Discriminative Regularization for Latent Variable Models with Applications to Electrocardiography

Andrew C. Miller [1] Ziad Obermeyer [2] John P. Cunningham [3] Sendhil Mullainathan [4]

Abstract

Generative models often use latent variables to represent structured variation in high-dimensional data, such as images and medical waveforms. However, these latent variables may ignore subtle, yet meaningful features in the data. Some features may predict an outcome of interest (e.g. heart attack) but account for only a small fraction of variation in the data. We propose a generative model training objective that uses a black-box discriminative model as a regularizer to learn representations that preserve this predictive variation. With these discriminatively regularized latent variable models, we visualize and measure variation in the data that influence a black-box predictive model, enabling an expert to better understand each prediction. With this technique, we study models that use electrocardiograms to predict outcomes of clinical interest. We measure our approach on synthetic and real data with statistical summaries and an experiment carried out by a physician.

1. Introduction

Machine learning research has led to extraordinary prediction results across many data types but less success in deriving human-level insight from (or incorporating that insight into) our models. Two such problems include *model diagnosis*: using expert human knowledge to understand and improve upon model shortcomings; and *model-based discovery*: understanding what human-interpretable features drive predictive performance. To be concrete, consider these biomedical applications:

(a) A newly trained model detects a common heart condition (e.g. atrial fibrillation or *afib*, a condition observable in an EKG) with reasonable accuracy, but below the level of a skilled cardiologist. How can we use human expertise to diagnose model shortcomings?

(b) By contrast, using EKG waveforms as inputs to a black-box predictor can outperform a cardiologist in predicting future cardiac events, such as heart attacks. It is critical to understand what EKG features that predictor is using, insight which can be used to motivate future experiments, treatments, and care protocols.

Though these scenarios are specific, they raise a general problem. Given a black-box predictive model $m(\mathbf{x})$ (e.g. a neural network) that converts a high-dimensional signal $\mathbf{x}$ (e.g. an EKG) into a prediction score of output $\mathbf{y}$ (e.g. afib), we would like to show domain experts what the algorithm "sees" — what variation in $\mathbf{x}$ influences predictions of $\mathbf{y}$.

In recent years, model interpretability has received considerable attention (Ribeiro et al., 2016; Doshi-Velez & Kim, 2017; Kindermans et al., 2017; Narayanan et al., 2018). One way to interrogate a predictive model is to examine perturbations of an input $\mathbf{x}$ that influence the prediction the most.[1] For example, we can compute the gradient of $m(\mathbf{x})$ with respect to model input (e.g. pixels) at point some data point $\mathbf{x}$. This will produce a *saliency map* that identifies which parts of the input most influence the outcome of interest (Baehrens et al., 2010; Simonyan et al., 2013). One shortcoming of this approach is that the gradient contains only local information; for high-dimensional structured inputs, following the gradient flow for any moderate distance will produce meaningless inputs that bear little resemblance to real observations.

We can improve upon saliency maps by restricting our perturbations to remain on the "data manifold." For instance, the *activation maximization* approach uses a deep generative model to represent the space of natural images with a latent variable $\mathbf{z}$ (Nguyen et al., 2016). To interrogate an input, a new observation is synthesized by finding the $\mathbf{z}$ that maximizes the prediction score for some classifier of interest. This data manifold constraint enables the visualization of realistic image features that influence the prediction.

[1] Data Science Institute, Columbia University, New York, NY, USA. [2] School of Public Health, UC Berkeley, Berkeley, CA, USA. [3] Department of Statistics, Columbia University, New York, NY, USA. [4] Booth School of Business, University of Chicago, Chicago, IL, USA.. Correspondence to: ACM <am5171@columbia.edu>.

Proceedings of the 36th International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

[1] For clarity, we will always use the term *prediction* to refer to the output of $m(\mathbf{x})$. Further, we assume the discriminative model $m(\mathbf{x})$ has been given to us — its parameters have been fit and fixed.

One issue with this approach is that a generative model may ignore subtle, yet important features of the data. Generative models fit with maximum likelihood optimize for *reconstruction error*, which focuses on representing directions of high variation. This can be problematic if the variation we are interested in visualizing and exploring are *predictive features* of $\mathbf{x}$ — potentially subtle, structured variation in the data that influence $m(\cdot)$ the most. We show that the standard (approximate) maximum likelihood objective can lead to under-representation of predictive features in $\mathbf{x}$.

To address these issues, we propose the *discriminitively regularized variational autoencoder* (DR-VAE), a simple way to constrain latent variables to represent the variation that a black-box model $m(\mathbf{x})$ uses to form predictions. To fit a DR-VAE, we augment the typical maximum likelihood-based objective with a regularization term that depends on the discriminative model $m(\mathbf{x})$. This constraint gives the user more control over the semantic meaning of the latent representation even when we train on a separate set of unlabeled data.

The contributions of this work are: (i) a constrained generative model objective that retains targeted, semantically meaningful properties;[2] (ii) an information theoretic interpretation of the regularizer; (iii) an empirical study of tradeoffs in a synthetic, a toy data, and a real setting with EKGs that demonstrates the advantages of incorporating such constraints; (iv) the use of model-morphed data to interpret black-box predictors in a real biomedical informatics application; and (v) validation of the technique using statistical summaries and physician assessment.

2. Background and Problem Setup

We are given an already-trained and fixed discriminative model $m : \mathcal{X} \mapsto \mathbb{R}$ that maps a data point $\mathbf{x} \in \mathcal{X}$ to a real-valued prediction score (e.g. the conditional log-odds for a binary classifier) for an outcome of interest $\mathbf{y}$. In our motivating example, $\mathbf{x}$ is a D-dimensional EKG, where $D = 300$ (3 leads × 100 samples per lead) and $\mathbf{y} = 1$ indicates the patient exhibits “ST elevation,” a clinically important indicator of potential heart attack (myocardial ischemia). The discriminative model computes the conditional probability that a patient has ST elevation given the EKG observation:

$$m(\mathbf{x}) = \text{logit}(Pr(\mathbf{y} = 1 \,|\, \mathbf{x})) \triangleq \mathbf{s}\,. \tag{1}$$

An example of such a discriminative model $m(\cdot)$ is a neural network that transforms $\mathbf{x}$ into intermediate representations (e.g. hidden layers) and eventually a scalar prediction. We often ask how a neural network arrives at its predictions. Given a neural network model trained and validated on data from another hospital system but fails in our hospital setting, how do we diagnose these model shortcomings?

