ConvNeXt applied to Pneumonia Detection

Yoann ROCH
CentraleSupélec
Gif-sur-Yvette, France
yoann.roch@student-cs.fr

Abstract

Pneumonia is a leading cause of death among children worldwide, resulting in a significant number of child fatalities each year. It accounts for 14% of deaths among children under 5, and, in 2019, 740 180 deaths were attributed to the disease. This highlights the severity of pneumonia as a public health concern, particularly in regions such as South Asia and Sub-Saharan Africa where it is highly prevalent.

Even in developed countries like the United States, pneumonia remains a significant cause of mortality, ranking among the top 10 causes of death. However, early detection and prompt treatment of pneumonia can substantially reduce mortality rates, especially in countries with a high prevalence of the disease. By implementing effective strategies for early identification and treatment, the impact of pneumonia on child mortality can be significantly mitigated.

Hence, this paper compares the ConvNeXt model, introduced in 2022, to the standard ResNet50 and shows its superior viability.

Introduction

Pneumonia is a respiratory infection that can be caused by bacteria, viruses, or fungi present in the air we inhale. When a person develops pneumonia, their lung's air sacs become inflamed and filled with fluid or pus, leading to severe breathing difficulties. The severity of pneumonia can range from mild cases to life-threatening situations. Detecting and treating pneumonia in a timely manner is crucial for effectively managing the high mortality rates associated with this infectious disease, particularly among children, in both developing and developed countries. This highlights the importance to improve constantly our models, to gain accuracy and time.

Convolutional Neural Networks (CNNs) draw inspiration from the visual cortex of the brain and are specifically designed to tackle complex pattern recognition tasks in image data. One of the key advantages of CNNs is their effectiveness in image classification, which is attributed to their ability to learn and recognize both linear and nonlinear patterns.
One of the reasons CNNs excel in image classification is their ability to achieve high accuracy with a relatively smaller number of parameters and connections compared to other types of neural networks. This characteristic makes training CNNs easier and more efficient, as it reduces the computational complexity and the risk of overfitting.
CNN's widespread success in the field of computer vision is not accidental. In numerous application scenarios, a "sliding window" strategy is inherent to visual processing, especially when dealing with high-resolution images. CNNs possess inherent biases that make them highly suitable for a diverse range of computer vision tasks. One of the most crucial biases is translation equivariance, which is a desirable property for tasks such as object detection. This property allows CNNs to effectively capture and recognize objects regardless of their position or location within an image, making them well-suited for various computer vision applications.

This paper presents the ConvNeXt model, developed in 2022 by Zhuang Liu *et al*. [1], and applies it to Pneumonia Detection. This model is compared to ResNet50, one of the state-ot-the-art CNN when it comes to image classification [2]. The ChestXray2017 dataset, containing 5,863 X-ray images and 2 categories (Pneumonia/Normal) was used.

Methodology

This section provides an overview of the methodology of this paper and introduces the ConvNeXt model.

Problem Formulation

The pneumonia detection task is a binary classification problem, where the input is a frontal-view chest Xray image X and the output is a binary label $y \in \{0,1\}$ indicating the absence or presence of pneumonia respectively. Due to a strong imbalance between the number of Pneumonia and Normal images, Weighted Cross Entropy Loss was preferred to a standard Cross Entrpy Loss.

$L_{WCE}(X, y) = -\omega_+ \cdot y \log(p) - \omega_- \cdot (1-y)\log(1-p)$

where p is the probability that the network assigns 1 to the data, $\omega_+ = |N|/(|P| + |N|)$ and $\omega_- = |P|/(|P| + |N|)$ with $|P|$ and $|N|$ the number of positive and negative cases of pneumonia in the training set respectively.

Metrics

Dealing with a binary classification problem, the most important tool of this paper will be the confusion matrix. It provides a good summary of the performances of the model and its ability to predict the class of each sample. The accuracy metric doesn't seem relevant here due to the imbalance of our data, therefore the F1-Score will be preferred. It computes the average between precision, the number of true positives divided by the number of samples predicted positive; and the recall, the number of true positive divided by the number of samples that should have been identified as positive.

