Leveraging Importance Weights in Subset Selection

Gui Citovsky[1], Giulia DeSalvo[1], Sanjiv Kumar[1], Srikumar Ramalingam[1], Afshin Rostamizadeh[1], Yunjuan Wang[2]

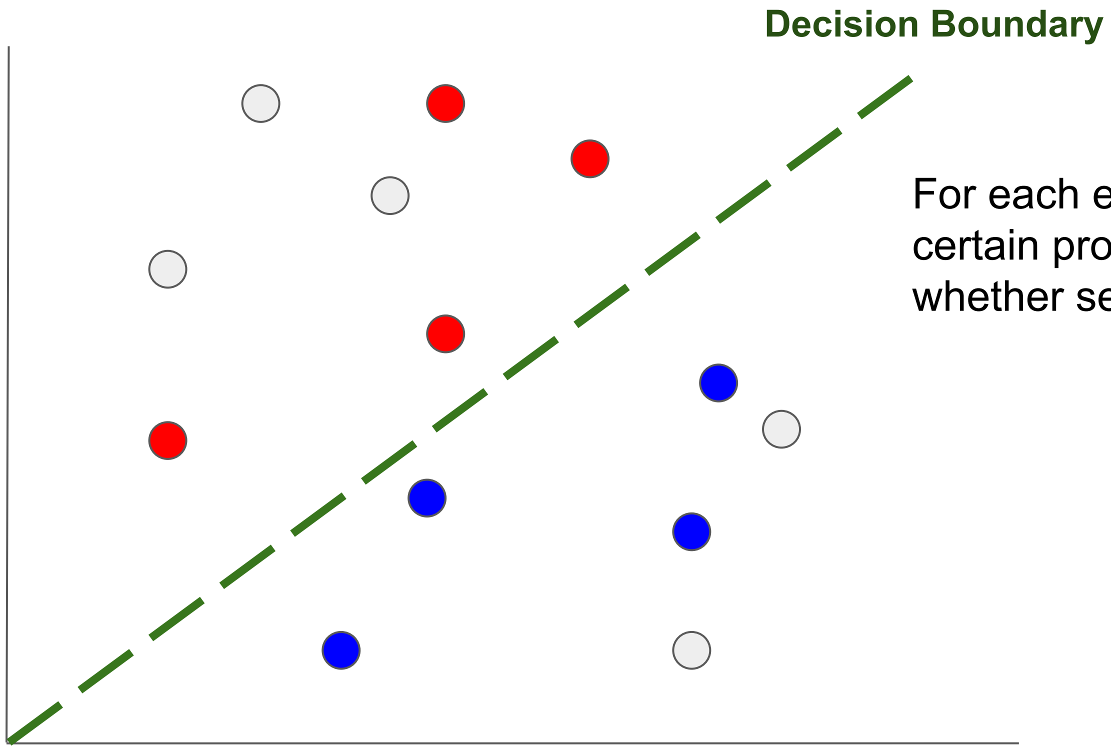
[1] Google Research [2] Johns Hopkins University

ICLR 2023

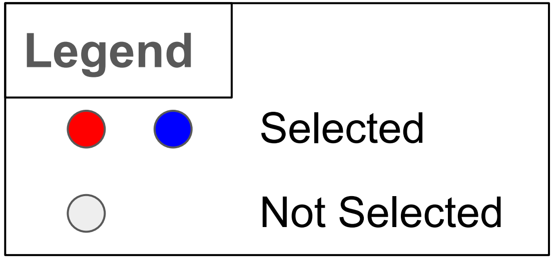
Motivation

- **Problem:** Training modern deep networks on large labeled datasets incur high computational costs.
- **Idea:** Find **the most informative subset** from the large labeled pool to approximate (or even improve upon) training with the entire training set.
- **One approach:** Weighted subsets of a dataset that can act as the proxy for the whole dataset.
  - Most competitive subset selection algorithms do not assign weights to the selected examples.
- **Goal:** Design an efficient subset selection algorithm that can
  1. Work for general loss functions and hypothesis classes,
  2. Select examples by importance sampling.

