CS 109B/Stat 121B/AC 209B/CSCI 109B: Final Project
Glickman, Protopapas
Dor Baruch, Michaela Kane, David Loving, & Brandon Walker

Cancer Diagnosis in Medical Imaging

Problem Statement

In the treatment and prevention of cancer, early detection plays a crucial and often life-saving role. One of the most common methods of early cancer detection is the CT (computer tomography) scan, which is used to detect anomalies in the form of pre-cancerous or cancerous nodules in body tissue. Human radiologists then study these scans and assess the severity and danger of any irregularities they observe.

However, the detection of nodules and their correct diagnosis is extremely challenging, as pre-cancerous nodules are often small, and even when they are visible, they often look like surrounding benign tissue formations [1]. In an attempt to assist radiologists, recent years have seen the development and use of neural networks to help classify CT scans and other medical images for the sake of more accurate, early diagnosis and cancer prevention.

These networks are not infallible themselves, unfortunately, and there are instances in which humans place too much confidence in artificial intelligence without enough critical thought [2]. As a result, there is an increasing demand for networks with a degree of explainability. If researchers and physicians can gain a better understanding of how a network makes its decisions, or what changes to the input variables have on the network's output, then it would be easier for them to accept the network's diagnosis. In the end, this makes for faster, more accurate assessments.

For this project, we attempted to address this issue of explainability through the combination of visualization and Bayesian methods. First, we trained a U-Net on CT Scans from LUng Nodule Analysis (LUNA) to create a network that could ascertain the presence of and locate lesions in lung CT Scans. We then trained InceptionResNetV2, an architecture with proven success in other image recognition tasks, to classify Mammographies from the Digital Database for Screening Mammography (DDSM). In both cases, after training the large models, we make Bayesian adaptations of the model to provide insight into the model's uncertainty and combine that with visualizations that emphasize the most salient parts of the input data.

Data Recources

1. **LUng Nodule Analysis (LUNA) CT Scans**
   The LUNA data we used contains 888 3-dimensional CT scans of patients' lungs (601 patients with diagnosed cancer, 287 without diagnosed cancer), where each scan has a slice thickness smaller than 2.5mm. Each 3D CT Scan consists of a variable number of 2-Dimensional "slices" of 512 by 512 pixels.

2. **Digital Database for Screening Mammography (DDSM) Mammographies** Contains over 10,000 mammographies (4,505 with lesions, 6,206 without lesions). Also includes a CSV file denoting lesion types, shapes, and pathologies (i.e., whether or not a lesion is malignant, benign, benign without callback, or unknown).
   * We ended up using a **preprocessed version** of the DDSM dataset, in which each image was

classified into one of 5 classes. We then changed the classes to malignant and not malignant, to later perform binary classification on the dataset.

Literature Review

In preparation for this project, we conducted research on both image classification neural networks, as well as on the importance and history of explainability in neural networks.

TODO: explain model research most relevant to our model

A recurring theme we found during our research was that of the importance of explainability in models. According to Doshi-Velez et al., by exposing the logic behind a decision, errors can be avoided through correction, and a higher level of trust for the model can be built [2]. In other words, if researchers and doctors understand the "thought process" of a model, that can allow them to use the network's output to inform their own decisions rather than dictating diagnoses.

Human experts often communicate not only their best opinion but also the level of their confidence in that opinion and those of competing explanations. Bayesian models allow the opportunity to sample predictions to gain similar information, but the training of deep Bayesian networks is still often prohibitively difficult. However, recently it has been shown that training a network with dropout can be understood as an approximation to a deep Gaussian process and model uncertainty can be provided by sampling from such a network with dropout still active [3].

Modeling Approach

U-Net

While many popular images segmentation algorithms generate bounding boxes for object localization, in the biomedical field, this is often not good enough. While bounding boxes provide the general location of an object, it still falls short of offering the object's exact location. Going beyond a bounding box to pixel level classification can save an enormous amount of time waiting for experts to annotate biomedical images. Additionally, many approaches require thousands of images to train. Generating this amount of training data is typically not possible in most biomedical settings. Therefore, developing an approach that can be trained on a relatively small amount of training data would be hugely beneficial.

