Компактные и эффективные нейронные сети для распознавания изображений на основе обучаемого двумерного разделимого преобразования

М.И. Вашкевич, Е.А. Кривальцевич
vashkevich@bsuir.by

Белорусский государственный университет
информатики и радиоэлектроники
Кафедра электронных вычислительных средств
Минск, Беларусь

27-я конференция DSPA'2025
Цифровая обработка сигналов и её применение
Москва, Россия

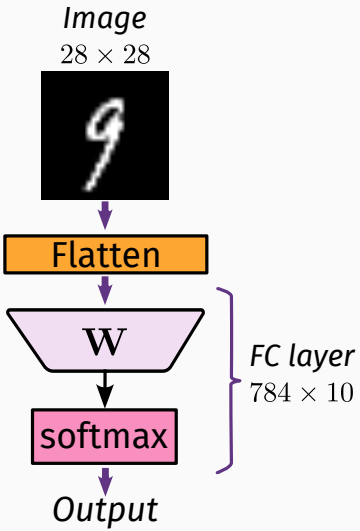
26 марта, 2025

Содержание

Введение

Реализация нейронной сети на ПЛИС

- Ставилась задача реализации на ПЛИС нейронной сети для распознавания изображений (рукописных цифр из базы MNIST)
- Простая однослойная нейронная сеть (7850 параметров) позволяет достичь относительно невысокой точности 92,5%
- При добавлении скрытых слоев число параметров сети стремительно увеличивается



Разделимое двумерное обучаемое преобразование

Двумерное разделимое преобразование

- **Двумерные разделимые преобразования** применяются в обработке изображений для снижения вычислительносй сложности при пространственной фильтрации. Ядро преобразования имеет вид:

$$\mathbf{W} = \mathbf{v} \times \mathbf{h}^T,$$

где $\mathbf{W} \in \mathbb{R}^{n \times n}$, $\mathbf{v}, \mathbf{h} \in \mathbb{R}^{n \times 1}$.

- Разделимое преобразование $\mathbf{W}$ имеет $2n$ независимых параметров, вместо n^2 параметров, которые имеет обычное преобразование.
- Пример разделимого преобразования – фильтр Собеля:

$$\begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} \times \begin{bmatrix} 1 & 2 & 1 \end{bmatrix}.$$

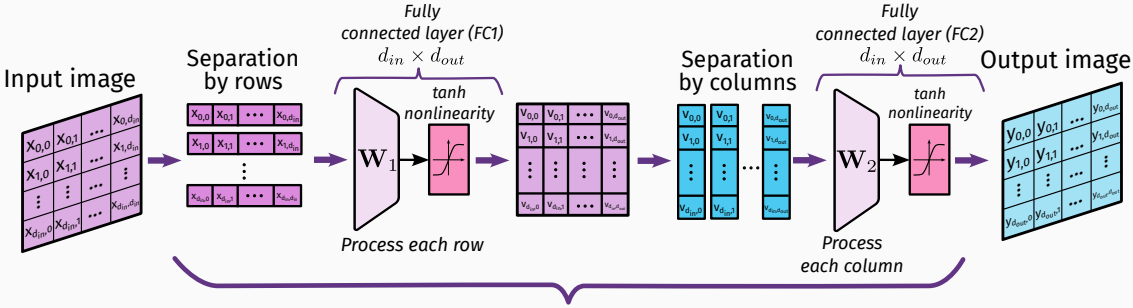
Двумерное разделимое обучаемое преобразование

- Предлагаемое обучаемое преобразование ($\mathrm{LST}_{\mathrm{2D}}$) обрабатывает изображение вначале по строкам, а затем по столбцам.
- Преобразование $\mathrm{LST}_{\mathrm{2D}}$ обрабатывает изображение $\mathbf{X}$ размера $d_{in} \times d_{in}$ при этом на выходе получается изображение $\mathbf{Y}$ размера $d_{out} \times d_{out}$:

