Automated Image Quality Evaluation of Structural Brain Magnetic Resonance Images using Deep Convolutional Neural Networks

Sheeba Sujit, Refaat Gabr, Ivan Coronado, Melvin Robinson, Sushmita Datta and Ponnada Narayana

November 21, 2018

Automated Image Quality Evaluation of Structural Brain Magnetic Resonance Images using Deep Convolutional Neural Networks

Sheeba J. Sujit [1], Refaat E. Gabr [1], Ivan Coronado [1], Melvin Robinson [2], Sushmita Datta [1], Ponnada A. Narayana [1]

[1] Department of Diagnostic and Interventional Imaging, The University of Texas Health Science Center at Houston (UTHealth)
[2] Department of Electrical Engineering, The University of Texas at Tyler

Abstract - **Automated evaluation of image quality is essential to assure accurate diagnosis and effective patient management. This is particularly important for multi-center studies, typically employed in clinical trials, in which the data are acquired on different machines with different protocols. Visual quality assessment of magnetic resonance imaging (MRI) data is subjective and impractical for large datasets. Data-intensive deep learning methods such as convolutional neural networks (CNNs) are promising tools for processing large-scale imaging datasets for automated quality assessment. In this study, we evaluate a CNN-based method for quality assessment of the Autism Brain Imaging Data Exchange (ABIDE) structural brain MRI dataset acquired from 17 sites on more than a thousand subjects. The CNN architecture consisted of an input layer, four convolution layers, two fully connected layers, and an output layer. A balanced set of 348 image volumes was used in the study. 60% of the data was used for training, 15% for validation, and 25% for testing. The results of the automated approach were compared with the evaluation by the radiologist. Performance of the CNN was assessed using the confusion matrix. The concordance in image quality labels between the expert and CNN was 86% (sensitivity = 81%, specificity = 92%). The present study shows that the proposed model can evaluate the image quality of brain MRI with higher classification accuracy compared to previous state-of-the-art classical machine learning algorithms.**

I. Introduction

Magnetic resonance imaging (MRI) is the most common imaging modality for evaluating neurological disorders. Unfortunately, MRI is prone to image artifacts arising from both intrinsic and extrinsic factors. Evaluation of the image quality is critical for accurate diagnosis. Visual inspection of the images is the most common way for image quality evaluation. However, this is not a viable option for evaluating large amounts of data that are typically collected in multicenter studies. There is thus a critical need for tools for automated evaluation of MRI images.

Recent studies have applied machine learning for automating image quality evaluation in large-scale and multi-center studies [1], [2]. In these studies supervised learning methods such as support vector machines (SVMs), random forests, and Gaussian naïve Bayes classifiers were applied [1], [2]. These supervised learning methods use hand-crafted metrics which probe different aspects of the image quality such as noise, ghosting, and nonuniformity. These metrics are used as features to train a classifier. For example, MRIQC extracts a vector of 64 image quality metrics per input image [1] and the UK Biobank computes 190 features to assess T1-weighted (structural) images [2]. Pizarro et al., [3] used three volumetric features and three artifact-specific features to train a SVM classifier. The possible challenge in these methods is the development of good quality metrics to drive the automated classification since the accuracy of classification depends on the features used [2], [3].

Recently, deep learning (DL) algorithms have rapidly become popular for analyzing medical images without a need for manually selecting the features [4]. Convolutional neural networks (CNNs) are one of the most popular algorithms for image analysis due to their self-learning ability [5], [6]. They achieve generalizability by training on large amounts of data [7]. These algorithms have been successfully used for image classification, localization, object detection, segmentation, registration, and other related tasks [8].

Recent studies have demonstrated the feasibility of automated assessment of medical image quality using CNNs [9][10]. Kustner et al., [9] proposed extracting patches from T1-weighted MR images of 16 healthy volunteers from a single site to train the CNN. Esses et al., [10] used 522 T2 –weighted Liver MRI images acquired from a single site. It is unclear that these single site results can be generalized to data acquired at different centers. The above considerations highlight the need for a model that can be generalizable to data that are typically acquired in multi-center clinical trials.

