Virtual Augmentation Supported Contrastive Learning of Sentence Representations

Dejiao Zhang∗ Wei Xiao Henghui Zhu Xiaofei Ma Andrew O. Arnold
AWS AI Labs, New York

Abstract

Despite profound successes, contrastive representation learning relies on carefully designed data augmentations using domain-specific knowledge. This challenge is magnified in natural language processing, where no general rules exist for data augmentation due to the discrete nature of natural language. We tackle this challenge by presenting a Virtual augmentation Supported Contrastive Learning of sentence representations (VaSCL). Originating from the interpretation that data augmentation essentially constructs the neighborhoods of each training instance, we, in turn, utilize the neighborhood to generate effective data augmentations. Leveraging the large training batch size of contrastive learning, we approximate the neighborhood of an instance via its K-nearest in-batch neighbors in the representation space. We then define an instance discrimination task regarding the neighborhood and generate the virtual augmentation in an adversarial training manner. We access the performance of VaSCL on a wide range of downstream tasks and set a new state-of-the-art for unsupervised sentence representation learning.

1 Introduction

Universal sentence representation learning has been a long-standing problem in Natural Language Processing (NLP). Leveraging the distributed word representations (Bengio et al., 2003; Mikolov et al., 2013; Collobert et al., 2011; Pennington et al., 2014) as the base features to produce sentence representations is a common strategy in the early stage. However, these approaches are tailored to different target tasks, thereby yielding less generic sentence representations (Yessenalina and Cardie, 2011; Socher et al., 2013; Kalchbrenner et al., 2014; Cho et al., 2014).

∗ The code and pretrained checkpoints can be found at https://github.com/amazon-research/sentence-representations. Correspondence to Dejiao Zhang <dejiaoz@amazon.com>.

This issue has motivated more research efforts on designing generic sentence-level learning objectives or tasks. Among them, supervised learning on the Natural Language Inference (NLI) datasets (Bowman et al., 2015a; Williams et al., 2017; Wang et al., 2018) has established benchmark transfer learning performance on various downstream tasks (Conneau et al., 2017; Cer et al., 2018; Reimers and Gurevych, 2019a; Zhang et al., 2021). Despite promising progress, the high cost of collecting annotations precludes its wide applicability, especially when the target domain has scarce annotations but differs significantly from the NLI datasets (Zhang et al., 2020).

On the other hand, unsupervised learning of sentence representations has seen a resurgence of interest with the recent successes in self-supervised contrastive learning. These approaches rely on two main components, data augmentation and an instance-level contrastive loss. The popular contrastive learning objectives Chen et al. (2020); He et al. (2020) and their variants thereof have empirically shown their effectiveness in NLP. However, the discrete nature of the text makes it challenging to establish universal rules for effective text augmentation generation.

Various contrastive learning based approaches have been proposed for sentence representation learning, where the main difference lies in how the augmentations are generated (Fang and Xie, 2020; Giorgi et al., 2020; Wu et al., 2020; Meng et al., 2021; Yan et al., 2021; Kim et al., 2021; Gao et al., 2021). Somewhat surprisingly, a recent work (Gao et al., 2021) shows that Dropout (Srivastava et al., 2014), i.e., augmentations obtained by feeding the same instance to the encoder twice, outperforms common data augmentations obtained by operating on the text directly, including cropping, word deletion, or synonym replacement. Again, this observation validates the inherent difficulty of attaining effective data augmentations in NLP.
This paper tackles the challenge by presenting a neighborhood-guided virtual augmentation strategy to support contrastive learning. In a nutshell, data augmentation essentially constructs the neighborhoods of each instance, with the semantic content being preserved. We take this interpretation in the opposite direction by leveraging the neighborhood of an instance to guide augmentation generation. Benefiting from the large training batch of contrastive learning, we approximate the neighborhood of an instance via its K-nearest in-batch neighbors. We then define an instance discrimination task within this neighborhood and generate the virtual augmentation in an adversarial training manner. We run in-depth analyses and show that our VaSCL model leads to a more dispersed representation space with the data semantics at different granularities being better captured. We evaluate our model on a wide range of downstream tasks and show that our model consistently outperforms the previous state-of-the-art results by a large margin.

