Gated Attention Network for Sequence Data

Keya Desai
Rutgers University
Email: kd706@scarletmail.rutgers.edu

Twisha Naik
Rutgers University
Email: tn268@scarletmail.rutgers.edu

Prakruti Joshi
Rutgers University
Email: phj15@scarletmail.rutgers.edu

***Abstract*—**Attention mechanisms have proved to be very useful for improving the performance of deep learning models. A lot of recent research focuses on the dynamic mechanism that adapts the network based on the input provided. We have re-implemented the recent technically challenging paper "Not All Attention Is Needed: Gated Attention Network for Sequence Data" [1]. The paper applies the proposed model GANet for the task of text classification. We have compared the GANet model with baseline models such as LSTM, BiLSTM and also attention models such as soft attention. We have evaluated the models based on two parameters - the classification accuracy and the density of the attention network. The GANet model of outperforms all the baselines using lesser computation by giving attention to only specific parts of the sequence. Hence, it proves to be a potential improvement in more complex applications and networks like the transformers.

Key Words - attention model, text classification, gated attention

I. Introduction

Deep learning has proved its prowess in multiple domains and applications. Though traditional architectures have fixed network, the advent of attention mechanism has enhanced the performance of deep learning models, specially in natural language processing and computer vision. The attention mechanism dynamically adapts the computation based on the input. Another parallel research in case of CNNs is the dynamic network configuration [2] having dynamic connections which are selectively activated in an input-dependent fashion. There are less attempts in the field of RNN or LSTM to apply similar configurations.

The proposed model Gated Attention Network (GANet) in the paper aims to combine both the forms of dynamic mechanism for the task of language modelling. The main motivation is to emphasis that not all the computation is needed in the traditional attention mechanism. Local attention tries to overcome this computational aspect by fixing a window to attend. However, in longer sequences, considering only a part of the input is not sufficient. This Gated attention mechanism combines the advantages of soft attention and local attention by providing a middle ground. The model processes the entire input sequence, however dynamically selects the parts of the input to be included in the computation. Similar to human visual system where only part of the scene is required to analyse certain questions, the theory levitates on the fact that only certain parts of the input are related to the output and subsequently lesser computation is involved. The dynamic attention connections leads to a sparser network.

The detailed explanation of the model is given in Section III. In summary, our work on the re-implementation is as follows:

- Coding of the GANet model along with the Gumbel softmax with configurations similar to the paper for text classification. The datasets have been selected such that the model can be analysed for both shorter sentences (TREC) [1] and longer sentences (IMDb)[2].
- Comparison of the GANet model with baseline sequence models of LSTM, BiLSTM and BiLSTM with soft attention. Further experiments have been conducted on different configurations of Auxiliary network. Analysis on accuracy and density of the model and its interpretability on short and long sentences has been given.
- As a novel approach, potential problems with the loss function used in the paper have been identified. And experiments have been conducted with variations of the loss function.
- Determination of the optimal hyperparameters and MLP network for each dataset. These details are missing in the original paper. Analysis of the datasets have been employed in order to determine the ideal sequence length and batch size. The study of classification results has been extended from the paper.

II. Related Work

The groundbreaking research of [3] and [4] describes the attention mechanism and the transformer model. The dynamic network configuration has been applied for CNNs [2]. An interesting line of research in the context of Natural Language Processing is the Gated Attention Mechanism. This method adjusts attention connections in attention networks. The early work on gated attention mechanisms experiments the weighted inter-alignment of the sequence coupled with the multi-hop architecture [5], [6]. One of the recent research paper [7], which is an advancement from the compare and aggregate method [8]; implements a gated self-attention model. The gated mechanism is combined with a memory network yielding the architecture: Gated Self-Attention Memory Network (GSAMN).

