Transfer Learning Without Knowing: Reprogramming Black-box Machine Learning Models with Scarce Data and Limited Resources

Yun-Yun Tsai [1] (s107062548@m107.nthu.edu.tw)



Joint work with
Pin-Yu Chen [2], Tsung-Yi Ho [1]

[1] Department of Computer Science, National Tsing Hua University, Taiwan
[2] IBM Research

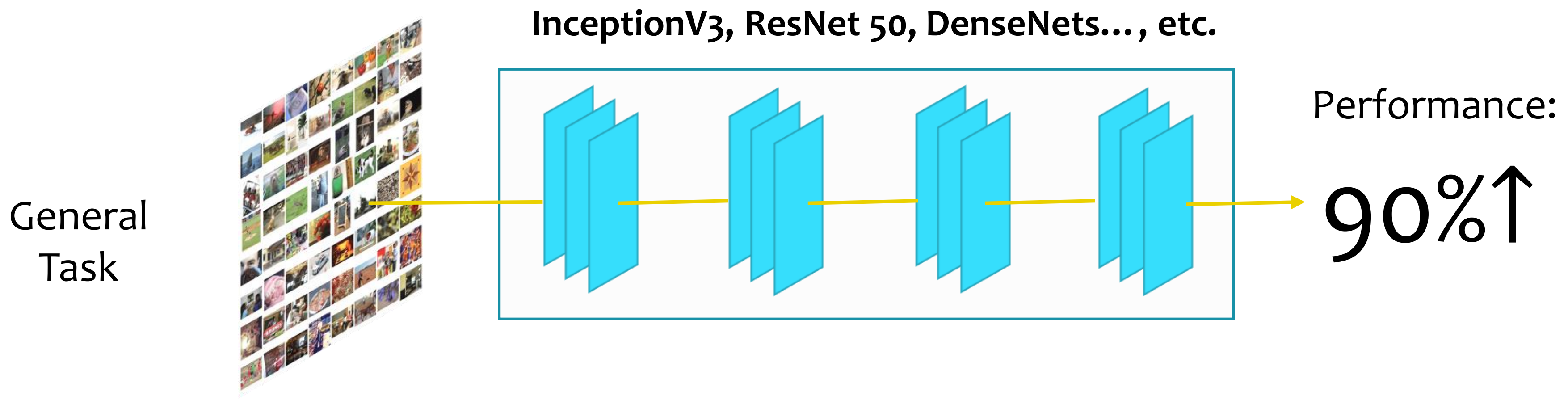
Outline

- Motivation & Main Idea
- Related Works
- Proposed Method
- Evaluation
- Conclusions

Motivation & Main Idea

Motivation

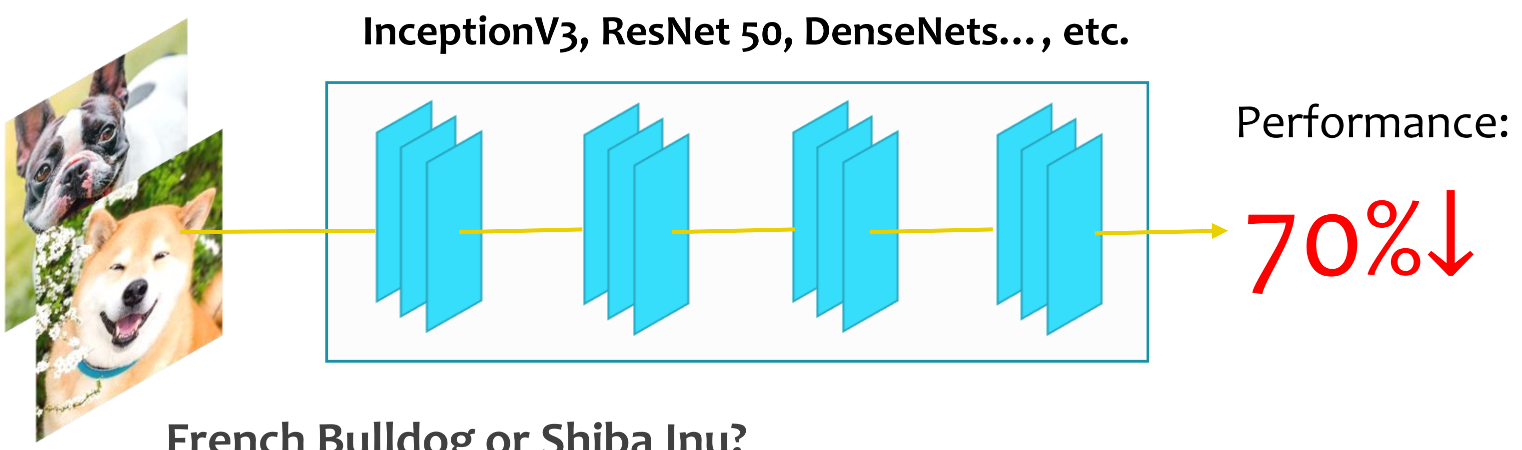
- **General classification task:**
  - MNIST, CIFAR10, Fashion MNIST…, etc.
  - Train from scratch with State-Of-The-Art DNN model.



Motivation

- **General classification task:**
  - MNIST, CIFAR10, Fashion MNIST…, etc.
  - Train from scratch with State-Of-The-Art DNN model.
- **Specific classification task:**
  - Relevance data is **scarce** and **limited**.
  - **Not enough feature** for training large-scale DNN.

Specific Task

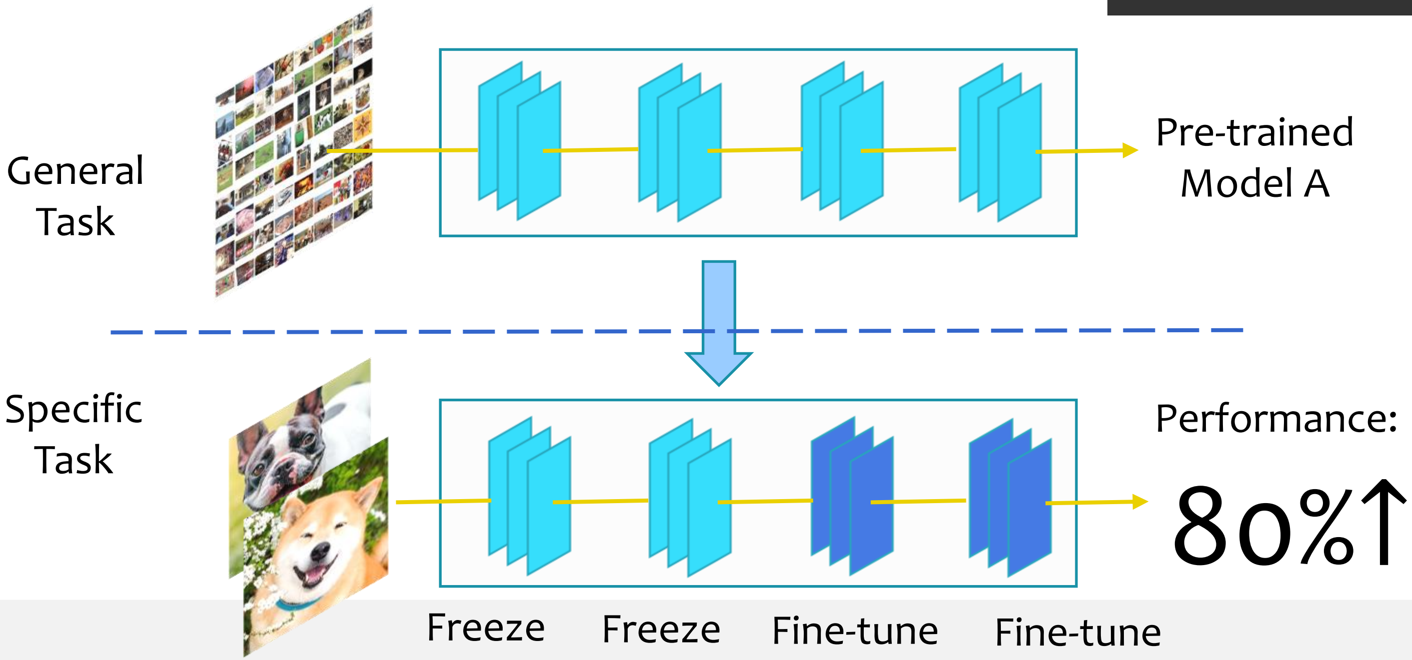


French Bulldog or Shiba Inu?

Motivation

- **Specific classification task:**
  - Relevance data is **scarce** and **limited**.
  - **Not enough feature** for training large-scale DNN.

