Uncovering the Connections Between Adversarial Transferability and Knowledge Transferability

Kaizhao Liang [*1] Jacky Y. Zhang [*1] Boxin Wang [1] Zhuolin Yang [1] Oluwasanmi Koyejo [1] Bo Li [1]

Abstract

Knowledge transferability, or transfer learning, has been widely adopted to allow a pre-trained model in the source domain to be effectively adapted to downstream tasks in the target domain. It is thus important to explore and understand the factors affecting knowledge transferability. In this paper, as the first work, we analyze and demonstrate the connections between knowledge transferability and another important phenomenon–adversarial transferability, *i.e.*, adversarial examples generated against one model can be transferred to attack other models. Our theoretical studies show that adversarial transferability indicates knowledge transferability, and vice versa. Moreover, based on the theoretical insights, we propose two practical adversarial transferability metrics to characterize this process, serving as bidirectional indicators between adversarial and knowledge transferability. We conduct extensive experiments for different scenarios on diverse datasets, showing a positive correlation between adversarial transferability and knowledge transferability. Our findings will shed light on future research about effective knowledge transfer learning and adversarial transferability analyses. All code and data are available here.

1. Introduction

Knowledge transfer is quickly becoming the standard approach for fast learning adaptation across domains. Also known as transfer learning or learning transfer, knowledge transfer has been a critical technology for enabling several real-world applications, including object detection (Zhang et al., 2014), image segmentation (Kendall et al., 2018), multi-lingual machine translation (Dong et al., 2015), and language understanding evaluation (Wang et al., 2019a), among others. For example, since the release of ImageNet (Russakovsky et al., 2015), pretrained ImageNet models (e.g., on TensorFlow Hub or PyTorch-Hub) have become the default option for the knowledge transfer source due to its broad coverage of visual concepts and compatibility with various visual tasks (Huh et al., 2016). Motivated by its importance, many studies have explored the factors associated with knowledge transferability. Most recently, Salman et al. (2020) showed that more robust pretrained ImageNet models transfer better to downstream tasks, which reveals that *adversarial training* helps to improve knowledge transferability.

In the meantime, *adversarial transferability* has been extensively studied—a phenomenon that an adversarial instance generated against one model has high probability attack another one without additional modification (Papernot et al., 2016; Goodfellow et al., 2014; Joon Oh et al., 2017). Hence, adversarial transferability is widely exploited in black-box attacks (Ilyas et al., 2018; Liu et al., 2016; Naseer et al., 2019). A line of work has been conducted to bound the adversarial transferability based on model (gradient) similarity (Tramèr et al., 2017b). Given that both *adversarial transferability* and *knowledge transferability* are impacted by certain model similarity and adversarial ML properties, in this work, we aim to conduct the *first* study to analyze the connections between them and ask,

> *What is the fundamental connection between knowledge transferability and adversarial transferability? Can we measure one and indicate the other?*

Technical Contributions. In this paper, we take the *first* step towards exploring the fundamental relation between adversarial transferability and knowledge transferability. We make contributions on both theoretical and empirical fronts.

- We formally define the adversarial transferability for the *first* time by considering all potential adversarial perturbation vectors. We then conduct thorough and novel theoretical analysis to characterize the precise connection between adversarial transferability and knowledge transferability based on our definition.

[*]Equal contribution [1]Department of Computer Science, the University of Illinois at Urbana-Champaign, Urbana, USA. Correspondence to: Oluwasanmi Koyejo <sanmi@illinois.edu>, Bo Li <lbo@illinois.edu>.

Proceedings of the 38th International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

- In particular, we prove that high adversarial transferability will indicate high knowledge transferability, which can be represented as the distance in an inner product space defined by the Hessian of the adversarial loss. In the meantime, we prove that high knowledge transferability will indicate high adversarial transferability.
- Based on our theoretical insights, we propose two practical adversarial transferability metrics that quantitatively measure the adversarial transferability in practice. We then provide simulational results to verify how these metrics connect with the knowledge transferability in a bidirectional way.
- Extensive experiments justify our theoretical insights and the proposed adversarial transferability metrics, leading to our discussion on potential applications and future research.

Related Work There is a line of research studying different factors that affect knowledge transferability (Yosinski et al., 2014; Long et al., 2015; Wang et al., 2019b; Xu et al., 2019; Shinya et al., 2019). Further, empirical observations show that the correlation between learning tasks (Achille et al., 2019; Zamir et al., 2018), the similarity of model architectures, and data distribution are all correlated with different knowledge transfer abilities. Interestingly, recent empirical evidence suggests that adversarially-trained models transfer better than non-robust models (Salman et al., 2020; Utrera et al., 2020), suggesting a connection between the adversarial properties and knowledge transferability. On the other hand, several approaches have been proposed to boost the adversarial transferability (Zhou et al., 2018; Demontis et al., 2019; Dong et al., 2019; Xie et al., 2019). Beyond the above empirical studies, there are a few existing analyses of adversarial transferability, which explore different conditions that may enhance adversarial transferability (Athalye et al., 2018; Tramèr et al., 2017b; Ma et al., 2018; Demontis et al., 2019). In this work, we aim to bridge the connection between adversarial and knowledge transferability, both of which reveal interesting properties of ML model similarities from different perspectives.

2. Adversarial Transferability and Knowledge Transferability

This section introduces the preliminaries and the formal definitions of the knowledge and adversarial transferability, and formally defines our problem of interest.

Notation. Sets are denoted in blackboard bold, e.g., $\mathbb{R}$, and the set of integers $\{1 \dots n\}$ is denoted as $[n]$. Distributions are denoted in calligraphy, e.g., $\mathcal{D}$, and the support of a distribution $\mathcal{D}$ is denoted as $\text{supp}(\mathcal{D})$. Vectors are denoted as bold lower case letters, e.g., $\boldsymbol{x} \in \mathbb{R}^n$, and matrices are denoted as bold uppercase letters, e.g., $\boldsymbol{W}$. We denote the entry-wise product operator between vectors or matrices as $\odot$. The Moore–Penrose inverse of a matrix $\boldsymbol{W}$ is denoted as $\boldsymbol{W}^\dagger$. We use $\|\cdot\|_2$ to denote Euclidean norm induced by Euclidean inner product $\langle \cdot, \cdot \rangle$. The standard inner product of two matrices is defined as $\langle \boldsymbol{W}, \boldsymbol{M} \rangle = \text{tr}(\boldsymbol{W}^\top \boldsymbol{M})$, where $\text{tr}(\cdot)$ is the trace of a matrix. The Frobenius norm $\|\cdot\|_F$ is induced by the standard matrix inner product. Moreover, in the (semi-)inner product space defined by a positive (semi-)definite matrix $\boldsymbol{S}$, the (semi-)inner product of two vectors or matrices is defined by $\langle \boldsymbol{v}_1, \boldsymbol{v}_2 \rangle_{\boldsymbol{S}} = \boldsymbol{v}_1^\top \boldsymbol{S} \boldsymbol{v}_2$ or $\langle \boldsymbol{W}, \boldsymbol{M} \rangle_{\boldsymbol{S}} = \text{tr}(\boldsymbol{W}^\top \boldsymbol{S} \boldsymbol{M})$, respectively. Given a vector $\boldsymbol{v}$, we define its normalization as $\widehat{\boldsymbol{v}} = \boldsymbol{v} / \|\boldsymbol{v}\|_2$. When using a denominator $\|\cdot\|_*$ other than Euclidean norm, we denote the normalization as $\widehat{\boldsymbol{v}}|_*$.

Given a (vector-valued) function f, we denote $f(\boldsymbol{x})$ as its evaluated value at $\boldsymbol{x}$, and f represents the function itself in the corresponding Hilbert space. Composition of functions is denoted as $g \circ f(\boldsymbol{x}) = g(f(\boldsymbol{x}))$. We use $\langle \cdot, \cdot \rangle_{\mathcal{D}}$ to denote the inner product induced by distribution $\mathcal{D}$ and inherited from Euclidean inner product, i.e., $\langle f_1, f_2 \rangle_{\mathcal{D}} = \mathbb{E}_{\boldsymbol{x} \sim \mathcal{D}} \langle f_1(\boldsymbol{x}), f_2(\boldsymbol{x}) \rangle$. Accordingly, we use $\|\cdot\|_{\mathcal{D}}$ to denote the norm induced by the inner product $\langle \cdot, \cdot \rangle_{\mathcal{D}}$, i.e., $\|f\|_{\mathcal{D}} = \sqrt{\langle f, f \rangle_{\mathcal{D}}}$. When the inherited inner product is defined by $\boldsymbol{S}$, we denote $\langle f_1, f_2 \rangle_{\mathcal{D}, \boldsymbol{S}} = \mathbb{E}_{\boldsymbol{x} \sim \mathcal{D}} \langle f_1(\boldsymbol{x}), f_2(\boldsymbol{x}) \rangle_{\boldsymbol{S}}$, and similarly for $\|f\|_{\mathcal{D}, \boldsymbol{S}}$.

