Efficient, privacy aware federated model sharing

Andreas Grammenos*
University of Cambridge
Cambridge, UK
ag926@cl.cam.ac.uk

Cecilia Mascolo*
University of Cambridge
Cambridge, UK
cm542@cl.cam.ac.uk

Jon Crowcroft*
University of Cambridge
Cambridge, UK
jac22@cl.cam.ac.uk

INTRODUCTION

Neural networks are becoming an invaluable tool for performing accurate reasoning in modern, complex datasets, yet they are extremely demanding both in terms of computation and memory requirements. Furthermore, sharing pre-trained models across different devices is becoming increasingly difficult as network size increases due to the large number of entries present in the network layers. In this work we try to address this issue by proposing a low-rank decomposition of the final network layers which has tunable approximation quality guarantees that can be used to address the aforementioned problems.

PROBLEM STATEMENT

Deep Neural networks can be viewed as a composition of several layers of transformations of the form $\mathbf{h} = g(\mathbf{vW})$, where $\mathbf{v}$ is the input vector in $\mathbb{R}^{n_v}$, $\mathbf{h}$ is the output in $\mathbb{R}^{n_h}$, and $\mathbf{W}$ is an $n_v \times n_h$ matrix of the model parameters for that layer. If the network is deep and complex, this constraint usually prevents sharing the complete model at the edge but also discourages creating an edge periodic feedback loop that could be used to improve the global model due to the communication costs involved.

Practically speaking, replacing $\mathbf{W}$ with a factored version is straightforward, as we can just replace matrix $\mathbf{W}$ in the objective function it with a suitable decomposition which can give a low-rank approximation. In terms of optimizing the learning time similar work has been performed that exploit this property in order to reduce the number of parameters required for training and hence drastically improve training time [1, 5].

Moving further, it is a reasonable assumption to expect a large number of transformations in a complex network and each to have a different $\mathbf{h}$, namely: $\mathbf{h_i} \forall [1, ..., n_l]$, where n_l is the number of layers the network has. Our goal is to find for each $\mathbf{h_i}$ a low-rank representation of the optimized $\mathbf{W_i}$ that has the following desirable properties: i) can be computed in an efficient way and ii) has tunable approximation quality guarantees.

PRELIMINARY DESIGN

One of the predominant tools to transform a matrix to its corresponding low-rank approximation is the use of Singular Value Decomposition (SVD) and Principal Component Analysis (PCA). Recent advances in how these techniques can be performed in a streaming fashion [3, 4] enable us to perform fast, parallel, and accurate transformations for the aforementioned $\mathbf{W_i}$ matrices which are computationally extremely efficient and, when operating under the assumption of no-missing data, the resulting low-rank approximation is very close to the theoretical limit, which is the offline SVD [4]; this provides us with an invaluable tool to perform the low-rank approximation operation but both of the previously mentioned techniques require the rank number r to be fixed and provided beforehand.

Finding a succinct, fast, and accurate way of progressively measuring the reconstruction quality of each $\mathbf{W_i}$ is then desirable in order to be able to *dynamically* adjust the number of rank r_i during its computation. In our preliminary design we employ energy thresholding and fix a range in which we desire our total energy captured to be; the adjusting algorithm in intentionally kept simple in order to have predictable results.

Finally the low-rank approximations W_i can be propagated to edge nodes and thus "transferring" the model to the rest of the network in a very efficient way.

PRIVACY VIA APPROXIMATION

Normally, sharing the exact model representation is desirable but today due to privacy concerns there are many instances that an approximation of that model that obfuscates the personally identifiable information while preserving the actual model is preferred.

Preserving user-privacy by using added noise has been well studied in literature [2, 6] with exceptional results. This principles can be applied to our scheme where using a tunable approximation quality we can obfuscate personally identifiable details which can make our model easier to share with third parties, especially with the new data regulations that have been recently introduced.

ONGOING WORK

Currently our work is focusing on formalizing the implementation details as well as constructing representative tests cases for the thorough evaluation of our proposal.

REFERENCES

[1] Misha Denil, Babak Shakibi, Laurent Dinh, Nando De Freitas, et al. 2013. Predicting parameters in deep learning. In *Advances in neural information processing systems*. 2148–2156.

[2] Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. 2006. Our data, ourselves: Privacy via distributed noise generation. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*. Springer, 486–503.

[3] Armin Eftekhari, Laura Balzano, Dehui Yang, and Michael B Wakin. 2016. SNIPE for memory-limited PCA from incomplete data. *arXiv preprint arXiv:1612.00904* (2016).

[4] Armin Eftekhari, Raphael A Hauser, and Andreas Grammenos. 2018. MOSES: A Streaming Algorithm for Linear Dimensionality Reduction. *arXiv preprint arXiv:1806.01304* (2018).

[5] Song Han, Huizi Mao, and William J Dally. 2015. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149* (2015).

[6] Baik Hoh, Marco Gruteser, Hui Xiong, and Ansaf Alrabady. 2007. Preserving privacy in gps traces via uncertainty-aware path cloaking. In *Proceedings of the 14th ACM conference on Computer and communications security*. ACM, 161–171.

*Author is also affiliated with The Alan Turing Institute