In this work, we propose a data-driven image quality analysis using CNN that is capable of self-learning image quality features from the input data, and can provide better generalization to multi-site data. Brain MRIs from the open-access Autism Brain Imaging Data Exchange (ABIDE) are used to train and evaluate the performance of the proposed technique. In section II of this paper, we describe the datasets,

This work is supported by the National Institutes of Health, grant 1R56NS105857-01 and Endowed Chair in Biomedical Engineering.

the network architecture and evaluation metrics in detail. The results are presented in section III and discussed in section IV.

II. METHODS

A. Image Database

The present study was carried out on structural brain images from the ABIDE database. ABIDE is a consortium that provides previously collected images for the purpose of data sharing within the scientific community [11]. The ABIDE dataset is publicly available [12] and contains images acquired at 17 sites with diverse acquisition settings and parameters. Many forms of image degradation in the ABIDE database are participant-specific or arise from practical settings (examples shown in Fig. 1). This heterogeneity makes it a valuable resource to train machine learning models that can be generalized to MRI data from other sites. The dataset includes structural brain MRI images, resting state functional MRI data, and phenotypic information for each patient. The phenotypic information includes manual image quality assessment by multiple experts [13]. In this study our focus is on structural MRI. We considered evaluation by one expert based on the general image quality as the ground truth.

B. Data preprocessing

The dataset contains a diverse set of images with different matrix sizes and image resolution acquired with variable scanner settings and parameters. Before using these images as input to the CNN, they were re-sampled to isotropic resolution of 1 x 1 x 1 mm^3 and matrix size 256 x 256 x 256. The middle 200 slices covering the brain were extracted. The image intensity was normalized between 0 and 1.

C. Deep Learning Architecture

The architecture of the multilayer CNN used in this study is shown in Fig. 2. It consisted of an input layer, four convolution layers, two fully connected (dense) layers and an output layer. A sigmoid function in the final layer provided the probability for the quality class. Each convolution layer consisted of N filter kernels of size M x L (Fig. 2), an activation unit, and a maximum pooling unit. The convolution operation produced feature maps by convolving a kernel

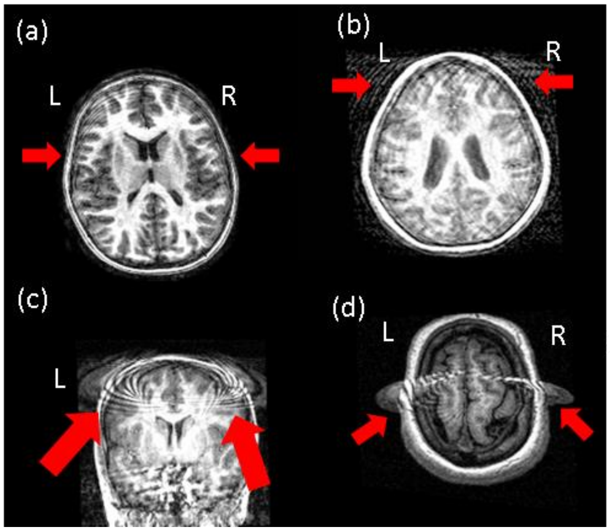

Figure 1. Example MRI scans from the ABIDE dataset with prominent artifacts: (a,b) severe motion artifacts due to head motion (c,d) severe coil artifacts.

across the input image or output from the previous convolution layers. The rectified linear unit (ReLu) defined as f(x) = max(0,x) was used as the activation function. Pooling layers were used to down-sample the output of preceding convolution operation [7]. The network was trained by minimizing the binary cross-entropy loss function with a learning rate of 10^{-4}. Optimization used the Adam method, performing stochastic optimization with adaptive gradient moment estimation [14], with associated parameters of β_1 = 0.9, β_2 = 0.999 and $\epsilon = 10^{-8}$. The analysis was performed on a balanced set of 348 images. Training used 2D slices as the input to the CNN. In total, our dataset contained 69,600 2D images out of which 34,800 were labeled '*exclude*' and 34,800 were labelled '*include*'. These were split into 41,760 slices (60%) for training, 10,440 slices (15%) for validation, and 17,400 slices (25%) for testing. Dropout was used to regularize the network weight updates to avoid overfitting. The maximum number of epochs was 300. A grid search was performed to optimize the learning rate and drop-out factors in the fully connected layers. All processing was done on the

