Abstract
Integrating deep learning with latent state space models has the potential to yield temporal models that are powerful, yet tractable and interpretable. Unfortunately, current models are not designed to handle missing data or multiple data modalities, which are both prevalent in real-world data. In this work, we introduce a factorized inference method for Multimodal Deep Markov Models (MDMMs), allowing us to filter and smooth in the presence of missing data, while also performing uncertainty-aware multimodal fusion. We derive this method by factorizing the posterior \( p(z|x) \) for non-linear state space models, and develop a variational backward-forward algorithm for inference. Because our method handles incompleteness over both time and modalities, it is capable of interpolation, extrapolation, conditional generation, label prediction, and weakly supervised learning of multimodal time series. We demonstrate these capabilities on both synthetic and real-world multimodal data under high levels of data deletion. Our method performs well even with more than 50% missing data, and outperforms existing deep approaches to inference in latent time series.

Introduction
Virtually all sensory data that humans and autonomous systems receive can be thought of as multimodal time series—multiple sensors each provide streams of information about the surrounding environment, and intelligent systems have to integrate these information streams to form a coherent yet dynamic model of the world. These time series are often asynchronously or irregularly sampled, with many timepoints having missing or incomplete data. Classical time series algorithms, such as the Kalman filter, are robust to such incompleteness: they are able to infer the state of the world, but only in linear regimes (Kalman 1960). On the other hand, human intelligence is robust and complex: we infer complex quantities with nonlinear dynamics, even from incomplete temporal observations of multiple modalities—for example, intended motion from both eye gaze and arm movements (Dragan, Lee, and Srinivasa 2013), or desires and emotions from actions, facial expressions, speech, and language (Baker et al. 2017; Ong, Zaki, and Goodman 2015).

There has been a proliferation of attempts to learn these nonlinear dynamics by integrating deep learning with traditional probabilistic approaches, such as Hidden Markov Models (HMMs) and latent dynamical systems. Most approaches do this by adding random latent variables to each time step in an RNN (Chung et al. 2015; Bayer and Osendorfer 2014; Fabius and van Amersfoort 2014; Fraccaro et al. 2016). Other authors begin with latent sequence models as a basis, then develop deep parameterizations and inference techniques for these models (Krishnan, Shalit, and Sontag 2017; Archer et al. 2015; Johnson et al. 2016; Karl et al. 2016; Doerr et al. 2018; Lin, Khan, and Hubacher 2018; Chen et al. 2018). Most relevant to our work are the Deep Markov Models (DMMs) proposed by Krishnan, Shalit, and Sontag (2017), a generalization of HMMs and Gaussian State Space Models where the transition and emission distributions are parameterized by deep networks.

Unfortunately, none of the approaches thus far are designed to handle inference with both missing data and multiple modalities. Instead, most approaches rely upon RNN inference networks (Krishnan, Shalit, and Sontag 2017; Karl et al. 2016; Che et al. 2018), which can only handle missing data using ad-hoc approaches such as zero-masking (Lipton, Kale, and Wetzel 2016), update-skipping (Krishnan, Shalit, and Sontag 2017), temporal gating mechanisms (Neil, Pfeiffer, and Liu 2016; Che et al. 2018), or providing time stamps as inputs (Chen et al. 2018), none of which have intuitive probabilistic interpretations. Of the approaches that do not rely on RNNs, Fraccaro et al. handle missing data by assuming linear latent dynamics (Fraccaro et al. 2017), while Johnson et al. (Johnson et al. 2016) and Lin et al. (Lin, Khan, and Hubacher 2018) use hybrid message passing inference that is theoretically capable of marginalizing out missing data. However, these methods are unable to learn nonlinear transition dynamics, nor do they handle multiple modalities.

We address these limitations—the inability to handle missing data, over multiple modalities, and with nonlinear dynamics—by introducing a multimodal generalization of Deep Markov Models, as well as a factorized inference method for such models that handles missing time points and modalities by design. Our method allows for both fil-
Here, \( \Pi \) is the Gaussian distribution, and \( \Pi \) is some arbitrary emission distribution. The distribution parameters \( \mu_{\theta}, \Sigma_{\theta} \) and \( \kappa_{\gamma}^{m} \) are functions of either \( z_{t-1} \) or \( z_{t} \). We learn these functions as neural networks with weights \( \theta \). We also use \( z_{t_1:t_2} \) to denote the time-series of \( z \) from \( t_1 \) to \( t_2 \), and \( x_{t_1:t_2}^{m_1:m_2} \) to denote the corresponding observations from modalities \( m_1 \) to \( m_2 \). We omit the modality superscripts when all modalities are present (i.e., \( x_{t} = x_{t}^{1:M} \)).

