SCARF: Self-Supervised Contrastive Learning using Random Feature Corruption

Dara Bahri, Heinrich Jiang, Yi Tay, Donald Metzler
Google Research
{dbahri,heinrichj,yitay,metzler}@google.com

Abstract

Self-supervised contrastive representation learning has proved incredibly successful in the vision and natural language domains, enabling state-of-the-art performance with orders of magnitude less labeled data. However, such methods are domain-specific and little has been done to leverage this technique on real-world *tabular* datasets. We propose SCARF, a simple, widely-applicable technique for contrastive learning, where views are formed by corrupting a random subset of features. When applied to pre-train deep neural networks on the 69 real-world, tabular classification datasets from the OpenML-CC18 benchmark, SCARF not only improves classification accuracy in the fully-supervised setting but does so also in the presence of label noise and in the semi-supervised setting where only a fraction of the available training data is labeled. We show that SCARF complements existing strategies and outperforms alternatives like autoencoders. We conduct comprehensive ablations, detailing the importance of a range of factors.

1 Introduction

In many machine learning tasks, unlabeled data is abundant but labeled data is costly to collect, requiring manual human labelers. The goal of self-supervised learning is to leverage large amounts of unlabeled data to learn useful representations for downstream tasks such as classification. Self-supervised learning has proved critical in computer vision [25, 50, 28, 72] and natural language processing [68, 78, 61]. Some recent examples include the following: Chen et al. [10] showed that training a linear classifier on the representations learned by their proposed method, SimCLR, significantly outperforms previous state-of-art image classifiers and requires 100x fewer labels to achieve state of art results; Brown et al. [7] showed through their GPT-3 language model that by pre-training on a large corpus of text, only few labeled examples were required for task-specific fine-tuning for a wide range of tasks.

A common theme of these advances is learning representations that are robust to different views or distortions of the same input; this is often achieved by maximizing the similarity between views of the same input and minimizing those of different inputs via a contrastive loss. However, techniques to generate views or corruptions have thus far been, by and large, domain-specific (e.g. color distortion [92] and cropping [10] in vision, and token masking [68] in NLP). Despite the importance of self-supervised learning, there is surprisingly little work done in finding methods that are applicable across domains and in particular, ones that can be applied to tabular data.

In this paper, we propose SCARF, a simple and versatile contrastive pre-training procedure. We generate a view for a given input by selecting a random subset of its features and replacing them by random draws from the features' respective empirical marginal distributions. Experimentally, we test SCARF on the OpenML-CC18 benchmark [75, 5, 19], a collection of 72 real-world classification datasets. We show that not only does SCARF pre-training improve classification accuracy in the fully-supervised setting but does so also in the presence of label noise and in the semi-supervised setting

Preprint. Under review.

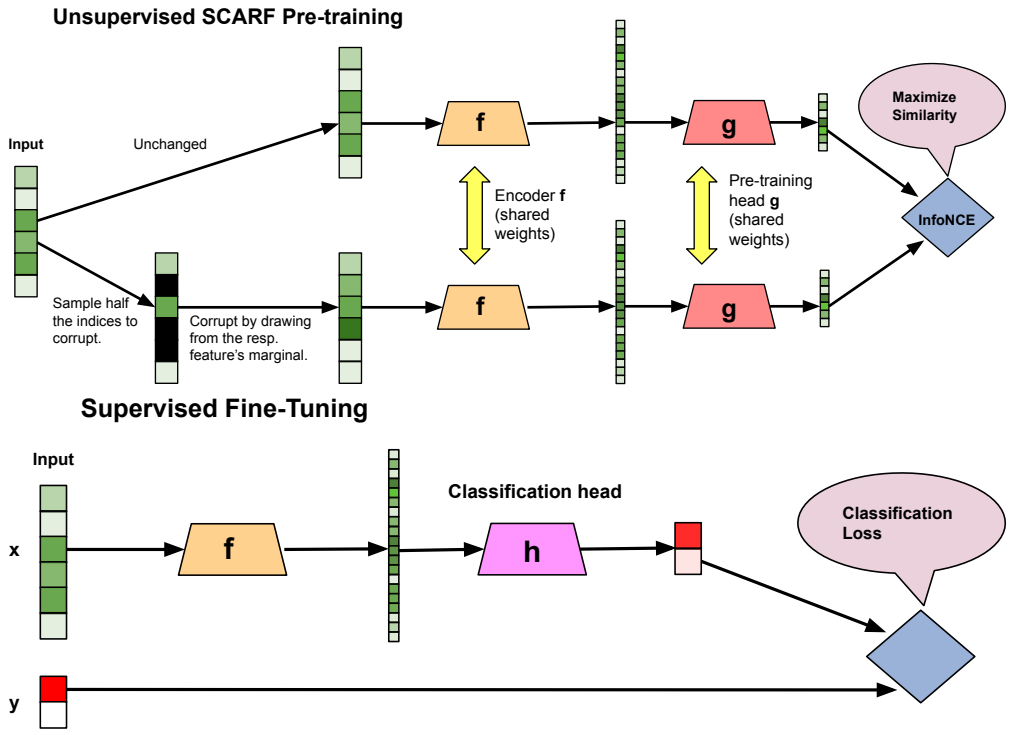

Figure 1: Diagram showing unsupervised SCARF pre-training (**Top**) and subsequent supervised fine-tuning (**Bottom**). During pre-training, networks f and g are learned to produce good representations of the input data. After pre-training, g is discarded and a classification head h is applied on top of the learned f and both f and h are subsequently fine-tuned for classification.

where only a fraction of the available training data is labeled. Moreover, we show that combining SCARF pre-training with other solutions to these problems further improves them, demonstrating the versatility of SCARF and its ability to learn effective task-agnostic representations. We then conduct extensive ablation studies, showing the effects of various design choices and stability to hyperparameters. Our ablations show that SCARF's way of constructing views is more effective than alternatives. We show that SCARF is less sensitive to feature scaling and is stable to various hyperparameters such as batch size, corruption rate, and softmax temperature.

2 Related Works

A number of self-supervised learning techniques have been proposed in computer vision [89, 74, 35]. One framework involves learning features based on generated images through various methods, including using a GAN [24, 17, 60, 9], predicting pixels [39], predicting colorizations [92, 40], ensuring local and global consistency [33], and learning synthetic artifacts [34]. Another line of work involves training on generated labels with semantic meanings that can be obtained without any human labelers (e.g. using segmentation masks) [63, 43, 56]. Leveraging the spatial relations of image patches and temporal ones of videos has also been successful. Such approaches include the image jigsaw puzzle [53, 1], context prediction [16], clustering [82, 8, 84], and geometric transformation recognition [41, 83]. Finally, most related to our approach, are the similarity-based approaches that contrast different views of the image [72, 27, 54, 29, 43, 28, 6, 80, 22]. In particular, our framework is similar to SimCLR [10], which involves generating views of a single image via image-based corruptions like random cropping, color distortion and blurring; however, we generate views that are applicable to tabular data.

Self-supervised learning has had an especially large impact in language modeling [59]. One popular approach is masked language modeling, wherein the model is trained to predict input tokens that have been intentionally masked out [15, 61, 67] as well as enhancements to this approach [45, 18, 4, 37, 36]

Algorithm 1 SCARF pre-training algorithm.

