

H-ensemble: an Information Theoretic Approach to Reliable Few-Shot Multi-Source-Free Transfer

Yanru Wu¹, Jianning Wang², Weida Wang¹, Yang Li^{1*}

¹Tsinghua Shenzhen International Graduation School, Tsinghua University[†]

²School of Computer Science and Engineering, Harbin Institute of Technology, Shenzhen
{wu-yr21, wangwd19}@mails.tsinghua.edu.cn, wangjianning@stu.hit.edu.cn, yangli@sz.tsinghua.edu.cn

Abstract

Multi-source transfer learning is an effective solution to data scarcity by utilizing multiple source tasks for the learning of the target task. However, access to source data and model details is limited in the era of commercial models, giving rise to the setting of *multi-source-free* (MSF) transfer learning that aims to leverage source domain knowledge without such access. As a newly defined problem paradigm, MSF transfer learning remains largely underexplored and not clearly formulated. In this work, we adopt an information theoretic perspective on it and propose a framework named *H-ensemble*, which dynamically learns the optimal linear combination, or ensemble, of source models for the target task, using a generalization of maximal correlation regression. The ensemble weights are optimized by maximizing an information theoretic metric for transferability. Compared to previous works, H-ensemble is characterized by: 1) its adaptability to a novel and realistic MSF setting for few-shot target tasks, 2) theoretical reliability, 3) a lightweight structure easy to interpret and adapt. Our method is empirically validated by ablation studies, along with extensive comparative analysis with other task ensemble and transfer learning methods. We show that the H-ensemble can successfully learn the optimal task ensemble, as well as outperform prior arts.

Introduction

The scarcity of annotated data is a principal challenge to machine learning algorithms in a variety of real-world scenarios, where transfer learning has emerged as an effective remedy, harnessing information and insights from well-established source tasks to augment data-deficient domains (see ???). Historically, transfer learning has centered on a singular source-task and target-task relationship. However, the spotlight is gradually shifting towards multi-source transfer learning, capitalizing on multiple source tasks to facilitate the training of the target task (see ?). A variety of works have contributed to both the theoretical foundation and practical algorithms on this topic. Yet, the trend of decentralization and growing privacy concerns have impeded access to source data, particularly when compared to the

models they’ve nurtured (?). This evolving landscape underscores the need for the novel paradigm of *multi-source-free* (MSF) transfer learning (?).

MSF transfer learning, sometimes referred to as MSF domain adaptation (?), entails transferring knowledge from multiple sources to a target task without accessing source data, as depicted in Fig. 1. Despite not being clearly defined in most previous papers and still largely underexplored, MSF transfer has been explored by several works these years. Preliminary efforts centered around simple strategies like averaging source models or reducing the problem to single source-free transfers, by source selection using empirical validations or transferability metrics (e.g. ?). Contemporary research has pivoted towards sophisticated model ensembles, assigning weights to sources based on perceived importance. For instance, approaches range from correlating source relevance to the target (?), to conditional entropy minimization (?), and even leveraging source-similarity and class-relationship perception modules for weight determination (??). Beyond these linear ensemble techniques, some researchers also introduced auxiliary models to endow cross-domain capabilities (?).

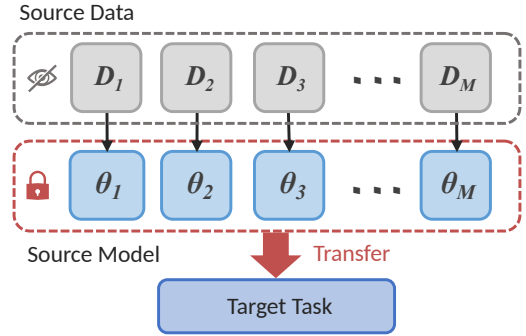


Figure 1: **Problem Setting of Multi-Source-Free Transfer Learning.** The transfer to the target task is based on the source models (whose access in our redefined MSF setting is also restricted) trained on the inaccessible source data.

*Corresponding author.

[†]Yanru Wu, Weida Wang and Yang Li are from the Shenzhen Key Laboratory of Ubiquitous Data Enabling, SIGS, Tsinghua. Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Motivation

While these methods have their merits, they still present three distinct limitations. Firstly, they mainly evaluate the

source importance as an independent characteristic of each source instead of viewing them as a whole, neglecting the likely interaction effects when combining. Secondly, these works focus on the unsupervised MSF transfer, while a more common setting of data scarcity in practice is *few-shot* learning, where we can only access a few of the target data but with labels (?). Thirdly, most of them hinge on accessing intermediate layer outputs or necessitate tuning of source models, which tend to become inaccessible in the era of proprietary large models. Also, despite their effectiveness, these transfer algorithms have quite complicated methodologies, making them hard to interpret or adapt.