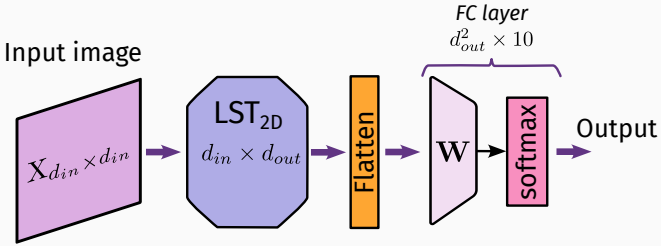
$$\mathbf{Y} = \mathrm{LST}_{d_{in} \times d_{out}}(\mathbf{X}) = \tanh(\mathbf{W}_2 \tanh(\mathbf{W}_1 \mathbf{X}^T)),$$

  где $\mathbf{W}_1$, $\mathbf{W}_2$ – матрицы весов слоев $\mathrm{FC1}$ и $\mathrm{FC2}$, соответственно, а d_{in} и d_{out} – это гиперпараметры преобразования, оперделяющие общее число обучаемых параметров $N_{params} = 2 \cdot (d_{in} + 1) \cdot d_{out}$.
- Мы используем нотацию $\mathrm{LST}_{d_{in} \times d_{out}}$ для обозначения слоя обучаемого преобразования, которое принимает на вход изображение $d_{in} \times d_{in}$ и выдает изображение размера $d_{out} \times d_{out}$.

Двумерное разделимое обучаемое преобразование



Нейронная сеть LST-1

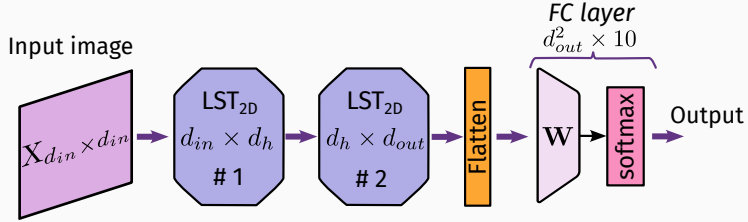


- LST_{2D} можно рассматривать, как базовый блок для построения компактных нейронных сетей для распознавания изображений
- Нейронная сеть LST-1 – простейший вариант нейронной сети, использующий блок LST_{2D}.
- Число параметров LST-1:

$$N_{params} = 2 \cdot (d_{in} + 1) \cdot d_{out} + (d_{out}^2 + 1) \times 10$$

- Для $d_{in} = d_{out} = 28$ число параметров модели $N_{params} = 9\,474$.

Нейронная сеть LST-2

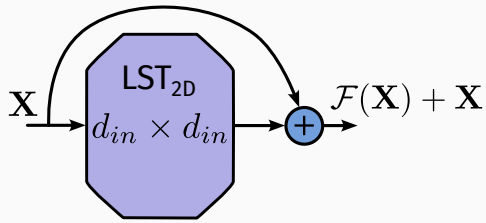


- Модель LST-2 имеет в своей основе два последовательно включенных блока LST$_{2D}$, что делает её более глубокой и выразительной
- Модель LST-2 имеет дополнительный параметр d_h – размерность «изображения» на скрытом слое. Число параметров модели LST-2

$$N_{params} = 2 \cdot (d_{in} + 1) \cdot d_h + 2 \cdot (d_h + 1) \cdot d_{out} + (d_{out}^2 + 1) \times 10$$

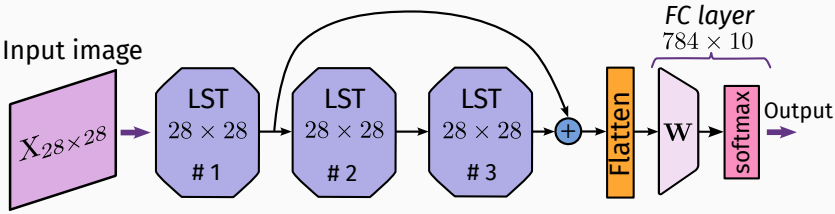
- Для $d_{in} = d_h = d_{out} = 28$ число параметров модели $N_{params} = 11\,098$.