input: unlabeled training data $\mathcal{X} \subseteq \mathbb{R}^M$, batch size N, temperature τ, corruption rate c, encoder network f, pre-train head network g.
let $\widehat{\mathcal{X}_j}$ be the uniform distribution over $\mathcal{X}_j = \{x_j : x \in \mathcal{X}\}$, where x_j denotes the j-th coordinate of x.
let $q = \lfloor c \cdot M \rfloor$ be the number of features to corrupt.
for sampled mini-batch $\left\{x^{(i)}\right\}_{i=1}^{N} \subseteq \mathcal{X}$ **do**
    for $i \in [N]$, uniformly sample subset $\mathcal{I}_i$ from $\{1, ..., M\}$ of size q and define $\tilde{x}^{(i)} \in \mathbb{R}^M$ as follows: $\tilde{x}_j^{(i)} = x_j$ if $j \notin \mathcal{I}_i$, otherwise $\tilde{x}_j^{(i)} = v$, where $v \sim \widehat{\mathcal{X}_j}$. # generate corrupted view.
    let $z^{(i)} = g\left(f\left(x^{(i)}\right)\right)$, $\tilde{z}^{(i)} = g\left(f\left(\tilde{x}^{(i)}\right)\right)$, for $i \in [N]$. # embeddings for views.
    let $s_{i,j} = {z^{(i)}}^{\top} \tilde{z}^{(j)} / \left(\|z^{(i)}\|_2 \cdot \|\tilde{z}^{(j)}\|_2\right)$, for $i, j \in [N]$. # pairwise similarity.
    define $\mathcal{L}_{\text{cont}} := \frac{1}{N} \sum_{i=1}^{N} -\log\left(\frac{\exp(s_{i,i}/\tau)}{\frac{1}{N}\sum_{k=1}^{N} \exp(s_{i,k}/\tau)}\right)$.
    update networks f and g to minimize $\mathcal{L}_{\text{cont}}$ using SGD.
end for
return encoder network f.

and variations involving permuting the tokens [85, 68]. Denoising autoencoders have been used by training them to reconstruct the input from a corrupted version (produced by, for example, token masking, deletion, and infilling) [42, 78, 20]. Finally, most related to our approach, is contrastive learning, which involves learning based on similarity of pairs of inputs. Such approaches include randomly replacing words and distinguishing between real and fake phrases [12, 51], random token replacement [49, 11], and adjacent sentences [36, 38, 14].

Recently, Yao et al. [86] adapted the similarity-based contrastive framework to large-scale recommendation systems in a similar way to our approach. The key difference is in the way the methods generate multiple views. Yao et al. [86] proposes masking random features in a correlated manner and applying a dropout for categorical features, while our approach involves randomizing random features based on the features' respective marginal training distribution (in an uncorrelated way). Generating such views for a task is a difficult problem: there has been much work done in understanding and designing them [81, 58, 23, 46, 57, 55] and learning them [62, 13, 32, 44, 93, 73, 71]. In this work, we present a simple method to generate multiple views that's effective for unsupervised representation learning of tabular data, which may be of interest to other applications that benefit from such data augmentation.

Within the contrastive learning framework, the choice of loss function is significant. InfoNCE [26, 54], which can be interpreted as a non-parametric estimation of the entropy of the representation [79], is a popular choice. Since then there have been a number of proposals [88, 25, 31]; however, we show that InfoNCE is effective for our framework.

3 SCARF

We now describe our proposed method (Algorithm 1), which is also described in Figure 1. For each mini-batch of examples from the unlabeled training data, we generate a corrupted version $\tilde{x}^{(i)}$ for each example $x^{(i)}$ as follows. We sample some fraction of the features uniformly at random and replace each of those features by a random draw from that feature's empirical marginal distribution, which is defined as the uniform distribution over the values that feature takes on across the training dataset. Then, we pass both $x^{(i)}$ and $\tilde{x}^{(i)}$ through the encoder network f, whose output we pass through the pre-train head network g, to get $z^{(i)}$ and $\tilde{z}^{(i)}$ respectively. Note that the pre-train head network ℓ_2-normalizes the outputs so that they lie on the unit hypersphere – this has been found to be crucial in practice [10, 79]. We train on the InfoNCE contrastive loss, encouraging $z^{(i)}$ and $\tilde{z}^{(i)}$ to be close for all i and $z^{(i)}$ and $\tilde{z}^{(j)}$ to be far apart for $i \neq j$, and we optimize over the parameters of f and g via SGD.

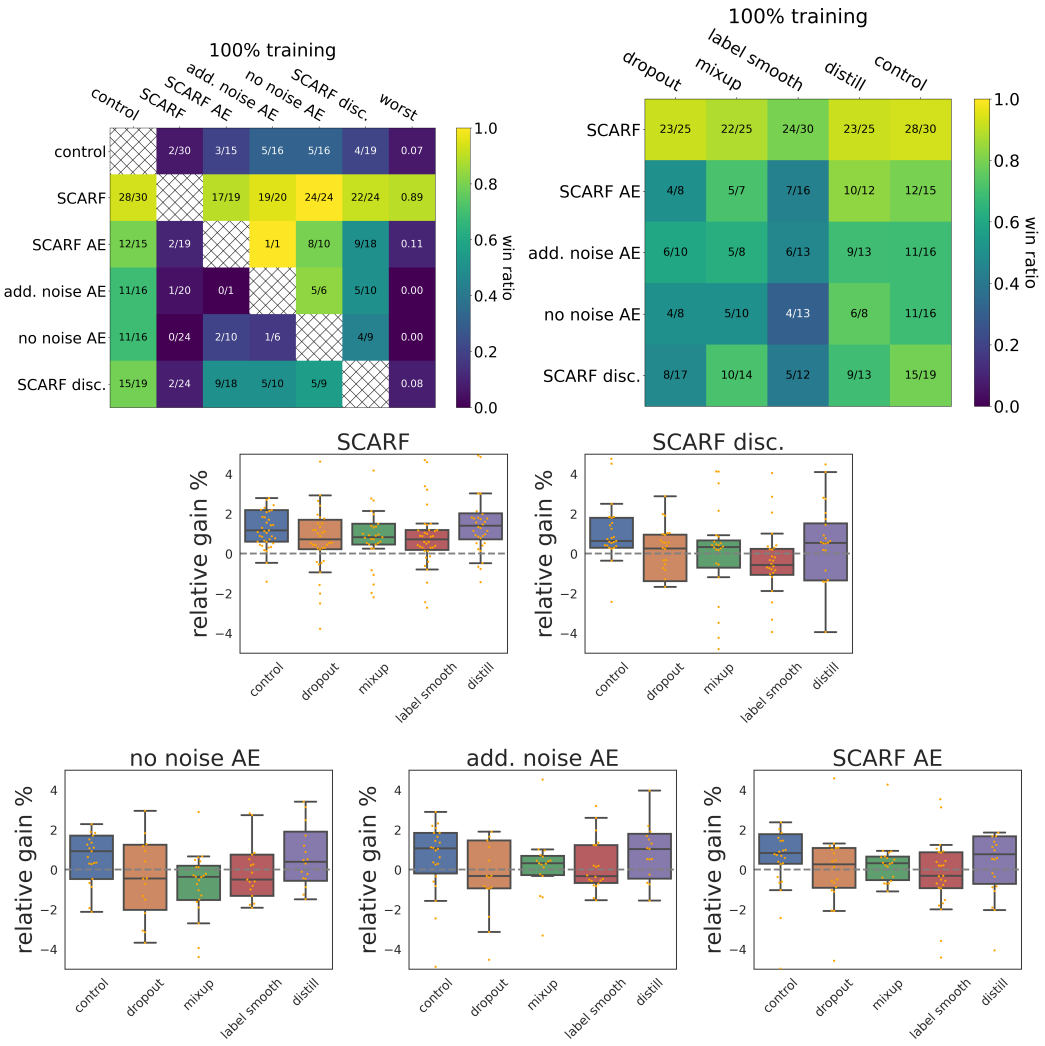

Figure 2: **Top**: Win matrices comparing pre-training methods against each other, and their improvement to existing solutions. **Bottom**: Box plots showing the relative improvement of different pre-training methods over baselines (y-axis is zoomed in). We see that SCARF pre-training adds value *even* when used in conjunction with known techniques.

Then, to train a classifier for the task via fine-tuning, we take the encoder network f and attach a classification head h which takes the output of f as its input and predicts the label of the example. We optimize the cross-entropy classification loss and tune the parameters of both f and h.

While pre-training can be run for a pre-determined number of epochs, much like normal supervised training, how many is needed largely depends on the model and dataset. To this end, we propose using early stopping on the validation InfoNCE loss. Given unlabeled validation data, we cycle through it for some number of epochs, running our proposed method to generate $\left(x^{(i)}, \tilde{x}^{(i)}\right)$ pairs. Once built, the loss on this static set is tracked during pre-training. Prototypical loss curves are shown in the Appendix.

4 Experiments

We evaluate the impact of SCARF pre-training on test accuracy after supervised fine-tuning in three distinct settings: on the full dataset, on the full dataset but where only 25% of samples have labels and the remaining 75% do not, and on the full dataset where 30% of samples undergo label corruption.