We want to jointly learn the parameters \( \theta \) of the generative model \( p_{0}(z_{1:T}, x_{1:T}) = p_{0}(x_{1:T}|z_{1:T})p_{0}(z_{1:T}) \) and the parameters \( \phi \) of a variational posterior \( q_{\phi}(z_{1:T}|x_{1:T}) \) which approximates the true (intractable) posterior \( p_{0}(z_{1:T}|x_{1:T}) \). To do so, we maximize a lower bound on the log marginal likelihood \( L(x; \theta, \phi) \leq p_{0}(x_{1:T}) \), also known as the evidence lower bound (ELBO):

\[
L(x; \theta, \phi) = \mathbb{E}_{q_{\phi}(z_{1:T}|x_{1:T})}[\log p_{0}(x_{1:T}|z_{1:T})] - \mathbb{E}_{q_{\phi}(z_{1:T}|x_{1:T})}[\text{KL}(q_{\phi}(z_{1:T}|x_{1:T}) || p_{0}(z_{1:T}))]
\]

In practice, we can maximize the ELBO with respect to \( \theta \) and \( \phi \) via gradient ascent with stochastic backpropagation (Kingma and Welling 2014; Rezende, Mohamed, and Wierstra 2014). Doing so requires sampling from the variational posterior \( q_{\phi}(z_{1:T}|x_{1:T}) \). In the following sections, we derive a variational posterior that factorizes over time-steps and modalities, allowing us to tractably infer the latent states \( z_{1:T} \) even when data is missing.

**Factorized Posterior Distributions**

In latent sequence models such as MDMMs, we often want to perform several kinds of inferences over the latent states. The most common of such latent state inferences are:

**Filtering** Inferring \( z_{t} \) given past observations \( x_{1:t} \).

**Smoothing** Inferring some \( z_{t} \) given all observations \( x_{1:T} \).

**Sequencing** Inferring the sequence \( z_{1:T} \) from \( x_{1:T} \).

Most architectures that combine deep learning with state space models focus upon filtering (Fabius and van Amersfoort 2014; Chung et al. 2015; Hafner et al. 2018; Buesing et al. 2018), while Krishnan et al. optimize their DMM for sequencing (Krishnan, Shalit, and Sontag 2017). One of our contributions is to demonstrate that we can learn the filtering, smoothing, and sequencing distributions within the same framework, because they all share similar factorizations (see Figures 1b and 1c for the shared inference structure of filtering and smoothing). A further consequence of these factorizations is that we can naturally handle inferences given missing modalities or time-steps.

To demonstrate this similarity, we first factorize the sequencing distribution \( p(z_{1:T}|x_{1:T}) \) over time:

\[
p(z_{1:T}|x_{1:T}) = p(z_{1}|x_{1:T}) \prod_{t=2}^{T} p(z_{t}|z_{t-1}, x_{1:t})
\]
This factorization means that each latent state \( z_t \) depends only on the previous latent state \( z_{t-1} \), as well as all current and future observations \( x_{t:T} \), and is implied by the graphical structure of the MDM (Figure 1a). We term \( p(z_{t}|z_{t-1}, x_{t:T}) \) the conditional smoothing posterior, because it is the posterior that corresponds to the conditional prior \( p(z_{t}|z_{t-1}) \) on the latent space, and because it combines information from both past and future (hence ‘smoothing’).

Given one or more modalities, we can show that the conditional smoothing posterior \( p(z_{t}|z_{t-1}, x_{t:T}) \), the backward filtering distribution \( p(z_{t}|x_{t:T}) \), and the smoothing distribution \( p(z_{t}|x_{t:T}) \) all factorize almost identically:

**Backward Filtering**

\[
p(z_{t}|x_{t:T}) \propto p(z_{t}|x_{t+1:T}) \prod_{m=1}^{M} \frac{p(z_{t}|x_{t+1:M}^{m}|z_{t})}{p(z_{t})}
\]