2 Related Work

Universal Sentence Representation Learning
Arguably, the simplest and most common approaches for attaining sentence representations are bag-of-words (Harris, 1954) and variants thereof. However, bag-of-words suffers from data sparsity and a lack of sensibility to word semantics. In the past two decades, the distributed word representations (Bengio et al., 2003; Mikolov et al., 2013; Collobert et al., 2011; Pennington et al., 2014) have become the more effective base features for producing sentence representations. The downside is that these approaches are tailored to the target tasks (Yessenalina and Cardie, 2011; Socher et al., 2013; Kalchbrenner et al., 2014; Cho et al., 2014), and thereby the resulting sentence representations attain limited transfer learning performance.

More recent efforts focus on directly designing the sentence-level learning objectives or tasks. On the supervised learning regime, Conneau et al. (2017); Cer et al. (2018) empirically show the effectiveness of leveraging the NLI task (Bowman et al., 2015a; Williams et al., 2017) to promote generic sentence representations. The task involves classifying each sentence pair into one of three categories: entailment, contradiction, or neutral. Reimers and Gurevych (2019b) further bolster the performance by using the pre-trained transformer (Devlin et al., 2018; Liu et al., 2019) as backbone. On the other end of the spectrum, Hill et al. (2016); Bowman et al. (2015b) propose using the denoising or variational autoencoders for sentence representation learning. Kiros et al. (2015); Hill et al. (2016) extend the distributional hypothesis to the sentence level and train an encoder-decoder to construct the surrounding context for each sentence. Alternatively, Logeswaran and Lee (2018) present a model that learns to discriminate the target context sentences from all contrastive ones.

Contrastive Learning
Contrastive learning has been the pinnacle of recent successes in sentence representation learning. Gao et al. (2021); Zhang et al. (2021) substantially advance the previous state-of-the-art results by leveraging the entailment sentences in NLI as positive pairs for optimizing the properly designed contrastive loss functions. Nevertheless, we focus on unsupervised contrastive learning and form the positive pairs via data augmentation since such methods are more cost-effective and applicable across different domains and languages. Along this line, several approaches have been proposed recently, where the augmentations are obtained via dropout (Yan et al., 2021; Gao et al., 2021), back-translation (Fang and Xie, 2020), surrounding context sampling (Logeswaran and Lee, 2018; Giorgi et al., 2020), or perturbations conducted at different semantic-level (Wu et al., 2020; Yan et al., 2021; Meng et al., 2021).

Consistency Regularization
Our work is also closely related to consistency regularization, which is often used to promote better performance by regularizing the model output to remain unchanged under plausible input variations that are often induced via data augmentations. Bachman et al. (2014); Sajjadi et al. (2016); Samuli and Timo (2017); Tarvainen and Valpola (2017) show randomized data augmentations such as dropout, cropping, rotation, and flipping yield effective regularization. Berthelot et al. (2019, 2020); Verma et al. (2019) improve the performance by applying Mixup (Zhang et al., 2017) and its variants on top of stochastic data augmentations. However, data augmentation has long been a challenge in NLP as there are no general rules for effective text transformations. An alternative that comes to light when considering the violation of consistency regularization can in turn be used to find the most sensitive perturbation for a model. Therefore, we utilize consistency regularization to promote informative virtual augmen-
3 Method

3.1 Preliminaries

Self-supervised contrastive learning often aims to solve the instance discrimination task. In our scenario, let $f$ denote the transformer encoder that maps the $i$-th input sentence $x_i$ to its representation vector $e_i = f(x_i)$. Further let $h$ be the contrastive learning head and $z_i = h(f(x_i))$ denote the final output for $x_i$. Let $B = \{i, i'\}_{i=1}^M$ denote the indices of a randomly sampled batch of paired examples, where $x_i, x_{i'}$ are two independent variations of the $i$-th instance. A popular loss function (Chen et al., 2020) for contrastive learning is defined as follows,