Datasets. We use 69 datasets from the public OpenML-CC18 benchmark[1] under the CC-BY licence. It consists of 72 real-world classification datasets that have been manually curated for effective benchmarking. Since we're concerned with tabular datasets in this work, we remove MNIST, Fashion-MNIST, and CIFAR10. For each OpenML dataset, we form $70\%/10\%/20\%$ train/validation/test splits, where a different split is generated for every trial and all methods use the same splits. The percentage used for validation and test are never changed and only training labels are corrupted for the label noise experiments.

Dataset pre-processing. We represent categorical features by a one-hot encoding, and most of the corruption methods explored in the ablations are on these one-hot encoded representations of the data (with the exception of SCARF, where the marginal sampling is done before one-hot encoding). We pre-process missing data as follows: if a feature column is always missing, we drop it. Otherwise, if the feature is categorical, we fill in missing entries with the mode, or most frequent, category computed over the full dataset. For numerical features, we inpute it with the mean. We explore scaling numerical features by z-score, min-max, and mean normalization. We find that for a vanilla network (i.e. control), z-score normalization performed the best for all but three datasets (OpenML dataset ids 4134, 28, and 1468), for which no scaling was optimal. We thus do not scale these three datasets and z-score normalize all others.

Model architecture and training. Unless specified otherwise, we use the following settings across all experiments. As described earlier, we decompose the neural network into an encoder f, a pre-training head g, and a classification head h, where the inputs to g and h are the outputs of f. We choose all three component models to be ReLU networks with hidden dimension 256. f consists of 4 layers, whereas both g and h have 2 layers. Both SCARF and the autoencoder baselines use g (for both pre-training and co-training, described later), but for autoencoders, the output dimensionality is the input feature dimensionality, and the mean-squared error reconstruction loss is applied. We train all models and their components with the Adam optimizer using the default learning rate of 0.001. For both pre-training and fine-tuning we use 128 batch size. Unsupervised pre-training methods all use early stopping with patience 3 on the validation loss, unless otherwise noted. Supervised fine-tuning uses this same criterion (and validation split), but classification error is used as the validation metric for early stopping, as it performs slightly better. We set a max number of fine-tune epochs of 200 and pre-train epochs of 1000, We use 10 epochs to build the static validation set. Unless otherwise noted, we use a corruption rate c of 0.6 and a temperature τ of 1, for SCARF-based methods. All runs are repeated 30 times using different train/validation/test splits. Experiments were run on a cloud cluster of CPUs, and we used about one million CPU core hours in total for the experiments.

Evaluation methods. We use the following to effectively convey the results across all datasets.

Win matrix. Given M methods, we compute a "win" matrix W of size $M \times M$, where the (i, j) entry is defined as:

$$W_{i,j} = \frac{\sum_{d=1}^{69} \mathbb{1}[\text{method } i \text{ beats } j \text{ on dataset } d]}{\sum_{d=1}^{69} \mathbb{1}[\text{method } i \text{ beats } j \text{ on dataset } d] + \mathbb{1}[\text{method } i \text{ loses to } j \text{ on dataset } d]}.$$

"Beats" and "loses" are only defined when the means are not a statistical tie (using Welch's t-test with unequal variance and a p-value of 0.05). A win ratio of $0/1$ means that out of the 69 (pairwise) comparisons, only one was significant and it was a loss. Since $0/1$ and $0/69$ have the same value but the latter is more confident indication that i is worse than j, we present the values in fractional form and use a heat map. We add an additional column to the matrix that represents the minimum win ratio across each row.

Box plots. The win matrix effectively conveys how often one method beats another but does not capture the degree by which. To that end, for each method, we compute the relative percent improvement over some reference method on each dataset. We then build box-plots depicting the distribution of the relative improvement across datasets, plotting the observations as small points. We only consider datasets where the means of the method and the reference are different with p-value 0.20. We use a larger p-value here than when computing the win ratio because otherwise some methods would not have enough points to make the box-plots meaningful.

Baselines. We use the following baselines:

[1] `https://docs.openml.org/benchmark/`

- *Label smoothing* [70], which has proved successful for accuracy [52] and label noise [47]. We use a weight of 0.1 on the smoothing term.
- *Dropout*. We use standard dropout [69] using rate 0.04 on all layers. Dropout has been shown to improve performance and robustness to label noise [66].
- *Mixup* [90], using $\alpha = 0.2$.
- *Autoencoders* [65]. We use this as our key ablative pre-training baseline. We use the classical autoencoder ("no noise AE"), the denoising autoencoder [76, 77] using Gaussian additive noise ("add. noise AE") as well as SCARF's corruption method ("SCARF AE"). We use MSE for the reconstruction loss. We try both pre-training and co-training with the supervised task, and when co-training, we add 0.1 times the autoencoder loss to the supervised objective. We discuss co-training in the Appendix as it is less effective than pre-training.
- SCARF *data-augmentation*. In order to isolate the effect of our proposed feature corruption technique, we skip pre-training and instead train on the corrupted inputs during supervised fine-tuning. We discuss results for this baseline in the Appendix as it is less effective than the others.
- *Discriminative* SCARF. Here, our pre-training objective is to discriminate between original input features and their counterparts that have been corrupted using our proposed technique. To this end, we update our pre-training head network to include a final linear projection and swap the InfoNCE with a binary logistic loss. We use classification error, not logistic loss, as the validation metric for early stopping, as we found it to perform slightly better.
- *Self-distillation* [30, 91]. We first train the model on the labeled data and then train the final model on both the labeled and unlabeled data using the first models' predictions as soft labels for both sets.
- *Deep k-NN* [3], a recently proposed method for label noise. We set $k = 50$.
- *Bi-tempered loss* [2], a recently proposed method for label noise. We use 5 iterations, $t_1 = 0.8$, and $t_2 = 1.2$.
- *Self-training* [87, 48]. A classical semi-supervised method – each iteration, we train on pseudo-labeled data (initialized to be the original labeled dataset) and add highly confident predictions to the training set using the prediction as the label. We then train our final model on the final dataset. We use a softmax prediction threshold of 0.75 and run for 10 iterations.
- *Tri-training* [94]. Like self-training, but using three models with different initial labeled data via bootstrap sampling. Each iteration, every model's training set is updated by adding only unlabeled points whose predictions made by the other two models agree. It was shown to be competitive in modern semi-supervised NLP tasks [64]. We use same hyperparameters as self-training.

4.1 SCARF pre-training improves predictive performance

Figure 2 shows our results. From the first win matrix plot, we see that all five pre-training techniques considered improve over no pre-training (control), and that SCARF outperforms the others and has more statistically significant wins. The second win matrix shows that SCARF pre-training boosts the performance of mixup, label smoothing, distillation, and dropout, and it does so better than alternatives. In other words, pre-training *complements* existing solutions, suggesting that a distinct mechanism is at play here. The box plots expand on the second win matrix, showing the relative improvement that each of the pre-training strategies confers over the baselines. Table 1 summarizes the box-plots by the average relative gain. We observe that SCARF generally outperforms the alternatives and adds a 1-2% relative gain across the board.

