Dual-View Distilled BERT for Sentence Embedding

Xingyi Cheng
fanyin.cxy@alibaba-inc.com
Ant Group
Hanzhou, Zhejiang, China

ABSTRACT

Recently, BERT realized significant progress for sentence matching via word-level cross sentence attention. However, the performance significantly drops when using siamese BERT-networks to derive two sentence embeddings, which fall short in capturing the global semantic since the word-level attention between two sentences is absent. In this paper, we propose a Dual-view distilled BERT (DvBERT) for sentence matching with sentence embeddings. Our method deals with a sentence pair from two distinct views, i.e., Siamese View and Interaction View. Siamese View is the backbone where we generate sentence embeddings. Interaction View integrates the cross sentence interaction as multiple teachers to boost the representation ability of sentence embeddings. Experiments on six STS tasks show that our method outperforms the state-of-the-art sentence embedding methods.

CCS CONCEPTS

• **Information systems → Similarity measures**.

KEYWORDS

Semantic Matching, BERT, Natural Language Inference, Dense Retrieval

ACM Reference Format:

Xingyi Cheng. 2021. Dual-View Distilled BERT for Sentence Embedding. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21), July 11–15, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3404835.3463057

1 INTRODUCTION

Recent sentence representation models like BERT [12] achieved state-of-the-art results on sentence-pair regression/classification tasks, such as question answering, natural language inference (NLI) [5, 23], and semantic textual similarity (STS) [1–4]. However, it has a low computational efficiency when candidate sentence-pairs are not given ahead, leading to a massive computational overhead. For example, seeking the most relevant sentence-pair of a collection requires pairing all sentences. The $O(n^2)$ computational complexity is an obstacle preventing many retrieval applications from adopting the technology.

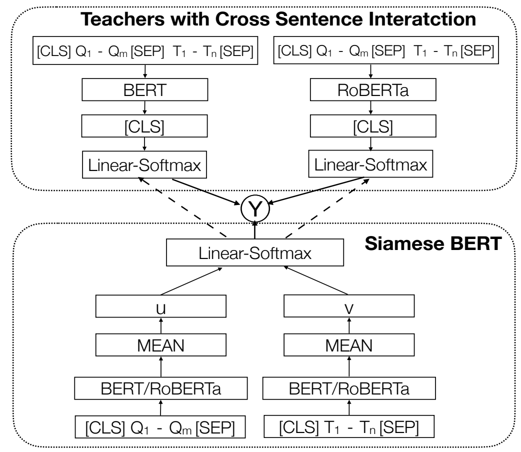

Figure 1: Overview of Dual View Distilled BERT. Dash lines indicate distillation.

A standard method to reduce the computations is separately encoding each sentence into a vector representation and then compare any two of them by similarity distance. However, in contrast to the standard BERT model, the performance of sentence matching is constrained. For instance, SBERT [20] using the siamese BERT-networks that decreased the performance by 3-4 points evaluated by Spearman correlation [19] on STS-Benchmark [6], which implies room for improvement. We argue that the siamese BERT-networks are limited to capture the full complexity of global semantic matching, neglecting the word-level interaction features across two sentences. The feature has been proved vital for predicting matching degrees [17, 25].

Motivated by these observations, we propose a Dual-view distilled BERT (DvBERT) by incorporating the word-level interaction features into sentence embeddings while maintains the same efficiency as siamese BERT-networks. We take inspiration from Multi-view learning [10, 24] and train the sentence matching model from two views: (1) Siamese View, we start with the siamese BERT-networks as a backbone to derive sentence embeddings, to be able to capture semantics similarity efficiently by calculating distances on the two fixed-size vectors. (2) Interaction View, the standard pre-trained models with cross-sentence interactions are utilized, acting as multiple teachers that generate predictions about the training set provided to the siamese networks to learn. The association between the two views acts as a regularization term that trains a student with soft targets from the multiple teacher's output distributions, making the procedure similar to knowledge distillation [14]. In contrast of other distilled versions of BERT [21, 22], our method aims to optimize sentence embedding representations

SIGIR '21, July 11–15, 2021, Virtual Event, Canada

ACM ISBN 978-1-4503-8037-9/21/07...$15.00
https://doi.org/10.1145/3404835.3463057

with two heterogeneous networks, together with multi-task knowledge distillation [18], neither distilling large models into a small model [8, 15] nor born-again networks [9, 13]. Besides, we compared the loss weighting and teacher annealing strategy [9] during the distillation process, suggesting that the latter was more efficient. Experiments demonstrate that DvBERT can achieve superior performance than siamese BERT-networks on six STS datasets.

