Latent Constituents via Self-Attention

August 2, 2018

1 Introduction

2 Problem

We would like to learn a generative model over source sentences $\mathbf{x} = \{x_0, x_1, \ldots\}$, using a distribution over latent trees $\mathbf{z}$. (Shen et al., 2018) Yin et al. (2018)

3 PRPN (Shen et al., 2018)

The Parsing-Reading-Predict Network (Shen et al., 2018) is an approach to unsupervised tree induction that uses a soft approximation to a latent variable to a degree of success. The paper is inspired by the following hypothesis: given a binary tree, a token x_t only requires information up to a token x_{l_t} that satisfies either of the following conditions:

(a) its leftmost sibling

(b) or, if the token x_t is a leftmost child, its parent's left sibling's leftmost child.

Although the hypothesis itself is not tested and seems to only serve as inspiration, the model does see empirical success in the task of language modeling. The model is realized through the following insight: given a ranking of tokens $\mathbf{x} = \{x_0, \ldots, x_T\}$, recursively splitting $\mathbf{x}$ using the following procedure induces a binary tree where the first token to the left of token x_t that has higher rank, denoted x_{l_t}, also satisfies condition (a) or (b): given token i with the next highest rank, create a subtree $(x_{<i}, (x_i, x_{>i}))$ and recursively perform this procedure until only terminal nodes remain.

3.1 The Model

Let x_t be the current token, $\mathbf{x}_{0:t-1}$ all previous tokens, z_t the prediction for the current score, and $\tilde{\mathbf{z}}_{0:t-1} \in \mathbb{R}_+^{t-1}$ be all previous scores. The model takes the form a language model, where the distribution over the next token

$$p(x_t \mid \mathbf{x}_{0:t-1}, \tilde{\mathbf{z}}_{0:t-1}, z_t) = f(h_t)$$

is parameterized by either a linear projection or additional residual blocks applied to the output of modified LSTMN (LSTM Memory-Network) h_t. f is referred to as the Predict Network. The LSTMN, referred to as the Reading Network, is as follows: at each time step, the hidden and cell input are given by

$$\{\tilde{h}_t, \tilde{c}_t\} = \sum_{i=1}^{t-1} s_i^t \{h_i, c_i\},$$

a convex combination of the previous hidden and cell outputs. Then

$$h_t, c_t = \texttt{LSTM}(\tilde{h}_t, \tilde{c}_t).$$

The coefficients s_i^t are computed by first predicting a key

$$k = W_h h_{t-1} + W_x x_t$$

and querying all previous hidden states in order to compute the intermediate attention

$$\tilde{s_i^t} = \text{softmax}(\frac{h_i k_t^T}{\sqrt{\delta_k}}),$$

where δ_k is the dimension of the hidden state. We then incorporate a form of multiplicative gating g_i^t that is a function of the induced ranking of the tokens, which we described in the previous section. Finally,

$$s_i^t = \frac{g_i^t}{\sum_j g_j^t} \tilde{s_i^t},$$

so at every timestep t the self-attention is renormalized using the gates $g_{l_t:t-1}^t$.

3.2 Gating Self-attention

Recall that the gates g_i^t are used to both induce a latent tree and modulate self-attention (due to the previously stated insight that modulating self-attention using a ranking induces a binary tree). At every timestep for token x_t, we define a latent variable l_t which corresponds to the index of the token satisfying either condition (a) or (b). We then would have

$$g_i^t = \begin{cases} 1, & l_t \leq i < t \\ 0, & \text{otherwise.} \end{cases}$$

4 Model

4.1 Generative Model

Parameterize with $d_t \sim \mathcal{N}(\mu_t, \sigma_t)$. Reparameterize comparisons with gumbel softmax? Leave all self attentions as is?

- $p(z)$
- $p(x|z)$

5 Training and Inference

References

Yikang Shen, Zhouhan Lin, Chin wei Huang, and Aaron Courville. Neural language modeling by jointly learning syntax and lexicon. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=rkgOLb-0W`.

Pengcheng Yin, Chunting Zhou, Junxian He, and Graham Neubig. StructVAE: Tree-structured latent variable models for semi-supervised semantic parsing. In *The 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, Melbourne, Australia, July 2018. URL `https://arxiv.org/abs/1806.07832v1`.