УДК 004.032.26 + 004.383.3

АППАРАТНО-ПРОГРАММНАЯ РЕАЛИЗАЦИЯ КЛАССИФИКАТОРА НА ОСНОВЕ МОДЕЛИ LST-1 И СПЕКТРАЛЬНЫЙ АНАЛИЗ ЕГО ПРИЗНАКОВОГО ПРОСТРАНСТВА

Тондель Д.А., Вашкевич М.И.

Белорусский государственный университет информатики и радиоэлектроники, Минск, Республика Беларусь, vashkevich@bsuir.by

Аннотация: рассмотрена реализация бинарного классификатора изображений на основе модели LST-1 на кристалле Xilinx Zynq-7000 XCZ010. Модель обучена на двух классах базы FashionMNIST («t-shirt» и «shirt»). Для аппаратной реализации модели её параметры квантовались в формат Q4.12. Спектральный анализ графа признаков LST-слоя показал, что 784 признака сводятся к 106 спектральным направлениям без потери точности, что указывает на избыточность признакового пространства модели и определяет направление ее дальнейшего сокращения. Сравнивались два варианта реализации модели: программный на процессоре ARM Cortex-A9 и аппаратно-программный с реализацией LST-слоя в виде IP-ядра на Vitis HLS. Перенос LST-слоя в программируемую логику снижает время классификации одного изображения с 542.6 до 221.8 мкс (ускорение в 2.4 раза) при сохранении точности 86.0%.

Ключевые слова: LST-1, ПЛИС, Zynq-7000, HLS, квантование, спектральная теория графов, классификация изображений.

I. ВВЕДЕНИЕ

Компактные нейросетевые модели на основе обучаемого двумерного разделимого преобразования (англ. *learned separable transform, LST*) пригодны для встраиваемых систем [1], однако их размещение на ПЛИС требует ответа на два вопроса: какая точность арифметики достаточна, и какая часть модели должна быть вынесена в программируемую логику.

Цель данной работы реализовать и экспериментально сравнить два варианта инференса модели LST-1 на Zynq-7000 (программный и аппаратно-программный) и оценить избыточность признакового пространства модели методами спектральной теории графов [2-3].

II. МОДЕЛЬ LST-1 И ЕЕ ПОДГОТОВКА К АППАРАТНОЙ РЕАЛИЗАЦИИ

LST-1 – это семейство компактных нейросетевых архитектур [1], основанных на обучаемом двумерном разделимом преобразовании (ОДРП). LST-слой заменяет полносвязный слой парой преобразований – по строкам и по столбцам, что резко сокращает число параметров при сохранении точности классификации. Вычисления, осуществляемые при выполнении ОДРП выражены формулой (1).

$$Y = tanh(W_2\, tanh(W_1 X^T + b_1)^T + b_2), \tag{1}$$

где X – входное изображение, W_1 и b_1– матрицы весов и смещений полносвязного слоя FC1 слоя LST (обработка по строкам), W_2 и b_2 – матрицы весов и смещений полносвязного слоя FC2 слоя LST (обработка по столбцам).

Архитектура модели LST-1 представлена на рисунке 1.

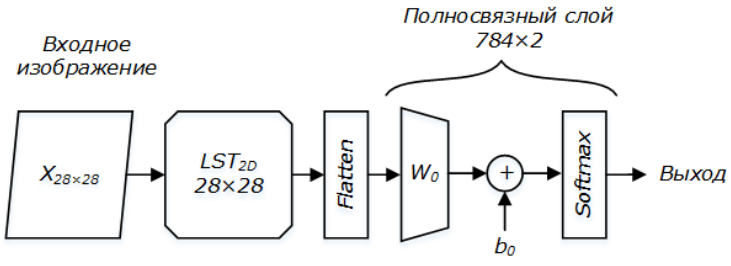


Рисунок 1. Архитектура LST-1

Входное изображение размером 28×28 поступает на LST-слой, выходом которого является вложение размером 28×28. Выход ОДРП представляется в виде вектора-столбца длиной 784 элемента, который поступает на полносвязный слой W_0 модели размером 784x2. Выходом модели является вектор-столбец длиной 2 элемента, содержащий вероятности принадлежности изображения к одному из двух классов.

