Visualizing Salient Network Activations in Convolutional Neural Networks for Medical Image Modality Classification

Sivaramakrishnan Rajaraman and Sameer Antani

Lister Hill National Center for Biomedical Communications,
National Library of Medicine, Bethesda, MD 20894, USA
sivaramakrishnan.rajaraman@nih.gov

Abstract. Convolutional neural network (CNN) has become the architecture of choice for visual recognition tasks. However, these models are perceived as black boxes since there is a lack of understanding of their learned behavior from the underlying task of interest. This lack of transparency is a drawback since poorly understood model behavior could adversely impact subsequent decision-making. Researchers use novel machine learning (ML) tools to classify the medical imaging modalities. However, it is poorly understood how these algorithms discriminate the modalities and if there are implicit opportunities for improving visual information access applications in computational biomedicine. In this study, we visualize the learned weights and salient network activations in a CNN based Deep Learning (DL) model to determine the image characteristics that lend themselves for improved classification with a goal of developing informed clinical question-answering systems. To support our analysis we cross-validate model performance to reduce bias and generalization errors and perform statistical analyses to assess performance differences.

Keywords: Image modality · Classification · Visualization · Saliency · Deep learning · Machine learning

1 Introduction

Medical images serve as a vital source of information for disease screening/diagnosis and an indispensable source of knowledge in clinical decision making/research [1]. The size of medical image repositories has been growing exponentially with the widespread use of digital imaging in clinics and hospitals. These voluminous repositories provide opportunities for researchers to extract meaningful information and develop computerized tools for screening/diagnosis. Medical images have also become an imperative part of the content in several biomedical publications [2–6]. This makes automated medical image classification and retrieval significant in a clinical decision support system, as well as for educational purposes. The images are from diverse medical imaging modalities such as common radiological imagery (e.g., X-rays, Computed Tomography (CT), Magnetic Resonance Imaging (MRI), ultrasound, among others), other medical images that often captured in the visual image spectrum as “photographs” but have distinct clinical classes (e.g., retinal funduscopy, endoscopy,

K. C. Santosh and R. S. Hegadi (Eds.): RTIP2R 2018, CCIS 1036, pp. 42–57, 2019.
https://doi.org/10.1007/978-981-13-9184-2_4

different kinds of microscopic images, surgical photography, photographs of medical devices and systems, among others), and also other graphical drawings and statistical charts of clinical importance found in biomedical journals.

Traditional image classification/retrieval methods tend to be text-based that index images based using descriptive metadata and manual annotations. While these are useful, the textual information may not be consistent with the visual content and modality classification becomes a concern. Also, text-based image classification doesn't consider the representation of image content. Images from different modalities have unique visual patterns, not contained in text labels. Thus, classification tools that arrive at decisions based on visual similarity are a meaningful complement and key for hybrid (text + image) information access systems. Conventionally, rule-based, hand-crafted features are extracted from the images toward visual recognition tasks [7]. However, the process is laborious, computationally expensive, demands iterative labeling and calibration. Hand-engineered features are specific to the modality and are often not transferable to other tasks resulting in suboptimal classification. Under these circumstances, data-driven machine learning (ML) approaches like Deep Learning (DL) becomes a handy tool [8–13].

To overcome challenges of devising hand-engineered features that capture variations in the underlying data, convolutional neural networks (CNN), a class of DL models are used in visual recognition tasks, to discover/learn hierarchical feature representations from raw input pixels [14]. CNNs learn these representations through their inherent mechanisms of receptive fields, shared weights and pooling [15]. These models learn to abstract lower-level features to construct higher-level features and learn complex, non-linear decision-making functions toward performing simultaneous feature extraction and classification. In 2012, the AlexNet model proposed by Krizhevsky [15] won the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [16] and brought the initial breakthrough in visual recognition tasks. This led to the use of several CNNs with varying architecture and depths including VGGNet [17], Inception [18], ResNet [19], Xception [20], and DenseNet [21]. At present, CNNs are delivering promising results in visual recognition tasks at levels exceeding human performance [22]. The encouraging performance of CNNs is accompanied by the availability of a large amount of annotated data. When annotated data are sparse, like in the case of medical images, transfer learning methods are preferred. In this approach, the CNNs are trained on a large selection of stock photographic images like ImageNet that contain more than 1.2 million annotated stock photographic images across 1000 categories [16]. The rich feature representations learned in the form of generic features from these large-scale datasets are transferred to the current task. The pre-trained model weights are fine-tuned in the deeper layers to learn the representations from the new task. The number of layers to fine-tune depends on the availability of annotated data. Literature studies have demonstrated that fine-tuning is a robust and effective strategy to adopt for a variety of classification/recognition tasks [23]. Fine-tuning would be a promising option for medical modality classification, the pre-trained weights could serve as a promising initialization for the new task [24]. The pre-trained CNNs are also used as feature extractors to extract the learned features from the current task. The extracted features are used to train a classifier to make predictions [25].

While state-of-the-art studies elucidate diverse methods for classification/recognition, none of them provide insights into the learned behavior of CNNs or rationalize their performance. It is equitable to declare that the current studies leave room for progress in visualizing and understanding CNN models. The unsettled issue of interpreting the learned behavior of CNNs toward classifying image modalities has become significant and is the focus of this work. In this study, we visualized the learned weights and salient network activations in an optimally trained CNN model applied to the task of medical modality classification. We evaluated the performance of state-of-the-art CNNs including VGG-16, ResNet-50, Xception, Inception-V3, and DenseNet-121, customized for the current task. We cross-validated the performance of the models to reduce bias/generalization errors and statistically validated for the presence/absence of significant differences in their performance. The following paper is organized as follows: Sect. 2 discusses the related work, Sect. 3 particularizes on the materials and methods, Sect. 4 discusses the results, and Sect. 5 concludes the work.

