TOWARDS ROBUST TEXTUAL REPRESENTATIONS WITH DISENTANGLED CONTRASTIVE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Although the self-supervised pre-training of transformer models has resulted in the revolutionizing of natural language processing (NLP) applications and the achievement of state-of-the-art results with regard to various benchmarks, this process is still vulnerable to small and imperceptible permutations originating from legitimate inputs. Intuitively, the representations should be similar in the feature space with subtle input permutations, while large variations occur with different meanings. This motivates us to investigate the learning of robust textual representation in a contrastive manner. However, it is non-trivial to obtain opposing semantic instances for textual samples. In this study, we propose a disentangled contrastive learning method that separately optimizes the uniformity and alignment of representations without negative sampling. Specifically, we introduce the concept of momentum representation consistency to align features and leverage power normalization while conforming the uniformity. Our experimental results for the NLP benchmarks demonstrate that our approach can obtain better results compared with the baselines, as well as achieve promising improvements with invariance tests and adversarial attacks.

1 INTRODUCTION

The self-supervised pre-training of transformer models has revolutionized natural language processing (NLP) applications. Such pre-training with language modeling objectives provides a useful initial point for parameters that generalize well to new tasks with fine-tuning. However, there is a significant gap between task performance and model generalizability. Previous approaches have indicated that neural models suffer from poor **robustness** when encountering *randomly permuted contexts* Ribeiro et al. (2020) and *adversarial examples* Jin et al. (2019).

To address this issue, several studies have attempted to leverage data augmentation or adversarial training into pre-trained language models (LMs) Wei & Zou (2019); Jin et al. (2019); Ng et al. (2020), which has indicated promising directions for the improvement of robust textual representation learning. Such methods generally augment textual samples with synonym permutations or back translation and fine-tune downstream tasks on those augmented datasets. Representations learned from instance augmentation approaches have demonstrated expressive power and contributed to the performance improvement of downstream tasks in robust settings. However, the previous augmentation approaches mainly focus on the supervised setting and neglect large amounts of unlabeled data. Moreover, it is still not well understood whether a robust representation has been achieved or if the leveraging of more training samples have contributed to the model robustness.

Specifically, a robust representation should be similar in the feature space with subtle permutations, while large variations occur with different semantic meanings. This motivates us to investigate robust textual representation in a contrastive manner. It is intuitive to utilize data augmentation to generate positive and negative instances for learning robust textual representation via auxiliary contrastive objects. However, it is non-trivial to obtain opposite semantic instances for textual samples. For example, given the sentence, "Obama was born in Honululu," we are able to retrieve a sentence such as, "Obama was living in Honululu," or, "Obama was born in Hawaii." There is no guarantee that these randomly retrieved sentences will have negative semantic meanings that contradict the original sample.

In this study, we propose a novel disentangled contrastive learning (DCL) method for learning robust textual representations. Specifically, we disentangle the contrastive object using two subtasks: feature alignment and feature uniformity Wang & Isola (2020). We introduce a unified model architecture to optimize these two sub-tasks jointly. As one component of this system, we introduce momentum representation consistency to align augmented and original representations, which explicitly shortens the distance between similar semantic features that contribute to feature alignment. As another component of this system, we leverage power normalization to enforce the unit quadratic mean for the activations, by which the scattering features within the same batch implicitly contribute to the feature uniformity. Our DCL approach is a unified, unsupervised, and model-agnostic approach, and therefore it is orthogonal to existing approaches. We conduct numerous experiments on NLP benchmarks, which demonstrate the effectiveness of this approach in normal and robust settings. The contributions of this study can be summarized as follows:

- We investigate robust textual representation learning problems and introduce a disentangled contrastive learning approach, which is unsupervised and model-agnostic.
- We introduce a unified model architecture to optimize the sub-tasks of feature alignment and uniformity, as well as providing theoretical intuitions.
- Extensive experimental results related to NLP benchmarks demonstrate the effectiveness of our method in the robust setting; we performed invariance tests and adversarial attacks, and verify that our approach can enhance state-of-the-art pre-trained language model methods.

