Research, Society and Development, v. 9, n. 10, e889108382, 2020
(CC BY 4.0) | ISSN 2525-3409 | DOI: http://dx.doi.org/10.33448/rsd-v9i10.8382

Classification of Pneumonia images on mobile devices with Quantized Neural Network

Classificação de imagens de Pneumonia em dispositivos móveis com Rede Neural Quantizada

Clasificación de imágenes de neumonía en dispositivos móviles con red neuronal cuantificada

Received: 09/11/2020 | Reviewed: 09/13/2020 | Accept: 09/17/2020 | Published: 09/19/2020

José Vigno Moura Sousa
ORCID: https://orcid.org/0000-0002-5164-360X
Universidade Brasil, Brasil
E-mail: josevigno@prp.uespi.br
Vilson Rosa de Almeida
ORCID: https://orcid.org/0000-0001-9077-2941
Universidade Brasil, Brasil
E-mail: vilson.almeida@universidadebrasil.edu.br
Aratã Andrade Saraiva
ORCID: https://orcid.org/0000-0002-3960-697X
Universidade Estadual do Piauí, Brasil
E-mail: aratasaraiva@gmail.com
Domingos Bruno Sousa Santos
ORCID: https://orcid.org/0000-0003-4018-242X
Universidade Estadual do Piauí, Brasil
E-mail: domingosbruno@prp.uespi.br
Pedro Mateus Cunha Pimentel
ORCID: https://orcid.org/0000-0002-5291-0810
Universidade Estadual do Piauí, Brasil
E-mail: pedrocunha@prp.uespi.br
Luciano Lopes de Sousa
ORCID: https://orcid.org/0000-0003-0551-4804
Universidade Estadual do Piauí, Brasil
E-mail: lucianolps@prp.uespi.br

Abstract

This paper presents an approach for the classification of child chest X-ray images into two classes: pneumonia and normal. We employ Convolutional Neural Networks, from pre-trained networks together with a quantization process, using the platform TensorFlow Lite method. This reduces the processing requirement and computational cost. Results have shown accuracy up to 95.4% and 94.2% for MobileNetV1 and MobileNetV2, respectively. The resulting mobile app also presents a simple and intuitive user interface.

Keywords: Classification; Images; Quantization; Mobile Devices; Pneumonia.

Resumo

Este artigo apresenta uma abordagem para a classificação de imagens de radiografias de tórax de crianças em duas classes: pneumonia e normal. Empregamos Redes Neurais Convolucionais, a partir de redes pré-treinadas em conjunto com um processo de quantização, utilizando o método da plataforma TensorFlow Lite. Isso reduz a necessidade de processamento e o custo computacional. Os resultados mostraram precisão de até 95,4% e 94,2% para MobileNetV1 e MobileNetV2, respectivamente. O aplicativo móvel resultante também apresenta uma interface de usuário simples e intuitiva.

Palavras-chave: Classificação; Imagens; Quantização; Dispositivos Móveis; Pneumonia.

Resumen

Este artículo presenta un enfoque para clasificar las imágenes de rayos X de tórax de los niños en dos clases: neumonía y normal. Usamos redes neuronales convolucionales, de redes pre-entrenadas junto con un proceso de cuantificación, utilizando el método de la plataforma TensorFlow Lite. Esto reduce los requisitos de procesamiento y el costo computacional. Los resultados mostraron una precisión de hasta 95,4% y 94,2% para MobileNetV1 y MobileNetV2, respectivamente. La aplicación móvil resultante también cuenta con una interfaz de usuario sencilla e intuitiva.

Palabras clave: Clasificación; Imágenes; Cuantización; Dispositivos móviles; Neumonía.

1. Introduction

Pneumonia is the leading infectious cause of death among children worldwide. According to the World Health Organization it killed 920,136 children under 5 years of age in 2015, which accounts for 16% of all deaths. Pneumonia affects children and families

everywhere, but it is more prevalent in South Asia and sub-Saharan Africa (World, 2016).

Chest X-rays are often used to assess cases of pneumonia and are the most commonly used diagnostic tests for chest-related diseases. A very small dose of ionizing radiation is used to produce breast imaging (Kermany, et al, 2018).

Pneumonia causes a pulmonary consolidation, meaning that the pulmonary alveoli are full of inflammatory fluid, instead of air (Iorio, et al., 2018). The image identification of pneumonia, as shows in Figure 1, is related to the opacities seen on the radiography. Normal lungs exhibit darker parts near the spine (bronchi filled with air (Kunz, et al., 2018)), whereas abnormal lungs show lighter (opaque) patches, as alveoli are filled with fluid.

The low accuracy in the diagnosis of pneumonia may lead to excessive prescription of antibiotics, which is harmful to patients, and is also a cause of inventory waste. Antibiotics also kill beneficial bacteria, causing unintended health problems (Kurt, Unluer, Evrin, Katipoglu, & Eser, 2018). Moreover, the excessive use of antibiotics may lead to the proliferation of drug resistant bacteria.