**Forward Smoothing**

\[
p(z_{t}|x_{t:T}) \propto p(z_{t}|x_{t+1:T}) \prod_{m=1}^{M} \frac{p(z_{t}|x_{t+1:M}^{m}|z_{t})}{p(z_{t})} \frac{p(z_{t}|x_{t+1:T})}{p(z_{t})}
\]

**Conditional Smoothing Posterior**

\[
p(z_{t}|z_{t-1}, x_{t:T}) \propto p(z_{t}|x_{t+1:T}) \prod_{m=1}^{M} \frac{p(z_{t}|x_{t+1:M}^{m}|z_{t})}{p(z_{t})} \frac{p(z_{t}|x_{t+1:T})}{p(z_{t})}
\]

Equations 5–7 show that each distribution can be decomposed into (i) its dependence on future observations, \( p(z_{t}|x_{t+1:T}) \), (ii) its dependence on each modality \( m \) in the present, \( p(z_{t}|x_{t+1:M}^{m}|z_{t}) \), and, excluding filtering, (iii) its dependence on the past \( p(z_{t}|z_{t-1}) \) or \( p(z_{t}|x_{t-1:T}) \). Their shared structure is due to the conditional independence of \( x_{t:T} \) given \( z_{t} \) from all prior observations or latent states. Here we show only the derivation for Equation 7, because the others follow by either dropping \( z_{t-1} \) (Equation 5), or replacing \( z_{t-1} \) with \( x_{t-1} \) (Equation 6):

\[
p(z_{t}|z_{t-1}, x_{t:T}^{1:M}) = p(x_{t+1:T}^{1:M}|z_{t}) p(x_{t}^{1:M}|z_{t}) \frac{p(z_{t}|z_{t-1})}{p(x_{t+1:T}^{1:M}|z_{t-1})}
\]

\[
\propto p(x_{t+1:T}^{1:M}|z_{t}) \prod_{m=1}^{M} \frac{p(x_{t+1:M}^{m}|z_{t})}{p(x_{t+1:M}^{m}|z_{t-1})} \frac{p(z_{t}|z_{t-1})}{p(z_{t})}
\]

\[
= \frac{p(z_{t}|x_{t+1:T}^{1:M}) p(x_{t}^{1:M}|z_{t})}{p(z_{t})} \prod_{m=1}^{M} \frac{p(z_{t}|x_{t+1:M}^{m}|z_{t})}{p(z_{t})} \frac{p(z_{t}|z_{t-1})}{p(z_{t})}
\]

\[
\propto p(z_{t}|x_{t+1:T}^{1:M}) \prod_{m=1}^{M} \frac{p(z_{t}|x_{t+1:M}^{m}|z_{t})}{p(z_{t})} \frac{p(z_{t}|z_{t-1})}{p(z_{t})}
\]

The factorizations in Equations 5–7 lead to several useful insights. First, they show that any missing modalities \( \mathcal{M} \in \{1, M\} \) at time \( t \) can simply be left out of the product over modalities, leaving us with distributions that correctly condition on only the modalities \( \{1, M\} \backslash \mathcal{M} \) that are present. Second, they suggest that we can compute all three distributions if we can approximate the dependence on the future, \( q(z_{t}|x_{t+1:T}) \approx p(z_{t}|x_{t+1:T}) \), learn approximate posteriors \( q(z_{t}|x_{t+1:M}^{m}) \approx p(z_{t}|x_{t+1:M}^{m}) \) for each modality \( m \), and know the model dynamics \( p(z_{t}|z_{t-1}) \).

**Multimodal Fusion via Product of Gaussians**

However, there are a few obstacles to performing tractable computation of Equations 5–7. One obstacle is that it is not tractable to compute the product of generic probability distributions. To address this, we adopt the approach used for the MVAE (Wu and Goodman 2018), making the assumption that each term in Equations 5–7 is Gaussian. If each distribution is Gaussian, then their products or quotients are also Gaussian and can be computed in closed form. Since this result is well-known, we state it in the supplement (see Wu and Goodman (2018) for a proof).

This Product-of-Gaussians approach has the added benefit that the output distribution is dominated by the input Gaussian terms with lower variance (higher precision), thereby fusing information in a way that gives more weight to higher-certainty inputs (Cao and Fleet 2014; Ong, Zaki, and Goodman 2015). This automatically balances the information provided by each modality \( m \), depending on whether \( p(z_{t}|x_{t+1:M}^{m}) \) is high or low certainty, as well as the information provided from the past and future through \( p(z_{t}|z_{t-1}) \) and \( p(z_{t}|x_{t+1:T}) \), thereby performing multimodal temporal fusion in a manner that is uncertainty-aware.

