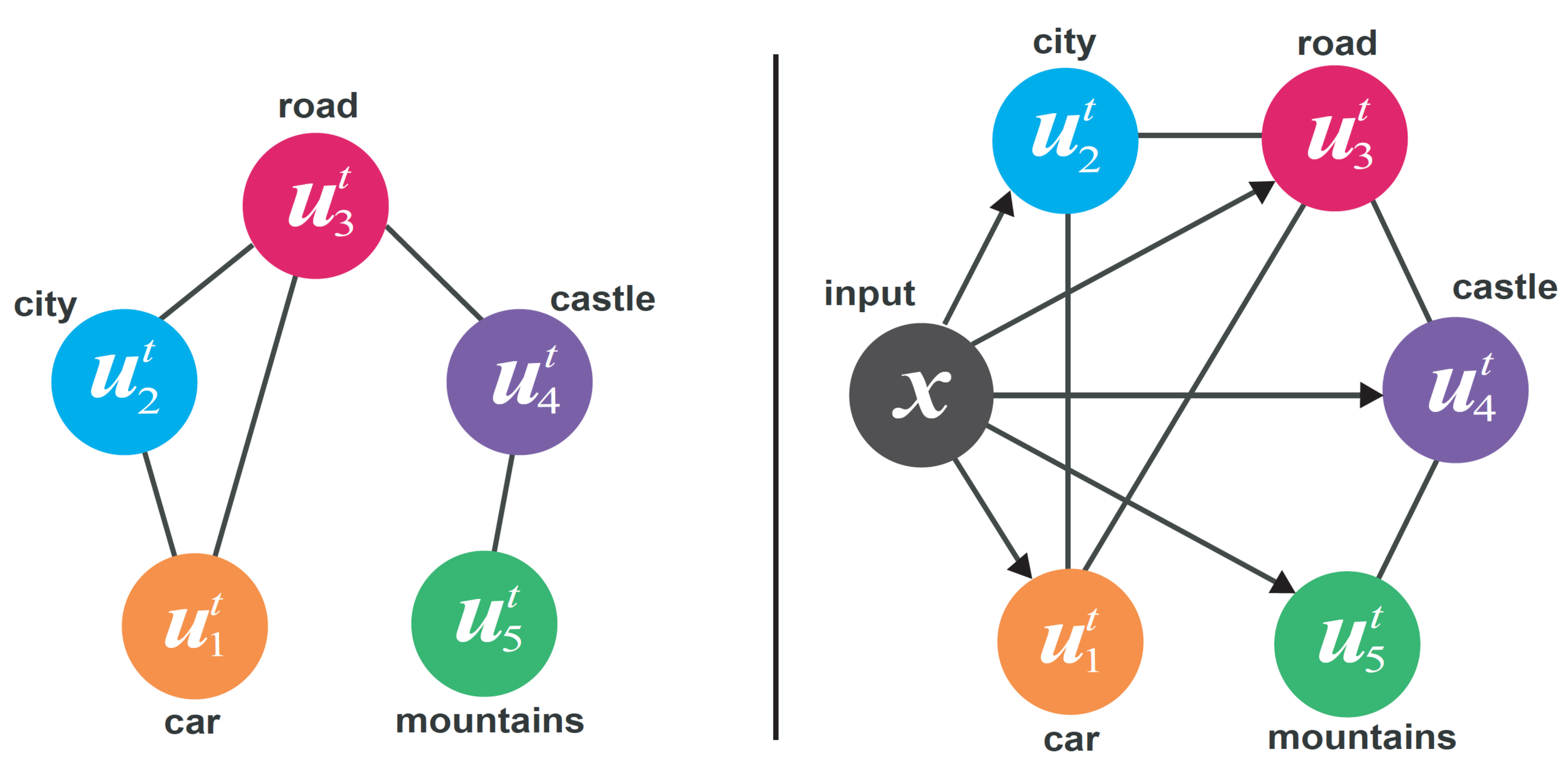
Neural Message Passing for Multi-Label Classification

Jack Lanchantin, Arshdeep Sekhon, Yanjun Qi
University of Virginia, Department of Computer Science

Overview

- We propose Label Message Passing (LaMP) Networks to model the joint prediction of labels by treating labels as nodes on a graph

Message Passing Neural Networks (MPNNs)



- Joint representations of nodes and edges are modelled using message passing rather than explicit probabilistic formulations, allowing for efficient inference
- Hidden state $v_i^t \in \mathbb{R}^d$ of node $i \in G$ is updated based on messages m_i^t from neighboring nodes $\{v_{j \in \mathcal{N}(i)}^t\}$ defined by neighborhood $\mathcal{N}(i)$:

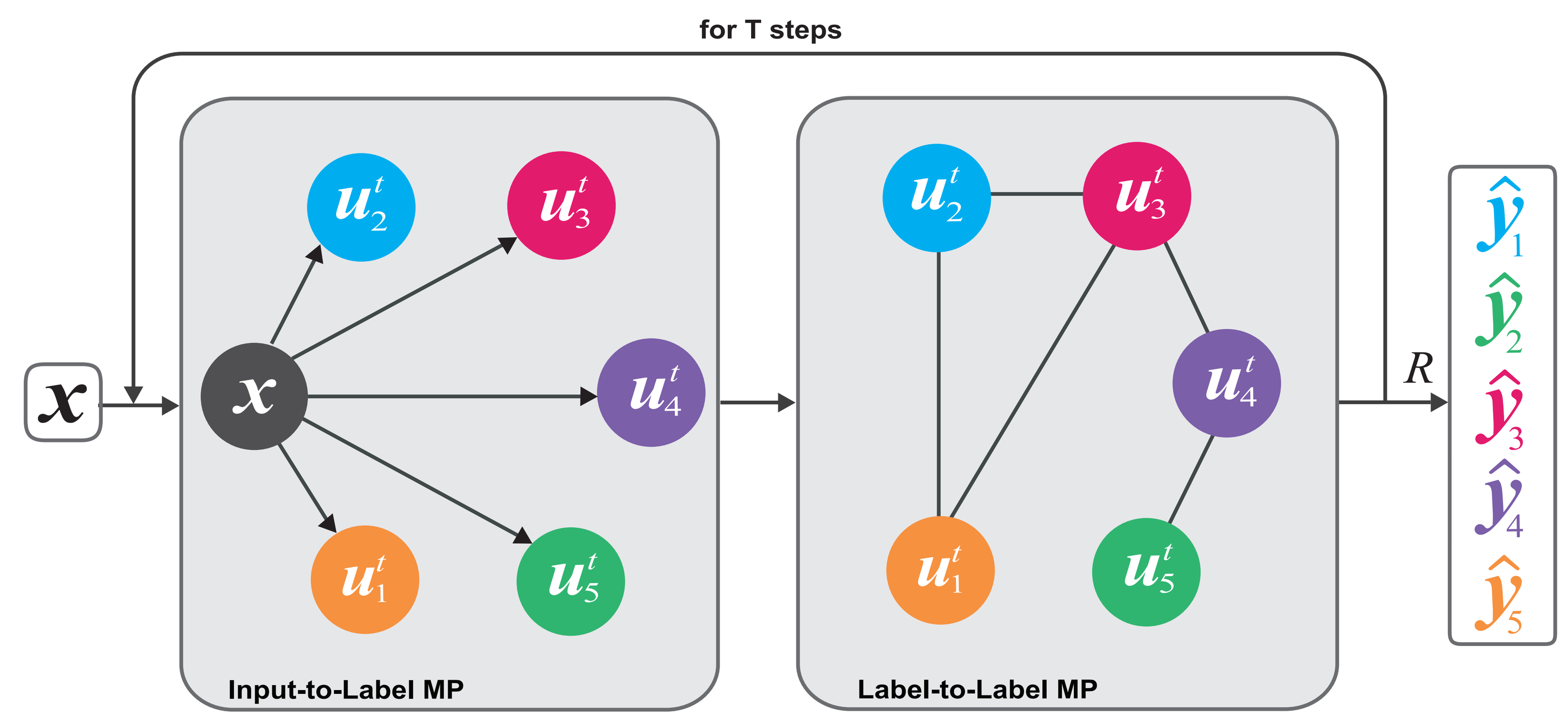
$$m_i^t = \Sigma_{j \in \mathcal{N}(i)} F_m(v_i^t, v_j^t),$$

$$v_i^{t+1} = F_u(m_i^t)$$

Multi-Label Classification Setup

- Goal: predict the set of labels $\{y_1, y_2, \ldots, y_L\}$, $y_i \in \{0, 1\}$ given x
- We represent the input x as feature vector feature vector $x \in \mathbb{R}^d$
- Labels first represented as embedded vectors $\{u_1^{t=0}, u_2^{t=0}, \ldots, u_L^{t=0}\}$, $u_i^t \in \mathbb{R}^d$
- The key idea of LaMP networks is that labels are represented as nodes in a label-interaction graph G_{yy} where nodes are vectors $\{u_{1:L}^t\}$
- Given x, LaMP models the conditional dependencies between label embeddings $\{u_1^t, u_2^t, \ldots, u_L^t\}$ using Message Passing Neural Networks

Label Message Passing Networks (LaMP)



