

477 A Proof of Proposition 1

478 *Proof.* Let \mathcal{T}_{m_i, n_j} be a low-resource, where we only know $q(x_{m_i}, y)$. We want to show that
 479 we can recover $p(x, y)$ from alignment information. By the assumption of visibility, for task
 480 j , there is a strongly aligned modality $m_k \neq m_i$ for which we know $p(x_{m_i}, x_{m_j})$. By Bayes'
 481 rule $p(x_{m_i}, x_{m_k}, y) = p(x_{m_i}|x_{m_k}, y)p(x_{m_k}, y)$, but x_{m_i} is conditionally independent of y
 482 if x_{m_k} is known due to the existence of one-to-one mapping between them. Therefore, we
 483 can calculate $p(x_{m_i}, x_{m_k}, y) = p(x_{m_i}|x_{m_k})p(x_{m_k}, y)$ recover the desired label $p(x_{m_i}, y) =$
 484 $\int dx_{m_k} p(x_{m_i}, x_{m_k}, y)$. Now we can replace $q(x, y)$ by the recovered $p(x, y)$ in the loss func-
 485 tion, thus achieving perfect generalization on this task. \square

486 A.1 Concerning Weak Alignment

487 For weak alignment, this property may not hold and perfect generalization may not be achievable.
 488 Therefore, one needs to tradeoff the error induced by weak alignment with the error from minimizing
 489 q directly (i.e. few-shot supervised learning). This does not necessarily mean that weak alignment will
 490 hurt generalization: if $p(x_1|x_2, y) = p(x_1|x_2)$ holds, then perfect generalization can still be achieved.
 491 Of course, one might differentiate between the *perfect weak alignment* problem, where the statement
 492 $p(x_1|x_2, y) = p(x_1|x_2)$ holds (or, requiring one additional assumption) and *proper weak alignment*,
 493 where it does not. One can therefore prove the following corollary.

494 **Corollary 1.** *Assuming perfect weak alignment, one can achieve perfect generalization error.*

495 The proof follows directly from Proposition 1.

496 B Alignment vs Supervision

497 In section 3.2 of the main text we provided an analysis of learning in a low-resource target modality
 498 by relying on alignment with an abundant source modality. Proposition 1 implies that if strong
 499 alignment is achievable, then one can achieve perfect generalization in the low-resource subset \mathcal{M} .
 500 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 the case
 501 of weak supervision, this property does not hold and perfect generalization is no longer achievable
 502 due to inherent errors in modeling alignment using only weakly paired data. Therefore, one needs to
 503 tradeoff the additional error induced by weak alignment with the error from minimizing q directly
 504 (i.e. few-shot supervised learning in the low-resource target modality). In this section, we provide
 505 both theoretical analysis regarding this tradeoff to help practitioners decide on whether to rely on
 506 cross-modal alignment or within-modality supervised learning when faced with a low-resource target
 507 modality.

508 B.1 Theoretical Analysis

509 We begin with some mathematical guidelines on applying our method. Future theoretical work will be
 510 directed at formalizing the discussion in this section. For example, rigorous bounds on the minimizers
 511 can be derived when the models used are Lipschitz-continuous.

512 We focus on providing a understanding weak alignment before extending the analysis to cover the
 513 case of strong alignment. Let S denote the total number of weak-alignment sets, each with ρ^2 inner-set
 514 variance, and N_t be the number of target data points with supervision, then, clearly, a *tradeoff* in ρ^2
 515 and $\frac{1}{N_t}$ exists: direct supervised learning results in a generalization error proportional to $\frac{1}{N_t}$, while
 516 weak supervision results in error proportional to ρ^2 . Dividing N data points into S nearest neighbor
 517 sets, the resulting sets each have roughly N/S data points. If the original data points are drawn from
 518 a uniform distribution, then, each set will have variance proportional to $\frac{1}{S}$. Then, performing weak
 519 alignment is better than doing supervised learning if

$$\frac{c_s}{S} < \frac{c_t}{N_t} \quad (5)$$

520 for some architecture and task dependent constants c_s, c_t . This means that if the number of anchored
 521 sets is large or when the number of supervised data point is very small, then one should opt for using
 522 weak-alignment.

