twoSIX TECHNOLOGIES

Automatic Prioritization of Incident Reports

Chae Clark

chae.clark@twosixtech.com

TWOSIXTECH.COM

This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA). The views, opinions, and/or findings expressed are those of the author(s) and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

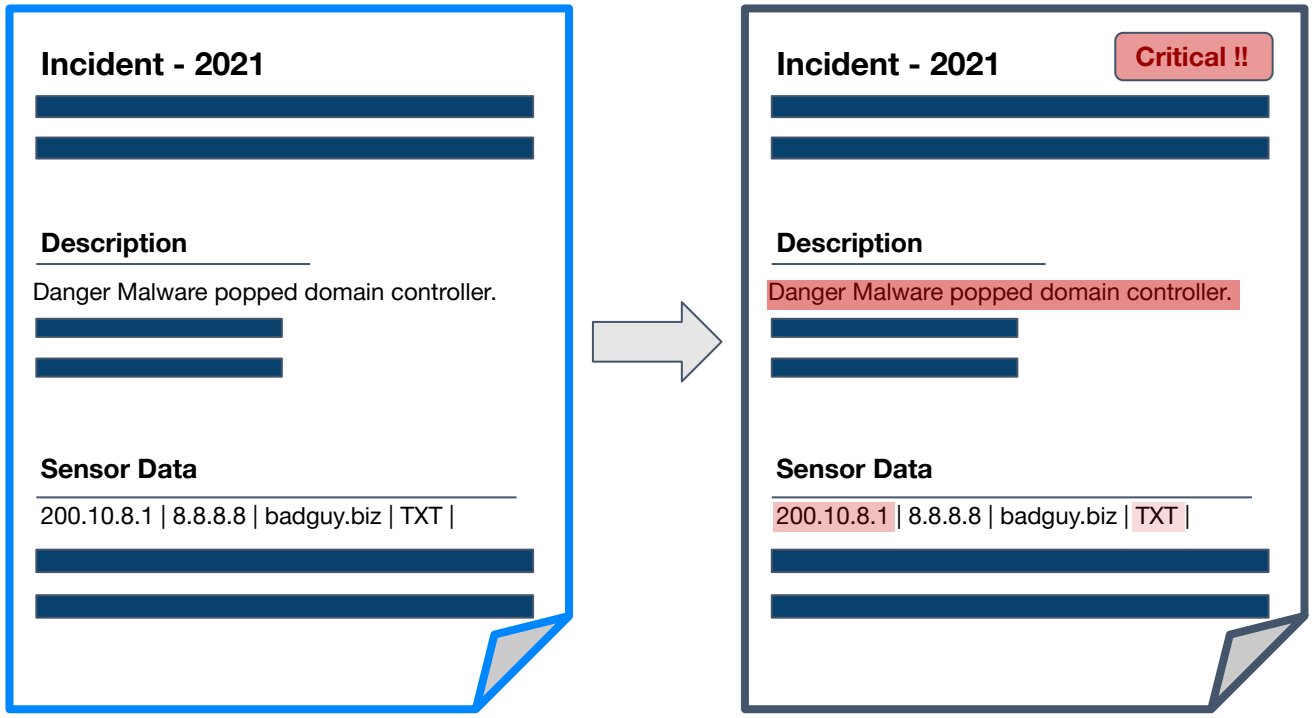
Prioritizing Network Incidents

Prioritizing and Explaining

We develop a Neural Network Regressor to prioritize network incidents

We employ a self-attention layer to produce explainable predictions

Important distinction is that we are NOT “detecting” malicious activity



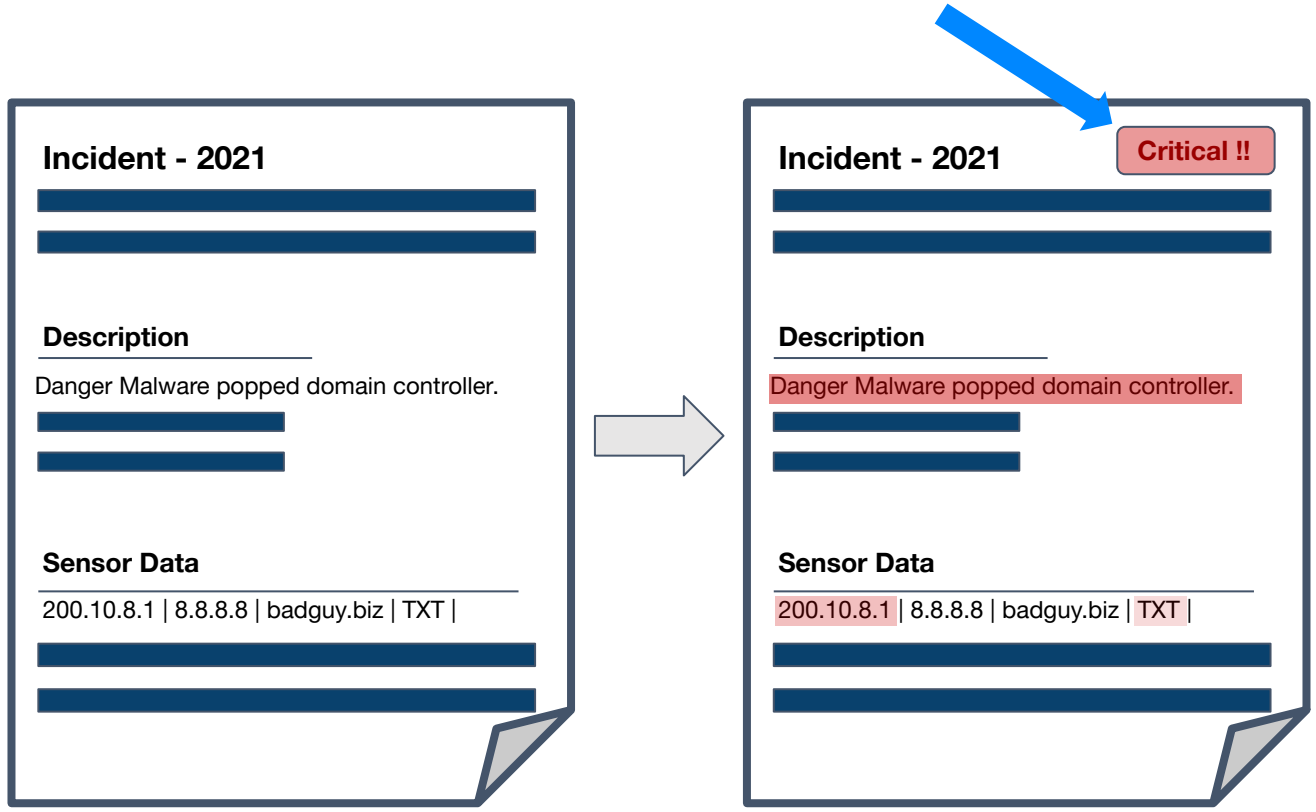
