Language Models Can Explain Neurons In Language Models

Language models can explain neurons in language models

BEN'S BITES a daily feed of AI product launches and news

Language Models Can Explain Neurons in Language Models: Unraveling the Inner Workings of AI

Introduction:

The world of artificial intelligence is constantly evolving, with language models at the forefront of this revolution. These sophisticated systems, capable of generating human-quality text and translating languages, often feel like black boxes. But what if we could peer inside and understand how these models actually think? This post delves into the fascinating concept of using language models themselves to explain the internal mechanisms, specifically the "neurons," that drive their behavior. We'll explore how this meta-analysis offers a powerful new lens for understanding the complexities of large language models (LLMs) and paves the way for future advancements in AI research and development. We'll examine current methods, limitations, and the potential for future breakthroughs in this exciting field.

H2: Deconstructing the "Neuron" in Language Models

Before diving into self-explanation, let's clarify what we mean by "neurons" in the context of language models. Unlike biological neurons, these are not physical cells but rather computational units within the model's architecture. Typically, these are the nodes within a neural network, each receiving input, performing a calculation, and passing the result to other nodes. The intricate network of connections and interactions between these nodes allows the model to process information and generate output. These "neurons" fire with different strengths depending on the input data, forming a complex pattern of activation that underpins the model's understanding and generation of text.

H3: The Challenge of Interpretability

Understanding the internal workings of LLMs is notoriously challenging. Their immense size and complexity often make it impossible to directly trace the path of information flow to understand how a particular output is generated. This lack of transparency, often referred to as the "black box" problem, hinders further development and trust in these powerful tools. We need methods to make these models more interpretable.

H3: Language Models as Interpreters of Themselves

Here's where the intriguing concept of using language models to explain themselves comes in. By carefully designing prompts and analyzing the model's responses, researchers can attempt to decipher the internal states and processes. This approach leverages the model's own ability to process and generate language to shed light on its internal "neural" activity. Imagine asking a sophisticated language model: "Explain how your internal representations processed the sentence 'The quick brown fox jumps over the lazy dog.'" The response, while not a perfect map of the neural network's activity, could reveal valuable insights into the model's interpretation of the sentence.

H2: Methods for Self-Explanation in Language Models

Several techniques are currently being explored to facilitate this self-explanation:

H3: Prompt Engineering for Insight

Carefully crafted prompts are crucial. Instead of general questions, specific, targeted prompts focusing on particular aspects of the model's behavior can elicit more informative responses. For example, asking about the activation levels of specific "neurons" or the influence of different input words on the output can yield deeper insights.

H3: Attention Mechanisms Analysis

Many modern language models utilize "attention mechanisms," which highlight the parts of the input text that are most relevant to generating each word in the output. Analyzing these attention weights can reveal how the model focuses on different parts of the input, providing clues about its internal processing.

H3: Gradient-based Explanation Methods

By analyzing the gradients during the model's training, researchers can gain insights into how changes in input affect the model's output. This provides a quantitative measure of the influence of different neurons on the final decision.

H2: Limitations and Future Directions

While promising, this approach faces challenges. The explanations generated by the models might not always be accurate or complete. The model's ability to self-explain is dependent on its training data and architecture, potentially leading to biased or incomplete interpretations. Further research is needed to develop more robust and reliable methods for extracting meaningful information about the internal workings of LLMs. The development of more sophisticated prompt engineering techniques, advanced visualization tools, and more interpretable model architectures will be vital in overcoming these challenges.

Conclusion

The idea of using language models to explain the "neurons" within language models represents a significant step forward in AI interpretability. While challenges remain, the potential benefits are substantial. This meta-analytical approach promises a deeper understanding of these complex systems, leading to improved model design, enhanced reliability, and a greater understanding of how AI processes information. Future research in this area will undoubtedly shape the future of AI development, ultimately leading to more trustworthy and effective AI systems.

FAQs

1. Are these explanations truly accurate? The accuracy of self-explanations is currently under investigation. While promising, they are not a perfect representation of the model's internal workings and require further validation.

2. Can this help improve the performance of language models? Understanding internal processes can inform improvements in model architecture and training techniques, potentially leading to more accurate and efficient models.

3. What are the ethical implications? The ability to understand LLMs better raises important ethical considerations, particularly concerning bias and potential misuse. Careful consideration of these aspects is crucial.

4. How can I contribute to this research? Participation in open-source projects focused on AI interpretability, sharing datasets, and contributing to the development of new techniques are valuable ways to get involved.

5. What are the next major breakthroughs expected in this field? Expect advances in explainable AI (XAI) techniques specifically designed for LLMs, improved visualization tools for understanding internal representations, and a deeper theoretical understanding of the relationship between internal neural activity and model output.

