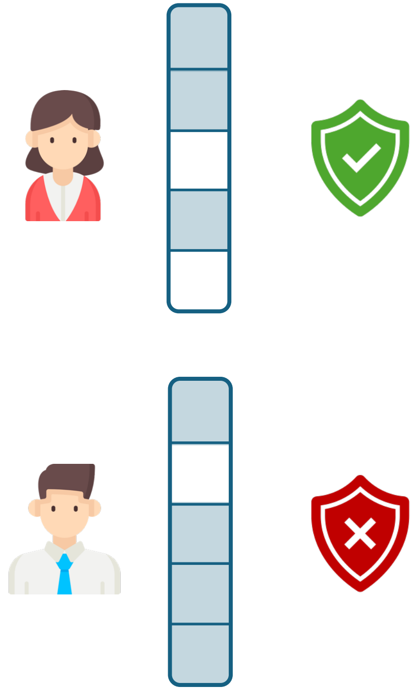
Estimating Conditional Mutual Information for Dynamic Feature Selection

Soham Gadgil*, Ian Covert*, Su-In Lee
Paul G. Allen School of Computer Science & Engineering
University of Washington

Introduction

- Dynamic Feature Selection: paradigm where we **sequentially query features** to make predictions with a minimal budget



Introduction

- Dynamic Feature Selection: paradigm where we **sequentially query features** to make predictions with a minimal budget

- Important in settings like **emergency medicine** where not all features are available, are costly to acquire, and best selections differ between predictions

Introduction

- Dynamic Feature Selection: paradigm where we **sequentially query features** to make predictions with a minimal budget

- Important in settings like **emergency medicine** where not all features are available, are costly to acquire, and best selections differ between predictions

- We propose **an information-theoretic** approach which selects features based on their **conditional mutual information (CMI)** with the target variable

Contribution

- We develop a learning approach (DIME) to estimate the CMI in a **discriminative fashion** and prove that our objective recovers the exact CMI at optimality

- Most works assume uniform feature costs; we adapt DIME to scenarios with **non-uniform feature costs**

- We analyze the role of **variable feature budgets** between samples and how they enable an improved **cost-accuracy tradeoff** through multiple stopping criteria

- DIME provides **consistent gains** across all the datasets tested compared to many recent methods

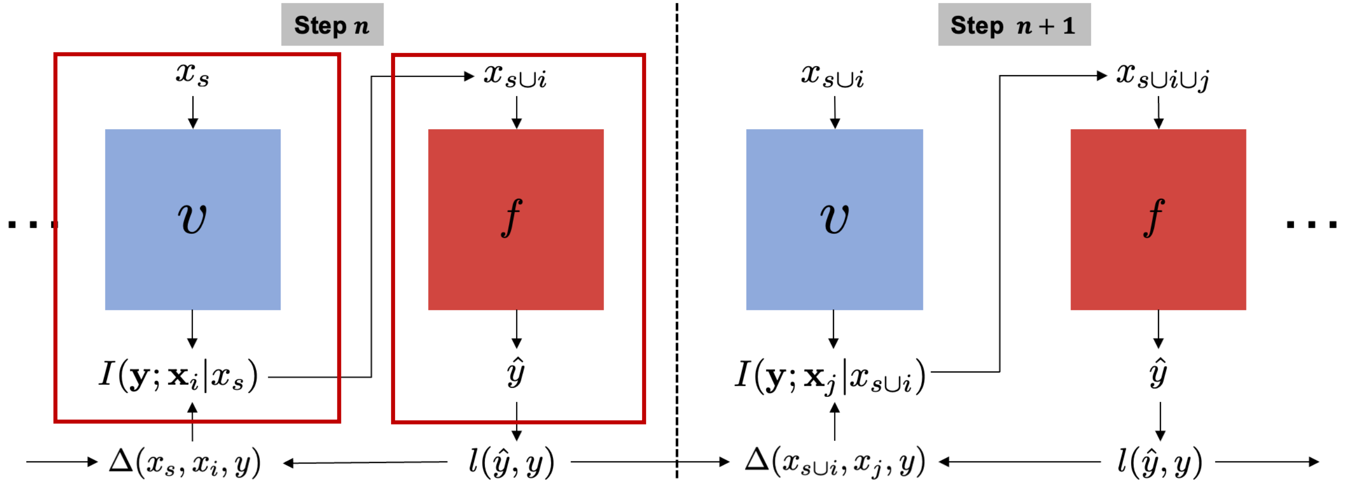
Proposed Method

- CMI, denoted as $I(y; x_i|x_s)$, shows how much information an unknown feature x_i provides about the target y given the current set of selected features x_s

- Given a value network that accurately predicts CMI, we can use it greedily select the next feature

- This is identical to performing greedy uncertainty minimization

Training Approach



- Two networks: the **value network** $\boldsymbol{v}$ and **predictor network** $\boldsymbol{f}$
- At each selection step n the value network $v(x_s;\phi)$ predicts the CMI $I(y;x_i|x_s)$ for each candidate feature
- The feature x_i which maximizes the CMI is used for the next prediction $f(x_{S\cup i};\theta)$

