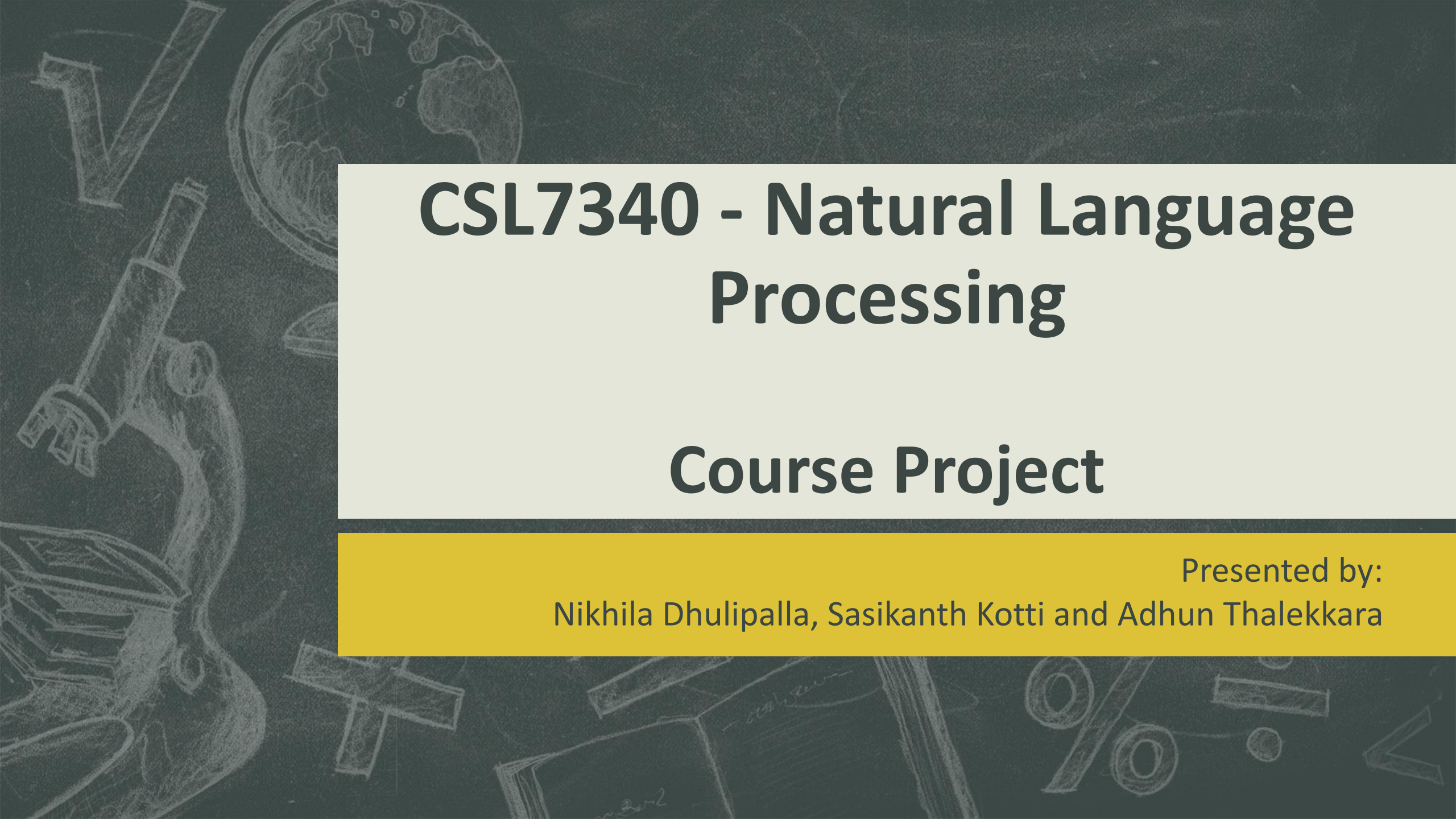


# CSL7340 - Natural Language Processing

## Course Project

Presented by:
Nikhila Dhulipalla, Sasikanth Kotti and Adhun Thalekkara

# Presentation Subsections

- Project task selected
- Paper – 1
- Paper – 2
- Paper – 3
- Architecture implemented by us
- Application live demo
- Conclusion

# Project task selected

- We have selected the task as Machine Translation for the course project.

