基於多階層 ReRAM 的壓縮超維向量運算

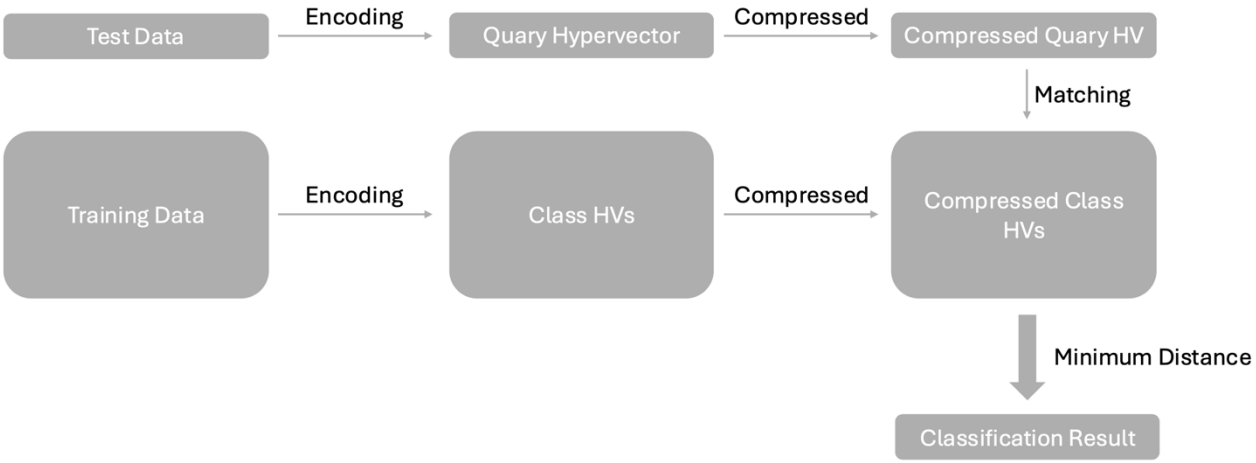
Hyperdimensional Computing with Compressed Hypervectors on Multi-Level ReRAM

指導教授：謝昀珊
專題成員：柯兆陽
開發工具：Python 3.11.5, C
測試環境：Ubuntu 24.04.2 LTS, NueroSIM

一、 簡介：

本專案針對高維計算（Hyperdimensional Computing, HDC）於實際應用中面臨的儲存與運算負擔進行優化。HDC 藉由高維超向量（hypervector）進行資料表徵及處理，具備高容錯性與運算簡化特性，然而大規模的向量維度（常見的維度為數千至數萬）導致需要較大的儲存需求和能耗開銷，不利於記憶體資源受限的邊緣裝置或低功耗環境使用。

在傳統運算架構下，資料需要不斷在記憶體與 CPU 間搬遷，在 HDC 的情況中更為顯著，因超向量的維度極高而造成大量的記憶體讀寫與資料移動，不僅會增加延遲，也大幅提升能耗。在實際運行上，搬運資料的耗能會因此遠高於運算本身。為解決此問題，本專題提出一種導入 Multi-Level ReRAM 的 PIM（Processing-in-Memory） 架構，系統架構如下圖一，將超向量的壓縮與比對運算直接於 ReRAM 記憶體陣列中完成，在不顯著犧牲推論準確率的前提下，可有效減少資料搬遷所造成的效能問題，並降低硬體複雜度。



圖一： System Flow Overview

本系統架構包含三大核心：

1. HDC 編碼：將輸入資料（如影像像素）透過隨機編碼與位置映射，轉換為高維 bipolar 超向量。

2. 向量壓縮與映射：採用 4 bit 多態態元（multi-level cell）設計，對原始 bipolar 向量進行量化與 壓縮，使其可直接對應至 ReRAM 儲存狀態。

3. PIM 架構：將模型向量預寫入 ReRAM 陣列，推論階段則透過輸入向量與儲存向量在 Crossbar 產生類比電流，並以 ADC 加總輸出相似度分數，完成向量內積計算。推論過程以 NeuroSim 模擬平台進行，完整考慮子陣列結構、記憶體特性、誤差與元件雜訊等物理參數。

二、 測試結果：

為了驗證本專題所提出之 Multi-Level ReRAM-PIM 架構的推論效能，選用三個公開資料集進行實驗：MNIST（影像分類）、ISOLET（語音辨識）與 Beijing Multi-Site Air Quality（環境感測資料分類）。這些資料集涵蓋視覺、語音與時間序列等不同領域，可有效評估系統的通用性與穩定性。

實驗結果顯示，在將原始 bipolar 超向量量化為 4-bits 表示後，模型整體推論準確率僅出現約 1-3% 以內的下降幅度，證實本架構具備良好的精度保持性：

Dataset	1 bit	2 bits	4 bits
MNIST	80.99%	80.95%	80.73%
ISOLET	82.88%	82.50%	82.69%
Beijing Air Quality	67.60%	65.20%	64.00%

表一：實驗結果

儘管採用了 4-bit 的量化設計，本架構在多種資料型態與應用場景下仍可保持與高精度模型相當的準確率，準確度下降幅度控制在 1-3% 以內，驗證其在低功耗環境下的可行性與實用價值。