2 DUAL VIEW DISTILLED BERT

We first present DvBERT and describe how these views can be combined with multi-task knowledge distillation.

2.1 Siamese BERT-networks

For a given dataset $\mathcal{D}_l$, Siamese BERT-networks aims to predict a label $y \in \mathbf{Y}$ by leveraging similarity measure between the sentence embeddings, where $\mathbf{Y} = \{entailment, contradiction, neutral\}$ in natural language inference. For any sentence-pairs, the siamese BERT converts the two sentences into sequential vectors individually, and then pool these two vectors into two sentence embeddings $\mathbf{u}$ and $\mathbf{v}$. SBERT [20] compares different pooling strategies from multiple datasets, and gives the result that the `MEAN` strategy is significantly superior to `MAX` and `[CLS]` token strategy. Hereafter, the `MEAN` pooling is our default configuration. For classification tasks, such as NLI, we concatenate $\mathbf{u}$, $\mathbf{v}$, and $|\mathbf{u} - \mathbf{v}|$ followed by a fully-connected layer, which projects the hidden size into a probability distribution.

$$p(y|\mathbf{u}, \mathbf{v}; \theta) = \mathtt{softmax}(W[\mathbf{u}, \mathbf{v}, |\mathbf{u} - \mathbf{v}|]),$$

where θ represents all learnable parameters from BERT, shared for $\mathbf{u}, \mathbf{v}$. And $W \in \mathbb{R}^{3d \times n}$ is the parameter of the fully-connected layer. d is the dimension of the sentence embeddings. We optimize the standard cross-entropy loss.

2.2 Cross Sentence Interaction

We use multiple teachers from different pre-trained models to introduce interaction matrices across words to enrich the word-level interactive features. Each model first pre-trains with labeled data, then re-labeling the data and adds it to a new training set. Specifically, as illustrated in Fig 1 (top), we concatenate the sentence-pair $Q = \{Q_i\}_{i=1,\dots,m}$ and $T = \{T_i\}_{i=1,\dots,N}$ into a text sequence `[[CLS]`Q`[SEP]`T`[SEP]]`. The `[CLS]` token is regarded as an aggregated semantic gap of the input sentence-pair since it is used to predict whether a sentence-pair coherent or not during pre-training. Let $\mathbf{z}^c_k$ be the `[CLS]` token from the kth pre-trained model, which followed a single fully-connected layer culminating in a softmax layer as our classifier:

$$q(y|\mathbf{z}^c_k; \phi_k) = softmax(O\mathbf{z}^c_k)),$$

where ϕ_k and $O \in \mathbb{R}^{d \times n}$ are the model parameters. The siamese BERT learns from the hard targets as well as soft targets from the teachers. Supposing that the ϕ_k and O has been optimized by cross-entropy loss, DvBERT trains the siamese BERT by minimizing

$$\mathcal{L}(\theta, W) = \sum_{k=1}^{K} D(q(y|\mathbf{z}^c_k; \phi_k), p(y|\mathbf{u}, \mathbf{v}; \theta)),$$

where D is a distance function between probability distributions, here we use the KL-divergence. K is the number of teachers. We hold the teacher predictions $q(y|\mathbf{z}^c_k; \phi)$ fixed when training the student. The BERT from the two views are not sharing since early experiments with sharing did not improve results.

2.3 Teacher Annealing

We leverage teacher annealing [9] strategy, which mixes the teacher prediction with the gold label during training. Teacher annealing progressively reduces the weight of soft targets as the training advances, making the student learning from the teacher to hard targets. This method ensures the student gets a rich training signal early in training but is not constrained to only simulating the teacher. Specifically, summarizing the siamese BERT, and other K BERT-related pre-trained model with cross sentence attention, the objective can be written as:

$$\mathrm{L}\,(\theta, W) = \textstyle\sum_{k=1}^{K} D(\lambda y + (1-\lambda) q(y|\mathbf{z}^c_k; \phi_k), p(y|\mathbf{u}, \mathbf{v}; \theta))$$

where λ is increase linearly from 0 to 1. In the beginning, $\lambda = 0$, which means the model is trained entirely based on the soft targets from teachers. As the model converges gradually, the model learns from hard targets with more confidence.