523 We can also rewrite this in terms of the number of data points for each set N_s we have. Since S is
 524 the number of anchor points, one expects that the error in alignment decreases as $\frac{1}{S}$. Let $N = SN_s$

denote the total number of data points. for some architecture and task dependent constant c_s, c_t . The above inequality is equivalent to

$$\frac{N_t}{S} = \frac{N_s N_t}{N} < c, \quad (6)$$

for some constant c . If we keep both the number of datapoints in each set and the supervised datapoints constant, then the trade-off depends only on N . If the number of total datapoints is large, one should use weak-alignment. What is the difference between learning with strong alignment and weak alignment? Intuitively, one would expect the generalization error to vanish when $N \rightarrow \infty$ for strong alignment, since the perfect one-to-one mapping between the target and the source can be discovered in this case. For weak alignment, however, one does not achieve vanishing generalization error in principle, since a fundamental uncertainty of order ρ^2 exists regarding the pairing relationship between different points within a given pair of anchored sets even if $N \rightarrow \infty$.

B.2 How to Choose S ?

In the previous section, we assumed that the center for each set is known. However, it might come as a problem in practice if the sets are not given *a priori* and if one has to resort to clustering methods such as k -means for finding the desired sets and estimating their centers. In this case, one has fix N , but variable S and N_s . The error in alignment now depends on both S and N_s : (1) as S gets small, then the error, as discussed in the previous section, increases as $\frac{1}{S}$; (2) smaller N_s makes it harder for us to estimate the center of each set, and the by the law of large numbers, we can estimate the center at error of order $\frac{1}{N_s}$. This incurs an error of order

$$\frac{c_1}{S} + \frac{c_2}{N_s} = \frac{c_1}{S} + \frac{c_2 S}{N} > 0$$

for some constants c_1, c_2 . One can take derivative to find the optimal S^* such that the error is minimized:

$$-\frac{c_1}{S^2} + \frac{c_2}{N} = 0 \rightarrow S^* = \sqrt{\frac{c_1 N}{c_2}}, \quad (7)$$

i.e. S^* should scale with \sqrt{N} .

C Cross-modal Generalization vs Domain Adaptation

In this section we make both methodological and empirical comparisons with a related field of work, domain adaptation.

C.1 Methodological Differences

At a high-level, the core differences between cross-modal generalization and domain adaptation lies in the fact that domain adaptation assumes that both source and target data are from the same modality (e.g. image-only). As a result, these models are able to share encoders for both source and target domains [58]. This makes the alignment problem straightforward for this simplified version of the problem.

By sharing the encoders, however, these domain adaptation methods do not directly model $p(x_s, x_t | m_s, m_t)$ for the two different modalities we are considering, which therefore makes it unable to obtain the generalization guarantees we derived in Proposition 1. Without alignment, a domain adaptation method is unlikely to work well since $p(x_s, x_t | m_s, m_t)$ is not modeled directly except on a few anchor points that some methods uses explicitly [65]. On the other hand, our approach explicitly models $p(x_s, x_t | m_s, m_t)$ using meta-alignment which in turn provides the guarantees in Proposition 1, thereby helping cross-modal generalization to low-resource modalities and tasks.

C.2 Empirical Differences

To further emphasize these methodological differences, we modify several classical domain adaptation methods for our task to verify that it is indeed necessary to use separate encoders and perform explicit alignment for cross-modal generalization. In particular, we implement the following baselines:

1. **Shared:** We share encoders for both modalities as much as much possible. The only non-shared parameter is a linear layer that maps data from the target modality’s input dimension to the source so

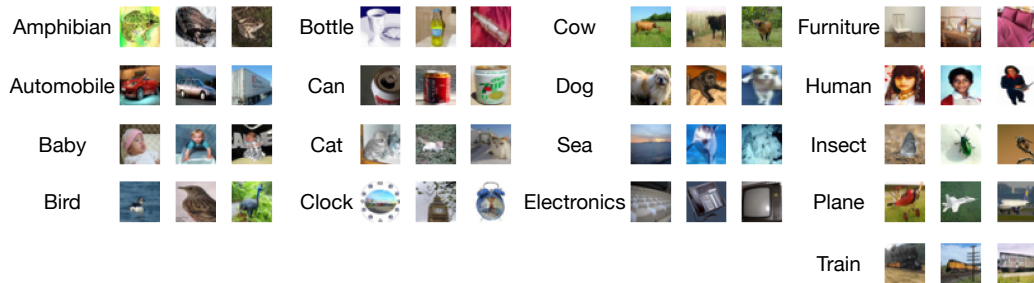



Figure 4: The 17 concepts shared across image and audio classification tasks that were used for weak alignment. Note that although we only show the images since, the audio spectrograms make up the second modality.

