УДК 004.932:004.023

Чабан О.Р., Манзюк Е.А.

Хмельницький національний університет

МЕТОД ДИСТИЛЯЦІЇ ЗНАНЬ ВІД МОДЕЛЕЙ-ВЧИТЕЛІВ ДО МОДЕЛІ-УЧНЯ ГЛИБОКОГО НАВЧАННЯ

У роботі запропоновано метод дистиляції знань від моделей-вчителів до моделі-учня глибокого навчання, який призначений для покращення класифікації медичних зображень, зокрема МРТ серця. Метод враховує проблему доменного зсуву та обмежену кількість анотованих даних, комбінуючи адаптивну дистиляцію знань, доменну адаптацію та підходи до збереження приватності. Метод складається з трьох блоків: навчання моделей-вчителів із доменною адаптацією, агрегування знань для створення узагальнених ознак та навчання моделі-учня із застосуванням псевдоанотацій та збереженням конфіденційності. Обчислювальні експерименти показали, що запропонований метод злегка перевершує сучасні підходи за точністю та узагальненням знань за різними доменами даних.

In this work, we propose a knowledge distillation method from teacher models to a student model to improve the classification of cardiac magnetic resonance images. The method addresses the domain shift problem and the limited amount of annotated data by combining adaptive knowledge distillation, domain adaptation, and privacy-preserving approaches. The method consists of three blocks: training teacher models with domain adaptation, aggregating knowledge to create generalized features, and training the student model using pseudo-annotations while preserving privacy. Computational experiments demonstrated that the proposed method slightly outperforms state-of-the-art approaches regarding accuracy and knowledge generalization across different data domains.

Запропонований метод дистиляції знань використовує кілька моделей-вчителів для передання знань до моделі-учня, щоби покращити її здатність до класифікації медичних зображень. Метою роботи є підвищення точності класифікації, здатності узагальнювати знання за різних доменів даних та використання великих обсягів неанотованих даних для навчання моделей глибокого навчання. Метод передбачає кілька важливих етапів, як от навчання моделей-вчителів із доменною адаптацією, адаптивну дистиляцію знань та збереження приватності даних пацієнтів.

На рисунку 1 наведено структуру методу, що складається з трьох основних блоків.

Блок 1: Навчання моделей-вчителів із доменною адаптацією – вчителі навчаються за різними наборами даних із доменними відмінностями, використовуючи комбіновану втрату, яка враховує класифікацію та адаптацію до домену.

Блок 2: Адаптивна дистиляція знань – зібрані знання накопичуються від різних вчителів для формування єдиного вектора-ознак, що передається моделі-учню.

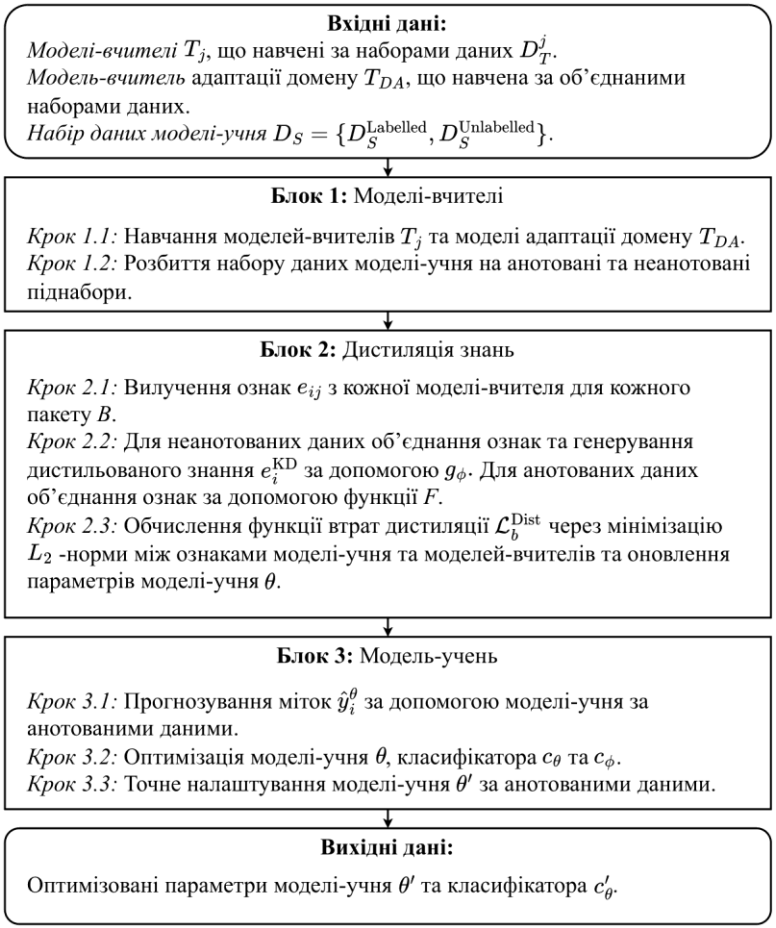

Рисунок 2.2 – Покрокова схема запропонованого методу дистиляції знань від моделей-вчителів до моделі-учня глибокого навчання

Блок 3: Навчання моделі-учня зі збереженням приватності – модель-учень використовує агреговані знання для навчання, з можливістю застосування псевдоанотацій і збереження приватності.

Нижче наведемо детальний опис кожного блоку.

