

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
一、緒論	1
1.1 研究動機	1
1.2 研究目的	2
1.3 論文架構	2
二、背景知識以及文獻回顧	3
2.1 背景知識	3
2.1.1 卷積神經網路	3
2.1.2 人如何感知彩色影像	3
2.2 文獻回顧	3
2.2.1 可解釋性人工智慧的演進與分類	3
2.2.2 對於 Inherently Interpretable 可解釋性模型之研究	4
2.2.3 對於 Post-hoc 可解釋性模型之研究	4
2.2.4 近年可解釋性模型趨勢之研究	4
2.2.5 以卷積神經網路為基礎的具可解釋性的深度學習模型	4

三、	研究方法	5
3.1	對以卷積神經網路為基礎的具可解釋性的深度學習模型之改進	5
3.1.1	優化模型流程與新增平行處理	5
3.1.2	優化空間位置保留機制之設計	5
3.1.3	優化放射狀基底函數	5
3.2	以卷積神經網路為基礎的 RGB 三通道可解釋性模型	5
3.2.1	模型架構	5
3.2.2	模型參數	5
3.2.3	RGB 三通道卷積模組設計與實現	5
3.2.4	模型流程	5
3.3	量化推論成果之方法	5
四、	實驗設計與結果	6
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	6
4.1.1	資料集介紹	6
4.1.2	實驗設計	6
4.1.3	實驗結果	6
4.2	模型保留空間位置特徵之臉部驗證實驗	6
4.2.1	實驗背景與目的	6
4.2.2	資料集介紹	6
4.2.3	模型架構與參數	6
4.2.4	實驗結果	6
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	7
4.3.1	資料集介紹	7
4.3.2	實驗設計	7

4.3.3	實驗結果	7
4.4	實際用於現實瘧疾影像上的效果	7
4.4.1	資料集介紹	7
4.4.2	模型架構與參數	7
4.4.3	實驗結果	7
五、	總結	8
5.1	結論	8
5.2	未來展望	8
	參考文獻	9

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

1.3 論文架構

二、 背景知識以及文獻回顧

2.1 背景知識

2.1.1 卷積神經網路

2.1.2 人如何感知彩色影像

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: [5] [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 [6]，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。[7]

2.2.5 以卷積神經網路為基礎的具可解釋性的深度學習模型

三、 研究方法

3.1 對以卷積神經網路為基礎的具可解釋性的深度學習模型之改進

3.1.1 優化模型流程與新增平行處理

3.1.2 優化空間位置保留機制之設計

3.1.3 優化放射狀基底函數

3.2 以卷積神經網路為基礎的 RGB 三通道可解釋性模型

3.2.1 模型架構

3.2.2 模型參數

3.2.3 RGB 三通道卷積模組設計與實現

3.2.4 模型流程

3.3 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)*, 2016.
- [3] European Parliament and Council of the European Union. “Regulation (EU) 2016/679 of the European Parliament and of the Council,” of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). (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] L. Rokach, “Decision forest: Twenty years of research,” *Information Fusion*, vol. 27, pp. 111–125, 2016.
- [6] 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.
- [7] 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.