

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

|   | 頁次  |
|---|-----|
| 摘要  | i   |
| Abstract                                  | ii  |
| 誌謝  | iii |
| 目錄  | iv  |
| <br>                                      |     |
| 一、緒論                                      | 1   |
| 1.1 研究動機 .....                            | 1   |
| 1.2 研究目的 .....                            | 2   |
| 1.3 論文架構 .....                            | 3   |
| <br>                                      |     |
| 二、背景知識與文獻回顧                               | 4   |
| 2.1 背景知識 .....                            | 4   |
| 2.1.1 人如何感知彩色影像 .....                     | 4   |
| 2.1.2 CNN-based Interpretable Model.....  | 12  |
| 2.2 文獻回顧 .....                            | 15  |
| 2.2.1 可解釋性人工智慧的演進分類 .....                 | 15  |
| 2.2.2 對於 Ante-hoc explainable 模型之研究 ..... | 16  |
| 2.2.3 對於 Post-hoc 可解釋性模型之研究 .....         | 17  |
| 2.2.4 近年可解釋性模型趨勢之研究 .....                 | 17  |

|           |   |           |
|-----------|---|-----------|
| <b>三、</b> | <b>研究方法</b>                             | <b>18</b> |
| 3.1       | 以卷積神經網路為基礎的 RGB 彩色可解釋性模型 .....          | 18        |
| 3.1.1     | 模型架構 .....                              | 18        |
| 3.1.2     | 演算法流程 .....                             | 19        |
| 3.1.3     | 模型符號說明 .....                            | 20        |
| 3.2       | 色彩提取區塊設計與實現 .....                       | 21        |
| 3.2.1     | Filter 初始化 .....                        | 21        |
| 3.2.2     | 彩色卷積模組 .....                            | 22        |
| 3.2.3     | 訓練過程的正規化 .....                          | 22        |
| 3.3       | 輪廓感知區塊之前處理設計 .....                      | 23        |
| 3.4       | 特徵傳遞區塊之優化設計 .....                       | 24        |
| 3.4.1     | 高斯卷積模組優化設計 .....                        | 24        |
| 3.4.2     | 特徵增強模組之優化設計 .....                       | 24        |
| 3.4.3     | 空間位置保留機制之優化設計 .....                     | 25        |
| 3.4.4     | 模型流程的精簡 .....                           | 26        |
| 3.5       | 可解釋性 .....                              | 27        |
| 3.5.1     | FM、RM、CI 的意義 .....                      | 27        |
| 3.5.2     | 色彩感知區塊之可解釋性 .....                       | 28        |
| 3.5.3     | 色彩特徵傳遞區塊之可解釋性 .....                     | 28        |
| 3.5.4     | 輪廓感知區塊和特徵學習區塊之可解釋性 .....                | 29        |
| <b>四、</b> | <b>實驗設計與結果</b>                          | <b>30</b> |
| 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 ..... | 30        |
| 4.1.1     | 資料集介紹 .....                             | 30        |
| 4.1.2     | 實驗設計 .....                              | 30        |
| 4.1.3     | 實驗結果 .....                              | 30        |

|           |                                      |           |
|-----------|--------------------------------------|-----------|
| 4.2       | 模型保留空間位置特徵之臉部驗證實驗 .....              | 30        |
| 4.2.1     | 實驗背景與目的 .....                        | 30        |
| 4.2.2     | 資料集介紹 .....                          | 30        |
| 4.2.3     | 模型架構與參數 .....                        | 30        |
| 4.2.4     | 實驗結果 .....                           | 30        |
| 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 ..... | 31        |
| 4.3.1     | 資料集介紹 .....                          | 31        |
| 4.3.2     | 實驗設計 .....                           | 31        |
| 4.3.3     | 實驗結果 .....                           | 31        |
| 4.4       | 實際用於現實瘡疾影像上的效果 .....                 | 31        |
| 4.4.1     | 資料集介紹 .....                          | 31        |
| 4.4.2     | 模型架構與參數 .....                        | 31        |
| 4.4.3     | 實驗結果 .....                           | 31        |
| <b>五、</b> | <b>總結</b>                            | <b>32</b> |
| 5.1       | 結論 .....                             | 32        |
| 5.2       | 未來展望 .....                           | 32        |
|           | <b>參考文獻</b>                          | <b>33</b> |

# 圖目錄

|   | 頁次 |
|---|----|
| 2.1 詳細視覺路徑圖 [6] . . . . .               | 5  |
| 2.2 視網膜整體架構 . . . . .                   | 6  |
| 2.3 視網膜對影像進行不同的平行處理 [8] . . . . .       | 7  |
| 2.4 眼球跳動示意圖 [8] . . . . .               | 8  |
| 2.5 視網膜與外膝體的對應關係 [7] . . . . .          | 9  |
| 2.6 Visual Pathway 的認知過程 [10] . . . . . | 10 |
| 2.7 皮層計算模組 [10] . . . . .               | 11 |
| 2.8 CIM 架構圖 [5] . . . . .               | 12 |
| 2.9 合併方式示意圖 [5] . . . . .               | 13 |
| 2.10 FM-RM-CI[5] . . . . .              | 14 |
| 3.1 模型架構圖 . . . . .                     | 18 |



# 表目錄

頁次

|                             |    |
|-----------------------------|----|
| 2.1 特徵圖對應法示意圖 [5] . . . . . | 15 |
|-----------------------------|----|

# 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

要了解人如何感知色彩我們必須要先了解將彩色影像的這條 Central Visual Pathway 會經過哪些的部位與流程，根據《Neuroscience》[6] 所介紹，彩色影像在 Central Visual Pathway 會經過的部位總共可以分為三個重要部位：視網膜 (Retina)、外側膝狀體 (外膝體，Lateral Geniculate Nucleus)、視覺皮層 (Visual Cortex)，彩色影像從視網膜進入後會送入外膝體，外膝體在收到兩側眼球的資訊後會將不同的資訊平行傳輸至不同的視覺皮層，視覺皮層則負責對這些資訊進行分層的整合與感知。詳細的 Central Visual Pathway 如圖 2.1