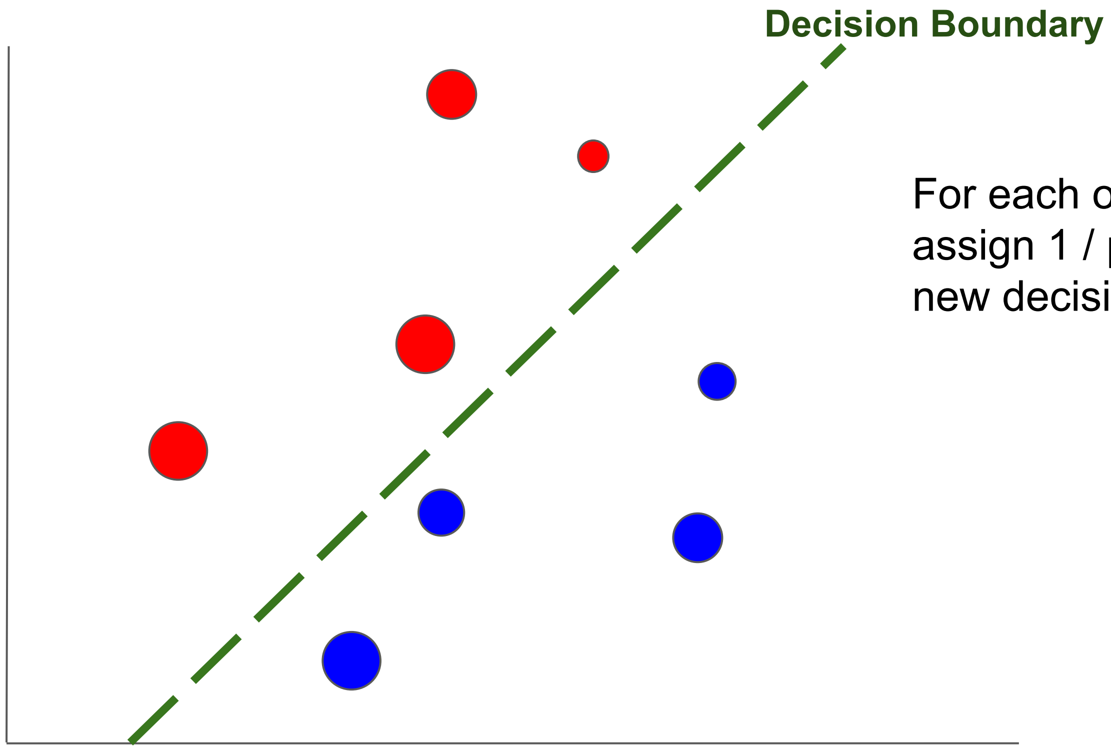
Importance Sampling



For each example (x, y), flip a coin with certain probability p(x, y) to decide whether select the example or not.



Importance Sampling



For each of the selected example (x, y), assign 1 / p(x, y) weight, and learn a new decision boundary.

Legend

Larger size point have larger weight.

Importance **We**ighted **S**ubset Selection (IWeS)

Labeled pool $\mathcal{P}$. Weight cap u. Initialize a seed set $\mathcal{S}_0$ uniformly selected at random from $\mathcal{P}$, with each examples having weight 1. Initialize the subset $\mathcal{S}$=$\mathcal{S}_0$. Remove $\mathcal{S}_0$ from $\mathcal{P}$.

For each sampling iteration $\boldsymbol{r}$:

- **Training step:**
  - Train two models $\boldsymbol{f_r}, \boldsymbol{g_r}$ on $\mathcal{S}$ with independent random initializations using the importance-weighted loss; i.e., $f_r = \arg\min_{h \in \mathcal{H}} \sum_{(\mathrm{x},y,w)\in\mathcal{S}} w \cdot \ell(h(\mathrm{x}), y)$
- **Sampling step:**
  - Select example (x, y) uniformly at random from $\mathcal{P}$.
  - Set the sampling probability $p(\mathrm{x}, y)$ using ***entropy-based disagreement*** or ***entropy*** criteria (defined later).
  - $Q \sim$ Bernoulli($p(\mathrm{x}, y)$).
    - If Q=1, include (x, y) inside $\mathcal{S}$ with weight $\min\left(\frac{1}{p(\mathrm{x},y)}, u\right)$, remove it from $\mathcal{P}$.
  - Keep sampling until select enough examples.

Sampling probability in IWeS

- **Entropy-based Disagreement (IWeS-dis).** The sampling probability is based on the disagreement on two functions with respect to entropy restricted to (x, y).

$$p(\mathrm{x}, y) = |\mathrm{P}_{f_r}(y|\mathrm{x}) \log_2 \mathrm{P}_{f_r}(y|\mathrm{x}) - \mathrm{P}_{g_r}(y|\mathrm{x}) \log_2 \mathrm{P}_{g_r}(y|\mathrm{x})|$$

  - ➢ Need **labeled examples**, $\mathrm{P}_{f_r}(y|\mathrm{x})$ is the probability of class y with f_r given example $\mathbf{x}$.
  - ➢ Need to **train two models**. If they disagree on (x, y), $p(\mathrm{x}, y)$ will be large, (x, y) is likely to be selected.

