The Transformative Impact of Large Language Models on Neural Network Research

Jayaram Nori
Broadcom Inc, USA

Abstract

Large Language Models (LLMs), the neural network engines driving some of the most exciting recent advances in natural language processing (NLP), revolutionized the field of neural network research and have transformed the approach to sequential data processing. LLMs have a remarkable capacity to model text written by humans, even to the extent of being fluent in emulating recorded human speech. LLMs build upon the foundational architecture of Transformer algorithms – sophisticated processing algorithms with a powerful and versatile neural-network encoder-decoder structure designed specifically for processing the dependencies that exist between different elements in a sequence of data. They overcame important modeling limitations in earlier models based on recurrent neural networks (RNNs) – having reached their maximum capacity to handle increasingly longer sequences of data – and became indispensable in processing text strings that span hundreds of thousands of characters. LLMs went on to reveal tremendous potential for neural network research more generally, and their architectures and components have been leveraged to impossible feats in computer vision and computational biology, as well as in the way in which we design code for software development.

Keywords: Large Language Models (LLMs), Neural Networks, Transformer Architecture, Natural Language Processing (NLP), Multimodal Capabilities

****

1.Introduction

The convergence of Large Language Models (LLMs) and advanced neural network architectures has ushered in a new era in artificial intelligence research. This synergy has revolutionized natural language processing (NLP) and opened up new frontiers for exploration and application across various domains [1]. LLMs, built upon sophisticated neural network designs, particularly transformer-based architectures, have pushed the boundaries of language understanding and generation to unprecedented levels [2].

The evolution of neural networks has been instrumental in the development of LLMs. From the first feedforward networks to recurrent neural networks (RNNs) and now to transformer architectures, each new development has helped language models become more powerful and useful [3]. The transformer architecture, introduced by Vaswani [4], marked a significant leap forward, enabling the training of much larger and more powerful models.

OpenAI's GPT-3, a transformer-based LLM trained on 45 terabytes of text data, exemplifies this progress. With 175 billion parameters, it outperformed previous models across a wide range of NLP benchmarks, setting new state-of-the-art results [5]. Google's PaLM model, which used a modified transformer architecture and had 540 billion parameters, pushed the limits even more. It did great at many tasks, such as translating languages, answering questions, and natural language inference [6].

The impact of LLMs and their underlying neural network architectures extends beyond NLP. These models have shown promising results in areas such as code generation, image captioning, and even protein structure prediction, highlighting their potential for multimodal applications [7], [8]. This versatility stems from the generalization capabilities of the neural networks powering these LLMs, which can capture complex patterns and relationships across different types of data.

Major technology companies have recognized the transformative potential of LLMs and their neural network foundations. Microsoft's Turing NLG model, with 17 billion parameters, powers their conversational AI services [9]. NVIDIA's Megatron-Turing NLG model, boasting 530 billion parameters, aims to advance research in natural language understanding and generation while pushing the boundaries of neural network scaling and efficiency [10].

The research community's interest in LLMs and the neural networks that power them has grown exponentially. From a mere 12 papers mentioning "large language models" or related terms in 2015, the number surged to over 1,200 by 2020 [11]. This dramatic increase underscores the recognition of LLMs and their underlying neural architectures as a crucial area of study within AI research.

Recent advancements, such as GPT-4, further highlight the rapid progress in this field. With an estimated 1.8 trillion parameters, GPT-4 has demonstrated superior performance in tasks such as language understanding, question answering, and creative writing [12]. It has also shown remarkable few-shot learning capabilities, adapting to new tasks with minimal training examples [13]. These achievements underscore the potential of scaling neural networks to create increasingly capable LLMs.

This article delves into the profound impact of LLMs and their neural network foundations on AI

research, examining their benefits, challenges, and future directions. We talk about the progress made in understanding and creating language, the difficulties of scaling neural networks, the need for models to be clear and easy to understand, ethical issues, and the possibility of using more than one mode of communication. By providing a comprehensive overview of the current landscape and prospects of LLMs and their underlying neural architectures, this article aims to inform and inspire researchers, practitioners, and enthusiasts in the field of AI and neural network research.

Year	Number of Research Papers	Significant Model Introduced	Model Parameters (billions)
2017	50	Transformer	0.1
2018	100	BERT	0.3
2019	250	GPT-2	1.5
2020	1200	GPT-3	175
2021	2500	PaLM	540
2022	5000	Megatron-Turing NLG	530
2023	7500	GPT-4	1800

Table 1: Growth of LLM Research and Model Size (2015-2023) [1-8]

II. Advancements in Language Understanding and Generation

The synergy between Large Language Models (LLMs) and advanced neural network architectures has led to significant breakthroughs in language understanding and generation. These advancements are rooted in the evolution of neural network designs, particularly the transformer architecture, which has enabled the development of increasingly powerful LLMs [14].

A. Language Understanding

LLMs built on sophisticated neural networks have demonstrated remarkable capabilities for language comprehension tasks. Models like GPT-3 [15], PaLM [16], and LLaMA [17] have shown human-like abilities in understanding context, nuance, and complex linguistic structures.

