УДК: 004.85

DOI: 10.15407/fmmit2025.40.089

Аналіз можливостей застосування моделі Vision Grid Transformer для аналізу структури документів українського бухгалтерського обліку

Максим Коростіль[1], Ілона Лагун[2]

[1] аспірант, Національний університет «Львівська політехніка», 79013, м. Львів, вул. Степана Бандери, 12, e-mail: maksym.i.korostil@lpnu.ua

[2] кандидат технічних наук, Національний університет «Львівська політехніка», 79013, м. Львів, вул. Степана Бандери, 12, e-mail: ilona.i.lahun@lpnu.ua

У сучасних умовах цифрової трансформації зростає потреба в автоматизації обробки бухгалтерських документів, зокрема в Україні, де значна частина первинної документації зберігається у паперовому вигляді або у форматі сканованих зображень. Ефективне вилучення інформації з таких документів вимагає застосування передових методів штучного інтелекту, зокрема глибокого навчання та мультимодального аналізу даних. У статті розглянуто можливість застосування модклі Vision Grid Transformer (VGT) для аналізу структури українських бухгалтерських документів. Модель VGT поєднує в собі два інформаційні потоки – візуальний (на основі Vision Transformer, ViT) та текстово-просторовий (на основі Grid Transformer, GiT), що забезпечує комплексне представлення документа як за зовнішнім виглядом, так і за змістом. Додаткову гнучкість моделі забезпечують методи попереднього навчання – MGLM (Masked Grid Language Modeling) та SLM (Segment Language Modeling), які дозволяють вивчати як локальні, так і глобальні контекстуальні залежності між текстовими елементами. У дослідженні акцентовано увагу на особливостях адаптації моделі VGT до українського контексту. Окреслено головні виклики, серед яких – відсутність якісних публічних анотованих датасетів українською мовою, необхідність високоточного оптичного розпізнавання символів (OCR) для кириличних шрифтів, проблеми з розпізнаванням рукописного тексту (HTR), а також складнощі, пов'язані з бухгалтерською термінологією та абревіатурами.

Ключові слова: *аналіз структури документа, глибоке навчання, оптичне розпізнавання символів, рукописне розпізнавання тексту, автоматизація бухгалтерського обліку, україномовні бухгалтерські документи*.

Вступ. Штучний інтелект (ШІ), зокрема, моделі глибокого навчання продемонструють значний потенціал у вирішенні завдань, пов'язаних з обробкою та аналізом неструктурованих даних – як текстових, так і візуальних. Успішне застосування таких моделей у сфері аналізу документів для різних мов та галузей створює передумови для їхнього використання в українській бухгалтерській практиці. Ключовими аспектами такого аналізу є розпізнавання структури документа (Document Layout Analysis, DLA) та виймання інформації (Information Extraction, IE) [1]. Перший етап передбачає виявлення складових документа — текстових блоків, заголовків, таблиць, зображень тощо. Надалі, на основі цих структурних компонентів, виконується вилучення релевантних даних.

Серед сучасних підходів до аналізу документів помітну увагу дослідників

привертає архітектура Vision Grid Transformer (VGT) – мультимодальна модель, здатна обробляти як візуальні, так і текстові характеристики документів [2]. Це робить її потенційно придатною для роботи з комплексними макетами бухгалтерських документів, де важлива як візуальна структура, так і змістовий контекст.

Наразі більшість наявних досліджень орієнтовані на англомовні документи, структура яких часто регламентується правовими або адміністративними нормами інших країн. Водночас адаптація подібних моделей до аналізу україномовних документів, особливо з бухгалтерського обліку, залишається актуальним та недостатньо дослідженим напрямом.

VGT відноситься до класу двопотокових мультимодальних моделей, які поєднують візуальні й текстові потоки для досягнення повного розуміння документа [2]. Її функціональність базується на сучасних методах попереднього навчання, що дозволяє моделі вивчати закономірності структури документів навіть на непроанотованих даних.

Основною метою є дослідження можливості використання архітектури VGT для аналізу українських бухгалтерських документів, які мають значну варіативність у структурі та оформленні.

