38th Conference on Uncertainty in Artificial Intelligence
Eindhoven, Netherlands
August 1-5, 2022
uai2022

Variational Message Passing Neural Network for Maximum-A-Posteriori (MAP) Inference

Zijun Cui, Hanjing Wang, Tian Gao, Kartik Talamadupula, Qiang Ji
cuiz3@rpi.edu jiq@rpi.edu

CONTRIBUTIONS

- ❑ Instead of relying on a fixed and pre-defined variational distribution, we propose a neural-augmented free energy where **variational distribution is parameterized via a neural network**. An optimal variational distribution is explored during training.
- ❑ Minimization of the neural-augmented free energy is achieved through a message passing neural network (MPNN). The training of the MPNN is guided by the neural-augmented free energy, **without requiring labeled training data**.
- ❑ We achieve **outstanding inference performance** compared to both training-free methods and training-based methods.

INTRODUCTION

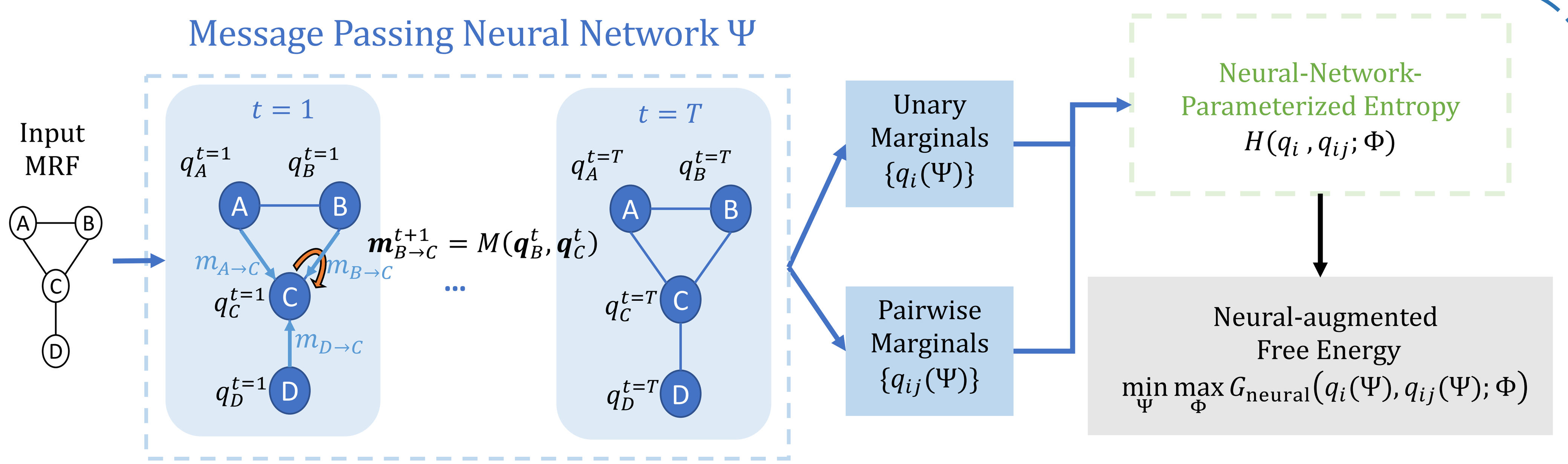
❑ MAP in Markov Random Fields (MRFs)

- For a set of N random variables $\{x_i\}_{i=1}^N$, an MRF $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ captures its joint distribution with $|\mathcal{V}| = N$ and $|\mathcal{E}| = M$. M is the total number of edges in the MRF. The joint distribution is defined as
$$p(\boldsymbol{x}) \propto \exp(\sum_{i\in\mathcal{V}} \theta_i(x_i) + \sum_{(i,j)\in\mathcal{E}} \theta_{ij}(x_i, x_j))$$
where $\boldsymbol{\theta} = \{\theta_i(x_i), \theta_{ij}(x_i, x_j)\}$ refers to probability parameters.
- MAP inference in MRF is formulated as
$$\boldsymbol{x}^* = arg\max_{\boldsymbol{x}} p(\boldsymbol{x}) = arg\max_{\boldsymbol{x}} \sum_{i\in\mathcal{V}} \theta_i(x_i) + \sum_{(i,j)\in\mathcal{E}} \theta_{ij}(x_i, x_j)$$

