Preliminary Survey on Foundation Language Models

Vyom Pathak

Introduction

- Learning representation of natural language from data is an open problem
- Statistical NLP models depend on engineered features to learn representations
- Neural language models use dense representation capturing various aspects of text
- Word embeddings are shallow models that lack contextual information
- Contextual word embeddings are deep models that can be adapted to various tasks
- Deep NNs overfit on low-resource datasets
- Pre-training language models (PTLM) on large text corpus can improve downstream tasks
- Foundation models or large language models (LLMs) are the latest trend in NLP using pre-training, and an aspect of fine-tuning for target task
- We present a preliminary survey on language models to connect all these models
- Criterion:
  - Differ based on how they learn representation from text
  - Vary in-terms of size, and complexity based on the data, and architecture
  - Addressing different research questions, and diverse tasks in NLP
  - Applications in real-world scenarios

Background - Language Representation Learning

- Language representation learning aims to capture the meaning of text using low-dimensional vectors
- There are two types of embeddings:
  - Non-contextual embeddings (shallow models)
  - Contextual embeddings (deep models)
- Non-contextual embeddings are static and have limitations such as handling out-of-vocabulary words
  - Representations are directly feed to train a model for target task (encoder-only)
- Contextual embeddings are dynamic and capture the context of the word
  - Representations are adapted to target tasks depending on training framework, scale, and architecture (encoder-only, decoder-only, encoder-decoder)
- Contextual embeddings can be further classified into following:
  - Sequential models
  - Non-sequential models
- Sequential models use convolutional or recurrent networks to capture the local context in order
- Non-sequential models use tree or graph structures to capture syntactic or semantic relations
- Self-attention mechanism is a non-sequential model that learns the connection weights dynamically

Background - Training Frameworks

- Large language models require large amounts of labeled data, which are costly and scarce
- Pre-training on large unlabeled corpora can improve model initialization, generalization, and regularization
- Pre-trained learning framework can be used as:
  - Non-contextual models (shallow) such as NNLM, CBOW, Skip-Gram, GLoVe
  - Contextual models (deep) such as LSTMs, ELMo, CoVe, biLM, ULMFiT, modern LLMs (BERT, GPTs)
- Pre-training tasks, and architectures have evolved learning representations from shallow to deep language models

Background - Pre-training tasks

- Pre-training tasks can be categorized into three broad types:
  - Supervised learning - Learning with labeled data
  - Unsupervised learning - Learning with unlabeled data (learning distribution)
  - Self-supervised learning - Learning with unlabeled data by generating labels (masked language modeling)
- For NLP, datasets of most supervised tasks are not large enough to train good pre-trained models.
- Unsupervised learning tasks include probabilistic language modeling (LM), bidirectional LM,
- Masked LM (MLM), and seq2seq MLM can be considered as semi-supervised
- Enhanced MLM tasks include dynamic masking, UniLM, translation LM etc.
- Permuted LM is a self-supervised task that uses random permutations of input sequences to generate the original sequence
- Next Sentence Prediction (NSP) is a sentence completion task that predicts whether two sentences are continuous or not
- Denoising Autoencoder (DAE) is an encoder-decoder task that recovers the original sequence from a partially corrupted input

Background - Adaptation to downstream task

- Effective adaptation of pre-trained models to downstream tasks is a challenging task
- Transfer learning is a common adaptation method that uses pre-trained models for different tasks
  - Choosing appropriate pre-training tasks, model architectures, and prospective corpus
  - Involves selecting which layers of the model to use for the downstream task
  - Deciding whether to tune or not to tune the pre-trained models
- Feature extraction (freezed encoders) and not tune the model
  - Requires more complex task specific layers
- Tune the model by performing fine-tuning
- Fine-tuning methods include:
  - Two-stage tuning - finetune on unlabeled task data, and fine-tuned on target task
  - Multi-task fine-tuning - finetune on multiple tasks at once
  - Model distillation - Using fewer layers with general information, and fine-tune them
  - Gradual unfreezing - Freezing some layers and fine-tuning, and unfreezing gradually
  - Planned sequential unfreezing - Unfreezing groups of layers based on representations
- Fine-tuning is fragile and requires careful hyper-parameter tuning