III. Method

A. GANet model

The architecture of the proposed GA-Net for classification tasks is as shown in Fig. 1. On the right is the Backbone attention network which processes the input sequence converted

[1]https://rdrr.io/cran/textdata/man/dataset$_trec.html$

[2]https://ai.stanford.edu/ amaas/data/sentiment/

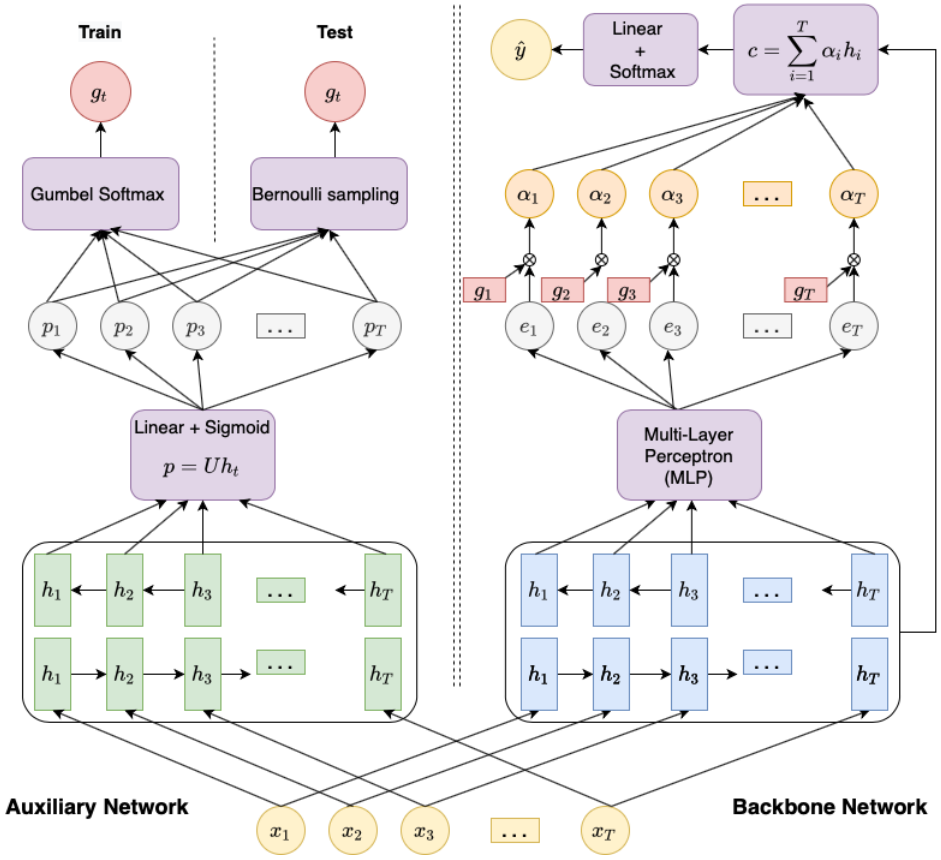

Fig. 1: GA-Net Architecture

into word embedding using pre-trained Glove and produces an output classification target label. In the paper [3], the backbone network is a BiLSTM network with attention mechanism. It has an attention weight associated with each individual input word similar to soft attention. The network on the left is a smaller BiLSTM auxiliary network which processes the same input sequence to dynamically select the important parts of the input to generate the output. The auxiliary network generates a binary gate for each word of sentence to control which input words the backbone network should attend. The gates g_t are sampled using Bernoulli sampling from the probability values computed by the auxiliary network. Gate is open when g_t = 1, otherwise closed. The weighted combination of the attention weights for which the gates are open are used to produce the final output. In the paper, the hidden dimensions of backbone BiLSTM (2-layer) and auxiliary BiLSTM (1-layer) are both 100. We have implemented the model with the same parameters as mentioned.

B. Training and Testing

1) Gumbel Softmax: The gates returned by the Auxiliary network are a result of a stochastic process and are binary. Hence, backpropagation through these gates while training is not possible. To make the learning possible, Gumbel Softmax [9] has been used which removes the stochastic node of Bernoulli sampling from the computational graph using the reparameterization trick and moving the stochasticity to a secondary node (details in Appendix VI-B).

