Scientist

Scientist, 2022, 22(4), 130-132.

Материалы VII итоговой научно-практической конференции НОМУИС
23-25 мая 2022 года, г. Барнаул
Алтайский государственный медицинский университет

ВЛИЯНИЕ КАЧЕСТВА ИСХОДНОГО НАБОРА ДАННЫХ ДЛЯ МАШИННОГО ОБУЧЕНИЯ НА ТОЧНОСТЬ ДИАГНОЗА

Алтайский государственный медицинский университет, г. Барнаул

Сигаева Д.В.

Научный руководитель: Логинов М.С., к.ф.-м.н.

E-mail: maxim.loginov@mail.ru

IMPACT OF TRAINING DATASET QUALITY ON DIAGNOSIS IN MACHINE LEARNING

Altai State Medical University, Barnaul

Sigaeva D.V.

Supervisor: Loginov M.S., PhD

Использование машинного обучения в медицинских информационных системах накладывает повышенные требования к качеству подготовки исходных данных. При обучении с учителем это требует корректно размеченного набора изображений с верным указанием диагнозов на снимках. Ошибки в исходных тренировочных данных могут понижать точность постановки диагноза с помощью искусственного интеллекта.

Ключевые слова: *машинное обучение, искусственный интеллект, анализ изображений, рентгенография, диагностика.*

Keywords: *machine learning, artificial intelligence, image analisys, X-ray.*

Введение

Для машинного обучения в диагностике заболеваний часто используется обучение с учителем, когда алгоритм (модель) сначала получает подготовленные или размеченные данные, а в процессе обучения модель анализирует их и

находит характерные признаки. При анализе изображений это приводит к задаче классификации. В качестве исходных данных в данной работе использовались снимки флюорографии, по каждому из них диагнозы (метки) были поставлены профессиональными врачами.

Обучение с учителем включает два этапа: использование исходного набора данных для тренировки модели, а потом проверку обученной модели на тестовых данных из другого набора размеченных данных. При этом возникает проблема подготовки корректных исходных данных: если в тренировочном наборе данных не все диагнозы поставлены верно или часть диагнозов не пропущена, то модель не сможет обучиться корректно, и вероятность постановки правильного диагноза будет ниже. **Целью** данной работы было исследование влияния исходных данных на точность постановки диагноза.

Материалы и методы

В качестве исходных данных для обучения использовались изображения из набора данных PadChest [1], которые обрабатывались с помощью пакета для анализа рентгенографических изображений TorchXRayVision [2]. Основой модели служила сверточная нейронная сеть класса DenseNet. Для обучения модели было отобрано 139 фронтальных изображений, они были разделены случайным образом на тренировочный набор (90% изображений) и тестовый набор (10%). Модель обучалась на тренировочном наборе, а далее точно диагноза проверялась на тестовом наборе.

Рассматривалось две гипотетические ситуации некорректных исходных данных: пропуск диагноза на снимке и неверно поставленный диагноз. Для этого были подготовлены еще три набора данных: с частичным удалением меток (из исходного набора случайным образом удалено 10% меток) и с частичной заменой меток 10 и 20% меток.

Результаты

Обучение на корректном наборе данных показало вероятность правильной постановки диагноза в 55-73% для трех отобранных заболеваний. Такой результат может быть обусловлен как недостаточной проработкой алгоритма модели, так и малым размером набора данных для обучения. При обучении модели на искусственно созданных некорректных наборах данных точность для

этих трех диагнозов снизилась на 4-19%. Для других диагнозов точность менялась менее значительно или не менялась вовсе. Скорее всего это обусловлено недостаточно большим размером тренировочного набора данных.

Выводы

Появление ошибок в исходных тренировочных данных для машинного обучения приводит к снижению точности, которое может достигать 19%. Это снижение скорее характеризует общую тенденцию, для получения более точных количественных данных требуется дальнейшее исследование на больших размерах тренировочного набора данных.

Список литературы:

1. A. Bustos, A. Pertusa, JM. Salinas, M. de la Iglesia. “PadChest: A large chest x-ray image dataset with multi-label annotated reports”. Medical Image Analysis, 2020 https://bimcv.cipf.es/bimcv-projects/padchest/

2. J. P. Cohen, M. Hashir, R. Brooks, H. Bertrand. "On the limits of cross-domain generalization in automated X-ray prediction. Medical Imaging with Deep Learning 2020" https://arxiv.org/abs/2002.02497

Как цитировать:
Сигаева Д.В. (2022). Влияние качества исходного набора данных для машинного обучения на точность диагноза. Материалы VII итоговой научно-практической конференции НОМУИС, 23-25 мая 2022 года, г. Барнаул, Алтайский государственный медицинский университет. Scientist, 22 (4), 130-132.