TTIC 31230, Fundamentals of Deep Learning

David McAllester

Masked Language Modeling (MLM)

Gibbs Sampling

and Pseudo-Likelihood

Masked Language Models (MLMs)

BERT: Pre-training of Deep Bidirectional Transformers ...
Devlin et al., October 2018

Masked Language Models (MLMs)

Autoregressive text generation requires the words to be generated one at a time (sequentially).

MLM allows the words to be generated in parallel.

Parallel generation of novel text is very low quality.

However, parallel generation in machine translation can have comparable performance to autoregressive translation.

Parallel generation in translation can be faster than autoregressive translation.

Masked Language Models

Consider a probability distribution on a block of text.

$$y = (w_1, \ldots, w_T)$$

In BERT 15% of the words in a block of text are masked and the masked words are predicted from the unmasked words using a transformer model.

Pseudo-Likelihood

MLM is closely reated to Pseudo-Likelihood (1975) and Gibbs Sampling (1984).

For $y = (w_1, \ldots, w_T)$ define

$$y_{-i} = (w_1, \ldots, w_{i-1}, M, w_{i+1}, \ldots w_T)$$

where M is a fixed mask.

For a probability distribution P on strings we define the pseudo-liklihood $\tilde{P}$ by

$$\tilde{P}(y) = \prod_i P(w_i \,|y_{-i})$$

Pseudo-Likelihood

$$\tilde{P}(y) = \prod_i P(w_i \mid y_{-i})$$

Pseudo-likelihood is particularly relevant to training Markov random fields (graphical models).

But pseudo-likelihood corresponds to the objective function of MLMs with one mask per text block.

$$\begin{aligned}\Phi^* &= \operatorname*{argmin}_{\Phi} \; E_{y\sim\mathrm{Pop}} \; -\ln \tilde{P}_\Phi(y) \\ &= \operatorname*{argmin}_{\Phi} \; \sum_i \; E_{y\sim\mathrm{Pop}} \;\; -\ln P_\Phi(w_i|y_{-i})\end{aligned}$$

Pseudo-Likelihood

$$\Phi^* = \operatorname*{argmin}_{\Phi} \sum_i E_{y \sim \mathrm{Pop}} \quad - \ln P_\Phi(w_i | y_{-i})$$

Assuming universality we get

$$P_{\Phi^*}(w_i | y_{-i}) = \mathrm{Pop}(w_i \mid y_{-i})$$

Gibbs Sampling

$$P_{\Phi^*}(w_i|y_{-i}) = \mathrm{Pop}(w_i \mid y_{-i})$$

The ability to compute conditional probabilities does not immediately provide any way to compute $P_\Phi(y)$ or to sample y from $P_\Phi(y)$.

In principle sampling can be done with an MCMC process called Gibbs sampling.

Gibbs Sampling

Let $y[i \leftarrow w]$ be the word sequence resulting from replacing the ith word in the word sequence y by the word w.

Gibbs sampling is defined by stochastic state transition

$$\begin{aligned} y^{t+1} &= y^t[i \leftarrow w] \\ i &\sim \text{uniform on } \{1, \ldots, T\} \\ w &\sim P_\Phi(w_i \mid y_{-i}) \end{aligned}$$

Markov Processes

A Markov chain is an autoregressive probability distribution on infite sequences $s_1, s_2, s_3, \ldots$ defined by

$$\begin{aligned} P(s_1) &= P_0(s_1) \\ P(s_{t+1}|s_1, \ldots, s_t) &= P_T(s_{t+1}|s_t) \end{aligned}$$

Here we are interested in the case where s_t is is the translation sentence after t rounds of parallel updates.

This process defines a probability distribution $P_t(s)$ on sentences after t rounds of updates.

Markov Processes

For a distribution Q on states (sentences) define $P(Q)$ to be the distribution on sentences defined by

$$P(Q)(s) = P(s_{t+1} = s \mid s_t), \quad s_t \sim Q$$

A stationary distribution of a Markov process is a distribution Q (on sentences in this example) such that $P(Q) = Q$. for

Any Markov chain (defined by transition probabilities on states) that is "ergotic" in the sense that every state can reach every state has a unique stationary distribution.

Gibbs Sampling and Pseudo-Liklihood

Pseudo-liklihood defines a Gibbs Sampling Markov chain.

It is a theorem that if this Markov Chain is ergotic then its stationary distribution equals the populaiton distribution.

Markov Processes

If the conditional distributions allow any state (sentence) to reach any state then the conditional probabilities determine a unique distribution on strings with the given conditional probabilities.

Furtermore, we can in principle sample from this distribution by running the Gibbs Markov chain sufficiently long.

Gibbs Sampling

For langauge modeling Gibbs sampling mixes too slowly — it does not reach its stationary distribution in feasible time.

However, in the case of translation the distribution on y given x is lower entropy and Gibbs sampling seems practicle.

END