Assignment 2: E1-246
Seq2Seq Model for Machine Translation

Aadesh Magare
aadeshmagare@iisc.ac.in

Abstract

This is a report for assignment 2 of E1-246 course, Natural Language Understanding. The task was to implement a machine translation system for English to German and English to Hindi translations. The report contains experimental details and observations.

1 Introduction

A Seq2Seq model with attention was implemented for the translation task. several attention mechanisms like scaled dot product, multiplicative, additive and key value attention were tried. The report contains details about the experiments and their comparison. further self attention for encoder and decoder was also explored and compared against.

2 Datasets

The English-German dataset was taken from WMT14 (WMT), which included Europarl v7, Common Crawl, News Commentary sources. It had over 4.5 M lines of text, out of which 0.15M were dropped for being exceedingly long length ($>$20 words). English-Hindi dataset was taken from (ufa) which had total 0.25 M lines of text out of which around 9k were dropped ($>$ 30 words). Developement and Test datasets were taken from WMT14. Corpus BLEU score is used as evaluation metric.

3 Model

The starter kit for project is taken from Stanford CS224n. (224) The Seq2Seq model has bidirectional LSTM encoder and a unidirectional LSTM decoder. The final model is trained with parameters in Table 1.

In the interest of training time En-Hi model is trained for 5 epocs while En-De model is trained for 1 epoc, thus the BLEU score is less than what is expected but it should suffice for our comparison task.

Parameter	Value
Batch Size	32
Hidden Size	300
Embedding Dimensions	300

Table 1: Final Model Parameters

4 Task 1: En-Hi Translation

4.1 Multiplicative Attention

The model with multiplicative attention trained for 5 epocs achieves a BLEU score of 0.6138. Training loss during the process initially drops rapidly to the level of 40 - 35 as seen in Figure 1.

Figure 1: Initial Training Loss (Multiplicative)

The loss further keeps dropping at a slow rate. Similar behaviour is seen for Perplexity, it rapidly drops to the level of 500-450 as seen in Figure 2 and further keeps dropping at slow rate.

4.2 Scaled Dot Product Attention

The model with scaled dot product attention trained for 5 epocs achieves a BLEU score of

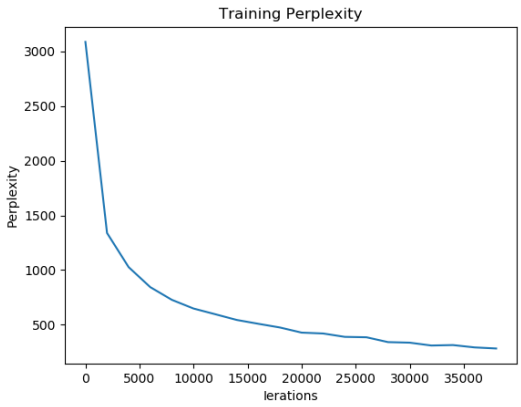

Figure 2: Initial Perplexity (Multiplicative)

2.1176. Training loss during the process initially drops rapidly to the level of 40 - 30 as seen in Figure 3.

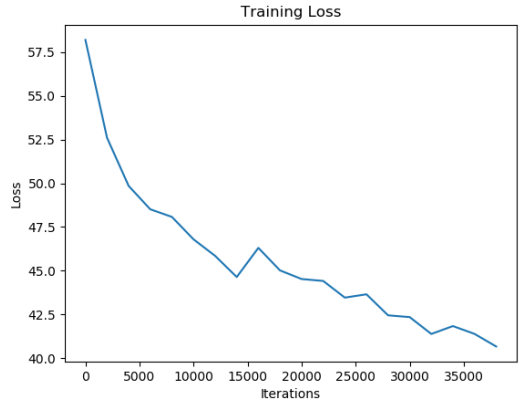

Figure 3: Initial Training Loss (Dot Product)

The loss further keeps dropping at a slow rate. Similar behaviour is seen for Perplexity, it rapidly drops to the level of 450-400 as seen in Figure 4 and further keeps dropping at slow rate.

4.3 Additive Attention

The model with additive attention trained for 5 epocs achieves a BLEU score of 0.8248. Training loss during the process initially drops rapidly to the level of 55 - 50 as seen in Figure 5.

The loss further keeps dropping at a slow rate. Similar behaviour is seen for Perplexity, it rapidly drops to the level of 250-200 as seen in Figure 6 and further keeps dropping at slow rate.

