IOWA STATE UNIVERSITY
Department of Computer Science

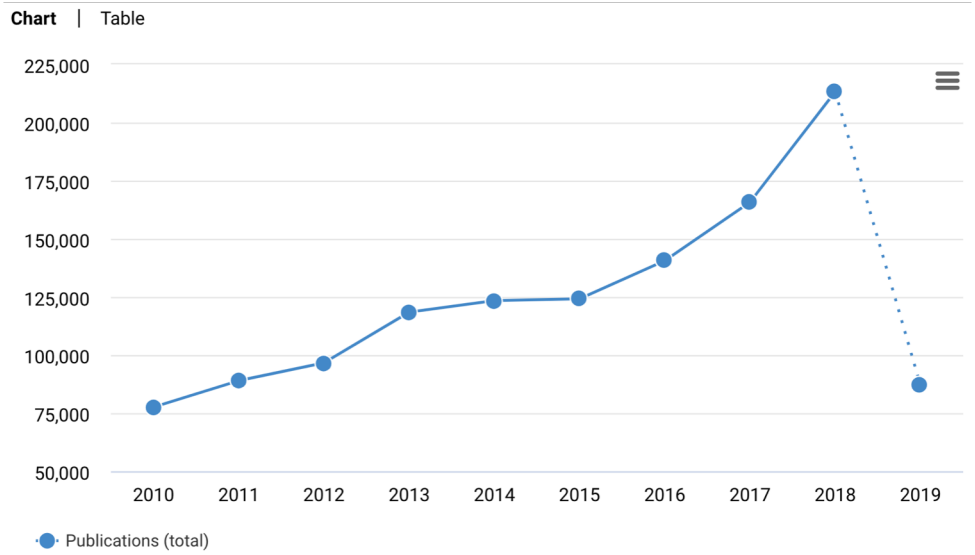
Boa Meets Python: A Boa Dataset of Data Science Software in Python Language

Sumon Biswas, Md Johirul Islam, Yijia Huang and Hridesh Rajan

http://boa.cs.iastate.edu

Data Science Everywhere

Trend of publications with topic “machine-learning”



https://app.dimensions.ai/discover/publication

Top 5 courses in GitHub in 2018

1. Stanford TensorFlow Tutorials
2. Deep Learning Specialization on Coursera
3. Creative Applications of Deep Learning with Tensorflow
4. Practical RL: A course in reinforcement learning in the wild
5. Data Science Coursera

* based on forks

https://github.blog/2018-03-20-top-10-courses-on-githu

Data Science Everywhere

- Data Science projects are growing very fast

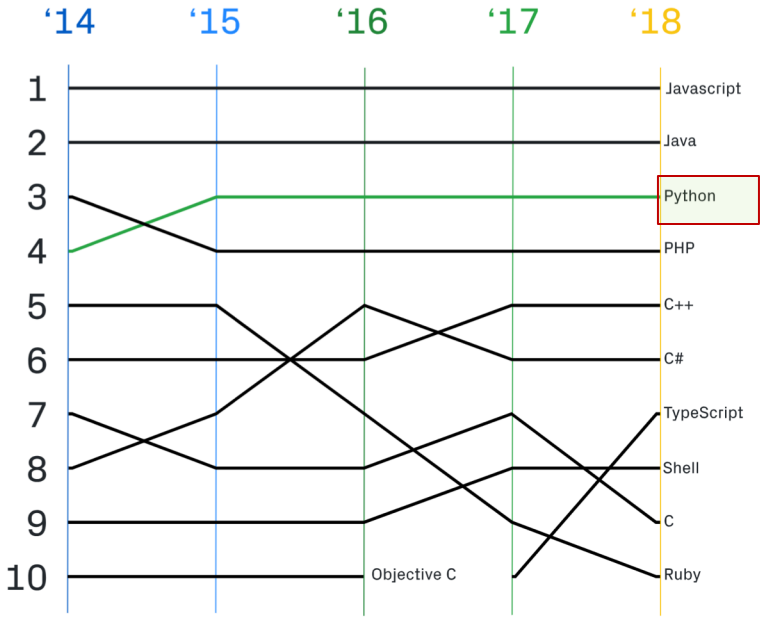
Top topics in **GitHub**

1. react
2. android
3. nodejs
4. docker
5. ios
6. linux
7. angular
8. **machine-learning**
9. electron
10. api