- We have chosen En-Hi corpus and it is taken from the given link - https://indicnlp.ai4bharat.org/samanantar/.

# Paper – 1:
## On the use of BERT for Neural Machine Translation

- The work in this paper has 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 NMT model, 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.
- BERT is used to encoder as prior knowledge for NMT models which in turn provides a very good contextual word embeddings learnt from monolingual corpus.
- 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.

**Results of the BLEU scores:**

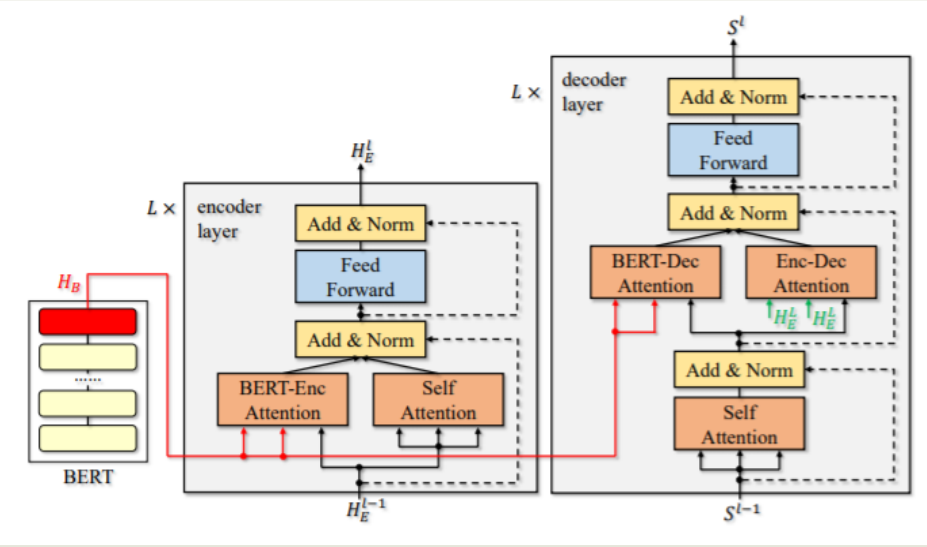
| Models | news14 |
|---|---|
| Baseline | 27.3 |
| NMTsrc.FT | 27.7 |
| Wiki.FT | 27.7 |
| News.FT | **27.9** |
| News.Emb | 27.7 |

**Advantages:**

- 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.

# Paper – 2:
## Incorporating BERT into neural machine translation

- 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.




- 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 $H_B$ (the output of the last layer in BERT) and the green color $H_E^L$ denote the output of 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.

## Results of the BLEU scores:

Table 2: BLEU of all IWSLT tasks.

| | Transformer | BERT-fused |
|---|---|---|
| En→De | 28.57 | 30.45 |
| De→En | 34.64 | 36.11 |
| En→Es | 39.0 | 41.4 |
| En→Zh | 26.3 | 28.2 |
| En→Fr | 35.9 | 38.7 |

Table 3: BLEU scores of WMT'14 translation.

| Algorithm | En→De | En→Fr |
|---|---|---|
| DynamicConv (Wu et al., 2019) | 29.7 | 43.2 |
| Evolved Transformer (So et al., 2019) | 29.8 | 41.3 |
| Transformer + Large Batch (Ott et al., 2018) | 29.3 | 43.0 |
| Our Reproduced Transformer | 29.12 | 42.96 |
| Our BERT-fused model | 30.75 | 43.78 |

Table 4: BLEU of document-level translation.

| | En→De | De→En |
|---|---|---|
| Sentence-level | 28.57 | 34.64 |
| Our Document-level | 28.90 | 34.95 |
| Miculicich et al. (2018) | 27.94 | 33.97 |
| Sentence-level + BERT | 30.45 | 36.11 |
| Document-level + BERT | 31.02 | 36.69 |

## Advantages:

- 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 BERT-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.

