國 立 中 央 大 學

資 訊 工 程 學 系

碩 士 論 文

灰階至 RGB：卷積神經網路可解釋性模型的優化與擴展

From Gray to RGB: Optimization and Extension of CNN-based Interpretable Model

研 究 生：涂建名

指導教授：蘇木春 博士

中 華 民 國 一百一十三年 六 月

灰階至 RGB：卷積神經網路可解釋性模型的優化與擴展

摘要

關鍵字：可解釋的人工智慧, 深度學習, 色彩感知, 性能提升

From Gray to RGB: Optimization and Extension of CNN-based Interpretable Model

Abstract

Keywords: Explainable Artificial Intelligence, Deep Learning, Color Perception, Performance Enhancement

誌謝

目錄

圖目錄

頁次

表目錄

頁次

一、 緒論

1.1 研究動機

自 1998 年 LeNet[1] 問世以來，隨著深度學習的蓬勃發展，人工智慧應用範圍也逐步融入到人們日常生活的方方面面。然而儘管人工智慧的發展如此蓬勃，實際上我們對於人工智慧的實際運作過程與做出決策的理由仍然存在著許多未知的地方，目前，大部分模型彷彿是一個黑盒子，我們雖了解其運作理論，但卻無法得知其每個決策的具體理由和依據。

當人工智慧開始運用到各行各業時，人們開始發覺在某些領域或是應用情境 (如：醫療決策、軍事領域、金融決策等) 下，單單只有高準確度是無法讓使用者具備足夠的信心採用人工智慧所預測的決策，這些領域所需要的決策往往需要合理的理由或是因果關係的推論支撐才足以讓使用者有足夠的信心採用，在此情況下，具備可解釋性的深度學習模型做出令使用者有信心採用的決策。

隨著美國國防部 MAPPA 在 2016 年將可解釋性人工智慧（XAI）列為 third-wave AI systems 列為 DARPA 計畫項目之一 [2]、歐盟也在同年通過了《European Union's General Data Protection Regulation (GDPR) 》裡面規範使用者有獲得有關於推論資訊的"meaningful information about the logic involved" 的權利 [3],[4]。這些重要的政策舉措使得可解釋性的深度學習模型成為了全球範圍內的熱門研究，不僅在學術界，也在企業界甚至國家層面都被視為重要的發展項目。

1.2 研究目的

本論文旨在深入研究 2023 年由 J.-F. Yang 等人所提出之 CNN-based Interpretable Model(以下簡稱 CIM) [5]，在此基礎上進行效能改進並進一步開發出一個不只適用於灰階影像而能更廣泛的適用於 RGB 彩色影像之可解釋性模型，使其在保持原來 CIM 模型的高準確度與高可解釋性的水準下可以應用於更多現實影像分類任務。

透過研究人眼如何辨識彩色影像，我們希望設計出用於模擬人眼感知色彩機構的色彩感知層和感知輪廓的輪廓感知層，使模型可以模仿人眼感知色彩的過程，並將兩者資訊結合後輸入 CIM。藉由 CIM 模型的階層式時序處理，模擬人腦多層皮質資訊傳遞，每一層都將擁有色彩和輪廓的特徵資訊，最終形成一個完整的影像特徵資訊，並輸入全連接層以學習每個分類的特徵。

此外本論文也希望開發出來的適用於 RGB 彩色可解釋性模型能夠針對每一層的輸出之特徵進行分析，並找出各層輸出特徵與最後預測分類之間的關係，以理解該模型是根據何種特徵做出分類判斷，從而形成一個使用者可以接受之解釋。

1.3 論文架構

本論文分為五個章節，架構如下:

第一章：緒論，敘述本論文的研究動機、目的和架構

第二章：背景知識與文件回顧，介紹本論文所需之背景知識與回顧可解釋性人工智慧的演進與各個分類的重要論文

第三章：研究方法，介紹本論文對以卷積神經網路為基礎的 RGB 彩色可解釋性模型的架構與方法

第四章：實驗設計與結果，對本論文所提出的方法在不同資料集上的效果進行實驗與觀察

第五章：總結，對本論文之結果做出結論並提出未來可行之研究方向

二、 背景知識與文獻回顧

2.1 背景知識

本章節將會介紹本論文所需的背景知識，可以幫助讀者更好地理解本論文提出的論文的概念和出發點，內容包含：人如何感知彩色影像、大腦皮質的運作、卷積神經網路與以卷積神經網路為基礎的可解釋性深度學習模型。

2.1.1 人如何感知彩色影像

2.1.2 皮質的運作

2.1.3 卷積神經網路

2.1.4 以卷積神經網路為基礎的可解釋性深度學習模型

2.2 文獻回顧

2.2.1 可解釋性人工智慧的演進與分類

Decision Tree: [6] **grinsztajn2022treebased**

介紹可解釋性人工智慧的歷程，分類，各分類著名的論文的簡介可解釋性人工智慧的研究最早可以追蹤到 1991 年的專家系統時代 W Swartout, C Paris 等人便開始對可解釋性人工智慧進行研究 [7]，但是

