають з окремих точок і об'єднують кластери, тоді як низхідні підходи починають з одного кластера і розбивають його. Ієрархія забезпечує гнучкість, але для великих наборів даних може вимагати значних обчислювальних витрат.

- Методи на основі щільності: Ці алгоритми визначають кластери як щільні області точок даних, розділені більш розрідженими областями. Вони можуть знаходити кластери довільної форми та ефективно обробляти шум. DBSCAN є відомим прикладом, який групує точки на основі параметрів локальної щільності. Вони не потребують k, але можуть бути чутливими до вибору параметрів і різної щільності.

- Методи, засновані на розподілі: Цей підхід передбачає, що дані генеруються із суміші базових розподілів ймовірностей (наприклад, гаусівських розподілів у моделях гаусівської суміші - GMM),

де кожен розподіл представляє кластер. Алгоритми, такі як EM-алгоритм, оцінюють параметри розподілу. Вони пропонують статистичну основу та гнучкість у формуванні кластерів, але покладаються на припущення щодо розподілу та можуть бути складними в обчислювальному плані.

Різноманітність цих методів підкреслює, що жоден з них не є універсально найкращим. Вибір залежить від характеристик та типу даних, очікуваної форми кластерів та обчислювальних ресурсів. Спільною проблемою для багатьох методів є визначення відповідної кількості кластерів, k.

2.1.2 Оцінювання якості кластеризації.

Оцінка якості результатів кластеризації має вирішальне значення через неконтрольовану природу завдання і різноманітність алгоритмів. Метрики оцінки кількісно визначають «якість» кластеризації, дозволяючи порівнювати алгоритми і налаштування параметрів. Метрики широко класифікуються як внутрішні або зовнішні.

Внутрішні метрики оцінювання: Оцінюють якість кластеризації, використовуючи лише дані та призначення кластерів, без зовнішніх міток. Зазвичай вони вимірюють компактність (схожість всередині кластерів) та відокремленість (відмінність між кластерами).

- Коефіцієнт силуету: Вимірює, наскільки точка схожа на власний кластер порівняно з іншими. Оцінки варіюються від -1 до 1, де 1 - найкращий показник. Оцінка для точки i дорівнює:

$$s(i) = \frac{b(i) - a(i)}{max(a(i), b(i))},$$

де $a(i)$ - середня внутрішньокластерна відстань, а $b(i)$ - найменша середня міжкластерна відстань.

- Індекс Девіса-Болдіна (DBI): Вимірює середній коефіцієнт схожості між кожним кластером та найбільш схожим кластером. Чим нижчі значення, тим краще.

- Індекс Калінського-Харабаша: Відношення міжкластерної дисперсії до внутрішньокластерної. Чим вищі значення, тим краще.

Зовнішні метрики оцінювання: Ці метрики оцінюють продуктивність алгоритму кластеризації, порівнюючи отримане розбиття з класифікацією, наданою ззовні. Це можливо в сценаріях, де відомі істинні мітки класів, часто в дослідницьких умовах для бенчмаркінгу алгоритмів на маркованих наборах даних. Зовнішні метрики оцінюють ступінь узгодженості між згенерованими алгоритмом кластерами та істинними класами.

- Скоригований індекс Ренда (ARI): вимірює схожість між передбаченими кластерами та істинними класами, розглядаючи пари вибірок та коригуючи випадковість узгодження. Оцінки варіюються від -1 до 1, де 1 - ідеальна збіжність.

- Оцінки на основі взаємної інформації (NMI, AMI): Ці метрики кількісно оцінюють узгодженість між двома розділами, використовуючи поняття з теорії інформації, зокрема, взаємну інформацію, яка вимірює спільну інформацію між двома розмітками. Нормалізована взаємна інформація (NMI) масштабує оцінку MI до діапазону. Скоригована взаємна інформація (AMI) додатково коригує показник MI для випадкового узгодження, подібно до ARI, також, як правило, в діапазоні від 0 (випадкове узгодження) до 1 (ідеальне узгодження). NMI буде детально обговорено пізніше.

- Однорідність, повнота та V-міра: Це також метрики, засновані на теорії інформації. Однорідність вимірює, чи кожен кластер містить лише представників одного класу. Повнота вимірює, чи всі

члени певного класу належать до одного кластера. V-міра є середнім гармонійним значенням однорідності та повноти, забезпечуючи єдиний показник, який збалансовує обидва аспекти.

Вибір залежить від наявності розмічених даних. Внутрішні метрики використовуються, коли мітки невідомі, тоді як зовнішні метрики оцінюють відновлення відомих структур. Жодна метрика не є ідеальною; використання декількох метрик дає більш цілісне уявлення.

2.1.3 Метрики оцінювання на основі взаємної інформації.

Теорія інформації пропонує математично обґрунтовану основу для кількісної оцінки подібності між двома різними кластерами набору даних. Метрики, отримані на основі взаємної інформації (Mutual Information, MI), є одними з найпоширеніших принципів зовнішнього оцінювання, що оцінюють відповідність між кластерами, отриманими за допомогою алгоритму, і мітками істинних класів.

Поняття ентропії є фундаментальним для теорії інформації і, відповідно, для метрик, заснованих на MI. Ентропія, позначена через H, вимірює кількість невизначеності, випадковості або «інформативності», пов'язаної з випадковою величиною або її розподілом ймовірностей. Для дискретної випадкової величини X, яка може набувати значень $x_1, x_2, ..., x_n$ з ймовірностями $P(x_i)$, ентропія визначається наступним чином:

$$H(X) = -\sum_{i=1}^{n} P(x_i) log(P(x_i)).$$

Вище значення ентропії означає більшу невизначеність щодо результату випадкової величини.

В контексті оцінки кластеризації ми розглядаємо два розділення даних: базова істина $U = \{U_1, ..., U_R\}$ та передбачення $V = \{V_1, ..., V_C\}$. Кожне розбиття можна розглядати як визначення розподілу ймовірностей між групами. Ентропія $H(U)$ кількісно виражає невизначеність,

пов'язану з істинною належністю до класу випадково вибраної точки даних, тоді як $H(V)$ кількісно виражає невизначеність у її прогнозованому кластерному розподілі.

Взаємна інформація, що позначається як $I(U;V)$, вимірює взаємну залежність між двома випадковими величинами U і V. Вона кількісно визначає кількість інформації, яку одна змінна надає про іншу, або, еквівалентно, зменшення невизначеності щодо однієї змінної, досягнуте завдяки знанню значення іншої. У контексті порівняння кластеризацій U і V, MI вимірює інформацію, що міститься в істинних розподілах класів і прогнозованих розподілах кластерів.

MI можна виразити через ентропію кількома еквівалентними способами :

$$I(U;V) = H(U) - H(U|V)$$

$$I(U;V) = H(V) - H(V|U)$$

$$I(U;V) = H(U) + H(V) - H(U,V)$$

Тут $H(U|V)$ і $H(V|U)$ представляють умовні ентропії (невизначеність, що залишилася в одному розбитті, враховуючи інше), а $H(U,V)$ є спільною ентропією (повна невизначеність об'єднаної системи).

MI має кілька важливих властивостей: вона симетрична, $I(U;V) = I(V;U)$, і невід'ємна, $I(U;V) \geq 0$. Важливо, що $I(U;V) = 0$ тоді і тільки тоді, коли два розбиття U і V є статистично незалежними. Вищі значення MI вказують на сильнішу залежність, а отже, на більшу узгодженість між двома кластеризаціями.

Глибше розуміння MI випливає з його зв'язку з дивергенцією Кульбака-Лейблера. MI можна інтерпретувати як дивергенцію між спільним розподілом ймовірностей $P_{(U,V)}$ і добутком відособлених розподілів $P_U P_V$:

$$I(U;V) = D_{KL}(P_{(U,V)} || P_U P_V).$$

Розподіл $P_U P_V$ представляє спільний розподіл, який мав би місце, якби U та V були незалежними. Таким чином, MI кількісно визначає, наскільки фактичний спільний зв'язок між кластерами відрізняється від незалежного, вимірюючи «інформаційний здобуток», отриманий при врахуванні їхньої залежності.

Однак, необроблені показники MI створюють проблеми в інтерпретації. Максимально можливе значення MI обмежене індивідуальними ентропіями, $min(H(U), H(V))$, які залежать від кількості кластерів та їх розподілу за розмірами. Отже, значення MI не можуть бути масштабовані до фіксованого діапазону, і порівняння оцінок MI між різними наборами даних або кластеризаціями з різною кількістю груп може ввести в оману. Вищий показник MI може виникнути просто через наявність більшої кількості кластерів в одному або обох розділах, а не відображати справді кращу узгодженість. Така залежність від інших чинників, окрім якості згоди, вимагає процедур нормалізації або коригування для отримання більш зрозумілих і порівнянних показників, що призводить до появи таких показників, як нормалізована взаємна інформація (NMI) та скоригована взаємна інформація (AMI).

2.1.4 Нормалізована взаємна інформація (NMI).

Нормалізована взаємна інформація (NMI) - це адаптація показника взаємної інформації (MI), спеціально розроблена для вирішення проблеми залежності від масштабу. Його основна мета - нормалізувати значення MI до послідовного діапазону, як правило, між 0 і 1, тим самим полегшуючи більш просту інтерпретацію і порівняння узгодженості кластеризації в різних сценаріях, наприклад, з різною кількістю кластерів або наборів даних.

Нормалізація досягається шляхом ділення необробленої оцінки взаємної інформації, $I(U;V)$, на коефіцієнт нормалізації, отриманий з ентропій окремих розділів, $H(U)$ і $H(V)$. Існує декілька схем нормалізації,

але загальним і широко прийнятим підходом, особливо в таких реалізаціях, як scikit-learn (починаючи з версії 0.22), є використання середнього арифметичного значення ентропій як знаменника :

$$NMI(U,V) = \frac{I(U;V)}{\frac{H(U)+H(V)}{2}}$$

Інтерпретація:

- NMI ≈ 1: Висока узгодженість або ідеальна кореляція між істинними мітками (U) та передбачуваними (V);
- NMI ≈ 0: Кластери значною мірою незалежні; мало спільної інформації, окрім випадкової.

NMI успадковує деякі бажані властивості MI: він симетричний, тобто $NMI(U,V) = NMI(V,U)$, та інваріантний до перестановок міток кластерів. Це означає, що назви кластерів не впливають на оцінку. Однак ключовим обмеженням є те, що стандартний NMI не враховує випадковість. Випадкові кластеризації можуть давати ненульові оцінки NMI. Скоригована взаємна інформація (AMI) коригує це базове узгодження, і їй часто надають перевагу для точних порівнянь. Незважаючи на це, NMI залишається широко використовуваним. Його розрахунок може сприйматися як простіший, а його потенційна похибка, зумовлена випадковістю, може бути менш значущою, якщо порівнювати кластери з подібною кількістю кластерів або якщо основною метою є відносне порівняння, а не абсолютна оцінка порівняно з випадковим базовим показником.

