Lung Cancer Detection Using Chest X-rays

Jackson Kehoe
Queen's University
jkehoe00@gmail.com

John Zhou
Queen's University
johnzhou7913@gmail.com

David Nguyen
Queen's University
david.nguyen@queensu.ca

Frank Siyung Cho
Queen's University
20fsc@queensu.ca

Abstract*—Lung cancer is the leading cause of cancer deaths in Canada. Early diagnosis is critical for recovery from lung cancer. The process for diagnosing usually involves x-ray imaging and after a CT scan. The team will be exploring the application of machine learning algorithms to x-ray images because CT scans are not easily accessible for the entire population. The team is using the NIH chest x-ray image data set where the data is narrowed down to masses and nodules found in the lungs. The team further preprocessed the data for an even distribution of cancer and non-cancer images. The 3 different machine learning models that will be explored are an AlexNet, a VGG-16 and a ResNet 50 model. To evaluate the model, the team will measure the accuracy, precision, recall and F1 scores of the model. A smaller subset of the data provided has bounding boxes, the bounding boxes will be included if there is small loss amounts. The team was able to achieve as high as 73.26% accuracy on the testing data, 0.70 precision, 1.00 recall score, and a 0.85 F1 score. Although the algorithm doesn't achieve high accuracies, it is important to consider that for each patient there are often as many as 20 x-ray images included, many of which don't provide clear imaging on cancerous cells. The motivation of the project is to create an aid for medical professionals, and if the algorithm can detect a cancerous image earlier, it can be an efficient and useful tool. Further work on the project includes exploring different models to improve the accuracy, obtaining more bounding box data to create a localization feature, and creating an interactive application that can be used in a medical setting.

I. Introduction

This is the paper for the QMind team exploring machine learning algorithm applications to detect lung cancer. This paper contains all the information achieved through the design process and findings of the team.

A. Motivation

Lung cancer is a leading cause of cancer related deaths in Canada. It is estimated that 30 000 Canadians were diagnosed with lung cancer in 2022 and 20 700 will die from it, representing 24% of all cancer deaths that year. [1] Early detection of cancer is crucial in administering proper treatment and improving survival rates.

Canadian health care has been crippled due to the COVID-19 pandemic, causing health care workers to be burnt out due to the stress of working for more hours and doing more tasks. In a survey of health care workers conducted in 2021, 83% of physicians said they felt more stressed, 65% stated they had an increase of workload, and 46% have had to work additional hours. [2] Due to the environment, 11% of doctors intend to leave health care or change careers within the next 3 years. [2] This loss in practitioners could further strain the system.

Both of these issues can be alleviated with the use of convolutional neural networks trained on chest x-rays to search for cancerous masses. Using this technology, abnormalities that could be easily overlooked can be detected and doctors could spend less time analysing x-rays by getting a quick and accurate second opinion.

B. Problem Definition

Pre-existing work in the area to diagnose lung cancer from imaging uses computed tomography (CT) scans. CT scans are not as easily accessible for the entire population and can cause delays in diagnosis. X-rays have are far more accessible and results are obtained nearly instantly.

The problem is to correctly identify possible cancerous areas on the lungs based on x-ray imaging. If possible with a high enough accuracy to additionally identify with bounding boxes the area of cancerous tissue. By scanning x-rays for cancerous tissues on the lung, the diagnosis process and accuracy can be improved. The medical field is a constantly expanding field for AI, it is important to keep the de-identify the patients and ensure ethical practices with the data.

C. Related Works

In 2018, Ausawalaithong et al. used a DenseNet model with ImageNet initial weightings and 121 layers to classify lung cancer through chest x-rays. [3] They trained their model on the same data set used in this paper. They differed by training their model on 108 899 images with 4992 nodule images being the positives and deemed cancerous, but kept a 50/50 split of cancer non-cancer for the validation and testing set. They did not include masses as positives. 108 899 images were used for training, 2048 images were used for validation, and 532 images were used for testing. They trained on an unspecified number of epochs. On the data set, they achieved an accuracy of 84.02%, a specificity of 85.34%, and a recall value of 82.71%.

