\section{Deep Averaging Network}

\subsection{ Architecture } (5 points)

*Describe the architecture you implemented at a level of detail that one could re-implement your model from the description.*

*Include what variants you experimented with and how you validated which worked best.*

The architecture implemented for our Deep Averaging Network involved three layers:

1. Embedded layer
2. Masked mean layer that calculates the ratio between the sum of tokens after eliminating those not in the mask, and the total number of tokens surviving since the last layer.
3. Linear feedforward layer with an input dimension of 300 and output dimension of 2.

We experimented with changing the dimensions of the hidden layers and dropout rates.

\subsection{ Results } (15 points)

*Report a table with all variants you validated and their performance.*

*Experiment with at least two major hyper-parameters.*

\subsection{Discussion} (5 points)

Discuss how you selected the final model to evaluate and which variants worked better.

Were there setting that were never successful and why do you think this is the case?

Similarly, were there any settings that were particularly good and why do you think this is the case?

\section{Recurrent Neural Network}

\subsection{ Architecture } (5 points)

Describe the architecture you implemented at a level of detail that one could re-implement your model from the description.

Include what variants you RNNs you experimented with and how you validated which worked best.

Include how you extracted a feature vector from your RNNs for classification.

\subsection{ Results }(15 points)

Report a table with all variants you validated and their performance.

Discuss how you selected the final model to evaluate.

Experiment with at least two major hyper-parameters and at least two different style of RNNs.

\subsection{Discussion} (5 points)

Discuss how you selected the final model to evaluate and which variants worked better.

Were there setting that were never successful and why do you think this is the case?

Similarly, were there any settings that were particularly good and why do you think this is the case?

\section{Qualitative Analysis}

\subsection{Importance of the GLOVE initialization}

In class we discussed that word embeddings offer a good initialization for features of sentences.

In this section, we will analyze how true this claim is.

Experimentation using DAN of a single Linear layer and lr of 0.0001 and dropout of 0.

\subsubsection{Training with alternative word embedding initialization}

Perform the following ablation experiments and report the results in a table, comparing them to the your results above:

\begin{itemize}

\item (2 points) Train a DAN but do not update the GLoVE word embeddings during learning

Val accuracy 82.3467

\item (2 points) Train a DAN but initialize the word embeddings to random vectors and update them during learning.

Val accuracy 77.7507

\item (2 points) Train a DAN but initialize the word embeddings to random vectors and {\bf do not update} them during learning.

Val accuracy 67.7522

\item (2 points) Train a DAN but initialize the word embeddings to random vectors and {\bf do not update} them during learning. Evaluate how increasing the depth of the DAN changes performance.

This is with linear, relu, linear

Val accuracy 78.0475

\end{itemize}

\subsubsection{Discussion} (7 points)

Discuss your results, with a focus toward answering the overall question of how important is GLoVE initialization to classification performance. Answer the following questions:

\begin{itemize}

\item Is updating embeddings away from their initialization crucial? Provide at least one hypothesis for explaining the results of your experiment.

\item What is the overall effect of initializing with pretrained embeddings? Are there any factors to consider beyond final performance?

\item What experiments could you run to further analyze the role of word embedding initialization?

\end{itemize}

\subsection{Model behavior as a function of sentence length} (10 points)

Different models may perform for differently on different types of input.

Often the length of the input sentence can be used as a simple proxy of complexity.

Analyze the performance of your best DANs and RNNs as function of the length of the input sentence.

Plot performance as a function of the length of the input sentence for both models on the same plot.

One way to do this is simply to bin the input sentences by length, and compute accuracy within each bin.

Discuss your results, answering the following questions

\begin{itemize}

\item Where do the models both perform well?

\item Are there types of sentences that one model excels at while the other fails?

\item What is your hypothesis for explaining the results you observe?

\end{itemize}

\section{Extra Credit} (up to 10 points)

Manually construct failure cases (inputs) for classifier (eg: negation and metaphor).

Consider both false positive and false negative cases.

Extra credit will be awarded if you can describe a pattern for the new failure cases you have discovered, and report at least 5 instances not from the dataset that demonstrate the pattern and that your model fails to classify correctly.

We will award credit based on how interest and coherent of errors you find.