The Benchmark Lottery

Вступление

Бенчмарк - это акт выполнения компьютерной программы, набора программ или других операций для оценки относительной производительности, обычно путем проведения ряда стандартных тестов и испытаний.

Целью статьи было показать сообществу, как небольшое изменение бенчмарков или их некорректный выбор может сильно влиять на итоговый результат, что может приводить к необьективному результату и замедлению прогресса, а также как этого можно избежать.

Жизненный цикл бенчмарка

1. Начало - определение бенчмарка (явно или неявно). Недавняя тенденция заключается в том, что создается доска лидеров или веб-сайт соревнований, на котором могут быть представлены результаты, полученные от сообщества.
2. Популярность или затухание - бенчмарк набирает популярность (из-за по количества использований, цитирований и тд), либо затухает и остается неизвестным.
3. Активная фаза разработки новых алгоритмов - считается самым активным этапом, где любой прогресс или самый современный результат может вызвать большой интерес и огласку. Самый большой прогресс также, как правило, достигается здесь.
4. Зрелость или устаревание - бенчмарк достигает зрелости, когда комьюнити считает задачу решенной, либо если какого-либо значительного прогресса достичь уже не выйдет.
5. Обновление - когда бенчмарк устаревает или достигает зрелости, далеко не редкость, что автор обновляет его. Это может быть решение проблем с существующей версией, или просто улучшение задачи, сложности или других аспектов бенчмарка.

Предвзятость и зависимость от состояния

Отсутствие согласованности в том, как именно человек что-либо оценивает (например, из-за разного уровня опыта, когнитивных предубеждений или даже присущей задаче неоднозначности) может внести большую вариативность в сравнения моделей и, как следствие, уменьшить обьективность (пример с SuperGLUE, VTAB).

В любой момент времени попытка новой идеи превзойти определенный критерий зависит от информации, собранной из предыдущих материалов и публикаций. Это естественный способ добиться прогресса в решении данной проблемы. Но если смотреть на это с другой точки зрения, то видно, как создается другой вид лотереи.

Другим аспектом того, что бенчмарки отслеживают состояние, является то, что участие в общих задачах на более позднем этапе сильно отличается от момента его создания

Подтасовка результатов

Зачастую во многих задач уже есть установленные и проверенные временем бенчмарки.

Бывают и несоответствия в заявленных бенчмарках или метриках. Из-за этого сравнение моделей бывает затруднено, иногда просто не существует стандарта, так как проблема может быть новой. Часто исследователи хотят показать свою модель только с лучшей стороны, не демонстрируя неудачные эксперименты.

Пример: Рекомендательные системы

В отличие от областей NLP, для рекомендательных систем не существует устоявшихся бенчмарков, которые предоставляют канонически ранжированные списки эффективности модели. Несмотря на то, что существует знаменитый Netflix Prize8 , этот набор данных не был широко использован в академических исследованиях или для бенчмарка новых моделей. В результате во многих статьях, посвященных рекомендательным системам, наборы оценок, как правило, произвольны.

Что можно сделать?

Был создан специальный чек-лист, основные аспекты:

1. Разбиение бенчмарка на блоки - правильно поставленная проблема или задача играет главную роль при разработке нового бенчмарка.
2. Вознаграждение за простоту и легкость адаптации - различные бенчмарки и задачи, возможно, есть способы определить количественную метрику, которая будет учитываться при ранжировании моделей.
3. Руководство по использованию бенчмарк - например, точную настройку, которую бенчмарк должен использовать для оценки или как следует сообщать о результатах.
4. Лучшее сравнение - проверка на вменяемость для новых моделей и просто эффективным способом сравнения с несколькими бенчмарками.
5. Регулярное обновление набора данных - алгоритмы со временем становятся слишком адаптированными к набору данных, по сути, запоминая все его особенности и теряя способность к обобщению
6. Различные статистические оценки - средняя производительность модели при нескольких случайных параметрах, тестирование значимости, игнорировать дисперсию на отдельных наборах данных и тд

Заключение

Несмотря на то, что постоянная разработка новых бенчмарков - это, возможно, признак продолжающегося прогресса, но с другой стороны, есть опасность застрять в порочном круге инвестиций в создание неизменных бенчмарков, которые вскоре будут отвергнуты из-за разных недостатков, типа негибких настроек или недостаточной общности и возможности расширения и улучшения. Из-за чего прогресс может сильно замедлиться или вовсе остановиться.

Тем не менее, есть много причин радоваться будущему - сообщество постоянно предпринимает позитивные изменения, которые способствуют устранению проблем с измерением прогресса в эмпирическом машинном обучении.

Пример чек-листа

Benchmarking checklist for reviewers and area chairs

- ☐ If there is written dissatisfaction about the author's choice of baselines, tasks, or benchmarks in the reviews, are there rationals beyond the fact that these requested datasets are "must-have" benchmarks?
- ☐ Are the reviews considering potential benefits like efficiency, fairness, and simplicity of the proposed model outside the commonly evaluated performance metrics (e.g., accuracy)?
- ☐ Are there any negative points in the reviews due to the paper proposing a method that deviates from the current trend/hype. If so, are there rational justifications for this?
- ☐ If the reviews penalizing the paper due to the proposed method not performing well only on a subset of tasks, is there enough logical elaboration on such criticism in the reviews?
- ☐ Are the reviews assessing the evaluation strategy in terms of studying the effect of different sources of variance (e.g., multiple splits, multiple random seeds, etc.)?
- ☐ If there are analyses on statistical significance testing, are they appreciated in the reviews? If there is no such analysis, are there recommendations on this provided in the reviews?
- ☐ If the paper is claiming SOTA or improvements over baselines on a benchmark, are there ablations on how much such improvement is secured by the tricks that are not tied to the main contributions?
- ☐ If the reviews are asking for more experiments, analysis, or evaluation on more benchmarks, are the potential blockers are considered for such requests? E.g. those experiments being out of reach in terms of computing budget (pre-training or extremely large datasets).
- ☐ If the paper is proposing a new idea while deviating from the common paradigms, is the "out of the hype" thinking valued in the reviews as opposed to solely recognizing SOTA performance?