4.2 SCARF pre-training improves performance in the presence to label noise

To show how pre-training improves model robustness when the data's labels are unreliable, we do as follows. Firstly, label noise robustness is often studied in two distinct settings - (1) when some subset of the training data is guaranteed to be uncorrupted and that set is known in advance, (2) when the entire dataset is untrustworthy. For simplicity, we consider setting 2 in our experiments. We corrupt labels as follows: leaving the validation and test splits uncorrupted, we select a random 30% percent of the training data to corrupt and for each datapoint, we replace its label by one of

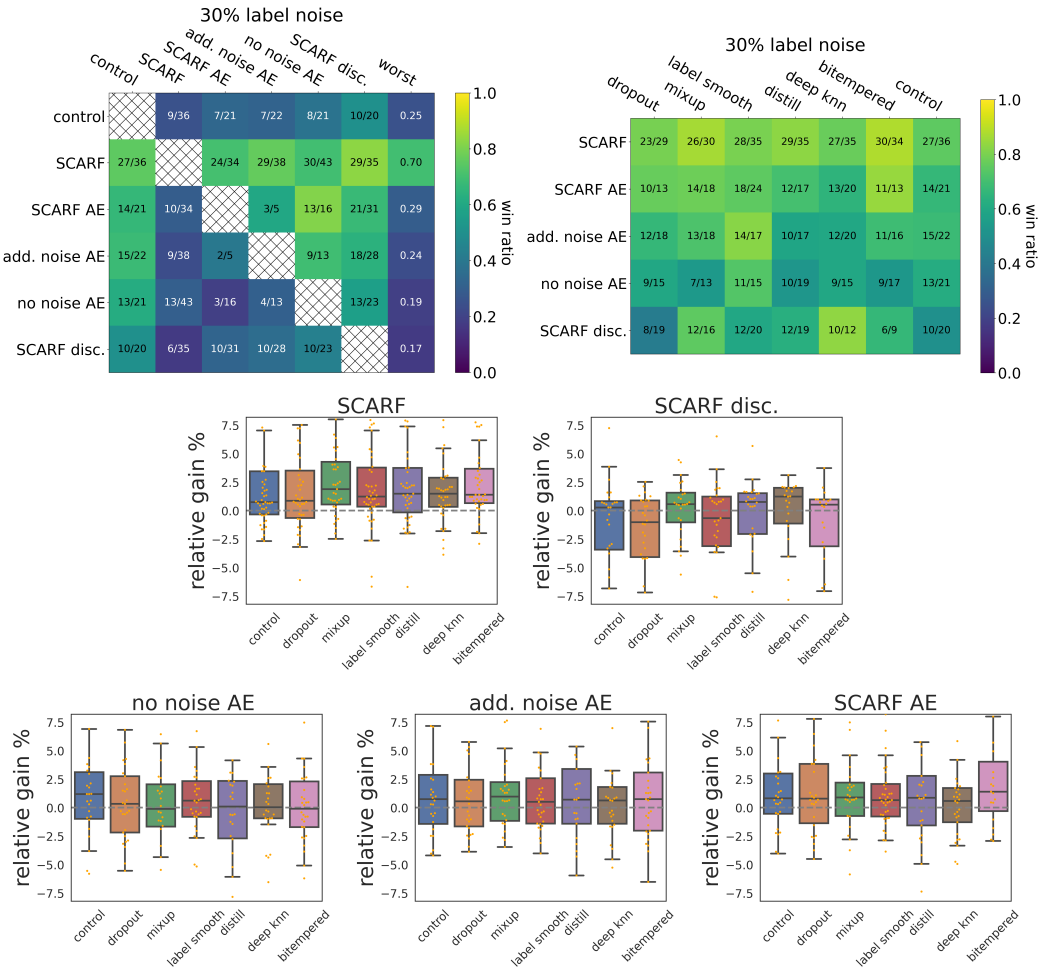

Figure 3: SCARF boosts baseline performance even when 30% of the training labels are corrupted. Notably, it improves state-of-the-art label noise solutions like Deep k-NN.

the classes, uniformly over all classes (including the datapoint's true class). Results are shown in Figure 3 and Table 1. Again, we see SCARF outperforms the rest and boosts all baselines by 2-3%.

4.3 SCARF pre-training improves performance when labeled data is limited

To show how pre-training helps when there is more unlabeled data than labeled ones, we remove labels in the training split so that only 25% of the original split remains labeled. Autoencoders, SCARF, self-training and tri-training all leverage the unlabeled remainder. Results are shown in Figure 4 and Table 1. SCARF outperforms the rest, adding a very impressive 2-4% to all baselines.

4.4 Ablations

We now detail the importance of every factor in SCARF. Due to space, we can only show some of the results here and the rest are in the Appendix.

Other corruption strategies are less effective and are more sensitive to feature scaling. Here, we ablate the marginal sampling corruption technique we proposed, replacing it with the following other promising strategies, while keeping all else fixed.

1. *No corruption.* We do not apply any corruption - i.e. $\tilde{x}^{(i)} = x^{(i)}$, in Algorithm 1. In this case, the cosine similarity between positive pairs is always one and the model is learning to make negative pairs as orthogonal as possible. Under the recent perspective [79] that the

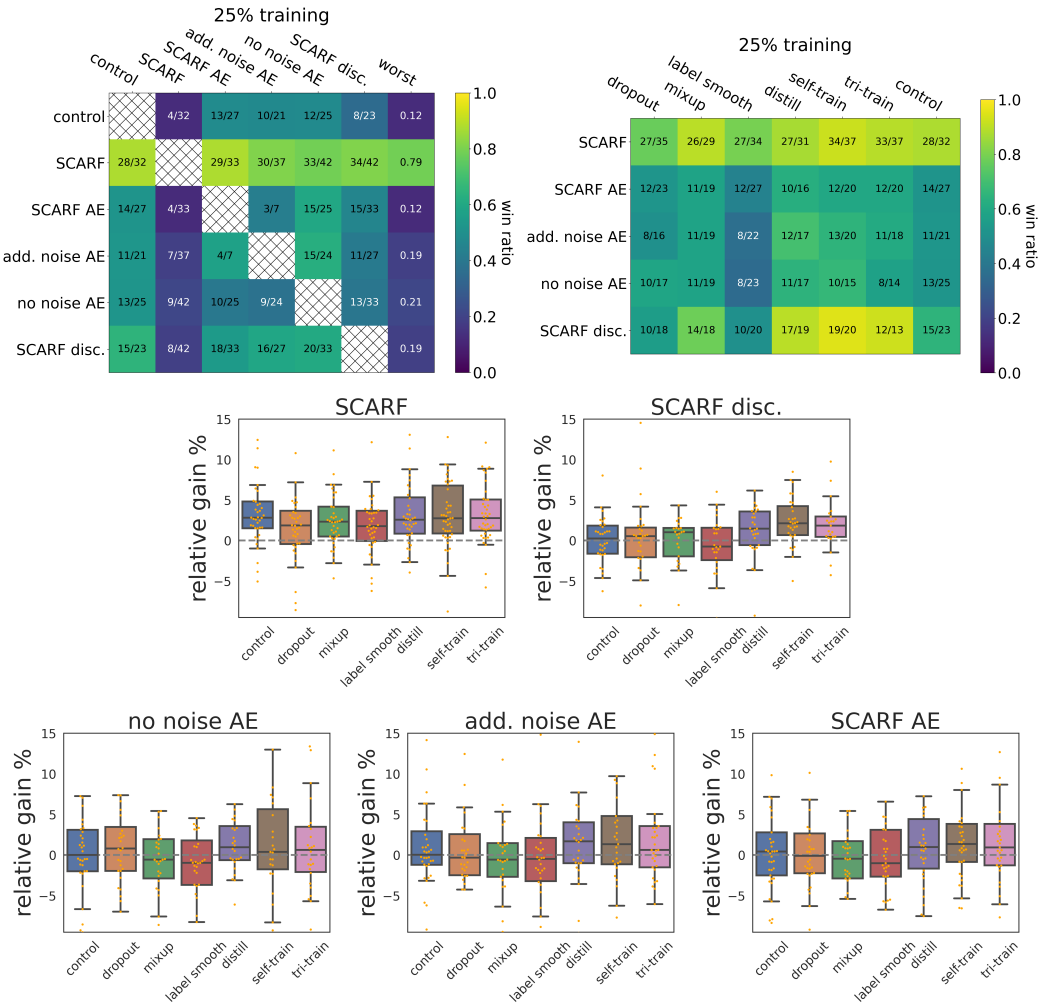

Figure 4: SCARF shows even more significant gain in the semi-supervised setting where 25% of the data is labeled and the remaining 75% is not. Strikingly, pre-training with SCARF boosts the performance of self-training and tri-training by several percent.

contrastive loss comprises two terms – one that encourages alignment between views of the same example – and one that encourages the hypersphere embeddings to be uniformly spread out – we see that with no corruption, pre-training may just be learning to embed input examples uniformly on the hypersphere.

