An Information Minimization Based Contrastive Learning Model for Unsupervised Sentence Embeddings Learning

Shaobin Chen[1], **Jie Zhou**[2], **Yuling Sun**[1*] and **Liang He**[1]
[1]School of Computer Science and Technology, East China Normal University Shanghai, China
[2]School of Computer Science, Fudan Univerisity Shanghai, China
shaobin_chen@stu.ecnu.edu.cn, jie_zhou@fudan.edu.cn, {ylsun, lhe}@cs.ecnu.edu.cn

Abstract

Unsupervised sentence embeddings learning has been recently dominated by contrastive learning methods (e.g., SimCSE), which keep positive pairs similar and push negative pairs apart. The contrast operation aims to keep as much information as possible by maximizing the mutual information between positive instances, which leads to redundant information in sentence embedding. To address this problem, we present an information minimization based contrastive learning (`InforMin-CL`) model to retain the useful information and discard the redundant information by maximizing the mutual information and minimizing the information entropy between positive instances meanwhile for unsupervised sentence representation learning. Specifically, we find that information minimization can be achieved by simple contrast and reconstruction objectives. The reconstruction operation reconstitutes the positive instance via the other positive instance to minimize the information entropy between positive instances. We evaluate our model on fourteen downstream tasks, including both supervised and unsupervised (semantic textual similarity) tasks. Extensive experimental results show that our `InforMin-CL` obtains a state-of-the-art performance. Code is made available. [1]

1 Introduction

How to learn universal sentence embeddings by large-scale pre-trained models (Devlin et al., 2019; Liu et al., 2019), such as BERT, has been studied extensively in the literature (Gao et al., 2021; Reimers et al., 2019). Recently, contrastive learning has been used widely to learn better sentence embeddings (Meng et al., 2021; Gao et al., 2021). Generally, contrastive learning uses various data augmentation methods to generate different views of the input sentences and samples positive instances and negative instances from views. Contrastive learning aims to learn effective embeddings by pulling positive instances together and pushing positive and negative instances apart. This operation focus on maximizing the mutual information between positive instances to retain as much information as possible, which includes both the useful and useless information. Previous studies like (Meng et al., 2021; Gao et al., 2021) ignore the redundant information stored in views, which has a bad impact on the performance of downstream tasks, as proved by (Achille and Soatto, 2018; Tian et al., 2020).

Table 1: Training texts may contain various kinds of redundant information, such as stop words, restatement, capitalization, and hyphen.

Original	Where is the party, it sounds great.
Stop words	Where is the party, it sounds great.
Restatement	The party sounds great, where is it.
Capitalization	Where Is The Party, It Sounds Great.
Hyphen	Where-is-the-party, it-sounds-great.

Table 1 gives an example of redundant information stored in training texts. In training texts, the sentences contain much redundant information which is not favorable to the downstream task. The redundant information may be stop words and the style of the sentence (e.g., restatement, capitalization, and hyphen). The existing study (Tian et al., 2020) also shows that discarding redundant information in views can help to improve the performance of the downstream task. Thus, we arise the following question: how to discard this redundant information by choosing the optimal views?

It is natural to solve the above questing via information bottleneck (IB), which has been utilized as an effective and simple method for learning a good embedding by keeping the important information and forgetting redundant information in various tasks (TISHBY, 1999; Chen and Ji, 2020; Tishby and Zaslavsky, 2015). Prompted by this, we at-

[*] Yuling Sun is the corresponding authors of this paper.
[1]https://github.com/Bin199/InforMin-CL

Proceedings of the 29th International Conference on Computational Linguistics, pages 4821–4831
October 12–17, 2022.

tend to solve the problem by drawing inspiration from the information minimization principle (an idea in IB theory) (Tian et al., 2020): A good set of views share the minimal information necessary to perform well at the downstream task. This method aims to retain useful information and forget redundant information. In this paper, we explore how to address the shortcomings (i.e., ignoring the redundant information stored in views) of previous work for unsupervised sentence representations via the information minimization principle.

We propose an information minimization based contrastive learning (`InforMin-CL`) model for unsupervised sentence embedding learning. `InforMin-CL` incorporates the information minimization principle into contrastive learning to not only learn the important information but also drop the redundant information. We optimize our `InforMin-CL` model from two perspectives: contrast and reconstruction. Firstly, we learn the useful information by a contrast task to maximize the mutual information. This task manages to attract positive pairs and repulse negative pairs. Attracting positive pairs stands for maximizing mutual information between positive instances. Secondly, we propose a reconstruction task that encourages the model to reconstruct the representation of the positive instance via the other one in the same pair.

The algorithm of `InforMin-CL` is easy to understand and can be implemented with just several lines of code. Moreover, our method does not change the major network structure, so it is model-agnostic and can be applied to any representation learning neural networks based on contrastive learning. Experiments in Section 4 show that `InforMin-CL` can help the model learn effective representations that improve downstream task performance. Our main contributions are summarized as follows:

- We propose an `InforMin-CL` model to learn good sentence embeddings by keeping useful information and getting rid of redundant information between positive instances.
- We achieve our model via simple contrast and reconstruction tasks and prove that the reconstruction task can drop redundant information by minimizing the information entropy.
- We achieve new state-of-the-art results on seven supervised tasks and seven unsupervised tasks, which indicates the great advantages of our proposed model.

