Learning Modality Knowledge Alignment for Cross-Modality Transfer

Wenxuan Ma [1] **Shuang Li** [1] **Lincan Cai** [1] **Jingxuan Kang** [2]

Abstract

Cross-modality transfer aims to leverage large pretrained models to complete tasks that may not belong to the modality of pretraining data. Existing works achieve certain success in extending classical finetuning to cross-modal scenarios, yet we still lack understanding about the influence of modality gap on the transfer. In this work, a series of experiments focusing on the source representation quality during transfer are conducted, revealing the connection between larger modality gap and lesser knowledge reuse which means ineffective transfer. We then formalize the gap as the knowledge misalignment between modalities using conditional distribution $P(Y|X)$. Towards this problem, we present **Mo**dality k**N**owledge **A**lignment (MoNA), a meta-learning approach that learns target data transformation to reduce the modality knowledge discrepancy ahead of the transfer. Experiments show that out method enables better reuse of source modality knowledge in cross-modality transfer, which leads to improvements upon existing finetuning methods

1. Introduction

Transferring knowledge from past experience to new tasks is a fundamental ability of human intelligence (Pan & Yang, 2010; Long et al., 2018; Zhuang et al., 2020). Such an ability to acquire and reuse knowledge is continuously pursued in machine learning community, aiming to build artificial intelligence systems that predicts more accurately and learns more data-efficiently. Today, as large fundation models that are trained on massive data are widely available (Bai et al., 2023; Touvron et al., 2023; Liu et al., 2023), using such pretrained model as powerful feature extractor for new tasks has become a common practice of tranfer learning (Ding et al., 2023; Zhang et al., 2023a). Naturally, the pretrained
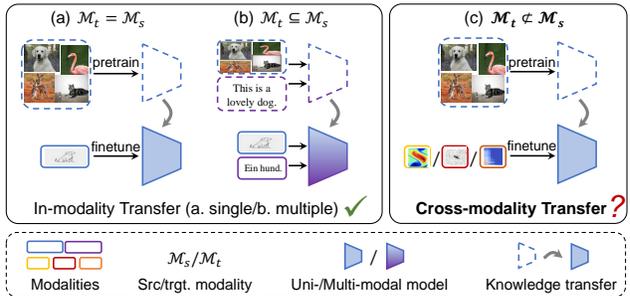
Figure 1. Comparison between in-modality finetuning (a)(b) and cross-modal finetuning (c). Both unimodal and multimodal finetuning are considered to be in-modality finetuning because the target modality is in the scope of the pretrained model's modality. In contrast, cross-modal finetuning exploits the pretrained model on target modalities that the pretrained model is not trained on.

model and the downstream task come from the same modality, e.g., the model is a vision transformer pretrained on ImageNet (Dosovitskiy et al., 2020) and the task is CIFAR-100 classification (Krizhevsky et al., 2009). However, recent stuides have been attempt to broaden this boundary to cross-modality transfer, using vision transformer for audio classification (Lin et al., 2023), and finetuning language model for tabular data (Dinh et al., 2022; Zhou et al., 2023). Fig. 1 illustrates the difference between in-modality and cross-modality transfer.

The motivation of such cross-modal transfer is easy to comprehend, especially when the target modality is data scarce. Scientific tasks like electrocardiogram classification (Clifford et al., 2017) and protein distance prediction (Adhikari, 2020) find difficulties in collecting large amount of training data, and further requires much expensive annotating costs from human experts. In such cases, it is desirable to leverage the pretrained model from other modalities like vision and language, in which data are easier to collect, to help the target modality tasks. However, the cross-modal transfer is not as straightforward as the in-modality transfer due to two challenges: 1) the input space and the label space are different across modalities, and 2) the knowledge required for addressing tasks in different modality may also differ.

Previous works tackle the first challenge by designing modality-specific embedders and predictors to interface with

---

[1]Beijing Institute of Technology, Beijing, China [2]University of Illinois Urbana-Champaign, USA. Correspondence to: Shuang Li <shuangli@bit.edu.cn>.

the pretrained model from input to output. However, the second challenge have not yet been well addressed. Some approaches (Zhang et al., 2023b; Han et al., 2023) argue that the large pretrained model can be served as universal encoder and thus freeze the pretrained model during finetuning. Other methods (Lu et al., 2022; Shen et al., 2023; Dinh et al., 2022) finetune the pretrained model along with modality-specific components. Both line of works empirically show that the pretrained model can be transferred to other modalities. Still, the key problem of *what knowledge from source modality is transferred via the pretrained model and how does it benefit the target modality* remains unsolved. For instance, ORCA (Shen et al., 2023) observes that training the model from scratch on some target modality tasks is even better than the vanilla finetuning of the pretrained model, which indicates that the knowledge contained in the pretrained model may not improve target performance if it is not properly transferred.

In this work, we delve deeper into this second challenge of cross-modal transfer. We begin with experiments investigating how target modality finetuning affects the representation quality of the source modality data. It is observed that finetuning a pretrained Swin Transformer (Liu et al., 2021) on some target modality tasks can help the Swin encoder to extract more discriminative features for images, while finetuning on other modalities impairs such ability. This empirical observation shows that there may exist aspects of knowledge, which we refer to as *modality semantic knowledge*, that differ between modalities in different degree and affect the validity of cross-modal transfer.

To specify such aspect of difference between modalities, we interpret the modality semantic knowledge as the conditional probability distribution $P(Y|X)$. We modify the conditional distribution of the source modality according to the tasks in target modality to make the two comparable. Consequently, we are able to formalize the *modality knowledge discrepancy* in terms of the divergence between conditional distributions of source and target modality. When the target conditional distribution is similar to the modified source conditional distribution, we say that the modality semantic knowledge is aligned and the source discriminative function learned by the pretrained model can be reused for the target modality. On the opposite, the modality semantic knowledge contradict each other and may not be mutually beneficial, which explains the observation in ORCA.

Our interpretation provides a new perspective towards understanding the effectiveness of the two-stage tuning pipeline proposed by previous cross-modal transfer works (Dinh et al., 2022; Shen et al., 2023): viewing the first stage as an implicit data transformation learning for target modality such that the conditional distribution on the transformed data are more aligned with source. As a result, it enlightens us

to directly learn a proper target embedding function ahead of finetuning, which helps minimize the knowledge misalignment. To this end, we propose a new method, MoNA, that improves the cross-modal transfer with two-stage training. In the first stage, MoNA leverages meta learning to learn an optimal target embedder which, when served as an initialization along with the pretrained weights for the full finetuning, allows a maximum reuse of the source modality knowledge during full finetuning. In the second stage, using the learned target embedder as the starting point, we follow the vanilla finetuning approach and update all the parameters to adapt to the target task while maximally leveraging source knowledge.

We conduct extensive experiments on two cross-modal transfer benchmarks, NAS-Bench-360 (Tu et al., 2022) and PDEBench (Takamoto et al., 2022), to validate our hypothesis and the effectiveness of our proposed methods. Both benchmarks focus on scientific problem related modalities, in which the training data scarcity is particularly acute. Comparisons of MoNA against previous methods are made, in which the results show that our method performs superior.

2. Problem Formulation and Analysis

In this section, we propose to test an assumption commonly made by previous cross-modal transfer approaches (Lu et al., 2022; Shen et al., 2023; Dinh et al., 2022; Zhang et al., 2023b) that the pretrained model can serve as a universal encoder for different modalities. Our experiments lead to an intuitive conclusion that the knowledge gap between modalities are not the same, and thus the assumption should take the modality knowledge discrepancy into consideration.

2.1. Introduction to basic notations and architecture

We consider the knowledge transfer between source modality $\mathcal{M}^s$ and target modality $\mathcal{M}^t$. Data in source modality, such as vision or language, is easier and cheaper to obtain, and large pretrained models are publicly available. Instead, the target modality considered in this paper has insufficient data to pretrain its own large models. The two modalities differ in both input space and label space, i.e., $\mathcal{X}^s \neq \mathcal{X}^t$, $\mathcal{Y}^s \neq \mathcal{Y}^t$. Cross-modal transfer aims to leverage a source pretrained model, parameterized by $\boldsymbol{\theta}^S$, to help a given target task with a small set of labeled data $\{\boldsymbol{x}_i^t, y_i^t\}_{i=1}^{n_t}$.

Following previous works (Lu et al., 2022; Shen et al., 2023), our model architecture $g_{\boldsymbol{\theta}}$ includes an embedder $e(\cdot; \boldsymbol{\theta}_e)$, a transformer encoder $f(\cdot; \boldsymbol{\theta}_f)$ and a predictor $h(\cdot; \boldsymbol{\theta}_h)$, and the parameter of the full model is denoted as $\boldsymbol{\theta} = \{\boldsymbol{\theta}_e, \boldsymbol{\theta}_f, \boldsymbol{\theta}_h\}$. Particularly, the pretrained transformer has its own embedder and predictor, and thus we denote the pretrained weights of the source model as $\boldsymbol{\theta}_0^S = \{\boldsymbol{\theta}_{e_0}^S, \boldsymbol{\theta}_{f_0}^S, \boldsymbol{\theta}_{h_0}^S\}$.

