Generative Multimodal Learning for Reconstructing Missing Modality

Ankur Agarwal [*] [1] **Nishant Mishra** [*] [2]

Group-7

Abstract

Multimodal learning with latent space models has the potential to help learn deeper, more useful representations that help getting better performance, even in missing modality scenarios. In this project we leverage latent space based model to perform inference models and reconstruction in all missing modality combinations. We trained a Multimodal Variational Encoder which uses a product of Experts based inference network on three different modalities consisting of MNIST handwritten digit images in two languages and spoken digit recordings for our experiments. We trained the model in a subsampled training paradigm using an ELBO loss that comprised the modality reconstruction losses, label cross-entropy loss as well as the Kullback-Leibler divergence for the latent distribution. We evaluated the total ELBO loss, individual reconstruction losses, classification accuracy and visual reconstruction outputs as part of our analysis. We observed encouraging results both in terms of successful convergence as well as accurate reconstructions.

1. Introduction

Multimodal Learning involves learning complex, representations using multiple independent modalities of a data to be able to solve a problem better. Multimodal Learning leads to more generalizable representations for the kind of data or task involving multiple forms. It has a number of applications such as Image Captioning, Visual Question Answering etc.

The problem we set out to address in this project was to train a latent variable based variational inference model on multimodal data, and using it to train similar models with a subset of the possible modalities in order to perform inference with all possible combinations of missing modalities provided as well as get a reconstruction of all modalities.

Such models can find application in a number of tasks. For example such a model trained on Image, speech and corresponding text transcripts as modalities when trained with sufficient data can be used with an efficient video calling application. Being trained to handle multiple combination of modalities and perform in missing modality scenarios, it will help reduce the bandwidth requirement of the video call app, by eliminating the need to transmit the video stream along with the associated audio and text modalities. We could only send the latter two and the video/face can be generated/reconstructed on the go using the model, hence saving a lot in terms of data at both the sender and receiver ends. Other applications could be in solving popular problems such as Image Captioning, Visual Question Answering, emotion recognition, inter-conversions among different modalities, time series predictions when future data is missing etc.

Latent Variable based models are a very popular branch of statistical modeling which rely on the mapping of input observations to hidden or latent representations that can be optimized for various tasks. Latent Variable models are also popular for unsupervised learning based tasks such as clustering, Principle Component Analysis.

Variational Autoencoders (Doersch, 2016) are generative models that allow us to model a data from a latent distribution obtained from the observations. A VAE is trained by optimizing a lower bound, an optimization technique called Variational Inference.

For our project, we needed a model with latent representation to be able to store information from all modalities, as well as generative properties for reconstruction of the different modalities using the latent representation.

Hence we approached the missing modality reconstruction and classification based problem using a Multimodal Variational Autoencoder(MVAE) inspired from the paper (Wu & Goodman, 2018) Our model used a tree like graph where the different modalities define the observation nodes. It consists of parallel fully connected encoder and decoder networks associated with each modality as part of a VAE and a prod-

[*] Equal contribution [1] Universite de Montreal, MILA [2] McGill University. Correspondence to: <Group 7>.

uct of experts technique for late fusion of the respective latent distribution parameters from each encoder to get a final representation. An additional linear decoder branch was used for label classification.Each modality has its own inference network. This model was trained by optimizing an estimated lower bound (ELBO) on the marginal likelihood of observed data, i.e reconstructions of the modalities as well as the classification loss. We also used a sampling based training scheme such that for each training example containing modalities, we obtained the loss for all combinations of modalities given to the model, this ensured the learned model generalized to perform well in reconstructing given any combination set of the modalities.

We used three modalities for experimentation and trained the model on a MNIST dataset with images in two languages, Farsi and Kannada as first two modalities and speech utterances of the MNIST digits as the third modality. The model performed well in terms of the convergence of ELBO loss, individual reconstruction losses, classification accuracy as well as the final visual reconstructions of the modalities. We also performed various analyses in terms of hyperparameter tuning, reconstruction under different modality combinations as well as analysis of disentanglement of representation property.

All the steps, experiments and results have been discussed in detail in the following sections of the report. Section 2 explores a bit of a background and highlights related work in the domain, followed by a detailed description of the Dataset we used in Section 3. Section 4 details the methodology including the model architecture, the training scheme as well as the various hyperparameters and design choices. The results of our experiments are tabulated and demonstrated in Section 5 along with detailed discussion of these results followed by final conclusions and future work in Sections 6 and 7 respectively.

