基於無梯度通道重要性評估之通道剪枝策略，應用於 OS2D 單樣本物件偵測模型之研究

專題編號：114-1-CSIE-S011

執行期限：113 年第 1 學期至 114 年第 1 學期
指導教授：陳彥霖教授
專題參與人員： 111820006 陳羿錦

壹、摘要

本研究旨在透過定位感知通道剪枝（LCP），結合創新的無梯度通道重要性評估方法，實現高效的 OS2D 物件偵測模型壓縮。

傳統剪枝技術在評估通道重要性時，需要大量梯度計算，導致 GPU 記憶體需求暴增並常發生溢出，使得大型模型壓縮難以執行。為克服此問題，本研究提出一套無梯度評估框架。此框架將計算複雜度從 O (NCHW·L) 大幅降至 O(CHW)，實現數千倍的運算加速，並將記憶體需求降低至僅需前向傳播，徹底解決大型模型剪枝的記憶體限制。

研究選用 OS2D (One-Stage One-Shot Object Detection) 作為基礎框架，此模型具備單張範例影像的快速學習能力。我們深入分析 LCP 定位感知理論，實現包含分類與回歸損失的完整數學模型，並建構了完整的模型壓縮流程，包含通道選擇、BN 同步剪枝、殘差連接處理與模型自動重建。

透過 GroZi-3.2K 資料集評估，實驗結果顯示，經過剪枝與微調後，模型仍能維持一定的準確度與學習能力，驗證了本技術的有效性。本研究期望為資源受限的邊緣運算平台，提供一個輕量化的物件偵測解決方案。

關鍵字：單樣本物件偵測、通道剪枝、無梯度近似、定位感知壓縮、輕量化

貳、緣由及目的

隨著無人機與自駕車技術的發展，對於在邊緣設備上運行的輕量化物件偵測模型需求日增。傳統深度學習模型如 YOLOv5 雖準確，但龐大的參數與對大量標註資料的依賴，使其難以部署。OS2D (One-Stage One-Shot Object Detection) 應運而生，解決了資料需求問題，能透過單張範例影像學習新物件，但模型本身仍舊龐大。為此，LCP (Localization-aware Channel Pruning) 技術被視為有效的模型壓縮方案，但其在實際應用上存在嚴重的 技術瓶頸。

LCP 原始理論需透過梯度計算評估通道重要性，導致記憶體需求暴增，不符實際研發環境。此問題源於 LCP 對反向傳播計算的依賴，使得這項技術因記憶體限制而無法實施。

本研究的核心目的，便是要解決這個困境。我們透過 數學理論推導，將原始的梯度計算公式，轉換為一套無梯度評估框架，消除對龐大記憶體的依賴。

藉由這項技術突破，本研究期望為無人機巡檢等邊緣運算應用，提供一個高效且記憶體友善的模型輕量化解決方案。

參、使用技術方法

為解決傳統 LCP 需大量 GPU 記憶體的瓶頸，我們將其通道重要性評估公式，透過數學推導轉換為 無梯度統計量，利用通道的 L1 範數、方差與稀疏性等特徵，來取代複雜的反向傳播運算。此方法大幅降低了計算複雜度與記憶體需求使大型模型剪枝在一般硬體上成為可能最後，我們設計完整的剪枝與微調流程確保模型在壓縮後仍能維持良好的偵測準確率。

肆、架構流程

本研究提出一個整合 OS2D 單樣本物件偵測與無梯度 LCP 剪枝的輕量化框架，整體流程由三大模組串接而成。

1. OS2D 單樣本物件偵測開發系統：此模組以預訓練的 OS2D 模型為基礎，能夠透過單張範例影像，快速學習並識別新物件。其核心流程包含特徵提取、餘弦相似度計算，以及後處理的非極大值抑制 (NMS)，確保模型在不需大量標註資料的情況下，具備高效偵測能力。

2. LCP 剪枝與無梯度通道重要性運算框架：此模組旨在將模型輕量化，並透過建立無梯度通道選擇器解決傳統剪枝的記憶體瓶頸。此框架包含 區域特徵對齊、LCPPruner 工具，並設計專用的 微調損失函數，確保剪枝後模型仍能維持良好的分類與定位準確性。