$$\ell_B(z_i, z_{i'}) = \frac{1}{\tau} \log \frac{e^{\text{sim}(z_i, z_{i'})}}{e^{\text{sim}(z_i, z_{i'})} + \sum_{j \in B \setminus \{i, i'\}} e^{\text{sim}(z_i, z_j)}} \; ,$$

where $\tau$ is the temperature hyper-parameter and $\text{sim}(\cdot, \cdot)$ denotes the cosine similarity, i.e., $\text{sim}(\cdot, \cdot) = z_i^T z_{i'}/\|z_i\|_2 \|z_{i'}\|_2$. Similarly, $\ell_B(z_{i'}, z_i)$ is defined by exchanging the roles of $z_i$ and $z_{i'}$ in the above equation. Intuitively, Equation (1) defines the log-likelihood of classifying the $i$-th instance as its positive $i'$ among all $2M-1$ candidates within

---

Figure 1: Illustration of VaSCL. For each instance $x_i$, in a randomly sampled batch, we optimize (i) an instance-wise contrastive loss with the dropout induced augmentation obtained by forwarding the same instance twice, i.e., $x_i$ and $x_{i'}$ denote the same text example; and (2) a neighborhood constrained instance discrimination loss with virtual augmentation (see Section 3.2).
of a randomly sampled batch with \( M \) examples. We first approximate the neighborhood \( \mathcal{N}(i) \) of the \( i \)th instance as its K-nearest neighbors in the representation space,
\[
\mathcal{N}(i) = \{ k : \mathbf{e}_k \text{ has the top-K similarity with } \mathbf{e}_i \text{ among all other M-1 instances in } \mathcal{B} \}
\]

We then define an instance-level contrastive loss regarding the \( i \)th instance and its neighborhood as follows,
\[
\ell_{\mathcal{N}(i)}(\mathbf{z}_i^\delta, \mathbf{z}_i) = -\log \frac{e^{\text{sim}(\mathbf{z}_i^\delta, \mathbf{z}_i)/\tau}}{\sum_{k \in \mathcal{N}(i)} e^{\text{sim}(\mathbf{z}_i^\delta, \mathbf{z}_k)/\tau}}.
\]

Putting it all together Therefore, for each randomly sampled minibatch \( \mathcal{B} \) with \( M \) samples, we minimize the following:
\[
\mathcal{L}_{\text{VaSCL}} = \frac{1}{2M} \sum_{i=1}^{M} \{ \ell_g(\mathbf{z}_i, \mathbf{z}_i') + \ell_g(\mathbf{z}_i', \mathbf{z}_i) + \ell_{\mathcal{N}_A(i)}(\mathbf{z}_i, \mathbf{z}_i^\ast) + \ell_{\mathcal{N}_A(i)}(\mathbf{z}_i^\ast, \mathbf{z}_i) \}
\]

The last two terms of the right hand side are defined in Equation 4. Notice that, \( \ell_g(\mathbf{z}_i, \mathbf{z}_i') \) is defined in the same way as Equation (1) except that \( \mathbf{z}_i, \mathbf{z}_i' \) are retained by feeding the \( i \)th instance in \( \mathcal{B} \) to the encoder twice. In summary, two instance discrimination tasks are posed for each training example: i) discriminating each instance and its dropout induced variation from the other in-batch instances; and ii) separating each instance and its virtual augmentation from its K nearest neighbors and their associated virtual augmentations.

4 Experiment

In this section, we mainly evaluate VaSCL against SimCSE (Gao et al., 2021) which leverages the dropout (Srivastava et al., 2014) induced noise as data augmentation. We show that VaSCL consistently outperforms SimCSE on various downstream tasks that involve semantic understanding at different granularities. We carefully study the regularization effects of VaSCL and empirically demonstrate that VaSCL leads to a more dispersed representation space with semantic structure better encoded. Please refer to Appendix A for details of our implementations and the dataset being used.

