Domain Adaptation :
Under what conditions can a model perform well on a dataset with a different distribution?

DMQA Open Seminar (2024. 03. 29)

Data Mining & Quality Analytics Lab.

김지현

발표자 소개

❖ 김지현 (Jihyun Kim)

- 고려대학교 산업경영공학과 대학원 재학
- Data Mining & Quality Analytics Lab. (김성범 교수님)
- Ph.D. Student (2022.03 ~ Present)

❖ Research Interest

- Domain Adaptation
- Domain Generalization

❖ Contact

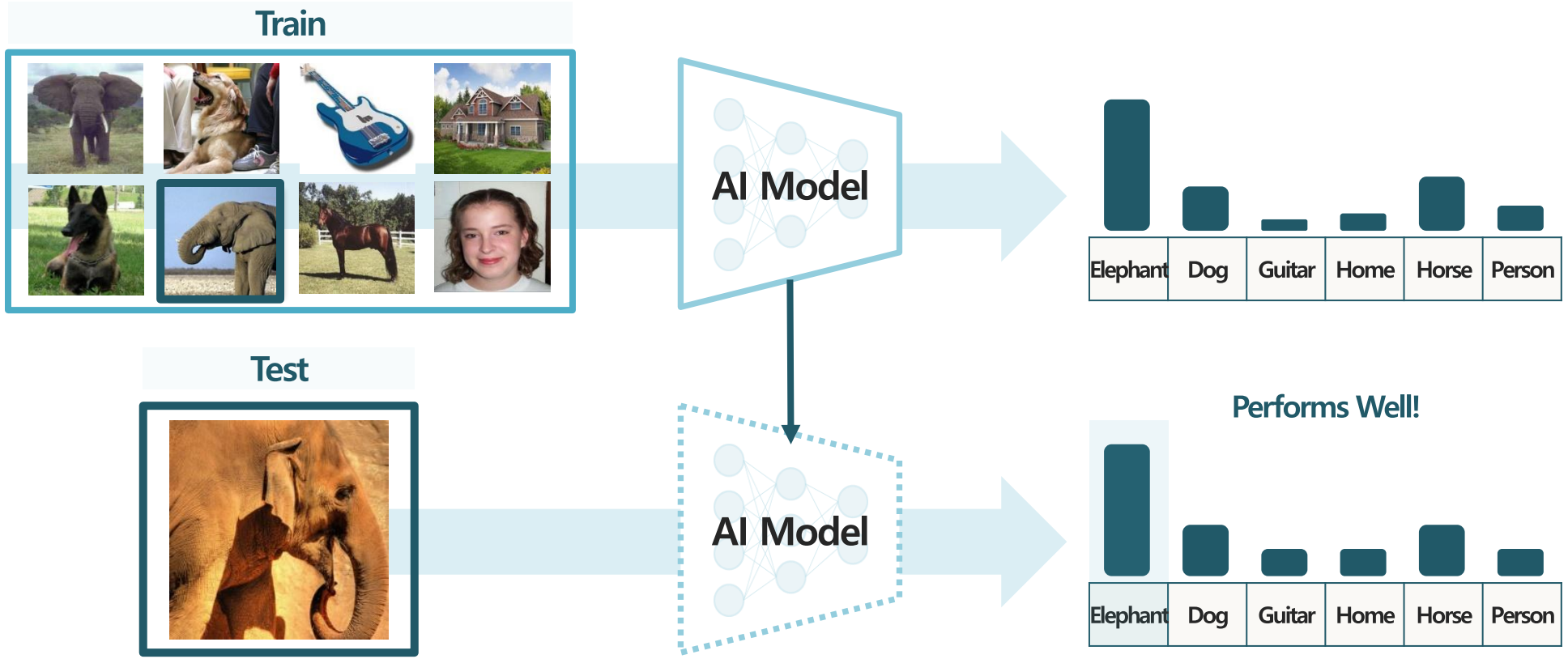
- jihyun_k@korea.ac.kr

Contents

Introduction

Background on Domain Adaptation

We can train an AI model on the training data and directly apply it to the test data!

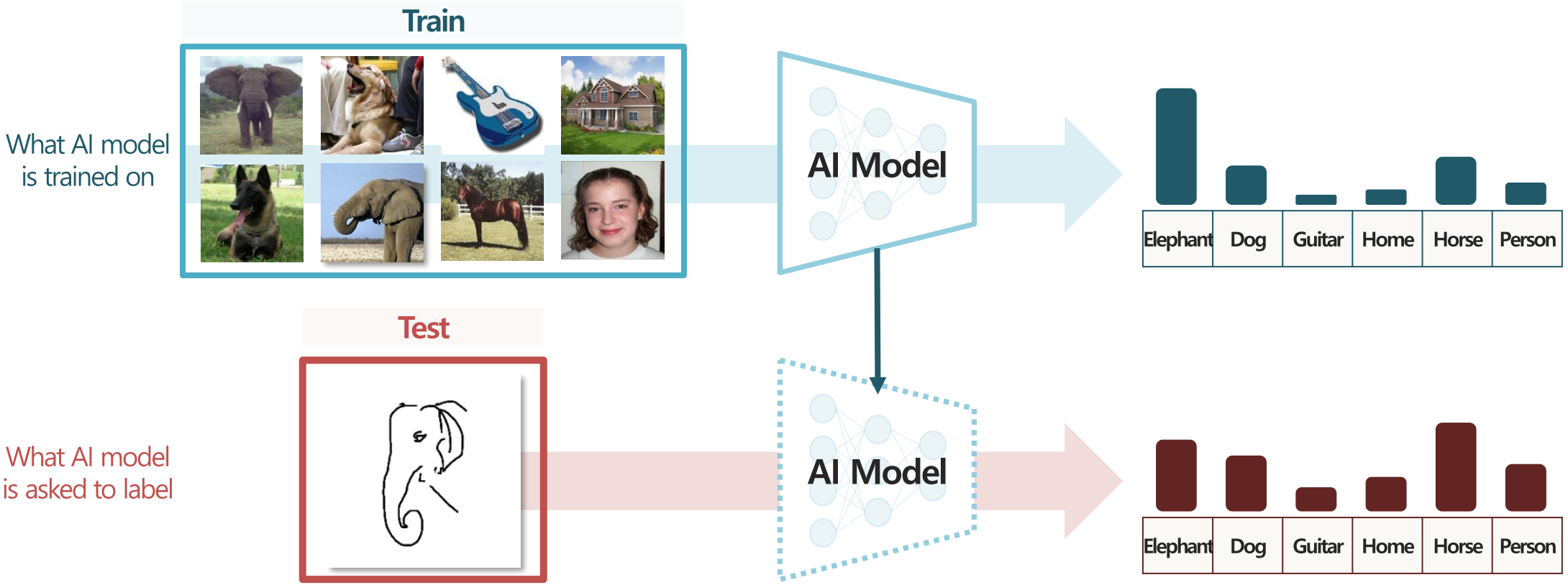


Introduction

Background on Domain Adaptation

We can train an AI model on the training data and directly apply it to the test data!

So, is AI solved?

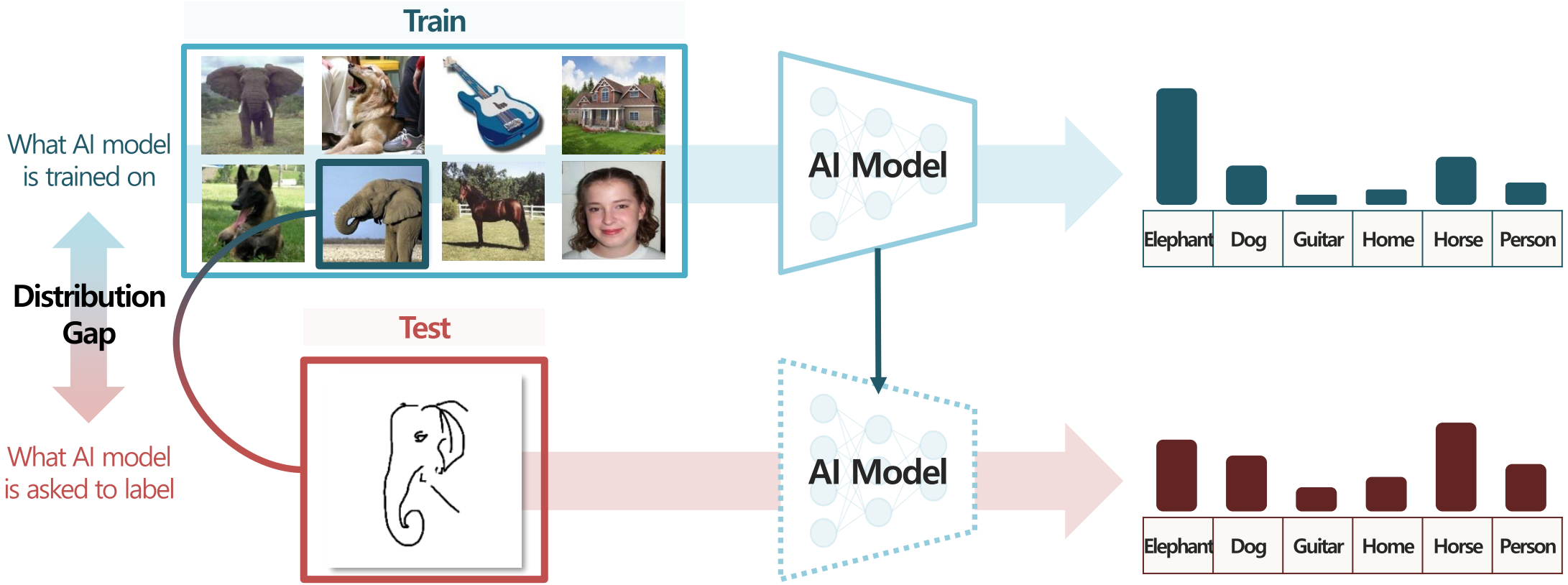


Introduction

Background on Domain Adaptation

We can train an AI model on the training data and directly apply it to the test data!

So, is AI solved?

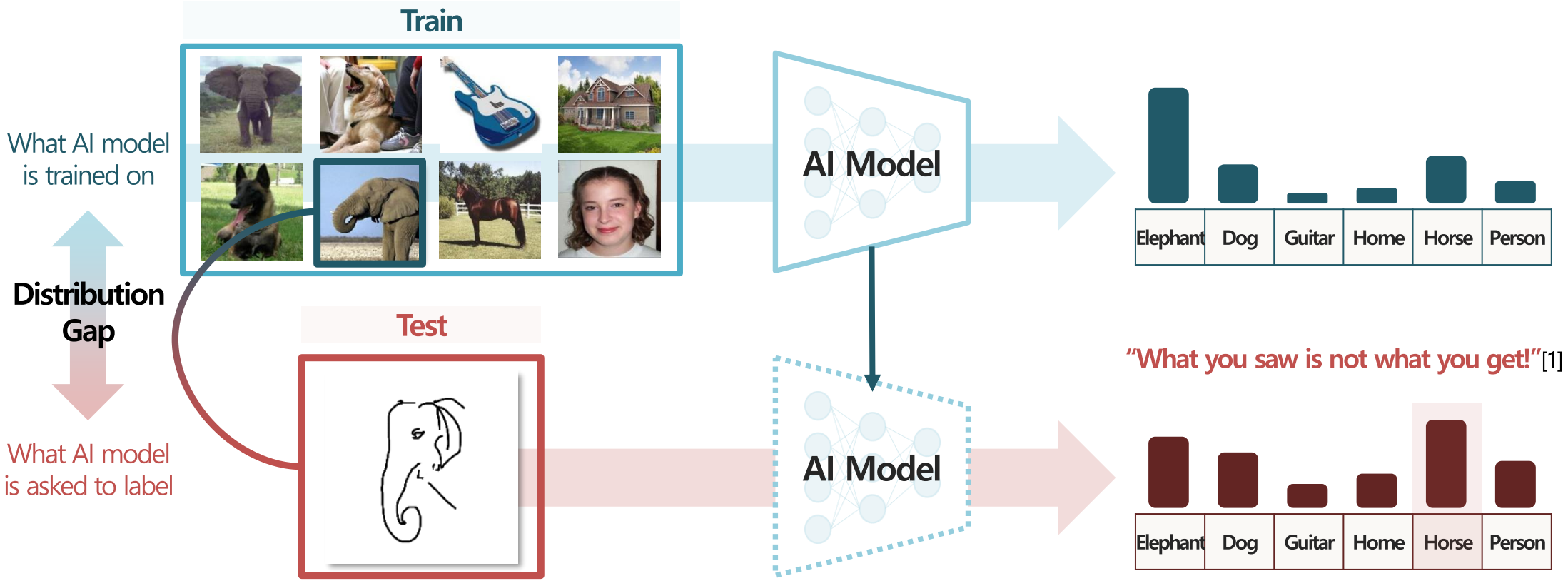


Introduction

Background on Domain Adaptation

We can train an AI model on the training data and directly apply it to the test data!

So, is AI solved?



"What you saw is not what you get!"[1]

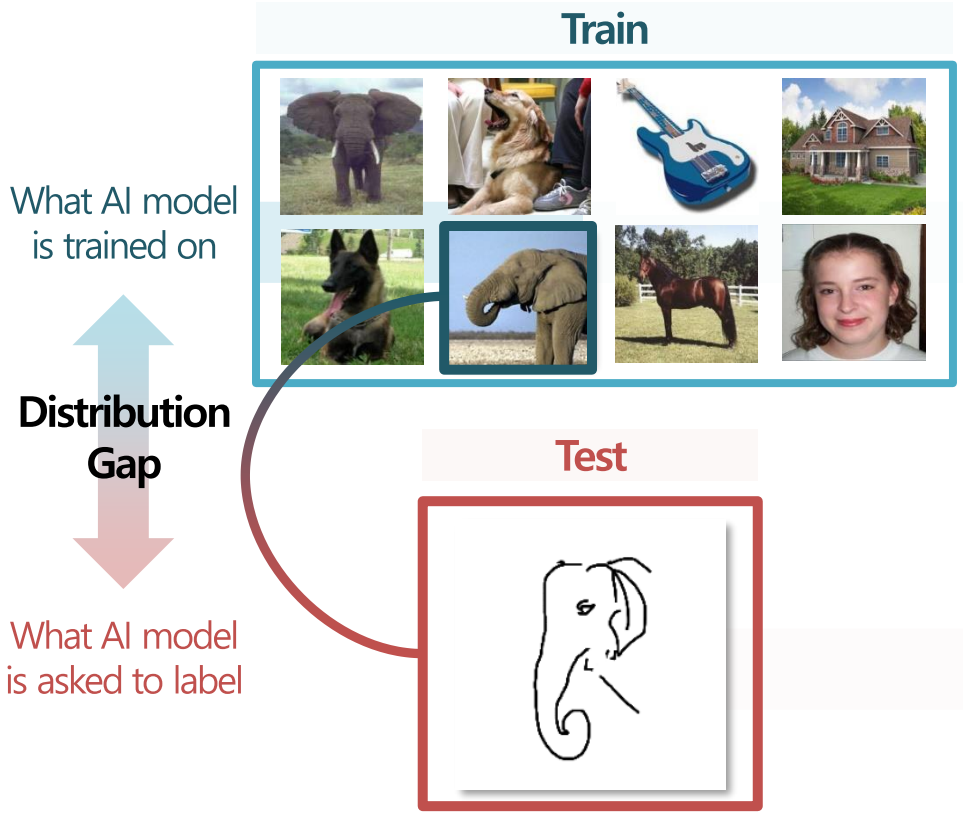
[1] B. Kulis, K. Saenko and T. Darrell, "What you saw is not what you get: Domain adaptation using asymmetric kernel transforms," CVPR 2011, Colorado Springs, CO, USA, 2011, pp. 1785-1792, doi: 10.1109/CVPR.2011.5995702.

Introduction

Background on Domain Adaptation

We can train an AI model on the training data and directly apply it to the test data!

So, is AI solved?



Violate i.i.d assumption, i.e.,
Training and Testing data are NOT
identically and independently distributed.

같은 분포로부터 수집되는 데이터

Domain shift!

Introduction

Background on Domain Adaptation

We can train an AI model on the training data and directly apply it to the test data!

So, is AI solved?

Domain

A dataset $\mathcal{D}$ consisting of input data x and output label y, drawn from a specific joint probability distribution $P(x, y)$.

Source Domain $\mathcal{D}_S = \{X_S, Y_S\} = \{(x_S^i, y_S^i)\}_{i=1}^{N_S} \sim P_S(x, y)$,

Target Domain $\mathcal{D}_T = \{X_T, Y_T\} = \{(x_T^j, y_T^j)\}_{j=1}^{N_T} \sim P_T(x, y)$,

where N_S and N_T are the number of source and target samples, respectively.

Covariate Shift, where $P_S(y|x) = P_T(y|x)$ for all x, but $P_S(x) \neq P_T(x)$;

Label Shift, where $P_S(x|y) = P_T(x|y)$ for all y, but $P_S(y) \neq P_T(y)$.

모델은 $X \rightarrow Y$ 관계를 학습하는데, **이 관계가 도메인과 무관하게 변하지 않음**

→ **x의 분포 차이만 해결**하면 source로 모델을 학습시켜도 **target에서 잘 동작 가능**

Violate i.i.d assumption, i.e.,
Training and Testing data are NOT
identically and independently distributed.

같은 분포로부터 수집되는 데이터

Domain shift!

Introduction

Background on Domain Adaptation

Finetuning a pretrained source model for the target domain!

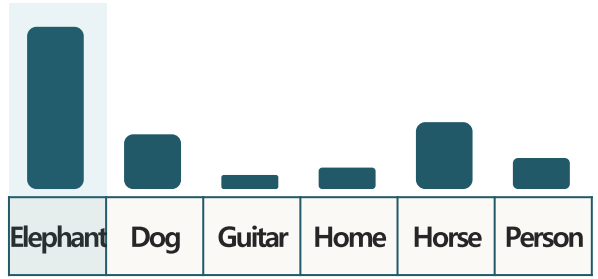
Labeled Source Domain으로 사전학습!

Domain Shift

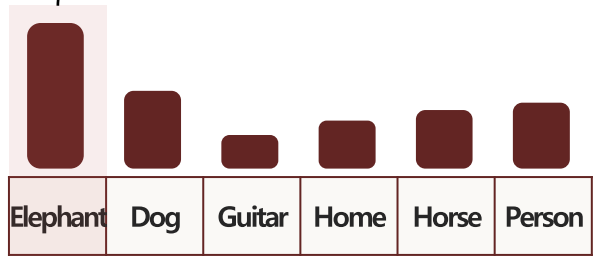
Target Domain으로 미세조정!

AI Model

AI Model



새로운 분포의 Target Domain에 적합하게 모델 파라미터 업데이트 가능!



Introduction

Background on Domain Adaptation

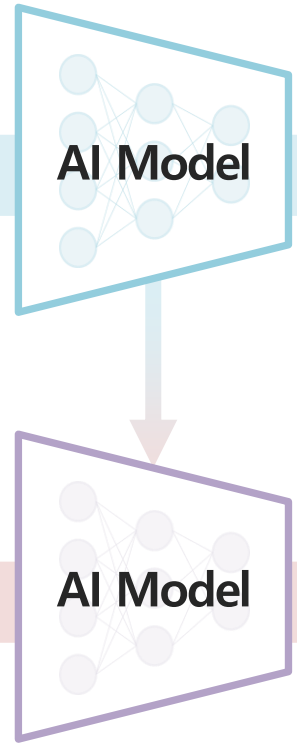
Finetuning a pretrained source model for the target domain!

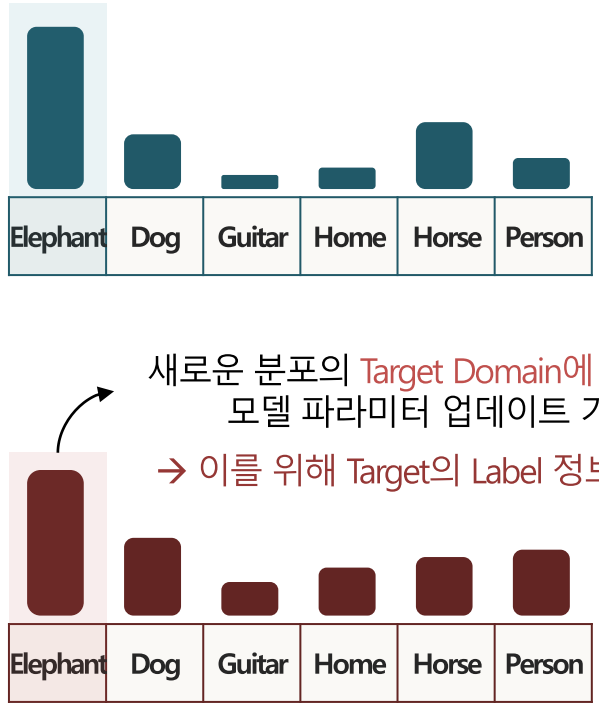
→ Requiring Large-scale labeled data for every new target domain

Labeled Source Domain으로 사전학습!

Domain Shift

Labeled Target Domain으로 미세조정!





새로운 분포의 Target Domain에 적합하게 모델 파라미터 업데이트 가능!

→ 이를 위해 Target의 Label 정보가 필요

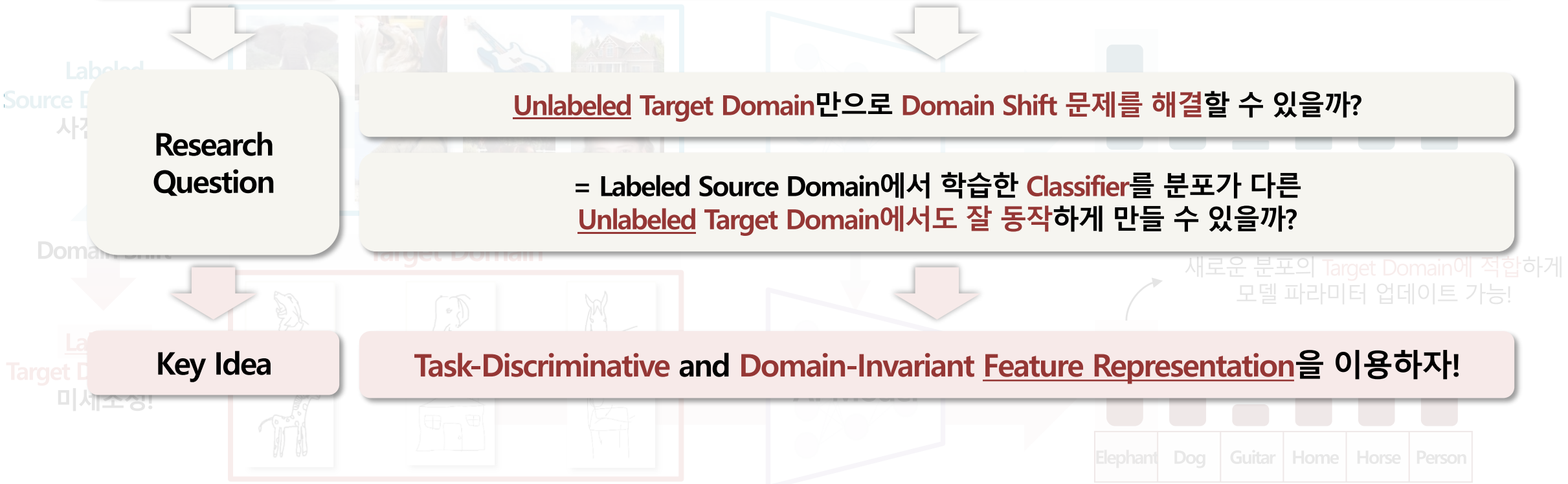
Introduction

Background on Domain Adaptation

Target에 Label 정보가 없을 때 Domain 차이 완화

Unsupervised Domain Adaptation!

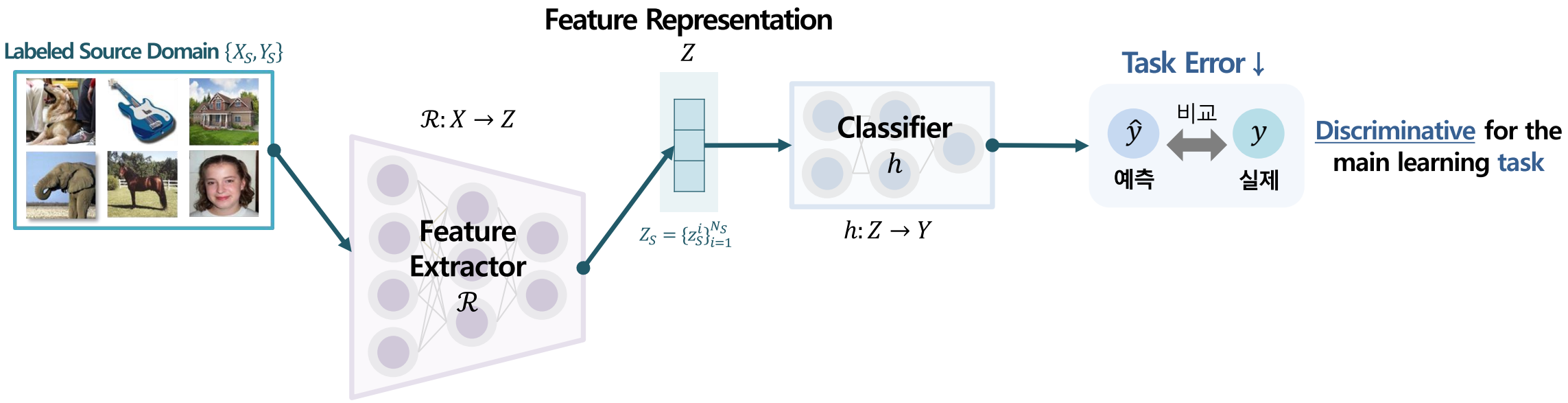
Limitation	Domain Shift 문제를 해결을 위해서 Labeled Target Domain이 필요! (비용↑)
Research Question	Unlabeled Target Domain만으로 Domain Shift 문제를 해결할 수 있을까? = Labeled Source Domain에서 학습한 Classifier를 분포가 다른 Unlabeled Target Domain에서도 잘 동작하게 만들 수 있을까?
Key Idea	Task-Discriminative and Domain-Invariant Feature Representation을 이용하자!



Preliminary

Cross-domain Generalization

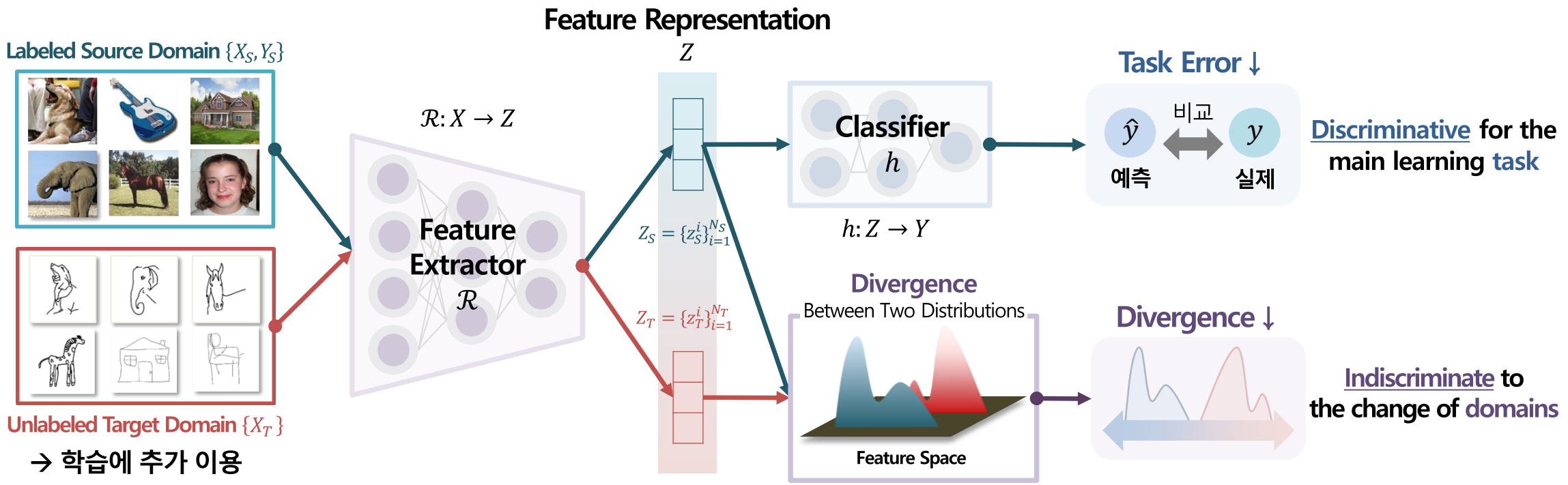
목적	Source Domain에서 학습한 Classifier $h: Z \rightarrow Y$가 Target Domain에서도 잘 동작하게 만들자
방법	Task-Discriminative and Domain-Invariant Feature Representation을 이용하자!



Preliminary

Cross-domain Generalization

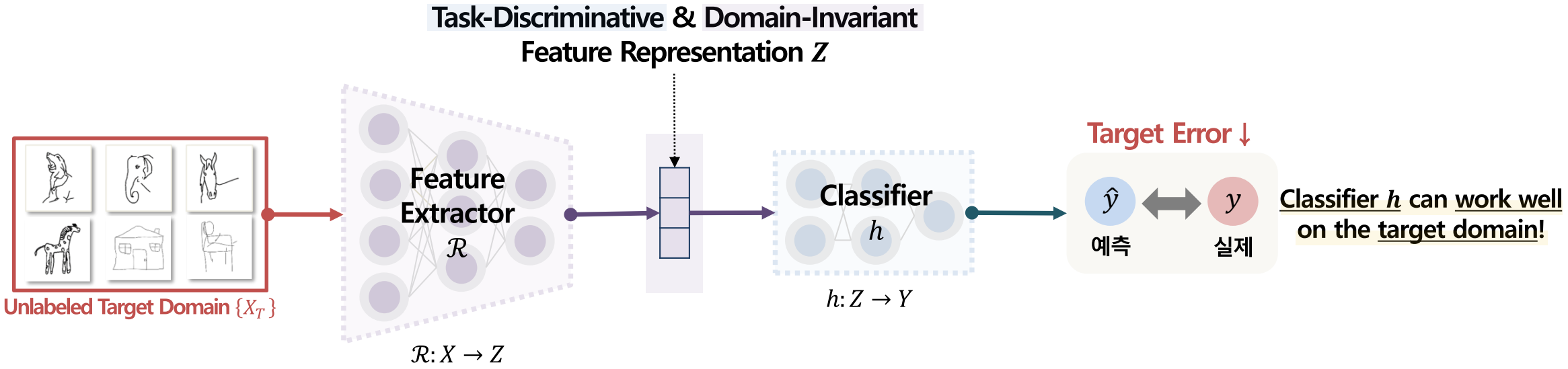
목적	Source Domain에서 학습한 Classifier $h: Z \rightarrow Y$가 Target Domain에서도 잘 동작하게 만들자
방법	Task-Discriminative and Domain-Invariant Feature Representation을 이용하자!



Preliminary

Cross-domain Generalization

목적	Source Domain에서 학습한 Classifier $h: Z \rightarrow Y$가 Target Domain에서도 잘 동작하게 만들자
방법	Task-Discriminative and Domain-Invariant Feature Representation을 이용하자!



