Deep Learning – Assignment 2

1 Data Preparation

Preprocessing:

To ensure compatibility with the pre-trained ResNet architecture, all input images were resized to a uniform resolution of 256X256 pixels. Subsequently, pixel intensity values were normalized using standard ImageNet statistics (mean=[0.485, 0.456, 0.406], std=[0.229, 0.224, 0.225]). This standardization ensures that input features share a common scale, thereby stabilizing and accelerating the gradient descent optimization process.

Dataset split:

The dataset, consisting of 3,179 images, was randomly partitioned into three distinct subsets to ensure a robust evaluation:

- **Training Set:** 70% (2,225 images) – utilized for model optimization.
- **Validation Set:** 15% (476 images) – utilized for hyperparameter tuning and learning rate scheduling.
- **Test Set:** 15% (478 images) – reserved exclusively for the final performance assessment to prevent data leakage.

To prevent overfitting, Data Augmentation was applied to the training set. The validation and test sets remained unmodified.

2 Model Design

Learning task:

The objective is Single-Object Localization, formulated as a regression problem. The model is trained to predict the four continuous coordinates (xmin, ymin, xmax, ymax) defining the bounding box of the primary animal in the image.

Loss function:

“Smooth L1 Loss” was selected as the objective function. This function is preferred over Mean Squared Error (MSE) for bounding box regression because it is less sensitive to outliers and prevents exploding gradients when the difference between predicted and ground truth coordinates is large, resulting in a more stable training convergence.

Performance measure(s):

The primary metric for evaluation is the “Intersection over Union” (IoU), averaged over all samples in the test set. Additionally, to provide a binary measure of success, the standard “PASCAL VOC criterion” was adopted. According to this standard, a prediction is considered correct (a "True Positive") if the IoU with the ground truth exceeds 0.50. This threshold is widely accepted in object detection literature as the boundary where the predicted bounding box significantly overlaps with the target object, effectively distinguishing successful localizations from failures.

Model(s) choice:

The ResNet18 architecture, pre-trained on ImageNet, was employed as the feature extractor (backbone). This model offers a balance between feature extraction capability and computational efficiency for the given dataset size. The original classification head was replaced with a regression head consisting of:

1. A Dropout layer (p=0.5) for regularization.
2. A Fully Connected (Linear) layer mapping the 512 input features to the 4 output coordinates.

Hyperparameter selection:

Hyperparameters were tuned based on validation performance. The final configuration includes:

- Batch Size: 32.
- Optimizer: ‘AdamW’ was chosen for its effective decoupling of weight decay from gradient updates.
- Learning Rate: Initialized at 10^-3. The ‘ReduceLROnPlateau’ scheduler was implemented to multiply the learning rate by a factor of 0.1 if the validation loss stalled for 3 consecutive epochs.

Regularization :

To mitigate overfitting on the limited training data, three regularization strategies were implemented:

1. Data Augmentation: During training, ‘RandomHorizontalFlip’ (p=0.5, with coordinate adjustment) and ColorJitter were applied to artificially increase data diversity.
2. Weight Decay: A coefficient of 10^-4 was set in the optimizer.
3. Dropout: Integrated before the final output layer to prevent neuron co-adaptation.

Baseline choice :

The performance is evaluated against a standard detection threshold of IoU > 0.50, commonly used in challenges such as PASCAL VOC. A qualitative baseline is established by visual inspection of the overlap correctness, ensuring the model performs significantly better than a random or mean-box predictor.

