Кафедра системого программирования СПбГУ

Выпускная квалификационная работа

Разработка на ПЛИС нейронной сети, реализующей распознавание объектов интереса на изображениях

Автор:
Холод Николай Григорьевич, 646 гр.

Научный рукводитель:
ст. преп. Смирнов М. Н.

Рецензент:
Гл. науч. сотр. АО НПП АМЭ,
к.т.н. Крюков С. Н.

Введение

Сверточные нейронные сети показывают отличные результаты в задаче классификации изображений

Вычислительная сложность современных нейронных сетей высока

Использование GPU в малогабаритных встраиваемых системах невозможно

Альтернатива — ПЛИС

Цель работы

Выработка и апробация подхода к построению нейросетевого классификатора изображений, предназначенного для работы на ПЛИС.

Постановка задачи

- Изучить, какие типы нейронных сетей используются в существующих реализациях на ПЛИС
- Разработать архитектуру нейронной сети, допускающую эффективную реализацию процедуры распознавания
- Произвести обучение и подбор оптимальных гиперпараметров нейронной сети на основе имеющихся данных
- Реализовать модули, необходимые для запуска обученной нейронной сети на ПЛИС
- Произвести тестирование и измерение производительности реализованной сети

Сжатые нейронные сети

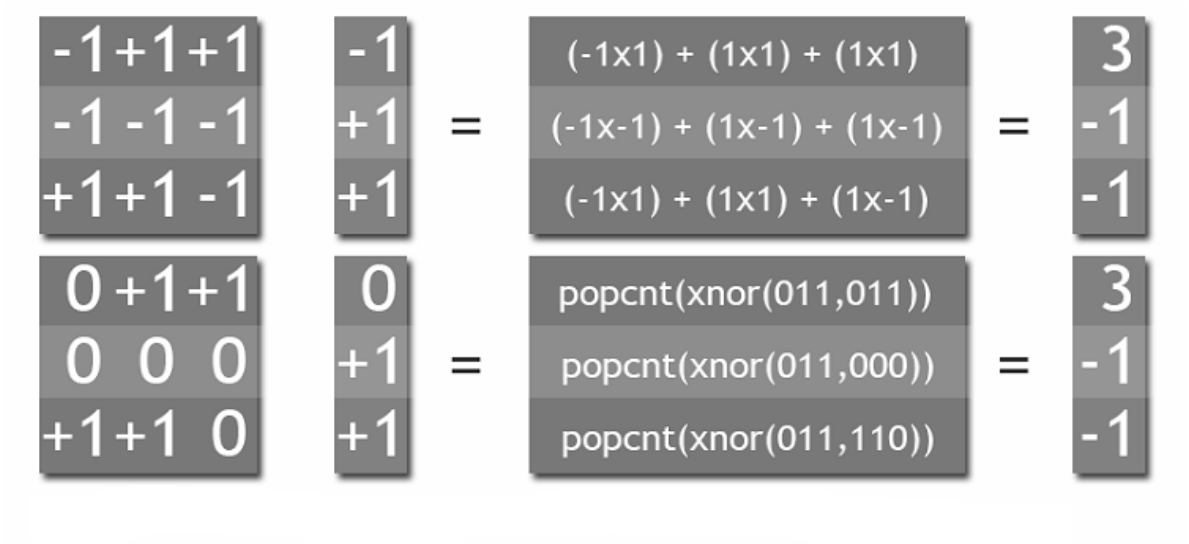
- Бинарные
  - Веса принимают значения $\{-1, 1\}$
  - Веса и активации принимают значения $\{-1, 1\}$
- Тернарные
  - Веса принимают значения $\{-1, 0, 1\}$
  - Веса и активации принимают значения $\{-1, 0, 1\}$
- Произвольной квантификации

  Веса и активации принимают целочисленные значения из фиксированного диапазона

Существующие библиотеки для запуска нейронных сетей на ПЛИС

- FINN
  - Бинарные нейронные сети
  - Для устройств производства Xilinx
- RebNet
  - бинарные веса, M-битные активации
  - Для устройств производства Xilinx
- BNN-PYNQ
  - Поддержка однобитных и двухбитных весов
  - Для устройств производства Xilinx
- OpenVino
  - 8-битные веса и активации
  - Поддерживается только Intel Arria 10

