Robustness in Multimodal Learning under Train-Test Modality Mismatch

Brandon McKinzie [1] **Vaishaal Shankar** [1] **Joseph Cheng** [2] **Yinfei Yang** [3] **Jonathon Shlens** [1] **Alexander Toshev** [1]

Abstract

Multimodal learning is defined as learning over multiple heterogeneous input modalities such as video, audio, and text. In this work, we are concerned with understanding how models behave as the type of modalities differ between training and deployment, a situation that naturally arises in many applications of multimodal learning to hardware platforms. We present a multimodal robustness framework to provide a systematic analysis of common multimodal representation learning methods. Further, we identify robustness shortcomings of these approaches and propose two intervention techniques leading to $1.5\times$-$4\times$ robustness improvements on three datasets, AudioSet, Kinetics-400 and ImageNet-Captions. Finally, we demonstrate that these interventions better utilize additional modalities, if present, to achieve competitive results of 44.2 mAP on AudioSet 20K.

1. Introduction

Machine learning models in the real world operate on a wide range of hardware platforms and sensor suites. Deployed models must operate on platforms ranging from wearable devices to autonomous vehicles in which a diverse suite of sensors provide a continuous commentary about the environment. Building a traditional machine learning model in this setting is challenging because *jointly* measuring data across *all* sensors might be infeasible. Likewise, sensor modalities may be added (or fail) at any time indicating that the tacit assumption of *i.i.d.* data may not occur in the real world.

Hence, properties of robustness across modalities become paramount when deploying a machine learning system to operate in a multimodal setting. First, a model should be able to operate on modalities not explicitly observed during training. For instance, we hope that the presence of additional modalities with no explicit labels may still benefit overall predictive performance. Second, models should gracefully degrade in the absence of modalities at test time. Both properties are unique to the multimodal setting.

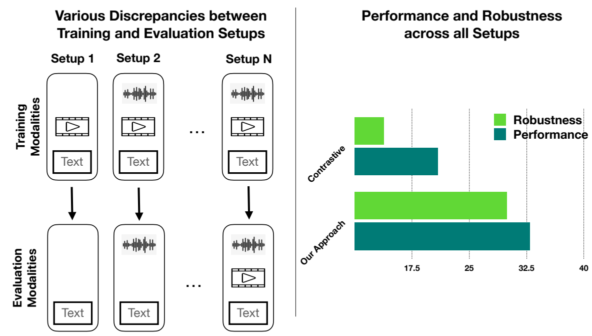

Figure 1. **Multimodal experimental setup and results.** We study representation learning for multimodal models which exhibit discrepancies between training and evaluation modalities. We define an analysis framework for this setup, study existing pretraining methods, and propose methods to improve robustness and performance.

To address these challenges, we study the problem of multimodal robustness. *How do models behave when arbitrary combinations of modalities may be added or removed at test time?* Supervised learning typically trains a model on a labeled dataset and examines how performance deteriorates as the hold-out validation set diverges from the training set (Recht et al., 2019; Shankar et al., 2021; Hendrycks & Dietterich, 2019). In our setting, we wish to instead build models in which one may flexibly swap in or out individual modalities that the model has seen during pretraining, downstream training, or both (Fig. 1).

One approach for achieving a flexible and performant representation to a suite of modalities is to have a model to learn a shared representation invariant to the modality identities – and subsequently train a discriminative model on top of that learned representation (Goodfellow et al., 2016). Several approaches for learning a shared representation have been explored in the literature, but recently, two prominent approaches – masked autoencoders (Gong et al., 2022; Geng et al., 2022) and contrastive learning (Radford et al., 2021; Wu et al., 2022b) – have demonstrated extraordinary promise in the setting of multimodal representations (Akbari et al., 2021). We focus our work on benchmarking robustness in representation learning, and ask how to improve

[1] Apple ML Research [2] Work done while at Apple [3] Apple. Correspondence to: Alexander Toshev <toshev@apple.com>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

such representations through new training strategies.

In this work we introduce a framework for measuring robustness in multimodal settings. We define a new robustness metric to capture variability across modalities by focusing on both average and worst-case performance across training and evaluation setups. Furthermore, we stratify these metrics across common scenarios such as adding, dropping, or completely swapping the modalities fed into the model.

We focus our experiments on representation learning with the AudioSet dataset (Gemmeke et al., 2017) in which three prominent modalities – audio, video and text – may be systematically manipulated. Additionally, we explore the generality of our results on Kinetics-400 (Kay et al., 2017) and ImageNet-Captions (Fang et al., 2022a).

We measure average and worst case performance when modalities are added or dropped at test time. To alleviate these degradations, we introduce two approaches to improve representation learning in a multimodal setting. The first approach — derived from knowledge distillation (Hinton et al., 2015) – termed *Modality Augmented Self-Distillation* (MASD), encourages consistency in the learned representations across labeled and unlabeled modalities. The second approach, derived from WiseFT (Wortsman et al., 2022), leverages a weighted combination of finetuned downstream weights and the initialization pretrained weights to induce robustness. We summarize our contributions as follows:

1. Introduce metrics and characterize performance in a multimodal setting on several datasets in terms of worst and average case performance.
2. Demonstrate training interventions (e.g. MASD, WiseFT) may additively lead to $1.5\times$-$4\times$ improvement of robustness on AudioSet, Kinetics-400 and ImageNet-Captions.
3. Increasing the number of modalities used to learn a representation improves downstream performance. In particular, we obtain SOTA results (44.2 mAP) on AudioSet-20K by leveraging text as an additional pretraining modality.

We hope that these results may accelerate the field of multimodal learning by offering simple, standard metrics and strong benchmarks for future improvements.

2. Related Work

2.1. Robustness

Robust machine learning has been a subject of study for decades. The support vector machine algorithm was presented as a "robust" prediction method (Boser et al., 1992) by finding the maximum margin classifier. Recently however there has been a push towards more practical forms of robustness for models operating on vision, natural language, speech and other modalities.

Worst case adversarial examples have been extensively studied in many domains (Szegedy et al., 2013; Alzantot et al., 2018; Carlini & Wagner, 2018) and while many effective "defense" methods have been proposed (Madry et al., 2017; Carlini et al., 2022; Carmon et al., 2019) it has been shown that these defenses reduce benign (non-adversarial) accuracy and don't generalize to *other* more natural forms of robustness (Taori et al., 2020). A similar story arises with synthetic "corruption" robustness (Hendrycks & Dietterich, 2019; Geirhos et al., 2018) where robust methods have been proposed but they fail to generalize to non synthetic corruptions.