2. *Mean corruption.* After determining which features to corrupt, we replace their entries with the empirical marginal distribution's mean.
3. *Additive Gaussian noise.* We add i.i.d $\mathcal{N}(0, 0.5^2)$ noise to features.
4. *Joint sampling.* Rather than replacing features by random draws from their marginals to form $\tilde{x}^{(i)}$, we instead randomly draw $\hat{x}^{(i)}$ from training data $\mathcal{X}$ – i.e. we draw from the empirical (joint) data distribution – and then set $\tilde{x}_j^{(i)} = \hat{x}_j^{(i)}\ \forall j \in \mathcal{I}_i$.
5. *Missing feature corruption.* We mark the selected features as "missing" and add one learnable value per feature dimension to our model. When a feature is missing, it's filled in with the corresponding learnable missing value.
6. *Feature dropout.* We zero-out the selected features.

We also examine the corruption strategies under different ways of scaling the input features. To that end, in addition to z-score scaling, we look at:

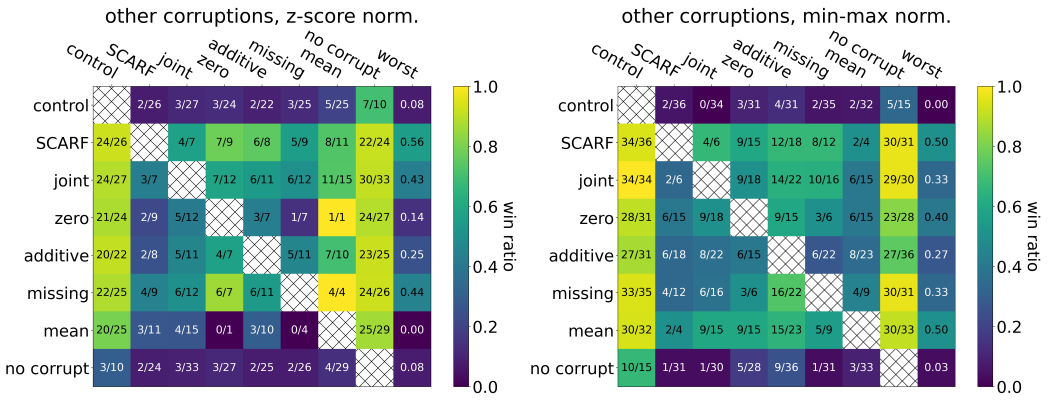

Figure 5: **Left:** Win matrix comparing different corruption strategies when z-score feature normalization is used in the fully labeled, noiseless setting. **Right:** The same matrix but when min-max feature scaling is used. We see that SCARF is better than alternative corruption strategies for different types of feature scaling.

100% labeled training	SCARF	SCARF AE	no noise AE	add. noise AE	SCARF disc.
control	**2.352**	2.244	1.107	1.559	0.574
dropout	**1.609**	1.196	0.623	1.228	-1.312
mixup	**1.72**	1.183	-0.377	0.971	-0.307
label smooth	**1.522**	0.711	-0.002	1.04	-0.894
distill	**2.392**	2.186	0.823	1.431	-0.394
25% labeled training					
control	**3.692**	1.702	0.777	1.662	0.233
dropout	**2.212**	1.848	2.013	1.155	-0.322
mixup	**2.809**	0.73	0.106	0.439	0.466
label smooth	**2.303**	0.705	-0.564	0.196	-0.206
distill	**3.609**	2.441	1.969	2.263	1.795
self-train	**3.839**	2.753	1.672	2.839	2.559
tri-train	**3.549**	2.706	1.455	2.526	1.92
30% label noise					
control	**2.261**	1.988	0.914	1.612	-1.408
dropout	2.004	**2.058**	0.9	1.471	-2.54
mixup	**2.739**	1.723	0.116	1.409	0.189
label smooth	**2.558**	1.474	0.703	1.395	-1.337
distill	**2.881**	2.296	-0.239	1.659	-0.226
deep knn	**2.001**	1.281	0.814	1.348	0.088
bitempered	2.68	**2.915**	0.435	1.387	-1.147

Table 1: Results using the fully labeled training data, only 25% of the labeled training data, and the full training data subject to 30% label noise. Shown is the average relative gain in accuracy when adding the pre-training methods (columns) to the reference methods (rows). Like the box-plots, we filter out datasets using p-value 0.20. We see that SCARF consistently outperforms alternatives, not only in improving control but also in improving methods designed specifically for the setting.

1. *Min-max scaling.* Here, $x_j = [x_j - \min(\mathcal{X}_j)] \,/\, [\max(\mathcal{X}_j) - \min(\mathcal{X}_j)]$.
2. *Mean scaling.* $x_j = [x - \text{mean}(\mathcal{X}_j)] \,/\, [\max(\mathcal{X}_j) - \min(\mathcal{X}_j)]$.

Figure 5 shows the results for z-norm and min-max scaling. SCARF marginal sampling generally outperforms the other corruption strategies for different types of feature scaling. Marginal sampling is neat in that in addition to not having hyperparameters, it is invariant to scaling and preserves the

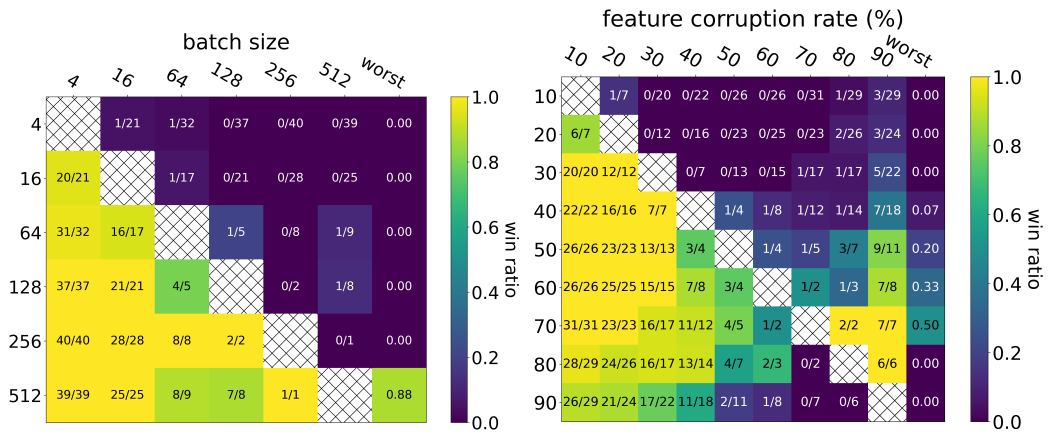

Figure 6: Win matrix for various batch sizes (**Left**) and corruption rates (**Right**) for the fully labeled, noiseless setting.

"units" of each feature. In contrast, even a simple multiplicative scaling requires the additive noise to be scaled in the same way.

SCARF is *not* sensitive to batch size. Contrastive methods like SimCLR [10] have shown consistent improvements upon increasing the batch size, N. There is a tight coupling between the batch size and how hard the contrastive learning task is, since, in our case, the loss term for each example i involves 1 positive and $N-1$ negatives. The need for large (e.g. 5000) batch sizes has motivated engineering solutions to support them [21] and have been seen as grounds for adopting other loss functions [88]. Figure 6 compares a range of batch sizes. We see that increasing the batch size past 128 did not result in significant improvements. A reasonable hypothesis here is that higher capacity models and harder tasks benefit more from negatives.

SCARF is fairly insensitive to the corruption rate and temperature. We study the effect of the corruption rate c in Figure 6. We see that performance is stable when the rate is in the range $50\% - 80\%$. We thus recommend a default setting of 60%. We see a similar stability with respect to the temperature hyperparameter (see the Appendix for more details). We recommend using a default temperature of 1.

Tweaks to the corruption do not work any better. The Appendix shows the effect of four distinct tweaks to the corruption method. We do not see any reason to use any of them.

Alternatives to InfoNCE do not work any better. We investigate the importance of our choice of InfoNCE loss function and see the effects of swapping it with recently proposed alternatives Alignment and Uniformity [79] and Barlow Twins [88]. We found that these alternative losses almost match or perform worse than the original and popular InfoNCE in our setting. See the Appendix for details.

5 Conclusion

Self-supervised learning has seen profound success in important domains including computer vision and natural language processing, but little progress has been made for the general tabular setting. We propose a self-supervised learning method that's simple and versatile and learns representations that are effective in downstream classification tasks, even in the presence of limited labeled data or label noise. Potential negative side effects of this method may include learning representations that reinforce biases that appear in the input data. Finding ways to mitigate this during the training process is a potential direction for future research.