2. Related Work

Application of Generative Models for Multimodal Learning is a popular area of focused research. Many different variants of MVAEs have been used to train generative models of the form $p(x_2|x_1)$, where x_1 and x_2 are different modalities, such as Conditional Variational Autoencoders (CVAEs) (Sohn et al., 2015) and conditional multi-modal autoencoders (CMMA) (Pandey & Dukkipati, 2017). One of the seminal works that forms a foundation for our project was the concept of Multimodal Variational Autoencoders (MVAEs) introduced in the paper Multimodal Generative Models for Scalable Weakly-Supervised Learning (Wu & Goodman, 2018). MVAE uses a product of experts based inference network and a sub-sampled training paradigm which enables it to generalize. It has shared parameters across modalities in order to learn any combination of missing modalities.It was also shown to be highly effective in weakly supervised learning.

Another recent work titled Factorized Inference in Deep Markov Models for Incomplete Multimodal Time Series (Zhi-Xuan et al., 2019) introduces a factorized inference method for Multimodal Deep Markov Models (MDMMs). It is a multimodal latent space variation of Hidden Markov Models that is trained using a variational forward-backward algorithm. It is capable of handling incompleteness in terms of both temporal information as well as modalities.

Joint Multi-modal VAE (JMVAE) used in (Suzuki et al., 2016) explicitly modeled the joint distribution $p(x_1, x_2)$. The JMVAE collectively trains the joint inference network $q(z|x_1, x_2)$ with two other inference networks $q(z|x_1)$ and $q(z|x_2)$, to handle missing modality. It is inferior to MVAE as it basically has an inference network for each subset of modalities, which is computationally intractable for large number of modalities.

Product of Experts used for the fusion of the distributions obtained from all the individual inference networks to get the final latent distribution is similar to a Restricted Boltzmann machine (RBM), another latent variable model that has been applied to multi-modal learning(Srivastava & Salakhutdinov, 2014).

Apart from VAEs, Generative Adversarial Networks (GANs) are also being adapted for multimodal representation learning or clustering(Pandeva & Schubert, 2019).

From application perspective MVAEs have been succesfully applied to different tasks. (Khattar et al., 2019) utilised a bimodal VAE for the task of fake news detection.

3. Data

For our experiments, we decided to use three modalities and chose to use MNIST as our dataset. This was decided based on both the computational complexity, availability as well as the statistical complexity. MNIST acts as a suitable dataset to test hypotheses on through experimentation. Our three modalities were binary handwritten digit images in Farsi, Binary Handwritten digit image in Kannada and Digit speech recordings of the digits. As part of our training pipeline, we made a random sampler that sampled triplets corresponding to a giving label/digit from each of the three modalities that acted as our training or test sample. Except for the spoken data, the other datasets are sampled to form a triplet without replacement because of class imbalance. We obtained a training set of 60,000 unique triplets and training and validation sets of 10,000 unique triplets each.

3.1. MNIST Images- Multi-Language (m_1, m_2)

We decided to use binary images of MNIST digits $0-9$ in different languages as our first two modalities. MNIST-MIX (Jiang, 2020), an opensource multi-language handwritten digit recognition dataset was used for this purpose. It is a dataset of handwritten digits in 10 different languages curated in the same format as the original MNIST dataset. The 10 different languages have different number of training and testing samples. We chose to go ahead with Farsi and Kannada handwritten digits as two of our three modalities for class balance and appropriate data sampling. The details of the dataset are given in Table 1 and samples shown in Figure 1. These images were normalized and then flattened before being passed on for training.

Dataset Name	Language	Training Size	Testing Size
FARSI	Farsi	60,000	20,000
Kannada-mnist	Kannada	60,000	20,240

Table 1. Dataset list with training and testing sizes

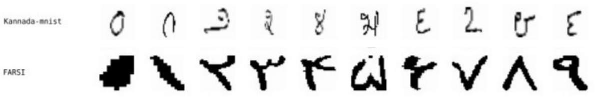

Figure 1. Samples from handwritten digit dataset representing digits 0-9

3.2. MNIST- Speech recording (m_3)

As part of our third modality we used the Free Spoken digit dataset. It is another open source dataset that contains recordings of spoken digits in .wav format at 8kHz trimmed to ensure minimal silence at the beginning and ending. It contains a total of 3000 recordings, 50 utterances for each individual digit by 6 different speakers, hence 300 recordings per digit. In order to use it for training our MVAE, we preprocessed the speech data by converting these recordings to 13 dimensional Mel Frequency Cepstral Coefficients (MFCC) feature vectors for better processing.

