I See, Therefore I Do: Estimating Causal Effects for Image Treatments

Abhinav Thorat, Ravi Kolla*, Niranjan Pedanekar

Sony Research India

Introduction

Background

Under Rubin-Neyman potential outcomes framework, Individual Treatment Effect (ITE) is defined as:

$$\text{ITE} = \mathrm{E}[Y_1 - Y_0 | X = x]$$

Research gap

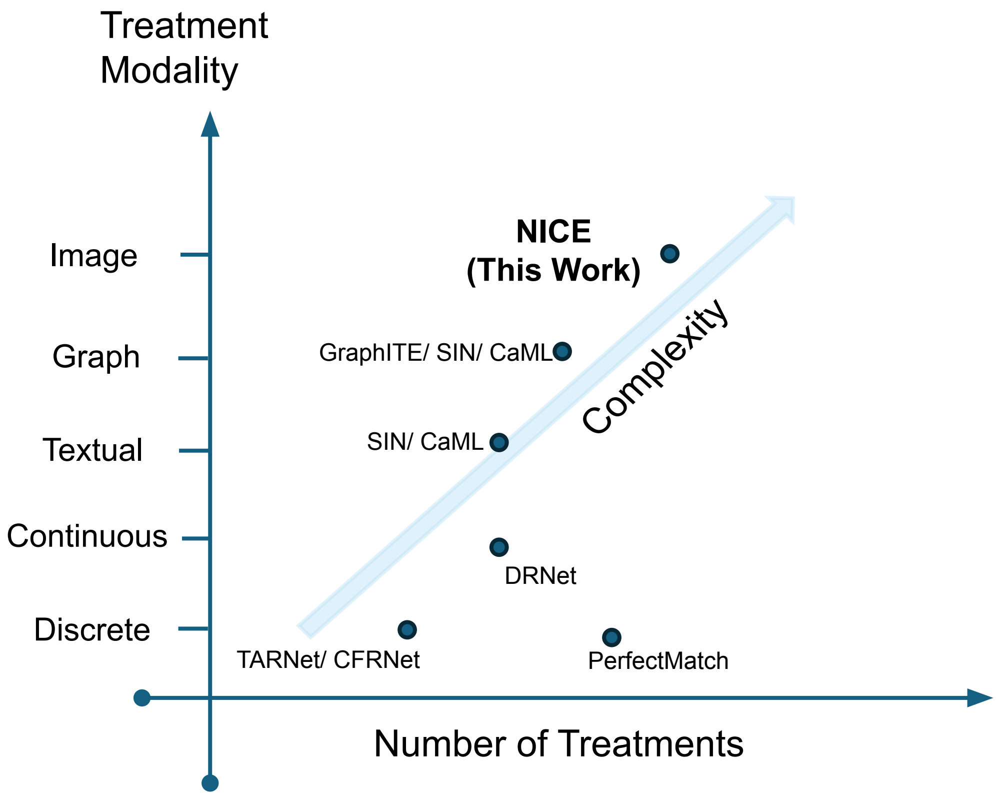
Majority of the ITE estimation literature **does not consider treatment information in the ITE estimation,** and merely represents treatments in scalar form

Problem statement

This work addresses **ITE estimation for Image treatments** by **utilizing auxiliary treatment information** in the estimation under **multiple treatments** setting

Practical use

Thumbnail personalization in video streaming and e-commerce platforms etc.



Evolution of Causal Effect Estimation Algorithms

Assumptions

Unconfoundedness

Conditional on observed covariates, potential outcomes are independent of treatment assignment

$$(Y_1, Y_2, \ldots Y_k) \perp t \mid x.$$

Positivity

Each user has positive probability of receiving any available treatment

$$0 < P(t = a \mid x = x') < 1 \; \forall 1 \leq a \leq k.$$

SUTVA

Each user's observed outcome depends only on the treatment they received, independent of other users' assigned treatments.