For the class of *natural* corruptions or distribution shifts recent large scale multimodal image-text models (Radford et al., 2021; Pham et al., 2021) have shown unprecedented robustness when evaluated in a *zero-shot* manner (Recht et al., 2019; Barbu et al., 2019; Shankar et al., 2021; Gu et al., 2019). Subsequent work has demonstrated improvements for robustness in fine-tuned models (Wortsman et al., 2022).

2.2. Multimodal Learning

A natural way to learn a representation in a self-supervised manner from streams of multimodal data is to (1) have a set of modality encoders and an aggregator producing a single representation from all available modalities and (2) to consider paired modalities as positive examples. This way of thinking naturally lends itself to contrastive learning that embeds different modalities in a common space (Radford et al., 2021; Sohn, 2016). Most of the current work focuses on image and text only (Radford et al., 2021; Alayrac et al., 2022; Yuan et al., 2021; You et al., 2022) with a number of recent efforts in including video, audio, and even tabular data (Akbari et al., 2021; Alayrac et al., 2020; Liang et al., 2022).

An alternative to contrastive learning is a masked reconstruction objective. Most previous approaches have focused on single modalities, such as text (Devlin et al., 2018), images (He et al., 2022), videos (Feichtenhofer et al., 2022), and audio (Baade et al., 2022; Chong et al., 2022). More recently, this approach has also been adopted in multimodal settings (Geng et al., 2022; Wang et al., 2022). Other works employ both masked reconstruction and contrastive objectives (Gong et al., 2022; Yang et al., 2022; Fang et al., 2022b; Singh et al., 2021).

From a model architecture perspective, it remains an open question how best to fuse information from different modalities (Dou et al., 2021; Liu et al., 2018). The flexibility of transformers (Vaswani et al., 2017) enables them to be readily adapted to other modalities beyond language (Akbari et al., 2021; Liang et al., 2022; Jaegle et al., 2021; Nagrani

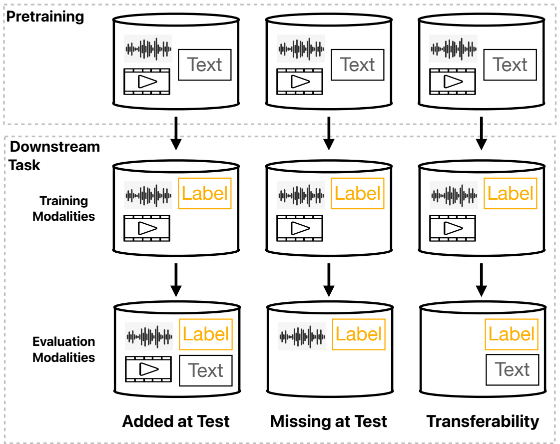

Figure 2. **Multimodal experiment setup:** pretraining, downstream task training, and evaluation (see Sec. 3), using as an example three modalities: video, audio, and text. The task at hand is classification, hence the presence of a label. At pretraining all modalities are present, while for a task only a subset is present. We describe three important setup corresponding to whether the evaluation contains more or less modalities, or a completely different set of modalities than at training: *Missing at Test*, *Added at Test*, and *Transferability* (see Sec. 4).

et al., 2021; Yang et al., 2022).

Increasing the number of modalities poses a challenge in training and in understanding the models. In supervised learning, the greedy nature of learning can be observed and quantified (Wu et al., 2022a; Hessel & Lee, 2020), as well as intra-modality and inter-modality heterogeneity (Liang et al., 2022).

3. Evaluation of Multimodal Representations

3.1. Setup and Notation

In this work we make several assumptions for our data that hold for a wide range of applications (see Fig. 2). First, we assume that we have readily available multimodal data consisting of several parallel input streams of different aligned modalities. Second, the above data can be acquired independently of the tasks of interest, although it might be related to it, and thus does not contain supervision.

We will refer to these data as *unsupervised pretraining data* D and the set of n modalities present in it by $M = \{m_1, \ldots, m_n\}$. Since we focus on subsets of modalities, it will be useful to refer to the data points x and datasets D restricted to a set of modalities $m \subseteq M$ by:

$$x|_m \quad \text{and} \quad D|_m = \{x|_m, x \in D\} \tag{1}$$

Further, for a downstream task we have data with supervision for both training and evaluation. It is reasonable to expect that the data with supervision are substantially smaller in quantity than the pretraining data. We refer to these data as *downstream training data* D_T with training modalities $M_T \subseteq M$, and *downstream evaluation data* D_E with evaluation modalities $M_E \subseteq M$. Importantly, the training and evaluation modality sets are allowed to be different, $M_T \neq M_E$, leading to robustness issues as shown later.

Downstream Task. Denote by $f_\theta(x)$ the downstream task model with weights θ. Note that f is multimodal, i.e. it can be applied on any subset $m \subseteq M$ of modalities, and such an application is denoted by $f_\theta(x|_m)$.

The parameters of the model are estimated by training for the downstream task on D_T using a task specific loss L:

$$L_{\text{task}}(D_T|_{M_T}) = \sum_{x \in D_T} L(f_\theta(x|_{M_T}) \tag{2}$$

where we explicitly say that the model is applied on x using only the modalities in M_T.

3.2. Multimodal Robustness Metrics

It is fair to assume that the downstream task of interest has a well established performance score p that can be measured for our model f. If this score is computed on the evaluation data $D_E|_{M_E}$ using modalities M_E after the model has been trained on $D_T|_{M_T}$ using modalities M_T, we denote this performance score by $p(M_E; M_T)$, where for brevity we skip the model and dataset notation.

Given a set of training modalities M_T, we propose to measure two aspects across all evaluation setups. The first is the average score, called performance, and represents how well the modalities M_T train a model when evaluated across all possible circumstances:

$$P(M_T) = \operatorname*{avg}_{M_E \subseteq M} p(M_E; M_T) \tag{3}$$

The second is the the worst score, called robustness, representing the worst possible deployment scenario for the model trained on M_T:

$$R(M_T) = \min_{M_E \subseteq M} p(M_E; M_T) \tag{4}$$

To produce a single set of metrics for a model across all possible training setups M_T, we propose to aggregate the above average and worst case performances in two ways. First, if one has control over picking an optimal training set, it makes sense to find the best performance and robustness. If we would like to evaluate on all possible training sets, then it makes sense to compute the average across training setups. We will refer to the former metrics as best Performance (P_{best}) and Robustness (R_{best}), and to the latter as simply

Performance (P) and Robustness (R):

$$P_{\text{best}} = \max_{M_T \subseteq M} P(M_T), \quad R_{\text{best}} = \max_{M_T \subseteq M} R(M_T) \qquad (5)$$

$$P = \operatorname*{avg}_{M_T \subseteq M} P(M_T), \quad R = \operatorname*{avg}_{M_T \subseteq M} R(M_T) \qquad (6)$$

Stratification of Performance and Robustness. The above metrics are originally defined over all possible evaluation modality sets $M_E \subseteq M$ for each training set . However, as motivated in Sec. 1 there can be various types of discrepancies. To better capture this, we refine $P(M_T)$ and $R(M_T)$ to be computed over a subset of possible evaluation modality sets M_E (Fig. 2):

1. **Missing at Test**: Testing modalities are a strict subset of the training modalities: $M_E \subset M_T$. This setup corresponds to having incomplete information at test time.
2. **Added at Test**: Testing modalities are a strict superset of the training modalities: $M_T \subset M_E$. This setup corresponds to modalities not present during training.
3. **Transferability**: Testing and training modalities are completely distinct: $M_T \cap M_E = \emptyset$. This is the most extreme setup, and tests the ability to transfer a task learned on one set to a completely different set of modalities.

