NARLabs
國家實驗研究院
National Applied Research Laboratories

經費來源：□01 公務 ■02 非公務

機密(E)：□是 ■否

分項計畫名稱：

財團法人國家實驗研究院
資安卓越中心規劃建置計畫
結案報告書

編　　號　：

成果報告適用期間：111.06.01~111.12.15

研 究 生 姓 名 ：周信廷

指 導 教 授 ：游家牧 游家牧 111 年 11 月 27 日

計畫共同主持人：游家牧 游家牧 111 年 11 月 27 日

計 畫 主 持 人 ：　　111 年　月　日

NARLabs
國家實驗研究院
National Applied Research Laboratories

一、基本資料

計畫名稱	財團法人國家實驗研究院 資安卓越中心規劃建置計畫	
研究生	周信廷	
指導教授	游家牧	
就讀學校	學校：	國立陽明交通大學
	科系：	財務金融研究所資料科學組
	年級：	二
研究題目	Deepfake 偵測分析技術	
研究或相關學習事項摘要(50-200字內)	1. 建立深度學習模型分析各類 Deepfake 影片，學習影片中之換臉特徵，用於偵測影像或監視錄影是否受到 Deepfake 技術之攻擊。 2. 將建立之深度學習模型部署至自行架設之網站，以提供社會大眾使用。	

每月獎助金(研究津貼)	□ 博士 ■ 碩士	20,000	元	執行期間	111 年	6	月	1	日
					至				
					111 年	12	月	15	日止

二、 研究主題：Deepfake 偵測分析技術

三、 研究問題及先前研究調查比較(Survey)：

深度臉部偽造技術（Deepfake）是一種將目標人臉移植到影片中的原始人臉，造成侵犯版權、資訊混淆甚至造成公眾恐慌等嚴重問題的惡意技術。由於人臉包含了豐富的個人資訊，濫用 Deepfake 將成為一種威脅。

自從 Deepfake 日益進步並引起嚴重的社會問題以及國家安全問題，學者們積極發展深度臉部偽造檢測技術（Deepfake Detection）以對抗 Deepfake。

到目前為止，Deepfake 的檢測方法大致可以分為兩種。第一種主要關注影片單一幀中的缺陷與破綻。第二種考慮了時間相關特徵。然而，有一些方法主要針對 Deepfake 技術的非本質缺陷，例如異常眨眼或不同顏色的虹膜，反過來又刺激了 Deepfake 影片合成的進步。所以現在成為了 Deepfake 與 Deepfake Detection 的技術角力戰。

在本計畫上半年期間已經完成原訂的目標，建立好深度學習模型並且部署至網站。因此，下半年我們將目標著重在提升深度學習模型的泛用性。

如上所述，Deepfake 與 Deepfake Detection 的攻防進入白熱化階段，新的攻擊方法層出不窮，為了讓模型同時針對現有以及全新的 Deepfake 進行偵測，我們將連續學習（Continual Learning）[1]技術結合現有的 Deepfake Detection 模型。

一般的深度學習模型在訓練過程中，會逐漸忘記過去學習過的任務特徵。即原本有一個以任務一（現有的 Deepfake 技術）的訓練資料進行訓練的模型，以任務一的測試資料進行測試有 99% 的正確率，在沒有看過任務二（新的 Deepfake 技術）資料的情況下，以任務二的測試資料進行測試，會有 90%正確率。

在任務一訓練完成後，將同一個模型以任務二的訓練資料繼續進行訓練，若以任務二的測試資料進行測試，會有 97%的正確率。然而以任務一的測試資料再次測試時，正確率會大幅下降至 60%，此現象稱為災難性遺忘[2][3][4]。

連續學習是模型模仿人類在整個生命週期中不斷從資料流中學習、微調的能力，能夠在有效學習新任務的同時維持在歷史任務上的表現，以避免災難性遺忘的現象。連續學習是模型適應快速變化的現實情景的關鍵，對於實現真正的人工智慧十分重要。

