Neural Machine Translation

> **Deep Learning Project Phase-2.1**

> 200968108 DSE-A 27 Batch-1

Problem Statement :

The ability to communicate with one another is a fundamental part of our daily life. There are nearly 7,000 different languages worldwide. As our world becomes increasingly connected, language translation provides a critical cultural and economic bridge between people from different countries and ethnic groups. Some of the more obvious use-cases include:

- **Business**: international trade, investment, contracts, finance
- **commerce**: travel, purchase of foreign goods and services, customer support
- **media**: accessing information via search, sharing information via social networks, localization of content and advertising
- **education**: sharing of ideas, collaboration, translation of research papers
- **government**: foreign relations, negotiation

To meet these needs, technology companies are investing heavily in machine translation. This investment and recent advancements in deep learning have yielded major improvements in translation quality.

According to Google, switching to deep learning produced a 60% increase in translation accuracy compared to the phrase-based approach previously used in Google Translate. Today, Google and Microsoft can translate over 100 different languages and are approaching human-level accuracy for many of them.

However, while machine translation has made lots of progress, it's still not perfect. 😬

Dataset :

The dataset used in this project for the task of converting/translating one language to another language is taken from the below website.

Tab-delimited Bilingual Sentence Pairs from the Tatoeba Project (Good for Anki and Similar Flashcard Applications)

If you don't already use Anki, vist the website at http://ankisrs.net/ to download this free application for Macintosh, Windows or Linux. Any flashcard program that can import tab-delimited text files, such as Anki (free) can use these files. Warning! There are errors in the Tatoeba Corpus.

http://www.manythings.org/anki/

We downloaded an English-German dataset that consists of bilingual sentence pairs from the Tatoeba Project. In this project, text in ***English*** language is being translated into ***German*** Language.

Metadata of the Dataset :

Each line in the dataset is a tab-delimited text consisting of an English text sequence, the translated German text sequence and some information regarding attribution. Note that each text sequence can be just one sentence, or a paragraph of multiple sentences. In this machine translation problem where English is translated into German, English is called the *source language* and German is called the *target language.*

Exploratory Analysis :

The text data available at the above mentioned website is present in the raw form(consists of punctuations, symbols, letters in both lower & upper case). So, pre-processing has been performed on the downloaded dataset and has been made ready! to be used for implementing various Deep Learning models and architectures.

There are 2 columns. One column has English words/sentences and the other one has German words/sentences. And this dataset can now be used for language translation task.

Google Colaboratory

https://colab.research.google.com/drive/1fAlkxthovv399UqXnTu3l9AbupVFb_IH?usp=sharing

Preprocessing Pipeline :

Below are a few pre-processing steps implemented on the downloaded dataset:

▼ 1. Load & examine the data.

```
Go.     Geh.    CC-BY 2.0 (France) Attribution: tatoeba.org #2877272 (CM) & #8597805 (Roujin)
Hi.     Hallo!  CC-BY 2.0 (France) Attribution: tatoeba.org #538123 (CM) & #380701 (cburgmer)
```

2. Cleaning the data. (includes the following):

▼ Converting the data into an array for easy implementation.

```
array([['Go.', 'Geh.',
        'CC-BY 2.0 (France) Attribution: tatoeba.org #2877272 (CM) & #8597805 (Roujin)'],
       ['Hi.', 'Hallo!',
        'CC-BY 2.0 (France) Attribution: tatoeba.org #538123 (CM) & #380701 (cburgmer)'],
       ['Hi.', 'Grüß Gott!',
        'CC-BY 2.0 (France) Attribution: tatoeba.org #538123 (CM) & #659813 (Esperantostern)'],
       ...,
       ["They're coming again.", 'Sie kommen wieder.',
        'CC-BY 2.0 (France) Attribution: tatoeba.org #2243605 (CK) & #6643044 (Felixjp)'],
       ["They're coming again.", 'Die kommen wieder.',
        'CC-BY 2.0 (France) Attribution: tatoeba.org #2243605 (CK) & #6645385 (Felixjp)'],
       ["They're doing it now.", 'Sie tun es jetzt.',
        'CC-BY 2.0 (France) Attribution: tatoeba.org #3740220 (CK) & #3815719 (nGerman)']],
      dtype='<U537')
```