**Approximate Filtering with Missing Data**

Another obstacle to computing Equations 5–7 is the dependence on future observations, \( p(z_{t}|x_{t+1:T}) \), which does not admit further factorization, and hence does not readily handle missing data among those future observations. Other approaches to approximating this dependence on the future rely on RNNs as recognition models (Krishnan, Shalit, and Sontag 2017; Che et al. 2018), but these are not designed to work with missing data.

To address this obstacle in a more principled manner, our insight was that \( p(z_{t}|x_{t+1:T}) \) is the expectation of \( p(z_{t}|z_{t+1}) \) under the backwards filtering distribution, \( p(z_{t+1}|x_{t+1:T}) \):

\[
p(z_{t}|x_{t+1:T}) = \mathbb{E}_{p(z_{t+1}|x_{t+1:T})} [p(z_{t}|z_{t+1})]
\]

For tractable approximation of this expectation, we use an approach similar to assumed density filtering (Huber, Beutler, and Hanebeck 2011). We assume both \( p(z_{t}|x_{t+1:T}) \) and \( p(z_{t}|z_{t+1}) \) to be multivariate Gaussian with diagonal covariance, and sample the parameters \( \mu, \Sigma \) of \( p(z_{t}|z_{t+1}) \) under \( p(z_{t+1}|x_{t+1:T}) \). After drawing \( K \) samples, we approximate the parameters of \( p(z_{t}|x_{t+1:T}) \) via empirical moment-matching:

\[
\hat{\mu}_{z_{t}|x_{t+1:T}} = \frac{1}{K} \sum_{k=1}^{K} \mu_{k}
\]

\[
\hat{\Sigma}_{z_{t}|x_{t+1:T}} = \frac{1}{K} \sum_{k=1}^{K} (\Sigma_{k} + \mu_{k} \mu_{k}^{T}) - \hat{\mu}_{z_{t}|x_{t+1:T}}^{2}
\]

Approximating \( p(z_{t}|x_{t+1:T}) \) by \( p(z_{t}|z_{t+1}) \) led us to three important insights. First, by substituting the expectation from Equation 8 into Equation 5, the backward filtering distribution becomes:

\[
p(z_{t}|x_{t:T}) \propto \mathbb{E}_{p(z_{t}|z_{t+1})} [\prod_{m=1}^{M} \frac{p(z_{t}|x_{t+1:M}^{m}|z_{t})}{p(z_{t})}]
\]

In other words, by sampling under the filtering distribution for time \( t+1 \), \( p(z_{t+1}|x_{t+1:T}) \), we can compute the filtering distribution for time \( t \), \( p(z_{t}|x_{t:T}) \). We can thus recursively compute \( p(z_{t}|x_{t:T}) \) backwards in time, starting from \( t = T \).

Second, once we can perform filtering backwards in time, we can use this to approximate \( p(z_{t}|x_{t+1:T}) \) in the smoothing distribution (Equation 6) and the conditional smoothing distribution (Equation 7).
Algorithm 1 A variational backward algorithm for approximate \textit{backward} filtering.

\begin{algorithm}
\caption{BackwardFiltering($x_{1:T}$, $K$)}
\begin{algorithmic}
\Function {BackwardFilter}{$x_{1:T}$, $K$}
\State Initialize $q(z_t|x_{1:T}, K) \propto p(z_t)$
\For{$t = T$ to 1}
\State Let $\mathcal{M} \subseteq [1, M]$ be the observed modalities at $t$
\State $q(z_{t}|x_{1:T}) \leftarrow q(z_{t}|x_{t+1:T}) \prod_{k \in \mathcal{M}} q(z_{t}^{m_{k}})
\State Sample $K$ particles $z_{t}^{k} \sim q(z_{t}|x_{1:T})$ for $k \in [1, K]$
\State Compute $p(z_{t-1}|z_{t}^{k})$ for each particle $z_{t}^{k}$
\State $q(z_{t-1}|x_{t}) \leftarrow \frac{1}{K} \sum_{k=1}^{K} p(z_{t-1}|z_{t}^{k})$
\EndFor
\Return \{$q(z_{t}|x_{1:T}), q(z_{t}|x_{t+1:T})$ for $t \in [1, T]$\}
\EndFunction
\end{algorithmic}
\end{algorithm}

