CAM: A Combined Attention Model for Natural Language Inference

Amit Gajbhiye[†], Sardar Jaf[†], Noura Al Moubayed[†], Steven Bradley[†] and A. Stephen McGough[‡]
[†]*Department of Computer Science, Durham University, Durham, UK*
{*amit.gajbhiye, sardar.jaf, noura.al-moubayed, s.p.bradley*}*@durham.ac.uk*
[‡] *School of Computing, Newcastle University, Newcastle upon Tyne, UK*
stephen.mcgough@ncl.ac.uk

***Abstract*—Natural Language Inference (NLI) is a fundamental step towards natural language understanding. The task aims to detect whether a premise entails or contradicts a given hypothesis. NLI contributes to a wide range of natural language understanding applications such as question answering, text summarization and information extraction. Recently, the public availability of big datasets such as Stanford Natural Language Inference (SNLI) and SciTail, has made it feasible to train complex neural NLI models. Particularly, Bidirectional Long Short-Term Memory networks (BiLSTMs) with attention mechanisms have shown promising performance for NLI. In this paper, we propose a Combined Attention Model (CAM) for NLI. CAM combines the two attention mechanisms: intra-attention and inter-attention. The model first captures the semantics of the individual input premise and hypothesis with intra-attention and then aligns the premise and hypothesis with inter-sentence attention. We evaluate CAM on two benchmark datasets: Stanford Natural Language Inference (SNLI) and SciTail, achieving 86.14% accuracy on SNLI and 77.23% on SciTail. Further, to investigate the effectiveness of individual attention mechanism and in combination with each other, we present an analysis showing that the intra- and inter-attention mechanisms achieve higher accuracy when they are combined together than when they are independently used.**

***Keywords*-Natural Language Inference, Textual Entailment, Deep Learning, Attention Mechanism, SNLI dataset, SciTail dataset.**

I. Introduction

Natural Language Inference (NLI) is a fundamental step towards natural language understanding. NLI is the task of determining whether a sentence called hypothesis can be inferred from a given sentence called the premise. From the algorithmic perspective, NLI is a multi-class classification problem. The three classes are Entailment (inferred to be true), Contradiction (inferred to be false) and Neutral (truth value unknown).

Traditional approaches to NLI range from machine learning based, lexical and semantic similarity based, to the methods that extracts structured information such as discourse commitments [1]. However, traditional approaches require extensive feature engineering. Moreover, these approaches do not generalise well because of the complexity and domain dependence nature of the feature engineering task.

Machine learning has been a dominant approach to NLI. However, the machine learning research for NLI is severely limited in performance by the lack of gold-standard premise-hypothesis pairs [2]. The field has renewed prosperity by the recent introduction of big datasets such as Stanford Natural Language Inference (SNLI) [2] and SciTail [3]. The public availability of these big datasets has made it feasible to train complex neural network models for NLI. Recurrent Neural Networks (RNNs), particularly bidirectional LSTMs [4] in combination with attention mechanisms [5] have shown state-of-the-art results on the SNLI dataset [6].

Attention mechanisms allow RNNs to automatically search for the most relevant parts of an input sequence and assigns weights to those parts. These weights are used for creating the attention-weighted representation of the input sequence [5]. The two broad categories of attention in research literature are: intra-attention and inter-attention. The intra-attention mechanism, known as self-attention [7], involves applying attention to the input sentence itself. During training, the model learns to assign higher weight to those parts of the input sentence which are important to its semantics. The attention-weighted sentence representations thus generated also capture the global context of the sentence [8]. In inter-attention mechanism, attention is applied between the input sentences. The attention-weighted sentence representation of one sentence is generated based on the contents of another sentence. In the sentence representation, the information that is important with respect to other sentences is assigned higher weights.

Attention mechanism has helped in achieving state-of-the-art performance for NLI task [9]. However, the current models that employ only intra-attention [8] do not utilize information from another sentence. The models utilizing inter-attention [10], [11] do not exploit contexts in individual sentences. We propose a Combined Attention Model (CAM) which employs intra-attention in conjunction with inter-attention to utilize the benefits from both mechanisms.

Our experiments on the SNLI and SciTail dataset show that intra- and inter-attention mechanisms work constructively and achieve higher accuracy when they are combined together in the same model than using them independently. By combining the intra- and inter-attention mechanism we achieve an accuracy of 86.14% on SNLI and 77.23% on SciTail datasets. The model performs exceptionally well on SciTail outperforming the prominent ESIM model [6] and decomposable attention model [12] by 6.6% and 4.9% respectively.

