Poster Abstract: Attentive Multimodal Learning on Sensor Data using Hyperdimensional Computing

Quanling Zhao
Computer Science and Engineering
University of California San Diego
La Jolla, California, USA
quzhao@ucsd.edu

Xiaofan Yu
Computer Science and Engineering
University of California San Diego
La Jolla, California, USA
x1yu@ucsd.edu

Tajana Rosing
Computer Science and Engineering
University of California San Diego
La Jolla, California, USA
tajana@ucsd.edu

ABSTRACT

With the continuing advancement of ubiquitous computing and various sensor technologies, we are observing a massive population of multimodal sensors at the edge which posts significant challenges in fusing the data. In this poster we propose *MultimodalHD*, a novel Hyperdimensional Computing (HD)-based design for learning from multimodal data on edge devices. We use HD to encode raw sensory data to high-dimensional low-precision hypervectors, after which the multimodal hypervectors are fed to an attentive fusion module for learning richer representations via inter-modality attention. Our experiments on multimodal time-series datasets show *MultimodalHD* to be highly efficient. *MultimodalHD* achieves 17x and 14x speedup in training time per epoch on HAR and MHEALTH datasets when comparing with state-of-the-art RNNs, while maintaining comparable accuracy performance.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; • **Computer systems organization** → *Embedded systems.*

KEYWORDS

Hyperdimensional Computing, Multimodal Learning.

ACM Reference Format:
Quanling Zhao, Xiaofan Yu, and Tajana Rosing. 2023. Poster Abstract: Attentive Multimodal Learning on Sensor Data using Hyperdimensional Computing. In *The 22nd International Conference on Information Processing in Sensor Networks (IPSN '23), May 9–12, 2023, San Antonio, TX, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3583120.3589824

1 INTRODUCTION

Along with the development of ubiquitous computing and sensing technologies, an increasing number of sensor types are integrated into one edge device to detect and collect richer data in the environment. For example, to monitor human activities, various types of sensors such as gyroscopes, accelerometers, or even EEG sensors are used. However, the multimodal nature of data presents unique challenges to processing and learning, especially from two aspects:

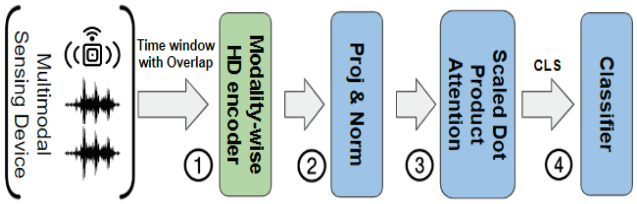

Figure 1: Overview of *MultimodalHD* with 4 stages: (1) A HD encoder to extract mod-wise sequence information. (2) Projection for feature extraction and dimensionality reduction. (3) Attention based multimodal fusion. (4) Classification on fused representations.

(i) how to exploit the multi-modality to facilitate better learning outcomes and (ii) how to process large quantity of data efficiently on resource-constrained embedded devices. Previous work on learning from temporal sensor data largely relies on the use of LSTMs [6]. However, due to the recurrence and sequential computation inherent to RNNs, it is extremely computational intensive and inefficient to train.

Hyperdimensional Computing (HD) is a novel brain-inspired lightweight computing paradigm where all data are first encoded into high-dimensional low-precision hypervectors (e.g., a 10K-bit binary vector). Then, associative learning is performed on those hypervectors through hardware-efficient and easily parallellizable operations, such as binding and bundling. HD with specialized hardware has been reported to be 3.6x-223x faster than the RTX 3090 GPU with learning performances comparable to conventional NNs [5]. We aim to combine the efficiency of HD and the capability of deep learning to learn from multimodal sensor data, excelling in both effectiveness and efficiency.

In this poster, we propose a new HD-based method named *MultimodalHD* to efficiently learn from multimodal data at the edge, with the complete flow shown in Fig. 1. *MultimodalHD* utilizes a static permutation-based HD encoder to effectively encode time-series data from different sensor types into hypervectors with uniform dimension. Unlike previous work that directly bundles the hypervectors of different modalities [3, 9], we propose an attention module that learns intermodality correlations and creates richer representations for downstream classification.

2 METHOD OVERVIEW

The proposed *MultimodalHD* has two main components: (1) a static modality-wise HD encoder that requires no training, and (2) an attentive multimodal fusion neural network that extracts information-rich representations from multimodal sensor data.

IPSN '23, May 9–12, 2023, San Antonio, TX, USA

ACM ISBN 979-8-4007-0118-4/23/05.
https://doi.org/10.1145/3583120.3589824

Modality-wise HD encoder: We first use permutation-based HD encoding to efficiently extract information from sequential and temporal sensor data, bypassing conventional RNNs. The static HD encoder is shown as stage 1 in Fig. 1. The goal of the HD encoder is to map high-precision, low-dimensional real-valued sensor readings to lower-precision, high-dimensional hypervectors in the HD space. To achieve multimodal fusion in the later stage, we keep the different modalities separate during encoding.

Attentive Multimodal Fusion: Previous work on HD-based multimodal fusion directly bundled hypervectors from different modality, which implicitly assign equal weight to each modality [3, 7]. However, as demonstrated in recent literature [8], depending on the underlying physical activity, some modality might be more useful than others.