Table 6: Performance on text to speech generalization on the Wilderness dataset. We compare MAGMA with some standard domain adaptation baselines and observe that none of them perform well on cross-modal generalization. Although domain confusion and alignment do improve upon standard encoder sharing, they still fall short of our approach. This serves to highlight the empirical differences between cross-modal generalization and domain adaptation.

TYPE	APPROACH	1-SHOT	5-SHOT	10-SHOT	#SPEECH (LABELED)
DA	Shared	55.6 ± 10.2	75.2 ± 8.4	81.9 ± 3.9	4395(0)
	Shared + Align [32]	59.7 ± 7.6	78.4 ± 6.2	84.3 ± 1.5	4395(0)
	Shared + Domain confusion [58]	59.5 ± 7.2	76.3 ± 9.4	83.9 ± 1.8	4395(0)
	Shared + Target labels [31]	57.3 ± 9.3	76.2 ± 8.4	84.0 ± 1.9	4395(4395)
Cross	Align + Classify [10, 23, 48, 57, 60]	61.1 ± 6.0	74.8 ± 2.1	86.2 ± 0.7	4395(0)
	Align + Meta Classify [51]	65.6 ± 6.1	89.9 ± 1.5	93.0 ± 0.5	4395(0)
	MAGMA  (ours)	67.9 ± 6.6	90.6 ± 1.5	93.2 ± 0.2	4395(0)

that all subsequent encoder layers can be shared. This reflects classical work in domain adaptation and transfer learning [29, 56].

2. Shared + Align: We share encoders for both modalities and further add our alignment loss (contrastive loss) on top of the encoded representations, in a manner similar to our meta-alignment model (a similar reference in the domain adaptation literature would be [32]).

3. Shared + Domain confusion: We share encoders for both modalities and further add a domain confusion loss on top of the encoded representations [58].

4. Shared + Target labels: Finally, we share encoders for both modalities and also use target modality labels during meta-training, in a manner similar to supervised domain adaptation [31].

Results: We present these results in Table 6 and observe that none of them perform well on cross-modal generalization. Although domain confusion and alignment do improve upon standard encoder sharing, they still fall short of our approach. Our method also outperforms the Shared + Target labels baseline which uses target modality labels to train the shared encoder during meta-training. This serves to highlight the empirical differences between cross-modal generalization and domain adaptation. Therefore, we conclude that **1. separate encoders** and **2. explicit alignment** is important for cross-modal generalization which distinguishes it from domain adaptation.

D Experimental Details

The code for running our experiments can be found in the supplementary material. We also provide some experimental details below. Since there are no established benchmarks in cross-modal generalization, we create our own by merging and preprocessing several multimodal datasets. We believe that these two benchmarks for assessing cross-modal generalization (image to audio and text to speech) will also be useful to the broader research community and hence we also open-source all data and data processing code.

D.1 Image to Audio

Data: To construct our generalization dataset, we combine 100 classes from CIFAR-100 and 10 classes from CIFAR-10 [36] to form 110 image classes, as well as 50 audio classes from ESC-50 [46]. The tasks across these modalities are different (i.e. different classification problems) which requires

Table 7: Table of hyperparameters for generalization experiments on image to audio task. Batchsize 4/8/16 indicates the batchsize used for 1/5/10-shot experiments respectively.

Model	Parameter	Value
Image Encoder	Shared layer size	256
	Batchsize	4/8/16
	Activation	ReLU
	Meta Optimizer	SGD
	Optimizer	Adam
	Meta Learning rate	$1e-1$
	Align Learning rate	$1e-3$
	Classifier Learning rate	$1e-3$
	Iterations	800
	Number of evaluation tasks	16
	Loss Margin	0.1
	Width Factor	1.3
	Number of Layers	4
	Blocks Per Layer	4, 5, 24, 3
Audio Encoder	Shared layer size	256
	Batchsize	4/8/16
	Activation	ReLU
	Meta Optimizer	SGD
	Optimizer	Adam
	Meta Learning rate	$1e-1$
	Align Learning rate	$1e-3$
	Classifier Learning rate	$1e-3$
	Iterations	800
	Number of evaluation tasks	16
	Loss Margin	0.1
	Intermediate Pooling Function	Max
	Final Pooling Function	Average
	Stride	1
Audio Decoder	Num hidden layers	1
	Hidden layer size	256
	Batchsize	4/8/16
	Activation	ReLU
	Meta Optimizer	SGD
	Optimizer	Adam
	Meta Learning rate	$1e-1$
	Align Learning rate	$1e-3$
	Classifier Learning rate	$1e-3$
	Iterations	800
	Number of evaluation tasks	16
	Loss	MSE
	Teacher forcing rate	0.5

cross-modal generalization. To bridge these two modalities with partially related label spaces, we define 17 shared classes across the 2 datasets for weak concept alignment. We show the 17 clustered concepts we used for weak alignment in Figure 4. These clusters are obtained by mapping similar classes between the datasets using similarities from WordNet [39] and text cooccurrence. The number of shared classes in train, val, and test, respectively is 12, 8, and 9, and the number of samples is 920, 580, 580, respectively.

