Discovering Order in Unordered Datasets: Generative Markov Networks

Yao-Hung Hubert Tsai1 Han Zhao1 Nebojsa Jojic2 Ruslan Salakhutdinov1

Abstract

The assumption that data samples are independently identically distributed is the backbone of many learning algorithms. Nevertheless, datasets often exhibit rich structure in practice, and we argue that there exist some unknown order within the data instances. In this paper, we introduce a Generative Markov Network (GMN) that can be used to extract the order of data instances in an unsupervised fashion. Specifically, we assume that the instances are sampled from a Markov chain. Our goal is to learn the transitional operator of the Markov chain as well as the order by maximizing the generation probability under all possible data permutations. We use neural networks as a compact and soft lookup table for approximating the possibly huge, but discrete transition matrix. This strategy allows us to amortize the space complexity with a single model. Furthermore, this simple and compact representation also provides a short description to the dataset and generalizes to unseen instances as well. To ensure that the learned Markov chain is ergodic, we propose a greedy batch-wise permutation scheme that allows fast training. Empirically, we show that GMNs are able to discover an order among data instances. We also extend the proposed GMN to one-shot recognition task and demonstrate comparable results to the state-of-the-art methods.

1. Introduction

Recent advances in deep neural networks offer a great potential for machines to learn automatically without human intervention. For instance, Convolutional Neural Networks (CNNs) (Krizhevsky et al., 2012) provided an automated way for learning image feature representations. Compared to hand-crafted features, such as SIFT and SURF, these hierarchical deep features demonstrate superior performance in recognition (Xu et al., 2015; Finn & Levine, 2017) and transfer learning (Glorot et al., 2011) problems. Another example is learning to learn for automatic parameter estimation. Andrychowicz et al. (2016) proposed to update model parameters without any pre-defined update rule, such as stochastic gradient descent (SGD) or ADAM (Kingma & Ba, 2014). Surprisingly, this update-rule-free framework showed better performance and faster convergence on both object recognition and image style transformation tasks. In this paper, we investigate the following question: given an unordered dataset where instances may exhibit some implicit order, can we find this order in an unsupervised fashion?

We argue that such order often exists even when dealing with the data that are naturally thought of as being i.i.d. sampled from a common, though complex, distribution. For example, let’s consider a dataset consisting of the joint locations on the body of the same person taken on different days. The data i.i.d. assumption is justified since postures of a person taken on different days are likely unrelated. However, we can arrange the data instances so that the joints follow an articulated motion or a set of motions in a way that makes each pose highly predictable given the previous one. Although this arrangement depends on the person as, for example, ballerina’s pose might obey different dynamics than the pose of a tennis player, the simultaneous inference on the pose dynamics can lead to a robust model that explains the correlations among joints. To put differently, if we reschedule the frames of a video clip, the data can now be modeled by an i.i.d. model. Nevertheless, reconstructing the order leads to an alternative model where transitions between the frames are easier to fit. The ballerina’s dancing, if sampled very sparsely, can be thought of as a rescheduled video sequence that needs to be reordered such that a temporal model can generate it.

Another example for discovering the order in an unsupervised manner is to study the evolution or dynamics of a system that cannot be easily recorded, such as the evolution of a galaxy. Usually, the process takes millions or billions of years; and it is impossible for a human to collect serial data. A model, capable of extracting the order, can potentially recover the evolution through a snapshot of thousands of
Discovering Order in Unordered Datasets: Generative Markov Networks

We term our model Generative Markov Networks (GMNs). One of the key ideas in the design of GMNs is to use neural networks, the transitional operator of GMNs can also model the data instances as if they were generated from a Markov chain. We propose a greedy batch-wise permutation scheme that allows fast training. Fig. 1 illustrates our proposed model.

To approach this problem, we model the data instances by treating them as if they were generated from a Markov chain. We propose to simultaneously train the transition operator and find the order given the current instance that is generated at the $t$-th step of the Markov chain, starting at each of the support instances, evaluate the query instance under generated samples, and assign the label with the highest log-likelihood probability.

Empirically, we demonstrate that GMNs are able to discover implicit orders among data instances while performing comparably with many of the current state-of-the-art methods on one-shot recognition tasks.

2. Generative Markov Networks

Let $\{s_i\}_{i=1}^{n}$ denote our training data which are assumed to be generated from an unknown Markov chain. Our goal is to jointly recover the unknown Markov chain as well as the order of the generation process. Note that since the generation order is unknown, even if the true Markov chain was given, it would still be computationally intractable to find the optimal order that best fits our data. To get around this intrinsic difficulty, as we show in Sec. 2.2, we propose a greedy algorithm to find an order given the current estimation of the transition operator.

We denote the underlying data order to be a permutation $\pi$. $\pi = \{\pi(t)\}_{t=1}^{n}$, where $\pi(t)$ represents the index of the instance that is generated at the $t$-th step of the Markov chain. In other words, a Markov chain is formed as follows:

$s_{\pi(1)} \rightarrow s_{\pi(2)} \rightarrow \cdots \rightarrow s_{\pi(n)}$.

We consider all the possible permutations $\pi$ and arbitrary distribution over these permutations, which leads to a joint log-likelihood estimation problem:

$$\max_{\theta, \pi} \log(P(\pi) \mathcal{P}(\{s_i\}_{i=1}^{n}, \pi; \theta))$$

**Figure 1.** Illustration of the proposed Generative Markov Networks (GMNs). The data are assumed to be generated from a Markov chain, and GMNs learn the permutation and the transitional operator simultaneously. The figure illustrates that we first train GMNs for the data in green circles (without knowing the data order) to recover the data permutation. Then, we apply the trained transition operator to the data in blue circles (without knowing the data order) to infer data permutation.

data points (by collecting images from different galaxies at the same time), allowing us to better understand how the galaxy changes over time. Furthermore, consider modeling molecular or cellular data instances. These processes are often too fast to permit entire trajectories or too destructive to allow multiple sampling. On the other hand, our model can provide an alternative way to learn the dynamics of the biological system.

