Tractable Probabilistic Multimodal Learning

Jian Vora, Pranay Reddy Samala, Siddharth Chandak

CS 231N, Stanford University

INTRODUCTION

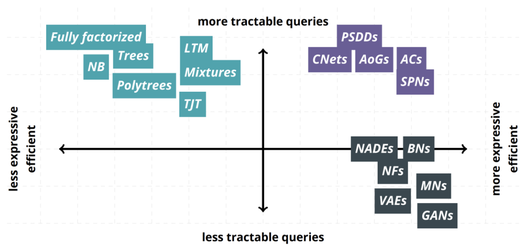
Given M modalities $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_M$ we aim to learn a generative model/joint distribution over all the modalities

$$p_\theta(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_M)$$

We want p_θ to allow us for tractable inference queries such as:

1. Marginalisation and conditioning on arbitrary modalities
2. Efficient sampling from the joint or any of the conditionals
3. Exact likelihood evaluation of any evidence

TRACTABLE PROBABILISTIC MODELS



Tractable Probabilistic Models are a class of generative models which allow for tractable inference for the following queries:

1. Complete Evidence (**EVI**): Assign the likelihood of a data point
2. Marginal Queries (**MAR**): Assign the likelihood and allow sampling from any marginal distribution
3. Conditional Queries (**CON**): Assign the likelihood and allow sampling from any conditional distribution derived from the joint
4. Maximum A Posterior (**MAP**), Marginal MAP (**MMAP**): Find variable assignment which maximizes some likelihood

Probabilistic Circuits (**PCs**) or Sum-Product Networks (**SPNs**) are a class of TPMs which are a computational graph encoding the joint as a multilinear polynomial of the input distributions using sum and product operations only

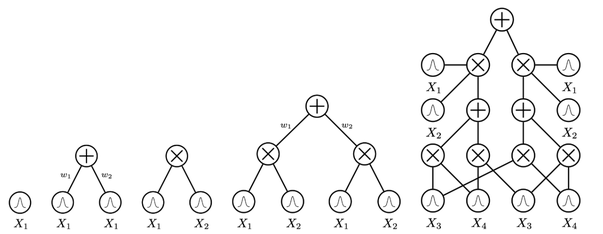


Figure 1. Various components of a probabilistic circuit: Sequentially building a full-fledged PC (rightmost) by composing sum and product nodes as a computational graph. The final output is the joint likelihood of the variables taking the assigned values. The above image was taken from http://web.cs.ucla.edu/~guyvdb/slides/AAAI20.pdf.

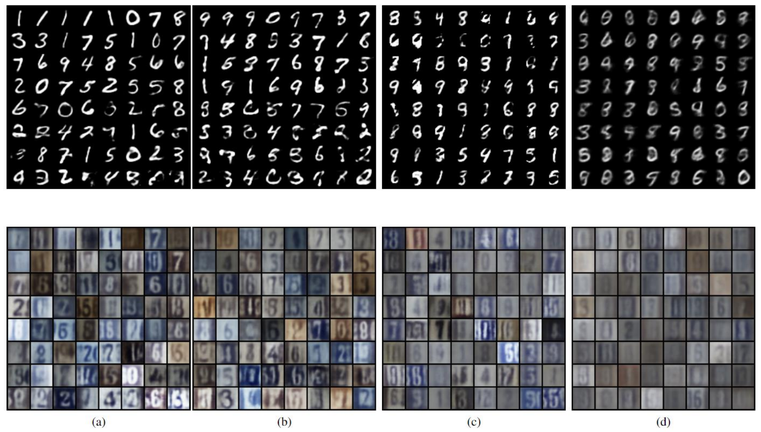
METHOD

Training Time

1. For each modality i, we train independent encoder-decoder pairs $\mathbf{E}_i$ and $\mathbf{D}_i$ and denote the latent code for each modality as $\mathbf{z}_i = \mathbf{E}_i(\mathbf{x}_i)$ and the reconstruction $\hat{\boldsymbol{x}}_i = \mathbf{D}_i(\mathbf{z}_i)$
2. Concatenate all the latent vectors to get $\mathbf{z} = \text{Concat}(\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_M)$
3. We learn a sum-product network $\mathcal{S}$ over the latent space $\mathbf{z}$ by maximizing likelihood of having seen the training data. Thus, our complete likelihood model is simply $\mathcal{S}(\text{Concat}(\mathbf{E}_1(\mathbf{x}_1), \mathbf{E}_2(\mathbf{x}_2), \ldots, \mathbf{E}_M(\mathbf{x}_M))$

