VisionAI Viva Voce Preparation: 25 FAQs

I. Core Theoretical & Model Questions (Sections 1, 2, 4, 5)

1. **What is the fundamental difference between Semantic Segmentation and Instance Segmentation?**
    - *Answer Focus:* Semantic labels every pixel but treats multiple instances of the same class (e.g., five cars) as one entity. Instance segmentation detects and labels **each separate instance** of an object individually, giving a unique mask to each.
2. **Why did you choose Mask R-CNN over a single-stage model like YOLACT or SOLO for this project?**
    - *Answer Focus:* The project prioritized **maximal segmentation fidelity (APmask)** and accuracy over ultra-high speed. Mask R-CNN, being a two-stage detector with **RoIAlign**, provides significantly better pixel-level precision for mask boundaries.
3. **Explain the role of the RPN (Region Proposal Network) in the Mask R-CNN architecture.**
    - *Answer Focus:* The RPN scans the feature maps to efficiently propose preliminary object bounding boxes (**proposals**). It essentially filters the search space, determining *where* objects might be, before the main heads determine *what* they are and *how* to segment them.
4. **What is the "quantization problem," and how does RoIAlign solve it compared to RoIPool?**
    - *Answer Focus:* RoIPool forces feature maps into integer coordinates, causing misalignment (quantization error) that harms mask accuracy. **RoIAlign** uses **bilinear interpolation** to accurately extract feature values at floating-point coordinates, ensuring precise pixel-to-pixel alignment.
5. **Mask R-CNN is a multi-task network. What are the three parallel tasks it optimizes, and why is the mask prediction handled differently?**
    - *Answer Focus:* Tasks are **Classification**, **Bounding Box Regression**, and **Mask Prediction**. The mask prediction head uses a small FCN and applies a **per-pixel binary cross-entropy loss** independently for *each* class, decoupling it from classification/regression to achieve high mask quality.

II. Implementation & Technical Deep Dive (Sections 6, 7, 8, 11, 13)

6. **Why was the ResNet50 backbone chosen, and what role did Transfer Learning play?**
    - *Answer Focus:* **ResNet50** uses **residual connections** to allow for deep, effective training. **Transfer Learning** (using ImageNet pre-trained weights) dramatically reduced the required training time and data, providing a strong starting point with robust feature extraction capabilities.
7. **Detail the primary mechanism by which the Feature Pyramid Network (FPN) achieves scale invariance.**

- *Answer Focus:* FPN uses a **top-down pathway** to propagate high-level semantic information (from deep layers) down to the higher-resolution, low-level feature maps via **lateral connections**. This ensures that features used for detecting small objects still contain rich semantic context.

8. **The loss function has three components. How did you handle the weighting (λ)?**
   - *Answer Focus:* The total loss is a weighted sum. Typically, the classification and regression components are balanced, and the mask loss is treated separately. The default Mask R-CNN implementations usually weigh RPN and RPN box losses, keeping the mask loss weight at 1.0 (or near 1.0) as the binary cross-entropy implicitly provides balanced gradient signals. *(Refer to Section 13.)*
9. **Describe the Data Preprocessing you performed. Why normalize to the [−1,1] range instead of just dividing by 255?**
   - *Answer Focus:* Preprocessing included **Normalization** and **Resizing (1024×1024)**. Normalizing to [−1,1] (or subtracting the mean and dividing by the standard deviation) helps center the data distribution around zero, which is known to accelerate the convergence of optimization algorithms like SGD.
10. **Your report mentions Batch Normalization (BN) Freezing. Why freeze BN layers during fine-tuning with a small batch size?**
    - *Answer Focus:* BN layers compute running mean/variance on the current batch. With a very small batch size (e.g., 2 images per GPU, as noted in Section 14), these statistics are unstable. Freezing the BN layers means they retain the stable, general statistics learned from the large ImageNet pre-training, preventing erratic weight updates.

III. Results, Evaluation, and Errors (Sections 15, 17, 20)

11. **Define the APmask metric and explain why it's more crucial than APbox for your project.**
    - *Answer Focus:* APmask averages Average Precision across 10 IoU thresholds (0.50 to 0.95) using the **mask overlap**, not the box overlap. It is the most crucial metric because the project's goal is **instance segmentation** (pixel accuracy), not just detection.
12. **Your quantitative results show a large gap between APS (Small Objects) and APL (Large Objects). What causes this, and how can it be addressed?**
    - *Answer Focus:* Small objects are represented by very few pixels in high-level feature maps. Even with FPN, the effective receptive field for small objects is limited. It can be addressed by increasing the input image resolution, further tuning anchor box scales, or adopting high-resolution detection heads.
13. **What were the three primary failure modes identified in your Error Analysis?**
    - *Answer Focus:* 1) **Localization/Boundary Errors** (low IoU FPs, often at high thresholds). 2) **Inter-Class Confusion** (minimal confusion between similar classes like 'Canned Food' and 'Boxed Goods'). 3) **Dense Packing Errors** (False Negatives where NMS suppresses occluded but valid proposals).
14. **How does NMS (Non-Maximum Suppression) contribute to the 'Dense Packing Errors' you observed?**

- *Answer Focus:* NMS aims to remove redundant bounding boxes. In densely packed areas, if two instances overlap heavily, NMS might incorrectly suppress the proposal for the less confident (often occluded) object, resulting in a **False Negative**.