Embedding of Source Modality Data (CIFAR-10) Generated by Encoders Fine-tuned on Different Target Tasks

| After Finetuning on CIFAR-100 | After Finetuning on Spherical | Pretrained on ImageNet | After Finetuning on NinaPro | After Finetuning on FSD50K |
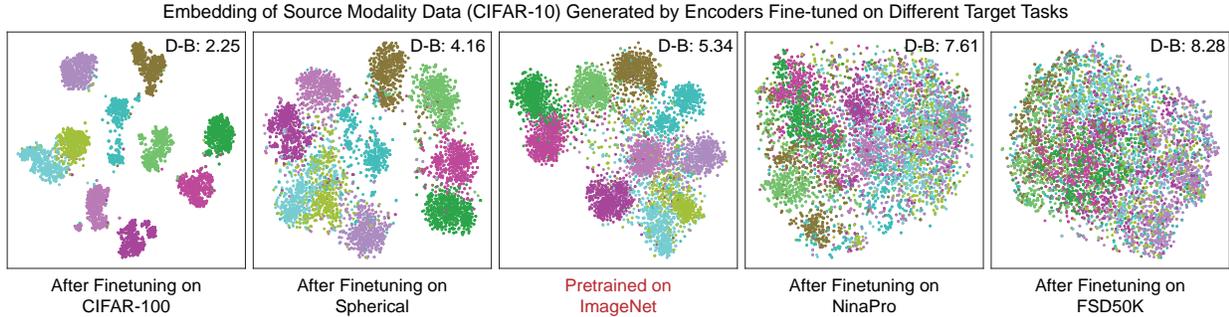
Figure 2. T-SNE visualization showing that the gap between different modalities are not the same. Figure in the middle depicts the embeddings of CIFAR-10 generated by an ImageNet pretrained Swin Transformer. The rest four figures are the embeddings of CIFAR-10 generated by the same model after being finetuned on different modalities. **None of these models are trained on CIFAR-10 directly**. Nevertheless, finetuning on CIFAR-100 and Spherical improve the visual representation from pretrained model while finetuning on NinaPro and FSD50K distort it. Davies-Bouldin indexes are shown at upper-right corner. Smaller index means better clustering.

The embedder maps the input data into a shared input embedding space $\hat{\mathcal{X}}$, and the encoder extracts features from the embedded input. The predictor is a linear layer that maps the encoder output to the label space. For our target model $g_{\boldsymbol{\theta}^{\mathcal{T}}} : \boldsymbol{\theta}^{\mathcal{T}} = \{\boldsymbol{\theta}_e^{\mathcal{T}}, \boldsymbol{\theta}_f^{\mathcal{T}}, \boldsymbol{\theta}_h^{\mathcal{T}}\}$, both embedder and the predictor are specifically re-designed to accommodate the input and label space of the target task, while we use $\boldsymbol{\theta}_{f_0}^{\mathcal{S}}$ to initialize the encoder weight $\boldsymbol{\theta}_f^{\mathcal{T}}$.

The flexibility of such architecture enables end-to-end training on the target task. **Vanilla finetuning** simply updates all the parameters of the target model by minimizing the task-specific loss on the given training dataset:

$$\boldsymbol{\theta}_{\mathcal{T}}^* = \arg\min_{\boldsymbol{\theta}_{\mathcal{T}}} \sum_{i=1}^{n_t} \ell\left(g_{\boldsymbol{\theta}^{\mathcal{T}}}(\boldsymbol{x}_i^t), y_i^t\right), \qquad (1)$$

where ℓ is the task loss function such as cross-entropy. Learning directly from target supervision in this way encourages the model to learn knowledge that help discriminate the target data. As the pretrained model already contains source discriminative knowledge, it is natural for cross-modal transfer to expect that source and target knowledge share similarities in some aspects so that the source knowledge can be reused to promote target learning. In the following, we 1) conduct experiments to show that this similarity depends on the modality, and 2) provide interpretation of modality knowledge and formalize the knowledge discrepancy.

2.2. Distortion of learned source modality knowledge

We look for a quantitative way to compare the extent of knowledge reuse among various cross-modal transfer scenarios. In this section, we select image modality as knowledge source and choose four target tasks from different modalities. We include two tasks closely related to images: CIFAR-100 (Krizhevsky et al., 2009), Spherical (Cohen et al., 2018) that contains spherically projected images, and two tasks dissimilar to image modality: NinaPro (Atzori et al., 2012) that represents hand gestures with electromyography signals, FSD50K (Fonseca et al., 2017b) that contains

audio clips of sound events. To be specific, we adopt Swin Transformer Base pretrained on ImageNet-22k as the source model and examine the properties of the model after finetuning it on different tasks.

Given that the comparison is conducted across distinct modalities, there lacks a general metric measuring the degree of knowledge reuse during transfer. Therefore, we turn to compare the distortion of source knowledge. Specifically, we would expect smaller distortion if more source knowledge is reused to solve target task, and vise versa. So we leverage the pretrained source model to extract the visual representation of CIFAR-10, a surrogate image dataset unseen by the model. Samples in this particular source dataset are denoted as $\{\boldsymbol{x}_i^s, y_i^s\}$ and their corresponding feature set $\{\boldsymbol{f}_i^s = f(e(\boldsymbol{x}_i^s; \boldsymbol{\theta}_{e_0}^{\mathcal{S}}); \boldsymbol{\theta}_{f_0}^{\mathcal{S}})\}$. Then, we finetune the pretrained model on the four target tasks using Eq. (1) respectively. After the finetuning process, we once again extract the representation of CIFAR-10 using the finetuned encoder and obtain $\{\boldsymbol{f}_i^s(\mathcal{M}_t) = f(e(\boldsymbol{x}_i^s; \boldsymbol{\theta}_{e_0}^{\mathcal{S}}); \boldsymbol{\theta}_f^{\mathcal{T}}, \mathcal{M}_t)\}$. Fig. 2 shows the T-SNE visualization results of the five different sets of CIFAR-10 image features.

The figure illustrates that encoders finetuned on CIFAR-100 or Spherical maintain or even improve their discriminability on image samples in CIFAR-10, whereas the encoders finetuned on NinaPro and FSD50K can no longer extract class discriminative features for images. Considering that finetuning on target modality makes the encoder focus on classifying target data and learning target discriminative function, what this observation indicates is that the knowledge required for discriminating samples in CIFAR-100 and Spherical are more aligned with that for CIFAR-10, compared to the latter two modalities. Such conclusion aligns with our intuition, since CIFAR-100 is vision dataset and Spherical is originated from natural images, whereas NinaPro and FSD50K are less relevant to images.

On the flip side, the results shows that CIFAR-100 and Spherical can better reuse the source knowledge in the pre-

trained encoder for task solving while NinaPro and FSD50K require the encoder to make greater adjustments in order to adapt to target tasks.

To investigate the source knowledge reuse (or distortion) during cross-modal transfer more quantitatively, we use linear probing on CIFAR-10 to evaluate the quality of extracted representations with encoders finetuned 1) on different target modalities, 2) with different epochs, and 3) with different transfer methods. Additional to vanilla finetuning, we consider the following two baselines:

- **ORCA** (Shen et al., 2023) adds an embedder training stage before finetuning. This first stage only updates the target embedder parameters $\theta_e^{\mathcal{T}}$ to minimize the Optimal Transport Dataset Distance (Alvarez-Melis & Fusi, 2020) between source and target embeddings in the shared input space $\hat{\mathcal{X}}$.

- We propose another baseline modified from previous works (Kumar et al., 2022; Zhang et al., 2023b), Embedder warmup (**Emb**), which is also a two-stage training method. The first stage solely updates the target embedder using the same task loss as vanilla finetuning while keeping the rest of the network frozen. The second stage finetunes the full network.

Fig. 3 shows the *error* rate of linear probing, where the dash line shows the linear probing results on pretrained encoder as a reference. Note that all these results are error rates on CIFAR-10 dataset that reflects the extent to which the model retains the source modality knowledge. The performance comparison on target modalities is not what we concern right now and can be found later in table. 2. From experiments we observe that modality has the greatest effect on linear probing results. Finetuning on FSD50K significantly distorts the encoder and impairs its discriminability on image data. We also notice that tuning on target dataset for more epochs leads to larger distortion of source knowledge on all target modalities except for image modality (CIFAR-100). These observation leads to the conclusion that knowledge for discriminating samples in different modalities differ in varying degrees, which we refer to as the misalignment of modality semantic knowledge. We argue that a large discrepancy may hinder the effectiveness of cross-modal transfer, and thus the assumption that source modality pretraining is beneficial to target modality should depends on the discrepancy.

We make additional observations about the source knowledge preservation effect of two-stage training methods. We observe that both ORCA and Emb achieves lower source error compared to vanilla finetuning, and Emb performs better than ORCA. This suggests that the target embedder trained in their first stage implicitly learns a mapping from $\mathcal{X}^t$ to $\hat{\mathcal{X}}$ that mitigates the knowledge misalignment between target

and source, and thus reduces the model distortion during its adaptation towards target tasks.

The above experiments motivate us to formalize the discrepancy of modality semantic knowledge, and to propose an improved objective for training target embedders that reduce such discrepancy.
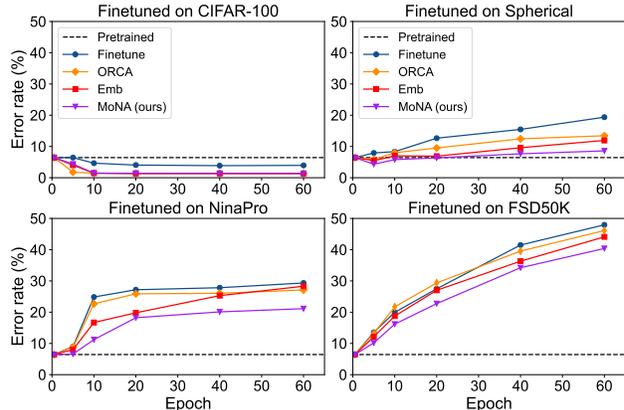
Figure 3. Linear probing results on CIFAR-10 using representations extracted by vision encoders finetuned on four different modalities and with different finetuning methods. "Pretrained" refers to the baseline that directly uses pretrained vision encoder.

2.3. Modality semantic knowledge discrepancy

We consider representing the semantic knowledge within a modality using the conditional distribution $P(Y|X)$, which describes the relationship between raw data space and semantic space of the modality. This is because for neural networks, acquiring the semantic knowledge means learning a mapping from data space to semantic space that resembles the true conditional distribution.