4.4 Key Value Attention

The model with key value attention trained for 5 epocs achieves a BLEU score of 0.6174. Training

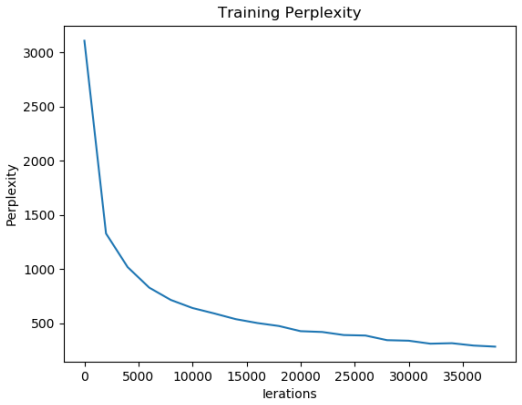

Figure 4: Initial Perplexity (Dot Product)

Figure 5: Initial Training Loss (Additive)

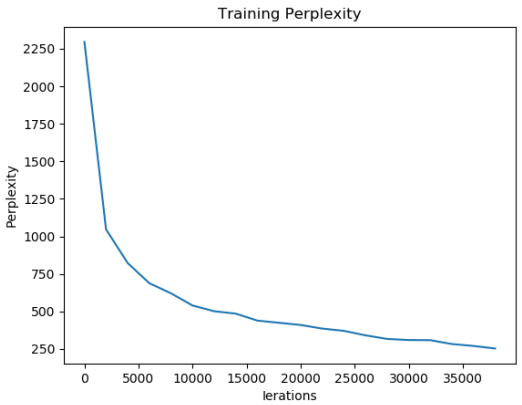

Figure 6: Initial Perplexity (Additive)

loss during the process initially drops rapidly to the level of 50 - 40 as seen in Figure 7.

Figure 7: Initial Training Loss (Key-Value)

The loss further keeps dropping at a slow rate. Similar behaviour is seen for Perplexity, it rapidly drops to the level of 400-300 as seen in Figure 8 and further keeps dropping at slow rate.

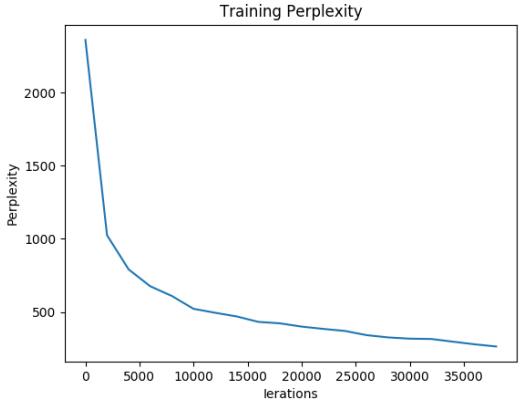

Figure 8: Initial Perplexity (Key-Value)

Performance of attention schemes for the task can be ordered as:
Scaled Dot Product > Additive > Key Value > Multiplicative

5 Task 2: En-De Translation

5.1 Multiplicative Attention

The model with multiplicative attention trained for 1 epoc achieves a BLEU score of 0.1428. Training loss during the process initially drops rapidly to the level of 50 - 45 as seen in Figure 9.

The loss further keeps dropping at a slow rate. Similar behaviour is seen for Perplexity, it rapidly drops to the level of 230-200 as seen in Figure 10 and further keeps dropping at slow rate.

Figure 9: Initial Training Loss (Multiplicative)

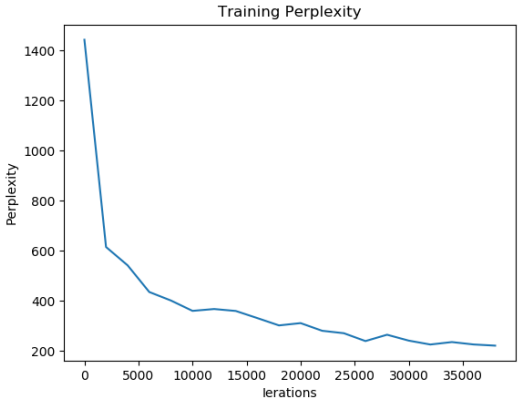

Figure 10: Initial Perplexity (Multiplicative)

5.2 Scaled Dot Product Attention

The model with scaled dot product attention trained for 1 epoc achieves a BLEU score of 5.529. Training loss during the process initially drops rapidly to the level of 40 - 35 as seen in Figure 11.