Knowledge Transferability Given a pre-trained *source* model $f_S : \mathbb{R}^n \to \mathbb{R}^m$ and a *target* domain $\boldsymbol{x} \in \mathbb{R}^n$ with data distribution $\boldsymbol{x} \sim \mathcal{D}$ and *target* labels $y(\boldsymbol{x}) \in \mathbb{R}^d$, *knowledge transferability* is defined as the performance of fine-tuning f_S on $\mathcal{D}$ to predict y. Concretely, knowledge transferability can be represented as a loss $\mathcal{L}(\,\cdot\,, y, \mathcal{D})$ after fine-tuning by composing the fixed source model with a trainable function $g : \mathbb{R}^m \to \mathbb{R}^d$, typically from a small function class $g \in \mathbb{G}$, *i.e.*,

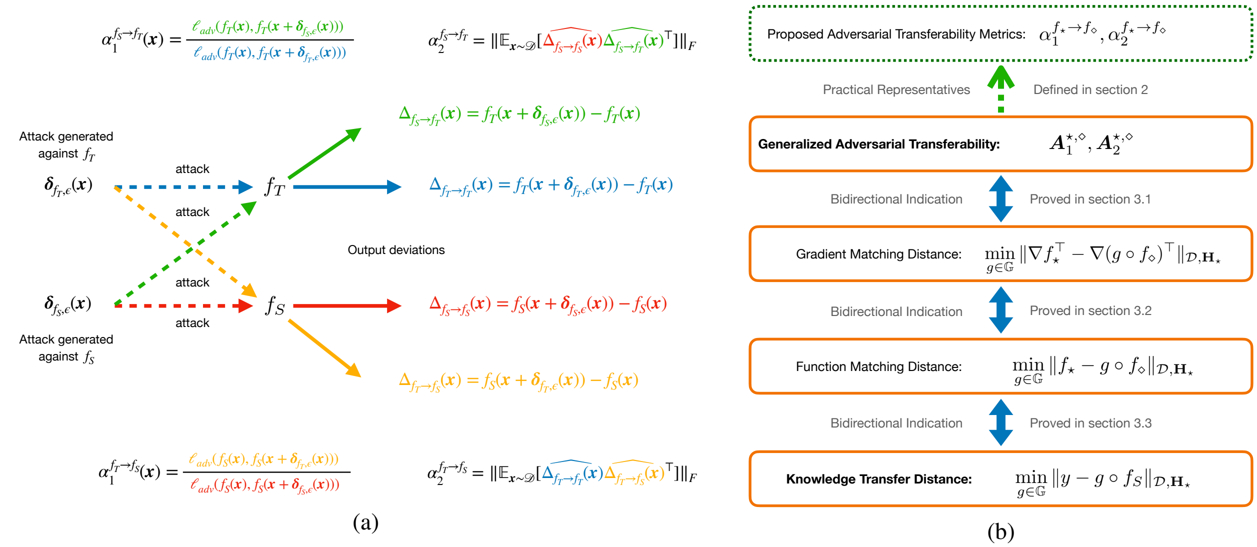
$$\min_{g \in \mathbb{G}} \quad \mathcal{L}(g \circ f_S, y, \mathcal{D}), \tag{1}$$

where the loss function $\mathcal{L}$ measures the error between $g \circ f_S$ and the ground truth y under the *target* data distribution $\mathcal{D}$. For example, for neural networks it is usual to stack on and fine-tune a linear layer; here $\mathbb{G}$ is the affine function class. We will focus on the affine setting in this paper.

For our purposes, a more useful measure of transfer is to compare the quality of the fine-tuned model to a model trained directly on the target domain $f_T : \mathbb{R}^n \to \mathbb{R}^d$. Thus, we study the following surrogate of knowledge transferability, where the ground truth target is replaced by a reference target model f_T:

$$\min_{g \in \mathbb{G}} \quad \mathcal{L}(g \circ f_S, f_T, \mathcal{D}). \tag{2}$$

Adversarial Attacks. For simplicity we consider untargeted attacks that seeks to maximize the deviation of model



$\alpha_1^{f_S\to f_T}(x) = \frac{\ell_{adv}(f_T(x), f_T(x+\delta_{f_S,\epsilon}(x)))}{\ell_{adv}(f_T(x), f_T(x+\delta_{f_T,\epsilon}(x)))}$
$\alpha_2^{f_S\to f_T} = \|\mathbb{E}_{x\sim\mathcal{D}}[\widehat{\Delta_{f_S\to f_S}(x)}\widehat{\Delta_{f_S\to f_T}(x)}^\top]\|_F$
Attack generated against f_T
$\delta_{f_T,\epsilon}(x)$
attack
attack
f_T
$\Delta_{f_S\to f_T}(x) = f_T(x+\delta_{f_S,\epsilon}(x)) - f_T(x)$
$\Delta_{f_T\to f_T}(x) = f_T(x+\delta_{f_T,\epsilon}(x)) - f_T(x)$
Output deviations
attack
attack
$\delta_{f_S,\epsilon}(x)$
Attack generated against f_S
f_S
$\Delta_{f_S\to f_S}(x) = f_S(x+\delta_{f_S,\epsilon}(x)) - f_S(x)$
$\Delta_{f_T\to f_S}(x) = f_S(x+\delta_{f_T,\epsilon}(x)) - f_S(x)$
$\alpha_1^{f_T\to f_S}(x) = \frac{\ell_{adv}(f_S(x), f_S(x+\delta_{f_T,\epsilon}(x)))}{\ell_{adv}(f_S(x), f_S(x+\delta_{f_S,\epsilon}(x)))}$
$\alpha_2^{f_T\to f_S} = \|\mathbb{E}_{x\sim\mathcal{D}}[\widehat{\Delta_{f_T\to f_T}(x)}\widehat{\Delta_{f_T\to f_S}(x)}^\top]\|_F$
(a)
Proposed Adversarial Transferability Metrics: $\alpha_1^{f_\star\to f_\diamond}, \alpha_2^{f_\star\to f_\diamond}$
Practical Representatives
Defined in section 2
Generalized Adversarial Transferability: $A_1^{\star,\diamond}, A_2^{\star,\diamond}$
Bidirectional Indication
Proved in section 3.1
Gradient Matching Distance: $\min_{g\in\mathbb{G}} \|\nabla f_\star^\top - \nabla(g\circ f_\diamond)^\top\|_{\mathcal{D},\mathbf{H}_\star}$
Bidirectional Indication
Proved in section 3.2
Function Matching Distance: $\min_{g\in\mathbb{G}} \|f_\star - g\circ f_\diamond\|_{\mathcal{D},\mathbf{H}_\star}$
Bidirectional Indication
Proved in section 3.3
Knowledge Transfer Distance: $\min_{g\in\mathbb{G}} \|y - g\circ f_S\|_{\mathcal{D},\mathbf{H}_\star}$
(b)

Figure 1. (a) An illustration of the two proposed adversarial transferability metrics α_1, α_2 under different adversarial transferability settings, *i.e.*, $\alpha_1^{f_S\to f_T}, \alpha_1^{f_T\to f_S}, \alpha_2^{f_S\to f_T}$, and $\alpha_2^{f_T\to f_S}$. (b) An overview of the theoretical analysis framework, and its practical inspirations, where $\star, \diamond \in \{T, S\}$ and $\star \neq \diamond$. The three blue double-headed arrows are the bidirectional indication relationships proved in our theory section, and the dashed green arrow shows in practice how the two proposed adversarial transferability metrics are measured as representatives of the generalized adversarial transferability based on our theory.

output as measured by a given adversarial loss function $\ell_{adv}(\cdot,\cdot)$. The targeted attack can be viewed as a special case. Without loss of generality, we assume the adversarial loss is *non-negative*. Given a datapoint $\boldsymbol{x}$ and model f, an adversarial example of magnitude ϵ is denoted by $\boldsymbol{\delta}_{f,\epsilon}(\boldsymbol{x})$, computed as:

$$\boldsymbol{\delta}_{f,\epsilon}(\boldsymbol{x}) = \arg\max_{\|\boldsymbol{\delta}\|\le\epsilon} \ell_{adv}(f(\boldsymbol{x}), f(\boldsymbol{x}+\boldsymbol{\delta})). \tag{3}$$

We note that in theory $\boldsymbol{\delta}_{f,\epsilon}(\boldsymbol{x})$ may not be unique, and its generalized definition and its discussion are provided in our theoretical analysis (Section 3).