One naive and obvious way to find the order in a dataset is to perform sorting based on a predefined distance metric; e.g., the Euclidean distance between image pixel values. However, the distance metrics have to be predefined differently according to distinct types/characteristics of the datasets at hand. A proper distance metric for one domain may not be a good one for other domains. For instance, the $p-$distance is a good measure for DNA/RNA sequences (Nei & Kumar, 2000) while it does not characterize the semantic distances between images. We argue that the key component of the ordering problem lies in the discovery of proper distance metric automatically and adaptively.
Discovering Order in Unordered Datasets: Generative Markov Networks

where \( P^{(1)}(\cdot) \) is the initial distribution of the Markov chain and \( T(s'|s; \theta) \) is the transition operator parametrized by model parameters \( \theta \). Note that the effect of the initial distribution \( P^{(1)}(\cdot) \) diminishes with the increase of the data size \( n \). Hence, without loss of generality, we assume \( P^{(1)}(\cdot) \) is uniform over all possible states, leading to the following optimization problem:

\[
\max_{\theta, \pi} \log \left( P(\pi) \prod_{t=2}^{n} T(s_{\pi(t)}|s_{\pi(t-1)}; \theta) \right),
\]

where \( \pi \in \Pi(n) \) and \( \Pi(n) \) is the set of all possible permutations over \([n]\). Unfortunately, direct optimization of (1) is computationally intractable. For each fixed \( \theta \) and \( P \), the number of all possible permutations (i.e., \(|\Pi(n)|\)) is \( n! \).

Proposition 1. Given the transition operator \( T \) and a set of instances \( \{s_i\}_{i=1}^{n} \), finding the optimal order in a Markov chain is NP-hard.

Proof. See Supplementary for the complete proof. In short, we can prove it by constructing a polynomial time reduction from a variant of the traveling salesman path problem (TSP). And since TSP is NP-complete, this shows that finding optimal order in a Markov chain is NP-hard. \( \square \)

2.1. Parametrized transition operator via Neural Networks

In practice, when the state space is large, we cannot often afford to maintain the tabular transition matrix directly, which takes up to \( O(d^2) \) space, where \( d \) is the number of states in the chain. For example, if the state refers to a binary image \( I \in \{0, 1\}^d \), the size of the state space is \( d = 2^d \) which is nearly infeasible to compute. Before optimizing (1), we should first find a family of functions to parametrize the transition operator \( T(\cdot|\cdot) \).

Being universal function approximators (Hornik et al., 1989), neural networks could be used to efficiently approximate the discrete structures which led to the recent success of deep reinforcement learning (Mnih et al., 2013). In our case, we utilize a neural network to approximate the discrete tabular transition matrix. The advantages are two-fold: first, it significantly reduces the space complexity by amortizing the space required by each separate state into a unified model. Since all the states share the same model as the transition operator, there is no need to store the transition vector for each separate state explicitly. Second, neural networks allow better generalization for the transition probabilities across states. The reason is that, in most real-world applications, states, represented as feature vectors, are not independent from each other. As a result, the differentiable approximation to a discrete structure has the additional smoothness properties, which allows the transition operator to have a good estimate even for the unseen states.

Let \( \theta \) be the parameters of the neural network. We can define

\[
f_\theta(s, s') = T(s'|s; \theta) : \mathbb{R}^d \times \mathbb{R}^d \to [0, 1]
\]

to be the transition function that takes two states \( s \) and \( s' \) as its input and returns the corresponding transition probability. Recall that \( s' \) and \( s \) are a vector of features. Note that one can consider each discrete transition operator as a lookup table; for example, we use \( s \) and \( s' \) to locate the corresponding row and column of the table and read out its probability. From this perspective, the neural network works as a soft lookup table that outputs the transition probability given two states (features).

2.2. Greedy Approximation of the Optimal Order

As mentioned above, the direct evaluation of Eq. (1) is computationally intractable given \( P \) and \( \theta \). Here, we develop a coordinate ascent style training algorithm to optimize Eq. (1) efficiently. The key insight comes from the observation that given \( \theta \), the optimization problem over \( \pi \) approximates to finding the optimal permutation \( \pi^\ast \) that achieves the maximum likelihood of the generating data:

\[
\max_{\theta, \pi} \log \left( P(\pi) \prod_{t=2}^{n} T(s_{\pi(t)}|s_{\pi(t-1)}; \theta) \right)
\]

= \[ \max_{\theta} \log \left( \prod_{t=2}^{n} T(s_{\pi^\ast(t)}|s_{\pi^\ast(t-1)}; \theta) \right)
\]

with \( \pi^\ast = \arg \max_{\pi \in \Pi(n)} \sum_{t=2}^{n} \log T(s_{\pi(t)}|s_{\pi(t-1)}; \theta) \).

The proof is straightforward and we leave it in Supplementary.

However, without further assumptions on the structure of the transition operator, this is still a hard problem which takes time \( O(n!) \) (see Proposition 1). Instead, we propose a greedy algorithm to approximate the optimal order, which takes time \( O(n^2 \log n) \). We list the pseudocode in Alg. 1.

At first, Alg. 1 enumerates all the possible states appearing in the first time step. For each of the following steps, it finds the next state by maximizing the transition probability at the current step, i.e., a local search to find the next state. The final approximate order is then defined to be the maximum of all these \( n \) orders. A naive implementation of this algorithm has time complexity \( O(n^3) \). However, we can reduce
Algorithm 1 Greedy Approximate Order

**Input:** Input data \( \{ s_i \}_{i=1}^n \) and transition operator \( \mathcal{T}(s|s;j;\theta) \)

1: \( v^* \leftarrow -\infty \)
2: for \( i = 1 \) to \( n \) do
3: \( \pi_i(1) \leftarrow i \)
4: for \( j = 2 \) to \( n \) do
5: \( \pi_i(j) \leftarrow \max_{s_k \in \{s_i(1),\ldots,s_i(j-1)\}} \mathcal{T}(s_k | s_{pi_i(j-1)};\theta) \)
6: end for
7: \( v_i \leftarrow \sum_{t=1}^n \log \mathcal{T}(s_{\pi_i(t-1)}|s_{\pi_i(t)};\theta) \)
8: if \( v_i > v^* \) then
9: \( v^* \leftarrow v_i \)
10: \( \hat{\pi} \leftarrow \pi_i \)
11: end if
12: end for

**Output:** \( \hat{\pi} \)

Algorithm 2 Batch-Wise Permutation Training

**Input:** \( \{ s_i \}_{i=1}^n, b, \ell, \gamma \)

1: Initialize \( \theta^{(0)} \), \( \{ x_i^{(0)} \}_{i=1}^b \)
2: for \( k = 1 \) to \( \infty \) do
3: if \( k \equiv 1 \pmod{\ell} \) then
4: Sample \( \{ x_i^{(k)} \}_{i=1}^b \sim \{ s_i \}_{i=1}^n \)
5: \( \{ x_i^{(k)} \}_{i=1}^b = \{ x_i^{(k)} \}_{i=1}^b \}
6: end if
7: Compute \( \hat{\pi}^{(k)} \) using the Greedy Approximate Order (Alg. 1)
8: Compute \( \nabla_\theta \log \mathcal{P}(\{ x_i \}_{i=1}^b ;\theta^{(k-1)}) = \partial_\theta \sum_{t=1}^b \log \mathcal{T}(x_{\hat{\pi}_t(k-1)}|x_{\hat{\pi}_t(k-1)};\theta^{(k-1)}) \)
9: \( \theta^{(k)} = \theta^{(k-1)} + \gamma \nabla_\theta \log \mathcal{P}(\{ x_i \}_{i=1}^b ;\theta^{(k-1)}) \)
10: end for

