AIN: Fast and Accurate Sequence Labeling with Approximate Inference Network

Xinyu Wang, Yong Jiang, Nguyen Bach, Tao Wang, Zhongqiang Huang, Fei Huang, Kewei Tu

School of Information Science and Technology, ShanghaiTech University
DAMO Academy, Alibaba Group

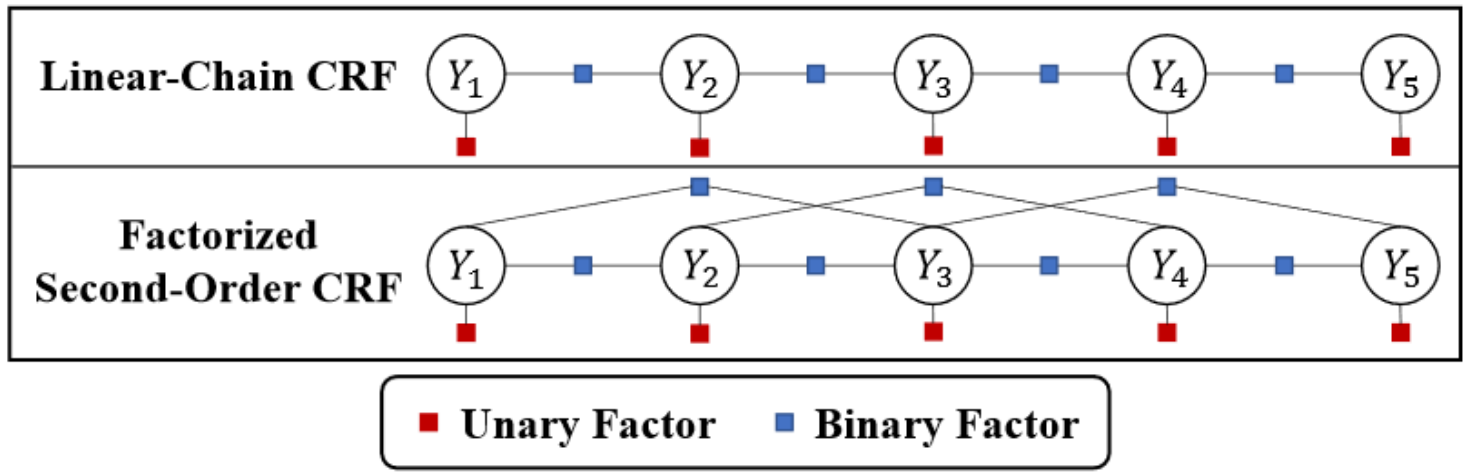
上海科技大学
ShanghaiTech University

DAMO
ALIBABA DAMO ACADEMY

Motivation

- Very fast sequence labelers are sometimes required for training and prediction
- BiLSTM-CRF model is one of the most successful approach to sequence labeling
- However, the CRF layer contains sequential computation. Which is difficult to be parallelized on GPU