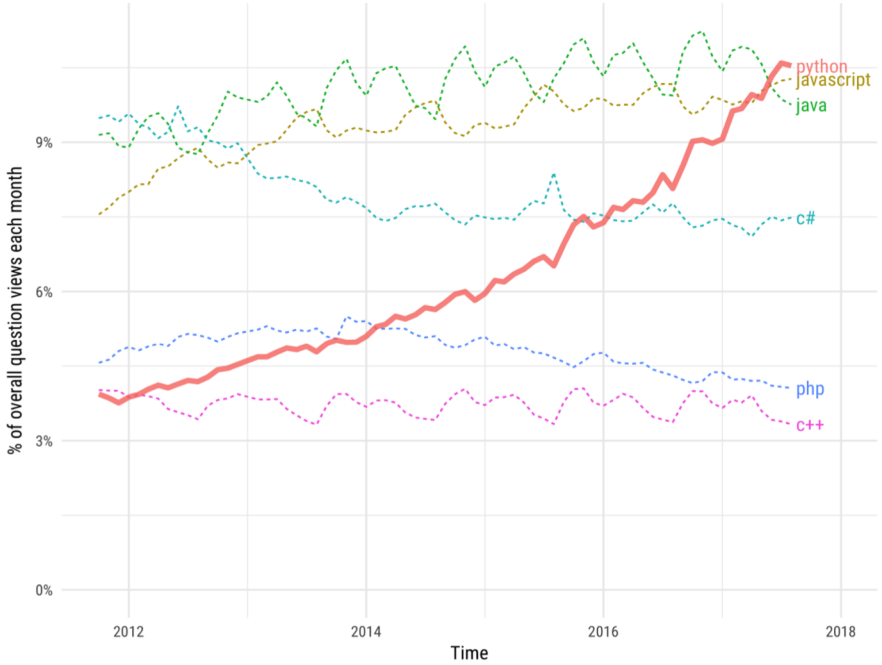
Top growing topics in **GitHub**

1. hacktoberfest
2. **pytorch**
3. **machine-learning**
4. dapp
5. gatsby
6. cryptocurrency
7. terraform-provider
8. easy-to-use
9. smart-contracts
10. exchange

Python in Data Science



Top languages over time in GitHub
https://octoverse.github.com/projects



Growth of programming languages in StackOverflow
https://stackoverflow.blog/2017/09/06/incredible-growth-python/

Motivation

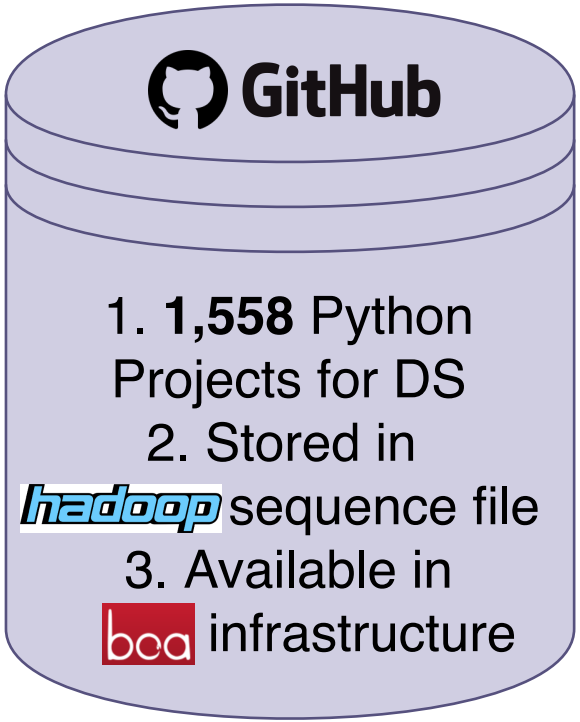
- Lots of Data Science (DS) software
- *Python* is one of the most used languages in DS
  - Lots of packages, easy-to-learn
- **MSR** have been very successful in software engineering
- Availability of benchmarks has historically accelerated research on a topic
  - e.g., Allamanis and Sutton's Java, DaCapo [1], Qualitas [2], etc.

