Proceedings of the 2023 International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/41/20230729

Investigation related to detection of Intracranial Hemorrhage based on edge impulse enhanced CT scanning

Chengmin Li

Department of Computer Science, Beijing Jiaotong University, Beijing, 100044, China

20722057@bjtu.edu.cn

Abstract. Intracranial Hemorrhage (ICH) is a critical medical condition demanding rapid and precise diagnosis, typically achieved through Computerized Tomography (CT) scans. This research investigates the potential of the Edge Impulse platform, a symbol of progress in edge computing, for the automatic detection of Intracranial Hemorrhage (ICH). The study leverages RGB images extracted from CT scans, employing transfer learning techniques. By utilizing the "brain ct hemorrhage AMINE dataset" available on Kaggle, this research combines Convolutional Neural Networks (CNNs) with the efficiency and adaptability offered by the MobileNet framework in a novel approach to address this diagnostic challenge. To ensure the models strength, robustness, applicability and a useful approach has been used, this study tested setups of the neural network to find the most effective ones. These setups involved changing parameters like resolution (ρ) and width multipliers (α) which greatly impact the model's diagnostic performance. The remarkable result was observed in a configuration, with a resolution of 160x160 pixels and a width multiplier of 0.5. After optimization this specific setup achieved an outstanding diagnostic accuracy rate of 99.8% with negligible loss. This accomplishment highlights how edge computing, through Edge Impulse can significantly improve and speed up ICH diagnostic procedures.

Keywords: Intracranial Hemorrhage, Edge Impulse, Diagnostic Automation, Machine Learning.

1. Introduction

It is widely acknowledged that the Intracranial Hemorrhage, or ICH, is a formidable health anomaly that can lead to severe neurological deficits if not promptly and accurately diagnosed. Among the widely employed diagnostic modalities, Computerized Tomography (CT) scans emerge as a cornerstone for the visualization of brain structures, offering crucial insights into potential pathological disturbances [1]. CT scans are indispensable for delineating the subtle nuances in the cerebral anatomy and detecting irregularities like ICH. This precision becomes particularly salient, considering the myriad subtypes of ICH, encompassing subarachnoid, intraventricular, subdural, epidural, and Intraparenchymal Hemorrhages, each with its distinct therapeutic implications [1].

The medical repercussions of ICH are severe. Ranging from minor symptoms to dire consequences such as permanent neurological damage, coma, or even death, the stakes in ICH diagnosis and management are exceptionally high. The heterogeneity in ICH manifestations necessitates impeccable accuracy in its detection, making the role of neuroradiologists pivotal [2]. These professionals, traditionally, have anchored their diagnostic conclusions on a meticulous analysis of non-contrast CT

DOI: 10.54254/2755-2721/41/20230729

scans, deciphering the hemorrhage's nature based on its location, morphology, and relationships with surrounding cerebral structures [1].

However, the recent deluge of biomedical data, especially from digital health tools, electronic health records, and imaging modalities, has both overwhelmed and presented an opportunity for innovative interventions [3]. Herein lies a limitation; despite the abundance of data, only a fractional segment is harnessed for diagnostic enhancements [3]. The current manual interpretations, though robust, are labor-intensive and time-consuming, underscoring the exigency for automated solutions.

Recent strides in AI have ushered in a transformative era for medical imaging. AI-powered diagnostic systems have displayed an adeptness in interpreting CT and MRI scans, matching, or even surpassing human experts in specific tasks [2,3]. These systems not only amplify the diagnostic accuracy but also dramatically slash the diagnostic turnaround times, engendering enhanced patient outcomes. Yet, the scalability of these AI solutions remains tethered to vast computational resources, often restricting their widespread clinical implementation [4].

When it comes to solutions of computational solutions, Edge Impulse is notable for its focus on edge computing. The concept of edge computing revolves around processing data in proximity to its origin, which reduces the reliance on centralized data centers. By analyzing data it enables real time analysis reduces latency and minimizes the expenses associated with transferring data. In the field of imaging these benefits result in faster diagnostics without the need, for complex computational setups.