As a solution to these issues, we focus on the few-shot MSF transfer challenge, proposing a novel transfer framework tailored to this setting. Given the scarcity of target samples and the inaccessibility of source details, we base our methodology on the optimization of features instead of the analysis of input data as other works did. Specifically, following the idea of linear source ensemble, we define the target feature as a weighted combination of transferred features and introduce an information theoretic metric, *H-score*, to gauge the quality of this synthesis, which can be alternatively interpreted as the transferability of the whole source model ensemble. The optimal source weights will then be simply determined by maximizing this metric, leveraging the robustness of H-score under few-shot settings as well as its computational efficiency and theoretical reliability (?). Our derived target generalizes the maximal correlation regression (MCR) by replacing its feature by an ensemble of source feature extractors and modifying the model accordingly. A detailed explanation of the underlying theoretic basis and intuition of H-ensemble framework will be presented in the Methodology section, showing that: 1) under the maximal correlation framework, the weights maximizing our adapted H-score are theoretically optimal for the target task; 2) the resulting optimal question is of a benign form and can be reliably solved by gradient descend based algorithms.

Contributions

In this paper, we propose an information theoretic MSF transfer framework named *H-ensemble*, where the source models are assembled using a generalization of MCR, and the source weights are obtained by maximizing an H-score based metric. Our contributions mainly lie in three aspects:

- We formulate the few-shot MSF transfer problem mathematically using a source-free setting that is more common in practical scenarios.
- An efficient and theoretically reliable framework derived from maximal correlation regression is introduced for MSF transfer. The framework is designed to be lightweight and easy to interpret and adapt, possible to serve as a *plug-in* module.
- We present both theoretical guarantee and experimental support to the effectiveness of our methodology.

Related Works

Maximal Correlation Regression and H-Score

Maximal correlation regression (MCR) network (?) refers to a specially designed network trained based on the optimization of feature quality by maximizing H-score. Here, H-score is a metric determining whether certain features are informative for the task (more formally introduced by ?)¹. Its derivation is rooted in the maximal correlation interpretation of deep neural networks (?), where given the extracted features, the formula of H-score is actually the optimal network performance under an information theoretic measurement. Previous work also extended H-score to a transferability metric and verified its effectiveness with extensive experiments (??), suggesting the potential of H-score in transfer learning.

Notably, as the basis of both MCR network and H-score, maximal correlation analysis originates from the works of Hirschfeld, Gebelein and Renyi (???) and has been followed and further explored by a broad spectrum of successive work. A detailed overview of maximal correlation analysis can be found given by ? (?).

Multi-Source Transfer Learning

Multi-source transfer learning, a well-established concept, encompasses approaches like naive multi-source transfer by feature extractor concatenation (?) and theoretical explorations (????). Contemporary algorithms can be classified into three main categories: re-weighting approaches that estimate and assign source importance (???); source selection methods using transferability metrics to identify optimal combinations (?); and boosting-based solutions that apply ensemble learning techniques to transfer learning (???). Further insights on this topic are available in related surveys, such as (?).

Problem Definition

Setting Comparison to Previous Works

Based on the change of settings mentioned in the Introduction section, there are mainly three terms to emphasize in our setting: multi-source, source-free and few-shot. Specially, to distinguish from previous notions, source-free is categorized into source-data-free and source-model-free. We compare few-shot MSF transfer with previous settings in these four aspects and summarize the main divergence in Table 1.

Mathematical Formulation

Consider the transfer learning setting with M source tasks $\{S_j\}_{j=1}^M$ and a target task T , we list the main notations in Table 2. For simplification, we assume that the sources and target tasks share the same input space X and all tasks are classification problems.²

¹The H-score introduced by Bao et al. is actually the one-sided form of the score in MCR (both firstly defined by Huang et al. in 2019). They are also differed in their definition of variables and optimal objects.

²This assumption is without loss of generality, for in most cases, model inputs can be mapped to any size by preprocessing (e.g. im-

Setting	Multi-Source	No Source Data	No Source Model Details	Target
Few-Shot transfer (?)	✗	✗	✗	Few-shot
Unsupervised DA (?)	✗	✗	✗	Unsupervised
Multi-Source DA (?)	✓	✗	✗	Unsupervised
Multi-Source-Free DA (?)	✓	✓	✗	Unsupervised
Few-Shot Multi-Source-Free Transfer (Ours)	✓	✓	✓	Few-shot

Table 1: **Comparison across Relative Problem Settings**, involving the aspects of whether there are multi-sources, no source data, no source model details, and different target learning approaches. The abbreviation ‘DA’ represents domain adaptation, ‘✓’ represents obtaining the corresponding aspects, while ‘✗’ the opposite. Among all settings, ours is relatively restrictive.

