INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Relation-based Counterfactual Data Augmentation and Contrastive Learning for Robustifying Natural Language Inference Models

Heerin Yang[1,3*], *Seung-won Hwang*[2], *Jungmin So*[1]

[1]Dept. of Computer Science and Engineering, Sogang University, Korea
[2]Dept. of Computer Science and Engineering, Seoul National University, Korea
[3]LG Electronics, Korea

heerin.yang@lge.com, seungwonh@snu.ac.kr, jso1@sogang.ac.kr

Abstract

Although pre-trained language models show good performance on various natural language processing tasks, they often rely on non-causal features and patterns to determine the outcome. For natural language inference tasks, previous results have shown that even a model trained on a large number of data fails to perform well on counterfactually revised data, indicating that the model is not robustly learning the semantics of the classes. In this paper, we propose a method in which we use token-based and sentence-based augmentation methods to generate counterfactual sentence pairs that belong to each class, and apply contrastive learning to help the model learn the difference between sentence pairs of different classes with similar contexts. Evaluation results with counterfactually-revised dataset and general NLI datasets show that the proposed method can improve the performance and robustness of the NLI model.

Index Terms: natural language inference, counterfactual data augmentation, contrastive learning

1. Introduction

A recently popular approach to solving natural language processing (NLP) problems is to use a pre-trained language model such as BERT [1] and RoBERTa [2], then fine-tune the model on a downstream task such as text classification. Although trained models achieve outstanding performance in various tasks such as sentiment analysis [3, 4] and natural language inference (NLI) [5, 6], it is well-known that these models often make decisions based on spurious patterns and correlations and therefore do not generalize well to other datasets. For example, NLI classifiers may learn that a sentence pair having significant lexical overlap is a sign that they are in an entailment relationship, which is not necessarily true [7].

Kaushik et al. [8] showed that a model trained on the original dataset performs poorly on a counterfactually revised dataset, which is another evidence that the model is relying on spurious patterns to classify data. For collecting counterfactually revised data, human workers were asked to edit given data samples to produce new samples that have different labels than the original ones. For example, if the given NLI sentence pair is "A man in a boom lift bucket welds. A man is working. (entailment)", then the worker writes counterfactual samples by revising the premise such as "A woman in a boom lift bucket welds. A man is working (contradiction)" or "A person in a boom lift bucket welds. A man is working. (neutral)". A classifier trained on the original dataset classifies all three pairs as entailment.

In this paper, we consider automatically generating counterfactual data for NLI tasks. While Kaushik et al. [8] claims that the counterfactually-revised train sets by human workers could improve model performance on the challenge sets, human annotation is costly. Our goal is to make the NLI model more robust to counterfactually revised data without getting help from human annotators. Compared to other NLP tasks where a single sentence or passage is considered as input, NLI poses a unique challenge where a sentence pair is given as input and its relation is an important feature for classification. However, existing augmentation methods such as EDA [9] regards an NLI sentence pair as a single unit of input without considering their relation. In contrast, we counterfactually augment hypothesis sentences for a fixed premise and vice versa, and represent their relation more explicitly, as a distance, to minimize or maximize during contrastive learning.

Specifically, we apply contrastive learning with the generated set, pulling the original pair and the generated pair with the same label together while pushing the original pair and other generated pairs away, in the embedding space. We empirically find that this method is more effective than applying supervised contrastive learning with unrelated sentence pairs [10]. There are other recent methods [11, 12, 13] using automatic data augmentation and contrastive learning to make the model more robust, but their improvements are limited mostly because they do not consider the unique characteristics of NLI where the inputs are pairs and their relations are important. The experimental results show that the proposed method achieves better accuracy compared to other robust text classification methods on counterfactually revised NLI datasets [8] as well as general NLI datasets[1].

