Final Project Presentation:
Cross-lingual Summarization

Inderjeet Jayakumar Nair, 170020013
Saurabh Jayesh Parekh, 170100016
Manas Jain, 170040068

Problem Statement

Consider two Languages: Source Language(L_s) and Target Language(L_T). We are given with a sentence S_s as input from L_s which may be assumed to be grammatically valid. We define our problem statement as proposing a model that generates a sentence S_T(grammatically valid) as output in L_T such that

a) S_T semantically captures the most relevant / important section of S_s
b) The number of tokens in S_T is less than the number of tokens in a valid machine translated output of S_s from L_s to L_T

If $L_s = L_T$, the task becomes automatic summarization.

We want to perform this task using shared encoder-decoder architecture.

Dataset Information

a) **HindiEnCorp 2.0**:
   i) Hindi-English Parallel Corpora with 127,607 sentences
   ii) Pre-trained models with this dataset to learn the language model as the pretrained models for hindi were difficult to find
   iii) Hypothesis behind using the dataset: If the language model is learnt in the encoder and decoder, cross lingual summarization dataset can just fine tune the attention weights in order to capture important segments, thereafter the encoder with learned language model can just obtain valid output corresponding to the important segments.

b) **Daily News Dataset**
   i) News-headlines dataset in English with 90,000 training sentences and 4,515 test sentences
   ii) Machine translated headlines to Hindi by using free Google Translate API
   iii) The machine translation was done in a batch of 4,500 sentences since the API was unstable

Dataset Examples

a) HindiEnCorp 2.0:

i) Humans destroyed the commons that they depended on : मानवों ने उन ही साझे संसाधनों को नष्ट किया जिन पर वो आधारित थे।

ii) Premchand is the author of the Hindi literature era : प्रेमचंद हिन्दी साहित्य के युग प्रवर्तक हैं |

b) Daily News Cross lingual summarized

i) The police on Saturday arrested three persons from a gang of five alleged criminals in Greater Noida, after a shootout in which one of them was injured and two managed to flee. The police had acted on a tip-off and discovered later that the gang from Bulandshahr, was involved in a robbery in Noida's Kasna this month : ग्रेटर नोएडा में पुलिस के साथ गोलीबारी के बाद हुए 3 लुटेरे

ii) Finance Minister Arun Jaitley on Friday said that the financial year may soon begin from January in sync with the calendar year. ""The matter of changing financial year is under consideration of the government,"" he told the Lok Sabha. In May, Madhya Pradesh became India's first state to shift its financial year format to January-December from the present April-March cycle.: जनवरी से शुरू हो सकता है वित्तीय वर्ष: जेटली

Approaches:

a) Sequence-2-Sequence GRU with attention mechanism
b) Transformer Architecture
c) Multi-Tasking Transformer Architecture