2.2 Autoencoders.

Автокодувальники (Autoencoders, AE) - це окремий клас нейронних мереж в парадигмі навчання без вчителя. Основна мета автокодувальників полягає у вивченні ефективних, часто стислих, представлень (кодувань) вхідних даних без використання явних міток. Ця здатність

робить їх потужними інструментами для таких завдань, як зменшення розмірності, вивчення особливостей, виявлення аномалій та генеративне моделювання. Виникнувши з концепцій, пов'язаних зі стисненням даних і нелінійними узагальненнями аналізу головних компонент (PCA) [6], AE перетворилися на фундаментальні будівельні блоки в сучасному глибокому навчанні. Фундаментальні дослідження таких піонерів, як LeCun (1987), Bourlard і Kamp (1988), Hinton і Zemel (1994) та Kramer (1991), встановили основні принципи, продемонструвавши потенціал нейронних мереж для вивчення значущих кодів даних шляхом самореконструкції. Всебічний розгляд AE в ширшому контексті глибокого навчання надано Goodfellow та ін. (2016). [7]

Квінтесенція архітектури AE складається з двох основних компонентів нейронної мережі: енкодера і декодера. Енкодер, позначений як функція f, приймає вхідні дані високої розмірності $x \in \mathbb{R}^d$ і відображає їх у внутрішнє представлення нижчої розмірності $h \in \mathbb{R}^p$, де зазвичай латентний вимір p менший за вхідний вимір d ($p << d$). Це стиснене представлення h називається кодом, латентним представленням або пляшкове горло, яке знаходиться в латентному просторі. Декодер, представлений функцією g, згодом приймає цей латентний код h як вхідні дані і намагається відновити початкові вхідні дані, виробляючи вихід $r = \hat{x} \in \mathbb{R}^d$. Таким чином, всю роботу автокодувальника можна описати як композицію $r = g(f(x))$.

В той час як механізм роботи AE передбачає навчання мережі для мінімізації різниці між входом x і реконструйованим виходом r, тобто $r \approx x$, кінцева мета часто виходить за рамки простої реконструкції. Якби AE міг ідеально відтворювати вхідні дані (тобто, вивчити функцію тотожності $g(f(x)) = x$ скрізь) без жодних обмежень, то вивчений прихований код h може бути не дуже інформативним або корисним для наступних завдань. Отже, AE зазвичай розробляються з внутрішніми обмеженнями, які запобігають тривіальному копіюванню. Найпоширенішим обмеженням є неповний ботлнек (шар латентного, стисненого

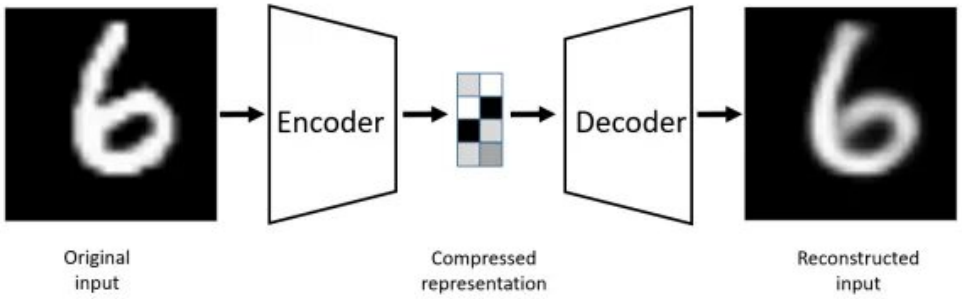


Рис. 1: Схема автокодувальника.

представлення), коли латентний вимір p значно менший за вхідний вимір d. Це структурне обмеження змушує кодер вивчати стиснене представлення, яке фіксує лише найсуттєвіші та найважливіші варіації або фактори, присутні у розподілі вхідних даних, відкидаючи шум і надмірність. Інші обмеження, такі як штрафи регуляризації, що застосовуються до ваг або активацій мережі, слугують аналогічній цілі. Завдання реконструкції, таким чином, виступає як проксі-мета або форма самоконтролю: навчаючись точно реконструювати дані, незважаючи на накладені обмеження, кодер неявно вивчає значуще, стиснене представлення h, яке витягує основний інформаційний зміст вхідних даних. Це вивчене представлення h часто є основним результатом, що представляє інтерес, слугуючи вхідними характеристиками для наступних завдань машинного навчання, таких як класифікація, генерація або, як доречно в даному випадку, кластеризація.

2.2.1 Математичне формулювання та постановка цільового функціоналу навчання.

Формально, енкодер і декодер є нейронними мережами, параметри яких задаються наборами ваг і зсувів, позначених θ_e і θ_d відповідно. Процес кодування має вигляд $h = f(x; \theta_e)$, а процес декодування - $r = g(h; \theta_d)$. Навчання автокодувальника передбачає пошук оптимальних параметрів θ_e і θ_d, які мінімізують обрану функцію втрат реконструкції:

$$L(x, r) = L(x, g(f(x; \theta_e); \theta_d).$$

Ця функція втрат кількісно визначає розбіжність або помилку між початковим входом x і реконструйованим виходом r. Мінімізація зазвичай досягається за допомогою стохастичного градієнтного спуску (SGD) або його модифікації (наприклад, Adam, RMSprop), де градієнти обчислюються за допомогою алгоритму зворотного розповсюдження похибки (backpropagation).

Вибір функції втрат L є важливим і значною мірою залежить від характеристик вхідних даних x. Дві найпоширеніші функції втрат - це середньоквадратична помилка (MSE) та бінарна перехресна ентропія (BCE).

Середньоквадратична похибка (MSE): Також відома як L2-loss, MSE є вибором за замовчуванням для автокодувальників, коли вхідні дані складаються з неперервних елементів з дійсними значеннями. Вона обчислює середню квадратичну різницю між вхідними даними та реконструкцією за всіма вимірами. Для окремої вибірки даних x втрати MSE задаються формулою:

$$L_{MSE}(x, r) = ||x - r||^2 = \sum_{j=1}^{d} (x_j - r_j)^2$$

При усередненні за міні-батчами (міні-партії) або за всім набором даних N вибірок вона стає $\frac{1}{N} \sum_{i=1}^{N} ||x_i - r_i||^2$. MSE сильніше карає за

більші помилки через операцію зведення в квадрат, але може бути чутливою до викидів у даних.

2.2.2 Згорткові Автокодувальники (CAEs) для зображень.

Стандартні автокодувальники, що використовують повноз'єднані шари, є універсальними, але не можуть явно врахувати просторову структуру, притаманну таким типам даних, як зображення. Зображення демонструють сильну просторову локальність (сусідні пікселі сильно корелюють) і статистичні властивості, які часто є інваріантними до переміщення. Повністю з'єднані шари обробляють кожен вхідний піксель незалежно в початкових шарах, фактично ігноруючи цю важливу 2D-структуру і призводячи до вибуху кількості параметрів при роботі з зображеннями високої роздільної здатності [14].

Для подолання цих обмежень було розроблено згорткові автокодувальники (CAE) як спеціалізовану архітектуру, пристосовану для обробки зображень. CAE замінюють повністю з'єднані шари стандартного AE на згорткові та пов'язані з ними шари, таким чином, вносячи індуктивні зміщення, придатні для обробки зображень.

Архітектура типового CAE відповідає структурі кодер-декодер, але використовує шари, призначені для просторових даних:

- Енкодер: Енкодер CAE складається зі стеку згорткових шарів. Ці шари застосовують фільтри (ядра), що навчаються, до вхідного зображення, виявляючи локальні шаблони та особливості (краї, текстури тощо). Ключовою властивістю згорткових шарів є спільне використання параметрів: один і той самий фільтр застосовується в різних просторових точках, що значно зменшує кількість параметрів порівняно з повноз'єднаними шарами і сприяє еквівалентності переміщенню. Коли дані проходять через кодер, просторові розміри (висота і ширина) зазвичай зменшуються, тоді як кількість каналів ознак часто збільшується. Зазвичай це дося-

гається за рахунок зменшення дискретизації (downsampling):

  - Шар об'єднання (Pooling Layer): Такі операції, як максимальне об'єднання або середнє об'єднання, явно зменшують просторову роздільну здатність, підсумовуючи регіони особливостей.
  - Згортки з кроком (Strided Convolution): Встановлення кроку згортки шару, що згортується, на значення більше 1 забезпечує зменшення вибірки безпосередньо під час операції згортки.

- Декодер: Декодер CAE має на меті відновити розміри та деталі оригінального зображення на основі стисненого латентного представлення. Зазвичай він використовує транспоновані згорткові шари (які часто називають деконволюційними шарами). Ці шари виконують операцію підвищення дискретизації (upsampling), збільшуючи просторову роздільну здатність, одночасно навчаючись генерувати дрібніші деталі зображення. Вони можуть бути розглянуті як вивчення зворотного просторового перетворення відповідних згорткових шарів у кодері.

Важливим аспектом проектування в CAEs є метод просторового зменшення дискретизації (downsampling) та підвищення дискретизації (upsampling). Хоча об'єднання шарів (наприклад, максимальне об'єднання) є поширеним у стандартних згорткових нейронних мережах (CNN) для класифікації, альтернативний підхід у CAE полягає в тому, щоб повністю покладатися на крокові згортки для дискретизації в кодері та відповідні крокові транспоновані згортки для дискретизації в декодері. Аргументом на користь використання кроків є те, що це дозволяє мережі навчитися оптимальному способу зменшення або збільшення просторової роздільної здатності в процесі навчання фільтра, замість того, щоб покладатися на фіксовану, розроблену вручну операцію об'єднання. Та-

ке навчене перетворення може запропонувати більшу гнучкість і потенційно зберегти більше релевантної інформації для точної реконструкції або для вилучення дискримінантних ознак, необхідних для наступних завдань, таких як кластеризація.

Розвиток і успіх CAEs [14] підкреслюють ширший принцип глибокого навчання: адаптація мережевих архітектур до конкретних характеристик і структури вхідних даних має вирішальне значення для досягнення високої продуктивності. У той час як стандартні AEs забезпечують загальну основу для неконтрольованого навчання репрезентації, CAEs враховують індуктивні упередження просторової локалізації, спільне використання параметрів та ієрархічне вилучення ознак, що робить їх значно ефективнішими та результативнішими для даних зображень. Ця архітектурна спеціалізація є фундаментальною для їхньої корисності в задачах, пов'язаних із зображеннями.

2.3 Моделі сумішей зі змінними концентраціями.

В статистичному аналізі моделі сумішей природнім чином виникають в більшості галузей дослідження. Скінченні сумішеві моделі (Finite Mixture Models, FMM) є потужним та широко використовуваним інструментом для моделювання гетерогенних даних. Фундаментальне припущення FMM полягає в тому, що спостережувані дані генеруються з сукупності, яка складається з скінченного числа M неспостережуваних підгруп або компонент. Кожен компонент характеризується власним розподілом регресорів, помилок, тощо. Щільність скінченної суміші для спостереження у записується як:

$$f(y;\Psi) = \sum_{i=1}^{M} \pi_i f_i(y;\theta_i),$$

де π_i - це апріорні ймовірності (або ваги, концентрації) належності спостереження до i-ї компоненти ($\pi_i \geq 0, \sum_{i=1}^{M} \pi_i = 1$), $f_i(y;,\theta_i)$ - це функція щільності i-ї компоненти, параметризована вектором параметрів θ_i, а

$\Psi = (\pi_1, ..., \pi_M, \theta_1, ..., \theta_M)$ - це повний вектор параметрів моделі. FMM успішно застосовуються в різноманітних галузях, включаючи кластерний аналіз, латентно-класовий аналіз, розпізнавання образів та медичну візуалізацію. [8]

Однак ключовим обмеженням класичних FMM є припущення про сталість концентрацій π_i для всіх спостережень у вибірці. Це припущення може бути надто обмежувальним для багатьох реальних наборів даних, де ймовірність належності до певної компоненти може змінюватися від одного спостереження до іншого під впливом зовнішніх факторів або характеристик самого спостереження.