language models can explain neurons in language models: Transformers for Natural Language Processing and Computer Vision Denis Rothman, 2024-02-29 The definitive guide to LLMs, from architectures, pretraining, and fine-tuning to Retrieval Augmented Generation (RAG), multimodal Generative AI, risks, and implementations with ChatGPT Plus with GPT-4, Hugging Face, and Vertex AI Key Features Compare and contrast 20+ models (including GPT-4, BERT, and Llama 2) and multiple platforms and libraries to find the right solution for your project Apply RAG with LLMs using customized texts and embeddings Mitigate LLM risks, such as hallucinations, using moderation models and knowledge bases Purchase of the print or Kindle book includes a free eBook in PDF format Book DescriptionTransformers for Natural Language Processing and Computer Vision, Third Edition, explores Large Language Model (LLM) architectures, applications, and various platforms (Hugging Face, OpenAI, and Google Vertex AI) used for Natural Language Processing (NLP) and Computer Vision (CV). The book guides you through different transformer architectures to the latest Foundation Models and Generative AI. You'll pretrain and fine-tune LLMs and work through different use cases, from summarization to implementing question-answering systems with embedding-based search techniques. You will also learn the risks of LLMs, from hallucinations and memorization to privacy, and how to mitigate such risks using moderation models with rule and knowledge bases. You'll implement Retrieval Augmented Generation (RAG) with LLMs to improve the accuracy of your models and gain greater control over LLM outputs. Dive into generative vision transformers and multimodal model architectures and build applications, such as image and video-to-text classifiers. Go further by combining different models and platforms and learning about AI agent replication. This book provides you with an understanding of transformer architectures, pretraining, fine-tuning, LLM use cases, and best practices.What you will learn Breakdown and understand the architectures of the Original Transformer, BERT, GPT models, T5, PaLM, ViT, CLIP, and DALL-E Fine-tune BERT, GPT, and PaLM 2 models Learn about different tokenizers and the best practices for preprocessing language data Pretrain a RoBERTa model from scratch Implement retrieval augmented generation and rules bases to mitigate hallucinations Visualize transformer model activity for deeper insights using BertViz, LIME, and SHAP Go in-depth into vision transformers with CLIP, DALL-E 2, DALL-E 3, and GPT-4V Who this book is for This book is ideal for NLP and CV engineers, software developers, data scientists, machine learning engineers, and technical leaders looking to advance their LLMs and generative AI skills or explore the latest trends in the field. Knowledge of Python and machine learning concepts is required to fully understand the use cases and code examples. However, with examples using LLM user interfaces, prompt engineering, and no-code model building, this book is great for anyone curious about the AI revolution.

language models can explain neurons in language models: Navigating the Circular Age of a Sustainable Digital Revolution Tanveer, Umair, Ishaq, Shamaila, Huy, Truong Quang, Hoang, Thinh Gia, 2024-08-26 In the face of rapid digitalization and environmental challenges, the world stands at a critical juncture. The relentless pace of technological advancement has brought unparalleled convenience and efficiency but has also contributed to unsustainable consumption patterns, resource depletion, and environmental degradation. Despite growing awareness, many industries need help integrating sustainable practices into their operations, hindered by a lack of understanding, resources, and clear guidelines. Moreover, the complexity of the circular economy and the ethical dimensions of digitalization pose significant challenges, requiring innovative solutions and comprehensive guidance. Navigating the Circular Age of a Sustainable Digital Revolution offers a timely and comprehensive solution to these pressing challenges. By exploring the intricate relationship between technology and sustainability, this book provides a roadmap for businesses, policymakers, and individuals to embrace sustainable practices in the digital era. Researchers and scholars gain profound insights from this book into the dynamics between digitalization and sustainable practices while policymakers find nuanced analyses to shape regulatory frameworks. Business leaders and professionals discover practical guidance for sustainable business models and digital transformation, and technology practitioners align their fields with sustainable advancements. Ultimately, the book empowers individuals and organizations to shape a future where technology and sustainability coexist, fostering a more sustainable and prosperous world.

language models can explain neurons in language models: Intelligent Systems Design and Applications Ajith Abraham,

language models can explain neurons in language models: Revolutionizing Communication Raquel V. Benítez Rojas, Francisco-Julián Martínez-Cano, 2024-10-22 Revolutionizing Communication: The Role of Artificial Intelligence explores the wide-ranging effects of artificial intelligence (AI) on how we connect and communicate, changing social interactions, relationships, and the very structure of our society. Through insightful analysis, practical examples, and knowledgeable perspectives, the book examines chatbots, virtual assistants, natural language processing, and more. It shows how these technologies have a significant impact on cultural productions, business, education, ethics, advertising, media, journalism, and interpersonal interactions. Revolutionizing Communication is a guide to comprehending the present and future of communication in the era of AI. It provides invaluable insights for professionals, academics, and everyone interested in the significant changes occurring in our digital age.

language models can explain neurons in language models: Artificial Neural Networks in Pattern Recognition Ching Yee Suen,

language models can explain neurons in language models: *Advances in Knowledge Discovery and Data Mining* De-Nian Yang,

language models can explain neurons in language models: Sprachmodelle verstehen Hans-Peter Stricker, 2024-05-30 Dieses Buch befasst sich mit Fragen rund um Sprachmodelle wie ChatGPT und um das Verstehen: Verstehen Chatbots, was wir ihnen sagen und meinen? Wie können uns Chatbots helfen, etwas besser zu verstehen - einen Text oder ein Konzept? Verstehen Sprachmodelle sich selbst - was sie sagen und warum sie es sagen? Können wir Sprachmodelle verstehen und wie? Das Buch richtet sich an technisch und philosophisch interessierte Laien, aber auch an Didaktiker aller Couleur, von der Lehrkraft bis zu Wissenschaftsjournalist:innen.

language models can explain neurons in language models: Interpretable Machine Learning Christoph Molnar, 2020 This book is about making machine learning models and their decisions interpretable. After exploring the concepts of interpretability, you will learn about simple, interpretable models such as decision trees, decision rules and linear regression. Later chapters focus on general model-agnostic methods for interpreting black box models like feature importance and accumulated local effects and explaining individual predictions with Shapley values and LIME.

