# UNDERSTANDING DEEP LEARNING REQUIRES RETHINKING GENERALIZATION

Chiyuan Zhang, Samy Bengio, Moritz Hardt,
Benjamin Recht, Oriol Vinyals

(26 Feb 2017)

Жилкина Ксения

Факультет Компьютерных Наук, ПМИ
НИУ ВШЭ

31 ноября 2019

# Оглавление

# Оглавление

# Generalization error

**Generalization error** - разница между ошибкой на обучающей и тестовой выборках, то есть насколько хорошо модель умеет "обобщать"

# Постановка проблематики

- Традиционные методы:
  умение "обобщать" – свойство семейства моделей или результат применения регуляризации на этапе обучения

Что отличает хорошо обобщающие модели от моделей с большой ошибкой обобщения?

# Оглавление

# Условия эксперимента

- Датасеты:

  **CIFAR10** (60000 32x32 colour images in 10 classes, 3 color channels, with 6000 images per class; 50000 training images and 10000 test images)

  **ImageNet ILSVRC 2012** (1000 classes; 1.2mln training, 50000 validation images; 299x299 images with 3 color channels)
- Обучение: SGD без тюнинга гиперпараметров

# Условия эксперимента

**CIFAR10:**

- Модели: упрощенная Inception (Szegedy et al., 2016), Alexnet(Krizhevsky et al., 2012), а также стандартный multi-layer perceptrons с разным кол-вом слоев
- Функция активации: везде одна, ReLU
- Обучение: SGD, momentum parameter = 0.9

**ImageNet:**

- Модели: Inception V3 (Szegedy et al., 2016)
- Предобработка данных и настройки эксперимента: переиспользованы из пакета TENSORFLOW
- Реализация эксперимента: distributed asynchronized SGD system with 50 workers

