COMPARATIVE ANALYSIS OF TRANSFORMER BASED LANGUAGE MODELS

Aman Pathak

Department of Computer Science Engineering,
Medi-Caps University, Indore, India

ABSTRACT

Natural language processing (NLP) has witnessed many substantial advancements in the past three years. With the introduction of the Transformer and self-attention mechanism, language models are now able to learn better representations of the natural language. These attention-based models have achieved exceptional state-of-the-art results on various NLP benchmarks. One of the contributing factors is the growing use of transfer learning. Models are pre-trained on unsupervised objectives using rich datasets that develop fundamental natural language abilities that are fine-tuned further on supervised data for downstream tasks. Surprisingly, current researches have led to a novel era of powerful models that no longer require fine-tuning. The objective of this paper is to present a comparative analysis of some of the most influential language models. The benchmarks of the study are problem-solving methodologies, model architecture, compute power, standard NLP benchmark accuracies and shortcomings.

KEYWORDS

Natural Language Processing, Transformers, Attention-Based Models, Representation Learning, Transfer Learning.

1. INTRODUCTION

Over the past few years, there has been rapid progress in the field of language modeling. The new generation of NLP models introduced from late 2018 has drastic performance improvements on many language understanding benchmarks, with few achieving near-human level accuracies. This significant rate of progress can also be comprehended from the need for more rigorous benchmarks as these models have outperformed on the existing benchmarks. The shift from GLUE [1] to SuperGLUE [2] is one of the prominent examples. The two major contributors behind this success are Transformer and Transfer learning.

Before the introduction of the Transformer [3], Recurrent Neural Networks were the state-of-the-art solution for any NLP tasks. Later, for many years LSTM [4] became the go-to architecture for developing sequence to sequence models. The introduction of gating mechanism and attention mechanisms successfully mitigated the vanishing gradients problem and enhanced the performance on longer sequence length. Nevertheless, these modifications could not resolve the shortcoming posed by its sequential nature, i.e., it inhibited parallelization within training examples. Transformer embarked a new beginning by introducing parallelization which facilitated encoding all parts of the input sequence together at the same time. This considerably reduced the time to train the model. The success of the Transformer model comes from the self-attention mechanism that understands the essential elements in a sentence by evaluating the relationships between all words. A few months later, GPT [5], one of the early implementations

David C. Wyld et al. (Eds): AIAP, SIGML, CNSA, NIAI - 2021
pp. 165-176, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.110111

of the transformer-based language model, successfully demonstrated the effectiveness of using transformers by outperforming several task-specific state-of-the-art models. This overwhelming success is predominantly attributed to the adaptation of Transformers. Consequently, Transformers became a conventional preference for many upcoming language models.

In the last few years, NLP has witnessed a rise in the adaptation of transfer learning. It has been vastly explored in the field of computer vision, where a Convolutional Neural Network (CNN) is pre-trained on a rich image dataset such as ImageNet [6] followed by fine-tuning on task-specific data. Similarly, in NLP, pre-training a language model on considerably large unlabeled text corpora has become the new standard approach. Rather than randomly initializing the model parameters, it helps the model to learn fundamental language abilities that can be valuable in any NLP task. Enormous text repositories like Wikipedia, books, articles, social media platforms, etc., act as the primary source of plain text. Transfer learning has been helpful to many NLP tasks that have an inadequate amount of training data. It would have been nearly impossible to reach the remarkable state-of-the-art language models without transfer learning.

The paper covers a detailed comparative analysis among various Transformer based Language models. It is divided into the following sections. Section II is a literature review. This section describes a broad summary of the proposed language models and their significant contributions. Section III presents the inferences drawn from the model. Section IV provides a detailed comparative analysis. Section V concludes the paper with insights from this analysis.

