Generalizing Error Correction for NLP Neural Networks

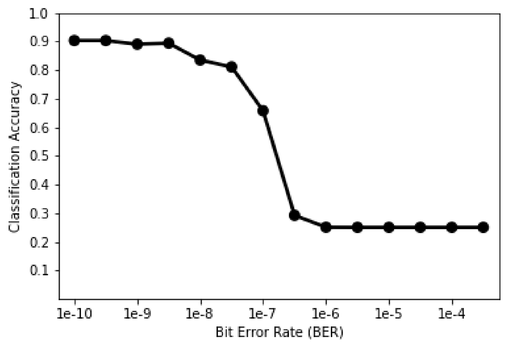
Michael Baluja, Simeng Zheng, and Paul H. Siegel University of California, San Diego

Abstract

An important concern for deep learning models is in regard to robust performance. One aspect of this involves protecting network weights from errors that occur when storing the weights in hardware. Research has been conducted on applications of selective protection for the weights of neural networks used for image classification. We extend those results for text classification through the Very Deep Convolutional Network (VDCNN), and generalize across different data sets. Experimental results have shown near-optimal performance when applying masks across different data sets. Current work looks to further understand the properties of the network weights across our different trained networks.

Introduction

It is important to ensure that the weights of a network are robust against errors that occur when storing the weights in hardware devices. One method used to ensure robustness involves using error correction in different areas of the network, such as the input or the weights. In order to ensure robustness, we utilize error correcting codes (ECCs) in order to try and detect when errors have occurred in the binary network weights, and if possible, fix the errors.
Since modern neural networks contain many millions of weights, it is important to optimize the redundancy to performance ratio by applying selective protection. An optimal trade-off can be recognized by utilizing deep reinforcement learning (DRL) to capture important areas of a network that are more prone to cause performance degradation when erroneous.
To understand noisy-performance, we present the following figure for the Floating Point Representation of the AG News data set:



Background

From Huang et al. we adopt the following conventions:

- Redundancy:

$$\frac{\text{\# parity check bits}}{\text{\# bits in representation}}$$

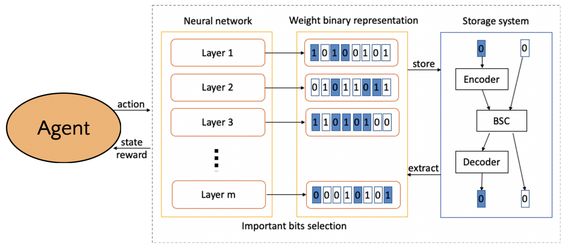
- Floating Point Weight Representation: 32-bit IEEE 754 Floating Point Standard
- Fixed Point Weight Representation: 8-bit linearly quantized binary representation

Background (Cont.)

As an indicator of cross-dataset generalizability, we define:

$$\%\ \text{loss} = 100 * (1 - \frac{\text{accuracy}}{\text{original accuracy}})$$

Prior works analyze the effectiveness of using various weight error detection/correction techniques for ensuring robust performance in image classification convolutional networks [2], [3]. We seek to understand the generalizability of these prior results for other convolutional nets, namely text classification networks, using the VDCNN present in [1]. We consider the AG News, Sogou News, and Yelp Polarity data sets, which are respectively English and Chinese News Categorization, and Sentiment Analysis sets.
The error detection/correction techniques to adopt include a DRL-based Selective Protection framework as well as weight nulling when a weight has too many errors to correct.
For the DRL framework, we present the following schematic.



Problem Statement

The weights of a neural network are prone to suffer from errors when stored in hardware. In order to ensure the weights do not degrade and cause performance issues, we must implement some selective protection to properly balance the trade-off between redundancy and network performance. This selective protection will be accomplished through the use of the noted error correction and detection techniques.

Hypothesis

We hypothesize that similar results to [2] will be achieved by our classifiers, allowing high performance for text classification networks while only selectively protecting the network weights. We also hypothesize that interchanging masks on networks trained on different data sets will ensure performance higher than non-protected noisy performance.

Procedure

To create our error-correction masks, we integrate a prebuilt DRL framework by [2]. This framework creates network masks to overlay on binary weights in order to indicate to the ECCs of which bits to protect. Under the hood, the framework consists of dual actor-critic networks using a Deep Deterministic Policy Gradient. The learning process consists of adding noise to the weights and reexamining performance. Noise addition is simulated through bit flips occurring at a preset standard Bit Error Rate (BER) of 0.01.
In order to test effectiveness of masks across data sets, we apply the mask from one data set onto a different data set with the same technique, and again validate noisy performance.

Results

Network training of AG News, Sogou News, and Yelp Polarity Review data sets on the VDCNN produced models with the following accuracies:

Dataset	Reported Accuracy	Achieved Accuracy
AG News	90.17%	90.28%
Sogou News	96.42%	95.78%
Yelp Polarity	94.73%	94.47%

To verify performance of the DRL-based framework for our use, we present the following performance results when adding noise to the AG News data set.

