РАССМОТРЕНИЕ СПОСОБОВ ФОРМИРОВАНИЯ НАБОРОВ ДАННЫХ ДЛЯ ОБУЧЕНИЯ НЕЙРОННЫХ СЕТЕЙ

Окунев С.В. Email: Okunev680@scientifictext.ru

Окунев Сергей Витальевич – студент,
кафедра информатики и вычислительной техники,
Сибирский государственный университет науки и технологий им. академика М.Ф. Решетнева,
г. Красноярск

Аннотация: *в данной статье рассмотрена тема формирования больших наборов данных для обучения нейронных сетей, описаны основные термины и понятия. Представлены основные способы формирования и преумножения наборов данных, в частности с помощью аугментации, а также генеративно-состязательной нейронной сети. Описаны способы формирования обучающих выборок в зависимости от класса поставленной задачи. Приведены принципы и схема работы сетей GAN, описан алгоритм работы генератора и дискриминатора, рассмотрены трудности их применения и возникающие сложности при обучении.*

Ключевые слова: *нейронные сети, наборы данных, обучающая выборка, аугментация, генеративно-состязательная нейронная сеть, генератор, дискриминатор.*

CONSIDERATION OF METHODS FOR FORMING DATA SETS FOR TRAINING NEURAL NETWORKS

Okunev S.V.

Okunev Sergey Vitalievich – Student,
DEPARTMENT OF INFORMATICS AND COMPUTER ENGINEERING,
SIBERIAN STATE UNIVERSITY OF SCIENCE AND TECHNOLOGY ACADEMICIAN M.F. RESHETNEV,
KRASNOYARSK

Abstract: *this article discusses the topic of forming large data sets for training neural networks, describes the basic terms and concepts. The main methods for the formation and augmentation of data sets are presented, in particular with the help of augmentation, as well as a generative-competitive neural network. The methods of forming data sets depending on the class of the job are described. The principles and the scheme of operation of GAN networks are described, the algorithm of the generator and discriminator operation is described, the difficulties of their application and the difficulties encountered in learning are considered.*

Keywords: *neural networks, data sets, training set, augmentation, generative-competitive neural network, generator, discriminator.*

УДК 004.85

В настоящее время нейронные сети стали очень популярны в нашей жизни, так как они позволяют решать сложные задачи и улучшать уже существующие решения. Особую важность в процессе работы с нейронными сетями представляет этап их обучения. Именно от него зависит получаемая точность работы в решении поставленной задачи. Для точного и качественного обучения необходимо иметь подготовленный набор данных. Однако зачастую возникает проблема, связанная с трудностью поиска наборов данных, особенно для частных задач.

Поиск необходимого набора данных не всегда приводит к успеху и тогда приходится самостоятельного собирать данные и формировать из них обучающие выборки. А также существует проблема при нехватке данных, особенно это актуально для глубоких нейронных сетей, так как они очень требовательны к большим объемам данных для сходимости обучения. Таким образом, рассмотрение способов формирования наборов данных является актуальной задачей.

Процесс подготовки набора данных представляет собой важный этап, в результате которого получается обработанный набор очищенных данных, пригодный для обработки алгоритмами машинного обучения. Такой набор называется *dataset*, необходимый для обучения модели нейронной сети и использования ее в конкретных задачах.

Dataset для обучения нейронной сети – это обработанная и структурированная информация в табличном виде. Строки такой таблицы называются объектами, а столбцы – признаками. Различают 2 вида признаков [1]:

– независимые переменные – предикторы;

– зависимые переменные – целевые признаки, которые вычисляются на основе одного или нескольких предикторов.

Признаки характерны для задач классификации, так как имеется конечное множество объектов, принадлежащих определенному классу.

Первичный набор исходных данных принято называть генеральной совокупностью. Выборка – это конечное подмножество элементов генеральной совокупности, изучив которое можно понять поведение исходного множества. Например, генеральная совокупность состоит из 150 тысяч посетителей сайта, а в выборку попали 250 из них [2, с. 45].