# 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) - proposed a variant method of LSTM called Multi-Dimensional Long Short-Term Memory (MDLSTM) for translation modelling.
- The implementation extends the current sequence to sequence backbone NMT models to a 2D structure in which the source and target sentences are aligned with each other in a 2D grid.

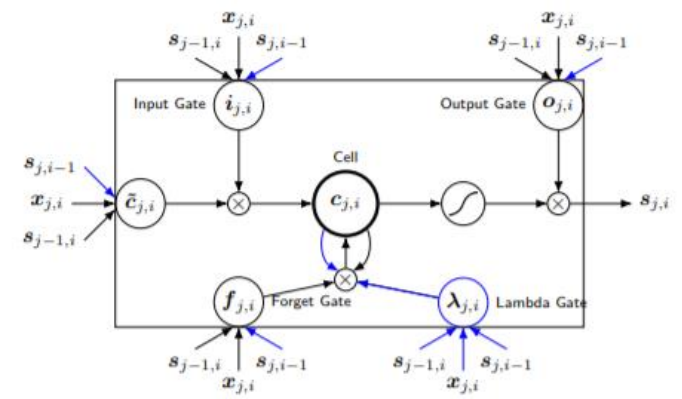



Figure 1: 2DLSTM unit. The additional links vs. standard LSTM are marked in blue.

- It maintains the state information in an internal cell state and apart from input, forget and the output gates that all control information flows, 2DLSTM employs an extra lambda gate - used to weight the two predecessor cells before passing them through the forget gate.

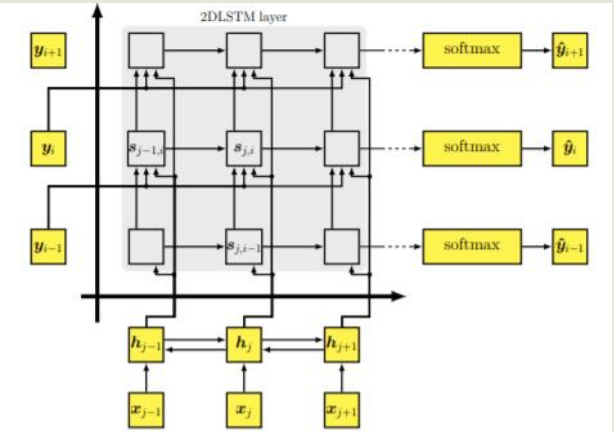



Figure 2: Two-dimensional sequence to sequence model (2D-seq2seq).

- Given a source sequence and a target sequence, we scan the source sequence from left to right and the target sequence from bottom to top.
- In the 2D-seq2seq model the horizontal-axis is the encoder and vertical axis is the decoder.
- As a pre-step before the 2DLSTM, in order to have the whole source context, a BiLSTM scans the input words once from left to right and once from right to left to compute a sequence of encoder states.
- 2DLSTM receives both encoder state and the last target embedding vector as an input. It repeatedly updates the source information, while generating new target word.

**Results of the BLEU scores:**

| | Models | Hidden Size | De→En | | | | | En→De | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | devset | newstest2016 | | newstest2017 | | devset | newstest2016 | | newstest2017 | |
| | | | PPL | BLEU | TER | BLEU | TER | PPL | BLEU | TER | BLEU | TER |
| 1 | attention | n=500 | 7.3 | 31.9 | 48.6 | 27.5 | 53.1 | 7.0 | 27.0 | 53.9 | 22.1 | 60.5 |
| 2 | 2D-seq2seq | | 6.5 | **32.6** | 47.8 | **28.2** | 52.7 | 6.1 | **27.5** | 53.8 | **22.4** | 60.6 |
| 3 | + weighting | | 6.5 | 32.3 | **47.1** | 27.9 | **51.7** | 6.3 | **27.5** | **53.3** | **22.4** | **60.0** |
| 1 | attention | n=1000 | 6.4 | 33.1 | 47.5 | 29.0 | **51.9** | 6.5 | 27.4 | 53.9 | 22.9 | 60.2 |
| 2 | 2D-seq2seq | | 5.7 | **33.7** | **46.9** | **29.3** | **51.9** | 5.3 | **28.9** | **52.6** | **23.2** | **59.5** |
| 3 | + weighting | | 6.1 | 32.7 | 47.1 | 28.0 | **51.9** | 5.7 | 27.8 | 53.0 | 22.7 | 60.0 |
| 4 | coverage | n=1000 | 6.3 | 33.1 | 47.5 | 28.7 | 51.9 | 5.8 | 28.6 | 52.4 | 23.0 | 59.4 |
| 5 | fertility | | 6.2 | 33.4 | 46.9 | 28.9 | 51.6 | 5.8 | 28.4 | 52.1 | 23.2 | 59.1 |

