TTIC 31230, Fundamentals of Deep Learning

David McAllester, Winter 2020

Language Modeling

Natural Language Understanding

GLUE: General Language Understanding Evaluation

ArXiv 1804.07461

Corpus	|Train|	|Test|	Task	Metrics	Domain
			Single-Sentence Tasks		
CoLA	8.5k	**1k**	acceptability	Matthews corr.	misc.
SST-2	67k	1.8k	sentiment	acc.	movie reviews
			Similarity and Paraphrase Tasks		
MRPC	3.7k	1.7k	paraphrase	acc./F1	news
STS-B	7k	1.4k	sentence similarity	Pearson/Spearman corr.	misc.
QQP	364k	**391k**	paraphrase	acc./F1	social QA questions
			Inference Tasks		
MNLI	393k	**20k**	NLI	matched acc./mismatched acc.	misc.
QNLI	105k	5.4k	QA/NLI	acc.	Wikipedia
RTE	2.5k	3k	NLI	acc.	news, Wikipedia
WNLI	634	**146**	coreference/NLI	acc.	fiction books

Table 1: Task descriptions and statistics. All tasks are single sentence or sentence pair classification, except STS-B, which is a regression task. MNLI has three classes; all other classification tasks have two. Test sets shown in bold use labels that have never been made public in any form.

BERT and GLUE

	Rank	Name	Model	URL	Score
	1	T5 Team - Google	T5		90.3
	2	ERNIE Team - Baidu	ERNIE		90.1
	3	Microsoft D365 AI & MSR AI & GATECH	MT-DNN-SMART		89.9
+	4	王玮	ALICE v2 large ensemble (Alibaba DAMO NLP)		89.7
+	5	Microsoft D365 AI & UMD	FreeLB-RoBERTa (ensemble)		88.4
	6	Junjie Yang	HIRE-RoBERTa		88.3
	7	Facebook AI	RoBERTa		88.1
+	8	Microsoft D365 AI & MSR AI	MT-DNN-ensemble		87.6
	9	GLUE Human Baselines	GLUE Human Baselines		87.1

BERT and SuperGLUE

	Rank	Name	Model	URL	Score
	1	SuperGLUE Human Baselines	SuperGLUE Human Baselines		89.8
+	2	T5 Team - Google	T5		89.3
	3	Zhuiyi Technology	RoBERTa-mtl-adv		85.7
	4	Facebook AI	RoBERTa		84.6
	5	IBM Research AI	BERT-mtl		73.5

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 $w_T =$ `<EOS>` and $w[t] \neq$ `<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