PAIR-LEVEL SUPERVISED CONTRASTIVE LEARNING FOR NATURAL LANGUAGE INFERENCE

Shu'ang Li[1], *Xuming Hu*[1], *Li Lin*[1], *Lijie Wen*[2*]

School of Software, Tsinghua University
[1]{lisa18, hxm19, lin-l16}@mails.tsinghua.edu.cn, [2]wenlj@tsinghua.edu.cn

ABSTRACT

Natural language inference (NLI) is an increasingly important task for natural language understanding, which requires one to infer the relationship between the sentence pair (**premise** and **hypothesis**). Many recent works have used contrastive learning by incorporating the relationship of the sentence pair from NLI datasets to learn sentence representation. However, these methods only focus on comparisons with sentence-level representations. In this paper, we propose a **Pair**-level **S**upervised **C**ontrastive **L**earning approach (PairSCL). We adopt a cross attention module to learn the joint representations of the sentence pairs. A contrastive learning objective is designed to distinguish the varied classes of sentence pairs by pulling those in one class together and pushing apart the pairs in other classes. We evaluate PairSCL on two public datasets of NLI where the accuracy of PairSCL outperforms other methods by 2.1% on average. Furthermore, our method outperforms the previous state-of-the-art method on seven transfer tasks of text classification.[1]

Index Terms— supervised contrastive learning, natural language inference, pair-level representation

1. INTRODUCTION

Natural Language Inference (NLI) is a fundamental problem in the research field of natural language understanding [1, 2], which could help tasks like questions answering, reading comprehension, summarization and relation extraction [3, 4, 5, 6]. In NLI settings, the model is presented with a pair of sentences, namely **premise** and **hypothesis** and is asked to reason the relationship between them from a set of relationships, including `entailment`, `contradiction` and `neutral`. In the last several years, large annotated datasets were made available, e.g., the SNLI [7] and MultiNLI datasets [8], which made it feasible to train rather complicated neural network-based models [9, 10].

However, these methods only use the feature of the sentence pair itself to predict the class, without considering the comparison between the sentence pairs in different classes. Many recent works explored using contrastive learning to tackle this problem. Contrastive learning is a popular technique in computer vision area [11, 12, 13] and the core idea is to learn a function that maps positive pairs closer together in the embedding space, while pushing apart negative pairs. A contrastive objective is used by [14] to fine-tune pre-trained language models to obtain sentence embeddings with the relationship of sentences in NLI, and achieved state-of-the-art performance in sentence similarity tasks. However, this approach can't distinguish well between the representation of sentence pairs in different classes. For example, two sentence pairs are in the same class of `entailment` from NLI dataset (P_1: Two men on bicycles competing in a race. H_1: People are riding bikes. P_2: Two dogs are running. H_2: There are animals outdoors). They simply consider H_1 as the positive set and H_2 as the negative set for P_1 without taking into account that these two pairs are in the same class.

Given this scenario, we propose a pair-level supervised contrastive learning approach. The pair-level representation is obtained by cross attention module which can capture the relevance and well characterize the relationship between the sentence pair. Therefore, the pair-level representation can perceive the class information of sentence pairs. Then we use the pair-level representations for contrastive learning by capturing the similarity between pairs in one class and contrasting them with pairs in other classes. The model is trained with a combined objective of a supervised contrastive learning term and a cross-entropy term. We evaluate PairSCL on two public datasets of NLI where the accuracy of PairSCL outperforms other methods by 2.1% on average. Furthermore, our method outperforms the previous state-of-the-art method on seven transfer tasks of text classification.

2. APPROACH

In this section, we describe our approach PairSCL. Figure 1 shows a high-level general view of PairSCL. PairSCL comprises the following three major components: an encoder that computes sentence representations for input text, a cross attention module to capture the relationship between the sentence pair and a joint-training layer including a cross-entropy term and supervised contrastive learning term.

[*] Corresbonding Author

[1] Our code will be available at https://github.com/THU-BPM/PairSCL.

