UNDERSTANDING THE ROBUSTNESS OF MULTI-MODAL CONTRASTIVE LEARNING TO DISTRIBUTION SHIFT

Yihao Xue, Siddharth Joshi, Dang Nguyen, Baharan Mirzasoleiman
Department of Computer Science,
University of California, Los Angeles
yihaoxue@g.ucla.edu, sjoshi804@cs.ucla.edu,
nguyentuanhaidang@gmail.com, baharan@cs.ucla.edu

ABSTRACT

Recently, multimodal contrastive learning (MMCL) approaches, such as CLIP (Radford et al., 2021), have achieved a remarkable success in learning representations that are robust against distribution shift and generalize to new domains. Despite the empirical success, the mechanism behind learning such generalizable representations is not understood. In this work, we rigorously analyze this problem and uncover two mechanisms behind MMCL's robustness: *intra-class contrasting*, which allows the model to learn features with a high variance, and *inter-class feature sharing*, where annotated details in one class help learning other classes better. Both mechanisms prevent spurious features that are over-represented in the training data to overshadow the generalizable core features. This yields superior zero-shot classification accuracy under distribution shift. Furthermore, we theoretically demonstrate the benefits of using rich captions on robustness and explore the effect of annotating different types of details in the captions. We validate our theoretical findings through experiments, including a well-designed synthetic experiment and an experiment involving training CLIP models on MSCOCO (Lin et al., 2014)/Conceptual Captions (Sharma et al., 2018) and evaluating them on shifted ImageNets.

1 INTRODUCTION

Learning classifiers that generalize under distribution shifts and across various domains has long been a challenge in machine learning. A key reason is that modern models are highly susceptible to learning simple, domain-dependent spurious features in the training data instead of more complex but generalizable features (Zhu et al., 2016; Sagawa et al., 2019; Xiao et al., 2020; Taori et al., 2020). Recently, Multimodal Contrastive Learning (MMCL) has demonstrated significant robustness in zero-shot image classification. It trains vision and language encoders using a contrastive loss to align representations of paired images and text, while pushing apart the representations of images and texts from different pairs (e.g., CLIP (Radford et al., 2021), ALIGN (Jia et al., 2021)).

Radford et al. (2021) have shown that models trained with CLIP, an exemplar of MMCL algorithms, exhibit better Out-of-Distribution (OOD) generalization (*c.f.* their Fig 13). Specifically, CLIP-trained zero-shot classifiers achieve higher OOD accuracy compared to classifiers with equivalent In-Distribution (ID) accuracy that are trained using various supervised learning techniques, including existing robust methods. Interestingly, the advantage of CLIP diminishes if any label supervision is introduced, e.g., through linear probing with labeled data on CLIP's image encoder (*c.f.* Fig 14 of (Radford et al., 2021)). Both findings suggest that the zero-shot classifier produced by CLIP is more robust to distribution shifts than any classifier trained with label supervision.

However, a comprehensive understanding of the reasons behind does not yet exist in the literature. Existing theoretical studies have only examined MMCL's in-distribution generalization (Nakada et al., 2023; Zhang et al., 2023a), but have not explored its OOD-robustness. In this paper, we aim to explain how CLIP, and more broadly MMCL, produces a zero-shot classifier with superior robustness, while demystifying the contributions of MMCL loss and image captions. This study is conducted by comparing the zero-shot classifier learned through MMCL with classifiers learned via supervised learning. The latter is arguably representative of all other non-zero-shot methods of classifier training, as supervised learning is essentially involved (e.g., end-to-end supervised learning, and linear probing on representations learned by unsupervised/self-supervised algorithms).

Specifically, we demonstrate that the MMCL loss accompanied by rich image captions enables at least two mechanisms providing robustness to zero-shot classification. (1) the *intra-class contrasting* between image-text pairs within the same latent class enables MMCL to easily learn generalizable features with a high variance, when they are annotated by text. In contrast, SL is highly prone to learning simple spurious features instead of the more generalizable features with a higher variance (Sagawa et al., 2020). For example if the majority of cow images with diverse shapes and colors appear on a simple green grass background, SL learns the grass to predict 'cow', but MMCL successfully learns the cow; (2) the *inter-class feature sharing* enabled by MMCL loss allows learning information about a class that only exists and is annotated in other classes. For example, if all the images of the tree class have green leaves but an image in the wolf class has a tree without leaves in its background, MMCL can disassociate the green leaf from the tree class. In contrast, SL cannot leverage this information and learns the green leaves as an indistinguishable part of 'tree'. Hence, it fails to classify trees without green leaves in the test data. Both mechanisms together enable MMCL to learn representations that are robust against distribution shift and generalize to unseen domains.

Furthermore, to emphasize the importance of captions, we analyze the effect of varying caption richness and show that rich captions are essential for achieving robustness. As an extreme case, if the captions are merely labels, no gains in robustness can be achieved.

We further support our theoretical findings with experiments, including a well-designed synthetic experiment and experiments on real datasets, including MSCOCO, Conceptual Captions, and shifted versions of ImageNet. The results demonstrate the crucial roles of the MMCL loss function and rich captions in achieving robustness, further validating our theoretical findings.

2 Related Works

Distribution shift. There is a long line of work on dealing with different types of distribution shift. This includes sub-population shift and domain generalization among others, where distribution of sub-populations in training and test data is different, and some sub-populations may be underrepresented or missing in the training data (Cai et al., 2021; Yang et al., 2023; Santurkar et al., 2020), (Gulrajani & Lopez-Paz, 2020; Joshi et al., 2023; Zhang et al., 2023b; Hu et al., 2020; Fahrbach et al., 2023), or a hybrid of both Koh et al. (2021). Another line of research focuses on evaluating models on natural variations in the source of data collection, with the precise category of shift typically unformalized or unknown. For example, a dataset that contains art and cartoon renditions of ImageNet classes (Hendrycks et al., 2021a), and other variations of ImageNet (Barbu et al., 2019; Recht et al., 2019; Shankar et al., 2021). Despite the diversity of settings, extensive studies (Sagawa et al., 2019; 2020; Xiao et al., 2020; Ilyas et al., 2019) revealed a common theme across these subfields: deep learning models often rely heavily on *spurious correlations* that are specific to the training data but do not hold in general, e.g., those between certain object classes and backgrounds/textures in the image (Zhu et al., 2016; Geirhos et al., 2018).

Multi-modal (contrastive) learning. Learning better representations based on multiple modalities has been a long pursuit (Ngiam et al., 2011; Srivastava & Salakhutdinov, 2012). Numerous methods for learning joint vision-language representations (Li et al., 2019; Lu et al., 2019; Tan & Bansal, 2019; Li et al., 2020; Yao et al., 2021) have emerged. Among them, MMCL (Radford et al., 2021; Jia et al., 2021; Mu et al., 2022; Goel et al., 2022; Pham et al., 2023) has stood out by achieving SOTA performance in various tasks. Notably, (Radford et al., 2021) showed that MMCL on large image-text datasets achieves a significant improvement in robustness to distribution shift. The empirical investigations of (Fang et al., 2022) suggests that this is only attributed to the large diverse image training data, with MMCL loss and text supervision contributing little. We show *provably* that it is not *only* the diverse image data that contributes to superior robustness of MMCL. Indeed, MMCL loss and richness of text annotations are crucial factors.