# Подробнее о модели

## Inception

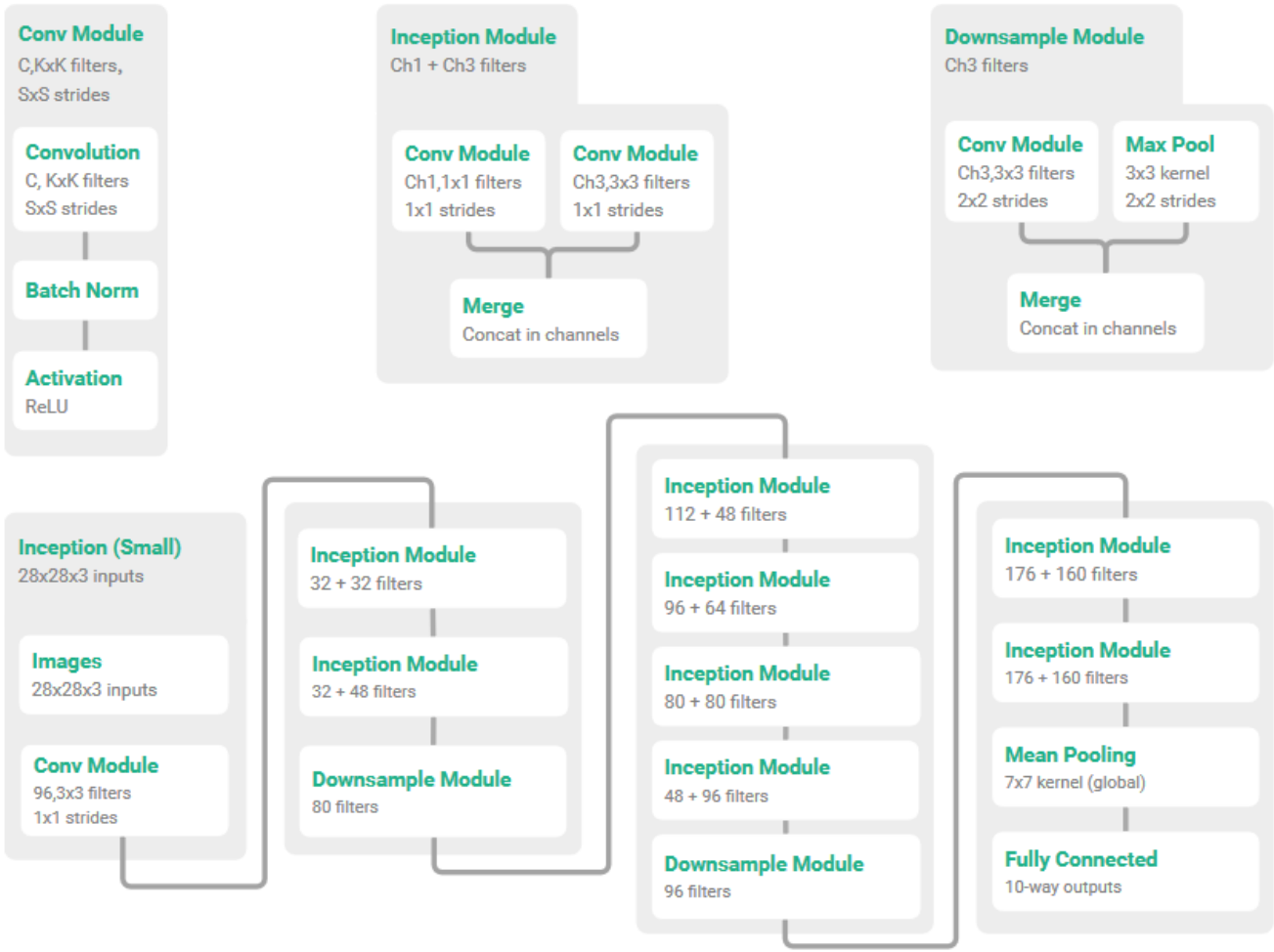
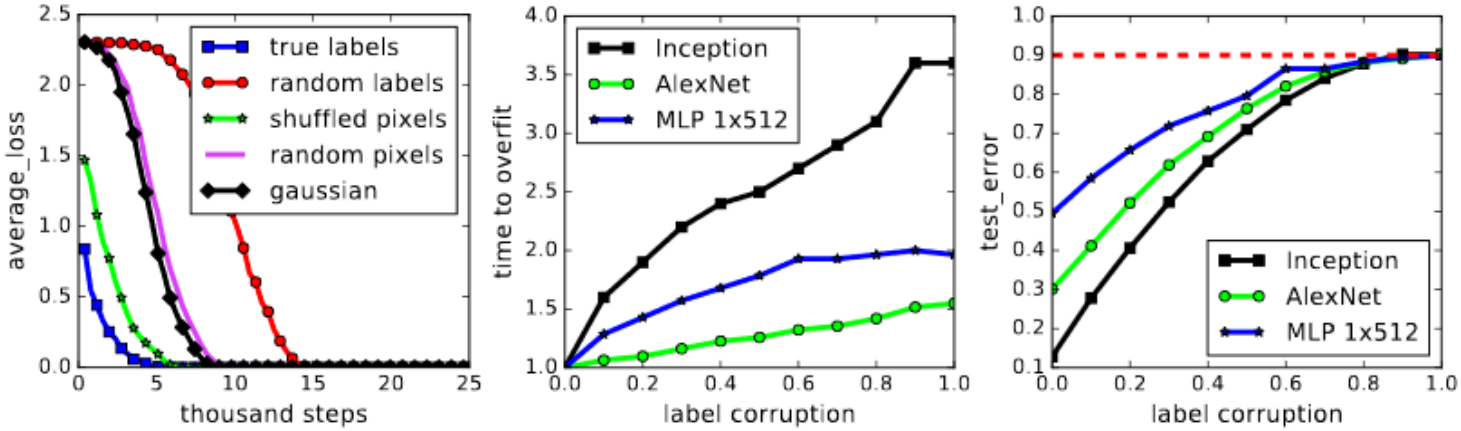



Figure 3: The small Inception model adapted for the CIFAR10 dataset. On the left we show the `Conv` module, the `Inception` module and the `Downsample` module, which are used to construct the Inception architecture on the right.