2 Related Work

2.1 Sentence Representation Learning

Learning sentence embeddings as an important problem in NLP has been widely studied. Recent work focus on leveraging the power of BERT (Devlin et al., 2019) to learn effective sentence embeddings, which are free from artificially supervised signals. BERT-flow (Li et al., 2020) transforms the anisotropic sentence embedding distribution into a smooth isotropic Gaussian distribution through normalizing flows. BERT-whitening (Su et al., 2021) further presents a whitening operation to enhance the isotropy of sentence embeddings and achieves better results.

Then, the contrastive learning approach is applied for sentence embedding learning. IS-BERT (Zhang et al., 2020) proposes a model with a feature extractor on top of BERT and an objective that maximizes the mutual information between global sentence embeddings and local sentence embeddings. CLEAR (Wu et al., 2020) employs multiple sentence-level augmentation strategies to learn sentence representation. Coco-LM (Meng et al., 2021) employs an auxiliary language model to corrupt text sequences, upon which it constructs a token-level task and a sequence-level task for pretraining the main model. Gao *et al.* (Gao et al., 2021) presents an unsupervised approach that predicts input itself with dropout noise and a supervised approach utilizing natural language inference datasets. SCD (Klein and Nabi, 2022) leverages the self-contrast of augmented samples obtained by dropout, which eliminates the reliance on negative pairs. However, all these studies lack consideration of discarding redundant information stored in views. In our work, we consider and solve the problem in the framework of the information minimization principle.

2.2 Information Minimization Principle

Information minimization principle (Tian et al., 2020) has been proposed to retain the minimal information necessary. In recent years, researchers utilize the information minimization principle to improve image representations (Tian et al., 2020; Tsai et al., 2020). Furthermore, information bottleneck is used to improve the interpretability of the

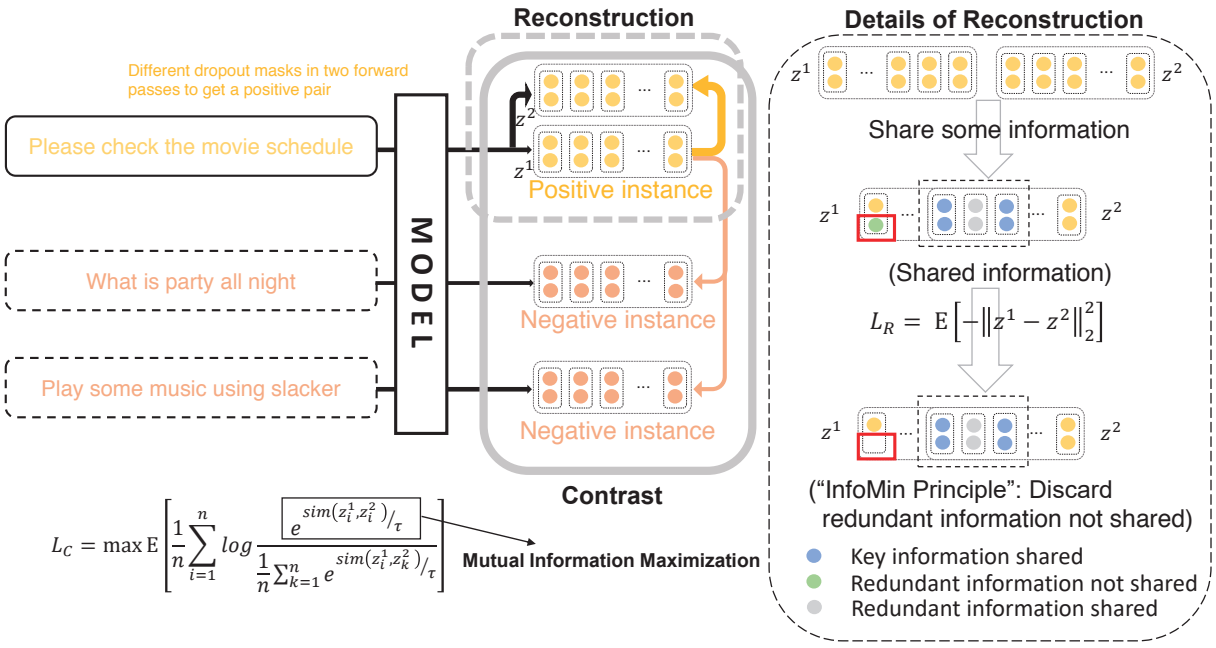


Figure 1: The architecture of our proposed framework. Contrast: Negative pairs are pushed apart while positive pairs are pulled together, which suggests maximizing the mutual information between positive instances. Reconstruction: We drop redundant information in z^1 not shared with z^2 (marked by the green point) by reconstructing one positive instance via the other positive instance.

attention-based models (Zhou et al., 2021). Tian *et al.* (Tian et al., 2020) shows good views for a given task in a contrastive representation learning framework should retain task-relevant information while minimizing irrelevant nuisances. However, it focuses on eliminating the task-irrelevant information via downstream datasets. Tsai *et al.* (Tsai et al., 2020) focuses on the multi-view setting between input and self-supervised signals and adopts self-supervised signals to reconstruct learned representations to discard task-irrelevant information. Our work also differs from (Tsai et al., 2020) in two perspectives: 1) we discard redundant information in texts while (Tsai et al., 2020) drops noise information stored in images, and additionally different self-supervised signals are used in two work; 2) (Tsai et al., 2020) validates their method by learning visual features evaluated by supervised tasks while sentence embedding learning evaluates embeddings via not only supervised tasks but also unsupervised tasks.

