An Image Patch Row-Column Ranking Method Using the Feature Accumulation Matrix to Explain Decisions of a Convolutional Neural Network

Luna M. Zhang
Department of Computer Science
Stony Brook University
Stony Brook, NY 11794-2424, USA
lunamzhang@gmail.com

1. Introduction

Various 2D methods [1–15], such as Grad-CAM [2] and the Patch Ranking Map [16], use 2D image information to explain image classifications of a machine learning model using all extracted features. To sufficiently interpret decisions of a Convolutional Neural Network (CNN), it is necessary to additionally use 1D image information associated with image patch rows and image patch columns. For example, a medical doctor may need to know about which image patch rows and image patch columns are most important for diagnosis. Since feature selection (FS) is useful in not only improving the model performance but also in interpreting a deep neural network [12, 13], we develop an efficient and accurate CNN model by adding a novel layer called the "FS Layer." In this paper, we create an image patch row-column ranking method to rank the top image patch rows and top image patch columns to use 1D image information for explaining decisions of the CNN model with the FS layer.

2. Image Patch Row-Column Ranking

A CNN extracts n $H \times W$ feature maps F^l (assuming the feature map shape is $H \times W \times n$) with features f^l_{ij} for $i = 0, 1, \ldots, H-1$, $j = 0, 1, \ldots, W-1$, and $l = 0, 1, \ldots, n-1$. The n feature maps are converted to m flattened features for $m = n \times H \times W$. The m features have m feature index numbers (0, 1, …, $m-1$). A FS layer selects the top k features from the m features. The k selected features have k feature index numbers I_p for $I_p \in 0, 1, \ldots, m-1$ for $p = 0, 1, \ldots, k-1$. A top feature with index I_p is associated with a feature map F^{q_p} where $q_p = I_p \bmod n$ for $p = 0, 1, \ldots, k-1$. Let $\bar{Q} = \{q_0, q_1, \ldots, q_{k-1}\}$. After eliminating duplicated elements in Q_S, we get Q with distinct elements for $Q \subseteq \bar{Q}$ [16].

Definition 1: Let the top feature map T^q have features t^q_{ij} for $i = 0, 1, \ldots, H-1$, $j = 0, 1, \ldots, W-1$, and $q \in Q$. If f^q_{ij} in a feature map F^q is a selected feature, then $t^q_{ij} = f^q_{ij}$, otherwise $t^q_{ij} = 0$.

Definition 2: Let the "feature binary matrix" B have binary numbers b_{ij} for $i = 0, 1, \ldots, H-1$, and $j = 0, 1, \ldots, W-1$. If f_{ij} is a selected feature, then $b_{ij} = 1$, otherwise $b_{ij} = 0$.

Definition 3: Let the "feature accumulation matrix" A have elements called "feature accumulators" a_{ij} for $i = 0, 1, \ldots, H-1$ and $j = 0, 1, \ldots, W-1$, where $a_{ij} = \sum_{s=1}^{N} b^s_{ij}$ where b^s_{ij} is an element of the feature binary matrix B^s, and N is the number of feature maps.

Algorithm 1 Image Patch Row-Column Ranking

Input: A feature accumulation matrix A.

Output: Top N patch rows for $1 \le N < H$ and top M patch columns $1 \le M < W$.

1: Step 1: Calculate the patch row-wise feature accumulation number A^i_{row}: $A^i_{row} = \sum_{j=0}^{W-1} a_{ij}$ for $i = 0, 1, \ldots, H-1$.
2: Step 2: Sort $\{A^0_{row}, A^1_{row}, \ldots, A^{H-1}_{row}\}$ to get top N patch row numbers.
3: Step 3: Calculate the patch column-wise feature accumulation number A^j_{column}: $A^j_{column} = \sum_{i=0}^{H-1} a_{ij}$ for $j = 0, 1, \ldots, W-1$.
4: Step 4: Sort $\{A^0_{column}, A^1_{column}, \ldots, A^{W-1}_{column}\}$ to get top M patch column numbers.

3. Performance Analysis

The Alzheimer's MRI preprocessed dataset with $6,400$ 128×128 images [17] are used for 4-class classification performance analysis for determining the 4 stages of Alzheimer's Disease (AD). The $6,400$ images are divided into $5,119$ training images, 642 test images, and 639 validation images. The final MaxPooling layer of the fine-tuned pretrained ResNet50 model generates 64 16×16 feature maps. The 16×16 feature map has 256 features associated with 256 8×8 patches of an original 128×128 image

that has 16 patch rows and 16 patch columns. A patch row or a patch column has 16 patches. We used seven different FS methods, including three FS algorithms using Chi2 [18], f_regression [19] and f_classif [20] respectively, the FS algorithm using f_regression and the RFE, the FS algorithm using Chi2, f_classif and RFE [21, 22], the FS algorithm using Chi2, mututal_info_regression [23], f_classif and the RFE, and the FS algorithm using Chi2, f_regression, mutual_info_regression, f_classif and the RFE. The RFE method is applied to rank the top features for each top feature set.