# Ход эксперимента

- Данные **без модификаций**
- **Частично модифицированные ответы**: с вер-ю p ответ каждого изображения независимо менялся на случайный класс из возможных, выбираясь равновероятно
- **Полностью случайные ответы**
- **Перестановка пикселей**: выбиралась одна перестановка и применялась ко всем изображений из train и test выборок
- **Случайные перестановки пикселей**: перестановка своя для каждого изображения
- **Семлирование всех пикселей из нормального распределения** с матожиданием и дивергенцией оригинальных данных

# Результаты эксперимента



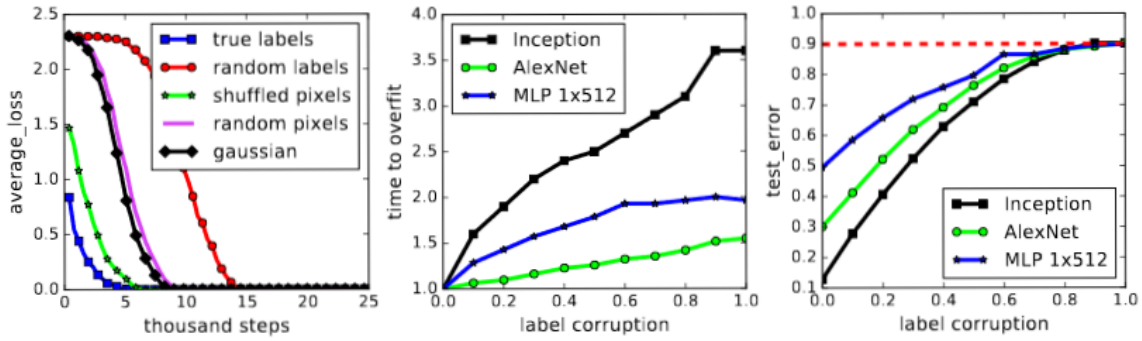

(a) learning curves (b) convergence slowdown (c) generalization error growth

Figure 1: Fitting random labels and random pixels on CIFAR10. (a) shows the training loss of various experiment settings decaying with the training steps. (b) shows the relative convergence time with different label corruption ratio. (c) shows the test error (also the generalization error since training error is 0) under different label corruptions.

## Почему результаты интересны

- Не нужно даже подстраивать изменение learning rate
- Быстрая сходимость на всех данных (хотя, очевидно, сначала темп замедляется, но градиент дает большие шаги)
- Идеальное переобучение! Функция хорошо описана

# И еще результаты




(a) learning curves (b) convergence slowdown (c) generalization error growth

- Random pixels и Gaussian сходятся медленнее random labels (измененные изображения легче отделить друг от друга)
- Даже на данных ImageNet accuracy впечатляющая

| data aug | dropout | weight decay | top-1 train | top-5 train | top-1 test | top-5 test |
|---|---|---|---|---|---|---|
| ImageNet 1000 classes with the original labels | | | | | | |
| yes | yes | yes | 92.18 | 99.21 | 77.84 | 93.92 |
| yes | no | no | 92.33 | 99.17 | 72.95 | 90.43 |
| no | no | yes | 90.60 | 100.0 | 67.18 (72.57) | 86.44 (91.31) |
| no | no | no | 99.53 | 100.0 | 59.80 (63.16) | 80.38 (84.49) |
| Alexnet (Krizhevsky et al., 2012) | | | - | - | - | 83.6 |
| ImageNet 1000 classes with random labels | | | | | | |
| no | yes | yes | 91.18 | 97.95 | 0.09 | 0.49 |
| no | no | yes | 87.81 | 96.15 | 0.12 | 0.50 |
| no | no | no | 95.20 | 99.14 | 0.11 | 0.56 |

Table 2 shows the performance on Imagenet with true labels and random labels, respectively.

