Kelly Gothard
CS 395 Deep Learning
HW 2

Dataset

The dataset used to train the character-level models below was a collection of 10000 of the most recent comments from the popular Reddit forum r/The_Donald. The dataset was collected from pushshift.io, an opensource curator of Reddit comments, using the pushshift API.

The comments were divided into sequences of 10 characters, and those sequences were encoded with a vocabulary size of 256, representing the ASCII integer encoding for characters.

Results

Three model types - simple RNN, LSTM, and GRU - were trained, validated, and tested on the above dataset with an 80-20 training-testing split and a 90-10 training-validation split. Additionally, each model was trained, validated, and testing using 3 different values of hidden units: 75, 125, and 150 hidden units. Table 1 below shows the training and validation accuracy, as well as the precision and recall from testing for each of the 9 total models.

For each model type, all metrics increased as the number of hidden units increased, with the exception of the GRU models with 125 units and 150 units. Of the model types, LSTM had the highest accuracy metrics for each hidden unit value. The GRU models had the next best metrics, and the simple RNNs had the lowest.

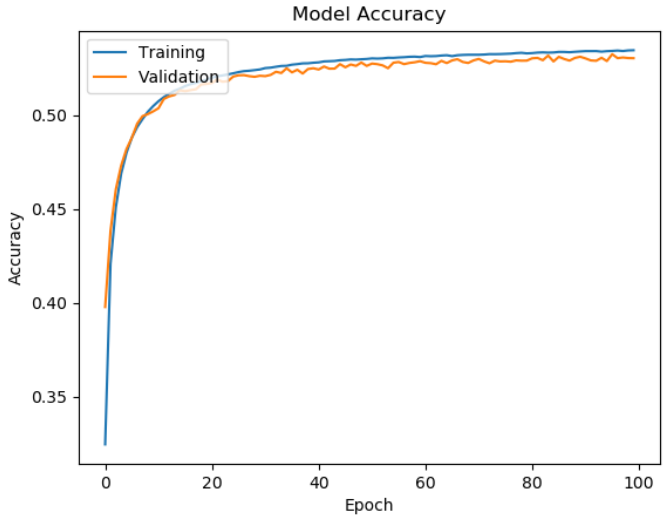
Table 1: Results Summary

Model Type	# Hidden Units	Train Acc.	Val. Acc.	Test Precision	Test Recall
RNN	75	0.54	0.53	0.53	0.53
	125	0.57	0.56	0.56	0.56
	150	0.58	0.57	0.57	0.57
LSTM	75	0.59	0.58	0.58	0.58
	125	0.63	0.60	0.60	0.60
	150	0.64	0.61	0.61	0.61
GRU	75	0.58	0.57	0.57	0.57
	125	0.62	0.60	0.60	0.60
	150	0.63	0.60	0.60	0.60

Simple RNN

A simple RNN with 75 hidden units was trained on sequences of 10 characters, in batches of size 512, to predict the next character in the sequence. 10 percent of the sequences were saved for testing, and 10 percent of the remaining training data was used for validation.

Figure 1: Training and Validation Accuracy over 100 Training Epochs for Simple RNN with 75 Hidden Units



The training and validation accuracy behave very similarly over the training epochs, suggesting that the model is not overfitting to the data. Additionally, checkpoints at every 20th epoch were tested with a small section of the training data to qualitatively assess the models improvement as training progressed. The checkpoints below show that the model was unable to successfully predict an entire word, but it did predict that there were three space delimited words that followed "I'm pretty", after having predicted just two words in epochs 40 and 60.

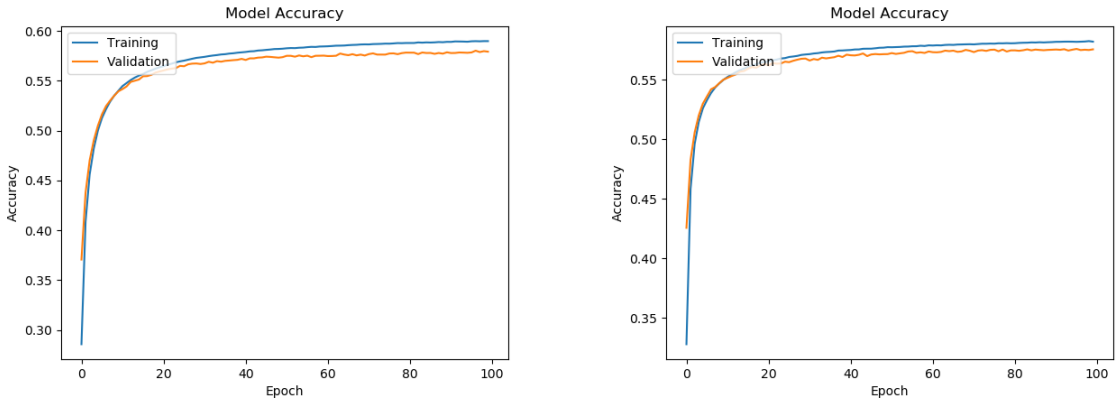
Checkpoints

Actual: I'm pretty sure he knows
Epoch 20: I'm pretty aopv aa wnow
Epoch 40: I'm pretty tope terwnow
Epoch 60: I'm pretty aope terwnow
Epoch 80: I'm pretty tope te wnow
Epoch 100: I'm pretty aope te wnow