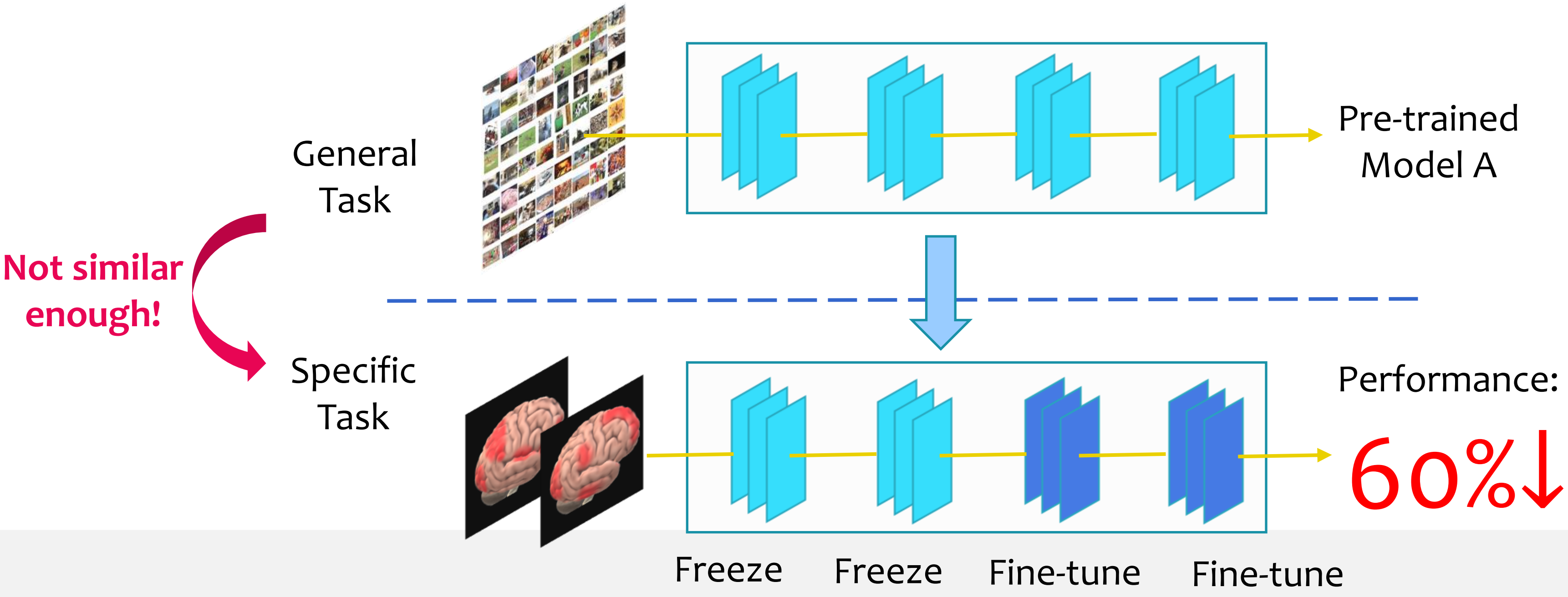
- **General solution: Transfer learning**
  - Sharing pretrained models' parameters.
  - Learning representation faster.



Limitations & Challenges

- **General solution: Transfer learning**
  - Sharing pretrained models' parameters.
  - Learning representation faster.

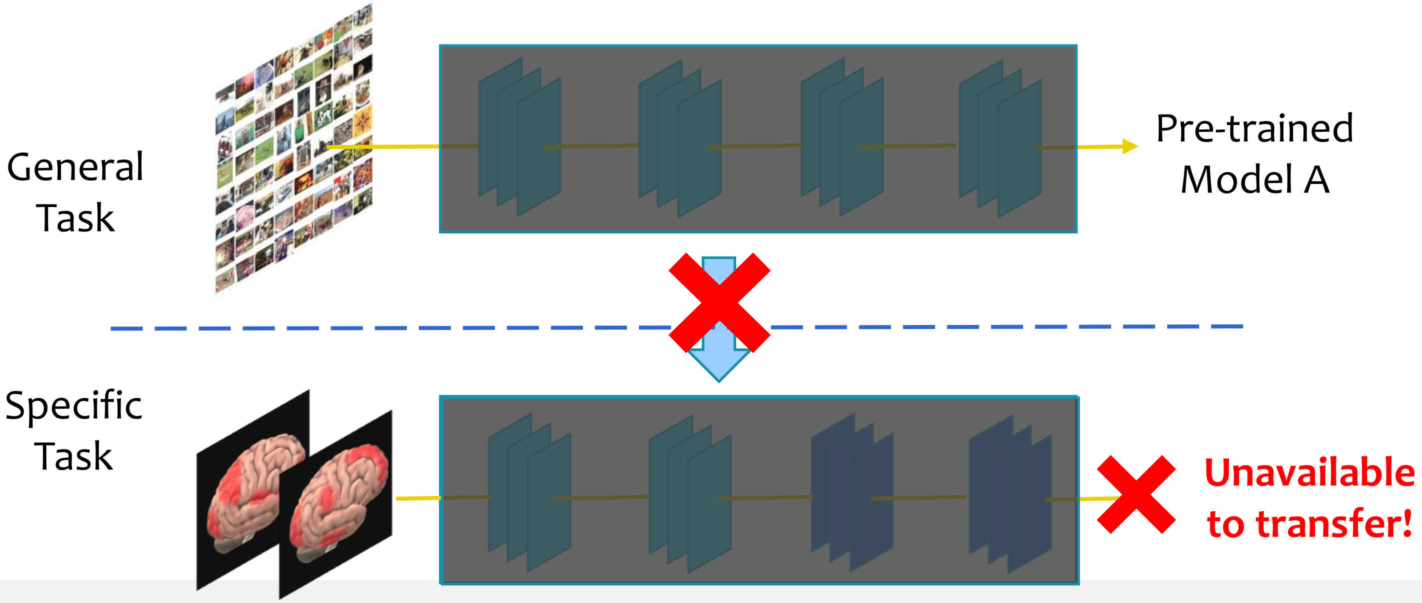
- **Limitations:**
  - **Negative transfer.**
  - **Domain shift should be small.**



Limitations & Challenges

- **General solution: Transfer learning**
  - Sharing pretrained models' parameters.
  - Learning representation faster.

- **Challenges:**
  - It can only transfer by modifying and finetuning a well-known network.
  - **Black-box networks** are unavailable, such as **Google AutoML, Microsoft Custom Vision…**, etc.



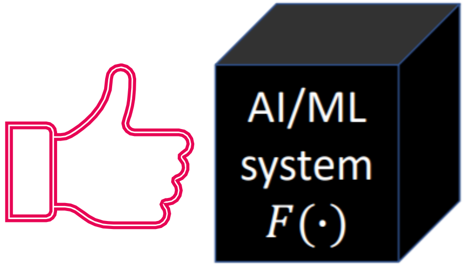
Main Idea

Black-box Adversarial Reprogramming

For those black-box models which have high performance and powerful learning ability, they might have great potential for transfer learning .

Google AutoML
Microsoft Custom Vision

Performance
&
Learning ability



Is it possible to transfer learning?

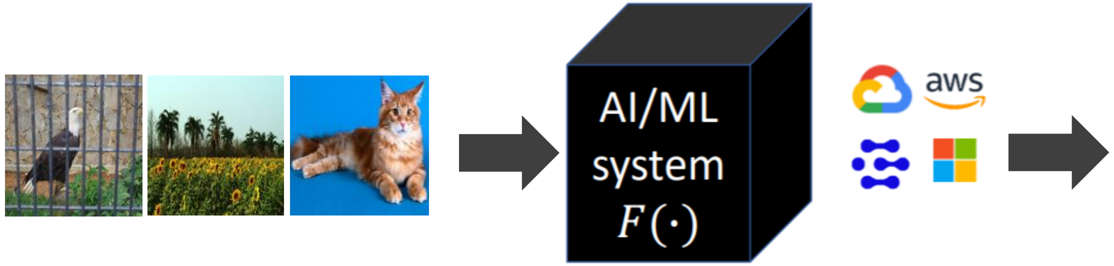
Main Idea

Black-box Adversarial Reprogramming

For those black-box models which have high performance and powerful learning ability, they might have great potential for transfer learning .

- **Black-box Adversarial Reprogramming (BAR):**
  - **Re-purposing black-box DNN** model for different classification tasks.
  - It can **handle domain shift problem** better than transfer learning.

Google AutoML
Microsoft Custom Vision



General Task
(Animal, Vehicle…, etc.)

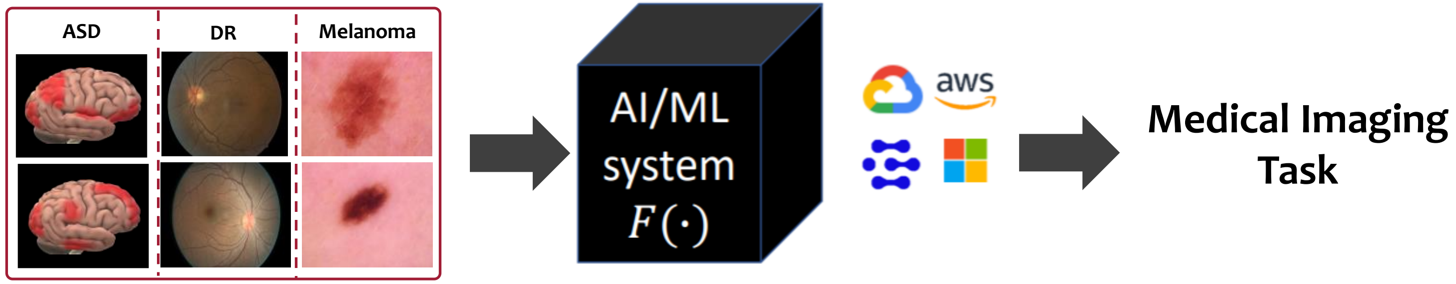
Main Idea

Black-box Adversarial Reprogramming

For those black-box models which have high performance and powerful learning ability, they might have great potential for transfer learning .