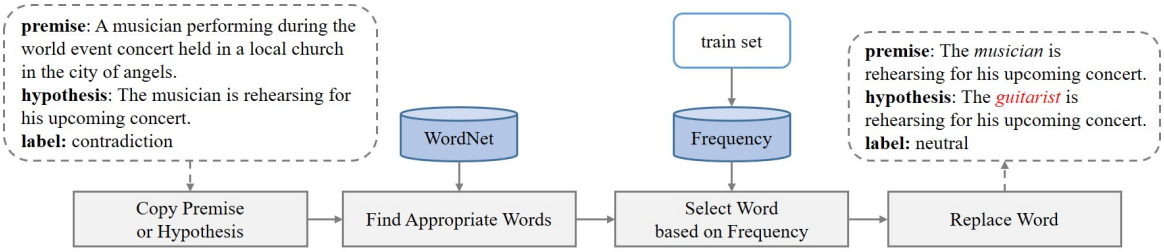
2. Related work

Data augmentation for NLP tasks can be divided into token-level and sentence-level augmentation. Token-level augmentation modifies individual words, such as substituting a word with synonyms [14, 15], randomly inserting, deleting, or swapping tokens [9]. Language models can be used for augmentation, by masking a particular word and using the model to fill in the blank [16, 17]. The quality of token-based augmentation depends on selecting which token to insert or remove, such as finding the rationale tokens and replacing them [11, 12, 13]. Sentence-level augmentation generates an entire sentence rather than modifying tokens from the original text. Examples include back-translation [18], paraphrasing [19], and conditional generation [20]. While sentence-level augmentation can generate more diverse text compared to token-level augmentation, it is more difficult to assign labels or determine the quality of generated data. Therefore, filtering methods based on teacher models are often used to select good quality data [21].

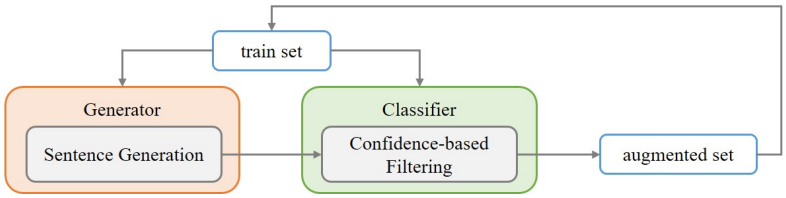
*This work is done when the student was at Songang University.

[1]The codes are available at https://github.com/hryang06/rda-rcl.

10.21437/Interspeech.2023-277



(a) *Process of token-level data augmentation. The word "musician" is replaced to "guitarist", and the label is changed to "neutral".*



(b) *Process of sentence-level data augmentation. The classifier and the generator are iteratively trained with augmented data.*

Figure 1: *Our proposed data augmentation framework.*

Contrastive learning is recently recognized as an effective method to improve model performance [22, 23]. It is shown to make the model more robust to perturbations and improve its generalization ability. In unsupervised contrastive learning, an original input is paired with a slightly modified input to form a positive pair and paired with a different sample to form a negative pair. For example, in C^2L [13], a negative pair is created by masking keyword tokens from the original text, while a positive pair is created by masking non-keyword tokens. It is also possible to use contrastive learning in a supervised learning context, gathering same-class samples together in the feature space, while separating different-class samples [10].

3. Proposed Method

3.1. Relation-based Counterfactual Data Augmentation

In our proposed method, we first generate a set of entailment, neutral, and contradiction sentence pairs for each sentence pair in the train set. We apply two major data augmentation approaches, token-level and sentence-level augmentation, tailored for NLI tasks to generate factual and counterfactual data.

3.1.1. Token-level Data Augmentation

While simple methods such as synonym replacement [9] can be used to generate class-preserving data, it is not trivial to generate counterfactual data. Suppose the original premise-hypothesis pair is "A man is walking down the street. A man is outside walking. (entailment)" Changing the hypothesis to "A woman is outside walking." will make the relation contradictory. However, if the original premise was "A person is walking down the street.", changing the hypothesis as such will not alter the label (neutral). In our proposed method, we take only one sentence from the original pair and copy the sentence to make an entailment pair (e.g. "A man is outside walking. A man is outside walking."). From this pair, we apply word substitution on either premise or hypothesis to generate sentence pairs that belong to the three classes.

