ECGR 5106 – Real Time Machine Learning (Spring 2023)

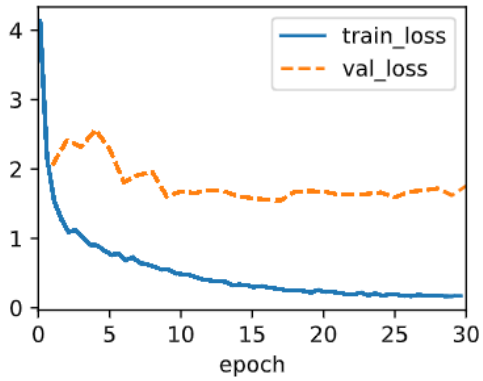
Nahush D. Tambe – 801060297

https://github.com/Ntambe25

Homework # 6

Problem 1:

For Problem # 1, the baseline Transformer model for Machine Translation problem was plotted using the code from d2l book and the lecture notes. Figure 1 below shows the same. The baseline model used 256 hidden units, 2 transformer blocks, 64 hidden units in the feed-forward neural network, and 4 attention heads in the multi-headed attention mechanism. After plotting the baseline model, a few more experiments were carried out with a deeper transformer model by changing the number of transformer blocks from 2 to 4, 6, and 8. Figures 2 through 8 show the results consisting of the model plots, their Training and the results of those trained model used to translate few English sentences into French.




```
Training Time (Baseline Transformer Model):  29.949615001678467
```

Figure 1: Results for Baseline Transformer Model for Machine Translation Problem

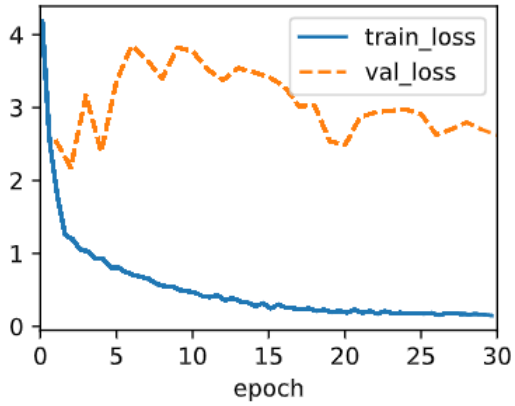
As mentioned earlier, Figure 1 above shows the plot for baseline Transformer model for Machine Translation problem. It can be clearly seen from the plot that the training loss starts at around 4 and goes down to almost 0.2. The validation loss on the other hand, starts out at 2, goes to around 2.75, comes down to about 1.75 and then is stable at 1.75 all throughout.

The plot does show a significant amount of generalization gap. The training time for this model was about 30 seconds.

```
go . => ['va', '!'], bleu,1.000
i lost . => ['je', 'perdu', '.'], bleu,0.687
he's calm . => ['il', 'est', 'calme', '.'], bleu,1.000
i'm home . => ['je', 'suis', 'chez', 'moi', '.'], bleu,1.000
```

Figure 2: Translation Results for Trained Baseline Transformer Model

Figure 2 above shows the translation results for baseline Transformer Model for Machine Translation Problem and their BLEU scores. For ¾ of the sentences, the score was an 100%, while for one of the sentences it was around 69%. Overall, the model has been trained well.




```
Training Time (Deeper Transformer Model (4 Blks)):  43.20629405975342
```

Figure 3: Results for Deeper Transformer Model with 4 Transformer Blocks

Figure 3 above shows the plot for a deeper Transformer model for Machine Translation problem with 4 transformer blocks instead of 2. It can be clearly seen from the plot that the training loss starts at around 4 and goes down to almost 0.2. The validation loss on the other hand, starts out at 2.5, goes to around 4.0, and then finally comes down at around 3.0. Compared to the baseline model, the training loss showed the same trend, but the validation loss was quite high.

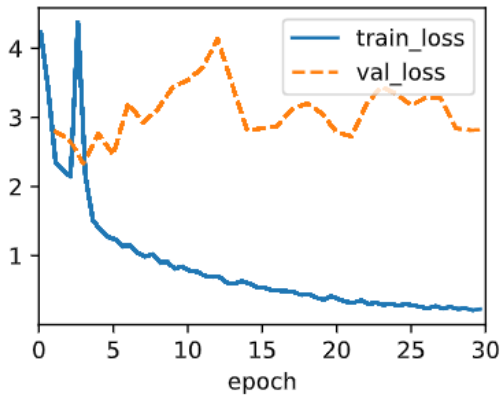
The plot does show a significant amount of generalization gap. The training time for this model was about 43 seconds, which is high compared to the baseline model, which was expected with an increased number of transformer blocks.

```
go . => ['va', '!'], bleu,1.000
i lost . => ["j'ai", 'perdu', '.'], bleu,1.000
he's calm . => ['il', 'est', 'mouillé', '.'], bleu,0.658
i'm home . => ['je', 'suis', 'chez', 'moi', '.'], bleu,1.000
```

