# POSTER: Applying Deep Learning to Object Store Caching

Effi Ofer
IBM Research - Haifa
effio@il.ibm.com

Amir Epstein
Work done while at IBM Research
amirepst@gmail.com

Dafna Sadeh
IBM Research- Haifa
dafna.sadeh@ibm.com

Danny Harnik
IBM Research - Haifa
dannyh@il.ibm.com

## ABSTRACT

Cache replacement policies comprise one of the oldest and most researched topic in computer science. But recent advances in the fields of artificial intelligence and machine learning introduce novel insight and new opportunities which can benefit prefetching and cache replacement policies.

In recent years the capabilities of artificial intelligence based algorithms have vastly expanded. State of the art artificial intelligence systems are capable of recognize images, predict human behavior, and beat the world champion in the ancient game of Go, by utilizing machine and deep learning techniques that can identify patterns within vast quantities of data. While cache replacement algorithms, such as LRU, ARC, LFU, and others, have been extensively studied, utilizing machine learning techniques to identify what and when to prefetch into the cache, remains an underexplored area. In this paper we make use of machine learning techniques to implement pattern based caching, an algorithm where we are able to identify which objects to prefetch from a multitenant cloud based object storage service into a shared cache before they are requested.

Our prefetching solution is based on a deep neural network that includes a word embedding phase and a deep learning phase. Word embedding is a technique to map words into vectors in high dimensional space. These vectors, which are capable of capturing syntactic and semantic relationships between words, have been shown to boost natural language processing tasks. Rather than convert words to vectors, we use word embedding to convert objects in the object store into vectors. Objects identifiers, similar to words in a text, have temporal relationships [1]. Thus, a sequence of objects identifiers requests can be analyzed in the same manner as a large corpus of text using word2vec type algorithms to produce object embeddings. Objects vectors are positioned in the vector space such that objects that have time correlations are in close proximity to each other in high dimensional space. Once we generate our object embeddings we then use recurrent neural networks (RNN) to predict relative likelihood of sequence of objects. This provide us with a model which we can then use to predict next object requests given a previous sequence of requests.

We tested our approach using simulations on traces taken from a real world publicly available cloud based multi-tenant object store - IBM Cloud Object Storage on the IBM Cloud. Object stores are uniquely fitting for machine learning based prefetching since each object is available with its meta data, enabling machine learning algorithms to take advantage of the semantic relationships contained within.

We have implemented our algorithm and tested it on real world data. We studied the benefits and issues involved and compare our results to other cache replacement policies. We built a simulation to compared its hit rate performance to that of an LRU based cache and found that under the right condition it outperforms the LRU. In this poster we will dive into the details of our neural network based algorithm and describe our insight into when our machine learning based prefetching outperforms regular cache replacement algorithms and when it is worse.

## REFERENCES

[1] Dong Dai, Forrest Sheng Bao, Jiang Zhou, and Yong Chen. 2016. Block2Vec: A Deep Learning Strategy on Mining Block Correlations in Storage Systems. *2016 45th International Conference on Parallel Processing Workshops (ICPPW)* (2016), 230–239.

*SYSTOR '18, June 4–7, 2018, HAIFA, Israel*
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-5849-1/18/06…$15.00
https://doi.org/10.1145/3211890.3211909