Thoughts on the Applicability of Machine Learning to Scientific Discovery and Possible Future Research Directions (Perspective)

Anonymous Author(s)
Affiliation
Address
email

Abstract

This is a short perspective paper discussing the potential use of machine learning for scientific discovery. Optimizing and streamlining the development of scientific knowledge is a critical issue for the future of humanity. In recent years, machine learning have begun to accelerate scientific progress. However, many of them have been about automation specific to individual scientific domains. In this paper, we emphasize the importance of discussing how to apply machine learning to more general scientific processes as well. We then briefly discuss some possible future research directions to automate scientific discovery by machine learning.

1 Introduction

Science has contributed greatly to the progress of humankind. Our affluent lifestyles supported by information technology, various materials, and electricity are all based on science. Scientific advances will continue to bring us a deeper understanding of nature. And it will bring further possibilities to humans that are unthinkable today. Optimization and efficiency of scientific discovery is therefore a crucial issue for the future of humanity.

Although science appears to adhere to a rigorously formulated method of knowledge production, the formulation of scientific practice is a very recent phenomenon in the history of science. For example, it is only since the 19th century that there has been active support for the hypothetico-deductive method, which is the basis of current science [1]. On the contrary, the very nature of scientific practice is still changing [2]. Indeed, we still do not know how the discovery of a good problem or hypothesis, which has supported human breakthroughs, is done, or how optimal intellectual production can be achieved [3]. There are many practices that remain in actual scientific practice that are more historically influenced than optimized for knowledge production. In this sense, the scientific process still leaves much room for optimization.

The use of machine learning technology can be a key to optimize and improve the efficiency of science. While there have been pioneering attempts to automate scientific discovery using artificial intelligence for many years [4, 5, 6], recent developments in machine learning, especially deep learning, have greatly accelerated the automation of scientific discovery. For example, scientists have applied machine learning to protein structure prediction [7, 8], physical properties prediction [9], drug discovery [10], quantum many-body problem [11], cosmology and quantum computation [12], genetics [13], nuclear fusion [14], and many other scientific domains and achieved significant results. In particular, in response to the recent COVID-19 pandemic, researchers in various fields have accelerated their efforts to apply machine learning techniques to scientific discoveries related to the COVID-19 [15].

Submitted to NeurIPS 2021 AI for Science Workshop, Attention Track.

These studies have provided very important insights and have greatly advanced science. On the other hand, the application of machine learning to individual-specific scientific problems has been the main focus. As we move forward with automation/optimization of science, the scientific community as a whole should discuss more potential applications of machine learning to more general scientific methods that are not limited to disciplines. It is also necessary to deepen the discussion among researchers on how to improve current machine learning to optimize their research. This means aiming for artificial intelligence that can do science, rather than automate individual tasks in science. This would achieve the qualitative and quantitative leaps in intellectual production that would be unprecedented.

Therefore, in this paper, I would like to offer a view on the applicability of machine learning to the scientific method more generally, and more specifically, what scientists should think about to realize artificial intelligence that can do science. In particular, we intend to discuss the gap between the current state of machine learning and the goals we should be aiming for and give some suggestions on the general direction we should be heading in.

2 Preliminary

2.1 On Research

We see research as the activity of making the unknown known. Strictly speaking, what is unknown depends on the subject who perceives it. However, if it can be immediately connected to the known, then there is no point in doing research, so in most cases, some leap of faith is required to make the unknown known. Therefore, reasoning about the unknown is necessary, which corresponds to hypothesis generation. And, when the inference to the unknown undergoes some procedure and satisfies certain criteria, the validity of the inference is justified and it becomes known. In the empirical sciences, this justification is determined by hypothesis testing and Bayesian posterior probability; in mathematics, by repeated deductive operations that are consistent with the axiomatic system; in engineering, by confirmation of the assumed behavior; in history, by the existence of multiple strong historical sources, and so on. This rigor of justification is why science is science.

Although we mentioned making the unknown known, strictly speaking, an event is not something that can be dichotomized into unknown or known. To be precise, the mainstream view is that verification is what produces a difference in the degree of belief in certainty that people have at a given point in time (Bayesian confirmation theory [16]). In other words, we may view research as an act of changing the degree of certainty (or reducing uncertainty) about an event.

In addition, although we treated the unknown as a given above, in the actual practice of science, *discovery of the unknown* itself is also an important research process. No matter how great a hypothesis you can come up with, it will not be very important research if the problem to be tackled is of little significance. For example, "What would happen if I took a step now?" and "How to realize an artificial intelligence that can use language?" are both unknowns, but the latter seems to be a much more important question for the scientific community and humanity.

