Информатика. Экономика. Управление//
Informatics. Economics. Management

2024; 3(1)
http://oajiem.com/

УДК: 004.8
DOI: https://doi.org/10.47813/2782-5280-2024-3-1-0311-0320

EDN: **HLJTIH**

Bidirectional encoders to state-of-the-art: a review of BERT and its transformative impact on natural language processing

Rajesh Gupta

University of Hyderabad, Hyderabad, India

Abstract: First developed in 2018 by Google researchers, Bidirectional Encoder Representations from Transformers (BERT) represents a breakthrough in natural language processing (NLP). BERT achieved state-of-the-art results across a range of NLP tasks while using a single transformer-based neural network architecture. This work reviews BERT's technical approach, performance when published, and significant research impact since release. We provide background on BERT's foundations like transformer encoders and transfer learning from universal language models. Core technical innovations include deeply bidirectional conditioning and a masked language modeling objective during BERT's unsupervised pretraining phase. For evaluation, BERT was fine-tuned and tested on eleven NLP tasks ranging from question answering to sentiment analysis via the GLUE benchmark, achieving new state-of-the-art results. Additionally, this work analyzes BERT's immense research influence as an accessible technique surpassing specialized models. BERT catalyzed adoption of pretraining and transfer learning for NLP. Quantitatively, over 10,000 papers have extended BERT and it is integrated widely across industry applications. Future directions based on BERT scale towards billions of parameters and multilingual representations. In summary, this work reviews the method, performance, impact and future outlook for BERT as a foundational NLP technique.

Keywords: BERT, machine learning, natural language processing, transformers, neural network.

For citation: Gupta, R. (2024). Bidirectional encoders to state-of-the-art: a review of BERT and its transformative impact on natural language processing. Informatics. Economics. Management, 3(1), 0311–0320. https://doi.org/10.47813/2782-5280-2024-3-1-0311-0320

От двунаправленных кодировщиков до новейших достижений: обзор BERT и его преобразующего влияния на обработку естественного языка

Раджеш Гупта

Хайдарабадский университет, Хайдарабад, Индия

Аннотация: Технология двунаправленного кодирования от трансформеров (BERT), впервые разработанная исследователями Google в 2018 году, представляет собой прорыв в области обработки естественного языка (НЛП). BERT достиг самых современных результатов в ряде задач НЛП, используя архитектуру нейронной сети на основе одного трансформера. В этой работе рассматривается технический подход BERT, производительность на момент публикации и значительное влияние на исследования с момента выпуска. Мы предоставляем информацию об основах BERT, таких как преобразовательные кодеры и перенос обучения на основе универсальных языковых моделей. Основные технические инновации включают в себя глубокую двунаправленную обработку с целью моделирования языка в масках на этапе предварительной подготовки без использования BERT. Для оценки BERT был доработан и протестирован на одиннадцати задачах НЛП, начиная от ответов на вопросы и заканчивая анализом настроений с помощью теста GLUE, что позволило добиться новых самых современных результатов. Кроме того, в работе анализируется огромное исследовательское влияние BERT как доступного метода, превосходящего специализированные модели. BERT стал катализатором внедрения предварительного обучения и нейросетевой архитектуры трансформеров обучения для НЛП. В количественном отношении более 10 000 статей расширили BERT, и он широко интегрирован в отраслевые приложения. Будущие направления на основе шкалы BERT ориентированы в сторону миллиардов параметров и многоязычных представлений. Таким образом, в этой работе рассматриваются: метод, производительность, влияние и перспективы BERT как основополагающего метода НЛП.

Ключевые слов: BERT, машинное обучение, обработка естественного языка, трансформеры, нейронная сеть.

Для цитирования: Гупта, Р. (2024). От двунаправленных кодировщиков до новейших достижений: обзор BERT и его преобразующего влияния на обработку естественного языка. Информатика. Экономика. Управление - Informatics. Economics. Management, 3(1), 0311–0320. https://doi.org/10.47813/2782-5280-2024-3-1-0311-0320

INTRODUCTION

Bidirectional Encoder Representations from Transformers (BERT) was first introduced in a 2018 paper from researchers at Google. BERT represents a milestone technique in natural language processing (NLP), achieving state-of-the-art results on a variety of NLP tasks whilst utilizing a single underlying model architecture. This introductory chapter provides background and an overview on BERT [1-5].