2 Related Work

Modality classification has become an imperative research task in recent years. Evaluation campaigns like Image Cross Language Image Retrieval (ImageCLEF) has been providing collections of annotated medical images for a variety of challenges including modality classification, compound figure separation, image captioning, and visual question and answering [26–30]. However, the collections are sparse in comparison to ImageNet and other large-scale data collections. Conventional methods for modality classification use rule-based, handcrafted feature descriptors toward representing the image characteristics. In [31], the authors used the ImageCLEF2011 modality classification dataset and a combination of SIFT and Bag-of-Colors based feature descriptors toward classifying the modalities. The authors demonstrated a classification accuracy of 72.5% and highlighted the importance of using color descriptors. In another study [32], the authors used the ImageCLEF2015 modality classification dataset and extracted SIFT and Bag-of-Words features to train a multi-class support vector machine (SVM) and obtained a classification accuracy of 60.91%. The authors in [33] used the same dataset, extracted 1st and 2nd order color features, and a manifold-based sparse classification method toward obtaining an accuracy of 73.24%. In another study [34], the authors participated in the ImageCLEF2016 sub-figure classification task, extracted features including contrast, gradient orientations, color, and local pattern distributions and trained a multi-class neural network classifier to obtain a classification accuracy of 72.46%. Literature reveals that the studies used rule-based features that suffer from limitations owing to lack of sufficient human experts to perform manual annotations, inter-/intra-observer variability, inter-class similarity and intra-class variances [35].

At present, DL is delivering promising results as compared to hand-engineered feature extraction in visual recognition tasks. The authors in [36] used multiple, customized CNNs to classify the ImageCLEF2013 medical modality collection with an accuracy of 74.90%. The authors in [37] used an ensemble of fine-tuned CNNs including AlexNet and GoogLeNet and obtained an accuracy of 82.48% in classifying the modality data collection from the ImageCLEF2016 evaluation challenge. The authors in [38] used a combination of pre-trained and customized CNNs toward classifying the sub-figure classification dataset from ImageCLEF2015 and ImageCLEF2016 evaluation challenges and obtained an accuracy of 76.87% and 87.37% respectively. In another study [39], the authors extracted features using a pre-trained ResNet and trained an SVM classifier to obtain an accuracy of 85.38% toward ImageCLEF2016 modality classification task. The authors in [40] participated in the ImageCLEF2016 subfigure classification challenge, employed a synergic signal method to merge the features of multiple fine-tuned ResNet models, and obtained an accuracy of 86.58%. While current literature explains promising methods for modality classification using CNNs, none of them interprets their learned behavior. The indispensable need for visualizing and interpreting the learned behavior of CNNs toward classifying image modalities is the motivation behind this study.

3 Materials and Methods

3.1 Data Collection and Preprocessing

In this study, an 11-class modality classification was performed, the modalities include Computed Tomography (CT), Magnetic Resonance Imaging (MRI), Positron Emission Tomography/CT fusion (PET/CT), ultrasound, 2-D radiography (X-ray), Scanning Electron Microscopy (SEM), Fluorescence Microscopy (FM), Light Microscopy (LM), retinal funduscopy, colonoscopy and statistical graphs. We pooled data from various resources including ImageCLEF2013 modality classification challenge [31], Open Access Biomedical Image Search Engine (OpenI®) and the Internet. For PET/CT, we collected additional images belonging to soft-tissue sarcoma [41], breast [42], head and neck cancer [43], non-small cell lung cancer [44] and prostate cancer [45], from the Cancer Imaging Archive [46]. For colonoscopy, we collected colonoscopy procedure videos from the Web and applied a frame grabbing algorithm to generate images. Additional retinal funduscopy images were pooled from the Messidor [47] dataset that includes 1200 images, acquired using color CCD cameras on a non-mydriatic retinograph. Sample images from different image modalities (Fig. 1). The distribution of data across the modalities is tabulated in Table 1. We evaluated the performance of state-of-the-art CNNs including VGG-16, ResNet-50, Xception, Inception-V3, and DenseNet-121 through five-fold cross-validation toward optimal model selection. As observed in Table 1, the distribution of data was imbalanced across the classes. To mitigate this issue, data was augmented by introducing class-specific perturbations in the training samples [15]. Data augmentation has been shown to improve the robustness, generalization ability of the predictive models, reduce bias and overfitting. The perturbations were reproducible that did not alter the image semantics but helped in generating new samples during model training. The data was augmented with rotations in the range (–3, 3), width/height shifts (–2, 2), horizontal and vertical flips. The categorical cross-entropic loss was modified by observing the distribution of class labels and producing weights to equally penalize the under/over-represented classes to assist the models to learn equally well across the classes [48]. Images were resampled to 300 × 300 pixel resolutions and mean-normalized to assist the models in faster convergence. The models were trained and tested on an NVIDIA DGX-1 system having Tesla V-100 GPUs with computational tools including Python® 3.6.3, Keras® 2.1.2 with Tensorflow® 1.4.0 backend, and CUDA 8.0/cuDNN 5.1 dependencies for GPU acceleration.