Архитектура модели LST-1 была описана на языке Python. Для её обучения были использованы классы 0 («t-shirt») и 6 («shirt») базы данных FashionMNIST, размер входных данных составил 28×28 отсчетов. В качестве функции потерь была использована CrossEntropyLoss, оптимизатор – AdamW, планировщик скорости обучения – CosineAnnealingWarmRestarts. Модель обучалась на протяжении 200 эпох. Точность классификации обученной модели на тестовой выборке составила 86,1%.

Основная вычислительная операция в нейронной сети – скалярное произведение вектора входных данных на вектор весов. В формате Q4.12 каждое произведение дает результат в формате Q8.24 и помещается в 32-битную переменную, которая играет роль аккумулятора – в ней накапливается сумма всех произведений. После завершения суммирования накопленное значение нужно сдвинуть вправо на 12 бит, чтобы вернуться к формату Q4.12. Используя обучающую выборку, было установлено, что 32-битного аккумулятора достаточно, чтобы избежать переполнения при суммировании, так как промежуточные значения в аккумуляторе находятся в диапазоне [-8; 8).

III. СПЕКТРАЛЬНЫЙ АНАЛИЗ ПРИЗНАКОВОГО ПРОСТРАНСТВА

Для входного изображения размера 28×28 модель генерирует D=784 признака. Пропустив весь обучающий набор из N примеров через модель, была получена матрица признаков F, столбцы $f_j = [f_{1,j}, \quad f_{2,j}, \quad \dots, \quad f_{N,j}]^T$ $(j = 1, \dots, D)$ которой описывают поведение одного признака по всей выборке. По матрице признаков была построена матрица сходства A путем попарного сравнения ее столбцов по формуле 2.

$$A_{ij} = s(f_i, f_j) = \frac{1}{1+d(f_i,f_j)}, \tag{2}$$

где $d(f_i, f_j)$ – функция расстояния между векторами f_i и f_j. В качестве функции расстояния было выбрано косинусное расстояние, учитывающее только направление векторов, игнорируя их масштаб. Визуализация полученной матрицы сходства приведена на рисунке 2.

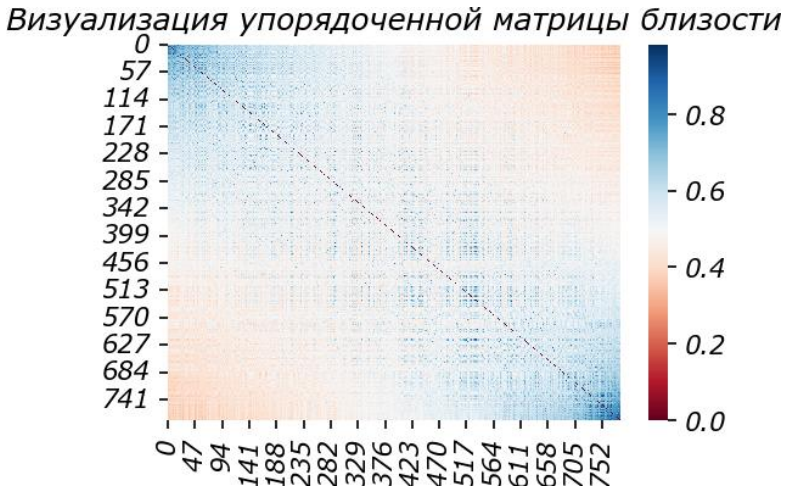


Рисунок 2. Упорядоченная матрица сходства признаков LST-слоя

Для матрицы близости были найдены ее невзвешенный и симметричный нормализованный лапласианы. Спектр лапласиана позволяет разбить D признаков на группы таким образом, чтобы признаки внутри одной группы были сильно похожи, а признаки из разных групп – слабо. Для этого необходимо вычислить k собственных значений лапласиана и соответствующие им собственные векторы v_k, при этом собственные векторы образуют базис $V_k = [v_2, \ v_3, \dots, v_{k+1}] \in R^{D \times k}$. Для каждого сэмпла $x_n \in R^D$ (строка матрицы признаков F) получаем эмбеддинг по формуле 3.