Edge Impulse, originally devised for applications in the Internet of Things (IoT) devices, has showcased its versatility across diverse domains [5]. From industrial machinery health monitoring to smart agriculture interventions, Edge Impulse has demonstrated its mettle in harnessing data for actionable insights, underlining its prospective utility in health care.

When envisioning Edge Impulse's foray into medical imaging, it's instrumental to consider its prior successes. For instance, in industrial settings, Edge Impulse's predictive maintenance models have preemptively detected machinery faults, obviating costly downtimes [5]. In the agricultural sphere, its models have identified pest infestations in real-time, guiding timely interventions. These success narratives, stemming from disparate sectors, collectively underscore Edge Impulse's potential in revolutionizing medical imaging, particularly in automating ICH detection from CT scans.

This study delves into an analysis of the Edge Impulse platform investigating its intricate features and powerful algorithms. The major target of this study is to shed light on the potential of Edge Impulse in identifying and diagnosing ICH from CT scans. By conducting assessments and empirical validations the aim is to assess how effective Edge Impulse is in automating ICH detection its alignment, with conventional diagnostic methods and its potential to improve clinical decision making by speeding up the process and enhancing accuracy.

2. Method

This research has harnessed the public Intracranial Hemorrhage dataset, colloquially referred to as the "brain-ct-hemorrhage-AMINE-dataset", which is on Kaggle [6]. From this extensive repository, a selection was made, while incorporating images from both the test and training sets, aggregating to an exact count of 1,022 images. Furthermore, al images are based on the RGB format. From a classification vantage, the dataset operates on a binary paradigm. Images are distinctly labeled either as 'normal patient' or 'Hemorrhagic patient'. While this bifurcation simplifies the classification schema, it concurrently elevates the exigency for precision, especially given the critical ramifications associated with intracranial hemorrhages. Illustrative samples from each category have been provided in Figure 1. Adopting an adept compression methodology, images were resized to a dimension of 160x160 pixels.

DOI: 10.54254/2755-2721/41/20230729

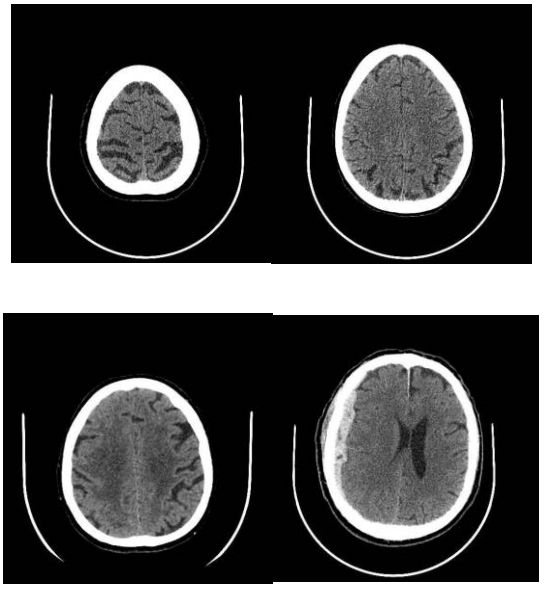

Figure 1. The sample images provided in this dataset [6].

2.1. Edge impulse-based MobileNet model

2.1.1. Introduction of Edge Impluse

Edge Impulse stands as a formidable force in the arena of Machine Learning Operations (MLOps) platforms with cloud-based, particularly tailored for crafting embedded and edge ML (TinyML) systems. With the increasing trend in TinyML, developers faced the conundrum of fragmented software stacks and a plethora of deployment hardware, like limited capabilities of computation [7]. This made the optimization of ML models a challenging affair due to a lack of portability. To alleviate these challenges, Edge Impulse emerged as a practical MLOps platform for crafting TinyML systems on a grand scale. By offering support for various software and hardware optimizations, Edge Impulse paves the way for a more portable and extensible software stack, adaptable to a multitude of embedded systems [8].

With the rapid growth and assimilation of machine learning into embedded systems, technologies such as wake word detection, anomaly detection, and visual object detection became prevalent in low-power devices [9]. However, despite these advancements, the embedded ML development ecosystem struggled to keep up with the soaring demand. Traditional ML development for embedded systems demanded a specific skill set, often necessitating developers to juggle between a new suite of tools and managing conflicting library dependencies. Edge Impulse created a platform to simplify this process by offering a comprehensive space, for gathering some significant models, such as data training models, deep learning models, and then Edge Impulse deploy them to embedded and edge computing devices to fit their user's demand [8].

