Введение

1. Катастрофическая забывчивость - что это.
2. 2017 год - решение проблемы by DeepMind. Оригинальный EWC-Fisher, EWC-метод аналогичен закрепление весов к эталонам на резинках разной упругости.
3. Придуманы другие варианты расчета важностей весов – SI (по изменению loss в процессе обучения), MAS (по модулю градиента сумм квадратов выходов), EWC-SIG (по суммарному по модулю сигналу, прошедшему через связь в НС)
4. WVA – веса не привязываются к эталонам, а имеют разную инерцию (в зависимости от важности) при смещении. Строго математически WVA это предельный случай EWC, где размер обучения датасету равен одному шагу на одном батче.

Зачем EWC для слонов? Почём EWC для слонов?

1. Зачем? Сохранение навыков при transfer-learning, fine-tuning для готовых языковых моделей под прикладные задачи.
2. Методы расчета важностей Fisher, MAS, SIG. Их достоинства и недостатки.
    - Fisher и MAS - задаем только функцию и считаем градиенты
      достоинства: просто реализовать код
      недостатки: трудно посчитать, т.к. tf и torch не считают градиенты батчами
    - SIG
      достоинства: важности считаются в прямом проходе - можно считать батчами
      недостатки: надо реализовать весь код расчета сигнала внутри модели - переработка кода модели
3. Методы закрепления весов:
    - EWC
      достоинства: лучшее качество, простота реализации - дорабатывается только loss
      недостатки: кушает 3x (размер модели) памяти
    - WVA
      достоинства: кушает 2x (размер модели) памяти
      недостатки: чуть хуже качество, необходимо дорабатывать код оптимизатора
4. Реализован EWC-Fisher, EWC-MAS для любых оптимизаторов и WVA-Fisher, WVA-MAS для SGD (без моментов) и Adam.

Эксперименты:

1. Структура экспериментов: берем 10 датасетов, учим глубокую НС последовательно, попутно меряем *среднюю* точность на всех изученных датасетах после каждого обучения
2. К чему применять WVA-ослабление к градиентам или приращениям весов?
   AA-тест для SGD без моментов – результат одинаков.
3. К чему применять WVA-ослабление к градиентам или приращениям весов?
   AB-тест для SGD, Adam-приращения, Adam-градиенты

Выводы: Adam-градиенты вообще не работает для WVA, Adam-приращения работает чуть хуже, чем SGD без моментов.

4. Какая WVA-вязкость лучше - гиперболическая 1/(1+imp) или экспоненциальная exp(-imp)? Для разных оптимизаторов?
   Выводы: exp всегда лучше, но в пределах погрешности и ее труднее считать, так что выбор - гиперболическая. Все это выполнено для обоих оптимизаторов SGD и Adam.
5. Оптимальный коэффициент ослаблений - почему он есть? Как его найти? Как зависит оптимальный коэффициент WVA-ослабления от количества датасетов?
   Выводы.
   Интуитивно: когда коэффициент маленький, то сеть хорошо учит текущий датасет, но быстрее забывает предыдущие датасеты, а когда коэффициент большой, сеть хорошо сохраняет навыки с выученных датасетов, но плохо обучается новым датасетам. Соответственно, должен быть оптимум где-то по средине.
   Экспериментально: оптимальный коэффициент есть, он не зависит от кол-ва последовательных датасетов, находится перебором по сетке.
6. Какой способ расчета важностей для WVA лучше: SIG, MAS или Fisher? В статьях нет нормального сравнения, нет расчета оптимальных гиперпараметров даже перебором.
   Выводы: экспериментально Fisher чуть лучше MAS, а MAS чуть лучше SIG, но все в пределах доверительного интервала.

Статья про реализации EWC на Хабре: «Вспомнить все» или решение проблемы катастрофической забывчивости для чайников

wva-plots.ipynb