4.1 Evaluation Datasets

In addition to the popular semantic textual similarity (a.k.a STS) related tasks, we evaluate two additional downstream tasks, short text clustering and few-shot learning based intent classification. Our motivation is twofold. First, these two tasks provide a new evaluation aspect that complements the pairwise similarity-oriented STS evaluation by assessing the high-level categorical semantics encoded in the representations. Second, two desired challenges are posted as short text clustering requires more effective representations due to the weak signal each text example manifests; and intent classification often suffers from data scarcity since the intents can vary significantly over different dialogue systems and the intent examples are costly to collect.
<table>
<thead>
<tr>
<th>Model</th>
<th>STS12</th>
<th>STS13</th>
<th>STS14</th>
<th>STS15</th>
<th>STS16</th>
<th>SICK-R</th>
<th>STS-B</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>RoBERTa_{distil}</td>
<td>54.41</td>
<td>46.85</td>
<td>56.96</td>
<td>65.79</td>
<td>64.22</td>
<td>61.10</td>
<td>59.01</td>
<td>58.33</td>
</tr>
<tr>
<td>SimCSE_{distil}</td>
<td>65.58</td>
<td>77.42</td>
<td>70.17</td>
<td>79.31</td>
<td>78.45</td>
<td>67.66</td>
<td>77.98</td>
<td>73.79</td>
</tr>
<tr>
<td>VaSCL_{distil}</td>
<td>67.68</td>
<td>80.61</td>
<td>72.19</td>
<td>80.92</td>
<td>78.59</td>
<td>68.81</td>
<td>77.32</td>
<td>75.16</td>
</tr>
<tr>
<td>RoBERTa_{base}</td>
<td>53.95</td>
<td>47.42</td>
<td>55.87</td>
<td>64.73</td>
<td>63.55</td>
<td>62.94</td>
<td>58.40</td>
<td>58.12</td>
</tr>
<tr>
<td>SimCSE_{base}</td>
<td>68.88</td>
<td>80.46</td>
<td>73.54</td>
<td>80.98</td>
<td>80.68</td>
<td>69.54</td>
<td>80.29</td>
<td>76.34</td>
</tr>
<tr>
<td>VaSCL_{base}</td>
<td>69.02</td>
<td>82.38</td>
<td>73.93</td>
<td>82.54</td>
<td>80.96</td>
<td>69.40</td>
<td>80.52</td>
<td>76.96</td>
</tr>
<tr>
<td>RoBERTa_{large}</td>
<td>55.00</td>
<td>50.14</td>
<td>54.87</td>
<td>62.14</td>
<td>62.99</td>
<td>58.93</td>
<td>54.56</td>
<td>56.95</td>
</tr>
<tr>
<td>SimCSE_{large}</td>
<td>69.83</td>
<td>81.29</td>
<td>74.42</td>
<td>83.77</td>
<td>79.79</td>
<td>68.89</td>
<td>80.66</td>
<td>76.95</td>
</tr>
<tr>
<td>VaSCL_{large}</td>
<td>73.36</td>
<td>83.55</td>
<td>77.16</td>
<td>83.25</td>
<td>80.66</td>
<td>72.96</td>
<td>82.36</td>
<td>79.04</td>
</tr>
</tbody>
</table>

Table 1: Spearman rank correlation between the cosine similarity of sentence representation pairs and the ground truth similarity scores.

Semantic Textual Similarity  The semantic textual similarity (STS) tasks are the most commonly used benchmark for evaluating sentence representations. STS consists of seven tasks, namely STS 2012-2016 (Agirre et al., 2012, 2013, 2014, 2015, 2016), the STS Benchmark (Cer et al., 2017), and the SICK-Relatedness (Marelli et al., 2014). For each sentence pair in these datasets, a fine-grained similarity score ranges from 0 to 5 is provided.

Short Text Clustering  Compared with general text clustering, short text clustering has its own challenge due to lack of signal. Nevertheless, texts containing only a few words grow at unprecedented rates from a wide range of popular resources, including Reddit, Stackoverflow, Twitter, and Instagram. Clustering those texts into groups of similar texts plays a crucial role in many real-world applications such as topic discovery (Kim et al., 2013), trend detection (Mathioudakis and Koudas, 2010), and recommendation (Bouras and Tsogkas, 2017). We evaluate six benchmark datasets for short text clustering. As shown in Table 4, the datasets present the desired diversities regarding both the cluster sizes and the number of clusters contained in each dataset.