2.1.2. Introduction of CNN

CNNs, where people often called Convolutional Neural Networks, have revolutionized the field of learning particularly in areas like natural language processing and computer vision. CNNs form the core of Edge Impulses approach. They are highly effective when dealing with grid like data structures such as images.

The CNN architecture consists of three kinds of layers, including pooling layers, fully connected layers, and convolutional layers [10, 11]. Convolutional layers perform convolution operations on input data detecting patterns and features. Pooling layers then reduce the size ensuring spatial invariance to certain image changes or distortions. Finally connected layers interpret these features and generate the final output. In the realm of learning CNNs shine due to their unique advantages and transformative

DOI: 10.54254/2755-2721/41/20230729

applications. They have proven their prowess in tasks like image classification by setting standards with models such, as LeNet 5, AlexNet and ResNet [10, 12].

2.1.3. Introduction of MobileNet

MobileNet, an innovation in the realm of deep learning and neural networks, serves as the primary computing method when calling upon Edge Impulse in the study. Originating as a response to the challenge posed by the intricate and vast neural networks required for image recognition, MobileNet emerged as a savior for devices with limited computational resources [13-15]. The standard neural networks, despite being highly accurate, are computationally intensive and require a large number of model parameters. This makes them ill-suited for mobile and embedded devices that operate under constraints of memory and processing power [15].

MobileNet further fine-tunes its efficiency by adapting two hyperparameters; the resolution multiplier (ρ) and the width multiplier (α) [15]. By modulating these hyper-parameters, MobileNet can strike a balance between the computational cost, model parameter size, and accuracy. Though adjusting these values can lead to a marked reduction in computational demand and model size, it often comes at the expense of reduced accuracy [15].

The genius behind MobileNet lies in its ability to drastically reduce computational demands without compromising significantly on accuracy. Instead of the standard convolution, a cornerstone in the domain of deep learning, MobileNet introduces depthwise separable convolution. This type of convolution breaks down the convolution into two separate stages; a depthwise convolution is performed first followed by a pointwise convolution [10]. The former ensures that each filter uses only one channel when input, and it is used for convolution, while the latter integrates the results from the convolution from different kernels, executing it with a form where 1*1 one is conducted. This method results in a computational cost that is approximately eight to nine times less than the conventional convolution, paving the way for faster processing with minimal loss in accuracy [15].

2.2. Implementation details

Utilizing transfer learning, this study adapted pre-trained MobileNetV2 models to the specific task, and the Hemorrhagic patients would be considered as positive sample.

The rate of positives refers to the likelihood of patients being diagnosed with Intracranial Hemorrhage when they are actually perfectly healthy. On the hand the rate of false negatives represents the probability of patients being diagnosed as normal when they actually have Intracranial Hemorrhage. Lastly the true negative rate indicates the proportion of patients who are correctly identified as normal when they are indeed normal.

In this study, specific attention was given to the hyperparameters. Training encompassed 10 cycles, adopting a learning rate of 0.001. The model's final layer was designed with 8 neurons and incorporated a dropout rate of 0.1. This configuration was specifically chosen to target efficient training within the 40-minute constraint posed by the Edge Impluse.

Moreover, to balance the computational overhead with accuracy, which is crucial for mobile and embedded systems, this study leveraged two distinct MobileNet V2 configurations. The first, with an image input size of 96x96 and α set to 0.35, aimed at reducing the parameter count. Another configuration with the same resolution but an α of 0.1 offered even more parameter savings.

On the other hand, a 160x160 input size was experimented with, having an α of 0.35. This configuration struck a balance between image resolution and model intricacy. Furthermore, a configuration with a resolution of 160x160, but an α of 0.5, was tested, offering a heightened resolution without a significant increase in parameters.

