Robust Learning under Distribution Shifts

Machine Learning Summer School, Okinawa

Mar. 8th, 2024

Han Zhao
hanzhao@illinois.edu
Assistant Professor
Department of Computer Science
University of Illinois Urbana-Champaign

Robustness

Domain-Invariant Representations
Fundamental limit in domain adaptation

Invariant Risk Minimization
An efficient algorithm for IRM via post-processing

Gradual Domain Adaptation
Learning under continuous distribution shifts

Robust Multitask Learning
Understanding multi-objective optimization

Learning Fair Representations
Tradeoff between fairness & accuracy

Fairness

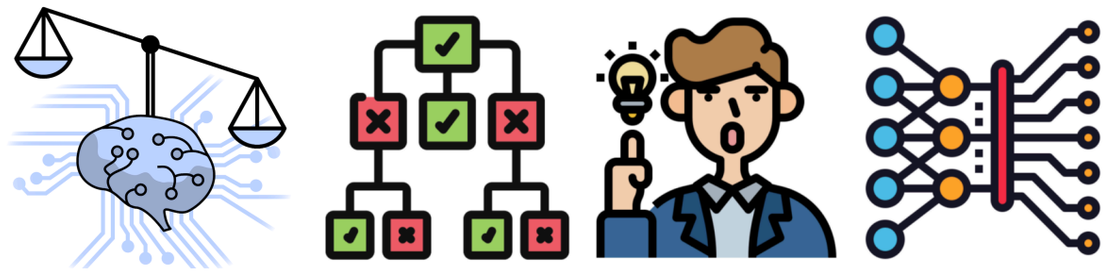
Invariant Representation Learning
Information-theoretic analysis between invariance & accuracy, with applications in distributional robustness and fairness

Tradeoff between robustness and fairness
Understanding the impact of adversarial robustness on fairness

Fair and Optimal Classification (I)
An optimal post-processing algorithm for demographic parity

Fair and Optimal Classification (II)
An optimal post-processing algorithm for equalized odds

Trustworthy Machine Learning



Differentially-Private and Fair Regression
Design fair & private regression algorithm

Privacy-Preserving Learning
Learning under attribute-inference attack

Machine Unlearning
Removing a subset of specified data from the learned model

Efficiency & Accuracy

Maximum Influence Subset
Understand and select out the subset of data with maximum influence to the learned model

Structured Representations
Learning representations with class hierarchical information

Privacy-Preserving Learning for Graph Neural Networks
Provide defenses strategies under attribute-inference attacks over graphs

Interpretability

Privacy

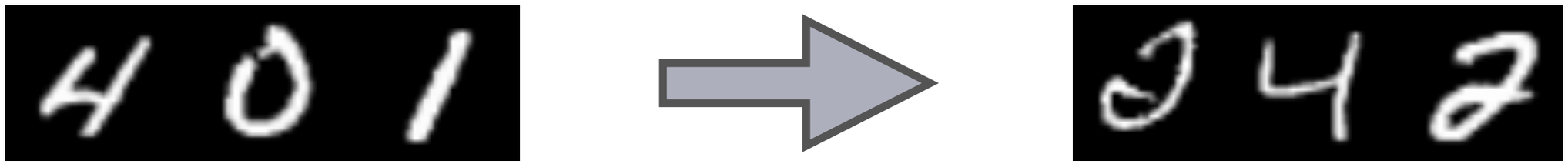
Robustness:

- Domain adaptation / generalization / Out-of-distribution generalization / transfer learning
- Invariant representation learning / invariant causal predictors

Key Factors underlying the Success

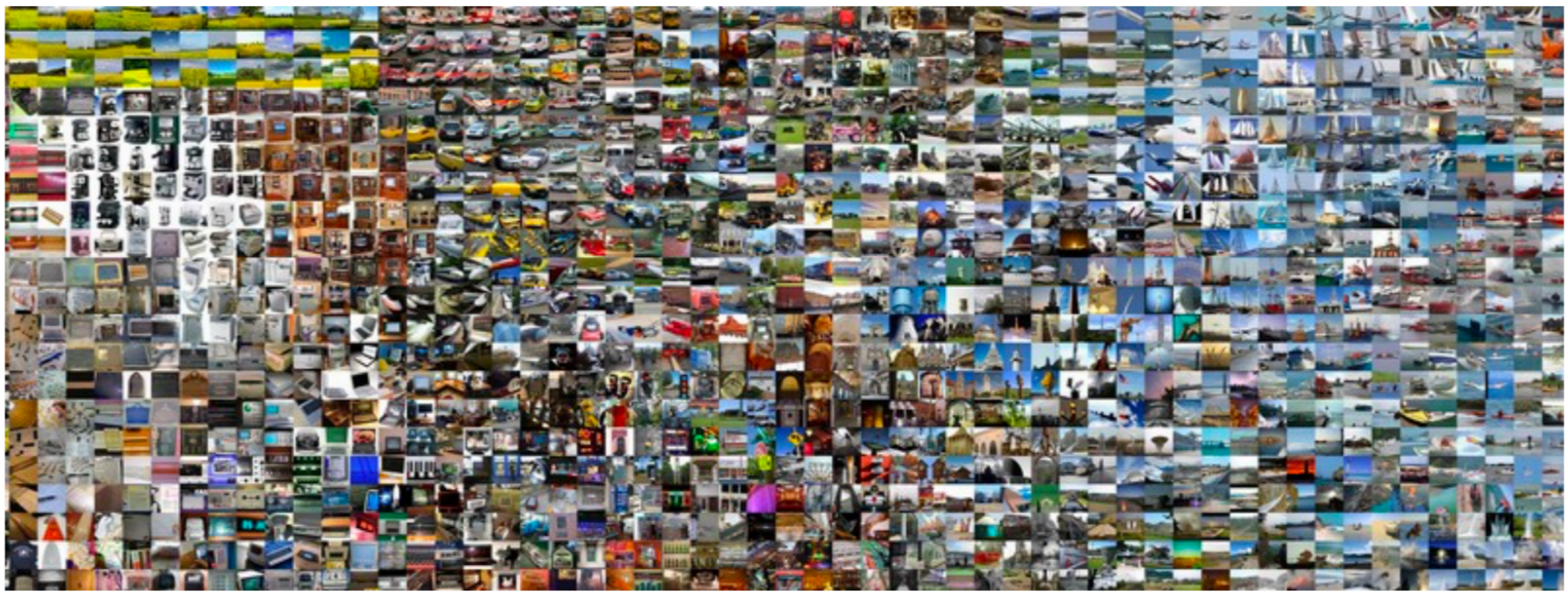
- Stationary distributions
- Large-scale labeled datasets & Powerful computation

Source (Train) = Target (Test)



Source

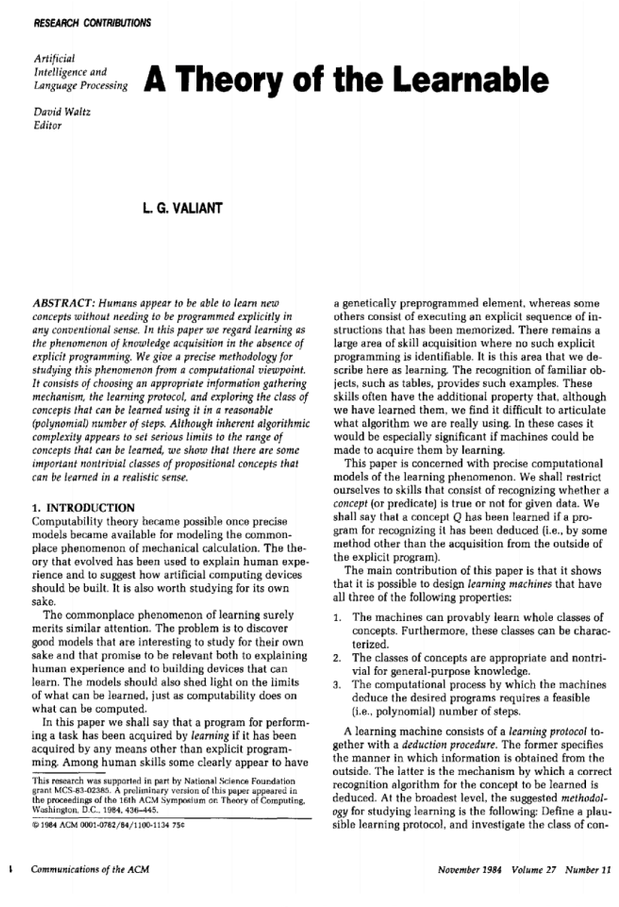
Target



Probably Approximately Correct (PAC)

Probably Approximately Correct (PAC, Valiant, CACM 1984)

- (Informal) A framework to quantify the meaning of learning a concept from samples
- With high probability (P), the learned predictor will have low generalization error (AC)
- No distributional assumption



RESEARCH CONTRIBUTIONS

Artificial Intelligence and Language Processing

David Waltz Editor

A Theory of the Learnable

L. G. VALIANT

ABSTRACT: Humans appear to be able to learn new concepts without needing to be programmed explicitly in any conventional sense. In this paper we regard learning as the phenomenon of knowledge acquisition in the absence of explicit programming. We give a precise methodology for studying this phenomenon from a computational viewpoint. It consists of choosing an appropriate information gathering mechanism, the learning protocol, and exploring the class of concepts that can be learned using it in a reasonable (polynomial) number of steps. Although inherent algorithmic complexity appears to set serious limits to the range of concepts that can be learned, we show that there are some important nontrivial classes of propositional concepts that can be learned in a realistic sense.

1. INTRODUCTION

Computability theory became possible once precise models became available for modeling the commonplace phenomenon of mechanical calculation. The theory that evolved has been used to explain human experience and to suggest how artificial computing devices should be built. It is also worth studying for its own sake.

The commonplace phenomenon of learning surely merits similar attention. The problem is to discover good models that are interesting to study for their own sake and that promise to be relevant both to explaining human experience and to building devices that can learn. The models should also shed light on the limits of what can be learned, just as computability does on what can be computed.

In this paper we shall say that a program for performing a task has been acquired by *learning* if it has been acquired by any means other than explicit programming. Among human skills some clearly appear to have a genetically preprogrammed element, whereas some others consist of executing an explicit sequence of instructions that has been memorized. There remains a large area of skill acquisition where no such explicit programming is identifiable. It is this area that we describe here as learning. The recognition of familiar objects, such as tables, provides such examples. These skills often have the additional property that, although we have learned them, we find it difficult to articulate what algorithm we are really using. In these cases it would be especially significant if machines could be made to acquire them by learning.

This paper is concerned with precise computational models of the learning phenomenon. We shall restrict ourselves to skills that consist of recognizing whether a *concept* (or predicate) is true or not for given data. We shall say that a concept Q has been learned if a program for recognizing it has been deduced (i.e., by some method other than the acquisition from the outside of the explicit program).

The main contribution of this paper is that it shows that it is possible to design *learning machines* that have all three of the following properties:

1. The machines can provably learn whole classes of concepts. Furthermore, these classes can be characterized.
2. The classes of concepts are appropriate and nontrivial for general-purpose knowledge.
3. The computational process by which the machines deduce the desired programs requires a feasible (i.e., polynomial) number of steps.

A learning machine consists of a *learning protocol* together with a *deduction procedure*. The former specifies the manner in which information is obtained from the outside. The latter is the mechanism by which a correct recognition algorithm for the concept to be learned is deduced. At the broadest level, the suggested *methodology* for studying learning is the following: Define a plausible learning protocol, and investigate the class of con-

This research was supported in part by National Science Foundation grant MCS-83-02385. A preliminary version of this paper appeared in the proceedings of the 16th ACM Symposium on Theory of Computing, Washington, D.C., 1984, 426–445.

A Theory of the Learnable, Communications of the ACM, 1984

Probably Approximately Correct (PAC)

With VC dim as the complexity measure, we have the following uniform generalization bound:

Given $\mathcal{D} = \{(x^{(i)}, y^{(i)})\}_{i=1}^{n} \sim \mu$ be a dataset of iid samples. Let $\mathcal{F}$ be a hypothesis class of finite VC-dim, i.e., $\mathrm{VCdim}(\mathcal{F}) < \infty$, then for $0 < \delta < 1$, with probability at least $1 - \delta$, for all $f \in \mathcal{F}$:

$$\varepsilon_\mu(f) \leq \hat{\varepsilon}_{\mathcal{D}}(f) + O\left(\sqrt{\frac{\mathrm{VCdim}(\mathcal{F}) + \log(1/\delta)}{n}}\right)$$

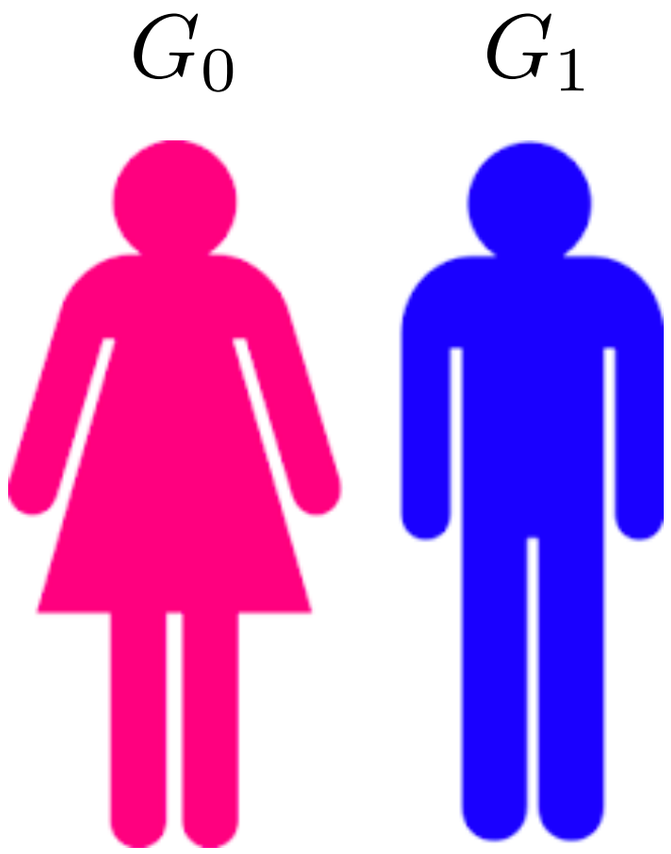
Note:

- The bound above gives the generalization error, and we can use the generalization error bound to provide an upper bound on the excess risk as well, i.e., $\varepsilon_\mu(f) - \inf_{f' \in \mathcal{F}} \varepsilon_\mu(f')$
- There are other forms of complexity measures to characterize the expressiveness/richness/ powerfulness of a given hypothesis class, e.g., Rademacher complexity, covering number, algorithmic stability, etc, but the high-level form is the same
- The bound above could be loose, i.e., the generalization error could be larger than 1 for classification problems
- Other kinds of generalization bounds exist as well, i.e., instead of providing high-probability bounds, providing guarantees on expected generalization error (PAC-Bayes, mutual information, etc)

Distribution Shift in Practice

Commercial face recognition systems

Some of the key findings from the paper:



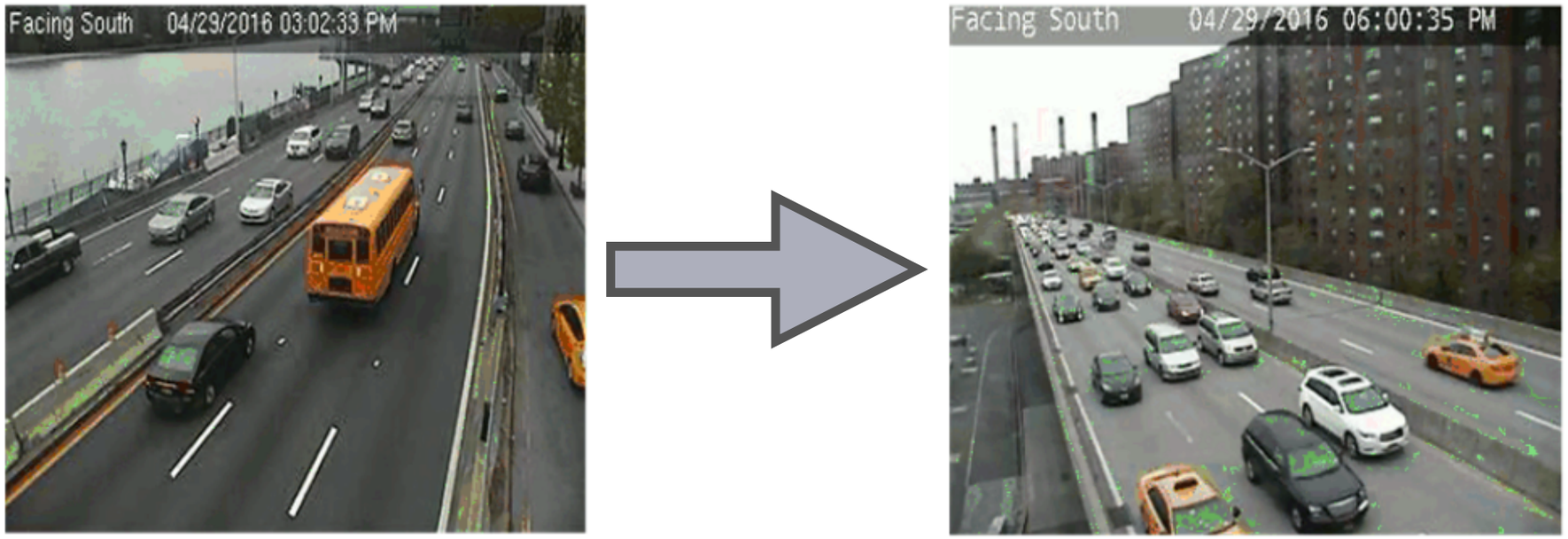
- All classifiers perform better on male faces than female faces (8.1% – 20.6% difference in error rate)
- All classifiers perform better on lighter faces than darker faces (11.8% – 19.2% difference in error rate)
- All classifiers perform worst on darker female faces (20.8% – 34.7% error rate)
- Microsoft and IBM classifiers perform best on lighter male faces (error rates of 0.0% and 0.3% respectively)
- Face++ classifiers perform best on darker male faces (0.7% error rate)
- The maximum difference in error rate between the best and worst classified groups is 34.4%

Accuracy disparity:

$$\Delta_{\mathrm{Err}}(h) := |\Pr(h(X) \neq Y \mid X \sim G_0) - \Pr(h(X) \neq Y \mid X \sim G_1)|$$

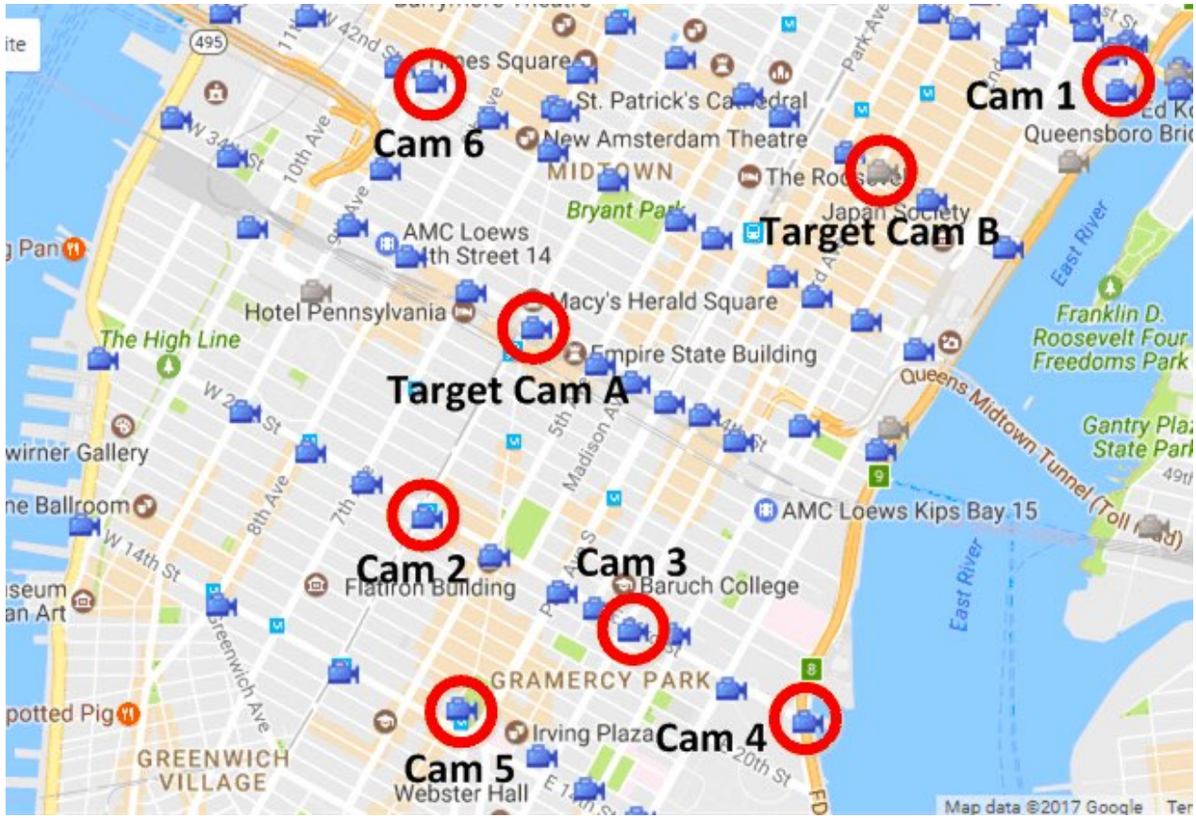
Domain Adaptation / Generalization

Machine learning models could be brittle



Source (with Labels)

Target (No Labels)



Domain Adaptation / Generalization

Domain adaptation: given unlabeled data from the target domain + labeled data from the source domain, can we do better?

Note: closely related to the setting of semi-supervised learning, but with a key difference:

Semi-supervised learning:

Training distribution $=$ Test distribution

Domain adaptation:

Training distribution $\neq$ Test distribution

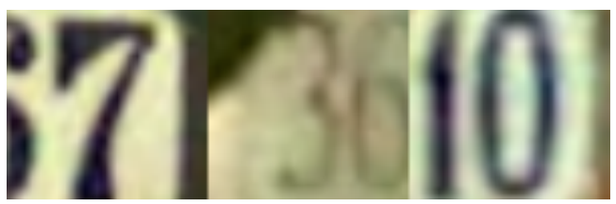
Domain Adaptation / Generalization

Domain adaptation: Training phase

Source domain:

(4, 8, 5)

• • •



(7, 3, 0)

Target domain:

• • •

Domain Adaptation / Generalization

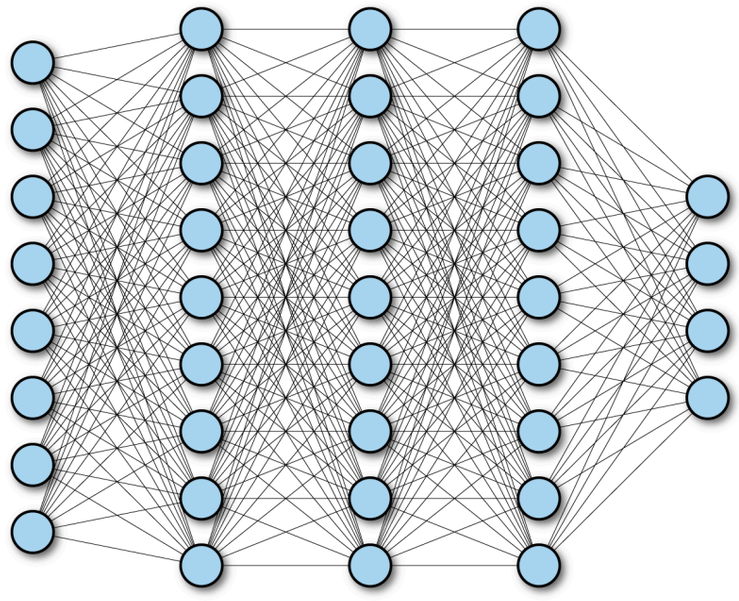
Domain adaptation: Training phase

Source domain:

(4, 8, 5) (7, 3, 0)

+

Target domain:

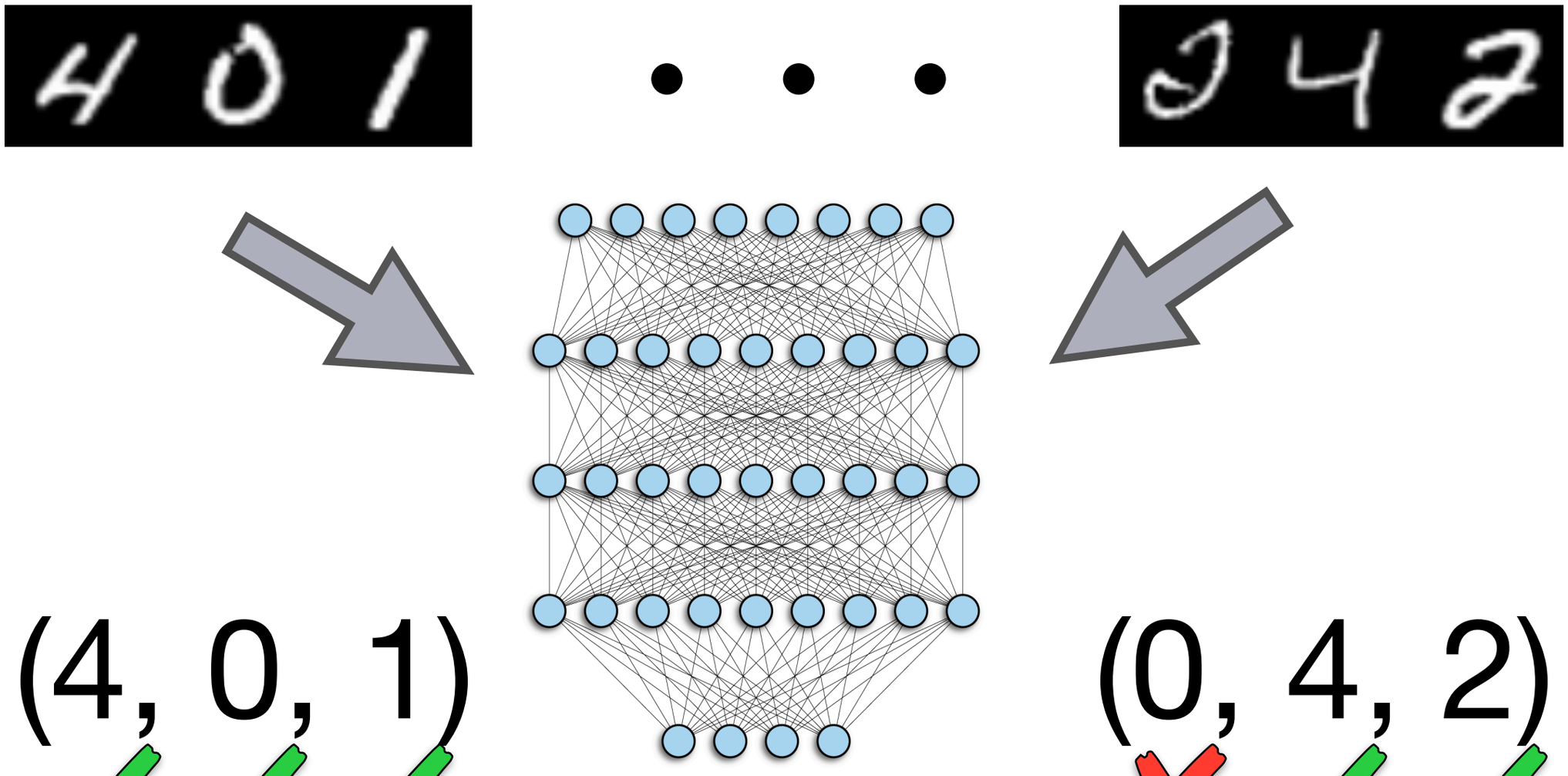


Classifier

Domain Adaptation / Generalization

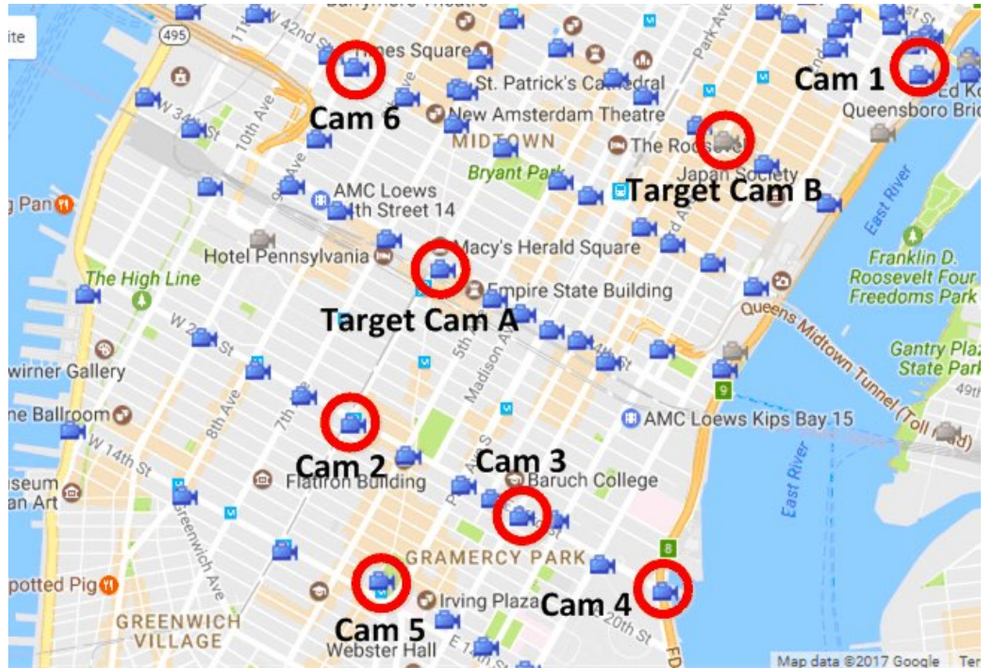
Domain adaptation: Testing phase

Target domain:



Distribution Shift in Practice

Mismatched distributions: Source $\neq$ Target



Is minimizing the source error sufficient for generalization on the target domain?