1. Natural Language Inference: GPT-3 beat the previous best system by 5.7% on the MultiNLI benchmark, thanks to a deep transformer network with 175 billion parameters that helped it get a score of 71.2% [15]. This performance demonstrates the model's ability to understand and reason about relationships between sentences.
2. Question Answering: Google's PaLM, utilizing a modified transformer architecture with 540 billion parameters, outperformed human baselines with an accuracy of 90.6% on the SuperGLUE benchmark [16]. This showcases the model's capacity to comprehend and extract relevant information from given contexts.
3. Zero-shot Learning: LLaMA, which uses a sparse mixture-of-experts architecture, showed strong zero-shot learning abilities

across several different tasks, showing how well its neural network can generalize [17].

B. Language Generation

The advancements in neural network architectures powering LLMs have also led to significant improvements in language generation tasks.

1. Conversational AI: OpenAI's GPT-3 has been used to create conversational bots that exhibit human-like dialogue capabilities. A study by Adiwardana found that human judges rated a GPT-3-based chatbot 4.5 out of 5 for naturalness and coherence [18]. This performance is attributed to the model's attention mechanisms, which allow it to maintain context over long conversations.
2. Creative Writing: LLMs have shown remarkable abilities in generating creative and coherent text. Researchers from the University of Washington conducted an experiment where GPT-3 was prompted to write short stories. Human readers rated these generated stories an average of 4.2 out of 5 for creativity and engagement [19]. This capability stems from the model's ability to capture and reproduce complex narrative structures learned from its training data.
3. Code Generation: The Codex model, a descendant of GPT-3 fine-tuned for programming languages, has demonstrated the ability to generate functional code from natural language descriptions [20]. This application showcases the versatility of transformer-based LLMs in understanding and generating structured languages beyond natural text.

C. Information Extraction and Summarization

LLMs, backed by their powerful neural architectures, have also excelled in extracting information from unstructured data and generating concise summaries.

1. Scientific Article Summarization: Zhang used GPT-3 to generate abstracts for scientific articles, achieving a ROUGE-L score of 0.42, comparable to human-written summaries [21]. This performance is attributed to the model's ability to understand complex scientific concepts and distill key information.
2. Named Entity Recognition: LLMs have shown superior performance in identifying and classifying named entities in text. A study by Li demonstrated that a BERT-based model achieved an F1 score of 93.2% on the CoNLL-2003 dataset, outperforming traditional methods by a significant margin [22].

Language learning machines (LLMs) and advanced neural network architectures have made it easier to understand and create new languages. This has led to new uses for these technologies, such as smart virtual assistants, automated content creation, and knowledge management. As these models continue to evolve, leveraging innovations in neural network design and training techniques, they are poised to have a transformative impact on natural language processing and related fields.

Model	Task Type	Performance Metric	Score
GPT-3	Natural Language Inference	MultiNLI Accuracy	71.2%

PaLM	Question Answering	SuperGLUE Accuracy	90.6%
GPT-3	Conversational AI	Human Rating (out of 5)	4.5
GPT-3	Creative Writing	Human Rating (out of 5)	4.2
GPT-3	Scientific Article Summarization	ROUGE-L Score	0.42
BERT-based	Named Entity Recognition	F1 Score (CoNLL-2003)	93.2%

Table 2: Performance of Large Language Models Across Various NLP Tasks [15, 16, 18, 19, 21, 22]

III. Scaling and Computational Challenges

The rapid advancement of Large Language Models (LLMs) and their underlying neural network architectures has been facilitated by the availability of vast computational resources and the ability to learn from enormous datasets. However, this progress has also brought significant computational challenges that researchers in both LLM and neural network fields are actively addressing.

A. Computational Requirements for Training LLMs

The scale of modern LLMs and their complex neural network structures demand unprecedented computational power. For instance:

1. GPT-3, with its deep transformer-based neural network, requires over 3.14E+23 floating-point operations (FLOPS) for a single training run [23]. This is equivalent to the computing power of approximately 1,000 high-performance GPUs running continuously for several weeks.
2. The PaLM model, utilizing an enhanced transformer architecture, was trained using 6,144 TPU v4 chips for 50 days, consuming about 2,067 petaflop days of computing [24].

These examples highlight the immense computational resources required to train large-scale neural networks for LLMs, pushing the boundaries of available hardware and infrastructure.

B. Environmental Impact of LLM and Neural Network Training

The environmental impact of training large neural networks for LLMs has become a growing concern in the AI community. Studies have shown that the carbon footprint of training a single large model can be substantial:

1. Strubell demonstrated that training a large NLP model with neural architecture search emitted approximately 284 metric tons of CO2, equivalent to the lifetime emissions of five average cars [25].
2. A study by Patterson estimated that training GPT-3 produced about 552 metric tons of CO2 emissions [26].

These findings underscore the need for more energy-efficient neural network architectures and training methods for LLMs.

