WEI-CHIH CHANG Q36111281
March 20, 2023

Abstract

設計非監督式的深度學習模型應用於偵測影片轉換場景的時機，並在深度學習模型輸出的結果後加上低通濾波器，使得最後得到的影片轉場時機結果是不受模型偵測的高頻誤差影響。

1. Related Work

這次的資料分別是 1780 幀的氣候變遷新聞畫面(climate)、770 幀的影片剪輯軟體宣傳影片(ftfm)以及 1380 幀的國際新聞影片(news)。news 和 climate 的場景轉換基本都是直接切換到下個場景，場景轉換較為簡單，但這兩個資料在相同場景中都會有持續切換的文字跑馬燈。而 ftfm 資料則是運用了大量的轉場技巧，同時有大量的運鏡，因此在偵測影片轉換場景時機上會較前兩者困難。

而在這次副專案中，我希望使用深度學習的演算法來完成判斷場景變換的任務。若使用監督式學習的方式，需要大量有標註的訓練資料，但現今常見的開源資料庫都是針對影像辨識進行標註，若應用於場景變換的任務中，則可能出現 A 場景是 a 廠牌的汽車影像、B 場景是 b 廠牌的汽車影像，但用影像辨識資料集訓練出來模型卻將兩場景判斷為相同場景的情況發生。同時在使用基本尺寸的資料集進行訓練時會需要耗費大量的訓練時間以及硬體成本。

為此我採用非監督式學習的演算法完成任務，雖然該演算法同樣需要大量時間，但是該方法可以避免使用到外部資料集，只需要使用提供的三種影片即可完成場景變換的偵測。為了完成此任務，我透過非監督式學習的演算法對某一影片進行分類的訓練，讓模型學習如何對每一幀畫面分類，完成訓練後接著將同一影片輸入到模型中預測出模型對每一幀影像的預測分類，透過該分類我們就可以將不同分類的邊界視為影片轉換場景的邊界。但由於非監督式學習無法知道影片中有多少種場景，因此會發生不同分類對應到相同場景的情況發生，使得模型在相同場景的分類結果會在數個分類中震盪，為此我們在模型分類結果後加上低通濾波器減少上述問題的發生，最後再計算出場景變化的邊界。

2. Auxiliary Classifier GAN

在該任務中，我使用生成對抗網路(GAN)作為我的非監督式學習模型。該模型由兩個網路所構成，分別是生成網路 (Generative Network) 和鑑別網路 (Discriminating Network)，該模型的精華是利用生成網路生成結果並將其標註為偽，將來自資料集的輸入標註為真，將兩者合併成平衡的新資料集後交由鑑別網路進行監督式學習的訓練。

