以深度學習神經網路實現特徵點遮蔽之人臉偵測

Feature Points Obscured Face Detection Using Deep Convolutional Neural Networks

林琛絜 [1] 劉遠楨 [2]
LIN, CHEN JIE [1] LIU, YUAN CHEN[2]
[1] 國立臺北教育大學 資訊科學系研究所 研究生
[1] National Taipei University of Education Graduate School of Information Science Student
E-mail：ckdcshadow@gmail.com
[2] 國立臺北教育大學 資訊科學系研究所 教授
[2] National Taipei University of Education Graduate School of Information Science Professor
E-mail：liu@tea.ntue.edu.tw

摘要

新冠肺炎病毒極高的傳播率使全球各國醫療資源供不應求，為了避免群聚感染而實施隔離更對經濟、運輸、教育等方面都造成嚴重影響，疫情爆發至今仍未見擴散情況被控制，可預見防疫將是一項需長期進行且不容疏忽的日常工作。

鑒於戴口罩為目前行之有效的防疫方法，而當前的臉部偵測模型對於遮蔽了半張臉的戴著口罩的人臉成效不彰，本研究將利用深度學習卷積神經網路克服以上的問題，希望能為防疫貢獻一份心力。

關鍵字：新冠肺炎、卷積神經網路、臉部偵測、遮蔽

Abstract

The extremely high transmission rate of the COVID-19 has made the supply of medical resources in countries around the world in short supply. The implementation of quarantine in order to avoid group infections has a serious impact on the economy, transportation, education and other aspects. Epidemic prevention will be a routine task that needs to be carried out for a long time and cannot be neglected.

In view of the fact that wearing a mask is currently an effective method of epidemic prevention, and the current face detection model is not effective for masked faces that cover half of the face, this study will use deep learning convolutional neural networks to overcome the above I hope to contribute to the prevention and control of the epidemic.

Keywords：COVID-19, Convolutional Neural Network, Face Detection, Masked

壹、 緒論

隨著科學技術的發展，機器學習在各領域有了長足的進展與多方面的應用，如資料探勘、搜尋引擎、自然語言的處理、還有圖像影像及音訊的偵測與辨識等等，都有著許多技術上的突破，其中卷積神經網路(Convolutional neural network, CNN)[1]的貢獻可謂功不可沒。

從 2019 年 11 月開始，一種全新冠狀病毒於中國武漢被首先發現，並導致嚴重特殊傳染性肺炎 (COVID-19) 疫情爆發，此種病毒具備高度傳染性，雖主要通過人與人近距離接觸傳染，但也被發現能夠通過被汙染的物品表面進行傳播，並在兩個月間迅速擴散至全球多國至今，截至 2021 年 4 月 27 日，全球已有累計逾 1.47 億例確診，其中逾 312 萬人死亡，目前尚無完成完整臨床試驗並且可供大眾安全使用的疫苗，期間世界衛生組織(World Health Organization, WHO)宣布它為致命疾病，且疫情構成全球大流行後，世界各國各種防疫宣導、規定、政策甚至封鎖措施紛紛出爐，如進出公務或非公務機關採實聯制、保持社交距離、入境隔離措施、出入高感染傳播風險場域需量體溫、戴口罩等等，曾造成全球醫療與民生用品供給失衡，從此，無論是自願或非自願，保護自身不被傳染、防止疫情繼續傳播成為人們生活中不可或缺的一環。

隨著被證明能確實減少飛沫和氣融膠傳播病毒的數量後[2]，口罩成為人們出門必備的物品，如不戴口罩連大眾運輸工具都無法搭乘，因為戴口罩而造成生活不便的情況也開始浮現，如呼吸不順暢、眼鏡起霧、部分具備人臉識別功能的設備因口罩的遮擋失靈等等。

另外，研究者觀察行人發現部分民眾因各種原因雖有戴口罩卻並未依正確方式佩戴，如口罩未攤開拉至鼻樑及下巴、鼻樑片未壓至與鼻樑貼合等，導致其口罩防護能力下降，具潛在染疫風險。