Adversarial Transferability. The process of adversarial transfer involves applying the adversarial example generated against a model f_1 to another model f_2. Thus, adversarial transferability from f_1 to f_2 measures how well $\boldsymbol{\delta}_{f_1,\epsilon}$ attacks f_2. We propose two metrics, namely, α_1 and α_2 that characterize adversarial transferability from complementary perspectives. To provide a visual overview of our definitions for the proposed adversarial transferability metrics, we present an illustration in Figure 1 (a).

Definition 1 (The First Adversarial Transferability)**.** *The first adversarial transferability from f_1 to f_2 at data sample $\boldsymbol{x}\sim\mathcal{D}$, is defined as*

$$\alpha_1^{f_1\to f_2}(\boldsymbol{x}) = \frac{\ell_{adv}(f_2(\boldsymbol{x}), f_2(\boldsymbol{x}+\boldsymbol{\delta}_{f_1,\epsilon}(\boldsymbol{x})))}{\ell_{adv}(f_2(\boldsymbol{x}), f_2(\boldsymbol{x}+\boldsymbol{\delta}_{f_2,\epsilon}(\boldsymbol{x})))}.$$

Taking the expectation, the first adversarial transferability is defined as

$$\alpha_1^{f_1\to f_2} = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\alpha^{f_1\to f_2}(\boldsymbol{x})\right].$$

Observe that the first adversarial transferability characterize how well the adversarial attacks $\boldsymbol{\delta}_{f_1,\epsilon}$ generated against f_1 perform on f_2, compared to f_2's whitebox adversarial attacks $\boldsymbol{\delta}_{f_2,\epsilon}$. Thus, high α_1 indicates high adversarial transferability. Note that the two attacks use the same magnitude constraint ϵ.

Recall that $\ell_{adv}(f(\boldsymbol{x}), f(\boldsymbol{x}+\boldsymbol{\delta}))$ measures the effect of the attack $\boldsymbol{\delta}$ on the model output $f(\boldsymbol{x})$. α_1 characterizes the relative magnitude of this deviation. However, this magnitude information is incomplete, as the direction of the deviation also encodes information about the adversarial transfer process. To this end, we propose the second adverserial metric, inspired by our theoretical analysis, which characterizes adversarial transferability from the directional perspective.

Definition 2 (The Second Adversarial Transferability)**.** *The second adversarial transferability from f_1 to f_2, under data distribution $\boldsymbol{x}\sim\mathcal{D}$, is defined as*

$$\alpha_2^{f_1\to f_2} = \|\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}[\widehat{\Delta_{f_1\to f_1}(\boldsymbol{x})}\widehat{\Delta_{f_1\to f_2}(\boldsymbol{x})}^\top]\|_F,$$

where

$$\Delta_{f_1\to f_1}(\boldsymbol{x}) = f_1(\boldsymbol{x}+\boldsymbol{\delta}_{f_1,\epsilon}(\boldsymbol{x})) - f_1(\boldsymbol{x})$$
$$\Delta_{f_1\to f_2}(\boldsymbol{x}) = f_2(\boldsymbol{x}+\boldsymbol{\delta}_{f_1,\epsilon}(\boldsymbol{x})) - f_2(\boldsymbol{x})$$

are deviations in model output given the adversarial attack $\boldsymbol{\delta}_{f_1,\epsilon}(\boldsymbol{x})$ generated against f_1, and $\widehat{\cdot}$ denotes the corresponding unit-length vector.

To further clarify the second adversarial transferability metric, consider the following alternative form of α_2.

Proposition 2.1. *The $\alpha_2^{f_1\to f_2}$ can be reformulated as*

$$(\alpha_2^{f_1\to f_2})^2 = \mathbb{E}_{\boldsymbol{x}_1,\boldsymbol{x}_2}\left[\theta_{f_1\to f_1}(\boldsymbol{x}_1,\boldsymbol{x}_2)\theta_{f_1\to f_2}(\boldsymbol{x}_1,\boldsymbol{x}_2)\right],$$

where $\boldsymbol{x}_1,\boldsymbol{x}_2 \overset{i.i.d.}{\sim} \mathcal{D}$, and

$$\theta_{f_1\to f_1}(\boldsymbol{x}_1,\boldsymbol{x}_2) = \langle \widehat{\Delta_{f_1\to f_1}(\boldsymbol{x}_1)}, \widehat{\Delta_{f_1\to f_1}(\boldsymbol{x}_2)}\rangle$$
$$\theta_{f_1\to f_2}(\boldsymbol{x}_1,\boldsymbol{x}_2) = \langle \widehat{\Delta_{f_1\to f_2}(\boldsymbol{x}_1)}, \widehat{\Delta_{f_1\to f_2}(\boldsymbol{x}_2)}\rangle$$

We can see that high α_2 indicates that it is more likely for the two inner products (*i.e.*, $\theta_{f_1\to f_1}$ and $\theta_{f_1\to f_2}$) to have the same sign. Given that the direction of f_1's output deviation indicates its attack $\boldsymbol{\delta}_{f_1,\epsilon}$, and the direction of f_2's output deviation indicates the transferred attack $\boldsymbol{\delta}_{f_1,\epsilon}$, high α_2 implies that the two directions will rotate by a similar angle as the data changes.

α_1 and α_2 represent complementary aspects of the adversarial transferability: α_1 can be understood as how often the adversarial attack transfers, while α_2 encodes directional information of the output deviation caused by adversarial attacks. An example is provided in the appendix section A to illustrate the necessity of both the metrics in characterizing the relation between adversarial transferability and knowledge transferability. To jointly take the two adversarial transferability metrics into consideration, we propose the following metric as the combined value of α_1 and α_2.

$$(\alpha_1 * \alpha_2)^{f_1\to f_2} = \left\|\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}[\alpha_1^{f_1\to f_2}(\boldsymbol{x})\widehat{\Delta_{f_1\to f_1}(\boldsymbol{x})}\widehat{\Delta_{f_1\to f_2}(\boldsymbol{x})}^\top]\right\|_F.$$

We defer the justification for the combined adversarial transferability metric in the next section, and move on to state a useful proposition.

Proposition 2.2. *The adversarial transferabililty metrics $\alpha_1^{f_1\to f_2}$, $\alpha_2^{f_1\to f_2}$ and $(\alpha_1 * \alpha_2)^{f_1\to f_2}$ are in $[0,1]$.*

So far, we have defined knowledge transferability, and two adversarial trasferability metrics. We can now analyze their connections more precisely.

Problem of Interest. Given a *source* model $f_S : \mathbb{R}^n \to \mathbb{R}^m$, the *target* data distribution $\boldsymbol{x} \sim \mathcal{D}$, the ground truth target $y : \mathbb{R}^n \to \mathbb{R}^d$, and a *target* reference model $f_T : \mathbb{R}^n \to \mathbb{R}^d$, we aim to study how the adversarial transferability between f_S and f_T, characterized by the two proposed adversarial transferability metrics, connects to the knowledge transfer loss $\min_{g\in\mathbb{G}} \mathcal{L}(g\circ f_S, y, \mathcal{D})$ with affine functions $g \in \mathbb{G}$ (equation 1).

3. Theoretical Analysis

In this section, we present the theoretical analysis on how the adversarial transferability and the knowledge transfer process are tied together. To simplify the discussion, as the objects studied in this section are specifically focused on the source domain S and the target domain T, we can use $\star$ or $\diamond$ as a placeholder for either S or T throughout this section.

Theoretical Analysis Overview. In subsection 3.1, we define the two *generalized adversarial transferabilities*, (*i.e.*, $\boldsymbol{A}_1$, $\boldsymbol{A}_2$), and present Theorem 3.1 showing that $\boldsymbol{A}_1$, $\boldsymbol{A}_2$ together determine a gradient matching distance $\min_{g\in\mathbb{G}} \|\nabla f_\star - \nabla g \circ f_\diamond\|$, between the Jacobian matrices of the source and target models in an inner product space defined by the Hessian of the adversarial loss function. In the same subsection, we also show that α_1 and α_2 represent the most influential factors in $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$, respectively. Next, we explore the connection to knowledge transferability in subsection 3.2 via Theorem 3.2 which shows the gradient matching distance approximates the function matching distance, *i.e.*, $\min_{g\in\mathbb{G}} \|f_\star - g\circ f_\diamond\|$, with a distribution shift up to a Wasserstein distance. Finally, in subsection 3.3 we complete the analysis by outlining the connection between the function matching distance and the knowledge transfer loss. A visual overview is shown in Figure 1 (b).