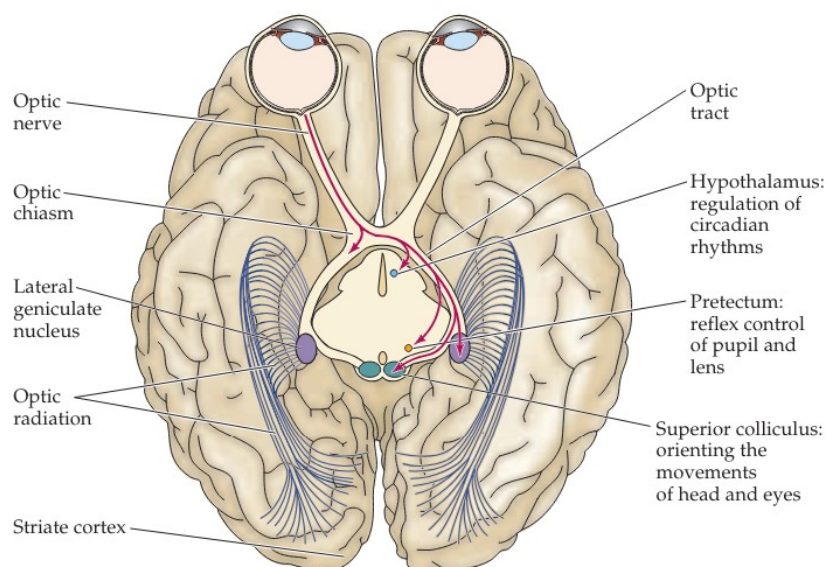
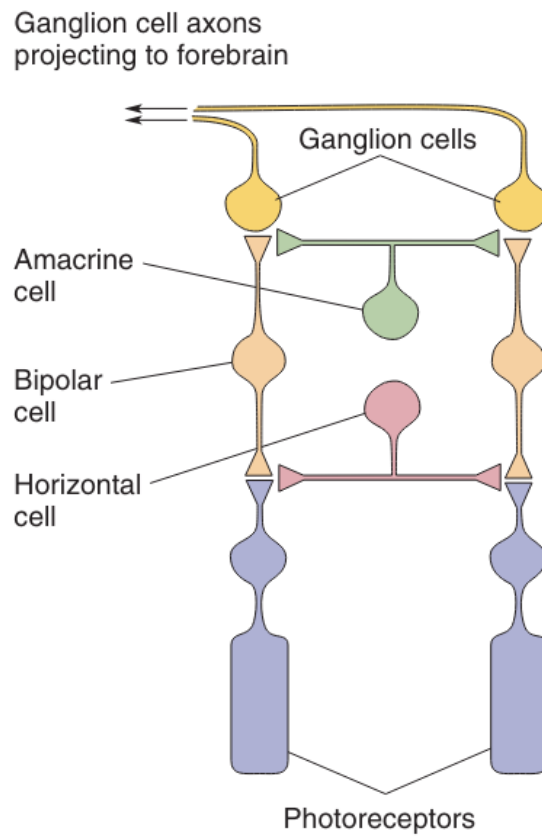


圖 2.1: 詳細視覺路徑圖 [6]

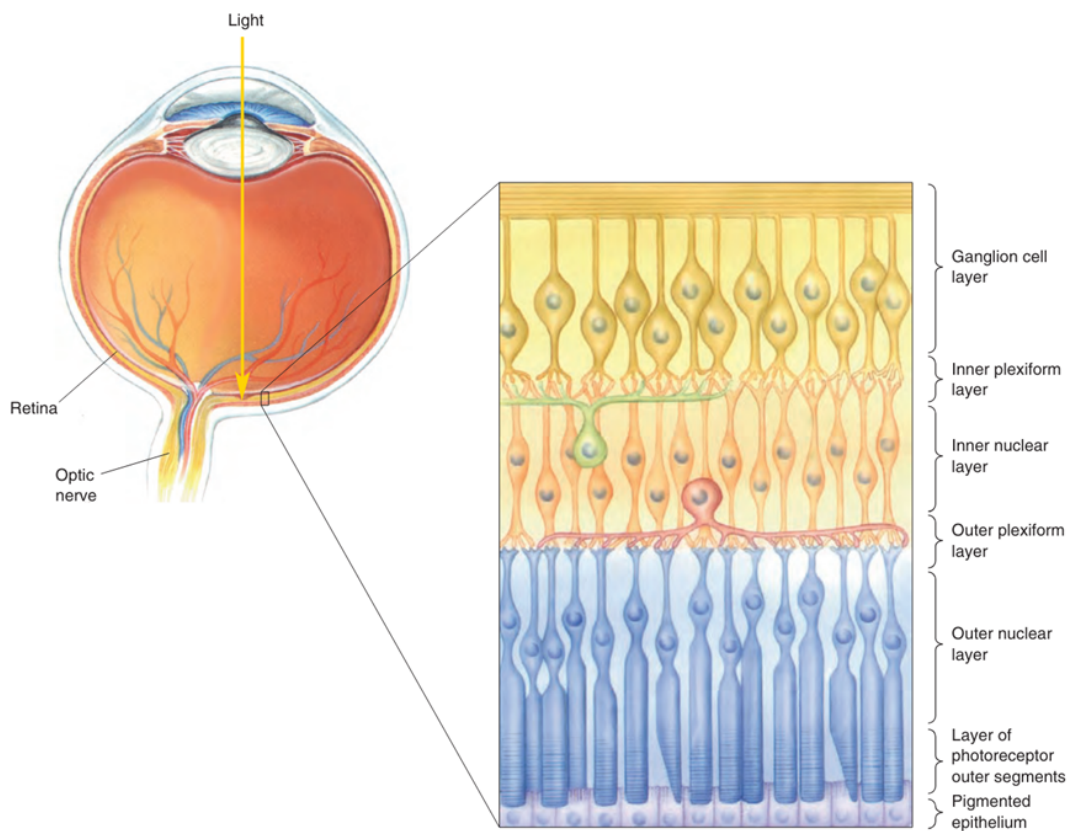
### 2.1.1.1 視網膜

外界事物將光反射進入眼睛當中，透過眼珠中的角膜與水晶體等透明的光學介質進行折射最終聚焦於視網膜表面的感光層上形成影像。在光聚焦於視網膜表層時，視網膜會將光轉換成動作電位並透過視神經傳輸至外膝體，也因此 Central Visual Pathway 才會將視網膜視為感知彩色影像的第一個重要部位。

根據《Neuroscience-Exploring the Brain》[7] 我們知道視網膜的基礎架構是由五類細胞組成如圖 2.2a，分別是：感光細胞、水平細胞、無長突細胞、雙極細胞、神經節細胞。感光細胞包含我們常聽到的視錐細胞等負責將輸入的光轉化為動作電位、雙極細胞負責將會將感光細胞的電位傳送到神經節細胞、這個傳輸的過程有些則是由水平細胞和無長突細胞協助傳輸，神經節細胞則負責將最後的資訊傳輸到外膝體之中。



(a) 視網膜基本架構 [7]



(b) 視網膜層級架構 [7]

圖 2.2: 視網膜整體架構

上面講的基礎架構只是最基本的情況，事實上視網膜的各類細胞遠比上面要架構要複雜許多，上述的五大類細胞還可以再更細的分為不同功能的變種細胞，由此組成了極度複雜的視網膜層級系統如圖 2.2b。如此複雜的視網膜系統，其功能當然不只負責影像的感知與電位傳輸，事實上在 2013 年的 [8]，就已經發現在感光細胞將光轉換為動作電位後會將電位傳輸到層級架構的 Inner Plexiform Layer(IPL) 中不同變種的雙極細胞，這些不同變種的雙極細胞對收到的影像資訊進行不同的平行處理，最終輸出影像中不同的方面的要素到神經節細胞，例如：紅藍綠不同色彩、明暗的變化、影像輪廓等等，這同時也是本論文在設計彩色影像感知時的核心概念。

