The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Multi-Entity Dependence Learning with Rich Context via Conditional Variational Auto-Encoder

Luming Tang*
Tsinghua University, China
tlm14@mails.tsinghua.edu.cn

Yexiang Xue*
Cornell University, USA
yexiang@cs.cornell.edu

Di Chen
Cornell University, USA
dc874@cornell.edu

Carla P. Gomes
Cornell University, USA
gomes@cs.cornell.edu

Abstract

Multi-Entity Dependence Learning (MEDL) explores conditional correlations among multiple entities. The availability of rich contextual information requires a nimble learning scheme that tightly integrates with deep neural networks and has the ability to capture correlation structures among exponentially many outcomes. We propose MEDL_CVAE, which encodes a conditional multivariate distribution as a generating process. As a result, the variational lower bound of the joint likelihood can be optimized via a conditional variational auto-encoder and trained end-to-end on GPUs. Our MEDL_CVAE was motivated by two real-world applications in computational sustainability: one studies the spatial correlation among multiple bird species using the *eBird* data and the other models multi-dimensional landscape composition and human footprint in the Amazon rainforest with satellite images. We show that MEDL_CVAE captures rich dependency structures, scales better than previous methods, and further improves on the joint likelihood taking advantage of very large datasets that are beyond the capacity of previous methods.

Introduction

Learning the dependencies among multiple entities is an important problem with many real-world applications. For example, in the sustainability domain, the spatial distribution of one species depends on other species due to their interactions in the form of mutualism, commensalism, competition and predation (MacKenzie, Bailey, and Nichols 2004). In natural language processing, the topics of an article are often correlated (Nam et al. 2014). In computer vision, an image may have multiple correlated tags (Wang et al. 2016).

The key challenge behind dependency learning is to capture correlation structures embedded among exponentially many outcomes. One classic approach is the Conditional Random Fields (CRF) (Lafferty, McCallum, and Pereira 2001). However, to handle the intractable partition function resulting from multi-entity interactions, CRFs have to incorporate approximate inference techniques such as contrastive divergence (Hinton 2002). In a related applicational domain called multi-label classification, the classifier chains (CC) approach (Read et al. 2009) decomposes the joint likelihood into a product of conditionals and reduces a multi-label classification problem to a series of binary prediction problems. However, as pointed out by (Dembczyński, Cheng, and Hüllermeier 2010), finding the joint mode of CC is also intractable, and to date only approximate search methods are available (Dembczyński et al. 2012).

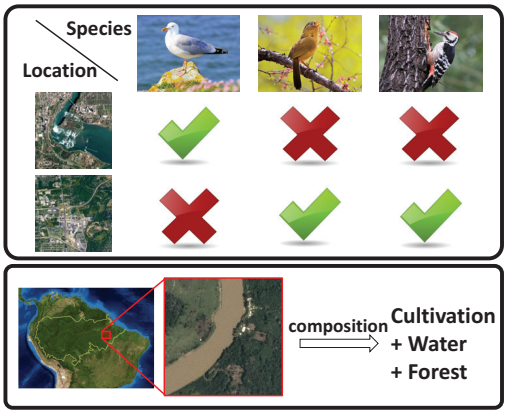

Figure 1: Two computational sustainability related applications for MEDL_CVAE. The first application is to study the interactions among bird species in the crowdsourced *eBird* dataset and environmental covariates including those from satellite images. The second application is to tag satellite images with a few potentially overlapping landscape categories and track human footprint in the Amazon rainforest.

The availability of rich contextual information such as millions of high-resolution satellite images as well as recent developments in deep learning create both opportunities and challenges for multi-entity dependence learning. In terms of opportunities, rich contextual information creates the possibility of improving predictive performance, especially when it is combined with highly flexible deep neural networks.

The challenge, however, is to design a nimble scheme that can both tightly integrate with deep neural networks and

*indicates equal contribution. The work was done when Tang, L. was a visiting student at Cornell University.

capture correlation structures among exponentially many outcomes. Deep neural nets are commonly used to extract features from contextual information sources, and can effectively use highly parallel infrastructure such as GPUs. However, classical approaches for structured output, such as sampling, approximate inference and search methods, typically cannot be easily parallelized.

Our contribution is **an end-to-end approach to multi-entity dependence learning based on a conditional variational auto-encoder, which handles high dimensional space effectively, and can be tightly integrated with deep neural nets to take advantages of rich contextual information**. Specifically, (i) we propose a novel *generating process* to encode the conditional multivariate distribution in multi-entity dependence learning, in which we bring in a set of hidden variables to capture the randomness in the joint distribution. (ii) The novel conditional generating process allows us to work with imperfect data, capturing noisy and potentially *multi-modal responses*. (iii) The generating process also allows us to encode the entire problem via *a conditional variational auto-encoder*, tightly integrated with deep neural nets and implemented end-to-end on GPUs. Using this approach, we are able to leverage rich contextual information to enhance the performance of MEDL that is beyond the capacity of previous methods.

We apply our Multi-Entity Dependence Learning via Conditional Variational Auto-encoder (MEDL_CVAE) approach to **two sustainability related real-world applications** (Gomes 2009). In the first application, we study the interaction among multiple bird species with crowd-sourced *eBird* data and environmental covariates including those from satellite images. As an important sustainable development indicator, studying how species distribution changes over time helps us understand the effects of climate change and conservation strategies. In our second application, we use high-resolution satellite imagery to study multi-dimensional landscape composition and track human footprint in the Amazon rainforest. See Figure 1 for an overview of the two problems. Both applications study the correlations of multiple entities and use satellite images as rich context information. We are able to show that our MEDL_CVAE (i) **captures rich correlation structures among entities**, therefore **outperforming approaches that assume independence among entities** given contextual information; (ii) **trains in an order of magnitude less time than previous methods** because of a full implementation on GPUs; (iii) achieves a better joint likelihood by incorporating deep neural nets to **take advantage of rich context information, namely satellite images**, which are beyond the capacity of previous methods.