Although the substance of specific hypotheses and methods of verification will differ greatly from study to study, the general framework of discovering the unknown, inference about the unknown, and justifying the inference should remain largely the same. Therefore, the process of inference about the unknown, or plausible inference [17], itself should remain the same, regardless of whether a human or a machine performs it. Of course, there can be a difference in whether each step is performed explicitly or implicitly. For example, when a machine learning model predicts a known directly from a given unknown, it looks different from the traditional explicit process of hypothesis-generating and testing. However, since the prediction is an evaluation criterion of validity, in this sense, direct prediction can be seen as an implicit inference and justification of the unknown. Since machine learning is the study of prediction and generalization, it is in principle possible for machine learning to do science.

2.2 On Scientific Discovery

In the philosophy of science, there is agreement that scientific discovery is impossible to formulate in a completely logical/formal way, while there is also a common understanding that it is possible to

evaluate its goodness in some form [18, 3]. Until the 20th century, the main debate has been about the scientific justification in the philosophy of science, but since the middle of the 20th century, a debate about scientific discovery has emerged [3]. For example, some consider discovery as abduction [19], some consider it as a search problem (problem solving) [20, 21], and some consider it as the result of social processes [22, 23], some consider it as that based on a certain strategy [24], etc.[3].

Among these, many have emphasized the importance of search problems in the automation of scientific discovery through machine learning in recent years [25, 26, 27, 28] Machine learning efficiency of search in science has been used in areas such as experimental design using Bayesian statistics [29, 30]. The idea is that multiple candidate hypotheses (or experimental conditions, etc.) are prepared in advance, and machine learning assists in determining the next efficient candidate solution to be explored. The recent claims mentioned above are not very different from these problem setups in terms of the main framework. On the other hand, recent claims seem to emphasize the importance of efficiently searching for as exhaustive and diverse solutions as possible [26, 27, 31, 28]. This is because new and important discoveries are brought about by the search for solutions that have not been explored before. Since search itself has long been studied as an important topic in machine learning, especially reinforcement learning [32], it is expected that these findings will further contribute to the efficiency of science.

Another major attempt to automate scientific discovery is the use of data from published papers. The earliest seminal work dates back to the 1980s by Swason [33]. Since then, there have been attempts to identify scientific discoveries and their key elements from the data of published papers [34, 35, 36]. In particular, in recent years, machine learning has been used to predict research trends and issues[37, 38], and automatically generate future work,[39].Research in natural language processing has developed rapidly since the advent of Transformer [40], and research using large scale data and large scale models has been [41, 42] achieving amazing results. Since the results of this research have been remarkable, we expect that research using scientific data for scientific discovery will develop further in the future.

3 Proposition

In this section, we will discuss what would be important in advancing the automation of scientific discovery regardless of a particular domain. First, we should promote the research on exploration issues, such as that described in the previous section. The exhaustive experimental conditions search does have brought about some scientific discoveries that won the Nobel Prize [26]. Since discovery inevitably involves some kind of search, in the end, research into machine learning that can do a better search will be essential to automate scientific discovery.

On the other hand, narrowing down the myriad of possible hypotheses to a finite number of candidate hypotheses before the search is inevitable for hypotheses search. Furthermore, the way for coming up with candidate hypotheses and generating the representation of them [43] is nontrivial but crucial. In other words, how to automate the stage before the search by machine learning is the key point. Nevertheless, discussion on how to automate this preliminary step seems to lack. Therefore, we propose that the machine learning research community should discuss research directions to proceed with this automation.

3.1 Analogy

One of the key research directions to advance pre-search automation would be the realization of neural networks that can learn analogies. The analogy is the application of one known relationship to another by focusing on structural similarity [44]. The analogy is important because formulating reasonable hypotheses needs to connect the structure of the target problem to a known structure.

The research community has long noted the importance of analogies [44] to human cognition [45, 46] and specifically scientific discovery [47, 48, 49, 50, 51]. Analogies have enabled several scientific discoveries [52], such as Kepler's laws [53]. On the other hand, current machine learning models have yet to be able to make analogies fully [54]. Therefore, the development of machine learning models that can think analogically would be critical for automatic scientific discovery.

Analogies are not substantially different from other types of pattern recognition, such as image recognition or language understanding, in that they are acquired inductively from data. The difference

is that they require higher-order abstract, structural, and systematic correspondences rather than lower-order perceptual correspondences. Humans would have been unable to perform plausible inference without such abstraction and symbolization because of cognitive limitations [55]. Analogies were therefore essential for humans to reason about the unknown.

