Distributed Evaluations: Ending Neural Point Metrics

Daniel Cohen Scott Jordan W. Bruce Croft
University of Massachusetts Amherst, Amherst, MA, USA
{dcohen,sjordan,croft}@cs.umass.edu

ABSTRACT

With the rise of neural models across the field of information retrieval, numerous publications have incrementally pushed the envelope of performance for a multitude of IR tasks. However, these networks often sample data in random order, are initialized randomly, and their success is determined by a single evaluation score. This is exacerbated by neural models achieving incremental improvements from previous neural baselines, leading to multiple near state of the art models that are difficult to reproduce and quickly become deprecated. As neural methods are starting to be incorporated into low resource and noisy collections that further exacerbate this issue, we propose evaluating neural models both over multiple random seeds and a set of hyperparameters within ϵ distance of the chosen configuration for a given metric.

KEYWORDS

deep learning, evaluation

ACM Reference Format:
Daniel Cohen Scott Jordan W. Bruce Croft. 2018. Distributed Evaluations: Ending Neural Point Metrics. In *Proceedings of ACM Conference, Washington, DC, USA, July 2017 (Conference'17),* 2 pages.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

As neural methods have become some of the most effective models for learning representations where traditional hand crafted features have failed to perform [6, 8, 10], there has been a large increase in publications using these approaches. This has allowed the field to move from handcrafting features to handcrafting larger architectures that can learn relevance with millions of parameters. While this approach has made significant strides in the field of IR, reproducible results have become a significant concern within the community [5]. Often, these state of the art results cannot be replicated due to a small issue such as batch size, data preprocessing, random seed, or other hyperparameters of the model. While Choromanska et al. [3] have demonstrated that local minimas are sufficiently close to the global minimum, this is not calibrated with local minimas being a sufficient in evaluation space such as mean precision or recall [2]; a model that achieves a similar loss value is therefore not calibrated to a similar ranking score.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Thus, we propose addressing this issue by introducing a new evaluation method for neural retrieval. Rather than pointwise comparisons of single scores, models would be reported with a probability density function over random seeds. This would allow future work to not only compare the mean performance score, but to examine the sensitivity of new architectures or training methods.

2 VOLATILITY OF NEURAL MODELS

Table 1: Sensitivity over two retrieval models across CQA and WikiQA collections using MAP as evaluation metric over multiple random seeds

Method	CQA	WikiQA
LSTM	.665 ± .004	.592 ± .021
Multitask LSTM	.615 ± .009	.572 ± .060

As Dür et al. [5] have demonstrated, state of the art neural models are extremely susceptible to small changes in hyperparameters, initialization and even random seeds. As IR neural models are often trained and evaluated over a limited number of training queries, this variance is not uncommon. To exemplify this, we conduct a small experiment over multiple random seeds by evaluating a conventional short text retrieval architecture [4, 13] compared with the same model with an additional multitask component to predict part of speech information [7]. This experiment was conducted over two collections. *CQA* which is the combination of nfl6 [4] and Yahoo's *manner* collection commonly referred to as L4 [11]. This combined collection has close to 200,000 individual queries. The other is WikiQA [16], which consists of approximately 2000 training queries.
As seen in Table 1, the large amount of data available within the CQA collection to evaluate these two methods results in a relatively stable performance across random seeds. However, moving to a lower resource collections results in a much higher variance across initialization. Of particular interest is that Multitask LSTM could be portrayed as the superior model under a certain set of random initial conditions.

As recent work has started using reinforcement learning to handle noisy approaches [14], the importance of fully documenting a proposed model's performance becomes an even greater issue. The REINFORCE algorithm [15], used in [14] is known to have exceptionally high variance in the gradient estimates, which translates to high variance in the performance metrics. To demonstrate the importance of using distributed evaluations, we implement a RL approach that has been shown empirically to be more stable than the one used in IRGAN [9, 12, 14]. However, even with these new algorithms, the stochastic optimization process has high variance

and has led to issues with reproducibility [1]. Any IR model using reinforcement learning needs to be evaluated over many trials to accurately convey the results. As seen in Figure!1, we show the sensitivity of two state-of-the-art reinforcement learning algorithms NAC-s [12] and PPO [9] on a common benchmark task, pendulum. The performance is measure as the average sum of reward the RL agent sees over its lifetime. We plot the inverse CDF of agents performance after running 125 thousand different settings of the hyper-parameters as random seeds.

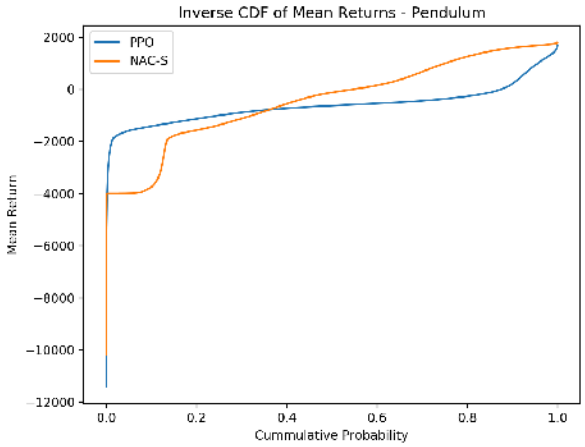

Figure 1: Full performance distribution of NACs and PPO on the pendulum swing and balance task

3 DISTRIBUTED EVALUATIONS

To circumvent the outlined issues in the previous section, we propose a two fold evaluation approach to neural models. First, final evaluation scores should be conducted over multiple random seeds. This creates a distribution of scores, and provides an illustration of the sensitivity of the proposed model to noise. Second, a subsequent set of scores would be evaluated over a small ϵ-ball of the top hyperparameters of the best performing model. The impact of small changes in the hyperparameter space reveals the robustness of the model over the small perturbations to architecture choices.

