Reward is Unnecessary

Abstract

In this article, we respond to the paper "Reward is Enough" by Silver et al. We hypothesize that pursuit of reward is not an optimal strategy to achieve human level A.I. We propose a simple definition of intelligence and contrast this "ideal" intelligence with other systems observed in humans, animals, and current A.I. In our formulation, the problem of intelligence can be simplified to that of a world model. This model learns the underlying joint probability distribution in the "world" it intends to model. We explore the advantages and limitations of this kind of model by idealizing it as a database search. We theorize that, in an idealized context, a world model matching our formulation acts both as a perfect general purpose data compression algorithm and is indistinguishable from an intelligent actor.

Introduction

Solving intelligence is a highly complex problem, in part because it is nearly impossible to get any significant number of people to agree about what intelligence actually means. We eliminate this dilemma by choosing to ignore any kind of consensus instead defining it as *"the ability to predict unknown information given known information"*. To put it more put it more simply, we define intelligence as a model of the world. We provide this definition this to eliminate the possibility of writing an entire article about how to achieve general intelligence without ever explicitly defining intelligence[1]. We acknowledge that many readers may find this definition simplistic and incomplete. Throughout the rest of this article we will try to make a case about why these readers are wrong. We will do this by exploring the concept of an ***Ideal World Model*** which learns to model the probabilistic relationships which define the underlying structure of any non-random data.

This type of model is not only theoretically capable of Artificial General Intelligence but can also act as an ideal general purpose data compression algorithm, these two problems being equivalent in this framework. Unfortunately, it suffers from the minor issue that due to its training complexity no one can train it. Nor will it become trainable at any point in the foreseeable future even as processing speed improves. It can, however, be approximated to a useful degree using models which can feasibly be trained. Approximations of this ideal model represent the most promising path toward human level A.I.

Existing Sequence Models Approximate Ideal World Models

Almost all data contains an underlying structure. This is to say that every piece of information in a specific sample is likely to affect the probability of every other piece. Given an image containing half of a cat, an average human will be able to predict that the other half of the image likely contains the other half of the cat. They will also be able to make reasonable predictions about the pose, fur color, number of appendages, and background information. This is because an average human has a reasonably accurate model of the world which allows them to extrapolate from incomplete information. This model happens to include information about cats, the environments which they are found in, and the visual appearance of both in three-dimensional space. This ability to model the world underpins all intelligent behavior from the formation of abstract concepts to problem solving.

An ***Ideal World Model (IWM)*** is a theoretical type of model which represents an unholy attempt to combine the properties of a GAN and a masked language model. IWMs predict the relative

likelihoods of different possible permutation of an arbitrary set of positions within a sample given the values of a different arbitrary set of positions. The goal is to directly model the joint probability distribution using a semi-supervised framework. This results in a model which can produce the most likely completion for any unknown part of a sample given any other part of that sample.

For example, given an image with an arbitrary set of missing pixels, a perfect IWM would be able to predict the relative probabilities of every possible permutation of pixels in the missing set given any subset of the known pixels. The reason why this type of model is entirely theoretical and will stay that way is due to its training complexity. For a dataset with samples consisting of N tokens the IWM must learn to predict the probabilities of all subsets of a set of size N given all subsets of a set of size N. The naive training complexity is therefore 2^{2N} per training example. This means that the number of permutations per training example exceeds one trillion when N=20.

While a realistic attempt to train this kind of model would not necessarily need to explore all 2^{2N} perumtations, exploring a significant proportion of them would be necessary to achieve a useful model. Thus, the exact implementation details of this kind of model are not extremely important given that we aren't actually attempting to train one. However, just because the model is untrainable doesn't mean it's useless, at least from conceptual standpoint.

Successful existing transformer sequence models represent useful approximations IWMs. Encoder only Masked Language Models (MLMs) such as BERT[2] mask (and sometimes corrupt) a limited proportion of tokens during each training step and learn to predict their independent probability given the set of unmasked tokens. Decoder only models such as GPT-3[3] learn to model the relationship between the set of all previous tokens and the next token in the sequence. GPT-3's training scheme results in N permutations per training example. BERT's theoretical training complexity is much higher but its exclusive focus on permutations where a supermajority of tokens are unmasked allows BERT to avoid the most difficult cases. Prediction becomes progressively easier as more information is present. This allows BERT-like models to achieve useful results in specific tasks without ever exploring anywhere near all possible permutations of their training sequences. There is nothing theoretically stopping a masked language model from being used for GPT-like text generation. However, due to the lack of exploration of cases where most or nearly all tokens are masked, performance of MLMs on text generation tasks is usually terrible.