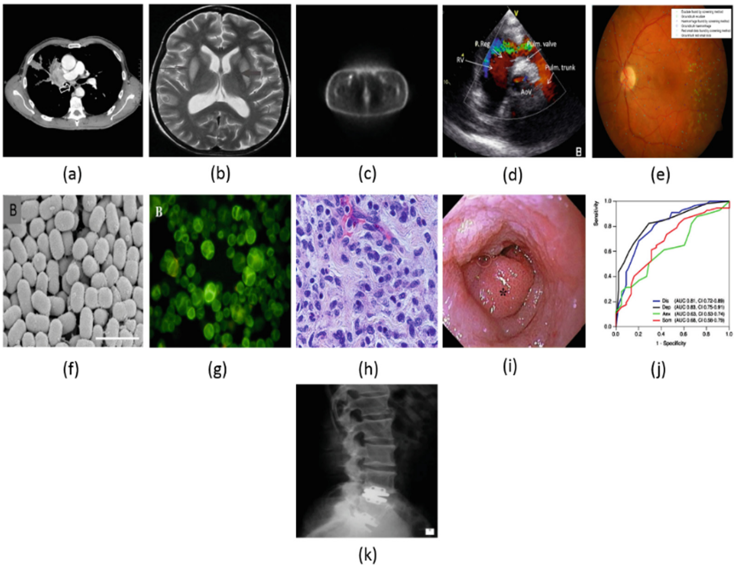

Fig. 1. Sample images from different modalities: (a) CT, (b) MRI, (c) PET/CT, (d) ultrasound, (e) retinal funduscopy, (f) SEM, (g) FM, (h) LM, (i) colonoscopy, (j) statistical graphs, (k) X-ray.

Table 1. Distribution of data across the image modalities.

Modality	#Images
CT	17,055
MRI	12,720
PET/CT	5,510
X-ray	20,030
Ultrasound	6,230
Colonoscopy	14,000
SEM	2,200
FM	5,000
LM	3,900
Retinal funduscopy	2,650
Statistical graphs	2,190

3.2 Model Configuration

The performance of CNNs including VGG-16, ResNet-50, Xception, Inception-V3, and DenseNet-121 was evaluated in this study. The models were initialized with pre-trained ImageNet weights and truncated at their deepest convolutional layer. A convolutional layer with 1024 filters of 3 × 3 dimensions, followed by a global average pooling (GAP) and Softmax layer was added to the truncated models as shown in Fig. 2. The pre-trained weights were fine-tuned using small weight updates to learn the representations of the image modalities, as established in the following steps: (a) instantiating the convolutional base of the pre-trained models and loading their weights; (b) truncating the models at the deepest convolutional layer; (c) adding the convolutional, GAP layer and top-level classifier; and (d) fine-tuning the models alongside the newly added layers to learn the representations for the current task. The models were optimized for hyperparameters by a randomized grid search method [49]. Search ranges including [1e−3 10e−2], [0.8 0.95] and [1e−10 10e−2] were used for the learning rate, stochastic gradient descent (SGD) momentum and L2-regularization respectively. A mini-batch size of 10 was used, 9148 iterations were performed per epoch. The performance of the predictive models were evaluated through five-fold cross-validation with the following performance metrics: (a) accuracy, (b) area under receiver operating characteristic (ROC) curve, (c) sensitivity, (d) specificity, (e) F1-score, and (f) Matthews Correlation Coefficient (MCC) [50].

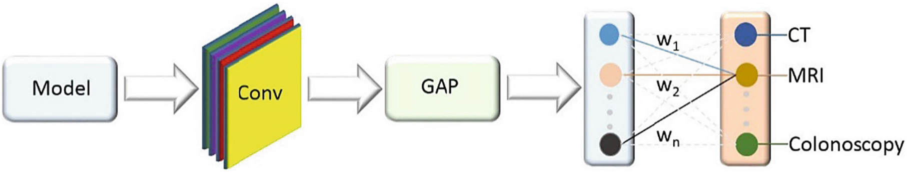

Fig. 2. Model architecture.

3.3 Visualization Studies

DL models are perceived as black boxes since the learned representations are difficult to extract and present in a human-readable form. However, the representation learned by CNNs are highly amenable to visualization because they represent visual concepts. The convolutional layers learn template matching filters whose output gets maximized on observing similar patterns in the input samples [51]. The layers at the beginning of the model are simple to interpret by visualizing the weights as an image. Filters in the deeper layers abstract the outputs from the earlier layers, making interpretation hard. Several methods have been developed for visualizing and interpreting the learned representations of the CNN models including visualizing the learned weights and intermediate layer activations. Visualizing the weights helps in interpreting the visual patterns/concepts learned by the filters in a given layer. The process helps to identify the robustness and generalization ability of the trained model by generating input images that maximize the output of the given layer corresponding to the expected class.

This is accomplished by performing gradient ascent in the input space by applying gradient descent to the value of the input image pixels in order to maximize the response of a specific filter. A loss function is constructed to maximize the value of the filter in a given convolution layer and is minimized during iterations of gradient descent. Visualizing intermediate layer activations helps in understanding successive data transformations and gaining an intuition of the learned patterns. Given an input, activation visualization helps to display the feature maps that are output by the convolutional and pooling layers toward interpreting the relatively independent features learned by the filters. The learned feature maps are visualized by independently plotting the filter contents as a 2D image.

3.4 Statistical Analyses

Statistical analyses help to validate the performance of the predictive models by describing the nature of the data, identifying the trends, and summarizing their relationships. Skewness and kurtosis give a measure of data variability [52]. Skewness is a measure of symmetry. Kurtosis is a measure of whether the samples are heavy-tailed/light-tailed relative to the normal distribution. Skewness and kurtosis measures for a normal distribution should be as close to zero as possible. Statistical tests like one-way analysis of variance (ANOVA) help in identifying the statistically significant differences between the means of two or more unrelated groups [53]. The null hypothesis (H0) infers the samples in the unrelated groups are drawn from populations with similar means. The test produces an F-statistic given by the ratio of the population variance calculated among the sample means to the variance within the samples. A higher value infers that the samples were drawn from populations with varying values for the mean. Under these circumstances, H0 is rejected and the alternate hypothesis (H1) is accepted. One-way ANOVA is a parametric test that requires the underlying data to satisfy assumptions of normality, independence of observations, and homogeneity of variances [54]. A Shapiro-Wilk test [55] is performed to test for data normality and Levene's test [56] to observe the homogeneity of variances. One-way ANOVA is an omnibus test that doesn't reveal where exactly the statistically significant difference exists. A post-hoc analysis like Tukey post-hoc is performed to identify the models that demonstrate statistically significant differences in the mean values for the different performance metrics under study [57].