- Black-box Adversarial Reprogramming (BAR):
  - **Re-purposing black-box DNN** model for different classification tasks.
  - It can **handle domain shift problem** better than transfer learning.

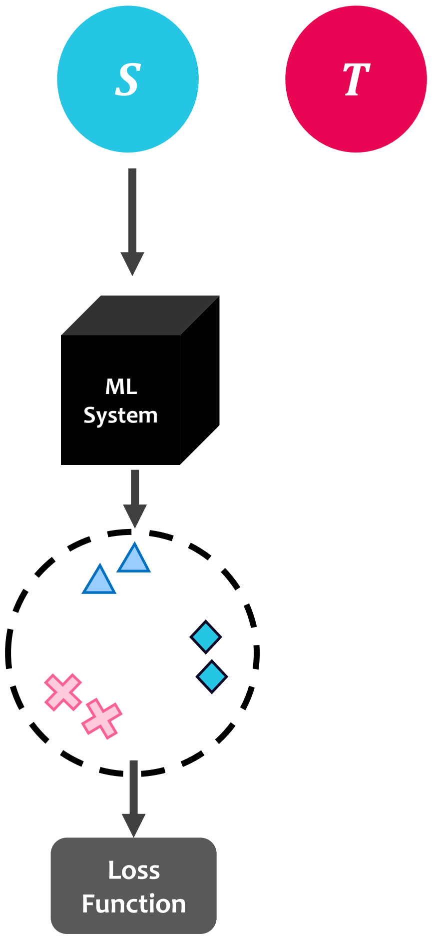
Google AutoML
Microsoft Custom Vision



Related Works

Related Works

- **Idea of Adversarial Reprogramming (AR)**
  - **Adversarial Reprogramming of Neural Networks [Elsayed et al., 2019]**
- Black-box attack methods
  - ZOO: Zeroth Order Optimization Based Black-box Attacks to Deep Neural Networks without Training Substitute Models [Chen et al., 2017]
  - AutoZOOM: Autoencoder-Based Zeroth Order Optimization Method for Attacking Black-Box Neural Networks [Tu et al., 2019]



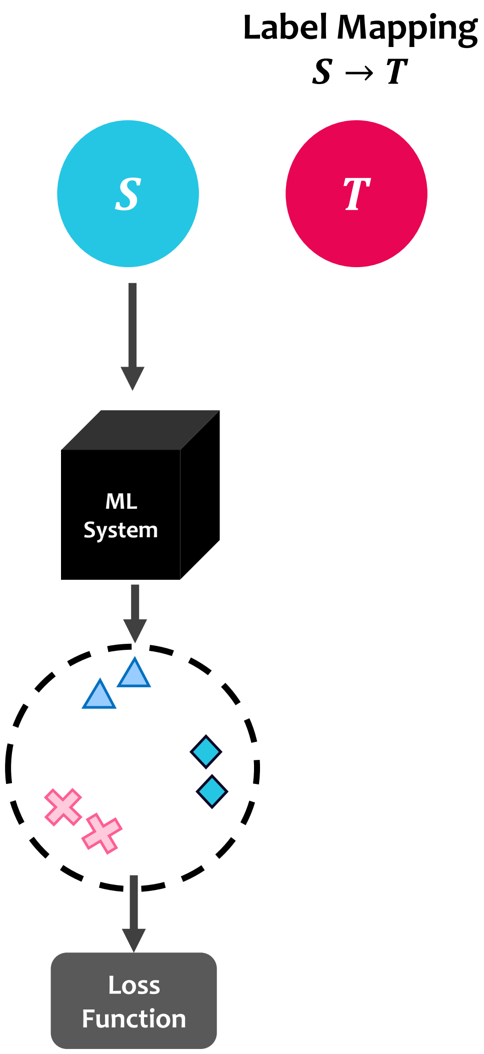
- **Assumptions of AR:**
  1. Cannot modify the target model, and instead must find a universal adversarial perturbation that can be added to all test-time inputs.
  2. Training an adversarial program in **White-box setting**.

Elsayed, G. F., Goodfellow, I., and Sohl-Dickstein, J. Adversarial reprogramming of neural networks. In International Conference on Learning Representations, 2019.

Related Works

- **Idea of Adversarial Reprogramming (AR)**
  - **Adversarial Reprogramming of Neural Networks [Elsayed et al., 2019]**
- **Black-box attack methods**
  - ZOO: Zeroth Order Optimization Based Black-box Attacks to Deep Neural Networks without Training Substitute Models [Chen et al., 2017]
  - AutoZOOM: Autoencoder-Based Zeroth Order Optimization Method for Attacking Black-Box Neural Networks [Tu et al., 2019]

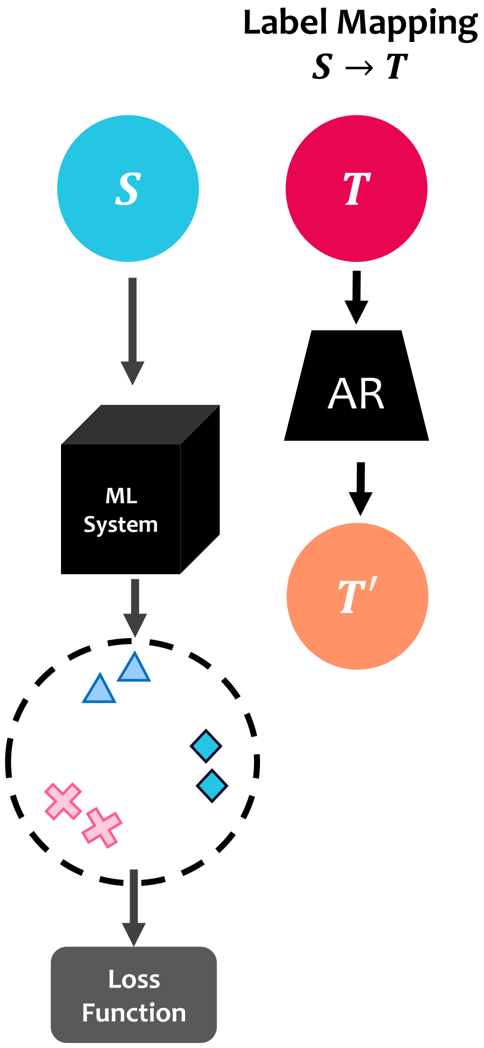
- **Assumptions of AR:**
  1. Cannot modify the target model, and instead must find a universal adversarial perturbation that can be added to all test-time inputs.
  2. Training an adversarial program in **White-box setting**.



Elsayed, G. F., Goodfellow, I., and Sohl-Dickstein, J. Adversarial reprogramming of neural networks. In International Conference on Learning Representations, 2019.

Related Works

- **Idea of Adversarial Reprogramming (AR)**
  - **Adversarial Reprogramming of Neural Networks [Elsayed et al., 2019]**
- **Black-box attack methods**
  - ZOO: Zeroth Order Optimization Based Black-box Attacks to Deep Neural Networks without Training Substitute Models [Chen et al., 2017]
  - AutoZOOM: Autoencoder-Based Zeroth Order Optimization Method for Attacking Black-Box Neural Networks [Tu et al., 2019]

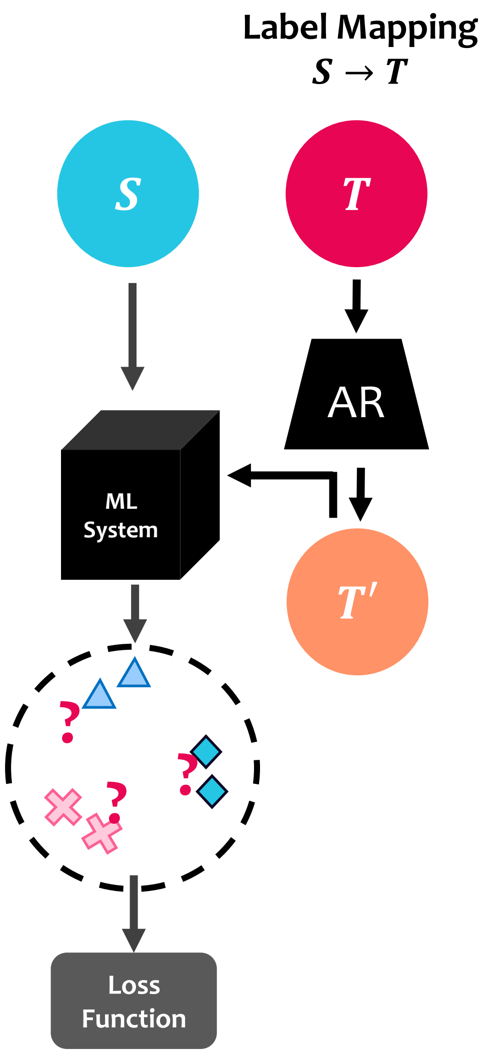


- **Assumptions of AR:**
  1. Cannot modify the target model, and instead must find a universal adversarial perturbation that can be added to all test-time inputs.
  2. Training an adversarial program in **White-box setting**.

Elsayed, G. F., Goodfellow, I., and Sohl-Dickstein, J. Adversarial reprogramming of neural networks. In International Conference on Learning Representations, 2019.

