
Learning in Low Resource Modalities via Cross-Modal Generalization

Anonymous Author(s)

Affiliation

Address

email

Abstract

The natural world is abundant with underlying concepts expressed naturally in multiple heterogeneous sources such as the visual, acoustic, tactile, and linguistic modalities. Despite vast differences in these raw modalities, humans seamlessly perceive multimodal data, learn new concepts, and show extraordinary capabilities in *generalizing* across input modalities. Much of the existing progress in multimodal learning, however, focuses primarily on problems where the same set of modalities are present at train and test time, which makes learning in low-resource modalities particularly difficult. In this work, we propose a general algorithm for cross-modal generalization: a learning paradigm where data from more abundant source modalities is used to learn useful representations for scarce target modalities. Our algorithm is based on meta-alignment, a novel method to align representation spaces across modalities while ensuring quick generalization to new concepts. Experimental results on generalizing from image to audio classification and from text to speech classification demonstrate strong performance on classifying data from an entirely new target modality with only a few (1-10) labeled samples. In addition, our method works particularly well when the target modality suffers from noisy or limited labels, a scenario particularly prevalent in low-resource modalities.

1 Introduction

One of the hallmarks of human intelligence is the ability to generalize seamlessly across sensory inputs and cognitive tasks [9]. We see objects, hear sounds, feel texture, smell odors, and taste flavors, all the while reinforcing our understanding of the world and the underlying concepts present in it [5]. Much of the existing progress in multimodal learning, however, has focused primarily on a fixed set of predefined modalities and tasks [11, 40, 35]. As a result, it is unclear how to transfer knowledge from models trained for one modality (e.g. vision) to another (e.g. audio) at test time, which makes learning in low-resource modalities difficult. In this work, we propose algorithms for cross-modal generalization, a learning paradigm to train a model that can quickly perform new tasks defined in a target modality despite only being trained for tasks in a different source modality. This is particularly useful in leveraging high-resource source modalities to learn representations for low-resource target modalities, where unlabeled data is scarce and labeled data is even harder to obtain (e.g. audio from low-resource languages [38], real-world environments [43], and medical images [15]).

In comparison with existing work in domain adaptation, transfer learning, and multi-task learning, the problem of cross-modal generalization brings fundamental differences regarding how data is expressed across different modalities. We highlight two crucial distinctions: 1) the different input spaces consist of extremely high-dimensional, complex, and heterogeneous source and target modalities, and 2) there exist task-discrepancies between source and target modalities, such as the inherent differences between the label spaces when transferring from image to audio classification tasks. These discrepancies in both input and output spaces introduce *new concepts* expressed in *new modalities*.

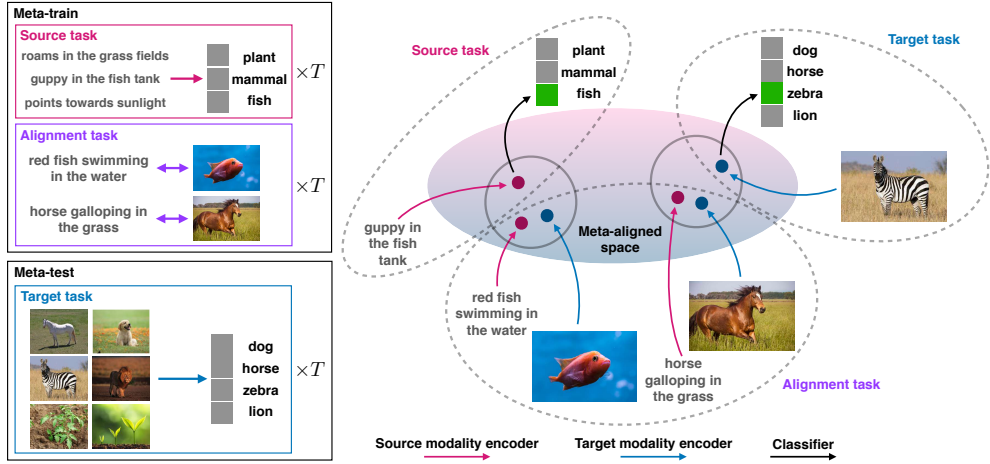


Figure 1: Cross-modal generalization leverages data from abundant source modalities (text) for scarce target modalities (images). At the heart of our approach is *meta-alignment*: a method to align representation spaces across modalities while ensuring quick generalization to new concepts. This enables generalization to the target having only seen labeled data in the source, as assessed by a series of few-shot classification and alignment tasks.

To account for these challenges, we formalize the conditions required for generalization and show that *cross-modal alignment* is necessary under partial observability across modalities and tasks. We propose a new algorithm (see Figure 1) based on *meta-alignment* which captures a space where representations of similar concepts in different modalities are close together while ensuring quick generalization to new tasks (i.e. seeing just a few labels in the target modality). We call it MAGMA (Meta-Alignment for Generalizing across ModAlities) and show how meta-alignment can be achieved using both strongly and weakly paired multimodal data abundant on the internet [64, 45, 35], which allows us to learn a classifier for transfer from source to target tasks. Finally, we quantify the trade-offs between labeling more data in the target modality and obtaining better source-target alignment.

We present experiments on two cross-modal tasks: generalizing from image to audio classification and from text to speech classification with the goal of classifying data from a new target modality when our approach gets only a few (1-10) labeled samples. Surprisingly, it is competitive with single modality baselines that have seen thousands of labeled examples from that target modality. In addition, we consider settings where the target modality suffers from noisy or limited data, a scenario particularly prevalent in low-resource modalities, languages, and concepts [21]. While this setting makes it difficult to directly train few-shot models in the target modality, we demonstrate that our cross-modal approach works particularly well.

2 Related Work

We highlight contrasts with recent work from related areas. Table 1 offers a summary of the data requirements across different related methods.

Few-shot learning: Many deep networks require large amounts of data to train [30, 13] which poses a problem in applications with limited labeled data [20, 21]. Few-shot learning [8] has enabled strong performance for these settings, with techniques spanning data augmentation [2], metric learning [59, 52], and learning better initializations [49, 41]. In the latter, **meta-learning** has recently emerged as a popular choice due to its simplicity in combination with gradient-based methods [17].

