

Article

The Pipeline of Processing fMRI data with Python Based on the Ecosystem NeuroDebian

Qiang Li ^{1,2†,‡} , Rong Xue ^{1,2,3†,‡,*}

¹ Sino-Danish College, University of Chinese Academy of Sciences, Beijing, 100049, China

² State Key Laboratory of Brain and Cognitive Science, Beijing MRI Center for Brain Research, Institute of Biophysics, Chinese Academy of Sciences, Beijing, 100101, China

³ Beijing Institute for Brain Disorders, Beijing, 100053, China

* Correspondence: rxue@bcslab.ibp.ac.cn

† Current address: 15 Datun Road, Chaoyang District, Beijing, 100101, China

‡ These authors contributed equally to this work.

Abstract: In the neuroscience research field, specific for medical imaging analysis, how to mining more latent medical information from big medical data is significant for us to find the solution of diseases. In this review, we focus on neuroimaging data that is functional Magnetic Resonance Imaging (fMRI) which non-invasive techniques, it already becomes popular tools in the clinical neuroscience and functional cognitive science research. After we get fMRI data, we actually have various software and computer programming that including open source and commercial, it's very hard to choose the best software to analyze data. What's worse, it would cause final result imbalance and unstable when we combine more than software together, so that's why we want to make a pipeline to analyze data. On the other hand, with the growing of machine learning, Python has already become one of very hot and popular computer programming. In addition, it is an open source and dynamic computer programming, the communities, libraries and contributors fast increase in the recent year. Through this review, we hope that can make neuroimaging data analysis more easy, stable and uniform base the one platform system.

Keywords: neuroscience; big data; functional Magnetic Resonance (fMRI); pipeline; one platform system

0. Introduction

High-Field structural and functional MRI (Magnetic Resonance Imaging), this technique can non-invasively detect brain signal and has substantially high spatial resolution compared with EEG(Electroencephalography, MEG(magnetoencephalogram), etc. However, neuroimaging is a complex field that explores inter-disciplinary studies including physics, engineering, biological science, clinical medicine, physiology, statistics, and much more. The pure (f)MRI data only provide limited information and feedback about the brain, so data mining is a necessary and significant step for us to get more quantitative and broad functional information. When I started to analysis my first batch of (f)MRI data, there existed a lot of software and scripts on the Internet. Unfortunately, First, the source code was extremely messed. Second, the script was written by a variety of programming language, and it hardly connects all code together. Third, according to prior research, if you set up different parameters in the same software, it would affect the final results and make them unstable. Even worse, it's hard for us to combine all kind of software or scripts together, then batch processing (f)MRI data (Eklund et al. 2015; Pauli et al. 2016; Della-Maggiore et al. 2002). Fourth, the main computer programming is Matlab for the (f)MRI data analysis, but with the boost of machine learning and deep learning, Python is gradually beyond Matlab. In addition, Python is a totally open-source computer programming language, so compared to Matlab, it can accommodate the huge community contributes to the Python. What's more, Python has already formed a complete software and powerful community

in the (f)MRI data analysis. Finally, compared to other computer programming languages, Python script is an interpretable and dynamic programming language, the source code is more simple and understandable (Fig. 1).