\( \hat{\pi} \) can be seen as a source of randomness (i.e., \( \mathcal{P}(s, z) = \mathcal{P}(s)\mathcal{P}(z) \)), and we use Variational Bayes Inference (Wainwright et al., 2008) for approximating \( \log \mathcal{T}(s'|s;\theta) \) as follows:

\[
\log \mathcal{T}(s'|s;\theta) = \int_z \mathcal{P}(s'|z, s;\psi) \mathcal{P}(z) dz
\]

\[
= \log \int_z \mathcal{Q}(z|s;\phi) \mathcal{P}(s'|z, s;\psi) \mathcal{P}(z)
\]

\[
\geq \int_z \mathcal{Q}(z|s;\phi) \log \frac{\mathcal{P}(s'|z, s;\psi) \mathcal{P}(z)}{\mathcal{Q}(z|s;\phi)} (\text{Jensen’s inequality})
\]

\[
= \mathbb{E}_{z \sim \mathcal{Q}(z|s;\phi)} \left[ \log \mathcal{P}(s'|z, s;\psi) \right] - KL \left( \mathcal{Q}(z|s;\phi) \| \mathcal{P}(z) \right)
\]

where \( \mathcal{T}(s'|s;\theta) \) has been replaced by a distribution \( \mathcal{P}(s'|s, z;\psi) \) parametrized by \( \psi \), which takes the inputs \( s \) and \( z \); \( KL \)-divergence, \( \mathcal{Q}(z|s;\phi) \) is an encoder function parametrized by \( \phi \) that encodes latent code \( z \) given current state \( s \), and \( \mathcal{P}(z) \) is a fixed prior which we take to be a Gaussian distribution \( \mathcal{N}(0,1) \). We use reparametrized trick to draw \( \mathcal{Q}(z|s;\phi) \) from Gaussian \( \mathcal{N}(\mu_{Q,s}(s),\sigma_{Q,s}^2(s)I) \) where \( \mu_{Q,s}(s) \) and \( \sigma_{Q,s}(s) \) are learnable functions.

We also adopt the conditional independence assumption for \( \mathcal{T}(s'|s) \) and \( \mathcal{P}(s'|s) \), i.e.,

\[
\mathcal{T}(s'|s) = \prod_{l=1}^p \mathcal{T}(s'|s) \quad \text{and} \quad \mathcal{P}(s'|s) = \prod_{l=1}^p \mathcal{P}(s'|s, z),
\]

where \( s' \in \mathbb{R}^p \) and \( s' \) is its \( l_{th} \) feature.

Then, \( \mathcal{T}(s'|s) \) is ensured to be a proper transition operator. Next, we consider two types of distribution family for \( \mathcal{P}(s'|s, z;\theta) \): Bernoulli and Gaussian. If \( s \in \{0,1\}^p \) (i.e., a binary image), we define \( \log \mathcal{P}(s'|s, z;\psi) \) as:

\[
s' \odot \log \left( g_\psi(s, z) \right) + (1-s') \odot \log \left( 1 - g_\psi(s, z) \right),
\]

where \( \odot \) is element-wise multiplication and \( g_\psi(s, z) : \{0,1\}^p \times \mathbb{R}^z \rightarrow \{0,1\}^p \).
Discovering Order in Unordered Datasets: Generative Markov Networks

If \( s \in \mathbb{R}^p \) (i.e., a real-valued feature vector), we choose \( P(s'|s, z; \psi) \) to be fixed variance factored Gaussian \( \mathcal{N}(\mu_{s', \psi}(s, z), \sigma_{s', \psi}^2 I) \), where \( \mu_{s', \psi}(s, z) : \mathbb{R}^{p+x} \to \mathbb{R}^p \) and \( \sigma_{s'} \) is a fixed variance. We simply choose \( \sigma_{s'} \) in all the experiments. \( \log P(s'|s, z; \theta) \) can thus be defined as

\[
-\frac{1}{2\sigma_{s'}} ||s' - \mu_{s', \psi}(s, z)||^2_2 + \text{const.}
\]

For simplicity, we set \( \theta = \{\psi \cup \phi\} \). Please refer to Supplementary for more details.

3. Experiments

In this section, we first qualitatively evaluate Generative Markov Networks (GMNs) on the Horse (Borenstein & Ullman, 2002) and MSR SenseCam (Jojic et al., 2010) datasets (in which data were assumed to be i.i.d.) to discover their implicit orders. Next, we use UCF-CIL Action dataset (Shen & Foroosh, 2008) (with order information) to quantitatively evaluate GMN and its generalization ability. Last, we adopt GMN on miniImageNet (Vinyals et al., 2016; Ravi & Larochelle, 2017) dataset for one-shot image recognition task. Additional experiments on MNIST (LeCun et al., 1990) and Moving MNIST (Srivastava et al., 2015) datasets are provided in Supplementary.

3.1. Discovering Implicit Order

3.1.1. Datasets Details

<Horset> Horse dataset (Borenstein & Ullman, 2002) consists of 328 horse images collected from the Internet. Each horse is centered in a 30x40 image. For the preprocessing, the author applied foreground-background segmentation and set the pixel value to 1 and 0 for object and background, respectively. Examples are shown Supplementary.

<MSR_SenseCam> MSR SenseCam (Jojic et al., 2010) is a dataset consisting of images taken by SenseCam wearable camera. It contains 45 classes with approximately 150 images per class. Each image has size 480x640. We resize each image into 224x224 and extract the feature from VGG-19 network (Simonyan & Zisserman, 2014). In this dataset, we consider only office category which has 362 images.

3.1.2. Implicit Orders

We apply Alg. 2 to train GMNs. When the training converges, we plot the images following permutation \( \pi \) in Alg. 1. Note that \( \pi \) can be seen as the implicit order suggested by GMNs. For comparison, we also plot the images following nearest neighbor sorting using Euclidean distances. Hyperparameters and network architectures for parameterizing \( T(\cdot; \cdot; \theta) \) are specified in Supplementary.

The results are shown in Fig. 2 (a). We first observe that consecutive frames in the order extracted by the GMN have visually high autocorrelation, implying that GMNs can discover some implicit order for these datasets. However, it is hard to qualitatively compare the results with Nearest Neighbor’s. To address this problem, we perform another evaluation in the following subsection and also perform quantitative analysis in Sec. 3.2.