Prioritizing Network Incidents

Prioritizing and Explaining

We develop a Neural Network Regressor to prioritize network incidents

We employ a self-attention layer to produce explainable predictions

Important distinction is that we are NOT “detecting” malicious activity



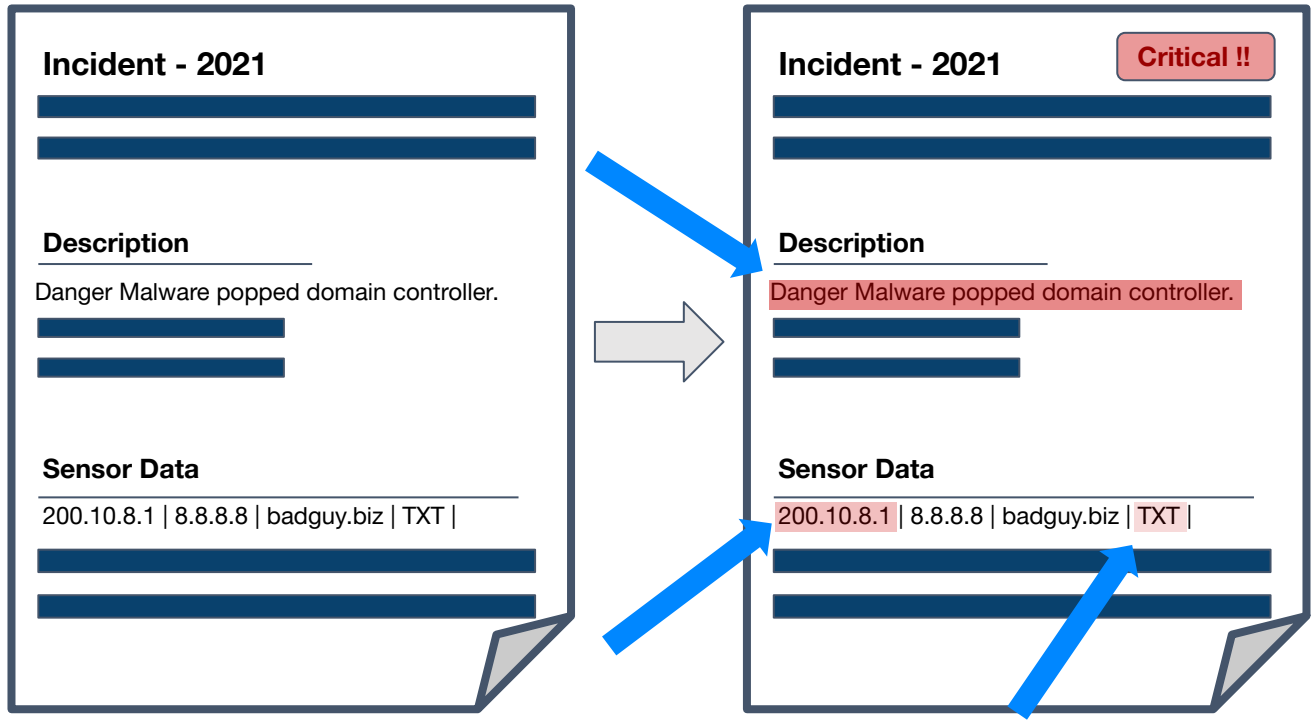
Prioritizing Network Incidents

Prioritizing and Explaining

We develop a Neural Network Regressor to prioritize network incidents

We employ a self-attention layer to produce explainable predictions

Important distinction is that we are NOT “detecting” malicious activity



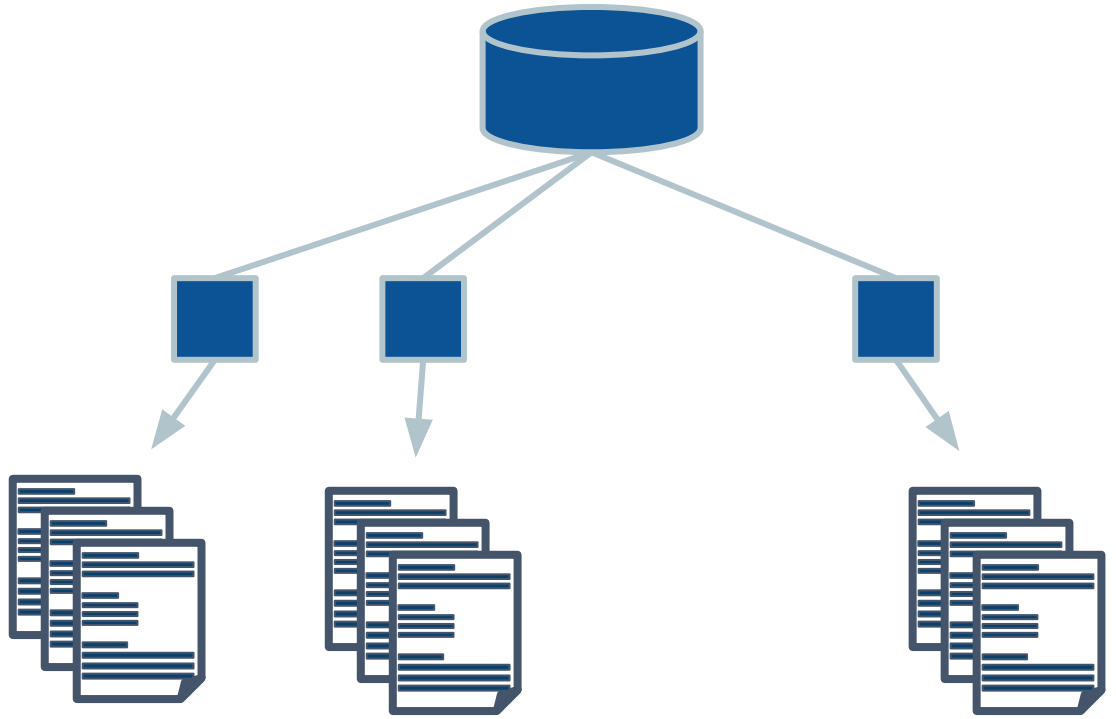
Historical Incident Reports

