Lung Cancer Classification with Convolutional Neural Networks

David García

BU MET College

Department of
Computer Science

MET CS 767:
Advanced Machine
Learning and Neural
Networks

Professor Zoran
Djordjevic

Fall, 2024

Abstract

The objective of this project is to create a deep learning model to classify medical scans of lungs into two types of cancer, which are Adenocarcinoma and Squamous Cell Carcinoma. Convolutional Neural Networks were used to achieve the previously mentioned goal. Diverse tests were conducted to test different model architectures and parameters, with the aim of achieving the optimal accuracy and cost function.

The dataset that was used for this project is public and was obtained online, titled *Lung-PET-CT-DX*, which stands for Lung PET and CT scan Diagnosis. It includes more than 250,000 medical scans of lungs from 355 different subjects. For the purposes of this project, only 20,000 images belonging to the A and G classes were used to train the models.

In order to train the models, the scans were read into pixel data and then split into train (70%), validation (15%), and test (15%) sets. The percentage of the data that was used for each set is indicated in parentheses. A total amount of 20,000 scans were used during the whole process, and to avoid bias, the same amount of images of each class were assigned to each set. In other words, the data split was 50-50 between the A and G classes. Two different tests were carried out:

1. Custom CNN Architecture
2. Transfer Learning with VGG16

The results are considered great. As expected, VGG performed better than the baseline model, but not by much. The best model achieved a 95.17% validation accuracy and a 0.1356 validation loss, with a 95.03% test accuracy and 0.1428 test loss.

Two of the biggest challenges while completing this project were the massive amount of data and the lack of domain knowledge. The big amount of data caused slow and extensive computation and slow-running programs, which spanned an extensive period of time. My

knowledge of medicine and interpreting scans is limited, so I relied on research I conducted online to gain a deeper understanding of the topic.

Introduction

Healthcare has become one of the most intriguing use cases for disruptive technologies, including Artificial Intelligence. Some of the latest applications of these state-of-the-art technologies include drug development, personalized medicine, and medical document transcription. These technological advancements have reshaped the industry by improving efficiency, lowering costs, and assisting medical professionals in providing the best possible service and diagnosis to their patients. Forbes magazine estimates that AI can save up to $100 billion annually, and regarding medical imaging, it can provide a 10% increase in diagnosis success rate as well as a 20% decrease in time and costs, over a span of two years.
The scope of this project involves diagnosing diseases using deep learning techniques, specifically convolutional neural networks (CNNs). CNNs are specialized for image and video processing, making them an ideal choice for this application.

Methodology

This study required a detailed and strategic planning. The project involved a considerable amount of research, beginning with the search for an appropriate dataset, investigating domain knowledge, and exploring the best tools and frameworks to use. A v2-8 TPU was utilized in Google Colab to train the models, using Python, and the data was stored in Google Drive. To train the Neural Networks, the Sequential API from Keras was implemented.

The utilized dataset, as described previously, contains CT and PET annotated scans of lungs with cancer. The images belong to 4 different classes, one representing a different type of lung cancer, which are labeled as “A” – Adenocarcinoma, “B” – Small Cell Carcinoma, “E” – Large Cell Carcinoma, and “G” – Squamous Cell Carcinoma. The annotations were done by five academic thoracic radiologists, who are experts in lung cancer. The reason the B and E classes were excluded was the great disparity that exists between the number of samples in relation to the A and G classes, which could inject bias into the models. The E class had only five subjects, and

while the B class had 44, a lot of these images were lost when downloading the dataset or when uploading to Google Drive.

Before diving into the training, an exploratory data analysis was carried out, in which the images were evaluated, and various pieces of information were obtained. The fundamental elements were addressed first, for instance, looking at the information of the scans and reading the metadata and available information. Their content was analyzed, where information about the method by which the scan was conducted, demographical information about subjects, and regarding the pixel composition and computational aspects of the scan was exhibited. CT scans had dimensions of 512x512 pixels, while PET scans came in the shape of 200x200 pixels. XML files were included in the dataset, which provided information about the location and size of tumor and its corresponding class, for each subject and each scan. Images were displayed, to obtain a visualization and achieve a deeper understanding of the nature of the observations.