References

[1] Unaiza Ahsan, Rishi Madhok, and Irfan Essa. Video jigsaw: Unsupervised learning of spatiotemporal context for video action recognition. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 179–189. IEEE, 2019.

[2] Ehsan Amid, Manfred K Warmuth, Rohan Anil, and Tomer Koren. Robust bi-tempered logistic loss based on bregman divergences. *arXiv preprint arXiv:1906.03361*, 2019.

[3] Dara Bahri, Heinrich Jiang, and Maya Gupta. Deep k-nn for noisy labels. In *International Conference on Machine Learning*, pages 540–550. PMLR, 2020.

[4] Hangbo Bao, Li Dong, Furu Wei, Wenhui Wang, Nan Yang, Xiaodong Liu, Yu Wang, Jianfeng Gao, Songhao Piao, Ming Zhou, et al. Unilmv2: Pseudo-masked language models for unified language model pre-training. In *International Conference on Machine Learning*, pages 642–652. PMLR, 2020.

[5] Bernd Bischl, Giuseppe Casalicchio, Matthias Feurer, Frank Hutter, Michel Lang, Rafael G Mantovani, Jan N van Rijn, and Joaquin Vanschoren. Openml benchmarking suites. *arXiv preprint arXiv:1708.03731*, 2017.

[6] Piotr Bojanowski and Armand Joulin. Unsupervised learning by predicting noise. In *International Conference on Machine Learning*, pages 517–526. PMLR, 2017.

[7] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[8] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 132–149, 2018.

[9] Ting Chen, Xiaohua Zhai, and Neil Houlsby. Self-supervised gan to counter forgetting. *arXiv preprint arXiv:1810.11598*, 2018.

[10] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[11] Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*, 2020.

[12] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *Journal of machine learning research*, 12(ARTICLE):2493–2537, 2011.

[13] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 702–703, 2020.

[14] Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. Bertje: A dutch bert model. *arXiv preprint arXiv:1912.09582*, 2019.

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[16] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, pages 1422–1430, 2015.

[17] Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. *arXiv preprint arXiv:1605.09782*, 2016.

[18] Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. Unified language model pre-training for natural language understanding and generation. *arXiv preprint arXiv:1905.03197*, 2019.

[19] Matthias Feurer, Jan N. van Rijn, Arlind Kadra, Pieter Gijsbers, Neeratyoy Mallik, Sahithya Ravi, Andreas Mueller, Joaquin Vanschoren, and Frank Hutter. Openml-python: an extensible python api for openml. *arXiv*, 1911.02490, 2019. URL `https://arxiv.org/pdf/1911.02490.pdf`.

[20] Markus Freitag and Scott Roy. Unsupervised natural language generation with denoising autoencoders. *arXiv preprint arXiv:1804.07899*, 2018.

[21] Luyu Gao and Yunyi Zhang. Scaling deep contrastive learning batch size with almost constant peak memory usage. *arXiv preprint arXiv:2101.06983*, 2021.

[22] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018.

[23] Raphael Gontijo-Lopes, Sylvia J Smullin, Ekin D Cubuk, and Ethan Dyer. Affinity and diversity: Quantifying mechanisms of data augmentation. *arXiv preprint arXiv:2002.08973*, 2020.

[24] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014.

[25] 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.

[26] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 297–304. JMLR Workshop and Conference Proceedings, 2010.

[27] Kaveh Hassani and Amir Hosein Khasahmadi. Contrastive multi-view representation learning on graphs. In *International Conference on Machine Learning*, pages 4116–4126. PMLR, 2020.

[28] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.

[29] Olivier Henaff. Data-efficient image recognition with contrastive predictive coding. In *International Conference on Machine Learning*, pages 4182–4192. PMLR, 2020.

[30] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[31] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

[32] Daniel Ho, Eric Liang, Xi Chen, Ion Stoica, and Pieter Abbeel. Population based augmentation: Efficient learning of augmentation policy schedules. In *International Conference on Machine Learning*, pages 2731–2741. PMLR, 2019.

[33] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. *ACM Transactions on Graphics (ToG)*, 36(4):1–14, 2017.

[34] Simon Jenni and Paolo Favaro. Self-supervised feature learning by learning to spot artifacts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2733–2742, 2018.

[35] Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[36] Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77, 2020.

[37] Guillaume Lample and Alexis Conneau. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*, 2019.

[38] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.

[39] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Learning representations for automatic colorization. In *European conference on computer vision*, pages 577–593. Springer, 2016.

[40] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Colorization as a proxy task for visual understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6874–6883, 2017.

[41] Hsin-Ying Lee, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Unsupervised representation learning by sorting sequences. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 667–676, 2017.

[42] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

[43] Yin Li, Manohar Paluri, James M Rehg, and Piotr Dollár. Unsupervised learning of edges. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1619–1627, 2016.

[44] Sungbin Lim, Ildoo Kim, Taesup Kim, Chiheon Kim, and Sungwoong Kim. Fast autoaugment. *arXiv preprint arXiv:1905.00397*, 2019.

[45] 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.

[46] Raphael Gontijo Lopes, Dong Yin, Ben Poole, Justin Gilmer, and Ekin D Cubuk. Improving robustness without sacrificing accuracy with patch gaussian augmentation. *arXiv preprint arXiv:1906.02611*, 2019.

[47] Michal Lukasik, Srinadh Bhojanapalli, Aditya Menon, and Sanjiv Kumar. Does label smoothing mitigate label noise? In *International Conference on Machine Learning*, pages 6448–6458. PMLR, 2020.

[48] David McClosky, Eugene Charniak, and Mark Johnson. Effective self-training for parsing. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 152–159, 2006.

[49] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. *arXiv preprint arXiv:1310.4546*, 2013.

[50] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6707–6717, 2020.

[51] Andriy Mnih and Koray Kavukcuoglu. Learning word embeddings efficiently with noise-contrastive estimation. *Advances in neural information processing systems*, 26:2265–2273, 2013.

[52] Rafael Müller, Simon Kornblith, and Geoffrey Hinton. When does label smoothing help? *arXiv preprint arXiv:1906.02629*, 2019.

[53] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pages 69–84. Springer, 2016.

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

[55] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*, 2019.

[56] Deepak Pathak, Ross Girshick, Piotr Dollár, Trevor Darrell, and Bharath Hariharan. Learning features by watching objects move. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2701–2710, 2017.

[57] Luis Perez and Jason Wang. The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:1712.04621*, 2017.

[58] Senthil Purushwalkam and Abhinav Gupta. Demystifying contrastive self-supervised learning: Invariances, augmentations and dataset biases. *arXiv preprint arXiv:2007.13916*, 2020.

[59] Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, pages 1–26, 2020.

[60] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[61] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*, 2019.

[62] Alexander J Ratner, Henry R Ehrenberg, Zeshan Hussain, Jared Dunnmon, and Christopher Ré. Learning to compose domain-specific transformations for data augmentation. *Advances in neural information processing systems*, 30:3239, 2017.

[63] Zhongzheng Ren and Yong Jae Lee. Cross-domain self-supervised multi-task feature learning using synthetic imagery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 762–771, 2018.

[64] Sebastian Ruder and Barbara Plank. Strong baselines for neural semi-supervised learning under domain shift. *arXiv preprint arXiv:1804.09530*, 2018.

[65] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science, 1985.

[66] Andrzej Rusiecki. Standard dropout as remedy for training deep neural networks with label noise. In *International Conference on Dependability and Complex Systems*, pages 534–542. Springer, 2020.

[67] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mass: Masked sequence to sequence pre-training for language generation. *arXiv preprint arXiv:1905.02450*, 2019.

[68] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. *arXiv preprint arXiv:2004.09297*, 2020.

[69] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[70] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[71] Alex Tamkin, Mike Wu, and Noah Goodman. Viewmaker networks: Learning views for unsupervised representation learning. *arXiv preprint arXiv:2010.07432*, 2020.

[72] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*, 2019.

[73] Toan Tran, Trung Pham, Gustavo Carneiro, Lyle Palmer, and Ian Reid. A bayesian data augmentation approach for learning deep models. *arXiv preprint arXiv:1710.10564*, 2017.