Figure 1a shows our token-level data augmentation process. We first choose a random noun word in the sentence using `spaCy`[2]. Then, we use `WordNet`[3] to find the substitution words. Table 1 shows how the substitution words are selected based on the revised sentence and the target class. For example, we choose a synonym or a hypernym to make an entailment sentence, a hyponym to make a neutral sentence, and an antonym or co-hyponym to make a contradiction sentence. Among candidate words, we sample a word based on its frequency in the train set. In the case where no candidate substitution is found, the sentence pair is omitted from contrastive learning. Table 2 shows the sentences generated by four different configurations. One limitation of our scheme is that we only substitute nouns in the sentence. Substituting words other than nouns for counterfactual data generation is left for future work.

Target Label	Revise Premise	Revise Hypothesis
entailment	synonym, hyponym	synonym, hypernym
neutral	hypernym	hyponym
contradiction	antonym, co-hyponym	

Table 1: *Relation types used in word substitution to generate a sample of the target label.*

3.1.2. Sentence-level Data Augmentation

Conditional generation techniques such as LAMBADA [20] can be used to generate the hypothesis sentence conditioned on the premise sentence (and the label), and vice versa. We follow the basic approach of LAMBADA, but instead of generating independent samples, we let the generator create a set of entailment, neutral, and contradiction sentences for each input sentence.

Pre-trained sequence-to-sequence language models such as GPT-2 [24], BART [25], and T5 [26] can be used as a sentence generator, and we use T5 model to generate counterfactual premise or hypothesis sentences. The problem with using a generator model is that the generated sentence pairs may have incorrect labels. A typical method to address this problem is to evaluate the generated data samples on a classifier model trained on the original data, and filter out samples that have low confidence in the target class [11, 21].

[2]https://spacy.io
[3]https://wordnet.princeton.edu

Token-level Data Augmentation	
CPRP (Copy-Premise-Revise-Premise)	**CHRP** (Copy-Hypothesis-Revise-Premise)
PE: A *man* sitting down hold 2 items and a camera around their neck. **PN:** A person sitting down hold 2 items and a camera around their *body*. **PC:** A person sitting down hold 2 items and a camera around their *head*. **H:** A *person* sitting down hold 2 items and a camera around their *neck*.	**PE:** The *puppy* is heavily panting as he runs in circles **PN:** The *animal* is heavily panting as he runs in circles **PC:** The *wolf* is heavily panting as he runs in circles **H:** The *dog* is heavily panting as he runs in circles
CPRH (Copy-Premise-Revise-Hypothesis)	**CHRH** (Copy-Hypothesis-Revise-Hypothesis)
P: A *man* in a pointed hat carries many large bags full of food down a city street. **HE:** A *person* in a pointed hat carries many large bags full of food down a city street. **HN:** A *boy* in a pointed hat carries many large bags full of food down a city street. **HC:** A *woman* in a pointed hat carries many large bags full of food down a city street.	**P:** The *musician* is rehearsing for his upcoming concert. **HE:** The *person* is rehearsing for his upcoming concert. **HN:** The *guitarist* is rehearsing for his upcoming concert. **HC:** The *cowboy* is rehearsing for his upcoming concert.
Sentence-level Data Generation	
P → H (Generate Hypothesis from Premise and Label)	**H → P** (Generate Premise from Hypothesis and Label)
P: A man in a sweatshirt is riding a bicycle holding some long thin planks of wood **HE:** *A man is riding a bike.* **HN:** *A man is biking to work* **HC:** *A man is sitting on a couch.*	**PE:** *A woman in a sweater is smiling.* **PN:** *A woman sitting on the floor looking at something.* **PC:** *A man in a blue jacket and hat talks to an audience over ice.* **H:** The woman is wearing a sweater.

Table 2: *Counterfactual data generated using our data augmentation methods. Replaced or generated words are marked in red.*