Popular approaches examine variation in $\mathbf{x}$ that change the prediction $m(\mathbf{x})$ by following the gradient of $m(\mathbf{x})$. One limitation, however, is that the gradient is local and ignores global structure in $\mathcal{X}$-space. Following the gradient flow defined by $\nabla_{\mathbf{x}} m(\mathbf{x})$ to increase the prediction score, we will generate physiologically implausible inputs (e.g. meaningless images or EKGs). We can constrain this exploration by restricting our search over $\mathcal{X}$ with a generative latent variable model (Nguyen et al., 2016).

Generative latent variable models A latent variable model is a probabilistic description of a high-dimensional data vector, $\mathbf{x}$, using a low-dimensional representation, $\mathbf{z}$. A common model for continuous-valued $\mathbf{x} \in \mathbb{R}^D$ uses a parametric mean function $g_{\boldsymbol{\theta}}(\mathbf{z})$, resulting in the model

$$\mathbf{z} \sim p(\mathbf{z})\,, \qquad \mathbf{x} \sim g_{\boldsymbol{\theta}}(\mathbf{z}) + \sigma \cdot \epsilon \tag{2}$$

where $\epsilon \sim \mathcal{N}(0, 1)$ and σ is the observation noise scale. When $g_{\boldsymbol{\theta}}(\cdot)$ is a deep neural network, we call this a *deep generative model* (DGM). One way to fit a DGM to data is the autoencoding variational Bayes (AEVB) framework (Kingma & Welling, 2013; Rezende et al., 2014). The AEVB framework fits DGMs via optimization of the *evidence lower bound* (ELBO), a tractable lower bound to the marginal likelihood of the data, $p(\mathbf{x} \,|\, \boldsymbol{\theta})$ (Jordan et al., 1999; Blei et al., 2017). For efficiency, AEVB introduces an *inference network* parameterized by a global variational parameter $\boldsymbol{\lambda}$ that specifies the posterior approximation as a conditional distribution $q_{\boldsymbol{\lambda}}(\mathbf{z} \,|\, \mathbf{x})$. Together, the generative network and inference network are referred to as a variational autoencoder (VAE). For a single observation $\mathbf{x}$, the VAE objective is

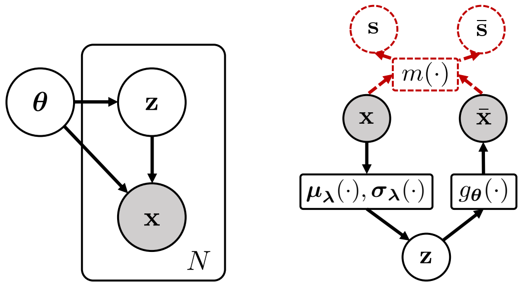
$$\mathscr{L}_{VAE}(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \mathbb{E}_{q_{\boldsymbol{\lambda}}(\mathbf{z}|\mathbf{x})}\left[\ln p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}) - \ln q_{\boldsymbol{\lambda}}(\mathbf{z}|\mathbf{x})\right] \leq \ln p_{\boldsymbol{\theta}}(\mathbf{x})\,.$$

We maximize this objective with respect to generative parameters $\boldsymbol{\theta}$ and inference network parameters $\boldsymbol{\lambda}$. A generative model captures the distribution of the data $\mathbf{x}$ via the latent representation $\mathbf{z}$. By varying $\mathbf{z}$, we can explore the *structural variation* in $\mathbf{x}$ (represented by the generative model) that influence $m(\mathbf{x})$.

3. Discriminative Regularization

Fitting a generative model requires choosing what variation in $\mathbf{x}$ is structure and what is noise. A deep generative model implicitly makes this choice by explaining variation in $\mathbf{x}$ with either $g_{\boldsymbol{\theta}}(\mathbf{z})$ (i.e. structure) or $\sigma \cdot \epsilon$ (i.e. noise). What is variation and what is noise, however, is not always obvious and not always easy to specify. A discriminative model $m(\mathbf{x})$ may be influenced by subtle features in $\mathbf{x}$. If the maximum likelihood objective targets reconstruction error (e.g. as a

[2] To be concrete, throughout we use “semantically meaningful information” to mean information used by the predictor to form prediction $m(\mathbf{x})$. We measure this with *discriminative reconstruction error* defined in Equation 19.


$\boldsymbol{\theta}$
$\mathbf{z}$
$\mathbf{x}$
N
$\mathbf{s}$
$\bar{\mathbf{s}}$
$m(\cdot)$
$\mathbf{x}$
$\bar{\mathbf{x}}$
$\boldsymbol{\mu}_{\boldsymbol{\lambda}}(\cdot), \boldsymbol{\sigma}_{\boldsymbol{\lambda}}(\cdot)$
$g_{\boldsymbol{\theta}}(\cdot)$
$\mathbf{z}$

(a) DGM (b) DR-VAE

Figure 1. Comparison of the VAE and DR-VAE. *Left*: Graphical representation of a DGM — low-dimensional latent variable $\mathbf{z}$ generates observed data $\mathbf{x}$ for all N observations. *Right*: Computational graph of a VAE (solid black) and a DR-VAE (additional dotted red). A VAE feeds data $\mathbf{x}$ through the recognition network (i.e. the "encoder") which yields an approximate posterior distribution over $\mathbf{z}$. A sample from this posterior is then mapped through the generative model $g_{\theta}(\mathbf{z})$ (i.e. the "decoder") producing the synthetic observation $\bar{\mathbf{x}}$ — a loss penalizes the divergence between $\mathbf{x}$ and $\bar{\mathbf{x}}$. The DR-VAE adds an additional penalty — the predictive function we care about $m(\cdot)$ must also be close for $\mathbf{x}$ and $\bar{\mathbf{x}}$. The strength of this penalty is determined by a tuning parameter $\beta > 0$.

Gaussian likelihood does), then subtle variation in $\mathbf{x}$ can be ignored — the optimization procedure pushes this variation into the noise term and $g_{\theta}(\mathbf{z})$ ignores these features.