In this article, X-ray scans were the primary medical imaging procedure used to identify potentially cancerous nodules. However, other imaging techniques, such as PET scans or MRI scans, can be used in order to obtain accurate results when applying machine learning models [4]. In fact, in 2021, Idrahim et al. used a data set comprised of 33 676 X-ray and CT scans to 4 different image classification models, a VGG19-CNN, a ResNet152V2, a ResNet152V2 + Gated Recurrent Unit (GRU), and ResNet152V2 + Bidirectional GRU (Bi-GRU). Of the 4 techniques the VGG-19 model was deemed

to have the most accurate results with an accuracy of 98.05%, a precision of 98.43%, a recall of 98.05% and a F1 score of 98.24% [4].

II. Methodology

A. Data

The data used is from the Chest-Xray8 data set that is publicly available online [5]. The data contains multiple x-ray images from patients that have been de-identified by name and given subject IDs. The data has labels associated with each image for different lung and heart afflictions. The data was first tested using all lung-related labels. After achieving undesirable results, the data was narrowed down to explore the accuracy in predicting lung cancer. The labels of masses and nodules were the best indicators for predicting lung cancer on chest x-rays. The provided data online had already been split into training, testing and validation data. Initially the data had approximately 1000 cancer positive images with nearly 50 000 cancer negative images. This resulted in the model predicting negative every time. The data was further processed for an even dispersion of data. Shown below in Figure 1 is a graph showing the dispersion between cancerous and non-cancerous images.

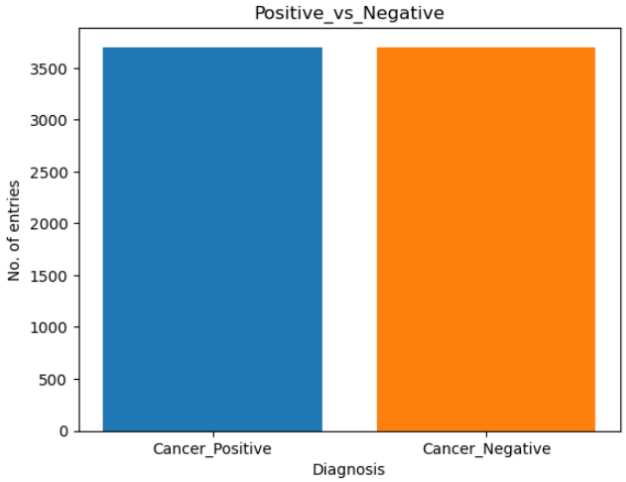

Fig. 1. Dispersion of the data used.

The images used for training and validation had a scale, shear, zoom, rotation, brightness range, horizontal and vertical flip applied to it. For testing, only a scale was applied. From these images, 6300 images were used for the training set, 44 for the validation set, and 1969 for the testing set.

B. Proposed Solutions

The proposed solution was to explore the accuracy of 3 different neural networks in predicting lung cancer. The chosen models were AlexNet, ResNet50 and a VGG-16 model due to their high accuracy in other machine vision applications. Each model will use the adaptive moment estimation(adam) as their optimizer with a learning rate of 1e-4. An AlexNet is based off a convolution neural network with 5 dense layers, 3 max pooling layers, 2 normalization layers, 2 fully connected layers and 1 softmax layer. The architecture can be seen in Figure 2.

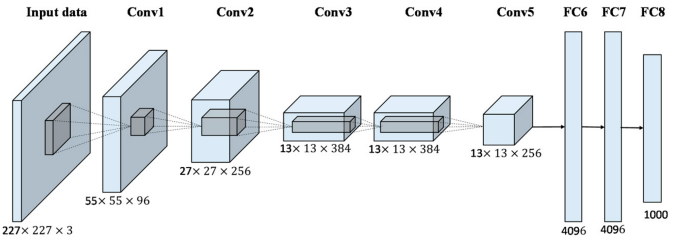

Fig. 2. Architecture of AlexNet[6].

The ResNet50 is a ResNet model consisting of 48 convolution layers with a MaxPool and Average Pool layer for a total of 50 layers. The architecture can be seen in Figure 3.

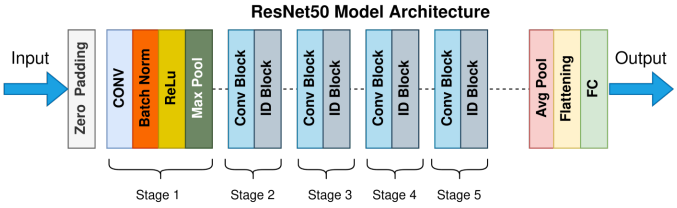

Fig. 3. Architecture of ResNet50[7].

