Deep Learning for Enterprise API Traffic

Mamoon Yunus, Forum Systems

Poster Presentation Video: https://youtu.be/oQV7Dwxs5-w

What are APIs?

- HTTP-based distributed communication
- Cloud Amazon EC2, IoT Nest, Fortune 500
- API Gateways used to control traffic flow

Problem Statement

- HTTP-based API traffic is growing
- API management is coarse-grain with labor-intensive policy authoring
- DNNs provide fine-grain, autonomous policy creation for better QoS & Security

Data Pre-processing

- $\sim$ 5 Million lines of API gateway logs
- $\sim$ 20,000 HTTP request-response pairs
- $\sim$ 40 HTTP categorical features
- $\sim$ 45 minutes of real-time data

Sample Data & Features

	Protocol	Scheme	Method	Client	short_URL
0	HTTP/1.1	https	POST	10.78.36.12	/CareCredit/getAccount
1	HTTP/1.1	https	POST	10.78.36.78	/MWEncryptD/services/TokenService
2	HTTP/1.1	https	POST	10.78.36.78	/MWEncryptD/services/TokenService
3	HTTP/1.1	https	POST	10.78.36.12	/CareCredit/iOvationRegister
4	HTTP/1.1	https	POST	10.78.36.12	/CareCredit/GetAllAcco
5	HTTP/1.1	https	POST	10.78.36.12	/CareCredit/getAccount

Features	Encoding	Embedding	Combined
37	19085	7400	26485
17	10317	3400	13717
8	322	1600	1922

Target: Remote Latency

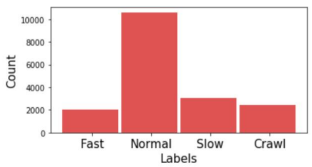
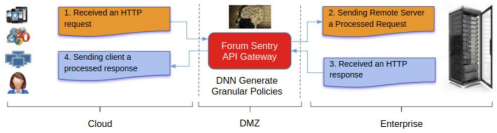
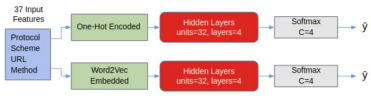


Figure 1: Remote Latency Classification

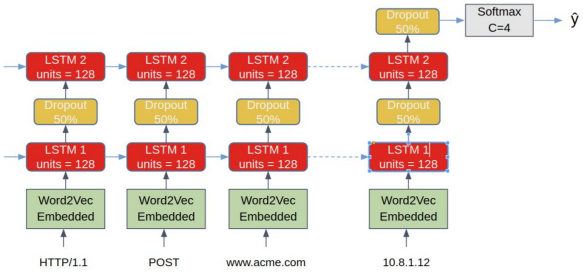
DNN-based API Gateway Deployment Architecture



Encoded & Embedded DNN Architecture



LSTM API Latency Sentiment Architecture



Performance Comparison

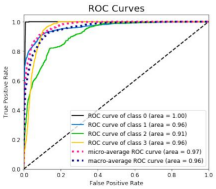


Figure 2: 1-hot encoding

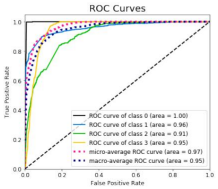


Figure 3: Word embedding

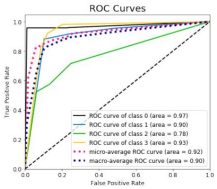


Figure 4: LSTM Sentiment

Experiments

- Adjust for class imbalance
- Remove non-essential features (Cookies, Auth)
- Input Features impact (8-37)
- Embedding Vector Size function of vocabulary
- Embedding Context Window Size
- Dropout vs. Regularization

PCA for Word2Vec

Accuracy Results

Model	Train	Test
Encoded-Regression ($\lambda = 5.0$)	0.81	0.79
Embedded Regression ($\lambda = 0$)	0.81	0.80
Encoded-Classification ($\lambda = 0.005$)	0.85	0.85
Embedded-Classification ($\lambda = 0.0$)	0.84	0.85
LSTM latency Sentiment (dr = 0.5)	0.82	0.83

Class	Precision	Recall	F-1
0	0.94	0.99	0.97
1	0.93	0.91	0.92
2	0.71	0.55	0.62
3	0.64	0.84	0.72

Conclusions

- Predictive performance of **3x** reduced Dense vs Sparse HTTP headers is similar
- Dense representation has a regularization effect
- Calibrate dense vector size with vocabulary

Future Work

- Try GloVe, CNN with API-embedding
- LSTM models with Attention Mechanism
- Additional data, targets & deeper HTTP inputs