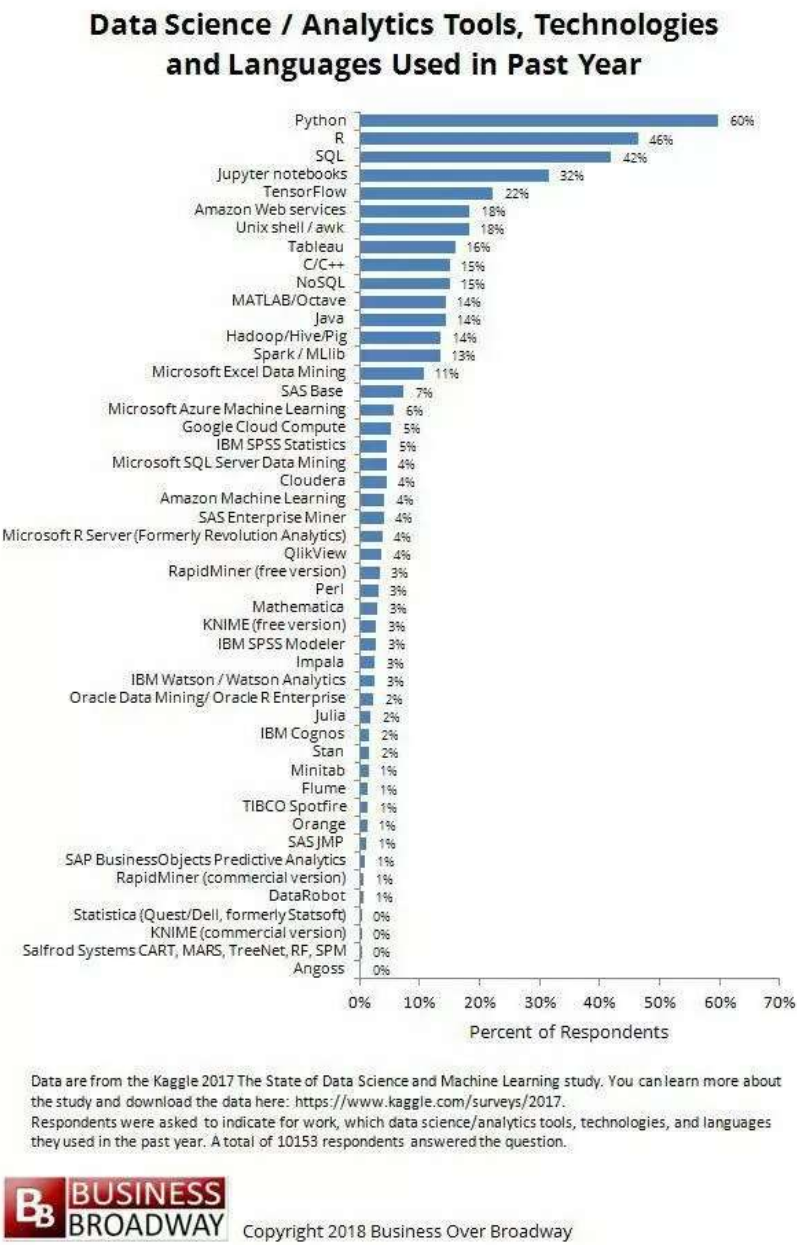


Figure 1. Python compared with other main popular computer language used in data science. The main analytic computer languages for data analysis in the neuroscience and neuroimaging field also fit this trend. The image is adapted from Business Broadway.

With the advent of online neuroimaging databases, people can mine more features from these big data. This is also especially important for students at smaller colleges and universities without an MRI scanner. My goal is to make neuroimaging analysis understandable and accessible for anyone who is interested in neuroscience and neuroimaging. Above all, I aim to make fMRI analysis more precise and integrative.

1. The Ultimate Neuroscience Software Platform-Ecosystem

NeuroDebian¹ embrace a lot of popular neuroscience research software for the Debian operating system as well as Ubuntu and other derivatives. Popular packages include AFNI², FSL³, PyMVPA⁴, SPM⁵ and many others (Halchenko and Hanke 2012).

1.1. Install NeuroDebian

1. Windows(64 bit), first choose a download server or imaging file, before you install it, you should install VirtualBox⁶. After installing VirtualBox, then finish configure it.
2. Mac OS X, the same way with Windows(64 bit).
3. Linux (Ubuntu, Debian etc), it has two conditions:

- Only install software by user chosen

```
wget -O http://neuro.debian.net/lists/bionic.cn-bj1.libre
| sudo tee /etc/apt/sources.list.d/neurodebian.sources.list

sudo apt-key adv --recv-keys --keyserver
hkp://pool.sks-keyservers.net:80 0xA5D32F012649A5A9

sudo apt-get update
```

- All software install once

```
wget -O http://neuro.debian.net/lists/bionic.cn-bj1.full
| sudo tee /etc/apt/sources.list.d/neurodebian.sources.list

sudo apt-key adv --recv-keys --keyserver
hkp://pool.sks-keyservers.net:80 0xA5D32F012649A5A9

sudo apt-get update
```

Based on the NeuroDebian system, you can download software and data that you want by one simple command line. e.g.

```
sudo apt-get install ###pkgname###
```

1.2. Manage VirtualBox

HashiCorp Vagrant⁷ provides the easy, friendly workflow regardless of your role as a new user, designer or professional programmer. It use configuration file which describes all your program or software condition requirements or libraries, operating system configuration, dependency etc.

1. Install Vagrant. More details and info see here (<https://www.vagrantup.com/docs/installation/>).
2. Configure the system environment. The mechanism of Vagrant is very similar to Github. It can help you manage your system(e.g. VirtualBox).

¹ <http://neuro.debian.net/>

² <https://afni.nimh.nih.gov/>

³ <https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/>

⁴ http://www.py_mvpa.org/

⁵ <https://www.fil.ion.ucl.ac.uk/spm/>

⁶ <https://www.virtualbox.org/wiki/Downloads>

⁷ <https://www.vagrantup.com/>

2. The Data Management Tools—Data Portal

DataLad⁸ providing a data portal and a versioning system for data download or transmit, DataLad lets you can control and share your data more flexible. As I have mentioned above, with the advent of online neuroimaging databases, some famous datasets are listed as follows:

- OpenfMRI - free and open sharing of raw magnetic resonance imaging (MRI) datasets. (The number of currently available datasets: 95. The number of subjects across all datasets: 3372. Until 5. Nov 2018)
- Neurodata.io - the database of large-scale connectomics data.
- Core Nets - the substantial database of macaque tract-tracing data.
- Preprocessed Connectomes Project - preprocessed human fMRI data.
- NITRC - another portal that provides the ability to search across many different data sets and databases.

