Deep Averaging Network and Gated Recurrent Unit

112685069, Abdullah Mitkar

MODEL REPRESENTATION

Deep Averaging Network

Input: Vector sequence (all sentences with words' in vector representation)
Output: last_representation(representation after last hidden layer)
layer_representation(representation after every hidden layer)

Algorithm for DAN

Part A: Implementing word dropout

i. Calculate dropout array from Bernoulli distribution using normal disctribution. Dimension size [1 x no_of_words]
ii. Replicate in a list for batch_size times. Dimension size [batch_size x no_of_words]
iii. Only during training, do; dropout_mask x sequential_mask and create one filter. Dimension size [batch_size x no_of_words]
    - The two tasks are multiplied because we need to apply dropout and sequential one after other, so we can apply both together

Part B: Averaging of Word Vectors

i. For every sentence representation, calculate the average of every word representation [batch_size x embedding_size]
ii. For number or layers:
    - Add the average to a list. [num_layer x batch_size x embedding size]
iii. Save the last representation [batch_size x embedding_size]
iv. Reorder the dimensions to make it [batch_size x num_layers x embedding_size]
v. Return last_representation, layer_representations.

Gated Recurrent Unit

Algorithm for GRU

<u>i.</u> vector sequence, unit_output=GRU(vector sequence)
<u>ii.</u> For no_of_GRU_units:
    - sequence, unit_output = GRU(unit_output)

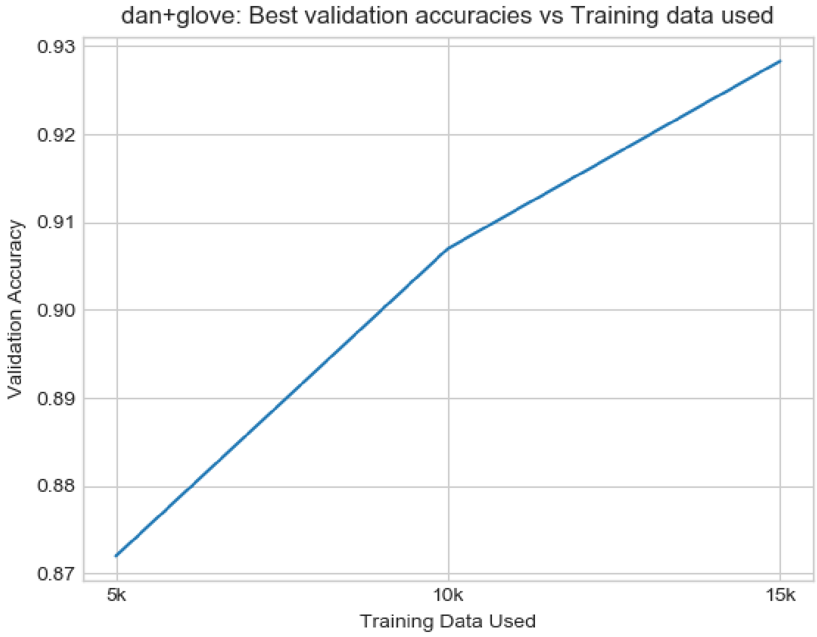
- LayerRepresentation.add(unit_output)
- Reorder the dimensions to make it [batch_size x num_layers x embedding_size]
- Return last_representation, layer_representations.