Algorithm 2 A variational backward-forward algorithm for approximate \textit{forward} smoothing.

\begin{algorithm}
\caption{ForwardSmooth($x_{1:T}$, $K_f$, $K_b$)}
\begin{algorithmic}
\Function {ForwardSmooth}{$x_{1:T}$, $K_f$, $K_b$}
\State Initialize $p(z_{1}|x_{1}) \leftarrow 1$
\State Collect $q(z_{T}|x_{T})$ from BackwardFilter($x_{1:T}$, $K_b$)
\For{$t = 1$ to $T$}
\State Let $\mathcal{M} \subseteq [1, M]$ be the observed modalities at $t$
\State $q(z_{t}|x_{1:T}) \leftarrow q(z_{t}|x_{t+1:T}) \prod_{k \in \mathcal{M}} q(z_{t}^{m_{k}}) q(z_{t}|x_{1:T-1}) p(x_{t}|z_{t}^{k})$
\State Sample $K_f$ particles $z_{t} \sim q(z_{t}|x_{1:T})$ for $k \in [1, K_f]$
\State Compute $p(z_{t+1}|z_{t}^{k})$ for each particle $z_{t}^{k}$
\State $q(z_{t+1}|x_{t+1}) \leftarrow \frac{1}{K_f} \sum_{k=1}^{K_f} p(z_{t+1}|z_{t}^{k})$
\EndFor
\Return \{$q(z_{t}|x_{1:T}), q(z_{t}|x_{t-1:T})$ for $t \in [1, T]$\}
\EndFunction
\end{algorithmic}
\end{algorithm}

While Algorithm 1 approximates the filtering distribution $q(z_{t}|x_{t:T})$, by setting the number of particles $K_f = 1$, it effectively computes the (backward) conditional filtering posterior $q(z_{t}|z_{t+1}, x_{t})$ and (backward) conditional prior $p(z_{t}|z_{t+1})$ for a randomly sampled latent sequence $z_{1:T}$. Similarly, while Algorithm 2 approximates smoothing by default, when $K_f = 1$, it effectively computes the (forward) conditional smoothing posterior $q(z_{t-1}|z_{t}, x_{1:T})$ and (forward) conditional prior $p(z_{t}|z_{t-1})$ for a random latent sequence $z_{1:T}$. These quantities are useful not only because they allow us to perform sequencing, but also because we can use them to compute the ELBO for both backward filtering and forward smoothing:

\begin{align}
L_{\text{filter}} &= \sum_{t=1}^{T} \left[ \mathbb{E}_{q(z_{t}|x_{1:T})} \log p(x_{t}|z_{t}) - \mathbb{E}_{q(z_{t}|x_{1:T})} \text{KL}(q(z_{t}|z_{t+1}, x_{t})||p(z_{t}|z_{t+1})) \right] \\
L_{\text{smooth}} &= \sum_{t=1}^{T} \left[ \mathbb{E}_{q(z_{t}|x_{1:T})} \log p(x_{t}|z_{t}) - \mathbb{E}_{q(z_{t-1}|x_{1:T})} \text{KL}(q(z_{t}|z_{t-1}, x_{1:T})||p(z_{t}|z_{t-1})) \right]
\end{align}

$L_{\text{filter}}$ is the filtering ELBO because it corresponds to a ‘backward filtering’ variational posterior $q(z_{t:T}|x_{1:T}) = \prod_{t=1}^{T} q(z_{t}|z_{t+1}, x_{t})$, where each $z_{t}$ is only inferred using the current observation $x_{t}$ and the future latent state $z_{t+1}$. $L_{\text{smooth}}$ is the smoothing ELBO because it corresponds to the correct factorization of the posterior in Equation 4, where each term combines information from both past and future. Since $L_{\text{smooth}}$ corresponds to the correct factorization, it should theoretically be enough to minimize $L_{\text{smooth}}$ to learn...
good MDMM parameters \( \theta, \phi \). However, in order to compute \( L_{\text{smooth}} \), we must perform a backward pass which requires sampling under the backward filtering distribution. Hence, to accurately approximate \( L_{\text{smooth}} \), the backward filtering distribution has to be reasonably accurate as well. This motivates learning the parameters \( \theta, \phi \) by jointly maximizing the filtering and smoothing ELBOs as a weighted sum. We call this paradigm backward-forward variational inference (BFVI), due to its use of variational posteriors for both backward filtering and forward smoothing.

