NLP Assignment-2

Chaitra Hegde (cvh255)

October 31, 2018

1 Introduction

Two models, namely a CNN and RNN models are built to see if two sentencesare entail, contradict or neutral to each other. While building the model, varioushyper parameter are tuned, namely-size of hidden dimension, kernel size of CNNand regularization.

The code can be found at `https://github.com/HegdeChaitra/Stanford_Natural_Language_Inference_dataset_classification`

Due to limited space, the loss and accuracy curve for the best model is put up here. But the curves for every single experiment conducted could be seen in the github page

2 RNN

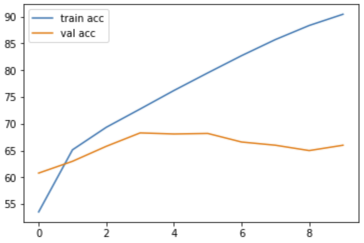
Figure no.	Feature maps	Hidden layers	Weight decay	Drop out	val acc	No of parameters
mod.1	100	1	None	None	62.5%	30223203
mod.2	200	1	None	None	65.5%	30503803
mod.3	500	1	None	None	68.2%	31705603
mod.4	700	1	None	None	66.5%	32806803
mod.5	500	2	None	None	65.6%	33208603
mod.6	500	2	0.00001	None	**68.3%**	33208603
mod.7	500	2	0.0001	None	67.8%	33208603
mod.8	500	1	0.00001	None	67.4%	31705603
mod.9	500	2	None	0.5	66%	33208603
mod.10	500	2	None	0.8	66.4%	33208603
mod.11	500	2	None	0.3	67.3%	33208603
mod.12	500	2	0.000001	0.2	66.5%	33208603

2.1 Number of Feature Maps

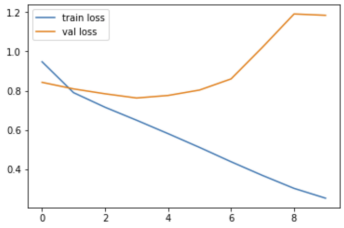
Number of hidden layers ios fixed to 1 and Feature Map(fm) fm=100,200, 500 and 700 were tried. For 100, accuracy of x was achieved. for 200, accuracy of 65.5% was achieved. while the fm=500 gave 68.2% accuracy. fm= 700 gave accuracy of 66.5%. It shows that increasing the no of feature maps until certain extent as it helps in learning more number of features about the dataset being trained. Hence, here on wards we are using the feature map of 500. Too many feature maps(eg:fm=700) is leading to over fitting on train set due to more parameters being learnt.

2.2 Number of hidden dimension

With fm=500, experiment was conducted to find the number of hidden dimension (hd). with hd=1, we had achieved accuracy of 68.2%. With hd=2, accuracy of 65.6% was achieved. Hence,with more hidden layers, the model is over fitting again. Hence, all experiments in attempt to increase the number of parameters is leading to over fitting. Time to try regularizing the model.



(a) RNN accuracy



(b) RNN Loss

Figure 1: RNN Best Model

```
predicted 0
Actual 1
sentence1:  her voice was doubtful .
sentence2 :  she sounded doubtful about it .

----------------------------------------------------------------------------------------------------
predicted 0
Actual 1
sentence1:  they do n't call them immigrants anymore that was back during my granddaddy 's day
sentence2 :  they used to call them immigrants .

----------------------------------------------------------------------------------------------------
predicted 1
Actual 0
sentence1:  this provides insight into the important japanese concept of katachi ( form ) , the rough equivalent of i
t is n't what you do ; it 's the way that you do it .
sentence2 :  all japanese people abide by the concept of katachi .

----------------------------------------------------------------------------------------------------
```

(a) RNN incorrect prediction

```
predicted 0
Actual 0
sentence1 :  pro-choicers point out that these close-up images literally cut the fetus 's context -- the woman -- out
of the picture .
sentence2 :  pro-choices say the close-up images are unfair to women .

----------------------------------------------------------------------------------------------------
predicted 2
Actual 2
sentence1 :  however , assuming the procedural requirements of chapter 36 are met , changes negotiated by the postal
service and a mail user for their mutual benefit may merit recommendation under the applicable statutory standards .
sentence2 :  changes negotiated by the postal service are too regular and prohibit my postal services .

----------------------------------------------------------------------------------------------------
predicted 1
Actual 1
sentence1 :  no . i guess i 'm going too .
sentence2 :  i 'll come along .

----------------------------------------------------------------------------------------------------
```