Transfer learning focuses on transferring knowledge from external data (e.g. larger datasets [29], unlabeled data [13], and knowledge bases [33]) to downstream tasks, where labeled data is more expensive [54]. **Domain adaptation** is a similar field which focuses on changing data distributions [26, 14]. However, existing works focus on data within the same modality (i.e. image domain adaptation [56], language transfer learning [13]) which simplifies the alignment problem.

Cross-modal learning studies knowledge transfer across modalities. For example, cross-modal data programming [15] uses weak labels in a source modality to train a classifier in the target modality. Cross-modal transfer learning aims to classify the same task from different input modalities [27, 62]. Finally, few-shot learning within target modalities (e.g. images) has been shown to benefit from additional multimodal information (e.g. word embeddings) [53, 61, 55]. However, this still requires labeled data from the target modality during meta-training (i.e. from a different domain). In a parallel vein, **co-learning** [5, 63] studies how external information from another modality can help prediction in a source modality, so both training and testing focuses on prediction in the source and is unable to

Table 1: Data requirements for generalization in a scarce target modality where both data and labels are rare. We use (meta-)train and (meta-)test to generalize conventional train and test stages in order to emphasize the availability of data and labels in the source and target. Cross-modal generalization leverages data from abundant source modalities for low-resource target modalities, thereby requiring only a *few samples* and *no labels* in the target beyond those used for few-shot fine-tuning.

APPROACHES	(META-)TRAIN			(META-)TEST		
	Modality	Data	Labels	Modality	Data	Labels
Transfer learning/Unsupervised pre-training [3, 13]	Target	Many	None	Target	Few	Few
Unsupervised meta-learning [25]	Target	Many	None	Target	Few	Few
Domain adaptation, Few-shot learning [17, 29, 42, 56]	Target	Many	Many	Target	Few	Few
Within modality generalization + cross-modal learning [15, 53, 55, 61, 63]	Source	Many	None	Target	Few	Few
	Target	Many	Many			
Cross-modal generalization (this work)	Source	Many	Many	Target	Few	Few
	Target	Few	None			

76 solve problems in an unseen target modality. To the best of our knowledge, our approach is the first
 77 to tackle *generalization* from a source to target modality.

78 Finally, **cross-modal retrieval** is also concerned with aligning spaces from 2 modalities, and methods
 79 have been proposed to handle learning from weak cross-modal supervision [7, 28]. However, such
 80 approaches do not actually study cross-modal generalization from, for example, image classification
 81 to audio classification tasks. Cross-modal generalization is harder since: 1) one has to learn not just the
 82 associations between modalities in retrieval but also associations to labels, 2) there is weak supervision
 83 both the target modality and in the label space (see Table 1), 3) tasks in different modalities have
 84 different (but semantically related) label spaces, and 4) the presence of new data and labels in the
 85 low-resource target modality which have to be learned using only a few samples.

86 3 Formalizing Multimodal Meta-Learning and Alignment

87 We build upon the definition of meta-learning in [24] and extend it to study multiple input modal-
 88 ities and the role of cross-modal alignment. The goal of meta-learning can be broadly defined
 89 as using the data for existing tasks to learn representations that enable fast learning on unseen
 90 tasks [34]. To reason over multiple tasks, we define a *task-distribution* $p(\mathcal{D}_m^x, \mathcal{D}_n^y)$ over tasks, where
 91 \mathcal{D}_m^x denotes the input space and \mathcal{D}_n^y the label space. For shorthand, we write this as a distribu-
 92 tion over the indices of the domains: $p(m, n) := p(\mathcal{D}_m^x, \mathcal{D}_n^y)$. Each task $\mathcal{T}(m, n)$ is defined as a
 93 pair of domains with a specified target pairing $p(x_m, y_n|m, n)$, $\mathcal{T}(m, n) = (\mathcal{D}_m, \mathcal{D}_n, p(x_m, x_n))$.
 94 We treat $p(m, n)$ as a marginal distribution obtained by integrating over a *meta-distribution*,
 95 $p(x_1, \dots, x_M, y_1, \dots, y_N, m_1, \dots, m_M, n_1, \dots, n_N)$, which we would like to model.

96 Within each task is an underlying pairing function, which maps the inputs to the labels or other
 97 inputs through $p_{m,n}(x, y) := p(x, y|m, n)$ for all $x \in \mathcal{D}_m^x$, $y \in \mathcal{D}_n^y$. To model this distribution, we
 98 learn a function from \mathcal{D}_m to \mathcal{D}_n using a meta-learning model f_w with parameters w . To account for
 99 generalization over all tasks, the overall meta-learning objective as follows:

100 **Definition 1.** A *meta-learning problem* is a minimization problem

$$\min_w \mathcal{L}[f_w] := \arg \min_w \mathbb{E}_{n, m \sim p(m, n); \mathbb{E}_{x, y \sim p_{m, n}(x, y)} - \log \left[\frac{f_w(x, y, m, n)}{p(x, y|m, n)} \right]. \quad (1)$$

101 When $p(n)$ is a delta distribution, we say that the problem is *single task*; otherwise, it is *multi-task*.
 102 $p(m)$ is the distribution over the source domains, and can take various kinds of distributions.

103 We call equation (1) the generalization loss and the goal of any model we consider is to minimize
 104 this loss. Notice that this loss is lower bounded by 0, and is achievable when $f_w(x, y, m, n) =$
 105 $p(x, y|m, n)$. A model f_w that achieves 0 loss in equation (1) is said to achieve perfect generalization.
 106 We are now ready to define a few-shot or zero-shot meta-learning problem.

107 **Definition 2.** Let \mathcal{M} be a subset of all the possible pairings of modalities and targets spaces. A
 108 meta-learning problem is said to be (partially) *low resource* if for all $m, n \in \mathcal{M}$, $p(x, y|m, n)$ is not
 109 known exactly, and has to be estimated using $q(x, y|m, n) \neq p(x, y|m, n)$.

