Оптимизация нейронных сетей

Как было рассмотрено на предыдущем занятии, каждый слой нейронной сети из нашего первого примера преобразует данные следующим образом:

output = relu(dot(W, input) + b)

В этом выражении W и b — тензоры, являющиеся атрибутами слоя. Они называются весами, или обучаемыми параметрами слоя (атрибуты kernel и bias соответственно). Эти веса содержат информацию, извлеченную сетью из обучающих данных.

Первоначально эти весовые матрицы заполняются небольшими случайными значениями (этот шаг называется случайной инициализацией). Начальные представления не несут никакого смысла, но они служат начальной точкой. Далее, на основе сигнала обратной связи, происходит постепенная корректировка весов. Эта постепенная корректировка, которая также называется обучением, составляет суть машинного обучения.

Ниже перечислены шаги, выполняемые в так называемом цикле обучения, который повторяется столько раз, сколько потребуется:

1. Извлекается пакет обучающих экземпляров x и соответствующих целей y .
2. Сеть обрабатывает пакет x (этот шаг называется прямым проходом) и получает пакет предсказаний y_pred.
3. Вычисляются потери сети на пакете, дающие оценку несовпадения между y_pred и y.
4. Корректируются веса сети так, чтобы немного уменьшить потери на этом пакете.

В конечном итоге получается сеть, имеющая очень низкие потери на тренировочном наборе данных: малое несовпадение предсказаний y_pred с ожидаемыми целями y . Сеть «научилась» отображать входные данные в правильные конечные значения. В процессе обучения шаг 1 по сути это операция ввода/вывода. Шаг 2 и 3 это набор операций с тензорами. Наиболее сложный шаг – это шаг 4, корректировка весов сети. Разберемся, как по весам сети понять какой и насколько вес должен изменяться, чтобы модель обучалась.

Градиент — это производная операции с тензором, обобщение понятия производной на функции с многомерными входными данными, то есть на функции, принимающие на входе тензоры.

Рассмотрим входной вектор x , матрицу W , цель y и функцию потерь loss . Вы можете с помощью W вычислить приближение к цели y_pred и определить потери или несоответствие, между кандидатом y_pred и целью y:

```
y_pred = dot(W, x)
loss_value = loss(y_pred, y)
```

Если входные данные x и y зафиксированы, тогда это можно интерпретировать как функцию, отображающую значения W в значения потерь:

```
loss_value = f(W)
```

Допустим, что W0 — текущее значение W . Тогда производной функции f в точке W0 будет тензор gradient(f)(W0) с той же формой, что и W , в котором каждый элемент gradient(f) (W0)[i, j] определяет направление и величину изменения в loss_value , наблюдаемого при изменении W0[i, j] . Тензор gradient(f)(W0) — это градиент функции f(W) = loss_value в W0.

Производную некоторой функции g(x) от единственного аргумента можноинтерпретировать как наклон кривой этой функции. Аналогично gradient(f)(W0) можно интерпретировать как тензор, описывающий кривизну f(W) в окрестностях W0. Соответственно, как и в случае с функцией g(x), значение которой можно уменьшить, немного сместив x в направлении, противоположном производной, функцию f(W) тензора также можно уменьшить, сместив W в направлении, противоположном градиенту: например, W1 = W0 - step * gradient(f)(W0) (где step — небольшой по величине множитель). Это означает, что для снижения нужно идти против кривизны. Обратите внимание: множитель step необходим, потому что gradient(f)(W0) лишь аппроксимирует кривизну в окрестностях W0 , поэтому очень нежелательно уходить слишком далеко от W0.

Стохастический градиентный спуск

Как известно, минимум функции — это точка, где производная равна 0. То есть остается только найти все точки, где производная обращается в 0, и выяснить, в какой из этих точек функция имеет наименьшее значение.

Применительно к нейронным сетям это означает аналитический поиск комбинации значений весов, при которых функция потерь будет иметь наименьшее значение. Этого можно добиться, решив уравнение gradient(f)(W) = 0 для W . Это полиномиальное уравнение с N переменными, где N — количество весов в сети. Решить уравнение для случая N = 2 или N = 3 не составляет труда, но превращается в практически неразрешимую задачу для нейронных сетей, в которых количество параметров редко бывает меньше нескольких тысяч и часто достигает нескольких десятков миллионов.

