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

SOPHOS

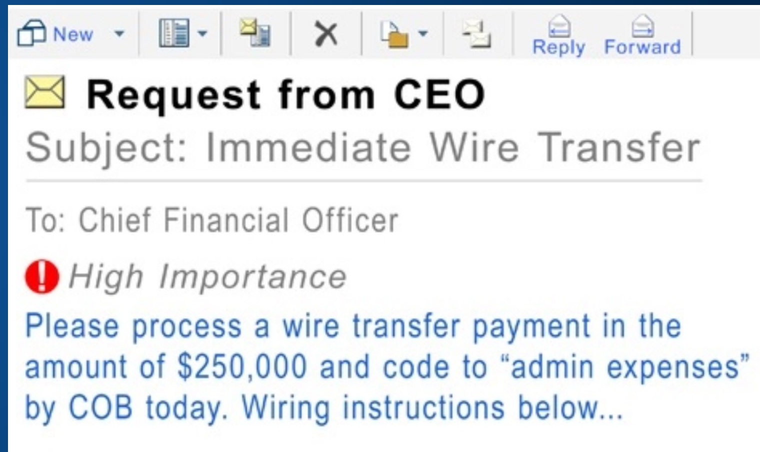
The problem we're solving

Targeted Phishing/Business Email Compromise(BEC)

https://www.fbi.gov/scams-and-safety/common-scams-and-crimes/business-email-compromise

Detecting phishing is hard because language is hard

- Hard NLP problems
  - Co-reference resolution
  - Word polysemy
  - Sentiment detection
- Social engineering attacks
  - Hand-written, individually targeted emails.
  - Incorporate background research on their targets.
  - Find ways to bypass existing detection mechanisms.



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

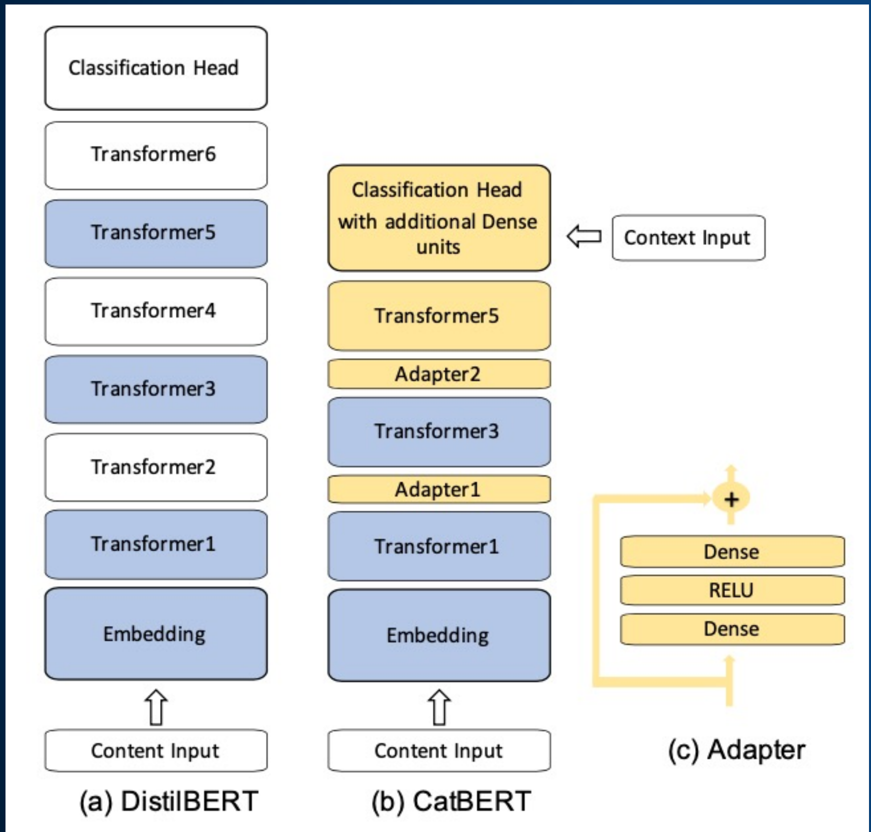
The Transformer revolution

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

- Transformers
  - Words are given contextual representation
  - Attention mechanisms build into models
  - Use efficient sub-word representations
  - Take advantage of modern neural net ideas and technologies
  - ***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