Ignoring predictive features in $\mathbf{x}$ will prevent model-based exploration of $m(\mathbf{x})$ with latents $\mathbf{z}$ — $g_{\boldsymbol{\theta}}(\cdot)$ and $\mathbf{z}$ will not be able to reconstruct certain features in $\mathbf{x}$ that are important to $m(\mathbf{x})$ because the generative model is focused on reconstructing features that are unimportant to $m(\mathbf{x})$.

Additionally, deep generative models can be overly flexible. This flexibility results in an equivalence class of generative models with similar likelihood performance, but with different reconstruction properties. Our goal is learn a DGM that uses this flexibility to represent potentially subtle, yet predictive features with the latent variable $\mathbf{z}$. We propose maximizing the following objective

$$\mathscr{L}^{(m)}_{\textit{DR-VAE}}(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \underbrace{\mathscr{L}_{VAE}(\boldsymbol{\theta}, \boldsymbol{\lambda})}_{\text{model likelihood}} - \underbrace{\beta \cdot \mathscr{L}^{(m)}_{disc.}(\boldsymbol{\theta}, \boldsymbol{\lambda})}_{\text{discrim. regularizer}} . \tag{3}$$

We refer to a DGM fit with this objective as a *discriminatively regularized VAE* (DR-VAE). This objective augments the typical evidence lower bound with a penalty on parameters that do not result in good *discriminative reconstruction* for predictor $m(\mathbf{x})$. This regularizer penalizes differences in the discriminative model score between real data and reconstructed data

$$\mathscr{L}^{(m)}_{disc.}(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \mathbb{E}_{\mathbf{z} \sim q_{\lambda}(\mathbf{z} \mid \mathbf{x})}\left[D\left(m(\mathbf{x}) \,||\, m(\bar{\mathbf{x}})\right)\right] \tag{4}$$

$$\bar{\mathbf{x}} \triangleq g_{\boldsymbol{\theta}}(\mathbf{z}) . \tag{5}$$

Here $D(\cdot \,||\, \cdot)$ is a divergence function whose specific form depends on the output of $m(\cdot)$ — we consider binary and continuous outcomes below. The discriminative penalty can be viewed as a type of posterior predictive check (Gelman et al., 2013). During model training, this term ensures that simulated data from our approximate posterior encode features we care about — the discriminative model predictions.

Note that to optimize Equation 3 with gradient-based updates, we will need to compute the gradient of $m(\mathbf{x})$ with respect to the input — a similar requirement for saliency maps (Baehrens et al., 2010; Simonyan et al., 2013) and activation maximization (Nguyen et al., 2016).

Binary outcomes For binary outcomes, the discriminative model $m(\mathbf{x})$ outputs the conditional probability $m(\mathbf{x}) = Pr(\mathbf{y} = 1 \mid \mathbf{x})$, a scalar value. To constrain the difference between $m(\mathbf{x})$ and $m(\bar{\mathbf{x}})$ we penalize the Kullback-Leibler divergence between the two conditional distributions

$$D_{KL}(q \,||\, p) = q \cdot \ln \frac{q}{p} + (1 - q) \cdot \ln \frac{1 - q}{1 - p} , \tag{6}$$

where $q = m(\mathbf{x})$ and $p = m(\bar{\mathbf{x}})$. We also consider the squared loss in the logit-probability values

$$D_{logit}(q \,||\, p) = \left(\ln \frac{q}{1 - q} - \ln \frac{p}{1 - p} \right)^2 . \tag{7}$$

Continuous outcomes For models that predict a continuous outcome, the predictive model encodes the conditional expectation, $m(\mathbf{x}) = \mathbb{E}[\mathbf{y} \mid \mathbf{x}]$. Assuming the model is homoscedastic, the predictive variance $\mathbb{V}(\mathbf{y} \mid \mathbf{x}) = \sigma^2$ is the same for all $\mathbf{x}$. Further assuming the outcome is conditionally Gaussian distributed, the KL-divergence between the predictive distribution $m(\mathbf{x})$ and $m(\bar{\mathbf{x}})$ is simply a function of the squared difference in expectations

$$D_{KL}(m(\mathbf{x}) \,||\, m(\bar{\mathbf{x}})) = \frac{1}{2\sigma^2} \left(m(\mathbf{x}) - m(\bar{\mathbf{x}})\right)^2 \tag{8}$$

$$= \frac{1}{2\sigma^2} \left(\mathbb{E}[\mathbf{y} \mid \mathbf{x}] - \mathbb{E}[\mathbf{y} \mid \bar{\mathbf{x}}]\right)^2 . \tag{9}$$

Valid evidence lower bound The divergence functions described above are all bounded below by zero (and only equal to zero when $m(\mathbf{x})$ and $m(\bar{\mathbf{x}})$ are equal). Fixing the regularization coefficient $\beta > 0$ ensures the loss remains a valid lower bound to the marginal likelihood

$$\mathscr{L}^{(m)}_{\textit{DR-VAE}}(\boldsymbol{\theta}, \boldsymbol{\lambda}) \leq \ln p_{\boldsymbol{\theta}}(\mathbf{x}) . \tag{10}$$

This eliminates potential optimization pathologies (e.g. the optimizer only maximizing the regularization term).

Relationship to mutual information For additional intuition, we can relate the discriminative penalty to information theoretic quantities. The conditional mutual information is

$$I(\mathbf{x}; \mathbf{y} \mid \mathbf{z}) = \mathbb{E}_{p(\mathbf{x}, \mathbf{z})}\left[D_{KL}(p(\mathbf{y} \mid \mathbf{z}, \mathbf{x}) \,||\, p(\mathbf{y} \mid \mathbf{z}))\right] . \tag{11}$$

