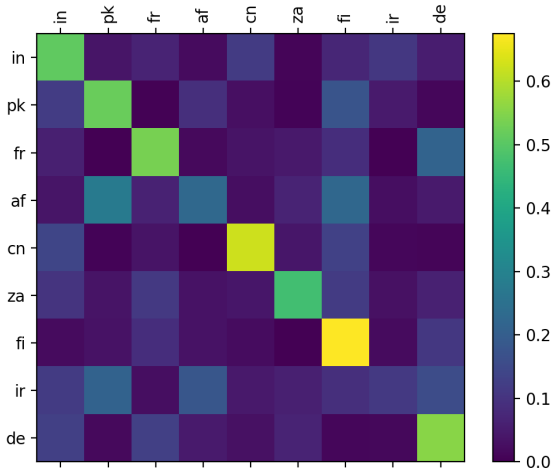
CS505 HW3 Writeup
Yuhao He
Part I
(1) ***Write code to output accuracy on the validation set. Include your best validation accuracy in the report. Discuss where your model is making mistakes and use a confusion matrix plot to support your answer.***

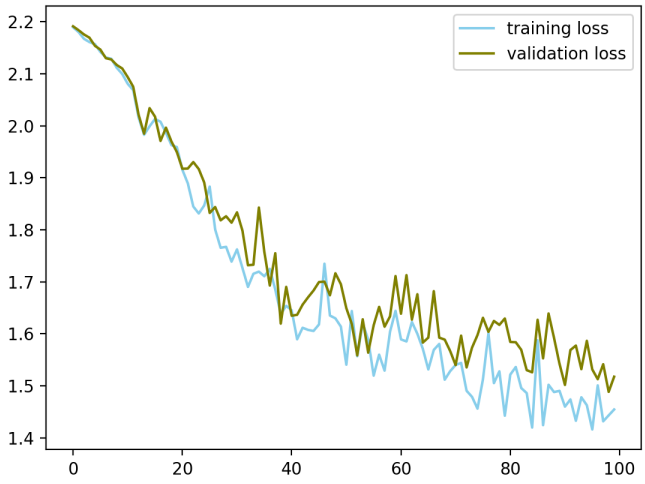
The best model for me is n_hidden = 512, learning rate = 0.0015, below is the confusion matrix it produced.



According the Confusion Matrix is shown above, the model performs well on classifying categories like cn and fi(>0.6), but bad on ir (<0.2). It seems that it might make confusion on ir and af as they have similar heat color on the matrix.

(2) ***Modify the training loop to periodically compute the loss on the validation set, and create a plot with the training and validation loss as training progresses. Is your model overfitting? Include the plot in your report.***

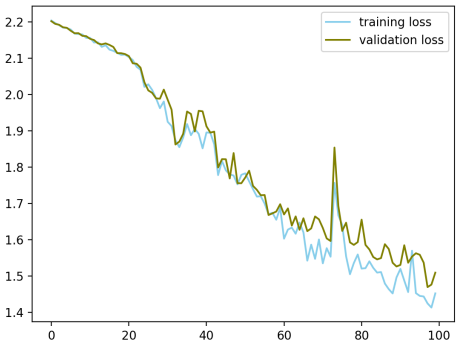
By creating the training set and validation set, and keeping all other parameters the same as the tutorial, we get the graph for the losses of train and validation set as blow.

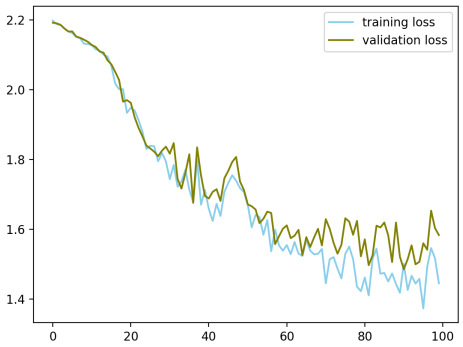


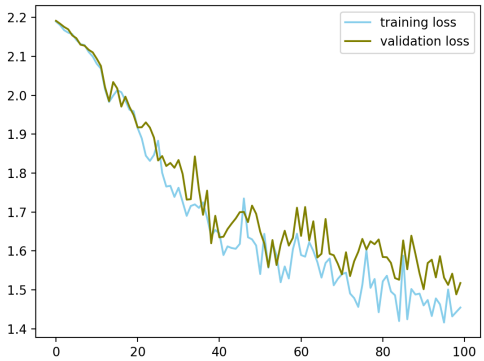
The above graph shows that the losses for training and validation are both decrease while the number of epoch increases, and the validation loss is slightly greater than the training loss for the most of time, but they are follow the same trend, which means the model is not overfitting.