1. Архітектура моделі Vision Grid Transformer (VGT)

Архітектура VGT (рис.1) ґрунтується на двох головних принципах: мультимодальності та ієрархічному розумінні документа.

Мультимодальність забезпечує виокремлення візуальних ознак, що відображають зовнішній вигляд документа (шрифти, лінії, розташування блоків), за допомогою Vision Transformer (ViT). Одночасно текстові й просторові ознаки обробляються Grid Transformer (GiT) [2]. Такий підхід дозволяє сформувати комплексне представлення документа.

Ієрархічне розуміння документа дозволяє моделювати семантичні зв'язки в документі на різних рівнях деталізації. Це забезпечує більш повне та контекстуалізоване розуміння документа.

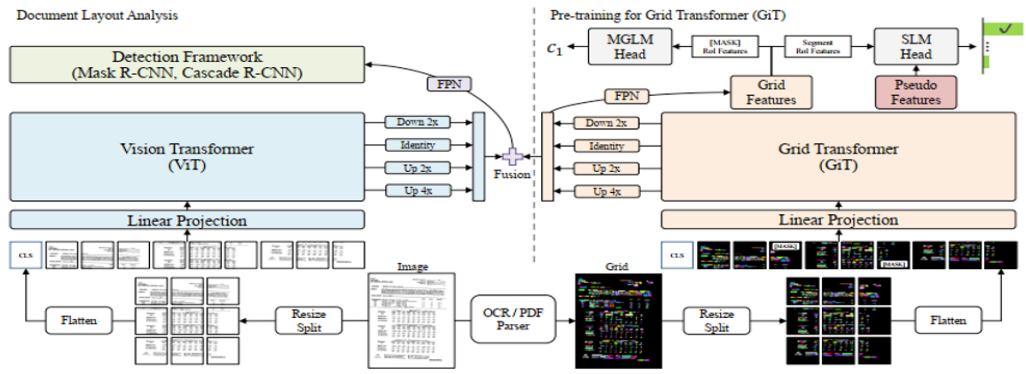

Рис. 1. Архітектура моделі VGT [2]

Потік Vision Transformer (ViT Stream) призначений для обробки візуальної інформації та відповідає за аналіз документа як зображення. Документ розглядається як візуальна сцена, де розташування елементів, їх розміри, шрифти та інші графічні

ознаки несуть важливу інформацію.

Зображення документа спочатку розбивається на послідовність прямокутних фрагментів (патчів), що взаємно не накладаються. Кожен патч потім лінійно проектується у векторне представлення (токен) фіксованої розмірності. До цих токенів додаються позиційні вбудовування (positional embeddings), що кодують інформацію про відносне розташування патчів, та спеціальний класифікаційний токен, що є частиною механізмів уваги (self-attention) і може використовуватися для агрегації інформації про весь документ. Для екстракції візуальних ознак можуть використовуватися попередньо навчені моделі ViT, такі як Document Image Transformer (DiT) [3], що дозволяє переносити знання, отримані на великих масивах зображень документів. Таким чином, цей компонент здатний ідентифікувати візуально відокремлені блоки, такі як заголовки документів ("Рахунок-фактура", "Акт"), табличні частини з переліком товарів чи послуг, області з підписами та печатками, а також окремі реквізити, що мають чітке візуальне оформлення (наприклад, виділені рамкою або розташовані в окремих секціях). Це особливо важливо для документів з варіативними шаблонами, де точне положення елементів може змінюватися.

Потік Grid Transformer (GiT Stream) аналізує текстовий вміст документа та його просторову організацію. Текст документа представляється у вигляді двовимірної ґратки (grid), де кожен елемент ґратки (наприклад, окреме слово, символ або фрагмент тексту) має як своє текстове значення, так і координати, що визначають його положення на сторінці [2]. Враховуючи таку двовимірну природу розташування тексту на сторінці, GiT може ефективно фіксувати взаємозв'язки між сусідніми та віддаленими текстовими елементами в контексті їхнього візуального розміщення [2]. Це дає змогу вловлювати дрібні особливості типографії, такі як стиль шрифту, нахил тощо, які можуть бути важливими для розуміння структури документа.

