Рецензия на рукопись
Регуляризация траектории оптимизации параметров модели глубокого обучения на основе дистилляции знаний
М. Горпинич.

1. Статья содержит некоторое количество орфографических ошибок, опечаток. Предлагается использовать онлайн сервисы для быстрого поиска опечаток. (пример https://text.ru/spelling).
2. Выборка разбивается на train и validation. Необходима еще часть test для проверки итогового качества после подбора параметров и гиперпараметров. Это уберегает нас от переобучения самих себя и модели на валидации.
3. Непонятна фраза «Машинное обучение и анализ данных, 2017. Том ??, № ??» в конце каждой страницы.
4. На графиках точности классификации от количества итераций полосы дисперсий сильно перекрывают друг друга и сливаются в единый серый цвет. Кроме того, разница между самими кривыми очень маленькая и сложно различимыми (приходится всматриваться). Возможно стоит выбрать больший масштаб и/или отрисовывать на одном графике меньше кривых и/или поиграть с параметром прозрачности.
5. Для рисунков 6,7,8 выбраны неудачные цвета, все точки выглядят одинаково. Лучше выбрать максимально непохожие цвета (например, red, green, blue). То, что какие-то точки обозначены крестиком, а какие-то кружочком не улучшает ситуацию, так как в первую очередь глаз замечает различия в цветах.
6. В выводе к графику 9 написано, что достигается гораздо большая точность при использовании дистиляции, чем без нее. На графике действительно видно различие в 5%, но непонятно, почему это считается «гораздо большей точностью». Дисперсионные интервалы почти не перекрываются и это хорошо, конечно. Предлагается запустить несколько известных архитектур нейросетей (а может и не нейросетевых моделей) и показать что все они отстают от модели с дестиляцией на эти 5%, а может еще и больше. (но в целом можно просто написать, что достигнута «большая» точность, а не «гораздо большая».
7. Хотелось бы еще сравнить затраченное время на работу алгоритмов, так как в статье утрверждается, что градиентная оптимизация долгая, а линейное приближение и случайный выбор - быстрая. Хочется увидеть разницу.
8. Хотелось бы увидеть сравнение моделей на разных метриках, все-таки accuracy плохо отражает действительность на, на пример, несбалансированных классах. Стоит посмотреть в сторону таблички TP, TN, FP, FN и всяких ROC-AUC кривых.
9. Не понятно, почему в формулах 5, 6 в одном случае argmin, а во втором argmax, вроде мы хотим всегда минимизировать лосс. Ну мб это я туплю конечно, или просто не очевидно почему они разные.
10. Не понятно зачем использовать два разных коэффициента beta в лоссе, если по факту нас интересует только доля вклада каждого слагаемого в лосс. А домножением всего лосса на константу ничего не изменится. То есть можно положить beta_1 равной единичке, а beta_2 будет играть роль коэффициента доли.
11. В параграфе про обсуждение рисунка 7 написано «По графику видно, что значение температуры уменьшается при увеличении логарифма температуры,». Это кажется оговорка, так как логарифм — функция монотонно возрастающая, и при увеличении аргумента должен расти и логарифм аргумента.
12. В параграфах про графики 6, 7, 8 ведутся разговоры о зависимости одних параметров от других, но не хватает вывода по этой части — какие параметры то лучше брать, исходя из графиков. Судя по всему надо лучше всего получается при T=0 и как можно меньшим beta. Может этому есть какое-то теоретическое обоснование. Вообще кажется, что эти параметры можно не оптимизировать совсем а зафиксировать изначально нулем.

13. Нет таблицы1, рисунков 10, 11. Возможно появятся еще конечно.

Работа топ, есть списочек мелких косяков и советов того, что можно улучшить, но в целом я бы рекомендовал работу к публикации. Получены результаты, согласующиеся с экспериментом.

Рецензент:
Кулаков Я.М.