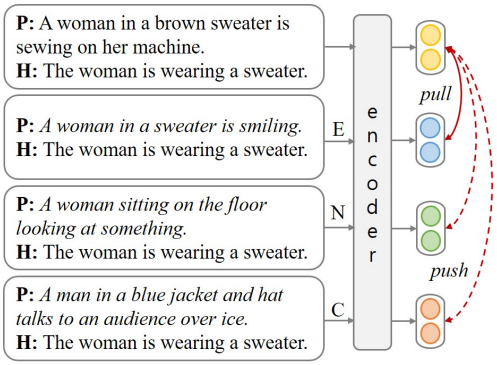


Figure 2: *Contrastive learning with augmented data.*

Figure 1b illustrates the sentence-level augmentation process. We first train a classifier model and a generator model with the original train set. Then, the generator model generates three sentences for each sentence pair in the original set. For the generated sentence pairs, we apply confidence-based filtering and drop samples with model confidence lower than a threshold τ. We go through an iterative process where the augmented set becomes the train set which is used to train the classifier and the generator. This iterative process goes on until we have obtained a full set (entailment, neutral, contradiction, plus original pair) for over 95% of samples in the original set. The samples that could not construct a full set during the augmentation stage are omitted from relation-based contrastive learning. Table 2 shows the sentences generated using our method.

3.2. Relation-based Contrastive Learning

Once each sentence pair is augmented with sentence pairs corresponding to all three classes, we train the classifier with the augmented set. The model is first trained with the contrastive learning objective. A set of four sentence pairs (original, entailment, neutral, contradiction) is passed through the encoder to obtain the sentence embedding vectors. Then, cosine similarity is measured between the original embedding vector and the embedding vectors of other sentence pairs. Finally, the contrastive loss $\mathcal{L}_{CL}$ is calculated according to Eq. 1.

$$\mathcal{L}_{CL} = -log\frac{exp(sim(x, x_y)/T)}{\sum_{c=0}^{C} exp(sim(x, x_c)/T)} \quad (1)$$

The contrastive learning process is shown in Figure 2. Suppose the original label is entailment. Then, the distance between the embedding vectors of the original and entailment pair is minimized, while the distances between the embedding vectors of the original and other pairs are maximized. After contrastive learning, the model is trained using cross-entropy loss.

4. Experimental Results

4.1. Experiment Setup

We use the counterfactually augmented SNLI dataset (CF-SNLI), which is also used by previous works for testing the robustness of NLI models [11, 12, 13]. CF-SNLI set contains "original" train and test sets sampled from SNLI [5]. It also has "revised premise" (RP) and "revised hypothesis" (RH) set, where the premise and hypothesis sentences are revised by human workers to produce sentence pairs with relations other than the original pair. We evaluate with all CF-SNLI test sets and also general NLI datasets– SNLI test set, MNLI dev-matched set, and MNLI dev-mismatched set [6].

We use BERT (bert-base-uncased) and RoBERTa (roberta-base) as pre-trained language models. For BERT, the model is trained with contrastive loss for 10 epochs (lr=1e-5), followed by cross-entropy loss for 3 epochs (lr=3e-5). For RoBERTa, the model is trained with contrastive loss for 10 epochs (lr=2e-6), followed by cross-entropy loss for 5 epochs (lr=1e-5). We use 0.1 as the temperature T in Eq. 1. In sentence generation, the threshold τ is empirically tuned to 0.9. The results were not sensitive to τ, unless we choose a very low number.

We compare the performance of our method with other recent methods based on counterfactual data. SSMBA [11] uses a corruption function to perturb the original text and a reconstruction function to generate a new text in the underlying data manifold. MASKER [12] selects keywords in the text using attention scores or gradients and applies masked keyword reconstruction to help the model learn the context rather than relying on particular tokens. C^2L [13] generates factual and counter-

Model	CF-SNLI			
	Original	RP	RH	RP & RH
BERT-base	75.5 ±1.4	41.8 ±2.6	64.5 ±2.0	53.1 ±2.2
+ SSMBA [11] *	75.8 ±1.5	42.5 ±0.9	65.0 ±0.3	53.8 ±0.5
+ MCL (grad+SL) [12] *	78.3 ±1.1	40.0 ±1.3	64.5 ±1.3	52.2 ±1.3
+ C^2L [13] *	76.2 ±1.7	43.1 ±2.5	65.8 ±1.7	54.5 ±2.1
+ SCL	75.7 ±1.1	42.3 ±1.2	65.9 ±0.9	54.1 ±1.0
+ RDA *(Ours)*	77.7 ±1.1	46.5 ±0.8	67.3 ±1.6	56.9 ±1.0
+ RDA-RCL *(Ours)*	**79.3** ±1.0	**47.5** ±0.9	**68.0** ±0.5	**57.8** ±0.6
RoBERTa-base	81.4 ±1.9	51.5 ±0.5	68.2 ±1.4	59.8 ±0.9
+ SCL	82.0 ±1.2	51.5 ±0.7	68.6 ±1.3	60.1 ±1.0
+ RDA *(Ours)*	84.5 ±1.0	59.3 ±0.8	73.3 ±0.8	66.3 ±0.4
+ RDA-RCL *(Ours)*	**84.7** ±1.3	**59.6** ±0.8	**73.6** ±0.4	**66.6** ±0.6