Preliminary

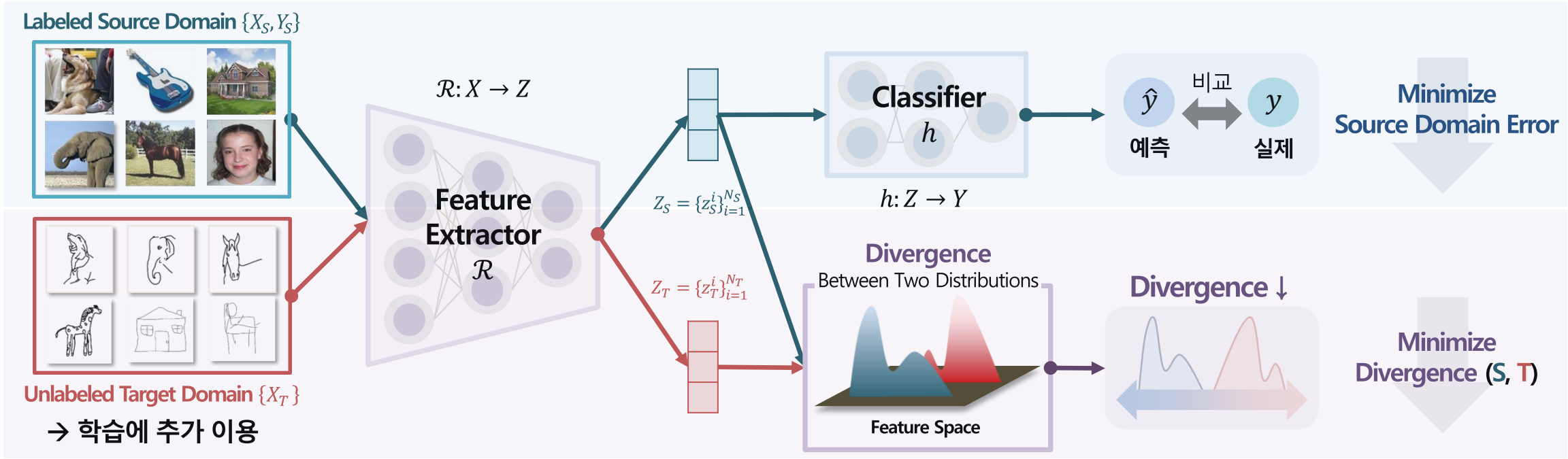
Cross-domain Generalization

목적: Source Domain에서 학습한 Classifier $h: Z \rightarrow Y$가 Target Domain에서도 잘 동작하게 만들자

방법: Task-Discriminative and Domain-Invariant Feature Representation을 이용하자!

Cross-domain Generalization

Target Domain Error ≤ Source Domain Error + Divergence(Source, Target)



Preliminary

Cross-domain Generalization

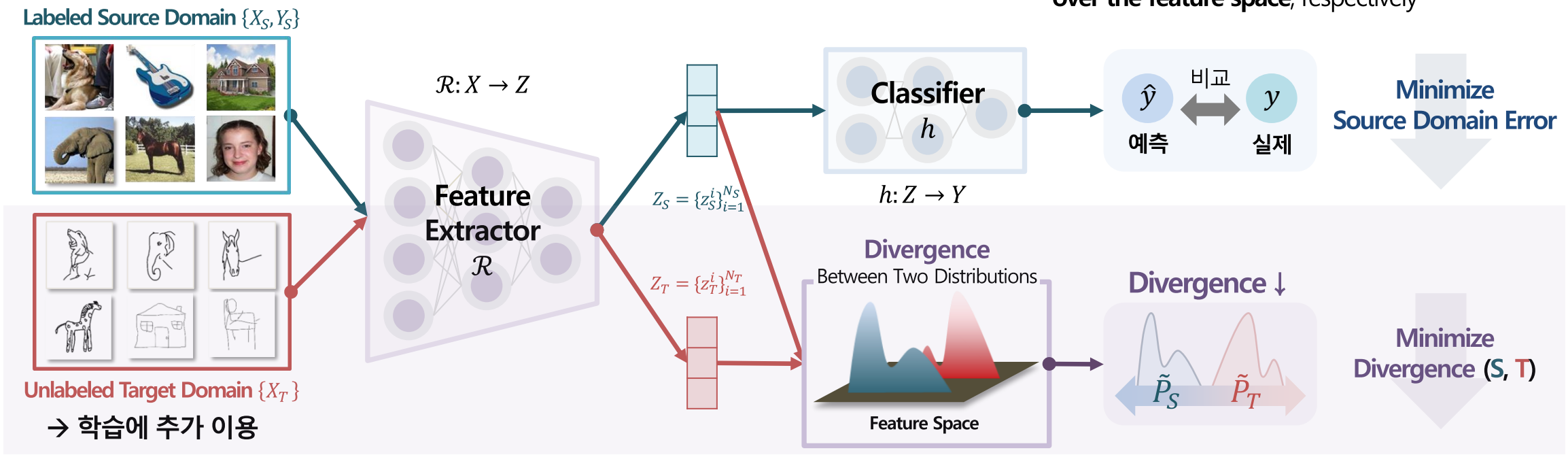
목적: Source Domain에서 학습한 Classifier $h: Z \rightarrow Y$가 Target Domain에서도 잘 동작하게 만들자

방법: Task-Discriminative and Domain-Invariant Feature Representation을 이용하자!

Cross-domain Generalization

Target Domain Error ≤ Source Domain Error + Divergence($\tilde{P}_S$, $\tilde{P}_T$)

$\tilde{P}_S$ and $\tilde{P}_T$ denotes the induced source and target distribution **over the feature space**, respectively



Preliminary

Cross-domain Generalization

목적 Source Domain에서 학습한 Classifier $h: Z \rightarrow Y$가 Target Domain에서도 잘 동작하게 만들자

방법 Task-Discriminative and Domain-Invariant Feature Representation을 이용하자!

Cross-domain Generalization

Target Domain Error ≤ Source Domain Error + Divergence($\tilde{P}_S$, $\tilde{P}_T$)

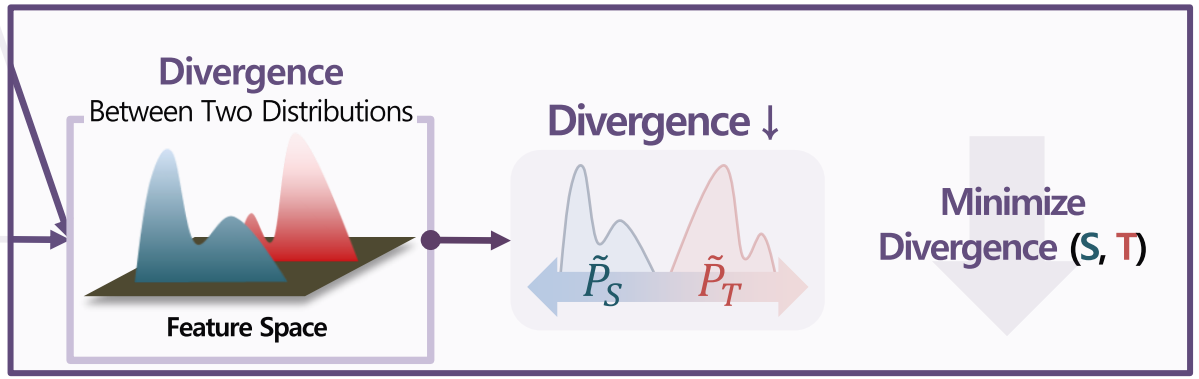
Labeled Source Domain $\{X_S, Y_S\}$

Unlabeled Target Domain $\{X_T\}$

→ 학습에 추가 이용

Feature Extractor ϕ

어떤 Divergence 지표를 사용하면 좋을까?
→ Lead to different methods



Preliminary

Cross-domain Generalization

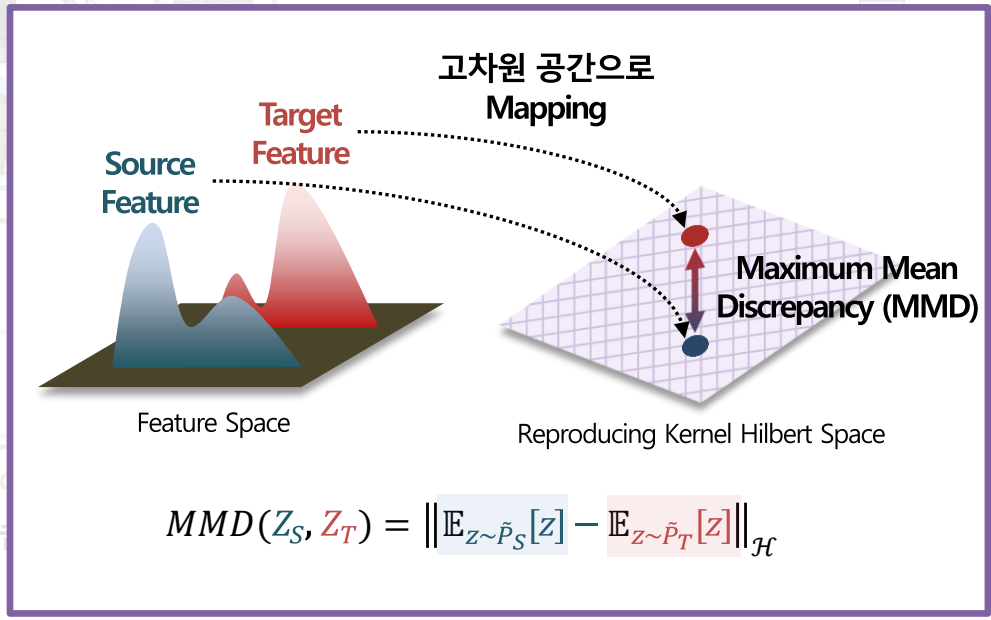
목적: Source Domain에서 학습한 Classifier $h: Z \rightarrow Y$가 Target Domain에서도 잘 동작하게 만들자

방법: Task-Discriminative and Domain-Invariant Feature Representation을 이용하자!

Cross-domain Generalization

Target Domain Error ≤ Source Domain Error + Divergence($\tilde{P}_S$, $\tilde{P}_T$)

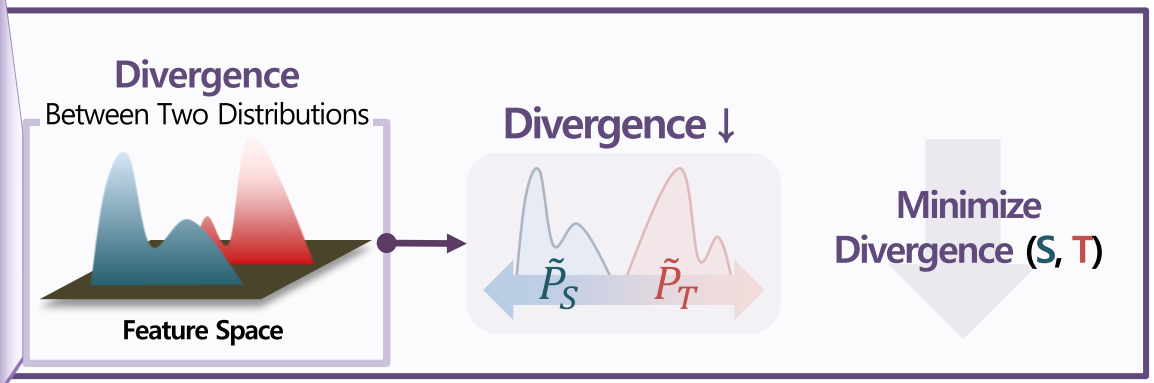
Moment Matching; Maximum Mean Discrepancy[2]



$$MMD(Z_S, Z_T) = \left\| \mathbb{E}_{Z \sim \tilde{P}_S}[Z] - \mathbb{E}_{Z \sim \tilde{P}_T}[Z] \right\|_{\mathcal{H}}$$

Moment Matching

Learn domain-invariant feature representations by
explicitly matching a moment-based distribution discrepancy



$$Total\ Loss = Source\ Error + \lambda MMD$$

[2] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. arXiv:1412.3474, 2014.

Preliminary

Cross-domain Generalization

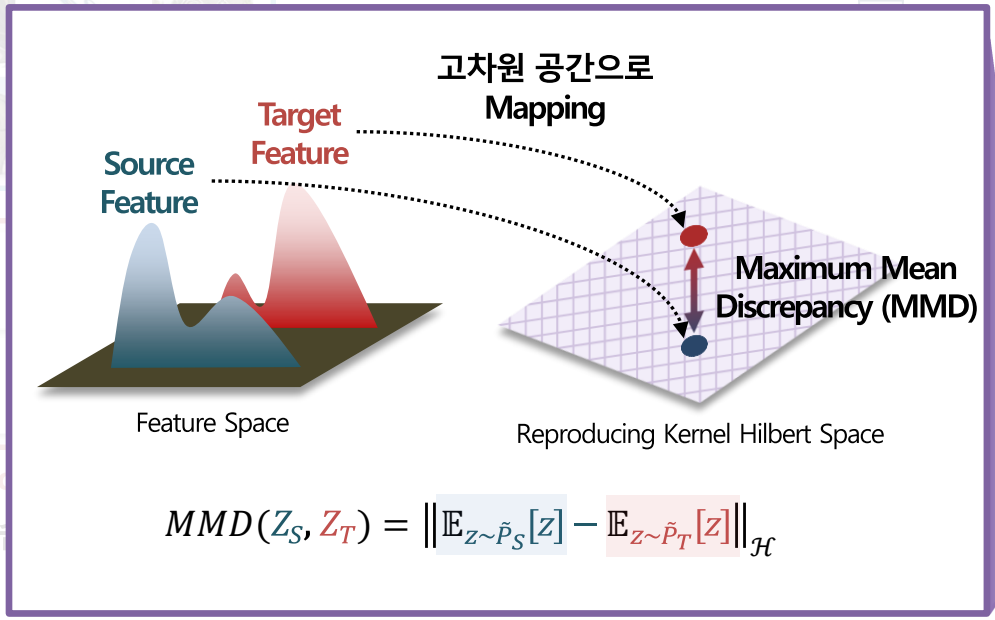
목적 | Source Domain에서 학습한 Classifier $h: Z \rightarrow Y$가 Target Domain에서도 잘 동작하게 만들자

방법 | Task-Discriminative and Domain-Invariant Feature Representation을 이용하자!

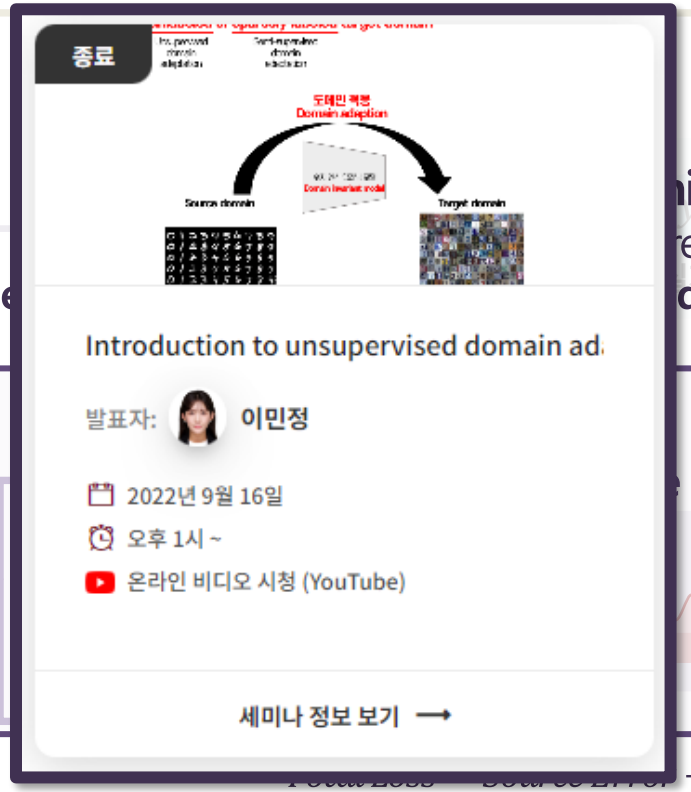
Cross-domain Generalization

Target Domain Error ≤ Source Domain Error + Divergence($\tilde{P}_S$, $\tilde{P}_T$)

Moment Matching; Maximum Mean Discrepancy[2]



$$MMD(Z_S, Z_T) = \left\| \mathbb{E}_{Z \sim \tilde{P}_S}[Z] - \mathbb{E}_{Z \sim \tilde{P}_T}[Z] \right\|_{\mathcal{H}}$$



ing
epresentations by
distribution discrepancy

Minimize
Divergence (S, T)

$+ \lambda MMD$

[2] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. arXiv:1412.3474, 2014.

Preliminary

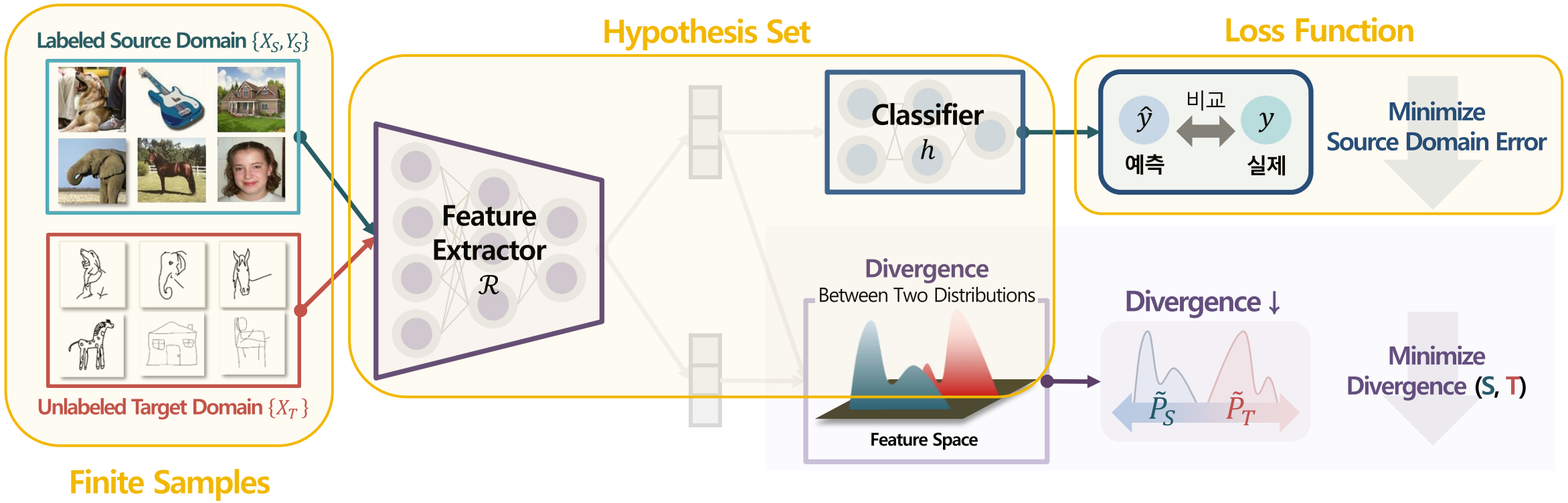
Cross-domain Generalization

목적: Source Domain에서 학습한 Classifier $h: Z \rightarrow Y$가 Target Domain에서도 잘 동작하게 만들자

방법: Task-Discriminative and Domain-Invariant Feature Representation을 이용하자!

Cross-domain Generalization

Target Domain Error ≤ Source Domain Error + Divergence($\tilde{P}_S$, $\tilde{P}_T$)



Methods

Analysis of Representations for Domain Adaptation, NIPS, 2006

❖ **Analysis of Representations for Domain Adaptation** (Ben-David et al, 2006)[3]

- Domain Adaptation을 위한 Generalization Bound를 이론적으로 정식화
- 유한한 데이터 표본만으로도 계산 가능한 Divergence 지표 제안 ($\mathcal{H}$-Divergence, $d_{\mathcal{A}}$-Distance)

NIPS, 24년 3월 기준 2419회 인용

Analysis of Representations for Domain Adaptation

Shai Ben-David
School of Computer Science
University of Waterloo
shai@cs.uwaterloo.ca

John Blitzer, Koby Crammer, and Fernando Pereira
Department of Computer and Information Science
University of Pennsylvania
{blitzer, crammer, pereira}@cis.upenn.edu

Abstract

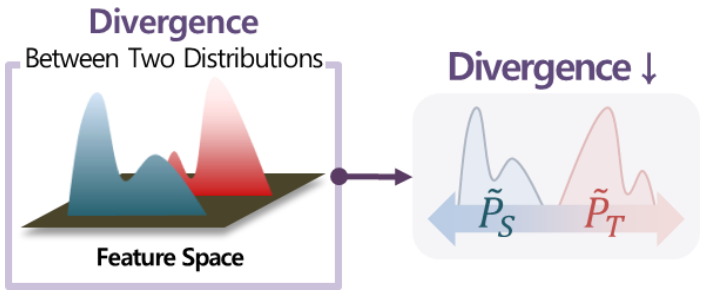
Discriminative learning methods for classification perform well when training and test data are drawn from the same distribution. In many situations, though, we have labeled training data for a *source* domain, and we wish to learn a classifier which performs well on a *target* domain with a different distribution. Under what conditions can we adapt a classifier trained on the source domain for use in the target domain? Intuitively, a good feature representation is a crucial factor in the success of domain adaptation. We formalize this intuition theoretically with a generalization bound for domain adaption. Our theory illustrates the tradeoffs inherent in designing a representation for domain adaptation and gives a new justification for a recently proposed model. It also points toward a promising new model for domain adaptation: one which explicitly minimizes the difference between the source and target domains, while at the same time maximizing the margin of the training set.

Cross-domain Generalization

Target Domain Error $\leq$ Source Domain Error + Divergence($\tilde{P}_S$, $\tilde{P}_T$)

Empirical $\mathcal{H}$-Divergence[3]

$$\hat{d}_{\mathcal{H}}(\tilde{Z}_S, \tilde{Z}_T) = 2\left(1 - \min_{h' \in \mathcal{H}} \left[\mathbb{E}_{z \sim \tilde{P}_S} I[h'(z) = 1] + \mathbb{E}_{z \sim \tilde{P}_T} I[h'(z) = 0]\right]\right)$$



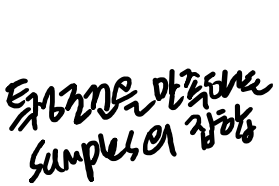
[3] Ben-David, S., Blitzer, J., Crammer, K., & Pereira, F. (2006). Analysis of representations for domain adaptation. Advances in neural information processing systems, 19.

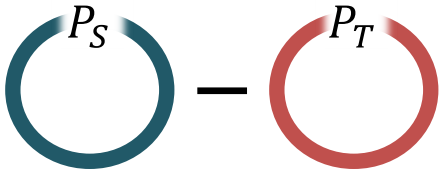
Methods

목적 | Source Domain에서 학습한 Classifier $h: Z \rightarrow Y$가 Target Domain에서도 잘 동작하게 만들자

방법 | Task-Discriminative and Domain-Invariant Feature Representation을 이용하자!

Analysis of Representations for Domain Adaptation, NIPS, 2006





사건 B가 관측될 확률

$\Pr_{P_S}[B] - \Pr_{P_T}[B]$

분포 P_S와 P_T 간 사건 B에 대한 관측 확률 차이를 모든 가능한 사건에 대해 조사 → 이 차이의 최대 값이 L_1-Divergence!

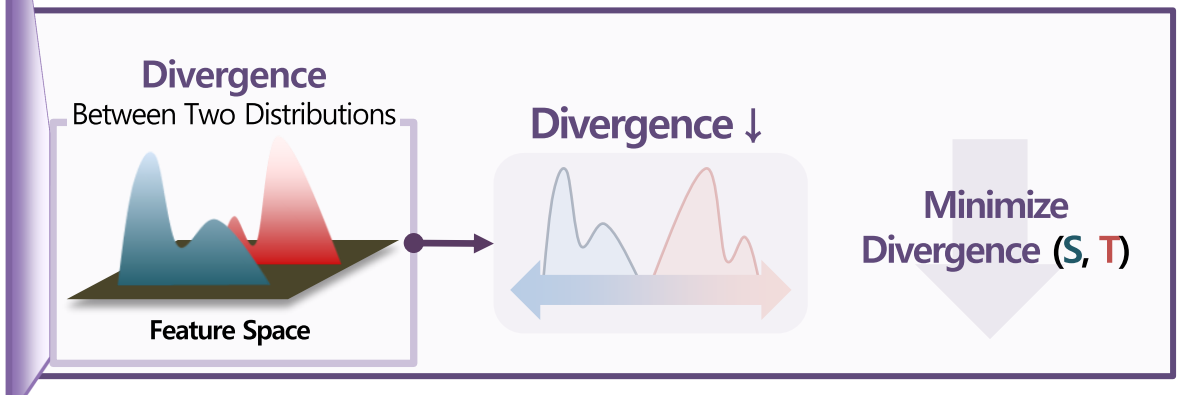
Definition of L_1-Divergence[3]

Given two domain distributions P_S and P_T over X, and let $\mathcal{B}$ be the set of measurable subsets under P_S and P_T, the L_1-divergence between P_S and P_T is

$$d_1(P_S, P_T) = 2 \sup_{B \in \mathcal{B}} \left| \Pr_{P_S}[B] - \Pr_{P_T}[B] \right|.$$

① L_1-Divergence

The absolute difference between two probability distributions P_S and P_T.



[3] Ben-David, S., Blitzer, J., Crammer, K., & Pereira, F. (2006). Analysis of representations for domain adaptation. Advances in neural information processing systems, 19.

Methods

목적 Source Domain에서 학습한 Classifier $h: Z \rightarrow Y$가 Target Domain에서도 잘 동작하게 만들자

방법 Task-Discriminative and Domain-Invariant Feature Representation을 이용하자!

Analysis of Representations for Domain Adaptation, NIPS, 2006

Cross-domain Generalization

Target Domain Error ≤ Source Domain Error + Divergence(Source, Target)

- $\mathcal{R}: X \rightarrow Z$ $\quad \mathcal{R}(x) = z$
- $f: X \rightarrow \{0,1\}$ True labeling function
- $\tilde{f}: Z \rightarrow \{0,1\}$ Induced image of f under $\mathcal{R}$ ← 모사
- $h: Z \rightarrow \{0,1\}$ Hypothesis
- $\epsilon(h, \tilde{f}) = \mathbb{E}_{z \sim \tilde{P}}[|h(z) - \tilde{f}(z)|] = \epsilon(h)$

L_1-Divergence[3]

$$d_1(P_S, P_T) = 2 \sup_{B \in \mathcal{B}} |\Pr_{P_S}[B] - \Pr_{P_T}[B]| .$$

Theorem 1. (Ben-David et al., 2010) For a hypothesis h,

$$\epsilon_T(h) \leq \epsilon_S(h) + d_1(P_S, P_T)$$

두 분포에서 Labeling function의 차이

$$+ \min\{\mathbb{E}_{x \sim P_S}[|f_S(x) - f_T(x)|], \mathbb{E}_{x \sim P_T}[|f_S(x) - f_T(x)|]\} .$$

This dissimilarity will need to be small for adaptation to be possible!

[3] Ben-David, S., Blitzer, J., Crammer, K., & Pereira, F. (2006). Analysis of representations for domain adaptation. Advances in neural information processing systems, 19.
[4] Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., & Vaughan, J. W. (2010). A theory of learning from different domains. Machine learning, 79, 151-175..

Methods

목적 Source Domain에서 학습한 Classifier $h: Z \rightarrow Y$가 Target Domain에서도 잘 동작하게 만들자

방법 Task-Discriminative and Domain-Invariant Feature Representation을 이용하자!

Analysis of Representations for Domain Adaptation, NIPS, 2006

Cross-domain Generalization

Target Domain Error ≤ Source Domain Error + Divergence(Source, Target)

Limitations

1) 임의의 분포로부터 추출된 **유한 개의 샘플로는 정확한 추정 불가**
   → DA을 위한 Representation 조건을 연구하기 위한 지표로 적합하지 않음
2) 너무 엄격한 지표이기에 **실제 minimum risk보다 큰 값을 상계로 제시**

L_1-Divergence[3]

$$d_1(P_S, P_T) = 2 \sup_{B \in \mathcal{B}} \left| \Pr_{P_S}[B] - \Pr_{P_T}[B] \right|.$$

분포 P_S 와 P_T 간 **가능한 모든 사건 B에 대한** 관측 확률 차이의 최대값

Theorem 1. (Ben-David et al., 2010) For a hypothesis h,

$$\epsilon_T(h) \leq \epsilon_S(h) + d_1(P_S, P_T)$$

$$+ \min\left\{ \mathbb{E}_{x \sim P_S}[|f_S(x) - f_T(x)|], \mathbb{E}_{x \sim P_T}[|f_S(x) - f_T(x)|] \right\}.$$

[3] Ben-David, S., Blitzer, J., Crammer, K., & Pereira, F. (2006). Analysis of representations for domain adaptation. Advances in neural information processing systems, 19.
[4] Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., & Vaughan, J. W. (2010). A theory of learning from different domains. Machine learning, 79, 151-175.

Methods

목적 | Source Domain에서 학습한 Classifier $h: Z \rightarrow Y$가 Target Domain에서도 잘 동작하게 만들자

방법 | Task-Discriminative and Domain-Invariant Feature Representation을 이용하자!

Analysis of Representations for Domain Adaptation, NIPS, 2006

Recap : L_1-Divergence

$$d_1(P_S, P_T) = 2 \sup_{B \in \mathcal{B}} \left| \Pr_{P_S}[B] - \Pr_{P_T}[B] \right|.$$

Definition of $\mathcal{H}$-Divergence[3]

Definition 1, Ben-David et al., 2006, 2010; Kifer et al. 2004

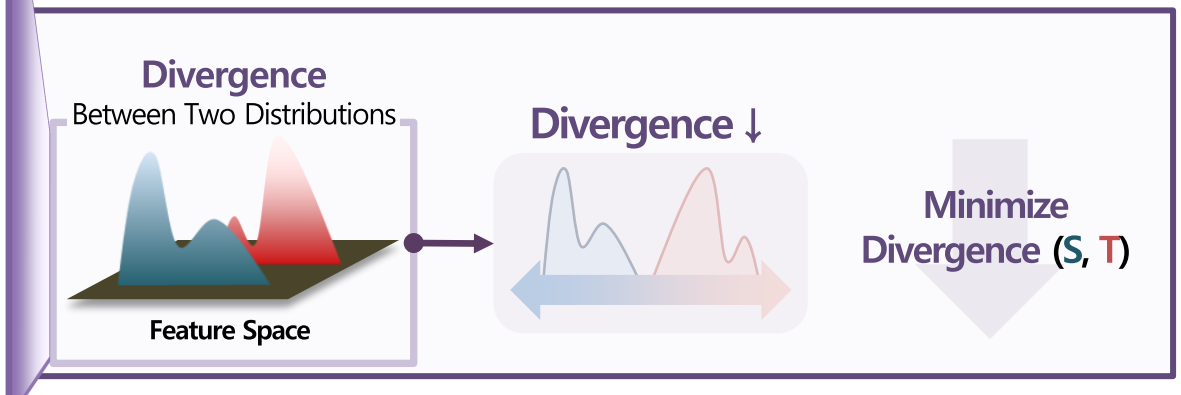
Let $\mathcal{R}$ be a fixed representation function from X to Z and $\mathcal{H} \subseteq \{h' : Z \rightarrow \{0, 1\}\}$ be a hypothesis class, given two domain distributions P_S and P_T over X, the $\mathcal{H}$-divergence between P_S and P_T is defined as

$$d_{\mathcal{H}}(P_S, P_T) = 2 \sup_{h' \in \mathcal{H}} \left| \Pr_{P_S}[h'(\mathcal{R}(x)) = 1] - \Pr_{P_T}[h'(\mathcal{R}(x)) = 1] \right|.$$

② $\mathcal{H}$-Divergence

Can **account for the hypothesis set**!

