Great Minds Think Alike: The Universal Convergence Trend of Input Salience

Yipei Wang, Jeffrey Mark Siskind, Xiaoqian Wang
Elmore Family School of Electrical and Computer Engineering
Purdue University
West Lafayette, IN 47907
wang4865,qobi,joywang@purdue.edu

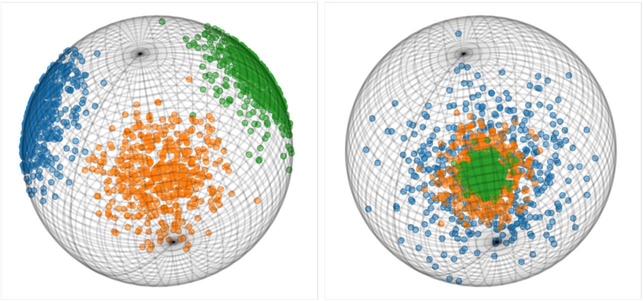
Abstract

Uncertainty is introduced in optimized DNNs through stochastic algorithms, forming specific distributions. Training models can be seen as random sampling from this distribution of optimized models. In this work, we study the distribution of optimized DNNs as a family of functions by leveraging a pointwise approach. We focus on the input saliency maps, as the input gradient field is decisive to the models' mathematical essence. Our investigation of saliency maps reveals a counter-intuitive trend: two stochastically optimized models tend to resemble each other more as either of their capacities increases. Therefore, we hypothesize several properties of these distributions, suggesting that (1) Within the same model architecture (e.g., CNNs, ResNets), different family variants (e.g., varying capacities) tend to align in terms of their population mean directions of the input salience. And (2) the distributions of optimized models follow a convergence trend to their shared population mean as the capacity increases. Furthermore, we also propose semi-parametric distributions based on the Saw distribution to model the convergence trend, satisfying all the counter-intuitive observations. Our experiments shed light on the significant implications of our hypotheses in various application domains, including black-box attacks, deep ensembles, etc. These findings not only enhance our understanding of DNN behaviors but also offer valuable insights for their practical application in diverse areas of deep learning.

1 Introduction

The advancement in computational power has significantly enhanced the capabilities of Deep Neural Networks (DNNs), leading to their unparalleled expressiveness and success in a multitude of applications across various fields (Krizhevsky et al., 2012; He et al., 2016; Rajkomar et al., 2018; Berner et al., 2019; Rombach et al., 2022; Padmaja et al., 2023; Thirunavukarasu et al., 2023). Despite these achievements, DNNs remain enigmatic, not only to end-users but also to researchers and practitioners (Ribeiro et al., 2016; Rudin, 2018; Preece et al., 2018). Due to the over-parameterization nature of modern DNNs, they are capable of reaching zero loss in the training distribution (Goodfellow et al., 2014b; Allen-Zhu et al., 2019; Du et al., 2019). Furthermore, the inherent stochastic nature of training algorithms means that even when using the same training data, DNNs tend to converge to various minima (Huang et al., 2017; Liu et al., 2020). Thus even though these models may exhibit comparable performance in terms of metrics like testing loss or accuracy, their underlying mechanisms can still differ significantly. Because of the stochastic nature of the training procedure, optimized DNNs collectively form a distribution over the functional space $\mathcal{C}^1(\mathcal{X})$, and training DNNs from scratch is thereby equivalent to randomly sampling from such a distribution without any guarantee. This inherent opacity, combined with the high dimensionality and nonlinearity, limits our understanding of the internal mechanisms of DNNs.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).



(a) Random Distribution (b) Convergent Distribution

Figure 1: A synthetic illustration of the distribution of the directional gradients of stochastically optimized models of the same input data. The subfigures demonstrate (a) an intuitive, stochastic scenario, where the distributions of different model families are not closely dependent. and (b) the converging distribution trend introduced by our hypothesis. Different colors represent different model families, and points represent different optimized models.

In response to these challenges, we study the aforementioned distributions. By adopting a pointwise approach, our focus is on the distribution of input salience (Simonyan et al., 2013) from the context of eXplainable Artificial Intelligence (XAI), which aims to demystify the inner workings of these complex models (Gunning and Aha, 2019; Arrieta et al., 2020; Van der Velden et al., 2022). Saliency maps, particularly in the form of *input gradients*, represent the data points within the gradient fields of DNNs. Thus the study of gradients can offer a deterministic view of the landscape of model predictions. This approach allows us to examine the intricate nuances of DNNs in a more structured and analytical manner.

For clarity, in the following context, we distinguish between the term **model architecture** (e.g. skip/direct connections) from the term **model family**. The latter refers to a specific collection of models $\mathcal{F}$, that differ only in capacity as determined by width and depth. Two models are said to be in the same family if they differ only in parameter values. Given an input, varying model families result in distinct distributions. A synthetic visualization of such distributions is shown in Figure 1(a). Different models are depicted by the points. However, the relationship between different model families, represented by various colors, remains elusive. In this work, we introduce and verify several hypotheses to uncover a striking pattern. (1) Within the same model architecture (e.g., CNNs, ResNets), different family variants (e.g., varying capacities) tend to align in terms of their population mean direction. (2) As the model capacities increase, the variance within the distribution of the same family diminishes. This leads to a converging trend of the distributions. Both hypotheses are illustrated in Figure 1(b). Additionally, we introduce a semi-parametric approach to model these distributions, providing detailed quantification of the convergence.

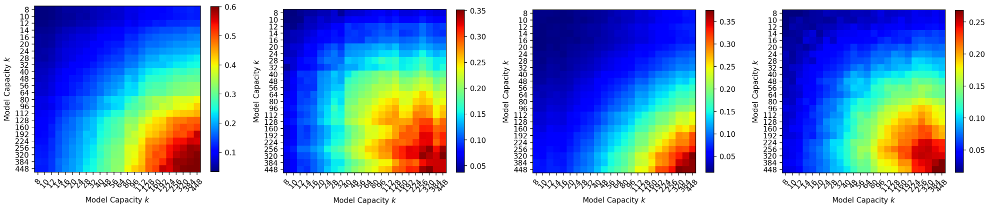
The similarities observed in input salience have direct implications for understanding the important vulnerability of DNNs regarding gradient attacks (Szegedy et al., 2013; Goodfellow et al., 2014a). In particular, in black-box attack settings, the gradients of the target model are not directly accessible. A higher degree of salience similarity naturally enhances transferability (Chen et al., 2023). Our findings elucidate why models with larger capacity consistently exhibit superiority in terms of adversarial robustness compared to smaller models (Madry et al., 2017; Gustafsson et al., 2020; Li et al., 2020; Bubeck and Sellke, 2021). Moreover, given that the mean direction is aligned across different models, it is possible to approximate this mean direction by randomly sampling from a set of independently optimized models. We demonstrate that these estimated mean directions can attain a near-perfect cosine similarity of almost 1.0, even between completely independent models or ensembles, in a high-dimensional space. Moreover, note that deep ensembles essentially calculate this population mean direction (Lee et al., 2015; Lakshminarayanan et al., 2017; Fort et al., 2019; Kondratyuk et al., 2020), where the mean of a group of independently trained models can improve the performance. As a consequence, the insights of our hypotheses also shed light on this phenomenon which, although empirically successful, has been somewhat enigmatic in terms of their source of capability (Lobacheva et al., 2020; Deng and Shi, 2021; Abe et al., 2022; Theisen et al., 2023). Furthermore, since deep ensembles approximate the aligned mean directions much faster than scaling up single models, this also demystifies the significant black-box attack transferability of deep ensembles (Yang et al., 2021; Chen et al., 2023). Our research thus not only advances the understanding of model behavior in practical applications but also contributes to the broader field of AI trustworthiness and efficiency. Our main contributions can be summarized as follows:

- We reveal an appealing phenomenon where the mean distribution directions of input salience across different model families have extremely high resemblance.
- We empirically demonstrate the distribution converges towards the mean direction as model capacity increases.
- Incorporating both empirical observations and theoretical analysis, we hypothesize distributional properties of optimized models, quantifying the aforementioned phenomena.
- The hypotheses effectively explain many hitherto unclear phenomena such as black-box attack transferability, the efficacy of deep ensemble methods, etc.

2 The Convergence of Input Saliency

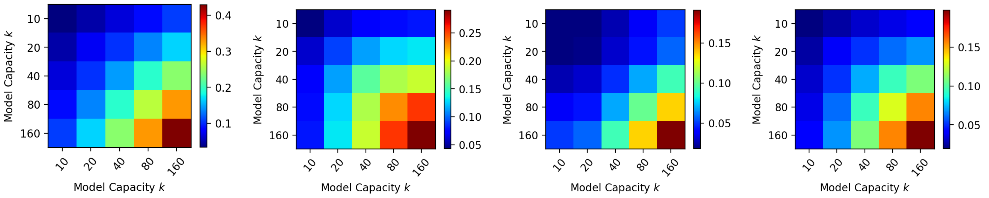
2.1 Salience Similarities

Notation. Let $\mathcal{X} \times \mathcal{Y} = \mathcal{D}$ denote the dataset, where $\mathcal{X} \subset \mathbb{R}^d$ is the input set and $\mathcal{Y} = [c]$ is the set of labels and $c \in \mathbb{N}_+$ is the number of classes. Following the benign overfitting phenomenon (Bartlett et al., 2020; Papyan et al., 2020; Cao et al., 2022), we let $\mathcal{F} = \{f | \mathcal{L}(f; \mathcal{X}_{\texttt{train}}, \mathcal{Y}_{\texttt{train}}) < 10^{-3}\}$ denote a family of optimized models, distinguished by different architectures, such as vanilla sequential CNNs, skipping blocks, etc. $\mathcal{L}$ denotes the expected cross-entropy loss for the training distribution. For simplicity, we focus on $f : \mathbb{R}^d \to \mathbb{R}$, which predicts the logit specifically for the *targeted class*. This is to stay consistent with XAI methods. We demonstrate in Appendix A that the difference between logit and probability (Wang and Wang, 2022) does not affect the observed phenomena. Unless otherwise indicated, experiments are carried out over the test set $\mathcal{X} = \mathcal{X}_{test}, \mathcal{Y} = \mathcal{Y}_{test}$. Within the same architecture, model capacity is determined by both the width and the depth. Since it is more difficult to model depth as a single variable, we model varying depth as different families $\mathcal{F}$ but model varying width k as a parameter of the family, $\mathcal{F}(k)$.



(a) CNN, CIFAR-10 (b) ResNet, CIFAR-10 (c) CNN, CIFAR-100 (d) ResNet, CIFAR-100

Figure 2: The individual similarity $\rho_{ind}(f^{(1)}, f^{(2)}) = \mathbb{E}_{\boldsymbol{x} \in \mathcal{X}}[\texttt{CosSim}(\nabla_{\boldsymbol{x}} f^{(1)}(\boldsymbol{x}), \nabla_{\boldsymbol{x}} f^{(2)}(\boldsymbol{x}))]$, where $f^{(1)} \in \mathcal{F}(k_1), f^{(2)} \in \mathcal{F}(k_2)$. CIFAR-10/100 and CNN & ResNets are tested here.



(a) CNN, CIFAR-10 (b) ResNet, CIFAR-10 (c) CNN, CIFAR-100 (d) ResNet, CIFAR-100

Figure 3: The expected similarity $\rho(k_1, k_2)$ between model families of varying capacities k_1, k_2. Here the datasets are CIFAR-10/100, and the models are CNN and ResNets.

The Increasing Similarity. Let $\texttt{CosSim} : \mathbb{R}^d \times \mathbb{R}^d \to [-1, 1]$ denote the cosine similarity metric, then the individual similarity between the input salience of two given models $f^{(1)} \in \mathcal{F}(k_1), f^{(2)} \in \mathcal{F}(k_2)$ given input $\boldsymbol{x}$ is $\rho_{ind}(f^{(1)}, f^{(2)}; \boldsymbol{x}) = \texttt{CosSim}(\nabla_{\boldsymbol{x}} f^{(1)}(\boldsymbol{x}), \nabla_{\boldsymbol{x}} f^{(2)}(\boldsymbol{x}))$. Taking the entire testing set into consideration, denote $\rho_{ind}(f^{(1)}, f^{(2)}) = \mathbb{E}_{\boldsymbol{x} \in \mathcal{X}}[\rho_{ind}(f^{(1)}, f^{(2)}; \boldsymbol{x})]$. In Figure 2, the expectations over the testing set $\mathbb{E}_{\boldsymbol{x} \in \mathcal{X}} \texttt{CosSim}(\nabla_{\boldsymbol{x}} f^{(1)}(\boldsymbol{x}), \nabla_{\boldsymbol{x}} f^{(2)}(\boldsymbol{x}))$ with varying $k_1, k_2 \in K$ are illustrated. Here, we define $K = \{j2^i : 4 \le j \le 7, 1 \le i \le 6\} = \{8, 10, 12, 14, 16, 20, \cdots, 384, 448\}$ to balance between finer linear scaling and coarser exponential scaling. It can be observed that the similarity between two stochastically optimized models $f^{(1)}, f^{(2)}$ has an increasing trend with respect to both k_1, k_2. Two different architectures CNN and ResNet are included. To rule out the influence of any single model, we define the similarity between families $\mathcal{F}(k_1), \mathcal{F}(k_2)$ for a given input $\boldsymbol{x}$ by taking the expectation of the two models:

$$\rho(k_1, k_2; \boldsymbol{x}) := \mathbb{E}_{f^{(1)} \in \mathcal{F}(k_1), f^{(2)} \in \mathcal{F}(k_2)} \texttt{CosSim}(\nabla_{\boldsymbol{x}} f^{(1)}(\boldsymbol{x}), \nabla_{\boldsymbol{x}} f^{(2)}(\boldsymbol{x})) \tag{1}$$

The global similarity between models of widths k_1, k_2 is then denoted by $\rho(k_1, k_2) = \mathbb{E}_{\boldsymbol{x} \in \mathcal{X}} \rho(k_1, k_2; \boldsymbol{x})$. Note that estimating this value requires training numerous $f \in \mathcal{F}(k)$ for each $k \in K$. Therefore, we carry out the experiments over $K' = \{j2^i : j = 5 i = 1, 2, 3, 4, 5\} = \{10, 20, 40, 80, 160\} \subset K$. For each $k_1, k_2 \in K'$, 100 models are sampled respectively to empirically estimate the expectation over $\mathcal{F}(k_1), \mathcal{F}(k_2)$. As observed in Figure 3, $\rho(k_1, k_2)$ has an increasing trend w.r.t. both k_1, k_2. Compared with Figure 3, the average similarity over the model families is similar to the individual cosine similarity for the same k values. As a result, studying the similarity of two randomly sampled models instead of the expectation over $\mathcal{F}$s can significantly alleviate the computational burden. This is further discussed in detail in Section 3.2. Besides, It trivially follows that $\forall k_1 > k_2, \rho(k_1, k_1) > \rho(k_1, k_2) > \rho(k_2, k_2)$, which suggests that *larger models tend to resemble smaller models even more than smaller models themselves.* – Even if the two smaller models only differ in the random seeds during training. Please refer to Appendix A for the results of more datasets, where such increasing trends still exist.

