7 Ways to Handle Large Data Files for Machine Learning

by **Jason Brownlee** on May 29, 2017 in **Machine Learning Process**

Exploring and applying machine learning algorithms to datasets that are too large to fit into memory is pretty common.

This leads to questions like:

- How do I load my multiple gigabyte data file?
- Algorithms crash when I try to run my dataset; what should I do?
- Can you help me with out-of-memory errors?

In this post, I want to offer some common suggestions you may want to consider.

7 Ways to Handle Large Data Files for Machine Learning
Photo by Gareth Thompson, some rights reserved.

1. Allocate More Memory

Some machine learning tools or libraries may be limited by a default memory configuration.

Check if you can re-configure your tool or library to allocate more memory.

A good example is Weka, where you can increase the memory as a parameter when starting the application.

2. Work with a Smaller Sample

Are you sure you need to work with all of the data?

Take a random sample of your data, such as the first 1,000 or 100,000 rows. Use this smaller sample to work through your problem before fitting a final model on all of your data (using progressive data loading techniques).

I think this is a good practice in general for machine learning to give you quick spot-checks of algorithms and turnaround of results. **Start Machine Learning**

You may also consider performing a sensitivity analysis of the amount of data used to fit one algorithm compared to the model skill. Perhaps there is a natural point of diminishing returns that you can use as a heuristic size of your smaller sample.

3. Use a Computer with More Memory

Do you have to work on your computer?

Perhaps you can get access to a much larger computer with an order of magnitude more memory.

For example, a good option is to rent compute time on a cloud service like Amazon Web Services that offers machines with tens of gigabytes of RAM for less than a US dollar per hour.

I have found this approach very useful in the past.

See the post:

- How To Develop and Evaluate Large Deep Learning Models with Keras on Amazon Web Services

4. Change the Data Format

Is your data stored in raw ASCII text, like a CSV file?

Perhaps you can speed up data loading and use less memory by using another data format. A good example is a binary format like GRIB, NetCDF, or HDF.

There are many command line tools that you can use to transform one data format into another that do not require the entire dataset to be loaded into memory.

Using another format may allow you to store the data in a more compact form that saves memory, such as 2-byte integers, or 4-byte floats.

5. Stream Data or Use Progressive Loading

Does all of the data need to be in memory at the same time?

Perhaps you can use code or a library to stream or progressively load data as-needed into memory for training.

This may require algorithms that can learn iteratively using optimization techniques such as stochastic gradient descent, instead of algorithms that require all data in memory to perform matrix operations such as some implementations of linear and logistic regression.

For example, the Keras deep learning library offers this feature for progressively loading image files and is called flow_from_directory.

Another example is the Pandas library that can load large CSV files in chunks.

6. Use a Relational Database

Relational databases provide a standard way of storing and accessing very large datasets.

Internally, the data is stored on disk can be progressively loaded in batches and can be queried using a standard query language (SQL).

Free open source database tools like MySQL or Postgres can be used and most (all?) programming languages and many machine learning tools can connect directly to relational databases. You can also use a lightweight approach, such as SQLite.

I have found this approach to be very effective in the past for very large tabular datasets.

Again, you may need to use algorithms that can handle iterative learning.

7. Use a Big Data Platform

In some cases, you may need to resort to a big data platform.

That is, a platform designed for handling very large datasets, that allows you to use data transforms and machine learning algorithms on top of it.

Two good examples are Hadoop with the Mahout machine learning library and Spark wit the MLLib library.

I do believe that this is a last resort when you have exhausted the above options, if only for the additional hardware and software complexity this brings to your machine learning project.

Nevertheless, there are problems where the data is very large and the previous options will not cut it.

Summary

In this post, you discovered a number of tactics that you can use when dealing with very large data files for machine learning.

Are there other methods that you know about or have tried?
Share them in the comments below.

Have you try any of these methods?
Let me know in the comments.

About Jason Brownlee

Jason Brownlee, PhD is a machine learning specialist who teaches developers how to get results with modern machine learning methods via hands-on tutorials.
View all posts by Jason Brownlee →

‹ On the Suitability of Long Short-Term Memory Networks for Time Series Forecasting

How to Evaluate the Skill of Deep Learning Models ›

40 Responses to *7 Ways to Handle Large Data Files for Machine Learning*

Chris May 29, 2017 at 6:59 pm #

REPLY

If the raw data is seperated by line break, such as csv EDIFACT, ect. Then there is a feature in almost every language I am aware of that will read only 1 line at a time using a socket stream. Which typically how any (buzzword alert) big data solution does it under the hood, nothing magic, hard, or revolutionary about it actually you'll find pretty much any simple GitHub repo doing it if they read files.
Any beginner coder should encounter this and universities should absolutely be teaching such a basic concept in any computer science related degree where you are required to read from a file..

Just thought I'd shed some light on this fact, the 7 ways are actually 7 things that if you see them as an example in blog posts you should immediately leave the site and never return.