Related Works

- **Idea of Adversarial Reprogramming (AR)**
  - **Adversarial Reprogramming of Neural Networks [Elsayed et al., 2019]**
- **Black-box attack methods**
  - ZOO: Zeroth Order Optimization Based Black-box Attacks to Deep Neural Networks without Training Substitute Models [Chen et al., 2017]
  - AutoZOOM: Autoencoder-Based Zeroth Order Optimization Method for Attacking Black-Box Neural Networks [Tu et al., 2019]

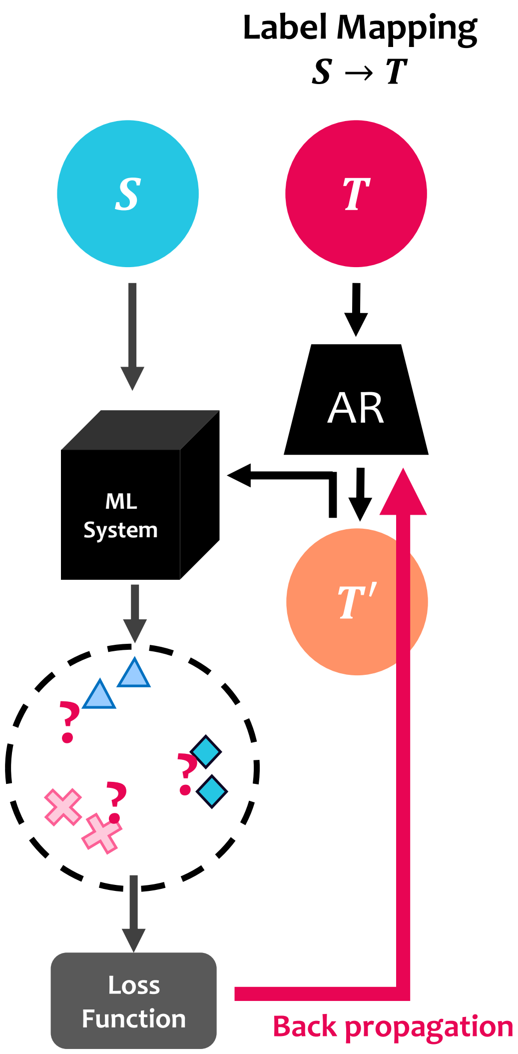


- **Assumptions of AR:**
  1. Cannot modify the target model, and instead must find a universal adversarial perturbation that can be added to all test-time inputs.
  2. Training an adversarial program in **White-box setting**.

Elsayed, G. F., Goodfellow, I., and Sohl-Dickstein, J. Adversarial reprogramming of neural networks. In International Conference on Learning Representations, 2019.

Related Works

- **Idea of Adversarial Reprogramming (AR)**
  - **Adversarial Reprogramming of Neural Networks [Elsayed et al., 2019]**
- **Black-box attack methods**
  - ZOO: Zeroth Order Optimization Based Black-box Attacks to Deep Neural Networks without Training Substitute Models [Chen et al., 2017]
  - AutoZOOM: Autoencoder-Based Zeroth Order Optimization Method for Attacking Black-Box Neural Networks [Tu et al., 2019]

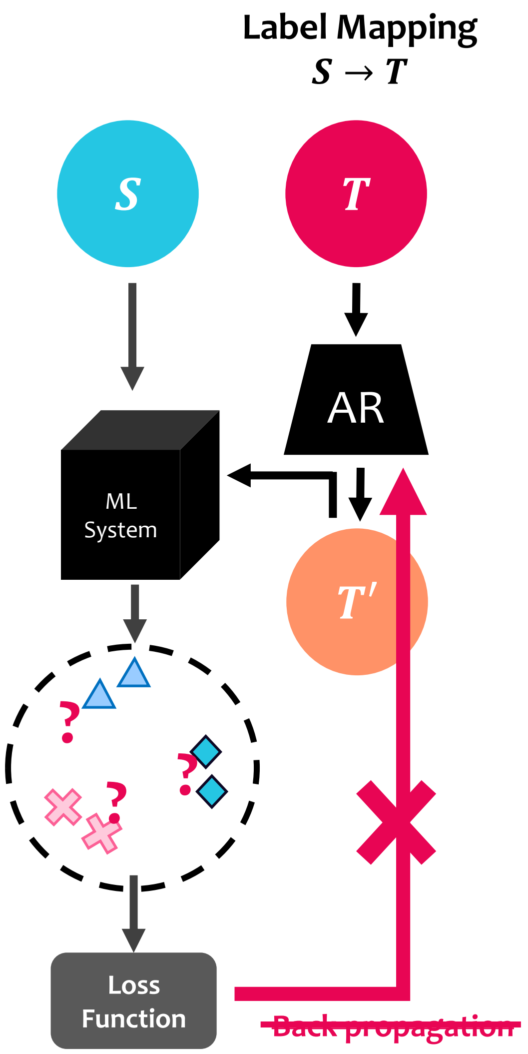


- **Assumptions of AR:**
  1. Cannot modify the target model, and instead must find a universal adversarial perturbation that can be added to all test-time inputs.
  2. Training an adversarial program in **White-box setting**.

Elsayed, G. F., Goodfellow, I., and Sohl-Dickstein, J. Adversarial reprogramming of neural networks. In International Conference on Learning Representations, 2019.

Related Works

- Idea of Adversarial Reprogramming (AR)
  - Adversarial Reprogramming of Neural Networks [Elsayed et al., 2019]
- **Black-box attack methods**
  - **ZOO: Zeroth Order Optimization Based Black-box Attacks to Deep Neural Networks without Training Substitute Models [Chen et al., 2017]**
  - AutoZOOM: Autoencoder-Based Zeroth Order Optimization Method for Attacking Black-Box Neural Networks [Tu et al., 2019]



- Derivative-free optimization method, where only the objective function value $f(x)$ at any input x is needed.
- Suitable for problems where the gradients (or back-propagation) are:
  - Unavailable
  - Un-computable

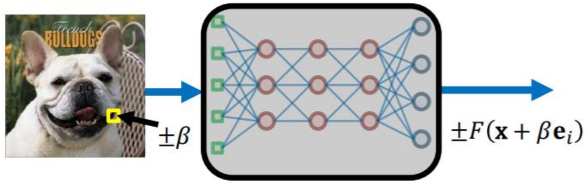
Chen, P.-Y., Zhang, H., Sharma, Y., Yi, J., and Hsieh, C.-J. ZOO: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In ACM Workshop on Artificial Intelligence and Security, pp. 15–26, 2017a.

Related Works

- Idea of Adversarial Reprogramming (AR)
  - Adversarial Reprogramming of Neural Networks [Elsayed et al., 2019]
- **Black-box attack methods**
  - **ZOO: Zeroth Order Optimization Based Black-box Attacks to Deep Neural Networks without Training Substitute Models [Chen et al., 2017]**
  - AutoZOOM: Autoencoder-Based Zeroth Order Optimization Method for Attacking Black-Box Neural Networks [Tu et al., 2019]

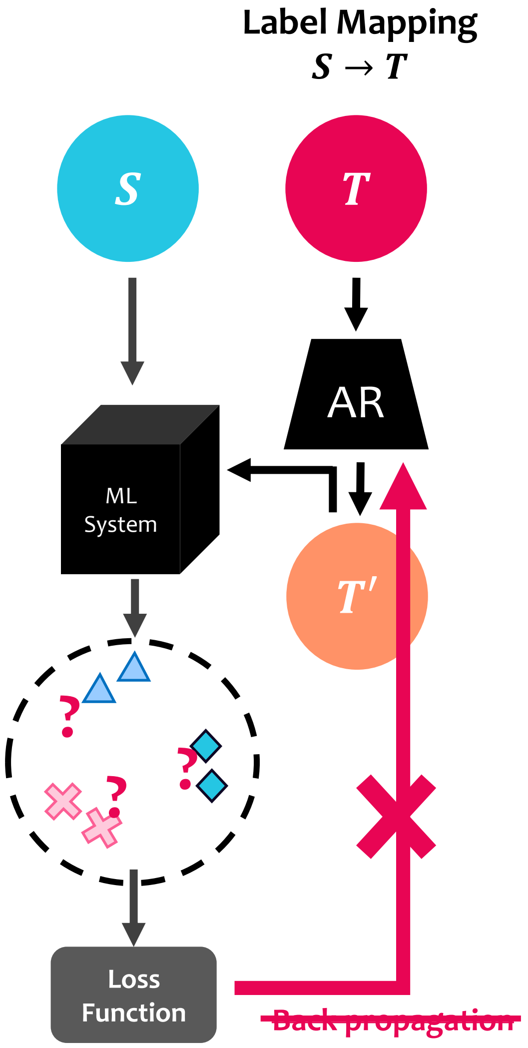
- Introduce a basis vector e_i for every pixel in x,

$$\mathbb{E}\left[\frac{f(x+\varepsilon e_i)-f(x)}{\varepsilon}\right]=\frac{f(x+\varepsilon e_i)-f(x-\varepsilon e_i)}{2\varepsilon}\approx\frac{\partial f(x)}{\partial x}$$



If $x \in \mathbb{R}^p$, it requires $2p$ queries to estimate gradient.