4 Results and Discussion

4.1 Performance Metrics Evaluation

Training was stopped after 30 epochs ($\approx$274K iterations) when the validation accuracy ceased to improve. From the randomized grid search, the optimal values for the learning rate, SGD momentum, and L2 regularization were found to be 1e−4, 0.9 and 1e−6 respectively. The CNNs converged to an optimal solution due to hyper-parameter optimization and regularization. It was observed from the cross-validation studies that

VGG-16 gave promising results for accuracy, sensitivity, precision, and F1-score compared to other models, as shown in Table 2.

4.2 Statistical Analyses Interpretation

While performing statistical analyses, it was observed that the skewness and kurtosis measures were close to 0 to signify that the values of the performance metrics were a little skewed and kurtotic but did not significantly differ from normality. It was also observed that the results of Levene's test were not statistically significant ($p > 0.05$) to signify that the homogeneity of variances was not violated. Shapiro-Wilk test ($p > 0.05$) and a visual inspection of the histograms and box plots showed that the values for the different performance metrics were approximately normally distributed. Hence, the parametric one-way ANOVA analysis was performed to observe for the significant differences in the performance metrics for the models under study. The values for the statistical measures and the consolidated results of one-way ANOVA are shown in Table 3. It was observed that, in terms of accuracy, no statistically significant difference in performance existed between the different models (F (4, 20) = 0.788, $p = .547$). Similar results were observed for AUC (F (4, 20) = 2.221, $p = .103$), sensitivity (F (4, 20) = 0.814, $p = .531$), precision (F (4, 20) = 0.777, $p = .553$), F1-score (F (4, 20) = 0.805, $p = .537$) and MCC (F (4, 20) = 0.692, $p = .606$). However, in terms of the minimum and maximum values, VGG-16 outperformed the other CNNs except for AUC and MCC where ResNet-50 and DenseNet-121 demonstrated higher values respectively. Considering the balance between sensitivity and precision as demonstrated by F1-score, VGG-16 delivered promising results than the other models under study.

Table 2. Performance metrics computed for the predictive CNN models.

Models	Accuracy	AUC	Sensitivity	Precision	F1-score	MCC
VGG-16	0.990 ± 0.003	0.998 ± 0.001	0.990 ± 0.01	0.990 ± 0.007	0.990 ± 0.008	0.985 ± 0.008
ResNet-50	0.986 ± 0.009	1.0 ± 0.001	0.986 ± 0.009	0.987 ± 0.009	0.986 ± 0.009	0.995 ± 0.009
Xception	0.984 ± 0.01	1.0 ± 0.001	0.984 ± 0.01	0.985 ± 0.01	0.984 ± 0.01	0.988 ± 0.011
Inception-V3	0.988 ± 0.008	0.999 ± 0.002	0.989 ± 0.008	0.988 ± 0.008	0.988 ± 0.008	0.990 ± 0.007
DenseNet-121	0.980 ± 0.013	0.994 ± 0.009	0.980 ± 0.013	0.981 ± 0.012	0.980 ± 0.013	0.986 ± 0.015

4.3 Visualizing Convolutional and Dense Layers

Visualization studies were performed with the optimally trained VGG-16 model for the current task. Inputs that maximized the filter activations in different layers of the model were visualized. It was observed that each layer learned a collection of filters that got increasingly refined and complex in the deeper layers. As observed in Fig. 3, the filters in the first convolutional layer (block1-conv1) encoded simple directional edges and colors/colored edges. The filters in the deeper layers (Fig. 4) abstracted the features from the earlier layers to form complex patterns.

Table 3. Summary of statistical measures.

Performance metrics	Models	Skewness		Kurtosis		Shapiro-Wilk	One-way ANOVA
Accuracy	VGG-16	−0.398	0.913	−1.052	2.0	0.478	F(4,20) = 0.788, *p* = .547
	ResNet-50	−0.782	0.913	0.094	2.0	0.814	
	Xception	−0.734	0.913	−0.378	2.0	0.772	
	Inception-V3	−0.439	0.913	−0.743	2.0	0.735	
	DenseNet-121	−0.778	0.913	−1.271	2.0	0.403	
AUC	VGG-16	−1.540	0.913	2.581	2.0	0.113	F(4,20) = 2.221, *p* = .103
	ResNet-50	−0.512	0.913	−2.963	2.0	0.167	
	Xception	−0.588	0.913	−2.898	2.0	0.111	
	Inception-V3	−0.932	0.913	−1.290	2.0	0.148	
	DenseNet-121	−1.586	0.913	2.520	2.0	0.145	
Sensitivity	VGG-16	−0.398	0.913	−1.052	2.0	0.478	F(4,20) = 0.814, *p* = .531
	ResNet-50	−0.818	0.913	0.268	2.0	0.806	
	Xception	−0.725	0.913	−0.382	2.0	0.782	
	Inception-V3	−0.313	0.913	−1.077	2.0	0.747	
	DenseNet-121	−0.775	0.913	−1.264	2.0	0.413	
Precision	VGG-16	−0.418	0.913	−1.077	2.0	0.522	F(4,20) = 0.777, *p* = .553
	ResNet-50	−0.698	0.913	−0.204	2.0	0.844	
	Xception	−0.654	0.913	−0.562	2.0	0.815	
	Inception-V3	−0.506	0.913	−0.559	2.0	0.730	
	DenseNet-121	−0.673	0.913	−1.741	2.0	0.397	
F1-Score	VGG-16	−0.398	0.913	−1.066	2.0	0.483	F(4,20) = 0.805, *p* = .537
	ResNet-50	−0.762	0.913	0.064	2.0	0.829	
	Xception	−0.689	0.913	−0.468	2.0	0.799	
	Inception-V3	−0.410	0.913	−0.845	2.0	0.723	
	DenseNet-121	−0.765	0.913	−1.357	2.0	0.401	
MCC	VGG-16	−1.538	0.913	3.068	2.0	0.209	F(4,20) = 0.692, *p* = .606
	ResNet-50	1.214	0.913	2.302	2.0	0.457	
	Xception	−1.285	0.913	1.747	2.0	0.450	
	Inception-V3	0.440	0.913	1.422	2.0	0.747	
	DenseNet-121	−1.697	0.913	3.152	2.0	0.154	