110 Therefore, the subset \mathcal{M} can be called the low-resource subset, and any task associated with \mathcal{M} is
 111 a low-resource task. Note that this definition is equivalent to a situation where we have infinitely
 112 many data points for the high resource tasks, and a finite number of data points for the low-resource
 113 tasks. In practice, $q(x, y|m, n)$ is an (imperfect) estimation of $p(x, y|m, n)$ due to limited labeled

data. Mathematically, for a few-shot meta learning problem, the optimization objective becomes

$$\mathcal{L}_q[f_w] := -\mathbb{E}_{n, m \sim p(m, n)} \left\{ \underbrace{\mathbf{1}_{(m, n) \notin \mathcal{M}} \mathbb{E}_{x, y \sim p_{m, n}(x, y)} \log \left[\frac{f_w(x, y, m, n)}{p(x, y|m, n)} \right]}_{\text{high resource tasks}} + \underbrace{\mathbf{1}_{(m, n) \in \mathcal{M}} \mathbb{E}_{x, y \sim q_{m, n}(x, y)} \log \left[\frac{f_w(x, y, m, n)}{q(x, y|m, n)} \right]}_{\text{low resource tasks}} \right\}, \quad (2)$$

where $\mathbf{1}$ is the indicator function. This new optimization objective no longer matches the generalization objective \mathcal{L} in equation (1). The minimizer of this equation is $f_w(x, y, m, n) = q(x, y|m, n)$, which has an generalization error $\Pr\{(m, n) \in \mathcal{M}\} \text{KL}(p; q)$, where $\text{KL}(\cdot; \cdot)$ is the KL-divergence.

3.1 Cross-modal Meta-learning

To extend this definition to an input space spanning *multiple modalities*, we need to first introduce how different modalities are connected to each other. We assume the existence of an underlying *joint probability space*, with density $p(x_1, \dots, x_M, y_1, \dots, y_N | m_1, \dots, m_M, n_1, \dots, n_N)$ over M input modalities and N output tasks. This probability explicitly gives the underlying relationships between all modalities and tasks. However, this space is usually not fully observed, and what is usually observed are often marginal distributions over some of variables (e.g. labeled tasks for individual modalities in image classification [12], or paired data across modalities in videos [11]). To be able to generalize over arbitrary modalities and tasks when observing only partial subsets, we first define the minimum requirements on observed data, which we call the *minimum visibility assumption*:

Assumption 1. (*Minimum visibility*) For every task n , there is at least one domain m such that $p(x, y|m, n)$ is known. Likewise, for every domain m , there is at least one task n such $p(x, y|m, n)$ is known. All the single variable marginal distributions $p(x)$, $p(y)$ are also known.

This is the minimum assumption required to ensure that all modalities and tasks are accessible. We are now ready to define *multimodal meta-learning*.

Definition 3. We say that a meta-learning problem is *multimodal* if $p(m)$ is a discrete, i.e. a multinomial distribution, and if the pairwise conditional joint distribution $p(x_1, x_2|m_1, m_2)$ is known for all $x_1 \in \mathcal{D}_{m_1}^x$, $x_2 \in \mathcal{D}_{m_2}^x$ and for all m, n in the support of $p(m, n)$.

In practice, we say that a distribution is known if it can be accurately estimated. To achieve cross-modal generalization to new modalities and tasks in \mathcal{M} , it is important to leverage $p(x_1, x_2|m_1, m_2)$ to “bridge” modalities that are each labeled for only a subset of tasks. We formally explain how this can be used to enable cross-modal generalization in the next section.

3.2 Cross-modal Alignment

We use the word *alignment* to refer to any algorithm that bridges multiple modalities using cross-modal information $p(x_1, x_2|m_1, m_2)$. We first differentiate between *strong* and *weak* alignment:

Definition 4. Let $p(x_i, x_j)$ be known for $x_i \in \mathcal{D}_{m_i}^x$, $x_j \in \mathcal{D}_{m_j}^x$ and $i \neq j$. If both $p(x_i|x_j)$ and $p(x_j|x_i)$ are delta distributions, i.e., if there is a one-to-one mapping between x_i and x_j , we say that there is a strong alignment between modality m_i and m_j . Otherwise, there is only weak alignment.

We now show that strong alignment across modalities can achieve optimal generalization error for tasks in the low-resource subset \mathcal{M} .

Proposition 1. (*Benefit of strong alignment*). Let all the modalities be pairwise strongly-aligned, then we can define a surrogate loss function $\tilde{\mathcal{L}}[f_w]$ such that $\mathcal{L}[\arg \min_{f_w} \tilde{\mathcal{L}}[f_w]] = 0$.

The proof is provided in Appendix A. This implies that if strong alignment is achievable, then one can achieve perfect generalization in the low-resource subset \mathcal{M} . We also note that a key property we used in the proof is that $p(x_1|x_2) = p(x_1|x_2, y)$. For weak alignment, this property does not hold and perfect generalization is no longer achievable, and one needs to tradeoff the error induced by weak alignment with the error from minimizing q directly (i.e. few-shot supervised learning). We further explain and qualitatively analyze this trade-off in Appendix B.

Therefore, unlabeled cross-modal information $p(x_1, x_2|m_1, m_2)$ allows us to bridge modalities that are each labeled for only a subset of tasks and achieve cross-modal generalization to new modalities and tasks in \mathcal{M} . In practice, however, $p(x_1, x_2|m_1, m_2)$ is unknown and needs to be estimated from data. In the following section, we explain a specific algorithm based on contrastive learning [18] to estimate $p(x_1, x_2|m_1, m_2)$ and show how this achieves cross-modal generalization.

