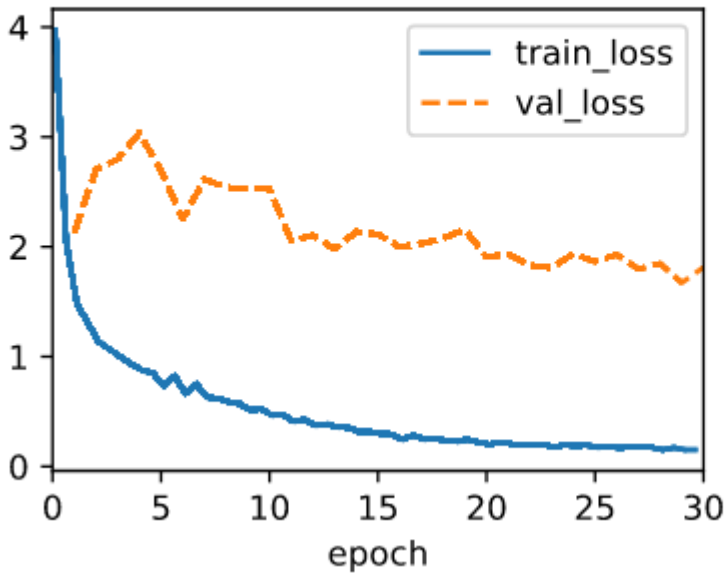
https://github.com/Christian-Martens-UNCC/ECGR-4106/tree/main/Homework_6-NLP_&_Vision_Transformers

Problem 1:

A) By altering the number of heads and encoder blocks, we are able to increase the complexity of the model compared to the model shown in the lecture example. Below are some of the specific results which are more interesting. Overall, I did not find that, for this dataset, increasing the model's complexity added any significant benefit to accuracy to the model. The BLEU scores and answers all seemed to be equivalent to one another and increasing the complexity only increased the training time. However, all of these models were much more accurate than the models trained for homework 5.

Model 1.4 – 4 Heads, 2 Blocks

```
Total Training Time : 54.5228 s
Estimated Average Training Time per Epoch : 1.81743 s
```

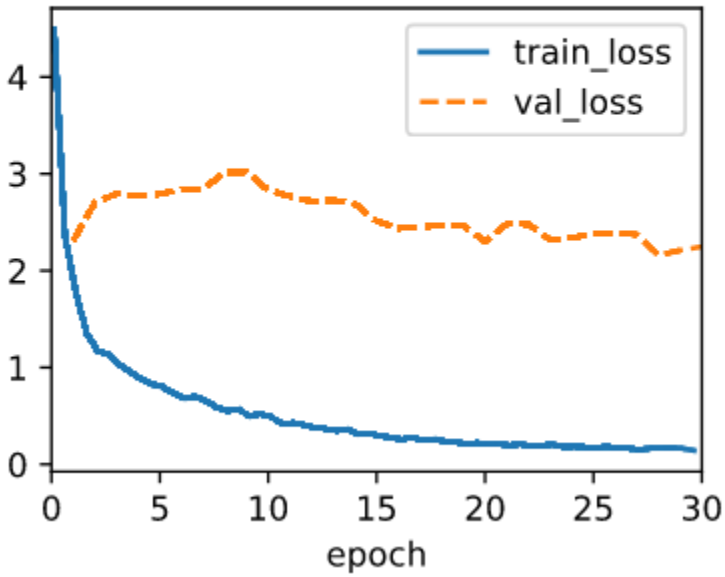



```
go . => ['va', '!'], bleu,1.000
i lost . => ["j'ai", 'perdu', '.'], bleu,1.000
he's calm . => ['il', 'est', 'mouillé', '.'], bleu,0.658
i'm home . => ['je', 'suis', 'chez', 'moi', '.'], bleu,1.000
```

This is an equivalent model to the one shown in the lecture examples. It's validation loss is around 2 and it trained in about 54 seconds.

Model 1.7 – 8 Heads, 2 Blocks

```
Total Training Time : 50.91092 s
Estimated Average Training Time per Epoch : 1.69703 s
```

