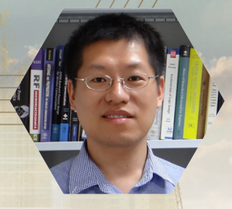
設計組
Design Group

D18-008

應用於癲癇偵測具有即時支持向量機學習核心之低功耗處理器

An SVM Processor with On-Chip Active Learning for Closed-Loop Epileptic Seizure Control

隊伍名稱　羅密歐與傅立葉 Romeo and Fourier
隊　　長　黃碩安 / 臺灣大學電子工程學研究所



指導教授

楊家驤
臺灣大學電機工程學系

美國加州大學洛杉磯分校電機博士，現為臺灣大學電機工程學系副教授。曾任教於交通大學電子工程學系。

研究領域

生醫訊號處理器、基頻通訊積體電路、與機器學習處理晶片設計。實驗室致力於開發低功耗之客制化晶片以提升資料處理速度與能量效率。

作品摘要

創作動機

支撐向量機（Support vector machine, SVM）為監督式學習的分類器，在預測推論（inference）上因權重稀疏性具有低運算量的特性，經常運用於穿戴式裝置或端運算上多種應用來滿足低功耗的需求，如初步圖像分類、生理疾病偵測、動態追蹤等。現今文獻中一般是將已訓練好的權重置入硬體裝置，但這樣的方式在動態生理訊號中（如腦波訊號）並不適當，因隨時間巨量變化的訊號會導致訓練好的模型無法得到良好效果。因此，根據最新數據進行模型更新有其必要性，但在晶片外訓練模型會有安全性與資料流失風險。由於晶片上訓練比起推論的複雜度增加許多，文獻上尚未有具有SVM訓練功能之晶片。本研究提出之SVM晶片可同時實現訓練與推論，並大幅降低訓練複雜度以達到低功耗需求與縮短運算時間。此晶片應用於癲癇偵測，可將最新的腦波資料進行即時訓練來更新偵測模型，進而提高分辨精度與降低假警報率。硬體架構與晶片實現技巧可應用於物聯網裝置，在裝置上蒐集資料、訓練並作適應性的更新，則可於終端裝置進行模型訓練，進而降低傳輸功耗與縮短反應時間。

設計特點

SVM在硬體上的訓練複雜度與計算延遲比起推論來的複雜許多，因此我們針對三種層面來進行最佳化設計：

- 演算法運算複雜度最佳化
  1. 訓練演算法採用ADMM最佳化，同時可以達到高度平行計算，並可以達到快速收斂的特性，比起常用的SMO加快78%。
  2. 採用mRMR特徵選取、低秩近似演算法，達到99.4%的運算複雜度降低與90.4%記憶體空間減少。
- 硬體面積最佳化
  1. 在腦波特徵擷取上，使用256點複數FFT同時利用實數與複數的資料完成512點實數FFT，減少34%的使用面積。
  2. 設計多功能的CORDIC-based運算單元，單一運算單元可支援六種不同的線性與非線性運算，節省46%硬體面積。
  3. 特徵值分解使用硬體摺疊來達到硬體共用，可以節省20%硬體面積。
- 硬體計算延遲最佳化
  1. 採用Approximate Jacobi method來達到特徵值分解，比起原先的cyclic Jacobi rotation減少89%的運算延遲時間。
  2. 使用8x的平行化運算，特徵值分解計算時間加快87.5%。

晶片實現成果

本作品以台積電40奈米CMOS製程實現，晶片面積為4.53mm^2，邏輯閘數為3.8M，操作頻率為130KHz。晶片操作在0.58V，晶片在SVM訓練與推論的功耗分別為2.9mW與1.9mW。SVM訓練與推論的計算延遲為0.78秒與0.71秒。此晶片具有以下特點：

1. 文獻上第一顆晶片可以同時支援SVM晶片上的訓練與推論。
2. 達到即時訓練與即時推論，兩者延遲均小於1秒鐘。
3. 訓練與推論的功耗小於3mW，可應用於低功耗的應用。
4. 應用於癲癇偵測，適應性的更新模型可降低50%的假警報率。
5. 訓練能量效率比使用高階CPU（Intel i7-2600）降低1.5x10^{5}倍、單位面積吐吞量增加364倍。

Abstract

Motivation

Support vector machine (SVM) is a supervised-learning-based classifier that features sparse weights and low computational complexity in inference. SVM has been deployed in many edge devices for low-power applications, such as image classification, melody detection, and motion tracking. Conventionally, a pre-trained model is loaded for the entire succeeding inference procedure. However, such an approach fails to support robust detection performance for highly-dynamic signals, such as EEG. Therefore, model adaptation, which adjusts the weights dynamically, is required and crucial for such applications. Off-chip model adaptation has been proposed to tackle the problem by re-training the model off-line. However, data have to be transmitted wirelessly to the outside body, introducing the risk of data loss and security issue. This makes on-chip model adaptation necessary. In this work, we propose an SVM chip that can support both training and inference. The chip achieves both low power and low latency through significant complexity reduction. Used in epileptic seizure detection, the detection rate and false alarm rate are improved by applying on-chip adaptation. The proposed SVM processor can be used in many IoT devices that demand realtime, energy-efficient classification along with dynamic adaptation.

Design Features

The computational complexity of on-chip SVM training is much higher than that of on-chip SVM inference. To reduce the complexity, optimization is performed across from three different aspects:

- Computational complexity
  1. An alternating direction method of multipliers (ADMM) optimizer is exploited because of its highly-parallel structure and fast convergent rate.
  2. Minimum redundancy and maximum relevance (mRMR) feature selection and low-rank approximation reduce 99.4% of computational complexity and 90.4% of memory storage requirement compared to the direct-mapped design.
- Hardware complexity
  1. For EEG feature extraction, a 256-point complex-valued fast Fourier transform is used to efficiently realize the 512-point real-valued FFT by fully utilizing two data streams, resulting in 34% of hardware area reduction.
  2. Reconfigurable CORDIC-based processing unit is designed to support six distinct linear and non-linear functions, saving 46% of hardware area.
- Processing latency
  1. Approximate Jacobi method is adopted in eigenvalue decomposition (EVD) to improve the convergence speed compared to original cyclic Jacobi method, resulting in 89% latency reduction.
  2. 8x parallel computing for disjointed rows and columns further reduces 87.5% of latency.

Chip Implementation

Designed and fabricated in TSMC 40nm CMOS technology, the chip integrates 3.8M gates in an area of 4.53mm^2. The proposed processor operates at a clock frequency of 130KHz from 0.58V. The power dissipation for SVM training and inference are 2.9mW and 1.9mW, respectively. The latency for SVM inference is 0.71s and the model can be retrained within 0.78s. The technical breakthroughs and contributions of this work are summarized as follows.

1. This work demonstrates the first SVM-based seizure detector for both detection and adaptation.
2. The chip achieves both SVM training and inference, each with a latency less than 1 second.
3. The power consumption of the chip is less than 3mW for both training and inference, allowing for low-power applications.
4. The false alarm rate is reduced by 50% through model adaptation for epileptic seizure detection.
5. The energy efficiency and throughput-to-area ratio of this chip are respectively 1.5x10^5 and 364 times higher than a high-end CPU (Intel i7-2600).