2.2 The Spherical Distribution of the Salience

Since the cosine similarity can be written as the inner product between $\frac{\nabla_{\boldsymbol{x}} f^{(1)}(\boldsymbol{x})}{\|\nabla_{\boldsymbol{x}} f^{(1)}(\boldsymbol{x})\|}$ and $\frac{\nabla_{\boldsymbol{x}} f^{(2)}(\boldsymbol{x})}{\|\nabla_{\boldsymbol{x}} f^{(2)}(\boldsymbol{x})\|}$, which are high-dimensional unit vectors, we explore the properties and potential distributions of $\mathcal{F}$ through the perspective of spherical statistics. Given an input $\boldsymbol{x} \in \mathcal{X}$, we denote by $\mathcal{G}_k(\boldsymbol{x})$ the set of all possible gradient directions of input $\boldsymbol{x}$ regarding the models f in $\mathcal{F}(k)$. Formally, let

$$\mathcal{G}_k(\boldsymbol{x}) = \big\{\boldsymbol{u} = \nabla_{\boldsymbol{x}} f(\boldsymbol{x}) / \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x})\| \big| \forall f \in \mathcal{F}(k)\big\}, \forall \boldsymbol{x} \in \mathcal{X} \tag{2}$$

Then the similarity is re-written as the inner product $\rho(k_1, k_2; \boldsymbol{x}) = \mathbb{E}_{\boldsymbol{u}_1 \in \mathcal{G}_{k_1}(\boldsymbol{x}), \boldsymbol{u}_2 \in \mathcal{G}_{k_2}(\boldsymbol{x})}[\boldsymbol{u}_1^T \boldsymbol{u}_2]$.

The Intra-Family vs. Cross-Family Paradox. An interesting paradox is raised as $\rho(\cdot, \cdot)$ increases with both inputs. Naturally, one would reasonably deduce that two models $f^{(1)}, f^{(2)} \in \mathcal{F}(k_1)$ should resemble each other since they are from the same family (i.e. having the exact same structure and only differ in training seeds). However, since $\rho(k_2, k_1) > \rho(k_1, k_1)$, the cross-model family similarity becomes greater than the intra-model family similarity. To uncover the mystery of the observations, we present the intuitive understanding and the rigorously analyzed hypotheses as follows.

The Intra-Family Hypothesis. Note that for intra-family scenario, $\boldsymbol{u}, \boldsymbol{v} \in \mathcal{G}_k(\boldsymbol{x})$ are i.i.d., the similarity can be written as $\rho(k, k; \boldsymbol{x}) = (\mathbb{E}_{\mathcal{G}_k(\boldsymbol{x})}[\boldsymbol{u}])^T (\mathbb{E}_{\mathcal{G}_k(\boldsymbol{x})}[\boldsymbol{v}]) = \|\mathbb{E}_{\mathcal{G}_k(\boldsymbol{x})}[\boldsymbol{u}]\|_2^2$, which denotes the square of the *population mean resultant length* (Mardia et al., 2000) of $\mathcal{G}_k(\boldsymbol{x})$. The population mean resultant length $\sqrt{\rho(k, k; \boldsymbol{x})}$ quantifies the degree of dispersion of $\mathcal{G}_k(\boldsymbol{x})$, where a larger length suggests a more concentrated distribution. In directional statistics, the degree of dispersion is usually quantified by the spherical variance $2(1 - \sqrt{\rho(k, k; \boldsymbol{x})})$ or the total variation $1 - \rho(k, k; \boldsymbol{x})$. Since $\rho(k, k; \boldsymbol{x})$ also increases w.r.t. k, this suggests an increasing concentration of input salience of models as the width k of the model increases. In conclusion, *the larger the models are, the smaller the spherical variance of the salience is.* Formally, we propose the following hypothesis.

- **Hypothesis I (H1)**: Let k denote the width (capacity) of the model and $\mathcal{G}_k(\boldsymbol{x}) = \big\{\boldsymbol{u} = \frac{\nabla_{\boldsymbol{x}} f(\boldsymbol{x})}{\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x})\|} \big| \forall f \in \mathcal{F}(k)\big\}$ denote the set of input gradient directions regarding $\boldsymbol{x}$. Then

$\mathbb{E}_{\mathcal{G}_k(\boldsymbol{x})}[\boldsymbol{u}] = \sqrt{\rho(k,k;\boldsymbol{x})}\boldsymbol{\mu}(k;\boldsymbol{x})$ and $\rho(k,k;\boldsymbol{x})$ *increases with* k. Here $\boldsymbol{\mu}(k,\boldsymbol{x}) = \frac{\mathbb{E}_{\mathcal{G}_k(\boldsymbol{x})}[\boldsymbol{u}]}{\|\mathbb{E}_{\mathcal{G}_k(\boldsymbol{x})}[\boldsymbol{u}]\|}$ denotes the unit mean direction of $\mathcal{G}_k(\boldsymbol{x})$.

Note that *H1 also holds for the change of model depths*, which is positively related to the dispersion degree of the distribution. However, the change in model depth inevitably affects model width. Thus, we only provide empirical verification in Section 3 but do not include it as a part of H2.

The Cross-Family Hypothesis. Unlike the intra-family similarity, the increasing cross-family similarity is where the phenomenon becomes counter-intuitive. Then due to the increasing intra-family similarities, when k_1, k_2 increase, $\boldsymbol{u}_1 \in \mathcal{G}_{k_1}(\boldsymbol{x})$ becomes closer to $\boldsymbol{\mu}(k_1;\boldsymbol{x})$, $\boldsymbol{u}_2 \in \mathcal{G}_{k_2}(\boldsymbol{x})$ becomes closer to $\boldsymbol{\mu}(k_2;\boldsymbol{x})$. However, the cross-familty similarities suggest that as $\boldsymbol{u}_1, \boldsymbol{u}_2$ approach their respective mean directions, their similarity increases as well. This indicates that the mean directions $\boldsymbol{\mu}(k_1;\boldsymbol{x}), \boldsymbol{\mu}(k_2;\boldsymbol{x})$ are similar, too. Formally, this intuition is considered as follows. For $k_1 > k_2$, when $\boldsymbol{\mu}(k_1;\boldsymbol{x})$ and $\boldsymbol{\mu}(k_2;\boldsymbol{x})$ are sufficiently similar, $\boldsymbol{\mu}(k_1;\boldsymbol{x})^T\boldsymbol{\mu}(k_2;\boldsymbol{x}) \approx \|\boldsymbol{\mu}(k_1;\boldsymbol{x})\|\|\boldsymbol{\mu}(k_1;\boldsymbol{x})\|$. Thus we have

$$\begin{aligned}\rho(k_1,k_2;\boldsymbol{x}) =& \mathbb{E}_{\boldsymbol{u}_1\in\mathcal{G}_{k_1}(\boldsymbol{x}),\boldsymbol{u}_2\in\mathcal{G}_{k_2}(\boldsymbol{x})}[\boldsymbol{u}_1^T\boldsymbol{u}_2] = \mathbb{E}_{\boldsymbol{u}_1\in\mathcal{G}_{k_1}(\boldsymbol{x})}[\boldsymbol{u}_1]^T\mathbb{E}_{\boldsymbol{u}_2\in\mathcal{G}_{k_1}(\boldsymbol{x})}[\boldsymbol{u}_2] \\ \approx& \|\mathbb{E}_{\boldsymbol{u}_1\in\mathcal{G}_{k_1}(\boldsymbol{x})}[\boldsymbol{u}_1]\|\|\mathbb{E}_{\boldsymbol{u}_2\in\mathcal{G}_{k_1}(\boldsymbol{x})}[\boldsymbol{u}_2]\| = \sqrt{\rho(k_1,k_1;\boldsymbol{x})\rho(k_2,k_2;\boldsymbol{x})},\end{aligned} \tag{3}$$

which is monotonic w.r.t. both k_1, k_2. Formally, this is summarized as

- **Hypothesis II (H2)**: Let $\mathcal{G}_{k_1}(\boldsymbol{x}), \mathcal{G}_{k_2}(\boldsymbol{x})$ denote the input gradient directions of two model families where $k_1 \neq k_2$. Then $\boldsymbol{\mu}(k_1;\boldsymbol{x}) \approx \boldsymbol{\mu}(k_2;\boldsymbol{x})$ regardless of k_1, k_2.

The two hypotheses are both empirically verified. For a smoother flow of the presentation, we defer the detailed experiments to Section 3. The basic ideas of H1 and H2 are illustrated in Figure 4 (a).

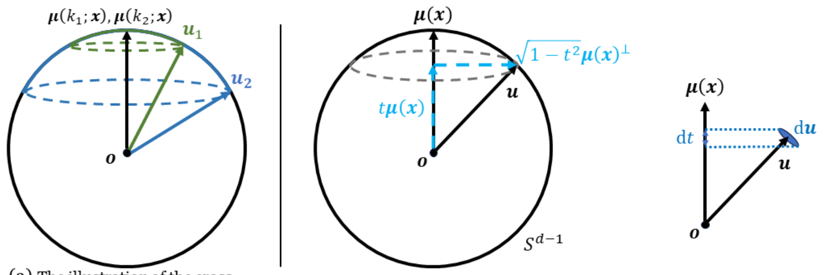
2.3 The Directional Distribution of Gradients

Given the analysis and hypothesis above, one can have an overview of the models' internal mechanisms. As the model capacity increases, models are distributed in a more concentrated manner, while the mean direction stays almost invariant. To better understand the models' behavior with the stochasticity, we delve into the distribution of $\mathcal{G}_k(\boldsymbol{x})$ and present a semi-parametric analysis with experimental verification. A general form of centralized symmetric distribution over hypersphere is known as the Saw distribution (Fisher et al., 1993) $p(\boldsymbol{u};\boldsymbol{\mu}) = \frac{\psi(\boldsymbol{u}^T\boldsymbol{\mu})}{Z}$, where $\boldsymbol{\mu}$ is the mean direction with $\|\boldsymbol{\mu}\| = 1$, $\psi \in \mathcal{C}([-1,1])$, and $Z = \int_{S^{d-1}}\psi(\boldsymbol{u}^T\boldsymbol{\mu})\mathrm{d}\boldsymbol{u}$ is the normalization term for distributions. Due to the symmetry assumption, the shape of the distribution is solely determined by ψ. For example, a monotonically increasing ψ suggests that $\boldsymbol{u}$ is distributed more densely near the mean direction and sparsely distant from the mean direction. Considering the concentration trend of gradients, we hypothesize that $\psi_k(\cdot)$ of $\mathcal{G}_k(\boldsymbol{x})$ not only monotonically increases with the input, but also increases faster with greater k values.

Marginalization. For $\forall\boldsymbol{u} \in \mathcal{G}_k(\boldsymbol{x})$, it can be decomposed to $\boldsymbol{u} = t\cdot\boldsymbol{\mu}(\boldsymbol{x}) + \sqrt{1-t^2}\cdot\boldsymbol{\mu}(\boldsymbol{x})^{\perp}$, where $\boldsymbol{\mu}(\boldsymbol{x})^{\perp}$ is a unit tangent to S^{d-1} at $\boldsymbol{\mu}$. Then $t = \boldsymbol{u}^T\boldsymbol{\mu}(\boldsymbol{x})$. This is shown in Figure 4 (b). Note that $\boldsymbol{\mu}(\boldsymbol{x})^{\perp}$ is independent from t, then the distribution of t is the marginal distribution over the intersection between S^{d-1} and the hyperplane spanned by $\boldsymbol{\mu}(\boldsymbol{x})^{\perp}$, which is a $(d-2)$-dimensional hypersphere. According to the symmetry assumption of Saw distribution, conditioned on a fixed similarity t, the distribution of $\boldsymbol{u}|t$ over the dashed S^{d-2} does not affect ψ. Therefore, we focus on the marginalized distribution of t. Note that the radius of the intersection S^{d-2} is $\sqrt{1-t^2}$, we thus have $\mathrm{d}\boldsymbol{u} = \frac{2\pi^{(d-1)/2}(1-t^2)^{(d-3)/2}}{\Gamma((d-1)/2)}\mathrm{d}t$, where the density of t is observed by the integral over the corresponding $(d-2)$-hypersphere. As a result, the marginal distribution of t has the PDF

$$p_k(t;\boldsymbol{x}) = \psi_k(t;\boldsymbol{x})(1-t^2)^{(d-3)/2}/Z' \tag{4}$$

where Z' is a constant normalization term. Note that $(1-t^2)^{(d-3)/2}$ is a symmetric bell curve centered at $t = 0$. Equation (4) can thus be viewed as using $\psi_k(t;\boldsymbol{x})$ to reweight its PDF $p_{\text{origin}}(t) = (1-t^2)^{(d-3)/2}\frac{\Gamma(d/2)}{\sqrt{\pi}\Gamma((d-1)/2)}$. Note that here $p_k(t;\boldsymbol{x})$ is the distribution of $t = \boldsymbol{u}^T\boldsymbol{\mu}(k;\boldsymbol{x})$, which can be empirically estimated, the shape of the function ψ_k becomes empirically accessible with varying k values. The empirical studies and verification are provided in Section 3.3.


$\boldsymbol{\mu}(k_1;\boldsymbol{x}), \boldsymbol{\mu}(k_2;\boldsymbol{x})$
$\boldsymbol{u}_1$
$\boldsymbol{u}_2$
$\boldsymbol{o}$
$\boldsymbol{\mu}(\boldsymbol{x})$
$\sqrt{1-t^2}\boldsymbol{\mu}(\boldsymbol{x})^{\perp}$
$\boldsymbol{u}$
$t\boldsymbol{\mu}(\boldsymbol{x})$
$\boldsymbol{o}$
S^{d-1}
$\boldsymbol{\mu}(\boldsymbol{x})$
$\mathrm{d}t$
$\mathrm{d}\boldsymbol{u}$
$\boldsymbol{u}$
$\boldsymbol{o}$

(a) The illustration of the cross-family distributions ($k_1 > k_2$).