$$z_n = x_n^T V_k \in R^k. \tag{3}$$

На полученных эмбеддингах был обучен классификатор по методу линейного дискриминантного анализа (LDA). Обучение проводилось для всех возможных $k \in [1; 784]$ собственных векторов. Зависимость точности от используемой размерности эмбеддинга приведена на рисунке 3.

Оптимальной была признана размерность $k = 106$ при использовании симметричного нормализованного лапласиана, так как она дала существенное снижение размерности при минимальных потерях точности (86,05%).

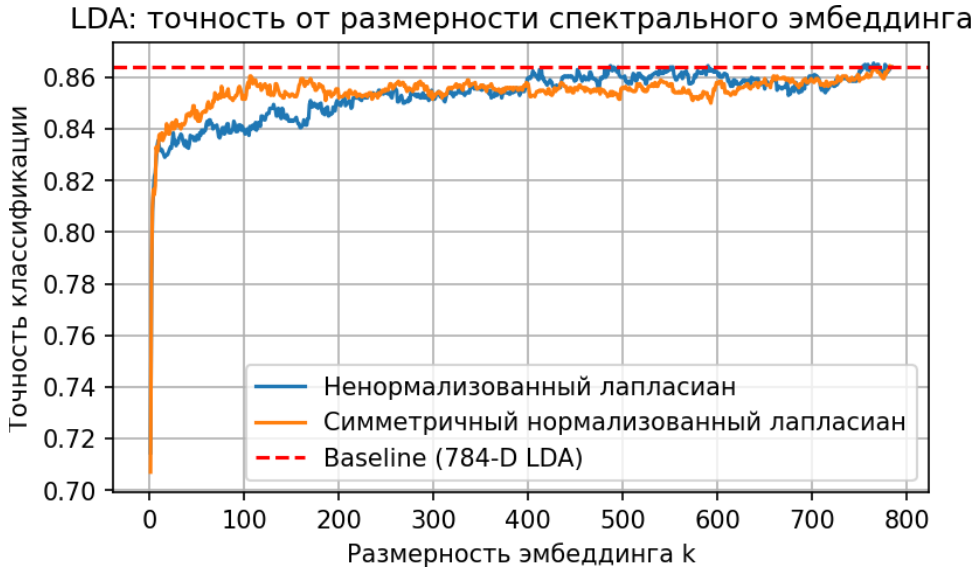


Рисунок 3. Зависимость точности классификации от размерности спектрального эмбеддинга для комбинаторного и нормализованного лапласианов

IV. АППАРАТНО-ПРОГРАММНАЯ РЕАЛИЗАЦИЯ НА ZYNQ-7000

а. Вариант A – программный инференс

Классификатор на основе модели LST-1 был реализован исключительно программным образом на языке C. Инференс был оформлен как разделяемая библиотека, компилируемая непосредственно на плате Zybo Z7 и вызываемая из Python через систему быстрого прототипирования PYNQ.

б. Вариант B – LST-слой в виде IP-ядра

При аппаратно-программной реализации наиболее вычислительно емкая часть – LST-слой – синтезирована в Vitis HLS как IP-ядро с интерфейсами AXI-Stream и AXI-Lite и обменом через DMA. При этом выходной полносвязный слой W_0 и функция argmax (заменяющая softmax в процессе инференса) выполнялись программно на процессоре ARM используя тот же C-код. Такой подход обеспечивал корректность сравнения вариантов. Структурная схема данной реализации приведена на рисунке 4.