2) Loss function: The loss mentioned in the paper is as follows:

$$L = -\sum_{k} y_k log y_k + \frac{\lambda ||G||_1}{T}$$

where, T is input sequence length

First term is the Cross-entropy (CE) loss which measures the classification accuracy where y_k is the ground-truth label for k-th class. The second term is associated with the number of gates that are open. It is an L_1 norm regularizer over all gates, where λ decides the trade-off between CE loss and L_1 norm. The second term aims to make the network more sparse by turning on less number of gates.

Issue: Even with a very small value of λ, the loss function penalises all the gates to be closed by making the probability values associated with it small. This results in cases when all the gates are predicted to be closed, leading to erroneous predictions.

Solution: The paper does not acknowledge this potential practical problem of stochastic behavior. As a solution the following approaches are tried to modify the loss function with the aim of achieving desired behavior.

1) Naive approach: Whenever it is the case that all gates are closed during testing, make all the gates open explicitly. This ensures results at par with soft attention but with increased density than GANet.
2) Threshold number of open gates: In the loss function, instead of simply minimizing the number of open gates, keep a lower limit to the number of the open gates (k), making the model learn the behaviour of keeping certain gates open. However, this approach does not bring significant improvement in the results.
3) Average sampling results: To avoid all the gates being closed and neutralizing the stochasticity involved, we sample the gates multiple times and average the results during training and testing. This idea is inspired from the concept of ensemble models [10]. The problem is avoided and the results are slightly better with little additional computational cost as the model learns the expected behavior.

3) Convergence criteria: Initially, the stopping criterion for training was set to a predefined number of epochs (100). Then it was set to stop the training when the validation loss did not decrease for certain number of epochs (10 or 20), similar to what is done in the course assignments.

4) Evaluation: During testing and validation, the gates generated by auxiliary network are sampled using the Bernoulli distribution of the probabilities to make the gates binary. Evaluation is done using training, validation, testing loss and accuracy of the classification output. In addition, density of the network is considered, which is an important factor highlighting the motivation of the paper. It gives a notion about the average number of words that the network attends to and gives an idea about the spareness of network. For class wise analysis, precision, recall, F1 score and confusion matrix are computed. (Analysis in Appendix VI-E)

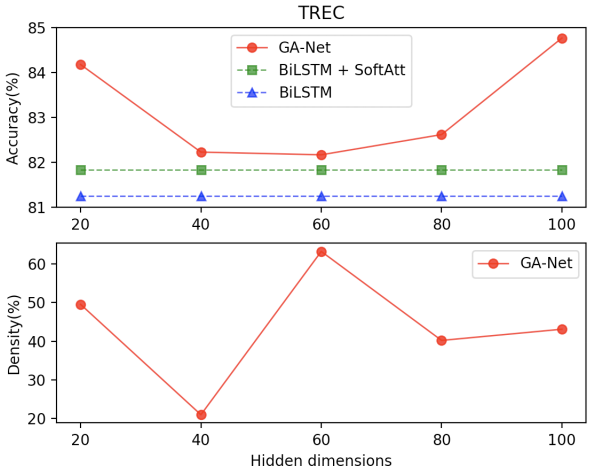

Fig. 2: TREC classification accuracy and attention density for different hidden size of auxiliary network

IV. Experiments

1) Framework: We have implemented the model using Pytorch. Adam optimizer is used to adjust the learning rate. For the inputs, pre-trained 100-dimensional GloVe word vectors [11] are used to generate a word embedding. Text preprocessing was done using *torchtext*.

2) Hyperparameter tuning: The optimal values of the hyperparameters are determined by adapting cross-validation for each dataset which the original paper does not provide. Experiments have been performed for 4 hyperparameters, over a range, as shown in Table I. The optimal value for each hyperparameter is given in the table. Detailed analysis of sequence length is given in Appendix VI-A.

