Hypothesis Disparity Regularized Mutual Information Maximization

Qicheng Lao, [1 2 3] **Xiang Jiang,** [3] **Mohammad Havaei** [3]
[1] West China Biomedical Big Data Center, West China Hospital of Sichuan University, Chengdu, China
[2] MILA, Université de Montréal, Montreal, Canada
[3] Imagia,* Montreal, Canada
qicheng.lao@gmail.com, xiang.jiang@imagia.com, mohammad@imagia.com

Abstract

We propose a hypothesis disparity regularized mutual information maximization (HDMI) approach to tackle unsupervised hypothesis transfer—as an effort towards unifying hypothesis transfer learning (HTL) and unsupervised domain adaptation (UDA)—where the knowledge from a source domain is transferred solely through hypotheses and adapted to the target domain in an unsupervised manner. In contrast to the prevalent HTL and UDA approaches that typically use a single hypothesis, HDMI employs multiple hypotheses to leverage the underlying distributions of the source and target hypotheses. To better utilize the crucial relationship among different hypotheses—as opposed to unconstrained optimization of each hypothesis independently—while adapting to the unlabeled target domain through mutual information maximization, HDMI incorporates a hypothesis disparity regularization that coordinates the target hypotheses jointly learn better target representations while preserving more transferable source knowledge with better-calibrated prediction uncertainty. HDMI achieves state-of-the-art adaptation performance on benchmark datasets for UDA in the context of HTL, without the need to access the source data during the adaptation.

Introduction

Mutual information (MI) maximization has been shown as a promising approach in unsupervised learning, as manifested in discriminative clustering (Bridle, Heading, and MacKay 1992; Krause, Perona, and Gomes 2010) and unsupervised representation learning (Hu et al. 2017; Hjelm et al. 2018; Tian, Krishnan, and Isola 2019; Oord, Li, and Vinyals 2018). Recently, it has also been applied to unsupervised domain adaptation (UDA), achieving new state-of-the-art performance even in the more restricted context of hypothesis transfer learning (HTL [1]) where the transfer of the knowledge from a source domain to a target domain is achieved solely through hypotheses (Liang, Hu, and Feng 2020). Hypothesis transfer has the notable privacy-preserving property that respects the privacy of the source domain by eliminating the need to access the source data while transferring knowledge to the target domain, and is favored by both theoretical analysis (Kuzborskij and Orabona 2013; Ben-David and Urner 2013; Perrot and Habrard 2015; Kuzborskij and Orabona 2017) and many empirical applications (Fei-Fei, Fergus, and Perona 2006; Yang, Yan, and Hauptmann 2007; Orabona et al. 2009; Jie, Tommasi, and Caputo 2011; Tommasi, Orabona, and Caputo 2013; Du et al. 2017; Fernandes and Cardoso 2019). However, HTL has been mostly explored in the supervised learning setting where the target labels are available except the work in (Liang, Hu, and Feng 2020).

In this paper, we further tackle the problem of unsupervised hypothesis transfer (Figure 1 (d)) under the umbrella of MI maximization, as an effort to bridge the gap between HTL (Figure 1 (b)) and UDA (Figure 1 (c)). In contrast to the prevalent approaches for HTL and UDA that tend to use a single hypothesis failing to uncover different modes of the model distribution, we propose to transfer knowledge from a set of *source hypotheses* learned from the source domain to a corresponding set of *target hypotheses* by means of MI maximization on the unlabeled target data. The employment of multiple hypotheses is especially relevant to domain adaptation with out-of-distribution examples, and has a pronounced impact on the uncertainty calibration as well as the final adaptation or transfer performance, as has been demonstrated in previous work such as deep ensembles (Lakshminarayanan, Pritzel, and Blundell 2017).

Furthermore, we highlight the crucial limitation of multiple independent MI maximization where different target hypotheses can be optimized in an unconstrained manner due to the lack of supervision, resulting in undesirable disagreements on the target label predictions as well as instability during the optimization process. To overcome such limitations and better take advantage of the relationship among different hypotheses, we propose a hypothesis disparity (HD) regularization to align the target hypotheses in a way such that the uncertainty manifested through different source hypotheses is taken into account while the undesirable disagreements are marginalized out. The HD regularization also shares the similar idea in unsupervised discriminative clustering with regularized information maximization, where a complexity penalty term is shown to be indispensable (Krause, Perona, and Gomes 2010). We term the proposed multiple hypotheses MI maximization with HD regularization as Hypothesis Dis-

*A US provisional patent application has been filed for protecting at least one part of the innovation disclosed in this article. This work was done at Imagia, funded through MEDTEQ grant 10-30 IA Multicentriq.

[1] HTL was first introduced in (Kuzborskij and Orabona 2013).

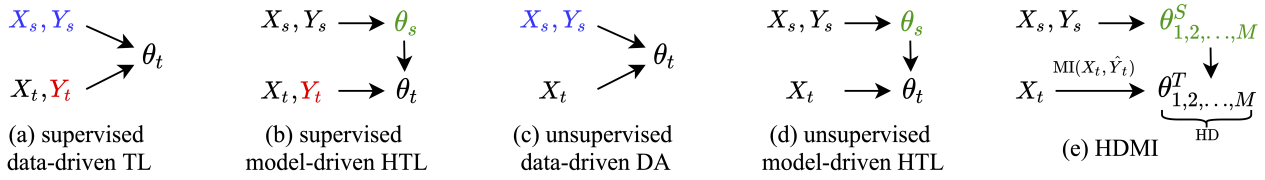


Figure 1: Graphical models for transfer learning. (a)-(d) Settings of transfer learning depending on the transfer medium and the availability of target labels; (e) Our proposed HDMI for unsupervised hypothesis transfer learning. The colors indicate direct access to the source data, target labels and source hypotheses during target adaptation.

parity regularized Mutual Information maximization (HDMI), illustrated in Figure 1 (e).

Finally, we evaluate the proposed HDMI approach on three benchmark datasets for domain adaptation in the context of unsupervised hypothesis transfer. We show that (i) the proposed HD regularization is effective in minimizing the undesirable disagreements among different target hypotheses and stabilizing the MI maximization process; (ii) Compared to direct MI maximization with single hypothesis or multiple hypotheses, the HD regularization facilitates the positive transfer of multiple modes from source hypotheses, and as a result, the target hypotheses obtained by HDMI preserve more transferable knowledge from each source hypothesis; (iii) HDMI uses well-calibrated predictive uncertainty to achieve effective MI maximization; and (iv) HDMI works through learning better representations shared by different target hypotheses. Overall, HDMI achieves new state-of-the-art performance in unsupervised hypothesis transfer learning.