Setting. As adversarial perturbations are constrained in a small ϵ-ball, it is reasonable to approximate the deviation of model outputs by its first-order Taylor approximation. Specifically, in this section we consider the Euclidean ϵ-ball. Therefore, the output deviation of a function f at $\boldsymbol{x}$ given a small perturbation $\|\boldsymbol{\delta}_\epsilon\|_2 \leq \epsilon$ can be approximated by

$$f(\boldsymbol{x}+\boldsymbol{\delta}_\epsilon) - f(\boldsymbol{x}) \approx \nabla f(\boldsymbol{x})^\top \boldsymbol{\delta}_\epsilon,$$

where $\nabla f(\boldsymbol{x})$ is the Jacobian matrix of f at $\boldsymbol{x}$.

We consider a convex and twice-differentiable adversarial loss function $\ell_{adv}^\star(\cdot)$ that measures the deviation of model output $f_\star(\boldsymbol{x}+\boldsymbol{\delta}_\epsilon) - f_\star(\boldsymbol{x})$, with minimum $\ell_{adv}^\star(\boldsymbol{0}) = 0$, for $\star \in \{S, T\}$. We note that we should treat the adversarial loss on f_S and f_T differently, as they may have different output dimensions. Accordingly, the adversarial attack (equation 3) can be written as

$$\boldsymbol{\delta}_{f_\star,\epsilon}(\boldsymbol{x}) = \underset{\|\boldsymbol{\delta}\|_2\leq\epsilon}{\arg\max}\ \ell_{adv}^\star(\nabla f_\star(\boldsymbol{x})^\top \boldsymbol{\delta}). \quad (4)$$

Another justification of the small-ϵ approximation follows the literature; since the ideal attack defined in equation 3 is often intractable to compute, much of the literature uses the proposed formulation (4) in practice, *e.g.*, see (Miyato et al., 2018), with experimental results suggesting similar behaviour as the standard definition.

The Small-ϵ Regime. Recall that the adversarial loss $\ell_{adv}^\star(\cdot)$ studied in this section is convex, twice-differentiable, and achieves its minimum at $\mathbf{0}$, thus in the small ϵ regime:

$$\begin{aligned}\ell^{\star}_{adv}(\nabla f_\star(\boldsymbol{x})^\top \boldsymbol{\delta}_\epsilon) &= \left(\boldsymbol{\delta}_\epsilon^\top \nabla f_\star(\boldsymbol{x}) \boldsymbol{H}_\star \nabla f_\star(\boldsymbol{x})^\top \boldsymbol{\delta}_\epsilon\right)^{1/2} \\ &= \|\nabla f_\star(\boldsymbol{x})^\top \boldsymbol{\delta}_\epsilon\|_{\boldsymbol{H}_\star},\end{aligned}$$

which is the norm of $f_\star$'s output deviation in the inner product space defined by the Hessian $\boldsymbol{H}_\star$ of the squared adversarial loss $(\ell^{\star}_{adv})^2$.

Accordingly, the adversarial attacks (4) can be written as

$$\boldsymbol{\delta}_{f_\star,\epsilon}(\boldsymbol{x}) = \operatorname*{arg\,max}_{\|\boldsymbol{\delta}\|_2 \le \epsilon} \|\nabla f_\star(\boldsymbol{x})^\top \boldsymbol{\delta}\|_{\boldsymbol{H}_\star}, \tag{5}$$

and we can measure the output deviation of $f_\diamond$'s caused by $f_\star$'s adversarial attack $\boldsymbol{\delta}_{f_\star,\epsilon}(\boldsymbol{x})$, denoted as:

$$\Delta_{f_\star \to f_\diamond,\epsilon}(\boldsymbol{x}) = \nabla f_\diamond(\boldsymbol{x})^\top \boldsymbol{\delta}_{f_\star,\epsilon}(\boldsymbol{x}). \tag{6}$$

Note that in the small-ϵ regime, the actual value of ϵ becomes trivial (*e.g.,* α_1), consequently we will omit the ϵ for notational ease:

$$\alpha_1^{f_\star \to f_\diamond}(\boldsymbol{x}) = \frac{\|\Delta_{f_\star \to f_\diamond}(\boldsymbol{x})\|_{\boldsymbol{H}_\diamond}}{\|\nabla f_\diamond(\boldsymbol{x})\|_{\boldsymbol{H}_\diamond}}.$$

Similarly, α_2 can be computed using (6) in Definition 2, *i.e.*,

$$\alpha_2^{f_\star \to f_\diamond} = \|\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}[\widehat{\Delta_{f_\star \to f_\star}(\boldsymbol{x})}\widehat{\Delta_{f_\star \to f_\diamond}(\boldsymbol{x})}^\top]\|_F.$$

With these insights, next we will derive our first theorem.

3.1. Adversarial Transfer Indicates the Gradient Matching Distance, and Vice Versa

We present an interesting finding in this subsection, *i.e.*, the generalized adversarial transferabilities $\boldsymbol{A}_1, \boldsymbol{A}_2$ have a direct connection to the gradient matching distance between the source model $f_S : \mathbb{R}^n \to \mathbb{R}^m$ and target model $f_T : \mathbb{R}^n \to \mathbb{R}^d$. The gradient matching distance is defined as the smallest distance an affine transformation can achieve between their Jacobians $\nabla f_T : \mathbb{R}^n \to \mathbb{R}^{n\times d}$ and $\nabla f_S : \mathbb{R}^n \to \mathbb{R}^{n\times m}$ in the inner product space defined by $\boldsymbol{H}_\star$ and data sample distribution $\boldsymbol{x} \sim \mathcal{D}$, as shown below.

$$\min_{g\in\mathbb{G}} \quad \|\nabla f_\star^\top - \nabla(g \circ f_\diamond)^\top\|_{\mathcal{D},\boldsymbol{H}_\star}, \tag{7}$$

where $g \in \mathbb{G}$ are affine transformations. Note that $g : \mathbb{R}^m \to \mathbb{R}^d$ if $(\star,\diamond) = (T,S)$, and $g : \mathbb{R}^d \to \mathbb{R}^m$ if $(\star,\diamond) = (S,T)$. We defer the analysis of how the gradient matching distance approximates the knowledge transfer loss, and focus on its connection to adversarial transfer.

A Full Picture of Adversarial Transferability. A key observation is that the adversarial attack (equation 5) is the singular vector corresponding to the largest singular value of the Jacobian $\nabla f_\star(\boldsymbol{x})$ in the $\boldsymbol{H}_\star$ inner product space. Thus, information regarding other singular values that are not revealed by the adversarial attack. Therefore, we can consider other singular values, corresponding to smaller signals than the one revealed by adversarial attacks, to complete the analysis. We denote $\boldsymbol{\sigma}_{f_\star,\boldsymbol{H}_\star}(\boldsymbol{x}) \in \mathbb{R}^n$ as the descending (in absolute value) singular values of the Jacobian $\nabla f_\star(\boldsymbol{x})^\top \in \mathbb{R}^{\cdot\times n}$ in the $\boldsymbol{H}_\star$ inner product space. In other words, we denote $\boldsymbol{\sigma}_{f_\star,\boldsymbol{H}_\star}(\boldsymbol{x}) \in \mathbb{R}^n$ as the square root of the descending eigenvalues of $\nabla f_\star(\boldsymbol{x})\boldsymbol{H}_\star\nabla f_\star(\boldsymbol{x})^\top$, *i.e.*,

$$\boldsymbol{\sigma}_{f_\star,\boldsymbol{H}_\star}(\boldsymbol{x}) = [\sigma_{f_\star}^{(1)}(\boldsymbol{x}), \ldots, \sigma_{f_\star}^{(n)}(\boldsymbol{x})]^\top. \tag{8}$$

Note that the number of non-zero singular values may be less than n, in which case we fill the rest with zeros such that vector is n-dimensional.