3 EXPERIMENTS

In this section, we present our approach to NLI and STS datasets.

3.1 Dataset

The NLI dataset consists of SNLI [5] and MultiNLI [23], annotated with the labels contradiction, entailment, and neutral. STS [4] assesses the matching degree to which two sentences are semantically equivalent to each other, which are human-annotated with a level of equivalence from 1 to 5. We follow the previous works [7, 11] to merge the training and test datasets in both NLI data as pre-training datasets of 940k sentence pairs. STS 2012-2016 datasets have no training data but 26k test data, so the datasets are used to evaluate the pre-trained DvBERT on NLI. STS-B is a collection of 8.6k sentence pairs and contains training, development, and test sets drawn from heterogeneous sources.

3.2 Training and Evaluation Settings

We pre-train DvBERT with a 3-way softmax classifier for one epoch on NLI datasets. The batch size is set to 16, and the dropout rate is set to 0.1 for all modules. We use Adam optimizer [16] for model training. We set the initial learning rate to 2e-5 with a decay ratio of 1.0, a linear learning rate warm-up over 10 percent. For fune-tuning STS-B, we replace the $(u, v, |u - v|)$ to $cosine(u, v)$, and set the distance metric to the mean square error loss for regression training. The epoch numbers were set to 4, and other hyperparameters keep the same as the NLI task setting. Basically, We keep hyper-parameters consistent with SBERT. Our two default teachers are standard BERT, RoBERTa. We also evaluate the performance of DvRoBERTa by replacing the siamese BERT to RoBERTa.

3.3 Unsupervised STS

We apply STS 2012 - 2016 and STS-B test data to evaluate the performance without any task-specific training data. We use the

	STS12	STS13	STS14	STS15	STS16	STS-B	Avg.
BERT Avg. embedding	38.78	57.98	57.98	63.15	61.06	46.35	54.22
BERT [CLS] embedding	20.16	30.01	20.09	36.88	38.08	16.50	26.95
SBERT-base	70.4	71.77	70.66	78.67	74.11	76.28	73.64
SRoBERTa-base	71.70	73.43	71.47	**80.79**	75.99	77.02	75.06
DvBERT-base	70.52	73.17	71.18	79.88	75.08	77.96	74.63
DvRoBERTa-base	**72.42**	**73.44**	**72.21**	80.43	**76.52**	**78.32**	**75.56**
SBERT-large	71.68	72.79	72.20	80.32	76.45	78.00	75.24
SRoBERTa-large	72.14	76.69	74.12	79.81	75.97	78.60	76.22
DvBERT-large	72.95	72.26	71.87	79.27	76.16	78.28	75.13
DvRoBERTa-large	**74.99**	**76.16**	**73.34**	**81.93**	**78.77**	**79.61**	**77.47**

Table 1: Spearman correlation of STS tasks without fine-tuning on task-specific data.

Spearman correlation between the cosine similarity of the sentence embeddings and the gold labels. The results are reported in Table 1. The first two lines show BERT without training on NLI get rather poor performance pooled by `MEAN` or `[CLS]` token. Especially for `[CLS]` token, as it is mainly used to distinguish the segment-pair, whether coherence or not, there is a discrepancy in single sentence representations. We evaluate our approach compared with SBERT (SRoBERTa) on six STS datasets. We can observe that the models with pre-training on NLI improve a large margin than those are not. The dual-view method substantially impacts the performance of the two pre-trained models, obtaining 0.56%-1.9% improvement on average.

	Base models	Large models
BERT-NLI	87.33 ± 0.23	89.09 ± 0.36
RoBERTa-NLI	**89.77** ± 0.47	**91.12** ± 0.17
SBERT	84.57 ± 0.2	84.72 ± 1.01
SRoBERTa	84.89 ± 0.34	86.13 ± 0.94
DvBERT	84.67 ± 0.23	85.31 ± 0.21
DvRoBERTa	**85.31** ± 0.37	**86.23** ± 0.67
SBERT-NLI	85.01 ± 0.17	85.91 ± 0.58
SRoBERTa-NLI	85.40 ± 0.2	86.15 ± 0.35
DvBERT-NLI	85.15 ± 0.24	86.21 ± 0.13
DvRoBERTa-NLI	**86.05** ± 0.22	**86.98** ± 0.46

Table 2: Spearman correlation of STS tasks. The average of 10 runs with different random seeds is reported. "-NLI" indicates the model is pre-trained on NLI data.