Notation	Meaning
$S_j; T$	Source/Target Task
$d_{S_j}; d_T$	Source/Target Class Number
$N_{S_j}; N_T$	Source/Target Data Size
$D_{S_j} = \{(x_{S_j}^i, y_{S_j}^i)\}_{i=1}^{N_{S_j}}$	Source Data
$D_T = \{(x_T^i, y_T^i)\}_{i=1}^{N_T}$	Target Data

Table 2: **Notation.** We omit $\{\cdot\}_{j=1}^M$ for simplification.

Notably, in our setting we have no access to source data D_{S_j} and only partial access to target data D_T , so here $N_{S_j} = 0$, $N_T = k \times d_T$ for a k -shot MSF transfer.³ For the source model consisting of a backbone feature extractor $\mathbf{f}_{S_j} : X \rightarrow \mathbb{R}^{d_f}$ and a classification head $\mathbf{g}_{S_j} : \mathbb{R}^{d_f} \rightarrow \mathbb{R}^{d_{S_j}}$ (d_f denotes the feature dimension)⁴, where the source-free restriction leaves the detailed model structure unknown.

In computation, we estimate the distribution P_X^T, P_Y^T of target data x_T, y_T with empirical distribution \hat{P}_X^T, \hat{P}_Y^T :

$$\hat{P}_X^T(x) = \frac{1}{N_T} \sum_{i=1}^{N_T} \mathbb{I}\{x_T^i = x\}; \hat{P}_Y^T(y) = \frac{1}{N_T} \sum_{i=1}^{N_T} \mathbb{I}\{y_T^i = y\}, \quad (1)$$

where $\mathbb{I}\{\cdot\}$ is the indicator function (?). Similarly, the conditional distribution of x_T given y_T $P_{X|Y}^T$ is estimated by its empirical distribution $\hat{P}_{X|Y}^T$:

$$\hat{P}_{X|Y}^T(x|y) = \frac{\sum_{i=1}^{N_T} \mathbb{I}\{x_T^i = x, y_T^i = y\}}{\sum_{i=1}^{N_T} \mathbb{I}\{y_T^i = y\}}. \quad (2)$$

When estimating the underlying distributions with empirical distributions of samples, Eqns. (1) and (2) may be poorly defined and intractable for most target task types due to the high dimension of input x . However, this problem can be

age resizing, pooling, certain embedding layers, etc.), and classification is one of the most representative tasks in machine learning.

³ k samples for each of the d_T classes

⁴ d_f is set to the same across all tasks considering that we could add a fully connected layer to derive any output dimension.

avoided, as in our framework we only need to compute the expectation of features over these distributions.

The goal of few-shot MSF problem is that, by utilizing the partially accessible data and black-box source models only, find the optimal parameters θ_T that minimizes certain divergence \mathcal{L} between the real conditional distribution $P_{Y|X}^T$ and the approximated $P_{Y|X}^{(\theta_T)}$, i.e.,

$$\theta_T^* = \arg \min_{\theta_T} \mathcal{L}(P_{Y|X}^{(\theta_T)}, P_{Y|X}^T). \quad (3)$$

Methodology

Following the analysis in the Motivation section, we model the target feature extractor \mathbf{f}_T by a weighted combination of source feature extractors \mathbf{f}_{S_j} :

$$\mathbf{f}_T = \sum_{j=1}^M \alpha_j \cdot \mathbf{f}_{S_j}, \quad \sum_{j=1}^M \alpha_j = 1. \quad (4)$$

We then need to efficiently evaluate target features to guide the optimization of weights α , where the maximal correlation analysis (MCA) is introduced. In MCA, a neural network is viewed as a feature extractor \mathbf{f} and a classifier \mathbf{g} , and the informativeness of features extracted by \mathbf{f} can be estimated by H-score.

Leveraging the robustness of H-score under few-shot settings as well as its computational efficiency and theoretical reliability, we use H-score as the feature quality measurement and set up the transfer framework under the same theory scheme as a generalization of MCR. Following previous works (?), the divergence \mathcal{L} in Eqn. (3) is defined as variational chi-squared divergence:

$$\mathcal{L}(\theta) = \sum_x P_X(x) \sum_y \frac{[P_{Y|X}^{(\theta)}(y|x) - P_{Y|X}(y|x)]^2}{P_Y(y)}. \quad (5)$$

We illustrate our MSF transfer framework H-ensemble in Fig. 2 and introduce each module in the sections below, showing how the divergence \mathcal{L} is minimized.