Наприклад, GiT може допомогти відрізнити поля "Код ЄДРПОУ Покупця" від "Код ЄДРПОУ Постачальника", навіть якщо вони мають схоже візуальне оформлення, аналізуючи навколишній текст (наприклад, слова "Покупець" або "Постачальник" поруч). Також GiT може розпізнавати логічні зв'язки між полями, наприклад, належність певної адреси конкретному контрагенту.

Для формування єдиного представлення документа модуль мультимодального злиття (Multi-modal Fusion) об’єднує обидва потоки – візуальний і текстово-просторовий, що дозволяє краще диференціювати елементи з подібним виглядом, але різним значенням. При цьому використовуються спеціальні механізми багатомасштабного злиття (multi-scale fusion) для інтеграції інформації з обох потоків [2]. Таким чином, якщо візуально два блоки тексту виглядають однаково, але мають різний текстовий зміст, що вказує на їхню різну семантичну роль (наприклад, адреса доставки та юридична адреса), мультимодальний аналіз дозволить правильно їх розрізнити. І навпаки, якщо текстовий зміст поля неоднозначний (наприклад, просто набір цифр), його візуальне розташування (наприклад, поруч з написом "ІПН") допоможе правильно його класифікувати.

Двопотокова архітектура VGT, що окремо обробляє візуальний та текстово-просторовий потоки з їх подальшим злиттям, є ключовою перевагою для аналізу бухгалтерських документів. У таких документах візуальне оформлення (наявність

ліній, таблиць, специфічних шрифтів для певних полів) та текстовий зміст (назви полів, числові значення, коди, одиниці виміру) є однаково важливими для правильної інтерпретації. Наприклад, ViT-компонент може ефективно розпізнати візуальні блоки, що відповідають реквізитам сторін, табличній частині або підписам, тоді як GiT-компонент може проаналізувати текстовий зміст цих блоків та їх взаємне розташування. Це дозволяє, наприклад, відрізнити поле "Сума без ПДВ" від поля "Сума ПДВ" не лише за їх візуальним розташуванням, але й за текстовим контекстом, навіть якщо візуально ці поля оформлені схоже. Злиття інформації з обох потоків допомагає уникнути помилок, які могли б виникнути при використанні лише однієї модальності.

2. Методи попереднього навчання для VGT

Ефективність VGT значною мірою залежить від стратегій попереднього навчання, які дозволяють моделі вивчати загальні закономірності структури та змісту документів на великих непроанотованих наборах даних [3, 4, 5].

VGT використовує два методи попереднього навчання – MGLM (Masked Grid Language Modeling) та SLM (Segment Language Modeling), які дозволяють моделі відновлювати контекстно масковані елементи в сітці та моделювати розташування більших текстових блоків [2]. Це особливо важливо для документів із варіативними шаблонами, як у випадку з українськими бухгалтерськими формами.

MGLM [2] – метод маскування окремих токенів у grid та відновлення їх із контексту. Під час MGLM обробки частина текстових елементів (токенів) у вхідній ґратці маскується (замінюється спеціальним символом або випадковим токеном), і завдання моделі полягає у передбаченні оригінальних замаскованих елементів на основі контексту – тобто, оточуючих видимих елементів у ґратці. Це допомагає GiT вивчати складні контекстуальні залежності між текстовими елементами в їхньому 2D-розташуванні.

SLM [2] (Segment Language Modeling) – метод маскування цілих сегментів (рядків/блоків) для кращого розуміння сегментної семантики, що спрямоване на навчання моделі розумінню взаємного розташування більших текстових сегментів або блоків у документі.