Can be **approximated empirically** given samples!



[3] Ben-David, S., Blitzer, J., Crammer, K., & Pereira, F. (2006). Analysis of representations for domain adaptation. Advances in neural information processing systems, 19.

Methods

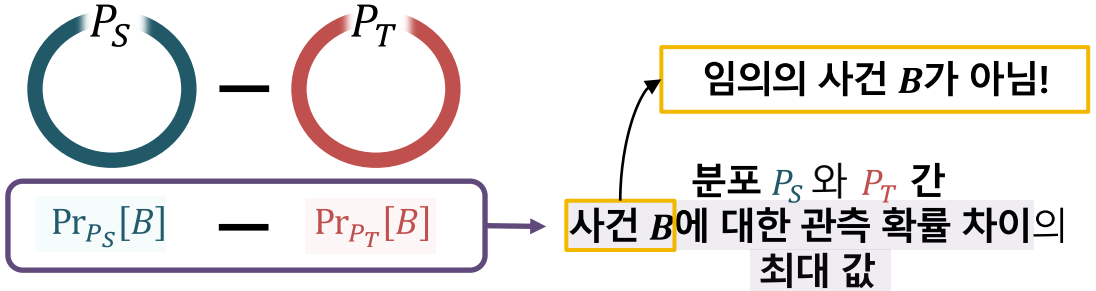
목적 | Source Domain에서 학습한 Classifier $h: Z \rightarrow Y$가 Target Domain에서도 잘 동작하게 만들자

방법 | Task-Discriminative and Domain-Invariant Feature Representation을 이용하자!

Analysis of Representations for Domain Adaptation, NIPS, 2006

Recap : L_1-Divergence

$$d_1(P_S, P_T) = 2 \sup_{B \in \mathcal{B}} \left| \Pr_{P_S}[B] - \Pr_{P_T}[B] \right|.$$



Definition of $\mathcal{H}$-Divergence[3]

Definition 1, Ben-David et al., 2006, 2010; Kifer et al. 2004

Let $\mathcal{R}$ be a fixed representation function from X to Z and $\mathcal{H} \subseteq \{h' : Z \rightarrow \{0, 1\}\}$ be a hypothesis class, given two domain distributions P_S and P_T over X, the $\mathcal{H}$-divergence between P_S and P_T is defined as

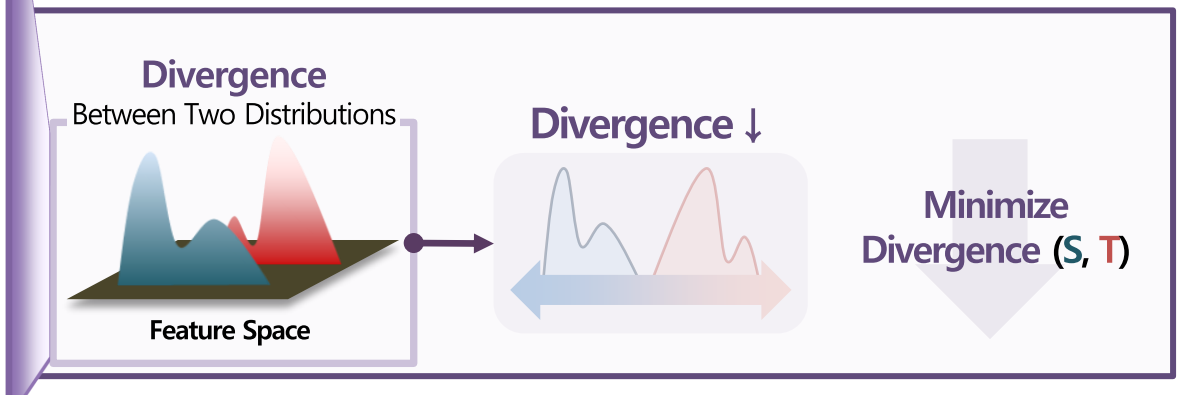
$$d_{\mathcal{H}}(P_S, P_T) = 2 \sup_{h' \in \mathcal{H}} \left| \Pr_{P_S}[h'(\mathcal{R}(x)) = 1] - \Pr_{P_T}[h'(\mathcal{R}(x)) = 1] \right|.$$

- 특정한 모델 h'에 대한 사건
- h'에 의해 z가 1로 분류될 사건으로 특정

② $\mathcal{H}$-Divergence

Can **account for the hypothesis set**!

Can be **approximated empirically** given samples!



[3] Ben-David, S., Blitzer, J., Crammer, K., & Pereira, F. (2006). Analysis of representations for domain adaptation. Advances in neural information processing systems, 19.

Methods

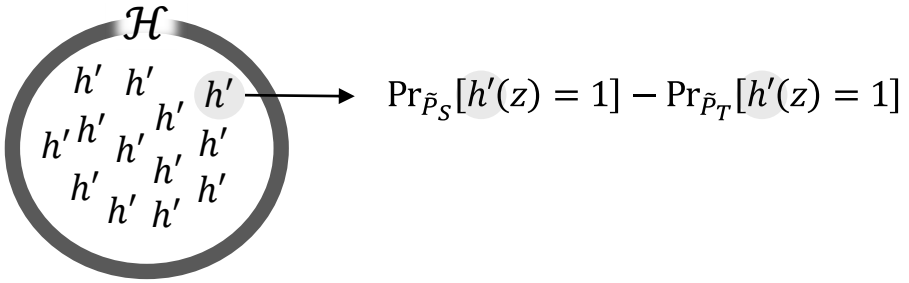
목적 Source Domain에서 학습한 Classifier $h: Z \rightarrow Y$가 Target Domain에서도 잘 동작하게 만들자

방법 Task-Discriminative and Domain-Invariant Feature Representation을 이용하자!

Analysis of Representations for Domain Adaptation, NIPS, 2006

Recap : L_1-Divergence

$$d_1(P_S, P_T) = 2 \sup_{B \in \mathcal{B}} \left| \Pr_{P_S}[B] - \Pr_{P_T}[B] \right|.$$



Definition of $\mathcal{H}$-Divergence[3]

Definition 1, Ben-David et al., 2006, 2010; Kifer et al. 2004

Let $\mathcal{R}$ be a fixed representation function from X to Z and $\mathcal{H} \subseteq \{h': Z \rightarrow \{0,1\}\}$ be a hypothesis class, given two domain distributions P_S and P_T over X, the $\mathcal{H}$-divergence between P_S and P_T is defined as

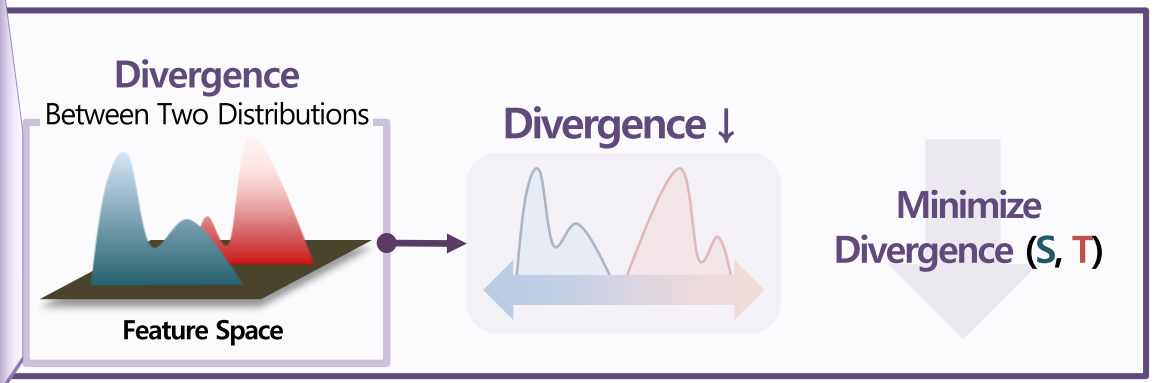
$$d_{\mathcal{H}}(P_S, P_T) = 2 \sup_{h' \in \mathcal{H}} \left| \Pr_{P_S}[h'(\mathcal{R}(x)) = 1] - \Pr_{P_T}[h'(\mathcal{R}(x)) = 1] \right|.$$

- 특정한 모델 h'에 대한 사건
- h'에 의해 z가 1로 분류될 사건으로 특정

② $\mathcal{H}$-Divergence

Can **account for the hypothesis set**!

Can be **approximated empirically** given samples!



[3] Ben-David, S., Blitzer, J., Crammer, K., & Pereira, F. (2006). Analysis of representations for domain adaptation. Advances in neural information processing systems, 19.

Methods

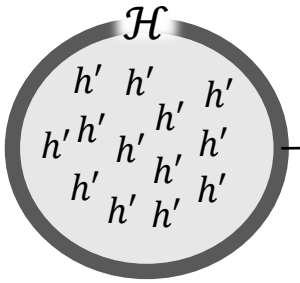
목적 | Source Domain에서 학습한 Classifier $h: Z \to Y$가 Target Domain에서도 잘 동작하게 만들자

방법 | Task-Discriminative and Domain-Invariant Feature Representation을 이용하자!

Analysis of Representations for Domain Adaptation, NIPS, 2006

Recap : L_1-Divergence

$$d_1(P_S, P_T) = 2 \sup_{B \in \mathcal{B}} \left| \Pr_{P_S}[B] - \Pr_{P_T}[B] \right|.$$



$$\sup_{h' \in \mathcal{H}} \left| \Pr_{\tilde{P}_S}[h'(z) = 1] - \Pr_{\tilde{P}_T}[h'(z) = 1] \right|$$

이 값이 0이라면?
→ 적어도 이 Hypothesis set에 대해서는 **domain에 상관 없이 h가 잘 동작**한다는 의미! (우리의 목적)

Definition of $\mathcal{H}$-Divergence[3]

Definition 1, Ben-David et al., 2006, 2010; Kifer et al. 2004

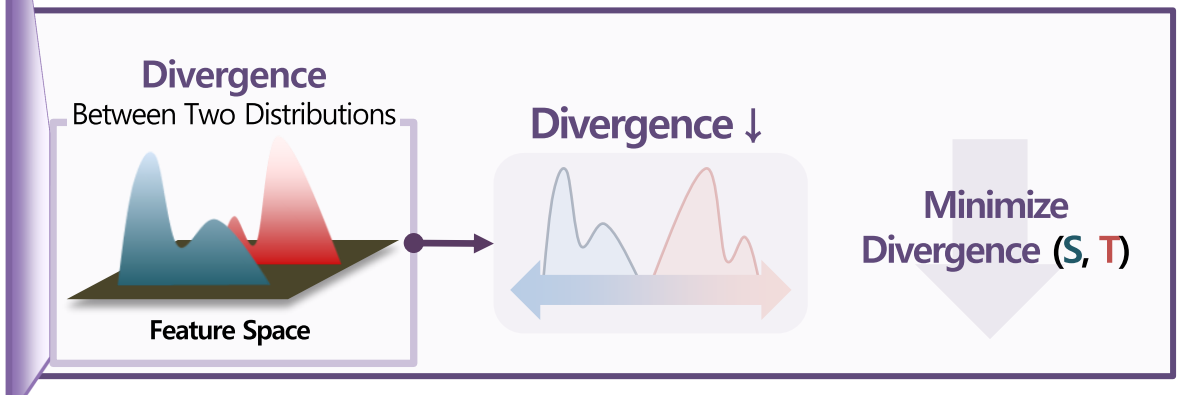
Let $\mathcal{R}$ be a fixed representation function from X to Z and $\mathcal{H} \subseteq \{h' : Z \to \{0, 1\}\}$ be a hypothesis class, given two domain distributions P_S and P_T over X, the $\mathcal{H}$-divergence between P_S and P_T is defined as

$$d_{\mathcal{H}}(P_S, P_T) = 2 \sup_{h' \in \mathcal{H}} \left| \Pr_{P_S}[h'(\mathcal{R}(x)) = 1] - \Pr_{P_T}[h'(\mathcal{R}(x)) = 1] \right|.$$

- 특정한 모델 h'에 대한 사건
- h'에 의해 z가 1로 분류될 사건으로 특정

② $\mathcal{H}$-Divergence

Can **account for the hypothesis set**!
Can be **approximated empirically** given samples!



[3] Ben-David, S., Blitzer, J., Crammer, K., & Pereira, F. (2006). Analysis of representations for domain adaptation. Advances in neural information processing systems, 19.

Methods

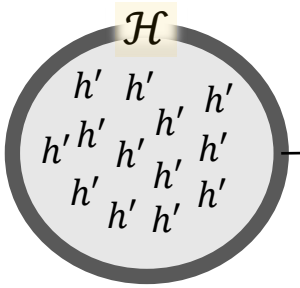
목적 Source Domain에서 학습한 Classifier $h: Z \rightarrow Y$가 Target Domain에서도 잘 동작하게 만들자

방법 Task-Discriminative and Domain-Invariant Feature Representation을 이용하자!

Analysis of Representations for Domain Adaptation, NIPS, 2006

Recap : L_1-Divergence

$$d_1(P_S, P_T) = 2 \sup_{B \in \mathcal{B}} \left| \Pr_{P_S}[B] - \Pr_{P_T}[B] \right|.$$



어떤 Hypothesis Set?

$$\sup_{h' \in \mathcal{H}} \left| \Pr_{\tilde{P}_S}[h'(z) = 1] - \Pr_{\tilde{P}_T}[h'(z) = 1] \right|$$

이 값이 0이라면?

→ 적어도 이 Hypothesis set에 대해서는 domain에 상관 없이 h가 잘 동작한다는 의미! (우리의 목적)

Definition of $\mathcal{H}$-Divergence[3]

Definition 1, Ben-David et al., 2006, 2010; Kifer et al. 2004

Let $\mathcal{R}$ be a fixed representation function from X to Z and $\mathcal{H} \subseteq \{h': Z \rightarrow \{0,1\}\}$ be a hypothesis class, given two domain distributions P_S and P_T over X, the $\mathcal{H}$-divergence between P_S and P_T is defined as

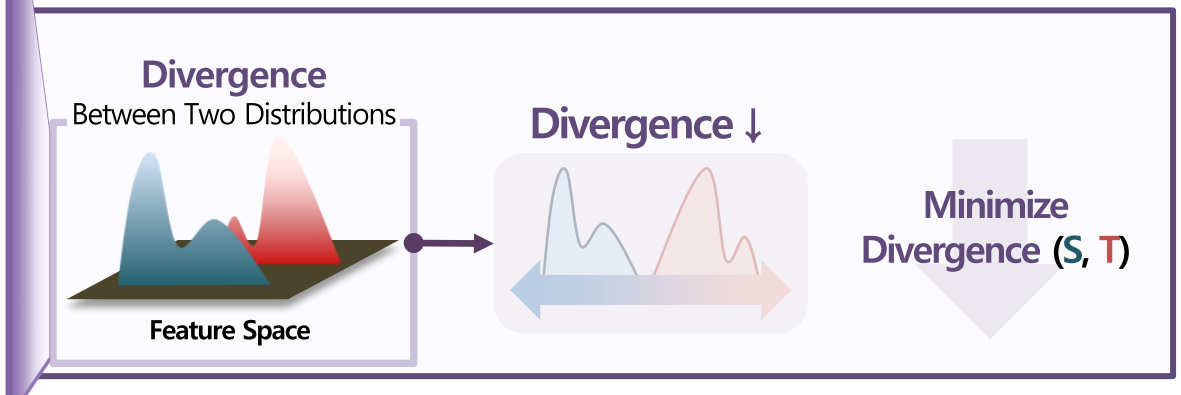
$$d_{\mathcal{H}}(P_S, P_T) = 2 \sup_{h' \in \mathcal{H}} \left| \Pr_{P_S}[h'(\mathcal{R}(x)) = 1] - \Pr_{P_T}[h'(\mathcal{R}(x)) = 1] \right|.$$

- 특정한 모델 h'에 대한 사건
- h'에 의해 z가 1로 분류될 사건으로 특정

② $\mathcal{H}$-Divergence

Can **account for the hypothesis set**!

Can be **approximated empirically** given samples!



[3] Ben-David, S., Blitzer, J., Crammer, K., & Pereira, F. (2006). Analysis of representations for domain adaptation. Advances in neural information processing systems, 19.

Methods

목적 | Source Domain에서 학습한 Classifier $h: Z \to Y$가 Target Domain에서도 잘 동작하게 만들자

방법 | Task-Discriminative and Domain-Invariant Feature Representation을 이용하자!

Analysis of Representations for Domain Adaptation, NIPS, 2006

Recap : L_1-Divergence

$$d_1(P_S, P_T) = 2 \sup_{B \in \mathcal{B}} \left| \Pr_{P_S}[B] - \Pr_{P_T}[B] \right|.$$

$\mathcal{H}$의 복잡도가 유한한 VC-dimension 인 경우!

→ $\sup_{h \in \mathcal{H}} \left| \Pr_{\tilde{P}_S}[h'(z) = 1] - \Pr_{\tilde{P}_T}[h'(z) = 1] \right|$

어떤 Hypothesis Set?

이 값이 0이라면?

→ 적어도 이 Hypothesis set에 대해서는 domain에 상관 없이 h가 잘 동작한다는 의미! (우리의 목적)

Definition of $\mathcal{H}$-Divergence[3]

Definition 1, Ben-David et al., 2006, 2010; Kifer et al. 2004

Let $\mathcal{R}$ be a fixed representation function from X to Z and $\mathcal{H} \subseteq \{h' : Z \to \{0, 1\}\}$ be a hypothesis class, given two domain distributions P_S and P_T over X, the $\mathcal{H}$-divergence between P_S and P_T is defined as

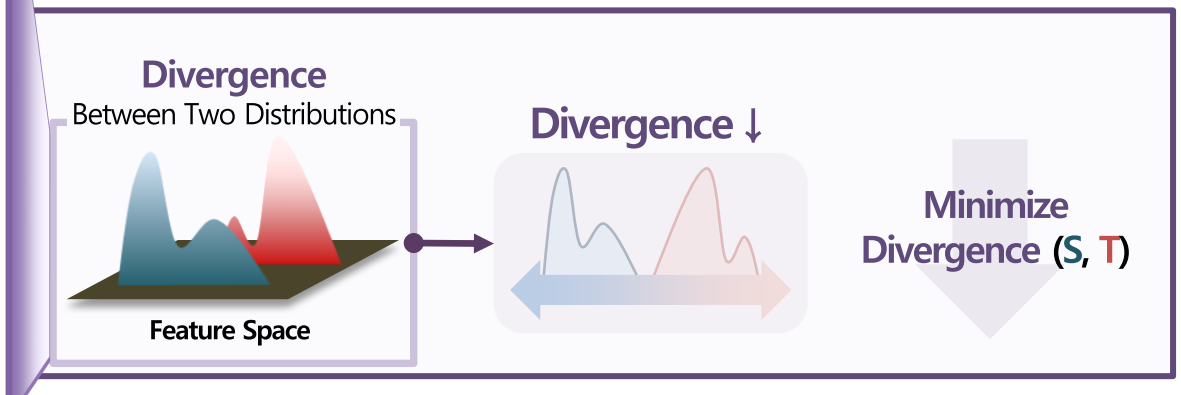
$$d_{\mathcal{H}}(P_S, P_T) = 2 \sup_{h' \in \mathcal{H}} \left| \Pr_{P_S}[h'(\mathcal{R}(x)) = 1] - \Pr_{P_T}[h'(\mathcal{R}(x)) = 1] \right|.$$

- 특정한 모델 h'에 대한 사건
- h'에 의해 z가 1로 분류될 사건으로 특정

② $\mathcal{H}$-Divergence

Can **account for the hypothesis set**!

Can be **approximated empirically** given samples!



[3] Ben-David, S., Blitzer, J., Crammer, K., & Pereira, F. (2006). Analysis of representations for domain adaptation. Advances in neural information processing systems, 19.

Methods

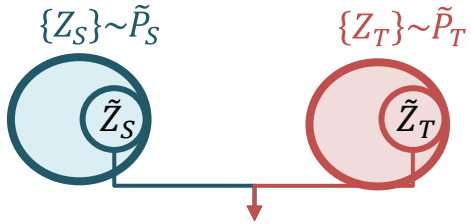
목적 | Source Domain에서 학습한 Classifier $h: Z \rightarrow Y$가 Target Domain에서도 잘 동작하게 만들자

방법 | Task-Discriminative and Domain-Invariant Feature Representation을 이용하자!

Analysis of Representations for Domain Adaptation, NIPS, 2006

Recap : $\mathcal{H}$-Divergence

$$d_{\mathcal{H}}(P_S, P_T) = 2 \sup_{h' \in \mathcal{H}} \left| \Pr_{P_S}[h'(\mathcal{R}(x)) = 1] - \Pr_{P_T}[h'(\mathcal{R}(x)) = 1] \right|$$



m개의 samples을 뽑아
경험적으로 $\mathcal{H}$-divergence 값을 구하자

$\hat{d}_{\mathcal{H}}(\tilde{Z}_S, \tilde{Z}_T)$

Empirically Estimated $\mathcal{H}$-Divergence[4]

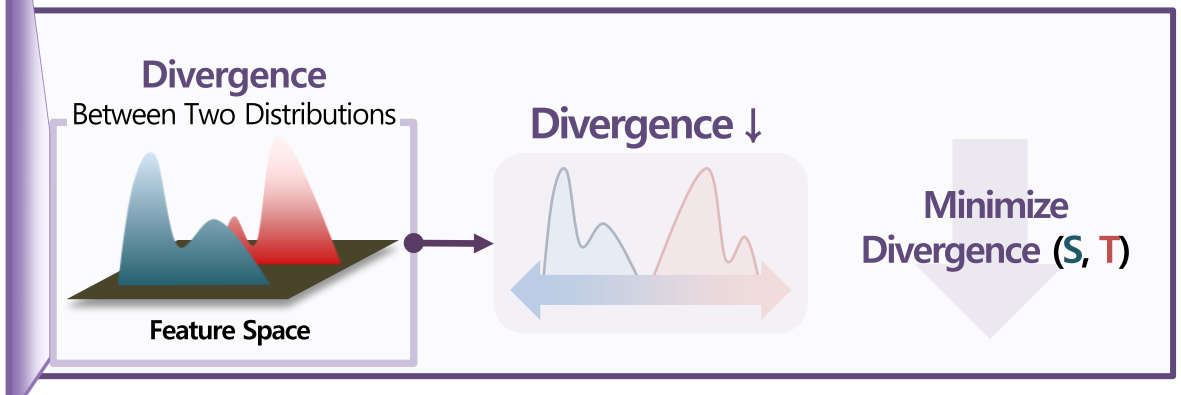
Lemma 1, Ben-David et al., 2010

Let $\mathcal{H}$ be a hypothesis class of VC dimension d.
Suppose we have two samples $\tilde{Z}_S$ and $\tilde{Z}_T$,
each of size m iid from $\tilde{P}_S$ and $\tilde{P}_T$, respectively, and
$\hat{d}_{\mathcal{H}}(\tilde{Z}_S, \tilde{Z}_T)$ is the empirical $\mathcal{H}$-divergence between samples,
then for any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$d_{\mathcal{H}}(P_S, P_T) \leq \hat{d}_{\mathcal{H}}(\tilde{Z}_S, \tilde{Z}_T) + 4\sqrt{\frac{d \log(2m) + \log\frac{2}{\delta}}{m}}.$$

③ Empirical $\mathcal{H}$-Divergence

Can **account for the hypothesis set**!
Can be **approximated empirically** given samples!



[3] Ben-David, S., Blitzer, J., Crammer, K., & Pereira, F. (2006). Analysis of representations for domain adaptation. Advances in neural information processing systems, 19.
[4] Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., & Vaughan, J. W. (2010). A theory of learning from different domains. Machine learning, 79, 151-175..

Methods

목적 | Source Domain에서 학습한 Classifier $h: Z \rightarrow Y$가 Target Domain에서도 잘 동작하게 만들자

방법 | Task-Discriminative and Domain-Invariant Feature Representation을 이용하자!

Analysis of Representations for Domain Adaptation, NIPS, 2006

Recap : $\mathcal{H}$-Divergence

$$d_{\mathcal{H}}(P_S, P_T) = 2 \sup_{h' \in \mathcal{H}} \left| \Pr_{P_S}[h'(\mathcal{R}(x)) = 1] - \Pr_{P_T}[h'(\mathcal{R}(x)) = 1] \right|$$

$$\Pr_{\tilde{P}_S}[h'(z) = 1] - \Pr_{\tilde{P}_T}[h'(z) = 1]$$

$$= \Pr_{\tilde{P}_S}[h'(z) = 1] + \Pr_{\tilde{P}_T}[h'(z) = 0] - 1$$

$$= 1 - \left(\left[\mathbb{E}_{\tilde{P}_S} I[h'(z) = 1] + \mathbb{E}_{\tilde{P}_T} I[h'(z) = 0] \right] \right)$$

Empirically Estimated $\mathcal{H}$-Divergence[4]

Lemma 2, Ben-David et al., 2010

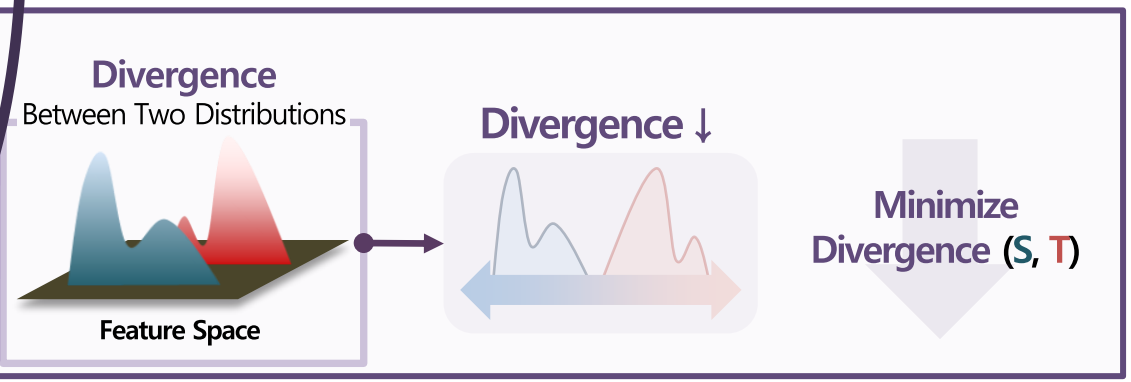
For a symmetric hypothesis set $\mathcal{H}$ (one where for every $h' \in \mathcal{H}$, the inverse hypothesis $1 - h'$ is also in $\mathcal{H}$) and $I[\cdot]$ is the binary indicator function,

$$\hat{d}_{\mathcal{H}}(\tilde{Z}_S, \tilde{Z}_T) = 2 \left(1 - \min_{h' \in \mathcal{H}} \left[\mathbb{E}_{z \sim \tilde{P}_S} I[h'(z) = 1] + \mathbb{E}_{z \sim \tilde{P}_T} I[h'(z) = 0] \right] \right)$$

$z \in \tilde{Z}_S$ $z \in \tilde{Z}_T$

③ Empirical $\mathcal{H}$-Divergence

Compute empirical $\mathcal{H}$-Divergence by **finding a classifier which attempts to separate one domain from the other!**



[3] Ben-David, S., Blitzer, J., Crammer, K., & Pereira, F. (2006). Analysis of representations for domain adaptation. Advances in neural information processing systems, 19.
[4] Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., & Vaughan, J. W. (2010). A theory of learning from different domains. Machine learning, 79, 151-175..

Methods

목적 | Source Domain에서 학습한 Classifier $h: Z \rightarrow Y$가 Target Domain에서도 잘 동작하게 만들자

방법 | Task-Discriminative and Domain-Invariant Feature Representation을 이용하자!

Analysis of Representations for Domain Adaptation, NIPS, 2006

Recap : $\mathcal{H}$-Divergence

$$d_{\mathcal{H}}(P_S, P_T) = 2 \sup_{h' \in \mathcal{H}} \left| \Pr_{P_S}[h'(\mathcal{R}(x)) = 1] - \Pr_{P_T}[h'(\mathcal{R}(x)) = 1] \right|$$

Empirically Estimated $\mathcal{H}$-Divergence[4]

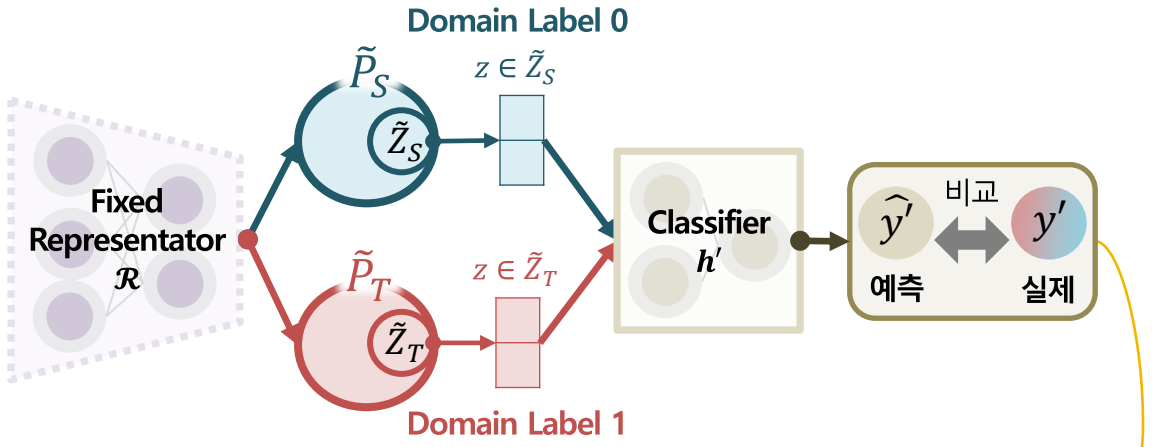
Lemma 2, Ben-David et al., 2010

For a symmetric hypothesis set $\mathcal{H}$ (one where for every $h' \in \mathcal{H}$, the inverse hypothesis $1 - h'$ is also in $\mathcal{H}$) and $I[\cdot]$ is the binary indicator function.

$$\hat{d}_{\mathcal{H}}(\tilde{Z}_S, \tilde{Z}_T) = 2\left(1 - \min_{h' \in \mathcal{H}} \left[\mathbb{E}_{z \sim \tilde{P}_S} I[h'(z) = 1] + \mathbb{E}_{z \sim \tilde{P}_T} I[h'(z) = 0] \right]\right).$$

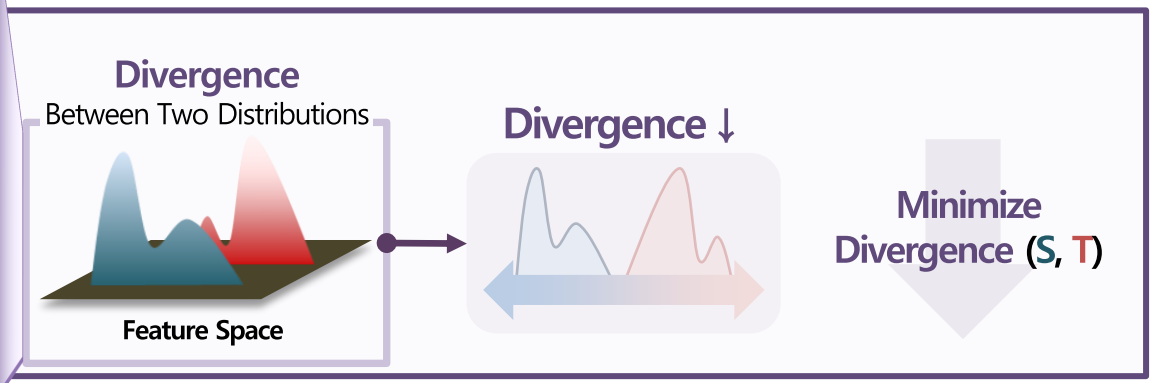
Domain Label 0 → $h'(z) = 1$이면 error

Domain Label 1 → $h'(z) = 0$이면 error



③ Empirical $\mathcal{H}$-Divergence

Compute empirical $\mathcal{H}$-Divergence by **finding a classifier which attempts to separate one domain from the other!**



[3] Ben-David, S., Blitzer, J., Crammer, K., & Pereira, F. (2006). Analysis of representations for domain adaptation. Advances in neural information processing systems, 19.
[4] Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., & Vaughan, J. W. (2010). A theory of learning from different domains. Machine learning, 79, 151-175..