Преимущество бинарных нейронных сетей



Основные этапы разработки нейросетевого классификатора

1. Обучение
   - GPU
   - keras
2. Запуск на ПЛИС
   - Конвертация весов сети
   - Реализация модулей поддержки запуска

Обучение бинарных нейронных сетей

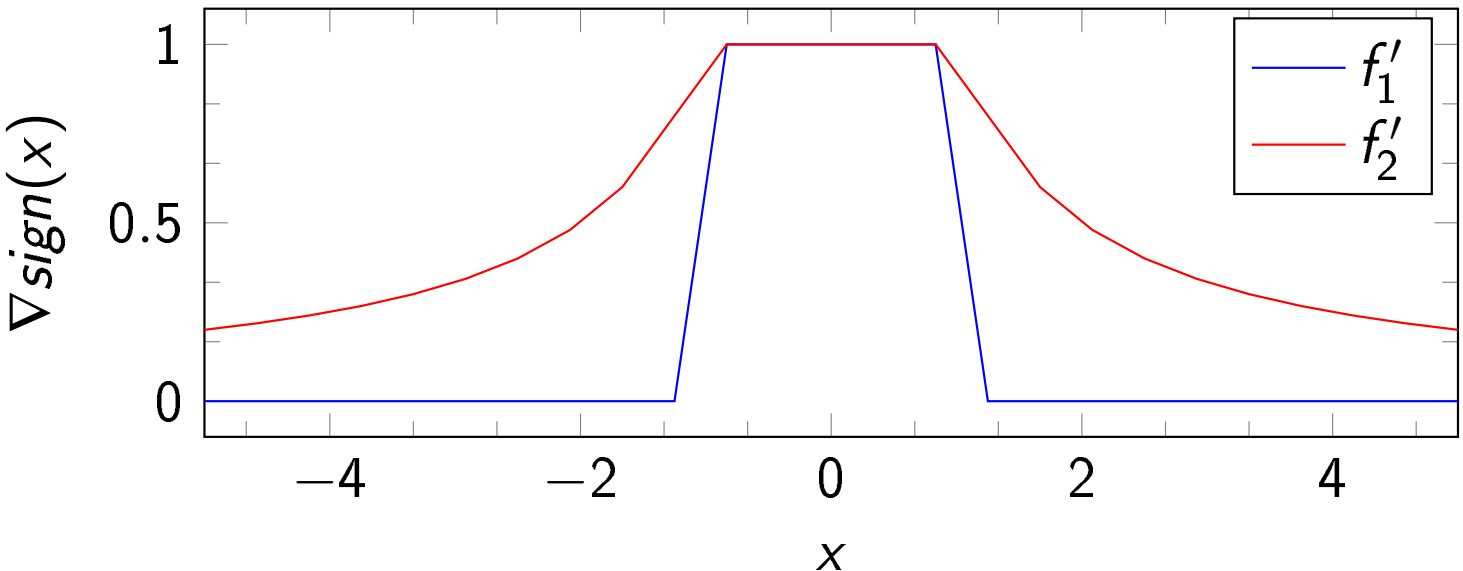
Straight through estimator

Обычно используемый вариант :