We can interpret the discriminative penalty as

$$\mathcal{L}_{disc.}(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \frac{1}{N}\sum_{n=1}^{N} \mathbb{E}_{\mathbf{z}\sim q_{\lambda}(\mathbf{z}\,|\,\mathbf{x}_n)}\left[D_{KL}(m(\mathbf{x})\,||\,m(\bar{\mathbf{x}}))\right] \quad (12)$$

$$\approx \mathbb{E}_{p(\mathbf{x})q_{\lambda}(\mathbf{z}\,|\,\mathbf{x})}\left[D_{KL}\left(p(\mathbf{y}\,|\,\mathbf{x})\,||\,p(\mathbf{y}\,|\,\mathbf{z})\right)\right] \quad (13)$$

$$= I(\mathbf{x};\mathbf{y}\,|\,\mathbf{z})\,. \quad (14)$$

where we have made two assumptions: (i) $\mathbf{y}$ is conditionally independent of $\mathbf{z}$ given $\mathbf{x}$ (e.g. $\mathbf{z}$ adds no information about $\mathbf{y}$ above $\mathbf{x}$), and (ii) the term $m(\bar{\mathbf{x}})$ is approximately

$$p(\mathbf{y}\,|\,\mathbf{z}) = \mathbb{E}_{p(\mathbf{x}\,|\,\mathbf{z})}\left[p(\mathbf{y}\,|\,\mathbf{x})\right] \approx p(\mathbf{y}\,|\,\mathbb{E}_{p(\mathbf{x}\,|\,\mathbf{z})}\left[\mathbf{x}\right]) = m(\bar{\mathbf{x}})\,.$$

The discriminative regularizer can be interpreted as a penalty on the approximate conditional mutual information between $\mathbf{x}$ and $\mathbf{y}$ given $\mathbf{z}$. This interpretation offers a key insight: minimizing the value of Equation 14 encourages the generative model to minimize the informativeness of $\mathbf{x}$ about $\mathbf{y}$ when we know the value of $\mathbf{z}$ — this forces $\mathbf{y}$ to be explained by $\mathbf{z}$ alone and not by variation described by $\sigma \cdot \epsilon$.

3.1. Model Morphings

A generative model for data $\mathbf{x} \in \mathcal{X}$ with latent variable $\mathbf{z}$ equips us with new tools. We can use this representation to find similar patients (e.g. patients close in $\mathbf{z}$-space). We can compute the probability of new observations, enabling us to cluster patients, detect outliers, or compare models.

We can also use the model to synthesize new samples — a task we focus on in this section. Given a generative model $g_{\boldsymbol{\theta}}(\mathbf{z})$, we can manipulate $\mathbf{z}$ in various ways to generate new instances of data. Synthesizing new data can help us understand the patient-specific features that are predictive of a particular outcome. This may enable the discovery of new features or reveal the physiological roots of a phenomenon that can be targeted with more precise measurements, therapies, or further analysis. Synthesized data can also be used for pedagogical purposes, highlighting subtle features identifiable to an expert but difficult to discern for a novice. And as stated before, synthesized data can help us understand what features are important to a black-box predictor. However, for any of these efforts to be successful, that latent space must encode semantically meaningful variation with respect to the outcome of interest.

To explore features of $\mathbf{x}$ important to predictor $m(\cdot)$, we propose following the gradient of $m(\cdot)$ with respect to the input along the subspace $\bar{\mathcal{X}} = \{\mathbf{x} : \mathbf{x} = g_{\boldsymbol{\theta}}(\mathbf{z}), \mathbf{z} \in \mathbb{R}^K\}$. Starting at some $\mathbf{x}^{(0)}$ with some $\mathbf{z}^{(0)}$, we trace out a *model-morphed* trajectory by recursively computing

$$\mathbf{z}^{(t+1)} \leftarrow \mathbf{z}^{(t)} + \delta \cdot \frac{\partial m}{\partial \mathbf{x}}\frac{\partial \mathbf{x}}{\partial \mathbf{z}}(\mathbf{z}^{(t)}) \quad (15)$$

$$\tilde{\mathbf{x}}^{(t+1)} \leftarrow g_{\boldsymbol{\theta}}(\mathbf{z}^{(t+1)}) \quad (16)$$

for some small step size δ over T steps. This yields a model-morphed sequence, $\tilde{\mathbf{x}}^{(0)}, \ldots, \tilde{\mathbf{x}}^{(T)}$, that gradually increases the predictive score $m(\cdot)$ (or decreases if $\delta < 0$). By construction, each element in the sequence is constrained to be on the data manifold defined by the generative model. This allows us to move along long paths away from the initial datum $\mathbf{x}^{(0)}$. Further, because the generative model has been constrained to reproduce predictive features, we will see that the DR-VAE model-morphings can capture much more predictive variation than the standard VAE.

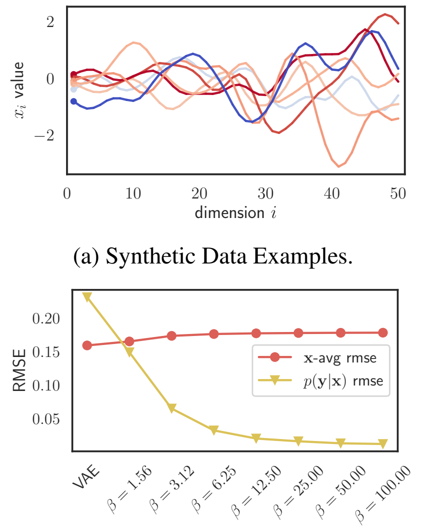
3.2. Related Work

Trade-offs between generative and discriminative models for classification performance are well-studied (Ng & Jordan, 2002; Bouchard & Triggs, 2004). Developing generative models with good discriminative properties is also an active area of research. Recent research has explored the use of discriminative models to influence generative model training and performance. As an example, Hughes et al. (2018) present an approach that constrains generative model learning to maintain good discriminative proprieties, and apply them to a commonly used latent factor model.

Incorporating the activations from discriminative models into generative model training has been explored as a method to produce sharper natural image samples (Lamb et al., 2016), and improve classification performance (Kuleshov & Ermon, 2017).

Interpreting complex predictive models is another area of related research. Work built on saliency maps (Simonyan et al., 2013) and plug-and-play activation maximization (Nguyen et al., 2016) use similar ideas to explore latent spaces of generative models. For instance, (Killoran et al., 2017) extend the activation maximization framework to synthesize discrete DNA sequences with desired properties by using a pre-trained generative model. Similarly, (Gómez-Bombarelli et al., 2018) apply this model-based latent space optimization to generate novel molecular compounds. We extend these methods by incorporating a semantically meaningful constraint into the generative model training itself and show that this captures relevant information in $\mathbf{x}$ that may otherwise be ignored.

We also note that our approach is similar in spirit to CycleGANs (Zhu et al., 2017), which incorporates a *cycle consistency loss* into the objective. This cycle consistency loss ensures that the original image can be recovered after being translated from one domain to another. Similarly, our discriminative penalty ensures the original prediction can be recovered after the image has been encoded into a low-dimensional latent representation. Additional discussion of related work is in the supplement.