Related Work

Relation to Hypothesis Transfer Learning Approaches for transfer learning can be categorized into data-driven approaches (*e.g.*, instance weighting, feature transformation) and model-driven approaches (Pan and Yang 2009; Zhuang et al. 2019; Fernandes and Cardoso 2019). The differences between the two categories are illustrated in Figure 1 (a) and (b) in a supervised transfer learning setting. In this work, we focus on the model-driven category, which is also referred to as HTL in the literature (Kuzborskij and Orabona 2013). HTL was first introduced in (Kuzborskij and Orabona 2013) with theoretical analysis on a regularized least squares problem, and later on extended to a general regularized empirical risk minimization problem (Kuzborskij and Orabona 2017). Approaches for HTL have also been proposed based on transformation functions (Du et al. 2017) and model structure similarity (Fernandes and Cardoso 2019). However, all the previous HTL approaches assume access to the labeled data in the target domain, *i.e.*, supervised HTL, whereas we explore the possibility of unsupervised HTL in this work.

Relation to Unsupervised Domain Adaptation Unsupervised domain adaptation, also considered as a form of transfer learning (transductive transfer learning (Pan and Yang 2009)), aims to adapt a target domain to a source domain without requiring target labels, and has also been extensively studied in the past few years (Ganin and Lempitsky 2015; Long et al. 2015; Tzeng et al. 2017; Hoffman et al. 2018; Saito et al. 2018; Zhang et al. 2019). While most work assumes simultaneous access to both source and target data during adaptation, a few argue against the reality and necessity of such assumption (Chidlovskii, Clinchant, and Csurka 2016; Liang et al. 2019; Liang, Hu, and Feng 2020). We follow the setting of (Liang, Hu, and Feng 2020), where the source data is inaccessible during target adaptation, and the source knowledge is transferred solely through hypotheses.

Relation to Mutual Information Maximization The mutual information maximization can be achieved between input and output (Bridle, Heading, and MacKay 1992; Krause, Perona, and Gomes 2010), between input and intermediate representation or context (Hu et al. 2017; Hjelm et al. 2018; Oord, Li, and Vinyals 2018), or between representations from different views (Tian, Krishnan, and Isola 2019; Bachman, Hjelm, and Buchwalter 2019). In this work, we maximize the MI between the input data of the target domain and the corresponding pseudo-labels predicted by the target hypothesis. In addition, we extend MI maximization to multiple hypotheses and also introduce a regularization term for MI maximization with multiple hypotheses. Unlike the regularized information maximization (Krause, Perona, and Gomes 2010) with the penalty term on the complexity of a single hypothesis, we emphasize on the disparity among multiple hypotheses; Compared with previous work in HTL that places the regularization directly between the source and target hypotheses (Kuzborskij and Orabona 2013; Fernandes and Cardoso 2019), our proposed HD regularization for MI maximization is an indirect and relaxed form of regularization that is only among the target hypotheses.

MI maximization has also been demonstrated to have better performance in discriminative clustering (Krause, Perona, and Gomes 2010), compared with conditional entropy minimization (Grandvalet and Bengio 2005) that was proposed for semi-supervised learning. In domain adaptation, approaches have been proposed based on MI maximization (Liang, Hu, and Feng 2020) and conditional entropy minimization, *e.g.*, integrating with a minmax game (Saito et al. 2019), virtual adversarial training [2] (Shu et al. 2018; Kumar et al. 2018), or correlation alignment (Morerio, Cavazza, and Murino 2017). HDMI is closely related to (Liang, Hu, and Feng 2020) in the sense that both are MI-based; however, instead of using a pseudo-label based self-training strategy to overcome the limitations of MI maximization in UDA, we propose to directly improve on MI maximization itself.

[2] VAT was originally proposed in (Miayto, Dai, and Goodfellow 2017) for semi-supervised text classification.

Relation to Auxiliary Classifiers and Ensemble Methods The multiple hypotheses used in HDMI also have a connection to auxiliary classifiers used in multi-task learning (Ruder 2017), domain adaptation from multiple sources (Mansour, Mohri, and Rostamizadeh 2009; Duan et al. 2009; Peng et al. 2019) and hypothesis transfer from auxiliary hypotheses (Yang, Yan, and Hauptmann 2007; Tommasi, Orabona, and Caputo 2013; Kuzborskij and Orabona 2017). While the auxiliary classifiers aim to leverage the knowledge learned from multiple different tasks or sources, our multiple hypotheses in this work focus on covering different modes of the underlying hypothesis distribution learned from a single source domain for a single task. In addition, HDMI also shares some appealing properties with ensemble methods, such as deep ensemble that improves uncertainty calibration (Lakshminarayanan, Pritzel, and Blundell 2017; Fort, Hu, and Lakshminarayanan 2019), and ensemble adversarial training for robustness (Tramèr et al. 2017). However, HDMI differs from ensemble methods by exploiting a hypothesis disparity regularization during the unsupervised optimization process, and we show later in our experiments that two hypotheses suffice HDMI to benefit from this regularization.