Simulations indicate that the fine-tuned ResNet50-FS using 800 selected features with test accuracy of 0.9891 is better than the conventional ResNet50 using all $16,384$ features with test accuracy of 0.9642. Top 188 common features based on average feature rankings from seven 800-feature sets generated by the seven FS methods are selected. Fig. 1 shows row-wise feature accumulation numbers (highlighted in blue) and column-wise feature accumulation numbers (highlighted in green). The more features a patch row or a patch column has, the more important it is for decision making. For example, patch row 0 and patch row 6 having row-wise feature accumulation numbers 0 (minimum) and 37 (maximum) are the least important and the most important for decisions, respectively. Every element of a feature accumulation matrix for the 64 16×16 feature maps of the ResNet50 model without FS is 64, so all row-wise and column-wise feature accumulation numbers are $1,024$. Thus, the traditional feature accumulation matrix is not useful for ranking patch rows and patch columns.

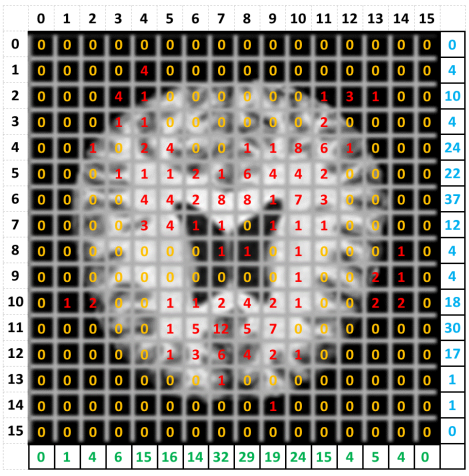

Figure 1. A feature accumulation matrix for 188 common features.

3.1. Relationship among Top Patch Rows, Top Patch Columns and Relevant Brain Areas

Four typical training images (non_67.jpg, verymild_777.jpg, mild_123.jpg, and moderate_2.jpg) for four classes are used to find relevant brain areas by using the ebrains software tool (Human Brain Atlas) [24]. 59 relevant brain areas are found. They include 10 brain areas associated with AD, 43 brain areas that may be associated with AD, and 6 brain areas that may not be associated with AD based on the information [25–27]. The image patch row-column ranking algorithm finds two top-ranked patch rows 6 and 11 and two top-ranked patch columns 7 and 8.

Table 1 shows that patch rows 6 and 11 have 6 brain areas among the 10 brain areas associated with AD and 9 brain areas among the 43 brain areas that may be associated with AD. Table 1 also shows that patch columns 7 and 8 have 5 brain areas among the 10 brain areas associated with AD and 6 brain areas among the 43 brain areas that may be associated with AD. The top patch rows and the top patch columns are not associated with any brain areas that may not be associated with AD. Thus, they are associated with important brain areas related to AD. A doctor may use 1D information in the top patch rows and top patch columns to partially interpret a diagnosis.

Table 1. Brain Areas Related to the Top Rows and Top Columns ("front to occi" = "frontal to occipital", red: associated with AD, blue: may be associated with AD)

Row 6	Row 11	Column 7	Column 8
front to occi	front to occi	front to occi	front to occi
temporal to parietal	33 (ACC)	33 (ACC)	33 (ACC)
PFcm (IPL)	p24c	p24ab	p24ab
TPJ (STG/SMG)	p24ab	p24c	p24c
TE 2.2 (STG)	p32	p32	p32
PF (IPL)	45 (IFG)	Fp1	Fp1
PFm (IPL)	frontal-I	Fp2	Fp2
PFop (IPL)	IFS1 (IFS)	hOc1	hOc1
		hOc2	hOc2
		hOc3d	hOc3d
		frontal-I	frontal-I

4. Conclusions and Future Works

The top-ranked patch rows 6 and 11 and the top-ranked patch columns 7 and 8 have the most top-ranked common features, 6 brain areas associated with AD, and 15 brain areas that may be associated with AD. The new patch row-column ranking method can generate useful 1D image-row-column information to interpret decisions of a CNN model.

The new method will be developed to analyze the new relationship among a brain disease such as AD, top patches [16], top patch rows, and top patch columns. It is useful to use hybrid 2D-1D information in important brain areas associated with the top patches, top patch rows, and top patch columns to make robust and explainable decisions such as medical diagnosis.

References

[1] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. *2016 IEEE Conference on Computer Vision and Pattern Recognition (2016 CVPR)*, pages 2921–2929, 2016.

[2] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *2017 IEEE International Conference on Computer Vision (2017 ICCV)*, pages 618–626, 2017.