Hyperparameters: We show the hyperparameters used in Table 7.

D.2 Text to Speech

Data: The dataset is composed of paired text-speech data from a 99-language subset of the Wilderness dataset [6]. The dataset was collected using text and speech from the Bible. We preprocessed the data so that every language corresponded to a different set of chapters, maximizing the independence

Table 8: Table of hyperparameters for generalization experiments on text to speech task. Batchsize 4/8/16 indicates the batchsize used for 1/5/10-shot experiments respectively.

Model	Parameter	Value
Text Encoder	Bidirectional	True
	Embedding dim	256
	Num hidden layers	1
	Hidden layer size	256
	Batchsize	4/8/16
	Activation	ReLU
	Meta Optimizer	SGD
	Optimizer	Adam
	Meta Learning rate	$1e-1$
	Align Learning rate	$1e-3$
	Classifier Learning rate	$1e-4$
	Iterations	800
	Loss Margin	0.1
	Number of evaluation tasks	16
Model	Parameter	Value
Speech Encoder	Embedding dim	40
	Num hidden layers	2
	Hidden layer size	256
	Batchsize	4/8/16
	Activation	ReLU
	Meta Optimizer	SGD
	Optimizer	Adam
	Meta Learning rate	$1e-1$
	Align Learning rate	$1e-3$
	Classifier Learning rate	$1e-4$
	Iterations	800
	Loss Margin	0.1
	Number of evaluation tasks	16
Model	Parameter	Value
Speech Decoder	Num hidden layers	1
	Hidden layer size	256
	Batchsize	4/8/16
	Activation	ReLU
	Meta Optimizer	SGD
	Optimizer	Adam
	Meta Learning rate	$1e-1$
	Align Learning rate	$1e-3$
	Classifier Learning rate	$1e-4$
	Iterations	800
	Loss	MSE
	Teacher forcing rate	0.5
	Number of evaluation tasks	16

between datapoints across languages. We chose a random $0.8 - 0.1 - 0.1$ split for train-val-test with respect to language for (79 languages, 9 languages, 10 languages), and the number of samples is 4395, 549, 549 for meta-train, meta-validation, and meta-test respectively. There is no overlap between the data used for source classification, target classification, and alignment tasks.

Hyperparameters: We show the hyperparameters used in Table 8

Table 9: 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)
	Unsup. meta-learning [25] (weak labels)	45.6 \pm 1.3	74.2 \pm 0.3	83.7 \pm 0.1	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)

E Additional Results

E.1 Image to Audio

Extra results: We implement one more baseline derived as variations from existing work and adapted to our cross-modal generalization task. We adapt unsupervised meta-learning [25] which uses the aforementioned 17 weak clusters as prediction targets for the target modality during meta-training. This gives more discriminative training signal than the self-supervised reconstruction objective discussed in main text while still not explicitly using target modality labels during meta-training. We show these results in Table 9. While this baseline does do better than the reconstruction version of unsupervised meta-learning, it still underperforms as compared to MAGMA.

E.2 Text to Speech

Extra results: We present some extra results by comparing with existing baselines in domain adaptation and transfer learning (see Section C.2) in Table 6. We observe that none of them perform well on cross-modal generalization. Although domain confusion and alignment do improve upon standard encoder sharing, they still fall short of our approach. This serves to highlight the empirical differences between cross-modal generalization and domain adaptation. Therefore, we conclude that **1. separate encoders** and **2. explicit alignment** is important for cross-modal generalization and distinguishes it from domain adaptation.

F Response to NeurIPS reviews

[Meta reviewer other tasks] We added a new experimental setup on text to speech generalization (section 5.2) to show that our approach works for another set of 2 different modalities. This shows that we can easily extend our approach beyond the CIFAR-ESC dataset we constructed. We also showed robustness to noisy target labels, so it is possible to align spaces for other cross-modal tasks as well. Recent progress in unsupervised cross-modal alignment can also help in this regard.