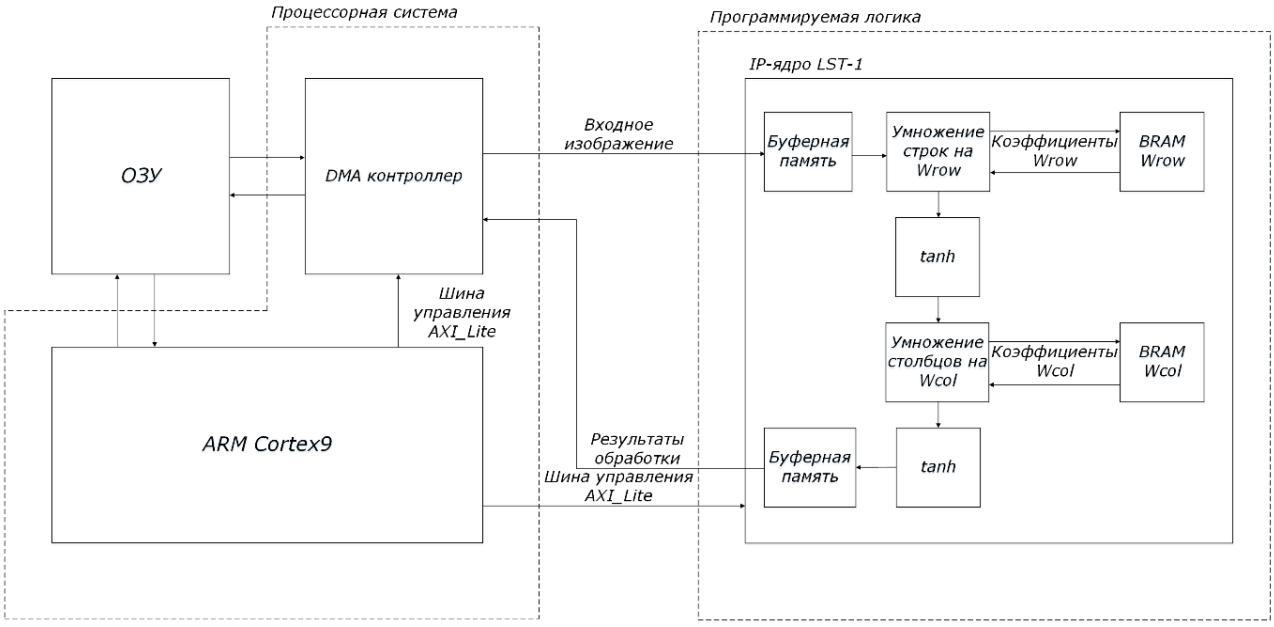


Рисунок 4. Структурная схема аппаратно-программной реализации (вариант B)

V. РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТА

Перенос LST-слоя в программируемую логику снижает медианную латентность с 538,3 до 221,8 мкс на изображение (ускорение в 2,4 раза, пропускная способность 1722 против 4244 изобр./с) при полном сохранении точности 86,0 %. Значения латентности для обеих реализаций приведены на рисунке 5. Полное сравнение обеих реализаций приведен в таблице 1.

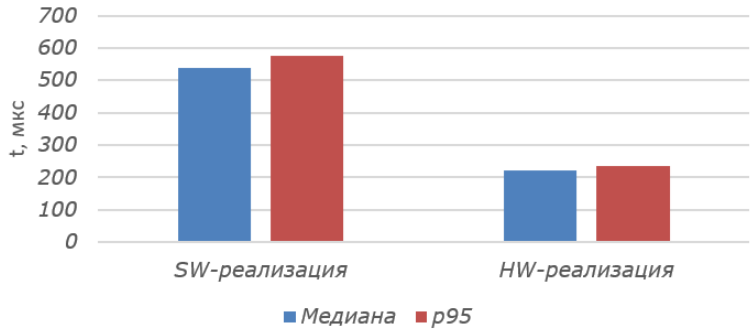


Рисунок 5. Распределение латентности инференса для вариантов реализации A и B

Таблица 1. Сравнение вариантов A и B: латентность, пропускная способность, точность, ресурсы ПЛИС и мощность