LSTM and GRU

We then substituted a simple RNN layer for an LSTM layer, and then once again for a GRU layer in another model, while keep the number of hidden layers fixed at 75. Figure 2 shows the training and validation accuracy for each of these models over 100 epochs. The training and validation accuracy for these models has a very similar behavior to the simple RNN model above, but with slightly higher accuracies.

Figure 2: Training and Validation Accuracy over 100 Training Epochs for LSTM (left) and GRU (right) with 75 Hidden Units



Once again, we tested a small sample of the training data at the 100th epoch checkpoint for each model, and we see that although the LSTM and GRU models performed slightly better than the simple RNN model, we are still unable to successfully predict one word.

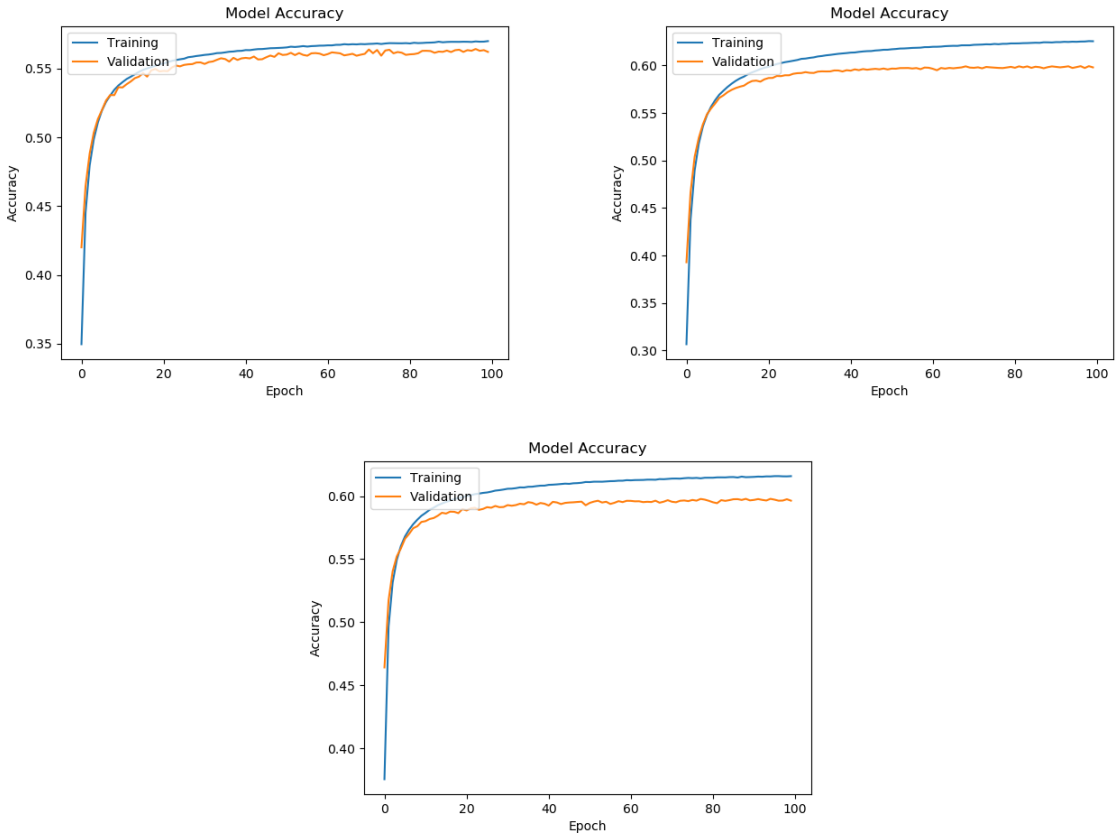
Checkpoints

Actual: I'm pretty sure he knows
RNN, 75 hidden units: I'm pretty aope te wnow
LSTM, 75 hidden units: I'm pretty supe te inown
GRU, 75 hidden units: I'm pretty supv te seow

Varying Hidden Units

We explored more models by increasing the number of hidden units in the RNN, LSTM, and GRU models to 125 and 150 hidden units. We found that increasing the number of hidden units tended to increase the accuracy in each of the models. Figure 3 shows the training and validation accuracies of each of the three model types with 125 hidden units, and checkpoints at the 100th epoch are included below.