4. Methodology

4.1. Model Architecture

Taking inspiration from the model described in (Wu & Goodman, 2018) we followed the model design of individual encoders for each inputs as separate experts. The structure of the model follows a tree-structured graph where the different modalities define the observation nodes. The encoding from the experts go through late fusion to combine their individual information extracts. The decoders for label classification and reconstruction use this bottle-neck as a common point.

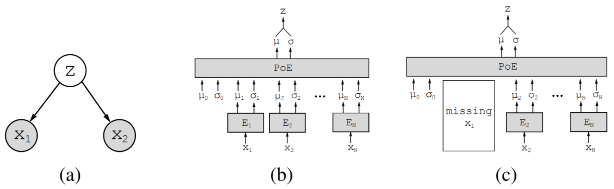

Figure 2. (a) Graphical model of the MVAE. (b) MVAE architecture with N modalities. μ_i and σ_i represent the i-th variational parameters; μ_0 and σ_0 represent the prior parameters. The product-of-experts (PoE) combines all variational parameters (c) If a modality is missing during training, we drop the respective inference network

4.1.1. Late Fusion - Encoding Inputs with Experts

Each modality passed as input is processed by separate encoder models. The model structure of late fusion allows many modalities to be added on later. Currently we have three modalities consisting of two image modalities and one for speech. Each encoder is supposed to return a latent variable parameters of mean and log variance. These latent variables are the compressed representation of each modality.

4.1.2. Fusion Technique - Product of Experts (PoE)

The fusing technique of taking the product of posteriors as the joint posterior is possible due to the fact that the modalities are independent of each other given the instance of the digit they represent. This is also the supporting cause for the tree structure of the model where each input modality is the observation node.

$$p(z|x_1, x_2, ..., x_N) = \frac{p(x_1, ..., x_N|z)p(z)}{p(x_1, ..., x_N)}$$

Since the modalities $(x_1, ..., x_N)$ are independent of each other given the common representation of the digit as z,

$$\begin{aligned}\implies p(z|x_1, x_2, ..., x_N) &= \frac{p(z)}{p(x_1, ..., x_N)} \prod_{i=1}^{N} p(x_i|z)\\ &= \frac{p(z)}{p(x_1, ..., x_N)} \prod_{i=1}^{N} \frac{p(z|x_i)p(x_i)}{p(z)}\\ &= \frac{\prod_{i=1}^{N} p(z|x_i)}{\prod_{i=1}^{N-1} p(z)} . \frac{\prod_{i=1}^{N} p(x_i)}{p(x_1, ..., x_N)}\end{aligned}$$

Ignoring the quotient term $\frac{\prod_{i=1}^{N} p(x_i)}{p(x_1,...,x_N)}$ and approximating $p(z|x_i)$ with $q(z|x_i) \equiv \tilde{q}(z|x_i)p(z)$ where our inference network model is $\tilde{q}(z|x_i)$.

$$\begin{aligned}\implies p(z|x_1, ..., x_N) &= \frac{\prod_{i=1}^{N} \tilde{q}(z|x_i)p(z)}{\prod_{i=1}^{N-1} p(z)}\\ &= p(z) \prod_{i=1}^{N} \tilde{q}(z|x_i)\end{aligned}$$

The joint posterior $p(z|x_1, ..., x_N)$ is calculated by taking the product of individual experts. We assume that our prior expert $p(z)$ and our experts $\tilde{q}(z|x_i)$ are Gaussian. We use the analytical solution and sampling with reparameterization technique to get the latent 'z'.

As the number of modalities increase the information in this latent representation densifies. Removing individual modalities does not completely destroy the stability of the latent. This advantage of the PoE favours our use case of reconstruction with missing modalities later during inference.