Approach

Problem Formalization

Let $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{H}$ be the input space, the output space, and the hypothesis space, respectively. We denote the source domain data as $\mathcal{D}_s = \{(x_i^S, y_i^S)\}_{i=1}^{N_s}$, where $x_i^S \in \mathcal{X}^S$ and $y_i^S \in \mathcal{Y}^S$. Similarly, the unlabeled target data is denoted as $\mathcal{D}_t = \{(x_i^T)\}_{i=1}^{N_t}$, where $x_i^T \in \mathcal{X}^T$ and $\mathcal{X}^T \neq \mathcal{X}^S$. In unsupervised HTL, the assumption is that the learning task $\mathcal{T}$ remains the same between the source and target domain where $\mathcal{Y}^S = \mathcal{Y}^T$ and $P_S(Y|X) = P_T(Y|X)$, which is also a typical assumption in closed-set UDA (Panareda Busto and Gall 2017). Suppose a source hypothesis $h_s : \mathcal{X}^S \to \mathcal{Y}^S$ and a target hypothesis $h_t : \mathcal{X}^T \to \mathcal{Y}^T$, we have the following posterior predictive distribution from a Bayesian perspective:

$$p(Y_t^*|\mathcal{D}_t, \mathcal{D}_s) = \int_{h_t} p(Y_t^*|\mathcal{D}_t, h_t) \int_{h_s} p(h_t|\mathcal{D}_t, h_s) p(h_s|\mathcal{D}_s) dh_s dh_t, \tag{1}$$

with the goal to predict the unobserved target labels Y_t^* by marginalizing over the posterior probabilities of both source and target hypotheses. Note that the target posterior $p(h_t|\mathcal{D}_t, h_s)$ is consistent with Figure 1 (d) where h_t is conditioned on h_s and $\mathcal{D}_t$, without direct access to the source data $\mathcal{D}_s$. In contrast to the prevalent UDA approaches that assume $h_t = h_s$ and HTL approaches that typically use a single source and target hypothesis [3], we propose to employ multiple hypotheses to better utilize the underlying distributions of source hypothesis h_s and target hypothesis h_t.

Below we describe the proposed approach that leverages the idea of multiple hypotheses, *i.e.*, given multiple source hypotheses trained on the source domain, we extract target knowledge from the unlabeled target data via mutual information maximization, with the constraint of minimized target hypothesis disparity, thus resulting in HDMI.

[3] Cases with multiple source domains are not included here.

Learning Multiple Source Hypotheses

As a first step, we learn a set of source hypotheses $\{h_i^S \in \mathcal{H}^S : \mathcal{X}^S \to \mathcal{Y}^S\}_{i=1}^M$ on the source data $\mathcal{D}_s$. We consider a set of source hypotheses $\{h_i^S : h_i^S = f_i^S \circ \psi^S\}_{i=1}^M$ that use a shared feature extractor ψ^S but M independent classifiers $\{f_i^S\}_{i=1}^M$ trained from different random initialization. Learning multiple source hypotheses is similar to *deep ensemble* (Fort, Hu, and Lakshminarayanan 2019) with the notable ability to learn diverse functions from different modes, whereas single hypothesis learning with maximum a posteriori only uncovers a single mode of the posterior $p(h_s|\mathcal{D}_s)$. In addition, compared with approximate Bayesian approaches like Monte Carlo dropout (MC-dropout) (Gal and Ghahramani 2016), deep ensemble with random initialization is shown to produce well-calibrated uncertainty estimations that are more robust to domain shift (Lakshminarayanan, Pritzel, and Blundell 2017). This is especially relevant to HTL and has a pronounced impact on the transfer performance.

For ease of exposition, given an input x, we denote $h(x) = p(y|x; h)$ as the output label probability distribution predicted by a hypothesis h, where $y \in \{1, \dots K\}$ and K is the number of classes. The multiple source hypotheses are learned jointly by minimizing the following objective function:

$$\mathcal{L}_{source} = \mathbb{E}_{h \in \mathcal{H}^S, (x,y) \in \mathcal{X}^S \times \mathcal{Y}^S} \left[\ell_{\text{CE}}(h(x), y) \right], \tag{2}$$

where ℓ_{CE} denotes the cross entropy loss function. In practice, we use the average of M hypotheses to approximate the expectation in Eq. 2.

Learning Target Hypotheses via Mutual Information Maximization

Given the unlabeled target data $\mathcal{D}_t = \{(x_i^T)\}_{i=1}^{N_t}$ and a set of previously learned source hypotheses $\{h_i^S\}_{i=1}^M$, we aim to adapt the source hypotheses into a set of corresponding target hypotheses $\{h_i^T \in \mathcal{H}^T\}_{i=1}^M$ by maximizing the mutual information between the empirical target input distribution and the predicted target label distribution induced by the target hypotheses. Let X^T denote the random variable for the target input, and $\hat{Y}^T$ denote the random variable of the model prediction inferred from hypothesis h with the empirical label distribution $p(\hat{y}^T; h) = \frac{1}{N} \sum_i p(y|x_i^T; h)$. The MI between the target input X^T and the output $\hat{Y}^T$ can be written as:

$$I(X^T; \hat{Y}^T) = H(\hat{Y}^T) - H(\hat{Y}^T|X^T). \tag{3}$$

With a set of target hypotheses $\{h_i^T\}_{i=1}^M$, the optimization process can be expressed as jointly maximizing the expectation of MI given in Eq. 3 over the target hypotheses:

$$\max_{\psi^T} \mathbb{E}_{h \in \mathcal{H}^T} \left[I(X^T; h(X^T)) \right], \tag{4}$$

where ψ^T denotes the shared feature extractor among the M target hypotheses and $h(X^T) := \hat{Y}^T$. Similar to (Liang, Hu, and Feng 2020), we fix the parameters of the classifiers for the target hypotheses (*i.e.*, $f_i^T = f_i^S$) while updating ψ^T, due to the fact that both source and target domains share the same label space.

Target Hypothesis Disparity Regularization

In addition to the multiple hypotheses MI maximization (referred to as *MI ensemble*) proposed in Eq. 4, we introduce a hypothesis disparity regularization to better take advantage of the relationship among different hypotheses. The proposed regularization is motivated by the crucial limitation of *MI ensemble* where different hypotheses can be optimized in an unconstrained manner, resulting in undesirable disagreements on the target label predictions due to the unsupervised adaptation procedure. This is also in alignment with the finding of using MI for unsupervised discriminative clustering, where it is shown that a complexity penalty term is indispensable for MI maximization (Krause, Perona, and Gomes 2010). The proposed regularization aims to take into account the uncertainty manifested through different hypotheses so as to marginalize out the undesirable disagreements resulted from *MI ensemble*.

Here, we define hypothesis disparity (HD) as a dissimilarity measure of the predicted label probability distributions among different hypotheses over the input space $\mathcal{X}$:

$$\text{HD}_{h_i,h_j\in\mathcal{H},i\neq j}(h_i,h_j)=\int_{\mathcal{X}} d(h_i(x),h_j(x))p(x)dx, \quad (5)$$

where $d(\cdot)$ can be any divergence measure between the predicted label probability distributions from the two hypotheses, *e.g.*, $-\sum_K h_i(x)\log h_j(x)$ if using cross entropy as the divergence measure with K unique labels. We discuss the relationship between cross entropy and Kullback–Leibler divergence as $d(\cdot)$ in MI maximization in the supplementary material, and provide empirical comparison in Table 5. We use cross entropy for $d(\cdot)$ throughout this paper.