Intent Classification  Intent classification aims to identify the intents of user utterances, which is a critical component of goal-oriented dialog systems. Attaining high intent classification accuracy is an important step towards solving many downstream tasks such as dialogue state tracking (Wu et al., 2019; Zhang et al., 2019) and dialogue management (Gao et al., 2018; Ham et al., 2020). A practical challenge is data scarcity because different systems define different sets of intents, and it is costly to obtain enough utterance samples for each intent. Therefore, few-shot learning has attracted much attention under this scenario, which is also our main focus. We evaluate four intent classification datasets originating from different domains. We summarize the data statistics in Appendix B.1.

4.2 Main Results

4.2.1 Evaluation Setup

Semantic Textual Similarity. Same as Reimers and Gurevych (2019b); Gao et al. (2021), in Table 1 we report the Spearman correlation\(^3\) between the cosine similarity of the sentence representation pairs and the ground truth similarity scores.

Short Text Clustering. We evaluate the sentence representations using K-Means (MacQueen et al., 1967; Lloyd, 1982) given its simplicity and report the clustering accuracy\(^4\) averaged over ten independent runs in Table 2.

Intent Classification. We freeze the transformer and fine-tune a linear classification layer with the softmax-based cross-entropy loss. We merge the training and validation sets, from which we sample K training and validation samples per class. We report the mean and standard deviation of the testing classification accuracy evaluated over five different splits in Table 3.\(^5\) We set the learning rate to 1e-04 and batch size to 32. For each task, we train the model with 1000 iterations.

\(^3\)Same as Reimers and Gurevych (2019b); Gao et al. (2021), we concatenate all the topics and report the overall Spearman’s correlation.

\(^4\)The clustering accuracy is computed by using the Hungarian algorithm (Munkres, 1957).

\(^5\)In each setting, we fix the five different splits for all models.
We report the evaluation results in Tables 1, 2, and 3. We report the testing accuracy on the checkpoint achieving the best validation accuracy.

### 4.2.2 Evaluation Results

We report the evaluation results in Tables 1, 2, and 3. As we can see, both SimCSE and VaSCL largely improve the performance of the pre-trained language models, while VaSCL consistently outperforms SimCSE on most tasks. To be more specific, we attain 0.6% – 2.1% averaged absolute improvement over SimCSE on seven STS tasks and 1.8% – 6.9% averaged absolute improvement on six short text clustering tasks. We also achieved considerable improvement over SimCSE on intent classification tasks under different few-shot learning scenarios. We do not include the evaluation on ATIS in Table 3 as this dataset is highly imbalanced with one single class account for more than 73% of the data. Please refer to Appendix C for details.

### 4.3 Analysis

To better understand what enables the good performance of VaSCL, we carefully analyze the representations at different semantic granularities.

#### Neighborhood Evaluation on Categorical Data

We first evaluate the neighborhood statistics on StackOverflow (Xu et al., 2017) which contains 20 balanced categories, each with 1000 text instances. For each instance, we retrieve its K nearest (top-K) neighbors in the representation space, among which those from the same class as the instance itself as treated as positives. In Figure 2a, we report both the percentage of true positives and the average distance of an instance to its top-K neighbors. For each top-K value, the evaluation is averaged over all 20,000 instances.

As indicated by the small distance values reported in Figure 2a, the representation space of the original RoBERTa model is tighter and is incapable of uncovering the categorical structure of data. In contrast, both VaSCL and SimCSE are capable of scattering representations apart while better capturing the semantic structures. Compared with SimCSE, VaSCL leads to even more dispersed representations with categorical structures being better encoded. This is also demonstrated by the better performance attained on both clustering and few-shot learning reported in Tables 2&3.

