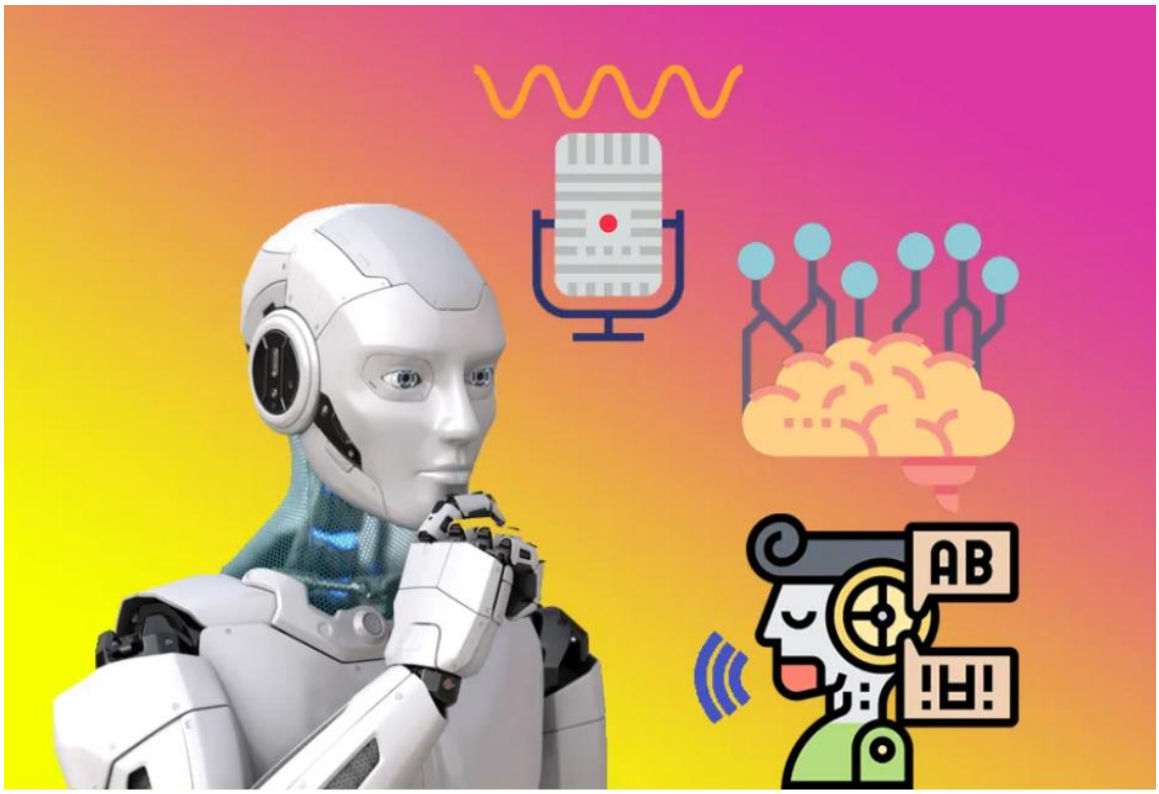
Course Project

Natural Language understanding



Submitted By:
Sasikanth Koti | MT19AIE308
Nikhila Dhulipalla | MT19AIE270
Adhun Thalekkara | MT19AIE205

Machine Translation: (that translates sentences from one language to another)

As part of this project English to Hindi language translation is implemented and evaluated using deep learning architectures.

Evaluation is carried out on these trained models using BLEU scores. The model is also evaluated quantitatively by utilizing some sample translations.

Along with this translation is also carried on the reviews provided in the given Kaggle set for additional task.

Implementation Details:

For all the questions, pytorch and hugging face frameworks are used.

The training is carried out using Colab environment.

Datasets :

https://www.kaggle.com/snap/amazon-fine-food-reviews

English to Hindi corpus from https://indicnlp.ai4bharat.org/samanantar/ is used for training the models.

Code :

https://colab.research.google.com/drive/1nRGGEOUul7w18z0u03LLNXAeDC-h6UmZ?usp=sharing

https://colab.research.google.com/drive/1d8nG9X1MQMaM_C60OQvAJV_1BfTtboNT?usp=sharing

https://colab.research.google.com/drive/1ILuonoqdSN1TwE0C-Vvf10RGME1XnFp-?usp=sharing

As part of the project, two architectures were trained and evaluated using different methodologies.

Part 1:

In this part, a pre-trained transformer model is considered for fine-tuning and it is implemented using Hugging face framework. "*Helsinki-NLP/opus-mt-en-hi*" is the pre-trained model considered for fine-tuning. This is a transformer-align family trained using *MarianMT* framework. This transformer encoder-decoder consists of 6 layers in each component and the pre-trained model is trained originally with OPUS dataset , where source language is English and target language is Hindi.