However, to measure the degree of alignment or "similarity" of such knowledge between two modalities is quite challenging. The difficulty lies in the fact that both the data space $\mathcal{X}$ and the label space $\mathcal{Y}$ are different and even non-overlapping across modalities.

Therefore, we need to modify the conditional distribution to make it comparable across modalities. Modifying the input space is rather easy, as we can simply embed the inputs into a shared space $\hat{\mathcal{X}}$ using modality-specific embedders. However, modifying the label space is more sophisticated.

Considering that source modalities, like vision and language, having large pretrained models are both rich in semantics, we make the following assumption: *The cardinality of source modality label space is larger than the cardinality of the label space of the target modality, i.e., $|\mathcal{Y}^s| \geq |\mathcal{Y}^t|$.*

This assumption is easily satisfied in practice. For example, vision transformers trained on ImageNet learns a discriminative function of one thousand categories whereas only four

classes are considered in an electrocardiogram classification task. With the assumption, we can select a subset of the source modality label space $\mathcal{Y}_{\mathcal{B}}^s \subset \mathcal{Y}^s$ such that $|\mathcal{Y}_{\mathcal{B}}^s| = |\mathcal{Y}^t|$. We further introduce a category permutation $\pi(\cdot)$ that adjusts the order of source classes. To this end, we can define a new label space of the source modality, namely the source subset after permutation $\mathcal{Y}_{\pi,\mathcal{B}}^s \triangleq \pi(\mathcal{Y}_{\mathcal{B}}^s)$. By measuring the discrepancy between the modified conditional distributions $P(Y_{\pi,\mathcal{B}}^s|\hat{X})$ and $P(Y^t|\hat{X})$, we can formalize the degree of alignment of the modality semantic knowledge as follows:

Definition 2.1. (Modality semantic knowledge discrepancy). *Given the source modality $\mathcal{M}^s$ and the target modality $\mathcal{M}^t$ satisfying the assumption, let $\hat{mathcalX}$ be the shared input space generated from raw data spaces by modality-specific embedders, and let $P(Y^s|\hat{X})$, $P(Y^t|\hat{X})$ be the conditional distribution for the source and target modality. Then, the modality semantic knowledge discrepancy between the two modalities is*

$$D(\mathcal{M}^s, \mathcal{M}^t) = \inf_{\pi,\mathcal{B}} d(P(Y_{\pi,\mathcal{B}}^s|\hat{X}), P(Y^t|\hat{X})),$$

where $d(\cdot, \cdot)$ is an arbitrary discrepancy measure between two conditional distributions.

The definition basically says that, if we can find an optimal subset within source semantics that, with a proper one-to-one matching between source semantics and target semantics, shares similar conditional distribution with target modality, then the knowledge discrepancy is considered to be small. The source model should be able to correctly distinguish target samples like the way it discriminates source samples within the subset.
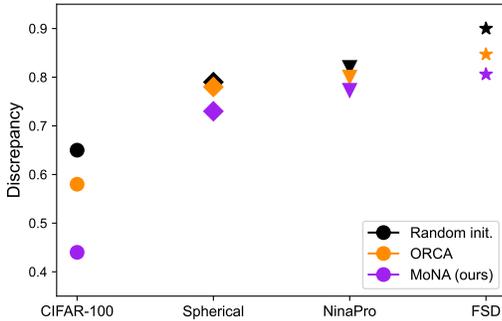
Figure 4. Modality knowledge discrepancy between image modality and four target modalities. Computation uses approximation.

With the definition, we calculate the modality semantic knowledge discrepancy between image modality and the four target tasks using an extreme approximating algorithm. Here we only demonstrate the results in Fig. 4 while leaving the implementation details in supplementary. Our calculation aligns with previous observations, showing that modalities do have different degree of knowledge discrepancy, and FSD50K is the most dissimilar modality from image modality among the four.

3. Modality Knowledge Alignment

Discovering that the modality knowledge may not be well aligned and its consequence of insufficient source knowledge reuse, we propose a new method that improves the modality knowledge alignment and the effectiveness of cross-modality transfer.

3.1. Embedder Warmup

In the previous experiments we find that Embedder warmup, in spite of its simplicity in training objective, preserves source knowledge better than other methods. Correspondingly, we turn to examine its performance on target modalities. Table 2 shows that Emb likewise surpasses its counterparts. We argue that during the embedder warmup, in order to minimize the task loss, the embedder are forced explicitly to project target original inputs into embeddings that are distinguishable by the source model, which is frozen and extract features according to the source knowledge.

Combined with our previous analysis, we hypothesize that the key to effective transfer is to learn a target embedding function $e^{\mathcal{T}} : \mathcal{X} \to \hat{\mathcal{X}}$ that makes the target conditional distribution $P(Y^t|\hat{X})$ more aligned with the source knowledge. Consequently, we propose to train the target embedder solely using objective in the next section to learn such embedding function ahead of the full finetuning process.

3.2. Learning to Align Modality Knowledge

Since we cannot estimate the target conditional probability without training a model, adopting the modality knowledge divergence directly as a objective for optimization is difficult. As an alternative, we propose to leverage meta learning pipeline to simulate the process in Fig. 3 and optimize the representation quality of source data after finetuning. Specifically, an ideal target embedder aligns the modality knowledge, allowing the encoder to retain its discriminability on image data during target finetuning. Therefore, if we use a source dataset to evaluate the finetuned encoder that is initialized by this ideal target embedder, we would obtain minimal error on the source data.

Such process is a standard bi-level optimization problem widely studied in meta-learning (Finn et al., 2017). Particularly in our case, the outer-loop updates the target embedder based on the outer-loop loss, which is computed using target encoder after inner-loop optimization. Fig. 5(a) illustrates a single step update of the embedder parameter ϕ_e in the outer-loop during meta learning, and Fig. 5(b) shows the process of bi-level optimization.

More specifically, the inner-loop is the optimization of the model on target dataset, subjected to the condition that the target embedder is initialized by ϕ_e, which is

$$\boldsymbol{\theta}^{\mathcal{T}^*}(\boldsymbol{\phi}_e) = \arg\min_{\boldsymbol{\theta}^{\mathcal{T}}} \mathcal{L}_{inner}(\boldsymbol{x}^t, y^t; \boldsymbol{\phi}_e), \qquad (2)$$

(a) Gradient computation of a single step update of ϕ_e

(b) Illustration of the bi-level optimization
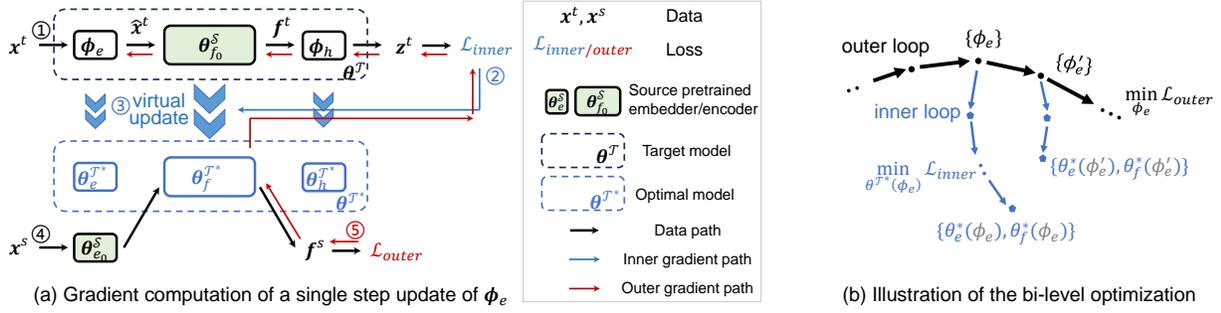
Figure 5. The framework of our proposed method. (a) illustrates a single update step of the embedder ϕ_e in the first stage of MoNA. The target data is forward propagated first to compute the inner-loop loss $\mathcal{L}_{inner}$, and the gradient is backpropagated to virtually update the full target model. Then, the updated encoder $\theta_f^{\mathcal{T}^*}$ receives source data embeddings from pretrained source embedder $\theta_{e_0}^{\mathcal{S}}$, and the outer-loop loss is computed using source features. Finally, the outer-loop gradient is used to update the embedder while the virtually updated model is discarded. (b) illustrated the bi-level optimization where the outer-loop updates ϕ_e according to inner-loop results.

where $\mathcal{L}_{inner}$ is the same loss as in Eq. (1), and

$$\boldsymbol{\theta}_{e_t}^{\mathcal{T}} = \boldsymbol{\theta}_{e_{t-1}}^{\mathcal{T}} - \alpha \nabla \mathcal{L}_{inner}, \quad \boldsymbol{\theta}_{e_0}^{\mathcal{T}} = \phi_e. \quad (3)$$

This inner-loop optimization simulates the full finetuning process in the second stage, and returns an encoder that is already adapted to target modality. Note that the whole optimal target model in the inner-loop depends on the initialization of the target embedder. Therefore we have $\boldsymbol{\theta}^{\mathcal{T}^*}(\phi_e) = \{\boldsymbol{\theta}_e^{\mathcal{T}^*}(\phi_e), \boldsymbol{\theta}_f^{\mathcal{T}^*}(\phi_e), \boldsymbol{\theta}_h^{\mathcal{T}^*}(\phi_e)\}$.

The outer-loop is an optimization problem with respect to the target embedder. Our goal is to find optimal embedder parameters ϕ_e^* such that the resulting optimal target encoder $\boldsymbol{\theta}_f^{\mathcal{T}^*}(\phi_e^*)$ generates high quality representations of source data. To calculate the loss, we leverage a small labeled dataset $\{\boldsymbol{x}_i^s, y_i^s\}$ in source modality as a surrogate and compute their features $\{\boldsymbol{f}_i^s = f(e(\boldsymbol{x}_i^s; \boldsymbol{\theta}_{e_0}^{\mathcal{S}}); \boldsymbol{\theta}_f^{\mathcal{T}^*}(\phi_e))\}$. Then we normalize these features onto a unit sphere and measure the alignment and uniformity of the source features (Wang & Isola, 2020). In particular, the alignment loss measures whether features from the same class are close, and the uniformity loss measures whether features from different classes are evenly distributed on the sphere.