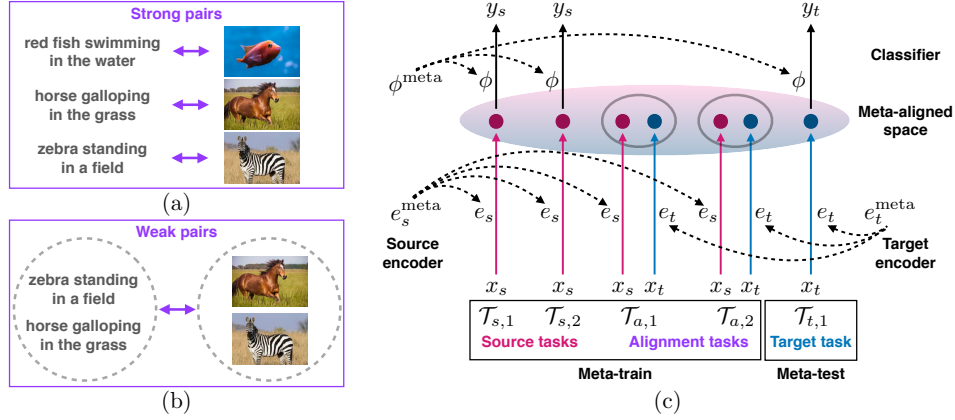


Figure 2: While strong pairs (a) provide exact, one-to-one correspondences across modalities, weak pairs (b) represent coarse semantic groupings which better reflect many-to-many cross-modal mappings and leverage weakly paired multimodal data available on the internet (e.g. videos, image captions). (c) During meta-training, meta-parameters e_s^{meta} , e_t^{meta} , ϕ^{meta} are trained using source modality classification tasks \mathcal{T}_s and alignment tasks \mathcal{T}_a . Meta-testing uses trained meta-parameters for few-shot generalization to target modality tasks \mathcal{T}_t .

4 MAGMA: Meta-Alignment for Generalizing across ModALities

In practice, we focus on the generalization problem from a source modality s to a target modality t . Insights from our theory leads to a practical algorithm involving two main steps: 1) learning a multimodal space via meta-alignment (§4.1), and 2) learning a cross-modal classifier (§4.2). We call our method MAGMA, short for Meta-Alignment for Generalizing across ModALities.

4.1 Meta-Alignment

We first simplify the problem by assuming access to strong pairs across modalities of the form (x_s, x_t) which makes it easier to learn *strong alignment* (see Figure 2(a)). At the same time, this is not an excessively strong assumption: many multimodal datasets contain paired multimodal data (e.g. activity recognition from audio and video [1] and emotion recognition from speech and gestures [4]).

In practice, alignment is achieved by modeling $p(x_t|x_s)$ instead of $p(x_s, x_t)$ since $p(x_s)$ can be estimated by the empirical distribution of the input. Directly learning a translation model $p_\theta(x_t|x_s)$ mapping each x_s to its corresponding x_t is unlikely to work in practice since x_s and x_t are extremely high-dimensional and heterogeneous data sources, and it is much easier if the alignment is achieved on a shared low-dimensional space. Therefore, we use Noise Contrastive Estimation (NCE) which learns a binary classifier to distinguish paired samples $(x_s, x_t) \in \mathcal{D}$ from unpaired negative samples $x_{t,\text{neg}}$. [16] showed that NCE in the asymptotic limit is an unbiased estimator of $p(x_t|x_s)$.

However, the vanilla NCE objective does not handle new concepts at meta-test time. We propose *meta-alignment* to capture an aligned space while ensuring quick generalization to new concepts across different modalities. Given strongly paired data, meta-alignment trains encoders e_s, e_t for source and target modalities across multiple alignment tasks $\{\mathcal{T}_{a,1}, \dots, \mathcal{T}_{a,T}\}$ into an aligned space. Each alignment task \mathcal{T}_a consists of a set of different concepts expressed as paired data across source and target modalities. We explicitly train for generalization to new concepts by training meta-alignment parameters e_s^{meta} and e_t^{meta} that are used to initialize instances of alignment models for each task [17]. When presented with a new task, we first initialize task parameters using meta parameters $e_s := e_s^{\text{meta}}$, $e_t := e_t^{\text{meta}}$ before training on the task by optimizing for the NCE loss with negative samples:

$$\mathcal{L}_{\text{strong align}} = \sum_{(x_s, x_t) \in \mathcal{T}_a} \left(-e_s(x_s)^\top e_t(x_t) + \sum_{x_{t,\text{neg}}} e_s(x_s)^\top e_t(x_{t,\text{neg}}) \right). \quad (3)$$

where $x_{t,\text{neg}}$ denotes unpaired negative samples. The NCE objective has a nice interpretation as capturing a space where the representations of similar concepts expressed in different modalities are close together, and different concepts in different modalities are far apart [18, 44]. The meta-parameters e_s^{meta} and e_t^{meta} are also updated using first-order gradient information [42] so that they gradually become better initialization parameters for new alignment tasks spanning new concepts.

Weak pairs: We now relax the data requirements from strong to *weak pairs*. Instead of one-to-one correspondences across modalities, weak pairs represent coarse groupings of semantic correspondence (see Figure 2(b)). This better reflects real-world multimodal data since cross-modal mappings are often

Algorithm 1 MAGMA: Meta-Alignment for Generalizing across ModAlities

Initialize meta-alignment encoders e_s^{meta} and e_t^{meta} , meta-classifier ϕ^{meta} .
for iteration = 1, 2, ... **do**
 Sample alignment task \mathcal{T}_a with train $\mathcal{D}_{\text{train}}^{\mathcal{T}_a}$ and test data $\mathcal{D}_{\text{test}}^{\mathcal{T}_a}$ of pairs $\{x_s, x_t\}$.
 Initialize $e_s := e_s^{\text{meta}}$, $e_t := e_t^{\text{meta}}$ and compute alignment loss using (4) on train data $\mathcal{D}_{\text{train}}^{\mathcal{T}_a}$.
 Compute \tilde{e}_s and \tilde{e}_t after gradient updates wrt alignment loss.
 Update meta-alignment encoders $e_s^{\text{meta}} \leftarrow e_s^{\text{meta}} + \epsilon(\tilde{e}_s - e_s^{\text{meta}})$, $e_t^{\text{meta}} \leftarrow e_t^{\text{meta}} + \epsilon(\tilde{e}_t - e_t^{\text{meta}})$.
 Sample source modality task \mathcal{T}_s with train $\mathcal{D}_{\text{train}}^{\mathcal{T}_s}$ and test data $\mathcal{D}_{\text{test}}^{\mathcal{T}_s}$ of pairs $\{x_s, y_s\}$.
 Initialize $\phi := \phi^{\text{meta}}$ and compute classification loss on train data $\mathcal{D}_{\text{train}}^{\mathcal{T}_s}$.
 Compute $\tilde{\phi}$ after gradient updates on task \mathcal{T}_s wrt classification loss.
 Update meta-classifier $\phi^{\text{meta}} \leftarrow \phi + \epsilon(\tilde{\phi} - \phi)$.

many-to-many (e.g. many ways of describing an image, many ways of speaking the same sentence). We consider two examples of weak pairs: 1. using large banks of weakly paired multimodal data available on the internet such as videos (weak pairs of image, audio, text [64, 45]) and image captions (weak pairs of images and words [35]), and 2. obtaining auxiliary pairs via WordNet [39], knowledge graphs [47], and scene graphs [35] that relate concept similarities across modalities.