3 A Framework for Comparing MMCL and SL

In this section, we present a general framework for comparing MMCL and SL, along with the corresponding notations. We start by modeling the multimodal data, and then formalize the MMCL and SL pipelines and their evaluation for robustness to distribution shift. We will formulate and analyze specific types of distribution shift in the next section.

3.1 MODELING MULTIMODAL DATA

To model multimodal data, it is essential to capture the fact that inputs from different modalities can represent the same abstract notion. For instance, both text and an image can represent 'a cow on grass'. We define *underlying feature vectors* to model this abstract notion, and model each input in a specific modality as a projection of the underlying feature vector onto that modality's input space.

Underlying feature . There is an underlying feature space shared among different modalities, where abstract notions reside. We model this as a vector space $\mathbb{R}^l$, where each vector is termed an *underlying feature vector* (e.g., 'a cow on grass'), and each element within the vector is referred to as an *underlying feature* (e.g., 'cow').

Latent classes and labels. Each $\boldsymbol{z}$ is associated with a *latent class*, represented by a *label* y. We note that the labels are only used by SL but not by MMCL.

Inputs in each modality. Each input example in a modality is an instantiation of an abstract notion. We model this as a projection from an underlying feature vector to another space where this modality's inputs live. Formally, let M represent a modality. Given a underlying feature vector $\boldsymbol{z}$, a corresponding input $\boldsymbol{x}_M$ in this modality is generated as: $\boldsymbol{x}_M = \boldsymbol{D}_M \boldsymbol{\mu}_M(\boldsymbol{z}) + \boldsymbol{\xi}_M$, where $\boldsymbol{\mu}_M(\boldsymbol{z}) \in \mathbb{R}^l$ is a random vector that depends on $\boldsymbol{z}$. It can be interpreted as a possibly distorted version of the original feature vector $\boldsymbol{z}$. Note that setting $\boldsymbol{\mu}_M(\boldsymbol{z}) = \boldsymbol{z}$ implies no distortion in the features when represented in this modality. $\boldsymbol{\xi}_M \in \mathbb{R}^{d_M}$ is a random noise drawn from $\mathcal{N}(0, \frac{\sigma_\xi^2}{d_M}\mathbf{I}_{d_M})$. The matrix $\boldsymbol{D}_M \in \mathbb{R}^{d_M \times l}$ ($d_M > l$) is a matrix with orthonormal columns that can be interpreted as a dictionary matrix. It captures the transformation from the lower dimensional feature space to the higher dimensional input space. Different modalities can have different $\boldsymbol{D}_M$ matrices, reflecting the idea that the same underlying feature may be instantiated differently in each modality (e.g., colors are represented differently in images and texts). Modeling modalities as above is consistent with (Nakada et al., 2023).

In this paper, for clarity and illustration, we focus on the popular vision and language modalities. We let I denote the modality for images, and T denote the modality for texts. However, we note that our framework and results directly apply to other modalities.

Distribution shift. We define two joint distributions between underlying features and latent classes: $\mathscr{P}^\star$, representing the 'ground-truth' in the real world, and $\mathscr{P}^{Tr}$, from which our training data are drawn. We let $\mathscr{P}^{Tr}$ exhibit spurious correlations between certain features and latent classes which do not hold in the ground-truth $\mathscr{P}^\star$. This setup captures the underlying reason for the performance drop observed in various types of distribution shift scenarios, as we will discuss in Section 2.

3.2 MULTI-MODAL CONTRASTIVE LEARNING (MMCL)

Unlike traditional supervised learning, MMCL does not see the input data's latent classes, but is instead given pairs of inputs from two modalities and aims to learn the correspondence between them.

Training dataset. The training dataset comprises n image-text pairs, denoted as $\{(\boldsymbol{x}_{I,i}, \boldsymbol{x}_{T,i})\}_{i=1}^n$, where for each index i, both $\boldsymbol{x}_{I,i}$ and $\boldsymbol{x}_{T,i}$ are generated based on the same underlying feature vector $\boldsymbol{z}_i$. In practice, the texts are usually captions accompanying the images. The feature vectors $\{\boldsymbol{z}_i\}_{i=1}^n$ are drawn from the training distribution $\mathscr{C}^{Tr}$.

Linear encoders. The encoders for modalities I and T are denoted as $g_I : \mathbb{R}^{d_I} \to \mathbb{R}^p$ and $g_T : \mathbb{R}^{d_T} \to \mathbb{R}^p$ respectively. We consider linear models for the encoders, given by $g_I(\boldsymbol{x}) = \boldsymbol{W}_I \boldsymbol{x}$ and $g_T(\boldsymbol{x}) = \boldsymbol{W}_T \boldsymbol{x}$, where $\boldsymbol{W}_I \in \mathbb{R}^{d_I \times p}$ and $\boldsymbol{W}_T \in \mathbb{R}^{d_T \times p}$ with $p \geq l$ are the corresponding encoder parameters. Linear encoders are employed widely in previous studies of MMCL (Nakada et al., 2023; Ren & Li, 2023) and general feature learning (Jing et al., 2021; Tian et al., 2021; Ji et al., 2021; Wu et al., 2022; Tian, 2022; Xue et al., 2023), to facilitate the analysis. In Section 6, we will empirically confirm that our findings extend to non-linear models.

Representation learning with MMCL. MMCL learns representations for both modalities in a shared latent space. We consider the linearized contrastive loss function from Nakada et al. (2023):

$$\mathcal{L}_{\text{MMCL}}(\boldsymbol{W}_I, \boldsymbol{W}_T) = \frac{1}{2n(n-1)} \sum_i \sum_{j \neq i} (s_{ij} - s_{ii}) + \frac{1}{2n(n-1)} \sum_i \sum_{j \neq i} (s_{ji} - s_{ii}) + \frac{\rho}{2} \|\boldsymbol{W}_I^\top \boldsymbol{W}_T\|_F^2,$$

where $s_{ij} \coloneqq g_I(\boldsymbol{x}_{I,i})^\top g_T(\boldsymbol{x}_{T,j}) = (\boldsymbol{W}_I \boldsymbol{x}_{I,i})^\top \boldsymbol{W}_T \boldsymbol{x}_{T,j}$ is the similarity (measured by inner product) between representations of an image and a text. This loss encourages the model to *align* each

image-text pair by increasing their representation similarity (s_{ii}) and *contrast* between images and texts that are not paired together by reducing their representation similarity ($s_{ij}, i \neq j$). The last term is a regularization term with $\rho > 0$. The linear loss and its uni-modal counterpart are widely used in analysis of CL, as they closely captures the dynamics of popular contrastive losses (Ji et al., 2021; Tian, 2022; Nakada et al., 2023), such as CLIP, as we will experimentally confirm in Section 6.

