Study Exercises: Explainability and Interpretability

1. What is the difference between explainability and interpretability of a model ?
2. Why is explainability important in the context of neural networks and CNNs?
3. What are the potential consequences of using a neural network without explainability in high-stakes applications?
4. Can you explain the difference between local and global explainability in neural networks?
5. How do saliency maps help us understand the predictions made by CNNs?
6. What is the role of activation maps in interpreting CNNs? Provide an example.
7. Describe the concept of integrated gradients and how it can be applied to CNNs.
8. How do SHAP values work, and why are they valuable for understanding model predictions?
9. What are some common tools and libraries used for implementing explainability in neural networks?
10. Explain the ethical considerations when using explainability techniques in machine learning.
11. How can explainability help detect and address biases in neural network models?
12. Discuss the trade-offs between model complexity and interpretability in neural networks.
13. What is the difference between feature importance and feature attribution in CNNs?
14. Describe the principles behind LIME (Local Interpretable Model-agnostic Explanations).
15. How does Grad-CAM (Gradient-weighted Class Activation Mapping) work, and when is it useful?
16. Why is it important to provide a coherent and globally consistent explanation for model predictions (as in SHAP values)?
17. Can you explain how explainability can be used to identify overfitting in a CNN model?
18. In what ways can explainability enhance collaboration between data scientists and domain experts?
19. Discuss the challenges of achieving both high accuracy and high explainability in a neural network model.
20. How can explainability methods assist in troubleshooting and improving model performance?
21. Provide an example of a situation where the lack of explainability in a model's decision could lead to a catastrophic outcome.

22. Why is it essential to consider explainability as a fundamental part of model development and not just an optional feature?
23. Explain the concept of model-agnostic explainability and its advantages.
24. How does interpretability in machine learning models contribute to model trust and adoption by end-users?
25. What role does explainability play in ensuring compliance with regulations, such as GDPR, in AI and machine learning applications?
26. Discuss the limitations and potential biases associated with explainability methods in neural networks.
27. Explain how explainability can be used to validate that a neural network is learning meaningful patterns in the data.
28. Can you identify scenarios in which black-box models might be preferred over transparent models, and vice versa?
29. What techniques can be used to visualize and communicate complex model explanations to non-technical stakeholders?
30. How does explainability facilitate the process of hyperparameter tuning in neural networks?
31. In what ways can explainability be integrated into an organization's machine learning pipeline for ongoing model monitoring and improvement?
32. How might the use of explainable AI impact the acceptance and adoption of machine learning in industries where it's currently underutilized?
33. Can you think of examples where AI models might produce ethically or socially problematic results even if they have high accuracy? How could this be addressed?
34. Consider the ethical dilemmas posed by autonomous vehicles. How should they be programmed to make decisions in no-win scenarios?