We denote a weak pair as *sets* X_s and X_t . While the exact contrastive loss for strong pairs no longer applies, we modify it by taking the expectation over pairs across the sets, i.e. $x_s, x_t \in X_s \times X_t$:

$$\mathcal{L}_{\text{weak align}} = \sum_{(X_s, X_t) \in \mathcal{T}_a} \left(- \sum_{(x_s, x_t) \in X_s \times X_t} e_s(x_s)^\top e_t(x_t) + \sum_{x_t, \text{neg}} e_s(x_s)^\top e_t(x_t, \text{neg}) \right) \quad (4)$$

and refer to this process as *weak alignment*. In practice, we approximate this expectation by sampling several $x_s \in X_s$ and $x_t \in X_t$ to treat as paired samples instead of enumerating all pairs. Obtaining negative pairs $x_{t, \text{neg}}$ are straightforward by sampling outside of the paired sets.

4.2 Cross-modal Generalization

Since a well-aligned space is now modality-agnostic, we train a single classifier parametrized by a set of meta-parameters ϕ^{meta} on top of the aligned representations to classify concepts across modalities. The joint set of classification tasks consists of tasks $\{\mathcal{T}_{s,1}, \dots, \mathcal{T}_{s,T}\}$ in the source modality and tasks $\{\mathcal{T}_{t,1}, \dots, \mathcal{T}_{t,T}\}$ in the target. When presented with a new task, we first initialize the classifier using meta parameters $\phi := \phi^{\text{meta}}$ before training on the task by optimizing for the cross-entropy loss. The meta-parameters ϕ^{meta} are updated using first-order gradient information [42] towards better initialization parameters to classify new concepts. Overall, meta-training consist of alignment tasks \mathcal{T}_a and classification tasks in the source modality \mathcal{T}_s . Tasks in the target modality \mathcal{T}_t are only presented during meta-testing. Each task consists of k labeled pairs to simulate an episode of k -shot learning. We show the full training algorithm in Algorithm 1 and a visual diagram in Figure 2(c).

During testing, a task \mathcal{T}_t is sampled in the target modality. We initialize a new model with the trained meta-alignment encoder e_t^{meta} and meta-classifier ϕ^{meta} , and perform gradient updates with the k labeled samples in the target modality. Note that throughout the entire training process, only k labeled samples in the target modality are presented to MAGMA, which better reflects scarce target modalities in the real-world where even labeled data for different tasks are difficult to obtain.

5 Experiments

We test the ability of our proposed approach to generalize from image to audio and text to speech classification tasks. Anonymized code is included in the supplementary.

5.1 Image to Audio Generalization

Dataset: We build a benchmark for cross-modal generalization by combining two large unimodal classification datasets over images (CIFAR-10 and CIFAR-100 [36]) and audio (ESC-50 [46]) with partially related label spaces. This allows us to leverage complementary information from both modalities while testing for the introduction of new concepts across modalities. To obtain weak pairs, we map similar classes between the datasets using similarities from WordNet [39] and text cooccurrence. This yields 17 unique clusters of weak pairs (Appendix D.1 lists all the clusters).

Table 2: Performance on image to audio concept classification from CIFAR-10 and CIFAR-100 to ESC-50. MAGMA is on par with the oracle few-shot audio baseline that has seen a thousand of labeled audio samples and outperforms existing unimodal and cross-modal baselines. #Audio (labeled) denotes the number of audio samples and labels used during meta-training.




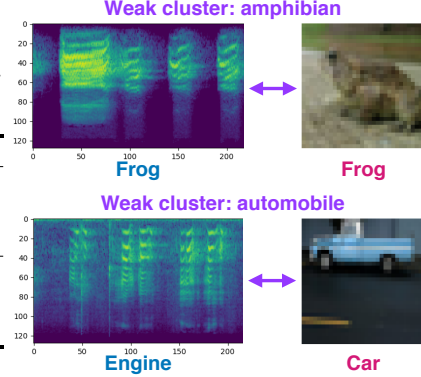
TYPE	APPROACH	1-SHOT	5-SHOT	10-SHOT	#AUDIO (LABELED)
Uni	Unsup. pre-training [3, 13]	44.2 \pm 0.8	72.3 \pm 0.3	77.4 \pm 1.7	0(0)
	Unsup. meta-learning [25] (reconstruct)	36.3 \pm 1.8	67.3 \pm 0.9	76.6 \pm 2.1	920(0)
Cross	Align + Classify [10, 23, 48, 57, 60]	45.3 \pm 0.8	73.9 \pm 2.1	78.8 \pm 0.1	920(0)
	Align + Meta Classify [51]	47.2 \pm 0.3	77.1 \pm 0.7	80.4 \pm 0.0	920(0)
	MAGMA  (ours)	47.5 \pm 0.2	85.9 \pm 0.7	92.7 \pm 0.4	920(0)
Oracle	Within modality generalization [17, 42]	45.9 \pm 0.2	89.3 \pm 0.4	94.5 \pm 0.3	920(920)

Table 3: **Left:** MAGMA yields better alignment scores than the baselines, indicating that meta-alignment can align new concepts using only weakly paired data. **Right:** samples of retrieved images given an audio sample. Despite being trained on weak pairs, meta-alignment can perform cross-modal retrieval at fine granularities.

