Hyperdimensional Computing for Text Classification

Fateme Rasti Najafabadi*, Abbas Rahimi†, Pentti Kanerva†, Jan M. Rabaey†
*Sharif University of Technology †University of California, Berkeley

***Abstract*—Hyperdimensional computing explores the emulation of cognition by computing with hypervectors as an alternative to computing with numbers. Hypervectors are high-dimensional, holographic, and (pseudo)random with i.i.d. components. These properties provide an opportunity for efficient computing. We focus on an application of hyperdimensional computing for text classification. Accordingly, we present an algorithm for classifying news articles from a stream of input letters. Using pentagrams of letters, the algorithm achieves a classification accuracy above 94% on eight news topics, surpassing other techniques reported in the literature, including Bayes, K-NN, and SVM. We demonstrate a software framework that enables efficient execution of such algorithms on contemporary hardware fabrics.**

I. Introduction

Brain-inspired hyperdimensional computing [1], [2] depends on very high dimensionality, for instance 10,000 dimensions, and randomness. It represents things in hypervectors that are manipulated by operations that produce new hypervectors. Hypervectors are random and holographic, i.e., every piece of information contained in the vector is distributed equally over all the components of the vector. These properties allow hyperdimensional computing to achieve efficiencies that are out of reach of traditional von Neumann architecture at the expense of slight, or no functional performance degradation.

The algorithm that forms the basis of this work has been used for identifying the language of unknown sentences from 21 European languages [3]. Here we extend its use to classifying news articles from a stream of input letters. The text classification algorithm shares a common property with the language recognition algorithm: streaming input is represented by vectors of the same dimensionality. Based solely on pentagrams of letters, the algorithm exhibits a recognition accuracy above 94% for unknown news from eight categories.

II. Hyperdimensional Computing

Hyperdimensional computing represents information by projecting data onto vectors in a high-dimensional space. There exist a huge number of different, nearly orthogonal vectors in such a space. This lets us combine two hypervectors into a new hypervector using well-defined vectorspace operations, while keeping the information of the two with high probability. We consider the multiplication, addition, and permutation (MAP) coding described in [4] to define the hyperdimensional vector space. The hypervectors are initially taken from a 10,000-dimensional space and have an equal number of randomly placed 1s and -1s. Such hypervectors are used to represent the basic input elements, i.e., the 26 letters of the Latin alphabet and the (ASCII) space.

The MAP operations on such hypervectors are defined as follows. Elementwise addition of two hypervectors A and B, is denoted by $A + B$. Information from a pair of hypervectors A and B is stored and utilized in a single hypervector by exploiting the addition operation. That is, the sum of two separate hypervectors naturally preserves unique information from each hypervector because of the mathematical properties of vector addition. Similarly, elementwise multiplication is denoted by $A * B$. For representing a sequence of hypervectors, we use a permutation operation ρ of the hypervector coordinates. For instance, a sequence of three consecutive letters of A-B-C, or a trigram, is stored as the hypervector $\rho((\rho A) * B) * C = \rho\rho A * \rho B * C$. This efficiently distinguishes the sequence A-B-C from, for instance, A-C-B.

To measure the similarity of two hypervectors, we use cosine similarity. Cosine similarity measures similarity between two hypervectors by measuring the cosine of the angle between them using an inner product.

A. Text Classification Algorithm

Our algorithm uses a similar procedure for recognizing a text's language by generating and comparing the hypervectors. The algorithm has two main functions: encoding and similarity search. The encoding function summarizes the letter sequences of the input text by summing all pentagram hypervectors into a single hypervector. During the training phase, such a hypervector is generated from a known topic sample, and is referred to as topic hypervector. During the testing phase, such a hypervector is generated in a similar vein from an unknown topic sample, and is referred to as query hypervector. Then, the query hypervector is passed to the similarity search function for concurrent comparison with a set of *pre-stored* topic hypervectors of known topic samples. Finally, the search function returns the topic that its hypervector has the closet match, measured as the cosine angle, with the query hypervector.

We used Reuters newswire dataset with eight classes of topics. It has 5.5K documents for training and 2.2K documents for testing. The algorithm shows an accuracy of 93% using pentagrams when we consider "all terms" in the input text, hence there is no need for pre-processing. This accuracy is measured as the microaveraging that gives equal weight to each per-document classification decision, rather than per-class. The classification accuracy is further increased to 94% by using light-weight pre-processing that removes the frequent "stop" words and words with less than 3 letters.

References

[1] Kanerva, P. Hyperdimensional computing: An introduction to computing in distributed representation with high-dimensional random vectors. *Cognitive Computation*, 1(2):139–159, 2009.

[2] Kanerva, P. Computing with 10,000-bitwords. In *52nd Annual Allerton Conference on Communication, Control, and Computing*, 2014.

[3] Joshi, A., et al. Language recognition using random indexing. *Computation and Language archive: arXiv:1412.7026v2*, 2014.

[4] Gayler, R. Multiplicative binding, representation operators, and analogy. *Advances in analogy research*, p. 405, Sofia, Bulgaria: New Bulgarian University, 1998.