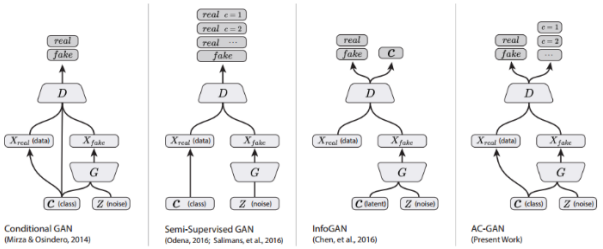


figure 1. 四種 GAN 架構圖(1)C-GAN (2)Semi-Supervised GAN (3)InfoGAN (4)AC-GAN。

而 AC-GAN 則是鑑別網路在判斷資料真偽的同時判斷該影像的類別，生成網路則被修改成透過輸入隨機向量和「類別」來生成影像，藉此來將影片中的不同場景進行分類。

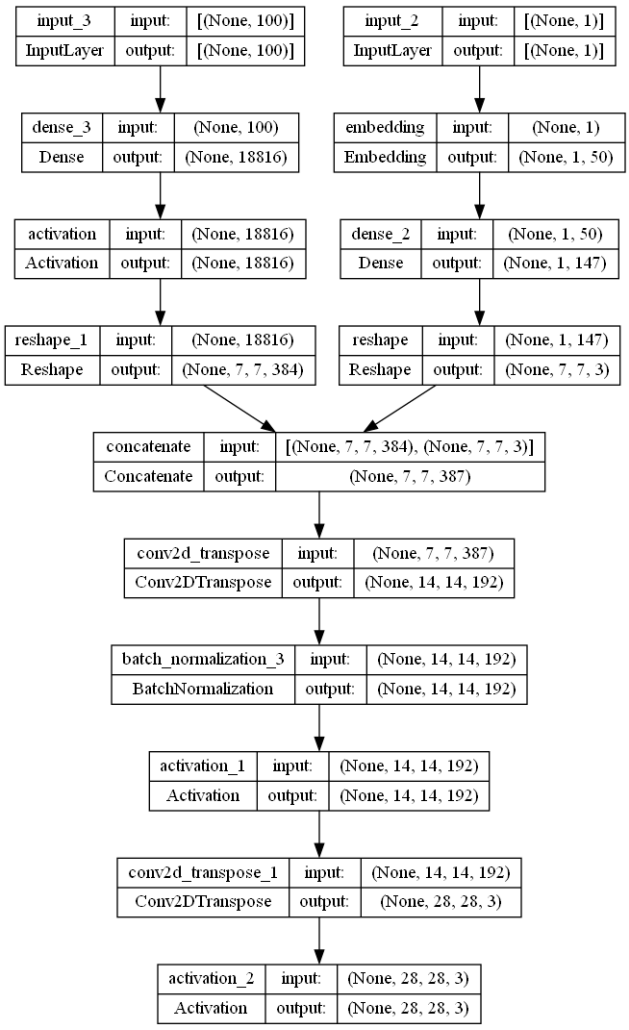


figure 2. 生成網路架構圖。

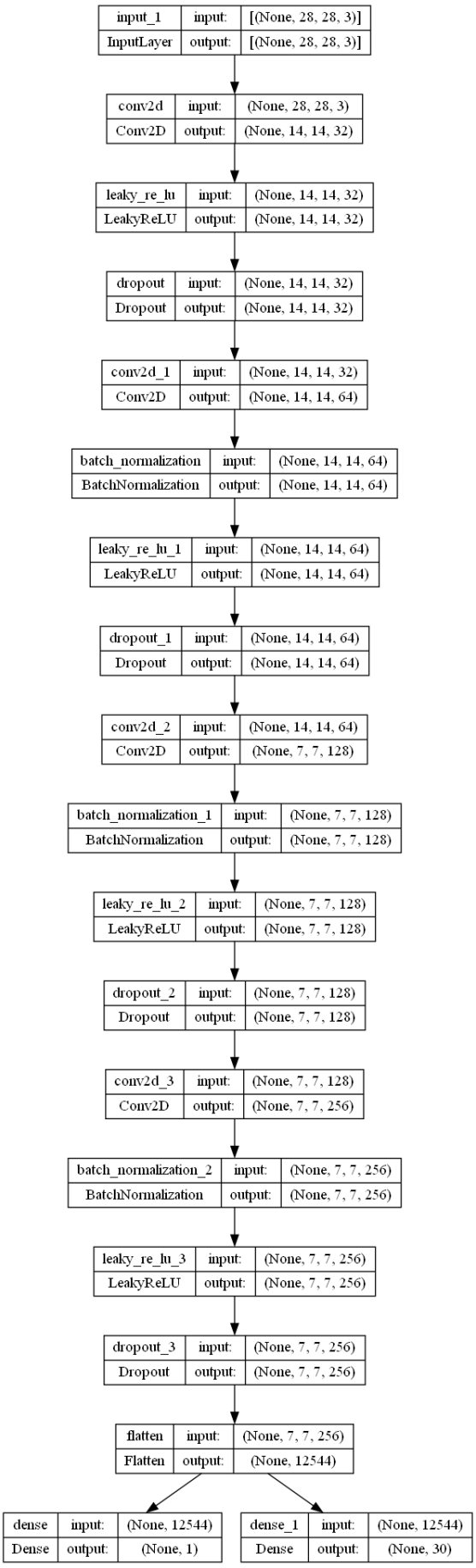


figure 3. 鑑別網路架構圖。

3. Visual Features

由於生成網路生成影像的原理是將 latent code 中取出的 Latent variable 經由轉置卷積轉換成能欺騙鑑別網路的影像，而鑑別器在鑑別不同尺寸的影像時，合理地在鑑別大尺寸的影像遠比小尺寸的影像來的容易。這就導致鑑別網路在學習鑑別大尺寸影像的學習效率過高，鑑別網路的 Loss 過早收斂至零，意味著鑑別網路幾乎能百分之百地辨別資料的真偽，使得生成網路連「偶然」都無法騙過鑑別網路，最終生成網路就會停止學習，因為生成網路無法從偶然的成功中學習到有用的權重。

因為轉置卷積生成網路的限制，我們在訓練模型前會將輸入的資料透過 cv2 函式庫，壓縮成 28*28 像素點的 RGB 影像，選用 RGB 影像而非灰階影像的原因是因為我們相信在鑑別網路輔助判斷資料類別時，色彩是其中一個重要特徵。雖然比起灰階，三通道的 RGB 影像會造成生成網路相較單通道的灰階影像較難訓練，但我們相信該改動能使 AC-GAN 在類別判斷上有較高的精準度。

4. Low-pass Filter for result of AC-GAN

AC-GAN 在判斷影片中某幀的的類別時，會因為將不同分類對應到相同類別的影像，從而導致出現高頻的震盪產生。為了有效解決該問題，我們在類別判斷結果後加上低通濾波器，從而使得類別的判斷變的平滑，最後再根據該結果判斷影格的轉換時機。該低通濾波器是採用前後類別頻率投票設計，根據該幀類別與前後 15 幀的類別進行類別的出現頻率進行排序，頻率最高的類別則為該幀類別。若該幀為類別 3，但前後幀出現頻率最高的類別是 12，則將該幀類別判斷修改為類別 12。