2 Related Work

Recently, studies have shown that pre-trained models (PTMs) Devlin et al. (2019a); Liu et al. (2019) on the large corpus are beneficial for downstream NLP tasks, such as in GLUE Wang et al. (2018), SQuAD Rajpurkar et al. (2016b), and SNLI Bowman et al. (2015). The application scheme of these systems is to fine-tune the pre-trained model using the limited labeled data of specific target tasks. Since training distributions often do not cover all of the test distributions, we would like a supervised classifier or model to perform well on. Therefore, a key challenge in NLP is learning robust textual representations. Previous studies have explored the use of data augmentation and adversarial training to improve the robustness of pre-trained language models. Wei & Zou (2019) proposed easy data augmentation techniques for boosting performance on text classification tasks. Li & Qiu (2020) introduced a novel text adversarial training with token-level perturbation to improve the robustness of pre-trained language models. However, supervised instance-level augmentation approaches ignore those unlabeled data and do not guarantee the occurrence of real robustness in the feature space.

Our work is motivated by contrastive learning (Saunshi et al., 2019; Oord et al., 2018), which aims at maximizing the similarity between the encoded query q and its matched key k^+, while distancing randomly sampled keys $\{k_0^-, k_1^-, k_2^-, ...\}$. By measuring similarity with a score function $s(q, k)$, a form of contrastive loss function is considered as follows:

$$\mathcal{L}_{contrast} = -\log \frac{\exp(s(q, k^+))}{\exp(s(q, k^+)) + \sum_i \exp(s(q, k_i^-))}, \tag{1}$$

where k^+ and k^- are positive and negative instances, respectively. The score function $s(q, k)$ is usually implemented with the cosine similarity $\frac{q^T k}{\|q\| \cdot \|k\|}$. q and k are often encoded by a learnable neural encoder (e.g., BERT (Devlin et al., 2019b)). Contrastive learning have increasingly attracted attention, which is beneficial for unsupervised or self-supervised learning from computer vision (Wu et al., 2018; Oord et al., 2018; Ye et al., 2019; Tian et al., 2019; He et al., 2019; Chen et al., 2020b) to natural language processing (Mikolov et al., 2013; Mnih & Kavukcuoglu, 2013; Devlin et al., 2019b; Clark et al., 2020). Chi et al. (2020) formulate cross-lingual language model pre-training as maximizing mutual information between multilingual-multi-granularity texts. Clark et al. (2020) utilized a discriminator to predict whether a token is replaced by a generator given its surrounding context. Iter et al. (2020) proposed to pre-train language models with contrastive sentence objectives to predict the surrounding sentences given an anchor sentence. Wei et al. (2020) proposed to encourage parallel cross-lingual sentences to obtain an identical semantic representation and distinguish whether a specific word is contained within these sentences. To the best of our knowledge, this is the first study to apply contrastive learning to robust textual representation learning.

3 PRELIMINARIES ON LEARNING ROBUST TEXTUAL REPRESENTATIONS

Definition 1.Robust textual representation indicates that the representation is vulnerable to small and imperceptible permutations originating from legitimate inputs. Formally, we have the following:

$$g\left(X+z\right)=g(X),\text{ and }\operatorname{Sim}\left(f(X+z),f(X)\right)\geq\epsilon, \tag{2}$$

where z refers to the random or adversarial permutation of the input text and $g(.)$ takes input from x and outputs a valid probability distribution for tasks. $f(.)$ is the feature encoder, such as BERT. We are interested in deriving methods for pre-training representations that provide guarantees for the movement of inputs such that they are robust to permutations. Therefore, a robust representation should be similar in the feature space with subtle permutations, while large variations are observed for different semantic meanings. Such constraints are related to the well-known contrastive learning Arora et al. (2019); Chen et al. (2020a) schema as follows:

Remark. Robust representation is closely related to regularizing the feature space with the following constraints:

$$L_{contrast}=\sum(\sum_{1}^{m}|f(X)-f(X+z)|-\sum_{1}^{n}|f(X)-f(X')|) \tag{3}$$

where m and n are the number of positive and negative instances, respectively, regarding the original input, X, $X+z$ and X' are the positive and negative instances, respectively. Note that we can obtain $X+z$ via off-the-shelf tools such as data augmentation or back-translation. However, it is non-trivial to obtain negative instances for textual samples. Previous approaches Chen et al. (2020a); Giorgi et al. (2020); Fang & Xie (2020); Chi et al. (2020); Wei et al. (2020) regard random sampling of the remaining instances from the corpus as negative instances; however, there is no guarantee that those random instances are semantically irrelevant. Recent semantic-based information retrieval approaches Xiong et al. (2020) can obtain numerous similar semantic sentences via an approximate nearest neighbor Liu et al. (2005), which further indicates that negative sampling for sentences may result in noise.