Для подолання цього обмеження були запропоновані моделі сумішей зі змінними концентраціями (Mixture with Varying Concentrations, MVC). MVC є прямим узагальненням FMM, яке дозволяє концентраціям компонент варіювати для кожного окремого спостереження.

Формально, нехай $\xi_1, ..., \xi_N$ – це вибірка незалежних спостережень. Вважається, що кожне спостереження $\xi_j (j = 1, ..., N)$ походить з однієї з M компонент, що характеризуються невідомими функціями розподілу $F_1, ..., F_M$. На відміну від FMM, ймовірність того, що спостереження ξ_j належить до m-ї компоненти $(m = 1, ..., M)$, позначається як $p^m_{j;N}$ і може залежати від індексу спостереження j. Ці ймовірності задовольняють умову $\sum_{m=1}^{M} p^m_{j;N} = 1$ для кожного j. Тоді функція розподілу $P_j(x)$ спостереження ξ_j визначається як:

$$P_j(x) = P\{\xi_j < x\} = \sum_{m=1}^{M} p^m_{j;N} F_m(x)$$

Ця залежність концентрацій $p^m_{j;N}$ від конкретного спостереження надає моделі значно більшу гнучкість. Це особливо важливо при моделюванні складних гетерогенних даних, де контекстуальна інформація, пов'язана з кожним спостереженням (наприклад, час спостереження, демографічні характеристики пацієнта, географічне положення), може впливати на ймовірність його належності до тієї чи іншої прихованої

групи. Можливість моделювати концентрації як функції відомих коваріат або індексів спостережень робить MVC більш реалістичною та потужною для аналізу таких даних, що підтверджується її застосуваннями в генетиці, соціології, нейронауках та інших галузях. Фундаментальні роботи з теорії та застосувань MVC належать Р. Майбороді, В. Мірошниченко, О. Сугакова, зокрема праці [9].

2.3.1 Непараметричне оцінювання в моделях MVC.

Значна частина досліджень MVC, зосереджена на непараметричній постановці задачі. У цій постановці основне припущення полягає в тому, що функції розподілу компонент $F_1, ..., F_M$ є повністю невідомими і не належать до якогось певного параметричного сімейства (наприклад, нормального). Основною статистичною задачею в такому випадку стає оцінювання цих невідомих функцій розподілу F_m, а також їхніх характеристик (таких як математичні сподівання, дисперсії, щільності або інші функціональні моменти) на основі спостережень $\xi_1, ..., \xi_N$, що походять із суміші.

Для оцінювання невідомих функцій розподілу компонент F_m в MVC моделях, коли концентрації $p^m_{j;N}$ вважаються відомими, використовується підхід, що базується на зважених емпіричних функціях розподілу (Weighted Empirical Distribution Functions, WEDF). Ідея полягає в тому, щоб "розділити"внесок кожного спостереження ξ_j між компонентами відповідно до відомих концентрацій. Оцінка для $F_m(x)$ має вигляд:

$$\hat{F}^m_N(x) = \sum_{j=1}^{N} a^m_{j;N} I(\xi_j < x)$$

де I – індикаторна функція, а $a^m_{j;N}$ – це спеціально підібрані ваги.

Р.Є. Майборода показав [15], що оптимальні (в сенсі мінімаксності та незміщеності) лінійні оцінки для F_m можна отримати, використовуючи так звані мінімаксні ваги. Для визначення цих ваг вводиться матриця концентрацій P_n розміром $M \times N$, де елемент $P_n[m, j] = p^m_{j;N}$, та

матриця Грама $\Gamma_N = P_n P_n^T$ (розміром $M \times M$). Якщо матриця Γ_N є невиродженою, то вектор ваг для j-го спостереження $a_{j;N} = (a_{j;N}^1, ..., a_{j;N}^M)^T$ пов'язаний з вектором концентрацій $p_{j;N} = (p_{j;N}^1, ..., p_{j;N}^M)^T$ через обернену матрицю Грама. Матриця A_n (розміром $M \times N$), що містить ці ваги (де $A_n[m, j] = a_{j;N}^m$), може бути записана у вигляді:

$$A_n = (P_n^T P_n)^{-1} P_n.$$

Ці ваги забезпечують оптимальні лінійні незміщені оцінки за певних умов [10].

Для цих оцінок доведені властивості конзистентності та асимптотичної нормальності за певних умов регулярності, таких як невиродженість матриці Грама Γ_N. [10] Асимптотична нормальність дозволяє будувати довірчі інтервали та перевіряти статистичні гіпотези щодо параметрів компонент.

2.3.2 Релевантність для сумішей автокодувальників у задачі кластеризації зображень

Теоретичні основи моделей сумішей зі змінними концентраціями (MVC) мають пряме відношення до задачі розробки архітектури суміші автокодувальників для кластеризації зображень. Набори даних зображень часто характеризуються значною гетерогенністю. Належність зображення до певного кластера може залежати не лише від його семантичного змісту чи візуальних ознак, але й від контекстуальних факторів, таких як умови освітлення, ракурс зйомки, наявність часткових затулень або стиль зображення.

Гнучкість MVC, що полягає у можливості мати індивідуальні для кожного зображення j концентрації $p_{j;N}^m$ (ймовірності належності до кластера m), надає природний спосіб моделювання такої гетерогенності. Концентрація $p_{j;N}^m$ для зображення j потенційно може бути пов'язана з ознаками, витягнутими автокодувальником з цього зображення (наприклад, з його латентним представленням), або з іншими доступними

метаданими зображення. Це дозволяє враховувати індивідуальні особливості кожного зразка при визначенні його ймовірнісної належності до різних кластерів.

Більше того, непараметричний характер методів оцінювання, розроблених для MVC (зокрема, припущення про невідомість розподілів компонент F_m), добре узгоджується з підходами глибокого навчання, такими як автокодувальники. Автокодувальники навчаються складним, керованим даними представленням без жорстких апріорних припущень щодо форми розподілу даних у просторі ознак або в латентному просторі. Таким чином, непараметрична природа MVC дозволяє поєднувати її з потужністю автокодувальників у вивченні представлень.

Включення принципів MVC до архітектури суміші автокодувальників може суттєво покращити її здатність моделювати складні кластерні структури в гетерогенних наборах зображень. Наприклад, методи, натхненні роботами Р.Є. Майбороди [9], такі як використання оптимальних ваг ($A_n = (P_n^T P_n)^{-1} P_n.$) для комбінування інформації від різних зображень при оновленні параметрів кластерів, або ідеї, подібні до методу найменших квадратів для зв'язку концентрацій з латентними ознаками, можуть стати основою для розробки нових, більш ефективних алгоритмів навчання сумішей автокодувальників. Це створює теоретичне підґрунтя для методологічної частини даної наукової роботи.

2.4 Mixture of Experts (MoE)

Mixture of Experts (MoE) - це метод ансамблевого навчання, який реалізує ідею розбиття моделі на спеціалізовані експертні мережі (expert network), при цьому паралельна розподільна мережа (gating network) визначає відповідного експерта для заданих вхідних даних.

Цей модульний підхід особливо корисний для задач з нелінійними та нерівномірними зв'язками між входами та виходами, таких як розпізнавання мови, класифікація зображень та інших задач високої роз-

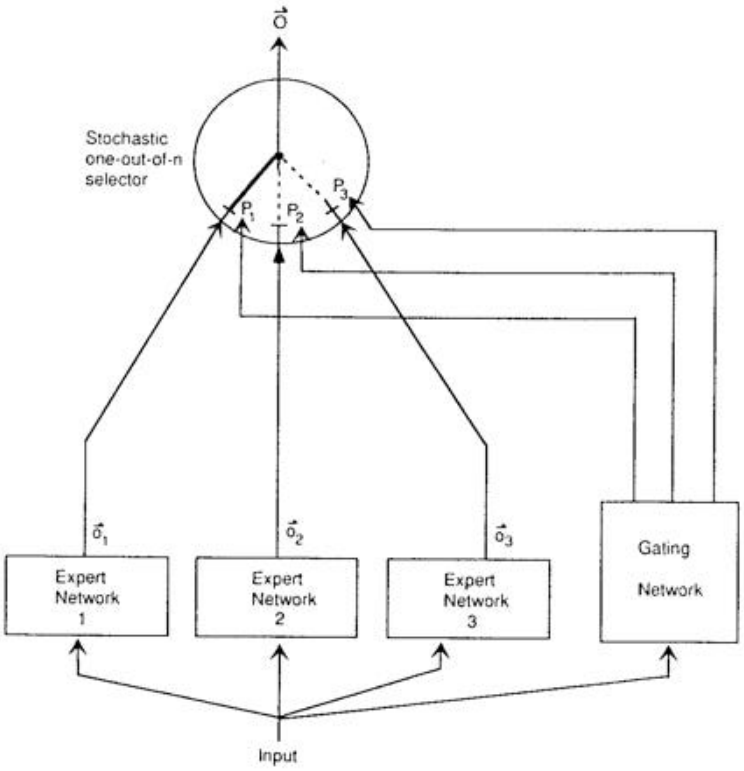


Рис. 2: Система експертів та розподільника.

мірності.

Фундаментальна праця «Adaptive Mixtures of Local Experts», опублікована в 1991 році, є наріжним каменем у цій галузі. Автори статті - Роберт А. Джейкобс, Майкл І. Джордан, Стівен Новлан і Джеффрі Е. Хінтон - представили процедуру керованого навчання для систем, що складаються з декількох окремих мереж, кожна з яких навчається обробляти підмножину навчальних прикладів [4]. У статті MoE позиціонується як модульна версія багатошарових перцептронів і як асоціативна версія змагального навчання, таким чином поєднуючи два, здавалося б, непорівнянних підходи.

2.4.1 Архітектура моделі MoE

Модель MoE складається з двох основних компонентів:

- Expert Network: Експертні мережі - це окремі нейронні мережі, кожна з яких призначена для спеціалізації на певній області вхідних

даних. Вони можуть бути дуже простими, наприклад, одношаровими перцептронами або неглибокими нейронними мережами, що є ключовою особливістю, яка робить можливим масштабованість та інтерпретованість. Кожен експерт отримує вхідні дані X і продукує вихідні передбачення y_k, де k позначає індекс експерта.

- Gating Network: Розподільна мережа - це додаткова паралельна нейронна мережа, яка обробляє вхідні дані X і видає вектор ймовірностей $g_k(X)$, де $\sum_{k=1}^{K} g_k(X) = 1$, і K - це кількість експертів. Ці ймовірності вказують на вірогідність того, що k-й експерт повинен бути використаний для даного входу X. Нейронна мережа ефективно виконує м'яке призначення, дозволяючи комбінувати прогнози декількох експертів.

