Deep Hybrid Models: Bridging Discriminative and Generative Approaches

Volodymyr Kuleshov, Stefano Ermon

Department of Computer Science, Stanford University

Highlights

We propose a new framework for training hybrid models based on coupling latent variables.

- Our framework offers greater modeling flexibility.
- It can handle complex models (incl. LV models)
- It is compatible with modern deep learning models
- Improves semi-supervised accuracy.

Instantiating the framework with neural networks gives rise to **deep hybrid models**.

Hybrids via Parameter Coupling

McCallum et al. (2006) propose new objective:

1. User specifies a joint probability model $p(x, y)$.
2. We maximize the *multi-conditional likelihood*

$$\mathcal{L}(x, y) = \alpha \cdot \log p(y|x) + \beta \cdot \log p(x).$$

where $\alpha, \beta > 0$ are hyper-parameters.

Observe that:

- When $\alpha = \beta = 1$, we have a generative model.
- When $\beta = 0$, we have a discriminative model.

Bayesian Parameter Coupling

The coupling prior objective approach (Lasserre, Bishop, Minka, 2006) optimizes the model

$$p(x, y, \theta_d, \theta_g) = p_{\theta_d}(y|x) p_{\theta_g}(x) p(\theta_d, \theta_g),$$

where the parameter coupling prior has the form

$$\log p(\theta_d, \theta_g) = \lambda ||\theta_d - \theta_g||$$

for some $|| \cdot ||$ and hyper-parameter $\lambda > 0$.

- $\lambda = 0$ yields a discriminative model
- As $\lambda \to \infty$ we get a generative model

Limitations of Existing Approaches

Crucially, both approaches work because $p(y|x)$, $p(x)$ *share weights*!

This poses two types of limitations:

- **Modeling**: e.g. can we make $p(y|x)$ be a convolutional neural network and $p(x)$ a VAE?
- **Computational**: marginal $p(x)$, posterior $p(y|x)$ need to be tractable

Generative Models

A generative model p specifies a joint probability $p(x, y)$ over both x and y.

Example: Naive Bayes

- Provides a richer prior
- Admits general queries (e.g. imputing features x)

Discriminative Models

A discriminative model p specifies a conditional probability $p(y|x)$ over y, given an x.

Example: Logistic regression.

- Lower asymptotic error
- Focus on prediction; fewer modeling assumptions

It well well-known that both Naive Bayes and logistic regression have a linear decision boundary

The difference is only training objective! It make sense to optimize between the two.

A New Framework For Hybrid Models By Coupling Latent Variables

1. User specifies p with a generative and a discriminative component and latent z

$$p(x, y, z) = p(y|x, z) \cdot p(x, z).$$

The $p(y|x, z)$, $p(x, z)$ can be very general; they only share latent z, not parameters!

2. We train both components using a multi-conditional objective

$$\alpha \cdot \mathbb{E}_{q(x,y)} \mathbb{E}_{q(z|x)} \underbrace{\ell\left(y, p(y|x, z)\right)}_{\text{discriminative loss } (\ell_2, \log)} + \beta \cdot \underbrace{D_f \left[q(x, z) || p(x, z)\right]}_{\text{f-divergence (KL, JS)}}$$

where $q(x, y)$ is data distribution and $\alpha, \beta > 0$ are hyper-parameters.

An Application: Deep Hybrid Models

Instantiating our framework with neural networks gives rise to deep hybrid models.

Explicit Density Models

- We maximize marginal multi-conditional log-likelihood

$$\log \int_{z \in \mathcal{Z}} p(y|x, z)^\gamma p(x, z) dz \geq \mathcal{L}.$$

- Applying the variational principle, we obtain:

$$\mathcal{L} = \mathbb{E}_{q(z|x)} \left[\gamma \log p(y|x, z) + \log p(x, z) - \log q(z|x)\right].$$

- This is a special case of our framework with:

$$L_D = \text{expected log loss} \quad L_G = \text{KL}\left(q(x, z) || p(x, z)\right).$$

