UE22AM351B (JAN - MAY 2025) - Deep Learning **Given on: March 11, 2025**
LAB #8 (Orange Problem) **Due on: March 16, 2025**

In this assignment, your task is to perform machine translation from English to your mother tongue. The dataset is available here. For example, the file *kan.txt* consists of English-Kannada sentence pairs. The models to be used are

1. Encoder decoder model without attention
2. Encoder decoder model with attention

Follow the below instructions.

1.1 Task

Step 0: Report the language into which you are translating the English sentences.

Step 1: Data Preprocessing

- **Text Cleaning:** Remove punctuations.
- **Tokenization:** Split sentences into words.
- **Text to Sequence Conversion:** Convert to a sequence of integers.
- **Padding Sequences:** Ensure uniform input size by padding sequences to the same length.

Step 2: Preparing the Dataset

- **Splitting Data:** Divide data into train, validation, and test datasets.

Step 3: Build the encoder-decoder Model

- **Build Encoder LSTM**
- **Build Decoder LSTM**
  - Without attention
  - With attention as described in class (which is called Bahdanau Attention or Additive Attention)

Step 4: Model Training

- **Set Hyperparameters**
- **Train the two models**
- **Monitor Performance**

Step 5: Evaluation of the two models

- **Use test data:** Evaluate the two models using the test data.
- **Visualisation:** Plot loss curves for the two models.
- **Prediction:** For the two models, tabulate actual and predicted sentences for any 10 sentences selected randomly from the test dataset

- **Report your observations/inferences made from the two models**

Step 6: Visualize Attention Weights

- Select a test sentence and generate a translation using the model trained with attention.
- Extract and plot attention weights as a heatmap.
- Label the heatmap with input and translated words to analyze attention alignment.

1.2 Submission

- Upload a .zip file (or .rar or .tar) in the shared link. The zipped file should contain
    1. report.pdf
    2. train.py: the code
    3. Any other Python scripts that you have written

  The name of the zipped file for submitting your assignment should be as follows: pes1ug19am139_8.zip, where pes1ug19am139 is your roll number and 8 implies that you are submitting your eighth lab assignment.
- Strictly adhere to the submission guidelines. If not, your submissions will not be graded.
- Any copying on assignments will result in a zero on the assignment. We will be using JPlag (a plagiarism tool) to detect similarities among multiple sets of source code files/reports.