3 Method

We propose an `InforMin-CL` model for unsupervised sentence representation learning (Figure 1). There are two main steps, contrast and reconstruction. We first present the contrast objective to learn the useful information: we push apart positive instances and negative instances while pulling positive instances together, which implies maximizing mutual information between positive instances. Later we present the reconstruction task to drop the useless information: we minimize the conditional information entropy of one positive instance given the other positive instance. Algorithm 1 provides the pseudo-code of *InforMin-CL*.

Algorithm 1 Pseudocode of *InforMin-CL* in a PyTorch-like style.

Input: batch size N, temperature τ, structure of f and Γ.
Output: encoder network $f(\cdot)$.

for sampled minibatch $\{x_k\}_{k=1}^{N}$ **do**
  for all $k \in \{1, \ldots, N\}$ **do**
    draw two augmentation functions $t \sim \Gamma$, $t' \sim \Gamma$
    $\tilde{x}_{2k-1} = t(x_k)$
    $z_{2k-1} = f(\tilde{x}_{2k-1})$
    $\tilde{x}_{2k} = t'(x_k)$
    $z_{2k} = f(\tilde{x}_{2k})$
  end for
  for all $i \in \{1, \ldots, 2N\}$ and $j \in \{1, \ldots, 2N\}$ **do**
    $s_{i,j} = z_i^{\mathrm{T}} z_j / (\|z_i\| \, \|z_j\|)$
  end for
  $L_C = \max \mathrm{E}\left[\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(s_{2i-1,2i}/\tau)}{\frac{1}{N}\sum_{k=1}^{N}\exp(s_{2i-1,2k}/\tau)}\right]$
  $L_R = \mathrm{E}\left[-\|z_{2k-1} - z_{2k}\|_2^2\right]$
  $L = L_C + L_R$
  update f to minimize L
end for

3.1 Contrast

Contrastive learning attends to learning effective representation by pulling positive sample pairs together and pushing apart negative sample pairs. We build upon the recent success of unsupervised SimCSE (Gao et al., 2021) and take the embeddings derived from the same sentence with independently different dropout masks as positive instances. We adopt the dropout mask (with default dropout probability p = 0.1) as an augmentation skill, which is proved to outperform other skills (Gao et al., 2021), to obtain positive pairs. The positive pair takes the same sentence, and their embeddings only differ in dropout masks. Other sentences in the same mini-batch are seen as negative instances.

We denote input, instance, and self-supervised signal as X, Z, and S. We feed the same input x to the encoder twice by applying different dropout masks and then get positive instances z^1 and z^2. In our work, we take one instance z^2 in positive pair as a self-supervised signal. The information required for downstream tasks is referred to as "key information": T. I and H represent mutual information and information entropy.

Let Z^{sup} be the sufficient supervised representation and $Z^{sup_{\min}}$ be the minimal and sufficient supervised representation:

$$\begin{aligned} Z^{\text{sup}} &= \arg\max_{Z} I\left(Z;T\right) \\ Z^{sup_{\min}} &= \arg\min_{Z} H\left(Z|T\right) \\ s.t.\ \ & I\left(Z;T\right) is \,\text{maximized} \end{aligned} \tag{1}$$

Let Z^{ssl} be the sufficient self-supervised representation and $Z^{ssl_{\min}}$ be the minimal and sufficient self-supervised representation:

$$\begin{aligned} Z^{ssl} &= \arg\max_{Z} I\left(Z;S\right) \\ Z^{ssl_{\min}} &= \arg\min_{Z} H\left(Z|S\right) \\ s.t.\ \ & I\left(Z;S\right) is \,\text{maximized} \end{aligned} \tag{2}$$

Then, we give theorem 1. For proof, please refer to (Tsai et al., 2020).

Theorem 1 *The supervised learned representations contain all the key information in the input (i.e. $I\left(X;T\right)$). The self-supervised learned representations contain all the key information in the input with a potential loss ε:*

$$\begin{aligned} I\left(X;T\right) = I\left(Z^{\text{sup}};T\right) &= I\left(Z^{\text{sup}_{\min}};T\right) \\ &\geq I\left(Z^{ssl};T\right) \\ &\geq I\left(Z^{ssl_{\min}};T\right) \\ &\geq I\left(X;T\right) - \varepsilon \end{aligned} \tag{3}$$

The contrastive learning objective maximizes the dependency between positive instance z^1 and self-supervised signal z^2, which suggests maximizing the mutual information $I\left(z^1;z^2\right)$. Theorem 1 suggests that maximizing $I\left(z^1;z^2\right)$ results in z_1 containing almost all the information required for downstream tasks from the input x. Note that T is utilized only for describing our method and in practice, no downstream datasets are used in the pre-training phase.

We use a contrastive learning objective similar to that in (Oord et al., 2018), which is a mutual information lower bound with low variance:

$$\mathcal{L}_C = \max \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{sim\left(z_i^1,z_i^2\right)/\tau}}{\frac{1}{N}\sum_{k=1}^{N}e^{sim\left(z_i^1,z_k^2\right)/\tau}}\right] \tag{4}$$

where $\left(z_1^1,z_1^2\right),\ldots,\left(z_N^1,z_N^2\right) \sim P^N\left(Z^1,Z^2\right)$, z_i^1, z_i^2 are two positive instances of the i-th examples. N refers to batch size and P refers to the statistical distribution of $\left(Z^1,Z^2\right)$.

3.2 Reconstruction

The details of the reconstruction are illustrated in the right of Figure 1. The positive instances z^1 and z^2 contain independent information while sharing some information. The noise information (marked as the green point) in z^1 is expected to be discarded by the reconstruction task. We prove that this task can discard the useless information by minimizing the information entropy.