2. LITERATURE REVIEW

Jacob Devlin et al. (Oct 2018) [7] proposed BERT, which stands for Bidirectional Representations from Transformers. It marks the beginning of one of the most influential concepts that changed the entire NLP scenario, a pre-trained deep bidirectional model. The term bidirectional emphasizes BERT's context-sensitive word embedding generated by combining the left and right contexts. To train such representation, the model uses a masked language model (MLM), which is inspired by the cloze task [8]. It also utilizes next sentence prediction (NSP), a second pre-training task along with MLM that builds a better understanding of the relationship between the sentences. To sum up, BERT attains new state-of-the-art results on eleven NLP tasks and establishes a firm foundation for the upcoming language models.

Zihang Dai et al. (Jan 2019) [9] proposed Transformer-XL, an improved transformer architecture capable of learning dependencies beyond the maximum sequence length T, or a fixed-length segment. To achieve this, the model suggests two changes in the original transformer architecture. It introduces recurrence between segment levels and a novel relative positional encoding. Transformer-XL learns dependencies that are 450% longer and attains 1874x speed up during evaluation than the transformer.

Zhilin Yang et al. (June 2019) [10] proposed XLNet, a generalized autoregressive pretraining method that uses permutation language modeling to enable an autoregressive model to learn bidirectional contexts. The bidirectional nature comes from the permutation of the factorization order, which allows it to see tokens occurring to the left and right in the same sequence. This makes XLNet perform better on language understanding tasks like question answering, reading comprehensions, etc.

Yinhan Liu et al. (July 2019) [11] proposed RoBERTa, which stands for Robustly optimized BERT approach. RoBERTa enhances the BERT model by providing a set of modifications in its pre-training implementation, some of which are inspired by the previous works in neural machine translation and are proven to enhance end-task performances. Experiments on GLUE [1],

SQUAD v1.1/v2.0 [12], RACE [13], reveals that RoBERTa outperforms BERT and establishes new state-of-the-art results.

Zhenzhong Lan et al. (Sept 2019) [14] proposed ALBERT, A Lite BERT, that achieves a new state-of-the-art solution on RACE, GLUE and SQUAD benchmarks. Although ALBERT-large (comparable to BERT-large) falls short in performance, it is 18x smaller and trains 1.7x faster. It disentangles the memory limitation and communication overhead that are inevitable when pre-trained models are scaled. To accomplish this, the model introduces two modifications in BERT, factorized embedding parameterization and cross-layer parameter sharing. Along with this, Sentence order prediction (SOP), a novel pre-training objective, is introduced that replaces NSP. These developments lead ALBERT to scale to ALBERT-xxlarge, a much larger model that outperforms BERT with 70% fewer parameters.

Colin Raffel et al (Oct 2019) [15], proposed T5, which stands for Text-to-Text Transfer Transformer. It is the result of combining the best design traits from past work in Language Modelling. The model introduces a unified text-to-text framework where NLP tasks are treated as a text-to-text problem. This novel framework enables T5 to solve diverse downstream tasks using a single pre-trained model, as opposed to fine-tuned task-specific models. T5 achieved the state-of-the-art solution on GLUE, CNN/Daily Mail [16] with achieving near-human score on SuperGLUE.

Jingqing Zhang et al. (Dec 2019) [17] proposed PEGASUS, a specialized pre-training model for abstractive text summarization. It is pre-trained on a novel self-supervised objective, Gap Sentences Generation (GSG), that bears a striking resemblance to the targeted downstream task. As a result, PEGASUS provides faster fine-tuning and greater performance on low resource summarization, i.e., accomplishing state-of-the-art results on six datasets with just 1000 examples.

Kevin Clark et al. (Mar 2020) [18] proposed ELECTRA, a two transformer text encoder (generator and discriminator) model that is pre-trained by distinguishing the real tokens in the original sentence. ELECTRA emphasizes stronger compute efficient models. Based on this vision, 12.5% trained and 50% trained ELECTRA-Small efficiently outsmarts BERT-Small and GPT, respectively. Replaced token detection is remarkably computationally efficient than MLM. Although the model architecture bears a similarity with Generative Adversarial Training (GAN) [19], the paper briefly explains the difference in their implementation and why it is an instance of Adversarial Contrastive Estimation [20] instead of GAN.