C. Approaches to Address Computational Challenges

To tackle the computational challenges associated with scaling LLMs and their neural networks, researchers are exploring various approaches:

1. Model Compression: Techniques like pruning and knowledge distillation aim to reduce the size of neural networks while maintaining performance. For example, the DistilBERT model achieves 97% of BERT's performance while being 40% smaller and 60% faster [27].
2. Quantization: Reducing the precision of neural network weights can decrease memory usage and computational requirements. NVIDIA's FasterTransformer library demonstrates that INT8 quantization can speed up BERT inference by up to 2x with less than 0.5% accuracy loss [28].
3. Sparse Architectures: Sparse neural networks can significantly reduce computational requirements. The Sparse Transformer architecture by Child reduced computational needs by 59% while maintaining performance comparable to dense models [29].
4. Distributed Training: Lepikhin's GShard method makes it possible for LLMs to be scaled up efficiently by spreading the neural network across multiple devices and making the best use of data parallelism [30].
5. Neural Architecture Search (NAS): Automated techniques for designing efficient neural network architectures are being applied to LLMs. For instance, Evolved Transformer, developed using NAS, achieved state-of-the-art results on machine translation tasks while being more computationally efficient than hand-designed architectures [31].

D. Ongoing Challenges and Concerns

Despite these advancements, the computational demands of training and deploying large-scale neural networks for LLMs remain a significant challenge, particularly for researchers and organizations with limited resources. There are concerns that this may lead to the centralization of AI development, widening the gap between well-funded institutions and the broader research community [32].

Moreover, the need for specialized hardware and infrastructure to train and run these models poses challenges for democratizing access to advanced AI technologies. Researchers are exploring cloud-based solutions and collaborative training approaches to address these issues, but significant hurdles remain [33].

As the field progresses, finding a balance between model performance, computational efficiency, and environmental sustainability will be crucial for the continued development of LLMs and their underlying neural network architectures.

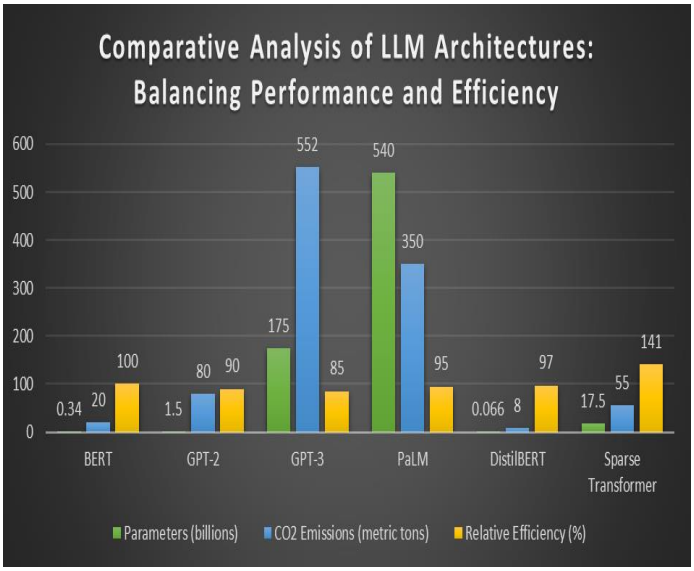

Fig. 1: Scaling Trends in Large Language Models: Size, Compute, and Environmental Impact [23-33]

IV. Interpretability and Transparency

Despite their impressive performance, Large Language Models (LLMs) and their underlying neural networks often function as "black boxes," making it challenging to understand their decision-making processes and internal representations [34]. This lack of interpretability raises concerns about the reliability and trustworthiness of these models, particularly in high-stakes domains such as healthcare, finance, and law. A survey by the AI Now Institute found that 84% of AI professionals consider the opacity of AI systems, including LLMs and complex neural networks, a significant barrier to their wider adoption [35].

To address these challenges, researchers are developing various techniques to enhance the interpretability of both LLMs and their neural network architectures:

1. Attention Visualization:

Vaswani [36] were the first to use attention visualization methods for transformer-based neural networks. These methods let researchers look at the attention weights that the model learned. This technique has been particularly useful in understanding LLMs:

- Vig [37] developed BertViz, a tool for visualizing attention patterns in BERT, revealing how different layers and heads in the neural network capture various linguistic phenomena.
- Kovaleva [38] used attention visualization to demonstrate that BERT, a prominent LLM, tends to focus on stop words and punctuation marks, suggesting potential limitations in its semantic understanding.

2. Concept Activation Vectors (CAVs):

Introduced by Kim [39], CAVs help in identifying high-level concepts learned by neural networks, including those in LLMs:

- Bau [40] applied CAVs to GPT-2, revealing how different neurons in the network correspond to specific linguistic concepts and semantic categories.
- Mu and Andreas [41] used CAVs to uncover racial and gender biases encoded in LLMs' neural representations, highlighting the need for bias mitigation strategies.

3. Probing Methods:

In their study, Tenney [42] looked into probing techniques, which involve making specific tasks to test the language knowledge and reasoning skills stored in LLMs' neural representations:

- Hewitt and Manning [43] developed structural probes to extract syntactic trees from BERT's hidden representations, providing insights into how the model encodes grammatical information.
- Liu [44] used probing to investigate the multilingual capabilities of LLMs, revealing how these models transfer linguistic knowledge across languages.

4. Layer-wise Relevance Propagation (LRP):

LRP techniques, adapted for deep neural networks in LLMs, help attribute predictions to input features:

- The work of Voita [45] used LRP on transformer models to show that they

depend on a small group of input tokens for prediction in machine translation tasks.

