Good afternoon professor,

Today, I am going to present my scientific project. The focus of my project is on optimizing and interpreting CNN models on the CIFAR-10 dataset using XAI techniques. Thank you for your attention.

Slide2

Here is an overview of my project:

**In this project, I will cover the following:**

1. **First, I will briefly introduce the CIFAR-10 dataset.**
2. **In the Convolutional Neural Network section, I will introduce the Baseline Model and methods for improving the model, including Cross Validation, Data Augmentation, Batch Normalization.**
3. **In the third and fourth parts, I will explain how I used two main XAI techniques: DeepSHAP and Layerwise Relevance Propagation (LRP), and the metrics I used to evaluate them.**
4. **Finally, I will list the references and resources that I used for this project.**

**Slide3**

**The CIFAR-10 dataset is made up of sixty thousand color images, each sized thirty two times thirty two pixels. Its large dataset and low-resolution characteristics will be well demonstrated in the XAI methods, I will introduce later. These images are divided into 10 different classes,鼠标移动到十个类别）**

**Slide4**

**Now, let's discuss my Convolutional Neural Network model, . I'll start with an overview of the CNN architecture.**

**First, we have the convolutional layers. These layers use a kernel that moves over the 2D input data, multiplying parts of the input(手指) and then summing up the results into one output pixel. This helps in extracting features from the images.**

**Next, we have the activation layers. I used the ReLU function,**

**The third layer is the pooling layer, which helps control overfitting**

**After several cycles of convolution, activation, and pooling layers, I flatten the output and pass it to dense layers. Finally, I chose Softmax as the activation function**

**Slide5**

**My baseline CNN model is a VGG-style network with three convolutional blocks, each containing two layers followed by a MaxPooling layer. After training for fifty epochs and setting early stopping, I found the following results:**

**The accuracy of the Base Model is: Seventy-two-point five eight percent**

**The Loss of the Base Model is: One point four eight**

**However, the validation loss is increasing significantly, which means my model is overfitting. next steps I will try to reduce overfitting and increase accuracy**

**Slide6**

**We can use various techniques such as Cross Validation and Dropout to improve the model. Cross Validation helps decrease the loss value，and more importantly, since I created a custom CNN model instead of using a pre-trained model, 手指it helps me determine the optimal hyperparameters.**

**We can also adopt methods like Dropout, Data Augmentation, and Batch Normalization. For example, 手指图片using Batch Normalization ensures that outputs are more evenly distributed after passing the activation function.**

**Slide7**

**After applying Cross Validation, The accuracy of the model did not change，but The loss of the has decreased compared to the basic model, it means Cross Validation can better generalization and reduce overfitting.**

**Slide 8: Improved Model with Dropout**

**Using the cross-validated optimized parameters, i continued improving the model by adding Dropout.**

**We see an accuracy increase of 4%, and the loss value has slightly decreased.**

**Slide 9: Model with Data Augmentation and Dropout**

**Next, we applied Data Augmentation along with Dropout.**

**We see the accuracy remain almost the same, the loss value slightly decreased. This improvement is due to Data Augmentation increases the diversity of the training data, it can help the model generalize better.**

**Slide 10: Model with Dropout and Batch Normalization**

**Then, we used Dropout along with Batch Normalization. The accuracy increased significantly（increase almost 10%）, It is because that Batch Normalization, as previously explained, can help reduce the vanishing(vanishing) gradient problem by scaling data to a specific range.**

**Slide 11:**

**Finally, we combined Dropout, Data Augmentation, and Batch Normalization.**

**At this point, the accuracy reached 90%, the loss value was at its lowest, and no signs of overfitting.**

**Slide 12: Accuracy Comparison**

**Here is a comparison of the accuracy of the baseline model and the improved models. From the chart, we can see that the accuracy improved from Seventy-two percent to ninety percent。 Batch Normalizations手指 and Dropout had the most significant impact on accuracy. This is because Batch Normalization stabilizes and speeds up the training process, and Dropout prevents overfitting by randomly dropping neurons.**

**Slide13**

**Now, let's move on to PART Ⅲ: XAI techniques**

**First, let me explain why we use XAI methods here. it helps us understand and interpret the decisions made by my complex model. It provides insights into which features or parts are most influential in the model's predictions.**

**DeepSHAP combines SHAP values with DeepLIFT to provide explanations. The basic principle behind DeepSHAP is to calculate the contribution of each feature by comparing the model's output with and without the feature.**

**Slide14**

**DeepSHAP can also be optimized by tuning parameters and functions. I found inspiration through the Issues forum/** **fɔːrəm/ on the SHAP GitHub repository /rɪˈpɒzɪtəri/.**

**手指Initially, the results were not good—DeepSHAP did not effectively distinguish** **/dɪsˈtɪŋɡwɪʃ/ between different classes and the explanations were complex and noisy.**

