

Основы практического использования нейронных сетей.

Лекция 5. Эффективность НС.

Дмитрий Буряк.
к.ф.-м.н.
dyb04@yandex.ru

Эффективность обучения НС

Основные причины низкой эффективности обучения НС

- ☐ Низкая эффективность = большая ошибка на тестовых данных.
- ☐ Проблемы с данными;
- ☐ Несоответствие архитектуры НС сложности задачи;
- ☐ Неоптимальные значения гиперпараметров;
- ☐ Переобучение;
- ☐ Ошибки в реализации.

Обозначения

- ❑ S_{train} – обучающая выборка; S_{test} – тестовая выборка;
- ❑ E_{train} – ошибка на обучающей выборке, E_{test} – ошибка на тестовой выборке, E_{goal} – целевое значение ошибки.
- ❑ $E_{test} > E_{goal}$

Анализ ошибки на обучающей выборке

☐ $E_{train} > E_{goal}$

- ☐ Увеличить размер НС;
- ☐ Улучшить алгоритм обучения
- ☐ Оптимизировать значения гиперпараметров алгоритма обучения
- ☐ Анализ качества исходных данных
 - низкое значение сигнал-шум;
 - ошибки алгоритма предобработки;
 - недостоверные референсные значения;
 - несбалансированная выборка.

Анализ ошибки на тестовой выборке

❑ $E_{train} < E_{goal}$ и $E_{test} > E_{goal}$

❑ Увеличить размер S_{train} ;

❑ Уменьшить размер НС;

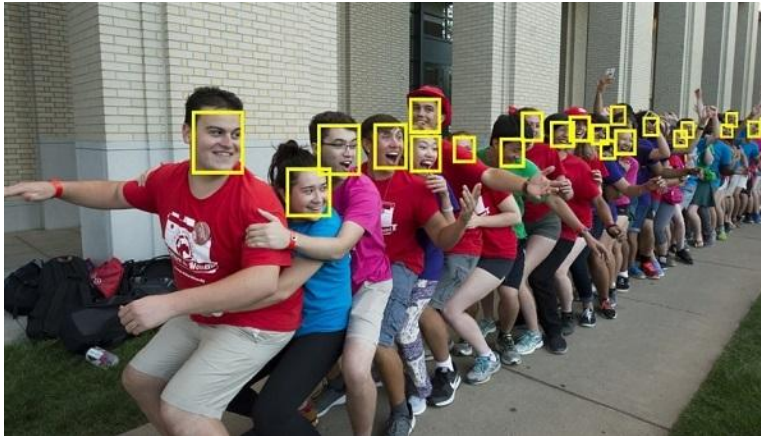
❑ Оптимизировать значения гиперпараметров НС (регуляризация);

❑ Подбор алгоритма обучения;

❑ Несоответствие S_{train} и S_{test} .

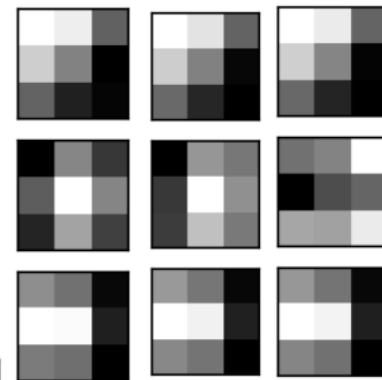
Анализ вычислений в НС

- ❑ Визуализация результатов:
 - согласованность статистических показателей и практики применения
- ❑ Визуализация, анализ результатов с наибольшими ошибками;



Анализ вычислений в НС (2)

- ☐ Обучение на подвыборке меньшего размера;
- ☐ Анализ внутренних состояний сети:
 - матрица весов/фильтры сверточного слоя;
 - карты признаков;
 - гистограмма выходов нейронов;
- ☐ Deconvnet