In this study, inspired by the approach utilized by Wang & Isola (2020), we disentangle the contrast loss with the two following properties:

- *Alignment*: two samples forming a positive pair should be mapped to nearby features, and therefore be (mostly) invariant to unneeded noise factors.
- *Uniformity*: feature vectors should have an approximately uniform distribution on the unit hypersphere, thereby preserving as much information of the data as possible.

$$\begin{aligned}
L_{\text{contrast}} =&\mathbb{E}\left[-\log\frac{e^{f_x^T f_y/\tau}}{e^{f_x^T f_y/\tau}+\sum_i e^{f_x^T f_{y_i^-}/\tau}}\right]\\
=&\mathbb{E}\left[-f_x^T f_y/\tau\right]+\mathbb{E}\left[\log\left(e^{f_x^T f_y/\tau}+\sum_i e^{f_z^T f_{y_i^-}/\tau}\right)\right]\\
\overset{\mathbf{P}[f_v=f_y)]=\mathbf{1}}{=}&\underbrace{\mathbb{E}\left[-f_x^T f_y/\tau\right]}_{\text{positive alignment}}+\underbrace{\mathbb{E}\left[\log\left(e^{1/\tau}+\sum_i e^{f_x^T f_{y_i^-}/\tau}\right)\right]}_{\text{uniformity}}
\end{aligned} \tag{4}$$

The alignment loss can be defined straightforwardly as follows:

$$\mathcal{L}_{\text{align}}\left(f;\alpha\right)\triangleq-\underset{(x,y)\sim p_{\text{pos}}}{\mathbb{E}}\left[\|f(x)-f(y)\|_2^{\alpha}\right],\quad\alpha>0 \tag{5}$$

Where $f(.)$ is the feature encoder and x,y are positive instance pairs. The uniformity metric refers to optimizing this metric should converge to a uniform distribution. Note that feature uniformity should be empirically reasonable with a finite number of points and asymptotically correct. Therefore, the loss can be defined with the radial basis function (RBF) kernel $G_t:\mathcal{S}^d\times\mathcal{S}^d\rightarrow\mathbb{R}_+$ Wang & Isola

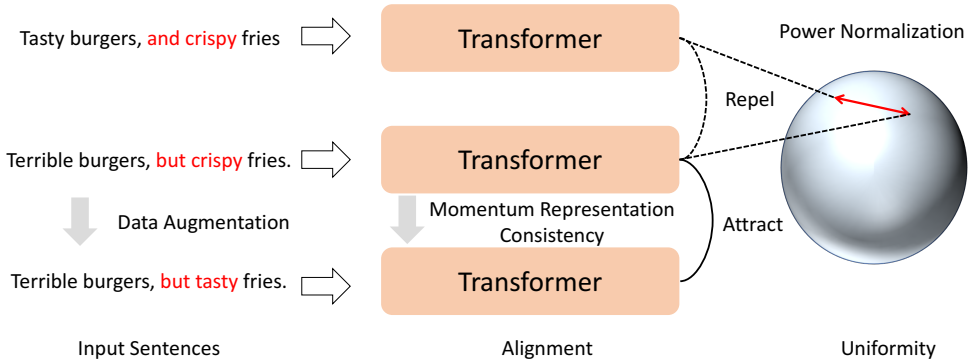


Figure 1: Disentangled contrastive learning for robust textual representations.

(2020). Formally, we have:

$$\begin{aligned} L_{uniform}(f;t) &\triangleq \log \mathop{\mathbb{E}}_{x,y \overset{\text{ii.d.}}{\mathbb{E}}} [G_t(u,v)] \\ &= \log \mathop{\mathbb{E}}_{x,y \overset{\text{id.d.}}{p_{\text{data}}}} \left[e^{-t\|f(x)-f(y)\|_2^2} \right], \quad t > 0 \end{aligned} \tag{6}$$

where t is a fixed parameter.

