ELEC6211 Project Preparation – Project Plan

Title	Efficient Training of Deep Neural Networks – Machine Translation
Student name:	Richa Ranjan
Supervisor name:	Mahesan Niranjan

Aims/research question and Objectives

The aim of this project is to showcase efficient training techniques with respect to Deep Neural Networks. More specifically, a Machine Translation task is chosen to explore possibilities in this regard. The idea for this research is inspired by the growing need of language translators. Several European languages have come into picture with global business expansions. Also, with social media gaining popularity, communication across cultures have become common. Thus, the need to have a common mode of communication is increasing, and hence Machine Translation, as an application has recently gained immense popularity. I, being an Indian, and my mother tongue being *Hindi*, I am keen on using Neural Networks to build Hindi-English translators and vice-versa. The idea is to help ease out communications for a large section of the Indian population, spread across different countries. With the hope of using Deep Neural Networks to build a translator at par with the Google translator (or even better, if possible), the main objectives of the project are summarized below:

1. To showcase techniques which are efficient in training a Deep Neural Network.
2. To implement the Sequence to Sequence approach recommended by Google.
3. To showcase how Recurrent Neural Networks(RNNs) and Long Short-Term Memory (LSTMs) can be effective in training a Deep Network with huge datasets.
4. To evaluate the Sequence to Sequence and Recurrent Neural Network results against previously applied techniques for the same application.
5. To discover additional feasible approaches (if any) that can be an advancement to the baseline model i.e. the Sequence to Sequence implementation.

One of the important aspects of this research would be to compare the state-of-the-art results and to showcase how optimized implementation of learning algorithms shows improvement in results. The baseline model will be an end-to-end LSTM which translates Hindi to English language. As this work deals in Machine Translation, accuracies from previously built models will be tested on the same data set and a comparison of techniques can be made.

Summary of proposed research and analysis methodology

This project will use an empirical process of training a Deep Neural Network to obtain maximum efficiency with the said combination of inputs and parameters. First step would be to build a state-of-the-art Deep Network that performs Machine Translation task on the Hindi and English languages. Going ahead, a range of gradient descent optimization algorithms such as *Adam*, Gradient Descent with Momentum, *RMSProp* can be used to optimize the results and observe accuracies to find the closest possible translation. Additionally, Hyperparameters like number of layers, epochs etc. can be tuned and tested how different values affect results.

While the above methods are empirical ways of tuning and obtaining results, more specific approaches can be thought of in this light. One of the most commonly used approaches for text data is the Recurrent Neural Network (RNN). The Long Short-Term Memory(LSTM) approach has provided significant results in Sequence to Sequence and Encoder-Decoder approaches by Google. This idea of this research is to build an LSTM-based model and to execute the same on the vectorized language data set. Two separate LSTMs will be used, one for input sequence, and the other for output sequence. Next, the source sentences will be reversed in the input order, as previous researches have shown better communication between input and output when the sentences are reversed. This will be the baseline model.