Preliminaries

We consider modeling the dependencies among multiple entities on problems with rich contextual information. Our dataset consists of tuples $D = \{(x_i, y_i)|i = 1, \ldots, N\}$, in which $x_i = (x_{i,1}, \ldots, x_{i,k}) \in \mathcal{R}^k$ is a high-dimensional contextual feature vector, and $y_i = (y_{i,1}, \ldots, y_{i,l}) \in \{0, 1\}^l$ is a sequence of l indicator variables, in which $y_{i,j}$ represents whether the j-th entity is observed in an environment characterized by covariates x_i. The problem is to learn a conditional joint distribution $Pr(y|x)$ which maximizes the conditional joint log likelihood over N data points:

$$\sum_{i=1}^{N} \log Pr(y_i|x_i).$$

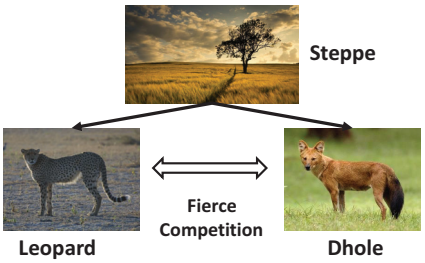

Figure 2: Leopards and dholes both live on steppes. Therefore, the probability that each animal occupies a steppe is high. However, due to the competition between the two species, the probability of their co-existence is low.

Multi-entity dependence learning is a general problem with many applications. For example, in our species distribution application where we would like to model the relationships of multiple bird species, x_i is the vector of environmental covariates of the observational site, which includes a remote sensing picture, the national landscape classification dataset (NLCD) values (Homer et al. 2015) etc. $y_i = (y_{i,1}, \ldots, y_{i,l})$ is a sequence of binary indicator variables, where $y_{i,j}$ indicates whether species j is detected in the observational session of the i-th data point. In our application to analyze landscape composition, x_i is the feature vector made up with the satellite image of the given site, and y_i includes multiple indicator variables, such as atmospheric conditions (clear or cloudy) and land cover phenomena (agriculture or forest) of the site.

Our problem is to capture rich correlations between entities. For example, in our species distribution modeling application, the distribution of multiple species are often correlated, due to their interactions such as cooperation and competition for shared resources. As a result, we often cannot assume that the probability of each entity's existence are mutually independent given the feature vector, i.e.,

$$Pr(y_i|x_i) \neq \prod_{j=1}^{l} Pr(y_{i,j}|x_i). \quad (1)$$

See Figure 2 for a specific instance. As a baseline, we call the model which takes the assumption in the righthand side of Equation (1) an independent probabilistic model.

Our Approach

We propose MEDL_CVAE to address two challenges in multi-entity dependence learning.

The first challenge is the noisy and potentially multi-modal responses. For example, consider our species distribution modeling application. One bird watcher can make slightly different observations during two consecutive visits to the same forest location. He may be able to detect a song bird during one visit but not the other if, for example, the bird does not sing both times. This suggests that, even under the very best effort of bird watchers, there is still noise inherently associated with the observations. Another, perhaps more complicated phenomenon is the multi-modal response, which results from intertwined ecological processes such as mutualism and commensalism. Consider, territorial species such as the Red-winged and Yellow-headed Blackbirds, both of which live in open marshes in Northwestern United States. However, the Yellowheads tend to chase the Redwings out of their territories. As a result, a bird watcher would see either Red-winged or Yellow-headed Blackbirds at an open marsh, but seldom both of them. This suggests that, conditioned on an open marsh environment, there are two possible modes in the response, seeing the Red-winged but not the Yellow-headed, or seeing the Yellow-headed but not the Red-winged.

The second challenge comes from the incorporation of rich contextual information such as remote sensing imagery, which provides detailed feature description of the underlying environment, especially in conjunction with the flexibility of deep neural networks. Nevertheless, previous multi-entity models, such as in (Chen et al. 2017; Guo and Gu 2011; Sutton and McCallum 2012), rely on sampling approaches to estimate the partition function during training. It is difficult to incorporate such sampling process into the back-propagation of the deep neural networks. This limitation poses serious challenges to taking advantage of the rich contextual information.

To address the aforementioned two challenges, we propose a conditional generating model, which makes use of hidden variables to represent the noisy and multi-modal responses. `MEDL_CVAE` incorporates this model into an end-to-end training pipeline using a conditional variational autoencoder, which optimizes for a variational lower bound of the conditional likelihood function.

Conditional Generating Process Unlike classic approaches such as probit models (Chib and Greenberg 1998), which have a single mode, we use a conditional generating process, which models noisy and multi-modal responses using additional hidden variables. The generating process is depicted in Figure. 3.

