The benchmark lottery

Докладчик: Латышев Александр
Рецензент: Седашов Данила
Практик-исследователь: Адыгамов Ильяс
Хакер: Павлов Вадим

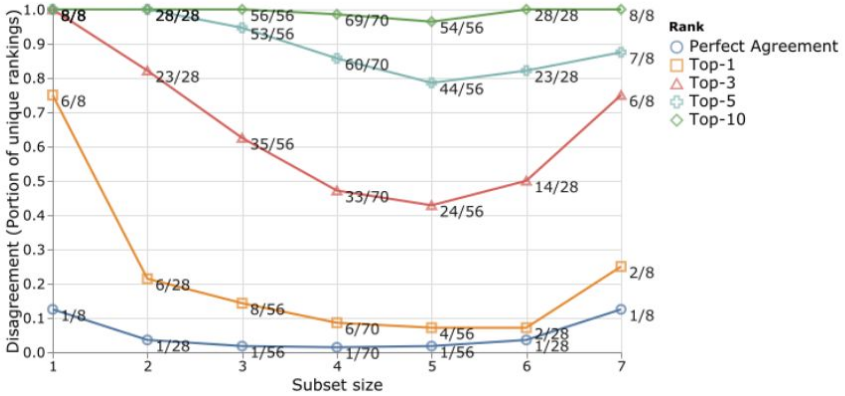
Идея возникновения

- Benchmark lottery - случайность нахождения эталонного результата.
- Достижение наименьшей ошибки часто зависит от многих неопределенностей.
- Относительные результаты одних и тех же моделей могут варьироваться в зависимости от постановки задачи.
- Ознакомить других авторов с потенциальными проблемами, которые могут затруднить оценку их модели и дать им варианты решения описанных в статье проблем.

Проблемы

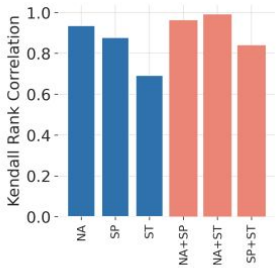
- Не всегда лучший алгоритм показывает лучший результат.
- Некоторые критерии оценок моделей не подвергаются сомнению, хотя должны.
- Ошибки в выборе лучших моделей могут замедлять развитие области в целом.
- Каждая группа людей в своей статье выбирает “удобные” методы, чтобы показать качество своей модели.

SuperGLUE

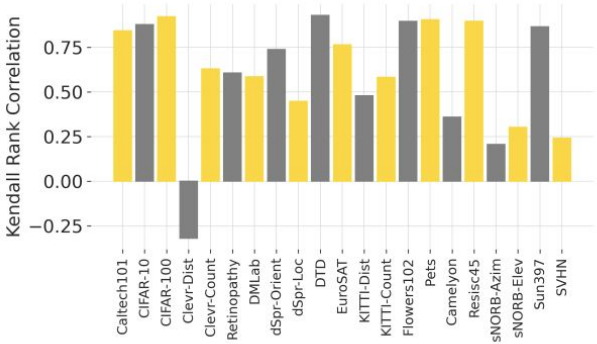


Один из примеров авторов, как рейтинг качества моделей может меняться в зависимости от выбора задач.

VTAB



(a) Different categories.



(b) Different task.

Пример согласованности результатов для моделей компьютерного зрения.

Long Range Arena

Table 3: Top 3 performing models on LRA depending on which subset of tasks we select.

Task	Best Model	Rank-2	Rank-3
t_1(Text only)	Linear Transformers	Performer	Transformer
t_2(Retrieval only)	Sparse Transformers	BigBird	Longformer
t_3(ListOps only)	Reformer	Synthesizer	Transformer
t_4 (Image only)	Sparse Transformer	Performer	Transformer
t_5(Path only)	Performer	Linformer	Linear Transformers
$t_1 + t_2$	BigBird	Sparse Transformer	Transformer
$t_1 + t_3$	Transformer	BigBird	Synthesizer
$t_1 + t_4$	Linear Transformer	Performer	Transformer
$t_1 + t_5$	Performer	BigBird	Transformer
$t_2 + t_3$	BigBird	Transformer	Longformer
$t_2 + t_4$	Sparse Transformer	BigBird	Transformer
$t_2 + t_5$	BigBird	Sparse Transformer	Performer
$t_3 + t_5$	Linformer	BigBird	Transformer
$t_3 + t_4$	Transformer	Synthesizer	Longformer
$t_4 + t_5$	Performer	Linear Transformer	Sparse Transformer
$t_1 + t_2 + t_3$	BigBird	Transformer	Synthesizer
$t_1 + t_2 + t_4$	Sparse Transformer	Transformer	BigBird
$t_1 + t_2 + t_5$	Performer	Linear Transformer	Transformer
$t_2 + t_3 + t_4$	Transformer	Longformer	Synthesizer
$t_2 + t_3 + t_5$	BigBird	Transformer	Longformer
$t_3 + t_4 + t_5$	BigBird	Transformer	Longformer
$t_1 + t_2 + t_3 + t_4$	Transformer	BigBird	Longformer
$t_1 + t_3 + t_4 + t_5$	BigBird	Transformer	Longformer
$t_1 + t_2 + t_4 + t_5$	Sparse Transformer	Performer	BigBird
$t_2 + t_3 + t_4 + t_5$	BigBird	Transformer	Longformer
$t_1 + t_2 + t_3 + t_4 + t_5$ (LRA Score)	BigBird	Transformer	Longformer

Не все так просто

- Простого набора различных задач недостаточно и усреднения их результатов. Веса могут быть разными.
- Есть различные варианты, какие веса ставить задачам. Например, разделяя их на темы.
- Человек тоже предвзят. И он принимает финальное решение в выборе задач.
- Общественное мнение оказывает влияние на направление развития области.