Загальний результат моделі розраховується як зважена сума експертних оцінок:

$$y = \sum_{k=1}^{K} g_k(X) y_k$$

Таке формулювання гарантує, що модель може використовувати сильні сторони декількох експертів, при цьому розподільна мережа динамічно підлаштовує внесок кожного експерта на основі вхідних даних. Використання м'якого розподілення, на відміну від жорсткого призначення, дозволяє більш плавні переходи між експертними областями та краще обробляти області, що перетинаються у вхідному просторі.

Акцент на локальності експертів свідчить про те, що експерти мають спеціалізуватися на локалізованих областях вхідного простору, чому сприяє здатність мережі зіставляти вхідні дані з відповідними експертами. Такою мережею може бути багатошаровий перцептрон, який виводить softmax ймовірності для експертів, виходячи з сучасної практики.

2.4.2 Алгоритм навчання та процедура навчання

Навчання моделі суміші експертів відбувається за парадигмою навчання з вчителем, метою якого є мінімізація розбіжності між виходом моделі y та істинним цільовим виходом для кожного входу. Функція втрат, як правило, середня квадратична помилка (MSE) для регресії або перехресна ентропія для класифікації.

Процедура навчання використовує метод зворотного поширення помилки (backpropagation) для оновлення параметрів як експертних мереж, так і розподільної мережі. Градієнти обчислюються відносно функції втрат, а ваги коригуються за допомогою градієнтного спуску або його різновидів. Важливим аспектом є взаємодія між експертами та розподільною мережею: розподільна мережа навчається призначати вищі ймовірності $g_k(X)$ тим експертам, які краще прогнозують вихід для певних вхідних даних, тим самим мінімізуючи загальні втрати. Цей адаптивний процес вразливий до нестабільності роутера, що призводить до незбалансованих експертів, бо розподільна мережа неефективно маршрутизує вхідні дані.

Цей підхід базується на градієнті, на відміну від пізніших розробок, які можуть використовувати EM-алгоритм для MoE моделей з параметричними розподілами експертів, такими як гаусові суміші. Навчання на основі градієнта є особливо актуальним для реалізації нейронних мереж.

2.4.3 Переваги та теоретичні висновки

Модульна структура моделі MoE має кілька переваг над монолітними моделями:

- Масштабованість: Використовуючи прості експертні мережі, модель може масштабуватися для обробки великих і складних наборів даних, при цьому кожен експерт фокусується на підмножині.

- Інтерпретованість: Природній поділ праці між експертами дає розуміння того, як обробляються різні частини вхідного простору, що підвищує інтерпретованість моделі.

- Гнучкість: Здатність розподільної мережі адаптивно маршрутизувати вхідні дані гарантує, що модель може ефективно обробляти нерівномірні та нелінійні розподіли даних.

Теоретично, позиціонування MoE як мосту між модульними мережами та змагальним навчанням забезпечує новий погляд на ансамблеві методи. Вона пов'язує контрольоване навчання з принципами конкурентного навчання, де експерти змагаються за представлення підмножин даних, але в контрольованому контексті.

2.4.4 Релевантність MoE для кластеризації зображень

Хоча оригінальна модель MoE призначена для навчання із вчителем, її принципи можуть бути поширені на завдання навчання без вчителя, такі як кластеризація зображень, за допомогою концепції суміші автокодувальників (Mixture of Autoencoders).

У контексті кластеризації зображень Mixture of Autoencoders (MoA) може включати кілька автокодувальників, кожен з яких навчений представляти певний кластер зображень. Це аналогічно моделі MoE, де кожна експертна мережа спеціалізується на підмножині даних. Механізм розподільника у цьому неконтрольованому середовищі може бути методом призначення кожному зображенню найбільш відповідного автокодувальника на основі його особливостей, що полегшує кластеризацію.

Наприклад, у статті «Deep Unsupervised Clustering Using Mixture of Autoencoders»пропонується метод глибокої кластеризації, де кожен кластер представлений автокодувальником, а кластерна мережа трансформує дані в інший простір, щоб вибрати відповідний автокодувальник для реконструкції [5]. Цей підхід мінімізує похибку відбудови, уни-

каючи потреби в умовах регуляризації для запобігання колапсу даних, і демонструє значні покращення на корпусах зображень і текстів.

Зв'язок полягає в модульному та адаптивному характері як MoE, так і MoA. У MoE розподільна мережа динамічно спрямовує вхідні дані до експертів; у MoA подібний механізм розподіляє зображення між автокодувальниками, уможливлюючи ефективну кластеризацію. Ця аналогія підкреслює, як принципи спеціалізації та розподілу праці в MoE можуть бути адаптовані до неконтрольованого навчання, особливо для даних високої розмірності, таких як зображення.

3 Модель даних.

Для кластеризації зображень, набір даних розглядається як результат складного генеративного процесу, що формує розподіл суміші. Кожне зображення в цьому розподілі характеризується індивідуальною концентрацією латентних кластерів.

Щоб ефективно змоделювати цю притаманну структуру та виконати точну кластеризацію, ми вводимо нову архітектуру глибокого навчання: Суміш автокодувальників зі змінними концентраціями (Mixture of Autoencoders with Varying Concentrations, **MoA**+). Ця модель спеціально розроблена для обробки різного впливу різних латентних груп у наборі зображень.

MoA+ досягає цього шляхом синтезу принципів з двох ключових областей:

- Структури суміші експертів (MoE) 2.4, яка забезпечує потужну архітектуру для розподілу складних завдань між спеціалізованими підмоделями.

- Статистичної теорії моделей сумішей зі змінними концентраціями (MVC) 2.3, яка пропонує інструменти, зокрема мінімаксні ваги, для обробки неоднорідних внесків від компонентів суміші.

Інтегруючи ці концепції, безпосередньо шляхом модифікації цільового функціоналу тренування, MoA+ прагне подолати значні обмеження, які часто виникають під час застосування стандартних моделей MoE до завдань кластеризації без вчителя. Основною проблемою, яка розглядається, є «домінування експертів», коли деякі експерти непропорційно переважають під час навчання, перешкоджаючи виявленню збалансованих та значущих кластерів. Таким чином, модель даних MoA+ є варіантом класичної MoE, але з різницею в loss-функції, що дозволяє ефективно навчати роутер.

3.1 Огляд архітектури.

Загальна архітектура MoA+ відповідає парадигмі суміші експертів (MoE) 2.4, що складається з кількох експертних мереж та розподільної мережі, яка спрямовує вхідні дані до відповідних експертів. Цей архітектурний шаблон, зокрема використання автокодувальників як експертів для неконтрольованих завдань, таких як кластеризація, є визнаним підходом у цій галузі, дослідженим у різних дослідженнях (наприклад, [11], що базується на загальній концепції MoE). MoA+ адаптує цю структуру зі спеціалізованими компонентами та унікальним цільовим функціоналом навчання.

3.1.1 Експерти: Згорткові Автокодувальники (CAEs).

Експерти в моделі MoA+ реалізовані як згорткові автокодувальники (CAE). Кожен експерт ($k = 1, ..., M$) — це незалежна нейронна мережа, призначена для навчання стисненого представлення (кодування) вхідного зображення, а потім реконструкції зображення з цього представлення (декодування).

- Вхід: Вхідними даними для кожного експертного CAE є зображення X_i. Зазвичай воно представлене як багатовимірний тензор. Наприклад, кольорове зображення може бути 3-канальним

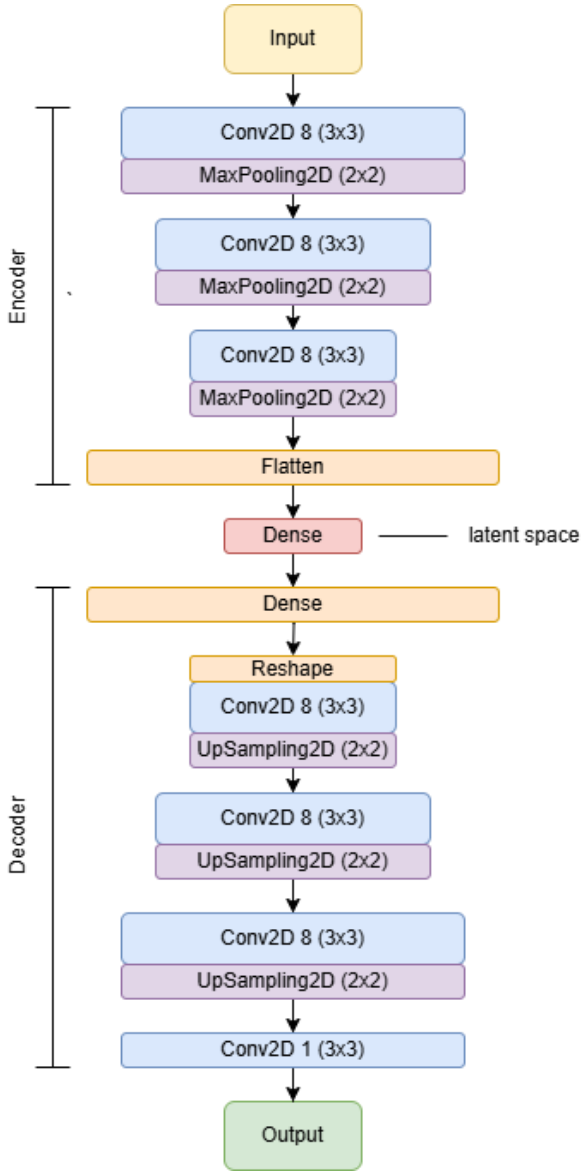


Рис. 3: Архітектура експерта CAE.

(RGB) тензором розмірності $(H, W, 3)$, де H – висота, W – ширина. Для зображень у градаціях сірого, таких як ті, що в наборі даних MNIST, це буде одноканальний тензор $(H, W, 1)$.

- Архітектура: CAE має класичну архітектуру кодер-декодер, використовуючи шари, спеціалізовані для просторових даних:
  - Encoder: Кодер побудований з трьох повторюваних блоків. Кожен блок складається з:
    * Шар Conv2D з 8 фільтрами та розміром ядра (3x3). Цей шар витягує просторові ознаки зі своїх вхідних даних (краї, текстури, форми). В якості функції активації використовується ReLU.
    * Шар MaxPooling2D з розміром пулу (2x2). Цей шар виконує зменшення дискретизації, зменшуючи її просторові розміри (висота та ширина) вдвічі, що допомагає створювати більш надійні ознаки (стійкі до невеликих зсувів та спотворень) та зменшує обчислювальне навантаження.

    Після третього шару об'єднання (pooling layer), шар Flatten перетворює кінцеву 3D-карту ознак (висота x ширина x фільтри) на 1D-вектор. Цей сплющений вектор потім пропускається через повноз'єднаний (Dense) шар. Вихід цього шару являє собою стиснутий вектор латентного простору, що фіксує важливу інформацію з вхідного зображення у низьковимірній формі.
  - Decoder: Декодер починається з іншого повноз'єднаного шару, який приймає вектор латентного простору як вхідні дані та трансформує його, потенційно готуючи до переформування. Шар переформування (Reshape) перетворює вихідний 1D-вектор із шару Dense назад у 3D-тензор з відповідними просторовими вимірами та глибиною (кількістю каналів),

щоб розпочати процес згорткової реконструкції. Потім декодер використовує три повторювані блоки, спрямовані на масштабування карт ознак до вихідних розмірів зображення. Кожен блок складається з:

* Шар Conv2D (8 фільтрів, ядро 3x3), аналогічний кодеру, що використовується тут для уточнення ознак під час реконструкції.
* Шар UpSampling2D розміром (2x2). Це подвоює висоту та ширину карти ознак, шляхом повторення рядків та стовпців, скасовуючи ефект шарів MaxPooling2D у кодувальнику.