4.2. Training Scheme

4.2.1. Loss objective and metrics of evaluation

With our autoencoders and PoE in place, we had a good enough setup for a structured model to process the inputs well and decode to classify the label outcome with appropriate reconstructions. To train this model with assurance of retaining the important reconstruction features at the latent state, we use a combination of ELBO losses.For training we used a subsampling strategy where We define a powerset of possible combinations of the input modalities. In general we should have $2^N - 1$ combinations for N modalities. In our case, with three modalities we had 7 such combinations.For each combination we calculate the ELBO loss that consists of reconstruction for all the modalities, label cross entropy loss and the divergence between true and model latent distribution, this aspect of training was different from the MVAE paper (Wu & Goodman, 2018). This method improves generalization of information at the latent and the decoders for reconstruction of all given any set of combined input. To this we add the classification objective loss in each combination.

$$Loss = \sum_{X \in P(x_1,...,x_N)} ELBO(X) + CE(label\ onehot)$$

$$\begin{aligned} where, ELBO(X) = E_{q_\phi(z|X)}[\sum_{x_i \in X} \log p_\theta(x_i|z)]\\ -KL[q_\phi(z|X)||p(z)]\end{aligned}$$

For evaluating the reconstruction of the MNIST images of binary values, the binary cross entropy was used. To evaluate the speech 13 dimensional MFCC feature dense vector we used the mean squared error. Since the label classes of the digits were onehots, we used cross entropy loss for it. Coefficients were provided to the reconstruction losses to balance out the difference in the output ranges. The ELBO losses for all 7 combinations of input modalities for a given training sample (triplet) were added to get the final training loss which was optimized.

$$\operatorname{ELBO}(x_1, \ldots, x_N) + \sum_{i=1}^{N} \operatorname{ELBO}(x_i) + \sum_{j=1}^{k} \operatorname{ELBO}(X_j)$$

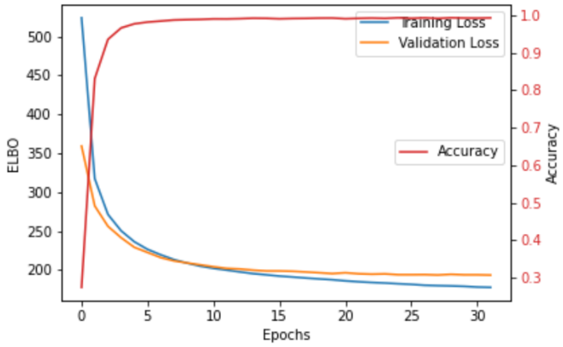

Figure 3. Loss and Label Accuracy convergence

4.2.2. Hyperparameters and design choices

We trained the model with a latent variable of 512 dimensions with a batch size of 128 using the Adam optimizer with a learning rate of 10^{-3} for 500 epochs. The reconstruction loss of speech was given a coefficient of 100 while unity for the image reconstructions. We used the swish activation function (Ramachandran et al., 2017) with fully connected linear layers as an alternative to ReLU with batchnormalization. For the Linear layers for log variance, we initialized the weights with zeros as a trick which solved the gradient explosion problem while training.

5. Results and Discussion

The accuracy of the different combinations are all close to cent percent. In general the classification task is easily solvable in the 3 datasets. As we add modalities the results get better as expected. The advantage we discussed of PoE for the use case of missing modality during inference not effecting the model much can be seen in the results table 2 and reconstructions in figure 4.