Analogies are necessary not only for humans but also for machines to do science. The reason is that one of the main goals of science is to understand the laws (rules) of nature, and the ability to learn highly abstract, systematic correspondences is essential for acquiring such knowledge generalizable to a wide range of tasks. Enabling neural networks to learn targeted patterns is a major problem [56, 57] because current neural networks are known to learn spurious correlations [58, 59]. While neural networks, in principle, can make predictions directly from raw experimental data, they can easily get trapped in task-specific spurious correlations without care. In particular, one well-known method of hypothesis generation, abduction [19], has more uncertainty for inference than induction, and the effects of spurious correlations can be more severe. Although machines need not make analogies the same as humans, the ability to retain higher-order abstract correspondences, learn systematic correspondences, and learn correspondences of structure would be vital for machines to make valid inferences about the unknown.

Various studies have emerged that attempt to use machine learning to learn analogies and abstract relations [60, 61, 62, 63, 64, 65, 66, 67, 68]. Although some studies have reported that neural networks can learn abstract relations for some simple datasets, it is far from perfect analogical reasoning. In particular, to our best knowledge, few studies have applied machine learning to analogical reasoning in the scientific domain. Therefore, it is urgent to advance the research in this direction. For challenges that machine learning faces concerning analogies in general, see, for example, this paper [54].

What seems to be the leading driver in advancing the acquisition of analogous capabilities specific to the scientific domain is the creation of datasets and the development of tasks. A reason that neural networks were able to develop so rapidly in such a short period is that benchmark datasets were well-developed. For example, ImageNet [69] for image processing and benchmark tasks like GLUE [70] that measure general-purpose language capability for NLP has greatly contributed to the research community. Prior research does point out that the dataset and the way to present it matter for learning analogies as well [62]. Several data and tasks for measuring general analogical reasoning capability have been proposed [71, 72, 73, 74]. However, little research has proposed those specific to science to our best knowledge. Therefore, the machine learning research community needs to move forward with appropriate problem setting, data set creation, and benchmarking tasks for science-specific analogy reasoning.

A promising first step in this direction is an approach that uses research paper data to learn analogical relations [75, 76]. This is because there are tons of research paper data and language models have developed rapidly in recent years. For example, one previous study considers a single paper as a pair of "problem" and "solution" and proposes a search engine for scientific papers using analogies between them [76, 75]. Specifically, they train a learner that classifies the abstract sentences of a given paper into "problem," "mechanism," and "the other" (e.g., "background" or "discovery"), and then train an encoder such that the more similar the embedding vectors of the "problem" and "mechanism" sentences are, the closer they are. This learning allows us to find scientific papers with the same "problem" but different "mechanisms". Although we need to design more abstract and general tasks for more general intelligence, it would be good to first design research-specific analogy tasks in this direction.

For data-driven research automation, it would be desirable in the long run that all data from the research process, including each research task and its intermediate products, will be recorded in a structured manner on the web. In current machine learning research, researchers publish research results as a repository, but what researchers consider in the research process does not remain as data. Therefore, no data to generate a hypothesis remains, making it hard to learn data-driven hypothesis generation patterns using these data. More to the point, researchers hesitate to release rejected research hypotheses during pre-study. In other words, less negative data remains, making hypothesis data imbalanced. Other fields, such as psychology, started to store information about the research process on the web with the tool Open Science Framework (OSF)[77]. Machine learning may also benefit from using such a tool to store the research process first. Or, since machine learning research goes up to the GitHub repository in the first place, it might be a good idea to save other research processes directly on the GitHub repository as well. In any case, the machine learning research

community needs to work to establish a culture in which researchers make the research process public.

3.2 Goal-conditioned Planning

Another research direction that will advance pre-exploration automation would be to realize neural networks with the ability to develop a strategy given a goal. Given a long-term goal, humans consider steps, or sub-goals, to reach the goal from the current point. The problem to tackle in a study is the sub-goal closest to the current point in this chain of sub-goals. This repeated application of plausible inference leads to problem discovery and hypothesis generation.