經過數十年的蓬勃發展，人臉識別技術被廣泛應用於如手機解鎖、監控系統、門禁管理、智慧零售等項目，但防疫方面目前多以紅外線熱像儀感應人體溫度，輔以人工方式判斷口罩是否佩戴及是否正確佩戴，於人潮眾多處難以一一發現及提醒，尤其在不得已必須與人群接觸的環境中，正確地佩戴口罩成為非常重要的保護自身及他人的方法。

在影像處理領域，戴著口罩的臉部偵測對於當前的臉部偵測模型具有相當大的挑戰性，因為口罩遮蔽了部分的臉部特徵，且口罩的種類、款式與花色繁多，還有訓練樣本過少，都對現行的臉部偵測模型造成一定的阻礙。

本研究希望克服目前因口罩遮擋人臉之特徵點的困難，訓練一不只能正確判斷目標是否佩戴口罩，進一步確認口罩是否以正確方式配戴之機器學習模型，期望能對防疫工作做出貢獻。

貳、 文獻探討

一、 卷積神經網路

(一)簡介

卷積神經網路(Convolutional Neural Network, CNN)是目前深度神經網路(Deep Neural network)領域的發展主力，在影像辨別方面卓有成效，以模仿人類大腦的認知方式所建立而成，現行的許多影像辨識模型都基於卷積神經網路的架構做延伸，具有自動學習、歸納特徵的特性，解決了原先因影像所需要處理的數據量過大以及影像因數位化難以保留特徵的問題。

(二)架構

典型的卷積神經網路由卷積層、池化層與全連接層構成。

卷積層將原始影像與特徵檢測濾波器做卷積運算，負責提取影像的局部特徵，其中特徵檢測濾波器隨機產生若干種；池化層用來降低維數，減少像素數量與網路計算的次數且保留特徵的關鍵訊息，目的為縮短訓練時間並防止過擬合；最後全連接層配合權重輸出預測的結果。

二、 Faster-RCNN

(一) 簡介

Region-based Convolutional Network method(R-CNN) [3] 先使用選擇性搜尋(Selective Search)預先篩選出約 2000 個可能包含重要特徵的區域(region proposal)，再將這些區域壓縮放入 CNN 提取特徵並分類提高準確性。

Fast Region-based Convolutional Network method(Fast R-CNN)［4］則將 R-CNN 中重複運算 CNN 的部分縮減至一次，再將擷取的特徵讓 2000 個 region proposal 運用，再使用 Region of Interest Pooling(ROIpooling)對應到 Feature map 上，解決了 R-CNN 運算較慢的問題。

Faster Region-based Convolutional Network method(Faster R-CNN)［5］使用 Region Proposal Network(RPN)來生成 region proposal，並整合 bounding box 與 regression 技術，大幅優化了 Fast R-CNN 的效能。

(二) 架構

Faster R-CNN 的架構大致可分為四個部分：卷積層(conv layers)在預處理時也會記錄特徵資訊給後面的池化層、Region Proposal Network 部分使用 RPN 生成 region proposals 的同時也用 bbox regression 校正 anchor 的位置、ROIPooling 整合 Feature Map 及 RPN 的資訊、最後的分類(Classification)則用 Softmax 函式計算被提出的區塊(Proposals)屬於哪個類別，並再次使用 bbox regression 技術取得更準確的區塊。

三、 You Only Look Once

相對於 Faster R-CNN 等 Two-stage 類型演算法將物件的位置偵測與分類分開進行，You Only Look Once(YOLO)[6]演算法將物體偵測轉化為一個 regression 問題考慮，開創了 One-stage 類型演算法的先河，大幅縮減了偵測所需時間，達到能即時偵測的程度。

