URL Classification using Bag of Features (BoF) of URL bitstream

Keiichi Shima & Hiroshi Abe (IIJ Innovation Institute, Inc.)
Daisuke Miyamoto (Nara Institute of Science and Technology)
Tomohiro Ishihara, Kazuya Okada & Yuji Sekiya (The University of Tokyo)
Yusuke Doi & Hirochika Asai (Preferred Networks, inc.)

CAIDA-WIDE Workshop on 2017-11-20 @ Keio University

Outstanding AI works

- In recent years, AI, more specifically, Deep Learning (DL), is getting notable attention

- Especially in media recognition fields, such as image, voice recognition, etc.

- Some researchers are also trying to apply DL in different fields (e.g. factory robots, games, etc)

- Back to our works, are we getting a benefit from AI technologies?

Difficulties

- DL (or Machine Learning (ML) also) requires information to be converted into vectors
- We call it as a feature vector
- Designing the model of the feature vector requires deep knowledge of the target information domains

Why is DL so hot?

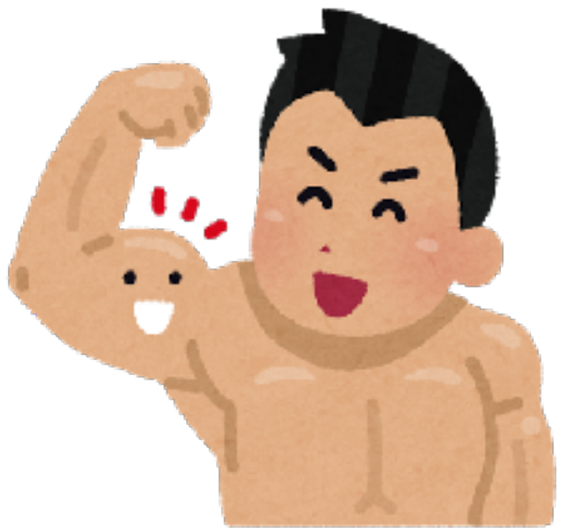
- Because recent DL applications don’t require to extract features manually

- A neural network learns which parts of information are important from a lot of examples

- For example, we can just throw the binary photo data into a neural network and that’s it

- Well, it is not that simple, anyway :)

What we are



We are not good at feature extraction

We have computers

Don’t think Just try

We’ve established the Muscle Learning (ML) team in WIDE

What we try to achieve

- We are thinking if we can apply the similar approach used for image recognition to network information

- Just put (almost) raw data and let the machines extract features

- No need to achieve domain specific deep knowledge before analyzing

Back to URLs

- Phishing is one of the major techniques to steal personal information
  - 1,220,523 attacks were reported in 2016 (*1)
- There are several services to defend
  - URL whitelisting
  - Contents investigation

(*1) Anti Phishing WG report: http://docs.apwg.org/reports/apwg_trends_report_q4_2016.pdf

URL features?

- Challenges
  - Is there any hidden features in the URL strings used for phishing sites?
  - Is it possible to distinguish “white” URLs and “black” URLs by just looking at the URL strings?
- We try to vectorize URLs to use as input information of ML methods without any specific domain knowledge

How to vectorize?

`www.iij.ad.jp/index.html`

↓ Split characters

`w w w . i i j . a d . j p / i n d e x . h t m l`

↓ Convert the URL into HEX values

`7777772E69696A2E61642E6A703F696E6465782E68746D6C`

↓ Extract 8-bits values by shifting 4 bits in the HEX values

```
77,77,77,77,77,72,2E,        3F,F6,69,96,6E,E6,64,
E6,69,96,69,96,6A,A2,        46,65,57,78,82,2E,E6,
2E,E6,61,16,64,42,2E,        68,87,74,46,6D,D6,6C
E6,6A,A7,70
```

Count the number of unique values for the host part and the URL path part respectively (Bag of features)

How to vectorize?