Нарешті, застосовується завершальний шар Conv2D з 1 фільтром та ядром (3x3). Цей шар відображає ознаки з попереднього шару на кінцевий єдиний вихідний канал, створюючи реконструйоване зображення у градаціях сірого. У цьому заключному шарі для реконструкції зображення використовується сигмоїдна функція активації, щоб забезпечити масштабування значень пікселів між 0 та 1.

- Вихід: Виходом k-го експерта CAE є реконструйоване зображення $\hat{X}_{i,k}$, в ідеалі з такими ж розмірами (H, W, C), як і вхід X_i. Мета полягає в тому, щоб $\hat{X}_{i,k}$ було якомога ближчим до X_i.

- Мета: Кожен експерт CAE, дотримуючись цієї специфічної архітектури, навчається реконструювати вхідні зображення, пропускаючи їх через процес кодування-декодування. У рамках MoA+ мета навчання заохочує різних експертів спеціалізуватися на ефективній реконструкції певних підмножин даних, тим самим вивчаючи представлення, корисні для розрізнення різних кластерів.

3.1.2 Роутер: розподільна мережа.

Роутер, або розподільна мережа, працює паралельно з експертами. Його роль полягає не в реконструкції, а у визначенні відповідності кожного експерта для заданого вхідного зображення.

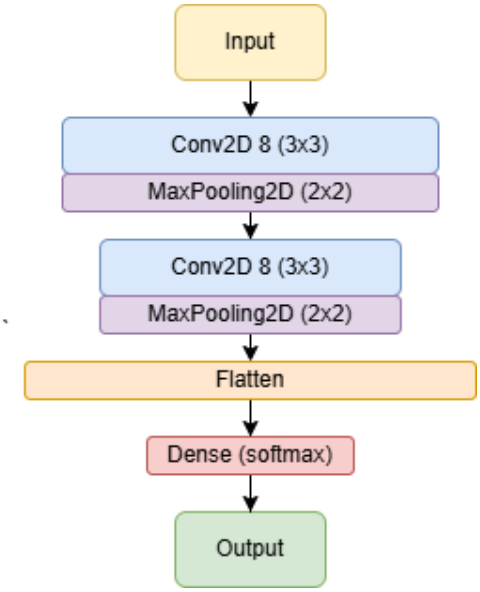


Рис. 4: Архітектура згорткового роутера.

- Вхід: Роутер приймає те саме вхідне зображення X_i (наприклад, тензор розмірності (H, W, C)), що й експерти.
- Архітектура: Розподільна мережа обробляє вхідне зображення за допомогою структури згорткової нейронної мережі (CNN), призначеної для вилучення ознак, релевантних для експертного відбору. Мережа починається з двох повторюваних блоків для вилучення ознак. Кожен блок містить:
  - Шар Conv2D з 8 фільтрами та розміром ядра (3x3) з ReLU функцією активації, що відповідає за визначення просторових шаблонів у вхідних даних.
  - Шар MaxPooling2D з розміром пулу (2x2), який виконує зменшення дискретизації карти ознак, зменшуючи її просторові

розміри та забезпечуючи певну інваріантність щодо переміщення.

Після другого шару об'єднання, отримана 3D-карта ознак пропускається через шар Flatten, перетворюючи її на 1D-вектор. Цей вектор потім безпосередньо подається в кінцевий вихідний шар: повноз'єднаний (Dense) шар з M вузлами, де M – кількість експертів. Найважливіше, що цей повноз'єднаний шар використовує функцію активації $softmax$, як показано на діаграмі. Функція $softmax$ гарантує, що вихідні дані шару є невід'ємними та дорівнюють 1, утворюючи коректний розподіл ймовірностей.

- Вихід: Вихід роутера для заданого вхідного значення X_i є вектором ймовірностей $(p^1_{i:n}, ..., p^M_{i:n})$.Кожен елемент $p^k_{i:n}$ представляє ймовірність, призначену роутером, того, що експерт k є найбільш підходящим вибором для вхідного значення X_i.

- Матриця ймовірностей P_n: Вектори ймовірностей, згенеровані роутером для всіх n зображень у наборі даних, збираються в матрицю $P_n = (p^k_{i:n})$. Ця матриця є важливою, оскільки вона формує основу для обчислення мінімаксних вагових коефіцієнтів, що використовуються у цільовому функціоналі навчання MoA+ (loss function).

3.1.3 Цільовий функціонал навчання MoA+.

Вихідні дані Експертів (реконструкції X_i) та Роутера (ймовірності P_n) об'єднуються у loss-функцію, яка керує процесом навчання. Щоб зрозуміти інновації в MoA+, корисно спочатку розглянути типову loss-функцію, яка використовується в стандартних моделях Mixture of Experts (Autoencoders).

Стандартна MSE-loss: Поширеним підходом у стандартних налаштуваннях MoE є мінімізація середньоквадратичної похибки (MSE)

між вхідним зображенням та зваженою комбінацією реконструкцій від усіх експертів. Як ваги зазвичай використовують прямі ймовірності p_i^k, що надаються роутером. Ця функція втрат може бути записана як:

$$\text{MSE}_{\text{standard}} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \sum_{k=1}^{M} p_i^k \hat{X}_{i,k})^2$$

У такому формулюванні модель спрямована на створення змішаної реконструкції на основі ймовірностей роутера. Однак, як зазначалося раніше і обговорювалося в літературі [12], такий підхід може призвести до домінування експерта, коли процес оптимізації надає перевагу одному експерту, якщо його початкові реконструкції або призначені ймовірності є дещо кращими, потенційно нехтуючи підготовкою інших експертів.

MoA+ loss-function: Модель MoA+ вводить спеціальну loss - функцію, розроблену з метою зменшення домінування експертів та заохочення збалансованих внесків експертів. Замість того, щоб зважувати реконструкції, MoA+ зважує індивідуальну помилку реконструкції кожного експерта, використовуючи мінімаксні ваги $a_{i:n}^k$, які виводяться з ймовірностей роутера P_n. MoA+ loss визначається як:

$$\text{loss}_{\text{MoA+}} = \left| \sum_{k=1}^{M} \sum_{i=1}^{n} a_{i:n}^k (X_i - \hat{X}_{i,k})^2 \right|.$$

Тут, $(X_i - \hat{X}_{i,k})^2$ представляє середньоквадратичну похибку (MSE) для реконструкції зображення X_i експертом k. Вирішальним елементом є вага $a_{i:n}^k$, мінімаксна вага, отримана з вихідних ймовірностей роутера P_n за допомогою $A_n = (P_n^T P_n)^{-1} P_n$.

Інтуїція та зв'язок з кластеризацією: Це модифіковане loss - формулювання сприяє збалансованій кластеризації. В той час як стандартні втрати можуть неявно поводитися як м'яка версія K-середніх (присвоюючи бали в першу чергу «найкращому» експерту на основі зваженої реконструкції), втрати MoA+ забезпечують баланс більш чітко. Включаючи глобальну структуру присвоєння балів завдяки $P_n^T P_n$ з мінімаксними вагами $a_{i:n}^k$ втрата карає конфігурації, де ймовірності

сильно сконцентровані на одному експерті по всьому набору даних. Терм інверсії матриці $(P_n^T P_n)^{-1}$ ефективно діє як регуляризатор, змушуючи модель розподіляти обов'язки більш рівномірно. Це гарантує, що всі експерти отримують достатню кількість навчальних сигналів, що концептуально схоже на те, що всі кластери K-середніх залишаються активними і представляють окремі частини даних, а не допускають розпаду кластерів.

Покращення та вирішення проблем: Використовуючи ці мінімаксні ваги, отримані на основі теорії сумішей зі змінними концентраціями, цільовий функціонал навчання MoA+ активно бореться з проблемою домінування експертів, що спостерігається у стандартному MoE. У результаті це призводить до:

- Більш стабільне та збалансоване навчання всіх експертів.
- Покращене розділення між вивченими кластерами, оскільки кожен експерт заохочується до ефективної спеціалізації.
- Зрештою, покращується продуктивність неконтрольованої кластеризації, що відображається в таких показниках, як NMI.

3.2 Тренування моделі.

Навчання моделі MoA+ передбачає одночасну оптимізацію параметрів як експертних згорткових автокодувальників (CAE), так і розподільної мережі, використовуючи спеціалізовану функцію втрат MoA+, призначену для сприяння збалансованій кластеризації.

3.2.1 Цільовий функціонал та проблеми оптимізації.

Основне завдання під час тренування - мінімізувати функцію втрат MoA+:

$$\text{loss}_{\text{MoA+}} = \left| \sum_{k=1}^{M} \sum_{i=1}^{n} a_{i:n}^{k} (X_i - \hat{X}_{i,k})^2 \right|.$$

Як раніше було визначено, X_i - це вхідне зображення, $\hat{X}_{i,k}$ - це реконструкція зображення CAE-експертом k, а $a^k_{i:n}$ - мінімаксні ваги, отримані з вихідних ймовірностей роутера $P_n = (p^k_{i:n})$ за допомогою $A_n = (P_n^T P_n)^{-1} P_n$.

Пряма оптимізація цієї функції втрат є складнішою порівняно зі стандартними loss-функціями глибокого навчання. Обчислення мінімальних ваг $a^k_{i:n}$ передбачає інверсію матриці $(P_n^T P_n)^{-1}$, де елементи P_n безпосередньо залежать від виходу розподільної мережі. Ця залежність робить обчислення градієнта під час зворотного поширення (backprop) неможливим за відсутності необхідних формул для $M > 4$, ніж для loss-функцій, де ваги є фіксованими або безпосередньо виводяться мережевим шаром.

Щоб уникнути наведених проблем оптимізації в MoA+ loss і керувати оптимізацією, використаємо відоме верхнє обмеження для $M = 2$ як цільовий функціонал оптимізації [16]:

$$\text{loss}_{\text{MoA+}} \leq \sqrt{\sum_{k=1}^{M} \frac{1}{\lambda_k^2}} \sqrt{\sum_{k=1}^{M} \left(\sum_{i=1}^{n} p^k_{i:n} (X_i - \hat{X}_{i,k})^2 \right)^2},$$

де λ_k - власні числа матриці Грама $\Gamma_n = P_n^T P_n$. Хоча ця границя невірна для $M > 2$, вона може слугувати ефективним евристичним або аналітичним інструментом для розуміння поведінки втрат або потенційно спрямовувати стратегії оптимізації. Вона пов'язує модифіковану loss-функцію з умовами, що включають початкові ймовірності $p^k_{i:n}$ та індивідуальними помилками реконструкції експертів.

3.2.2 Процедура тренування та оцінювання ефективності.

На продуктивність моделі MoA+ під час навчання впливають кілька гіперпараметрів, які необхідно встановити належним чином, часто шляхом експериментів:

- Кількість кластерів/експертів (M): Цей параметр є фундамен-

тальним і повинен бути обраний на основі попередніх знань або вивчення набору даних. Кількість експертів M відповідає бажаній кількості кластерів для неконтрольованої задачі. Процес навчання спрямований на те, щоб кожен експерт спеціалізувався на реконструкції даних, що належать переважно до одного з M латентних кластерів у наборі даних.