The reconstruction task encourages the self-supervised signal z^2 to reconstruct the learned representation z^1, which suggests maximizing the log conditional likelihood $\mathbb{E}_{P_{Z^1,Z^2}}\left[\log P\left(Z^1|Z^2\right)\right]$. We know that

$$-H\left(Z^1|Z^2\right) = \mathbb{E}_{P_{Z^1,Z^2}}\left[\log P\left(Z^1|Z^2\right)\right] \tag{5}$$

Thus, this reconstruction also means minimizing $H\left(Z^1|Z^2\right)$.

Theorem 2 *The sufficient self-supervised representation contains more redundant information*

in the input than the sufficient and minimal self-supervised representation. The latter contains an amount of the information, $I\left(X;S|T\right)$, that cannot be discarded from the input:

$$\begin{aligned} I\left(Z^{ssl};X|T\right) &= I\left(X;S|T\right) + I\left(Z^{ssl};X|S,T\right) \\ &\geq I\left(Z^{ssl_{\min}};X|T\right) = I\left(X;S|T\right) \\ &\geq I\left(Z^{\sup_{\min}};X|T\right) = 0 \end{aligned} \tag{6}$$

Theorem 2 (please refer to (Tsai et al., 2020) for proof) indicates that Z^{ssl} contains two parts of redundant information while $Z^{ssl_{\min}}$ contains one part of redundant information, discarding $I\left(Z^{ssl};X|S,T\right)$.

Thus, if z^2 can perfectly reconstruct z^1 for any

$$\left(z^1, z^2\right) \sim P_{Z^1,Z^2} \tag{7}$$

under the constraint that $I\left(z^1;z^2\right)$ is maximized, we get $z^{1_{ssl_{\min}}}$ according to Eq. 2. And then z^1 discards redundant information, excluding $I\left(z^1;z^2|t\right)$ (i.e., the amount of redundant information in the shared information between two positive instances z^1 and z^2). For easier optimization, we use $\mathbb{E}_{P_{Z^1,Z^2}}\left[\log Q_\Phi\left(Z^1|Z^2\right)\right]$ as the lower bound of $\mathbb{E}_{P_{Z^1,Z^2}}\left[\log P\left(Z^1|Z^2\right)\right]$. In our deployment, we utilize the design in Eq. 4 and let $Q_\Phi\left(Z^1|Z^2\right)$ be Gaussian $N\left(Z^1|Z^2,\sigma\mathrm{I}\right)$ with $\sigma\mathrm{I}$ as a diagonal matrix. Hence, we obtain the reconstruction objective as follows:

$$\mathcal{L}_R = \mathbb{E}_{z^1,z^2\sim P_{Z^1,Z^2}}\left[-\left\|z^1 - z^2\right\|_2^2\right] \tag{8}$$

We combine two objectives as a total objective:

$$\mathcal{L} = \mathcal{L}_C + \lambda * \mathcal{L}_R \tag{9}$$

where λ is a hyper-parameter. Training model with the total loss enables us to discard redundant information in views.

4 Experiment

4.1 Evaluation Setup

We conduct our experiments on seven standard supervised tasks and also seven unsupervised tasks. We use the SentEval Toolkit (Conneau and Kiela, 2018) for evaluation. Following (Reimers et al., 2019; Gao et al., 2021), we take unsupervised tasks as the main comparison of the sentence embedding approaches and supervised results for reference.

Unsupervised Tasks We evaluate representations on seven semantic textual similarity (STS) tasks: STS 2012-2016 (Agirre et al., 2012, 2013, 2014, 2015, 2016), STS Benchmark (Cer et al., 2017), and SICK-Relatedness (Marelli et al., 2014) and compute the cosine similarity between sentence embeddings. All the unsupervised experiments are fully unsupervised, which means no STS training datasets are used and all embeddings are fixed once they are trained. For the sake of comparability, we follow the evaluation protocol of (Gao et al., 2021), employing Spearman's rank correlation and aggregation on all topic subsets.

Supervised Tasks We evaluate representations on seven supervised tasks: MR (Pang and Lee, 2005), CR (Hu and Liu, 2004), SUBJ (Pang and Lee, 2004), MPQA (Wiebe et al., 2005), SST-2 (Socher et al., 2013), TREC (Voorhees and Tice, 2000) and MRPC (Dolan and Brockett, 2005). A logistic regression classifier is trained on the top of (frozen) sentence embeddings produced by different methods. We follow default configurations from SentEval and use accuracy as the metric.

Training Details We start from pre-trained BERT (Devlin et al., 2019) (uncased) or RoBERTa (Liu et al., 2019) (cased). Similar to (Gao et al., 2021), we train our `InforMin-CL` in an unsupervised fashion on 10^6 randomly sampled sentences from English Wikipedia. During training, we add an MLP layer on the top of the [CLS] representation as sentence embeddings and directly take the [CLS] representation as sentence embeddings at testing time. A masked language modeling (MLM) objective (Devlin et al., 2019) is added as an optional auxiliary loss to the Eq. 9: $\mathcal{L} + \beta * \mathcal{L}_{MLM}$ (β is a hyper-parameter). For all results, we use the following hyper-parameters: epoch: 1, temperature τ: 0.05, optimizer: Adam (Kingma and Ba, 2015)). We carry out grid-search of batch size $\in \{64, 128, 256\}$ and learning rate $\in \{1e-5, 3e-5, 5e-5\}$ on STS-B development sets. During the training process, we save the checkpoint with the highest score on the STS-B development set to find the best hyperparameters. We adopt the hyperparameter settings listed in Table 3. For all results, we use a PC with a GeForce RTX 3090 GPU (CUDA 11, PyTorch 1.7.1).