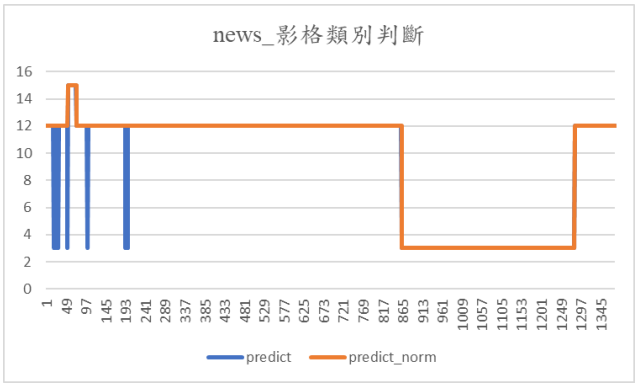


figure 4. news 影格資料的原始判斷與經過低通濾波器後的比較。

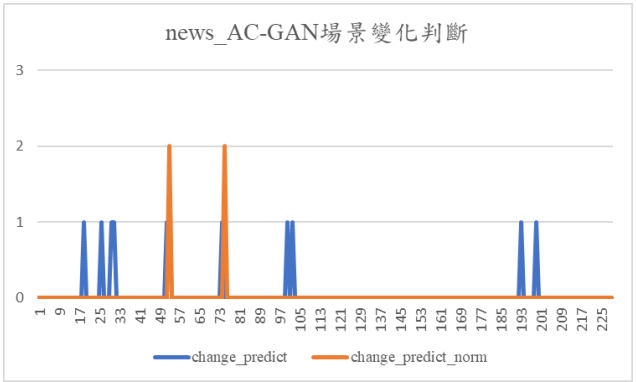


figure 5. 從圖中可以看出低通濾波器能有效消除高頻震盪。

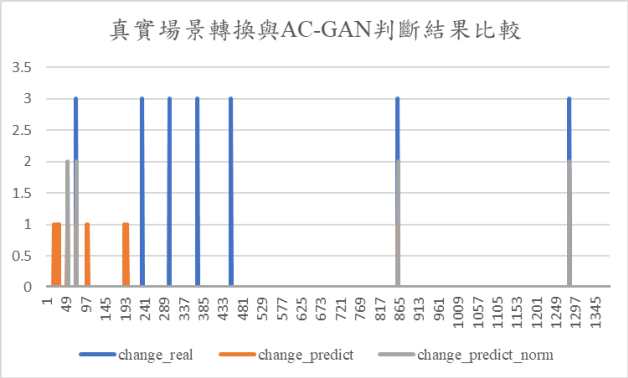


figure 6. 真實場景轉換與 AC-GAN 判斷結果比較。

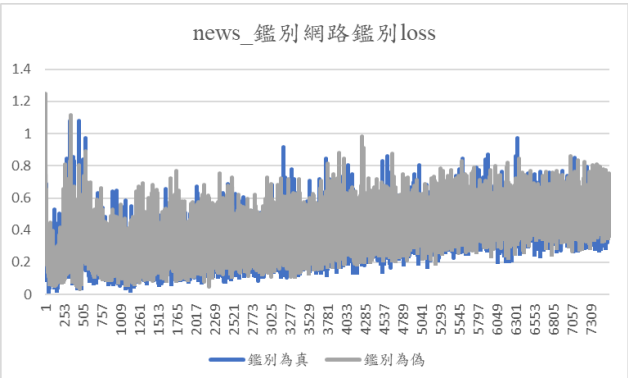
5. Performance

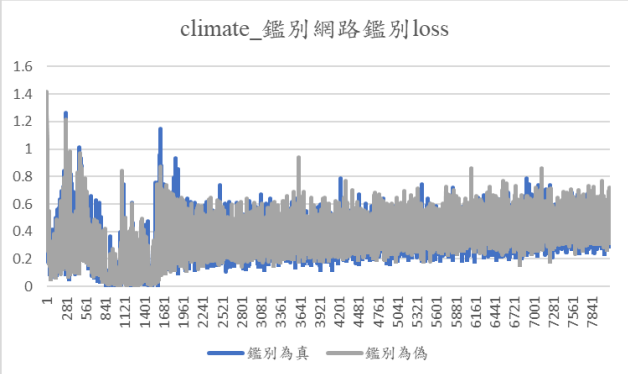