Using these two approaches, it now becomes viable to create a smoothed distribution of scores from a model and evaluate a novel architecture with the additional information. Using KL-divergence, $KL(P||Q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} dx$, one can not only examine point statistics such as mean and variance, but also how similar each model's sensitivity to randomness and hyperparameters.

4 CONCLUSION

In this paper, we address the issue of under reporting the performance of models that are highly susceptible to noise both in the training data, but also within the model itself. While the proposed distributed evaluation requires greater computation than taking the result of a single run, hyperparameter tuning within a small convex hull is common practice when fine tuning a model for a collection. Thus one need only include these results in the final paper and not incur additional overhead.

With the recent push to release code for the public, setting a standard of distributed results would bring the field one step closer to allowing these methods to be reproducible.

5 ACKNOWLEDGEMENTS

This work was supported in part by the Center for Intelligent Information Retrieval. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

REFERENCES

[1] [n. d.].
[2] Clément Calauzènes, Nicolas Usunier, and Patrick Gallinari. 2012. On the (Non-)existence of Convex, Calibrated Surrogate Losses for Ranking. In *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (Eds.). Curran Associates, Inc., 197–205. http://papers.nips.cc/paper/4646-on-the-non-existence-of-convex-calibrated-surrogate-losses-for-ranking.pdf
[3] Anna Choromanska, Mikael Henaff, Michaël Mathieu, Gérard Ben Arous, and Yann LeCun. 2014. The Loss Surface of Multilayer Networks. *CoRR* abs/1412.0233 (2014). arXiv:1412.0233 http://arxiv.org/abs/1412.0233
[4] Daniel Cohen and W. Bruce Croft. [n. d.]. End to End Long Short Term Memory Networks for Non-Factoid Question Answering. In *ICTIR '16*.
[5] Alexander Dür, Andreas Rauber, and Peter Filzmoser. 2018. Reproducing a Neural Question Answering Architecture Applied to the SQuAD Benchmark Dataset: Challenges and Lessons Learned. In *Advances in Information Retrieval - 40th European Conference on IR Research, ECIR 2018, Grenoble, France, March 26-29, 2018, Proceedings*. 102–113. https://doi.org/10.1007/978-3-319-76941-7_8
[6] Jiafeng Guo, Yixing Fan, Qingyao Ai, and W. Bruce Croft. 2016. A Deep Relevance Matching Model for Ad-hoc Retrieval. In *CIKM '16*. ACM, New York, NY, USA, 55–64. https://doi.org/10.1145/2983323.2983769
[7] Mingsheng Long and Jianmin Wang. 2015. Learning Multiple Tasks with Deep Relationship Networks. *CoRR* abs/1506.02117 (2015). arXiv:1506.02117 http://arxiv.org/abs/1506.02117
[8] Bhaskar Mitra, Fernando Diaz, and Nick Craswell. 2017. Learning to match using local and distributed representations of text for web search. In *WWW 17*. 1291–1299.
[9] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal Policy Optimization Algorithms. *CoRR* abs/1707.06347 (2017). arXiv:1707.06347 http://arxiv.org/abs/1707.06347
[10] Aliaksei Severyn and Alessandro Moschitti. 2015. Learning to Rank Short Text Pairs with Convolutional Deep Neural Networks. In *SIGIR (SIGIR '15)*. ACM, New York, NY, USA, 373–382. https://doi.org/10.1145/2766462.2767738
[11] Mihai Surdeanu, Massimiliano Ciaramita, and Hugo Zaragoza. 2008. Learning to rank answers on large online QA collections. In *ACL:HLT*. 719–727.
[12] Philip Thomas. 2014. Bias in Natural Actor-Critic Algorithms. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*. 441–448. http://jmlr.org/proceedings/papers/v32/thomas14.html
[13] Di Wang and Eric Nyberg. 2015. A Long Short-Term Memory Model for Answer Sentence Selection in Question Answering. In *ACL-IJCNLP, ACL 2015, July 26-31, 2015, Beijing, China, Volume 2: Short Papers*. 707–712. http://aclweb.org/anthology/P/P15/P15-2116.pdf
[14] Jun Wang, Lantao Yu, Weinan Zhang, Yu Gong, Yinghui Xu, Benyou Wang, Peng Zhang, and Dell Zhang. 2017. IRGAN: A Minimax Game for Unifying Generative and Discriminative Information Retrieval Models. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017*. 515–524. https://doi.org/10.1145/3077136.3080786
[15] Ronald J. Williams. 1992. Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning. *Machine Learning* 8 (1992), 229–256. https://doi.org/10.1007/BF00992696
[16] Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. WikiQA: A Challenge Dataset for Open-Domain Question Answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*. 2013–2018. http://aclweb.org/anthology/D/D15/D15-1237.pdf