On the Importance of High-Rank Kernel Approximation

Anonymous Author(s)
Affiliation
Address
email

Abstract

We investigate the problem of training kernel approximation models under memory constraints. Our first step toward this goal is to compare the performance of two leading kernel approximation methods, the Nyström method [3] and random Fourier features (RFFs) [2], in this memory constrained setting. Although the Nyström method is generally known to outperform RFFs for a fixed number of features [4], we find that RFFs are able to outperform Nystrom features in this constrained setting. This holds true in spite of the fact that the Nyström method often has smaller kernel approximation error than the RFF approximation, in terms of both the spectral norm and the Frobenius norm, under the same budget. In order to explain this phenomenon, we leverage the recent work of Avron et al.[1]; in this work, they consider an approximation $\tilde{K}$ to a kernel matrix K, which satisfies $(1-\Delta)(K+\lambda\mathbb{1}) \preceq \tilde{K}+\lambda\mathbb{1} \preceq (1+\Delta)(K+\lambda\mathbb{1})$, and they bound the generalization performance of the regression model trained using $\tilde{K}$, with regularization constant λ, in terms of Δ. We demonstrate across a number of tasks that this metric Δ correlates much more strongly with generalization performance than other measures of kernel approximation error, and show that under a fixed memory budget, random Fourier features systematically attains lower Δ values. We explain that this is due to the fact that the RFF approximation is higher rank than the Nyström approximation, when under the same memory budget. Taking inspiration from these results, we propose using *low-precision* random Fourier features (LP-RFFs) in order to generate an even higher rank approximation under the same memory budget. We build on the theory of Avron et al. to analyze the generalization performance of LP-RFFs, proving that performance decays as ...make formal statement. Empirically, we show the we are able to get within xx% of the performance of full-precision RFF models, using xx times less memory during training, on the TIMIT speech recognition dataset.

1 Introduction

- Kernel methods are powerful, but don't scale well. Need kernel approximation methods.
- For kernel approximation methods to do well, generally need lots of features. When performing GPU training (limited memory), or when considering model deployment, the memory requirement for these models becomes a large bottleneck. An example of this is the recent work in large-scale kernel methods for speech recognition, which are constrained in terms of the number of features they can fit on GPU.
- As a result, we perform a thorough empirical comparison of Nystrom and RFF, two leading kernel approximation methods, in the memory constrained setting. Surprisingly, we

Submitted to 32nd Conference on Neural Information Processing Systems (NIPS 2018). Do not distribute.

find that although the Nystrom method often has lower kernel approximation error (Frobenius/Spectral norm), it performs worse than RFF in this setting. We note that this runs counter to the common belief that Nystrom is better than RFF.

- We explain these observations using the recent theoretical results of Avron et al. We note that the bounds in this paper differ in important ways from prior bounds which are generally based on the kernel approximation error (and thus useless to explain the observed phenomenon).
- We note that the primary insight provided by these recent bounds is that it is crucial to have a high-rank decomposition whose spectrum roughly aligns with the exact spectrum, even if it has high kernel approximation error. Note that this runs contrary to a lot of the literature kernel approximation literature, which is often narrowly concerned with reducing the kernel approximation error. Furthermore, we consider the computational cost, in terms of memory, of computing these features, another aspect which is generally not considered when comparing Nystrom and RFF.
- We leverage this insight to propose LP-RFF.
- We analyze LP-RFF theoretically, showing that you can upper bound the value Δ for LP-RFF in terms of the number of bits per feature used.
- We show that LP-RFF are able to perform on par with full-precision features, at a fraction of the memory.
- This opens the door for training kernel approximation models with many more features than are currently being used. To fully leverage these insights to attain faster, lower-memory, and lower-energy training, it will be necessary to implement efficient version of our algorithm (mention TPUs, GPUs, FPGAs, etc). AM: Are TPUs already being used for neural net training in low-precision?
- Include sexy plot in intro summarizing the main results of the paper? Generalization performance vs. memory usage on TIMIT, for Nyström , RFF, and LP-RFF?

2 Related Work

- Give an overview of kernel approximation generalization bounds. Note that the central object they generally use to bound the generalization error is $\|K - \tilde{K}\|$. Discuss the more recent results which approach the problem differently AM: We should read recent papers closely.
- Discuss Nystrom vs. RFF paper. Our main differentiators: 1) we consider larger/more challenging datasets, 2) we consider much larger numbers of features, 3) we consider memory budget.
- Discuss low-precision, and recent interest in doing low-precision training.
- Discuss fastfood/circulant RFF, and other attempts at reducing memory footprint of projection matrix. Note that this doesn't address the memory of the features themselves, which can take a ton of space for moderate size mini-batches. Furthermore, our method can be combined with these methods in order to reduce the memory footprint even further (in fact, that is what we do!).
- Discuss other attempts at improving RFF (orthogonal RFF, Spherical Structured Feature Maps, Gaussian quadrature, quasi-monte carlo, etc.), and why we are different. Discuss that while these methods often improve the kernel approximation error for a fixed number of features, they do not increase the rank for a fixed amount of memory.
- Perhaps discuss methods of improving Nystrom? Sampling distributions, etc.