Key contributions

Dataset simulation

- ❖ As there are no existing datasets, created new semi-synthetic datasets
- ❖ Image treatments are real, and covariates and potential outcomes are simulated

Neural Network Architecture

- ❖ Proposed NICE architecture with shared representation learning, MSE and MMD losses
- ❖ Capability of handling multiple treatments and zero shot (novel treatment) scenarios

Empirical evaluation

- ❖ Demonstrated the superior performance of NICE against across various experimental setups

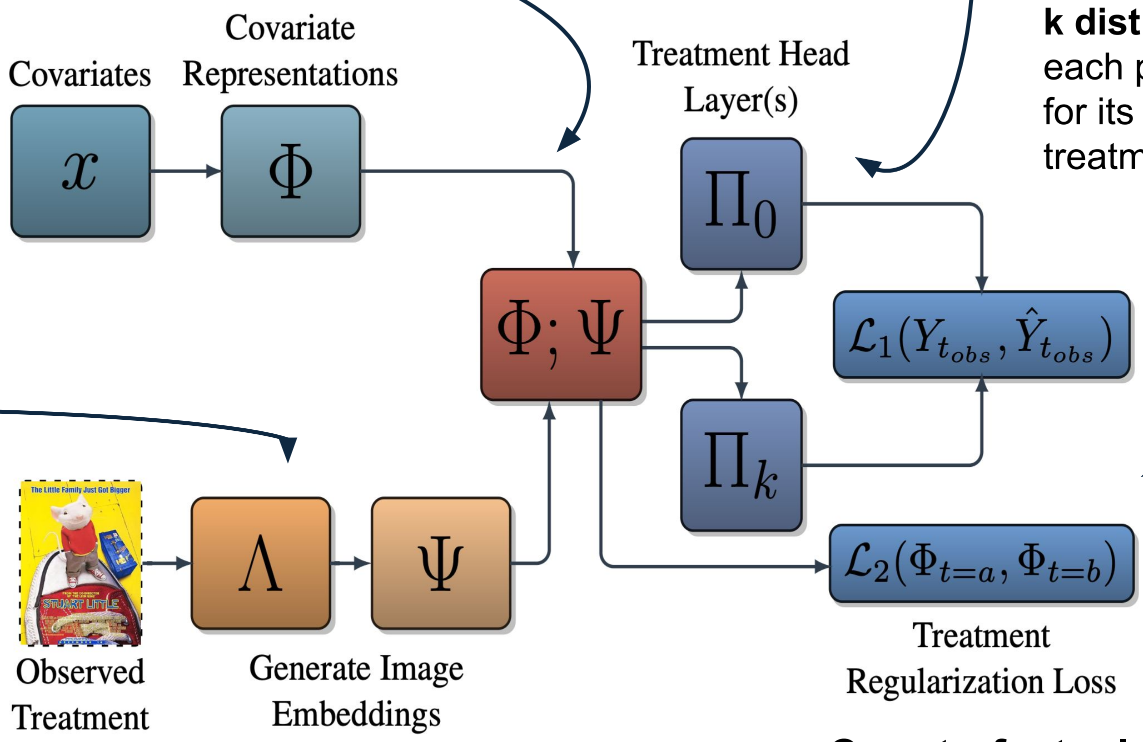
Theoretical guarantees

- ❖ Derived an upper bound on the PEHE error metric for ITE estimation

Proposed Model

Joint Representation Learning

Learns low-dimensional embeddings for both **user covariates** and **image treatments** using separate fully connected networks.



Treatment-Specific Outcome Heads

For k possible treatments, NICE uses **k distinct neural network heads**, each predicting the potential outcome for its corresponding image treatment.