Prompts for zero-shot classification. We test the model's capability in zero-shot classification, where a text prompt is created for each label (e.g., 'a photo of a *dog*'), and the prediction is determined by the prompt with the highest representation similarity with the given image. To formalize this, we define the prompt $\boldsymbol{p}_y$ for each latent class y as $\boldsymbol{p}_y = \boldsymbol{D}_T \bar{\boldsymbol{z}}_y$, where $\bar{\boldsymbol{z}}_y := \mathbb{E}_{(\boldsymbol{z},y)\sim\mathscr{P}^\star}[\boldsymbol{z}|y]$. That is, the prompt is 'the center of all underlying feature vectors with label y in the true distribution' represented in modality T. This closely resembles real world practices where the representation of multiple texts with engineered templates like `'a bad photo of a {}'`, `'a good photo of a {}'` are averaged (Radford et al., 2021).

Robustness evaluation. Given two encoders g_I and g_T with parameters $\boldsymbol{W}_I$ and $\boldsymbol{W}_T$, respectively, we evaluate the zero-shot performance on the true distribution $\mathscr{P}^\star$. Given an image $\boldsymbol{x}_I$, the prediction is $\hat{y}(\boldsymbol{x}_I) = \arg\max_y g_I(\boldsymbol{x}_I)^\top g_T(\boldsymbol{p}_y)$. The test accuracy, denoted by $\mathrm{Acc}^{\mathrm{MMCL}}_{\mathscr{P}^\star}(\boldsymbol{W}_I, \boldsymbol{W}_T)$, is

$$\mathrm{Acc}^{\mathrm{MMCL}}_{\mathscr{P}^\star}(\boldsymbol{W}_I, \boldsymbol{W}_T) = \mathbb{E}_{(\boldsymbol{z},y)\in\mathscr{P}^\star, \boldsymbol{x}_I = \boldsymbol{D}_I\boldsymbol{\mu}(\boldsymbol{z})+\boldsymbol{\xi}_I}[\mathbb{1}(\hat{y}(\boldsymbol{x}_I) = y)], \tag{1}$$

where $\mathbb{1}(\cdot)$ denotes the indicator function.

Relation to feature cross-covariance. We utilize the connection between the cross-covariance between images and captions, and the MMCL objective for our analysis.

Definition 3.1. *We define $\boldsymbol{C}^{Tr}$ as the cross-covariance between images' and texts' feature vectors $\boldsymbol{C}^{Tr} := \mathbb{E}_{\boldsymbol{z}\in\mathscr{C}^{Tr}, \boldsymbol{\mu}_I(\boldsymbol{z}), \boldsymbol{\mu}_T(\boldsymbol{z})}[\boldsymbol{\mu}_I(\boldsymbol{z})\boldsymbol{\mu}_T(\boldsymbol{z})^\top]$. When $\boldsymbol{\mu}_I(\cdot)$ and $\boldsymbol{\mu}_T(\cdot)$ both are identity, $\boldsymbol{C}^{Tr}$ is the covariance of the original feature vector.*

Lemma 3.2 (Informal)**.** *Given an image with feature $\boldsymbol{\mu}'$ and a text with feature $\boldsymbol{\mu}''$, the similarity (inner product of representations) between them, computed using encoders trained on the training set, is: **similarity score** $\approx \boldsymbol{\mu}'^\top \boldsymbol{C}^{Tr} \boldsymbol{\mu}'' = \sum_{i=1}^{l}\sum_{j=1}^{l} C^{Tr}_{ij}\mu'_i\mu''_j$.*

That is, the image-text similarity is a weighted sum of products between the features in image and text inputs. The weights are determined by the feature cross-covariance matrix of training data, whose i,j-th element is the covariance between feature i in images and feature j in texts.

Importance of zero-shot. We emphasize that using zero-shot classification instead of training a linear classifier on the representations is crucial for achieving robustness in MMCL. The latter essentially involves SL, which falls short for the same reasons as shown in our analysis for SL in Section 4.

3.3 Supervised Learning (SL)

Standard SL has access to each input's label and the inputs are from a single modality (i.e., images). Let $\{(\boldsymbol{x}_{I,i}, y_i)\}_{i=1}^n$ be the training dataset with n inputs $\boldsymbol{x}_{I,i}$ and their labels y_i, we train a linear model $f(\boldsymbol{x}) = \boldsymbol{W}^\top \boldsymbol{x}$ with weights $\boldsymbol{W} \in \mathbb{R}^{d_I \times q}$, where $q = 1$ for binary classification and $q = \#$classes for multi-class classification. We consider minimizing logistic loss for binary classification, and Cross-Entropy loss for multiclass classification, with gradient descent at a sufficiently small step size.

Robustness evaluation. Given a model with weights $\boldsymbol{W}$, the accuracy, denoted by $\mathrm{Acc}^{\mathrm{SL}}_{\mathscr{P}^\star}(\boldsymbol{W})$, is evaluated on the true distribiution $\mathscr{P}^\star$ as $\mathrm{Acc}^{\mathrm{SL}}_{\mathscr{P}^\star}(\boldsymbol{W}) = \mathbb{E}_{(\boldsymbol{z},y)\in\mathscr{P}^\star, \boldsymbol{x}_I = \boldsymbol{D}_I\boldsymbol{\mu}(\boldsymbol{z})+\boldsymbol{\xi}_I}[\mathbb{1}(\hat{y}(\boldsymbol{x}_I) = y)]$, where $\hat{y}(\boldsymbol{x}_I) = \mathrm{sign}(\boldsymbol{W}^\top \boldsymbol{x}_{I_j})$ for binary classification, and $\hat{y}(\boldsymbol{x}_I) = \arg\max_j [\boldsymbol{W}^\top \boldsymbol{x}_I]_j$ for multi-class classification, with $[\boldsymbol{W}^\top \boldsymbol{x}_I]_j$ denoting the j-th element in the vector $\boldsymbol{W}^\top \boldsymbol{x}_I$.

4 Two Mechanisms behind the Robustness of MMCL

Next, we explore two scenarios illustrating MMCL's superior robustness to distribution shift, compared to SL. First, we consider the scenario where generalizable core feature has a higher variance than domain-dependent spurious feature. Then, we consider the data distribution where each latent class has a core feature, that co-occurs with a strong spurious feature in the training data. These features can occur in other latent classes as well, independently of each other. For clarity, we set $\boldsymbol{\mu}_I(\boldsymbol{z}) = \boldsymbol{z}, \boldsymbol{\mu}_T(\boldsymbol{z}) = \boldsymbol{z}, \forall \boldsymbol{z} \in \mathbb{R}^l$ in this section.