Determining strategies under a given goal has been studied in reinforcement learning (RL) [78, 79, 80]. However, determining the problem in research involves a large gap between the goal and the actual problem to be addressed, and hence the reasoning that connects the two involves a great deal of uncertainty. Therefore, whether explicitly expressed or not, the agent must divide the goal into a series of sub-goals and connect them with inferences between them. The RL researchers have long studied this generation of sub-goals and the *temporal abstraction* for it [81, 82, 83, 84, 85]. While there are studies that generate sub-goals under which goals are explicitly given [86], to the best of our knowledge, there is no study yet of applying them to reasoning as is done in more abstract studies. Reinforcement learning tasks require fine control, but the reasoning in science involves more logical reasoning. The application of reinforcement learning to automated theorem proving has increased in recent years [87], and the insights gained in this research domain may be useful in the future for learning logical reasoning in the scientific domain as well. In any case, it would be necessary to design a task more specific to scientific reasoning.

In addition, the determination of the unknown that a study address critically influences the value of the study. No matter how great a hypothesis you can come up with, if the problem to be addressed is of little significance, it will not be a very important study. However, it is difficult to evaluate how important each sub-goals is in realizing a goal. Therefore, learning the patterns of the importance of the problems using machine learning is necessary. It may be useful to use the findings of studies that evaluate the impact of research from data [88] and especially from scientific papers [89].

3.3 Language Understanding

Language is one of the key elements that distinguish humans from other animals. Anything that arises in the process of reasoning, such as hypotheses generation and problem decision we mentioned above, would be best done in language whenever possible. Therefore, developing machine learning models that can understand language would matter for research automation.

While machine learning in NLP has made great strides in recent years, there are still several challenges to be resolved. One of the essentials, especially as it relates to research, is the ability to map linguistic understanding with the understanding of nature. This is because symbolic manipulation in an abstract language does not necessarily mean understanding nature. Therefore, we should put more resource on the research of *grounded language learning*, which maps language to the real world [90, 91, 92, 93, 94]. The fact that multimodal learning of images and language has led to deeper language understanding in recent years also shows the importance of incorporating information from modalities other than language [95, 96]. Slightly off-topic, but research aimed at acquiring a physical common sense is also an important area related to the realization of intelligence capable of understanding language related to science in the same sense [97].

Another essential research direction related to language understanding for science is the acquisition of systematic generalization abilities. Human intelligence can show systematic generalization from experience to unseen events, which is key for language acquisition. Furthermore, systematicity is indispensable in that it is also key to acquiring analogies, strategy formulation, and the discovery and application of natural laws. This is because these require symbol-like systematic reasoning.

Neural networks are known to be poor at this systematic generalization [98]. However, several studies have been conducted in recent years for neural networks to achieve systematic generalization [99, 100, 101, 102, 103, 104, 105, 106, 107, 108]. While simple models can do so on simple datasets [106, 107], task-specific heuristics still need to be devised for complex data sets [108]. Therefore, we should consider how to achieve a generic systematic generalization capability.

4 Discussion

In this paper, we discussed some important factors for the goal of automating science regardless of a specific domain, especially those that may be important in the process leading up to the generation of candidate hypotheses. In particular, we raised the importance of analogy, goal-conditioned strategy formulation, and language understanding for the automation of science.

On the other hand, as we explained in Section 3.1, developing the data and task for the scientific research process would be a key driver for advancing scientific discovery automation. It is also important to re-examine the nature of science and to work from a scientific/engineering perspective to understand what scientific discovery, problem decision, and hypothesis generation are like. For this purpose, we should promote automation of scientific discovery in both specific science domains and the general scientific process. Keeping discussion on what is needed for the automation/optimization of science across research areas would be worthwhile.

References

[1] Howard Sankey. Scientific method. In Martin Curd, Stathis Psillos, et al., editors, *The Routledge Companion to Philosophy of Science*, chapter 23, pages 248–258. Routledge London, 2014.

[2] Anthony JG Hey, Stewart Tansley, Kristin Michele Tolle, et al. *The fourth paradigm: data-intensive scientific discovery*, volume 1. Microsoft research Redmond, WA, 2009.

[3] Jutta Schickore. Scientific discovery. `https://plato.stanford.edu/entries/scientific-discovery/`, 2018. Accessed: 2022-09-24.

[4] Robert K Lindsay, Bruce G Buchanan, Edward A Feigenbaum, and Joshua Lederberg. Dendral: a case study of the first expert system for scientific hypothesis formation. *Artificial intelligence*, 61(2):209–261, 1993.

[5] Pat Langley, Herbert A Simon, Gary L Bradshaw, and Jan M Zytkow. *Scientific discovery: Computational explorations of the creative processes*. MIT press, 1987.

[6] Ross D King, Kenneth E Whelan, Ffion M Jones, Philip GK Reiser, Christopher H Bryant, Stephen H Muggleton, Douglas B Kell, and Stephen G Oliver. Functional genomic hypothesis generation and experimentation by a robot scientist. *Nature*, 427(6971):247–252, 2004.

