

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

Докладчик: Тондель Дарья Анатольевна

Научный руководитель: доцент, д.т.н. Вашкевич Максим Иосифович

Актуальность исследования

- Задачи классификации изображений во встраиваемых системах являются ограниченными по ресурсам.
- Классические полносвязные сети обладают избыточностью по числу параметров.
- Требуется компромисс между точностью, числом параметров, скоростью вычислений и энергопотреблением.

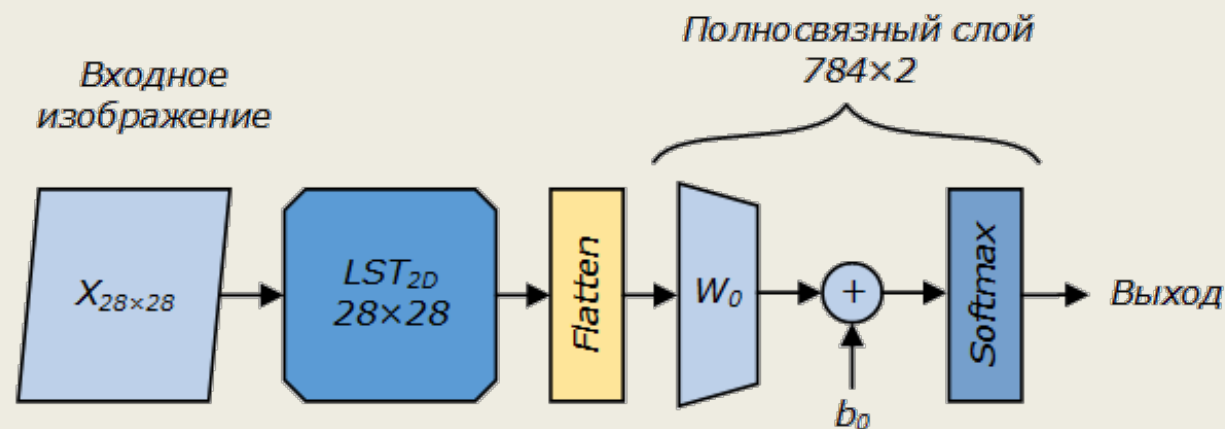
Цели исследования

- Реализовать и экспериментально сравнить два варианта инференса модели LST-1 на Zynq-7000 (программный и аппаратно-программный).
- Оценить избыточность признакового пространства модели методами спектральной теории графов.