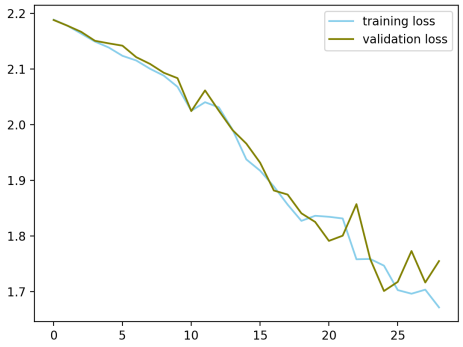
(3) ***Experiment with the learning rate (at least 5 different learning rates). You can try a few different learning rates and observe how this affects the loss. Use plots to explain your experiments and their effects on the loss.***

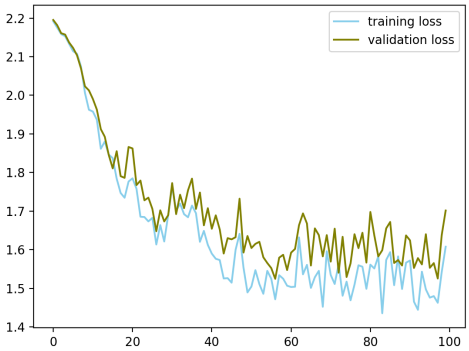
The learning rates I tried are 0.001, 0.0015, 0.002, 0.0025, 0.003, while other configuration remain the same.







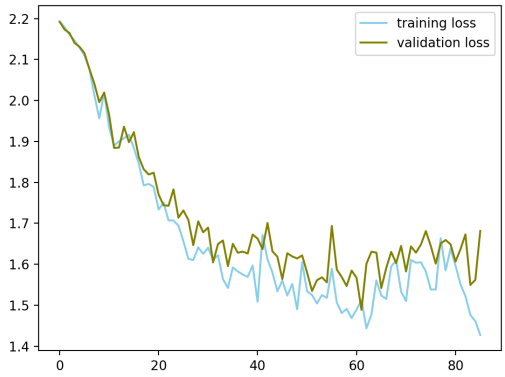


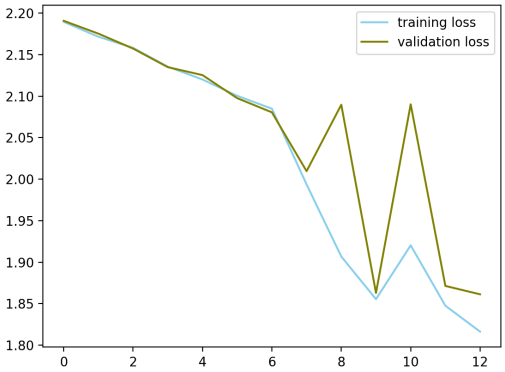


As we can see from above, the larger the learning rate, the faster the losses decrease. It's obvious to see that in fig4 (count from left to right, start at the top), the loss falls to 1.8 while in the figures before it, their losses are higher than this value.
Moreover, a higher learning rate, also results unstable losses. We can see from the final graph, its loss value wave between 1.5 to 1.7 after 40, which indicates that we may overshoot during the optimization process.

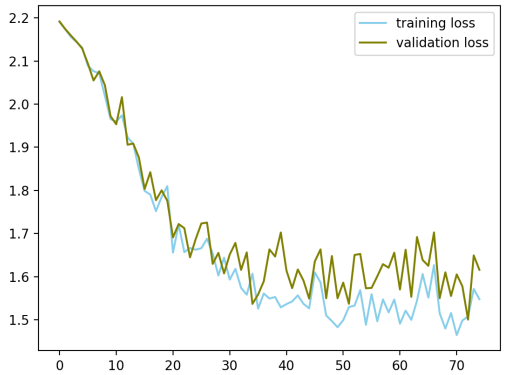
(4) Experiment with the size of the hidden layer or the model architecture (at least 5 different sizes and/or modifications). How does this affect validation accuracy?

