Технічні науки

УДК 004.853

Попеляєв Денис Павлович
студент
Національний технічний університет України
«Київський політехнічний інститут ім. Ігоря Сікорського»
Попеляев Денис Павлович
студент
Национальный технический университет Украины
«Киевский политехнический институт им. Игоря Сикорского»
Popeliaev Denis
student
National technical university of Ukraine
«Igor Sikorsky Kyiv Polytechnic Institute»

МЕТОДИ РОЗШИРЕННЯ ОБ'ЄМУ НАВЧАЛЬНИХ ДАНИХ ДЛЯ ШТУЧНИХ НЕЙРОННИХ МЕРЕЖ
МЕТОДЫ РАСШИРЕНИЯ ОБЪЕМА УЧЕБНЫХ ДАННЫХ ДЛЯ ИСКУССТВЕННЫХ НЕЙРОННЫХ СЕТЕЙ
METHODS OF TRAINING DATA AUGMENTATION FOR ARTIFICIAL NEURAL NETWORKS

Анотація. В даній роботі описані проблеми, що притаманні навчальним множинам для штучних нейронних мереж та тестування підходу до їх вирішення – штучного додавання даних у навчальну множину (data augmentation).

Ключові слова: Штучні нейронні мережі, навчальна множина, штучне додавання даних, незбалансовані навчальні множини, перенавчання.

Аннотация. В данной работе описаны проблемы, присущие учебным множествам для искусственных нейронных сетей и тестирование

подхода к их решению - искусственного добавления данных в учебное множество (data augmentation).

Ключевые слова: Искусственные нейронные сети, обучающее множество, искусственное добавление данных, несбалансированные учебные множества, переобучение.

Summary. This paper describes the problems inherent in the training datasets for artificial neural networks and approach testing to solving these problems - adding artificial training data (data augmentation).

Keywords: Artificial neural networks, dataset, data augmentation, imbalanced datasets, overfiting.

1. Вступ

За останні роки машинне навчання не тільки набрало популярності, а й сформувалося як окрема дисципліна з досліджень розпізнавання образів та теорії обчислювального навчання в галузі штучного інтелекту. Репрезентативні вхідні дані високої якості є ключем до якісних моделей машинного навчання, а їх дефіцит може перешкоджати розвитку моделі. Формування множини навчальних даних має принципово важливе значення для успішного вирішення завдань машинного навчання. Часто завдання машинного навчання зводяться саме до правильного формування навчальної множини. Помилки у формуванні навчальної множини зазвичай виявляються критичними і здатні звести нанівець ефективність самих алгоритмів навчання. Серед фахівців по машинному навчанню загальновизнаним вважається, що наявність якісних навчальних даних часто набагато важливіше якості алгоритму навчання. У зв'язку з активним розвитком глибоких нейронних мереж в останнє десятиліття питання формування множини навчальних даних приймає особливо важливе значення, оскільки в багатьох задачах глибокі нейронні мережі демонструють якість, що істотно перевершує інші алгоритми, однак, щоб

отримати подібний виграш в якості, необхідно використовувати навчальну множину дуже великого розміру (до декількох мільйонів зображень, при цьому навчання вимагає великого обсягу обчислювальних ресурсів і може займати кілька тижнів на багатопроцесорному кластері), а також спеціальні методи розширення та імітації розширення навчальної множини, які будуть розглянуті далі. У той же час, в сучасній літературі по машинному навчанню, питанню формування навчальної множини приділяється недостатня увага, недостатньо розвинена теоретична база, що пояснює явища, що виникають в процесі формування множини навчальних даних.

2. Можливі проблеми при формуванні навчальної множини

Для оцінки якості навчальної множини зазвичай використовується її обсяг (кількість навчальних прикладів). Однак дана метрика не особливо інформативна. По-перше, даних може бути дуже багато, але всі вони - однакові, по-друге, навіть якщо всі об'єкти - різні, деякі області простору ознак можуть залишитися незаповненими (так звані "sparse areas"), і, по-третє, в самій процедурі формування навчальної множини можуть бути закладені помилки. Розглянемо деякі можливі проблеми і помилки при формуванні навчальної множини.

2.1. Фонові закономірності

У завданнях машинного навчання об'єкт може бути заданий набором значень ознак і значеннями цільових змінних. Завдання машинного навчання - знайти закономірності між значеннями спостережуваних ознак і цільових змінних. При цьому на основі кожного конкретного навчального об'єкта, не беручи до уваги інші об'єкти, будь-яку залежність, характерну для даного об'єкта, можна рахувати завжди істинною. При розгляді великої кількості різноманітних об'єктів з усіх можливих закономірностей характерними залишиться лише невелика кількість дійсно значимих закономірностей. Зауважимо, що на підставі малого числа даних немає

ніякого способу відрізнити правильну закономірність від помилкової. Фахівцями машинного навчання було виведено емпіричне правило, згідно якого нейронна мережа завжди в першу чергу буде навчатися відрізняти найяскравіші закономірності. Помилкові закономірності, що виникають в результаті браку даних, називають фоновими закономірностями. По суті, деякі види перенавчання (overfitting) полягають в завчанні фонових закономірностей. Приклад фонової закономірності - залежність між класом зображення і кольором одного конкретного пікселя [1].

