Краткий отчет

Было исследовано поведение следующих методов оптимизации при обучении нейронной сети без и с использованием *Batch Normalization (BN)*: SGD, SGDm (momentum), Adam, Adagrad, Adadelta, RMSProp.

Эксперименты проводились на полносвязной сети с тремя скрытыми слоями по 100 нейронов с функцией активации ReLU и 10 нейронами на последнем слое с softmax'ом на выходе. Для обучения использовался датасет MNIST. При обучении минимизировалась кросс-энтропийная функция ошибок.

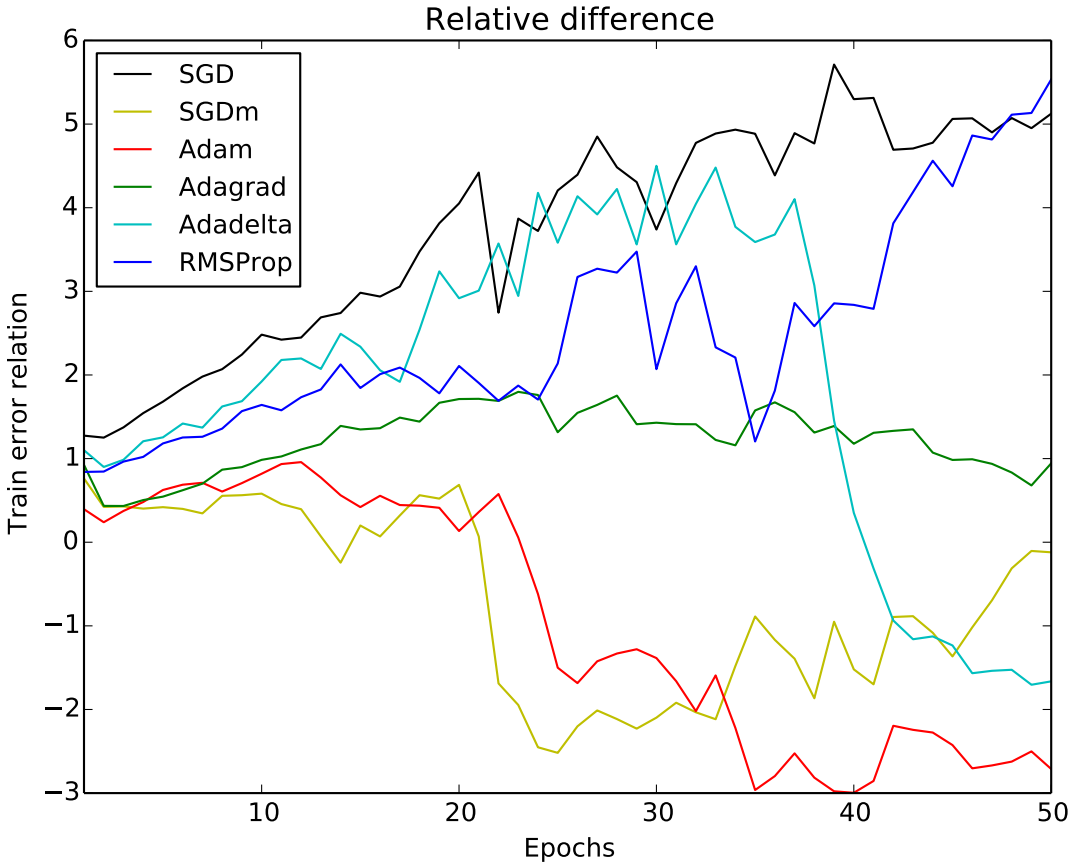

Рис. 1: Относительное уменьшении ошибки при использовании BN

На графике 1 отложено значение $\log \frac{x}{y}$ после каждой эпохи обучения, где x — ошибка обучения метода без использования BN, y — с использованием BN, то есть отрицательные значения показывают, что второй работает хуже (ошибка больше).

Из графика 1 видно, что в худшую сторону выделяются два метода: SGDm и Adam, а так же Adadelta падает на последних эпохах.

Если же посмотреть на график 2 ошибки методов без использования BN, то видно, что эти три метода оптимизации работают лучше всего.