Factors of the CRF layer

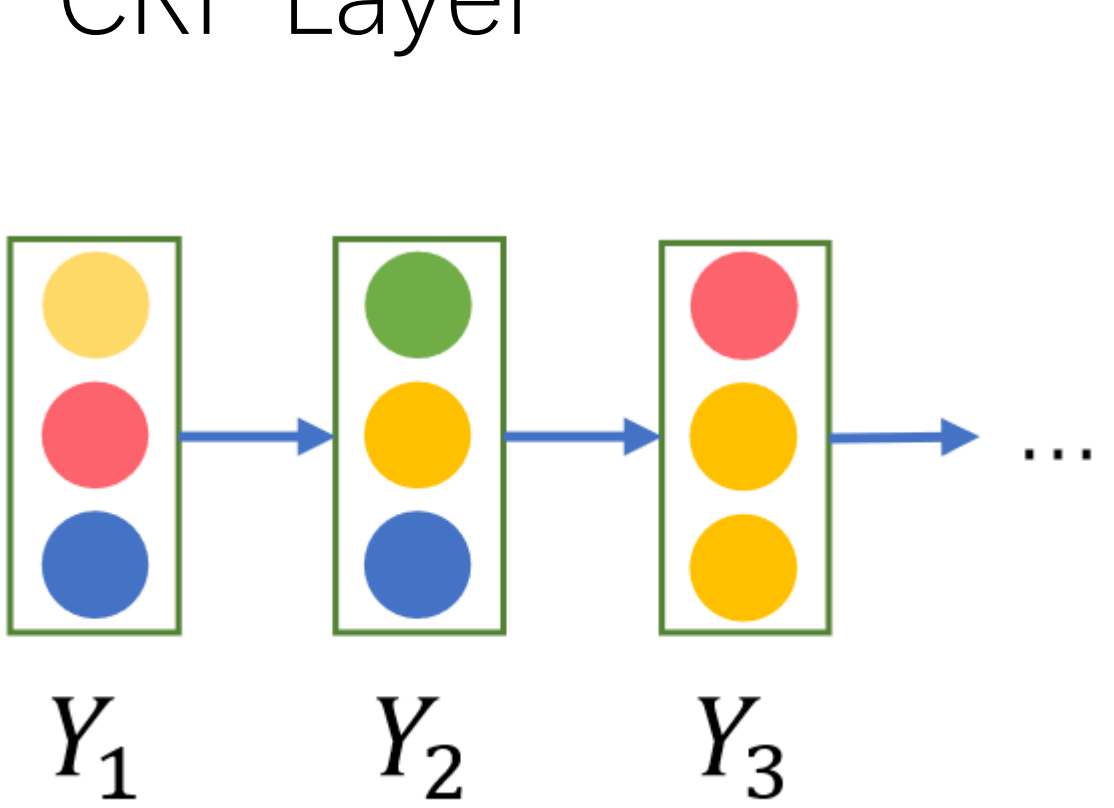


$$\psi(\mathbf{x}, \mathbf{y}, i) = \psi_u(\mathbf{x}, y_i) + \psi_b(y_{i-1}, y_i)$$

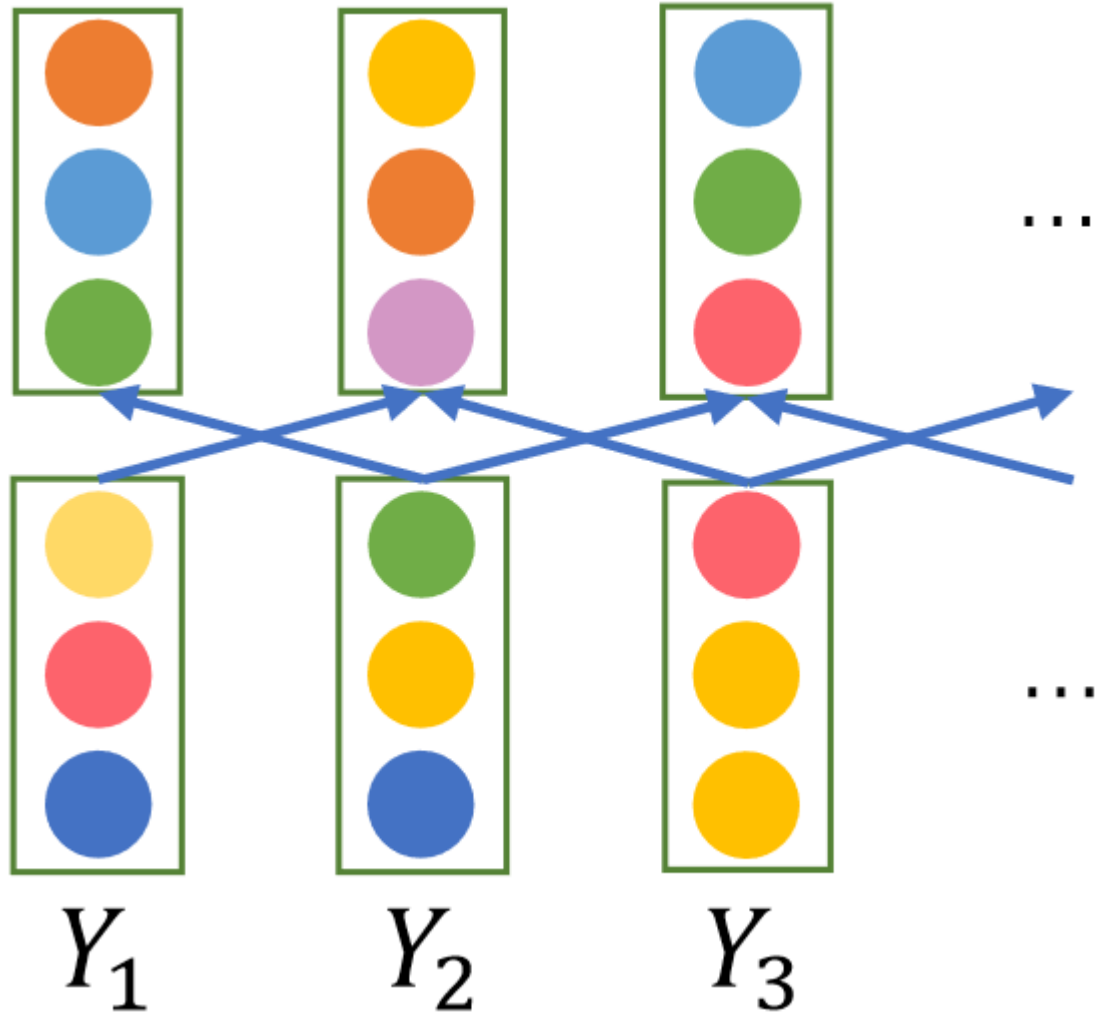
$$\psi(\mathbf{x}, \mathbf{y}, i) = \psi_u(\mathbf{x}, y_i) + \psi_t(y_{i-2}, y_{i-1}, y_i)$$

$$\psi_t(y_{i-2}, y_{i-1}, y_i) = \psi_{b'}(y_{i-2}, y_i) + \psi_b(y_{i-1}, y_i)$$