We show with the empirical evidence that minimizing the HD among target hypotheses can effectively regularize the target hypotheses to maximally agree with each other, and help to coordinately learn better representation through the shared feature extractor resulting in better performance.

HDMI

We now present our proposed approach—HDMI—by integrating the HD regularization into the *MI ensemble*. We denote $R(\mathcal{H})$ as a general form of regularization imposed on the hypothesis space $\mathcal{H}=\mathcal{H}^S\cup\mathcal{H}^T$. Then we have the following objective function for MI-based unsupervised HTL:

$$\mathbb{E}_{h\in\mathcal{H}^T}\left[-I(X^T;h(X^T))\right]+\lambda R(\mathcal{H}), \quad (6)$$

and the proposed HDMI can be given as:

$$\mathbb{E}_{h\in\mathcal{H}^T}\left[-I(X^T;h(X^T))\right]+\lambda\mathbb{E}_{h_{i,j}\in\mathcal{H}^T,i\neq j}\left[\text{HD}(h_i,h_j)\right]. \quad (7)$$

For computational efficiency, we set an anchor hypothesis that is randomly chosen from target hypotheses, and compute the average of the disparity between the anchor hypothesis and the rest $M-1$ hypotheses. In addition, if the HD among target hypotheses is minimized, the posterior predictive distribution in Eq. 1 that computes the expectation of label predictions over the target hypothesis space can be equivalently simplified to the prediction from any hypothesis sampled from the target posterior:

$$p(Y_t^*|\mathcal{D}_t,\mathcal{D}_s)\simeq p(Y_t^*|\mathcal{D}_t,h_t), h_t\sim p(h_t|\mathcal{D}_t,\{h_i^S\}_{i=1}^M). \quad (8)$$

Therefore, we report the performance of the anchor hypothesis by using Eq. 8 as our final HDMI performance, as compared with *HDMI ensemble* that uses Eq. 1 for the target predictive performance. Experimental results (Table 1 and Figure 3) also empirically confirm the two are equivalent.

Our proposed HDMI with HD regularization is related to previous methods that use other forms of regularization: if $R(\mathcal{H})=\boldsymbol{w}_t^\intercal\boldsymbol{w}_t$, with $\boldsymbol{w}_t$ denoting the parameters of a target hypothesis, we reach the regularized information maximization approach proposed for discriminative clustering in a single domain (Krause, Perona, and Gomes 2010); if $R(\mathcal{H})=\|\boldsymbol{w}_t-\boldsymbol{w}_s\|_2^2$, we obtain the typical L_2 regularization between the source and target hypotheses (Kuzborskij and Orabona 2013; Fernandes and Cardoso 2019). We empirically find that the proposed HD regularization is superior to both of them (denoted as L_2 and L_2 *source*, respectively, in Table 1 and Table 5).

Experiments

Setup

We validate HDMI on three benchmark datasets for UDA in the context of HTL, and compare the adaptation/transfer performance with various state-of-the-art UDA and HTL methods as baselines.

Datesets **Office-31** (Saenko et al. 2010) has three domains: Amazon (A), DSLR (D) and Webcam (W), with 31 classes and 4,652 images. **Office-Home** (Venkateswara et al. 2017) is a more challenging dataset with 65 classes and 15,500 images in four domains: Artistic images (Ar), Clip art (Cl), Product images (Pr) and Real-World images (Rw). **VisDA-C** (Peng et al. 2018) is a large-scale dataset with 12 classes, with 152,397 Synthetic images in the source domain and 55,388 Real images in the target domain.

Baselines The baseline methods can be divided into two categories depending on whether the model has access to both source and target domain data during adaptation. Most of the previous unsupervised domain adaptation methods (*e.g.*, DAN (Long et al. 2015), DANN (Ganin and Lempitsky 2015), rRevGrad+CAT (Deng, Luo, and Zhu 2019), CDAN+BSP (Chen et al. 2019), CDAN+TransNorm (Wang et al. 2019), SAFN+ENT (Xu et al. 2019), MDD (Zhang et al. 2019)) require *source data access* during adaptation, whereas SHOT-IM (Liang, Hu, and Feng 2020) and SHOT (Liang, Hu, and Feng 2020) are unsupervised HTL methods without the *source data access* constraint. We also report *Source only*, which directly applies the source hypothesis to obtain the target predictions without any adaptation, and *MI ensemble*, which uses multiple hypotheses for MI maximization but without the HD regularization. In addition, we also report the results of two other regularization approaches, namely *MI ensemble* + L_2 and *MI ensemble* + L_2 *source*. Our HDMI with

Source	# of Hypotheses	Method	A→D	A→W	D→A	D→W	W→A	W→D	Avg.
✓	single	DAN (Long et al. 2015)	78.6	80.5	63.6	97.1	62.8	99.6	80.4
		DANN (Ganin and Lempitsky 2015)	79.7	82.0	68.2	96.9	67.4	99.1	82.2
		SAFN+ENT (Xu et al. 2019)	90.7	90.1	73.0	98.6	70.2	99.8	87.1
		rRevGrad+CAT (Deng, Luo, and Zhu 2019)	90.8	94.4	72.2	98.0	70.2	**100.**	87.6
		CDAN+BSP (Chen et al. 2019)	93.0	93.3	73.6	98.2	72.6	**100.**	88.5
		MDD (Zhang et al. 2019)	93.5	**94.5**	**74.6**	98.4	72.2	**100.**	88.9
✗	single	Source only	79.7	75.7	61.2	96.0	59.8	98.2	78.4
		MI maximization	90.2	92.3	73.0	96.5	73.1	95.0	86.7
		SHOT (Liang, Hu, and Feng 2020)	93.1	90.9	74.5	98.8	74.8	99.9	88.7
	multiple*	Source only	81.1	77.2	61.2	96.5	60.7	98.4	79.2
		MI ensemble	91.0	93.0	72.3	96.5	73.7	97.4	87.3
		MI ensemble + L_2	93.6	93.2	70.4	96.0	72.5	97.6	87.2
		MI ensemble + L_2 source	92.0	91.7	68.7	97.9	66.1	99.8	86.0
		HDMI (λ=0.5)	**94.4**	94.0	73.7	**98.9**	**75.9**	99.8	**89.5**
		HDMI ensemble (λ=0.5)	**94.4**	94.0	73.6	**98.9**	**75.9**	99.8	89.4

* Two hypotheses as an illustration. More examples are shown in Table 4.

