Tractable Probabilistic Circuits

Guy Van den Broeck

Beyond Bayes: Paths Towards Universal Reasoning Systems - Jul 21, 2022
a unifying framework for tractable models
Probabilistic circuits

*computational graphs* that recursively define distributions

Simple distributions are tractable “black boxes” for:

- **EVI**: output $p(x)$ (density or mass)
- **MAR**: output 1 (normalized) or $Z$ (unnormalized)
- **MAP**: output the mode
Probabilistic circuits

computational graphs that recursively define distributions

\[ p(X_1) = w_1 p_1(X_1) + w_2 p_2(X_1) \]
⇒
mixtures

\[ p(X) = \begin{cases} p(Z = 1) \cdot p_1(X|Z = 1) \\ + p(Z = 2) \cdot p_2(X|Z = 2) \end{cases} \]
Probabilistic circuits

*computational graphs* that recursively define distributions

\[ p(X_1) = w_1 p_1(X_1) + w_2 p_2(X_1) \]

\[ p(X_1, X_2) = p(X_1) \cdot p(X_2) \]

⇒ *mixtures*

⇒ *factorizations*
Likelihood

\[ p(X_1 = -1.85, X_2 = 0.5, X_3 = -1.3, X_4 = 0.2) \]
Likelihood \[ p(X_1 = -1.85, X_2 = 0.5, X_3 = -1.3, X_4 = 0.2) \]
Likelihood $\ p(X_1 = -1.85, X_2 = 0.5, X_3 = -1.3, X_4 = 0.2)$
Tractable marginals

A sum node is *smooth* if its children depend on the same set of variables.

A product node is *decomposable* if its children depend on disjoint sets of variables.

---

Darwiche and Marquis, “A Knowledge Compilation Map”, 2002
Smoothness + decomposability = tractable MAR

If \( p(x) = \sum_i w_ip_i(x) \), (smoothness):

\[
\int p(x)dx = \int \sum_i w_ip_i(x)dx = \\
= \sum_i w_i \int p_i(x)dx
\]

\( \Rightarrow \) integrals are “pushed down” to children
Smoothness + decomposability = tractable MAR

If \( p(x, y, z) = p(x)p(y)p(z) \), (decomposability):

\[
\int \int \int p(x, y, z) \, dx \, dy \, dz = \\
= \int \int \int p(x)p(y)p(z) \, dx \, dy \, dz = \\
= \int p(x) \, dx \int p(y) \, dy \int p(z) \, dz
\]

\[\Rightarrow \text{integrals decompose into easier ones}\]
**Smoothness** + **decomposability** = **tractable MAR**

Forward pass evaluation for MAR

\[ \Rightarrow \text{linear in circuit size!} \]

E.g. to compute \( p(x_2, x_4) \):

- Leafs over \( X_1 \) and \( X_3 \) output \( Z_i = \int p(x_i) dx_i \)
  \[ \Rightarrow \text{for normalized leaf distributions: } 1.0 \]
- Leafs over \( X_2 \) and \( X_4 \) output **EVI**
- Feedforward evaluation (bottom-up)
Smoothness + decomposability = tractable MAR

Forward pass evaluation for MAR

\[ \Rightarrow \text{linear in circuit size!} \]

E.g. to compute \( p(x_2, x_4) \):
- leafs over \( X_1 \) and \( X_3 \) output \( Z_i = \int p(x_i) \, dx_i \)
  \[ \Rightarrow \text{for normalized leaf distributions: 1.0} \]
- leafs over \( X_2 \) and \( X_4 \) output EVI
- feedforward evaluation (bottom-up)
Learning Expressive Probabilistic Circuits

Hidden Chow-Liu Trees

Learned CLT structure captures strong pairwise dependencies

Learning Expressive Probabilistic Circuits

Hidden Chow-Liu Trees

Learned CLT structure captures strong pairwise dependencies

Learned HCLT structure

Compile into an equivalent PC

Learning Expressive Probabilistic Circuits

Hidden Chow-Liu Trees

Learned CLT structure captures strong pairwise dependencies

Learned HCLT structure

Compile into an equivalent PC

Mini-batch Stochastic Expectation Maximization

Lossless Data Compression

Expressive probabilistic model \( p(x) \) + Efficient coding algorithm

Determines the theoretical limit of compression rate

How close we can approach the theoretical limit

Lossless Neural Compression with Probabilistic Circuits

Probabilistic Circuits
- Expressive → SoTA likelihood on MNIST.
- Fast → Time complexity of en/decoding is $O( |p| \log(D) )$, where $D$ is the # variables and $|p|$ is the size of the PC.