Methods

목적 | Source Domain에서 학습한 Classifier $h: Z \rightarrow Y$가 Target Domain에서도 잘 동작하게 만들자

방법 | Task-Discriminative and Domain-Invariant Feature Representation을 이용하자!

Analysis of Representations for Domain Adaptation, NIPS, 2006

Recap : $\mathcal{H}$-Divergence

$$d_{\mathcal{H}}(P_S, P_T) = 2 \sup_{h' \in \mathcal{H}} \left| \Pr_{P_S}[h'(\mathcal{R}(x)) = 1] - \Pr_{P_T}[h'(\mathcal{R}(x)) = 1] \right|$$

Empirically Estimated $\mathcal{H}$-Divergence[4]

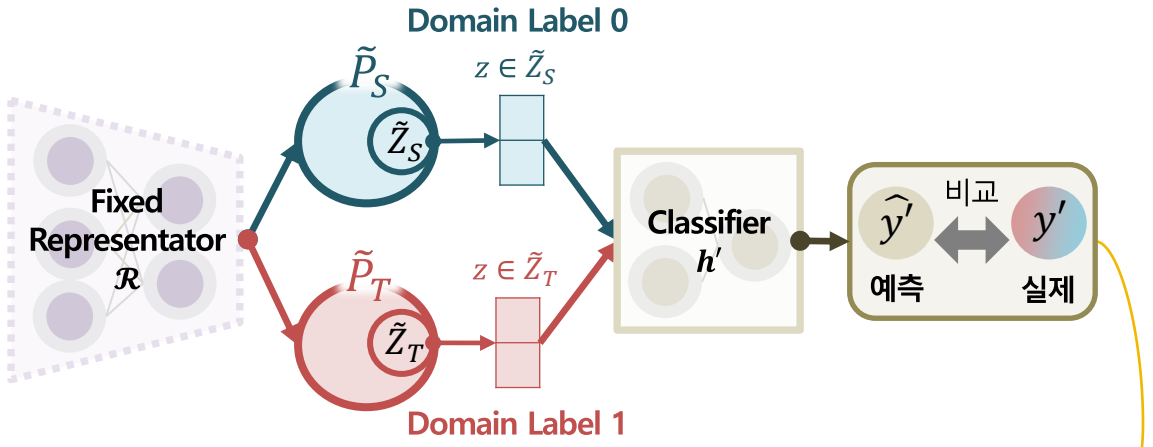
Lemma 2, Ben-David et al., 2010

For a symmetric hypothesis set $\mathcal{H}$ (one where for every $h' \in \mathcal{H}$, the inverse hypothesis $1 - h'$ is also in $\mathcal{H}$) and $I[\cdot]$ is the binary indicator function.

$$\hat{d}_{\mathcal{H}}(\tilde{Z}_S, \tilde{Z}_T) = 2\left(1 - \min_{h' \in \mathcal{H}} \left[\mathbb{E}_{z \sim \tilde{P}_S} I[h'(z) = 1] + \mathbb{E}_{z \sim \tilde{P}_T} I[h'(z) = 0] \right]\right).$$

Domain Label 0 → $h'(z) = 1$이면 error

Domain Label 1 → $h'(z) = 0$이면 error

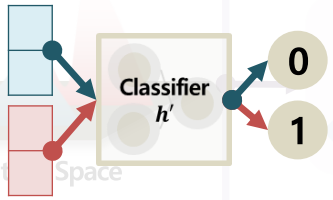


③ Empirical $\mathcal{H}$-Divergence

Compute empirical $\mathcal{H}$-Divergence by **finding a classifier which attempts to separate one domain from the other!**

정확히 구분한다면?

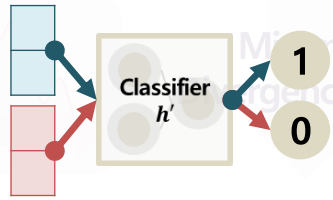
$\hat{d}_{\mathcal{H}}(\tilde{Z}_S, \tilde{Z}_T) = 2(1 - 0) = 2$



$\mathbb{E}_{z \sim \tilde{P}_S} I[h'(z) = 1] + \mathbb{E}_{z \sim \tilde{P}_T} I[h'(z) = 0] = 0$

구분을 못 한다면?

$\hat{d}_{\mathcal{H}}(\tilde{Z}_S, \tilde{Z}_T) = 2(1 - 2) = -2$



$\mathbb{E}_{z \sim \tilde{P}_S} I[h'(z) = 1] + \mathbb{E}_{z \sim \tilde{P}_T} I[h'(z) = 0] = 2$

[3] Ben-David, S., Blitzer, J., Crammer, K., & Pereira, F. (2006). Analysis of representations for domain adaptation. Advances in neural information processing systems, 19.
[4] Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., & Vaughan, J. W. (2010). A theory of learning from different domains. Machine learning, 79, 151-175.

Methods

목적 | Source Domain에서 학습한 Classifier $h: Z \rightarrow Y$가 Target Domain에서도 잘 동작하게 만들자

방법 | Task-Discriminative and Domain-Invariant Feature Representation을 이용하자!

Analysis of Representations for Domain Adaptation, NIPS, 2006

Recap : $\mathcal{H}$-Divergence

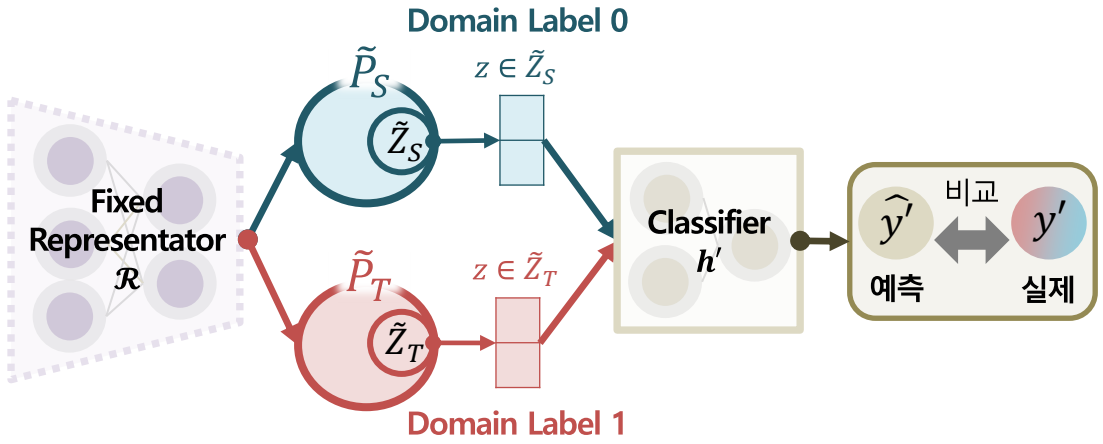
$$d_{\mathcal{H}}(P_S, P_T) = 2 \sup_{h' \in \mathcal{H}} \left| \Pr_{P_S}[h'(\mathcal{R}(x)) = 1] - \Pr_{P_T}[h'(\mathcal{R}(x)) = 1] \right|$$

$\hat{d}_{\mathcal{A}}$-Distance[3]

Proxy $\mathcal{A}$-distance (PAD)

Define the error of a classifier h' on the task of discriminating between points sampled from different distributions as $\epsilon(h') = \frac{1}{2m} \sum_{i=1}^{2m} \left| h(z_i) - \mathrm{I}[z_i \in \tilde{Z}_T] \right|$, where $I[z_i \in \tilde{Z}_T]$ is the indicator function, then:

$$\hat{d}_{\mathcal{A}}(\tilde{Z}_S, \tilde{Z}_T) = 2\left(1 - 2 \min_{h' \in \mathcal{H}} \epsilon(h')\right).$$

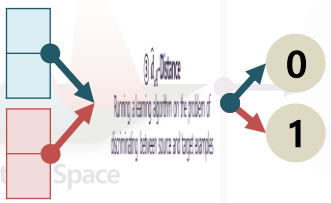


③ $\hat{d}_{\mathcal{A}}$-Distance

Running a learning algorithm on the problem of discriminating between source and target examples

정확히 구분한다면?

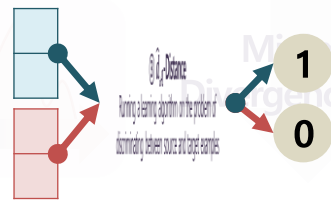
$\hat{d}_{\mathcal{H}}(\tilde{Z}_S, \tilde{Z}_T) = 2(1-0) = 2$



$\mathbb{E}_{z \sim \tilde{P}_S} I[h'(z) = 1] + \mathbb{E}_{z \sim \tilde{P}_T} I[h'(z) = 0] = 0$

구분을 못 한다면?

$\hat{d}_{\mathcal{H}}(\tilde{Z}_S, \tilde{Z}_T) = 2(1-2) = -2$



$\mathbb{E}_{z \sim \tilde{P}_S} I[h'(z) = 1] + \mathbb{E}_{z \sim \tilde{P}_T} I[h'(z) = 0] = 2$

[3] Ben-David, S., Blitzer, J., Crammer, K., & Pereira, F. (2006). Analysis of representations for domain adaptation. Advances in neural information processing systems, 19.
[4] Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., & Vaughan, J. W. (2010). A theory of learning from different domains. Machine learning, 79, 151-175..

Methods

목적	Source Domain에서 학습한 Classifier $h: Z \rightarrow Y$가 Target Domain에서도 잘 동작하게 만들자
방법	Task-Discriminative and Domain-Invariant Feature Representation을 이용하자!

Analysis of Representations for Domain Adaptation, NIPS, 2006

Cross-domain Generalization

Target Domain Error ≤ **Source Domain Error** + **Divergence($\tilde{P}_S$, $\tilde{P}_T$)**

- $\mathcal{R}: X \rightarrow Z$
- $f: X \rightarrow \{0,1\}$ True labeling function
- $\tilde{f}: Z \rightarrow \{0,1\}$ Induced image of f under $\mathcal{R}$ ← 모사
- $h: Z \rightarrow \{0,1\}$ Hypothesis
- $\epsilon(h, \tilde{f}) = \mathbb{E}_{z \sim \tilde{P}}\left[\left|h(z) - \tilde{f}(z)\right|\right] = \epsilon(h)$

Empirical $\mathcal{H}$-Divergence[3]

$$\hat{d}_{\mathcal{H}}(\tilde{Z}_S, \tilde{Z}_T) = 2\left(1 - \min_{h' \in \mathcal{H}}\left[\mathbb{E}_{z \sim \tilde{P}_S} I[h'(z) = 1] + \mathbb{E}_{z \sim \tilde{P}_T} I[h'(z) = 0]\right]\right)$$

Theorem 2. (Ben-David et al., 2006)

Let $\mathcal{H}$ is a hypothesis class of VC dimension d. With probability at least $1 - \delta$, for every $h \in \mathcal{H}$:

$$\epsilon_T(h) \leq \hat{\epsilon}_S(h) + \sqrt{d \log \frac{2em}{d} + \log \frac{4}{\delta}} + \hat{d}_{\mathcal{H}}(\tilde{Z}_S, \tilde{Z}_T) + 4\sqrt{\frac{d \log(2m) + \log \frac{4}{\delta}}{m}} + \lambda,$$

$$with\ \lambda \geq \inf_{h^* \in \mathcal{H}}[\epsilon_S(h^*) + \epsilon_T(h^*)]. \rightarrow$$ 두 분포에서 모두 잘 동작하는 이상적인 classifier h^*가 $\mathcal{H}$에 존재할 때 $\epsilon_T(h)$ 최소화 가능

[3] Ben-David, S., Blitzer, J., Crammer, K., & Pereira, F. (2006). Analysis of representations for domain adaptation. Advances in neural information processing systems, 19.
[4] Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., & Vaughan, J. W. (2010). A theory of learning from different domains. Machine learning, 79, 151-175..

Methods

목적 Source Domain에서 학습한 Classifier $h: Z \rightarrow Y$가 Target Domain에서도 잘 동작하게 만들자

방법 Task-Discriminative and Domain-Invariant Feature Representation을 이용하자!

Analysis of Representations for Domain Adaptation, NIPS, 2006

Cross-domain Generalization

Target Domain Error ≤ Source Domain Error + Divergence($\tilde{P}_S$, $\tilde{P}_T$)

- $\mathcal{R}: X \rightarrow Z$
- $f: X \rightarrow \{0,1\}$ True labeling function
- $\tilde{f}: Z \rightarrow \{0,1\}$ Induced image of f under $\mathcal{R}$ ← 모사
- $h: Z \rightarrow \{0,1\}$ Hypothesis
- $\epsilon(h, \tilde{f}) = \mathbb{E}_{z \sim \tilde{P}}[|h(z) - \tilde{f}(z)|] = \epsilon(h)$

Empirical $\mathcal{H}$-Divergence[3]

$$\hat{d}_{\mathcal{H}}(\tilde{Z}_S, \tilde{Z}_T) = 2\left(1 - \min_{h' \in \mathcal{H}}\left[\mathbb{E}_{z \sim \tilde{P}_S} I[h'(z) = 1] + \mathbb{E}_{z \sim \tilde{P}_T} I[h'(z) = 0]\right]\right)$$

Theorem 2. (Ben-David et al., 2006)

Let $\mathcal{H}$ $\epsilon_S(h) \leq \hat{\epsilon}_S(h) + \sqrt{d \log \frac{2em}{d} + \log \frac{4}{\delta}}$ VC dimension d. With probability at least $1 - \delta$, for every $h \in \mathcal{H}$:

$$\epsilon_T(h) \leq \hat{\epsilon}_S(h) + \sqrt{d \log \frac{2em}{d} + \log \frac{4}{\delta}} + \hat{d}_{\mathcal{H}}(\tilde{Z}_S, \tilde{Z}_T) + 4\sqrt{\frac{d \log(2m) + \log \frac{4}{\delta}}{m}} + \lambda,$$

$$with\ \lambda \geq \inf_{h^* \in \mathcal{H}}[\epsilon_S(h^*) + \epsilon_T(h^*)].$$ → 두 분포에서 모두 잘 동작하는 이상적인 classifier h^*가 $\mathcal{H}$에 존재할 때 $\epsilon_T(h)$ 최소화 가능

[3] Ben-David, S., Blitzer, J., Crammer, K., & Pereira, F. (2006). Analysis of representations for domain adaptation. Advances in neural information processing systems, 19.
[4] Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., & Vaughan, J. W. (2010). A theory of learning from different domains. Machine learning, 79, 151-175.

Methods

목적 | Source Domain에서 학습한 Classifier $h: Z \rightarrow Y$가 Target Domain에서도 잘 동작하게 만들자

방법 | Task-Discriminative and Domain-Invariant Feature Representation을 이용하자!

Analysis of Representations for Domain Adaptation, NIPS, 2006

Cross-domain Generalization

Target Domain Error ≤ Source Domain Error + Divergence($\tilde{P}_S$, $\tilde{P}_T$)

Analysis of Representations for Domain Adaptation

Shai Ben-David
School of Computer Science
University of Waterloo
shai@cs.uwaterloo.ca

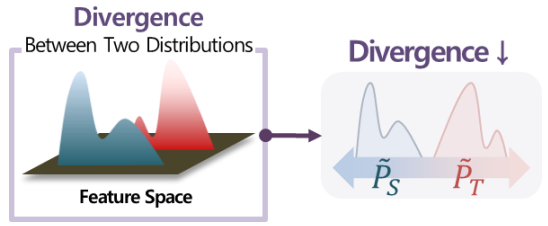
John Blitzer, Koby Crammer, and Fernando Pereira
Department of Computer and Information Science
University of Pennsylvania
{blitzer, crammer, pereira}@cis.upenn.edu

Abstract

Discriminative learning methods for classification perform well when training and test data are drawn from the same distribution. In many situations, though, we have labeled training data for a *source* domain, and we wish to learn a classifier which performs well on a *target* domain with a different distribution. Under what conditions can we adapt a classifier trained on the source domain for use in the target domain? Intuitively, a good feature representation is a crucial factor in the success of domain adaptation. We formalize this intuition theoretically with a generalization bound for domain adaption. Our theory illustrates the tradeoffs inherent in designing a representation for domain adaptation and gives a new justification for a recently proposed model. It also points toward a promising new model for domain adaptation: one which explicitly minimizes the difference between the source and target domains, while at the same time maximizing the margin of the training set.

Empirical $\mathcal{H}$-Divergence[3]

$$\hat{d}_{\mathcal{H}}(\tilde{Z}_S, \tilde{Z}_T) = 2\left(1 - \min_{h' \in \mathcal{H}} \left[\mathbb{E}_{z \sim \tilde{P}_S} I[h'(z) = 1] + \mathbb{E}_{z \sim \tilde{P}_T} I[h'(z) = 0]\right]\right)$$



$\mathcal{H}$-Divergence가 작을 때,
= Representation을 Classifier h'가 구분할 수 없을 때,
= Classifier h'의 error가 큰 Domain-Invariant Representation에서,
↓
성공적으로 Domain Adaptation 수행 가능

[3] Ben-David, S., Blitzer, J., Crammer, K., & Pereira, F. (2006). Analysis of representations for domain adaptation. Advances in neural information processing systems, 19.
[4] Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., & Vaughan, J. W. (2010). A theory of learning from different domains. Machine learning, 79, 151-175..

Methods

DANN, JMLR, 2016

목적	Source Domain에서 학습한 Classifier $h: Z \rightarrow Y$가 Target Domain에서도 잘 동작하게 만들자
방법	Learn a Task-Discriminative and Domain-Invariant Feature Representation!

❖ Domain Adversarial Training of Neural Networks (Ganin et al, 2016)[5]

- Domain Adaptation을 위해 선행연구[3]에서 제안했던 $\mathcal{H}$-Divergence를 최적화
- GAN[6]에서 제안된 적대적 학습 (Adversarial Learning) 개념을 차용하여 학습 수행

JMLR, 24년 3월 기준 8689회 인용

Domain-Adversarial Training of Neural Networks

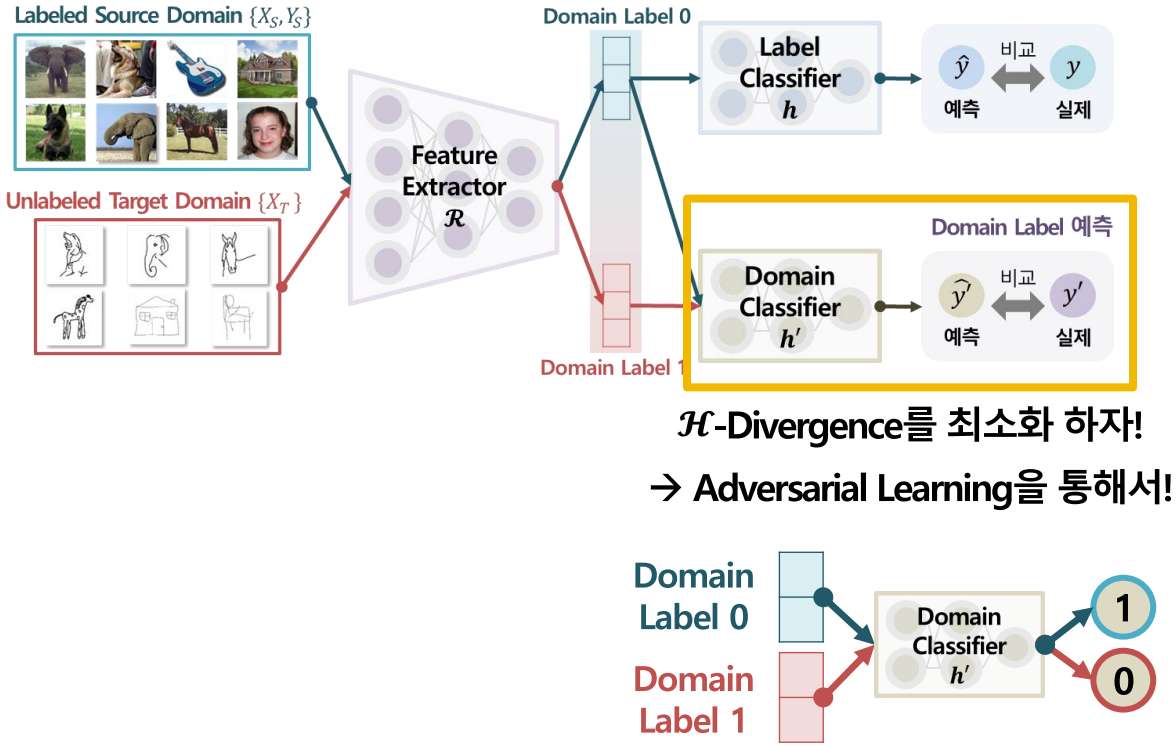
Abstract

We introduce a new representation learning approach for domain adaptation, in which data at training and test time come from similar but different distributions. Our approach is directly inspired by the theory on domain adaptation suggesting that, for effective domain transfer to be achieved, predictions must be made based on features that cannot discriminate between the training (source) and test (target) domains.

The approach implements this idea in the context of neural network architectures that are trained on labeled data from the source domain and unlabeled data from the target domain (no labeled target-domain data is necessary). As the training progresses, the approach promotes the emergence of features that are (i) discriminative for the main learning task on the source domain and (ii) indiscriminate with respect to the shift between the domains. We show that this adaptation behaviour can be achieved in almost any feed-forward model by augmenting it with few standard layers and a new *gradient reversal* layer. The resulting augmented architecture can be trained using standard backpropagation and stochastic gradient descent, and can thus be implemented with little effort using any of the deep learning packages.

We demonstrate the success of our approach for two distinct classification problems (document sentiment analysis and image classification), where state-of-the-art domain adaptation performance on standard benchmarks is achieved. We also validate the approach for descriptor learning task in the context of person re-identification application.

Keywords: domain adaptation, neural network, representation learning, deep learning, synthetic data, image classification, sentiment analysis, person re-identification



[3] Ben-David, S., Blitzer, J., Crammer, K., & Pereira, F. (2006). Analysis of representations for domain adaptation. Advances in neural information processing systems, 19.
[5] Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., ... & Lempitsky, V. (2016). Domain-adversarial training of neural networks. The journal of machine learning research, 17(1), 2096-2030.
[6] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. Advances in neural information processing systems, 27.

Methods

DANN, JMLR, 2016

목적 | Source Domain에서 학습한 Classifier $h: Z \rightarrow Y$가 Target Domain에서도 잘 동작하게 만들자

방법 | Learn a Task-Discriminative and Domain-Invariant Feature Representation!

❖ **Domain Adversarial Training of Neural Networks** (Ganin et al, 2016)[5]

- Domain Adaptation을 위해 선행연구[3]에서 제안했던 $\mathcal{H}$-Divergence를 최적화
- GAN[6]에서 제안된 적대적 학습 (Adversarial Learning) 개념을 차용하여 학습 수행

JMLR, 24년 3월 기준 8689회 인용

Domain-Adversarial Training of Neural Networks

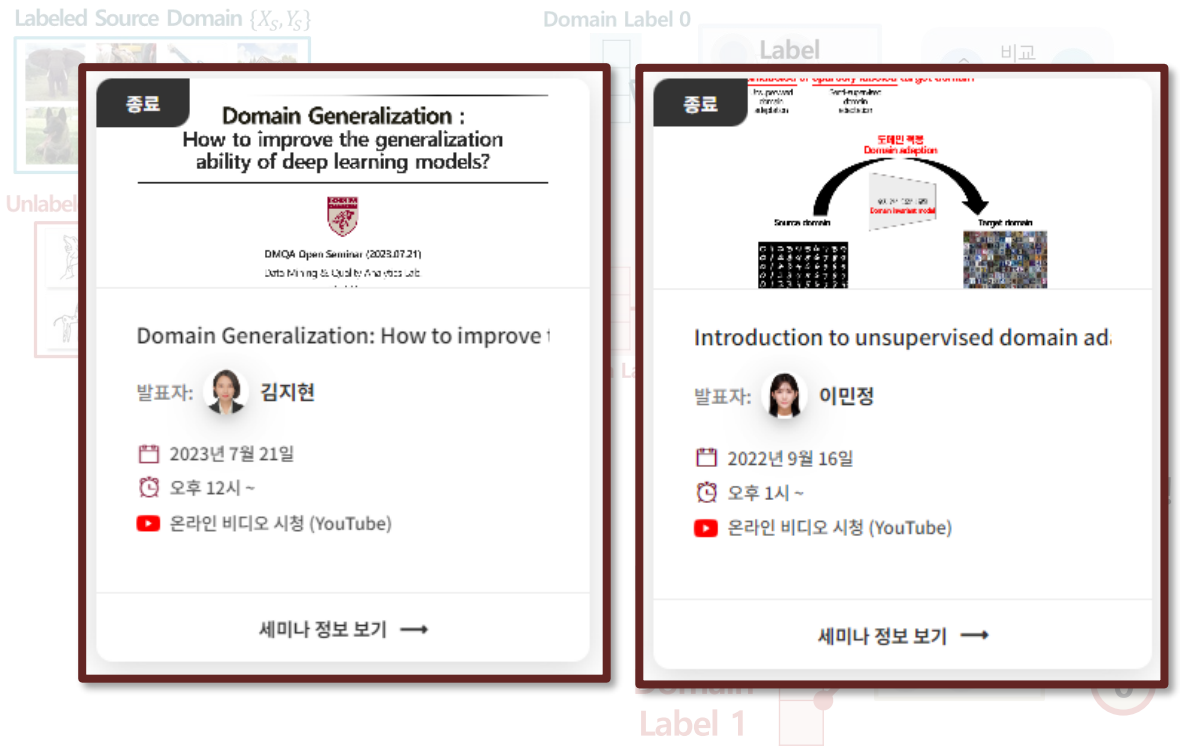
Abstract

We introduce a new representation learning approach for domain adaptation, in which data at training and test time come from similar but different distributions. Our approach is directly inspired by the theory on domain adaptation suggesting that, for effective domain transfer to be achieved, predictions must be made based on features that cannot discriminate between the training (source) and test (target) domains.

The approach implements this idea in the context of neural network architectures that are trained on labeled data from the source domain and unlabeled data from the target domain (no labeled target-domain data is necessary). As the training progresses, the approach promotes the emergence of features that are (i) discriminative for the main learning task on the source domain and (ii) indiscriminate with respect to the shift between the domains. We show that this adaptation behaviour can be achieved in almost any feed-forward model by augmenting it with few standard layers and a new *gradient reversal* layer. The resulting augmented architecture can be trained using standard backpropagation and stochastic gradient descent, and can thus be implemented with little effort using any of the deep learning packages.

We demonstrate the success of our approach for two distinct classification problems (document sentiment analysis and image classification), where state-of-the-art domain adaptation performance on standard benchmarks is achieved. We also validate the approach for descriptor learning task in the context of person re-identification application.

Keywords: domain adaptation, neural network, representation learning, deep learning, synthetic data, image classification, sentiment analysis, person re-identification



[3] Ben-David, S., Blitzer, J., Crammer, K., & Pereira, F. (2006). Analysis of representations for domain adaptation. Advances in neural information processing systems, 19.
[5] Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., ... & Lempitsky, V. (2016). Domain-adversarial training of neural networks. The journal of machine learning research, 17(1), 2096-2030.
[6] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. Advances in neural information processing systems, 27.

Methods

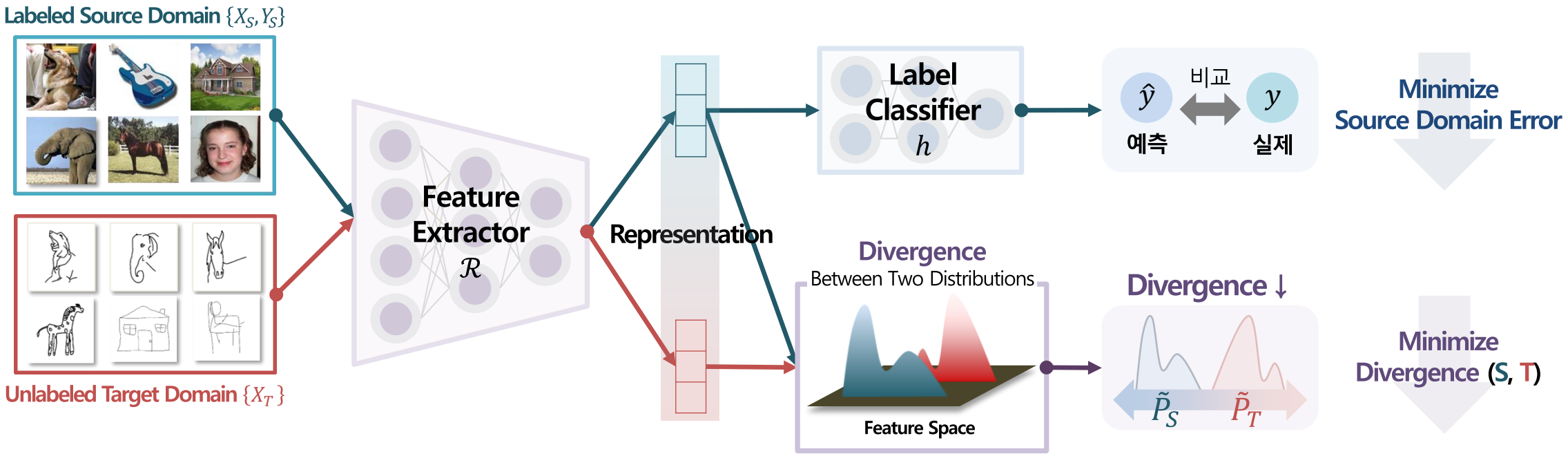
DANN, JMLR, 2016

목적 Source Domain에서 학습한 Classifier $h: Z \rightarrow Y$가 Target Domain에서도 잘 동작하게 만들자

방법 Learn a Task-Discriminative and Domain-Invariant Feature Representation!