For the pre-training of the model, about 50,000 sentence pairs are considered as the total data set. The dataset is split into 0.9 train data and 0.1 test data. So, the train data consists of 45000 sentence pairs and test set consists of 5000 sentence pairs. It is also to be noted that these sentence pairs are selected such that the length is between 2 and 10 words.

The dataset is tokenized using sentence piece tokenizer available from this pre-trained model and then further pre-processed to embed the label information. This dataset is then utilized to fine-tune the model. Fine-tuning is carried out with learning rate of 2e-5, weight decay of 0.01 and batch size of 64 for 15 epochs.

Fine-tuning progress is shown below :

Epoch	Training Loss	Validation Loss	Bleu	Gen Len
1	3.211800	2.853307	14.991700	9.581600
2	2.883300	2.731435	15.693600	9.531200
3	2.544300	2.668295	16.151300	9.369200
4	2.431900	2.628693	16.823000	9.488400
5	2.264000	2.605112	16.919500	9.430600
6	2.161800	2.590049	17.274300	9.329200
7	2.105300	2.580123	17.234400	9.303000
8	1.989700	2.574540	17.551400	9.348600
9	1.945200	2.572600	17.608000	9.370000

10	**1.877100**	**2.572255**	**17.777300**	**9.295600**
11	**1.831800**	**2.572932**	**17.565300**	**9.318400**
12	**1.812600**	**2.574233**	**17.611300**	**9.289200**
13	**1.767200**	**2.576111**	**17.709200**	**9.295400**
14	**1.758500**	**2.576909**	**17.748900**	**9.302600**
15	**1.737700**	**2.578087**	**17.714500**	**9.310000**

BLEU score on the hold-out or test dataset is used for model evaluation.

BLEU score for the pre-trained mode on hold-out dataset is **11.0621.**

BLEU score for the best model after fine-tuning on hold-out dataset is **17.7773.**

Shown below are some translations from English to Hindi:

Simple sentences:

1. **English** : Have you seen a spider moving on a wall?

Model Translation : एक मकड़ी दीवार पर चला हुआ देखा है?

Reference : क्या आपने मकड़ी को दीवार पर चलते हुए देखा है?

2. **English** : Aisha likes to watch the TV show, "The Little Monkey".

Model Translation : टीवी शो देखना पसंद करते हैं, 'छोटे गेंद'

Reference : आयशा को टीवी शो "द लिटिल मंकी" देखना पसंद है।

3. **English** : Because it does nothing to help us and it is lazy.

Model Translation : क्योंकि यह हमारी मदद करने और आलसीपन की बात नहीं है।

Reference : क्योंकि यह हमारी मदद करने के लिए कुछ नहीं करता है और यह आलसी है।

4. **English** : Cricket is an outdoor game while Snakes and Ladders is an indoor game.

Model Translation : क्रिकेट एक बाहरी खेल है, जबकि गेंदबाजों और टायरों का खेल है।

Reference : क्रिकेट एक आउटडोर खेल है जबकि सांप और सीढ़ी एक इनडोर खेल है।

Complex sentences:

1. **English** : Coronavirus-induced lockdown brings country to a grinding halt.

Model Translation : कोरोना वायरस-इन्ड लॉकडाउन से देश कोना बंद हो जाता है।

Reference : कोरोनावायरस-प्रेरित लॉकडाउन देश को पीसने की स्थिति में लाता है।

2. **English** : Since winter is coming, I think I'll knit a warm sweater, because I'm always cold.

Model Translation : सर्दियों से आ रहा है, मुझे लगता है कि मैं एक गर्म मिर्च पहनूंगा, क्योंकि मैं हमेशा ठंडा हूं।

Reference : चूंकि सर्दी आ रही है, मुझे लगता है कि मैं एक गर्म स्वेटर बुनूंगा, क्योंकि मैं हमेशा ठंडा रहता हूं।

3. **English** : I really didn't like the movie even though the acting was good.

Model Translation : मुझे फिल्म पसंद नहीं थी, भले ही काम अच्छा रहा हो।

Reference : अभिनय अच्छा होने के बावजूद मुझे वास्तव में फिल्म पसंद नहीं आई।

4. **English** : After being apart for years, he still had feelings for her.

Model Translation : सालों से अलग रहने के बाद भी उसे उनके लिए सहानुभूति थी।

Reference : सालों तक अलग रहने के बाद भी उसके मन में उसके लिए भावनाएँ थीं।

5. **English** : We also found minor boys and girls there.