In the generating process, we are given contextual features x_i, which for example, contain a satellite image. Then we assume a set of hidden variables z_i, which are generated based on a normal distribution conditioned on the values of x_i. The binary response variables $y_{i,j}$ are drawn from a Bernoulli distribution, whose parameters depend on both the contextual features x_i and hidden variables z_i. The complete generating process becomes,

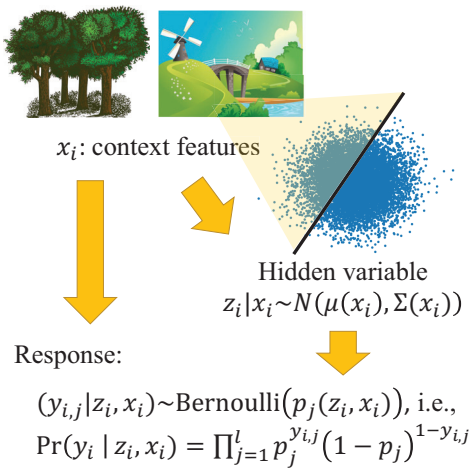

Figure 3: Our proposed conditional generating process. Given contextual features x_i such as satellite images, we use hidden variables z_i conditionally generated based on x_i to capture noisy and multi-modal response. The response y_i depends on both contextual information x_i and hidden variables z_i. See the main text for details.

$$x_i : \text{contextual features}, \tag{2}$$

$$z_i|x_i \sim N\left(\mu_d(x_i), \Sigma_d(x_i)\right), \tag{3}$$

$$y_{i,j}|z_i, x_i \sim \text{Bernoulli}\left(p_j(z_i, x_i)\right). \tag{4}$$

Here, $\mu_d(x_i)$, $\Sigma_d(x_i)$ and $p_j(z_i, x_i)$ are general functions depending on x_i and z_i, which are modeled as deep neural networks in our application and learned from data. We denote the parameters in these neural networks as θ_d. The machine learning problem is to find the best parameters that maximize the conditional likelihood $\prod_i Pr(y_i|x_i)$.

This generating process is able to capture noisy and potentially multi-modal distributions. Consider the Red-winged and the Yellow-headed Blackbird example. We use $y_{i,1}$ to denote the occurrence of Red-winged Blackbird and $y_{i,2}$ to denote the occurrence of Yellow-headed Blackbird. Conditioned on the same environmental context x_i of an open marsh, the output $(y_{i,1} = 0, y_{i,2} = 1)$ and $(y_{i,1} = 1, y_{i,2} = 0)$ should both have high probabilities. Therefore, there are two modes in the probability distribution. Notice that it is very difficult to describe this case in any probabilistic model that assumes a single mode.

Our generating process provides the flexibility to capture multi-modal distributions of this type. The high-level idea is similar to mixture models, where we use hidden variables z_i to denote which mode the actual probabilistic outcome is in. For example, we can have $z_i|x_i \sim N(0, I)$ and two functions $p_1(z)\, p_2(z)$, where half of the z_i values are mapped to $(p_1 = 0, p_2 = 1)$ and the other half to $(p_1 = 1, p_2 = 0)$. Figure 3 provides an example, where the z_i values in the region with a yellow background are mapped to one value, and the remaining values are mapped to the other value. In this way, the model will have high probabilities to produce both outcomes $(y_{i,1} = 0, y_{i,2} = 1)$ and $(y_{i,1} = 1, y_{i,2} = 0)$.

Conditional Variational Autoencoder Our training algorithm is to maximize the conditional likelihood $Pr(y_i|x_i)$. Nevertheless, a direct method would result in the following optimization problem:

$$\max_{\theta_d} \sum_i \log Pr(y_i|x_i) = \sum_i \log \int Pr(y_i|x_i, z_i) Pr(z_i|x_i) \mathrm{d}z_i$$

which is intractable because of a hard integral inside the logarithmic function. Instead, we turn to maximizing variational lower bound of the conditional log likelihood. To do this, we use a variational function family $Q(z_i|x_i, y_i)$ to approximate the posterior: $Pr(z_i|x_i, y_i)$. In practice, $Q(z_i|x_i, y_i)$ is modeled using a conditional normal distribution:

$$Q(z_i|x_i, y_i) = N(\mu_e(x_i, y_i), \Sigma_e(x_i, y_i)). \tag{5}$$

Here, $\mu_e(x_i, y_i)$ and $\Sigma_e(x_i, y_i)$ are general functions, and are modeled using deep neural networks whose parameters are denoted as θ_e. We assume Σ_e is a diagonal matrix in our formulation. Following similar ideas in (Kingma and Welling 2013; Kingma et al. 2014), we can prove the following variational equality:

$$\begin{aligned} &\log Pr(y_i|x_i) - D\left[Q(z_i|x_i, y_i)||Pr(z_i|x_i, y_i)\right] \\ =&\mathrm{E}_{z_i \sim Q(z_i|x_i, y_i)}\left[\log Pr(y_i|z_i, x_i)\right] - D\left[Q(z_i|x_i, y_i)||Pr(z_i|x_i)\right] \end{aligned} \tag{6}$$

On the left-hand side, the first term is the conditional likelihood, which is our objective function. The second term is the Kullback-Leibler (KL) divergence, which measures how close the variational approximation $Q(z_i|x_i, y_i)$ is to the true posterior likelihood $Pr(z_i|x_i, y_i)$. Because Q is modeled using a neural network, which captures a rich family of functions, we assume that $Q(z_i|x_i, y_i)$ approximates $Pr(z_i|x_i, y_i)$ well, and therefore the second KL term is almost zero. And because the KL divergence is always non-negative, the right-hand side of Equation 6 is a tight lower bound of the conditional likelihood, which is known as the variational lower bound. We therefore directly maximize this value and the training problem becomes:

$$\begin{aligned} \max_{\theta_d, \theta_e} \quad & \sum_i \mathrm{E}_{z_i \sim Q(z_i|x_i, y_i)}\left[\log Pr(y_i|z_i, x_i)\right] - \\ & D\left[Q(z_i|x_i, y_i)||Pr(z_i|x_i)\right]. \end{aligned} \tag{7}$$