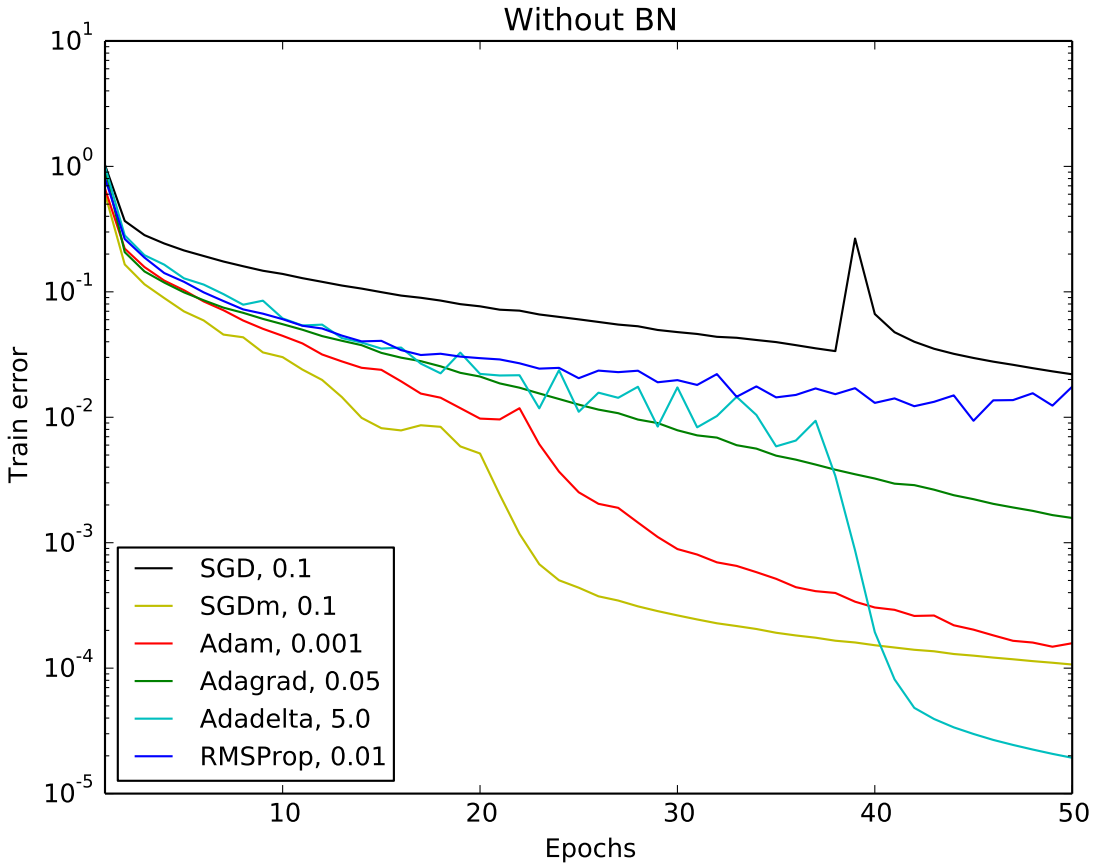

Рис. 2: Величина ошибки без использования BN

Тогда можно предположить, что BN не дает прибавки методам, просто потому что они и так быстро сходятся.

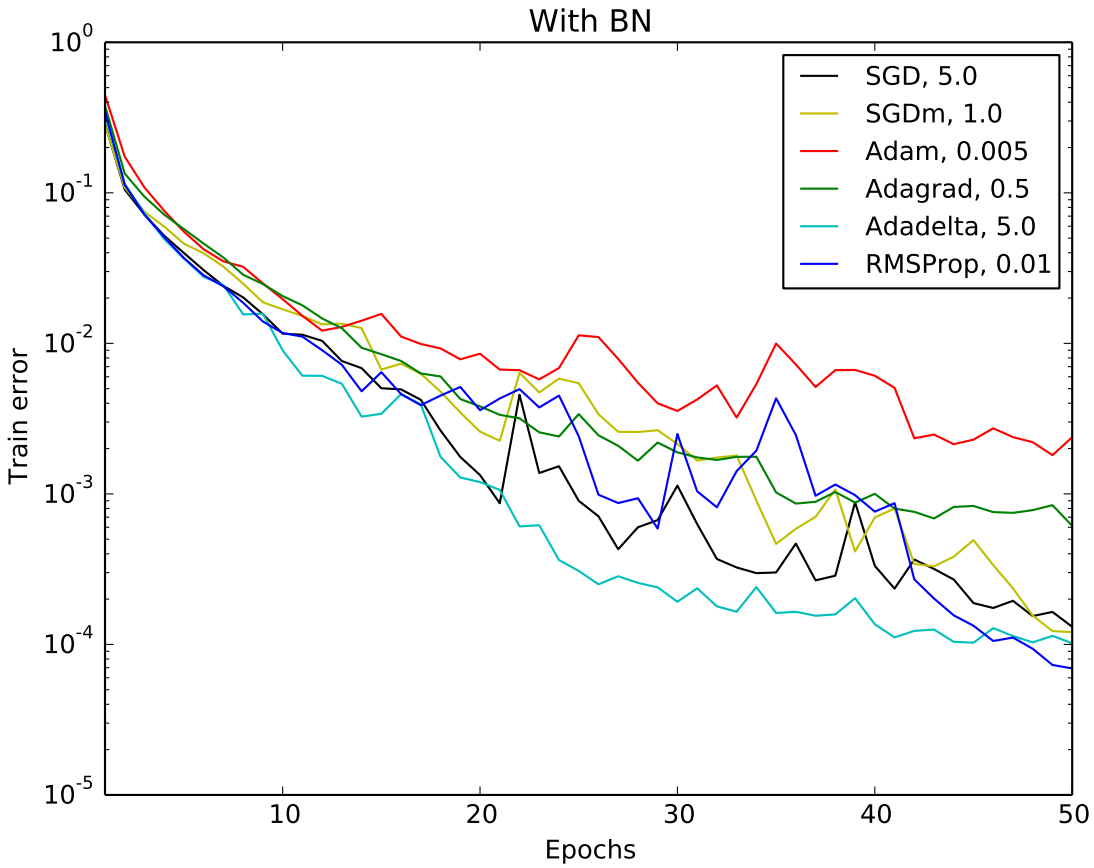

Рис. 3: Величина ошибки при использовании BN

Однако на графике 3 мы видим, что Adam при использовании BN работает хуже всех остальных методов, когда для SGDm и Adadelta это совсем не так (отметим, что для каждого метода оптимизации мы подобрали лучший шаг α отдельно без и с использованием BN).

Поэтому возникает предположение о несовместимости метода Adam и использовании *Batch Normalization*.