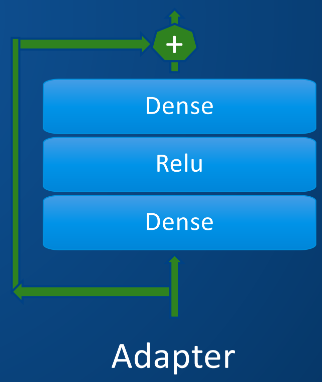
Inputs for Context-Aware BERT

- 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

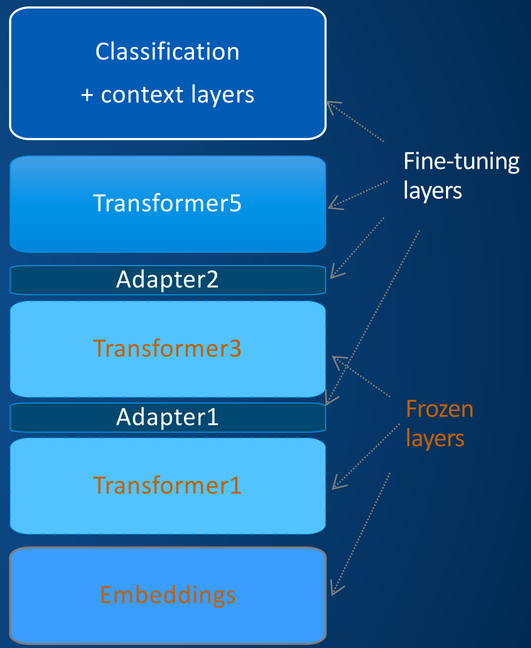
Model Compression with Adapters

- 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.