[74] Hsiao-Yu Fish Tung, Hsiao-Wei Tung, Ersin Yumer, and Katerina Fragkiadaki. Self-supervised learning of motion capture. *arXiv preprint arXiv:1712.01337*, 2017.

[75] Joaquin Vanschoren, Jan N. van Rijn, Bernd Bischl, and Luis Torgo. Openml: Networked science in machine learning. *SIGKDD Explorations*, 15(2):49–60, 2013. doi: 10.1145/2641190.2641198. URL `http://doi.acm.org/10.1145/2641190.2641198`.

[76] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103, 2008.

[77] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, and Léon Bottou. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(12), 2010.

[78] Liang Wang, Wei Zhao, Ruoyu Jia, Sujian Li, and Jingming Liu. Denoising based sequence-to-sequence pre-training for text generation. *arXiv preprint arXiv:1908.08206*, 2019.

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

[80] Xiaolong Wang and Abhinav Gupta. Unsupervised learning of visual representations using videos. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2015.

[81] Zhirong Wu, Yuanjun Xiong, Stella Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance-level discrimination. *arXiv preprint arXiv:1805.01978*, 2018.

[82] Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *International conference on machine learning*, pages 478–487. PMLR, 2016.

[83] Dejing Xu, Jun Xiao, Zhou Zhao, Jian Shao, Di Xie, and Yueting Zhuang. Self-supervised spatiotemporal learning via video clip order prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10334–10343, 2019.

[84] Jianwei Yang, Devi Parikh, and Dhruv Batra. Joint unsupervised learning of deep representations and image clusters. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5147–5156, 2016.

[85] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*, 2019.

[86] Tiansheng Yao, Xinyang Yi, Derek Zhiyuan Cheng, Felix Yu, Ting Chen, Aditya Menon, Lichan Hong, Ed H Chi, Steve Tjoa, Jieqi Kang, et al. Self-supervised learning for large-scale item recommendations. *arXiv preprint arXiv:2007.12865*, 2020.

[87] David Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *33rd annual meeting of the association for computational linguistics*, pages 189–196, 1995.

[88] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. *arXiv preprint arXiv:2103.03230*, 2021.

[89] Xiaohua Zhai, Avital Oliver, Alexander Kolesnikov, and Lucas Beyer. S4l: Self-supervised semi-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1476–1485, 2019.

[90] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[91] Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3713–3722, 2019.

[92] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer, 2016.

[93] Xinyu Zhang, Qiang Wang, Jian Zhang, and Zhao Zhong. Adversarial autoaugment. *arXiv preprint arXiv:1912.11188*, 2019.

[94] Zhi-Hua Zhou and Ming Li. Tri-training: Exploiting unlabeled data using three classifiers. *IEEE Transactions on knowledge and Data Engineering*, 17(11):1529–1541, 2005.

A Appendix

We now present findings held out from the main text. Unless noted otherwise, we use the same hyper-parameters (i.e. $c = 0.6$, temperature $\tau = 1$, etc.).

A.1 Pre-training loss curves

Figure 7 shows prototypical pre-training training and validation loss curves for SCARF. We see that both losses tend to decrease rapidly at the start of training and then diminish slowly until the early stopping point. We also notice the noisy (high variance) nature of the training loss - this is due to the stochasticity of our corruption method.

A.2 Ablations Continued

We present more ablation results, where the metric is accuracy using 100% of the labeled data.

Impact of temperature

Figure 8 shows the impact of the temperature term. While prior work considers temperature an important hyperparameter that needs to be tuned, we see that a default of 1 (i.e. just softmax) works the best in our setting.

More corruption ablations

Figure 11 shows the following points.

- Corrupting one view is better than corrupting both the views. The likely explanation for this is that at a corruption rate of 60%, corrupting both views i.i.d means that the fraction of feature indices that contain the same value for both views is small - in other words, there is less information between the two views and the contrastive task is harder.
- Corrupting the same set of feature indices within the mini-batch performs slightly worse than random sampling for every example in the mini-batch.
- An alternative way to select feature indices to corrupt is to use Bernoulli's: for each feature index, corrupt it with probability (corruption rate) c, ensuring that at least one index is corrupted. In the selection method we describe in Algorithm 1, a constant number of indices are corrupted, whereas here it is variable. This alternative way of selecting indices performs roughly the same.
- Once the feature indices to corrupt are determined, rather than corrupting them by drawing from the empirical marginal distribution, we can instead draw a random example from the training set and use its feature entries for corrupting *all* views for the mini-batch. This performs worse.

Alternative losses

Figure 8 shows the impact of using two recently-proposed alternatives to the InfoNCE loss: Uniform and Align [79] and Barlow Twins [88]. We use 5e-3 for the hyperparameter in Barlow, and equal weighting between the align and uniform loss terms. We find no benefit in using these other losses.

SCARF pre-training outperforms SCARF co-training

Here we address the question: is it better to pre-train using SCARF or to apply it as a term in the supervised training loss? In particular, $\mathcal{L}_{\text{co-train}} = \mathcal{L}_{\text{supervised}} + \lambda_{\text{cont}}\mathcal{L}_{\text{cont}}$, where $\mathcal{L}_{\text{cont}}$ is as described in Algorithm 1. Figure 9 shows that pre-training outperforms co-training for a range of different λ_{cont}. We see this is also the case for additive noise autoencoders.

SCARF pre-training outperforms SCARF data augment

Figure 10 shows that using SCARF only for data augmentation during supervised training performs far worse than using it for pre-training.

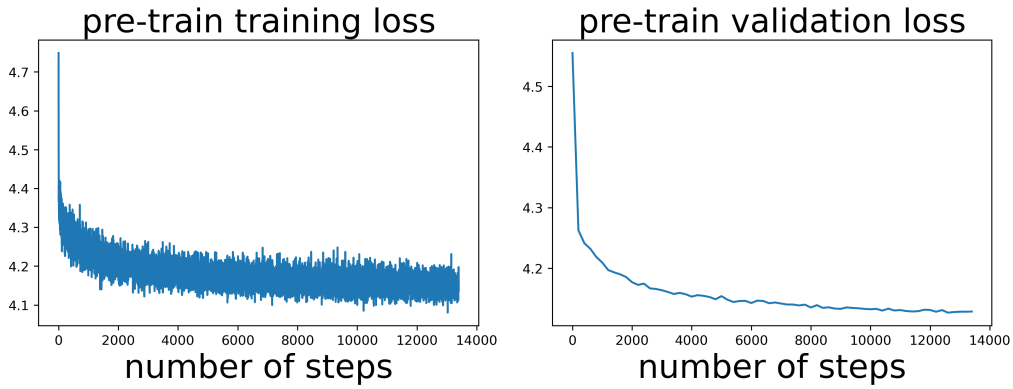

Figure 7: Training and validation loss curves for SCARF pre-training on Phonemes (dataset id 1489). We observe that both losses drop rapidly initially and then taper off. The training curve is jittery because of the random corruptions, but the validation curve isn't because the validation dataset is built once before training and is static throughout training.

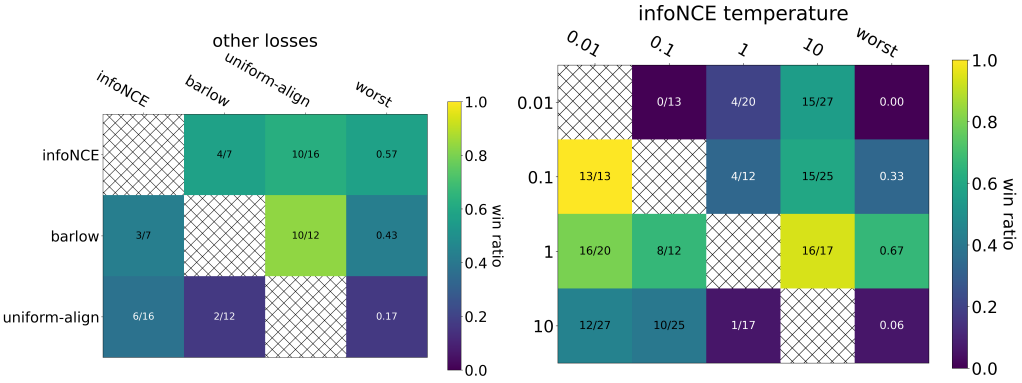

Figure 8: **Left**: Barlow Twins loss performs similar to InfoNCE while Uniform-Align performs worse. **Right**: InfoNCE softmax temperature 1 performs well.