(b) The decomposition (left) and marginalization (right) of $\boldsymbol{u} \in \mathcal{G}_k(\boldsymbol{x})$.

Figure 4: (a) presents an illustration of H1 and H2. Blue and green caps represent $\boldsymbol{u}_1 \in \mathcal{G}_{k_1}(\boldsymbol{x})$ (green) and $\boldsymbol{u}_2 \in \mathcal{G}_{k_2}(\boldsymbol{x})$ (blue) regions [2]. H1: larger ks lead to smaller spherical variances; H2: the mean directions are extremely similar. (b) illustrates (left) the decomposition of $\boldsymbol{u}$ to the mean direction and the orthogonal direction; and (right) the marginalization of the distribution from $\boldsymbol{u}$ to t.

3 Empirical Verification of Hypotheses

In this section, we provide comprehensive experimental results to verify the aforementioned hypotheses. First, we introduce the detailed setups of our experiments. They are carried out on Intel(R) Core(TM) i9-9960X CPU @ 3.10GHz with Quadro RTX 6000 GPUs.

Datasets & Models. Due to the massive size of experiments, here we mainly follow the setups of the benign overfitting (Nakkiran et al., 2021), which also present a comprehensive study of optimized DNNs through CIFAR-10 and CIFAR-100 (Krizhevsky et al., 2009). Besides, we also include TinyImagenet-200 (Le and Yang, 2015) as a compromise between the computational efficiency and the dataset complexity. As for models, we include CNNs and ResNets as in (Nakkiran et al., 2021). These two models represent the two typical architectures – the direction connection and the skip connection. We also notice a striking capacity gap between them in the original implementation. Therefore, we term them CNNSmall (CS) and ResNetLarge (RL), respectively, and include CNNLarge (CL), and ResNetSmall (RS) to fill the gap. The comparison between the small and large families also shows the influence of depths. As for the training process, following Nakkiran et al. (2021), we use stochastic gradient descent (SGD) as the solver, with a batch size of 128. The input data are normalized, but not augmented. We start with the initial learning rate $\gamma_0 = 0.1$ and update it with $\gamma_t = \gamma_0/\sqrt{1+t}$, where t is the epoch. Please refer to Appendix B for more experimental details.

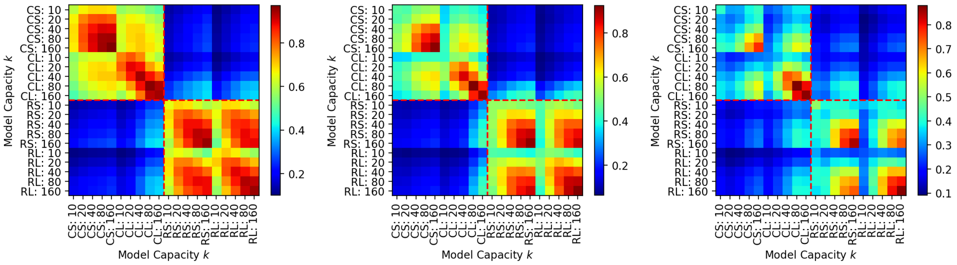
3.1 The Mean Direction Similarity (H2)

H2 can be verified independently from H1 and can provide simplifications and insights to verifying H1. We hence focus on H2 first. As stated in H2, the mean directions of different model families are consistently aligned, i.e., $\boldsymbol{\mu}(k_1;\boldsymbol{x}) \approx \boldsymbol{\mu}(k_2;\boldsymbol{x})$. For each family $\mathcal{F}(k)$ where $k \in K'$, $M = 100$ models with different random seeds are trained. The population mean is then estimated through

$$\tilde{\boldsymbol{\mu}}(k;\boldsymbol{x}) = \Big(\frac{1}{M}\sum_{i=1}^{M}\frac{\nabla_{\boldsymbol{x}} f_i(\boldsymbol{x})}{\|\nabla_{\boldsymbol{x}} f_i(\boldsymbol{x})\|}\Big)\Big/\Big\|\frac{1}{M}\sum_{i=1}^{M}\frac{\nabla_{\boldsymbol{x}} f_i(\boldsymbol{x})}{\|\nabla_{\boldsymbol{x}} f_i(\boldsymbol{x})\|}\Big\| \approx \boldsymbol{\mu}(k;\boldsymbol{x}), f_i \in \mathcal{F}(k). \tag{5}$$

Then the similarity of mean directions are naturally $(\tilde{\boldsymbol{\mu}}(k_1;\boldsymbol{x}))^T\tilde{\boldsymbol{\mu}}(k_2;\boldsymbol{x})$. Note that when $k_1 = k_2$, the 100 models are partitioned by the seeds to avoid overlapping. The results of the expectation over $\mathcal{X}$ are visualized in Figure 5. It can be found that the mean directions have extremely high resemblance within each architecture, as proved by the high cosine similarities. It should be noted that with high dimensionality (e.g. $d = 3072$ for CIFAR), a cosine similarity close to 1 is an extremely significant result. We demonstrate this with the uniform distribution on the hypersphere in Appendix D. The observations verify the hypothesis all $\mathcal{G}_k(\boldsymbol{x})$ almost share the same mean direction within the model architecture. This not only hold across different widths determined by k, but also holds across different depths (i.e. CS vs. CL, RS vs. RL). Therefore, the mean direction is mostly related to the certain model architecture instead of any single model $f \in \mathcal{F}$, making it an intrinsic

[2] The caps are to illustrate the variance differences. Actual distributions are over the entire hypersphere.



(a) CIFAR-10 (b) CIFAR-100 (c) TinyImagenet

Figure 5: The heatmap visualization between the estimated population mean directions from different model families. Each entry is computed by $\mathbb{E}_{\boldsymbol{x}\in\mathcal{X}}\mathtt{CosSim}(\tilde{\boldsymbol{\mu}}_j(\boldsymbol{x};\mathcal{F}),\tilde{\boldsymbol{\mu}}_{j'}(\boldsymbol{x};\mathcal{F}'))$ The results are generated from CIFAR-10/100 and TinyImagenet datasets. $\mathcal{F}\in$ {CS, CL, RS, RL}.

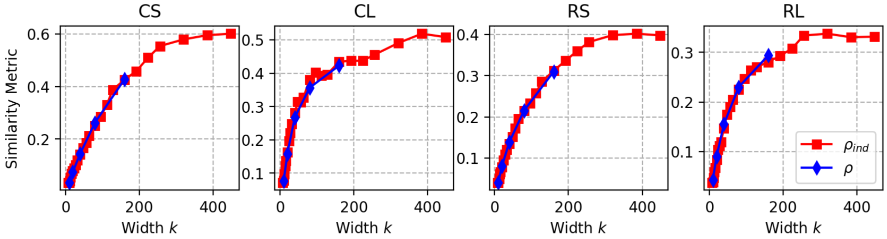


Figure 6: Illustration of (red) $\rho_{ind}(f^{(1)},f^{(2)}), f^{(1)},f^{(2)}\in\mathcal{F}(k)$ and (blue) $\rho(k,k)$ on CIFAR-10.

property of the model architecture. With this property, how different model architectures differ in mechanisms can be studied by looking deeper into the population mean direction of saliency maps. For instance, it can be observed that the ResNet architecture admits a closer relation between models of different depths, compared with the CNN architecture.

3.2 The Decreasing Spherical Variance with k (H1)

Expectation over $\mathcal{F}$ vs. Conditioned on $f\in\mathcal{F}$. As previously discussed, the spherical variance of distributions over the hypersphere can be measured by the population mean resultant length $\sqrt{\rho(k,k;\boldsymbol{x})}$, which, unfortunately, requires an estimation of the mean directions. This can be expensive to study for a comprehensive set K of k values. The experiments on a subset $K'=\{10,20,40,80,160\}$ are already carried out in Figure 3, shown as the diagonal elements. As k increases, the resultant length increases monotonically, indicating a decreasing spherical variance and a more concentrated distribution around the mean directions.

The computational burden of taking the expectation over $\mathcal{F}$ can be alleviated by considering randomly picked f. In order to compare ρ and ρ_{ind}, we consider the model-dependent set $\mathcal{S}(f)=\{\rho_{ind}(f,f;\boldsymbol{x}):\boldsymbol{x}\in\mathcal{X}\}$ for each $f\in\mathcal{F}$. Here we compute the expected Wasserstein distance $\mathbb{E}_{f^{(1)},f^{(2)}\in\mathcal{F}(k)}\mathtt{WD}(\mathcal{S}(f^{(1)}),\mathcal{S}(f^{(2)}))$. This is estimated by the $\binom{M}{2}$ distinct pairs of models. The distances of all 60 (dataset, model family) pairs lie below 0.035. Such observation suggests that after taking the expectation over $\mathcal{X}$, the differences across individual models can be mitigated. Please refer to Table 1 for comprehensive results on all model families and datasets. As a consequence, it suffices to use $\rho_{ind}(f,f)$ for some $f\in\mathcal{F}(k)$ to approximate $\rho(k,k)$. This is in fact the diagonal elements of Figure 2. A comparison between the diagonal elements $\rho_{k,k}$ and $\rho_{ind}(f^{(1)},f^{(2)}), f^{(1)},f^{(2)}\in\mathcal{F}(k)$ over CIFAR-10 is presented in Figure 6. Please refer to the appendix for other datasets. ρ_{ind} is evaluated over $k\in K$, while ρ is evaluated over $k\in K'\subset K$. It can be found that after taking the average over $\mathcal{X}$, even though ρ is a little smoother than ρ_{ind}, they are very consistent. This verifies that the resultant length increases with $k\in K$ in a much more comprehensive set of models. Thus H1 is empirically verified.

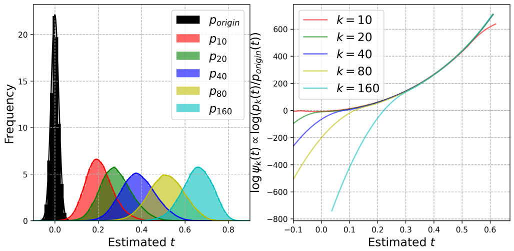


Figure 8: (left) The illustration of the frequency of the mixture $\mathcal{T}_k$, where $k \in \{10, 20, 40, 80, 160\}$. Specifically, the black histogram represents the distribution p_{origin}. The dashed curves are the approximated PDF p_k obtained by KDE. The results are generated using CNNSmall and CIFAR-10. (right) The illustration of $\log \frac{p_k}{p_{\text{origin}}}$, which is linearly related to $\log \psi_k$.

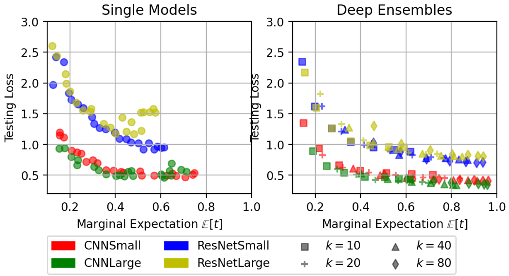


Figure 9: The illustration of the relation between the expected testing loss $\mathbb{E}_{\mathcal{X}}[\mathcal{L}]$ and the marginal expectation $\mathbb{E}_{\mathcal{X}}[t]$, where models are from (a) single models with varying capacities (b) deep ensembles with varying member #. Each color represents a model family. In particular, in (b), different marker shapes indicate different $k \in [10, 20, 40, 80]$ of the ensembles.

3.3 The Shape of the Saw Distribution

Given $\mathcal{G}_k(\boldsymbol{x})$, the marginalized distribution $p_k(t; \boldsymbol{x})$ can be approximated by $\mathcal{T}_k(\boldsymbol{x}) = \{\tilde{t} = \boldsymbol{u}^T \tilde{\boldsymbol{\mu}}(k; \boldsymbol{x}) | \boldsymbol{u} \in \mathcal{G}_k(\boldsymbol{x})\}$. In order to obtain the global results over test dataset $\mathcal{X}$, consider the unions of different input samples $\mathcal{T}_k = \cup_{\boldsymbol{x} \in \mathcal{X}} \mathcal{T}_k(\boldsymbol{x})$. This is an approximation to the mixture distributions $p_k(t) = \frac{1}{|\mathcal{X}|} \sum_{\boldsymbol{x} \in \mathcal{X}} p_k(t; \boldsymbol{x})$. We plot the histogram of $\mathcal{T}_k$ with $k \in K'$ for CNNSmall and CIFAR-10 in Figure 8. The left figure shows p_{origin} and the estimated p_k, visualized by different colors. Qualitatively, $p_k(t)$ has higher means with greater k values. The reason why the density of p_k is not centered at $t = 1$ (i.e. $\boldsymbol{u} = \boldsymbol{\mu}$) is because as t increases, the size of the $(d-2)$-hypershpere decreases with $p_{\text{origin}}(t) = (1 - t^2)^{(d-3)/2} \frac{\Gamma(d/2)}{\sqrt{\pi}\Gamma((d-1)/2)}$, which is shown as black histograms. This is much faster than the increase of ψ_k. The shape of ψ_k is determined by p_k / p_{origin} with normalization. From the right figure, by comparing p_k with $p_{\texttt{PDF}}$, it can be empirically verified that $\psi_k(t)$ is increasing vastly. It is also observed that larger ks lead to a faster increase of ψ_k and higher $\mathbb{E}[t]$. This also provides a quantitative understanding of H1 and H2. The results of other model architectures and datasets can be found in the appendix.

Verification of the Symmetry. Saw distributions study the marginalized value $t = \boldsymbol{u}^T \boldsymbol{\mu}$ to directly focus on the degree of concentration of the gradients. This naturally leads to rotationally symmetric distributions since the distribution on the intersection between S^{d-1} and the hyperplane does not affect the distribution of t. We thus carry out an empirical study of the distribution on the intersection (i.e. conditioned on t). Specifically, we train 1000 CNN models with $K = 40$ and seeds 1-1000 on CIFAR-10 and compute t regarding each test sample. The distribution of the first sample is visualized in Figure 7. We partition the range of t into 10 intervals by every 10 percent of the frequency, and inspect the direction of the mean of the gradients in each interval, each direction is estimated by 100 models. If these conditional mean directions are consistent with the population mean direction, then the gradients are symmetrically distributed on each S^{d-2} hypersphere (R7(right)), thus verifying the rotational symmetry. We investigate the cosine similarities between the conditional and unconditional mean directions on the first 1000 samples. The 10×1000 similarity values have a mean and std at approximately 0.970 and 0.013 respectively. Thus the rotational symmetry is empirically verified.