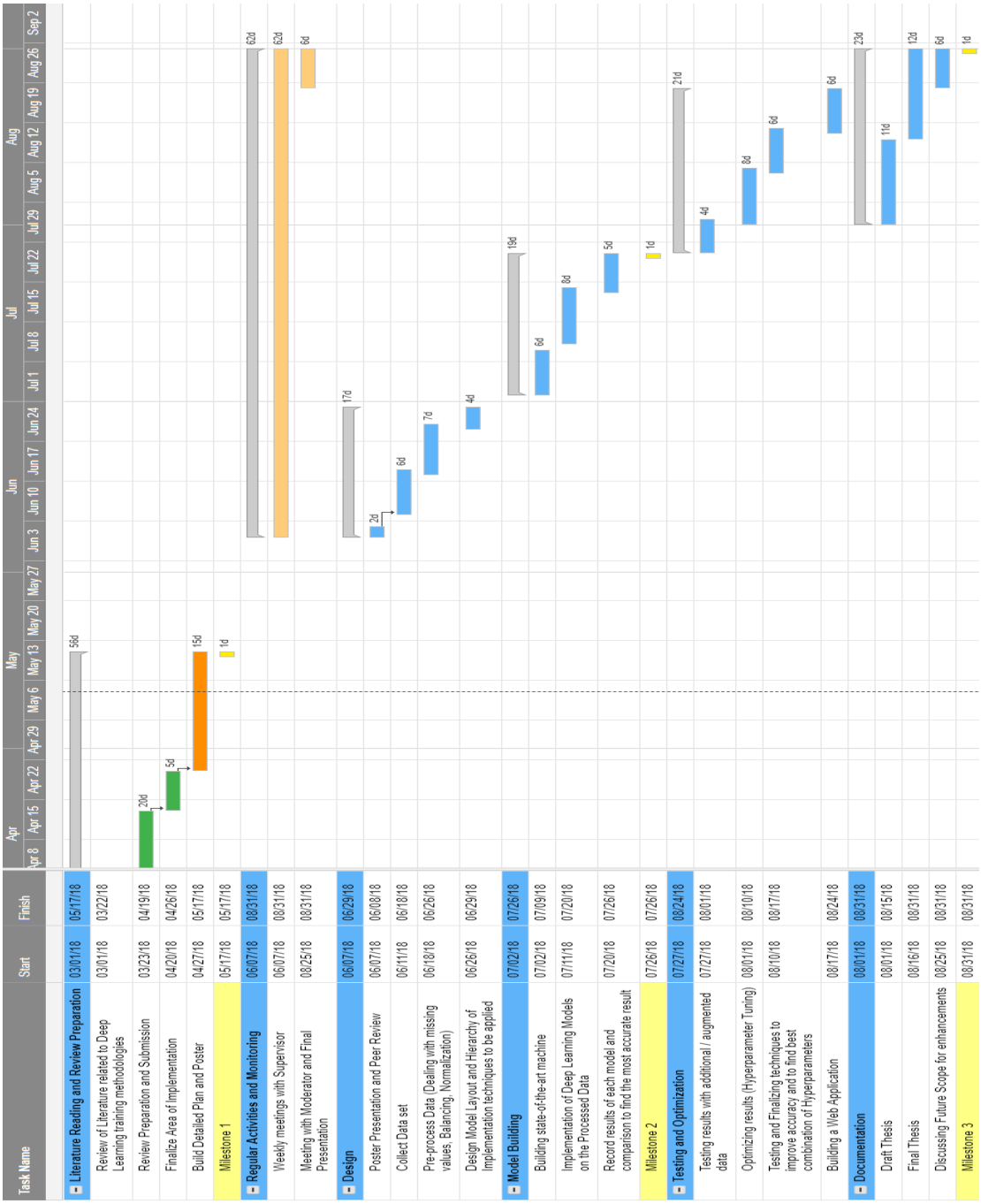
The Data will be obtained from the '*The IIT Bombay English-Hindi Parallel Corpus'* for the baseline implementation. Later, if needed, this data can be augmented to extend the data set size. Going further, depending upon the accuracy of the model, the same can be applied on some of the minority languages to test if the translation accuracy stands at par with the previous language translations. Initially, a low confidence model will be built using a part of the dataset, which will later be exposed to human translation. If better performance is achieved quicker than the state-of-the-art model, the same process will be continued. This is the idea of Active Learning. Implementing the active learning process on LSTMs would have interesting insights, and hopefully, better translation accuracies.

On this model, further enhancements like parallelization and phrase-based approach can be implemented. Once the baseline accuracy is available, there can be multiple methods to tune the parameters and to improve the accuracies. Also, a web application is a part of the project plan. This application would provide an interface for users to enter their sentences/phrases/words in Hindi, and to get the equivalent English translation.

The final thesis is going to be a documented version of the research, which will be followed by a final presentation with the Moderator. The ongoing activity would be to regularly meet with the supervisor and to perform regular monitoring on the progress.