▼ Reducing the size of dataset to save the computation cost.(Only in this case)

```
print("Raw data has",deu_eng.shape[0],"examples.")
```

```
Raw data has 255817 examples.
```

```
deu_eng = deu_eng[:50000,:]
deu_eng
```

▼ Removing irrelevant text like attribution details

```
deu_eng1 = np.delete(deu_eng, 2 ,1)
deu_eng1
```

▼ Splitting each sample/text into English-German pairs.

```
array([['Go.', 'Geh.'],
       ['Hi.', 'Hallo!'],
       ['Hi.', 'Grüß Gott!'],
       ...,
       ["They're coming again.", 'Sie kommen wieder.'],
       ["They're coming again.", 'Die kommen wieder.'],
       ["They're doing it now.", 'Sie tun es jetzt.']], dtype='<U537')
```

▼ Removing punctuations.

```
array([['Go', 'Geh'],
       ['Hi', 'Hallo'],
       ['Hi', 'Grüß Gott'],
       ...,
       ['Theyre coming again', 'Sie kommen wieder'],
       ['Theyre coming again', 'Die kommen wieder'],
       ['Theyre doing it now', 'Sie tun es jetzt']], dtype='<U537')
```

▼ Converting the text to lower case.

```
array([['go', 'geh'],
       ['hi', 'hallo'],
       ['hi', 'grüß gott'],
       ...,
       ['theyre coming again', 'sie kommen wieder'],
       ['theyre coming again', 'die kommen wieder'],
       ['theyre doing it now', 'sie tun es jetzt']], dtype='<U537')
```

3. Tokenizing & vectorizing the text into numerical sequences.
4. Padding those sequences with 0's to bring them to same length.

Objectives :

- *The goal is to have machines translate content well enough for human translators to understand its meaning and easily improve upon the text.*
- convey the original tone and intent of a message, taking into account cultural and regional differences between source and target languages

---

Literature Review

Title of the paper	Authors	Conference published in	Date of Publication	Models/Algorithms Used	Architecture	Pros & Cons
Machine Translation using Deep Learning: An Overview.	*By Shashi Pal Singh, Ajai Kumar, Hemant Darbari*	2017 International Conference on Computer, Communications and Electronics (Comptelix)	01-02 July 2017	R2NN(Recursive & Recurrent NN), LSTM, RAE, BLEU	Encoder - Decoder	Difficult to train RNN due to long dependencies but LSTM avoid the problems occurred with RNN. Uses back Propagation with time algo. to learn model parameters.
Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation	*Kyunghyun Cho , Bart van Merrienboer , Caglar Gulcehre , Dzmitry Bahdanau , Fethi Bougares , Holger Schwenk , Yoshua Bengio*	arXiv:1406.1078	3 Sep 2014	RNN, - CGM(Convolutional n-gram Model), CSLM(Continuous Space Language Model)	RNN Encoder–Decoder	The architecture used is able to capture linguistic irregularities in phrase-pairs well and propose well-formed target phrases. It has large potential for further improvement and analysis. One approach that was not investigated here is to replace the whole, or a part of the phrase table by letting the RNN Encoder–Decoder propose target phrases.
Fast and Robust Neural Network Joint Models for Statistical Machine Translation	*Jacob Devlin, Rabih Zbib, Zhongqiang Huang, Thomas Lamar, Richard Schwartz, and John Makhoul*	Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics	June 2014	NNJM(Neural Network Joint Model), NNLMT, BLEU		One issue with the S2T NNJM is that the probability is computed over every target word, so it does not explicitly model NULL-aligned source words. In order to assign a probability to every source word during decoding, we also train a neural network lexical translation model (NNLMT).

