結構化剪枝於異常檢測模型(PatchCore)之優化

Structured Pruning-Based Optimization for Anomaly Detection: A Comparative Study on PatchCore

指導教授：許舒涵教授
專題成員：耿于恆 陳致希
開發工具：Python
測試環境：Linux NvidiaA4000

一、簡介：

本研究旨在透過結構化剪枝技術，探討其對異常檢測模型 PatchCore 中特徵萃取能力的影響，尤其是針對以 ResNet18 為 backbone 的情境。我們希望在大幅降低模型參數量與提升推論速度的同時，仍能保持高水準的異常檢測效能。本研究選用工業標準的異常檢測數據集 MVTec 作為評估基準，並使用 Anomalib 函式庫進行 PatchCore 實作。

我們採用包括 magnitude、batch-norm、growing-reg、layer-adaptive 及 group-norm 等五種不同的剪枝方式，分別進行結構化剪枝，並與未剪枝（origin）的模型進行比較分析。此外，我們還使用了 MobileNetV3 Small 作為額外的骨幹架構，並透過 magnitude pruning 方法進行消融實驗，以評估其在異常檢測中的表現。我們發現 MobileNetV3 Small 採用的 depthwise separable convolution 顯著降低了模型的計算量（MACs），並且在 MVTec 資料集上的表現略優於 ResNet18。剪枝後進行重新訓練，以恢復模型效能，並透過核心集合取樣技術，進一步壓縮 PatchCore 的記憶庫空間，以降低推論時的計算成本與記憶體使用量。

系統架構流程包含：選用 ResNet18 及 MobileNetV3 Small 作為 backbone；應用結構化剪枝技術於 backbone；使用 ImageNet 資料集進行重新訓練；將訓練好的剪枝模型整合至 PatchCore 中，透過萃取特徵並建構記憶庫進行異常檢測。在推論階段，我們進一步優化計算方式，以節省 GPU 記憶體使用。

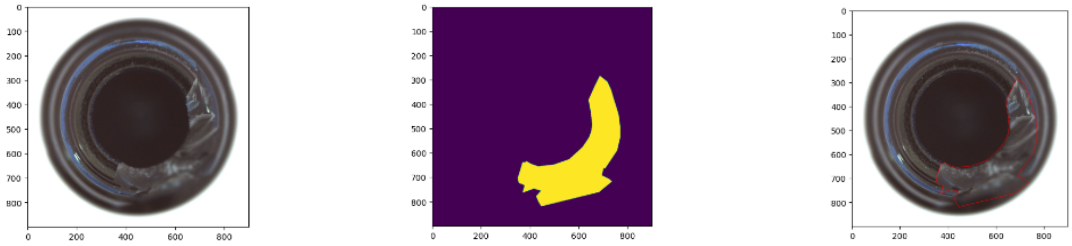
二、測試結果：

我們在 MVTec 數據集上測試不同剪枝方法與比例後的模型效能，結果顯示，適度剪枝比例（0.1-0.3）之模型與未剪枝模型的效能幾乎一致，而較高的剪枝比例（0.5-0.9）則會逐漸降低像素層級的異常檢測精確度。但整體而言，即使在大幅剪枝後，影像層級的檢測效能仍維持相當高水準，參數量明顯減少（從

11.69M 減少至最低 0.16M）。具體數據如下表所示：

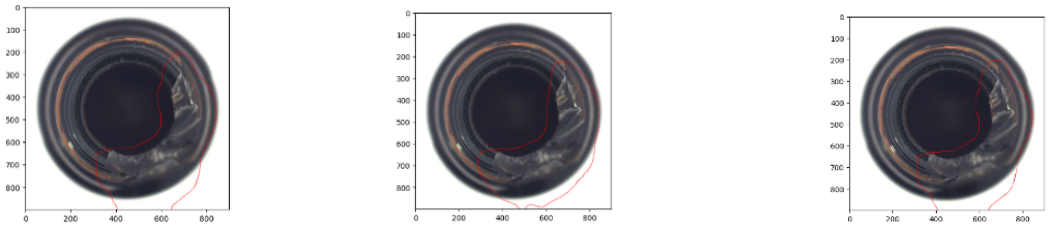
方法與剪枝比例	Image AUROC	Image F1-score	Pixel AUROC	Pixel F1-score	參數量
origin	1.0000	0.9920	0.9783	0.7120	11.69M
magnitude_0.5	1.0000	0.9920	0.9756	0.6934	3.06M
batch_norm_0.5	0.9783	0.9589	0.9654	0.6467	3.06M
growing_reg_0.5	0.9812	0.9609	0.9687	0.6495	3.06M
layer_adaptive_0.5	0.9804	0.9597	0.9678	0.6483	3.06M
group_norm_0.5	0.9795	0.9593	0.9667	0.6475	3.06M

圖一 ： 不同程度剪枝後 ResNet-18 骨幹網路的評估結果



(a) Original image (b) Ground truth mask (c) Ground truth image

Fig. 1: Imaged used for inference



(a) Pruning Ratio = 0 (b) Pruning Ratio = 0.3 (c) Pruning Ratio = 0.9

Fig. 2: Inference result for different pruning ratios

綜合上述結果，本研究驗證了 Lottery Ticket Hypothesis 的有效性，即模型中存在可透過結構化剪枝提取的優化子網路，在維持異常檢測高效能的同時，顯著降低參數數量與計算需求，特別適合資源受限之工業環境的部署。此外，MobileNetV3 Small 骨幹架構的使用，更進一步證實了 depthwise separable convolution 技術對於提高模型效率的有效性。