2.2. Відсутність навчальних об'єктів певного виду

Найпростіший приклад помилки при формуванні навчальної множини - якщо в ній відсутні дані певного виду (не покрита деяка область простору об'єктів), алгоритм не зможе правильно навчитися їх класифікувати. Логічно було б додати сюди і недостатню кількість навчальних об'єктів певного виду, однак в різних випадках достатнім є різне число об'єктів.

2.3. Відсутність даних певного виду щодо признакової системи

Признакова система породжує деяке розбиття множини даних на кейси, кожному кейсу відповідає деякий вузький набір значень ознак, при цьому кейсів тим більше, чим більш різноманітні і складні ознаки. Якщо деякий з даних варіантів не буде покритий об'єктами з навчальної множини або імовірнісний розподіл всередині варіанта буде невірно відображати властивості того що моделюється, навчання може виявитися некоректним. При ускладненні признакової системи підвищуються вимоги до навчальної множини.

2.4. Деякі з генеруючих змінних не варіюються

Важливий окремий випадок проблеми відсутності даних певного виду. Дуже часто при формуванні навчальної множини частина генеруючих змінних має завжди одні й ті ж значення або дуже вузький діапазон значень.

2.5. Розбалансування

Нерозумне з семантичної точки зору порушення співвідношень кількості даних різного виду в використовуваній множині даних, що приводить до необґрунтованого завищення впливу на результат одних і заниження впливу або повного ігнорування інших даних, і, як наслідок, до прийняття неоптимальних рішень. Найпростіший приклад - через особливості формування навчальної множини, прикладів одного класу або типу набагато більше, ніж прикладів іншого класу або типу. Різні алгоритми навчання мають різну стійкість до подібних проблем. Розбалансування особливо критичне при використанні дерев рішень. Зауважимо, що розбалансування - досить загальний клас явищ, які можуть виникати не тільки в процесі формування навчальної множини. Розбалансування також можливе в задачах чисельної оптимізації і при тестуванні алгоритмів розпізнавання. Часто проблема розбалансування вирішується за допомогою різних видів нормалізації.

3. Способи додавання даних в навчальну множину

Додавання даних є одним з найпростіших і ефективних способів поліпшити якість навчальної множини та вирішення описаних проблем. При цьому просте додавання даних довільного виду не завжди ефективно, часто потрібно додати дані певного різновиду для підвищення якості розпізнавання. Поширеним підходом є програмна генерація. У разі використання синтетичних навчальних даних зручніше всього згенерувати відсутні навчальні приклади. Однак не у всіх завданнях допустимо використання програмно згенерованих даних. У таких випадках доводиться застосовувати складніші методи додавання даних. Більш ефективним підходом є так зване "збільшення даних" (data augmentation). Модифікація наявних зображень з метою розширити навчальної множини. Активно застосовується при навчанні глибоких нейронних мереж, а також в умовах дефіциту розмічених даних. Застосовуються стиснення,

розтягування, горизонтальне відображення, поворот, випадковий зсув в колірному просторі, випадкова або закономірна зміна деяких пікселів, обрізання частини зображення. Вважається, що додавання повністю випадкового шуму неефективно, слід додавати шум, обумовлений даними (який потенційно можливий в реальних даних). Цей метод дозволяє гнучко доповнювати простір навчальних даних саме тими значеннями, які є дефіцитними, тобто заповнювати "пробіли" ("sparse areas").

Оскільки глибинні мережі повинні бути навчені на величезній кількості тренувальних даних для досягнення задовільних результатів, якщо початковий набір зображень досить обмежений, то є доцільним застосувати методи збільшення даних для підвищення продуктивності. Збільшення даних стає звичним і навіть необхідним етапом роботи при підготовці сучасної глибокої мережі.

Є багато способів виконати збільшення даних, найпопулярніші з яких вже були згадані вище. Крім того, ефективним є поєднання декількох різних обробок, наприклад, роблячи обертання і випадкове масштабування одночасно. Крім того, існують більш складні техніки, наприклад можна підняти насиченість і значення (компоненти S і V колірного простору HSV) всіх пікселів зображення в интервалі між 0,25 і 4, помножити ці значення на коефіцієнт від 0,7 і 1.4, і додати до них значення від -0,1 до 0,1. Крім того, можна додати значення між -0.1, 0.1 до компоненти відтінку (H компонента HSV) всіх пікселів в зображенні / контурі [2].