Since the adversarial attack $\boldsymbol{\delta}_{f_\star,\epsilon}(\boldsymbol{x})$ corresponds to the largest singular value $\sigma_{f_\star}(\boldsymbol{x})^{(1)}$, we can also generalize the adversarial attack by including all the singular vectors. *i.e.*,

$$\boldsymbol{\delta}_{f_\star}^{(i)}(\boldsymbol{x}) \quad \text{corresponds to} \quad \sigma_{f_\star}^{(i)}(\boldsymbol{x}), \quad \forall i \in [n]. \tag{9}$$

Loosely speaking, one could think $\boldsymbol{\delta}_{f_\star}^{(i)}(\boldsymbol{x})$ as the adversarial attack of $f_\star(\boldsymbol{x})$ in the subspace orthogonal to all the previous attacks, *i.e.*, $\boldsymbol{\delta}_{f_\star}^{(j)}(\boldsymbol{x})$ for $\forall j \in [i-1]$.

Accordingly, for $\forall i \in [i]$ we denote the output deviation as

$$\Delta_{f_\star \to f_\diamond}^{(i)}(\boldsymbol{x}) = \nabla f_\diamond(\boldsymbol{x})^\top \boldsymbol{\delta}_{f_\star}^{(i)}(\boldsymbol{x}). \tag{10}$$

As a consequence, we *generalize the first adversarial transferability* to be a n-dimensional vector $\boldsymbol{A}_1^{\star,\diamond}(\boldsymbol{x})$ including the adversarial losses of all of the generalized adversarial attacks, where the i^{th} element in the vector is

$$\boldsymbol{A}_1^{\star,\diamond}(\boldsymbol{x})^{(i)} = \frac{\|\Delta_{f_\star \to f_\diamond}^{(i)}(\boldsymbol{x})\|_{\boldsymbol{H}_\diamond}}{\|\nabla f_\diamond(\boldsymbol{x})\|_{\boldsymbol{H}_\diamond}}. \tag{11}$$

Note that the first entry of $\boldsymbol{A}_1^{\star,\diamond}(\boldsymbol{x})$ is the original adversarial transferability, *i.e.*, $\boldsymbol{A}_1^{\star,\diamond}(\boldsymbol{x})^{(1)}$ is the same as the $\alpha_1^{f_\star \to f_\diamond}(\boldsymbol{x})$ in Definition 1.

With the above generalization that captures the full picture of the adversarial transfer process, we able to derive the following theorem.

Theorem 3.1. *Given the target and source models $f_\star, f_\diamond$, where $(\star,\diamond) \in \{(S,T),(T,S)\}$, the gradient matching distance (equation 7) can be written as*

$$\min_{g\in\mathbb{G}} \quad \|\nabla f_\star^\top - \nabla(g \circ f_\diamond)^\top\|_{\mathcal{D},\boldsymbol{H}_\star} = \tag{12}$$

$$\sqrt{1 - \frac{\mathbb{E}[\boldsymbol{v}^{\star,\diamond}(\boldsymbol{x}_1)^\top \boldsymbol{A}_2^{\star,\diamond}(\boldsymbol{x}_1,\boldsymbol{x}_2)\boldsymbol{v}^{\star,\diamond}(\boldsymbol{x}_2)]}{\|\nabla f_\star^\top\|^2_{\mathcal{D},\boldsymbol{H}_\star} \cdot \|\boldsymbol{J}^\dagger\|^{-1}_{\boldsymbol{H}_\diamond}}}\,\|\nabla f_\star^\top\|_{\mathcal{D},\boldsymbol{H}_\star},$$

where the expectation is taken over $\boldsymbol{x}_1, \boldsymbol{x}_2 \overset{i.i.d.}{\sim} \mathcal{D}$, and

$$\begin{gathered}\boldsymbol{v}^{\star,\diamond}(\boldsymbol{x}) = \sigma_{f_\diamond,\boldsymbol{H}_\diamond}^{(1)}(\boldsymbol{x})\boldsymbol{\sigma}_{f_\star,\boldsymbol{H}_\star}(\boldsymbol{x}) \odot \boldsymbol{A}_1^{\star,\diamond}(\boldsymbol{x}) \\ \boldsymbol{J} = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}[\nabla f_\diamond(\boldsymbol{x})^\top \nabla f_\diamond(\boldsymbol{x})].\end{gathered}$$

Moreover, $\boldsymbol{A}_2^{\star,\diamond}(\boldsymbol{x}_1,\boldsymbol{x}_2)$ is a matrix, and its element in the i^{th} row and j^{th} column is

$$\begin{aligned}\boldsymbol{A}_2^{\star,\diamond}(\boldsymbol{x}_1,\boldsymbol{x}_2)^{(i,j)} = \langle \widehat{\Delta^{(i)}_{f_\star\to f_\star}(\boldsymbol{x}_1)}\big|_{\boldsymbol{H}_\star}, \widehat{\Delta^{(j)}_{f_\star\to f_\star}(\boldsymbol{x}_2)}\big|_{\boldsymbol{H}_\star}\rangle \\ \cdot \langle \widehat{\Delta^{(i)}_{f_\star\to f_\diamond}(\boldsymbol{x}_1)}\big|_{\boldsymbol{H}_\diamond}, \widehat{\Delta^{(j)}_{f_\star\to f_\diamond}(\boldsymbol{x}_2)}\big|_{\boldsymbol{H}_\diamond}\rangle_{\widehat{\boldsymbol{J}^\dagger}|_{\boldsymbol{H}_\diamond}}.\end{aligned}$$

Recall the alternative representation of the second adversarial transferability α_2, and we can immediately observe that α_2 is determined by $\boldsymbol{A}_2$. Therefore, both α_1 and α_2 appear in this relation. Let us interpret the theorem, and justify the two proposed adversarial transferability metrics.

Interpretation of Theorem 3.1. First, we consider components that are not directly related to the adversarial transfer in the RHS of (12). The $\|\nabla f_\star^\top\|_{\mathcal{D},\boldsymbol{H}_\star}$ outside represents the overall magnitude of the loss. In the fraction, the $\|\nabla f_\star^\top\|_{\mathcal{D},\boldsymbol{H}_\star}$ in the denominator normalizes the $\boldsymbol{\sigma}_{f_\star}$ in the numerator. Similarly, though more complicated, the $\|\boldsymbol{J}^\dagger\|_2^{-1}$ in the denominator corresponds to the $\sigma_{f_\diamond}^{(1)}$ in the numerator. We note that these are properties of $f_\star, f_\diamond$.

Next, observe that the components directly related to the adversarial transfer process are the *generalized adversarial transferability* $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$. Let us neglect the superscript $^{(i)}$ or $^{(j)}$ for now, so we can see that their interpretations are the same as we introduced for α_1 and α_2 in section 2. That is, $\boldsymbol{A}_1$ captures the magnitude of the deviation in model outputs caused by adversarial attacks, while $\boldsymbol{A}_2$ captures the direction of the deviation. A minor difference between α_2 and $\boldsymbol{A}_2$ is that the second inner product in the elements of $\boldsymbol{A}_2$ is defined by a positive semi-definite matrix $\widehat{\boldsymbol{J}^\dagger}$. For practical implementation, we choose to neglect this term, and use the standard Euclidean inner product in α_2, which can be understood as a stretched version of the $\widehat{\boldsymbol{J}^\dagger}$ inner product space.

Moreover, as the singular vector $\boldsymbol{\sigma}_{f_\star}$ has descending entries, we can see that in the vector $\boldsymbol{A}_1$ and the matrix $\boldsymbol{A}_2$, the elements with superscript $^{(1)}$ have the most influence in the relations. In other words, the two proposed adversarial transferability metrics, α_1 and α_2, are the most influential factors in equation 12. We can also see that the combined metric $(\alpha_1 * \alpha_2)$ also stems from here by only considering the components with the first superscript.

To interpret the relation between the gradient matching distance and the adversarial transferabilities, we introduce the following proposition. This shows that, in general, $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ with their elements closer to 1 can serve as a bidirectional indicator of a smaller gradient matching distance.

Proposition 3.1. *In Theorem 3.1,*

$$0 \le \frac{\mathbb{E}[\boldsymbol{v}^{\star,\diamond}(\boldsymbol{x}_1)^\top \boldsymbol{A}_2^{\star,\diamond}(\boldsymbol{x}_1,\boldsymbol{x}_2)\boldsymbol{v}^{\star,\diamond}(\boldsymbol{x}_2)]}{\|\nabla f_\star^\top\|^2_{\mathcal{D},\boldsymbol{H}_\star}\cdot\|\boldsymbol{J}^\dagger\|^{-1}_{\boldsymbol{H}_\diamond}} \le 1.$$

In conclusion, Theorem 3.1 reveals a bidirectional relation between the adversarial transfer process and the gradient matching distance, where the adversarial transfer process can be encoded by the generalized adversarial transferabilities, *i.e.*, $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$. Moreover, α_1 and α_2 play the most influential role in their generalization, *i.e.,* $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$.