2.2.2 對於 Inherently Interpretable 可解釋性模型之研究

2.2.2.1 基於多層自我映射圖之可視覺化深度學習模型

2.2.3 對於 Post-hoc 可解釋性模型之研究

2.2.3.1 Local Interpretable Model-agnostic Explanations(LIME)

2.2.3.2 Shapley Additive Explanations(SHAP)

2.2.4 近年可解釋性模型趨勢之研究

2.2.4.1 Tabnet: Attentive interpretable tabular learning

2.2.4.2 Building more explainable artificial intelligence with argumentation

XAI 的新趨勢使用論證的方式來解釋，特別是計算論證有助於理解理性決策的所有步驟以及在不確定性下進行推理。[8]

三、 研究方法

3.1 以卷積神經網路為基礎的 RGB 彩色可解釋性模型

3.1.1 模型架構

此章節將介紹本論文所提出的可解釋性模型整體架構與每個部分的功能，並說明資料在模型中的運作方式，模型架構圖如圖 3.1。整個模型可以分成三個部分，色彩感知層、輪廓感知層和特徵傳遞層。

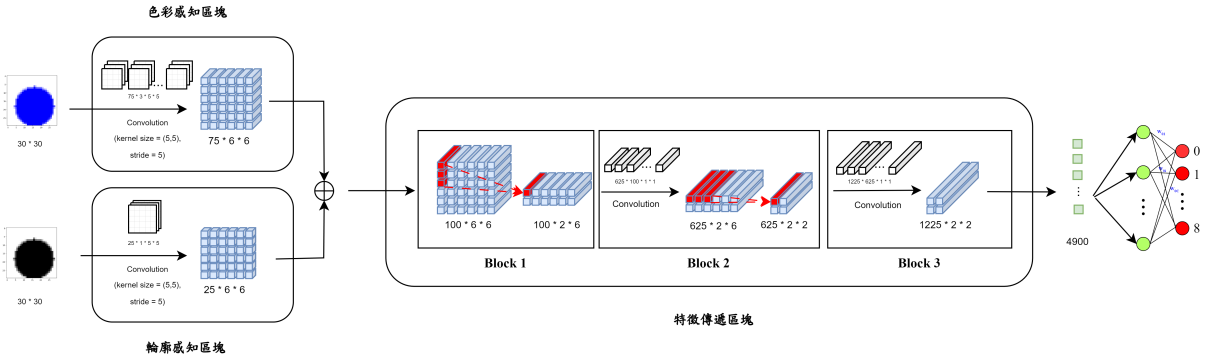

圖 3.1: 模型架構圖

色彩感知區塊基於 Thomas Young 所提出的三色視覺理論 (Trichromacy Theory)**<empty citation>**，透過模擬眼球中的三種類型的視錐細胞，提取與學習影像中不同區塊的 RGB 的比例，以提取影像中每個區塊的色彩特徵。輪廓感知區塊透過將影像進行灰階化後使用高斯卷積層來提取影像中輪廓和邊緣的特徵。特徵感知區塊使用了 CIM 模型模擬了大腦皮

質的運作模式並對其進行優化，每一層的 Block 都會將底層的特徵資訊整合並傳遞到下一層進行學習。特徵傳遞層的 Block 可以分成三個部分，高斯卷積模組，特徵增強模組，空間合併模組，高斯卷積模組負責學習與提取輸入的特徵，特徵增強模組負責過濾不重要的特徵，空間合併模組則模擬皮層的資訊合併，融合眼球跳動的概念，將輸入的資訊根據空間位置關係進行合併。

3.1.2 演算法流程

Step 1：決定整個模型的架構與參數

Step 2：對輸入的彩色影像做灰階化產生對應的灰階影像

Step 3：將彩色影像和灰階影像分別輸入色彩提取層和輪廓提取層提取出色彩特徵與輪廓特徵

Step 4：在獲得色彩特徵與輪廓特徵時將他們 concat 在一起形成一個綜合特徵

Step 5：將綜合特徵輸入特徵傳遞層進行綜合特徵的學習與合併

Step 6：將完整的特徵資訊輸入全連接層學習分類特徵

Step 7：將色彩提取區塊的 weight 正規化至 [0,1] 之間

Step 8：計算 loss value 並進行反向傳播

3.2 色彩提取區塊設計與實現

此章節將說明本論文所提出的色彩提取區塊設計與實現，該區塊主要是透過將 filter 的初始化為不同的 RGB 色彩值來作為該區塊的 weights，並使用高斯卷積模組計算出影像中不同區塊之色彩與 filter 的相似度，後將結果送入特徵增強模組，最終形成影像的色彩特徵。這樣的方法目的在於模擬人眼中的三類視錐細胞基於 RGB 值來感知不同外界的色彩的過程，並透過卷積操作來去模擬人眼的眼球跳動，從而重現人眼獲得色彩特徵的完整流程。以下將針對 filter 初始化、高斯卷積模組、特徵增強模組優化設計三個部分進行詳細說明。