Our outer-loop objective that measures the source discriminability of the induced encoder takes the following form:

$$\mathcal{L}_{outer} = \mathcal{L}_{align} + \mathcal{L}_{uniform}$$
$$= - \mathbb{E}_{i,j: y_i^s = y_j^s}[||\boldsymbol{f}_i^s - \boldsymbol{f}_j^s||_2^2] - \log \mathbb{E}_{i,j}\left[e^{-2||\boldsymbol{f}_i^s - \boldsymbol{f}_j^s||_2^2}\right]. \quad (4)$$

Notably, the source knowledge cannot be well-preserved at the beginning of the embedder training. To prevent the embedder from overly focusing on source modality, and also to keep the optimization process stable, we strike a balance between source and target knowledge learning by jointly minimizing the two objectives with a trade-off parameter λ:

$$\mathcal{L}'_{outer} = \lambda \mathcal{L}_{outer} + \mathcal{L}_{inner}. \quad (5)$$

Algorithm 1 MoNA: Modality Knowledge Alignment

Input: Source pretrained model $g_{\boldsymbol{\theta}_0^{\mathcal{S}}}$; Learning rate α, β; Maximum iterations I_1, I_2.
Output: Model for the target task: $g_{\boldsymbol{\theta}^{\mathcal{T}}}$.
Stage 1: Target embedder training
for $iter = 1, 2, \cdots, I_1$ **do**
    Initialize the target model $g_{\boldsymbol{\theta}^{\mathcal{T}}}$ with ϕ_e and $\boldsymbol{\theta}_{f_0}^{\mathcal{S}}$.
    Virtually update: $\boldsymbol{\theta}^{\mathcal{T}^*} = \boldsymbol{\theta}^{\mathcal{T}} - \alpha \nabla_{\boldsymbol{\theta}^{\mathcal{T}}} \mathcal{L}_{inner}$.
    Compute source features $\{\boldsymbol{f}_i^s\}$ with $\boldsymbol{\theta}_{e_0}^{\mathcal{S}}$ and $\boldsymbol{\theta}_f^{\mathcal{T}^*}$.
    Obtain outer-loop loss using Eq. (4).
    Update target embedder: $\phi_e \leftarrow \phi_e - \beta \nabla_{\phi_e} \mathcal{L}'_{outer}$.
end for
Stage 2: Full finetuning
for $iter = 1, 2, \cdots, I_2$ **do**
    Initialize the target model $g_{\boldsymbol{\theta}^{\mathcal{T}}}$ with ϕ_e and $\boldsymbol{\theta}_{f_0}^{\mathcal{S}}$.
    Update target model towards Eq. (1).
end for

In practice, we adopt single step update in the inner loop, a simplification that will be discussed in the analytical experiment section. This enables us to reuse the loss $\mathcal{L}_{inner}$ calculated during the inner-loop virtual update to compute this combined objective $\mathcal{L}'_{outer}$ efficiently. To this end, our proposed MoNA updates the target embedder in the first stage using:

$$\phi_e^* = \arg\min_{\phi_e} \mathcal{L}'_{outer}. \quad (6)$$

With the modality knowledge being better aligned, MoNA conducts vanilla finetuning in the second stage. The complete algorithm of MoNA is demonstrated in Alg. 1.

4. Experiments

In this section, we show experiments conducted on two cross-modal benchmarks. We follow the test protocol in ORCA (Shen et al., 2023) and evaluate MoNA on NAS-Bench-360 (Tu et al., 2022) and PDEBench (Takamoto

Table 1. Prediction errors (↓) on ten tasks of NAS-Bench-360.

	CIFAR-100 0-1 error (%)	Spherical 0-1 error (%)	Darcy Flow relative l_2	PSICOV MAE8	Cosmic 1-AUROC	NinaPro 0-1 error (%)	FSD50K 1-mAP	ECG 1-F1 score	Satellite 0-1 error (%)	DeepSEA 1-AUROC
Hand-designed	19.39	67.41	8.00E-3	3.35	0.127	8.73	0.62	0.28	19.8	0.3
NAS-Bench-360	23.39	48.23	2.60E-3	2.94	0.229	7.34	0.6	0.34	12.51	0.32
DASH	24.37	71.28	7.90E-3	3.3	0.19	**6.60**	0.6	0.32	12.28	**0.28**
Perceiver IO	70.04	82.57	2.40E-2	8.06	0.485	22.22	0.72	0.66	15.93	0.38
FPT	10.11	76.38	2.10E-2	4.66	0.233	15.69	0.67	0.5	20.83	0.37
ORCA	6.53	29.85	7.28E-3	1.91	0.152	7.54	0.56	0.28	11.59	0.29
MoNA	**6.48**	**27.13**	**6.80E-3**	**0.99**	**0.121**	7.28	**0.55**	**0.27**	**11.13**	**0.28**

Table 2. Comparing various cross-modal transfer methods on NAS-Bench-360.

	CIFAR-100 0-1 error (%)	Spherical 0-1 error (%)	Darcy Flow relative l_2	PSICOV MAE8	Cosmic 1-AUROC	NinaPro 0-1 error (%)	FSD50K 1-mAP	ECG 1-F1 score	Satellite 0-1 error (%)	DeepSEA 1-AUROC
Train-from-scratch	50.87	76.67	8.00E-2	5.09	0.50	9.96	0.75	0.42	12.38	0.39
Finetuning	7.67	55.26	7.34E-3	1.92	0.17	8.35	0.63	0.44	13.86	0.51
Finetuning++	6.60	33.17	7.50E-3	1.91	0.168	8.00	0.63	0.35	12.73	0.38
Frozen-encoder	10.02	59.62	6.84E-3	3.43	0.481	35.20	0.72	0.37	19.71	0.36
ORCA	6.53	29.85	7.28E-3	1.91	0.152	7.54	0.56	0.28	11.59	0.29
Emb	6.52	28.76	7.50E-3	1.35	0.139	7.74	0.56	0.28	11.40	0.29
MoNA	**6.48**	**27.13**	**6.80E-3**	**0.99**	**0.121**	7.28	**0.55**	**0.27**	**11.13**	**0.28**

et al., 2022). NAS-Bench-360 is a comprehensive benchmark that contains diverse tasks from ten different modalities. PDEBench covers a wide range of partial differential equations (PDEs) including challenging physical problems.

Each benchmark involves both tasks with 1D and 2D inputs. Following previous works, we adopt pretrained language model RoBERTa (Liu et al., 2019) and pretrained vision model Swin Transformer (Liu et al., 2021) for 1D and 2D tasks respectively. Following ORCA, We use CoNLL-2003 and CIFAR-10 as the source modality datasets to compute the outer-loop meta loss. Hyper-parameters for each task are in supplementary.

4.1. Results on NAS-Bench-360

Several experiments are conducted on this benchmark. The benchmark involves seven tasks with 2D inputs and three tasks with 1D inputs. Table 1 shows the results comparison using the *full* training set in each modality. We compare different kinds of baselines including training hand-designed and NAS-based architectures (NAS-Bench-360, DASH (Shen et al., 2022)) solely on target modalities, general purpose networks Perceiver IO (Jaegle et al., 2022), and cross-modal transfer methods FPT (Lu et al., 2022) and ORCA. We observe that MoNA achieves top performance on nine out of ten tasks, and surpasses previous cross-modal transfer methods on all tasks.

We also compares several variants of cross-modal transfer. In single stage methods, finetuning refers to the vanilla finetuning in Eq. (1). Finetuning++ initializes the target embedder using source embedder weights under certain modifications to map the dimension. Frozen encoder resembles to previous works (Zhang et al., 2023b) and it keeps the pretrained weight frozen during finetuning. In two-stage methods where a warmup stage is conducted before finetuning, we consider ORCA, Emb and MoNA. These

Table 3. Performance Comparison (↓) with Different Backbones.

Method	Pretrained Model (Size)	Pretrained Dataset	Spherical 0-1 error (%)	NinaPro 0-1 error (%)	FSD 1-mAP
Finetuning	ViT-Base (86M)	IN-22K	47.24	15.63	0.74
ORCA	ViT-Base (86M)	IN-22K	36.52	8.78	0.63
MoNA	ViT-Base (86M)	IN-22K	**33.34**	**8.00**	**0.62**
Finetuning	CLIP ViT-B/16 (86M)	WIT-400M	57.47	13.81	0.77
ORCA	CLIP ViT-B/16 (86M)	WIT-400M	43.12	8.50	0.69
MoNA	CLIP ViT-B/16 (86M)	WIT-400M	**41.35**	**7.59**	**0.67**
Finetuning	Swin-Base (88M)	IN-22K	55.26	8.35	0.63
ORCA	Swin-Base (88M)	IN-22K	29.85	7.54	0.56
MoNA	Swin-Base (88M)	IN-22K	**27.13**	**7.28**	**0.55**

experiments serves as an ablation study of our method and their results is reported in table 2.

The results lead to following conclusions: 1) finetuning the encoder helps the pretrained model adapt to target modality and performances better than frozen encoder. 2) Two-stage methods are generally superior than single stage methods, indicating that a proper target embedding function leads to better knowledge transfer. 3) Combined with the results on source knowledge preservation experiment in Fig. 3, we see that methods that better align modality knowledge achieves higher cross-modal transfer performance.