Alec Radford et al. (June 2020) [21] proposed GPT-3, the third- generation autoregressive language model in the GPT series. The model consists of 175B parameters, the largest in any model till date and is 100x bigger in comparison to GPT-2 [22]. The model implements few-shot learning, i.e., capable of performing a novel NLP task by looking at a limited number of examples/instructions. That is, the model does not update any parameter-weights and can be applied directly on any NLP tasks. In most cases, this eliminates the need to fine-tune the model further, which can be challenging at times due to the unavailability of domain-specific data. GPT-3 is task-agnostic and surpasses the previous state-of-the-art fine-tuned model in few downstream tasks.

3. INFERENCES DRAWN

3.1. BERT

Although BERT outperformed on eleven NLP tasks, its pre-training phrase can be optimized further. It is pre-trained on two unsupervised tasks one of which is MLM. It trains the model to utilize bidirectionality by predicting a few masked tokens in a sentence. The masking strategy creates a mismatch in the fine-tuning tasks as [MASK] tokens never occur in fine-tuning data until created externally. To soften this, the [MASK] token was used only 80% of the time. Also, in sentences with more than one masked token, the model treats them independent of each other, i.e., the second masked token is assumed to be semantically independent of the first masked token. This assumption hurts the prediction in cases of high dependence between the tokens. The second pre-training objective, NSP, improves the performance on Natural Language Interface and Question Answering tasks.

3.2. Transformer-XL

The vanilla transformer model did not emphasize retaining contexts between fragments of segments exceeding the maximum sequence length (T=512 tokens).In other words, the information flow between segments of the same documents was not captured, causing context fragmentation. To deal with such limitation, Transformer-XL proposed a segment-level recurrence mechanism that enabled reusing the hidden layer of the previous segment. This ensured a proper flow of information, thus accurately capturing long term dependencies. Absolute positional embeddings (used in the vanilla model) had to be replaced by the relative positional encodings as it was incompatible with segment recurrence. Such encoding enabled self-attention layers to figure out the relative distance between tokens and where to attend them.

3.3. XLNet

Inspired by recent successes in bidirectional context representation and segment recurrence, XLNet integrates both into a generalized autoregressive pre-trained model. Using the permutation LM pretraining objective, samples of different orders of the factorization are obtained using different attention masks. This permutation allowed such orders that required contexts of both the left and the right tokens for predicting a token, equipping a bidirectional behavior. To facilitate such behavior, XLNet's Two-Stream self-attention architecture produces two kinds of token representations instead of one, i.e., content and query representation. The content representation is the same as in the vanilla transformer and is used when predicting other tokens. On the other hand, the novel query representation is used when the token itself is being predicted. It includes contextual information of tokens occurring before the current token in the order and the positional information of the token itself.

3.4. RoBERTa

This paper focuses on an in-depth analysis of BERT and finding new alternatives to improve performance. As a result, new modifications in the pre-training were discovered. Following those, RoBERTa is pre-trained on a dataset that is 10x larger, a larger vocabulary, and larger mini-batch sizes. The dynamic masking scheme used in MLM ensures that different masked versions of the same sentence occur across different epochs, leading to slightly better pre-training. RoBERTa excludes NSP, as experiments prove that it no longer remains critical with the proposed set of modifications, and instead, eliminating it improves performances on downstream NLI tasks such as SQuAD 1.1/2.0, MNLI-m [23], RACE, etc.

3.5. ALBERT

Factorization of embedding parameters decomposes the word embedding size E and hidden layer size H. One-hot vectors are projected firstly into the context-independent word embedding space and then into context-dependent hidden space, significantly reducing O(V ×H) parameters to O(V × E + E × H). On the contrary, the benchmark scores with cross-layer parameter sharing are worse than no sharing at all, indicating that there exists a trade-off between the number of parameters and model performance. The parameter-reduction mechanism significantly reduces the number of unique parameters, enabling the ALBERT-xxlarge scale to even larger hidden layer size with having around 70% of parameters of BERT-large.

