TTIC 31230, Fundamentals of Deep Learning

David McAllester, Autumn 2022

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

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.

The intuition is that MLMs achieve a better "understanding" of the robability distribution because each prediction looks at both the past and the future.

Masked Language Models (MLMs)

The BERT paper presents experiments showing that fine-tuning MLMs for downstream tasks (such as the the GLUE benchmark) outperformed fine-tuning of autoregressive models on those same tasks.

Masked Language Modeling

However, the more recent application of MLMs has been in machine translation.

The translation y of a given sentence x is generated by a word-parallel initialization followed by some number of rounds of word-parallel Gibbs Sampling.

This word-parallel sampling is faster on parallel hardware than auto-regressive sampling.

Pseudo-Likelihood

Here we will give a theoretical analysis of MLMs in terms of 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

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

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

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

Gibbs Sampling

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.

This implies that if the conditional distributions allow any state 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