(a) Synthetic Data Examples.

(b) Error summary

Figure 2. The DR-VAE can trade little generative error for improved discriminative error. (a) Example data — 50-dimensional correlated Gaussian draws with a linear increase in marginal variance, $\Sigma_{1,1} = .1$ to $\Sigma_{50,50} = 1$. Predictive information is contained *entirely* in the first dimension (marked by ● on the left) — blue indicates lower probability and red indicates higher probability. (b) Generative (red ●) and discriminative (yellow ▼) test reconstruction error for a sequence of increasingly regularized models. As β grows, we trade some generative accuracy ($\mathbf{x}$-avg rmse) for substantial discriminative accuracy ($p(\mathbf{y}\,|\,\mathbf{x})$ rmse).

4. Experiments

4.1. Synthetic Data

We apply DR-VAEs to a synthetic data set that, by construction, has predictive power in the dimension of least variation. We generate data according to the following model

$$\mathbf{x} \sim \mathcal{N}(0, \Sigma)\,, \qquad \mathbf{y}\,|\,\mathbf{x} \sim \mathrm{Bern}(f(\mathbf{x}; \boldsymbol{\phi}))\,, \tag{17}$$

where $f(\mathbf{x}; \boldsymbol{\phi})$ is a multi-layer perceptron and Σ is a full rank, structured covariance matrix.

We construct the covariance matrix with two properties: (i) smooth sample paths, and (ii) increasing marginal variance. The Σ matrix is constructed such that the marginal variances (along the diagonal) increase; $\Sigma_{1,1} = .1$ increases linearly to $\Sigma_{D,D} = 1$. Examples of the synthetic data are visualized in Figure 2a and experiment details are in the supplement.

The black-box predictor, $f(\mathbf{x}; \boldsymbol{\phi})$, is a multi-layer perceptron that depends only on x_1, the first dimension of $\mathbf{x}$ and the dimension of least variation. Reconstruction of directions of maximal variation will only carry predictive information about $\mathbf{y}$ through correlations with x_1. In this synthetic setup, we examine the generative and discriminative reconstruction trade-off as we increase β.

We fit a linear variational autoencoder (i.e. factor analysis) to learn the distribution of $\mathbf{x}$ with discriminative regularization. We limit the latent $\mathbf{z}$ dimension to be $K < D$; we set $K = 10$ and $D = 50$ in this example. To accurately reconstruct the discriminative score, $p(\mathbf{y} = 1\,|\,\mathbf{x})$, the model of $\mathbf{x}$ needs to contain information about the first dimension of $\mathbf{x}$.

We optimize the objective in Equation 3 for increasing values of β and compute the generative reconstruction error (gen-err) and the discriminative reconstruction error (disc-err) on held out test data

$$\text{gen-err} = \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}\left[\frac{1}{D}\sum_d (\mathbf{x}_d - \bar{\mathbf{x}}_d)^2\right]^{1/2} \tag{18}$$

$$\text{disc-err} = \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}\left[(m(\mathbf{x}) - m(\bar{\mathbf{x}}))^2\right]^{1/2}. \tag{19}$$

In Figure 2b we visualize these errors as a function of the regularization strength. Note that we report discriminative error instead of predictive accuracy because the object of our study is not the outcome itself but the predictive model $m(\cdot)$. We want our generative model reconstructions to result in the same predictions (and including the same mistakes) from the discriminative model. With these highly structured Gaussian observations, we trade little generative reconstruction error to obtain nearly perfect discriminative reconstruction. The discriminative regularizer competes with the ELBO term for model capacity. When $\beta = 0$ (i.e. a VAE), the model ignores the directions of least variation. As we increase β, the discriminative regularizer competes with the ELBO, encouraging the generative model to represent the directions of variation that influence the discriminative model.

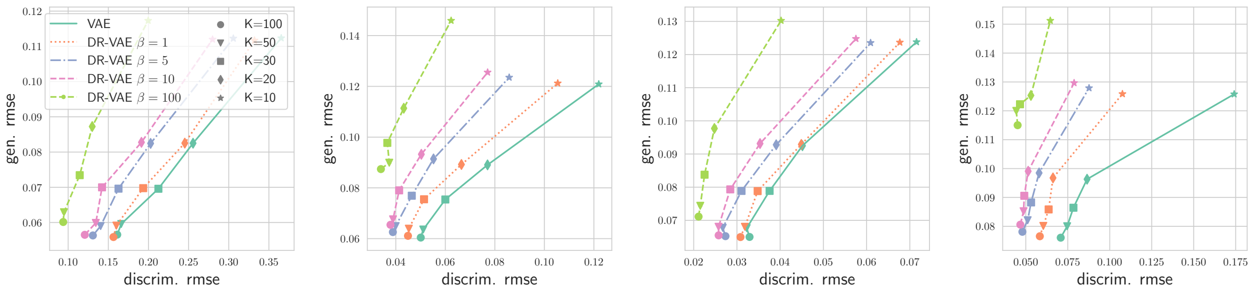
Digit Experiment For additional empirical analysis in a semi-synthetic setting, we include an additional experiment on a modified MNIST data set in the supplement.[3]

4.2. EKG Data

We now apply DR-VAEs to a data set of clinical EKGs, using tracings from three leads V1, II, and V5 (analogous to channels, each recording the heart's electrical activity from a different point on the chest), which contain a 10-second sequence of individual beats. We train discriminative models using segmented EKG beats (Christov, 2004), depicted in Figure 4, to predict the following outcomes

- ST elevation (binary): A subtle feature interpreted by eye that indicates heart attack. We use this outcome to illustrate model-morphings.
- Bundle Branch Block (binary): A delay or blockage of the electrical system in the heart that can be labeled from an EKG observation by eye.
- Major adverse cardiac events with six months (MACE) (binary): These events include myocardial infarction, car-

[3] Code available at `https://github.com/andymiller/DR-VAE`.