We impose the above constraints on M_T and M_E in the computation of P and R in Eq. (3) and Eq. (4), and by proxy in Eq. (4).

Note that when the data has only two modalities, i.e. $|M| = 2$, for *Added at Test* and *Transferability* robustness and performance are identical $R = P$, as for every training modality set, there is only one evaluation modality set satisfying *Added at Test* and *Transferability* combinations. Then, the average and minimum opertions in Eq. (3) and Eq. (4) result in the same values.

4. Multimodal Self-Supervised Learning

4.1. Models

Pretraining. Multimodal data, as paired streams of different modalities, is a natural candidate for self-supervised learning as it is reasonable to assume that different modalities present different views of the same underlying content. This can be operationalized using contrastive (Radford et al., 2021; Jia et al., 2021) or masked reconstruction (He et al., 2022) objectives.

For the multimodal setup, we encode the different modalities with modality-specific encoders. In the case of contrastive learning, we follow closely the VATT architecture by Akbari et al. (2021), and formulate pair-wise InfoNCE losses (Gutmann & Hyvärinen, 2010; Oord et al., 2018) across all possible pairs of input modalities. This objective tries to learn per-modality representations that are as similar as possible for paired modalities. For MAE, we closely follow

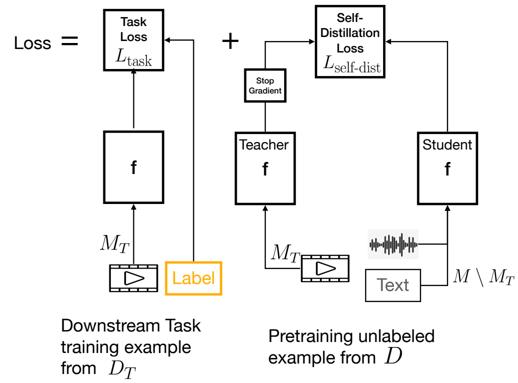

Figure 3. Diagram of Modality Augmented Self-Distillation. The Downstream task loss on the left receives labeled examples with M_T modalities (in this example Video), while the self-distillation loss receives unlabeled examples with all modalities, M_T are routed to the teacher network and $M \setminus M_T$ to the student (in this example, Audio and Text).

the AV-MAE baseline architecture described in Gong et al. (2022). Although masked reconstruction does not explicitly enforce a shared representation space for modalities, the hope is that the final shared-modality encoder layer contains information transferable from one modality to another. For further details of the formulation as well as architecture, we refer the reader to the Appendix and Sec. 5.

Downstream Training. After learning a representation using SSL, we apply it for a downstream task. In particular, denote by E_i the encoder for modality m_i that embeds an input $x|_{m_i}$ of this modality into a Euclidean space $E_i(x|_{m_i}) \in \mathbb{R}^d$ (see Sec. 3.1 for notation). Suppose, at downstream training or inference time, the data $D|_{M'}$ have a subset of modalities $M' \subseteq M$. Then, the final representation for $x \in D$:

$$E(x) = \frac{1}{|M'|} \sum_{m' \in M'} E_{m'}(x|_{m'}) \qquad (7)$$

This representation is used, for example in the case of a classification downstream task, to learn a classifier.

4.2. Improving Multimodal Robustness

We hypothesize that during downstream task learning, we see only a subset of all possible modalities, and as such this learning can ‘damage’ the pretrained model and diminish its ability to deal with the modalities not seen during downstream training. To address this challenge, we propose to apply ideas from transfer learning.

4.2.1. Modality Augmented Self-Distillation

One way to mitigate the problem is to use the pretraining data that contain all modalities but no supervision. These data can be used to regularize the performance of the model on all the modalities, even if this model is trained with a subset of the modalities present in the downstream training data. To achieve this, we draw inspiration from (Li & Hoiem, 2017; Castro et al., 2018; Hou et al., 2018; Rebuffi et al., 2017) to use Knowledge Distillation (Hinton et al., 2015) on the pretraining data.

In more detail, assume that the downstream task is classification and the model $f_\theta(y, x)$ produces probabilities over labels y for a given input x. Then, the teacher model $f_\theta(y, x|_{M_T})$ is the same model trained over the downstream training modalities M_T and data D_T. The student model is the same model $f_\theta(y, x|_{M \setminus M_T})$ as well (same weights), however, restricted over the modalities $M \setminus M_T$ *not* present in the downstream training data. Since the student and teacher models share the same weights, but have different input modalities, we call this loss self-distillation:

$$L_{\text{self-dist}}(D) = -\sum_{x \in D} \sum_{y} f_\theta(y, x|_{M_T}) \log \left(f_\theta(y, x|_{M \setminus M_T}) \right)$$

The final objective of MASD combines the above loss with the downstream task loss from Eq. (2) (see Fig. 3):

$$L_{MASD} = L_{\text{task}}(D_T|_{M_T}) + L_{\text{self-dist}}(D_{SD}) \qquad (8)$$

where the self-distillation loss is defined of a subset $D_{SD} \subset D$ of the pre-training data.

Since both the student and teacher model share the same weights θ, the above loss makes sure that the model is well behaved across all modalities M. Note that for training stability we stop the gradient flow through the teacher.

4.2.2. Applying WISE-FT to MASD models

There has been a recent line of work on improving the distributional robustness of finetuned large scale image-text models by *weight-space ensembling* (WISE-FT) the finetuned models and its pretrained (non finetuned) counterpart (Wortsman et al., 2022; Ilharco et al., 2022). While prior work used this procedure to obtain robustness on out-of-distribution test sets, we use the procedure to improve the robustness of our model when there is a difference between the train and test modalities.

Denote by θ_{masd} be the weights obtained by MASD and θ_{lp} be the weights obtained via linear probing. We compute our new weights by taking a weighted average:

$$\theta_{\text{wise}} = \alpha \theta_{\text{masd}} + (1 - \alpha) \theta_{\text{lp}} \qquad (9)$$

The only deviation from Wortsman et al. (2022) is that they averaged the finetuned image network with the pretrained network weights and "zero-shot" weights induced by the text embeddings of the class names. Since we finetune all the encoders and want a procedure that is modality agnostic we replace the text based zero-shot weights with linear probe weights. While the choice of α can be tuned with cross-validation we find a constant value of $\alpha = 0.75$ works well for our experiments.