Вероятностная модель порождения данных предполагает, что выборка из генеральной совокупности формируется случайным образом. Простая выборка является математической моделью серии независимых опытов и, как правило, используется для машинного обучения. При этом для каждого этапа обучения необходим свой набор данных:

– обучающая выборка;

– тестовая выборка;

– проверочная (валидационная) выборка.

Способы формирования обучающих и оценочных выборок зависят от класса задачи, решаемой с помощью машинного обучения:

– для задач классификации данные следует разделять так, чтобы в полученных наборах численное соотношение объектов разных классов было таким же, как в исходной генеральной совокупности;

– для задач регрессионного анализа необходимо одинаковое распределение целевой переменной в полученных наборах, которые будут использоваться для обучения и контроля качества.

В настоящее время в открытом доступе представлено большое множество информации, которую можно легко использовать для создания наборов данных. Однако данный процесс усложняется при наличии каких-либо дополнительных требований предъявляемых к данным или при недостаточном их количестве. Например, чтобы обучить нейронную сеть с миллионом параметров, нужно очень много обучающих примеров, которые не всегда легко найти.

Можно выделить несколько случаев:

– нужного набора данных нет в открытом доступе;

– имеется набор данных, но недостаточного размера.

При отсутствии готового к использованию набора данных ситуация немного усложняется и может затянуться во времени. Процесс будет состоять в поиске и структуризации открытых данных. Информацию можно брать с веб-платформ, предоставляющих статистику, со сторонних сайтов с помощью парсинга, а также можно попробовать найти похожий набор данных и попытаться переформировать под собственные требования. Для данной цели нет универсальных решений и необходимо основываться на конкретном случае. Можно прибегнуть к помощи дополнительных программ. Как правило, реализуется простейший алгоритм, который обрабатывает данные и собирает из них *dataset*. Существует готовые библиотеки для создания собственных наборов данных, например, *ArcGIS Pro* [3, с. 180].

Для увеличения набора данных можно использовать методы преумножения, особенно актуально это для цифровых наборов. Изображения можно искажать, переворачивать, менять тон. С помощью такого способа можно значительно приумножить выборку изображений [4].

А также существует еще один эффективный способ преумножения изображений – это применение генеративно-состязательной нейронной сети (*GAN*), представляющей собой архитектуру, которая состоит из генератора и дискриминатора. Архитектура данной сети состоит из двух разных сетей. Одна нейронная сеть – генератор, создает случайные новые экземпляры данных, а другая — дискриминатор, оценивает их на подлинность. То есть дискриминатор принимает решение, относится ли экземпляр данных к набору тренировочных данных или нет. Так же существует разновидность данной архитектуры, называемой *DGAN* (*Deep Convolutional Generative Adverserial Networks*) – сверточные генеративно-состязательные сети. Эта модель заменяет сверточными слоями полностью соединённые слои генеративной состязательной сети. Данную сеть можно эффективно применить при уже имеющемся наборе данных.

Генератор создает новые изображения, которые он передает на оценку дискриминатору. Цель генератора состоит в том, чтобы генерировать такие данные, которые будут приняты дискриминатором. Цель дискриминатора – определить, является ли изображение подлинным (Рисунок 1).

При этом генератор не имеет представления о том, что представляют собой исходные данные и обучается на основе ответов дискриминатора, с каждой итерацией изменяя результаты своей работы. За основу генератор берет вектор случайного шума и на основании его генерирует данные.

При использовании сети *GAN* для генерации изображений существует определенная трудность с обучением данной сети. Существует ряд правил, которых стоит придерживаться, например, при обучении дискриминатора необходимо удерживать значения генератора постоянными и наоборот. То есть каждая сеть должна тренироваться против статичного «противника».

