Experimental Study on Sentiment Analysis with Dynamic Attention Mechanism in DPCNN

NTU-AI6127: Project Final Report

Name
School of Computer Science and Engineering
Nanyang Technological University
{Hew Bork Hang} hewb0002@ntu.edu.sg
{Rani Vigneshkumar R.} vigneshk001@ntu.edu.sg

Abstract

The applications of Sentiment Analysis has made revolutionary progress to people's lives and it still bears much more potential. Our paper took an interest in combining Bidirectional Encoder Representations from Transformers (BERT) with Deep Pyramid Convolutional Neural Network (DPCNN) before introducing dynamic attention mechanism to this combined model which mainly lies in having attention mechanism implemented in the varying sizes of data in the model. Hence, three models (self-attention, global attention and base mode with no attention) are designed in our experiments. Two datasets (finance and tweets) are used in our experiments due to them having only 3 classes of negative, neutral and positive in the labels. The attention-based models have positive impact for finance data negligible impact for tweet data. Evaluation of the attention maps have provided insights that suggest possible causes such as the dilution of convoluted trigram vectors through convolution and misalignment of convoluted trigram sequences.

1 Introduction

In sentiment analysis, the main objective is to understand the judgement, mood or evaluation from each text within a collection of corpuses. [1] A simple illustration is that a classification of positive, neutral or negative is predicted on an analysed text by a trained model. Some of its applications are brand monitoring which has high commercial value and social studies projects such as prediction of human behaviours. To better obtain the semantic relationship of tokens in sentiment analysis, there has been shift towards deep learning techniques by both researches and business community. Techniques such as transformers, Bert and GPT have revolutionised the field of NLP. Aligning with the similar objective of improving model performance through semantic relationship, our paper seeks to investigate the coupling of Bidirectional Encoder Representations from Transformers (BERT) and Deep Pyramid Convolutional Neural Network (DPCNN) with dynamic attention mechanism.

Our paper provides the following contributions:

1) We introduced two modes of attention mechanism (global and self) in the DPCNN which we believe they have not been attempted on in DPCNN based on our research.

2) The global and self-attention mechanisms in DPCNN are dynamic as the sizes of data varies along the model.

3) In global attention mechanism in DPCNN, we introduced linear layers in parallel to output fixed features of certain size from the varying data sizes.

2 Related Work (or Background)

The source of our project's inspiration is a conference document "MIHNet: Combining N-gram, Sequential and Global Information for Text Classification" by Yingxin Song on using BERT + Deep Pyramid Convolutional Neural Network (DPCNN) for text categorisation. [2] This article experimented on BERT with other neural network models (CNN variants) for text classification. [3] In the experimentation, CNN variants such as TextCNN and DPCNN captured n-gram features while RCNN used LSTM to capture global sequential information. These models were combined with BERT for improving model's performance. These combined models were tested on eight different datasets such as IMDB for performance evaluation. It was found that BERT + RCNN tends to get best result for sentiment analysis and question analysis while BERT + DPCNN performs best in topic classification. Hence, the paper's results showed that combining n-gram, sequential and global information yielded good performance in text classification.

DPCNN contains N times of repeat blocks for repeated convolutions to produce varying n-gram sequences with fixed number of feature maps till the sequence length is reduced to 1 through downsampling.[4] Each repeated block contains maxpooling (for downsampling) followed by 2 convolution layers of kernel size (3,1) with Rectified Linear Unit (ReLU) as activation function. The above descriptions are visualized in the Figure 1 below.

The downsampling from maxpooling reduces the sequence length of n-grams by half at every repeat block. The inputs to DPCNN are embeddings from BERT with size of (batchsize x 1 x sequence length x dimension). The output is of size (batchsize x feature map size x 1 x 1). In addition, there is a skip connection involved in the repeat block which adds the identity matrix to minimise gradient vanishing issues and allows features to be passed on to deeper networks.[4] Padding is required to keep sequence length constant for skip connection.

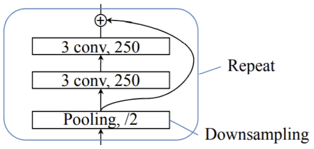


Figure 1: Repeat Block Architecture [V1]

What inspired us is the architecture of DPCNN that can capture multiple n-gram sequences of various lengths due to downsampling. With curiosity, we decided to apply attention mechanism to the DPCCN during decoding to study its impact on overall performance on classification task. The challenge expected here is to have the attention mechanism dynamic as the size of sequences varies.

3 Approach (or Method)

There are a total of 3 models - base model, self attention model (model 1) and global attention model (model 2).

3.1 Base Model

Our base model consists of pre-trained bert integrated with DPCNN. The following Figure 2 displays the architecture of base model.