978-1-6654-0540-9/22/$31.00 ©2022 IEEE

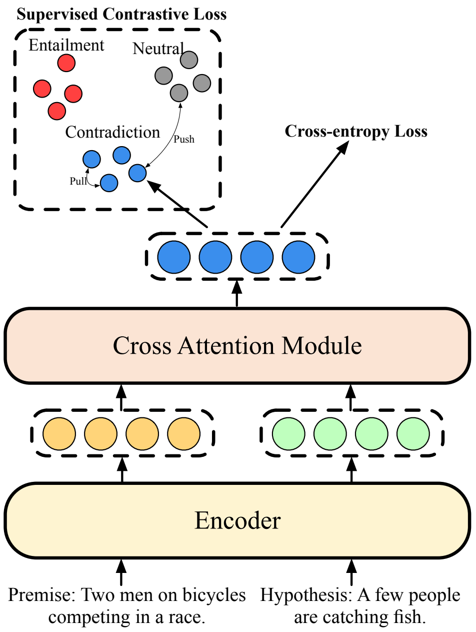
Supervised Contrastive Loss
Entailment
Neutral
Contradiction
Push
Pull
Cross-entropy Loss
Cross Attention Module
Encoder
Premise: Two men on bicycles competing in a race.
Hypothesis: A few people are catching fish.

Fig. 1. The framework of PairSCL.

2.1. Text Encoder

Each instance in a NLI dataset consists of two sentences and a label indicating the relation between them. Formally, we denote **premise** as $X^{(p)} = \{x_1^{(p)}, x_2^{(p)}, \cdots, x_m^{(p)}\}$ and **hypothesis** as $X^{(h)} = \{x_1^{(h)}, x_2^{(h)}, \cdots, x_n^{(h)}\}$, where m and n are length of the sentences respectively. The instance in the batch $\mathcal{I}$ is denoted as $(X^{(p)}, X^{(h)}, y)_{i \in \mathcal{I}}$, where $i = \{1, \ldots, K\}$ is the indices of the samples and K is the batch-size. The encoder (e.g., BERT, RoBERTa) takes $X^{(p)}, X^{(h)}$ as inputs and computes the semantic representations, denoted as $\mathbf{S}^{(p)} = \{\mathbf{s}_i^{(p)} | \mathbf{s}_i^{(p)} \in \mathbb{R}^k, i = 1, 2, \cdots, m\}$ and $\mathbf{S}^{(h)} = \{\mathbf{s}_j^{(h)} | \mathbf{s}_j^{(h)} \in \mathbb{R}^k, j = 1, 2, \cdots, n\}$, where k is the dimension of the encoder's hidden state.

2.2. Cross Attention Module

Different from single sentence classification, we need a proper interaction module to better clarify the sentences pair's relationship for NLI task. In practice, we need to compute token-level weights between words in **premise** and **hypothesis**. Therefore, we introduce the cross attention module to calculate the co-attention matrix $\mathbf{C} \in \mathbb{R}^{m \times n}$ of the token level. Each element $\mathbf{C}_{i,j} \in \mathbb{R}$ indicates the relevance between the i-th word of **premise** and the j-th word of **hypothesis**:

$$\mathbf{C}_{i,j} = \mathbf{P}^T tanh(\mathbf{W}(\mathbf{s}_i^{(p)} \odot \mathbf{s}_j^{(h)})), \tag{1}$$

where $\mathbf{W} \in \mathbb{R}^{d \times k}$, $\mathbf{P} \in \mathbb{R}^d$, and $\odot$ denotes the element-wise production operation. Then the attentive matrix could be formalized as:

$$\mathbf{c}_i^{(p)} = softmax(\mathbf{C}_{i,:}), \quad \mathbf{c}_j^{(h)} = softmax(\mathbf{C}_{:,j}), \tag{2}$$

$$\mathbf{s}_i^{(p)'} = \mathbf{S}^{(h)} \cdot \mathbf{c}_i^{(p)}, \quad \mathbf{s}_i^{(h)'} = \mathbf{S}^{(p)} \cdot \mathbf{c}_j^{(h)}, \tag{3}$$

