Assignment 3: 3: An Interesting transition driven by AI

Jiayu Li

February 25, 2021

Listed in the course are various recordings on AI-Driven transition in areas such as transportation/mobility, commerce, space, banking, health (click on the menu at the left of https://cybertraining-dsc.github.io/modules/ai-first/2021/course_lectures/ (Links to an external site.).

We also discussed this in the introduction to the course (week one lecture).

Choose one area from industry, or medicine, or scientific research, or consumer activities where AI is driving a transition. This could self-driving cars (industry), the design of Covid vaccines (medicine), Deep Mind's work on protein structure (science), or interpersonal interactions (consumer), giving a "random" example from four categories. Your choice does not need to be in my lectures, but it can be. Describe in 2-3 pages the nature of transition and the AI needed. This could become a final class project but needn't be!

Submit descriptions to Canvas and GitHub.

1 The Protein structure prediction problem

The problem of protein structure prediction, i.e., given amino acid sequences, calculate the three-dimensional structure of a protein composed of these amino acids.

Previous studies include Professor David Baker of the University of Washington, who developed a program called Rosetta[2] to predict protein structures. However, due to limited computational power, it is impossible to calculate the energy states of all molecules precisely. Therefore, these computational programs make many compromises; they can only calculate proteins with a small number of amino acids and a relatively simple arrangement.

2 Alpha Fold 2

On November 30, 2020, AlphaFold2[1], a program developed by DeepMind, an artificial intelligence company owned by google, scored amazingly well in CASP 14, a protein structure prediction competition in 2020. This is the first time in history that people have protein structure prediction software that is close to the level of use.

The advantage of AlphaFold is that it learns a model that reproduces the real world (protein folding) to a large extent, combining the digital world in the computer with the complex real world.

This overlap then makes the search process, which would otherwise be inching along in the real world, thousands of times faster, and not only that, it can easily introduce various search algorithms that are already in the AI to further improve the efficiency of the search. And this can all be done on a computer, without the need to operate instruments or go into a laboratory. This is something that occurs not only in biology, but also in other so-called sinkhole professions. Finding a good combination to obtain a material with a particular property, for example, again requires a lot of repetitive experiments and then a human search for better results through the years of experience of the scientists.

If a very accurate model exists, then the number of experiments in reality, can be reduced substantially, and the efficiency of the whole iteration will be a qualitative leap. And AI may be able to find some unimaginable combinations to obtain unexpected performance and also broaden the horizons of researchers. Of course the prerequisite for reaching this ideal situation is that the model should be accurate enough, and it is better not to have the loopholes of misclassification. Otherwise, once the model is used to start a search according to a certain criterion, it is entirely possible that the optimal protein sequence it gives will have an actual folding scheme that is completely different from the prediction.

I believe that the partners who do model-based RL have experience with this: look at the average error of the model is quite low, but in some states the error can be so large that the strategy trained with the learned model exploits the model's loopholes and leads to complete invalidation.

So there is actually still a long way to go, but no matter what, there is only one real world, and the model will only get better and better.

3 Conclusion

There is no doubt that the predictions of AlphaFold 2 are very impressive. But it is not a substitute for human research work. Protein structure prediction", "protein folding" and, abstracting the protein folding problem as a combinatorial optimization problem, are three completely different problems. It is a common misconception that AlphaFold 2 solves the "protein folding" problem, but we are still far from solving the real protein folding problem (including folding path, folding energy surface, folding rate, etc.). In my opinion, AlphaFold 2 will be an important tool for structural biologists rather than a replacement for structural biologists.

References

[1] M. AlQuraishi. Alphafold at casp13. *Bioinformatics*, 35(22):4862–4865, 2019.

[2] C. A. Rohl, C. E. Strauss, K. M. Misura, and D. Baker. Protein structure prediction using rosetta. *Methods in enzymology*, 383:66–93, 2004.