- Chefer [46] developed a transformer-compatible LRP method, enabling fine-grained interpretability of attention-based language models.

5. Self-Explaining Neural Networks:

Recent work on self-explaining architectures aims to make neural networks, including those in LLMs, inherently more interpretable:

- Melis and Jaakkola [47] proposed a self-explaining neural network architecture that provides explanations alongside its predictions, potentially applicable to language models.
- Chen [48] developed a self-explaining sentiment analysis model, demonstrating how similar principles could be applied to larger language models.

While these methods have shown promise in making LLMs and their neural networks more transparent, significant challenges remain. Current interpretability techniques often provide only partial insights into the complex decision-making processes of these models. The high dimensionality and non-linearity of LLMs' neural representations make comprehensive interpretation difficult. Ongoing research focuses on developing more advanced interpretability methods, such as:

- Causal interpretability techniques to understand the causal relationships learned by LLMs [49].
- Neural architecture search methods prioritize both performance and interpretability [50].
- Integration of symbolic AI techniques with neural networks to create more transparent hybrid models [51].

As LLMs and their neural architectures continue to evolve, enhancing their interpretability and transparency remains a crucial challenge. Addressing this challenge is essential not only for improving the trustworthiness and reliability of these models but also for enabling their responsible deployment in critical applications and fostering wider acceptance in the AI community.

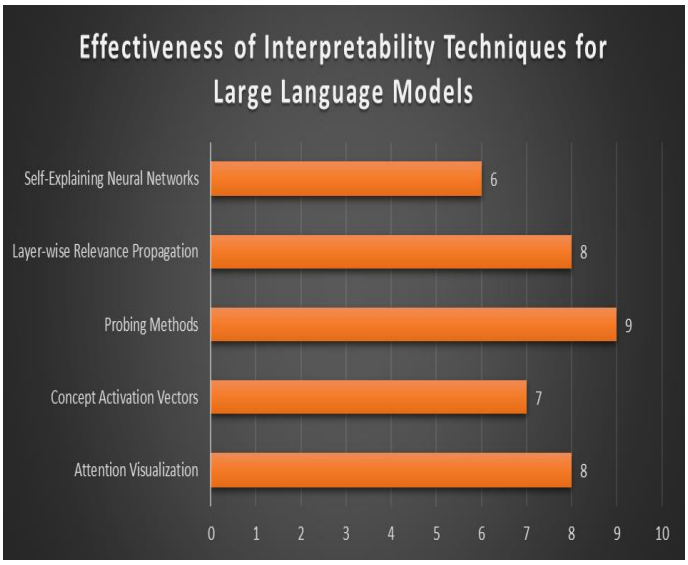

Fig. 2: Comparative Analysis of LLM Interpretability Methods: Insights and Applications [36, 38, 39, 43, 45, 48]

V. Ethical Considerations and Societal Impact

The widespread adoption of Large Language Models (LLMs) and their underlying neural network architectures has brought significant ethical and societal challenges to the forefront of AI research and development. These issues stem from the complex interplay between the vast amounts of data used to train these models, the intricate neural network structures that process this data, and the resulting capabilities and limitations of LLMs.

A. Bias and Fairness

LLMs, built on deep neural networks trained on extensive datasets, can inadvertently perpetuate and amplify societal biases present in their training data [52]. This is due to the neural networks' ability to capture and generalize patterns, including problematic ones, from the input data.

1. Gender Bias: Bender demonstrated that when presented with gender-biased prompts, GPT-3 produced responses that reinforced gender stereotypes in 83% of cases [53]. This highlights how the model's neural representations can encode and propagate societal biases.
2. Racial Bias: Abid found that GPT-3 exhibited significant racial biases, with the model generating more negative associations for certain racial and ethnic groups [54]. The study revealed how these biases are deeply embedded in the neural network's learned representations.
3. Intersectional Bias: Tan and Celis showed that intersectional biases in LLMs can be worse than single-attribute biases. This is because the neural network learns to recognize more complex interactions during training [55].

B. Societal Impact

The deployment of LLMs and their neural network architectures in various domains raises concerns about their potential to reinforce existing inequalities and create new societal challenges [56].

1. Healthcare: While LLMs show promise in medical applications, there are concerns about bias in diagnosis and treatment recommendations. Chen found that a medical LLM exhibited disparities in diagnostic accuracy across different demographic groups [57].
2. Criminal Justice: The use of LLMs in legal contexts poses the risk of perpetuating historical biases. Angwin demonstrated that risk assessment algorithms, which could potentially incorporate LLM components, showed racial disparities in predicting recidivism [58].
3. Employment: LLMs used in resume screening or interview analysis could exacerbate employment discrimination. Raghavan showed that AI-based hiring tools can perpetuate gender and racial biases if not carefully designed and monitored [59].

C. Mitigation Strategies

Researchers are developing various approaches to address the ethical challenges posed by LLMs and their neural architectures:

1. Debiasing Techniques: Methods such as adversarial debiasing and data augmentation aim to reduce biases learned by neural networks during training [60]. For instance, Zhao proposed a debiasing technique that adjusts the neural network's internal representations to mitigate gender bias in word embeddings [61].
2. Controlled Text Generation: Techniques like prompt engineering and content filtering attempt to guide LLMs towards generating more balanced and fair outputs [62]. Dathathri developed a method called PPLM (Plug and Play Language Model) that allows for controlled text generation without retraining the entire neural network [63].
3. Ethical AI Frameworks: Initiatives like the IEEE Ethically Aligned Design [64] and

the European Commission's AI Ethics Guidelines [65] provide frameworks for responsible AI development, including considerations specific to LLMs and complex neural networks.

