NN and ML Proposal

Daniel, Fishbein df3622@rit.edu
Jivitesh, Debata jd9039@rit.edu

1 Motivation

As AI gets more integrated into our day to daily life it is important to understand why AI makes the decisions it makes. If there is a plane crash where people are injured or killed it is critical to know why the AI made the decisions it did. The motivation for this proposal stems from the very fact of how intelligence deals with unknowns. An emerging problem with neural nets is we can't observe the inner workings of neural nets very clearly. When trained for a specific task and then given new data that is out of its "training data set" current models give outputs "solely" dependent on its training data. To learn how to deal with new data, the model has to be "retrained". This has computation costs and in some cases, the time to retrain and deploy the model needs to happen in real-time. Humans approach these types of situations with "caution".

We define caution first as the care taken to avoid danger or mistakes. Caution is a tool that humans have used to survive dangerous situations.

In life-threatening situations, humans have exercised "caution" to minimize risk moving forward. Indian culture coins the term " jugaad" which loosely translates to jerry rigging. The idea is to do something good enough but not well when it is known that well is not possible. In situations where the risk of being wrong far outweighs the benefits of being partially right, we wish to err on the side of being right.

Qualitatively we can look at the DeepMind AlphaStar which played the competitive RTS StarCraft 2. While playing at the grandmaster level, players quickly discovered that by intentionally playing unoptimally, it was possible to trick the AI into making game-losing decisions. A simple example of this was to not attack when it was advantageous. It was discovered that if a player waited and built up defenses, eventually the AI would unoptimally attack a very defined position.

A player that intentionally played unoptimally was possible in the dataset of AlphaStar however it did not come across any in its training. Having a logical way to dismantle complex situations that were possible but not within its training data was simply beyond its capabilities.

Similarly, in FPV drones that travel at 150 miles per hour if the target object that the drone is following is lost then the current models instruct the drone to either fail graciously or try following based on motion prediction. Then as data improves more situations of losing the target are learned. But in that first instance, it doesn't correctly function as "trying" to follow that human target.

2 Expariment

In our paper, we propose a metric of the emergent quality of caution. When given a situation how reckless can the AI be to optimize its minima rather than fall back on its truths of the world to identify something safely rather than correctly.

Using this metric we can identify the AI's internal biases and why it sometimes makes seemingly erroneous decisions. We want to explore why it making hilariously awful decisions from a qualitative and quantitative perspective.

32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

Quantitatively we want to define a loss function that can reduce "success bias" in a system in favor of a more quantitatively right answer when the environment is less explored.

Our goal with this metric is to have caution on one end and recklessness/comfort on the other.

NOTE: Grammarly does not like this sentence and insists it should be: "Our goal with this metric is to have a caution on one end and recklessness/comfort on the other." This is a situation where the Grammarly AI should be cautious

Through the lens of an application like OpenSet, we have a justly difficult problem where we can repeatedly test the limits of AI in situations that have an adaptable number of outputs and predictable inputs.

A possible experiment is to train various models on MNIST with classes 0-9 as possible outputs however only classes 0-5 are given during training. During testing, all 10 numbers will be shown. We anticipate all classes 0-9 to be put in classes 0-5. Our goal of this project is to have the model identify within the first instance, that an 8 does not belong in 0-5 and if possible organize the remaining 6-9 classes without fully retaining the model.

This experiment can be expanded into color-shifted images and tiny datasets.

Other experiments could also include testing existing engines such as Chat GPT on questions that do not make sense and Grammerly on sentences that use words in unusual ways.

3 Timeline

- 7 weeks left in the semester

- 1 week to gather models we wish to test
- 1 week to make a pipeline from various models to various datasets
- 2 weeks to develop a loss function
- 1 week to organize findings and write a paper
- This leaves us with a flexible 2 weeks for extended development if some tasks take longer than expected.

References

W. J. Scheirer, A. de Rezende Rocha, A. Sapkota and T. E. Boult, "Toward Open Set Recognition," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 7, pp. 1757-1772, July 2013, doi: 10.1109/TPAMI.2012.256.

Bendale, A., Boult, T. (2015). Towards Open Set Deep Networks. arXiv. https://doi.org/10.48550/ARXIV.1511.06233

Lyu, Z., Gutierrez, N. B., Beksi, W. J. (2022). MetaMax: Improved Open-Set Deep Neural Networks via Weibull Calibration. arXiv. https://doi.org/10.48550/ARXIV.2211.10872

Xia, Z., Dong, G., Wang, P., Liu, H. (2021). Spatial Location Constraint Prototype Loss for Open Set Recognition. CoRR, abs/2110.11013. https://arxiv.org/abs/2110.11013

Arulkumaran, K., Cully, A., Togelius, J. (2019). AlphaStar: an evolutionary computation perspective. Proceedings of the Genetic and Evolutionary Computation Conference Companion.

Vaze, Sagar, et al. “Open-Set Recognition: A Good Closed-Set Classifier Is All You Need?” ArXiv.org, 13 Apr. 2022, https://arxiv.org/abs/2110.06207.