3.6. T5

T5 combines the results of a rigorous study done across different aspects of a model. It is a deep bidirectional pre-trained model that uses both the encoder and decoder model of the original transformer. The model is pre-trained on Colossal Clean Crawled Corpus (C4), a large dataset created to avoid repetitions, as models underperform when trained on a limited size dataset due to memorization of the repeated examples. A predefined prefix is added before every input sequence to perform a certain downstream task. Although STS-B [24] is a regression task (outputs a number between 1-5), T5 successfully transforms it into a 21-class classification problem. The output is rounded and assigned a class name, thereby fitting a regression task into the framework. Currently, T5 tops the STS-B leaderboard with T5-11B achieving new state-of-the-art solution.

3.7. PEGASUS

PEGASUS is a perfectly curated pre-trained model specialized in generating abstractive text summarization. It is a task-specific architecture where every component in pre-training maps closely to text summarization. The model is pre-trained on C4 [25] and HugeNews datasets, both consisting of large volumes of articles. Therefore, covering a depth of various domains for text summarization. The pre-training objective, Gap Sentence Generation (GSG), is also similar to the target downstream task, i.e., generating abstractive summaries from an input of masked sentences. The masking scheme in GSG suggests that masking important spans of sentences of an article leads to a better performance. The model is adaptable and can be fine-tuned on limited data, achieving human performance.

3.8. ELECTRA

The replaced token detection (RTD) is a sample-efficient pre-training objective as it leads to faster training by classifying all the input tokens instead of working on just 15% masked tokens, as in MLM. It is noticeable that the model consistently performs better at CoLA [26], as it roughly matches RTD's purpose. During pre-training, the generator (G) performs MLM, followed by replacing the [MASK] token with convincing substitutes using the maximum likelihood. The discriminator (D) then detects whether tokens in the modified sentence have been replaced or not. Finally, after the pre-training, the generator is removed from the architecture. The model is optimized by sharing embeddings between the two encoders and selecting an efficient generator size that reduces training compute. The paper further explores training ELECTRA ++ models by training them longer on a bigger dataset and creating more efficient models by distilling them.

3.9. GPT-3

GPT-3 is a remarkable language model capable of achieving near human performance in Natural Language Generation (NLG) downstream tasks such as writing poems, songs, and even a novel. On the contrary, it performs notably dull on natural language inference (NLI) tasks such as ANLI [26], WIC [27] and on many reading comprehension tasks such as DROP [28], MultiRC [29], RACE [13] and QuAC [30]. Surprisingly, GPT-3 performs worse than GPT on RACE-H and RACE-M datasets, scoring 10.6 and 4.8 points less, respectively. GPT-3 focuses less on research value and more on occupying top scores on the leaderboard. To conclude, GPT-3 has exceptional skills for text generative tasks though it lacks language inference capabilities and needs development in semantic understanding of the natural language.

4. COMPARATIVE ANALYSIS

The comparative analysis of the language model spans the following topics:

Table 1. Brief statistics about various language models.

Model	Year	Part of the Transformer used	Pretraining Objective used	Dataset Size	Total Train Compute*	Vocabulary Embeddings and its size
BERT	2018, Oct	Encoder	Denoising (MLM + NSP)	16GB	1.9E+20	WordPiece -30k
Transformer-XL	2019, Jan	Decoder	Autoregressive LM	*Not Specified*	*Not Specified*	*Not Specified*
XLNet	2019, June	Decoder	Permutation LM	158GB	3.9E+21	SentencePiece - 32k
RoBERTa	2019, July	Encoder	Denoising (Dynamic MLM)	160GB	3.2E+21	Byte-Pair Encoding - 50k
ALBERT	2019, Sept	Encoder	Denoising (MLM + SOP)	16 GB	3.1E+22	SentencePiece - 30k
T5	2019, Oct	Encoder + Decoder	Denoising (Span based MLM)	750GB	3.30E+22	SentencePiece – 32k
PEGASUS	2019, Dec	Encoder + Decoder	GSG	4.5 TB	*Not Specified*	Unigram – 96k
ELECTRA	2020, Mar	Encoder	Discriminating (MLM + RTD)	16GB	3.1E+21	WordPiece – 30k
GPT-3	2020, June	Decoder	Autoregressive LM	45 TB (unfiltered)	3.14E+23	Byte-Pair Encoding

*Total train compute corresponds to the computing power consumed by the largest variant of the model.