4 Disentangled Contrastive Learning

In this section, we present a preliminary study on how to learn robust textual representation via disentangled contrastive learning, as represented in Figure 1. Because the aforementioned analysis shows that contrastive learning can be disentangled with feature alignment and uniformity, it is intuitive to optimize the representation learning method with separated objects, thereby learning without negative textual instances.

4.1 Feature Alignment with Momentum Representation Consistency

There are multiple ways to align a textual representation. We utilize two transformers with a consistent momentum representation to explicitly guarantee feature alignment Grill et al. (2020). The two networks are defined by a set of weights θ and ξ. We use the exponential moving average of the parameters θ to get ξ. Formally, we have:

$$\xi \leftarrow \tau\xi + (1-\tau)\theta \tag{7}$$

Given a sentence $\mathcal{X}$ and its augmentation $\mathcal{X}'$ (e.g, via data augmentation) from the first original network, we may obtain output representations $q \triangleq f_\theta(X)$ and $p \triangleq f_\theta(X')$. Note that previous works Chen et al. (2020b); Grill et al. (2020) indicates that an projection p in feature space improve the performance. We then leverage a projection function $g(p_\theta)$ and ℓ_2-normalize both $g(p_\theta)$ and q_ξ to $\bar{g}(p_\theta) \triangleq g/\|g(p_\theta)\|_2$ and $\bar{q}_\xi \triangleq q_\xi/\|q_\xi\|_2$, respectively. We leverage the mean squared loss as follows:

$$\mathcal{L}_{\text{align}} \triangleq \|\overline{g}\left(q\right) - \bar{p}_\xi\|_2^2 = 2 - 2 \cdot \frac{\langle g\left(q_\theta\right), p_\xi \rangle}{\|g\left(q_\theta\right)\|_2 \cdot \|p_\xi\|_2} \tag{8}$$

Additionally, we make the losses symmetrical $\mathcal{L}_{\text{align}}$ by feeding X to the augmented network and X', separately. We optimize $\mathcal{L}_{\text{align}} + \widetilde{\mathcal{L}}_{\text{align}}$ with respect to θ only, but *not* ξ, via the stop-gradient.

4.2 FEATURE UNIFORMITY WITH POWER NORMALIZATION

To ensure that feature vectors should have an approximately uniform distribution, we can directly optimize the Eq. 6. However, different from computer vision, in the original loss of BRET Devlin et al. (2019a), we have already utilized the next sentence prediction loss. Such a contrastive object has explicitly made the sentence representation $f(.)$ scattered in the feature space; thus, the model may quickly collapse without learning. Inspired by Santurkar et al. (2018), we argue that batch normalization can identify the common-mode between examples of a mini-batch and removes it using the other representations in the mini-batch as implicit negative examples. We can, therefore, view batch normalization as a novel method of implementing feature uniformity on embedded representations. Because vanilla batch normalization will lead to significant performance degradation when naively used in NLP, we leverage an enhanced power normalization Shen et al. (2020) to guarantee feature uniformity. Specifically, we leverage the unit quadratic mean rather than the mean/variance of running statistics with an approximate backpropagation method to compute the corresponding gradient. Formally, we have the following:

$$\begin{aligned} \widehat{\boldsymbol{X}}^{(t)} &= \frac{\boldsymbol{X}^{(t)}}{\psi^{(t-1)}} \\ \boldsymbol{Y}^{(t)} &= \gamma \odot \widehat{\boldsymbol{X}}^{(t)} + \beta \\ \left(\psi^{(t)}\right)^2 &= \left(\psi^{(t-1)}\right)^2 + (1-\alpha)\left(\psi_B^2 - \left(\psi^{(t-1)}\right)^2\right) \end{aligned} \tag{9}$$

Note that we compute the gradient of the loss regarding the quadratic mean of the batch. In other words, we utilize the running statistics to conduct backpropagation, thus, resulting in bounded gradients, which is necessary for convergence in NLP (see proofs in Shen et al. (2020)).