Seq-2-seq GRU with attention

Trained on our dataset

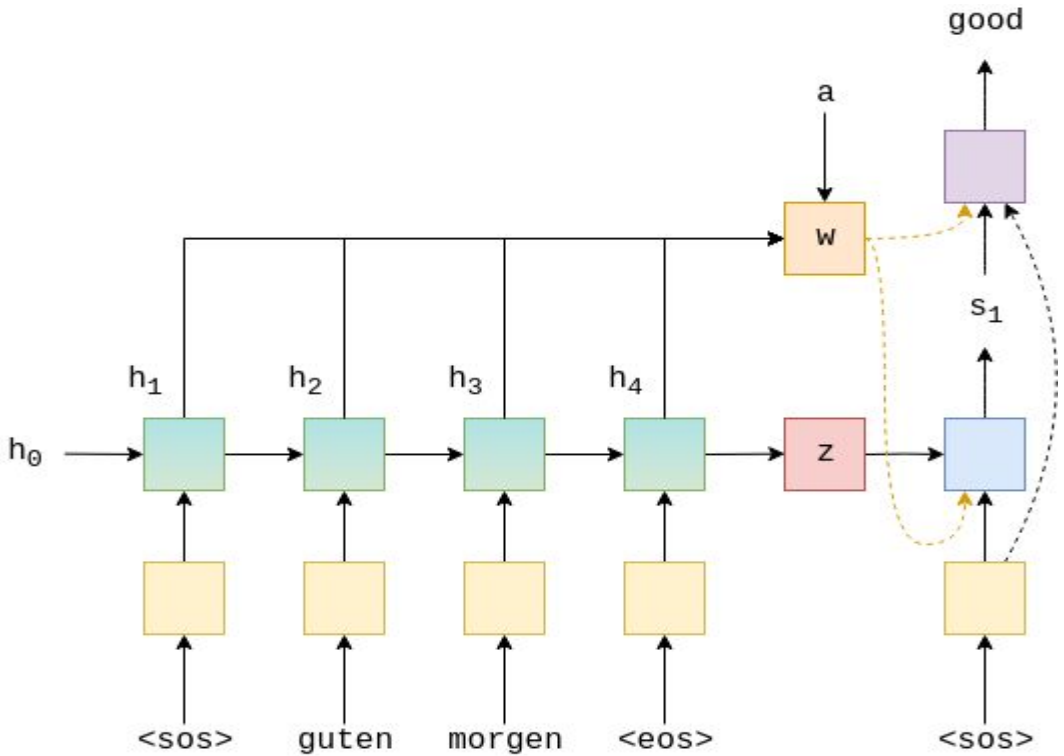
Similar to machine translation task

Encoder

We use a single layer GRU, however we now use a *bidirectional RNN*. With a bidirectional RNN, we have two RNNs in each layer. A *forward RNN* going over the embedded sentence from left to right (shown below in green), and a *backward RNN* going over the embedded sentence from right to left (teal).

Attention

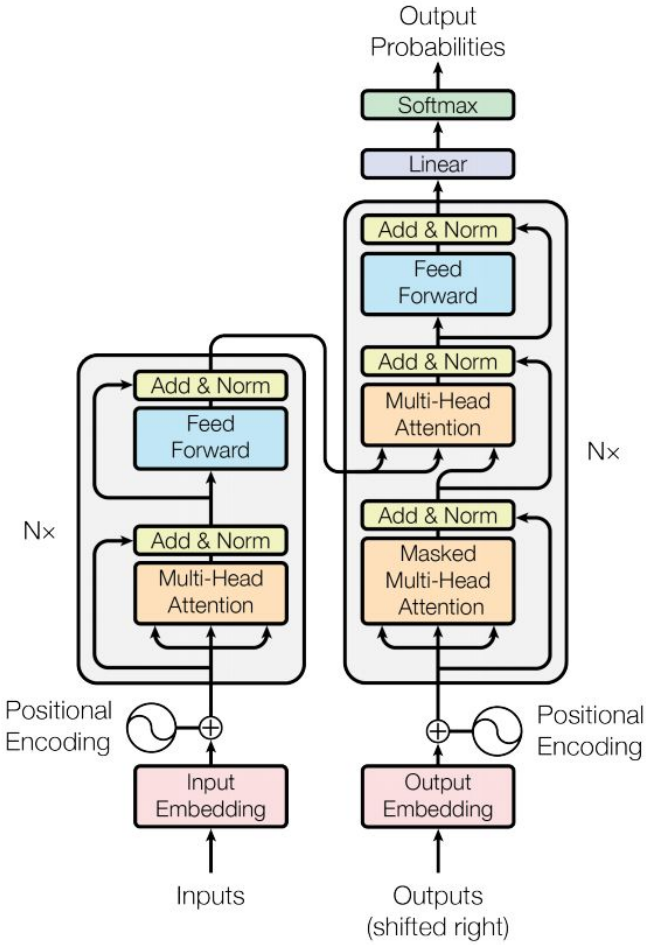
This will take in the previous hidden state of the decoder and all of the stacked forward and backward hidden states from the encoder, The layer will output an attention vector, , that is the length of the source sentence, each element is between 0 and 1 and the entire vector sums to 1. Intuitively, this layer takes what we have decoded so far, , and all of what we have encoded to produce a vector that represents which words in the source sentence we should pay the most attention to in order to correctly predict the next word to decode



Transformer Architecture

- Pretrained on HindiEnCorp 2.0
- Trained on Daily News
- Total Pretraining time = 3 days
- Total Training time = 3 hrs

