kNN-STUFF: Performance Models (v0.2)

João Vieira

December 29, 2021

Single Accelerator

$$\#\text{cycles per classification} = (M + M \times N) \times \varepsilon_{\text{DMA}} + 17 + k$$

Parallel Configuration 0

$$\#\text{cycles per classification} = \frac{(M \times C \times A + M \times N) \times \varepsilon_{DMA} + 17 + k \times C \times A}{C \times A}$$

Parallel Configuration 1

$$\#\text{cycles per classification} = \frac{(M \times A + \frac{M \times N}{C}) \times \varepsilon_{DMA} + 17 + k \times C \times A}{A}$$

Notes

1. When using parallel configuration 1, there is an additional software component that merges the results calculated in hardware, incurring in an overhead per classification modeled by the equation $f(M, k, C) = k \times C \times (M + k)$.
2. All models assume a perfect CPU capable of fully exploiting kNN-STUFF capabilities which, for smaller datasets, may not be accurate.
3. When using parallel configurations, the maximum performance can only be achieved if:
   - (when using parallel configuration 0) the number of testing samples is a multiple of $A \times C$;
   - (when using parallel configuration 1) the number of testing samples is a multiple of A.
4. The bigger the dataset, the more reliable are the considered models.

Symbols

- M: number of features per sample;
- N: number of training samples;
- ε_{DMA}: Xilinx DMA efficiency (≈ 1);
- k: k nearest neighbors;
- C: number of clusters;
- A: number of accelerators per cluster.