Model Translation After fine-tuning : वहीं, हमें लड़कियों और नाबालिग लड़कों की भी खबर है।

Reference (from train dataset): हमने वहां नाबालिग लड़के और लड़कियों को पाया।

Before Fine-tuning : हमें वहाँ छोटे लड़के और लड़कियाँ भी मिलीं ।

In these above-mentioned samples, the reference translations are obtained using google translate service.

In Simple sentences samples:

2nd and 3rd sentences appear to be different from that of reference translations.

In 2nd sentence, the model is unable to translate the name "*Aisha*" at all.

In 3rd sentence, the translation is giving a different context on whole.

In 4th sentence, "*Snakes and Ladders is an indoor game*" is not translated correctly by the model.

In Complex sentences samples:

In 1st sentence, the model doesn't to capture the context, this can be observed from the translation of "*brings country to a grinding halt*". Here the model translated as "*देश कोना बंद हो जाता है।*. The input represents something like a sense of economy and people's lives are halted but this is not conveyed in the translation.

In 2nd sentence, the part "*knit a warm sweater*" was translated as "*एक गर्म मिर्च पहनूंगा"* possibly because the model is not aware of the translation for sweater - as model vocabulary may not have had a Hindi translation for sweater.

In 4th sentence, the translation for the word "*feelings*" was done as "*सहानुभूति*", instead of "*भावनाएँ*". Although this may reflect feeling , the appearance of this word in Hindi translation appears a complex translation by the model.

The 5th sentence is used for perform comparison before and after fine-tuning. Before fine-tuning the words "*minor boys*" translated as "*छोटे लड़के*". This doesn't reflect the context; 'minor' is taken as 'small' during translation. But after fine-tuning, the context for translation is reflected by the translation "*नाबालिग लड़कों*".

Part 2:

Here again it has two parts:

PART A:

This consists of training a seq2seq model with attention from scratch. The encoder GRU cells and decoder consists of attention with GRU.

Attention is applied to the outputs of the encoder and then this is combined with the last hidden state, also called as skip thought vector. This is then provided as input to the decoder to obtain the translations.

For training 50000 sentence pairs are considered as the total dataset. The dataset is split into 0.9 train data and 0.1 test data. So, the train data consists of 45000 sentence pairs and test data consists of 5000 sentence pairs. It is also to be noted that these sentence pairs are selected such that the length is between 2 and 10 words.

To obtain better convergence, teacher forcing is also used. It is to be noted that we attempted to train this seq2seq network from scratch. To improve the training process, fast text word embeddings are used for both encoder and attention-decoder.

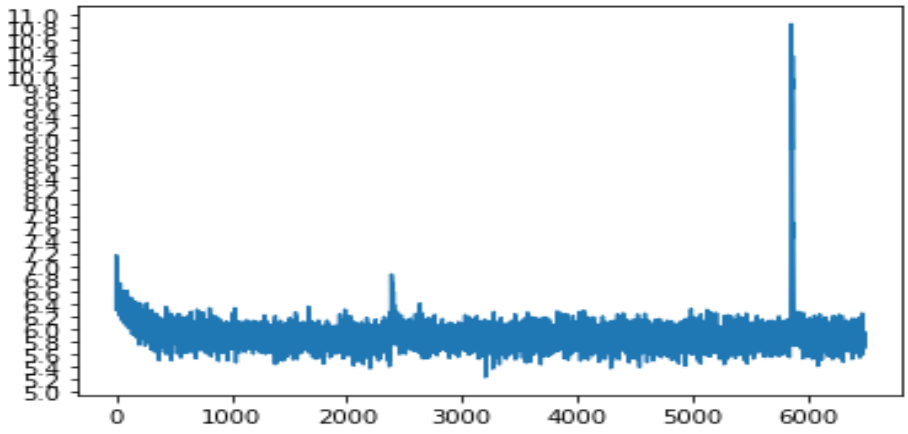
English fast text word embeddings are used in the encoder and Hindi fast text word embedding are used in the attention-decoder.

Training is carried out with the following hyper-parameters :

learning rate of **0.01** , weight decay of **0.001** and momentum **0.9**.

Training is carried out for about **650000** iterations:

The loss curve is shown below :



Below are the translations for some of the samples in train dataset:

> Police arrested five, including the husband and husbands brotherin-law.

= पति और सास-ससुर सहित पांच गिरफ्तार

< पुलिस के पर और और <EOS>

> So, when we do sampling with replacement.

= तो, जब हम प्रतिस्थापन के साथ नमूना(sample) करते हैं।

< साथ हमें साथ साथ साथ साथ साथ साथ साथ <EOS>

> Other symptoms include:

= अन्य लक्षणों की बात करें, तो इसमें:

< आप और में <EOS>

> But that didnt happen in this case.

= लेकिन इस दवा के मामले में ऐसा नहीं हुआ.