www.iij.ad.jp

16	1	2E	3
42	1	61	1
64	1	69	2
6A	2	70	1
72	1	77	5
96	2	A2	1
A7	1	E6	3

256 dimensional sparse vector

index.html

2E	1	46	1
57	1	65	1
68	1	6C	1
6D	1	74	1
78	1	82	1
87	1	D6	1
E6	1		

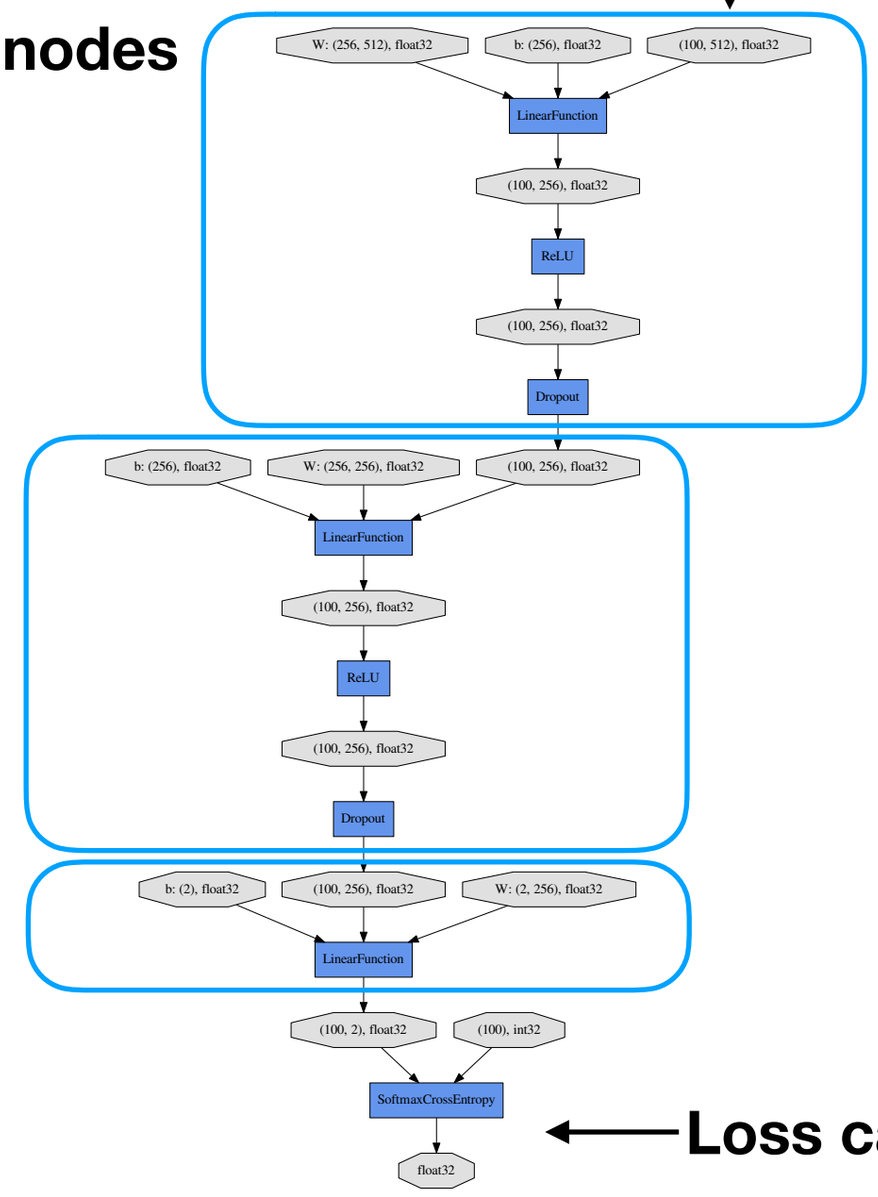
256 dimensional sparse vector

512 dimensional sparse vector

Neural network topology

A 512 dimensional vector generated from a URL string

Linear mapping to 256 nodes



Linear mapping to 256 nodes

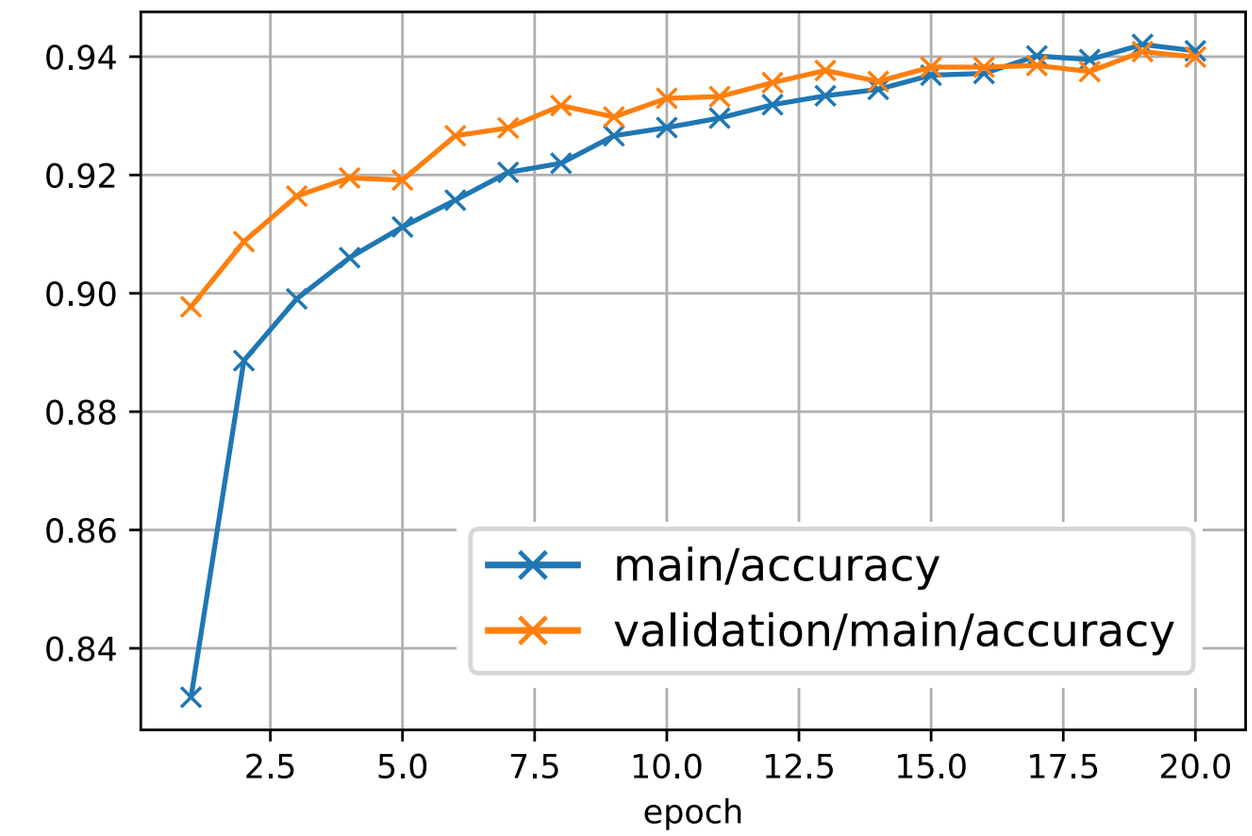
Reduction to 2 nodes

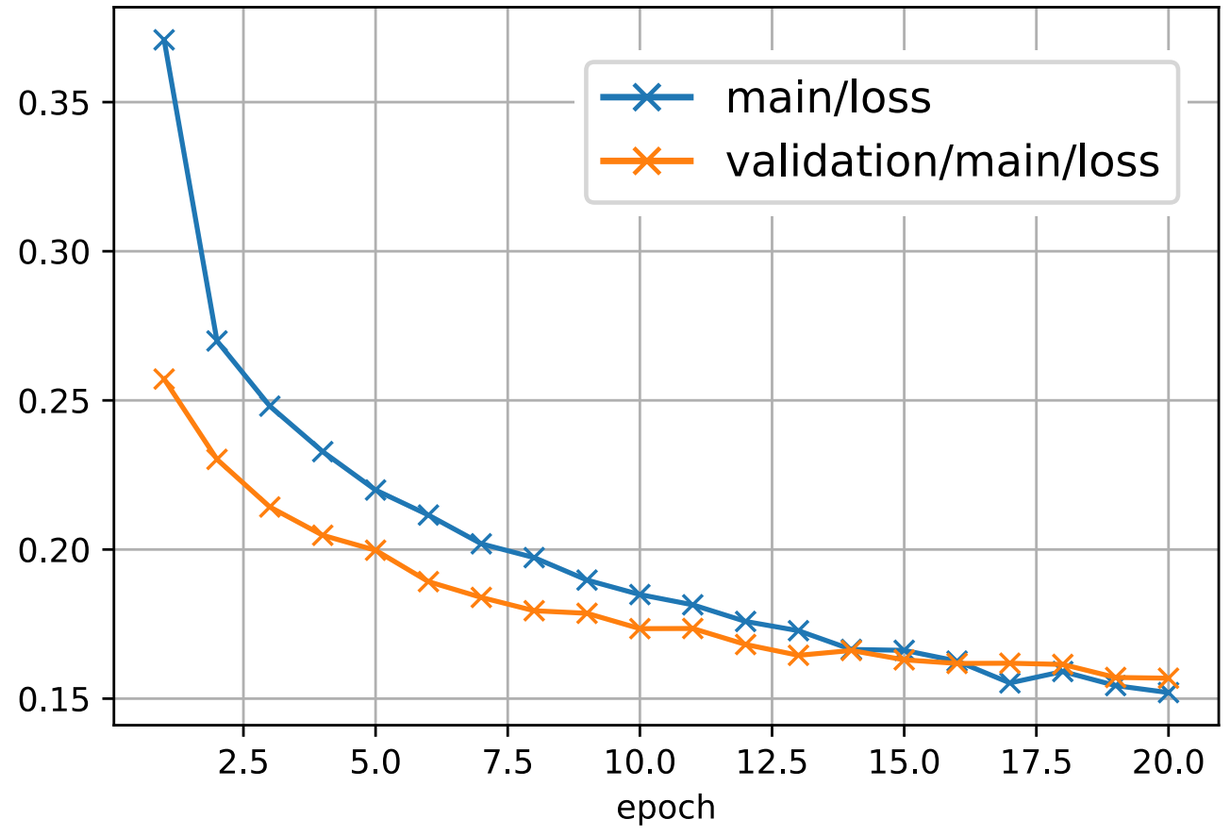
Loss calculation

Classify using the neural network

- Datasets