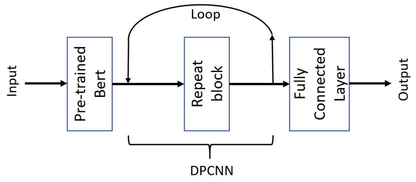


Figure 2: Folded base model architecture

The following components and their functions are common in all 3 models.

1. ***Encoder: BERT*** - BERT here refers to the embeddings and encoder **without** decoder. Due to the nature of BERT, its embeddings contain semantic and synthetic information with the help of Transfer Learning. Thus, the results will be better if these embeddings are applied.

Pre-trained Bert is used as the number of datasets for this project is much less as compared to the number of datasets that Bert has been pre-trained on.

2. ***Decoder: DPCNN*** - Modification is done to DPCNN by keeping kernel size (3,1) as constant in convolution layers. Thus, this project focuses on convolutions of 3-grams (trigram) sequences in DPCNN. The code integration of BERT encoder and DPCNN decoder had been implemented by us and not outsourced. Majority of the code modifications occur when applying the attention mechanisms.
3. ***Fully Connected Layer*** - The last linear layer is for translating hidden neurons into the scores for each number of classes.

3.2 Self Attention Model (model 1)

In our model, we introduce self attention to DPCNN. The intention here is to experiment if relationship development among the convoluted trigrams within each repeat block can impact model performance for classification. Self attention is widely used in other CNN architectures but it has yet to be integrated according to best of our knowledge. Calculation of self attention is similar to existing scaled dot product methodology as summarised in the Figure 3.

The query matrix,key and values are of the same sizes from the same convoluted trigram sequence.The dot product is between every query in query matrix and key (matrix) and are then converted into probabilities through softmax function. The output of matrix multiplication between softmax results and the values produces the trigram embedding sequences with attention.

Hence, the intra-relationship within the trigram sequences of feature maps is embedded in the output. The self attention mechanism is located at the end of every repeat blocks as illustrated below in Figure 4.

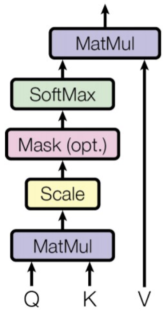


Figure 3: Scaled Dot Product [5]

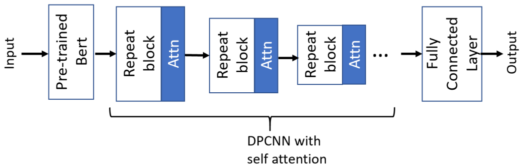


Figure 4: Self attention mechanism (unfolded)

The additional condition here to existing methods is that the size of repeat blocks changes due to the downsampling of sequences. Therefore the self attention mechanism has to be dynamic. This is, however, solved by accessing and temporarily storing the size of convoluted trigram sequence before the repeat block in memory.

3.3 Global Attention Model (model 2)

The global attention mechanism works on dot product attention calculation as observed in self attention previously but on the global features instead of local features. The intention here is to experiment if certain convoluted trigram sequences have higher significance than others that can have impact on model performance for classification. The global features are the varying trigram sequences (with corresponding feature maps) after each repeat block. These are extracted from every repeat block as shown in Figure 5.

As the trigram sequences with feature maps are of different sizes, we additionally introduced

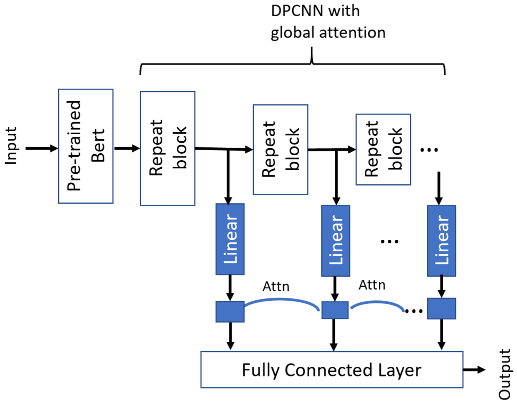


Figure 5: Global attention mechanism (unfolded)

linear layers to output a fixed feature size of 10 for each repeat block. These are then used for scaled dot product attention calculation as summarized in Figure 3. The fixation here avoids unfair advantages to initial repeat blocks as they have more features as representation as compared to the repeat blocks at the end of the process.

The challenging aspect of this dynamic global attention as compared to dynamic self attention is that the output size of every block plot has to known prior to training in order to initialize the linear weights to obtain fixed feature size of 10. Hence, pre-calculation is required to determine the number of repeat blocks and their corresponding output sizes.

4 Experiments

4.1 Data

In this project, 2 datasets were utilised and they were airline tweets [6] and financial data [7] which have 3 classes - 0: negative, 1: neutral and 2: positive. These datasets with 3 classes introduced a certain level of complexity. In all of the datasets, there are additional information such as location, user id and business id. As our project's focus is on the text alone, we only extracted the texts and their corresponding sentiment classes.