< लेकिन इससे इससे इससे इससे इससे इससे इससे इससे इससे

> But people are clever.

= लेकिन लोग होशियार हैं.

< लेकिन लोग लोग लोग लोग हैं। <EOS>

> Are you satisfied with this?

= आप इससे संतुष्ट हैं?

< आप आप आप आप आप आप हैं? हैं? <EOS>

> Some Ailments That Have Been Linked to Stress

= कुछ बीमारियाँ जिनका संबंध तनाव से जोड़ा गया है

< इससे कुछ कुछ कुछ कुछ कुछ कुछ <EOS>

> Rajasthan Subordinate and Ministerial Services Selection Board (RSMSSB)

= विभाग - राजस्थान अधीनस्थ एवं मंत्रालय सेवा चयन बोर्ड

< और और और और और <EOS>

> This caused much inconvenience to the passengers.

= जिसके चलते यात्रियों को काफी परेशानी होने लगी.

< इससे इससे इससे इससे इससे इससे इससे इससे इससे इससे

> Read the second part:

= इसके आगे का पार्ट पढ़ें-

< आप के की नहीं <EOS>

Below are the translations for some of the samples in test dataset

Actual :

[['This fact cannot be denied.', 'इस तथ्य पर यकीन नहीं किया सकता.'],

['So far, 11 have been arrested.',

'इस बार 11 शहजादों को गिरफ्तार किया गया है.']]

Predictions :

[['इससे', 'इससे', 'इससे', 'इससे', 'इससे', 'इससे', 'इससे', 'इससे', 'इससे'],

['तो', 'इसकी', 'के', 'में', 'में', 'में', 'में', 'हैं।']]

A BLEU score of **0.00853** is obtained on the test dataset.

PART B:

In addition to this, training of seq2seq model is also carried out by selecting sentences whose length lies between 2 words to 5 words, along with using 2 GRU layers both in encoder and attention decoder.

We observed that this didn't improve the BLEU score on the hold out set. Sample translations on the hold-out/test dataset and BLEU scores are as below.

A BLEU score of **0.0076** is obtained on the test dataset.

Below are the translations for some of the samples in train dataset:

> PF interest rate up

= पीएफ पर बढा ब्याज

< कृषि समय का <EOS>

> Congress leader Udit Raj

= कांग्रेस नेता उदित राज

< कांग्रेस कांग्रेस कांग्रेस <EOS>

> Look at me.

= मेरी ओर देखो।

< बीजेपी को <EOS>

> The case against Walmart

= वालमार्ट के खिलाफ मुकदमा

< कम के के <EOS>

> Statistics testify to this.

= आंकड़े इसकी गवाह हैं।

< कम का का <EOS>

> In 1 hour.

= 1 घंटे में।

< कम 1 में <EOS>

> Watch the song below.

= नीचे देखें गाना।

< देखें देखें देखें <EOS>

> Central Pay Commissions

= केंद्रीय वेतन आयोग

< कांग्रेस का का <EOS>

> to next key change

= अगले कुंजी परिवर्तन पर

< कम से से <EOS>

> The Center of Administration

= प्रशासन का केंद्र

< कम की की <EOS>

From the above samples , it can be observed and concluded that the seq2seq model with attention and using GRU cells needs huge amount of data and training time to obtain decent BLEU score on hold-out dataset when trained from scratch.

Comparing our results with the papers reviewed:

ALGORITHM	TRANSLATION	BLEU SCORE	TRANSLATION	BLEU SCORE
DYNAMICCONV (WU ET AL., 2019	En → De	29.7	En → Fr	43.2
EVOLVED TRANFORMER (SO ET AL., 2019)	En → De	29.8	En → Fr	41.3
TRANFORMER +LARGE BATCH (OTT ET AL., 2018)	En → De	29.3	En → Fr	43.0

REPRODUCED TRANSFORMER	En → De	29.12	En → Fr	43.96
BERT-FUSED MODEL	En → De	30.75	En → Fr	43.78
ATTENTION MODEL (HIDDENSIZE 500) **NEWSTEST2016** **NEWSTEST2017**	En → De	27.0 22.1	En → Fr	31.9 27.5
2D-SEQ2SEQ MODEL (HIDDENSIZE 500) **NEWSTEST2016** **NEWSTEST2017**	En → De	27.5 22.4	En → Fr	32.6 28.2
2D-SEQ2SEQ MODEL +WEIGHTING (HIDDENSIZE 500) **NEWSTEST2016** **NEWSTEST2017**	En → De	27.5 22.4	En → Fr	32.3 27.9
ATTENTION MODEL (HIDDENSIZE 1000) **NEWSTEST2016** **NEWSTEST2017**	En → De	27.4 22.9	En → Fr	33.1 29.0
2D-SEQ2SEQ MODEL (HIDDENSIZE 1000) **NEWSTEST2016** **NEWSTEST2017**	En → De	28.9 23.2	En → Fr	33.7 29.3
2D-SEQ2SEQ MODEL +WEIGHTING (HIDDENSIZE 1000) **NEWSTEST2016** **NEWSTEST2017**	En → De	27.8 22.7	En → Fr	32.7 28.0
COVERAGE (HIDDENSIZE 1000)	En → De		En → Fr	