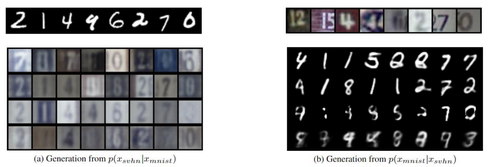
RESULTS: MNIST-SVHN

Model	Joint Qua(↓)	Joint Coh(↑)	$\text{Mod}_1 \rightarrow \text{Mod}_2$ FID(↓)	$\text{Mod}_1 \rightarrow \text{Mod}_2$ Acc(↑)	$\text{Mod}_2 \rightarrow \text{Mod}_1$ FID(↓)	$\text{Mod}_2 \rightarrow \text{Mod}_1$ Acc(↑)
MVAE	220.56	28.15	92.58	54.60	94.27	27.45
MMVAE	112.49	34.75	101.58	72.25	35.98	**59.02**
PPPC	**87.71**	**38.10**	**64.60**	**75.67**	**21.34**	47.35
AE	76.49	78.38	55.52	81.48	17.42	98.15



(a) (b) (c) (d)

Figure 3. The above figure shows unconditional samples for (a): GMM, (b): SPN, (c): MMVAE, (d): MVAE arranged as a column drawn from the joint distribution $p(x_{\text{mnist}}, x_{\text{svhn}})$. These samples are arranged in order of their likelihoods in a decreasing order



(a) Generation from $p(x_{svhn}|x_{mnist})$ (b) Generation from $p(x_{mnist}|x_{svhn})$

Figure 4. Samples drawn from conditional distribution with each row indicating a model in the order of (1): GMM, (2): SPN, (3): MMVAE, (4): MVAE. Qualitatively samples from SPNs are the most in sync with the digit classes and also show the least blurry artifacts unlike MVAE/MMVAE

Figure 2. Using the model to sample $\boldsymbol{x}_{k+1}, \ldots, \boldsymbol{x}_M$ given $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$

RESULTS: CELEBA ATTRIBUTES

(a) (b) (c)

Figure 5. The above figure shows unconditional samples (only the image modality) for (a): PPPC, (b): MMVAE, (c): MVAE drawn from the joint distribution arranged in order of decreasing likelihoods for each method

(a) (b) (c) (d)

Figure 6. The above figure shows conditional samples for (a): Blond Hair, (b): Glasses, (c): Male, (d): Mouth Open attributes true for PPPC. Samples from MVAE and MMVAE are in the Appendix as their performance is significantly worse.

Figure 7: Row 1: Blonde + Glasses ; Row 2: Male + Mouth open
Row 3: Blonde + Woman + Mouth closed; Row 4: Blonde + Woman + Mouth Open

Model	Joint Qua(↓)	Joint Coh(↑)
MVAE	70.264	0.252
MMVAE	93.031	0.236
PPPC	**66.713**	**0.126**
AE	58.871	0.0365

Model	Blond FID(↓)	Blond Acc(↑)	Glasses FID(↓)	Glasses Acc(↑)	Male FID(↓)	Male Acc(↑)	Mouth Open FID(↓)	Mouth Open Acc(↑)
MVAE	81.16	20.4	120.74	6.5	93.59	2.9	69.61	19.3
MMVAE	118.02	6.2	137.35	7.1	104.55	41.8	99.72	34.7
PPPC	**72.47**	**81.2**	**89.27**	**50.7**	**74.64**	**87.4**	**63.83**	**57.4**