  - 26722 “black” URLs downloaded from www.phishtank.com which are active phishing site URLs as of 2017-4-24
  - 175290 “white” URLs captured at a research network
- Method
  - Convert all the URLs into vectors and shuffle them
  - 10% of them were used for the DNN training and the rest were used for validation

Accuracy and Loss





(a) Our method (optimizer = Adam)

Related Work

- S. Garera, N. Provos, M. Chew, and A. D. Rubin, “A framework for detection and measurement of phishing attacks,” in Proceedings of the 2007 ACM Workshop on Recurring Malcode, ser. WORM ’07. New York, NY, USA: ACM, November 2007, pp. 1–8.

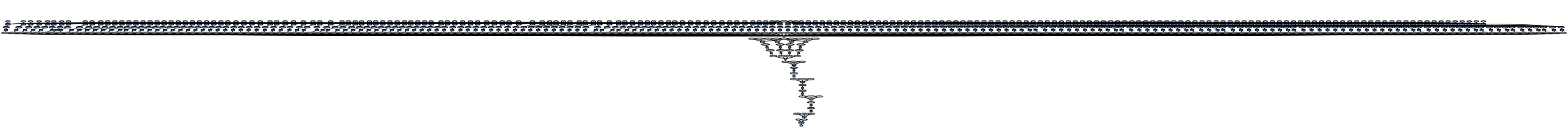
- J. Ma, L. K. Saul, S. Savage, and G. M. Voelker, “Beyond blacklists: Learning to detect malicious web sites from suspicious URLs,” in Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ser. KDD ’09. New York, NY, USA: ACM, June 2009, pp. 1245–1254.

- P. Prakash, M. Kumar, R. R. Kompella, and M. Gupta, “PhishNet: Predictive blacklisting to detect phishing attacks,” in 2010 Proceedings IEEE INFOCOM, ser. INFOCOM, 2010, pp. 1–5.

- B. Sun, M. Akiyama, T. Yagi, M. Hatada, and T. Mori, “AutoBLG: Automatic URL blacklist generator using search space expansion and filters,” in 2015 IEEE Symposium on Computers and Communication, ser. ISCC, July 2015, pp. 625–631.