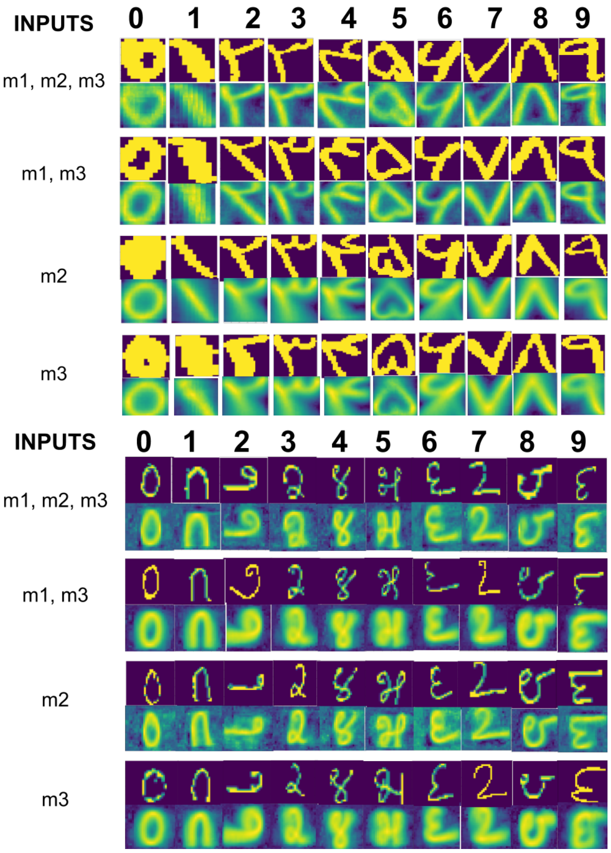

Figure 4. Original Image and Image reconstruction outputs in various combination of modalities. (Top: Farsi MNIST reconstruction), (Bottom :Kannada MNIST reconstruction)

The reconstructions of the two MNIST are pretty convincing. In the cases when the respective modality which is considered for reconstruction are missing in the input, the cross entropy loss goes up but the visual inspection shows that this is caused by blurry image reconstruction. The recreated images evidently show the respective language MNIST. From table 2 we can notice that the BCE for farsi reconstruction without it being input goes as high as 348 while for kannada it goes upto 135. The BCE for reconstruction when their inputs being present are as low as 89 and 68 which is comparable. The reason for this 3 times shooting in farsi compared to kannada when the respective modalities are missing is because the scribes of farsi are thicker then kannada. Thick scribes being blurred cause a higher loss.

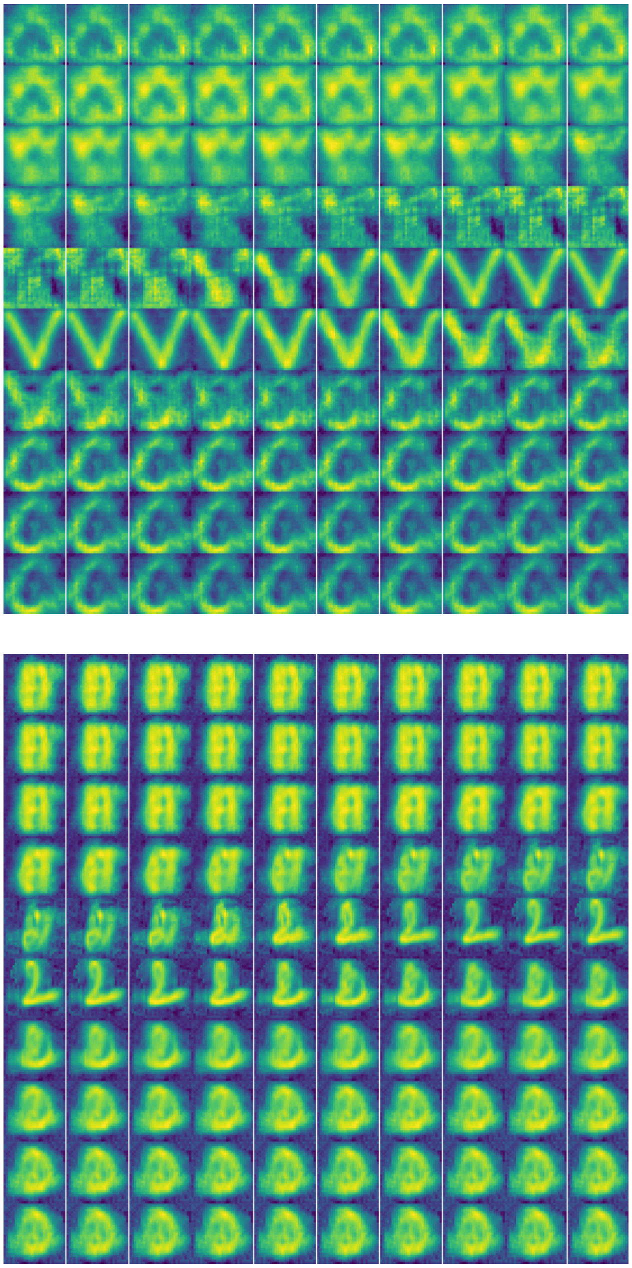

Figure 5. Output Reconstructions obtained by perturbing index 201 of latent variable by amounts -1500 to 1500 in intervals of 30 shown in 10x10 grid. Output reconstruction transitions as [5,3,7,0], original input label 7. (Top: Farsi), (Bottom :Kannada)

