УТЕЧКА ДАННЫХ ПРИ ОБУЧЕНИИ НЕЙРОННЫХ СЕТЕЙ НА ПРИМЕРЕ АНАЛИЗА МЕДИЦИНСКИХ ИЗОБРАЖЕНИЙ

Е.О. Шубкин, аспирант гр. А0-39
Томский политехнический университет
E-mail: eos5@tpu.ru

Введение

Выявление патологий и диагностирование заболеваний путем анализа медицинских изображений является одной из наиболее актуальных задач последних лет. В настоящее время самым популярным решением этой задачи является машинное обучение. Особенно большой скачок в этом направлении произошел в 2020 году, когда для диагностирования COVID-19 было предложено использовать компьютерное зрение вместе со сверточными нейронными сетями.

В настоящее время обучающие выборки становятся все больше и вместе с ними растет и точность моделей. Но при практическом применении полученных моделей мы сталкиваемся с такими проблемами как переобучение, утечка данных, нахождение моделями обходных путей, неверная логика рассуждений, а также проблема неконкретизации при постановке задач.

Целью данной работы является обзор одной из проблем современного машинного обучения – утечки данных, на примере анализа медицинских изображений.

Проблемы современного машинного обучения

Во многом в перечисленных выше проблемах кроется неспособность моделей машинного обучения корректно работать на более разнообразных примерах чем те что встречались им при обучении. Из-за ограниченного разнообразия данных (не имеет ничего общего с маленьким объемом обучающей выборки) модели ищут обходные пути для получения правильного результата, и происходит переобучение или утечка данных. Это приводит к хорошим показателям модели при обучении и абсолютно несоответствующим им практическим и тестовым результатам.

В настоящее время для повышения точности обучения используют увеличение размеров моделей – путем усложнения архитектуры и добавления большего количества слоев. Добавление слоев, а точнее dropout слоя отлично подходит для решения проблемы переобучения модели.

Переобучение модели (overfitting) – это использование во время обучения моделью признаков, с помощью которых модель может получать высокий результат на обучающей выборке и только на ней. Таким образом результаты на тестовой выборке (из того же распределения) будут значительно хуже.

Второй, наиболее часто используемый для повышения точности модели, способ – это увеличение объема обучающих данных. Да, увеличение размеров обучающей выборки (а вместе с этим и увеличение разнообразия данных) является самым простым способом, но подходит не во всех случаях. Один из таких случаев – утечка данных, о которой будет речь далее.

Утечка данных (data leakage) – это использование во время обучения моделью информации, которая не будет доступна в последующем применении на практике. Также к утечке данных можно отнести и shortcut learning [1]. Это явление при котором модель руководствуется правилами которые показывают хорошие результаты на обучающей выборке (и на тестовых данных, так как зачастую они берутся из одного распределения), но не подходят для более сложных условий тестирования (то есть для практического применения). Рассмотрим эту проблему подробнее на примере анализа медицинских изображений.

Утечка данных

В анализе медицинских изображений, таких как рентгеновские снимки и снимки КТ, существуют некоторые особенности при подготовке данных, которые могут повлечь за собой утечку данных. Приведем некоторые из них:

1. Обрезка изображений. Частая ошибка, когда на снимках остается лишняя информация, заставляющая модель менять логику на ошибочную. Например, в исследовании рака кожи, на некоторых снимках обучающей выборки была линейка. Ее используют дерматологи, при подозрении на опухоль, чтобы точно измерить ее размер. Как правило, данная проверка

применяется только при поражениях, вызывающих беспокойство. В связи с этим модель при получении изображения с линейкой в большинстве случаев признавала опухоль злокачественной. Утечка данных в данном примере состоит в том, что единственная причина по которой модель признавала опухоль злокачественной – наличие линейки как основание для наличия рака [2].

2. Настройка изображения (распределение пикселей, яркость, контрастность). Данные настройки изображения могут быть незаметны для человеческого глаза, но классифицированы моделью как дополнительная информация. Например, при диагностике заболеваний по рентгеновским снимкам. Если при обучении выборка собрана с разных больниц, рентгеновские аппараты которых могут иметь различные настройки яркости\контрастности снимков. Таким образом модель может обучиться определять больницу по тональности снимка, а вместе с этим и наиболее вероятное для нее заболевание.
3. Объединение разрозненных наборов данных. Этот пункт зачастую может соединять в себе другие. Рассмотрим набор данных COVIDx созданный для обучения модели COVID-Net для обнаружения положительного COVID-19 [3]. Набор COVIDx объединил в себе две выборки – положительные образцы были взяты из открытой базы снимков COVID-19, а отрицательные из набора данных детской пневмонии. В данном примере есть несколько причин для утечки данных. Первая – на большинстве снимков из отрицательной выборки был ярлык R, что дало повод модели по умолчанию считать снимок с меткой отрицательным на COVID-19. Вторая – изначально различие детских снимков от выборки со снимками COVID-19.
   Аналогичная проблема возникла и в другом исследовании по обнаружению положительного COVID-19. Приняв во внимание ошибки в наборе COVIDx отрицательную выборку заменили на выборку радиологического общества Северной Америки (RSNA) по обнаружении пневмонии. Данные смешались намного лучше, но утечка данных произошла из-за ярлыков и различной обрезки. В этот раз на снимках отрицательной выборки присутствовал ярлык L, что также было интерпретировано моделью как основной признак отрицательного COVID-19. Второй причиной стала обрезка, так как на большинстве снимков RSNA не были обрезаны плечи.

Перечисленные особенности играют большую роль при составлении обучающей выборки. Так как при попытках увеличить обучающую выборку путем объединения нескольких выборок из разных источников непременно придется столкнуться с разной обрезкой, разметкой и тональностью.

Заключение

В результате проведенного исследования можно сделать вывод что модели, связанные с медицинскими изображениями и имеющие более обширную обучающую выборку (выходящую за пределы одного набора данных), ожидаемо столкнуться с проблемой утечки данных. Это подчеркивает важность подготовки и предварительной подготовки данных.

В связи с этим также осложняется процесс корректной работы моделей машинного обучения на практических примерах.

Список использованных источников

1. Geirhos, R., Jacobsen, JH., Michaelis, C. et al. Shortcut learning in deep neural networks. Nat Mach Intell 2, 665–673 (2020)
2. Esteva, A., Kuprel, B., Novoa, R. et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 542, 115–118 (2017)
3. L. Wang, A. Wong COVID-Net: A Tailored Deep Convolutional Neural Network Design for Detection of COVID-19 Cases from Chest Radiography Images (2020)