Значення попереднього навчання для VGT важко переоцінити, оскільки вона є однією з перших архітектур, що активно використовує попереднє навчання на основі ґраткового представлення тексту для глибинного 2D семантичного розуміння документів. Це дозволяє моделі ефективно навчатися на великих обсягах непроанотованих документів, що є критично важливим, враховуючи значний дефіцит великих, якісно анотованих датасетів, особливо для специфічних доменів, таких як українська бухгалтерська документація. Методи попереднього навчання MGLM та SLM можуть бути особливо корисними для аналізу типових, та водночас варіативних шаблонів українських бухгалтерських документів. Хоча існують стандартизовані форми для деяких документів, підприємства часто мають право розробляти власні шаблони, дотримуючись лише переліку обов'язкових реквізитів. MGLM, маскуючи та передбачаючи окремі слова, числа або коди в сітці, може допомогти моделі вивчити типові значення та формати для певних полів (наприклад, характерний формат дати, довжину коду ЄДРПОУ, типові одиниці виміру). SLM, моделюючи взаємне

розташування сегментів (наприклад, блоку реквізитів постачальника відносно блоку реквізитів покупця, або табличної частини відносно підсумкових сум), може допомогти моделі зрозуміти загальну логічну структуру документа, навіть якщо конкретні поля дещо зміщені, мають різне формулювання або візуальне оформлення в документах різних підприємств. Це підвищує стійкість моделі до варіацій у форматах та покращує її здатність до узагальнення.

Ефективність GiT-компонента також залежить від обраної гранулярності ґратки. Для бухгалтерських документів, де важливі як окремі цифри та короткі коди, так і багатослівні назви полів (наприклад, "Найменування підприємства, від імені якого було складено документ"), вибір оптимального розміру елемента ґратки (окремий символ, слово, чи, можливо, фраза) та методу агрегації інформації в комірках цієї ґратки стає критичним. Якщо токен – це слово, то для числових полів або кодів це може бути оптимально. Однак, якщо поле містить декілька слів, необхідно визначити, як це буде представлено в ґратці і як GiT оброблятиме такі багатослівні комірки. Неправильно обрана гранулярність може призвести до втрати важливої інформації або до неможливості точно розрізнити близько розташовані, але семантично різні поля, що негативно вплине на точність розпізнавання.

3. Переваги та недоліки архітектури VGT

Дослідження, у яких згадується архітектура VGT, демонструють її високу ефективність на низці відомих публічних бенчмарків для аналізу структури документів [2]. Зокрема, VGT досягла нових state-of-the-art результатів на таких датасетах, як PubLayNet, DocBank, та новоствореному D^4LA. Наприклад, на бенчмарку PubLayNet [6] було продемонстровано покращення середньої точності (mAP) з 95.7% до 96.2%, на DocBank [7] з 79,6 % до 84,1 % і на власному D^4LA-датасеті – із 67,7 % до 68,8 % mAP [2].

До ключових переваг VGT, що сприяють таким результатам, можна віднести:

1. Покращене розуміння текстово-орієнтованих класів. Завдяки ефективному використанню текстових ознак, виокремлених GiT-компонентом, VGT демонструє кращі результати в розпізнаванні елементів структури, для яких текстовий зміст є визначальним (наприклад, заголовки, списки, абзаци тексту).
2. Здатність вивчати семантику на рівні токенів та сегментів. Завдяки завданням попереднього навчання MGLM та SLM, VGT здатна моделювати семантичні зв'язки як на мікрорівні (окремі слова/токени), так і на макрорівні (більші текстові блоки/сегменти) [2].

Незважаючи на значні переваги, архітектура VGT, як і будь-яка складна модель глибокого навчання, має певні обмеження та потенційні недоліки. Моделі на основі Transformer, до яких належить і VGT, відомі своєю високою обчислювальною складністю та значними вимогами до ресурсів, особливо під час навчання на великих датасетах, та інференсу на зображеннях документів з високою роздільною здатністю або великою кількістю тексту. Мультимодальні методи, що поєднують кілька потоків обробки, часто є повільнішими порівняно з одномодальними підходами [5]. Наприклад, за деякими даними, VGT при використанні графічного процесора (GPU) GTX 1070 обробляє одну сторінку документа приблизно за 1.75 секунди, тоді як на центральному процесорі (CPU) i7-8700 цей час зростає до 13.5 секунд [6]. Такі

показники швидкодії можуть бути неприйнятними для деяких сценаріїв використання, що вимагають обробки в реальному часі або обробки дуже великих обсягів документів.