The final model of VGG-16 is an improvement to the AlexNet model. It has 13 convolutional layers and 3 fully connected layers, doubling the layers when compared with AlexNet. The architecture can be seen in Figure 4.

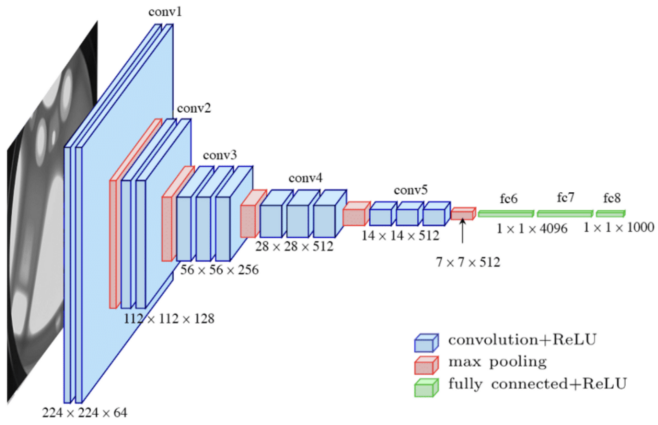

Fig. 4. Architecture of VGG-16[8].

C. Evaluation

The models are evaluated based on their accuracy when predicting positive and negative instances of cancer. 3 different accuracy measures are collected for the training, validation

and testing. Additionally, measures for the precision, recall and F1-score will be collected to evaluate the effectiveness of each model. The equations for the accuracy measures are shown respectively below.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

D. Bounding Boxes

Bounding boxes will be applied to the model if a regression model is able to accurately predict the area of concern. The data for bounding boxes is limited for the x-rays with approximately 160 entries for mass and nodules annotated. Additionally, applying bounding boxes to x-rays is difficult as it doesn't provide perfectly clear images when compared to CT scans. Shown below in Figure 5 is the annotated bounding box on an image.

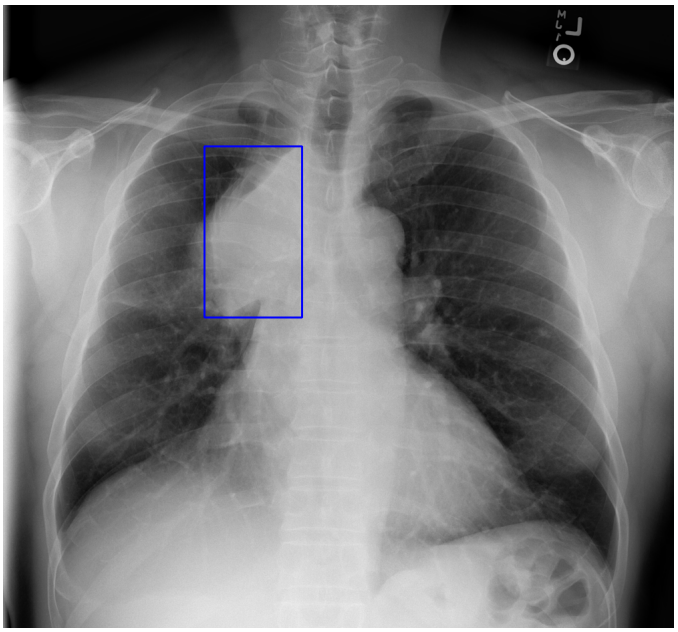

Fig. 5. Bounding Box on Training Image.

III. Results

A. Classification

The performance of the 3 models were decided based on 4 metrics: Accuracy, Precision, Recall and F1. These 4 metrics were decided to be the most relevant however, other metrics such as Specificity or Negative Predictive Value (NPV) could have been used.

The accuracy metric was used to determine how many images the model could correctly classify as having masses/nodules from the total amount of sample images; this is demonstrated below in Equation 1.

The precision metric was used to determine the number of samples which were predicted to have masses/nodules from the total amount of samples with masses/nodules; this is demonstrated in Equation 2.