Feature-to-Label Message Passing

Passes messages from input x to each label embedding u_i^t

$$m_i^t = F_m(u_i^t, x),$$
$$u_i^{t'} = F_u(m_i^t).$$

Label-to-Label Message Passing

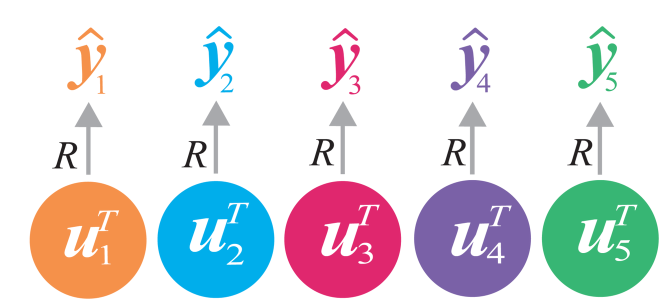
Passes messages between label embeddings to update their states conditioned on x

$$m_i^{t'} = \Sigma_{j \in \mathcal{N}(i)} F_m(u_i^{t'}, u_j^{t'}),$$
$$u_i^{t+1} = F_u(m_i^{t'}).$$

Readout Layer

Predicts the probabilities of each label being positive $\{\hat{y}_1, \ldots \hat{y}_L\}$

$$\hat{y}_i = R(u_i^T; \mathbf{W}^o) = \text{sigmoid}(\mathbf{W}_i^o u_i^T).$$

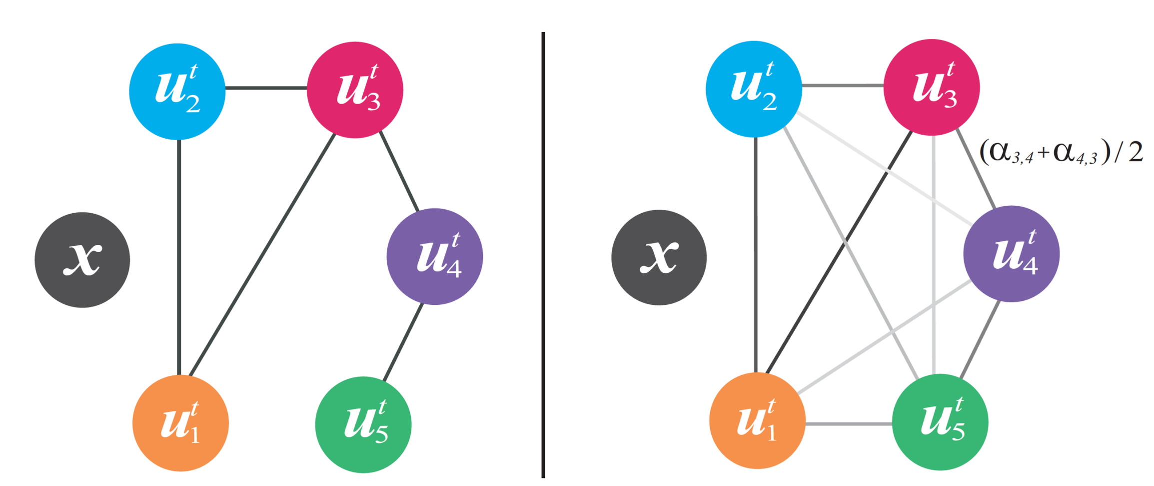


Loss Function

$$Loss(y, \hat{y}^t) = \frac{1}{T} \sum_{t=0}^{T} \frac{1}{L} \sum_{i=1}^{L} -(y_i \log(\hat{y}_i^t) + (1 - y_i) \log(1 - \hat{y}_i^t))$$

Label Graph Structure

- Prior: Use known label structure or place edges between co-occurring labels
- Fully Connected: Use attention to learn the graph while training the classifier



Results

- We validate the benefits of LaMP on eight real-world MLC datasets
- Three LaMP variants: LaMP$_{el}$ uses an edgeless label graph assuming no label dependencies, LaMP$_{fc}$ uses a fully connected label graph, and LaMP$_{pr}$ uses a prior label graph

