Dictionary Based Saliency Feature Extractors for Explaining Vision Transformers and ResNets

Steve Mendoza[1], Sean Wu[2], and Fabien Scalzo[2,3]

[1] Laboratory of FMRI Technology, Mark and Mary Stevens Neuroimaging and Informatics Institute, Keck School of Medicine, University of Southern California
steveame@usc.edu
[2] Seaver College, Pepperdine University, Malibu, CA 90265
[3] Department of Computer Science, University of California, Los Angeles (UCLA), CA 90095

Abstract. Deep neural networks and transformer networks have found utility in medical imaging but determining what regions are salient remains difficult. We use the theory of dictionary learning as a possible solution to this problem. We use the Big Healthy Brains (BHB) dataset, a collection of 10 different sites of healthy T1 MRI brain images. It can show subtle salient features that are likely used in the models trained using ResNet and Vision Transformers. We use our methodology to analyze gender classification in brain MRI using one axial slice. we train a model to achieve 84 percent accuracy using ResNet 18 and a model using visual transformers with 73 percent accuracy. With subtle model trimming we heavily influence the performance with Resnet showing 71 percent but transformers showing 71 percent as well. We hypothesize that transformer and Resnet models may use different features and that transformers might be more robust in feature selection.

1 Introduction

Deep learning has shown promise in diagnosis and data processing in medicine but determining how it works is still unknown. A typical deep learning uses 100-1000 images [11], and may involve multi-modal datasets. However, results interpretation is challenging. Model dependent tools such as Grad-CAM, [9] exist for convolutional neural networks (CNN). [2] We argue that dictionary learning can interpret deep learning models as diverse as CNN and transformer networks [10]. The dictionary learning directly applies as a model agnostic interpretation approach applying for CNN and transformer networks.

The authors base their work partly on. [7] In this work we compare the classical dictionary learning algorithm [1] with the L4 dictionary [12]. L4 dictionary based norm is a recent algorithm [12] that uses the L4 norm as compared to L1 norm [1] The details are described in the aforementioned work. In short, we have two populations, population A and B. For each member of population A, we take difference images comparing the population A member to each member in population B. Out of all of these images, we keep the 3 closest images, as

measured using Euclidean distance metric. Once we have built the dataset, the dataset size would be the number of members of Population A times the number of neighbors that there are in our case it would be 3.

1.1 Related Works

The principle behind our approach comes from [4]. A paper [5] implements [4] for brain vasculature. Our approach emphasizes interpretability of visual features used in the work, while keeping the same mathematical framework. Another work that used K-SVD and dictionary learning is [8], analyzing fMRI subtraction images.

1.2 Comparing the two dictionary learning algorithms

In this work, we consider two algorithms and treat them as general feature extractors for any medical image with defined anatomy. Both algorithms solve the decomposition problem:

$$Y = DX \tag{1}$$

In the equation above, the Y matrix represents our image data, represented as vectors, D is a sparse dictionary, and X is a sparsity matrix.

To solve this equation, we solve the following norm:

$$\max_{B_k} \quad \|E_k - b_k c_k\| \tag{2}$$

The L4 norm dictionary extracts information globally rather than column wise. We address this limitation with the K-SVD by considering the top 6 dictionary atoms and averaging them to come up with a representative dictionary. The L4 provides us with certain reproducible segments for us to consider.

2 Methodology

2.1 Dictionary Learning Processing and Dataset

We start with the data set with Big Healthy Brains (BHB). [3], a large multi-center study with 3208 images, with preprocessed data. We use the K-SVD method as described in previous work [**?**]. For all experiments we only use one axial slice as shown in Figure 1. The process involves making difference images of population A- population B, and keeping 3 images for each image of the closest relative to the image in the other population. We train a dictionary learning algorithm to differentiate the populations. The end result is 6 dictionary atoms. We then build a dictionary by using Population B - Population A, the inverse of the first operation and repeat the same steps. We then combine the most prominent features of both these dictionaries and then threshold them to make a mask to mask out these features. We use a validation fold in the data provided

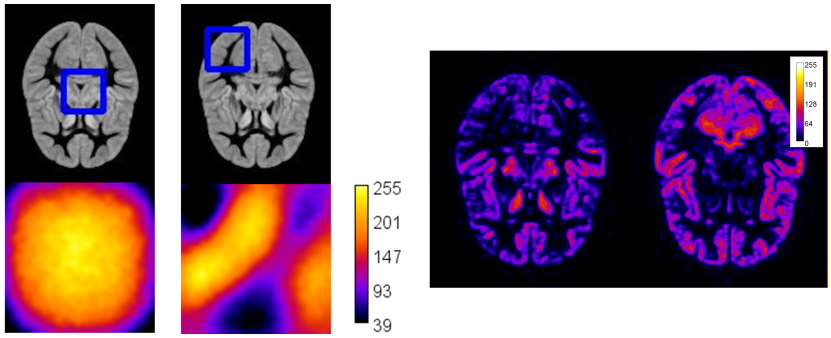

Fig. 1: Left: L4 dictionary patch based atoms. Right: KSVD Dictionary learning atoms all combined, with the brighter regions corresponding to more salient regions.

for most of our tests and the test set in the BHB for the external test set given in the results section.