Серед недоліків також можна виділити залежність від якості оптичного розпізнавання символів (OCR). Ефективність GiT-потоку, який відповідає за аналіз текстово-просторової інформації, безпосередньо залежить від точності попереднього етапу – OCR та правильного визначення координат розпізнаного тексту для формування 2D ґратки. Помилки OCR (неправильно розпізнані символи, пропущені слова, неправильне розбиття на слова) можуть суттєво погіршити якість вхідних даних для GiT і, як наслідок, знизити загальну точність аналізу структури документа.

Крім цього існує потреба у великій кількості даних для попереднього навчання. Хоча стратегії самонавчання (self-supervised learning), такі як MGLM та SLM, частково вирішують проблему необхідності проанотованих даних, ефективне попереднє навчання VGT все ж таки потребує доступу до дуже великих наборів документів для вивчення загальних закономірностей.

3. Критичні виклики при застосуванні VGT до аналізу документів українського бухобліку

Загалом, застосування передових архітектур, таких як VGT, для аналізу україномовної бухгалтерської документації пов'язане з низкою специфічних викликів, які можуть суттєво вплинути на кінцеву ефективність системи.

Якість вхідних даних для текстового потоку VGT (GiT-компонента) безпосередньо залежить від точності попереднього етапу – оптичного розпізнавання символів (OCR) для друкованого тексту та розпізнавання рукописного тексту (HTR) для рукописних вставок. В контексті цього, перед застосуванням VGT постають наступні виклики:

1. Якість сканування та зображень. Низька роздільна здатність документів, поганий контраст між текстом та фоном, наявність шумів, перекосів, плям або інших артефактів сканування можуть значно ускладнити роботу OCR-систем, призводячи до помилок у розпізнаванні символів, слів або цілих рядків. Це, в свою чергу, призведе до передачі некоректної або неповної інформації до GiT-компонента VGT.
2. Специфічні символи української кирилиці. Українська абетка містить унікальні символи, такі як 'і', 'ї', 'є', 'ґ'. Важливо, щоб OCR-система коректно розпізнавала ці символи та не плутала їх зі схожими символами з інших кириличних алфавітів (наприклад, російської), особливо якщо OCR-модель навчалася на змішаних даних або недостатньо адаптована до української мови.
3. Проблеми з HTR. Хоча переважна більшість бухгалтерських документів є друкованими, деякі важливі елементи, такі як підписи відповідальних осіб, прізвища та ініціали, резолюції керівників, або навіть суми прописом, можуть бути заповнені рукописним способом. Якість HTR для української мови, особливо для коротких фрагментів тексту або нерозбірливого почерку, залишається значним викликом [8]. Дослідження показують, що навіть для спеціалізованих моделей HTR українських рукописів, частка помилок на рівні символів (Character Error Rate, CER) може становити близько 4.2% [8], що є

недостатнім для точного виймання критично важливих даних з бухгалтерських документів. Загальні виклики, пов'язані з OCR кириличних шрифтів, особливо в історичних або низькоякісних документах [9], також можуть бути частково релевантними. Інтеграція сучасних мовних моделей для постобробки та корекції результатів OCR/HTR може дещо покращити ситуацію, але не вирішує проблему повністю [8].

Якість OCR є своєрідним "вузьким місцем" для мультимодальних моделей типу VGT. Навіть якщо ViT-компонент ідеально розпізнає візуальну структуру документа (наприклад, точно локалізує поле для суми ПДВ), помилки OCR у текстовому потоці, що надходить до GiT-компонента, можуть призвести до неправильної класифікації цього поля або виймання невірних даних. Наприклад, якщо OCR розпізнає напис "Сума ПДВ" як "Сума ПДБ" або неправильно прочитає числове значення, то GiT отримає спотворений вхід, і навіть ідеальна робота ViT та модуля злиття не зможе гарантувати коректний кінцевий результат. Це особливо актуально для української мови з її унікальними кириличними символами та потенційною нестачею високоякісних OCR-рушіїв, спеціально навчених на шрифтах та типах документів, що використовуються в українській бухгалтерії.