The first term of the objective function in Equation 7 can be directly formalized as two neural networks concatenated together – one encoder network and the other decoder network, following the reparameterization trick, which is used to backpropagate the gradient inside neural nets. At a high level, suppose $r \sim N(0, I)$ are samples from the standard Gaussian distribution, then $z_i \sim Q(z_i|x_i, y_i)$ can be generated from a "recognition network", which is part of the "encoder network": $z_i \leftarrow \mu_e(x_i, y_i) + \Sigma_e^{1/2}(x_i, y_i)r$. Notice that z_i is a hidden variable. Its value depends on the parameters in μ_e and Σ_e, which are learned from data, without any human intervention. The "decoder network" takes the input of z_i from the encoder network and feeds it to the neural network representing the function $Pr(y_i|z_i, x_i) = \prod_{j=1}^{l} \left(p_j(z_i, x_i)\right)^{y_{i,j}} \left(1 - p_j(z_i, x_i)\right)^{1-y_{i,j}}$ together with x_i. The second KL divergence term can be calculated in a close form. The entire neural network structure is shown as Figure 4. We refer to $Pr(z|x)$ as the *prior network*, $Q(z|x, y)$ as the *recognition network* and $Pr(y|x, z)$ as the *decoder network*. These three networks are all multi-layer fully connected neural networks. The fourth *feature network*, composed of multi-layer convolutional or fully connected network, extracts high-level features from the contextual source. All four neural networks are trained simultaneously using stochastic gradient descent. In our experiment, the convolutional neural networks used in the feature network are composed of multiple "Convolution→Batch Normalization→Relu→Pooling" layers, which is a classical setting in many applications. The feature network for each context input has its own hyperparameter settings, such as the number of layers and hidden layer size. According to our experiment, the performance of our algorithm is not sensitive to the network structure and hyperparameter settings.

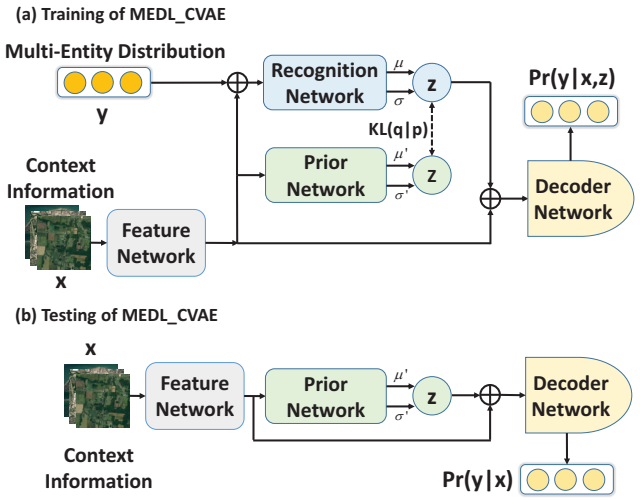

Figure 4: Overview of the neural network architecture of MEDL_CVAE for both training and testing stages. $\oplus$ denotes a concatenation operator.

Related Work

Multi-entity dependence learning was studied extensively for prediction problems under the names of multi-label classification and structured prediction. Our applications, on the other hand, focus more on probabilistic modeling rather than classification. Along this line of research, early methods include k-nearest neighbors (Zhang and Zhou 2005) and dimension reduction (Zhang and Zhou 2010; Li et al. 2016).

Classifier Chains (CC) First proposed by (Read et al. 2009), the CC approach decomposes the joint distribution into the product of a series of conditional probabilities. Therefore the multi-label classification problem is reduced to l binary classification problems. As noted by (Dembczyński, Cheng, and Hüllermeier 2010), CC takes a greedy approach to find the joint mode and the result can be arbitrarily far from the true mode. Hence, Probabilistic Classifier Chains (PCC) were proposed which replaced the greedy

strategy with exhaustive search (Dembczyński, Cheng, and Hüllermeier 2010), ϵ-approximate search (Dembczyński et al. 2012), beam search (Kumar et al. 2013) or A* search (Mena Waldo et al. 2015). To address the issue of error propagating in CC, Ensemble of Classifier Chains (ECC) (Liu and Tsang 2015) averages several predictions by different chains to improve the prediction.

Conditional Random Field (CRF) (Lafferty, McCallum, and Pereira 2001) offers a general framework for structured prediction based on undirected graphical models. When used in multi-label classification, CRF suffers from the problem of computational intractability. To remedy this issue, (Xu et al. 2011) applied ensemble methods and (Deng et al. 2014) proposed a special CRF for problems involving specific hierarchical relations. In addition, (Guo and Gu 2011) proposed using Conditional Dependency Networks, although their method also depended on the Gibbs sampling for approximate inference.

Ecological Models Species distribution modeling has been studied extensively in ecology and (Elith and Leathwick 2009) presented a nice survey. For single-species models, (Phillips, Dudík, and Schapire 2004) proposed maxentropy methods to deal with presence-only data. By taking imperfect detection into account, (MacKenzie, Bailey, and Nichols 2004) proposed occupancy models, which were further improved with a stronger version of statistical inference (Hutchinson, Liu, and Dietterich 2011). Species distribution models have been extended to capture population dynamics using cascade models (Sheldon and Dietterich 2011) and non-stationary predictor response models (Fink, Damoulas, and Dave 2013).

Multi-species interaction models were also proposed (Jun et al. 2011; Harris 2015). Recently, Deep Multi-Species Embedding (DMSE) (Chen et al. 2017) uses a probit model coupled with a deep neural net to capture inter-species correlations. This approach is closely related to CRF and also requires MCMC sampling during training.