Considering this scenario, computational systems capable of providing fast and accurate Pneumonia diagnosis are of great importance and are becoming increasingly common (Manogaran, Varatharajan, & Priyan, 2018). Used as an aid tool, they can minimize errors (Malmir, Amini, & Chang, 2017), while screening potential infected patients.

A recent trend in classification is the use of deep learning techniques (especially Convolutional Neural Networks - CNN's) that can deliver high classification accuracy at the expenses of high computing cost. To reduce this cost, several quantization schemes have gained attention recently, with some focusing on quantization of weight and others focusing on the activation quantizations (Choi, et al., 2018).

As a result, extensive research on weight quantification and activation to minimize CNN's computing and storage costs has been conducted, making it possible to effectively host such solutions on platforms with limited resources (for example, mobile devices) (Choi, et al., 2018).

This paper describes a mobile device system capable of classifying children's chest X-ray images into two classes: Pneumonia and Normal. Samples from a pre-trained CNN are subject to a quantization stage through the TensorFlow Lite platform (Jacob, et al., 2018), considerably reducing the computational cost and processing times.

The proposed method uses two pre-trained neural networks, known as MobileNetV1 (Howard, et al., 2017) and MobileNetV2 (Sandler, Howard, Zhu, Zhmoginov, & Chen, 2018), for the construction of a mobile application aiming at greater mobility. As a result, fast and

accurate diagnosis of childhood pneumonia, especially in remote areas with precarious conditions can be attained.

This paper comprises four sections. In section 2 presents materials and methods, results and conclusion are given in Sections 3 and 4, respectively.

2. The Proposed Method

We now present the proposed methodology for the training and classification of pneumonia from x-ray images on mobile devices.

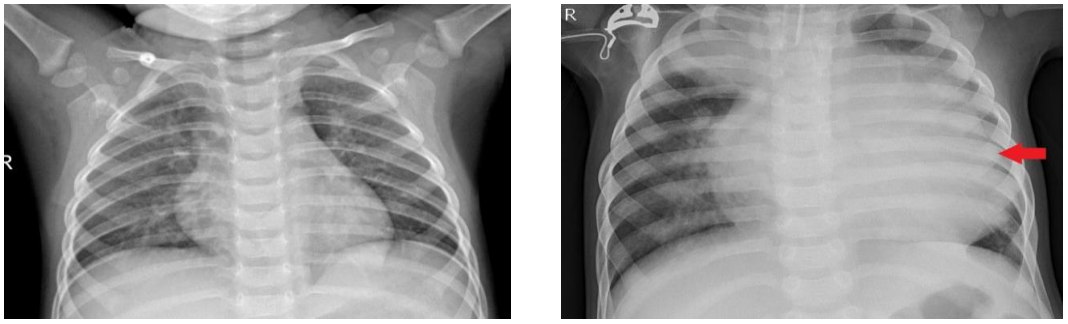
2.1. Dataset

We start by describing the dataset used in the experiments. The images come from the Guangzhou Women and Children Medical Center, taken from pediatric patients aged one to five years. They are all part of the routine clinical procedure (Kermany, Zhang, & Goldbaum, 2018). It contains 5856 chest X-ray images (anteroposterior), categorized as: Viral Pneumonia (1493), Bacterial Pneumonia (2780) and Normal (1583). The dataset possesses quality control, with garbled and low-quality images removed. The diagnosis was given by two specialist physicians and checked by a third one in order to minimize errors (Kermany, Zhang, & Goldbaum, 2018).

In Figure 2 it is possible to analyze how the dataset was divided into training and validation and also the number of images in each class can be analyzed. The first two columns represent the training and test division, the blue column represents the amount of training with 70% of the images, while in orange the amount of test images with 30% is presented. In the last two columns is represented the number of images for each class, in blue is represented the Normal class with 27% of the images, while in orange is represented the Pneumonia class with 73% of the images.

Figure 1- A lung x-ray: a) normal and b) pneumonia. Font: (Kermany, Zhang, & Goldbaum, 2018).

Figure 2 - A lung x-ray: a) normal and b) pneumonia.



Font: Kermany, Zhang, & Goldbaum, (2018).

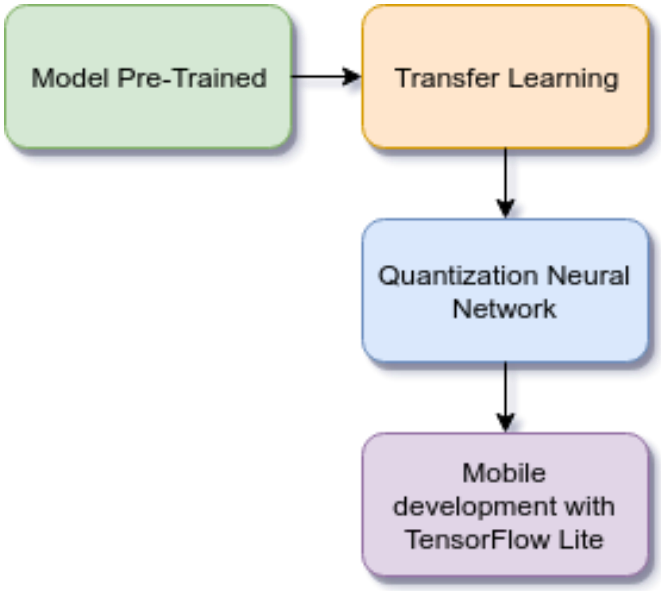
Figure 3 - Dataset Division.