Figure 11: Initial Training Loss (Dot Product)

The loss further keeps dropping at a slow rate. Similar behaviour is seen for Perplexity, it rapidly drops to the level of 220-200 as seen in Figure 12 and further keeps dropping at slow rate.

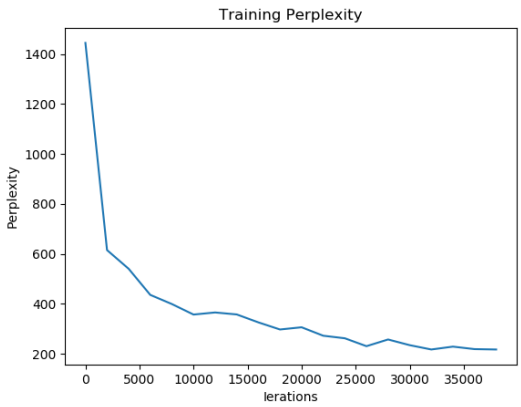

Figure 12: Initial Perplexity (Dot Product)

5.3 Additive Attention

The model with additive attention trained for 1 epoc achieves a BLEU score of 0.2283. Training loss during the process initially drops rapidly to the level of 45 - 40 as seen in Figure 13.

Figure 13: Initial Training Loss (Additive)

The loss further keeps dropping at a slow rate. Similar behaviour is seen for Perplexity, it rapidly drops to the level of 220-200 as seen in Figure 14 and further keeps dropping at slow rate.

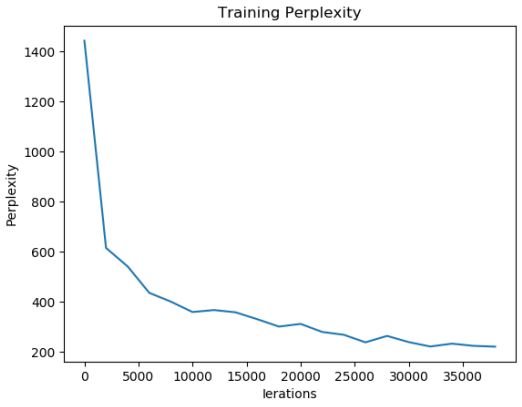

Figure 14: Initial Perplexity

5.4 Key Value Attention

The model with key value attention trained for 1 epoc achieves a BLEU score of 0.1294. Training loss during the process initially drops rapidly to the level of 40 - 35 as seen in Figure 15.

Figure 15: Initial Training Loss

The loss further keeps dropping at a slow rate. Similar behaviour is seen for Perplexity, it rapidly drops to the level of 220-200 as seen in Figure 16 and further keeps dropping at slow rate.

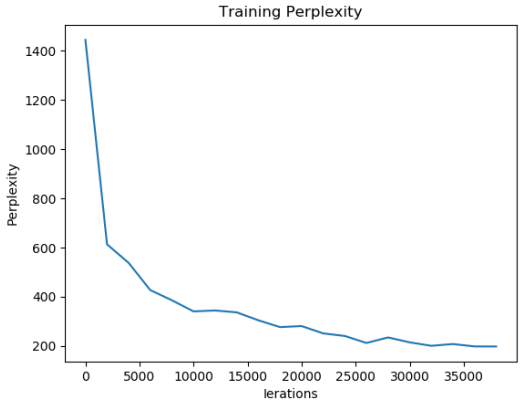

Figure 16: Initial Perplexity

Performance of attention schemes for the task can be ordered as:
Scaled Dot Product > Additive > Multiplicative > Key Value

6 Self-Attention:

Self-attention was implemented for encoder and decoder for the best performing model from previous sections. The training loss and perplexity follows similar expected behaviour as earlier where it initially drops rapidly and then slow down towards end of training. Self-attention however improves the BLEU score of model and shows a 1.2537 point improvement for En-Hi translation task and 0.13 point improvement for En-De task.

Repository:

The project is available at (Git) which contains the code for experiments and additional plots obtained during the process.

References

Cs224n: Natural language processing with deep learning.

Hindencorp: Hindi-english and hindi-only corpus.

Seq2seq nlu assignment 2.

Wmt14: Machine translation task.