Training Data

We have access to over 1.5 decades of network incidents.

These (roughly) 30,000 incidents cover several enterprise networks.

They contain a severity designation made by human analysts (along with other metadata)



Input Features

We extract 8 features from the reports

TTP - the categorized attack as detailed in the MITRE framework

Connection Success - whether the intrusion/exfil/etc. was successful

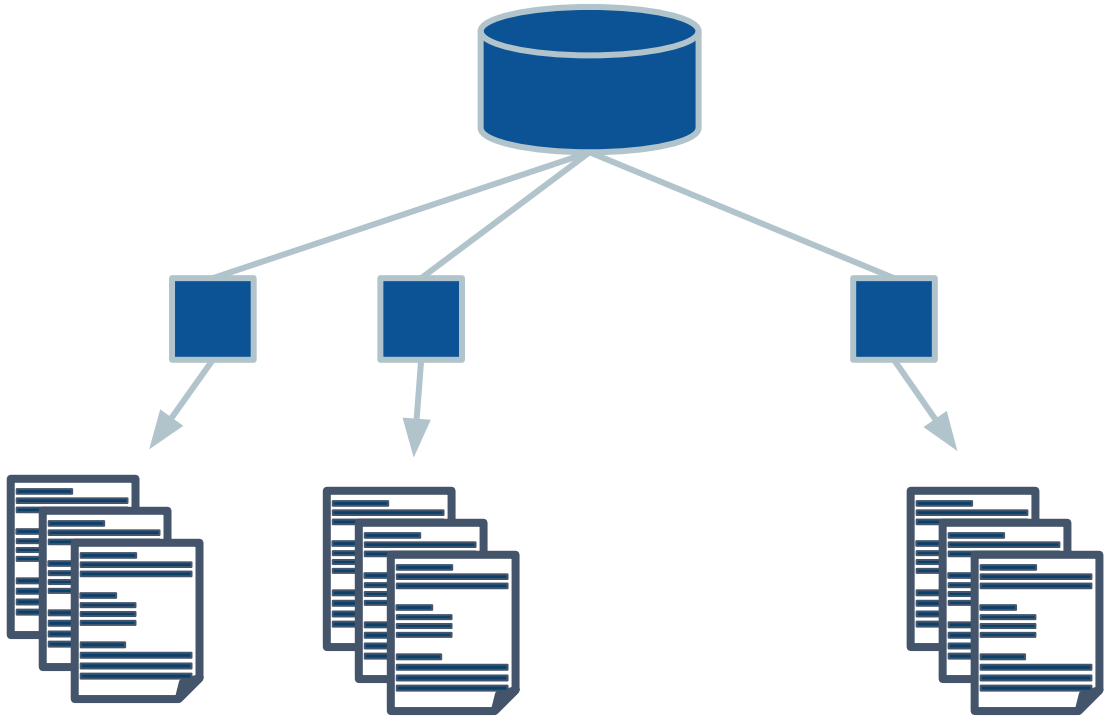
Duration - the amount of time the attack was active on the system

Src./Dst. Role - role within the enterprise (admin, server, external, etc.)

Service - resource used during the alert (http, dns, ssh, etc.)

