Постановка задачи

Целью работы было оценить практическую применимость статьи [1]. Были выбранны следующие критерии оценки:

1. Стабильность обучения.
2. Качество обучения.
3. Скорость сходимости сети.
4. Обобщающая способность.

Для проведения тестов использовались следующие датасеты:

1. 2 концентрические, 100-мерные сферы различного радиуса.
2. 2 100-мерных набора сэмплов из нормальных распределений с одинаковыми дисперсиями и разными средними.
3. Подмножество датасета CIFAR (классы 0 и 1). Размер датасета: 2000 - train, 2000 - test, 2000 - validation.
4. Подмножество датасета MNIST (цифры 1 и 7). Размер датасета: 13000 - train, 2000 - test, 2000 - validation.

Во всех сетях в качестве функции потерь рассматривался модифицированный Hinge Loss:

$$\mathfrak{L}(y_{pred}, y_{real}) = (max(-y_{real}y_{pred} + 1; 0))^3 \tag{1}$$

Так же, в соответствии со статьёй [1], использовался регуляризатор веса α в ЭН:

$$\mathfrak{L}_{reg}(\alpha) = \lambda\frac{\alpha^2}{2} \tag{2}$$

Где λ - гипер параметр, отвечающий за степень регуляризации.

Стабильность

В ходе обучения сети с экспоненциальным нейроном (ЭН) проявлялся эффект нестабильности обучения - в любой момент ошибка на обучающей выборке может "взорваться"и затем, как следствие, расходятся градиенты и веса сети. Причина этого в чрезвычайной чувствительности ЭН к малым изменениям весов в нём. Это подтверждает такой эксперимент: к весам нейрона добавляется нормальный шум из $N(0, \sigma)$, где σ в 10 раз меньше по величине чем среднее значение модулей весов в нейроне. Такое изменение ЭН сразу приводит к расхождению сети, хотя в случае нейрона с не экспоненциальной функцией активации (ReLU, Sigmoid, ...) такого не происходит. Частично проблему "взрыва"градиента удаётся решить нормализацией входных данных, за счёт чего на ранних итерациях обучения не происходит расхождения сети. Так же, уменьшение learning rate с числом пройденых итераций делает обучение стабильнее. Однако полностью проблему решить не удалось.

Качество

Основным результатом статьи [1] было доказательство отсутствия локальных минимумов у получившейся сети. Было выдвинуто предположение, что существуют задачи, в которых обучение сети без ЭН приводит к "застреванию"в локальном минимуме, в то время как добавление ЭН приведёт к попаданию в глобальный минимум и, как следствие лучшему качеству на обучающей выборке. Однако, проведённые эксперименты эту гипотезу не доказывают:

Датасет	Train Size	Без ЭН	С ЭН	Архитектура сети
Сферы, $R_1 = 1.0, R_2 = 1.01$	10000	0.000\0	0.000\0	4
Сферы, $R_1 = 1.0, R_2 = 1.01$	15000	44.3\7	3693.2\1096	4
Сферы, $R_1 = 1.0, R_2 = 1.01$	20000	8311.8\2338	335.8\35	4
Гауссианы, $\mathbb{E}_1 = 0.0, \mathbb{E}_2 = 0.1$	15000	0.000\0	0.000\0	4
Гауссианы, $\mathbb{E}_1 = 0.0, \mathbb{E}_2 = 0.1$	20000	41.0\0	53.9\0	4
MNIST	13000	0.000\0	0.000\0	6
CIFAR	2000	0.431\1	0.994\2	5

Таблица 1: Результаты обучения. Loss \ misclassification rate.

Как видно из таблицы 1 нельзя однозначно сказать, что ЭН улучшает или ухудшает качество, т.к. наблюдаются результаты как в одну , так и в другую сторону.

При проведении экспериментов возникли следующие сложности:

1. На реальных данных (MNIST/CIFAR) задача бинарной классификации решается с 0 misclassification rate даже сетью с очень маленьким числом обучаемых параметров и, как следствие такие датасеты нельзя использовать для сравнения качества. Поэтому для оценки приходилось использовать датасеты с трудноразделимыми классами с большим числом объектов. Как следствие, обученные сети, по сути, переобучались, запоминая примеры обучающей выборки на весах, и говорить о том, что в реальных задачах результаты будут аналогичны — нельзя.

2. Чтобы обучить сеть с ЭН приходилось подбирать параметр λ и стратегию уменьшения learning rate , т.к. в ином случае сеть рано или поздно расходилась.