We then combine these and we keep the top 1 percent of features. Once we have the top 1 percent of features, we then take this mask and multiply each case in the validation data by the mask. So we remove 1 percent of the features, by which in this case are the pixels, and compare the validation accuracy of the original image and by applying this mask.

L4 dictionary preprocessing We implement the L4 dictionary as described in earlier work using 32x32 patches, as shown in Figure 1. We consider the top 8 atoms. We use the atoms as templates and try to find them on the raw images. The dictionary atom that looks circular was the most signficant dictionary atom, corresponding to the center of the image. We then find a region shown in Figure 1 that has unique characteristics and is a specific brain region.

2.2 Results

We find that model trimming using the learned features significantly effects classification performance. The effect is model dependent, with CNN (ResNet18) affected more than transformers (DeiT) despite its better performance on the original dataset.

Through our model trimming experiments, we offer some experimental evidence to suggest CNN and transformers have different mechanisms behind their

Raw T1 MRI Training	ROC-AUC	PR-AUC	Average Precision
ResNet18 Raw T1 MRI	84.44%	83.77%	83.82%
ResNet18 1% Significant Pixel Subtraction	71.20%	71.25%	71.34%
ResNet18 L4 Dictionary Mask	57.66%	57.90%	58.10%
ResNet18 L4 Dictionary Mask Pt.2	82.22%	82.35%	82.39%
ResNet18 Raw T1 MRI External Test Set	84.07%	85.08%	85.12%
DeiT Raw T1 MRI	73.47%	70.61%	70.79%
DeiT 1% Significant Pixel Subtraction	71.48%	69.75%	69.86%
DeiT L4 Dictionary Mask	67.36%	66.64%	66.77%
DeiT L4 Dictionary Mask Pt.2	73.13%	71.76%	71.60%
DeiT Raw T1 MRI External Test Set	67.92%	70.99%	71.10%

feature extraction, with transformers showing more reslience to these perturbations. We offer this could explain similar work in adversiaral attacks in which adversarial attacks are model dependent in terms of CNN and transformer networks. [6].

References

1. Aharon, M., Elad, M., Bruckstein, A.: K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. IEEE Transactions on signal processing **54**(11), 4311–4322 (2006)
2. Al-Haija, Q.A., Adebanjo, A.: Breast cancer diagnosis in histopathological images using resnet-50 convolutional neural network. In: 2020 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS). pp. 1–7. IEEE (2020)
3. Dufumier, B., Grigis, A., Victor, J., Ambroise, C., Frouin, V., Duchesnay, E.: Openbhb: a large-scale multi-site brain mri data-set for age prediction and debiasing. NeuroImage **263**, 119637 (2022)
4. Gaonkar, B., Pohl, K., Davatzikos, C.: Pattern based morphometry. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2011: 14th International Conference, Toronto, Canada, September 18-22, 2011, Proceedings, Part II 14. pp. 459–466. Springer (2011)
5. Kwitt, R., Pace, D., Niethammer, M., Aylward, S.: Studying cerebral vasculature using structure proximity and graph kernels. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013: 16th International Conference, Nagoya, Japan, September 22-26, 2013, Proceedings, Part II 16. pp. 534–541. Springer (2013)
6. Mahmood, K., Mahmood, R., Van Dijk, M.: On the robustness of vision transformers to adversarial examples. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7838–7847 (2021)
7. Mendoza, S., Scalzo, F., Chien, A.: Determining and validating population differences in magnetic resonance angiography using sparse representation. In: 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp. 3101–3108. IEEE (2022)
8. Ramezani, M., Marble, K., Trang, H., Johnsrude, I.S., Abolmaesumi, P.: Joint sparse representation of brain activity patterns in multi-task fmri data. IEEE Transactions on Medical Imaging **34**(1), 2–12 (2014)

9. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)
10. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jegou, H.: Training data-efficient image transformers amp; distillation through attention. In: International Conference on Machine Learning. vol. 139, pp. 10347–10357 (July 2021)
11. Wang, F., Casalino, L.P., Khullar, D.: Deep learning in medicine—promise, progress, and challenges. JAMA internal medicine **179**(3), 293–294 (2019)
12. Zhai, Y., Yang, Z., Liao, Z., Wright, J., Ma, Y.: Complete dictionary learning via l 4-norm maximization over the orthogonal group. The Journal of Machine Learning Research **21**(1), 6622–6689 (2020)