Figure 4: Translation Results for Trained Deeper Transformer Model with 4 Transformer Blocks

Figure 4 above shows the translation results for deeper Transformer Model for Machine Translation Problem with 4 transformer blocks and their BLEU scores. For ¾ of the sentences,

the score was an 100%, while for one of the sentences it was around 66%. Overall, the model has been trained well, but the baseline mode showed a slightly better result.




```
Training Time (Deeper Transformer Model (6 Blks)):  46.027830839157104
```

Figure 5: Results for Deeper Transformer Model with 6 Transformer Blocks

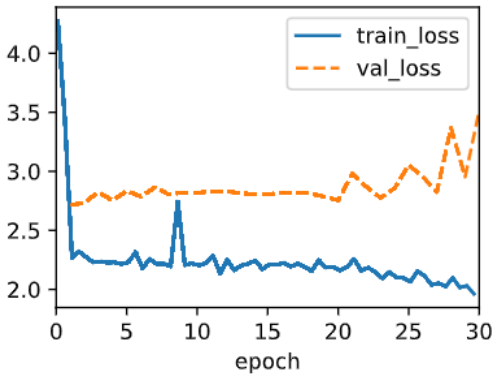
Figure 5 above shows the plot for a deeper Transformer model for Machine Translation problem with 6 transformer blocks instead of 2. It can be clearly seen from the plot that the training loss first starts at around 4, comes down to about 2, goes back to 4.25, and then finally shows a decreasing trend going down to almost 0.2. The validation loss on the other hand, starts out at 2.5, goes to around 4.0, and then finally comes down at around 3.0. Compared to the baseline and deeper model with 4 transformer blocks, the training loss showed a more unstable trend, but the validation loss was around the same as the deeper model with 4 transformer blocks.

The plot does show a significant amount of generalization gap. The training time for this model was about 46 seconds, which is high compared to the baseline model, which was expected with an increased number of transformer blocks.

```
go . => ['va', '!'], bleu,1.000
i lost . => ["j'ai", 'perdu', '.'], bleu,1.000
he's calm . => ['<unk>', '.'], bleu,0.000
i'm home . => ['je', 'suis', 'chez', 'moi', '.'], bleu,1.000
```

Figure 6: Translation Results for Trained Deeper Transformer Model with 6 Transformer Blocks

Figure 6 above shows the translation results for deeper Transformer Model for Machine Translation Problem with 6 transformer blocks and their BLEU scores. For ¾ of the sentences, the score was an 100%, while for one of the sentences it was 0. Overall, the model showed mediocre results and is not great as the two earlier models.




```
Training Time (Deeper Transformer Model (8 Blks)):  39.53871965408325
```

Figure 7: Results for Deeper Transformer Model with 8 Transformer Blocks

Figure 7 above shows the plot for a deeper Transformer model for Machine Translation problem with 8 transformer blocks instead of 2. It can be clearly seen from the plot that the training loss first starts at around 4, comes down to about 2.75, then is almost stable at the same value, and 1astly ends at 2.0. The validation loss on the other hand, starts out at 2.5 and slowly goes to about 3.5. Compared to the baseline and deeper model with 4 and 6 transformer blocks, the training loss showed a more unstable trend, but the validation loss was around the same as the deeper model with 4 transformer blocks, but also quite stable compared to the deeper model with 6 transformer blocks.

The plot does show a significant amount of generalization gap. The training time, surprisingly taking the fact into consideration that this model was significantly deeper than the baseline model, was only about 39.5 seconds, which is comparatively lower than all the previous models.

```
go . => ['<unk>', '<unk>', '<unk>', '<unk>', '<unk>', '<unk>', '<unk>', '<unk>', '<unk>'], bleu,0.000
i lost . => ['<unk>', '<unk>', '<unk>', '<unk>', '<unk>', '<unk>', '<unk>', '<unk>', '<unk>'], bleu,0.000
he's calm . => ['<unk>', '<unk>', '<unk>', '<unk>', '<unk>', '<unk>', '<unk>', '<unk>', '<unk>'], bleu,0.000
i'm home . => ['<unk>', '<unk>', '<unk>', '<unk>', '<unk>', '<unk>', '<unk>', '<unk>', '<unk>'], bleu,0.000
```

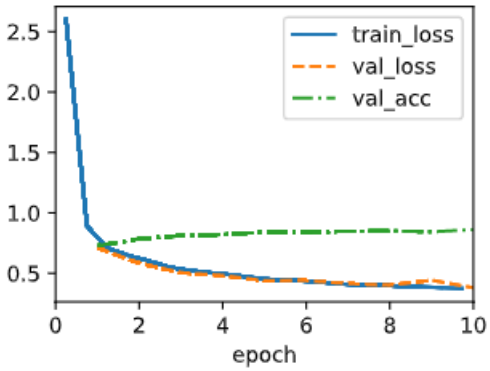
Figure 8: Translation Results for Trained Deeper Transformer Model with 8 Transformer Blocks

Figure 8 above shows the translation results for deeper Transformer Model for Machine Translation Problem with 8 transformer blocks and their BLEU scores. For all of the sentences, the score was not calculated itself. Looking at the scores, the model performed the worst, compared to all the earlier models.

