Adversarial Transformer Network for Machine Translation

Akhila Josyula, Vatsal Shah and Swarup Satish

ajosyula@cs.umass.edu, vhshah@cs.umass.edu, ssatish@cs.umass.edu

University of Massachusetts, Amherst

Problem Statement

We look to perform Machine Translation from German to English using the IWSLT 2016 DE-EN dataset. Machine translation has numerous practical applications in the areas of movie subtitle generation, live speech translation, assistance to human translators. We aim to train an MT model using adversarial training, wherein our Generator and Discriminator are both self-attention based networks.

Dataset

We are using the International Workshop on Spoken Language Translation (IWSLT) 2016 de-en dataset. Some of the statistics for the data are provided below.

- Number of sentences pairs : 196884
- Number of German tokens : 3130282
- Number of English tokens : 3371117
- Average length a German sentence : 15 tokens
- Average length of an English sentence : 17 tokens

Further exploration of the dataset revealed that only 338 sentence pairs were of length greater than 100 (0.17% of the dataset). Thus, we omitted them, and only included sentences of length less than or equal to 100.

Our Approach and Baseline

- The baseline model is a sequence to sequence 2 layer LSTM with encoder attention on decoder. Each layers consist of 500 hidden-size and a dropout of 0.3.
- Our adversarial model as in figure 1 consists of a transformer network as the generator. It consists of a 6 layer encoder decoder network where each layer is a 8 headed self attention with the decoder also receiving source side attention as in figure 2.
- The discriminator is made up of a self-attention based classifier which discriminates between machine and human generated sentences.
- We make use of a cross entropy loss for the generator output to make it produce translations close to the ground truth, as well as the discriminator objective to make it produce more human-like sentences.

Our Contributions

- To make adversarial training stable, we need to pretrain the Generator and Discriminator.
- We currently have a trained Transformer model which achieves a BLEU score of 26.8 on the dev set.
- We are now implementing the adversarial training for the model.
- To produce a sentence from the Generator output, we need to sample/greedy decode from the distribution produced by the Generator. This step is non-differentiable, and hence it is difficult to backprop gradients into the Generator, from the loss of the Discriminator.
- We are looking at methods to get around this.

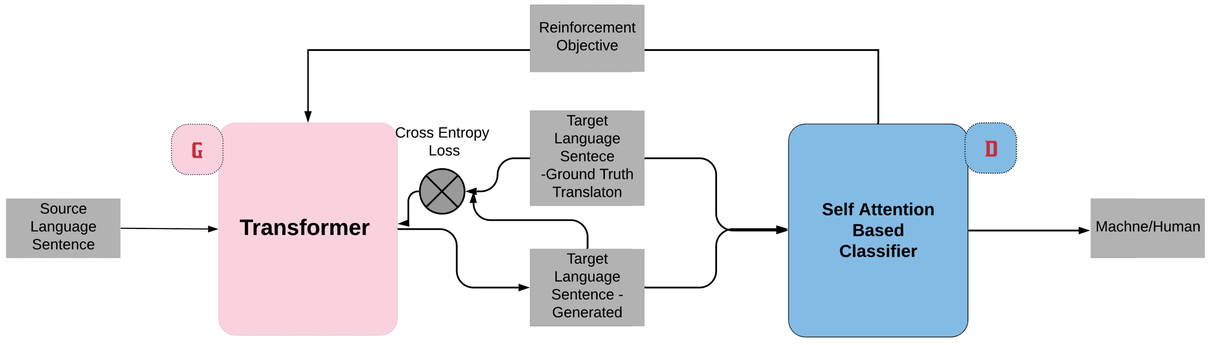

Figure 1: System architecture

Ground truth translation	LSTM output	Transformer output
It's Christmas Eve, you're at the mall, you're driving around in circles looking for parking, and what do you see?	And it 's a short time before Christmas in the 2030s , . room and what looks like a parking lot of the search	It 's Christmas Eve , you 're at the mall , you 're driving around in circles looking for parking , and what do you see ?
If all the answers are yes , I feel as though I've won .	When all the answers are Ja , I feel like I won as if I won .	If all the answers are loud , I feel like I won .

Table 1: Comparing outputs from our baseline model and Transformer

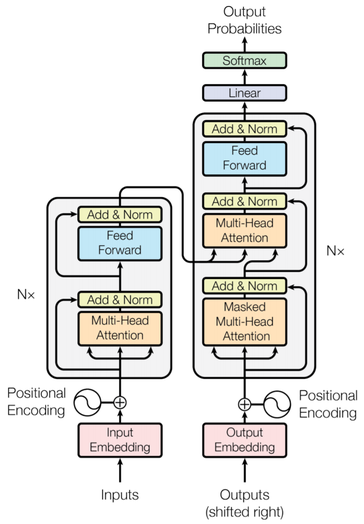

Figure 2: Transformer architecture

Experiments & Analysis

- We currently have two models, a baseline LSTM sequence-sequence model, and a Transformer model.
- On the dev set, the baseline model achieved a BLEU score of 16.3, while the Transformer model achieved a BLEU score of 26.8.
- The transformer model can generate sentences which preserve structure and meaning of the ground-truth sentence much better than the LSTM model as can be seen in the first output of Table 1.
- From the second example in Table 1, we can see that although transformer does better than LSTM based model, it still doesn't completely preserve meaning. We expect to see mitigate this using our adversial architecture.