Experimentation

The below are all the hyperparameters for training

```
## Hyperparameters
epochs = 10             # 10
lr = 0.001              # acc to other parameters
batch_size = 32         # fixed time optimality

num_layers = 2          # fixed as assingment
expansion_factor = 2    # strong enough no need to experiment
n_heads = 4             # vairable (4, 8)
embed_dim = 300         # variable (100, 300)
```

We will experiment with `n_head` and `embed_dim` keeping all other hyperparamters constant as they are insignificant.

- `n_head` takes either 4, 8
- `embed_dim` takes either 100, 300

Results

n_head	embed_dim	Test Loss	Perplexity
4	100	2.885	17.90
4	300	2.360	10.59
6	100	2.816	16.70
6	300	1.603	4.96

Interpretation

- The model accuracy increases with increase in `n_head` or `embed_dim`.
- I am unable to reach the optimal value of `n_head` and `embed_dim` as the model does not fit in the GPU memory for larger values.
- This model is trained on a rather small dataset 30,000 sentence pairs. The perplexity will decrease as the dataset size increases.
- The above observations support the correctness of the model.