Image Embedding via Pre-trained Models

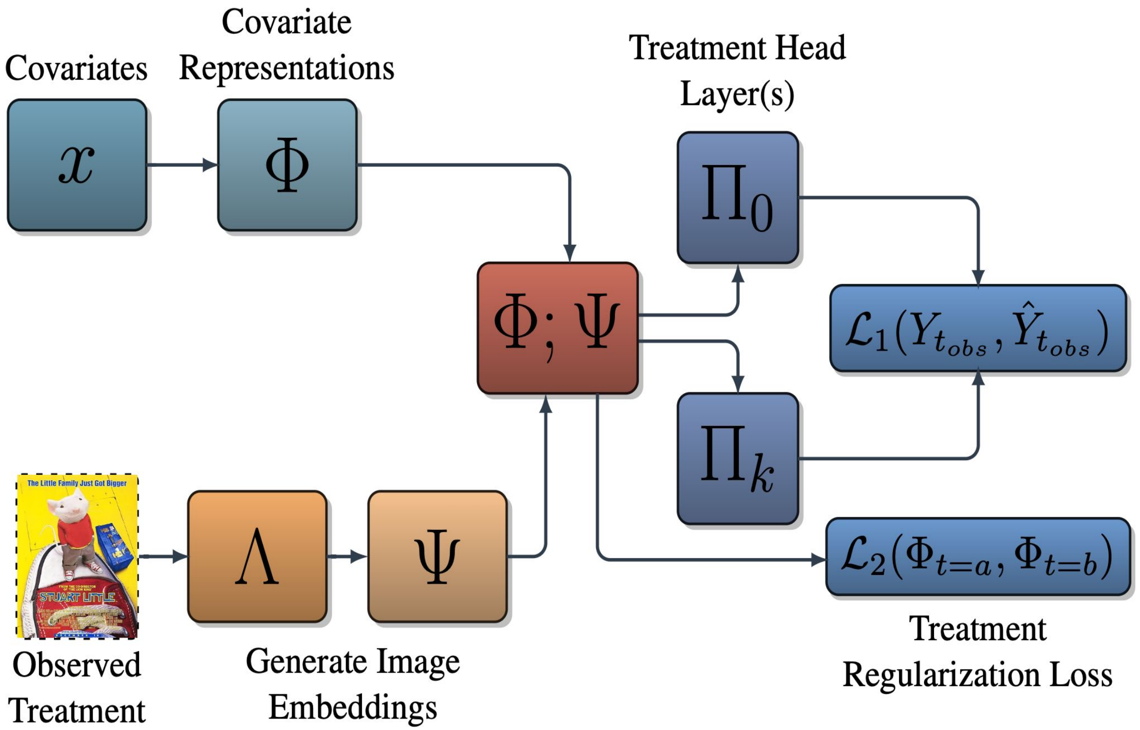
Uses pre-trained models (e.g., **ResNet**, **VGG**) to extract semantic embeddings from treatment images, which are then refined by a learnable network.

Model Agnostic wrt Pretrained Model

Counterfactual Estimation with Regularization

Combines **mean squared error (MSE)** for factual outcomes with **Maximum Mean Discrepancy (MMD)** to reduce treatment assignment bias across embeddings.

Loss Functions



MSE Loss $L_1 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i^{t_i})^2$

The **MSE loss** ensures the model learns to **accurately predict factual outcomes** by minimizing the error between observed and predicted values.

Loss Function: $L = \alpha \cdot L_1 + \beta \cdot L_2$

Treatment Regularization Loss

$$L_2 = \frac{1}{\binom{k}{2}} \sum_{a=1}^{k} \sum_{b=1}^{a-1} \mathrm{MMD}(\{\Phi;\Psi\}_{t=a}, \{\Phi;\Psi\}_{t=b})$$