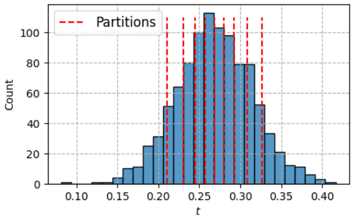
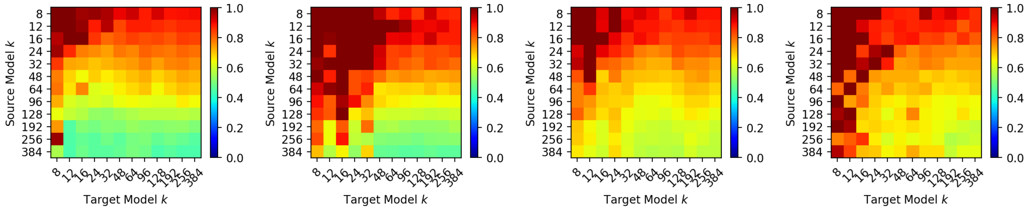
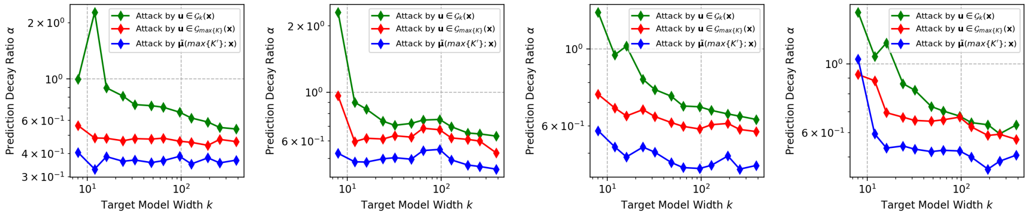


Figure 7: The marginal distribution of t of the first test sample of CIFAR-10. Red dashed lines partition the range of t every 10 percent of the frequency.



(a) CIFAR-10, CS (b) CIFAR-10, CL (c) CIFAR-10, RS (d) CIFAR-10, RL

Figure 10: The results of single model black-box attack. The value of each entry is $\alpha(k_1, k_2)$ for different model capacities, where k_1 is the width parameter of the source model and k_2 is the width parameter of the target model.



(a) CIFAR-10, CS (b) CIFAR-10, CL (c) CIFAR-10, RS (d) CIFAR-10, RL

Figure 11: The comparison between the single-model attack from the largest model (red), the single-model attack from the very same capacity (green) and the attack by the mean direction (blue).

4 Applications of Hypotheses

4.1 Deep Ensemble: Why Does It Work?

After verifying the hypotheses, we explore possible applications and implications of the discovered phenomena. The deep ensemble method makes use of the stochasticity to of models by incorporating the predictions from m members. While deep ensembles have been verified to be effective in improving performance, the source of such capability remains mysterious. Note that ensemble members are i.i.d. optimized models with SGD, which correspond to the population mean of our hypothesis. We thus provide another perspective in understanding the capability of deep ensembles.

For single models, as the model capacity increases, benign overfitting suggests that the testing loss decreases, too. We deduce that this is because the distribution of larger models becomes more concentrated, and combined with H2, the closeness to the aligned population mean is highly related to the models' testing performance. As shown in Figure 9(a), it can be observed that the expected loss $\mathbb{E}_{\mathcal{X}}[\mathcal{L}]$ and the marginal expectation $\mathbb{E}_{\mathcal{X}}[t]$ are highly correlated. Similarly, deep ensemble approximates the population mean much more effectively by increasing the number m of members than scaling up a single model by k. We thus scale up the deep ensemble by changing the number of ensemble members. The results are shown in Figure 9(b). It can be found that (1) the correlation between $\mathbb{E}_{\mathcal{X}}[\mathcal{L}]$ and $\mathbb{E}_{\mathcal{X}}[t]$ is not only significant, and (2) the correlation pattern is shared between two completely different scaling mechanisms, single model scaling and model ensembles.

4.2 Black-Box Attack via Saliency Similarity

The understanding of adversarial attacks can benefit from the behaviors of the input salience of optimized models given their close relation to input gradients. We verify the aforementioned similarities through the black-box attacks, where the adversarial samples are generated from the gradients of source models while the gradients of the target models are not available. Let $f^{(1)} \in \mathcal{F}(k_1)$ denote the source model and $f^{(2)} \in \mathcal{F}(k_2)$ denote the target model. We define the attack rate from

$f^{(1)}$ to $f^{(2)}$ similar to ρ_{ind} as

$$\alpha(f^{(1)}, f^{(2)}) = \mathop{\mathbb{E}}_{\boldsymbol{x}\in\mathcal{X}}\Big[f^{(2)}\big(\boldsymbol{x} - \epsilon \cdot \texttt{sign}(\nabla_{\boldsymbol{x}} f^{(1)}(\boldsymbol{x}))\big)/f^{(2)}(\boldsymbol{x})\Big] \tag{6}$$

which is the performance decay of $f^{(2)}$ when attacked by model $f^{(1)}$. Small $\alpha(k_1, k_2)$ values suggest successful attack from $f^{(1)}$ to $f^{(2)}$. The results are shown in Figure 10, where the attack step is set to $\epsilon = 0.05$. In each heatmap figure, the y-axis represents the width of the source models, while the x-axis represents the width of the target models. It can be observed that larger models succeed in attacking smaller models, but the opposite is not true. To attack a large model, the gradient needs to be generated from a model of a comparable level in terms of capacity.

Mean Direction Attack. According to the verified hypothesis H2, for any two individual models $f^{(1)} \in \mathcal{F}(k_1), f^{(2)} \in \mathcal{F}(k_2)$ the mean directions $\boldsymbol{\mu}(k;\boldsymbol{x})$ is closer to both of them than themselves, regardless of k, k_1, k_2. It is then suggested that using the mean gradient can perform more successful black-box attacks. We employ the mean direction $\tilde{\boldsymbol{\mu}}(160, \boldsymbol{x})$ to attack models of different capacities, and compare the results with the attack from the largest single model (red) and the attack from the models of the identical structure (green). The results are shown in Figure 11, where it can be observed that the mean direction of salience transfers much more successfully than single models.

5 Conclusions

In this paper, we introduce hypotheses to explain the observations on the input salience convergence w.r.t. the model capacities. Under the same model architecture, stochastic algorithms such as SGD, result in certain distributions of optimized models. We hypothesize and use pointwise methods to verify that such distribution follows a Saw distribution with **aligned population means**, which is invariant from the model families. Besides, the variance of the distribution decreases as the model capacity increases, suggesting a convergence trend of the models' internal mechanism – the larger the models are, the less variant they tend to be affected by the randomness from the stochastic algorithm during the training phase. Furthermore, since the distributions converge towards the aligned population mean direction, the limiting points can be estimated by the population mean of models. Based on this, we present comprehensive experiments on the properties of the limiting model and demonstrate its capability in various domains, such as the black-box attack transferability, and the explanation of the effectiveness of deep ensembles. However, it is admitted that, due to the high computational burden, although improved from CIFAR-10/100 to TinyImagenet compared to (Nakkiran et al., 2021), our experiments are limited to rather small datasets.

Our introduced hypotheses also lead to various interesting topics. Note that the aligned mean direction stays invariant to the model families, which indicates such population mean is more related to the essence of the dataset itself rather than any single model. Leveraging this property can bring a deeper and more comprehensive understanding of the relation between data distributions and models.

6 Related Work

In terms of the convergence trend of DNNs, existing works focus on the convergence of single models throughout the training process. The parameters of DNNs have been demonstrated to converge to global minima throughout the training progress (Goodfellow et al., 2014b; Li et al., 2018; Allen-Zhu et al., 2019; Liu et al., 2020; Damian et al., 2021; Refinetti et al., 2023; Suh and Cheng, 2024). Recent years, the studies of benign overfitting also suggest that increasing model capacities can improve the performance instead of exacerbating the overfitting issue (Bartlett et al., 2020; Nakkiran et al., 2021; Cao et al., 2022). While the studies of input gradients span into an abundant but extremely complicated spectrum. Among them, the area that is the most related to our work is the XAI domain, where the input gradient and its variants are crucial in revealing the models' internal mechanisms (Simonyan et al., 2013; Springenberg et al., 2014; Selvaraju et al., 2017; Sundararajan et al., 2017; Adebayo et al., 2018; Shah et al., 2021). On the other hand, the studies of the distribution of optimized models have received little attention. Such topics are slightly dipped in the efforts to demystify the source of capability of deep ensembles (Lee et al., 2015; Fort et al., 2019; Allen-Zhu and Li, 2020; Kobayashi et al., 2021; Abe et al., 2022; Ganaie et al., 2022; Theisen et al., 2023) and their implications (Lakshminarayanan et al., 2017; Geiger et al., 2020; Yang et al., 2021; Chen et al., 2023). Thus, to our knowledge, the studies on the distribution of optimized models remain a novel topic.

Acknowledgement

This work was supported, in part, by the Defense Advance Research Projects Agency (Prime contract award number: HR0011222003, subcontract award number: 2103299-01, grant number: 13001129), the EMBRIO Institute, contract #2120200, a National Science Foundation (NSF) Biology Integration Institute, and NSF IIS #1955890, IIS #2146091, IIS #2345235. The content of the information does not necessarily reflect the position of the US Government. No official endorsement should be inferred. Approved for public release; distribution is unlimited.

References

Abe, T., Buchanan, E. K., Pleiss, G., Zemel, R., and Cunningham, J. P. (2022). Deep ensembles work, but are they necessary? *Advances in Neural Information Processing Systems*, 35:33646–33660.

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., and Kim, B. (2018). Sanity checks for saliency maps. *Advances in neural information processing systems*, 31.

Allen-Zhu, Z. and Li, Y. (2020). Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. *arXiv preprint arXiv:2012.09816*.

Allen-Zhu, Z., Li, Y., and Song, Z. (2019). A convergence theory for deep learning via over-parameterization. In *International conference on machine learning*, pages 242–252. PMLR.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al. (2020). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115.

Bartlett, P. L., Long, P. M., Lugosi, G., and Tsigler, A. (2020). Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070.

Berner, C., Brockman, G., Chan, B., Cheung, V., Dębiak, P., Dennison, C., Farhi, D., Fischer, Q., Hashme, S., Hesse, C., et al. (2019). Dota 2 with large scale deep reinforcement learning. *arXiv preprint arXiv:1912.06680*.

Bubeck, S. and Sellke, M. (2021). A universal law of robustness via isoperimetry. *Advances in Neural Information Processing Systems*, 34:28811–28822.

Cao, Y., Chen, Z., Belkin, M., and Gu, Q. (2022). Benign overfitting in two-layer convolutional neural networks. *Advances in neural information processing systems*, 35:25237–25250.

Chen, H., Zhang, Y., Dong, Y., and Zhu, J. (2023). Rethinking model ensemble in transfer-based adversarial attacks. *arXiv preprint arXiv:2303.09105*.

Damian, A., Ma, T., and Lee, J. D. (2021). Label noise sgd provably prefers flat global minimizers. *Advances in Neural Information Processing Systems*, 34:27449–27461.

Deng, D. and Shi, E. B. (2021). Ensembling with a fixed parameter budget: When does it help and why? In *Asian Conference on Machine Learning*, pages 1176–1191. PMLR.

Du, S., Lee, J., Li, H., Wang, L., and Zhai, X. (2019). Gradient descent finds global minima of deep neural networks. In *International conference on machine learning*, pages 1675–1685. PMLR.

Fisher, N. I., Lewis, T., and Embleton, B. J. (1993). *Statistical analysis of spherical data*. Cambridge university press.

Fort, S., Hu, H., and Lakshminarayanan, B. (2019). Deep ensembles: A loss landscape perspective. *arXiv preprint arXiv:1912.02757*.

Ganaie, M. A., Hu, M., Malik, A., Tanveer, M., and Suganthan, P. (2022). Ensemble deep learning: A review. *Engineering Applications of Artificial Intelligence*, 115:105151.

Geiger, M., Jacot, A., Spigler, S., Gabriel, F., Sagun, L., d'Ascoli, S., Biroli, G., Hongler, C., and Wyart, M. (2020). Scaling description of generalization with number of parameters in deep learning. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(2):023401.

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

Goodfellow, I. J., Vinyals, O., and Saxe, A. M. (2014b). Qualitatively characterizing neural network optimization problems. *arXiv preprint arXiv:1412.6544*.

Gunning, D. and Aha, D. (2019). Darpa's explainable artificial intelligence (xai) program. *AI magazine*, 40(2):44–58.

Gustafsson, F. K., Danelljan, M., and Schon, T. B. (2020). Evaluating scalable bayesian deep learning methods for robust computer vision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 318–319.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Huang, G., Li, Y., Pleiss, G., Liu, Z., Hopcroft, J. E., and Weinberger, K. Q. (2017). Snapshot ensembles: Train 1, get m for free. *arXiv preprint arXiv:1704.00109*.

Kobayashi, S., von Oswald, J., and Grewe, B. F. (2021). On the reversed bias-variance tradeoff in deep ensembles. ICML.

Kondratyuk, D., Tan, M., Brown, M., and Gong, B. (2020). When ensembling smaller models is more efficient than single large models. *arXiv preprint arXiv:2005.00570*.

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

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30.

Le, Y. and Yang, X. (2015). Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3.

Lee, S., Purushwalkam, S., Cogswell, M., Crandall, D., and Batra, D. (2015). Why m heads are better than one: Training a diverse ensemble of deep networks. *arXiv preprint arXiv:1511.06314*.

Li, H., Xu, Z., Taylor, G., Studer, C., and Goldstein, T. (2018). Visualizing the loss landscape of neural nets. *Advances in neural information processing systems*, 31.

Li, Y., Bai, S., Zhou, Y., Xie, C., Zhang, Z., and Yuille, A. (2020). Learning transferable adversarial examples via ghost networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11458–11465.