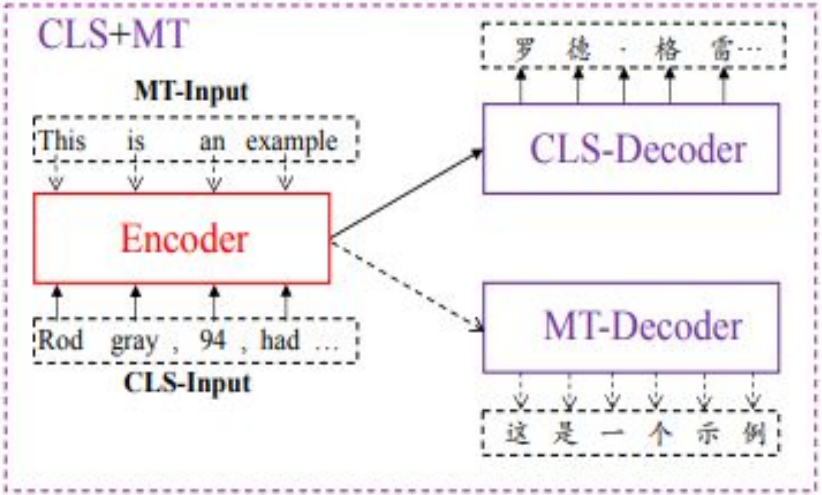
Performance: Bleu Score = 0.4136



Multi-Tasking Transformer Architecture

In this approach, a multi-task loss function objective is set up to train cross-lingual summarizer. The transformer consists of one encoder and two decoders for the task of machine translation and cross-lingual summarization. Since MT can be regarded as a special case of CLS with compression ratio 1:1. Jointly training the model makes the encoder highly specialized in language understanding.

- Implemented the model from scratch in Pytorch
- Jointly trained on HindiEnCorp 2.0 MT dataset and Daily News CLS dataset
- The model is still under training due to the requirement of high computational power for such specialized architecture
- Achieved 0.3406 BLEU score when just trained for 3 epoch



Some examples from the Trained Transformer Model

Input: Four labourers on Monday were reportedly injured after a tree branch fell on them at Dombivli station road in Mumbai. They were admitted to hospital with injuries and were later declared out of danger. Reportedly, tree fall cases are on rise in Kalyan-Dombivli. Last year fewer cases were reported. We have been getting complaints of tree falls daily.

Output: मुंबई में पेड़ की शाखा गिरने से 4 मजदूर घायल हो गए

Input: The Bombay High Court on Monday summoned the Maharashtra Women and Child Development Secretary after 42 children went missing over the last three years from a Mumbai remand home. The court criticised the Maharashtra government for lack of 'proactive action' in the matter. The Bombay High Court is hearing a PIL on the allegations of corruption in the remand home.

Output: 42 बच्चों के घर से बाहर निकलने के बाद बॉम्बे <unk> ने सचिव को बुलाया

Some examples from the Trained Transformer Model

Input: As many as 76 passengers were rescued from cable cars suspended over a river in German city Cologne after a gondola crashed into a support pillar on Sunday. Passengers were left stranded, and children were seen clinging to parents while dangling as many as 40 metres above the river. The fire department lowered them to safety from the cable cars

Output: केबल कार <unk> से उतरने के बाद 76 यात्री निलंबित

Input: An 11-year-old tribal boy allegedly committed suicide on Tuesday by hanging himself near his school, after he was caught stealing ?30 from his classmate in Maharashtra's Mokhada. The boy was reportedly ashamed of his act and had tried to force a classmate to commit suicide with him, but he refused. Police said the boy has a history of criminal activities.

Output: 19 - वर्षीय आदिवासी लड़के ने <unk> पकड़ा , 30 से 30 पकड़े जाने के बाद आत्म हत्या कर ली

Challenges

- **Dataset Generation**: Unable to find dataset for English-Hindi CLS
- **Limited Computation power**:
  - This made us use lesser number of words in the vocabulary as the number of parameters increases with the size of the vocabulary. Thus in many of the previous examples we obtained <UNK> token corresponding to unknown word.
  - Huge training time is required for seq2seq and Multi task Transformer and thus they are being trained
  - We are unable to use the CFILT parallel Hindi English Corpus for pretraining due to large memory requirements
- **Google's Rate Limiter**: The free API makes several HTTP requests to the Google's machine translation server to retrieve the machine translated output. As a result several times, our IP was blocked from using Google Services
- **Unstable Session of Colab**: When the code is executing for a large period, the session collapses to limit the usage. Colab prevents us from over using their GPU resources by disabling our access to GPU