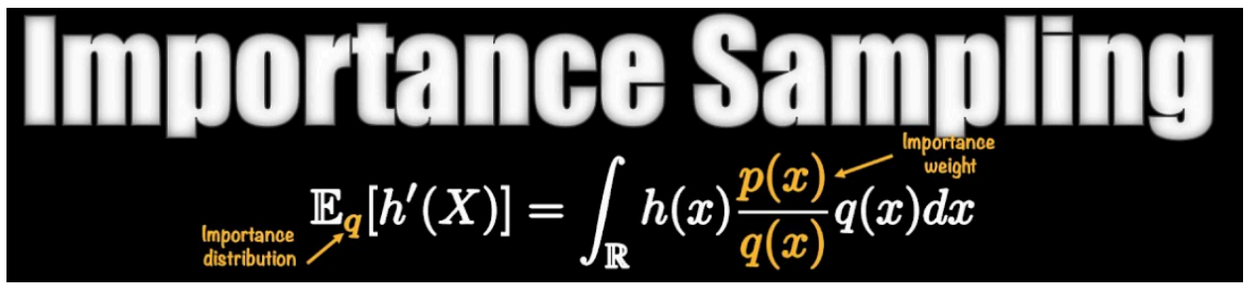
What we want: Generalization on the Target domain

What we have: Labeled data from Source domain + unlabeled data from Target domain

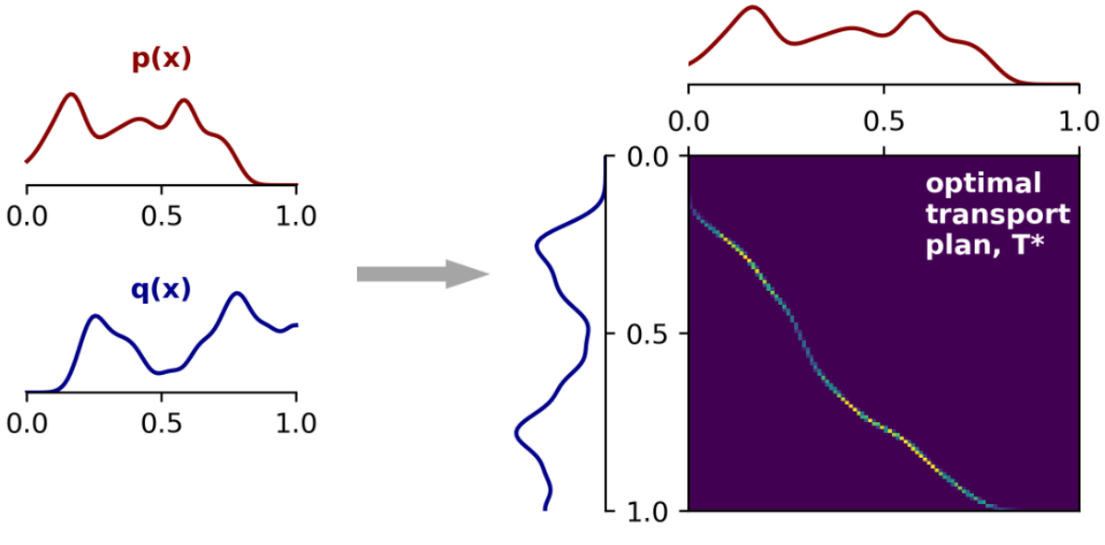
Distribution Shift in Practice

Mismatched distributions: Source $\neq$ Target

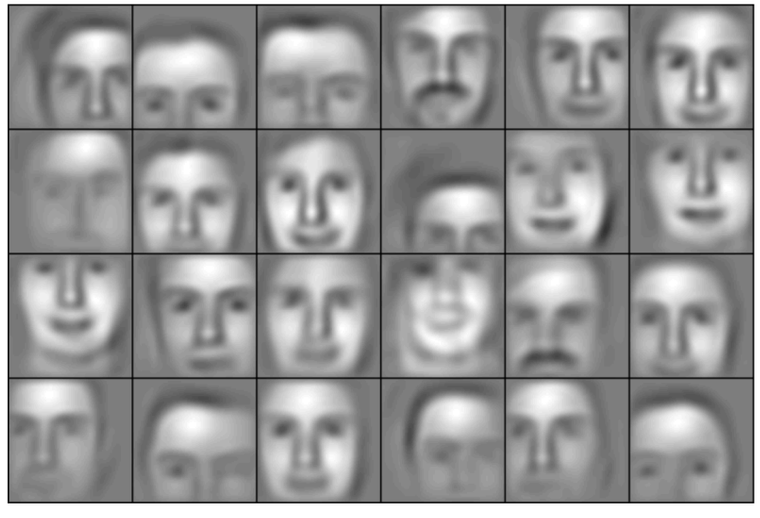
Is minimizing the source error sufficient for generalization on the target domain?



Importance sampling



Optimal transport



Feature/Representation learning

Domain-Invariant Representation Learning

How to bridge the gap?

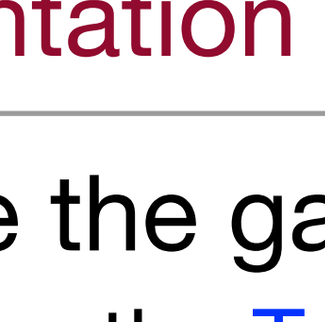
What we want: Generalization on the Target domain

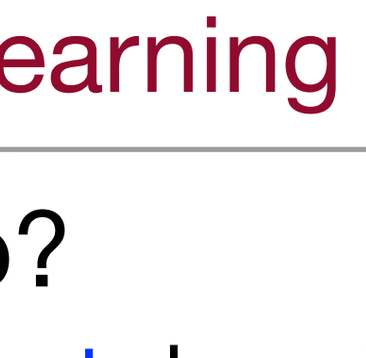
What we have: Labeled data from Source domain + unlabeled data from Target domain

Make the two domains closer

Source

$g_S(\cdot)$

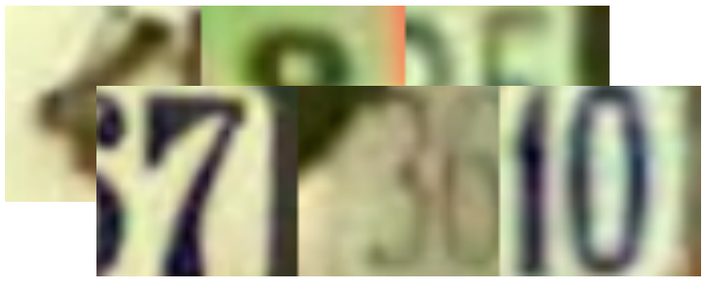




Target

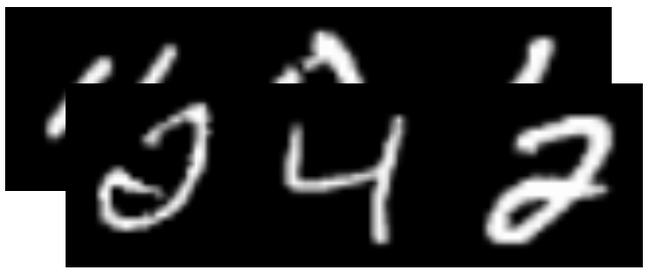
$g_T(\cdot)$

Domain-Invariant Representation Learning



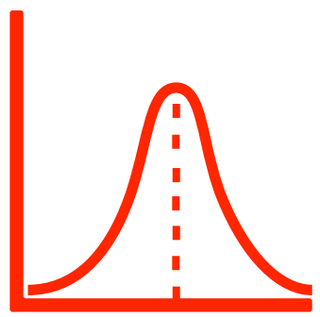
Source

$g_S(\cdot)$

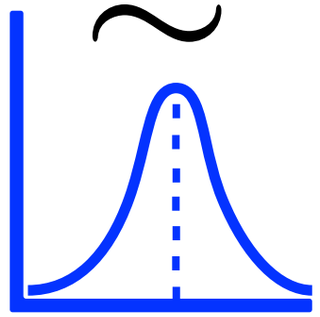


Target

$g_T(\cdot)$



$\approx$



$h(\cdot)$

$\varepsilon_S(h \circ g_S) = \ell(h(g_S(\mathbf{x})), y)$

Supervised training

$= h \circ g_T$

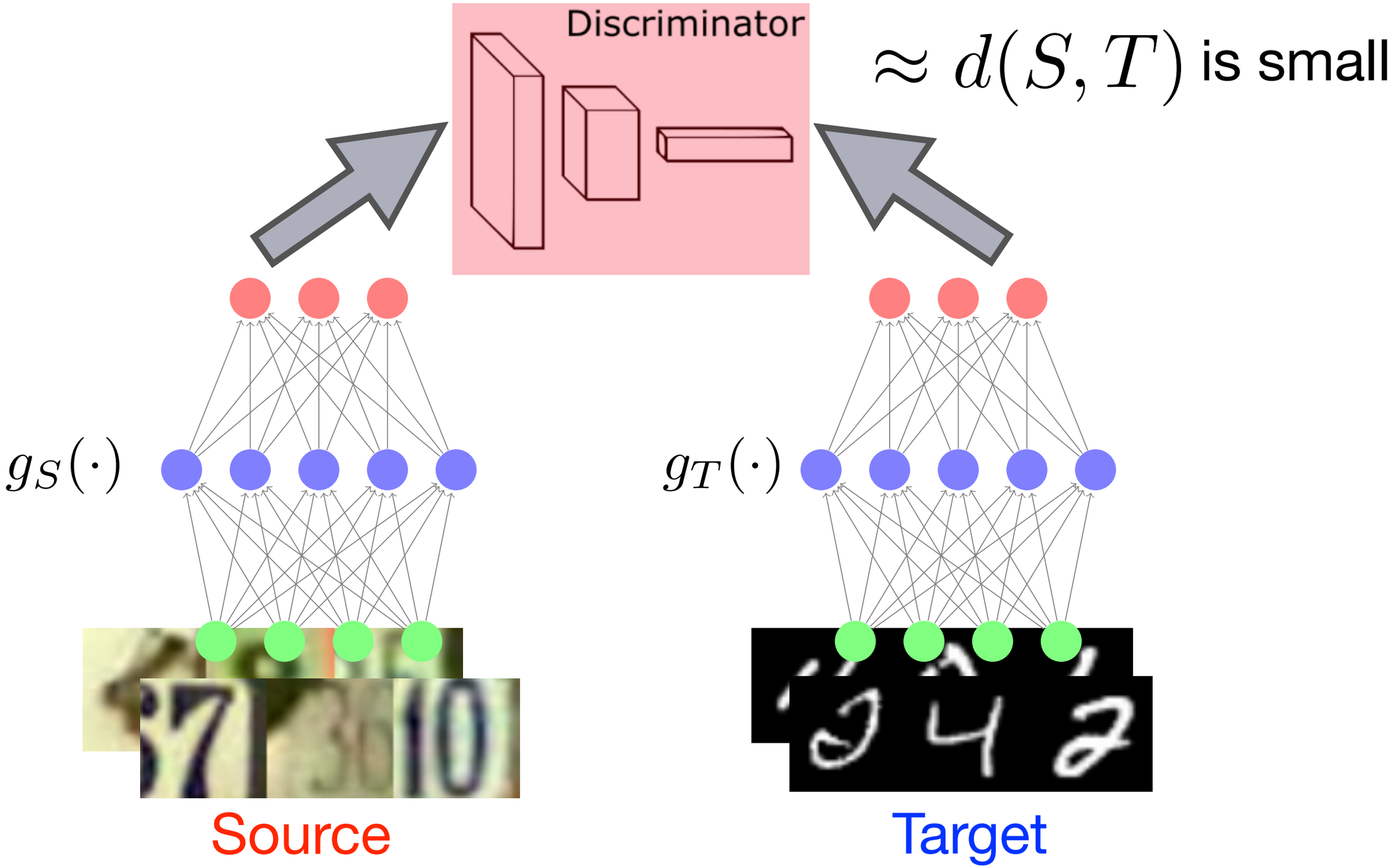
Matching feature distribution

Adversarial discriminator [Ganin et al. ICML15]
Maximum-mean discrepancy [Long et al. ICML 15]
Transportation distance [Shen et al. AAAI 18]
Multiple-domain extensions [Zhao et al. NeurIPS 18]

Goal: generalization on target domain

Domain-Invariant Representation Learning

[Ganin et al. ICML15] Source or Target?

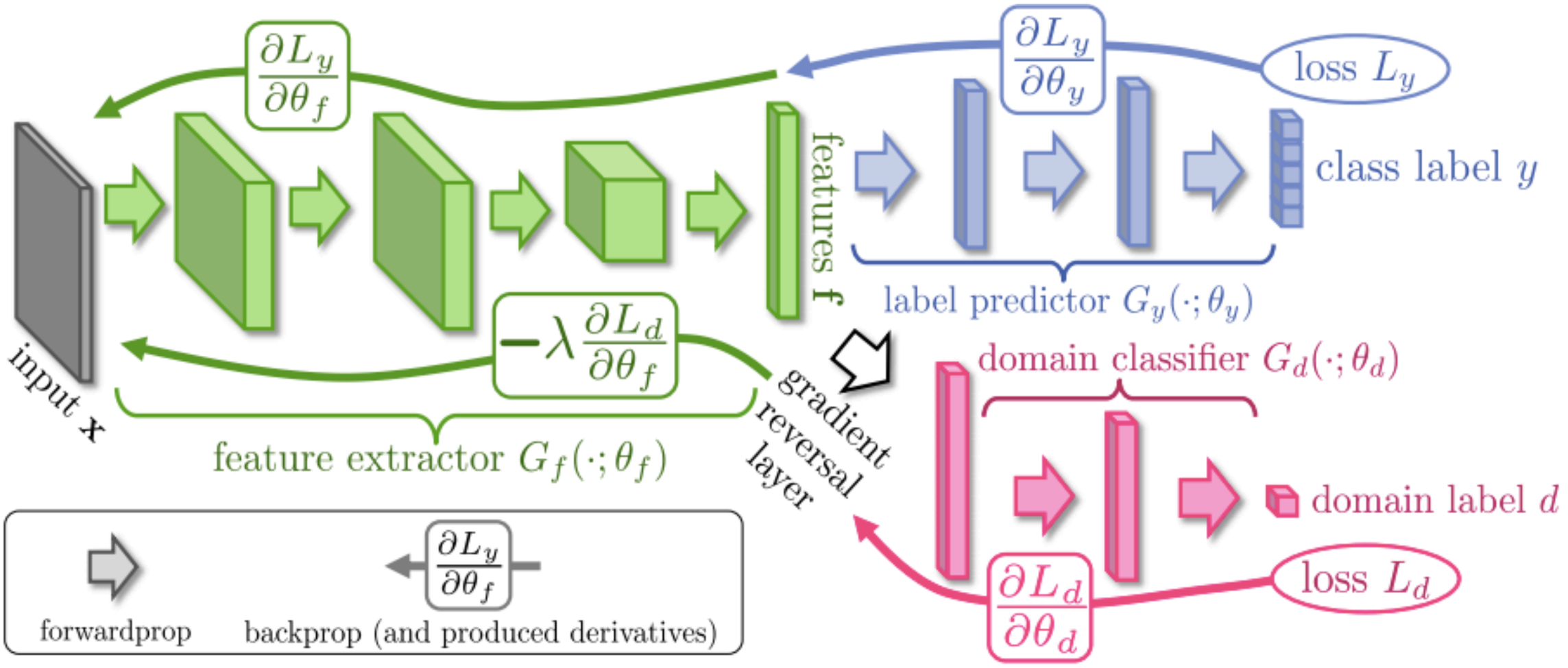


"Unsupervised Domain Adaptation by Backpropagation", Ganin and Lempitsky, ICML' 15

Domain-Invariant Representation Learning

Domain-Adversarial Neural Networks (DANN):

- Goal 1: Learn discriminative features for the target task of interest
- Goal 2: Learn domain-invariant features to confuse the domain classifier

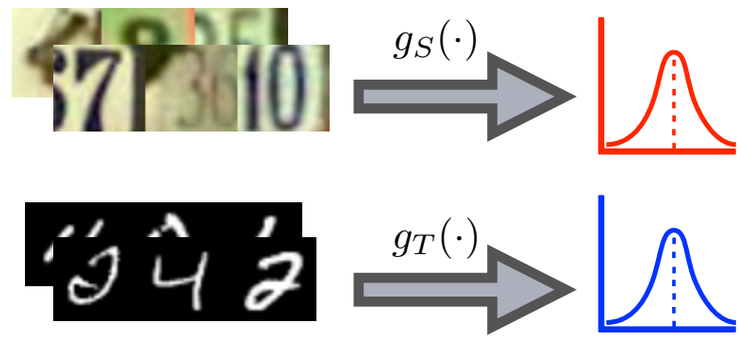


[Figure credit: Ganin et al.' 16]

"Domain-Adversarial Training of Neural Networks", Ganin et al., JMLR' 16

A Theory of Learning from Different Domains

Theorem (Ben-David et al.' 07):



$$\varepsilon_T(h) \leq \varepsilon_S(h) + \frac{1}{2} d(S;T) + \lambda^*$$

- $\varepsilon_T(h)/\varepsilon_S(h)$: true target/source errors
- $d(S;T)$: divergence between target/source input distributions
- $\lambda^* := \min_{h' \in \mathcal{H}} \varepsilon_S(h') + \varepsilon_T(h')$: optimal joint error