Source: Authors.

2.2. Pipeline Method's

The diagram illustrated in Figure 3 shows the method's main constituent parts. Is comprised four main modules: a) a pre-trained model; b) a transfer learning process in which x-ray lung images are trained; c) quantization through the TensorFlow Lite (Hubara, Courbariaux, Soudry, El-Yaniv, & Bengio, 2017) which aims at optimizing the model for the mobile application and d) the android app for the final classification of x-ray images.

Figure 4 - Pipeline of the proposed method.



Source: Authors.

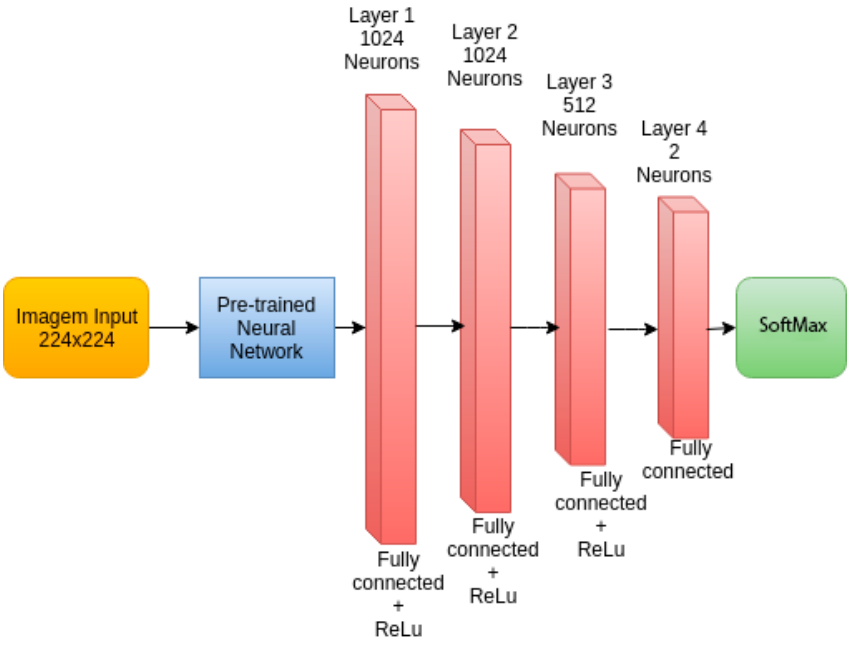
2.3. Transfer learning

Transfer Learning is a common trend in Deep Learning which aims at storing knowledge gained while solving one problem and applying it to a different but related problem. It is present in many applications such as: (Abidin, et al., 2018) (Douarre, Schielein, Frindel, Gerth, & Rousseau, 2018) (Khatami, et al., 2018) (Baltruschat, Nickisch, Grass, Knopp, & Saalbach, 2018) (Chen, Dou, Chen, & Heng, 2018). The technique consists in using a pre-trained model with distinct classes of the problem to be solved (Wu, Qin, Pan, & Yuan, 2018), this becomes an advantage in the use of small data sets (Shallu & Mehra, 2018) because there is a difficulty in getting large enough sets of data for specific problems (Ramalingam & Garzia, 2018), making it has to train complex models such as: VGG19, Xception, Inception V3, among others.

Transfer learning normally preserves the initial and intermediates layers, while the final layer is replaced and trained again (Ramalingam & Garzia, 2018). Figure 4 illustrates the transfer learning process.

For the training of neural networks, all weights are defined as non-trainable, since they were trained with the ImageNet data set. Hence, the last layer of the networks is removed and four dense layers are added, with the latter having the same number of neurons as the number of classes to be classified. The SoftMax function is used to activate the last layer of the networks in Figure 4.

Figure 5 - Transfer learning Architecture.



Source: Authors.

The hyperparameters for each of the networks are shown in Table 1. The epoch parameter used for both networks was set to 100. It indicates the number of times the data set is analyzed at each layer of the network. This epoch values gives MobileNetV1 and edge over MobileNetV2 in relation to the training time which is shorter in the former. The Batch Size parameter (hyperparameter that defines the number of samples to work before updating the internal model parameters) form MobileNetV1 and MobileNetV2 is set to 30 and 40, respectively.

Table 1 - Training hyperparameters of MobileNetV1 and MobileNetv2.

CNN	Learning Rate	Optimizer	Batch Size	Epochs
MobileNetV1	0.0001	Adam	30	100
MobileNetV2	0.0001	Adam	40	100

Source: Authors.

The processing was performed using a GeForce GTX 1060 video card, with 1280 CUDA cores (processors), 6 GB of dedicated memory, 12 GB of RAM and a fourth generation Core i5 processor. The training time for the networks was as follows: 150 minutes for MobileNetV1 and 200 minutes for MobileNetV2.