NEWSTEST2016 **NEWSTEST2017**		28.6 23.0		33.1 28.7
FERTILITY (HIDDENSIZE 1000) **NEWSTEST2016** **NEWSTEST2017**	En → De	28.4 23.2	En → Fr	33.4 28.9
BASELINE	En →De			
NMTSRC.FT	En →De			
WIKI.FT	En →De			
NEWS.FT	En →De			
NEWS.EMB	En →De			
FINE-TUNED "HELSINKI-NLP/OPUS-MT-EN-HI" TRANSFORMER (6 LAYERS IN BOTH ENCODER-DECODER) **45000 TRAIN (EN TO HI)** **5000 TESTS (EN TO HI)**	En → Hi	17.7773		
SEQ2SEQ - ATTENTION (1 LAYER GRU IN BOTH ENCODER-DECODER) SAMANANTAR ENGLISH TO INDIAN LANGUAGES. **45000 TRAIN (EN TO HI)** **5000 TEST (EN TO HI)**	En → Hi	0.00853		

We conclude by saying - with the power of fine-tuning , transfer learning and transformer architecture we have obtained a decent BLEU for English-Hindi translation task even with small dataset.

Detailed Paper Reviews:

PAPER 1:

On the use of BERT for Neural Machine Translation.

Recently, neural machine translation (NMT) tasks have gained a lot of visibility by using larger pre-trained models. These models depend on the transformer model, and a feed – forward network depending on attention mechanism based on recurrent neural nets. The BERT model has a transformer model (reused to learn bi-directional language model for large text corpora) for predicting the masked word in a given sentence by understanding the context and predicting the next sentence for deciding if the two sentences are successive or not. The author of this paper shows that the BERT pre-training model can be used to supervised NMT, and by studying the results of monolingual data, various methods of integrating the pre-training BERT model with the NMT model are also compared. This work is carried out in two phases: the first phase involves the systematic comparison of the performance of different BERT + NMT architectures on a standard supervised NMT, and the second phase involves evaluating the data set in the domain beyond the BLEU score.

The encoder – decoder architecture is adopted by the typical NLT model. Here the encoder used for contextualizing the word embeddings from the input sentence and decoder is used for generating the translation as output from left to right. The pretrained Language Models (LM) like BERT can be used to encoder as prior knowledge for NMT models which in turn provides a very good contextual word embeddings learnt from monolingual corpus. The main aim is to train such models once and reuse them on different language pairs. There are two advantages to using BERT pre-training-one of them is that it helps to solve the monolingual task of source sentence encoding, and the second advantage is for big data. It also helps in learning the translation – related information in the encoder – decoder model. The comparison is carried out with the below models:

Baseline - A transformer-big model has a shared decoder input-output embedding parameters.

Embedding (Emb) – For the baseline model the embedding layer is replaced by the BERT parameters (thus having 6 + 6 encoder layers). The model is then fine-tuned almost like the ELMO setting.

Fine-Tuning (FT): For the baseline model, encoder is initialized by the BERT parameters.

Freeze: For the baseline model, encoder is initialized by the BERT parameters and frozen.

Three monolingual corpora belonging to different size and domains were used for pretraining BERT models – NMT-scr, Wiki and News. The motivation behind using these are – the NMT-scr is used to evaluate the robustness after training it on source corpora, the Wiki corpora is considered as out of domain data in comparison with News dataset and the News dataset consists of huge data and is considered to be in domain data.

	Lines	Tokens
NMT-src	4.5M	104M
Wiki	72M	2086M
News	210M	3657M

Table 1: Monolingual (English) training data

The BLEU score cannot be determined on the in-domain data so authors have introduced the additional evaluation which helps in determining the impact on LM pretrained models on various out of domain data and also to evaluate the robustness on different types of noises. There are 4 types of noise introduced: Typos – synthetic noise is added to the test data randomly by swapping characters and/or by insert or delete operation on the characters and/or upper casing the tokens. Unk – this deal with the unknown characters being introduced either at the beginning or end of the sentence for source and target. ChrF – It is the distribution of difference in sentences charf between the original and noise test data. It is the corpus-level metric, it measures the sentence count $\Delta(\text{chrF})(m, n, s) = \text{chrF}(m(n(s)), r) - \text{chrF}(m(s), r)$. NMT models are trained on English-German and English-Russian IWSLT MT datasets to understand the masked LM encoder pretrained for lower resource settings.