Multi-Source-Free Transfer

Overall, the MSF transfer framework follows the generalized MCR(?) network and has two modules, feature extractor $\mathbf{f}_T = \sum_{j=1}^M \alpha_j \cdot \mathbf{f}_{S_j}$ and classifier \mathbf{g}_T derived using

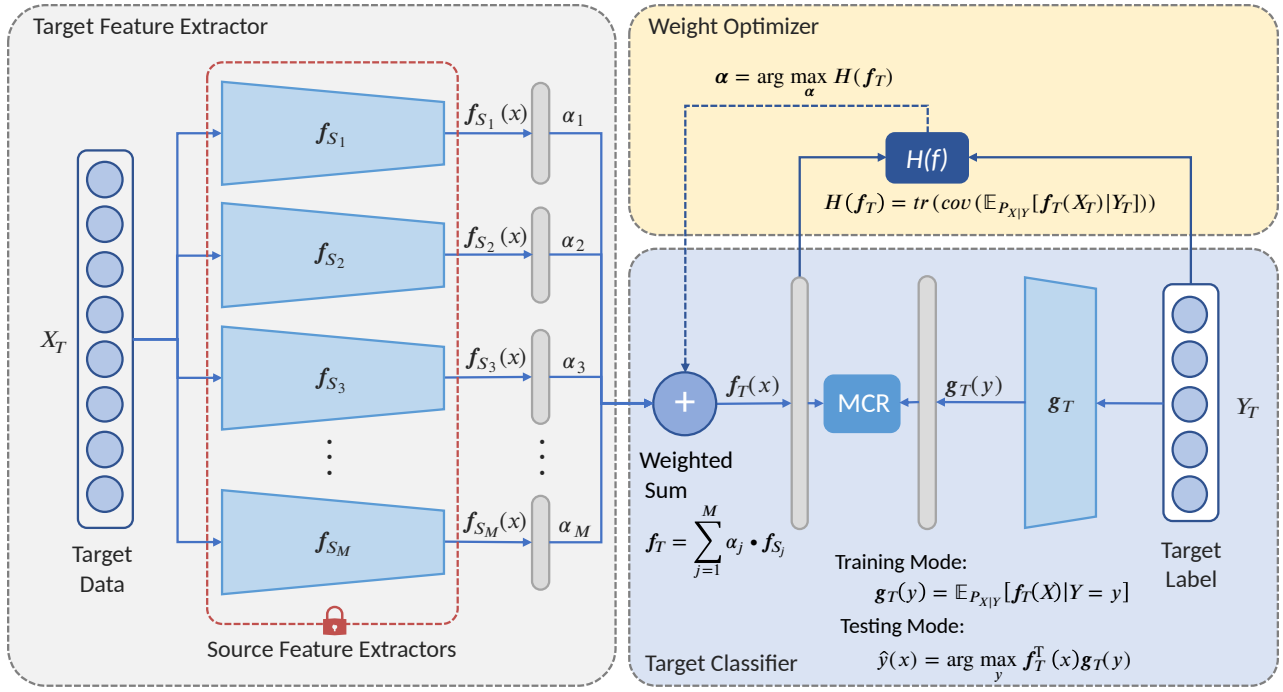


Figure 2: **Framework of H-Ensemble.** The framework consists of three modules. The target data firstly flow into Target Feature Extractor. Then the Weight Optimizer will utilize the outputs of source feature extractors and target label to derive the optimal source weight α , which makes the parameter in deriving target feature. Finally the Target Classifier will be trained and used together with the extractor for test according to a generalization of maximal correlation regression (MCR).

MCR training mode as below. The approximated conditional distribution is defined as

$$P_{Y|X}^{\theta_T}(y|x) = P_Y^T(y)(1 + \mathbf{f}_T^T(x)\mathbf{g}_T(y)). \quad (6)$$

Specifically, given input data $x \sim P_X^T$, the features extracted by source models are $\mathbf{f}_{s_j}(x)$. To simplify the deduction, we assume that $\mathbb{E}_{P_X^T}[\mathbf{f}_{s_j}(x)] = 0$ and $\mathbb{E}_{P_X^T}[\mathbf{f}_{s_j}(x) \cdot \mathbf{f}_{s_j}(x)^T] = \mathbf{I}$ for all j , as a normalization layer could be efficiently added to extractors. Subsequently, \mathbf{f}_T is also a normalized feature with $\mathbb{E}_{P_X^T}[\mathbf{f}_T(x)] = 0$ and $\mathbb{E}_{P_X^T}[\mathbf{f}_T(x) \cdot \mathbf{f}_T(x)^T] = \mathbf{I}$. After deriving the target feature extractor \mathbf{f}_T , the optimal target classifier \mathbf{g}_T minimizing $\mathcal{L}(\theta_T)$ can be directly calculated by generalizing MCR to a multiple extractor combination scenario.