4.1 Robustness via Intra-class Contrasting

We start by analyzing the first scenario which illustrates how MMCL can learn features that are challenging for SL to learn. Consider the case where the majority or all images of a 'cow' appear on 'grass'. Here, grass is a spurious feature with high correlation with cow. Grass is often a simple green surface without a high variation. But, cows can vary a lot in their appearance. This makes cows more difficult to learn than grass. Below, we will formalize this scenario and demonstrate that SL learns the spurious feature (grass) but MMCL learns the generalizable feature (cow) and obtains a superior robustness.

4.1.1 Distribution of features

The following definition simulates the aforementioned scenario.

Definition 4.1 (**Data Model** 1)**.** *In both $\mathscr{P}^\star$ and $\mathscr{P}^{Tr}$, each label y is uniformly drawn from $\{-1, 1\}$ and the corresponding feature vector $\boldsymbol{z} \in \mathbb{R}^2$ is generated as $\boldsymbol{z} = [z_{core}, z_{spu}]^T$ where $z_{core} \sim \mathcal{N}(y, \sigma_{core}^2)$, represents the core feature that contains information of the label y, and $z_{spu} \sim \mathcal{N}(a, \sigma_{spu}^2)$. In the true distribution $\mathscr{P}^\star$, a is uniformly drawn from $\{-1, 1\}$ and is independent of the label y, making the feature z_{spu} irrelevant to the label. However, in the training distribution $\mathscr{C}^{Tr}$, there is a strong correlation between a and y, s.t. $\Pr(a = y) = p_{spu}$, where $1 \geq p_{spu} > 1/2$.*

Recall from Section 3.1 that the inputs in each modality are generated based on feature vectors. In SL, where we have only one modality, the situation becomes equivalent to the one analyzed in (Sagawa et al., 2020). Similar variants are studied in (Wald et al., 2021; Aubin et al., 2021; Yao et al., 2022) to investigate distribution shift and out-of-domain generalization. Despite its simplicity, this setup reflects key aspects of general distribution shift. Here, z_{core} is the core feature and z_{spu} is the spurious feature, such as 'grass' in the aforementioned example, or texture/backgrounds in ImageNet.

We assume that the core feature has a larger variance than the spurious feature, indicated by the values of σ_{core} and σ_{spu}. This is detailed in below, along with some additional assumptions.

Assumption 4.2. *The gap between the variances of the core and spurious features is significant: $\sigma_{core} = \Theta(1)$, $\sigma_{core} \geq 1$ and $\sigma_{spu} = O(\frac{1}{\sqrt{\log n}})$. The spurious correlation is large: $p_{spu} = 1 - o(1)$. We consider the high-dimensional (overparameterized) setting where $n = \omega(1)$, $d_I = \Omega(n)$ and $d_T = \Omega(n)$. The noise levels are not too large: $\sigma_{\xi,I} = O(\log n)$ and $\sigma_{\xi,T} = O(\log n)$.*

4.1.2 Comparing Robustness of SL and MMCL

Under Assumption 4.2, SL tends to associate labels mostly with the spurious feature, as they appear to be more stable and reliable for prediction compared to the core feature. This results in low accuracy when tested on the ground-truth distribution, as demonstrated in the following theorem.

Theorem 4.3 (Theorem 1 from (Sagawa et al., 2020))**.** *Let $\boldsymbol{W}^*$ represent the model trained using SL as described in Section 3.3. Assuming that Assumption 4.2 holds, and n and d_I are sufficiently large, with a high probability, the accuracy of $\boldsymbol{W}^*$ on the true distribution satisfies $Acc^{SL}_{\mathscr{P}^\star}(\boldsymbol{W}^*) \leq 2/3$. Additionally, the model's test accuracy on examples where $a \neq y$ is $\leq 1/3$, worse than random chance.*

Next, we examine MMCL. From Lemma 3.2, we know that the similarity between an image with feature $\boldsymbol{z}$ and a text with feature $\boldsymbol{z}'$ is approximately $[z_{\text{core}}\ z_{\text{spu}}] \begin{bmatrix} 1+\sigma_{\text{core}}^2 & 2p_{\text{spu}}-1 \\ 2p_{\text{spu}}-1 & 1+\sigma_{\text{spu}}^2 \end{bmatrix} \begin{bmatrix} z'_{\text{core}} \\ z'_{\text{spu}} \end{bmatrix}$, showing that the variance of features ensures that image and text features, that share the underlying core features, have a higher similarity score. Furthermore, if we let $\boldsymbol{z}'$ be the feature $\bar{\boldsymbol{z}}_{y'} = [y'\ \ 0]^\top$ in label y''s corresponding prompt $\boldsymbol{p}_{y'}$, we deduce that the similarity to the prompt is approximately $(1+\sigma_{\text{core}}^2)y'z_{\text{core}} + (2p_{\text{spu}}-1)y'z_{\text{spu}}$. Here, the core feature carries more weight when the variance is large. In essence, since the MMCL loss contrasts images and unpaired texts in the same latent classes, learning features that have high variance is encouraged; this is in contrast with SL, where features that have low variance are preferred. With the above observation, after bounding the effect of noise, we arrive at the following theorem (with proof in Appendix C.2).

Theorem 4.4. *Let $\boldsymbol{W}_I^*$ and $\boldsymbol{W}_T^*$ be the weights of the encoders trained using MMCL as described in Section 3.2. Under Assumption 4.2 [1], with a high probability of at least $1 - O(\frac{1}{\text{poly}(n)}) = 1 - o(1)$,*

[1]The theorem holds under more relaxed assumptions about the variances and spurious correlation level; see details in Appendix C.2), but here we use Assumption 4.2 to keep consistency with Theorem 4.3

the encoders achieve the following zero-shot accuracy on the true distribution

$$Acc^{MMCL}_{\mathscr{P}^\star}(\boldsymbol{W}_I^*, \boldsymbol{W}_T^*) \geq 1 - \frac{1}{2}\boldsymbol{\Phi}(\kappa_1) - \frac{1}{2}\boldsymbol{\Phi}(\kappa_2) - o(1),$$

where $\kappa_1 = \frac{2p_{spu} - 2 - \sigma_{core}^2}{\sqrt{(1+\sigma_{core}^2)^2\sigma_{core}^2 + (2p_{spu}-1)^2\sigma_{spu}^2}}$, $\kappa_2 = \frac{-2p_{spu} - \sigma_{core}^2}{\sqrt{(1+\sigma_{core}^2)^2\sigma_{core}^2 + (2p_{spu}-1)^2\sigma_{spu}^2}}$ *and* $\boldsymbol{\Phi}$ *denotes the CDF of the standard normal distribution. Meanwhile, the model's test accuracy on examples where* $a \neq y$ *is lower bounded by* $1 - \boldsymbol{\Phi}(\kappa_1) - o(1)$.

Corollary 4.5. *With* $\sigma_{core} = 1$*, for sufficiently large* n*,* $Acc^{MMCL}_{\mathscr{P}^\star}(\boldsymbol{W}_I^*, \boldsymbol{W}_T^*) \geq 81\%$*, Moreover, in this case, no model can achieve an accuracy higher than 85%.*

This, compared with Theorem 4.3, demonstrates that MMCL can outperform SL by a large margin, and comes close to achieving the best possible accuracy of 85%.