[1] S. M. Blackburn, R. Garner, C. Hoffmann, A. M. Khang, K. S. McKinley, R. Bentzur, A. Diwan, D. Feinberg, D. Frampton, S. Z. Guyer et al., "The DaCapo benchmarks: Java benchmarking development and analysis," in ACM Sigplan Notices, vol. 41, no. 10. ACM, 2006

[2] E.Tempero,C.Anslow,J.Dietrich,T.Han,J.Li,M.Lumpe,H.Melton, and J. Noble, "The Qualitas corpus: A curated collection of Java code for empirical studies," in Software Engineering Conference (APSEC), 2010 17th Asia Pacific. IEEE, 2010

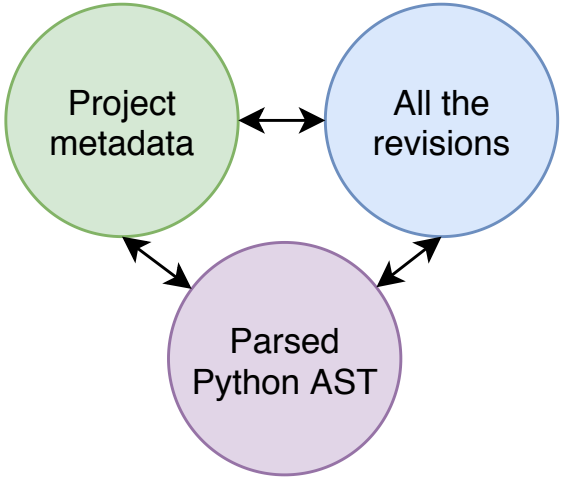
Contributions

1. A large dataset for analyzing *Python* DS projects
2. Efficiently store the dataset in Hadoop sequence file
   - make it memory efficient and
   - parallelly accessible
3. Dataset is publicly available on *Boa* infrastructure



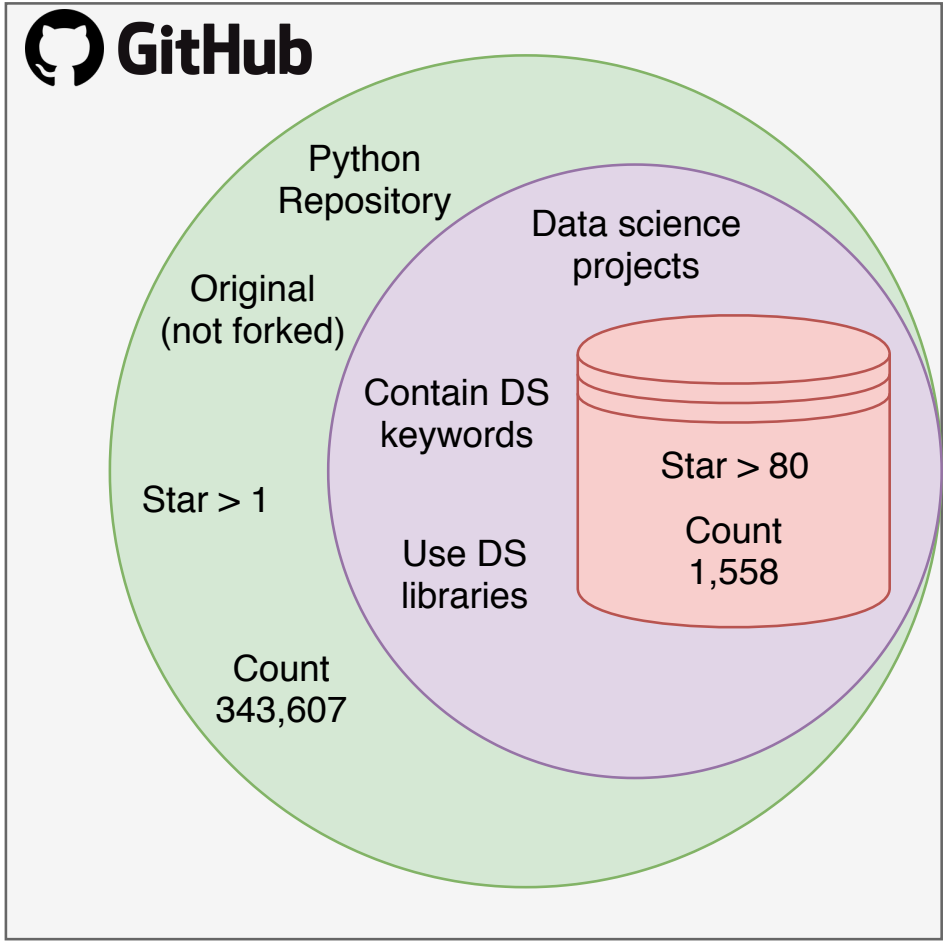
Dataset Metrics

- Top rated projects: Tensorflow, Keras, Pandas, Spacy, Theano etc.
- Projects use at least 33 DS libraries including Pytroch, Caffe, Keras, Tensorflow, XGBoost, NLTK etc.