3 Preliminaries and Notation

3.1 Background: Kernel approximation

- Give a very brief overview of kernel approximation, Nystrom and RFF.
- Discuss the amount of memory each of these methods uses during training. It is important to be very clear about how we are measuring memory.

3.2 Avron Generalization Bound

- Present their bound, and discuss in moderate detail.

$$R(\tilde{f}) \leq \frac{1}{1-\Delta}\hat{R}_K(f) + \frac{\Delta}{1+\Delta} \cdot \frac{rank(\tilde{K})}{n} \cdot \sigma_\nu^2$$

- Mention that we can understand Δ in terms of $D = \max_i |\log(\sigma_i + \lambda) - \log(\tilde{\sigma}_i + \lambda)|$ ($1 + \Delta \geq \exp(D)$), as a way of visualizing the bound; if you plot both spectra $+\lambda$ in log-scale, the Δ is the maximum gap between the two lines (if eigenvectors are the same. So even if the eigenvectors are different, this us gives lower-bound for the true value of Δ).
- Discuss their result which bounds the number of random Fourier features needed to get a specific Δ value.

3.3 Notation

- Present notation for Nystrom and RFF features, kernel matrices, etc.
- Specify that n denotes number of datapoints, d is data dimension, m is number of random features, etc.

4 Nyström vs. Random Fourier Features

- In this section, we will present a number of experiments comparing the performance of Nyström and RFF. We show that under a fixed memory budget, RFF outperforms Nyström, and this is in spite of the fact that Nyström often attains lower kernel approximation error (Frobenius/spectral).
- we will explain this using fixed design theory, and the Δ parameter from [1].

4.1 Nyström vs. RFF: Empirical comparison

- Across various classification/regression datasets (TIMIT, yearpred, CovType, Census, Adult), we compare Nyström vs. RFF, for a wide range of numbers of features.
- We compare these two methods in various ways: 1) For a fixed # of features, 2) For a fixed amount of memory, 3) for a fixed kernel approximation error (x-axis = spectral/Frobenius norm). For the y-axis in these plots, we consider kernel approximation error, and downstream performance.
- We observe that RFF outperforms Nyström for a fixed memory budget, and for a fixed kernel approximation error.

4.2 Nyström vs. RFF: "Δ-analysis"

- Across a few small datasets (Census, Adult), we measure the Δ parameter, and plot the performance of Nyström vs. RFF in terms of this parameter.
- We observe that Nyström and RFF models that have similar Δ values, have similar generalization performance. We show that this correlation is much stronger than the correlation between kernel approximation error and generalization performance.
- We plot the value of Δ attained by Nyström and RFF as a function of memory and number of features. We show that for a fixed memory budget, RFF attains much smaller values of Δ.

5 Low-Precision Random Fourier Features (LP-RFF)

- We introduce low-precision RFF AM: Should we do this earlier, in a "Methods" section?.
- We present our theoretical results about LP-RFF, bounding it's approximation error (maybe push this to appendix), and bounding it's Δ, as a function of the number of quantized features, and the precision being used.
- We show our experiments, which have three parts: (1) Explore the performance on big tasks, (2) Explain trends on big experiments by doing detailed analysis on small tasks, (3) show low-precision training doesn't degrade performance much relative to FP training.

5.1 LP-RFF: Method

- Present LP-RFF method in detail.

5.2 LP-RFF: Theory

- Discuss bounds on kernel approximation performance (if we have space).
- Discuss bounds on Δ.

5.3 LP-RFF: Experiments

- We show performance on the "big tasks".
- We explain these results in terms of the "small tasks", where we measure Δ, showing that below a certain number of bits, Δ can increase a lot. If we have space, we explain the trade-offs here a bit, showing in the fixed design setting that higher noise means higher λ^*, which means fewer bits can be used until Δ increases a lot.
- We briefly present low-precision training, showing that performance doesn't degrade much relative to full-precision training.

6 Conclusion

- We have shown the practical significance of the recent theory of [1]. This came in two parts: (1) It explains why RFF outperforms Nyström, even when they both have the same kernel approximation error; a direct consequence is that under a fixed memory budget, RFF is better. (2) It inspired us to define, analyze, and test the performance of LP-RFF.
- We showed that we can do much better, for a fixed memory budget, using LP-RFF, because they are a much higher-rank decomposition of the kernel matrix.
- This opens the door for larger-scale kernel approximation experiments, which can leverage recent advances in chips, etc, for fast training. For future work, we plan to implement efficient version of our LP-RFF training, to be able to apply this to large/challenging tasks (e.g., speech recognition, computer vision).

References

[1] Haim Avron, Michael Kapralov, Cameron Musco, Christopher Musco, Ameya Velingker, and Amir Zandieh. Random fourier features for kernel ridge regression: Approximation bounds and statistical guarantees. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pages 253–262, 2017.

[2] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems 20, Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 3-6, 2007*, pages 1177–1184, 2007.

[3] C.K.I. Williams and M. Seeger. Using the Nyström method to speed up kernel machines. In T.K. Leen, T.G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 682–688. MIT Press, 2001.

[4] Tianbao Yang, Yu-Feng Li, Mehrdad Mahdavi, Rong Jin, and Zhi-Hua Zhou. Nyström method vs random fourier features: A theoretical and empirical comparison. In *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States.*, pages 485–493, 2012.