Protection	Representation	Code	Ratio	Accuracy
Bitmask	Float	Ideal	0.2507	90.21%
Bitmask	Float	BCH	0.2499	90.36%
Bitmask	Fixed	Ideal	0.0331	90.16%
Bitmask	Fixed	BCH	0.0397	89.93%
Topbits	Float	Ideal	0.2355	90.49%
Topbits	Float	BCH	0.2496	90.42%
Topbits	Fixed	Ideal	0.2224	90.50%
Topbits	Fixed	BCH	0.1283	90.43%

We display the following generalizability results for AG News on Sogou news (News → News), AG News on Yelp (News → Sentiment), and Yelp on AG News (Sentiment → News). Similar results appear when substituting Sogou News for AG News.

Technique	Type	Accuracy	% Loss
BitMask Float Ideal	AG Sogou	94.73%	1.10%
	AG Yelp	93.64%	0.88%
	Yelp AG	87.54%	3.03%
BitMask Float BCH	AG Sogou	95.34%	0.46%
	AG Yelp	94.14%	0.35%
	Yelp AG	88.40%	2.09%
BitMask Fixed Ideal	AG Sogou	93.42%	2.46%
	AG Yelp	93.49%	1.03%
	Yelp AG	88.96%	1.47%
BitMask Fixed BCH	AG Sogou	92.85%	3.06%
	AG Yelp	93.10%	1.45%
	Yelp AG	87.71%	2.85%
TopBits Float Ideal	AG Sogou	95.47%	0.32%
	AG Yelp	94.13%	0.36%
	Yelp AG	90.19%	0.10%
TopBits Float BCH	AG Sogou	95.66%	0.13%
	AG Yelp	94.42%	0.06%
	Yelp AG	90.12%	0.18%
TopBits Fixed Ideal	AG Sogou	95.61%	0.18%
	AG Yelp	94.41%	0.07%
	Yelp AG	89.85%	0.47%
TopBits Fixed BCH	AG Sogou	92.32%	3.61%
	AG Yelp	94.29%	0.19%
	Yelp AG	89.76%	0.57%

For all masks, their relative similarities have been calculated, and we see minimum similarities of 62.50%, maximum of 98%, and a mean of 80.76%. As an example of the relative similarities between masks, we present the following mask comparison for the TopBits Fixed Point Ideal Code Technique on the Sogou News and Yelp Polarity data sets.

Results (Cont.)

The entries represent the number of bits protected in the recorded layer.

Dataset	Layers 1-12	Layer 13	Layer 14	Layer 15
Sogou	8	1	1	7
Yelp	8	0	0	2

Conclusion

We acknowledge a few noteworthy results. The first involves the similarity of the performance between adding noise to the entire weighted network and adding noise solely to the linear layers in the network. It is important to draw attention to the delicate trade-off between where protection is given. With such a large number of weights in the linear layers, adding protection to even a single linear bit per layer may mean removing protection from entire convolutional layers earlier in the network. Due to the vast possible number of configurations for protection over these DNNs, it would likely be nearly impossible to reach a similar convergence on performance in a traditional way, and because of that, this project is a paragon for the usages of DRL.
It is also of interest to note that in many cases, we see protection for all or nearly all bits in our convolutional layers (1-12), while the linear layers (13-15) are sparsely protected.
Additionally, it is of interest to note that while the news generalizability masks seem to work well on all tested datasets, the Yelp masks do not appear to be as generalizable.
In terms of anomolous results, we note the high percent loss present in the TopBits Fixed BCH result for the AG Mask on the Sogou Network. For all other similar techniques, the average percent loss is 0.464%.
While preliminary results show high compatibility between masks of different datasets, additional work should be done in order to understand the reasoning behind the results. It is important to analyze the datasets in terms of understanding their similarities so that it is easier to understand the reasoning behind the generalizabillity.

References

[1] A. Conneau, H. Schwenk, L. Barrault, and Y. Lecun, "Very Deep Convolutional Networks for Text Classification," Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, 2017.
[2] K. Huang, P. Siegel and A. Jiang, "Functional Error Correction for Robust Neural Networks", (accepted for publication in Journal on Selected Area in Information Theory, 2020). Currently Available: https://arxiv.org/abs/2001.03814
[3] M. Qin, C. Sun, and D. Vucinic, "Robustness of Neural Networks against Storage Media Errors," Journal of Information and Graphics, vol. 6, pp. 181–186, Dec. 2018.

Acknowledgment

The work was supported by the Center for Memory and Recording Research (CMRR), Faculty Mentor Program (FMP), and the California Louis Stokes Alliance for Minority Participation (CAMP) Program at UC San Diego.