We further enhance the collected local semantic information:

$$\mathbf{s}_i^{(p)''} = [\mathbf{s}_i^{(p)}; \mathbf{s}_i^{(p)'}; \mathbf{s}_i^{(p)} - \mathbf{s}_i^{(p)'}; \mathbf{s}_i^{(p)} \odot \mathbf{s}_i^{(p)'}], \tag{4}$$

$$\tilde{\mathbf{s}}_i^{(p)} = ReLU(\mathbf{W}_i^{(p)} \mathbf{s}_i^{(p)''} + \mathbf{b}_i^{(p)}), \tag{5}$$

where $[\cdot; \cdot; \cdot; \cdot]$ refers to the concatenation operation. $\mathbf{s}_i^{(p)} - \mathbf{s}_i^{(p)'}$ indicates the difference between the original representation and the **hypothesis**-information enhanced representation of **premise**, and $\mathbf{s}_i^{(p)} \odot \mathbf{s}_i^{(p)'}$ represents their semantic similarity. Both values are designed to measure the degree of semantic relevance between the sentences pair. The smaller the difference and the larger the semantic similarity, the sentences pair are more likely to be classified into Entailment category. The difference and element-wise product are then concatenated with the original vectors $(\mathbf{S}^{(p)}, \mathbf{S}^{(p)'})$. We expect that such operations could help enhance the pair-level information and capture the inference relationships of **premise** and **hypothesis**. We get the new representation containing **hypothesis**-guided inferential information for **premise**:

$$\tilde{\mathbf{S}}^{(p)} = (\tilde{\mathbf{s}}_1^{(p)}, \tilde{\mathbf{s}}_2^{(p)}, \ldots, \tilde{\mathbf{s}}_m^{(p)}), \tag{6}$$

$$\hat{\mathbf{S}}^{(p)} = LayerNorm(\tilde{\mathbf{S}}^{(p)}), \tag{7}$$

where $LayerNorm(.)$ is a layer normalization. The result $\hat{\mathbf{S}}^{(p)}$ is a 2D-tensor that has the same shape as $\mathbf{S}^{(p)}$. The representation of **hypothesis** $\hat{\mathbf{S}}^{(h)}$ is calculated in the same way. We aggregate these representations and the pair-level representation $\mathbf{Z}$ for the sentence pair is obtained as follows:

$$\mathbf{Z} = [\hat{\mathbf{S}}^{(p)}; \hat{\mathbf{S}}^{(h)}; \hat{\mathbf{S}}^{(p)} - \hat{\mathbf{S}}^{(h)}; \hat{\mathbf{S}}^{(p)} \odot \hat{\mathbf{S}}^{(h)}], \tag{8}$$

As described, the cross attention module can capture the relevance of the sentence pair and well characterize the relationship. Therefore, the pair-level representation can perceive the class information of sentence pairs.

2.3. Training Objective

Supervised contrastive loss A contrastive loss brings the latent representations of samples belonging to the same class closer together, by defining a set of positives (that should be closer) and negatives (that should be further apart). In [13], the authors extended the above loss to a supervised contrastive loss by regarding the samples belonging to the same class as positive set. Inspired by this, we adopt supervised contrastive learning objective to align the pair-level representation obtained from cross attention module to distinguish sentence pairs from different classes.

In the training stage, we randomly sample a batch $\mathcal{I}$ of K examples $(X^{(p)}, X^{(h)}, y)_{i \in \mathcal{I}=\{1,\ldots,K\}}$ as denoted in Section 2.1. We denote the set of positives as $\mathcal{P} = \{p : p \in \mathcal{I}, y_p = y_i \wedge p \neq i\}$, with size $|\mathcal{P}|$. The supervised contrastive loss on the batch $\mathcal{I}$ is defined as:

$$\ell_{i,p} = \frac{\exp(\mathbf{Z}_i \cdot \mathbf{Z}_p/\tau)}{\sum_{k \in \mathcal{I}/i} \exp(\mathbf{Z}_i \cdot \mathbf{Z}_k/\tau)}, \tag{9}$$

$$\mathcal{L}_{SCL} = \sum_{i \in \mathcal{I}} -\log \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \ell_{i,p}, \tag{10}$$

where $\ell_{i,p}$ indicates the likelihood that pair i is most similar to pair p and τ is the temperature hyper-parameter. Larger values of τ scale down the dot-products, creating more difficult comparisons. $\mathbf{Z}_i$ is the pair-level representation of pair $(X^{(p)}, X^{(h)})_i$ from the cross attention module. Supervised contrastive loss $\mathcal{L}_{SCL}$ is calculated for every sentence pair among the batch $\mathcal{I}$. To minimize contrastive loss $\mathcal{L}_{SCL}$, the similarity of pairs in the same class should be as large as possible, and the similarity of negative examples should be as small as possible. In this way, we can map positive pairs closer together in the embedding space, while pushing apart negative pairs.

