TTIC 31230, Fundamentals of Deep Learning

David McAllester, Winter 2020

Language Modeling

Natural Language Understanding

GLUE: General Language Understanding Evaluation

ArXiv 1804.07461

BERT and GLUE

BERT and SuperGLUE

Language Modeling

The recent progress on NLP benchmarks is due to pretraining on language modeling.

Langauge modeling is based on unconditional cross-entropy minimiztion.

$$\Phi^* = \operatorname*{argmin}_{\Phi} \; E_{y \sim \mathrm{Pop}} \; - \ln P_{\Phi}(y)$$

In language modeling y is a sentence (or fixed length block of text).

Language Modeling

Let W be some finite vocabulary of tokens (words).

Let Pop be a population distribution over W^* (sentences).

We want to train a model $P_\Phi(y)$ for sentences y

$$\Phi^* = \operatorname*{argmin}_{\Phi} \; E_{y \sim \mathrm{Pop}} \; -\ln P_\Phi(y)$$

Autoregressive Models

A structured object, such as a sentence or an image, has an exponentially small probability.

An autoregressive model computes conditional probability for each part given "earlier" parts.

$$P_\Phi(w_0, w_1, \cdots, w_T) = \prod_{t=0}^{T} P_\Phi(w_t \mid w_1, \ldots, w_{t-1})$$

The End of Sequence Token `<EOS>`

We want to define a probability distribution over sentence of different length.

For this we require that each sentence is "terminated" with an end of sequence token `<EOS>`.

We requite $w_T = \texttt{<EOS>}$ and $w[t] \neq \texttt{<EOS>}$ for $t < T$.

This allows

$$P_\Phi(w_0, w_1, \cdots, w_T) = \prod_{t=0}^{T} P_\Phi(w_t \mid w_1, \ldots, w_{t-1})$$

To handle sequences of different length.

Standard Measures of Performance

Bits per Character: For character language models performance is measured in bits per character. Typical numbers are slightly over one bit per character.

Perplexity: It would be natural to measure word language models in bits per word. However, it is traditional to measure them in perplexity which is defined to be 2^b where b is bits per word. Perplexities of about 60 were typical until 2017.

According to Quora there are 4.79 letters per word. 1 bit per character (including space characters) gives a perplexity of $2^{5.79}$ or 55.3.

The State of the Art (SOTA)

As of March 2020 the state of the art neural language models yield perplexities of about 10.

END