❑ Variational Belief Propagation (BP) for MAP inference

- Variational belief propagation approach formulates MAP inference as an optimization problem where a variational distribution is obtained by minimizing a variational free energy under a variational assumption.
- In MRF, it is natural to assume a variational distribution $q(x)$ is a function of $\{q_i(x_i)\}_{i\in\mathcal{V}}$ and $\{q_{ij}(x_i, x_j)\}_{(i,j)\in\mathcal{E}}$, referred to as *pairwise assumption*.
- Under pairwise assumption, we have the variational free energy of form:
$$G_{\text{pairwise}}(\{q_i\}, \{q_{ij}\}) = U(\{q_i\}, \{q_{ij}\}) - \epsilon(\sum_{i\in\mathcal{V}} c_i H(q_i) + \sum_{(i,j)\in\mathcal{E}} c_{ij} H(q_i, q_j))$$
with average energy $U(\{q_i\}, \{q_{ij}\}) = -\sum_{i\in\mathcal{V}} q_i(x_i)\theta_i(x_i) - \sum_{(i,j)\in\mathcal{E}} q_{ij}(x_i, x_j)\theta_{ij}(x_i, x_j)$. ϵ is a sufficient small value. $H(q_i)$ denotes the entropy of q_i. $H(q_i, q_j)$ denotes the entropy of $q_{ij}(x_i, x_j)$.
- An optimal variational distribution is obtained as
$$\{q_i^*, q_{ij}^*\} = arg\min_{\{q_i, q_{ij}\}} G_{pairwise}(\{q_i\}, \{q_{ij}\})$$
MAP inference is performed as $x_i^* = argmax\, q_i^*(x_i)$.
- Each of the variational BP algorithms is specific to a family of variational distributions, leading to an entropy approximation (i.e., a set of c_i and c_{ij}). The performance of MAP inference is limited by the corresponding variational assumption.

METHODS



❑ Neural-augmented Free Energy

- We propose a neural-augmented free energy G_{neural} where we parameterize variational distribution families through neural network parameters Φ as
$$G_{\text{neural}}(\boldsymbol{q}^{node}, \boldsymbol{q}^{edge}; \Phi) = U(\boldsymbol{q}^{node}, \boldsymbol{q}^{edge}) - \epsilon H(\boldsymbol{q}^{node}, \boldsymbol{q}^{edge}; \Phi)$$
with $\boldsymbol{q}^{node} = \{q_i(x_i)\}_{i\in\mathcal{V}}$ and $\boldsymbol{q}^{edge} = \{q_{ij}(x_i, x_j)\}_{(i,j)\in\mathcal{E}}$. Parameterization of variational distribution families is implicitly achieved via the neural-network-parameterized entropy $H(\boldsymbol{q}^{node}, \boldsymbol{q}^{edge}; \Phi)$.
- Theoretically performance guarantee with G_{neural} is provided through three propositions:
[Proposition 1]: Neural-augmented free energy G_{neural} is provable convex with a strictly concave neural-network-parameterized approximation $H(\boldsymbol{q}^{node}, \boldsymbol{q}^{edge}; \Phi)$.
[Proposition 2]: MAP inference error $\Delta_{map}(q_\Phi^*, p)$ is upper bounded by an entropy approximation scaled by ϵ, i.e., $\Delta_{map}(q_\Phi^*, p) \leq \epsilon H(q_\Phi^*; \Phi)$. The minimal MAP inference error is hence upper bounded by an optimal entropy approximation with $\Phi^* = arg\min_\Phi H(q_\Phi^*; \Phi)$.
[Proposition 3]: Neural-augmented free energy subsumes existing variational distribution families as a strict generalization. The optimal MAP inference performance achieved with neural-augmented free energy is superior or comparable to existing variational distribution families, i.e., $\Delta_{map}(q_{\Phi^*}^*, p) \leq \Delta_{map}(q_{\Phi^{fix}}^*, p)$.

❑ Minimization of Neural-augmented Free Energy with MPNN

- To minimize G_{neural} , we employ MPNN which performs inference through message passing with messages parameterized via neural network parameters Ψ. Each node in MPNN is mapped to a variable in MRF. Node feature $\{\boldsymbol{h}_i\}_{i=1}^N$ corresponds to the unary marginal estimation $\{\boldsymbol{q}_i\}_{i=1}^N$ in logarithmic .
- At each iteration t, i-th node receives a message from its neighbor j-th node through message function $\mathcal{M}$ as
$$\boldsymbol{m}_{j\to i}^{t+1} = \mathcal{M}(\boldsymbol{h}_i^t, \boldsymbol{m}_{i\to j}^t, \theta_{ij})$$
$\mathcal{M}$ is realized through MLP containing free parameters Ψ to be learned. Each node then update its feature as $\boldsymbol{h}_i^{t+1} = \boldsymbol{m}_i^{t+1} + \theta_i - \ln(z_i^{t+1})$ with the aggregated message $\boldsymbol{m}_i^{t+1} = \sum_{j\in\mathcal{N}(i)} \boldsymbol{m}_{j\to i}^{t+1}$. $z_i^{t+1} = \sum_{x_i} \exp(\boldsymbol{m}_i^{t+1} + \theta_i)$. The update process is repeated until convergence. In the end, unary and pairwise marginal estimations are extracted by following BP's belief equation.

