УДК 004.93'1

Оптимизация быстродействия методом послойной аппроксимации глубоких нейронных сетей

Е.Е. Лимонова[1], Д.А. Ильин[2], Д.П. Николаев[3]

[1]Московский физико-технический институт (государственный университет)

[2]Институт системного анализы Российской Академии Наук

[3]Институт проблем передачи информации им. А.А. Харкевича Российской Академии Наук

Нейронные сети широко используются для распознавания образов, которое на сегодняшний день получает все большее распространение. Становится возможным решать задачи распознавания образов, такие как распознавание текста, речи, потоковая классификация объектов, не только на персональных компьютерах, но и на мобильных устройствах, и во встроенных системах. Зачастую подобные устройства обладают ограниченными ресурсами, поэтому используемые алгоритмы распознавания образов должны работать не только достаточно качественно, но и достаточно быстро.

В нейронных сетях, используемых для распознавания образов, нейроны чаще всего расположены по слоям. Вычисления в каждом слое можно представить в виде матричных операций [1]:

$$\mathbf{y} = \varphi(W\mathbf{x} + \mathbf{b}), \tag{1}$$

где $\mathbf{x}$ – входной вектор слоя, $\mathbf{y}$ – выходной вектор, W – матрица весов, $\mathbf{b}$ – вектор смещений нейронной сети, φ – функция активации.

Одним из методов увеличения скорости работы нейронной сети является использование целочисленной арифметики. Это означает, что весовые коэффициенты представляются в виде целых чисел путем масштабирования. Операции над целыми числами маленького размера (в том числе запись и чтение из памяти) выполняются процессором быстрее, чем операции над вещественными числами. В результате нейронная сеть работает быстрее. Целочисленная арифметика эффективно работает в нейронных сетях с ограниченными функциями активации, поскольку все промежуточные вычисления остаются ограниченными и целочисленный тип данных не переполняется [2].

Однако целочисленная арифметика отличается от вещественной по ряду причин. Первая из них заключается в том, что вещественные коэффициенты не всегда имеют точное целочисленное представление. Кроме того, при определении арифметических операций, в частности, умножения, возникает ряд нюансов, связанных с округлением. Например, можно округлять с отбрасыванием дробной части, к ближайшему целому, к нулю и т.д. На разных архитектурах разные реализации округления могут иметь различную

эффективность. Таким образом, арифметические операции, определенные в целочисленной арифметике, лишь приближают соответствующие операции над исходными коэффициентами.

Рассмотрим более общий случай: замену какой-либо используемой операции на ее аппроксимации. При этом качество работы нейронной сети может упасть. Особенно это заметно для глубоких нейронных сетей, в которых ошибка будет увеличиваться с ростом числа слоев. Обучение новой нейронной сети с новой моделью вычислений трудно осуществимо на практике, поскольку готовые программные пакеты не поддерживают изменение модели вычислений. Предложенный метод заключается в послойной замене точных операций их аппроксимациями и последующим дообучением остальной части сети, в которой все еще используются точные операции. То есть, на первом шаге исходная обучающая выборка подается на вход нейронной сети с измененным первым слоем. Данные на выходе первого входа сохраняются и используются для дообучения оставшейся части сети, после чего меняется модель вычислений второго слоя и т.д.

В нашей работе для проведения экспериментов использовались нейронные сети из 5 слоев (порядка 10000 весовых коэффициентов в слое) со сверточной архитектурой, предназначенные для распознавания латинских букв и цифр. Исследование выполнялось на процессоре Samsung Exynos 5422. Процессоры со схожими характеристиками зачастую используются на мобильных устройствах и во встроенных системах. Мы использовали 8 бит для представления коэффициентов нейронной сети, сложение было реализовано с насыщением, а умножение – с отбрасыванием дробной части. Такая реализация позволила ускорить матричные операции практически в 2.5 раза, распознавание в целом – на 45%. Несмотря на то, что после применения такой аппроксимации качество распознавания упало в 50 раз, предложенный метод позволил восстановить его до приемлемого значения.

Таким образом, послойное дообучение нейронной сети позволяет использовать различные аппроксимации вычислений без потери качества работы, что может быть важно для ускорения вычислений. Кроме того, предложенный метод позволяет осуществлять дообучение нейронной сети без изменения инфраструктуры обучения и смены используемых алгоритмов обучения.

Работа частично финансово поддержана грантом РФФИ номер 13-07-12172.

Литература

1. *Хайкин С.* Нейронные сети. Полный курс. 2-е изд.: пер. англ. — М.: Вильямс, 2006. — 1104 с.
2. *Vanhoucke V., Senior A., Mao M. Z.* Improving the speed of neural networks on CPUs // Deep Learning and Unsupervised Feature Learning Workshop, NIPS 2011.— 2011