3.2. The Gradient Matching Distance indicates the Function Matching Distance, and Vice Versa

To bridge the gradient matching distance to the knowledge transfer loss, an immediate step is to connect the gradient distance to the function distance which directly serves as a surrogate knowledge transfer loss as defined in (equation 2). Specifically, in this subsection, we present a connection between the function matching distance, *i.e.*,

$$\min_{g\in\mathbb{G}} \|f_\star - g\circ f_\diamond\|_{\mathcal{D},\boldsymbol{H}_\star}, \tag{13}$$

and the gradient matching distance, *i.e.*,

$$\min_{g\in\mathbb{G}} \|\nabla f_\star^\top - \nabla(g\circ f_\diamond)^\top\|_{\mathcal{D},\boldsymbol{H}_\star}, \tag{14}$$

where $g \in \mathbb{G}$ are affine transformations.

For intuition, consider a point $\boldsymbol{x}_0$ in the input space $\mathbb{R}^n$, a path $\gamma_{\boldsymbol{x}} : [0,1] \to \mathbb{R}^n$ such that $\gamma_{\boldsymbol{x}}(0) = \boldsymbol{x}_0$ and $\gamma_{\boldsymbol{x}}(1) = \boldsymbol{x}$. Then, denoting γ as the function of $\boldsymbol{x}$, we can write the difference between the two functions as

$$\begin{aligned}f_\star - g\circ f_\diamond = \int_0^1 (\nabla f_\star(\gamma(t)) - \nabla(g\circ f_\diamond(\gamma(t))))^\top \dot\gamma(t)\,\mathrm{d}t \\ + (f_\star(\boldsymbol{x}_0) - g\circ f_\diamond(\boldsymbol{x}_0)).\end{aligned}$$

Noting that the function difference is a path integral of the gradient difference, we should expect a distribution shift when characterizing their connection, *i.e.*, the integral path affects the original data distribution $\mathcal{D}$. Accordingly, as the integral path may leave the support of $\mathcal{D}$, it is necessary to assume the smoothness of the function, as shown below.

Denoting the optimal $g \in \mathbb{G}$ in (13) as $\tilde g$, and one of the optimal $g \in \mathbb{G}$ in (14) as $\tilde g'$, we define

$$h_{\star,\diamond} := f_\star - \tilde g\circ f_\diamond \quad\text{and}\quad h'_{\star,\diamond} := f_\star - \tilde g'\circ f_\diamond, \tag{15}$$

and we can see that the gradient matching distance and the function matching distance can be written as

$$(13) = \|h_{\star,\diamond}\|_{\mathcal{D},\boldsymbol{H}_\star} \quad\text{and}\quad (14) = \|\nabla {h'_{\star,\diamond}}^\top\|_{\mathcal{D},\boldsymbol{H}_\star}.$$

Assumption 1 (β-smoothness)**.** *We assume $h_{\star,\diamond}$ and $h'_{\star,\diamond}$ are both β-smooth,* i.e.*,*

$$\|\nabla h_{\star,\diamond}^\top(\boldsymbol{x}_1) - \nabla h_{\star,\diamond}^\top(\boldsymbol{x}_2)\|_{\boldsymbol{H}_\star} \le \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2,$$

and similarly for $h'_{\star,\diamond}$.

With this assumption, we can prove that the gradient matching distance and the function matching distance can bound each other.

Theorem 3.2. *With the notation defined in equation 15, assume the β-smoothness assumption holds. Given a data distribution $\mathcal{D}$ and $\tau > 0$, there exist distributions $\mathcal{D}_1, \mathcal{D}_2$ such that the type-1 Wasserstein distance $W_1(\mathcal{D}, \mathcal{D}_1) \leq \tau$ and $W_1(\mathcal{D}, \mathcal{D}_2) \leq \tau$ satisfying*

$$\frac{1}{2B^2}\|h_{\star,\diamond}\|^2_{\mathcal{D},\boldsymbol{H}_\star} \leq \|\nabla h'^{\top}_{\star,\diamond}\|^2_{\mathcal{D}_1,\boldsymbol{H}_\star} + \beta^2(B-\tau)^2_+$$
$$\frac{1}{3n}\|\nabla h'^{\top}_{\star,\diamond}\|^2_{\mathcal{D},\boldsymbol{H}_\star} \leq \frac{2}{\tau^2}\|h_{\star,\diamond}\|^2_{\mathcal{D}_2,\boldsymbol{H}_\star} + \beta^2\tau^2,$$

where n is the dimension of $\boldsymbol{x} \sim \mathcal{D}$, and $B = \inf_{\boldsymbol{x}_0 \in \mathbb{R}^n} \sup_{\boldsymbol{x} \in supp(\mathcal{D})} \|\boldsymbol{x} - \boldsymbol{x}_0\|_2$ is the radius of $supp(\mathcal{D})$.

We note that the above theorem compromises some tightness in exchange for a cleaner presentation without losing its core message, which is discussed in the proof of the theorem.

Interpretation of Theorem 3.2. The theorem shows that under the smoothness assumption, the gradient matching distance indicates the function matching distance, and vice versa, with a distribution shift bounded in Wasserstein distance. As the distribution shift is in general necessary, we conjecture that using different data distributions for adversarial transfer and knowledge transfer can also be applicable.

3.3. The Function Matching Distance Indicates Knowledge Transferability, and Vice Versa

To complete the story, it remains to connect the function matching distance to knowledge transferability. As the adversarial transfer is symmetric (*i.e.*, either from $f_S \to f_T$ or $f_T \to f_S$), we are able to use the placeholders $\star, \diamond \in \{S, T\}$ all the way through. However, as the knowledge transfer is asymmetric (*i.e.*, $f_S \to y$ to the target ground truth), we need to instantiate the direction of adversarial transfer to further our discussion.

Adversarial Transfer from $f_T \to f_S$. As we can see from the $\boldsymbol{A}_1^{\star,\diamond}$ in Theorem 3.1, this direction corresponds to $(\star, \diamond) = (T, S)$. Accordingly, the function matching distance (equation 13) becomes

$$\min_{g \in \mathbb{G}} \|f_T - g \circ f_S\|_{\mathcal{D},\boldsymbol{H}_T}. \tag{16}$$

We can see that equation 16 directly translates to the surrogate knowledge transfer loss that uses the "pseudo ground truth" from the target reference model f_T.

In other words, the function matching distance serves as an approximation of the knowledge transfer loss defined as their distance in the inner product space of $\boldsymbol{H}_T$, *i.e.*,

$$\min_{g \in \mathbb{G}} \|y - g \circ f_S\|_{\mathcal{D},\boldsymbol{H}_T}. \tag{17}$$

The accuracy of the approximation depends on the performance of f_T, as shown in the following theorem.

Theorem 3.3. *The surrogate transfer loss (16) and the true transfer loss (17) are close, with an error of $\|f_T - y\|_{\mathcal{D},\boldsymbol{H}_T}$.*

$$-\|f_T - y\|_{\mathcal{D},\boldsymbol{H}_T} \leq (17) - (16) \leq \|f_T - y\|_{\mathcal{D},\boldsymbol{H}_T}$$

Adversarial Transfer from $f_S \to f_T$. This direction corresponds to $(\star, \diamond) = (S, T)$. Accordingly, the function matching distance (equation 13) becomes

$$\min_{g \in \mathbb{G}} \|f_S - g \circ f_T\|_{\mathcal{D},\boldsymbol{H}_S}. \tag{18}$$

Since the affine transformation g acts on the target reference model, it can not be directly viewed as a surrogate transfer loss. However, interesting interpretations can be found in this direction, depending on the output dimension of $f_S : \mathbb{R}^n \to \mathbb{R}^m$ and $f_T : \mathbb{R}^n \to \mathbb{R}^d$.

That is, when the direction of adversarial transfer is from $f_S \to f_T$, the indicating relation between it and knowledge transferability would possibly be unidirectional, depending on the dimensions. More discussion is included in the appendix section B due to space limitation.

