Parallel Computing Conference, ParCo, Bologna, 12-15 September 2017

Multi-GPU k-Nearest Neighbor search in the context of data embedding

Adrian Kłusek [a,1] and Witold Dzwinel [a,2]

[a] *AGH University of Science and Technology, WIET, Department of Computer Science, al. A Mickiewicza 30, 30-059 Kraków, Poland*

Abstract. The k-nearest neighbor (k-NN) search is the rudimentary procedure widely used in machine learning and data embedding techniques. Herein we present a new multi-GPU/CUDA implementation of the brute-force (BF) k-NN algorithm. We demonstrate its advantages over currently the fastest GPU/CUDA implementations of BF k-NN, e.g., [1] both in terms of computational time and memory requirements. Unlike its competitors, our code scales linearly with the number of GPUs (up to 8 units) what allows for scrutinizing much larger high-dimensional datasets. We also present a new GPU implementation of the approximate k-NN algorithm used in the LargeVis data embedding algorithm [2], which is more than two orders of magnitude faster than its original CPU version. We discuss its limitations in terms of accuracy, efficiency and usefulness in data embedding.

Keywords. k-Nearest Neighbor, multi-GPU CUDA implementation

1. Introduction

Finding k-nearest neighbors (k-NN) in a set of M N-dimensional vectors is the basic problem in diverse scientific disciplines from molecular and agent simulations to pattern recognition and machine learning. Developing the fast k-NN algorithms is of principal interest in big data science, where not only the number of samples M but also their dimensionality N are very challenging. The high-dimensional data are usually positioned on a manifold $\sum$ embedded in the high-dimensional feature space $\mathbf{Y}$, such that $\dim\sum < \dim \mathbf{Y}$. In general, the dimension of $\sum$, its size and shape are unknown and have to be guessed from data. Only the local Euclidean distances from a given vector to its k- nearest neighbors can approximate the real distances on the manifold. Thus, the data embedding (DE) methods based on the Isomap algorithm [3] assume that the pair-wise distances between the nearest neighboring vectors are only known while all other distances in the manifold $\sum$ can be approximated by using, e.g., the FloydWarshall algorithm [3].

The growing popularity of visual analysis of medium-sized data through the data embedding [2–9] opens the debate of employing these methods also to visual inspection of truly big data. In this case, the concept based on excellent t-SNE embedding algorithm [4] cannot be used due to its high computational complexity and memory load. However,

[1]corresponding Author, E-mail: klusek@agh.edu.pl.

[2]E-mail: dzwinel@agh.edu.pl.