

# FPGA реализация нейронной сети прямого распространения для распознавания рукописных чисел

---

Е.А. Кривальцевич    М.И. Вашкевич

`krivalcevi4.egor@gmail.com, vashkevich@bsuir.by`

Белорусский государственный университет  
информатики и радиоэлектроники  
Кафедра электронных вычислительных средств



XIV Международная научная конференции  
«Информационные технологии и системы»  
Минск, Республика Беларусь



20 ноября, 2024

1. Прототипирование нейронных сетей на FPGA
2. Постановка задачи
3. Обучение нейронной сети
4. Аппаратная реализация нейронной сети
5. Использование PYNQ для прототипирования и тестирования нейронной сети
6. Описание эксперимента и результаты

# Введение

---

# Прототипирование нейронных сетей на FPGA

- Вычислительной платформой для обучения и эксплуатации нейросетевых моделей чаще всего выступают графические процессоры, которые содержат множество вычислительных ядер, способных обрабатывать потоки данных параллельно.
- FPGA (Field Programmable Gate Array) представляют собой реконфигурируемые вычислительные платформы, позволяющие реализовывать параллельно-поточные архитектуры НС.
- При реализации НС на базе FPGA появляется возможность использовать для представления параметров НС типов данных, обеспечивающих различную точность.

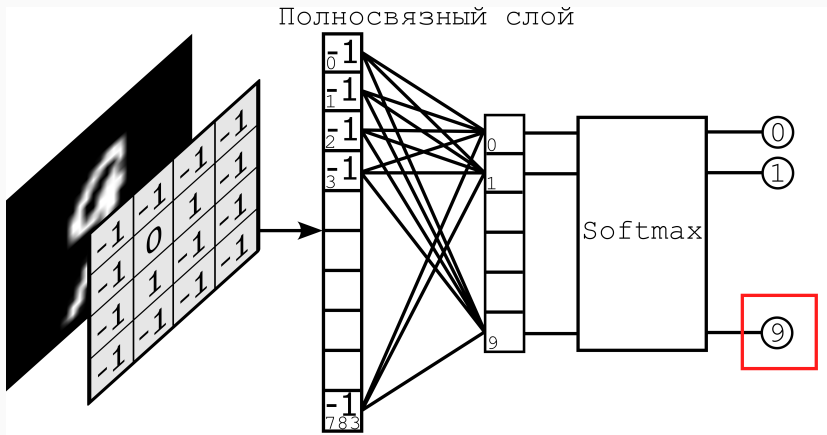
## Цель исследования

- Получить аппаратно реализованную НС прямого распространения для распознавания рукописных цифр
- Выяснить влияние разрядности представления весовых коэффициентов НС на точность определения цифр и аппаратные затраты
- Оценить наиболее оптимальную реализацию НС

# Обучение нейронной сети

---

# Архитектура нейронной сети

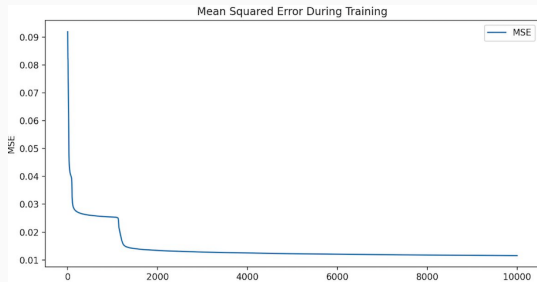


# Параметры для обучения

## Основные блоки

- Входные данные приводятся к диапазону  $[-1, 1]$  и устанавливается их среднеквадратическое отклонение(СКО) равным 0,5
- Оптимизация производилась с использованием метода стохастического градиентного спуска (SGD) (скорость обучения  $\eta = 3 \cdot 10^{-3}$ , число эпох – 10000, моментум  $\gamma = 0,9$ )

## График обучения ...





# Аппаратная реализация нейронной сети

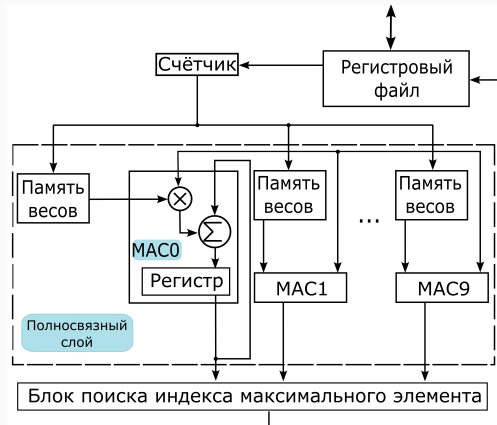
---

# Структурная схема IP-блока

## Основные блоки

- Регистровый файл
- Счётчик
- Полносвязный слой
- Блок поиска максимального индекса