All interpretation methods are explained in depth and discussed critically. How do they work under the hood? What are their strengths and weaknesses? How can their outputs be interpreted? This book will enable you to select and correctly apply the interpretation method that is most suitable for your machine learning project.

language models can explain neurons in language models: *Pattern Recognition* Ullrich Köthe,

language models can explain neurons in language models: Fundamentals of Neural Network Modeling Randolph W. Parks, Daniel S. Levine, Debra L. Long, 1998 Provides an introduction to the neural network modeling of complex cognitive and neuropsychological processes. Over the past few years, computer modeling has become more prevalent in the clinical sciences as an alternative to traditional symbol-processing models. This book provides an introduction to the neural network modeling of complex cognitive and neuropsychological processes. It is intended to make the neural network approach accessible to practicing neuropsychologists, psychologists, neurologists, and psychiatrists. It will also be a useful resource for computer scientists, mathematicians, and interdisciplinary cognitive neuroscientists. The editors (in their introduction) and contributors explain the basic concepts behind modeling and avoid the use of high-level mathematics. The book is divided into four parts. Part I provides an extensive but basic overview of neural network modeling, including its history, present, and future trends. It also includes chapters on attention, memory, and primate studies. Part II discusses neural network models of behavioral states such as alcohol dependence, learned helplessness, depression, and waking and sleeping. Part III presents neural network models of neuropsychological tests such as the Wisconsin Card Sorting Task, the Tower of Hanoi, and the Stroop Test. Finally, part IV describes the application of neural network models to dementia: models of acetycholine and memory, verbal fluency, Parkinsons disease, and Alzheimer's disease. Contributors J. Wesson Ashford, Rajendra D. Badgaiyan, Jean P. Banquet, Yves Burnod, Nelson Butters, John Cardoso, Agnes S. Chan, Jean-Pierre Changeux, Kerry L. Coburn, Jonathan D. Cohen, Laurent Cohen, Jose L. Contreras-Vidal, Antonio R. Damasio, Hanna Damasio, Stanislas Dehaene, Martha J. Farah, Joaquin M. Fuster, Philippe Gaussier, Angelika Gissler, Dylan G. Harwood, Michael E. Hasselmo, J, Allan Hobson, Sam Leven, Daniel S. Levine, Debra L. Long, Roderick K. Mahurin, Raymond L. Ownby, Randolph W. Parks, Michael I. Posner, David P. Salmon, David Servan-Schreiber, Chantal E. Stern, Jeffrey P. Sutton, Lynette J. Tippett, Daniel Tranel, Bradley Wyble

language models can explain neurons in language models: Inteligência Artificial e ChatGPT Fabrício Carraro, 2023-12-11 A Inteligência Artificial está abrindo um novo mundo. Estamos presenciando uma revolução multidisciplinar com a adoção em tempo recorde do ChatGPT, Bard, Midjourney e muitas outras. São as IAs generativas e elas já estão muito presentes no nosso dia a dia mesmo sem percebermos. Porém, para pessoas de tecnologia que querem dominar o tema, e não apenas utilizar essas ferramentas fantásticas, é preciso entender com profundidade como elas operam por baixo dos panos, como aplicar técnicas de Engenharia de Prompt e, ainda, pensar com responsabilidade nos aspectos éticos e nos desafios desta revolução. Neste livro, Fabrício Carraro explica detalhadamente como é o funcionamento por dentro dessas Inteligências Artificiais. Você partirá dos conceitos de Machine Learning, com os diferentes tipos de aprendizado, redes neurais artificiais e Deep Learning, chegando aos modelos de linguagem, como os famosos LLMs (Large Language Models), e algoritmos de Processamento de Linguagem Natural. O ChatGPT será visto com especial atenção, desde como foi realizado o treinamento do modelo até tópicos como tokens, temperatura, alucinações e parâmetros de calibragem da OpenAI. O livro aborda ainda as melhores práticas para gerar prompts e obter respostas mais precisas ao lidar com LLMs utilizando conceitos de Engenharia de Prompt. Por último, o autor levanta questões como segurança, direitos autorais, fake news, viés e as futuras implicações que a Inteligência Artificial pode provocar no mundo.

language models can explain neurons in language models: Speech & Language Processing Dan Jurafsky, 2000-09

language models can explain neurons in language models: Learning Deep Learning

Magnus Ekman, 2021-07-19 NVIDIA's Full-Color Guide to Deep Learning: All You Need to Get Started and Get Results To enable everyone to be part of this historic revolution requires the democratization of AI knowledge and resources. This book is timely and relevant towards accomplishing these lofty goals. -- From the foreword by Dr. Anima Anandkumar, Bren Professor, Caltech, and Director of ML Research, NVIDIA Ekman uses a learning technique that in our experience has proven pivotal to success—asking the reader to think about using DL techniques in practice. His straightforward approach is refreshing, and he permits the reader to dream, just a bit, about where DL may yet take us. -- From the foreword by Dr. Craig Clawson, Director, NVIDIA Deep Learning Institute Deep learning (DL) is a key component of today's exciting advances in machine learning and artificial intelligence. Learning Deep Learning is a complete guide to DL. Illuminating both the core concepts and the hands-on programming techniques needed to succeed, this book is ideal for developers, data scientists, analysts, and others--including those with no prior machine learning or statistics experience. After introducing the essential building blocks of deep neural networks, such as artificial neurons and fully connected, convolutional, and recurrent layers, Magnus Ekman shows how to use them to build advanced architectures, including the Transformer. He describes how these concepts are used to build modern networks for computer vision and natural language processing (NLP), including Mask R-CNN, GPT, and BERT. And he explains how a natural language translator and a system generating natural language descriptions of images. Throughout, Ekman provides concise, well-annotated code examples using TensorFlow with Keras. Corresponding PyTorch examples are provided online, and the book thereby covers the two dominating Python libraries for DL used in industry and academia. He concludes with an introduction to neural architecture search (NAS), exploring important ethical issues and providing resources for further learning. Explore and master core concepts: perceptrons, gradient-based learning, sigmoid neurons, and back propagation See how DL frameworks make it easier to develop more complicated and useful neural networks Discover how convolutional neural networks (CNNs) revolutionize image classification and analysis Apply recurrent neural networks (RNNs) and long short-term memory (LSTM) to text and other variable-length sequences Master NLP with sequence-to-sequence networks and the Transformer architecture Build applications for natural language translation and image captioning NVIDIA's invention of the GPU sparked the PC gaming market. The company's pioneering work in accelerated computing--a supercharged form of computing at the intersection of computer graphics, high-performance computing, and AI--is reshaping trillion-dollar industries, such as transportation, healthcare, and manufacturing, and fueling the growth of many others. Register your book for convenient access to downloads, updates, and/or corrections as they become available. See inside book for details.