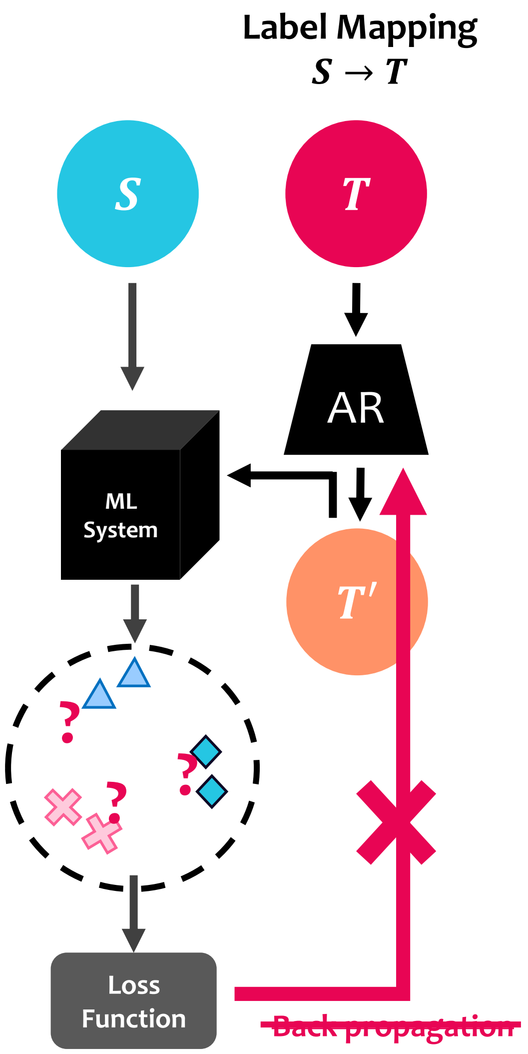
Query inefficient!



Chen, P.-Y., Zhang, H., Sharma, Y., Yi, J., and Hsieh, C.-J. ZOO: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In ACM Workshop on Artificial Intelligence and Security, pp. 15–26, 2017a.

Related Works

- Idea of Adversarial Reprogramming (AR)
  - Adversarial Reprogramming of Neural Networks [Elsayed et al., 2019]
- **Black-box attack methods**
  - ZOO: Zeroth Order Optimization Based Black-box Attacks to Deep Neural Networks without Training Substitute Models [Chen et al., 2017]
  - **AutoZOOM: Autoencoder-Based Zeroth Order Optimization Method for Attacking Black-Box Neural Networks [Tu et al., 2019]**



- Solving query inefficient problem of previous ZOO method.
  1. **An autoencoder (AE)** to learn reconstruction from a dimension-reduced representation.
  2. **An adaptive random gradient estimation strategy** to balance number of queries and distortion.

Tu, C.-C., Ting, P., Chen, P.-Y., Liu, S., Zhang, H., Yi, J., Hsieh, C.-J., and Cheng, S.-M. Autozoom: Autoencoder-based zeroth order optimization method for attacking black-box neural networks. AAAI, 2019.

Proposed Framework

Proposed Framework

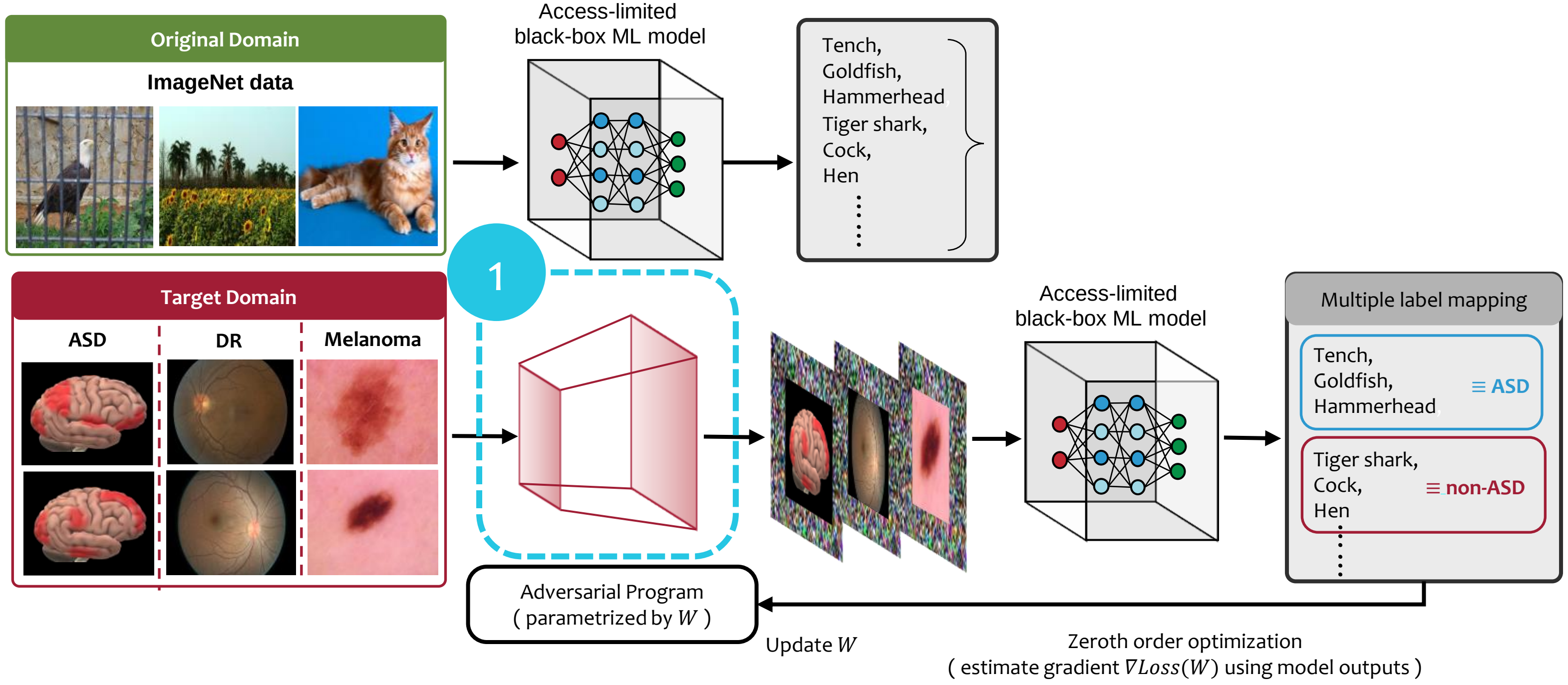
Step:

1. **Generate adversarial example from adversarial program.**

$$\tilde{X}_i = \{T_i\}_{padding} + P, \text{ and}$$

$$P = tanh(\boldsymbol{W} \odot M)$$

Trainable parameters:
$W \in \mathbb{R}^d$



Proposed Framework

Step:

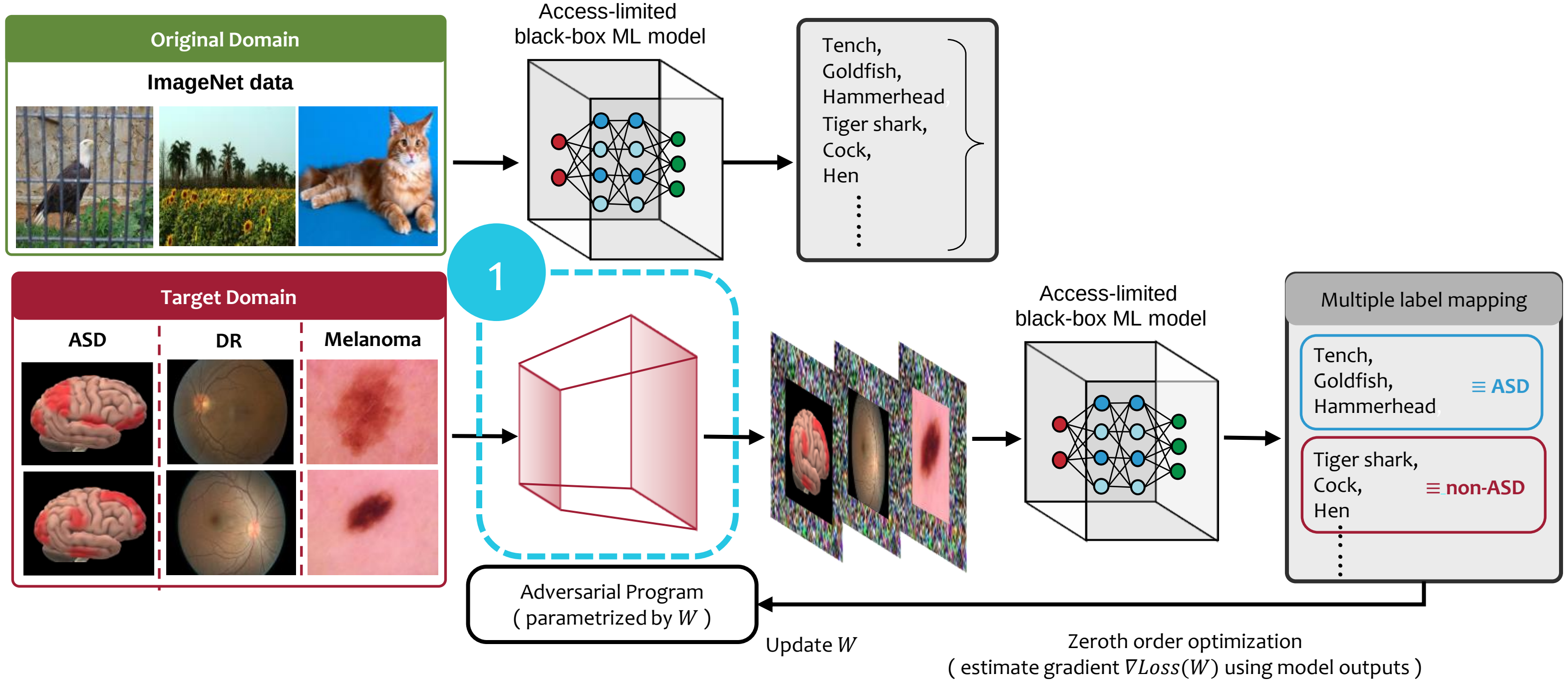
1. **Generate adversarial example from adversarial program.**