A. Krizhevsky і ін. запропонували техніку "fancy PCA" при підготовці знаменитого Alex-Net в 2012 році. Fancy PCA змінює інтенсивності RGB каналів в тренувальних зразках. На практиці, по-перше виконується PCA на множині піксельних значень RGB над множиною тренувальних образів. І потім, для кожного тренувального образу, додається деяка величина до кожного пікселя зображення RGB, яка є випадковою величиною взятою з гауссового розподілу з нульовим середнім і стандартним відхиленням 0,1.

Важливо, що кожне відхилення використовується тільки один раз для всіх пікселів конкретного тренувального зображення та до тих пір, поки зображення не буде використовуватися для навчання знову. Тобто, коли модель отримуватиме те ж навчальне зображення знову, для нього буде випадковим чином згенероване інше відхилення для збільшення даних. При використанні цієї техніки на конкурсі ImageNet 2012 року, було отримано зниження помилки "топ-1" більш ніж на 1% [3].

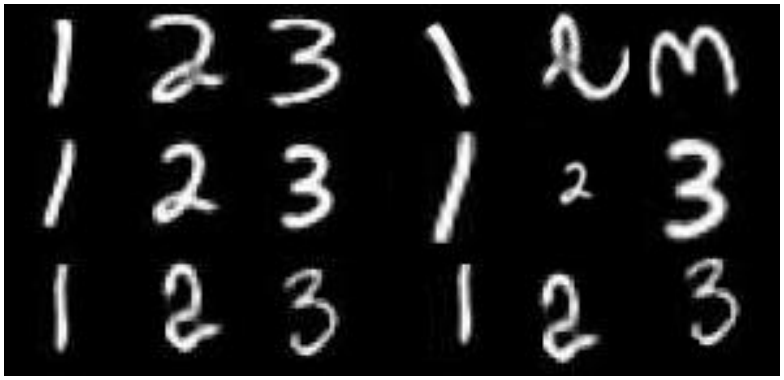
4. Результати підходу data augmentation

Для тестування підходу була взята CNN (convolutional neural network) з класичною архітектурою та публічній набір данних, що складався з зразків рукописних цифр [4]. Вхідний набір був розбитий на навчальну та тестову вибірку. При навчанні на такому наборі, без внесення ніяких змін, було досягнуто посимвольної помилки в 1.63%. Як бачимо, конволюційні нейронні мережі вже є достатньо потужним інструметом для рішення подібних задач. Наступним етапом є додання до початкового набору (до кожної з двох вибірок) зображень, що мали наступні зміни:

- *нахил* (на випадкову величину від 0 до 360 градусів з кроком в 30 градусів)
- *масштабування* (на випадкову величину від 0.5 до 1.5 з кроком 0.1)
- *зрушення* (на випадкову величину від -5 до 5 пікселів вертикально та горизонтально з кроком в 1 піксель)

На рисунку 1 співставленні оригінальні зображення та змінені нахилом, масштабуванням та зрушенням відповідно з першого по третій рядок.

Рисунок 1. складено автором на основі [4]



Результати навчання з використанням цих модифікацій показані в табл.1.

Таблиця 1. [розробка автора]

Тип вибірки	Посимвольна помилка
Оригінальна вибірка	1.63%
Збільшення вибірки за рахунок зображень з нахилом	1.27%
Збільшення вибірки за рахунок зображень з нахилом та масштабуванням	1.12%
Збільшення вибірки за рахунок зображень з нахилом, масштабуванням та зрушенням	1.08%

Комбінування декількох модифікацій дає збільшення ефективності класифікації мережею зображень з тестової вибірки за рахунок віддалення межі перенавчання (overfitting). Для отримання кращих результатів необхідно переглянути архітектуру мережі.

5. Висновок

Були розглянуті підходи до збільшення розміру навчальних даних для машинного навчання та перевірена їх ефективність для усунення основних проблем при формуванні навчальної множини. Варто відзначити, що в кожному конкретному випадку необхідно вибирати свій алгоритм отримання навчальних даних. В деяких випадках можна використовувати навчальну вибірку, створену тільки з реальних даних. Але якщо реальних

даних не дуже багато, або якщо в навчальній множині відсутні дані певного виду (не покрита деяка область простору об'єктів), оптимальний підхід - data augmentation.

Література:

1. Kaftannikov I.L., Parasich A.V. Problems of Training Set's Formation in Machine Learning Tasks. Bulletin of the South Ural State University. Ser. Computer Technologies, Automatic Control, Radio Electronics, 2016, vol. 16, no. 3, pp. 15–24. DOI: 10.14529/ctcr160302
2. K. He, X. Zhang, S. Ren, and J. Sun. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification, ICCV, 2015.
3. A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In NIPS, 2012
4. Yann LeCun, Corinna Cortes, Christopher J.C. Burges. The MNIST database of handwritten digits, http://yann.lecun.com/exdb/mnist/