Liu, S., Papailiopoulos, D., and Achlioptas, D. (2020). Bad global minima exist and sgd can reach them. *Advances in Neural Information Processing Systems*, 33:8543–8552.

Lobacheva, E., Chirkova, N., Kodryan, M., and Vetrov, D. P. (2020). On power laws in deep ensembles. *Advances In Neural Information Processing Systems*, 33:2375–2385.

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

Mardia, K. V., Jupp, P. E., and Mardia, K. (2000). *Directional statistics*, volume 2. Wiley Online Library.

Muller, M. E. (1959). A note on a method for generating points uniformly on n-dimensional spheres. *Communications of the ACM*, 2(4):19–20.

Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B., and Sutskever, I. (2021). Deep double descent: Where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124003.

Padmaja, B., Moorthy, C. V., Venkateswarulu, N., and Bala, M. M. (2023). Exploration of issues, challenges and latest developments in autonomous cars. *Journal of Big Data*, 10(1):61.

Papyan, V., Han, X., and Donoho, D. L. (2020). Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663.

Preece, A., Harborne, D., Braines, D., Tomsett, R., and Chakraborty, S. (2018). Stakeholders in explainable ai. *arXiv preprint arXiv:1810.00184*.

Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., Liu, P. J., Liu, X., Marcus, J., Sun, M., et al. (2018). Scalable and accurate deep learning with electronic health records. *NPJ digital medicine*, 1(1):18.

Refinetti, M., Ingrosso, A., and Goldt, S. (2023). Neural networks trained with sgd learn distributions of increasing complexity. In *International Conference on Machine Learning*, pages 28843–28863. PMLR.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

Rudin, C. (2018). Please stop explaining black box models for high stakes decisions. *Stat*, 1050:26.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626.

Shah, H., Jain, P., and Netrapalli, P. (2021). Do input gradients highlight discriminative features? *Advances in Neural Information Processing Systems*, 34:2046–2059.

Simonyan, K., Vedaldi, A., and Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Smith, I., Ortmann, J., Abbas-Aghababazadeh, F., Smirnov, P., and Haibe-Kains, B. (2023). On the distribution of cosine similarity with application to biology. *arXiv preprint arXiv:2310.13994*.

Springenberg, J. T., Dosovitskiy, A., Brox, T., and Riedmiller, M. (2014). Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*.

Suh, N. and Cheng, G. (2024). A survey on statistical theory of deep learning: Approximation, training dynamics, and generative models. *arXiv preprint arXiv:2401.07187*.

Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. (2013). Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Theisen, R., Kim, H., Yang, Y., Hodgkinson, L., and Mahoney, M. W. (2023). When are ensembles really effective? *arXiv preprint arXiv:2305.12313*.

Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., and Ting, D. S. W. (2023). Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

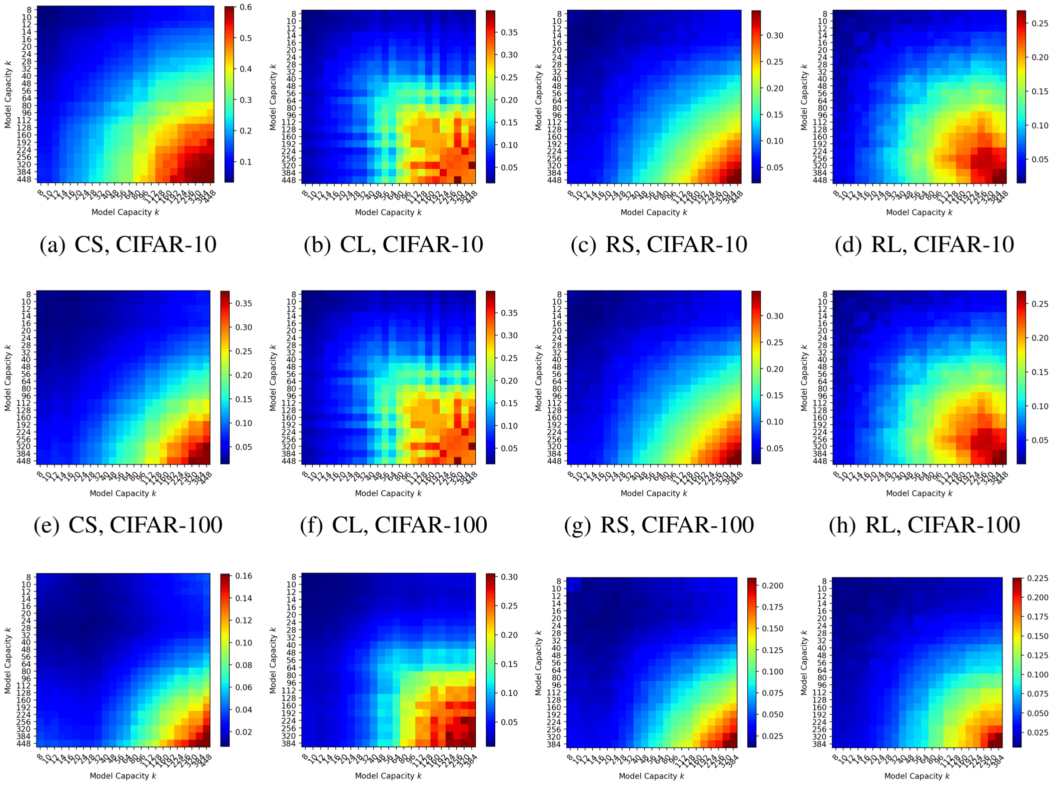
Van der Velden, B. H., Kuijf, H. J., Gilhuijs, K. G., and Viergever, M. A. (2022). Explainable artificial intelligence (xai) in deep learning-based medical image analysis. *Medical Image Analysis*, 79:102470.

Wang, Y. and Wang, X. (2022). “why not other classes?”: Towards class-contrastive back-propagation explanations. *Advances in Neural Information Processing Systems*, 35:9085–9097.

Yang, Z., Li, L., Xu, X., Zuo, S., Chen, Q., Zhou, P., Rubinstein, B., Zhang, C., and Li, B. (2021). Trs: Transferability reduced ensemble via promoting gradient diversity and model smoothness. *Advances in Neural Information Processing Systems*, 34:17642–17655.

A Additional Results of the Similarities

The illustration of CIFAR-100 & TinyImagenet. In addition to the cosine similarity $\rho_{ind}(k_1, k_2)$ shown in Figure 2, we include more sophisticated datasets such as CIFAR-100 and TinyImagenet in Figure 12. It is observed that the observed phenomena, where $\rho_{ind}(k_1, k_2)$ tend to increase w.r.t. both k_1, k_2, hold across different datasets. It is also worth noticing that compared with the results of CIFAR-10 shown in Figure 2, the resulting $\rho_{ind}(k_1, k_2)$ of CIFAR-100 and TinyImagnet (Figure 12) shows another peak at around $k \approx 10$ (see Figure 12(e)). This is related to the deep double descent phenomenon (Nakkiran et al., 2021), where when the complexities of the model and the dataset are comparable, the overfitting issue is at peak. For smaller such as CIFAR-10 or larger models such as ResNet, this issue becomes much less significant, as a very small k value is already sufficient for the data distribution. Also, as the complexity of the data distribution increases, the cosine similarity decreases inevitably for the same model complexity. This suggests a less concentrated distribution of the optimized models of the same capacity compared with less complex dataset. Correspondingly, the results of the expectation over the model sets for $k \in K' = [10, 20, 40, 80, 160]$ are shown in Figure 13.



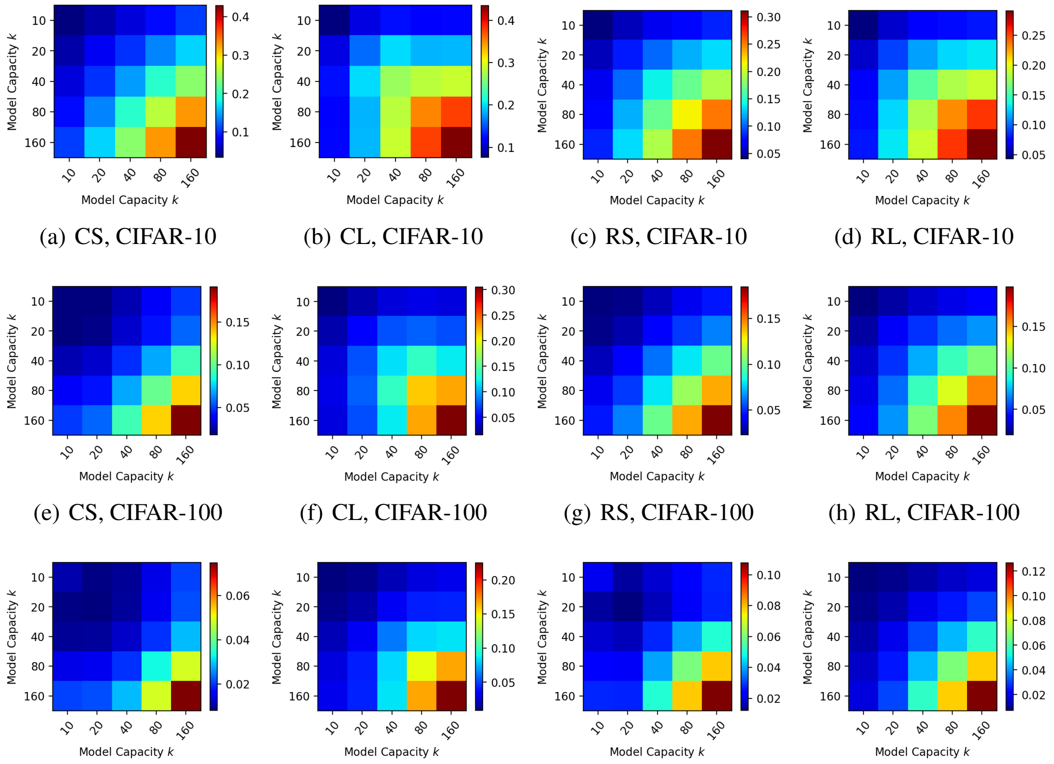
(a) CS, CIFAR-10 (b) CL, CIFAR-10 (c) RS, CIFAR-10 (d) RL, CIFAR-10

(e) CS, CIFAR-100 (f) CL, CIFAR-100 (g) RS, CIFAR-100 (h) RL, CIFAR-100

(i) CS, TinyImagenet (j) CL, TinyImagenet (k) RS, TinyImagenet (l) RL, TinyImagenet

Figure 12: The expected similarity $\rho(k_1, k_2)$ between models of varying widths k_1, k_2. Here we include CNNSmall, CNNLarge, ResNetSmall, and ResNetLarge as $\mathcal{F}$. The values of k_1, k_2 determine the widths in each layer. Here the datasets are CIFAR-10 (top), CIFAR-100 (middle) and tinyImagenet (bottom).

The Ablation of the Training Process. To verify that the observed increasing trends of $\rho_{ind}(k_1, k_2)$ with model capacities are caused by the training process of DNNs instead of some normalization issue, we compare the similarity for models with initialized parameters. The results are shown in Figure 14. All three datasets and four model families are included. It can be clearly observed that, when the model parameters are initialized, the similarity between input saliency maps of different models are distributed randomly. The cosine similarity values are very concentrated around 0, which is the mean of random distribution. This verifies that the aforementioned increasing trends are caused by the optimization of models instead of normalization process.



(a) CS, CIFAR-10 (b) CL, CIFAR-10 (c) RS, CIFAR-10 (d) RL, CIFAR-10

(e) CS, CIFAR-100 (f) CL, CIFAR-100 (g) RS, CIFAR-100 (h) RL, CIFAR-100

(i) CS, TinyImagenet (j) CL, TinyImagenet (k) RS, TinyImagenet (l) RL, TinyImagenet

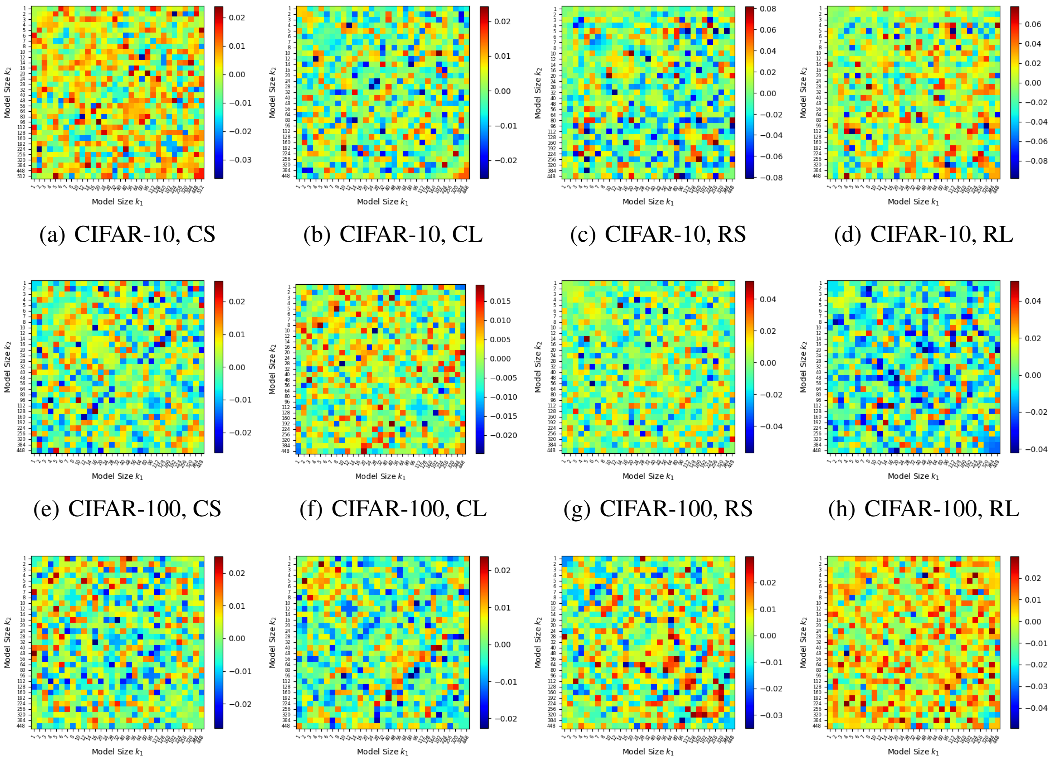
Figure 13: The expected similarity $\rho(k_1, k_2)$ between models of varying widths k_1, k_2. Here we include CNNSmall, CNNLarge, ResNetSmall, and ResNetLarge as $\mathcal{F}$. The values of $k_1, k_2 \in [10, 20, 40, 80, 160]$ determine the widths in each layer. Here the datasets are CIFAR-10 (top), CIFAR-100 (middle), and tinyImagenet (bottom).