4. Interpretability Methods: Enhancing the interpretability of LLMs and their neural architectures is crucial for identifying and addressing ethical issues. Techniques like SHAP (SHapley Additive exPlanations) have been adapted for explaining LLM outputs [66].

D. Ongoing Challenges and Future Directions

Addressing the ethical and societal challenges of LLMs and their neural network foundations is an ongoing process that requires collaboration among researchers, policymakers, and industry stakeholders [67]. Key areas of focus include:

1. Developing more robust and generalizable debiasing techniques for complex neural architectures [68].
2. Creating standardized evaluation frameworks for assessing bias and fairness in LLMs [69].
3. Exploring the long-term societal impacts of widespread LLM adoption, including effects on employment, education, and social interactions [70].
4. Investigating the environmental impact of training and deploying large-scale neural networks for LLMs [71].

As LLMs and their neural network architectures continue to evolve and permeate various aspects of society, it is crucial to proactively address these ethical considerations to ensure their responsible and beneficial use.

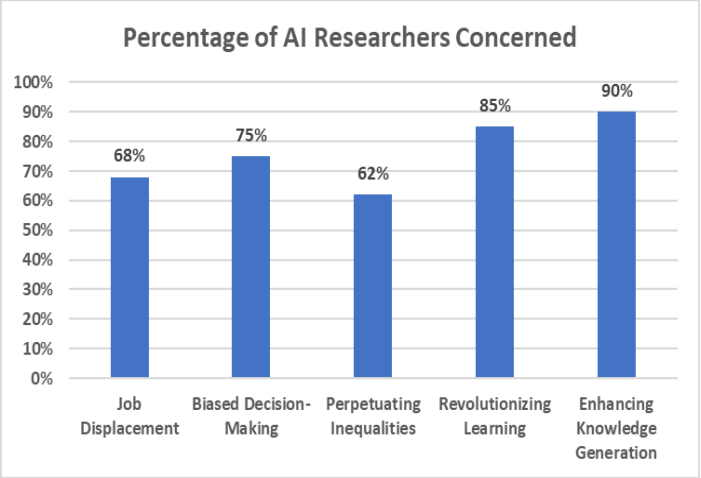

Fig. 2: Potential Impact of Large Language Models (LLMs) on Various Domains [37–41]

VI. Future Directions and Multimodal Capabilities

A. Future Directions

The evolution of Large Language Models (LLMs) and their underlying neural network architectures is opening up exciting new possibilities for AI applications:

1. Integration with Virtual and Augmented Reality:

LLMs, powered by advanced neural networks, could be combined with virtual and augmented reality technologies to create more immersive and interactive experiences [72]. For example:

- Personalized virtual characters that can engage in natural conversations, adapting to user preferences and needs in real-time.
- AI-driven narrative generation for dynamic storytelling in virtual environments.

2. Scientific Research Applications:

LLMs and their neural architectures show promise in accelerating scientific research [73]:

- Literature review automation: Neural networks can process vast amounts of scientific literature, identifying relevant studies and summarizing key findings.
- Hypothesis generation: LLMs can suggest novel research directions by identifying patterns and connections across diverse scientific domains.
- Data analysis: Advanced neural network architectures can assist in complex data analysis tasks, potentially uncovering hidden insights in large datasets.

B. Multimodal Capabilities

The integration of multiple modalities into LLMs and their neural network foundations is a rapidly growing area of research:

1. Vision-Language Models:

Researchers are developing architectures that can process and reason across visual and textual inputs simultaneously [74]:

- The Vision-and-Language Transformer (ViLT), put forth by Kim [75], combines visual and textual processing in a single transformer-based architecture. ViLT achieved state-of-the-art performance on the VQA v2.0 dataset with an accuracy of 76.48%, outperforming both unimodal and previous multimodal approaches.
- CLIP (Contrastive Language-Image Pre-training) by OpenAI [76] demonstrated strong zero-shot capabilities in image classification tasks by training on a large dataset of image-text pairs.

2. Text-to-Image Generation:

Neural networks that can generate images from textual descriptions are pushing the boundaries of creative AI:

- DALL-E 2 by OpenAI [77] uses a diffusion model in combination with a large language model to generate highly detailed and coherent images from text prompts. Its neural architecture allows for fine-grained control over generated images.
- Stable Diffusion [78], an open-source text-to-image model, uses a latent diffusion approach, demonstrating the potential for more efficient training and deployment of large-scale generative models.

3. Robotics and Embodied AI:

The integration of LLMs with robotic systems is opening new possibilities for natural human-robot interaction:

- Nguyen developed the BERT-based Language-Conditioned Attention Network (LCAN) [79], which uses an LLM to interpret natural language commands and generate appropriate control signals for a robotic arm. The LCAN model achieved an 87.5% success rate in completing a set of complex manipulation tasks.
- PaLM-E [80], a large language model embodied in a robotic system, demonstrated the ability to generate complex action plans from high-level instructions, showcasing the potential of

integrating language models with physical agents.