Figure 3: Training and Validation Accuracy over 100 Training Epochs for RNN (top left), LSTM (top right) and GRU (bottom) with 125 Hidden Units



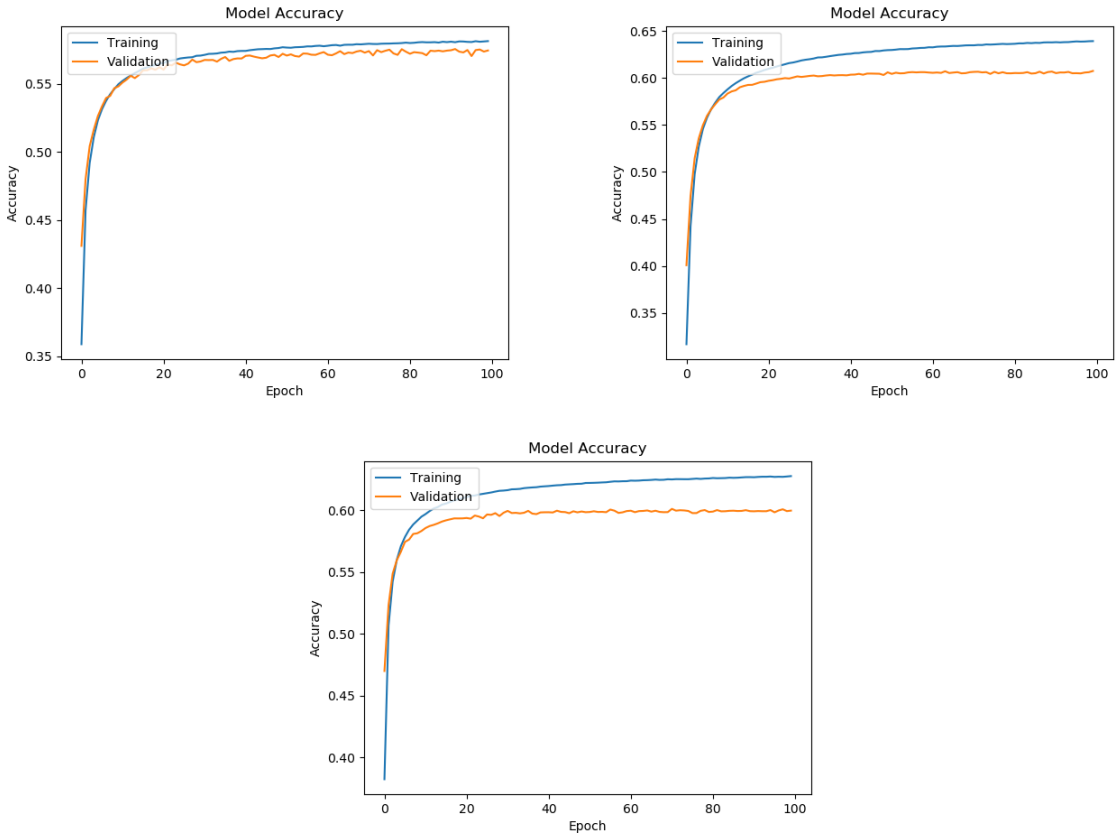
The LSTM and GRU models with 125 hidden units were finally able to successfully predict the word "sure" after our initial set of 10 characters. This is exciting, however, the models are unsuccessful at predicting the subsequent words.

Checkpoints

Actual: I'm pretty sure he knows
RNN, 125 hidden units: I'm pretty sape ta wnow
LSTM, 125 hidden units: I'm pretty sure ye'inows
GRU, 125 hidden units: I'm pretty sure te wnows

Each model type with 150 hidden units had slighly higher accuracy than with fewer units, apart from the GRU. At this point, the GRU appears to be stagnating. Figure 4 shows each models' training and validation accuracies, and tests at the 100th checkpoint are included below.

Figure 4: Training and Validation Accuracy over 100 Training Epochs for RNN (top left), LSTM (top right) and GRU (bottom) with 150 Hidden Units



These checkpoint tests show that LSTM and GRU at 150 units are also able to predict the word "sure" after "pretty", but they cannot successfully predict the next word.

Checkpoints

Actual: I'm pretty sure he knows
RNN, 150 hidden units: I'm pretty iope te wnow
LSTM, 150 hidden units: I'm pretty sure te inows
GRU, 150 hidden units: I'm pretty sure te inows

Normalized confusion matrices are included below for reference.

Figure 5: Normalized Confusion Matrices for RNN, LSTM, and GRU with 75, 125, and 150 Hidden Units (Rows from top: 75 units, 125 units, 150 units; Columns from left: RNN, LSTM, GRU)