3.1.3. Image Propagation

Now, we provide an alternative evaluation for GMNs with a proposed term: image propagation, which is defined as follows: given an image \( s \), the propagated image \( s' = \arg \max_{s' \in \{s_n\} \setminus \{s\}} T(s'|s; \theta) \).
Discovering Order in Unordered Datasets: Generative Markov Networks

3.2. Recovering Orders in Ordered Datasets

UCF-CIL Action Dataset: UCF-CIL Action Dataset (Shen & Foroosh, 2008) is a dataset containing different action videos by various subjects, distinct cameras, and diverse viewpoints. In addition to videos, the dataset also provides 11 tracking points (presented in x and y axis) for head, right shoulder, right elbow, right hand, left shoulder, left elbow, left hand, right knee, right foot, left knee, and left foot. We represent a frame by concatenating 11 tracking points at the end of pre-extracted features from VGG-19 network (Simonyan & Zisserman, 2014). In this dataset, we select ballet fouette actions to evaluate GMNs. Hyper parameters choice, model architecture, and additional experiments on tennis serve actions are provided in Supplementary.

We use Kendall Tau-b metric (Kendall et al., 1946) to measure the distance between two permutations, which outputs the value from −1 to 1. The larger the Kendall Tau-b value, the more similar the two permutations are. For example, if Kendall Tau-b value is 1, two permutations are exactly the same. In the following, we evaluate GMNs under two settings: (1) train GMNs and evaluate the order for the same subject, and (2) train GMNs from one subject but evaluate on the different subject. The results are reported for random 20 trials with mean and standard deviation.

3.2.1. Orders Recovery

Even though by simulating Markov chain generation process, GMNs can jointly discover order in the dataset, this order is strictly implicit. Hence, we investigate how well it does in cases when we do have ground truth ordering. In this experiment, we take a video sequence and then reshuffle its frames, then attempt to reconstruct the order.

In Tbl. 1, we report the Kendall Tau-b values between the true order and the recovered order from GMNs/ Nearest Neighbor (NN) with Euclidean distance. GMNs are trained on different subjects with ballet fouette actions.

Next, we examine the generalization ability for GMNs. In most of the cases, when we apply the trained GMNs (i.e., apply the trained transition operator in GMNs) for the different subject, we can achieve higher Kendall Tau-b values comparing to Nearest Neighbor. The exceptions are highlighted as red color in Tbl. 1. This result implies that GMNs has the ability to learn the transition operator that can generalize from one subject to another.

3.3. One-Shot Recognition

miniImageNet (Vinyals et al., 2016; Ravi & Larochelle, 2017) is a benchmark dataset designed...
for the evaluation of one-shot learning (Vinyals et al., 2016; Ravi & Larochelle, 2017). Being a subset of ImageNet (Russakovsky et al., 2015), it contains 100 classes and each class has 600 images. Each image is downsampled to size 84x84. As suggested in (Ravi & Larochelle, 2017), the dataset is divided into three parts: 64 classes for training, 16 classes for validation, and 20 classes for testing.

Identical to (Ravi & Larochelle, 2017), we consider the 5-way 1-shot problem. That is, from testing classes, we sample 5 classes with each class containing 1 labeled example. The labeled examples refer to support instances. Then, we randomly sample 500 unlabeled query examples in these 5 classes for evaluation. We repeat this procedure for 10,000 times and report the average with 95% confidence intervals in Tbl. 2.

### 3.3.1. Training Details

Same as (Vinyals et al., 2016; Ravi & Larochelle, 2017), we train GMNs on training classes and then apply it to testing classes. For each training episode, we sample 1 class from the training classes and let \( \{ s_i \}_{i=1}^{20} \) be all the data from this class. We consider 3,000 training episodes.

On the other hand, for each testing episode, we apply GMNs to generate a chain from each support instance:

\[
\hat{s}_1 \sim \mathcal{T}(\cdot|s_0^c; \theta), \quad \hat{s}_2 \sim \mathcal{T}(\cdot|\hat{s}_1^c; \theta), \quad \ldots, \quad \hat{s}_k \sim \mathcal{T}(\cdot|\hat{s}_{k-1}^c; \theta),
\]

where \( s_0^c \) is the support instance belonging to class \( c \) and \( \hat{s}^c \) is the generated samples from the Markov chain.

Next, we fit each query example into each chain by computing the average approximating log-likelihood. Namely, the probability for generating the query sample \( q \) in the chain of class \( c \) is

\[
P(s_q|c) := \frac{1}{k+1} \left( \log \mathcal{T}(s_q|s_0^c; \theta) + \sum_{i=1}^{k} \log \mathcal{T}(s_q|\hat{s}_i^c; \theta) \right).
\]

In a generative viewpoint, the predicted class \( \hat{c} \) for \( s_q \) is determined by

\[
\hat{c} = \arg \max_c P(s_q|c).
\]

For fair comparisons, we use the same architecture specified in (Ravi & Larochelle, 2017) to extract 1600-dimensional features. We pretrain the architecture using standard softmax regression on image-label pairs in training and validation classes. The architecture consists of 4 blocks. Each block comprises a CNN layer with 64 3x3 convolutional filters, Batch Normalization (Ioffe & Szegedy, 2015) layer, ReLU activation, and 2x2 Max-Pooling layer. Then, we train GMNs based on these 1, 600 dimensional features. Hyperparameters and network architecture are specified in Supplementary.

### 3.3.2. Results

We compare GMNs and the related approaches in Tbl. 2, in which Basic model refers the architecture in (Ravi & Larochelle, 2017) and Advanced models refer to more complicated designs. Generally, it is not fair to compare the methods using different architecture designs; therefore, we only discuss the methods using Basic model in the following. Please refer to Sec. 4 for the details and discussions on the related approaches.

First, we compare GMNs with Parametric methods, which are meta-learning methods. The best result is reported by Meta-SGD (Li et al., 2017) with 50.47 ± 1.87. Although GMN suffers from the performance drop, it requires a much less computational budget. The reason is that the meta-learning approaches (Ravi & Larochelle, 2017; Finn et al., 2017; Munkhdalai & Yu, 2017; Li et al., 2017; Mishra et al., 2017) rely on huge networks to manage complicated intersections between meta and base learners, while parameters for GMN exist only in \( \theta \) which is a relatively tiny network.
On the other hand, the best performance reported in the Nonparametric approaches is Prototypical Networks (Snell et al., 2017) with 49.42 ± 0.78. Note that these Nonparametric approaches are classified as distance-metric learning approaches (see Sec. 4). While GMNs suffers from performance deterioration, it enjoys more flexibility without the need of defining any distance metric as in (Vinyals et al., 2016; Koch et al., 2015; Triantafillou et al., 2017; Snell et al., 2017; Shyam et al., 2017; Mehrotra & Dukkipati, 2017). The reason is that we determine the class for query sample by fitting it into the simulated Markov chain for support instances.