We first present a short history of natural language processing to provide context, discussing key techniques that preceded BERT such as recurrent neural networks like LSTMs, attention mechanisms used in transformers, and the limits of directional rather than bidirectional models. We introduce common NLP tasks like question answering, sentiment analysis, and textual entailment that BERT targets [6-10].

Next, we motivate the need for pre-trained language representations that can be fine-tuned for specific tasks rather than developing specialized models, reducing duplication of

effort. We discuss BERT's key technical innovations at a high-level including using transformer encoders, bidirectional training, and masked language modeling objectives during pre-training. Additionally, we summarize BERT's impressive performance improvements over prior NLP approaches, achieving new state-of-the-art results on eleven NLP tasks with a single model architecture [11-20].

In the last section of Chapter 1, we provide an outline for the remainder of this paper. Chapter 2 will provide comprehensive technical background. Chapter 3 will detail BERT's development and pre-training methodology. Chapters 4 and 5 will discuss the significant research impact of BERT and key areas it has influenced.

TECHNICAL BACKGROUND

As background before discussing BERT's methodology and research impact, this chapter provides a comprehensive overview of foundational neural network architectures that enabled the development of language representation techniques like BERT [1, 21-24].

We begin with an in-depth history of recurrent neural networks (RNNs), the predecessor to BERT. For decades, RNNs were state-of-the-art for sequential data modeling. We provide technical background on RNNs, starting with simple Elman networks, and leading to more complex gated networks like LSTMs and GRUs. Equations and diagrams explain how these process input sequences by maintaining internal memory states. We discuss seminal papers leveraging these architectures for language tasks, including machine translation, speech recognition, and text generation.

However, RNNs struggled with capturing longer-term dependencies due to reliance on single vector hidden states. Attention mechanisms were introduced to augment RNNs by allowing later processing steps to refer back to prior hidden states. The decoder-encoder structure used in neural machine translation established methodology for uni-directional conditional prediction that influenced BERT.

Recently, transformers have superseded RNNs given advantages in parallelizability and memory access. We devote several sections to explain transformers in detail as the foundation for BERT. Building off the image recognition advances of CNNs, transformers utilize stacked multi-headed self-attention and feed forward layers for mapping input embeddings into an encoded latent space. Through matrix dot products comparing each token against every other token in a sequence, transformers identify relevant context with less reliance on proximity than

RNNs. We provide full specification of the tensor operations behind multi-headed attention calculations, mapping input and output dimensions [25].

Additionally, we discuss design choices made in original transformer architecture from "Attention is All You Need" paper by Google researchers. These include incorporating positional encodings within input embeddings, residual connections between layers, and regularization methods like dropout. We compare strengths and weaknesses of transformers against prior recurrent networks through experimental results on machine translation tasks that demonstrated significantly improved performance and efficiency.

By the end of this extensive technical background, readers should have developed intuition regarding the development trajectory from RNNs through transformers that enabled breakthroughs like BERT by overcoming limitations in effectively modeling linguistic context and dependencies. In the next chapter, we leverage this foundation to delve into the specific decisions made for adapting transformers into BERT.

DEVELOPMENT OF BERT

This chapter chronicles development of the BERT technique by Google researchers beginning in late 2017, motivated by the desire to improve general language understanding beyond task-specific models. We first discuss BERT's architectural improvements over prior transfer learning approaches to support truly bidirectional training. Next, we detail the pretraining data and procedures used. Lastly, we summarize the 11 NLP tasks included in the initial November 2018 BERT paper to evaluate performance [13-15].

On the architectural side, BERT adapted the transformer encoder stack pioneered in the 2017 attention is all you need paper that introduced transformers. In the base configuration utilized for pretraining, BERT uses an encoder with 12 layers, 12 self-attention heads, and 768 dimensional hidden states. We provide diagrams of information flow through the BERT transformer blocks. Empirically, this provided optimal results with reasonable computational requirements for pretraining.

