CS533: Natural Language Processing
Project Proposal
Gated Attention Network for Sequence Data

Keya Desai kd706
Twisha Naik tn268
Prakruti Joshi phj15

Abstract

Attention mechanisms have proved to be very useful for improving the performance of deep learning models across various domains specifically in language modelling. Traditional attention mechanisms either focus on the entire sequence or just a small span of words in the vicinity (local attention). The paper we aim to replicate (Xue et al., 2019) introduces the idea of gated attention which adds a binary gated mechanism that zeros out the attention of less relevant part of the input sequence. The output of the gate is determined by an auxiliary network which is a BiLSTM model in the original paper.

1 Problem

With the advancements in the neural network, the many tasks in computer vision and natural language processing are done efficiently. Having achieved great success in the accuracy of tasks, a lot of recent research focuses on the dynamic mechanism that adapts itself based on the input provided. As a part of the project for the course of CS533 Natural Language Processing, we propose to implement one such dynamic method for attention called gated attention. We will follow the methods and models described in the reference paper (Xue et al., 2019).

The basic limitation of the attention model is that it attends to all the parts of the input sequence even though majority of the input parts might not be relevant to the output target. Thus, such detailed attention is not needed especially for longer sequences. Smaller weights may be assigned to the words that are actually pertinent to the output target due to distribution of the weight to all the parts of the input sequence. The importance of the problem is evident from the results achieved by the reference paper as shown in Fig. 1.

2 Goal

Our goal is to re-implement the technically challenging NLP paper "Not All Attention Is Needed: Gated Attention Network for Sequence Data" (Xue et al., 2019) and test the model on text classification tasks of SST-2 and TREC datasets.

The implementation sub-tasks are:

- There will be two different networks in the complete model. One will be auxiliary network to predict the values of gate that defines what attention components are necessary. The second model is what the paper calls the backbone network which is a bi-directional LSTM model with attention. This model will take the input from auxiliary network and predict the final classification output. We aim to implement the different combinations of the backbone and the auxiliary network for comparison as described in the paper.

- In the auxiliary model, the gates are binary. This makes the normal backpropagation method invalid due to discontinuity. To overcome this problem, the training is done by implementing Gumbel Softmax idea proposed in 2016 (Jang et al., 2016).

- Both the main and auxiliary networks are trained together by using a combined final loss function given by:

$$Loss = \sum_{k} y_k log(\hat{y}_k) + \frac{\lambda \left\| G \right\|_1}{T}$$

The first term in the equation is the cross-entropy loss where y_k is the target label and $\hat{y}_k$ is the predicted label. The second term in the equation is an l_1 norm regulizer over all gates (G) where T is the length of input

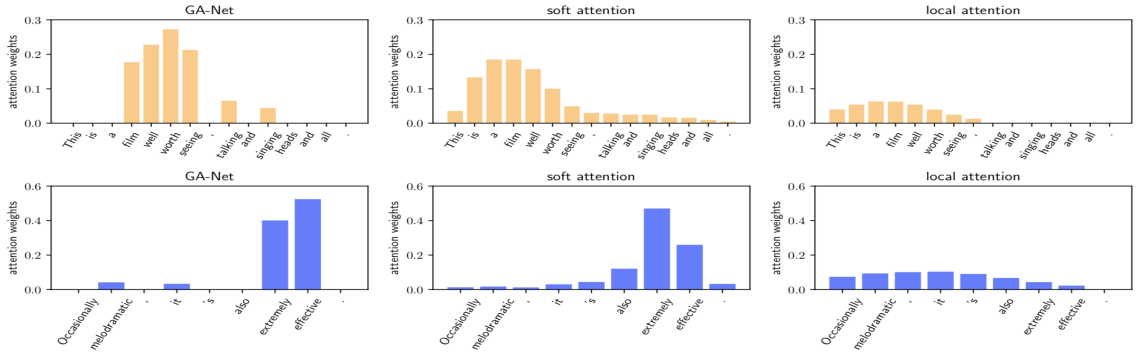

Figure 1: Results of proposed GA-Net model

sequence and λ is a hyperparameter deciding the trade-off between two terms.

3 Achievability

- **Computing Resources**:
  To train the model, we would be utilising the GPU resources provided by Google Cloud.

- **Dataset**:
  We aim to run experiments of the following datasets:

  1. **TREC** This dataset is for question classification. The task is to classify each question into 6 categories - Abbreviation, Description and abstract concepts, Entities, Human beings, Locations and numeric values. The data has 5452 training samples and 500 testing samples. [1]
  2. **SST-2** Stanford Sentiment Treebank v2. It is a collection of movie reviews and the task is to classify positive/negative reviews. Version 2 of the dataset, which we will be using, has no neutral reviews, resulting in binary labels. The data contains 9,613 training samples and 1,821 testing samples. [2]

- **Code framework**:
  We will use Python with PyTorch to implement the network.

- **Proposed Model Architecture**:
  The architecture of the proposed GA-Net for classification tasks is as shown in Fig. 2.