The use of a simple F1-Score lines up with other papers, but notice that in this problem, one would rather predict more frequently Pneumonia and be wrong, than predicting Normal and be wrong. A weighted score would be interesting to consider.

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision}+\text{recall}}$$

$$F_\beta = (1+\beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{\beta^2\text{precision}+\text{recall}}$$

ResNet50

ResNet, short for residual network, is a widely used architecture primarily employed for image classification tasks. The convolutional layers of the ResNet network utilize 3x3 filters, and downsampling is achieved through convolutional layers with a stride of 2. The final layer of the network consists of a fully connected layer with 256 units and two channels, employing ReLU and softmax activation functions, respectively.

One notable feature of ResNet is the incorporation of shortcut connections, which address issues such as accuracy degradation and vanishing gradients commonly encountered in deep neural networks. These connections allow the network to skip over layers that it deems less relevant during training. As a result, training error is reduced, and the network converges faster compared to other architectures. Figure 1 illustrates the functioning of shortcut connections in the ResNet50 model.

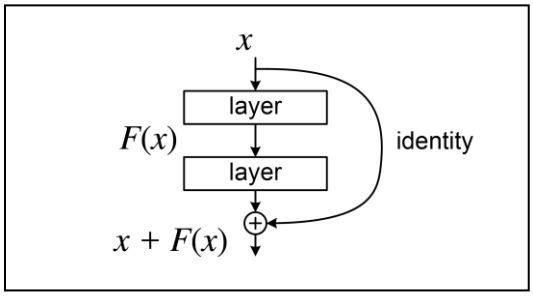

Figure1: Shortcut connection in standard ResNet

ResNet has several variants, including ResNet50, ResNeXt, ResNet34, and ResNetV2. For the classification of chest X-ray images into two classes, the ResNet50 variant was utilized. ResNet50 is a residual network composed of 50 layers.

ConvNeXt

Zhuang Liu *et al.* [1] gradually transcend the standard ResNet50 toward the design of a vision Transformer, leading to a new family of CNNs called ConvNeXts. These models can compete with ViT, introduced in early 2020 [3], and Swin Transformer as well [4].

In this section, we'll just recall some steps made by the authors to improve the ResNet50 network.

Macro-design: change in the stage compute ratio, adopting the 1:1:3:1 ratio of the Swin Transformer ((3,4,6,3) -> (3,3,9,3)) and the stem, first layer of the model that downsamples the input image, switching to a "patchify" stem, meaning the input images are embed in patches.

ResNeXt-ify: ResNeXt [5] employs grouped convolution in the BottleNeck layer to reduce FLOPS. In ConvNext, they use depth-wise convolution. Depth-wise convs are

grouped convolutions where the number of groups is equal to the number of input channels.

Inverted bottleneck: The original bottleneck first reduces the features via a 1x1 conv, then it applies the heavy 3x3 conv and finally expands the features to the original size. An inverted bottleneck block does the opposite. This design can be found in every Transformer block.

Large Kernel Sizes: The modern Vision Transformer models, such as Swin, incorporate a larger kernel size (7x7) compared to traditional Convolutional Neural Networks. However, increasing the kernel size also leads to a higher computational cost. To address this, the authors propose moving up the depth-wise convolution, resulting in a reduction in the number of channels. This approach is reminiscent of the Transformers model, where the Multihead Self Attention (MSA) operation is performed before the MLP (Multi-Layer Perceptron) layers.

Micro-design:

- The Gaussian Error Linear Unit, or GELU replaces ReLU
- Fewer activation functions
- Fewer normalization layers
- LayerNorm instead of BatchNorm
- Separate downsampling layers

A comparison between Swin, ResNet and ConvNeXt block can be found in Figure 2.