[Meta reviewer, R3 novelty and comparisons] We added a long discussion with related work in section 2. In summary, both Huang et al. [1] and Cao et al. [2] are designed for weakly supervised cross-modal retrieval and **DO NOT actually study cross-modal generalization** from, for example, image classification to audio classification tasks. **Cross-modal generalization for classification tasks are harder since:** 1) one has to learn not just the associations between modalities in retrieval but also associations to labels, 2) there is weak supervision both the target modality and in the label space (see Table 1), 3) tasks in different modalities have different (but semantically related) label spaces, and 4) the presence of new data and labels in the target modality which have to be learned using only a few samples. As a result, we find that conventional contrastive objectives with weak supervision (used in retrieval and similar in spirit to [1,2]) **DO NOT** work well (see ablations in Table 2, Align + Classify and Align + Meta Classify). Meta-alignment is necessary for cross-modal generalization. However, we acknowledge that their ideas are also related since they also tackle problems related to having only weak supervision, and have included appropriate discussions in the paper (section 2).

[Meta reviewer labeled samples] From Tables 3 and 4 we see that **our approach uses orders of magnitude less labeled data in the target modality** as compared to the oracle baseline (i.e. meta-learning within the target modality), which is consistent with the setting claimed in Introduction (labeled data in target modality is scarce). Although domain knowledge from Wordnet or other

sources are helpful in automatically mining for weak clusters, this is still much easier to obtain than actually labeling data in the target modality. Furthermore weak clusters can also be obtained without domain knowledge - for example one can use videos as weak alignment between visual, audio, and text modalities for cross-modal generalization.

[Meta reviewer clarity] We significantly edited the paper to main the message clearer and also changed the title - we want to emphasize the core contribution towards learning in low-resource modalities (e.g. speech and text in rare languages) by leveraging high-resource modalities. We also clarified the notation in the problem formulation (section 3) that was confusing some reviewers, and better linked the theoretical insights with the algorithm design (link between sections 3 and 4).

[R1, R3 weak pairs] Weak pairs can be generally found for any modality without large amounts of labeled data since they are obtained by using external knowledge bases relating different modalities e.g. scene graphs relating images and word concepts, videos relating coarse image frames and corresponding text and audio. We also emphasize that the domain of weakly-paired modalities are usually different from the domains of labeled source and target modalities, and our proposed algorithm can handle this type of domain shift between alignment and classification tasks.

[R1 equation 1] We missed a minus sign in front of Eq. 1 and fixed this error. With this minus sign in place, Eq.1 can be written as a KL loss between two distributions, and KL divergence is always positive.

[R1 single-modal] Yes, we do see applications of this method in single-modal domain adaptation which are interesting directions for future work but outside the scope of this paper. We believe that explicit meta-alignment can help for domain adaptation as well.

[R2 motivation] We added a long discussion with co-learning in section 2. Essentially, co-learning is unable to solve problems in an unseen target modality. Rather, co-learning 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 modality**. Therefore, co-learning falls under the same category as 'Within modality generalization + cross-modal learning' in Table 1. To the best of our knowledge, our approach is the **first to tackle generalization from a source to target modality**.

[R2 definition] x_i indexes each of the M modalities, y_j indexes each of the N tasks, m_i indexes a prior over the M modalities and n_j indexes a prior over the N tasks. The goal is to transfer a model that performs well on classifying one modality (e.g. text) to one that quickly learns to classify another modality (e.g. audio). The entire problem is then called generalization from a text classification to audio classification problem. We clarified this in section 3.

[R2 table 1] In Table 1, we use meta-train and meta-test to generalize conventional train and test, but the same guidelines for data and labels hold for transfer learning etc. We chose this taxonomy to make the train and test splits clear for both conventional and meta-learning approaches. We clarified this point in Table 1.

[R2 theory] The theoretical setup informs us that modeling alignment is necessary to learn across modalities and tasks (see Figure 2 and section 3.2). The theory proves that perfect alignment can achieve perfect generalization in the low-resource modalities (proposition 1). This allows us to design an approach based on alignment which we find achieves strong performance on several cross-modal generalization tasks.

[R2 symbols] Our indexing and notation is consistent. M is the total number of modalities and N is the number of tasks. For example, $p(x_i; m_i)$ is the conditional distribution of x_i given that x_i is drawn from the m_i modality; the subscript i serves as a link between n_i and x_i , and, therefore, the subscript i to x_i and m_i needs to appear together. Likewise, n_j refers to the distribution of task for the random variable y_j and so on.