2.4. Quantized Neural Networks

Quantized Neural Networks (QNNs) use low accuracy weights and activations. These networks are trained from scratch in an arbitrary fixed-point precision. Where in iso-precision, QNNs that use fewer bits require deeper and wider network architectures than networks that use more accurate operators, while requiring less complex arithmetic and fewer bits per weight (Moons, Goetschalckx, Van Berckelaer, & Verhelst, 2017).

A method was introduced to train quantized neural networks (QNNs) with weights and activations of extremely low precision (for example, 1 bit) at runtime. During the training stage, quantized weights and activations are used to calculate the parameter gradients. During the next steps, QNNs dramatically reduce memory size and access, replacing most arithmetic operations with bit-by-bit operations (Hubara, Courbariaux, Soudry, El-Yaniv, & Bengio, 2017).

A quantization scheme that allows inference to be performed using integer-only arithmetic was proposed in (Jacob, et al., 2018). It can be implemented more efficiently than floating-point inference in commonly available hardware-only integers.

In our approach the weights of an existing trained model are loaded and adjusted for quantization. We used the pre-trained meshes MobileNetV1 and MobileNetV2. After being trained with the images of Pneumonia, quantization of the TensorFlow Lite was applied. Results are given in Table 1.

The quantization scheme is an integer mapping q for real numbers r, that is, of the form (Jacob, et al., 2018):

Equation 1

$$r = S(q - Z)$$

This scheme consists in the multiplication of two square arrays $N \times N$ of real numbers, r_1 e r_2 with its product represented by $r_3 = r_1 r_2$. We denote the entries of each of these matrices $r_\alpha (\alpha = 1{,}2{,}\vee 3)$ as $r_\alpha^{(i,j)}$ for ii, jN, and the quantization parameters with which they are quantified as (S_α, Z_α). We denote the inputs quantized by $q_\alpha^{(i,j)}$. Then, Equation 1 become becomes:

Equation 2

$$r_\alpha^{(i,j)} = S_\alpha \left(q_\alpha^{(i,j)} - Z_\alpha \right)$$

From the definition of the multiplication of matrices, we have:

Equation 3

$$S_3\left(q_3^{(i,k)} - Z_3\right) = \sum_{j=1}^{N} S_1\left(q_1^{(i,j)} - Z_1\right) S_2\left(q_2^{(j,k)} - Z_2\right)$$

which can be rewritten as:

Equation 4

$$q_3^{(i,k)} = Z_3 + M \sum_{j=1}^{N} \left(q_1^{(i,j)} - Z_1\right)\left(q_2^{(j,k)} - Z_2\right)$$

Post-training quantization was then performed, thereby reducing the model size while improving CPU and hardware accelerator latency, with little degradation in model accuracy. These techniques can be performed on an already trained TensorFlow floating model and applied during TensorFlow Lite conversion. The models have been fully quantized, i.e., weights and activations. Table 2 shows how the resulting models were fully quantized. We still keep the float input and output for convenience.

Table 2 - Post training quantization options.

Technique	Benefits	Hardware
Post training "hybrid"	4x smaller, 2-3x speedup, accuracy	CPU
Post training integer	4x smaller, More speedup	CPU, Edge TPU, etc.
Post training fp16	2x smaller, Potential GPU acceleration	CPU/GPU

Source: Authors.

2.5. MobileNetV1

This network features a class of efficient models called MobileNets for mobile and integrated vision applications. MobileNets are based on a simplified architecture that uses separable convolutions in depth to build light, deep neural networks. Where two simples global hyperparameters are introduced that switch efficiently between latency and precision. These hyper-parameters allow the model builder to choose the correct size model for their application based on constraints of the problem (Howard, et al., 2017).

The MobileNet model is based on depth-separable convolutions, which are forms of factorized convolutions that factorize a standard convolution into a convolution in depth and a

convolution of 1×1 called convolution point. For MobileNets, deep convolution applies a single filter to each input channel. The point convolution then applies a convolution of 1×1 to combine the convolution outputs in depth. Then the depth convolution with one filter per input channel can be written in Equation 5 (Howard, et al., 2017).