Theorem 1. *Given feature extractors of function form $\mathbf{f}_i(x)$, $i \in \{1, \dots, n\}$ and the linearly combined feature extractor $\mathbf{f}(x) = \sum_{i=1}^n \alpha_i \cdot \mathbf{f}_i(x)$, where $\mathbb{E}_{P_X}[\mathbf{f}_i(x)] = 0$, $\mathbb{E}_{P_X}[\mathbf{f}_i(x) \cdot \mathbf{f}_i(x)^T] = \mathbf{I}, \forall i$ and $\sum_{i=1}^n \alpha_i = 1$. The optimal classifier $\mathbf{g}^*(y)$ will be of the form:*

$$\mathbf{g}^*(y) = \sum_{i=1}^n \alpha_i \cdot \mathbb{E}_{P_{X|Y}}[\mathbf{f}_i(X)|Y = y], \quad (7)$$

where the proof is provided in the Appendix.I.

By using Eqn. (7), we can skip training and directly set the target classifier $\mathbf{g}_T(y)$ to the theoretical optimal $\mathbf{g}_T^*(y) =$

$\sum_{i=1}^n \alpha_i \cdot \mathbb{E}_{P_{X|Y}}[\mathbf{f}_i(X)|Y = y]$, thereby achieving high efficiency in transfer.

Notably, the form of $\mathbf{g}_T(y)$ does not conform to the usual classifiers that take in extracted features and output the class probabilities (as in the source classifiers $\mathbf{g}_{s_j} : \mathbb{R}^{d_f} \rightarrow \mathbb{R}^{d_{s_j}}$). On the contrary, the target classifier $\mathbf{g}_T(y)$ acts as a label encoder, which maps labels y to a newly defined y-feature, i.e. $\mathbf{g}_T : \mathbb{R}^{d_T} \rightarrow \mathbb{R}^{d_f}$. In testing mode, the classification result is given by maximizing the correlation between the features extracted, i.e. $\hat{y}(x) = \arg \max_y \mathbf{f}_T^T(x)\mathbf{g}_T(y)$. More details of this network architecture are discussed in previous works (?).

H-Score for Estimating Feature Transferability

Defining the target feature as $\mathbf{f}_T = \sum_{j=1}^M \alpha_j \cdot \mathbf{f}_{s_j}$, we introduce the information theoretic metric H-score (?) to measure the feature transferability for weight optimization. In this section, we give a detailed formulation of H-score as well as its intuitive interpretation. From MCA, the mathematical form of H-score is defined as follows.

Definition 1. *With input data x , label y , feature extractor $\mathbf{f}(x)$ and classifier $\mathbf{g}(y)$ (both zero-mean functions). The H-score of \mathbf{f} with regard to the task casting x to y is:*

$$H(\mathbf{f}, \mathbf{g}) = \mathbb{E}_{P_{XY}}[\mathbf{f}^T(X)\mathbf{g}(Y)] - \frac{1}{2}\text{tr}(\text{cov}(\mathbf{f}(X))\text{cov}(\mathbf{g}(Y))). \quad (8)$$

Algorithm 1: H-ensemble: Training

Input: Target data $D_T = \{(x_T^i, y_T^i)\}_{i=1}^{N_T}$, source feature extractors $\{f_{S_j}\}_{j=1}^M$

Parameter: Learning rate λ

Output: Target classifier g_T , source weight α

Randomly initialize $\alpha = \{\alpha_1, \alpha_2, \dots, \alpha_M\}$, $\sum_{j=1}^M \alpha_j = 1$;

```
for  $y \leftarrow 1$  to  $d_T$  do                                     // Compute empirical conditional expectation
     $n_y \leftarrow \sum_{i=1}^{N_T} \mathbb{I}\{y = y_T^i\}$ ;
     $\mathbb{E}_{P_{X|Y}^T}[f_{S_j}(X_T)|Y_T = y] \leftarrow \frac{1}{n_y} \sum_{i=1}^{N_T} f_{S_j}(x_T^i) \cdot \mathbb{I}\{y = y_T^i\}$ ,  $j = 1, \dots, M$ 
end
repeat                                                         // Optimize weight
     $H(f_T; \alpha) \leftarrow \text{tr}(\text{cov}(\sum_{j=1}^M \alpha_j \mathbb{E}_{P_{X|Y}^T}[f_{S_j}(X_T)|Y_T]));$ 
     $\alpha \leftarrow \alpha + \lambda \nabla_{\alpha} H(f_T; \alpha);$ 
    for  $j \leftarrow 1$  to  $M$  do                                     // Project weight to hyperplane  $\sum_{j=1}^M \alpha_j = 1$ 
         $\alpha_j \leftarrow \alpha_j - \frac{1}{M} \sum_{j=1}^M \alpha_j + \frac{1}{M}$ 
    end
until  $\alpha$  converges;
for  $y \leftarrow 1$  to  $d_T$  do                                     // Compute classifier function
     $g_T(y) = \sum_{j=1}^M \alpha_j \cdot \mathbb{E}_{P_{X|Y}}[f_{S_j}(X_T)|Y_T = y]$ 
end
```

Since we have defined our classifier $g_T(y)$ as the optimal classifier $g_T^*(y)$ in the previous section, the H-score can be simplified to the one-sided H-score.

