mathematics & mechanics

Классификация изображений бинарными нейронными сетями с расширенной информацией

Чанчиков Антон Юрьевич. Бакалавриат, 4 курс, группа 19122

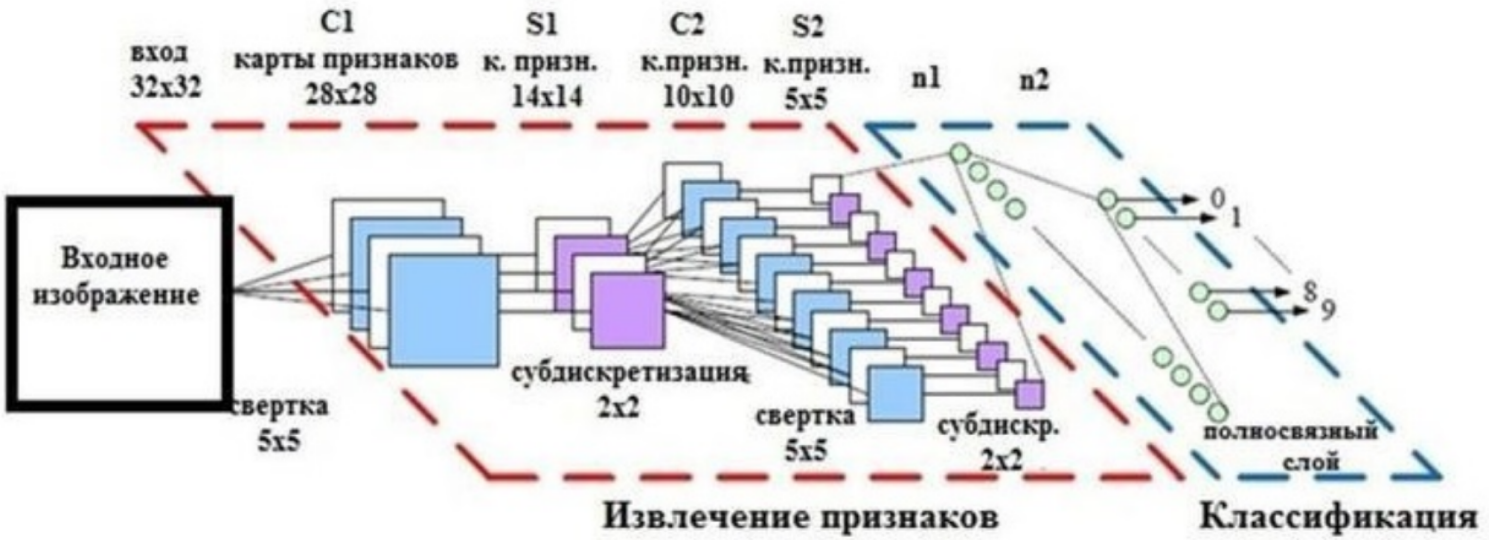
Научные руководители:

- Тарков Михаил Сергеевич. Профессор кафедры Вычислительных систем ММФ НГУ, к.т.н., доцент
- Городничев Максим Александрович. Старший преподаватель кафедры Вычислительных систем ММФ НГУ

Содержание

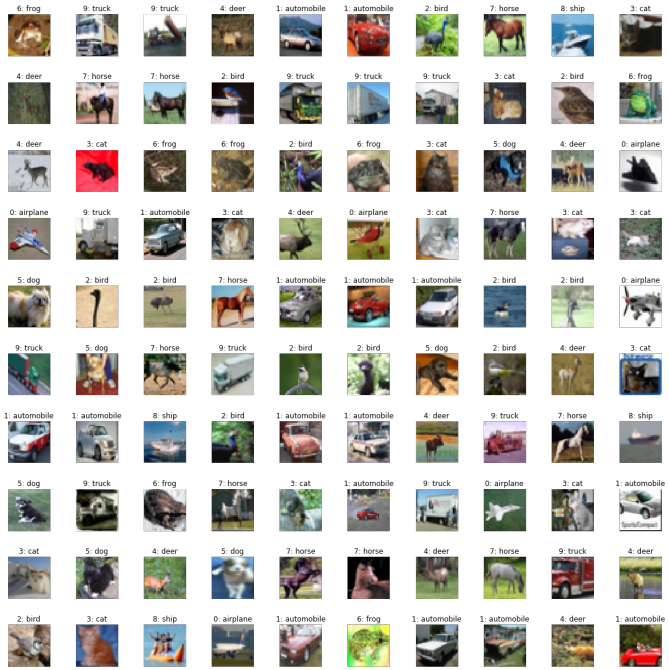
Введение

Сверточные нейронные сети



Задача классификации изображений

Постановка задачи: определить, к какому классу наиболее вероятно принадлежит входное изображение.



