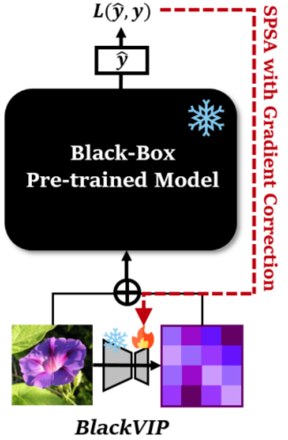
BAR: Black-Box Adversarial Reprogramming

Paper Review by Ravialdy

Main source: [Tsai'20;ICML] Tsai *et al.*, "Reprogramming Black-box Machine Learning Models with Scarce Data and Limited Resources." International Conference on Machine Learning, 2020

Brief Recap about BlackVIP :



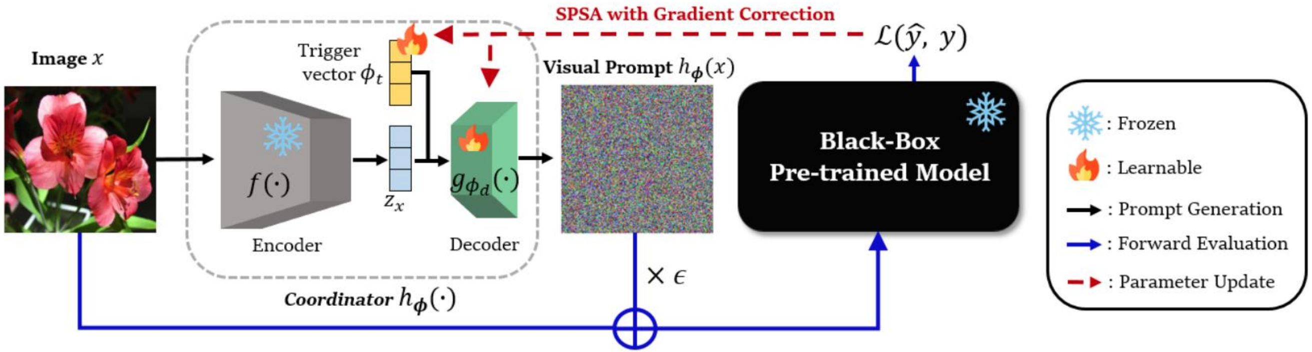


Figure 1. Illustration of BlackVIP complete architecture.

- **BlackVIP** is the **first work** for efficient transfer learning in **black-box setting** that **uses visual prompting**.
- However, **BlackVIP** is **not the very first method** to explore **black-box fine-tuning**.

Main source: [Oh'23;CVPR] Oh *et al.*, "BlackVIP: Black-Box Visual Prompting for Robust Transfer Learning." IEEE Computer Vision and Pattern Recognition(CVPR), 2023

Motivation :

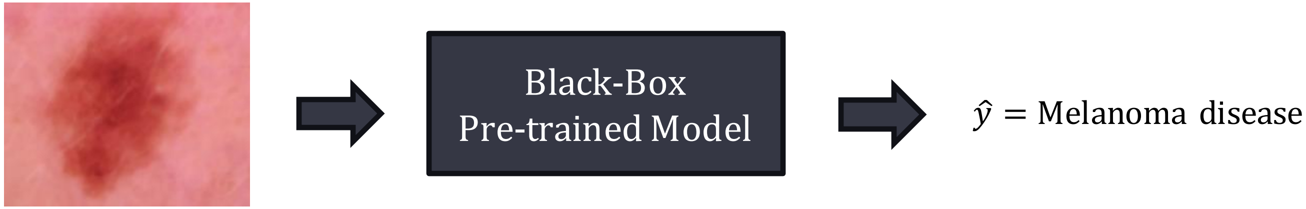


Figure 2. Illustration of implementing black-box Pre-trained Model (PTM) in medical imaging classification task.

- Collecting data in **medical domain** is very **expensive** and involves many experts -> PTM helps to improve accuracy!
- There exists some high-performing PTM, but those are **often in form of APIs** or **proprietary softwares** (e.g., Clarifai.com and Microsoft Custom Vision API).
- Is it possible to fine-tune those models without having the access into its parameters (black-box setting)?