4.3 IMPLEMENTATION DETAILS

We leverage synonyms from WordNet categories to conduct data augmentation for computation efficiency. We combine all the momentum representation consistency and power normalization results in a unified architecture with the mask language model object. We leverage the same architecture of the BERT-base Devlin et al. (2019a). We first pre-train the model in a large-scale corpus unsupervisedly (e.g., the same corpus and training steps with BERT) and then fine-tune the model using task datasets.

5 EXPERIMENT

We evaluated our method using NLP benchmarks, including tasks of text classification, natural language inference, machine reading comprehension, and the GLUE series of language understanding tasks. We conduct experiments on the normal test set as well as robust settings (e.g., invariance tests and adversarial attacks). The code and datasets are available at `anonymous`.

5.1 DATASETS AND SETTING

We conducted experiments on three benchmarks: GLUE (Wang et al., 2019), SQuAD(Rajpurkar et al., 2016a), and SNLI Bowman et al. (2015).

GLUE Wang et al. (2019) is an NLP benchmark aimed at evaluating the performance of downstream tasks of the pre-trained models. Notably, we leverage nine tasks in GLUE, including CoLA, RTE, MRPC, STS, SST, QNLI, QQP, and MNLI-m/mm. We follow the same setup as the original BERT for single sentence and sentence pair classification tasks. We leverage a multi-layer perception with a softmax layer to obtain the predictions.

SQuAD is a reading comprehension dataset constructed from Wikipedia articles. We report results on SQuAD 1.1. Here also, we follow the same setup as the original BERT model and predict an answer span—the start and end indices of the correct answer in the correct context.

SNLI is a collection of 570k human-written English sentence pairs that have been manually labeled for balanced classification with entailment, contradiction, and neutral labels, thereby supporting the

task of natural language inference (NLI). We add a linear transformation and a softmax layer to predict the correct label of NLI.

To evaluate the robustness of our approach, we also conduct invariance testing with CheckList[1] Ribeiro et al. (2020) and adversarial attacks[2]. To generate label-preserving perturbations, we used WordNet categories (e.g., synonyms and antonyms). We selected context-appropriate synonyms as permutation candidates. To generate adversarial samples, we leverage a probability-weighted word saliency (PWWS) Ren et al. (2019) method based on synonym replacement. We manually evaluate the quality of the generated instances. We also conduct experiments that apply data augmentation and adversarial training to the BERT model.

We utilize PyTorch Paszke et al. (2019) to implement our model. We use Adam optimizer with a cosine decay learning rate schedule. We set the initial learning rate as 1e-5. We use a batch size of 32 over eight Nvidia 1080Ti GPUs. With this setup, training takes approximately one month. We leverage the grid search to find optimal hyper-parameters in the development set. We ran each experiment five times and calculated the average performance.

5.2 Results and Analysis

Table 1: Summary of results on GLUE.

Model		CoLA	SST-2	MRPC	QQP	MNLI (m/mm)	QNLI	RTE	GLUE Avg
Normal	BERT	56.8	92.3	89.7	89.6	84.6/85.2	91.5	69.3	82.3
	BERT+DA	58.6	93.2	86.5	86.7	84.2/84.4	91.1	68.9	81.7
	DCL	**60.9**	93.0	**89.7**	**90.0**	84.7/84.6	**91.7**	**69.7**	**83.0**
Robust	BERT	46.4	91.8	88.1	84.9	81.6/82.2	89.2	67.1	78.9
	BERT+DA	53.8	92.9	85.6	85.5	83.1/83.4	90.7	66.3	80.1
	DCL	48.4	**92.4**	86.0	**85.5**	82.5/82.7	89.7	**68.8**	79.5

Main Results

From Table 1 and 2, we can observe the following: 1) Vanilla BERT achieves poor performance in the robust set on both GLUE and SQUAD, which indicates that the previous fine-tuning approach cannot obtain a robust textual representation. This will lead to performance decay with permutations. 2) With data augmentation, BERT can obtain improved performance in the robust set; however, a slight performance decay is observed in the original test set. We argue that data augmentation can obtain better performance by fitting to task-specific data distribution; there is no guarantee that more data will result in robust textual representations. 3) Our DCL approach achieves improved performance in both the original test set and robust set compared with vanilla BERT. Note that our DCL is an unsupervised approach, and we leverage the same training instances with BERT. The performance improvements indicate that our approach can obtain more robust textual representations that enhance the performance of the system.