To address the issue, in *MultimodalHD*, we introduce an attention mechanism inspired from [10] in *MultimodalHD*. We apply standard scaled dot product attention (stage 3 in Fig. 1). However, due to the large dimensionality of hypervectors and its holographic property, it would be inefficient to use the standard dot product directly. Therefore, we apply a projection layer with normalization (stage 2 in Fig. 1) before the dot product attention to (i) reduce the dimensionality of input for subsequent attention computation and (ii) extract features from hypervectors of different modality. After projection, a learnable CLS token similar to ViT [4] is concatenated to the projected hypervectors.

3 EXPERIMENTS

Experimental Setup. We test *MultimodalHD* on two commonly used public human activity recognition datasets, HAR [1] and MHEALTH [2], both of which contain continuous multimodal sensor readings. The HAR dataset contains time-series accelerometer and gyroscope readings of 30 subjects performing 6 common daily activities. The MHEALTH dataset has 13 common daily activities, including data from accelerometer, gyroscope and magnetometer data. In this experiment, we use fixed time windows of 2.56s and split the dataset into individual multimodal time-series samples with 75% overlap.

We evaluate *MultimodalHD* in comparison with the traditional HD method that bundles the hypervector of different modalities [9], and LSTM. For LSTM based classifier, each modality is processed through an LSTM encoder and then concatenate together, after which a fully connected layer is used as classifier.

To evaluate performance and efficiency, we use the weighted F1 score and training time per epoch as our main evaluation metrics. All experiments were implemented using PyTorch running on an Intel Core i5 11400H CPU.

Results. As shown in Fig.2, *MultimodalHD* achieves comparable F1 scores to LSTM, while surpassing the traditional HD method with a significant margin even under a lower dimensionality. This demonstrates the effectiveness of our attentive multimodal fusion module. Additionally, *multimodalHD* achieves faster convergence on both datasets, which is a benefit of using HD encoder and learning on hyervectors. In terms of training time, the LSTM model, on average, uses 24 seconds on HAR and 46 seconds on MHEALTH per training epoch. *MultimodalHD* only uses 1.4 and 3.1 seconds, achieving a speedup of 17x and 14x respectively.

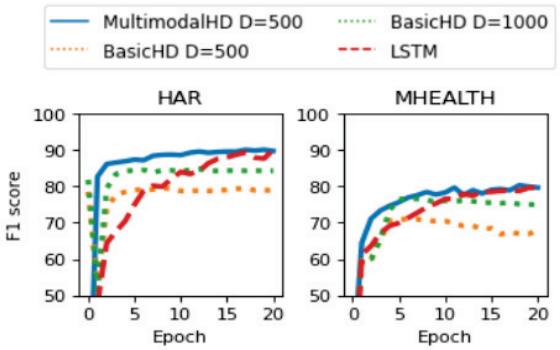

Figure 2: Accuracy results of *MultimodalHD*.

4 CONCLUSION

In this work we proposed *MultimodalHD*, a novel design for efficient and accurate learning on multimodal sensory data. *MultimodalHD* uses HD encoder and an attention mechanism to achieve multimodal fusion across different sensor modalities, which significantly reduces training time while maintaining comparable accuracies to state-of-the-art LSTM. We believe that HD is promising for multimodal learning at the edge, and we are actively working on expanding *MultimodalHD* to a federated setting.

ACKNOWLEDGMENTS

This work was supported in part by National Science Foundation under Grants #2003279, #1826967, #2100237, #2112167, #1911095, #2112665, and in part by SRC under task #3021.001. This work was also supported in part by PRISM and CoCoSys, centers in JUMP 2.0, an SRC program sponsored by DARPA.

REFERENCES

[1] Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra, Jorge Luis Reyes-Ortiz, et al. 2013. A public domain dataset for human activity recognition using smartphones.. In *Esann*.
[2] Oresti Banos, Rafael Garcia, Juan A Holgado-Terriza, Miguel Damas, Hector Pomares, Ignacio Rojas, Alejandro Saez, and Claudia Villalonga. [n. d.]. mHealthDroid: a novel framework for agile development of mobile health applications. In *IWAAL 2014*.
[3] En-Jui Chang, Abbas Rahimi, Luca Benini, and An-Yeu Andy Wu. 2019. Hyperdimensional computing-based multimodality emotion recognition with physiological signals. *AICAS*.
[4] Alexey Dosovitskiy et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* (2020).
[5] Arpan Dutta, Saransh Gupta, Behnam Khaleghi, Rishikanth Chandrasekaran, Weihong Xu, and Tajana Rosing. [n. d.]. Hdnn-pim: Efficient in memory design of hyperdimensional computing with feature extraction. In *Proceedings of the Great Lakes Symposium on VLSI 2022*.
[6] Isibor Kennedy Ihianle, Augustine O Nwajana, Solomon Henry Ebenuwa, Richard I Otuka, Kayode Owa, and Mobolaji O Orisatoki. 2020. A deep learning approach for human activities recognition from multimodal sensing devices. *IEEE Access* (2020).
[7] Yeseong Kim, Mohsen Imani, and Tajana S Rosing. 2018. Efficient human activity recognition using hyperdimensional computing. In *Proceedings of the 8th International Conference on the Internet of Things*.
[8] Xiaomin Ouyang, Xian Shuai, Jiayu Zhou, Ivy Wang Shi, Zhiyuan Xie, Guoliang Xing, and Jianwei Huang. 2022. Cosmo: contrastive fusion learning with small data for multimodal human activity recognition. In *Mobicom*.
[9] Kenny Schlegel, Florian Mirus, Peer Neubert, and Peter Protzel. 2021. Multivariate time series analysis for driving style classification using neural networks and hyperdimensional computing. In *2021 IEEE Intelligent Vehicles Symposium (IV)*.
[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *NIPS* (2017).