The authors have successfully carried out the below experiments and results are satisfactory for the argument made.

- Masked language model task is advantageous compared to the next sentence prediction task carried out in BERT. BERT can be trained on any one source language and be reused on various other translation pairs which provides better initialization point and performance.
- Pretraining encoders have better initialization for NMT encoders when trained BERT on source corpora. Also, the training data size decreases and enables to train on bigger model.

- Finetuning the BERT pretrained encoder is more convenient as it helps in retaining the same model size than reusing the BERT in the embedding layer which might increase the model size significantly.
- Using the pretrained encoders doesn't need to consider just the BLEU scores but it can be generalized to be used in new domains and also in terms of robustness.

Limitations:

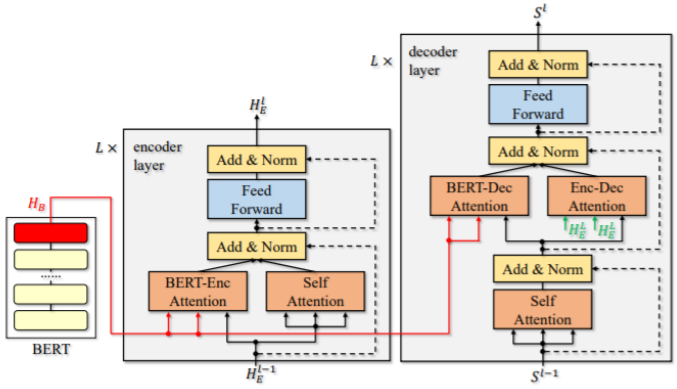
- Though the BERT pretrained model's performance is better on noise and domain test data but there are no clear results shown regarding this.
- The robustness tests carried out are not showing evident results to the experiments conducted which are supposed to be more robust than expected.

PAPER 2:

Incorporating BERT into neural machine translation

BERT which was recently proposed turned out to be much powerful on the various NLU tasks but the application of it in Nature Machine Translation (NMT) hit a hard rock. More often BERT is used for fine tuning rather than for contextual embeddings, so authors have proposed new algorithm – BERT-fused model. Here the representation from BERT is carried out by inserting it into all layers and attention mechanism is adapted to observe how each layer will be interacting with the representations. This is carried out to as the BERT and NMT modules might use different word segment rules which in turn results in different sequence lengths. There are two attention modules for BERT, one is encoder attention and other is decoder attention. Experiments are conduction on supervised which involves sentence-level and document level translations, semi-supervised and unsupervised MT on sever datasets.

The proposed algorithm is carried out in couple of steps explained below:



Step1: given the input, BERT model will first try to encode it into representations H_B = BERT(input) where H_B is considered as the output og the last layer in BERT.

Step 2: Hidden representation of l-th layer of the encoder is denoted as H_Z^l and the word embeddings in the sequence x is denoted as H_E^0.

$$\tilde{h}_i^l = \frac{1}{2}\left(\mathtt{attn}_S(h_i^{l-1}, H_E^{l-1}, H_E^{l-1}) + \mathtt{attn}_B(h_i^{l-1}, H_B, H_B)\right), \forall i \in [l_x]$$

where the $attn_S$ and $attn_B$ are attention models having different parameters and the encoder will produce the output from the last layer.

Step 3: $S_{<t}^l$ be the hidden state at the decoder with the time step as t:

$$\hat{s}_t^l = \mathtt{attn}_S(s_t^{l-1}, S_{<t+1}^{l-1}, S_{<t+1}^{l-1});$$
$$\tilde{s}_t^l = \frac{1}{2}\left(\mathtt{attn}_B(\hat{s}_t^l, H_B, H_B) + \mathtt{attn}_E(\hat{s}_t^l, H_E^L, H_E^L)\right),\ s_t^l = \mathtt{FFN}(\tilde{s}_t^l).$$

$attn_S$, $attn_B$ and $attn_E$ are the self-attention model, BERT decorder attention model and the encoder-decoder model respectively and these keep iterating over the layers to obtain S_t^l
which is then mapped via a linear transformation and softmax for obtained the t-th prediction word. This decoding process iterates till the end of the sentence token is obtained.

In the above given architecture, BERT is at the left side, encoder and decoder model at the right and the dash lines are residual connections. The red color in the left side denote H_B and the green color H_E^L denote the output from the last layer from the BERT and encoder. The output from the BERT is an external sequence representation and attention model is used in NMT model to leverage the pre-trained model rather than the tokenization way.

