自編碼器在異常檢測的可靠度

Anomaly Detection Reliability of Autoencoders

指導教授: 許舒涵
專題成員: 白淳文
開發工具: Anaconda、Python、Pytorch
測試環境: Linux CentOS 7.9.2009 (Core)

一、 簡介：

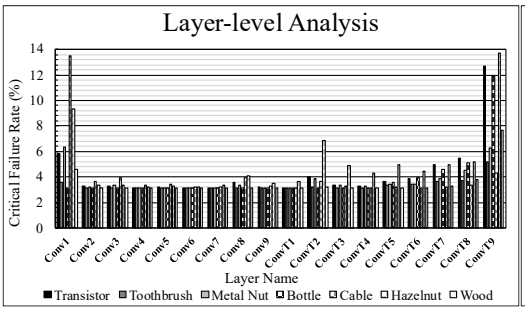
異常檢測是智慧製造中的一項關鍵任務，透過機器學習和數據分析技術，及時識別生產過程中的異常情況，例如定位缺陷產品，從而幫助企業大幅降低成本。然而，許多企業往往僅關注機器學習模型的準確性，卻忽略了模型的可靠性，這可能導致意外故障和損失。在推論過程中，機器學習模型所面臨的主要可靠性問題來自位元翻轉，也稱為單事件擾動（SEUs），這對電子硬體構成了關鍵的可靠性挑戰，可能由缺陷、輻射或瞬時擾動引起。單事件擾動通常會翻轉記憶體單元或邏輯電路中的單一位元，在深度神經網絡推理過程中可能引發計算錯誤，並損害模型的準確性，對敏感的嵌入式系統或關鍵設備（如航天器、醫療設備、自駕車等）造成重大影響，甚至可能導致系統不穩定或災難性故障。因此，我們在本研究中探索了深度學習模型在晶體管、牙刷、金屬螺母、寶特瓶、電纜、堅果、木頭七種物體異常檢測任務的可靠度，特別是對存儲訓練後網絡參數的權重記憶體的影響，參數數值形式以 IEEE-754 單精度浮點數儲存。

自編碼器（Autoencoders）是廣泛應用於異常偵測的模型，並且在數據壓縮和特徵提取中也扮演著重要角色。本研究中，我們使用基於軟體的錯誤（位元翻轉）注入方法來測量自編碼器的可靠性，並將該方法應用於評估自編碼器在工業數據集上進行異常偵測的抗故障性，從不同粒度層級進行分析。我們發現，自編碼器的抗故障性在模型的各層和各個位元位置之間差異較大。根據實驗結果，我們提供了一個實用的工作流程來評估和增強模型的可靠度，並在僅增加 3%的記憶體額外開銷的情況下，實現了平均 43%的可靠度提升。

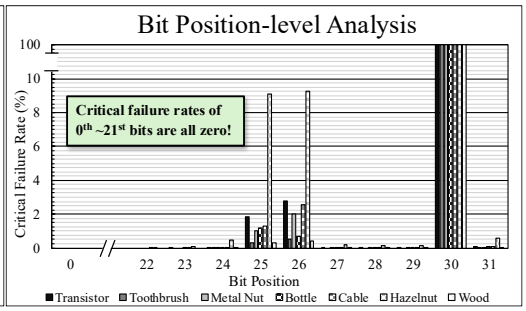
二、測試結果：

這份研究的主要貢獻有三。首先，我們驗證了統計性錯誤注入法在自編碼器的有效性，這個方法由[1]提出，旨在大幅降低錯誤注入所需的時間；在驗證其有效性後，我們將這個方法套用到異常檢測模型中，評估模型在七個物體上的異常檢測任務可靠度；最後，我們基於實驗結果，提出了一個實用的工作流程來評估和增強模型的可靠度。礙於篇幅限制，我們只在此份報告中說明我們的實驗結果。

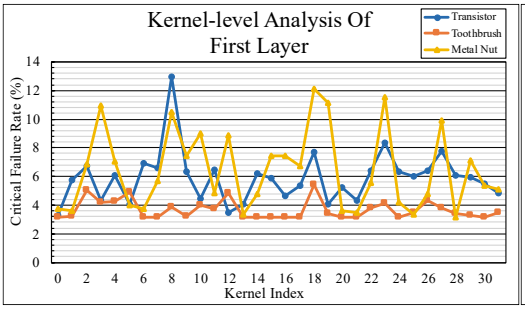
以下是我們的實驗結果，我們定義關鍵錯誤比率(Critical Failure Rate)為導致異常檢測產生錯誤結果的位元翻轉所佔的比例；例如，如果將500個位元翻轉單獨注入到特定層中，其中50個位元翻轉導致模型中異常檢測錯誤，那麼該層的關鍵錯誤比率為10%。圖一展示了各層關鍵錯誤比率，顯示了自編碼器在一和最後一層最為脆弱；圖二呈現了各個位元的關鍵錯誤比率，說明對應浮點數中負責儲存指數位的位元較為脆弱；圖三和四則呈現了各個模型第一和最後一層中各個卷積核的關鍵錯誤比率，顯示自編碼器在不同卷積核間也有顯著的可靠度差異。以上實驗結果顯示了針對性保護模型中的關鍵脆弱部分是必須的。



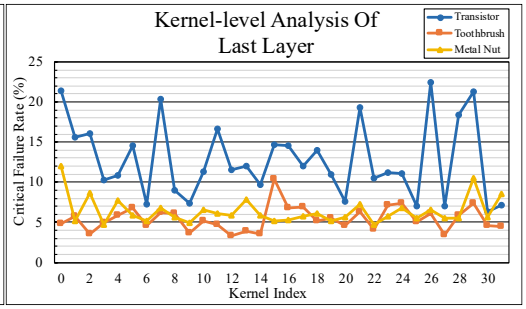
圖一



圖二



圖三



圖四

引用資料

[1] A. Ruospo *et al.*, "Assessing Convolutional Neural Networks Reliability through Statistical Fault Injections," in *Des. Autom. Test. Eur. Conf. Exhib. (DATE)*, 2023, pp. 1-6.