### Experiments

We compare BFVI against state-of-the-art RNN-based inference methods on two multimodal time series datasets over a range of inference tasks. **F-Mask** and **F-Skip** use forward RNNs (one per modality), using zero-padding and update skipping respectively to handle missing data. They are thus multimodal variants of the ST-L network in (Krishnan, Shalit, and Sontag 2017), and similar to the variational RNN (Chung et al. 2015) and recurrent SSM (Hafner et al. 2018). **B-Mask** and **B-Skip** use backward RNNs, with masking and skipping respectively, and correspond to the Deep Kalman Smoother in (Krishnan, Shalit, and Sontag 2017). The underlying MDMM architecture is constant across inference methods. Architectural and training details can be found in the supplement. Code is available at https://git.io/Jeoze.

### Datasets

**Noisy Spirals.** We synthesized a dataset of 1000 noisy 2D spirals (600 train / 400 test) similar to Chen et al. (Chen et al. 2018), treating the \( x \) and \( y \) coordinates as two separate modalities. Spiral trajectories vary in direction (clockwise or counter-clockwise), size, and aspect ratio, and Gaussian noise is added to the observations. We used 5 latent dimensions, and two-layer perceptrons for encoding \( q(z_t | x_t^m) \) and decoding \( p(x_t^m | z_t) \). For evaluation, we compute the mean squared error (MSE) per time step between the predicted trajectories and ground truth spirals.

**Weizmann Human Actions.** This is a video dataset of 9 people each performing 10 actions (Gorelick et al. 2007). We converted it to a trimodal time series dataset by treating silhouette masks as an additional modality, and treating actions as per-frame labels, similar to He et al. (2018). Each RGB frame was cropped to the central \( 128 \times 128 \) window and resized to \( 64 \times 64 \). We selected one person's videos as the test set, and the other 80 videos as the training set, allowing us to test action label prediction on an unseen person. We used 256 latent dimensions, and convolutional / deconvolutional neural networks for encoding and decoding. For evaluation, we compute the Structural Similarity (SSIM) between the input video frames and the reconstructed outputs.

### Inference Tasks

We evaluated all methods on the following suite of temporal inference tasks for both datasets: **reconstruction:** reconstruction given complete observations; **drop half:** reconstruction after half of the inputs are randomly deleted; **forward extrapolation:** predicting the last 25% of a sequence when the rest is given; and **backward extrapolation:** inferring the first 25% of a sequence when the rest is given. When evaluating these tasks on the Weizmann dataset, we provided only video frames as input (i.e. with neither the silhouette masks nor the action labels), to test whether the methods were capable of unimodal inference after multimodal training.

We also tested cross-modal inference using the following conditional generation / label prediction tasks: **conditional generation (Spirals):** given \( x \)- and initial 25% of \( y \)-coordinates, generate rest of spiral; **conditional generation (Weizmann):** given the video frames, generate the silhouette masks; and **label prediction (Weizmann):** infer action labels given only video frames.

<table>
<thead>
<tr>
<th>Method</th>
<th>Recon. (MSE, SD)</th>
<th>Drop Half (MSE, SD)</th>
<th>Fwd. Extra. (MSE, SD)</th>
<th>Bwd. Extra. (MSE, SD)</th>
<th>Cond. Gen. (MSE, SD)</th>
<th>Label Pred. (MSE, SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BFVI (ours)</td>
<td><strong>0.02 (0.01)</strong></td>
<td><strong>0.04 (0.01)</strong></td>
<td>0.12 (0.10)</td>
<td>0.07 (0.03)</td>
<td><strong>0.26 (0.26)</strong></td>
<td>–</td>
</tr>
<tr>
<td>F-Mask</td>
<td><strong>0.02 (0.01)</strong></td>
<td>0.06 (0.02)</td>
<td><strong>0.10 (0.08)</strong></td>
<td>0.18 (0.07)</td>
<td>1.37 (1.39)</td>
<td>–</td>
</tr>
<tr>
<td>F-Skip</td>
<td>0.04 (0.01)</td>
<td>0.10 (0.05)</td>
<td>0.13 (0.11)</td>
<td>0.19 (0.06)</td>
<td>1.51 (1.54)</td>
<td>–</td>
</tr>
<tr>
<td>B-Mask</td>
<td><strong>0.02 (0.01)</strong></td>
<td><strong>0.04 (0.01)</strong></td>
<td>0.18 (0.14)</td>
<td><strong>0.04 (0.01)</strong></td>
<td>1.25 (1.23)</td>
<td>–</td>
</tr>
<tr>
<td>B-Skip</td>
<td>0.05 (0.01)</td>
<td>0.19 (0.05)</td>
<td>0.32 (0.22)</td>
<td>0.37 (0.15)</td>
<td>1.64 (1.51)</td>
<td>–</td>
</tr>
</tbody>
</table>