4. Related Works and Discussions

The proposed Generative Markov Networks assumes a Markov chain data generation process, and one of the fundamental ideas is utilizing a deep neural network for amortizing the large state space in the transitional operator. In this Section, we provide brief discussions on related deep generative models. Since we have also applied GMNs into one-shot learning tasks, we will discuss the differences between our model and recent researches in one-shot learning.

**Deep Generative Models:** Deep generative models can be categorized into two classes (Song et al., 2017) based on the sampling methods: ancestral sampling and iterative sampling. Instances of ancestral sampling models are Variational Auto-encoders (VAEs) (Kingma & Welling, 2013), Generative Adversarial Networks (GANs) (Goodfellow et al., 2014), and their families. In the inference phase, these approaches generated one sample from the model by performing a single inference pass from latent variables to data. Lots of attention are attracted in supervised (Lin et al., 2017), semi-supervised (Dai et al., 2017), and unsupervised (Hsu et al., 2017) applications.

As a comparison, methods based on iterative sampling performed multiple and iterative passes through all the variables in the corresponding graphical models. Usually, these methods involved simulating a Markov chain in the entire state space, and they aimed at improving quality of generated samples by mixing the underlying chain. Recent works on this line of research included (Salakhutdinov et al., 2007; Hinton, 2009; Salakhutdinov & Larochelle, 2010; Bengio et al., 2013; 2014; Sohl-Dickstein et al., 2015; Bordes et al., 2017; Song et al., 2017).

Our approach can be categorized as an iterative sampling-based model. However, it has three significant differences comparing to previous works. First, existing works assumed training instances are i.i.d. sampled from the stationary distribution of a Markov chain. On the other hand, we assume the data are sequentially generated from a Markov in an unknown order. The order can be recovered by our proposed GMN and thus help us understand the implicit data relationships. Second, prior approaches were proposed based on the notion of denoising models. In other words, their goal was generating high-quality images; on the other hand, we aim at discovering an implicit order in the dataset. Third, to the best of our knowledge, all the existing works were implicit models in the sense that they only admitted efficient sampling schemes. In contrast, the proposed GMN is an explicit model where besides an efficient sampling procedure, the model maintains a tractable likelihood function that can be computed efficiently.

**One-Shot Learning:** Here, we focus on deep one-shot learning models which could be divided into two categories: nonparametric and parametric approaches. The former aimed at either learning a similarity measurement between instance pairs (Koch et al., 2015) or applying specific metric loss based on cosine distance (Vinyals et al., 2016) or Euclidean distance (Snell et al., 2017). Usually, these methods relied heavily upon human design.

As a comparison, methods in the second category offered more generalities. Typically, this type of methods tackled the problem using a meta-learning framework to train the parametric classifiers. Precisely, they considered two levels of learning: the first stage is to update base learners’ parameters and the second stage is to update parameters for the meta learner. Recent works (Ravi & Larochelle, 2017; Kaiser et al., 2017; Finn et al., 2017) belonged to this category.

GMNs is an instance of the first category and has the following difference. Instead of designing a specific loss for one-shot learning purpose, we employ Markov chain data generation assumption and determine the label of query instance by fitting it into the simulated Markov chain from support instances.

5. Conclusion

In this paper, we argue that data i.i.d. assumption is not always the case in most of the datasets. Often, data instances exhibit implicit order which may benefit our understanding and analysis of the dataset. To observe the implicit order, we propose a novel Generative Markov Networks which considers a Markov chain data generation scheme. Specifically, we simultaneously learn the transitional operator as a generative model in the Markov chain as well as find the optimal orders of the data under all possible permutations. In lots of experiments, we show that our model is able to observe implicit order from unordered datasets, recover order from ordered datasets, and perform well on the one-shot recognition task.
Discovering Order in Unordered Datasets: Generative Markov Networks

References

Andrychowicz, Marcin, Denil, Misha, Gomez, Sergio, Hoffman, Matthew W, Pfau, David, Schaul, Tom, and de Freitas, Nando. Learning to learn by gradient descent by gradient descent. In Advances in Neural Information Processing Systems, pp. 3981–3989, 2016.

Bengio, Yoshua, Yao, Li, Alain, Guillaume, and Vincent, Pascal. Generalized denoising auto-encoders as generative models. In Advances in Neural Information Processing Systems, pp. 899–907, 2013.

Bengio, Yoshua, Laufer, Eric, Alain, Guillaume, and Yosinski, Jason. Deep generative stochastic networks trainable by backprop. In International Conference on Machine Learning, pp. 226–234, 2014.

Bordes, Florian, Honari, Sina, and Vincent, Pascal. Learning to generate samples from noise through infusion training. arXiv preprint arXiv:1703.06975, 2017.

Borenstein, Eran and Ullman, Shimon. Class-specific, top-down segmentation. In European conference on computer vision, pp. 109–122. Springer, 2002.

Dai, Zihang, Yang, Zhilin, Yang, Fan, Cohen, William W, and Salakhutdinov, Ruslan R. Good semi-supervised learning that requires a bad gan. In Advances in Neural Information Processing Systems, 2017.

Finn, Chelsea and Levine, Sergey. Deep visual foresight for planning robot motion. In Robotics and Automation (ICRA), 2017 IEEE International Conference on, pp. 2786–2793. IEEE, 2017.

Finn, Chelsea, Abbeel, Pieter, and Levine, Sergey. Model-agnostic meta-learning for fast adaptation of deep networks. arXiv preprint arXiv:1703.03400, 2017.

Glorot, Xavier, Bordes, Antoine, and Bengio, Yoshua. Domain adaptation for large-scale sentiment classification: A deep learning approach. In Proceedings of the 28th international conference on machine learning (ICML-11), pp. 513–520, 2011.

Goodfellow, Ian, Pouget-Abadie, Jean, Mirza, Mehdi, Xu, Bing, Warde-Farley, David, Ozair, Sherjil, Courville, Aaron, and Bengio, Yoshua. Generative adversarial nets. In Advances in neural information processing systems, pp. 2672–2680, 2014.

Hinton, Geoffrey E. Deep belief networks. Scholarpedia, 4 (5):5947, 2009.

Hornik, Kurt, Stinchcombe, Maxwell, and White, Halbert. Multilayer feedforward networks are universal approximators. Neural networks, 2(5):359–366, 1989.

Hsu, Wei-Ning, Zhang, Yu, and Glass, James. Unsupervised learning of disentangled and interpretable representations from sequential data. In Advances in Neural Information Processing Systems, 2017.

