CatBERT: Context-Aware Tiny BERT for Detecting Targeted Social Engineering Emails

Authors: Younghoo Lee, Joshua Saxe, Richard Harang @ Sophos AI

1st KDD Workshop on AI-enabled Cybersecurity Analytics
The problem we’re solving
Targeted Phishing/Business Email Compromise (BEC)

Step 1: Identify a Target
- Organized crime groups target US and European businesses, exploiting information available online to develop a profile on the company and its executives.

Step 2: Grooming
- Spear phishing emails and/or telephone calls target victim company officials (typically an individual identified in the finance department).
- Perpetrators use persuasion and pressure to manipulate and exploit human nature.
- Grooming may occur over a few days or weeks.

Step 3: Exchange of Information
- The victim is convinced online is conducting a legitimate business transaction. The unwitting victim is then provided with instructions.

Step 4: Wire Transfer
- Upon transfer, the funds are steered to a bank account controlled by the organized crime group.

Business E-Mail Compromise Timeline
An outline of how the business e-mail compromise is executed by some organized crime groups

Detecting phishing is hard because language is hard

• Hard NLP problems
  o Co-reference resolution
  o Word polysemy
  o Sentiment detection

• Social engineering attacks
  o Hand-written, individually targeted emails.
  o Incorporate background research on their targets.
  o Find ways to bypass existing detection mechanisms.

https://www.fbi.gov/news/stories/business-e-mail-compromise
The Transformer revolution

• Pre-Transformer approaches
  o Mostly didn’t consider words in context
  o Mostly didn’t provide attention mechanisms
  o Mostly operated on either words (too coarse) or characters (too fine grained)

• Transformers
  o Words are given contextual representation
  o Attention mechanisms build into models
  o Use efficient sub-word representations
  o Take advantage of modern neural net ideas and technologies
  o However, full-sized models are computationally expensive and slow to use in high volume.
CatBERT
(Context-Aware Tiny BERT)
Model Architecture and Goals

- **High Accuracy**
  - By combined content and context input

- **Fast Inference**
  - By downsizing with Adapters with no cost to accuracy
Inputs for Context-Aware BERT

- Context input from header
  - From, To, CC, Reply-To fields provide context information about the communication
  - Features include communication type, number of recipients and CCs

- Content input from text
  - Text data from Subject + body provides the intention of the communication

```
| From:     | Jun.Jardon@fungamex.com |
| To:       | Jag.Dost@fungame.com    |
| Reply-To: | Jue.Jardon@gmail.com    |
| Subject:  | Wire Transfer           |
|           | Hi Jag,                 |
|           | I will need you to process an urgent transfer payment, which needs to go out today. Let me know when you are set to proceed with the payment. |
|           | Regards,                |
|           | Jun                     |
```
Model Compression with Adapters

- **Adapter Block**
  - 2 Dense units with a non-linear activation unit.
  - Dimension of Dense units is same as Transformer’s output one.
  - There is a skip connection to bypass the block if necessary.
Model Compression with Adapters

- Partial Fine-tuning
  - Lower blocks (Embeddings, Transformer1 and 3) are fixed to minimize forgetting of learned representations.
  - Upper blocks (Transformer5 and head) and Adapters are jointly fine-tuned.
Results
Datasets and Performance Metrics

• Dataset
  o Benign emails: 3.8M emails
  o Malicious emails: 407K Phishing and 1K BEC emails
  o Data time split by 70%, 15%, 15% as a training, validation and test set

• Baseline models
  o DistilBERT6: distilled BERT with 6 Transformers from BERT base with 12 Transformers
  o LSTM: RNN model with BERT’s embedding layer
  o LR: Logistic Regression model with TF-IDF features

• Training
  o Baseline code from HuggingFace’s PyTorch version
  o Trained on a p3.8xlarge/AWS instance which has 4 NVIDIA Telsa V100 GPUs

• Performance metrics
  o ROC curves and AUC
  o Inference speed and model size
CATBERT outperformed baseline models.
Ablation Study with Adapter and Context Input

Adapter and Context layers improved performance.
Performance against Adversarial attacks

CatBERT is more robust than baseline models.
## Model Size and Inference Speed

<table>
<thead>
<tr>
<th>Model Type</th>
<th># of Transformer blocks</th>
<th># of Embedding layer parameters</th>
<th># of Transformer blocks parameters</th>
<th># of total parameters</th>
<th>CPU inference speed (milliseconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multilingual DistilBERT</td>
<td>6</td>
<td>92 million (100%)</td>
<td>42 million (100%)</td>
<td>135 million (100%)</td>
<td>130 (100%)</td>
</tr>
<tr>
<td>Multilingual CatBERT</td>
<td>3</td>
<td>92 million (100%)</td>
<td>23 million (54%)</td>
<td>117 million (85%)</td>
<td>79 (60%)</td>
</tr>
</tbody>
</table>

CatBERT is smaller and faster than the baseline model.
Summary

• An efficiently downsized CaTBERT achieved both high speed and high accuracy in detecting hand-crafted social engineering email attacks.
  o By fine-tuning a highly pruned BERT with Adapters
  o By combining email text content with header context information