Recap: Cross-domain Generalization

Target Domain Error ≤ Source Domain Error + Divergence($\tilde{P}_S$, $\tilde{P}_T$)



Methods

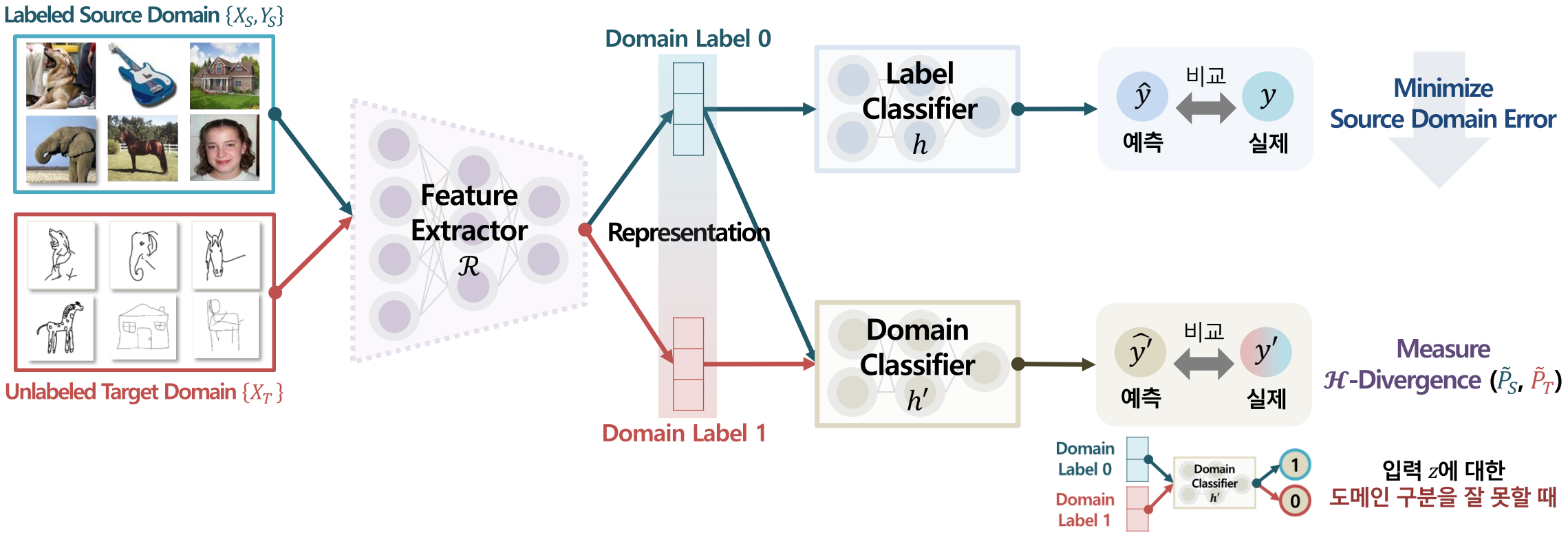
DANN, JMLR, 2016

목적: Source Domain에서 학습한 Classifier $h: Z \rightarrow Y$가 Target Domain에서도 잘 동작하게 만들자

방법: Learn a Task-Discriminative and Domain-Invariant Feature Representation!

Recap: Cross-domain Generalization

Target Domain Error ≤ Source Domain Error + $\mathcal{H}$-Divergence($\tilde{P}_S$, $\tilde{P}_T$)



[3] Ben-David, S., Blitzer, J., Crammer, K., & Pereira, F. (2006). Analysis of representations for domain adaptation. Advances in neural information processing systems, 19.

Methods

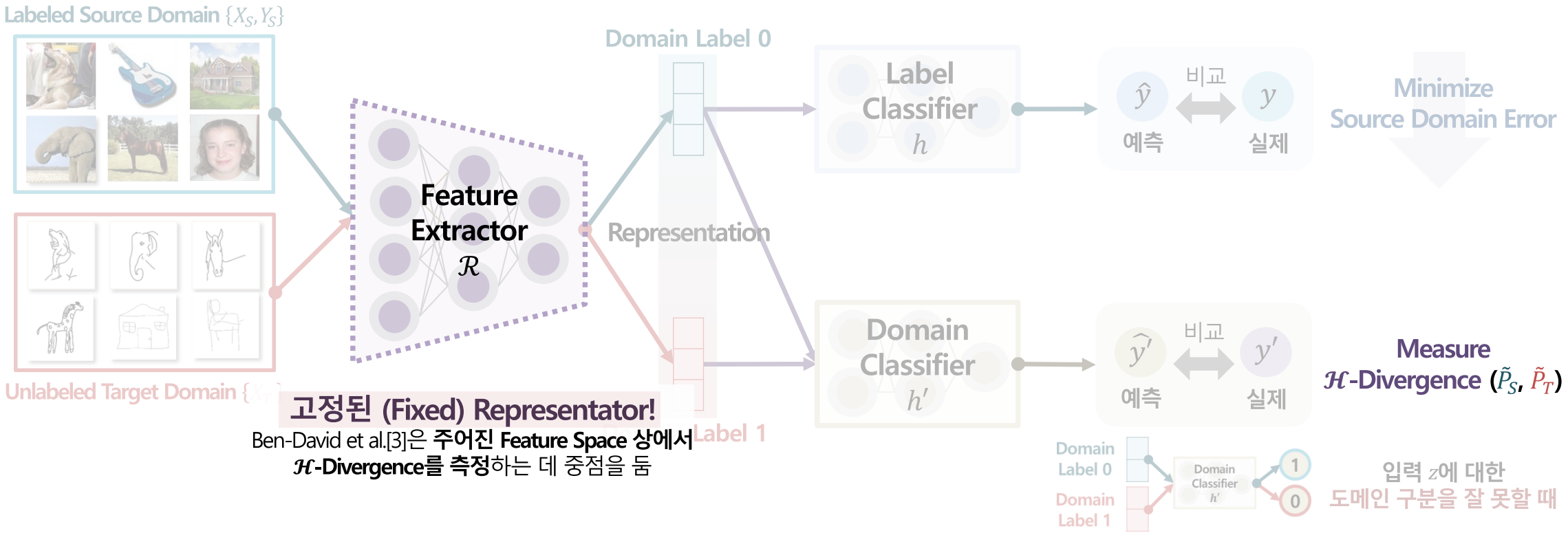
DANN, JMLR, 2016

목적 Source Domain에서 학습한 Classifier $h: Z \rightarrow Y$가 Target Domain에서도 잘 동작하게 만들자

방법 Learn a Task-Discriminative and Domain-Invariant Feature Representation!

Recap: Cross-domain Generalization

Target Domain Error ≤ Source Domain Error + $\mathcal{H}$-Divergence($\tilde{P}_S$, $\tilde{P}_T$)



고정된 (Fixed) Representator!

Ben-David et al.[3]은 **주어진 Feature Space 상에서 $\mathcal{H}$-Divergence를 측정**하는 데 중점을 둠

[3] Ben-David, S., Blitzer, J., Crammer, K., & Pereira, F. (2006). Analysis of representations for domain adaptation. Advances in neural information processing systems, 19.

Methods

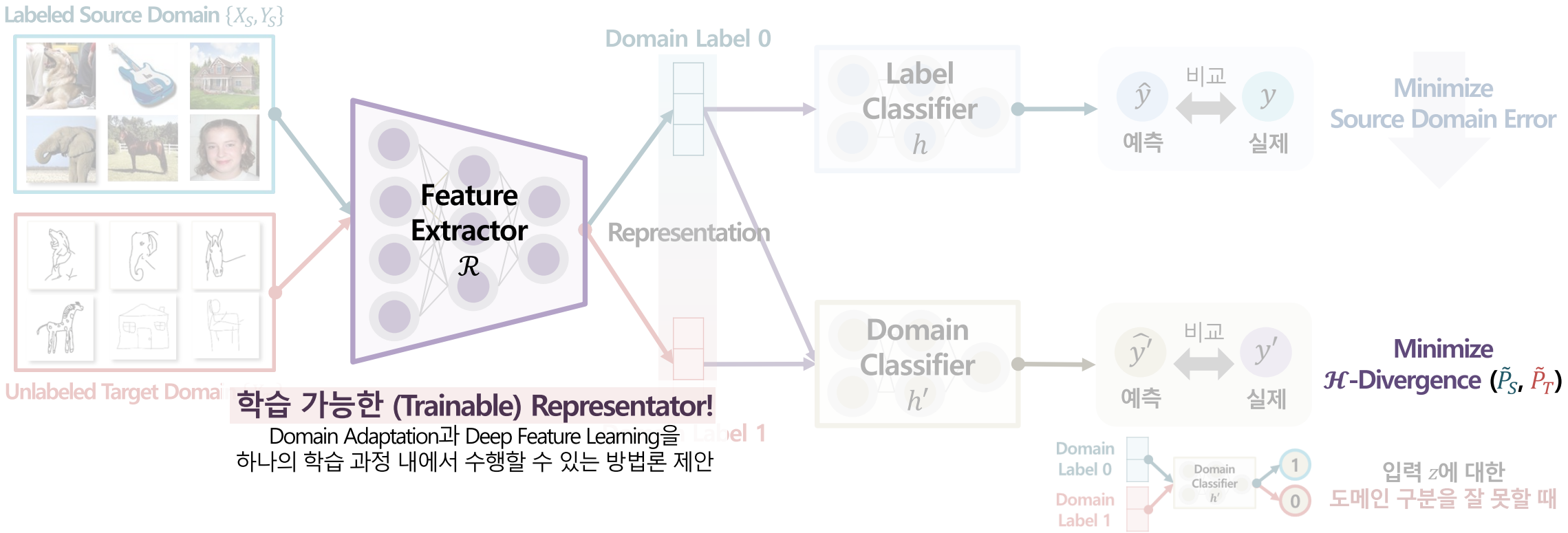
DANN, JMLR, 2016

목적 Source Domain에서 학습한 Classifier $h: Z \rightarrow Y$가 Target Domain에서도 잘 동작하게 만들자

방법 Learn a Task-Discriminative and Domain-Invariant Feature Representation!

Recap: Cross-domain Generalization

Target Domain Error ≤ Source Domain Error + $\mathcal{H}$-Divergence($\tilde{P}_S$, $\tilde{P}_T$)



학습 가능한 (Trainable) Representator!

Domain Adaptation과 Deep Feature Learning을
하나의 학습 과정 내에서 수행할 수 있는 방법론 제안

[3] Ben-David, S., Blitzer, J., Crammer, K., & Pereira, F. (2006). Analysis of representations for domain adaptation. Advances in neural information processing systems, 19.
[5] Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., ... & Lempitsky, V. (2016). Domain-adversarial training of neural networks. The journal of machine learning research, 17(1), 2096-2030.

Methods

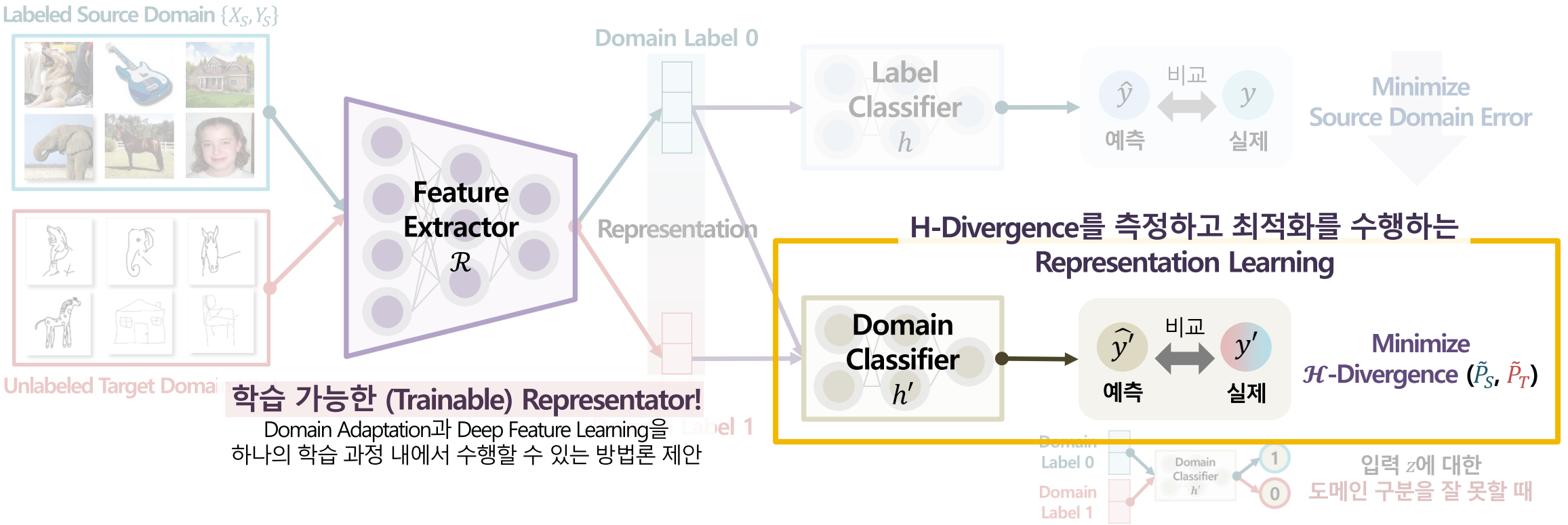
DANN, JMLR, 2016

목적 | Source Domain에서 학습한 Classifier $h: Z \rightarrow Y$가 Target Domain에서도 잘 동작하게 만들자

방법 | Learn a Task-Discriminative and Domain-Invariant Feature Representation!

❖ Domain Adversarial Training of Neural Networks (Ganin et al, 2016)[5]

- **Motivation** : To embed domain adaptation into the process of representation learning
  - $\mathcal{H}$-Divergence를 최적화하여 Task-Discriminative 하면서도 Domain-invariant한 Representation Learning 수행[5]



[3] Ben-David, S., Blitzer, J., Crammer, K., & Pereira, F. (2006). Analysis of representations for domain adaptation. Advances in neural information processing systems, 19.
[5] Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., ... & Lempitsky, V. (2016). Domain-adversarial training of neural networks. The journal of machine learning research, 17(1), 2096-2030.

Methods

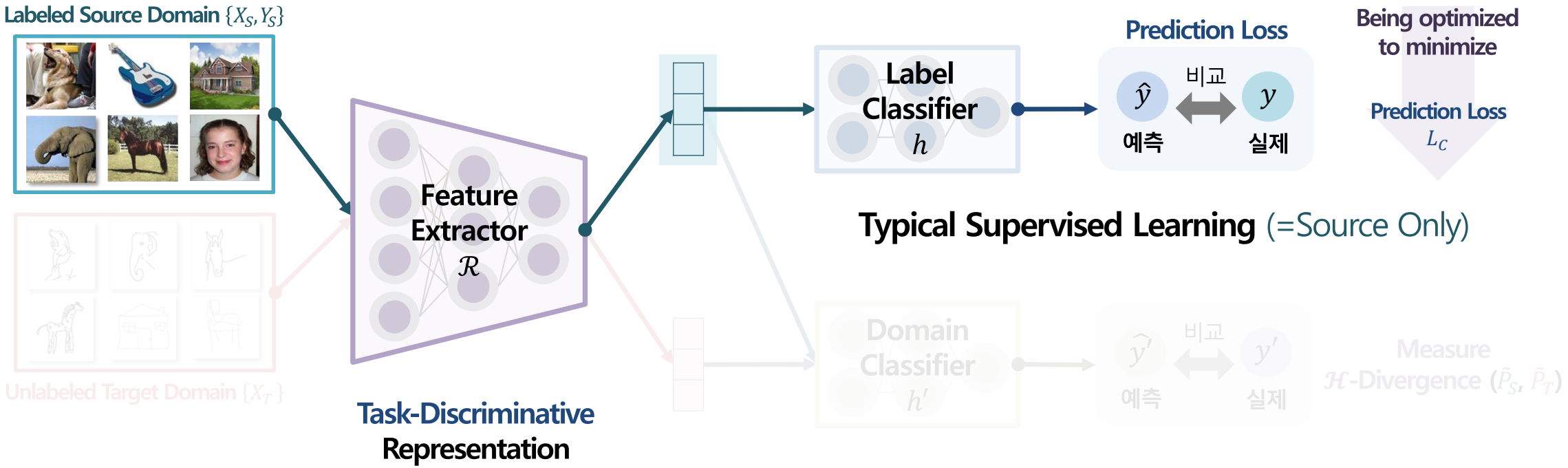
DANN, JMLR, 2016

목적 | Source Domain에서 학습한 Classifier $h: Z \rightarrow Y$가 Target Domain에서도 잘 동작하게 만들자

방법 | Learn a Task-Discriminative and Domain-Invariant Feature Representation!

❖ Domain Adversarial Training of Neural Networks (Ganin et al, 2016)[5]

- **Feature Extractor 학습 : (1) Source Domain Error ↓** (2) $\mathcal{H}$-Divergence ↓
  - Class Label이 있는 Source Domain을 기반으로, Task-Discriminative Feature 추출을 위해 Prediction (Source) Loss 최소화



[3] Ben-David, S., Blitzer, J., Crammer, K., & Pereira, F. (2006). Analysis of representations for domain adaptation. Advances in neural information processing systems, 19.
[5] Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., ... & Lempitsky, V. (2016). Domain-adversarial training of neural networks. The journal of machine learning research, 17(1), 2096-2030.

Methods

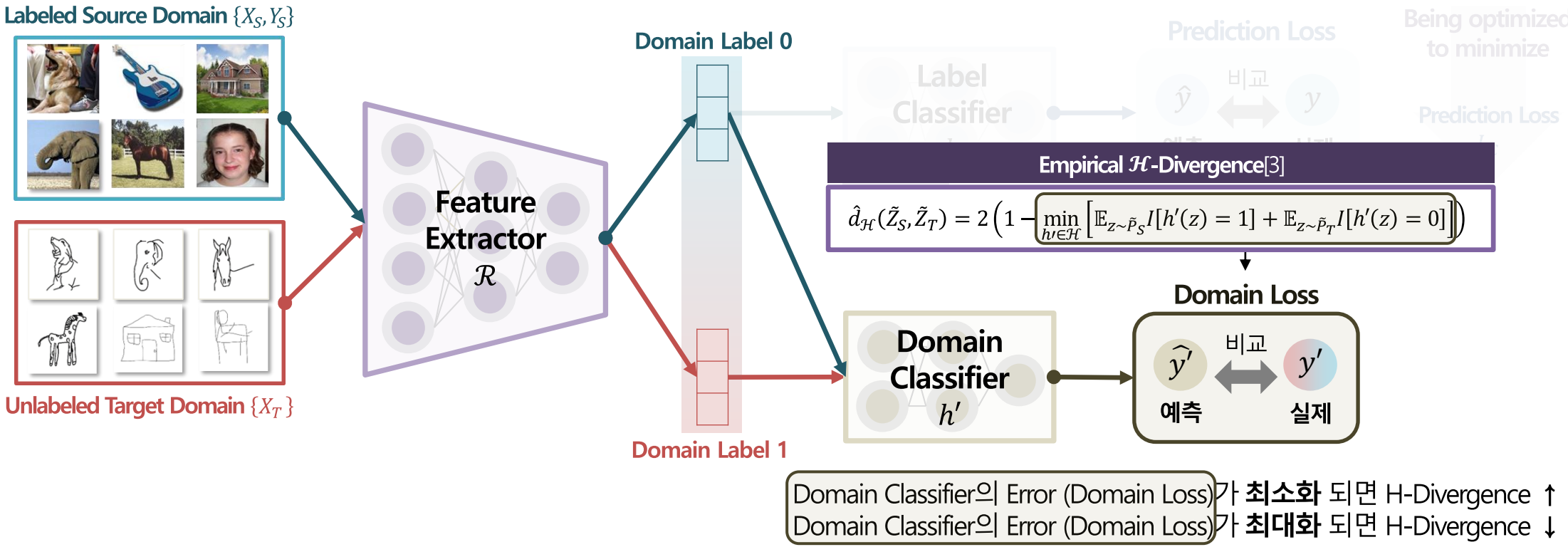
DANN, JMLR, 2016

목적 | Source Domain에서 학습한 Classifier $h: Z \rightarrow Y$가 Target Domain에서도 잘 동작하게 만들자

방법 | Learn a Task-Discriminative and Domain-Invariant Feature Representation!

❖ Domain Adversarial Training of Neural Networks (Ganin et al, 2016)[5]

- **Feature Extractor 학습** : (1) Source Domain Error ↓ **(2) $\mathcal{H}$-Divergence ↓**
  - Domain-Invariant Feature 추출을 위해 Domain Classifier의 입력 z에 대한 Domain 구분 능력 약화 = **Domain Loss 최대화**



Domain Classifier의 Error (Domain Loss)가 **최소화** 되면 H-Divergence ↑
Domain Classifier의 Error (Domain Loss)가 **최대화** 되면 H-Divergence ↓

[3] Ben-David, S., Blitzer, J., Crammer, K., & Pereira, F. (2006). Analysis of representations for domain adaptation. Advances in neural information processing systems, 19.
[5] Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., ... & Lempitsky, V. (2016). Domain-adversarial training of neural networks. The journal of machine learning research, 17(1), 2096-2030.

Methods

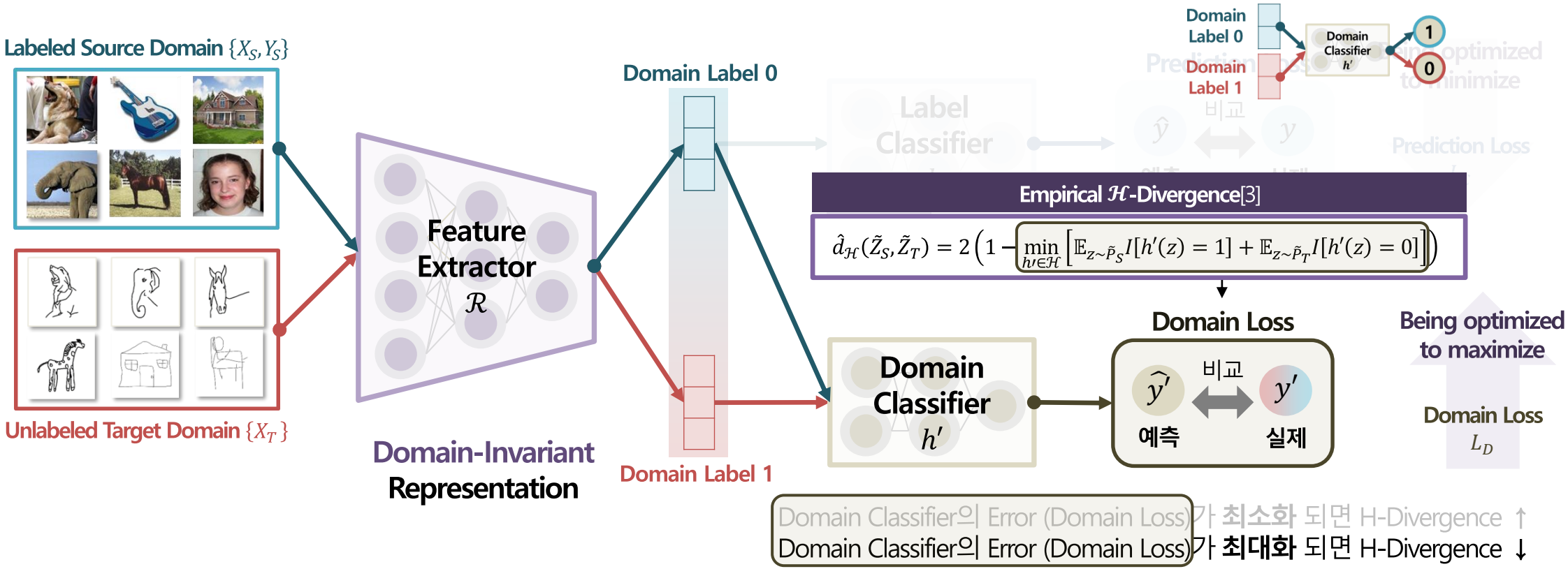
DANN, JMLR, 2016

목적 | Source Domain에서 학습한 Classifier $h: Z \rightarrow Y$가 Target Domain에서도 잘 동작하게 만들자

방법 | Learn a Task-Discriminative and Domain-Invariant Feature Representation!

❖ Domain Adversarial Training of Neural Networks (Ganin et al, 2016)[5]

- **Feature Extractor 학습** : (1) Source Domain Error ↓ **(2) $\mathcal{H}$-Divergence ↓**
  - Domain-Invariant Feature 추출을 위해 Domain Classifier의 입력 z에 대한 Domain 구분 능력 약화 = **Domain Loss 최대화**



Domain Classifier의 Error (Domain Loss)가 **최소화** 되면 H-Divergence ↑

Domain Classifier의 Error (Domain Loss)가 **최대화** 되면 H-Divergence ↓

[3] Ben-David, S., Blitzer, J., Crammer, K., & Pereira, F. (2006). Analysis of representations for domain adaptation. Advances in neural information processing systems, 19.
[5] Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., ... & Lempitsky, V. (2016). Domain-adversarial training of neural networks. The journal of machine learning research, 17(1), 2096-2030.

Methods

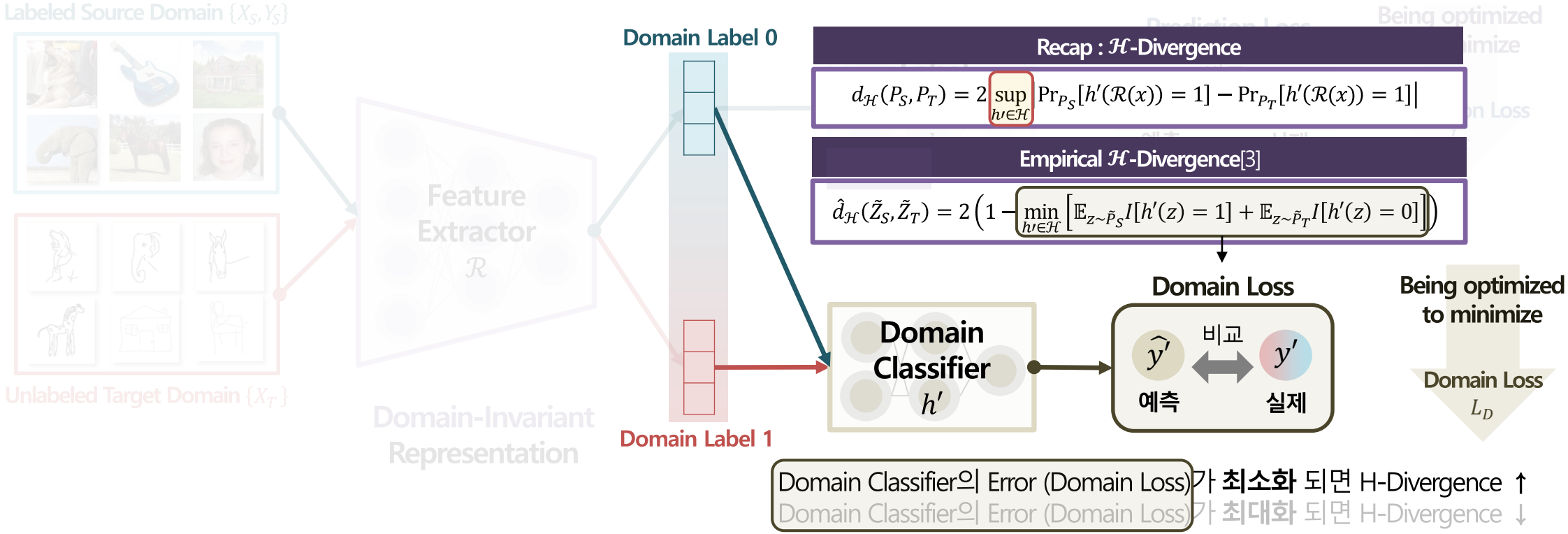
DANN, JMLR, 2016

목적 | Source Domain에서 학습한 Classifier $h: Z \rightarrow Y$가 Target Domain에서도 잘 동작하게 만들자

방법 | Learn a Task-Discriminative and Domain-Invariant Feature Representation!

❖ Domain Adversarial Training of Neural Networks (Ganin et al, 2016)[5]

- **Domain Classifier 학습 : $\mathcal{H}$-Divergence ↑**
  - 입력 z에 대한 Domain 구분 능력 강화 (=**Domain Loss 최소화**)
  - Feature Extractor의 Feature Representation이 충분히 Domain-Invariant 해질 때까지 피드백 제공



[3] Ben-David, S., Blitzer, J., Crammer, K., & Pereira, F. (2006). Analysis of representations for domain adaptation. Advances in neural information processing systems, 19.
[5] Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., ... & Lempitsky, V. (2016). Domain-adversarial training of neural networks. The journal of machine learning research, 17(1), 2096-2030.

Methods

DANN, JMLR, 2016

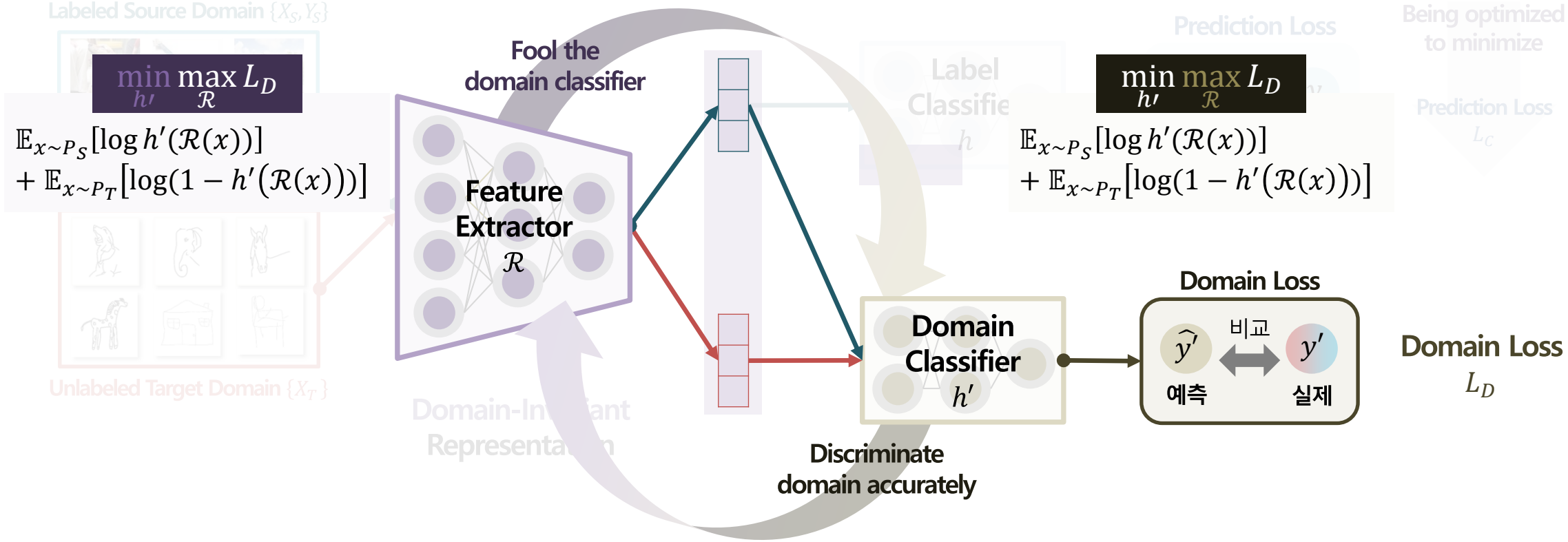
목적 | Source Domain에서 학습한 Classifier $h: Z \rightarrow Y$가 Target Domain에서도 잘 동작하게 만들자

방법 | Learn a Task-Discriminative and Domain-Invariant Feature Representation!

❖ Domain Adversarial Training of Neural Networks (Ganin et al, 2016)[5]

- **적대적 학습 (Adversarial Learning)을 통한 $\mathcal{H}$-Divergence 최적화**
  - **Feature Extractor** : 입력 z에 대한 Domain 구분 능력 약화 (=**Domain Loss 최대화**)
  - **Domain Classifier**: 입력 z에 대한 Domain 구분 능력 강화 (=**Domain Loss 최소화**)

GAN[6] with minimax loss와 유사



[3] Ben-David, S., Blitzer, J., Crammer, K., & Pereira, F. (2006). Analysis of representations for domain adaptation. Advances in neural information processing systems, 19.
[5] Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., ... & Lempitsky, V. (2016). Domain-adversarial training of neural networks. The journal of machine learning research, 17(1), 2096-2030.
[6] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. Advances in neural information processing systems, 27.