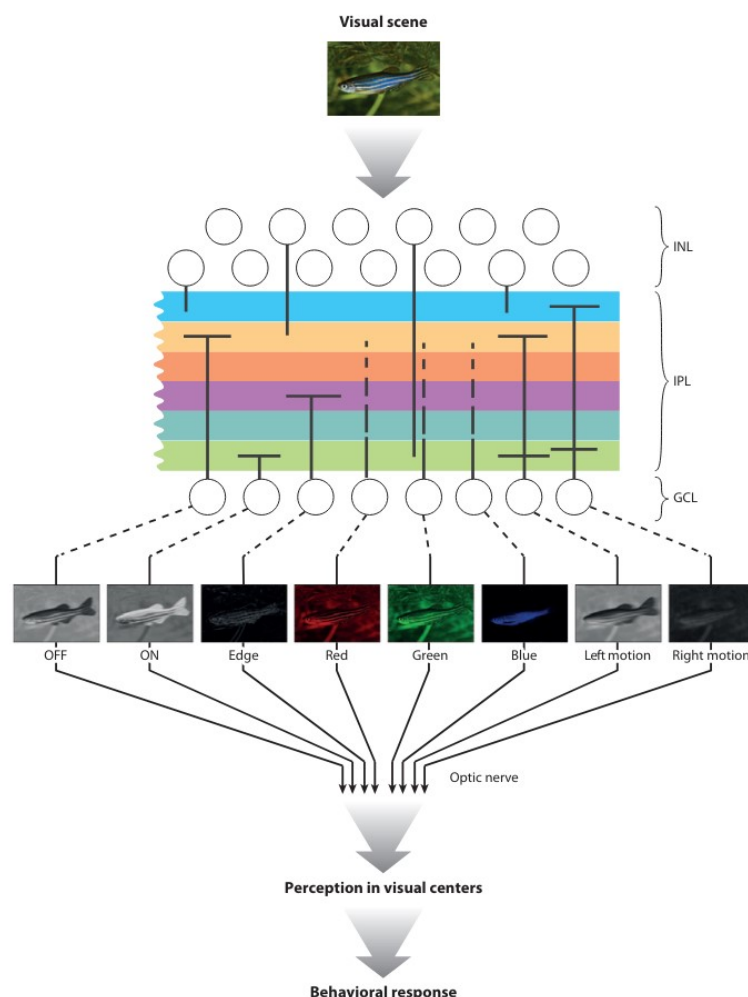


圖 2.3: 視網膜對影像進行不同的平行處理 [8]

此外，在 Jeff Hawkins 的著作 [9] 和《Principles of Neural Science》[10]



中均有提到，視網膜能接收到的影像資訊不只是上述的空間性資訊，同時也具有不同時間性的資訊，這代表同樣的影像進入視網膜的型態是會隨時間而改變。具體而言，眼睛會在每一秒鐘快速移動視線焦點三到四次(如圖 2.4)，但人類的認知上不會有所感覺，這個行為被稱為「眼球跳動」(Saccade)。眼球跳動使得同一個影像產生時間上的變化，形成時間性的影像資訊。這個概念也被運用於本論文的空間位置保留機制中的時間遺忘參數上，使得影像特徵可以呈現時序上的不同。



圖 2.4: 眼球跳動示意圖 [8]

### 2.1.1.2 外側膝狀體

外側膝狀體 (外膝體) 主要負責將視網膜不同的方面資訊 (如: 色彩、輪廓、運動方向…) 傳輸到對應的初級視覺皮質，其中的細胞分層排列，每層分布排列著不同種類的細胞。除此之外，外膝體中不同的細胞層也對應著不同視野的半個視網膜形成如圖 2.5 的對應關係，這也表示視網膜中相鄰區域的同種類的影像資訊在外膝體中很可能在同一個細胞層，這個性質也保證了外膝體在資訊傳輸的過程可以保留資訊的空間位置資訊。由於視覺皮層會將影像資訊從初級到高級逐漸進行資訊整合和學習，因此外膝體的存在可以協助視覺皮層將不同的影像資訊傳輸到對應的視覺皮質層，這對視覺皮質能夠進行平行處理與整合起到至關重要的作用。

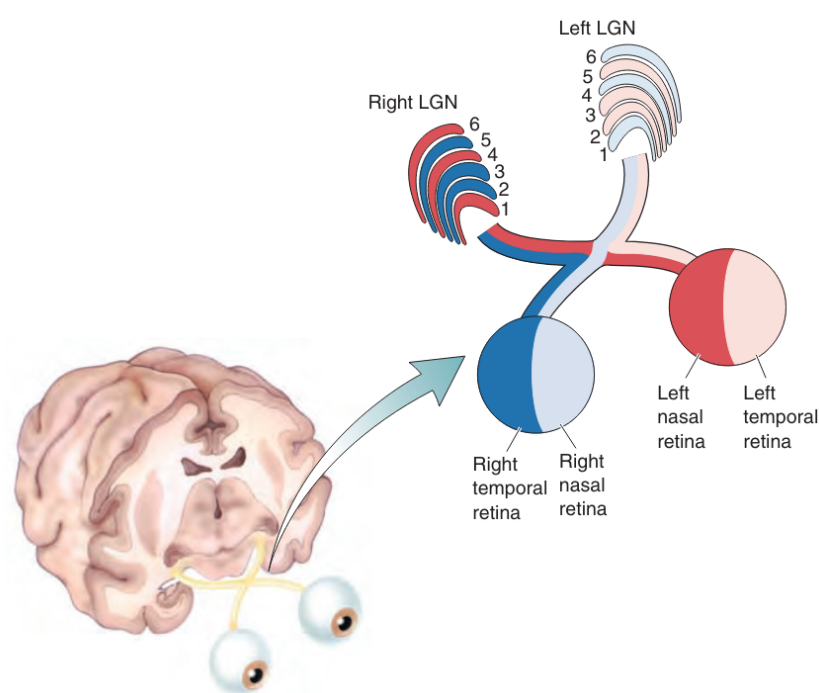


圖 2.5: 視網膜與外膝體的對應關係 [7]

### 2.1.1.3 視覺皮層

皮層可以說是大腦最重要的地方，高級認知功能例如：記憶、思考、感覺等均在此發生。根據 [10] 中的說明，人類的大腦中皮質以分層結構存在的，而視覺皮層又可以分為初級視覺皮層 (Primary visual cortex, V1) 和紋外皮層 (V2、V4、MT 層) 這四層。初級視覺皮層負責接收 LGN 傳輸來的資訊並開始處理顏色、方向、輪廓等視覺資訊，其餘皮層則負責整合和傳遞資訊給下面的皮層，因此隨著皮層的深入，所得到的視覺資訊也會越來越完整。

[10] 提出一種影像資訊處理架構，示意圖如圖 2.6，其將皮質中的資訊傳輸路徑分成兩類，Ventral Pathway 和 Dorsal Pathway, Ventral Pathway 由 V1、V2、V4 組成負責處理影像的色彩、形狀等資訊，而 Dorsal Pathway 由 V1、V2、MT 組成負責處理影像的運動方向的資訊。但同時也強調這個分類只是一個大致分類，實際上不同的視覺特徵資訊 (如顏色、運動方向等) 在皮層中也會相互聯繫，也就是存在所謂的側向連結，最終才能形成統一的影像感知。

