ВИКОРИСТАННЯ CNN ДЛЯ БАГАТОКЛАСОВОЇ КЛАСИФІКАЦІЇ ТА КЛАСИФІКАЦІЇ З БАГАТЬМА МІТКАМИ ЗОБРАЖЕНЬ ЇЖІ

Мінухін С.В., Шапошник М.В.

Кафедра інформаційних систем,
Харківський національний економічний університет ім. С.Кузнеця, Ук
Україна

E-mail: serhii.minukhin@hneu.net
maxym.shaposhnyk@hneu.net

Abstract

The article analyzes the problem of the "information bottleneck" in hybrid modular generative architectures for the image-to-recipe generation task. The analysis focuses on the theoretical flaws of a "naïve" modular approach, which, due to the high intra-class variability of the food domain, fails to provide the generative component with sufficient details, leading to the generation of template-based, non-relevant recipes. To solve this problem, an advanced architecture is proposed that implements the "Explicit Semantic Supervision" method. This approach is based on combining multiclass classification (CNN-1, identifying the dish class) and multilabel classification (CNN-2, identifying ingredients), implemented via parallel "heads" with different loss functions (CCE and BCE). The analysis confirms that this dual-component approach, which acts as an ensemble method, provides the generative component with visually-grounded details (class + ingredients) and allows for the generation of relevant descriptions, thus overcoming the bottleneck.

Довільна форма презентації страв вимагає відходу від жорсткої класифікації до детального розпізнавання їх елементів (інгредієнтів), що враховувало б варіативність у структурі певного рецепту [1].

Розпізнавання їжі є значно складнішою задачею, ніж просте анотування зображень. Це зумовлено істотною внутрішньокласовою мінливістю (багато варіацій однієї страви), а також значними деформаціями, що відбуваються під час її приготування. Приготовані страви часто роблять свої інгредієнти незрозумілими та невиразними [2].

Це вимагає від моделі розрізнення "тонких відмінностей" (fine-grained differences), наприклад, таких як типи інгредієнтів або способи нарізки. Сучасні VLM, такі як CLIP [3], хоч і є потужними, проте часто не здатні відрізнити таких критично важливих деталей у харчовому домені [4].

Наївний підхід — наприклад, просте використання CNN для класифікації страви теоретично некоректним. Така архітектура створює критичне "інформаційне вузьке місце" (information bottleneck). CNN надає одну мітку загального класу (наприклад, "піца"), що не дає достатньої інформації про розпізнаний об’єкт на зображенні.

Метою даної роботи є розробка вдосконаленої модульної архітектури, яка долає цей "інформаційний дефіцит". Пропонується впровадження двоетапної класифікації, що складається з двох ключових компонентів: використання першої CNN (DenseNet121 [5]) для визначення загального класу страви та одночасне використання другої CNN для ідентифікації конкретних інгредієнтів страви, видимих на зображенні (класифікація з декількома мітками). Об’єднання цих двох потоків даних (клас + інгредієнти) при формуванні промпту для LLM у подальшому дозволить генерувати опис рецептів приготування, що є значно більш обґрунтованим у візуальних деталях.

Для вирішення проблеми "інформаційного вузького місця" пропонується вдосконалена конвеєрна архітектура. Цей конвеєр складається з трьох ключових етапів, які обробляють візуальну інформацію на різних рівнях деталізації, перш ніж надати її до великої мовної моделі (LLM).

Визначення Класу: вхідне зображення страви обробляється першою згортковою нейронною мережею (CNN-1) DenseNet121. Цей компонент вирішує задачу багатокласової класифікації (multiclass classification), де кожне зображення зіставляється з однією, найбільш ймовірною, загальною категорією страви (напр., "салат", "піца", "паста").

Визначення Інгредієнтів: паралельно або послідовно те саме зображення обробляється другою CNN (CNN-2) DenseNet121. Цей компонент вирішує більш складну задачу - багатозначної класифікації (multilabel classification). Його мета — ідентифікувати набір (a set) усіх візуально присутніх інгредієнтів (наприклад, ["помідор", "сир", "пепероні", "тісто"]).