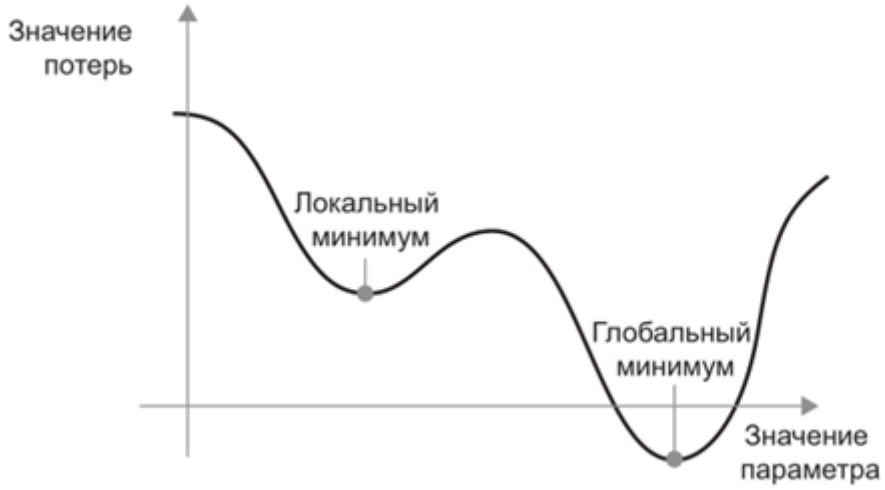
Поэтому на практике используется алгоритм из четырех шагов, представленный в начале этого занятия: вы можете понемногу изменять параметры, опираясь на текущие значения потерь в случайном пакете данных. Поскольку функция дифференцируема, можно вычислить ее градиент, который позволяет эффективно реализовать шаг 4, который называется обратный проход. И в итоге в алгоритме 4 шаг разбивается на 2:

1. Вычисляется градиент потерь для параметров сети (обратный проход).
2. Параметры корректируются на небольшую величину в направлении, противоположном градиенту, например W -= step * gradient , и тем самым снижаются потери.

Выбор разумной величины шага step имеет большое значение. Если выбрать его слишком маленьким, спуск потребует большого количества итераций и может застрять в локальном минимуме. Если шаг будет слишком большим, ваши корректировки могут приобретать нецеленаправленный характер и приводить в случайные точки на кривой.

По сути, описанный выше алгоритм является стохастическим градиентным спуском на небольших пакетах (mini-batch stochastic gradient descent, minibatch SGD). Термин «стохастический» отражает тот факт, что каждый пакет данных выбирается случайно (в науке слово «стохастический» считается синонимом слова «случайный»). На самом деле истинный SGD это когда в каждой итерации используется единственный образец и цель, а не весь пакет данных.

Существует также множество вариантов стохастического градиентного спуска, которые отличаются тем, что при вычислении следующих приращений весов принимают в учет не только текущие значения градиентов, но и предыдущие приращения. Примерами могут служить такие алгоритмы, как SGD с импульсом, Adagrad, RMSProp и некоторые другие. Эти варианты известны как методы оптимизации, или оптимизаторы. В частности, внимания заслуживает идея импульса, которая используется во многих этих вариантах. Импульс вводится для решения двух проблем SGD: невысокой скорости сходимости и попадания в локальный минимум. Рассмотрим функцию, которая имеет форму, представленную на рисунке



Как видите, для значения данного параметра имеется локальный минимум: движение из этой точки влево или вправо повлечет увеличение потери. Если корректировка рассматриваемого параметра осуществляется методом градиентного спуска с маленьким шагом обучения, тогда процесс оптимизации может застрять в локальном минимуме, не найдя пути к глобальному минимуму. Таких проблем можно избежать, если использовать идею импульса, заимствованную из физики. Вообразите, что процесс оптимизации — это маленький шарик, катящийся вниз по кривой потерь. Если шарик имеет достаточно высокий импульс, он не застрянет в мелком овраге и окажется в глобальном минимуме. Импульс реализуется путем перемещения шарика на каждом шаге, исходя не только из текущей величины наклона (текущего ускорения), но также из текущей скорости (набранной в результате действия силы ускорения на предыдущем шаге). На практике это означает, что приращение параметра w определяется не только по текущему значению градиента, но также по величине предыдущего приращения параметра:

```
past_velocity = 0.
momentum = 0.1
while loss > 0.01:
    w, loss, gradient = get_current_parameters()
    velocity = past_velocity * momentum + learning_rate * gradient
    w = w + momentum * velocity - learning_rate * gradient
    past_velocity = velocity
    update_parameter(w)
```

Объединение производных: алгоритм обратного распространения ошибки

В предыдущем алгоритме мы произвольно предположили, что, если функция дифференцируема, мы можем явно вычислить ее производную. На практике функция нейронной сети состоит из множества последовательных операций с тензорами, объединенных в одну цепочку, каждая из которых имеет простую, известную производную. Например, пусть есть сеть f , состоящая из трех операций с тензорами a , b и c и весовыми матрицами W1 , W2 и W3 :

f(W1, W2, W3) = a(W1, b(W2, c(W3)))

Формула сообщает нам, что такую цепочку функций можно получить с использованием следующего тождества, которое называется цепным правилом: f(g(x)) = f'(g(x)) * g'(x) . Применение цепного правила к вычислению значений градиента нейронной сети приводит к алгоритму, который называется обратным распространением ошибки (Backpropagation), или обратным дифференцированием. Обратное распространение начинается с конечного значения потери и

движется в обратном направлении, от верхних слоев к нижним, применяя цепное правило для вычисления вклада каждого параметра в значение потери.

С другими методами обучения можно ознакомиться по данной **ссылке**