15. **What did your qualitative visualizations confirm about the model's performance?**
    - *Answer Focus:* They confirmed the high boundary fidelity of the masks (due to RoIAlign) and the model's robustness in handling partial occlusion. They visually validated the quantitative metrics, showing failures primarily with very small/distant objects.

IV. Optimization and Deployment (Sections 21, 22, 23, 24)

16. **Explain the process and benefit of using Mixed Precision Training.**
    - *Answer Focus:* It uses a blend of 16-bit (Float16) and 32-bit (Float32) floating-point numbers. The **benefit** is a significant reduction in memory usage and near **doubling of theoretical inference speed** on modern GPU hardware with minimal loss in accuracy.
17. **You used Dynamic Range Quantization. How does this compare to Full Integer Quantization, and why did you choose Dynamic Range?**
    - *Answer Focus:* **Dynamic Range** quantizes only the weights to 8-bit integers, keeping activations in floating-point. **Full Integer** quantizes both weights and *activations*. Dynamic Range was chosen because it offered a great balance: **≈75% model size reduction with zero loss in accuracy**, whereas Full Integer caused an accuracy drop (APmask fell to 0.355).
18. **Why is Dockerization essential for MLOps and the reproducibility of your model?**
    - *Answer Focus:* Docker wraps the model, code, dependencies (TensorFlow, NumPy, etc.), and environment (OS) into a single, isolated container. This guarantees that the training environment is an **exact replica** of the deployment/testing environment, eliminating "it worked on my machine" issues.
19. **What is ONNX, and why is it crucial for your Edge Deployment Strategy?**
    - *Answer Focus:* **ONNX (Open Neural Network Exchange)** is an open standard format for representing deep learning models. It's crucial because it allows the VisionAI model to be ported seamlessly to various high-performance edge inference engines (like NVIDIA TensorRT or specialized AI accelerators) without being locked into the original TensorFlow framework.
20. **Describe the function of the `/health` check endpoint in your Docker deployment.**
    - *Answer Focus:* It's a basic API endpoint that reports the status of the model loading and the API server itself. Container orchestration platforms (like Kubernetes) rely on this endpoint to confirm that the service is running, ready to serve traffic, and hasn't crashed internally.

V. Application and Project Management (Sections 3, 25, 31, 33)

21. **What was your quantitative target objective for this project, and did you meet it?**
    - *Answer Focus:* The target objective was to achieve a mask Average Precision (APmask) **above 0.35**. The model achieved **0.378**, successfully meeting the project goal. *(Refer to Section 17.)*
22. **What specific value does Instance Segmentation provide over Object Detection in a Healthcare application like Lesion Delineation?**
    - *Answer Focus:* Object detection only gives a bounding box (a rectangle). Instance segmentation provides the **precise pixel-level boundary (mask)**, which is essential for accurate calculation of lesion volume, boundary irregularity, and input for critical clinical procedures like radiotherapy planning.
23. **How does your Streamlit Demo contribute to the project's overall goal?**
    - *Answer Focus:* The Streamlit Demo serves as the user-facing demonstration and visualization layer, proving the end-to-end functionality of the system. It showcases the model's output in real-time and allows users to interactively test inference controls like **Confidence Threshold** and **NMS IoU Threshold**.
24. **In your Future Work section, you mention Swin Transformer. Why consider a Transformer-based architecture over a CNN like ResNet50?**
    - *Answer Focus:* CNNs (like ResNet) excel at local feature extraction. **Swin Transformers** (or Vision Transformers) are better at capturing **global contextual relationships** across the entire image. They often achieve better performance in dense prediction tasks and can potentially improve handling of complex scenes and small objects.
25. **If deploying this model for public surveillance (Security/Surveillance application), what are your two most immediate Ethical/Privacy concerns?**
    - *Answer Focus:* 1) **Data Privacy:** Ensuring all training data is anonymized and that the system does not store identifiable information (faces/licenses) in production logs. 2) **Bias Mitigation:** Ensuring the model performs equally well across diverse lighting conditions, demographics, and environments to prevent unfair performance degradation against certain groups.