When we face these big data, the size of one subject scanning data can be larger than 1GB. Assume we want to transmit multi-subjects scanning data for remote services or share with other universities or people, that will be a big challenge for researchers. DataLad can fix these problems. You can download, upload, consume these reproducibility data with other people.

2.1. Install DataLad

DataLad can easily be installed via pip.

```
pip install —user datalad
```

If your developing system is Linux(e.g. Ubuntu, Debian etc.)

```
sudo apt-get install datalad
```

More information about how to install DataLad, link here⁹.

3. Basic Unix Command Line

It is a necessary step for us to study image analysis. You need to know some basic and frequently-used command lines in the Unix system. e.g. ls, cd, rm, grep etc. More tutorial about Unix commands line here¹⁰.

4. Version Control System

Git¹¹ is a free, powerful and open source distributed version control system designed to operate and cooperate projects from small to very large with speed and efficiency.

First, you can choose Git version based on your system. More information on how to install Git link here (<https://git-scm.com/downloads>). Here I will show you a simple case of how to use Git to share your scripts or control your version system.

The first step: build an account on the Github.com and configure your environment. More details link here¹². Then build a new repository and initial it. You can clone or download this repository into the local directory.

```
git clone ####The Link of Repository####
```

⁸ <https://www.datalad.org/>

⁹ https://www.datalad.org/get_datalad.html

¹⁰ http://fsl.fmrib.ox.ac.uk/fslcourse/unix_intro/

¹¹ <https://git-scm.com/>

¹² <https://git-scm.com/doc>

Now, you can modify, add, change and remove files or folders in your local directory. What has been changed in your repository, you can check it.

```
#Check current condition
git status
```

Before you submit to your remote repository, you need to add comments for what kind of information is updated. Then you can push your local repository into remote services.

```
git commit -m "add info for what has been changed"
git push -u origin master
%For security, you need login your name and password%
%Then you can go back your Github to check your updated repository%
```

5. Docker

Docker¹³ is a lightweight container software, it provides container software that is ideal and easy for users who are looking to occupy small computer memory and handle data with container-based applications. It has totally different principle with Virtual machine, because docker share the computer hardware resources and operating system with host machine, it can easy to make dynamic distribution of resources.

The docker can make each software or app or imaging run separately in the container. You also can build you imaging according to your needs in your work. More basic tutorials about Docker link here¹⁴. Docker is also very popularly used in the neuroimaging analysis recently.

6. Brain Imaging Data Structure(BIDS)

BIDS¹⁵, a format for organizing and describing neuroscience data structure of neuroimaging experiments (Gorgolewski et al. 2016). One of the big problems for (f)MRI data analysis is data organization in the directory. In the fMRI experiment, it consists of BOLD data, anatomy data, behavioral data (Fig.2). The researchers can use BIDS app to make their data in a more standard format and order. The benefit can be listed as follows:

1. It will be easier for sharing your data.
2. It can make your data more integrity and order.
3. People will read and understand easier.
4. Many data analysis software all support BIDS formats.

PyBIDS is one of BIDS library that can centralize interactions with datasets. For more information about BIDS visit here¹⁶. More command lines with Python link here¹⁷.

Here is an example for explaining how to use PyBIDS to reorganize (f)MRI data. More detailed information and tutorials you can check here¹⁸.

```
import bids.layout
import bids.tests
import os
```

The last data organization after BIDS is shown as follows:

¹³ <https://www.docker.com/>

¹⁴ <https://neurohackweek.github.io/docker-for-scientists/>

¹⁵ <http://bids.neuroimaging.io/>

¹⁶ <http://bids.neuroimaging.io>

¹⁷ <https://bids-standard.github.io/pybids/>

¹⁸ <https://github.com/bids-standard/pybids/blob/master/examples/pybids%20tutorial.ipynb>