Hyperparameter	Values	Optimal Value
Learning Rate (α)	[2e-5, 1e-4, **2e-4**, 5e-4, 1e-3, 2e-3, 5e-3]	2e-4
Temperature (τ)	[0.5, 1, **1.5**, 2.0]	1.5
Regularization parameter (λ)	[4e-6, 5e-6, 1e-5, 1e-4, **4e-4**, 5e-4]	4e-4
Batch Size	[8, **16, 32**, 64, 128]	16 (Trec), 32 (IMDb)
Sequence length	[8, **10**, 12] [100, 200, **300**, 500]	10 (Trec), 300 (IMDb)

TABLE I: Hyperparamter search

3) Comparison with Baseline models: To validate the significance of the GANet model, we have compared its results with that of the baseline models, as done in the paper. The comparison is done with the basic sequence models of LSTM, BiLSTM and with soft attention where attention is given to all the words in the input sequence. The results for the accuracy and the density are given in Table II. All the networks are trained and tested using the same configurations. GANet outperforms all the baseline models.

4) Auxiliary configuration variations:

- Hidden dimension (BiLSTM): Hidden size of the auxiliary network is varied to understand the complexity of the auxiliary network required to generate the desired results. The results for different hidden dimension size is given in Fig. 2. It can be observed that even a small network with hidden dimension of 20 surpasses the accuracy of the baseline models.
- Model variations: In the paper, the authors have experimented with different configurations of auxiliary network. Following which we experimented with a 2-layer Feedforward network, LSTM, BiLSTM and soft attention model as the auxiliary model. All the variations of auxiliary model outperforms the baseline model (Table II). Using Soft Attention as auxiliary model works the best in terms of accuracy. BiLSTM gives comparable accuracy but with lesser density making it the most suitable auxiliary network.

5) Interpretability: One of the aims of this project is to analyse the functioning of the gates in short and long sentences. The attention weights of GANet are compared with soft attention weights for the TREC data in Fig 3. The class of the first example is *location*. GANet selectively assigns more weight to the words *(where, located)* necessary for understanding the input. The weight distribution is sparse. Whereas soft attention assigns weights even to the non-important words like *is, the , university* which do not determine the output class.

