Do solutions for catastrophic forgetting degrade model robustness to naturalistic corruption and adversarial attacks?

CMSC 35200 Project Proposal

Tuan Pham tuanph18@uchicago.edu
Amit Pradhan pradhanak@uchicago.edu
Sebastian Barrios sebastian.barrios@chicagobooth.edu

Developing a deep learning model, for example in image classification, is not only about achieving the best classification accuracy. Even with increasing computational power, there are multiple issues that also need to be solved, including - to name a few - fast inference time, manageable memory size, continual learning and robustness to data/model corruption. An attractive solution for fast computation, efficient power consumption and potentially easier hardware implementation is binary neural networks (BNN) [1], in which only the signs of the model's hidden weights are utilized during inference.

A recent study takes inspiration from biological metaplasticity to solve catastrophic forgetting (CF) problems and continual learning problems for BNN [2]. More specifically, by creating a form of multiplicative gating during learning for hidden weights to represent weight consolidation, they are able to solve the permuted-MNIST task, sequential learning with CIFAR-10/100 dataset, and stream learning with these datasets. This approach shares certain similarities with another study in regular neural networks, also inspired by neuroscience literature, which takes into account (a) weight stabilization based on task importance, combined with (b) context-dependent gating allowing for more sparse, non-overlapping population activations, to facilitate learning and remembering a large number of tasks [3].

Back to BNNs, due to such quantization, they can be quite sensitive to corruption, for example data adversarial attacks [4] or potential soft errors in hardware accelerators [5]. Hence, we wish to ask whether previous solutions for CF [2] could further degrade model robustness to data corruption or help ameliorate it.

In conclusion, we propose a plan to (i) replicate the selected studies that tackle the topic of catastrophic forgetting, and (ii) observe whether solutions to avoid CF lead to models that have a higher level of degradation when different types of naturalistic corruption to input data are introduced [6]. We will focus on this form of corruption mainly; but we are also interested in the resulting vulnerability of these models to adversarial attacks [4], which we will attempt if time allows.

Plan

1. Literature dive - we will be familiarizing ourselves with literature on (a) binary networks training, (b) different solutions for catastrophic forgetting in both continuous-valued and binary NN, (c) benchmarking robustness against naturalistic image corruption, adversarial attacks, and their proposed solutions (e.g [3] for BNN)

2. Replication of known solutions to CF - we will be replicating the study for BNN to solve for (a) permuted MNIST (10 max) dataset using MLP and, if *time allows*, (b) CIFAR-10 class splitting using CNN. For (a), we will be testing at least for a few variations of the gating parameters and network sizes. We will first focus our attempt (a) using metaplasticity in BNN [2] and tentatively, if *time allows*, (b) implementing context-dependent switching and weight stabilization (partial gating, split network and context dependent gating variations).

3. Test against naturalistic corruption - for each task solved, we will be applying different forms of corruption (gaussian noise, blurring, …) on all considered tasks to observe whether solutions to CF could further increase test errors due to such corruption, or salvage it.

4. *Tentative* test against adversarial attacks - this is completely *tentative* depending on the time we have. The design of this is similar to #3, but we have to research on the specific form of adversarial attacks to focus on. Additionally, [4] proposes a Lipschitz regularization term for the hidden weight matrices to safeguard against such attacks, which we could integrate in addition to CF solutions. Such regularization also would be interesting to test again naturalistic corruption.

Side note expectations: One could hypothesize that the metaplasticity/weight stabilization solutions to CF allows the network to develop a certain form of robustness against data perturbation, and prevents neurons to develop selectivity to single classes, which could easily be quantified and tested after training if *time allows*.

If so, informed by a previous work on the tradeoff between model selectivity and robustness [7], we hypothesize that the metaplasticity solution might actually guard BNN against naturalistic corruption to some degree but the network might further suffer under adversarial attacks.

References

[1] Courbariaux, Matthieu, Itay Hubara, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. 2016. “Binarized Neural Networks: Training Deep Neural Networks with Weights and Activations Constrained to +1 or -1.” *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1602.02830.

[2] Laborieux, Axel, Maxence Ernoult, Tifenn Hirtzlin, and Damien Querlioz. 2021. “Synaptic Metaplasticity in Binarized Neural Networks.” *Nature Communications* 12 (1): 2549.

[3] Masse, Nicolas Y., Gregory D. Grant, and David J. Freedman. 2018. “Alleviating Catastrophic Forgetting Using Context-Dependent Gating and Synaptic Stabilization.” *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1802.01569.

[4] Lin, Ji, Chuang Gan, and Song Han. 2019. “Defensive Quantization: When Efficiency Meets Robustness.” *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1904.08444.

[5] Khoshavi, Navid, Connor Broyles, and Yu Bi. 2020. “Compression or Corruption? A Study on the Effects of Transient Faults on BNN Inference Accelerators.” In *2020 21st International Symposium on Quality Electronic Design (ISQED)*, 99–104.

[6] Hendrycks, Dan, and Thomas G. Dietterich. 2018. “Benchmarking Neural Network Robustness to Common Corruptions and Surface Variations.” *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1807.01697.

[7] Leavitt, Matthew L., and Ari S. Morcos. 2020. “On the Relationship between Class Selectivity, Dimensionality, and Robustness.” *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/2007.04440.

More detailed plan (tentative)

- Week 1: Read papers + receive comments on proposal, change plan as needed
- Week 2+3.5(4?): Need to coordinate and split into task
    - Implement metaplasticity
        - Develop system for permuted MNIST
        - Decide whether to do also CIFAR+convnet and/or weight stabilization
        - Would be interesting to see how difficult it is to put in an option to do Lipschitz regularization (but don’t do anything yet)
    - Figure out how to benchmark on naturalistic corruption on a simple model
        - Does it make sense for MNIST data? or do we actually have to turn to CIFAR?
        - Need decide on a system to just take in any model file + different datsets
- Week 4-6:
    - Start to integrate the models after metaplasticity to test corruption robustness