Start Machine Learning

REPLY

Jason Brownlee June 2, 2017 at 12:24 pm #

Thanks for the input Chris, there are a lot of different types of machine learning practitioners out there.

MicrobicTiger June 2, 2017 at 1:24 pm #

REPLY

Hi Chris,

What if your data were geographic points with values, each line represented a different point and you were looking to recognize patterns across clusters of points with varying cluster geometries? How would line by line source data reading help me here?

David Severson March 10, 2019 at 1:32 pm #

REPLY

Sorry for late reply. You are correct that files can be processed one line at a time. However, various algos need to reference almost any other piece of data in the set or maybe massive pieces of intermediate date created in the process. As a result those algos are much more difficult to reduce their memory requirements. Off the top of my head they do these in code by using some of these techniques….

Sometimes multiple passes through data on disk
Sometimes intermediate date is small enough to squeeze into memory
Sometime none of the above are possible and algo must go to disk to reference a date element so they put that data in some sort of indexed structure like an HBase is done.
Sometimes they can do it in pieces and use something like gradient descent as noted here

It is a serious engineering problem when the size of training set gets to large and the algo can't progressively process as you propose here

Jason Brownlee March 11, 2019 at 6:46 am #

REPLY

Yes, it's a totally different beast!

felipe almeida May 30, 2017 at 4:43 pm #

REPLY

Some of the tips are a little bit obvious but overall it's good. You could give more examples in each topic, such as "use file formats like X", Also, you could mention

things like using stochastic gradient descent or other kinds of online learning, where you feed the examples one at a time.

Jason Brownlee June 2, 2017 at 12:33 pm #

REPLY

Thanks for the suggestion.

felipe almeida May 30, 2017 at 4:45 pm #

REPLY

Oh yeah, you could also mention using sparse (rather than dense) matrices, as they take much less space and some algorithms (like SVM) can handle sparse feature matrices directly. Here's a link explaining that for sklearn.

Jason Brownlee June 2, 2017 at 12:34 pm #

REPLY

Great suggestion.

Peter Marelas May 30, 2017 at 9:29 pm #

REPLY

A few things I would suggest if you are a python user.

For out-of-core pre-processing:

– Transform the data using a dask dataframe or array (it can read various formats, CSV, etc)
– Once you are done save the dask dataframe or array to a parquet file for future out-of-core pre-processing (see pyarrow)

For in-memory processing:

– Use smaller data types where you can, i.e. int8, float16, etc.
– If it still doesn't fit in-memory convert the dask dataframe to a sparse pandas dataframe

For Big Data try Greenplum (free) https://greenplum.org/. It is a derivative of Postgres. Benefit being queries are processed across cores in parallel. Also has a mature machine learning plugin called MADlib.

Jason Brownlee June 2, 2017 at 12:33 pm #

REPLY

Great suggestion Peter, thanks.

Lee Zee June 20, 2017 at 4:37 am #

REPLY

Can feature selection applications identify features that are comprised of parts of multiple columns in a large datasets? Or, will each identified predictive feature be restricted to data from a single column of data?

Jason Brownlee June 20, 2017 at 6:42 am #

REPLY

Often they focus on single columns. Perhaps you can dip into research and find some more complex methods.

Dan August 14, 2017 at 11:58 am #

REPLY

Hi Jason,

I have encountered a problem when using NLTK to analysis text based on Hadoop/Spark environment, and the problem is the NLTK data (corpora) can't be find on each worker node (I only download the NLTK data in worker node, and I can't download these data on each worker node due to access limitation.

Could you give me some suggestion about how to conduct NLP analysis with NLTK data on each worker node without download the NLTK data in each worker node?

Thanks in advance.

Jason Brownlee August 15, 2017 at 6:28 am #

REPLY

Sorry, I have not used NLTK in Hadoop/Spark, I cannot give you good advice.

Azhaar Hussain August 30, 2017 at 5:43 am #

REPLY

Hi Jason,

I wanted to understand how we put the machine learning in use in practice? Let say, I have developed a model for prediction and it works, how do I put this in production?

Thanks,
Azhaar

Jason Brownlee August 30, 2017 at 6:22 am #

REPLY

See this post:
http://machinelearningmastery.com/deploy-machine-learning-model-to-production/

debraj August 31, 2017 at 7:23 pm #

REPLY

i have a log data with lot of procedure in it. how i can apply ML for a new log data to predict what and all procedure are followed

Jason Brownlee September 1, 2017 at 6:44 am #

REPLY

Perhaps this post will help you define your problem:
http://machinelearningmastery.com/how-to-define-your-machine-learning-problem/

Daniel November 20, 2017 at 3:56 am #

REPLY

Hi Jason,

Thanks for this great post !!!
I am curious why you said Big Data platform such as Hadoop and Spark is the last resort ?
What's the reason !

Thank you,

Daniel

Jason Brownlee November 20, 2017 at 10:21 am #

REPLY

It is a lot of overhead to bring to the table only when it is really needed, e.g. only when you exhaust other options and you truly need a big data platform.