Нейронная сеть с остаточными связями на базе LST



- Наличие остаточных связей (англ. *residual connection*) в нейронной сети позволяет решить проблему «затухающих» градиентов и получать глубокие модели
- Если размерность изображения на входе и на выходе блока LST совпадают, то появляется возможность построения нейронной сети с остаточными связями
- Таким образом, LST можно рассматривать, как базовый блок для построения глубоких нейронных сетей с небольшим числом обучаемых параметров

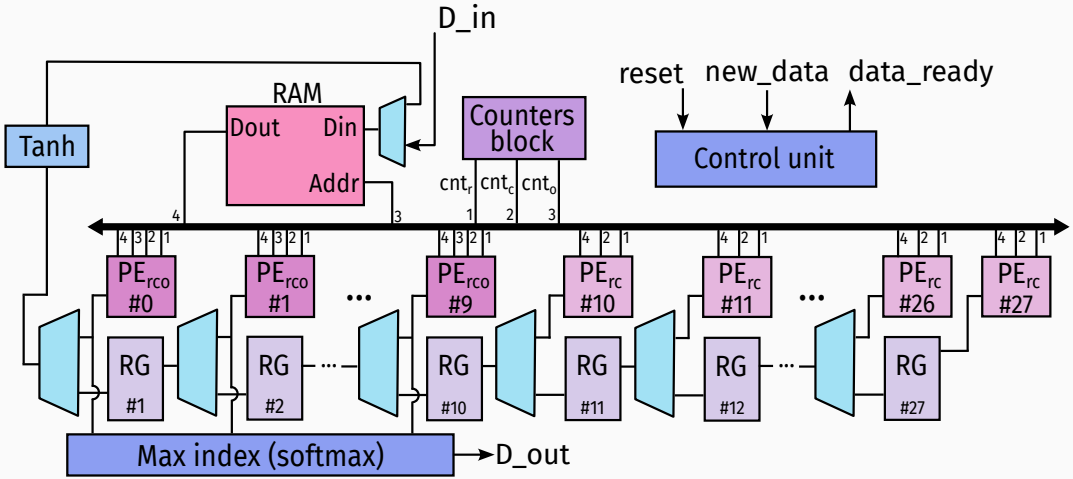
Нейронная сеть с остаточными связями на базе ResLST-3



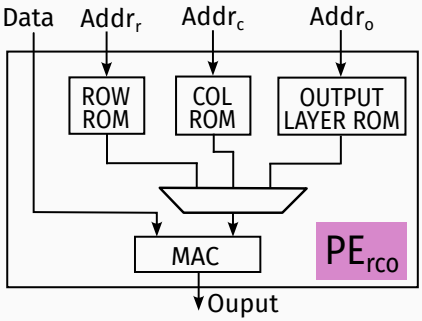
- В работе предложена нейронная ResLST-3, состоящая из трех блоков LST, и имеющая одну остаточную связь
- Внутреннее представление изображения имело размерность 28×28
- Общее число параметров модели $N_{params} = 12\ 722$

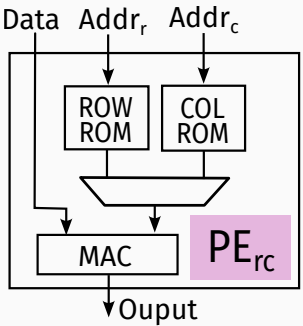
Реализация на FPGA

Реализация LST-1 на FPGA



Вычислительные блоки модели LST-1





- В состав вычислителя LST-1 входят 10 блоков PE_{rco} и 18 блоков PE_{rc}
- На первом этапе вычисления LST_{2D} задействуются все процессорные элементы PE. В каждом PE в памяти «ROW ROM» хранится один столбец матрицы $\mathbf{W}_1$ (для обработки строк изображения), а в памяти «COL ROM» хранятся столбцы матрицы $\mathbf{W}_2$ (для обработки столбцов изображения)

Результаты синтеза

- Вычислитель LST-1 описан на языке SystemVerilog и реализован на отладочной плате Xilinx Zybo Z7 (FPGA XC7Z010)
- Для организации процесса тестирования использовался дистрибутив Linux PYNQ, который запускался на ARM ядре кристалла XC7Z010.
- Вычислитель LST-1 был реализован в виде IP-ядра с использованием 12-разрядного представления чисел.