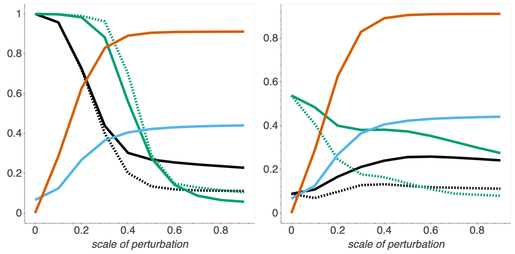
4. Synthetic Experiments

The synthetic experiment aims to bridge the gap between theory and practice by verifying some of the theoretical insights that may be difficult to compute for large-scale experiments. Specifically, the synthetic experiment aims to verify: first, how influential are the two proposed adversarial transferability metrics α_1, α_2 comparing to the other factors in the generalized adversarial attacks (equation 9); Second, how does the gradient matching distance track the knowledge transfer loss. The dataset ($N = 5000$) is generated by a Gaussian mixture of 10 Gaussians. The ground truth target is set to be the sum of 100 radial basis functions. The dimension of $\boldsymbol{x}$ is 50, and the dimension of the target is 10. Details of the datasets are defer to appendix section F.

Models Both the source model f_S and target model f_T are one-hidden-layer neural networks with sigmoid activation.

Methods First, sample $D = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^N$ from the distribution, where $\boldsymbol{x}$ is 50-dimensional, $\boldsymbol{y}$ is 10-dimensional. Then we train a target model f_T on D. To derive the source models, we first train a target model on D with width $m = 100$. Denoting the weights of a target model as $\boldsymbol{W}$, we randomly sample a direction $\boldsymbol{V}$ where each entry of $\boldsymbol{V}$ is sampled from $U(-0.5, 0.5)$, and choose a scale $t \in [0, 1]$. Subsequently, we perturb the model weights of the clean source model as $\boldsymbol{W'} := \boldsymbol{W} + t\boldsymbol{V}$, and define the source model f_S to be a one-hidden-layer neural network with weights $\boldsymbol{W'}$. Then, we compute each of the quantities we care about, including α_1, α_2 from both $f_S \to f_T$ and $f_T \to f_S$, the gradient matching distance (equation 7), and the actual knowledge transfer distance (equation 17). We use the standard ℓ_2 loss as the adversarial loss function.

Results We present two sets of experiment in Figure 2. The indication relations between adversarial transferability and knowledge transferability can be observed. Moreover: 1. the metrics α_1, α_2 are more meaningful if using the regular attacks $\boldsymbol{\delta}_{f_\star}^{(1)}$; 2. the gradient matching distance tracks the actual knowledge transferability loss; 3. the directions of $f_T \to f_S$ and $f_S \to f_T$ are similar.



(a) $\boldsymbol{\delta}_{f_\star}^{(1)}$ (b) $\boldsymbol{\delta}_{f_\star}^{(2)}$

Figure 2. As defined in equation 9, (a) corresponds to the regular adversarial attacks, while (b) the secondary adversarial attack. That is, (b) represents the other information in the adversarial transferring process compared with the first. The x-axis shows the scale of perturbation $t \in [0, 1]$ that controls how much the source model deviates from the target model. There are in total 6 quantities reported. Specifically, $\alpha_1^{f_T \to f_S}$ is **black solid**; $\alpha_1^{f_S \to f_T}$ is **black dotted**; $\alpha_2^{f_T \to f_S}$ is **green solid**; $\alpha_2^{f_S \to f_T}$ is **green dotted**; the gradient matching loss is **red solid**; and the knowledge transferability distance is **blue solid**.

5. Experimental Evaluation

We present the real-data experiments based on both image and natural language datasets in this section, and discuss the potential applications.

Adversarial Transferability Indicating Knowledge Transferability. In this experiment, we show how to use adversarial transferability to identify the optimal transfer learning candidates from a pool of models trained on the same source dataset. We first train 5 different architectures (AlexNet, Fully connected network, LeNet, ResNet18, ResNet50) on cifar10 (Krizhevsky et al., 2009). Then we perform transfer learning to STL10 (Coates et al., 2011) to obtain the knowledge transferability of each, measured by accuracy. At the same time, we also train one ResNet18 on STL10 as the target model, which has poor accuracy because of the lack of data. To measure the adversarial transferability, we generate adversarial examples with PGD (Madry et al., 2017) on the target model and use the generated adversarial examples to attack each source model. The adversarial transferability is expressed in the form of α_1 and α_2. Our results in Table 1 indicate that we can use adversarial tarnsferability to forecast knowledge transferability, where the only major computational overheads are training a naive model on the target domain and generating a few adversarial examples. In the end, We further evaluate the significance of our results by Pearson score. More details about training and generation of adversarial examples can be found in the appendix G.

Model	Knowledge Trans.	α_1	α_2	$\alpha_1 * \alpha_2$
Fully Connected	28.30	0.346	0.189	0.0258
LeNet	45.65	0.324	0.215	0.0254
AlexNet	55.09	0.337	0.205	0.0268
ResNet18	76.60	0.538	0.244	0.0707
ResNet50	77.92	0.614	0.234	0.0899

Table 1. Knowledge transferability (Knowledge Trans.) among different model architectures. Our correlation analysis shows Pearson score of -0.51 between the transfer loss and α_1. Lower transfer loss corresponds to higher transfer accuracy. More details can be found in fig 4 in the Appendix G

To further validate our idea, we also conduct experiments on the NLP domain. We first finetune 5 different BERT classification models on different data domain (IMDB, Moview Review (MR), Yelp, AG, Fake). We refer the models trained on MR, Yelp, AG and Fake datasets as the source models, and take the model trained on IMDB datset as the target model. To measure the knowledge transferability, we finetune the source models with new linear layers on the target dataset for one epoch. We report the accuracy of the transferred models on the target test set as the metric to indicate the knowledge transferability. In terms of the adversarial transferability, we generate adversarial examples by the state-of-the-art whitebox attack algorithm T3 (Wang et al., 2020) against the target model and transfer the adversarial examples to source models to evaluate the adversarial transferability. Following our previous experiment, we also calculate α_1 and α_2. Experimental results are shown in Table 2. We observe that source models with larger adversarial transferability, measured by α_1, α_2 and $\alpha_1 * \alpha_2$, indeed tend to have larger knowledge transferability.

Model	Knowledge Trans.	α_1	α_2	$\alpha_1 * \alpha_2$
MR	89.34	0.743	0.00335	3.00e-3
Yelp	88.81	0.562	0.00135	8.87e-4
AG	87.58	0.295	0.00021	8.56e-5
Fake	84.06	0.028	0.00032	5.58e-6

Table 2. Knowledge transferability (Knowledge Trans.) from the Source Models (MR, Yelp, AG, Fake) to the Target Model (IMDB). Adversarial transferability is measured by using the adversarial examples generated against the Target Model (IMDB) to attack the Source Models and estimate α_1 and α_2. The correlation analysis shows Pearson Score of 0.27 between the transfer confidence and α_1. Higher transfer confidence indicates higher knowledge transferability. More details can be found in Figure 6 in Appendix §G.

Knowledge Transferability Indicating Adversarial Transferability. In addition, we are interested in the impact of knowledge transferability on adversarial transferability.

Similarity	Knowledge Trans.	α_1	α_2	$\alpha_1 * \alpha_2$
0%	45.00	0.310	0.146	0.0169
25%	45.68	0.318	0.305	0.0383
50%	59.09	0.338	0.355	0.0436
75%	71.62	0.337	0.312	0.0402
100%	81.84	0.358	0.357	0.0489

Table 3. Knowledge transferability (Knowledge Trans.) of different source model. Similarity indicates how similar the source distributions are with the target distribution. Our correlation analysis shows Pearson score of -0.06 between the transfer loss and α_1. Lower transfer loss corresponds to higher knowledge transferability. More details can be found in fig 8 in the Appendix G.

Model	Knowledge Trans.	α_1	α_2	$\alpha_1 * \alpha_2$
MR	89.34	0.584	0.00188	2.32e-3
Yelp	88.81	0.648	0.00120	9.52e-4
AG	87.58	0.293	0.00016	4.35e-6
Fake	84.06	0.150	0.00073	3.55e-5

Table 4. Knowledge transferability (Knowledge Trans.) from the Source Models (MR, Yelp, AG, Fake) to the Target Model (IMDB). Adversarial transferability is measured by using the adversarial examples generated against the Source Models to attack the Target Models and estimate α_1 and α_2. The correlation analysis shows Pearson Score of 0.27 between the transfer confidence and α_1. Higher transfer confidence indicates higher knowledge transferability. More details can be found in Figure 9 in Appendix §G.

As predicted by our theory, the more knowledge transferable a source model is to the target domain, the more adversarial transferable it is.

We split cifar10 into 5 different subsets containing different percentages of animals and vehicles. We train a resNet18 on each of them as source models, which are later finetuned to obtained the knowledge transferability measured by accuracy. Then we train another resNet18 on a subset of stl10 that only contains vehicles. Different from the last experiment, we generate adversarial examples with PGD on each of the source models and transfer them to the target model. Table 3 shows, the source model that transfers knowledge better generates more transferable adversarial examples. This implies we can use this relation to facilitate blackbox attack against a hidden target model, given some knowledge about the source and target domains. More details of training and generation of adversarial examples can be found in the appendix.