SOPHOS

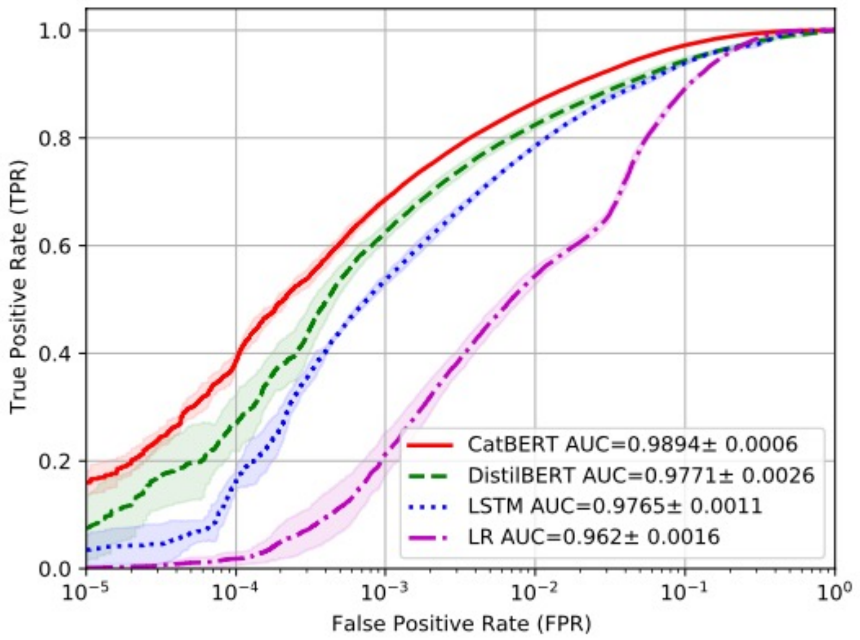
Results

SOPHOS

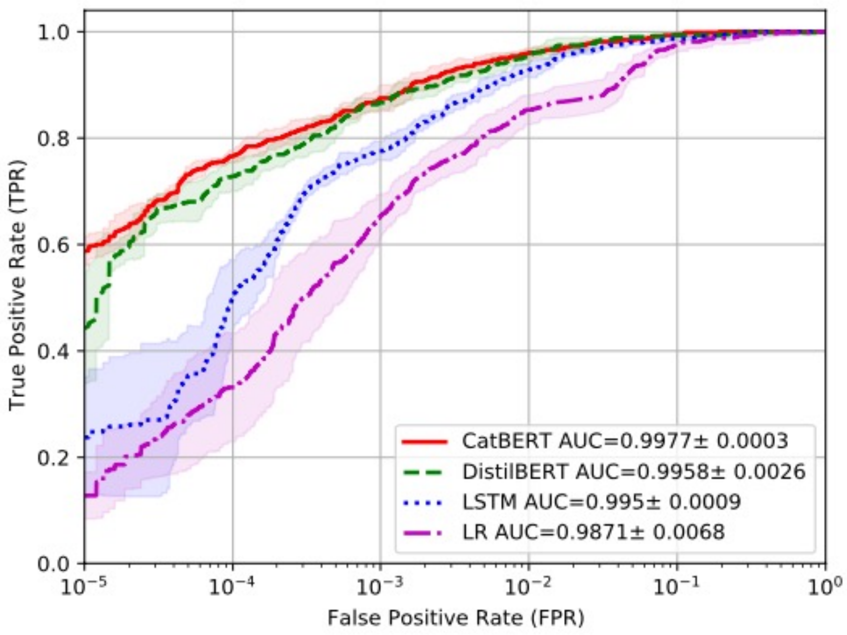
Datasets and Performance Metrics

- Dataset
  - Benign emails: 3.8M emails
  - Malicious emails: 407K Phishing and 1K BEC emails
  - Data time split by 70%, 15%, 15% as a training, validation and test set
- Baseline models
  - DistilBERT6: distilled BERT with 6 Transformers from BERT base with 12 Transformers
  - LSTM: RNN model with BERT's embedding layer
  - LR: Logistic Regression model with TF-IDF features
- Training
  - Baseline code from HuggingFace's PyTorch version
  - Trained on a p3.8xlarge/AWS instance which has 4 NVIDA Telsa V100 GPUs
- Performance metrics
  - ROC curves and AUC
  - Inference speed and model size