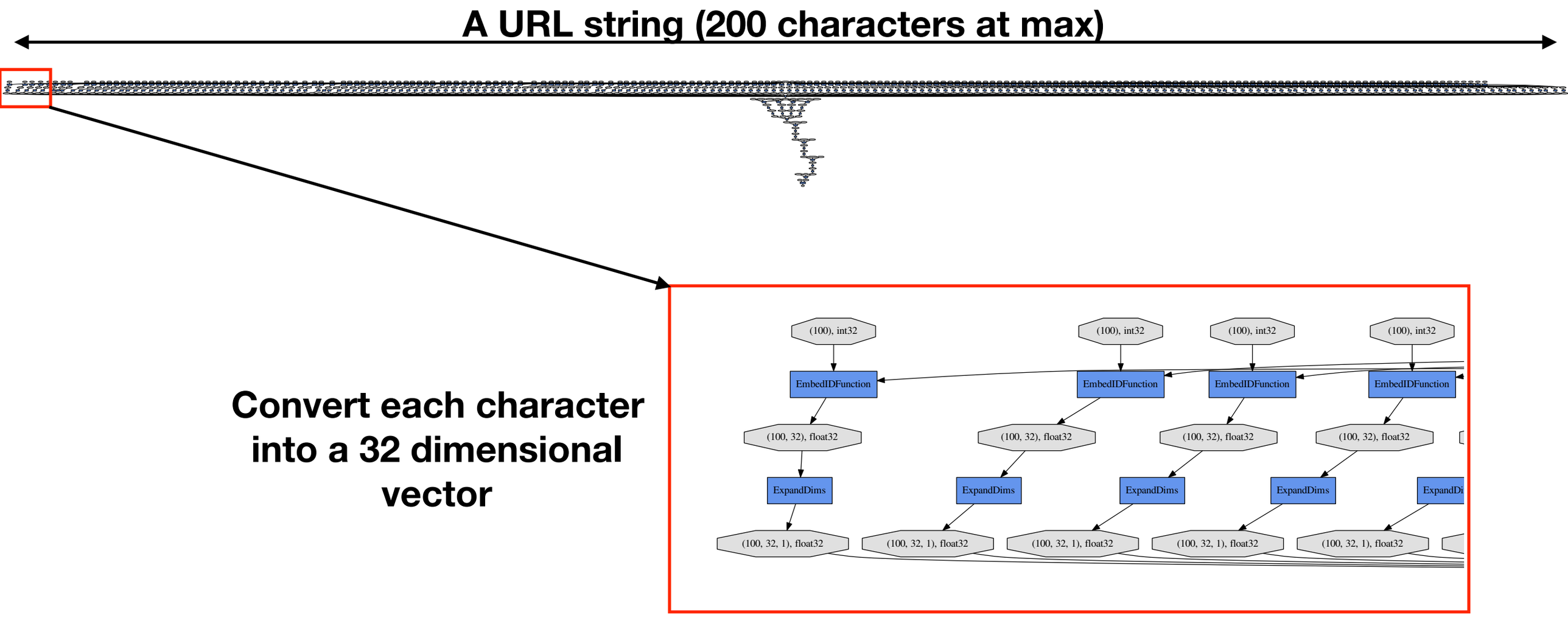
- J. Saxe and K. Berlin, “eXpose: A character-level convolutional neural network with embeddings for detecting malicious URLs, file paths and registry keys,” CoRR, vol. abs/1702.08568, February 2017.

The neural network topology of eXpose



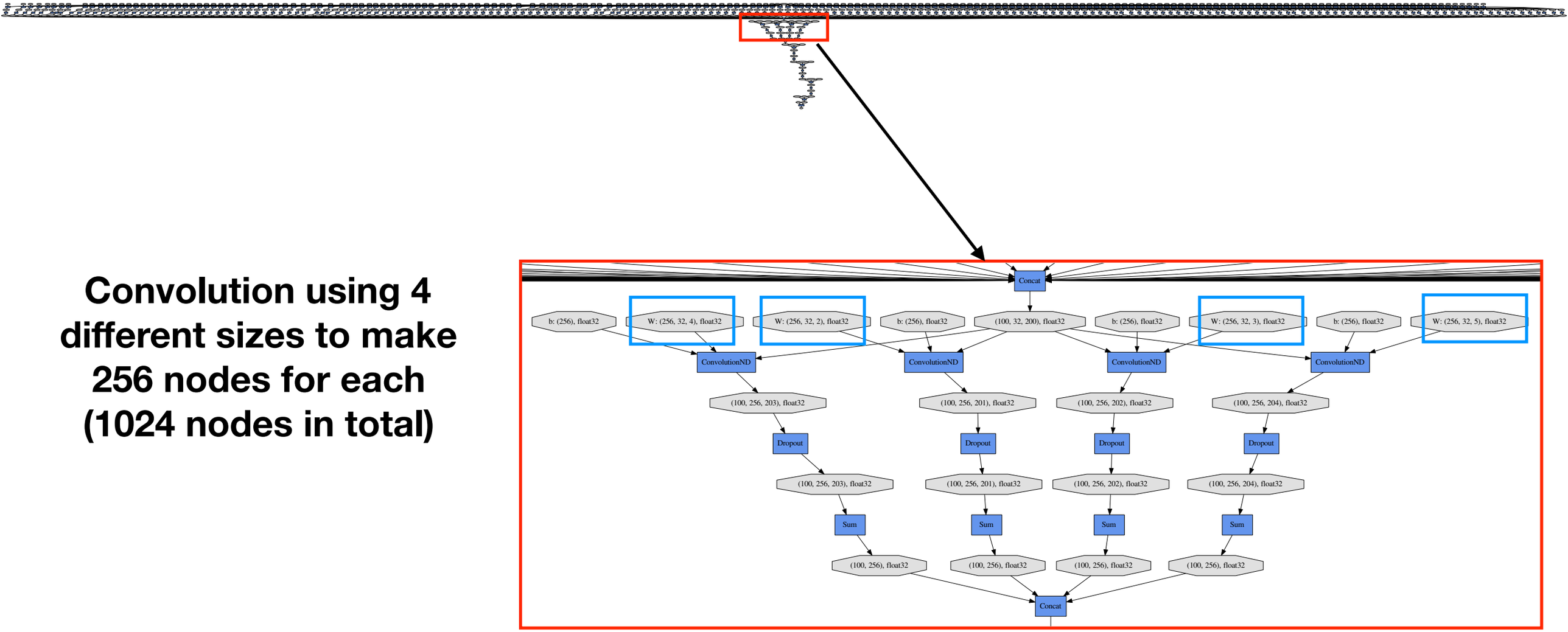
?