Background - Adaptation to downstream task

- Prompt-based technique as a tuning method
- Prompt-based methods can be discrete or continuous
- Discrete prompts are sequences of words inserted into the input text to help the pre-trained model converge faster
  - Manually or automatically generated
- Continuous prompts - Words that are combined with word-type embeddings
  - Outperform discrete prompts on relation-oriented tasks
- Prompt-based methods can achieve similar performance to fine-tuning as the model size scales up

Background - Task Capability

- Classical NLP tasks include
  - Question answering - Answering questions based on a given text or knowledge base
    - Can be one-shot or multi-round, extractive or generative, single-hop or multi-hop
  - Sentiment analysis - Detecting the polarity or emotion of a text
    - Use datasets such as SST-2
  - Machine translation - Translating text from one language to another
    - Use datasets such as WMT
  - Information extraction - Extracting structured information from unstructured text
    - Includes tasks such as named entity recognition, relation extraction, event extraction, etc.
  - Summarization - Generating shorter text that preserves the meaning of the longer text
    - Can be extractive or abstractive, single-document or multi-document
- Large-scale benchmarks are required for evaluation of LLMs on these tasks
  - GLUE
  - SuperGLUE

Catalogue

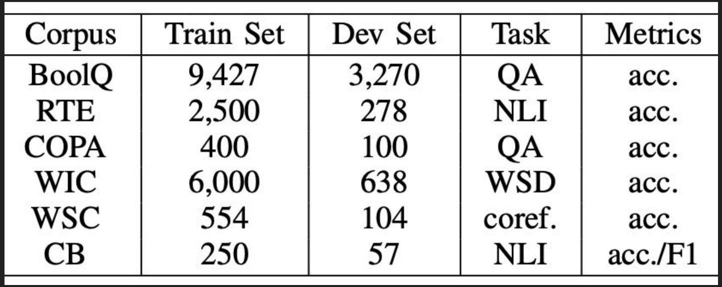
Model	Architecture	Self-Attention	Pre-Training Tasks	Pre-Training Corpus	Parameters (M - Million, B - Billion)	Applications
RoBERTa	Encoder-Only Transformer	Bi-directional	MLM with dynamic masking	BooksCorpus, English Wikipedia, CC News, OpenWebText, Stories	125M 355M	NLU and QA
DeBERTa	Encoder-Only Transformer	Disentangled attention mechanism	MLM with dynamic masking	BooksCorpus, English Wikipedia, and RealNews	144M 350M 700M	NLU, QA, NLI, and SA
GPT-2	Decoder-Only Transformer	Uni-directional Attention	MLM	WebText, Text with high Reddit karma scores	117M 355M 762M	NLG, TS, MT, TC, and Finetuned for NLU
Transformer-XL	Decoder-Only Transformer with segment-level recurrence, and relative positional encoding	Relative positioned Uni-directional attention mechanism	MLM	Wikitext-103	355M	NLG, TS, and Finetuned for NLU

Catalogue

Model	Architecture	Self-Attention	Pre-Training Tasks	Pre-Training Corpus	Parameters (M - Million, B - Billion)	Applications
Bart	Transformer with BERT as encoder, and GPT as decoder	Bidirectional self-attention for encoder, and uni-directional self-attention for decoder	Denoising Autoencoder with Span Corruption	BooksCorpus, English Wikipedia, CC News, OpenWebText, Stories	10% bigger than BERT (355M)	NLG, NLU - TC, MT, TC
T5	Transformer with relative positional encoding, and text-to-text format	Scaled up transformer style self-attentions	Denoising Autoencoder with Span Corruption	C4	60M 220M 770M 3B 11B	NLG based MT, QA, AS, and TC