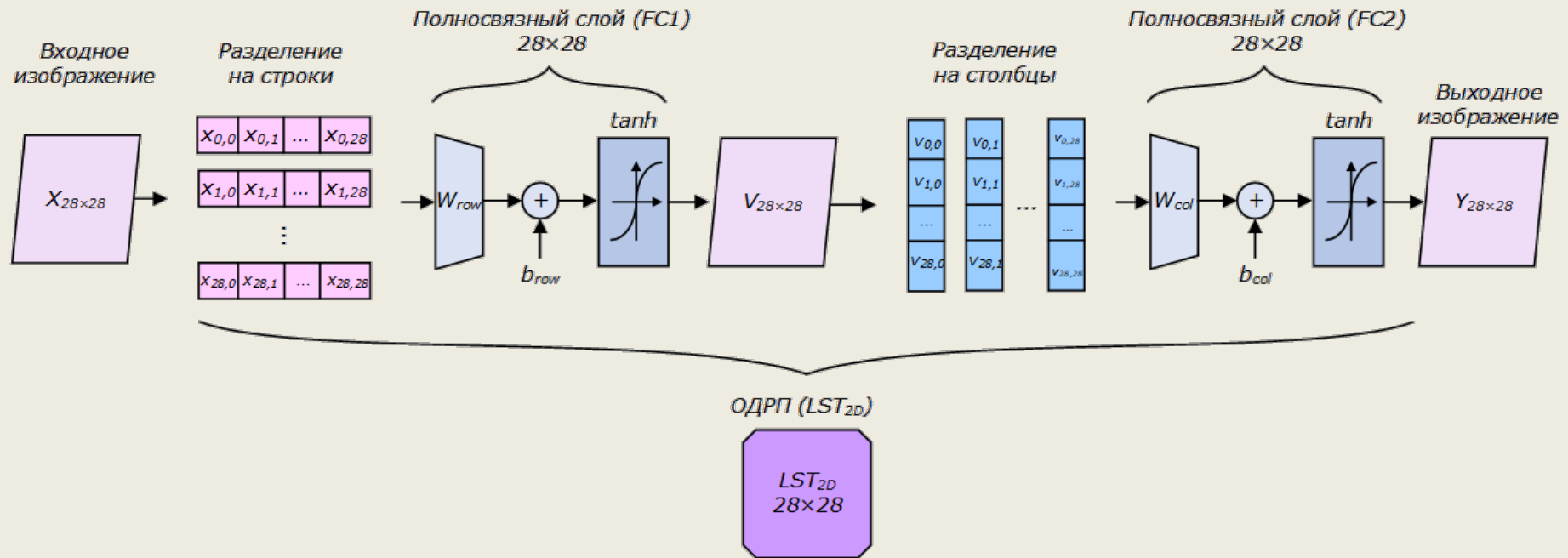
Модель LST-1

LST-1 – это семейство компактных нейросетевых архитектур, 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.



Представление ОДРП



Вычисления, осуществляемые при выполнении ОДРП выражены следующей формулой:

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

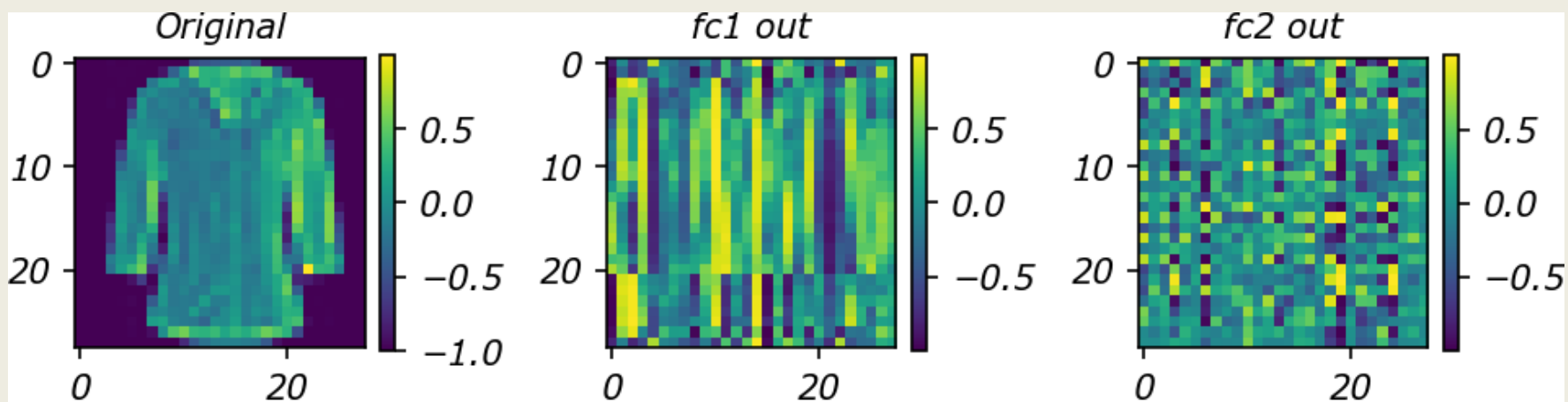
где X – входное изображение,

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).