Location - the physical or virtual location targeted in the intrusion

Description - the full textual description of the event



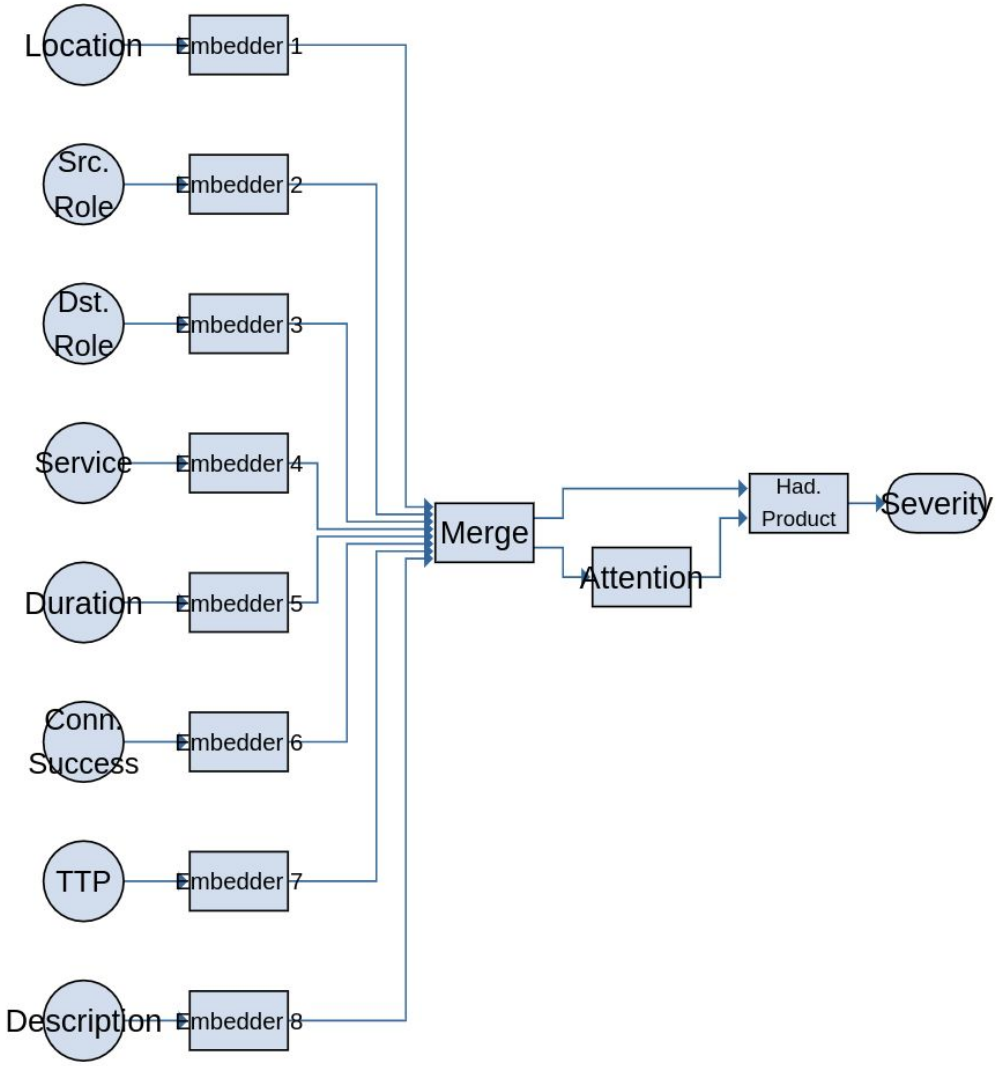
Model Architecture

Prioritizing and Explaining

We develop a Neural Network Regressor to prioritize network incidents

We employ a self-attention layer to produce explainable predictions

We use a single output node instead of a classifier due to the outputs being ordinal



