МЕТОДЫ АВТОМАТИЗАЦИИ РАЗМЕТКИ ДАННЫХ ДЛЯ КОМПЬЮТЕРНОГО МОДЕЛИРОВАНИЯ

И. А. Батраева, Д. С. Пантелеев,

Саратовский национальный исследовательский государственный университет им. Н. Г. Чернышевского, Россия
E-mail: tlanaufp@yandex.ru, pantelev00@gmail.com

В данной статье будут описаны существующие и потенциальные методы и подходы к автоматизации процесса разметки данных для дальнейшего моделирования, их достоинства и недостатки. В современном мире компьютерное моделирование всевозможных процессов стало, если не само собой разумеющимся, то очевидным явлением. Каждое моделирование нуждается в данных, и эти данные должны быть категоризированы и размечены в соответствии с темой и предметной областью моделирования. Поэтому задача автоматизации процесса разметки моделируемых данных является актуальной, так как ручная разметка может занимать во много раз больше времени, чем само моделирование.

DATA MARKUP AUTOMATION METHODS FOR COMPUTER MODELING

I. A. Batraeva, D. S. Panteleev

This paper will describe existing and potential methods and approaches to automating the data markup process for further modeling, their advantages and disadvantages. In the modern world, computer simulation of all kinds of processes has become, if not a matter of course, then an obvious phenomenon. Every modeling need data, and that data must be categorized and label according to the theme and domain of the simulation. Therefore, the task of automating the process of labeling modeling dataset is relevant, since manual labeling can take many times longer than the simulation itself.

В настоящее время, основным способом решения задач по моделированию экономических и других процессов является компьютерное моделирование, которое осуществляется на основе анализа больших объемов данных. При работе с большими данными можно выделить две основные проблемы – автоматизация сбора данных – эта проблема в принципе уже решена, и проблема разметки имеющихся данных, которая до сих пор во многих случаях осуществляется вручную.

Разметка данных является жизненно важным этапом предварительной обработки данных для дальнейшего моделирования. Каждая ошибка или неточность в этом процессе может негативно сказаться на «качестве» датасета. Более того, общая производительность прогностической модели может быть нарушена, что приведет к неправильному толкованию результатов моделирования. Таким образом, одной из основных задач является построение такого алгоритма, который мог бы размечать дискретные данные.

Существует три варианта разметки данных [2]: ручная, полуавтоматическая и полностью автоматическая разметка. На данный момент основным под-

ходом к разметке данных является ручной подход. Когда человек или группа людей вручную выделяют анализируемые объекты в «сырых» данных, либо помечают их, относя к определенной категории. И если объем данных достаточно большой приходится либо отдавать это на аутсорсинг, либо собирать команду людей и размечать данные самостоятельно.

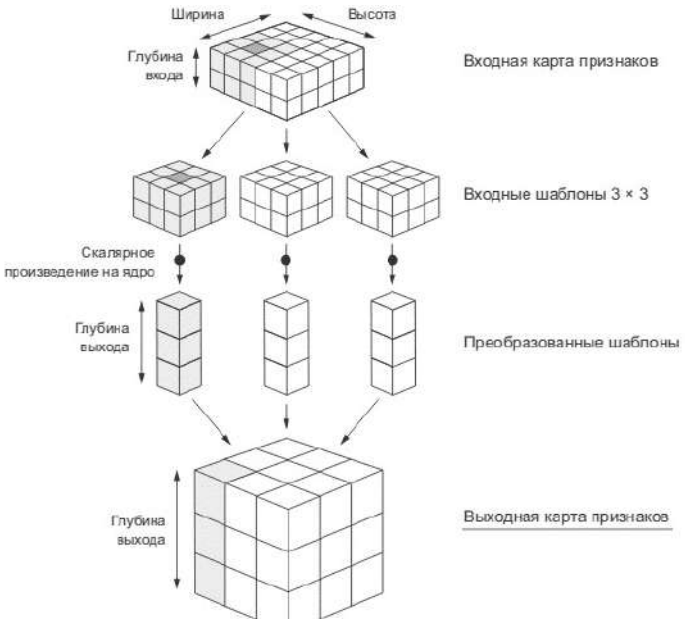
Данный подход имеет свои плюсы, например, в случае привлечения к разметке людей из соответствующей данным предметной области, можно ожидать довольно высокий уровень точности, как разметки, так и дальнейшего моделирования. Основным минусом такого подхода, очевидно, являются трудозатраты, как материальные, так и временные.