Блок 1 передбачає навчання кількох моделей-вчителів за різними наборами даних, кожен із яких відображає певний домен. Домени відрізняються за джерелами даних, протоколами створення медичних зображень та умовами, що зумовлює проблему доменного зсуву. Для успішного навчання моделі використовуються комбіновані втрати, які включають класифікаційну втрату та втрату адаптації:

$$L_{\text{teacher}} = L_{\text{class}} + \lambda L_{\text{domain}}, \tag{1}$$

де L_{class} – класифікаційна втрата, L_{domain} – доменна втрата, що сприяє вивченню домен-інваріантних ознак, а λ – гіперпараметр.

В *Блоці 2* методу під назвою «Адаптивна дистиляція знань» використовуються агреговані вектори ознак, які отримуються від моделей-вчителів. Вплив кожної моделі-вчителя визначається за допомогою адаптивних ваг:

$$w_t = \frac{\exp\left(\frac{\alpha P_t\left(y|x\right)}{T}\right)}{\sum_{k=1}^{N}\exp\left(\frac{\alpha P_k\left(y|x\right)}{T}\right)}, \tag{2}$$

де T – параметр, що контролює розгладжування розподілу ваг.

Після обчислення адаптивних ваг об'єднані ознаки використовуються для навчання моделі-учня.

Блок 3 відповідає за навчання моделі-учня зі збереженням приватності на основі отриманих агрегованих ознак. У цьому блоці використовуються псевдоанотації для великої кількості даних у такий спосіб:

$$\hat{y}_i = \arg\max\left\{P_{\text{student}}\left(y|x_i\right)\right\} \quad \text{if} \quad \max\left\{P_{\text{student}}\left(y|x_i\right)\right\} > \tau, \tag{3}$$

де τ – поріг для псевдоанотацій.

Навчання відбувається зі збереженням приватності за допомогою додавання гауссівського шуму до градієнтів:

$$\tilde{g} = g + \mathcal{N}\left(0, \sigma^2\right). \tag{4}$$

Запропонований метод був перевірений за наборами даних зображень магнітно-резонансної томографії (МРТ) серця, як от Automated Cardiac Diagnosis Challenge (ACDC) [1] та Multi-Center Multi-Vendor and Multi-Disease Cardiac Image Dataset (M&Ms-2) [2]. Зображення МРТ із цих наборів суттєво відрізняються між собою за умовами знімання, що створює значний доменний зсув. У процесі проведення експериментальних тестувань, методами для порівняння були такі підходи, як одна модель-вчитель (STM) [3], а також сучасні методи напівкерованого навчання (SSL) [4] і дистиляції знань з адаптацією доменів (KDDA) [5].

Результати показали, що запропонований метод досяг вищої точності як за початковим доменом (ACDC), так і за цільовим доменом (M&Ms-2), що вказує на його здатність добре узагальнювати знання. Таблиця 1 демонструє порівняння рівня результативності за цільовим доменом (M&Ms-2).

Таблиця 1 – Результати та порівняння за метриками класифікації за цільовим доменом M&Ms-2 (у відсотках, %)

Модель	Accuracy	Precision	Recall	F_1-score	AUC-ROC
STM [3]	71.0	69.5	70.0	69.7	76.0
SSL [4]	80.5	79.2	79.8	79.5	85.0
KDDA [5]	84.0	83.0	83.5	83.2	88.0
Запропонований метод	**85.5**	**83.4**	**84.0**	**84.7**	**88.5**

З таблиці 1 бачимо, що запропонований метод злегка перевершив інші підходи. Наприклад, точність запропонованого методу становила 85.5 %, що вище, ніж у базового підходу STM (71.0 %) і навіть у сучасних методів SSL (80.5 %) та KDDA (84.0 %).

У підсумку, основні переваги запропонованого методу дистиляції знань полягають у такому:

– адаптивне зважування – динамічне коригування ваг дало змогу інтегрувати знання від різних моделей-вчителів до одного учня, підвищивши точність класифікації зображень МРТ у порівнянні з аналогами;

– адаптація доменів – вилучення доменно-інваріантних ознак сприяло подоланню зсуву між різними наборами даних;

– збереження приватності – додавання гауссівського шуму до градієнтів забезпечило збереження конфіденційних даних без суттєвих втрат точності класифікації.

Перелік посилань

1. Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved? / O. Bernard et al. IEEE Transactions on Medical Imaging. 2018. Vol. 37, no. 11. P. 2514–2525. URL: https://doi.org/10.1109/tmi.2018.2837502
2. Deep learning segmentation of the right ventricle in cardiac MRI: The M&ms Challenge / C. Martín-Isla et al. IEEE Journal of Biomedical and Health Informatics. 2023. P. 1–14. URL: https://doi.org/10.1109/jbhi.2023.3267857
3. Data efficient unsupervised domain adaptation for cross-modality image segmentation / C. Ouyang et al. Lecture Notes in Computer Science. Cham, 2019. P. 669–677. URL: https://doi.org/10.1007/978-3-030-32245-8_74
4. ACPL: Anti-curriculum pseudo-labelling for semi-supervised medical image classification / F. Liu et al. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) : Proceedings, New Orleans, LA, USA, 18–24 June 2022. New York, NY, USA, 2022. P. 20697–20706. URL: https://doi.org/10.1109/cvpr52688.2022.02004
5. Multiple teachers-meticulous student: a domain adaptive meta-knowledge distillation model for medical image classification / S. Nabavi. arXiv: arXiv:2403.11226. P. 1–26. URL: https://doi.org/10.48550/arXiv.2403.11226