[7] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

[8] Andrew W Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Alexander WR Nelson, Alex Bridgland, et al. Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792):706–710, 2020.

[9] Rampi Ramprasad, Rohit Batra, Ghanshyam Pilania, Arun Mannodi-Kanakkithodi, and Chiho Kim. Machine learning in materials informatics: recent applications and prospects. *npj Computational Materials*, 3(1):1–13, 2017.

[10] Jessica Vamathevan, Dominic Clark, Paul Czodrowski, Ian Dunham, Edgardo Ferran, George Lee, Bin Li, Anant Madabhushi, Parantu Shah, Michaela Spitzer, et al. Applications of machine learning in drug discovery and development. *Nature reviews Drug discovery*, 18(6):463–477, 2019.

[11] Giuseppe Carleo and Matthias Troyer. Solving the quantum many-body problem with artificial neural networks. *Science*, 355(6325):602–606, 2017.

[12] Giuseppe Carleo, Ignacio Cirac, Kyle Cranmer, Laurent Daudet, Maria Schuld, Naftali Tishby, Leslie Vogt-Maranto, and Lenka Zdeborová. Machine learning and the physical sciences. *Reviews of Modern Physics*, 91(4):045002, 2019.

[13] Maxwell W Libbrecht and William Stafford Noble. Machine learning applications in genetics and genomics. *Nature Reviews Genetics*, 16(6):321–332, 2015.

[14] Jonas Degrave, Federico Felici, Jonas Buchli, Michael Neunert, Brendan Tracey, Francesco Carpanese, Timo Ewalds, Roland Hafner, Abbas Abdolmaleki, Diego de Las Casas, et al. Magnetic control of tokamak plasmas through deep reinforcement learning. *Nature*, 602(7897):414–419, 2022.

[15] Connor Shorten, Taghi M Khoshgoftaar, and Borko Furht. Deep learning applications for covid-19. *Journal of big Data*, 8(1):1–54, 2021.

[16] Vincenzo Crupi. Confirmation. `https://plato.stanford.edu/entries/confirmation/`, 2020. Accessed: 2022-09-24.

[17] George Polya. *Mathematics and plausible reasoning, Volume 1: Induction and analogy in mathematics*. Princeton University Press, 1954.

[18] Thomas Nickles. Scientific method. In Martin Curd, Stathis Psillos, et al., editors, *The Routledge Companion to Philosophy of Science*, chapter 42, pages 442–451. Routledge London, 2014.

[19] Norwood Russell Hanson. *Patterns of discovery: An inquiry into the conceptual foundations of science*. CUP Archive, 1965.

[20] Herbert A Simon. Does scientific discovery have a logic? *Philosophy of science*, 40(4):471–480, 1973.

[21] Herbert Alexander Simon, Pat Langley, and Gary L Bradshaw. Scientific discovery as problem solving. 47:1–28, 1981.

[22] Thomas S Kuhn. *The structure of scientific revolutions*, volume 111. Chicago University of Chicago Press, 1970.

[23] Imre Lakatos. *The methodology of scientific research programmes: Volume 1: Philosophical papers*, volume 1. Cambridge university press, 1980.

[24] Lindley Darden. *Theory change in science: Strategies from Mendelian genetics*. Oxford University Press on Demand, 1991.

[25] Connor W Coley, Natalie S Eyke, and Klavs F Jensen. Autonomous discovery in the chemical sciences part i: Progress. *Angewandte Chemie International Edition*, 59(51):22858–22893, 2020.

[26] Hiroaki Kitano. Nobel turing challenge: creating the engine for scientific discovery. *npj Systems Biology and Applications*, 7(1):1–12, 2021.

[27] Yoshua Bengio. Machine learning for scientific discovery. `https://www.youtube.com/watch?v=NwYxZygO1XM`, 2022. Accessed: 2022-09-25.

[28] Alexander Lavin, Hector Zenil, Brooks Paige, David Krakauer, Justin Gottschlich, Tim Mattson, Anima Anandkumar, Sanjay Choudry, Kamil Rocki, Atılım Güneş Baydin, et al. Simulation intelligence: Towards a new generation of scientific methods. *arXiv preprint arXiv:2112.03235*, 2021.

[29] Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Statistical Science*, pages 273–304, 1995.

[30] Bobak Shahriari, Kevin Swersky, Ziyu Wang, Ryan P Adams, and Nando De Freitas. Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE*, 104(1):148–175, 2015.