The recall metric also known as sensitivity or true positive rate was used to determine how many predicted images and actual images had masses/nodules from the total amount of images with masses/nodules; this is demonstrated in Equation 3.

Finally, the F1 metric was used to determine the number of correct predictions the model made; this is demonstrated in Equation 4.

The results for the classification task for all models can be found in Table 1 below.

TABLE I
METRICS FOR BINARY CLASSIFICATION USING DIFFERENT MODELS

Model	Testing Accuracy	Precision	Recall	F1
VGG-16	73.26%	0.70	1.00	0.85
ResNet50	57.29%	0.50	0.85	0.62
AlexNet	49.21%	0.49	1.00	0.66

Out of the models we tested for binary classification, VGG-16 performed the best overall. VGG-16 achieved an accuracy of 73.26%, precision of 0.73 and F-1 score of 0.85 for predicting cancer versus no cancer on a given X-ray.

Our results have several limitations. Firstly, the models likely underperformed due to underfitting, as our models were only trained to 20-30 epochs due to computational restrictions. With further training, the models' accuracy would likely have increased. This was further limited by the distribution of data and low amount of data available for cancer positive versus negative. With the 50/50 distribution we utilized to disperse data, the models were not exposed to the entire range of X-rays for training and were thus not generalizable.

It is also notable that past attempts at mass and nodule classification in the literature achieved significantly lower accuracy compared to other diseases in the same data set (Chest-Xray8).

The training results of the models can be seen through figures 6 to 8. All 3 differed in accuracies for training and validation. It is important to reiterate that the validation set only had 44 images, so interpreting the results should mainly be focused on the training accuracy.