[R2 initialization] We certainly do not claim novelty in meta-learning the general initialization parameters, and we provide appropriate citations to Finn et al. at places where we applied their method. The novelty is in the formalization of the new cross-modal generalization problem, the algorithm based on meta-alignment with strong and weak supervision, and the new cross-modal evaluation tasks.

[R3 alignment] We have improved the exposition greatly in section 3. In essence, alignment is formulated in section 3.2 with the aim of learning an estimator $p(x_s|x_t)$ mapping related pairs from

705 target to source modalities. In section 4.1, we approximate $p(x_s|x_t)$ using a contrastive estimator
706 from positive and negative pairs and extend it to **meta-alignment** by defining multiple alignment
707 tasks and explicitly optimizing for generalization from one task to another.

708 **[R5 encoder]** Both source and target encoders are trained using NCE loss during meta-alignment
709 tasks. The source encoder is also using supervised loss during meta-training in the source modality.
710 The target modality is trained using supervised loss during meta-testing in the target modality with
711 small number of target modality labeled data. We added this to the paper in section 4 (also see figure
712 2).

View Meta-Reviews

Paper ID

8716

Paper Title

Cross-Modal Generalization via Meta Alignment

META-REVIEWER #1

META-REVIEW QUESTIONS

1. Please recommend a decision for this submission.

Reject

3. Please provide a meta-review for this submission. Your meta-review should explain your decision to the authors. It should augment the reviews and communicate how the reviews, author response, and discussion were used to arrive at a decision. Dismissing or ignoring a review is not acceptable unless you have a good reason for doing so. If you want to make a decision that is not clearly supported by the reviews, perhaps because the reviewers did not come to a consensus, please justify your decision appropriately, including, but not limited to, reading the submission in full and writing a detailed meta-review that explains your decision.

The key contribution of the work is the proposed meta-alignment framework for the generalization capability between modalities. The weakly-paired (i.e. many-vs-many) data, is promising direction in small-data scenarios.

Weakness:

- strong or weak paired data can be easily obtained for other tasks beyond CIFAR and ESC?
- * Novelty : prior work, e.g. Huang et al. [1], Cao et al. [2].
- a number of labeled samples in target modality are required, which seems to be inconsistent with the setting claimed in Introduction (labeled data in target modality is scarce). Also, external knowledges from WordNet are required.

Reject. Improve Clarity of the paper and its motivation and how it is different from previous approaches such as co-learning

View Reviews

Paper ID

8716

Paper Title

Cross-Modal Generalization via Meta Alignment

Reviewer #1

Questions

1. Summary and contributions: Briefly summarize the paper and its contributions.

This paper describes MAGMA, a meta learning based approach for cross-modal generalization. In order to enable meta learning for different modalities, the paper first proposes a meta-alignment that can project inputs from different modalities into an aligned space, where an NCE loss is defined and used based on strong or weak paired data. In order to let NCE mechanism handle unseen inputs at test time, meta-alignment parameters are used to initialize the alignment models for each downstream tasks. Based on the learned aligned space, the method then trains a single classifier on top of the aligned representations to classify inputs across modalities. Evaluations are conducted on CIFAR and ESC datasets and outperforms several baselines.

The key contribution of the work is the proposed meta-alignment framework for the generalization capability between modalities.

2. Strengths: Describe the strengths of the work. Typical criteria include: soundness of the claims (theoretical grounding, empirical evaluation), significance and novelty of the contribution, and relevance to the NeurIPS community.

The theoretical definitions in Section 3 are clear and the proposed method is detailed enough for other researchers to reproduce.

The experimental numbers on the two used datasets look good and significantly outperform 4 different baselines.

3. Weaknesses: Explain the limitations of this work along the same axes as above.

Not sure whether strong or weak paired data can be easily obtained for other tasks beyond CIFAR and ESC. This could be further discussed besides the current discussions from line 201 to line 204.

4. Correctness: Are the claims and method correct? Is the empirical methodology correct?

The claims and the method described in this paper are sound and convincing.

5. Clarity: Is the paper well written?

The paper is well written.

6. Relation to prior work: Is it clearly discussed how this work differs from previous contributions?

the differences between this work and previous work are clearly discussed in Section 2.

7. Reproducibility: Are there enough details to reproduce the major results of this work?

Yes

8. Additional feedback, comments, suggestions for improvement and questions for the authors:

(1) Could the authors explain more about equation 1: why the loss is lower bounded by 0? Can $f_w(x,y,m,n)$ be smaller than $p(x,y|m,n)$?

(2) Does the proposed method work for single-modal tasks as well?