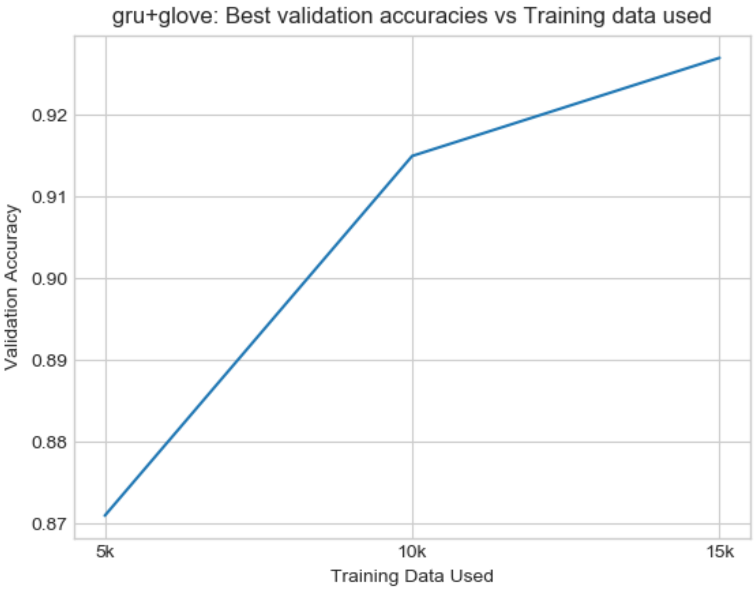
ANALYSIS

Learning Curves

Increasing the training data

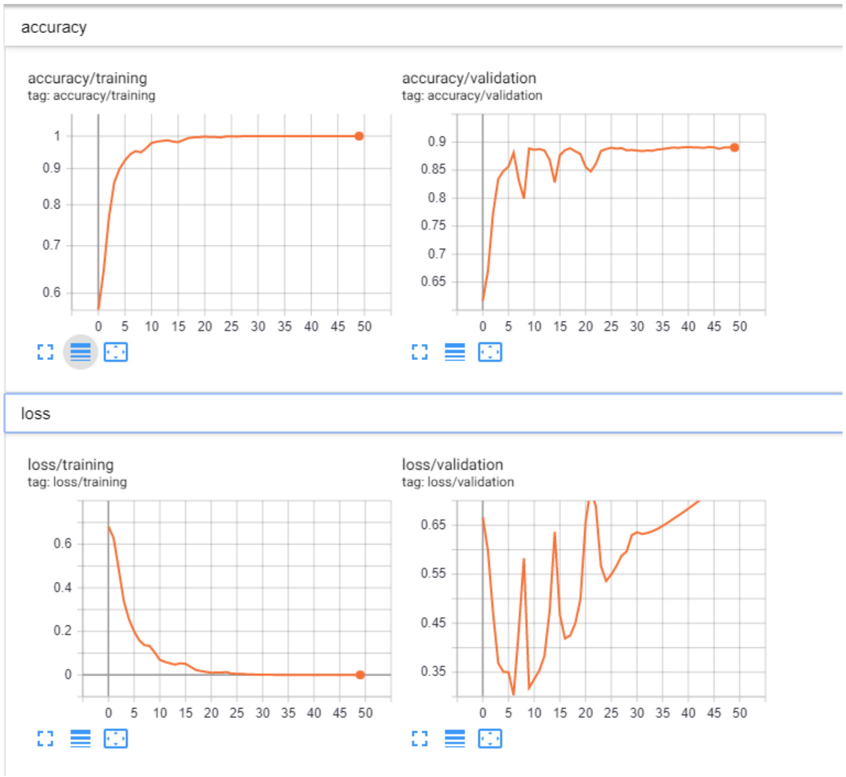
The accuracy increases as the training data increases. However, this increase is more in case of GRU when you compare it to DAN. For training data of size 10k, the accuracy of GRU is more as compared to DAN. This is so because GRU can perform better leaning when compared to DAN over smaller training samples. This can be ascribed to its complex neural networks.





Increase no of epochs

Increasing no of epoch increases accuracy exponentially to a certain number of epochs. However, the accuracy of the validation set drops after epoch 6 and consequently the loss shoots. The accuracy then fluctuates but then attains a certain value. The corresponding loss also fluctuates.



Error Analysis

DAN over GRU

One advantage of DAN over GRU is that it takes much lesser time to train.

GRU over DAN

DAN essentially averages the word vectors and in doing that it may happen that essentially important single words that change the meaning of the sentence may be averaged out. For eg., I like NLP and I hardly NLP. There is only a single word difference. But since most of the words are similar, the vectors are averaged out and hence the meaning is lost. GRU does a pretty good job in maintaining information contained in words vectors.

GRU fails

Test Case: {"text": "The actors were great. They acted really nice. The actress was good. She was nice. The actors perfectly portrayed the character to great perfection. It really takes the acting to a whole new level. I can saw only good things about the actor. Although, the movie was bad.", "label": 0}

GRU labels it positive even though it is a negative review which goes into saying that GRU can wrongly associate words to movie even though they are about actor.