❑ Training Objectives

- The total training objective is based on neural-augmented free energy
$$\min_\Psi \max_\Phi G_{\text{neural}}(\boldsymbol{q}^{node}(\Psi), \boldsymbol{q}^{edge}(\Psi); \Phi)$$
- Two phase alternative update is considered for effective training. At each iteration r, we firstly update Ψ as
$$\Psi^{r+1} = arg\min_\Psi G_{\text{neural}}(\boldsymbol{q}^{node}(\Psi), \boldsymbol{q}^{edge}(\Psi); \Phi^r)$$
We then update Φ as
$$\Phi^{r+1} = arg\max_\Phi G_{\text{neural}}(\boldsymbol{q}^{node}(\Psi^{r+1}), \boldsymbol{q}^{edge}(\Psi^{r+1}); \Phi) = arg\min_\Phi H(\boldsymbol{q}^{node}(\Psi^{r+1}), \boldsymbol{q}^{edge}(\Psi^{r+1}); \Phi)$$
According to the proposition 2, Φ is updated in the direction of minimizing the MAP inference error.
- After training, only MPNN module with optimal parameter Ψ^* is required for MAP inference. MAP configuration is obtained as $x_i^* = argmax\, q_i(x_i; \Psi^*)$.

EXPERIMENTS

❑ Compared to training-free methods

- Training-free methods refer to optimization algorithms.
- V-MPNN is better particularly on complex and larger graphs by leveraging neural-augmented free energy.

Graph	N=9				N=15			
	BP	TRW-BP	MPLP	V-MPNN	BP	TRW-BP	MPLP	V-MPNN
STAR	**1.0**	.99	**1.0**	.93	**1.0**	**1.0**	**1.0**	.74
TREE	**1.0**	.99	**1.0**	.96	**1.0**	**1.0**	**1.0**	.93
PATH	**1.0**	**1.0**	**1.0**	.97	**1.0**	**1.0**	**1.0**	.93
CYCLE	**.91**	.76	.90	.85	.84	.84	**.89**	.87
LADDER	.68	.66	.72	**.77**	.63	.61	.67	**.72**
2D GRID	.57	.48	**.74**	**.74**	.56	.50	.63	**.69**
CIRCULAR LADDER	.62	.50	.76	**.83**	.61	.53	.63	**.73**
BARBELL	.57	.55	.67	**.71**	.60	.57	.64	**.66**
LOLLIPOP	.59	.60	.61	**.88**	.62	.55	.58	**.67**
WHEEL	.56	.44	.62	**.70**	.58	.50	.62	**.69**
BIPARTITE	.54	.52	.62	**.74**	.62	.56	.55	**.64**
TRIPARTITE	.57	.62	.52	**.68**	.52	.55	.51	**.65**
COMPLETE	.56	.60	.49	**.65**	.54	.54	.53	**.60**
MEAN	.71	.67	.73	**.80**	.70	.67	.69	**.73**

❑ Compared to training-based methods

- Training-based methods refer to neural-network-based models that require exact inference results for training.
- V-MPNN is better particularly on simple and sparse graphs by leveraging the injected well-established theories.

Graph	N=9		N=15	
	Node-GNN	V-MPNN	Node-GNN	V-MPNN
STAR	.65	**.93**	.52	**.74**
TREE	.77	**.96**	.75	**.93**
PATH	.81	**.97**	**.73**	**.93**
CYCLE	.79	**.85**	.75	**.87**
LADDER	.72	**.77**	.69	**.72**
2D GRID	.72	**.74**	**.74**	.69
C-LADDER	.81	**.83**	.71	**.73**
BARBELL	**.72**	.71	**.71**	.66
LOLLIPOP	.72	**.88**	**.69**	.67
WHEEL	.68	**.70**	**.70**	.69
BIPARTITE	**.75**	.74	**.74**	.64
TRIPARTITE	**.73**	.68	**.72**	.65
COMPLETE	**.82**	.65	.70	.60
MEAN	.75	**.80**	.70	**.73**

CONCLUSION

- ❑ A Variational message passing neural network (V-MPNN) is proposed, leveraging both the power of neural network (in both modeling complex functions and conducting message passing mechanism), and the well-established algorithmic theories on variational belief propagation.
- ❑ Outstanding inference performance is achieved compared against both training-free and training-based methods.

[Acknowledgement] This work is supported by the Rensselaer-IBM AI Research Collaboration (http://airc.rpi.edu), part of the IBM AI Horizons Network (http://ibm.biz/AIHorizons)