language models can explain neurons in language models: The Principles of Deep Learning Theory Daniel A. Roberts, Sho Yaida, Boris Hanin, 2022-05-26 This volume develops an effective theory approach to understanding deep neural networks of practical relevance.

language models can explain neurons in language models: Neural Machine Translation Philipp Koehn, 2020-06-18 Learn how to build machine translation systems with deep learning from the ground up, from basic concepts to cutting-edge research.

language models can explain neurons in language models: *The NEURON Book* Nicholas T. Carnevale, Michael L. Hines, 2006-01-12 The authoritative reference on NEURON, the simulation environment for modeling biological neurons and neural networks that enjoys wide use in the experimental and computational neuroscience communities. This book shows how to use NEURON to construct and apply empirically based models. Written primarily for neuroscience investigators, teachers, and students, it assumes no previous knowledge of computer programming or numerical methods. Readers with a background in the physical sciences or mathematics, who have some knowledge about brain cells and circuits and are interested in computational modeling, will also find it helpful. The NEURON Book covers material that ranges from the inner workings of this program, to practical considerations involved in specifying the anatomical and biophysical properties that are to be represented in models. It uses a problem-solving approach, with many working examples that readers can try for themselves.

language models can explain neurons in language models: Action to Language via the Mirror Neuron System Michael A. Arbib, 2006-09-07 In this book, internationally recognised experts from child development, computer science, linguistics, neuroscience, primatology and robotics discuss the role of the mirror neuron system for the recognition of hand actions and the evolutionary basis for the brain mechanisms that support language.

language models can explain neurons in language models: Deep Learning Ian Goodfellow, Yoshua Bengio, Aaron Courville, 2016-11-10 An introduction to a broad range of topics in deep learning, covering mathematical and conceptual background, deep learning techniques used in industry, and research perspectives. “Written by three experts in the field, Deep Learning is the only comprehensive book on the subject.” —Elon Musk, cochair of OpenAI; cofounder and CEO of Tesla and SpaceX Deep learning is a form of machine learning that enables computers to learn from experience and understand the world in terms of a hierarchy of concepts. Because the computer gathers knowledge from experience, there is no need for a human computer operator to formally specify all the knowledge that the computer needs. The hierarchy of concepts allows the computer to learn complicated concepts by building them out of simpler ones; a graph of these hierarchies would be many layers deep. This book introduces a broad range of topics in deep learning. The text offers mathematical and conceptual background, covering relevant concepts in linear algebra, probability theory and information theory, numerical computation, and machine learning. It describes deep learning techniques used by practitioners in industry, including deep feedforward networks, regularization, optimization algorithms, convolutional networks, sequence modeling, and practical methodology; and it surveys such applications as natural language processing, speech recognition, computer vision, online recommendation systems, bioinformatics, and videogames. Finally, the book offers research perspectives, covering such theoretical topics as linear factor models, autoencoders, representation learning, structured probabilistic models, Monte Carlo methods, the partition function, approximate inference, and deep generative models. Deep Learning can be used by undergraduate or graduate students planning careers in either industry or research, and by software engineers who want to begin using deep learning in their products or platforms. A website offers supplementary material for both readers and instructors.

language models can explain neurons in language models: Discovering the Brain National Academy of Sciences, Institute of Medicine, Sandra Ackerman, 1992-01-01 The brain ... There is no other part of the human anatomy that is so intriguing. How does it develop and function and why does it sometimes, tragically, degenerate? The answers are complex. In Discovering the Brain, science writer Sandra Ackerman cuts through the complexity to bring this vital topic to the public. The 1990s were declared the Decade of the Brain by former President Bush, and the neuroscience community responded with a host of new investigations and conferences. Discovering the Brain is based on the Institute of Medicine conference, Decade of the Brain: Frontiers in Neuroscience and Brain Research. Discovering the Brain is a field guide to the brainâ€an easy-to-read discussion of the brain's physical structure and where functions such as language and music appreciation lie. Ackerman examines: How electrical and chemical signals are conveyed in the brain. The mechanisms by which we see, hear, think, and pay attentionâ€and how a gut feeling actually originates in the brain. Learning and memory retention, including parallels to computer memory and what they might tell us about our own mental capacity. Development of the brain throughout the life span, with a look at the aging brain. Ackerman provides an enlightening chapter on the connection between the brain's physical condition and various mental disorders and notes what progress can realistically be made toward the prevention and treatment of stroke and other ailments. Finally, she explores the potential for major advances during the Decade of the Brain, with a look at medical imaging techniquesâ€what various technologies can and cannot tell usâ€and how the public and private sectors can contribute to continued advances in neuroscience. This highly readable volume will provide the public and policymakersâ€and many scientists as wellâ€with a helpful guide to understanding the many discoveries that are sure to be announced throughout the Decade of the

Brain.