II. Related Work

The intra-attention based model proposed by Liu et al. [8], applies attention to premise and hypothesis itself in order to identify the parts of sentences that are important to sentence semantics. Average pooling is first applied to the outputs of word-level BiLSTM and then intra-attention mechanism is employed to replace average pooling on the same sentence for better sentence representation. The authors applied various input strategies and achieved the maximum accuracy of 85.0%.

Rocktäschel et al. [10] first applied inter-attention to NLI models. The model is based on word-by-word attention and reasons entailment or contradiction over aligned word- and phrase-pairs. The eminence of inter-attention for NLI task is further established in the state-of-the-art models of Chen et al. [6], Tay et al. [9] and Parikh et al. [12]. The key idea of modeling inter-sentence attention is to soft-align the sub-phrases of premise and hypothesis. Tay et al. [9] and Parikh et al. [12] employs a standard projection layer with ReLU activation function whereas Chen et al. [6] utilize the similarity between the output hidden states of BiLSTMs of premise and hypothesis.

The closest work to our research is that by Parikh et al. [12]; they augmented inter-attention with intra-attention gaining 0.5% in accuracy by employing feed-forward neural-network at both the intra- and inter-attention layers. Our model fundamentally differs from the model proposed by Parikh et al. [12] both at the intra- and inter-attention layers. They have employed feed-forward neural-network at both the intra- and inter-attention layers. However, we used inner-attention mechanism [8] for intra-attention and dot attention mechanism [13] at the inter-attention layers.

Parikh et al. [12] have shown the effectiveness of using combined attention mechanisms, however, the possibility of using different attention mechanisms at intra- and inter-attention layers has not been explored to the best of our knowledge. We experimented with various combinations of attention mechanisms at intra- and inter-attention layer and found that not all combined attention mechanisms work constructively to achieve competitive accuracy for NLI task. We explored the possibility of employing inner-attention [8] and word-attention [14] at the intra-attention layer in combination with each of the dot, general and concate attention mechanisms [13] at inter-attention layer. We achieved the highest accuracy for the proposed combination of inner-attention and dot attention mechanisms. Furthermore, with each attention mechanism at intra- and inter-attention layers we experimented with the feed-forward neural network of [12], however that did not further improve the accuracy of our model.

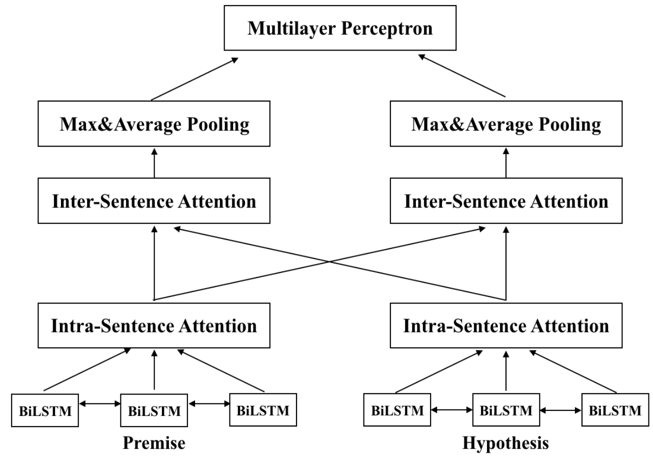

Figure 1: A high level layered architecture of Combined Attention Model (CAM).

III. Proposed Model

The proposed model combines intra-attention and inter-attention for modeling the interaction between premise-hypothesis pairs. Fig. 1 demonstrates the high-level view of the proposed model. The layered architecture is composed of the following layers: input encoding, intra-attention, inter-attention, composition and pooling.

In our notations, given the word sequence of premise **a** = $(a_1, \ldots, a_n)$ and hypothesis **b** = $(b_1, \ldots, b_m)$ with lengths n and m respectively. Each a_i, $b_j \in \mathbb{R}^r$, is a word embedding of r-dimensional, which can be initialized with pre-trained embeddings vectors, such as Glove [15].