Experiments

Datasets and Implementation Details

We evaluate our method on two computational sustainability related datasets. The first one is a crowd-sourced bird observation dataset collected from the *eBird* citizen science project (Munson et al. 2012). Each record in this dataset is referred to as a checklist in which the bird observer reports all the species he detects together with the time and the geographical location of an observational session. Crossed with the National Land Cover Dataset (NLCD) (Homer et al. 2015), we get a 15-dimension feature vector for each location which describes the nearby landscape composition with 15 different land types such as water, forest, etc. In addition, to take advantages of rich external context information, we also collect satellite images for each observation site by matching the longitude and latitude of the observational site to Google Earth[1] . From the upper part of Figure 5, the satellite images of different geographical locations are quite different. Therefore these images contain rich geological information. Each image covers an area of 12.3km^2 near the observation site. For the use of training and testing, we transform all this data into the form (x_i, y_i), where x_i denotes the contextual features including NLCD and satellite images and y_i denotes the multi-species distribution. The whole dataset contains all the checklists from the Bird Conservation Region (BCR) 13 (Committee and others 2000) in the last two weeks of May from 2004 to 2014, which has 50,949 observations in total. Since May is a migration season and lots of non-native birds fly over BCR 13, this dataset provides a good opportunity to study these migratory birds using this dataset. We choose the top 100 most frequently observed birds as the target species which cover over 95% of the records in our dataset. A simple mutual information analysis reveals rich correlation structure among these species.

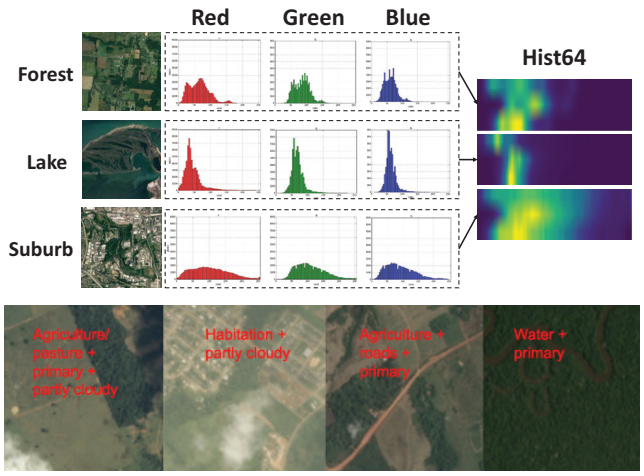

Figure 5: (Top) Three satellite images (left) of different landscapes contain rich geographical information. We can also see that the histograms of RGB channels for each image (middle) contain useful information and are good summary statistics. (Bottom) Examples of sample satellite image chips and their corresponding landscape composition.

Our second application is the Amazon rainforest landscape analysis[2], in which we tag satellite images with a few landscape categories. Some categories are about atmospheric conditions, such as clear or cloudy and others are about landcover, such as agricultural or forest land. Raw satellite images were derived from Planet's full-frame analytic scene products using 4-band satellites in the sun-synchronous orbit and the International Space Station orbit. The organizers at Kaggle used Planet's visual product processor to transform raw images to 3-band jpg format. Each satellite image sample in this dataset contains an image chip covering a ground area of 0.9 km^2. The chips were analyzed using the Crowd Flower[3] platform to obtain ground-truth

[1]https://www.google.com/earth/. Google Earth has already conducted preprocessing including cloud removing on the satellite images.

[2]https://www.kaggle.com/c/planet-understanding-the-amazon-from-space.

[3]https://www.crowdflower.com/

Dataset	Training Set Size	Test Set Size	# Entities
eBird	45855	5094	100
Amazon	30383	4048	17

Table 1: The statics of the *eBird* and the *Amazon* dataset

landscape composition. There are 17 composition label entities and they represent a reasonable subset of phenomena of interest in the Amazon basin and can broadly be broken into three groups: atmospheric conditions, common land cover phenomena and rare land use phenomena. Each chip has one or more atmospheric label entities and zero or more common and rare label entities. Sample chips and their composition are demonstrated in the lower part of Figure 5. There exists rich correlations between label entities, for instance, agriculture has a high probability to co-occur with water and cultivation. We randomly choose 34,431 samples for training, validation and testing. The details of the two datasets are listed in table 1.

We propose two different neural network architectures for the *feature network* to extract useful features from satellite images: multi-layer fully connected neural network (MLP) and convolutional neural network (CNN). We also rescale images into different resolutions: Image64 for 64*64 pixels and Image256 for 256*256 pixels. In addition, we experiment using summary statistics such as the histograms of image's RGB channels (upper part of Figure 5) to describe an image (denoted as Hist). Inspired by (You et al. 2017) that assumes permutation invariance holds and only the number of different pixel type in an image (pixel counts) are informative, we transfer each image into a matrix $\mathbf{H} \in \mathbb{R}^{d\times b}$, where d is the number of bands and b is the number of discrete range section, thus $H_{i,j}$ indicates the percentage of pixels in the range of section j for band i. We use RGB so $d = 3$. We utilize histogram models with two different b settings, Hist64 for $b = 64$ and Hist128 for $b = 128$.

All the training and testing process of our proposed MEDL_CVAE are performed on one NVIDIA Quadro 4000 GPU with 8GB memory. The whole training process lasts 300 epochs, using batch size of 512, Adam optimizer (Kingma and Ba 2014) with learning rate of 10^{-4} and utilizing batch normalization (Ioffe and Szegedy 2015), 0.8 dropout rate (Srivastava et al. 2014) and early stopping to accelerate the training process and to prevent overfitting.