4.1. Autoregressive Models (AR)

GPT is the first transformer implementation of the Autoregressive Language Model. Transformer-XL performs far better than GPT because the segment-level recurrence captures longer dependencies than vanilla Transformer, on which GPT is based. XLNet introduces bidirectionality in the autoregressive model and integrates Transformer-XL's recurrence and

outperforms many AR models. GPT-3's enormous model has a relatively simple pre-training procedure, yet the few-shot learning approach achieves state-of-the-art result in many downstream tasks.

4.2. Autoencoding Models (AE)

BERT is the first transformer model to be pre-trained on the MLM objective giving rise to the first deep bidirectional pre-trained model. The dynamic masking strategy proposed by RoBERTa is better than the static masking strategy employed in BERT. It produces different corrupted versions of the masked sentences that help in avoiding repetitions of examples. ALBERT and T5 improve it further by masking spans of words. This strategy provides significant speedup during training and also ensures a balance between the length of the masked spans. Too much masking would fail to provide enough context for prediction whereas too little masking would make it computationally expensive. While PEGASUS is a sequence to sequence model, it masks whole sentences which makes it even more challenging.

4.3. Sequence to Sequence Models (SeqtoSeq)

SeqtoSeq models include both the parts of the Transformer, i.e., encoder and decoder. These models are generally used for language translations and generating abstractive summaries. The performance of PEGASUS, T5, and the vanilla Transformer on the CNN/DM [16] benchmark test is compared.

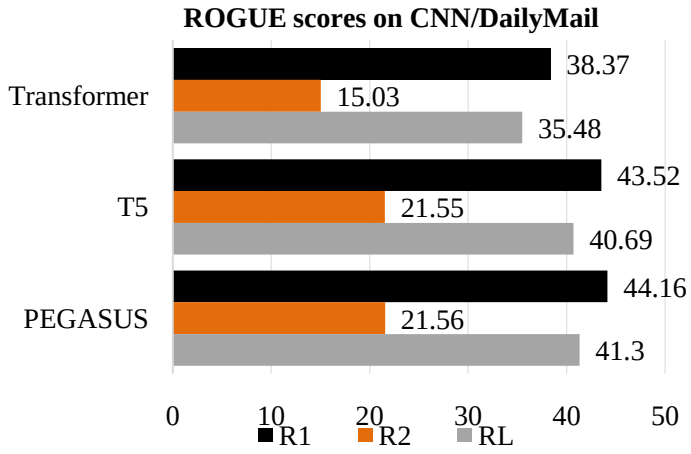

Fig. 1. ROGUE scores on CNN/DailyMail on different language models. Best viewed in colour.

PEGASUS outperforms every model because of two significant reasons. Firstly, it is a pre-trained model fine-tuned for text summarization tasks whereas, T5 is a multitask language model. Secondly, it is fine-tuned on extra training data that provides an edge.

4.4. Performance on Machine Translation tasks

On the WMT benchmark, although T5 and GPT-3 achieve a significant increase over the vanilla transformer, they are still behind the state-of-the-art solution. As for both the models a large portion of the training data is in English, therefore, pre-training on multilingual data can be beneficial. Also, the state-of-the-art models for machine translation are specially trained with back-translation [31][32], a data augmentation technique that obtains substantial improvement in machine translation tasks.