It's important to note that the choice in configurations was bounded by MobileNet's limitations. In Edge Impluse, the maximum permissible α for a 96x96 model is 0.35, and the least allowed for a 160x160 model is also 0.35. Consequently, no controls were set for the 160x160, α=0.5 and 96x96, α=0.1 configurations. This study noted FP as False Positive, TP as True Positive, TN as True Negative, and FN as False Negative.

DOI: 10.54254/2755-2721/41/20230729

3. Results and discussion

The result tables are shown in Table 1, Table 2, Table 3, Table 4 and Table 5:

Table 1. The Result of Data, ρ = 96*96, α=0.35.

Loss	Accuracy	FP	TP	FN	TN
0.31	92.1%	9.0%	93.5%	6.5%	91.0%

Table 2. The Result of Data, ρ = 160*160, α=0.35.

Loss	Accuracy	FP	TP	FN	TN
0.21	94.0%	10.1%	100%	0%	89.9%

Table 3. The Result of Data, ρ = 96*96, α=0.1.

Loss	Accuracy	FP	TP	FN	TN
0.11	96.6%	3.5%	96.8%	3.2%	96.5%

Table 4. The Result of Data, ρ = 96*96, α=0.05.

Loss	Accuracy	FP	TP	FN	TN
0.09	96.8%	3.0%	96.5%	3.5%	97.0%

Table 5. The Result of Data, ρ = 160*160, α=0.5.

Loss	Accuracy	FP	TP	FN	TN
0.01	99.8%	0.3%	100%	0%	99.7%

3.1. Effect of Configuration

The model that uses a configuration of 160x160. An alpha value of 0.5 shows the best overall performance achieving an accuracy of 99.8% with a minimal loss of 0.01. It excels in detecting positives and has zero instances of missing any positives.

In contrast the model with a configuration of 96x96 pixels and an alpha value of 0.35 has the accuracy at 92.1%. Its false positive rate is relatively high at 9.0% compared to configurations indicating more instances where it mistakenly identifies negatives as positives.

3.2. Effect of Resolution

When the resolution is increased from 96x96 to 160x160 pixels (with an alpha value of 0.35) the accuracy improves from 92.1% to 94.0%. Notably the higher resolution model achieves a true positive rate (100%). However, this higher resolution model (with a configuration of 160x160 pixels and an alpha value of 0.35) does have a higher false positive rate at 10.1% compared to its counterpart with a resolution of 96x96 pixels (, with an alpha value of 9%).

3.3. Effect of Width Multiplier (α)

Decreasing the width multiplier from α=0.35 to α=0.1 at a resolution of 96x96 boosts accuracy from 92.1% to 96.6%. The model with α=0.1 exhibits better metrics across all categories, especially in terms of reduced loss and False Positive rate.

Reducing the width multiplier to a value of α=0.05 at the resolution only shows slight improvements. This suggests that as α decreases there are diminishing returns in terms of accuracy and other metrics.

3.4. Effect of False Positive and Negative rate

Among all configurations, the model with ρ = 160x160 and α=0.35 achieves a perfect true positive rate. However, it also has the false positive rate at 10.1%. On the hand both models with configurations ρ =

DOI: 10.54254/2755-2721/41/20230729

160x160, α=0.5 and ρ = 96x96 α=0.1 have zero false negatives indicating that they don't miss any positive samples.

3.5. Effect of Loss

As the width multiplier α reduces or resolution increases, the loss tends to decrease. The model with ρ = 160x160 and α=0.5 has the lowest loss, which corresponds to its high accuracy. While all configurations perform overall considering its superior accuracy and minimal loss values makes the model with ρ = 160x160 and α=0.5 stand out as an optimal choice.

3.6. Clinical Implications

From a clinical perspective, the model's high true positive rate is significant. Missing an ICH diagnosis can lead to severe consequences, including even death. Therefore, a model that detects all positive ICH cases, like the one with ρ = 160x160 and α=0.5, would be immensely valuable in a clinical setting. Indeed, it's also crucial to minimize false positives, as they could lead to unnecessary treatments or interventions which may carry their own set of risks.

3.7. Model Efficiency and Deployment

Considering the deployment on edge devices, it's essential to weigh the trade-off between accuracy and computational efficiency. While the ρ = 160x160, α=0.5 configuration offered the best performance, it might also require more computational resources. In contrast, the ρ = 96x96, α=0.1 model, which offers relatively high accuracy with presumably less computational demand, could be more suitable for real-time applications on constrained devices.