The U-Net Architecture [5] addresses each of the concerns above. The U-Net localizes objects within an image at the pixel level and responds very well to data augmentation for training, drastically reducing the amount of training data necessary to achieve good performance.

We applied the U-Net to the LUNA Dataset [1] in an attempt to automate the localization task performed by radiologists. For simplicity, we extracted a single 2-Dimensional slice from each 3-Dimensional scan. While extracting, we also used the PyLIDC [2] Python package to generate the annotated masks for each of the extracted slices. The network then learns to reproduce the annotated mask by optimizing the Dice coefficient.

Because the U-Net was trained with dropout, we can predict with dropout to draw samples from the network. Rather than provide a simple masked output, we use a combination of brightness and saturation to visually represent the mean and standard deviation of the samples at each pixel. We can also provide a visualization of the distribution per-pixel, which we do for the pixel with the highest mean score. This approach allows a medical professional to look at the output image

[1] LUNA16 Grand Challange

[2] https://pylidc.github.io

and immediately understand where lesions are likely and other areas that may warrant their close attention.

Hybrid Network

The search for better image-classification networks is a large on-going effort, with reference implementations of successful entries often being provided for popular deep learning frameworks. Because our second dataset lacks annotations, we tested multiple Keras' implementations of common network architectures like VGG15, ResNet50, and InceptionResNetV2 to perform binary (malignant/benign) classification on entire images. We first tried to perform transfer learning on VGG16 and ResNet50 by loading Imagenet pre-trained weights for the feature-extraction (convolutional) layers and training addition classification layers that we added to the models. For both networks, we also tried to re-train the networks on the dataset. However, using transfer learning with VGG16 and ResNet50 provided poor accuracy likely because our mammography data is dramatically different from the imagenet classes. Eventually, we got the best accuracy by using InceptionResNetV2 and training the entire network from scratch.

To gain insight into model uncertainty, we used the last pre-dense layer from the network to create new features to be used as inputs in a simple Bayesian neural network. While this is not identical to a fully-Bayesian neural network, the hybrid solution of a traditional convolutional network with a Bayesian classifier at the end still allows us to sample from the posterior predictive distribution and therefore gain some insight into uncertainty. We augment this by also computing saliency maps with the Keras-Vis library which highlight the areas of the image that are most relevant to the model's beliefs. Combined, this provides easily-digested insight into the model's decision-making process.

Results and Interpretation

U-Net

The U-Net visualizations frequently perform well at identifying the locations of lesions. Our sample distributions often were strongly bi-modal, identifying cases where the network finds it very credible that a particular area contains a lesion even when the most frequent prediction is that it almost certainly does not. These are very important because it provides a human expert the opportunity to look closer at such regions for the final determination.

Hybrid Network

The ROC curve shows that this model is not especially useful as a final diagnostic tool, and the posterior predictive distributions often show a very high level of uncertainty. The saliency maps do not show especially strong responses, also indicating that the model is still unsure of how exactly to determine classification. We believe that the model may not be an optimal one for this task or the training regime needs to be modified for better performance. However, we are satisfied that our hybrid network and saliency maps visualization have offered us insight into why the model makes mistakes and performs poorly.

Conclusion and Future Work

References

1. Baker, Darren, et al. *Predicting Lung Cancer Incidence from CT Imagery*. Stanford University, 2017.

2. Doshi-Velez, Finale, et al. "Accountability of AI Under the Law: The Role of Explanation." *SSRN Electronic Journal*, 2017, doi:10.2139/ssrn.3064761.

3. Gal, Yarin, and Zoubin Ghahramani. *Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning*. University of Cambridge, 2016.

4. He, Kaiming, et al. "Deep Residual Learning for Image Recognition." *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, doi:10.1109/cvpr.2016.90.

5. Ronneberger, Olaf, et al. "U-Net Convolutional Networks for Biomedical Image Segmentation." *Informatik Aktuell Bildverarbeitung F??r Die Medizin*, 2017, doi:10.1007/978-3-662-54345.