4.2 Evaluation method

During training, validation loss from the validation dataset is used to determine best combination of hyper parameters as well as the end of training. The lower the validation, the better the model is assumed to be.

As the objective of this project is on the model's performance on classification, metrics such as precision, recall and f1 scores from the f1 table are utilised for evaluation of results. The higher the f1 score, precision and recall, the better the model is. Precision shows ratio of true positives to the sum of true and false positives. Recall shows the ratio of true positives to the sum of true positives and false negatives. F1 score is calculated by this equation: (2 x precision x recall)/(precision + recall).

4.3 Experimental details

All the 3 datasets have been split into 70% train set, 15% validation set and 15% test set. Class imbalances have been handled with additional proportional weights in the loss function. Hence, more importance is given to loss belonging to class with lower number of text data and vice versa. The determination of sequence length for each dataset is determined visually as observed in Appendix A.1: 25 for tweets data and 40 for financial data.

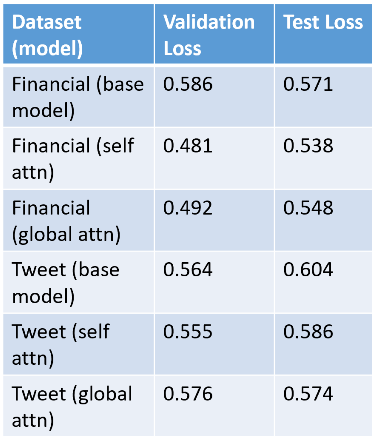
During training, the size of feature maps (CNN channel size) was noted to be not as the major determinant. In two experiments with channel size of 4096 and 50, similar evaluation and test loss were reached. Instead, longer training time was incurred with a high feature map size. The two key hyperparameters were noted to be the learning rate and the corresponding epoch. This can be seen in several of the training experiments where a relatively large learning rate with respect to the number of undergone epoch caused the model to overfit and generalise poorly. For model stability, a low learning rate and small epoch was set. The best combination of hyerparameters are shown in Appendix A.2.

4.4 Results

This section shows 2 forms of results - loss and F1 table. As we have seen in Transformers that attention mechanisms play a significant role in elevating model performance, it is expected to have positive outcome. Nonetheless, we are neutral as we are investigating out of curiosity.

Validation and Test Losses:

Based on the Table in Figure 6 on the right, the validation and test losses for both self and global attention mechanisms are lower than the those of the base model in both the datasets. This suggests that the attention mechanisms in DPCNN have improved the results based on initial observation. For the financial data, the test loss is lower for model with self attention mechanism than with global attention. However, this is vice versa for the tweet data.

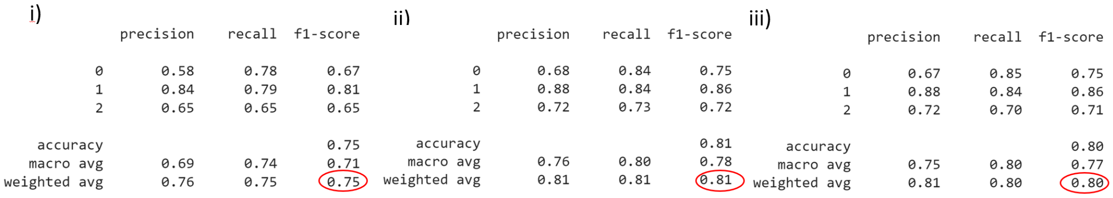


Dataset (model)	Validation Loss	Test Loss
Financial (base model)	0.586	0.571
Financial (self attn)	0.481	0.538
Financial (global attn)	0.492	0.548
Tweet (base model)	0.564	0.604
Tweet (self attn)	0.555	0.586
Tweet (global attn)	0.576	0.574

Figure 6: Val and Test Loss

F1 table:

The precision scores, recall scores and f1 scores are tabulated for finance data which is shown in Figure 7 below. As we are dealing with imbalanced dataset, the weighted average f1 scores are compared (indicated by the red circles). The models with self attention (0.81) and global attention (0.80) have attained higher f1 score as compared to the base model (0.75). In fact, the precision and recall scores for each respective class for the models with attention mechanism are greater than the corresponding scores in the base model. These indicate that the attention mechanisms in DPCNN have proven to improve model performance for this financial dataset.



i)
```
              precision    recall  f1-score

           0       0.58      0.78      0.67
           1       0.84      0.79      0.81
           2       0.65      0.65      0.65

    accuracy                           0.75
   macro avg       0.69      0.74      0.71
weighted avg       0.76      0.75      0.75
```

ii)
```
              precision    recall  f1-score

           0       0.68      0.84      0.75
           1       0.88      0.84      0.86
           2       0.72      0.73      0.72

    accuracy                           0.81
   macro avg       0.76      0.80      0.78
weighted avg       0.81      0.81      0.81
```