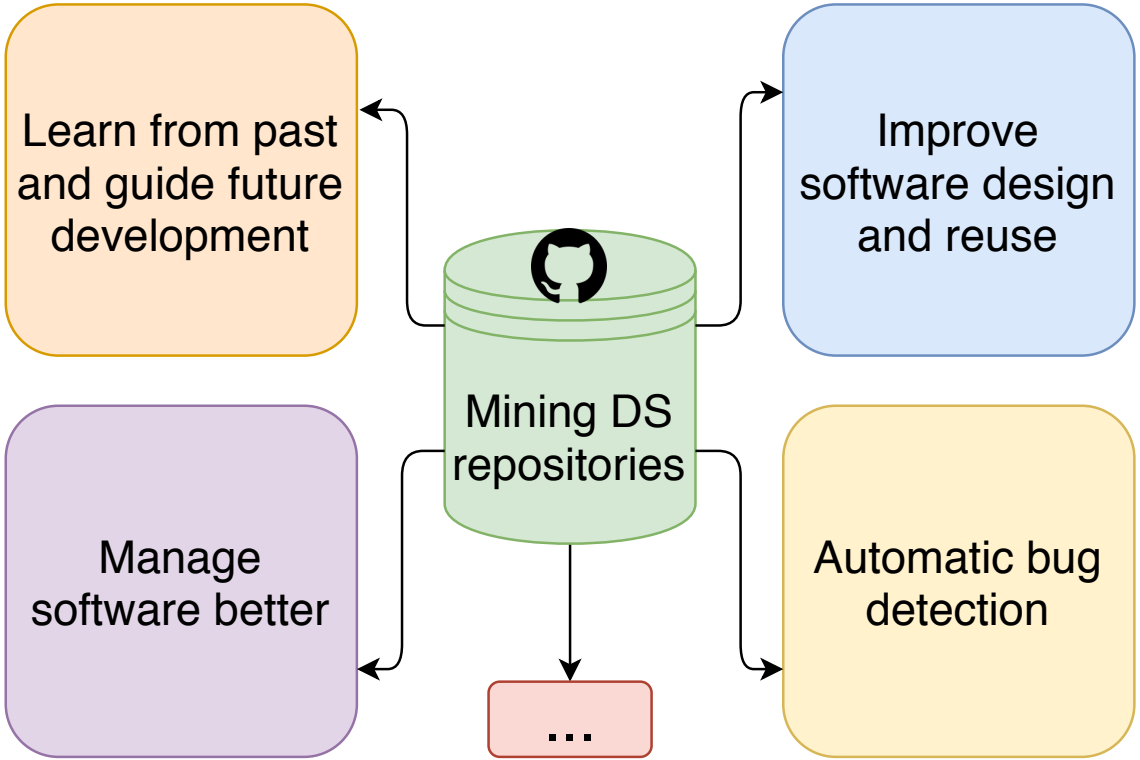


Metric			Count
All repositories	Owner	Organization	350
		Individual user	1,208
	Total		**1,558**
Developers			9,839
Revisions			557,311
Python files (latest snapshot)			86,321
Python files (all revisions)			4,977,680