Table 2: Summary of results on SQuAD 1.1.

Model		F1	EM
Normal	BERT	88.5	80.8
	BERT+DA	88.2	80.4
	DCL	**88.4**	**81.0**
Robust	BERT	86.7	77.8
	BERT+DA	87.8	79.9
	DCL	86.8	78.1

Adversarial Attack Results

From Table 3, we can observe the following: 1) Vanilla BERT achieves a poor performance with adversarial attacks; BERT with adversarial training can obtain a good performance. However, we notice that there exists a performance decay for adversarial training in the original test set. Note that adversarial training methods would lead to standard performance degradation Wen et al. (2019), i.e., the degradation of natural examples. 2) Our DCL approach achieves improved performance in the

[1] https://github.com/marcotcr/checklist.git

[2] https://github.com/thunlp/OpenAttack

Table 3: Summary of results on CoLA and SNLI.

Model		CoLA	SNLI
NORMAL	BERT	56.8	91.0
	BERT+Adv	55.0	91.1
	DCL	**59.0**	91.0
ADVERSARIAL	BERT	47.7	90.1
	BERT+Adv	55.1	91.1
	DCL	48.9	90.5

test set with and without an adversarial attack, which further demonstrates that our approach can obtain robust textual representations that are stable for different types of permutations.

Quantitative Analysis of Textual Representation

As we hypothesize that power normalization can implicitly contribute to feature uniformity, we conduct further experiments to analyze the effects of normalization Abe & Josh. Specifically, we random sample instances and leverage the cosine similarity of the original input projection vectors and the augmented projection vectors. We calculate the average cosine similarity between positive instances (in blue) and random instances (in red) with different strategies, including without normalization (No Norm), batch normalization (BN), and power normalization (DCL).

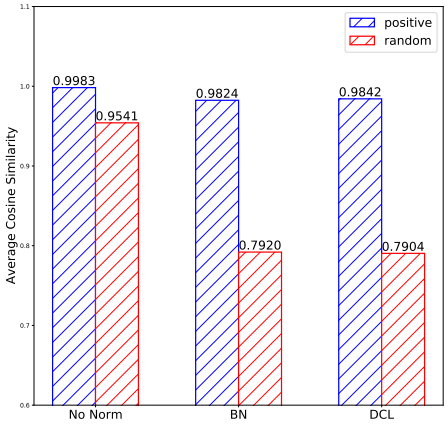


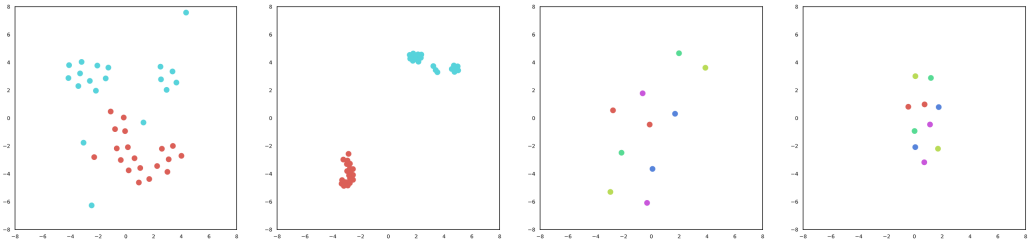
Figure 2: Cosine similarity of the original input projection vectors with the augmented input projection vectors.

From Figure 2, we observe that with no normalization in p or q, the feature space is aligned for both positive and negative instances, which shows that there exists a feature collapse for textual representation learning. Considering DCL training (i.e., with power normalization), we notice that the textual representations are relatively more similar between the positive instances (0.9842) than random (negative) ones (0.7904); thus, we can obtain different vectors.

Next, we give an intuitive explanation of preventing feature collapse for textual representation learning. Given an input instance without negative examples, the model may always output the projection vector z with $[0, 1, 0, 0, \dots]$. Thus, the model can achieve a perfect prediction through learning a simple identity function, which, in other words, collapse in the feature space. With normalization, the output vector z cannot obtain such singular values. Since the outputs will be redistributed regarding the learned mean and standard deviation, we can implicitly learn robust textual representations.