Additionally, in terms of performance on examples where the spurious correlation does not hold, i.e., $a \neq y$, it's evident that MMCL excels. As Theorem 4.3 shows, SL's accuracy is even worse than random chance. In contrast, Theorem 4.4 demonstrates that MMCL consistently performs better than random chance. It maintains random chance even in the worst-case scenario, as indicated by $\boldsymbol{\Phi}(\kappa_1) \leq \frac{1}{2}$, owing to $2p_{\text{spu}} - 2 - \sigma_{\text{core}}^2 \leq 0$. When $\sigma_{\text{core}} = 1$, it achieves an accuracy of 69%.

4.2 Robustness via Inter-class Feature Sharing

Next, consider the second scenario and demonstrate how MMCL benefits from annotated details in some latent classes to disassociate spurious correlations in other latent classes, while SL fails to grasp these details. For example, typical images of a 'tree' have green leaves. However, trees in in the background of images of 'wolf' or 'ski resort' may appear without leaves. SL, which only observes the labels, tends to overlook the trees without leaves as they do not contribute to learning 'wolf' and 'ski resort', thus incorrectly correlating trees with the color green. In contrast, in MMCL, if the trees without leaves are annotated in the captions, the model can disassociate the green leaves from tree.

4.2.1 Distribution of features

We first present the underlying feature distributions and then compare MMCL's robustness with SL.

Definition 4.6 (**Data Model** 2)**.** ***True distribution*** $\mathscr{P}^\star$**.** *We have* $2m$ *latent classes in total, with labels* $1, \ldots, 2m$*. For each label* y*, we define a unique alias* (k, c)*:* $k = \lfloor (y+1)/2 \rfloor$*, and* $c = 1$ *if* y *is odd, and* $c = -1$ *if* y *is even. The label is sampled uniformly. Let* $\beta \in [0, 1)$*. Given a label alias* (k, c)*, the corresponding feature vector* $\boldsymbol{z} = [z_1, z_2, \ldots, z_{2m}]^\top$ *is generated as:*

$$\begin{array}{llll} \forall j \leq m, & \text{if } j = k \text{ then } z_j = c & & \text{if } j \neq k \text{ then } z_j \sim U(\{-\beta, +\beta\}) \\ \forall j > m, & \text{if } j = k+m \text{ then } z_j \sim U(\{-\alpha, +\alpha\}) & & \text{if } j \neq k+m \text{ then } z_j \sim U(\{-\beta\alpha, \beta\alpha\}) \end{array}$$

where $U(S)$ *denotes the uniform distribution over set* S*.*

Training distribution $\mathscr{P}^{Tr}$**.** *The training distribution is similar to the true distribution, but with* z_{k+m} *always equal to* $c\alpha$*, making it appear as if the* $k+m$*-th coordinate also indicates the label.*

Here, each feature vector in latent class (k, c) (e.g., 'tree') has a core feature at coordinate k (characteristics of a tree) and a spurious feature at coordinate $k+m$ that correlates with the latent class in the training distribution but not the true distribution (e.g., the color green). With a large α, such a spurious feature has a larger magnitude than the true feature, making it easier to be learned. There are also other features at different coordinates that do not correlate with the label; these features are weaker (indicated by $\beta < 1$) so that they do not change the latent class. One observation is that examples in latent class other than (k, c) would show no correlation between the k-th and $k+m$-th features, hinting at their independence from each other (e.g., trees are not necessarily green). We will show that unlike SL, MMCL can leverage such a hint to obtain a superior robustness.

4.2.2 Comparing Robustness of SL and MMCL

The theorem below demonstrates that SL achieves a low accuracy under distribution shift when the spurious feature is strong, i.e., when α is large.

Theorem 4.7. *Assuming that the input noise in each modality is zero, i.e., $\sigma_{\xi,I} = \sigma_{\xi,T} = 0$, and all possible feature vectors in $\mathscr{P}^{Tr}$ uniformly appear in the training dataset.* [2] *Let $\boldsymbol{W}^*$ be the model trained using SL as described in Section 3.3. The accuracy on the true distribution has the following upper bound: $Acc^{SL}_{\mathscr{P}^\star}(\boldsymbol{W}^*) \leq 50\% + \frac{2}{(1+\alpha^2)(1-\beta)^2-8}$. For example, if $\alpha = 10$ and $\beta = 1/3$, then $Acc^{SL}_{\mathscr{P}^\star}(\boldsymbol{W}^*) \leq 60\%$.*

Next, we will examine how MMCL leverages the information about independence of core and spurious features in each latent class, which is hidden in other latent classes. First, recall the conclusion in Lemma 3.2, and obtain that the similarity between an image with features $\boldsymbol{z}$ and a text with features $\boldsymbol{z}'$ is given by $\boldsymbol{z}^\top \begin{bmatrix} \frac{1+(m-1)\beta^2}{m}\mathbf{I}_m & \frac{\alpha}{m}\mathbf{I}_m \\ \frac{\alpha}{m}\mathbf{I}_m & \frac{1+(m-1)\beta^2}{m}\alpha^2\mathbf{I}_m \end{bmatrix} \boldsymbol{z}'$. The fact that β only appears on the diagonal and not on off-diagonal elements shows that the occurrence of core feature of a given latent class in other latent classes increases the weight for same-feature products between images and texts rather than different-feature products. For example, trees without green leaves in classes other than tree increase the covariance between texts and images of tree, but do not contribute to the correlation between tree and green. Hence appearance of green in any image has a limited impact on its similarity to a text describing a tree. More precisely, when computing the similarity between a given image and the prompt for a tree, a weight $\frac{1+(m-1)\beta^2}{m}$ is assigned to 'the true characteristic of a tree' and a weight $\frac{\alpha}{m}$ is assigned to 'green'. Here, a larger β leads to more weight placed on the core feature, highlighting how MMCL utilizes shared features between classes to enhance robustness. This insight leads us to the following theorem demonstrating the superior performance of MMCL under distribution shift.

Theorem 4.8. *Under the same assumption as in Theorem 4.7 Let $\boldsymbol{W}_I^*$ and $\boldsymbol{W}_T^*$ be the weights of encoders trained using MMCL as described in Section 3.2. Then as long as $\beta^2 m > \frac{\alpha^2(1+\beta)}{1-\beta} - 1 + \beta^2$, the model has 100% zero-shot accuracy on the true distribution, i.e., $Acc^{MMCL}_{\mathscr{P}^\star}(\boldsymbol{W}_I^*, \boldsymbol{W}_T^*) = 100\%$.*

We also observe that if the features were not shared between classes, i.e., $\beta = 0$, it would be impossible for the model to achieve such performance. This once again emphasizes the role of shared features.

