Uyanga Jung

CAPSTONE PROJECT PROPOSAL

AWS Machine Learning Engineer Nanodegree program

I. DOMAIN BACKGROUND

Dementia is an umbrella term for loss of memory and other thinking abilities severe enough to interfere with daily life. Dementia is known as old people's disease. However, in the recent years, the younger people who are in their 40s get diagnosed with dementia or mild cognitive disorder and the number is increasing.

As of 2021, 9.5 million people in the USA age of 65 and older have some form of dementia and this number is projected to reach 20 million by 2050 according to Alzheimer's Association's report. The main known dementia type is Alzheimer's disease which constitutes 60-80% of the cases and Lewy Body, Vascular, Frontotemporal, and other forms of dementia such as Parkinson's and Huntington's constitute the rest of the cases. In addition, dementia can be mixed or can appear as mix of one or more of previously mentioned types. Only in the USA, 6.2 million people age of 65 and older were diagnosed with Alzheimer's disease and this number is estimated to reach 13.8 million by 2060.

Dementia is one of the leading causes of elderly death not only in the USA but also in the world. A 57.4 million people have dementia globally and 75% of people with dementia are not diagnosed globally. The researches suggest that there are risk factors that contribute to the progression of the disease from young age. Due to slow and long progression of the disease, it's important to diagnose dementia as early as possible to either prevent from going severe or delay the progression of the disease. Dementia is driving the demand and health care cost for elderly care and the healthcare annual spending is projected to increase $350 billion in 2021 to $1.1 trillion by 2050 in the USA.

Therefore, early diagnosis of dementia is important to provide timely care, decrease cost and slow down the progression of the disease.

II. PROBLEM STATEMENT

There are many factors that contribute to dementia throughout the lifetime of a person. The diagnosis of dementia is typically made by a physician or neurologist based on considering genetic components, multiple lab tests, neurological exams such as brain MRI or CAT scans, and psychiatric assessment of mental health & cognitive skills. The researchers and scientists have been using brain images (MRI, X-ray, CAT scan etc.) to predict whether the patient will develop dementia in the future. Since Alzheimer's is the disease that majority of dementia patients get diagnosed and with aforementioned reasons, in this project, I determined to build AI models to predict severity of Alzheimer's disease based on MRI images.

III. DATASETS AND INPUTS

Although it's ideal to use patient's historical information and multiple MRI scan (if possible) to determine the progression of the disease, due to the sensitive nature of medical data, such dataset is scarce. Therefore, labeled brain MRI image dataset was selected to be used in this project. The dataset is from Kaggle platform and can be accessed here.

The dataset consists of train and test images, total of 5,000 images. The images were labeled as one of the following severity levels:

1. Mild Demented
2. Moderate Demented
3. Non Demented
4. Very Mild Demented

Each category is given as a folder with the corresponding images. Typically, medical images are stored as **DICOM** format, however, this dataset is given as **.jpg** format. It's assumed that images are confirmed Alzheimer's disease patients and it's unknown whether it contains any mixed dementia such as Alzheimer's with Lewy body or Vascular dementia.

IV. SOLUTION STATEMENT

Our problem is image classification problem. Therefore, image classification models will be used. I'm planning to utilize pretrained model as well as building a CNN model.

The following solution is proposed.

1. Use Sagemaker script mode
2. Tune the hyperparameters
3. Use debugger and profiler
4. Utilize GPU and
5. Use multi-instance training, if model training job exceed 1 hour GPU training
6. Create endpoint for inference and test the models
7. Try and deploy the model using Docker (optional – if I have extra time)

V. BENCHMARK MODEL

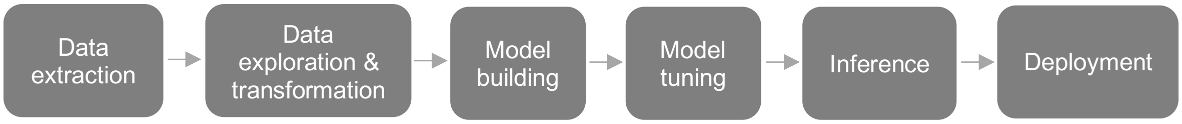
The data has been available on Kaggle platform since 2020. Many people have used the data to build models and utilize predictive tools to achieve high accuracy. The codes can be found here. Depending on the type of framework used, for example, pre-trained (large or small) or AutoML, the results vary. The highest accuracy of 91-97% were obtained by automl tools – Fastai and transfer learning models (VGG model were commonly used) yielded accuracy of 80-87%. Therefore, the goal of this project is to exceed 80% accuracy.

VI. EVALUATION METRICS

The main evaluation metric will be Accuracy score. Depending on the results of data exploration and label imbalance, the metric can be changed to Recall and Precision score.

VII. PROJECT DESIGN

The project will follow the following process.



- • • • -