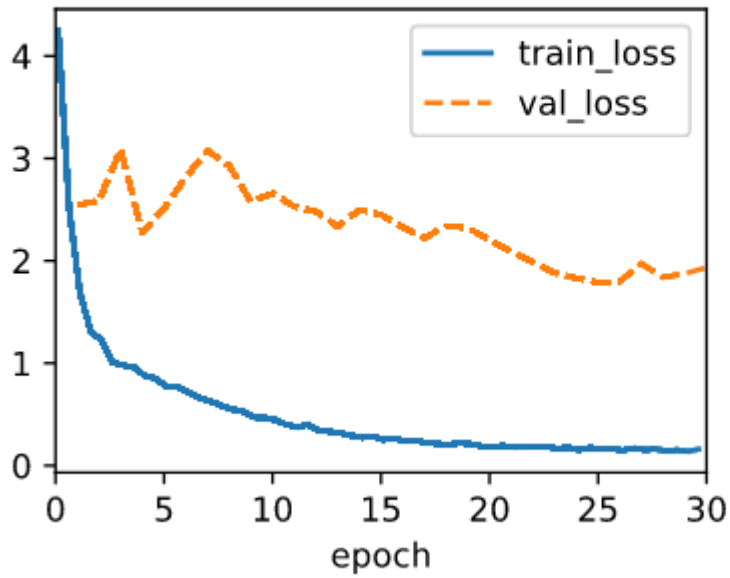



```
go . => ['va', '!'], bleu,1.000
i lost . => ["j'ai", 'perdu', '.'], bleu,1.000
he's calm . => ['<unk>', '.'], bleu,0.000
i'm home . => ['je', 'suis', 'chez', 'moi', '.'], bleu,1.000
```

This was a more complex model that I tested using 8 heads as opposed to 4. As shown above, the model trained faster but had a higher loss. The resulting BLEU scores also aren't as good as the default model.

Model 1.9 – 8 Heads, 4 Blocks

```
Total Training Time : 87.00588 s
Estimated Average Training Time per Epoch : 2.9002 s
```

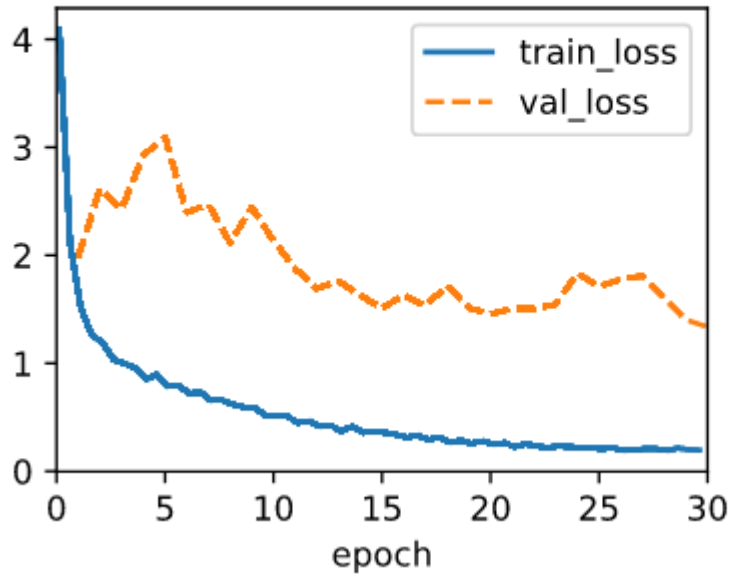



```
go . => ['va', '!'], bleu,1.000
i lost . => ["j'ai", 'perdu', '.'], bleu,1.000
he's calm . => ['il', 'est', 'mouillé', '.'], bleu,0.658
i'm home . => ['je', 'suis', 'chez', 'moi', '.'], bleu,1.000
```

The most complex model that was trained. As seen above, the model took over 1.5 times as much time to train compared to the base model but has nearly identical loss and BLEU scores.

Model 1.1 – 2 Heads, 2 Blocks

```
Total Training Time : 49.71535 s
Estimated Average Training Time per Epoch : 1.65718 s
```




```
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
```

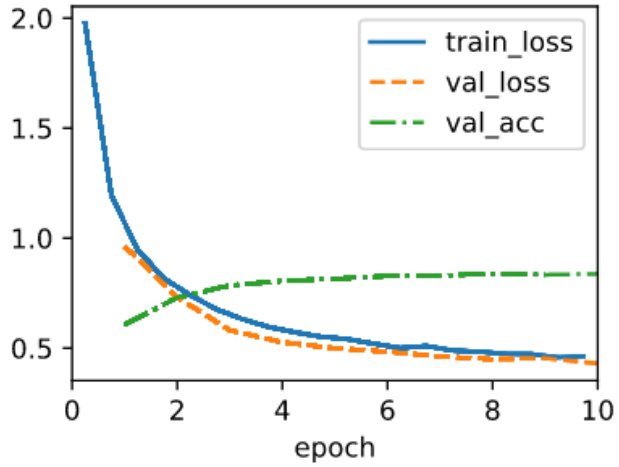
The first model that I trained and also a simpler model than the lecture. This model, surprisingly, had a lower loss than any other model that was trained at around 1.5. The lower loss and reduced training time led me to conclude that, for this application, this model was the best model that I trained.