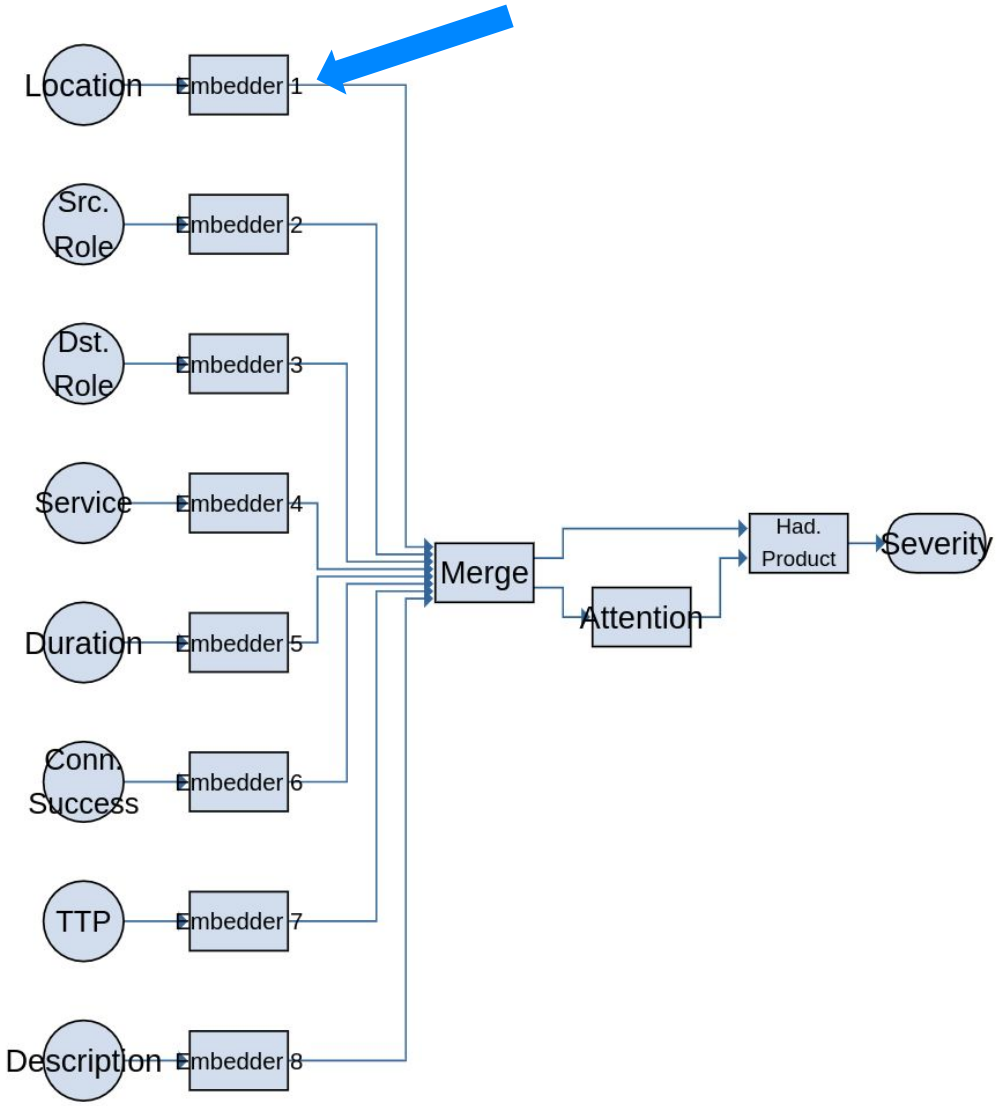
Model Architecture

Prioritizing and Explaining

We develop a Neural Network Regressor to prioritize network incidents

We employ a self-attention layer to produce explainable predictions

We use a single output node instead of a classifier due to the outputs being ordinal



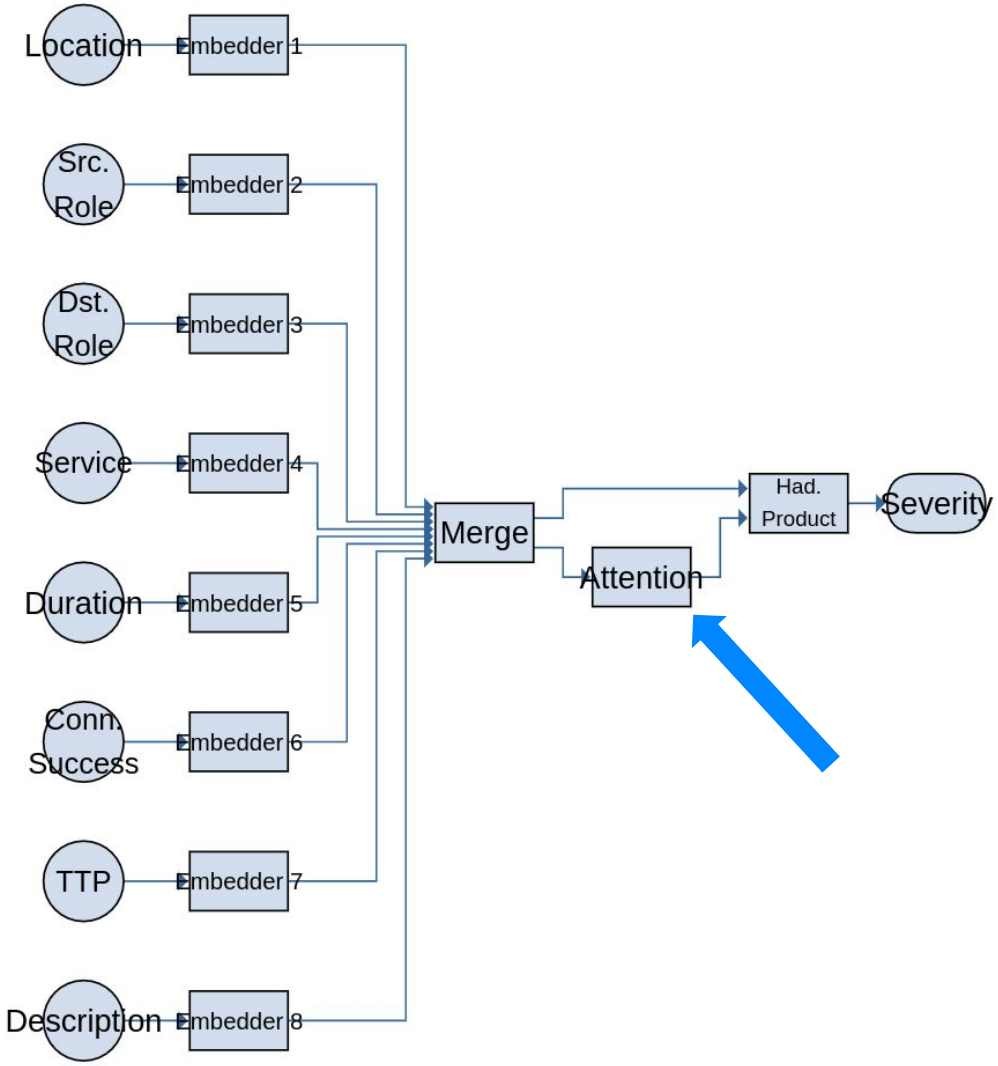
Model Architecture

Prioritizing and Explaining

We develop a Neural Network Regressor to prioritize network incidents

We employ a self-attention layer to produce explainable predictions

We use a single output node instead of a classifier due to the outputs being ordinal