# Оглавление

# Rademacher complexity и VC-dimension

- Широко используется, гибко оценивает сложность класса гипотиз
- Формула для класса гипотез $\mathcal{H}$ на данных $x_1, ...x_n$ (где $\sigma_1, ...\sigma_n$ =-1, +1 - равновероятно):

$$\hat{\mathfrak{R}}_n(\mathcal{H}) = \mathbb{E}_\sigma \left[ \sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i h(x_i) \right]$$

- Измеряет возможность $\mathcal{H}$ описать случайные бинарные лейблы
- Поскольку эксперименты говорят, что случайные лейблы описываются семейством идеально, ожидаем верхнюю границу 1. А она тривиальна из формулы и ничего не дает
- С VC-dimension аналогично

# Uniform stability

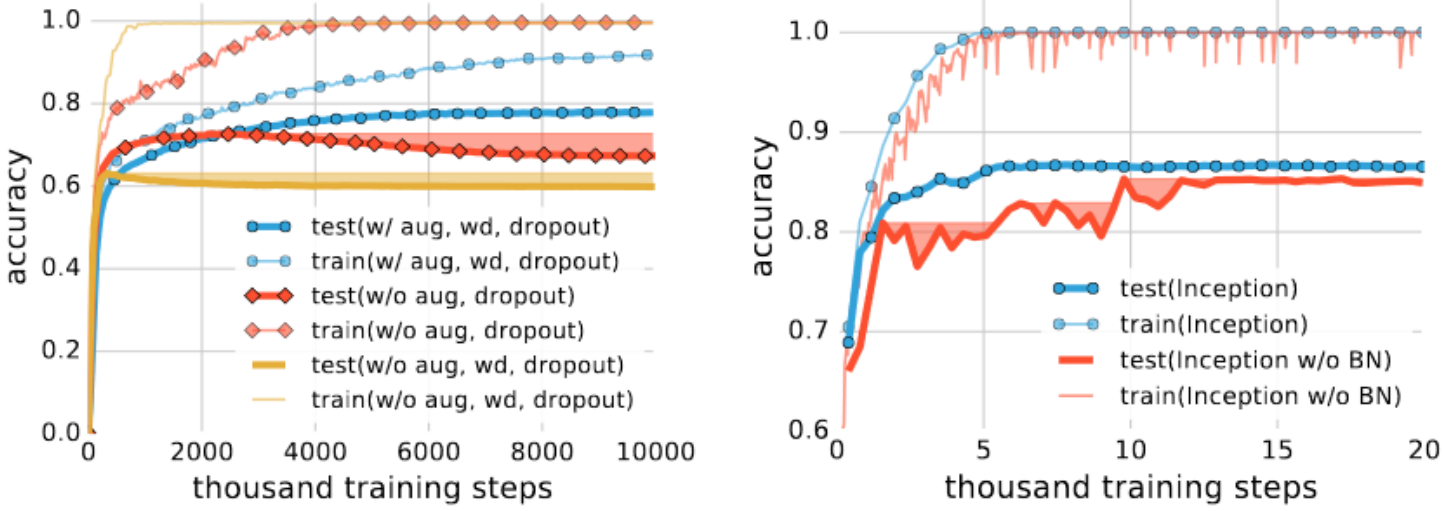
- Концентрирует внимание на алгоритме обучения
- Измеряет, насколько алгоритм чувствителен к замене одного примера из данных для обучения
- Мера не обращает внимания на особенности данных, поэтому не подходит для описания результатов эксперимента

# Оглавление

# Явная регуляризация

- Аугментация данных: случайные обрезания, экспозиционные и цветовые настройки
- Weight decay: эквивалентен l-2 регуляризации на весах
- Dropout: вероятностная маска на выходах слоя, в эксперимерименте использована только для Inception V3 на ImageNet