ROC Curves



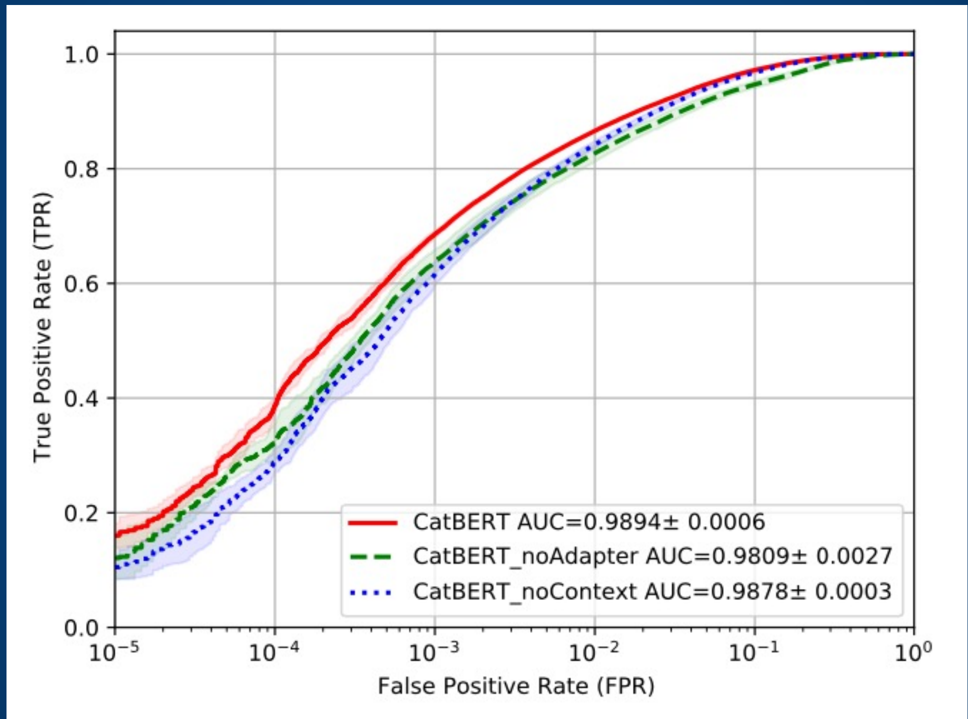
(a) All samples



(b) BEC samples

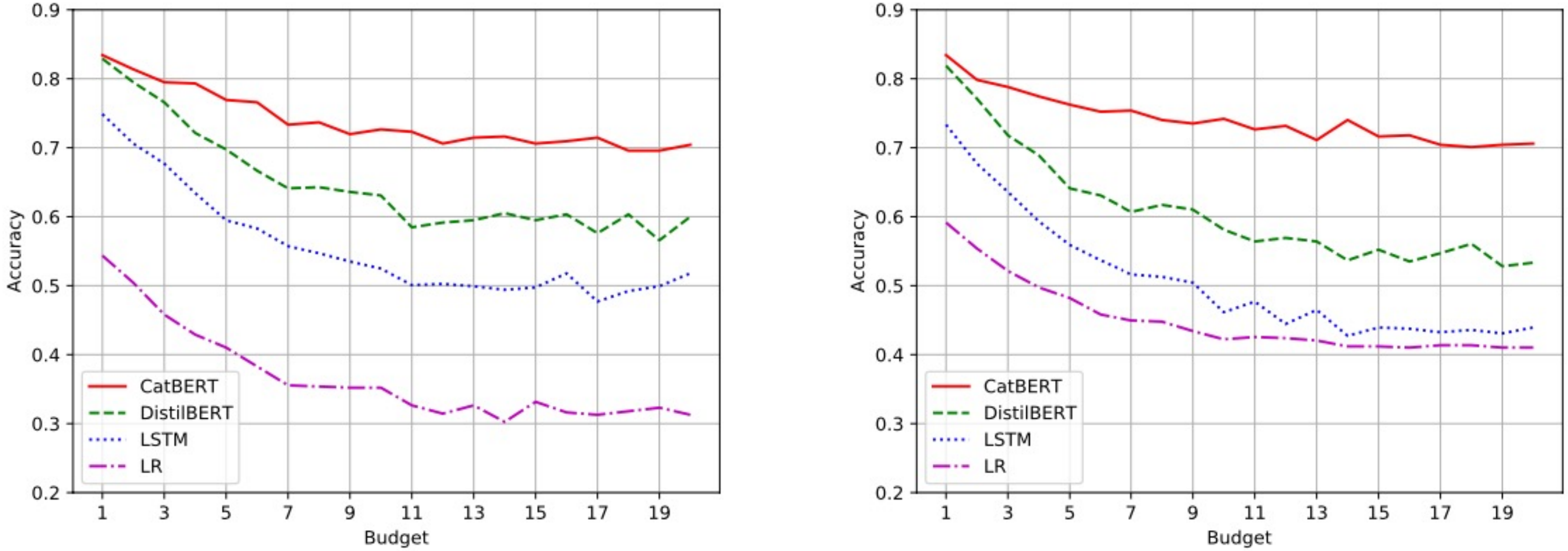
CATBERT outperformed baseline models.

Ablation Study with Adapter and Context Input



Adapter and Context layers improved performance.

Performance against Adversarial attacks



(a) Synonyms attacks **(b)** Typo attacks

CatBERT is more robust than baseline models.

Model Size and Inference Speed

	# of Transformer blocks	# of Embedding layer parameters	# of Transformer blocks parameters	#of total parameters	CPU inference speed (milliseconds)
Multilingual DistilBERT	6	92 million (100%)	42 million (100%)	135 million (100%)	130 (100%)
Multilingual CaTBERT	3	92 million (100%)	23 million (54%)	117 million (85%)	79 (60%)

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.
  - By fine-tuning a highly pruned BERT with Adapters
  - By combining email text content with header context information