(b) RNN correct prediction

Figure 2: RNN prediction example

2.3 Regularization

I tried two regularization methods on two experiments- one that was clearly over fitting (fm=500, dh=2) and other model which performed best so far (fm=500, dh=1). First with weight decay. with fm=500, hd=2 and weight decay = 0.00001 was used to get accuracy of 68.3%. As we had hoped, the same model with weight decay gave almost 3% increase in accuracy. In order to see if increasing regularization by increasing the weight decay results in better performance, weight decay = 0.0001 was tried, and it gave accuracy of 67.8%. Hence, its too much of regularization. Hence, we will be switching back to weight decay=0.00001. Now, we pick a model which was performing very well before using regularization (fm=500, hd=1),and try weight decay=0.00001, gave performance of 67.4% which is not as good as the same model without regularization. Its because the regualarization in this case is restricting learning on train set that in a way is also affecting validation performance.

Drop-outs were used for regularization as well. Too much of drop out is affecting the learning it self while slight drop out probability is giving result better than model without any regularization

2.4 Best RNN Model and MNLI performance

The model with hidden size =2 with feature maps = 500 with weight decay=0.00001 is giving the best performance on SNLI validation dataset = 68.3%

The performance on MNLI dataset genre wise is as follows

Genre	Val Accuracy
Fiction	45.32%
Telephone	45.07%
Slate	41.01%
Government	41.04%
Travel	42.15%

As a whole MNLI dataset validation performance is 42.94%

3 CNN

Fig no.	channels1	Channels2	Weight decay	Drop out	kernel size	val accuracy	No of parameters
mod.1	200	100	None	None	5	65.6%	30502903
mod.2	500	100	None	None	5	64.1%	31103203
mod.3	300	100	None	None	5	64.6%	30703003
mod.4	600	100	None	None	5	64.1%	31303303
mod.5	300	200	None	None	5	64.0%	30953103
mod.6	300	300	None	None	5	65.4%	31203203
mod.7	200	100	0.00001	None	5	65.9%	30502903
mod.8	300	300	0.00001	None	5	65.7%	31203203
mod.9	200	100	0.00005	None	5	65.2%	30502903
mod.10	200	100	0.00001	None	3	66.8%	30382903
mod.11	200	100	None	0.5	3	66.1%	30382903
mod.12	200	100	None	0.3	3	67.2%	30382903
mod.13	200	100	None	0.2	3	65%	30382903
mod.14	200	100	None	0.3	7	64.7%	30622903
mod.15	200	100	0.000001	0.2	3	**67.4%**	30382903

3.1 Number of Channels

The CNN model consists of two convolution layer. Hence deciding the number of channels/ feature maps for each of the convolution is crucial. With fixed kernel size of 5 and without any regularization, various combination is tried out. More number of feature maps leads to more number of trainable parameters of the model. channel1(c1)=200 and channel2(c2=100) gave performance of 65.6%. c1=300, c2=100 gave performance of 65.4% accuracy. All other combinations with (c1,c2) = (500,100),(300,100),(600,100),(300,300) are giving below 65% performance. The possible reason in case of larger feature maps is over fitting - for which we will try regularization in next section. In remaining cases, the low performance is due to inappropriate combination for c1 and c2- i,e either c1 is too large and c2 is small which doesnt allow good flow of information.

Hence only models that are performing well are (c1,c2)=(200,100) and (300,300)

3.2 Regularization and Kernel Size

With kernel size =5, and for two configuration of models (c1,c2)=(200,100) and (300,300), weight decay and l2 regularizations are tried. First weight decay(wd)=0.00001 is tried for both configuration to get performance of 65.9% and 65.7% respectively. Both the values are very close, hence we will be moving forward with c1,c2=200,100 as it has lesser parameters than the other.