Methods

DANN, JMLR, 2016

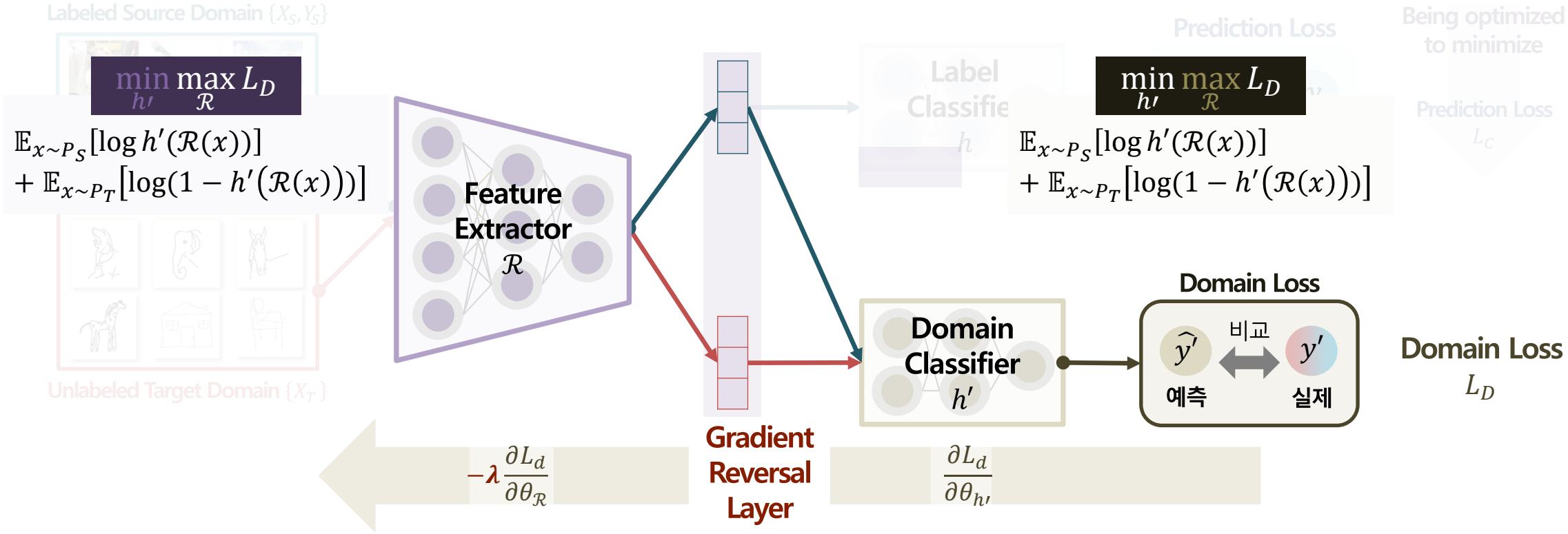
목적 | Source Domain에서 학습한 Classifier $h: Z \rightarrow Y$가 Target Domain에서도 잘 동작하게 만들자

방법 | Learn a Task-Discriminative and Domain-Invariant Feature Representation!

❖ Domain Adversarial Training of Neural Networks (Ganin et al, 2016)[5]

- **적대적 학습 (Adversarial Learning)을 통한 $\mathcal{H}$-Divergence 최적화**
  - **Feature Extractor** : 입력 z에 대한 Domain 구분 능력 약화 (=**Domain Loss 최대화**)
  - **Domain Classifier**: 입력 z에 대한 Domain 구분 능력 강화 (=**Domain Loss 최소화**)

역전파 시 gradient에 negative scalar($-\lambda$)를 곱합



[3] Ben-David, S., Blitzer, J., Crammer, K., & Pereira, F. (2006). Analysis of representations for domain adaptation. Advances in neural information processing systems, 19.
[5] Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., ... & Lempitsky, V. (2016). Domain-adversarial training of neural networks. The journal of machine learning research, 17(1), 2096-2030.

Methods

DANN, JMLR, 2016

목적 | Source Domain에서 학습한 Classifier $h: Z \rightarrow Y$가 Target Domain에서도 잘 동작하게 만들자

방법 | Learn a Task-Discriminative and Domain-Invariant Feature Representation!

❖ Domain Adversarial Training of Neural Networks (Ganin et al, 2016)[5]

- **적대적 학습 (Adversarial Learning)을 통한 $\mathcal{H}$-Divergence 최적화**
  - **Feature Extractor** : 입력 z에 대한 Domain 구분 능력 약화 (=**Domain Loss 최대화**)
  - **Domain Classifier**: 입력 z에 대한 Domain 구분 능력 강화 (=**Domain Loss 최소화**)

Domain-Adversarial Training of Neural Networks

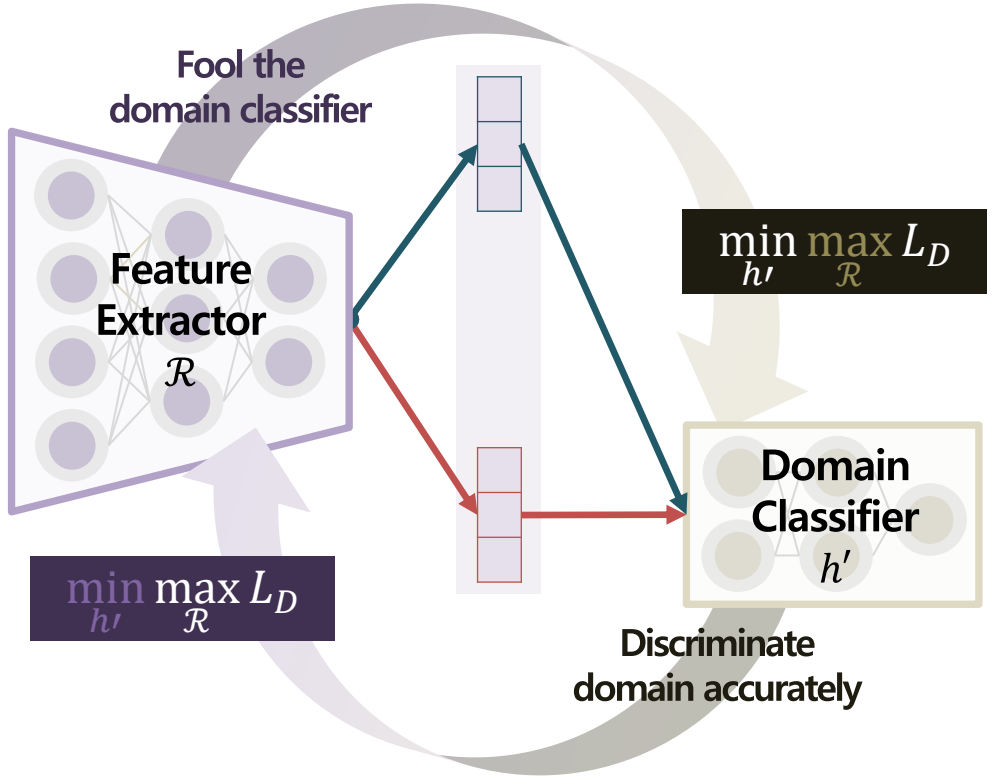
Abstract

We introduce a new representation learning approach for domain adaptation, in which data at training and test time come from similar but different distributions. Our approach is directly inspired by the theory on domain adaptation suggesting that, for effective domain transfer to be achieved, predictions must be made based on features that cannot discriminate between the training (source) and test (target) domains.

The approach implements this idea in the context of neural network architectures that are trained on labeled data from the source domain and unlabeled data from the target domain (no labeled target-domain data is necessary). As the training progresses, the approach promotes the emergence of features that are (i) discriminative for the main learning task on the source domain and (ii) indiscriminate with respect to the shift between the domains. We show that this adaptation behaviour can be achieved in almost any feed-forward model by augmenting it with few standard layers and a new *gradient reversal* layer. The resulting augmented architecture can be trained using standard backpropagation and stochastic gradient descent, and can thus be implemented with little effort using any of the deep learning packages.

We demonstrate the success of our approach for two distinct classification problems (document sentiment analysis and image classification), where state-of-the-art domain adaptation performance on standard benchmarks is achieved. We also validate the approach for descriptor learning task in the context of person re-identification application.

Keywords: domain adaptation, neural network, representation learning, deep learning, synthetic data, image classification, sentiment analysis, person re-identification



[3] Ben-David, S., Blitzer, J., Crammer, K., & Pereira, F. (2006). Analysis of representations for domain adaptation. Advances in neural information processing systems, 19.
[5] Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., ... & Lempitsky, V. (2016). Domain-adversarial training of neural networks. The journal of machine learning research, 17(1), 2096-2030.

Methods

A Theory of Learning From Different Domains, ML, 2010

❖ **A Theory of Learning From Different Domains** (Ben-David et al, 2010)[4]

- 선행연구[3]에서 제안했던 $\mathcal{H}$-Divergence에서 고도화된 지표인 $\mathcal{H}\Delta\mathcal{H}$-Divergence 지표 제안 (Lemma 3)
- 이를 기반으로 Target Domain에서의 일반화 오류 계산 시 더 나은 상계 값을 얻을 수 있음 (Theorem 2)

Machine Learning, 24년 3월 기준 3559회 인용

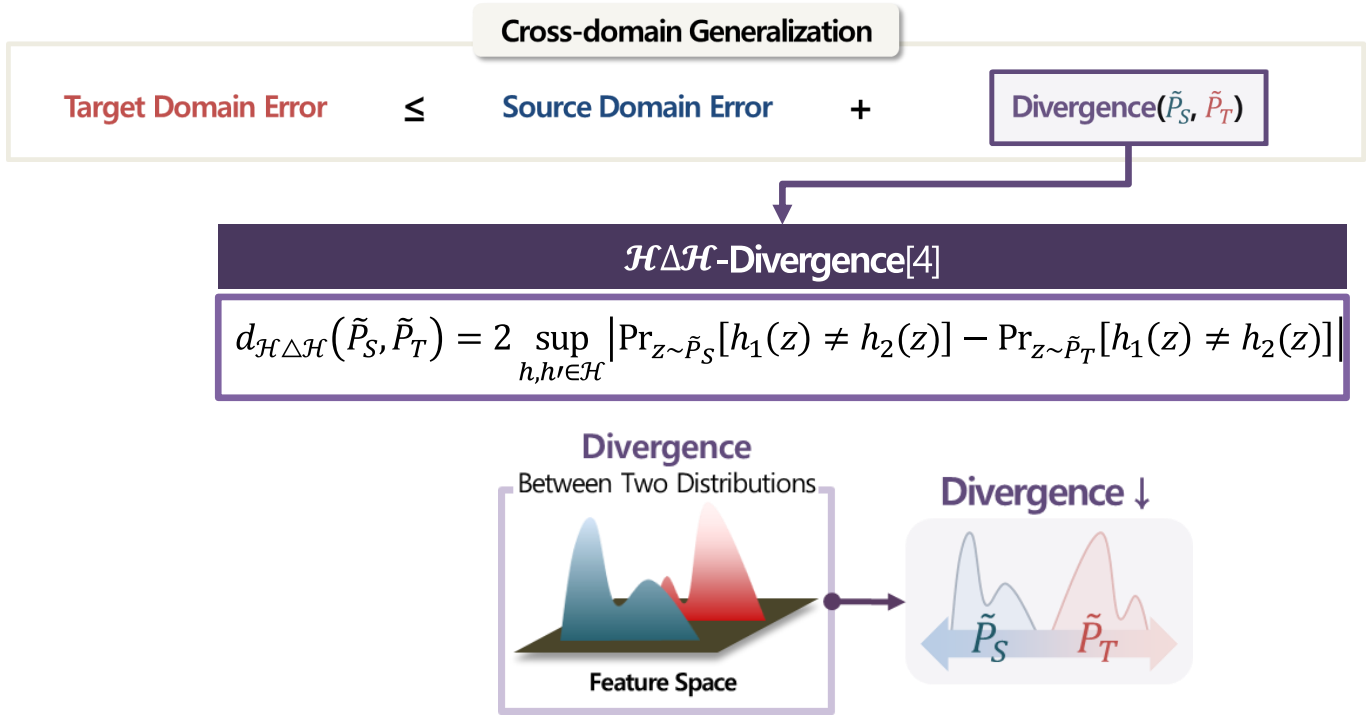
Mach Learn (2010) 79: 151–175
DOI 10.1007/s10994-009-5152-4

A theory of learning from different domains

Shai Ben-David · John Blitzer · Koby Crammer · Alex Kulesza · Fernando Pereira · Jennifer Wortman Vaughan

Received: 28 February 2009 / Revised: 12 September 2009 / Accepted: 18 September 2009 /
Published online: 23 October 2009

Abstract Discriminative learning methods for classification perform well when training and test data are drawn from the same distribution. Often, however, we have plentiful labeled training data from a *source* domain but wish to learn a classifier which performs well on a *target* domain with a different distribution and little or no labeled training data. In this work we investigate two questions. First, under what conditions can a classifier trained from source data be expected to perform well on target data? Second, given a small amount of labeled target data, how should we combine it during training with the large amount of labeled source data to achieve the lowest target error at test time?



[3] Ben-David, S., Blitzer, J., Crammer, K., & Pereira, F. (2006). Analysis of representations for domain adaptation. Advances in neural information processing systems, 19.
[4] Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., & Vaughan, J. W. (2010). A theory of learning from different domains. Machine learning, 79, 151-175..

Methods

A Theory of Learning From Different Domains, ML, 2010

Recap : $\mathcal{H}$-Divergence

$$d_{\mathcal{H}}(P_S, P_T) = 2 \sup_{h' \in \mathcal{H}} \left| \Pr_{P_S}[h'(\mathcal{R}(x)) = 1] - \Pr_{P_T}[h'(\mathcal{R}(x)) = 1] \right|$$

Recap : Empirical $\mathcal{H}$-Divergence[3]

$$\hat{d}_{\mathcal{H}}(\tilde{Z}_S, \tilde{Z}_T) = 2\left(1 - \min_{h' \in \mathcal{H}} \left[\mathbb{E}_{z \sim \tilde{P}_S} I[h'(z) = 1] + \mathbb{E}_{z \sim \tilde{P}_T} I[h'(z) = 0] \right]\right)$$

Symmetric Difference Hypothesis Space $\mathcal{H}\Delta\mathcal{H}$[4]

Definition 3, Ben-David et al., 2010

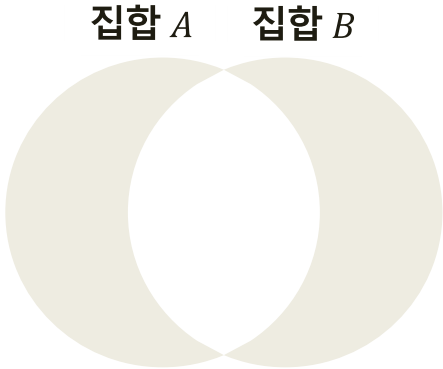
For a hypothesis space $\mathcal{H}$, the *symmetric difference hypothesis space* $\mathcal{H}\Delta\mathcal{H}$ is the set of hypotheses

$g \in \mathcal{H}\Delta\mathcal{H} \iff g(z) = h_1(z) \oplus h_2(z)$ for some $h_1, h_2 \in \mathcal{H}$,

where $\oplus$ is the XOR function. In words, every hypothesis $g \in \mathcal{H}\Delta\mathcal{H}$ is the set of disagreements between two hypotheses h_1 and h_2 in $\mathcal{H}$.

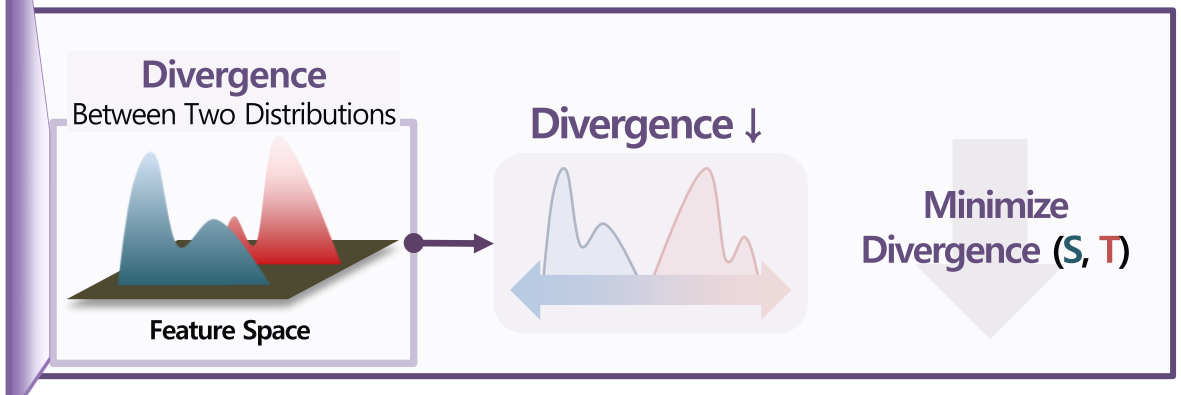
Symmetric Difference of set A and B

$A\Delta B = (A - B) \cup (B - A)$



④ $\mathcal{H}\Delta\mathcal{H}$-Divergence

Very useful in reasoning about error!
(When we give a bound on the target error)



[4] Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., & Vaughan, J. W. (2010). A theory of learning from different domains. Machine learning, 79, 151-175..

Symmetric Difference of set A and B

$A \Delta B = (A - B) \cup (B - A)$

Methods

A Theory of Learning From Different Domains, ML, 2010

Recap : $\mathcal{H}$-Divergence

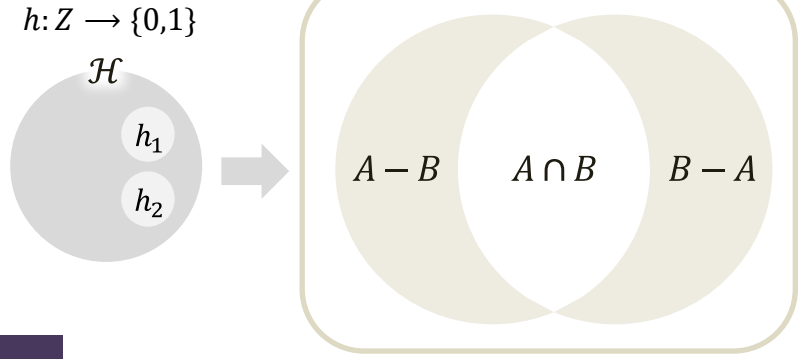
$$d_{\mathcal{H}}(P_S, P_T) = 2 \sup_{h \in \mathcal{H}} \left| \Pr_{P_S}[h'(\mathcal{R}(x)) = 1] - \Pr_{P_T}[h'(\mathcal{R}(x)) = 1] \right|$$

Recap : Empirical $\mathcal{H}$-Divergence[3]

$$\hat{d}_{\mathcal{H}}(\tilde{Z}_S, \tilde{Z}_T) = 2\left(1 - \min_{h \in \mathcal{H}} \left[\mathbb{E}_{z \sim \tilde{P}_S} I[h'(z) = 1] + \mathbb{E}_{z \sim \tilde{P}_T} I[h'(z) = 0] \right]\right)$$

$A = \{z | h_1(z) = 1\}$ $B = \{z | h_2(z) = 1\}$

$h: Z \longrightarrow \{0,1\}$



$A \cap B = \{z | h_1(z) = h_2(z) = 1\}$

$A - B = \{z | h_1(z) = 1 \, and \, h_2(z) = 0\}$

$B - A = \{z | h_1(z) = 0 \, and \, h_2(z) = 1\}$

$(A \cup B)^c = \{z | h_1(z) = h_2(z) = 0\}$

Symmetric Difference Hypothesis Space $\mathcal{H}\Delta\mathcal{H}$[4]

Definition 3, Ben-David et al., 2010

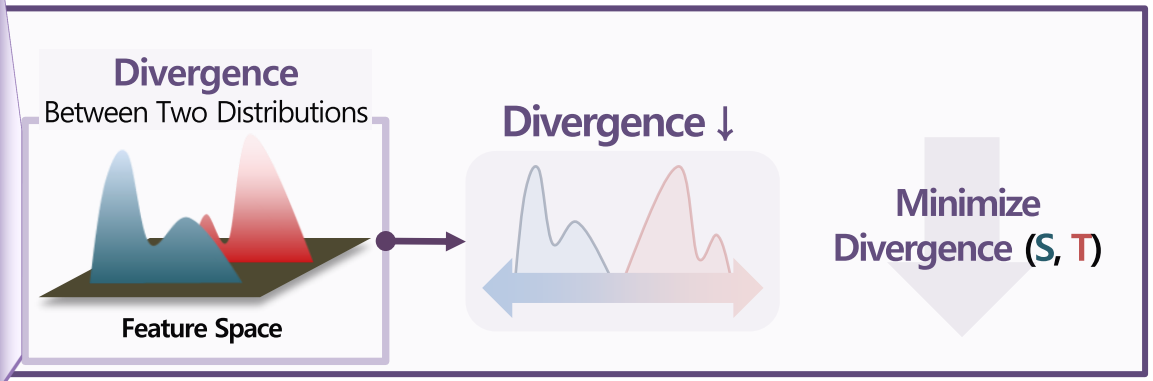
For a hypothesis space $\mathcal{H}$, the *symmetric difference hypothesis space* $\mathcal{H}\Delta\mathcal{H}$ is the set of hypotheses

$g \in \mathcal{H}\Delta\mathcal{H} \iff g(z) = h_1(z) \oplus h_2(z)$ for some $h_1, h_2 \in \mathcal{H}$,

where $\oplus$ is the XOR function. In words, every hypothesis $g \in \mathcal{H}\Delta\mathcal{H}$ is the set of disagreements between two hypotheses h_1 and h_2 in $\mathcal{H}$.

④ $\mathcal{H}\Delta\mathcal{H}$-Divergence

Very useful in reasoning about error!
(When we give a bound on the target error)



[4] Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., & Vaughan, J. W. (2010). A theory of learning from different domains. Machine learning, 79, 151-175..

Symmetric Difference of set A and B

$A \Delta B = (A - B) \cup (B - A)$

Methods

A Theory of Learning From Different Domains, ML, 2010

Recap : $\mathcal{H}$-Divergence

$$d_{\mathcal{H}}(P_S, P_T) = 2 \sup_{h \in \mathcal{H}} \left| \Pr_{P_S}[h'(\mathcal{R}(x)) = 1] - \Pr_{P_T}[h'(\mathcal{R}(x)) = 1] \right|$$

Recap : Empirical $\mathcal{H}$-Divergence[3]

$$\hat{d}_{\mathcal{H}}(\tilde{Z}_S, \tilde{Z}_T) = 2\left(1 - \min_{h \in \mathcal{H}} \left[\mathbb{E}_{z \sim \tilde{P}_S} I[h'(z) = 1] + \mathbb{E}_{z \sim \tilde{P}_T} I[h'(z) = 0] \right]\right)$$

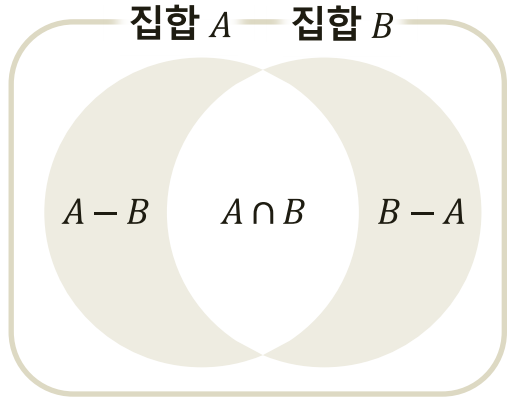
$A = \{z | h_1(z) = 1\}$ $B = \{z | h_2(z) = 1\}$

$h: Z \longrightarrow \{0,1\}$

$\mathcal{H}$

h_1

h_2



$A \cap B = \{z | h_1(z) = h_2(z) = 1\}$

$A - B = \{z | h_1(z) = 1 \, and \, h_2(z) = 0\}$

$B - A = \{z | h_1(z) = 0 \, and \, h_2(z) = 1\}$

$(A \cup B)^c = \{z | h_1(z) = h_2(z) = 0\}$

Symmetric Difference Hypothesis Space $\mathcal{H}\Delta\mathcal{H}$[4]

Definition 3, Ben-David et al., 2010

For a hypothesis space $\mathcal{H}$, the *symmetric difference hypothesis space* $\mathcal{H}\Delta\mathcal{H}$ is the set of hypotheses

$g \in \mathcal{H}\Delta\mathcal{H} \iff g(z) = h_1(z) \oplus h_2(z)$ for some $h_1, h_2 \in \mathcal{H}$,

where $\oplus$ is the XOR function. In words, every hypothesis $g \in \mathcal{H}\Delta\mathcal{H}$ is the set of disagreements between two hypotheses h_1 and h_2 in $\mathcal{H}$.

④ $\mathcal{H}\Delta\mathcal{H}$-Divergence

Very useful in reasoning about error!
(When we give a bound on the target error)

$$g(z) = \begin{cases} 0, & h_1(z) = h_2(z) \quad z \in (A \triangle B)^c = (A \cap B) \cup (A \cup B)^c \\ 1, & h_1(z) \neq h_2(z) \quad z \in A \triangle B = (A - B) \cup (B - A) \end{cases}$$

같은 입력 z에 대해 가설들(h_1, h_2)이 얼마나 다르게 동작하는가?

$g \in \mathcal{H}\Delta\mathcal{H}$

$\mathcal{H}$에 속한 두 가설 h_1와 h_2 의 예측 불일치(disagreements)를 포착하는 가설 g의 집합

[4] Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., & Vaughan, J. W. (2010). A theory of learning from different domains. Machine learning, 79, 151-175..

Methods

A Theory of Learning From Different Domains, ML, 2010

Recap : $\mathcal{H}$-Divergence

$$d_{\mathcal{H}}(P_S, P_T) = 2 \sup_{h' \in \mathcal{H}} \left| \Pr_{P_S}[h'(\mathcal{R}(x)) = 1] - \Pr_{P_T}[h'(\mathcal{R}(x)) = 1] \right|$$

Recap : Empirical $\mathcal{H}$-Divergence[3]

$$\hat{d}_{\mathcal{H}}(\tilde{Z}_S, \tilde{Z}_T) = 2\left(1 - \min_{h' \in \mathcal{H}} \left[\mathbb{E}_{z \sim \tilde{P}_S} I[h'(z) = 1] + \mathbb{E}_{z \sim \tilde{P}_T} I[h'(z) = 0] \right]\right)$$

Definition of $\mathcal{H}\Delta\mathcal{H}$-Distance[4]

Lemma 3, Ben-David et al., 2010

For any hypotheses $h_1, h_2 \in \mathcal{H}$,

$$d_{\mathcal{H}\Delta\mathcal{H}}(\tilde{P}_S, \tilde{P}_T)$$

$$= 2 \sup_{h_1, h_2 \in \mathcal{H}} \left| \Pr_{\tilde{P}_S}[h_1(z) \neq h_2(z)] - \Pr_{\tilde{P}_T}[h_1(z) \neq h_2(z)] \right|$$

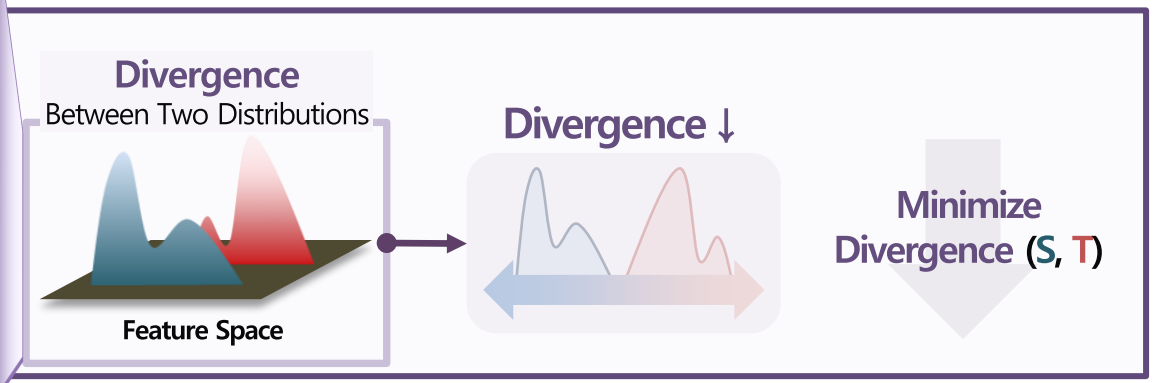
$$= 2 \sup_{h_1, h_2 \in \mathcal{H}} |\epsilon_S(h_1, h_2) - \epsilon_T(h_1, h_2)| \qquad \epsilon(h_1, h_2) = \mathbb{E}_{z \sim \tilde{P}}[|h_1(z) - h_2(z)|]$$

$$\geq 2|\epsilon_S(h_1, h_2) - \epsilon_T(h_1, h_2)|$$

$$\frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(\tilde{P}_S, \tilde{P}_T) \geq |\epsilon_S(h_1, h_2) - \epsilon_T(h_1, h_2)|$$

④ $\mathcal{H}\Delta\mathcal{H}$-Divergence

Very useful in reasoning about error!
(When we give a bound on the target error)



[4] Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., & Vaughan, J. W. (2010). A theory of learning from different domains. Machine learning, 79, 151-175..

Methods

A Theory of Learning From Different Domains, ML, 2010

Recap : $\mathcal{H}$-Divergence

$$d_{\mathcal{H}}(P_S, P_T) = 2 \sup_{h' \in \mathcal{H}} \left| \Pr_{P_S}[h'(\mathcal{R}(x)) = 1] - \Pr_{P_T}[h'(\mathcal{R}(x)) = 1] \right|$$

Recap : Empirical $\mathcal{H}$-Divergence[3]

$$\hat{d}_{\mathcal{H}}(\tilde{Z}_S, \tilde{Z}_T) = 2\left(1 - \min_{h' \in \mathcal{H}} \left[\mathbb{E}_{z \sim \tilde{P}_S} I[h'(z) = 1] + \mathbb{E}_{z \sim \tilde{P}_T} I[h'(z) = 0] \right]\right)$$

Definition of $\mathcal{H}\Delta\mathcal{H}$-Distance[4]

Lemma 3, Ben-David et al., 2010

For any hypotheses $h_1, h_2 \in \mathcal{H}$,

$$d_{\mathcal{H}\triangle\mathcal{H}}(\tilde{P}_S, \tilde{P}_T)$$

$$= 2 \sup_{h_1, h_2 \in \mathcal{H}} \left| \Pr_{\tilde{P}_S}[h_1(z) \neq h_2(z)] - \Pr_{\tilde{P}_T}[h_1(z) \neq h_2(z)] \right|$$

$$= 2 \sup_{h_1, h_2 \in \mathcal{H}} |\epsilon_S(h_1, h_2) - \epsilon_T(h_1, h_2)| \qquad \epsilon(h_1, h_2) = \mathbb{E}_{z \sim \tilde{P}}[|h_1(z) - h_2(z)|]$$

$$\geq 2|\epsilon_S(h_1, h_2) - \epsilon_T(h_1, h_2)|$$

$$\frac{1}{2} d_{\mathcal{H}\triangle\mathcal{H}}(\tilde{P}_S, \tilde{P}_T) \geq |\epsilon_S(h_1, h_2) - \epsilon_T(h_1, h_2)|$$

$$d_{\mathcal{H}\triangle\mathcal{H}}(\tilde{P}_S, \tilde{P}_T) = 2 \sup_{h_1, h_2 \in \mathcal{H}} |\epsilon_S(h_1, h_2) - \epsilon_T(h_1, h_2)|$$

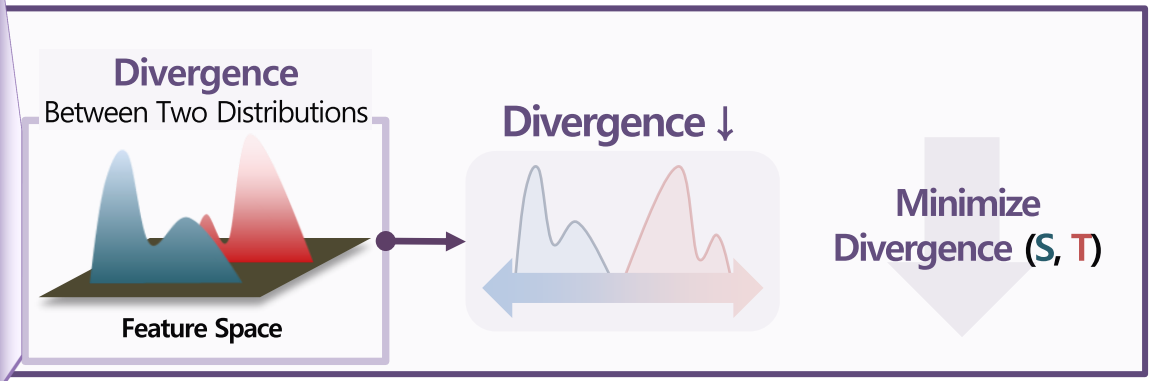
Source와 Target 간에 가설 쌍들이 얼마나 유사하게 동작할까?