(a) CIFAR-10. Трехканальные цветные изображения, разрешение 32x32

Бинарные нейронные сети

Идея:

$$b_w = \text{sign}(w) = \begin{cases} +1, \text{ если } w \geq 0 \\ -1, \text{ иначе} \end{cases}, \quad b_x = \text{sign}(x) = \begin{cases} +1, \text{ если } x \geq 0 \\ -1, \text{ иначе} \end{cases}$$

где w и x - веса и активации полноточной нейронной сети,
b_w и b_x - бинарной нейронной сети.

Тогда выход сети представляет из себя побитово примененные операции XNOR и POPCOUNT к множествам B_w и B_x

$$Y = (B_w \oplus B_x) \cdot \alpha$$

Бинарные нейронные сети

Преимущества

- Уменьшение нагрзуки на вычисления
- Снижение веса модели

Недостатки

- Большие потери информации при бинаризации

Цель и задачи

Цель и задачи

Цель - разработка и сравнительное исследование методов оптимизации бинарных нейронных сетей с расширенной информацией.

Для достижения указанной цели были поставлены следующие задачи:

- Выбрать архитектуры сверточных нейронных сетей для классификации изображений.
- Провести их бинаризацию и улучшение.
- Оценить качество, время работы и количество выполняемых операций в процессе классификации изображений новыми бинарными сетями.
- Сравнить полученные результаты для разных сверточных сетей.

Обзор

Связанные работы

С 2016 по сентябрь 2022 г. - не менее 239 фундаментальных разработок в теме бинарных нейронных сетей.

- ReActNet предлагает для бинаризации использовать функции с обучаемыми порогами по входным каналам, что позволяет подавать на вход сети подавать больше различной информации для анализа.

$$RSign(x^i) = \begin{cases} +1, \text{ если } x^i \geq \beta^i \\ -1, \text{ если } x^i < \beta^i \end{cases}$$

- IR-Net перед бинаризацией весов предлагает их сбалансировать и стандартизовать для уменьшения эффекта потери информации.

$$w_{std} = \frac{\widehat{w}}{\sigma(\widehat{w})}, \ \widehat{w} = w - \bar{w}$$

БНС с улучшенной информацией

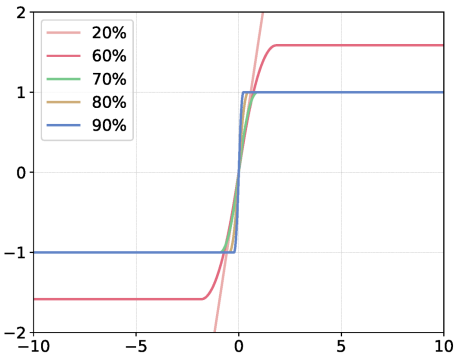
- Использование K функций RSign

$$b_x^{i,k} = RSign^k(x^i) = \begin{cases} +1, \text{ если } x^i \geq \beta^{i,k} \\ -1, \text{ если } x^i < \beta^{i,k} \end{cases}$$

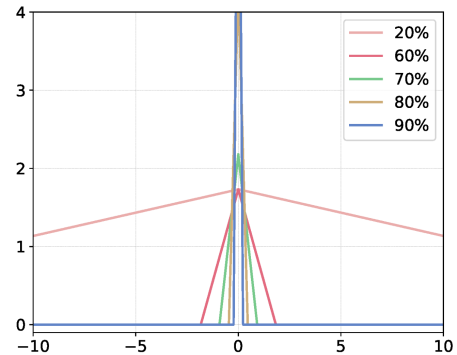
- Для обработки пакетов вводится обобщенная двочиная свертка

$$Y^k = BConv(B_w, B_x^k) = (B_w \oplus B_x^k) \cdot \alpha$$

- Стандартизация весов, как в IR-Net, и использование особой функции IEE для бинаризации весов. Особенность в том, что она постепенно в процессе обучения аппроксимирует функцию знака, придавая градиентам весов ненулевые значения