language models can explain neurons in language models: Mirror Neurons and the Evolution of Brain and Language Maxim I. Stamenov, Vittorio Gallese, 2002-12-17 The emergence of language, social intelligence, and tool development are what made homo sapiens sapiens differentiate itself from all other biological species in the world. The use of language and the management of social and instrumental skills imply an awareness of intention and the consideration that one faces another individual with an attitude analogical to that of one’s own. The metaphor of ‘mirror’ aptly comes to mind.Recent investigations have shown that the human ability to ‘mirror’ other’s actions originates in the brain at a much deeper level than phenomenal awareness. A new class of neurons has been discovered in the premotor area of the monkey brain: ‘mirror neurons’. Quite remarkably, they are tuned to fire to the enaction as well as observation of specific classes of behavior: fine manual actions and actions performed by mouth. They become activated independent of the agent, be it the self or a third person whose action is observed. The activation in mirror neurons is automatic and binds the observation and enaction of some behavior by the self or by the observed other. The peculiar first-to-third-person ‘intersubjectivity’ of the performance of mirror neurons and their surprising complementarity to the functioning of strategic communicative face-to-face (first-to-second person) interaction may shed new light on the functional architecture of conscious vs. unconscious mental processes and the relationship between behavioral and communicative action in monkeys, primates, and humans. The present volume discusses the nature of mirror neurons as presented by the research team of Prof. Giacomo Rizzolatti (University of Parma), who originally discovered them, and the implications to our understanding of the evolution of brain, mind and communicative interaction in non-human primates and man.(Series B)

language models can explain neurons in language models: Jefferson Himself Thomas Jefferson, 1970

language models can explain neurons in language models: Scientific and Technical Aerospace Reports , 1994

language models can explain neurons in language models: From Neurons to Neighborhoods National Research Council, Institute of Medicine, Board on Children, Youth, and Families, Committee on Integrating the Science of Early Childhood Development, 2000-11-13 How we raise young children is one of today's most highly personalized and sharply politicized issues, in part because each of us can claim some level of expertise. The debate has intensified as discoveries about our development-in the womb and in the first months and years-have reached the popular media. How can we use our burgeoning knowledge to assure the well-being of all young children, for their own sake as well as for the sake of our nation? Drawing from new findings, this book presents important conclusions about nature-versus-nurture, the impact of being born into a working family, the effect of politics on programs for children, the costs and benefits of intervention, and other issues. The committee issues a series of challenges to decision makers regarding the quality of child care, issues of racial and ethnic diversity, the integration of children's cognitive and emotional development, and more. Authoritative yet accessible, From Neurons to Neighborhoods presents the evidence about brain wiring and how kids learn to speak, think, and regulate their behavior. It examines the effect of the climate-family, child care, community-within which the child grows.

language models can explain neurons in language models: Speaking, Writing and Communicating , 2023-05-24 Speaking, Writing and Communicating, Volume 78 in The Psychology of Learning and Motivation series, provides the latest release in this important resource that features empirical and theoretical contributions in cognitive and experimental psychology. - Presents the latest information in the highly regarded Psychology of Learning and Motivation series - Provides an essential reference for researchers and academics in cognitive science - Contains information relevant to both applied concerns and basic research

language models can explain neurons in language models: *Weakly Connected Neural Networks* Frank C. Hoppensteadt, Eugene M. Izhikevich, 2012-12-06 Devoted to local and global analysis of weakly connected systems with applications to neurosciences, this book uses bifurcation theory and canonical models as the major tools of analysis. It presents a systematic and well motivated development of both weakly connected system theory and mathematical neuroscience, addressing bifurcations in neuron and brain dynamics, synaptic organisations of the brain, and the nature of neural codes. The authors present classical results together with the most recent developments in the field, making this a useful reference for researchers and graduate students in various branches of mathematical neuroscience.

language models can explain neurons in language models: Computation, Learning, and Architectures , 1992

language models can explain neurons in language models: The Neural Simulation Language Alfredo Weitzenfeld, Michael A. Arbib, Amanda Alexander, 2002 Simulation in NSL - Modeling in NSL - Schematic Capture System - User Interface and Graphical Windows - The Modeling Language NSLM - The Scripting Language NSLS - Adaptive Resonance Theory - Depth Perception - Retina - Receptive Fields - The Associative Search Network: Landmark Learning and Hill Climbing - A Model of Primate Visual-Motor Conditional Learning - The Modular Design of the Oculomotor System in Monkeys - Crowley-Arbib Saccade Model - A Cerebellar Model of Sensorimotor Adaptation - Learning to Detour - Face Recognition by Dynamic Link Matching - Appendix I : NSLM Methods - NSLJ Extensions - NSLC Extensions - NSLJ and NSLC Differences - NSLJ and NSLC Installation Instructions.

language models can explain neurons in language models: How Smart Machines Think Sean Gerrish, 2018-10-30 Everything you've always wanted to know about self-driving cars, Netflix recommendations, IBM's Watson, and video game-playing computer programs. The future is here: Self-driving cars are on the streets, an algorithm gives you movie and TV recommendations, IBM's Watson triumphed on Jeopardy over puny human brains, computer programs can be trained to play Atari games. But how do all these things work? In this book, Sean Gerrish offers an engaging and accessible overview of the breakthroughs in artificial intelligence and machine learning that have made today's machines so smart. Gerrish outlines some of the key ideas that enable intelligent machines to perceive and interact with the world. He describes the software architecture that allows self-driving cars to stay on the road and to navigate crowded urban environments; the million-dollar Netflix competition for a better recommendation engine (which had an unexpected ending); and how programmers trained computers to perform certain behaviors by offering them treats, as if they were training a dog. He explains how artificial neural networks enable computers to perceive the world—and to play Atari video games better than humans. He explains Watson's famous victory on Jeopardy, and he looks at how computers play games, describing AlphaGo and Deep Blue, which beat reigning world champions at the strategy games of Go and chess. Computers have not yet mastered everything, however; Gerrish outlines the difficulties in creating intelligent agents that can successfully play video games like StarCraft that have evaded solution—at least for now. Gerrish weaves the stories behind these breakthroughs into the narrative, introducing readers to many of the researchers involved, and keeping technical details to a minimum. Science and technology buffs will find this book an essential guide to a future in which machines can outsmart people.