**Advantages:**

- The Novel 2D sequence to sequence model (2D-seq2seq), network that applies a 2DLSTM unit to read both the source and the target sentences jointly.
- In each decoding step, the network implicitly updates the source representation conditioned on the generated target words so far.

# Our architecture:

- We have implemented the following in our project:

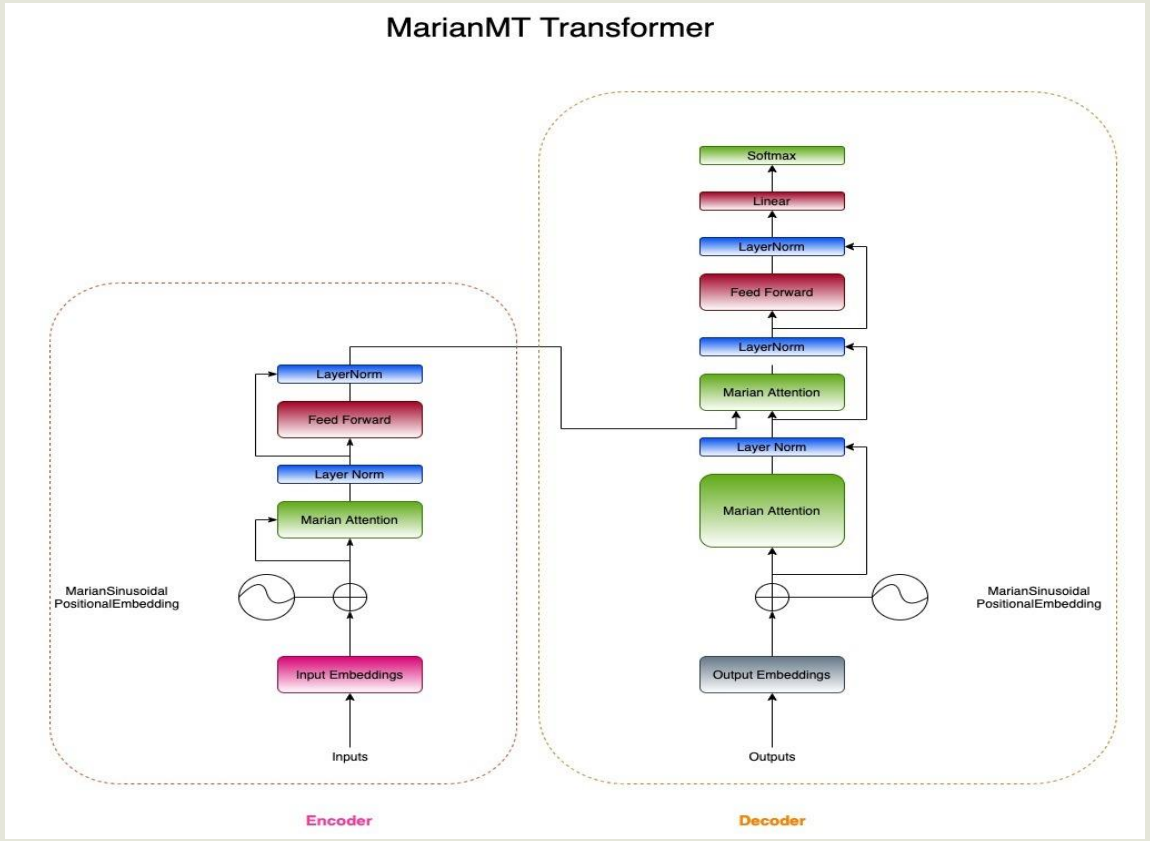
1. A pre-trained transformer model is considered for fine-tuning and it is implemented using Hugging face framework. "*Helsinki-NLP/opus-mt-en-hi*" is the pre-trained model considered for fine-tuning. It is a transformer-align family trained using MarianMT framework. This transformer encoder-decoder 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.