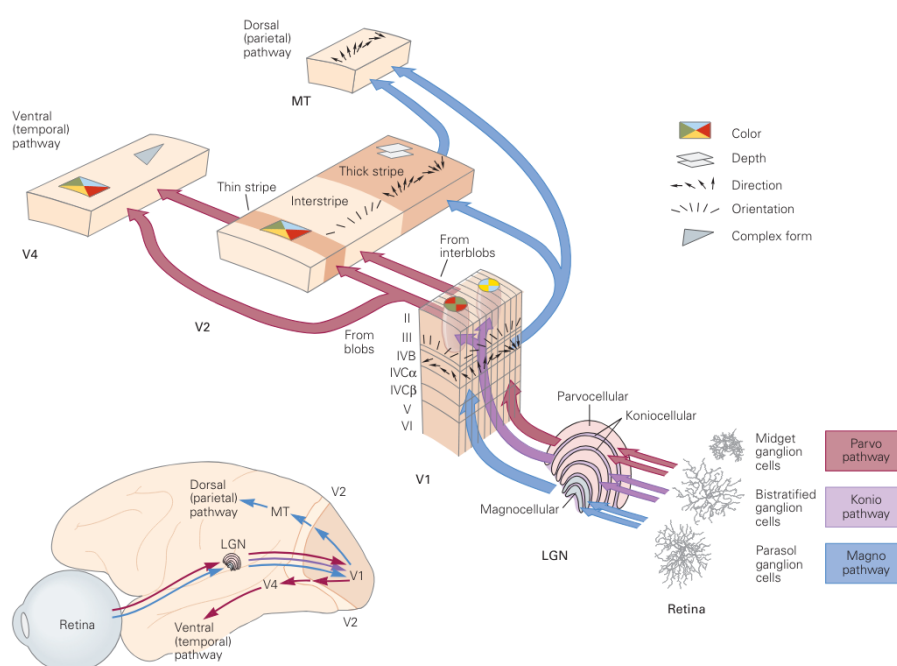


圖 2.6: Visual Pathway 的認知過程 [10]

在皮層中，V1 將相同功能的神經束聚在一起形成不同功能的柱狀結構(例如顏色柱、方向柱、線條柱等) 這些柱狀體負責處理各自的視覺資訊並按照固定規則排列成層，並且將左右眼的柱狀層交錯排列形成一個皮層計算模組 (Cortical Computational Module)，如圖 2.7，重複上百次來覆蓋左右眼看到的所有影像。這樣設計的目的是希望不管從 V1 中的任何位置和方向移動都可以漸進的對每塊視野的顏色、輪廓、方向、深度等特徵資訊進行充分的分析。以上的兩個概念被本論文應用在合併色彩與輪廓特徵與學習綜合特徵中，使我們的模型可以也可以充分學習顏色與輪廓兩個特徵，並隨著層數的深入逐漸組成更完整的影像特徵資訊。

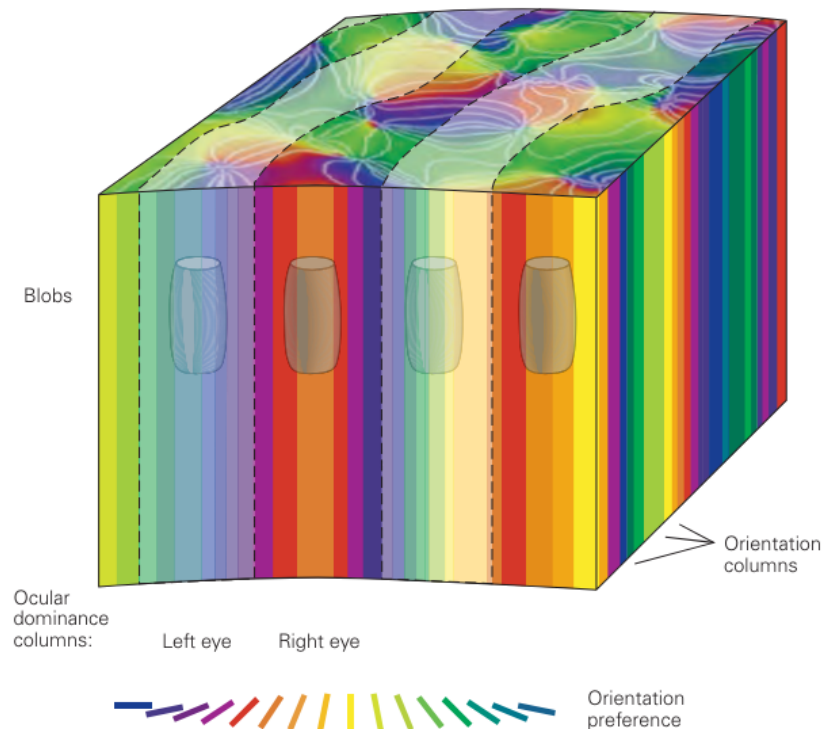


圖 2.7: 皮層計算模組 [10]



## 2.1.2 CNN-based Interpretable Model

CNN-based Interpretable Model(CIM)[5] 為 Yang 等人於 2023 年提出的可解釋性模型，該模型是模擬大腦視覺皮層的階層架構和影像的時序性關係來解釋深度學習的決策的過程，目的是希望使模型的決策過程透明化解決黑箱決策的問題。然而該模型目前只能運用於灰階影像上而無法處理彩色影像的問題，本論文的目標是基於此模型進行改進，開發出一個更適用於現實彩色影像的新模型。

### 2.1.2.1 模型架構

該模型採用多層結構，每層由三個部分組成：高斯卷積模組、特徵增強模組、空間位置保留合併模組，整體模型架構如圖 2.8，

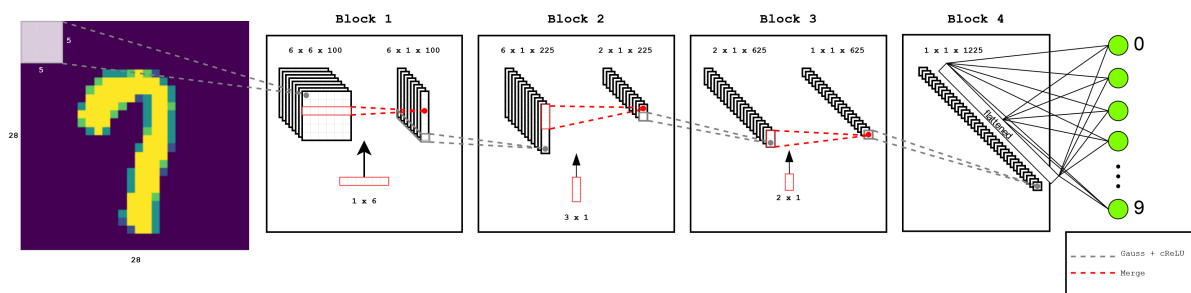


圖 2.8: CIM 架構圖 [5]