Скорость обучения

Так как и сеть без ЭН, и с ЭН зависят от гипер параметров (learning rate, reg. lambda), то для сравнения скорости сходимости использовался следующий подход: гипер параметры перебираются из заданного множества и для каждой их комбинации считается число итераций до достижения 0 misclassification rate. Лучший результат (наименьшее число итераций) определяет скорость сходимости сети.

Датасет	Train Size	Без ЭН	С ЭН	Архитектура сети
Сферы, $R_1 = 1.0, R_2 = 1.01$	1000	20	23	4
Сферы, $R_1 = 1.0, R_2 = 1.01$	5000	28	35	4
Сферы, $R_1 = 1.0, R_2 = 1.01$	10000	77	126	4
Гауссианы, $\mathbb{E}_1 = 0.0, \mathbb{E}_2 = 0.1$	10000	45	67	4
Гауссианы, $\mathbb{E}_1 = 0.0, \mathbb{E}_2 = 0.1$	15000	95	131	4
MNIST	13000	15	20	6
CIFAR	2000	20	23	5

Таблица 2: Скорость обучения. Количество итераций до сходимости.

Как видно из таблицы 2 добавление ЭН не улучшает скорость сходимости сети, а в некоторых случаях ухудшает.

Обобщающая способность

Для оценки обобщающей способности был проведён следующий эксперимент: аналогично предыдущей секции, перебирались всевозможные комбинации гипер параметров. Для каждой комбинации, на которой за фиксированное число итераций на обучающей выборке достигался нулевой misclassification rate, вычислялась ошибка на тестирующей выборке. Затем, для набора гипер параметров, на котором достигнут минимум ошибки на тестирующей выборке вычислялась ошибка на валидационном множестве. Эта ошибка и считалась характеристикой обобщающей способности сети.

Датасет	Train Size	Без ЭН	С ЭН	Архитектура сети
Сферы, $R_1 = 1.0, R_2 = 1.5$	5000	2684.8\658	6901.4\796	4
Сферы, $R_1 = 1.0, R_2 = 5.0$	5000	678.1\140	2351.7\188	4
Гауссианы, $\mathbb{E}_1 = 0.0, \mathbb{E}_2 = 0.5$	5000	384.2\56	788.6\63	4
Гауссианы, $\mathbb{E}_1 = 0.0, \mathbb{E}_2 = 1.0$	5000	1.1\0	4.6\0	4
MNIST	13000	0.2\5	0.6\4	6
CIFAR	2000	3901.5\314	4516.7\313	5

Таблица 3: Результаты обучения. Loss \ misclassification rate.

Как видно из таблицы 3, добавление ЭН только ухудшило обобщающую способность.

Выводы

Проведённые эксперименты показывают, что добавление ЭН не позволяет улучшить ни одну из характеристик, которую обычно оптимизируют при использовании машинного обучения (ни качество на тестовой и обучающей выборках, ни скорость обучения), а с учётом значительного ухудшения стабильности обучения и необходимостью дополнительно подбирать оптимальное значение ещё одного гипер параметра λ делает практическое применение данной модификации невыгодным.

Список литературы

[1] Shiyu Liang, Ruoyu Sun, Jason D. Lee, R. Srikant, Adding One Neuron Can Eliminate All Bad Local Minima, arXiv:1805.08671

Приложение 1

Архитектура сетей

Название модуля	Входной размер	Выходной размер
X	—	100×1
Linear	100×1	50×1
Sigmoid		
Linear	50×1	25×1
Sigmoid		
Linear	25×1	1×1

Таблица 4: Архитектура полносвязной сети 1

Название модуля	Входной размер	Выходной размер
X	—	3072×1
Linear	3072×1	1536×1
Sigmoid		
Linear	1536×1	768×1
Sigmoid		
Linear	768×1	1×1

Таблица 5: Архитектура полносвязной сети 2

Название модуля	Входной размер	Выходной размер	Kernel Size	Stride	Padding
X	—	$1 \times 28 \times 28$			
Conv	$1 \times 28 \times 28$	$64 \times 27 \times 27$	4×4	0	1
PReLU					
MaxPool	$64 \times 27 \times 27$	$64 \times 9 \times 9$	3		
Conv	$64 \times 9 \times 9$	$128 \times 8 \times 8$	4×4	0	1
PReLU					
MaxPool	$128 \times 8 \times 8$	$128 \times 2 \times 2$	3		
Conv	$128 \times 2 \times 2$	$256 \times 1 \times 1$	4×4	0	1
PReLU					
Linear	256×1	1×1			

Таблица 6: Архитектура свёрточной сети