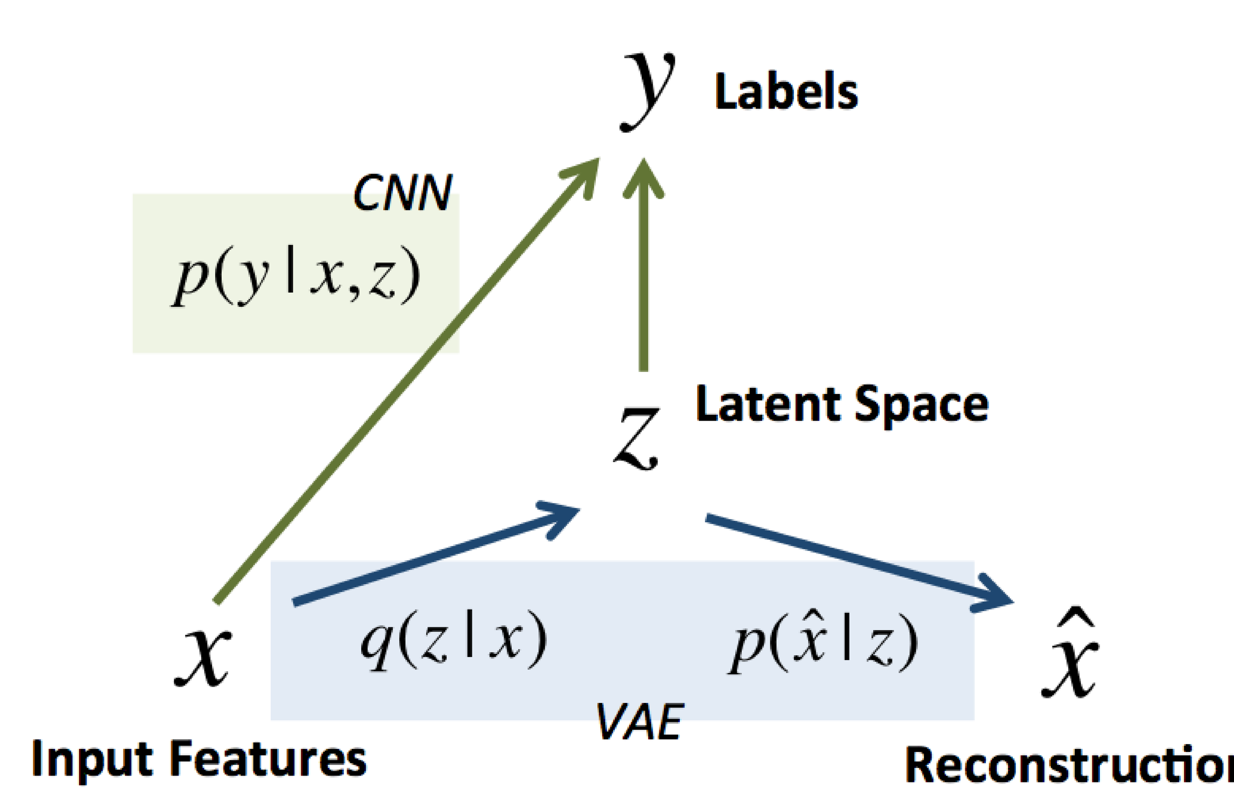
Implicit Density Models

- We may also choose $p(x, z)$ to be a GAN. Then:

$$L_D = \text{expected log loss} \quad L_G = \text{JS}\left(q(x, z) || p(x, z)\right).$$

- We use a discriminator D to optimize L_G

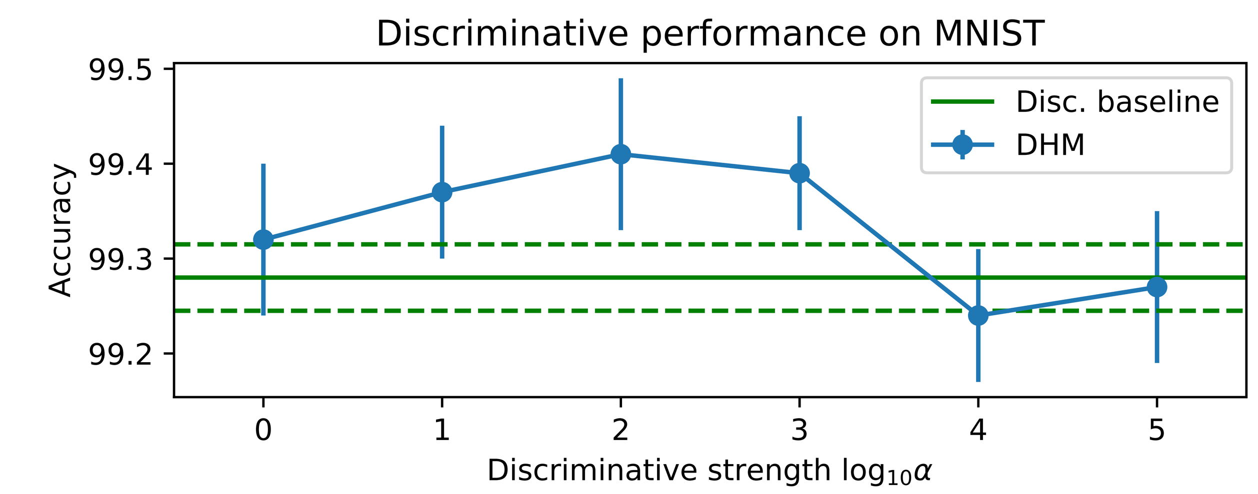
$$L_G \approx \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{q(z|x_i)} \log D(x_i, z) + \mathbb{E}_{p(x,z)} \log(1 - D(x, z))$$

