Generalized Language Models: Common Tasks & Datasets

April 24, 2019 by Lilian Weng

EDITOR'S NOTE: Generalized Language Models is an extensive four-part series by Lillian Weng of OpenAI.

- Part 1: CoVe, ELMo & Cross-View Training
- Part 2: ULMFiT & OpenAI GPT
- Part 3: BERT & OpenAI GPT-2
- Part 4: Common Tasks & Datasets

Do you find this in-depth technical education about language models and NLP applications to be useful? *Subscribe below to be updated when we release new relevant content*.

This article finalizes the series on generalized language models:

- Metric: Perplexity
- Common Tasks and Datasets
- Reference

Metric: Perplexity

Perplexity is often used as an intrinsic evaluation metric for gauging how well a language model can capture the real word distribution conditioned on the context.

A perplexity of a discrete proability distribution p is defined as the exponentiation of the entropy:

$$H(s) = -\sum_{i=1}^{N} P(w_i) \log_2 p(w_i) = -\sum_{i=1}^{N} \frac{1}{N} \log_2 p(w_i)$$

The perplexity for the sentence becomes:

$$2^{H(s)} = 2^{-\frac{1}{N}\sum_{i=1}^{N} \log_2 p(w_i)} = (2^{\sum_{i=1}^{N} \log_2 p(w_i)})^{-\frac{1}{N}} = (p(w_1) \dots p(w_N))^{-\frac{1}{N}}$$

A good language model should predict high word probabilities. Therefore, the smaller perplexity the better.

Common Tasks and Datasets

Question-Answering

- SQuAD (Stanford Question Answering Dataset): A reading comprehension dataset, consisting of questions posed on a set of Wikipedia articles, where the answer to every question is a span of text.
- RACE (ReAding Comprehension from Examinations): A large-scale reading comprehension dataset with more than 28,000 passages and nearly 100,000 questions. The dataset is collected from English examinations in China, which are designed for middle school and high school students.

Commonsense Reasoning

- Story Cloze Test: A commonsense reasoning framework for evaluating story understanding and generation. The test requires a system to choose the correct ending to multi-sentence stories from two options.
- SWAG (Situations With Adversarial Generations): multiple choices; contains 113k sentence-pair completion examples that evaluate grounded common-sense inference

Natural Language Inference (NLI): also known as **Text Entailment**, an exercise to discern in logic whether one sentence can be inferred from another.

- RTE (Recognizing Textual Entailment): A set of datasets initiated by text entailment challenges.
- SNLI (Stanford Natural Language Inference): A collection of 570k human-written English sentence pairs manually labeled for balanced classification with the labels `entailment`, `contradiction`, and `neutral`.
- MNLI (Multi-Genre NLI): Similar to SNLI, but with a more diverse variety of text styles and topics, collected from transcribed speech, popular fiction, and government reports.
- QNLI (Question NLI): Converted from SQuAD dataset to be a binary classification task over pairs of (question, sentence).
- SciTail: An entailment dataset created from multiple-choice science exams and web sentences.

Named Entity Recognition (NER): labels sequences of words in a text which are the names of things, such as person and company names, or gene and protein names

- CoNLL 2003 NER task: consists of newswire from the Reuters, concentrating on four types of named entities: persons, locations, organizations and names of miscellaneous entities.
- OntoNotes 0.5: This corpus contains text in English, Arabic and Chinese, tagged with four different entity types (PER, LOC, ORG, MISC).
- Reuters Corpus: A large collection of Reuters News stories.
- Fine-Grained NER (FGN)

Sentiment Analysis

- SST (Stanford Sentiment Treebank)
- IMDb: A large dataset of movie reviews with binary sentiment classification labels.

- CoNLL-2004 & CoNLL-2005

Sentence similarity: also known as *paraphrase detection*

- MRPC (MicRosoft Paraphrase Corpus): It contains pairs of sentences extracted from news sources on the web, with annotations indicating whether each pair is semantically equivalent.
- QQP (Quora Question Pairs) STS Benchmark: Semantic Textual Similarity

Sentence Acceptability: a task to annotate sentences for grammatical acceptability.

- CoLA (Corpus of Linguistic Acceptability): a binary single-sentence classification task.

Text Chunking: To divide a text in syntactically correlated parts of words.

- CoNLL-2000

Part-of-Speech (POS) Tagging: tag parts of speech to each token, such as noun, verb, adjective, etc. the Wall Street Journal portion of the Penn Treebank (Marcus et al., 1993).

Machine Translation: See Standard NLP page.

- WMT 2015 English-Czech data (Large)
- WMT 2014 English-German data (Medium)
- IWSLT 2015 English-Vietnamese data (Small)

Coreference Resolution: cluster mentions in text that refer to the same underlying real world entities.