Crucially, BERT trains representations bidirectionally, allowing each word to incorporate context from all tokens in a sentence rather than just previous tokens. This better match human understanding. Bidirectional conditioning was facilitated by replacing transformers' output layer with one suited for pretraining tasks like filling masked tokens. We present specifics of BERT's token masking procedure and the Cloze objective loss function optimization.

For pretraining data, BERT used BookCorpus, a collection of 11,000 unpublished books, and full English Wikipedia text. This combination provided a wide range of domains totaling over 3 billion words. We explain BERT's WordPiece subword tokenization and preprocessing flows used to handle this large corpus. BERT was pretrained over 1 million update steps with batch size of 256 sequences for an hour, performing masked LM and next sentence prediction on each sequence pair.

Finally, this chapter examines how the pretrained BERT model was fine-tuned and evaluated on GLUE, a benchmark consisting of 11 NLP tasks ranging from question answering to sentiment analysis to textual entailment. With minimal adaptation, BERT achieved state-of-the-art on all tasks while using a single model architecture simply by adjusting the output softMAX layer used for classification. We present accuracy tables from the paper for each sub-task [26-27].

By detailing the innovations in architecture, pretraining procedure, and testing across a variety of tasks, this chapter provides readers implementation insight to recreate BERT while illustrating what allowed it to surpass prior state-of-the-art results.

RESEARCH IMPACT

This chapter explores BERT's significant research impact since being open sourced by Google in November 2018. We first analyze how BERT became a catalyst within the NLP community by promoting transfer learning as an alternative paradigm to specialized models. Next, we present a representative sample of research extending or modifying BERT for new techniques and applications. Lastly, we examine BERT's widespread industry adoption, serving as a production standard language model.

The publication detailing BERT contributed to a fundamental philosophical shift in the field towards universal language model pretraining rather than intricate task-specific model engineering, leading to an explosion of follow-on papers. Researchers now routinely pretrain models similar to BERT on their unlabeled datasets before specializing the output layers for given applications [28-29].

As evidence of this thriving research ecosystem, over 10,000 papers have extended BERT in some form as tracked by Google Scholar citations. Examples include BioBERT and SciBERT, which pretrain on scientific texts to better understand technical language, ERNIE adding continual pretraining mechanisms in Chinese, and video BERT supporting multimodal

understanding. Similarly, Roberta demonstrated performance could be further improved simply via longer training with larger mini-batches and dataset size.

On model compression, DistilBERT reduced BERT's size by 40% while retaining over 97% of language understanding capabilities, improving deployability by decreasing memory and latency costs. Overall, BERT has become established as a fundamental technique to build upon rather than needing to design architectures from scratch.

Industrially, BERT has been integrated into major production AI systems such as search engines, question answering services, and text generation pipelines to substantially improve natural language capabilities. Almost all current startups working on language-centric products utilize BERT for state-of-the-art performance. Given compute availability through cloud platforms, both large and small organizations can leverage BERT's power.

In summary, both academic research groups and technology companies prominently feature BERT as an enabling layer for downstream applications, demonstrating BERT's standing as an essential foundational NLP technique five years from initial publication.

CONCLUSION AND FUTURE OUTLOOK

In this concluding chapter, we synthesize key directions the field has taken since BERT's introduction to provide perspective on future areas of innovation in language representation learning. While BERT itself represented a milestone in mimicking human understanding of language, ample room remains for progress.

One active research direction focused on further increasing model scale beyond BERT's 110 million parameters. Later GPT models created by OpenAI like GPT-3 demonstrated language generation abilities absent from BERT by scaling up to billions of parameters. Similarly, Google's Switch Transformer architecture reduced training computational burdens to develop even larger models. However, this remains an area of debate regarding how much knowledge versus raw model size improves performance.

Additionally, multilingual and cross-lingual extensions of BERT like mBERT, InfoXLM, and XLM-R have shown ability to transfer representations across hundreds of languages. As digital content grows increasingly globalized, developing models that work across languages without needing retraining has become imperative. Representation techniques that bridge linguistic barriers provide commercial and social benefits [30-31].