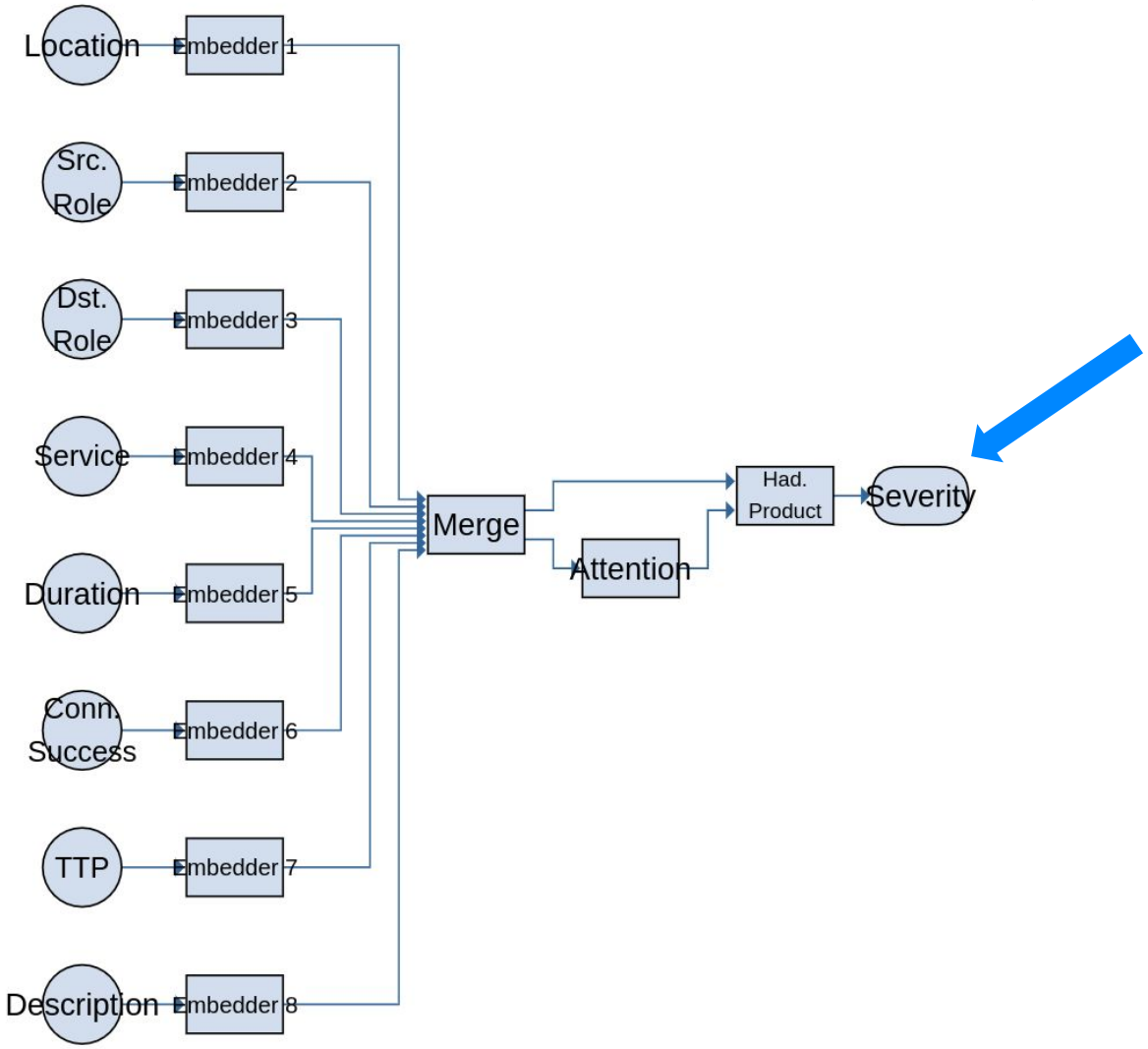
Model Architecture

Prioritizing and Explaining

We develop a Neural Network Regressor to prioritize network incidents

We employ a self-attention layer to produce explainable predictions

We use a single output node instead of a classifier due to the outputs being ordinal