3.4 Fine-tuning on STS-B

Since STS-B is a regression task, we adopt the cosine similarity followed mean square loss to take the place of both the fully-connected layer and cross-entropy loss from the NLI classification task. The experiment was divided into three setups. (1) Standard BERT/RoBERTa pre-train on NLI, then fine-tune on STS-B; (2) DvBERT trains only on STSb; (3) DvBERT first trained on NLI for all teachers and the student, then trains on STS-B. The report gives the average Spearman correlation and its standard error after ten runs, as shown in table 2. The first two lines are standard BERT and RoBERTa models, which are used as our teacher models to capture the word-level attention to each other and significantly achieve the best results. We can observe that the pre-training on NLI consistently improves the performance for the shown models since NLI enhances the models towards language understanding. The results demonstrate that DvBERT can improve generalization capability.

3.5 Effect of Dual View Distillation

In order to observe how the dual-view method generalizes to STS set from the NLI training set, we plot the SRoBERTa vs. DvRoBERTa spearman correlation with every 1000 steps for one epoch. The base model is configured with 12 layers, 12 self-attention heads, and the hidden size of 768 while the large model is set to 24 layers, 16 self-attention heads, and the hidden size of 1024. In Figure 2, we can see that both base and large models improve the ability to generalization. We find the large model with dual-view achieves more benefits than the base model because of the large model of teachers. Notably, DvRoBERTa has relatively poor performance in the early stage, as the student mainly learns from teachers, which are given higher weights to loss of soft targets.

3.6 Effect of Teacher Annealing

To verify the effect of teacher annealing strategy for DvBERT, we show the importance of teacher annealing vs. loss weighting strategy. The loss weighting strategy combined the loss of hard targets and soft targets by weighted summation. The hyper-parameter α from 0 to 1, it weights the loss of the soft target for optimization. As seen in Figure 3, the left eleven error bars show the cosine spearman correlation of DvRoBERTa-base with different the various α. Using pure hard targets without teacher annealing (i.e., $\alpha = 0$) performs no better than weighted distillation. It further illustrated the dual-views from sentence-pairs can boost the single view of siamese BERT. On the other hand, the teacher annealing strategy (the right bar) shows a better correlation than the loss weighting strategy.

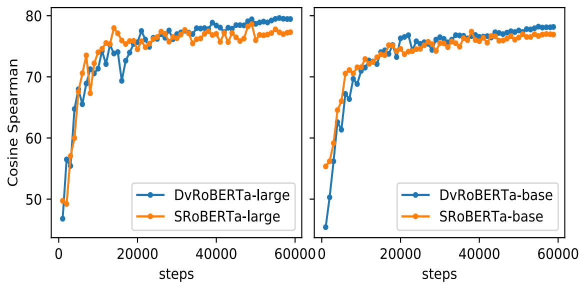

Figure 2: Spearman correlation for SRoBERTa and DvRoBERTa.

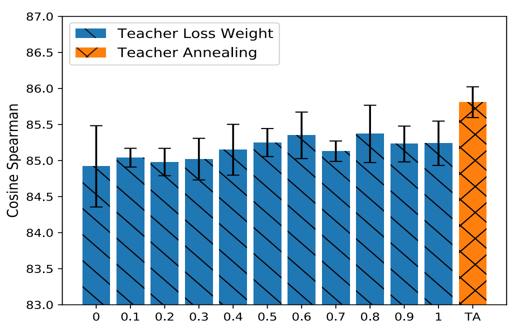

Figure 3: Comparison of teacher loss Weighting and teacher annealing

4 CONCLUSIONS

We proposed a dual-view approach that enhances sentence embeddings for matching, which adopt two heterogeneous networks to adapt to two views. Specifically, it allows siamese BERT-networks to effectively leverage the cross sentence interaction models while keeping the efficiency of using sentence embedding in retrieval tasks. The experiments on six STS datasets show that our models achieve consistent gains and outperform the performance of siamese BERT-networks.