language models can explain neurons in language models: Deep Learning for Natural Language Processing Jason Brownlee, 2017-11-21 Deep learning methods are achieving state-of-the-art results on challenging machine learning problems such as describing photos and translating text from one language to another. In this new laser-focused Ebook, finally cut through the math, research papers and patchwork descriptions about natural language processing. Using clear explanations, standard Python libraries and step-by-step tutorial lessons you will discover what natural language processing is, the promise of deep learning in the field, how to clean and prepare text data for modeling, and how to develop deep learning models for your own natural language processing projects.

language models can explain neurons in language models: Supervised Sequence Labelling with Recurrent Neural Networks Alex Graves, 2012-02-06 Supervised sequence labelling is a vital area of machine learning, encompassing tasks such as speech, handwriting and gesture recognition, protein secondary structure prediction and part-of-speech tagging. Recurrent neural networks are powerful sequence learning tools—robust to input noise and distortion, able to exploit long-range contextual information—that would seem ideally suited to such problems. However their role in large-scale sequence labelling systems has so far been auxiliary. The goal of this book is a complete framework for classifying and transcribing sequential data with recurrent neural networks only. Three main innovations are introduced in order to realise this goal. Firstly, the connectionist temporal classification output layer allows the framework to be trained with unsegmented target sequences, such as phoneme-level speech transcriptions; this is in contrast to previous connectionist approaches, which were dependent on error-prone prior segmentation. Secondly, multidimensional recurrent neural networks extend the framework in a natural way to data with more than one spatio-temporal dimension, such as images and videos. Thirdly, the use of hierarchical subsampling makes it feasible to apply the framework to very large or high resolution sequences, such as raw audio or video. Experimental validation is provided by state-of-the-art results in speech and handwriting recognition.

language models can explain neurons in language models: Spiking Neuron Models Wulfram Gerstner, Werner M. Kistler, 2002-08-15 Neurons in the brain communicate by short electrical pulses, the so-called action potentials or spikes. How can we understand the process of spike generation? How can we understand information transmission by neurons? What happens if thousands of neurons are coupled together in a seemingly random network? How does the network connectivity determine the activity patterns? And, vice versa, how does the spike activity influence the connectivity pattern? These questions are addressed in this 2002 introduction to spiking neurons aimed at those taking courses in computational neuroscience, theoretical biology, biophysics, or neural networks. The approach will suit students of physics, mathematics, or computer science; it will also be useful for biologists who are interested in mathematical modelling. The text is enhanced by many worked examples and illustrations. There are no mathematical prerequisites beyond what the audience would meet as undergraduates: more advanced techniques are introduced in an elementary, concrete fashion when needed.

language models can explain neurons in language models: *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, Klaus-Robert Müller, 2019-09-10 The development of “intelligent” systems that can take decisions and perform autonomously might lead to faster and more consistent decisions. A limiting factor for a broader adoption of AI technology is the inherent risks that come with giving up human control and oversight to “intelligent” machines. For sensitive tasks involving critical infrastructures and affecting human well-being or health, it is crucial to limit the possibility of improper, non-robust and unsafe decisions and actions. Before deploying an AI system, we see a strong need to validate its behavior, and thus establish guarantees that it will continue to perform as expected when deployed in a real-world environment. In pursuit of that objective, ways for humans to verify the agreement between the AI decision structure and their own ground-truth knowledge have been explored. Explainable AI (XAI) has developed as a subfield of AI, focused on exposing complex AI models to humans in a systematic and interpretable manner. The 22 chapters included in this book provide a timely snapshot of algorithms, theory, and applications of interpretable and explainable AI and AI techniques that have been proposed recently reflecting the current discourse in this field and providing directions of future development. The book is organized in six parts: towards AI transparency; methods for interpreting AI systems; explaining the decisions of AI systems; evaluating interpretability and explanations; applications of explainable AI; and software for explainable AI.

language models can explain neurons in language models: *Crossing Borders* Bernhard Kettemann, Georg Marko, 1999

language models can explain neurons in language models: Strengthening Deep Neural Networks Katy Warr, 2019-07-03 As deep neural networks (DNNs) become increasingly common in real-world applications, the potential to deliberately fool them with data that wouldn’t trick a human presents a new attack vector. This practical book examines real-world scenarios where DNNs—the algorithms intrinsic to much of AI—are used daily to process image, audio, and video data. Author Katy Warr considers attack motivations, the risks posed by this adversarial input, and methods for increasing AI robustness to these attacks. If you’re a data scientist developing DNN algorithms, a security architect interested in how to make AI systems more resilient to attack, or someone fascinated by the differences between artificial and biological perception, this book is for you. Delve into DNNs and discover how they could be tricked by adversarial input Investigate methods used to generate adversarial input capable of fooling DNNs Explore real-world scenarios and model the adversarial threat Evaluate neural network robustness; learn methods to increase resilience of AI systems to adversarial data Examine some ways in which AI might become better at mimicking human perception in years to come