[3] Score-cam++: Class discriminative localization with feature map selection. *Journal of Physics: Conference Series, vol. 2278, 2022 6th International Conference on Machine Vision and Information Technology (CMVIT 2022), doi:10.1088/1742-6596/2278/1/012018.*, 2022.

[4] S.S.Y. Kim, N. Meister, V.V. Ramaswamy, Fong R., and O. Russakovsky. Hive: Evaluating the human interpretability of visual explanations. *The 1st Explainable AI for Computer Vision (XAI4CV) Workshop at CVPR 2022*, pages 1–18, 2022.

[5] E. Tjoa and C. Guan. A survey on explainable artificial intelligence (xai): towards medical xai. *arXiv:1907.07374v5 [cs.LG]*, pages 770–778, 2020.

[6] A. Wang, W.-N. Lee, and X. Qi. Hint: Hierarchical neuron concept explainer. *The 1st Explainable AI for Computer Vision (XAI4CV)Workshop at CVPR 2022*, pages 1–50, 2022.

[7] Schöttl A. Improving the interpretability of gradcams in deep classification networks. *The 3rd International Conference on Industry 4.0 and Smart Manufacturing*, pages 620–628, 2020.

[8] Q. Zhang, Y.N. Wu, and S.-C. Zhu. Interpretable convolutional neural networks. *arXiv:1710.00935v4 [cs.CV]*, pages 770–778, 2020.

[9] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. *2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, USA, 2018*, pages 839–847.

[10] H. Wang, Z. Wang, M. Du, F. Yang, Z. Zhang, S. Ding, P. Mardziel, and Hu X. Score-cam: Score-weighted visual explanations for convolutional neural networks, 2020. arXiv:1910.01279 [cs.CV].

[11] A. Singh, S. Sengupta, and V. Lakshminarayanan. Explainable deep learning models in medical image analysis. *Journal of Imaging*, 6(52):1–19, 2020.

[12] S. Khanal, B. Brodie, X. Xing, A.-L. Lin, and N. Jacobs. Causality for inherently explainable transformers: Cat-xplain. *The 1st Explainable AI for Computer Vision (XAI4CV) Workshop at CVPR 2022*, pages 1–3, 2022.

[13] T. He, J. Guo, N. Chen, X. Xu, Z. Wang, K. Fu, L. Liu, and Z. Yi. Medimlp: Using grad-cam to extract crucial variables for lung cancer postoperative complication prediction. *IEEE Journal of Biomedical and Health Informatics*, pages 1762–1771, 2019.

[14] AD Arya, SS Verma, P Chakarabarti, T Chakrabarti, AA Elngar, AM Kamali, and M. Nami. A systematic review on machine learning and deep learning techniques in the effective diagnosis of alzheimer's disease. *Brain Informatics*, 10(1):17, 2023.

[15] A Sarica, F Aracri, MG Bianco, F Arcuri, A Quattrone, and A Quattrone. Explainability of random survival forests in predicting conversion risk from mild cognitive impairment to alzheimer's disease. *Brain Informatics*, 10(1):31, 2023.

[16] L.M. Zhang. Patch ranking map: Explaining relations among patches, top-ranked features and decisions of a convolutional neural network. *2024 International Joint Conference on Neural Networks (IJCNN 2024)*, 2024.

[17] Alzheimer mri preprocessed dataset, https://www.kaggle.com/datasets/sachinkumar413/alzheimer-mri-dataset. 2023.

[18] chi2. sklearn.feature_selection.chi2. https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.chi2.html. 2024.

[19] f_reg. sklearn.feature_selection.f_regression. https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.f_regression.html. 2024.

[20] f_cla. sklearn.feature_selection.f_classif. https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.f_classif.html. 2024.

[21] RFE. Feature ranking with recursive feature elimination. https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.rfe.html. 2024.

[22] I. Guyon, J. Weston, S. Barnhill, and Vapnik V. Gene selection for cancer classification using support vector machines. *Machine Learning*, pages 389–422, 2002.

[23] m_i_reg. sklearn.feature_selection.mutual_info_regression. https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.mutual_info_regression.html. 2024.

[24] ebrains. Human brain atlas, 2024. https://www.ebrains.eu/tools/human-brain-atlas.

[25] H. I.L. Jacobs, M. P.J. Van Boxtel, J. Jolles, F.R.J. Verhey, and H. B.M. Uylings. Parietal cortex matters in alzheimer's disease: An overview of structural, functional and metabolic findings, 2012. Neuroscience & Biobehavioral Reviews, vol. 36, issue 1, pp. 297-309, January 2012.

[26] S. Holroyd, M. L. Shepher, and J. H. Downs III. Occipital atrophy is associated with visual hallucinations in alzheimer's disease, 2012. The Journal of Neuropsychiatry and Clinical Neurosciences, https://doi.org/10.1176/jnp.12.1.25, Feb. 1, 2000.

[27] ChatGPT4. Chatgpt 4, 2024. https://chat.openai.com/.