Problem 2:

A) Using the d2l library, I altered the lecture model to become more complex. Due to the hardware specifications of my laptop, I had to reduce the number of hidden nodes to reduce training time from hours to minutes. However, all other parameters (other than the ones I am testing) stayed the same so the comparison is still viable. The resulting 4 graphs are below. As with problem 1, I found that the increase in complexity only, if at all, slightly increased the validation accuracy of the model, which was generally not worth the additional training time.

Model 2.1 – 8 Heads, 2 Blocks

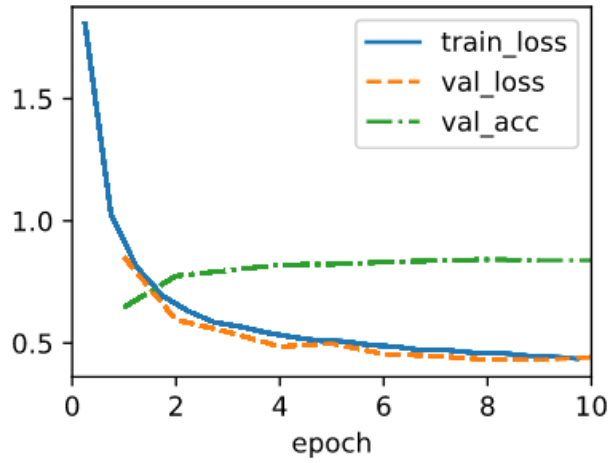
```
Total Training Time : 723.82303 s
Estimated Average Training Time per Epoch : 72.3823 s
```



This model is a replica of the model in the lecture. The validation accuracy is around 80% as well as the shortest training time at 72 sec/epoch

Model 2.2 – 8 Heads, 3 Blocks

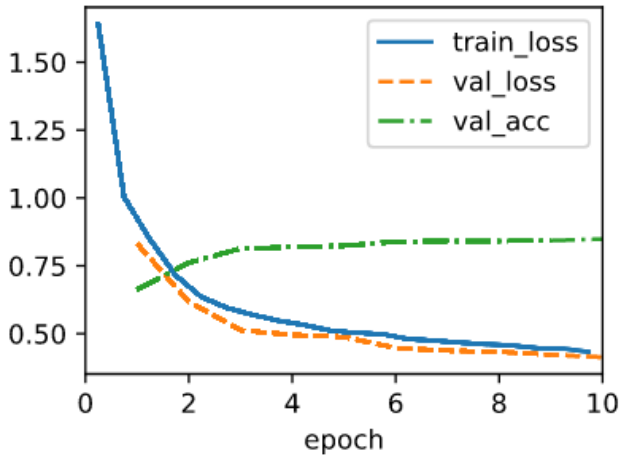
```
Total Training Time : 867.40374 s
Estimated Average Training Time per Epoch : 86.74037 s
```



This model was more complex, took longer to train, and had lower validation accuracy. Overall, this model was worse than the first on all accounts

Model 2.3 – 16 Heads, 2 Blocks

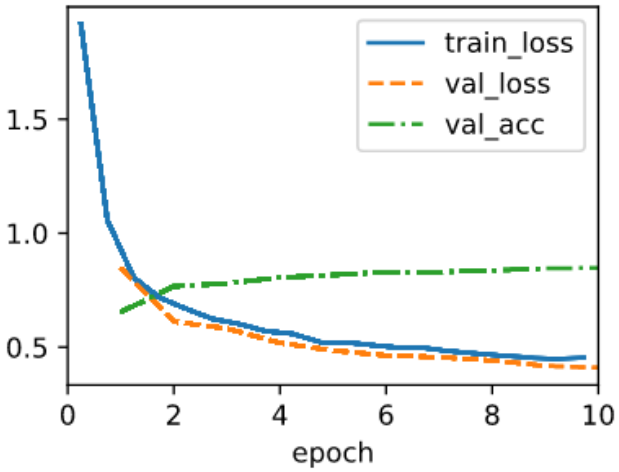
```
Total Training Time : 948.40636 s
Estimated Average Training Time per Epoch : 94.84064 s
```



This model had slightly higher validation accuracy compared to model 2.1. Although it took substantially longer to train comparatively, this model is the better model.

Model 2.4 – 16 Heads, 3 Blocks

```
Total Training Time : 1364.59887 s
Estimated Average Training Time per Epoch : 136.45989 s
```



This model took so long to train zero benefit. With smaller datasets like CIFAR10, it does not seem like you need crazy large transformers to still have a very competent model.