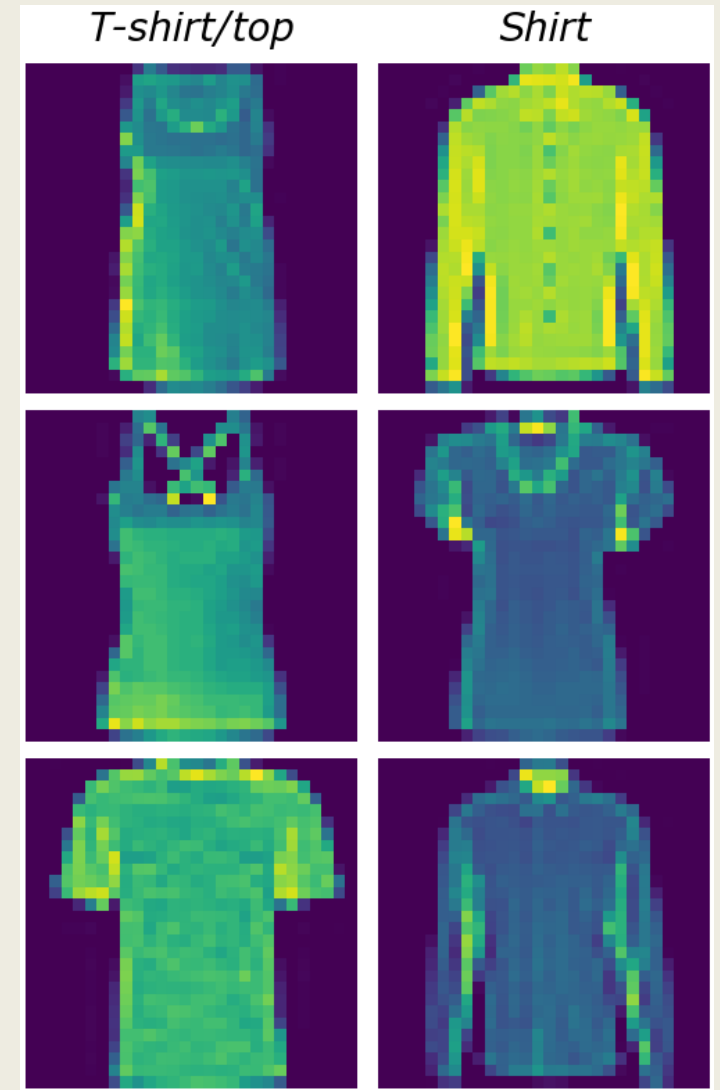
Представление ОДРП

Обработка изображения в нейронной сети LST-1-28



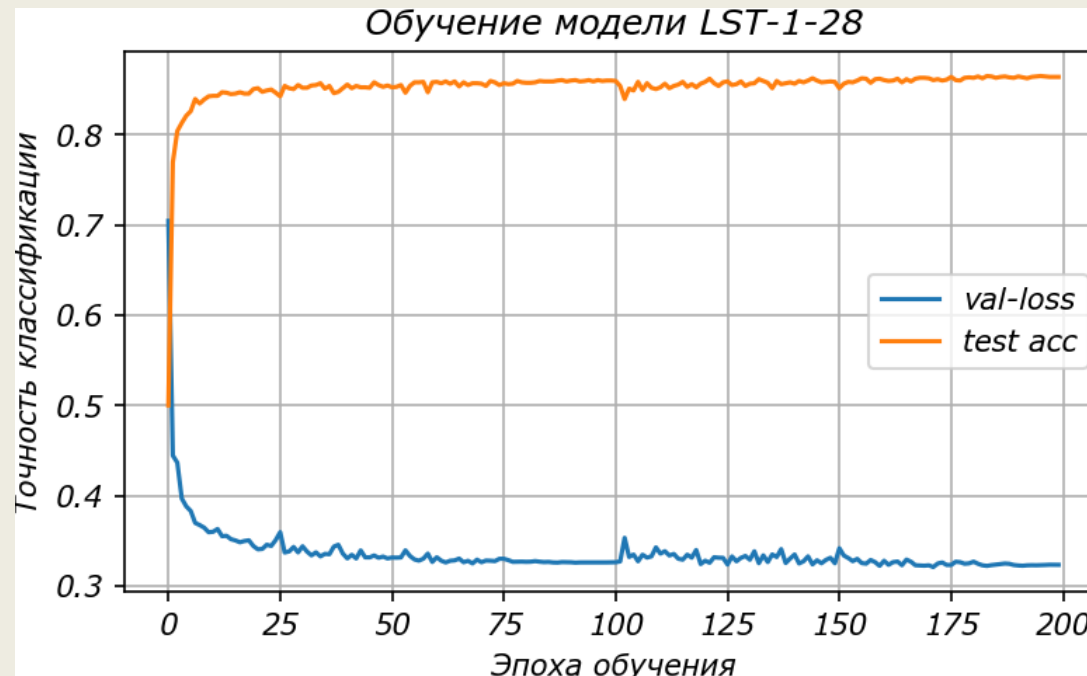
Набор данных FashionMNIST

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 28×28 пикселей.



Обучение модели LST-1

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%.