To show that MoNA achieves consistent performance improvement across different pretrained models, we conduct experiments on other two vision backbones, which are ViT-Base (Dosovitskiy et al., 2020) pretrained on ImageNet-22K and CLIP ViT-Base/16 (Radford et al., 2021) pretrained on WIT-400M. The results in table 3 show that, although the performance on target modalities varies due to the different capability of the pretrained models, MoNA consistently improves knowledge transfer and is superior to the State-of-the-Art method on different pretrained models.

4.2. Results on PDEBench

The benchmark includes eight PDEs with 1D/2D inputs, which we address similarly using language and vision

Table 4. Normalized Root Mean Squared Errors (nRMSEs, ↓) on 8 PDEBench tasks.

	Advection 1D	Burgers 1D	Diffusion-Reaction 1D	Diffusion-Sorption 1D	Navier-Stokes 1D	Darcy-Flow 2D	Shallow-Water 2D	Diffusion-Reaction 2D
PINN	0.67	0.36	0.006	0.15	0.72	0.18	0.083	0.84
FNO	0.011	**0.0031**	**0.0014**	0.0017	0.068	0.22	**0.0044**	**0.12**
U-Net	1.1	0.99	0.08	0.22	–	–	0.017	1.6
ORCA	0.0098	0.0120	0.0030	**0.0016**	0.062	0.081	0.0060	0.820
MoNA	**0.0088**	0.0114	0.0028	**0.0016**	**0.054**	**0.079**	0.0057	0.818

Table 5. Prediction errors (↓) on three classic benchmarks from different modalities.

	AudioSet-20k 1-mAP	ESC50 0-1 error (%)	UCF101 0-1 error (%)
From scratch	0.634	30.00	57.62
Finetuning	0.541	27.50	26.51
ORCA	0.538	12.75	16.86
MoNA	**0.523**	**9.64**	**13.45**

Table 6. Ablation Studies on Loss Design.

Loss Ablations	CIFAR-100	Spherical	NinaPro	FSD50K
MoNA w/o $\mathcal{L}_{inner}$ in Eq. (5)	8.00	28.76	7.74	0.58
MoNA	**6.48**	**27.13**	**7.28**	**0.55**
Contrastive Loss for $\mathcal{L}_{outer}$	6.51	27.90	7.44	0.55
Clustering Metric for $\mathcal{L}_{outer}$	7.09	28.02	8.19	0.56

Figure 6. Analytical experiments. Accuracies with varying value of trade-off λ (left two), and different inner-loop iterations (right).

Table 7. Comparison on Training Time in GPU Hour.

	CIFAR-100	Spherical	NinaPro	FSD50K
ORCA	11.57	12.40	1.42	16.20
MoNA	12.15	12.32	1.15	17.74
Performance Gain (relative)	+0.7%	+9.1%	+3.4%	+1.8%

pretrained models. We consider three baselines in the original paper, namely U-Net (Ronneberger et al., 2015), PINN (Raissi et al., 2019), FNO (Li et al., 2020), where the last two methods are specifically designed for PDEs. We also compare ORCA as the cross-modal transfer baseline. The result is shown in table 4.

We observe that MoNA achieves state-of-the-art on four out of eight tasks on PDEBench. Notably, it outperforms ORCA on seven tasks with significant improvements on Advection Equation and Navier-Stokes' Equation, and it achieves competitive results with specialized method FNO.

4.3. Results on Several Other Tasks

To further demonstrate the scalability of MoNA, we conduct experiments on more datasets and modalities, including AudioSet (Gemmeke et al., 2017) and ESC50 in audio modality, and UCF101 (Soomro et al., 2012) in video modality.

As shown in table 5, MoNA surpasses all baselines including ORCA. These results validate the scalability of our method, indicating that MoNA is a general cross-modality transfer method that can be applied on a wide range of tasks.

4.4. Analytical Experiments

Ablation Studies. To validate the effectiveness of our design, we conduct several ablation studies on four tasks in NAS-Bench-360. The first ablation study investigates the effect of different training objectives as shown in table 6. We begin with excluding the inner-loop loss $\mathcal{L}_{inner}$ from the total objective of the outer-loop in Eq. (5). Comparing to MoNA, the performance drop on all four tasks shows that

$\mathcal{L}_{inner}$ is crucial to ensure the adaptation of the embedder.

We move on to replace the alignment and uniformity loss (i.e., $\mathcal{L}_{outer}$) by two variants. The contrastive loss refers to the supervised contrastive loss (Khosla et al., 2020), and the clustering metric we adopt is the Davies-Bouldin index. The results in table 6 shows that contrastive loss achieves slightly worse performance than MoNA, while the clustering metric leads to much worse results. We hypothesize the reason is that the DB index becomes less informative when the feature dimension is high.

The ablations on different training strategies can be found in table 2. Comparing two-stage MoNA against one-stage finetuning method, we find that the knowledge alignment stage brings significant improvement.

MoNA reduces modality knowledge misalignment. We empirically validate that MoNA reduces the discrepancy between source modality and four target tasks (Fig. 4), and by which better retains source knowledge during finetuning than other methods (Fig. 3). Our empirical results align with our hypothesis that reducing knowledge misalignment between modalities leads to more effective transfer and higher performance on target tasks.

Parameter sensitivity of λ. This parameter balances the trade-off between adapting to target knowledge and preserving source knowledge. Fig. 6(a)(b) shows results on Spherical and NinaPro tasks with varying λ. We observe that optimal value lies in the range of $[0.3, 0.5]$, whereas much higher or lower values lead to performance drop.

Inner-loop update steps. We take different steps of gradi-

ent descent in the inner-loop and show the results in Fig. 6(c). We observe that increasing the inner-loop steps slightly improves the performance. However, it also causes an increasing accuracy variation as well as computational cost. Therefore, following standard gradient-based meta-learning algorithms like MAML (Finn et al., 2017), we adopt the single step update. This helps to reduce the optimization time as well as the requirements for large memory.

Training Efficiency. The double loop optimization paradigm of meta-learning has its intrinsic limitation. However, MoNA mitigates this issue in practice and achieves comparable efficiency with ORCA, as shown in table 7. The reason is two fold: first, we adopt single step update as discussed above to reduce the inner-loop optimization time. Second, MoNA requires only 5 epochs of target embedder training to achieve satisfactory performance, which is significantly lesser than ORCA, which requires more than 60 epochs in the first stage. Therefore, although MoNA cost longer time in training one epoch, the overall training time is comparable between the two methods.

5. Related Work

5.1. Cross-domain Transfer

Transfer learning is extensively investigated in the topic of domain shift (Pan & Yang, 2010). Domain Adaptation (DA) (Ben-David et al., 2010; Long et al., 2018; Ganin & Lempitsky, 2015; Li et al., 2021) studies the knowledge transfer between source and target domain that differ in data distribution yet share the same task. finetuning (Yosinski et al., 2014; Xuhong et al., 2018; You et al., 2020; Kumar et al., 2022; Ma et al., 2023) is another powerful pipeline that enables knowledge transfer from large source dataset to different downstream tasks. It initializes the target model for downstream tasks using weights pretrained on massive source data, achieving higher performances than training the model from scratch. Our work belongs to the second transfer pipeline, yet addressing the more challenging situation where knowledge required for solving target tasks may not be quite aligned with source.

5.2. Cross-modality Transfer

Cross-modal transfer extends the boundary of transfer learning from same modality to different modalities (Shen et al., 2023). The possibility of leveraging model pretrained on one modality to benefit tasks on other modalities gains increasing attention in recent years (Dinh et al., 2022; Lu et al., 2022; Reid et al., 2022; Mirchandani et al., 2023; Pang et al., 2023). One stream of works focuses on the knowledge reuse and transfer from source pretrained models. FPT (Lu et al., 2022) empirically shows that a pretrained language model (PLM) can benefit a variety of non-language downstream tasks, whereas other works transfer PLM to specific modalities (Zhou et al., 2023; Pang et al., 2023;

Reid et al., 2022). ORCA proposes a general workflow for cross-modal transfer that first aligns the data distribution between source and target embeddings and then finetunes the source model to adapt to target modality. Another line of researches aims to learn a general model for several modalities. Meta-transformer (Zhang et al., 2023b) proposes to design modality-specific embedders while keeping a multimodal pretrained model like CLIP (Radford et al., 2021) frozen as the unified backbone. OneLLM (Han et al., 2023) further adds projection layers on top of the pretrained encoder, interfacing it to large language models.

Despite the empirical success of previous works, we still lack understanding of the reason behind these success. As the target modalities can be greatly diverse, the common assumption that knowledge from source modality is beneficial to all target modalities requires further examination. In this work, we formalize the knowledge discrepancy between modalities as an interpretation to the actual feasibility of knowledge transfer between certain source to target modality, and propose a novel method to reduce modality knowledge misalignment.

5.3. Meta-Learning

Our work leverages the optimization-based meta-learning approaches that use bi-level optimization to embed learning procedures like gradient descent into the meta-optimization problem (Hospedales et al., 2021; Rajeswaran et al., 2019). MAML (Finn et al., 2017) is a representative work alone this line of meta-learning. It learns parameters that can serve as a general initialization to solve new downstream tasks with high efficiency. The method mimics the learning process on new tasks in the inner-loop, and updates the outer-loop parameters according to the final performances of each new task within the inner-loop. The spirit of such idea also inspires researches in domain generalization (Li et al., 2018; Balaji et al., 2018; Dou et al., 2019; Li et al., 2019; Liu et al., 2020), where the inner-loop simulates the model's generalization in unseen target domains.

6. Conclusion and Discussion

In this work, we empirically reveal the connection between the modality knowledge discrepancy and the effectiveness of cross-modal transfer. We provide interpretation of such discrepancy in terms of the divergence between conditional distributions. We further propose MoNA, a meta-learning based method to align source and target modality knowledge and improve from existing cross-modal transfer methods. Extensive experiments on two benchmarks with various modalities validate our approach.