Equation 5

$$G\hat{}_{k,l,m} = \sum_{i,j} K\hat{}_{i,j,m} \cdot F_{k+i-1,l+j-1,m}$$

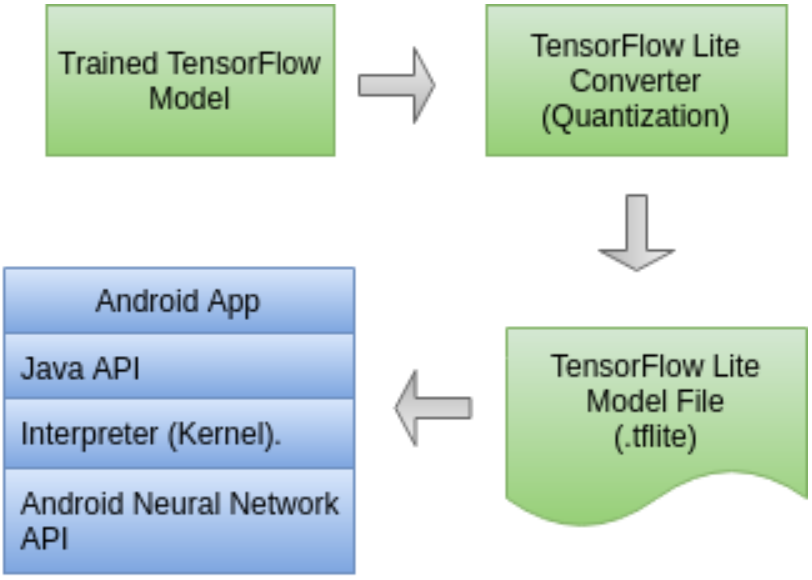
2.6. MobileNetV2

This network drives the state of the art for mobile-oriented computing vision models, significantly reducing the number of operations and memory required, while maintaining the same accuracy. The main contribution is a new layer module: the inverted waste with linear bottleneck. This module takes as input a compressed low-dimension representation that is first expanded to high dimension and filtered with a deep, light convolution (Sandler, Howard, Zhu, Zhmoginov, & Chen, 2018).

2.7. Mobile Application Development

The method chosen in our work uses the Java API of TensorFlow Lite (Jacob, et al., 2018), suitable for Android and IOS application development. TensorFlow Lite is TensorFlow's solution for lightweight models for mobile and embedded devices which allows to run a trained model on a mobile device. It also makes use of hardware acceleration on Android with the Machine Learning APIs (see Figure 5).

Figure 6 - TensorFlow Android.



Source: Authors.

This case the application was developed for the Android platform, which ranks thoracic images. The goal is to aid in the rapid and accurate diagnosis of Childhood Pneumonia. For this is developed a simple and intuitive interface, which consists of two functionalities, the first is: the option to search figure 6 which consists of loading an image present on the device, the second one is the sorting option, where the most likely classification for the image is displayed. The results of this method can be analyzed in the Figures 6, 7.

Figure 6 - App classification Normal images.

b- MobileNetV1

b- MobileNetV2

Source: Authors.

Figure 7 - App classification Pneumonia images.

b- MobileNetV1

b- MobileNetV2

Source: Authors.

3. Results and Discussion

In this section we present the results obtained in each stage of the development of this paper. We provide a comparison between the pre-trained networks MobileNetV1 and MobileNetV2, with a Batch Size parameter set to 30 and 40, respectively. Both networks employ Adam as optimizer, 100 epochs in each training and a rate of 0.0001 learning rate. MobileNetV1 took 150 minutes to be fully trained, while MobileNetV2 spent 200 minutes to complete.

3.1. Evaluation Metrics

The model precision can be estimated by Equation 6 in which Ac_f is the sum of the differences between the actual valuey_i and the expected value $\hat{y}_i$. This allow us to infer the generalization capacity of the network.

Equation 6

$$Ac_f = \sum_{i=1}^{k} (y_i - \hat{y}_i)$$

As a statistical tool, we have the confusion matrix that provides the basis for describing the accuracy of the classification as well as characterizing the errors, helping refine the accuracy (Saraiva, et al., 2018). The confusion matrix is formed by an array of squares of numbers arranged in rows and columns that express the number of sample units of a particular category, inferred by a decision rule, compared to the current category.

The measures derived from the confusion matrix are: total accuracy (used in this work), individual class precision, producer precision, user precision and Kappa index, among others.

The total accuracy is calculated by dividing the sum of the main diagonal of the error matrix x_{ii}, by the total number of samples collected n, according to Equation 7:

Equation 7

$$T = \frac{\sum_{i}^{a}{}_{1} x_{ii}}{n}.$$

To fully evaluate the effectiveness of the models, precision and recall are examined. Unfortunately, precision and recall are often in tension. That is, improving precision usually reduces recall and vice—versa.

Equation 8

$$Precision = \frac{TP}{TP + FP}$$

Equation 9

$$Recall = \frac{TP}{TP + FN}$$

F1 Score is a simple metric, which takes both Precision and Recall into account, so you can try to maximize that number to improve your model. This is simply the harmonic mean of precision and recall.

Equation 10

$$F1Score = \frac{2 * Precision * Recall}{Precision + Recall}$$

AUC - ROC Curve is a measure of performance for sorting problems in various threshold settings. ROC is a probability curve and AUC represent the degree or measure of separability (Bowers & Zhou, 2019).

3.2. Results

Before quantization, data amounts for 70.4 MB of storage in MobileNetV1. However, after quantization the size decreased considerably, reaching 23.3 MB. Likewise, in MobileNetV2, the initial size before quantization was 80.1 MB. Following the same procedure applied to MobileNetV1, data was reduced to 25.0 MB (see Table 1).

Table 3 - Model size before and after quantization.

CNN	Size before quantization	Size after quantization
MobileNetV1	70.4 MB	23.3 MB
MobileNetV2	80.1 MB	25.0 MB

Source: Authors.

This significant decrease in the model size is crucial for the development of the proposed mobile application as it also allows a crucial reduction in the computational cost necessary for the application to work on a mobile device.