The BERT – fused model is verified on the supervised setting which has both the low-resource and rich-resource scenarios and experiments are carried out in document level translation. Later the model is combined with back translation to observe the effectiveness on semi supervised NMT. The dataset used are IWSLT'14 and IWSLT'17, to match the dimensions of BERT and NMT models authors have considered $BERT_{base}$ for IWSLT and $BERT_{large}$ for WMT tasks. NMT model is first train till it convergences and then two attention models of BERT are initialized randomly. The results shown are evident on the BLEU scores increases on both datasets. For the translation with document level contextual information, BERT is able to capture the context between two sentences and also predict

if these sentences are consecutive or not. The algorithm for this is to translate a sentence to target domain by considering both sentence and its preceding sentence as inputs. The sentence is fed to encoder, similar to sentence level translation and for BERT the input is the concatenation of two sequences. The results appear to have achieve much better values compared to sentence level baseline. For Unsupervised NMT, the BERT is pretained on XLM model. Initially the we train the unsupervised NMT till it converges and then BERT-fused model is initialized with the obtained mode and training is carried out. It shows evident results on BLEU score that this model is much better than the previous proposed ones.

The authors have successfully carried out the below experiments and results are satisfactory for the argument made.

- Pre trained model's output features are fused in all layers in the NMT model making sure the pretrained featured are completely made use of.
- Attention model acts a bridge between the NMT model and the pretrained features of BERT.
- When the BERT-encoder and BRET-decoder were removed, the BLEU scores were dropped, this shows that output of BERT should be applied to both encoder and decoder for obtaining better scores.
- The concatenation of BERT and NMT model as BERT-fused model has shown promising results and also the experiments conducted on supervised, semi supervised and unsupervised NMT models shown the effectiveness.

Limitations:

- The pre-trained BERT models used were trained on huge datasets and the models build on such are somewhere indirectly making use of the additional training data.
- BERT and NMT models are making use of different subword vocabulary, so the models with BERT for initialization only face the problem of mismatch tokenization.

PAPER 3:

Towards Two-Dimensional Sequence to Sequence Model in Neural Machine Translation

The authors experimented on the alternative methods for Neural Machine Translation (NMT). As the part of the paper, they proposed that a variant method of LSTM called Multi-Dimensional Long Short-Term Memory (MDLSTM) can be

used for translation modelling. According to the authors the Machine translation is a two-dimensional mapping of MDLSTM. The method MDLSTM was initially implemented by some other researchers, even though the authors found the work is interesting since their application in Hand Writing Recognition and automatic speech recognition. There are many auto encoder techniques for machine translation which encodes the translating language and decodes it in to the required language. For the same purpose wide variety of CNN and RNN networks are there. In all these applications the Two languages i.e. the source and target languages are dealt separately in the form of one-dimensional sequences. Some methods make use of attention models which make the model to do selective focus on specific part of the language architecture. In these methods the encoder states are computed only at the beginning and left untouched progressing the translation, according to the authors of this paper this is one of the draw backs which is they overcame by introducing the 2DLSTM. In the proposed method at every decoding step encoding states are visited repeatedly in order to ensure more accuracy. This is done by applying source to train map on 2DLSTM as two-dimensional sequence and they named the model two-dimensional sequence to sequence model.

As a part of their experiment, they did translation from German to English and English to German. Datasets they used were Europarl-v7, News-commentary-v10 and Common-Crawl. They trained their model on samples having size with 4.6 Million training samples. They did tokenization and true-casing using a toolkit called Mosses. They trunked off the sentences with more than 50 sub words and batched them together. This action is fine because most of the sentences doesn't go beyond 50 sub words. They used Adam optimizer, dropout of 30%. They used case sensitive BLUE for evaluating the model. They experimented on attention model and their own 2D-seq2seq model with hidden size 500 and 1000. Their 2D seq2seq model outperformed the standard attention model by a considerable margin. But the margin is not too significant. To improve the score, they did try introducing a new normalized weight layer. But it did not help the performance significantly in terms of BLEU or TER. They extended their experiment by testing on coverage model and fertility model. Both models are based on the feedback arrangement, even though the 2D seq2seq model managed to keep up with these models with a slightly greater performance.

By the experiment they were able to introduce a novel method called 2D-seq2seq model. One of its drawbacks is it took 1 second to train 791 words and 1 second to decode 0.7 words at the same time attention model which they took as their competitive model could process 2944 and 48 words per second to train and

decode respectively. The authors mentioned that they used single gpu for training, but by redeveloping the architecture they will be able to accomplish a faster speed for training and decoding. The introduced 2D-seq2seq model has its own advantage since it using a tab on the source language while its decoding. Even though the architecture looks promising they were not able to show the full ability of their model. The comparatively worst speed also questions the model's need. Still the authors have a very clear vision about the future works on their model. One of them is changing the architecture compatible to the multi-gpu utilization, which can certainly improve the speed way better than that they achieved now. Also, they plan to build a two directional 2DLSTM and stacking 2DLSTMs to create a deeper model.

