Online Additive Quantization

Qi Liu
qiliu67@mail.ustc.edu.cn
University of Science and Technology of China

Jin Zhang
abczj@mail.ustc.edu.cn
University of Science and Technology of China

Defu Lian*
liandefu@ustc.edu.cn
University of Science and Technology of China

Yong Ge
yongge@arizona.edu
The University of Arizona

Jianhui Ma
jianhui@ustc.edu.cn
University of Science and Technology of China

Enhong Chen
cheneh@ustc.edu.cn
University of Science and Technology of China

ABSTRACT

Approximate nearest neighbor search (ANNs) plays an important role in many applications ranging from information retrieval, recommender systems to machine translation. Several ANN indexes, such as hashing and quantization, have been designed to update for the evolving database, but there exists a remarkable performance gap between them and retrained indexes on the entire database. To close the gap, we propose an online additive quantization algorithm (online AQ) to dynamically update quantization codebooks with the incoming streaming data. Then we derive the regret bound to theoretically guarantee the performance of the online AQ algorithm. Moreover, to improve the learning efficiency, we develop a randomized block beam search algorithm for assigning each data to the codewords of the codebook. Finally, we extensively evaluate the proposed online AQ algorithm on four real-world datasets, showing that it remarkably outperforms the state-of-the-art baselines.

CCS CONCEPTS

• **Information systems** → **Nearest-neighbor search**; • **Computing methodologies** → *Machine learning*.

KEYWORDS

Additive Quantization, Beam Search, Nearest Neighbor Search, Online Update

ACM Reference Format:
Qi Liu, Jin Zhang, Defu Lian*, Yong Ge, Jianhui Ma, and Enhong Chen. 2021. Online Additive Quantization. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '21), August 14–18, 2021, Virtual Event, Singapore.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3447548.3467441

*Corresponding author: Defu Lian

KDD '21, August 14–18, 2021, Virtual Event, Singapore.

ACM ISBN 978-1-4503-8332-5/21/08…$15.00
https://doi.org/10.1145/3447548.3467441

1 INTRODUCTION

The advent of Internet has led to massive information overload in recent decades. For example, Google indexes more than 1 trillion webpages, Twitter hosts hundreds of millions of tweets and Flickr has billions of images. One method to address the information overload is to search for relevant information in data oceans with queries, but it turns out to be a challenging task. It is boiled down to the nearest neighbor search problem in a given database [18, 32], which can be efficiently yet approximately solved by the Approximate Nearest Neighbor search (ANNs) techniques, including hashing [33], quantization [23], tree [30] or graph index-based approaches [20, 21]. In addition to efficient information filtering, ANNs also plays an important role in many other tasks, such as recommender systems [24, 26, 28, 31], machine translation [7, 36], and multi-class classification [11, 25].

Due to data generation at an unprecedented rate per day, databases are dynamically growing while the data distribution may evolve over time. For instance, Twitter receives over 100 million tweets per day, Flickr receives over 3,000 images per minute, and Youtube has more than 100 hours of videos uploaded per minute. Without incorporating newly generated data, ANNs may not provide highly accurate responses to achieve satisfactory performance. However, it is almost computationally impractical to train the ANNs method from scratch each time the new data comes in due to the large size of the database. Therefore, it is important to develop ANNs algorithms to handling incremental data with a low computational cost.

Several studies have been conducted to support online learning of ANNs algorithms. There are mainly two lines of directions for this task. One line of research direction is online hash [3–5, 8, 12, 13, 17], to adapt hashing-based ANNs to accommodating the incremental data. The main idea is to update the hash function with the new data and then update the hash codes of existing stored data via the new hash functions. The advantage of the hash-based ANNs methods lies in the low computational cost of search and low storage cost of binary codes. However, the problems of these methods include low accuracy of approximation due to low capacity of representation, and the high computation cost of code maintenance due to the high frequent update of the hash functions. Moreover, these methods require keeping the old data so that the new hash code of the old data can be updated. Another line of research direction is online quantization [6, 35], to update codebooks to incorporate incremental data. The most related work is online product