Input Encoding Layer We utilize BiLSTMs to encode the input premise and hypothesis sentences. The BiLSTM processes the input sequence in forward and backward directions to incorporate contextual information at each time step of processing a word in the input sequence. The hidden state output at any time step is the concatenation of forward and backward hidden states. The $\bar{a} \in \mathbb{R}^{n \times 2d}$ and $\bar{b} \in \mathbb{R}^{m \times 2d}$ in Equation (1) and (2) respectively, represents the $2d$-dimensional representation for each word in the premise and hypothesis. Where d is the dimension of hidden states of LSTMs.

$$\bar{a}_i = \text{BiLSTM}(a, i) \forall i \in [1, \ldots, n] \quad (1)$$

$$\bar{b}_j = \text{BiLSTM}(b, j) \forall j \in [1, \ldots, m] \quad (2)$$

Intra-Attention Layer This layer applies intra-attention [8] to premise and hypothesis sentence independently. Through attention weights, the intra-attention layer emphasizes the words important to the semantics of the input sentence. The attention-weighted sentence representation thus generated represent a more accurate and focused sentence representations of the input sentence. The attention-weighted sentence representation is generated according to Equations $(3)-(5)$

$$M = \tanh\left(W^y Y + W^h R_{avg} \otimes e_L\right) \tag{3}$$

$$\alpha = softmax\left(w^T M\right) \tag{4}$$

$$r = Y\alpha^T \tag{5}$$

where W^y and W^h are trained projection matrices, Y is the matrix of hidden output vectors of the BiLSTM layer, R_{avg} is obtained from the average pooling of Y, $e_L \in \mathbb{R}^L$ is a vector of 1s, w^T is the transpose of trained parameter vector w, α is a vector of attention weights and r is the attention-weighted sentence representation. The attention-weighted sentence representation of the premise and the hypothesis is repeated for the maximum sentence length (L) and is denoted as r_p and r_h respectively.

Inter-Attention Layer The inter-attention layer uses soft alignment to associate relevant sub-components between the attention weighted representations of premise and hypothesis. The inter-attention layer, first, computes the unnormalized attention weights as a similarity of hidden states of intra-attention weighted representations of premise and hypothesis following Equation (6).

$$e_{ij} = r_{pi}^T r_{hj} \tag{6}$$

Next, for each word in the intra-attention weighted representation of the premise, the relevant semantics based on hypothesis, is extracted following Equation (7). Similarly, this is done for hypothesis according to Equation (8).

$$\tilde{r}_{pi} = \sum_{j=1}^{m} \frac{\exp(e_{ij})}{\sum_{k=1}^{m} \exp(e_{ik})} r_{hj} \tag{7}$$

$$\tilde{r}_{hj} = \sum_{i=1}^{n} \frac{\exp(e_{ij})}{\sum_{k=1}^{n} \exp(e_{kj})} r_{pi} \tag{8}$$

$\tilde{r}_p$ represents the content in r_p which are relevant based on r_h. Similarly, $\tilde{r_h}$ represents the content in r_h which are important with respect to r_p. We enrich the collected inference information through the element-wise multiplication of the tuples $(r_p, \tilde{r}_p)$ and $(r_h, \tilde{r}_p)$ as shown in Equations (9) and (10).

$$f_p = \tilde{r_p} \odot r_p \tag{9}$$

$$f_h = \tilde{r_h} \odot r_h \tag{10}$$

Pooling Layer To facilitate the classification of the relationship between premise and hypothesis, a relation vector is formed from the average and max pooling of the encoding of premise and hypothesis generated previously by inter-attention layer in Equations (9) and (10). Pooling is performed according to Equations (11) and (12).

$$\begin{aligned} \mathrm{v}_{p,avg} &= \text{average } \{f_p, i\}_{i=1}^n \\ \mathrm{v}_{p,max} &= \max \{f_p, i\}_{i=1}^n \end{aligned} \tag{11}$$

$$\begin{aligned} \mathrm{v}_{h,avg} &= \text{average}\{f_h, i\}_{i=1}^m \\ \mathrm{v}_{h,max} &= \max\{f_h, i\}_{i=1}^m \end{aligned} \tag{12}$$

where $v_{p,avg}$ and $v_{p,max}$ represents the fixed length vector for premise sentences resulting from the average and max pooling over $\{f_p, i\}_{i=1}^n$. Similarly, the fixed length representations is generated for hypothesis according to Equation (12).

Classification Layer To classify the relationship between premise and hypothesis, we feed the concatenation of vectors obtained from Equations (11) and (12) to a multilayer perceptron (MLP) classifier. Specifically, the classifier input is composed as in Equation (13).

$$F_{relation} = [\mathrm{v}_{p,avg}; \mathrm{v}_{p,max}; \mathrm{v}_{h,avg}; \mathrm{v}_{h,max}] \tag{13}$$

The MLP classifier consists of a hidden layer with $tanh$ activation and a three-way $softmax$ output layer. The network is then trained in an end-to-end manner with the standard multi-class cross entropy loss.