9. Please provide an "overall score" for this submission.

7: A good submission; accept.

10. Please provide a "confidence score" for your assessment of this submission.

3: You are fairly confident in your assessment. It is possible that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work. Math/other details were not carefully checked.

11. Have the authors adequately addressed the broader impact of their work, including potential negative ethical and societal implications of their work?

Yes

Reviewer #2

Questions

1. Summary and contributions: Briefly summarize the paper and its contributions.

My understanding for this paper includes: (1) the task is cross-model generalization (while the "cross" is not clear to me after reading this paper); and (2) the proposed method is based on meta-learning and aims to learn improved alignment among/between multi-modal data spaces.

2. Strengths: Describe the strengths of the work. Typical criteria include: soundness of the claims (theoretical grounding, empirical evaluation), significance and novelty of the contribution, and relevance to the NeurIPS community.

1. Formulating multi-modal tasks is interesting. Moreover, the few-shot setting is interesting and challenging.

2. This paper proposes a method to make use of the weakly-paired (i.e. many-vs-many) data, which is very promising direction in small-data scenarios.

3. Weaknesses: Explain the limitations of this work along the same axes as above.

1. The motivation of this paper is not clear. In line 6, the authors mentioned the drawback of existing multimodal learning methods is the simple setting that modalities at train and test are the same, however, this is not true. Referring to 'Multimodal Machine Learning: A Survey and Taxonomy', there is one method of multimodal learning called co-learning, the aim of which is to solve the unseen target modality problem. This is an example against the statement "existing multimodal learning methods is the simple setting that modalities at train and test are the same".

2. The definition of cross-modal generalization problem is not easy to understand for me. For example, in line 89, what is n_N or m_M ? Is the n one of the variables within $1-N$? Overall question on this confusing definition is whether to solve the problem "a model trained on image-to-audio dataset does not perform well on a text-to-speech dataset" or "a model trained on classifying sentences should perform well on classifying images"??? Given the confusing description or formulation of the task, I am very concerned about the quality of this paper.

3. In table 1, I don't understand why should transfer learning, unsupervised pre-training, domain adaptation, cross-modal learning be framed in the setting of "meta-train and meta-test"? I think transfer learning, unsupervised pre-training, domain adaptation, and cross-modal learning are all containing conventional train and test stages. Meta-learning is based on multiple learning tasks and unseen test tasks (e.g. for handling the few-shot learning tasks).

4. What are the relations between the theory and the instantiation?

5. The meaning of the math symbols in the theory is unclear. For example, what's the difference between D^{x_m} and D_m . What's the meaning of x_1 , x_M , y_1 and y_N ?

6. In the instantiation, it is not novel to meta-learn the general initialization parameters. Please double check the related works from Chelsea Finn (eg. [17] "Model-agnostic meta-learning for fast adaptation of deep networks",

7. I guess a lot of `\vspace{-xxx}` are used for compressing the space before and after the titles of sections. This is not allowed in principle. Please keep the main content of the paper within 8 pages (final will be within 9 pages) without compressing any line space.

4. Correctness: Are the claims and method correct? Is the empirical methodology correct?

As this paper is not easy to read and understand, I am not sure about the correctness.

5. Clarity: Is the paper well written?

No.

6. Relation to prior work: Is it clearly discussed how this work differs from previous contributions?

Not exactly. For example, learning model initialization is not novel as it has been proposed in [17].

7. Reproducibility: Are there enough details to reproduce the major results of this work?

No

8. Additional feedback, comments, suggestions for improvement and questions for the authors:

Please refer to my detailed questions in "Weaknesses". I think solving those questions can improve the paper.

After rebuttal, sorry that I insist on my original rating, due to (1) the answers do not convince me, which means they lack motivation of doing such cross modality tasks (why not adapt same modality data/models that are definitely easier and more sensible to deep learning models); and (2) the paper is difficult to read and follow.

9. Please provide an "overall score" for this submission.

3: A clear reject.

10. Please provide a "confidence score" for your assessment of this submission.

4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.

11. Have the authors adequately addressed the broader impact of their work, including potential negative ethical and societal implications of their work?

Only partially, more discussion is needed.

Reviewer #3

Questions

1. Summary and contributions: Briefly summarize the paper and its contributions.

The paper proposes a cross-modal generalization algorithm that uses data from abundant source modalities to learn useful representations for scarce target modalities. Experimental comparisons show its superiority in the settings where the target modality suffers from noisy or limited data.