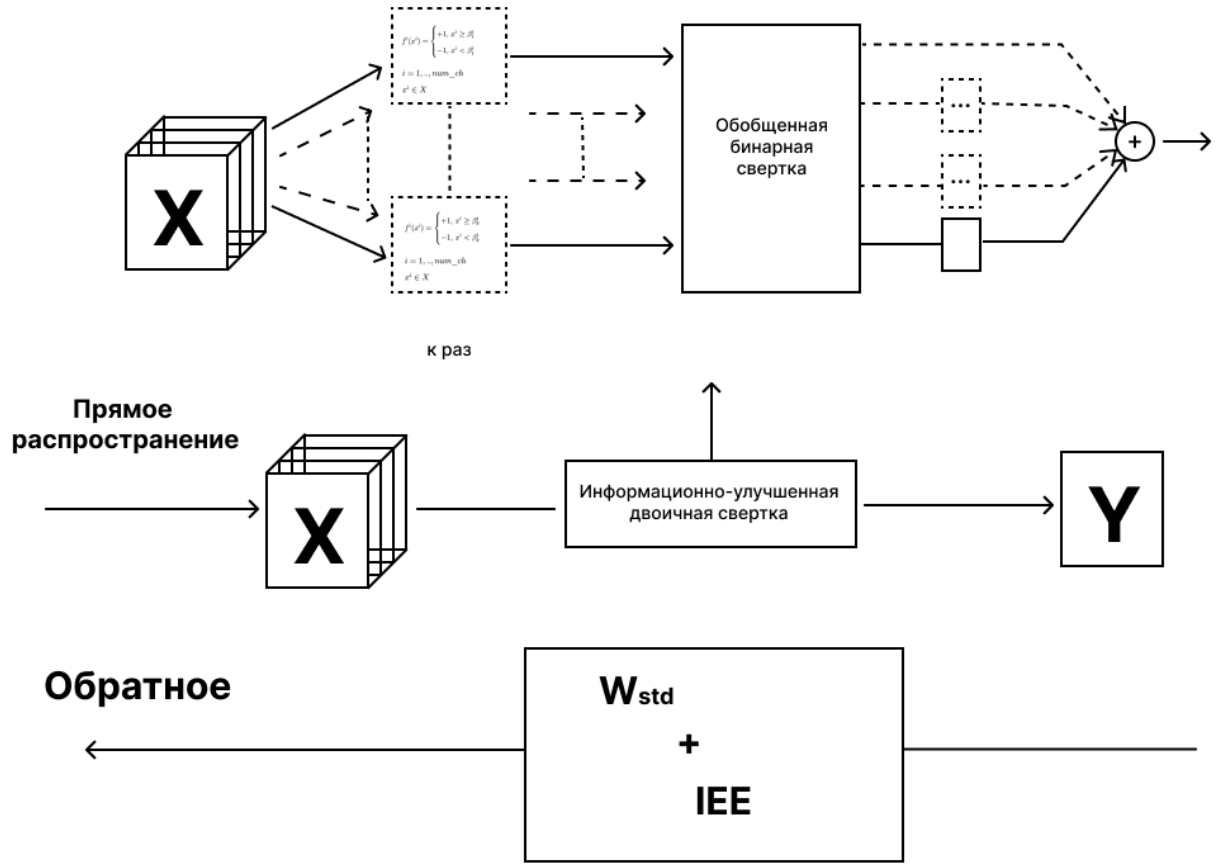


(a) $F(x)$



(b) $F'(x)$

БНС с улучшенной информацией



Структура БНС с улучшенной информацией

Исследовательская основа

Выбранные архитектуры

ResNet

- Высокая точность обработки изобрежний
- Требует больших вычислительных мощностей
- Глубокая сеть с большим количеством параметров