Compared with InceptionV3, used in the work of (Kermany, et al., 2018), with the same training dataset, we see an improvement in accuracy for both networks. MobileNetV1 had a 95.4% hit rate, while MobileNetV2 had an accuracy of 94.2%. The InceptionV3 (Kermany, et al., Identifying medical diagnoses and treatable diseases by image-based deep learning, 2018) presented an accuracy of 92.8%. This is a strong indication of the benefits of data quantization compression from pre-trained neural networks applied in the area of image classification. Results are presented in Table 4:

Table 4 - Classification Accuracy for Pneumonia × Normal data samples.

Reference paper	Neural Network	Accuracy
This	MobileNetV1	95.4%
This	MobileNetV2	94.2%
(Kermany, et al., 2018)	InceptionV3	92.8%

Source: Authors.

Moreover, these preliminary results encouraged us think of an efficient Android application, with a simple and intuitive user interface, capable of performing thoracic images classification for normal and pneumonia breast images. We aim at ease of use, mobility, accuracy of classification under low computational cost and energy constraints.

Table 5 shows quantitative results for the classification between Normal and Pneumonia data samples for the assessed metrics. The values were calculated after the quantization step, which for them MobileNetV1 had better performance compared to MobileNetV2, these results are obtained after the quantization method, where the two models had an accuracy of more than 94% in the data. of test. With this, it can be noted that the models performed well compared to other works, such as. (Kermany, et al., 2018) (Dittimi & Suen, 2019).

Table 5 - Performance Evaluation after quantization for the proposed metrics.

CNN	Accuracy	ROC AUC	Kappa	Recall	Precision	F1 Score
MobileNetV1	95.4%	94.1%	88.3%	95.8%	94.5%	95.6%
MobileNetV2	94.2%	93.9%	87.4%	97.9%	93.1%	95.5%

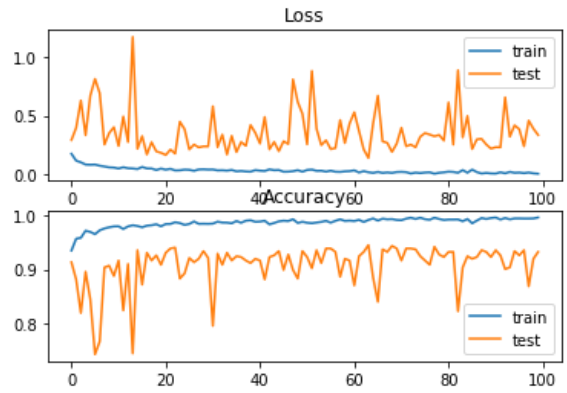
Source: Authors.

The Figures 6 and 7 show the mobile application interface model used in this paper, which demonstrates the efficiency of each pre-trained network used. In the tests performed, MobileNetV1 stands out over MobileNetV2, achieving an improvement of 2.5% and 3.1% in the Normal and Pneumonia class, respectively.

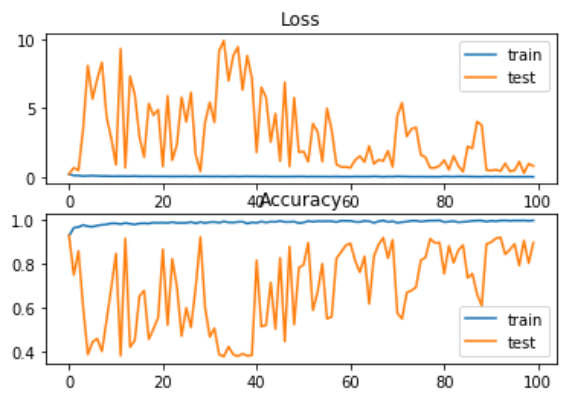
Figure 8 illustrate the training history of the proposed networks. It can be seen that the test accuracy of both models during the training is much larger than the training accuracy. Hence, it is possible to perceive the generalization power of the models when they are tested.

Figure 8 - Training progression MobileNetV1 and MobileNetV1.

b- MobileNetV1



b- MobileNetV2

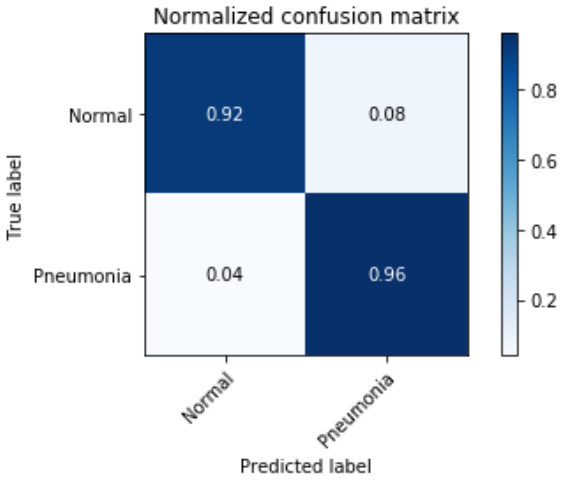


Source: Authors.