Based on our formulation of modality knowledge discrepancy, future work may involve evaluating different source modalities and pretrained models to find the most transferable source model to the given target task.

Acknowledgement

Impact Statement

The investigation of cross-modal transfer learning in this paper reveals the connection between modality knowledge discrepancy and cross-modality transfer effectiveness, which provides new insights to the applications when finetuning vision or language pretrained models to a variety of tasks such as PDE solving, cardiac disease prediction and hand gesture recognition. Additionally, the proposed meta-learning-based modality knowledge alignment method has the potential to enhance cross-modal transfer performance in diverse fields mentioned above and improve the real-world utility of deep neural networks.

References

Adhikari, B. A fully open-source framework for deep learning protein real-valued distances. *Scientific reports*, 10 (1):13374, 2020.

Afouras, T., Chung, J. S., and Zisserman, A. Asr is all you need: Cross-modal distillation for lip reading. In *ICASSP*, pp. 2143–2147, 2020.

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

Alvarez-Melis, D. and Fusi, N. Geometric dataset distances via optimal transport. In *NeurIPS*, pp. 21428–21439, 2020.

Atzori, M., Gijsberts, A., Heynen, S., Hager, A.-G. M., Deriaz, O., Van Der Smagt, P., Castellini, C., Caputo, B., and Müller, H. Building the ninapro database: A resource for the biorobotics community. In *BioRob*, pp. 1258–1265, 2012.

Aytar, Y., Vondrick, C., and Torralba, A. Soundnet: Learning sound representations from unlabeled video. In *NeurIPS*, 2016.

Bai, Y., Geng, X., Mangalam, K., Bar, A., Yuille, A., Darrell, T., Malik, J., and Efros, A. A. Sequential modeling enables scalable learning for large vision models. *CoRR, abs/2312.00785*, 2023.

Balaji, Y., Sankaranarayanan, S., and Chellappa, R. Metareg: Towards domain generalization using meta-regularization. In *NeurIPS*, 2018.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. A theory of learning from different domains. *Machine learning*, 79(1):151–175, 2010.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. On the opportunities and risks of foundation models. *CoRR, abs/2108.07258*, 2021.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. In *NeurIPS*, pp. 1877–1901, 2020.

Clifford, G. D., Liu, C., Moody, B., Li-wei, H. L., Silva, I., Li, Q., Johnson, A., and Mark, R. G. Af classification from a short single lead ecg recording: The physionet/computing in cardiology challenge 2017. In *CinC*, pp. 1–4, 2017.

Cohen, T. S., Geiger, M., Köhler, J., and Welling, M. Spherical cnns. In *ICLR*, 2018.

Dai, R., Das, S., and Bremond, F. Learning an augmented rgb representation with cross-modal knowledge distillation for action detection. In *ICCV*, pp. 13053–13064, 2021.

Ding, N., Qin, Y., Yang, G., Wei, F., Yang, Z., Su, Y., Hu, S., Chen, Y., Chan, C.-M., Chen, W., et al. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, 5(3):220–235, 2023.

Dinh, T., Zeng, Y., Zhang, R., Lin, Z., Gira, M., Rajput, S., Sohn, J.-y., Papailiopoulos, D., and Lee, K. Lift: Language-interfaced fine-tuning for non-language machine learning tasks. In *NeurIPS*, pp. 11763–11784, 2022.

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

Dou, Q., Coelho de Castro, D., Kamnitsas, K., and Glocker, B. Domain generalization via model-agnostic learning of semantic features. In *NeurIPS*, 2019.

Finn, C., Abbeel, P., and Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, pp. 1126–1135, 2017.

Fonseca, E., Pons Puig, J., Favory, X., Font Corbera, F., Bogdanov, D., Ferraro, A., Oramas, S., Porter, A., and Serra, X. Freesound datasets: a platform for the creation of open audio datasets. In *ISMIR*, pp. 486–493, 2017a.

Fonseca, E., Pons Puig, J., Favory, X., Font Corbera, F., Bogdanov, D., Ferraro, A., Oramas, S., Porter, A., and Serra, X. Freesound datasets: a platform for the creation of open audio datasets. In *ISMIR*, 2017b.

Ganin, Y. and Lempitsky, V. Unsupervised domain adaptation by backpropagation. In *ICML*, pp. 1180–1189, 2015.

Garcia, N. C., Morerio, P., and Murino, V. Modality distillation with multiple stream networks for action recognition. In *ECCV*, pp. 103–118, 2018.

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

Girshick, R., Donahue, J., Darrell, T., and Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, pp. 580–587, 2014.

Gupta, S., Hoffman, J., and Malik, J. Cross modal distillation for supervision transfer. In *CVPR*, pp. 2827–2836, 2016.

Han, J., Gong, K., Zhang, Y., Wang, J., Zhang, K., Lin, D., Qiao, Y., Gao, P., and Yue, X. Onellm: One framework to align all modalities with language. *CoRR, abs/2312.03700*, 2023.

Hospedales, T., Antoniou, A., Micaelli, P., and Storkey, A. Meta-learning in neural networks: A survey. *TPAMI*, 44 (9):5149–5169, 2021.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models. In *ICLR*, 2022.

Jaegle, A., Borgeaud, S., Alayrac, J.-B., Doersch, C., Ionescu, C., Ding, D., Koppula, S., Zoran, D., Brock, A., Shelhamer, E., et al. Perceiver io: A general architecture for structured inputs & outputs. In *ICLR*, 2022.

Kenton, J. D. M.-W. C. and Toutanova, L. K. Bert: Pretraining of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pp. 4171–4186, 2019.

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., and Krishnan, D. Supervised contrastive learning. In *NeurIPS*, 2020.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Kumar, A., Raghunathan, A., Jones, R., Ma, T., and Liang, P. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *ICLR*, 2022.

Li, D., Yang, Y., Song, Y.-Z., and Hospedales, T. Learning to generalize: Meta-learning for domain generalization. In *AAAI*, 2018.

Li, S., Xie, B., Lin, Q., Liu, C. H., Huang, G., and Wang, G. Generalized domain conditioned adaptation network. *TPAMI*, 2021.

Li, Y., Yang, Y., Zhou, W., and Hospedales, T. Feature-critic networks for heterogeneous domain generalization. In *ICML*, pp. 3915–3924, 2019.

Li, Z., Kovachki, N., Azizzadenesheli, K., Liu, B., Bhattacharya, K., Stuart, A., and Anandkumar, A. Fourier neural operator for parametric partial differential equations. In *ICLR*, 2020.

Lin, Y.-B., Sung, Y.-L., Lei, J., Bansal, M., and Bertasius, G. Vision transformers are parameter-efficient audio-visual learners. In *CVPR*, pp. 2299–2309, 2023.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. In *NeurIPS*, 2023.

Liu, Q., Dou, Q., and Heng, P.-A. Shape-aware meta-learning for generalizing prostate mri segmentation to unseen domains. In *MICCAI*, pp. 475–485, 2020.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *CoRR, abs/1907.11692*, 2019.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pp. 11012–11022, 2021.

Long, M., Cao, Y., Cao, Z., Wang, J., and Jordan, M. I. Transferable representation learning with deep adaptation networks. *TPAMI*, 41(12):3071–3085, 2018.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. *CoRR, abs/1711.05101*, 2017.

Lu, K., Grover, A., Abbeel, P., and Mordatch, I. Frozen pretrained transformers as universal computation engines. In *AAAI*, pp. 7628–7636, 2022.

Ma, W., Li, S., Zhang, J., Liu, C. H., Kang, J., Wang, Y., and Huang, G. Borrowing knowledge from pre-trained language model: A new data-efficient visual learning paradigm. In *ICCV*, pp. 18786–18797, 2023.

Mirchandani, S., Xia, F., Florence, P., Ichter, B., Driess, D., Arenas, M. G., Rao, K., Sadigh, D., and Zeng, A. Large language models as general pattern machines. *CoRR, abs/2307.04721*, 2023.

Pan, S. J. and Yang, Q. A survey on transfer learning. *TKDE*, 22(10):1345–1359, 2010.

Pang, Z., Xie, Z., Man, Y., and Wang, Y.-X. Frozen transformers in language models are effective visual encoder layers. *CoRR, abs/2310.12973*, 2023.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, pp. 8024–8035, 2019.

Petitjean, F., Inglada, J., and Gançarski, P. Satellite image time series analysis under time warping. *TGRS*, 50(8): 3081–3095, 2012.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *ICML*, pp. 8748–8763, 2021.

Raissi, M., Perdikaris, P., and Karniadakis, G. E. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational physics*, 378:686–707, 2019.

Rajeswaran, A., Finn, C., Kakade, S. M., and Levine, S. Meta-learning with implicit gradients. In *NeurIPS*, 2019.

Reid, M., Yamada, Y., and Gu, S. S. Can wikipedia help offline reinforcement learning? *CoRR, abs/2201.12122*, 2022.

Ren, S., Du, Y., Lv, J., Han, G., and He, S. Learning from the master: Distilling cross-modal advanced knowledge for lip reading. In *CVPR*, pp. 13325–13333, 2021.

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pp. 234–241, 2015.

Sarkar, P. and Etemad, A. Xkd: Cross-modal knowledge distillation with domain alignment for video representation learning. In *AAAI*, pp. 14875–14885, 2024.

Shen, J., Khodak, M., and Talwalkar, A. Efficient architecture search for diverse tasks. In *NeurIPS*, pp. 16151–16164, 2022.

Shen, J., Li, L., Dery, L. M., Staten, C., Khodak, M., Neubig, G., and Talwalkar, A. Cross-modal fine-tuning: Align then refine. In *ICML*, 2023.

Soomro, K., Zamir, A. R., and Shah, M. Ucf101: A dataset of 101 human actions classes from videos in the wild. *CoRR, abs/1212.0402*, 2012.