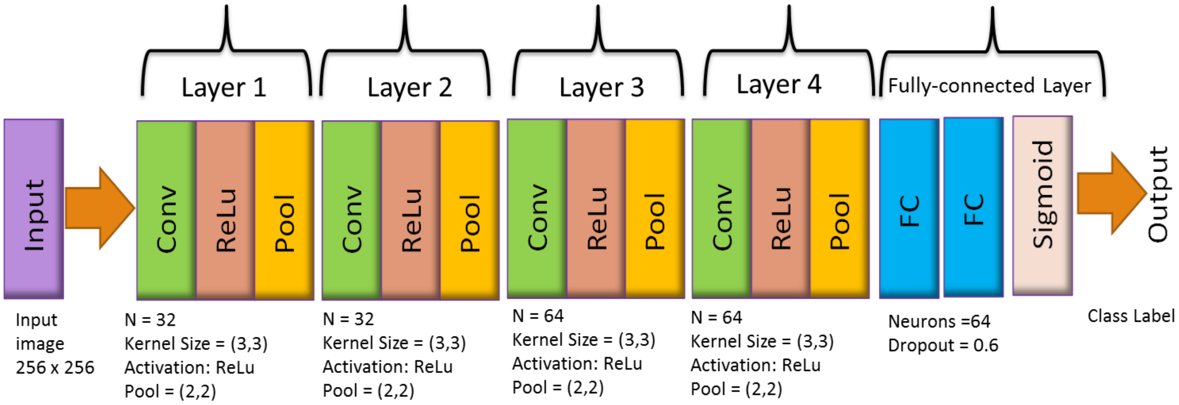

Figure 2. Architecture of the proposed CNN for evaluating image quality. Conv: convolution, FC: fully-connected, ReLU: rectified linear unit, Pool: maximum pooling.

Maverick 2 cluster at the Texas Advanced Computing Center (TACC) at Austin, Texas. NVIDIA Tesla GTX graphics processing unit (GPU) cards were used, and implementation was carried out in Python using the Keras library [15] and TensorFlow [16].

D. Classification Metrics

The performance of the CNN was evaluated using the confusion matrix which summarizes the predicted vs. expected results [17]. The sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), accuracy (ACC), and the F1-score (the harmonic mean of the PPV and sensitivity) were calculated as follow:

$$\text{Sensitivity} \; = \frac{TP}{(TP+FN)} \tag{1}$$

$$\text{Specificity} = \frac{TN}{(TN+FP)} \tag{2}$$

$$\text{PPV} = \frac{TP}{(TP+FP)} \tag{3}$$

$$\text{NPV} = \frac{TN}{(TN+FN)} \tag{4}$$

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{5}$$

$$\text{F1- score} = \frac{2.TP}{2.TP+FP+FN} \tag{6}$$

where *TP* = True Positive (both rater and CNN labels the image as “exclude”), *TN* = True Negative (both rater and CNN labels the image as “include”), *FP* = False Positive (rater labels the image as “include” but CNN labels as “exclude”) and *FN* = False Negative (rater labels the image as “exclude” but CNN labels as “include”).

III. Results

Table I shows the concordance between rater and CNN algorithm in identifying the image quality of the 2D slices from T1-weighted brain MR volume. The rater scored 42% (7306/17400) of cases as *‘include’* and 58% (10,094/17400) as ‘*exclude*’. CNN scored 50% (8648/17400) as ‘*include*’ and 50% (8752/17400) as ‘*exclude*’. There was agreement between the rater and CNN in 86% of the cases. The sensitivity, specificity, PPV and NPV of the CNN were 81%, 92%, 93% and 78% respectively. The F1-score of ‘*exclude*’ was 87% and ‘*include*’ was 84%. High F1-score for both classes shows good classification performance.

Table I. Concordance between CNN and rater in evaluating image quality.