Table 1: Target accuracy (%) on Office-31 with ResNet-50.

Method	Ar→Cl	Ar→Pr	Ar→Rw	Cl→Ar	Cl→Pr	Cl→Rw	Pr→Ar	Pr→Cl	Pr→Rw	Rw→Ar	Rw→Cl	Rw→Pr	Avg.
DAN (Long et al. 2015)	43.6	57.0	67.9	45.8	56.5	60.4	44.0	43.6	67.7	63.1	51.5	74.3	56.3
DANN (Ganin and Lempitsky 2015)	45.6	59.3	70.1	47.0	58.5	60.9	46.1	43.7	68.5	63.2	51.8	76.8	57.6
SAFN (Xu et al. 2019)	52.0	71.7	76.3	64.2	69.9	71.9	63.7	51.4	77.1	70.9	57.1	81.5	67.3
CDAN+TransNorm (Wang et al. 2019)	50.2	71.4	77.4	59.3	72.7	73.1	61.0	53.1	79.5	71.9	59.0	82.9	67.6
MDD (Zhang et al. 2019)	54.9	73.7	77.8	60.0	71.4	71.8	61.2	53.6	78.1	72.5	60.2	82.3	68.1
SHOT-IM (Liang, Hu, and Feng 2020)	52.8	72.9	78.4	65.4	73.8	74.1	64.6	50.8	78.9	72.7	53.5	81.2	68.3
SHOT (Liang, Hu, and Feng 2020)	56.9	**78.1**	81.0	**67.9**	78.4	78.1	**67.0**	54.6	81.8	73.4	58.1	**84.5**	71.6
Source only*	45.6	69.2	76.5	55.3	64.4	67.4	55.1	41.6	74.4	66.0	46.3	79.4	61.8
MI ensemble*	55.2	71.9	80.2	62.6	76.8	77.8	63.2	53.8	81.1	67.9	58.3	81.4	69.2
HDMI (λ=0.3)*	57.4	76.9	81.6	67.6	**79.1**	78.1	65.1	**56.0**	**82.5**	73.5	59.5	83.6	71.7
HDMI (λ=0.4)*	**57.8**	76.7	**81.9**	67.1	78.8	**78.8**	66.6	55.5	82.4	**73.6**	**59.7**	84.0	**71.9**

* Two hypotheses as an illustration.

Table 2: Target accuracy (%) on Office-Home with ResNet-50.

Source	Method	Avg. per-class accuracy
✓	JAN (Long et al. 2017)	61.6
	GTA(Sankaranarayanan et al. 2018)	69.5
	MCD (Saito et al. 2018)	71.9
	CDAN (Long et al. 2018)	70.0
	MDD (Zhang et al. 2019)	74.6
✗	SHOT-IM (Liang, Hu, and Feng 2020)	77.9
	SHOT (Liang, Hu, and Feng 2020)	79.6
	Source only	44.6
	MI ensemble (two hypotheses)	72.4
	HDMI (two hypotheses, λ=0.5)	**82.4**

Table 3: Target domain per-class average accuracy (%) on VisDA-C (Synthetic→Real) with ResNet-101.

independent classifiers (IC) from different random initialization is referred to as *HDMI-IC*, so as to be distinguished from that with MC-dropout sampled classifiers denoted as *HDMI-MC*.

Implementation Details We provide the details in the supplementary material. Note that we set the number of hypotheses $M = 2$ by default unless otherwise stated, since empirical results suggest that the HD regularization between two hypotheses suffices HDMI for better performance.

Results

State-of-the-Art Performance of HDMI We present the results of different methods on Office-31 in Table 1, Office-Home in Table 2, and VisDA-C in Table 3. The per-class accuracy for VisDA-C is detailed in Table 6 (supplementary material). As seen from all tables, the proposed HDMI achieves state-of-the-art performance on the target domains in all datasets, even outperforming the methods that have additional access to the source data during adaptation (methods for which "source" marked as ✓ in the tables). In unsupervised HTL setting (methods for which "source" marked as ✗ in the tables), HDMI also outperforms previous state-of-the-art methods SHOT-IM (also based on MI maximization) and SHOT (with an extra pseudo-label based self-training

# of Hypotheses (M)	Source only	MI maximization	HDMI					
			HDMI-IC					HDMI-MC
			$\lambda = 0.1$	$\lambda = 0.3$	$\lambda = 0.5$	$\lambda = 0.7$	$\lambda = 1.0$	$\lambda = 0.5$
2	81.1	91.0	92.2	94.2	94.4	93.6	94.2	93.6
3	81.5	91.6	93.2	94.0	95.2	95.6	95.6	92.2
4	82.3	92.8	93.6	94.4	95.0	96.4	94.6	91.0

Table 4: Robustness of HDMI (Target accuracy (%) on A→D, Office-31). HDMI is robust to the choices of the number of hypotheses used (M) and weight hyperparameter tuning (λ).

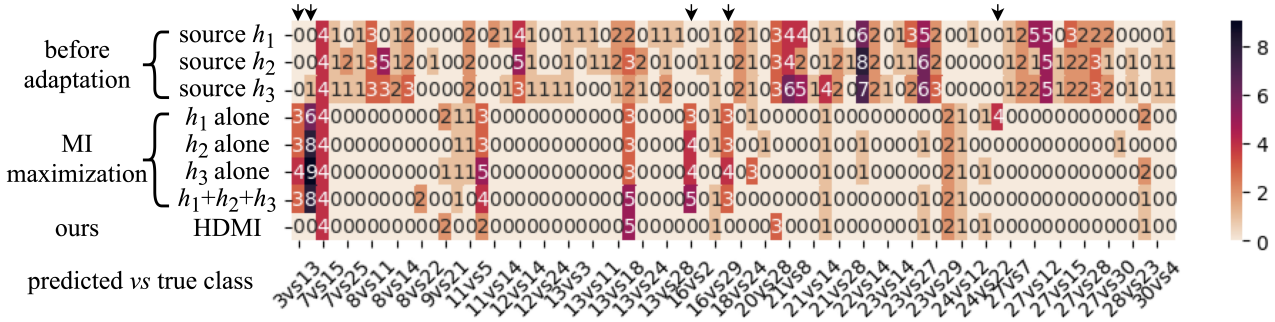


Figure 2: Target error analysis shows that HDMI (with three hypotheses) preserves more transferable source knowledge, as compared with using MI maximization alone (on A→D, Office-31).