Softmax Activations. Apart from the normalization concern, recent work in (Wang and Wang, 2022) demonstrate the difference between the input salience of the predicted logits and probabilities. As a result, we clarify that, although we define f as the predicted logit (before softmax activations), this choice does not affect the observed increasing trend, no matter when the input salience is generated concerning the logit, probability, or the loss. The results of $\rho_{ind}(k_1, k_2)$, generated from the saliency maps w.r.t. the predicted probability (after softmax activations), are shown in Figure 15. It can be found that $\rho_{ind}(k_1, k_2)$ still increases with both k_1 and k_2.

B Experiment Details

Model Details Throughout the experiments, we use CNNSmall, CNNLarge, ResNetSmall, and ResNetLarge as model families. Within each family, model width is controlled by the parameter k. And the model depths are controlled by the “Small" vs. “Large" suffixes. For CNNs, CNNSmall consists of convolutional layers with channels $[k, 2k, 4k, 8k]$, while CNNLarge repeats each layer twice: $[k, k, 2k, 2k, 4k, 4k, 8k, 8k]$. The details of CNNs are shown in Table 2. Additionally, for TinyImagenet, since the input data is of size 64×64, we increase the stride of the second `MaxPool2d` layer (Layer 10) to 4. As for ResNets, we modify the width of ResNet-10 for ResNetSmall and ResNet-18 for ResNetLarge. Note that $k = 64$ ResNetSmall corresponds to ResNet-10, while $k = 64$ ResNetLarge corresponds to ResNet-18. The sizes of models are illustrated in Figure 16 as the # of trainable parameters.

It should also be noted that, ideally, CNNSmall and CNNLarge, ResNetSmall and ResNetLarge are considered as the same families due to the same architecture. However, since widths can be adjusted independently of the depth, while the adjustment of depth inevitable affects the width, we split them.



(a) CIFAR-10, CS (b) CIFAR-10, CL (c) CIFAR-10, RS (d) CIFAR-10, RL

(e) CIFAR-100, CS (f) CIFAR-100, CL (g) CIFAR-100, RS (h) CIFAR-100, RL

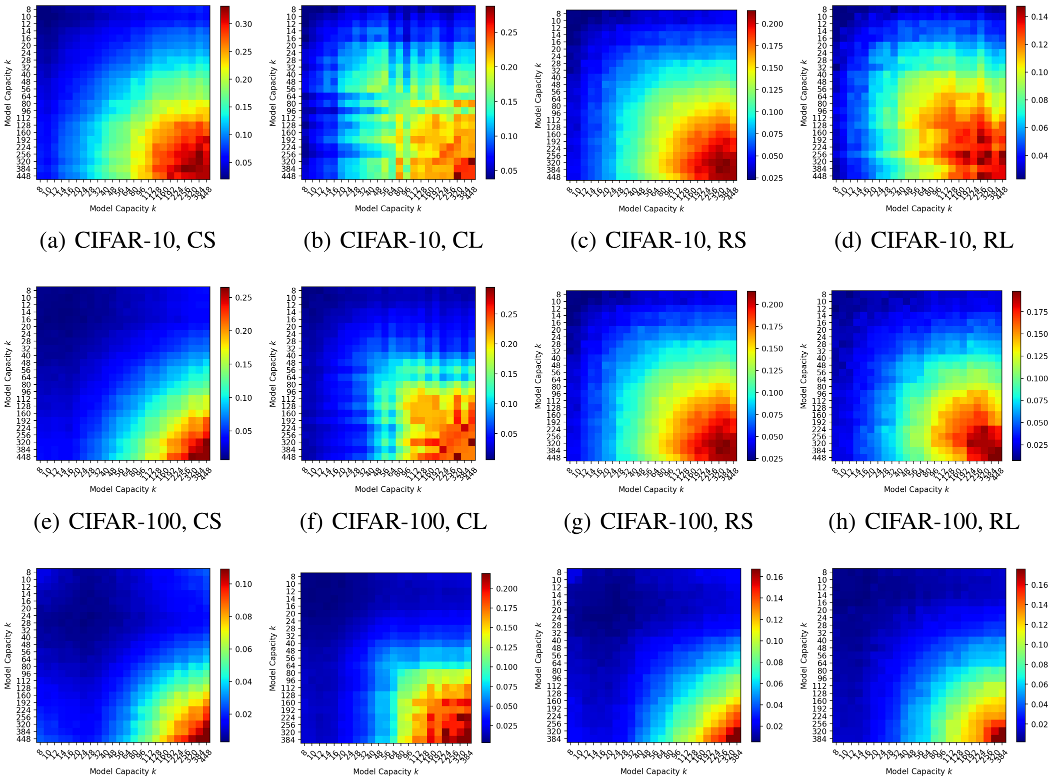
(i) TinyImagenet, CS (j) TinyImagenet, CL (k) TinyImagenet, RS (l) TinyImagenet, RL

Figure 14: The expected similarity $\rho(k_1, k_2)$ between models of varying widths k_1, k_2. Here we include CNNSmall (CS), CNNLarge (CL), ResNetSmall (RS), and ResNetLarge (RL) as $\mathcal{F}$. The values of k_1, k_2 determine the widths in each layer. All the models are initialized to random values without any optimizations. Here the datasets are CIFAR-10 (top), CIFAR-100 (middle), and TinyImagenet (bottom).

Table 1: The average Wasserstein distance $\mathbb{E}_{f^{(1)}, f^{(2)} \in \mathcal{F}(k)} \texttt{WD}(\mathcal{S}(f^{(1)}), \mathcal{S}(f^{(2)}))$ with the standard deivation for all model families {CS, CL, RS, RL}×{10, 20, 40, 80, 160} over CIFAR-10/100 and TinyImagenet datasets. Note that here the baseline should be 2 since the cosine similarity lies in $[-1, 1]$. It is observed that deeper models usually have larger distances (CS vs. CL, RS vs. RL). We deduce that this is because of the training for deeper models is more difficult.

CIFAR-10	$k = 10$	$k = 20$	$k = 40$	$k = 80$	$k = 160$
CS	0.0081 ± 0.0051	0.0100 ± 0.0060	0.0106 ± 0.0062	0.0135 ± 0.0094	0.0153 ± 0.0106
CL	0.0143 ± 0.0091	0.0171 ± 0.0144	0.0322 ± 0.0295	0.0334 ± 0.0287	0.0345 ± 0.0271
RS	0.0097 ± 0.0059	0.0092 ± 0.0055	0.0088 ± 0.0052	0.0073 ± 0.0043	0.0061 ± 0.0036
RL	0.0096 ± 0.0054	0.0107 ± 0.0065	0.0090 ± 0.0053	0.0090 ± 0.0057	0.0163 ± 0.0121
CIFAR-100	$k = 10$	$k = 20$	$k = 40$	$k = 80$	$k = 160$
CS	0.0078 ± 0.0052	0.0056 ± 0.0034	0.0054 ± 0.0031	0.0060 ± 0.0034	0.0085 ± 0.0056
CL	0.0098 ± 0.0059	0.0099 ± 0.0060	0.0156 ± 0.0106	0.0184 ± 0.0130	0.0222 ± 0.0146
RS	0.0071 ± 0.0034	0.0066 ± 0.0029	0.0062 ± 0.0030	0.0061 ± 0.0029	0.0055 ± 0.0029
RL	0.0087 ± 0.0042	0.0079 ± 0.0037	0.0078 ± 0.0037	0.0065 ± 0.0029	0.0078 ± 0.0044
TinyImagenet	$k = 10$	$k = 20$	$k = 40$	$k = 80$	$k = 160$
CS	0.0033 ± 0.0018	0.0021 ± 0.0010	0.0018 ± 0.0007	0.0028 ± 0.0013	0.0060 ± 0.0049
CL	0.0032 ± 0.0016	0.0032 ± 0.0014	0.0055 ± 0.0029	0.0129 ± 0.0083	0.0204 ± 0.0155
RS	0.0083 ± 0.0052	0.0058 ± 0.0038	0.0049 ± 0.0025	0.0045 ± 0.0023	0.0041 ± 0.0018
RL	0.0062 ± 0.0040	0.0060 ± 0.0035	0.0055 ± 0.0030	0.0055 ± 0.0031	0.0057 ± 0.0031

But our experiments in Section 3.1 verify that the depths do not affect the population mean of model



(a) CIFAR-10, CS (b) CIFAR-10, CL (c) CIFAR-10, RS (d) CIFAR-10, RL

(e) CIFAR-100, CS (f) CIFAR-100, CL (g) CIFAR-100, RS (h) CIFAR-100, RL

(i) TinyImagenet, CS (j) TinyImagenet, CL (k) TinyImagenet, RS (l) TinyImagenet, RL

Figure 15: The expected similarity $\rho(k_1, k_2)$ between models of varying widths k_1, k_2. Here we include CNNSmall, CNNLarge, ResNetSmall, and ResNetLarge as $\mathcal{F}$. The values of k_1, k_2 determine the widths in each layer. In particular, all the cosine similarities are between the input saliency maps of the predicted probabilities instead of the predicted logits. Here the datasets are CIFAR-10 (top), CIFAR-100 (middle) and tinyImagenet (bottom).

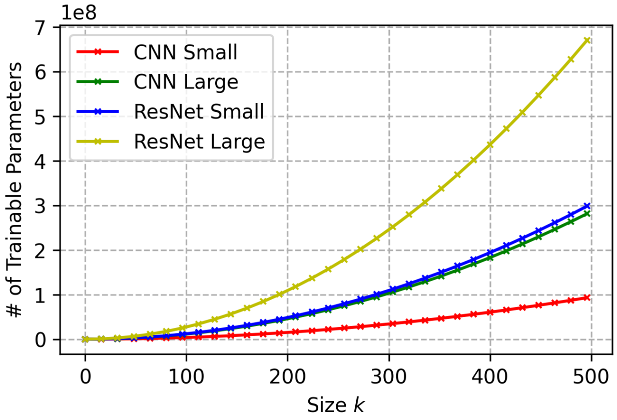
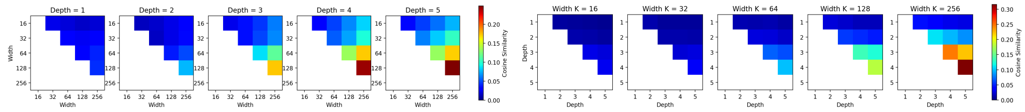


Figure 16: The # of trainable parameters of models vs. the width parameter k for each architecture.

Table 2: CNNSmall Model Details

Layer	Type	Parameters
0	Conv2d	3 inch, k outch, ks 3, stride 1, padding 1
1	BatchNorm2d	-
2	ReLU	-
3	Conv2d	k inch, $2k$ outch, ks 3, stride 1, padding 1
4	BatchNorm2d	-
5	ReLU	-
6	MaxPool2d	ks 2, stride 2
7	Conv2d	$2k$ inch, $4k$ outch, ks 3, stride 1, padding 1
8	BatchNorm2d	-
9	ReLU	-
10	MaxPool2d	ks 2, stride 2
11	Conv2d	$4k$ inch, $8k$ outch, ks 3, stride 1, padding 1
12	BatchNorm2d	-
13	ReLU	-
14	MaxPool2d	ks 8, stride 8
15	Flatten	-
fc	Linear	in_features=80, out_features n_class



(a) Depth-Wise Similarity

(b) Width-Wise Similarity

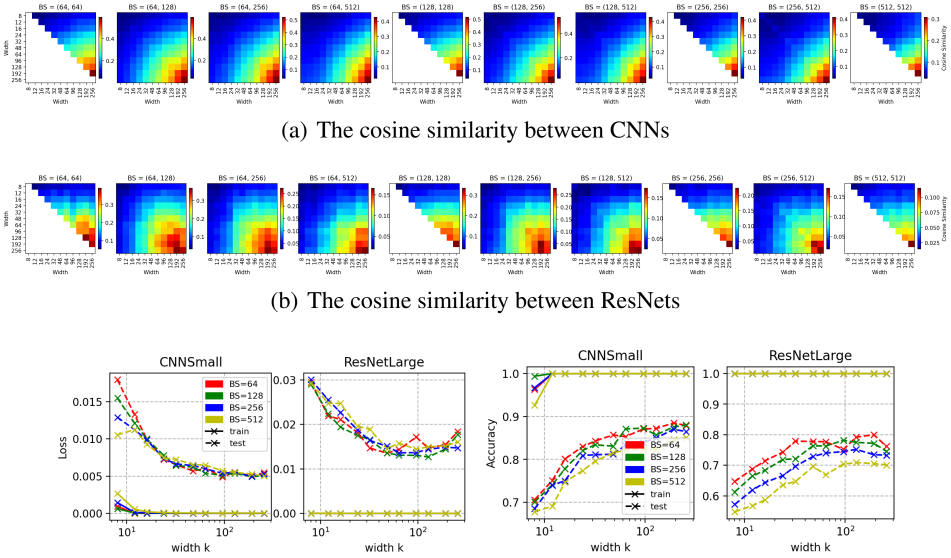
Figure 17: The cosine similarities between CNN models with different widths (left) and depths (right). Widths and depths are enumerated in $\{16, 32, 64, 128, 256\}$ and $\{1, 2, 3, 4, 5\}$. Note that to scale the depths without affecting widths, here we fix the widths of all layers to k instead of $[k, 2k, 4k, 8k]$ as the manuscript

distributions. In experiments, we set $\mathcal{X}$ as the first 1000 samples of the unshuffled testing set of each dataset (CIFAR-10/100, TinyImagenet).

C Additional Settings

This work aims to reveal the convergence trend of the distribution of model behaviors under the stochasticity of the training criterion. This does not limit the conclusion to the specific criterion described above. Distinct training criteria can lead to different distributions of trained models. However, these different distributions of trained models all satisfy the revealed trend. To verify this, we present additional experiments to investigate possible variants such as (1) depths and widths; (2) learning rates; (3) batch sizes; (4) solvers (5) initializations, and (6) other model architectures. Note that as studied in Figure 6 the manuscript, ρ_{ind} can be a computationally efficient compromise of ρ. Therefore, we studied ρ_{ind} in these additional experiments. Besides, there exists enumerous possible combinations of different variants of different aspects. As a result, here we only vary these settings partially since enumerating the entire grid is infeasible. The results as detailed as follows.