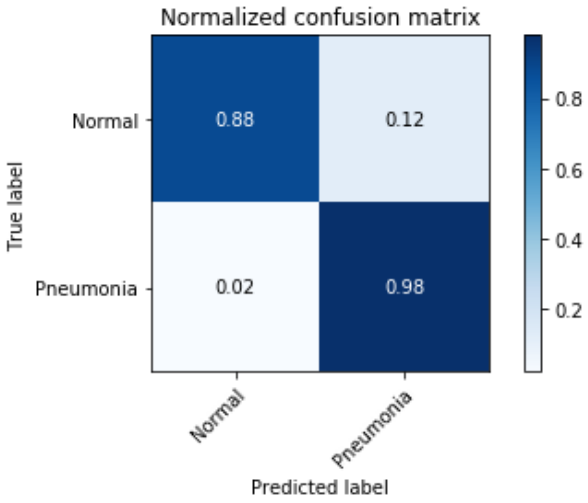
Figures 9 exhibit the confusion matrices for the classification. The results are good if compared with those of. (Kermany, et al., 2018) (Dittimi & Suen, 2019).

Figure 9 - Confusion Matrix MobileNetV1 and MobileNetV2.

b- MobileNetV1



b- MobileNetV2



Source: Authors.

4. Conclusion

This paper proposed a mobile application for the classification of x-ray images comprising normal and diseased images (pneumonia). We employed two pre-trained neural networks, MobileNetV1 and MobileNetV2, with learning transfer strategies together with quantization technique. We showed that the compression, result of the quantization process on

both MobileNetV1 and MobileNetV2 led to a substantial reduction in amount of data to be processed and, therefore, the possibility to efficiently run the classification process on a mobile device.

The mobile application also presents a simple and intuitive user interface and is capable of classifying thoracic images into either normal and abnormal (pneumonia) with an accuracy up to 95.4% and 94.2% for MobileNetV1 and MobileNetV2, respectively. This is an improvement over a similar method (Kermany, et al., 2018) with 92.8% accuracy. As future work, it is intended to carry out a classification with more classes, classifying the type of Pneumonia, which may be viral, bacterial or viral caused by Covid19.

References

Abidin, A. Z., Deng, B., DSouza, A. M., Nagarajan, M. B., Coan, P., & Wismüller, A. (2018). Deep transfer learning for characterizing chondrocyte patterns in phase contrast X-Ray computed tomography images of the human patellar cartilage. *Computers in Biology and Medicine, 95*, 24-33. doi:https://doi.org/10.1016/j.compbiomed.2018.01.008

Baltruschat, I. M., Nickisch, H., Grass, M., Knopp, T., & Saalbach, A. (2018). Comparison of Deep Learning Approaches for Multi-Label Chest X-Ray Classification. *CoRR, abs/1803.02315*. Fonte: http://arxiv.org/abs/1803.02315

Bowers, A. J., & Zhou, X. (2019). Receiver operating characteristic (ROC) area under the curve (AUC): A diagnostic measure for evaluating the accuracy of predictors of education outcomes. *Journal of Education for Students Placed at Risk (JESPAR), 24*, 20–46.

Chen, C., Dou, Q., Chen, H., & Heng, P.-A. (2018). Semantic-Aware Generative Adversarial Nets for Unsupervised Domain Adaptation in Chest X-ray Segmentation. *CoRR, abs/1806.00600*. Fonte: http://arxiv.org/abs/1806.00600

Chen, X., Hu, X., Zhou, H., & Xu, N. (2017). Fxpnet: Training a deep convolutional neural network in fixed-point representation. *2017 International Joint Conference on Neural Networks (IJCNN)*, 2494–2501.

Choi, J., Chuang, P. I.-J., Wang, Z., Venkataramani, S., Srinivasan, V., & Gopalakrishnan, K. (2018). Bridging the accuracy gap for 2-bit quantized neural networks (QNN). *arXiv preprint arXiv:1807.06964*.

Dittimi, T. V., & Suen, C. Y. (2019). Mobile Phone based ensemble classification of Deep Learned Feature for Medical Image Analysis. *IETE Technical Review*, 1–12.

Douarre, C., Schielein, R., Frindel, C., Gerth, S., & Rousseau, D. (2018). Transfer Learning from Synthetic Data Applied to Soil–Root Segmentation in X-Ray Tomography Images. *Journal of Imaging, 4*. doi:10.3390/jimaging4050065

Gavai, N. R., Jakhade, Y. A., Tribhuvan, S. A., & Bhattad, R. (2017). MobileNets for flower classification using TensorFlow. *2017 International Conference on Big Data, IoT and Data Science (BID)*, 154–158.

He, K., Girshick, R. B., & Dollár, P. (2018). Rethinking ImageNet Pre-training. *CoRR, abs/1811.08883*. Fonte: http://arxiv.org/abs/1811.08883

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

Hubara, I., Courbariaux, M., Soudry, D., El-Yaniv, R., & Bengio, Y. (2017). Quantized neural networks: Training neural networks with low precision weights and activations. *The Journal of Machine Learning Research, 18*, 6869–6898.

Iorio, G., Capasso, M., Prisco, S., De Luca, G., Mancusi, C., Laganà, B., Comune, V. (2018). Lung Ultrasound Findings Undetectable by Chest Radiography in Children with Community-Acquired Pneumonia. *Ultrasound in medicine & biology*.

Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., Kalenichenko, D. (2018). Quantization and training of neural networks for efficient integer-arithmetic-only inference. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (pp. 2704–2713).

Kermany, D. S., Goldbaum, M., Cai, W., Valentim, C. C., Liang, H., Baxter, S. L., et al. (2018). Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell, 172*, 1122–1131.

Kermany, D. S., Goldbaum, M., Cai, W., Valentim, C. C., Liang, H., Baxter, S. L., Zhang, K. (2018). Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. *Cell, 172*, 1122 - 1131.e9. doi:https://doi.org/10.1016/j.cell.2018.02.010

Kermany, D., Zhang, K., & Goldbaum, M. (2018). Large Dataset of Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images. *Mendeley Data V3*. doi:http://dx.doi.org/10.17632/rscbjbr9sj.3

Khatami, A., Babaie, M., Tizhoosh, H. R., Khosravi, A., Nguyen, T., & Nahavandi, S. (2018). A sequential search-space shrinking using CNN transfer learning and a Radon projection pool for medical image retrieval. *Expert Systems with Applications, 100*, 224-233. doi:https://doi.org/10.1016/j.eswa.2018.01.056

Kunz, W. G., Patzig, M., Crispin, A., Stahl, R., Reiser, M. F., & Notohamiprodjo, M. (2018). The Value of Supine Chest X-Ray in the Diagnosis of Pneumonia in the Basal Lung Zones. *Academic radiology*.

Kurt, I. S., Unluer, E. E., Evrin, T., Katipoglu, B., & Eser, U. (2018). Urine Dipstick of Sputum for the Rapid Diagnosis of Community Acquired Pneumonia. *Journal of the National Medical Association*.

Malmir, B., Amini, M., & Chang, S. I. (2017). A medical decision support system for disease diagnosis under uncertainty. *Expert Systems with Applications, 88*, 95–108.

Manogaran, G., Varatharajan, R., & Priyan, M. K. (2018). Hybrid recommendation system for heart disease diagnosis based on multiple kernel learning with adaptive neuro-fuzzy inference system. *Multimedia tools and applications, 77*, 4379–4399.

Moons, B., Goetschalckx, K., Van Berckelaer, N., & Verhelst, M. (2017). Minimum energy quantized neural networks. *2017 51st Asilomar Conference on Signals, Systems, and Computers*, 1921–1925.

Ramalingam, S., & Garzia, F. (10 de 2018). Facial Expression Recognition using Transfer Learning. *2018 International Carnahan Conference on Security Technology (ICCST)*, (pp. 1-5). doi:10.1109/CCST.2018.8585504

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4510–4520.

Saraiva, A., Melo, R. T., Filipe, V., Sousa, J. V., Ferreira, N. F., & Valente, A. (2018). Mobile multirobot manipulation by image recognition.

Saraiva., A. A., Ferreira., N. M., de Sousa., L. L., Costa., N. J., Sousa., J. V., Santos., D. B., Soares., S. (2019). Classification of Images of Childhood Pneumonia using Convolutional Neural Networks. *Proceedings of the 12th International Joint Conference on Biomedical Engineering Systems and Technologies. 2*, 112-119. SciTePress. doi:10.5220/0007404301120119

Saraiva., A. A., Santos., D. B., Costa., N. J., Sousa., J. V., Ferreira., N. M., Valente., A., & Soares., S. (2019). Models of Learning to Classify X-ray Images for the Detection of Pneumonia using Neural Networks. *Proceedings of the 12th International Joint Conference on Biomedical Engineering Systems and Technologies. 2*, 76-83. SciTePress. doi:10.5220/0007346600760083

Shallu, & Mehra, R. (2018). Breast cancer histology images classification: Training from scratch or transfer learning? *ICT Express, 4*, 247-254. doi:https://doi.org/10.1016/j.icte.2018.10.007

Ting, K. M. (2017). Confusion Matrix. Em C. Sammut, G. I. Webb (Eds.), *Encyclopedia of Machine Learning and Data Mining* (pp. 260–260). Boston, MA: Springer US. doi:10.1007/978-1-4899-7687-1_50

World, O. (2016). World, Health, Organization pneumonia. Retrieved from https://www.who.int/en/news-room/fact-sheets/detail/pneumonia

Wu, L. (2019). *Biomedical Image Segmentation and Object Detection Using Deep Convolutional Neural Networks.* Ph.D. dissertation, figshare.

Wu, Y., Qin, X., Pan, Y., & Yuan, C. (7 de 2018). Convolution Neural Network based Transfer Learning for Classification of Flowers. *2018 IEEE 3rd International Conference on Signal and Image Processing (ICSIP)*, (pp. 562-566). doi:10.1109/SIPROCESS.2018.8600536

Zoph, B., Vasudevan, V., Shlens, J., & Le, Q. V. (2018). Learning transferable architectures for scalable image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8697–8710.

Percentage of contribution of each author in the manuscript

Vigno Moura Sousa – 20%

Vilson Rosa de Almeida – 20%

Aratã Andrade Saraiva – 20%

Domingos Bruno Sousa Santos – 20%

Pedro Mateus Cunha Pimentel – 10%

Luciano Lopes de Sousa – 10%