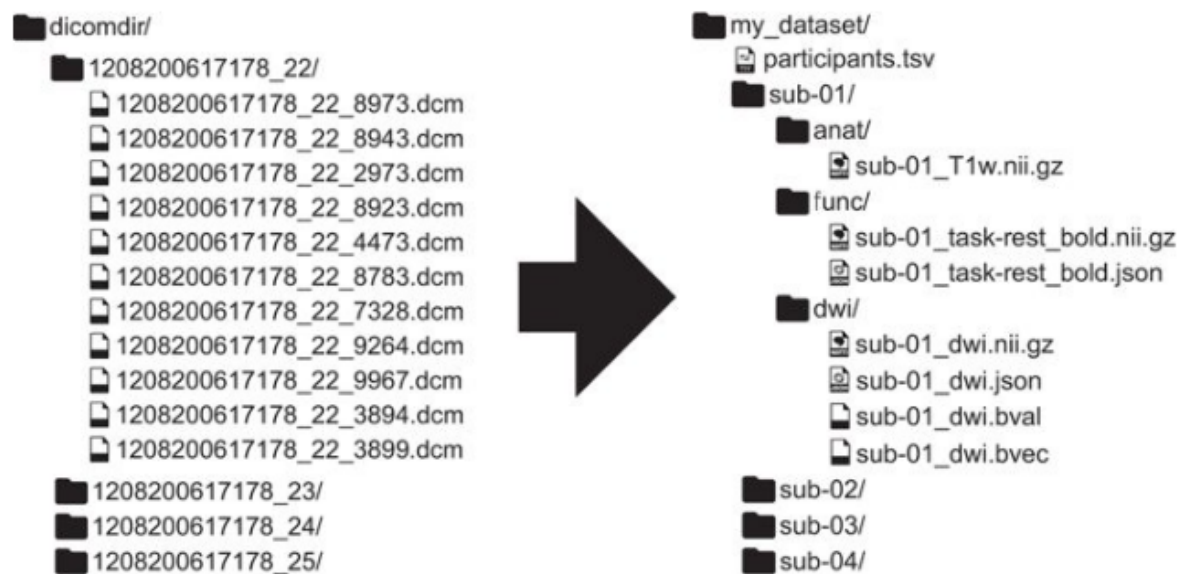


Figure 2. The BIDS processing result of (f)MRI data. The image is adapted from <http://bids.neuroimaging.io/>.

7. The Quality Control of (f)MRI data

This step is also very important before you start processing your raw data. In the Python, `mriqc` is a very powerful tool for checking your data quality and it also can provide a user-friendly and nice report for your data. More example for `mriqc`, you can check here¹⁹.

8. The Tools of Preprocessing of (f)MRI data Based on The NeuroDebian

In this section, we highly recommend two tools for preprocessing raw (f)MRI data with Python.

8.1. *fmripred: A Robust Preprocessing Pipeline for fMRI Data*

`Fmripred`²⁰ is a functional magnetic resonance imaging (fMRI) data preprocessing pipeline based Python, it provides very easy, friendly interface for users, it also allowed uses to input minimum parameters to drive the total pipeline works, but it providing more details output reporting (Esteban et al. 2018). Compared to SPM (Statistical Parametric Mapping, <https://www.fil.ion.ucl.ac.uk/spm/>) software, it would be more easy interface the next analysis with machine learning and other approaches.

The document of `fmripred` link here²¹. The total `fmripred` workflow can be shown as follows (Fig. 3):

For more tutorials on how to practice it, you can link here²².

8.2. *Nipype: Neuroimaging in Python Pipelines and Interfaces*

`Nipype`²³, an open source, community contribute based from NiPy, is a Python project that provides a uniform interface to existing neuroimaging software and facilitates interaction between these packages within a single workflow (Gorgolewski et al. 2011).