Sugiyama, M., Krauledat, M., and Müller, K.-R. Covariate shift adaptation by importance weighted cross validation. *JMLR*, 8(5), 2007.

Takamoto, M., Praditia, T., Leiteritz, R., MacKinlay, D., Alesiani, F., Pflüger, D., and Niepert, M. Pdebench: An extensive benchmark for scientific machine learning. In *NeurIPS*, pp. 1596–1611, 2022.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *CoRR, abs/2307.09288*, 2023.

Tu, R., Roberts, N., Khodak, M., Shen, J., Sala, F., and Talwalkar, A. Nas-bench-360: Benchmarking neural architecture search on diverse tasks. In *NeurIPS*, pp. 12380–12394, 2022.

Wang, T. and Isola, P. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *ICML*, pp. 9929–9939, 2020.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. Huggingface's transformers: State-of-the-art natural language processing. *CoRR, abs/1910.03771*, 2019.

Xu, X., Tao, C., Shen, T., Xu, C., Xu, H., Long, G., and Lou, J.-g. Re-reading improves reasoning in language models. *CoRR, abs/2309.06275*, 2023.

Xue, Z., Gao, Z., Ren, S., and Zhao, H. The modality focusing hypothesis: Towards understanding crossmodal knowledge distillation. In *ICLR*, 2023.

Xuhong, L., Grandvalet, Y., and Davoine, F. Explicit inductive bias for transfer learning with convolutional networks. In *ICML*, pp. 2825–2834, 2018.

Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. How transferable are features in deep neural networks? In *NeurIPS*, pp. 3320–3328, 2014.

You, K., Kou, Z., Long, M., and Wang, J. Co-tuning for transfer learning. In *NeurIPS*, pp. 17236–17246, 2020.

Zhang, K. and Bloom, J. S. deepcr: Cosmic ray rejection with deep learning. *The Astrophysical Journal*, 889(1): 24, 2020.

Zhang, S., Dong, L., Li, X., Zhang, S., Sun, X., Wang, S., Li, J., Hu, R., Zhang, T., Wu, F., et al. Instruction tuning for large language models: A survey. *CoRR, abs/2308.10792*, 2023a.

Zhang, Y., Gong, K., Zhang, K., Li, H., Qiao, Y., Ouyang, W., and Yue, X. Meta-transformer: A unified framework for multimodal learning. *CoRR, abs/2307.10802*, 2023b.

Zhou, J. and Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning–based sequence model. *Nature methods*, 12(10):931–934, 2015.

Zhou, Q.-L., Ye, H.-J., Wang, L.-Y., and Zhan, D.-C. Unlocking the transferability of tokens in deep models for tabular data. *CoRR, abs/2310.15149*, 2023.

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., and He, Q. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020.

A. Appendix

A.1. Benchmark Introduction and Training Details

We summarize the details of ten tasks in NAS-Bench-360 in table 8. The ten tasks can be divided into three groups: 2D point prediction (classification), 2D dense prediction and 1D classification.

CIFAR-100 (Krizhevsky et al., 2009) in a standard classification task on natural images. Spherical (Cohen et al., 2018) classifies spherical projections of the CIFAR-100 images which simulate distorted image signals. NinaPro (Atzori et al., 2012) moves away from the image modality to classify hand gestures indicated by electromyography signals. FSD50K (Fonseca et al., 2017b) is an audio classification task originated from the larger Freesound dataset (Fonseca et al., 2017a) with spectrogram as input and multiple labels as output.

Darcy-Flow (Li et al., 2020) is a regression task for learning a map from the initial conditions of Partial Differential Equations (PDEs) to the solution at a later time-step. PSICOV (Adhikari, 2020) predicts the inter-residual distance of a small set of protein structures. Cosmic (Zhang & Bloom, 2020) is the last dense prediction task in benchmark, aiming to identify comsic ray contamination in the images collected from the Hubble Space Telescope.

ECG (Clifford et al., 2017) is the classification task on electrocardiogram signals that is frequently used in heart disease diagnosis. Satellite (Petitjean et al., 2012) is the classification of land cover type giving the satellite image time series as inputs. DeepSEA (Zhou & Troyanskaya, 2015) predicts the functional effects from genetic sequences and makes prediction among 36 categories of chromatin protein behavior. Please find more detailed description of these tasks in the original paper (Tu et al., 2022).

The training configurations for vanilla finetuning (MoNA's second stage) for each task basically follow the setups in ORCA (Shen et al., 2023) and is summarized in table 9. For the first stage, we uniformly adopt AdamW (Loshchilov & Hutter, 2017) with learning rate 3e-5 and weight decay 0.1 since we find it works reasonably well on all tasks. We warmup the embedder with ten epochs before moving to the second stage.

Table 8. Introduction of the ten tasks in NAS-Bench-360.

	CIFAR100	Spherical	NinaPro	FSD50K	DarcyFlow	PSICOV	Cosmic	ECG	Satellite	DeepSEA
# training data	60K	60K	3956	51K	1.1K	3606	5250	330K	1M	250K
Input shape	2D	2D	2D	2D	2D	2D	2D	1D	1D	1D
Output type	Point	Point	Point	Point	Dense	Dense	Dense	Point	Point	Point
# classes	100	100	18	200	–	–	–	4	24	36
Loss	CE	CE	LpLoss	MSELoss	BCE	FocalLoss	BCE	CE	CE	BCE
Expert network	DenseNet-BC	S2CN	Attention Model	VGG	FNO	DEEPCON	deepCR-mask	ResNet-1D	ROCKET	DeepSEA

Table 9. Configuration of hyper-parameters and optimizers used in NAS-Bench-360.

	CIFAR100	Spherical	NinaPro	FSD50K	Darcy Flow	PSICOV	Cosmic	ECG	Satellite	DeepSEA
Batch Size	32	32	32	32	4	1	4	4	16	16
Epoch	60	60	60	100	100	10	60	15	60	13
Grad. Accum.	32	4	1	1	1	32	1	16	4	1
Optimizer	SGD	AdamW	Adam	Adam	AdamW	Adam	AdamW	SGD	AdamW	Adam
Learning Rate	1.00E-04	1.00E-04	1.00E-04	1.00E-04	1.00E-03	5.00E-06	1.00E-03	1.00E-06	3.00E-05	1.00E-05
Weight Decay	1.00E-03	1.00E-01	1.00E-05	5.00E-05	5.00E-03	1.00E-05	0.00E+00	1.00E-01	3.00E-06	0.00E+00

On PDEBench, we evaluate all tasks except 2D and 3D Navier-Stokes Equations which are too computational expensive. We provide simple introduction to each PDE and refer more details to the original paper (Takamoto et al., 2022).

1D Advection equation models pure advection behavior without non-linearity, with parameter β informing the constant advection speed. 1D Burgers' equation models the non-linear behavior and diffusion process in fluid dynamics. The parameter ν is the diffusion coefficient which is assumed constant. 1D Diffusion-Reaction equation combines a diffusion process and a rapid evolution from a source term, where two parameters ν, ρ control the degree of combination. 1D

Diffusion-Sorption equation models a diffusion process which is retarded by a sorption process. The equation is applicable to real world scenarios. 1D compressible Navier-Stokes equation describes the dynamics of compressible fluid, where η and ζ are the shear and bulk viscosity, respectively.

2D Darcy-Flow equation describes a steady-state solution of the flow dynamics over the unit square. The force term $f(x)$ is simplified as the constant β and it changes the scale of the solution. 2D shallow-water equations are derived from the general Navier-Stokes equation that presents a suitable framework for modelling free-surface flow problems. 2D Diffusion-Reaction equation extends the 1D equation by considering two non-linearly coupled variables.This task serves as a challenging problem since the coupling is non-linear and its real world application is huge.

Table 10. Introduction of eight tasks evaluated on PDEBench.

	Advection	Burgers	Diffusion-Reaction	Diffusion-Sorption	Navier-Stokes	Darcy-Flow	Shallow-Water	Diffusion-Reaction
Input shape	1D	1D	1D	1D	1D	2D	2D	2D
Output type				Dense				
Resolution	1024	1024	1024	1024	1024	128 * 128	128 * 128	128 * 128
Parameters	$\beta = 0.4$	$\nu = 1.0$	$\nu = 0.5, \rho = 1.0$	–	$\eta = 1.0, \zeta = 1.0$	$\beta = 0.1$	–	–
Loss				Normalized Root Mean Squared Errors (nRMSEs)				

Table 11. Configuration of hyper-parameters and optimizers used in NAS-Bench-360

	Advection	Burgers	Diffusion-Reaction	Diffusion-Sorption	Navier-Stokes	Darcy-Flow	Shallow-Water	Diffusion-Reaction
Batch Size	4	4	4	4	4	4	4	4
Epoch	200	200	200	200	200	100	200	200
Grad. Accum.	1	1	1	1	1	1	1	1
Optimizer	Adam	Adam	SGD	AdamW	AdamW	AdamW	AdamW	Adam
Learning Rate	1.00E-04	1.00E-05	1.00E-03	1.00E-04	1.00E-04	1.00E-04	1.00E-04	1.00E-04
Weight Decay	1.00E-05	1.00E-05	1.00E-05	0	1.00E-03	1.00E-05	0	1.00E-03

A.2. Detailed Explanation of the Model Architecture

In this section we provide a detailed explanation of the modality-specific embedders and predictors. Note that our implementation of these modules follows exactly the designs in ORCA.

The structure of the modality-specific embedder depends on whether the task is 2D or 1D.

- For 2D tasks, the embedder consists of a linear projection layer and a LayerNorm operation. For any input data with size $C \times H \times W$, where C,H and W are channels, height and width, we first resize it to $C \times 224^2$ and divide it into N patches of size $C \times 4^2$. Then the linear projection layer maps each patch into a token of size 128 and the LayerNorm operation is applied on all the projected patches. Therefore, the embedder can be formulated as a function $e_{2D} : \mathbb{R}^{N \times 16C} \to \mathbb{R}^{N \times 128}$.