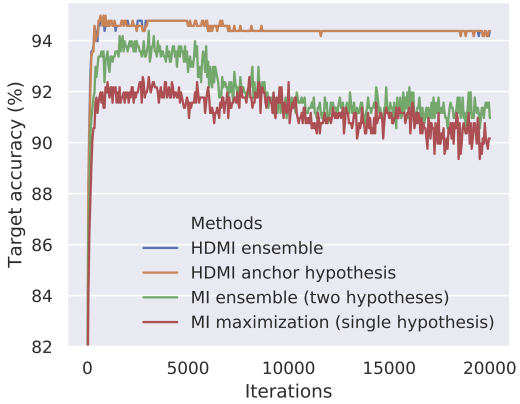


Figure 3: The hypothesis disparity regularization stabilizes the optimization for MI maximization (on A→D, Office-31).

strategy) (Liang, Hu, and Feng 2020). Compared with *MI ensemble*, adding the HD regularization effectively increases the target accuracy from 87.3% to 89.5% on Office-31, from 69.2% to 71.9% on Office-Home, and from 72.4% to 82.4% on VisDA-C. In addition, we also show in Table 1 that the proposed HD regularization in HDMI is superior to other forms of regularization such as those presented in *MI ensemble* + L_2 and *MI ensemble* + L_2 *source*.

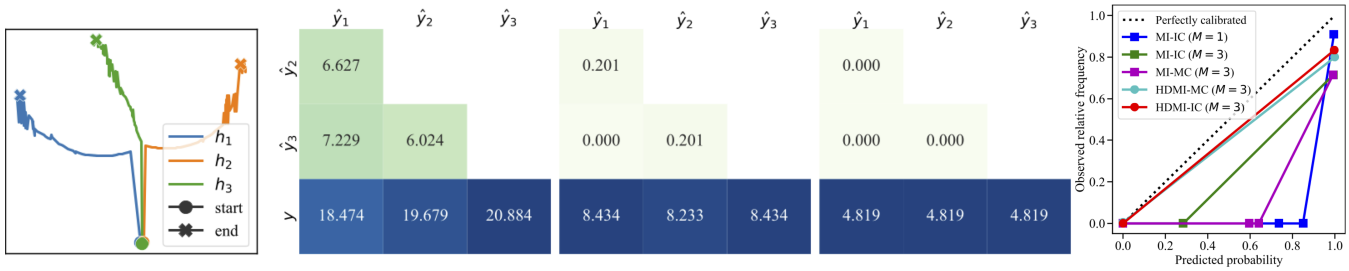
Robust Performance of HDMI To validate the robustness of HDMI in terms of the number of hypotheses M and the hyperparameter λ, we perform experiments on A→D, Office-31 with different configurations of M and λ, and summarize the results in Table 4. It is shown that HDMI consistently obtains improved performance over the MI maximization baseline without the HD regularization. More importantly, we show that using two hypotheses suffices HDMI for the improved performance. We also find the implementation of HDMI with independent classifiers (HDMI-IC) is preferable to that with MC-droupout (HDMI-MC) due to its ability to cover different modes in the hypothesis space.

Analyses

Here, we investigate how multiple hypotheses and HD regularization improves the MI maximization process.

The Hypothesis Disparity Regularizes MI Maximization Figure 3 shows the learning curves of the target accuracy for different mutual information based approaches with or without the HD regularization. As shown in the figure, the target performance degrades in "MI maximization (single hypothesis)" due to the lack of proper regularization. Furthermore, we find that the use of deep ensemble in "MI ensemble (two hypotheses)" does not help alleviate this performance degradation problem. This necessitates the proposal of our HD regularized MI maximization, where the transfer process is stable and effective.

HDMI Facilitates the Positive Transfer of Multiple Modes from the Source Hypotheses Figure 2 summarizes the fine-grained hypothesis-level prediction errors made by different approaches. The figure reveals that direct MI maximization (row 4-7) suffers from negative transfer and introduces new errors that were not present in the *Source only* models (row 1-3) before adaptation, *e.g.*, columns with arrows, indicating partial lost of the transferable source knowledge during adaptation. In contrast, HDMI (row 8) facilitates



(a) Source training trajectories (b) Source only (c) MI ensemble (d) HDMI (e) Reliability diagram

Figure 4: (a) t-SNE visualization comparing the trajectory of target predictions from different source hypotheses. We follow the same plotting procedure as in (Fort, Hu, and Lakshminarayanan 2019). (b-d) Disagreements between predictions from different hypotheses (%), where $\hat{y}_i$ denotes the predictions of h_i and y denotes the ground-truth labels. (e) Reliability diagram of the target domain after transfer (class 11 as the positive class). All figures are on A→D, Office-31.

positive transfer of different modes learned from the source domain, shown in Figure 4 (a), and results in stable and effective target adaptation. As a result, the HD regularization prevents negative transfer from the MI maximization and facilitates the positive transfer of multiple modes from source hypotheses.

HDMI Maximally Reduces the Disagreement Among Target Hypotheses Figure 4 (b)-(d) compare the disagreement among predictions from different target hypotheses and ground-truth, where HDMI (Figure 4 (d)) is shown to maximally reduce the disagreement compared with *Source only* (Figure 4 (b)) and *MI ensemble* (Figure 4 (c)), demonstrating the effectiveness of the HD regularization in bringing target hypotheses to align with each other. We have similar findings on the KL divergence of example-level predictions between target hypotheses (Figure 5, supplementary material).