[31] Emmanuel Bengio, Moksh Jain, Maksym Korablyov, Doina Precup, and Yoshua Bengio. Flow network based generative models for non-iterative diverse candidate generation. *Advances in Neural Information Processing Systems*, 34:27381–27394, 2021.

[32] Susan Amin, Maziar Gomrokchi, Harsh Satija, Herke van Hoof, and Doina Precup. A survey of exploration methods in reinforcement learning. *arXiv preprint arXiv:2109.00157*, 2021.

[33] Don R Swanson. Fish oil, raynaud's syndrome, and undiscovered public knowledge. *Perspectives in biology and medicine*, 30(1):7–18, 1986.

[34] Santo Fortunato, Carl T Bergstrom, Katy Börner, James A Evans, Dirk Helbing, Staša Milojević, Alexander M Petersen, Filippo Radicchi, Roberta Sinatra, Brian Uzzi, et al. Science of science. *Science*, 359(6379):eaao0185, 2018.

[35] Pankaj Agarwal and David B Searls. Can literature analysis identify innovation drivers in drug discovery? *Nature Reviews Drug Discovery*, 8(11):865–878, 2009.

[36] Naveen Donthu, Satish Kumar, Debmalya Mukherjee, Nitesh Pandey, and Weng Marc Lim. How to conduct a bibliometric analysis: An overview and guidelines. *Journal of Business Research*, 133:285–296, 2021.

[37] Dan Lahav, Jon Saad Falcon, Bailey Kuehl, Sophie Johnson, Sravanthi Parasa, Noam Shomron, Duen Horng Chau, Diyi Yang, Eric Horvitz, Daniel S Weld, et al. A search engine for discovery of scientific challenges and directions. AAAI, 2022.

[38] Tom Hope, Jason Portenoy, Kishore Vasan, Jonathan Borchardt, Eric Horvitz, Daniel S Weld, Marti A Hearst, and Jevin West. Scisight: Combining faceted navigation and research group detection for covid-19 exploratory scientific search. *arXiv preprint arXiv:2005.12668*, 2020.

[39] Qingyun Wang, Lifu Huang, Zhiying Jiang, Kevin Knight, Heng Ji, Mohit Bansal, and Yi Luan. Paperrobot: Incremental draft generation of scientific ideas. *arXiv preprint arXiv:1905.07870*, 2019.

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[41] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[42] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[43] William Bechtel and Robert C Richardson. *Discovering complexity: Decomposition and localization as strategies in scientific research*. MIT press, 1993.

[44] Paul Bartha. Analogy and analogical reasoning. `https://plato.stanford.edu/entries/reasoning-analogy/`, 2019. Accessed: 2022-09-28.

[45] Dedre Gentner. Structure-mapping: A theoretical framework for analogy. *Cognitive science*, 7(2):155–170, 1983.

[46] Keith J Holyoak. Analogical thinking and human intelligence. *Advances in the psychology of human intelligence*, 2:199–230.

[47] Mary Hesse. Models and analogies in science. *British Journal for the Philosophy of Science*, 16(62), 1965.

[48] Paul Thagard. Conceptual combination and scientific discovery. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, 1984(1):2–12, 1984.

[49] Dedre Gentner and Michael Jeziorski. The shift from metaphor to analogy in western science. *Metaphor and thought*, 447, 1993.

[50] Keith J Holyoak and Paul Thagard. *Mental leaps: Analogy in creative thought*. MIT press, 1996.

[51] Kevin Dunbar. How scientists think: On-line creativity and conceptual change in science. 1997.

[52] Douglas R Hofstadter and Emmanuel Sander. *Surfaces and essences: Analogy as the fuel and fire of thinking*. Basic books, 2013.

[53] Dedre Gentner. Analogy in scientific discovery: The case of johannes kepler. In *Model-based reasoning*, pages 21–39. Springer, 2002.

[54] Melanie Mitchell. Abstraction and analogy-making in artificial intelligence. *Annals of the New York Academy of Sciences*, 1505(1):79–101, 2021.

[55] Jacob Feldman. The simplicity principle in perception and cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, 7(5):330–340, 2016.

[56] Martin Arjovsky. *Out of distribution generalization in machine learning*. PhD thesis, New York University, 2020.

[57] Zheyan Shen, Jiashuo Liu, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. Towards out-of-distribution generalization: A survey. *arXiv preprint arXiv:2108.13624*, 2021.

[58] Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, and Debdeep Mukhopadhyay. Adversarial attacks and defences: A survey. *arXiv preprint arXiv:1810.00069*, 2018.

