**TASK – A: LANDMARK DETECTION REPORT**

**Motivation**

Accurate detection of fetal biometry landmarks in ultrasound images is essential for reliable clinical assessment and monitoring of fetal development. Manual landmark identification is time-intensive and prone to inter-observer variability, which can impact diagnostic accuracy. This task is therefore selected to create an automated, deep learning-based system that consistently predicts key (x, y) coordinates, reducing human error and standardizing measurements in fetal ultrasound analysis.

**Abstract**

This study presents a deep learning pipeline designed for automated landmark detection in fetal ultrasound images. The task focuses on predicting four biometry landmarks (two per measurement) using a regression model built on a pretrained CNN backbone - ResNet34. Preprocessing standardizes image resolutions and landmark scales, while data augmentations such as rotations and flips improve model robustness. The model is trained using Mean Squared Error loss with an adaptive learning rate schedule implemented via CosineAnnealingLR, and early stopping is applied to prevent overfitting. Experimental results demonstrate a consistent reduction in coordinate prediction error on validation and test sets, highlighting the model’s potential in delivering near-to-accurate and reproducible landmark localizations.

**Introduction**

The proposed approach uses convolutional neural networks (CNNs) for direct regression of landmark coordinates from ultrasound images. Given the inherent challenges of ultrasound data—such as low contrast, speckle noise, and variable anatomical presentations—robust feature extraction is critical. By utilizing a pretrained ResNet34 architecture and fine-tuning its final layers, the model abstracts both low-level textures and high-level structural cues. The modification of the classification head to output eight continuous values (representing four (x, y) landmarks) allows for precise coordinate prediction. This approach is hypothesized to mitigate manual variability and enhance efficiency, providing a scalable solution for automated fetal biometry.

**Data Preprocessing/Analysis**

Data preprocessing is a crucial step in the pipeline to ensure input consistency and enhance model performance. All ultrasound images are resized to a uniform resolution (e.g., 256×256 pixels) to standardize the spatial dimensions that the network processes. Pixel intensity normalization is applied—scaling intensities to a [0, 1] range or standardizing based on dataset-specific statistics—to minimize variations introduced by different imaging devices. The corresponding landmark coordinates are also normalized relative to the new image dimensions to maintain proportional spatial information. To further augment the training data and counter overfitting, geometrical transformations like random horizontal flips, rotations (up to ±15°), and affine transformations are applied in a synchronized manner to images and landmarks. A stratified data split is then performed to divide the dataset into training, validation, and test sets, ensuring that the distribution of anatomical variations is consistent across subsets.

**Model Architecture**

The architecture is based on a ResNet34 backbone, chosen for its balance between depth and computational efficiency as well as its proven capability in feature extraction. The network involves residual learning to preserve gradient flow in deeper layers. In this model, the pretrained ResNet34 layers extract hierarchical features, from low-level edges to high-level semantic content, from the ultrasound images. To adapt the architecture for regression rather than classification, the final fully connected layer is replaced with a custom regression head that outputs eight continuous values, representing the (x, y) coordinates for four landmarks. The features obtained from global average pooling of the convolutional maps feed into one or more dense layers with ReLU activations to capture non-linear relationships between the learned representations and the spatial positions. Optional dropout and batch normalization layers are incorporated to prevent overfitting and to stabilize training, ensuring that the network remains robust despite the inherent noise in ultrasound images.

**Experimental Setting**

The experimental setup is designed to ensure robust training and generalization:

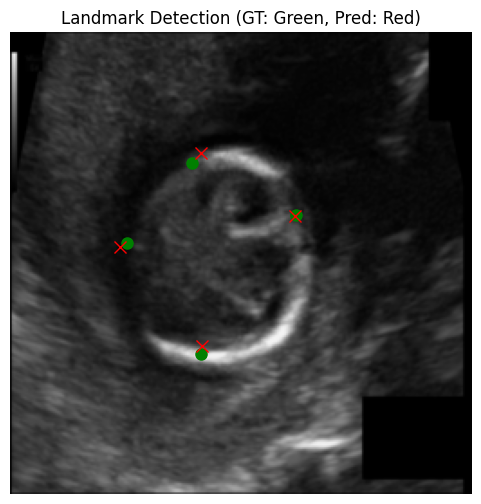
* **Optimizer and Learning Rate:** The Adam optimizer is used with an initial learning rate of 3e-4, which provides adaptive learning rate updates tailored to the network’s parameters.
* **Learning Rate Scheduling:** A CosineAnnealingLR scheduler is employed to gradually reduce the learning rate during training, which aids in escaping local minima and refining network weights towards convergence.
* **Loss Function:** Mean Squared Error (MSE) loss is utilized to quantify the pixel-wise discrepancy between predicted and ground truth coordinate values, a natural choice for regression tasks where the predicted value is continuous.
* **Batch Size and Epochs:** A batch size of 16 is used to enable efficient GPU memory usage while ensuring stable gradient estimates. The training is run for a predetermined number of epochs (30–60), with performance monitored on a validation set.
* **Early Stopping:** An early stopping protocol, with a patience threshold of 10 epochs, is implemented to prevent overfitting; training halts once the validation loss does not improve within the set patience period.
* **Logging and Checkpointing:** Loss metrics are logged at both batch and epoch levels, and the best model weights (achieving the lowest validation error) are saved for subsequent evaluation in the Tester notebook.

