UNDERSTANDING DEEP LEARNING REQUIRES RETHINKING GENERALIZATION

Халиков Даниил

МОП 162

Мотивация

- Generalization error – разница ошибок на обучающих и тестовых выборках.
- В чем различие между хорошо и плохо обобщающими сетями?
- Почему это важно – модели становятся более интерпретируемыми, понимание приводит к более надежному построению моделей.

Как можно понять «обобщение»?

- Рademахеровская сложность

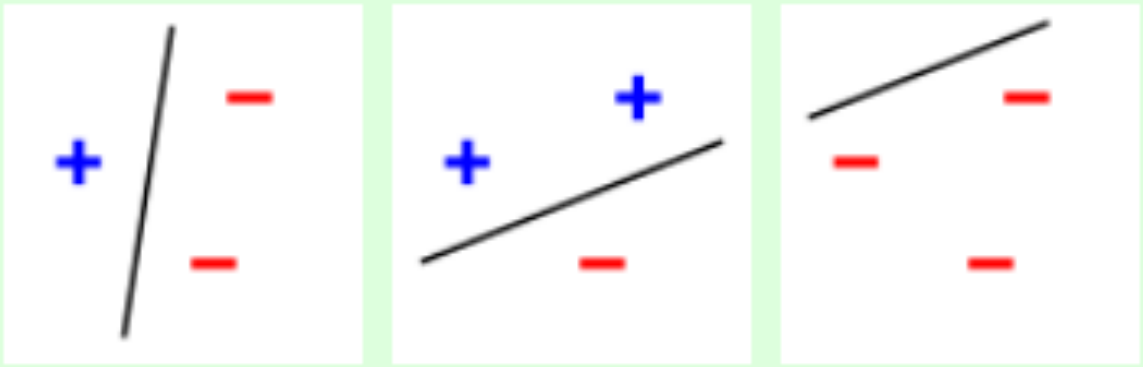
$$\mathrm{Rad}_S(F) = \frac{1}{m}\,\mathrm{E}\left[\sup_{f\in F}\sum_{i=1}^{m}\sigma_i f(z_i)\right]$$

- δ – независимые случайные величины равновероятно принимающие +1/-1.
- F – семейство функций, в нашем случае нейросеть.

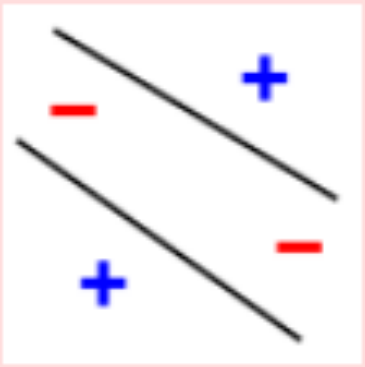
Как можно понять «обобщение»?

- Пусть классификатор $f(\theta)$ разбивает множество точек, если при любом присвоении лейблов существует такое θ, что f не делает ошибок на этом множестве.
- Размерность Вапника-Червоненкиса – такое максимальное количество точек, что f будет их разбивать.