[59] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

[60] Scott E Reed, Yi Zhang, Yuting Zhang, and Honglak Lee. Deep visual analogy-making. *Advances in neural information processing systems*, 28, 2015.

[61] Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. A simple neural network module for relational reasoning. *Advances in neural information processing systems*, 30, 2017.

[62] Felix Hill, Adam Santoro, David Barrett, Ari Morcos, and Timothy Lillicrap. Learning to make analogies by contrasting abstract relational structure. In *International Conference on Learning Representations*, 2019.

[63] David Barrett, Felix Hill, Adam Santoro, Ari Morcos, and Timothy Lillicrap. Measuring abstract reasoning in neural networks. In *International conference on machine learning*, pages 511–520. PMLR, 2018.

[64] Xander Steenbrugge, Sam Leroux, Tim Verbelen, and Bart Dhoedt. Improving generalization for abstract reasoning tasks using disentangled feature representations. *arXiv preprint arXiv:1811.04784*, 2018.

[65] Sjoerd Van Steenkiste, Francesco Locatello, Jürgen Schmidhuber, and Olivier Bachem. Are disentangled representations helpful for abstract visual reasoning? *Advances in Neural Information Processing Systems*, 32, 2019.

[66] Kecheng Zheng, Zheng-Jun Zha, and Wei Wei. Abstract reasoning with distracting features. *Advances in Neural Information Processing Systems*, 32, 2019.

[67] Youngsung Kim, Jinwoo Shin, Eunho Yang, and Sung Ju Hwang. Few-shot visual reasoning with meta-analogical contrastive learning. *Advances in Neural Information Processing Systems*, 33:16846–16856, 2020.

[68] Asahi Ushio, Luis Espinosa-Anke, Steven Schockaert, and Jose Camacho-Collados. Bert is to nlp what alexnet is to cv: can pre-trained language models identify analogies? *arXiv preprint arXiv:2105.04949*, 2021.

[69] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

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

[71] Ke Wang and Zhendong Su. Automatic generation of raven' s progressive matrices. In *Twenty-fourth international joint conference on artificial intelligence*, 2015.

[72] Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, and Song-Chun Zhu. Raven: A dataset for relational and analogical visual reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5317–5327, 2019.

[73] Sheng Hu, Yuqing Ma, Xianglong Liu, Yanlu Wei, and Shihao Bai. Stratified rule-aware network for abstract visual reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1567–1574, 2021.

[74] François Chollet. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019.

[75] Hyeonsu B Kang, Xin Qian, Tom Hope, Dafna Shahaf, Joel Chan, and Aniket Kittur. Augmenting scientific creativity with an analogical search engine. *ACM Transactions on Computer-Human Interaction*, 2022.

[76] Joel Chan, Joseph Chee Chang, Tom Hope, Dafna Shahaf, and Aniket Kittur. Solvent: A mixed initiative system for finding analogies between research papers. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–21, 2018.

[77] Erin D Foster and Ariel Deardorff. Open science framework (osf). *Journal of the Medical Library Association: JMLA*, 105(2):203, 2017.

[78] Leslie Pack Kaelbling. Learning to achieve goals. In *IJCAI*, volume 2, pages 1094–8. Citeseer, 1993.

[79] Tom Schaul, Daniel Horgan, Karol Gregor, and David Silver. Universal value function approximators. In *International conference on machine learning*, pages 1312–1320. PMLR, 2015.

[80] Minghuan Liu, Menghui Zhu, and Weinan Zhang. Goal-conditioned reinforcement learning: Problems and solutions. *arXiv preprint arXiv:2201.08299*, 2022.

[81] Richard S Sutton, Doina Precup, and Satinder Singh. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence*, 112(1-2):181–211, 1999.

[82] Doina Precup. *Temporal abstraction in reinforcement learning*. University of Massachusetts Amherst, 2000.

[83] Peter Dayan and Geoffrey E Hinton. Feudal reinforcement learning. *Advances in neural information processing systems*, 5, 1992.

[84] Jürgen Schmidhuber. Neural sequence chunkers. 1991.

[85] Alexander Sasha Vezhnevets, Simon Osindero, Tom Schaul, Nicolas Heess, Max Jaderberg, David Silver, and Koray Kavukcuoglu. Feudal networks for hierarchical reinforcement learning. In *International Conference on Machine Learning*, pages 3540–3549. PMLR, 2017.

[86] Elliot Chane-Sane, Cordelia Schmid, and Ivan Laptev. Goal-conditioned reinforcement learning with imagined subgoals. In *International Conference on Machine Learning*, pages 1430–1440. PMLR, 2021.