# Явная регуляризация



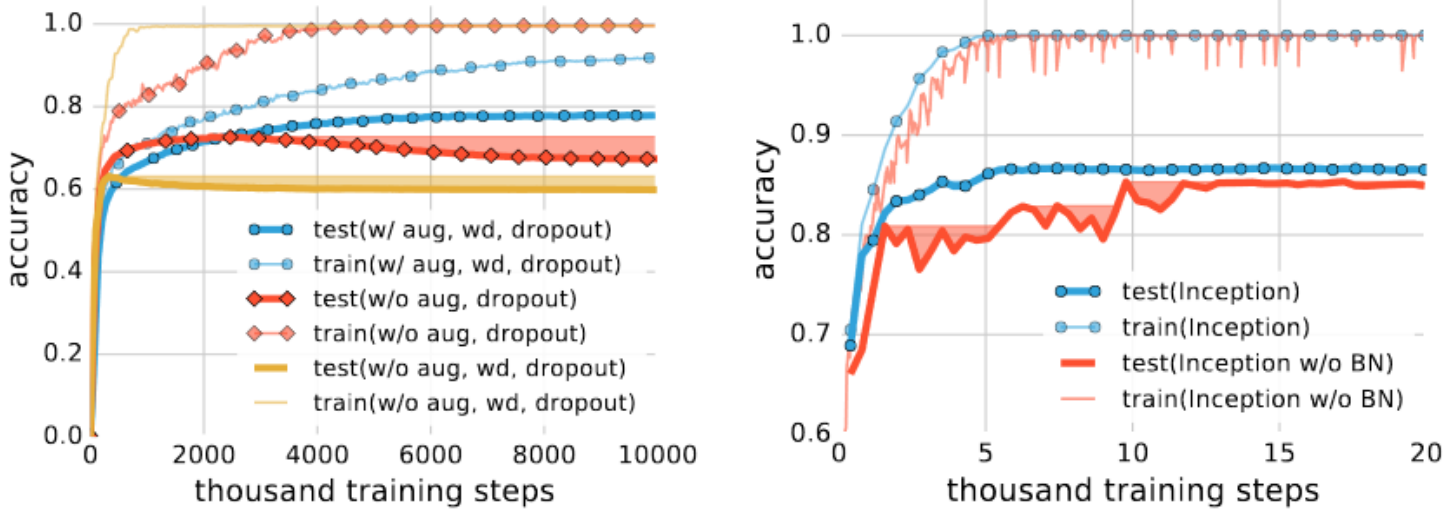

(a) Inception on ImageNet

(b) Inception on CIFAR10

Figure 2: Effects of implicit regularizers on generalization performance. aug is data augmentation, wd is weight decay, BN is batch normalization. The shaded areas are the cumulative best test accuracy, as an indicator of potential performance gain of early stopping. (a) early stopping could potentially improve generalization when other regularizers are absent. (b) early stopping is not necessarily helpful on CIFAR10, but batch normalization stablize the training process and improves generalization.

- Регуляризация действительно несколько увеличивает качество
- Но даже при отключении регуляризации модель показывает себя хорошо
- Авторы говорят о том, что большее значение играет именно подбор модели

# Неявная регуляризация




(a) Inception on ImageNet

(b) Inception on CIFAR10

Figure 2: Effects of implicit regularizers on generalization performance. aug is data augmentation, wd is weight decay, BN is batch normalization. The shaded areas are the cumulative best test accuracy, as an indicator of potential performance gain of early stopping. (a) early stopping could potentially improve generalization when other regularizers are absent. (b) early stopping is not necessarily helpful on CIFAR10, but batch normalization stablize the training process and improves generalization.

- Early stopping
- Batch normalization. Inception использует BatchNorm модули, поэтом создадим Inception без этих модулей

# Результаты на конечных наборах данных

**Сколь простая модель может описать функцию данных с рандомизированными ответами?**

**Теорема:**

- 2-слойная ReLU сеть с 2*n + d параметрами может описать любую функцию ответов на любых n d-мерных сэмплах (прошлый результат статьи 2014г - min размер O(dn))

# Выводы

- Нейронные сети легко подгоняются под функцию с рандомизированными ответами на сэмплах
- Явная регуляризация **может** улучшить обобщение, но она **не обязательна** для улучшения обобщения и **не является достаточной** для улучшения обобщения
- Существует 2-слойная ReLU сеть с 2*n + d параметрами, которая может описать любую функцию ответов на любых n d-мерных сэмплах