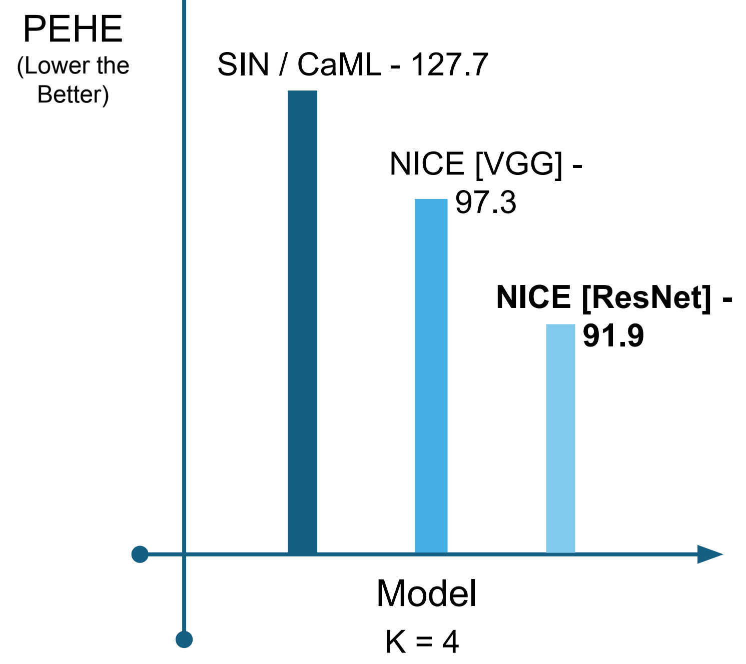
The **MMD-based regularization loss** promotes **balanced and unbiased representations** across treatments, enabling **reliable counterfactual estimation** even under **treatment assignment bias**.

NICE achieves lower PEHE across all number of treatments settings

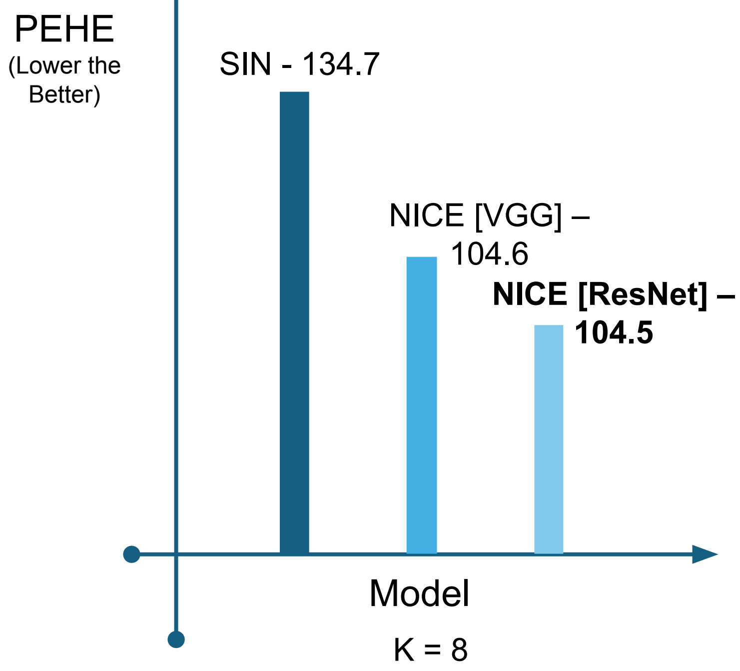
Setting
- ❖ No. of treatment (K) = {4, 8, 16}
- ❖ Moderate treatment assignment bias, κ =10 for all treatments

$$\epsilon_{\text{PEHE}} = \frac{1}{\binom{k}{2}} \sum_{a=1}^{k} \sum_{b=1}^{a-1} \left[\frac{1}{n} \sum_{i=1}^{n} \left(\hat{\tau}^{a,b}(x_i) - \tau^{a,b}(x_i) \right)^2 \right]$$