We evaluate the impact of knowledge transferability to adversarial transferability in the NLP domain as well. We mostly follow the setting describe in the previous section, where we have four source models and one target model, and the knowledge transferability from source models to the target model is measured by the accuracy of the transferred models on the target test set. The difference lies on the evaluation of the adversarial transferability, where we generate adversarial examples against the source models and evaluate their attack capability on the target model. As shown in Table 4, we note that when the source data domain is getting closer to the target data domain, the knowledge transferability grows, and the adversarial transferability also increases. More experimental details can be found in Appendix G.

Ablation Studies Following the settings in table 1, we conduct ablation studies (table 5) on two additional attack methods, MI (Tramèr et al., 2017a), PGD-L2 and two additional ϵ with PGD, 2/225, 4/255, we discover that neither the attack method nor ϵ has significant impact on our conclusion.

Model	Knowledge Trans.	α_1	α_2	$\alpha_1 * \alpha_2$
Fully Connected	28.30	0.0985	0.0196	0.00027
LeNet	45.65	0.2106	0.0259	0.00158
AlexNet	55.09	0.1196	0.0206	0.00037
ResNet18	76.60	0.2739	0.0413	0.00405
ResNet50	77.92	0.1952	0.0320	0.00172

$\epsilon = 2/255$. Pearson score is -0.45.

Model	Knowledge Trans.	α_1	α_2	$\alpha_1 * \alpha_2$
Fully Connected	28.30	0.0974	0.0225	0.00029
LeNet	45.65	0.2099	0.0309	0.00192
AlexNet	55.09	0.1283	0.0230	0.00048
ResNet18	76.60	0.2853	0.0481	0.00496
ResNet50	77.92	0.2495	0.0414	0.00337

$\epsilon = 4/255$. Pearson score is -0.49.

Model	Knowledge Trans.	α_1	α_2	$\alpha_1 * \alpha_2$
Fully Connected	28.30	0.1678	0.0379	0.0013
LeNet	45.65	0.0997	0.0503	0.0005
AlexNet	55.09	0.1229	0.0506	0.0009
ResNet18	76.60	0.2731	0.0630	0.0052
ResNet50	77.92	0.3695	0.0550	0.0081

Attack with MI. Pearson score is -0.45.

Model	Knowledge Trans.	α_1	α_2	$\alpha_1 * \alpha_2$
Fully Connected	28.30	0.0809	0.0175	0.00018
LeNet	45.65	0.2430	0.0190	0.00149
AlexNet	55.09	0.1101	0.0188	0.00031
ResNet18	76.60	0.3619	0.0303	0.00464
ResNet50	77.92	0.2506	0.0237	0.00179

ℓ_2 attack with $\epsilon = 1$. Pearson score is -0.40.

Table 5. With varying attack methods and ϵ, adversarial transferability is still correlated with knowledge transferability.

6. Conclusion

We theoretically analyze the relation between adversarial transferability and knowledge transferability. We provide empirical experimental justifications in pratical settings. Both our theoretical and empirical results show that adversarial transferability can indicate knowledge transferability and vice versa. We expect our work will inspire future work on further exploring other factors that impact knowledge transferability and adversarial transferability.

Acknowledgments This work is partially supported by NSF grant No.1910100, NSF CNS 20-46726 CAR, Amazon Research Award, and the Intel RSA 2020.

References

Achille, A., Lam, M., Tewari, R., Ravichandran, A., Maji, S., Fowlkes, C. C., Soatto, S., and Perona, P. Task2vec: Task embedding for meta-learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 6430–6439, 2019.

Athalye, A., Carlini, N., and Wagner, D. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International Conference on Machine Learning*, pp. 274–283, 2018.

Coates, A., Ng, A., and Lee, H. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 215–223, 2011.

Demontis, A., Melis, M., Pintor, M., Jagielski, M., Biggio, B., Oprea, A., Nita-Rotaru, C., and Roli, F. Why do adversarial attacks transfer? explaining transferability of evasion and poisoning attacks. In *28th {USENIX} Security Symposium ({USENIX} Security 19)*, pp. 321–338, 2019.

Dong, D., Wu, H., He, W., Yu, D., and Wang, H. Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1723–1732, 2015.

Dong, Y., Pang, T., Su, H., and Zhu, J. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4312–4321, 2019.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Huh, M., Agrawal, P., and Efros, A. A. What makes imagenet good for transfer learning? *arXiv preprint arXiv:1608.08614*, 2016.

Ilyas, A., Engstrom, L., Athalye, A., and Lin, J. Black-box adversarial attacks with limited queries and information. In *International Conference on Machine Learning*, pp. 2137–2146, 2018.

Joon Oh, S., Fritz, M., and Schiele, B. Adversarial image perturbation for privacy protection–a game theory perspective. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1482–1491, 2017.

Kariyappa, S. and Qureshi, M. K. Improving adversarial robustness of ensembles with diversity training. *arXiv preprint arXiv:1901.09981*, 2019.

Kendall, A., Gal, Y., and Cipolla, R. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7482–7491, 2018.

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

Liu, Y., Chen, X., Liu, C., and Song, D. Delving into transferable adversarial examples and black-box attacks. *arXiv preprint arXiv:1611.02770*, 2016.

Long, M., Cao, Y., Wang, J., and Jordan, M. Learning transferable features with deep adaptation networks. In *International Conference on Machine Learning*, pp. 97–105, 2015.

Ma, X., Li, B., Wang, Y., Erfani, S. M., Wijewickrema, S., Schoenebeck, G., Song, D., Houle, M. E., and Bailey, J. Characterizing adversarial subspaces using local intrinsic dimensionality. *arXiv preprint arXiv:1801.02613*, 2018.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Miyato, T., Maeda, S.-i., Koyama, M., and Ishii, S. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993, 2018.

Naseer, M. M., Khan, S. H., Khan, M. H., Khan, F. S., and Porikli, F. Cross-domain transferability of adversarial perturbations. In *Advances in Neural Information Processing Systems*, pp. 12885–12895, 2019.

Papernot, N., McDaniel, P., and Goodfellow, I. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277*, 2016.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3): 211–252, 2015.

Salman, H., Ilyas, A., Engstrom, L., Kapoor, A., and Madry, A. Do adversarially robust imagenet models transfer better? *arXiv preprint arXiv:2007.08489*, 2020.

Shinya, Y., Simo-Serra, E., and Suzuki, T. Understanding the effects of pre-training for object detectors via eigenspectrum. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 0–0, 2019.

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

Tramèr, F., Papernot, N., Goodfellow, I., Boneh, D., and McDaniel, P. The space of transferable adversarial examples. *arXiv preprint arXiv:1704.03453*, 2017b.

Utrera, F., Kravitz, E., Erichson, N. B., Khanna, R., and Mahoney, M. W. Adversarially-trained deep nets transfer better. *arXiv preprint arXiv:2007.05869*, 2020.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*, 2019a. URL https://openreview.net/forum?id=rJ4km2R5t7.

Wang, B., Pei, H., Pan, B., Chen, Q., Wang, S., and Li, B. T3: Tree-autoencoder constrained adversarial text generation for targeted attack. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6134–6150, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.495. URL https://www.aclweb.org/anthology/2020.emnlp-main.495.

Wang, Z., Dai, Z., Póczos, B., and Carbonell, J. Characterizing and avoiding negative transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 11293–11302, 2019b.

Xie, C., Zhang, Z., Zhou, Y., Bai, S., Wang, J., Ren, Z., and Yuille, A. L. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2730–2739, 2019.

Xu, R., Li, G., Yang, J., and Lin, L. Larger norm more transferable: An adaptive feature norm approach for unsupervised domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1426–1435, 2019.

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

Zamir, A. R., Sax, A., Shen, W., Guibas, L. J., Malik, J., and Savarese, S. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3712–3722, 2018.

Zhang, X., Zhao, J., and LeCun, Y. Character-level convolutional networks for text classification. *arXiv preprint arXiv:1509.01626*, 2015.

Zhang, Z., Luo, P., Loy, C. C., and Tang, X. Facial landmark detection by deep multi-task learning. In *European conference on computer vision*, pp. 94–108. Springer, 2014.

Zhou, W., Hou, X., Chen, Y., Tang, M., Huang, X., Gan, X., and Yang, Y. Transferable adversarial perturbations. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 452–467, 2018.