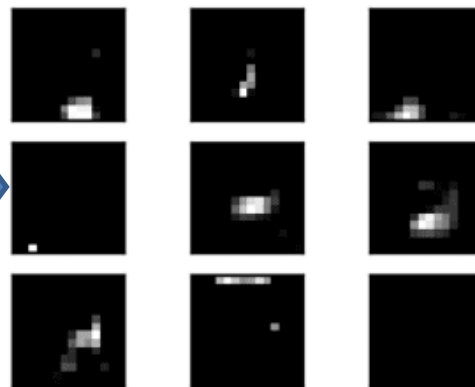
Фильтры
первого слоя VGG16



Входное изображение
VGG16



Примеры карт признаков
1-го слоя VGG16



Примеры карт признаков
внутренних VGG16

Стратегии повышения эффективности НС

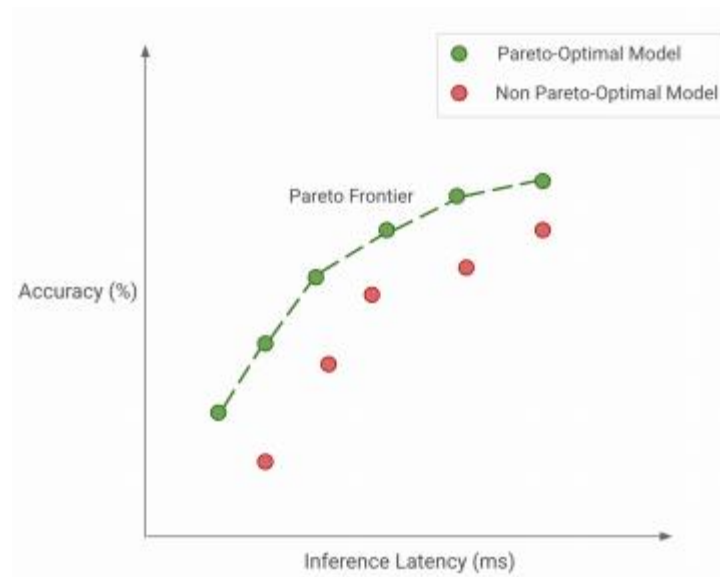
Эффективность НС

□ Эффективность обученной НС

- Размер сети
- Латентность
- Число MAC/FLOP

□ Эффективность проведения обучения

- Размер сети
- Точность



Технологии повышения эффективности НС

Areas

Compression
Techniques

Learning
Techniques

Automation

Efficient
Architectures

Description

- уменьшение размера НС
- удаление связей, нейронов, ...
- уменьшение разрядности представления НС

- эффективные алгоритмы обучения
- устойчивые функции потерь
- дистилляция

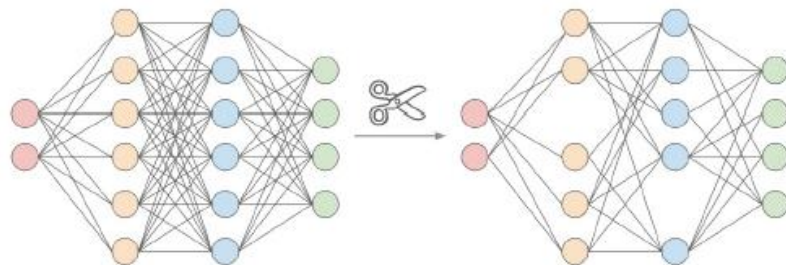
- автоматиз-ный поиск значений гиперпараметров
- подбор архитектуры

- разработка и применение эффективных архитектур

- библиотеки разработки и реализации НС
- высокопроизводительные вычислительные платформы

Прореживание (pruning) НС

- Удаление элементов архитектуры НС
 - Связи, нейроны, фильтры, ...



Algorithm 1: Standard Network Pruning with Fine-Tuning

Data: Pre-trained dense network with weights W , inputs X , number of pruning rounds N , fraction of parameters to prune per round p .

Result: Pruned network with weights W' .

```
1  $W' \leftarrow W$ ;  
2 for  $i \leftarrow 1$  to  $N$  do  
3    $S \leftarrow \text{compute\_saliency\_scores}(W')$ ;  
4    $W' \leftarrow W' - \text{select\_min\_k}(S, \frac{|W'|}{p})$ ;  
5    $W' \leftarrow \text{fine\_tune}(X, W')$   
6 end  
7 return  $W'$ 
```

Стандартный алгоритм прореживания обученной НС

Особенности алгоритмов прореживания НС

- ☐ Критерий выбора элементов для прореживания
 - оценка влияния на функцию потерь
 - абсолютная величина, вторая производная ...
- ☐ Элементы НС для удаления
- ☐ Распределение доли удаляемых элементов по НС
- ☐ Расписание проведения прореживания
- ☐ Возможность восстановления связей
- ☐ «Гипотеза о лотерейном билете»
 - В любой большой НС существует подсеть, которая может быть обучена с такой же эффективностью

