Disguise Adversarial Networks for Imbalanced Click-through Rate Prediction

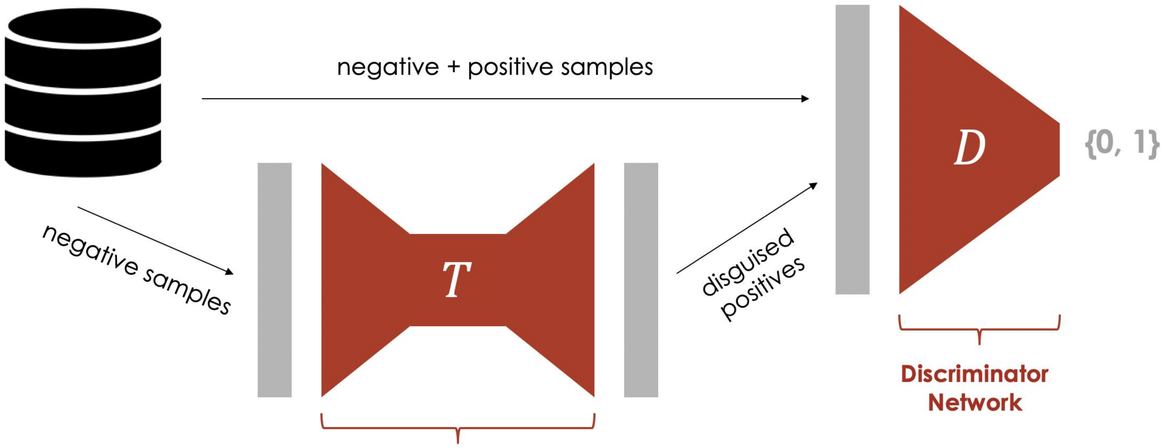
Data Science Institute
COLUMBIA UNIVERSITY

Derek Zhao, James Xue, Chaitanya Dara
Sai Pavan Kumar Unnam, Daniel Gomez
Industry Mentors: Prasad Chalasani, Aravind Sadagopan
Faculty Mentor: Eleni Drinea

Data Science Capstone Project
with
MediaMath

I. THE PROBLEM OF IMBALANCED DATA

A common issue in using machine learning to predict ad conversions in click-through rate datasets is the imbalanced classification problem: binary classifiers struggle to train effectively due to the lack of sufficient exposure to positive (minority) class samples. We explore the effectiveness of a novel neural architecture, the **Disguise Adversarial Network (DAN)**[1], a synthetic oversampling technique that transforms negative samples into positive samples, thus rebalancing the dataset.



$$\text{minimize } \mathcal{L}_1(T) = -\mathbb{E}_{x^- \sim P_{\Omega^-}}\left[\log D(T(x^-))\right] + \lambda \|T(x^-) - x^-\|_1$$

Want $D(T(x^-)) = 1$ — **Disguise Quality** assesses how well T transforms negative samples to look like positive samples

Want $T(x^-) \approx x^-$ — **Regularizer** restricts T from making too drastic of transformations to negative samples

$$\text{minimize } \mathcal{L}_2(D) = -\mathbb{E}_{x^- \sim P_{\Omega^-}}\left[\log(1 - D(x^-))\right] - \mathbb{E}_{x^+ \sim P_{\Omega^+}}\left[\log(D(x^+))\right] + \eta \mathbb{E}_{x^- \sim P_{\Omega^-}}\left[H\left(D(T(x^-))\right)\right]$$

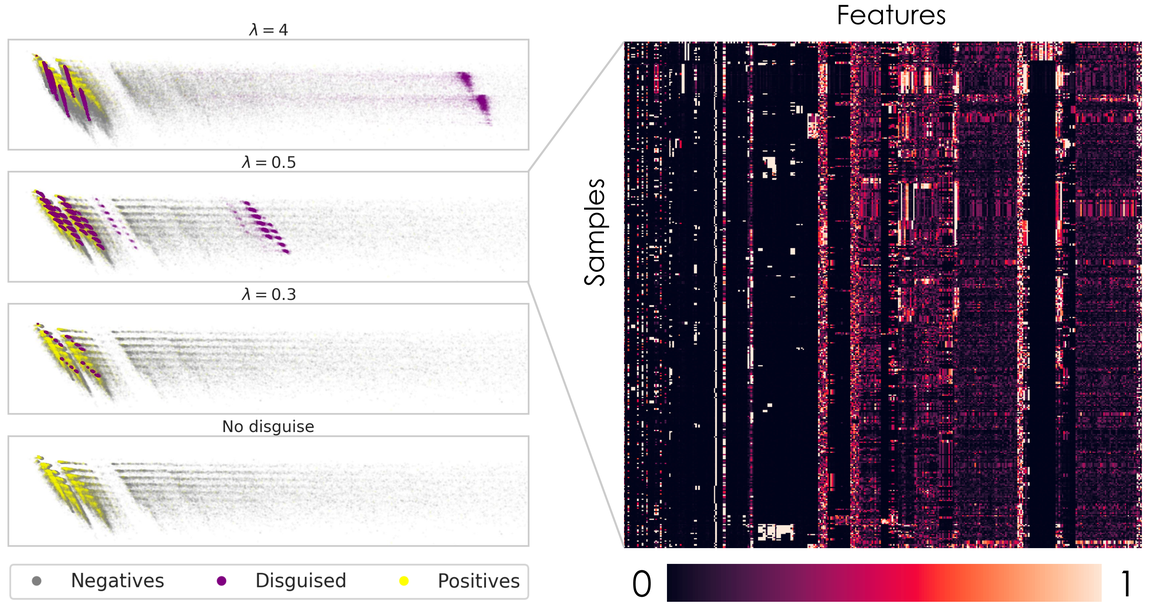
Want $D(x^-) = 0$
Want $D(x^+) = 1$
Want $D(T(x^-)) \approx$ 0 or 1 but not 0.5