¹⁹ <https://mriqc.readthedocs.io/en/stable/>

²⁰ <https://fmripred.readthedocs.io/en/stable/>

²¹ <https://fmripred.readthedocs.io/en/stable/>

²² <https://github.com/poldracklab/fmripred-notebooks>

²³ <https://nipype.readthedocs.io/en/latest/>

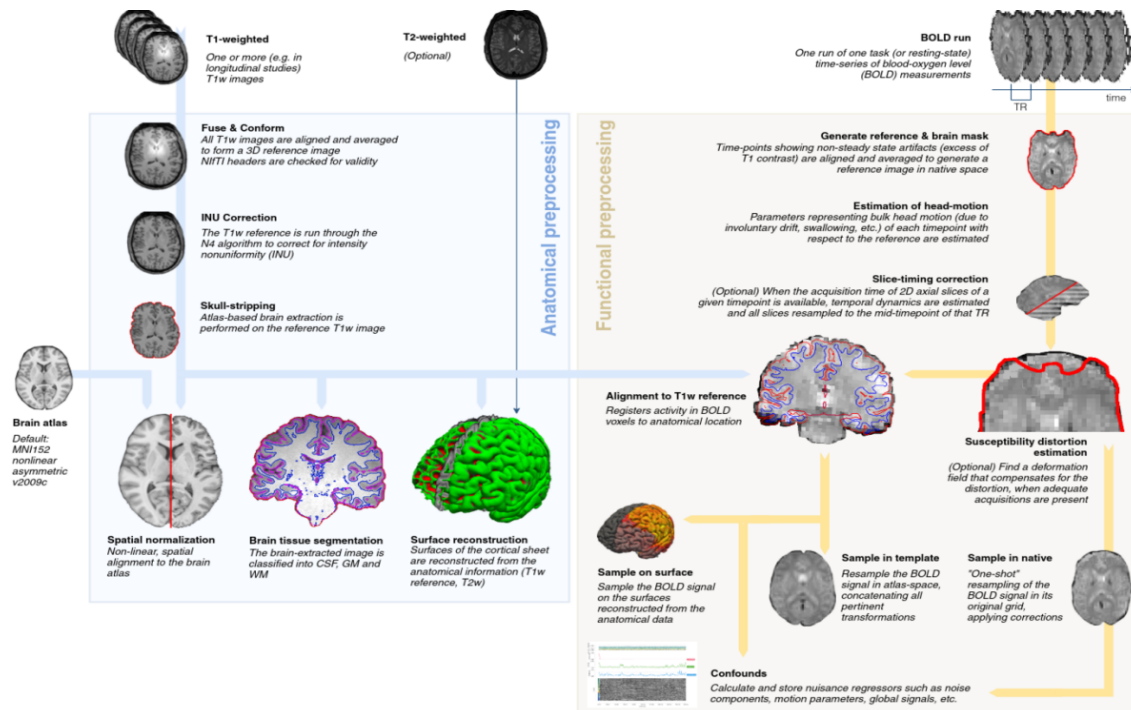


Figure 3. The fmripipeline workflow chart. In the left side, it is used for preprocessing anatomic MRI data, e.g. head motion correction, tissue segmentation, surface reconstruction etc. In the right side, it is used for preprocessing functional MRI data, e.g. slicing correction, smooth and statistics etc. The image is adapted from <https://fmripipeline.readthedocs.io/en/stable/>.

The mechanism of Nipype is very similar to fmripipeline. It also combined different packages together (e.g. ANTS, FreeSurfer etc) (Fig. 4). More document of Nipype can be found here²⁴.

9. Machine learning

After preprocessing (f)MRI data, we can do more exploration for these data and mining more medical information. Machine learning is one of the perfect examples to explain it. Here I will introduce some libraries with Python to do machine learning for neuroimaging.

9.1. Nilearn

Nilearn²⁵ is a Python module for fast and easy statistical learning on NeuroImaging data. It based on the scikit-learn²⁶ toolbox (popular machine learning package with Python) for multivariate statistics with applications such as classification, decoding, or connectivity analysis etc, it still growing and adding more function by large community contributors, in the future, it will have more function and property into it.

Among the best libraries of machine learning for neuroimaging with Python, I think Nilearn is the best choice for us. In the next part, I will show you some simple examples to help you quickly get into Nilearn. Nilearn also supplies interface that helps you to automatically download data from OpenfMRI and NeuroVault (a new home for all brain statistical maps).

```
#Import module from nilearn
from nilearn import datasets
```

²⁴ <https://nipype.readthedocs.io/en/latest/documentation.html>

²⁵ <https://nilearn.github.io/>

²⁶ <https://scikit-learn.org/stable/index.html>

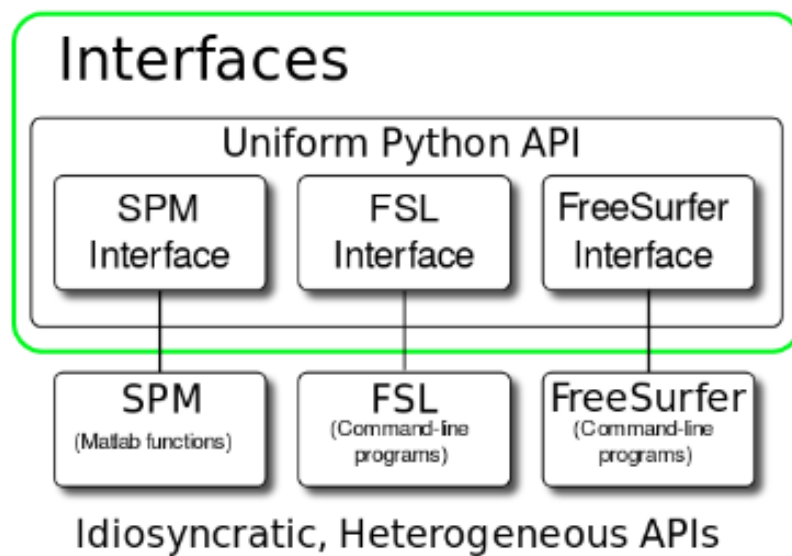


Figure 4. The framework of Nipype. It clusters SPM, FSL, FreeSurfer etc software together and supplies user-friendly interface with Python. The image is adapted from <https://nipype.readthedocs.io/en/latest/index.html>.

```
#Download the data from nilearn
motor_images = datasets.fetch_neurovault_motor_task()
stat_img = motor_images.images[0]

#Plotting the statistics images
from nilearn import plotting
plotting.plot_glass_brain(stat_img, threshold=3)

plotting.show()
```

The plotting result is shown as follows (Fig. 5). More examples with Nilearn, check here²⁷.

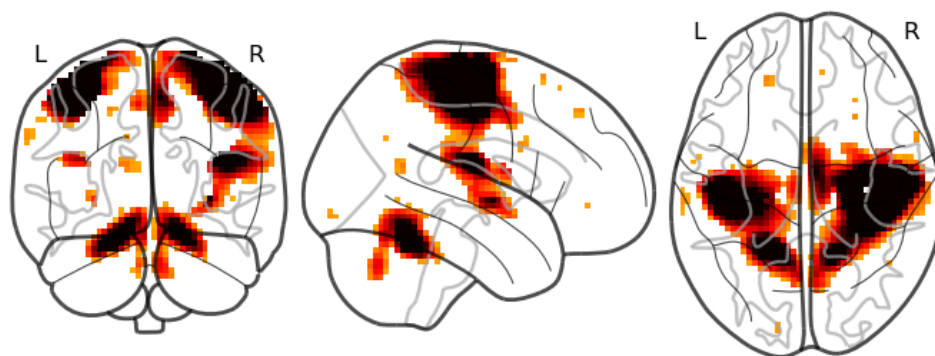


Figure 5. The statistical map of motor task is plotted with glass brain format.