四、 研究方法及步驟

Avalanche[5]套件是一款針對連續學習所設計出來的開源程式碼，本套件設計了各種連續學習中可能遇到的情境，並收錄了經典的連續學習演算法。

此套件主要分成 5 個部分：

1. Benchmarks: 負責將各式各樣的任務資料集轉換成連續學習中所需要的資料流。
2. Training: 收錄各種經典的連續學習演算法。
3. Evaluation: 提供各種深度學習訓練與測試時所需要用到的指標，以及連續學習領域中特定的衡量指標。
4. Models: 提供各種連續學習領域的經典深度學習模型。
5. Logging: 紀錄訓練過程的各種資訊。

我們將利用本套件進行連續學習的實驗，研究步驟如下：

1. 實驗主要分成四部分，實驗一二皆自行建立一個簡單的模型，並採用 Cifar10 資料集，其中有 50000 筆訓練資料與 10000 筆測試資料，資料總共有 10 類標籤。實驗三是要將 Avalanche 套件套用到目前現有的 Deepfake Detection 模型上。實驗四是要利用 Avalanche 套件結合不同的 Deepfake 資料集進行訓練。
2. 實驗一的每種方法各有 10 個**獨立的**深度學習模型，每個模型**實際輸入的訓練資料量逐漸增加**，例如：
   - Model_1 輸入 1 - 5000 筆訓練資料（獨立）
   - Model_2 輸入 1 - 10000 筆訓練資料（獨立）
   - …
   - Model_10 輸入 1 - 50000 筆訓練資料（獨立）
   - 測試資料皆為 10000 筆
3. 實驗二的每種方法各有 10 個的深度學習模型，每個新模

型都會**基於舊模型**繼續進行訓練，模型實際輸入的訓練資料量不變，**但曾經看過的訓練資料逐漸增加**，例如：

- Model_1 輸入 1 - 5000 筆訓練資料
- Model_2 輸入 5000 - 10000 筆訓練資料（基於 Model_1）
- …
- Model_10 輸入 45000 - 50000 筆訓練資料（基於 Model_9）
- 測試資料皆為 10000 筆

4. 實驗三會將 Avalanche 套件套用到目前現有的 Deepfake Detection 模型 DFDC_3 上，並觀察成效。
5. 實驗四會利用 Avalanche 套件結合不同的 Deepfake 資料集進行訓練，例如：CelebA、F2F、DFDC…等資料集。

五、 研究成果及效益

在實驗一中共有兩種方法，兩種方法各有 10 個模型（Model_1 到 Model_10），每個模型皆互相獨立，**實際輸入的訓練資料量逐漸增加**。

第一種方法 Data Aggregate 是由我們自行手動切割不同比例的 Cifar10 資料集並輸入到模型中。

第二種方法 Avalanche Naive，則是利用 Avalanche 套件將總資料量切分成 10 個情境（scenarios），並將每個 scenario 的資料輸入到模型中進行訓練。**在連續學習領域中，不同的 scenario 表示不同時間點流入模型的資料，前文提到連續學習是要讓模型模仿人類在整個生命週期中不斷從資料流中學習，而每個 scenario 即是整個生命週期中的不同時間點，可將每個 scenario 視為不同的任務。**

由於每個模型獨立，隨著每個模型輸入的資料量越來越多，測試正確率理當越來越高。在資料總量逐漸增加的情況下，Data Aggregate 測試正確率進步幅度為 15%，Avalanche Naive 測試正確率進步幅度為 14%。然而，我們觀察到測試正確率並非不斷提升，在 Model_4 到 Model_6 的測試期間，測試正確率有些微的下降。其中，Data Aggregate 方法中的 Model_10 理論上是所有方法的上限，如表一所示。

在實驗二中共有四種方法，每種方法皆是利用 Avalanche 套件中不同的演算法進行實驗，每種方法各有 10 個模型（Scenario_1 到 Scenario_10），每個新模型都會基於舊模型繼續進行訓練，**模型實際輸入的訓練資料量不變，但曾經看過的訓練資料逐漸增加**。實驗二沿用實驗一方法二 Model_10 之實驗設定，即總資料量為 50000 筆，切分成 10 個 scenarios，每個 scenario 有 5000 筆資料。