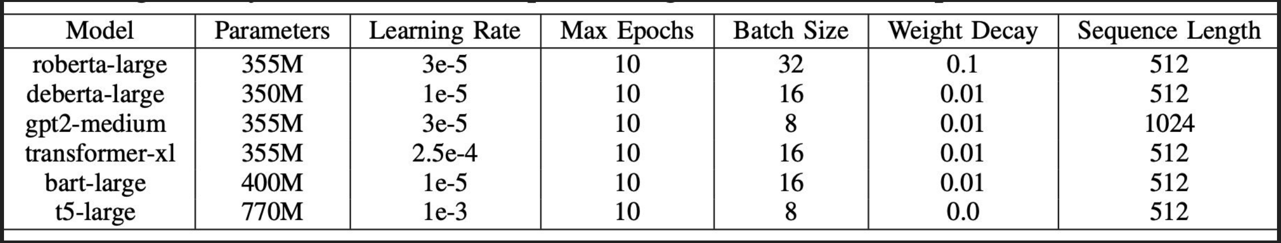
Experiment Setup - Datasets

- BoolQ - A question answering task that requires answering yes/no questions based on a passage from Wikipedia
- RTE - A natural language inference task that requires determining whether a pair of premises and hypotheses entail each other or not
- COPA - A causal reasoning task that requires identifying the cause or effect of a given premise from two choices
- WIC - A word sense disambiguation task that requires determining whether a polysemous word is used in the same sense in two sentences
- WSC - A coreference resolution task that requires commonsense reasoning to identify the antecedent of a pronoun in a sentence
- CB - A textual entailment task that requires classifying the degree of belief of an embedded clause in a short text

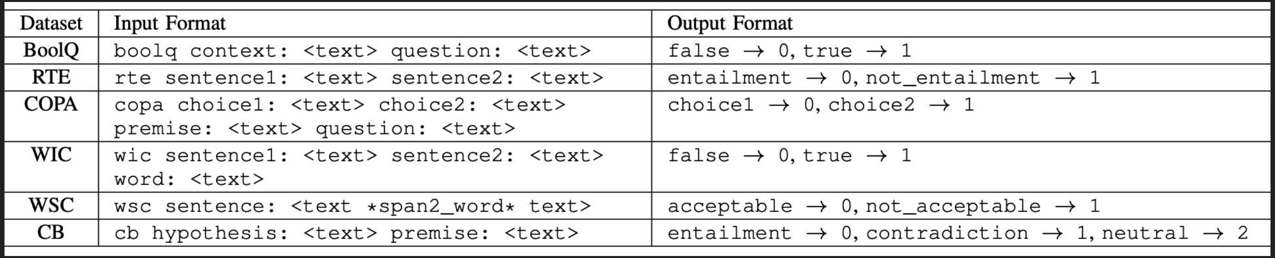


Corpus	Train Set	Dev Set	Task	Metrics
BoolQ	9,427	3,270	QA	acc.
RTE	2,500	278	NLI	acc.
COPA	400	100	QA	acc.
WIC	6,000	638	WSD	acc.
WSC	554	104	coref.	acc.
CB	250	57	NLI	acc./F1

Experiment Setup - Models

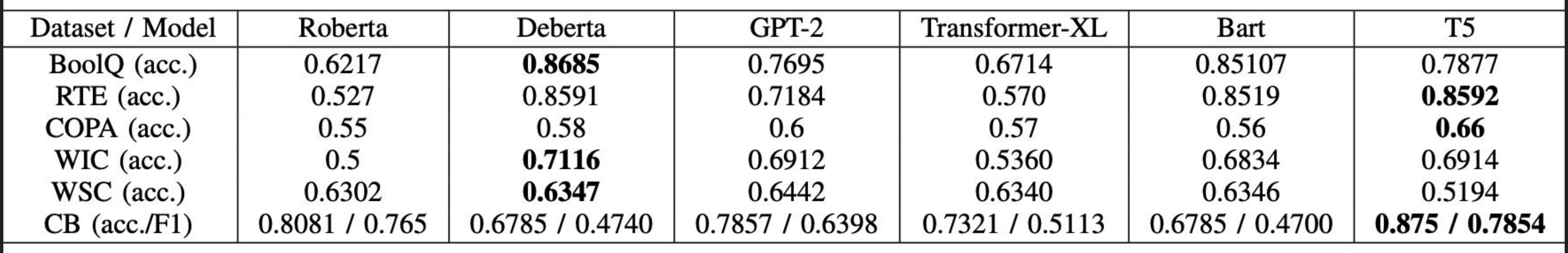


Model	Parameters	Learning Rate	Max Epochs	Batch Size	Weight Decay	Sequence Length
roberta-large	355M	3e-5	10	32	0.1	512
deberta-large	350M	1e-5	10	16	0.01	512
gpt2-medium	355M	3e-5	10	8	0.01	1024
transformer-xl	355M	2.5e-4	10	16	0.01	512
bart-large	400M	1e-5	10	16	0.01	512
t5-large	770M	1e-3	10	8	0.0	512