K	EXPERIMENT	R@1 \uparrow	R@5 \uparrow	R@10 \uparrow	RANK \downarrow	COS. \downarrow
5	No align	1.0%	2.0%	5.5%	101	0.428
	Align	2.0%	5.5%	8.5%	103	0.272
	MAGMA 	4.0%	19.5%	39.0%	13	0.003
10	No align	0.5%	3.0%	4.5%	101	0.399
	Align	1.5%	11.0%	18.5%	52	0.222
	MAGMA 	3.5%	17.5%	35.0%	15	0.004



Baselines: We used a ResNet pretrained on ImageNet [12] to encode the images, a convolutional network pretrained on AudioSet [19] to encode audio [50, 37], and a shared fully-connected classification network for prediction. We focus our comparison with three broad sets of baselines:

1) **Unimodal (Uni)** baselines only use unlabeled data from the target modality during meta-training following our low-resource assumption. The simplest baseline ignores meta-training and just fine-tunes on the tasks in meta-test. Since the audio encoder was already pretrained on AudioSet, we call this baseline **unsupervised pre-training** [3, 13]. To better leverage unlabeled target modality data, we also compare with the **unsupervised meta-learning** baselines in [25] where we perform self-supervised reconstruction of unlabeled audio samples using a sequential autoencoder during meta-training. In Appendix D.1 we also compare to unsupervised meta-learning methods.

2) **Cross-modal (Cross)** methods are those that fall under our cross-modal generalization framework. Although prior work has not studied this exact setting, we adapt several baselines from domain adaptation and meta learning that fall under the following categories: a) **Align + Classify** which uses supervised alignment methods such as adversarial learning [57] cycle reconstruction [10, 23] or contrastive loss [60] to align input spaces from multiple domains before training a shared classifier [48]. b) **Align + Meta Classify** which learns a shared space using standard supervised alignment [18] before meta-learning a classifier [51], and c) **MAGMA** which represents our full model of jointly training for generalization across alignment and classification tasks. Since all methods are agnostic to the specific alignment algorithm used, we choose to use contrastive loss with negative sampling as described in section 4.1 for fair comparison across all baselines.

3) **Oracle:** The ideal (but likely unrealistic) scenario where meta-training and meta-testing both involve labeled data in the target modality. We use large amounts of labeled data in the target modality during meta-training in a different domain and transfer to meta-test tasks using the Reptile algorithm [42]. Since there is the least amount of domain shift, we expect this method to perform best but with the requirement of large amounts of labeled data in the target modality.

Metrics: We evaluate few-shot ($k = 1, 5, 10$) audio classification accuracy by fixing 8 evaluation tasks, each comprised of 5 unseen audio classes during meta-testing. We compute the accuracy averaged across all 8 tasks and repeat all experiments 10 times to report mean and standard deviations.

Results: From Table 2, we observe that our cross-modal approach is on par (outperforms for $k = 1$, and within 2 – 3% for $k = 5, 10$) with the oracle baseline that has seen a thousand labeled audio examples during meta-training. We also outperform existing unimodal and cross-modal generalization baselines. In particular, we find that jointly training for generalization across alignment and classification improves upon standard supervised alignment methods commonly used in domain adaptation [51, 48].

Table 4: Performance on text to speech generalization on the Wilderness dataset. MAGMA outperforms the oracle few-shot speech baseline that has seen thousands of labeled speech samples and the existing unimodal baselines. #Speech (labeled) denotes the number of speech samples and labels used during meta-training.


TYPE	APPROACH	1-SHOT	5-SHOT	10-SHOT	#SPEECH (LABELED)
Uni	Supervised learning	55.2 \pm 8.6	73.1 \pm 3.4	84.3 \pm 0.1	0(0)
	Unsup. meta-learning [25] (reconstruct)	61.5 \pm 4.4	83.5 \pm 4.0	88.5 \pm 2.1	4395(0)
Cross	Align + Classify [10, 23, 48, 57, 60]	61.1 \pm 6.0	74.8 \pm 2.1	86.2 \pm 0.7	4395(0)
	Align + Meta Classify [51]	65.6 \pm 6.1	89.9 \pm 1.5	93.0 \pm 0.5	4395(0)
	MAGMA  (ours)	67.9 \pm 6.6	90.6 \pm 1.5	93.2 \pm 0.2	4395(0)
Oracle	Within modality generalization [17, 42]	61.3 \pm 11.2	77.0 \pm 0.3	87.5 \pm 0.6	4395(4395)

Table 5: Language classification predictions on low-resource speech samples after training for labeled text data. Despite seeing just 5 labeled speech samples, our method is able to accurately classify low-resource languages.

SPEECH (TEXT IN PARENTHESIS)	OURS	ORACLE
(Beda Yesus agot gu ofa oida Bua buroru Didif ojgomu)	Meax	Russian
(Ido hai Timotiu natile hampai moulou Aturana Musa)	Badaic	Jamaican Patois
(Mu habotu pa kali Mataoqu osolae vekoi Rau sari Mua kana pa kauru Nenemu gua)	Roviana	Avokaya

Alignment: We show alignment performance in Table 3 on recall@ k , rank, and cosine loss metrics [18]. Our model yields better alignment metric values than the baseline approach, indicating that meta-alignment successfully aligns new concepts in low-resource target modalities. We also show samples of retrieved data in the source modality (image) given an input in the target modality (audio). Despite being trained only on weak pairs, meta-alignment can perform retrieval at fine granularities.

5.2 Text to Speech Generalization

Dataset: We use the Wilderness dataset, a large-scale multimodal dataset composed of parallel multilingual text and speech data [6]. We use a subset of 99 languages for language classification from text (source) and speech (target) individually. The tasks are split such as there is no overlap between the text and speech samples used for classification and the pairs seen for strong alignment.

Baselines: We use LSTMs to encode both text and speech data. We experiment with the same set of baselines spanning unimodal, cross-modal, and oracle methods described in section 5.1.

Metrics: We report few-shot ($k = 1, 5, 10$) language classification accuracy from speech by fixing 8 evaluation tasks, each comprised of 5 unseen languages during meta-test. We compute accuracy across all 8 tasks and repeat experiments 10 times to report mean and standard deviations.