#### Fine-grained Semantic Understanding

We then compare VaSCL against SimCSE and RoBERTa on encoding more fine-grained semantic concepts. We randomly sample 20,000 premises from the combined set of SNLI (Bowman et al., 2015a) and MNLI (Williams et al., 2017), where

<table>
<thead>
<tr>
<th>Model</th>
<th>Ag News</th>
<th>Search Snippets</th>
<th>Stack Overflow</th>
<th>Biomedical</th>
<th>Tweet</th>
<th>Google News</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>RoBERTa.distil</td>
<td>59.32</td>
<td>33.18</td>
<td>14.16</td>
<td>24.69</td>
<td>37.10</td>
<td>58.05</td>
<td>37.75</td>
</tr>
<tr>
<td>SimCSE.distil</td>
<td>73.33</td>
<td>60.74</td>
<td>66.97</td>
<td>35.69</td>
<td>50.68</td>
<td>67.55</td>
<td>59.16</td>
</tr>
<tr>
<td>VaSCL.distil</td>
<td>71.71</td>
<td>62.76</td>
<td>73.98</td>
<td>38.82</td>
<td>51.35</td>
<td>67.66</td>
<td>61.05</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>5-Shot</th>
<th>10-Shot</th>
<th>20-Shot</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNIPS</td>
<td>BANK77</td>
<td>CLINIC150</td>
<td>HWU64</td>
</tr>
<tr>
<td>RoBERTa</td>
<td>76.71</td>
<td>71.4</td>
<td>67.86</td>
</tr>
<tr>
<td>SimCSE</td>
<td>79.51</td>
<td>72.2</td>
<td>68.24</td>
</tr>
<tr>
<td>VaSCL</td>
<td>83.51</td>
<td>76.5</td>
<td>71.71</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>5-Shot</th>
<th>10-Shot</th>
<th>20-Shot</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNIPS</td>
<td>BANK77</td>
<td>CLINIC150</td>
<td>HWU64</td>
</tr>
<tr>
<td>RoBERTa</td>
<td>58.71</td>
<td>53.6</td>
<td>49.56</td>
</tr>
<tr>
<td>SimCSE</td>
<td>62.93</td>
<td>51.55</td>
<td>45.11</td>
</tr>
<tr>
<td>VaSCL</td>
<td>66.09</td>
<td>61.57</td>
<td>69.04</td>
</tr>
</tbody>
</table>

Table 2: Clustering accuracy reported on six short text clustering datasets.

Table 3: Few-shot learning evaluation of Intent Classification. Each result is aggregated over 5 independent splits. We choose RoBERTa-base as backbone.
The performance of explicit text augmentation is evaluated using the standard dropout for training, i.e., "SimCSE w/ {WDel/WNet/CTxt})" in Figure 3. As Figure 3a shows, contrastive learning with moderate explicit text augmentations, i.e., augmentation strength less than 20%, does yield better sentence representations when compared with the original RoBERTa model. Nevertheless, both virtual augmentation strategies, i.e., SimCSE & VaSCL, substantially outperform all three explicit text augmentation strategies on almost all downstream tasks. Although a bit surprising, especially considering the performance gap between SimCSE and explicit augmentations, this comparison provides a new perspective on interpreting the underlying challenge of designing effective transformations that operate on the discrete text directly.

**Virtual Augmentation Performs Better** The performance of explicit text augmentation is evaluated using the standard dropout for training, i.e., "SimCSE w/ {WDel/WNet/CTxt})" in Figure 3. As Figure 3a shows, contrastive learning with moderate explicit text augmentations, i.e., augmentation strength less than 20%, does yield better sentence representations when compared with the original RoBERTa model. Nevertheless, both virtual augmentation strategies, i.e., SimCSE & VaSCL, substantially outperform all three explicit text augmentation strategies on almost all downstream tasks. Although a bit surprising, especially considering the performance gap between SimCSE and explicit augmentations, this comparison provides a new perspective on interpreting the underlying challenge of designing effective transformations that operate on the discrete text directly.

**VaSCL Outperforms SimCSE** Figure 3a also empirically demonstrates that VaSCL outperforms SimCSE no matter in the presence of explicit text augmentations or not. The only exception occurs when the explicit augmentation strength is too large, i.e., 20% of the words of each text are perturbed. One possible explanation is that undesired noises are generated by the large perturbations on discrete texts directly, which can violate the coherent semantics maintained by a neighborhood and hence make it hard for VaSCL to generate effective virtual augmentations.