高斯卷積模組使用高斯函數 (公式如式 (2.1)) 取代原始卷積操作中的內積操作來對輸入進行特徵提取，使得該模組計算的結果具有相似度的意義並且此特性也被運用在後續的可解釋性中。

高斯函數 (Gaussian function)：

$x$  為輸入值、 $m$  為中心點、 $\sigma$  為寬度參數

$$\phi(x) = \exp\left(-\frac{\|x - m\|^2}{2\sigma^2}\right) \quad (2.1)$$

特徵增強模組使用該論文中新提出變形的 ReLU 函數稱為 changed ReLU(簡稱為 cReLU) 來過濾不重要的特徵使得後續的解釋性可以有更好

的發揮。

change ReLU 的公式如式 (2.2)， $c$  為人工取的一個閾值

$$f(x) = \begin{cases} 0 & \text{if } x < c \\ x & \text{if } x \geq c \end{cases} \quad (2.2)$$

空間位置保留合併模組為該論文所提出來的全新機制，文中他們將一組輸入在經過高斯卷積和特徵增強後產生的輸出稱為  $RM$ ，而一張影像又擁有多組資訊，從而產生出多張屬於這個影像的  $RM$ 。使用合併公式式 (3.3) 並加入時間遺忘函數 (forgetting factor) $\alpha$ ，來進行合併來模擬特徵資訊在皮層中的時序性特徵資訊與皮層的逐層合併的現象，從而達到在合併的過程中保留特徵之間的時序關係，也可讓越後面的層數學習到更完整的特徵資訊。

合併公式：

$RM_c$  為合併後的  $RM$ ， $RM_k$  為第  $k$  張  $RM$ ， $n$  為輸入資訊的數量， $\alpha$  為一個可訓練的參

$$RM_c = \frac{1}{n} \sum_{k=0}^{n-1} \alpha^k \times RM_k \quad , \text{ where } n = H_{SF}^i \times W_{SF}^i \quad (2.3)$$

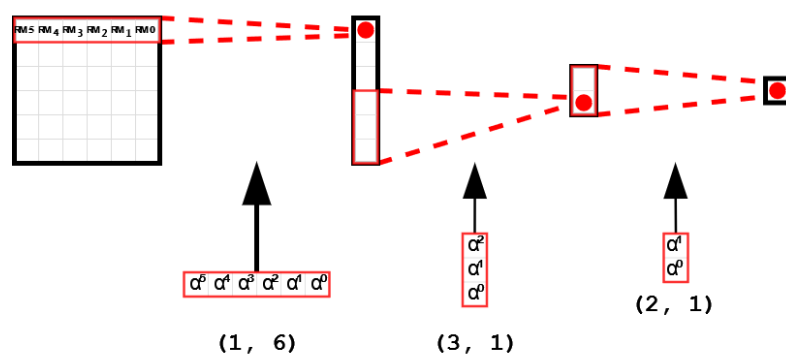


圖 2.9: 合併方式示意圖 [5]

### 2.1.2.2 可解釋性

文中將每層的高斯卷積模組的濾波器視為二階矩陣稱為特徵映射圖 (Feature Map,  $FM$ ), 一組輸入在每層經過高斯卷積模組和特徵增強模組後的輸出稱為特徵映射響應圖 (Response Map,  $RM$ ), 當模型訓練完畢後會讓每個卷積模組中的濾波器對資料集中的所有影像進行反應並且從所有反應中找到造成最大反應的影像, 每個濾波器都會有屬於自己的最大反映影像, 稱為特徵映射圖之對應影像 (Corresponding Image,  $CI$ ) 以上的  $FM, RM, CI$  便是模型提供可解釋性的核心要素, 其關係圖如圖 2.10

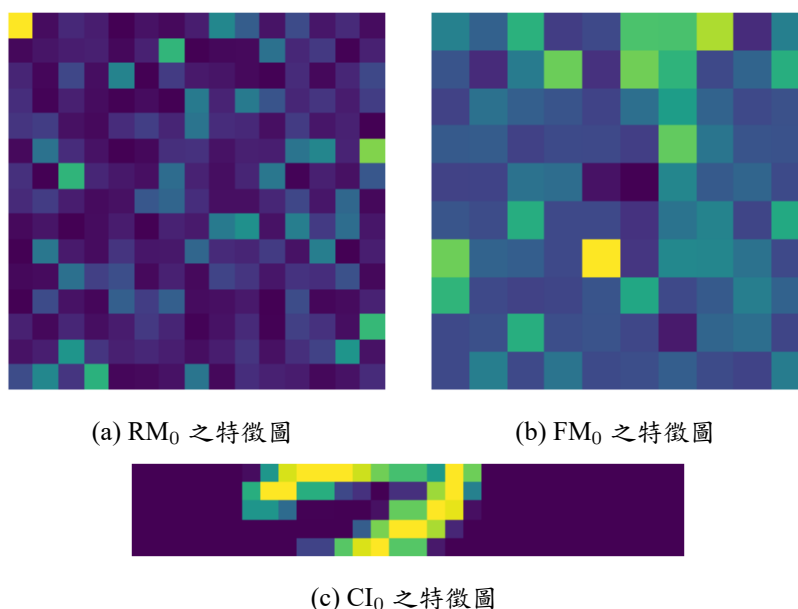
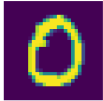
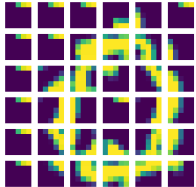
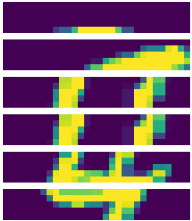
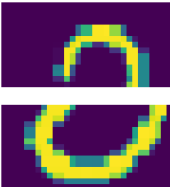
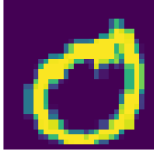


圖 2.10: FM-RM-CI[5]

在可解釋性的方法上, 論文中提出了兩個方法, 第一個方法為特徵圖對應法, 在將一張影像輸入進模型後第一步視覺化輸入影像中每個區塊在模型中每層的  $RM$  並找出  $RM$  中最大反應, 第二步從  $RM$  最大反應找出造成反應的  $FM$ , 第三步根據  $FM$  和  $CI$  的對應關係, 找出該  $FM$  對應之  $CI$ , 第四步利用每個區塊的  $CI$  根據原本的位置組合出輸入影像在模型中被分解的過程這個方法用於了解模型在經過高斯卷積、特徵增強、空間位置保留合併後所關注的特徵資訊。範例如表 2.1

表 2.1: 特徵圖對應法示意圖 [5]

| 原始影像  | RM-CI1  | RM-CI2  | RM-CI3  | RM-CI4  |
|---|---|---|---|---|
|  |  |  |  |  |

第二個方法為全連接層權重對應法，是透過將全連接層的輸入  $x$  和全連接層的權重  $w$  的乘積視為  $RM$  並且對應之  $FM$  為最後一個 Block 的  $FM$ ，透過此  $FM$  找到  $CI$  將全連接層轉成可以理解的資訊。