[87] Christian Szegedy. A promising path towards autoformalization and general artificial intelligence. In *International Conference on Intelligent Computer Mathematics*, pages 3–20. Springer, 2020.

[88] Dashun Wang and Albert-La´szlo´ Baraba´si. The science of impact. In Dashun Wang and Albert-La´szlo´ Baraba´si, editors, *The Science of Science*, chapter 23, pages 1–63. Cambridge University Press, 2021.

[89] Pavel Savov, Adam Jatowt, and Radoslaw Nielek. Identifying breakthrough scientific papers. *Information Processing & Management*, 57(2):102168, 2020.

[90] Laura Ruis, Jacob Andreas, Marco Baroni, Diane Bouchacourt, and Brenden M Lake. A benchmark for systematic generalization in grounded language understanding. *Advances in neural information processing systems*, 33:19861–19872, 2020.

[91] Maxime Chevalier-Boisvert, Dzmitry Bahdanau, Salem Lahlou, Lucas Willems, Chitwan Saharia, Thien Huu Nguyen, and Yoshua Bengio. Babyai: A platform to study the sample efficiency of grounded language learning. *arXiv preprint arXiv:1810.08272*, 2018.

[92] Cédric Colas, Tristan Karch, Olivier Sigaud, and Pierre-Yves Oudeyer. Autotelic agents with intrinsically motivated goal-conditioned reinforcement learning: a short survey. *Journal of Artificial Intelligence Research*, 74:1159–1199, 2022.

[93] Brenden M Lake and Gregory L Murphy. Word meaning in minds and machines. *Psychological review*, 2021.

[94] James L McClelland, Felix Hill, Maja Rudolph, Jason Baldridge, and Hinrich Schütze. Placing language in an integrated understanding system: Next steps toward human-level performance in neural language models. *Proceedings of the National Academy of Sciences*, 117(42):25966–25974, 2020.

[95] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

[96] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.

[97] Maarten Sap, Vered Shwartz, Antoine Bosselut, Yejin Choi, and Dan Roth. Commonsense reasoning for natural language processing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 27–33, 2020.

[98] Jerry A Fodor and Zenon W Pylyshyn. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2):3–71, 1988.

[99] Yinuo Guo, Zeqi Lin, Jian-Guang Lou, and Dongmei Zhang. Hierarchical poset decoding for compositional generalization in language. *Advances in Neural Information Processing Systems*, 33:6913–6924, 2020.

[100] Yinuo Guo, Hualei Zhu, Zeqi Lin, Bei Chen, Jian-Guang Lou, and Dongmei Zhang. Revisiting iterative back-translation from the perspective of compositional generalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7601–7609, 2021.

[101] Yu Gai, Paras Jain, Wendi Zhang, Joseph Gonzalez, Dawn Song, and Ion Stoica. Grounded graph decoding improves compositional generalization in question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1829–1838, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[102] Jonathan Herzig, Peter Shaw, Ming-Wei Chang, Kelvin Guu, Panupong Pasupat, and Yuan Zhang. Unlocking compositional generalization in pre-trained models using intermediate representations. *arXiv preprint arXiv:2104.07478*, 2021.

[103] Rajarshi Das, Manzil Zaheer, Dung Thai, Ameya Godbole, Ethan Perez, Jay Yoon Lee, Lizhen Tan, Lazaros Polymenakos, and Andrew McCallum. Case-based reasoning for natural language queries over knowledge bases. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9594–9611, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[104] Hao Zheng and Mirella Lapata. Disentangled sequence to sequence learning for compositional generalization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4256–4268, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[105] Daniel Furrer, Marc van Zee, Nathan Scales, and Nathanael Schärli. Compositional generalization in semantic parsing: Pre-training vs. specialized architectures. *arXiv preprint arXiv:2007.08970*, 2020.

[106] Santiago Ontanon, Joshua Ainslie, Zachary Fisher, and Vaclav Cvicek. Making transformers solve compositional tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3591–3607, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[107] Róbert Csordás, Kazuki Irie, and Juergen Schmidhuber. The devil is in the detail: Simple tricks improve systematic generalization of transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 619–634, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[108] Chenyao Liu, Shengnan An, Zeqi Lin, Qian Liu, Bei Chen, Jian-Guang Lou, Lijie Wen, Nanning Zheng, and Dongmei Zhang. Learning algebraic recombination for compositional generalization. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1129–1144, Online, August 2021. Association for Computational Linguistics.