²⁷ http://nilearn.github.io/auto_examples/index.html

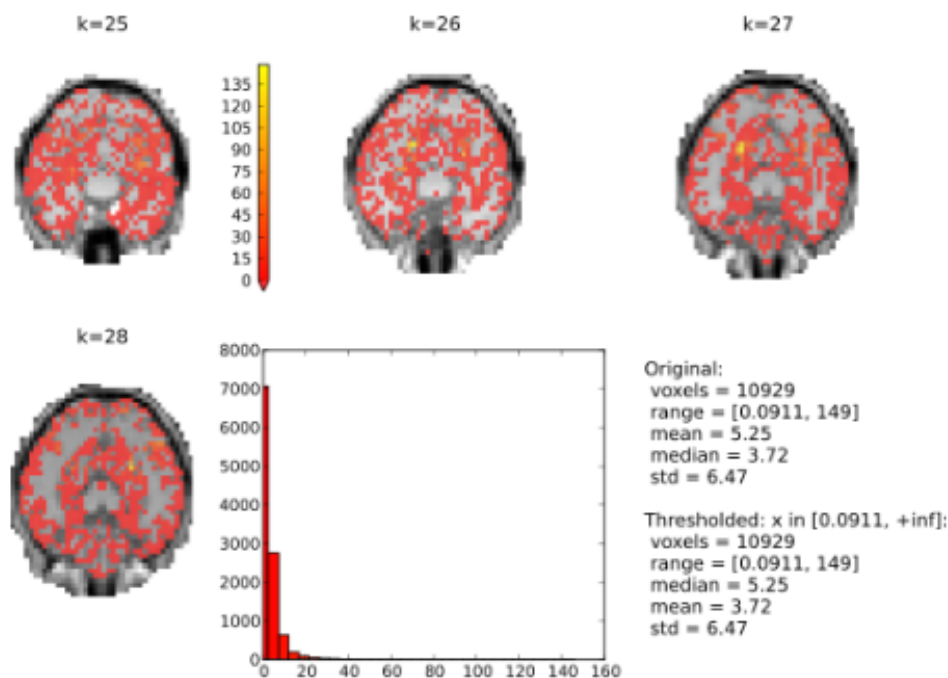


Figure 6. The basic f(MRI) plotting with Pymvpa

9.2. Pymvpa - MultiVariate Pattern Analysis (MVPA) in Python

PyMVPA²⁸ is a Python package intended to more statistical learning analyses of large datasets. It provides more enrich high-level interface to a broad range of algorithms for classification, regression, feature selection, result plot. It is designed to integrate well with related software packages, such as scikit-learn, shogun²⁹, etc. PyMVPA is free software and requires nothing but free-software to run (Hanke et al. 2009). The next code from PyMVPA official website.

The result image looks like as follows (Fig. 6). More source code link here³⁰.

10. Deep learning

As mentioned before, with the development of supercomputer and machine learning, more and more data is produced from society. In the neuroscience research field, the traditional data analysis approaches face a big challenge and bottleneck when the big data fade into it, so that's why deep learning make very popular in the neuroscience research area that specific for fMRI data analysis, through deep learning model that we can train enough data to dig more medical information or cognitive mechanism. When we back to see the published paper in the recent year, the research topic related to deep learning already become a hot topic. So what kind of tools that we can use to neuroscience data analysis, in the next section, I will introduce some famous, classical and friendly model or library that we can use in our research.

²⁸ <http://www.pymvpa.org>

²⁹ <http://www.shogun-toolbox.org/>

³⁰ http://www.pymvpa.org/examples/mri_plot.html

10.1. Tensorflow - Very popular open source deep learning model framework

Tensorflow³¹ is an open source software library for numerical computation using data flow graphs. It provides stable Python API and C APIs as well as without API backwards compatibility guarantee like C++, Java, JavaScript and Swift.

Tensorflow is fully supporting CPU and GPU, it also supports friendly interface with Python, you can through easy install or drive CPU and GPU works (Pattanayak 2017; Gad 2018; Pattanayak 2017).