Training Approach

- Predictor loss: cross entropy

$$\min_{\theta} \mathbb{E}_{xy}\mathbb{E}_s[\ell(f(x_s;\theta), y)]$$

- Value network loss: MSE

$$\min_{\phi} \mathbb{E}_{xy}\mathbb{E}_s\mathbb{E}_i\left[\left(v_i(x_s;\phi) - \Delta(x_s, x_i, y)\right)^2\right]$$

where $\Delta(x_s, x_i, y) = \ell(f(x_s), y) - \ell(f(x_s, x_i), y)$

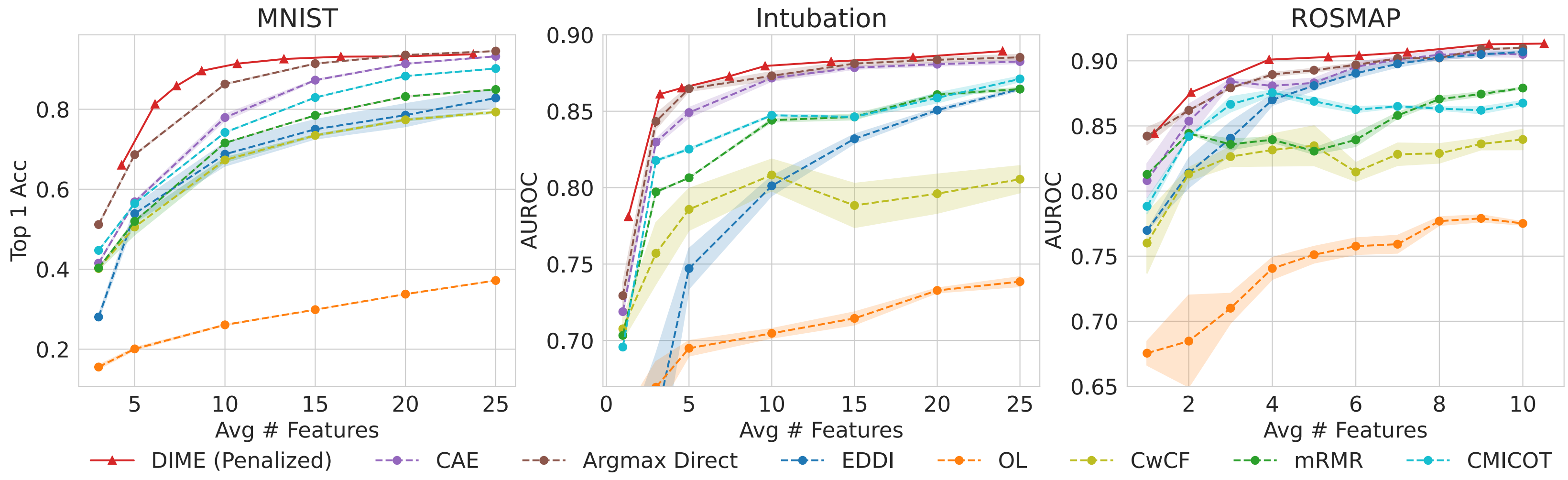
- Models are trained jointly, with selections being made using the ϵ-greedy approach

Datasets

- Tabular Datasets
  - MNIST (flattened, d = 784)
  - ROSMAP dataset for dementia onset prediction (d = 43)
  - Intubation dataset for predicting the need of respiratory support (d = 35)

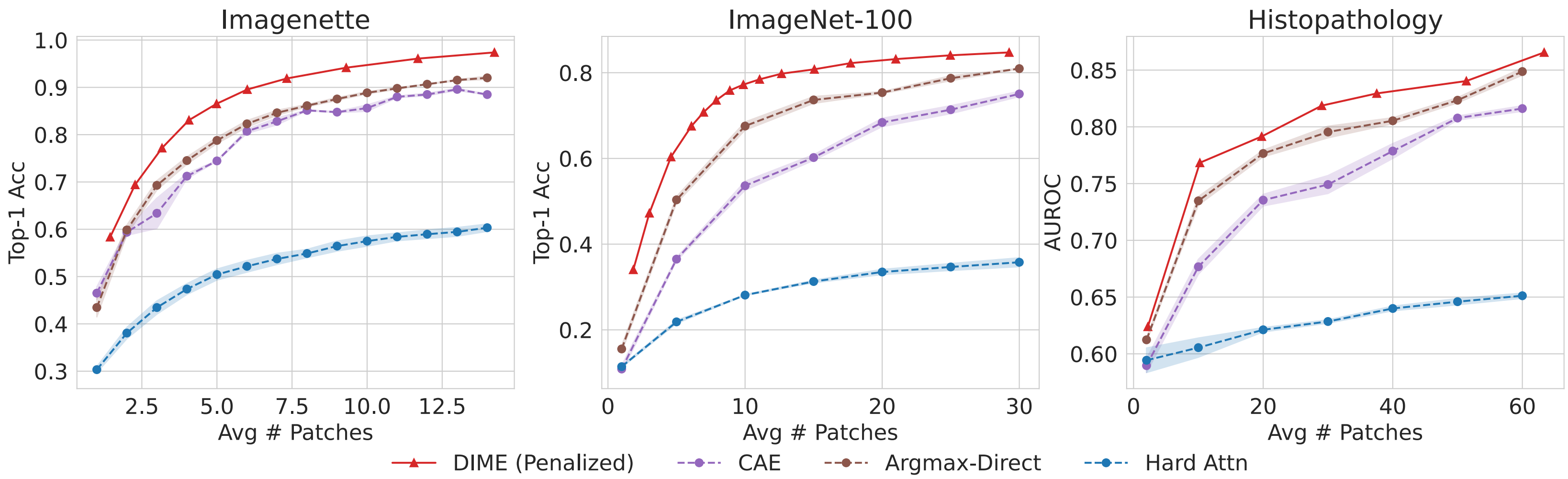
- Image Datasets
  - Imagenette, an Imagenet subset with 10 classes
  - Imagenet-100, an Imagenet subset with 100 classes
  - MHIST, a histopathology dataset to predict benign or pre-cancerous lesions