Text	Labels
Muzik tipiko di Korsou/ Traditional Curacao music	Flamenco, Music, Mandolin, Music of Latin America
Tie Down Roping - 2013 NFR Round 8	Bang
COUPLES YOGA CHALLENGE	Music, Speech, Breathing
Eventide Timefactor Delay Pedal Part 2	Effects unit, Guitar, Music, Musical instrument, Chorus effect, Plucked string instrument
Klakson, który zwala z nóg	Vehicle, Speech
A Cappella Pitch Perfect Mashup	Singing, Music, Choir, Vocal music, A capella

Table 1. Random examples of text and associated labels from AudioSet evaluation set.

5. Experimental Setup

AudioSet (Gemmeke et al., 2017) is a video, audio, and text multi-label audio classification dataset over 527 classes. Prior work has largely leveraged the audio and/or video, but we also include the title of the video as text. AudioSet consists of an unbalanced training set of 1,743,790 examples, used as unlabeled pretraining data; a training and evaluation sets of 18,649 and 17,065 examples respectively used for the downstream task.

Note that the title is related to the content but rarely contains the audio event label (in 25.5% of the training video titles we have the label word mentioned; for examples see Table 1). Thus, the text presents a strong modality but does not make the problem trivial.

Kinetics-400 (Kay et al., 2017) is a video and audio action recognition dataset over 400 classes. It consists of a training and evaluation sets of 246,245 and 40,000 examples respectively used for the downstream task.

ImageNet-Captions (Fang et al., 2022a) is an image-text dataset created by extracting Flickr captions for images from the original ILSVRC2012 training dataset. It contains 999/1000 of the original ImageNet classes. The dataset contains 448,896 examples which we randomly split into 359,116 training and 89,779 evaluation images.

Preprocessing. We employ standard preprocessing before inference and training for each modality (e.g. (Gong et al., 2021; Nagrani et al., 2021)). Briefly, audio is extracted as single-channel 8 sec snippet sampled at 16 kHz with necessary padding. We compute log Mel spectrograms (128 frequency bins, 25ms Hamming window, 10 ms stride), and extract 16×16 patches. During training, videos are randomly short-side rescaled between 256 and 320 pixels, and randomly cropped to 224×224. During inference, videos are fixed short-side rescaled to 256 pixels following by a center crop to 224×224.

Data Augmentations We largely follow (Huang et al., 2022) for audio, (Feichtenhofer et al., 2022) for video, and (Radford et al., 2019) for text. This also includes applying mixup (Zhang et al., 2017) with rate 0.5 on all inputs/labels except text, drop path (Huang et al., 2016) with drop rate 0.1, and SpecAug (Park et al., 2019) with time/frequency masking of 192/48.

For videos, during both pretraining and downstream training we also use color augmentations for brightness (max delta = 0.2), contrast (max delta=0.1), saturation (max delta=0.0), hue (max delta=0.025). Also, we ensure the 8 seconds of audio/video are aligned such that the audio segment begins at the first sampled video frame and ends at the last sampled video frame.

For pretraining MAE, we follow (Feichtenhofer et al., 2022) and adopt repeated sampling, where each batch is duplicated/repeated some number of times (for us, we set the number of repeats per batch to 2), which improves training throughput due to the high cost of loading audio/video. This only makes sense for MAE with high masking ratios (we mask out 80% of the audio and 90% of the video during pretraining).

Pretraining. We use a ViT-B/16 architecture (Dosovitskiy et al., 2020) for all three modalities with appropriate modality specific positional encodings for both contrastive learning and MAE.

In the case of contrastive, we initialize weights with CLIP ViT-B/16 (Radford et al., 2021). In paticular, for text, we load the CLIP text encoder as-is. For video, we adopt the separable positional encoding ((Feichtenhofer et al., 2022)) and initialize the spatial component with the weights from CLIP's image encoder. For audio, we perform bilinear interpolation of the positional encodings (Dosovitskiy et al., 2020) to accommodate the input audio shape of 800×128.

After the aforementioned initialization, for AudioSet and Kinetics-400 we learn multimodal representation using AudioSet 2M. For ImageNet-Captions we use the OpenAI released ViT-B-16 CLIP representation.

For MAE, due to architectural differences, we cannot easily initialize from CLIP, so those models are pretrained from scratch[1]. Following (Gong et al., 2022), our MAE consists of modality-separate encoders with the ViT-B/16 architectures, with the final layer shared across modalities. For audio, we use a fixed 2D sinusoidal positional encoding as described in (Huang et al., 2022) and (He et al., 2022). For text, we use fixed 1D sinusoidal positional encodings as described in (Vaswani et al., 2017).

We pretrain the models with a 1024 batch size using the AdamW optimizer (Loshchilov & Hutter, 2017) with a learning rate of 8e-4. We pretrain the MAE and contrastive models 256 and 32 epochs, respectively.

Downstream Training. In the MASD loss in Eq. 8 we need a modality complete unlabeled data D_{SD} for self distillation. For experiments on AudioSet D_{SD} is a random 20K sample from the pre-training AudioSet data. For experiments on Kinetics D_{SD} is either a random 20K sample from pre-training AudioSet data or 20% random sample from the Kinetics training data. In the latter case the downstream task training data consists of the remaining 80%.

We run the downstream training 30 epochs after pretraining. Note that before we train the full model we learn a linear classifier on top of the frozen pretrained weights (referred to as linear probing), which is then used for initializing the classifier for the downstream training. Following (He et al., 2022), during linear probing we include BatchNorm without the affine transformation before the final classifier.

For all hyperparameters consult Table 2

Config	Pretraining		Linear Probing		Finetuning	
	Contr.	MAE	Contr.	MAE	Contr.	MAE
global batch	1024	1024	256	128	128	64
learning rate	8e-4	8e-4	1e-2	1e-2	1e-4	1e-4
LR warmup	1000	2000	200	200	1000	2000
epochs	32	256	360	360	30	60
optimizer	AdamW	AdamW	AdamW	AdamW	AdamW	AdamW

Table 2. Training hyperparameters used for pretraining, linear probing, and finetuning.

6. Multimodal Robustness Analysis

In the following we provide an analysis of multimodal models focusing on the following high level questions:

1. How do different multimodal representation learning methods fare against discrepancies between downstream training and evaluation modalities?
2. What type of discrepancies have the strongest impact on peformance and/or robustness?
3. What is the effect of the proposed interventions from Sec. 4.2 on multimodal robustness?