We also studied the disentanglement property of the latent space representation. We perturbed the latent space at particular indices with varying noises, we observed a consistent pattern of variations in reconstruction output of 2 different reconstructions. We noticed that a very big noise value was required to be added to bring changes in the reconstruction. This shows that the representations are disentangled. In

Combination	Classification (Accuracy)(%)	ELBO	Reconstruction M1 (BCE)	Reconstruction M2 (BCE)	Reconstruction M3 (MSE)
m1	99.6	248.85	89.09	135	0.133
m2	99.7	436.9	348.67	68.96	0.134
m3	99.2	493.03	348.81	135.76	0.0095
m1, m2	99.93	187.64	89.05	69.04	0.133
m2, m3	99.94	427.44	346.19	69.1	0.011
m1, m3	99.88	239.11	89.33	134.8	0.013
m1, m2, m3	99.95	177.62	89.38	69.1	0.014

Table 2. Training performance at different combinations of the modalities and joint inference experts.
BCE: Binary Cross Entropy; MSE: Mean Squared Error; ELBO: Evidence Lower Bound; m1: MNIST Language 1 (Farsi); m2: MNIST Language 2 (Kannada); m3: Spoken MNIST (MFCC features).

Figure 5 we used a perturbation noise in the range -1500 and 1500 to get 100 variation reconstructions of the farsi and kannada MNIST for and input z corresponding to label 7. It was interesting to see the digits vary simultaneously in both the reconstructions. This showed that the latent z representation was disentangled and we found a variable that could tune between the digits. We observe that the model learns disentangled representations some of which are modality agnostic.

6. Conclusion

We conclude that the late fusion model using PoE as the fusion technique with the training scheme we were able to converge the loss objective to solve the classification task and reconstruction using multiple modality and even for missing modality during inference time. The PoE and training scheme do work as expected as for a generalized setup of different combination of input modalities. The reconstruction is sharp and clear in presence of all the modalities and gets blurry but still evident to show the MNIST digit as we remove modalities. Not only does the latent representation of the MVAE capture important robust features, it also learns to disentangle them some of which are even modality agnostic.

7. Future Work

The model architecture could be improved with the usage of CNNs as a replacement to Linear layers. Another model improvement is to try in future is GANs instead of VAE in the same multimodal setting.

The application of reconstructing missing modality in conversations would be interesting given the data modalities of video, audio and text for primary tasks like emotion and sentiment analysis. The current primary task of label classification is easily solved for the given modalities and so analysing the differences due to addition of modality could not be done. With complex modalities and complex tasks, there would be added scope to do some studies there.

References

Doersch, C. Tutorial on variational autoencoders, 2016.

Jiang, W. Mnist-mix: A multi-language handwritten digit recognition dataset, 2020.

Khattar, D., Goud, J. S., Gupta, M., and Varma, V. Mvae: Multimodal variational autoencoder for fake news detection. In *The World Wide Web Conference*, pp. 2915–2921, 2019.

Pandeva, T. and Schubert, M. MMGAN: generative adversarial networks for multi-modal distributions. *CoRR*, abs/1911.06663, 2019. URL http://arxiv.org/abs/1911.06663.

Pandey, G. and Dukkipati, A. Variational methods for conditional multimodal deep learning. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pp. 308–315. IEEE, 2017.

Ramachandran, P., Zoph, B., and Le, Q. V. Swish: a self-gated activation function. *arXiv preprint arXiv:1710.05941*, 7, 2017.

Sohn, K., Lee, H., and Yan, X. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28: 3483–3491, 2015.

Srivastava, N. and Salakhutdinov, R. Multimodal learning with deep boltzmann machines. *The Journal of Machine Learning Research*, 15(1):2949–2980, 2014.

Suzuki, M., Nakayama, K., and Matsuo, Y. Joint multimodal learning with deep generative models. *arXiv preprint arXiv:1611.01891*, 2016.

Wu, M. and Goodman, N. Multimodal generative models for scalable weakly-supervised learning. In *Advances in Neural Information Processing Systems*, pp. 5575–5585, 2018.

Zhi-Xuan, T., Soh, H., and Ong, D. C. Factorized inference in deep markov models for incomplete multimodal time series, 2019.