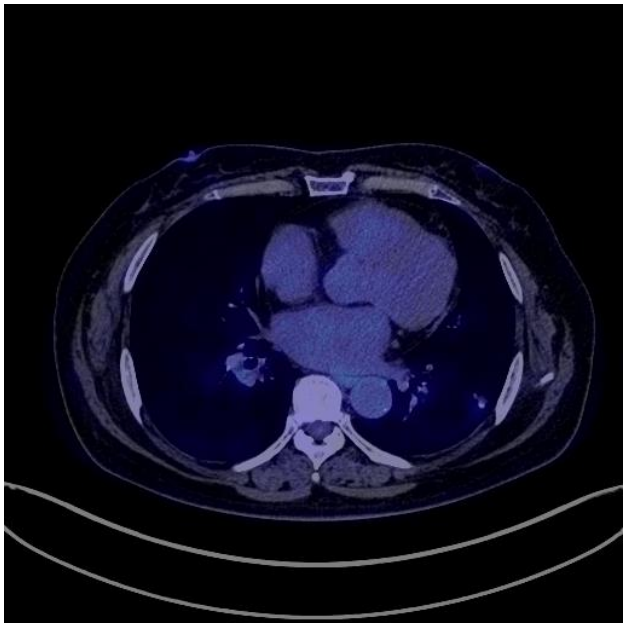

Figure 1: Colored CT scan of dimensions 512x512

After the analysis, images were read from the .dcm files, and were subsequently prepared for training. Scans were read into their pixel data as NumPy arrays, and divided in four categories by class and dimension: class A and 512x512 dimensions, class A and 200x200 dimensions, class G and 512x512 dimensions and class G and 200x200 dimensions. By performing this division, an

equal split of classes and type of scans was ensured. Of a total of 20,000 images used for training, 75% were CT scans and the remaining 25% PET scans, which are also equally divided among classes. PET scans were resized to 512x512 pixels, to match the dimensions of the CT scans, and all values were scaled to be between 0 and 1, dividing images by 255. Before normalization, values that were out of the range of 0 to 255 were clipped to remove noise.

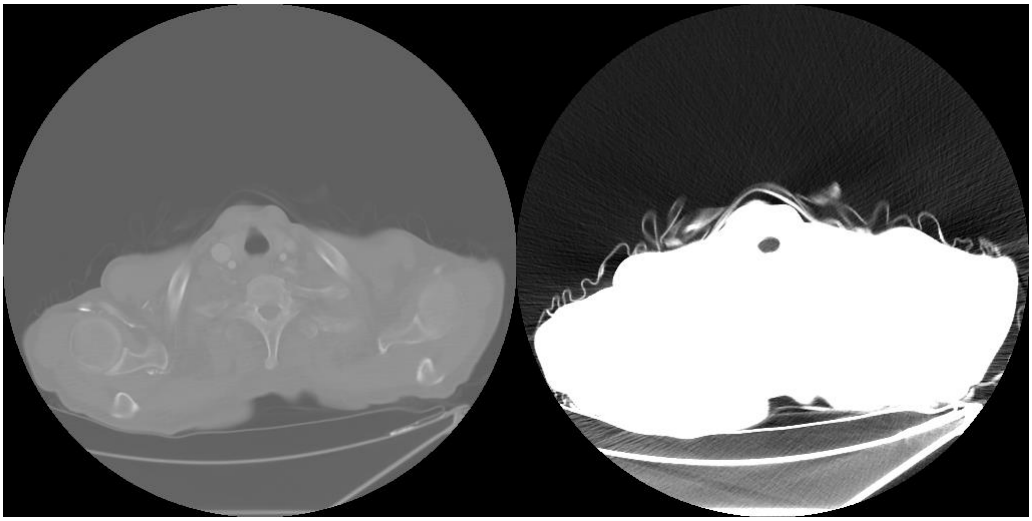

Figure 2: Comparison between normalized and non-normalized scans

Finally, a train-test split was done in the following manner: 70% of data was allocated to the training set, and 15% each to validation and test sets. A more detailed overview of estimations of the training data split can be found in Table 1, presented below.