- **Entropy (IWeS-ent).** The sampling probability is the normalized entropy of model prediction on $\mathbf{x}$

$$p(\mathrm{x}, \cdot) = -\sum_{y' \in \mathcal{Y}} \mathrm{P}_{f_r}(y'|\mathrm{x}) \log_2 \mathrm{P}_{f_r}(y'|\mathrm{x}) / \log_2 |\mathcal{Y}|.$$

  - ➢ $p(\mathrm{x}, \cdot)$ is high when f_r is not confident about its prediction as it effectively randomly selects a label from $\mathcal{Y}$.
  - ➢ Can be used in an **active learning setting** where the algorithm can only access the unlabeled examples.
  - ➢ **Only train one model**, save some computational cost.

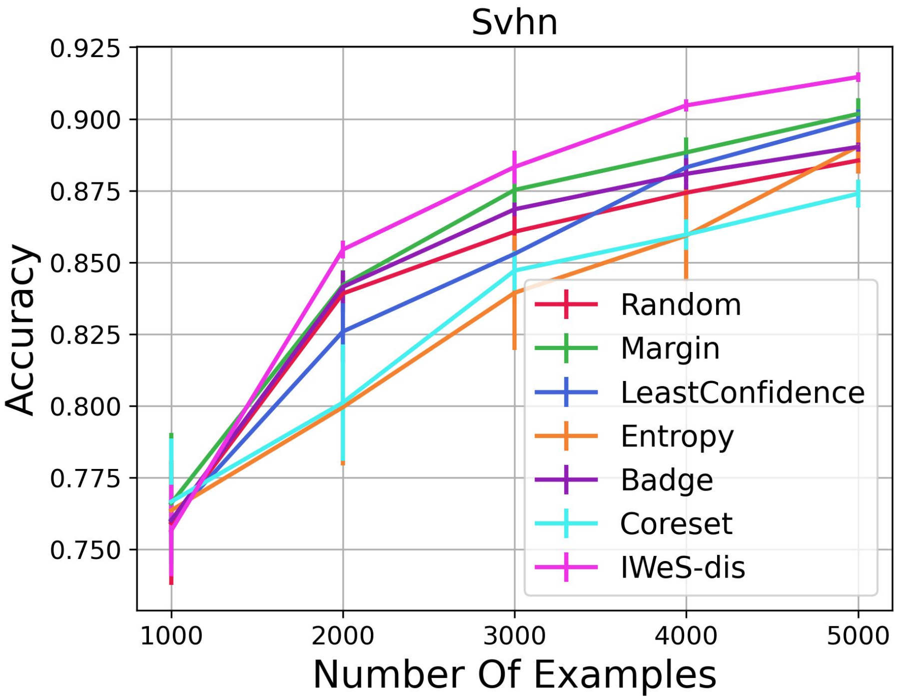
Theoretical Motivation

- The weighted loss $\frac{1}{|\mathcal{P}|}\sum_{i\in\mathcal{P}}\frac{Q_i}{p(\mathbf{x}_i,y_i)}\ell(f(\mathbf{x}_i),y_i)$ is **an unbiased estimator** of the population risk $\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\ell(f(\mathbf{x}),y)]$.
- A closely related algorithm IWeS-V operates on an i.i.d. exampl$(\mathbf{x}_t,y_t)$) , defines the sampling probability as $p_t = \max_{f,g\in\mathcal{H}_t} \ell(f(\mathbf{x}_t),y_t)-\ell(g(\mathbf{x}_t),y_t)$.
- **Generalization guarantee**: with probability at least $1-\delta$, the generalization gap is bounded by $\mathcal{O}(\sqrt{\log(T/\delta)/T})$, where *T* is the labeled pool size.
- **Expected sampling rate** bound of IWeS-V is tighter compared with IWAL algorithm that can only get access to the unlabeled examples.

Experiment Setup

- **Datasets:**
  - Small scale (multi-class) datasets: CIFAR10, CIFAR100, SVHN, EUROSAT, CIFAR10 Corrupted, Fashion MNIST.
  - Large scale (multi-label) dataset: Open Images v6.
- **Baselines:** Uncertainty Sampling (margin, entropy, least confidence), BADGE, Coreset, Random Sampling.
- **Models:** VGG-16 for small scale datasets, ResNet-101 for Open Images v6.