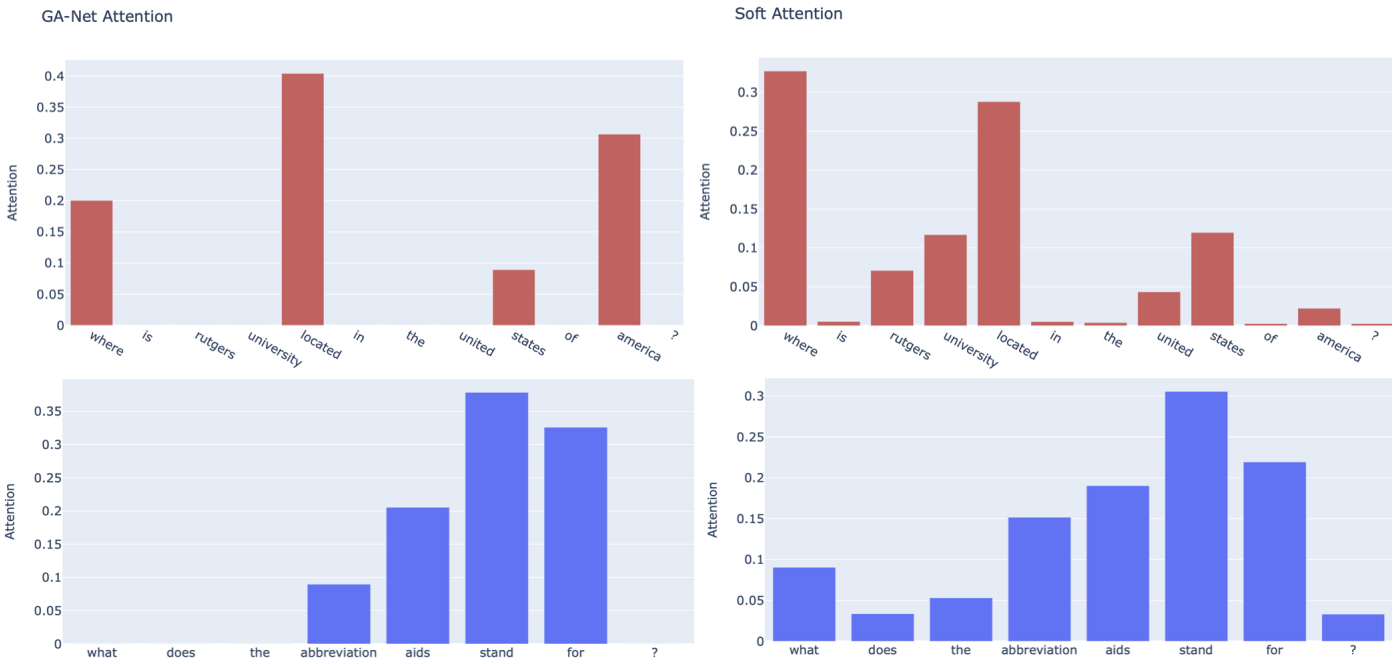

Fig. 3: Comparison of soft attention and GA-Net Attention for TREC data

Similarly, the class of the second sentence is *abbreviation*. GANet successfully captures the intuition by giving focused attention to a group of words. Local attention would have given similar results on shorter sentences where a window is able to capture essence of sentence (example 2 in Fig 3. However, it gives poor results in longer datasets like IMDb. The part of the sentences which gets attended to in longer sentences using GANet is given in Fig 4. GANet is able to correctly predict the output by using one third of the sentence and significantly reduces computations.

Example:

Now this is **more like** it! **One of the best movies** I have ever seen! **Despite** it made **very well on all aspects**, this movie was **put down** solely for **not being** too **historically accurate**. Loosen up! There are tons of historical movies out there that were forgiven for **not being too historically accurate** and many of them do not even come close to how grand, how **entertaining** and how **captivating** this movie was! Now this is what a movie ticket is all about! ..... If the viewer of this movie is open minded and has the ability to separate **politics** from art,you will find this movie not only **one of the best classics**, but also **one of the best movies** of all time. I rate it the **second best western ever**, right behind Wayne's The Cowboys

Prediction: Positive **Density**: 0.371

Fig. 4: Attention result on IMDB. The bold parts indicate the words for which gate is open.

6) MLP configuration: In the paper, the configuration of MLP network in the backbone network is not mentioned. Hence, we experimented with the number of layers, hidden states and the activation functions for each layer and found that two feedforward neural layers with 32 and 1 as hidden dimensions respectively generates good results. The activation function used between the two layers is *Tanh*.

Network	TREC (SL = 10)			IMDB (SL = 300)		
	Train	**Test**	**Density**	**Train**	**Test**	**Density**
LSTM	0.956	0.812	-	0.972	0.755	-
BiLSTM	0.969	0.815	-	0.973	0.765	-
BiLSTM (Soft Attn.)	0.972	0.821	1	0.979	0.772	1
GANet (LSTM + LSTM)	0.942	0.830	0.616	0.985	0.780	0.332
GANet (BiLSTM + FF)	0.937	0.832	0.452	0.985	0.781	0.327
GANet (BiLSTM + LSTM)	0.954	0.836	0.431	0.986	0.785	0.329
GANet (BiLSTM + Self Attn)	0.96	**0.845**	0.536	0.988	0.788	0.573
GANet (BiLSTM + BiLSTM)	**0.964**	0.842	**0.371**	**0.989**	**0.791**	**0.324**