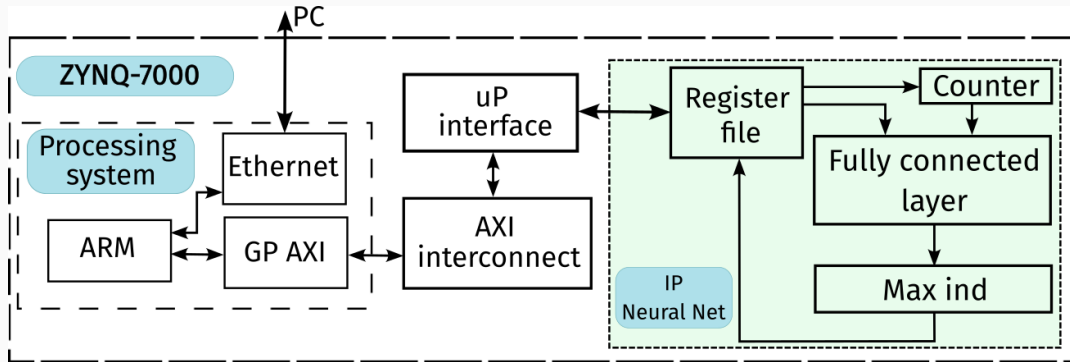
## Структурная схема IP-блока



# Использование PYNQ для прототипирования и тестирования нейронной сети

---

# Структурная схема проекта



- Подключение PL блока к PS осуществляется с помощью AXI4-Lite и uP интерфейсов.

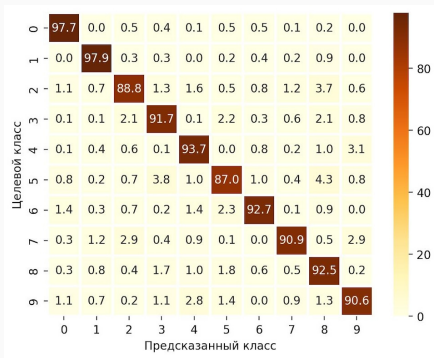
## Эксперимент и результаты

---

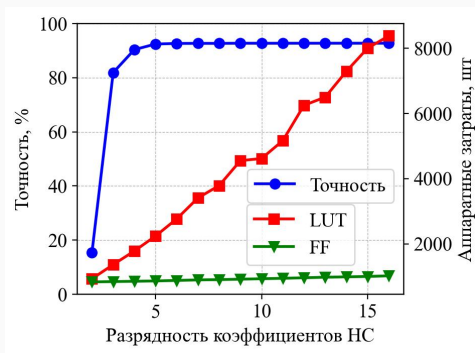
# Описание эксперимента

- Набор данных MNSIT (10 тыс. изображений рукописных цифр  $28 \times 28$ )
- Данные подаются последовательно из процессорной системы
- Результаты группируются в виде матриц спутывания
- Проведено 15 тестов с различными разрядностями весовых коэффициентов (от 2 до 16)
- Составлен график зависимости точности от разрядности
- Разложены классы весовых коэффициентов на битовые плоскости
- Проанализированы аппаратные затраты

## Матрица спутывания

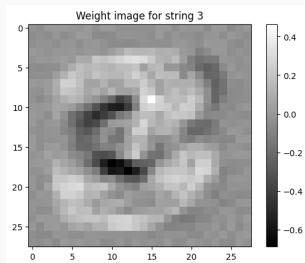


## Точность и затраты блоков LUT/FF

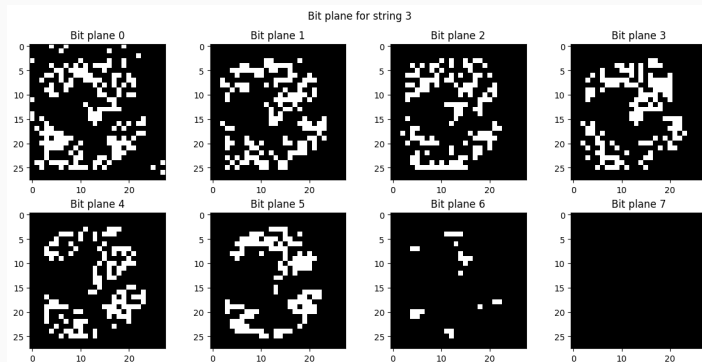


# Разложение на битовые плоскости

## Весовой класс 3



## Битовые плоскости





# Аппаратные затраты

Таблица 1: Аппаратные затраты для 5 битного представления коэффициентов

| Тип блока     | Использовано | Доступно | Соотношение,<br>% |
|---------------|--------------|----------|-------------------|
| LUT as logic  | 2180         | 17600    | 12.39             |
| LUT as memory | 60           | 6000     | 1                 |
| Flip Flop     | 862          | 35200    | 2.45              |
| RAMB18        | 10           | 120      | 8.33              |
| DSP           | 0            | 80       | 0                 |
| BUFG          | 1            | 32       | 3.13              |

- Рассмотренный эксперимент на основе НС прямого распространения с полносвязным слоем показывает, что формат представление весовых данных существенно влияет на точность определения до 5 битной разрядности. Дальнейшее увеличение разрядности не несет значительных изменений в точности.
- Предложенная структура НС показывает, что при увеличении разрядности наблюдается линейный рост в потреблении LUT и FF блоков FPGA.