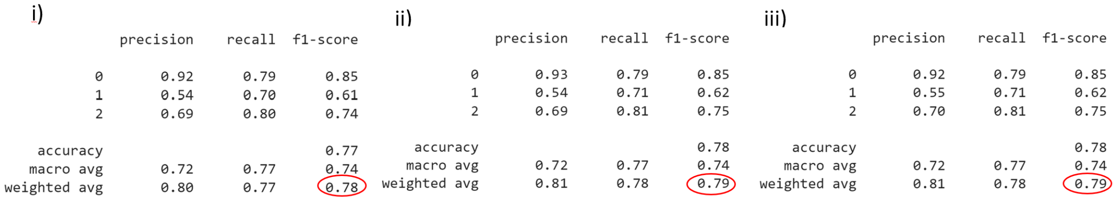
iii)
```
              precision    recall  f1-score

           0       0.67      0.85      0.75
           1       0.88      0.84      0.86
           2       0.72      0.70      0.71

    accuracy                           0.80
   macro avg       0.75      0.80      0.77
weighted avg       0.81      0.80      0.80
```

Figure 7: F1 table for i) base model ii) self attention iii) global attention on Financial data

On the contrary, the attention mechanisms in DPCNN seem to have negligible impact for tweet dataset. With reference to the Figure 8, the weighted average f1 scores (indicated by the circles), precision and recall scores remain rather constant (0.78-0.79). This showed that the model performance for this tweet dataset is invariant to the attention mechanisms.



i)
```
              precision    recall  f1-score

           0       0.92      0.79      0.85
           1       0.54      0.70      0.61
           2       0.69      0.80      0.74

    accuracy                           0.77
   macro avg       0.72      0.77      0.74
weighted avg       0.80      0.77      0.78
```

ii)
```
              precision    recall  f1-score

           0       0.93      0.79      0.85
           1       0.54      0.71      0.62
           2       0.69      0.81      0.75

    accuracy                           0.78
   macro avg       0.72      0.77      0.74
weighted avg       0.81      0.78      0.79
```

iii)
```
              precision    recall  f1-score

           0       0.92      0.79      0.85
           1       0.55      0.71      0.62
           2       0.70      0.81      0.75

    accuracy                           0.78
   macro avg       0.72      0.77      0.74
weighted avg       0.81      0.78      0.79
```

Figure 8: F1 table for i) base model ii) self attention iii) global attention on Tweet data

For a sample of visual comparison of the above, please refer to in Appendix A.4 on constituents of recall where the changes in shape of classes are apparent down the column for Financial data in Figure 16 but shapes remain same down the column for Tweet data in Figure 17.

The results for tweet dataset is very surprising to us as both self and global attention in DPCNN had no impact on the model performance. We will be analysing the attention maps in the next segment to possibly identify the reasons.

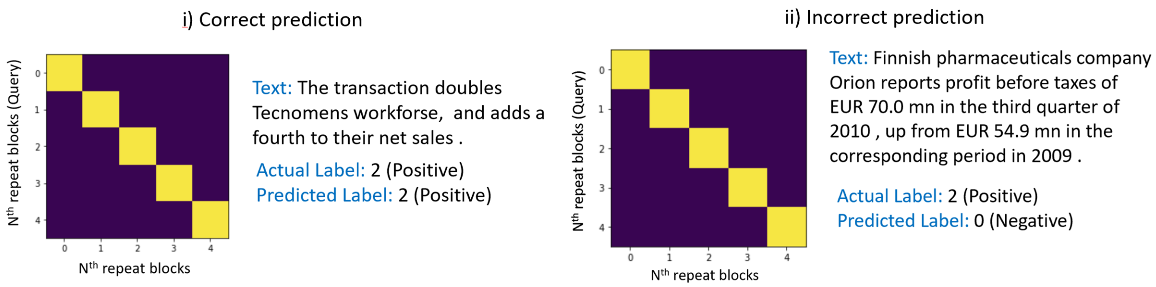
5 Analysis based on Attention Maps

The samples selected for error analysis are mostly the samples with incorrect classification with base model. For attentions on tweet data, some samples have been predicted with same results by base models as the ones with attentions due to negligible improvements. For recap, repeat block architecture can be observed from Figure 1. Only a sample is taken out for demonstration in the report for each subsequent subsections.

5.1 Global Attention

The inputs to global attention mechanism are vectors of 10 from linear layers as seen in Figure 5.

5.1.1 Financial data


i) Correct prediction

Text: The transaction doubles Tecnomens workforce, and adds a fourth to their net sales .

Actual Label: 2 (Positive)

Predicted Label: 2 (Positive)

ii) Incorrect prediction

Text: Finnish pharmaceuticals company Orion reports profit before taxes of EUR 70.0 mn in the third quarter of 2010 , up from EUR 54.9 mn in the corresponding period in 2009 .