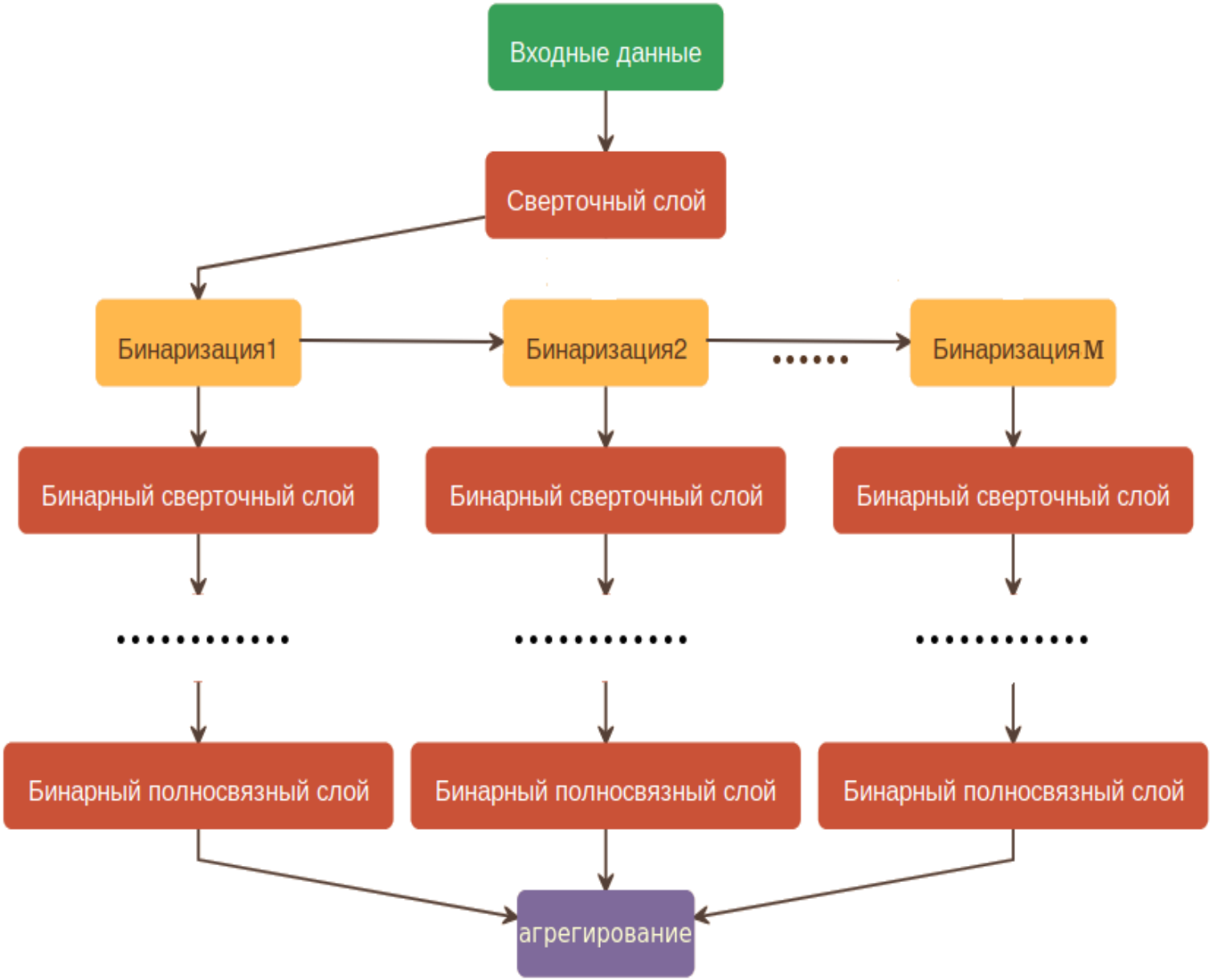
$$f_1'(x) = \begin{cases} 1, x \in [-1,1] \\ 0, |x| > 1 \end{cases} \tag{1}$$

Используемый в данной работе:

$$f_2'(x) = \begin{cases} 1, x \in [-1,1] \\ \frac{1}{x}, |x| > 1 \end{cases} \tag{2}$$



Предложенная архитектура



Агрегирование

- Усреднение
- Взвешенное суммирование
  - Обучение сети как в случае с усреднением
  - Логистическая регрессия на выходах сети