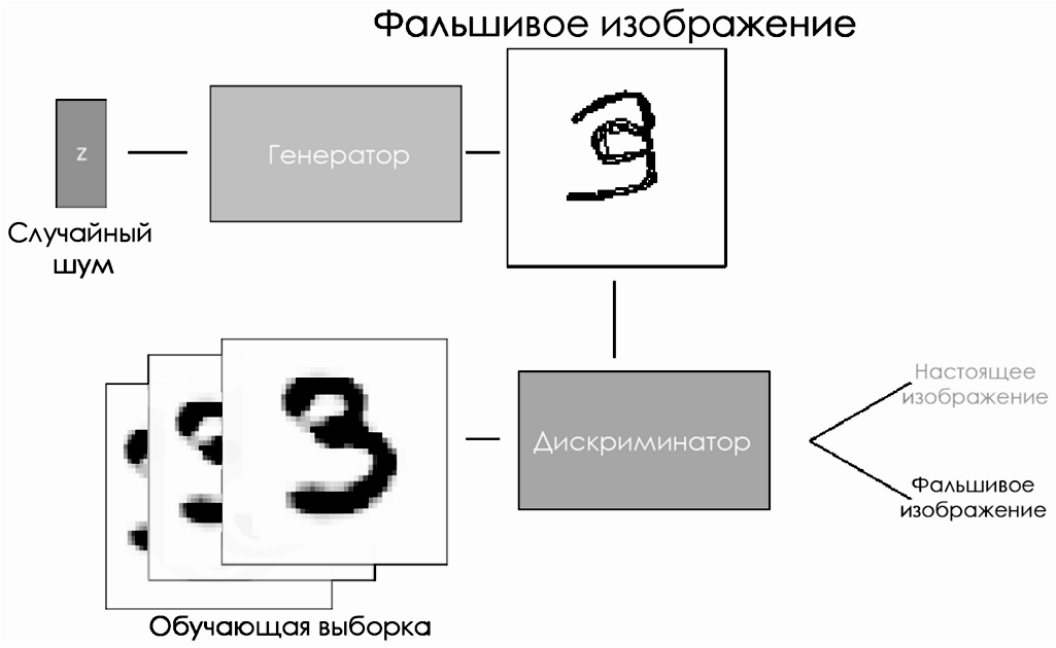

Рис. 1. Принцип работы GAN сети

Так же может возникнуть ситуация при неравномерном обучении, например, когда дискриминатор слишком хорошо обучен, то он будет возвращать значения очень близкие к 0 или 1 и генератор будет испытывать трудности при чтении вектора градиента. Если же генератор хорошо обучен, то он будет использовать неточности дискриминатора, которые будут приводить к ложному срабатыванию [5, с. 387].

Еще одной проблемой может стать длительность обучения. Необходимо иметь большие вычислительные мощности. Обучение на одном процессоре может занять целый день.

Для задачи пополнения набора данных можно применить обученные нейронные сети, результат работы которых может быть использован в качестве входных данных для обучаемой нейронной сети. Например, существуют сети для построения карт значимости в виде тепловой карты. Результаты работы таких сетей можно применить в сетях по отслеживанию объектов на видеопоследовательности или в сетях, предназначенных для сегментации изображений.

Хорошо подготовленный набор данных является очень важной составляющей качественного процесса обучения. В настоящее время найти нужную информацию не составит труда, однако возникает трудность при ее обработке в большом количестве. Процесс создания готового набора данных проходит в несколько этапов в зависимости от сложившейся ситуации. Только после длительного процесса сбора и структурирования информации ее можно применить в машинном обучении. Особую требовательность к большим наборам данных имеют глубокие сети, нуждающиеся в длительном процессе обучения.

Таким образом, изучение способов формирования больших наборов для обучения является актуальной задачей и в настоящее время требует больших затрат времени. Для автоматизации данного процесса можно применять различные методы обработки информации, в том числе и нейронные сети.

Список литературы / References

1. Обучение нейронной сети. [Электронный ресурс]. Режим доступа: https://www.bigdataschool.ru/bigdata/dataset-data-preparation.html/ (дата обращения: 23.12.2019).
2. *Medioni Gerard*. Sing Bing Kang Emerging Topics in Computer Vision. Издательство Prentice Hall Ptr, 2004. 45 с.
3. *Латыпова Р.* Нейронные сети [Текст]. М.: LAP Lambert Academic Publishing, 2012. 180 с.
4. *Radhakrishna A.* Frequency-tuned Salient Detection [Электронный ресурс]. Режим доступа: http://infoscience.epfl.ch/ (дата обращения: 05.05.2019).
5. *Гарсия Глория Буэно*. Обработка изображений с помощью OpenCV Г М.: ДМК Пресс, 2015. 387 с.