Actual Label: 2 (Positive)

Predicted Label: 0 (Negative)

Nth repeat blocks (Query)
0
1
2
3
4
Nth repeat blocks
0
1
2
3
4
Nth repeat blocks (Query)
0
1
2
3
4
Nth repeat blocks
0
1
2
3
4

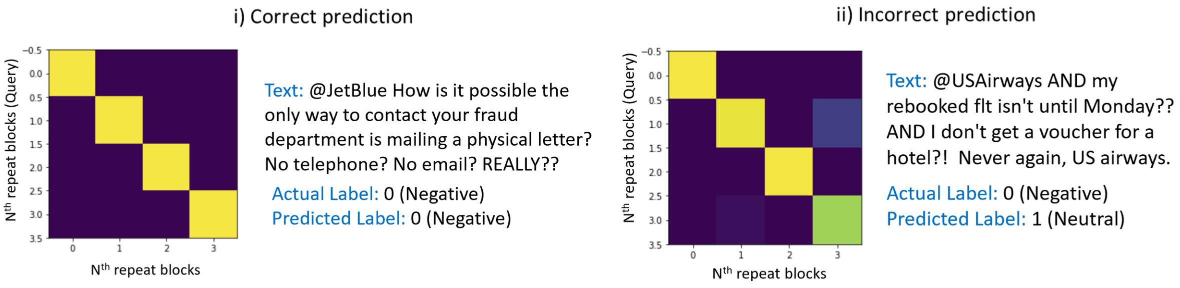
Figure 9: Global Attention Maps for a pair of financial samples

The attention maps in Figure 9 are of 5 by 5 as there are 5 repeat blocks corresponding to the sequence length of 40 as stated initially in Section 4.3. For both correct and incorrect classification results, the attention maps indicate the strong alignment only between their respective repeat blocks. This pattern has been observed for approximately 30 more samples. No relationships among the repeat blocks are observed based on the attention maps. This suggests that the cause of misclassification may not be due to lack of attention. This also justifies the relatively small improvement in the F1 scores in Figure 7 iii) in the previous section as large improvement may require other types of solution.

Upon closer look at the 2 texts, it can also be that the incorrect classification for sample on the right may be due to model's inability to interpret the numerical values unlike the sample on the left only contains words in the text. More statistical analysis is needed for justification.

Nonetheless, the small improvement indicates that the alignment with attention is necessary for better model performance.

5.1.2 Tweet data


i) Correct prediction

Text: @JetBlue How is it possible the only way to contact your fraud department is mailing a physical letter? No telephone? No email? REALLY??

Actual Label: 0 (Negative)

Predicted Label: 0 (Negative)

ii) Incorrect prediction

Text: @USAirways AND my rebooked flt isn't until Monday?? AND I don't get a voucher for a hotel?! Never again, US airways.

Actual Label: 0 (Negative)

Predicted Label: 1 (Neutral)

Nth repeat blocks (Query)
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Nth repeat blocks
0
1
2
3
Nth repeat blocks (Query)
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Nth repeat blocks
0
1
2
3

Figure 10: Global Attention Maps for a pair of tweet samples

The attention maps in Figure 10 are of 4 by 4 as there are 4 repeat blocks corresponding to the sequence length of 25 as stated initially in Section 4.3. Unlike what we observed for financial data, there is a distinct difference between the correct and incorrect classification attention maps based on 30 samples. The incorrectly classified attention map consists of at least 1 repeat block being asscoiated with another repeat repeat block other than itself. As observed on the right on Figure 10, the attention from repeat bock 4 constitutes of approximately 0.8 of itself and 0.2 of attention map from repeat block 2. Perhaps, other forms of attention such as multiplicative with learning parameters may be better for this dataset for better alignment. This probably may be one of the causes for negligible improvement in this dataset.

Upon closer look at the 2 texts, both texts dos not contain explicit words for emotional reference. In fact, there are numerous punctuations and capitalised words that highlight the extreme expression of negative emotion. Perhaps, the model may not be good at finding semantic relationship in such scenarios.

5.2 Self Attention

It it essential to note that the inputs to self attention are not word embeddings but convoluted trigram sequences of feature maps (channel size).