**Weizmann Video Dataset: SSIM or Accuracy* (SD)**

<table>
<thead>
<tr>
<th>Method</th>
<th>SSIM (MSE, SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BFVI (ours)</td>
<td><strong>.85 (.03)</strong></td>
</tr>
<tr>
<td>F-Mask</td>
<td>.68 (.18)</td>
</tr>
<tr>
<td>F-Skip</td>
<td>.70 (.12)</td>
</tr>
<tr>
<td>B-Mask</td>
<td>.79 (.04)</td>
</tr>
<tr>
<td>B-Skip</td>
<td>.80 (.04)</td>
</tr>
</tbody>
</table>

*For evaluation, we compute the mean structural similarity (SSIM) or label accuracy (higher is better) per time-step with respect to original videos. Means and Standard Deviations (SD) are across the test set.
In this paper, we introduced backward-forward variational inference (BFVI) as a novel inference method for Multimodal Deep Markov Models. This method handles incomplete data via a factorized variational posterior, allowing us to easily marginalize over missing observations. Our method is thus capable of a large range of multimodal temporal inference tasks, which we demonstrate on both a synthetic dataset and a video dataset of human motions. The ability to handle missing data also enables applications in weakly supervised learning, and this was the case even on the training set (not shown in Table 1). We suspect that this is because the RNN-based methods lack a principled approach to multimodal fusion, and hence fail to learn a latent space which captures the mutual information between action labels and images. In contrast, BFVI learns to both predict one modality from another, and to propagate information across time, as can be seen from the reconstruction and predictions in Figure 3.

Weakly Supervised Learning

In addition to performing inference with missing data test time, we compared the various methods ability to learn with missing data at training time, amounting to a form of weakly supervised learning. We tested two forms of weakly supervised learning on the Spirals dataset, corresponding to different conditions of data incompleteness. The first was learning with data missing uniformly at random. This condition can arise when sensors are noisy or asynchronous. The second was learning with missing modalities, or semi-supervised learning, where a fraction of the sequences in the dataset only has a single modality present. This condition can arise when a sensor breaks down, or when the dataset is partially unlabelled by annotators. We also tested learning with missing modalities on the Weizmann dataset.

Results for these experiments are shown in Figure 4, which compare BFVI’s performance on increasing levels of missing data against the next best method, averaged across 10 trials. Our method (BFVI) performs well, maintaining good performance on the Spirals dataset even with 70% uniform random deletion (Figure 4a) and 60% uni-modal examples (Figure 4b), while degrading gracefully with increasing missingness. This is in contrast to B-Mask, which is barely able to learn when even 10% of the spiral examples are uni-modal, and performs worse than BFVI on the Weizmann dataset at all levels of missing data (Figure 4c).

Conclusion

In this work, we introduced backward-forward variational inference (BFVI) as a novel inference method for Multimodal Deep Markov Models. This method handles incomplete data via a factorized variational posterior, allowing us to easily marginalize over missing observations. Our method is thus capable of a large range of multimodal temporal inference tasks, which we demonstrate on both a synthetic dataset and a video dataset of human motions. The ability to handle missing data also enables applications in weakly supervised learning of labelled time series. Given the abundance of multimodal time series data where missing data is the norm rather than the exception, our work holds great promise for many future applications.
Figure 4: Learning curves under various forms of weak supervision. (a) Learning with randomly missing data on the Spirals dataset. (b) Semi-supervised learning on the Spirals dataset, where some sequences are entirely missing the y-coordinate. (c) Semi-supervised learning on the Weizmann dataset, where some sequences have no action labels.

Acknowledgements

This work was supported by the A*STAR Human-Centric Artificial Intelligence Programme (SERC SSF Project No. A1718g0048).

References