The neural network topology of eXpose



Convert each character into a 32 dimensional vector

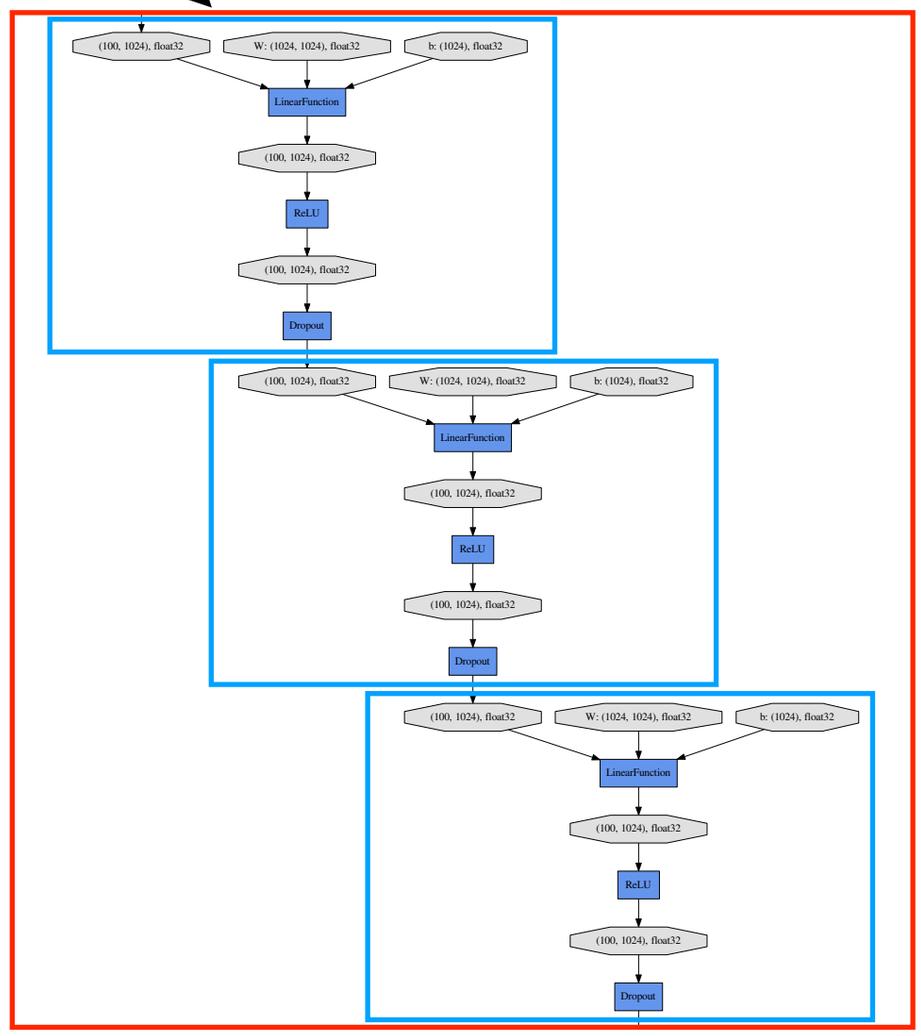
The neural network topology of eXpose



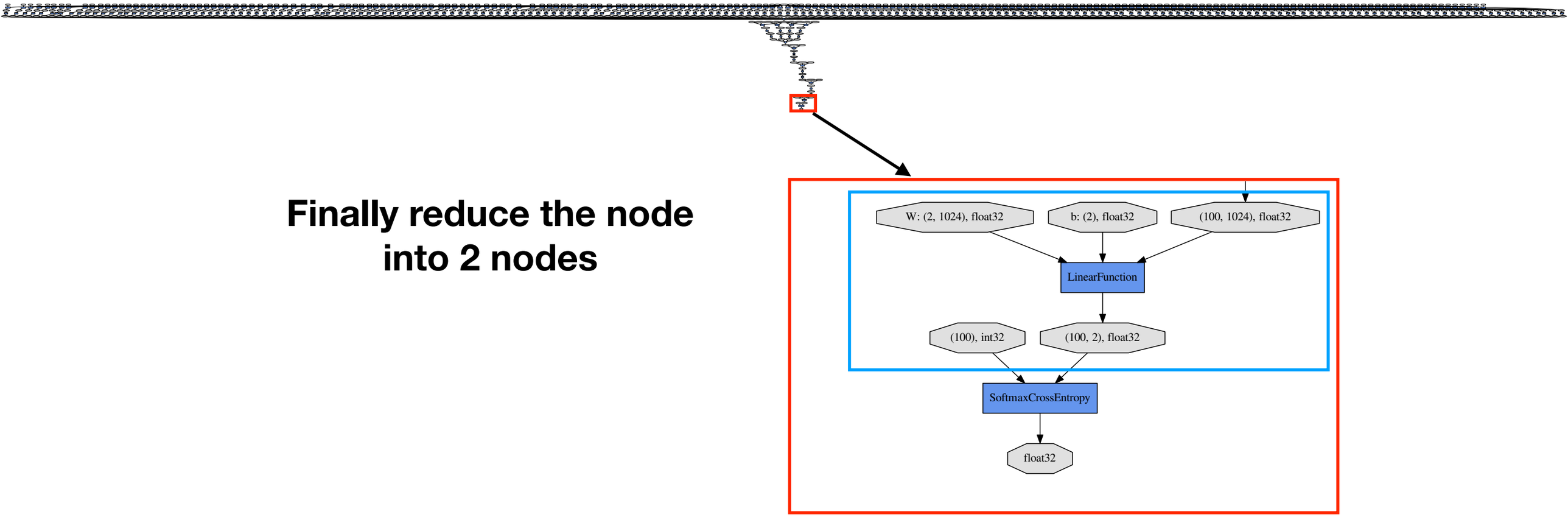
Convolution using 4 different sizes to make 256 nodes for each (1024 nodes in total)