4.2 Main Results

Baselines We compare `InforMin-CL` to previous typical sentence embedding methods, which

Table 2: Unsupervised task results (spearman's correlation). †: results from (Gao et al., 2021). ♡: results from (Klein and Nabi, 2022). All other results are reproduced and reevaluated by ourselves.

Model	STS12	STS13	STS14	STS15	STS16	STS-B	SICK-R	Avg.
GloVe embeddings (avg.)†	55.14	70.66	59.73	68.25	63.66	58.02	53.76	61.32
$\text{BERT}_{\text{base}}$ (first − last avg.)†	39.70	59.38	49.67	66.03	66.19	53.87	62.06	56.70
$\text{BERT}_{\text{base}}$−flow†	58.40	67.10	60.85	75.16	71.22	68.66	64.47	66.55
$\text{BERT}_{\text{base}}$−whitening†	57.83	66.90	60.90	75.08	71.31	68.24	63.73	66.28
IS − $\text{BERT}_{\text{base}}$†	56.77	69.24	61.21	75.23	70.16	69.21	64.25	66.58
CT − $\text{BERT}_{\text{base}}$†	61.63	76.80	68.47	77.50	76.48	74.31	69.19	72.05
SCD − $\text{BERT}_{\text{base}}$♡	66.94	78.03	69.89	78.73	76.23	76.30	**73.18**	74.19
SimCSE − $\text{BERT}_{\text{base}}$	67.01	82.14	73.76	80.49	79.01	77.04	69.94	75.63
InforMin-CL − $\text{BERT}_{\text{base}}$	**70.22**	**83.48**	**75.51**	**81.72**	**79.88**	**79.27**	71.03	**77.30**
$\text{RoBERTa}_{\text{base}}$ (first − last avg.)†	40.88	58.74	49.07	65.63	61.48	58.55	61.63	56.57
$\text{RoBERTa}_{\text{base}}$−whitening†	46.99	63.24	57.23	71.36	68.99	61.36	62.91	61.73
DeCLUTR − $\text{RoBERTa}_{\text{base}}$†	52.41	75.19	65.52	77.12	78.63	72.41	68.62	69.99
SCD − $\text{RoBERTa}_{\text{base}}$♡	63.53	77.79	69.79	80.21	77.29	76.55	72.10	73.89
SimCSE − $\text{RoBERTa}_{\text{base}}$	**70.32**	82.48	**74.84**	**82.13**	**82.14**	**81.57**	68.62	**77.44**
InforMin-CL − $\text{RoBERTa}_{\text{base}}$	69.79	**82.57**	73.36	80.91	81.28	81.07	**70.30**	77.04
SimCSE − $\text{RoBERTa}_{\text{large}}$	**72.64**	83.78	**75.83**	**84.24**	**80.12**	**81.10**	69.81	**78.22**
InforMin-CL − $\text{RoBERTa}_{\text{large}}$	70.91	**84.20**	75.57	82.26	79.68	**81.10**	**72.81**	78.08

Table 3: Batch sizes, learning rates and λ adopted for InforMin-CL.

	BERT	RoBERTa	
	base	base	large
Batch size	128	128	128
Learning rate	3e-5	1e-5	3e-5
λ	0.4	4	4

include averaging GloVe embeddings (Pennington et al., 2014), Skip-thought (Kiros et al., 2015) and average BERT or RoBERTa embeddings. We also compare to post-processing methods and methods using a contrastive objective. Post-processing methods include BERT-flow (Li et al., 2020) and BERT-whitening (Su et al., 2021). Methods using a contrastive objective include IS-BERT (Zhang et al., 2020), DeCLUTR (Giorgi et al., 2021), CT (Carlsson et al., 2020), SimCSE (Gao et al., 2021), and SCD (Klein and Nabi, 2022). IS-BERT maximizes the agreement between global and local features. DeCLUTR takes different spans from the same document as positive pairs. CT aligns embeddings of the same sentence from two different encoders. SimCSE takes the embedding of the same input with different dropouts as positive pairs. SCD leverages the self-contrast of augmented samples obtained by dropout.

Performance on Unsupervised Tasks Table 2 shows the evaluation results on seven STS tasks. InforMin-CL achieves comparable or better results than previous state-of-the-art baselines. For $\text{BERT}_{\text{base}}$ based models, our method outperforms the best approach with a large margin ($+1.67\%$) on average. For RoBERTa based model, InforMin-CL obtains comparable results (less than 0.4 points in average). All these indicate that our model can improve the performance of downstream tasks by forgetting the irrelevant information in pre-training phase.

Performance on Supervised Tasks Table 4 shows the evaluation results on seven supervised tasks. Results indicate that InforMin-CL performs on par or better than all baselines. Our method achieves better results on most of and even all tasks using RoBERTa, with an average gain 0.81% on $\text{RoBERTa}_{\text{base}}$ and 1.66% on $\text{RoBERTa}_{\text{large}}$ respectively. With an MLM task added, further gains on average results are observed for BERT and RoBERTa. It raises the average scores of InforMin-CL from 85.52% to 86.96% for $\text{BERT}_{\text{base}}$ and from 86.11% to 87.01% for $\text{RoBERTa}_{\text{base}}$. Particularly, InforMin-CL w/ MLM obtains the outperforms all the baselines in average.