Наступним критичним викликом є дефіцит даних та анотування. На сьогоднішній день спостерігається значний дефіцит великих, загальнодоступних датасетів саме українських первинних бухгалтерських документів, які були б анотовані для завдань DLA. Існуючі популярні бенчмарки, такі як PubLayNet, DocBank, D^4LA, на яких VGT демонструє високі результати, не містять специфічних українських бухгалтерських документів [2]. Датасет WordScape [10], хоча й містить багатомовні документи, має значно менше українських документів порівняно з російськими чи англійськими, і ці документи не обов'язково є бухгалтерськими. Деякі комерційні джерела [11] згадують про наявність "Ukrainian OCR Image Datasets", що включають рахунки-фактури, але ці датасети, ймовірно, призначені для навчання OCR-систем і не обов'язково містять детальну розмітку структури документа (логічні регіони, поля, таблиці) для DLA. Інші датасети українською мовою стосуються інших предметних областей (аналіз мирних переговорів, розпізнавання мовлення, загальні тексти) і не підходять для навчання моделей обробки бухгалтерських документів.

Створення якісних анотованих датасетів є надзвичайно трудомістким та дороговартісним процесом. Анотування бухгалтерських документів вимагає не лише уважності та точності, але й експертних знань у галузі бухгалтерського обліку для правильної ідентифікації та класифікації полів і структурних елементів. Крім цього, бухгалтерські документи містять конфіденційну комерційну інформацію (фінансові показники, дані про контрагентів тощо). Це створює значні перешкоди для збору таких документів у великих обсягах, їхньої публікації та використання для навчання моделей сторонніми розробниками.

Таким чином, головним бар'єром для ефективного використання VGT може стати не стільки сама архітектура моделі, скільки забезпечення її адекватними та релевантними навчальними даними для цільового домену.

Слід також згадати про мовно-специфічні нюанси адаптації моделі. Українська мова має свої лексичні, граматичні та орфографічні особливості, які необхідно враховувати. Більшість сучасних state-of-the-art моделей для обробки природної мови

та аналізу документів (включаючи базові компоненти, на яких може будуватися VGT, наприклад, BERT-подібні архітектури для GiT) попередньо навчаються на величезних корпусах англомовних текстів. Українська мова, порівняно з англійською, належить до низькоресурсних мов, для яких обсяги доступних текстових даних значно менші. Адаптація таких моделей для ефективної роботи з українською мовою є нетривіальним завданням і вимагає спеціальних підходів [12].

Українська бухгалтерська термінологія має свою специфіку, включаючи усталені терміни, професійний жаргон, численні абревіатури (наприклад, ПДВ, ЄДРПОУ, ІПН, МФО, ТМЦ) та особливості відмінювання числових значень прописом. Модель повинна бути здатною правильно розуміти та інтерпретувати ці мовні конструкції. Для ефективної роботи GiT-компонента VGT з українським текстом необхідні якісні україномовні векторні представлення слів (word embeddings) або речень (sentence embeddings). Існують певні розробки в цьому напрямку, наприклад, українські моделі на основі sentence-transformers [13], але їхня придатність та ефективність у складі VGT для бухгалтерського домену потребує дослідження. Створення бенчмарків для мультимодального розуміння українською мовою, таких як ZNO-Vision [14], є важливим кроком, але вони поки що не охоплюють специфіку бухгалтерських документів.

Іншим серйозним викликом являється адаптація до предметної області (Domain Adaptation). Навіть за наявності якісних загальних моделей для DLA та обробки української мови, необхідна їхня адаптація до специфіки саме бухгалтерських документів.

Моделі, навчені на шаблонах наукових статей (що використовуються в PubLayNet або DocBank) або на різноманітних веб-документах (D^4LA), можуть погано узагальнюватися на специфічні структури, макети та візуальні характеристики українських бухгалтерських документів. Дослідження показують, що поточні підходи до DLA, часто зосереджені на академічних працях, демонструють нижчу продуктивність на інших типах документів, таких як фінансові звіти. Робота "Source-Free Document Layout Analysis (SFDLA)" [9] особливо підкреслює проблеми адаптації моделей DLA до таких доменів, як фінансові документи та медичні записи, через чутливість даних, обмеженість ресурсів для анотування та значні відмінності в структурі. Пряме перенесення моделей, навчених на одному домені (наприклад, наукові статті), на інший (наприклад, фінансові звіти) часто призводить до значного падіння продуктивності, яке може сягати в середньому -32.64% [9].