## 2.2 文獻回顧

本章節將向讀者介紹可解釋性人工智慧，並回顧可解釋性人工智慧的發展歷程和分類，協助讀者可以更了解可解釋性人工智慧的現狀與研究方向，

### 2.2.1 可解釋性人工智慧的演進分類

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 (Explainable Artificial Intelligence, XAI) 的研究最早可以追蹤到 1991 年的專家系統時代 W Swartout, C Paris 等人便發覺在金融、醫療，軍事等重要領域中進行的決策中往往需要一個足以支撐結論的合理解釋，便從此開始對 XAI 進行研究 [11]。

XAI 從一開始的小眾領域在近些年開始蓬勃發展成活躍的研究領域，甚至每年都會有各式各樣的 Review Paper 被發表 [12] [13]，論其原因應該歸功於近年來人工智慧已被各行各業所應用於分類、物件偵測、資料分析等任務，隨著人工智慧的普及，在重要領域中 XAI 的地位和重要性也日益上升。



根據 [14] 和 [15] 的研究，我們將可解釋性人工智慧的研究分為 Ante-hoc 和 Post-hoc 兩大類，然而這兩類方法均有各自的優缺點，值得注意的是，在 [15] 有介紹到最近有些研究展現出具有解決這兩類缺點的潛力，但其解釋性仍須更進一步的研究，因此我們將其視為不同於前兩類的第三大類。以下讓我們介紹這些類別的特性與較為著名的研究。

## 2.2.2 對於 Ante-hoc explainable 模型之研究

Ante-hoc explainable 模型 ([14] 稱為 Interently Interpretable Models)，不同於 Black Box 模型，在模型設計時就會考量可解釋性的需求，使得模型本身便具備產生解釋性的能力，其最終目標便是產生一個既準確又可以讓人理解決策過程的深度學習模型。

這類模型最著名的便是機器學習中的 Decision Tree(DT)[16]，DT 透過一系列的規則去不斷進行抉擇並且可以很容易地被可視化，使得人們可以容易理解 DT 的決策邏輯，並且也經常被用在金融等常運用表格去表示資料的領域 [17]。除此之外也有：一系列由 DT 衍伸出來的模型，例如 NGE、ANFIS、FALCON、FMMC，HRCNN, SIMTree；由 CNN 為基礎發展出來的模型，例如基於多層自我映射圖之可視覺化深度學習模型；等等其他模型

然而這類模型存在一個問題便是必須再訓練模型前找出有效的特徵提取方法，模型的可解釋性與準確度都會因為使用不同的特徵提取方法而造成很大的影響。如何在使用不複雜的特徵提取的同時提高準確度與解釋性也是這類模型的很大的課題。

#### **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。[15]

## 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，色彩感知區塊、輪廓感知區塊和特徵傳遞區塊。

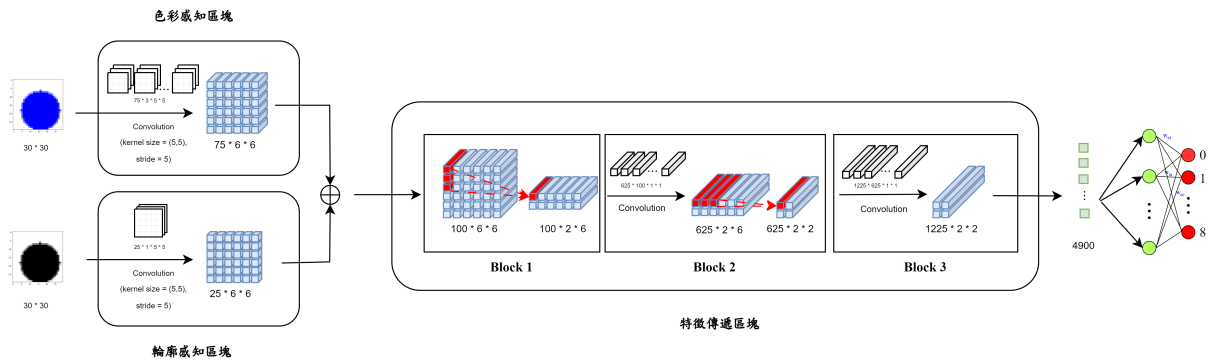


圖 3.1: 模型架構圖

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合併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

### 3.1.3 模型符號說明

由於本研究在可解釋性部分採用了 CIM 中的特徵圖解析方法來解析本論文模型的決策過程，因此在各個區塊我們均會產生特徵映射圖 (Feature Map, FM)、特徵映射響應圖 (Response Image, RM) 和特徵映射圖之對應影像 (Corresponding Image, CI) 作為可解釋性的核心要素。具體的產生過程將會在章節 3.5.1 去介紹。

$FM_{color}$ ：色彩感知區塊之高斯卷積模組的 FM

$RM_{color}$ ：色彩感知區塊之高斯卷積模組的 RM

$CI_{color}$ ：色彩感知區塊之高斯卷積模組的 CI

$FM_i^{color}$ ：色彩特徵傳遞區塊第 i 層之高斯卷積模組的 FM

$RM_i^{color}$ ：色彩特徵傳遞區塊第 i 層之高斯卷積模組的 RM

$CI_i^{color}$ ：色彩特徵傳遞區塊第 i 層之高斯卷積模組的  $FM_i$  的對應影像

$FM_{gray}$ ：輪廓感知區塊之高斯卷積模組的 FM

$RM_{gray}$ ：輪廓感知區塊之高斯卷積模組的 RM

$CI_{gray}$ ：輪廓感知區塊之高斯卷積模組的  $CI_{gray}$  的對應影像

$FM_i^{gray}$ ：輪廓特徵傳遞區塊第 i 層之高斯卷積模組的 FM

$RM_i^{gray}$ ：輪廓特徵傳遞區塊第 i 層之高斯卷積模組的 RM

$CI_i^{gray}$ ：輪廓特徵傳遞區塊第 i 層之高斯卷積模組的  $CI_i$  的對應影像

## 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濾波器 (Filter) 的初始化為不同的 RGB 色彩值來作為該區塊的權重，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、彩色卷積模組、訓練過程中的正規化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 [18] 中的方式將每個值初始化範圍為  $[-\sqrt{\frac{6}{fan\_in}}, \sqrt{\frac{6}{fan\_in}}]$ ， $fan\_in$  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 Uniform 相較於常用的 Uniform 初始化和 Xavier Uniform [19] 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  $\frac{1}{\sqrt{n}}$  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 [18] 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 彩色卷積模組

### 3.2.3 訓練過程的正規化

在訓練的過程中由於濾波器的數值可能會隨著梯度下降而產生小於零或著大於一的數值，但是由於輸入顏色的範圍為  $[0, 1]$  之間。因此為了確保濾波器也穩定在這個範圍內，並且當模型訓練完時需要將濾波器變成特徵映射圖 (FM) 來進行可視化呈現，我們在模型每次完成倒傳遞並更新完參數後，便對色彩提取區塊之高斯卷機模組的濾波器進行 max-min 正規化計算，使濾波器的範圍穩定維持在  $[0, 1]$  之間，從而保證顏色呈現的準確度。