Considering results of both supervised and unsupervised tasks, we present the following findings: 1) BERT-based InforMin-CL performs better on unsupervised tasks; 2) RoBERTa-based InforMin-CL achieves better results on supervised tasks. The difference in performance us-

Table 4: Supervised task results (accuracy). †: results from (Gao et al., 2021). ♡: results from (Klein and Nabi, 2022). All other results are reproduced and reevaluated by ourselves. w/ MLM: adding MLM as an auxiliary task with $\beta = 0.1$.

Model	MR	CR	SUBJ	MPQA	SST	TREC	MRPC	Avg.
GloVe embeddings (avg.)†	77.25	78.30	91.17	87.85	80.18	83.00	72.87	81.52
Skip − thought†	76.50	80.10	93.60	87.10	82.00	92.20	73.00	83.50
Avg. BERT embeddings†	78.66	86.25	94.37	88.66	84.40	**92.80**	69.54	84.94
BERT− [CLS] embeddings†	78.68	84.85	94.21	88.23	84.13	91.40	71.13	84.66
IS − $\text{BERT}_{\text{base}}$†	81.09	**87.18**	94.96	88.75	85.96	88.64	74.24	85.83
SCD − $\text{BERT}_{\text{base}}$♡	73.21	85.80	**99.56**	88.67	85.59	89.80	75.71	85.52
SimCSE − $\text{BERT}_{\text{base}}$	81.47	86.86	94.79	89.25	86.27	89.40	72.81	85.84
InforMin-CL − $\text{BERT}_{\text{base}}$	80.99	85.72	94.63	**89.47**	85.67	88.20	73.97	85.52
w/ MLM	**82.87**	87.05	95.22	88.43	**87.15**	92.20	**75.77**	**86.96**
SimCSE − $\text{RoBERTa}_{\text{base}}$	81.26	87.36	93.58	87.56	86.93	84.80	75.01	85.21
SCD − $\text{RoBERTa}_{\text{base}}$♡	82.17	87.76	93.67	85.69	88.19	83.40	76.23	85.30
InforMin-CL − $\text{RoBERTa}_{\text{base}}$	82.22	88.08	93.57	**87.75**	87.59	86.60	76.99	86.11
w/ MLM	**83.49**	**88.69**	**94.79**	86.81	**88.30**	**89.40**	**77.57**	**87.01**
SimCSE − $\text{RoBERTa}_{\text{large}}$	80.85	85.99	93.08	87.65	86.33	89.00	72.46	85.05
InforMin-CL − $\text{RoBERTa}_{\text{large}}$	**82.50**	**88.32**	**93.81**	**89.38**	**87.64**	**90.80**	**74.49**	**86.71**

ing BERT and RoBERTa is mainly caused by the difference in pre-training corpus. BERT is trained over 16 GB text (BooksCorpus (Zhu et al., 2015) and English Wikipedia) while RoBERTa is trained over totally 160 GB of uncompressed text (BooksCorpus (Zhu et al., 2015), English Wikipedia, CC-NEWS (Nagel, 2016), OPENWEBTEXT (Gokaslan et al., 2019), and STORIES (Trinh and Le, 2018)). Thus, the diverse large-scale high quality datasets enhance the RoBERTa to learn the important and useful information with limited parameters. InforMin-CL, which improves performance by discarding redundant information, struggles to represent its effects in this setting due to less noise information. This causes the unsupervised results of InforMin-CL are similar to competitors for RoBERTa-based models.

4.3 Ablation Study

Influence of λ We investigate how different reconstruction objectives with λ from 0.04 to 4 affect our model's performance. We report the average performance of unsupervised tasks and supervised tasks in this experiment. The results are obtained using $\text{BERT}_{\text{base}}$. Results demonstrate that InforMin-CL constantly works well over this wide range of λ. As shown in Table 5, with increasing λ, the performance of both unsupervised and supervised tasks rises first and falls later.

Influence of β We introduce one more optional variant which adds a masked language modeling (MLM) objective to the Eq. 9: $\mathcal{L} + \beta * \mathcal{L}_{MLM}$ (β is a hyper-parameter). We analyze how different β influence the performance on unsupervised and supervised tasks. As we show in Table 6, we find that adding MLM objectives with different β consistently helps improve performance on supervised tasks but brings a significant drop in STS tasks.

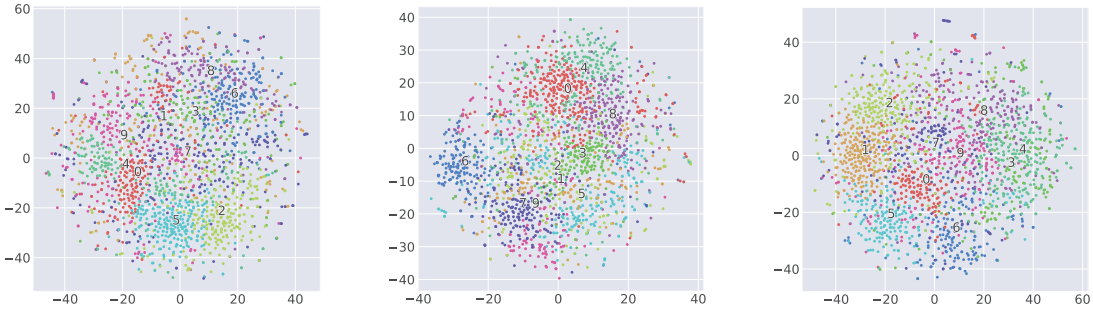
Influence of Batch Sizes To explore the impact of batch sizes, we report the average performance of downstream tasks with batch sizes (N in Eq. 4) from 64 to 256. In this experiment, only batch size changes while all other hyper-parameters keep unchanged. We use $\text{BERT}_{\text{base}}$ to evaluate on the test set of unsupervised and supervised tasks. As we show in Table 7, we find that InforMin-CL is not sensitive to batch size, similar to SimCSE, mainly caused by the good set of initial parameters.