In conclusion, as the number of transformer blocks increased, the results were more and more worst and the model with 8 transformer blocks was not able to translate any of the English sentences.

Problem 2:

For Problem # 2, first, the baseline Vision Transformer model was plotted using the code from d2l book and the lecture notes. The baseline model used 2 multiheaded self-attention blocks. After plotting the baseline model, the same code was used, but instead of 2 self-attention blocks, the model was plotted using 3 and 6 blocks. Figures 9 through 11 below show the results for the same.



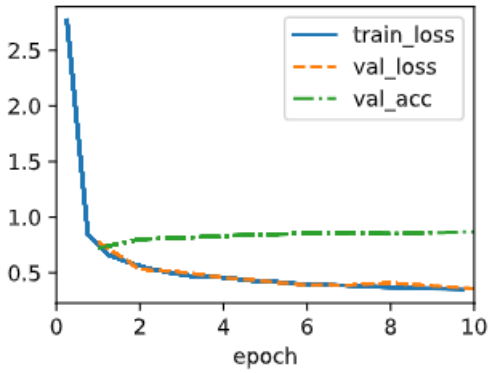

```
Training Time (Baseline Vision Transformer):  313.9881868362427

Number of FLOPs: 247.2 MMac
Number of parameters: 6.72 M
```

Figure 9: Results for Baseline Vision Transformer Model

It can be seen from the plot that the training loss starts around 2.5 and goes down to almost 0.25. The validation loss starts at around 0.75 and also goes down to about 0.25. The validation accuracy on the other hand is almost stable, only increasing slightly from about 0.75 to almost 0.9. The training and validation loss do not show a sign of a generalization gap. Figure 4 above shows the same results.

It took about 314 seconds to train this baseline Vision Transformer model with 2 multiheaded self-attention blocks. The number of FLOPS needed for this model were around 247 MMAC, while the total number of parameter count was 6.72 M.



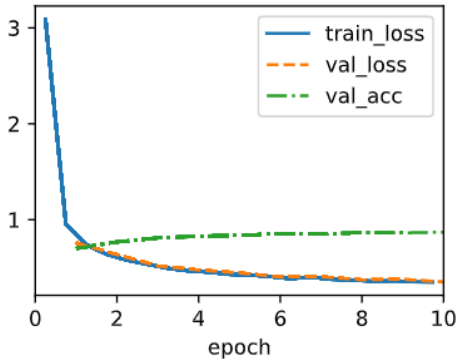

```
Training Time (Vision Transformer (3 Blks)):  471.20379614830017
```

```
Number of FLOPs: 354.27 MMac
Number of parameters: 9.61 M
```

Figure 10: Results for Deeper Vision Transformer Model (3 Blocks)

It can be seen from the plot that the training loss starts just a bit higher than the baseline model, from around 3.0 and goes down to almost 0.25. The validation loss starts at around 0.75 and also goes down to about 0.25. The validation accuracy on the other hand is almost stable, only increasing slightly from about 0.75 to almost 0.9. The training and validation loss do not show a sign of a generalization gap. Figure 10 above shows the same results. Except for some minor changes, the plot is almost identical to the plot of baseline Vision Transformer model.

It took about 471 seconds to train this deeper Vision Transformer model with 3 multiheaded self-attention blocks. The number of FLOPS needed for this model were around 354 MMAC, while the total number of parameter count was 9.61 M. Compared to the baseline Vision Transformer model, this model has a significantly higher number of FLOPS and parameter count, but with a higher count, the accuracy did not improve, rather stayed almost the same.




```
Training Time (Vision Transformer (6 Blks)):  904.2825357913971
```

```
Number of FLOPs: 703.79 MMac
Number of parameters: 19.06 M
```

Figure 11: Results for Deeper Vision Transformer Model (6 Blocks)

It can be seen from the plot that the training loss, validation loss and the validation accuracy showed almost the same trends as the earlier baseline and deeper model with 3 blocks. Figure 1 above shows the same results. Except for some minor changes, the plot is almost identical to the plot of baseline and the earlier Vision Transformer model.

It took about 904 seconds to train this deeper Vision Transformer model with 6 multiheaded self-attention blocks. The number of FLOPS needed for this model were around 704 MMAC, while the total number of parameter count was 19.06 M. Compared to the baseline Vision Transformer model, this model has a significantly higher number of FLOPS and parameter count, but with a higher count, the accuracy did not improve, rather stayed almost the same.

In conclusion, with an increased number of multiheaded self-attention blocks, the accuracy and overall performance of the model did not grow.