I was not trying to offend. If you're doing hadoop all day and want to run small data through it, then by all means.

David Severson March 10, 2019 at 1:40 pm #

REPLY

I want to second the commenter here. Some companies have these capabilities already setup and may already have much of the date in question. So spark may be the first place to go after an easy run on your desktop doesn't pan out. Forcing and undeposited desktop to process the problem may take considerably longer.

Also, spinning up EMR with or without spark on AWS can be pretty quick if you don't have in-house stuff running.

Too many business problems are getting too big for our laptops not to have some resources prepped and ready to go.

oksana December 3, 2017 at 4:58 am #

REPLY

Hello Dr. Brownlee,
Your recommendation No. 6 – 6. Use a Relational Database, this is what i tried to do using R and failed (not enough knowledge). I was being able to connect to SQL db certain columns of interest, but was unable to extract data I needed – is there a resource you recommend that i read? Any recommendation is highly appreciated.
thank you!

Jason Brownlee December 3, 2017 at 5:28 am #

REPLY

I have done this myself, but it was years ago. Sorry, I don't have a good resource to recommend other than some google searching.

Liam January 24, 2018 at 10:49 pm #

REPLY

Awesome. Thank you for posting your thoughts about this machine learning problem. 🙂

It helped me a lot for my current project.

Thank you

Liam

Start Machine Learning

Jason Brownlee January 25, 2018 at 5:56 am #

REPLY

You're welcome.

Bhavika Panara May 22, 2018 at 10:22 pm #

REPLY

Hi, Jason Brownlee

I want to feed very large image dataset which has 1200000 images and 15,000 classes to convolution neural network. But I am not able to feed all images to CNN However, I have GTX 1080ti 11 GB GPU and 32 GB CPU RAM.

how can I train my model on this very large image dataset on my limited computing resource?

Is there any technique available so I can train my model using multiple chunks of images.

Jason Brownlee May 23, 2018 at 6:27 am #

REPLY

Perhaps you can use progressive loading? I have found it to be very effective for large datasets on small computers.

Vincent July 11, 2019 at 10:25 pm #

REPLY

Do you mean lazy loading?

Jason Brownlee July 12, 2019 at 8:43 am #

REPLY

No, you can define a data generator to load/prepare one sample or a batch samples on demand for the model during training.

Akash Tyagi October 13, 2019 at 3:49 am #

Can you please guide more on this.

Start Machine Learning

Jason Brownlee October 13, 2019 at 8:31 am #

I have many examples, for image start here:
https://machinelearningmastery.com/how-to-load-large-datasets-from-directories-for-deep-learning-with-keras/

For a custom generator start here:
https://machinelearningmastery.com/develop-a-deep-learning-caption-generation-model-in-python/

Anjali Batra August 28, 2018 at 8:20 pm #

REPLY

increase the memory as a parameter : Linked provided in the first point doesnt work anymore. Can you please suggest some other url for the same purpose.

Jason Brownlee August 29, 2018 at 8:10 am #

REPLY

You can increase the memory for a Java application by adding the -Xmx flag, for example for 8GB use -Xmx8000m

Sintyadi Thong April 23, 2019 at 2:07 pm #

REPLY

Hi, Jason!
It is a very good article.

I am wondering, let's say I have 300Million rows of data.
Is it legit to use bootstrapping with samples from the 300 million rows?

let's say I use sample with replacement to get a bag of 30 million rows, and create N bags of it.
From each of them, I run a model..

so somewhat like bagging, but the number of rows in each bag is less than the actual number of rows.

Is it possible and logically correct?

Thanks!

Start Machine Learning

REPLY

Jason Brownlee April 23, 2019 at 2:33 pm #

Perhaps. It depends on how sensitive your model is to the amount of data, and how much time/compute you have available.

Eric Ngo May 26, 2019 at 6:56 am # REPLY

Hi Jason,
I have about 960 .csv files that each .csv file contains speech/voice of a person and 120 transcripts. Should I concatenate 960 .csv files into a single file?

Jason Brownlee May 27, 2019 at 6:35 am # REPLY

Perhaps, it really depends on how you intend to model the problem.

Ashley July 3, 2019 at 9:16 am # REPLY

What if I am using Orange? I am using the software for financial analysis because I am not a programmer and cannot code. Can Orange handle large sets of data?

Jason Brownlee July 4, 2019 at 7:36 am # REPLY

I don't know, sorry.

Leave a Reply

Start Machine Learning

Name (required)

Email (will not be published) (required)

Website

SUBMIT COMMENT

Welcome!

My name is *Jason Brownlee* PhD, and I **help developers** get results with **machine learning**.

Read more

Never miss a tutorial:

Picked for you:

Your First Deep Learning Project in Python with Keras Step-By-Step

Your First Machine Learning Project in Python Step-By-Step

How to Develop LSTM Models for Time Series Forecasting

Start Machine Learning

Why Machine Learning Does Not Have to Be So Hard

Machine Learning for Programmers