	Class A		**Class G**			
	CT Scans	**PET Scans**	**CT Scans**	**PET Scans**	**Total**	
Train	5,250	1,750	5,250	1,750	14,000	**70%**
Validation	1,125	375	1,125	375	3,000	**15%**
Test	1,125	375	1,125	375	3,000	**15%**
Total	7,500	2,500	7,500	2,500	**20,000**	
	37.5%	**12.5%**	**37.5%**	**12.5%**		
	50%		**50%**			**100%**

Table 1: Distribution of training data split

Two separate tests were conducted for training: the first using a custom CNN architecture,

referred to as the baseline model, and the second using the pre-trained VGG16 model. It is important to note that an RGB channel was added to the dataset to train the VGG16 model, as it expects the input data to be in those dimensions. The baseline model is comprised of four 2D max pooling and four 2D convolutional layers, as well as a flatten layer, one fully connected layer with 512 neurons and an output layer with sigmoid activation. In total, the baseline model contains 59 million parameters. In the other hand, the VGG16 contains twelve 2D convolution and five 2D max pooling layers, while the rest of the architecture is the same as The VGG16 model has close to 82 million parameters, 14 million that belong to the convolutional base.

As for training itself, both models had almost identical parameters: they were compiled with binary cross entropy and accuracy as metrics and Adam as optimizer. They were trained for 20 and 30 epochs, with early stopping with patience of 7. This would cause training to stop if no improvement or overfitting is shown, saving time and computing power. The batch size was of 64 and validation loss is the metric of interest. Table 2 below presents a more comprehensible breakdown of the parameters of each model.

Parameter	**Baseline**	**VGG16**
2D Max Pooling Layers	4	12
2D Convolutional Layers	4	5
Flatten Layer	1	
Dense Layer, Neurons	1, 512	
Output Layer Activation Function	sigmoid	
Total Parameters	59,223,681	81,824,577
Optimizer	Adam	
Loss Function	Binary Cross Entropy	
Metrics of Interest	Validation Loss, Validation Accuracy	
Batch Size	64	
Early Stopping Patience	7	
Epochs (with Early Stopping)	30 (5)	20 (20)
Size in MB	225.92	312.14