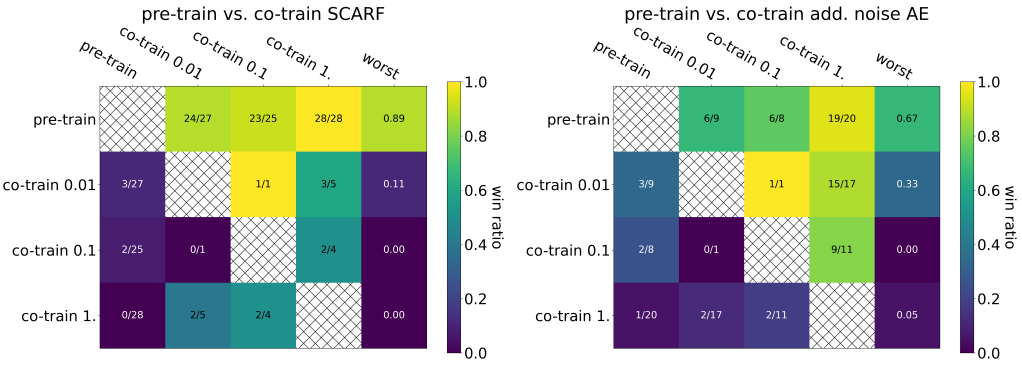

Figure 9: **Left:** Pre-training with SCARF beats co-training for a range of weights on the co-training term. **Right:** The same is true for additive noise autoencoders.

Using InfoNCE error as the pre-training validation metric is worse

SCARF uses the same loss function (InfoNCE) for training and validation. If we instead use the InfoNCE error for validation - where an error occurs when the model predicts an off-diagonal entry of the batch-size by batch-size similarity matrix - downstream performance degrades, as shown in Figure 10.

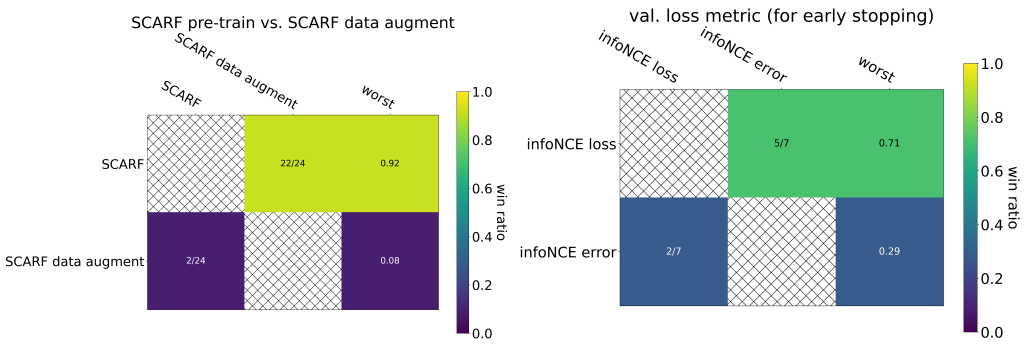

Figure 10: **Left**: Using SCARF only for data augmentation during supervised training performs worse than using it for pre-training. **Right**: Using InfoNCE error instead of InfoNCE loss as the validation metric for early stopping degrades performance.

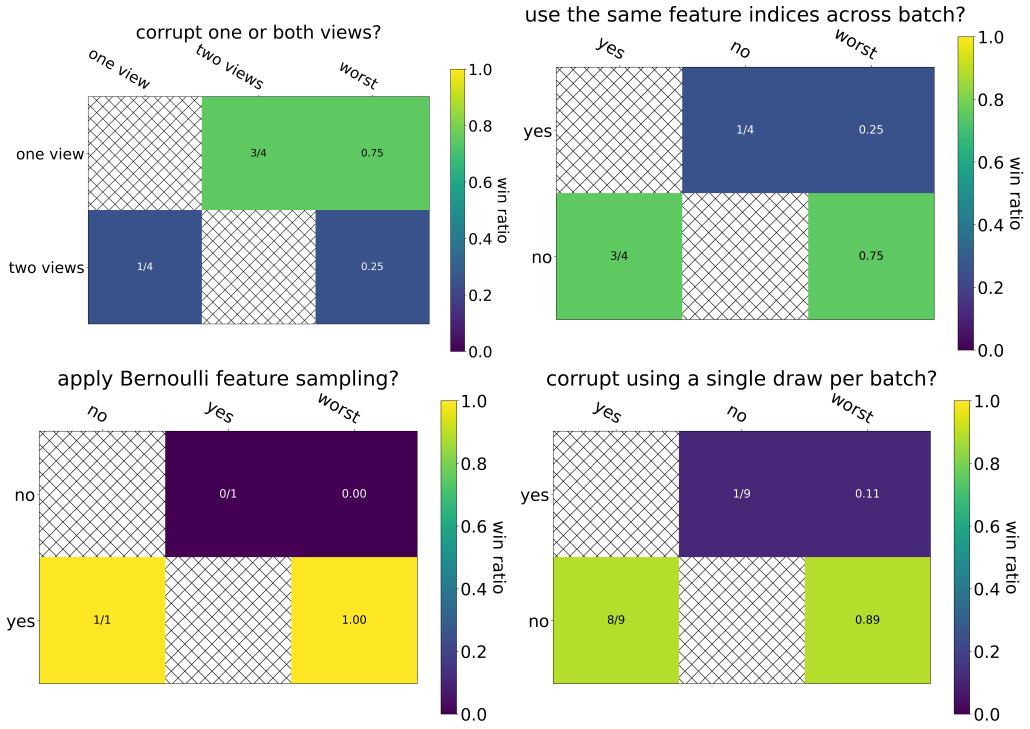

Figure 11: **Top left**: Corrupting one view is better than corrupting both. **Top right**: Using different random feature indices for each example in the mini-batch is better than using a same set across the batch. **Bottom left**: Selecting a variable number of feature indices via coin flips performs similar to the method described in Algorithm 1. **Bottom right**: Corrupting by replacing the features by the features of a *single* drawn example for the whole mini-batch performs worse.

More on feature scaling

Figure 12 compares the different corruption strategies when the input features are transformed using "mean" scaling. SCARF's marginal sampling remains competitive here as well.

More on uniform-align loss

Figure 12 compares SCARF using the uniform-align loss, for different weights between the align and uniform loss terms. The best performance is achieved with equal weight between the two, and

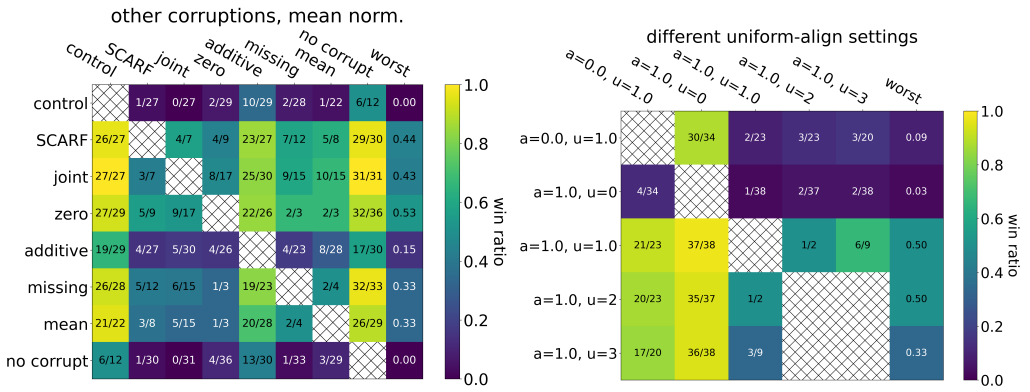

Figure 12: **Left:** SCARF's corruption strategy remains competitive for "mean" feature scaling, as was the case for both z-score and min-max scaling. **Right:** Comparison of different hyperparameters for the uniform-align loss. a and u are the weights on the align and uniform terms respectively.

Figure 8 shows that this underperforms vanilla InfoNCE. We thus recommend using the vanilla InfoNCE loss.