Main source: [Tsai'20;ICML] Tsai *et al.*, "Reprogramming Black-box Machine Learning Models with Scarce Data and Limited Resources." International Conference on Machine Learning, 2020

#Key Idea 1 : Use Adversarial Program

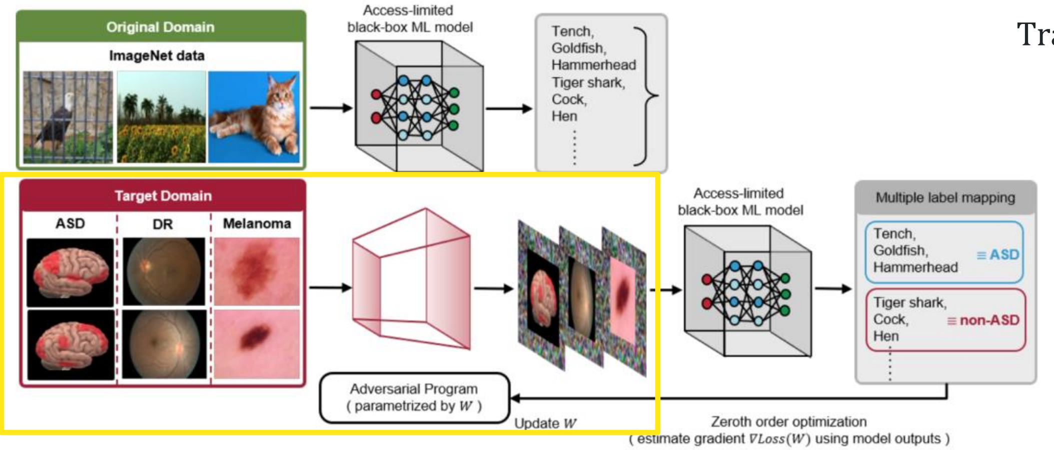


Figure 3. Diagram of adversarial program part in BAR model.

Transformed sampled data with adversarial program :

$$\tilde{X}_i = X_i + P \text{ and } P = \tanh(W \odot M)$$

Notation meanings :

D_i is target data for each sample $i =$ 1,2, …, n.

X_i is zero-padded data sample containing D_i.

$M \in \{0, 1\}$ is a binary mask function.

$W \in \mathbb{R}^d$ is a set of trainable parameters.

P is an adversarial program parametrized by W.

- Inspired by how adversarial attacks manipulates the prediction of a well-trained deep learning model.
- Sampled target data will be transformed with adversarial program parameterized by learnable W.

Main source : [Tsai'20;ICML] Tsai *et al.*, "Reprogramming Black-box Machine Learning Models with Scarce Data and Limited Resources." International Conference on Machine Learning, 2020

#Key Idea 2 : Use Multi-Label Mapping (MLM)

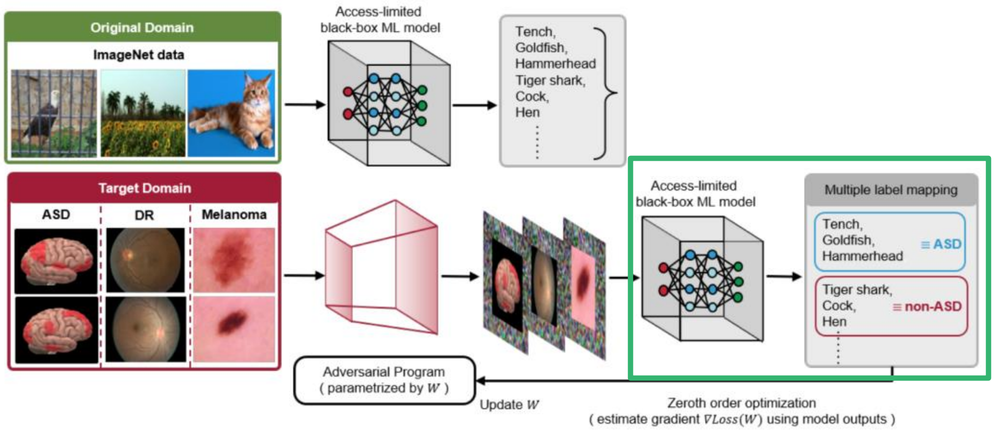


Figure 4. Diagram of Multi-Label Mapping (MLM) part in BAR model.

*Note : ASD = Autism Spectrum Disorder.
Assume this is ASD medical classification task.

- Take top-k probabilities and corresponding labels from logits generated by black-box model.
- If the resulting source label belongs to those top-k labels, then the predicted label will be ASD, otherwise is non-ASD.