CNN	Rater		
	(*Exclude*)	**(*Include*)**	Total
(*Exclude*)	8170	582	8752
(*Include*)	1924	6724	8648
Total	10,094	7,306	

IV. Discussion

In this work, we evaluated a DL model for reference-less image quality assessment of structural brain MRI data. By integrating multiple convolution layers, the CNN learned to produce feature maps with relevant information about image quality. The fully-connected layers combine these features to classify the input image. The classification performance of the proposed model cannot be easily compared to other methods when the datasets used for training and testing are different. However, we compared our results to the maximum classification accuracy achieved by other state-of-the-art methods listed in literature. The supervised classification framework proposed by Esteben et al., [1] used random forests classifiers on the ABIDE databased on 64 features extracted from each input image. While the accuracy was satisfactory on a held-out dataset (F1-score = 72%), the sensitivity (28%) was low. FidelAlfaro-Almagro et al., [2] extracted 190 features, and combined the output of three classifiers: Bayes network classifier, naïve Bayes classifier and MetaCost classifier. They achieved an accuracy of 84%. Pizarro et al., [3] defined 3 volumetric features and 3 artifact-specific features to train a SVM classifier, achieving an accuracy of 80%. Using CNNs, Kustner et al. [9] and Esses et al., [10] achieved accuracy of 97% and 79%, respectively, but data were limited to single site. The proposed CNN model achieved an accuracy of 86% on a multi-center image database, with images acquired from different scanners and with different scan parameters.

One advantage of deep learning methods is that, once a model is trained, classification of new image is very efficient, and typically takes less than one second. This feature makes it suitable for real-time decisions about image quality to determine if there is a need for re-acquisition before the patient leaves the MRI scanner.

Our model achieved high accuracy in predicating overall image quality. However, the model does not provide information on the type of the artifact (motion, flow, wrap-around, etc.), nor does it provide any spatial localization of the artifact. These are active areas of research, and other models are being developed to address these challenges.

Further development to improve the performance of the CNN will investigate deeper networks and pre-trained models. Building models to detect and classify various classes of image degradation and different modalities (e.g. T2-weighted, FLAIR, etc.) will also be pursued.

V. References

[1] O. Esteban et al., “MRIQC: Advancing the automatic prediction of image quality in MRI from unseen sites,” *PLoS One*, vol. 12, no. 9, p. e0184661, 2017.

[2] F. Alfaro-Almagro et al., “Image processing and Quality Control for the first 10,000 brain imaging datasets from UK Biobank,” *Neuroimage*, vol. 166, pp. 400–424, 2018.

[3] R. A. Pizarro et al., “Automated Quality Assessment of Structural Magnetic Resonance Brain Images Based on a Supervised Machine Learning Algorithm,” *Front. Neuroinform.*, vol. 10, no. 52, 2016.

[4] G. Litjens et al., “A survey on deep learning in medical image

analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, 2017.

[5] I. Goodfellow et al., *Deep learning*, vol. 1. MIT press Cambridge, 2016.

[6] Y. LeCun et al., "Deep learning," *Nature*, vol. 521, no. 7553, p. 436, 2015.

[7] Z. Akkus et al., "Deep Learning for Brain MRI Segmentation: State of the Art and Future Directions," *J. Digit. Imaging*, vol. 30, no. 4, pp. 449–459, 2017.

[8] J. Ker et al., "Deep learning applications in medical image analysis," *IEEE Access*, vol. 6, pp. 9375–9389, 2018.

[9] T. Küstner et al., "Automated reference-free detection of motion artifacts in magnetic resonance images," *Magnetic Resonance Materials in Physics, Biology and Medicine*, pp. 1–14, 2017.

[10] S. J. Esses et al., "Automated image quality evaluation of T2-weighted liver MRI utilizing deep learning architecture," *J. Magn. Reson. Imaging*, 2018.

[11] A. Di Martino et al., "The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism," *Mol. Psychiatry*, vol. 19, no. 6, p. 659, 2014.

[12] "http://fcon_1000.projects.nitrc.org/indi/abide/." .

[13] C. Craddock et al., "The Neuro Bureau Preprocessing Initiative: open sharing of preprocessed neuroimaging data and derivatives <br />," *Front. Neuroinform.*, no. 41.

[14] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, 2015.

[15] F. Chollet, "Keras," 2015. [Online]. Available: https://github.com/keras-team/keras.

[16] M. Abadi et al., "TensorFlow: A System for Large-Scale Machine Learning TensorFlow: A system for large-scale machine learning," in *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI '16)*, 2016.

[17] N. M. Balasooriya and R. D. Nawarathna, "A sophisticated convolutional neural network model for brain tumor classification," in *Industrial and Information Systems (ICIIS), 2017 IEEE International Conference on*, 2017, pp. 1–5.