實驗二中第一種方法 Naive 是最簡單也最不有效的連續學習策略。第二種方法 CWRStar 會將模型結構中特定層的權重凍結，減少模型遺忘過去看過的資訊。第三種方法 Replay 會從記憶體中讀取資料重新輸入給模型進行複習。第四種方法 Gradient Episodic Memory（GEM）會把不同情境的資料存起來，重新輸入給模型進行複習。

實驗二中隨著模型曾經看過的訓練資料逐漸增加，測試準確率也有上升的趨勢，Naive 進步幅度 3.67%，CWRStar 進步幅度 2.74%，Replay 進步幅度 4.44%，GEM 進步幅度 2.92%。實驗二測試正確率的提升過程相對不穩定，有震盪的趨勢。**由於每個 scenario 在連續學習領域中被視為不同的任務，透過實驗二的結果可以發現，利用 Avalanche 套件進行連續學習後，模型的測試正確率並沒有大幅下降，有效避免了災難性遺忘的現象。**

表一：實驗一之測試正確率

	Model_1	Model_2	Model_3	Model_4	Model_5
Data Aggregate	45%	50%	53%	54%	57%
Avalanche Naive	37%	41%	42%	44%	42%
	Model_6	Model_7	Model_8	Model_9	Model_10
Data Aggregate	56%	57%	59%	59%	60%
Avalanche Naive	42%	48%	46%	47%	51%

表二：實驗二之測試正確率

	Scenario_1	Scenario_2	Scenario_3	Scenario_4	Scenario_5
Naive	48.07%	46.04%	46.69%	49.46%	49.24%
CWRStar	46.7%	48.61%	48.41%	48.64%	48.77%
Replay	48.09%	47.61%	50.61%	49.38%	50.38%
GEM	48.1%	48.46%	48.47%	48.31%	48.34%
	Scenario_6	Scenario_7	Scenario_8	Scenario_9	Scenario_10
Naive	51.03%	50.73%	50.38%	51.41%	51.74%
CWRStar	49.4%	49.27%	49.16%	49.29%	49.44%
Replay	50.06%	51.64%	52.39%	52.63%	52.53%
GEM	49.16%	50.75%	50.74%	49.53%	51.02%

六、 困難、突破、及未來規劃

Avalanche 套件提供了各種經典的連續學習演算法，且從實驗結果也可以觀察到測試正確率確實有提升。目前的困難是實驗三需要將 Avalanche 套件套用到現有的 Deepfake Detection 模型 DFDC_3 上，然而 DFDC_3 的訓練方法經過特殊設計，相較於實驗一二複雜許多，需要自行修改 Avalanche 套件中的底層架構，使其可以採用 DFDC_3 的訓練方法。如果其他現有的 Deepfake Detection 模型的訓練方法也經過特殊設計，則需要陸續將它們加入到套件的底層架構中。

未來預計會將實驗三四完成，使現有的 Deepfake Detection 模型模仿人類不斷從資料流中學習。

七、 參考文獻

[1] 黃浴（2020）。介紹幾篇 incremental/continual/lifelong learning 的綜述論文。取自 https://zhuanlan.zhihu.com/p/336250745

[2] Z.H. Shen（2021）。機器終身學習 (Life Long Learning, LL)_災難性遺忘(Catastrophic Forgetting)_李弘毅_ML2021#14。取自 https://medium.com/@ZH_Shen/%E6%A9%9F%E5%99%A8%E7%B5%82%E8%BA%AB%E5%AD%B8%E7%BF%92-life-long-learning-ll-%E7%81%BD%E9%9B%A3%E6%80%A7%E9%81%BA%E5%BF%98-catastrophic-forgetting-%E6%9D%8E%E5%BC%98%E6%AF%85-ml2021-14-92d208f1df5c

[3] 秀的博客（2019）。連續學習介紹。取自 https://xiuyuli.com/blog/continue-learning/

[4] Hung-yi Lee（2021）。【機器學習 2021】機器終身學習 (Life Long Learning, LL) (一) - 為什麼今日的人工智慧無法成為天網？災難性遺忘(Catastrophic Forgetting)。取自 https://youtu.be/rWF9sg5w6Zk

[5] ContinualAI（2021）。Avalanche: an End-to-End Library for Continual Learning。取自 https://github.com/ContinualAI/avalanche