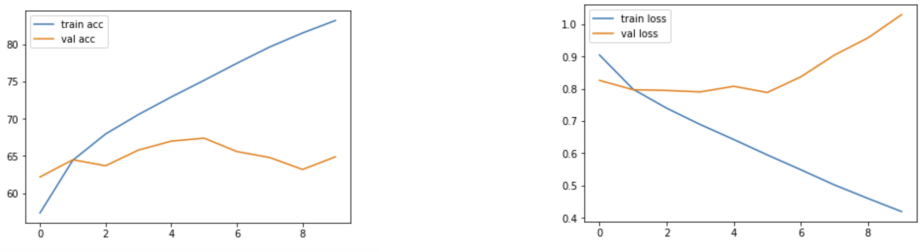
For the same configuration, the kernel size was reduced to 3 and the performance boosted to 66.8%. Hence kernel size=3 works better than 5. The possible reason being that very short term dependency capture/sequence info is sufficient than the dependency capture using kernel size=5

To see if increasing regularization works, wd=0.00005 was tried but performance dropped. Hence, wd=0.00001 was giving better performance.

Now, we will experiment with the drop out regularization, without weight decay for previous best configuration of c1,c2=200,100 and kernel size=3. dp=0.5 gave 66.1% accuracy and dp=0.3 gave 67.2% accuracy. decreasing dp helped here as we were regularizing too much with dp=0.5. To see if further lowering helps, dp=0.2 is tried but it turns out to be inefficient amount of regularization.

To see if capturing more sequence information with kernel size =7 helps for best config so far, kernel size=7 was tried but it was not good.

At last, combination of weight decay and drop out wd=0.000001 and dp=0.2 was tried which gave 67.4% accuracy which is the best model.



(a) CNN Accuracy

(b) CNN Loss

Figure 3: CNN Best model

```
predicted 1
Actual 2
sentence1:  the rustic bras-david picnic area , for example , is set alongside a burbling stream .
sentence2:  the picnic area is not near a stream .

---------------------------------------------------------------------------------------------------
predicted 0
Actual 2
sentence1:  his grandson akbar chose agra for his capital over delhi .
sentence2:  his grandson chose washington dc as the capital , not new york city .

---------------------------------------------------------------------------------------------------
predicted 1
Actual 0
sentence1:  the purpose of the diwan-i-khas is hotly disputed ; it is not necessarily the hall of private audience th
at its name implies .
sentence2:  the hall is not know many people .

---------------------------------------------------------------------------------------------------
```

(a) CNN incorrect prediction

```
predicted 1
Actual 1
sentence1 : the office of information and regulatory affairs of omb approved the
sentence2 :  something was approved by the office of affairs .

---------------------------------------------------------------------------------------------------
predicted 2
Actual 2
sentence1 : programs in michigan and the district of columbia received one-year grant terms for 2002 .
sentence2 :  programs in michigan receive no grants at all .

---------------------------------------------------------------------------------------------------
predicted 1
Actual 1
sentence1 : alternatively , there are sousa and goncalves ( rua do castanheiro , 47 ) and unibasket ( rua do carmo ,
42 ; tel . 291/226 925 ) , both in funchal .
sentence2 :  there are other places in funchal .

---------------------------------------------------------------------------------------------------
```

(b) CNN correct prediction

Figure 4: CNN prediction example

3.3 Best CNN model and MSNLI performance

The CNN model with channel1=200, channel2=100, with kernel size=3 and weight decay =0.000001, drop out =0.2 gave SNLI validation performance of 67.4%.

The MLNI performance of the best model category wise is as follows:

Genre	Val Accuracy
Fiction	43.015%
Telephone	41.39%
Slate	41.71%
Government	40.15%
Travel	42.66%

The final validation accuracy on MNLI dataset as a whole is 41.78%

4 Correct and Incorrect predictions

The correct predictions clearly shown that are correct - there is no ambiguity, no long term relation among words and contains very less out of vocabulary sentence

Where as the incorrect predictions seem to contain bit of a long term dependecy that is not necessarily captured by the model. Also, it has few out of vocabulary word which are less frequent than other words and hence excluded from the vocabulary. Good fix would be to increase the vocabulary size so that most of the words are included and try using heirarchial attention models to capture sentence level and word level dependency and point to most important sentences and words.