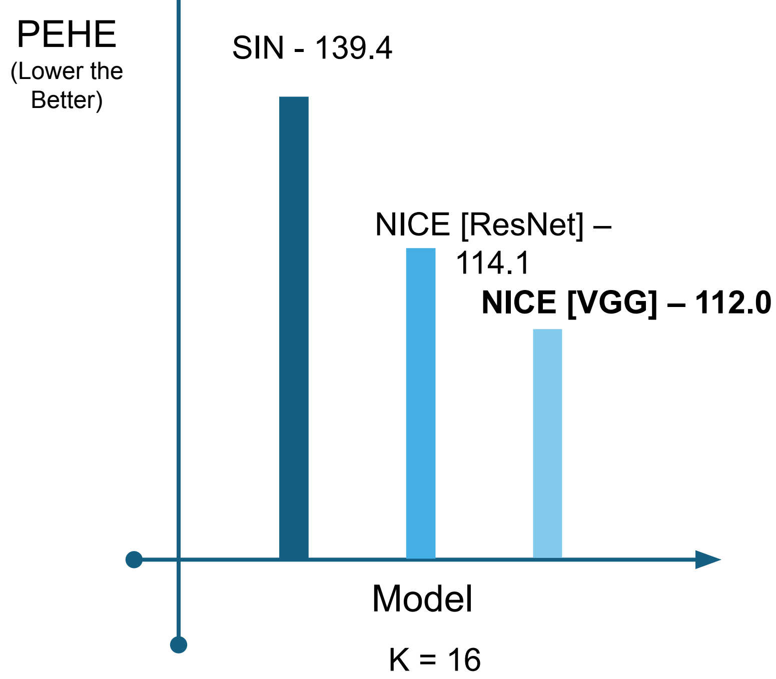
28% Lower than Next Best Baseline



22.4% Lower than Next Best Baseline



19.6% Lower than Next Best Baseline



NICE outperforms baselines in zero-shot scenarios

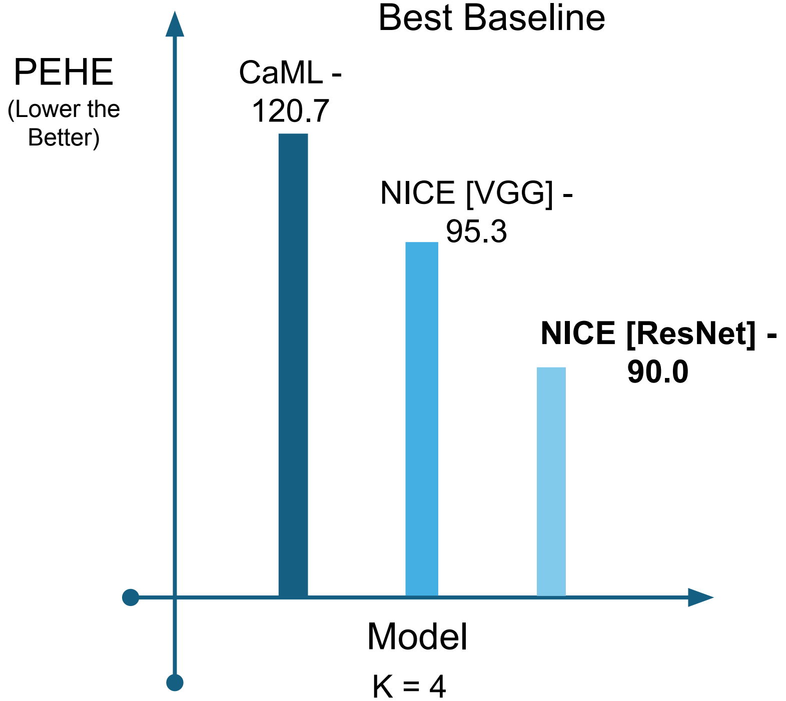
Setting
- ❖ No. of treatment (K) = {4, 8, 16}
- ❖ A treatment is called as zero-shot if its samples are not seen by a model during training
- ❖ Considered 1 zero-shot treatment
- ❖ Moderate treatment assignment bias, κ =10

$$\epsilon_{\text{PEHE}}^{\text{ZS}} = \frac{1}{k-1}\sum_{\substack{a=1\\ a\neq z}}^{k}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\tau}^{a,z}(x_i)-\tau^{a,z}(x_i)\right)^2\right]$$