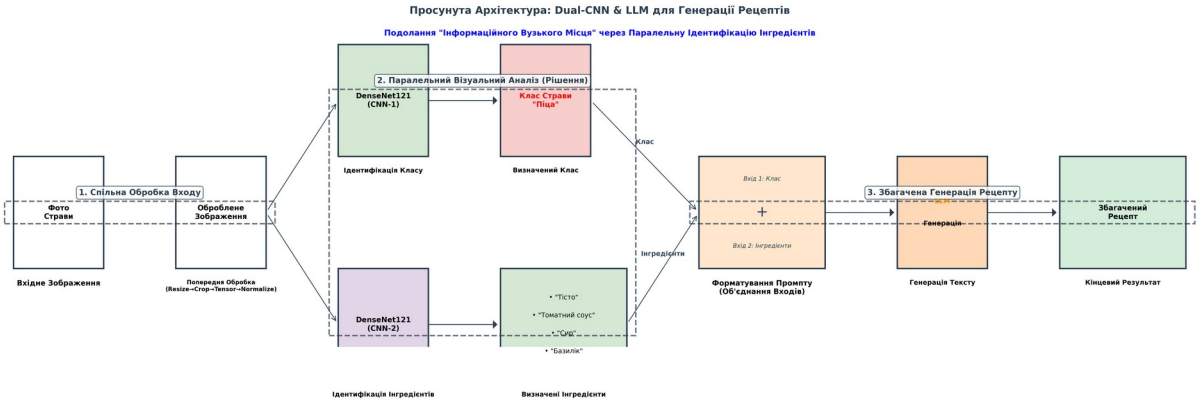

Рис. 1. Загальна схема запропонованої архітектури

У контексті даної роботи, багатокласова класифікація — це процес, при якому вхідному зображенню (фото страви) необхідно присвоїти одну мітку класу з фіксованого набору можливих категорій страв. Для експериментів використовуються стандартні набори даних, такі як Food-101 [6] і Food-256 [7], що містять 101 та 256 унікальних класів страв відповідно. Саме ця велика кількість класів визначає високу трудомісткість задачі багатокласової класифікації. Специфіка цих наборів даних ускладнюється високою внутрішньокласовою мінливістю (багато варіацій однієї страви) та, часто, низькою міжкласовою мінливістю (різні страви виглядають дуже схоже). Цей етап є критично важливим, оскільки він надає LLM початковий, загальний контекст ("Це паста").

Для вирішення цієї задачі було обрано архітектуру DenseNet121. Цей вибір обґрунтований двома наступними ключовими перевагами.

1. Ефективність Архітектури: DenseNet (Densely Connected Convolutional Networks) безпосередньо вирішує проблему "згасаючого градієнта" (vanishing gradient) шляхом з'єднання кожного шару з кожним іншим шаром. Це забезпечує стабільне навчання навіть у дуже глибоких мережах, що є критичним для складних задач розпізнавання.

2. Навчання з Передачею Знань (Transfer Learning): Модель DenseNet121 використовується з вагами, попередньо навченими на масивному наборі даних ImageNet. Це дозволяє моделі використовувати вже сформовані "знання" про базові візуальні ознаки (краї, текстури, форми), що значно прискорює процес навчання та підвищує кінцеву точність на специфічному домені (їжа).

Оскільки це задача багатокласової класифікації, стандартним та найбільш ефективним методом щодо оцінки продуктивості є мінімізація Категорійної Перехресної Ентропії (Categorical Cross-Entropy). Ця функція втрат вимірює "відстань" між двома розподілами ймовірностей: прогнозований розподіл: Вектор ймовірностей, який генерує модель (наприклад, [0.85, 0.1, 0.05] для класів ["піца", "паста", "салат"]); істинний розподіл: "One-hot" вектор, де правильний клас має 1, а всі інші — 0 (наприклад, [1.0, 0.0, 0.0]).

Мінімізуючи цю функцію втрат, змушуємо модель генерувати розподіл, максимально близький до істинного, тобто бути "впевненою" у правильному класі.

Визначення Категорійної Перехресної Ентропії використовує такий вираз:

$$L_{class} = -\sum_{i=1}^{C} y_i \log(\hat{y}_i),$$

(1)

де L_{class} – втрати для визначення класу страви;

C – загальна кількість класів;

y_i – індикатор того, чи є клас *i* правильним.

Для еталонної мітки y_i дорівнює 1 - для правильного класу, і 0 - для всіх інших (one-hot encoding).

$\hat{y}_i$ -ймовірність, яку модель спрогнозувала для класу *i* (вихід із шару SoftMax).

Якщо CNN-1 вирішує задачу багатокласової класифікації (одне зображення — одна мітка), то CNN-2 вирішує задачу класифікації з багатьма мітками (multilabel classification). Це означає, що кожне вхідне зображення може (і повинно) бути пов'язане з декількома мітками (інгредієнтами) одночасно з великого словника можливих інгредієнтів.

Саме цей компонент долає "інформаційне вузьке місце". Замість однієї загальної мітки ("піца"), він надає LLM детальний візуальний контекст ("тісто", "томатний соус", "сир", "базилік"), що дозволяє розрізнити "Маргариту" від "Пепероні".