Bound-minimizing algorithm:

$$\min \quad \varepsilon_S(h) + \frac{1}{2} d(S;T)$$

error minimization on the source domain

distribution matching between source and target domains

"A theory of learning from different domains", Ben-David et al., MLJ' 10

Domain-Invariant Representation Learning

Source

$g(\cdot)$

$h(\cdot)$

$$\varepsilon_S(h \circ g) = \ell(h(g(\mathbf{x})), y)$$

Supervised training

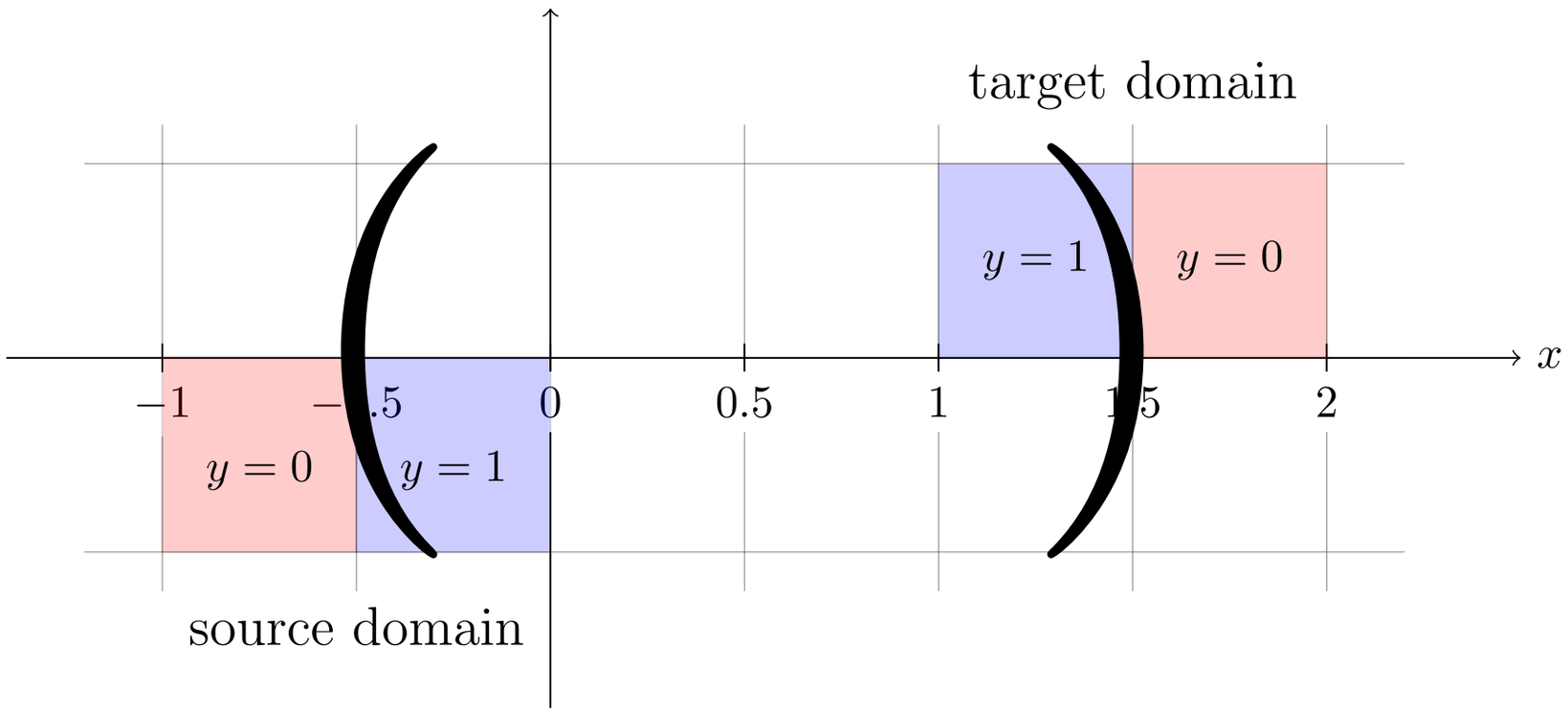
Target

$g(\cdot)$

Question: Is it guaranteed to generalize on target domain?

A simple 1D adaptation example

Before adaptation:



Source: $\mathcal{D}_S = U(-1, 0)$

Target: $\mathcal{D}_T = U(1, 2)$

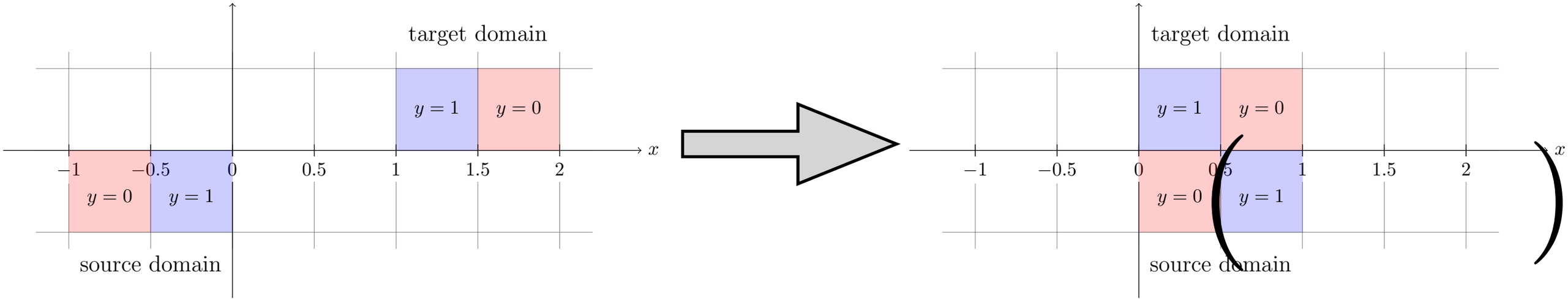
$$h^*(x) = 1 \text{ iff } x \in (-1/2, 3/2)$$

$$\lambda^* = \min_{h'} \varepsilon_S(h') + \varepsilon_T(h') = 0$$

"On Learning Invariant Representations for Domain Adaptation", Zhao et al., ICML' 19

A simple 1D adaptation example

After adaptation:



Bound-minimizing algorithm: $\min \quad \varepsilon_S(h) + \frac{1}{2} d(S;T)$

Source: $\mathcal{D}'_S = U(0,1)$

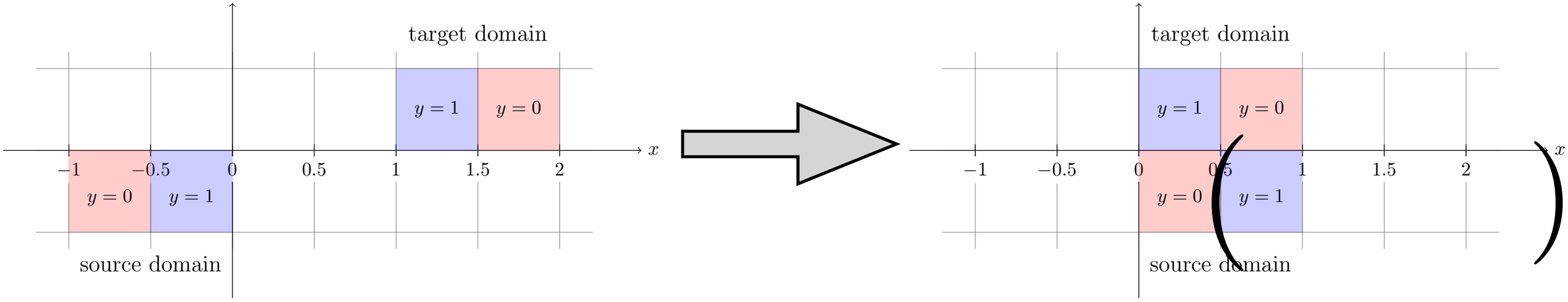
Target: $\mathcal{D}'_T = U(0,1)$

$$d(S,T) = 0$$

If $\varepsilon_S(h) = 0$, then $\varepsilon_T(h) = 1$!

A simple 1D adaptation example

After adaptation:



Bound-minimizing algorithm: $\min \quad \varepsilon_S(h) + \frac{1}{2} d(S;T)$

What's wrong with this example?

$$\varepsilon_T(h) \leq \varepsilon_S(h) + \frac{1}{2} d(S;T) + \lambda^*$$
$$\| \qquad\qquad \|$$
$$0 \qquad\qquad 0$$

In fact, $\forall h \in 2^{[0,1]}, \quad \varepsilon_S(h) + \varepsilon_T(h) = 1 \Rightarrow \lambda^* = 1$

Key Message: Matching features is not sufficient!

An information-theoretic lower bound

Well, that's just a worst-case example, in general it works…

Consider the general adaptation scenario:

$$X \xrightarrow{g} Z \xrightarrow{h} \hat{Y}$$

- $g(\cdot)$ (nonlinear) feature transformation
- $h(\cdot) \in \{0, 1\}$ (randomized) classification function

Theorem [Zhao et al., ICML 19]: suppose the Markov chain holds and $d(\mathcal{D}_S^Y, \mathcal{D}_T^Y) \geq d(\mathcal{D}_S^Z, \mathcal{D}_T^Z)$, then:

$$\varepsilon_S(h \circ g) + \varepsilon_T(h \circ g) \geq \frac{1}{2}\left(d(\mathcal{D}_S^Y, \mathcal{D}_T^Y) - d(\mathcal{D}_S^Z, \mathcal{D}_T^Z)\right)^2$$

Distance between marginal label distributions

Distance between feature distributions

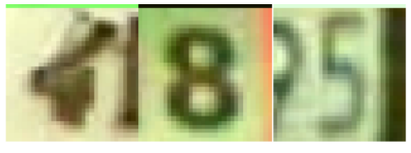
Key Message: the better the distribution alignment & the source risk, the worse the target risk

Empirical results

Digit classification:

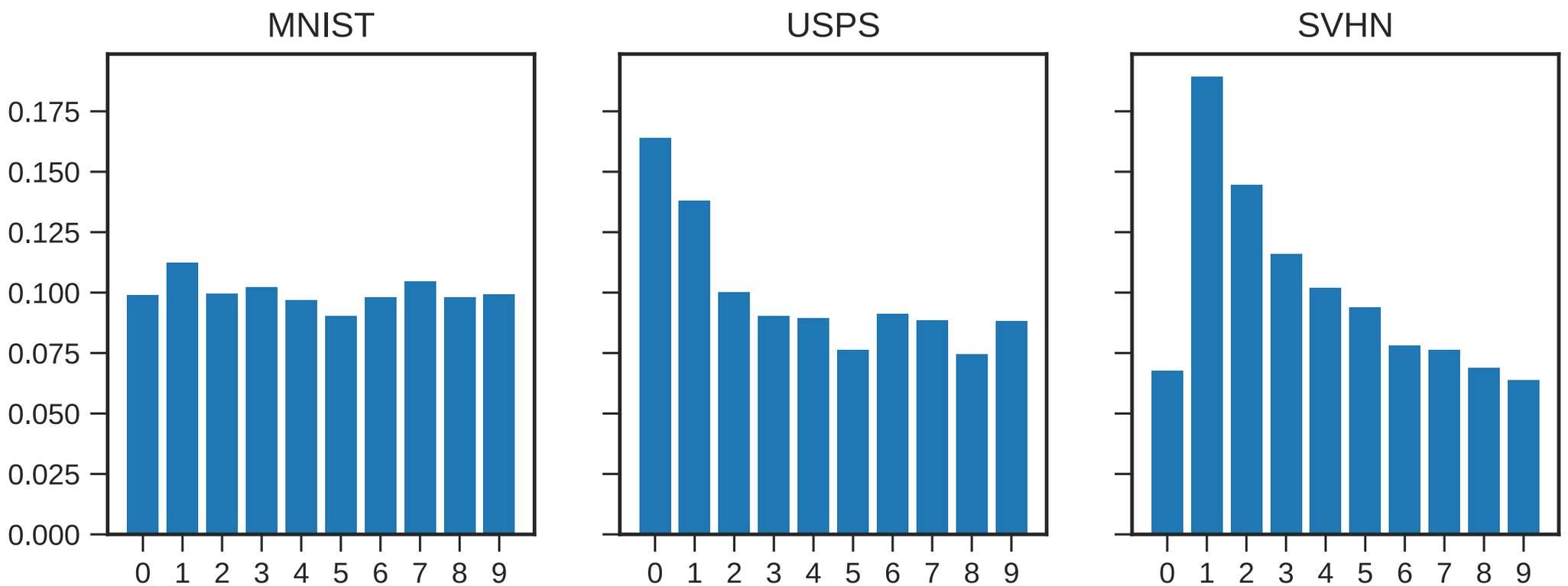
MNIST

SVHN



USPS

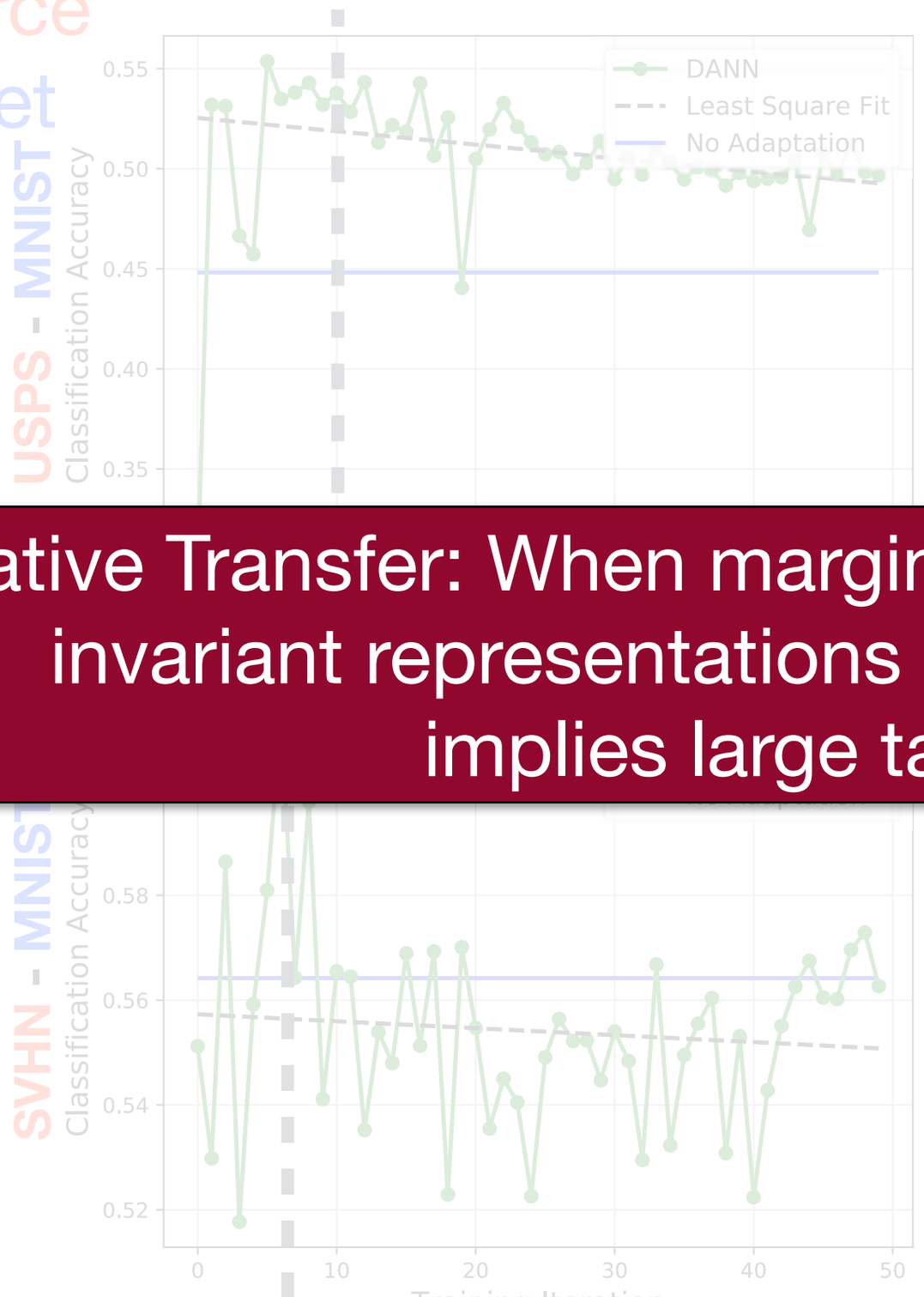
Marginal label distribution:



Negative transfer between source and target

Source

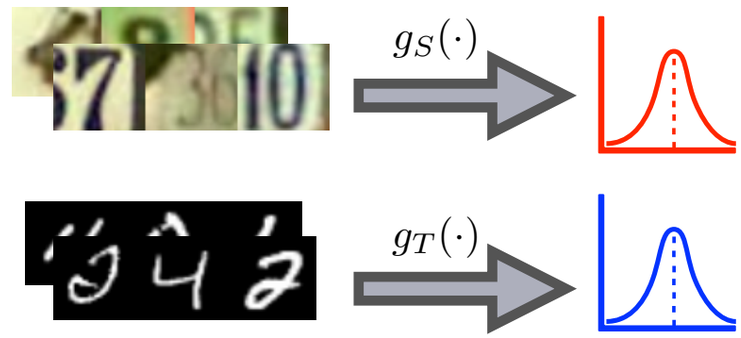
Target



Negative Transfer: When marginal label distributions differ, invariant representations & small training error implies large target error

A Theory of Learning from Different Domains

Theorem (Ben-David et al.' 07):



$$\varepsilon_T(h) \leq \varepsilon_S(h) + \frac{1}{2} d_{\mathcal{X}}(S,T) + \lambda^*_{\mathcal{X}}$$

- $\varepsilon_T(h)/\varepsilon_S(h)$: true target/source errors
- $d_{\mathcal{X}}(S,T)$: divergence between target/source input distributions over $\mathcal{X}$
- $\lambda^*_{\mathcal{X}} := \min_{h':\mathcal{X}\to\{0,1\}} \varepsilon_S(h') + \varepsilon_T(h')$: optimal joint error obtainable from $\mathcal{X}$

Feature transformation

$$\varepsilon_T(h) \leq \boxed{\varepsilon_S(h) + \frac{1}{2} d_{\mathcal{Z}}(S,T)} + \boxed{\lambda^*_{\mathcal{Z}}}$$

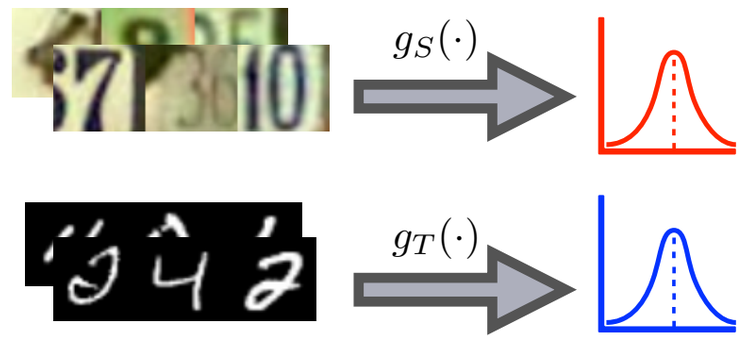
Minimization Blow up

"A theory of learning from different domains", Ben-David et al., MLJ' 10

A Theory of Learning from Different Domains

Theorem (Ben-David et al.' 07):

$$\varepsilon_T(h) \leq \varepsilon_S(h) + \frac{1}{2} d_{\mathcal{X}}(S, T) + \lambda^*_{\mathcal{X}}$$



- $\varepsilon_T(h)/\varepsilon_S(h)$: true target/source errors
- $d_{\mathcal{X}}(S, T)$: divergence between target/source input distributions over $\mathcal{X}$
- $\lambda^*_{\mathcal{X}} := \min_{h':\mathcal{X}\to\{0,1\}} \varepsilon_S(h') + \varepsilon_T(h')$: optimal joint error obtainable from $\mathcal{X}$

Feature transformation

$$\varepsilon_T(h) \leq \varepsilon_S(h) + \frac{1}{2} d_{\mathcal{Z}}(S, T) + \lambda^*_{\mathcal{Z}}$$

Minimization

Blow up

"A theory of learning from different domains", Ben-David et al., MLJ' 10

A Decomposition of the Error Difference

Theorem: Let $\langle \mathcal{D}_S, f_S \rangle$ and $\langle \mathcal{D}_T, f_T \rangle$ be the source and target domains, respectively. For any function class $\mathcal{H} \subseteq [0,1]^{\mathcal{X}}$, and $\forall h \in \mathcal{H}$, the following inequality holds:

$$\varepsilon_T(h) \leq \varepsilon_S(h) + d_{\tilde{\mathcal{H}}}(\mathcal{D}_S, \mathcal{D}_T) + \min\left\{\mathbb{E}_{\mathcal{D}_S}[|f_S - f_T|], \mathbb{E}_{\mathcal{D}_T}[|f_S - f_T|]\right\}$$

where $\tilde{\mathcal{H}} := \{\text{sgn}(|h(x) - h'(x)| - t) : h, h' \in \mathcal{H}, t \in [0,1]\}$

- $\mathcal{D}_S(\mathcal{D}_T)$ marginal distributions over features in source (target) domains
- $f_S(f_T)$ optimal predictors from features to labels in source (target) domains
- The output of the predictor could be probabilistic, i.e., [0, 1]

"A theory of learning from different domains", Ben-David et al., MLJ' 10
"On Learning Invariant Representations for Domain Adaptation", Zhao et al., ICML' 19

A Decomposition of the Error Difference

Theorem: Let $\langle \mathcal{D}_S, f_S \rangle$ and $\langle \mathcal{D}_T, f_T \rangle$ be the source and target domains, respectively. For any function class $\mathcal{H} \subseteq [0,1]^{\mathcal{X}}$, and $\forall h \in \mathcal{H}$, the following inequality holds:

$$\varepsilon_T(h) \leq \varepsilon_S(h) + d_{\tilde{\mathcal{H}}}(\mathcal{D}_S, \mathcal{D}_T) + \min\{\mathbb{E}_{\mathcal{D}_S}[|f_S - f_T|], \mathbb{E}_{\mathcal{D}_T}[|f_S - f_T|]\}$$

where $\tilde{\mathcal{H}} := \{\text{sgn}(|h(x) - h'(x)| - t) : h, h' \in \mathcal{H}, t \in [0,1]\}$

$d(\phi(X_S), \phi(X_T))$

$d(\mathbb{E}[Y_S|\phi(X_S)], \mathbb{E}[Y_T|\phi(X_T)])$

- An error decomposition theorem
- The first term ~ Invariant representations
- The second term ~ Invariant predictors
- Free of the original λ^* term

Could extend to a finite sample bound using standard generalization theory

Invariant Risk Minimization

Bayesian Network: a graph used to encode the conditional independence in a joint distribution

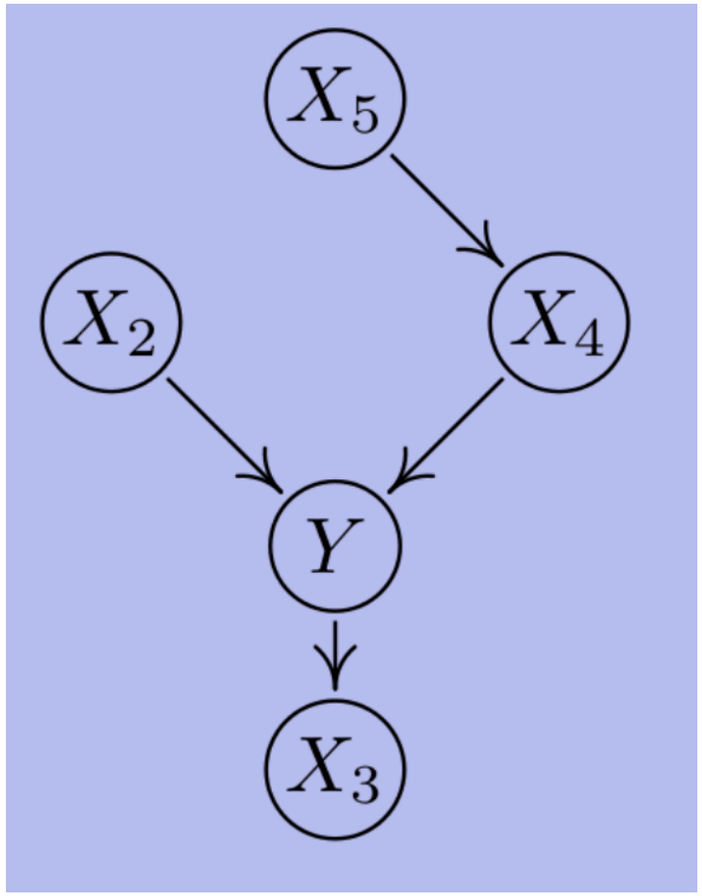
- A directed acyclic graph $\mathcal{G}$, where each node $X \in \mathcal{G}$ corresponds to an RV
- Let $\mathrm{Pa}(X)$ denote the parents of node X, then each node is associated with a conditional probability distribution (CPD): $\Pr(X \mid \mathrm{Pa}(X))$

Let $X_1, \ldots, X_n$ be a topological ordering of all the nodes in $\mathcal{G}$, then the following decomposition of the joint probability holds:

$$\Pr(X_1, \ldots, X_n) = \prod_{i=1}^{n} \Pr(X_i \mid X_1, \ldots, X_{i-1})$$

$$= \prod_{i=1}^{n} \Pr(X_i \mid \mathrm{Pa}(X_i))$$

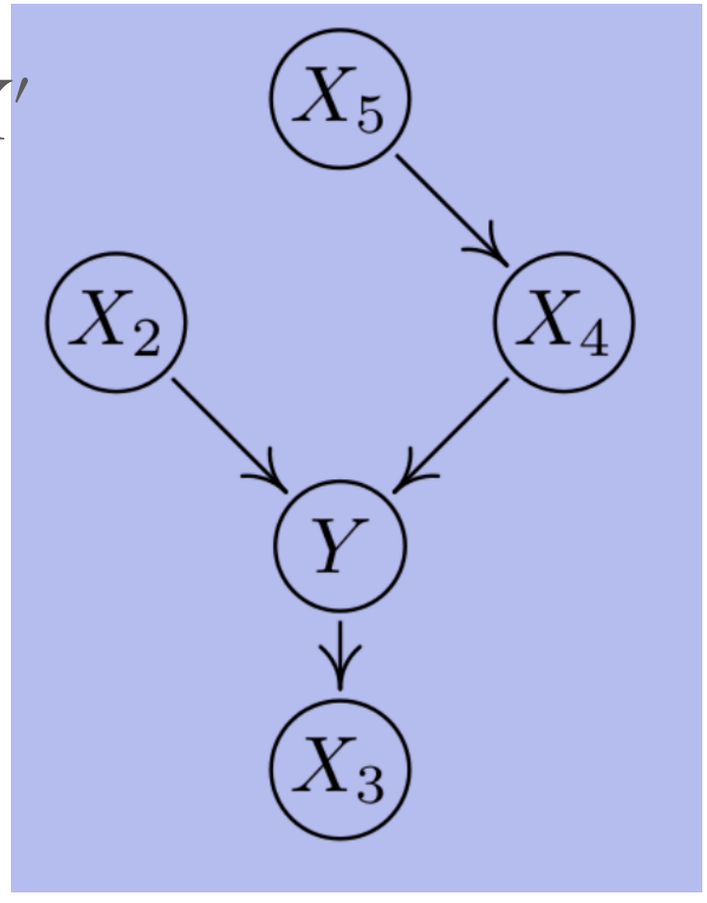
In short: every RV is conditionally independent of non-descendants given its parents.



Invariant Risk Minimization

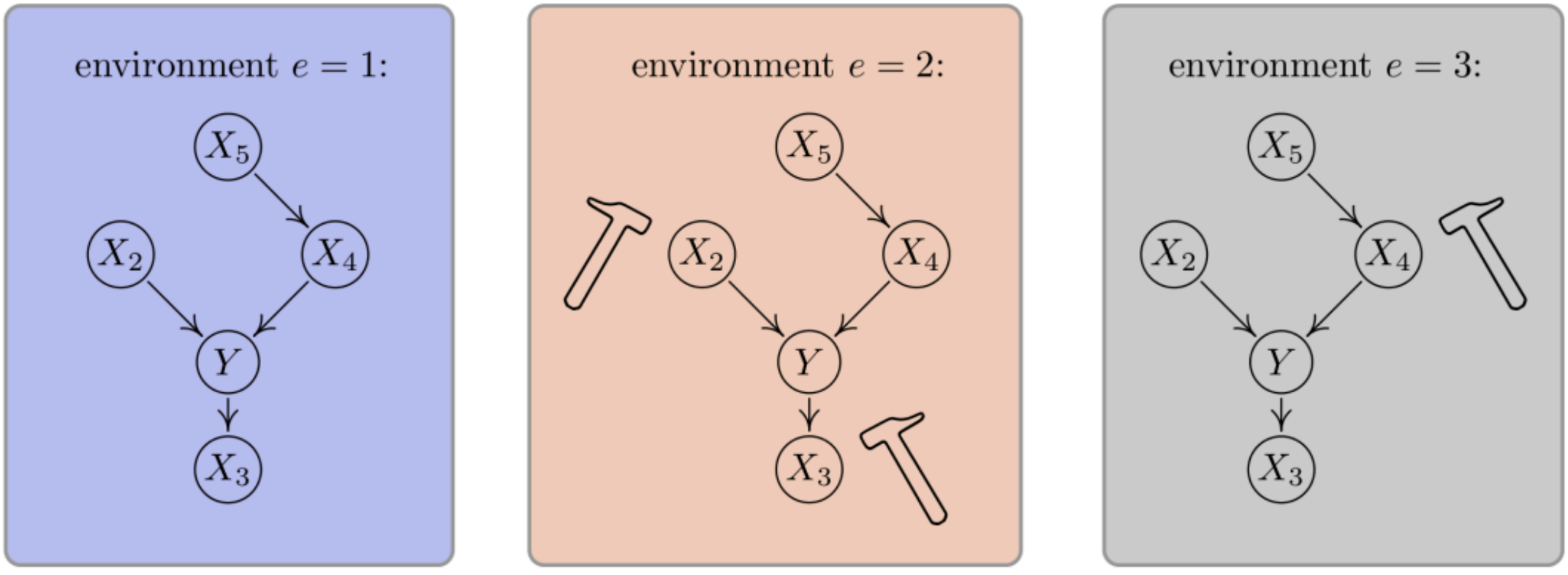
Structural Causal Model (SCM):

- A directed acyclic graph $\mathcal{G}$, where each node $X \in \mathcal{G}$ corresponds to an RV (called Endogenous variables)
- At each node X, there is also a noise RV U_X (called Exogenous variables)
- If $X \to X'$ in $\mathcal{G}$, then $\exists f : X \times U_X \to X'$ where $X' = f(X, U_X)$ (we call X is the cause of X'
- If $X \to X'$ then we say X has a causal effect on X'



Invariant Risk Minimization

Structural Causal Model (SCM):



$$Y = f(X_2, X_4)$$

Suppose from e = 1 we learned:

- What if in e = 2, we fix the value of X2?
- What if in e = 3, we fix the value of X4?
- More generally, what if we arbitrarily change the marginal distributions of X2, X4 and others (except Y)?

Invariant Risk Minimization

Given the causal graph, how to find the subset of causal features?