HDMI Presents Well-Calibrated Predictive Uncertainty Uncertainty calibration is especially important for the performance of hypothesis transfer between different source and target domains where better calibrated probabilities lead to more effective hypothesis transfer. To investigate whether HDMI benefits from uncertainty calibration, we plot the reliability diagram (DeGroot and Fienberg 1983; Niculescu-Mizil and Caruana 2005) of different approaches and confirm that HDMI is better calibrated than other approaches (Figure 4 (e)). In consistent with (Lakshminarayanan, Pritzel, and Blundell 2017), we also find that multiple hypotheses using independent classifiers (IC) is superior to that with MC-dropout sampled classifiers (MC) in both cases of *MI ensemble* and HDMI. Quantitative analysis of the uncertainty calibration confirms that HDMI has the best Brier score and the expected calibration error (ECE) score (Naeini, Cooper, and Hauskrecht 2015) (Table 7, supplementary material).

Ablation Study

We summarize the results of ablation studies in Table 5. We first evaluate the impact of shared feature extractor ψ among target hypotheses by comparing "MI ensemble (independent ψ)" with "HDMI (independent ψ)". We find that the HD regularization does not help MI maximization if the feature extractors are independent, suggesting that HD regularization works through learning better representations shared by different target hypotheses. In addition, we also find MI maximization performs better than conditional entropy minimization in unsupervised HTL, similar to the finding in discriminative clustering (Krause, Perona, and Gomes 2010). Lastly, we show cross entropy measure surrogates the proposed hypothesis disparity better than KL divergence. The detailed results are provided in the supplementary material (Table 8, Table 9 and Table 11).

Method*	Target avg. accuracy (%)
Source only	79.2
MI ensemble	87.3
HDMI	**89.5**
HDMI with KL	88.6
MI ensemble (independent ψ)	87.7
HDMI (independent ψ)	87.5
HD only	84.8
Conditional Entropy + HD	85.7
MI ensemble + L_2	87.2
MI ensemble + L_2 source	86.0

* With two hypotheses, λ=0.5.

Table 5: Ablation study (on Office-31).

Conclusion

In this paper, we tackle the problem of unsupervised hypothesis transfer learning to bridge the gap between unsupervised domain adaptation and hypothesis transfer learning. We propose a hypothesis disparity regularized mutual information maximization approach that not only employs multiple source and target hypotheses but also utilizes the relationship among different hypotheses to overcome the limitation of mutual information maximization with a single source and target hypothesis. Empirical results demonstrate that the proposed hypothesis disparity regularization minimizes undesirable disagreements among hypotheses and preserves more transferable knowledge from the source domain. Our approach achieves state-of-the-art performance on three benchmark datasets for unsupervised domain adaptation in the context of hypothesis transfer learning.

References

Bachman, P.; Hjelm, R. D.; and Buchwalter, W. 2019. Learning representations by maximizing mutual information across views. In *NeurIPS*, 15509–15519.

Ben-David, S.; and Urner, R. 2013. Domain adaptation as learning with auxiliary information. In *New directions in transfer and multi-task-workshop@ NIPS*.

Bridle, J. S.; Heading, A. J.; and MacKay, D. J. 1992. Unsupervised Classifiers, Mutual Information and'Phantom Targets. In *NeurIPS*, 1096–1101.

Chen, X.; Wang, S.; Long, M.; and Wang, J. 2019. Transferability vs. discriminability: Batch spectral penalization for adversarial domain adaptation. In *ICML*, 1081–1090.

Chidlovskii, B.; Clinchant, S.; and Csurka, G. 2016. Domain adaptation in the absence of source domain data. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 451–460.

DeGroot, M. H.; and Fienberg, S. E. 1983. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)* 32(1-2): 12–22.

Deng, Z.; Luo, Y.; and Zhu, J. 2019. Cluster Alignment with a Teacher for Unsupervised Domain Adaptation. In *ICCV*, 9944–9953.

Du, S. S.; Koushik, J.; Singh, A.; and Póczos, B. 2017. Hypothesis transfer learning via transformation functions. In *NeurIPS*, 574–584.

Duan, L.; Tsang, I. W.; Xu, D.; and Chua, T.-S. 2009. Domain adaptation from multiple sources via auxiliary classifiers. In *ICML*, 289–296.

Fei-Fei, L.; Fergus, R.; and Perona, P. 2006. One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence* 28(4): 594–611.

Fernandes, K.; and Cardoso, J. S. 2019. Hypothesis transfer learning based on structural model similarity. *Neural Computing and Applications* 31(8): 3417–3430.

Fort, S.; Hu, H.; and Lakshminarayanan, B. 2019. Deep Ensembles: A Loss Landscape Perspective. *arXiv:1912.02757* .

Gal, Y.; and Ghahramani, Z. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, 1050–1059.

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

Grandvalet, Y.; and Bengio, Y. 2005. Semi-supervised learning by entropy minimization. In *NeurIPS*, 529–536.

Hjelm, R. D.; Fedorov, A.; Lavoie-Marchildon, S.; Grewal, K.; Bachman, P.; Trischler, A.; and Bengio, Y. 2018. Learning deep representations by mutual information estimation and maximization. *arXiv:1808.06670* .

Hoffman, J.; Tzeng, E.; Park, T.; Zhu, J.-Y.; Isola, P.; Saenko, K.; Efros, A. A.; and Darrell, T. 2018. Cycada: Cycle-consistent adversarial domain adaptation. In *ICML*.

Hu, W.; Miyato, T.; Tokui, S.; Matsumoto, E.; and Sugiyama, M. 2017. Learning discrete representations via information maximizing self-augmented training. In *ICML*, 1558–1567. JMLR. org.

Jie, L.; Tommasi, T.; and Caputo, B. 2011. Multiclass transfer learning from unconstrained priors. In *ICCV*, 1863–1870. IEEE.

Krause, A.; Perona, P.; and Gomes, R. G. 2010. Discriminative clustering by regularized information maximization. In *NeurIPS*, 775–783.

Kumar, A.; Sattigeri, P.; Wadhawan, K.; Karlinsky, L.; Feris, R.; Freeman, B.; and Wornell, G. 2018. Co-regularized alignment for unsupervised domain adaptation. In *NeurIPS*, 9345–9356.

Kuzborskij, I.; and Orabona, F. 2013. Stability and hypothesis transfer learning. In *ICML*, 942–950.