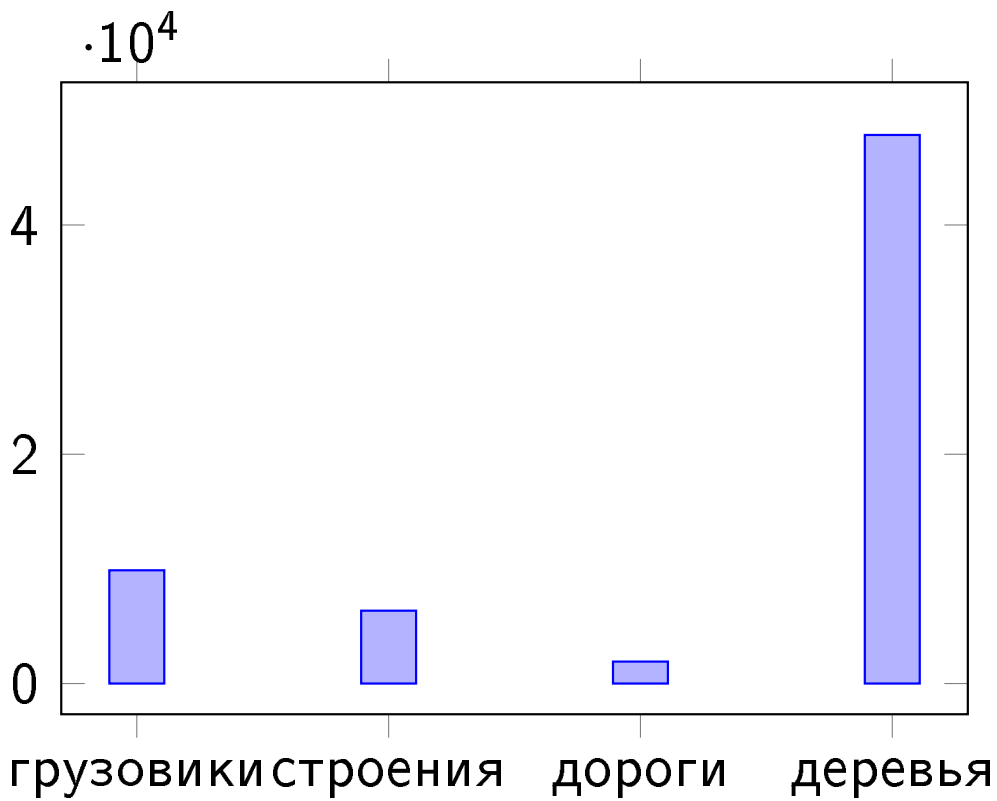
Сравнение с существующими методами

Набор данных Cifar10

Сеть	$M = 1$	$M = 2$	$M = 3$
FINN		80.1	
RebNet	80.59	85.94	86.98
Представленая архитектура	**84.37**	85.84	87.14
взвешенный ансамбль		**86.65**	**87.42**
Представленая архитектура , бинарная функция активации (1)	82.27	85.01	86.12
взвешенный ансамбль		85.62	86.53
Представленая архитектура, бинарные веса первого слоя	80.48	83.26	83.99
взвешенный ансамбль		83.87	84.54

Используемые данные

- 66000 изображений в градациях серого размером 48x48
- 4 класса
- 2 класса представляют собой объекты интереса(строения, грузовики)
- 2 класса фоновые (дороги и деревья)
- Выборка несбалансированна

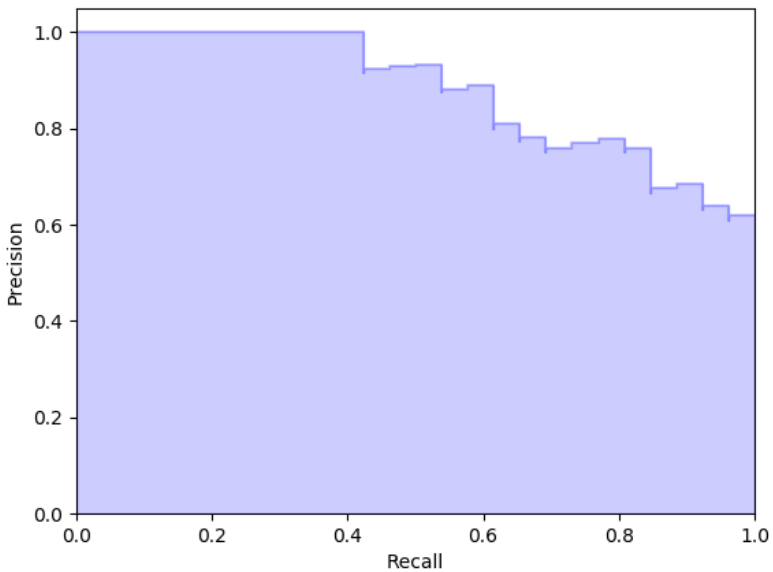


Precision–recall AUC

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

		Actual Values: Positive (1)	Actual Values: Negative (0)
Predicted Values	Positive (1)	TP	FP
Predicted Values	Negative (0)	FN	TN