Experimental Results

We compare the proposed MEDL_CVAE with two different groups of baseline models. The first group is *models assuming independence structures among entities*; i.e., the distribution of all entities are independent of each other conditioned on the feature vector. Within this group, we have tried different models with different feature inputs, including models that simply combine independent binary classifiers on each indicator $y_{i,j}$ as well as models with highly advanced deep neural net structure, ResNet (He et al. 2016). The second group is *previously proposed multi-entity dependence models* which have the ability to capture correlations among entities. Within this group, we compare with the recent proposed Deep Multi-Species Embedding (DMSE) (Chen et al. 2017). This model is closely related to CRF, representing a wide class of energy based approach. Moreover, it further improves classic energy models, taking advantages of the flexibility of deep neural nets to obtain useful feature description. Nevertheless, its training process uses classic MCMC sampling approaches, therefore cannot be fully integrated on GPUs. We also compare with Probabilistic Classifier Chains (PCC) (Dembczyński, Cheng, and Hüllermeier 2010), which is a representative approach among a series of models proposed in multi-label classification.

Our baselines and MEDL_CVAE are all trained using different feature network architecture as well as satellite im-

Method	Neg. JLL	Time (min)
NLCD+MLP	36.32	2
Image256+ResNet50	34.16	5.3 hrs
NLCD+Image256+ResNet50	34.48	5.7 hrs
NLCD+Hist64+MLP	34.97	3
NLCD+Hist128+MLP	34.62	4
NLCD+Image64+MLP	33.73	9
NLCD+MLP+PCC	35.99	21
NLCD+Hist128+MLP+PCC	35.07	33
NLCD+Image64+MLP+PCC	34.48	53
NLCD+DMSE	30.53	20 hrs
NLCD+MLP+MEDL_CVAE	30.86	9
NLCD+Hist64+MLP+MEDL_CVAE	28.86	20
NLCD+Hist128+MLP+MEDL_CVAE	28.71	22
NLCD+Image64+MLP+MEDL_CVAE	**28.24**	48

Table 2: Negative joint log-likelihood and training time of models assuming independence (first section), previous multi-entity dependence models (second section) and our MEDL_CVAE on the *eBird* test set. MEDL_CVAE achieves lower negative log-likelihood compared to other methods with the same feature network structure and context input while taking much less training time. Our model is also the only one among joint models (second and third section) which achieves the best log-likelihood taking images as inputs, while other models must rely on summary statistics to get good but suboptimal results within a reasonable time limit.

Method	Neg. JLL
Image64+MLP	2.83
Hist128+MLP	2.44
Image64+CNN	2.16
Image64+MLP+PCC	2.95
Hist128+MLP+PCC	2.60
Image64+CNN+PCC	2.45
Image64+MLP+MEDL_CVAE	2.37
Hist128+MLP+MEDL_CVAE	2.09
Image64+CNN+MEDL_CVAE	**2.03**

Table 3: Performance of baseline models and MEDL_CVAE on the *Amazon* dataset. Our method clearly outperforms models assuming independence (first section) and previous multi-entity dependence models (second section) with various types of context input and feature network structures.

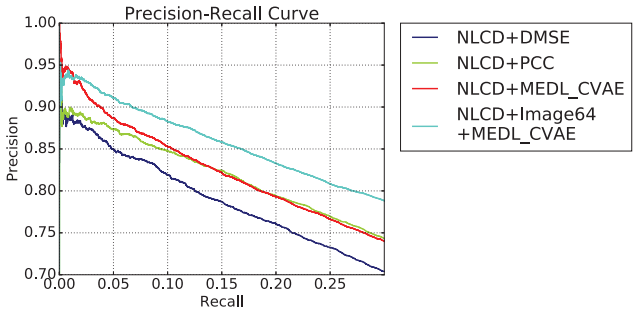

Figure 6: Precison-Recall Curves for a few multi-entity dependence models on the *ebird* dataset (better view in color). MEDL_CVAE utilizing images as input outperforms other joint methods without imagery.

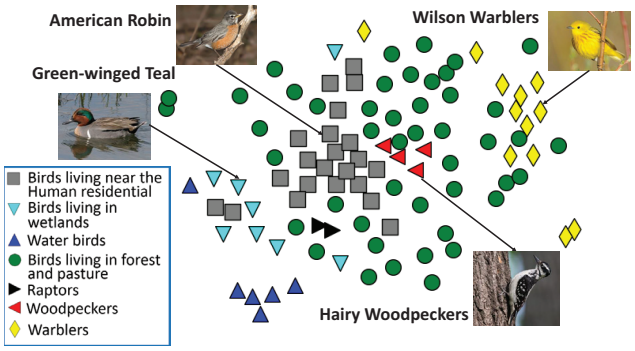

Figure 7: Visualization of the vectors inside decoder network's last fully connected layer gives a reasonable embedding of multiple bird species when our model is trained on the *eBird* dataset. Birds living in the same habitats are clustered together.

agery with different resolution and encoding. We use Negative Joint Distribution Loglikelihood (Neg. JLL) as the main indicator of a model's performance: $-\frac{1}{N}\sum_{i=1}^{N}\log Pr(y_i|x_i)$, where N is the number of samples in the test set. This indicator measures how likely the model produces the observations in the dataset. For MEDL_CVAE models, $Pr(y_i|x_i) = \mathrm{E}_{z_i\sim Pr(z_i|x_i)}\left[Pr(y_i|z_i,x_i)\right]$. We obtain 10,000 samples from the posterior $Pr(z_i|x_i)$ to estimate the expectation. We also double checked that the estimation is close and within the bound of the variational lower bound in Equation 6. The sampling process can be performed on GPUs within a couple of minutes. The experiment results on *eBird* and *Amazon* dataset are shown in Table 2 and 3, respectively.

