Final Paper

Michael Wacey

For Doctor Hamza Alsarhan

SEAS 8505: Applied Machine Intelligence

August 17, 2024

George Washington University

# Introduction

AI and Machine Learning has been around for a long time, has recently accelerated in what it can do, has been applied to many problems, and can be applied to many more problems. In this paper I look at the topics of Process, Ensemble Models, and Neural Networks as applied to Transformers. These are only a small part of the breath of AI and Machine Learning. But they are illustrative and provide a good framework. I show how these can be combined to produce real world results and discuss a recent project that I have been leading. Lastly, I talk about some of the research directions that can be taken.

# Topics

This course covered many important topics in Machine Intelligence. From these, I have selected three that I consider having both theoretical and practical significance. All of them should play a significant role long into the future. The topics are Process, Ensemble Models, and Neural Networks and Transformers.

Any tool such as Machine Learning is only useful when applied to a well-defined process. A process ensures that any issues that can come up are addressed in the most efficient way possible. Using a process over time allows for process improvement that can also add to efficiency. In addition, a well-defined process lets stake holders know what to expect and when to expect it.

In this class and previous ones, we have learned about many models. Each model has its pros and cons. By combining multiple models, it is possible to take advantage of pros while minimizing the cons. There are two additional features of ensemble models that make the particularly useful. The wisdom of the crowds is a book and the concept that many people are more effective at some tasks than one or a few experts. In addition, complex models tend to be at risk of overfitting on the training data. Ensemble methods take advantage of these two ideas by leveraging many, possibly thousands, of simpler versions of models and combining the output into one solution.

Lastly and possibly most importantly, Neural Networks and Transformers seems to be the future of Machine Learning now. There may be something that replaces transformers, but it is not currently clear what that would be. Tools like ChatGPT and Gemini have shown the power that Transformers have to do tasks that were historically considered to be in the human domain. Up until transformers, the work seemed to be to find a combination of CNN, LSTM, and Attention mechanisms. In the “Attention Is All You Need” paper, the team showed that Attention Mechanisms are powerful enough when combined in the right way. As powerful as they have proven, some of their shortcomings have been addressed through Retrieval Augmented Generation (RAG). There remain the issues of training time, training energy, training in pieces, to name a few.

# Integration and Takeaways

My observation from these three topics is that we have made great strides in several areas. We have learned how to apply the research to practical areas. Data analysis, sentiment analysis, fraud detection, and large language models (LLM) to name a few. But we still have a long way to go. There are a few tools to automate the machine learning process (e.g., AutoML). But they are new and will take time to mature. While LLMs are great at Prompt / Response interactions, they are not yet conversational. They do not ask clarifying questions, and they do not form new memories.

In my work, I had an interesting experience with leveraging Neural Networks to recognize products from pictures. Other companies in this space had trained CNNs on this task. In general, a CNN could be trained with several hundred pictures to recognize a product. Given the number of products in a typical supermarket (Over 30,000) and the rate of product label changes and new product introductions, we were not looking forward to the constant training. We needed an alternative.

We looked at the possibility of training a Neural Network to take the items on the pictures of the shelfs and look for them on the web. To do this we set up a reinforcement learning approach where there was a reward for getting the right product. The initial tests have been running for the past few weeks. We took several thousand pictures from the cough and cold aisles of Walmart’s around the country. Results have been very promising. Our next step is to include the price tags to improve the recognition and to report prices.

The above problem is a real issue for many brands across the US. The brands have no idea how much of their product or their competitors’ products are on the shelfs. There are companies that can send workers out to check. But this is very expensive for each brand. Being able to send a worker out to take photos and then process the photos can make it cost effective. This is just one example of how research is being used to solve real business problems.

# Emerging Trends

The biggest trend in machine learning these days is the application of Large Language Models and hence Transformers to almost every problem. They have caught research, business, and consumer community for the past two years. However, there are some core issues with the current set of tools. They often reflect the bias built in the data that they are trained on. They take considerable effort, time, and energy resources to train. They have significant propensity to “hallucinate”. At some point, they require complete retraining of the entire model to incorporate new data. Explainability of why a specific decision was made is often lacking. Some of these issues are being addressed by Retrieval Augmented Generation but there is still a long way to go.

I find that there are three areas where Machine Learning needs to focus. One is on improving the performance of the training. The second is on the ability to form new memories and incorporate them into the model. The last is the ability of an LLM to carry on an extended conversation. While these are not the only areas, they will go a long way to make Machine Learning more useful and flexible.

It seems to me that calling this work Machine Learning and Artificial Intelligence sets up unrealistic expectations. These are tools that can support and help human beings complete their jobs. In some ways it is very similar to how a person with a shovel can dig bigger holes than with their hands and with a backhoe can dig much bigger holes. Like any powerful tool, using Machine Learning needs to come with training and understanding of both the risks and possibilities.

A core piece of effectively using Machine Learning in real world situations is to have a well-defined process to implement them. This process should include metrics that people can understand within their domain. This will allow a buyer of machine learning offerings compare alternatives in a rational way. Research could be done to find projects and the traits that made them a success or failure. This type of research has been done in the past on large software projects to develop process and metrics that can be used.

# Conclusion

AI and Machine learning have been around for a long time. Alan Turing introduced what we now call the Turing Test in 1950. Much of Machine Learning is an outgrowth of Statistics. I found references to use of Statistics in Business as early as 1909 but could not find any academic sources to confirm this. They all seemed to cover the period from about 1995 to the present. So, todays AI and Machine learning has a long background and is not that new. We saw that when we studied the history of Artificial Neural Networks.

It is new that we have technology that can process millions or billions of neurons. That can perform extensive gradient descent across large neural networks. That can calculate the cross product of very large dimensional arrays. We have amazing power at our fingertips. In addition, papers like the “Attention Is All You Need” are showing powerful ways to organize processing to produce amazing results. This is also true in ensemble methods such as Random Forests that are producing significantly better predictions.

I take from this that we need to build on the work that has gone before us to explore new and different approaches. The two most important directions are in performance and capability. Performance will let us do more while using fewer resources. New capabilities such as forming new memories and conversation ill make them much more useful tools.

# Citations

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Advances in neural information processing systems, pages 5998-6008, 2017.

Jaime G. Carbonell, Ryszard S. Michalski, Tom M. Mitchell, 1 - AN OVERVIEW OF MACHINE LEARNING, Editor(s): Ryszard S. Michalski, Jaime G. Carbonell, Tom M. Mitchell, Machine Learning, Morgan Kaufmann, 1983, Pages 3-23, ISBN 9780080510545, <https://doi.org/10.1016/B978-0-08-051054-5.50005-4>.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in Neural Information Processing Systems, 33, 9459-9474.

Surowiecki, J. (2004). The wisdom of crowds: Why the many are smarter than the few, and how collective wisdom shapes business, economies, societies, and nations. New York, NY: Random House.

Turing, Alan (October 1950), "Computing Machinery and Intelligence" (PDF), Mind, LIX (236): 433–460, doi:10.1093/mind/LIX.236.433

Wagner, C., & Vinaimont, T. (2010). Evaluating the wisdom of crowds. *Proceedings of Issues in Information Systems*, *11*(1), 724-732.