IV. Experiments and Results

Data The datasets, SNLI [2] and SciTail [3] used for evaluating our model are well balanced across NLI classes. We used the standard train/dev/test splits, as shown in Table I.

Table I: Experimental Datasets

Dataset	Train	Validation	Test
SNLI	549, 367	9, 842	9, 824
SciTail	23, 596	1, 304	2, 126

Hyperparameters We use pre-trained 300-*D* Glove 840*B* vectors to initialize the word embeddings [15]. The out-of-vocabulary (OOV) words are initialized by uniform distribution between [-0.05, 0.05]. The hidden states of all the layers for SciTail and SNLI datasets are set to 100 and 300 respectively. The Adam optimizer [16] with an initial learning rate of 0.001 is used. Dropout with the rate of 0.4 is applied only to the input of BiLSTM layer for SNLI and to each feed forward connection with dropout rate 0.3 for SciTail dataset [17]. We tuned the batch size amongst [32, 256, 512] and L2 regularization amongst [1e-4, 1e-5]. Each model is optimized on development set for the best performance.

Results on SNLI Table II shows the performance of different models on SNLI benchmark. The first row presents the lexical classifier by Bowman et al. [2]. Sentence encoding based models are shown in the second group (from row 2 to 6) of Table II. Bowman et al. [2] used LSTMs to generate sentence encoding of premise and hypothesis. The sentence encodings are then fed to a multilayer perceptron to identify the relationship between premise and hypothesis. Following

Table II: Accuracies of the models on SNLI.

Models	Accuracy (%)	
	Train	Test
Lexical Classifier [2]	99.7	78.2
100D LSTM [2]	84.8	77.6
300D LSTM [18]	83.9	80.6
600D BiLSTM (intra-attention) [8]	84.5	84.2
600D Gumbel TreeLSTM [19]	93.1	86.0
Distance-based Self-Attention Network [20]	89.6	86.3
100D LSTMs word-by-word attention [10]	85.3	83.5
100D Deep Fusion LSTM [21]	85.2	84.6
600D BiLSTM (diversing input) [8]	85.9	85.0
50D Stacked TC-LSTMs [11]	86.7	85.1
300D MMA-NSE (attention) [22]	86.9	85.4
300D LSTMN (deep attention fusion) [23]	87.3	85.7
200D Decomposable attention (intra-attention) [12]	90.5	86.8
600D ESIM + 300D TreeLSTM [6]	93.5	88.6
ESIM + ELMo [24]	91.6	88.7
300D CAM (Our Approach)	**90.5**	**86.1**

this strategy various sentence encoders are proposed, as shown in the second group of models in Table II.

The third group of models (from row 7 to 15) used inter-attention mechanism to align the sub-phrases between premise and hypothesis. Peters et al. [24] holds the current state-of-the-art performance on SNLI among the inter-attention, non-ensemble models. Embeddings from Language Models (ELMo) word embeddings of Peters et al. [24], when used with ESIM model of Chen et al. [6] improved the accuracy from 88.6% to 88.7%.

Among the models employing inter-sentence attention, our model (Combined Attention Model (CAM)) achieves a competitive accuracy of 86.14% on the SNLI dataset. Our model outperforms the previous models proposed by Rocktäschel et al. [10], Liu et al. [21], Liu et al. [8], Liu et al. [11], Munkhdalai and Yu [22] and Cheng et al. [23]. CAM achieves higher accuracy than the intra-attention with diversing input model of Liu et al. [8] by 1.4%.

Results on SciTail SciTail dataset contains the labelled data for the classes of NLI - neutral and entailment. The NLI, thus transforms into binary classification task. Table III shows our empirical results on the SciTail dataset. The low accuracies of the state-of-the-art ESIM [6] and decomposable attention model [12] suggest that SciTail is a difficult dataset to model. The performance gain of CAM over the strong ESIM and decomposable attention model is 6.6% and 4.9% in terms of accuracy.

V. Analysis and Discussion

Ablation Analysis We evaluate the effectiveness of individual components of our model on SciTail and SNLI datasets. Table IV depicts the results. For SciTail, both of our intra-attention-only and inter-attention-only models outperform the models of Parikh et al. [12] and Chen et al. [6] by a large margin, as detailed below.