Cross-entropy loss Supervised contrastive loss mainly focuses on separating each pair apart from the others of different classes, whereas there is no explicit force in discriminating contradiction, neutral and entailment. Therefore, we adopt the softmax-based cross-entropy to form the classification objective:

$$\mathcal{L}_{CE} = CrossEntropy(\mathbf{WZ} + \mathbf{b}, y), \tag{11}$$

where $\mathbf{W}$ and $\mathbf{b}$ are trainable parameters. $\mathbf{Z}$ is the pair-level representation from the cross attention module and y is the corresponding label of the pair.

Overall loss The overall loss is a weighted average of CE and the SCL loss, denoted as:

$$\mathcal{L} = \mathcal{L}_{SCL} + \alpha \mathcal{L}_{CE}, \tag{12}$$

where α is a hyper-parameter to balance two objectives.

3. EXPERIMENTAL SETUP

3.1. Benchmark Dateset

We conduct our experiments on NLI task and other 7 transfer learning tasks.

Natural language inference task: We evaluate on two popular benchmarks: the Stanford Natural Language Inference (SNLI) [7] and the MultiGenre NLI Corpus (MultiNLI) [8] and compute classification accuracy as the evaluation metric. Detailed statistical information is shown in Table 1.

Dataset	Train	Dev	Test	Len(P)	Len(H)
SNLI	549K	9.8K	9.8K	14	8
MultiNLI(m)	392K	9.8K	9.8K	22	11
MultiNLI(mm)		9.8K	9.8K	22	11

Table 1. Statistics of datasets: SNLI, MultiNLI. Len(P) and Len(H) refer to the average length of two sentences respectively. MultiNLI(m) and MultiNLI(mm) indicate the matched and mismatched datasets respectively.

Transfer tasks: We also evaluate on the following transfer tasks: MR [15], CR [16], SUBJ [17], MPQA [18], SST-2 [19], TREC [20] and MRPC [21]. For single-sentence classification tasks, we train a logistic regression classifier on top of frozen BERT encoder representation **S**. In MRPC task, we use the pair-level representation **Z** obtained from the cross attention module for the sentence pair to map the semantic space. We follow default configurations from SentEval [22].

3.2. Implementation Details

We start from pre-trained checkpoints of BERT [23] (uncased) or RoBERTa [24] (cased). We implement PairSCL based on Huggingface's `transformers` package [25]. All experiments are conducted on 5 Nvidia GTX 3090 GPUs.

We train our models for 10 epochs with a batch size of 512 and temperature τ = 0.05 using an Adam optimizer [26]. The hyper-parameter α is set as 1 for combining objectives. The learning rate is set as 5e-5 for base models and 1e-5 for large models. The maximum sequence length is set to 128.

3.3. Baseline Models

To analyze the effectiveness of PairSCL on NLI, we select ESIM [9], KIM [27], ADIN [28], BERT [23] and RoBERTa [24] as baselines. They are all trained with NLI supervision.

For transfer tasks, we evaluate with SBERT, SRoBERTa [29] and SimCSE [14]. We directly report the results from [29], since our evaluation setting is the same with theirs.

Model	SNLI	MultiNLI(m)	MultiNLI(mm)
ESIM	88.0	72.3	72.1
KIM	88.6	77.2	76.4
ADIN	88.8	78.8	77.9
BERT	89.8	83.3	82.7
PairSCL-BERT$_{\mathrm{base}}$	**91.9**	**85.5**	**84.6**
RoBERTa$_{\mathrm{base}}$	91.2	90.8	90.2
PairSCL-RoBERTa$_{\mathrm{base}}$	**93.2**	**92.7**	**92.3**

Table 2. Performance on the test dataset. MultiNLI(m) and MultiNLI(mm) represent the accuracy on matched and mismatched datasets. The best performance is in **bold** among models with the same pre-trained encoder.