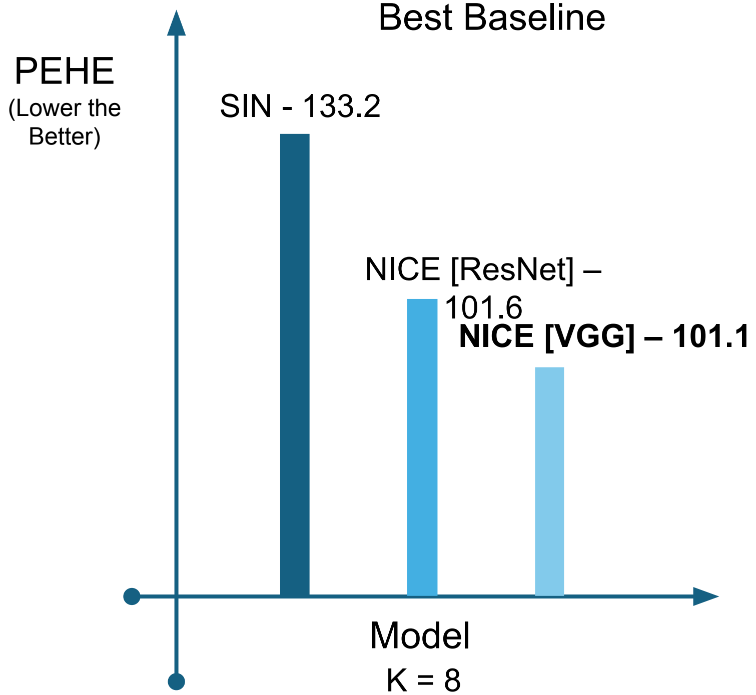
Next Best Baseline

NICE Variant

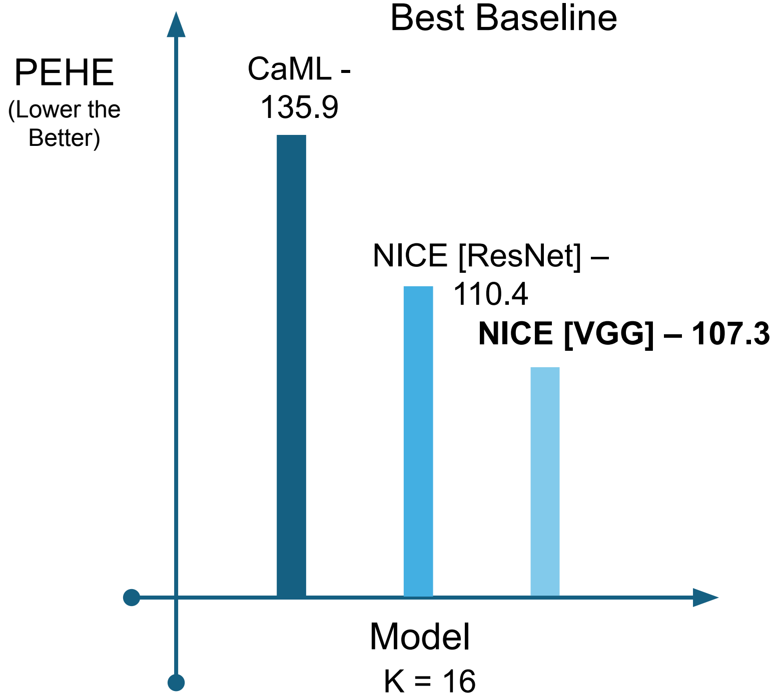
25.4% Lower than Next Best Baseline



24% Lower than Next Best Baseline



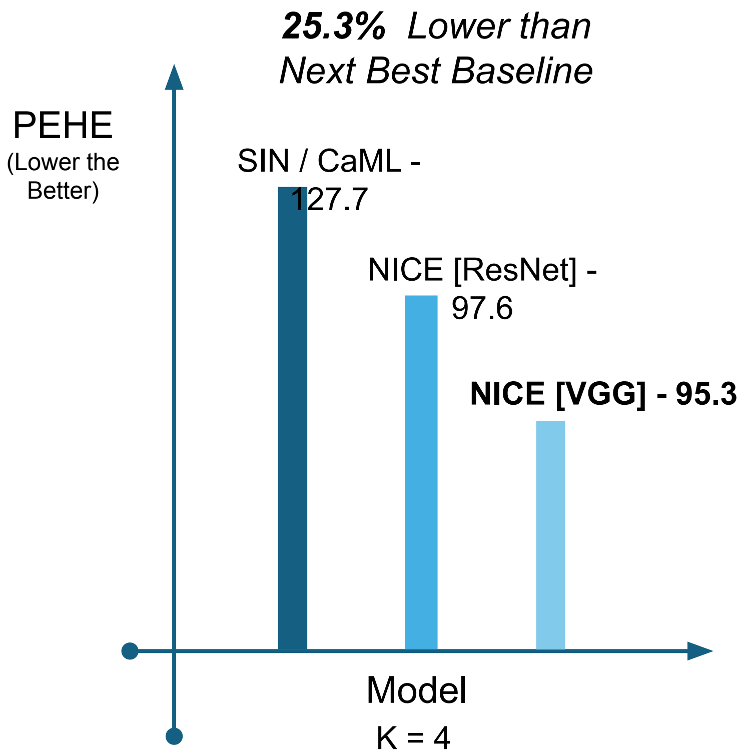
21% Lower than Next Best Baseline

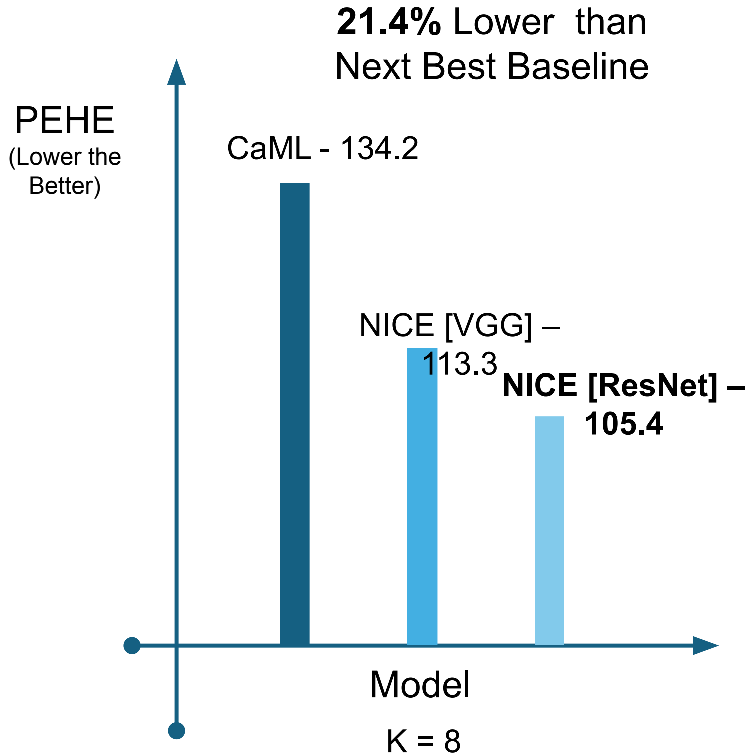


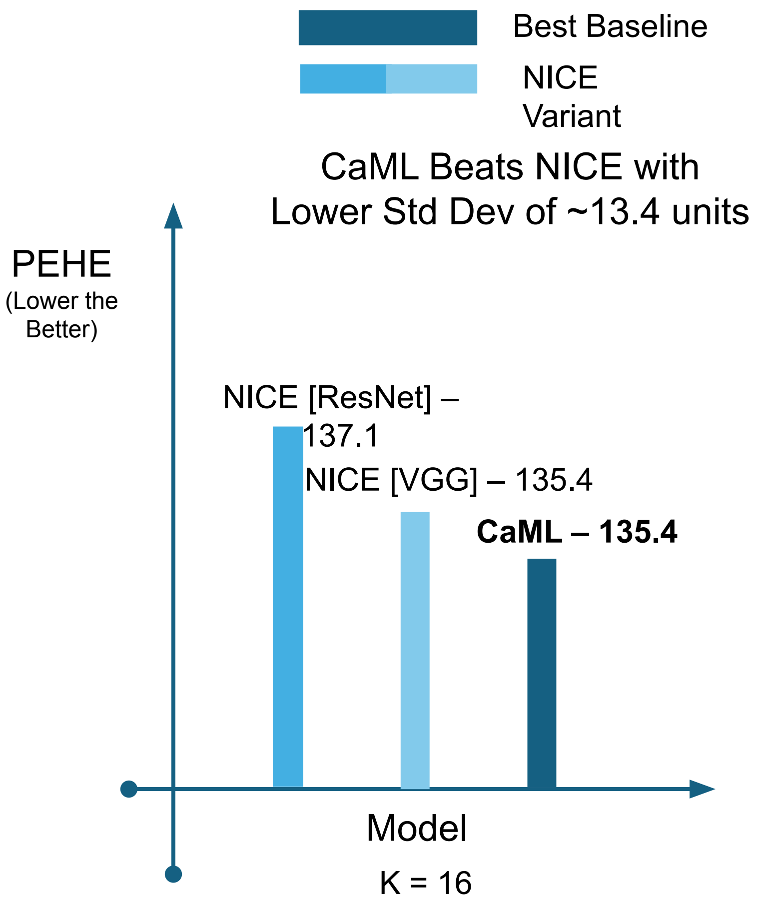
NICE performance under high treatment assignment bias scenario

Setting

- ❖ No. of treatment (K) = {4, 8, 16}
- ❖ Considered High treatment assignment bias scenario,
  κ = 100 for a treatment and 10 for all other treatments







Theoretical Guarantees of NICE

Theorem. Let $\Phi : \mathcal{X} \to \mathbb{R}_X$ and $\Psi : \mathcal{I} \to \mathcal{R}_I$ are twice differentible and invertible functions. Let