Размерность Вапника-Червоненкиса



Примеры разделения трёх точек на два класса

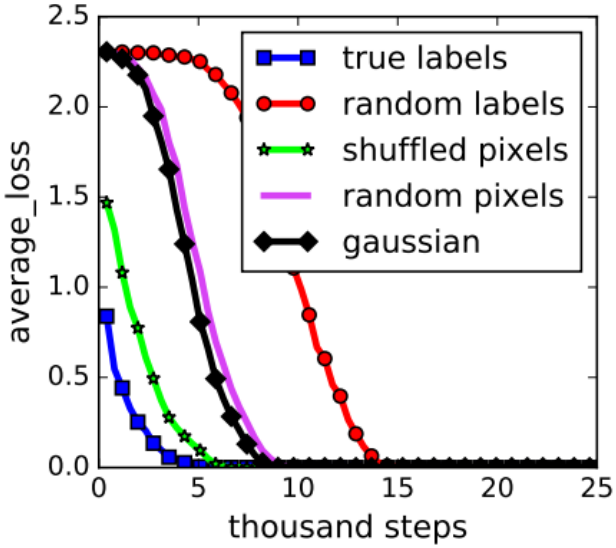


Разделение невозможно для этих четырёх точек

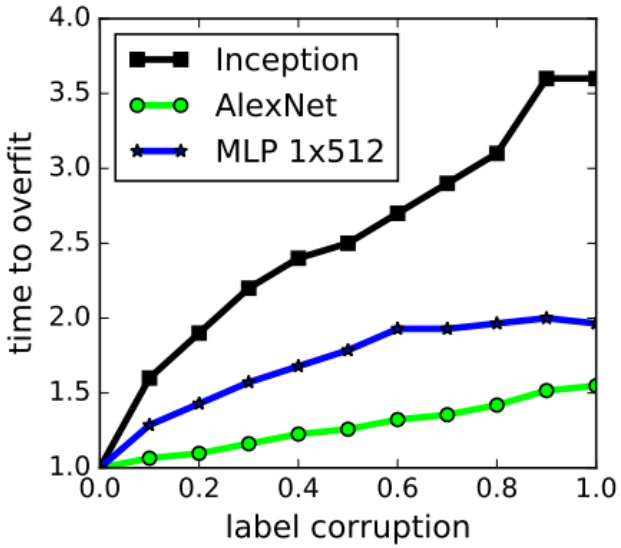
Randomization-тесты

- Partially corrupted labels
- Random labels
- Shuffled pixels
- Random pixels
- Gaussian

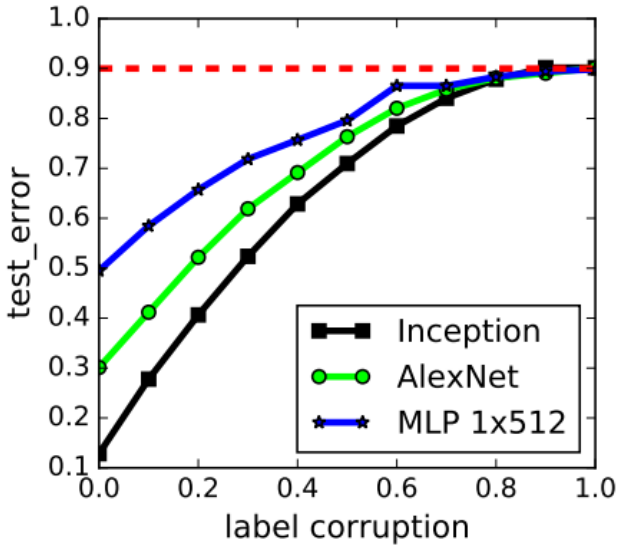
Результаты



(a) learning curves



(b) convergence slowdown



(c) generalization error growth

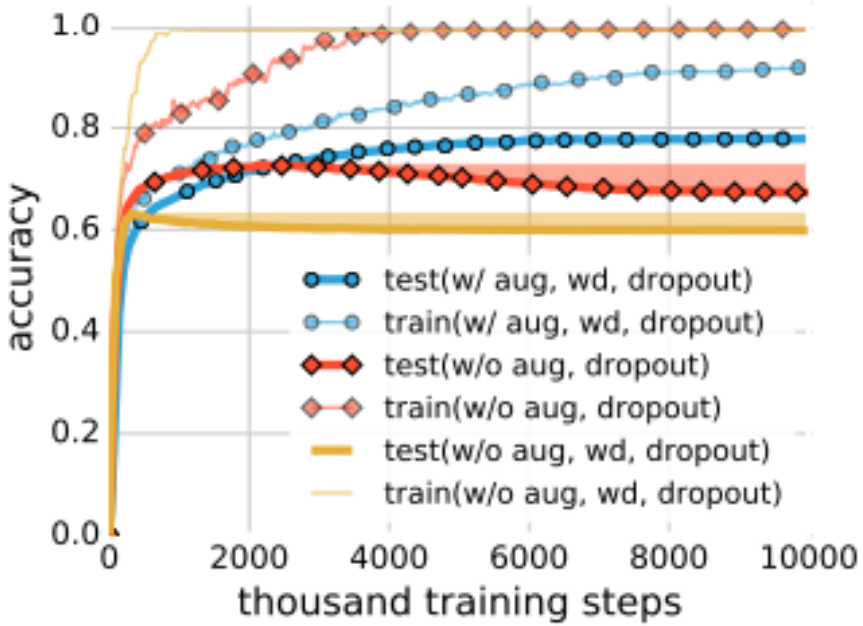
Результаты

- Глубинные нейросети легко достигают нулевой ошибки на обучающей выборке и способны «запомнить» ее.
- Радемахеровская сложность и VC-размерность не подходят для объяснения наблюдаемых явлений.
- Процесс обучения остается таким же легким, время обучения возросло на малую константу.