Model	MR	CR	SUBJ	MPQA	SST	TREC	MRPC	Avg.
SBERT$_{\texttt{base}}$♣	83.64	89.43	94.39	89.86	88.96	89.60	76.00	87.41
SimCSE-BERT$_{\texttt{base}}$♡	82.69	89.25	94.81	89.59	87.31	88.40	73.51	86.51
PairSCL-BERT$_{\texttt{base}}$	**83.80**	**89.69**	**94.94**	**89.92**	**89.23**	**89.87**	**76.67**	**87.73**
SRoBERTa$_{\texttt{base}}$♡	84.91	90.83	92.56	88.75	90.50	88.60	78.14	87.76
SimCSE-RoBERTa$_{\texttt{base}}$♡	84.92	92.00	94.11	89.82	91.27	88.80	75.65	88.08
PairSCL-RoBERTa$_{\texttt{base}}$	**85.38**	**92.67**	**95.12**	**90.56**	**92.14**	**89.06**	**79.65**	**89.23**

Table 3. Transfer task results of different snetence embedding models (measured as accuracy). ♣: results from [29]; ♡: results from [14]. The best performance is in **bold** among models with the same pre-trained encoder.

4. EXPERIMENT RESULTS

4.1. NLI Results

Table 2 shows the average results of different models on NLI task. Our approach outperforms state-of-the-art models by 2.1% on SNLI with BERT encoders. On MultiNLI dataset, we compare on two test sets (matched and mismatched). PairSCL-BERT$_{\texttt{base}}$ achieves 85.5% and 84.6% respectively. For RoBERTa encoders, PairSCL achieves outstanding performance, 93.2% on SNLI and 92.7%/92.3% on MultiNLI. For the results on two datasets, we conduct the students paired t-test and the p-value of the significance test between the results of PairSCL and RoBERTa is less than 0.01 and 0.05, respectively.

This performance gains are due to the stronger ability of PariSCL to learn pair-level representation with cross attention. PairSCL can capture pair-level semantics effectively by the specifically-designed contrastive signal – predicting whether two sentence pairs belong to the same class.

4.2. Transfer Tasks Results

Table 3 shows the evaluation results on transfer tasks. We can observe that PairSCL-BERT$_{\texttt{base}}$ outperforms several supervised baselines like InferSent and Universal Sentence Encoder, and keeps comparable to the strong supervised method SBERT$_{\texttt{base}}$. When further performing representation transfer with RoBERTa base architecture, our approach achieves even better performance. On average, our approach outperforms SimCSE-RoBERTa$_{\texttt{base}}$ with an improvement of 1.15% (from 88.08% to 89.23%).

As we argued earlier, it benefits from that our model can distinguish the sentences of different classes well by pulling the sentence from the same class together and pushing them of different classes further apart.

4.3. Ablation Study

To better understand the contribution of each key component of PairSCL, we conduct an ablation study on SNLI based on BERT encoders. The results are shown in Table 4.

After removing the cross attention mechanism, the model simply concat the representation of two sentences. The performance decreases by 1.6% on the test set which shows the joint representation obtained by cross attention can well characterize the relationship between the sentence pair. Remove the cross-entropy loss and the test accuracy decreases by 0.7%. Without the supervised contrastive learning loss, the accuracy of our model is decreased to 90.7%. The reason is that the contrastive learning objective can learn the discrepancy between the sentence pairs of different classes by pulling the sentence pairs from the same class together and pushing the pairs of different classes further apart.

Model	Accuracy
PairSCL (-CE loss)	91.2
PairSCL (-SCL loss)	90.7
PairSCL (-Cross attention)	90.3
PairSCL	**91.9**

Table 4. Ablation study on SNLI with BERT encoders.

5. CONCLUSION

In this paper, we propose a pair-level supervised contrastive learning approach. We adopt a cross attention module to learn the joint representations of the sentence pairs. A contrastive learning objective is designed to distinguish the varied classes of sentence pairs by pulling those in one class together and pushing apart the pairs in other classes. We evaluate PairSCL on two popular datasets: SNLI and MultiNLI. The experiment results show that PairSCL obtains new state-of-the-art performance compared with existing models. For the transfer tasks, PariSCL outperforms the previous state-of-the-art method with 1.2% averaged improvement. We carefully study the components of PairSCL, and show the effects of different parts.

6. ACKNOWLEDGMENTS

The work was supported by the National Key Research and Development Program of China (No. 2019YFB1704003), the National Nature Science Foundation of China (No. 62021002 and No. 71690231), Tsinghua BNRist and Beijing Key Laboratory of Industrial Big Data System and Application.

7. REFERENCES

[1] Bill MacCartney and Christopher D. Manning, "Modeling semantic containment and exclusion in natural language inference," in *Proc. of COLING*, Aug. 2008, pp. 521–528.