Для цього завдання використана згорткова нейронна мережа зі схожою архітектурою "основи" (shared backbone), наприклад, той же DenseNet, що й у CNN-1, для ефективного вилучення ознак. Однак, "голова" (head) цієї моделі кардинально відрізняється:

1. Замість одного виходу LogSoftmax на C класів, вона має N виходів (де N — загальна кількість інгредієнтів у словнику).

2. Замість Softmax (який змушує суму ймовірностей дорівнювати 1), на кожному з N виходів застосовується функція активації Sigmoid.

Оскільки Softmax тут не застосовується і кожен інгредієнт є незалежним (наявність "помідора" не виключає наявності "сиру"), не можливо використовувати Категорійну Перехресну Ентропію.

Стандартним рішенням для такої задачі є Бінарна Перехресна Ентропія (Binary Cross-Entropy, BCE), яка застосовується до кожного виходу (інгредієнта) незалежно. Отже розглядаємо цю задачу не як одну складну, а як N простих бінарних класифікацій ("Чи є на фото помідор? Так/Ні", "Чи є на фото сир? Так/Ні", і т.д.). Функція втрат L_{ingr} обчислює середнє значення BCE по всім N можливим інгредієнтам.

Отже, остаточна формула розрахунку втрат з використанням бінарної CE [8] має вигляд:

$$L_{Multi-BCE} = -\frac{1}{N}\sum_{j=1}^{S}\sum_{i=1}^{N}\left[y_{j,i}\log\left(\hat{y}_{j,i}\right) + \left(1 - y_{j,i}\right)\log\left(1 - \hat{y}_{j,i}\right)\right], \quad (2)$$

N – загальна кількість зображень;

S – загальна кількість інгредієнтів;

$y_{j,i}$ - імовірність того чи присутній інгредієнт *i* на зображенні (1 – так, 0 – ні);

$\hat{y}_{j,i}$ - ймовірність, яку модель спрогнозувала для присутності інгредієнта *i*.

Ідея одночасного розпізнавання класу страви та її інгредієнтів, що є задачею Multi-Task Learning (MTL), досліджена в наукових статтях. Існуючі роботи згадають «багатозадачне глибоке навчання…для розпізнаваня інгредієнтів та їжі» [9] або аналізують навчання у «одно-або багатозадачному режимі» [10].

Більшість сучасних робіт покладаються на механізми уваги [11]. Це моделі, які вчаться самостійно «дивитися» на правильні, дрібнозернисті частини зображення. Модель сама здогадується, що шматочки «пепероні» є важливими, але не має семантичного розуміння [12].

Інший підхід – це використання CNN зі складними функціями втрат, такими як Triplet Loss [13], Center Loss або SCloss (Subclass Center Loss) [14]. Ці функції змушують моделі групувати схожі варіації страви в один щільний кластер у просторі ознак.

Наукова новизна запропонованої двокомпонентної архітектури (CNN-1 для класу + CNN-2 для інгредієнтів) полягає у використанні методу "Явної Семантичної Супервізії" (Explicit Semantic

Supervision) для вирішення проблеми "візуального інформаційного вузького місця" та високої внутрішньокласової мінливості , що є специфікою домену їжі. На відміну від домінуючих SOTA-парадигм, які покладаються або на "неявну увагу" (Implicit Attention) (де модель самостійно вчиться фокусуватися на релевантних пікселях), або на "абстрактне метричне навчання" (Metric Learning) (де складні функції втрат, як Triplet Loss чи Center Loss, математично групують схожі зображення, не розуміючи їхньої суті), — запропонований підхід забезпечує семантичне та інтерпретоване розрізнення. Модель "розуміє", чому "Маргарита" та "Пепероні" є різними, оскільки компонент CNN-2 надає явний, керований людиною сигнал, ідентифікуючи їхні унікальні дрібнозернисті атрибути (напр., ['базилік'] проти ['пепероні']).

Технічно ця "явна супервізія" реалізується шляхом впровадження паралельної "голови" (head) для CNN-2, яка архітектурно відрізняється від CNN-1. Замість одного виходу Softmax для багатокласової класифікації, CNN-2 має N незалежних виходів, де N — загальна кількість інгредієнтів. На кожному з цих виходів застосовується функція активації Sigmoid , а вся "голова" оптимізується окремою функцією втрат — Бінарною Перехресною Ентропією (BCE) L_{ingr} — вирішуючи задачу класифікації з багатьма мітками. Саме ця паралельна оптимізація двох різних функцій втрат L_{class} та L_{ingr} з двома різними "головами" і є тим механізмом, що долає "інформаційне вузьке місце".