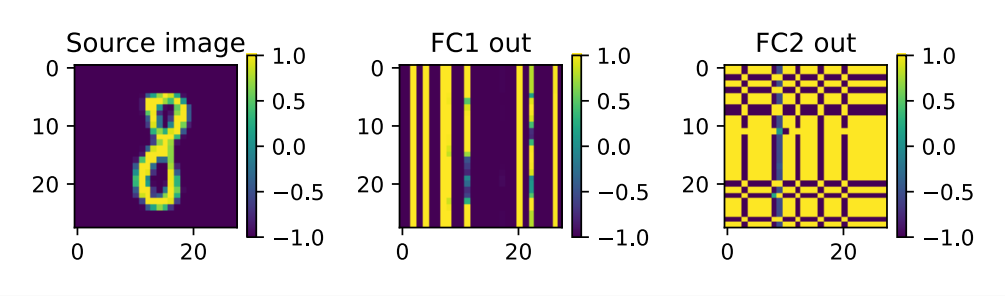
Тип блока	Использовано	Доступно	Соотношение, %
LUT as logic	6473	17600	36,8
Flip Flop	680	35200	1,9
RAMB18	29	120	24,2
DSP	0	80	0

Эксперименты и результаты

Описание эксперимента

- Набор данных MNIST (60+10 тыс. изображений рукописных цифр размером 28×28)
- Инициализация весов моделей выполнялась методом Ксавье
- Целевая функция – отрицательное логарифмическое правдоподобие (`torch.nn.NLLLoss`)
- Обучение производилось при помощи алгоритма Adam (скорость обучения $\eta = 2 \cdot 10^{-3}$, число эпох – 300, размер батча – 1000)
- Для оценки качества распознавания использовались метрика – точность (англ. *accuracy*)

Представление изображения в модели LST-1



- Модель LST-1 кодирует изображение, как нерегулярный шаблон похожий на шахматную доску.

Сравнение предложенных нейронных сетей на базе LST

Автор	Архитектура НС	# Число параметров	Точность
Liang, et al. (2018)	784-2048-2048-2048-10	10 100 000	98.32%
Umuroglu, et al. (2017)	784-1024-1024-10	1 863 690	98.40%
Medus, et al. (2019)	784-600-600-10	891 610	98.63%
Huynh [1]	784-126-126-10	115 920	98.16%
Huynh [1]	784-40-40-40-10	34 960	97.20%
Westby, et al. [2]	784-12-10	9 550	93.25%
$\mathrm{LST}_{\mathrm{2D}}$-1 [предложена]	$\mathrm{LST}_{28\times28}$-784-10	9 474	98.02%
$\mathrm{LST}_{\mathrm{2D}}$-2 [предложена]	$2 \times \mathrm{LST}_{28\times28}$-784-10	11 098	98.34%
ResLST-3 [предложена]	$3 \times \mathrm{LST}_{28\times28}$-784-10	12 722	98.53%

[1] T. V. Huynh, “Deep neural network accelerator based on FPGA,” in 4th NAFOSTED Conference on Information and Computer Science, 2017, pp. 254–257.

[2] I. Westby, et al. “FPGA acceleration on a multilayer perceptron neural network for digit recognition,” Journal of Supercomputing, 2021, vol. 77, no. 12, pp. 356–373.

Выводы

- Предложено двумерное обучаемое разделимое преобразование, которое может быть использовано в качестве базового блока для построения компактных нейронных сетей для распознавания изображений
- Предложены три нейронных сети на основе LST, которые имеют высокую точность распознавания рукописных цифр (более 98 %) и малое число обучаемых параметров (9–12 тыс.)
- Предложена архитектура вычислителя для реализации модели LST-1 на базе FPGA
- Предложенный блок LST можно рассматривать, как альтернативу полносвязному слою при реализации нейронных сетей прямого распространения (многослойных перцептронов)

LST – можно пробовать