Table III: Accuracies of the models on SciTail. The model accuracies are reported from [3] except for CAFE which is reported from [9]

Models	Test Accuracy(%)
Majority class	60.3
NGram	70.6
ESIM	70.6
DGEM w/o edges	70.8
Decomposable attention	72.3
DGEM	77.3
CAFE	83.3
CAM (our approach)	**77.2**

Table IV: Ablation analysis for SCI and SNLI datasets

Models	Test Accuracy(%)	
	SciTail	SNLI
Combined Attention	**77.23**	**86.14**
Intra-attention-only	75.49	80.27
Inter-attention-only	76.06	85.04

When we remove inter-attention mechanism from CAM, the intra-attention-only model has an accuracy of 75.49% and outperforms the decomposable attention model of Parikh et al. [12] and ESIM model of Chen et al. [6] (please refer Table III for the model accuracy of Parikh et al. [12] and Chen et al. [6]) by 3.1% and 4.9% respectively.

When we remove the intra-attention mechanism from CAM, the inter-attention-only model achieves an accuracy of 76.06%. The inter-attention-only model improves over the accuracy of decomposable attention of Parikh et al. [12] by 3.76% and by 5.46% over the ESIM model of Chen et al. [6]. CAM performs comparatively with DGEM model of Khot et al. [3].

For SNLI, the intra-attention-only model does not perform well and it achieves an accuracy of 80.27%. However, the inter-attention-only model achieves an accuracy of 85.04%, which is higher than the word-by-word attention model of Rocktäschel et al. [10] by 1.5% and deep fusion LSTM model of Liu et al. [21] by 0.4%. The inter-attention-only model performs competitively with the intra-attention with diversing input model of Liu et al. [8]

It is worth to note that the SciTail dataset contains longer premises and hypotheses than the SNLI dataset [3]. The results of the ablation analysis for SciTail suggest that for long sentences, it is crucial to first capture the semantics of the input sentence by intra-attention mechanism. The results on both of the datasets suggest that intra-attention and inter-attention work constructively and achieve high accuracy when they are combined.

Further Analysis To investigate the effectiveness of each attention mechanism individually and in combination with each other, we further analyse the performance of each model in Table IV. Fig. 2 present the result of the analysis.

For SNLI: The three models correctly classified 74% the test samples (central region (e) Fig. 2(a)). Combined attention model outperforms each of the individual attention mechanism by correctly classifying 2.2% of test cases individually (region(c) in Fig. 2(a)) as compared to 1.8% of intra-attention only and 2.1% of inter-attention only model. The inter-attention model and combined attention model correctly classify 7.0% of test samples whereas intra-attention and combined attention correctly classify 3.0% of test samples. This suggests that inter attention is crucial for the high performance on SNLI. The intra-attention and inter-attention correctly classifies 2.0% of test samples. There are 7.9% test samples which cannot be classified correctly by any of the three models.

For SciTail: The three models correctly classified 64% of test cases (central region, Fig. 2(b)). Similar to SNLI, the combined attention model gets the highest percent (3.3%) of test samples classified correctly. Unlike for SNLI, the intra-attention-only and combined attention models agree on a larger number of test cases (5.1%) than the inter-attention-only and combined attention model, which agree on 4.6% of test cases. Given the fact that SciTail is difficult to model [9], the result suggest that capturing the semantics of individual sequence first with intra-sentence attention is crucial for modeling complex datasets. Moreover, a significant number of test samples (13.4%) are not classified correctly by any of the model. This further indicates the high complexity SciTail.

Linguistic analysis of the test samples in each region of Fig. 2 is an interesting investigation to understand the behaviour of each model. Particularly, it is interesting to analyze syntax and semantics of the premise-hypothesis pairs, which are incorrectly classified by the intra-attention-only and inter-attention-only models but correctly classified by combined attention model. Region (c) in Fig. 2 depicts these test cases. A preliminary linguistic observation on the syntactic structure of premise-hypothesis pairs in this region suggest that for longer premises (word count $>$ 20) the combined attention model predicts the test classes correctly more often than the intra-attention-only and inter-attention-only models.

VI. Conclusion and Future Work

In this paper, we proposed a natural language inference model called Combined Attention Model (CAM), that benefits from intra-attention and inter-attention mechanisms. Experiments on two benchmark datasets: SNLI and SciTail demonstrate that CAM performs competitively to the previous models. CAM achieves an accuracy of 86.14% on SNLI and 77.23% on SciTail. We show that, CAM performs particularly effectively on the hard to model SciTail dataset and outperforms the state-of-the-art ESIM by 6.6% and decomposable attention models by 4.9%. Further, the results of ablation analysis shows that the intra-attention and inter-attention mechanism work constructively and achieve higher accuracy when they are combined together in the same model than when they are independently used. In future work, we will investigate the effectiveness of incorporating syntactic information such as part-of-speech tags and parse trees into the input sentences. We believe those linguistic features would further benefit the model to capture some semantic aspects of the sentences.