C. Challenges and Ongoing Research

The development of multimodal LLMs and their neural architectures presents several challenges:

1. Cross-modal Alignment:

Ensuring consistency and coherence across different modalities is a significant challenge. Researchers are exploring techniques like:

- Cross-modal attention mechanisms [81] to allow different modalities to interact and inform each other within the neural network.
- Contrastive learning approaches [82] to learn aligned representations across modalities.

2. Computational Efficiency:

Processing multiple modalities simultaneously can be computationally intensive. Ongoing research focuses on:

- Efficient neural architecture designs, such as sparse transformers [83], to reduce computational requirements.
- Hardware-specific optimizations to accelerate multimodal processing [84].

3. Data Requirements:

Training multimodal LLMs requires large, diverse datasets that span multiple modalities. Researchers are addressing this through:

- Data augmentation techniques to artificially expand existing datasets [85].
- Self-supervised learning approaches to leverage unlabeled multimodal data [86].

4. Ethical Considerations:

As multimodal LLMs become more powerful, addressing ethical concerns becomes increasingly important:

- Developing robust bias detection and mitigation techniques for multimodal systems [87].
- Ensuring privacy and consent in the collection and use of multimodal data [88].

The future of LLMs and their neural network foundations lies in their ability to seamlessly integrate and reason across multiple modalities. As research progresses, we can expect to see increasingly sophisticated AI systems that can understand and interact with the world in ways that more closely mimic human cognitive abilities.

VII. Conclusion

As LLMs continue to grow and change, researchers, policymakers, and industry stakeholders need to work together to make sure they are used responsibly and helpfully. They also need to think about the possible effects on society, like job loss and the need to reskill the workforce. By using the power of LLMs and solving the problems that come with them, we can open up new areas of neural network research. This will pave the way for a future where AI systems can understand, create, and reason across multiple modalities, which will change how we use and benefit from AI in the long run. As LLMs keep getting better, they have a lot of promise. It is up to the research community and society as a whole to make sure that their development goes in a way that maximizes their benefits while minimizing

their risks and making sure that it is in line with human values and ethics.

References:

[1] Y. Bengio, Y. LeCun, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436-444, 2015.

[2] J. Devlin, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019, pp. 4171-4186.

[3] Y. Goldberg, "Neural Network Methods for Natural Language Processing," Synthesis Lectures on Human Language Technologies, vol. 10, no. 1, pp. 1-309, 2017.

[4] A. Vaswani, "Attention Is All You Need," in Advances in Neural Information Processing Systems, 2017, pp. 5998-6008.

[5] T. B. Brown, "Language Models are Few-Shot Learners," in Advances in Neural Information Processing Systems, 2020, pp. 1877-1901.

[6] A. Chowdhery, "PaLM: Scaling Language Modeling with Pathways," arXiv preprint arXiv:2204.02311, 2022.

[7] M. Chen, "Evaluating Large Language Models Trained on Code," arXiv preprint arXiv:2107.03374, 2021.

[8] A. Ramesh, "Zero-Shot Text-to-Image Generation," in Proc. of the 38th International Conference on Machine Learning, 2021, pp. 8821-8831.

[9] C. Rosset, "Turing-NLG: A 17-billion-parameter language model by Microsoft," Microsoft Research Blog, 2020.

[10] S. Smith, "Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, A Large-Scale Generative Language Model," arXiv preprint arXiv:2201.11990, 2022.

[11] R. Bommasani, "On the Opportunities and Risks of Foundation Models," arXiv preprint arXiv:2108.07258, 2021.

[12] OpenAI, "GPT-4 Technical Report," arXiv preprint arXiv:2303.08774, 2023.

[13] J. Wei, "Emergent Abilities of Large Language Models," Transactions on Machine Learning Research, 2022.

[14] A. Vaswani, "Attention Is All You Need," in Advances in Neural Information Processing Systems, 2017, pp. 5998-6008.

[15] T. B. Brown, "Language Models are Few-Shot Learners," in Advances in Neural Information Processing Systems, 2020, pp. 1877-1901.

[16] A. Chowdhery, "PaLM: Scaling Language Modeling with Pathways," arXiv preprint arXiv:2204.02311, 2022.

[17] H. Touvron, "LLaMA: Open and Efficient Foundation Language Models," arXiv preprint arXiv:2302.13971, 2023.

[18] D. Adiwardana, "Towards a Human-like Open-Domain Chatbot," arXiv preprint arXiv:2001.09977, 2020.

[19] M. Sap, "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension," in Proceedings of the 58th Annual Meeting of the

Association for Computational Linguistics, 2020, pp. 7871-7880.

[20] M. Chen, "Evaluating Large Language Models Trained on Code," arXiv preprint arXiv:2107.03374, 2021.

[21] J. Zhang, "PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization," in Proceedings of the 37th International Conference on Machine Learning, 2020, pp. 11328-11339.

[22] X. Li, "A Survey on Deep Learning for Named Entity Recognition," IEEE Transactions on Knowledge and Data Engineering, vol. 34, no. 1, pp. 50-70, 2022.

