Assignment 5: Fit a Deep Learning Model to a HuggingFace Dataset

Dataset Description and Preprocessing

The dataset used for this project is the **Iris dataset** from the Hugging Face Datasets Library. This dataset consists of 150 samples, each with four numerical features representing sepal and petal length/width. The task is to classify these samples into three species: *Iris-setosa, Iris-versicolor,* and *Iris-virginica*.

Reason for Selection: The Iris dataset was chosen because it is a small and easy-to-use dataset, making it ideal for quickly implementing and testing deep learning models.

Preprocessing Steps:

- The dataset was converted into a Pandas DataFrame.
- The categorical species labels were encoded as numerical values (0, 1, 2).
- The dataset was split into an 80/20 training/testing split.
- Features were converted into tensors to be compatible with PyTorch.

Model Architecture

A fully connected deep neural network (DNN) was implemented using PyTorch. The architecture consists of:

- **Input Layer:** 4 neurons (one for each feature)
- **Hidden Layers:** Two layers with 8 neurons each, using ReLU activation
- **Output Layer:** 3 neurons (one for each class)

The model was optimized using the Adam optimizer with a learning rate of 0.01 and trained for 100 epochs using the cross-entropy loss function.

Justification for Hyperparameter Choices:

- **Learning Rate (0.01):** A moderately high learning rate was chosen to ensure faster convergence while avoiding instability.
- **Optimizer (Adam):** Adam was selected as it adapts learning rates dynamically and works well with minimal hyperparameter tuning.
- **Loss Function (Cross-Entropy):** This is the standard loss function for multi-class classification problems.
- **Epochs (100):** A sufficient number of epochs were used to allow the model to learn effectively without excessive overfitting.
- **Hidden Layer Size (8 neurons per layer):** This was chosen as a balance between computational efficiency and learning capacity given the dataset size.

Key Findings from Evaluation Metrics and Visualizations

The model achieved an **accuracy of 100%** on the test set, as seen in the classification report and confusion matrix. Each class had a precision, recall, and F1-score of **1.00**, indicating perfect classification.

- **Confusion Matrix:** No misclassifications were observed.
- **Loss Curve:** The loss function steadily decreased over the epochs, confirming effective training.

Insights and Interpretation

The results indicate that the model is highly effective at classifying Iris species. However, the perfect accuracy may suggest overfitting due to the small dataset size. Potential improvements could include regularization techniques or experimenting with a different train/test split. This project successfully demonstrates how to train and evaluate a deep learning model using PyTorch with the Hugging Face Datasets Library.