Uncovering the Connections Between Adversarial Transferability and Knowledge Transferability

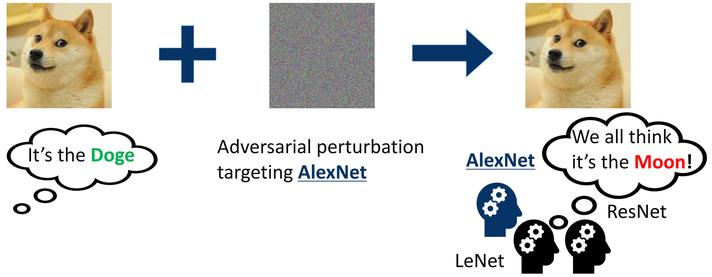
Kaizhao Liang* Jacky Y. Zhang* Boxin Wang Zhuolin Yang Oluwasanmi Koyejo Bo Li

University of Illinois at Urbana-Champaign

Background

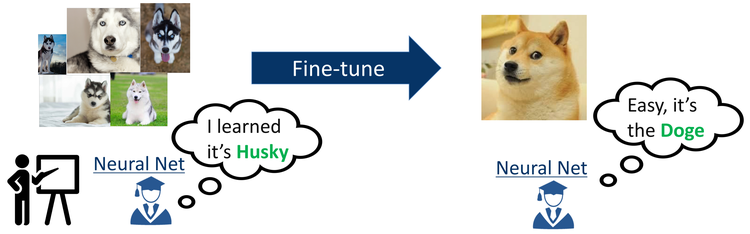
What is Adversarial Transferability?

Adversarial perturbation generated on one model can be used to attack the other.



What is Knowledge Transferability?

Model trained on one task can be used as a prior to train on another task



Motivations

Questions

Q1. Are there fundamental connections between the two transferabilities?

Q2. If Q1 is true, then can we measure one and indicate another?

Q3. If Q2 is true, then are there any potential applications?

To all the three questions: Yes

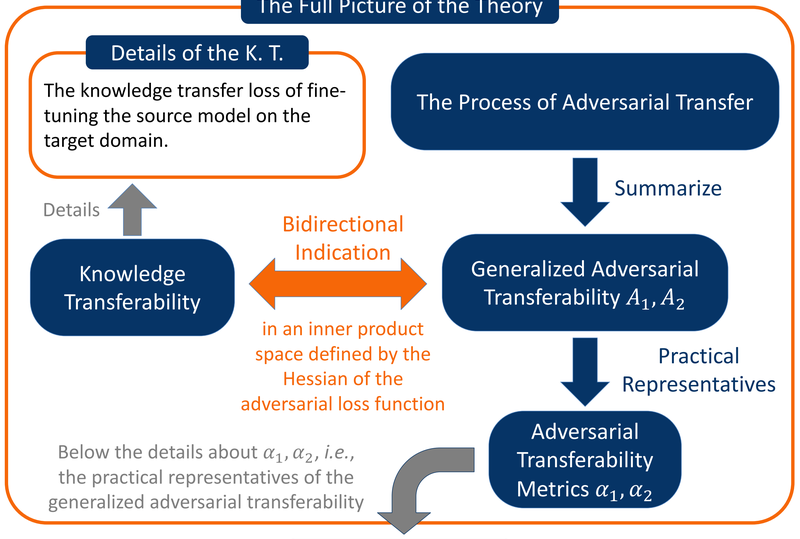
*Equal Contribution

Theoretical Analysis Sketch

Theorems Sketch

Knowledge transferability indicates the generalized adversarial transferability, and vice versa.

The Full Picture of the Theory



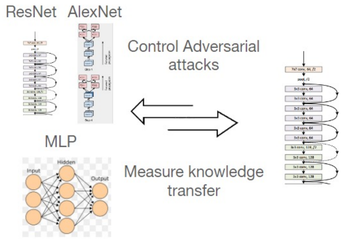
Details of α_1, α_2

- Notations: input data x; source model f_s: $\mathbb{R}^n \to \mathbb{R}^m$; target model f_T: $\mathbb{R}^n \to \mathbb{R}^d$; $\delta_{f_s}(x)$ denotes the adversarial perturbation of data x on model f_s; similar definition for $\delta_{f_T}(x)$.
- Setting: for example, we consider transfer $\delta_{f_s}(x)$ to f_T.
- $\alpha_1$characterizes the 'magnitude' of the difference between $f_T(x + \delta_{f_s}(x))$ and $f_T(x)$. A_1 generalizes α_1.
- $\alpha_2$characterizes the 'direction' of the difference between $f_T(x + \delta_{f_s}(x))$ and $f_T(x)$. A_2 generalizes α_2.
- $\alpha_1, \alpha_2 \in [0,1]$: the closer to 1, the better the adversarial transferability.

Experiments

From Adversarial transferability to knowledge transferability

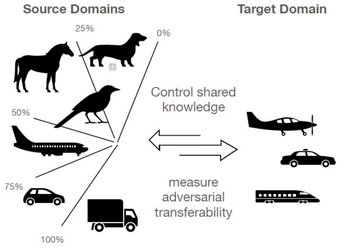
We pick 5 different architectures that have different levels of adversarial trasnferability to the target model. Then we measure each of their accuracy on the target domain after transfer learning.



Model	Knowledge Trans.	α_1	α_2	$\alpha_1 * \alpha_2$
MLP	28.30	0.35	0.19	0.026
LeNet	45.65	0.32	0.22	0.025
AlexNet	55.09	0.34	0.21	0.027
ResNet18	76.60	0.54	0.24	0.071
REsNet50	77.92	0.61	0.22	0.090

From Knowledge transferability to Adversarial transferability

We manually construct 5 different source data distributions. The percentage represents how close they are to the target domain. Then we measure how adversarial transferable their adversarial samples to the target model.



Similarity	Knowledge Trans.	α_1	α_2	$\alpha_1 * \alpha_2$
0%	28.30	0.31	0.15	0.017
25%	45.65	0.32	0.31	0.038
50%	55.09	0.34	0.36	0.044
75%	76.60	0.34	0.31	0.040
100%	77.92	0.36	0.36	0.049

Empirical Observation

Knowledge transferability and the proposed adversarial transferability metrics track each other.