PROBING

For the main model we use the last representation as the representation for the entire sentence.
For the probing model, we use the layer representation that we created.
The stack essentially contains the intermediate representation at the end of every intermediate state.
We use the representation of the layer under consideration and use a linear classification to evaluate its performance.

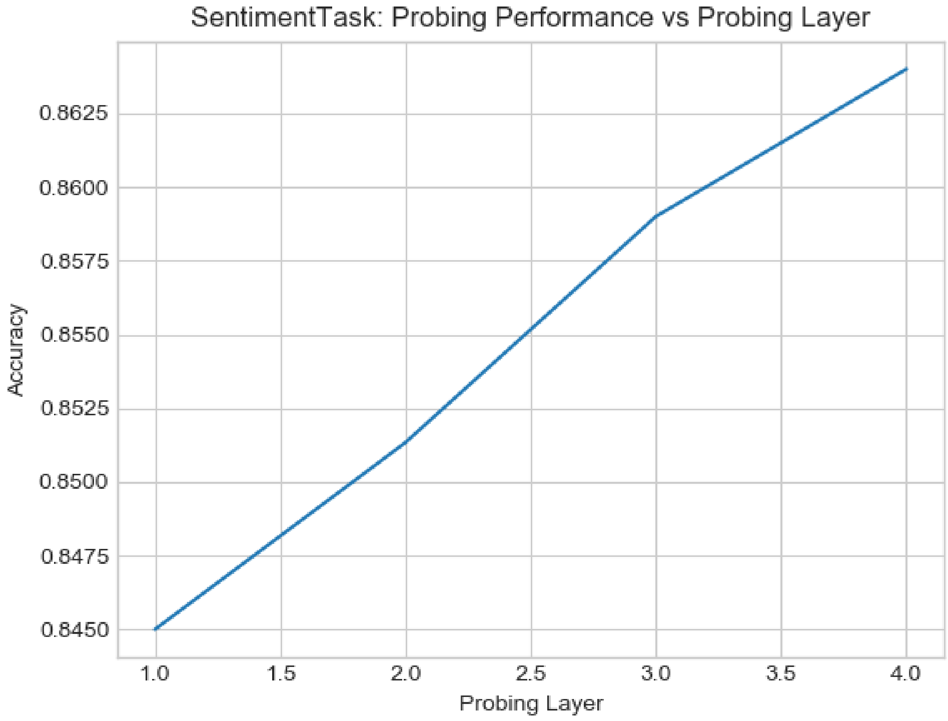
Probing for Sentiment Analysis

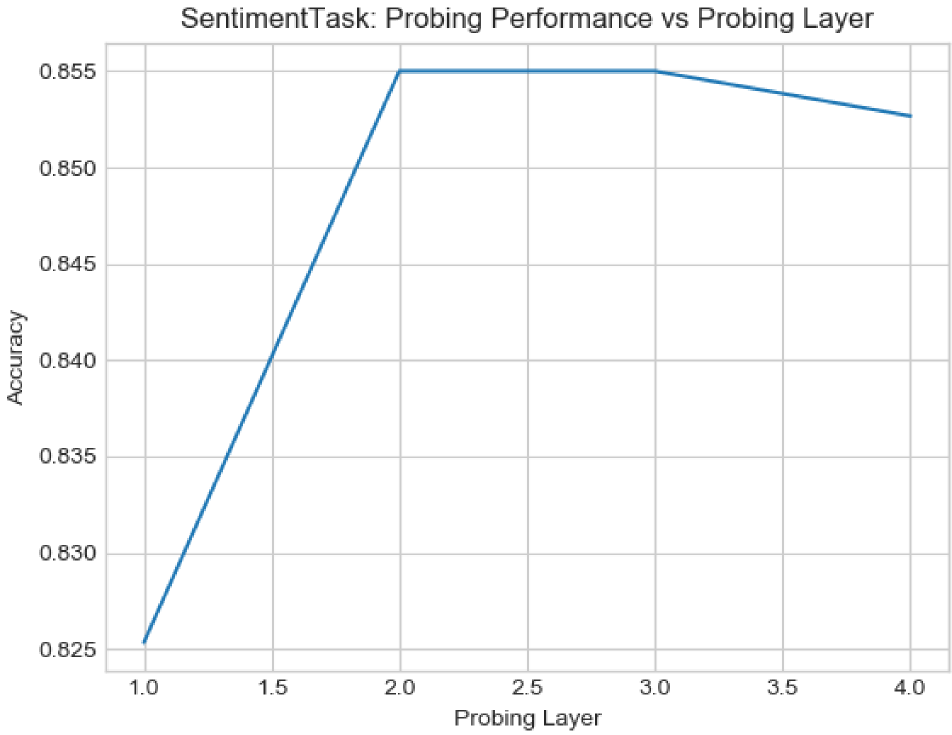
DAN