6.1. Analysis of multimodal learned representations

We focus on learned representations from standard contrastive learning and MAE presented in Sec. 4.1. Performance and Robustness metrics are presented in Tab. 3.

More modalities are better. To motivate the use of *multiple modalities during pretraining, training and evaluation* we measure the performance of contrastive learning, the better performing SSL model during both pretraining and

[1]This is partially why MAE is pretrained for 256 epochs, whereas the contrastive models are pretrained for 32 epochs. Another reason our contrastive method is pretrained for fewer epochs is because it was challenging to avoid overfitting if we pretrained any longer.

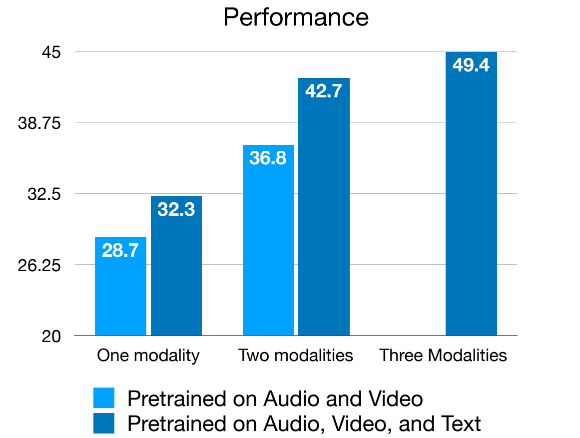

Figure 4. **Increasing the number of modalities at pretraining improves performance.** We consider two models pretrained using Contrastive Learning, one using audio and video, and a second using audio, video, and text. These are applied on a downstream task using 1-3 modalities. The performance numbers are averages across all possible combinations of 1-3 modalities, accordingly. Note, that the only the 3-modality pretrained model can be applied on 3 modalities, hence the right side of the plot has only one model.

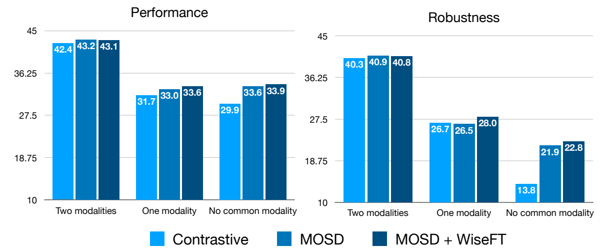

Figure 5. **MASD and WiseFT improves performance and robustness.** Average Performance (P) and Robustness (R) as the number of overlapping modalities between training/test goes from Two, to One, to None, for Contrastive Learning, Contrastive + MASD, and Contrastive + MASD + WiseFT.

downstream training, while maintaining $M_T = M_E$. We compute Performance per Eq. 6 where we average only across modality sets of a fixed size $|M_T| = |M_E| = k$. We vary $k \in \{1, 2, 3\}$.

Performance consistently improves as a model trains and tests on additional modalities (Fig. 4). Furthermore, the models benefit from more modalities at both pretraining and downstream training time. More specifically, pretraining on more modalities boosts performance further by 3.5 - 6.0 points (Fig. 4, light vs dark blue).

Multimodal representation struggles at downstream task for modalities not seen during training. The metrics introduced in Tab. 3 (*Overall*) aggregate across all possible training and evaluation combinations. To better understand which combinations challenge these models the most, we utilize the startified Performance and Robustness metrics *Added at Test*, *Missing at Test*, and *Transferability* defined in Sec. 3.2. Table 3 (right) shows results over these metrics.

The first observation is that the models are most robust when we have additional modalities at evaluation. In addition, the gap between robustness and average performance for both SSL methods is quite small in this case, which means that additional modalities during evaluation tend to only improve results. It's worth noting that, since the additional evaluation modalities were not present during downstream training, many of their associated parameters have not changed since pretraining, and yet they can still be combined with the fine-tuned parameters and improve evaluation performance. This is particular interesting for MAE, since all input modalities must pass through the final modality-shared encoder layer.

In the case of missing modalities at evaluation we see a small performance drop and a large robustness drop for both methods, although the degradation is worse for MAE [2]. Of course, some performance degradation is expected when modalities are removed. Ideally, performance should degrade *gracefully*, meaning it performs not significantly worse on the evaluation modalities than it would if those were the same modalities used during training.

In the case of completely different modalities at evaluation, we see that contrastive learning exhibits some transferability properties, but MAE collapses completely. This is again expected due to the difference in pretraining objectives, and since the only modality-shared parameters for contrastive models are the final linear classifier head whereas the MAE encoder also has modality-shared parameters in its final transformer layer. This seems to be the most challenging setup for all SSL methods.

6.2. Analysis of Robustness Interventions

The performance and robustness of propsed interventions from Sec. 4.2 and baseline models are shown in Table 3. These metrics are presented as an average across all training/evaluation modality combinations as well as across combination slices identified in Sec. 4.

MASD improve both performance and robustness As a first observation, MASD leads to a Performance improvement and substantial improvement of Robustness, for AudioSet, Kinetics-400, and ImageNet-Captions. Thus, MASD is addressing the weaknesses of original SSL methods. In particular, it reduced the degradation in case of *Added at Test* and *Transferability*, and in the case of Contrastive Learning, MASD doubles both Performance and Robustness. These results are consistent across both datasets, which demonstrates the generality of the learnings. The

[2]For example, on an AVT trained model, AV performance is 84.8% and 74.3% of AVT for contrastive and MAE, respectively.

Contrastive Loss Pretraining Method											
Dataset	Downstream Task Training	Overall				Missing at Test		Added at Test		Transferability	
		P_{best}	R_{best}	P	R	P	R	P	R	P	R
AudioSet	linear probe	33.7	22.1	28.0	15.0	29.6	24.0	34.2	33.5	16.0	13.9
	fine-tune	36.5	20.8	29.9	13.8	31.2	23.6	38.1	37.4	15.1	13.0
	WiseFT	37.3	22.3	29.5	13.5	31.4	24.5	37.6	36.9	14.0	12.0
	MASD	**37.4**	24.1	33.5	21.9	30.5	22.4	**40.4**	39.7	26.1	24.1
	MASD+WiseFT	37.3	**24.8**	**33.9**	**22.8**	**31.3**	**24.1**	40.2	**39.5**	**26.3**	**24.3**
	fine-tune on 2M	37.0	21.8	32.7	18.2	30.5	23.5	41.3	40.3	20.3	18.2
Kinetics-400	linear probe	42.2	21.7	34.7	17.0	34.4	**18.5**	36.8*		16.2*	
	fine-tune	45.5	11.1	36.2	6.1	29.1	11.1	47.8*		3.6*	
	MASD, distill-on-AS	49.8	23.5	40.6	18.5	**37.7**	17.5	49.0*		19.1*	
	MASD, distill-on-Kinetics	**52.0**	**26.9**	**45.2**	**19.9**	29.1	11.1	**59.0***		**33.7***	
ImageNet-Captions	linear probe	70.5	68.4	66.0	48.8	70.5	68.4	74.3*		39.1*	
	fine-tune	78.7	66.7	75.4	58.7	72.0	66.7	85.3*		54.7*	
	MASD	**84.3**	**80.8**	**82.4**	**76.0**	72.0	66.7	**90.9***		**80.8***	