Title of the paper	Authors	Conference published in	Date of Publication	Models/Algorithms Used	Architecture	Pros & Cons
Edinburgh Neural Machine Translation Systems for WMT 16	*Rico Sennrich, Barry Haddow, Alexandra Birch*	WMT 2016	2016	BPE(Byte Pair Encoding)-originally devised as a compression algorithm	Attentional Encoder-Decoder	For English↔German and English→Czech, we trained a right-to-left model with reversed target side, and we found reranking the system output with these reversed models helpful.
A Convolutional Encoder Model for Neural Machine Translation	*Jonas Gehring, Michael Auli, David Grangier, Yann N. Dauphin*	Association for Computational Linguistics(ACL) 2017	2017	BiLSTM, GRU, Bi Directional RNN, LSTM, Sigmoid & tanh activations in CNN, BLEU	Based on deep convolutional encoders by using residual connections	Single-layer CNN model can outperform a Uni-directional LSTM encoder & BiLSTM. Our architecture leads Large gen. speed improvements & can translate twice as fast as baselines with BiRNN encoders.
Convolutional Sequence to Sequence Learning	*Jonas Gehring Michael Auli David Grangier Denis Yarats Yann N. Dauphin*	ICML 2017	2017	RNN, CNN, Position Embeddings	Multi-step attention mechanism	Compared to recurrent networks, our convolutional approach allows to discover compositional structure in the sequences more easily since representations are built hierarchically. Our model relies on gating and performs multiple attention steps.
A Study of Machine Translation Methods	*Dr. John T. Abraham, Bijimol T.K*	NCILC 2014	Feb 2014	SMT, Corpus based MT, Dictionary based MT, RBMT, HMT, Example based MT(Memory based translation)	Study on various MT methods	Each of these has its own advantages and limitations as explained in this paper. It's a proven fact that no two translation system can produce identical translations of same text in the same language pair.

Title of the paper	Authors	Conference published in	Date of Publication	Models/Algorithms Used	Architecture	Pros & Cons
On the Properties of Neural Machine Translation: Encoder–Decoder	*Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, Yoshua Bengio*	SSST-8	25 Oct 2014	2 NMT models analyzed in this paper: RNN Encoder-Decoder and the other has the encoder replaced with **grConv**.	A new approach to SMT referred to as Neural MT, inspired by Deep Representational Learning.	Clearly 4We used Moses as a baseline, trained with additional the phrase-based SMT system still shows the superior performance over the proposed purely neural machine translation system, but we can see that under certain conditions (no unknown words in both source and reference sentences), the difference diminishes quite significantly. This analysis suggests that that the current neural translation approach has its weakness in handling long sentences.
A Transformer-Based Neural Machine Translation Model for Arabic Dialects That Utilizes Subword Units	*Laith H. Baniata, Isaac. K. E. Ampomah and Seyoung Park*	Sensors 2021	29 Sep 2021	Encoder-Decoder networks, Multi-Head Attention(MHA)	Transformer Based-NM	Due to the lack of standardization for the Arabic vernaculars, Conventional NMT methods for Arabic dialects are incapable of translating parts of input source sentences.

Title of the paper	Authors	Conference published in	Date of Publication	Models/Algorithms Used	Architecture	Pros & Cons
Effective Approaches to Attention-based Neural Machine Translation	*Minh-Thang Luong, Hieu Pham, Christopher D. Manning*	Conference on EMNLP 2015 held in Lisbon, Portugal.	17 Aug 2015	LSTM, GRU, RNN, BLEU	Our Attention based models classified into 2 categories: global & local.	The global approach which always looks at all source positions and the local one that only attends to a subset of source positions at a time. . For the English to German translation direction, our ensemble model has established new state-of-the-art results for both WMT'14 and WMT'15, outperforming existing best systems, backed by NMT models and n-gram LM ranks, by more than 1.0 BLEU. Our analysis shows that attention-based NMT models are superior to nonattentional ones in many cases, for example in translating names and handling long sentences.
Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation	*Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi*		26 Sep 2016	Wordpiece model, Using Deep LSTM with Residual Connections	GNMT Architecture	**GMNT reduces translation errors by more than 60%** compared to the **PBMT** (Phrase based Statistical Machine Translation) model.

Shortlist models for implementation:

From the above literature review, we can try and implement a few models that would give us good accuracy and efficiency along with good performance. We can also try implementing using the LSTM, BiRNN and Transformers architecture and compare the performances.

Baseline-model:

The baseline model in the project is built using the **Recurrent Neural Networks** as they are designed **to take sequences of text as inputs or return sequences of text as outputs, or both,** which also syncs with our requirement and the objective i.e.., **Sequence to Sequence** conversion.

.