Метрика	*SW*-реализация	*HW*-реализация
Латентность (медиана), мкс/*изобр.*	538,3	221,8
Латентность (p95), мкс/ *изобр.*	577,2	235,6
Пропускная способность, *изобр./с*	1722	4244
Точность	0,8600	0,8600
LUT, шт. (% от доступного количества)	–	7008 (39,82)
LUTRAM, шт. (% от доступного количества)	–	1343 (22,38)
FF, шт. (% от доступного количества)	–	7599 (21,59)
BRAM, шт. (% от доступного количества)	–	13,50 (22,50)
DSP, шт. (% от доступного количества)	–	57 (71,25)
BUFG, шт. (% от доступного количества)	–	2 (6,25)
Потребляемая статическая мощность, Вт	–	0,122
Потребляемая динамическая мощность, Вт	–	1,569

VI. ЗАКЛЮЧЕНИЕ

Показано, что модель LST-1 переносится на платформу Zynq-7000 в арифметике Q4.12 без потери точности, а перенос LST-слоя в программируемую логику дает ускорение в 2,4 раза; спектральный анализ графа признаков выявил их существенную избыточность (106 значимых направлений из 784), что определяет направление дальнейшего сокращения модели и ее аппаратной реализации.

БЛАГОДАРНОСТЬ

Авторы выражают благодарность резиденту ПВТ компании «Инженерный Центр Ядро», которая является одним из центров разработки YADRO, за предоставленное оборудование для проведения экспериментов в рамках работы совместной учебной лаборатории с БГУИР.

ЛИТЕРАТУРА

[1] Кривальцевич, Е. А. Нейронные сети на основе обучаемого двумерного разделимого преобразования для классификации изображений: теория и аппаратная реализация на FPGA / Е. А.

Кривальцевич, М. И. Вашкевич // Информатика. 2025. Т. 22, № 4. С. 36–54. DOI: 10.37661/1816-0301-2025-22-4-36-54.

[2] Belkin, M. Laplacian eigenmaps for dimensionality reduction and data representation / M. Belkin, P. Niyogi // Neural Computation. 2003. Vol. 15, No 6. P. 1373–1396.

[3] Von Luxburg, U. A tutorial on spectral clustering / U. von Luxburg // Statistics and Computing. 2007. Vol. 17, No 4. P. 395–416.

[4] Xilinx Inc. Vitis High-Level Synthesis User Guide (UG1399). 2023.

[5] Xilinx Inc. Zynq-7000 SoC Technical Reference Manual (UG585). 2018. Mode of access: https://docs.amd.com/r/en-US/ug585-zynq-7000-SoC-TRM.

HARDWARE-SOFTWARE IMPLEMENTATION OF AN LST-1 BASED CLASSIFIER AND SPECTRAL ANALYSIS OF ITS FEATURE SPACE

D.A. Tondel, M.I. Vashkevich

Belarusian State University of Informatics and Radioelectronics, Minsk, Republic of Belarus,
vashkevich@bsuir.by

Abstract: the paper considers the implementation of a binary image classifier based on the LST-1 model on a Xilinx Zynq-7000 XCZ010 chip. The model was trained on two classes from the FashionMNIST dataset (“t-shirt” and “shirt”). For the hardware implementation of the model, its parameters were quantized to the Q4.12 format. Spectral analysis of the feature graph of the LST layer showed that 784 features are reduced to 106 spectral directions without loss of accuracy, which indicates redundancy in the model’s feature space and determines the direction for its further reduction. Two inference variants were compared: a software one on an ARM Cortex-A9 processor and a hardware-software one with the LST layer implemented as an IP core in Vitis HLS. Porting the LST layer to programmable logic reduces the classification time of a single image from 542.6 to 221.8 μs (a 2.4 speedup) while fully preserving an accuracy of 86.0%.

Keywords: LST-1, FPGA, Zynq-7000, HLS, quantization, spectral graph theory, image classification.