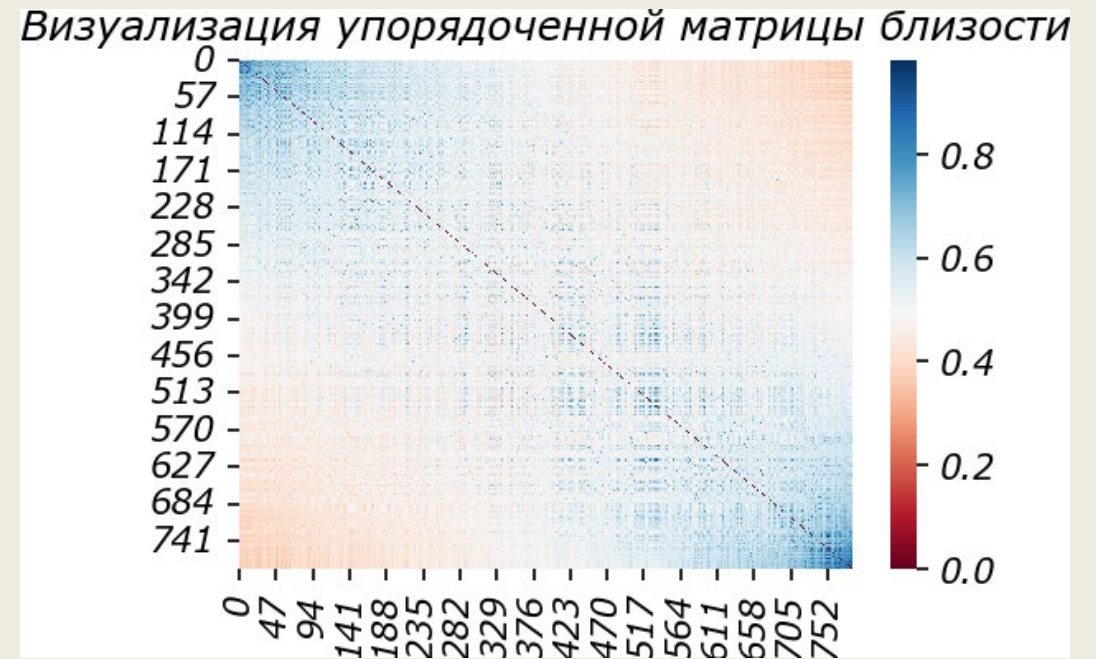
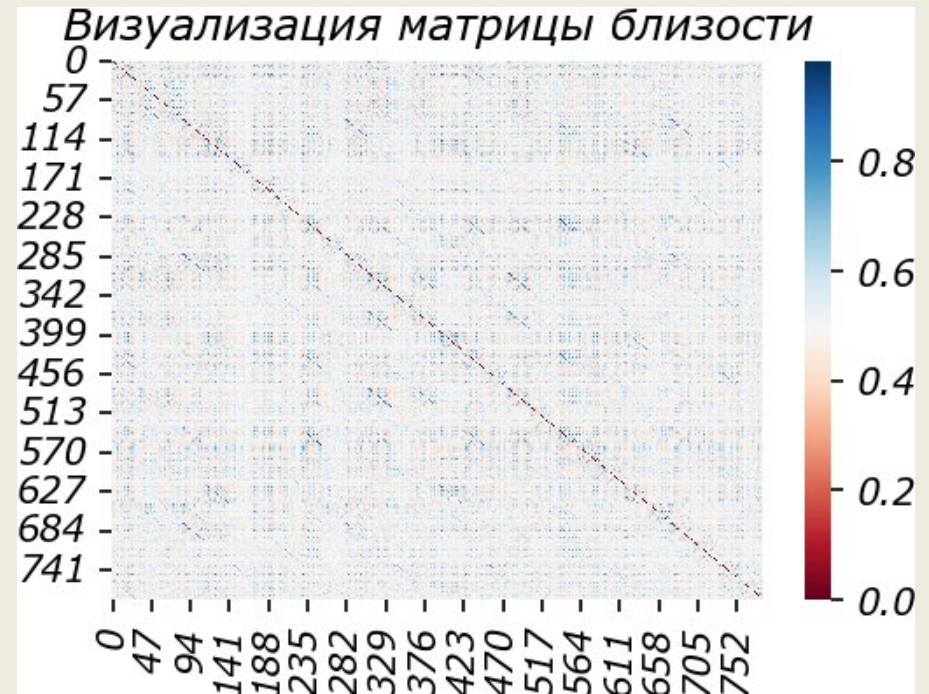


Спектральный графовый эмбединг

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 $f_j = [f_{1,j}, f_{2,j}, \dots, f_{N,j}]^T$ ($j = 1, \dots, D$) которой описывают поведение одного признака по всей выборке. По матрице признаков была построена матрица сходства A путем попарного сравнения ее столбцов по следующей формуле:

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

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.