在分析場景轉換時間的執行效果時，我們分成模型學習、資料進化和場景判斷三者進行探討。

5.1 Model learning Performance

我們分別使用三個資料集分別進行 AC-GAN 的訓練，會發現在三者在鑑別網路鑑別 loss、鑑別網路分類 loss 和生成網路生成 loss 都有相同的趨勢。

鑑別網路在鑑別任務上 loss 都在 1 以下，但都無法有效收斂，這是因為鑑別網路的真偽判斷權重會傳給生成網路，用以讓生成網路生成更逼真的影像用以欺騙鑑別網路。





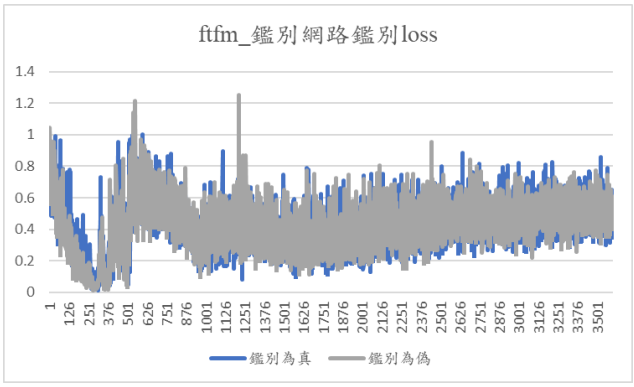
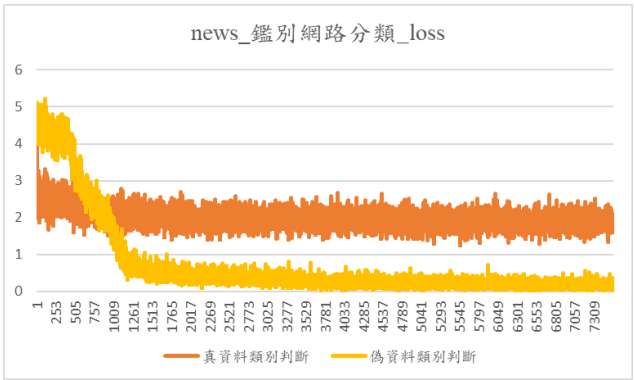
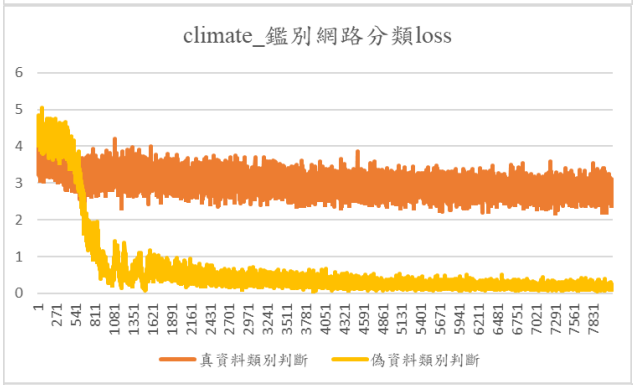