Methodology

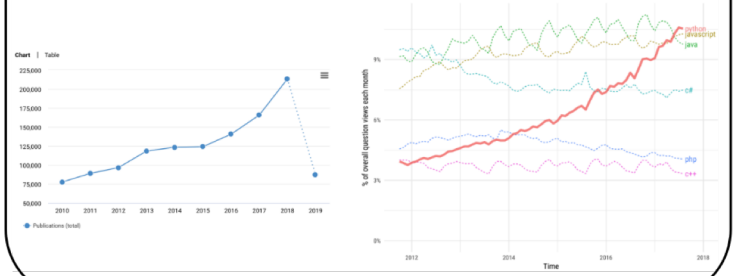


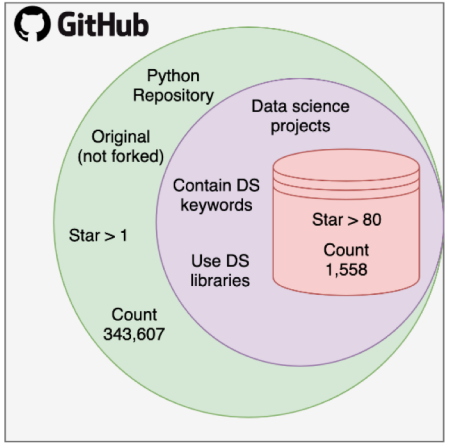
What to Do with the Dataset

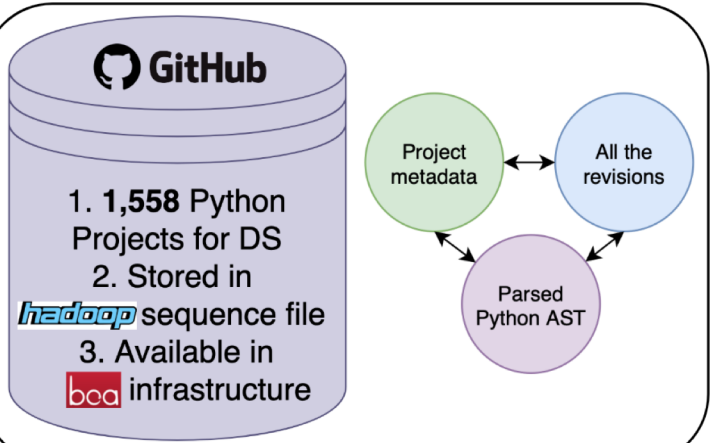


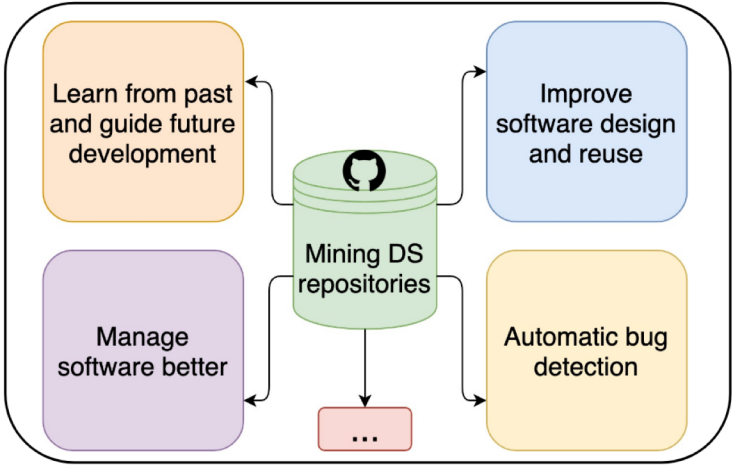
Summary

- DS software are increasing rapidly
- Python is the most popular in DS
- MSR in on DS software is needed