MobileNetV2

- Высокая точность обработки изобрежний
- Эффективное потребление вычислительных ресурсов
- Малый вес

Методы оптимизации

Методы оптимизации

- Регуляризация - методика ограничения модели для улучшения ее обобщающей способности
- Аугментация - расширение исходного набора данных, путем применения к изображению некоторых операций, таких как поворот, сдвиг, инвертация каналов и другие
- Дистилляция - тактика обучения небольшой модели используя знания другой, более масштабной сети
- Обрезание - удаление незначимых весов сети для облегчения ее веса при небольших потерях точности

Разработка фреймворка

Разработка фреймворка

Для реализации большого количества экспериментов с использованием различных моделей и методов было принято решение создать исследовательский фреймворк для более легкой, гибкой и быстрой работы.

Он включает в себя:

- Реализацию всех методов, вариантов моделей и других опциональных вещей, описанных ранее
- Возможность менять гиперпараметры, модели и наборы данных для исследований
- Возможность запускать несколько экспериментов разом из некоторого фиксированного пространства методов для мгновенного сравнения результатов

Эксперименты

Результаты ResNet18

Значение K	Точность	Время обучения с/эпоха (min, max)
Небинарная	0.89	25.9; 28.4
1	0.5958	37.9; 40.67
2	0.736	61.6; 65.1
3	**0.757**	**83.2; 90.9**
4	0.7292	105; 123.7
5	0.744	104.3; 107.4
6	0.737	122.2; 128.4

Тип сети	FLOPS	GPU Memory usage	Вес
Небинарная	$2.7 \cdot 10^9$	6.1Gb	42.26Mb
Бинарная ($K = 3$)	$2.43 \cdot 10^8$	4.8Gb	4.64Mb

Тип сети	W/o	+WD	+LS	+RA	-LS	+LS+DP
Небинарная	0.89	0.8573	0.8503	0.9201	0.9115	0.9150
Бинарная	0.757	0.7737	0.7641	0.8488	0.8566	0.8511

Результаты MobileNetV2

Значение K	Точность	Время обучения с/эпоха (min, max)
Небинарная	0.8211	49.53 ; 52.18
1	0.2915	85.57; 89.15
2	**0.4684**	**99.76; 105.0**
3	0.4571	117.36; 124.01
4	0.4603	141.5; 148.91
5	0.4725	178.34; 185.28
6	0.4518	194.57; 199.13

Тип сети	FLOPS	GPU Memory usage	Вес
Небинарная	$3 \cdot 10^9$	3.2Gb	6.53Mb
Бинарная ($K = 2$)	$1.9 \cdot 10^8$	2.6Gb	0.87Mb

Модель	W/o	+WD	+LS	+RA	-LS	+LS+DP
MobNetV2	0.8124	0.8289	0.8275	0.9031	0.9043	0.905

Перенос знаний

Модель	W/o	+WD	+LS	+RA	-LS	+LS+DP
Бинарная (K=3)	0.8105	0.8356	0.8491	0.89	0.8747	0.8944

Таблица: Дистилляция знаний полноточной ResNet18 на бинарный аналог

Выводы

- Сильное расширение информации не всегда способно повысить качество модели, то есть из увеличения гиперпараметра K не следует рост точности.
- Показано, что применение методов регуляризации в основном улучшает точность сетей, однако для каждой нужно искать индивидуальный путь их применения.
- Для бинарных моделей крайне эффективно работает тактика обучения учитель-ученик, стратегия обучения способствует повышению качества ученика.
- Теоретические ожидания не всегда оправдываются на практике, что связано с эффективностью программной реализации.

Заключение

1. Разработан план экспериментального исследования с целью повышения точности бинарных нейронных сетей с расширенной информацией, предложены методы оптимизации сетей: поиск оптимального количества информации для расширения, комбинация методов регуляризации, которую нужно подбирать вручную для разных архитектур, перенос знаний, который показал себя перспективно для практического применения.

2. Разработана схема автоматизации исследования и реализован программный фреймворк для ее выполнения.

3. Проведены эксперименты по изучению поставленных вопросов и показано, что применение некоторых из предложенных методов позволяет повысить точность модели, например, расширение информации, методы регуляризации и расширение датасета для бинарной ResNet18.

Q&A