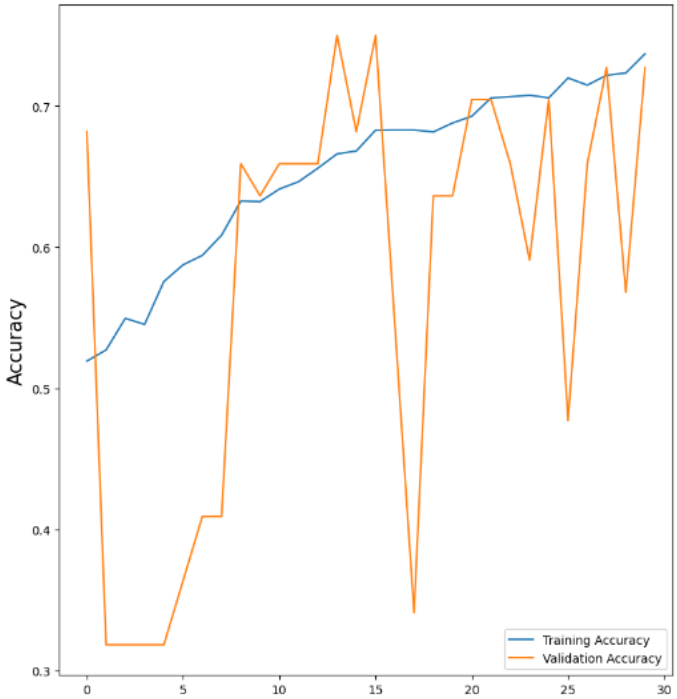

Fig. 6. ResNet50 Training and Validation Accuracy

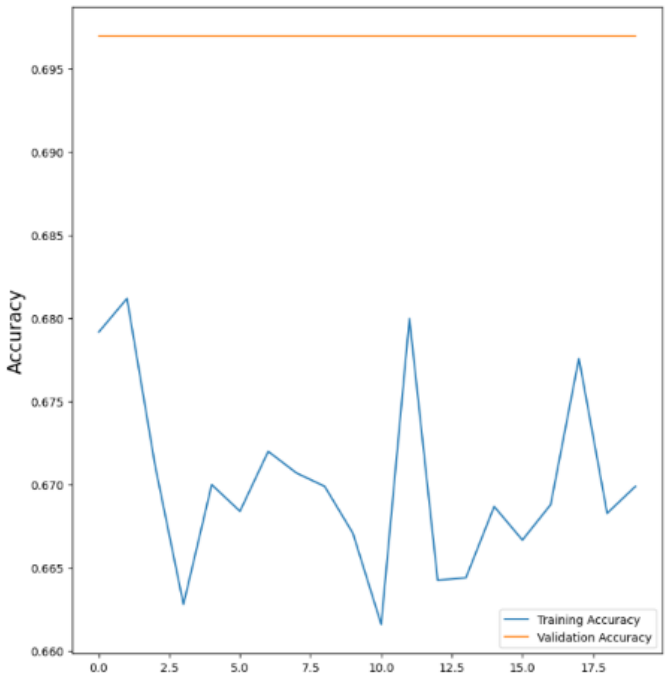

Fig. 8. VGG-16 Training and Validation Accuracy

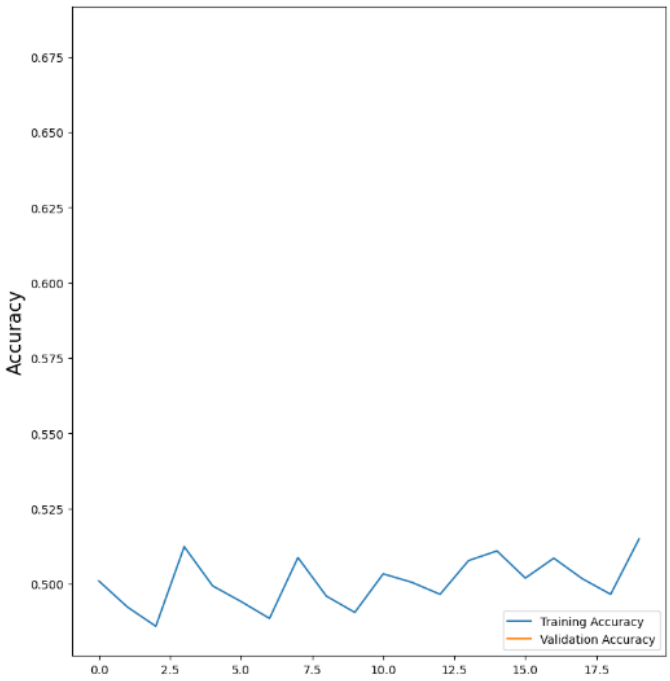

Fig. 7. AlexNet Training and Validation Accuracy

ResNet50 gradually improves its training accuracy from the low 50%s to low 70%s, but its validation accuracy remains inconsistent. As the accuracy hasn't plateaued by the end of training, the accuracy could further improve with more epochs. AlexNet training accuracy remains steady around 50% throughout training. Due to error in running, the validation was not tracked, but the consistent results suggest that it would be consistent like VGG-16's validation accuracy. VGG-16 produces consistent results throughout training, retraining a training accuracy around 67% and a validation accuracy around 70%. The AlexNet and VGG-16 consistent results suggest that the model have been overfitted early on and were not able to improve. This suggests that the models are too simple and could benefit from more layers, a different optimizer, a different loss metric, and/or a smaller learning rate.

B. Localization

The bounding boxes were trained on a basic CNN and loss was measured. The loss by the end of training was 8.97 X e-04 on the training data and 0.0362 on the validation data. The values on the bounding boxes were normalized before being passed into the model. So these values are considered to be large. Good results were unable to be achieved in the bounding boxes due to lack of data available and distinction within the x-ray images. The results of a sample bounding box prediction is shown in blue on the image with green representing the true bounding box in Figure 9.

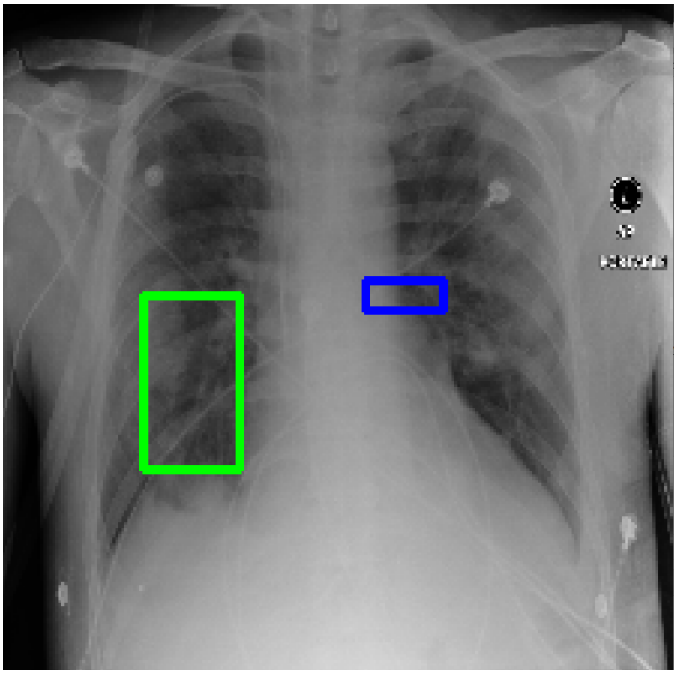

Fig. 9. Predicted bounding box shown in blue, actual bounding box shown in green.