language models can explain neurons in language models: Deep Learning and the Game of Go Kevin Ferguson, Max Pumperla, 2019-01-06 Summary Deep Learning and the Game of Go teaches you how to apply the power of deep learning to complex reasoning tasks by building a Go-playing AI. After exposing you to the foundations of machine and deep learning, you'll use Python to build a bot and then teach it the rules of the game. Foreword by Thore Graepel, DeepMind Purchase of the print book includes a free eBook in PDF, Kindle, and ePub formats from Manning Publications. About the Technology The ancient strategy game of Go is an incredible case study for AI. In 2016, a deep learning-based system shocked the Go world by defeating a world champion. Shortly after that, the upgraded AlphaGo Zero crushed the original bot by using deep reinforcement learning to master the game. Now, you can learn those same deep learning techniques by building your own Go bot! About the Book Deep Learning and the Game of Go introduces deep learning by teaching you to build a Go-winning bot. As you progress, you'll apply increasingly complex training techniques and strategies using the Python deep learning library Keras. You'll enjoy watching your bot master the game of Go, and along the way, you'll discover how to apply your new deep learning skills to a wide range of other scenarios! What's inside Build and teach a self-improving game AI Enhance classical game AI systems with deep learning Implement neural networks for deep learning About the Reader All you need are basic Python skills and high school-level math. No deep learning experience required. About the Author Max Pumperla and Kevin Ferguson are experienced deep learning specialists skilled in distributed systems and data science. Together, Max and Kevin built the open source bot BetaGo. Table of Contents PART 1 - FOUNDATIONS Toward deep learning: a machine-learning introduction Go as a machine-learning problem Implementing your first Go bot PART 2 - MACHINE LEARNING AND GAME AI Playing games with tree search Getting started with neural networks Designing a neural network for Go data Learning from data: a deep-learning bot Deploying bots in the wild Learning by practice: reinforcement learning Reinforcement learning with policy gradients Reinforcement learning with value methods Reinforcement learning with actor-critic methods PART 3 - GREATER THAN THE SUM OF ITS PARTS AlphaGo: Bringing it all together AlphaGo Zero: Integrating tree search with reinforcement learning

language models can explain neurons in language models: A Thousand Brains Jeff Hawkins, 2021-03-02 A bestselling author, neuroscientist, and computer engineer unveils a theory of intelligence that will revolutionize our understanding of the brain and the future of AI. For all of neuroscience's advances, we've made little progress on its biggest question: How do simple cells in the brain create intelligence? Jeff Hawkins and his team discovered that the brain uses maplike structures to build a model of the world—not just one model, but hundreds of thousands of models of everything we know. This discovery allows Hawkins to answer important questions about how we perceive the world, why we have a sense of self, and the origin of high-level thought. A Thousand Brains heralds a revolution in the understanding of intelligence. It is a big-think book, in every sense of the word. One of the Financial Times' Best Books of 2021 One of Bill Gates' Five Favorite Books of 2021

language models can explain neurons in language models: Demystifying the Brain V. Srinivasa Chakravarthy, 2018-12-07 This book presents an emerging new vision of the brain, which is essentially expressed in computational terms, for non-experts. As such, it presents the fundamental concepts of neuroscience in simple language, without overwhelming non-biologists with excessive biological jargon. In addition, the book presents a novel computational perspective on the brain for biologists, without resorting to complex mathematical equations. It addresses a comprehensive range of topics, starting with the history of neuroscience, the function of the individual neuron, the various kinds of neural network models that can explain diverse neural phenomena, sensory-motor function, language, emotions, and concluding with the latest theories on consciousness. The book offers readers a panoramic introduction to the “new brain” and a valuable resource for interdisciplinary researchers looking to gatecrash the world of neuroscience.

language models can explain neurons in language models: Language and Cognition Kuniyoshi L. Sakai, Leonid Perlovsky, 2015-07-07 Interaction between language and cognition remains an unsolved scientific problem. What are the differences in neural mechanisms of language and cognition? Why do children acquire language by the age of six, while taking a lifetime to acquire cognition? What is the role of language and cognition in thinking? Is abstract cognition possible without language? Is language just a communication device, or is it fundamental in developing thoughts? Why are there no animals with human thinking but without human language? Combinations even among 100 words and 100 objects (multiple words can represent multiple objects) exceed the number of all the particles in the Universe, and it seems that no amount of experience would suffice to learn these associations. How does human brain overcome this difficulty? Since the 19th century we know about involvement of Broca’s and Wernicke’s areas in language. What new knowledge of language and cognition areas has been found with fMRI and other brain imaging methods? Every year we know more about their anatomical and functional/effective connectivity. What can be inferred about mechanisms of their interaction, and about their functions in language and cognition? Why does the human brain show hemispheric (i.e., left or right) dominance for some specific linguistic and cognitive processes? Is understanding of language and cognition processed in the same brain area, or are there differences in language-semantic and cognitive-semantic brain areas? Is the syntactic process related to the structure of our conceptual world? Chomsky has suggested that language is separable from cognition. On the opposite, cognitive and construction linguistics emphasized a single mechanism of both. Neither has led to a computational theory so far. Evolutionary linguistics has emphasized evolution leading to a mechanism of language acquisition, yet proposed approaches also lead to incomputable complexity. There are some more related issues in linguistics and language education as well. Which brain regions govern phonology, lexicon, semantics, and syntax systems, as well as their acquisitions? What are the differences in acquisition of the first and second languages? Which mechanisms of cognition are involved in reading and writing? Are different writing systems affect relations between language and cognition? Are there differences in language-cognition interactions among different language groups (such as Indo-European, Chinese, Japanese, Semitic) and types (different degrees of analytic-isolating, synthetic-inflected, fused, agglutinative features)? What can be learned from sign languages? Rizzolatti and Arbib have proposed that language evolved on top of earlier mirror-neuron mechanism. Can this proposal answer the unknown questions about language and cognition? Can it explain mechanisms of language-cognition interaction? How does it relate to known brain areas and their interactions identified in brain imaging? Emotional and conceptual contents of voice sounds in animals are fused. Evolution of human language has demanded splitting of emotional and conceptual contents and mechanisms, although language prosody still carries emotional content. Is it a dying-off remnant, or is it fundamental for interaction between language and cognition? If language and cognitive mechanisms differ, unifying these two contents requires motivation, hence emotions. What are these emotions? Can they be measured? Tonal languages use pitch contours for semantic contents, are there differences in language-cognition interaction among tonal and atonal languages? Are emotional differences among cultures exclusively cultural, or also depend on languages? Interaction of language and cognition is thus full of mysteries, and we encourage papers addressing any aspect of this topic.