In both the cases of GPT-like models and BERT-like models the result is a useful, but limited, approximation of a full IWM. This approximation is capable of a certain subset of the tasks which fully trained and properly functioning IWM would be able to do. However, both types of models are also far more useful than an IWM because they can actually be trained in a non-theoretical context. The utility of exploring IWMs is not derived from the usefulness of a model which cannot be trained. It is derived from a better understanding of the what existing models are approximating and how these approximations can be improved.

Ideal World Models Approximate a Database Search

IWMs are, themselves, approximations of the "world" of data which they model. This world is the underlying joint probability distribution which defines the spatiotemporal structure of the data. We will define the world **W** as a multiset containing an infinite number of non-unique samples matching the distribution found in the world we are attempting to model.

W represents a database of not only every sample which exists in the real world, but also every sample which could exist. With this infinite database in hand, we no longer even need a model to predict the probability of some set of positions

given some other set of positions. We can just do a database search and count the results.

For those who are unfamiliar with it, image inpainting is a sub-field of image generation. The goal of image inpainting is to fill in missing pixels in an incomplete image by inferring them from the non-missing pixels. It also provides an excellent explanatory example of why it is useful to consider this problem from the perspective of a database search on **W**.

Given an image with missing patches of pixels, conducting a search of our infinite database **W** for all images which match the non-missing pixels will return an infinite number of images. However, these matching images will also contain every possible permutation of the unknown pixels. The number of matching images is infinite but the number of permutations is finite and some permutations are more common than others. The distribution of images in **W** perfectly matches the underlying distribution of all possible images and the images in it may be duplicated an infinite number of times. This means that the most likely completion of an image given no prior information is the most common image in **W**. This also means that there exists a permutation within the images matching the known pixels which is most frequent and therefore is the most likely completion of the image. This completion of the image represents the theoretically optimal solution for this problem.

IWMs approximate **W** by modeling the relationships between every subset of positions and every other subset of positions. A dataset with an infinite number of rows but a finite number of discrete features can be represented in finite storage by calculating the conditional probabilities of every feature given every other set of features and storing the calculated probabilities in a huge table. This works because the amount of variation present in the data is finite. This process records all variation found in the data which allows for lossless recreation of the original data. For this reason, perfectly training an IWM on an infinite dataset is essentially the same as memorizing the whole dataset by converting it to a compressed representation. Complete samples can be extracted from this compressed representation using partial samples as keys. Querying an IWM which perfectly models **W** and querying **W** directly are therefore equivalent when calculating conditional probability.

For non-infinite datasets, generalization occurs because of this compression. Simpler representations are forced to model more relationships and memorize fewer. More compact representations of the data are likely to generalize better. However, poor generalization only becomes a problem for models which can store more variation than exists in the training data. The conceptual table of all possible relationships found in real world data is comically large. While the bias variance tradeoff technically still exists for this model type, too much variance is unlikely to ever be a problem when training models which approximate IWMs on any non-trivial datasets. This is likely why increasing parameter counts continues to improve performance in large sequence models[4].

Problem Solving Without Reward

Current A.I. solutions to problem solving often make use of reinforcement learning. This is a reasonable choice. After all, it is vaguely equivalent to what is used by essentially every biological intelligence to encourage useful behavior and discourage harmful behavior. Given that human intelligence relies heavily on a reward system and that the goal of research into General A.I. is to achieve human level intelligence, it's obvious why people attempt to emulate it.

The main issue with this approach is that human intelligence is neither very general nor easily re-creatable because results from two successive levels of very complex and poorly understood inductive bias. The first level involves the pre-defined neurodevelopmental structure imposed by biology on human development. The second involves the complex system of rewards through

which biology incentivizes and discourages specific behaviors. Both are opaque and incomprehensibly complicated after being shaped by millions of years of natural selection. They do not come with documentation.