5.2.1 Financial data

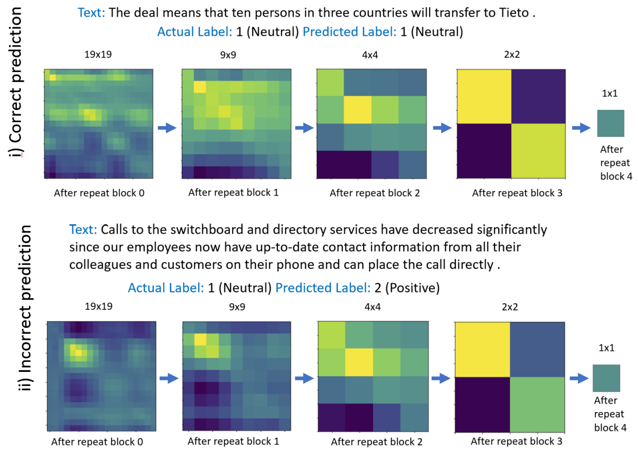


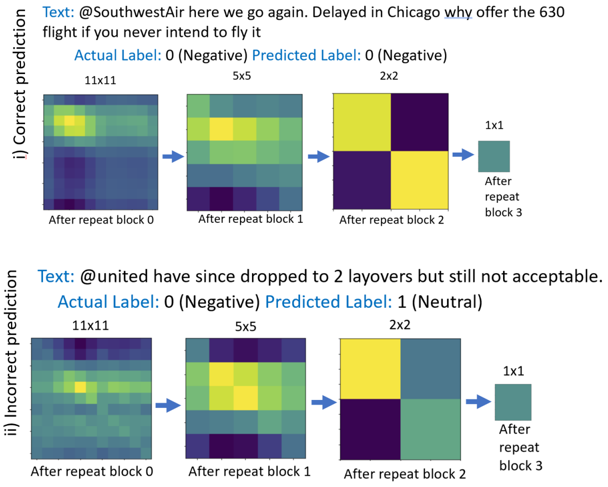
Figure 11: Self Attention Maps for a pair of finance samples

There are 5 attention maps shown for each sequence in financial data as the sequence length used is 40. It is not easy to determine the meaning of the associations here as we are not dealing with tokens directly but rather convoluted trigram sequences after BERT encoder and it is not possible to know from these maps what the these sequnces comprises of especially after downsampling. However, we can still observe generic patterns. Generally, the attention maps for correct prediction consisted higher values at the diagonal locations of the attention maps after initial repeat blocks but not distinct. In other words, the output after the attention is associated to various trigram features (within sequence) of different probabilities based on the maps.

As we progress towards the last few repeat blocks, there is a general trend that the attention maps for correct and incorrect predictions appear to be similar. The contents in the sequences at the nth repeat block will undoubtedly be different. The differences between correct and incorrect predictions are more distinct after initial attention maps. This is probably due to the fact that convolution causes vector representations to merge and thereby diluting the differences. Convoluting trigrams at a time with a stride of 2 means there is will always be 1 common term between the adjacent convoluted trigrams and this overlap can further increase the similarity between adjacent trigrams if convoluted recursively. The downsampling with maxpooling also selects trigram vectors with large values and thus causing the maps to be similar at the end.

In this pair of samples, it could be that the length of text in incorrect prediction is longer than 25 and thus part of text is cut off. On a personal note, we do feel this text has a certain level of positive sentiment as the text outlines the advantages. This brings up an issue of how sentiment analysis can be subjective among individuals.

5.2.2 Tweet data


Text: @SouthwestAir here we go again. Delayed in Chicago why offer the 630 flight if you never intend to fly it

Actual Label: 0 (Negative) Predicted Label: 0 (Negative)

Text: @united have since dropped to 2 layovers but still not acceptable.

Actual Label: 0 (Negative) Predicted Label: 1 (Neutral)

Figure 12: Self Attention Maps for a pair of tweet samples

There are 4 attention maps shown for each sequence in financial data as the sequence length used is 25. Similar observations as in financial data are obtained but with less proportion of higher values at the diagonal locations of the maps. In other words, the probabilities of the associations are scattered for correct prediction and even more scattered for incorrect predictions in the maps after initial repeat blocks (not optimised alignment).Perhaps, other attention mechanisms with learning parameters may provide better performance. However, this alone is not enough to deduce the reason for self attention mechanisms not to work on the tweet dataset.

6 Limitations

The unexpected result may be explained by the different nature of the datasets. The text in financial dataset was written in proper structured English while the text in tweets was expressed in unstructured English with trendy words not part of the English dictionary. More statistical text analysis and lexicon based analysis with POS tagging are required for error analysis to complement the observations made from the attention maps. Furthermore, the DPCNN used in this project may not be the optimised architecture. Hence, more experiments on DPCNN architecture with attention mechanisms is required to identify how the convolution variation such as having bigrams with varying strides affects the attention maps.

7 Conclusion

In conclusion, we have experimented the combined model of BERT and DPCNN with dynamic attention mechanism. The results of running these models on two datasets are contrasting. While the f1 scores of global and self attention mechanism models are significantly better than that of base model in the finance dataset, the f1 scores of attention based models and base model have no significant difference in tweet data. Possible future works include analysis with POS tagging, convolution with varying kernel sizes and removal of attentions for last repeat blocks in self attention.