- Швидкість навчання (learning rate): Контролює розмір кроку під час оновлення параметрів оптимізатором.
- Розмір партії (батчів): Кількість зразків, що обробляються на кожній ітерації навчання.
- Кількість епох: загальна кількість проходів через навчальні дані.

Навчання зазвичай відбувається з використанням ітеративних алгоритмів оптимізації, поширених у глибокому навчанні, таких як Adam.

Після завершення тренування, приналежність зображення X_i до певного кластера визначається за індексом експерта k, який отримав найвищу ймовірність $p^k_{i:n}$ від роутера, тобто $k = argmax_j p^j_{i:n}$.

Первинною метрикою, обраною для оцінки якості кластеризації навченої моделі MoA+, є нормалізована взаємна інформація (Normalized Mutual Information, NMI) 2.1.4. Такий вибір зумовлений міркуваннями, що NMI широко використовується і приймається як стандартна метрика оцінки в літературі з некерованої кластеризації. Використання NMI дозволяє проводити пряме і справедливе порівняння продуктивності MoA+ з іншими відомими алгоритмами кластеризації, таким чином, ефективно оцінюючи внесок і покращення, які пропонує підхід MoA+.

4 Моделювання.

У цьому розділі представлено результати експериментального дослідження запропонованої моделі MoA+ (3) та її порівняння з базовою

моделлю Mixture of Experts (MoE), яка відповідає класичній моделі Хінтона (2.4). Експерименти проводилися на двох різних наборах даних: класичному датасеті рукописних цифр MNIST та спеціалізованому датасеті аудіозаписів військових дій (MAD), що представлені як мелспектрограми (візуальне представлення спектру частот звукового сигналу, де частотна вісь трансформована за мел-шкалою, яка імітує нелінійне сприйняття висоти звуку людським вухом). Оцінка якості кластеризації проводилася за допомогою метрики Нормалізованої Взаємної Інформації (NMI), а якість реконструкції — за допомогою Середньоквадратичної Помилки (MSE). Для обох експериментів була задана однакова конфігурація гіперпараметрів та кількість кластерів/експертів $M = 2$.

4.1 Результати на MNIST

Датасет MNIST є стандартним бенчмарком для задач розпізнавання образів та кластеризації. Він складається з 70 000 зображень рукописних цифр (0-9) розміром 28x28 пікселів у градаціях сірого. Для наших експериментів ми розглядали задачу бінарної кластеризації (M=2), де модель має розділити дані на дві групи.

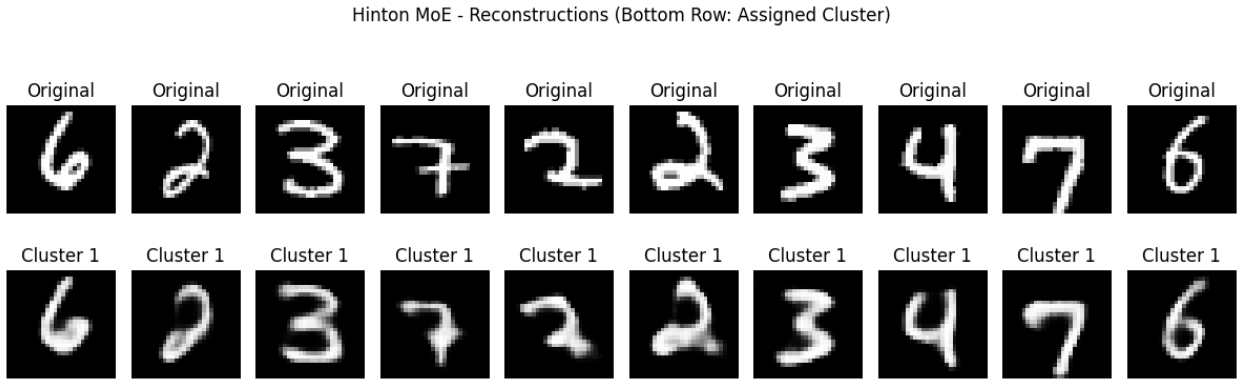
Таблиця 1: Результати кластеризації на датасеті MNIST (M=2)

Модель	**MSE**	**NMI**
Hinton MoE	0.013	0.000
MoA+	0.015	**0.276**

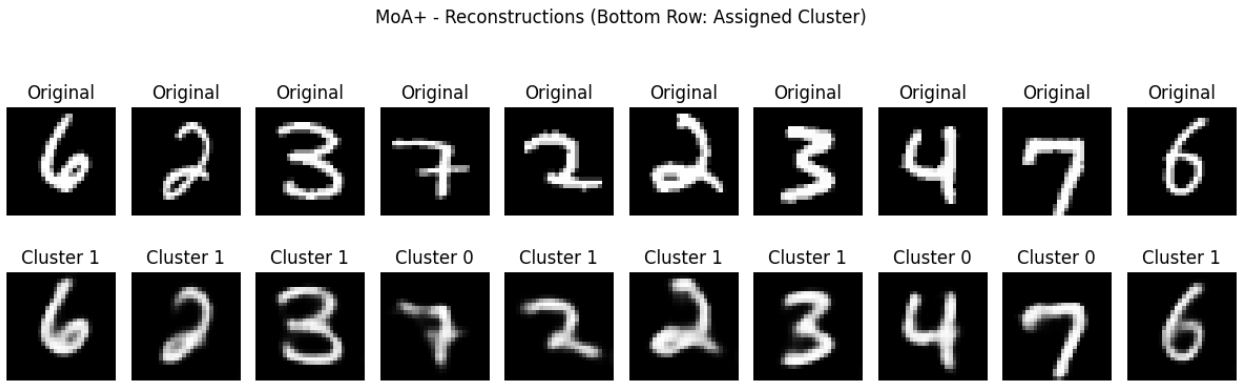
З Таблиці 1 видно, що модель MoA+ показує кращу якість кластеризації (NMI = 0.276) порівняно з близьким до нуля результатом моделі Hinton MoE. Це свідчить про те, що використання мінімаксних ваг у функції втрат MoA+ сприяє кращому розділенню класів. При цьому якість реконструкції (MSE) для обох моделей є порівнянною, з незначно вищим MSE для MoA+, що може бути пов'язано з тим, що модель MoA+ більше фокусується на збалансованому навчанні експер-

тів для кластеризації, аніж на мінімізації сумарної помилки реконструкції шляхом домінування одного експерта. Невелике погіршення MSE є прийнятною ціною за значне покращення якості кластеризації (NMI), що є основною метою роботи.

Візуальний аналіз результатів підтверджує ці висновки. На Рис. 5 показані приклади оригінальних зображень та їх реконструкцій моделями Hinton MoE та MoA+.



(а) Hinton MoE

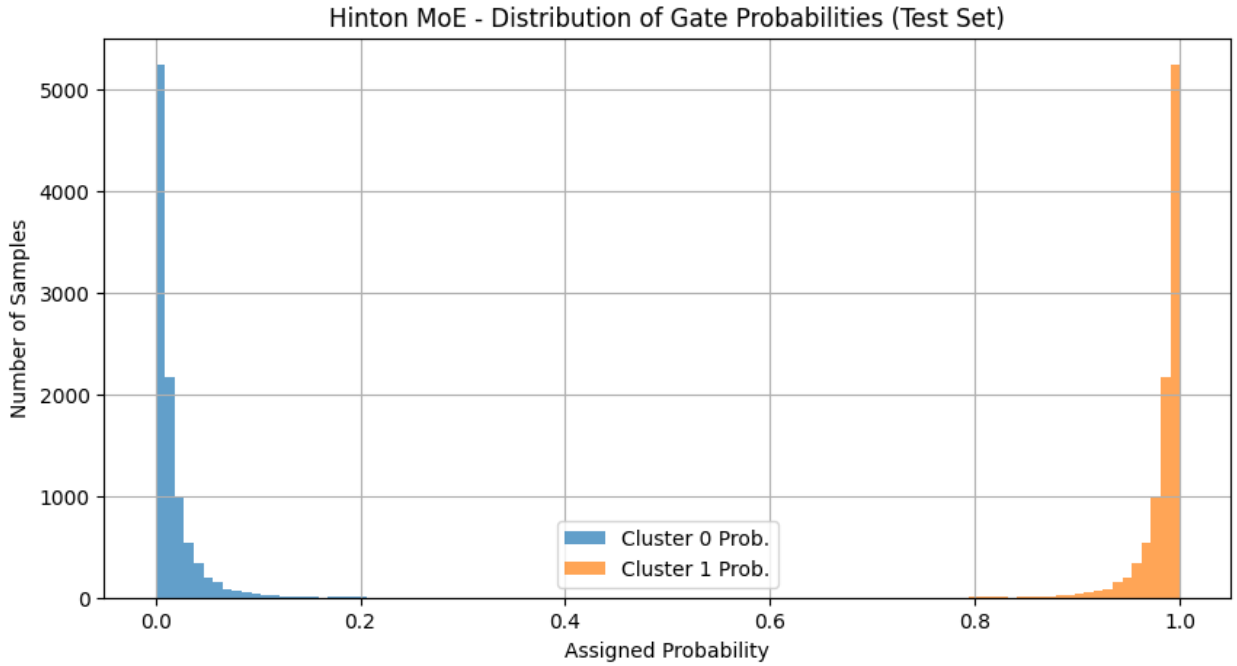


(б) MoA+

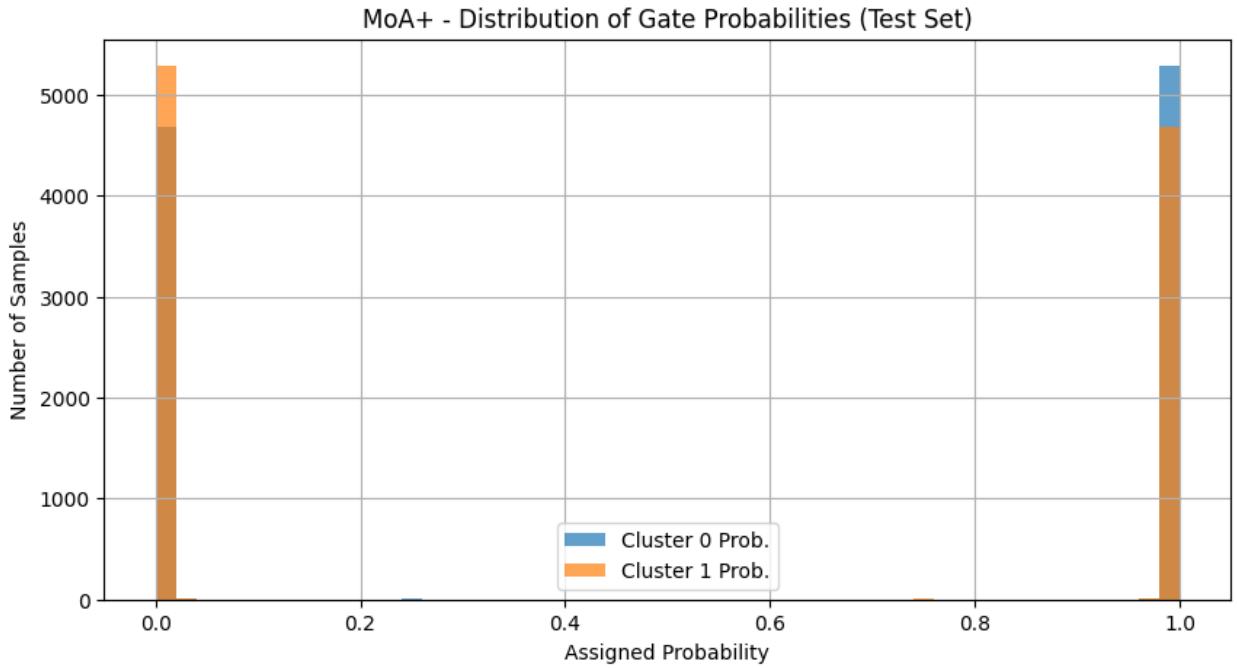
Рис. 5: Реконструкції зображень MNIST моделями Hinton MoE та MoA+. Нижній рядок показує призначений кластер.