Important Consideration about Robustness An important question is whether the improvement in accuracy under distribution shift is solely due to MMCL's improvement in in-distribution generalization. In Appendix E, we demonstrate that we control for in-distribution generalization in both theoretical examples. Specifically, in Data Model 1, SL has slightly better in-distribution accuracy, while in Data Model 2, both SL and MMCL achieve 100% in-distribution accuracy. Thus, MMCL's improvement solely results from enhanced robustness, and in fact, both relative and effective robustness as defined in Taori et al. (2020).

5 Understanding the Benefit of Rich Image Captions

In Section 4, we assumed that both $\boldsymbol{\mu}_I(\cdot)$ and $\boldsymbol{\mu}_T(\cdot)$ are identity, implying that the captions mentioned everything depicted in the image. However, in practice, captions often serve as annotations or illustrations accompanying the image, with certain details omitted. Empirical evidence suggests that rich captions are generally beneficial (Santurkar et al., 2022; Nguyen et al., 2023), but it remains unclear if richness of captions can affect robustness and, if so, how. In this section, we theoretically investigate this question by varying how much and what information is mentioned in captions. Specifically, we keep $\boldsymbol{\mu}_I(\cdot)$ as an identity function, while let $\boldsymbol{\mu}_T(\boldsymbol{z})$ represent a masked version of the original feature vector $\boldsymbol{z}$, where some information may not be reflected in caption.

Benefits of mentioning variations in the core features. Recall that in Section 4.1, utilizing Data Model 1 (Definition 4.1), we showed that MMCL can learn large-variance core features better than SL, resulting in less reliance on the spurious feature. Now, we use the same data model to explore what happens if the feature variance is not fully reflected in the captions. For example, when the caption only contains the word 'cows' or 'grass', without describing their appearance.

Definition 5.1 (Feature masking in data model 1 (Definition 4.1))**.** *Given a feature vector $\boldsymbol{z} = \begin{bmatrix} z_{core} \\ z_{spu} \end{bmatrix}$ with corresponding y and a, we let $\boldsymbol{\mu}_T(\boldsymbol{z}) = \begin{bmatrix} y + \psi_{core}(z_{core} - y) \\ a + \psi_{spu}(z_{spu} - a) \end{bmatrix}$, with ψ_{core} drawn from Bernoulli(π_{core}) and ψ_{spu} drawn from Bernoulli(π_{spu}). Both π_{core} and π_{spu} are $\in [0, 1]$.*

[2] Assumptions are made to simplify the analysis, but our analysis can be readily extended to show that same conclusions holds with high probability in broader settings with sufficient sample size and reasonable noise level.

Here, $z_{\text{core}} - y$ and $z_{\text{spu}} - a$ represent the variations in the core and spurious features, both of which are Gaussian random variables by Definition 4.1. This implies that the captions capture these variations with probabilities π_{core} and π_{spu}, respectively. When $\pi = 0$, caption ignores all the details and treats all features of the same kind as a single entity. The following theorem shows the effects of π_{core}, π_{spu}.

Theorem 5.2. *With data from data model 1 and $\boldsymbol{\mu}_T$ defined in Definition 5.1, with a high probability, the model trained using MMCL has a test accuracy on examples where the spurious correlation does not hold (i.e., $a \neq y$) given by $1 - \boldsymbol{\Phi}(\frac{2p-2-\pi_{core}^2\sigma_{core}^2}{\sqrt{(1+\pi_{core}^2\sigma_{core}^2)^2\sigma_{core}^2+(2p-1)^2\sigma_{spu}^2}}) \pm o(1)$. The non-negligible part of this accuracy increases as π_{core} increases and is independent of π_{spu}.*

The theorem reveals that the model exhibits less reliance on the spurious correlation when the caption mentions the variance in the core feature (e.g., appearance of the cow in each specific image). Additionally, we notice that mentioning variance in the spurious feature has minimal effect on the robustness, as it does not impact the correlation with the core feature.

Mentioning more features benefits robustness. Next, we utilize data model 2 to explore the effect of mentioning more features in the captions.

Definition 5.3 (Feature masking in data model 2 (Definition 4.6))**.** *For a feature vector $\boldsymbol{z}$ with label (k, c), let $\boldsymbol{\mu}_T(\boldsymbol{z}) = \boldsymbol{\psi} \odot \boldsymbol{z}$, where $\boldsymbol{\psi} = [\psi_1 \dots \psi_l]^\top$ with $\psi_k = 1$ and $\psi_j \sim$ Bernoulli(π) for $j \neq k$.*

Here, the caption always mentions the feature indicating the latent class, while other features are mentioned with a probability π. Note that $\pi = 0$ corresponds to the setting where the caption is just the same as the label. The following theorem demonstrates that the model can achieve robustness only when the caption sufficiently mentions features that are not directly related to the image's latent class.

Theorem 5.4. *With data model 2 and $\boldsymbol{\mu}_T$ defined in Definition 5.3, let $\boldsymbol{W}_I^*$ and $\boldsymbol{W}_T^*$ be the weights of encoders trained using MMCL. Then the model's accuracy on the true distribution satisfies $Acc_{\mathscr{D}^\star}^{MMCL}(\boldsymbol{W}_I^*, \boldsymbol{W}_T^*) = 100\%$ if $\pi > \tilde{\pi}$, and $Acc_{\mathscr{D}^\star}^{MMCL}(\boldsymbol{W}_I^*, \boldsymbol{W}_T^*) \leq 50\%$ if $\pi < \tilde{\pi}$, where $\tilde{\pi} := \frac{(1+\beta)\alpha^2-1+\beta}{(1-\beta)\beta^2(m-1)}$.*

As explained in Section 4.2, even if certain features do not directly indicate labels for a class, they can still help learn relationships between features (for example, not all trees are green), and this knowledge can be valuable for other classes. However, if these features are missing from the captions, they contribute less to the cross-covariance matrix used by the model for predictions (Lemma 3.2). In the extreme case where $\pi = 0$, captions reduce to labels used by SL, and robustness does not improve.

6 Experiments

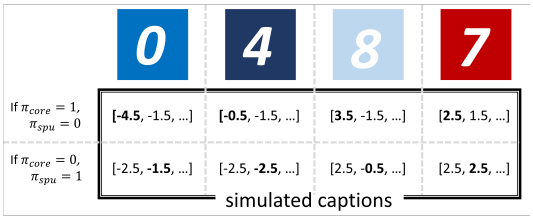


Figure 1: Construction of captions.

A Semi-synthetic Experiment. We conduct a carefully designed semi-synthetic binary classification experiment to showcase MMCL's robustness and the significance of rich captions. The task is to distinguish digits 0 to 4 (class 1) from digits 5 to 9 (class 2). In the training set, MNIST (Deng, 2012) digits are placed on colored backgrounds, including three types of blue and three types of red. As illustrated Figure 4, for digits 0-4, 99.5% of images have randomly selected shades of blue as the background, while the remaining 0.5% have random red backgrounds. The same applies to digits 5-9, but with blue and red swapped. In the test set, backgrounds are randomly chosen for all images. Therefore, digits represent the core feature, while colors serve as the spurious feature whose correlation with classes only exist in the training data. Captions are simulated as vectors, where the first coordinate contains digit information and the second contains color information.