Dataset	Input Format	Output Format
BoolQ	`boolq context: <text> question: <text>`	`false` → 0, `true` → 1
RTE	`rte sentence1: <text> sentence2: <text>`	`entailment` → 0, `not_entailment` → 1
COPA	`copa choice1: <text> choice2: <text> premise: <text> question: <text>`	`choice1` → 0, `choice2` → 1
WIC	`wic sentence1: <text> sentence2: <text> word: <text>`	`false` → 0, `true` → 1
WSC	`wsc sentence: <text *span2_word* text>`	`acceptable` → 0, `not_acceptable` → 1
CB	`cb hypothesis: <text> premise: <text>`	`entailment` → 0, `contradiction` → 1, `neutral` → 2

Results



Dataset / Model	Roberta	Deberta	GPT-2	Transformer-XL	Bart	T5
BoolQ (acc.)	0.6217	**0.8685**	0.7695	0.6714	0.85107	0.7877
RTE (acc.)	0.527	0.8591	0.7184	0.570	0.8519	**0.8592**
COPA (acc.)	0.55	0.58	0.6	0.57	0.56	**0.66**
WIC (acc.)	0.5	**0.7116**	0.6912	0.5360	0.6834	0.6914
WSC (acc.)	0.6302	**0.6347**	0.6442	0.6340	0.6346	0.5194
CB (acc./F1)	0.8081 / 0.765	0.6785 / 0.4740	0.7857 / 0.6398	0.7321 / 0.5113	0.6785 / 0.4700	**0.875 / 0.7854**

- Deberta models perform best on BoolQ, WIC, and WSC, which require robust representation for low-resource tasks
- T5 model performs best on RTE, COPA, and CB, which are entailment tasks that require natural language inference
- Decoder-only models perform poorly because they lack proper encoder representations
- Encoder-decoder models perform better than encoder-only and decoder-only models because they can learn and generate representations
- COPA and WSC are the most difficult tasks because of the low resource dataset and the task difficulty

Future Work

- Experimenting with more models and datasets, such as Big-bench
- Exploring different training frameworks, such as in-context learning, UL2, Megatron, instruction fine-tuning, etc.
- Studying the emergent abilities, interpretability, and reliability of pre-trained models
- Regulating the models to avoid ethical and social issues, such as bias, contamination, privacy, etc.
- Compressing the models to serve a large user base with minimum latency

Conclusion

- A systematic survey on language models that connects various models based on multiple criteria
- Examines different types of language models and analyzes their pre-training tasks, training frameworks, adaptation methods, and evaluation benchmarks
- Performs experiments on SuperGLUE tasks using six representative models and comparing their performance and efficiency
- Reveals that different models have different strengths and weaknesses depending on the task and the data
- Proposes some future directions to enhance the robustness and comprehensiveness of this survey

References

- Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. Pre-trained Models for Natural Language Processing: A Survey. Science China Technological Sciences, 63(10):1872–1897, October 2020. arXiv:2003.08271 [cs].
- Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. A Survey for In-context Learning, December 2022. arXiv:2301.00234 [cs]
- Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. In Advances in Neural Information Processing Systems, volume 32. Curran Associates, Inc., 2019
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach, July 2019. arXiv:1907.11692 [cs].
- Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. DeBERTa: Decoding-enhanced BERT with Disentangled Attention,October 2021. arXiv:2006.03654 [cs].
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. OpenAI blog, 1(8):9, 2019.
- Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V. Le, and Ruslan Salakhutdinov. Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context, June 2019. arXiv:1901.02860 [cs, stat].
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension, October 2019. arXiv:1910.13461 [cs, stat].
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research, 21(1):5485–5551, 2020.
- Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent Abilities of Large Language Models, October 2022. arXiv:2206.07682 [cs].