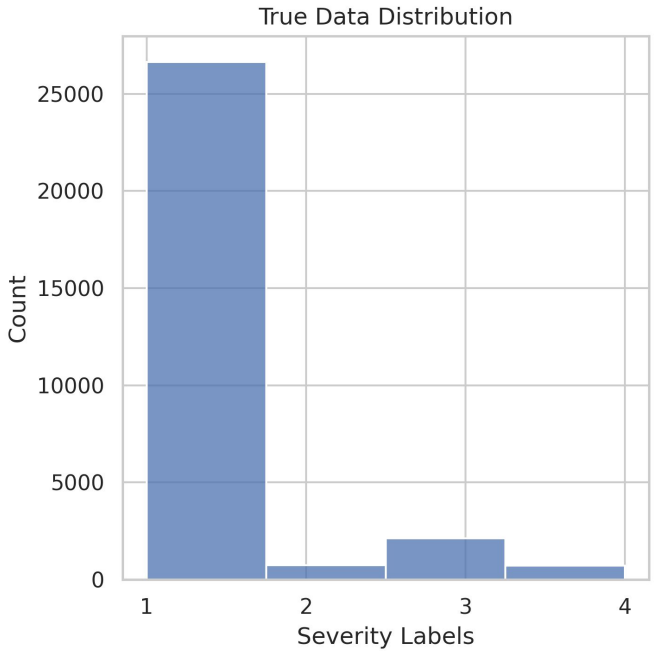
Experimental Setup

Training Parameters

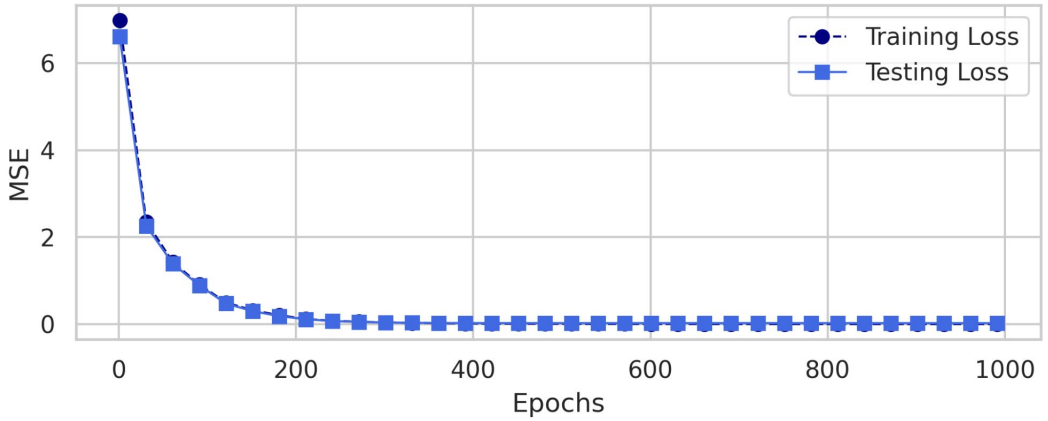
We balance the data set by sub-sampling to the smallest class:

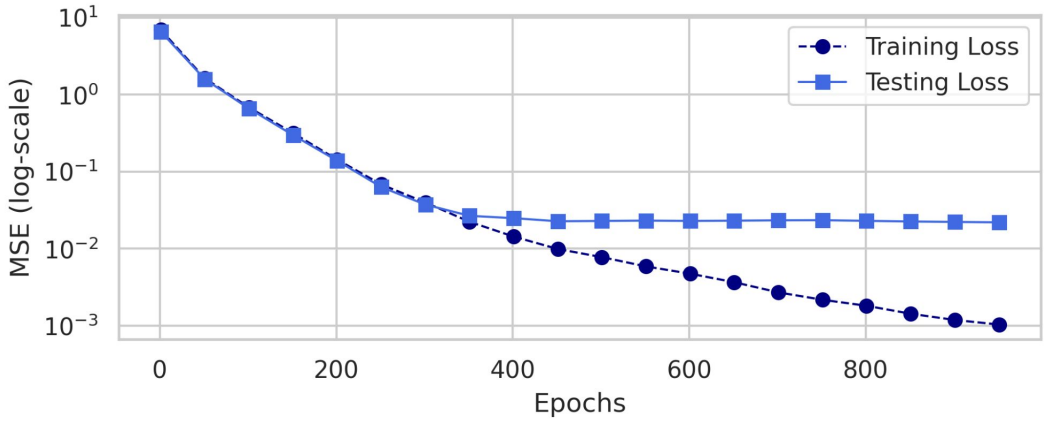
- **Medium Severity (~ 750)**
- **Critical Severity (~ 750)**

We then retain 80% of the balanced set for training and use all other reports as the evaluation set.



Results

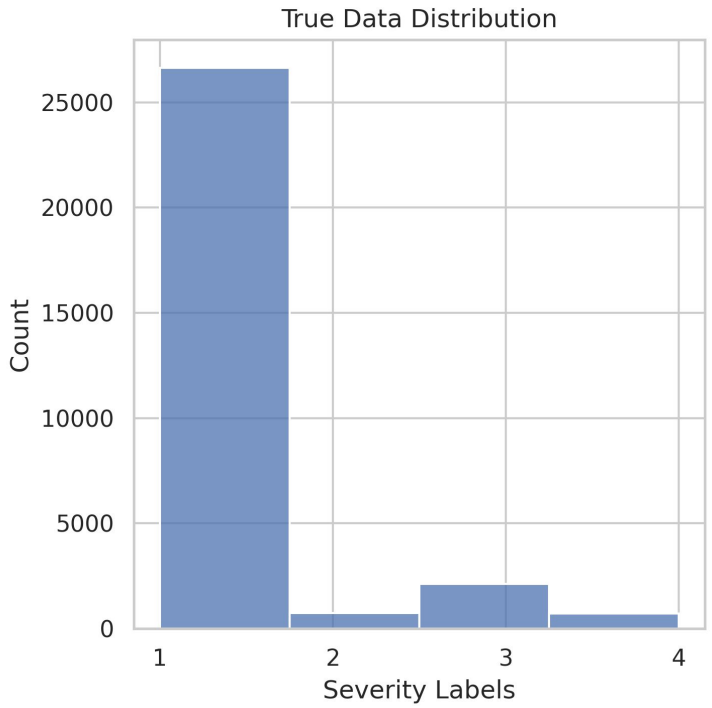

Training Loss
Testing Loss
MSE
0
2
4
6
0
200
400
600
800
1000
Epochs


Training Loss
Testing Loss
MSE (log-scale)
10^{1}
10^{0}
10^{-1}
10^{-2}
10^{-3}
0
200
400
600
800
Epochs