- Find a subset of the input data, such that the optimal predictor of the target given the subset does not change across environments

Invariant Risk Minimization [Arjovsky et al. 2019]

- Going beyond linear transformations (since subset selection = projection, which is linear)

IRM objective:

$$\min_{\substack{\Phi:\mathcal{X}\to\mathcal{H}\\ w:\mathcal{H}\to\mathcal{Y}}} \quad \sum_{e\in\mathcal{E}_{\text{tr}}} R^e(w\circ\Phi)$$

$$\text{subject to} \quad w \in \arg\min_{\bar{w}:\mathcal{H}\to\mathcal{Y}} R^e(\bar{w}\circ\Phi), \text{ for all } e\in\mathcal{E}_{\text{tr}}.$$

- The same optimal prediction rule applies to all the environments
- Find data representation $\Phi(\cdot)$ that elicits such representations
- Bi-level optimization formulation

"Invariant Risk Minimization", Arjovsky, Bottou, Gulrajani, Lopez-Paz, 2020

Invariant Risk Minimization

Follow-up works on relaxing/simplifying the bi-level optimization framework of IRM:

- Minimization of conditional mutual information: “Invariant Rationalization”, Chang, Zhang, Jaakkola, ICML’ 20
- Minimizing the variance of empirical risks from multiple domains: “Out-of-Distribution Generalization via Risk Extrapolation”, Krueger et al., ICLR’ 20
- Combination of information bottleneck principle:
  - “Invariant Information Bottleneck for Domain Generalization”, Li et al., AAAI’ 22,
  - “Invariance Principle Meets Information Bottleneck for Out-of-Distribution Generalization”, NeurIPS’ 21
- ……

But,

- Again, it does not guarantee small error difference
- Fails when different distributions are not from the same SCM
- Enforcing the invariant optimal predictor is only a necessary condition, but not sufficient

A Unified Perspective from Representation Learning

Can we combine invariant representations and invariant predictors to provide a unified theory for OOD generalization?

Key observation: Let $\phi : \mathcal{X} \to \mathcal{Z}$ be the feature map, E be the RV corresponding to domain index, then

Invariant Representations: $\phi(X) \perp E$

Invariant Predictors: $Y \perp E \mid \phi(X)$

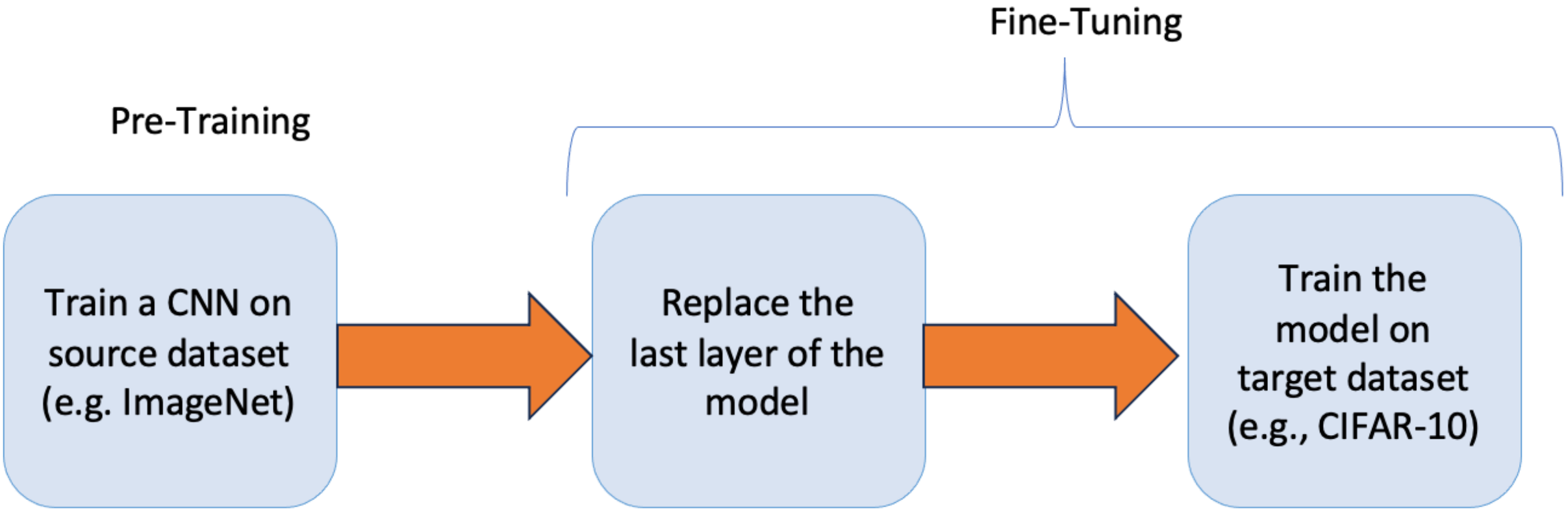
But $(\phi(X), Y)$ determines the joint distribution over features and labels, so if $(\phi(X), Y) \perp E$ then there would be no distributional shift.

In order to explain OOD generalization, we need to consider shifts in both $\phi(X)$ and $Y \mid \phi(X)$.

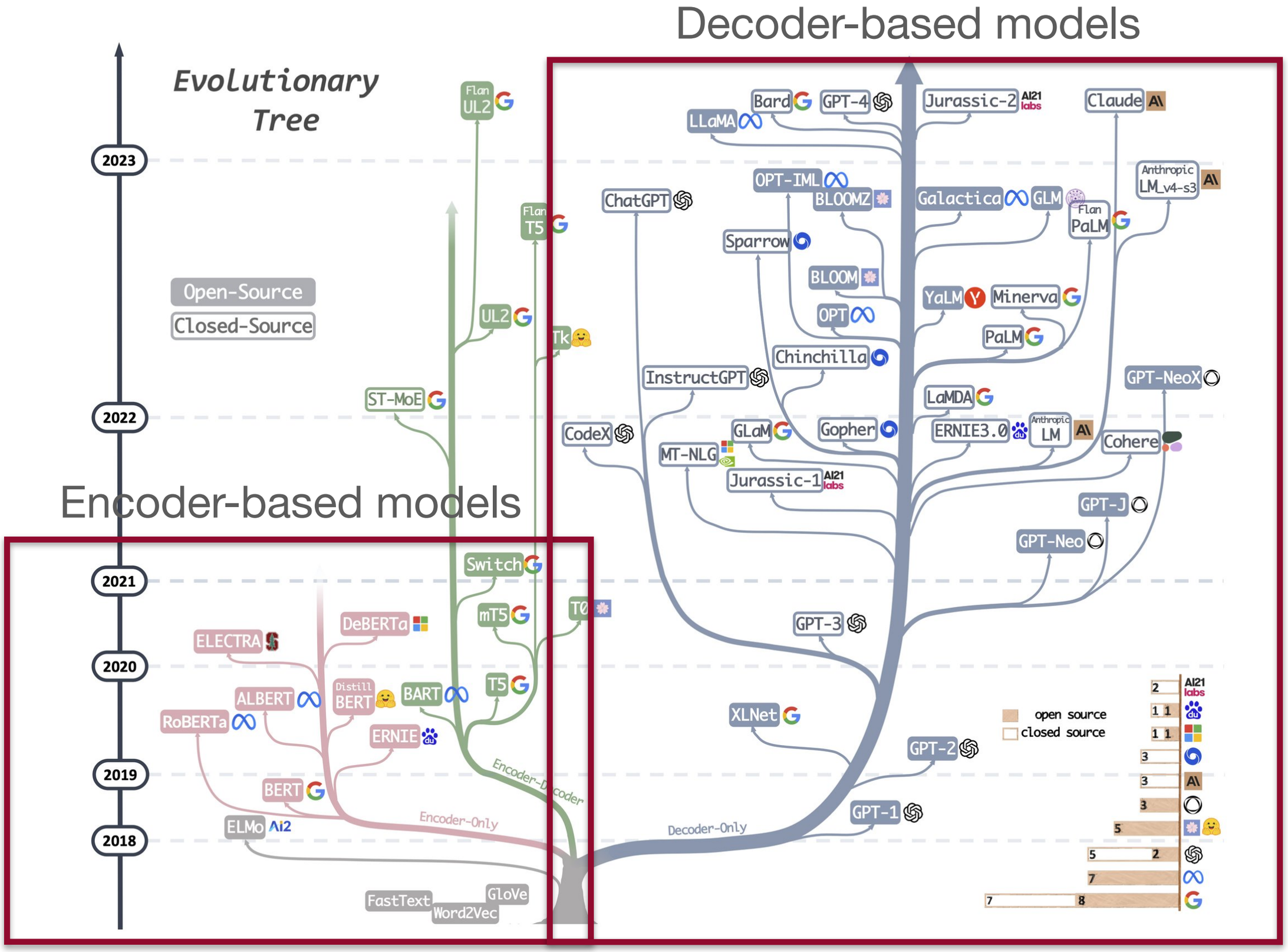
New Paradigms in Transfer Learning

Typical paradigm: pre-training + fine-tuning

1. Pre-training a feature encoder
2. Fine-tune task-specific header for prediction



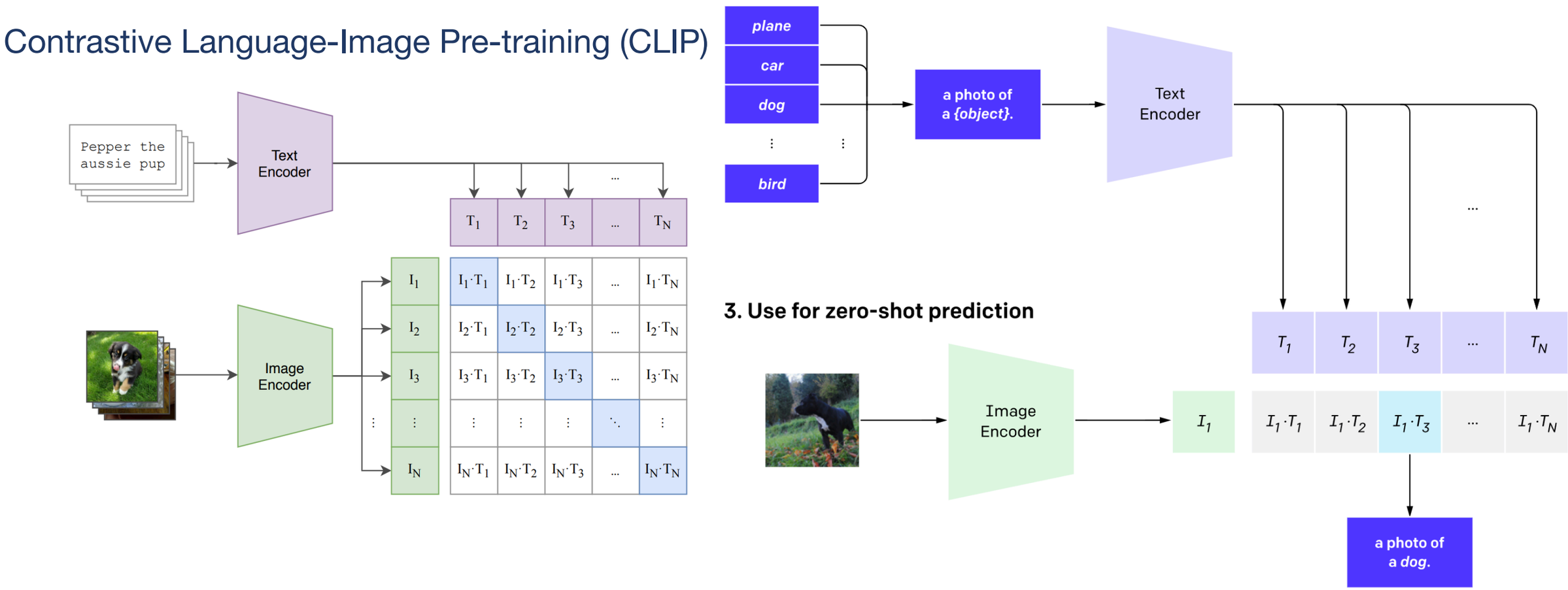
New Paradigms in Transfer Learning



New Paradigms in Transfer Learning

New paradigm: pre-training + task mapping through prompt

1. Pre-training a feature encoder
2. Mapping the class label to a text token, and ask the model to complete the prompt



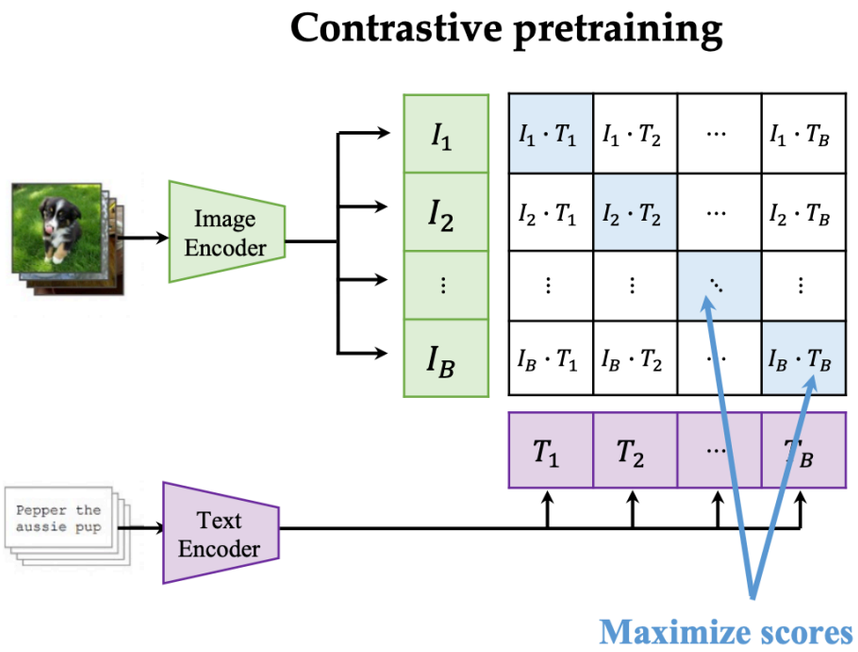
"Learning Transferable Visual Models From Natural Language Supervision", Radford et al., ICML' 21

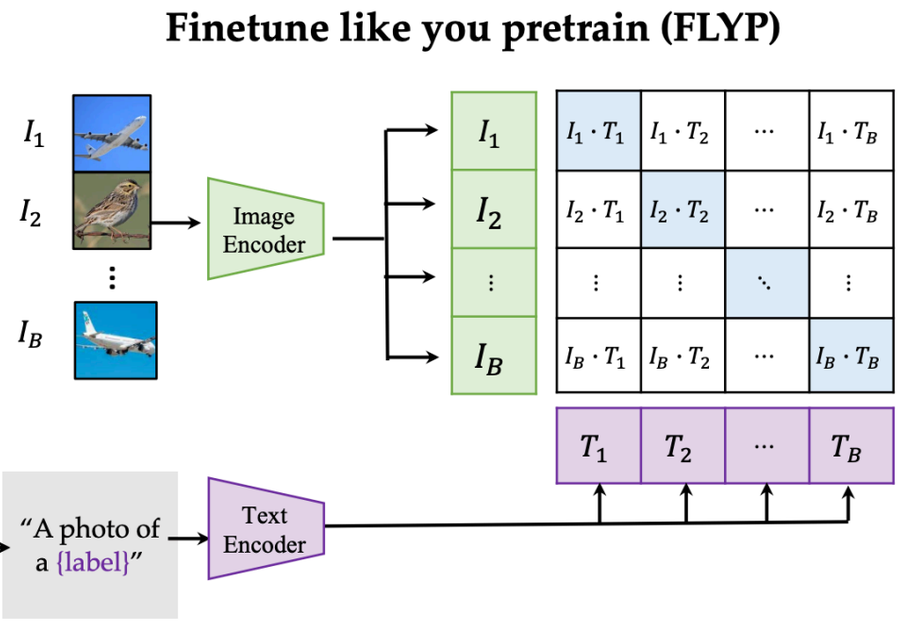
New Paradigms in Transfer Learning

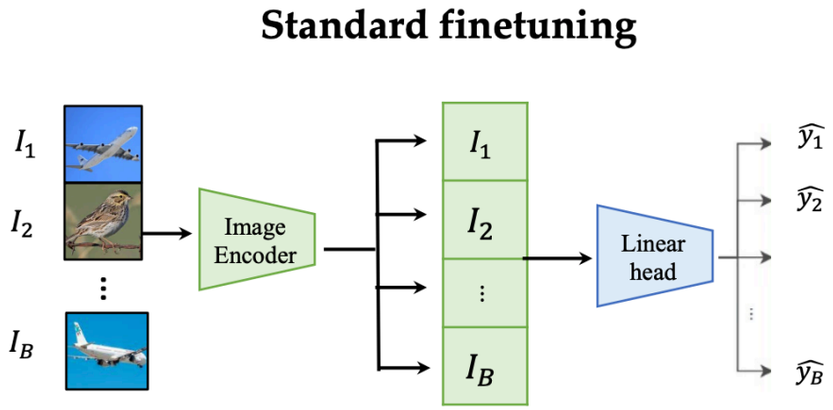
Weight interpolation of model parameters