References

[1] B. MacCartney, T. Grenager, M.-C. de Marneffe, D. Cer, and C. D. Manning, "Learning to recognize features of valid textual entailments," in *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, ser. HLT-NAACL '06, Stroudsburg, PA, USA, 2006, pp. 41–48.

[2] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, "A large annotated corpus for learning natural language inference," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2015, pp. 632–642.

[3] T. Khot, A. Sabharwal, and P. Clark, "Scitail: A textual entailment dataset from science question answering," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA, February 2-7, 2018*, 2018.

[4] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[5] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *CoRR*, vol. abs/1409.0473, 2014.

[6] Q. Chen, X. Zhu, Z.-H. Ling, S. Wei, H. Jiang, and D. Inkpen, "Enhanced lstm for natural language inference," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, vol. 1, 2017, pp. 1657–1668.

[7] Z. Lin, M. Feng, C. N. dos Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio, "A structured self-attentive sentence embedding," in *International Conference on Learning Representations 2017 (Conference Track)*, 2017.

[8] Y. Liu, C. Sun, L. Lin, and X. Wang, "Learning natural language inference using bidirectional lstm model and inner-attention," *arXiv preprint arXiv:1605.09090*, 2016.

[9] Y. Tay, L. A. Tuan, and S. C. Hui, "A compare-propagate architecture with alignment factorization for natural language inference," *arXiv preprint arXiv:1801.00102*, 2017.

[10] T. Rocktäschel, E. Grefenstette, K. M. Hermann, T. Kocisky, and P. Blunsom, "Reasoning about entailment with neural attention," in *International Conference on Learning Representations (ICLR)*, 2016.

(a) SNLI (b) SciTail

Figure 2: Venn diagram showing the percent of test samples correctly classified by each model in Table IV. The central overlapped region depicts the percent of correctly classified test samples by all the three models. The label adjoining each attention model shows the percent of test cases incorrectly classified by the individual model. The label at the left bottom shows the percent of test samples incorrectly classified by all the models. For instance, for SNLI (Fig. (a)) the three models classified 74.0% of test cases correctly. The combined attention model individually misclassified 5.9% of test cases and all the three models misclassified 7.9% of test cases.

[11] P. Liu, X. Qiu, Y. Zhou, J. Chen, and X. Huang, "Modelling interaction of sentence pair with coupled-lstms," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 1703–1712.

[12] A. Parikh, O. Täckström, D. Das, and J. Uszkoreit, "A decomposable attention model for natural language inference," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2016, pp. 2249–2255.

[13] T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2015, pp. 1412–1421.

[14] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 1480–1489.

[15] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[16] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014. [Online]. Available: http://arxiv.org/abs/1412.6980

[17] A. Gajbhiye, S. Jaf, N. Al-Moubayed, A. S. McGough, and S. Bradley, "An exploration of dropout with rnns for natural language inference." in *27th international Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, proceedings. Part III.*, ser. Lecture notes in computer science. Cham: Springer, October 2018, no. 11141, pp. 157–167. [Online]. Available: http://dro.dur.ac.uk/25791/

[18] S. R. Bowman, J. Gauthier, A. Rastogi, R. Gupta, C. D. Manning, and C. Potts, "A fast unified model for parsing and sentence understanding," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, vol. 1, 2016, pp. 1466–1477.

[19] J. Choi, K. M. Yoo, and S. goo Lee, "Learning to compose task-specific tree structures." AAAI, 2017.

[20] J. Im and S. Cho, "Distance-based self-attention network for natural language inference," *arXiv preprint arXiv:1712.02047*, 2017.

[21] P. Liu, X. Qiu, J. Chen, and X. Huang, "Deep fusion lstms for text semantic matching," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, vol. 1, 2016, pp. 1034–1043.

[22] T. Munkhdalai and H. Yu, "Neural tree indexers for text understanding," in *Proceedings of the conference. Association for Computational Linguistics. Meeting*, vol. 1. NIH Public Access, 2017, p. 11.

[23] J. Cheng, L. Dong, and M. Lapata, "Long short-term memory-networks for machine reading," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 551–561.

[24] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *Proc. of NAACL*, 2018.