(a) Age

(b) Bundle Branch Block

(c) MACE

(d) ST Elevation

Figure 3. DR-VAEs contain more predictive information. Depicted are discriminative (horizontal) vs. generative (vertical reconstruction error on held out data for the four outcomes (age, bundle branch block, MACE, and ST elevation). A single trace represents a single value of β; shapes (e.g. ●, ■, or ⧫) depict different latent dimensionality. We can see in some examples that there is a *constant frontier* — we can achieve better discriminative reconstruction accuracy and sacrifice no generative reconstruction accuracy. The discriminative regularizer is pushing the generative model to prefer this more meaningful representation over other, equivalent generative models. We also note that predictive accuracy of the actual outcome $\mathbf{y}$ remains very similar using $\mathbf{x}$ and $\bar{\mathbf{x}}$.

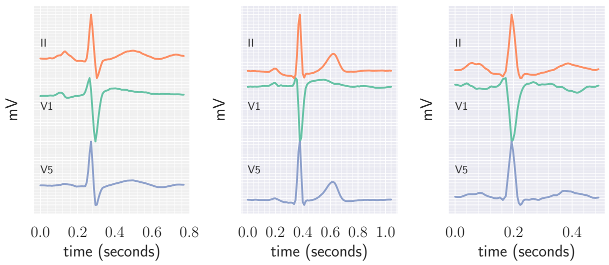


Figure 4. Example data: three EKG beats from three different patients. The EKG signal considered records three leads (V1, II, V5). Beats are segmented and placed on the unit interval.

diac arrest, stent, or a coronary artery bypass grafting. MACE risk is not predicted with EKGs.

- Age (continuous): Patient age has a noisy relationship to cardiac function. Visualizing the cardiac functional correlates of aging is potentially of clinical and scientific interest (Jones et al., 1990; Khane et al., 2011).

EKG data are a good fit for DR-VAEs as a subtle deviation from normal cardiac function can be predictive of a significant clinical outcome. Understanding ST elevation and bundle branch block predictors is model diagnosis category, while age and MACE predictors fall under model-based discovery (defined in Section 1).

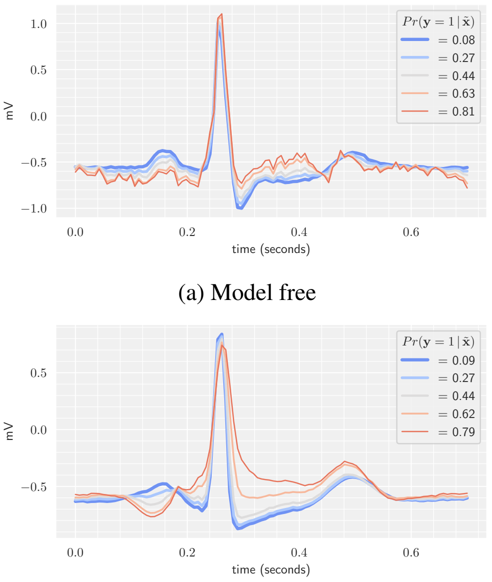
Data and model setup We construct a data set for each outcome with close to even base-rates of $\mathbf{y} = 1$. We split our cohort by patients into a training/validation development set (75% of patients) and a testing set (25% of patients) — no patients overlap in these two groups. We report predictive performance and model reconstruction statistics only on the held-out test patients. Further details and statistics of data used in each experiment are in the supplementary material.

For each outcome, discriminative model $m(\cdot)$ is a multi-layer perceptron classifier (or regressor) trained to minimize the cross entropy (or squared error) loss. Each discriminative model has two hidden layers of size 100 and a ReLU nonlinearity. Discriminative models are trained with dropout ($p = .5$) and stochastic gradient optimization with Adam (Kingma & Ba, 2014), starting with the learning rate set to .001 and halved every 25 epochs. We save the model with the best performance on the validation set.

Throughout our experiments we compare a standard VAE to the DR-VAE with values of $\beta = [1, 5, 10, 100]$ (note that a DR-VAE with $\beta = 0$ corresponds to a standard VAE). All deep generative models have one hidden layer with 500 units and a ReLU nonlinearity. We also train generative models with gradient-based stochastic optimization using Adam (Kingma & Ba, 2014), with an initial learning rate of .001 that is halved every 25 epochs. Accompanying code contains all model architecture and training details.

Discriminative vs. generative trade-offs For each outcome, we train all five generative models and compare both generative and discriminative reconstruction error. In Figure 3 we visualize the trade-off between generative and discriminative reconstruction error on held-out test data. In these experiments, there exists a trade-off between generative reconstruction and discriminative reconstruction. However, the standard VAE lies on a *nearly constant* part of the frontier — we can see this by comparing a shape (e.g. ●, ■, or ⧫) across the different traces (colored lines). Four of the models are essentially constant in terms of generative rmse from the VAE to DR-VAE with $\beta = 1$. For these outcomes, we can trade negligible generative performance for meaningful discriminative performance. With higher values of β, we see the trade-off emerge.

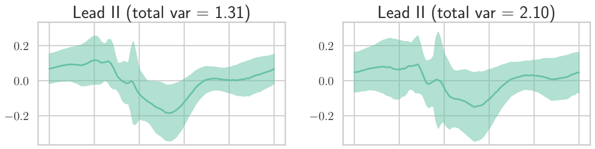
As a concrete example, for ST Elevation and latent size $K = 50$, the discriminative RMSE for the DR-VAE with $\beta = 1$ is 20% lower than the RMSE for a standard VAE, while



(a) Model free

(b) DR-VAE, $\beta = 5$

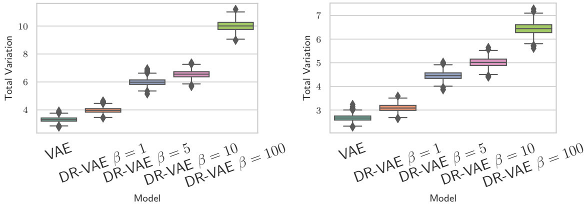
Figure 5. Model-based morphings can identify more meaningful features. Depicted are morphed examples (only lead II) comparing the (a) direct gradient and (b) DR-VAE with $\beta = 5$. The thick blue line indicates the starting EKG (low ST elevation), and the thin red line indicates the morphed EKG to $Pr(\mathbf{y}\,|\,\mathbf{x}) = .8$ (high ST elevation). Without a model, the EKG quickly starts to look unrealistic. The model preserves structure — smoothness, and characteristic EKG features.