language models can explain neurons in language models: Next Generation AI Language Models in Research Kashif Naseer Qureshi, Gwanggil Jeon, 2024-11-13 In this comprehensive and cutting-edge volume, Qureshi and Jeon bring together experts from around the world to explore the potential of artificial intelligence models in research and discuss the potential benefits and the concerns and challenges that the rapid development of this field has raised. The international chapter contributor group provides a wealth of technical information on different aspects of AI, including key aspects of AI, deep learning and machine learning models for AI, natural language processing and computer vision, reinforcement learning, ethics and responsibilities, security, practical implementation, and future directions. The contents are balanced in terms of theory, methodologies, and technical aspects, and contributors provide case studies to clearly illustrate the concepts and technical discussions throughout. Readers will gain valuable insights into how AI can revolutionize their work in fields including data analytics and pattern identification, healthcare research, social science research, and more, and improve their technical skills, problem-solving abilities, and evidence-based decision-making. Additionally, they will be cognizant of the limitations and challenges, the ethical implications, and security concerns related to language models, which will enable them to make more informed choices regarding their implementation. This book is an invaluable resource for undergraduate and graduate students who want to understand AI models, recent trends in the area, and technical and ethical aspects of AI. Companies involved in AI development or implementing AI in various fields will also benefit from the book’s discussions on both the technical and ethical aspects of this rapidly growing field.

language models can explain neurons in language models: Neural Network Methods for Natural Language Processing Yoav Goldberg, 2022-06-01 Neural networks are a family of powerful machine learning models. This book focuses on the application of neural network models to natural language data. The first half of the book (Parts I and II) covers the basics of supervised machine learning and feed-forward neural networks, the basics of working with machine learning over language data, and the use of vector-based rather than symbolic representations for words. It also covers the computation-graph abstraction, which allows to easily define and train arbitrary neural networks, and is the basis behind the design of contemporary neural network software libraries. The second part of the book (Parts III and IV) introduces more specialized neural network architectures, including 1D convolutional neural networks, recurrent neural networks, conditioned-generation models, and attention-based models. These architectures and techniques are the driving force behind state-of-the-art algorithms for machine translation, syntactic parsing, and many other applications. Finally, we also discuss tree-shaped networks, structured prediction, and the prospects of multi-task learning.

Change Gemini's language - Computer - Gemini Apps Help
Change Gemini's language You can choose the language Gemini Apps display, and in certain cases, understand in Language settings. This …

Fitbit-Hilfe - Google Help
Offizielle Fitbit-Hilfe, in der Sie Tipps und Lernprogramme zur Verwendung des Produkts sowie weitere Antworten auf häufig gestellte …

Download & use Google Translate
You can translate text, handwriting, photos, and speech in over 200 languages with the Google Translate app. You can also use Translate on …

Change your Gmail language settings
Change the language in Gmail Open Gmail. In the top right, click Settings . Click See all settings. In the "Language" section, pick a …

I want to change my Sheet language to English from the local one.
Mar 6, 2023 · Personal info > General preferences for the web > Language Check that your preferred language is listed first for ...

Change Gemini's language - Computer - Gemini Apps Help
Change Gemini's language You can choose the language Gemini Apps display, and in certain cases, understand in Language settings. This setting changes the language for the menu, notifications, and other text in Gemini Apps. It also affects the languages that you can talk to Gemini in when you say “Hey Google” or use the mic in the prompt field.

Fitbit-Hilfe - Google Help
Offizielle Fitbit-Hilfe, in der Sie Tipps und Lernprogramme zur Verwendung des Produkts sowie weitere Antworten auf häufig gestellte Fragen finden.

Download & use Google Translate
You can translate text, handwriting, photos, and speech in over 200 languages with the Google Translate app. You can also use Translate on the web.

Change your Gmail language settings
Change the language in Gmail Open Gmail. In the top right, click Settings . Click See all settings. In the "Language" section, pick a language from the drop-down menu. At the bottom of the page, click Save Changes. Type in another language Important: You can use input tools to type in languages like Hindi, Arabic, or Chinese.

I want to change my Sheet language to English from the local one.
Mar 6, 2023 · Personal info > General preferences for the web > Language Check that your preferred language is listed first for translation. The chosen language locale (country) determines the UI language of Google Docs/Sheets/etc, and if the Docs ruler is in centimeters or inches. Personal info > General preferences for the web > Input Tools Languages added here are ...

Translate written words - Computer - Google Translate Help
On your computer, open Google Translate. At the top of the screen, select the languages to translate. From: Choose a language or select Detect language . To: Select the language that you want the translation in. In the text box on the left, enter the text you want to translate. Choose what you want to do: Look up details: To check available details for each result, such as ...

Translate documents & websites - Computer - Google Help
In your browser, go to Google Translate. At the top, click Documents. Choose the languages to translate to and from. To automatically set the original language of a document, click Detect language. Click Browse your computer. Select the file you want to translate. Click Translate and wait for the document to finish translating. Click Download translation to download your ...

Change app language on your Android phone - Google Help
Change the language setting for a specific app Important: Apps that are set to follow the system default use the first supported language in the list. On your device, open your Settings app. Tap System Languages App Languages. Select the app you want to change. Choose a language.

Change Google Maps languages or domains
Change Google Maps languages or domains Google Maps automatically takes you to a country domain and shows place names in a country’s local languages. You can change the country domain or language shown in Google Maps.

Where can i do change language - Google Play Community
For detailed instructions with visual aids, you can visit the official Google Support page on changing the language in Google Play by clicking on the following link: Change your Google Play language settings I hope this helps! If you have any further questions, feel free to ask.

Back to Home