The core idea of tensorflow is based graphy drive model, more abstract and details that can be described as follows (Fig. 7):

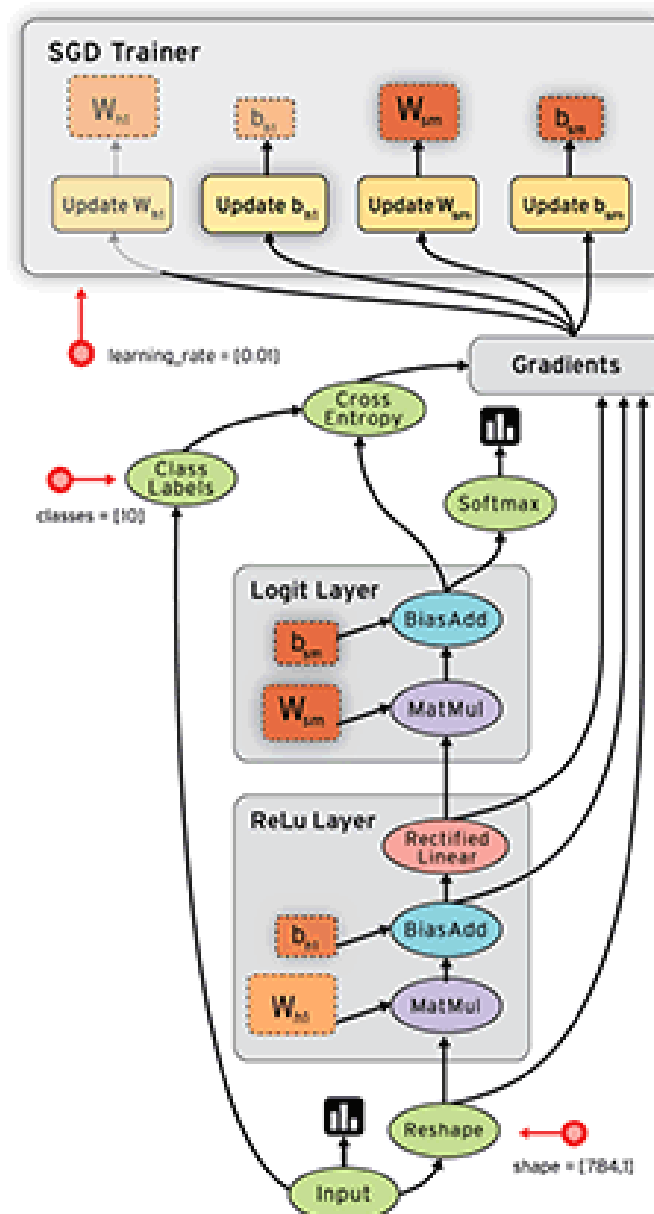


Figure 7. Tensorflow data flow chart. The graph nodes represent mathematical operations, while the graph edges represent the multidimensional data arrays (tensors) that flow between them. The image is adapted from <https://www.tensorflow.org/guide/graphs>.

³¹ <https://www.tensorflow.org/>

More example or tutorials that link <https://www.tensorflow.org/>, you can find more API s or samples about tensorflow.

10.2. Others- The Deep Learning framework

10.2.1. Keras

Keras³² is a high-level neural networks API, written in Python and capable of running on top of TensorFlow, CNTK³³(Microsoft Inc), or Theano³⁴. Keras is also a very popular deep learning framework, it offers a very easy interface with Python. More example and tutorials link <https://keras.io/>.

10.2.2. PyTorch

PyTorch³⁵, an open source deep learning platform that provides a seamless path from research prototyping to production deployment (Ketkar 2017; Chen et al. 2018). PyTorch is an end-to-end deep learning framework, the user of PyTorch is already over tensorflow and keras in some period time, it also has complete and perfect documents and tutorials for getting started PyTorch from zero. More examples link <https://pytorch.org/tutorials/>.

10.2.3. Brain.js

brain.js³⁶ is a library of neural networks written in JavaScript. Front developing is the main and popular software method for display some context, actually, more programmer love JavaScript and use it to training some neural networks, what's more, JavaScript is also the very popular computer programming compared others. So the brain.js that is a deep learning library that based with JavaScript (Wang, Cai, and Wei 2016; van der Spuy 2012). In order to you can have a direct impression for brain.js, more examples link <https://github.com/BrainJS/brain.js/tree/master/examples>.

In this section, I only list some famous and frequently used deep learn frameworks, actually, more deep learning framework is developed according to each demand in various areas today, but for neuroscience data analysis, you can use above frameworks is already can solve your problem, of course, you also can contribute to these deep learning frameworks through Github³⁷, all these frameworks are open sources, welcome to contribute your idea and make it more perfect.

11. Summary

This review that we generalize all methods and software from the ultimate system platform to a high-level deep learning framework, we hope that this review that can give some new student or researchers can easily walk into neuroscience data analysis. In general, the basic approaches or methods in the neuroscience data analysis(specific for neuroimaging data, such as fMRI etc.) are all covered on this paper.