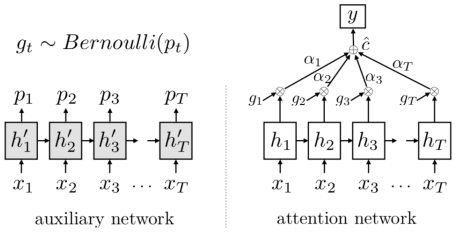

Figure 2: Proposed Architecture

On the right is the Backbone attention network and on the left, there is a small auxiliary network producing a series of probabilities. The gates g_t are sampled from the output of auxiliary network and are binary. Gate is open when g_t = 1, otherwise closed.

In the original paper, the hidden dimensions of backbone BiLSTM and auxiliary BiLSTM are both 100. We might have to scale it down based on the computation required for training. We will be using Adam optimiser as suggested in the paper. We will be fine-tuning the hyperparameters such as learning rate, batch size, temperature to be used in Gumbel-Softmax and λ used in the loss function based on our experiments. For the inputs, pre-trained 100-dimensional GloVe word vectors [3] will be used. A possible extension is applying transfer learning from pre-trained models like BERT, $BERT_{base}$.

- **Supervision setting**:
  We will be using labeled datasets for training and supervised learning technique to train the model.

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

[2] https://github.com/AcademiaSinicaNLPLab/sentiment$_d ataset$5.

[3] glove.6B, pretrained from Wikipedia 2014 and Gigaword 5. https://nlp.stanford.edu/projects/glove/

4 Related Work

The attention mechanism has enhanced the performance of deep learning models in various applications. The research papers (Vaswani et al., 2017) and (Luong et al., 2015) describe the attention mechanism and the transformer model utilising the attention mechanism. We researched about the current advancements in the attention model and its modifications. The dynamic network configuration has been applied for CNNs (Chen et al., 2018). It has a different mechanism than an attention mechanism where it selectively activates a part of the network at a time in an input-dependent manner.

An interesting line of research that we found in the context of Natural Language Processing is the Gated Attention Mechanism. This method combines the dynamic features explained above by dynamically adjusting 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 (Shen et al., 2017), (Dhingra et al., 2016). One of the recent research paper (Yoon et al., 2019), which is an advancement from the compare and aggregate method (Wang and Jiang, 2016); implements a gated self-attention model. The gated self attention model extends the attention mechanism by calculating a real vector gate to control the flow of information in terms of the attention weights, instead of a scalar value. This gated mechanism is combined with a memory network yielding the architecture: Gated Self-Attention Memory Network(GSAMN).

Another interesting ongoing research in NLP that we explored and considered for the project is compressing the huge pre-trained models like BERT (Devlin et al., 2018) to smaller models so that they can run efficiently even in absence of powerful computational resources. BERT serves as a benchmark model for a lot of NLP tasks with its novel learning technique of masking (MLM) and Next Sentence Prediction (NSP). We did a thorough research about various models such as TinyBERT (Jiao et al., 2019), DistilBERT (Sanh et al., 2019), MobilbeBERT (Sun et al.). Each of these models require significant computation resources for knowledge distillation and fine-tuning for task-specific data. Hence, due to lack of computational power, we decided to implement another idea.

5 Future Work

- Extend the idea of dynamic network configuration to more complex attention-based models, such as transformers and seq-to-seq models.
- Changing the structure of LSTM gates internally for better performance. (Gu et al., 2019).

References

Zhourong Chen, Yang Li, Samy Bengio, and Si Si. 2018. Gaternet: Dynamic filter selection in convolutional neural network via a dedicated global gating network. *arXiv preprint arXiv:1811.11205*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Bhuwan Dhingra, Hanxiao Liu, Zhilin Yang, William W Cohen, and Ruslan Salakhutdinov. 2016. Gated-attention readers for text comprehension. *arXiv preprint arXiv:1606.01549*.

Albert Gu, Caglar Gulcehre, Tom Le Paine, Matt Hoffman, and Razvan Pascanu. 2019. Improving the gating mechanism of recurrent neural networks. *arXiv preprint arXiv:1910.09890*.

Eric Jang, Shixiang Gu, and Ben Poole. 2016. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2019. Tinybert: Distilling bert for natural language understanding. *arXiv preprint arXiv:1909.10351*.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Gehui Shen, Yunlun Yang, and Zhi-Hong Deng. 2017. Inter-weighted alignment network for sentence pair modeling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1179–1189.

Zhiqing Sun, Hongkun Yu, Xiaodan Song, Renjie Liu, Yiming Yang, and Denny Zhou. Mobilebert: Task-agnostic compression of bert for resource limited devices.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Shuohang Wang and Jing Jiang. 2016. A compare-aggregate model for matching text sequences. *arXiv preprint arXiv:1611.01747*.

Lanqing Xue, Xiaopeng Li, and Nevin L Zhang. 2019. Not all attention is needed: Gated attention network for sequence data. *arXiv preprint arXiv:1912.00349*.

Seunghyun Yoon, Franck Dernoncourt, Doo Soon Kim, Trung Bui, and Kyomin Jung. 2019. A compare-aggregate model with latent clustering for answer selection. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 2093–2096.