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

Jojic, Nebojsa, Perina, Alessandro, and Murino, Vittorio. Structural epitome: a way to summarize ones visual experience. In Lafferty, J. D., Williams, C. K. I., Shawe-Taylor, J., Zemel, R. S., and Culotta, A. (eds.), Advances in Neural Information Processing Systems 23, pp. 1027–1035. Curran Associates, Inc., 2010. URL http://papers.nips.cc/paper/4092-structural-epitome-a-way-to-summarize-ones-visual-experience.pdf.

Kaiser, Łukasz, Nachum, Ofir, Roy, Aurko, and Bengio, Samy. Learning to remember rare events. arXiv preprint arXiv:1703.03129, 2017.

Kendall, Maurice George et al. The advanced theory of statistics. The advanced theory of statistics., (2nd Ed), 1946.

Kingma, Diederik and Ba, Jimmy. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

Kingma, Diederik P and Welling, Max. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114, 2013.

Koch, Gregory, Zemel, Richard, and Salakhutdinov, Ruslan. Siamese neural networks for one-shot image recognition. In ICML Deep Learning Workshop, volume 2, 2015.

Krizhevsky, Alex, Sutskever, Ilya, and Hinton, Geoffrey E. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, pp. 1097–1105, 2012.

LeCun, Yann, Boser, Bernhard E, Denker, John S, Henderson, Donnie, Howard, Richard E, Hubbard, Wayne E, and Jackel, Lawrence D. Handwritten digit recognition with a back-propagation network. In Advances in neural information processing systems, pp. 396–404, 1990.

Li, Zhenguo, Zhou, Fengwei, Chen, Fei, and Li, Hang. Meta-sgd: Learning to learn quickly for few shot learning. arXiv preprint arXiv:1707.09835, 2017.

Lin, Kevin, Li, Dianqi, He, Xiaodong, Sun, Ming-ting, and Zhang, Zhengyou. Adversarial ranking for language generation. NIPS, 2017.
Discovering Order in Unordered Datasets: Generative Markov Networks

Mehrotra, Akshay and Dukkipati, Ambedkar. Generative adversarial residual pairwise networks for one shot learning. *arXiv preprint arXiv:1703.08033*, 2017.

Mishra, Nikhil, Rohaninejad, Mostafa, Chen, Xi, and Abbeel, Pieter. Meta-learning with temporal convolutions. *arXiv preprint arXiv:1707.03141*, 2017.

Mnih, Volodymyr, Kavukcuoglu, Koray, Silver, David, Graves, Alex, Antonoglou, Ioannis, Wierstra, Daan, and Riedmiller, Martin. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

Munkhdalai, Tsendsuren and Yu, Hong. Meta networks. *arXiv preprint arXiv:1703.00837*, 2017.

Nei, Masatoshi and Kumar, Sudhir. *Molecular evolution and phylogenetics*. Oxford university press, 2000.

Ravi, Sachin and Larochelle, Hugo. Optimization as a model for few-shot learning. *ICLR*, 2017.

Russakovsky, Olga, Deng, Jia, Su, Hao, Krause, Jonathan, Satheesh, Sanjeev, Ma, Sean, Huang, Zhiheng, Karpathy, Andrej, Khosla, Aditya, Bernstein, Michael, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

Salakhutdinov, Ruslan and Larochelle, Hugo. Efficient learning of deep boltzmann machines. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 693–700, 2010.

Salakhutdinov, Ruslan, Mnih, Andriy, and Hinton, Geoffrey. Restricted boltzmann machines for collaborative filtering. In *Proceedings of the 24th international conference on Machine learning*, pp. 791–798. ACM, 2007.

Shen, Yuping and Foroosh, Hassan. View-invariant action recognition using fundamental ratios. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pp. 1–6. IEEE, 2008.

Shyam, Pranav, Gupta, Shubham, and Dukkipati, Ambedkar. Attentive recurrent comparators. *arXiv preprint arXiv:1703.00767*, 2017.

Simonyan, Karen and Zisserman, Andrew. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Snell, Jake, Swersky, Kevin, and Zemel, Richard S. Prototypical networks for few-shot learning. *arXiv preprint arXiv:1703.05175*, 2017.

Sohl-Dickstein, Jascha, Weiss, Eric A, Maheswaranathan, Niru, and Ganguli, Surya. Deep unsupervised learning using nonequilibrium thermodynamics. *arXiv preprint arXiv:1503.03585*, 2015.
Supplementary for Generative Markov Networks

Yao-Hung Hubert Tsai 1 Han Zhao 1 Nebojsa Jojic 2 Ruslan Salakhutdinov 1

1. Proof of Proposition 1

Here, we prove that finding the optimal order in a Markov chain with a given transitional operator and a set of instances is NP-hard. We do so by constructing a polynomial time reduction from a variant of the travelling salesman problem (TSP). To proceed, we first formally define the travelling salesman path problem (TSP) and our optimal order problem in Markov chain (ORDER).

Definition 1.1 (TSP). INSTANCE: A weighted undirected graph \( G = (V, E, w) \) and a constant \( p \). QUESTION: Is there a Hamiltonian path in \( G \) whose weight is at most \( C \)?

Definition 1.2 (ORDER). INSTANCE: A Markov chain \( M = (S, T, q) \), a sequence of states \( X \) and a constant \( p \), where \( S = [n] \) is the set of states, \( T \in \mathbb{Q}^{n \times n} \) is the transitional matrix and \( q \) is the initial distribution. QUESTION: Is there a permutation \( \pi \) for \( X \) under which \( M \) generates \( \pi(X) \) with probability at least \( p \)?

Theorem 1.1 ((Garey & Johnson, 2002)). TSP is NP-complete.

Theorem 1.2. TSP \( \leq_{P} \) ORDER.

Proof. Given an instance of TSP, \( (G = (V, E, w), C) \), we shall construct an instance of ORDER, \( (M = (S, T, q), X, p) \) in polynomial time such that the answer to the latter is “yes” if and only if the answer to the first is also “yes”.

Given \( (G = (V, E, w), C), n = |V| \), define \( \Delta := \max_{i \in [n]} \sum_{j=1}^{n} \exp(-w(e_{ij})) \). With \( \Delta \), we construct a Markov chain \( M \) where the state set is \( S = V \) and the transitional matrix is defined as:

\[
T_{ij} := \frac{\exp(-w(e_{ij}))}{\Delta}, \forall e_{ij} \in E
\]

and

\[
T_{ii} := 1 - \sum_{j \neq i} T_{ij}.
\]

Set the initial distribution of \( M \) to be uniform, i.e., \( q = 1/n \cdot 1_n \), where \( 1_n \) is a \( n \)-dimensional vector of all ones. Choose \( X = S = [n] \) and \( p = \exp(-C)/n\Delta^n \).