Bibliography

[1] M. S. J. M. A. A. Ms. Binju Saju, "Comprehensive Study on Sentiment Analysis: Types, Approaches, Recent," IEEE, 2020.

[2] T. Z. Rie Johnson, "Deep Pyramid Convolutional Neural Networks for Text Categorization".

[3] Y. Song, "MIHNet: Combining N-gram, Sequential and Global Information for Text Classification," Journal of Physics: Conference Series, 2020.

[4] R. Johnson and T. Zhang, “Deep Pyramid Convolutional Neural Networks for Text Categorization,” in Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vancouver, Canada, 2017, pp. 562–570, doi: 10.18653/v1/P17-1052.

[5] A. Vaswani et al., “Attention is All you Need,” p. 11.

[6] https://www.kaggle.com/crowdflower/twitter-airline-sentiment (this data here is cleaned after sourced from the following link) crowdflower-airline-twitter-sentiment

[7] https://www.kaggle.com/ankurzing/sentiment-analysis-for-financial-news, Malo, (in the link, they have this acknowledgement) P., Sinha, A., Korhonen, P., Wallenius, J., Takala, P. (2014). Good debt or bad debt: Detecting semantic orientations in economic texts. Journal of the Association for Information Science and Technology, 65(4), 782-796.

Team Contributions

Both team members provided equal contributions to this project.

A Appendix

The sequence length of the tweets data have been selected as 25 based on visual analysis of the Figure below.

A.1 Sequence Length

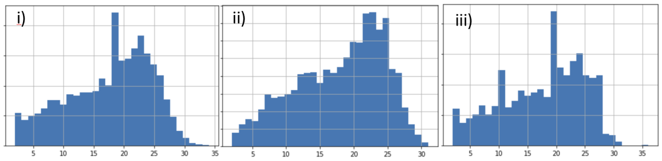


Figure 13: Seq Len Tweets i)train ii)val iii)test

The sequence length of the financial data have been selected as 40 based on visual analysis of the Figure 13 above.

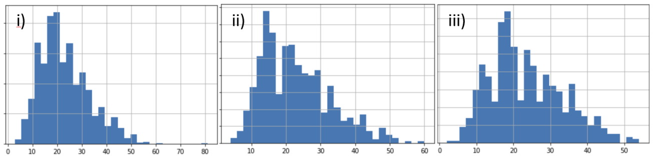
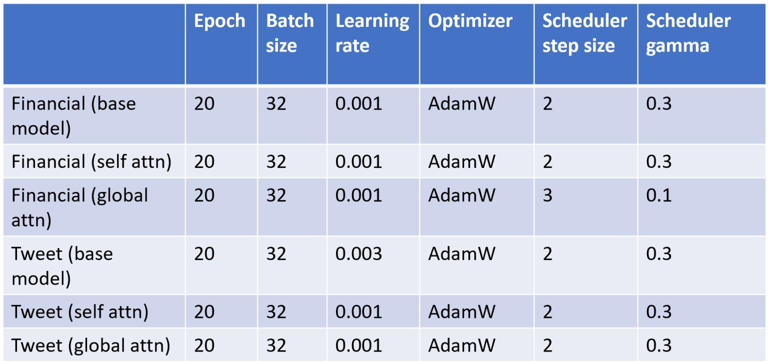


Figure 14: Seq Len Financial i)train ii)val iii)test

The sequence length of the amazon food review data have been selected as 25 based on visual analysis of the Figure 14 above.

A.2 Best Combination of HyperParameters



	Epoch	Batch size	Learning rate	Optimizer	Scheduler step size	Scheduler gamma
Financial (base model)	20	32	0.001	AdamW	2	0.3
Financial (self attn)	20	32	0.001	AdamW	2	0.3
Financial (global attn)	20	32	0.001	AdamW	3	0.1
Tweet (base model)	20	32	0.003	AdamW	2	0.3
Tweet (self attn)	20	32	0.001	AdamW	2	0.3
Tweet (global attn)	20	32	0.001	AdamW	2	0.3