2. Training a seq2seq model with attention from scratch and the encoder consists of GRU cells and decoder consists of attention with GRU. Here the 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.

3. 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.

4. Bonus - Translations are performed using fine-tuned transformer model from point 1 here.

# MarianMT transformer implemented




- The base model considered for this task is English-Hindi translation and fine tuned.
- We have implemented few modifications on the existing dataset – flores and used it for our architecture.
- Helsinki-NLP/opus-mt-en-hi was the pretrained model used for fine tunning and we have performed the evaluation for before and after finetuning of the model.
- After fine tuning, the model is able to give better BLEU score comparatively and we have chosen the best BLEU score from the epochs run.

```
{'eval_bleu': 11.0621,
 'eval_gen_len': 9.559,
 'eval_loss': 3.7714550495147705,
 'eval_mem_cpu_alloc_delta': 24780800,
 'eval_mem_cpu_peaked_delta': 2457600,
 'eval_mem_gpu_alloc_delta': 0,
 'eval_mem_gpu_peaked_delta': 1332053504,
 'eval_runtime': 149.6696,
 'eval_samples_per_second': 33.407,
 'init_mem_cpu_alloc_delta': 2338455552,
 'init_mem_cpu_peaked_delta': 0,
 'init_mem_gpu_alloc_delta': 305772544,
 'init_mem_gpu_peaked_delta': 0}
```

```
{'eval_bleu': 17.7773,
 'eval_gen_len': 9.2956,
 'eval_loss': 2.5722546577453613,
 'eval_mem_cpu_alloc_delta': 1523712,
 'eval_mem_cpu_peaked_delta': 1368064,
 'eval_mem_gpu_alloc_delta': 0,
 'eval_mem_gpu_peaked_delta': 1179455488,
 'eval_runtime': 138.0769,
 'eval_samples_per_second': 36.212,
 'init_mem_cpu_alloc_delta': 4096,
 'init_mem_cpu_peaked_delta': 0,
 'init_mem_gpu_alloc_delta': 305772544,
 'init_mem_gpu_peaked_delta': 0}
```

| 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 |

- We have considered few sentences and categorized them as simple and complex sentence and below are the English, model translated output and reference is the google translator.
- The 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.

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?

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।

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।

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।

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।

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।

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।

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।

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 ।

- Here it is used for perform comparison before and after fine-tuning.
- Before fine-tuning the words “*minor boys*” translated as “*छोटे लडके*”. 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 “*नाबाललग लडकों*”.

## Additional tasks which we tried to implement:

- We have trained a seq2seq model with attention from scratch. The encoder consists of GRU cells and decoder consists of attention with GRU.
- Attention is applied to the outputs of the encoder and is combined with the last hidden state, this also called as skip thought vector. This is then provided as input to the decoder to obtain the translations.
- The sentence pairs are selected such that the length is between 2 and 10 words.
- To obtain better convergence, teacher forcing is also used. 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.
- The BLEU score obtained on test data is **0.00853**.

```
> 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>
```

**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।']]

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

These are the sentence translated outputs of the seq2seq trained dataset.

Here the first sentence is the English input and next line is the model translated output and third line is the reference.

- 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.

```
> 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>
```

These are the sentence translated outputs of the seq2seq trained dataset.

Here the first sentence is the English input and next line is the model translated output and third line is the reference.