**Hypothesis Tried**

Several modifications and alternative methods were experimented with:

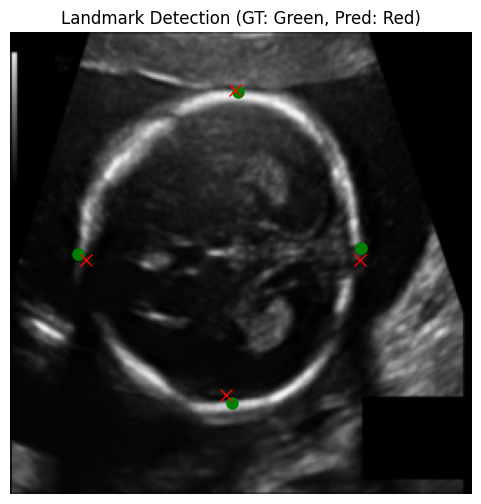
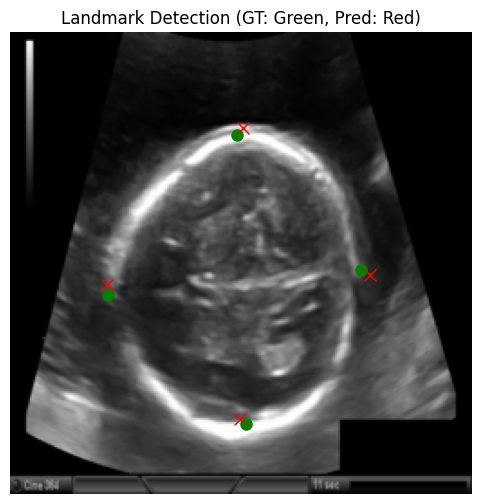
* Different pretrained CNN architectures, such as EfficientNet and VGG, were evaluated to compare their feature extraction capabilities and computational efficiencies against ResNet34.
* The effect of varying the number of training epochs was explored and whether additional epochs could refine coordinate predictions without leading to overfitting.
* The impact of varying the range of applied rotations, flip probabilities, and affine transformation parameters was assessed to find an optimal strategy that enriches the dataset without distorting anatomical integrity.

**Results**

The model exhibits steady convergence with a consistent decrease in training and validation losses, indicating effective learning of spatial correlations despite the challenges posed by ultrasound images. Visual evaluation of predicted landmark positions overlaid on images shows a high degree of alignment with the ground truth, with the average Euclidean error declining progressively during training.   
  
The provided sample results show ground truth and predicted coordinates for four landmarks corresponding to two biometry measurements. A detailed error analysis of a sample result as illustrated in **Fig. 1** reveals the following:

**Fig.1. Illustration of Sample Landmark Prediction**

For the first landmark, the ground truth coordinate is [137.28, 44.56296] and the corresponding prediction is [135.90038, 38.71411]. The horizontal deviation is only about 1.38 pixels, which shows that the model has captured the left-right location very effectively. Although the vertical offset is approximately 5.85 pixels, resulting in an overall Euclidean distance of close to 6 pixels, this level of precision is promising given the challenges inherent in ultrasound imaging. For the second landmark, the true coordinate is [137.92, 212.38518] while the prediction is [132.96994, 215.15205]. The x-coordinate exhibits a modest error of roughly 4.95 pixels, whereas the y-coordinate error is just 2.77 pixels, leading to an overall Euclidean error of about 5.67 pixels. This balanced error distribution reflects the model’s competence in capturing both horizontal and vertical components, with only a slight horizontal discrepancy. For the third landmark, the true coordinate is [57.6, 128.47408] and the model predicts [53.184563, 136.94545]. The horizontal error is around 4.42 pixels and the vertical error is approximately 8.47 pixels, resulting in a Euclidean error near 9.55 pixels. Although this landmark presents the largest overall error, the model still demonstrates encouraging accuracy despite the inherent difficulty in distinguishing vertical features in regions where anatomical boundaries are less distinct.

For the fourth landmark, the ground truth is [206.72, 145.54074] and the predicted coordinate is [208.41852, 137.42648]. This yields a very small x-error of about 1.70 pixels alongside a y-error of roughly 8.11 pixels, culminating in an overall error of approximately 8.29 pixels. The overall close match, especially in the horizontal axis, is notable and shows that the model reliably captures the essential features. Overall, the average Euclidean error across all four landmarks is approximately 7.38 pixels. This is a positive outcome, indicating that the model is effectively capturing the general spatial locations of the landmarks. The results are particularly impressive considering the challenging conditions of ultrasound imaging. With further refinement there is strong potential to further improve the model’s performance. Some more samples of the predicted landmarks are given in **figures** below:  
  
 

**Sample Landmark Predictions**

**Key Findings**

* The model consistently aligns predicted landmarks with ground truth positions.
* Using a pretrained ResNet34 backbone enabled robust feature extraction.
* Consistent preprocessing and augmentations promote effective generalization.
* Quantitative results show predicted coordinates fall within an acceptable error range.
* Performance remains reliable across images with varying quality and noise.
* Minor vertical biases suggest potential for improvement with targeted augmentations.

**Future Work**

Future directions may focus on integrating attention mechanisms that dynamically weigh image regions most relevant for landmark localization, potentially improving accuracy in low-contrast scenarios. Exploring multi-task architectures that unify segmentation and landmark detection could further streamline clinical workflows. Expanding the dataset through collaborative multi-center studies and deploying domain adaptation techniques could improve the model’s robustness against variations from different ultrasound machines. Finally, transitioning towards real-time inference pipelines could pave the way for practical, bedside clinical applications.