Biological systems favor this structure for intelligence because it is an efficient way to achieve Darwinian goals. Energy is generally not wasted processing information and acquiring skills which do not provide a fitness advantage for the organism. Performance in important areas is optimized, usually at the cost of performance in other areas. This is true for everything from insects to humans. The kinds of behavior our reward systems encourage and the types of skills our neurological structures allow us to acquire are shaped and limited by our biology. This is not the only way to achieve intelligent behavior.

We will describe systems of intelligence which pursue some reward or specific goal as **Motivated Intelligence (MI)**. In contrast, IWMs and models which approximate their behavior can be described as **Unmotivated Intelligence (UI)** because their learning is not inherently shaped by any specific goal or purpose. UI systems are conceptually a much purer form of intelligence as they are not biased to focus on anything specific. Unfortunately, compared to MI systems they also require many orders of magnitude more compute to train before they can solve the same problems because of their lack of focus. The advantage of UI systems is that they are conceptually simple to create. Existing human level MI systems, also known as humans, require much less compute to train but only work because a successful model and reward structure has been pre-defined. This was done by evolution over millions of years. If this structure cannot be extracted from human biology directly, re-creating something equivalent through simulation would require much more compute than simply training a UI system.

Problem solving in UI systems is conceptually very different from problem solving in a more conventional MI framework which humans are intrinsically familiar with. Instead of using reward to encourage specific behavior, we conceptualize problem solving as essentially equivalent to the previously mentioned problem of image inpainting.

Consider the case of an image divided into three sections where the first and third section are present but the middle is missing. To fill in the missing section, an image inpainter must simultaneously consider both the left and right sections of the image. This is because the middle section must join the two together in a seamless way.

Images are 2D but the same framework can be applied to data with any number of dimensions if it has an underlying spatiotemporal structure. The universe itself can be considered as a static 4D volume where future and past are no more different than up and down. Solving a problem can therefore be conceptualized as predicting the series of actions necessary to connect some initial state to a desired state within this volume. Future actions are simply unknown values to be predicted given that the state resulting from them must seamlessly connect the initial and desired states. This framework can be applied to any problem where a desired state can be defined using the data.

Previous work which leverages this concept exists. The Decision Transformer[5] represents a GPT-like approximation of an IWM applied to tasks typically handled by reinforcement learning. It works by encoding the world state, action, and the difference between the desired score and the current score at each time step. Learning to generate sequences matching this formulation allows a score target to be set at inference time. The authors refer to this desired state and progress towards it as “reward” but this is not conceptually accurate and has nothing in common with reward in a conventional reinforcement learning context.

Front loading the end goal is what allows GPT-like models to fill in the missing information in-between the initial and desired state. This is

conceptually similar to image generation in DALL-E[6]. DALL-E models sequences consisting of both natural language text descriptions and the images which they are describing. The result is a model which generates images based on natural language descriptions. Similarly, desired states are not limited to exact environment states. High level natural language descriptions can be used to define them.

Some minor caveats exist. Whereas in the image generation example the model is entirely in control of all information used to bridge the gap, when acting on an environment this is not necessarily true. Acting within a complex environment with incomplete information necessitates incorporating new information in real time. While the model would be perfectly capable of predicting a series of environmental states and actions which would seamlessly connect the initial and desired state, there is no guarantee that series of events would materialize in practice. This does not present a major problem but it does limit the number of actions which can be filled in simultaneously. As the future environment state is out of the model's control and is likely not perfectly predictable, the model is usually limited to operating in an essentially autoregressive context to incorporate new information at each new time step.

There is one critical limitation which must be understood when using this type of model to solve problems. To illustrate, consider the problem of chess. Playing chess by searching **W** is relatively straightforward. At each new move, we search **W** for a game matching all prior moves, where the opponent had an arbitrarily high ELO rating, and where our side was ultimately victorious. From these winning games, we select the most frequent next move. This move is the move which is most associated with victory when playing against opponents of the highest possible skill level for the game currently being played.

The issue is that extremely high ELO players represent a tiny proportion of all chess players. Failing to include the ELO rating of the opponent would mean that games resulting from our query would be sampled from games against opponents of all skill levels. This means that in this case there is no guarantee that the move most associated with victory in this case is ideal or even good, only that it was good enough to win against the typical opponent.