Критерии редко меняются

Авторы замечают, что в большинстве статей по одной теме все опираются на схожие функции ошибок на схожих наборах данных. Что со временем приводит к переобучению.

Повторное использование тестовой выборки или гиперпараметров других моделей искажает реальные результаты.

Больше попыток запустить модель на тестовых данных - больше шансов получить “лучшую” модель.

Рекомендательные системы

- Отсутствие первостепенных требований оценивания.
- В данных нет четкого разделения на тренировочную и тестовую выборки.
- Каждый автор показывает сильные стороны модели.
- Не получается определить, какая модель действительно лучшая.
- Сложность оценивания взаимодействия реального пользователя с системой лишь затрудняет процесс.

В результате, может создаваться иллюзия прогресса, когда последнего нет.

Arcade Learning Environment

Среда для разработки искусственного интеллекта игр, которую в подтверждение используют авторы статьи.

Сравнение различных алгоритмов происходит без заранее известных условий. Каждый автор выбирает удобные для себя.

Реальное сравнение моделей между собой невозможно.

Что можно улучшить

- Создать больше рекомендаций.
- Стандартизировать наборы данных и критерии оценивания.
- Пересмотр важности своевременного изменения критериев.
- Повышение важности простоты модели, а не только ее качества.
- Все тесты должны быть воспроизводимыми. Чтобы не было различий в одной и той же модели между публикациями.
- Разные ресурсы - разные результаты. Чем больше “денег и железа” и автора есть, тем выше его шансы получить лучшую модель.
- Оценивать качество моделей статистическими методами.

Что можно улучшить

- В большинстве случаев есть всего одно разбиение выборки на тренировочную/тестовую. Предлагается иметь несколько для уменьшения эффекта лотереи.
- Создание наборов данных по примеру VTAB и GLUE.
- Ограничить количество запросов каждой новой модели за тестовыми данными, либо вовсе варьировать последние. Чтобы не было переобучения. Превратить тест в “живой” организм(в пример авторы ставят GEM - критерий оценки генерации текстов).

Выводы

- Benchmark lottery - лишь один из вариантов рассмотрения проблемы переобучения моделей.
- Построение новых критериев сложный процесс и надо грамотно выделять на него ресурсы и время.
- Не смотря на все сказанные в статье недочеты, в последние годы меньше авторов пытаются выиграть в “лотерею”, и все больше областей имеют грамотно выбранные задачи.

Рецензент

Плюсы

- Много ссылок на существующие/собственные исследования. Строгие рассуждения (насколько это возможно)

- Большое количество экспериментов: 4 бенчмарка в разных областях, в каждом минимум 11 моделей

- Comprehensive survey по тому, что не так с бенчмарками. Подсвечивают новую проблему для community

Минусы

- Обзор литературы не сосредоточен в одном месте
- Местами сложно читаемый язык
- Нет всех результатов и гиперпараметров — приходится “верить на слово”

Оценка: 7, уверенность: 3

Практик-исследователь

Авторы

Все из Google. Из них 3 основных:

- *Mostafa Dehghani* - есть одна статья со словом benchmark в названии. Несколько статей формата “survey”
- *Yi Tay*
- *Alexey A. Gritsenko*

Статья опубликована летом 2021 года.

picardythird · 8 mo. ago

Oh look, Google writing a paper that will probably be published in a highly-ranked venue about a topic that literally everybody knows about.

10 Reply Share Report Save Follow

nonotan · 8 mo. ago

On the one hand, you're spot on. On the other hand, it's an important topic that isn't really being seriously tackled despite "everybody knowing about it". So if a Google paper puts it in the spotlight and gets people to act even a little faster, I can't complain too much.

5 Reply Share Report Save Follow

Статьи, которые цитируют эту статью

- https://arxiv.org/pdf/2112.01342.pdf - авторы(из РФ) меняют схему усреднения, после чего кардинально меняются значение метрики и положение моделей отнсительно друг друга
- https://proceedings.neurips.cc/paper/2021/file/f514cec81cb148559cf475e7426eed5e-Paper.pdf - авторы критикуют существующую схему оценки качества в RL и предлагают свою, основанную на статистической теории
- https://openreview.net/pdf?id=FBBWy2Sjwg - авторы поднимают ту же проблему и анализируют поведение моделей на граничных примерах, т.е. тех, которые можно отнести к обоим классам одновременно

Похожие статьи

- https://arxiv.org/pdf/2104.14337.pdf - авторы создают новый сильный динамично меняющийся бенчмарк, отмечая проблему плохих бенчмарков
- https://aclanthology.org/2021.naacl-main.385.pdf - авторы предлагают 4 критерия хороших бенчмарков, критикуют adversarial подход из предыдущей статьи