- For 1D tasks, the embedder consists of a linear projection layer, a LayerNorm operation and learnable positional embeddings. For any input data with size $C \times L$ where C and L are channels and sequence length respectively, we first divide it into N patches of size $C \times \frac{L}{N}$. Then the linear projection layer maps each patch into a token of size 768 and the LayerNorm operation is applied on all the projector patches. Finally, the positional embeddings are added to the patches. Therefore, the embedder can be formulated as a function $e_{1D} : \mathbb{R}^{CL} \to \mathbb{R}^{768N}$.

The structure of the modality-specific predictor depends on whether the task is classification or dense prediction.

- For classification tasks, the predictor consists of an average pooling layer and a linear projection layer. The average pooling layer averages the dense feature map of size $N' \times d$ to produce feature of size d, then the linear projection layer maps the feature to logits of size K, where d and K represent feature dimension and the number of categories. Therefore, the predictor can be formulated as a function $h_c : \mathbb{R}^{N'd} \to \mathbb{R}^K$.

- For dense prediction tasks, the predictor consists of a linear projection layer, a pixel rearrangement operation and two adaptive pooling layers. The linear projection layer takes the dense feature map of size $7^2 \times d$ as input and output a new feature of size $7^2 \times 3072$, which is then reorganized into shape $224^2 \times 3$. Next, two pooling operations are applied sequentially, turning the feature size from 3×224^2 to $K \times 224^2$ and finally $K \times H \times W$ which is in accordance with the input spatial dimension. Therefore, the predictor can be formulated as a function $h_d : \mathbb{R}^{49d} \to \mathbb{R}^{KHW}$.

A.3. Theoretical Foundation for the Meta-Learning Objective

We provide a preliminary theoretical analysis of our meta-learning objective. As in our paper, we denote the target embedder parameter as ϕ_e and the pretrained encoder parameter as θ_f. The objective of MoNA is:

$$\min_{\phi_e} \mathcal{L}_{inner}(\theta_f, \phi_e) + \lambda \mathcal{L}_{outer}(\theta_f - \alpha \nabla_{\theta_f} L_{inner}(\theta_f, \phi_e)). \tag{7}$$

We then get the approximation of the objective using Taylor expansion as:

$$\min_{\phi_e} \mathcal{L}_{inner}(\theta_f, \phi_e) + \lambda \mathcal{L}_{outer}(\theta_f) + \lambda \nabla \mathcal{L}_{outer}(\theta_f) \cdot (-\alpha \nabla \mathcal{L}_{inner}(\theta_f, \phi_e)). \tag{8}$$

Since it is the optimization problem of ϕ_e, the second term is neglected and we obtain the final form as:

$$\min_{\phi_e} \mathcal{L}_{inner}(\theta_f, \phi_e) - \lambda \alpha (\nabla \mathcal{L}_{outer}(\theta_f)) \cdot (\nabla \mathcal{L}_{inner}(\theta_f, \phi_e)). \tag{9}$$

We see that the first term directly minimizes the target task loss, whereas the second term maximizes the dot product of the source loss gradient and the target loss gradient, both with respect to θ_f. In other words, the second term updates the target embedder in a way that makes the gradient direction of the target task loss align with the gradient direction of the source modality loss. In the optimal situation where the two gradients point to the same direction, finetuning the encoder θ_f using target loss will maximally maintain the source knowledge. With this theoretical support, we argue the optimality of the target embedder update objective in terms of achieving cross-modal knowledge alignment.

A.4. Approximation Algorithm for Computing Modality Knowledge Discrepancy

We explain the algorithm used to approximate the modality knowledge discrepancy in definition 2.2. The results is shown in Figure 4. For each target sample $(\boldsymbol{x}_i^t, y_i^t)$, we use the one-hot embedding of its label as the target conditional distribution $p(y^t|\boldsymbol{x}_i^t)$. For the source conditional distribution, we leverage the complete source pretrained model (with the original classifier trained on ImageNet) to compute the logits $\boldsymbol{z}_i^s$. We then simplify the searching for optimal subset $\mathcal{B}$ in the definition as a random category selection. Number of the selected source classes are equal to the number of the target categories being compared. With the subset selected, we consider the maximum logit value within the subset and assign the source category as $y_{i,\mathcal{B}}^s = \arg\max_{k \in \mathcal{B}} [\boldsymbol{z}_i^s]_k$, and we also use the one-hot embedding of source predicted category to model the source conditional distribution $p(y_{\mathcal{B}}^s|\boldsymbol{x}_i^t)$. To this end, the discrepancy between two conditional distribution can be simplified as

$$d(p(y_{\mathcal{B}}^s|\boldsymbol{x}_i^t), p(y^t|\boldsymbol{x}_i^t)) = \mathbb{I}_{[y_{i,\mathcal{B}}^s \neq y_i^t]}, \tag{10}$$

Therefore, we can compute the modality knowledge discrepancy as

$$D(\mathcal{M}^s, \mathcal{M}^t) = \frac{1}{n_t} \sum_{i=1}^{n_t} d(p(y_{\mathcal{B}}^s|\boldsymbol{x}_i^t), p(y^t|\boldsymbol{x}_i^t)). \tag{11}$$

Still, we need to find the optimal permutation π that matches source and target categories one-to-one. Since it is too computational expensive to iteration through all the permutations, we opt to randomly permute the target label indexes. Therefore in practice, we conduct random experiment for 100,000 times. Each time we randomly select subset of the source and randomly permute the target label indexes. We compute the modality knowledge discrepancy using Eq. (11), and the final discrepancy is the minimum value during the whole process. Alg. 2 summarizes the complete process.

A.5. Cross-Modality Transfer Against Cross-Modality Knowledge Distillation

Cross-modal Knowledge Distillation (Xue et al., 2023) is an alternative paradigm for cross-modality knowledge sharing, and is proven to be useful on diverse applications including video representation learning (Sarkar & Etemad, 2024), action

---

Algorithm 2 Approximation Algorithm for Modality Knowledge Discrepancy

---

1: **Input:** Source pretrained model $g_{\boldsymbol{\theta}_0^S}$; Target data $\{\boldsymbol{x}_i^t, y_i^t\}$; Target task category number K; Maximum experiments I.
2: **Output:** Modality Knowledge Discrepancy $D(\mathcal{M}^s, \mathcal{M}^t)$.
3: Compute the source logit for each target sample: $\boldsymbol{z}_i^s = g_{\boldsymbol{\theta}_0^S}(\boldsymbol{x}_i^t)$.
4: $min_D \leftarrow 1$.
5: **for** $exp = 1, 2, \cdots, I$ **do**
6:     Randomly select a subset $\mathcal{B}$ of class index from source class indexes with the size equals to target class number, $|\mathcal{B}| = K$.
7:     $y_{i,\mathcal{B}}^s \leftarrow \arg\max_{k \in \mathcal{B}}[\boldsymbol{z}_i^s]_k$
8:     Randomly shuffle the index of target labels: $y^t \leftarrow \pi(y^t)$, (e.g., $\pi(1) = 2, \pi(2) = 1$.)
9:     Compute $D_{exp}(\mathcal{M}^s, \mathcal{M}^t)$ using Eq. (11).
10:    $min_D \leftarrow \min(D_{exp}, min_D)$.
11: **end for**
12: Return min_D.

---

recognition (Garcia et al., 2018; Dai et al., 2021), lip reading (Ren et al., 2021; Afouras et al., 2020), depth (Gupta et al., 2016), sound (Aytar et al., 2016) and etc. Specifically, XKD (Sarkar & Etemad, 2024) explores leveraging Maximum Mean Discrepancy to align video and image modality, ASR (Afouras et al., 2020) proposes a novel Connectionist Temporal Classification loss that enables learning sequence-to-sequence tasks without the need for explicit alignment of training targets to input frames, and Augmented RGB (Dai et al., 2021) also investigates the sequence-to-sequence knowledge distillation framework using a contrastive strategy.

A common thread among these methodologies is their reliance on paired data or multimodal representations of identical data points. This prerequisite, however, may not always be feasible or accessible in certain modalities, such as those involving Partial Differential Equations (PDEs) or protein structure prediction, thereby limiting their applicability.

In contrast, cross-modal transfer learning emerges as a more versatile and inclusive framework for knowledge sharing across modalities, primarily because it eschews the need for paired data, thereby casting a wider net in terms of application potential. However, this flexibility comes at the cost of an increased risk of ineffective transfer, particularly when faced with substantial modality knowledge discrepancies and in the absence of paired data to serve as a bridge between the disparate modalities.

Therefore, one of the primal purposes of this work is to describe the extent of modality discrepancy systematically. We hope that our effort can motivate research in modality discrepancy, which would eventually provide guidelines for better cross-modal knowledge transfer.

A.6. Limitation

Here we discuss a few limitations of our work as well as potential solution towards these limitations.

Firstly, our experimental framework adheres to the protocols established by ORCA, which involves utilizing CIFAR10 and CoNLL-2003 as surrogate source datasets for the vision and language modalities, respectively. This approach, while facilitating a direct comparison with established benchmarks, introduces a limitation in that the choice of surrogate datasets might influence the outcomes of cross-modal transfer learning. The potential variability in transfer performance attributed to different source datasets is a factor that our current analysis does not account for, presenting a critical area for future exploration.

Secondly, in the implementation of MoNA, we opted for a simplistic yet effective strategy akin to Model-Agnostic Meta-Learning (MAML), using a single-step update within the inner loop to balance the performance and computational cost. Noticed that with recent advancements in gradient-based meta-learning proposing improved algorithms for inner-loop optimization, we are committed to actively exploring these methods to improve the algorithm of MoNA.