**Actual :**

[['Deputations meet Chief Minister', 'प्रतिनिधिमंडल मिले मुख्यमंत्री से'],

['Offline Account Disabled', 'ऑफ़लाइन — खाता निष्क्रिय']]

**Predictions :**

[['कम', 'की', 'की'], ['कम', 'का', 'की']]

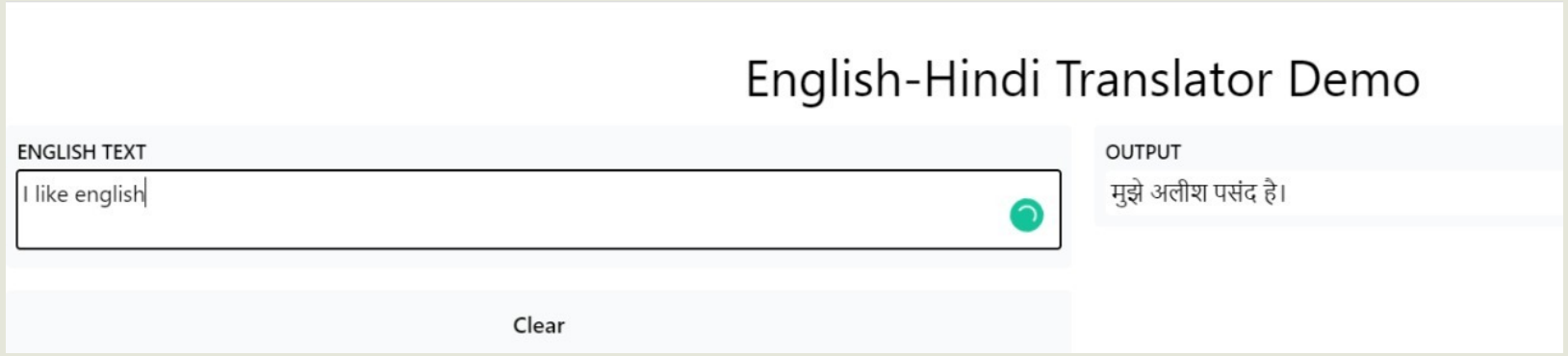
Here are the translations for some of the samples in test dataset

- From the samples shown in the slides, 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.

| Algorithm | Translation | BLEU Score |
|---|---|---|
| **Fine-tuned "helsinki-nlp/opus-mt-en-hi" transformer**<br>(6 layers in both encoder-decoder)<br>45000 train (en to hi)<br>5000 tests (en to hi) | En → Hi | **17.7773** |
| **Seq2seq – attention** (from scratch)<br>(1 layer gru in both encoder-decoder)<br>Samanantar english to indian languages.<br>45000 train (en to hi) 5000 test (en to hi) | En → Hi | 0.00853 |

- As part of viva, we have created a demo version of our model translation and its live on the link – https://50567.gradio.app

- Some of the samples are shown below:

## Bonus question

- In order to perform English to Hindi translations, the amazon-fine-food-reviews is obtained from Kaggle. As the number of reviews were huge (nearly 50K) only 200 reviews are considered to perform language translation.
- Translations are performed using fine-tuned transformer model from Part 1.
- 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.

| | |
|---|---|
| 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।'] |

| | |
|---|---|
| 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.<br />I have tried all of the flavors but really do like the grapefruit flavor... no after-taste and I actually like the slight carbonation.<br />I use other Hammer products and really like their whole product line. | ['यह उत्पाद मुझे लंबी या बाइक चलाने के बाद भी काम करता है.'] |
| 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 ।'] |

# Conclusion

- Fine-tuned "*helsinki-nlp/opus-mt-en-hi*" transformer with 6 layers in both encoder-decoder gave the better BLEU score compared to the other tasks.

- Seq2Seq model with attention from scratch has the encoder GRU cells and decoder consists of attention with GRU and the sentence pairs are selected such that the length is between 2 and 10 words. This is giving comparatively better score compared to the sentences whose length lies between 2 words to 5 words, along with using 2 GRU layers both in encoder and attention decoder.


THANK YOU!