1. **BERT (Bidirectional Encoder Representations from Transformers)**:- **2019**

**[Props]** Google,masked language models, Transformer Enc.

**[Reference]** J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” arXiv preprint arXiv:1810.04805, 2018.

**[Dataset]** WikiEn+BookCorpus

1. **XLNet (Generalized autoregressive pretraining):- 2019**

**[Props]** PLM, Transformer Enc.

**[Reference]** Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. XLNet: Generalized autoregressive pretraining for language understanding. In NeurIPS, pages 5754–5764, 2019.

**[Dataset]** WikiEn+ BookCorpus+Giga5+ClueWeb+Common Crawl

1. **RoBERTa (A robustly optimized BERT pretraining approach):-2019**

**[Props]** Facebook, Enhanced masked language models, Transformer Enc.

**[Reference]** 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 pretraining approach. arXiv preprint arXiv:1907.11692, 2019.

**[Dataset]** BookCorpus+CCNews+OpenWebText+STORIES

1. **ALBERT (A Lite BERT):- 2020 ICLR**

**[Props]**

**[Reference]** Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). Albert: A lite bert for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942.

**[Dataset]** Wikipedia + BookCorpus

1. **DistilBERT (A distilled version of BERT):- 2020, EMC^2: Co-located with NeurIPS’19**

**[Props]**

**[References]** Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.

**[Dataset]** Wikipedia + BookCorpus

1. **OpenAI GPT-3 (Generative Pre-training): - 2020**

**[Props]** autoregressive language models,Transformer Dec.

**[Reference]** Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, AdityaRamesh, DanielM.Ziegler, Jerey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are fewshot learners. In Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, 2020

**[Dataset]** CommonCrawl +WebText **+** books corpora + English-language Wikipedia

1. **T5 (Text-to-Text Transfer Transformer):- 2020**

**[Props]** encoder-decoder models

**[Reference]** Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. In Journal of Machine Learning Research.

**[Dataset]** Colossal Clean Crawled Corpus (C4)

1. **Google ELECTRA:- 2020 ICLR**

**[Props]**

**[Reference]** Clark, K., Luong, M. T., Le, Q. V., & Manning, C. D. (2020). Electra: Pre-training text encoders as discriminators rather than generators. arXiv preprint arXiv:2003.10555.

**[Dataset]** WikiEn+BookCorpus (for most of the experiment ) & WikiEn+ BookCorpus+Giga5+ClueWeb+Common Crawl (for large model)

1. **Facebook Muppet (Massive Multi-task Representations with Pre-Finetuning):- 2021 EMLP**

**[Props]**

**[Reference]** Aghajanyan, A., Gupta, A., Shrivastava, A., Chen, X., Zettlemoyer, L., & Gupta, S. (2021). Muppet: Massive Multi-task Representations with Pre-Finetuning. arXiv preprint arXiv:2101.11038.

**[Dataset]**

**Corpus and links:**

|  |  |
| --- | --- |
| **Corpus** | **Link** |
| **Transformer or Neural LM** | |
| WikiEn | <https://dumps.wikimedia.org/enwiki/> |
| Colossal Clean Crawled Corpus -C4 : LM T5 | <https://www.tensorflow.org/datasets/catalog/c4> |
| BooksCorpus (800M words) (16gb) | 1. <https://huggingface.co/datasets/bookcorpusopen#source-data> 2. <https://huggingface.co/datasets/bookcorpus> |
| ClueWeb 2012-B | <https://lemurproject.org/clueweb12/index.php> (not free) |
| Common Crawl | <http://commoncrawl.org/> |
| CC-NEWS (which we collected from the English portion of the CommonCrawl News dataset) | <https://huggingface.co/datasets/cc_news> |
| OPENWEBTEXT Corpus | <https://skylion007.github.io/OpenWebTextCorpus/>  <https://github.com/jcpeterson/openwebtext>  <https://openwebtext2.readthedocs.io/en/latest/background/> |
| WikiText-2 and WikiText-103  by Pointer Sentinel Mixture Models, Stephen Merity | <https://www.salesforce.com/products/einstein/ai-research/the-wikitext-dependency-language-modeling-dataset/> |
| **statistical language modelling** | |
| One billion word benchmark for measuring progress in **statistical language modelling** | <https://github.com/ciprian-chelba/1-billion-word-language-modeling-benchmark> |