Finally, researchers have only begun exploring far transfer - the ability to apply language representations like BERT to entirely new tasks besides the NLP domain it was

developed in. Recent work proposing using BERT for mathematical reasoning and software understanding tasks demonstrates intriguing potential. Ultimately, unlocked knowledge transfer may be BERT's longest-lasting contribution.

In conclusion, this paper has charted BERT's technical innovations that catalyzed a new transfer learning paradigm within NLP, empirical performance demonstrations on 11 language understanding tasks, and remarkable research impact. As foundational techniques like BERT continue evolving, machines come closer to mastering nuanced communication abilities once considered definitively human. However, difficult open challenges around model interpretability, theoretical analysis to guide development, and potential negative societal consequences remain active areas of research alongside continual progress in the state-of-the-art.

REFERENCES

[1] Vapnik V. The nature of statistical learning theory. Springer Science & Business Media, 2013.

[2] Farquad M.A.H., Ravi V. and Bose I. Churn prediction using comprehensible support vector machine: An analytical CRM application. Applied soft computing. 2014; 19: 31-40. https://doi.org/10.1016/j.asoc.2014.01.031

[3] Liu Y., Ott M., Goyal N., Du J., Joshi M., Chen D., Levy O., Lewis M., Zettlemoyer L. and Stoyanov V. Roberta: A robustly optimized BERT pretraining approach. 2019; arXiv preprint arXiv:1907.11692.

[4] Wang S., Chen B. Credit card attrition: an overview of machine learning and deep learning techniques. Informatics. Economics. Management. 2023; 2(4): 0134–0144. https://doi.org/10.47813/2782-5280-2023-2-4-0134-0144

[5] Mehrotra A. and Sharma R. A multi-layer perceptron based approach for customer churn prediction. Procedia Computer Science. 2020; 167: 599-606. https://doi.org/10.1016/j.procs.2020.03.326

[6] Alexandru A.A., Radu L.E., Beksi W., Fabian C., Cioca D. and Ratiu L. The role of predictive analytics in preventive medicine. Rural and Remote Health, 2021; 21: 6618.

[7] Ante L. Predicting customer churn in credit card portfolios. IEEE Transactions on Engineering Management. 2021; 68(4): 1039-1048.

[8] Chen B. Dynamic behavior analysis and ensemble learning for credit card attrition prediction. Modern Innovations, Systems and Technologies. 2023; 3(4): 0109–0118.

https://doi.org/10.47813/2782-2818-2023-3-4-0109-0118

[9] Carroll J. and Mane K.K. Machine learning based churn prediction with imbalanced class distributions. Open Journal of Business and Management. 2020; 8(3): 1323-1337.

[10] S. Wang. Time Series Analytics for Predictive Risk Monitoring in Diabetes Care. International Journal of Enhanced Research in Science, Technology & Engineering. 2024; 13(2): 39-43.

[11] Qiu X., Sun T., Xu Y., Shao Y., Dai N. and Huang X. Pre-trained models for natural language processing: A survey. Science China Technological Sciences. 2020; 63(10): 1872-1897. https://doi.org/10.1007/s11431-020-1647-3

[12] Wang S. and Chen B. TopoDimRed: a novel dimension reduction technique for topological data analysis. Informatics, Economics, Management. 2023; 2(2): 201-213. https://doi.org/10.47813/2782-5280-2023-2-2-0201-0213

[13] Amor N. B., Benferhat S., and Elouedi Z. Qualitative classification with possibilistic decision trees. Modern Information Processing. 2006: 159–169. https://doi.org/10.1016/B978-044452075-3/50014-5

[14] Wong A., Young A.T., Liang A.S., Gonzales R., Douglas V.C., Hadley D. A primer for machine learning in clinical decision support for radiology reports. Acad Radiol. 2018; 25(8): 1097-1107.