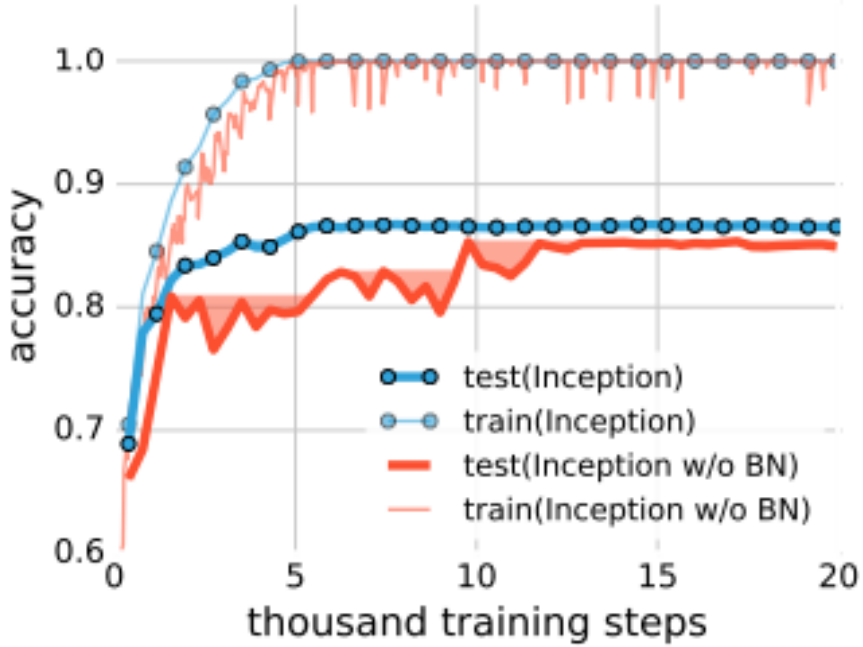
Регуляризация

- Используются:
  - Dropout
  - Batch-normalization
  - Data augmentation
  - Weight decay

Регуляризация



(a) Inception on ImageNet



(b) Inception on CIFAR10

Выводы

- Лучше менять саму архитектуру сети, чем добавлять регуляризаторы.
- Регуляризация может помочь обобщению, но она ни необходима, ни достаточна для нее.
- BN в целом улучшает обобщение.
- Ранняя остановка обучения иногда улучшает обобщение.

Finite-Sample Expressivity

- В большинстве работ рассматривается вся область выборок и какие функции могут или не могут быть представлены нейросетью.
- В целом k-глубинные сети в целом более успешны чем (k – 1)-глубинные.
- Более применимо на практике – выразительность на конечном сэмпле размера n.

Finite-Sample Expressivity

- Даже двухслойная нейросеть может представлять любую функцию на выбранном сэмпле размера n, если количество параметров p больше n.
- Теорема:
  - Существует двухслойная нейросеть с ReLu-активациями и 2n+d параметрами, что она может представить любую функцию на сэмпле размером n в d размерностях.

Приложение к линейным моделям

- Выборка из точек размера n вида (x, y) где x – d размерный вектор, y – лейбл.
- Решаем следующую задачу:

$$\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \text{loss}(w^T x_i, y_i)$$

- Если d >= n, то можем соответствовать любой разметке.

Приложение к линейным моделям

- X – n d матрица, где i строчка – x_i^T
- Если X имеет ранг n, то Xw = y имеет бесконечное множество решений.
- Все ли они обобщают одинаково хорошо?
- Исследуем к чему будет сходиться SGD:

$$w_{t+1} = w_t - \eta_t e_t x_{i_t}$$

- Если $w_0 = 0$,то $w = \sum_{i=1}^{n} \alpha_i x_i$, а значит $w = X^T \alpha$

Приложение к линейным моделям

- Подставляем

  $$XX^T\alpha = y$$

- Имеем единственное решение
- Такой “kernel trick” дает удивительно хорошие результаты, на MNIST без препроцессинга тест-ошибка всего 1,2%
- SGD сходится к решению с минимальной нормой.

Заключения

- Эффективная емкость успешных архитектур способна покрыть обучающую дату.
- Эти модели достаточны чтобы запомнить ее.
- Традиционные способы не способны объяснить обобщающую способность больших нейросетей.
- Оптимизация остается такой же легкой, даже если итоговая модель плохо обобщает.

Список используемой литературы

- https://arxiv.org/pdf/1611.03530.pdf

Список вопросов

- 1)Опишите, как SGD играет роль регуляризатора в линейной модели (описать шаги без подробностей).
- 2)Какую роль играют регуляризаторы в обобщающей способности сетей?
- 3)Напишите формулу Радемахеровской сложности, какой у нее тривиальный верхний предел для задачи бинарной классификации?