Definition 2. With input data x , label y and feature extractor $f(x)$ (a zero-mean feature function). The one-sided H-score of f with regard to the task casting x to y is:

$$H(f) = \text{tr}(\text{cov}(f(X))^{-1} \text{cov}(\mathbb{E}_{P_{X|Y}}[f(X)|Y])). \quad (9)$$

Theorem 2. For a task T with feature extractor f and classifier g , the minimization of \mathcal{L} in maximal correlation regression is equivalent to the maximization of H-score, i.e.

$$f^*, g^* = \arg \min_{f, g} \mathcal{L} = \arg \max_{f, g} H(f, g). \quad (10)$$

Moreover, if g is defined as the optimal classifier g^* , $H(f, g)$ is then equal to the one-sided H-score $H(f)$, i.e.

$$H(f, g^*) = H(f). \quad (11)$$

Theorem 2 indicates that the minimization of $\mathcal{L}(\theta)$ is equivalent to the maximization of the one-sided H-score $H(f)$, and therefore we can determine the optimal target feature in our method by simply maximizing $H(f_T)$.

The formula of one-sided H-score in Eqn. (9) can be intuitively interpreted as normalizing the inter-class feature variance $\text{cov}(\mathbb{E}_{P_{X|Y}}[f(X)|Y])$ with feature redundancy $\text{tr}(\text{cov}(f(X)))$, reflecting the efficiency of extracted feature in distinguishing different classes. The proof of Eqn. (8), Eqn. (9) and Theorem 2 can be found in (??). Considering that our target feature f_T is normalized, we can further simplify Eqn. (9) to:

$$H(f_T) = \text{tr}(\text{cov}(\mathbb{E}_{P_{X|Y}^T}[f(X_T)|Y_T])). \quad (12)$$

Eqn. (12) can be explicitly computed from extracted features and labels. Hence, using target data D_T , we could efficiently estimate in advance the transferred performance, or transferability, of target feature $f_T(x)$, and optimize the source weights accordingly. For clarity, in the following sections we refer to the score defined in Eqn. (12) as ‘H-score’.

Source Feature Re-weighting

With the source feature extractors and target samples fixed, the target feature f_T can be viewed as a function of the weight α defined in Eqn. (4) and therefore the same for its H-score $H(f_T)$. Hence, by maximizing $H(f_T; \alpha)$ with respect to α , we will be able to determine the optimal weight α for minimizing the loss $\mathcal{L}(\theta)$ in multi-source transfer, alternately interpreted as deriving the target feature of the highest transferability. We formulate our optimization problem as follows.

Definition 3. With input data x_T , label y_T and feature extractors $f_{S_j}(x)$ for $j \in \{1, \dots, M\}$ (all zero-mean, unit-variance feature functions), the optimal feature weight $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_M)^T \in \mathbb{R}^M$ will be given by:

$$\alpha^* = \arg \max_{\alpha} H \left(\sum_{j=1}^M \alpha_j \cdot f_{S_j} \right) \text{ s.t. } \sum_{j=1}^M \alpha_j = 1. \quad (13)$$

The linear constraint of α is obviously convex. We then verify the benign property of the proposed optimization problem by proving that the optimal object H-score is also a convex function of α .

Theorem 3. With input data x and label y , when weighted α_i summing up to 1 and fixed features \mathbf{f}_i being zero-mean, unit-variance ($i \in \{1, \dots, n\}$), the H-score of the weighted feature sum will be a convex quadratic form of α as below:

$$H(\mathbf{f}) = H\left(\sum_{i=1}^n \alpha_i \cdot \mathbf{f}_i\right) = \sum_{i=1, j=1}^{n, n} \alpha_i \alpha_j \cdot \text{tr}(\mathbb{E}_{P_Y}[\mathbb{E}_{P_{X|Y}}[\mathbf{f}_i(X)|Y] \cdot \mathbb{E}_{P_{X|Y}}[\mathbf{f}_j(X)|Y]^T]). \quad (14)$$

A detailed proof of Theorem 3 is presented in the Appendix.I. The optimization problem of α can therefore be reliably solved by gradient descent (GD) based algorithms with theoretical guarantee. We then use the optimal α and derive the optimal target model as defined in Eqn. (3) with Eqn. (4) and (7).