Table 5: Ablation studies of different hyper-parameters λ. The results are based on the test sets using $\text{BERT}_{\text{base}}$.

λ	Avg. Sup	Avg. Unsup
0.04	85.20	76.09
0.4	**85.52**	**77.30**
4	85.03	77.18

4.4 Uniformity and Alignment

We further conduct analysis to understand the inner workings of InforMin-CL.



(a) SimCSE (b) SCD (c) InforMin-CL

Figure 2: The t-SNE of sentence representations learned with SimCSE, SCD and InforMin-CL using $\text{BERT}_{\text{base}}$. The points are embeddings of sentences sampled from the IMDB dataset without fine-tuning.

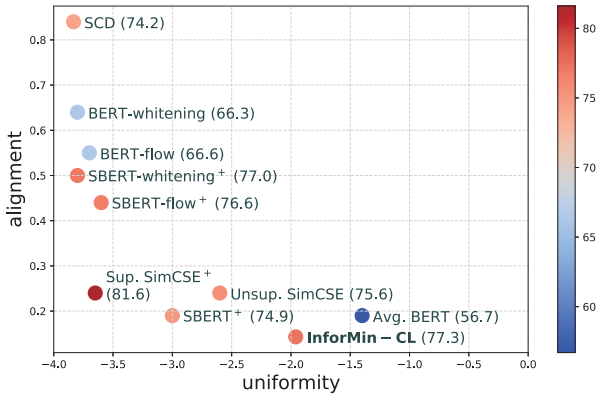


Figure 3: Quantitative analysis of embeddings - *alignment* vs. *uniformity* (the smaller, the better). The plot of models is based on $\text{BERT}_{\text{base}}$. Points represent average STS performance with Spearman's correlation color coded (+ corresponds to supervised methods).

Table 6: Ablation studies of the MLM objective based on the test sets using $\text{BERT}_{\text{base}}$.

Model	Avg. Sup	Avg. Unsup
w/o MLM	85.52	**77.30**
w/ MLM		
$\beta = 0.01$	86.46	**63.59**
$\beta = 0.1$ (ours)	86.96	63.25
$\beta = 1.0$	**87.04**	60.85

Table 7: Ablation studies of different batch sizes. The results are based on test sets using $\text{BERT}_{\text{base}}$.

Batch size	64	128	256
Avg. Sup	85.38	85.52	**85.77**
Avg. Unsup	76.64	**77.30**	76.14

Qualitative Analysis As shown in (Wang and Isola, 2020), the asymptotics of the contrastive learning objective (4) can be expressed by the following equation when the number of negative instances approaches infinity:

$$\mathcal{L}_C = \max \Big[\frac{1}{N} \sum_{i=1}^{N} \mathbb{E}\big[sim\left(z_i^1, z_i^2\right)/\tau\big] - \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}\Big[\log \frac{1}{N} \sum_{k=1}^{N} e^{sim\left(z_i^1, z_k^2\right)/\tau}\Big]\Big] \tag{10}$$

The first term in square brackets in Eq. 10 improves alignment of the space. The alignment performs better when the similarity score rises. While optimizing the reconstruction objective, z^1 and z^2 are pulled closer, which means that the similarity score of z^1 and z^2 becomes higher. In other words, InforMin-CL effectively improves alignment of pre-trained embeddings while keeping a good uniformity, which is the key to the success of InforMin-CL. We also follow (Wang and Isola, 2020) to use uniformity and alignment to measure the quality of representation space for InforMin-CL and other models. Figure 3 shows uniformity and alignment of different sentence embedding models along with their STS averaged results. InforMin-CL achieves the best in terms of *alignment* (0.143), which can be related to the strong effect of the reconstruction objective. In terms of *uniformity*, InforMin-CL is slightly inferior to unsupervised SimCSE. This is also reflected in the final results in the t-SNE plots.

Quantitative Analysis The t-SNE (Reif et al., 2019) plot in Figure 2 demonstrates the advantages of InforMin-CL. We sample 2000 sentences from IMDB (Maas et al., 2011) dataset and gen-

erate the embeddings of sentences using SimCSE, SCD and InforMin-CL. We use K-Means (Jain and Dubes, 1988) clustering to group similar sentence embeddings and form 10 clusters. Results indicate that similar sentence pairs (marked by same colors) generated by `InforMin-CL` are more aligned.

5 Limitations

Although our method outperforms baselines on both unsupervised and supervised tasks in most cases, there are still at least two limitations. First, we simply sample negative instances from other sentences in the mini-batch, which may lead to false negatives. Punishing false negatives during training by assigning lower weight for negatives with higher similarity may be a solution. Second, although redundant information is discarded, what redundant information forgets and remains is unknown. It would be interesting to explore this problem by integrating interpretation methods.