Both features exhibit variance; for example, there are four variations of digits between 0 and 4 and three variations of blue backgrounds. We use π_{core} and π_{spu} to control the specificity of the captions, determining how much the caption mentions the variance in each feature. Being 'specific' means mentioning the exact value (e.g., specifying a particular shade of blue), while 'not specific' means referring to a value that represents an entire category (e.g., using the mean value for three shades of blue to represent any blue). Figure 1 shows an example. For more details, please refer to Appendix G.1.

We plot the OOD accuracy in Figure 2, while varying the values of π_{core} and π_{spu}. We observe: (1) With sufficiently rich captions (high π_{core}), MMCL exhibits better robustness than SL (horizontal

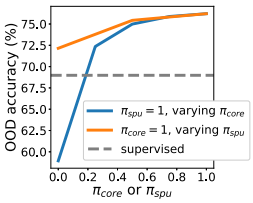
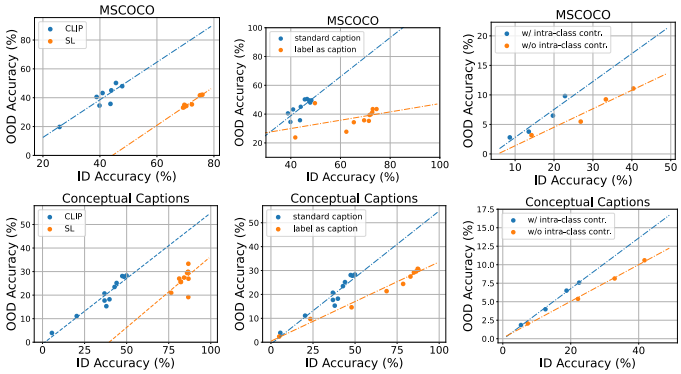


Figure 2: OOD accuracy on the semi-synthetic data. A large π_{core} is crucial for ensuring MMCL's superior robustness compared to SL, but the value of π_{spu} has minimal effect.



(a) MMCL vs SL (b) Caption richness (c) Intra-class contr.

Figure 3: (a) MMCL is more robust than SL. (b) Caption richness and (c) intra-class contrasting contribute to robustness. Note that (c) is in a different setup than (a)(b), as detailed in Appendices G.2 and G.3.

line). (2) A high π_{core}, indicating that the captions mentioning the variance of the core feature, is essential for achieving robustness, as reducing π_{core} significantly hurts the robustness. (3) In contrast, π_{spu} has minimal effect on robustness. It's worth noting that (2) and (3) directly validate the conclusions from Theorem 5.2. Additional discussion can be found in Appendix G.1.

Robustness on real data. We further coroborate our conclusions with experiments on MSCOCO (Lin et al., 2014) and Conceptual Captions (Sharma et al., 2018). We train models on these two datasets, and evaluate them on six shifted versions of ImageNets. See experimental details in Appendices G.2 and G.3).

MMCL is more OOD-robust than SL. We employ the widely used CLIP loss for MMCL and CE loss for SL to compare the robustness of the resulting models. Due to computational resource constraints, we adopt the simplified training setting from (Ren & Li, 2023), training a 3-layer MLP on top of frozen pretrained encoders. Following (Taori et al., 2020), we plot the ID-ODD accuracy relationship by varying the model size (width of the MLP). Fig 3a shows that models trained with MMCL demonstrate superior robustness on both datasets. Note that although the ID accuracy of CLIP is lower than that of SL, resulting in seemingly comparable OOD accuracies, we anticipate the actual advantage of CLIP to become more pronounced as the dataset size scales, similar to the original dataset used in (Radford et al., 2021). We also provide results regarding with other algorithms, including SimCLR and SupCon, and other backbone architectures in Appendix G.4.

Richness of captions is critical in achieving robustness. To demonstrate the impact of caption richness on robustness, we train an alternative version of CLIP wherein the captions are simplified to be the same as the labels. As depicted in Fig 3b, this modification leads to diminished robustness, which corroborates the theoretical conclusions in Section 5.

Intra-class contrasting contributes to robustness. To illustrate the mechanism theoretically presented in Section 4.1, we modify the CLIP loss to exclude pairs from the denominator if they are from the same class, thereby eliminating contrasting between images and texts of the same class. In this experiment, unlike the previous ones, we train the encoders from scratch and obtain different ID-OOD accuracy pairs by varying the training data size. This is because the effect of intra-class contrasting was not evident in the 3-layer MLP setting, likely due to the small model's limited capacity rendering it less sensitive to modifications in the loss. In Fig 3c, we observe that removing intra-class contrasting from the loss compromises robustness, confirming the importance of intra-class contrasting.

7 Conclusion

In this work, we provided the first theoretical explanation for MMCL's enhanced OOD robustness compared to SL. We showed conclusively that this robustness is attributed to aspects of the MMCL loss function, i.e. (1) intra-class contrasting (2) inter-class feature sharing, as well as the nature of multi-modal data i.e. (3) richness of captions. We confirmed our theoretical results using both synthetic and real-world experiments. Our findings could inspire the development of improved loss functions and data curation practices to further enhance MMCL's robustness.

Acknowledgements This research is partially supported by the National Science Foundation CAREER Award 2146492 and Cisco Systems.

REFERENCES

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

Benjamin Aubin, Agnieszka Słowik, Martin Arjovsky, Leon Bottou, and David Lopez-Paz. Linear unit-tests for invariance discovery. *arXiv preprint arXiv:2102.10867*, 2021.

Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. *Advances in neural information processing systems*, 32, 2019.

Tianle Cai, Ruiqi Gao, Jason Lee, and Qi Lei. A theory of label propagation for subpopulation shift. In *International Conference on Machine Learning*, pp. 1170–1182. PMLR, 2021.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations. February 2020. doi: 10.48550/arXiv.2002.05709. URL https://arxiv.org/abs/2002.05709v3.

Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. Debiased Contrastive Learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 8765–8775. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/63c3ddcc7b23daa1e42dc41f9a44a873-Abstract.html.

Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

Matthew Fahrbach, Adel Javanmard, Vahab Mirrokni, and Pratik Worah. Learning rate schedules in the presence of distribution shift. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 9523–9546. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/fahrbach23a.html.

Alex Fang, Gabriel Ilharco, Mitchell Wortsman, Yuhao Wan, Vaishaal Shankar, Achal Dave, and Ludwig Schmidt. Data determines distributional robustness in contrastive language image pre-training (clip). In *International Conference on Machine Learning*, pp. 6216–6234. PMLR, 2022.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018.

Shashank Goel, Hritik Bansal, Sumit Bhatia, Ryan Rossi, Vishwa Vinay, and Aditya Grover. Cyclip: Cyclic contrastive language-image pretraining. *Advances in Neural Information Processing Systems*, 35:6704–6719, 2022.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020.

Jeff Z HaoChen, Colin Wei, Adrien Gaidon, and Tengyu Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. *Advances in Neural Information Processing Systems*, 34:5000–5011, 2021.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8340–8349, 2021a.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. *CVPR*, 2021b.

Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs. *Advances in neural information processing systems*, 33:22118–22133, 2020.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. *Advances in neural information processing systems*, 32, 2019.

Wenlong Ji, Zhun Deng, Ryumei Nakada, James Zou, and Linjun Zhang. The power of contrast for feature learning: A theoretical analysis. *arXiv preprint arXiv:2110.02473*, 2021.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pp. 4904–4916. PMLR, 2021.

Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. *arXiv preprint arXiv:2110.09348*, 2021.

Siddharth Joshi, Yu Yang, Yihao Xue, Wenhan Yang, and Baharan Mirzasoleiman. Towards mitigating spurious correlations in the wild: A benchmark and a more realistic dataset, 2023.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2020.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pp. 5637–5664. PMLR, 2021.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pp. 121–137. Springer, 2020.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.

Stephen J Montgomery-Smith. The distribution of rademacher sums. *Proceedings of the American Mathematical Society*, 109(2):517–522, 1990.

Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. In *European Conference on Computer Vision*, pp. 529–544. Springer, 2022.

Ryumei Nakada, Halil Ibrahim Gulluk, Zhun Deng, Wenlong Ji, James Zou, and Linjun Zhang. Understanding multimodal contrastive learning and incorporating unpaired data. *arXiv preprint arXiv:2302.06232*, 2023.

Edwin G Ng, Bo Pang, Piyush Sharma, and Radu Soricut. Understanding guided image captioning performance across domains. *arXiv preprint arXiv:2012.02339*, 2020.

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pp. 689–696, 2011.

Thao Nguyen, Samir Yitzhak Gadre, Gabriel Ilharco, Sewoong Oh, and Ludwig Schmidt. Improving multimodal datasets with image captioning. *arXiv preprint arXiv:2307.10350*, 2023.

Hieu Pham, Zihang Dai, Golnaz Ghiasi, Kenji Kawaguchi, Hanxiao Liu, Adams Wei Yu, Jiahui Yu, Yi-Ting Chen, Minh-Thang Luong, Yonghui Wu, et al. Combined scaling for zero-shot transfer learning. *Neurocomputing*, 555:126658, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pp. 5389–5400. PMLR, 2019.

Yunwei Ren and Yuanzhi Li. On the importance of contrastive loss in multimodal learning. *arXiv preprint arXiv:2304.03717*, 2023.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115 (3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An investigation of why overparameterization exacerbates spurious correlations. In *International Conference on Machine Learning*, pp. 8346–8356. PMLR, 2020.

Shibani Santurkar, Dimitris Tsipras, and Aleksander Madry. Breeds: Benchmarks for subpopulation shift. *arXiv preprint arXiv:2008.04859*, 2020.

Shibani Santurkar, Yann Dubois, Rohan Taori, Percy Liang, and Tatsunori Hashimoto. Is a caption worth a thousand images? a controlled study for representation learning. *arXiv preprint arXiv:2207.07635*, 2022.

Nikunj Saunshi, Jordan Ash, Surbhi Goel, Dipendra Misra, Cyril Zhang, Sanjeev Arora, Sham Kakade, and Akshay Krishnamurthy. Understanding contrastive learning requires incorporating inductive biases. *arXiv preprint arXiv:2202.14037*, 2022.

Vaishaal Shankar, Achal Dave, Rebecca Roelofs, Deva Ramanan, Benjamin Recht, and Ludwig Schmidt. Do image classifiers generalize across time? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9661–9669, 2021.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2556–2565, 2018.

Galen R Shorack and GR Shorack. *Probability for statisticians*, volume 951. Springer, 2000.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1): 2822–2878, 2018.

Nitish Srivastava and Russ R Salakhutdinov. Multimodal learning with deep boltzmann machines. *Advances in neural information processing systems*, 25, 2012.

Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.

Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, 33:18583–18599, 2020.

Yuandong Tian. Understanding deep contrastive learning via coordinate-wise optimization. *Advances in Neural Information Processing Systems*, 35:19511–19522, 2022.

Yuandong Tian, Xinlei Chen, and Surya Ganguli. Understanding self-supervised learning dynamics without contrastive pairs. In *International Conference on Machine Learning*, pp. 10268–10278. PMLR, 2021.

Christopher Tosh, Akshay Krishnamurthy, and Daniel Hsu. Contrastive estimation reveals topic posterior information to linear models. *J. Mach. Learn. Res.*, 22:281–1, 2021a.

Christopher Tosh, Akshay Krishnamurthy, and Daniel Hsu. Contrastive learning, multi-view redundancy, and linear models. In *Algorithmic Learning Theory*, pp. 1179–1206. PMLR, 2021b.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

Yoav Wald, Amir Feder, Daniel Greenfeld, and Uri Shalit. On calibration and out-of-domain generalization. *Advances in neural information processing systems*, 34:2215–2227, 2021.

Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, pp. 10506–10518, 2019.

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

Per-Åke Wedin. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12:99–111, 1972.

Zixin Wen and Yuanzhi Li. Toward understanding the feature learning process of self-supervised contrastive learning. In *International Conference on Machine Learning*, pp. 11112–11122. PMLR, 2021.

Ruijia Wu, Linjun Zhang, and T Tony Cai. Sparse topic modeling: Computational efficiency, near-optimal algorithms, and statistical inference. *Journal of the American Statistical Association*, pp. 1–13, 2022.

Kai Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition. *arXiv preprint arXiv:2006.09994*, 2020.

Yihao Xue, Siddharth Joshi, Eric Gan, Pin-Yu Chen, and Baharan Mirzasoleiman. Which features are learnt by contrastive learning? on the role of simplicity bias in class collapse and feature suppression. *arXiv preprint arXiv:2305.16536*, 2023.

Yuzhe Yang, Haoran Zhang, Dina Katabi, and Marzyeh Ghassemi. Change is hard: A closer look at subpopulation shift. *arXiv preprint arXiv:2302.12254*, 2023.

Huaxiu Yao, Linjun Zhang, and Chelsea Finn. Meta-learning with fewer tasks through task interpolation. *arXiv preprint arXiv:2106.02695*, 2021.

Huaxiu Yao, Yu Wang, Sai Li, Linjun Zhang, Weixin Liang, James Zou, and Chelsea Finn. Improving out-of-distribution robustness via selective augmentation. In *International Conference on Machine Learning*, pp. 25407–25437. PMLR, 2022.

Qi Zhang, Yifei Wang, and Yisen Wang. On the generalization of multi-modal contrastive learning. *arXiv preprint arXiv:2306.04272*, 2023a.

Xingxuan Zhang, Yue He, Renzhe Xu, Han Yu, Zheyan Shen, and Peng Cui. Nico++: Towards better benchmarking for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16036–16047, 2023b.

Zhuotun Zhu, Lingxi Xie, and Alan L Yuille. Object recognition with and without objects. *arXiv preprint arXiv:1611.06596*, 2016.