Qualitative Analysis of Textual Representation

We randomly selected instances to visualize a sentence with T-SNE Maaten & Hinton (2008) to better understand the behaviors of textual representations. The different color refers to the different sentence pairs for both random permutation and adversarial attack settings. From Figure 3, it may be observed that our approach can obtain a relatively similar semantic representation with permutations in both invariant tests and adversarial attack settings. Note that we explicitly align the projection of the textual representation with a random permutation, thereby encouraging similar semantic instances to have relatively similar representations.



(a) BERT(Random) (b) DCL(Random) (c) BERT(Adv) (d) DCL(Adv)

Figure 3: T-SNE visualizations of sentence embeddings.

5.3 Discussion

Robust Representation with Contrastive Learning

Conventional approaches usually try to leverage instance-level augmentation aimed at achieving good performance on a robust set. However, there is no guarantee that robust textual representations will be obtained. Intuitively, directly aligning the representation of input tokens with slight permutations may contribute to robust representations. However, without any negative constraints, the model will easily collapse with a sub-optimal solution. In this study, we observe that power normalization identifies this common-mode between examples. In other words, it can remove those trivial samples by using the other representations in the batch as implicit negative instances. We can, therefore, view normalization as an implicitly contrastive learning method.

Limitations

This work is not without limitations. We only consider the synonym replacement as a data augmentation strategy due to the efficiency of processing a huge amount of data. Other strong data augmentation methods can also be leveraged. Another issue is representation alignment, as there are lots of augmentations. We cannot enumerate all positive pairs for alignments; thus, there is still some room for designing more efficient feature aligning algorithms. Moreover, as we utilize the square root loss, which is absolutely a Euclidean distance. Recent approaches Meng et al. (2019; 2020) indicates that Euclidean space may be sub-optimal for textual representations. Lastly, with power normalization, the network outputs are no longer learning a pure function of the corresponding inputs. Thus, it may be interesting to develop methods avoiding the use of power normalization during training. Moreover, it may be promising to investigate alternative methods, such as weight standardization with group normalization for textual representation learning.

6 Conclusions and Future Work

We investigated robust textual representation learning and proposed a disentangled contrastive learning approach. We introduced feature alignment with a momentum representation consistency and feature uniformity with power normalization. We empirically observed that our approach could obtain an improved performance compared with baselines in NLP benchmarks and achieve a robust performance with invariant tests and adversarial attacks. We also performed quantitative and qualitative analyses for learned textual representations, which indicated that our approach mitigates model collapse and can learn robust textual representations. This may provide a basis for future works concerning robust representation learning. Our approach is model-agnostic; therefore, it can be applied to any pre-trained language models.

Further research on robust textual representation learning may be conducted to investigate such topics as: 1) exploiting multi-task learning for robust representations; 2) investigating the essence of model robustness and proposing more efficient approaches to learn robust representations; and 3) incorporating more complex views (e.g., higher-order or skip n-grams, syntactic and semantic parses, etc.) and designing appropriate self-supervised tasks.

REFERENCES

Fetterman Abe and Albrecht Josh. Understanding self-supervised and contrastive learning with "bootstrap your own latent" (byol). https://untitled-ai.github.io/understanding-self-supervised-contrastive-learning.html.

Sanjeev Arora, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. A theoretical analysis of contrastive unsupervised representation learning. *arXiv preprint arXiv:1902.09229*, 2019.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. The snli corpus. 2015.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020a.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of Machine Learning and Systems 2020*, pp. 10719–10729, 2020b.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. Infoxlm: An information-theoretic framework for cross-lingual language model pre-training. *arXiv preprint arXiv:2007.07834*, 2020.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. ELECTRA: pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020*. OpenReview.net, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina 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, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019a. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://www.aclweb.org/anthology/N19-1423.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina 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, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019b. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://www.aclweb.org/anthology/N19-1423.

Hongchao Fang and Pengtao Xie. Cert: Contrastive self-supervised learning for language understanding. *arXiv preprint arXiv:2005.12766*, 2020.

John M Giorgi, Osvald Nitski, Gary D Bader, and Bo Wang. Declutr: Deep contrastive learning for unsupervised textual representations. *arXiv preprint arXiv:2006.03659*, 2020.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. *CoRR*, abs/1911.05722, 2019.