Depths and widths. The scale of depths is not as straightforward as the width since modifying depths may change widths as well. Therefore, in the manuscript we study the influence of depth by setting -small and -large variations. Here we present additional results that study the influence of depths continuously, with 1-5 layers, each of which is followed by a max-pooling layer with stride 2. Finally, an adaptive pooling layer is appended at the end. To rule out the influence of widths (channels), all layers have the same width, determined by k. e.g., For the 4-layer scenario, the intermediate layers have widths $[k, k, k, k]$ instead of $[k, 2k, 4k, 8k]$ in the manuscript. The results are shown in Figure 17. It can be found that (1) Given a fixed depth or width, the influence of the



(a) The cosine similarity between CNNs

(b) The cosine similarity between ResNets

(c) The Loss of Models

(d) The Accuracy of Models

Figure 18: (a) and (b) illustrate the cosine similarity between (a) CNNs and (b) ResNets with different batch sizes in $\{64, 128, 256, 512\}$. (c) shows the loss and (d) shows the accuracy of trained models.

other factor is similar when scaled up. (2) Depths are slightly different from widths. Larger widths lead to higher similarities, while closer structures in depths have higher similarities. For widths (left), the similarity always increases left-to-right and top-to-bottom. But for depths (right), pairs near the diagonal have higher similarities.

Batch Sizes. We investigate the influence of batch sizes, varying in $\{64, 128, 256, 512\}$ The results are shown in Figure 18. It can be observed that although different batch sizes lead to different performance (e.g. testing accuracy), the convergence trend holds in all scenarios.

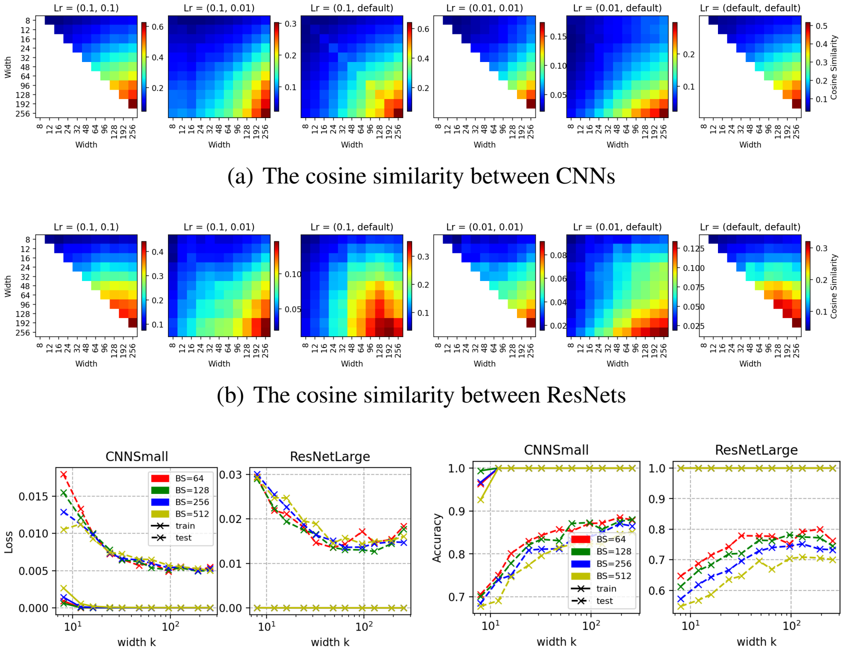
Learning Rates. We test different learning rates on how they affect the results. We include `1e-1`, `1e-2`, `default`, where "`default`" refers to the criterion used in the manuscript. As shown in Figure 19, the revealed trend is preserved in all learning rates. It is also worth noticing that learning rates affect ResNets more than CNNs.

Solvers. Apart from SGD, we include Adam, AdamW, and SGD w/ momentum. For Adam and AdamW we set the learning rate to 1e-3, while SGD w/ momentum uses a learning rate of 1e-1 with a momentum of 0.9. The results are shown in Figure 20. Although different solvers lead to models of different performances, they all preserve the same convergence trend.

Table 3: The comparison between the similarities between single models of different criteria.

	(a) Diff. Init.; Same Order	(b) Diff. Init.; Diff. Order	(c) Same Init. θ_0; Diff Order	(d) Same Init. θ_1; Diff Order
# of pairs	100	4095	4095	4095
mean of ρ_{ind}	0.0758	0.0753	0.0879	0.0855
std. of ρ_{int}	0.0038	0.0037	0.0042	0.0048

Initializations. Given a training scheme and model family $\mathcal{F}(k)$, the training procedure leads to a distribution of trained models $p(f)$. When the initialization is fixed to θ, the training procedure is essentially sampling from the conditional distribution $p(f|\theta)$ instead of the unconditional distribution $p(f)$. We then studied the difference between the unconditional distribution $p(f)$ and the conditional distribution $p(f|\theta_0)$. We focus on two conditional distributions $p(f|\theta_0)$ and $p(f|\theta_1)$, where θ_1



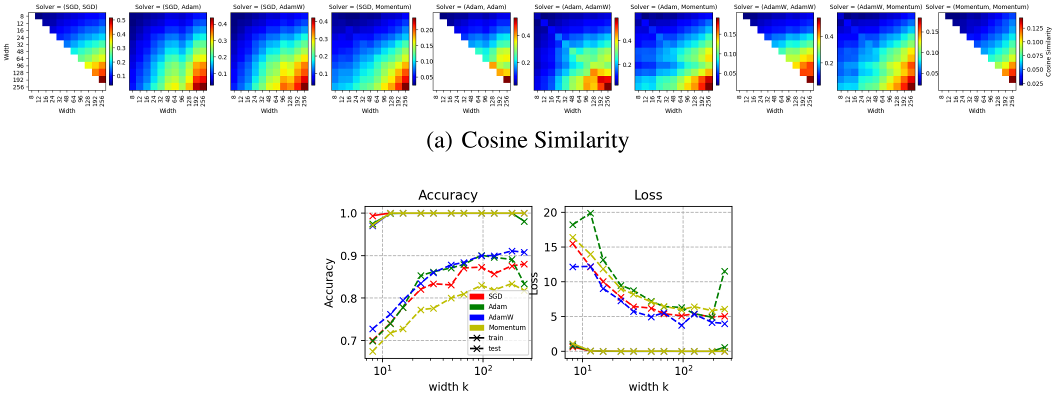
(a) The cosine similarity between CNNs

(b) The cosine similarity between ResNets

(c) The Loss of Models

(d) The Accuracy of Models

Figure 19: (a) and (b) illustrate the cosine similarity between (a) CNNs and (b) ResNets with different batch sizes in $\{64, 128, 256, 512\}$. (c) shows the loss and (d) shows the accuracy of trained models.



(a) Cosine Similarity

(b) Performance

Figure 20: (a) The cosine similarity between CNN models with different solvers in {`SGD, Adam, AdamW, SGD w/ Momentum`}. (b): The accuracy and loss of trained models.

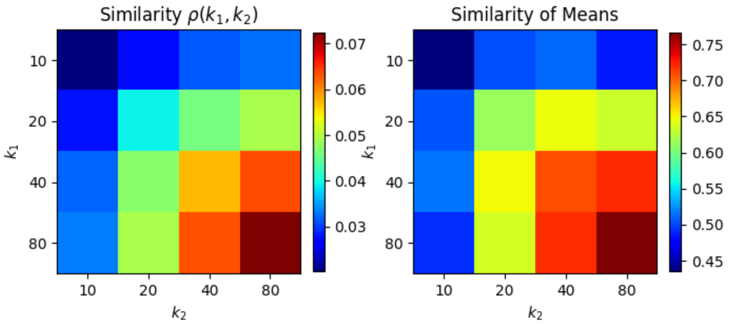


Figure 21: The cosine similarity between Vision Transformers (ViTs) on CIFAR-10. The capacity is controlled by $k \in \{10, 20, 40, 80\}$, where the embedding dimension is $4k$, separated to $k/2$ heads. The left subfigure shows the mean of the similarity. The right subfigure shows the similarity of the population mean.

represents the initializations under `seed=1` and θ_0 represents those under `seed=0`. Other settings are identical. We thus have $f_1^0, \cdots, f_{100}^0 \sim p(f|\theta_0)$ and $f_1^1, \cdots, f_{100}^1 \sim p(f|\theta_1)$. The superscript indicates the initialization seeds and the subscript indicates the training seeds.

First, we notice immediately that the training seeds for both θ_0 and θ_1 are 1 100. This means that $\forall i, f_i^0, f_i^1$ differ only in initializations. We inspect (a) $\rho_{ind}(f_i^0, f_i^1)$ (100 pairs) to see if they have exceptional similarity compared with (b) $\rho_{ind}(f_i^0, f_j^1), i \neq j$ ($\binom{100}{2} = 4950$ pairs). Besides, within the same condition, all models only differ in terms of the orders of the training batch. We thus also inspect the similarity of all models of the same condition: (c) $\rho_{ind}(f_i^0, f_j^0), i \neq j$ and (d) $\rho_{ind}(f_i^1, f_j^1), i \neq j$. Each of them has $\binom{100}{2} = 4950$ pairs.

As demonstrated in Table 3, (i) the comparison between (a) and (b) indicates that with different initializations, the same order of batches in the training procedure does not contribute to higher similarities. (ii) The comparison among (b)(c)(d) indicates that the same initialization indeed leads to higher similarity even though the order of batches is distinct. It should be noted that the contributions of batch orders and initializations are also affected by the number of epochs. Intuitively, more training epochs should lead to smaller contributions from the initializations but greater contributions from the batch orders.

Vision Transformers. Vision Transformers (ViT) have risen in recent years as another powerful architecture for CV tasks. Here we include a brief study of ViTs to demonstrate that the discovered phenomenon holds across different architectures.

Specifically, we train vision transformer (ViT) models on CIFAR-10 with varying capacities controlled by $k \in \{10, 20, 40, 80\}$. CIFAR-10 has an input size of 32×32 pixels, thus the patch size is set as 4×4, resulting in 8×8 patches. The embedding size is set to $4k$, divided by $k/2$ heads, and we set the depth to 8. The seeds vary in 1-100 and results in 100 trained models of each k. We study the mean of the similarity $\rho(k_1, k_2)$ (i.e. the same experiments as Figure 3 in the manuscript) and the similarity of the population mean (i.e. the same as Figure 5 in the manuscript). The results are shown in Figure 21. It can be observed that although distinct from convolutional layers, the transformer structure also has the discovered convergence trend. It can also be noted that the degree of dispersion of ViTs is much higher than CNN-based models.

In conclusion, although training schemes can affect the resulting distributions of models, the influence of the model capacity stays invariant across different criteria.

D Uniform Distributions on the Hypersphere

According to (Muller, 1959), due to the spherical symmetry property of the zero-mean Gaussian distribution, the cosine similarity between two Gaussian variables are actually uniformly distributed over the hypersphere $\mathcal{S}^{d-1}$ in $\mathbb{R}^d$.[3] Therefore, the cosine similarity between two i.i.d. multivariate Gaussian tensors is essentially the inner product between two i.i.d. uniform tensors over the hypersphere. Formally, suppose X, Y to be high-dimensional i.i.d. random variables of dimension d that follow the Gaussian distribution $\mathcal{N}(\mathbf{0}, I_{d\times d})$. The cosine similarity is $Z = \frac{X^TY}{\|X\|\cdot\|Y\|}$. WLOG, it suffices to consider the scenario where $\frac{Y}{\|Y\|} = \boldsymbol{e}_1 = (1, 0, \cdots, 0)$. And it is written as $Z = \boldsymbol{e}_i^T\frac{X}{\|X\|} = \frac{X_1}{\|X\|} = \frac{X_1}{\sqrt{\sum_{i=1}^d X_i^2}} = \sqrt{\frac{X_1^2}{X_1^2\sum_{i=1}^d X_i^2}}$. Note that $X_1^2 \sim \chi^2(1), \sum_{i=2}^d X_i^2 \sim \chi^2(d-1)$. As a result, $Z^2 = \frac{X_1^2}{X_1^2+\sum_{i=2}^d X_i^2}$ follows a beta distribution $Z^2 \sim Beta(\frac{1}{2}, \frac{d-1}{2})$. The pdf is thus $f_{Z^2}(x) = \frac{x^{-1/2}(1-x)^{(d-3)/2}}{B(1/2,(d-1)/2)}$. And then when $Z > 0$,

$$f_Z(x) = f_{Z^2}(x^2)|2x| \tag{7}$$

$$= \frac{(x^2)^{-1/2}(1-x^2)^{(d-3)/2}}{B(1/2,(d-1)/2)}|2x| \tag{8}$$

$$= \frac{2}{B(\frac{1}{2}, \frac{d-1}{2})}(1-x^2)^{(d-3)/2} \tag{9}$$

According to (Smith et al., 2023), let $u = (1+x)/2$,then $x^2 = (2u-1)^2$. Then this can be simplified to

$$f_U(u) = \frac{2}{B(\frac{1}{2}, \frac{d-1}{2})}(1-(2u-1)^2)^{(d-3)/2}\cdot\frac{\mathrm{d}x}{\mathrm{d}u} \tag{10}$$

$$= \frac{1}{B(\frac{1}{2}, \frac{d-1}{2})}2^{d-2}u^{(d-1)/2-1}(1-u)^{(d-1)/2-1} \tag{11}$$

$$= \frac{1}{B(\frac{1}{2}, \frac{d-1}{2})2^{2-d}}u^{(d-1)/2-1}(1-u)^{(d-1)/2-1} \tag{12}$$

$$= Beta(\frac{d-1}{2}, \frac{d-1}{2}) \tag{13}$$

This is because

$$B(\frac{d-1}{2}, \frac{d-1}{2}) = B(\frac{d-1}{2}, \frac{d+1}{2})\cdot 2 \tag{14}$$

$$= \frac{\Gamma(\frac{d-1}{2})\Gamma(\frac{d+1}{2})}{\Gamma(d)}\cdot 2 \tag{15}$$

$$= \frac{\Gamma(\frac{d-1}{2})\Gamma(\frac{d+1}{2})}{2^{d-1}\pi^{-1/2}\Gamma(\frac{d-1}{2})\Gamma(\frac{d+1}{2})}\cdot 2 \tag{16}$$

$$= 2^{2-d}\frac{\Gamma(\frac{d-1}{2})\Gamma(\frac{1}{2})}{\Gamma(\frac{d}{2})} = B(\frac{1}{2}, \frac{d-1}{2})2^{2-d} \tag{17}$$

Finally, we are able to conclude that $\frac{1+Z}{2} \sim Beta(\frac{d-1}{2}, \frac{d-1}{2})$, where Z is the cosine similarity between two i.i.d. d-dimensional Gaussian vectors. This result suggests that the distribution of Z will become very concentrated around 0. And this concentration exacerbates exponentially with the dimension d. Given a cosine similarity level ρ, the probability $\mathbb{P}(Z > \rho)$ can be extremely small, and also deceases exponentially with ρ, too. We visualize the relation between the probability $\mathbb{P}(Z > \rho)$ and ρ with varying dimensions d in Figure 22. In the low-dimensional space such as $d = 3$, the distribution of the cosine similarity is close to uniform as humans' intuition. However, as the dimension increases, the cosine similarity is very unlikely to maintain high, as demonstrated by the curves of $d = 3\times 8\times 8 = 48$ (orange) and $d = 3\times 32\times 32 = 3072$ (green).