Figure 15: Best hyperparameter combination for each model per dataset

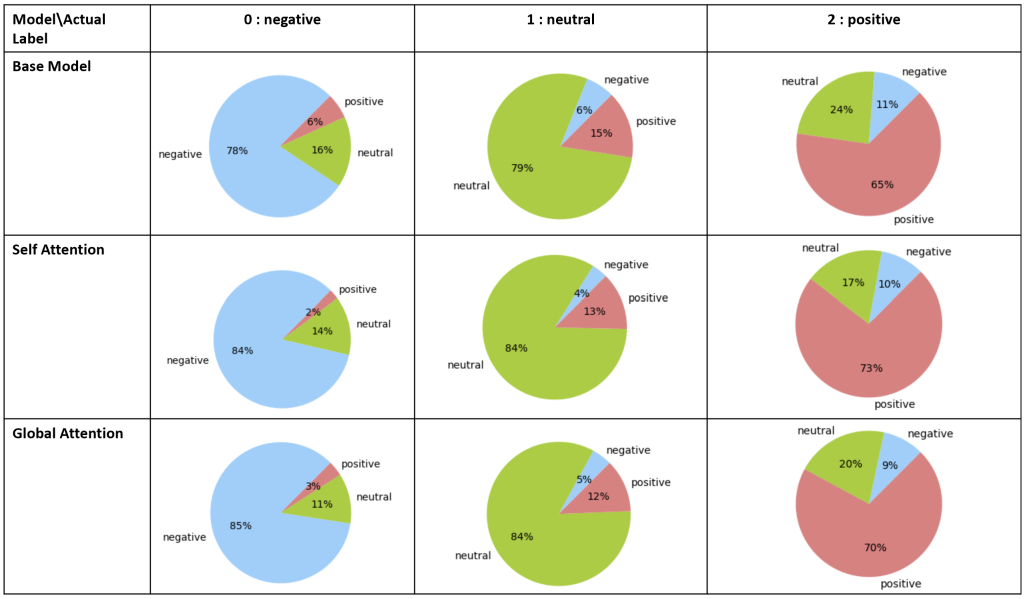
A.3 HyperParameter Tuning Samples

The link below constains some sample results recorded during training.

Link to access some samples:

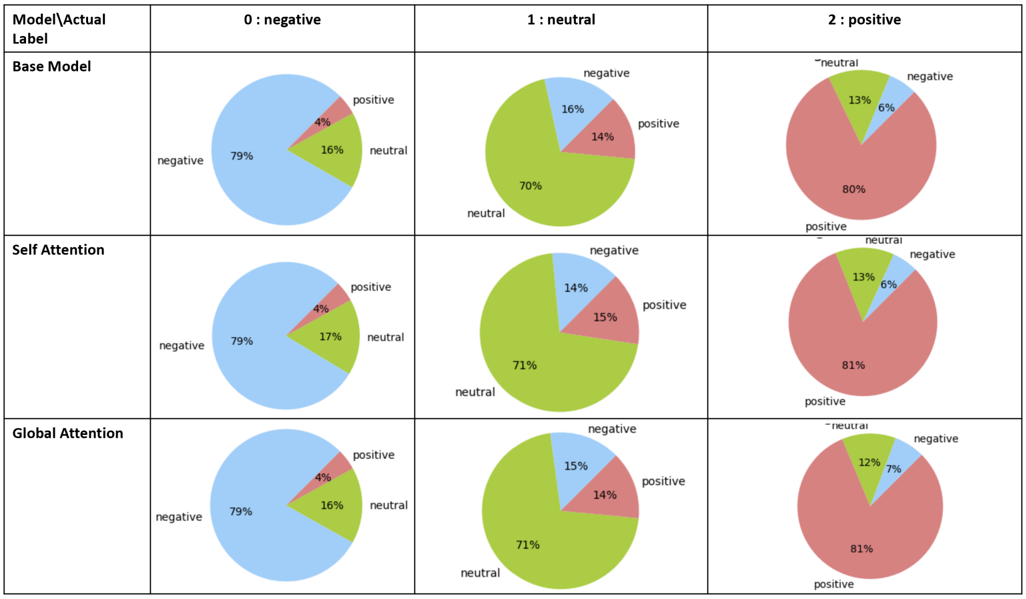
https://drive.google.com/file/d/1bdkXRhxxuUwy0OX1gPOhJY2cNMKcjYYn/view?usp=sharing

A.4 Constituents of Recall



Model\Actual Label	0 : negative	1 : neutral	2 : positive
Base Model	negative 78% positive 6% neutral 16%	negative 6% positive 15% neutral 79%	neutral 24% negative 11% positive 65%
Self Attention	positive 2% neutral 14% negative 84%	negative 4% positive 13% neutral 84%	neutral 17% negative 10% positive 73%
Global Attention	positive 3% neutral 11% negative 85%	negative 5% positive 12% neutral 84%	neutral 20% negative 9% positive 70%

Figure 16: recall on Financial data



Model\Actual Label	0 : negative	1 : neutral	2 : positive
Base Model	positive 4% neutral 16% negative 79%	negative 16% positive 14% neutral 70%	neutral 13% negative 6% positive 80%
Self Attention	positive 4% neutral 17% negative 79%	negative 14% positive 15% neutral 71%	neutral 13% negative 6% positive 81%
Global Attention	positive 4% neutral 16% negative 79%	negative 15% positive 14% neutral 71%	neutral 12% negative 7% positive 81%

Figure 17: recall on Tweet data