As a possible solution, we use projected gradient descent (PGD) for optimization and present the resulting training process of H-ensemble in Algorithm 1. The derivation of the projection formula in PGD is included in Appendix.I. The predicted result in testing will be efficiently computed by:

$$\hat{y}(x) = \arg \max_y \left(\sum_{j=1}^M \alpha_j \mathbf{f}_{S_j}^T(x) \right) \cdot \mathbf{g}_T(y). \quad (15)$$

Experimental Results

Experiment Setup

Datasets. To verify the effectiveness of H-ensemble in the few-shot MSF setting, we conduct extensive experiments on four benchmark datasets. **VisDA-2017** (?) is a visual domain transfer challenge dataset containing over 280,000 images across 12 categories in training (T, synthetic) and validation (V, real) domains respectively. **Office-31** (?) is a standard transfer learning dataset with 4,652 images and 31 unbalanced object categories in three domains: Amazon, DSLR and Webcam, consisting of objects commonly encountered in office scenarios. Developed from Office-31, **Office-Caltech** (?) contains 2,533 images covering four domains: Amazon, DSLR, Webcam and Caltech256, each with 10 categories. **Office-Home** (?) is a more challenging dataset with 15,599 images in 65 unbalanced categories collected from four domains: Artistic, Clip Art, Product, and Real-world.⁵

Task Setting. We split each domain into several tasks. For example, in VisDA-2017 we further divide both domain V and T into 4 tasks (*i.e.* v0 - v3 and t0 - t3), each with 3 classes. The resulting eight tasks, encompassing both domain and label differences, constitute a diverse task pool for the transfer experiment. We do the same on the Office series datasets and the details can be found in the Appendix.II.

Following the standard protocol of few-shot learning, the training data for k -shot is k samples per class randomly

⁵For an evaluation on large-scale datasets, we also compared H-ensemble against other methods on **DomainNet** (?) during rebuttal. The results are included in the Appendix.III.

selected from the target task. Source models are set to Resnet18 with hidden dim 256 trained on full training sets. More details of extra dataset, experimental setup and model implementation are discussed in the Appendix.II.

Comparison with Baselines

Baselines. For a general performance evaluation, we conduct comparison experiments using the first task synthesis strategy. Considering very few methods are designed for few-shot MSF transfer setting, we take SOTA works under similar settings and variations of our method as baselines. The compared methods include: 1) **Trivial Solutions:** Target Only, Single-Best, Single-Avg.; 2) **Relevant Approaches:** Average W., MultiFinetune; 3) **Prior SOTAs:** MCW (?), DECISION (?), DATE (?). Here, Single-* adapt one source model to target task using MCR and *-Best/Avg. stand for the best/average performance. Target Only finetunes Imagenet-pretrained model on few-shot target data only. Average W. replace the weighted sum in our framework with equal weights. MultiFinetune learns a fully connected classifier with the target extractor learned in H-ensemble. The Prior SOTAs are the SOTAs with the closest problem setting (MSFDA).

Results and Analysis. We take VisDA-2017 10-shot as an example⁶ and summarize the results in Table 3. Overall, H-ensemble performs the best among all the methods including the best single source transfer, especially in the more challenging real world image (V) domain. Here the significant degradation of DECISION and DATE is due to the change of problem setting from UDA to FS transfer.

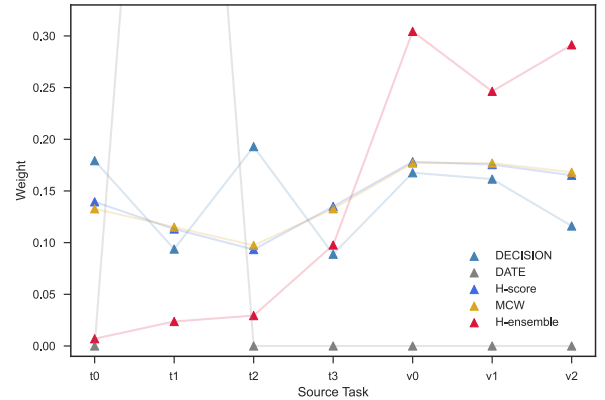


Figure 3: **Source Weights derived in Different Methods.** (VisDA-2017, target v3, 10-shot) The source weight in H-ensemble (red) conforms most to intuition, emphasizing on tasks from the same domain (v0, v1, v2).

For further analysis, we plot the source weights for task v3 derived in weighting-based methods in Fig. 3. It is clearly illustrated that given four tasks on domain V and T each, H-ensemble can successfully recognize the importance of integrating tasks from the same domain (v0, v1, v2) and attach

⁶Experiments on other datasets & shots and statistical results are listed in Appendix.III.