TABLE II: Results of classification on TREC and IMDB

V. Conclusions and Future Scope

The proposed model surpasses the baseline models in terms of classification accuracy. A small auxiliary network is able to significantly improve the results in terms of density, accuracy and interpretability, thus rendering a sparser network. The study in the paper is extended by experimenting with the loss function, hyperparameters and the classification results. On analysing the misclassification results on ambiguous sentences (Appendix VI), combining existing aspect based sentiment analysis with GANet seems to be a good research direction. Owing to the GANet's performance in text classification, it has a great potential in more complex language modelling networks such as tranformers and seq-to-seq applications like question answering, entity recognition etc.

REFERENCES

[1] L. Xue, X. Li, and N. L. Zhang, "Not all attention is needed: Gated attention network for sequence data," *arXiv preprint arXiv:1912.00349*, 2019.

[2] Z. Chen, Y. Li, S. Bengio, and S. Si, "Gaternet: Dynamic filter selection in convolutional neural network via a dedicated global gating network," *arXiv preprint arXiv:1811.11205*, 2018.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[4] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," *arXiv preprint arXiv:1508.04025*, 2015.

[5] G. Shen, Y. Yang, and Z.-H. Deng, "Inter-weighted alignment network for sentence pair modeling," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 1179–1189.

[6] B. Dhingra, H. Liu, Z. Yang, W. W. Cohen, and R. Salakhutdinov, "Gated-attention readers for text comprehension," *arXiv preprint arXiv:1606.01549*, 2016.

[7] S. Yoon, F. Dernoncourt, D. S. Kim, T. Bui, and K. Jung, "A compare-aggregate model with latent clustering for answer selection," in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 2019, pp. 2093–2096.

[8] S. Wang and J. Jiang, "A compare-aggregate model for matching text sequences," *arXiv preprint arXiv:1611.01747*, 2016.

[9] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with gumbel-softmax," *arXiv preprint arXiv:1611.01144*, 2016.

[10] Y. Yuan, Y. Lyu, X. Shen, I. W. Tsang, and D.-Y. Yeung, "Marginalized average attentional network for weakly-supervised learning," *arXiv preprint arXiv:1905.08586*, 2019.

[11] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543. [Online]. Available: http://www.aclweb.org/anthology/D14-1162

[12] C. Sun, L. Huang, and X. Qiu, "Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 380–385. [Online]. Available: https://www.aclweb.org/anthology/N19-1035

[13] M. Saeidi, G. Bouchard, M. Liakata, and S. Riedel, "SentiHood: Targeted aspect based sentiment analysis dataset for urban neighbourhoods," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. Osaka, Japan: The COLING 2016 Organizing Committee, Dec. 2016, pp. 1546–1556. [Online]. Available: https://www.aclweb.org/anthology/C16-1146

VI. APPENDIX

A. Data Description

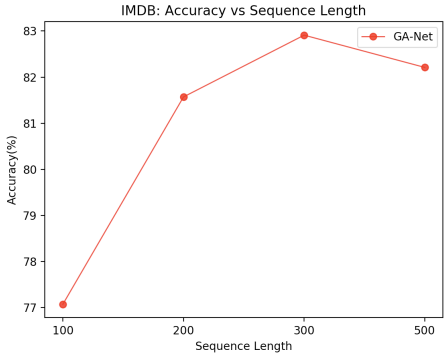

Fig. 5: IMDB Test Accuracy vs Sequence Length

Experiments are performed on two datasets - one with short sentences (TREC data) and second with much longer sentences (IMDb data). The descriptions about the datasets is given in Table III.