(a) VAE ($\beta = 0$)

(b) DR-VAE, $\beta = 5$

Figure 6. DR-VAEs describe more predictive variation. Depicted is the mean morphing delta, $\mathbb{E}[\bar{\mathbf{x}} - \tilde{\mathbf{x}}]$ (and marginal variance $\hat{\Sigma}_{morph}$) for a single lead (II) while varying β. The expectation was estimated on 1024 test beats using the ST Elevation predictor. On the left, the standard VAE exhibits smaller morphing variation. As we increase β, the model morphs exhibit more variation, indicating the DR-VAE represents a richer set of predictive features than the standard VAE. Note that generative reconstruction is similar for these models.



(a) *low-to-high*.

(b) *high-to-low*.

Figure 7. DR-VAEs capture more predictive variation. Depicted is the trace of the $\hat{\Sigma}_{morph}$ matrix defined in Equation 20 variation by model. As we increase β, the latent space finds more ways (over the $N = 1024$ test examples) to increase (left) and decrease (right) the probability of ST elevation according to $m(\cdot)$. Note that the VAE and $\beta = 1, 5$, and 10 all have roughly the same generative reconstruction error (see Figure 3).

the generative reconstruction error remains identical (seen in Figure 3d). This improvement indicates that we have learned a representation $\mathbf{z}$ that contains more information about $m(\cdot)$ than the standard VAE.

Exploring z-space with morphings We qualitatively and quantitatively examine the difference between the suite of generative models trained on EKG observations. To compare the $\mathbf{z}$ space of each generative model, we examine morphed pairs of test data. For a generative model $g_{\boldsymbol{\theta}}(\cdot)$, to compute a morphed pair for test example $\mathbf{x}_n$ we

- compute $\bar{\mathbf{x}}_n = \boldsymbol{\mu}_{\boldsymbol{\lambda}}(\mathbf{x}_n)$ via the encoder
- update $\bar{\mathbf{x}}_n$ into $\tilde{\mathbf{x}}_n$ via morphing $\mathbf{z}$ with the operation defined in Equation 16 using model $g_{\boldsymbol{\theta}}(\cdot)$
- stop updating when $m(\tilde{\mathbf{x}})$ has reached a certain value (e.g. $m(\tilde{\mathbf{x}})$ reaches the second highest decile)

This creates a pair, $\bar{\mathbf{x}}_n$ and $\tilde{\mathbf{x}}_n$, the latter of which has morphed from the former according to the gradient of model $m(\cdot)$ and $g_{\boldsymbol{\theta}}(\cdot)$. Figure 5b depicts a model-morphing trajectory. The initial EKG starts at the thick blue line (low ST elevation) and is morphed to the thin red line (high ST elevation). By comparison, Figure 5a depicts a model free morph, the result of following the gradient flow of $m(\mathbf{x})$ which produces unrealistic synthetic examples — particularly non-smooth sample paths and the elimination of physiologically characteristic waves. We repeat for N examples in our test set, yielding a collection of morphed pairs.

A key question is, does a DR-VAE have more capacity to explore predictive variation than a standard VAE? Intuitively, if two models achieve equal generative performance, but model a has more morphing variation than model b, model a has more capacity to explore predictive variation than model b. One could imagine using an entropy estimator to measure this capacity. We use a simpler summary to quantify this, the empirical covariance of the difference between $\bar{\mathbf{x}}$ and $\tilde{\mathbf{x}}$ according to generative model $g_{\boldsymbol{\theta}}(\mathbf{z})$

$$\hat{\Sigma}_{morph} = \text{cov}\left(\{\bar{\mathbf{x}}_n - \tilde{\mathbf{x}}_n\}_{n=1}^N\right). \qquad (20)$$

The trace of $\hat{\Sigma}$ summarizes how much predictive variation model $g_{\boldsymbol{\theta}}(\cdot)$ has captured with respect to discriminative model $m(\mathbf{x})$. We can also examine the spectrum of $\hat{\Sigma}_{morph}$ for a clearer picture of our model-morphs — e.g. how many effective dimensions of $\mathcal{X} = \mathbb{R}^D$ are used?

We conduct two experiments that center on statistics of morphed pairs: (i) *low-to-high*: we take test examples from the lowest decile of $m(\mathbf{x})$ (e.g test patients with the lowest probability of ST elevation) and morph them *up* such that $m(\tilde{\mathbf{x}}) = .75$ (e.g. high probability of ST elevation); (ii) *high-to-low*: we take test examples from the highest decile of $m(\mathbf{x})$ (e.g. test patients with the highest probability of ST elevation) and morph them *down* such that $m(\tilde{\mathbf{x}}) = .1$.

For each direction (low-to-high and high-to-low) we compute morphed pairs for all five generative models. In Figures 7a and 7b we plot the total variation of $\hat{\Sigma}_{morph}$ for the ST elevation outcome. We can see that as β increases, so does the total variation in the morphed pairs. In fact the total number of dimensions with non-zero variation also increases with β (visualized in the supplemental material). In Figure 6, we show the marginal variance of the model-morphings as we increase β (for lead V1), summarizing the total lead variation in the title. As β grows, the marginal variance of the morph also grows, indicating higher variation in predictive features induced by the morph.

These summaries are evidence that the DR-VAE is able to capture a wider range of EKG features that indicate high ST elevation (according to the predictor $m(\mathbf{x})$). This tool can be used to reveal potentially unknown ways in which $m(\mathbf{x})$ behaves — allowing us to diagnose the model or trace a prediction to its physiological origins. Using just a VAE (as in Nguyen et al. (2016)) will fail to visualize much of the predictive variation in $\mathbf{x}$-space.

Comparing variation in observational data may capture correlated modes of variation — if two EKG features frequently co-occur, an information-constrained latent space will learn to express them together. If the predictor were trained with a non-confounded data set, and the generative model were trained with a confounded data set, our exploratory approach may not draw a distinction between them.

Physician validation We validate our generative model of EKGs with an expert labeling experiment. Our goal is to measure how real the model-based EKGs appear and how convincing model-morphed features are. We construct a two-alternative forced choice labeling task. We present a physician with N trials, where each trial compares a pair of EKG beats — one real and one fake (in a randomized order). The fake beat has been constructed with a ST elevation-regularized DR-VAE ($\beta = 5$) in one of four ways: (i) a $\bar{\mathbf{x}}$ with low ST elevation (second decile), (ii) a $\bar{\mathbf{x}}$ with high ST elevation (ninth decile), (iii) a $\tilde{\mathbf{x}}$ morphed from low-to-high ST elevation, and (iv) a $\tilde{\mathbf{x}}$ morphed from high-to-low ST elevation. The morphed examples are morphed from the second to the ninth decile (or vice versa). Here, the decile is with respect to values of the predictor $m(\cdot)$.