3 Results

Training curves:

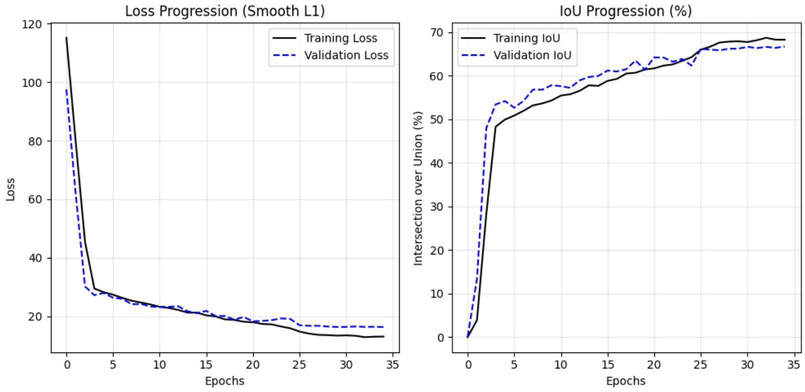

Figure 1, Training and validation curves for Smooth L1 Loss (left) and Intersection over Union (right).

The training and validation loss curves are presented in Figure 1. These curves demonstrate stable convergence throughout the training process. The model learns rapidly during the initial 5 epochs. Towards the end of the training (around epoch 24 and epoch 34), the learning rate scheduler reduced the learning rate. While this did not result in a sharp visual drop in the already converged loss curve, it facilitated the stabilization of weights and fine-tuning of the model, contributing to the final peak performance on the Test set. The validation IoU curve closely tracks the training IoU curve, indicating that the applied regularization techniques successfully maintained generalization capabilities and kept overfitting under control.

Inference:

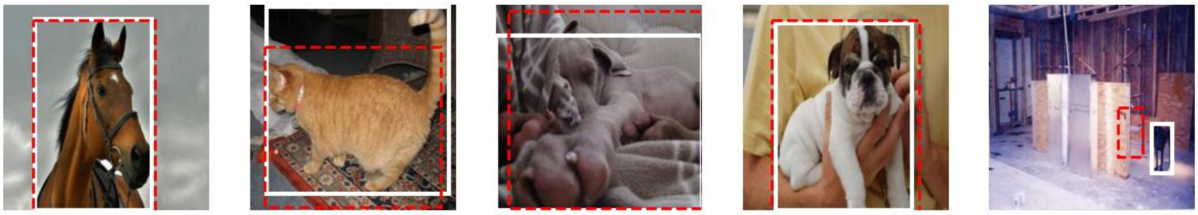

Figure 2, Inference samples from the test set.

Qualitative results on the unseen test set are displayed in Figure 2. The model demonstrates high precision in localizing animals (indicated by red dashed boxes) compared to the ground truth (solid white boxes). As observed in the figure, the model exhibits robustness against cluttered backgrounds and varying object scales. However, minor inaccuracies can be observed in cases where the animal is heavily occluded or situated at the image boundary.

Performance:

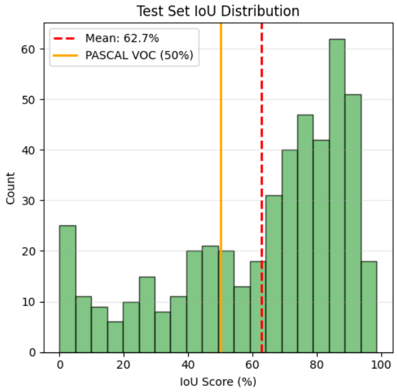

Figure 3, Distribution of Intersection over Union (IoU) scores on the Test Set.

The quantitative assessment of the model is detailed in Figure 3. The histogram illustrates the distribution of IoU scores on the unseen Test Set. The model achieved a Final Mean IoU of 62.7% (indicated by the red dashed line). Furthermore, the Accuracy (>0.5 IoU) is 71.13%, meaning approximately three out of four predictions are considered successful detections according to PASCAL VOC standards.

As observed in the histogram, the distribution is skewed towards higher IoU values, with a significant cluster of predictions in the 0.7–0.9 range. This indicates that when the model detects an object, it generally achieves a high degree of overlap, rather than just barely passing the 0.5 threshold.

4. Implementation details

Loss computation:

The training loss displayed in the curves is computed during the epoch by averaging the losses of individual batches as they are processed. The validation loss is calculated after the epoch ends by evaluating the model on the entire validation set without gradient updates.