Kuzborskij, I.; and Orabona, F. 2017. Fast rates by transferring from auxiliary hypotheses. *Machine Learning* 106(2): 171–195.

Lakshminarayanan, B.; Pritzel, A.; and Blundell, C. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *NeurIPS*, 6402–6413.

Liang, J.; He, R.; Sun, Z.; and Tan, T. 2019. Distant supervised centroid shift: A simple and efficient approach to visual domain adaptation. In *CVPR*, 2975–2984.

Liang, J.; Hu, D.; and Feng, J. 2020. Do We Really Need to Access the Source Data? Source Hypothesis Transfer for Unsupervised Domain Adaptation. *arXiv:2002.08546* .

Long, M.; Cao, Y.; Wang, J.; and Jordan, M. I. 2015. Learning transferable features with deep adaptation networks. *arXiv:1502.02791* .

Long, M.; Cao, Z.; Wang, J.; and Jordan, M. I. 2018. Conditional adversarial domain adaptation. In *NeurIPS*, 1640–1650.

Long, M.; Zhu, H.; Wang, J.; and Jordan, M. I. 2017. Deep transfer learning with joint adaptation networks. In *ICML*, 2208–2217. JMLR. org.

Mansour, Y.; Mohri, M.; and Rostamizadeh, A. 2009. Domain adaptation with multiple sources. In *NeurIPS*, 1041–1048.

Miayto, T.; Dai, A. M.; and Goodfellow, I. 2017. Virtual Adversarial Training for Semi-Supervised Text Classification. In *ICLR*.

Morerio, P.; Cavazza, J.; and Murino, V. 2017. Minimal-entropy correlation alignment for unsupervised deep domain adaptation. *arXiv:1711.10288* .

Naeini, M. P.; Cooper, G.; and Hauskrecht, M. 2015. Obtaining well calibrated probabilities using bayesian binning. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Niculescu-Mizil, A.; and Caruana, R. 2005. Predicting good probabilities with supervised learning. In *ICML*, 625–632.

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

Orabona, F.; Castellini, C.; Caputo, B.; Fiorilla, A. E.; and Sandini, G. 2009. Model adaptation with least-squares SVM for adaptive hand prosthetics. In *2009 IEEE International Conference on Robotics and Automation*, 2897–2903. IEEE.

Pan, S. J.; and Yang, Q. 2009. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering* 22(10): 1345–1359.

Panareda Busto, P.; and Gall, J. 2017. Open set domain adaptation. In *ICCV*, 754–763.

Peng, X.; Bai, Q.; Xia, X.; Huang, Z.; Saenko, K.; and Wang, B. 2019. Moment matching for multi-source domain adaptation. In *ICCV*, 1406–1415.

Peng, X.; Usman, B.; Kaushik, N.; Wang, D.; Hoffman, J.; and Saenko, K. 2018. Visda: A synthetic-to-real benchmark for visual domain adaptation. In *CVPR Workshops*, 2021–2026.

Perrot, M.; and Habrard, A. 2015. A theoretical analysis of metric hypothesis transfer learning. In *ICML*, 1708–1717.

Ruder, S. 2017. An overview of multi-task learning in deep neural networks. *arXiv:1706.05098* .

Saenko, K.; Kulis, B.; Fritz, M.; and Darrell, T. 2010. Adapting visual category models to new domains. In *European conference on computer vision*, 213–226. Springer.

Saito, K.; Kim, D.; Sclaroff, S.; Darrell, T.; and Saenko, K. 2019. Semi-supervised domain adaptation via minimax entropy. In *ICCV*, 8050–8058.

Saito, K.; Watanabe, K.; Ushiku, Y.; and Harada, T. 2018. Maximum classifier discrepancy for unsupervised domain adaptation. In *CVPR*, 3723–3732.

Sankaranarayanan, S.; Balaji, Y.; Castillo, C. D.; and Chellappa, R. 2018. Generate to adapt: Aligning domains using generative adversarial networks. In *CVPR*, 8503–8512.

Shu, R.; Bui, H. H.; Narui, H.; and Ermon, S. 2018. A dirt-t approach to unsupervised domain adaptation. In *ICLR*.

Tian, Y.; Krishnan, D.; and Isola, P. 2019. Contrastive multiview coding. *arXiv:1906.05849* .

Tommasi, T.; Orabona, F.; and Caputo, B. 2013. Learning categories from few examples with multi model knowledge transfer. *IEEE transactions on pattern analysis and machine intelligence* 36(5): 928–941.

Tramèr, F.; Kurakin, A.; Papernot, N.; Goodfellow, I.; Boneh, D.; and McDaniel, P. 2017. Ensemble adversarial training: Attacks and defenses. *arXiv:1705.07204* .

Tzeng, E.; Hoffman, J.; Saenko, K.; and Darrell, T. 2017. Adversarial discriminative domain adaptation. In *CVPR*, 7167–7176.

Venkateswara, H.; Eusebio, J.; Chakraborty, S.; and Panchanathan, S. 2017. Deep hashing network for unsupervised domain adaptation. In *CVPR*, 5018–5027.

Wang, X.; Jin, Y.; Long, M.; Wang, J.; and Jordan, M. I. 2019. Transferable Normalization: Towards Improving Transferability of Deep Neural Networks. In *NeurIPS*, 1951–1961.

Xu, R.; Li, G.; Yang, J.; and Lin, L. 2019. Larger Norm More Transferable: An Adaptive Feature Norm Approach for Unsupervised Domain Adaptation. In *ICCV*, 1426–1435.

Yang, J.; Yan, R.; and Hauptmann, A. G. 2007. Cross-domain video concept detection using adaptive svms. In *Proceedings of the 15th ACM international conference on Multimedia*, 188–197.

Zhang, Y.; Liu, T.; Long, M.; and Jordan, M. 2019. Bridging Theory and Algorithm for Domain Adaptation. In Chaudhuri, K.; and Salakhutdinov, R., eds., *ICML*, volume 97 of *Proceedings of Machine Learning Research*, 7404–7413. Long Beach, California, USA: PMLR.

Zhuang, F.; Qi, Z.; Duan, K.; Xi, D.; Zhu, Y.; Zhu, H.; Xiong, H.; and He, Q. 2019. A Comprehensive Survey on Transfer Learning. *arXiv:1911.02685* .