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.

Hybrids via Parameter Coupling
McCallum et al. (2006) propose new objective:
- User specifies a joint probability model p(x, y).
- We maximize the multi-conditional likelihood
  \[ \mathcal{L}(x, y) = \alpha \cdot \log p(x | y) + \beta \cdot \log p(x) \]
  where \( \alpha, \beta > 0 \) are hyper-parameters.

Bayesian Parameter Coupling
The coupling prior objective approach (Lasserre, Bishop, Minka, 2006) optimizes the model
\[ p(x, y, \theta_y, \theta_y) = \rho_y(y) \rho_x(x)p(y|x)\rho_y(y)\rho_x(x) \]
where the parameter coupling prior has the form
\[ \log p(\theta_y, \theta_y) = \lambda ||\theta_y - \theta_x|| \]
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!

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 \( x \).
Example: Logistic regression.
- Lower asymptotic error.
- Focus on prediction; fewer modeling assumptions.

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
   \[ \alpha \cdot \mathbb{E}_{q(x,y)} \mathbb{E}_{q(z|x)} \cdot \log p(y|x, z) + \beta \cdot D_{KL}(q(z|x) || p(z|x)) \]

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
  \[ \log \int_z p(y|x, z)^{\gamma} p(z) dz \geq \mathcal{L}. \]
- Applying the variational principle, we obtain:
  \[ \mathcal{L} = \mathbb{E}_{q(x,y)[} \left( \sum \log p(y|x, z) + \log p(x, z) - \log q(z|x) \right) \]
  - This is a special case of our framework with:
    \[ L_D = \text{expected log loss} \quad L_G = \text{KL}(q(z|x)||p(z|x)) \]

Implicit Density Models
- We may also choose \( p(x, z) \) to be a GAN. Then:
  \[ L_D = \text{expected log loss} \quad L_G = \text{IS}(q(z|x)||p(z|x)) \]
- We use a discriminator \( D \) to optimize \( L_G \)
  \[ L_G \approx \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{q(z|x)} \log D(x_i, z) + \mathbb{E}_{p(z|x)} \log(1-D(x_i, z)) \]

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

- 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.

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 \( \uparrow \) performance!

Method | SVHN Accuracy
--- | ---
VAE (Kingma et al.) | 36.02 ± 0.10%
SDGM (Maaloe et al.) | 16.61 ± 0.24%
Improved GAN (Salimans et al.) | 8.11 ± 1.33%
ALI (Dumoulin et al.) | 7.42 ± 0.65%
L-model (Aila et al.) | 5.45 ± 0.25%
Implicit DHM (ours) | 4.45 ± 0.35%