Table 2: Breakdown of model parameters

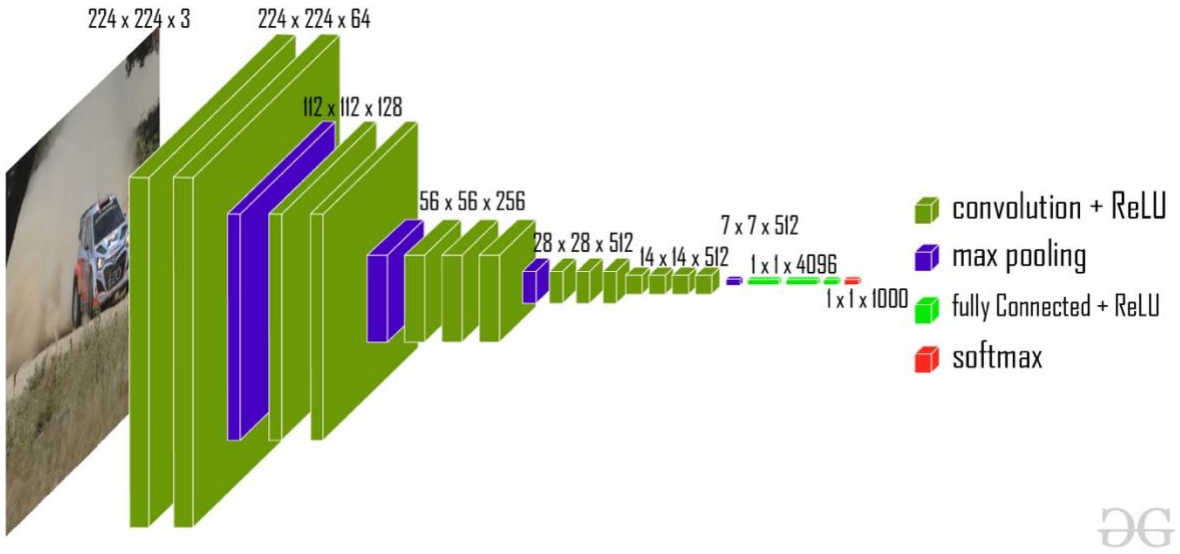

Figure 3: VGG16 architecture

Results

Both models performed considerably well, the VGG16 model performing slightly better among all metrics. It is evident that that the VGG16 model is more robust, generalizes better, and continues to improve as more epochs pass. In the other hand, early stopping proved to be effective with the baseline model, as after 5 epochs its performance did not improve. This clearly indicates that even though it is a strong model, it is not as robust as VGG16, and more trials could be done to improve its performance.

Below are the validation accuracy and loss charts, in Figures 4 and 5, which show the evolution of each model throughout the training process.

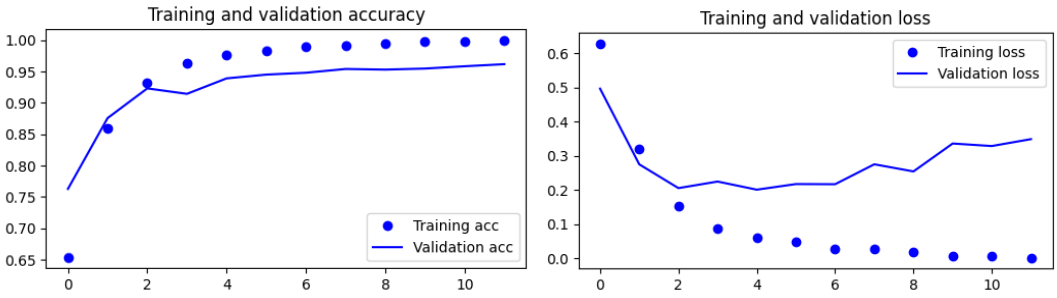

Figure 4: Baseline model results

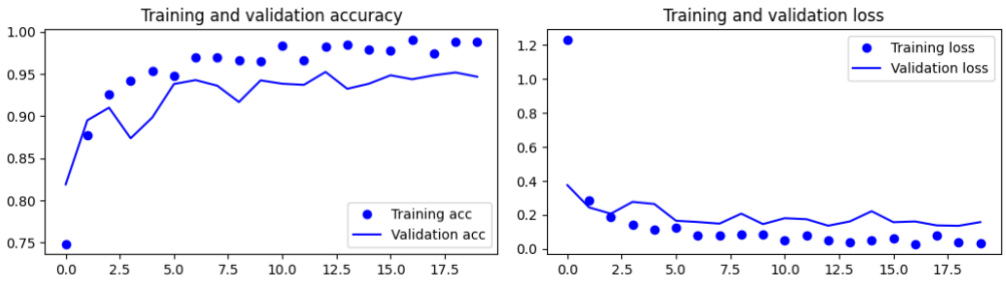

Figure 5: VGG16 model results

It is surprising that the baseline model has a slightly worse performance, considering it has around 23 million parameters less than the VGG16 model, and it trained for five epochs, while VGG16 trained for 20. One could even argue that the baseline model is better, because even though the VGG16 model is superior in all metrics, the baseline could be better to use in practice. The existing tradeoff between accuracy and computing efficiency comes into play: is that 1% difference in accuracy and .06 in loss a significant difference? This warrants further exploration. Below is a comparative table of the metrics of both models.

