# Problem 4: (English to French language translator)

**An implementation of English to French language translation using Long short-term memory (LSTM) architecture is demonstrated as a Google Colab notebook.**

<https://colab.research.google.com/drive/1QXtXjrEw0C0TSPgBh88UwxJl807d_RyD?usp=sharing>

Dataset: <https://drive.google.com/file/d/1_V6eSyHdwMIG5cJxXhK6GDqeYajVa9Gs/view?usp=sharing>

****

Figure: Flow diagram of a machine learning project

**Explain the design principles you considered**

I have drafted a Figure (shown above) to depict the steps/processes carried out to train a machine learning model. A complete production life cycle includes data curation, refinement, training, hyperparameter tuning, error analysis, validation, and application/service deployment (scalability, performance, and maintenance). Automatic differentiation frameworks such as Tensorflow, PyTorch and FastAi are used for model development.

1. Data pipeline: Machine learning model is highly dependent on quality of data provided for the training. Data pipeline involve tasks such as data collection, pre-processing, feature representation, bias removal and privacy.
2. Modelling and training: The nature of problem drives the decision of model selection i.e., either it is supervised, unsupervised or a reinforcement learning problem. For the present case, recurrent neural network (RNN) based architectures such as LSTM, Bi-LSTM, GRU, LSTM with attention, and Transformers can be a choice. Training a model with different hyperparameters settings results differently in terms of accuracy.
3. Deployment: of machine learning model is becoming an important concern as different devices/platforms are becoming clients of the model. Low powered IoT devices need machine learning model for real time prediction and classification. High end biomedical applications can handle large models. Mobile devices communicate with model through REST services.

**How the model can be trained, tested, and evaluated**

Training is the process of constructing an internal model that can perform prediction on the data. The model is initially fit on the training dataset and then tested with the test set. Following procedure is carried out during the language translation:

Training: Provided data in the form of CSV file is loaded and split into two parts. A training set and a test set. Several hyperparameters such as batch size, number of epochs, activation function, learning rate, momentum, regularization, loss function and optimizers have substantial effect on the training results. Data imbalance also effects the performance of the training process. The size of machine learning model (either shallow or deep) has performance related pros and cons.

Testing: There are several validation methods for testing the performance of trained model such as cross-validation and hold-out validation. In case of hold-out validation, the original data is split into training and test parts. Model is fitted on training data and the tested with the test data. Whereas, for cross-validation, the data is split into k-random sets and among that one set is used for testing while others are used for the training. This procedure is repeated until all parts are used as test set. Normally, the value is taken 5 or 10. Hold-out is suitable in conditions when a large dataset is available for training.

Evaluation: There are several performance measures for the evaluation of a machine learning model. These performance measures include loss, validation loss, receiving operating characteristics (ROC) analysis, confusion matrix, bilingual evaluation understudy (BLEU), error rate analysis, precision, recall, F-measure and ROUGE.

**How you would explain how the neural network works to the researcher.**

Neural network or artificial neural networks (ANN) is an algorithmic approach in machine learning that can model human like behaviour in recognizing patterns.

|  |  |  |
| --- | --- | --- |
|  |  |  |
| (a) | (b) | (c) |

Figure: (a) Traditional computing (b & c) machine learning

Neural networks work for researchers as it learns by itself a function or a program to perform a certain task against the provided data. Researchers do not need to write program themselves rather they just need to provide data and target output to the machine learning algorithm (as shown in the Figure above). Later, a trained machine learning model can predict output for an unseen data instance. In case of un-supervised learning when output is not provided with the data, the machine learning model will identify patterns to draw a conclusion.