

ОГЛЯД ПІДХОДІВ ДО ОПТИМІЗАЦІЇ ВИБОРУ НАВЧАЛЬНИХ ДАНИХ У ПРОЦЕСІ САМОНАВЧАННЯ МОДЕЛЕЙ ІДЕНТИФІКАЦІЇ ОБ'ЄКТІВ

У задачах ідентифікації об'єктів глибокі нейронні мережі демонструють високу ефективність, проте їх самонавчання є обчислювально затратним процесом, що ускладнює їх оперативну адаптацію до нових даних. Одним із ключових чинників, що визначає швидкість і ефективність самонавчання, є стратегія відбору навчальних прикладів. Традиційні підходи, зокрема випадковий або рівномірний вибір даних, не забезпечують оптимального використання обчислювальних ресурсів. Водночас у контексті реальних застосувань систем ідентифікації об'єктів критично важливо забезпечити ефективне навчання моделі на обмеженій вибірці, фокусуючись на найбільш інформативних зразках.

Одним із способів оптимізації вибору навчальних даних є метод активного навчання. Даний підхід дозволяє зменшити необхідний обсяг анотованих даних шляхом вибору найбільш інформативних навчальних прикладів. Його основна концепція полягає в наданні моделі можливості самостійно визначати зразки з великого пулу неанотованих даних, які, згідно з її оцінкою, матимуть найбільший вплив на покращення її продуктивності. Такий підхід сприяє зниженню вимог до обсягу навчальної вибірки, скороченню витрат на процес маркування даних та підвищенню ефективності навчання моделі [1,2].

Є декілька підходів для реалізації методу активного навчання оптимізації вибору даних. Одним із найбільш поширених підходів є **метод невизначеного вибору**, який ґрунтується на принципі відбору зразків, для яких модель демонструє найменшу впевненість у прогнозах. Основна ідея цього підходу полягає в тому, що навчання на найбільш невизначених прикладах дозволяє суттєво покращити узагальнювальні властивості моделі, скорочуючи кількість необхідних анотованих даних.

Є декілька способів реалізації методу невизначеного вибору:

- **Entropy-based sampling** – вибір прикладів, для яких ентропія розподілу ймовірностей прогнозованих класів є максимальною.
- **Margin Sampling** – вибір прикладів, для яких різниця між ймовірностями двох найбільш ймовірних класів є мінімальною.
- **Least Confident Sampling** – вибір прикладів, для яких ймовірність найбільш ймовірного класу є мінімальною.

Метод невизначеного вибору ефективний у випадках, коли модель стикається з раніше невідомими класами або варіаціями об'єктів, що дозволяє зосередити ресурси на найкритичніших зразках [1,3].

Ще один популярний підхід - метод **комітетного запиту**. Даний підхід використовує ансамбль моделей (так званий "комітет") для оцінки невизначеності щодо нового зразка даних. Основна ідея даного методу полягає в тому, що якщо моделі в комітеті мають значні розбіжності у прогнозах щодо певного прикладу, цей зразок є інформативним і повинен бути включений у вибірку для анотації.

Основні принципи методу **комітетного запиту**:

- Формування комітету моделей.
- Оцінка невизначеності на основі розбіжностей між прогнозами.
- Вибір найбільш суперечливих прикладів для анотації.

Поєднання методу **комітетного запиту** із самонавчанням дозволяє ефективніше адаптувати моделі до нових даних, мінімізуючи витрати на анотацію.

Отже, активне навчання є потужним підходом до оптимізації процесу самонавчання моделей ідентифікації об'єктів, дозволяючи суттєво зменшити обсяг необхідних анотованих даних та прискорити процес навчання. Метод невизначеного вибору дозволяє моделі зосередитися на випадках, у яких вона демонструє найменшу впевненість, тоді як комітетний підхід використовує розбіжності в прогнозах різних моделей для виявлення суперечливих прикладів. Подальші дослідження будуть спрямовані на розробку гібридних методів, що комбінують переваги різних стратегій активного навчання для досягнення ще вищої ефективності самонавчання моделей.

Список використаних джерел:

1. Settles B. Active Learning // Synthesis Lectures on Artificial Intelligence and Machine Learning. – Morgan & Claypool, 2012. – 123 p.
2. Olsson F. A literature survey of active machine learning in the context of natural language processing. – Swedish Institute of Computer Science, 2009. – 47 p.
3. Wang K., Zhang D., Li Y. Cost-effective active learning for deep image classification // IEEE Transactions on Circuits and Systems for Video Technology, 2020. P. 1441–1452.