We can observe that: (1)**MEDL_CVAE significantly outperforms all independent models** given the same feature network (CNN or MLP) and context information (Image or Hist), even if we use highly advanced deep neural net structures such as ResNet in independent models. It proves that our method is able to capture rich dependency structures among entities, therefore outperforming approaches that assume independence among entities. (2) Compared with previous multi-entity dependence models, **MEDL_CVAE trains in an order of magnitude less time**. Using low-dimensional context information NLCD, which is a 15-dimensional vector, PCC's training needs nearly twice the time of MEDL_CVAE and DMSE needs over 130 times (20 hours). (3) In each model group in Table 2, it is clear that adding satellite images improves the performance, which proves that **rich context is informative.** (4) Only our model MEDL_CVAE is able to take full advantage of the rich context in satellite images. Other models, such as DMSE, already suffer from a long training time with low-dimensional feature inputs such as NLCD, and cannot scale to using satellite images. It should be noted that NLCD+Image64+MLP+MEDL_CVAE can achieve much better performance with only 1/25 time of DMSE. PCC needs less training time than DMSE but doesn't perform well on joint likelihood. It is clear that **due to the end-to-end training process on GPUs, our method is able to take advantage of rich context input to further improve multi-entity dependence modeling**, which is beyond the capacity of previous models.

To further prove MEDL_CVAE's modeling power, we plot the precision-recall curve shown in Figure 6 for all dependency models on the *ebird* dataset. The precision and recall is defined on the marginals to predict the occurrence of individual species and averaged among all 100 species in the dataset. As we can see, our method outperforms other models after taking rich context information into account.

Latent Space and Hidden Variables Analysis In order to qualitatively confirm that our MEDL_CVAE learns useful dependence structure between entities, we analyze the latent space formed by the hidden variables in the neural network. Inspired by (Chen et al. 2017), each vector of decoder network's last fully connected layer can be treated as an embedding showing the relationship among species. Figure 7 visualizes the embedding using t-SNE (Maaten and Hinton 2008). We can observe that birds having similar environmental preferences tend to cluster together. In addition, previous work (Kingma and Welling 2013) has shown that the *recognition network* in Variational Auto-encoder is able to cluster high-dimensional data. Therefore we conjecture that the posterior of z from the *recognition network* should carry meaningful information on the cluster groups. Figure 8 visualizes the posterior of $z \sim Q(z|x,y)$ in 2D space using t-SNE on the *Amazon* dataset. We can see that satellite images of similar landscape composition also cluster together.

Conclusion

In this paper, we propose MEDL_CVAE for multi-entity dependence learning, which encodes a conditional multivariate distribution as a generating process. As a result, the variational lower bound of the joint likelihood can be optimized via a conditional variational auto-encoder and trained end-to-end on GPUs. Tested on two real-world applications in computational sustainability, we show that MEDL_CVAE captures rich dependency structures, scales better than previous methods, and further improves the joint likelihood taking ad-

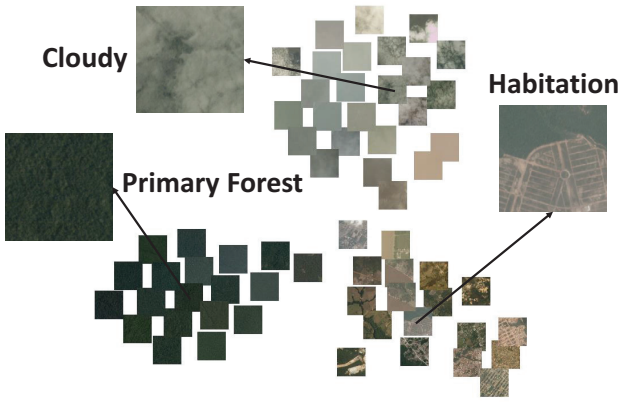

Figure 8: Visualization of the posterior $z \sim Q(z|x, y)$ gives a good embedding on landscape composition when our model is trained on the *Amazon* dataset. Pictures with similar landscapes are clustered together.

vantage of very rich context information that is beyond the capacity of previous methods. Future directions include exploring the connection between the current formulation of `MEDL_CVAE` based on deep neural nets and the classic multivariate response models in statistics.

Acknowledgements

Tang, L. would like to thank the Department of Physics, Tsinghua University for providing financial support for his stay at Cornell. The authors thank Bin Dai for discussions and all the reviewers for helpful suggestions. We are thankful to Kaggle competition organizers and the Cornell Lab of Ornithology for providing data for this research. This research was supported by National Science Foundation (Grant Number 0832782, 1522054, 1059284, 1356308), ARO grant W911-NF-14-1- 0498, the Leon Levy Foundation and the Wolf Creek Foundation.

References

Chen, D.; Xue, Y.; Fink, D.; Chen, S.; and Gomes, C. P. 2017. Deep multi-species embedding. In *IJCAI*.

Chib, S., and Greenberg, E. 1998. Analysis of multivariate probit models. *Biometrika* 85(2):347–361.

Committee, U. N., et al. 2000. North american bird conservation initiative: Bird conservation region descriptions, a supplement to the north american bird conservation initiative bird conservation regions map.

Dembczyński, K.; Waegeman, W.; Cheng, W.; and Hüllermeier, E. 2012. On label dependence and loss minimization in multi-label classification. *Machine Learning* 88(1-2):5–45.

Dembczyński, K.; Cheng, W.; and Hüllermeier, E. 2010. Bayes optimal multilabel classification via probabilistic classifier chains. In *ICML*.