$d_{\mathcal{H}\triangle\mathcal{H}}$ Source와 Target에 **무관**하게 가설 쌍들의 동작 방식이 **유사**

$d_{\mathcal{H}\triangle\mathcal{H}}$ Source와 Target **차이**에 따라 가설 쌍들의 동작 방식이 **상이**

④ $\mathcal{H}\Delta\mathcal{H}$-Divergence

Very useful in reasoning about error!
(When we give a bound on the target error)



[4] Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., & Vaughan, J. W. (2010). A theory of learning from different domains. Machine learning, 79, 151-175..

Methods

A Theory of Learning From Different Domains, ML, 2010

Cross-domain Generalization

Target Domain Error ≤ **Source Domain Error** + **Divergence($\tilde{P}_S$, $\tilde{P}_T$)**

- $\mathcal{R}: X \to Z$
- $f: X \to \{0,1\}$ True labeling function
- $\tilde{f}: Z \to \{0,1\}$ Induced image of f under $\mathcal{R}$
- $h: Z \to \{0,1\}$ Hypothesis

모사

$$\epsilon(h, \tilde{f}) = \mathbb{E}_{z\sim\tilde{P}}\left[\left|h(z) - \tilde{f}(z)\right|\right] = \epsilon(h)$$

$\mathcal{H}\Delta\mathcal{H}$-Distance[4]

$$d_{\mathcal{H}\triangle\mathcal{H}}(\tilde{P}_S, \tilde{P}_T) = 2 \sup_{h_1,h_2\in\mathcal{H}} |\epsilon_S(h_1,h_2) - \epsilon_T(h_1,h_2)|$$

$$\frac{1}{2} d_{\mathcal{H}\triangle\mathcal{H}}(\tilde{P}_S, \tilde{P}_T) \geq |\epsilon_S(h_1,h_2) - \epsilon_T(h_1,h_2)|$$

Theorem 2. (Ben-David et al., 2010)

Let $\mathcal{H}$ is a hypothesis class of VC dimension d. With probability at least $1-\delta$, for every $h \in \mathcal{H}$:

$$\epsilon_T(h) \leq \hat{\epsilon}_S(h) + \sqrt{d\log\frac{2em}{d} + \log\frac{4}{\delta}} + \frac{1}{2}\hat{d}_{\mathcal{H}\triangle\mathcal{H}}(\tilde{Z}_S, \tilde{Z}_T) + 4\sqrt{\frac{2d\log(2m)+\log\frac{2}{\delta}}{m}} + \lambda,$$

$$with\ \lambda \geq \inf_{h^*\in\mathcal{H}}[\epsilon_S(h^*) + \epsilon_T(h^*)].$$

Lemma 1, Ben-David et al., 2010

$$d_{\mathcal{H}}(P_S, P_T) \leq \hat{d}_{\mathcal{H}}(\tilde{Z}_S, \tilde{Z}_T) + 4\sqrt{\frac{d\log(2m)+\log\frac{2}{\delta}}{m}}$$

Note:
The VC dimension of $\mathcal{H}\triangle\mathcal{H}$ is at most twice the VC dimension of $\mathcal{H}$.

[4] Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., & Vaughan, J. W. (2010). A theory of learning from different domains. Machine learning, 79, 151-175..

Methods

A Theory of Learning From Different Domains, ML, 2010

Cross-domain Generalization

Target Domain Error ≤ **Source Domain Error** + **Divergence($\tilde{P}_S$, $\tilde{P}_T$)**

Mach Learn (2010) 79: 151–175
DOI 10.1007/s10994-009-5152-4

A theory of learning from different domains

Shai Ben-David · John Blitzer · Koby Crammer · Alex Kulesza · Fernando Pereira · Jennifer Wortman Vaughan

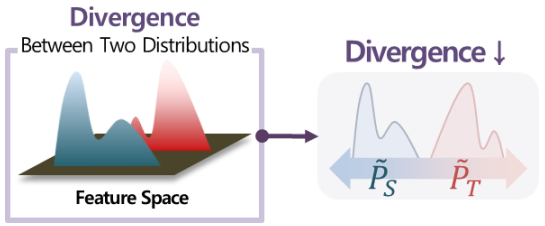
Received: 28 February 2009 / Revised: 12 September 2009 / Accepted: 18 September 2009 /
Published online: 23 October 2009

Abstract Discriminative learning methods for classification perform well when training and test data are drawn from the same distribution. Often, however, we have plentiful labeled training data from a *source* domain but wish to learn a classifier which performs well on a *target* domain with a different distribution and little or no labeled training data. In this work we investigate two questions. First, under what conditions can a classifier trained from source data be expected to perform well on target data? Second, given a small amount of labeled target data, how should we combine it during training with the large amount of labeled source data to achieve the lowest target error at test time?

$\mathcal{H}\Delta\mathcal{H}$-Distance[4]

$$d_{\mathcal{H}\triangle\mathcal{H}}(\tilde{P}_S, \tilde{P}_T) = 2 \sup_{h_1, h_2 \in \mathcal{H}} |\epsilon_S(h_1, h_2) - \epsilon_T(h_1, h_2)|$$

$$\frac{1}{2} d_{\mathcal{H}\triangle\mathcal{H}}(\tilde{P}_S, \tilde{P}_T) \geq |\epsilon_S(h_1, h_2) - \epsilon_T(h_1, h_2)|$$



$\mathcal{H}\Delta\mathcal{H}$-Divergence가 작을 때,
= Representation 차이를 Classifier h_1, h_2가 구분할 수 없을 때,
↓
성공적으로 Domain Adaptation 수행 가능

[4] Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., & Vaughan, J. W. (2010). A theory of learning from different domains. Machine learning, 79, 151-175..

Methods

MCD, CVPR, 2018

❖ **Maximum Classifier Discrepancy for Unsupervised Domain Adaptation** (Saito et al, 2018)[7]

- 선행연구[4]에서 제안했던 $\mathcal{H}\Delta\mathcal{H}$-Divergence 개념을 차용하여 새로운 Discrepancy Loss 제안
- Feature Extractor와 Two Task-Specific Classifiers 간의 적대적 학습 (Adversarial Learning)을 통해 Discrepancy 최소화

CVPR, 24년 3월 기준 2043회 인용

Maximum Classifier Discrepancy for Unsupervised Domain Adaptation

Kuniaki Saito[1], Kohei Watanabe[1], Yoshitaka Ushiku[1], and Tatsuya Harada[1,2]

[1]The University of Tokyo, [2]RIKEN

{k-saito,watanabe,ushiku,harada}@mi.t.u-tokyo.ac.jp

Abstract

In this work, we present a method for unsupervised domain adaptation. Many adversarial learning methods train domain classifier networks to distinguish the features as either a source or target and train a feature generator network to mimic the discriminator. Two problems exist with these methods. First, the domain classifier only tries to distinguish the features as a source or target and thus does not consider task-specific decision boundaries between classes. Therefore, a trained generator can generate ambiguous features near class boundaries. Second, these methods aim to completely match the feature distributions between different domains, which is difficult because of each domain's characteristics.

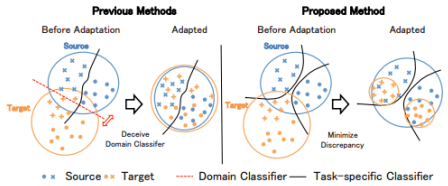
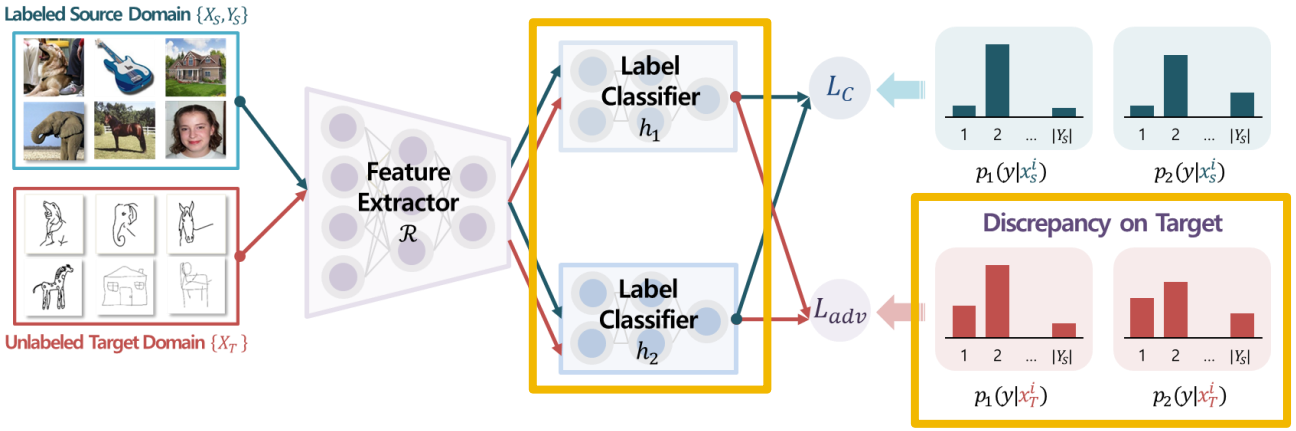


Figure 1. (Best viewed in color.) Comparison of previous and the proposed distribution matching methods.. **Left**: Previous methods try to match different distributions by mimicing the domain classifier. They do not consider the decision boundary. **Right**: Our proposed method attempts to detect target samples outside the support of the source distribution using task-specific classifiers.



2개의 Classifiers를 이용하여 $\mathcal{H}\Delta\mathcal{H}$-Divergence 근사

→ Adversarial Learning 기반 최적화

[7] Saito, K., Watanabe, K., Ushiku, Y., & Harada, T. (2018). Maximum classifier discrepancy for unsupervised domain adaptation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3723-3732).

Methods

MCD, CVPR, 2018

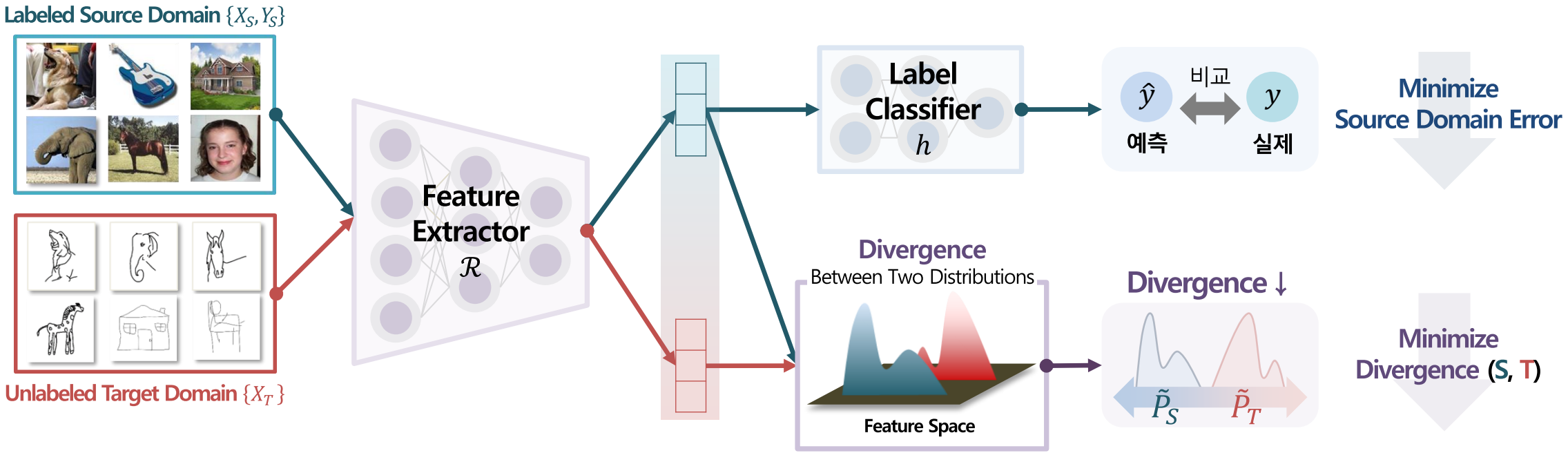
목적 Source Domain에서 학습한 Classifier $h: Z \rightarrow Y$가 Target Domain에서도 잘 동작하게 만들자

방법 Learn a Task-Discriminative and Domain-Invariant Feature Representation!

Recap: Cross-domain Generalization

Target Domain Error ≤ Source Domain Error + $d_{\mathcal{H}\triangle\mathcal{H}}(\tilde{P}_S, \tilde{P}_T) = 2 \sup_{h_1,h_2 \in \mathcal{H}} |\epsilon_S(h_1,h_2) - \epsilon_T(h_1,h_2)|$

Recap : $\mathcal{H}\Delta\mathcal{H}$-Divergence[4]



[4] Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., & Vaughan, J. W. (2010). A theory of learning from different domains. Machine learning, 79, 151-175..

Methods

MCD, CVPR, 2018

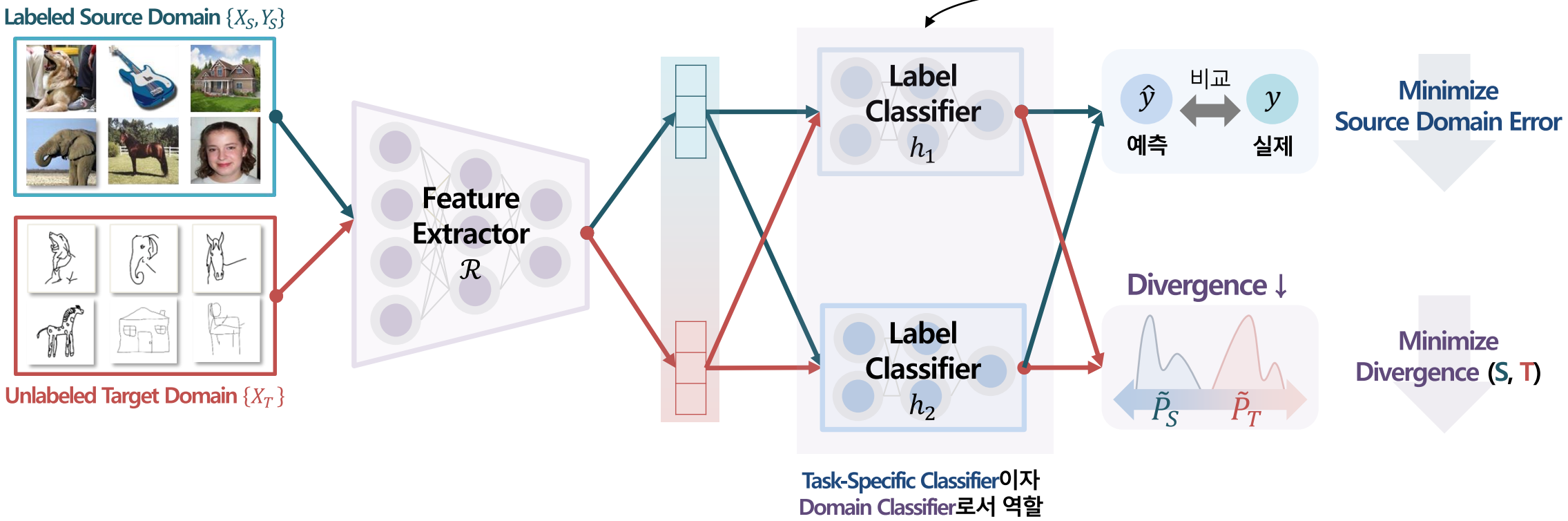
목적 Source Domain에서 학습한 Classifier $h: Z \rightarrow Y$가 Target Domain에서도 잘 동작하게 만들자

방법 Learn a Task-Discriminative and Domain-Invariant Feature Representation!

Recap: Cross-domain Generalization

Target Domain Error ≤ Source Domain Error + $d_{\mathcal{H}\triangle\mathcal{H}}(\tilde{P}_S, \tilde{P}_T) = 2 \sup_{h_1,h_2 \in \mathcal{H}} |\epsilon_S(h_1,h_2) - \epsilon_T(h_1,h_2)|$

Recap : $\mathcal{H}\Delta\mathcal{H}$-Divergence[4]



[4] Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., & Vaughan, J. W. (2010). A theory of learning from different domains. Machine learning, 79, 151-175.
[7] Saito, K., Watanabe, K., Ushiku, Y., & Harada, T. (2018). Maximum classifier discrepancy for unsupervised domain adaptation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3723-3732).

Methods

MCD, CVPR, 2018

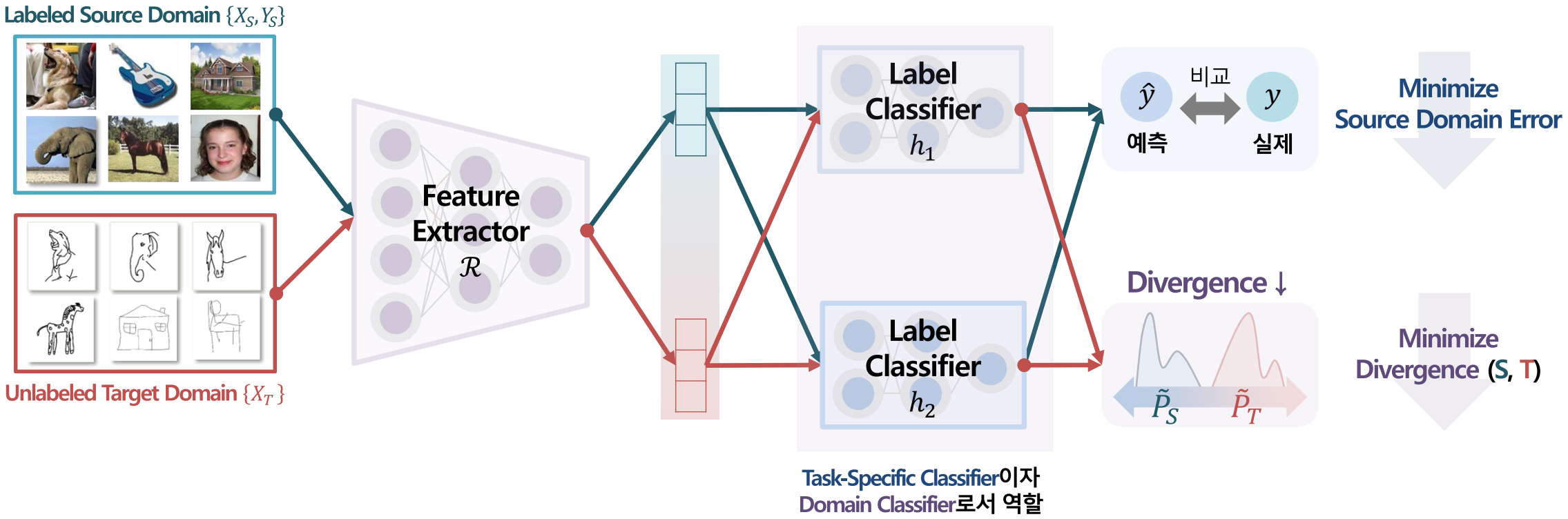
목적 | Source Domain에서 학습한 Classifier $h: Z \rightarrow Y$가 Target Domain에서도 잘 동작하게 만들자

방법 | Learn a Task-Discriminative and Domain-Invariant Feature Representation!

❖ **Maximum Classifier Discrepancy for Unsupervised Domain Adaptation** (Saito et al, 2018)[7]

- 두 개의 Task-Specific Classifiers (h_1, h_2)를 이용하여, $\mathcal{H}\Delta\mathcal{H}$-Divergence[4]에 기반한 새로운 Divergence 척도 제안
- **Discrepancy Loss** : Target Domain의 Feature Representation Z_T가 입력되었을 때, h_1, h_2 출력 값의 절대값 차이로 정의

$$Z_T = \{z_T^i\}_{i=1}^{N_T} \sim \tilde{P}_T$$



[4] Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., & Vaughan, J. W. (2010). A theory of learning from different domains. Machine learning, 79, 151-175.
[7] Saito, K., Watanabe, K., Ushiku, Y., & Harada, T. (2018). Maximum classifier discrepancy for unsupervised domain adaptation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3723-3732).

Methods

MCD, CVPR, 2018

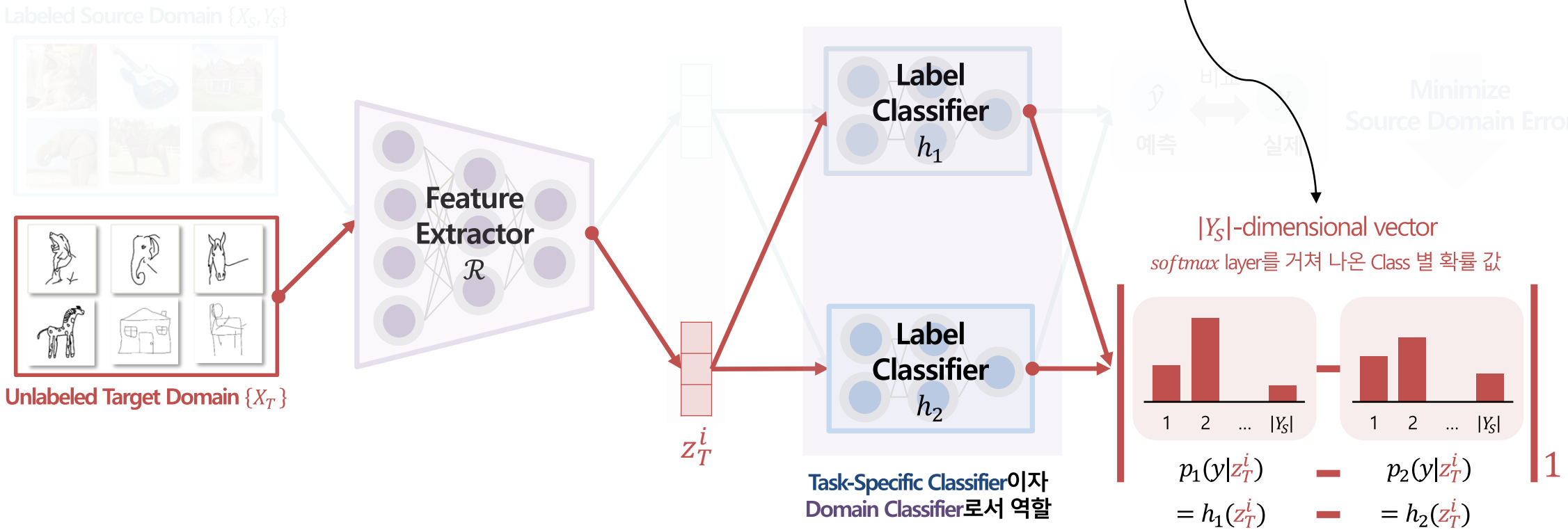
목적 | Source Domain에서 학습한 Classifier $h: Z \rightarrow Y$가 Target Domain에서도 잘 동작하게 만들자

방법 | Learn a Task-Discriminative and Domain-Invariant Feature Representation!

❖ **Maximum Classifier Discrepancy for Unsupervised Domain Adaptation** (Saito et al, 2018)[7]

- 두 개의 Task-Specific Classifiers (h_1, h_2)를 이용하여, $\mathcal{H}\Delta\mathcal{H}$-Divergence[4]에 기반한 새로운 Divergence 척도 제안
- **Discrepancy Loss** : Target Domain의 Feature Representation Z_T가 입력되었을 때, h_1, h_2 출력 값의 절대값 차이로 정의

$Z_T = \{z_T^i\}_{i=1}^{N_T} \sim \tilde{P}_T$



[4] Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., & Vaughan, J. W. (2010). A theory of learning from different domains. Machine learning, 79, 151-175.
[7] Saito, K., Watanabe, K., Ushiku, Y., & Harada, T. (2018). Maximum classifier discrepancy for unsupervised domain adaptation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3723-3732).

Methods

MCD, CVPR, 2018

목적 | Source Domain에서 학습한 Classifier $h: Z \rightarrow Y$가 Target Domain에서도 잘 동작하게 만들자

방법 | Learn a Task-Discriminative and Domain-Invariant Feature Representation!

❖ **Maximum Classifier Discrepancy for Unsupervised Domain Adaptation** (Saito et al, 2018)[7]

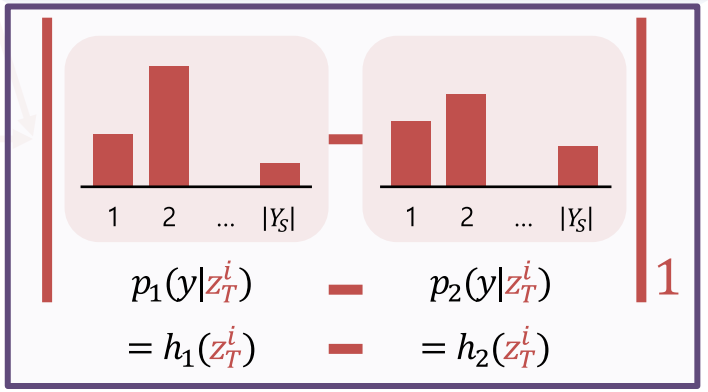
- 두 개의 Task-Specific Classifiers (h_1, h_2)를 이용하여, $\mathcal{H}\Delta\mathcal{H}$-Divergence[4]에 기반한 새로운 Divergence 척도 제안
- **Discrepancy Loss** : Target Domain의 Feature Representation Z_T가 입력되었을 때, h_1, h_2 출력 값의 절대값 차이로 정의

Discrepancy Loss[7]

Eq. (1), Saito et al., 2018

The absolute values of the difference between the two classifiers' probabilistic outputs:

$$d(p_1, p_2) = \left\| p_1(y|z_T^i) - p_2(y|z_T^i) \right\|_1$$



[4] Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., & Vaughan, J. W. (2010). A theory of learning from different domains. Machine learning, 79, 151-175.
[7] Saito, K., Watanabe, K., Ushiku, Y., & Harada, T. (2018). Maximum classifier discrepancy for unsupervised domain adaptation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3723-3732).

Methods

MCD, CVPR, 2018

목적 | Source Domain에서 학습한 Classifier $h: Z \rightarrow Y$가 Target Domain에서도 잘 동작하게 만들자

방법 | Learn a Task-Discriminative and Domain-Invariant Feature Representation!

❖ **Maximum Classifier Discrepancy for Unsupervised Domain Adaptation** (Saito et al, 2018)[7]

- 두 개의 Task-Specific Classifiers (h_1, h_2)를 이용하여, $\mathcal{H}\Delta\mathcal{H}$-Divergence[4]에 기반한 새로운 Divergence 척도 제안
- **Discrepancy Loss** : Target Domain의 Feature Representation Z_T가 입력되었을 때, h_1, h_2 출력 값의 절대값 차이로 정의

Recap : $\mathcal{H}\Delta\mathcal{H}$-Divergence[4]

$$d_{\mathcal{H}\Delta\mathcal{H}}(\tilde{P}_S, \tilde{P}_T) = 2 \sup_{h_1, h_2 \in \mathcal{H}} |\epsilon_S(h_1, h_2) - \epsilon_T(h_1, h_2)|$$

$$\sup_{h_1, h_2 \in \mathcal{H}} \mathbb{E}_{Z_T \sim \tilde{P}_T}[|h_1(Z_T) - h_2(Z_T)|]$$

Discrepancy Loss[7]

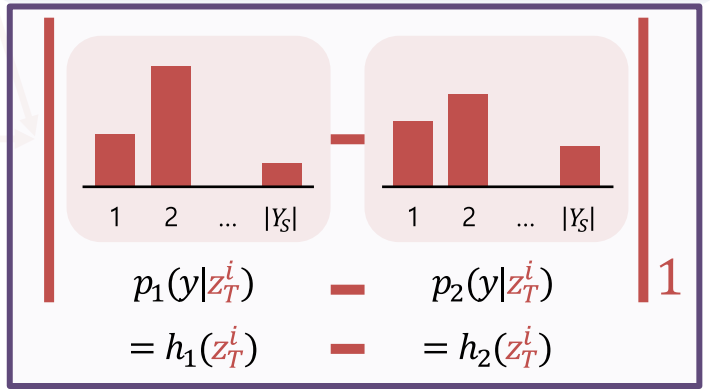
Eq. (1), Saito et al., 2018

The absolute values of the difference between the two classifiers' probabilistic outputs:

$$d(p_1, p_2) = \left\| p_1(y|z_T^i) - p_2(y|z_T^i) \right\|_1$$

Labeled Source Samples에 대해서는 $h_1(z) \neq h_2(z)$ 값이 낮을 것이라 가정

(Assume that h_1 and h_2 can classify source samples correctly[7])



[4] Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., & Vaughan, J. W. (2010). A theory of learning from different domains. Machine learning, 79, 151-175.
[7] Saito, K., Watanabe, K., Ushiku, Y., & Harada, T. (2018). Maximum classifier discrepancy for unsupervised domain adaptation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3723-3732).

Methods

MCD, CVPR, 2018

목적 | Source Domain에서 학습한 Classifier $h: Z \rightarrow Y$가 Target Domain에서도 잘 동작하게 만들자

방법 | Learn a Task-Discriminative and Domain-Invariant Feature Representation!