Results: From Table 4, we observe that MAGMA surprisingly outperforms the oracle baseline in addition to the unimodal and cross-modal methods. We hypothesize this is due to the fact that text data is cleaner than speech data and the community generally has better models for encoding text tokens than speech spectrograms. This implies that one can also leverage better models in abundant, well-studied, and cleaner source modalities using our approach. Consistent with Table 2, we find that the performance improvement is greatest for the 1-shot setting, suggesting that meta-alignment is particularly suitable for low-resource problems.

Model predictions: Finally, we show some samples of language classification predictions on low-resource speech samples in Table 5. Despite seeing just 5 labeled speech samples, our method is able to quickly generalize and classify low-resource languages.

Noisy labels: We evaluate the effect of noisy labels in the target modality since it is often difficult to obtain exact labels in low-resource modalities such as rare languages. To simulate label noise, we add symmetric noise [22] to target modality labels (meta-train and meta-test). Despite only seeing $k = 1, 5, 10$ labels in the target, MAGMA is more robust than the oracle baseline (see Figure 3).

6 Conclusion

In this work, we formalized and proposed the MAGMA algorithm for cross-modal generalization: a learning paradigm where abundant source modalities are used to learn useful representations for scarce target modalities. Our experiments demonstrate strong performance on classifying data from an entirely new target modality under limited samples and noisy labels, which is particularly useful for generalization to low-resource speech and languages.

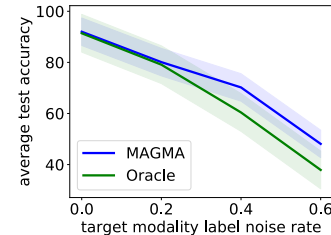


Figure 3: MAGMA is more robust to label noise in the target since it uses cross-modal information from source modalities. This makes it suitable for low-resource target modalities with imperfect annotations.

References

- [1] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Apostol (Paul) Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. In *arXiv:1609.08675*, 2016.
- [2] Antreas Antoniou, Amos Storkey, and Harrison Edwards. Data augmentation generative adversarial networks. *arXiv preprint arXiv:1711.04340*, 2017.
- [3] Alexei Baevski, Michael Auli, and Abdelrahman Mohamed. Effectiveness of self-supervised pre-training for speech recognition. *arXiv preprint arXiv:1911.03912*, 2019.
- [4] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246, Melbourne, Australia, July 2018. Association for Computational Linguistics.
- [5] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.
- [6] A. W. Black. Cmu wilderness multilingual speech dataset. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5971–5975, 2019.
- [7] Zhangjie Cao, Mingsheng Long, Jianmin Wang, and Qiang Yang. Transitive hashing network for heterogeneous multimedia retrieval. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pages 81–87, 2017.
- [8] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification. *arXiv preprint arXiv:1904.04232*, 2019.
- [9] François Chollet. On the measure of intelligence. *CoRR*, abs/1911.01547, 2019.
- [10] Safa Cicek and Stefano Soatto. Unsupervised domain adaptation via regularized conditional alignment. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1416–1425, 2019.
- [11] Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2054–2063, 2018.
- [12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
- [13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- [14] Nanqing Dong and Eric P Xing. Domain adaption in one-shot learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 573–588. Springer, 2018.
- [15] Jared Dunnmon, Alexander Ratner, Nishith Khandwala, Khaled Saab, Matthew Markert, Hersh Sagreiya, Roger E. Goldman, Christopher Lee-Messer, Matthew P. Lungren, Daniel L. Rubin, and Christopher Ré. Cross-modal data programming enables rapid medical machine learning. *CoRR*, abs/1903.11101, 2019.
- [16] Chris Dyer. Notes on noise contrastive estimation and negative sampling. *arXiv preprint arXiv:1410.8251*, 2014.
- [17] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1126–1135. JMLR. org, 2017.

- [18] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc’ Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*, pages 2121–2129, 2013.
- [19] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780. IEEE, 2017.
- [20] Marzyeh Ghassemi, Tristan Naumann, Peter Schulam, Andrew L. Beam, and Rajesh Ranganath. Opportunities in machine learning for healthcare. *CoRR*, abs/1806.00388, 2018.
- [21] Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor O. K. Li. Universal neural machine translation for extremely low resource languages. *CoRR*, abs/1802.05368, 2018.
- [22] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor W. Tsang, and Masashi Sugiyama. Co-sampling: Training robust networks for extremely noisy supervision. *CoRR*, abs/1804.06872, 2018.
- [23] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A. Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. *CoRR*, abs/1711.03213, 2017.
- [24] Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey. Meta-learning in neural networks: A survey. *arXiv preprint arXiv:2004.05439*, 2020.
- [25] Kyle Hsu, Sergey Levine, and Chelsea Finn. Unsupervised learning via meta-learning. *CoRR*, abs/1810.02334, 2018.
- [26] Yen-Chang Hsu, Zhaoyang Lv, and Zsolt Kira. Learning to cluster in order to transfer across domains and tasks. *arXiv preprint arXiv:1711.10125*, 2017.
- [27] Xin Huang, Yuxin Peng, and Mingkuan Yuan. Cross-modal common representation learning by hybrid transfer network. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence, IJCAI’17*, page 1893–1900. AAAI Press, 2017.
- [28] Xin Huang, Yuxin Peng, and Mingkuan Yuan. Cross-modal common representation learning by hybrid transfer network. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 1893–1900, 2017.
- [29] Mi-Young Huh, Pulkit Agrawal, and Alexei A. Efros. What makes imagenet good for transfer learning? *CoRR*, abs/1608.08614, 2016.
- [30] Forrest Iandola, Matt Moskewicz, Sergey Karayev, Ross Girshick, Trevor Darrell, and Kurt Keutzer. Densenet: Implementing efficient convnet descriptor pyramids. *arXiv preprint arXiv:1404.1869*, 2014.
- [31] Hal Daumé III. Frustratingly easy domain adaptation. *CoRR*, abs/0907.1815, 2009.
- [32] Pratik Jawanpuria, Mayank Meghwanshi, and Bamdev Mishra. Geometry-aware domain adaptation for unsupervised alignment of word embeddings. *arXiv preprint arXiv:2004.08243*, 2020.
- [33] Annervaz K M, Somnath Basu Roy Chowdhury, and Ambedkar Dukkipati. Learning beyond datasets: Knowledge graph augmented neural networks for natural language processing. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 313–322, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.
- [34] Mikhail Khodak, Maria-Florina Balcan, and Ameet Talwalkar. Provable guarantees for gradient-based meta-learning. *CoRR*, abs/1902.10644, 2019.