Table 3: *Accuracy of methods on counterfactually-augmented SNLI dataset. * results from Choi et al. [13].*

Model	SNLI	MNLI	
	test	dev-m	dev-mm
BERT-base	76.0 ±0.7	52.1 ±2.5	51.8 ±3.1
+ SCL	75.8 ±0.5	52.8 ±2.0	53.4 ±2.5
+ RDA	76.7 ±0.4	59.5 ±0.9	60.5 ±1.6
+ RDA-RCL	**77.8** ±0.6	**60.1** ±2.1	**61.5** ±2.9
RoBERTa-base	79.7 ±0.9	58.5 ±3.3	60.0 ±3.8
+ SCL	80.2 ±1.2	59.9 ±2.8	62.0 ±3.1
+ RDA	83.1 ±0.2	69.7 ±0.1	70.8 ±0.4
+ RDA-RCL	**83.1** ±0.4	**70.5** ±0.3	**71.6** ±0.5

Table 4: *Accuracy on SNLI test and MNLI dev sets.*

factual samples by masking non-causal and causal tokens in the original text, and applies contrastive learning to help the model learn to rely on causal tokens.

4.2. Results

In the tables, BERT-base and RoBERTa-base are baseline models fine-tuned with CF-SNLI original train set, and SCL refers to supervised contrastive learning [10], where contrastive learning is applied without data augmentation. RDA (Relation-based Data Augmentation) and RCL (Relation-based Contrastive Learning) are the components of our proposed method. RDA is the case where only data augmentation is applied, whereas RDA-RCL is the case where contrastive learning is also applied. We seek to answer the following research questions.

RQ1: Does the proposed method perform better than the baseline and other data augmentation methods? In Table 3, models trained with different methods were evaluated on CF-SNLI test sets. We can observe that the proposed method achieves higher accuracy over the baseline and other methods in all sets for both BERT and RoBERTa models. The performance improvement is 6-8% for the RP set and 3-5% for the RH set, respectively. The proposed method also achieves 3-4% improvement over the baseline on the original test set, which indicates that the method not only improves robustness to counterfactual revisions but helps the model performance in general.

RQ2: Does the proposed method show good performance on the general NLI sets? Since it is important to see whether the proposed method is effective in datasets other than CF-SNLI, we have evaluated the models on SNLI test set and MNLI dev sets. Since CF-SNLI original set is sampled from SNLI, we can say that SNLI is an in-domain set whereas MNLI is an out-of-domain set. Table 4 shows that the proposed method achieves significantly higher accuracy over baseline for both BERT and RoBERTa models. While the accuracy improvement is 2-4% for SNLI, our method achieves 8-12% higher accuracy over baseline on MNLI dev sets, which shows that the method is also effective in improving generalization performance.

RQ3: Is the proposed method better than general supervised contrastive learning? The relation-based contrastive learning applies supervised contrastive learning on sentence pairs with the common premise or hypothesis. The question is whether it is better than applying general SCL where contrastive learning is applied to different sentence pairs. Table 3 and 4 show that applying general SCL only achieves marginal improvement over baseline, while RDA-RCL shows significantly better results for different datasets as well as different models.

RQ4: Does applying relation-based contrastive learning helps improving model performance? Since we assign labels to counterfactually generated sentence pairs, augmenting them to the train set already helps improve model performance. However, applying relation-based contrastive learning further boosts performance. In Table 3 and 4, RDA-RCL achieves up to 2% higher accuracy over RDA for varying datasets and models, while there is no case where RCL degrades the performance.