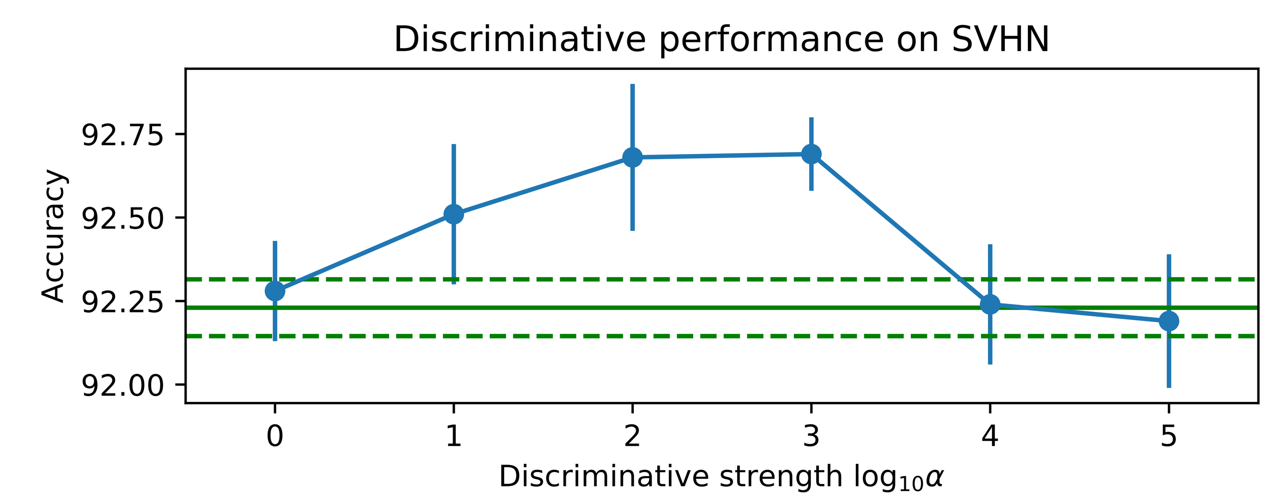


- We learn latent z useful for gen. and disc. tasks
- This is a form of multi-task learning
- It regularizes the model and improves features

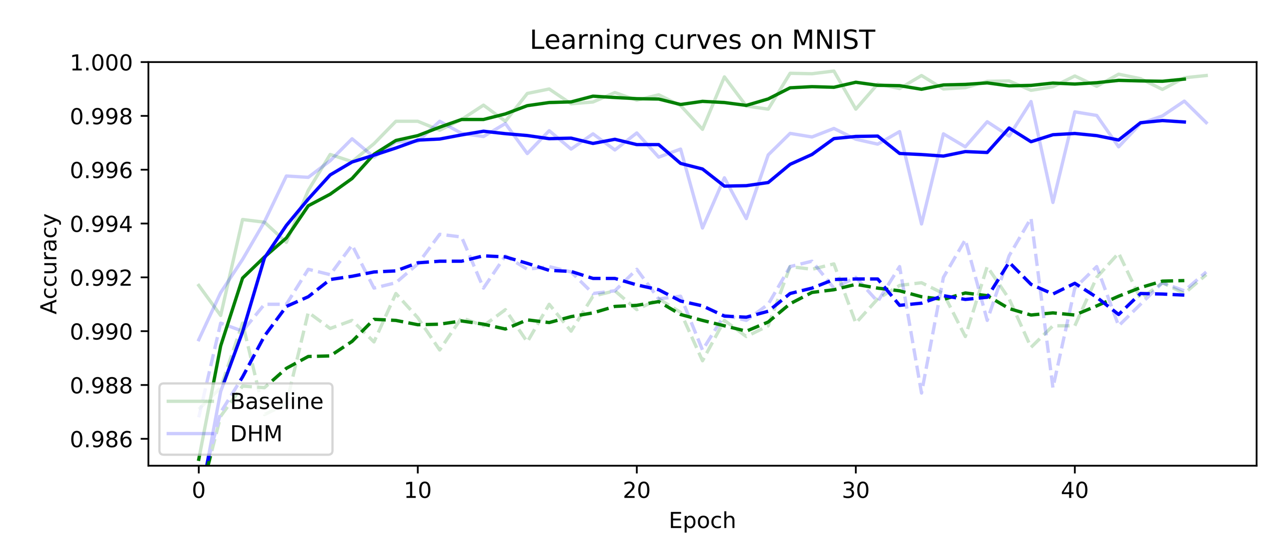
Interpolating Between Disc. and Gen.

- Hybrid models improve classification accuracy.





- Deep hybrid models overfit less.



Semi-Supervised Learning

There are two families of algorithms:

- Discriminative (transductive SVM, entropy reg.)
- Generative (VAEs, auxiliary variable DGMs)

Our framework allows applying both methods to the same model for ↑ performance!

Method	**SVHN** Accuracy
VAE (Kingma et al.)	$36.02 \pm 0.10\%$
SDGM (Maaloe et al.)	$16.61 \pm 0.24\%$
Improved GAN (Salimans et al.)	$8.11 \pm 1.3\%$
ALI (Dumoulin et al.)	$7.42 \pm 0.65\%$
Π-model (Aila et al.)	$5.45 \pm 0.25\%$
Implicit DHM (ours)	$4.45 \pm 0.35\%$