У даній роботі проаналізовано ключову проблему "інформаційного вузького місця" (information bottleneck) у модульних «image-to-recipe» архітектурах. Було продемонстровано, що "наївний" підхід, який поєднує одну CNN (для визначення загального класу) та LLM (для її опису), є теоретично неможливим, оскільки він не надає генеративній моделі достатньо «fine-grained» візуальних деталей, що й призводить до нерелевантних результатів.

Для вирішення цієї проблеми було запропоновано вдосконалену модульну архітектуру, що базується на багатоетапному візуальному аналізі. Запропонований конвеєр включає два паралельні компоненти:

CNN-1 (DenseNet121) - для вирішення задачі багатокласової класифікації (визначення загального класу страви);

CNN-2 (DenseNet121) - для вирішення задачі класифікації з багатьма мітками (ідентифікація конкретних інгредієнтів).

Фактично, запропонований підхід можна розглядати як реалізацію ансамблевого методу, де поєднання виходів двох спеціалізованих візуальних моделей (класифікатора класів та класифікатора інгредієнтів) забезпечує більш надійну та деталізовану основу для подальшої генерації тексту в кулінарному домені.

Об’єднання цих двох потоків інформації (клас та інгредієнти) у структурований промпт дозволить LLM генерувати візуально обґрунтовані та релевантні рецепти, долаючи обмеження "наївного" підходу.

Література

1. Salvador A., Hynes N., Aytar Y., Marin J., Ofli F., Weber I., Torralba A. Learning Cross-modal Embeddings for Cooking Recipes and Food Images // Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2017. P. 3068–3076.

2. Salvador A., Drozdzal M., Giro-i-Nieto X., Romero A. Inverse Cooking: Recipe Generation from Food Images // Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2019. P. 10445–10454.

3. Radford A., Kim J. W., Hallacy C., Ramesh A., Goh G., Agarwal S., Sastry G., Askell A., Mishkin P., Clark J., Kruege G., Sutskever I. Learning Transferable Visual Models From Natural Language Supervision // Proceedings of the 38th International Conference on Machine Learning (ICML). 2021. Vol. 139. P. 8748–8763.

4. Wahed M., Zhou X., Yu T., Lourentzou I.. Fine-Grained Alignment for Cross-Modal Recipe Retrieval // Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). 2024. P. 5584–5593.

5. Huang G., Liu Z., Maaten L., Weinberger K. Q. “Densely Connected Convolutional Networks // Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017. P. 2261–2269.

6. Bossard L., Guillaumin M., Van Gool L. Food-101 – Mining Discriminative Components with Random Forests // Computer Vision – ECCV 2014. Lecture Notes in Computer Science. 2014. Vol. 8694. P. 446–461.

7. Kawano Y., Yanai K. Automatic Expansion of a Food Image Dataset Leveraging Existing Categories with Domain Adaptation // Computer Vision - ECCV 2014 Workshops. ECCV 2014. Lecture Notes in Computer Science. Springer, Cham. 2015. Vol. 8927.https://doi.org/10.1007/978-3-319-16199-0_1

8. Katona T., Tóth G., Petró M., Harangi B. Advanced Multi-Label Image Classification Techniques Using Ensemble Methods // Machine Learning and Knowledge Extraction. 2024. Vol. 6(2). P. 1281-1297.

9. Weiqing M., Shuqiang J., Shuhui W., Jitao S., Shuhuan M. A Delicious Recipe Analysis Framework for Exploring Multi-Modal Recipes with Various Attributes // Proceedings of the 25th ACM international conference on Multimedia (MM '17). Association for Computing Machinery, New York, NY, USA. 2017. P. 402–410.

10. Chen J., Zhu B., Ngo C. W., Chua T. S., Jiang Y. G. A Study of Multi-Task and Region-Wise Deep Learning for Food Ingredient Recognition // IEEE transactions on image processing : a publication of the IEEE Signal Processing Society. 2021. Vol. 30. P. 1514–1526.

11. Fukui H., Hirakawa T., Yamashita T., Fujiyoshi H. Attention Branch Network: Learning of Attention Mechanism for Visual Explanation. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2018. P. 10697-10706.

12. Chen Z., Wang J., Wang Y. Enhancing Food Image Recognition by Multi-Level Fusion and the Attention Mechanism // Foods (Basel, Switzerland). 2025. Vol.14(3). P. 461.

13. Kim S., Seo M., Laptev I., Cho M., Kwak S. Deep Metric Learning Beyond Binary Supervision // 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA. 2019. P. 2283-2292.

14. Metwalli A-S., Wei S., Chase W. Food Image Recognition Based on Densely Connected Convolutional Neural Networks. 2020. P. 27-32.