Metric	Baseline	VGG16	Absolute Difference
Validation accuracy	0.9390	0.9517	0.0127
Validation loss	0.2005	0.1356	0.0649
Test accuracy	0.9460	0.9503	0.0043
Test loss	0.2025	0.1428	0.0597

Table 3: Comparison of model performance metrics

Conclusion

A state-of-the-art performance was achieved, as the resulting model can classify lung cancer scans into the proper cancer category with a 95% accuracy. In other words, 19 out of every 20 scan is classified correctly, which is an improvement from what experts can achieve. In practice, medicine professionals can use an algorithm of this nature to obtain faster and more accurate results. According to a study from Lund University, medical professionals using AI can detect cancer with 20% higher accuracy than if they were not using it, and reducing the workload by 44%. It is important to clarify that the purpose of AI models, especially in a rigorous industry

like healthcare, is to assist medicine professionals and not to replace them, taking into account ethical practices and the sensitive nature of the cases.

The project, however, did not came to fruition without overcoming various challenges. The most significant obstacle was the complexity of computation and the large size of the dataset, exceeding 100 GB. The process of downloading the data locally and uploading it to Google Drive lasted around 50 hours, with the computer running continuously for a few days. Additionally, the complexity of the algorithms caused a long run time during data preprocessing and especially when training the models. The training of the baseline model lasted for 3.5 hours, while the VGG16 model spent almost 10 hours in training. Furthermore, the runtime crashed more than once, which would reset the training process. It is estimated that between 80 and 100 computing units from Colab were used for this study, validating its complexity. An additional obstacle of this project was my lack of domain knowledge in medicine, which led me to rely heavily on online sources for research and information.

Even though the objective of this study was achieved, one could argue that there is room for improvement. For example, only two of the four classes were included. It should be possible to include all four classes through advanced techniques, like data augmentation for instance. Due to time and computing power limitations, only two models were tested, with a small portion of a dataset. It would have been interesting to undertake this study with access to more computing power, to be capable of developing more trials, which would include testing more complex architectures to the baseline model, or even more pretrained algorithms like ResNet50V2 or MedNet. This could improve the already impressive results that were achieved. The help of a medicine student or professional would have been of invaluable help, specially by assisting in the interpretation of scans and metadata, and ensuring every step taken is correct in the medical point of view.

Acknowledgements

I would like to express my gratitude to Dr. Zoran Djordjevic for his guidance during the project and the course, and whose knowledge was invaluable for the completion of this project, and to Katherine Dunkerley, Dr. Ming Zhang, and Dr. Guanglan Zhang for their support throughout the dataset search process.

Disclaimer

This report and the trained algorithms are by no means meant to be used in practice, given that no medical professionals were involved in the study. They are for academical purposes only.

Bibliography

Li, P., Wang, S., Li, T., Lu, J., HuangFu, Y., & Wang, D. (2020). A Large-Scale CT and PET/CT Dataset for Lung Cancer Diagnosis (Lung-PET-CT-Dx) [Data set]. The Cancer Imaging Archive. https://doi.org/10.7937/TCIA.2020.NNC2-0461

Perrone, M. (2024, May 14). *Will AI replace doctors who read X-rays, or just make them better than ever?*. AP News. https://apnews.com/article/ai-algorithms-chatgpt-doctors-radiologists-3bc95db51a41469c390b0f1f48c7dd4e

Viswa, C. A., Bleys, J., Leydon, E., Shah, B., & Zurkiya, D. (2024, January 9). *Generative AI in the pharmaceutical industry: Moving from hype to reality*. McKinsey & Company. https://www.mckinsey.com/industries/life-sciences/our-insights/generative-ai-in-the-pharmaceutical-industry-moving-from-hype-to-reality