3.2.1 Filter 初始化

由於色彩提取區塊的輸入為彩色影像其輸入通道分別為紅、綠、藍三色的通道，我們令輸出的通道數為 C_{out} 、kernel 的長、寬為 H_{kernel} 、W_{kernel}，因此，Filter 的形狀為 $(C_{out}, 3, H_{kernel}, W_{kernel})$。我們將 Filter 視為 C_{out} 個不同 RGB 色彩的 H_{kernel} * W_{kernel} 的色塊，這樣設置的目的是希望可以讓色彩提取區塊專注於學習影像中不同區域的色彩分布和色彩特徵，而不需要額外去學習輪廓特徵

在實作中，我們先使用 Kaiming Uniform 的方式將 RGB 三個通道分別初始化出 C_{out} 個，根據 kaiming 論文 [9] 中的方式將每個值初始化範圍為 $[-\sqrt{\frac{6}{\text{fan\_in}}}, \sqrt{\frac{6}{\text{fan\_in}}}]$, fan_in 為輸入通道數。此處的目的是希望初始化出 C_{out} 種不同的 RGB 色彩, 形成 $(C_{out}, 3)$ 的值，並且再將這 C_{out} 的色彩重複擴張成 H_{kernel} * W_{kernel} 的色塊。

我們在實作中選擇使用 Kaiming Uniform 的原因是因為 Kaiming Uniformm 相較於常用的 Uniform 初始化和 Xavier Uniform[10] 初始化多考慮了整流線性單位函數 (ReLU) 的存在，Uniform 初始化的方式無法解決隨著神經網路的增加而導致梯度消失的問題，Xavier Uniform 為了解決

梯度消失問題加入了 rescale 函數 $\frac{1}{7}\sqrt{n}$ 但卻只適用於激活函數為線性函數的情況下，而 Kaiming Uniform 在解決梯度消失的問題時同時考慮了激活函數為非線性函數的情況並在 [9] 透過實驗證明了 kaiming uniform 在神經網路在不影響準確度的同時更快收斂。由於我們的模型中在特徵增強模組中使用的非線性函數 ReLU 的變形去進行特徵增強，因此選擇了 Kaiming Uniform 來去進行後面的實驗。

3.2.2 高斯卷積模組

3.2.3 特徵增強模組之優化設計

3.2.4 訓練過程的正規化

3.3 特徵感知區塊的設計與實現

3.3.1 模型效能優化設計

3.3.2 空間位置保留機制之優化設計

3.4 可解釋性

3.4.1 色彩感知區塊之可解釋性

3.4.2 輪廓感知區塊之可解釋性

3.4.3 特徵感知區塊之可解釋性

3.5 量化推論成果之方法

四、 實驗設計與結果

4.1 灰階優化模型與以卷積神經網路為基礎的具可解釋性的深度學習模型之比較

4.1.1 資料集介紹

4.1.2 實驗設計

4.1.3 實驗結果

4.2 模型保留空間位置特徵之臉部驗證實驗

4.2.1 實驗背景與目的

4.2.2 資料集介紹

4.2.3 模型架構與參數

4.2.4 實驗結果

4.3 以卷積神經網路為基礎的 RGB 三通道可解釋性模型之效果驗證

4.3.1 資料集介紹

4.3.2 實驗設計

4.3.3 實驗結果

4.4 實際用於現實瘧疾影像上的效果

4.4.1 資料集介紹

4.4.2 模型架構與參數

4.4.3 實驗結果

五、 總結

5.1 結論

5.2 未來展望

參考文獻

[1] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[2] D. Gunning. "Explainable artificial intelligence (xai)." (Aug. 10, 2016), [Online]. Available: https://www.darpa.mil/program/explainable-artificial-intelligence.

[3] European Parliament and Council of the European Union. "Regulation (EU) 2016/679 of the European Parliament and of the Council." (May 4, 2016), [Online]. Available: https://data.europa.eu/eli/reg/2016/679/oj.

[4] B. v. d. S. Chris Jay Hoofnagle and F. Z. Borgesius, "The european union general data protection regulation: What it is and what it means*," *Information & Communications Technology Law*, vol. 28, no. 1, pp. 65–98, 2019.

[5] C.-F. YANG *et al.*, "A cnn-based interpretable deep learning model," Master's thesis, National Central University, 2023.

[6] L. Rokach, "Decision forest: Twenty years of research," *Information Fusion*, vol. 27, pp. 111–125, 2016.

[7] W. Swartout, C. Paris, and J. Moore, "Explanations in knowledge systems: Design for explainable expert systems," *IEEE Expert*, vol. 6, no. 3, pp. 58–64, 1991.

[8] L. Longo, M. Brcic, F. Cabitza, *et al.*, "Explainable artificial intelligence (xai) 2.0: A manifesto of open challenges and interdisciplinary research directions," *Information Fusion*, vol. 106, p. 102 301, 2024.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," *CoRR*, vol. abs/1502.01852, 2015.

[10] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, Y. W. Teh and M. Titterington, Eds., ser. Proceedings of Machine Learning Research, vol. 9, Chia Laguna Resort, Sardinia, Italy: PMLR, 13–15 May 2010, pp. 249–256.