REFERENCES

[1] Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Iñigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria, and Janyce Wiebe. 2015. SemEval-2015 Task 2: Semantic Textual Similarity, English, Spanish and Pilot on Interpretability. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. Association for Computational Linguistics, Denver, Colorado, 252–263. https://doi.org/10.18653/v1/S15-2045

[2] Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2014. SemEval-2014 Task 10: Multilingual Semantic Textual Similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*. Association for Computational Linguistics, Dublin, Ireland, 81–91. https://doi.org/10.3115/v1/S14-2010

[3] Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. *SEM 2013 shared task: Semantic Textual Similarity. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*. Association for Computational Linguistics, Atlanta, Georgia, USA, 32–43. https://www.aclweb.org/anthology/S13-1004

[4] Eneko Agirre, Daniel M. Cer, Mona T. Diab, and Aitor Gonzalez-Agirre. 2012. SemEval-2012 Task 6: A Pilot on Semantic Textual Similarity. In *Proceedings of the 6th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2012, Montréal, Canada, June 7-8, 2012*. 385–393. https://www.aclweb.org/anthology/S12-1051/

[5] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*. 632–642. https://doi.org/10.18653/v1/d15-1075

[6] Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Association for Computational Linguistics, Vancouver, Canada, 1–14. https://doi.org/10.18653/v1/S17-2001

[7] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal Sentence Encoder for English. In *EMNLP 2018: System Demonstrations, Brussels, Belgium, October 31 - November 4, 2018*. 169–174. https://doi.org/10.18653/v1/d18-2029

[8] Debajyoti Chatterjee. 2019. Making Neural Machine Reading Comprehension Faster. *CoRR* abs/1904.00796 (2019). arXiv:1904.00796 http://arxiv.org/abs/1904.00796

[9] Kevin Clark, Minh-Thang Luong, Urvashi Khandelwal, Christopher D. Manning, and Quoc V. Le. 2019. BAM! Born-Again Multi-Task Networks for Natural Language Understanding. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, Anna Korhonen, David R. Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, 5931–5937. https://doi.org/10.18653/v1/p19-1595

[10] Kevin Clark, Minh-Thang Luong, Christopher D. Manning, and Quoc V. Le. 2018. Semi-Supervised Sequence Modeling with Cross-View Training. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*. 1914–1925. https://doi.org/10.18653/v1/d18-1217

[11] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. In *EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*. 670–680. https://doi.org/10.18653/v1/d17-1070

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*. 4171–4186. https://doi.org/10.18653/v1/n19-1423

[13] Tommaso Furlanello, Zachary Chase Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. 2018. Born-Again Neural Networks. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018 (Proceedings of Machine Learning Research)*, Jennifer G. Dy and Andreas Krause (Eds.), Vol. 80. PMLR, 1602–1611. http://proceedings.mlr.press/v80/furlanello18a.html

[14] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the Knowledge in a Neural Network. *CoRR* abs/1503.02531 (2015). arXiv:1503.02531 http://arxiv.org/abs/1503.02531

[15] Weipeng Huang, Xingyi Cheng, Kunlong Chen, Taifeng Wang, and Wei Chu. 2019. Toward Fast and Accurate Neural Chinese Word Segmentation with Multi-Criteria Learning. *CoRR* abs/1903.04190 (2019). arXiv:1903.04190 http://arxiv.org/abs/1903.04190

[16] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[17] Wuwei Lan and Wei Xu. 2018. Neural Network Models for Paraphrase Identification, Semantic Textual Similarity, Natural Language Inference, and Question Answering. In *COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*. 3890–3902. https://www.aclweb.org/anthology/C18-1328/

[18] Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Improving Multi-Task Deep Neural Networks via Knowledge Distillation for Natural Language Understanding. *CoRR* abs/1904.09482 (2019). arXiv:1904.09482 http://arxiv.org/abs/1904.09482

[19] Leann Myers and Maria J Sirois. 2004. Spearman correlation coefficients, differences between. *Encyclopedia of statistical sciences* 12 (2004).

[20] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*. 3980–3990. https://doi.org/10.18653/v1/D19-1410

[21] 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* (2019).

[22] Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. 2019. Patient Knowledge Distillation for BERT Model Compression. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, 4322–4331. https://doi.org/10.18653/v1/D19-1441

[23] Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*. 1112–1122. https://doi.org/10.18653/v1/n18-1101

[24] Chang Xu, Dacheng Tao, and Chao Xu. 2013. A Survey on Multi-view Learning. *CoRR* abs/1304.5634 (2013). arXiv:1304.5634 http://arxiv.org/abs/1304.5634

[25] Weidi Xu, Xingyi Cheng, Kunlong Chen, and Taifeng Wang. 2020. Symmetric Regularization based BERT for Pair-wise Semantic Reasoning. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1901–1904.