The probing graph shows that accuracy of representation increases and subsequently adds more value across each hidden layer. And it is a gradual increase showing that DAN takes time to understand the and identify the weights but it continues to do a better job as it approaches the last layer.

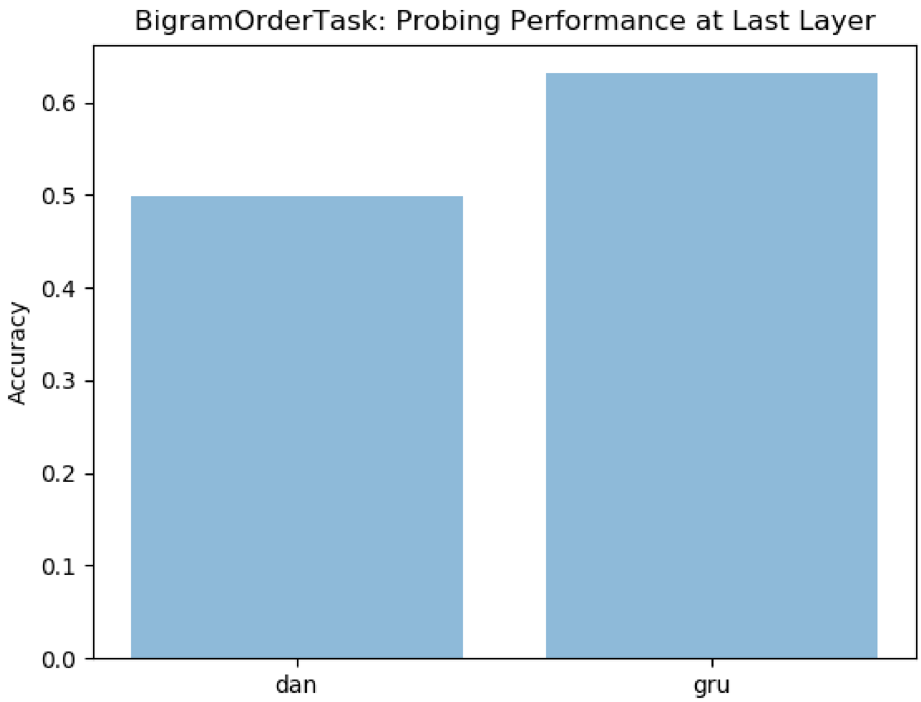
GRU

The probing graph shows that accuracy of representation increases exponentially up to a point which further adds on to the premise that GRU takes little time to correctly identify the values and does a good job at it.