Experimental Results – IWeS-dis compared with baselines



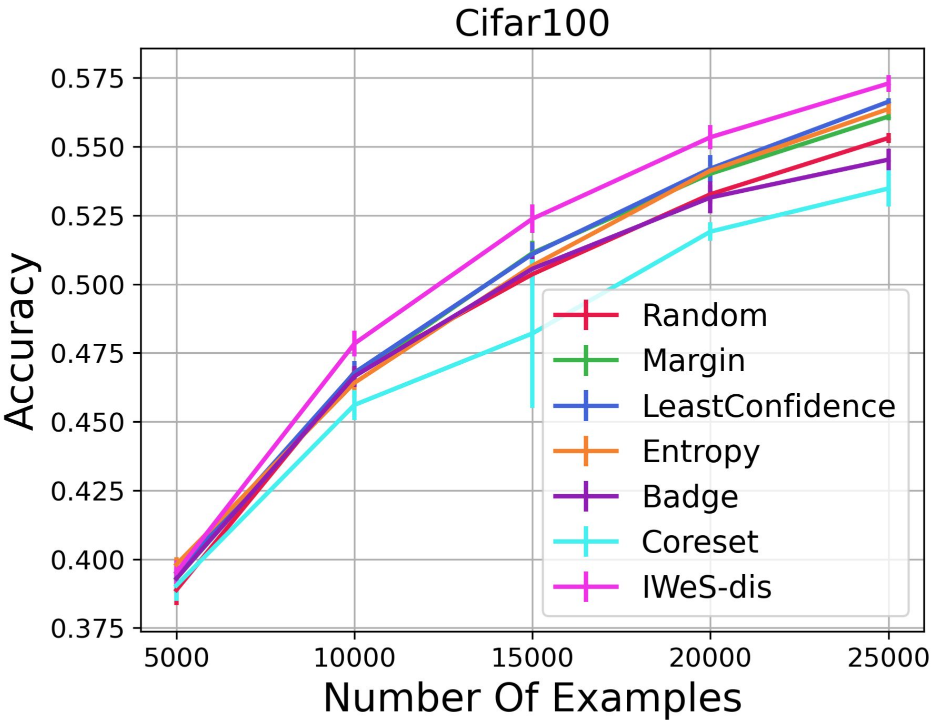


Figure: Accuracy of VGG16 when trained on examples selected by **IWeS-dis** and baseline algorithms.