**New Linguistic Patterns Are Required** Another observation drawn from Figure 3a is that both SimCSE and VaSCL attain worse performance on most downstream tasks when combining with explicit text augmentations. Although VaSCL does improve the performance of explicit augmentations
(a) Virtual augmentation vs. explicit augmentation. For each downstream task, we report the mean performance averaged over all its subtasks. The explicit augmentations are evaluated using SimCSE (dropout) for training, i.e., "SimCSE w/ [WDel/WNet/CTxt]".

(b) Cosine similarity between each original training example and its augmentation evaluated on the representation spaces of different models. From left to right, the augmentations are obtained via WDel, WNet, and CTxt. Each point is averaged over 20,000 randomly sampled training examples and their augmentations. We exclude "SimCSE w/ WNet" and "VaSCL w/ WNet" for better visualization. Please refer to Figure 4 in Appendix for the full plot.

Figure 3: Comparing and combining virtual augmentation with explicit augmentation.

... in most cases, this is undesired as we expect a win-win outcome that moderate explicit augmentations could further enhance VaSCL. We hypothesize that new and informative linguistic patterns are missing for the expected performance gain.

To validate our hypothesis, in Figure 3b we report the cosine similarity between each original training example and its augmentation evaluated on the representation spaces of different models. Our observation is twofold. First, the representations induced by RoBERTa and the one trained with contextual synonyms substitution ("SimCSE w/ CTxt") are very similar in all three settings, which also explains why "SimCSE w/ WDel" attains similar performance as RoBERTa on the downstream tasks. We attribute this to the fact that CTxt leverages the transformer itself to generate augmentations which hence carry limited unseen and effective linguistic patterns. Second, as indicated by the comparatively smaller similarity values in Figure 3b, the incorporation of explicit augmentations tightens the representation spaces of both SimCSE and VaSCL, which also results in a worse performance of downstream tasks. One possible explanation is that all the three explicit augmentations are weak and noisy, which harms both the instance discrimination force and the semantic relevance of each neighborhood.

5 Conclusion

In this paper, we present a virtual augmentation-oriented contrastive learning framework for unsupervised sentence representation learning. Our key insight is that data augmentation can be interpreted as constructing the neighborhoods of each training instance, which can, in turn, be leveraged to generate effective data augmentations. We evaluate VaSCL on a wide range of downstream tasks and substantially advance the state-of-the-art results. Moreover, we conduct in-depth analyses and show that VaSCL leads to a more dispersed representation space with the data semantics at different granularities being better encoded.

On the other hand, we observe a performance drop of both SimCSE and VaSCL when combined with the explicit text augmentations. We suspect this is caused by the linguistic patterns generated by explicit augmentations being less informative yet noisy. We hypothesize effective data augmentation operations on the discrete texts could complement our virtual augmentation approach if new and informative linguistic patterns are generated.
References


Yu Meng, Chenyan Xiong, Payal Bajaj, Saurabh Tiwary, Paul Bennett, Jiawei Han, and Xia Song. 2021. Coco-lm: Correcting and contrasting text sequences for language model pretraining. *arXiv preprint arXiv:2102.08473*.


A Implementation

Same as the original SimCSE work (Gao et al., 2021), we adopted $10^6$ randomly sampled sentences from English Wikipedia as training data. We implement our models with Pytorch (Paszke et al., 2017). We use the pre-trained RoBERTa models as the backbone. We choose a two-layer MLP with size $(d \times d, d \times 128)$ to optimize our contrastive learning losses, where $d$ denotes the dimension of the sentence representations. We use Adam (Kingma and Ba, 2015) as our optimizer with a constant learning rate of $5e-04$, which we scale to $5e-06$ for updating the backbones/transfomers. We set the virtual augmentation strength of VaSCL, i.e., $\Delta$ in Equation (3), to 15 for both DistilRoBERTa and RoBERTaBase, and 30 for RoBERTaLarge.