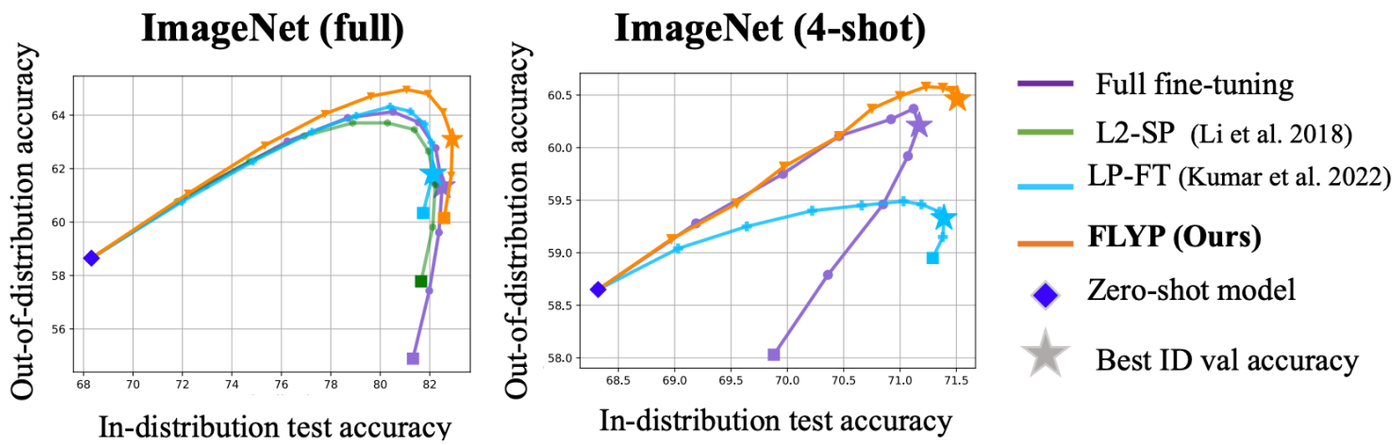
Finetune like you pretrain: Improved finetuning of zero-shot vision models

Sachin Goyal[1], Ananya Kumar[2], Sankalp Garg[1], Zico Kolter[1,3], and Aditi Raghunathan[1]





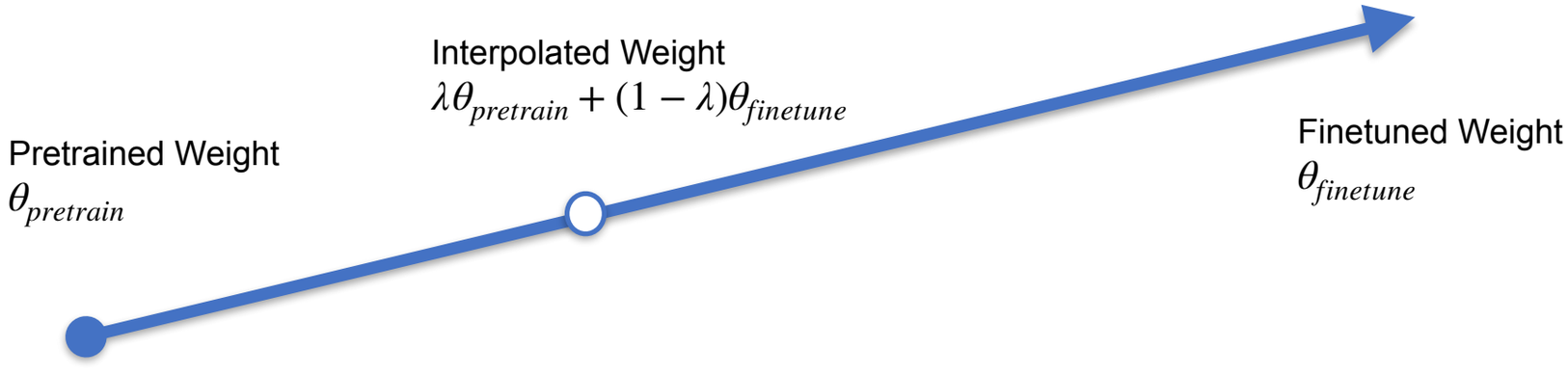




"Finetune like you pretrain: Improved finetuning of zero-shot vision models", Goyal et al., ICCV' 23

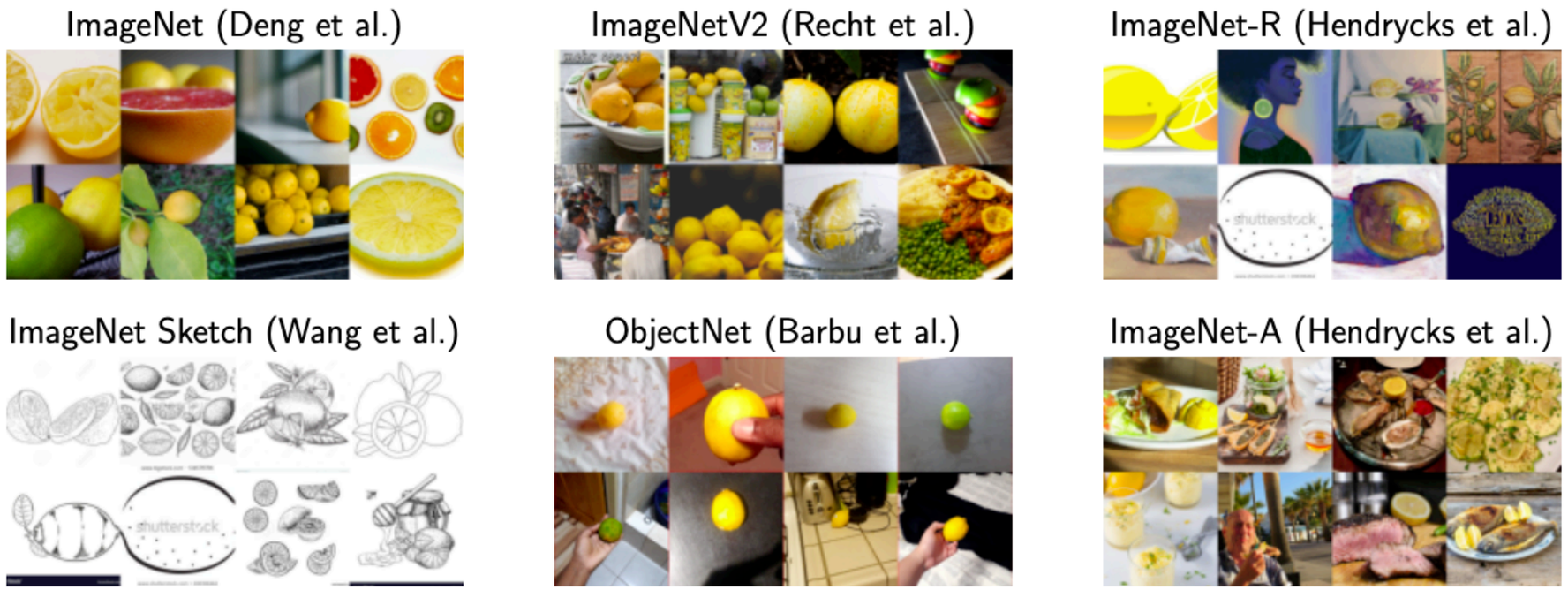
New Paradigms in Transfer Learning

Weight interpolation of model parameters



Weight Interpolation:

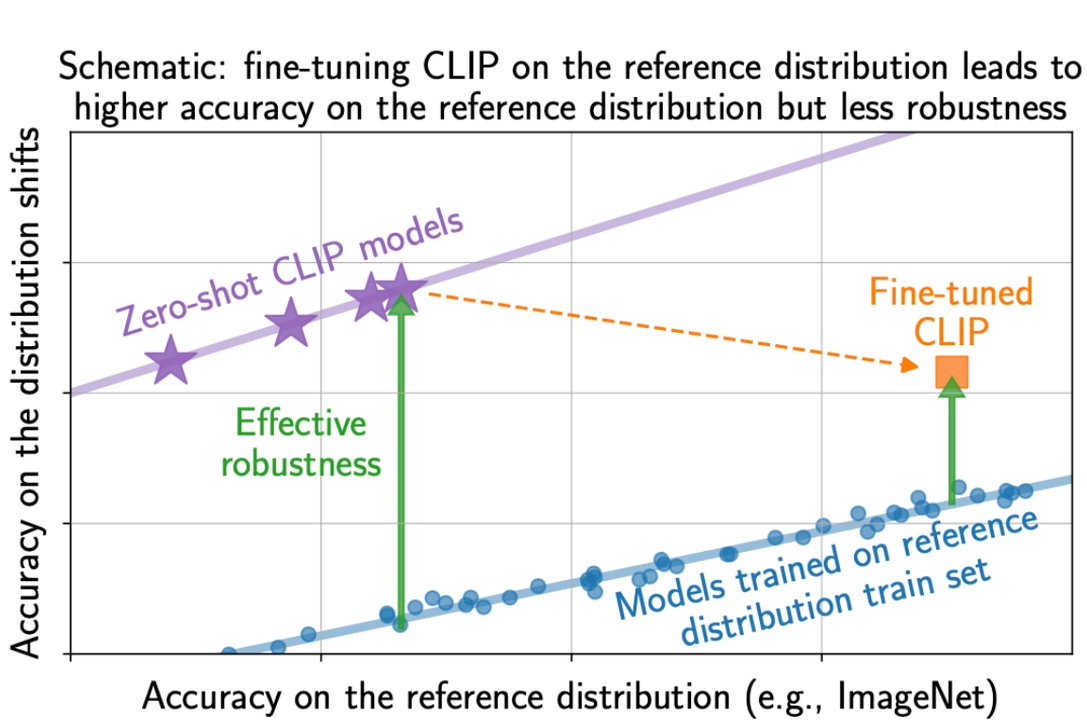
Simply take a linear combination of the pertained weight and fine-tuned weight in the model parameter space.

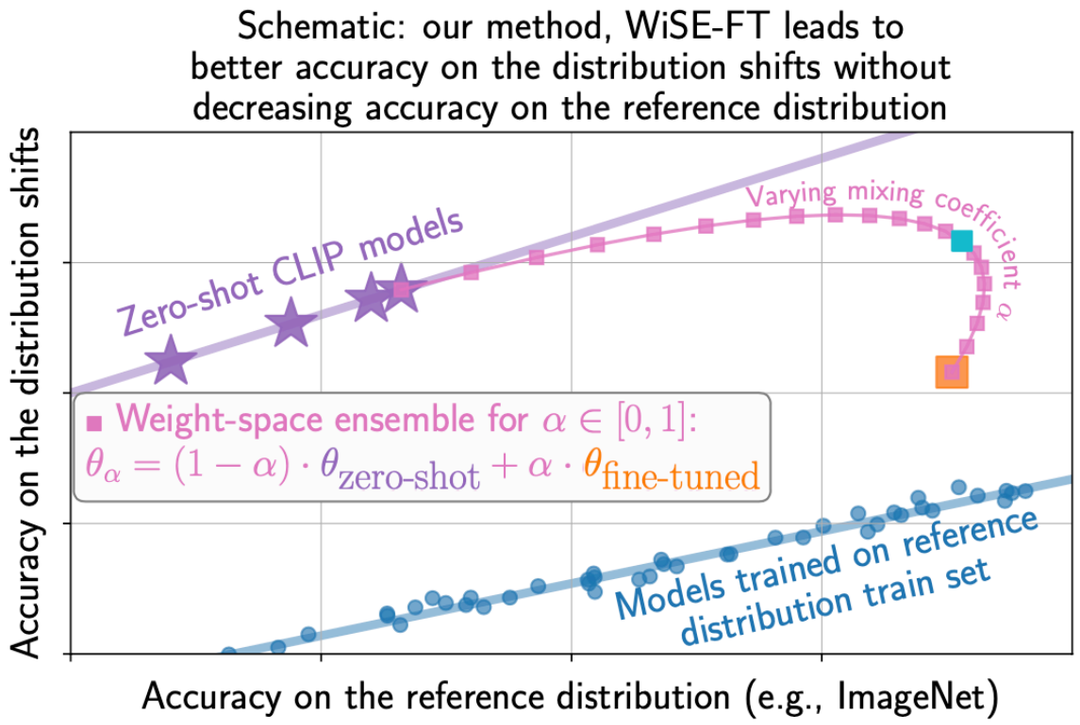


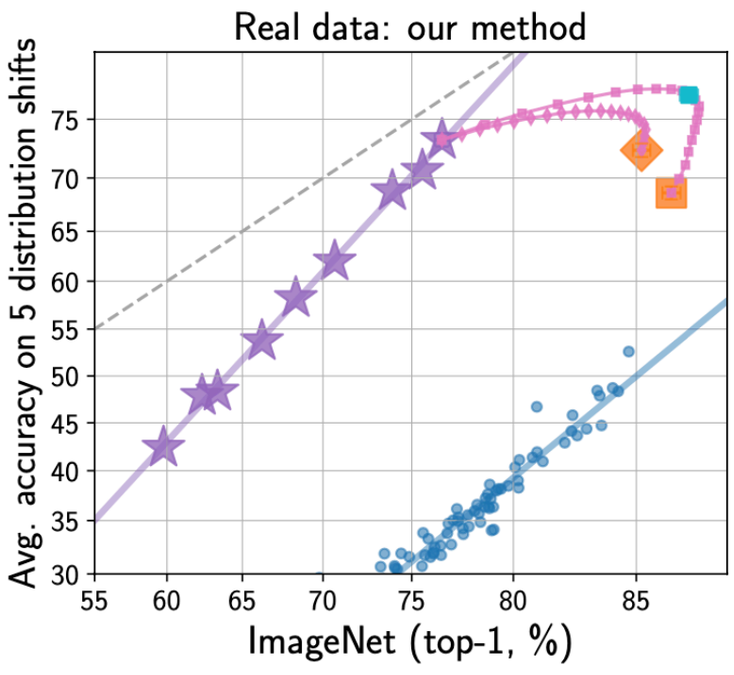
"Finetune like you pretrain: Improved finetuning of zero-shot vision models", Goyal et al., ICCV' 23