Method	$R \rightarrow t0$	$R \rightarrow t1$	$R \rightarrow t2$	$R \rightarrow t3$	$R \rightarrow v0$	$R \rightarrow v1$	$R \rightarrow v2$	$R \rightarrow v3$	Avg.
TargetOnly	0.9730	0.9675	0.9165	0.7260	0.9180	0.8285	0.8750	0.7645	0.8711
Single-Average	0.9697	0.9659	0.9266	0.7503	0.8416	0.7640	0.7753	0.7013	0.8368
Single-Best	0.9830	0.9725	<u>0.9405</u>	0.7760	0.9435	0.8470	<u>0.8915</u>	<u>0.8110</u>	<u>0.8956</u>
Average W.	0.9840	0.9770	0.9340	0.7510	0.9240	0.8370	0.8355	0.7910	0.8792
MultiFinetune	0.9695	0.9715	0.9400	0.7570	0.9105	0.8150	0.8240	0.7760	0.8704
MCW	0.9890	0.9585	0.9415	0.7675	<u>0.9460</u>	0.8160	0.8655	0.7600	0.8805
DECISION	0.5590	0.4635	0.7700	0.3125	0.3055	0.3245	0.5435	0.3040	0.4478
DATE	0.2490	0.3655	0.7740	0.2635	0.2500	0.4545	0.5155	0.1260	0.3748
H-ensemble*	<u>0.9855</u>	<u>0.9745</u>	0.9415	<u>0.7690</u>	0.9470	<u>0.8420</u>	0.8990	0.8300	0.8986

Table 3: **Accuracy Comparison on VisDA-2017 (10-shot)**. ‘R’ stands for the rest tasks. The highest/second-highest accuracy is marked in **Bold**/Underscore form respectively. Our method achieves the overall best performance.

them with high weights. It also correctly emphasizes on $t3$ (the only task sharing the same class labels with $v3$) the most among tasks on domain T, displaying an outstanding and interpretable ability to measure source importance.

Ablation Study

We further explore the capability of H-ensemble when transferring to harder tasks on Office datasets, as well as verify components in H-ensemble by ablating. The baselines are constructed by substituting each part with naive solutions. Specifically, we have:

- **Single MCR**: multi-source \rightarrow average performance of single-source transfer;
- **Average W.**: weighting strategy \rightarrow average weights;
- **MultiFinetune**: MCR classifier \rightarrow finetune classifier.

For each variety, we record the average transfer performance on each dataset, concluding the results of 8-shot⁷ in Table 4, where both the effectiveness of H-ensemble and the necessity of every module are explicitly supported. It is also notable that our whole framework is established on a united theory ground, and any substitution may break its theoretical reliability and interpretability. We also visualize the target feature extracted by random source, average weight and H-ensemble in Fig. 4, showing that H-ensemble extracts features with the strongest class discriminability for the given task.

Conclusion

In this work, we highlight a common yet largely underexplored scenario of multi-source transfer which we refer to as few-shot MSF transfer, and take the lead in giving a mathematical formulation of it. Addressing this problem, we take on an information theoretic perspective and propose our H-ensemble framework. Consisting of three main components, H-ensemble dynamically learns the optimal linear ensemble of source models for the target task, with the ensemble

⁷Compared to VisDA-2017, the shot reduces due to fewer samples per class in Office-* datasets. More results in Appendix.III.

Method	Office-31	Office-Caltech	Office-Home
Single MCR	<u>0.8483</u>	0.9645	0.6235
Average W.	0.8479	0.9716	<u>0.6702</u>
MultiFinetune	0.8414	0.9663	0.6571
H-ensemble*	0.8644	<u>0.9708</u>	0.6940

Table 4: **Ablation Study on Office Datasets (8-shot)**. Our weighting strategy and optimal MCR classifier both contribute to the overall effectiveness.

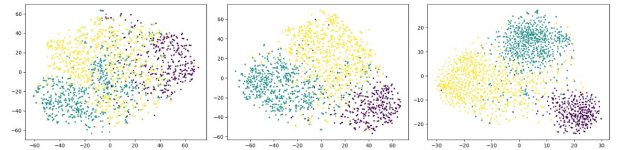


Figure 4: **Feature extracted by different models**, visualized by t-SNE. From left to right: random source, Average W. and H-ensemble. It turns out that features generated by our method have the greatest discriminability.

weights optimized by maximizing H-score and classifier determined by a generalization of MCR. We present detailed theoretic deduction and interpretation as well as extensive experimental validation for our method, showing that H-ensemble can effectively boost the learning on the target task in the few-shot MSF transfer scheme.

Acknowledgement

This study is supported in part by the Tsinghua SIGS Scientific Research Start-up Fund (Grant QD2021012C), Natural Science Foundation of China (Grant 62001266) and Shenzhen Key Laboratory of Ubiquitous Data Enabling (No. ZDSYS20220527171406015).

quasi possimus tempora repellendus modi velit, quod dicta dignissimos cumque deserunt perspiciatis, earum iure

vero modi provident neque aut sunt rerum consequatur accusamus itaque, reiciendis eum deserunt tempore?