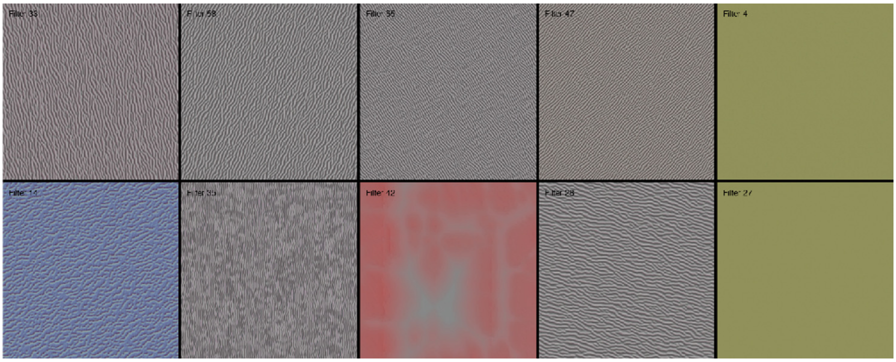

Fig. 3. Visualizing random filters in the first convolutional layer in the first convolutional block.

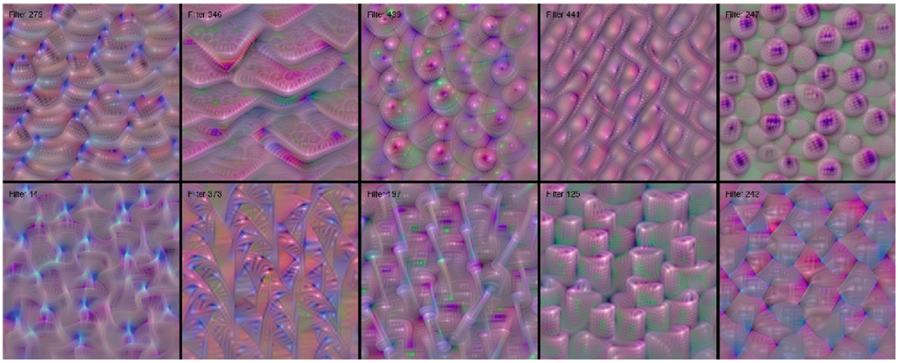

Fig. 4. Visualizing random filters in the third convolutional layer in the fifth convolutional block.

The filters functioned as a basis of vectors to compactly encode the inputs that became more intricate as they begun to incorporate information from an increasingly larger spatial extent, converging to specific patterns in the image modalities under study. The final dense layer was visualized as shown in Fig. 5. Filters 0 to 10 represent CT, SEM, colonoscopy, FM, retinal funduscopy, statistical graphs, LM, MRI, PET/CT, ultrasound, and X-ray respectively. It was observed that for CT, the filters maximally responded to the CT-like contour patterns in the input. For SEM and LM, higher activations were observed for blob-like cell patterns. For colonoscopy, the input patterns that mimicked tissue color and texture maximally activated the filters. For FM, the patterns exhibiting green fluorescence captured by the cells produced higher activations. Axes like patterns maximally activated the filters for the expected statistical graphs class. For MRI, the input patterns simulating the neo-cortical convolutions led to high values of activations. For PET/CT fusion, the input patterns mimicked scan contours. For ultrasound, the patterns mimicked the tissue texture and shape of image formation. For the X-ray class, the filters were maximally activated for rib and bone-like patterns in the input image.