6 Conclusion

In this work, we propose `InforMin-CL`, an effective contrastive learning approach, which improves state-of-the-art sentence embedding performance on downstream tasks. `InforMin-CL` discards redundant information stored in positive instances by encouraging one positive instance to reconstruct the other positive instance in the same pair. We test `InforMin-CL` on seven supervised and seven unsupervised tasks. Experimental results indicate our method outperforms all previous competitors.

Acknowledgements

We thank the anonymous reviewers for their valuable comments. We also thank Shanghai Science and Technology Innovation Action Plan Project (21511104500) for funding this research.

References

Alessandro Achille and Stefano Soatto. 2018. Emergence of invariance and disentanglement in deep representations. *The Journal of Machine Learning Research*, 19(1):1947–1980.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Inigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, et al. 2015. Semeval-2015 task 2: Semantic textual similarity, english, spanish and pilot on interpretability. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 252–263.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2014. Semeval-2014 task 10: Multilingual semantic textual similarity. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, pages 81–91.

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez Agirre, Rada Mihalcea, German Rigau Claramunt, and Janyce Wiebe. 2016. Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *SemEval-2016. 10th International Workshop on Semantic Evaluation; 2016 Jun 16-17; San Diego, CA. Stroudsburg (PA): ACL; 2016. p. 497-511.*

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. Semeval-2012 task 6: A pilot on semantic textual similarity. In ** SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 385–393.

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. * sem 2013 shared task: Semantic textual similarity. In *Second joint conference on lexical and computational semantics (* SEM), volume 1: proceedings of the Main conference and the shared task: semantic textual similarity*, pages 32–43.

Fredrik Carlsson, Amaru Cuba Gyllensten, Evangelia Gogoulou, Erik Ylipää Hellqvist, and Magnus Sahlgren. 2020. Semantic re-tuning with contrastive tension. In *International Conference on Learning Representations*.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14.

Hanjie Chen and Yangfeng Ji. 2020. Learning variational word masks to improve the interpretability of neural text classifiers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4236–4251.

Alexis Conneau and Douwe Kiela. 2018. Senteval: An evaluation toolkit for universal sentence representations. *arXiv preprint arXiv:1803.05449*.

Jacob Devlin, Ming-Wei Chang, and Kenton et al. Lee. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*.

William B Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases.

In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 6894–6910.

John Giorgi, Osvald Nitski, Bo Wang, and Gary Bader. 2021. DeCLUTR: Deep contrastive learning for unsupervised textual representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 879–895.

Aaron Gokaslan, Vanya Cohen, Ellie Pavlick, and Stefanie Tellex. 2019. Openwebtext corpus.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177.

Anil K Jain and Richard C Dubes. 1988. *Algorithms for clustering data*. Prentice-Hall, Inc.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Ryan Kiros, Yukun Zhu, Russ R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Advances in neural information processing systems*, pages 3294–3302.

Tassilo Klein and Moin Nabi. 2022. Scd: Self-contrastive decorrelation for sentence embeddings. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics (ACL).

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. On the sentence embeddings from pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 9119–9130.

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

Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 142–150.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A sick cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 216–223.

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. In *Conference on Neural Information Processing Systems*.

Sebastian Nagel. 2016. Cc-news. *URL: http://web. archive. org/save/http://commoncrawl. org/2016/10/newsdatasetavailable*.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Bo Pang and Lillian Lee. 2004. A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, pages 271–es.

Bo Pang and Lillian Lee. 2005. Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 115–124.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Emily Reif, Ann Yuan, Martin Wattenberg, Fernanda B Viegas, Andy Coenen, Adam Pearce, and Been Kim. 2019. Visualizing and measuring the geometry of bert. *Advances in Neural Information Processing Systems*, 32.

Nils Reimers, Iryna Gurevych, Nils Reimers, Iryna Gurevych, Nandan Thakur, Nils Reimers, Johannes Daxenberger, Iryna Gurevych, Nils Reimers, Iryna Gurevych, et al. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Jianlin Su, Jiarun Cao, Weijie Liu, and Yangyiwen Ou. 2021. Whitening sentence representations for better semantics and faster retrieval. *arXiv preprint arXiv:2103.15316*.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. 2020. What makes for good views for contrastive learning? In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

N TISHBY. 1999. The information bottleneck method. In *Proc. 37th Annual Allerton Conference on Communications, Control and Computing, 1999*, pages 368–377.

Naftali Tishby and Noga Zaslavsky. 2015. Deep learning and the information bottleneck principle. In *2015 ieee information theory workshop (itw)*, pages 1–5. IEEE.

Trieu H Trinh and Quoc V Le. 2018. A simple method for commonsense reasoning. *arXiv preprint arXiv:1806.02847*.

Yao-Hung Hubert Tsai, Yue Wu, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020. Self-supervised learning from a multi-view perspective. In *International Conference on Learning Representations*.

Ellen M Voorhees and Dawn M Tice. 2000. Building a question answering test collection. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 200–207.

Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939. PMLR.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*.

Zhuofeng Wu, Sinong Wang, Jiatao Gu, Madian Khabsa, Fei Sun, and Hao Ma. 2020. Clear: Contrastive learning for sentence representation. *arXiv preprint arXiv:2012.15466*.

Yan Zhang, Ruidan He, Zuozhu Liu, Kwan Hui Lim, and Lidong Bing. 2020. An unsupervised sentence embedding method by mutual information maximization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1601–1610.

Jie Zhou, Yuanbin Wu, Qin Chen, Xuan-Jing Huang, and Liang He. 2021. Attending via both fine-tuning and compressing. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2152–2161.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.