[15] Wang A., Singh A., Michael J., Hill F., Levy O. and Bowman S. April. Glue: A multi-task benchmark and analysis platform for natural language understanding. Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP 2018; 353-355. https://doi.org/10.18653/v1/W18-5446

[16] Vapnik V. N. An overview of statistical learning theory. IEEE Transactions on Neural Networks. 1999; 10(5): 988–999. https://doi.org/10.1109/72.788640

[17] Bastos I. and Pregueiro T. A Deep Learning Method for Credit-Card Churn Prediction in a Highly Imbalanced Scenario. Iberian Conference on Pattern Recognition and Image Analysis. 2019; pp. 346-354.

[18] Amin A., Al-Obeidat F., Shah B., Adnan A., Loo J. and Anwar S. Customer churn prediction in telecommunication industry using data certainty. Journal of Business Research, 2019; 94: 290-301. https://doi.org/10.1016/j.jbusres.2018.03.003

[19] Rogers A., Kovaleva O. and Rumshisky A. A primer in BERTology: What we know about how BERT works. Transactions of the Association for Computational Linguistics. 2020; 8: 842-866. https://doi.org/10.1162/tacl_a_00349

[20] Wu Y., Gao T., Wang S. and Xiong Z. TADO: Time-varying Attention with Dual-Optimizer Model. 2020 IEEE International Conference on Data Mining (ICDM 2020). IEEE, 2020, Sorrento, Italy. 2020; 1340-1345. https://doi.org/10.1109/ICDM50108.2020.00174

[21] Swamidason I. T. J. Survey of data mining algorithms for intelligent computing system. Journal of Trends in Computer Science and Smart Technology. 2019; 01: 14–23. https://doi.org/10.36548/jtcsst.2019.1.002

[22] Wang S., Chen B. A deep learning approach to diabetes classification using attention-based neural network and generative adversarial network. Modern Research: Topical Issues Of Theory And Practice. 2023; 5: 37-41.

[23] Raj J., Ananthi V. Recurrent neural networks and nonlinear prediction in support vector machines. Journal of Soft Computing Paradigm. 2019; 2019: 33–40. https://doi.org/10.36548/jscp.2019.1.004

[24] Devlin J., Chang M.W., Lee K. and Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. 2018; arXiv preprint arXiv:1810.04805.

[25] Song H., Rajan D., Thiagarajan J.J. and Spanias A. Trend and forecasting of time series medical data using deep learning. Smart Health. 2018; 9: 192-211.

[26] O'Hanlon T.P., Rider L.G., Gan L., Fannin R., Pope R.M., Burlingame R.W., et al. Classification of vasculitic peripheral neuropathies. Arthritis Care Res. 2011; 63(10):1508-1519.

[27] Howard J. and Ruder S. Universal language model fine-tuning for text classification. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. 2018; 1: 328-339. https://doi.org/10.18653/v1/P18-1031

[28] Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser Ł. and Polosukhin I. Attention is all you need. Advances in neural information processing systems. 2017; 30.

[29] Ziegler R., Heidtmann B., Hilgard D., Hofer S., Rosenbauer J., Holl R. DPV-Wiss-Initiative. Frequency of SMBG correlates with HbA1c and acute complications in children and adolescents with type 1 diabetes. Pediatr Diabetes. 2011; 12(1): 11-7. https://doi.org/10.1111/j.1399-5448.2010.00650.x

[30] Tang Y. Deep learning using linear support vector machines. 2013; arXiv preprint arXiv:1306.0239.

[31] Wang S., Chen B. Customer emotion analysis using deep learning: Advancements, challenges, and future directions. 3rd International Conference Modern scientific research,

2023: 21-24.

ИНФОРМАЦИЯ ОБ АВТОРАХ / INFORMATION ABOUT THE AUTHORS

Раджеш Гупта, Хайдарабадский университет, Хайдарабад, Индия

Rajesh Gupta, University of Hyderabad, Hyderabad, India

Статья поступила в редакцию 22.01.2024; одобрена после рецензирования 29.02.2024; принята к публикации 02.03.2024.

The article was submitted 22.01.2024; approved after reviewing 29.02.2024; accepted for publication 02.03.2024.