УДК 004.93

Метод снижения влияния исчезающего градиента в полносвязном ядре сверточной сети

А.А. Алексеев
(Университет ИТМО, г. Санкт-Петербург)
Научный руководитель – д. т. н., Ю.Н. Матвеев
(Университет ИТМО, г. Санкт-Петербург)

Одним из важнейших этапов при обучении сети является выбор ее параметра глубины. Известно, что теоретически возможно достичь получения сложных решающих функций путем каскадного соединения нешироких сетей. При этом может быть вычислительный выигрыш при тех же качественных показателях. На практике увеличение глубины на стандартных сетях в определенный момент приводит к ухудшению работы, по причине не связанной с переобучением. Данная проблема вызвана снижением прироста градиента при обучении методом обратного распространения ошибки. В литературе эта проблема определяется как проблема исчезающего (или взрывного) градиента.

Одним из переломных моментов при решении данной проблемы явилось появление сетей типа ResNet [1] и использование в них так называемых residual блоков. Основная цель использования данных блоков состоит в решении двух задач – обеспечении Identity функции от выхода сети к ее входу, а также обучения весов сети, необходимых для решения задач классификации или регрессии. Подобный подход позволил создавать сети с большой глубиной, например как в сети ResNet101. При детальном анализе сети [2] выяснилось, что набор базовых residual блоков эквивалентен ансамблю простых (shallow) сетей и не оптимален с точки зрения решения задачи создания сети очень большой глубины. Известные свойства ансамбля, такие как возможность перестановки и ее избыточность, подтвердились. Также исследование показало, что по-прежнему большее влияние на эффективность градиента оказывают короткие пути между выходным и входным слоями сети. В работе [3] было предложено решение проблемы для сверточной сети CNN, в которой при меньшей длине по сравнению с ResNet, была обеспечена одинаковая ошибка при ее работе.

В нашей работе мы хотим показать, что по аналогии с [3] для CNN, возможен отказ от Identity как параллельного пути, путем его внедрение в саму сеть полносвязного типа FC и получение меньшего влияния эффекта vanishing gradient descent, без использования явного ансамбля простых сетей как у ResNet. В нашей работе мы решили данную задачу путем случайной генерации связей между нейронами соседних уровней, в которых веса фиксируются и равны 1. Количество связей зависит от количества нейронов уровня Ln+1. Такой подход оправдан в том числе в случае неравного количества нейронов между уровнями. Чтобы полностью устранить влияние ансамбля простых сетей, мы обучили сеть с постепенным снижением влияния Identity связей между уровнями в процессе обучения. Таким образом начало обучения обеспечивает качественный прирост градиента, постепенно переводя сеть из параллельно в последовательно связанные уровни. Результаты анализа градиента на начальном уровне обучения сети показали его больший прирост на всех уровнях сети по сравнению со стандартным вариантом.

Литература

1. Deep Residual Learning for Image Recognition / K. He [и др.] // CoRR. —2015. — T. abs/1512.03385. — arXiv: 1512.03385. — URL: http://arxiv.org/abs/1512.03385..
2. Veit, A. Residual Networks are Exponential Ensembles of Relatively Shallow Networks / A. Veit, M. J. Wilber, S. J. Belongie // CoRR. — 2016. — T. abs/1605.06431. — arXiv: 1605 . 06431. — URL: http : / / arxiv . org / abs /1605.06431.
3. Zagoruyko, S. DiracNets: Training Very Deep Neural Networks Without Skip-Connections / S. Zagoruyko, N. Komodakis // CoRR. — 2017. — T. abs/1706.00388. — arXiv: 1706 . 00388. — URL: http : / / arxiv . org / abs /1706.00388.