Now we show \( \langle G = (V, E, w), C \rangle \in \mathcal{L}(TSP) \Rightarrow \langle M = (V, T, 1/n \cdot 1_n), \exp(-C)/n\Delta^n \rangle \in \mathcal{L}(ORDER) \). If \( \langle G = (V, E, w), C \rangle \in \mathcal{L}(TSP) \), then there exists a Hamiltonian path over \( V \) such that the weight of the path \( \leq C \). In other words, there exists a permutation \( \pi \) over \( V \) such that \( \sum_{i=2}^{n} w(\pi_{i-1}, \pi_{i}) \leq C \), but this in turns implies:

\[
\Pr(\pi(X); M) = \frac{1}{n} \prod_{i=2}^{n} \Pr(\pi(X_i) | \pi(X_{i-1})) = 1 - \frac{n}{n} \exp\left(\frac{\sum_{i=2}^{n} w(\pi_{i-1}, \pi_{i})}{\Delta}ight) = \frac{p}{\Delta^n}\]

which shows \( \langle M = (V, T, 1/n \cdot 1_n), \exp(-C)/n\Delta^n \rangle \in \mathcal{L}(ORDER) \). The other direction is exactly the same, i.e., if we know that \( \Pr(\pi(X); M) \geq 1 - \frac{C}{n\Delta^n} = p \), then we know that there is a Hamiltonian path in \( G \) with weight \( \leq C \).

Corollary 1.1. ORDER is NP-complete.

Proof. We have proved that TSP \( \leq_{P} \) ORDER. Along with the fact that TSP is NP-complete, this shows ORDER is NP-hard. Clearly, ORDER \( \in \mathbb{NP} \) as well, because a non-deterministic machine can first guess a permutation \( \pi \) and verify that \( \Pr(\pi(X); M) \geq p \), which can be done in polynomial time in both \( n \) and \( |X| \). This finishes our proof that ORDER is NP-complete.

2. Proof in Section 2.2

Proof. For any \( \pi \) over \( \Pi(n) \), we have:

\[
\log \left( \mathcal{P}(\pi) \prod_{t=2}^{n} T(s_{\pi(t)}|s_{\pi(t-1)}; \theta) \right)
\]

\[
\leq \log \left( \mathcal{P}(\pi^*) \prod_{t=2}^{n} T(s_{\pi^*(t)}|s_{\pi^*(t-1)}; \theta) \right)
\]

\[
= \log \left( \mathcal{P}(\pi^*) \prod_{t=2}^{n} T(s_{\pi^*(t)}|s_{\pi^*(t-1)}; \theta) \right)
\]

\[
= \log \left( \prod_{t=2}^{n} T(s_{\pi^*(t)}|s_{\pi^*(t-1)}; \theta) \right)
\]
where \( I(\pi = \pi^*) \) is the indicator function that takes value 1 iff \( \pi = \pi^* \) otherwise 0. And it is easy to check that, the optimal distribution, \( P(\pi^*) \) over all permutation is going to be a point mass distribution, which means \( P(\pi^*) = 1 \). Note that \( I(\pi = \pi^*) \) also defines a valid distribution over \( \Pi(n) \), which proves our claim.

3. Example Images for Horse Dataset

Horse dataset (Borenstein & Ullman, 2002) consists of images collected from Internet. Fig. 1 illustrates these RGB images. Object-background segmentations are applied on these images and the horses are centered in 30x40 images. The processed images are shown in Fig. 2.

4. Full Ordering Results for Horse and MSR_SenseCam

Fig. 3 and 4 show the results of the implicit order observed from GMN the order implied from Nearest Neighbor sorting. On the other hand, Fig. 5 and 6 illustrate the image propagation of GMN and Nearest Neighbor search.

5. MNIST

\(<MNIST>\) MNIST (LeCun et al., 1990) is a well-studied dataset that contains 60,000 training examples. Each example is a digit image with size 28x28. We rescale the pixel values to \([0, 1]\). Note that since MNIST contains a large number of instances, we perform the ordering in a randomly sampled batch to demonstrate our results.

Fig. 7 illustrates the results for the implicit order observed from Generative Markov Networks and the order inferred from Nearest Neighbor sorting using Euclidean distance. On the other hand, Fig. 8 shows the results for the image propagation with the proposed GMNs and Nearest Neighbor search with Euclidean distance.

6. Moving MNIST

\(<Moving\ MNIST>\) Moving MNIST (Srivastava et al., 2015) contains 10,000 sequences each of length 20 showing 2 digits moving in a 64x64 frame. We rescale the pixel values to \([0, 1]\). For each training episode, we apply Alg. 2 to train GMN on one randomly chosen sequence. We consider 6,000 training episodes. To evaluate the generalization ability of GMNs, we randomly sample a disjoint sequence from training sequences and observe the optimal permutation (implicit order) from Alg. 1.

Fig. 9 illustrates the results for the implicit order observed from Generative Markov Networks, the order inferred from Nearest Neighbor sorting using Euclidean distance, and the suggested explicit order. We find that both the orders observed from GMN and NN sorting manifest smooth motions for two digits in the frame. It is worth noting that our proposed GMN enjoys the freedom of not defining any distance metric.

The image propagation results for Moving MNIST dataset are shown in Fig. 10. We consider two approaches: the proposed Generative Markov Networks and Nearest Neighbor search. We find that, by learning the transition operator in a Markov chain as a generative model, GMN performs much better image propagation results than Nearest Neighbor search which is a discriminative model.

7. UCF_CIL Action

In the main text, we have provided the experiments on ballet fouette actions. Fig. 11 illustrates the comparison between (a) true order, (b) order recovered from Nearest Neighbor, (c) order recovered from GMN trained on different subject, and (d) order recovered from GMN trained on the same subject.

Next, we provide the experiments on tennis serve actions. The results are provided in Tbl. 1. We can clearly see that, in most of the cases, the order implied by GMNs enjoy better Kendall Tau-b values than Nearest Neighbor, which means our proposed model can recover more accurate orders.

8. Hyper Parameters

Fig. 3 lists the hyper parameters choice. Note that for smaller datasets (i.e., Horse, MSR_SenseCam), we can directly train GMNs on entire dataset. In other words, \( b_o = b \) and \( t = 1 \). Note that the number of frames is reduced to 20 and 30 for ballet fouette and tennis serve actions, respectively.

9. Network Architectures for Transition Operator

We elaborate the design of the transition operator in Fig. 12. In our design, \( U \) can be seen as a gating mechanism between input \( X_t \) and the learned update \( \tilde{X} \). More precisely, the output can be written as

\[
X_{t+1} = U \odot \tilde{X} + (1 - U) \odot X_t,
\]

where \( \odot \) denotes element-wise product. We specify each function \( f \) in Tbl. 2, 4, 5, 6, 7, and 8. Note that we omit the bias term for simplicity. We use ADAM (Kingma & Ba, 2014) with learning rate 0.001 and 0.2 dropout rate to train our \( T(\cdot; \theta) \).
Figure 1. RGB images of Horse Dataset.

