Catastrophic Forgetting versus Model Robustness in BNNs

Tuan Pham Amit Pradhan Sebastian Barrios

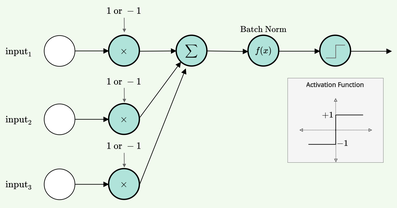
{tuanph18, pradhanak, sbarrios}@uchicago.edu

Motivation

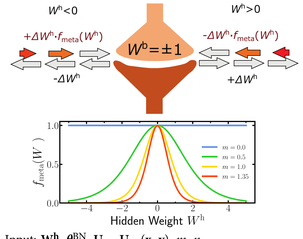
A recent study [1] takes inspiration from biological metaplasticity to solve catastrophic forgetting (CF) problems and continual learning problems for binary neural networks (BNN). More specifically, by creating a form of multiplicative gating during learning for hidden weights to represent weight consolidation, they are able to solve the permuted-MNIST task, sequential learning with CIFAR-10/100 dataset, and stream learning with these datasets. The metaplasticity method shows promise of reduced precision models in continual learning context. However, quantized neural networks are also known to be sensitive to input perturbation [2], specifically adversarial attacks [3]. This poses a question on whether binary networks trained to prevent catastrophic forgetting would generally be sensitive or robust towards different types of input perturbation, particularly natural corruptions [4] and adversarial attacks.

Binary Neural Networks (BNN)

An attractive solution for fast computation, efficient power consumption and potentially easier hardware implementation is BNN, in which only the signs of the model's hidden weights are utilized during inference.



Synaptic Metaplasticity [1]

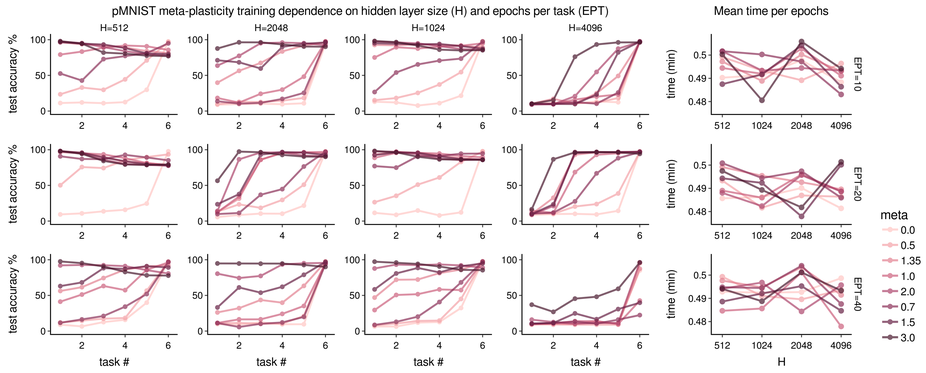


Input: $\mathbf{W^b}$, $\boldsymbol{\theta}^{\mathrm{BN}}$, $\mathbf{U_W}$, $\mathbf{U_\theta}$, $(\mathbf{x}, \mathbf{y})$, m, η.
Output: $\mathbf{W^b}$, $\boldsymbol{\theta}^{\mathrm{BN}}$, $\mathbf{U_W}$, $\mathbf{U_\theta}$.

$\mathbf{W^b} \leftarrow \mathrm{Sign}(\mathbf{W^b})$ ▷Computing binary weights
$\hat{\mathbf{y}}$, cache $\leftarrow$ Forward($\mathbf{x}$, $\mathbf{W^b}$, $\boldsymbol{\theta}^{\mathrm{BN}}$) ▷Perform inference
$C \leftarrow \mathrm{Cost}(\hat{\mathbf{y}}, \mathbf{y})$ ▷Compute mean loss over the batch
$(\partial_{\mathbf{W}} C, \partial_{\boldsymbol{\theta}} C) \leftarrow$ Backward(C, $\hat{\mathbf{y}}$, $\mathbf{W^b}$, $\boldsymbol{\theta}^{\mathrm{BN}}$, cache)
▷Cost gradients
$(\mathbf{U_W}, \mathbf{U_\theta}) \leftarrow \mathrm{Adam}(\partial_{\mathbf{W}} C, \partial_{\boldsymbol{\theta}} C, \mathbf{U_W}, \mathbf{U_\theta})$
for W^h in $\mathbf{W^h}$ **do**
if $U_W \cdot W^b > 0$ **then** ▷If U_W prescribes to decrease$|W^b|$
$W^h \leftarrow W^h - \eta U_W \cdot f_{\mathrm{meta}}(m, W^h)$ ▷Metaplastic update
else
$W^h \leftarrow W^h - \eta U_W$
end if
end for
$\boldsymbol{\theta}^{\mathrm{BN}} \leftarrow \boldsymbol{\theta}^{\mathrm{BN}} - \eta \mathbf{U_\theta}$
return $\mathbf{W^b}$, $\boldsymbol{\theta}^{\mathrm{BN}}$, $\mathbf{U_W}$, $\mathbf{U_\theta}$