[23] T. B. Brown, "Language Models are Few-Shot Learners," in Advances in Neural Information Processing Systems, 2020, pp. 1877-1901.

[24] A. Chowdhery, "PaLM: Scaling Language Modeling with Pathways," arXiv preprint arXiv:2204.02311, 2022.

[25] E. Strubell, A. Ganesh, and A. McCallum, "Energy and Policy Considerations for Deep Learning in NLP," in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 3645-3650.

[26] D. Patterson, "Carbon Emissions and Large Neural Network Training," arXiv preprint arXiv:2104.10350, 2021.

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

[28] NVIDIA Corporation, "FasterTransformer: Accelerated Transformer Inference on GPU," GitHub repository, 2021. [Online]. Available: https://github.com/NVIDIA/FasterTransformer

[29] R. Child, "Generating Long Sequences with Sparse Transformers," arXiv preprint arXiv:1904.10509, 2019.

[30] D. Lepikhin, "GShard: Scaling Giant Models with Conditional Computation and Automatic Sharding," arXiv preprint arXiv:2006.16668, 2020.

[31] D. R. So, C. Liang, and Q. V. Le, "The Evolved Transformer," in Proceedings of the 36th International Conference on Machine Learning, 2019, pp. 5877-5886.

[32] A. Luccioni, E. Baylor, and N. Duchene, "Analyzing Sustainability and Scaling Challenges for Transformers," in Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 2022, pp. 4529-4541.

[33] P. Rajpurkar, "AI for Global Health: Learning From a Decade of Digital Health Innovation at Stanford," arXiv preprint arXiv:2104.01915, 2021.

[34] Z. C. Lipton, "The Mythos of Model Interpretability," Queue, vol. 16, no. 3, pp. 31-57, 2018.

[35] AI Now Institute, "AI Now 2019 Report," New York University, 2019.

[36] A. Vaswani, "Attention Is All You Need," in Advances in Neural Information Processing Systems, 2017, pp. 5998-6008.

[37] J. Vig, "A Multiscale Visualization of Attention in the Transformer Model," in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 2019, pp. 37-42.

[38] O. Kovaleva, A. Romanov, A. Rogers, and A. Rumshisky, "Revealing the Dark Secrets of BERT," in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, 2019, pp. 4365-4374.

[39] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, and R. Sayres, "Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV)," in International Conference on Machine Learning, 2018, pp. 2668-2677.

[40] A. Bau, Y. Belinkov, H. Sajjad, N. Durrani, H. Dalvi, and J. Glass, "Identifying and Controlling Important Neurons in Neural Machine Translation," in International Conference on Learning Representations, 2019.

[41] J. Mu and J. Andreas, "Compositional Explanations of Neurons," in Advances in Neural Information Processing Systems, 2020, pp. 17153-17163.

[42] I. Tenney, D. Das, and E. Pavlick, "BERT Rediscovers the Classical NLP Pipeline," in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 4593-4601.

[43] J. Hewitt and C. D. Manning, "A Structural Probe for Finding Syntax in Word Representations," in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019, pp. 4129-4138.

[44] Q. Liu, "Multilingual Probing of Deep Pre-Trained Contextual Encoders," in Proceedings of the First NLPL Workshop on Deep Learning for Natural Language Processing, 2019, pp. 1-10.

[45] E. Voita, D. Talbot, F. Moiseev, R. Sennrich, and I. Titov, "Analyzing Multi-Head Self-Attention: Specialized Heads Do the Heavy Lifting, the Rest Can Be Pruned," in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 5797-5808.

[46] H. Chefer, S. Gur, and L. Wolf, "Transformer Interpretability Beyond Attention Visualization," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 782-791.

[47] D. A. Melis and T. Jaakkola, "Towards Robust Interpretability with Self-Explaining Neural Networks," in Advances in Neural Information Processing Systems, 2018, pp. 7775-7784.

[48] J. Chen, L. Song, M. Wainwright, and M. Jordan, "Learning to Explain: An Information-Theoretic Perspective on Model Interpretation," in International Conference on Machine Learning, 2018, pp. 883-892.

[49] M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-Precision Model-Agnostic Explanations," in Proceedings of the AAAI Conference on Artificial Intelligence, 2018, pp. 1527-1535.

[50] E. Real, A. Aggarwal, Y. Huang, and Q. V. Le, "Regularized Evolution for Image Classifier Architecture Search," in Proceedings of the AAAI Conference on Artificial Intelligence, 2019, pp. 4780-4789.

[51] T. Kojima, "Large Language Models are Zero-Shot Reasoners," in Advances in Neural Information Processing Systems, 2022.

[52] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" in Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 2021, pp. 610-623.

[53] S. L. Blodgett, S. Barocas, H. Daumé III, and H. Wallach, "Language (Technology) is Power: A Critical Survey of 'Bias' in NLP," in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 5454-5476.

[54] A. Abid, M. Farooqi, and J. Zou, "Persistent Anti-Muslim Bias in Large Language Models," in Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, 2021, pp. 298-306.

[55] Y. C. Tan and L. E. Celis, "Assessing Social and Intersectional Biases in Contextualized Word Representations," in Advances in Neural Information Processing Systems, 2019, pp. 13230-13241.