Results

Data Distribution

- **Balanced training data**
- **Model still achieves true data distribution on test set**



Results

Interpretation

- We probably have a bit of overfitting here
- We achieve reasonable F1 results on the evaluation set (with a caveat)

Training Set Metrics				
	Precision	Recall	F1-Score	Support
Low (1)	1.0	1.0	1.0	580
Medium (2)	0.99	0.99	0.99	580
High (3)	1.0	0.99	0.99	580
Critical (4)	1.0	1.0	1.0	580
Accuracy			1.0	2320
Macro Avg.	1.0	1.0	1.0	2320
Weighted Avg.	1.0	1.0	1.0	2320

Results

Interpretation

- We probably have a bit of overfitting here
- We achieve reasonable F1 results on the evaluation set (with a caveat)

Training Set Metrics				
	Precision	Recall	F1-Score	Support
Low (1)	1.0	1.0	1.0	580
Medium (2)	0.99	0.99	0.99	580
High (3)	1.0	0.99	0.99	580
Critical (4)	1.0	1.0	1.0	580
Accuracy			1.0	2320
Macro Avg.	1.0	1.0	1.0	2320
Weighted Avg.	1.0	1.0	1.0	2320

Results

Interpretation

- We probably have a bit of overfitting here
- We achieve reasonable F1 results on the evaluation set (with a caveat)

Testing Set Metrics				
	Precision	Recall	F1-Score	Support
Low (1)	1.0	1.0	1.0	26066
Medium (2)	0.33	0.99	0.49	150
High (3)	1.0	0.95	0.96	1530
Critical (4)	0.6	1.0	0.8	145
Accuracy			1.0	27891
Macro Avg.	0.7	1.0	0.8	27891
Weighted Avg.	1.0	1.0	1.0	27891

Results

Interpretation

We probably have a bit of overfitting here

We achieve reasonable F1 results on the evaluation set (with a caveat)

Testing Set Metrics				
	Precision	Recall	F1-Score	Support
Low (1)	1.0	1.0	1.0	26066
Medium (2)	0.33	0.99	0.49	150
High (3)	1.0	0.95	0.96	1530
Critical (4)	0.6	1.0	0.8	145
Accuracy			1.0	27891
Macro Avg.	0.7	1.0	0.8	27891
Weighted Avg.	1.0	1.0	1.0	27891

Results

Interpretation

We probably have a bit of overfitting here

We achieve reasonable F1 results on the evaluation set (with a caveat)

Testing Set Metrics				
	Precision	Recall	F1-Score	Support
Low (1)	1.0	1.0	1.0	26066
Medium (2)	0.33	0.99	0.49	150
High (3)	1.0	0.95	0.96	1530
Critical (4)	0.6	1.0	0.8	145
Accuracy			1.0	27891
Macro Avg.	0.7	1.0	0.8	27891
Weighted Avg.	1.0	1.0	1.0	27891

Attention

Interesting Observations

Duration is “never” important, but that’s a fault of the humans

Whether the connection was successful or blocked is less important for higher severity reports

This is also the case for the source IP/Host role

For the critical incidents “only” the type of attack is important

Attention Weights (Median %) [Test Data]					
	Overall	Low (1)	Medium (2)	High (3)	Critical (4)
Description	0.1	0.1	0.4	0.2	0.1
Successful	7.6	7.8	8.5	3.9	1.2
Duration	0	0	0	0	0
Source Role	90.5	90.8	55.4	30.2	2.6
Target Role	0.1	0.1	1.8	1.5	0
Service	0.1	0.1	0.4	0.2	0
Location	0.1	0.1	0.2	0.1	0
TTP	0.5	0.4	32.8	63.3	94.9
External	0	0	0.2	0.1	0

Attention

Interesting Observations

Duration is “never” important, but that’s a fault of the humans

Whether the connection was successful or blocked is less important for higher severity reports

This is also the case for the source IP/Host role

For the critical incidents “only” the type of attack is important

Attention Weights (Median %) [Test Data]					
	Overall	Low (1)	Medium (2)	High (3)	Critical (4)
Description	0.1	0.1	0.4	0.2	0.1
Successful	7.6	7.8	8.5	3.9	1.2
Duration	0	0	0	0	0
Source Role	90.5	90.8	55.4	30.2	2.6
Target Role	0.1	0.1	1.8	1.5	0
Service	0.1	0.1	0.4	0.2	0
Location	0.1	0.1	0.2	0.1	0
TTP	0.5	0.4	32.8	63.3	94.9
External	0	0	0.2	0.1	0

Attention

Interesting Observations

Duration is “never” important, but that’s a fault of the humans

Whether the connection was successful or blocked is less important for higher severity reports

This is also the case for the source IP/Host role

For the critical incidents “only” the type of attack is important

Attention Weights (Median %) [Test Data]					
	Overall	Low (1)	Medium (2)	High (3)	Critical (4)
Description	0.1	0.1	0.4	0.2	0.1
Successful	7.6	7.8	8.5	3.9	1.2
Duration	0	0	0	0	0
Source Role	90.5	90.8	55.4	30.2	2.6
Target Role	0.1	0.1	1.8	1.5	0
Service	0.1	0.1	0.4	0.2	0
Location	0.1	0.1	0.2	0.1	0
TTP	0.5	0.4	32.8	63.3	94.9
External	0	0	0.2	0.1	0

Questions

chae.clark@twosixtech.com