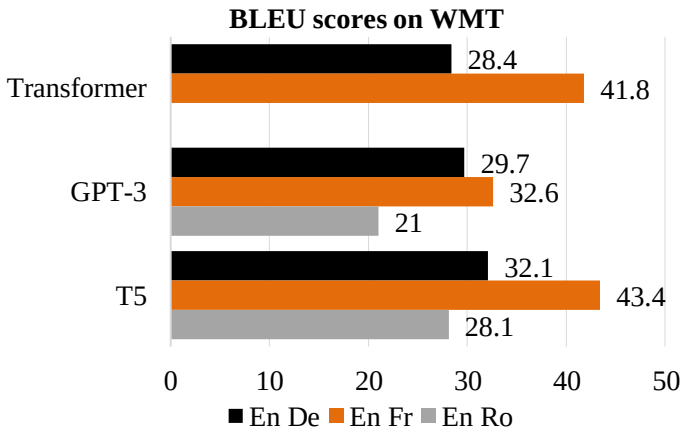

Fig. 2. BLEU scores on WMT for different language models. Best viewed in colour.

4.5. Performance on QA benchmarks

Question Answer tasks evaluate the logical reasoning abilities of the model. SQuAD 2.0, a reading comprehension benchmark is one of the prominent methods for testing QA abilities.

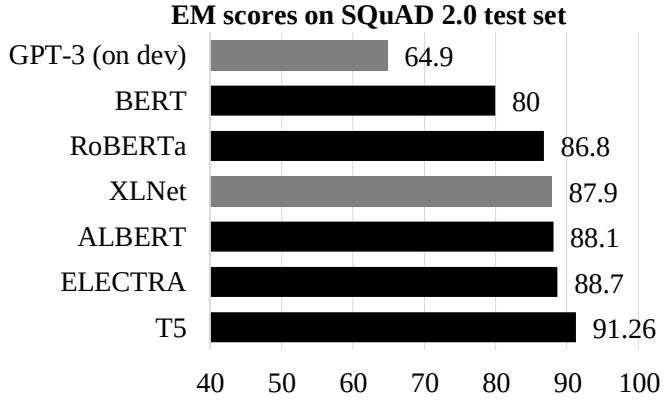

Fig. 3. Exact Match on SQUAD2.0 scored by various language models. Black bars represent Autoencoding (AE) models and grey bars represent autoregressive (AR) models.

AE models perform better than AR models because bidirectional context leads to better reasoning. XLNet scores marginally higher than RoBERTa, owing to two factors. Firstly, it is trained on additional data and secondly, it uses a layer-wise learning rate scheduler, while RoBERTa uses a uniform learning rate. It is noteworthy that ALBERT and ELECTRA are 30x and 47x smaller in comparison to T5, respectively. Also, ELECTRA consumes only 10% of the compute power required for pre-training T5. Considering this and the marginal performance gains of T5, ALBERT and ELECTRA emerge as strong competitors.

4.6. Performance on NLI benchmarks

Natural Language Inference tasks evaluate the semantical understanding between the sentences. Recognizing textual entailment (RTE) [33] is a common NLI task in GLUE and SuperGLUE, where the model predicts if meaning of one sentence can be inferred from another. The

conclusions are similar to the previous section. AE models perform better. GPT-3 lacks semantical understanding and performs comparably to BERT.

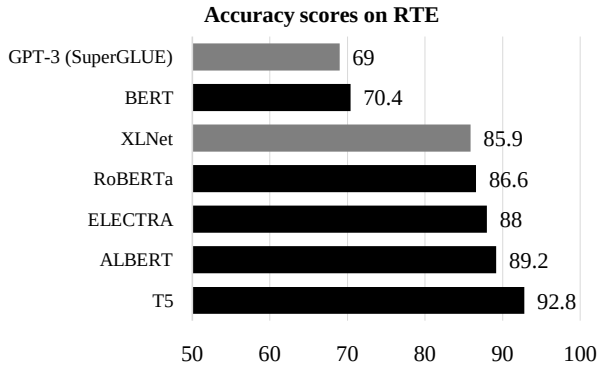

Fig. 4. Accuracy scores on RTE on different language models. Black bars represent AE models and grey bars represent AR models.

4.7. Increasing demand for computing power

It has become a fact that bigger models trained on large volumes of data are naturally better at NLP tasks. For such training, a lot of compute power is required. Below is the total computing power used by different language models in their pre-training phase.