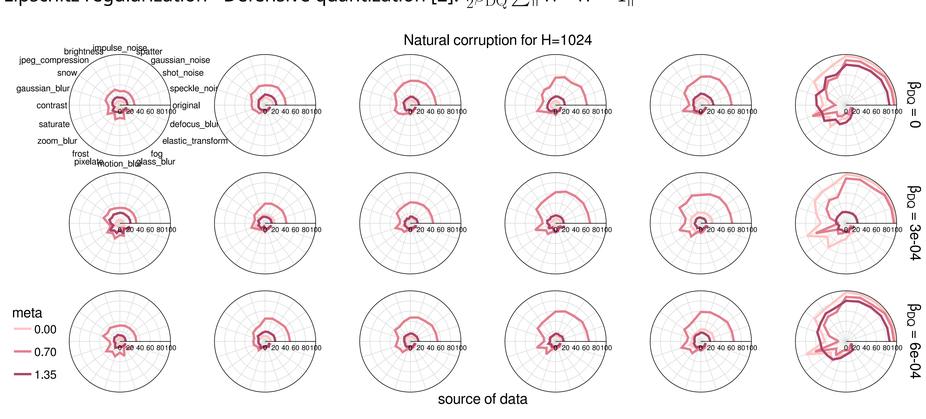
Permuted MNIST reproducibility

The values of `meta` show the effects similar to the paper: increased `meta` allows to prevent catastrophic forgetting for pMNIST task in a 2-layer MLP. However, the size effects were not reproducible as increased in size is counter-effective.

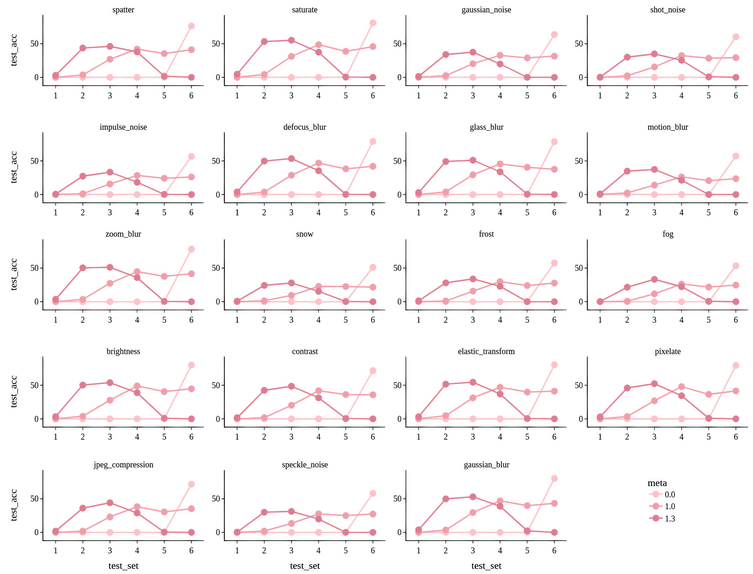


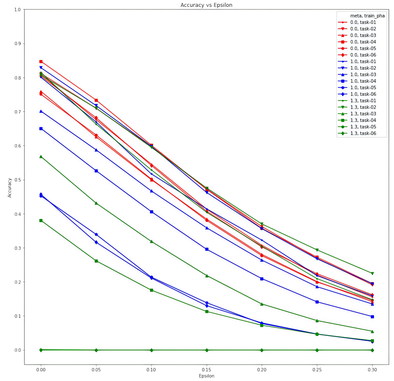
pMNIST - Robustness

Lipschitz regularization - Defensive quantization [2]: $\frac{1}{2}\beta_{\mathrm{DQ}} \sum \|\mathbf{W}^T\mathbf{W} - \mathbf{I}\|^2$



CIFAR split-class





Conclusions

While metaplasticity for BNN was useful to prevent catastrophic forgetting (pMNIST and split-class CIFAR tasks), the resulting models were not as robust. There appears to be a trade-off for continual learning and robustness - only medium values of `meta` would result in more robust models. Additionally, for the pMNIST task, Lipschitz regularization has mixed effects on model robustness for attacks and corruptions, for the pMNIST task.

However, there were also issues within the metaplasticity training: (1) dependence on the task-relevant batchnorm states, (2) irreproducible network size benefits as seen in the paper [1]. Plus, we did not have time to assess model compression benefits and speed in comparison with real-valued networks - one reason was the fact that metaplasticity was not effective in such networks. However, future experiments could involve (1) more continual learning tasks, (2) more systematic comparisons (performance, time, inference, memory cost, robustness) between different known solutions for catastrophic forgetting for neural networks across different quantization levels.

[1] A. Laborieux, M. Ernoult, T. Hirtzlin, and D. Querlioz, ``Synaptic metaplasticity in binarized neural networks,'' Nat. Commun., vol. 12, p. 2549, May 2021.

[2] J. Lin, C. Gan, and S. Han, ``Defensive quantization: When efficiency meets robustness,'' Apr. 2019.

[3] I. J. Goodfellow, J. Shlens, and C. Szegedy, ``Explaining and harnessing adversarial examples,'' Dec. 2014.

[4] D. Hendrycks and T. G. Dietterich, ``Benchmarking neural network robustness to common corruptions and surface variations,'' July 2018.