YOLO 的主要思路是將影像切分成數個網格(grid)，每格做若干個 bounding boxes 預測其屬於哪個類別並賦予一個分數，用卷積來判斷該格裡是否有物件的中心，並且同時也輸出每個 bounding box 屬於某物件的機率，最後利用 Non-Maximum Suppression(NMS)演算法選出最佳預測框，達成偵測影像中物件位置與類型的效果，但只能處理每個網格中最多只有一個物件的情況，且對影像中較小的物件偵測效果不佳。

YOLO 的作者在發表兩年後提出了 YOLO v2[7]，為了提高偵測定位的準確性，YOLO v2 將 Dataset 的預訓練分為兩個步驟，先用較低解析度影像進行訓練，再改為輸入較高解析度影像繼續訓練，使得預訓練模型能夠適應高解析度影像，並且對影像資料多做了歸一化等預處理，另外借鑒 Faster R-CNN 中使用卷積層與 RPN 來預測 Anchor Box 的思路，作者使用 5 個大小形狀不同的 Anchor Box 對物件邊框做偏移的預測來取代原先的全連接層，減少運算量的同時，達到可以在同一網格中進行多目標偵測的目的，以及保持運算速度的情況下對準確度的提升作出改進，並且在此基礎下提出一個可偵測約 9000 類物件的偵測系統 YOLO9000，對影像中所佔比例較小的物件偵測雖有改善但仍有較容易忽略的情況。

YOLO v3[8]在 YOLO v2 的基礎上更進一步，採用全卷積網路，捨棄池化層改用卷積中的 stride 降維，利用 feature map 做檢測，並且設置了不同尺度的 bounding box 來解決小物件的問題，另外 YOLO v3 把 softmax 函數改為多個 logistic 分類器對 anchor 評分。

YOLO v4[9]在 YOLO v3 各部位都做了相當程度的改進，性能大幅提升的同時對硬體的需求也有所降低，是 YOLO 系列的一次重大更新。

目標偵測模型通常由輸入一張影像(Input)、預訓練骨架(Backbone)、用來提取不同層級的特徵圖(Neck)還有用來預測對象類別和 bounding box 的檢測器 Head 所組成，YOLO v4 對於 dataset 做了許多增加訓練樣本變異性的增強，從幾何變形、亮度變化外還使用了隨機擦除、剪切、多圖組合等方式創造出新的訓練樣本，並使用自我對抗訓練(Self-Adversarial Training, SAT) 形成影像上沒有目標的假象，然後對修改後的影像進行正常的目標檢測，提升模型的穩固性(robustness)。

參、研究方法

一、 研究方法

YOLO v5 是仍在開發中的目標偵測模型，其架構與 YOLO v4 頗為相似，兩者優劣仍有許多爭議，但前者提供比後者較小架構的神經網路，符合追求實時偵測的本研究需求，因此我們決定使用 YOLO v5 作為提取特徵的卷積神經網路，針對我們自己蒐集、標籤的數據集來訓練，並比較兩者的差異選擇較適合的方式，調整其中的參數以提升本研究的準確度。

輸入測試影像時常用的做法是將影像調整至統一的大小，我們在此更進一步希望縮放後影像 padding 的邊框越少越好，以下是我們使用的自適應影像縮放方法公式：

$$\alpha = \frac{|G_h P - G_w P| \ mod\ 32}{2}$$

$$\mathrm{P} = \min(P_h, P_w)$$

$$P_h = \frac{e_h}{G_h}, P_w = \frac{e_w}{G_w}$$

其中α為需填充的像素行或列數。

除了使用 Mosaic 數據增強提升小目標偵測性能外，我們也使用旋轉影像、加入雜訊、調整亮度、調整對比度、縮放及傾斜旋轉等方式相互結合，盡可能擴充數據集的豐富度，降低類別不平衡的影響。

不同於 YOLO v4 訓練時使用的 Mish 激勵函數，我們採用更快速的 Leaky ReLU 激勵函數作為預訓練激勵函數，其公式如下：

$$\mathrm{LReLU}(x) = \begin{cases} x\ if\ x > 0 \\ \alpha x\ if\ x \le 0 \end{cases}$$