Exact and Approximate Inference on the CRF Layer



Viterbi



Mean Field Variational Inference

Approximate Inference on the CRF layer

- Approximates $P(\mathbf{y}|\mathbf{x})$ with a factorized distribution $Q(\mathbf{y}|\mathbf{x}) = \prod_{i=1}^{n} Q_i(y_i|\mathbf{x})$
- Iteratively minimize the KL divergence $KL(Q||P)$

Approximate Inference on the CRF layer

- Message Passing: $$s(i,j,k) := \sum_{y_i=1}^{L} Q_i^{k-1}(y_i|\mathbf{x})\psi_b(y_{\min\{i,j\}}, y_{\max\{i,j\}})$$

- Approximate inference on the linear-chain CRF

$$Q_i^m(y_i|\mathbf{x}) \propto \exp\{\psi_u(\mathbf{x}, y_i) + s(i-1, i, m) + s(i+1, i, m)\}$$

- Approximate inference on the second-order CRF

$$Q_i^m(y_i|\mathbf{x}) \propto \exp\{\psi_u(\mathbf{x}, y_i) + s'(i-2, i, m) + s(i-1, i, m) + s(i+1, i, m) + s'(i+2, i, m)\}$$

Time complexity

- Traditional CRF layer (forward-backward/Viterbi algorithms)
  - CPU: $O(nL^2)$
  - GPU: $O(n \log L)$
- AIN
  - CPU: $O(nL^2)$
  - GPU: $O(M \log L)$ (where M is a constant representing the inference steps and is typically very small)

Experiments

- Named Entity Recognition (NER)
- Chunking
- Part-Of-Speech (POS) Tagging
- Slot Filling

Encoders

- BiLSTM
- CNN
- Linear

Speed Comparison

	WORD-CHAR-BILSTM								WORD-CNN			
	Training				Prediction				Training		Prediction	
# Words	32		128		32		128		32	128	32	128
	All	Dec.	All	Dec.	All	Dec.	All	Dec.	All	All	All	All
MaxEnt*	6.8×	-	13.1×	-	3.0×	-	5.9×	-	12.9×	40.1×	6.3×	18.6×
AIN-1O	4.3×	7.7×	10.2×	31.4×	1.7×	2.4×	4.4×	12.7×	5.6×	21.5×	2.4×	6.8×
AIN-F2O	3.5×	5.3×	8.7×	20.1×	1.5×	1.9×	4.1×	10.6×	4.4×	16.7×	1.8×	5.5×

Table 1: Relative speedup over the **CRF** model with 10,000 sentences of 32/128 words. **All** represents the speed of the full model. **Dec.** represents the speed of decoder. ⋆: For reference.

Performance Comparison

	WORD-CHAR-BILSTM					WORD-CNN					WORD ONLY				
	NER	POS	Chunk	SF	Avg.	NER	POS	Chunk	SF	Avg.	NER	POS	Chunk	SF	Avg.
MaxEnt*	83.74	94.84	92.58	95.47	91.65	75.19	94.00	87.05	91.07	86.83	52.27	90.53	78.17	62.93	70.98
CRF	84.17	94.91	**92.88**	95.52	91.87	**79.44**	94.26	**89.21**	92.24	**88.79**	**72.28**	92.79	**89.39**	76.82	82.82
AIN-1O	**84.22**	**94.97**	92.87	**95.59**	**91.91**	78.47	94.29	88.86	92.18	88.45	70.23	92.84	88.69	88.76	85.13
AIN-F2O	84.11	94.91	92.85	95.58	91.86	78.71	**94.32**	88.75	**92.26**	88.51	71.16	**93.03**	88.80	**88.86**	**85.46**

Table 2: Averaged F1 score and accuracy on four tasks. **SF** represents the slot filling task. $\star$: For reference.

Conclusion

- We propose approximate inference networks that use Mean-Field Variational Inference instead of exact probabilistic inference algorithms such as the forward-backward and Viterbi algorithms for sequence labeling
- Empirical results show that AINs are significantly faster than traditional CRF and achieve competitive accuracy on 4 tasks with 15 datasets over three encoder types