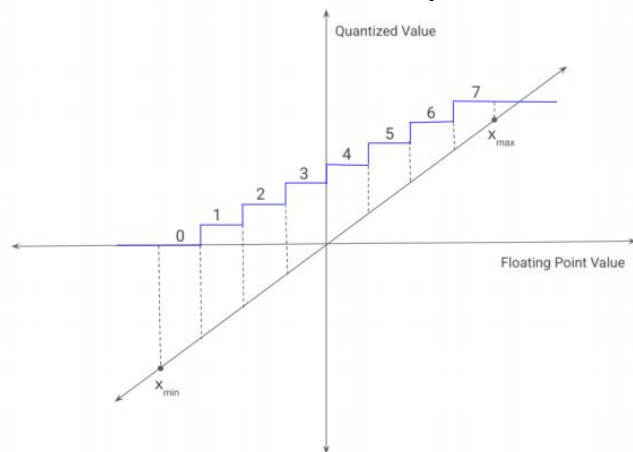
Квантизация

- ❑ Уменьшение разрядности представления весов, значений функции активации.
- ❑ Сокращение размера занимаемой памяти.
- ❑ Возможность реализации на специальных вычислительных платформах с сохранением высокой производительности

Квантизация весов

□ Квантизация весов обученной сети

- Деквантизации при проведении вычислений
- Сохранение точности при 8 битном представлении
- Эксперименты с 4, 3, 2 и 1 битными сетями
- Вычислительные затраты на деквантизацию



Квантизация вещественных весов
в значения с фиксированной точкой

Algorithm 2: Quantizing a given weight matrix X

Data: Floating-point tensor to compress X , number of precision bits b for the fixed-point representation.

Result: Quantized tensor X_q .

- 1 $X_{min}, X_{max} \leftarrow \min(X, 0), \max(X, 0)$;
- 2 $X \leftarrow \text{clamp}(X, X_{min}, X_{max})$;
- 3 $s \leftarrow \frac{X_{max} - X_{min}}{2^b - 1}$;
- 4 $z \leftarrow \text{round}\left(x_{q_{min}} - \frac{X_{min}}{s}\right)$;
- 5 $X_q \leftarrow \text{round}\left(\frac{X}{s}\right) + z$;
- 6 return X_q ;

Algorithm 3: Dequantizing a given fixed-point weight matrix X_q

Data: Fixed-point matrix to dequantize X_q , along with the scale s , and zero-point z values which were calculated during quantization.

Result: Dequantized floating-point weight matrix $\hat{X}$.

- 1 $\hat{X} \leftarrow s(X_q - z)$;
- 2 return $\hat{X}$;

Алгоритмы квантизации и деквантизации

Квантизация весов во время обучения

❑ Недостатки статической квантизации

- Единичные вылеты в значениях весов
- Разные распределения весов в интервале квантизации

❑ Quantization-Aware-Training (QAT)

- Симулирование квантизации во время обучения
- Вычисление значения функции потерь после проведения квантизации

$$\begin{aligned}\hat{X} &= \text{FakeQuant}(X) \\ &= \text{Dequantize}(\text{Quantize}(X)) \\ &= s \left(\left(\text{round} \left(\frac{\text{clamp}(X, x_{\min}, x_{\max})}{s} \right) + z \right) - z \right) \\ &= s \left(\text{round} \left(\frac{\text{clamp}(X, x_{\min}, x_{\max})}{s} \right) \right)\end{aligned}$$

Model Architecture	Quantization Type	Top-1 Accuracy	Size (MB)
MobileNet v2-1.0 (224)	Baseline	71.9%	14
	Post-Training Quantization	63.7%	3.6
	Quantization-Aware Training	70.9%	3.6

Сравнение статической квантизации и QAT

Квантизация активаций

- ❑ Реализация на специальных вычислительных платформах
 - Поддержка вычислений с фиксированной точностью
 - Увеличение скорости вычислений
 - Требуется реализация вычислений с фиксированной точностью

Вопросы

- ❑ Что может быть причиной, когда ошибка на обучающей выборке превосходит целевое значение ошибки ?
- ❑ Какие действия следует предпринимать, чтобы добиться уменьшения ошибки на тестовой выборке?
- ❑ Что такое прунинг?
- ❑ Какие бывают виды квантизации?