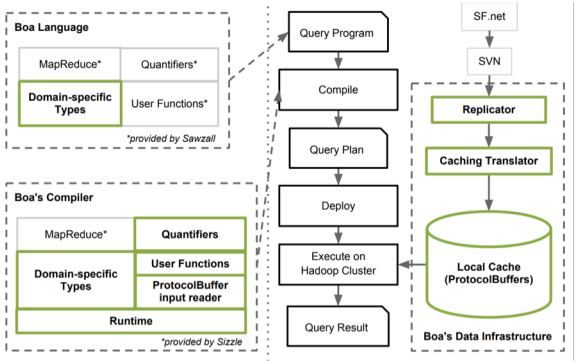




Appendix

Boa - Mining Large Scale Software Repositories

1. Infrastructure

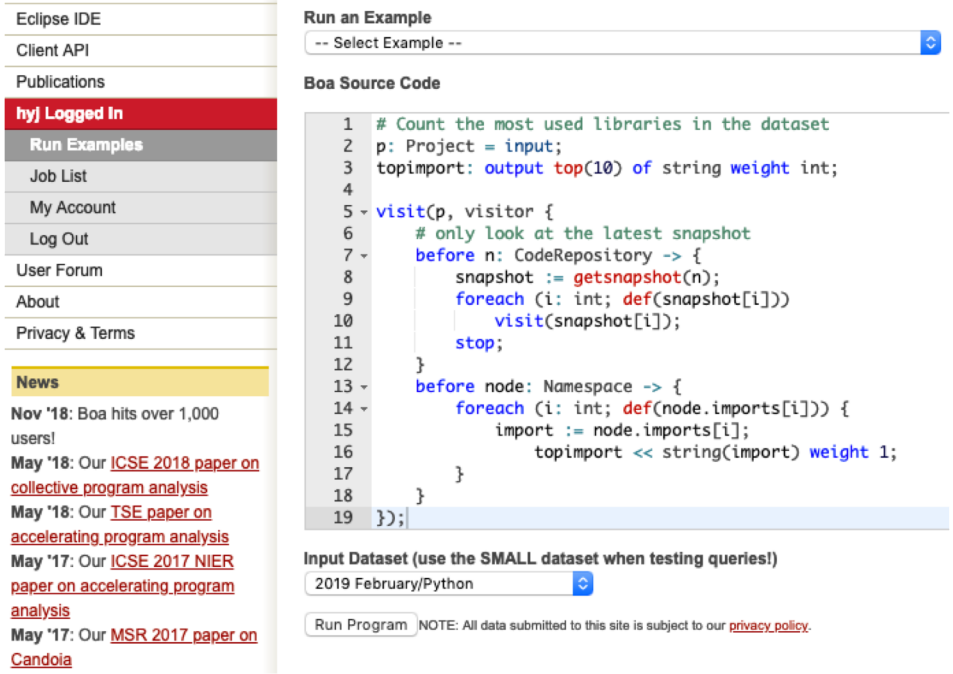


1. Domain-specific language

```
1  # What are the 10 largest projects, in terms of AST nodes?
2  p: Project = input;
3  top: output top(15) of string weight int;
4  astCount := 0;
5  visit(p, visitor {
6      before n: CodeRepository -> {
7          snapshot := getsnapshot(n);
8          foreach (i: int; def(snapshot[i]))
9              visit(snapshot[i]);
10         stop;
11     }
12     before _ -> astCount++;
13     before Project, ChangedFile -> ;
14 });
15
16 top << p.project_url weight astCount;
```

Robert Dyer, Hoan Anh Nguyen, Hridesh Rajan, and Tien N. Nguyen, "Boa: A Language and Infrastructure for Analyzing Ultra-Large-Scale Software Repositories", In the proceedings of the 35th International Conference on Software Engineering (ICSE 2013), May 22, 2013. San Francisco, CA.

Boa Web Based Interface



http://boa.cs.iastate.edu

Data Schema

Fields	Attributes
Project	id, name, created_date, code_repositories, ...
Repository	url, kind, revisions
Revision	id, log, committer, commit_date, files
Person	username, real_name, email
File	name, kind

Fields	Attributes
ASTRoot	imports, namespaces
Namespace	name, modifiers, declarations
Declaration	name, kind, modifiers, parents, fields, methods, ...
Type	name, kind
Method	name, modifiers, return_type, statements, ...
Variable	name, modifiers, initializer, variable_type
Statement	kind, condition, expression, statements, ...
Expression	kind, literal, method, is_postfix, ...
Modifier	kind, visibility, other, ...

Applications - API usage study

API Call Sequences	Count
add, Activation, add, Dropout, add, Dense, add, Activation	115
add, Dense, add, Activation, add, Dropout, add, Dense	114
Dense, add, Activation, add, Dropout, add, Dense, add	112
Conv2d, BatchNorm2d, ReLU, Conv2d, BatchNorm2d, ReLU, Conv2d, BatchNorm2d	103
Sequential, Conv2d, BatchNorm2d, ReLU, Conv2d, BatchNorm2d, ReLU, Conv2d	99
BatchNorm2d, ReLU, Conv2d, BatchNorm2d, ReLU, Conv2d, BatchNorm2d, Lambda	82
Conv2d, BatchNorm2d, ReLU, Conv2d, BatchNorm2d, Lambda, LambdaReduce, ReLU	82
LambdaMap, Sequential, Sequential, Conv2d, BatchNorm2d, ReLU, Conv2d, BatchNorm2d	82
ReLU, Conv2d, BatchNorm2d, ReLU, Conv2d, BatchNorm2d, Lambda, LambdaReduce	82
Sequential, LambdaMap, Sequential, Sequential, Conv2d, BatchNorm2d, ReLU, Conv2d	82