[3] This is because of the rotation-invariance. Let $O \in \mathbb{R}^{d\times d}$ be any orthonormal matrix, then after the rotation we obtain $\frac{OX}{\|X\|} = \frac{OX}{\|OX\|}$. Since $OX \sim \mathcal{N}(\mathbf{0}, I)$, too, we know that $\frac{OX}{\|X\|}$ and $\frac{X}{\|X\|}$ are from the identical distribution.

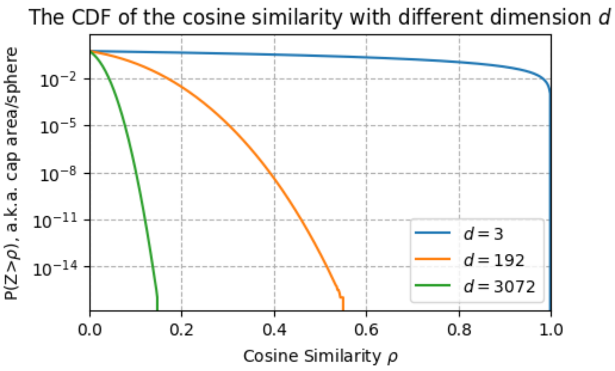


Figure 22: The relation between the probability $\mathbb{P}(Z > \rho)$ and the cosine similarity value ρ.

E Supplementary Experiment Results

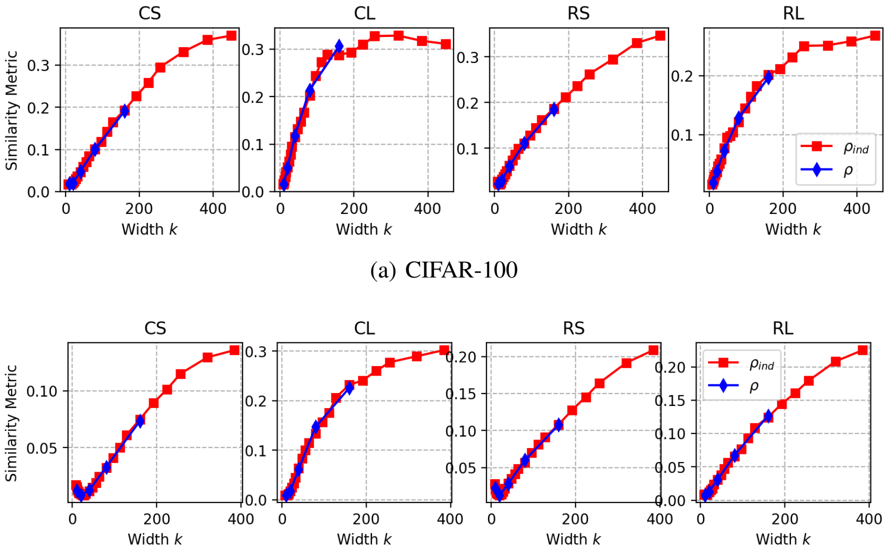
Note that all the experiments are carried across three datasets CIFAR-10/100 and TinyImagenet, along with four model families CNNSmall, CNNLarge, ResNetSmall, and ResNetLarge. Due to the space limit, only part of the selected results can be put in the manuscript. Therefore, we defer the results with different models/datasets into this section for the audience' reference. All the conclusions drawn from the experiment results shown in the manuscript hold for the results demonstrated here.

In Figure 23, we present the complementary results of Figure 6 on CIFAR-100 and TinyImagenet. It can be clearly observed that ρ_{ind} and ρ have very similar values, after taking the expectation over $\mathcal{X}$.

Figure 25 shows the results of black-box attack results between different models. All other settings are identical to the results shown in Figure 10, but with CIFAR-100 and TinyImagenet instead of CIFAR-10. It is observed that the capacity of models has significant influence to the models' robustness and black-box attack transferability. And this trend is highly correlated to the similarity ρ_{ind}, ρ as demonstrated in Figures 2, 3, 12, 13 and 15.

Similarly, Figure 26 shows supplementary results of Figure 11. All settings are identical except for the datasets. CIFAR-100 and TinyImagenet are tested instead of CIFAR-10. It can be observed that using the estimated mean gradient direction (blue), the performance drops much more significantly than the attack from single models of either the exact same family as the target model (green) or the largest single model tested ($k = 448$, red). Note that due to the complexity of TinyImagenet, in the overfitting phase (i.e. when the target models' capacities are comparable to the dataset (Nakkiran et al., 2021)), the single-model black-box attack results in opposite effect – the prediction actually increases.

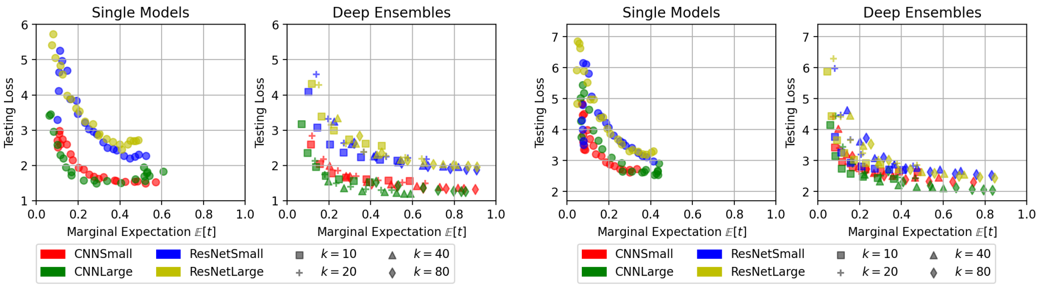
In Figure 24, we present CIFAR-100 and TinyImagenet results as supplementary of Figure 9. It can be observed that the testing loss is highly correlated to the expectation of $t = \boldsymbol{u}^T \boldsymbol{\mu}(\boldsymbol{x})$. Such phenomena are also consistent across different model families and datasets. For both single models and ensembles, the closer they are to the convergent limiting point (i.e. larger $\mathbb{E}[t]$), the higher testing performance they have. Note that here for the sake of consistency, we approximate $\boldsymbol{\mu}(\boldsymbol{x})$ through $\tilde{\boldsymbol{\mu}}(160; \boldsymbol{x})$. Therefore, in the ensemble experiments (left of each subfigure), $k = 160$ is omitted.



(a) CIFAR-100

(b) TinyImagenet

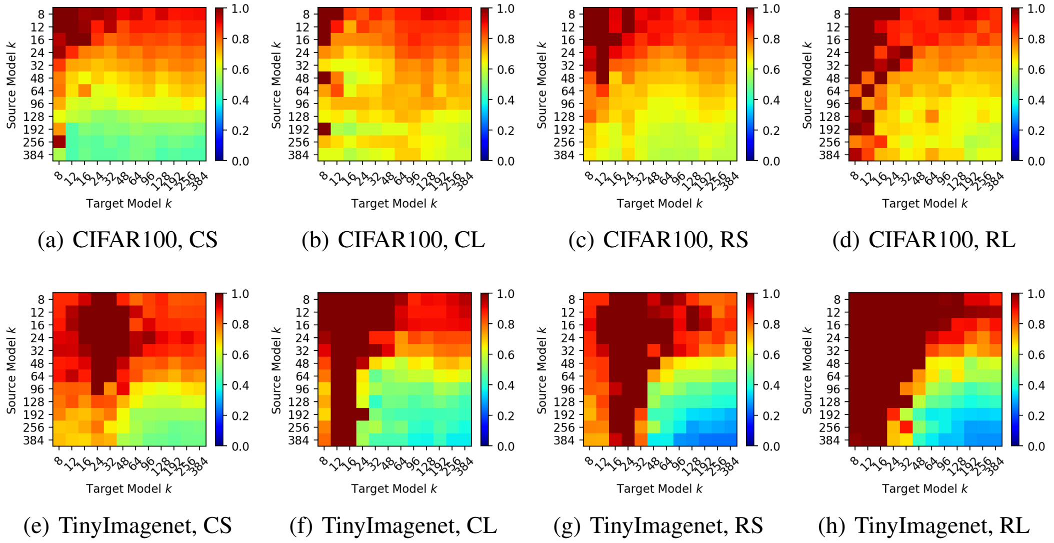
Figure 23: Illustration of (red) $\rho_{ind}(f^{(1)}, f^{(2)}), f^{(1)}, f^{(2)} \in \mathcal{F}(k)$ and (blue) $\rho(k, k)$ on (a) CIFAR-100 and (b) TinyImagenet.



(a) CIFAR-100

(b) TinyImagenet

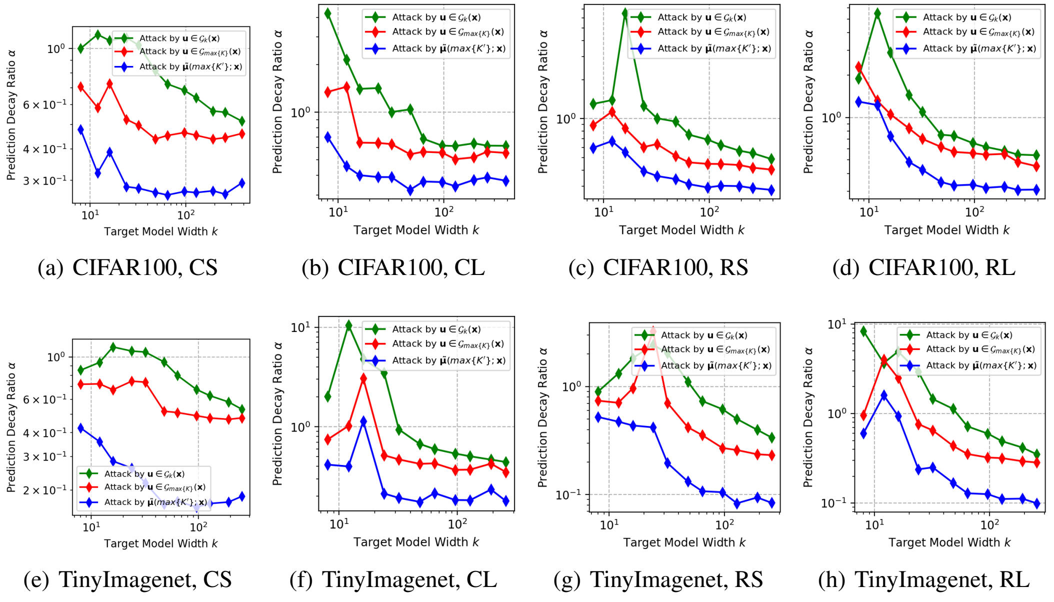
Figure 24: The illustration of the relation between the expected testing loss $\mathbb{E}_{\mathcal{X}}[\mathcal{L}]$ and the marginal expectation $\mathbb{E}_{\mathcal{X}}[t]$. Both (a) CIFAR-100 and (b) TinyImagenet results are shown as supplementary to Figure 9. Models are from (i) single models with varying structure; and (ii) deep ensembles with varying members. Each color represents a model family.



(a) CIFAR100, CS (b) CIFAR100, CL (c) CIFAR100, RS (d) CIFAR100, RL

(e) TinyImagenet, CS (f) TinyImagenet, CL (g) TinyImagenet, RS (h) TinyImagenet, RL

Figure 25: The results of single model black-box attack. The value of each entry is $\alpha(f^{(1)}, f^{(2)})$, $f^{(1)} \in \mathcal{F}(k_1), f^{(2)} \in \mathcal{F}(k_2)$ for different model capacities. Here k_1 is the width parameter of the source model and k_2 is the width parameter of the target model.



(a) CIFAR100, CS (b) CIFAR100, CL (c) CIFAR100, RS (d) CIFAR100, RL

(e) TinyImagenet, CS (f) TinyImagenet, CL (g) TinyImagenet, RS (h) TinyImagenet, RL

Figure 26: The comparison between the single-model attack from the largest model (red), the single-model attack from the very same structure (green), and the attack by the mean direction (blue). The top row shows the results of CIFAR-100, and the bottom row shows the results of TinyImagenet. Some figures' y-axis are set to logarithm for clarity.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract mainly points out the hypotheses made in this work. They are all properly discussed and empirically verified in the manuscript.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Please refer to the conclusion section for the limitations.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: The theoretical analysis in this paper is rigorous. Hypothesis are both theoretically explained and empirically verified.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in the appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The information needed to reproduce the experiments is discussed in both the manuscript and the appendix. The code is also provided in the supplementary materials. The repository will be publicized upon acceptance.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The code is provided as the supplementary material in the submission. The repository will be publicized upon acceptance.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Details of the experiments are discussed in both the manuscript and the appendix. The code is also provided in the supplementary materials.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The potential statistical significance lies in the variance among different ensemble members. This is studied in Table 1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Please refer to the first paragraph of Section 3

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The authors have reviewed the code of ethics and this work conforms with the code of ethics of NeurIPS.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: Focusing on fundamental, theoretical and general topics, this works does not have societal impact.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: There is no such risk of this work.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: This work only uses general packages such as `Numpy`, `PyTorch`, `torchvision`, etc.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: This work does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This work does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.