IWMs model their environments and make no distinctions between environmental and adversarial obstacles. Opponents and their behaviors are simply part of the environment to be predicted. In the absence of information about the opponent an IWM will initially assume the that the opponent is typical in every way. This may result in sub-optimal initial behavior which may be unrecoverable against a sufficiently skilled opponent.

Ideal World Models Are Optimal Data Compression Algorithms

IWMs have one final interesting property which, while not directly related to problem solving, is significant enough to discuss. They function as ideal data compression algorithms. Data compression algorithms attempt to reduce the number of bits necessary to represent information. They can be both lossy and lossless. Lossless data compression allows for exact re-creation of the original data whereas lossy data compression attempts to preserve conceptually important information. The minimum amount of information necessary to represent any data is the minimum amount of information from which all other information in that data can be inferred. IWMs represent an optimal solution to lossless general purpose data compression while maintaining perfect perceptual quality in lossy compression.

Consider the case of image compression. For the sake of explanation, we will use pixels even though it would technically be optimal to use individual bits. Querying **W** allows for retrieval of complete samples by using incomplete samples as queries. Thus, there exists a minimum

subset of the pixels in any image which can be used to retrieve the full image. For this query, the most likely completion of the image is the original image itself. This set of pixels represents very close to the minimum amount of information which would be necessary to represent the image.

The same process can also be used for lossy compression but to a much higher degree. As more and more pixels are removed the reconstructed image will deviate more and more from its ground truth. However, perceptual quality will not drop. Querying **W** with the known pixels will continue to return in infinite number of realistic completions, from which we select the most likely. Images reconstructed from even a tiny subset of the pixels will therefore be high quality and coherent even if their actual contents differs significantly from the original. In an optimal lossy context, the minimum amount of information which is necessary to represent data like images, video, and audio in form perceptually equivalent to the original is astoundingly small.

Both schemes can be applied to any kind of information, including types for which lossy compression is not normally applicable. This includes text and even bit streams. This works in some cases because the reconstructed data will be coherent, even if it is not exactly correct.

Discussion

In this article, we explore IWMs and other types of world models which approximate them. Sufficiently powerful world models can solve arbitrarily difficult problems. The primary limitation to their capability is simply compute. The better they can model the world, the greater their problem-solving potential. The larger the model, the better it can model the world.

It is unclear how many orders of magnitude model complexity will need to increase to achieve human level performance in relevant tasks. It is also unclear how model structure will continue to evolve to incorporate more types of data simultaneously. For text the $O(N^2)$ complexity of vanilla transformers is acceptable but for other types of data this may not hold true. When incorporating multiple types of complex data such as video, audio, and possible actions a complexity of $O(N^2)$ is usually non-viable. Current quantization methods which alleviate this problem by reducing the number of tokens are more of a Band-Aid than a long-term solution. Even the $O(N)$ complexity of proposed efficient transformers[7] may not be sufficient in the long run. It should be noted that to succeed more efficient models do not need to perform similarly to the original on a per-parameter bases, they only need to continue scaling as more parameters are added.

References

[1] D. Silver, S. Singh, D. Precup, and R. S. Sutton, "Reward Is Enough," *Artificial Intelligence,* p. 103535, 2021.

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

[3] T. B. Brown *et al.*, "Language models are few-shot learners," *arXiv preprint arXiv:2005.14165,* 2020.

[4] J. Kaplan *et al.*, "Scaling laws for neural language models," *arXiv preprint arXiv:2001.08361,* 2020.

[5] L. Chen *et al.*, "Decision transformer: Reinforcement learning via sequence modeling," *arXiv preprint arXiv:2106.01345,* 2021.

[6] A. Ramesh *et al.*, "Zero-shot text-to-image generation," *arXiv preprint arXiv:2102.12092,* 2021.

[7] Y. Tay, M. Dehghani, D. Bahri, and D. Metzler, "Efficient transformers: A survey," *arXiv preprint arXiv:2009.06732,* 2020.

Appendix

Q&A

Q: How should I cite this paper?

A: You shouldn’t

Q: Do you have any proof for any of these claims?

A: Nope

Q: Are you absolutely sure about everything in this paper?

A: Nope

Q: Is this even an academic paper?

A: No, this is basically a blog post but I don’t have a blog

Q: What’s your favorite flavor of ice cream?

A: Mint chocolate chip