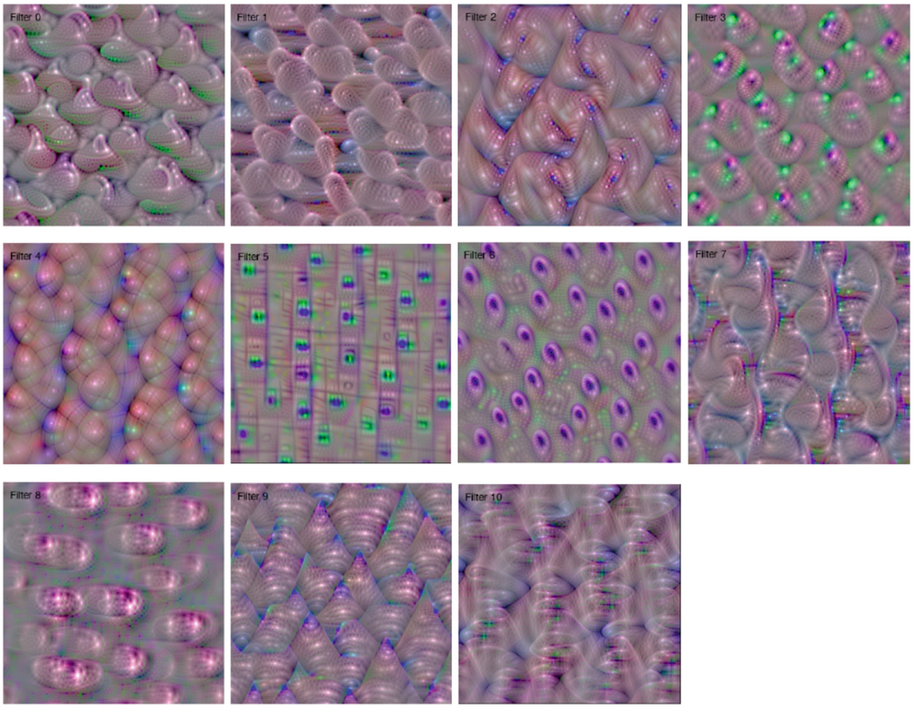

Fig. 5. Visualizing the final dense layer. (Color figure online)

The trained model's notion of the expected classes were found to be at the level of local textures to demonstrate that they do not understand the concept of the classes analogous to humans. The model learned the decomposition of the visual space as a hierarchical-modular network of filters, a probabilistic mapping between filter combinations and a set of class labels. However, the nature of human vision is not purely convolutional. The human visual cortex is complex, active, structured into cortical columns with multifaceted functionality, and involves motor control [58].

4.4 Visualizing Intermediate Activations

The activations of the intermediate layers of the trained VGG-16 model was visualized to gain further insight into its performance and learned behavior. Given an input image of CT and MRI for instance, random filters in the feature maps were extracted and plotted to visualize the activations as shown in Figs. 6 and 7. It was observed that the filters in the earlier layers acted as a collection of various edge detectors. The activations retained almost all of the information present in the original image. At the deeper layers, the activations became increasingly abstract and less visually interpretable. Higher representations carried increasingly less information about the visual contents, and increasingly more information related to the expected class.

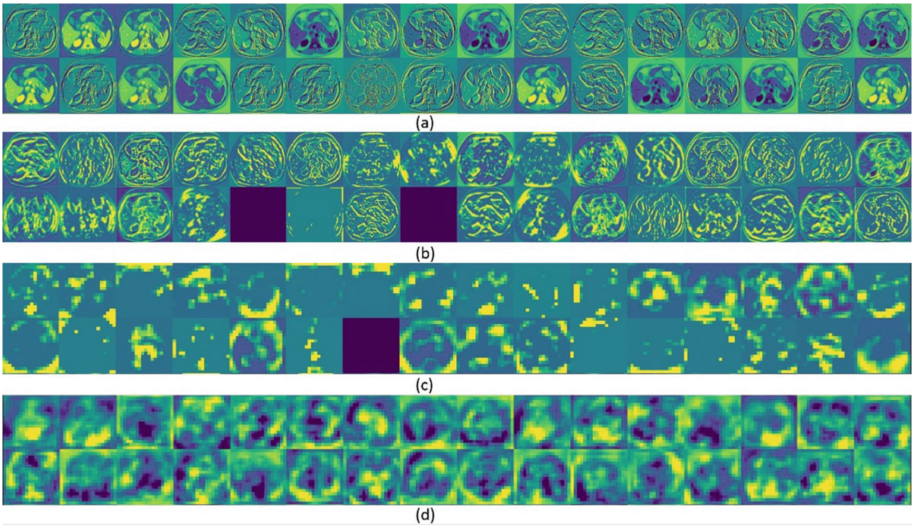

Fig. 6. Visualizing intermediate activations for an input CT image: (a) block1-conv1, (b) block3-conv1, (c) block5-conv3, and (d) deepest convolutional layer.

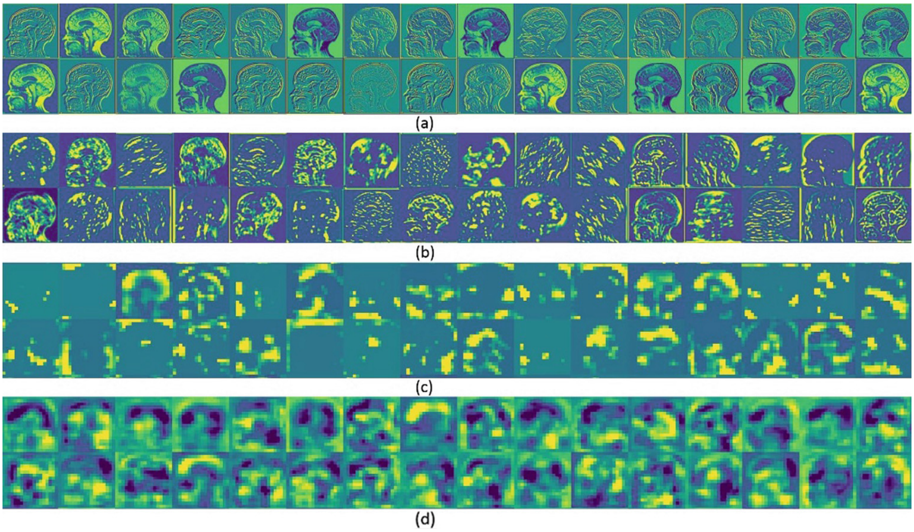

Fig. 7. Visualizing intermediate activations for an input MRI image: (a) block1-conv1, (b) block3-conv1, (c) block5-conv3, and (d) deepest convolutional layer.