Results: Tabular Datasets



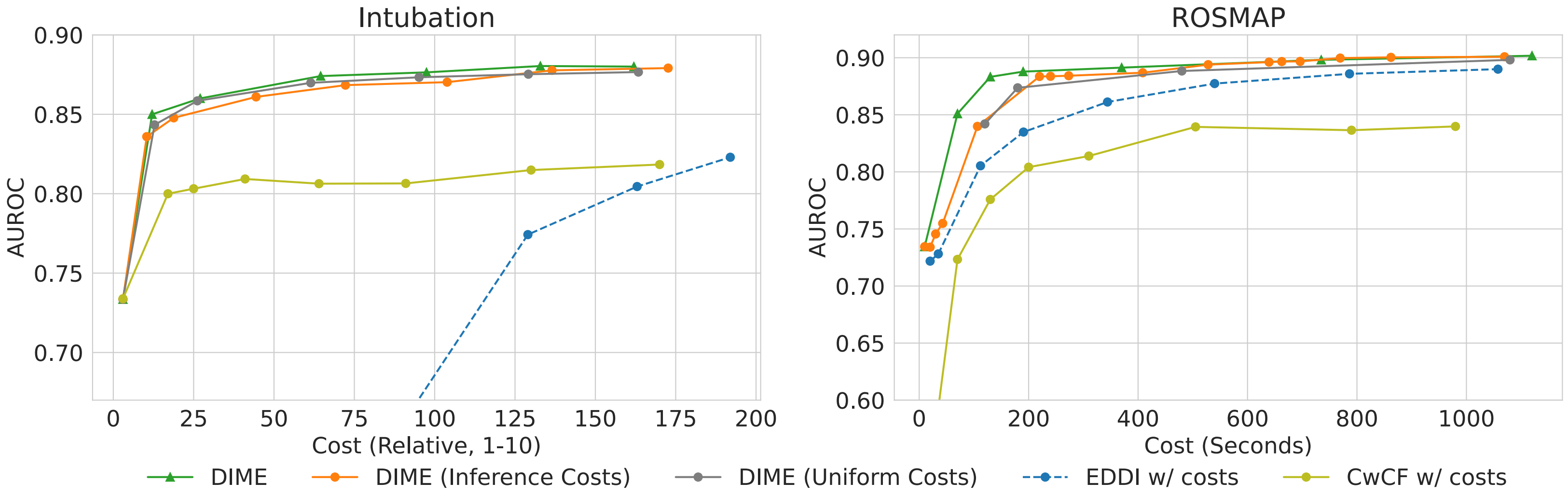
- Used multilayer perceptrons (MLPs) with two hidden layers and ReLU non-linearity
- DIME achieves the best results among all methods for both medical diagnosis tasks
- Performs the best on MNIST, achieving over 90% accuracy with an average of ~ 10/784 features (1.27%)

Results: Image Datasets



- Used Vision Transformers (ViT-small-patch-16) with a shared backbone
- Images are 256x256 with each feature being a 16x16 patch
- DIME with the penalized stopping criteria outperforms the baselines for all feature budgets
- Achieves nearly 97% accuracy on Imagenette with only ~15/196 patches (7.7%).

Results: Non-Uniform Costs



- For Intubation, relative costs are considered
- For ROSMAP, costs are expressed as the time required to obtain each feature
- DIME provides the best cost-accuracy tradeoff, reflecting the improved CMI estimation

Conclusion

- This work presents DIME, a new DFS approach enabled by estimating the CMI in a **discriminative fashion**

- Our approach involves learning value and predictor networks, trained in an **end-to-end fashion** with a straightforward regression objective

- We prove that our training approach recovers the exact CMI at optimality

- Empirically, DIME accurately estimates the CMI and enables an improved **cost-accuracy tradeoff**

- DIME beats prior methods, is robust to higher image resolutions, scales to more classes, and benefits from modern architectures