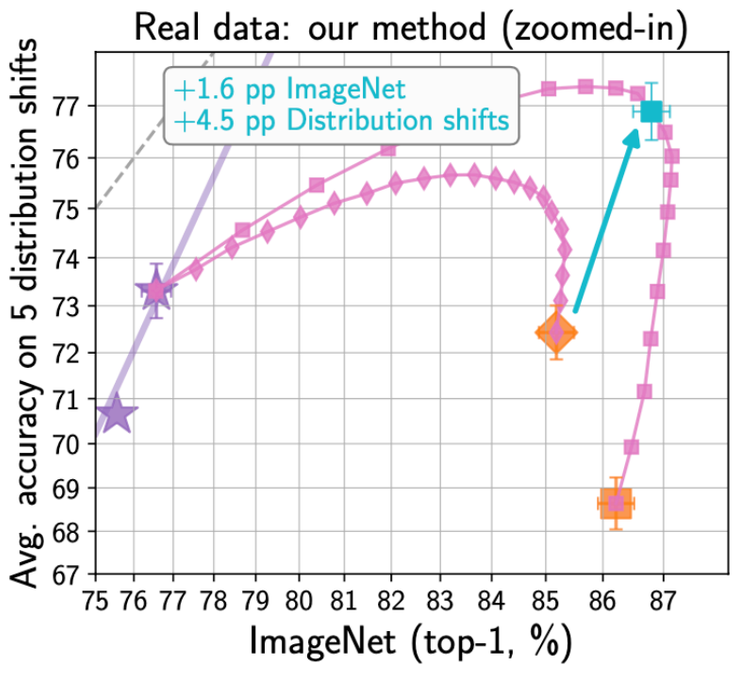
New Paradigms in Transfer Learning

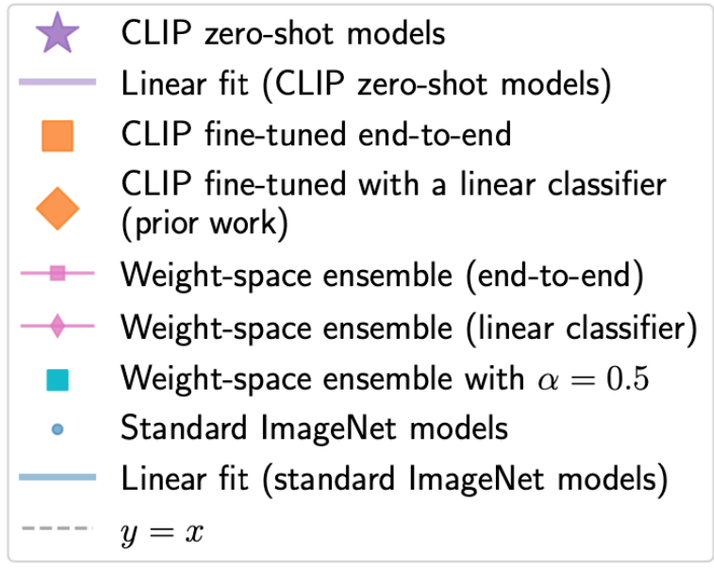
Weight interpolation of model parameters







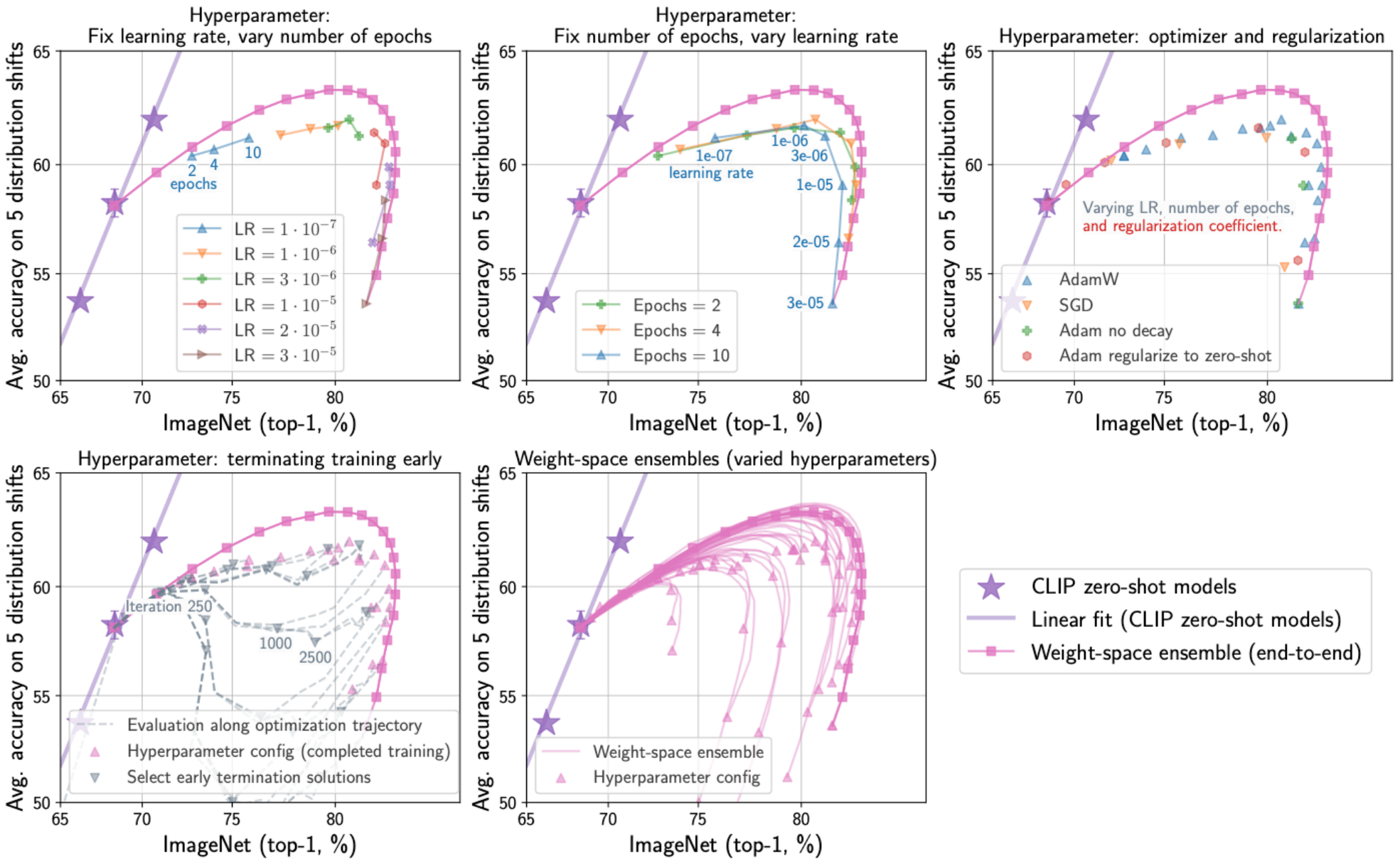




"Finetune like you pretrain: Improved finetuning of zero-shot vision models", Goyal et al., ICCV' 23

New Paradigms in Transfer Learning

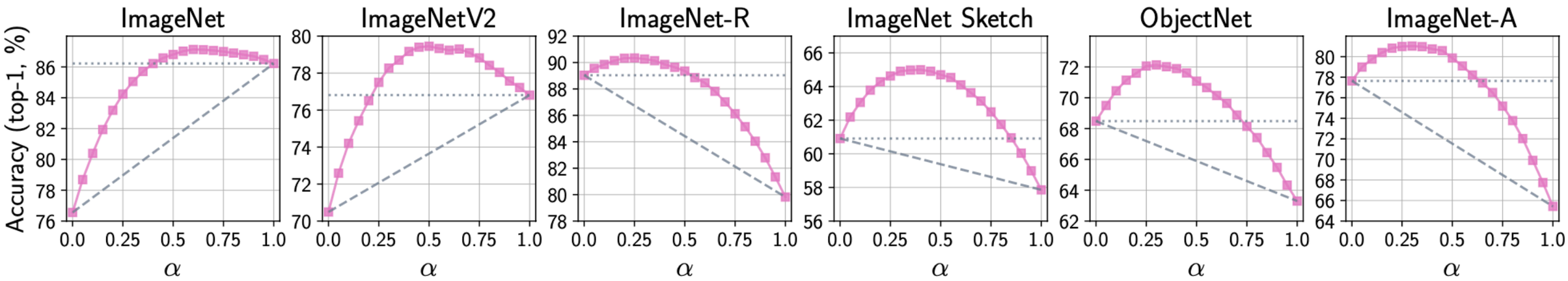
Weight interpolation of model parameters



"Finetune like you pretrain: Improved finetuning of zero-shot vision models", Goyal et al., ICCV' 23

New Paradigms in Transfer Learning

Weight interpolation of model parameters

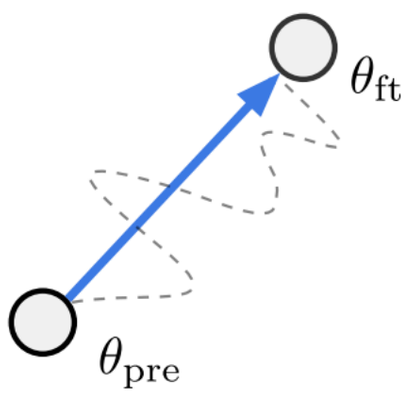


“Finetune like you pretrain: Improved finetuning of zero-shot vision models”, Goyal et al., ICCV’ 23

New Paradigms in Transfer Learning

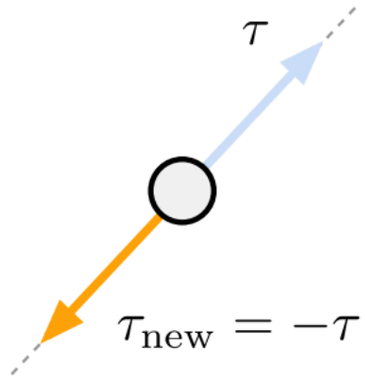
Interpret each task as the difference of model parameters

a) Task vectors



$$\tau = \theta_{\text{ft}} - \theta_{\text{pre}}$$

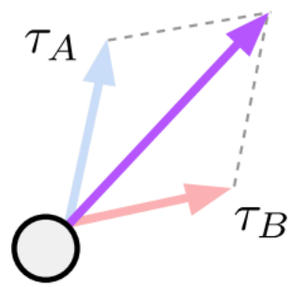
b) Forgetting via negation



Example: making a language model produce less toxic content

c) Learning via addition

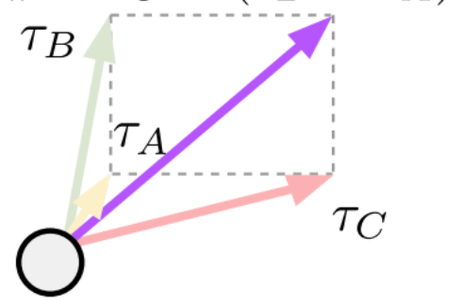
$$\tau_{\text{new}} = \tau_A + \tau_B$$



Example: building a multi-task model

d) Task analogies

$$\tau_{\text{new}} = \tau_C + (\tau_B - \tau_A)$$



Example: improving domain generalization

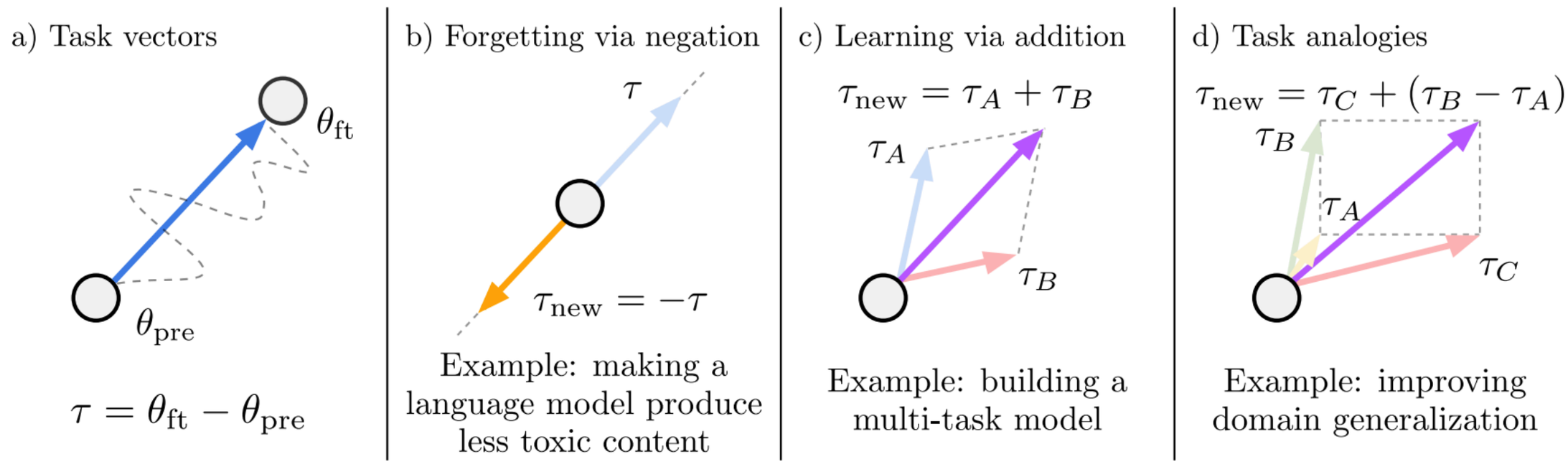
Table 3: **Improving performance on target tasks with external task vectors.** For four text classification tasks from the GLUE benchmark, adding task vectors downloaded from the Hugging Face Hub can improve accuracy of fine-tuned T5 models. Appendix D.6 shows additional details.

Method	MRPC	RTE	CoLA	SST-2	Average
Zero-shot	74.8	52.7	8.29	92.7	57.1
Fine-tuned	88.5	77.3	52.3	94.5	78.1
Fine-tuned + task vectors	89.3 (+0.8)	77.5 (+0.2)	53.0 (+0.7)	94.7 (+0.2)	78.6 (+0.5)

“Editing Models with Task Arithmetic”, Ilharco et al., ICLR’ 23

New Paradigms in Transfer Learning

Interpret each task as the difference of model parameters



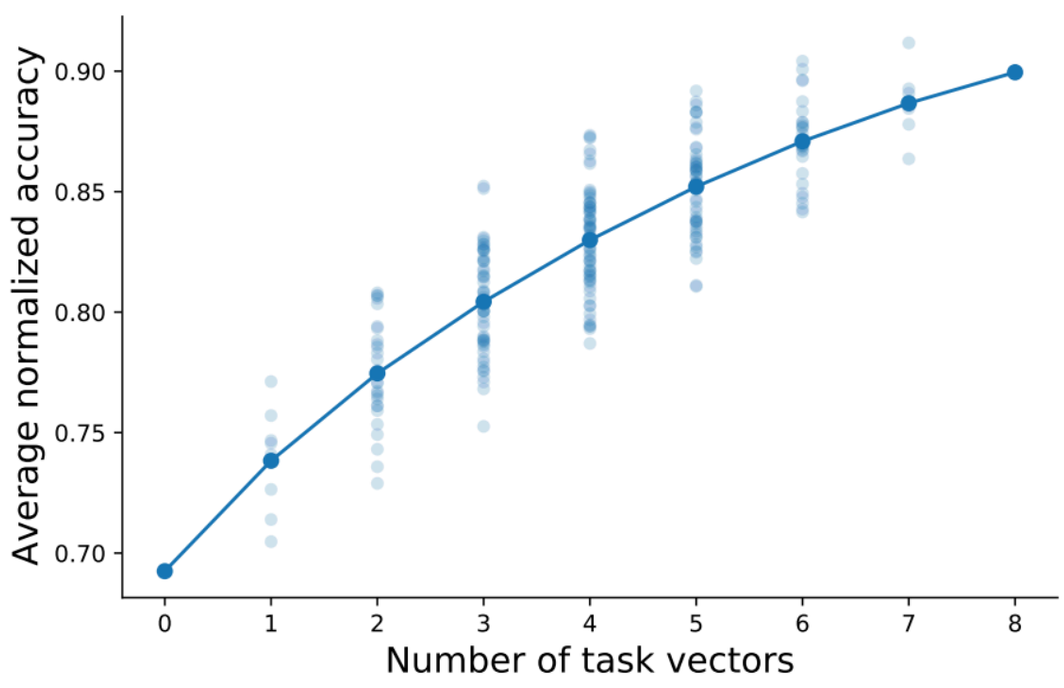


Figure 3: **Adding task vectors builds multi-task models** for image classification tasks. Accuracy is averaged over all downstream tasks. When more task vectors are available, better multi-task vectors can be built. Each point represents an experiment with a subset of the eight tasks we study, and the solid line connects the average performance for each subset size. Recall that the average normalized accuracy of using multiple fine-tuned models is always one. Additional details and experiments are in Appendix D.

"Editing Models with Task Arithmetic", Ilharco et al., ICLR' 23

Open Questions

Why and when task arithmetic work?

- Does this imply a linear relationship between task and model parameter?
- How to enable more fine-grained task arithmetic?
- Is it specific to “foundation models”?
- Connection to flat-minima? Sharpness-aware minimization?
- …