其中α在訓練中以反向傳播進行調整。

輸出 anchor box 時，YOLO v5 採用 GIOU_Loss 來更進一步提升運算速度，但因為我們調整過測試影像縮放的縮放方式，計算中考量了影像長寬比的 CIOU_Loss 更為適合，所以選擇其作為我們的損失函數。

$$\mathrm{CIOU_{Loss}} = 1 - \mathrm{CIOU} = 1 - (\mathrm{IOU} - \frac{{Distance_2}^2}{Distance_{C^2}} - \frac{v^2}{(1 - IOU) + v})$$

$$\mathrm{v} = \frac{4}{\pi^2}\left(\arctan\frac{\mathrm{w^{gt}}}{\mathrm{h^{gt}}} - \arctan\frac{\mathrm{w^{p}}}{\mathrm{h^{p}}}\right)^2$$

在神經網路的隱藏層我們使用 Leaky RELU，測試時我們則選擇 Sigmoid 作為激勵函數。

$$\mathrm{Sigmoid}(x) = \frac{1}{1 + e^{-x}}$$

二、 數據集

目前較為知名且包含戴口罩的公開人臉資料庫如下所列：

1. WIDER Face dataset[10]包含了 32203 張人臉影像及 393703 個人臉邊界標註框，測試子集有 3226 張人臉影像與 12880 個人臉邊界標註框，並分為三種偵測難度:容易、普通和困難，我們在其中挑選了 500 張影像，其中包含 221 張戴口罩人臉影像及 103 張人臉上有遮蔽物的影像。
2. Kaggle Face Mask Detection dataset 包含了 853 張分別標註了有戴口罩、沒戴口罩、口罩沒戴好的影像。
3. Kaggle YOLO medical mask dataset 包含了 631 張標註了戴口罩人臉邊界框影像。
4. Real-World Masked Face Dataset(RMFD)收集了包含 525 人的約五千張戴口罩人臉影像及九萬張人臉影像，還有將一般人臉影像後製成戴口罩人臉影像約 50 萬張。

雖然人臉資料庫數量眾多，但適合本研究的影像，尤其是口罩配戴不當的影像難尋，且其中有許多影像是人工創造而非自然影像，或是標註錯誤、影像只有人臉而沒有背景等等無法使用或訓練效果不佳的情況，因此除了使用現存的資料庫，我們也藉由拍攝、爬蟲取得更多的訓練影像並標註，然後使用前文所述之數據增強方法對其擴充，最後一共是 4294 張影像，其中包含 6428 個戴口罩人臉標註框、1614 個沒有戴口罩人臉標註框、398 個錯誤配戴口罩人臉標註框，並將其隨機以 4:1 比例分配給神經網路的訓練集與驗證集。

三、 演算法

我們將整個過程分為兩個演算法。

首先對數據集影像進行預處理，然後對數據集進行訓練，接著使用訓練出來的模型偵測測試影像中是否有配戴口罩的人臉及其是否配戴正確。

在訓練的部分將影像及像素值作為輸入，對影像調整大小並歸一化後進行數據增強以擴展數據集，降低過擬合的可能，然後將數據集隨機分為訓練集與驗證集，以 Adam 優化函數編譯整個模型後存檔。

接著將之前存檔的模型部屬到要偵測的影像或影片中，如果偵測到人臉，將會顯示一個邊框將其包圍，並顯示該人臉屬於正確配戴口罩、錯誤配戴口罩、沒有配戴口罩中哪一類。

肆、 研究結果

一、 研究設備介紹

本研究使用了配備 Intel i5-6500 處理器(3.2GHz)、2 x 8GB DDR3 記憶體及 NVIDIA GeForce GTX 1060 6GB 平行處理器的桌上型電腦，並且使用 Python3.7 環境下的 Jupyter Notebook 軟體實現與訓練神經網路。