- CoNLL-2012

Long-range Dependency

- LAMBADA (LAnguage Modeling Broadened to Account for Discourse Aspects): A collection of narrative passages extracted from the BookCorpus and the task is to predict the last word, which require at least 50 tokens of context for a human to successfully predict.
- Children's Book Test: is built from books that are freely available in Project Gutenberg. The task is to predict the missing word among 10 candidates.

Multi-task benchmark

- GLUE multi-task benchmark: https://gluebenchmark.com
- decaNLP benmark: https://decanlp.com

Unsupervised pretraining dataset

- Books corpus: The corpus contains "over 7,000 unique unpublished books from a variety of genres including Adventure, Fantasy, and Romance."
- 1B Word Language Model Benchmark
- English Wikipedia: ~2500M words

[2] Kevin Clark et al. "Semi-Supervised Sequence Modeling with Cross-View Training." EMNLP 2018.

[3] Matthew E. Peters, et al. "Deep contextualized word representations." NAACL-HLT 2017.

[4] OpenAI Blog "Improving Language Understanding with Unsupervised Learning", June 11, 2018.

[5] OpenAI Blog "Better Language Models and Their Implications." Feb 14, 2019.

[6] Jeremy Howard and Sebastian Ruder. "Universal language model fine-tuning for text classification." ACL 2018.

[7] Alec Radford et al. "Improving Language Understanding by Generative Pre-Training". OpenAI Blog, June 11, 2018.

[8] Jacob Devlin, et al. "BERT: Pre-training of deep bidirectional transformers for language understanding." arXiv:1810.04805 (2018).

[9] Mike Schuster, and Kaisuke Nakajima. "Japanese and Korean voice search." ICASSP. 2012.

[10] Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation

[11] Ashish Vaswani, et al. "Attention is all you need." NIPS 2017.

[12] Peter J. Liu, et al. "Generating wikipedia by summarizing long sequences." ICLR 2018.

[13] Sebastian Ruder. "10 Exciting Ideas of 2018 in NLP" Dec 2018.

[14] Alec Radford, et al. "Language Models are Unsupervised Multitask Learners.". 2019.

[15] Rico Sennrich, et al. "Neural machine translation of rare words with subword units." arXiv preprint arXiv:1508.07909. 2015.

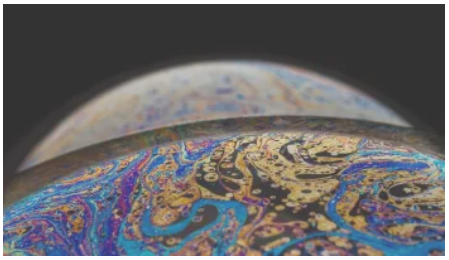
This article was originally published on Lil'Log and re-published to TOPBOTS with permission from the author.

- [ ] Chatbots & Conversational AI
- [ ] Computer Vision
- [ ] Ethics & Safety
- [ ] Robotics
- [ ] Machine Learning
- [ ] Deep Learning
- [ ] Reinforcement Learning
- [ ] Generative Models
- [ ] Other (Please Describe Below)

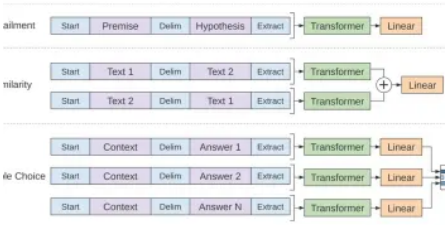
What is your biggest challenge with AI research? *

SUBMIT

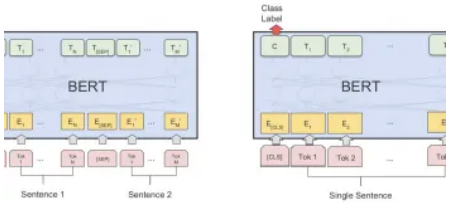
Related



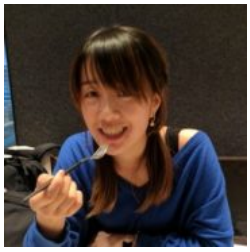
The Relationship Between Perplexity And Entropy In NLP



Generalized Language Models: ULMFiT & OpenAI GPT



Generalized Language Models: BERT & OpenAI GPT-2



About Lilian Weng

Lilian Weng is on the Robotics team at OpenAI. She writes code, reads papers, does research on deep learning models, and works on physical machines.

Leave a Reply

Name

Email

Website

POST COMMENT

POPULAR ARTICLES

NeurIPS 2021 – 10 Papers You Shouldn't Miss

A Guide To Knowledge Graphs

Why Graph Theory Is Cooler Than You Thought

10 Leading Language Models For NLP In 2021

BERT Inner Workings

Pretrain Transformers Models in PyTorch Using Hugging Face Transformers

More Articles

TOPICS

Bots

Brands

Business

China

Commerce

Computer Vision

Conversational AI

Customer Service

Cybersecurity