where $H(\hat{y}) = -[\hat{y}\log(\hat{y}) + (1 - \hat{y})\log(1 - \hat{y})]$

Entropy of Predictive Distribution
measures the degree of uncertainty in D's predictions on disguised samples

II. VISUAL INTUITIONS

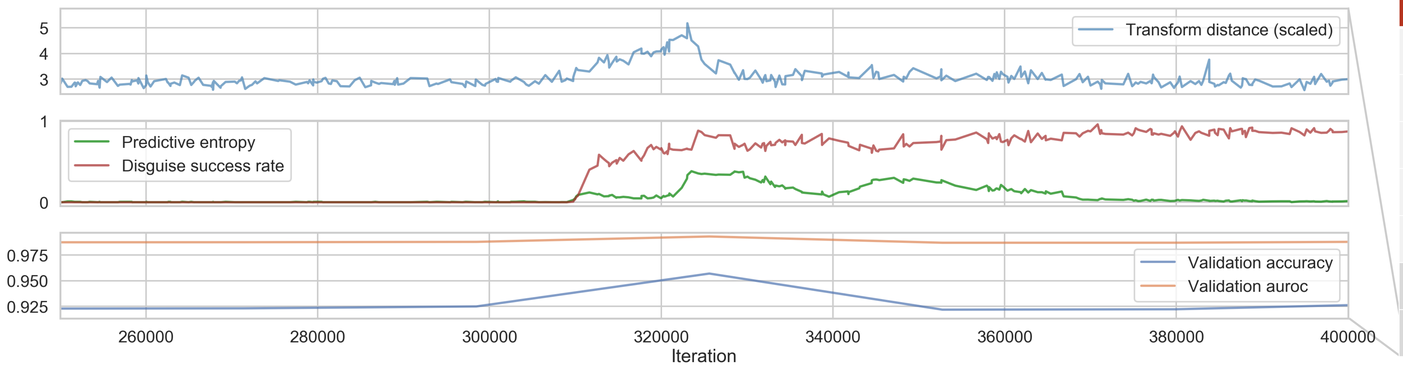
The disguise network attempts to learn a transformation on negative class data that satisfies two properties: 1) negative samples are transformed to look like positive samples and 2) the transformation is not too drastic. The hyperparameter λ balances these two competing interests, with higher values encouraging the disguise network to learn an identity transformation and lower values affording the network greater flexibility at the cost of reduced disguise diversity. This effect can be seen on the left using MediaMath's advertising data projected onto two dimensions via PCA.



The rightmost figure shows a heatmap of scaled transformation magnitudes across 400 random samples of negative data. Certain features are noticeably more relevant for performing successful transformations than others. The disguise network can be a powerful tool for inferring sample-specific feature importance.

III. MIXED RESULTS

The DAN does not appear to offer benefits in tasks where the data is linearly separable or where base models already predict with high recall. To account for this, we further skew MediaMath's training data from 11% positive to 1% positive while leaving validation and test sets unchanged. Under these conditions, the DAN is able to approximate the performance of models trained on the original dataset.



Positive class proportion (train)	Disg.	Discr.	Accuracy	AUROC	Precision	Recall
0.11	No	Logistic	**0.9173**	**0.9844**	0.5795	0.9789
0.11	No	64 32	**0.9173**	0.9843	**0.5800**	0.9754
0.11	256 x 4	64 32	0.9168	**0.9844**	0.5794	0.9653
0.5	No	Logistic	0.9155	0.9650	0.5732	0.9901
0.5	No	64 32	0.9155	0.9668	0.5728	**0.9933**
0.01	No	64 32	0.9073	0.9809	**0.6568**	0.3768
0.01	256 x 4	64 32	**0.9181**	**0.9838**	0.5899	**0.9041**

[1] Deng, Yue & Shen, Yilin & Jin Hongxia, 'Disguise Adversarial Networks for Click-through Rate Prediction', in *Proceedings of the Twenty-Sixth Joint Conference on Artificial Intelligence*, 2017.