Як зазначалося раніше, навіть у межах одного типу бухгалтерського документа (наприклад, рахунок-фактура) можуть існувати численні варіації шаблонів, що використовуються різними підприємствами. Модель повинна бути стійкою до таких варіацій і не "перенавчатися" на специфічні особливості лише одного чи кількох шаблонів. Ця варіативність вимагає від VGT не просто запам'ятовування шаблонів, а глибокого розуміння семантики та логічної структури документа. Це, в свою чергу, підсилює важливість ефективних стратегій попереднього навчання (таких як MGLM та SLM, що використовуються в VGT [2]) та, можливо, застосування технік аугментації даних, які можуть генерувати реалістичні варіації бухгалтерських документів для розширення навчальної вибірки.

У ситуаціях, коли вихідні дані, на яких попередньо навчалася базова модель

VGT, є недоступними (наприклад, якщо використовується готова попередньо навчена модель від стороннього розробника), а є лише сама модель, виникає потреба в методах адаптації без доступу до вихідних даних (Source-Free Domain Adaptation). Такі підходи, описані, наприклад, в роботі SFDLA [9], дозволяють адаптувати модель до цільового домену, використовуючи лише непроанотовані дані з цього домену.

Висновки. У статті проаналізовано потенціал використання моделі Vision Grid Transformer (VGT) для завдань аналізу україномовної бухгалтерської документації. Показано, що двопотокова мультимодальна структура VGT, а саме поєднання візуального (ViT) та текстово-просторового (GiT) аналізу, дозволяє досягати високої точності у розпізнаванні структурних елементів документів. Разом із методами попереднього навчання (MGLM, SLM), ця модель здатна узагальнювати структури навіть на варіативних шаблонах. Водночас виявлено низку викликів, пов'язаних з адаптацією моделі до української мови: обмежена кількість анотованих даних, залежність від якості OCR/HTR, специфіка кириличного письма та бухгалтерської термінології. Подолання цих бар'єрів вимагає створення спеціалізованих україномовних датасетів, оптимізації розпізнавання тексту та застосування методів адаптації без доступу до даних. З урахуванням цих факторів VGT має потенціал стати ефективним інструментом для цифровізації бухгалтерських процесів в Україні.

Література

1. Shehzadi, T., Stricker, D., & Afzal, M. Z. (2024). A Hybrid Approach for Document Layout Analysis in Document images. [ArXiv preprint] arXiv:2404.17888. https://doi.org/10.48550/arXiv.2404.17888
2. Da, C., Luo, C., Zheng, Q., & Yao, C. (2023). *Vision Grid Transformer for Document Layout Analysis* [ArXiv preprint]. arXiv. https://doi.org/10.48550/arXiv.2308.14978
3. Li, J., Xu, Y., Lv, T., Cui, L., Zhang, C., & Wei, F. (2022). *DiT: Self-supervised pre-training for Document Image Transformer* [ArXiv preprint]. arXiv. https://doi.org/10.48550/arXiv.2203.02378
4. Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., & Zhou, M. (2019). *LayoutLM: Pre-training of text and layout for document image understanding* [ArXiv preprint]. arXiv. https://doi.org/10.48550/arXiv.1912.13318
5. Shen, Z., Zhang, R., Dell, M., Lee, B. C. G., Carlson, J., & Li, W. (2021). *LayoutParser: A unified toolkit for deep learning-based document image analysis* [ArXiv preprint]. arXiv. https://doi.org/10.48550/arXiv.2103.15348
6. Zhong, X., Tang, J., & Yepes, A. J. (2019). *PubLayNet: Largest dataset ever for document layout analysis* [ArXiv preprint]. arXiv. https://doi.org/10.48550/arXiv.1908.07836
7. Li, M., Xu, Y., Cui, L., Huang, S., Wei, F., Li, Z., & Zhou, M. (2020, December). *DocBank: A Benchmark Dataset for Document Layout Analysis*. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)* (pp. 949–960). Barcelona, Spain (Online): International Committee on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.82
8. Tikhonov, A., & Rabus, A. (2024). *Handwritten Text Recognition of Ukrainian Manuscripts in the 21st Century: Possibilities, Challenges, and the Future of the First Generic AI-based Model*. *Kyiv-Mohyla Humanities Journal*, 11, 226–247. https://doi.org/10.18523/2313-4895.11.2024.226-247
9. Gruber, I., Picek, L., Hlaváč, M., Neduchal, P., Hrúz, M. (2024). Improving Handwritten Cyrillic OCR by Font-Based Synthetic Text Generator. In: Moosaei, H., Hladík, M., Pardalos, P.M. (eds) Dynamics of Information Systems. DIS 2023. Lecture Notes in Computer Science, vol 14321. Springer, Cham. https://doi.org/10.1007/978-3-031-50320-7_8
10. Weber, M., Siebenschuh, C., Butler, R. M., Alexandrov, A., Thanner, V. R., Tsolakis, G., Jabbar, H., Foster, I., Li, B., Stevens, R., & Zhang, C. (2023). WordScape: A pipeline to extract multilingual, visually rich documents with layout annotations from web crawl data [Paper presentation]. In Advances in Neural Information Processing Systems, 36 (Datasets and Benchmarks Track). NeurIPS. https://doi.org/10.48550/arXiv.2312.10188Tewes, S., Chen, Y., Moured, O., Zhang, J., & Stiefelhagen, R. (2025). *SFDLA: Source-Free Document Layout Analysis* [ArXiv preprint]. arXiv.