ELEC6211 Project Preparation – Project Plan

Research plan – Gantt chart or Pert chart



Task Name	Start	Finish
Literature Reading and Review Preparation	03/01/18	05/17/18
Review of Literature related to Deep Learning training methodologies	03/01/18	03/22/18
Review Preparation and Submission	03/23/18	04/19/18
Finalize Area of Implementation	04/20/18	04/26/18
Build Detailed Plan and Poster	04/27/18	05/17/18
Milestone 1	05/17/18	05/17/18
Regular Activities and Monitoring	06/07/18	08/31/18
Weekly meetings with Supervisor	06/07/18	08/31/18
Meeting with Moderator and Final Presentation	08/25/18	08/31/18
Design	06/07/18	06/29/18
Poster Presentation and Peer Review	06/07/18	06/08/18
Collect Data set	06/11/18	06/18/18
Pre-process Data (Dealing with missing values, Balancing, Normalization)	06/18/18	06/26/18
Design Model Layout and Hierarchy of Implementation techniques to be applied	06/26/18	06/29/18
Model Building	07/02/18	07/26/18
Building state-of-the-art machine	07/02/18	07/09/18
Implementation of Deep Learning Models on the Processed Data	07/11/18	07/20/18
Record results of each model and comparison to find the most accurate result	07/20/18	07/26/18
Milestone 2	07/26/18	07/26/18
Testing and Optimization	07/27/18	08/24/18
Testing results with additional / augmented data	07/27/18	08/01/18
Optimizing results (Hyperparameter Tuning)	08/01/18	08/10/18
Testing and Finalizing techniques to improve accuracy and to find best combination of Hyperparameters	08/10/18	08/17/18
Building a Web Application	08/17/18	08/24/18
Documentation	08/01/18	08/31/18
Draft Thesis	08/01/18	08/15/18
Final Thesis	08/16/18	08/31/18
Discussing Future Scope for enhancements	08/25/18	08/31/18
Milestone 3	08/31/18	08/31/18

Ethical statement

The main ethical concern in any project is privacy and integrity of the data and work. I, as a student of professional studies, would like to declare that the data used for my research will be obtained and maintained securely, with no malicious intentions, and I will not engage myself in any business involving data exposure to third parties, or making the project activities public. If the data contains Personally Identifiable Information (PII) or any other similar sensitive data, appropriate security measures will be applied to ensure that the application does not accidently leaks them. I would ensure to get my progress checked weekly by the Supervisor, so that any unintentional ethical concerns (if any) will be monitored regularly.

Also, this project aims at providing Machine Translation solutions on some widely spoken as well as less spoken languages. Thus, Inclusion/Exclusion bias is an ethical concern. I would ensure to the best of my capability, to keep a balance between languages used.

Additionally, if this project is used for future applications such as, interpretation of social media conversations, automated translations can have unintentional errors that can steer or influence unpleasant conversations. It is thus necessary to understand and appreciate the accuracy rate involved in translation tasks. I would therefore, set a more realistic expectation for future use, such as providing the accuracies of the model as part of the application specifications.

Legal and commercial aspects

Commercial: Machine Translation has been serving a wide range of applications for over half a century. In the last decade, Deep Neural Networks have shown significant improvements in accuracy and fluency of Machine-based Translations. Although there are several Machine Translators available currently, there is always a dearth for better accuracy models. The product of this research will consequently have great demands in the market of Natural Language Processing. In addition to the likes of Google Neural Machine Translation (GNMT), the model will be a useful advancement in applications like Social Media Posts across the globe. Customized corpus can also be built for further references in applications like Language Trainings, or even Business communication requirements. Hindi being a language spoken by a large section of the Indian population, translators are needed even for learning processes, for example, if a Hindi-speaking native Indian wishes to learn English, the global language, this application would be a good start. If the model is able to show better accuracies than Google translators, it will have market-wide demands.

Legal: To avoid copyright and intellectual property infringement, any additional data created, or novel algorithmic approaches should be appropriately trademarked. This would avoid further misuse of both data and model.