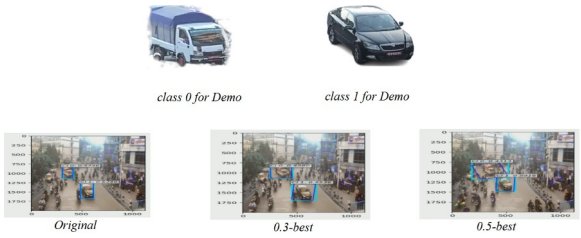
3. 實驗驗證與 Demo 系統：本模組透過 GroZi-3.2k 資料集 進行量化實驗，驗證不同剪枝率下的模型效能。此外，更開發 React + Python 的網頁版 Demo 系統，模擬實際應用場景，展示剪枝後模型在即時與非即時影片偵測中的可行性。

伍、實驗結果

本研究基於 Grozi-3.2k 資料集，對所提出的無梯度通道剪枝策略進行了嚴謹驗證。實驗結果顯示，相較於原始模型，剪枝率為 30% 的模型在經過微調後，其 mAP@0.50 達到了 0.73，成功壓縮模型體積至 28.05 MB，同時保持了與原始模型（0.83 mAP）相近的偵測效能。即使在更極端的 50% 剪枝率下，模型仍能透過微調將 mAP 恢復至 0.60，證明了此無梯度運算框架在顯著輕量化模型的同時，仍能保有良好的偵測與學習能力。這項成果驗證了所提出的方法在資源受限環境下的可行性與實用性。

表一、實驗結果

模型編號	Model Size	mAP@0.50
Original	39.95 MB	0.83
30%	28.05 MB	0.73
50%	20.52	0.6



圖一、Demo 結果展示

陸、結論與展望

本研究提出一種無梯度運算剪枝策略，搭配逐層微調方法應用於 OS2D 模型，並於 GroZi-3.2k 資料集上驗證其於單樣本物件偵測任務中的可行性。實驗結果顯示，在 30% 剪枝率下，模型能同時兼顧準確率與壓縮效果，有效降低參數量並提升資源效率；即使在 50% 剪枝率下，透過適度的微調仍可維持一定的偵測能力，展現本研究方法於模型效能與資源限制之間取得平衡的潛力。此外本研究亦於多平台環境中進行驗證，進一步支持其跨平台應用價值。未來展望方面，將針對不同卷積層設計差異化的微調策略，並結合資料增強與正則化技術，以進一步提升高比例剪枝模型的穩定性；同時亦將探索更先進的 backbone 結構（如 MobileNet、EfficientNet、Vision Transformer），並擴展至更大規模的資料集與多樣化應用場景，以驗證該方法的通用性與實務應用潛力。

柒、參考文獻

[1] A. Osokin, D. Sumin, and V. Lomakin, “OS2D: One-Stage One-Shot Object Detection by Matching Anchor Features,” *in Proc. European Conference on Computer Vision (ECCV)*, 2020, arXiv preprint arXiv:2003.06800.

[2] Z. Xie, L. Zhu, L. Zhao, B. Tao, L. Liu, and W. Tao, “Localization-aware Channel Pruning for Object Detection,” *Neurocomputing*, vol. 403, pp. 161-175, 2020.

[3] P. Molchanov, S. Tyree, T. Karras, T. Aila, and J. Kautz, “Pruning Convolutional Neural Networks for Resource Efficient Inference,” in *Proc. International Conference on Learning Representations (ICLR)*, 2017.

[4] Z. Liu, J. Li, Z. Shen, G. Huang, S. Yan, and C. Zhang, “Learning Efficient Convolutional Networks through Network Slimming,” in *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2736-2744.

[5] X. Liu, J. Cao, H. Yao, and W. Sun, “AdaPruner: Adaptive Channel Pruning and Effective Weights Inheritance,” in *Proc. AAAI Conference on Artificial Intelligence*, 2021.