Experimental Results – IWeS-dis compared with IWeS-ent

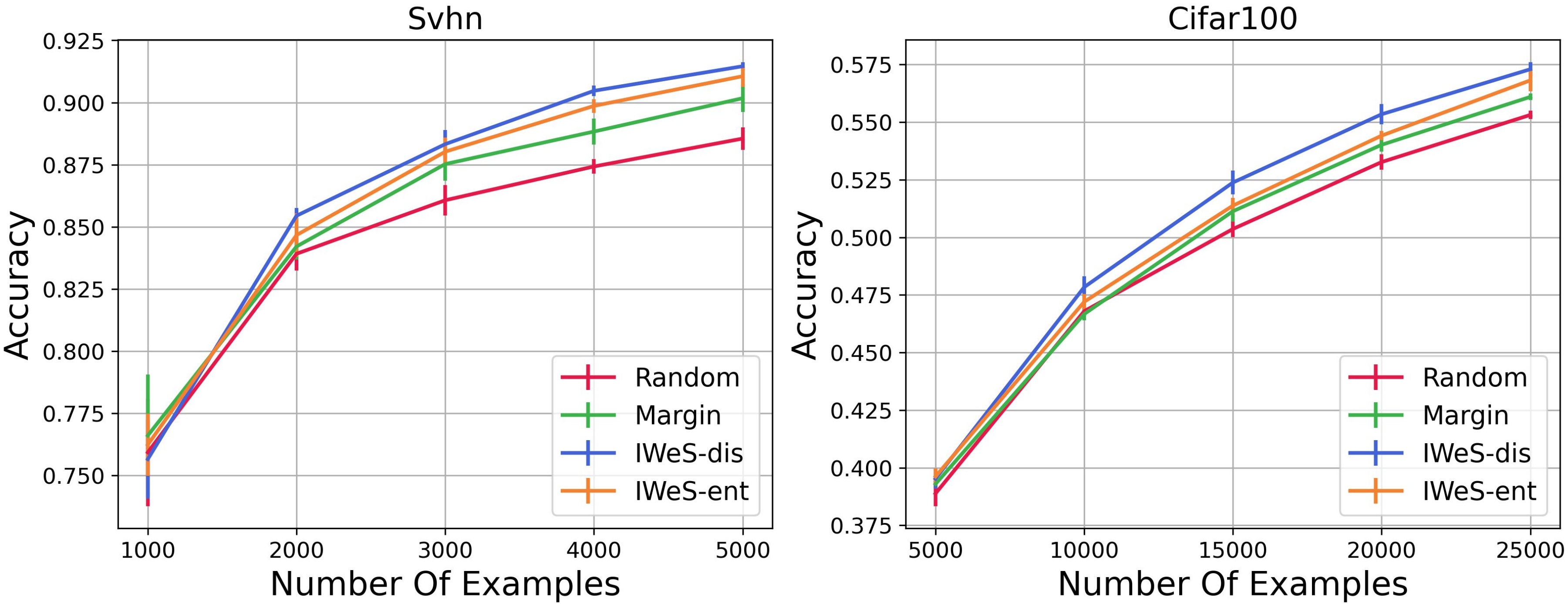
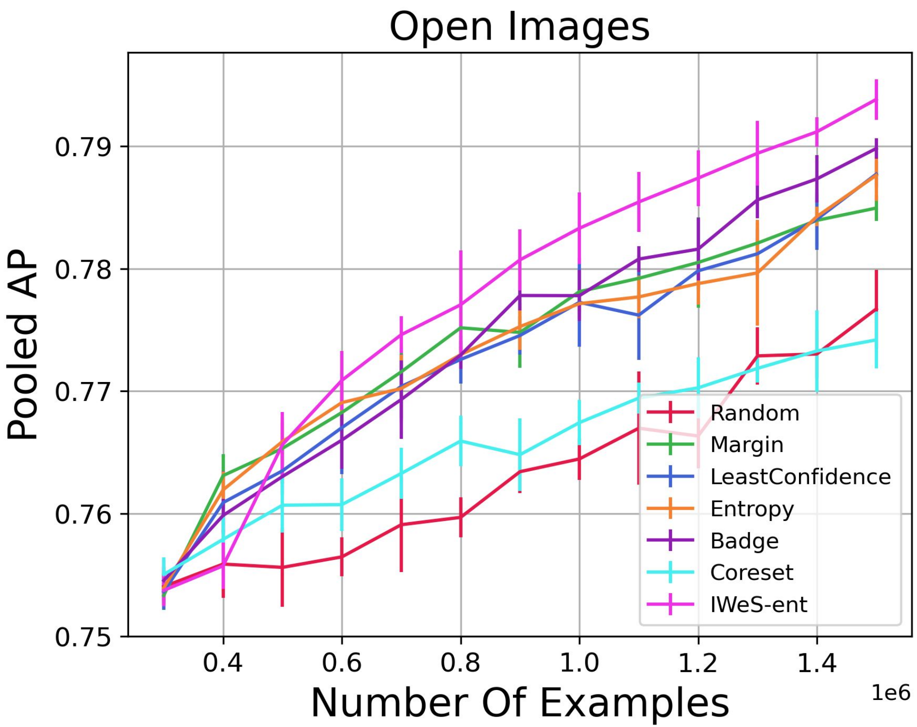


Figure: Accuracy of VGG16 when trained on examples selected by **IWeS-ent**, **IWeS-dis**, margin sampling and random sampling.

Experimental Results – Open Images

- IWeS-dis / IWeS-ent consistently outperform the baselines on different datasets.
- IWeS-dis algorithm slightly outperforms the IWeS-ent algorithm on a few datasets.
- IWeS-dis requires training two neural networks, which is computationally expensive for Open Image dataset.



Pooled Average Precision of ResNet101 trained on examples selected by **IWeS-ent** and the baseline algorithms.

Conclusion

- Developed **a novel subset selection algorithm**, IWeS, that selects examples by **importance sampling** where the sampling probability assigned to each example is based on the **entropy** of models trained on previously selected batches.
- Demonstrate IWeS achieves **significant improvement** over several baselines for six common multi-class datasets and one large-scale multi-label dataset.
- Provide an initial **theoretical analysis**, proving **generalization bound** $\mathcal{O}(1/\sqrt{T})$ that depends on the full training dataset size T, and showing **a tighter sampling rate** bound by leveraging label information compared with IWAL that does not use label information.