The sparsity of activations increased with the depth of the layer: in the first convolutional layer of the first convolutional block, almost all channels were activated by the input image; in the deeper layers, many of the channels were blank that inferred that the pattern encoded by these filters was not found in the input image. This led to the interpretation of the important, universal characteristic of the representations learned by CNNs; the features became increasingly abstract with the depth of the layer. The activations of deeper layers carried subtle information about the specific input being seen, and more information about the imaging modality. The CNN model acted as an information distillation pipeline that operated on the raw input image pixels to perform repeated transformations so that irrelevant information was filtered and useful information pertaining to the modality of the image was magnified and retained.

5 Conclusion

In this study, we visualized the learned weights and salient network activations in a trained CNN model toward understanding its learned behavior, applied to the challenge of medical modality classification. We interpreted how the model distilled information through transformations and retained characteristic features pertaining to the imaging modality toward decision-making. We also statistically validated the performance of the predictive CNNs for optimal model selection and deployment. We are currently performing pilot studies in deploying the trained model into mobile devices and analyzing performance. We currently worked with images containing a single modality. Often, figures contain multiple panels describing different modalities. In the future, we plan to work with multi-panel images, perform panel segmentation so that each subpanel contains a single modality and improve the classification performance.

References

1. Ben Abacha, A., Gayen, S., Lau, J.J., Rajaraman, S., Demner-Fushman, D.: NLM at ImageCLEF 2018 visual question answering in the medical domain. In: CEUR Workshop Proceedings, p. 2125 (2018)
2. Demner-Fushman, D., Antani, S., Thoma, G.R., Simpson, M.: Design and development of a multimodal biomedical information retrieval system. J. Comput. Sci. Eng. **6**, 168–177 (2012)
3. Rajaraman, S., Candemir, S., Kim, I., Thoma, G.R., Antani, S.: Visualization and interpretation of convolutional neural network predictions in detecting pneumonia in pediatric chest radiographs. MDPI Appl. Sci. **8**(10), 1715 (2018)
4. Rajaraman, S., et al.: Understanding the learned behavior of customized convolutional neural networks toward malaria parasite detection in thin blood smear images. J. Med. Imag. **5**(3), 034501 (2018)
5. Rajaraman, S., et al.: A novel stacked generalization of models for improved TB detection in chest radiographs. In: Proceedings of the International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 718–721 (2018)
6. Thamizhvani, T.R., Lakshmanan, S., Rajaraman, S.: Mobile application-based computer-aided diagnosis of skin tumours from dermal images. Imaging Sci J. **66**(6), 382–391 (2018)

7. Khan, S., Yong, S.P.: A comparison of deep learning and hand crafted features in medical image modality classification. In: Proceedings of the International Conference on Computer and Information Sciences, pp. 633–638 (2016)
8. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**, 436–444 (2015)
9. Rajaraman, S., et al.: Pre-trained convolutional neural networks as feature extractors toward improved malaria parasite detection in thin blood smear images. PeerJ **6**, e4568 (2018)
10. Rajaraman, S., et al.: Comparing deep learning models for population screening using chest radiography. In: Proceedings of the SPIE Medical Imaging: Computer-aided Diagnosis, p. 105751E (2018)
11. Rajaraman, S., Antani, S., Xue, Z., Candemir, S., Jaeger, S., Thoma, G.R.: Visualizing abnormalities in chest radiographs through salient network activations in deep learning. In: Proceedings of the IEEE Life Sciences Conference, pp. 71–74 (2017)
12. Rajaraman, S., Antani, S., Jaeger, S.: Visualizing deep learning activations for improved malaria cell classification. Proc. Mach. Learn. Res. **69**, 40–47 (2017)
13. Xue, Z., Rajaraman, S., Long, L.R., Antani, S., Thoma, G.R.: Gender detection from spine x-ray images using deep learning. In: Proceedings of the IEEE International Symposium on Computer-based Medical Systems, pp. 54–58 (2018)
14. Simard, P., Steinkraus, D., Platt, J.C.: Best practices for convolutional neural networks applied to visual document analysis. In: Proceedings of the 7th International Conference on Document Analysis and Recognition, pp. 958–963 (2003)
15. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Proceedings of the Advances in Neural Information Processing Systems, pp. 1–9 (2012)
16. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Li, F.F.: ImageNet: a large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255 (2009)
17. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2015)
18. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818–2826 (2016)
19. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
20. Chollet, F.: Xception: Deep Learning with Separable Convolutions. arXiv preprint arXiv:1610.02357 (2016)
21. Huang, G., Liu, Z., Weinberger, K.Q., van der Maaten, L.: Densely Connected Convolutional Networks. arXiv preprint arXiv:1608.06993 (2017)
22. Mnih, V., et al.: Human-level control through deep reinforcement learning. Nature **518**, 529–533 (2015)
23. Margeta, J., Criminisi, A., Lozoya, R.C., Lee, D.C., Ayache, N.: Fine-tuned convolutional neural nets for cardiac MRI acquisition plane recognition. Comput. Methods Biomech. Biomed. Eng. Imaging Vis. **5**, 339–349 (2017)
24. Lynch, S., Ng, A.: Why AI is the new electricity. https://news.stanford.edu/thedish/2017/03/14/andrew-ng-why-ai-is-the-new-electricity/
25. Razavian, A.S., Azizpour, H., Sullivan, J., Carlsson, S.: CNN features off-the-shelf: an astounding baseline for recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 512–519 (2014)
26. De Herrera, A., Schaer, R., Bromuri, S., Müller, H.: Overview of the ImageCLEF 2016 medical task. In: CEUR Workshop Proceedings, p. 1609 (2016)