З Рис. 5а видно, що модель Hinton MoE схильна призначати всі зображення до одного кластера, що свідчить про проблему домінування експерта. Натомість, Рис. 5б демонструє, що модель MoA+ призначає зображення до різних кластерів, що відповідає кращому значенню NMI.

Розподіл ймовірностей, що генеруються роутером (gating network), також ілюструє різницю в поведінці моделей (Рис. 9).



(а) Hinton MoE



(б) MoA+

Рис. 6: Розподіл ймовірностей роутера для моделей на MNIST.

Для Hinton MoE (Рис. 6а) розподіл ймовірностей чітко вказує на сильну незбалансованість роутера, який обирає завжди одного експерта. Навпаки, для MoA+ (Рис. 6б) спостерігаються два чіткі піки біля 0 та 1 для обох кластерів, що свідчить про більш впевнене та збалансоване призначення зображень до експертів.

Таблиці розподілу міток (Рис. 10) дають змогу глибше зрозуміти, як моделі групують справжні класи.

TrueLabel	Cluster 1
0	9.80
1	11.35
2	10.32
3	10.10
4	9.82
5	8.92
6	9.58
7	10.28
8	9.74
9	10.09

(а) Hinton MoE

TrueLabel	Cluster 0	Cluster 1
0	0.15	20.66
1	21.11	0.36
2	0.81	21.02
3	2.30	18.87
4	16.85	1.91
5	8.71	9.16
6	0.49	19.81
7	18.60	0.91
8	13.99	4.95
9	16.98	2.34

(б) MoA+

Рис. 7: Відсотковий розподіл справжніх міток MNIST по кластерах, отриманих моделями.

З Рис. 7а видно, що Cluster 1 моделі Hinton MoE містить всі приклади даних, формуючи один великий кластер. Для MoA+ (Рис. 7б), хоча ідеального розділення на два семантично значущі кластери всіх 10 цифр не досягнуто (що очікувано для M=2), спостерігається більш чітке групування. Наприклад, Cluster 1 переважно захоплює цифри 0, 2, 3, 6, тоді як Cluster 0 — інші. Це підтверджує вище значення NMI для MoA+.

4.2 Результати на MAD

Для оцінки моделі на більш складних та специфічних даних було використано датасет MAD. Цей датасет складається з мелспектрограм, отриманих з аудіозаписів, що містять звуки військових дій, такі як перестрілки та комунікація військових. Деякі записи містять суміш обох типів звуків. Дані походять з різних джерел, включаючи реальні записи з Російсько-Української війни, конфлікту в Афганістані, а також записи з військових тренувань. Складність цього датасету полягає у високій варіативності звуків, наявності шумів та перекритті різних звукових подій. Задача також розглядалася як бінарна кластеризація (M=2).

Результати порівняння моделей на зазначеному датасеті наведено в Таблиці 2.

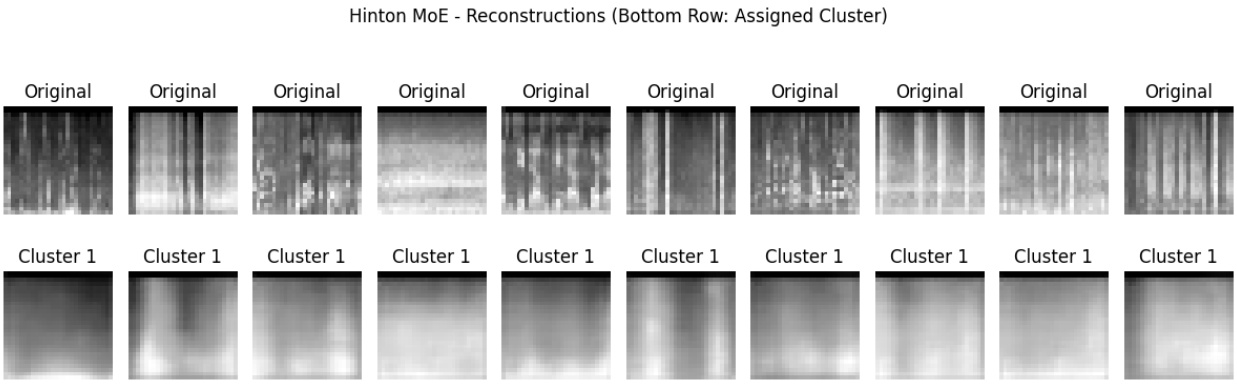
Таблиця 2: Результати кластеризації на датасеті MAD (M=2)

Модель	**MSE**	**NMI**
Hinton MoE	0.008	0.023
MoA+	0.010	**0.127**

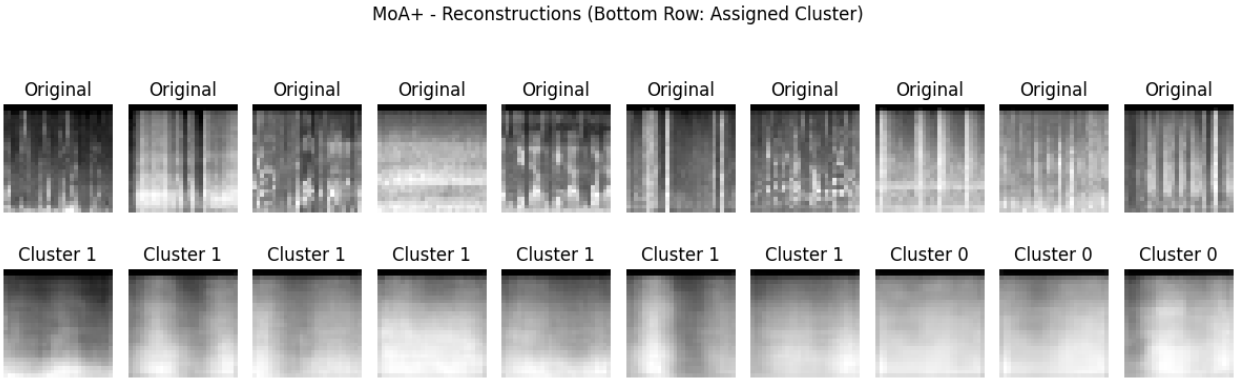
Як і на MNIST, модель MoA+ демонструє значно вищу якість кластеризації на датасеті MAD (NMI = 0.127) порівняно з базовою моделлю Hinton MoE (NMI = 0.023). Показники MSE для обох моделей знову є близькими. Незважаючи на те, що абсолютне значення NMI для MoA+ на MAD є нижчим, ніж на MNIST, приріст відносно базової моделі MoE є суттєвим, що підкреслює переваги запропонованого підходу на даних зі складною структурою.

Модель Hinton MoE (Рис. 8а) знову призначає більшість прикладів одному кластеру. Модель MoA+ (Рис. 8б) демонструє більш збалансоване призначення до двох кластерів.

Розподіл ймовірностей для MoA+ на MAD (Рис. 9б) показує, що хоча піки не такі чіткі, як на MNIST, все ж спостерігається тенденція до бімодального розподілу.



(а) Hinton MoE



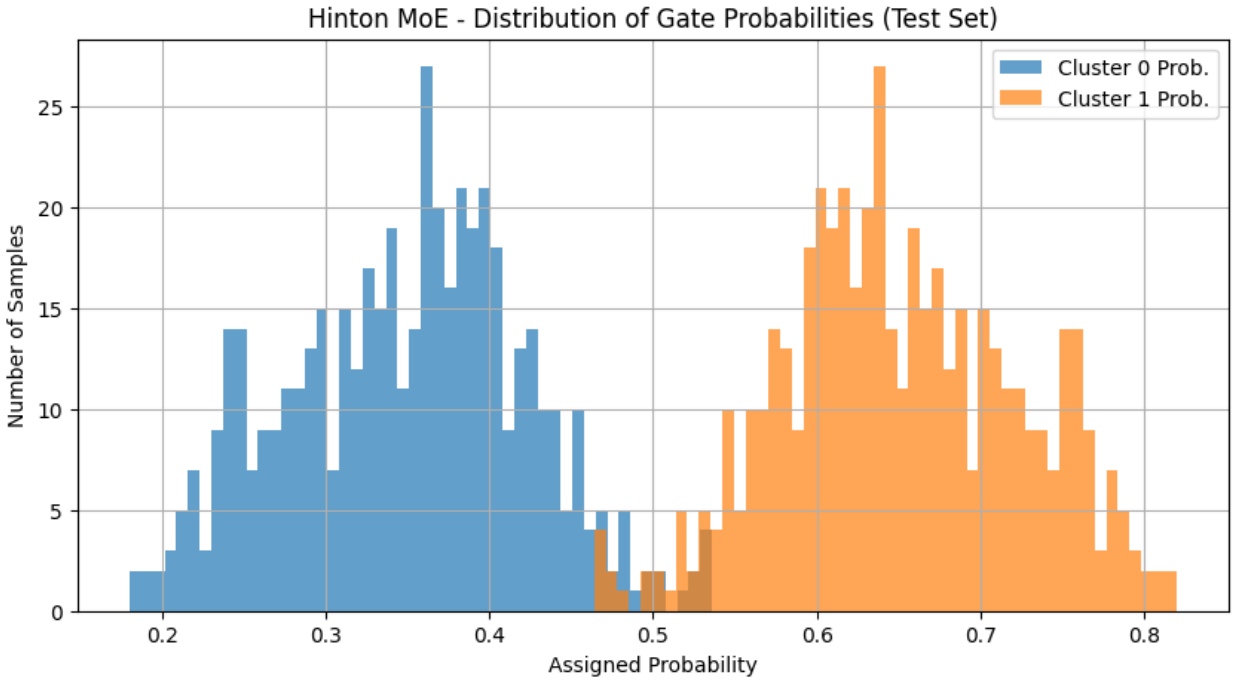
(б) MoA+

Рис. 8: Реконструкції мелспектрограм з датасету MAD моделями Hinton MoE та MoA+. Нижній рядок показує призначений кластер.