二、 評估指標

目標偵測神經網路常見的評估指標有以下幾種：

$$Precision = \frac{T_p}{T_p + F_p}$$

$$Recall = \frac{T_p}{T_p + F_n}$$

$$F1score = 2 * \frac{Recall * Precision}{Recall + Precision}$$

$$AP = \int_0^1 p_{smooth}(r)dr$$

其中T_p = True positive 表示標註為真且預測結果為真， F_p = False positive 表示標註為假但預測結果為真，F_n = False negative 則表示標註為真且預測結果為假，預測結果的真假則通過 CIOU_Loss 損失函數與一個閾值(threshold)來判斷。

Precision 為某一類別預測結果正確的準確率，本研究三類別中最低達到 86.1%，Recall 為某一類別預測結果正確佔所有此類樣本的比例，本研究三類別中最低達到 87%，F1 分數則為 Precision 與 Recall 的調和平均數，本研究三類別中最低達到 69%。

以 Precision 為縱軸、Recall 為橫軸畫出一 PR 曲線圖後，我們可以看出兩者呈現負相關，Average Precision(AP)為對此曲線平滑化後做積分做為評估指標，mAP 則為所有類別的 AP 值平均，本研究之 mAP 值達到 67.7%。

伍、 結論與未來展望

一、 結論

本研究針對戴口罩人臉的數據集經過嚴格的篩選與重新標註，完成了區分為三個類別的訓練及開發，可適用於輸入各種尺寸的影像與影片，參考了 YOLO v4、YOLO v5 與其他神經網路架構，我們的神經網路可以在兩小時內運行 300 次 epoch 完成訓練，並在 960x540 解析度的影片下達到平均每秒 72.09 幀影格的速度。

在疫情肆虐全球的現在，每個人都需要提高警覺，出門戴口罩更是必不可少，本研究訓練之深層神經網路模型可適用於車站、學校出入口等人流密集的地方，以其判斷快速的特性做到實時區分出對於戴口罩做得不夠確實的人群，提高大家的警惕心，降低群聚感染發生的可能，可望對於防疫做出些許貢獻。

二、 未來展望

目前的數據集若使用更深的神經網路架構容易因類別不平衡導致過擬合的情況發生，如果加入更多的真實影像數據應該可以使訓練模型的準確度更上一層樓。

另外也許可以結合紅外線熱像儀產生的紅外線影像，開發一能同時判斷是否配戴口罩、體溫是否正常之警報系統，讓防疫工作更加完善。

參考文獻

一、 中文部分

二、 英文部分

[1] LeCun, Yann, et al. Gradient-based learning applied to document recognition.(1998). Proceedings of the IEEE 86.11 : 2278-2324

[2] Howard, J., Huang, A., Li, Z., Tufekci, Z., Zdimal, V., van der Westhuizen, H., … Rimoin, A. W. (2021). Face Masks Against COVID-19: An Evidence Review. PNAS January 26, 2021 118(4) e2014564118; doi:10.1073/pnas.2014564118

[3] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in
Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, Jun.23-28, 2014, pp.580-587.

[4] R. Girshick. Fast R-CNN. (2015) arXiv:1504.08083

[5] S. Ren, K. He, R. Girshick, and J. Sun. (2015). Faster R-CNN: Towards
real-time object detection with region proposal networks. In NIPS.

[6] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. arXiv preprint arXiv:1506.02640, 2015.

[7] J. Redmon and A. Farhadi. Yolo9000: Better, faster, stronger. In Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on, pages 6517–6525. IEEE, 2017.

[8] Joseph Redmon and Ali Farhadi. YOLOv3: An incremental improvement. arXiv preprint arXiv:1804.02767, 2018.

[9] Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. YOLOv4: Optimal Speed and Accuracy of Object Detection. arXiv 2020, arXiv:2004.10934

[10] Shuo Yang, Ping Luo, Chen-Change Loy, and Xiaoou Tang. Wider face: A face detection benchmark. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 5525–5533, 2016. http://shuoyang1213.me/WIDERFACE/index.html.

附錄