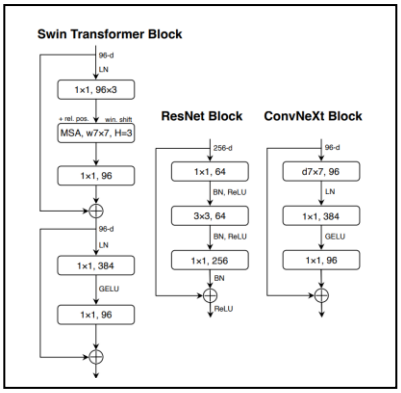

Figure2: Block designs for a ResNet, a Swin Transformer, and a ConvNeXt

Transfer Learning

Transfer Learning is a technique inspired by the way humans apply their existing knowledge to understand and solve new tasks. Similarly, neural networks can be trained and tested on various datasets, acquiring knowledge that can be leveraged for training and testing on new datasets. Transfer Learning involves utilizing the knowledge gained from previous tasks to solve newer tasks [6].

Experimental results

Both models were trained and evaluated on the Chest X-Ray Images (Pneumonia) dataset. The dataset comprised 5239 images for training and 624 images for testing. The performance of the models was assessed using Accuracy, Recall, and F1 as evaluation metrics. These measures were used to analyze and determine the best-performing models.

Data pre-processing

To address the overfitting problem, one effective strategy is to augment the dataset artificially. By applying various transformations to the existing training data while keeping the labels intact, we can create a larger and more diverse dataset. These techniques, known as data augmentation, involve modifying the array representation of the training data. Some commonly used augmentations include grayscale conversions, horizontal and vertical flips, random crops, color jittering, translations, rotations, and more. By implementing a few of these transformations, we can significantly increase the number of training examples and enhance the robustness of our model. Here we will apply the following:

1. Resize and crop to 224x224 as many images are of different sizes.
2. Data Augmentation: random horizontal flip, random rotation and random grayscale.
3. Convert images into PyTorch tensors.

Training step

We used a standard Adam optimizer for the Resnet and an an AdamW for the ConvNeXt. Exponential Learning Rate scheduler, with a base learning rate of 0.0001 was used for both models, and the training step took around 430 min for 10 epochs.

Results

Concerning ResNet's training results, the transfer learning appears tremendously efficient, getting to 95% of accuracy at the 3rd epoch on the validation set. ConvNeXt

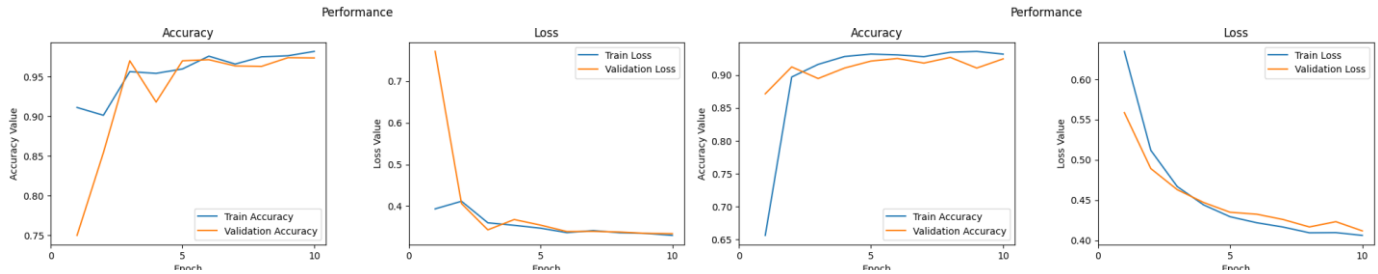

Figure3: Training of (a) ResNet (b) ConvNeXt

suffers from the lack of previous learning and seems to perform worse than ResNet in these tests. The dataset size hingers ConvNeXt: there aren't enough training data for it to catch up with the pretrained ResNet. However, it appears that the loss plateaus in the ResNet training, whereas it decreases efficiently in the ConvNeXt, which shows that the model has more potential. The training results can be found in Figure 3.