4. Conclusion

This study illuminated the promise of utilizing the Edge Impulse platform, specifically adapting MobileNet models, to achieve high diagnostic precision for ICH detection from CT scans. Through various configurations, the research underscored the potential trade-offs between computational efficiency and diagnostic accuracy. The configuration with ρ = 160x160 and α=0.5 emerged as the frontrunner in terms of performance, showcasing the capability of edge computing in medical imaging diagnostics. However, the ultimate choice of model configuration would invariably hinge on the specific clinical setting, resource availability, and the desired balance between accuracy and computational demand. Further studies could delve into deploying the optimized model in real-world clinical scenarios to gauge its effectiveness and utility. Additionally, integrating feedback loops where the model continually learns from new cases can further refine its diagnostic prowess.

References

[1] Flanders A E Prevedello L M Shih G Halabi S S Kalpathy-Cramer J and Nath J 2020 Construction of a Machine Learning Dataset through Collaboration: The RSNA 2019 Brain CT Hemorrhage Challenge RSNA-ASNR

[2] Matsoukas S Scaggiante J Schuldt B R et al 2022 Accuracy of artificial intelligence for the detection of intracranial hemorrhage and chronic cerebral microbleeds Radiol med 127 1106–1123

[3] O'Neill T J Xi Y Stehel E Browning T Ng Y S, Baker C and Peshock R M 2020 Active Reprioritization of the Reading Worklist Using Artificial Intelligence Has a Beneficial Effect on the Turnaround Time for Interpretation of Head CT with Intracranial Hemorrhage

[4] Waring J Lindvall C and Umeton R 2020 Automated machine learning: Review of the state-of-the-art and opportunities for healthcare Artificial Intelligence in Medicine 104

[5] Chauhan K Jani S Thakkar D Dave R Bhatia J Tanwar S and Obaidat M S 2020 Automated Machine Learning: The New Wave of Machine Learning IEEE 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA) 205–212

DOI: 10.54254/2755-2721/41/20230729

[6] Mahjoubi mohamed amine 2022 brain-ct-hemorrhage-AMINE-dataset Kaggle brain-ct-hemorrhage-AMINE-dataset Kaggle
[7] Banbury C Zhou C Fedorov I Matas R Thakker U Gope D Janapa Reddi V Mattina M and Whatmough P 2021 Micronets: Neural network architectures for deploying tinyml applications on commodity microcontrollers Proceedings of Machine Learning and Systems 3 517–532
[8] Hymel S Banbury C Situnayake D Elium A Ward C Kelcey M Baaijens M Majchrzycki M Plunkett J Tischler D Grande A Moreau L Maslov D Beavis A Jongboom J and Reddi V J 2023 Edge Impulse: An MLOps Platform for Tiny Machine Learning arXiv:2212.03332v3 [cs.DC]
[9] Gruenstein A Alvarez R Thornton C and Ghodrat M 2017 A cascade architecture for keyword spotting on mobile devices arXiv preprint arXiv:1712.03603
[10] Li Z Liu F Yang W Peng S and Zhou J 2021 A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects IEEE Transactions on Neural Networks and Learning Systems
[11] Hubel D H and Wiesel T N 1962 Receptive fields, binocular interaction and functional architecture in the cat's visual cortex J. Physiol. 160(1) 106–154
[12] Qiu Y et al 2019 Semantic segmentation of intracranial hemorrhages in head CT scans. In 2019 IEEE 10th International Conference on Software Engineering and Service Science (ICSESS) pp 112-115 IEEE
[13] Jin J Dundar A and Culurciello E 2014 Flattened convolutional neural networks for feedforward acceleration arXiv preprint arXiv:1412.5474
[14] Rastegari M Ordonez V Redmon J and Farhadi A 2016 Xnornet: Imagenet classification using binary convolutional neural networks arXiv preprint arXiv:1603.05279
[15] Wang M Liu B and Foroosh H 2016 Factorized convolutional neural networks arXiv preprint arXiv:1608.04337