Π be a hypothesis function. Let $\mathcal{G}$ denote a family of functions $g : \mathcal{R}_X \times (\mathcal{R}_I; \{0,1\})$. Assume the loss function L used to define $l_{\Pi,\Phi,\Psi}$ is the squared loss function. Further, assume that there exists a constant $D_{\Phi,\Psi} > 0$ s.t. the loss function $l(\cdot)$ satisfies the following: $\frac{1}{D_{\Phi,\Psi}} l_{\Pi,\Phi,\Psi}\left(\Phi^{-1}(r_x), \Psi^{-1}(r_{I_t}), t\right) \in \mathcal{G}$ for $t \in \{0,1\}$. Then, we have

$$\epsilon_{\text{PEHE}}(\Pi,\Phi,\Psi) \leq \underbrace{\frac{2}{k}\sum_{a=1}^{k} \epsilon_F^{t=a}(\Pi,\Phi,\Psi)}_{\text{MSE loss}} + \underbrace{\frac{2}{\binom{k}{2}} \sum_{a=1}^{k}\sum_{b=1}^{a-1} \left(D_{\Phi,\Psi}\, \mathtt{IPM}_{\mathcal{G}}\left(p_\Phi^{t=a}, p_\Phi^{t=b}\right)\right.}_{\text{Average IPM loss}} -2\min\{\sigma_{Y_a}^2, \sigma_{Y_b}^2\}\Big).$$

Provides an **upper bound** on the **PEHE** obtained by **NICE** as a function of **MSE loss** computed using **factual outcomes** and **average IPM loss** between all pairs of treatments

Conclusion

- Studied Individual Treatment Effect (ITE) estimation problem for Image treatments

- Proposed SOTA **NICE** framework that utilizes auxiliary treatment information to obtain improved causal effect estimates

- Demonstrated NICE’s superior performance against baselines across various setups including **zero-shot** and **high treatment assignment bias** scenarios

- Derived **an upper bound** on the PEHE error metric for NICE algorithm

Thank you