Arithmetic Coding:
$$p(X_1 < x_1)$$
$$p(X_1 \leq x_1)$$
$$p(X_2 < x_2 | x_1)$$
$$p(X_2 \leq x_2 | x_1)$$
$$p(X_3 < x_3 | x_1, x_2)$$
$$p(X_3 \leq x_3 | x_1, x_2)$$
..
Lossless Neural Compression with Probabilistic Circuits

SoTA compression rates

<table>
<thead>
<tr>
<th>Dataset</th>
<th>HCLT (ours)</th>
<th>IDF</th>
<th>BitSwap</th>
<th>BB-ANS</th>
<th>JPEG2000</th>
<th>WebP</th>
<th>McBIts</th>
</tr>
</thead>
<tbody>
<tr>
<td>MNIST</td>
<td>1.24 (1.20)</td>
<td>1.96 (1.90)</td>
<td>1.31 (1.27)</td>
<td>1.42 (1.39)</td>
<td>3.37</td>
<td>2.09</td>
<td>(1.98)</td>
</tr>
<tr>
<td>FashionMNIST</td>
<td>3.37 (3.34)</td>
<td>3.50 (3.47)</td>
<td>3.35 (3.28)</td>
<td>3.69 (3.66)</td>
<td>3.93</td>
<td>4.62</td>
<td>(3.72)</td>
</tr>
<tr>
<td>EMNIST (Letter)</td>
<td>1.84 (1.80)</td>
<td>2.02 (1.95)</td>
<td>1.90 (1.84)</td>
<td>2.29 (2.26)</td>
<td>3.62</td>
<td>3.31</td>
<td>(3.12)</td>
</tr>
<tr>
<td>EMNIST (ByClass)</td>
<td>1.89 (1.85)</td>
<td>2.04 (1.98)</td>
<td>1.91 (1.87)</td>
<td>2.24 (2.23)</td>
<td>3.61</td>
<td>3.34</td>
<td>(3.14)</td>
</tr>
</tbody>
</table>

Compress and decompress 5-40x faster than NN methods with similar bitrates

<table>
<thead>
<tr>
<th>Method</th>
<th># parameters</th>
<th>Theoretical bpd</th>
<th>Codeword bpd</th>
<th>Comp. time (s)</th>
<th>Decomp. time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PC (HCLT, $M = 16$)</td>
<td>3.3M</td>
<td>1.26</td>
<td>1.30</td>
<td>9</td>
<td>44</td>
</tr>
<tr>
<td>PC (HCLT, $M = 24$)</td>
<td>5.1M</td>
<td>1.22</td>
<td>1.26</td>
<td>15</td>
<td>86</td>
</tr>
<tr>
<td>PC (HCLT, $M = 32$)</td>
<td>7.0M</td>
<td>1.20</td>
<td>1.24</td>
<td>26</td>
<td>142</td>
</tr>
<tr>
<td>IDF</td>
<td>24.1M</td>
<td>1.90</td>
<td>1.96</td>
<td>288</td>
<td>592</td>
</tr>
<tr>
<td>BitSwap</td>
<td>2.8M</td>
<td>1.27</td>
<td>1.31</td>
<td>578</td>
<td>326</td>
</tr>
</tbody>
</table>
Lossless Neural Compression with Probabilistic Circuits

Can be effectively combined with Flow models to achieve better generative performance

<table>
<thead>
<tr>
<th>Model</th>
<th>CIFAR10</th>
<th>ImageNet32</th>
<th>ImageNet64</th>
</tr>
</thead>
<tbody>
<tr>
<td>RealNVP</td>
<td>3.49</td>
<td>4.28</td>
<td>3.98</td>
</tr>
<tr>
<td>Glow</td>
<td>3.35</td>
<td>4.09</td>
<td>3.81</td>
</tr>
<tr>
<td>IDF</td>
<td>3.32</td>
<td>4.15</td>
<td>3.90</td>
</tr>
<tr>
<td>IDF++</td>
<td><strong>3.24</strong></td>
<td>4.10</td>
<td>3.81</td>
</tr>
<tr>
<td>PC+IDF</td>
<td>3.28</td>
<td><strong>3.99</strong></td>
<td><strong>3.71</strong></td>
</tr>
</tbody>
</table>
PC Learners keep getting better! … stay tuned …