Furthermore, the confusion matrices (Figure 4.) show that ConvNeXt has a better accuracy, precision and is overall more balanced. Even though its F1-score (Table 1) doesn't reflect this improvement, ConvNeXt appears to be more viable with proper training.

Model	Accuracy	Precision	Recall	F1-score
ResNet	0.88	0.82	**0.99**	**0.90**
ConvNeXt	**0.88**	**0.86**	0.92	0.89

Table1: Metrics over test dataset

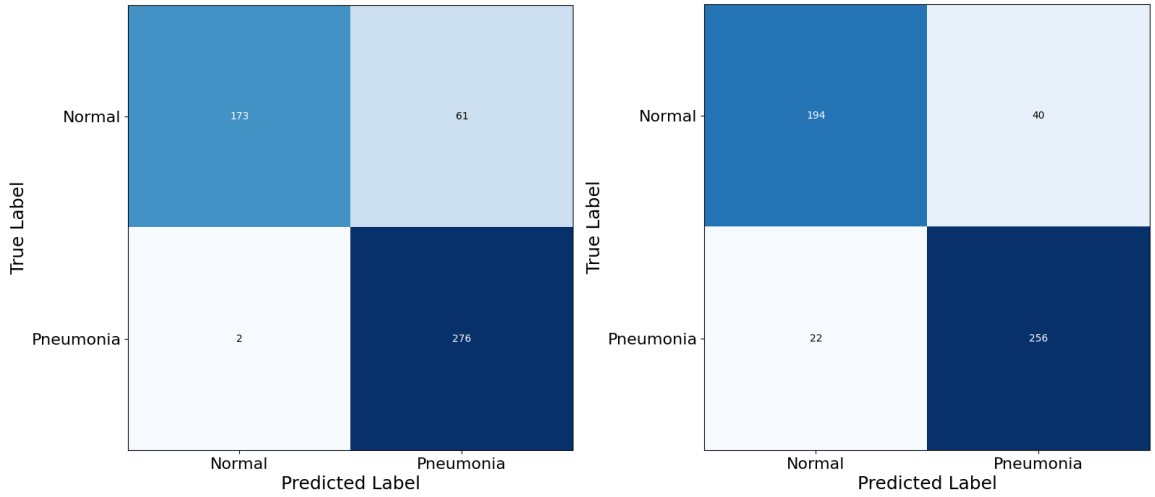

Figure4: Confusion matrix of (a) ResNet (b) ConvNeXt

Discussion and Conclusion

This paper introduces two neural networks designed for real-time applications, which demonstrate exceptional performance. Both models exhibit high accuracy and consistency, and present flaws and strenghts. The ResNet model possess high Recall, which is important as minimizing false negatives is crucial in medical imaging; whereas ConvNeXt is overall more stable and presents higher potential.

One goal of this paper would be to fine-tune the ConvNeXt to achieve better results and be able to outperform in every metric the ResNet.

In their study, Rajpurkar et al. [7] introduced the CheXNet model, which is a highly efficient and accurate model suitable for real-time applications. The models proposed in this paper can be further adapted to classify other diseases with a similar level of accuracy as achieved by CheXNet. To enhance the overall performance of the models, the utilization of larger datasets is recommended.

References

[1] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, Saining Xie, A ConvNet for the 2020s. *arXiv preprint arXiv: 2201.03545,* 2022.

[2] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby, An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv:2010.11929*, Oct 2020.

[4] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, Baining Guo, Swin Transformer: Hierarchical Vision Transformer using Shifted Windows.

[5] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, Kaiming He, Aggregated Residual Transformations for Deep Neural Networks. *arXiv:1611.05431*, 2016

[6] Wentao Mao, Ling Ding, Siyu Tian, Xihui Liang, Online detection for bearing incipient fault based on deep transfer learning. https://doi.org/10.1016/j.measurement.2019.107278, 2019.

[7] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, M.P. Lungren, Chexnet:Radiologist-level pneumonia detection on chest x-rays with deep learning, (2017). arXiv preprint arXiv:1711.05225.