³² <https://keras.io/>

³³ <https://docs.microsoft.com/zh-cn/cognitive-toolkit/>

³⁴ <http://www.deeplearning.net/software/theano/>

³⁵ <https://pytorch.org/>

³⁶ <https://brain.js.org>

³⁷ <https://github.com/>

Acknowledgments: This work was supported in part by the Ministry of Science and Technology of China grant (2015CB351701), the National Natural Science Foundation of China grant (31730039, 81871350), Beijing Natural Science Foundation grant L182055 and the Chinese Academy of Sciences grant (XDBS01000000).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

fMRI	functional Magnetic Resonance Imaging
EEG	Electroencephalography
MEG	Magnetoencephalogram
BIDS	Brain Imaging Data Structure
BOLD	Blood Oxygen Level Dependent
MVPA	MultiVariate Pattern Analysis

Appendix A

All software used in this review or more software in the neuroscience data analysis that we did a website that you can find by linking <https://sinodanish.github.io/brainsoftware/>.

References

1. Eklund, Anders, Thomas Nichols, Mats Andersson, and Hans Knutsson. 2015. "Empirically Investigating the Statistical Validity of SPM FSL and AFNI for Single Subject FMRI Analysis". In 2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI). IEEE. doi:10.1109/isbi.2015.7164132.
2. Pauli, Ruth, Alexander Bowring, Richard Reynolds, Gang Chen, Thomas E. Nichols, and Camille Maumet. 2016. "Exploring FMRI Results Space: 31 Variants of an FMRI Analysis in AFNI FSL and SPM". *Frontiers in Neuroinformatics* 10 (July). Frontiers Media SA. doi:10.3389/fninf.2016.00024.
3. Della-Maggiore, Valeria, Wilkin Chau, Pedro R. Peres-Neto, and Anthony R. McIntosh. 2002. "An Empirical Comparison of SPM Preprocessing Parameters to the Analysis of FMRI Data". *NeuroImage* 17 (1). Elsevier BV: 19–28. doi:10.1006/nimg.2002.1113.
4. Halchenko, Yaroslav O., and Michael Hanke. 2012. "Open Is Not Enough. Lets Take the Next Step: An Integrated Community-Driven Computing Platform for Neuroscience". *Frontiers in Neuroinformatics* 6. Frontiers Media SA. doi:10.3389/fninf.2012.00022.
5. Gorgolewski, Krzysztof J., Tibor Auer, Vince D. Calhoun, R. Cameron Craddock, Samir Das, Eugene P. Duff, Guillaume Flandin, et al. 2016. "The Brain Imaging Data Structure a Format for Organizing and Describing Outputs of Neuroimaging Experiments". *Scientific Data* 3 (June). Springer Nature: 160044. doi:10.1038/sdata.2016.44.
6. Esteban, Oscar, Christopher Markiewicz, Ross W Blair, Craig Moodie, Ayse Ilkay Isik, Asier Erramuzpe Aliaga, James Kent, et al. 2018. "fMRIPrep: a Robust Preprocessing Pipeline for Functional MRI", April. Cold Spring Harbor Laboratory. doi:10.1101/306951.
7. Gorgolewski, Krzysztof, Christopher D. Burns, Cindee Madison, Dav Clark, Yaroslav O. Halchenko, Michael L. Waskom, and Satrajit S. Ghosh. 2011. "Nipype: A Flexible Lightweight and Extensible Neuroimaging Data Processing Framework in Python". *Frontiers in Neuroinformatics* 5. Frontiers Media SA. doi:10.3389/fninf.2011.00013.
8. Hanke, Michael, Yaroslav O. Halchenko, Per B. Sederberg, Stephen José Hanson, James V. Haxby, and Stefan Pollmann. 2009. "PyMVPA: a Python Toolbox for Multivariate Pattern Analysis of FMRI Data". *Neuroinformatics* 7 (1). Springer Nature: 37–53. doi:10.1007/s12021-008-9041-y.
9. Pattanayak, Santanu. 2017. "Introduction to Deep-Learning Concepts and TensorFlow". In *Pro Deep Learning with TensorFlow*, 89–152. Apress. doi:10.1007/978-1-4842-3096-1_2.
10. Gad, Ahmed Fawzy. 2018. "TensorFlow Recognition Application". In *Practical Computer Vision Applications Using Deep Learning with CNNs*, 229–94. Apress. doi:10.1007/978-1-4842-4167-7_6.
11. Pattanayak, Santanu. 2017. *Pro Deep Learning with TensorFlow*. Apress. doi:10.1007/978-1-4842-3096-1.
12. Chen, Kathleen M, Evan M Cofer, Jian Zhou, and Olga G Troyanskaya. 2018. "Selene: a PyTorch-Based Deep Learning Library for Biological Sequence-Level Data", October. Cold Spring Harbor Laboratory. doi:10.1101/438291.
13. Wang, Yao, Wan-dong Cai, and Peng-cheng Wei. 2016. "A Deep Learning Approach for Detecting Malicious JavaScript Code". *Security and Communication Networks* 9 (11). Wiley: 1520–34. doi:10.1002/sec.1441.
14. van der Spuy, Rex. 2012. "Learning JavaScript". In *Foundation Game Design with HTML5 and JavaScript*, 59–110. Apress. doi:10.1007/978-1-4302-4717-3_2.