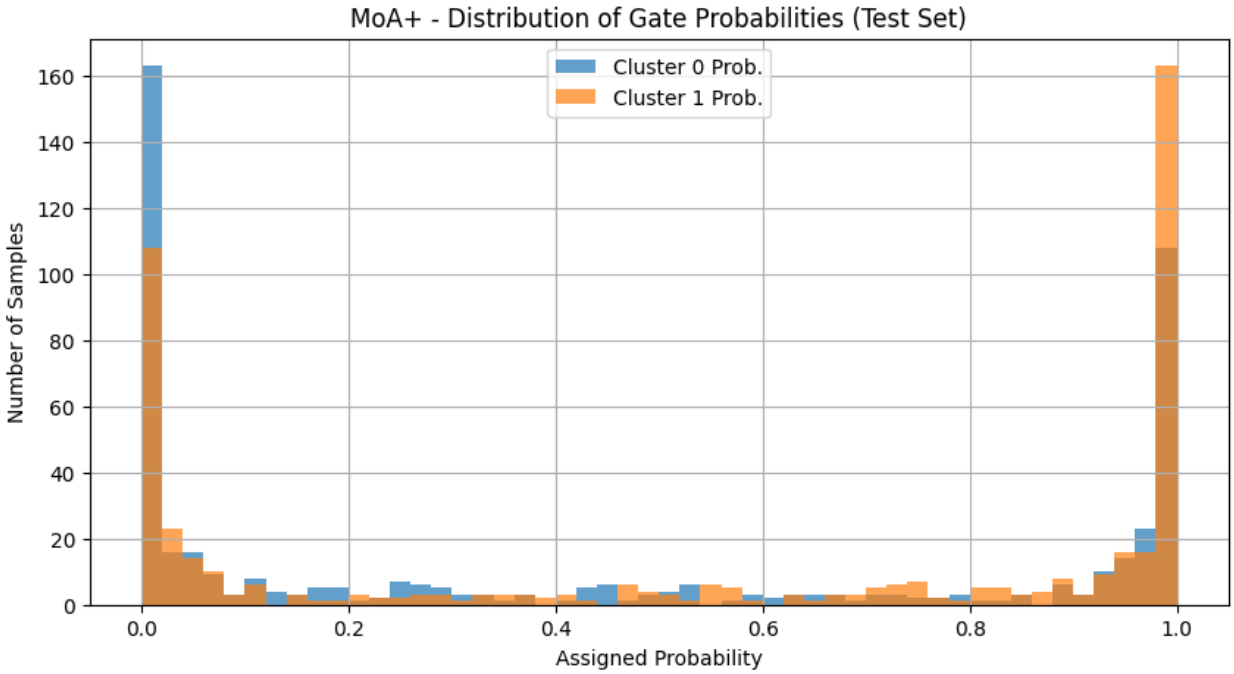
Таблиця розподілу міток для моделі MoA+ на датасеті MAD (Рис. 10б), де справжні відповідають двом основним типам звуків: перестрілки та комунікація, показує, як MoA+ намагається структурувати ці складні дані.

Рис. 10б для MoA+ на MAD показує, що Cluster 1 переважно захоплює приклади з міткою 0 - звуки комунікації військових, тоді як Cluster 0 має більшу частку прикладів з міткою 1 - звуки перестрілок. Це свідчить про те, що модель MoA+ знаходить характерні особливості в структурі даних.

Загалом, результати моделювання на обох датасетах демонструють перевагу запропонованої моделі MoA+ над стандартним підходом Hinton MoE з точки зору якості кластеризації. Це підтверджує ефективність використання мінімаксних ваг для боротьби з проблемою домінування



(а) Hinton MoE



(б) MoA+

Рис. 9: Розподіл ймовірностей роутера для моделей на MAD.

експертів та досягнення більш збалансованого навчання.

TrueLabel	Cluster 0	Cluster 1
0	88.89	41.63
1	11.11	58.37

(а) Hinton MoE

TrueLabel	Cluster 0	Cluster 1
0	19.23	59.86
1	80.77	40.14

(б) MoA+

Рис. 10: Відсотковий розподіл справжніх міток MAD по кластерах, отриманих моделями.

5 Висновки.

У даній роботі було представлено та досліджено нову архітектуру для неконтрольованої кластеризації зображень – MoA+ (Суміш автокодувальників зі змінною концентрацією). Основною метою роботи було розв'язання поширеної проблеми стандартних моделей типу Суміш Експертів (MoE), а саме – домінування експерта, коли один або декілька експертів стають надмірно активними, пригнічуючи навчання інших та призводячи до незбалансованої та неефективної кластеризації.

Ключовою інновацією моделі MoA+ є запровадження модифікованої loss-функції, яка інтегрує принципи з теорії сумішей зі змінними концентраціями (MVC), зокрема, шляхом використання мінімаксних ваг ($a^k_{i:n}$). Ці ваги, обчислені як $A_n = (P_n^T P_n)^{-1} P_n$ з матриці ймовірностей P_n, що генерується роутером, дозволяють зважувати індивідуальні помилки реконструкції кожного експерта. Такий підхід забезпечує більш рівномірний розподіл відповідальності між експертами, змушуючи їх спеціалізуватися на різних аспектах даних. Концептуально, це подібно до забезпечення активності всіх кластерів у методі K-середніх, запобігаючи їх колапсу або поглинанню одним домінуючим кластером. Таким чином, проблема домінування експерта ефективно поборена, що сприяє більш стабільному та збалансованому навчанню всіх компонент моделі.

Ефективність запропонованого підходу MoA+ була продемонстрована на двох наборах даних:

- Класичний бенчмарк MNIST: На цьому еталонному датасеті рукописних цифр модель MoA+ показала значне покращення якості кластеризації порівняно з базовою моделлю Hinton MoE. При M=2 експертах MoA+ досягла значення NMI 0.276, тоді як Hinton MoE має значення близько нуля, при порівнянних значеннях MSE (0.015 для MoA+ та 0.013 для Hinton MoE). Візуальний аналіз реконструкцій, розподілу ймовірностей роутера та розподілу справжніх міток по кластерах переконливо підтвердив, що MoA+ успішно долає домінування експерта та формує більш чіткі та збалансовані кластери.

- Спеціалізований датасет MAD (Military Activity Dataset): Цей набір даних, що складається з мелспектрограм аудіозаписів військових дій (перестрілки, комунікація військових під час Україно-Російської війни, конфлікту в Афганістані, тренувань), представляє собою значно складнішу задачу через високу варіативність, шуми та перекриття звукових подій. Навіть на цих складних реальних даних модель MoA+ продемонструвала свою перевагу: NMI склало 0.127, тоді як для Hinton MoE NMI був 0.023 (при MSE 0.010 для MoA+ та 0.008 для Hinton MoE). Суттєвий відносний приріст NMI вказує на здатність MoA+ виявляти структуру у складних, зашумлених даних, де традиційні підходи зазнають невдачі. Зокрема, модель показала здатність частково розрізняти звуки комунікації та перестрілок.

Успіх на датасеті MAD відкриває перспективи для практичного застосування подібних моделей. Наприклад, системи на основі MoA+ потенційно можуть використовуватися для автоматичного аналізу акустичної обстановки в зоні бойових дій у реальному часі. Це може включати ідентифікацію типів активності (наприклад, відрізнення звуків

перестрілки від пересування техніки або розмовної комунікації), що є критично важливим для ситуаційної обізнаності під час тактичних операцій, таких як розвідка, патрулювання або штурм укріплень чи лісосмуг. Здатність моделі кластеризувати нечіткі та змішані сигнали може допомогти у виявленні прихованих патернів та попередженні про потенційні загрози.

Таким чином, проведена робота демонструє, що запропонований метод MoA+, заснований на інноваційній функції втрат з мінімаксними вагами, є ефективним підходом для покращення якості неконтрольованої кластеризації зображень (та їх спектральних представлень) шляхом вирішення проблеми домінування експертів. Подальші дослідження можуть включати тестування моделі на ширшому спектрі даних, дослідження впливу більшої кількості експертів (M>2) та розробку адаптивних методів визначення оптимальної кількості кластерів.

6 Список літератури.

[1] Caron, Mathilde & Bojanowski, Piotr & Joulin, Armand & Douze, Matthijs. (2018). Deep Clustering for Unsupervised Learning of Visual Features. 10.48550/arXiv.1807.05520.

[2] Liu, Zhili, et al. "Task-customized masked autoencoder via mixture of cluster-conditional experts."arXiv preprint arXiv:2402.05382 (2024).

[3] Chazan, Shlomo & Gannot, Sharon & Goldberger, Jacob. (2018). Deep Clustering Based on a Mixture of Autoencoders. 10.48550/arXiv.1812.06535.

[4] R. A. Jacobs, M. I. Jordan, S. J. Nowlan and G. E. Hinton, "Adaptive Mixtures of Local Experts,"in Neural Computation, vol. 3, no. 1, pp. 79-87, March 1991, doi: 10.1162/neco.1991.3.1.79.

[5] Zhang, Dejiao et al. “Deep Unsupervised Clustering Using Mixture of Autoencoders.” ArXiv abs/1712.07788 (2017): n. pag.

[6] Mienye, Domor & Swart, Theo. (2025). Deep Autoencoder Neural Networks: A Comprehensive Review and New Perspectives. Archives of Computational Methods in Engineering. 1-20. 10.1007/s11831-025-10260-5.

[7] Ian Goodfellow and Yoshua Bengio and Aaron Courville. (2016). Deep Learning. MIT Press. http://www.deeplearningbook.org

[8] Zhang, Hanze & Huang, Yangxin. (2015). Finite Mixture Models and Their Applications: A Review. Austin Biometrics and Biostatistics. 2. 1013.

[9] Maiboroda, R., Miroshnychenko, V., & Sugakova, O. (2025). Estimation of Concentrations Parameters in the Model of Mixture with Varying Concentrations. Austrian Journal of Statistics, 54(1), 1–16. https://doi.org/10.17713/ajs.v54i1.1953

[10] Liubashenko, Daryna & Maiboroda, Rostyslav. (2015). Linear regression by observations from mixture with varying concentrations. Modern Stochastics: Theory and Applications. 2. 1. 10.15559/15-VMSTA41.

[11] L. Nalmpantis, I. Varlamis, ”Deep Unsupervised Clustering with Mixture of Autoencoders,”arXiv preprint arXiv:1910.07763v3, 2019.

[12] B. Zoph, I. Bello, S. Kumar, et al., ”ST-MoE: Designing Stable and Transferable Sparse Expert Models,”arXiv preprint arXiv:2202.08906, 2022.

[13] Wheeldon A, Serb A. A study on the clusterability of latent representations in image pipelines. Front Neuroinform. 2023 Feb

16;17:1074653. doi: 10.3389/fninf.2023.1074653. PMID: 36873564; PMCID: PMC9978803.

[14] Pintelas, Emmanuel & Livieris, Ioannis & Pintelas, P.. (2021). A Convolutional Autoencoder Topology for Classification in High-Dimensional Noisy Image Datasets. Sensors. 21. 7731. 10.3390/s21227731.

[15] Maiboroda R. E., Sugakova O. V. Statistics of mixtures with varying concentrations with application to DNA microarray data analysis // Journal of nonparametric statistics. 24 , No 1 201–205 (2012)

[16] Mikushova, O., & Miroshnychenko, V. (2025). Differentiable Bounds on Loss Functions in the Mixture Model with Varying Concentrations. In Proceedings of the XXIII International Scientific - Practical Conference «Shevchenkivska Vesna – 2025» (p. 33). Kyiv, Ukraine: Taras Shevchenko National University of Kyiv.