Figure 2. Pre-processed images of Horse Dataset.

Figure 3. For Horse dataset: (a) implicit order observed from GMNs (b) order implied from nearest neighbor sorting using Euclidean distance.

Figure 4. For office category in SenseCam dataset: (a) implicit order observed from GMNs (b) order implied from nearest neighbor sorting using Euclidean distance.

Table 1. Kendall Tau-b Metric between the true order and the recovered orders from GMN/ Nearest Neighbor (NN) with Euclidean distance. GMNs are trained on different subjects with tennis serve actions for 10 different subjects. The results are provided with mean and standard deviation from 20 random trials.

|         | Subject 1    | Subject 2    | Subject 3    | Subject 4    | Subject 5    | Subject 6    | Subject 7    | Subject 8    | Subject 9    | Subject 10   |
|---------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| GMN     | -0.085 ± 0.034 | -0.147 ± 0.093 | 0.652 ± 0.032 | 0.377 ± 0.030 | 0.541 ± 0.048 | 0.034 ± 0.028 | 0.292 ± 0.065 | 0.518 ± 0.036 | 0.308 ± 0.038 | -0.343 ± 0.044 |
| NN      | 0.039        | -0.039       | 0.669        | 0.159        | 0.480        | -0.172       | 0.209        | 0.370        | 0.126        | -0.163       |
Figure 5. For Horse dataset, *image propagation* from (a) learned transition operator in GMNs (b) nearest neighbor search using Euclidean distance.

Figure 6. For office category in SenseCam dataset, *image propagation* from (a) learned transition operator in GMNs (b) nearest neighbor search using Euclidean distance.

Figure 7. For MNIST dataset: (a) implicit order observed from GMNs (b) order implied from nearest neighbor sorting using Euclidean distance.
Figure 8. For MNIST dataset, image propagation from (a) learned transition operator in GMNs (b) nearest neighbor search using Euclidean distance.

Figure 9. For Moving MNIST dataset: (a) implicit order observed from GMNs (b) order implied from nearest neighbor sorting using Euclidean distance (c) suggested explicit order.

Figure 10. For Moving MNIST dataset, image propagation from (a) learned transition operator in GMNs (b) nearest neighbor search using Euclidean distance.

Figure 11. Ballet fouette action for subject 1 with: (a) true order, (b) order recovered from Nearest Neighbor, (c) order recovered from GMN trained on different subject, and (d) order recovered from GMN trained on the same subject.
Table 2. Details of functions for Horse experiments.

| function | details                  |
|----------|--------------------------|
| f1       | 1200x512 FC layer with ReLU |
| f21      | 512x128 FC layer          |
| f22      | 512x128 FC layer          |
| f3       | 1328x512 FC layer with ReLU |
| f41      | 512x1200 FC layer with sigmoid |
| f42      | 512x1200 FC layer with sigmoid |

Table 3. Hyper parameters choice.

| Hyper Parameters | b_o | b | t |
|------------------|-----|---|---|
| Horse            | 328 | 328 | 1  |
| MSR_SenseCam     | 362 | 362 | 1  |
| MNIST            | 50  | 500 | 600 |
| Moving MNIST     | 20  | 20  | 1  |
| UCF_CIL ballet fouette | 20  | 20  | 1  |
| UCF_CIL tennis serve | 30  | 30  | 1  |
| miniImageNet     | 20  | 100 | 10 |

Table 4. Details of functions for MSR_SenseCam experiments.

| function | details                  |
|----------|--------------------------|
| f1       | 4096x1024 FC layer with ReLU |
| f21      | 1024x256 FC layer          |
| f22      | 1024x256 FC layer          |
| f3       | 4352x1024 FC layer with ReLU |
| f41      | 1024x4096 FC layer with sigmoid |
| f42      | 1024x4096 FC layer          |

Table 5. Details of functions for MNIST experiments.

| function | details                  |
|----------|--------------------------|
| f1       | 784x512 FC layer with ReLU |
| f21      | 512x128 FC layer          |
| f22      | 512x128 FC layer          |
| f3       | 912x512 FC layer with ReLU |
| f41      | 512x784 FC layer with sigmoid |
| f42      | 512x784 FC layer with sigmoid |

Table 6. Details of functions for Moving MNIST experiments.

| function | details                  |
|----------|--------------------------|
| f1       | 4096x1024 FC layer with ReLU |
| f21      | 1024x256 FC layer          |
| f22      | 1024x256 FC layer          |
| f3       | 4352x1024 FC layer with ReLU |
| f41      | 1024x4096 FC layer with sigmoid |
| f42      | 1024x4096 FC layer          |

Table 7. Details of functions for UCF_CIL experiments.

| function | details                  |
|----------|--------------------------|
| f1       | 4118x1024 FC layer with ReLU |
| f21      | 1024x64 FC layer          |
| f22      | 1024x64 FC layer          |
| f3       | 4182x1024 FC layer with ReLU |
| f41      | 1024x4118 FC layer with sigmoid |
| f42      | 1024x4118 FC layer          |
Figure 12. Network design for $\mathcal{T} (\cdot | \cdot ; \theta)$.

Table 8. Details of functions for miniImageNet experiments.

| function | details |
|----------|---------|
| f1       | 1600x1024 FC layer with ReLU // 1024x512 FC layer with ReLU // 512x256 FC layer with ReLU |
| f21      | 256x64 FC layer |
| f22      | 256x64 FC layer |
| f3       | 1664x256 FC layer with ReLU |
| f41      | 256x512 FC layer with ReLU // 512x1024 FC layer with ReLU // 1024x1600 FC layer with sigmoid |
| f42      | 256x512 FC layer with ReLU // 512x1024 FC layer with ReLU // 1024x1600 FC layer |
**References**

Borenstein, Eran and Ullman, Shimon. Class-specific, top-down segmentation. In *European conference on computer vision*, pp. 109–122. Springer, 2002.

Garey, Michael R and Johnson, David S. *Computers and intractability*, volume 29. wh freeman New York, 2002.

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

LeCun, Yann, Boser, Bernhard E, Denker, John S, Henderson, Donnie, Howard, Richard E, Hubbard, Wayne E, and Jackel, Lawrence D. Handwritten digit recognition with a back-propagation network. In *Advances in neural information processing systems*, pp. 396–404, 1990.

Srivastava, Nitish, Mansimov, Elman, and Salakhutdinov, Ruslan. Unsupervised learning of video representations using LSTMs. In *ICML*, 2015.