Если говорить о полуавтоматической разметке данных, то мы сталкиваемся, хоть и не так остро, но с теми же проблемами, что и при ручной разметке, это трудозатраты. Основным «бутылочным горлышком» полуавтоматической разметки является человек, так что полуавтоматическая разметка не может похвастаться скоростью по сравнению с ручной.

Самым же перспективным вариантом разметки является автоматический. Можно выделить два основных метода автоматизации разметки на сегодняшний день [3]:

1. Первый метод заключается в обучении компьютерной сверточной нейронной сети [4] с подкреплением на небольшой части данных для дальнейшей обработки оставшихся данных этой нейронной сетью.

При этом, чем больше данных обработала сеть, тем больше данных она использовала для собственного дообучения, так называемое «подкрепление».



Архитектура сверточной нейронной сети

В конечном итоге, мы получаем нейронную сеть, которая по итогу разметки всех данных имеет более высокую точность разметки, нежели чем была в начале, поэтому нейросеть можно запустить повторно, уже без подкрепления

на полном объеме данных для повторной, более точной разметки. Такой способ был в частности реализован в Саратовском университете для решения задачи разметки текстовых корпусов для Саратовского диалектного корпуса [5]. В работе [6] исследовалось обучение нейронной сети для анализа текстов с точки зрения определения их жанровой принадлежности. Обычно эта работа выполняется вручную, поэтому обработка больших корпусов текстов является чрезвычайно длительным процессом.

Такие методы дают достаточно точные результаты, но требуют больших ресурсов для своей реализации.

2. Второй метод разметки можно охарактеризовать как наиболее консервативный. Данные размечаются согласно некоторой функции, которая, основываясь на определенных признаках производит разметку. Так, например, при исследовании [7] влияния глобального потепления на мировую экономику в период всеобщего локдауна ученые пришли к выводу, что замедление мировой экономики и соответствующее уменьшение выбросов CO_2 не сказалось на замедлило процесс глобального потепления. Для этого ученые проанализировали изменение ледовой шапки на северном полюсе по спутниковым снимкам и его корреляцию с темпами мировой экономики, и, в частности, количеством выбросов CO_2. Для разметки данных спутниковых снимков ученые создали и использовали алгоритм, основным признаком разметки которого являлся поиск белового цвета на изображении. Такой метод имеет следующие недостатки: отсутствие универсальности и точность. Если с последним все понятно, то первое подразумевает, что при необходимости разметить другой набор данных, придется писать новую функцию, так как старая создана исключительно для конкретных данных и к новым данным не подойдет.

Возможным решением проблемы стал бы метод, объединяющий в себе достоинства обоих вышеописанных. Предлагается размечать данные случайным образом с помощью соответствующего скрипта или функции, после чего выполнять проверку данных на точность разметки. Если алгоритм счел часть данных размеченными с достаточной точностью, он откладывает их, для остальных данный цикл повторяется, и так, пока точность разметки всех данных не будет удовлетворительной. Основной задачей при реализации такого метода является создание необходимого и достаточного алгоритма верификации точность полученной разметки. От этого алгоритма зависит конечная точность разметки данных данным способом.

СПИСОК ЛИТЕРАТУРЫ

1. *Меркулова Ю. В.* О сущности экономического моделирования // Международный журнал экспериментального образования. 2015. № 9. С. 149-151.

2. Как работает разметка данных. [Электронный ресурс]. URL: https://data.korusconsulting.ru/press-center/blog/kak-rabotaet-razmetka-dannykh/ (дата обращения: 16.10.2022).

3. *Namatevs I., Sudars K., Polaka I.* Automatic data labeling by neural networks for the counting of objects in videos // Procedia Computer Science. 2019. № 149. С. 151-158.

4. A Comprehensive Guide to Convolutional Neural Networks. [Электронный ресурс]. URL: https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53 (дата обращения: 16.10.2022).

5. *Батраева И. А., Гольдин В. Е., Крючкова О. Ю.* Поисковый механизм саратовского диалектного корпуса // Компьютерные науки и информационные технологии. Материалы Междун. науч. конф. 2009. С. 24-27.

6. Использование анализа семантической близости слов при решении задачи определения жанровой принадлежности текстов методами глубокого обучения // Вестник Томского государственного университета. Управление, вычислительная техника и информатика. 2020. № 50. С. 14-22.

7. Климат и коронавирус: война повесток? // Современная Европа. 2020. № 7. С. 87-100.