[2] Leixian Shen, Enya Shen, Yuyu Luo, Xiaocong Yang, Xuming Hu, Xiongshuai Zhang, Zhiwei Tai, and Jianmin Wang, "Towards natural language interfaces for data visualization: A survey," *arXiv preprint arXiv:2109.03506*, 2021.

[3] Ido Dagan, Dan Roth, Mark Sammons, and Fabio Massimo Zanzotto, "Recognizing textual entailment: Models and applications," *Synthesis Lectures on Human Language Technologies*, vol. 6, no. 4, pp. 1–220, 2013.

[4] Xuming Hu, Lijie Wen, Yusong Xu, Chenwei Zhang, and S Yu Philip, "Selfore: Self-supervised relational feature learning for open relation extraction," in *Proc. of EMNLP*, 2020, pp. 3673–3682.

[5] Xuming Hu, Chenwei Zhang, Yawen Yang, Xiaohe Li, Li Lin, Lijie Wen, and S Yu Philip, "Gradient imitation reinforcement learning for low resource relation extraction," in *Proc. of EMNLP*, 2021, pp. 2737–2746.

[6] Xuming Hu, Chenwei Zhang, Fukun Ma, Chenyao Liu, Lijie Wen, and Philip S. Yu, "Semi-supervised relation extraction via incremental meta self-training," in *Proc. of EMNLP: Findings*. 2021, pp. 487–496, Association for Computational Linguistics.

[7] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning, "A large annotated corpus for learning natural language inference," in *Proc. of EMNLP*, Sept. 2015, pp. 632–642.

[8] Adina Williams, Nikita Nangia, and Samuel Bowman, "A broad-coverage challenge corpus for sentence understanding through inference," in *Proc. of NAACL*, 2018, pp. 1112–1122.

[9] Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen, "Enhanced LSTM for natural language inference," in *Proc. of ACL*, July 2017, pp. 1657–1668.

[10] Yichen Gong, Heng Luo, and Jian Zhang, "Natural language inference over interaction space," in *ICLR*, 2018.

[11] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. of CVPR*, 2020, pp. 9729–9738.

[12] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. of ICML*. PMLR, 2020, pp. 1597–1607.

[13] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan, "Supervised contrastive learning," *NeurIPS*, vol. 33, 2020.

[14] Tianyu Gao, Xingcheng Yao, and Danqi Chen, "Simcse: Simple contrastive learning of sentence embeddings," *Proc. of EMNLP*, 2021.

[15] Bo Pang, "Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales," in *Proc. of ACL*, 2005, pp. 115–124.

[16] Minqing Hu and Bing Liu, "Mining and summarizing customer reviews," in *Proc. of SIGKDD*, 2004, pp. 168–177.

[17] Bo Pang and Lillian Lee, "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts," *Proc. of ACL*, 2004.

[18] Janyce Wiebe, Theresa Wilson, and Claire Cardie, "Annotating expressions of opinions and emotions in language," *Language resources and evaluation*, vol. 39, no. 2, pp. 165–210, 2005.

[19] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proc. of EMNLP*, 2013, pp. 1631–1642.

[20] Ellen M Voorhees and Dawn M Tice, "Building a question answering test collection," in *Proc. of SIGIR*, 2000, pp. 200–207.

[21] William B Dolan and Chris Brockett, "Automatically constructing a corpus of sentential paraphrases," in *Proc. of IWP*, 2005.

[22] Alexis Conneau and Douwe Kiela, "Senteval: An evaluation toolkit for universal sentence representations," in *Proc. of LREC*, 2018.

[23] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proc. of NAACL*, 2019, pp. 4171–4186.

[24] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[25] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al., "Huggingface's transformers: State-of-the-art natural language processing," *arXiv e-prints*, pp. arXiv–1910, 2019.

[26] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *ICLR*, 2015.

[27] Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Diana Inkpen, and Si Wei, "Neural natural language inference models enhanced with external knowledge," in *Proc. of ACL*, 2018, pp. 2406–2417.

[28] Di Liang, Fubao Zhang, Qi Zhang, and Xuan-Jing Huang, "Asynchronous deep interaction network for natural language inference," in *Proc. of EMNLP*, 2019, pp. 2692–2700.

[29] Nils Reimers and Iryna Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *Proc. of EMNLP-IJCNLP*, 2019.