Masked Autoencoder Pretraining Method											
Dataset	Downstream Task Training	Overall				Missing at Test		Added at Test		Transferab.	
		P_{best}	R_{best}	P	R	P	R	P	R	P	R
AudioSet	linear probe	23.4	5.5	14.0	1.8	17.2	7.7	17.8	17.0	1.5	1.2
	fine-tuned	28.9	3.8	20.0	1.3	21.4	10.0	30.8	30.3	1.1	0.9
	MASD	**30.6**	**15.1**	**26.6**	**9.5**	**21.6**	**10.3**	**35.4**	**34.4**	**18.5**	**15.1**
Kinetics-400	linear probe	30.6	11.0	19.8	3.9	19.6	11.0	7.1*		0.4*	
	fine-tuned	50.5	17.1	38.2	5.9	40.2	17.1	47.0*		0.3*	
	MASD, distill-on-AS	49.0	19.2	41.4	15.7	38.5	**19.2**	50.6*		14.0*	
	MASD, distill-on-Kinetics	**53.1**	**27.6**	**49.1**	**21.5**	40.2	17.1	**61.9***		**34.7***	

Table 3. Best Performance (P_{best}), Best Robustness (R_{best}), Average Performance (P) and Robustness (R) for two pretraining techniques with and without MASD, WiseFT: **top** is Contrastive Learning, **bottom** is Mask Autoenconder. We show results on AudioSet using audio, video, and text; Kinetics-400 using audio and video; and ImageNet-Captions with image and text. On the left side under *Overall* we show metrics computed over all possible training/evaluation modalities, on the right we show results for specific training/evaluation modality combinations (see Sec. 6.1). For Kinetics and AudioSet experiments we pretrain on AudioSet only. During self-distillation on Kinetics, we provide experiments by using AudioSet or a held-out portion of Kinetics.* For datasets with two modalities, per Sec. 6, the values for robustness and performance for these training/evaluation combinations are identical.

SSL Pretraining	Downstream Task Training	Downstream Task Training Modalities					
		V	A	T	AV	AT	VT
Contrastive	fine-tune	AVT	A	AVT	AVT	AT	AVT
Contrastive	MASD	AVT	AVT	AVT	AVT	AVT	AVT
Contrastive	MASD+WiseFT	AVT	AVT	AVT	AVT	AVT	AVT
MAE	fine-tune	VT	AT	T	AVT	AT	VT
MAE	MASD	AVT	AVT	AVT	AVT	AVT	AVT

Table 4. For each training modality set, we show the combination of evaluation modalities yielding the highest performance (see text). We abbreviate video=V, audio=A, text=T.

only degradation is in *Missing at Test* which is fixed by Wise-FT. Furthermore, our results show that MASD generalizes across three different types of modality sets across AudioSet, Kinetics-400, and ImageNet-Captions.

To further see the benefit of our proposed interventions we plot Robustness vs Performance for each possible training modality set in Fig. 6. While we see that Robustness is generally correlated with Performance, our interventions when combined consistently improve Robustness beyond the trend line. This is similar to a notion of "high effective robustness" as defined in (Taori et al., 2020).

Model	Pretrain	Training/Evaluation Modalities			
		A	V	AV	AVT
Contrastive, FT	AS2M	39.5	25.6	43.7	49.4
Contrastive, MASD+WiseFT	AS2M	39.5	30.0	44.2	49.4
MBT (Nagrani et al., 2021)	IN21K	31.3	27.7	43.9	
CAV-MAE (Gong et al., 2022)	AS2M	37.7	19.8	42.0	
VATT (Akbari et al., 2021)	IN	39.4			
Audio-MAE (Huang et al., 2022)	AS2M	37.0			

Table 5. Mean Average Precision on AudioSet 20K test for standard Contrastive Learning, MASD, and other competitive approaches on AudioSet. For pre-training, one can used either AS2M, ImageNet, or ImageNet 21K (Deng et al., 2009). Results not present in the literature are empty.

MASD improves robustness beyond supervised learning on more examples A natural question is whether downstream supervised training on larger labeled data can address multimodal robustness issues. In Table 3, we present downstream training on 2M labeled examples, which is $100\times$ than the labeled downstream training data for all other experiments. Although we see a 50% boost in robustness compared to regular downstream fine-tuning, this experiments still underperforms MASD on Robustness, in particular for

Transferability, while using substantially more labeling.

Robustness gains correlate with train-test modality gap. To better understand MASD, we compute metrics as we decrease the number of common modalities between training and evaluation. In Fig. 5, we show Performance and Robustness for $k = |M_T \cap M_E| \in \{0, 1, 2\}$ (see Eq. (6)), i.e. zero, one, or two common modalities. We can see that as the number of common modalities decreases, MASD degrades more gracefully compared to standard Contrastive Learning. WiseFT provides an additional stability in performance.

MASD helps better utilize all modalities at evaluation time. Another property of MASD is that it can utilize all modalities present at downstream evaluation, even if these are not available at downstream training. To see this, for each downstream training modality set M_T we identify the evaluation modality set M_E yielding highest performance: $\arg\max_{M_E \subseteq M} p(M_E; M_T)$ for each $M_T \subseteq M$.

We summarize the best evaluation modalities for each training modality set in Table 4. We can see that for the original Contrastive learning, in 2 out of 6 training setups the model attains best performance using the same evaluation modalities it has been trained on, $M_E = M_T$. However, for MASD we see that it always works best when we use all modalities, $M_E = \{A, V, T\}$. For MAE, we see an even bigger utilization – while in 5 cases the original model prefers a subset of the modalities at evaluation, with MASD the model in all 6 cases benefits from having all modalities at evaluation.

MASD achieves competitive performance compared to other approaches To better put MASD in perspective, we compare its performance to other approaches in the literature. In Table 5, we show results using the same training and evaluation modalities, we do so for four different modality sets: audio only, video only; audio and video; audio, video, and text. When using AudioSet 20K downstream training set as only labeled data, MASD achieves higher or equal performance to other reported approaches, across all studied modality combinations. Further, if using text, we obtain even superior performance (although other approaches do not use text). This shows that MASD not only fixes robustness issues for underlying SSL methods, but also keeps competitive results across various evaluation setups. We note that our AV number is the best reported number among all methods that only have access to the AS-20k labels.