1) TREC: Task is to classify each question into 6 categories - Abbreviation, Description, Entities, Human beings, Locations and numeric values. Distribution of classes in the training data is as follows: 'ENTY': 1250, 'HUM': 1223, 'DESC': 1162, 'NUM': 896, 'LOC': 835, 'ABBR': 86
2) IMDB: It is a sentiment analysis dataset with two positive and negative classes and the following data distribution: 'POS': 8808, 'NEG': 8692
3) Sequence length analysis:
Sequence length proved to be an important hyperparameter for language modelling. Using *torchtext* in Pytorch, the data is preprocessed, converted to lowercase and padded to a fixed length which is the input size feeded to the BiLSTM networks. We analysed the datasets and studied the distribution of the lengths of the sequences. The best results were achieved when the sequence length was between average length and the 75th quantile length. The result for IMDb can be seen in Fig 5

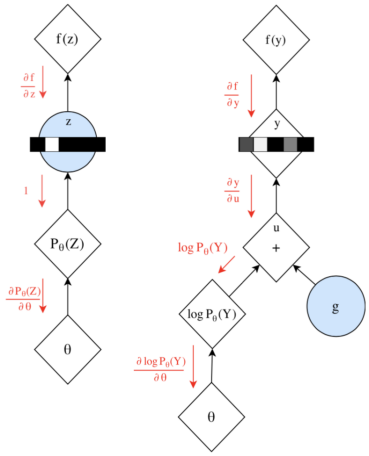

Fig. 6: Gumbel Softmax backpropagation

B. Gumbel Softmax

Gumbel softmax uses the reparameterization trick to remove the stochastic node (in blue color) and makes it possible to backpropagate the loss. Let p_t be the probability of gate g_t being 1 (open). While testing, we pick $g_t = Bernoulli(p_t)$. But for training the Gumbel Softmax is included as follows:

$$\begin{aligned} g_t &= one\_hot(argmax_i p_{t,i}, i = 0, 1) \\ p_{t,0} &= 1 - p_t, p_{t,1} = p_t \end{aligned} \tag{1}$$

Using Gumbel softmax, this can be written as:

$$\begin{aligned} \hat{g}_t &= [\hat{p}_{t,i}]_{i=0,1} \\ \hat{p}_{t,i} &= \frac{exp((log(p_{t,i} + \epsilon_i)/\tau)}{\sum_{j=0}^{1} exp((log(p_{t,j} + \epsilon_j)/\tau)} \end{aligned} \tag{2}$$

ϵ is a sample from Gumbel (0,1) distribution and τ is the temperature. When τ is less than 1, the distribution is close to one-hot and when it is greater than 1, distribution is close to uniform. Hence, the sampling is moved from Bernoulli to Gumbel.

Dataset	Train	Test	Classes	Type	Average length	Median length	75th quantile length
TREC	6k	500	6	Question classification	9.89	9.0	12
IMDB	25k	25k	2	Sentiment analysis	230.78	173	280

TABLE III: Dataset Details

C. Probability and gate relation

The gate probability p_t is used to sample the gate g_t using Bernoulli and distribute the attention among gates that are 1. From figure 7, it is observed that the gated attention is focused only on certain words for which the gated probabilities p_t is fairly high. NOTE: Due to the sampling involved in the process, it is possible that the gates corresponding to high probability are closed and vice-versa.

D. Training

The convergence criteria set while training was to stop after 10 consecutive iterations of increasing validation loss, in order to avoid over training of the model. Validation loss is used as a criteria rather than the accuracy because loss gives a better picture of both density and the prediction precision rather than just the accuracy. It can be seen from the graphs in figures 8 and 9,that the training loss keeps on decreasing but we save the model when the validation loss is the least.