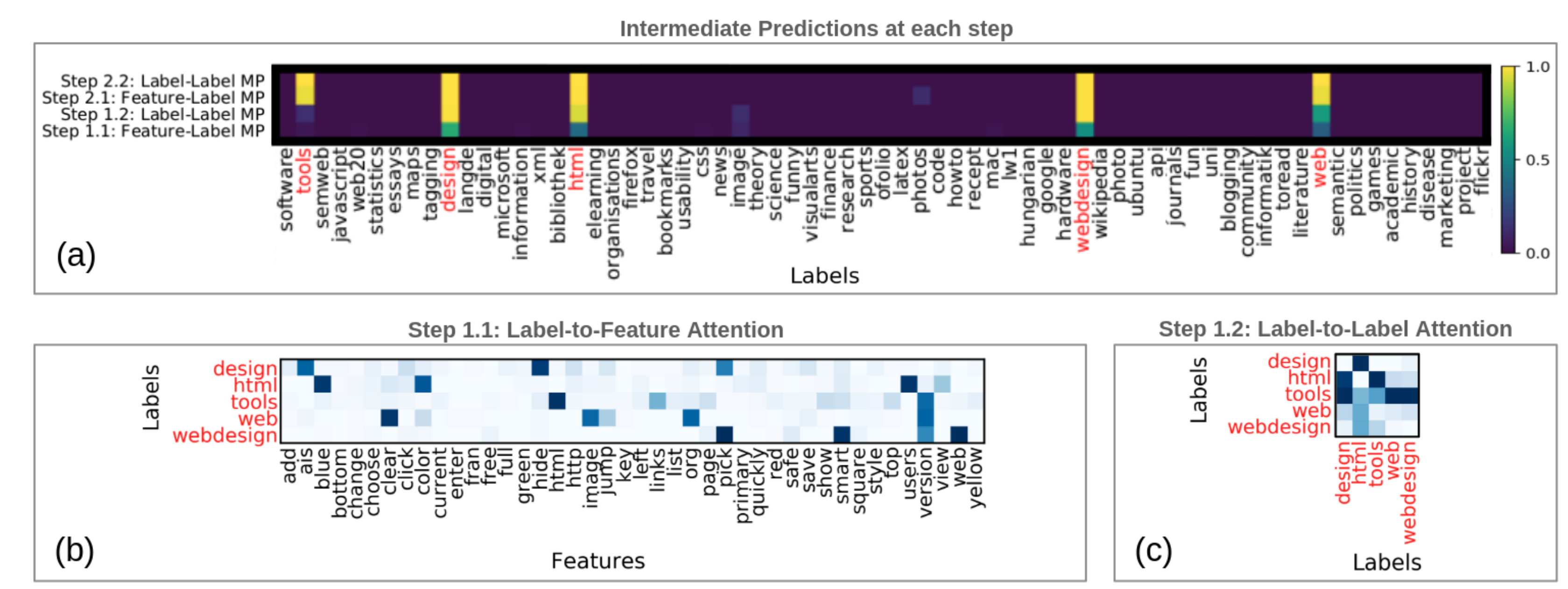
Performance

Example-based F1 scores across all 8 datasets

	Reuters	Bibtex	Bookmarks	Delicious	RCV1	TFBS	NUSWIDE	SIDER
FastXML[1]	-	-	-	-	0.841	-	-	-
Madjarov[2]	-	0.434	0.257	0.343	-	-	-	-
SPEN[3]	-	0.422	0.344	0.375	-	-	-	-
RNN Seq2Seq[4]	0.894	0.393	0.362	0.320	**0.890**	0.249	0.329	0.356
MLP	0.854	0.363	0.368	0.371	0.865	0.167	0.371	**0.766**
LaMP$_{el}$	0.883	0.435	0.375	0.369	0.887	0.310	0.376	**0.766**
LaMP$_{pr}$	0.902	**0.447**	0.386	**0.372**	0.887	0.321	0.372	**0.766**
LaMP$_{fc}$	**0.906**	0.445	**0.389**	**0.372**	0.889	**0.321**	0.376	0.764

Interpretability

Visualization of intermediate predictions and attention scores



Speed

Each column shows training or testing speed for LaMP in minutes per epoch. Speedups over RNN Seq2Seq are in parentheses

Dataset	Training	Testing
Reuters	0.788 (1.5x)	0.116 (2.1x)
Bibtex	0.376 (2.1x)	0.080 (2.1x)
Delicious	3.172 (1.1x)	0.473 (3.2x)
Bookmarks	9.664 (1.2x)	1.849 (1.3x)
RCV1	98.346 (1.2x)	1.003 (1.7x)
TFBS	187.14 (2.5x)	13.04 (4.2x)
NUS-WIDE	3.201 (1.2x)	0.921 (8.0x)
SIDER	0.027 (2.5x)	0.003 (21x)

References

[1] Y. Prabhu and M. Varma, "Fastxml: A fast, accurate and stable tree-classifier for extreme multi-label learning," in *ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, 2014.

[2] G. Madjarov, D. Kocev, D. Gjorgjevikj, and S. Džeroski, "An extensive experimental comparison of methods for multi-label learning," *Pattern recognition*, vol. 45, no. 9, 2012.

[3] D. Belanger and A. McCallum, "Structured prediction energy networks," in *International Conference on Machine Learning*, pp. 983–992, 2016.

[4] J. Nam, E. L. Mencía, H. J. Kim, and J. Fürnkranz, "Maximizing subset accuracy with recurrent neural networks in multi-label classification," in *Advances in Neural Information Processing Systems*, 2017.