We train SimCSE (Gao et al., 2021) using $3e-05$ for optimizing the contrastive learning head and the backbone. We also tried the default learning rate $1e-05$ (suggested in Gao et al. (2021)) as well as our learning rate setup for optimizing the RoBERTa models with SimCSE. We found $3e-05$ yields better performance. For both SimCSE and VaSCL, we set the batch size to 1024, train all models over five epochs and evaluate the development set of STS-B every 500 iterations. We report all our evaluations on the downstream tasks with the associated checkpoints attaining the best performance on the validation set of STS-B.

B Dataset Statistics

B.1 Intent Classification Dataset

We evaluate our model on four intent classification datasets: (1) SNIPS (Coucke et al., 2018) is a SLU benchmark that consists of 7 distinct intents. (2) BANKING77 (Casanueva et al., 2020) is a large fine-grained single banking domain intent dataset with 77 intent classes. (3) HWU64 (Liu et al., 2021) contains 25,716 examples for 64 intents in 21 domains. (4) CLINC150 (Larson et al., 2019) spans 150 intents and 23,700 examples across 10 domains. As we can see here, SNIPS are limited to only a small number of classes, which oversimplifies the intent detection task and does not emulate the true environment of commercial systems. The remaining three datasets contain much more diversity and are more challenging.

B.2 Short Text Clustering Dataset

<table>
<thead>
<tr>
<th>Dataset</th>
<th>N</th>
<th>W</th>
<th>C</th>
<th>ImN</th>
</tr>
</thead>
<tbody>
<tr>
<td>AgNews</td>
<td>8.0K</td>
<td>23</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>SearchSnippets</td>
<td>13.3K</td>
<td>18</td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>StackOverflow</td>
<td>20K</td>
<td>8</td>
<td>20</td>
<td>1</td>
</tr>
<tr>
<td>Biomedical</td>
<td>20K</td>
<td>13</td>
<td>20</td>
<td>1</td>
</tr>
<tr>
<td>GoogleNews</td>
<td>11.1K</td>
<td>29</td>
<td>152</td>
<td>143</td>
</tr>
<tr>
<td>Tweet</td>
<td>2.5K</td>
<td>8</td>
<td>89</td>
<td>249</td>
</tr>
</tbody>
</table>

Table 4: Statistics of six short text clustering datasets. N: number of text samples; $W$: average number of words each text example has; C: number of clusters; ImN: imbalance number defined as the size of the largest class divided by that of the smallest class.

- **SearchSnippets** is extracted from web search snippets, which contains 12340 snippets associated with 8 groups Phan et al. (2008).
- **StackOverflow** is a subset of the challenge data published by Kaggle, where 20000 question titles associated with 20 different categories are selected by Xu et al. (2017).
(a) Evaluating VaSCL in presence of different explicit data augmentation strategies.

(b) Cosine similarity between the representations of each original training example and its augmentation evaluated on different models. From left to right, the augmentations are obtained via WDel, WNet, and CTxt. Each point is averaged over 20,000 randomly sampled training examples.

Figure 4: Comparing and combining virtual augmentation with explicit text augmentations. (Full plot of Figure 3 in Section 4.4.)

- Biomedical is a subset of PubMed data distributed by BioASQ\textsuperscript{9}, where 20000 paper titles from 20 groups are randomly selected by Xu et al. (2017).

- AgNews is a subset of news titles (Zhang and LeCun, 2015), which contains 4 topics selected by Rakib et al. (2020).

- Tweet consists of 89 categories with 2472 tweets in total (Yin and Wang, 2016).

- GoogleNews contains titles and snippets of 11109 news articles related to 152 events (Yin and Wang, 2016).

C Full Evaluation of Intent Classification

ATIS (Hemphill et al., 1990) is a benchmark for the air travel domain. This dataset is highly imbalanced, with the largest class containing 73\% of all the training and validation examples. Moreover, more than 60\% classes have less than 20 examples. We thereby exclude this task in our evaluation.

\textsuperscript{9}http://participants-area.bioasq.org