E. TREC Classification Result analysis

To understand the class-wise results, the evaluation metrics of confusion matrix, precision, recall and F1-score have been employed. The following shows the test result for GANet on TREC data:

Confusion matrix: $\begin{bmatrix} 61. & 1. & 1. & 0. & 2. & 0. \\ 5. & 74. & 7. & 0. & 8. & 0. \\ 1. & 7. & 129. & 0. & 1. & 0. \\ 2. & 7. & 2. & 98. & 4. & 0. \\ 1. & 4. & 4. & 0. & 72. & 0. \\ 0. & 1. & 3. & 0. & 0. & 5. \end{bmatrix}$

Precision: $\begin{bmatrix} 0.87 & 0.78 & 0.88 & 1.0 & 0.83 & 1.0 \end{bmatrix}$

Recall: $\begin{bmatrix} 0.94 & 0.79 & 0.93 & 0.87 & 0.89 & 0.56 \end{bmatrix}$

F1 Score: 0.86

All the classes have a precision of atleast 78%, which indicates that the misclassification is on the lower end. Accept for the last class of 'Abbreviation', every other class has a recall of atleast 79%. The number of examples of the 'Abbreviation' class is low and hence not all sentences belonging to that class are identified correctly, resulting in a low recall of 55%. The F1-score is high at 86%. Hence, it can be concluded that GANet performs well in a classification task.

F. Misclassification Analysis

We extended the study for classification results to analyse the misclassifications in case of the sentiment analysis data (IMDb). We found some interesting misclassifications on ambiguous test sentences. One such example is:
"The acting was good but the direction was horrible. Even great performances could not save the bad script. The movie can be seen once."

The accuracy of GANet for such ambiguous sentences is better than other models. These misclassifications are a part of problem in language modelling known as *Aspect based Sentiment Analysis*. We researched about the problem and found the current advances in this domain: [12]. Due to time limitations, we could not apply GANet to Sentihood dataset [13]. However, integrating dynamic gated mechanism in the recent architectures for aspect based sentiment analysis is a potential research area.

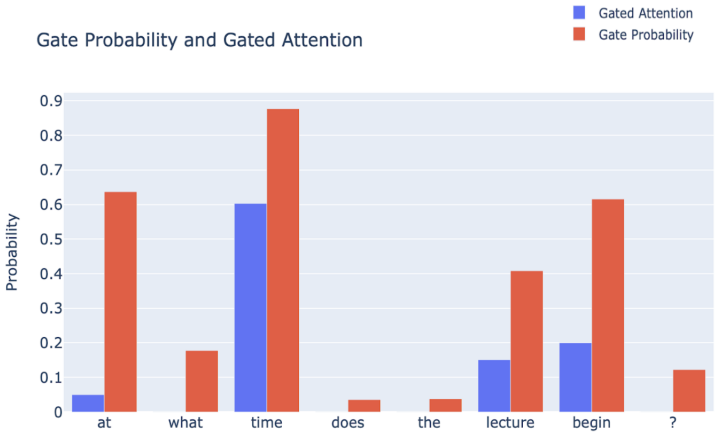

Fig. 7: Gate probabilities and final attention weights

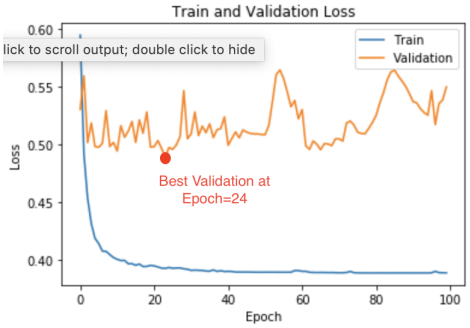

Fig. 8: IMDB Training Loss

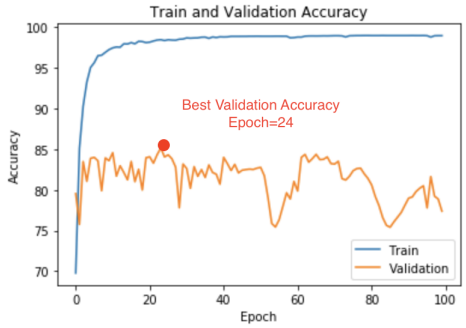

Fig. 9: IMDB Training Accuracy