Результаты

AUC

- multiclass — для оценки качества определения класса объекта
- binary — для оценки качества определения наличия объекта интереса

	Величина	A1	A2	A3	A4	A5	A6
M=1	auc binary	0.894	0.842	**0.907**	0.896	0.797	0.853
	auc multiclass	0.921	0.871	0.922	**0.932**	0.857	0.889
M=2	auc binary	0.901	0.875	**0.923**	0.919	0.818	0.874
	auc multiclass	0.929	0.908	0.938	**0.946**	0.884	0.912
взвеш.	auc binary	0.901	0.878	0.918	**0.948**	0.801	0.881
	auc multiclass	0.921	0.908	**0.932**	0.928	0.857	0.911
M=3	auc binary	0.916	0.884	**0.932**	0.929	0.832	0.892
	auc multiclass	0.936	0.914	0.943	**0.950**	0.891	0.923
взвеш.	auc binary	0.917	0.886	**0.931**	**0.931**	0.833	0.892
	auc multiclass	0.931	0.913	**0.951**	**0.951**	0.874	0.917

Выбранная архитектура

Сверточный слой 3x3x64
Многоуровневая бинаризация
Бинарный сверточный слой 3x3x64
Бинарная активация
Слой пуллинга
Бинарный сверточный слой 3x3x128
Бинарная активация
Бинарный сверточный слой 3x3x128
Бинарная активация
Слой пуллинга
Бинарный сверточный слой 3x3x256
Бинарная активация
Бинарный сверточный слой 3x3x256
Бинарная активация
Слой пуллинга
Бинарный полносвязный слой 512
Бинарная активация
Бинарный полносвязный слой 4

Intel HLS Compiler

Инструмент, генерирующий RTL описание из исходного кода на языке C

- Упрощается разработка
- Уменьшается порог вхождения для программистов

Реализация поддержки нейронной сети для ПЛИС

- Бинарный полносвязный слой + Бинарная активация
  - Модуль умножения битовой матрицы на битовый вектор с последующим сравнением с порогом
- Бинарный сверточный слой + Бинарная активация
  - Модуль скользящего окна
  - Модуль умножения битовой матрицы на битовый вектор с последующим сравнением с порогом
- Слой пуллинга
  - Модуль скользящего окна
  - Модуль пулинга
- Многоуровневая бинаризация
  1. Сравнение входных данных с порогом
  2. Замена входных данных на их модуль разности с порогом
- Вещественный сверточный слой

Тестирование

Для тестирования использовалась плата DE1-SoC с Cyclone V

1. Результаты классификации на ПЛИС сравнивались с результатами классификации на CPU
   Норма разницы между прогнозами $< 10^{-5}$
2. Пропускная способность составила около 900 изображений в секунду
   - В среднем требуется обрабатывать 3000 изображений в секунду

Результаты

- Проведен обзор типов нейронных сетей, используемых в существующих реализациях на ПЛИС
- Разработана архтектура бинарной нейронной сети, позволяющая незначительно увеличить точность распознавания по сравнению с точностью бинарных нейронных сетей с сопоставимой вычислительной сложностью
- Произведено обучение сети данной архитектуры на данных с датасета «АМЭ»
- Реализованы модули поддержки запуска бинарной нейронной сети на на ПЛИС
- Произведено тестирование нейронной сети на наборе данных АМЭ
- По промежуточным результатам работы представлена к публикуации статья в журнале «Известия ЮФУ ТЕХНИЧЕСКИЕ НАУКИ»