Limitation:

- The one thing authors missed is to validate their model on different language sets. Since we know all the languages does not keep the same structure and policies. Since their model is constantly looking on the source and target language model there is a chance for certain languages or language pairs this model can outperform the attention model in a great margin. But still, it is untouched by the authors.

- To improve their speed, they can use faster LSTM techniques. Also, they can try to implement their concept with GRU model.

Bonus:

Following is a link for fine food reviews - https://www.kaggle.com/snap/amazon-fine-food-reviews

In order to perform English to Hindi translations, the amazon-fine-food-reviews is obtained from Kaggle. Since the number of reviews are huge which are nearly 50K, 200 reviews are considered to perform language translation.

Translations are performed using fine-tuned transformer model from Part 1. The list of translations is provided in a file (translated_200.csv).

The translations are assessed manually by randomly picking up 20 samples. Since this is a review dataset, the language used to write the reviews is generic in nature and may not be structured. This is already reflected in different sentences of the reviews; hence the model may not perform a perfect translation.

Some translations were not performed for the entire review. This may be due to max input length setup for the model configuration.

Following are examples of some translations which appear fine :

I have bought several of the Vitality canned dog food products and have found them all to be of good quality. The product looks more like a stew than a processed meat and it smells better. My Labrador is finicky and she appreciates this product better than most.	['मैंने कई विकेट खरीद लिए हैं और सभी को अच्छे गुणवत्ता के रूप में पाया है। उत्पाद एक संसाधित मांस की तुलना में अधिक लग रहा है और इसका स्वाद अधिक बढ़ता है।']
This is a confection that has been around a few centuries. It is a light, pillowy citrus gelatin with nuts - in this case Filberts. And it is cut into tiny squares and then liberally coated with powdered sugar. And it is a tiny mouthful of heaven. Not too chewy, and very flavorful. I highly recommend this yummy treat. If you are familiar with the story of C.S. Lewis' "The Lion, The Witch, and The Wardrobe" - this is the treat that seduces Edmund into selling out his Brother and Sisters to the Witch.	['यह एक छोटा सा चम्मच होता है जो कि कुछ शताब्दियों से होता है। इस मामले में हल्दी कुमारी के साथ। और यह छोटे वर्ग में काट दिया जाता है और फिर निविंदीवाला होता है। और यह आकाश का एक छोटा सा मुंह है। मैं बहुत ही स्वादिष्ट नहीं, और मैं इसका बहुत आदर करता हूँ। यदि आप जानते हैं, तो इस मामले में प्रिय हैं।']

In the above although the 2nd review is fine, some words such as "*C.S. Lewis*" are not translated. This may be due the vocabulary and corpus that was used for training.

Following are translations which are not perfect:

Great taffy at a great price. There was a wide assortment of yummy taffy. Delivery was very quick. If your a taffy lover, this is a deal.	['बहुत भारी कीमत पर है। yyyoffy की एक व्यापक अभिनेत्री थी। वितरण बहुत तेज था। यदि आपका एक म्यूजिक प्रेमी, यह एक सौदा है।']
This product serves me well as a source of electrolytes during and after a long run or bike ride. I have tried all of the flavors but really do like the grapefruit flavor... no after-taste and I actually like the slight	['यह उत्पाद मुझे लंबी या बाइक चलाने के बाद भी काम करता है.']

carbonation. I use other Hammer products and really like their whole product line.	
Halloween is over but, I sent a bag to my daughters class for her share. The chocolate was fresh and enjoyed by many.	['उन्होंने कहा, ‘‘‘‘हैन ओवर है लेकिन मैंने अपनी बेटियों को बैग भी भेजा है ।']

In the above sentences, yummy taffy was not translated to Hindi. In the second row, electrolyte was not translated and in third row the translation is not appropriate as no context is reflected from the source in the translated text.

We hypothesize that the translation can be improved by training the model further with more data specifically collected from reviews. More advanced approaches such as "**Zero-Shot Translation**" can also be employed to ensure models translate sentences, phrases and words that were never seen before.

References:

https://shaojiejiang.github.io/post/en/transformer-align-model/

https://huggingface.co/transformers/model_doc/marian.html

https://huggingface.co/Helsinki-NLP/opus-mt-en-hi

https://marian-nmt.github.io

https://pytorch.org/tutorials/intermediate/seq2seq_translation_tutorial.html

https://indicnlp.ai4bharat.org/indicft/#downloads

https://fasttext.cc/docs/en/english-vectors.html

https://ai.googleblog.com/2016/11/zero-shot-translation-with-googles.html