Overall, the proposed method is an effective way to robustify NLI models against counterfactual revisions, as well as improve model accuracy and generalization performance.

5. Conclusions

This paper studied the effectiveness of relation-based data augmentation and contrastive learning on NLI tasks. For a given sentence pair, the proposed method applies token-based and sentence-based augmentation to generate a set of counterfactual sentence pairs for all classes. Relation-based contrastive learning is done using the set of counterfactual sentence pairs to help the model effectively learn the difference between classes. Empirical results show that our methods can improve the robustness of classifier models on NLI tasks. Since any sentences can be used as input to our methods, a possible future work can use our methods to create a large number of NLI sentence pairs using inputs outside the train set.

6. Acknowledgements

This work was supported by the NRF (National Research Foundation) of Korea under grant no. 2021S1A5A2A03064795.

7. References

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Jun. 2019, pp. 4171–4186.

[2] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[3] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, Oregon, USA: Association for Computational Linguistics, Jun. 2011, pp. 142–150.

[4] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Seattle, Washington, USA: Association for Computational Linguistics, Oct. 2013, pp. 1631–1642.

[5] 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, Sep. 2015, pp. 632–642.

[6] A. Williams, N. Nangia, and S. Bowman, "A broad-coverage challenge corpus for sentence understanding through inference," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Jun. 2018, pp. 1112–1122.

[7] T. McCoy, E. Pavlick, and T. Linzen, "Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Jul. 2019, pp. 3428–3448.

[8] D. Kaushik, E. Hovy, and Z. C. Lipton, "Learning the difference that makes a difference with counterfactually augmented data," *International Conference on Learning Representations (ICLR)*, 2020.

[9] J. Wei and K. Zou, "EDA: Easy data augmentation techniques for boosting performance on text classification tasks," 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)*. Association for Computational Linguistics, Nov. 2019, pp. 6382–6388.

[10] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 18 661–18 673.

[11] N. Ng, K. Cho, and M. Ghassemi, "SSMBA: Self-supervised manifold based data augmentation for improving out-of-domain robustness," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 1268–1283.

[12] S. J. Moon, S. Mo, K. Lee, J. Lee, and J. Shin, "Masker: Masked keyword regularization for reliable text classification," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 15, pp. 13 578–13 586, May 2021.

[13] S. Choi, M. Jeong, H. Han, and S.-w. Hwang, "C2l: Causally contrastive learning for robust text classification," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2022.

[14] W. Y. Wang and D. Yang, "That's so annoying!!!: A lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using #petpeeve tweets," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal, Sep. 2015, pp. 2557–2563.

[15] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," in *Proceedings of the 28th International Conference on Neural Information Processing Systems*, ser. NIPS'15, 2015, p. 649–657.

[16] S. Kobayashi, "Contextual augmentation: Data augmentation by words with paradigmatic relations," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, New Orleans, Louisiana, Jun. 2018, pp. 452–457.

[17] F. Gao, J. Zhu, L. Wu, Y. Xia, T. Qin, X. Cheng, W. Zhou, and T.-Y. Liu, "Soft contextual data augmentation for neural machine translation," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 5539–5544.

[18] S. Edunov, M. Ott, M. Auli, and D. Grangier, "Understanding back-translation at scale," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 489–500.

[19] M. Iyyer, J. Wieting, K. Gimpel, and L. Zettlemoyer, "Adversarial example generation with syntactically controlled paraphrase networks," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, New Orleans, Louisiana, Jun. 2018, pp. 1875–1885.

[20] A. Anaby-Tavor, B. Carmeli, E. Goldbraich, A. Kantor, G. Kour, S. Shlomov, N. Tepper, and N. Zwerdling, "Do not have enough data? deep learning to the rescue!" *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, pp. 7383–7390, Apr. 2020.

[21] Y. Wu, M. Gardner, P. Stenetorp, and P. Dasigi, "Generating data to mitigate spurious correlations in natural language inference datasets," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 2660–2676.

[22] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," *arXiv preprint arXiv:1911.05722*, 2019.

[23] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," *arXiv preprint arXiv:2002.05709*, 2020.

[24] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[25] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Jul. 2020, pp. 7871–7880.

[26] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1–67, 2020.