❖ Maximum Classifier Discrepancy for Unsupervised Domain Adaptation (Saito et al, 2018)[7]

- 두 개의 Task-Specific Classifiers (h_1, h_2)를 이용하여, $\mathcal{H}\Delta\mathcal{H}$-Divergence[4]에 기반한 새로운 Divergence 척도 제안
- **Discrepancy Loss** : Target Domain의 Feature Representation Z_T가 입력되었을 때, h_1, h_2 출력 값의 절대값 차이로 정의

Recap : $\mathcal{H}\Delta\mathcal{H}$-Divergence[4]

$$d_{\mathcal{H}\Delta\mathcal{H}}(\tilde{P}_S, \tilde{P}_T) = 2 \sup_{h_1, h_2 \in \mathcal{H}} |\epsilon_S(h_1, h_2) - \epsilon_T(h_1, h_2)|$$

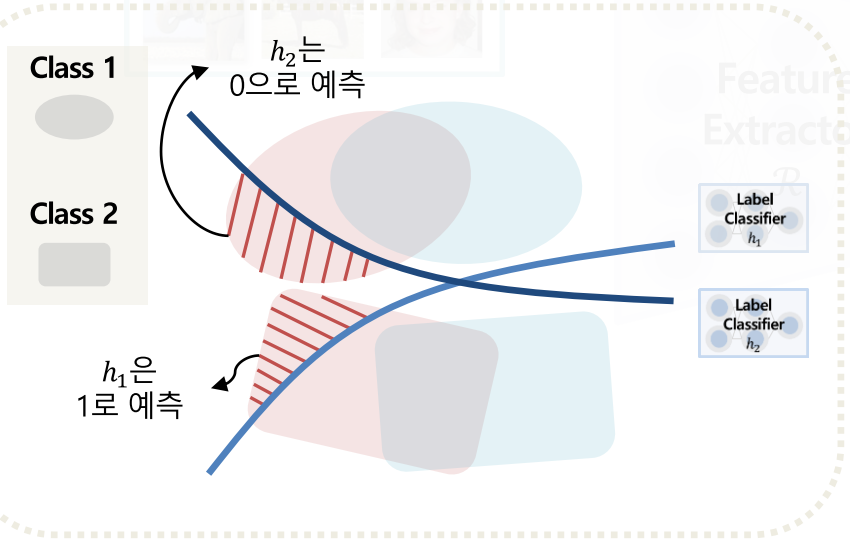
$$\sup_{h_1, h_2 \in \mathcal{H}} \mathbb{E}_{Z_T \sim \tilde{P}_T}[|h_1(Z_T) - h_2(Z_T)|]$$

Discrepancy Loss[7]

Eq. (1), Saito et al., 2018

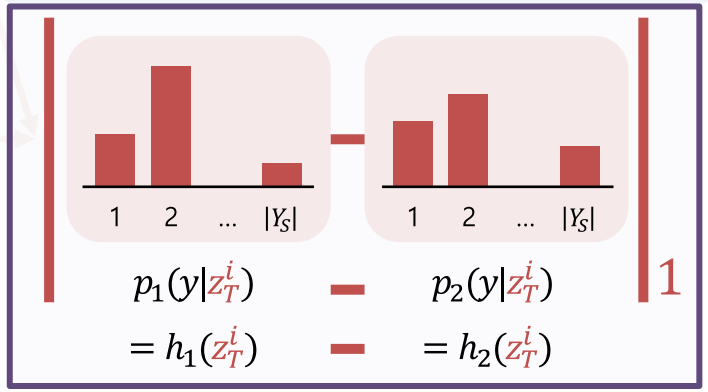
The absolute values of the difference between the two classifiers' probabilistic outputs:

$$d(p_1, p_2) = \| p_1(y|z_T^i) - p_2(y|z_T^i) \|_1$$



Unlabeled Target Samples에 대해서는 $h_1(z) \neq h_2(z)$ 값이 높을 것이라 가정

(Detect target samples misclassified by the h_1 and h_2 learned from source samples[7])



[4] Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., & Vaughan, J. W. (2010). A theory of learning from different domains. Machine learning, 79, 151-175.
[7] Saito, K., Watanabe, K., Ushiku, Y., & Harada, T. (2018). Maximum classifier discrepancy for unsupervised domain adaptation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3723-3732).

Methods

MCD, CVPR, 2018

목적	Source Domain에서 학습한 Classifier $h: Z \rightarrow Y$가 Target Domain에서도 잘 동작하게 만들자
방법	Learn a Task-Discriminative and Domain-Invariant Feature Representation!

❖ Maximum Classifier Discrepancy for Unsupervised Domain Adaptation (Saito et al, 2018)[7]

- 3번의 학습 과정을 통해 최적화 수행
  - **Step 1**. Feature Extractor와 Two Task-Specific Classifiers 모두 Prediction (Source) Loss 최소화

$\mathcal{R}$ h_1, h_2

Recap : $\mathcal{H}\Delta\mathcal{H}$-Divergence[4]

$$d_{\mathcal{H}\Delta\mathcal{H}}(\tilde{P}_S, \tilde{P}_T) = 2 \sup_{h_1,h_2 \in \mathcal{H}} |\epsilon_S(h_1,h_2) - \epsilon_T(h_1,h_2)|$$

$$\sup_{h_1,h_2 \in \mathcal{H}} \mathbb{E}_{z_T \sim \tilde{P}_T}[|h_1(z_T) - h_2(z_T)|]$$

Discrepancy Loss[7]

Eq. (1), Saito et al., 2018

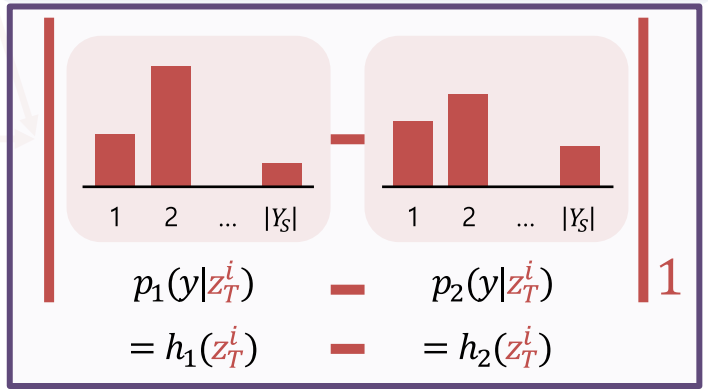
The absolute values of the difference between the two classifiers' probabilistic outputs:

$$d(p_1, p_2) = \left\| p_1(y|z_T^i) - p_2(y|z_T^i) \right\|_1$$

Step 1. $\min_{\mathcal{R},h_1,h_2} \hat{\epsilon}_S(h)$

Minimize Prediction Loss

→ To make classifiers and feature extractor obtain task-discriminative features!



[4] Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., & Vaughan, J. W. (2010). A theory of learning from different domains. Machine learning, 79, 151-175.
[7] Saito, K., Watanabe, K., Ushiku, Y., & Harada, T. (2018). Maximum classifier discrepancy for unsupervised domain adaptation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3723-3732).

Methods

MCD, CVPR, 2018

목적 | Source Domain에서 학습한 Classifier $h: Z \rightarrow Y$가 Target Domain에서도 잘 동작하게 만들자

방법 | Learn a Task-Discriminative and Domain-Invariant Feature Representation!

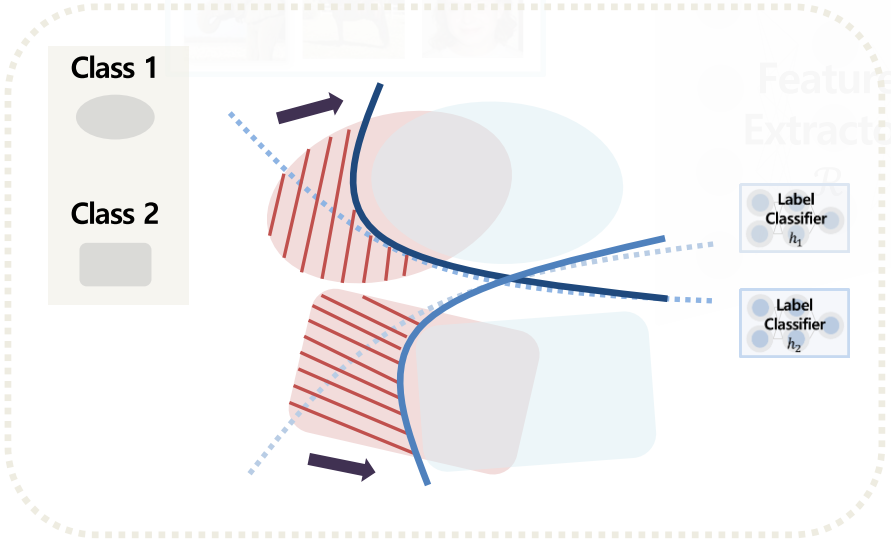
❖ Maximum Classifier Discrepancy for Unsupervised Domain Adaptation (Saito et al, 2018)[7]

- 3번의 학습 과정을 통해 최적화 수행
  - **Step 2**. Two Task-Specific Classifiers는 Discrepancy Loss **최대화** h_1, h_2

Recap : $\mathcal{H}\Delta\mathcal{H}$-Divergence[4]

$$d_{\mathcal{H}\Delta\mathcal{H}}(\tilde{P}_S, \tilde{P}_T) = 2 \sup_{h_1, h_2 \in \mathcal{H}} |\epsilon_S(h_1, h_2) - \epsilon_T(h_1, h_2)|$$

$$\sup_{h_1, h_2 \in \mathcal{H}} \mathbb{E}_{z_T \sim \tilde{P}_T}[|h_1(z_T) - h_2(z_T)|]$$



Step 2. $\min_{h_1, h_2} \hat{\epsilon}_S(h) - \hat{\epsilon}_T(h_1, h_2)$

Minimize Prediction Loss
+ Maximize Discrepancy Loss
→ To make classifiers detect the target samples far from the support of the source!

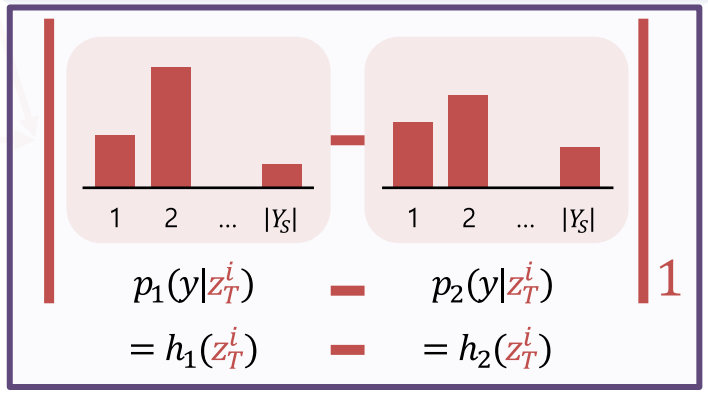
이 단계에서 Feature Extractor는 파라미터 업데이트 X

Discrepancy Loss[7]

Eq. (1), Saito et al., 2018

The absolute values of the difference between the two classifiers' probabilistic outputs:

$$d(p_1, p_2) = \left\| p_1(y|z_T^i) - p_2(y|z_T^i) \right\|_1$$



[4] Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., & Vaughan, J. W. (2010). A theory of learning from different domains. Machine learning, 79, 151-175.
[7] Saito, K., Watanabe, K., Ushiku, Y., & Harada, T. (2018). Maximum classifier discrepancy for unsupervised domain adaptation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3723-3732).

Methods

MCD, CVPR, 2018

목적 | Source Domain에서 학습한 Classifier $h: Z \rightarrow Y$가 Target Domain에서도 잘 동작하게 만들자

방법 | Learn a Task-Discriminative and Domain-Invariant Feature Representation!

❖ Maximum Classifier Discrepancy for Unsupervised Domain Adaptation (Saito et al, 2018)[7]

- 3번의 학습 과정을 통해 최적화 수행
  - **Step 3**. Feature Extractor $\mathcal{R}$는 Discrepancy Loss **최소화**

Recap : $\mathcal{H}\Delta\mathcal{H}$-Divergence[4]

$$d_{\mathcal{H}\Delta\mathcal{H}}(\tilde{P}_S, \tilde{P}_T) = 2 \sup_{h_1, h_2 \in \mathcal{H}} |\epsilon_S(h_1, h_2) - \epsilon_T(h_1, h_2)|$$

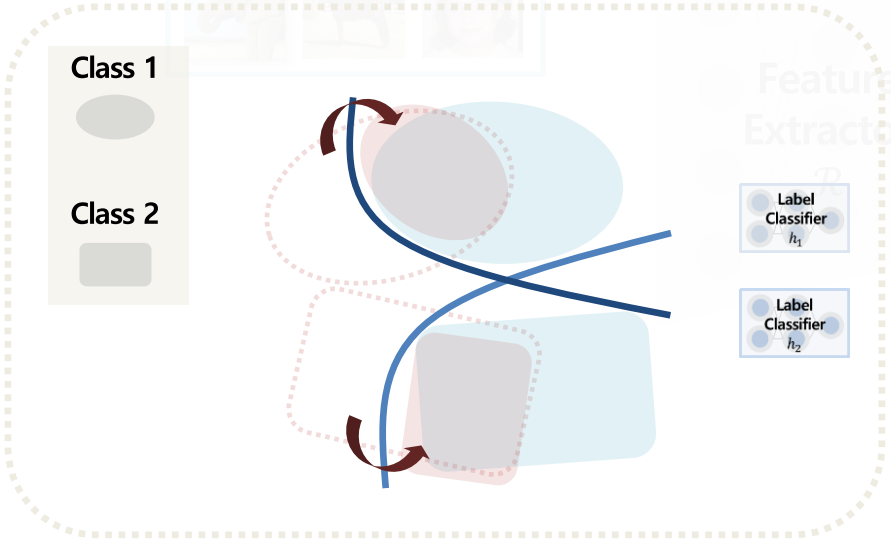
$$\sup_{h_1, h_2 \in \mathcal{H}} \mathbb{E}_{z_T \sim \tilde{P}_T}[|h_1(z_T) - h_2(z_T)|]$$

Discrepancy Loss[7]

Eq. (1), Saito et al., 2018

The absolute values of the difference between the two classifiers' probabilistic outputs:

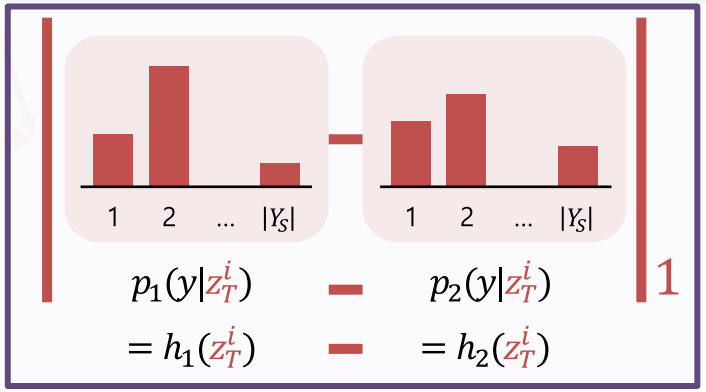
$$d(p_1, p_2) = \left\| p_1(y|z_T^i) - p_2(y|z_T^i) \right\|_1$$



Step 3. $\min_{\mathcal{R}} \hat{\epsilon}_T(h_1, h_2)$

Minimize Discrepancy Loss
→ To make feature extractor obtain domain-invariant feature representation!

(*같은 batch samples에 대해 n번 학습 반복)
이 단계에서 Classifiers는 파라미터 업데이트 X



[4] Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., & Vaughan, J. W. (2010). A theory of learning from different domains. Machine learning, 79, 151-175.
[7] Saito, K., Watanabe, K., Ushiku, Y., & Harada, T. (2018). Maximum classifier discrepancy for unsupervised domain adaptation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3723-3732).

Methods

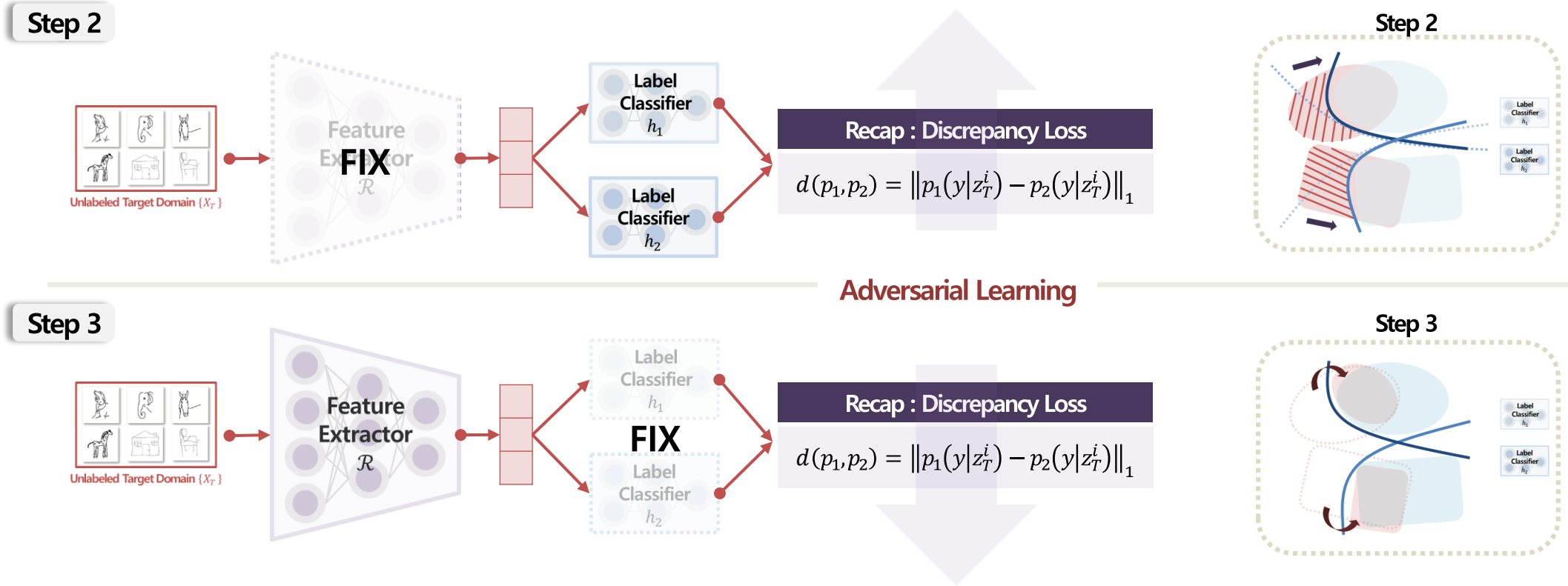
MCD, CVPR, 2018

목적 | Source Domain에서 학습한 Classifier $h: Z \rightarrow Y$가 Target Domain에서도 잘 동작하게 만들자

방법 | Learn a Task-Discriminative and Domain-Invariant Feature Representation!

❖ Maximum Classifier Discrepancy for Unsupervised Domain Adaptation (Saito et al, 2018)[7]

- 3번의 학습 과정을 통해 최적화 수행
  - **Step 2 & 3.** 적대적 학습 (Adversarial Learning)을 통한 Discrepancy Loss **최적화**



[4] Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., & Vaughan, J. W. (2010). A theory of learning from different domains. Machine learning, 79, 151-175.
[7] Saito, K., Watanabe, K., Ushiku, Y., & Harada, T. (2018). Maximum classifier discrepancy for unsupervised domain adaptation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3723-3732).

Methods

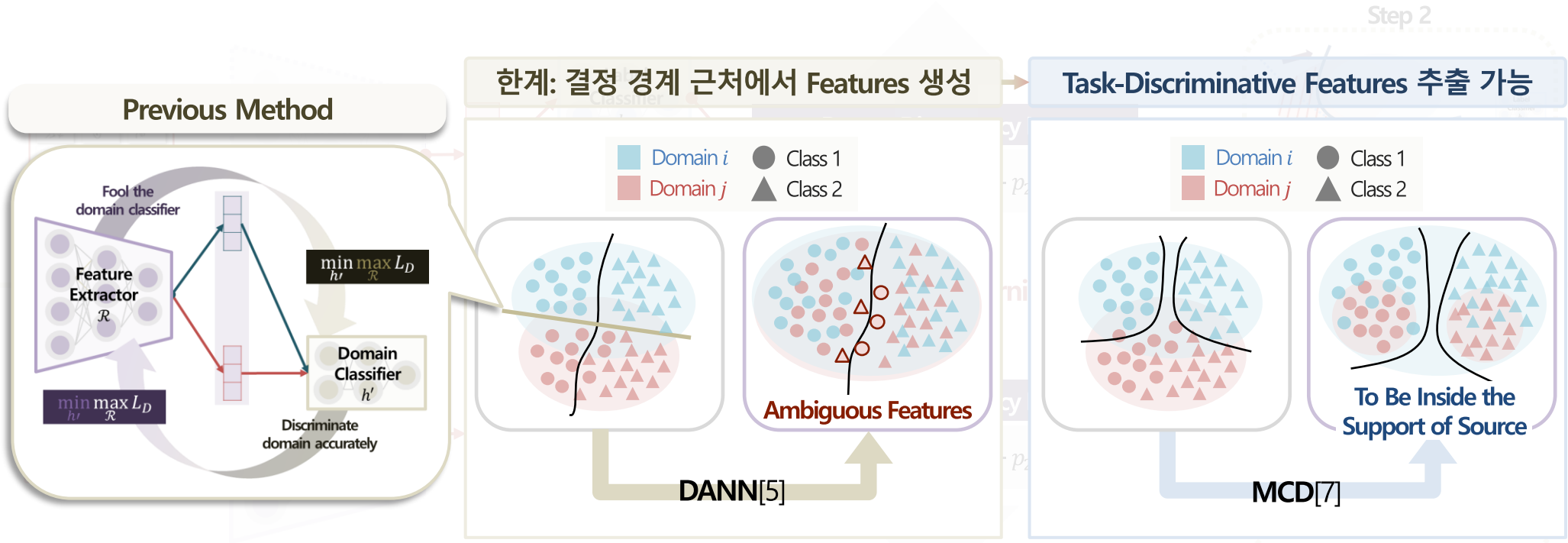
MCD, CVPR, 2018

목적	Source Domain에서 학습한 Classifier $h: Z \rightarrow Y$가 Target Domain에서도 잘 동작하게 만들자
방법	Learn a Task-Discriminative and Domain-Invariant Feature Representation!

❖ Maximum Classifier Discrepancy for Unsupervised Domain Adaptation (Saito et al, 2018)[7]

- 3번의 학습 과정을 통해 최적화 수행
  - **Step 2 & 3.** 적대적 학습 (Adversarial Learning)을 통한 Discrepancy Loss **최적화**

More accurately classify target domain samples!



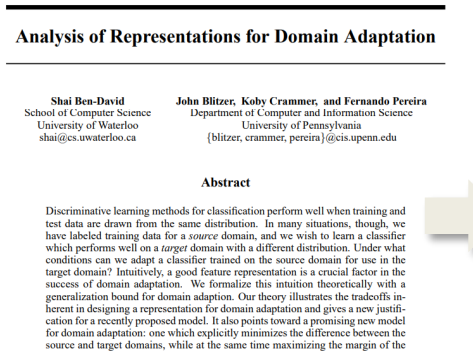
[4] Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., & Vaughan, J. W. (2010). A theory of learning from different domains. Machine learning, 79, 151-175.
[7] Saito, K., Watanabe, K., Ushiku, Y., & Harada, T. (2018). Maximum classifier discrepancy for unsupervised domain adaptation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3723-3732).

Conclusion

목적 Source Domain에서 학습한 Classifier $h: Z \rightarrow Y$가 Target Domain에서도 잘 동작하게 만들자

방법 Task-Discriminative and Domain-Invariant Feature Representation을 이용하자!

Target Domain Error ≤ **Source Domain Error** + **Divergence($\tilde{P}_S$, $\tilde{P}_T$)**



Analysis of Representations for Domain Adaptation

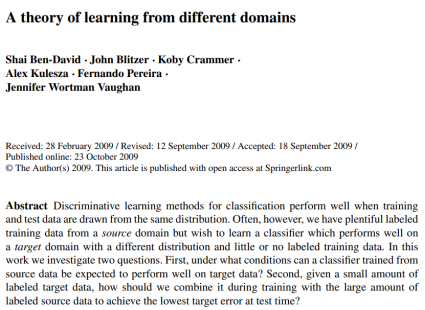
Shai Ben-David
School of Computer Science
University of Waterloo
shai@cs.uwaterloo.ca

John Blitzer, Koby Crammer, and Fernando Pereira
Department of Computer and Information Science
University of Pennsylvania
{blitzer, crammer, pereira}@cis.upenn.edu

Abstract

Discriminative learning methods for classification perform well when training and test data are drawn from the same distribution. In many situations, though, we have labeled training data for a *source* domain, and we wish to learn a classifier which performs well on a *target* domain with a different distribution. Under what conditions can we adapt a classifier trained on the source domain for use in the target domain? Intuitively, a good feature representation is a crucial factor in the success of domain adaptation. We formalize this intuition theoretically with a generalization bound for domain adaption. Our theory illustrates the tradeoffs inherent in designing a representation for domain adaptation and gives a new justification for a recently proposed model. It also points toward a promising new model for domain adaptation: one which explicitly minimizes the difference between the source and target domains, while at the same time maximizing the margin of the

How to measure?

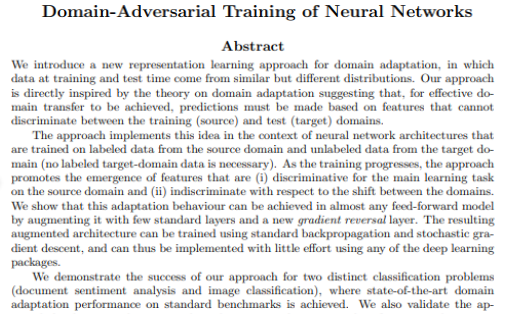


A theory of learning from different domains

Shai Ben-David · John Blitzer · Koby Crammer · Alex Kulesza · Fernando Pereira · Jennifer Wortman Vaughan

Received: 28 February 2009 / Revised: 12 September 2009 / Accepted: 18 September 2009 / Published online: 23 October 2009
© The Author(s) 2009. This article is published with open access at Springerlink.com

Abstract Discriminative learning methods for classification perform well when training and test data are drawn from the same distribution. Often, however, we have plentiful labeled training data from a *source* domain but wish to learn a classifier which performs well on a *target* domain with a different distribution and little or no labeled training data. In this work we investigate two questions. First, under what conditions can a classifier trained from source data be expected to perform well on target data? Second, given a small amount of labeled target data, how should we combine it during training with the large amount of labeled source data to achieve the lowest target error at test time?



Domain-Adversarial Training of Neural Networks

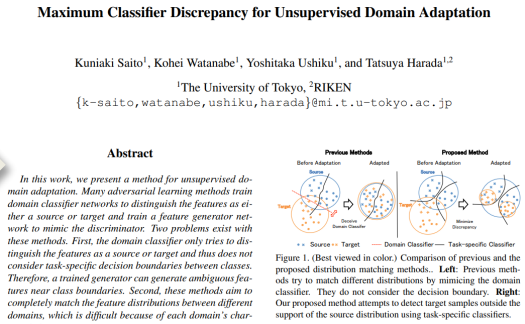
Abstract

We introduce a new representation learning approach for domain adaptation, in which data at training and test time come from similar but different distributions. Our approach is directly inspired by the theory on domain adaptation suggesting that, for effective domain transfer to be achieved, predictions must be made based on features that cannot discriminate between the training (source) and test (target) domains.

The approach implements this idea in the context of neural network architectures that are trained on labeled data from the source domain and unlabeled data from the target domain (no labeled target-domain data is necessary). As the training progresses, the approach promotes the emergence of features that are (i) discriminative for the main learning task on the source domain and (ii) indiscriminate with respect to the shift between the domains. We show that this adaptation behaviour can be achieved in almost any feed-forward model by augmenting it with few standard layers and a new *gradient reversal* layer. The resulting augmented architecture can be trained using standard backpropagation and stochastic gradient descent, and can thus be implemented with little effort using any of the deep learning packages.

We demonstrate the success of our approach for two distinct classification problems (document sentiment analysis and image classification), where state-of-the-art domain adaptation performance on standard benchmarks is achieved. We also validate the ap-

How to optimize?



Maximum Classifier Discrepancy for Unsupervised Domain Adaptation

Kuniaki Saito[1], Kohei Watanabe[1], Yoshitaka Ushiku[1], and Tatsuya Harada[1,2]
[1]The University of Tokyo, [2]RIKEN
{k-saito,watanabe,ushiku,harada}@mi.t.u-tokyo.ac.jp

Abstract

In this work, we present a method for unsupervised domain adaptation. Many adversarial learning methods train domain classifier networks to distinguish the features as either a source or target and train a feature generator network to mimic the discriminator. Two problems exist with these methods. First, the domain classifier only tries to distinguish the features as a source or target and thus does not consider task-specific decision boundaries between classes. Therefore, a trained generator can generate ambiguous features near class boundaries. Second, these methods aim to completely match the feature distributions between different domains, which is difficult because of each domain's characteristics.

Figure 1. (Best viewed in color.) Comparison of previous and the proposed distribution matching methods.. **Left**: Previous methods try to match different distributions by mimicking the domain classifier. They do not consider the decision boundary. **Right**: Our proposed method attempts to detect target samples outside the support of the source distribution using task-specific classifiers.

Empirical $\mathcal{H}$-Divergence[3]

$$\hat{d}_{\mathcal{H}}(\tilde{Z}_S, \tilde{Z}_T) = 2\left(1 - \min_{h' \in \mathcal{H}} \left[\mathbb{E}_{z \sim \tilde{P}_S} I[h'(z) = 1] + \mathbb{E}_{z \sim \tilde{P}_T} I[h'(z) = 0]\right]\right)$$

DANN $\min_{\mathcal{R}} \max_{h'} L_D$[5]

$$\mathbb{E}_{x \sim P_S}[\log h'(\mathcal{R}(x))] + \mathbb{E}_{x \sim P_T}[\log(1 - h'(\mathcal{R}(x)))]$$

$\mathcal{H}\Delta\mathcal{H}$-Divergence[4]

$$d_{\mathcal{H}\Delta\mathcal{H}}(\tilde{P}_S, \tilde{P}_T) = 2 \sup_{h_1, h_2 \in \mathcal{H}} |\epsilon_S(h_1, h_2) - \epsilon_T(h_1, h_2)|$$

Discrepancy Loss[7]

$$d(p_1, p_2) = \left\| p_1(y|z_T^i) - p_2(y|z_T^i) \right\|_1$$

Thank You

Appendix

Method

Analysis of Representations for Domain Adaptation (2006)

목적	Source Domain에서 학습한 Classifier $h: Z \rightarrow Y$가 Target Domain에서도 잘 동작하게 만들자
방법	Learn a Task-Discriminative and Domain-Invariant Feature Representation!

Recap : $\mathcal{H}$-Divergence

$$d_{\mathcal{H}}(P_S, P_T) = 2 \sup_{h' \in \mathcal{H}} \left| \Pr_{P_S}[h'(z) = 1] - \Pr_{P_T}[h'(z) = 1] \right|$$

Empirically Estimated $\mathcal{H}$-Divergence[4]

Lemma 2, Ben-David et al., 2010

(Cf. 원문에는 typo가 있습니다)

For a symmetric hypothesis set $\mathcal{H}$ (one where for every $h' \in \mathcal{H}$, the inverse hypothesis $1 - h'$ is also in $\mathcal{H}$) and $I[\cdot]$ is the binary indicator function,

$$\hat{d}_{\mathcal{H}}(\hat{P}_S, \hat{P}_T) = 2\left(1 - \min_{h' \in \mathcal{H}} \left[\mathbb{E}_{\hat{P}_S} I[h'(z) = 1] + \mathbb{E}_{\hat{P}_T} I[h'(z) = 0]\right]\right).$$

Hypothesis가 symmetric하다?

= hypothesis class $\mathcal{H}$ 내의 모든 hypothesis h에 대하여,
그 inverse hypothesis 1-h 또한 $\mathcal{H}$에 포함된다는 의미.

e.g., Binary classification problem:

- Hypothesis h 는 input z를 0 또는 1로 분류하는 함수
- 만약 어떤 hypothesis h가 특정 input z 를 1로 분류한다면, inverse hypothesis 1- h 는 같은 input z 를 0으로 분류할 것

이처럼 Hypothesis class $\mathcal{H}$가 symmetric하다는 것은, h가 $\mathcal{H}$에 포함되어 있다면 1-h도 반드시 $\mathcal{H}$에 포함되어 있다는 뜻.

linear classifier의 경우를 생각해 보면, 어떤 linear classifier $h(z) = \sin(w^T z)$가 hypothesis class $\mathcal{H}$에 포함되어 있다면, 그 inverse classifier인 $1 - h(z) = \sin(-w^T z)$ 역시 $\mathcal{H}$에 포함되어 있을 때, 이 $\mathcal{H}$를 symmetric하다고 말할 수 있음.

[3] Ben-David, S., Blitzer, J., Crammer, K., & Pereira, F. (2006). Analysis of representations for domain adaptation. Advances in neural information processing systems, 19.
[4] Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., & Vaughan, J. W. (2010). A theory of learning from different domains. Machine learning, 79, 151-175.