IV. Conclusion

In this paper, 3 convolutional neural networks, an AlexNet, a ResNet50, and a VGG-16, were trained on chest x-ray images from the Chest-Xray8 data set. After applying data augmentation on an even split of images showing cancer and no cancer and splitting the data into training, validation, and testing sets, none of these models were able to produce high accuracy, precision, recall, and F1 scores when classifying images. Of the 3, VGG-16 performed the best with an accuracy of 73.26%, a precision of 0.70, a recall of 1.00, and an F1 of 0.85. The bounding box regression model was also unable to produce bounding boxes that accurately mapped the area with masses or nodules.

As evident in the DenseNet model produced by Ausawalaithong et al. on the same data set used in this paper, high scores in classifications are possible. This suggests a number of changes that could improve results: (1) Add more layers in the network - the DenseNet used 121 layers, while the ResNet50, the largest model evaluated in this paper, had 50 layers. (2) Add more images to validation - the DenseNet had 2048 images for validation while the 3 models had 44 images to validate from. Validation isn't used to train the model, but provides insight on how it performs while training. With a clearer output, researchers will have a better understanding on how to improve it. (3) Run for more epochs - more epochs should help the models converge in accuracy. As seen in the ResNet50 accuracy plot, the model has large spikes throughout training and had not plateaued in its learning yet. (4) Use a different optimizer, loss funciton, and/or decrease the learning rate - the AlexNet and VGG-16 models converged early on and don't suggest improvement through more epochs. Changing 1 or all 3 could improve on this.

The same suggestions can be applied to the bounding box regression model, however this can also be attributed to the small size of the data set with exact positions of mass. More lung cancer chest x-ray data sets should be considered to supplement this.

References

[1] Lung and bronchus cancer statistics. Canadian Cancer Society. (2022, May). Retrieved May 2, 2023, from https://cancer.ca/en/cancer-information/cancer-types/lung/statistics

[2] Impacts experienced by health care workers during the COVID-19 pandemic, by occupation, Canada, September to November 2021. Statistics Canada. (2022, June 3) Retrieved March 2, 2023, from https://www150.statcan.gc.ca/n1/daily-quotidien/220603/cg-a001-eng.htm

[3] Ausawalaithong, W., Marukatat, S., Thirach A., Wilaiprasitporn, T. (2018, August 31) Automatic Lung Cancer Prediction from Chest X-ray Images Using Deep Learning Approach. Arxiv. Retrieved March 2, 2023, from https://arxiv.org/pdf/1808.10858.pdf

[4] Ibrahim DM, Elshennawy NM, Sarhan AM. Deep-chest: Multi-classification deep learning model for diagnosing COVID-19, pneumonia, and lung cancer chest diseases. Comput Biol Med. 2021 May;132:104348.

[5] CXR8. NIH Clinical Center. (n.d.). Retrieved February 24, 2023, from https://nihcc.app.box.com/v/ChestXray-NIHCC

[6] Han, X., Zhong, Y., Cao, L., &; Zhang, L. (2017, August 16). Pre-Trained Alexnet Architecture with pyramid pooling and supervision for high spatial resolution remote sensing image scene classification. MDPI. Retrieved February 25, 2023, from https://www.mdpi.com/2072-4292/9/8/848

[7] Mukherjee, S. (2022, August 18). The annotated resnet-50. Medium. Retrieved February 25, 2023, from https://towardsdatascience.com/the-annotated-resnet-50-a6c536034758

[8] Le, K. (2021, December 8). An overview of VGG16 and nin models. Medium. Retrieved February 25, 2023, from https://medium.com/mlearning-ai/an-overview-of-vgg16-and-nin-models-96e4bf398484