7. Discussion

In this paper we quantified the notion of robustness in a multimodal representation. We introduced several simple definitions of robustness based on average and worst case performance across subsets of modalities. We characterized the robustness of state-of-the-art learned representations based on contrastive learning and masked autoencoders.

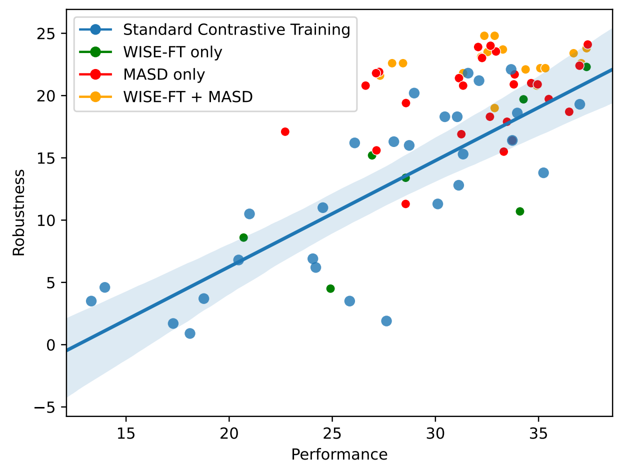

Figure 6. **Interventions improve robustness.** WiSE-FT + MASD provide substantial improvements to robustness across most training modality sets M_T. For each of the four methods and each possible training modality set $M_T \subseteq \{\text{audio}, \text{video}, \text{text}\}$, we plot robustness vs performance per Eq. (4) and (3).

We found that performance degrades with greater discrepancies between training and testing modalities, however these degradations may be alleviated with training improvements based on MASD distillation and WiseFT aggregation. Using these techniques we are able to improve upon state-of-the-art with AudioSet by leveraging multimodal data not available to the downstream task.

We observe several limitations for this current work, and opportunities for extensions and next steps. First, we focused our representation learning on homogenous multimodal data and it is unclear how this work will succeed in large scale hetergenous datasets. Further, although our benchmarks quantify the multimodal behavior on several datasets, it is unclear what is truly achievable given the structure and features of a given dataset. We strongly suspect that these results may be heavily dependent on the specifics of a given multimodal dataset but much work remains to characterize how the trends identified persist and how these benchmarks vary across typical multimodal conditions.

8. Acknowlegements

The authors would like to thank Jason Ramapuram and Tatiana Likhomanenko for useful suggestions regarding Knowledge Distillation; Jason Ramapuram, Devon Hjelm, Hadi Pour Ansari, and Barry Theobold for detailed feedback on the experiments, algorithm design, overall paper structure and writing; Oncel Tuzel, Sachin Mehta, Fartash Faghri, Alkesh Patel for ongoing feedback during the project; Tom Nickson and Angelos Katharopoulos for ongoing infrastructure support.

References

Akbari, H., Yuan, L., Qian, R., Chuang, W.-H., Chang, S.-F., Cui, Y., and Gong, B. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in Neural Information Processing Systems*, 34: 24206–24221, 2021.

Alayrac, J., Recasens, A., Schneider, R., Arandjelovic, R., Ramapuram, J., Fauw, J. D., Smaira, L., Dieleman, S., and Zisserman, A. Self-supervised multimodal versatile networks. *CoRR*, abs/2006.16228, 2020. URL `https://arxiv.org/abs/2006.16228`.

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022.

Alzantot, M., Sharma, Y., Elgohary, A., Ho, B.-J., Srivastava, M., and Chang, K.-W. Generating natural language adversarial examples. *arXiv preprint arXiv:1804.07998*, 2018.

Baade, A., Peng, P., and Harwath, D. Mae-ast: Masked autoencoding audio spectrogram transformer. *arXiv preprint arXiv:2203.16691*, 2022.

Barbu, A., Mayo, D., Alverio, J., Luo, W., Wang, C., Gutfreund, D., Tenenbaum, J., and Katz, B. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. *Advances in neural information processing systems*, 32, 2019.

Boser, B. E., Guyon, I. M., and Vapnik, V. N. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pp. 144–152, 1992.

Carlini, N. and Wagner, D. Audio adversarial examples: Targeted attacks on speech-to-text. In *2018 IEEE security and privacy workshops (SPW)*, pp. 1–7. IEEE, 2018.

Carlini, N., Tramer, F., Dvijotham, K., and Kolter, J. Z. (certified!!) adversarial robustness for free!, 2022. URL `https://arxiv.org/abs/2206.10550`.

Carmon, Y., Raghunathan, A., Schmidt, L., Duchi, J. C., and Liang, P. S. Unlabeled data improves adversarial robustness. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL `https://proceedings.neurips.cc/paper/2019/file/32e0bd1497aa43e02a42f47d9d6515ad-Paper.pdf`.

Castro, F. M., Marín-Jiménez, M. J., Guil, N., Schmid, C., and Alahari, K. End-to-end incremental learning. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 233–248, 2018.

Chong, D., Wang, H., Zhou, P., and Zeng, Q. Masked spectrogram prediction for self-supervised audio pre-training, 2022. URL `https://arxiv.org/abs/2204.12768`.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *CVPR*, pp. 248–255. Ieee, 2009.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Dou, Z.-Y., Xu, Y., Gan, Z., Wang, J., Wang, S., Wang, L., Zhu, C., Zhang, P., Yuan, L., Peng, N., Liu, Z., and Zeng, M. An empirical study of training end-to-end vision-and-language transformers, 2021. URL `https://arxiv.org/abs/2111.02387`.

Fang, A., Ilharco, G., Wortsman, M., Wan, Y., Shankar, V., Dave, A., and Schmidt, L. Data determines distributional robustness in contrastive language image pre-training (CLIP). In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S. (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 6216–6234. PMLR, 17–23 Jul 2022a. URL `https://proceedings.mlr.press/v162/fang22a.html`.

Fang, Y., Wang, W., Xie, B., Sun, Q., Wu, L., Wang, X., Huang, T., Wang, X., and Cao, Y. Eva: Exploring the limits of masked visual representation learning at scale, 2022b. URL `https://arxiv.org/abs/2211.07636`.

Feichtenhofer, C., Fan, H., Li, Y., and He, K. Masked autoencoders as spatiotemporal learners. *arXiv preprint arXiv:2205.09113*, 2022.

Geirhos, R., Temme, C. R. M., Rauber, J., Schütt, H. H., Bethge, M., and Wichmann, F. A. Generalisation in humans and deep neural networks. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural*

Information Processing Systems, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/0937fb5864ed06ffb59ae5f9b5ed67a9-Paper.pdf.

Gemmeke, J. F., Ellis, D. P., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., and Ritter, M. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 776–780. IEEE, 2017.