- [35] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *CoRR*, abs/1602.07332, 2016.
- [36] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.
- [37] Anurag Kumar, Maksim Khadkevich, and Christian Fugen. Knowledge transfer from weakly labeled audio using convolutional neural network for sound events and scenes. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 326–330, 2018.
- [38] Xinjian Li, Zhong Zhou, Siddharth Dalmia, Alan W Black, and Florian Metze. Santlr: Speech annotation toolkit for low resource languages. *arXiv preprint arXiv:1908.01067*, 2019.
- [39] George A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, November 1995.
- [40] Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi. Towards multimodal sentiment analysis: Harvesting opinions from the web. In *Proceedings of the 13th International Conference on Multimodal Interactions*, pages 169–176. ACM, 2011.
- [41] Tsendsuren Munkhdalai and Hong Yu. Meta networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2554–2563. JMLR. org, 2017.
- [42] Alex Nichol and John Schulman. Reptile: a scalable metalearning algorithm. *arXiv preprint arXiv:1803.02999*, 2:2, 2018.
- [43] OpenAI, Marcin Andrychowicz, Bowen Baker, Maciek Chociej, Rafal Józefowicz, Bob McGrew, Jakub W. Pachocki, Jakub Pachocki, Arthur Petron, Matthias Plappert, Glenn Powell, Alex Ray, Jonas Schneider, Szymon Sidor, Josh Tobin, Peter Welinder, Lilian Weng, and Wojciech Zaremba. Learning dexterous in-hand manipulation. *CoRR*, abs/1808.00177, 2018.
- [44] Akila Pemasiri, Kien Nguyen, Sridha Sridharan, and Clinton Fookes. Semantic correspondence: A hierarchical approach. *CoRR*, abs/1806.03560, 2018.
- [45] Veronica Perez-Rosas, Rada Mihalcea, and Louis-Philippe Morency. Utterance-Level Multimodal Sentiment Analysis. In *Association for Computational Linguistics (ACL)*, Sofia, Bulgaria, August 2013.
- [46] Karol J Piczak. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1015–1018, 2015.
- [47] Jay Pujara and Sameer Singh. Mining knowledge graphs from text. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, WSDM ’18*, 2018.
- [48] Anant Raj, Vinay P. Namboodiri, and Tinne Tuytelaars. Subspace alignment based domain adaptation for RCNN detector. *CoRR*, abs/1507.05578, 2015.
- [49] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. 2016.
- [50] Tal Ridnik, Hussam Lawen, Asaf Noy, and Itamar Friedman. Tresnet: High performance gpu-dedicated architecture. *arXiv preprint arXiv:2003.13630*, 2020.
- [51] Doyen Sahoo, Hung Le, Chenghao Liu, and Steven C. H. Hoi. Meta-learning with domain adaptation for few-shot learning under domain shift, 2019.
- [52] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in neural information processing systems*, pages 4077–4087, 2017.
- [53] Richard Socher, Milind Ganjoo, Christopher D. Manning, and Andrew Y. Ng. Zero-shot learning through cross-modal transfer. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 1, NIPS’13*, page 935–943, Red Hook, NY, USA, 2013. Curran Associates Inc.

- 443 [54] Chuanqi Tan, Fuchun Sun, Tao Kong, Wenchang Zhang, Chao Yang, and Chunfang Liu. A
444 survey on deep transfer learning. *CoRR*, abs/1808.01974, 2018.
- 445 [55] Yao-Hung Hubert Tsai and Ruslan Salakhutdinov. Improving one-shot learning through fusing
446 side information. *CoRR*, abs/1710.08347, 2017.
- 447 [56] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain
448 adaptation. *CoRR*, abs/1702.05464, 2017.
- 449 [57] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain
450 adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*,
451 pages 7167–7176, 2017.
- 452 [58] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain
453 confusion: Maximizing for domain invariance. *CoRR*, abs/1412.3474, 2014.
- 454 [59] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks
455 for one shot learning. In *Advances in neural information processing systems*, pages 3630–3638,
456 2016.
- 457 [60] Chang Wang and Sridhar Mahadevan. Heterogeneous domain adaptation using manifold
458 alignment. In *Proceedings of the Twenty-Second International Joint Conference on Artificial
459 Intelligence - Volume Volume Two, IJCAI’11*, page 1541–1546. AAAI Press, 2011.
- 460 [61] Chen Xing, Negar Rostamzadeh, Boris Oreshkin, and Pedro O O. Pinheiro. Adaptive cross-
461 modal few-shot learning. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc,
462 E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages
463 4847–4857. Curran Associates, Inc., 2019.
- 464 [62] Zhenguo Yang, Min Cheng, Qing Li, Yukun Li, Zehang Lin, and Wenyin Liu. Cross-domain
465 and cross-modality transfer learning for multi-domain and multi-modality event detection. In
466 Athman Bouguettaya, Yunjun Gao, Andrey Klimenko, Lu Chen, Xiangliang Zhang, Fedor
467 Dzerzhinskiy, Weijia Jia, Stanislav V. Klimenko, and Qing Li, editors, *Web Information Systems
468 Engineering – WISE 2017*, pages 516–523, Cham, 2017. Springer International Publishing.
- 469 [63] Amir Zadeh, Paul Pu Liang, and Louis-Philippe Morency. Foundations of multimodal co-
470 learning. *Information Fusion*, 64:188–193, 2020.
- 471 [64] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe
472 Morency. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable
473 dynamic fusion graph. In *ACL*, 2018.
- 474 [65] Qiming Zhang, Jing Zhang, Wei Liu, and Dacheng Tao. Category anchor-guided unsupervised
475 domain adaptation for semantic segmentation. In *Advances in Neural Information Processing
476 Systems*, pages 433–443, 2019.