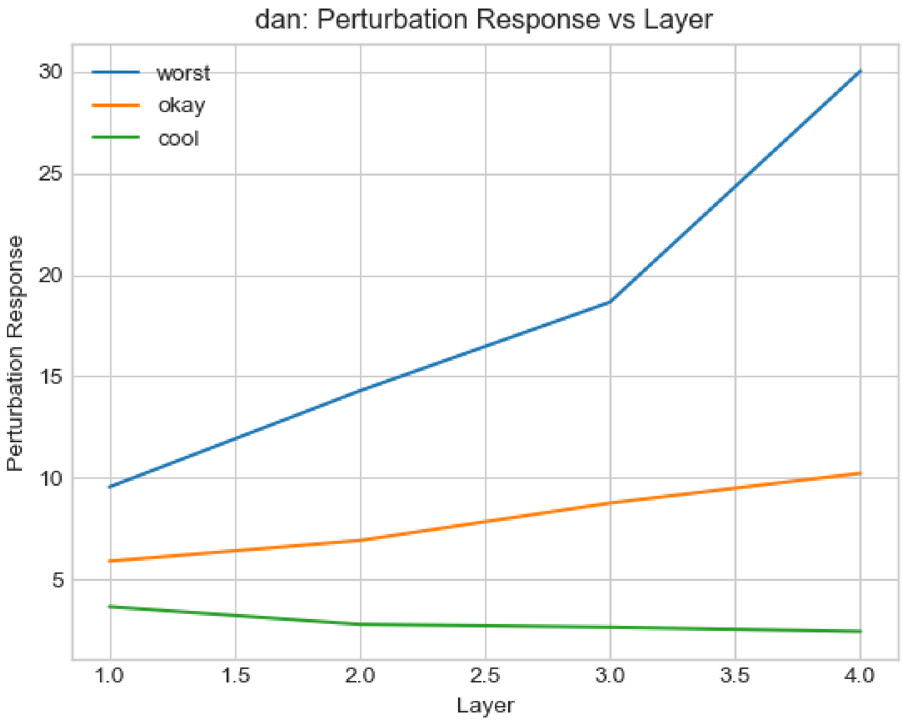

SentimentTask: Probing Performance vs Probing Layer
Accuracy
0.8450
0.8475
0.8500
0.8525
0.8550
0.8575
0.8600
0.8625
Probing Layer
1.0
1.5
2.0
2.5
3.0
3.5
4.0

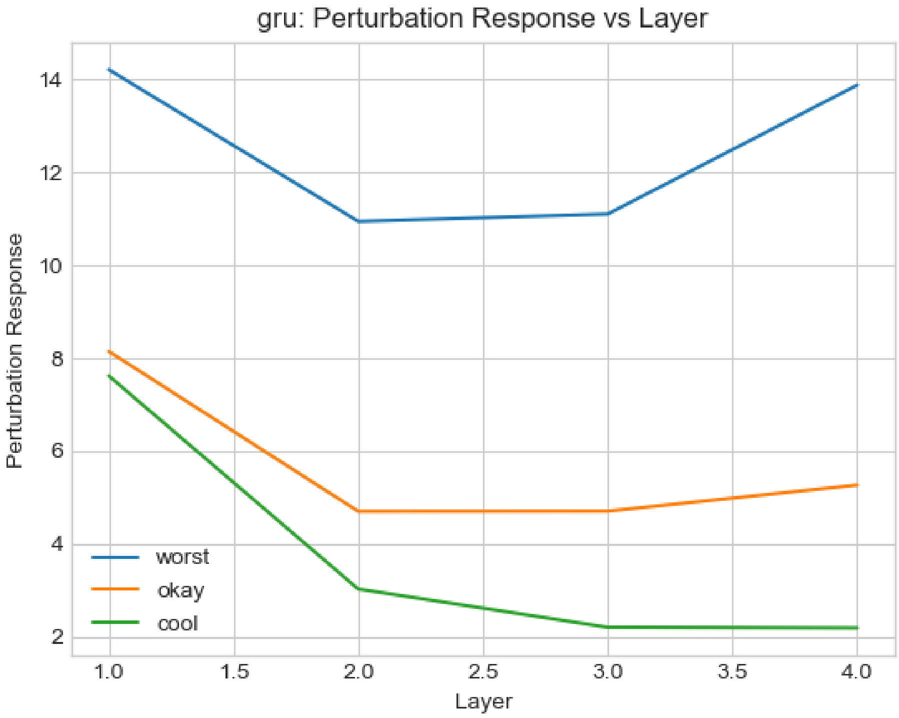

SentimentTask: Probing Performance vs Probing Layer
Accuracy
0.825
0.830
0.835
0.840
0.845
0.850
0.855
Probing Layer
1.0
1.5
2.0
2.5
3.0
3.5
4.0

Probing for Bigram Task



Analyzing Perturbing Response





Perturbation response shows that DAN initially represents all the words similarly in 1 normal form but as the layers' progress it does a better job representing it. GRU on the other hand does a great job right from the beginning in identifying the different between the words. This is due to its hidden layer and that goes into further adding to the statement that GRU does a better job in representing words right from the early stages.