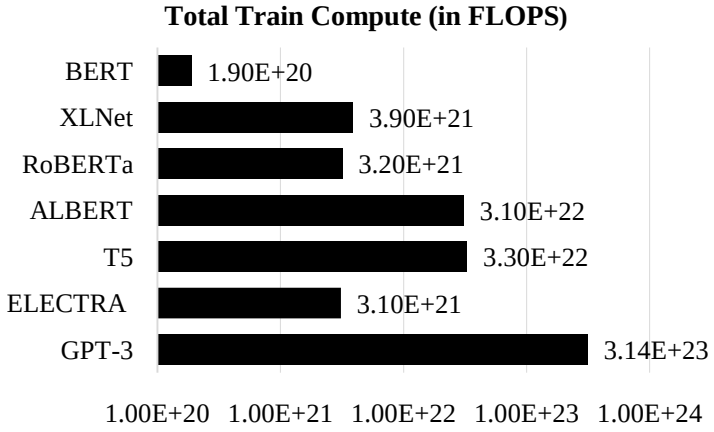

Fig. 5. The required training compute power for the largest variant of various language models. (plotted in log scale)

The increasing demand for energy resources can be observed as language models have evolved from BERT to GPT-3. ELECTRA efficiently demonstrates the balance between the computative power consumed and its remarkable capability.

4.8. Unwanted Gender Bias in models

The datasets used by language models are often compiled from online repositories such as Wikipedia, books, the internet, social media platforms, etc. In these sources, traces of human stereotypes and biases such as a male doctor and female nurse can be spotted easily. Winogender Schema [34], is a SuperGLUE diagnostic tool that evaluates a model's gender-parity and the accuracy with which a model can predict stereotypical sentences.

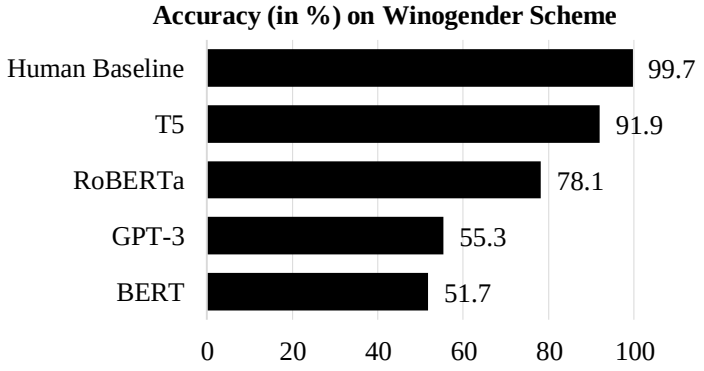

Fig. 6. Accuracy on Winogender scheme scored by various language models.

BERT and GPT-3 both suffer from gender bias, getting it right just over 50% of the time, whereas T5 tops the leader board. GPT-3's low scores can be accounted for as most occupations in the training data were associated with males. Unless the datasets are filtered thoroughly, biases will propagate into the language model.

4.9. Ineffectiveness of Next Sentence Prediction

NSP is a binary classification pre-training objective where the model learns sentence relationship by predicting if sentence 2 correctly follows sentence 1. The ineffectiveness raises from the kind of data used for training and testing. In the training data, sentences in the positive class correctly follow each other whereas sentences in the negative class are created by concatenating random sentences from two separate documents. Since the probability of mismatching topics is relatively higher, predicting negative class becomes easier, i.e., the model might excel in NSP without learning much about the sentimental relationship in sentences. Consequently, RoBERTa, XLNet, and ELECTRA chose not to include it and ALBERT replaced it with SOP, making it challenging by predicting the order of two consecutive sentences.

5. CONCLUSIONS

In this paper, a comparative analysis of transformer-based language models has been presented. The evaluation spans across various aspects such as language modeling methodology, design choices, compute power, accuracy scores on popular benchmarks, and shortcomings. With the introduction of Transformers, the process of developing language models has streamlined. The shift from pre-trained word embeddings to pre-trained language models marks the beginning of a new epoch. The tradition of pre-training massive models on large volumes of data to achieve state-of-the-art results has now become very common. It is also noteworthy that due to its valuable contribution to deep bidirectional representations, BERT has established itself as a baseline model. Despite the brilliance of state-of-the-art language models, prominent limitations such as the malicious utilization of language models, increasing consumption of energy resources, and the presence of stereotypical biases, lays the groundwork for future researches.