Main source : [Tsai'20;ICML] Tsai *et al.*, "Reprogramming Black-box Machine Learning Models with Scarce Data and Limited Resources." International Conference on Machine Learning, 2020

Zeroth Order Optimization (ZOO) for Black-box Setting

$$L(y, f(x)) = \max(0, 1 - yf(x))$$

Common Hinge loss.

$$f(\mathbf{x}, t) = \max\{\max_{i \neq t} \log[F(\mathbf{x})]_i - \log[F(\mathbf{x})]_t, -\kappa\}$$

Proposed Hinge-like loss by [Chen'17].

Notation meanings :

y is the ground-truth label.

$f(x)$ is the predicted label.

$\log[F(\mathbf{x})]_i$ is log of the confidence score for class i.

$\log[F(\mathbf{x})]_t$ is log of the confidence score for desired class t.

κ is a tuning parameter.

Figure 5. Example of ZOO method by [Chen'17] which is inspired by hinge loss.

- The true gradients of black-box models are infeasible to get -> Calculate the estimate gradients!
- There are already many ZOO methods, most of them are used in black-box adversarial attacks.
- The first ZOO method proposed by [Chen'17] for black-box adversarial attack in image classification.

Main source : [Tsai'20;ICML] Tsai *et al.*, "Reprogramming Black-box Machine Learning Models with Scarce Data and Limited Resources." International Conference on Machine Learning, 2020

Zeroth Order Optimization for BAR :

$$g_j = b \cdot \frac{f(W + \beta U_j) - f(W)}{\beta} \cdot U_j,$$

$$\bar{g}(W) = \frac{1}{q} \sum_{j=1}^{q} g_j,$$

$$W_{t+1} = W_t - \alpha_t \cdot \bar{g}(W_t),$$

Notation meanings :

$f(W)$ be the loss or objective function.

W is the optimization variables.

q is a perturbation constant.

$\bar{g}(W)$ is an averaged gradient estimator.

b is is a scalar balancing bias constant.

β is the smoothing parameter.

$U_j \in \mathbb{R}^d$ is random direction vector.

Figure 6. ZOO used in BAR model.

- BAR uses the one-sided averaged gradient estimator proposed by [Liu'18] which is the best ZOO method at that time.

Main source : [Tsai'20;ICML] Tsai *et al.*, "Reprogramming Black-box Machine Learning Models with Scarce Data and Limited Resources." International Conference on Machine Learning, 2020

Performance :

Model	Accuracy	Sensitivity	Specificity
ResNet 50 (BAR)	**70.33%**	**69.94%**	**72.71%**
ResNet 50 (AR)	72.99%	73.03%	72.13%
Train from scratch	51.55%	51.17%	53.56%
Transfer Learning (finetuned)	52.88%	54.13%	54.70%
Incept. V3 (BAR)	**70.10%**	**69.40%**	**70.00%**
Incept. V3 (AR)	72.30%	71.94%	74.71%
Train from scratch	50.20%	51.43%	52.67%
Transfer Learning (finetuned)	52.10%	52.65%	54.42%
SOTA 1. (Heinsfeld et al., 2018)	65.40%	69.30%	61.10%
SOTA 2. (Eslami et al., 2019)	69.40%	66.40%	71.30%

Predicted Class

Actual Class		Positive	Negative	
	Positive	True Positive (TP)	False Negative (FN) Type II Error	Sensitivity $\frac{TP}{(TP+FN)}$
	Negative	False Positive (FP) Type I Error	True Negative (TN)	Specificity $\frac{TN}{(TN+FP)}$
		Precision $\frac{TP}{(TP+FP)}$	Negative Predictive Value $\frac{TN}{(TN+FN)}$	Accuracy $\frac{TP+TN}{(TP+TN+FP+FN)}$

Figure 7. Performance comparison on ASD classification task.

Model	From Scratch	Finetuning	AR	BAR
ResNet 50*	73.44%	76.63%	80.48%	**79.33%**
Incept. V3	72.10%	74.20%	76.42%	**74.33%**
DenseNet 121	67.22%	71.29%	75.22%	**72.33%**

Figure 8. Performance comparison on diabetic retinopathy detection task. The notation ∗ denotes the network used in SOTA method.

- Notice that training from scratch and finetuning is not good in this case due to limited data.
- Also note that Adversarial Reprogramming (AR) is a white-box version of BAR.

Main source: [Tsai'20;ICML] Tsai *et al.*, "Reprogramming Black-box Machine Learning Models with Scarce Data and Limited Resources." International Conference on Machine Learning, 2020

Thank You