**After adding a softmax function to the output layer, the results became much clearer. This improvement is because the softmax function converts the network's output values (logits) into a probability distribution， in the interval [0, 1].**

**This example highlights that tuning parameters and choosing the appropriate functions are crucial for obtaining clear explanations with DeepSHAP.**

**Slide15**

**To show the performance of DeepSHAP, I selected a highlight examples.**

**This image clearly demonstrates the role of XAI. The reason it is identified as a bird rather than a horse手指 is that the bird's beak in this image is very sharp, and the positive relevance around the bird's beak is higher compared to the horse's heatmap. Therefore, the model classifies it as a bird.**

**Slide16**

**Now, let's explain the second XAI technique– Layerwise Relevance Propagation (LRP).**

**Gamma Rule (LRP-γ) Epsilon (LRP-𝟄)**

**Slide17**

**I applied different lrp methods individually。We can see in Cifar10 dataset, The explanation of lrp-0 is complex and noisy**

**Gamma Rule (LRP-γ) Epsilon (LRP-𝟄)**

**Slide 18**

**Due to the depth and complexity of my model, as I shown in the previous slide, a single LRP method cannot be effectively applied. I get inspiration from one GitHub project and combined different LRP methods to apply them in my model.**

**For the earliest convolutional layers (corresponding to layers 0-5 in my model), I used the Z-B Rule.**

**For the intermediate convolutional layers, I applied LRP-γ. Gamma Rule**

**For the deeper convolutional layers and the first fully connected layers (corresponding to layers 12-17 in my model), I used Epsilon (LRP-𝟄)**

**Finally, for the final fully connected layers, I applied LRP-0.**

**Slide 19:** **Slide 19: Effectiveness of Combining LRP Methods**

**We can see that the effect of combining LRP methods is clearly better than applying each one individually. This improvement is because different methods correspond to different layers and features of the model, it allowed the do more accurate and detailed explanations.**

**For example, we can observe that the top part of the ship's**

**cockpit** **['kɔkpit] and the outline of the ship appear in red, indicating positive relevance, and the waves caused by the ship are in blue, corresponding to noise or irrelevant background features.**

**So In deep、complex model and large image tasks, it is important to choose the appropriate LRP method. By combining methods, We can leverage the strengths of each approach**

**Slide 20:**

**In this part, I'll introduce some metrics to compare the quality of DeepSHAP and LRP. first, let me explain the concepts of Local and Global Explanations.**

**Local explanations aim to explain the prediction of a small subset of data. For example, in image classification, a local explanation might highlight which parts of the image were most important for classifying it as a frog.**

**Global explanations aim to provide an overview of the model’s behavior. For instance, in assessing** **/əˈsesɪŋ/ the risk of wound infection, SHAP values can be used to explain the impact of different features. We can see that blood transfusion has a very positive influence.**

**Slide21**

**lrp can better detect the outline of the airplane, and Deepshap is more complex and noisy. so in local explanations, applying LRP in my model is superior to DeepSHAP.**

**Even for misclassified images, LRP can detect more feature contributions in the original label. We can see Deepshap just detect few feature contributions in label bird。**

**Slide 22**

**Now let's evaluate the two models using the Consistency metric**

**Compared to LRP, DeepSHAP has much higher consistency. This means it’s better at figuring out how different features impact the model’s decisions.**

**For example, in assessing the risk of wound** **/wuːnd/ infection, DeepSHAP can effectively identify the impact of features like transfusion, sex type, age.**

**However, DeepSHAP is not as suitable for CIFAR-10 image recognition because it is difficult to list specific features like sex type and age for images.**

**Slide 23**

**Next, let's use the Stability metric to evaluate the models.**

**LRP’s higher stability means it gives more robust explanations for similar inputs.**

**In CIFAR-10, we can't effectively list different features, but when you need to understand how each pixel in the image contributes to the prediction, LRP is very useful.**

**This also proves that DeepSHAP is not suitable for large datasets like CIFAR-10.**

**Slide24**

**Here, I have listed the advantages and disadvantages of DeepSHAP and LRP.**

**As we discussed earlier, DeepSHAP has high consistency and can be applied to any machine learning model. On the other hand, LRP has low consistency and complex implementation. This makes DeepSHAP more suitable for scenarios  /sɪ'nɑ:ri:oz / where specific features can be listed, such as disease assessment and house price prediction.**

**And, LRP has high stability and is effective for deep models. It is more suitable like in my project，which involving complex models and large dataset. I can use different LRP methods for different neural network layers.**

**Overall, compared to DeepSHAP, LRP is more suitable for my project**

**Slide 25**

**In this section, I have listed some important references and resources. These include literature and GitHub repositories [rɪˈpɒzɪtri], they provided solutions and inspired innovative approaches for my project.**