Table 1: Density estimation performance on MNIST-family datasets in test set bpd.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Sparse PC (ours)</th>
<th>HCLT</th>
<th>RatSPN</th>
<th>IDF</th>
<th>BitSwap</th>
<th>BB-ANS</th>
<th>McBits</th>
</tr>
</thead>
<tbody>
<tr>
<td>MNIST</td>
<td><strong>1.14</strong></td>
<td>1.20</td>
<td>1.67</td>
<td>1.90</td>
<td>1.27</td>
<td>1.39</td>
<td>1.98</td>
</tr>
<tr>
<td>EMNIST(MNIST)</td>
<td><strong>1.52</strong></td>
<td>1.77</td>
<td>2.56</td>
<td>2.07</td>
<td>1.88</td>
<td>2.04</td>
<td>2.19</td>
</tr>
<tr>
<td>EMNIST(Letters)</td>
<td><strong>1.58</strong></td>
<td>1.80</td>
<td>2.73</td>
<td>1.95</td>
<td>1.84</td>
<td>2.26</td>
<td>3.12</td>
</tr>
<tr>
<td>EMNIST(Balanced)</td>
<td><strong>1.60</strong></td>
<td>1.82</td>
<td>2.78</td>
<td>2.15</td>
<td>1.96</td>
<td>2.23</td>
<td>2.88</td>
</tr>
<tr>
<td>EMNIST(ByClass)</td>
<td><strong>1.54</strong></td>
<td>1.85</td>
<td>2.72</td>
<td>1.98</td>
<td>1.87</td>
<td>2.23</td>
<td>3.14</td>
</tr>
<tr>
<td>FashionMNIST</td>
<td><strong>3.27</strong></td>
<td>3.34</td>
<td>4.29</td>
<td>3.47</td>
<td>3.28</td>
<td>3.66</td>
<td>3.72</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dataset</th>
<th>PC</th>
<th>Bipartite flow</th>
<th>AF/SCF</th>
<th>IAF/SCF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Penn Treebank</td>
<td><strong>1.23</strong></td>
<td>1.38</td>
<td>1.46</td>
<td>1.63</td>
</tr>
</tbody>
</table>
Expressive models without compromises
Queries as pipelines: KLD

$$\text{KLD}(p \parallel q) = \int p(x) \times \log((p(x)/q(x))) \, dx$$
Inference by tractable operations

systematically derive tractable inference algorithm of complex queries

<table>
<thead>
<tr>
<th>Query</th>
<th>Tract. Conditions</th>
<th>Hardness</th>
</tr>
</thead>
<tbody>
<tr>
<td>CROSS ENTROPY</td>
<td>Cmp, q Det</td>
<td>#P-hard w/o Det</td>
</tr>
<tr>
<td>SHANNON ENTROPY</td>
<td>Sm, Dec, Det</td>
<td>coNP-hard w/o Det</td>
</tr>
<tr>
<td>RéNYI ENTROPY</td>
<td>SD</td>
<td>#P-hard w/o SD</td>
</tr>
<tr>
<td>MUTUAL INFORMATION</td>
<td>Sm, Det, Det*</td>
<td>coNP-hard w/o SD</td>
</tr>
<tr>
<td>KULLBACK-LEIBLER DIV.</td>
<td>Cmp, Det</td>
<td>#P-hard w/o Det</td>
</tr>
<tr>
<td>RéNYI'S ALPHA DIV.</td>
<td>Cmp, q Det</td>
<td>#P-hard w/o Det</td>
</tr>
<tr>
<td>ITAKURA-SAITO DIV.</td>
<td>Cmp, Det</td>
<td>#P-hard w/o Det</td>
</tr>
<tr>
<td>CAUCHY-SCHWARZ DIV.</td>
<td>Cmp</td>
<td>#P-hard w/o Cmp</td>
</tr>
<tr>
<td>SQUARED LOSS</td>
<td>Cmp</td>
<td>#P-hard w/o Cmp</td>
</tr>
</tbody>
</table>

tractability is a spectrum
Learn more about probabilistic circuits?

Tutorial (3h)

Overview Paper (80p)

Probabilistic Circuits: A Unifying Framework for Tractable Probabilistic Models

YooJung Choi
Antonio Vergari
Guy Van den Broeck
University of California, Los Angeles
Los Angeles, CA, USA

Contents

1 Introduction .................................................. 3
2 Probabilistic Inference: Models, Queries, and Tractability ........................................ 4
   2.1 Probabilistic Models ........................................ 5
   2.2 Probabilistic Queries ....................................... 6
   2.3 Tractable Probabilistic Inference ....................... 8
   2.4 Properties of Tractable Probabilistic Models .......... 9

https://youtu.be/2RAG5-L9R70

Thanks

This was the work of many wonderful students/postdocs/collaborators!