After read all the comments from other reviews and the rebuttal of the authors. I believe the paper still needs further polishing. Therefore, I will remain my score unchanged.

2. Strengths: Describe the strengths of the work. Typical criteria include: soundness of the claims (theoretical grounding, empirical evaluation), significance and novelty of the contribution, and relevance to the NeurIPS community.

+The paper is well organized. The figures are illustrative.

+Comprehensive experiments show good performance improvement.

+Experimental comparisons show more robust performance over Oracle in noisy labels setting.

3. Weaknesses: Explain the limitations of this work along the same axes as above.

1. The paper is conceptually solid. However, it is hard to claim novelty in the paper. The idea of transferring task to cross-modal target domain from a different source modality using an auxiliary cross-modal (or single modal) dataset has been explored in prior work, e.g. Huang et al. [1], Cao et al. [2].
2. To construct weak pairs, a number of labeled samples in target modality are required, which seems to be inconsistent with the setting claimed in Introduction (labeled data in target modality is scarce). Also, external knowledges from WordNet are required.
3. The explanation needs to be improved. There are some details to the method which are not explained clearly. For example, the exact formulation of alignment task T_a during meta-training lacks clarity.

[1] "Cross-modal Common Representation Learning by Hybrid Transfer Network". Huang et al. IJCAI 2017.

[2] "Transitive Hashing Network for Heterogeneous Multimedia Retrieval". Cao et al. AAAI 2017.

4. Correctness: Are the claims and method correct? Is the empirical methodology correct?

Yes

5. Clarity: Is the paper well written?

Yes

6. Relation to prior work: Is it clearly discussed how this work differs from previous contributions?

-Please refer to weaknesses.

7. Reproducibility: Are there enough details to reproduce the major results of this work?

Yes

8. Additional feedback, comments, suggestions for improvement and questions for the authors:

-Please refer to weaknesses.

9. Please provide an "overall score" for this submission.

3: A clear reject.

10. Please provide a "confidence score" for your assessment of this submission.

4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.

11. Have the authors adequately addressed the broader impact of their work, including potential negative ethical and societal implications of their work?

Yes

Reviewer #5

Questions

1. Summary and contributions: Briefly summarize the paper and its contributions.

The authors propose a method of cross-modal generalization via meta alignment which can be seen as an application of meta-learning on domain adaptation. By aligning different modalities through alignment task and mapping the correspondence of the different modalities into a fine-tuned source modality classifier to get the final result. The authors also create a meta-learning dataset for this paper.

2. Strengths: Describe the strengths of the work. Typical criteria include: soundness of the claims (theoretical grounding, empirical evaluation), significance and novelty of the contribution, and relevance to the NeurIPS community.

1. The idea of combining meta-learning and domain adaptation is novel where the need for large human-annotated datasets might be reduced.
2. The paper is well-written so it is a pleasure to read it. Thanks. Especially, the formulation part.
3. The experiment part is thorough, while the method gets much better results than the compared methods.

3. Weaknesses: Explain the limitations of this work along the same axes as above.

Some unclear parts to me (may not be weaknesses):

1. If I understand correctly, the proposed method heavily relies on the performance of encoder structures. While to train a good encoder may need a lot of supervision. According to Algorithm 1, the encoder is updated only using NCE loss. Are there may other supervision? Or NCE loss is enough in your setting? Correct me if I am wrong.

Thanks.

2. While training the classifier, the encoder is fixed according to Algorithm 1. So, is this the case during training? (update classifier with encoder fixed)

4. Correctness: Are the claims and method correct? Is the empirical methodology correct?

To the best of my knowledge, yes.

5. Clarity: Is the paper well written?

Yes. The authors show clear motivation about meta alignment.

6. Relation to prior work: Is it clearly discussed how this work differs from previous contributions?

Yes. The paper has thorough Related work. Thanks.

7. Reproducibility: Are there enough details to reproduce the major results of this work?

Yes

8. Additional feedback, comments, suggestions for improvement and questions for the authors:

%post-rebuttal

Thanks for answering my question during rebuttal.

As the other reviewers suggested, the related work part is somewhat incomplete, Also, the motivation should be made clearer and more details are encouraged.

Therefore, I would like to keep my original score.

9. Please provide an "overall score" for this submission.

5: Marginally below the acceptance threshold.

10. Please provide a "confidence score" for your assessment of this submission.

3: You are fairly confident in your assessment. It is possible that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work. Math/other details were not carefully checked.

11. Have the authors adequately addressed the broader impact of their work, including potential negative ethical and societal implications of their work?

Yes