$$\tilde{X}_i = \{T_i\}_{padding} + P, \text{ and}$$

$$P = tanh(W \odot \boldsymbol{M})$$

Binary Mask: $M \in \mathbb{R}^d$,

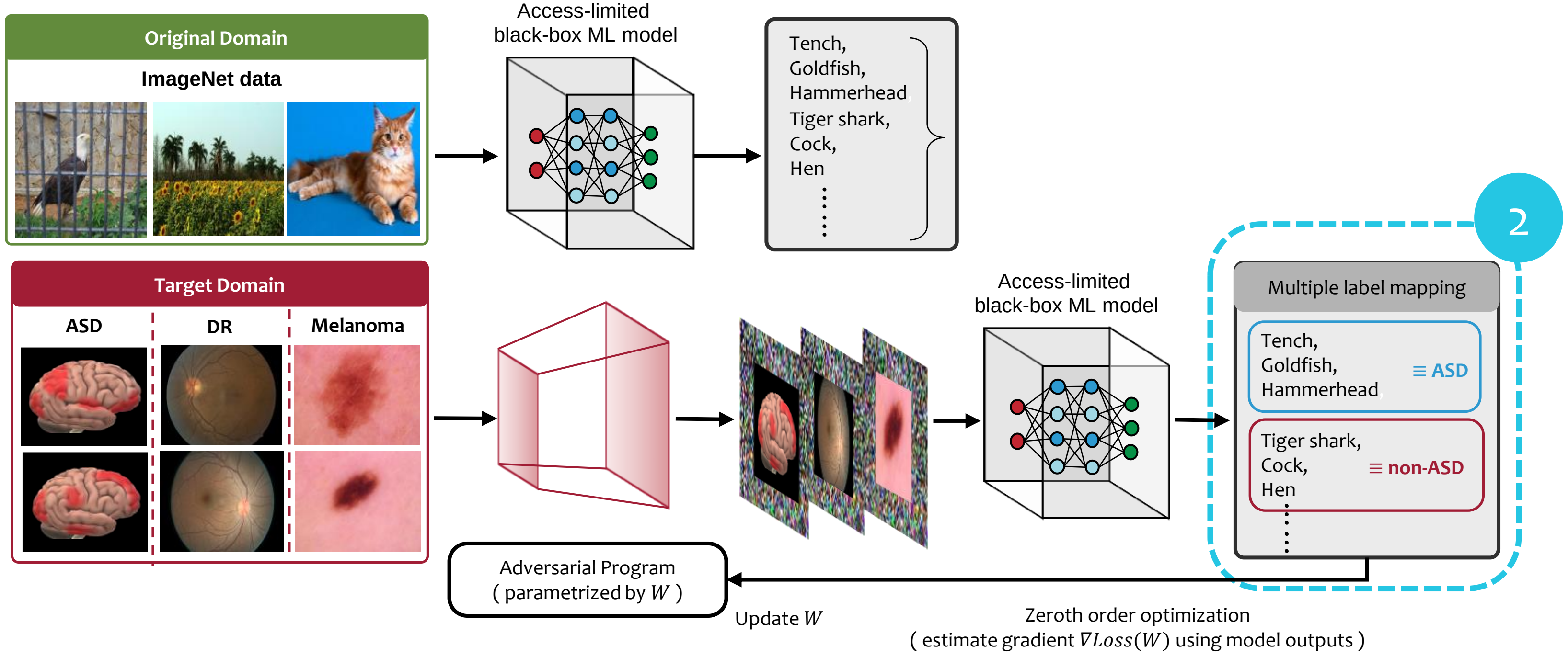
- Note: $M_j = 0$ means the area is used for embedding T_i, otherwise $M_j = 1$.



Proposed Framework

Step:

1. Generate adversarial example from adversarial program.
2. **Multi-map source label to target.**



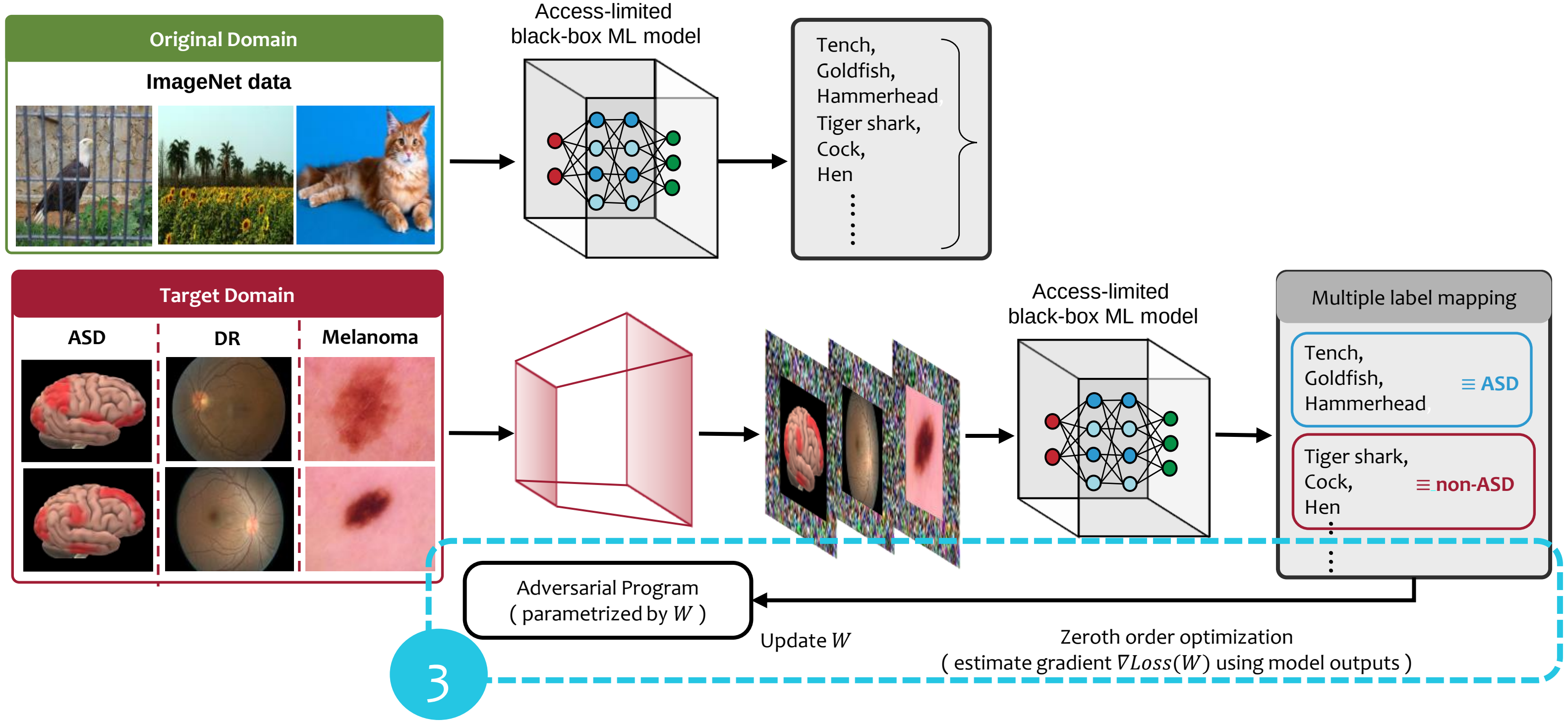
- Uses $h_j(\cdot)$ to denote m to 1 mapping function

$$h_{ASD}(F(X)) = \frac{F_{Tench}(X) + F_{Goldenfish}(X) + F_{Hammerhead}(X)}{3}$$

Proposed Framework

Step:

1. Generate adversarial example from adversarial program.
2. Multi-map source label to target.
3. **Optimize adversarial program with parameter W by ZOO method.**