figure 7. 鑑別網路鑑別 loss (1)news (2)climate (3)ftfm。

在分類任務上，偽資料的分類效果遠比真資料的分類效果更好，我們推斷是因為偽資料是透過生成網路根據 Latent variable 與「類別」進行生成，而類別的特徵有機會在轉置卷積的過程中洩漏，從而導致鑑別網路比起真實資料更容易對偽資料進行分類。





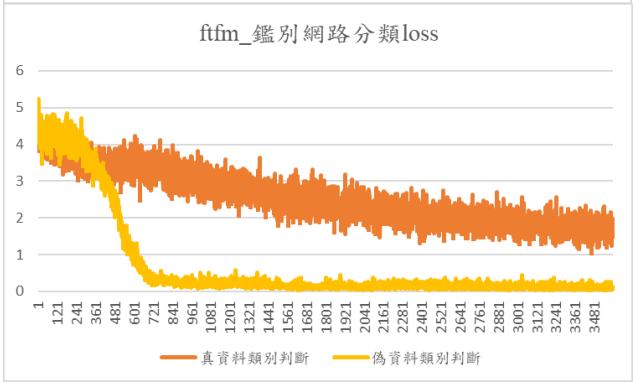


figure 8. 鑑別網分類 loss (1)news (2)climate (3)ftfm。

生成網路因為有部分權重來自鑑別網路的分類 loss，因此生成網路的 loss 趨勢和鑑別網路的分類 loss 成相關。

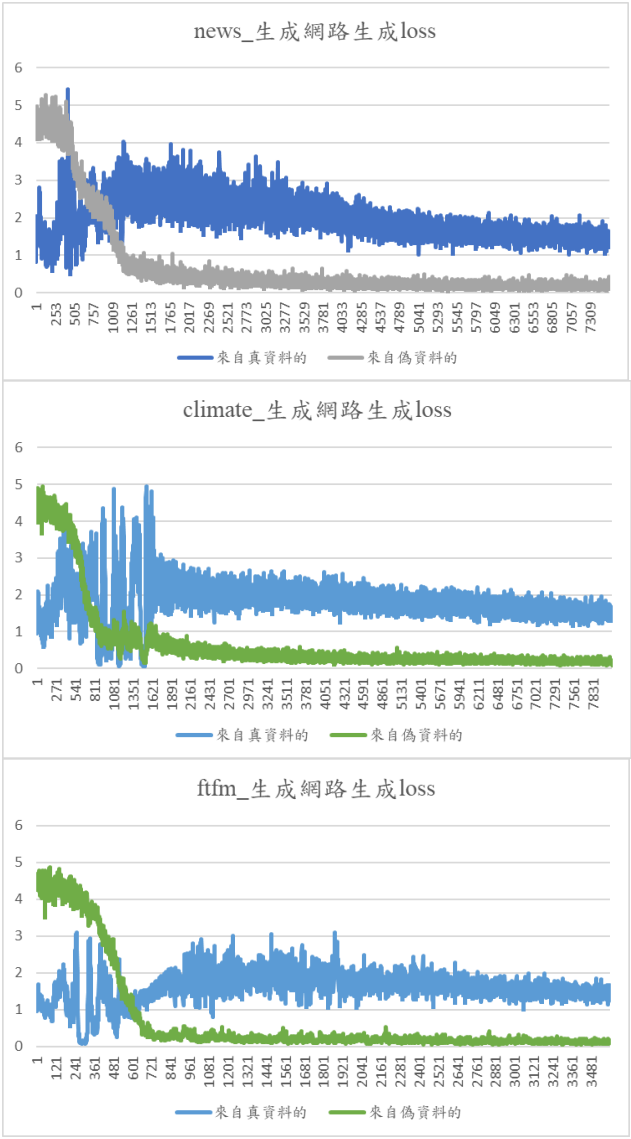


figure 9. 生成網路生成 loss (1)news (2)climate (3)ftfm。

5.2 Performance of Data Evolution from Generator

生成網路是透過類別和隨機向量作為初始值，根據 loss 來訓練轉置卷積的權重，從而達到從隨機向量生成有意義影像的效果。因此在最初的 epoch 生成網路生成的影像接近隨機噪音，但能看出影像下方的像素點代表新聞畫面的跑馬燈，隨著 epoch 的更迭，生成網路能夠生成出人像但分類上是錯誤的，最後到第 2940 個 epoch 時，生成網路生成的影像接近穩定，同時能夠人工判斷該分類生成的影像對應到實際資料的哪個場景。而在第 4620 個 epoch 時，生成的影像退化到與第 2520 個 epoch 的結果類似，這是因為輸入生成網路的初始值有非常高的比例來自隨機向量，在沒有對 Latent code 進行解耦前，若生成的隨機向量與分類不匹配，就會導致生成網路生成的影像同時帶有不同類別的特徵。

figure 10. 從左上到右下分別是 news 的生成網路在每 420 個 epoch 的生成結果，類別為 23。

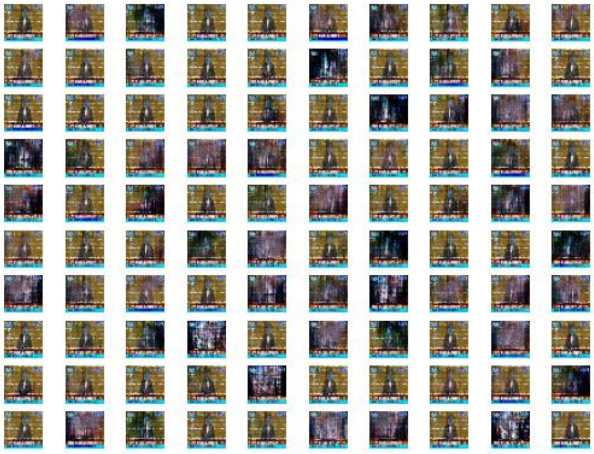


figure 11. 透過第 7560 個 epoch 的生成網路隨機生成 100 個類別為 0 的影像，可以發現約有八成的影像是符合類別的，但有兩成的影像是帶有其他類別特徵的。

5.3 Performance of Shot Change Detection

從圖表中可以看出，AC-GAN 在判斷 ftfm 的場景轉換時機是有最好的效果的，與人工標註時間的誤差差距不到一秒或是完全重疊。

然而從 climate 的場景轉換判斷結果，我們推論 AC-GAN 傾向將同一場景的長鏡頭的運鏡辨識為不同類別。climate 的最後 100 幀是長鏡頭的空拍運鏡，AC-GAN 就在這 100 幀的區間內，多次判斷場景轉換。又或是 news 開頭的幾幀，雖然沒有長鏡頭運鏡，但人物從柵欄中穿過，AC-GAN 也是判斷多次場景轉換。

從上述分析我們可以知道 AC-GAN 對影像的類別有極高的靈敏度，但同時這也代表 AC-GAN 在場景類別的辨識上容易有誤判。為了解決該問題我們可以在未來嘗試以下三種方法(1)將長短期記憶的概念添加到 AC-GAN 中 (2)將 AC-GAN 修改成能餵入更多像素點的影像 (3)對 Latent code 進行解耦。