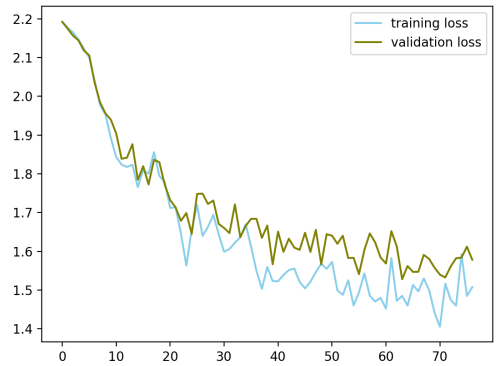
I chose to experiment with 5 different sizes of hidden layer which are 128, 256, 384, 512, 640



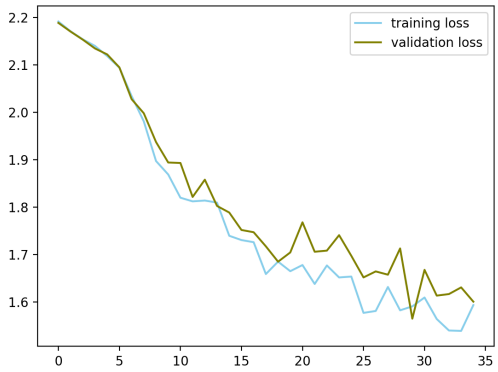


Left: n_hidden = 128 micro_acc = 0.488943 macro_acc = 0.487000
Right: n_hidden = 256 micro_acc = 0.111387 macro_acc = 0.115789





Left: n_hidden = 384 micro_acc = 0.4580999 macro_acc = 0.060184
Right: n_hidden = 512 micro_acc = 0.4796999 macro_acc = 0.059398



n_hidden = 640 micro_acc = 0.113499 macro_acc = 0.112568
From the graphs above, we can see that:

i. We can't guarantee that the more hidden states, would produce a higher accuracy
ii. When n_hidden >= 256, the training loss and validation loss seem to diverge. This might indicate that too many hidden layer would cause an overfitting

Part II

(1) ***Include a sample of the text generated by your model***

I used *Donald Trump Speeches* collection. The following is the output the model generated

```
[20m 58s (2800 56%) 1.4486]
Why, say in ceration and there's briegest in pupty and some womething. Sow
the peoplonestly we setter

[21m 43s (2900 57%) 1.7686]
What now when you say freat and was people want the Unitaw it can people have
deopsen - you know. You

[22m 28s (3000 60%) 1.5207]
What and they're doing our mose.
don't win and the he hups. This couple in comm to be a vetor again an
[29m 16s (3900 78%) 1.7700]
Whe wall of to week to weehrate spent a.
week, I have to disas president is going to do with wode a st
```

```
You know what what we have to be the people the because the world and the
people the because the world the people the because the world the because the
because the world what we have to be the people of the country and the world
what the world and the world the people the because the world in the people
of the because what we have to be the world in the world the tough of the
because the world the country the because the people the because the world
are the country the tough the because the world and the way the country of
the tell you know it to be the he was a country the people of the country our
money and the world the country and the world what we have the country of the
because the tough the tough to the because the world the country of the tough
the because the world the people of the tough the because the world the
because the people the tough the because the world the people of the people
things the people of the because the else the people of the because the world
the people of
You don't Jumbaness on over airma, for smart-ord which..?.D Juse ownes.. Uney
plan.
```

(2) Give a qualitative discussion of the results and where does it do well and where does it seem to fail?

From the output in the first question, we can see although it produces the word initial with "Wh" well in some case, it still gave us some random word that is not a real English word. This might result from that we set the temperature too high, which make the model produce words more randomly. The smaller a temperature, the more likely to produce a real word.

(3) Report perplexity on a couple validation texts that are similar and different to the training data

To avoid overflow and underflow errors, I take log and then take exponential to perplexity.

In evaluation function, there is a variable output_dist which can help us to calculate P(c|prime_str), and a function next_char_prob(c, prime_str) to compute next char probability and eventually to calculate the perplexity of a sentence.

Sentence for temperature as 1.4 with perplexity 813.7173915399036
    Wheter.
you?

cerfu vety negotia erientolas Hear isbie-ting actuewa and whefeerrrible he's goin our no

Sentence for temperature as 0.2 with perplexity 1.3015408492096683:
    Wher we're going to be a lot of the way the way it. The with the way and they're going to be a preside

The second sentence has a smaller perplexity than the first one, and it is true because the second one has more meaningful words than the previous sentence