https://doi.org/10.48550/arXiv.2503.18742
11. FutureBeeAI. (2025). Ukrainian Printed OCR Image Datasets [Data set]. Retrieved from https://www.futurebeeai.com/dataset/printed-ocr-image-data-sets/ukrainian
12. Lupașcu, M., Rogoz, A.-C., Stupariu, M. S., & Ionescu, R. T. (2025, February 8). Large Multimodal Models for Low-Resource Languages: A Survey [ArXiv preprint]. arXiv. https://doi.org/10.48550/arXiv.2502.05568
13. lang-uk. (2024, March 23). ukr-paraphrase-multilingual-mpnet-base [Sentence-Transformers model]. Hugging Face. Retrieved from https://huggingface.co/lang-uk/ukr-paraphrase-multilingual-mpnet-base
14. Paniv, Y., Kiulian, A., Chaplynskyi, D., Khandoga, M., Polishko, A., Bas, T., & Gabrielli, G. (2024, November 22). Benchmarking Multimodal Models for Ukrainian Language Understanding Across Academic and Cultural Domains [ArXiv preprint]. arXiv. https://doi.org/10.48550/arXiv.2411.14647

Analysis of the possibilities application of the Vision Grid Transformer model for document structure analysis of ukrainian accounting documents

Maksym Korostil, Ilona Lahun

In the context of ongoing digital transformation, the need for automation of accounting document processing is steadily increasing, particularly in Ukraine, where a significant portion of primary documentation still exists in paper form or as scanned images. Effective information extraction from such documents requires the application of advanced artificial intelligence techniques, especially deep learning and multimodal data analysis. This paper explores the applicability of the Vision Grid Transformer (VGT) architecture for analyzing the layout structure of Ukrainian accounting documents. The VGT model combines two information streams – a visual stream (based on the Vision Transformer, ViT) and a text-spatial stream (based on the Grid Transformer, GiT), enabling comprehensive document representation both in terms of visual layout and semantic content. The model's flexibility is further enhanced by pretraining methods such as Masked Grid Language Modeling (MGLM) and Segment Language Modeling (SLM), which help capture both local and global contextual dependencies among textual elements. The study emphasizes the specific challenges of adapting VGT to the Ukrainian language and accounting domain. These include the lack of high-quality publicly annotated Ukrainian datasets, the necessity of accurate optical character recognition (OCR) for Cyrillic scripts, issues related to handwritten text recognition (HTR), and the complexity of domain-specific terminology and abbreviations used in accounting.

Keywords: document layout analysis, deep learning, optical character recognition, handwritten text recognition, accounting automation, Ukrainian-language documents.

Отримано 08.07.2025