### 3.3 輪廓感知區塊之前處理設計

在輪廓感知區塊中，為了使該區塊能夠專注於提取輪廓特徵而不需考慮色彩因素，我們會對輸入影像進行前處理。首先，我們會將彩色影像轉換為灰階影像，轉換公式如式 (3.1)，Gray 代表灰階值，R、G、B 各代表紅、綠、藍三色通道的值。

$$Gray = 0.299 * R + 0.587 * G + 0.114 * B \quad (3.1)$$

這種加權方法主要考慮了人眼對於不同色彩的敏感度，從而能夠較為完整保留彩色圖像中的細節，接著為了消除不同色彩導致的灰度值差異，使該區塊能夠專注於輪廓資訊而不考慮色彩資訊，我們接著進行了 max-min 正規化，使每張影像的灰階值統一道 0 ~ 1 之間。透過上面步驟的前處理，我們從而可以確保輪廓感知區塊能夠專注於提取輪廓的特徵。

彩色圖片與灰階圖片



### 3.4 特徵傳遞區塊之優化設計

特徵傳遞區塊以 CIM 為基礎，對 CIM 的原始區塊進行深入分析和改進並對其進行優化，我們希望可以這些優化措施可以在可解釋性不變的同時，又可以對模型的效能和準確度進行提升，此外我們還希望這些優化措施可以將一些原本人工去分析資料集後指定的參數變成可訓練參數，進一步提高模型的自動化與適應能力。

#### 3.4.1 高斯卷積模組優化設計

補 CNN 的 cite 在高斯卷積模組中，CIM 的原始實作方法修改了傳統 CNN 中的卷積操作，將 CNN 卷積操作的內積轉換為放射狀積底函數的高斯函數，目的是希望以距離為基礎的高斯函數可以表達輸入與 Filter 之間的距離關係並將之視為相似度。

在實作上 CIM 採用了逐個計算每個 Windows 和濾波器之間的歐式距離並將結果輸入進高斯函數得出相似度。我們改進了這個實作過程，利用了 GPU 的多核優勢，將每次卷積的所有位置的 Windows 取出來後平行放入 GPU 的多個核心，同時計算這些 Windows 與濾波器的歐式距離再輸入高斯函數中得出相似度，這樣的實作方式使得整體模型的訓練速度可以大幅提高，也完整利用 GPU 的效能，提升了模型效率。

可以加入原本的卷積方式和平行處理的方式的示意圖

#### 3.4.2 特徵增強模組之優化設計

在特徵增強模組中，CIM 使用了自行設計的增強特徵的整流線性單位函數 (changed ReLU, 簡稱 cReLU)，希望透過閾值過濾不重要的特徵，其公式如式 (2.2) 所示。作為閾值的  $c$  值需要透過觀察資料集在高斯卷積模組的輸出來設定具體不同的  $c$  的數值來適應不同的資料集。然而一旦閾值設定過小則會削弱特徵增強的效果，閾值設定過大則會將大部分的

特徵都歸零導致無法學習出有效特徵，因此如何設定  $c$  則成為一個非常困難的問題，並且每次更換資料集都必須面對這個問題。

為了解決閥值設定的問題，我們對 cReLU 的公式進行改善，首先設定一個希望保留的  $RM$  元素的百分比參數  $p\%$ ， $RM$  的數目稱之為  $C_{RM}$  接著從同一影像在該層的所有  $RM$  中取出第  $p\% * C_{RM}$  個最大的元素稱之為  $P_{RM}$ ，並且將所有小於  $P_{RM}$  的數值歸零。如此一來，使用者只需要決定希望可以保留  $p\%$  的數值，無須糾結於設定閥值  $c$  具體數值。這種方法簡化了參數設定過程，同時也確保了特徵增強的有效性。改善後的 cReLU 的公式為

$$f(x) = \begin{cases} 0 & \text{if } x < P_{RM} \\ x & \text{if } x \geq P_{RM} \end{cases}, \text{ where } P_{RM} = RM_i[p\% * C_{RM}] \quad (3.2)$$

### 3.4.3 空間位置保留機制之優化設計

在空間保留機制模組上，CIM 為了保留  $RM$  之間的空間位置關係簡化了眼球跳動的方式，加入了一個可訓練的時間遺忘函數  $\alpha$  進入合併公式式 (3.3)。CIM 之合併公式：

$RM_c$  為合併後的  $RM$ ， $RM_k$  為第  $k$  張  $RM$ ， $n$  為輸入資訊的數量， $\alpha$  為一個可訓練的參數

$$RM_c = \frac{1}{n} \sum_{k=0}^{n-1} \alpha^k \times RM_k, \text{ where } n = H_{SF}^i \times W_{SF}^i \quad (3.3)$$

這樣會導致越早被看到(時間越早)的數值需要乘上的  $\alpha^k$  的  $k$  值越大，在合併時的加權也就越小。儘管這種做法確實可以使  $RM$  呈現出空間上的關係，然而當我們需要合併的  $RM_k$  越多時，隨著  $k$  值越來越大， $\alpha^k$  會快速變小。這樣便造成當我們將  $RM_k$  乘上過小的  $\alpha^k$  後，便會變得這個數值便會在  $RM_c$  中變得無足輕重。

#### 可以放入 $\alpha$ 的指數變化圖

對於大小較大的影像而言這種方法是有問題的，這會使得影像逐漸喪失一部分  $RM_k$  所擷取的特徵。因此我們提出了將  $\alpha$  的方法從  $\alpha^k$ ，改成從  $[0.9 \sim 0.99]$  之間進行等距採樣，得到  $k$  個值代替  $\alpha^k$  按照時序順序去乘以  $RM_k$ ，我們將第  $k$  個 RM 所對應的  $\alpha$  稱為  $\beta_k$ ，如此便可以改善當需要合併的  $RM$  數量過多時  $\alpha^k$  過小的問題，同時也保留了 CIM 當初設計時希望保留的  $RM$  之間的對應空間關係。

改善後的合併公式如式 (3.4)， $\beta_k$  從  $[0.9 \sim 0.99]$  之間取出的第  $k$  個值：

$$RM_c = \frac{1}{n} \sum_{k=0}^{n-1} \beta_k \times RM_k \quad , \text{ where } n = H_{SF}^i \times W_{SF}^i \quad (3.4)$$

#### 可以放等距取值示意圖 [5]

### 3.4.4 模型流程的精簡

我們將原來在 CIM 的模型訓練流程中每層接在空間位置保留合併模組後的維度轉置與 Reshape 步驟 (在 [5] 模型流程的 Step 4) 去除，CIM 會加入這個步驟的初衷在於方便之後輸出將濾波器輸出成可解釋性圖片，然而加入這個步驟導致一些不必要的問題：首先，CIM 原本在第一層高斯卷積後的高斯卷積模組都只是做 (1,1) 的卷積操作，但是加入了這個步驟就會將後續的高斯卷積層的濾波器大小改變，容易引起閱讀者的誤解。其次，這個步驟並不會增加準確度或是增加效能，反而是多一些所轉置與 Reshape 所需要的時間。基於以上原因我們決定將此步驟捨棄，使過程容易理解、減少所需時間。