We present the expert with fifty of each category — 200 in total — with the task of identifying which EKG beat is real and which has ST elevation.

For the real-vs-fake question, the expert was able to distinguish slightly above random guessing, with an overall accuracy rate of 60% [54-66% CI]. For a more detailed picture, we compute accuracy rates (and 95% confidence intervals) for each type of synthetic data. For non-morphed types, (i) and (ii), we observe rates of 60% [46-74%] and 64% [50-78%], respectively. For morphed types, (iii) and (iv), we observe rates of 76% [64-88%] and (iv) 42% [28-56%], respectively.

Model-morphed examples are interpreted asymmetrically — high ST elevation examples morphed to look like low ST elevation examples are nearly indistinguishable from real data. However, examples morphed to have high ST elevation are more likely to be flagged as fake. This may be due to model smoothness becoming more conspicuous or poor EKG feature reconstruction. Overall, our model was able to fool an expert between 40% and 24% of the time, depending on the construction of the synthetic data.

For the feature-inducing accuracy, the expert label of ST elevation almost always matched the model-morphing induced feature. The low-to-high ST elevation morphed examples were correctly labeled 88% [78-96%] of the time, and the high-to-low morphed examples were correctly labeled 94% [88-100%]. This result is evidence that model-morphs can induce clinically relevant features that are recognizable and visually interpretable to an expert. We view these results as a promising first step toward model-based feature discovery and understanding for black-box EKG predictors. Further experiment details are in the supplement.

5. Discussion and Conclusion

We described discriminitively regularized VAEs (DR-VAEs), a method to constrain the representation of a deep generative model to contain targeted information about a black-box predictor of interest. We motivated the regularizer from an information theoretic perspective. We applied DR-VAEs to synthetic and EKG data, and empirically showed that DR-VAE representations can contain more predictive information than standard VAEs. We measured how realistic synthesized EKGs from the deep generative model are with an expert labeling experiment. While the results were promising, the task of further developing and validating richer generative models remains open.

In future work, we will use this technique to highlight features of an EKG that are predictive of various cardiovascular diseases. We hope to use model-morphs to relate predictions back to the underlying cardiac physiology driving the prediction. Along this same thread, we would like to develop flexible generative models an inductive bias rooted in cardiology, uniting physiological plausibility with DGM flexibility.

Acknowledgments

ACM and JPC are funded by NSF NeuroNex Award DBI-1707398 and The Gatsby Charitable Foundation GAT3419. The authors thank Scott Linderman for taking a close look at an early draft and CJ Chang for the tireless interpretation of synthetic and real EKG tracings.

References

Baehrens, D., Schroeter, T., Harmeling, S., Kawanabe, M., Hansen, K., and MÃžller, K.-R. How to explain individual classification decisions. *Journal of Machine Learning Research*, 2010.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 2017.

Bouchard, G. and Triggs, B. The tradeoff between generative and discriminative classifiers. In *16th IASC International Symposium on Computational Statistics (COMPSTAT'04)*, pp. 721–728, 2004.

Christov, I. I. Real time electrocardiogram QRS detection using combined adaptive threshold. *Biomedical engineering online*, 2004.

Doshi-Velez, F. and Kim, B. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

Gelman, A., Stern, H. S., Carlin, J. B., Dunson, D. B., Vehtari, A., and Rubin, D. B. *Bayesian data analysis*. Chapman and Hall/CRC, 2013.

Gómez-Bombarelli, R., Wei, J. N., Duvenaud, D., Hernández-Lobato, J. M., Sánchez-Lengeling, B., Sheberla, D., Aguilera-Iparraguirre, J., Hirzel, T. D., Adams, R. P., and Aspuru-Guzik, A. Automatic chemical design using a data-driven continuous representation of molecules. *ACS central science*, 2018.

Hughes, M. C., Hope, G., Weiner, L., McCoy Jr, T. H., Perlis, R. H., Sudderth, E., and Doshi-Velez, F. Semi-supervised prediction-constrained topic models. In *International Conference on Artificial Intelligence and Statistics*, 2018.

Jones, J., Srodulski, Z., and Romisher, S. The aging electrocardiogram. *The American journal of emergency medicine*, 1990.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. An introduction to variational methods for graphical models. *Machine learning*, 1999.

Khane, R. S., Surdi, A. D., and Bhatkar, R. S. Changes in ECG pattern with advancing age. *Journal of basic and clinical physiology and pharmacology*, 2011.

Killoran, N., Lee, L. J., Delong, A., Duvenaud, D., and Frey, B. J. Generating and designing DNA with deep generative models. *arXiv preprint arXiv:1712.06148*, 2017.

Kindermans, P.-J., Hooker, S., Adebayo, J., Alber, M., Schütt, K. T., Dähne, S., Erhan, D., and Kim, B. The (un) reliability of saliency methods. *arXiv preprint arXiv:1711.00867*, 2017.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Kuleshov, V. and Ermon, S. Deep hybrid models: Bridging discriminative and generative approaches. In *Proceedings of the Conference on Uncertainty in AI*, 2017.

Lamb, A., Dumoulin, V., and Courville, A. Discriminative regularization for generative models. *arXiv preprint arXiv:1602.03220*, 2016.

Narayanan, M., Chen, E., He, J., Kim, B., Gershman, S., and Doshi-Velez, F. How do humans understand explanations from machine learning systems? An evaluation of the human-interpretability of explanation. *arXiv preprint arXiv:1802.00682*, 2018.

Ng, A. Y. and Jordan, M. I. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In *Advances in neural information processing systems*, 2002.

Nguyen, A., Dosovitskiy, A., Yosinski, J., Brox, T., and Clune, J. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. In *Advances in Neural Information Processing Systems*, 2016.

Rezende, D. J., Mohamed, S., and Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*, 2014.

Ribeiro, M. T., Singh, S., and Guestrin, C. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

Simonyan, K., Vedaldi, A., and Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision*, 2017.