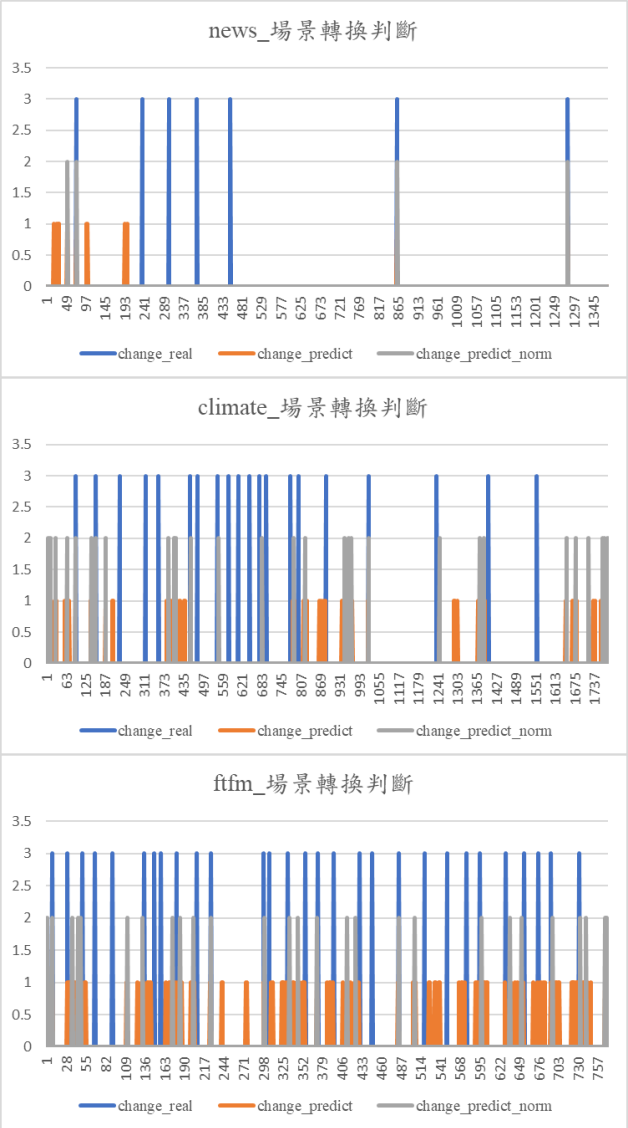


figure 12. 場景轉換判斷結果 (1)news (2)climate (3)ftfm。

6. Conclusion

會使用 AC-GAN 進行場景轉換的判斷，是因為 AC-GAN 能對場景進行分類，所以今天在遇到以下情境的影像時：兩個人在對話，場景切換到某個人的回憶，接著場景又切換回兩個人的對話。AC-GAN 就有能力辨別出雖然場景變化了兩次，但第一個場景和第三個場景是相同的場景。

本篇報告驗證 AC-GAN 確實能應用在影像的場景轉換上，但是精確度遠不如傳統根據前後幀的變化程度進行場景轉換辨識的方法，然而 AC-GAN 有著比起傳統方法更高的擴充性能，因此在未來若能根據 5.3 提出的三種方法進行模型的修正，讓 AC-GAN 的辨識精確度提高，降低誤判率，則 AC-GAN 就能在場景辨識上比肩傳統方法，甚至做到判斷出轉場種類或是辨識出兼用卡的能力，從而從相同的影像中分析出更多有用的特徵。

參考文獻

[1] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio: “Generative Adversarial Networks”, 2014; [http://arxiv.org/abs/1406.2661 arXiv:1406.2661].

[2] J. Brownlee Generative Adversarial Networks with Python: Deep Learning Generative Models for Image Synthesis and Image Translation Machine Learning Mastery, URL (2019) https://books.google.co.kr/books?id=YBimDwAAQBAJ

[3] A. Odena, C. Olah, and J. Shlens. Conditional image synthesis with auxiliary classifier gans. In ICML, 2017.