**CAPSTONE PROJECT (DLFA – Cohort 5)**

--------------------------------------------------------------------------------------------------

1. **Title - Explainability in Deep Models.**
2. **Brief problem statement-**

Despite widespread adoption, machine learning models remain mostly black boxes for end user. Understanding the reasons behind predictions is, however, quite important in assessing trust, which is fundamental if one plans to take action based on a prediction, or when choosing whether to deploy a new model.

1. **Background information:**

AI creates a strong foundation, but we need additional frameworks to help the user understand how the model makes decisions. Explainable AI techniques provide the means to try to unravel the mysteries of AI decision-making, helping end users easily understand and interpret model predictions.

For instance, suppose an AI model flags a warning that an email as fraudulent. Before user take any decision whether he has to keep or discard the email, user may want to know, why did the model take this decision, what are the features/variables triggered this conclusion? Explainable AI (XAI) can help answer these questions.

XAI is essential for ensuring transparency, trust, and accountability in AI, especially in high-stakes applications like healthcare, finance, and law. XAI enables responsible and human-centered AI deployment. XAI addresses concerns that arise from the "black-box" nature of many AI models, especially complex models like deep neural networks, by providing a means to understand and interpret how these models work.

Applications of Explainable AI: Healthcare, Finance, Autonomous Vehicles, Criminal Justice.

1. **Motivation for selection of the project**

To understand how exactly AI Models take decision based on different features and variables and take necessary action to correct them based on the results.

1. **Detailed dataset description and dataset source**

We are planning to use dataset like:

1. ImageNet – ImageNet is a large-scale, widely used database of labeled images, specifically designed to aid in visual object recognition research.

ImageNet contains millions of images organized by the WordNet hierarchy, covering over 20,000 categories.

1. Pascal VOC – The Pascal Visual Object Classes (VOC) dataset is another influential dataset in computer vision research, particularly for tasks like object detection, image classification, and segmentation. The dataset contains images of 20 primary object categories, such as person, bicycle, car, dog, bird, and more, which were carefully chosen to represent common objects in various scenes.
2. **Current benchmark:**

There are several explainable AI frameworks available depending on different techniques. Some of them are mentioned below:

* LIME (Local Interpretable Model-Agnostic Explanations) – It is based on the popular approximation in the perturbation-based methods - the local linear approximation.
* SHAP (SHapley Additive exPlanations)- It is an interpretability framework for machine learning models, based on game theory concepts.
* Grad-CAM (Gradient-weighted Class Activation Mapping)- It is a technique for visually interpreting decisions made by convolutional neural networks (CNNs), particularly in image classification and object detection tasks.
* DAX (distillation aided explainability)- is a framework that combines model distillation with explainability techniques to make complex, often opaque models (like deep neural networks) more interpretable. It is a gradient free framework.

1. **Proposed Plan:**
   1. Our plan is to show usefulness of XAI. We will implement couple of Explainable AI framework like LIME/ GradCAM on top of an AI model and explain them why XAI is useful. We will also try and understand model weights and how it allowing us to interpret features and inputs.
   2. stages with defined deliverables

i. Analysis:

* Defining the problem statement
* Choosing the NN models, XAI models and dataset

ii. Execution:

* Learning about different XAI frameworks and how they work.
* Using XAI model for the chosen NN models for extracting the explainability parameters.

iii. Conclusion:

* Comparison of the results based on the metrices.
* Submission
  1. methodology

1. packages and tools

Tools: Google Colab notebook for coding, Jupiter Notebook;

Packages: pytorch; torch.nn.functional; mathplotlib.pyplot; numpy; pandas; mark\_boundaries; sklearn; lime\_base; lime\_image; Image; models; transformers

1. algorithms

Explainability Algorithms: Grad-CAM; LIME

NN models: resnet50/resnet100; ViT

1. metrics

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | XAI Models / Data Set | | | |
| Parameters | LIME/ResNet | GradCAM/ResNet | LIME/ViT | GradCAM/ViT |
| Accuracy |  |  |  |  |
| Ease of use |  |  |  |  |
| Runtime |  |  |  |  |
| Generalization |  |  |  |  |

1. deployment plan [Optional]

NA

1. **Preliminary Exploratory Data Analysis**

Data analysis and training is not a part of this problem statement.

1. **Expected outcomes**

Comparing the results of couple of Explainable AI framework for

selected dataset and models.

**10. Project demonstration strategy (tentative plans)**

1. Defining the problem statement for the chosen project
2. Initial analysis for the problem statement
3. Defining the approach for the problem statement:
4. Analyzing and understanding the XAI models
5. Choosing the appropriate dataset, NN transformer model, and XAI model the observations
6. Using the XAI models to understand the NN model interpretability, outcome
7. Capture and document the comparative results of the chosen XAI models

IV. Submission of the project

**11. Proposed timeline of project stage executions:**

**12. Team members’ -**

* Ramprassath T C
* Vinod Kumar
* Rajesh Dey
* Sailaja
* Sharad Kumar Patesaria