Random Contrast Learning: A New Approach to Supercomputing and Machine Learning

Dr. Morten Middelfart [1] **Sam Martin** **Ben Martin**

Abstract

In this paper, we demonstrate that, compared to deep learning, random contrast learning (RCL) produces unsupervised language models with faster training, faster inference, and reduced size, all by orders of magnitude, while maintaining better recall. Thus far, we have applied our model to several small datasets. Our findings indicate a promising path toward broader applications in language and exhibit the power of RCL as a new paradigm in machine learning.

1. Introduction

Machine learning systems that rely on neural networks require large datasets and significant compute, are prone to drift toward common patterns, and have upper limits on the size to which they can scale. Large datasets have become increasingly accessible, but the cost of compute typically constrains the development of large language models to well-funded organizations. Moreover, despite their growing prevalence in the industry, deep learning models continue to behave unexpectedly in response to less common inputs. Each additional training parameter increases the computational complexity of these systems exponentially. The result is an ever-increasing need for data and computing power. Consequently, the need is apparent for a structure capable of sublinear to linear scalability.

RCL both addresses many shortcomings of current unsupervised machine learning systems and exhibits sublinear to linear scaling without theoretical limit.

2. Approach

We approach building a RCL language model with the following assumption: In the analysis of text, word combinations that make a passage unique constitute, or are, context.

Consider two books on the same topic. What distinguishes them? Each book's author, publisher, copyright date, or segments of conclusions diverging from those of the other book are examples of phrases that distinguish one book from the other. Rather than compare two similar or two different texts, we compare a given text to a random sample of text. The comparison yields an auditable model of word combinations that make the text unique. Because the given text is compared with a sufficiently varied sample of random text, coincidental and uninteresting patterns are filtered out. Random contrast thus identifies patterns that are both distinctive and interesting.

This method enables us to group texts according to context and to predict subsequent word combinations. The resulting model is a collection of n-gram corpora that mirror the directionality of context within the dataset — a structure totally different from a typical deep learning model. In the following sections, we present the results of our preliminary tests and draw some conclusions from them.

2.1 Training Datasets

We chose several small datasets to expedite our proof of concept. As is typical, the quality of a model depends on the dataset in relation to its intended purpose. We chose training sets ranging from unstructured tweets to semi-structured legal documents to demonstrate stable recall regardless of the quality of text.

[1] Lumina, Tampa, FL, USA. Correspondence to: RCL@lumina247.com

Dataset	Raw Size	Data Lake Storage	Inference Model	Compression Rate
Legal Documents	393	358	235	59.80%
Tweets	663	604	410	61.84%

Data shown in kilobytes.

2.2 CPU vs GPU

Unlike typical deep learning models, RCL runs faster on CPU than GPU. Furthermore, we observed an order of magnitude increase in speed when comparing physical machines (2x32 Core 2.3 GHz and 128 GB RAM Non-GPU enabled) to virtual machines (32 Core 2.1 GHz and 128 GB RAM Non-GPU enabled).

3. Experiments

The following metrics compare the RCL model to a Keras-framework neural network developed for the purpose of machine translation. We did not develop the inference for machine translation, but the following tests indicate the capacity of the RCL model to outperform neural networks in size, training speed, inference speed, and recall.

	Keras	RCL	Improvement
Size (KB)	242,325	1180	205.36x
Training	6d 3h 37m	0d 3h 8m	47.11x
Inference	27.0s	1.30s	20.77x
Recall	79.6%	96.3%	—

In testing, we trained a RCL model and a neural network on a physical machine (CPU Intel Core i7-6700K 4.00 GHz, GPU NVIDIA Ge-Force GTX 1080, 32GB RAM), utilizing CPU for RCL and GPU for training the neural network.

RCL produced a language model more than 200x smaller than the deep learning model. RCL also trained 47x faster. To test inference speed, we calculated the average run time per 1,000 queries. The RCL model outperformed the neural network in inference speed by 20x. Moreover, RCL reduces the upper bound for false negatives from 20.4% to 3.7%.

The deep learning model reached maximal use of RAM while training, and RCL did not. RCL minimizes the size of and coordination between processes in memory. RCL processes remain largely independent and are naturally smaller. This enables RCL to scale linearly across machines.

4. Specification vs Memorization

When we reduced the algorithm's sensitivity to distinctive word combinations, recall improved; when we raised the algorithm's sensitivity, recall declined. In application, recall may be exchanged for greater sensitivity to context, and vice versa. Future tests will shed more light on this dynamic.

5. Conclusion

When applied to developing an unsupervised language model, RCL outperforms neural networks in size and speeds of training and inference by 1-3 orders of magnitude and in recall. Because its processes are largely independent, unlike neural networks, RCL scales linearly across machines without theoretical limit. As the size of the RCL language model increases, inference speed remains near constant.

In this paper, we applied RCL to the development of an unsupervised language model. We have also sought to recommend RCL as a new paradigm in machine learning, with the potential to outperform neural networks generally. Beyond its application to language, we have successfully applied RCL techniques in imaging and other media that require the detection of weak signals.

Updated February 15, 2022[2]

[2] An earlier version of this paper did not include the day of February updated. This version includes the latest metrics as of the 15th of February.