The neural network topology of eXpose

Perform 1024 to 1024 linear mapping 3 times



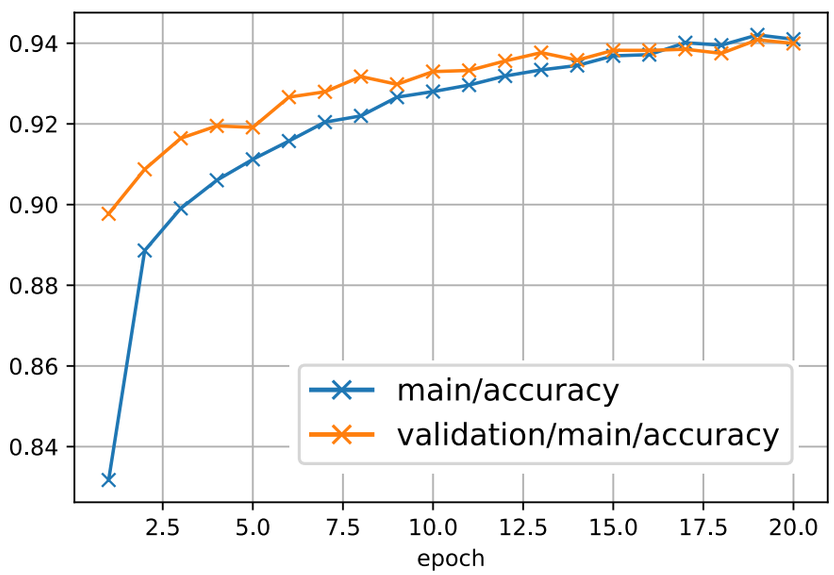
The neural network topology of eXpose

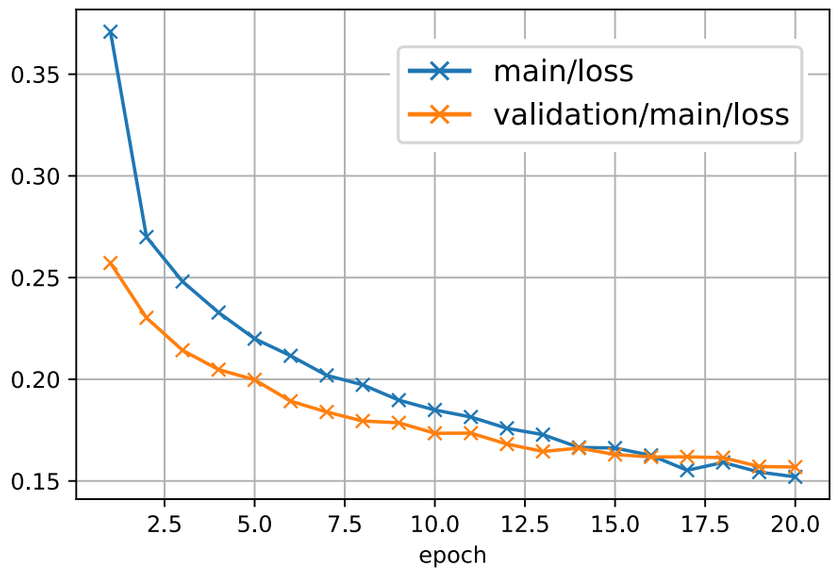


Finally reduce the node into 2 nodes

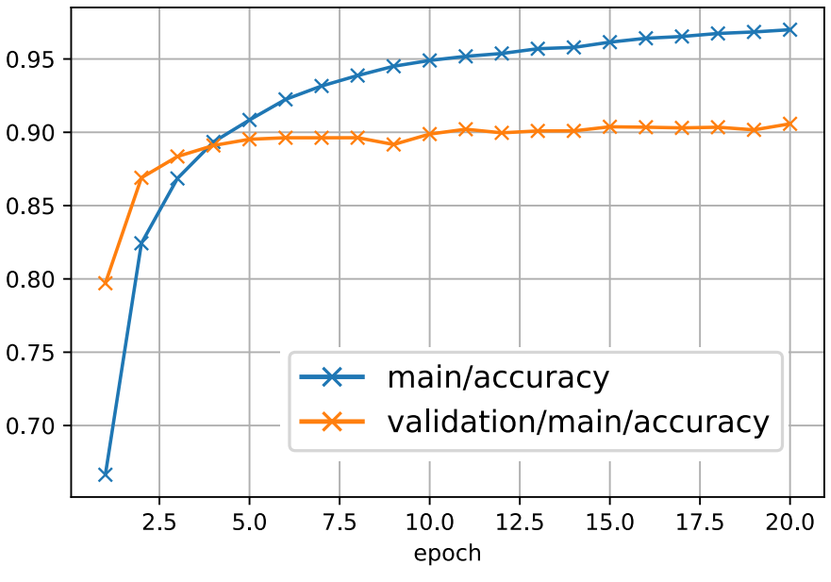
OK, great

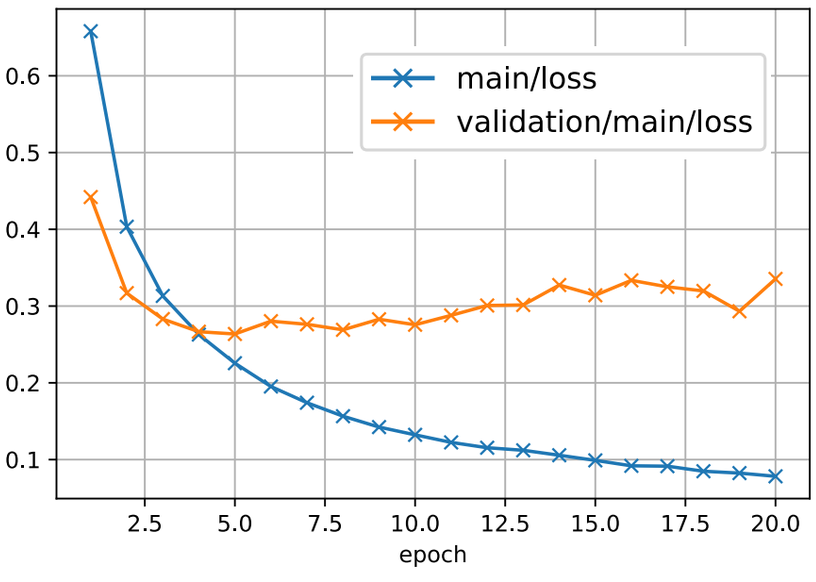
Comparison with the same dataset





(a) Our method (optimizer = Adam)





(b) eXpose (optimizer = Adam)

Summary

- We are trying to utilize Deep Learning technologies for network information
- The goal is to provide better vectorization mechanisms for network data that don’t require any domain specific knowledge
- The proposed URL vectorization works with some limited sets of data, but can be improved more
- We will explore further