Geng, X., Liu, H., Lee, L., Schuurams, D., Levine, S., and Abbeel, P. Multimodal masked autoencoders learn transferable representations. *arXiv preprint arXiv:2205.14204*, 2022.

Gong, Y., Chung, Y.-A., and Glass, J. Ast: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778*, 2021.

Gong, Y., Rouditchenko, A., Liu, A. H., Harwath, D., Karlinsky, L., Kuehne, H., and Glass, J. Contrastive audio-visual masked autoencoder. *arXiv preprint arXiv:2210.07839*, 2022.

Goodfellow, I., Bengio, Y., Courville, A., and Bengio, Y. *Deep learning*, volume 1. MIT Press, 2016.

Gu, K., Yang, B., Ngiam, J., Le, Q., and Shlens, J. Using videos to evaluate image model robustness, 2019. URL https://arxiv.org/abs/1904.10076.

Gutmann, M. and Hyvärinen, A. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 297–304. JMLR Workshop and Conference Proceedings, 2010.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022.

Hendrycks, D. and Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations, 2019. URL https://arxiv.org/abs/1903.12261.

Hessel, J. and Lee, L. Does my multimodal model learn cross-modal interactions? it's harder to tell than you might think! *arXiv preprint arXiv:2010.06572*, 2020.

Hinton, G., Vinyals, O., Dean, J., et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015.

Hou, S., Pan, X., Loy, C. C., Wang, Z., and Lin, D. Lifelong learning via progressive distillation and retrospection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 437–452, 2018.

Huang, G., Sun, Y., Liu, Z., Sedra, D., and Weinberger, K. Deep networks with stochastic depth, 2016. URL https://arxiv.org/abs/1603.09382.

Huang, P.-Y., Xu, H., Li, J., Baevski, A., Auli, M., Galuba, W., Metze, F., and Feichtenhofer, C. Masked autoencoders that listen, 2022. URL https://arxiv.org/abs/2207.06405.

Ilharco, G., Wortsman, M., Gadre, S. Y., Song, S., Hajishirzi, H., Kornblith, S., Farhadi, A., and Schmidt, L. Patching open-vocabulary models by interpolating weights, 2022. URL https://arxiv.org/abs/2208.05592.

Jaegle, A., Borgeaud, S., Alayrac, J.-B., Doersch, C., Ionescu, C., Ding, D., Koppula, S., Zoran, D., Brock, A., Shelhamer, E., et al. Perceiver io: A general architecture for structured inputs & outputs. *arXiv preprint arXiv:2107.14795*, 2021.

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q. V., Sung, Y., Li, Z., and Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. *ICML*, 2021.

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

Li, Z. and Hoiem, D. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017.

Liang, P. P., Lyu, Y., Fan, X., Mo, S., Yogatama, D., Morency, L.-P., and Salakhutdinov, R. Highmmt: Towards modality and task generalization for high-modality representation learning. *arXiv preprint arXiv:2203.01311*, 2022.

Liu, K., Li, Y., Xu, N., and Natarajan, P. Learn to combine modalities in multimodal deep learning, 2018. URL https://arxiv.org/abs/1805.11730.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Nagrani, A., Yang, S., Arnab, A., Jansen, A., Schmid, C., and Sun, C. Attention bottlenecks for multimodal fusion. *Advances in Neural Information Processing Systems*, 34: 14200–14213, 2021.

Oord, A. v. d., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D., and Le, Q. V. SpecAugment: A simple data augmentation method for automatic speech recognition. In *Interspeech 2019*. ISCA, sep 2019. doi: 10.21437/interspeech.2019-2680. URL https://doi.org/10.21437%2Finterspeech.2019-2680.

Pham, H., Dai, Z., Ghiasi, G., Kawaguchi, K., Liu, H., Yu, A. W., Yu, J., Chen, Y.-T., Luong, M.-T., Wu, Y., Tan, M., and Le, Q. V. Combined scaling for open-vocabulary image classification, 2021. URL https://arxiv.org/abs/2111.10050.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. *ICML*, 2021.

Rebuffi, S.-A., Kolesnikov, A., Sperl, G., and Lampert, C. H. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 2001–2010, 2017.

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. Do imagenet classifiers generalize to imagenet? In *ICML*, pp. 5389–5400. PMLR, 2019.

Shankar, V., Dave, A., Roelofs, R., Ramanan, D., Recht, B., and Schmidt, L. Do image classifiers generalize across time? In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9661–9669, October 2021.

Singh, A., Hu, R., Goswami, V., Couairon, G., Galuba, W., Rohrbach, M., and Kiela, D. Flava: A foundational language and vision alignment model, 2021. URL https://arxiv.org/abs/2112.04482.

Sohn, K. Improved deep metric learning with multi-class n-pair loss objective. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper/2016/file/6b180037abbebea991d8b1232f8a8ca9-Paper.pdf.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Taori, R., Dave, A., Shankar, V., Carlini, N., Recht, B., and Schmidt, L. Measuring robustness to natural distribution shifts in image classification. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 18583–18599. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/d8330f857a17c53d217014ee776bfd50-Paper.pdf.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *NeurIPS*, 30, 2017.

Wang, W., Bao, H., Dong, L., Bjorck, J., Peng, Z., Liu, Q., Aggarwal, K., Mohammed, O. K., Singhal, S., Som, S., et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022.

Wortsman, M., Ilharco, G., Kim, J. W., Li, M., Kornblith, S., Roelofs, R., Lopes, R. G., Hajishirzi, H., Farhadi, A., Namkoong, H., and Schmidt, L. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7959–7971, June 2022.

Wu, N., Jastrzebski, S., Cho, K., and Geras, K. J. Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks. In *International Conference on Machine Learning*, pp. 24043–24055. PMLR, 2022a.

Wu, Y., Chen, K., Zhang, T., Hui, Y., Berg-Kirkpatrick, T., and Dubnov, S. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. *arXiv preprint arXiv:2211.06687*, 2022b.

Yang, Z., Fang, Y., Zhu, C., Pryzant, R., Chen, D., Shi, Y., Xu, Y., Qian, Y., Gao, M., Chen, Y.-L., et al. i-code: An integrative and composable multimodal learning framework. *arXiv preprint arXiv:2205.01818*, 2022.

You, H., Zhou, L., Xiao, B., Codella, N., Cheng, Y., Xu, R., Chang, S.-F., and Yuan, L. Learning visual representation from modality-shared contrastive language-image pretraining, 2022. URL https://arxiv.org/abs/2207.12661.

Yuan, L., Chen, D., Chen, Y.-L., Codella, N., Dai, X., Gao, J., Hu, H., Huang, X., Li, B., Li, C., et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021.

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. mixup: Beyond empirical risk minimization, 2017. URL `https://arxiv.org/abs/1710.09412`.