- **Loss function:**

1. Maximize the probability of

$$p_t = P(h_j(y_{target})|X_{target})$$

2. Uses Focal loss

$$L_{focal}(p_t) = -\omega\boxed{(1-p_t)^{\gamma}}log(p_t)$$

Proposed Framework

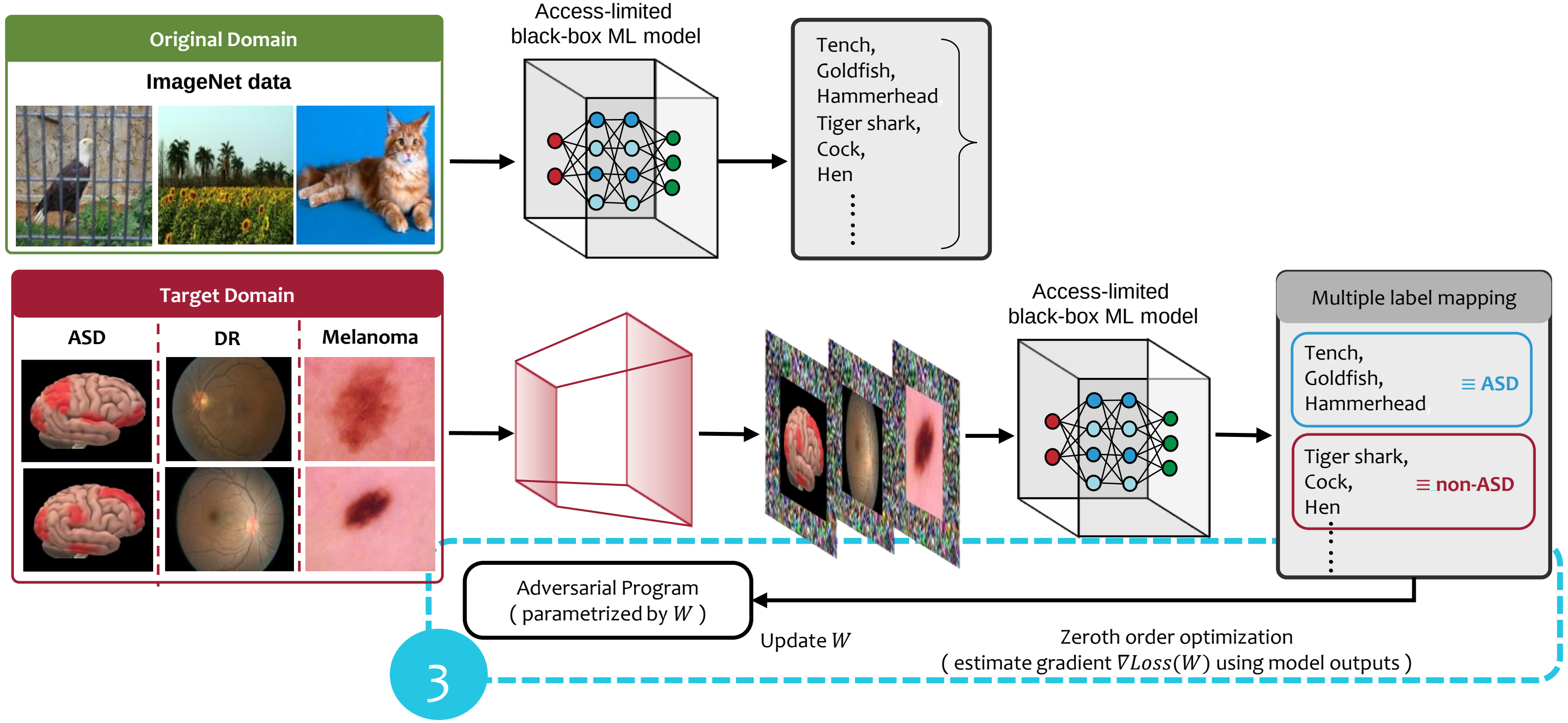
Step:

1. Generate adversarial example from adversarial program.
2. Multi-map source label to target.
3. **Optimize adversarial program with parameter W by ZOO method.**

Random vector-based estimation

$$g_j = b \cdot \frac{f(W + \varepsilon u) - f(W)}{\varepsilon} \cdot u$$

- Notes: b is a tuning parameter
  u is a random unit vector.



Proposed Framework

Step:

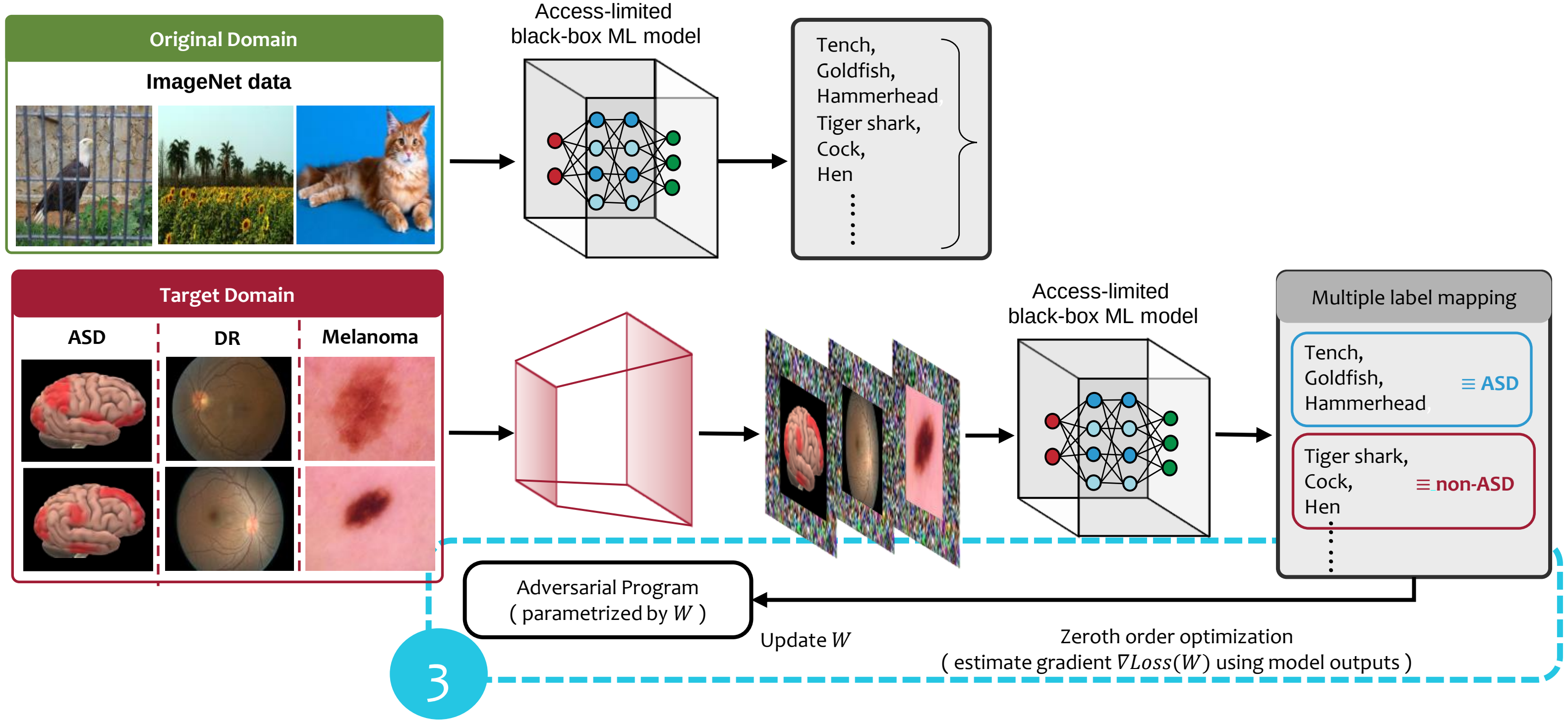
1. Generate adversarial example from adversarial program.
2. Multi-map source label to target.
3. **Optimize adversarial program with parameter W by ZOO method.**

Random vector-based estimation

One-time averaged gradient estimator

$$\nabla f(W) \approx \frac{1}{q}\sum_{j=1}^{q} g_j$$

- Notes: q is number of random unit vectors.



Proposed Framework

Step:

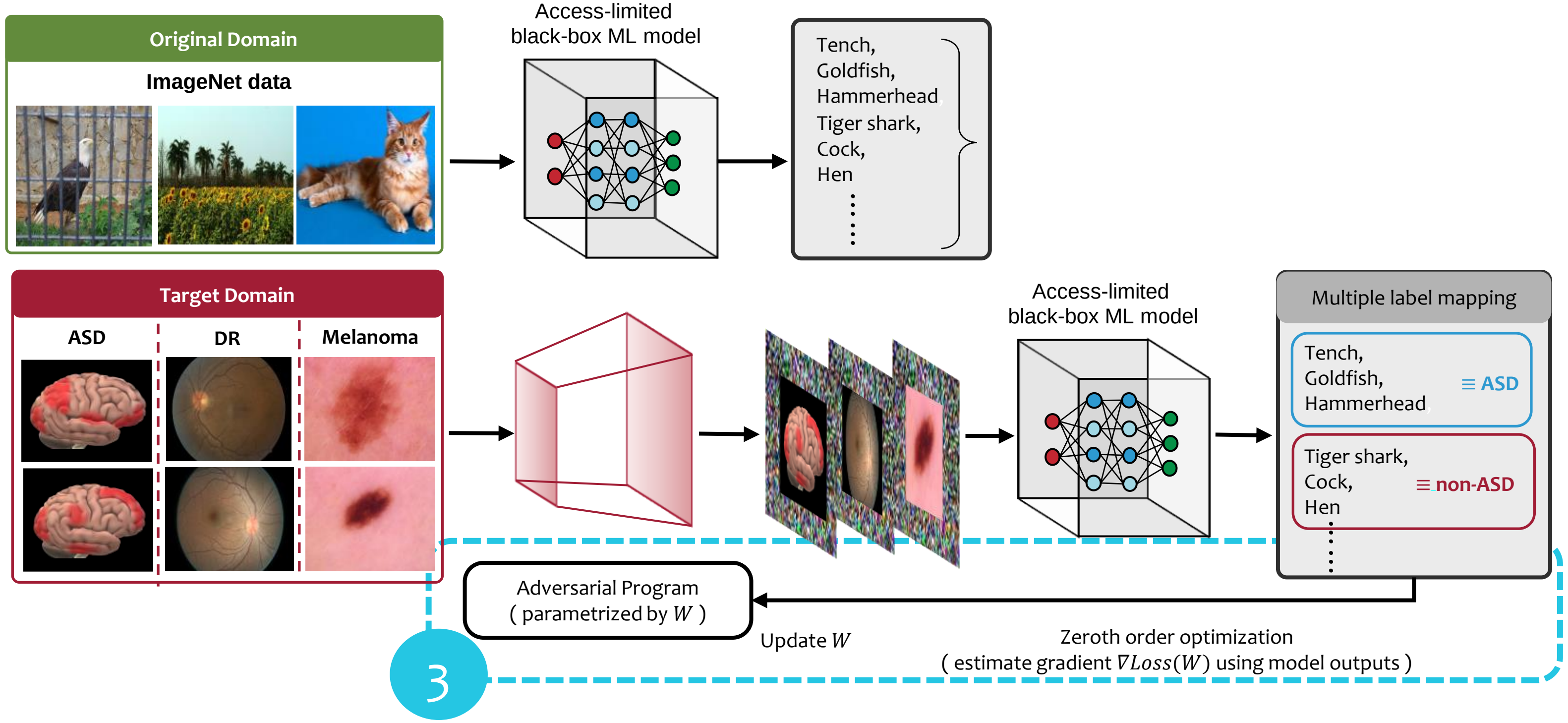
1. Generate adversarial example from adversarial program.
2. Multi-map source label to target.
3. **Optimize adversarial program with parameter W by ZOO method.**