Deng, J.; Ding, N.; Jia, Y.; Frome, A.; Murphy, K.; Bengio, S.; Li, Y.; Neven, H.; and Adam, H. 2014. Large-scale object classification using label relation graphs. In *ECCV*.

Elith, J., and Leathwick, J. R. 2009. Species distribution models: ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution, and Systematics* 40(1):677.

Fink, D.; Damoulas, T.; and Dave, J. 2013. Adaptive spatio-temporal exploratory models: Hemisphere-wide species distributions from massively crowdsourced ebird data. In *AAAI*.

Gomes, C. P. 2009. Computational sustainability: Computational methods for a sustainable environment, economy, and society. *The Bridge* 39(4):5–13.

Guo, Y., and Gu, S. 2011. Multi-label classification using conditional dependency networks. In *IJCAI*.

Harris, D. J. 2015. Generating realistic assemblages with a joint species distribution model. *Methods in Ecology and Evolution*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Identity mappings in deep residual networks. In *ECCV*.

Hinton, G. E. 2002. Training products of experts by minimizing contrastive divergence. *Neural computation* 14(8):1771–1800.

Homer, C.; Dewitz, J.; Yang, L.; Jin, S.; Danielson, P.; Xian, G.; Coulston, J.; Herold, N.; Wickham, J.; and Megown, K. 2015. Completion of the 2011 national land cover database for the conterminous united states–representing a decade of land cover change information. *Photogrammetric Engineering & Remote Sensing*.

Hutchinson, R. A.; Liu, L.-P.; and Dietterich, T. G. 2011. Incorporating boosted regression trees into ecological latent variable models. In *AAAI*.

Ioffe, S., and Szegedy, C. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, 448–456.

Jun, Y.; Wong, W.-K.; Dietterich, T.; Jones, J.; Betts, M.; Frey, S.; Shirley, S.; Miller, J.; and White, M. 2011. Multi-label classification for species distribution. In *Proceedings of the ICML 2011 Workshop on Machine Learning for Global Challenges*.

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

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

Kingma, D. P.; Mohamed, S.; Rezende, D. J.; and Welling, M. 2014. Semi-supervised learning with deep generative models. In *Advances in Neural Information Processing Systems*, 3581–3589.

Kumar, A.; Vembu, S.; Menon, A. K.; and Elkan, C. 2013. Beam search algorithms for multilabel learning. *Machine learning* 92(1):65–89.

Lafferty, J. D.; McCallum, A.; and Pereira, F. C. N. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01.

Li, C.; Wang, B.; Pavlu, V.; and Aslam, J. 2016. Conditional bernoulli mixtures for multi-label classification. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16.

Liu, W., and Tsang, I. 2015. On the optimality of classifier chain for multi-label classification. In *Advances in Neural Information Processing Systems*, 712–720.

Maaten, L. v. d., and Hinton, G. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research* 9(Nov):2579–2605.

MacKenzie, D. I.; Bailey, L. L.; and Nichols, J. D. 2004. Investigating species co-occurrence patterns when species are detected imperfectly. *Journal of Animal Ecology* 73.

Mena Waldo, D.; Montañés Roces, E.; Quevedo Pérez, J. R.; and Coz Velasco, J. J. d. 2015. Using a* for inference in probabilistic classifier chains. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015)*. Association for the Advancement of Artificial Intelligence.

Munson, M. A.; Webb, K.; Sheldon, D.; Fink, D.; Hochachka, W. M.; Iliff, M.; Riedewald, M.; Sorokina, D.; Sullivan, B.; Wood, C.; et al. 2012. The ebird reference dataset, version 4.0.

Nam, J.; Kim, J.; Mencía, E. L.; Gurevych, I.; and Fürnkranz, J. 2014. Large-scale multi-label text classificationrevisiting neural networks. In *Joint european conference on machine learning and knowledge discovery in databases*, 437–452. Springer.

Phillips, S. J.; Dudík, M.; and Schapire, R. E. 2004. A maximum entropy approach to species distribution modeling. In *Proceedings of the Twenty-first International Conference on Machine Learning*, ICML '04, 83–. New York, NY, USA: ACM.

Read, J.; Pfahringer, B.; Holmes, G.; and Frank, E. 2009. Classifier chains for multi-label classification. In *Machine Learning and Knowledge Discovery in Databases, European Conference, ECML PKDD 2009*.

Sheldon, D. R., and Dietterich, T. G. 2011. Collective graphical models. In *NIPS*.

Srivastava, N.; Hinton, G. E.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of machine learning research* 15(1):1929–1958.

Sutton, C., and McCallum, A. 2012. An introduction to conditional random fields. *Found. Trends Mach. Learn.* 4.

Wang, J.; Yang, Y.; Mao, J.; Huang, Z.; Huang, C.; and Xu, W. 2016. Cnn-rnn: A unified framework for multi-label image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2285–2294.

Xu, X.-S.; Jiang, Y.; Peng, L.; Xue, X.; and Zhou, Z.-H. 2011. Ensemble approach based on conditional random field for multi-label image and video annotation. In *Proceedings of the 19th ACM International Conference on Multimedia*.

You, J.; Li, X.; Low, M.; Lobell, D.; and Ermon, S. 2017. Deep gaussian process for crop yield prediction based on remote sensing data. In *AAAI*, 4559–4566.

Zhang, M.-L., and Zhou, Z.-H. 2005. A k-nearest neighbor based algorithm for multi-label classification. In *2005 IEEE International Conference on Granular Computing*, volume 2.

Zhang, Y., and Zhou, Z.-H. 2010. Multilabel dimensionality reduction via dependence maximization. *ACM Trans. Knowl. Discov. Data* 4(3).