Спектральный графовый эмбединг

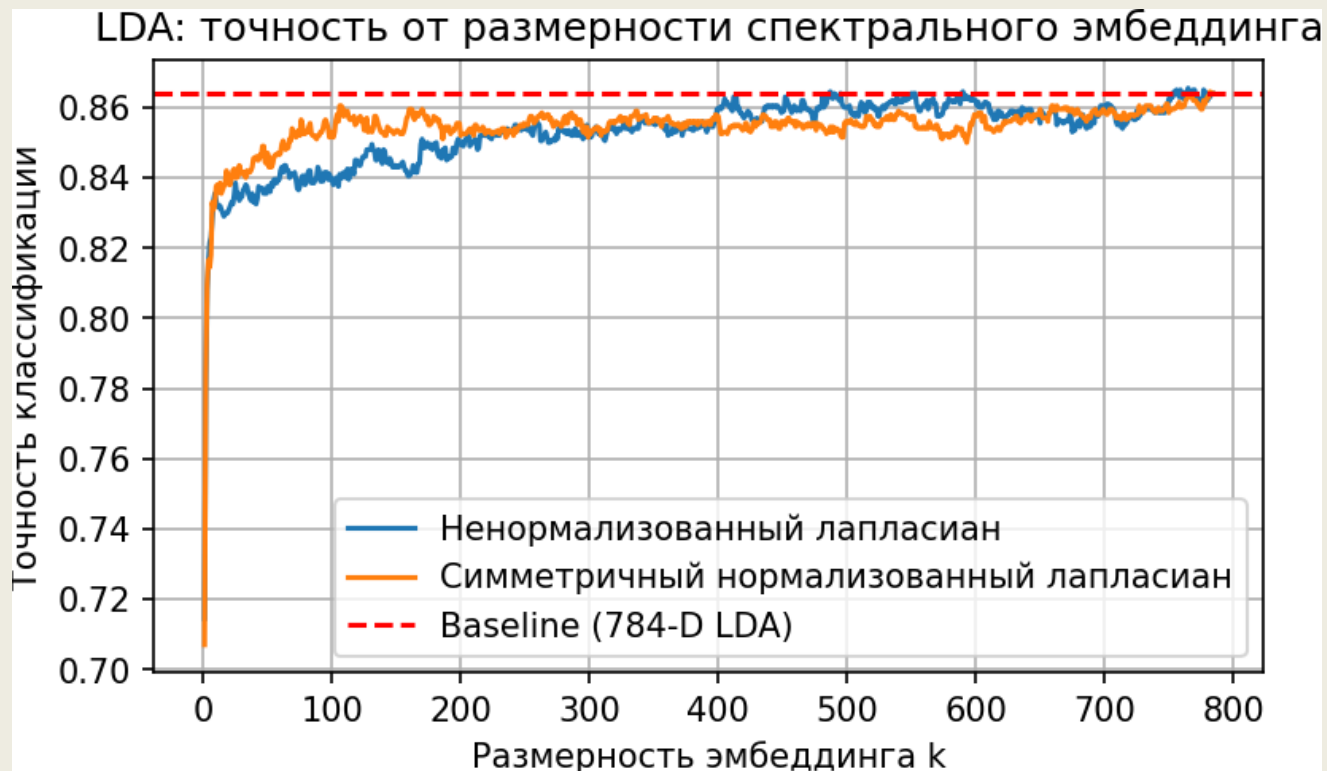
Для матрицы близости были найдены ее невзвешенный и симметричный нормализованный лапласианы.

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) получаем эмбединг по следующей формуле:

$$z_n = x_n^T V_k \in R^k$$

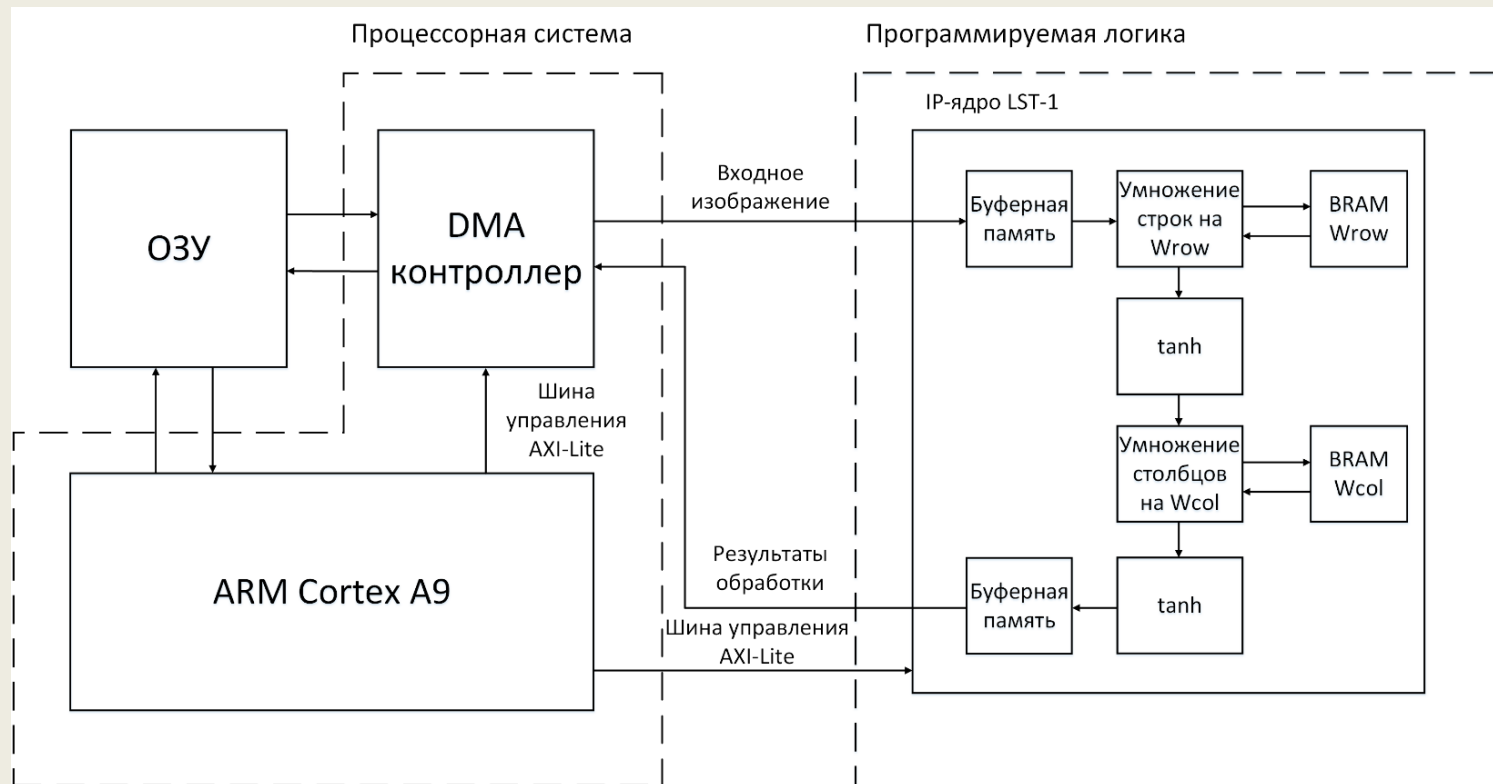
На полученных эмбедингах был обучен классификатор по методу линейного дискриминантного анализа (LDA).



Реализация модели LST-1 на FPGA

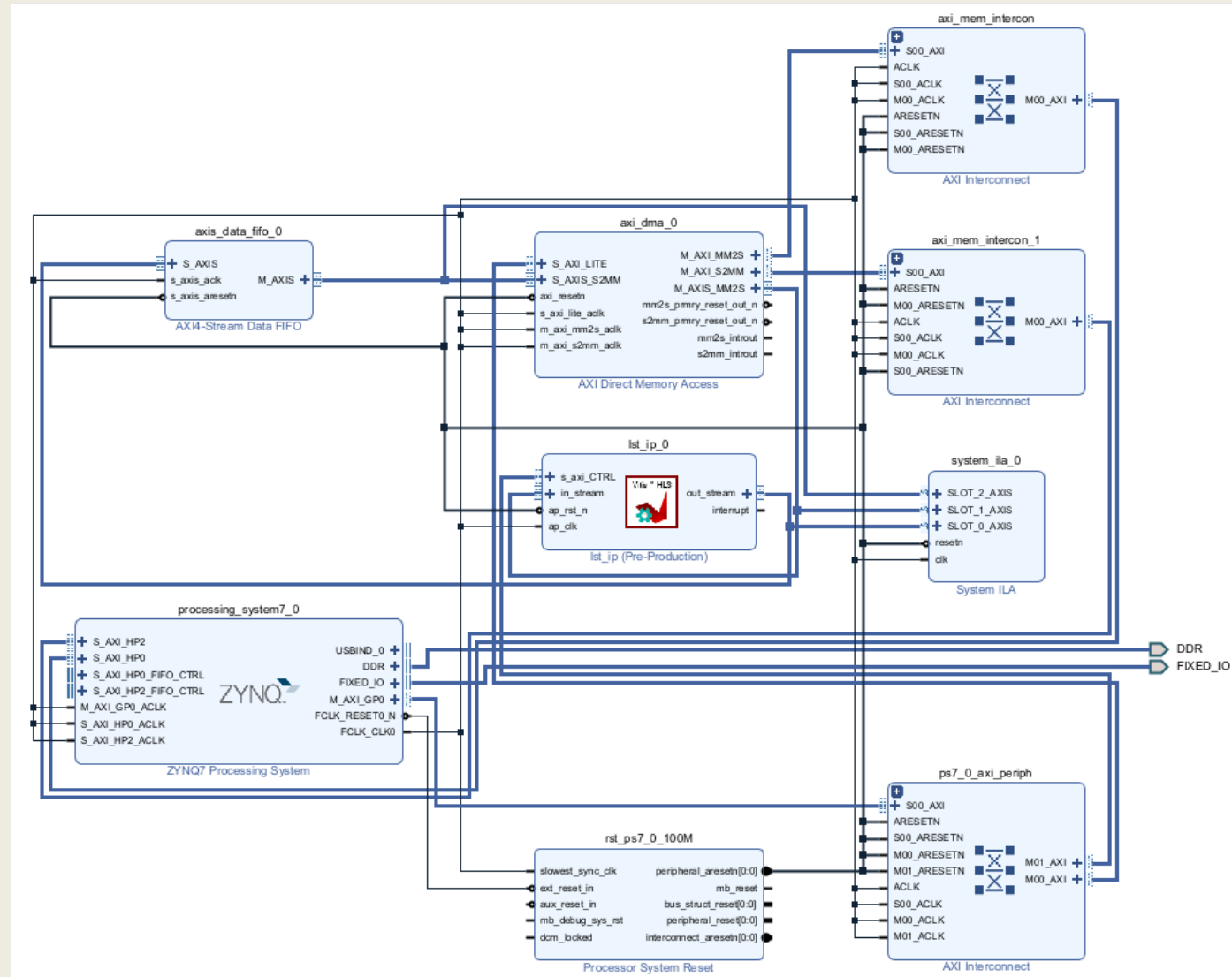
При аппаратно-программной реализации наиболее вычислительно емкая часть – LST-слой – синтезирована в Vitis HLS как IP-ядро с интерфейсами AXI-Stream и AXI-Lite и обменом через DMA. При этом выходной полносвязный слой W_0 и функция argmax (заменяющая softmax в процессе инференса) выполнялись программно на процессоре ARM, используя C-код.