#### 加入我們刪除了原先模型流程中的哪個步驟

## 3.5 可解釋性

本研究使用了 Yang 在 2023 年 CIM 論文中提出的特徵映射法做為解析模型的決策過程的基礎，使得模型的決策過程可以被視覺化出來，讓使用者了解模型關注的特徵資訊與決策背後的邏輯關係。

### 3.5.1 FM、RM、CI 的意義

特徵映射圖 (FM)、特徵映射響應圖 (RM)、特徵映射圖之對應影像 (CI)，為 Yang 在 CIM 論文中提出的名詞，也是特徵映射法的核心三要素。

特徵映射響應圖 (RM) 為當輸入進入高斯卷積模組後所得到的輸出，其代表的是輸入對於該高斯卷積模組的濾波器的相似度，也被視為輸入對該高斯卷機模組所有的濾波器的反應強度。同一層、不同位置的 RM 數值為輸入影像的不同位置對於某個濾波器的反應，不同層、同一位置的 RM 數值為輸入影像的同一位置對於不同濾波器的反應。

特徵映射圖 (FM) 為將高斯卷積模組的濾波器提取出來並 Reshape 成二維矩陣所形成，由於模型在高斯卷積模組為計算輸入濾波器與相似度，因此 FM 可以被視為模型學習到了輸入影像中的何種特徵。

特徵映射圖之對應影像 (CI) 為記錄資料集中所有影像對特定濾波器的反應，從中選出與該濾波器有最大的反應的影像，並該輸入影像視為該濾波器的對應影像。會需要 CI 的原因在於除了顏色感知層和輪廓感知層之外，後續的特徵傳遞層的輸入均為前一層所計算出來的相似度，人類已無法去解讀出 FM 的意義。因此我們需要透過 CI 來找出該濾波器與何種影像最相似，幫助使用者理解 FM 代表的特徵長相。

RMCI 為根據輸入影像在某層的高斯卷積模組的每一張 RM 的最大反應的位置來找到對應位置的 CI 影像。

### 3.5.2 色彩感知區塊之可解釋性

在色彩感知區塊的 FM 本身即可被視為  $C_{out}$  個不同 RGB 顏色，並且這個特性導致 FM 不會學習到輪廓的特徵。然而由於 CI 是由資料集中影像的部分區塊組成，因此 FM 對應的 CI 的影像反而卻同時存在顏色和輪廓，這會導致最後組合起來導致 CI 容易出現顏色相似，但輪廓與輸入影像完全不同的情況，看起來會十分怪異。為了解決這個問題，我們會取每個 CI 的平均色彩作為該 CI 的代表色，我們接著會將 CI 的代表色擴張為和原始 CI 相同大小色塊，並且使用這些色塊代替原本的 CI 進行呈現，從而更好的反應 FM 學習到的色彩特徵。

總結色彩感知區塊產生可解釋性圖片的流程如下：

- 1 輸入影像後，找出色彩感應區塊的 RM
- 2 從 RM 中找出每個 RM 的最大反應
- 3 根據 RM 的最大反應位置找出對應的 CI
- 4 找出對應 CI 的代表色並擴張成與原始 CI 相同大小的色塊
- 5 利用色塊組出輸入影像在該層的代表影像

### 3.5.3 色彩特徵傳遞區塊之可解釋性

在色彩特徵區塊上，由於經過空間合併模組的特徵合併，因此該區塊所學到的是比色彩感知區塊更加完整的特徵。為了準確地呈現 FM 對應的 CI 的不同的顏色與位置，我們採用以下的方法進行處理：

首先，我們找到輸入影像在該區塊的高斯卷積模組對應的 CI。接著，將這些 CI 切成與  $CI_{color}$  相同大小的小圖，然後，對這些小圖分別取平均顏色並將這些平均顏色擴張為和  $CI_{color}$  相同大小的色塊，用這些色塊代替原本的 CI 按照對應位置組成 RM-CI 影像。再加入了分割這個步驟後，

每個 CI 都能反映出不同部分對應的顏色和位置，從而更好的呈現 CI 的顏色分布。

總結色彩特徵傳遞區塊產生可解釋性圖片的流程如下：

- 1 輸入影像後，找出色彩感應區塊的 RM
- 2 從 RM 中找出每個 RM 的最大反應
- 3 根據 RM 的最大反應位置找出對應的 CI
- 4 將對應的 CI 分割成數個與  $CI_{color}$  相同大小的小圖
- 5 找出各個小圖的代表色並擴張成與原始 CI 相同大小的色塊
- 6 利用色塊組出輸入影像在該層的代表影像

#### 3.5.4 輪廓感知區塊和特徵學習區塊之可解釋性

## 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 實際用於現實瘡疾影像上的效果

### 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] D. Purves, G. J. Augustine, D. Fitzpatrick, *et al.*, “Neuroscience, 3rd ed.,” in Sinauer Associates Inc., 2004, ch. 10-11, pp. 229–282.
- [7] M. Bear, B. Connors, and M. Paradiso, *Neuroscience: Exploring the Brain*. Wolters Kluwer, 2016.
- [8] H. Baier, “Synaptic laminae in the visual system: Molecular mechanisms forming layers of perception,” *Annual Review of Cell and Developmental Biology*, vol. 29, no. Volume 29, 2013, pp. 385–416, 2013.
- [9] J. Hawkins and S. Blakeslee, *On Intelligence*. USA: Times Books, 2004.
- [10] E. R. Kandel, J. D. Koester, S. H. Mack, and S. A. Siegelbaum, in *Principles of Neural Science*, 6e. New York, NY: McGraw Hill, 2021.
- [11] 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.
- [12] S. A. and S. R., “A systematic review of explainable artificial intelligence models and applications: Recent developments and future trends,” *Decision Analytics Journal*, vol. 7, p. 100 230, 2023.

- [13] V. Chamola, V. Hassija, A. R. Sulthana, D. Ghosh, D. Dhingra, and B. Sikdar, “A review of trustworthy and explainable artificial intelligence (xai),” *IEEE Access*, vol. 11, pp. 78 994–79 015, 2023.
- [14] I. E. Nielsen, D. Dera, G. Rasool, R. P. Ramachandran, and N. C. Bouaynaya, “Robust explainability: A tutorial on gradient-based attribution methods for deep neural networks,” *IEEE Signal Processing Magazine*, vol. 39, no. 4, pp. 73–84, Jul. 2022.
- [15] 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.
- [16] L. Rokach, “Decision forest: Twenty years of research,” *Information Fusion*, vol. 27, pp. 111–125, 2016.
- [17] L. Grinsztajn, E. Oyallon, and G. Varoquaux, “Why do tree-based models still outperform deep learning on typical tabular data?” *Advances in neural information processing systems*, vol. 35, pp. 507–520, 2022.
- [18] 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.
- [19] 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.