REFERENCES

[1] Alex Wang, Amapreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. arXiv preprint arXiv:1804.07461, 2018.

[2] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. SuperGLUE: A stickier benchmark for generalpurpose language understanding systems. arXiv preprint arXiv:1905.00537, 2019b.

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Advances in neural information processing systems, 2017.

[4] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. Neural computation, 9(8):1735–1780, 1997.

[5] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understandingwith unsupervised learning. Technical report, OpenAI.

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. FeiFei. ImageNet: A large-scale hierarchical image database. In CVPR, 2009.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pretraining of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

[8] Wilson L Taylor. 1953. Cloze procedure: A new tool for measuring readability. Journalism Bulletin, 30(4):415–433.

[9] Zihang Dai, Zhilin Yang, Yiming Yang, William W Cohen, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. arXiv preprint arXiv:1901.02860, 2019.

[10] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. XLNet: Generalized autoregressive pretraining for language understanding. arXiv preprint arXiv:1906.08237, 2019.

[11] 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. arXiv preprint arXiv:1907.11692, 2019.

[12] Pranav Rajpurkar, Jian Zhang,Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In Empirical Methods in Natural Language Processing (EMNLP).

[13] Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. Race: Large-scale reading comprehension dataset from examinations. arXiv preprint arXiv:1704.04683.

[14] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A lite bert for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942, 2019.

[15] Colin Raffel and Noam Shazeer and Adam Roberts and Katherine Lee and Sharan Narang and Michael Matena and Yanqi Zhou and Wei Li and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. . arXiv preprint arXiv: 1910.10683, 2019.

[16] Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. In Advances in neural information processing systems, 2015.

[17] Jingqing Zhang and Yao Zhao and Mohammad Saleh and Peter J. Liu. PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. arXiv preprint arXiv: 1912.08777, 2020.

[18] Kevin Clark and Minh-Thang Luong and Quoc V. Le and Christopher D. Manning. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. arXiv preprint arXiv: 2003.10555, 2020.

[19] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In NeurIPS, 2014.

[20] Avishek Joey Bose, Huan Ling, and Yanshuai Cao. Adversarial contrastive estimation. In ACL, 2018.

[21] 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, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin and Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, Dario Amodei. Language Models are Few-Shot Learners. arXiv preprint arXiv:2005.14165, 2020

[22] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners, 2019.

[23] Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In North American Association for Computational Linguistics (NAACL).
[24] Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. arXiv preprint arXiv:1708.00055, 2017.
[25] Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. Neural network acceptability judgments. arXiv preprint arXiv:1805.12471, 2018.
[26] Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. Adversarial nli: A new benchmark for natural language understanding. arXiv preprint arXiv:1910.14599, 2019.
[27] Mohammad Taher Pilehvar and Jose Camacho-Collados. WIC: 10,000 example pairs for evaluating context-sensitive representations. arXiv preprint arXiv:1808.09121, 2018.
[28] Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. arXiv preprint arXiv:1903.00161, 2019.
[29] Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In Proceedings of North American Chapter of the Association for Computational Linguistics (NAACL), 2018.
[30] Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. Quac : Question answering in context. Arxiv, 2018.
[31] Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. Understanding back-translation at scale. arXiv preprint arXiv:1808.09381, 2018.
[32] Guillaume Lample and Alexis Conneau. Cross-lingual language model pretraining. arXiv preprint arXiv:1901.07291, 2019.
[33] Ido Dagan, Oren Glickman, and Bernardo Magnini. The PASCAL recognising textual entailment challenge. In Machine Learning Challenges Workshop, 2005.
[34] Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. Gender bias in coreference resolution. arXiv preprint arXiv:1804.09301, 2018.