Dan Iter, Kelvin Guu, Larry Lansing, and Dan Jurafsky. Pretraining with contrastive sentence objectives improves discourse performance of language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4859–4870, Online, 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.acl-main.439.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. *arXiv*, pp. arXiv–1907, 2019.

Linyang Li and Xipeng Qiu. Textat: Adversarial training for natural language understanding with token-level perturbation. *arXiv preprint arXiv:2004.14543*, 2020.

Ting Liu, Andrew W Moore, Ke Yang, and Alexander G Gray. An investigation of practical approximate nearest neighbor algorithms. In *Advances in neural information processing systems*, pp. 825–832, 2005.

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.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

Yu Meng, Jiaxin Huang, Guangyuan Wang, Chao Zhang, Honglei Zhuang, Lance Kaplan, and Jiawei Han. Spherical text embedding. In *Advances in Neural Information Processing Systems*, pp. 8208–8217, 2019.

Yu Meng, Yunyi Zhang, Jiaxin Huang, Yu Zhang, Chao Zhang, and Jiawei Han. Hierarchical topic mining via joint spherical tree and text embedding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1908–1917, 2020.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*, pp. 3111–3119, 2013.

Andriy Mnih and Koray Kavukcuoglu. Learning word embeddings efficiently with noise-contrastive estimation. In *Advances in Neural Information Processing Systems 26*, pp. 2265–2273, 2013.

Nathan Ng, Kyunghyun Cho, and Marzyeh Ghassemi. Ssmba: Self-supervised manifold based data augmentation for improving out-of-domain robustness. *arXiv preprint arXiv:2009.10195*, 2020.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018. URL http://arxiv.org/abs/1807.03748.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems*, pp. 8026–8037, 2019.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, Austin, Texas, November 2016a. Association for Computational Linguistics. doi: 10.18653/v1/D16-1264. URL https://www.aclweb.org/anthology/D16-1264.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016b.

Shuhuai Ren, Yihe Deng, Kun He, and Wanxiang Che. Generating natural language adversarial examples through probability weighted word saliency. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 1085–1097, 2019.

Marco Túlio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with checklist. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pp. 4902–4912. Association for Computational Linguistics, 2020. URL https://www.aclweb.org/anthology/2020.acl-main.442/.

Shibani Santurkar, Dimitris Tsipras, Andrew Ilyas, and Aleksander Madry. How does batch normalization help optimization? In *Advances in Neural Information Processing Systems*, pp. 2483–2493, 2018.

Nikunj Saunshi, Orestis Plevrakis, Sanjeev Arora, Mikhail Khodak, and Hrishikesh Khandeparkar. A theoretical analysis of contrastive unsupervised representation learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 5628–5637, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL http://proceedings.mlr.press/v97/saunshi19a.html.

Sheng Shen, Zhewei Yao, Amir Gholami, Michael W Mahoney, and Kurt Keutzer. Powernorm: Rethinking batch normalization in transformers. In *In the proceedings of the International Conference on Machine Learning (ICML)*, 2020.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *CoRR*, abs/1906.05849, 2019. URL http://arxiv.org/abs/1906.05849.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=rJ4km2R5t7.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. *arXiv preprint arXiv:2005.10242*, 2020.

Jason Wei and Kai Zou. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*, 2019.

Xiangpeng Wei, Yue Hu, Rongxiang Weng, Luxi Xing, Heng Yu, and Weihua Luo. On learning universal representations across languages. *arXiv preprint arXiv:2007.15960*, 2020.

Yuxin Wen, Shuai Li, and Kui Jia. Towards understanding the regularization of adversarial robustness on neural networks. 2019.

Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018*, pp. 3733–3742. IEEE Computer Society, 2018.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. Approximate nearest neighbor negative contrastive learning for dense text retrieval. *arXiv preprint arXiv:2007.00808*, 2020.

Mang Ye, Xu Zhang, Pong C. Yuen, and Shih-Fu Chang. Unsupervised embedding learning via invariant and spreading instance feature. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019*, pp. 6210–6219. Computer Vision Foundation / IEEE, 2019.