27. Apostolova, E., You, D., Xue, Z., Antani, S., Demner-Fushman, D., Thoma, G.R.: Image retrieval from scientific publications: Text and image content processing to separate multipanel figures. J. Am. Soc. Inf. Sci. Tec. **64**, 893–908 (2013)
28. Santosh, K.C., Aafaque, A., Antani, S., Thoma, G.R.: Line segment-based stitched multipanel figure separation for effective biomedical CBIR. Int. J. Pattern Recogn. Artif. Intell. **31**(6), 1757003 (2017)
29. Santosh, K.C., Xue, Z., Antani, S., Thoma, G.R.: NLM at ImageCLEF 2015: biomedical multipanel figure separation. In: CEUR Workshop Proceedings, p. 1391 (2015)
30. Santosh, K.C., Antani, S., Thoma, G.R.: Stitched multipanel biomedical figure separation. In: IEEE International Symposium on Computer-based Medical Systems, pp. 54–59 (2009)
31. De Herrera, A., Markonis, D., Müller, H.: Bag–of–colors for biomedical document image classification. In: Greenspan, H., Müller, H., Syeda-Mahmood, T. (eds.) MCBR-CDS 2012. LNCS, vol. 7723, pp. 110–121. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-36678-9_11
32. Pelka, O., Friedrich, C.M.: FHDO biomedical computer science group at medical classification task of ImageCLEF 2015. In: CEUR Workshop Proceedings, p. 1391 (2015)
33. Cirujeda, P., Binefa, X.: Medical image classification via 2D color feature based covariance descriptors. In: CEUR Workshop Proceedings, p. 1391 (2015)
34. Li, P., et al.: UDEL CIS at ImageCLEF medical task 2016. In: CEUR Workshop Proceedings, p. 1609 (2016)
35. De Herrera, A., Kalpathy-Cramer, J., Fushman, D.D., Antani, S., Müller, H.: Overview of the imageCLEF 2013 medical tasks. In: CEUR Workshop Proceedings, p. 1179 (2013)
36. Yu, Y., et al.: Modality classification for medical images using multiple deep convolutional neural networks. J. Comput. Inf. Syst. **11**(15), 5403–5413 (2015)
37. Kumar, A., Kim, J., Lyndon, D., Fulham, M., Feng, D.: An ensemble of fine-tuned convolutional neural networks for medical image classification. IEEE J. Biomed. Heal. Inf. **21**, 31–40 (2017)
38. Yu, Y., Lin, H., Meng, J., Wei, X., Guo, H., Zhao, Z.: Deep transfer learning for modality classification of medical images. MDPI Inf. **8**(3), 91 (2017)
39. Koitka, S., Friedrich, C.M.: Traditional feature engineering and deep learning approaches at medical classification task of ImageCLEF 2016. In: CEUR Workshop Proceedings, p. 1609 (2016)
40. Zhang, J., Xia, Y., Wu, Q., Xie, Y.: Classification of Medical Images and Illustrations in the Biomedical Literature Using Synergic Deep Learning. arXiv preprint arXiv:1706.09092 (2017)
41. Vallières, M., Freeman, C.R., Skamene, S.R., El Naqa, I.: A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities. Phys. Med. Biol. **60**, 5471–5496 (2015)
42. Bloch, B., Jain, A., Jaffe, C.: Data From BREAST-DIAGNOSIS. https://wiki.cancerimagingarchive.net/display/Public/BREAST-DIAGNOSIS#9e4592af79b249bfaff992eceebbf842
43. Vallières, M., et al.: Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer. Sci. Rep. **7**(1), 10117 (2017)
44. Gevaert, O., et al.: Non-small cell lung cancer: identifying prognostic imaging biomarkers by leveraging public gene expression microarray data-methods and preliminary results. Radiology **264**, 387–396 (2012)
45. Kurdziel, K.A., et al.: The kinetics and reproducibility of 18F-sodium fluoride for oncology using current pet camera technology. J. Nucl. Med. **53**, 1175–1184 (2012)
46. Clark, K., et al.: The cancer imaging archive (TCIA): maintaining and operating a public information repository. J. Digit. Imaging **26**, 1045–1057 (2013)

47. Decencière, E., et al.: Feedback on a publicly distributed image database: the messidor database. Image Anal. Stereol. **33**, 231–234 (2014)
48. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.: ChestX-ray8: hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–19 (2017)
49. Bergstra, J., Bengio, Y.: Random search for hyper-parameter optimization. J. Mach. Learn. Res. **13**, 281–305 (2012)
50. Matthews, B.W.: Comparison of the predicted and observed secondary structure of T4 phage lysozyme. BBA - Protein Struct. **405**, 442–451 (1975)
51. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) Computer Vision – ECCV 2014, LNCS, vol. 8689, pp. 818–833. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10590-1_53
52. Groeneveld, R.A., Meeden, G.: Measuring skewness and kurtosis. Statistician **33**, 391–399 (1984)
53. Rossi, J.S.: One-way ANOVA from summary statistics. Educ. Psychol. Meas. **47**, 37–38 (1987)
54. Daya, S.: One-way analysis of variance. Evid. Based Obstet. Gynecol. **5**, 153–155 (2003)
55. Shapiro, S.S., Wilk, M.B.: An analysis of variance test for normality (complete samples). Biometrika **52**, 591 (1965)
56. Gastwirth, J.L., Gel, Y.R., Miao, W.: The Impact of levene's test of equality of variances on statistical theory and practice. Stat. Sci. **24**, 343–360 (2009)
57. Kucuk, U., Eyuboglu, M., Kucuk, H.O., Degirmencioglu, G.: Importance of using proper post hoc test with ANOVA. Int. J. Cardiol. **209**, 346 (2016)
58. Bressler, S.L.: Large-scale cortical networks and cognition. Brain Res. Rev. **20**(3), 288–304 (1995)