[56] M. Whittaker, "AI Now Report 2018," AI Now Institute, New York University, 2018.

[57] I. Y. Chen, P. Szolovits, and M. Ghassemi, "Can AI Help Reduce Disparities in General Medical and Mental Health Care?" AMA Journal of Ethics, vol. 21, no. 2, pp. 167-179, 2019.

[58] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine Bias," ProPublica, May 23, 2016.

[59] M. Raghavan, S. Barocas, J. Kleinberg, and K. Levy, "Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices," in Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020, pp. 469-481.

[60] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai, "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings," in Advances in Neural Information Processing Systems, 2016, pp. 4349-4357.

[61] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang, "Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods," in Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2018, pp. 15-20.

[62] S. Dathathri, "Plug and Play Language Models: A Simple Approach to Controlled Text Generation," in International Conference on Learning Representations, 2020.

[63] S. Krause, S. Dathathri, T. Sherborne, and K. Narasimhan, "GEDI: Generative Discriminator Guided Sequence Generation," in Findings of the Association for Computational Linguistics: EMNLP 2021, 2021, pp. 4929-4952.

[64] The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, "Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems," IEEE, 2019.

[65] High-Level Expert Group on Artificial Intelligence, "Ethics Guidelines for Trustworthy AI," European Commission, 2019.

[66] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in

Advances in Neural Information Processing Systems, 2017, pp. 4765-4774.

[67] D. Schiff, "Principles to Practices for Responsible AI: Closing the Gap," arXiv preprint arXiv:2006.04707, 2020.

[68] S. Barocas, M. Hardt, and A. Narayanan, "Fairness and Machine Learning: Limitations and Opportunities," fairmlbook.org, 2019.

[69] R. K. E. Bellamy, "AI Fairness 360: An Extensible Toolkit for Detecting and Mitigating Algorithmic Bias," IBM Journal of Research and Development, vol. 63, no. 4/5, pp. 4:1-4:15, 2019.

[70] K. Crawford, "AI Now 2019 Report," AI Now Institute, New York University, 2019.

[71] E. Strubell, A. Ganesh, and A. McCallum, "Energy and Policy Considerations for Deep Learning in NLP," in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 3645-3650.

[72] M. Sra, A. Mottelson, and P. Maes, "Your Place and Mine: Designing a Shared VR Experience for Remotely Located Users," in Proceedings of the 2018 Designing Interactive Systems Conference, 2018, pp. 85-97.

[73] I. Laponogov, "ChemListem: Chemical Named Entity Recognition Using Large Language Models," Journal of Cheminformatics, vol. 13, no. 1, pp. 1-13, 2021.

[74] J. Lu, "ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks," in Advances in Neural Information Processing Systems, 2019, pp. 13-23.

[75] W. Kim, B. Son, and I. Kim, "ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision," in Proceedings of the 38th International Conference on Machine Learning, 2021, pp. 5583-5594.

[76] A. Radford, "Learning Transferable Visual Models From Natural Language Supervision," in Proceedings of the 38th International Conference on Machine Learning, 2021, pp. 8748-8763.

[77] A. Ramesh, "Hierarchical Text-Conditional Image Generation with CLIP Latents," arXiv preprint arXiv:2204.06125, 2022.

[78] R. Rombach, "High-Resolution Image Synthesis with Latent Diffusion Models," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 10684-10695.

[79] A. H. Nguyen, T. Vu, C. Nguyen, D. Nguyen, and S. Nahavandi, "A Vision and Language-Based System for Human-Robot Collaboration," IEEE Access, vol. 9, pp. 153343-153356, 2021.

[80] D. Ahn, "Do As I Can, Not As I Say: Grounding Language in Robotic Affordances," arXiv preprint arXiv:2204.01691, 2022.

[81] H. Tan and M. Bansal, "LXMERT: Learning Cross-Modality Encoder Representations from Transformers," in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, 2019, pp. 5100-5111.

[82] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A Simple Framework for Contrastive Learning of Visual Representations," in Proceedings of the 37th International Conference on Machine Learning, 2020, pp. 1597-1607.

[83] R. Child, "Generating Long Sequences with Sparse Transformers," arXiv preprint arXiv:1904.10509, 2019.

[84] N. P. Jouppi, "In-Datacenter Performance Analysis of a Tensor Processing Unit," in Proceedings of the 44th Annual International Symposium on Computer Architecture, 2017, pp. 1-12.

[85] C. Shorten and T. M. Khoshgoftaar, "A Survey on Image Data Augmentation for Deep Learning," Journal of Big Data, vol. 6, no. 1, pp. 1-48, 2019.

[86] P. Bachman, R. D. Hjelm, and W. Buchwalter, "Learning Representations by Maximizing Mutual Information Across Views," in Advances in Neural Information Processing Systems, 2019, pp. 15535-15545.

[87] S. Deng, S. Paul, and C. Dyer, "Towards Debiasing Multimodal Transformers," arXiv preprint arXiv:2206.08571, 2022.

[88] S. Ö. Arik and T. Pfister, "TabNet: Attentive Interpretable Tabular Learning," in Proceedings of the AAAI Conference on Artificial Intelligence, 2021, pp. 6679-6687.