Структурная схема устройства



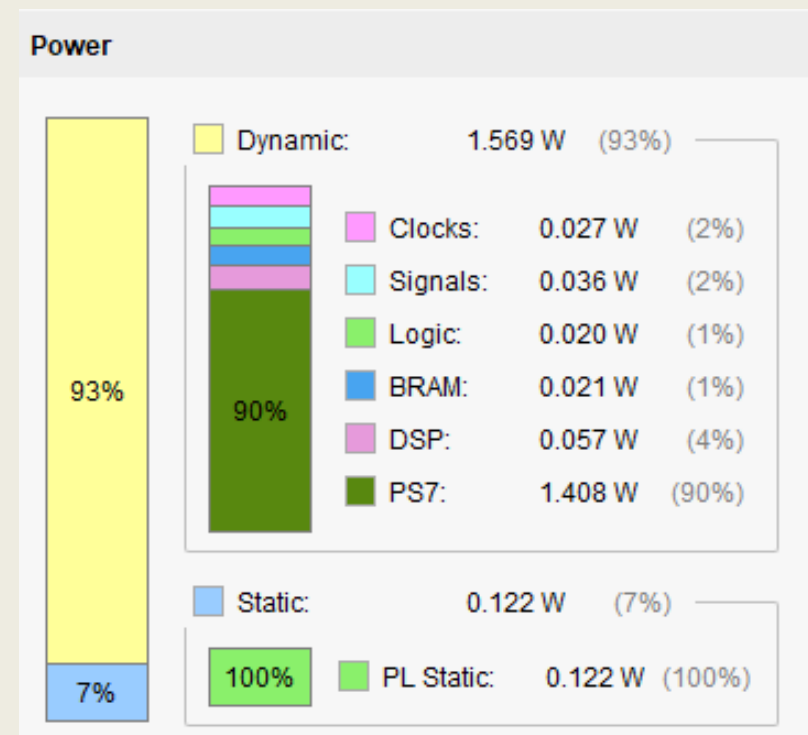
Реализация модели LST-1 на FPGA

Блочная диаграмма проекта



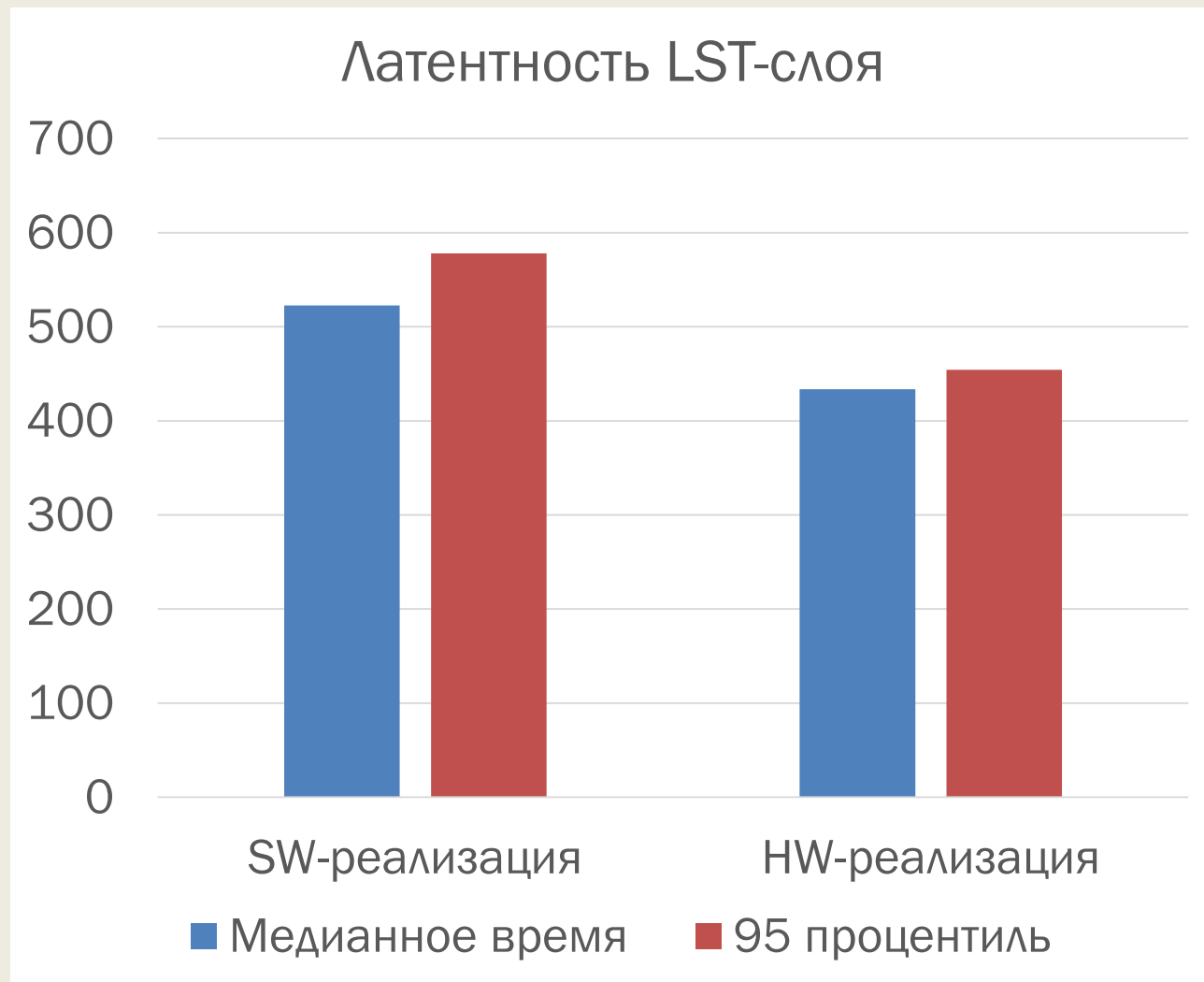
Результаты имплементации проекта в среде Vivado 2023.1

Ресурс	Использовано	Доступно	Использовано, %
LUT	7008	17600	39,82
LUTRAM	1343	6000	22,38
FF	7599	35200	21,59
BRAM	13,50	60	22,50
DSP	57	80	71,25
BUFG	2	32	6,25



Сравнение латентности двух реализаций

Также 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.



Выводы

- Обучена модель LST-1 на двух классах FashionMNIST. Точность классификации составила 86.1%.
- Спектральный анализ пространства признаков модели показал, что признаковое пространство, генерируемое LST-слоем является избыточным и размерность может быть снижена до 106 (в 7.4 раза) признаков при сохранении точности классификации 86.05% при использовании симметричного нормализованного лапласиана и LDA, хотя удаление шумовых компонент не позволяет увеличить точность классификации.
- Вычисления, производимые моделью, реализованы в виде программного и программно-аппаратного инференса, причем перенос части вычислений на ПЛИС увеличивает их скорость в 1.3 раза.

СПАСИБО ЗА ВНИМАНИЕ!