Random vector-based estimation

One-time averaged gradient estimator

Uses SGD to optimize W

$$W_{t+1} = W_t - \alpha_t \cdot \nabla f(W_t)$$



Evaluation

Evaluation

Model

pretrained ImageNet models

- ResNet 50
- Inception V3
- DenseNet 121

Dataset

Medical imaging classification

- Autism Spectrum Disorder (ASD)
  - 2 classes
- Diabetic Retinopathy (DR)
  - 5 classes
- Melanoma
  - 7 classes

Online ML APIs

Real-life Black-box ML Models

- Clarifai.com
- Microsoft Custom Vision API

Baselines

- Vanilla adversarial reprogramming (white-box AR)
- Transfer learning with finetuned
- Train from scratch
- State-of-the-art (SOTA) of each task

Autism Spectrum Disorder (ASD)

- Autism Brain Imaging Data Exchange (ABIDE) database.
- 503 ASD and 531 non-ASD.
- The data sample is a 200×200 brain-regional correlation graph of fMRI measurements.

Model	Accuracy	Sensitivity	Specificity
Resnet 50 (AR)	72.99%	73.03%	72.13%
Resnet 50 (BAR)	70.33%	69.94%	72.71%
Train from scratch	50.96%	50.13%	52.34%
Transfer Learning (finetuned)	52.88%	54.13%	53.50%
Incept.V3 (AR)	72.30%	71.94%	74.71%
Incept.V3 (BAR)	70.10%	69.40%	70.00%
Train from scratch	49.80%	50.40%	51.55%
Transfer Learning (finetuned)	50.10%	51.23%	47.42%
SOTA 1. (Heinsfeld et al., 2018)	65.40%	69.30%	61.10%
SOTA 2. (Eslami et al., 2019)	69.40%	66.40%	71.30%

- Heinsfeld, A. S., Franco, A. R., Craddock, R. C., Buchweitz, A., and Meneguzzi, F. Identification of autism spectrum disorder using deep learning and the abide dataset. In NeuroImage: Clinical, 2018.
- Eslami, T., Mirjalili, V., Fong, A., Laird, A. R., and Saeed, F. Asd-diagnet: A hybrid learning approach for detection of autism spectrum disorder using fmri data. Frontiers in Neuroinformatics, 13, Nov 2019.

Diabetic Retinopathy (DR)

Model	From Scratch	Finetuning	AR	BAR
Resnet 50	66.23%	76.63%	80.48%	79.33%
Incept.V3	63.00%	74.20%	76.42%	74.33%
DenseNet 121	64.12%	71.29%	75.22%	72.33%

- Notes: The performance of **SOTA is 81.36%,** which requires additional data augmentation with fine-tuning on single Resnet 50.

Melanoma

Model	From Stratch	Finetuning	AR	BAR
Resnet 50	59.01%	76.90%	82.05%	81.71%
Incept.V3	52.91%	58.63%	82.01%	80.20%
Densenet 121	52.28%	58.88%	80.76%	78.33%

- Notes: The performance of **SOTA is 78.65%,** which uses specifically designed data augmentation with finetuning on Densenet.

- Sarki et al. Convolutional neural networks for mild diabetic retinopathy detection: an experimental study. *bioRxiv*, 763136.
- Li, et al. Skin lesion analysis towards melanoma detection via end-to-end deep learning of convolutional neural networks. arXiv preprint arXiv:1807.08332, 2018.

Evaluation

Dataset

Medical imaging classification

- Autism Spectrum Disorder (ASD)
  - 2 classes

Online ML APIs

Real-life Black-box ML Models

- Clarifai.com
- Microsoft Custom Vision API

Baselines

- Vanilla adversarial reprogramming (white-box AR)
- Transfer learning with finetuned
- Train from scratch
- State-of-the-art (SOTA) of each task

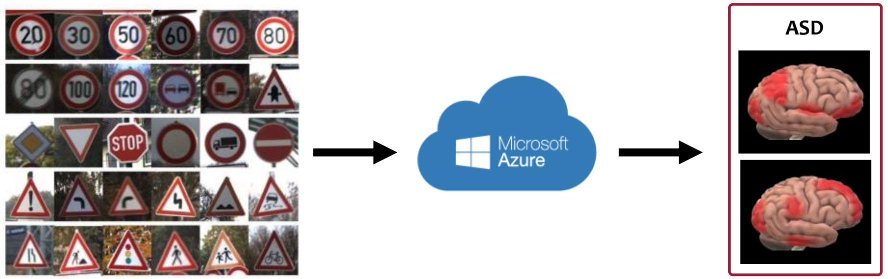
Reprogramming Real-life Prediction APIs

- **Clarifai NSFW API (2 classes)**
  - Recognize images or videos with inappropriate contents (e.g., “porn”, “sex”, or “nudity”).
  - Two output labels: **NSFW & SFW**.
- **Clarifai Moderation API (5 classes)**
  - Recognize images or videos have contents such as “gore”, “drugs”, “explicit nudity”, or “suggestive nudity”.

Orig. Task to New Task	q	# of query	Accuracy	cost
NSFW to ASD	15	12.8k	64.04%	$14.24
	25	24k	65.70%	$23.2
Moderation to ASD	15	11.9k	65.14%	$13.52
	25	23.8k	67.32%	$23.04

- **Microsoft Custom Vision API**
  - An online training platform for customized model.
  - Reprogramming **Traffic sign classification task → ASD task**.

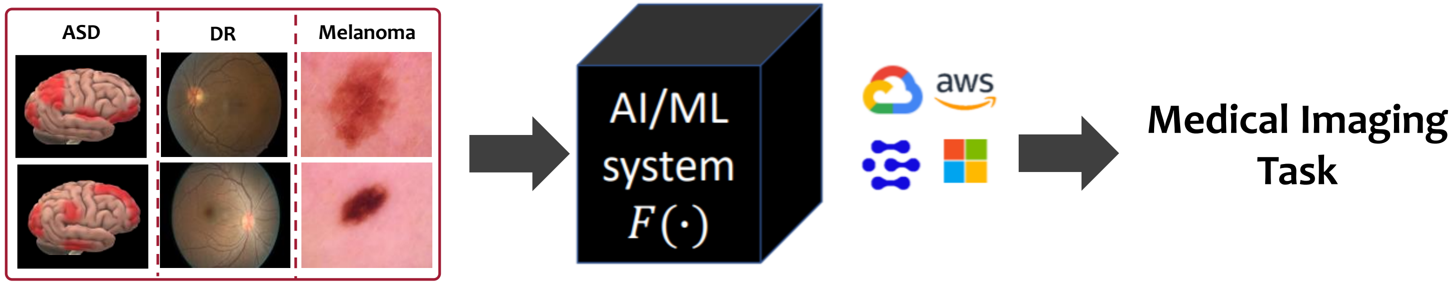
Original task to New task	q	# of query	Accuracy	Cost
Microsoft Custom Vision API to ASD	1	1.86k	48.15%	$3.72
	5	5.58k	62.34%	$11.16
	10	10.23k	69.15%	$20.46



- https://www.clarifai.com
- https://azure.microsoft.com/zh-tw/services/cognitive-services/custom-vision-service/

Conclusions

1. We proposed a novel **black-box adversarial reprogramming** framework for limited data classification tasks.
2. We used multi-label mapping and gradient-free approach to handle the infeasible gradient through black-box model.
3. Reprogrammed black-box ImageNet models for **three medical imaging tasks** and outperformed the general transfer learning methods.
4. We demonstrated the practicality of BAR by reprogramming online classification APIs from **Clarifai.com and Microsoft Custom Vision.**



Thanks for your attention