Microsoft

ICASSP2020 Barcelona

Texception: A Character/Word-Level Deep Learning Model For Phishing URL Detection

Farid Tajaddodianfar*, Jack W. Stokes**, Arun Gururajan**

* Currently at Amazon Inc., Seattle WA, USA. Farid performed this research while being at Microsoft
** Microsoft Corp., Redmond WA, USA

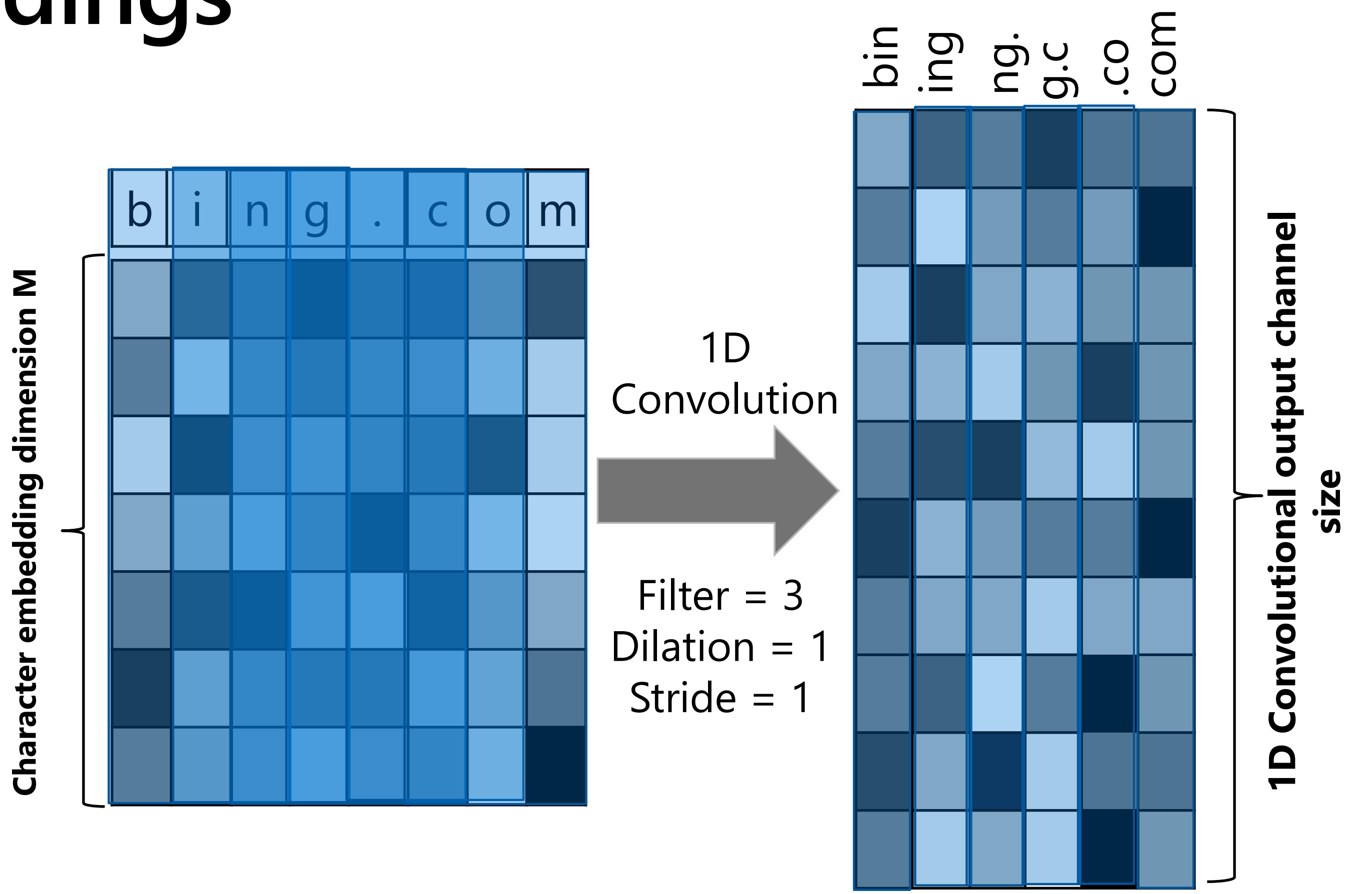
URL string to reveal phishing content

- URL string can be indicative of maliciousness of its host
  - Specific patterns in URL string can be used over and over by bad actors
- There are signals other than URL (e.g. HTML content, page DOM, etc)
  - Need significant level of transformation before being used
- URL can be used as a relatively strong and available feature to detect phishing attacks

Convolution on Character Embedding

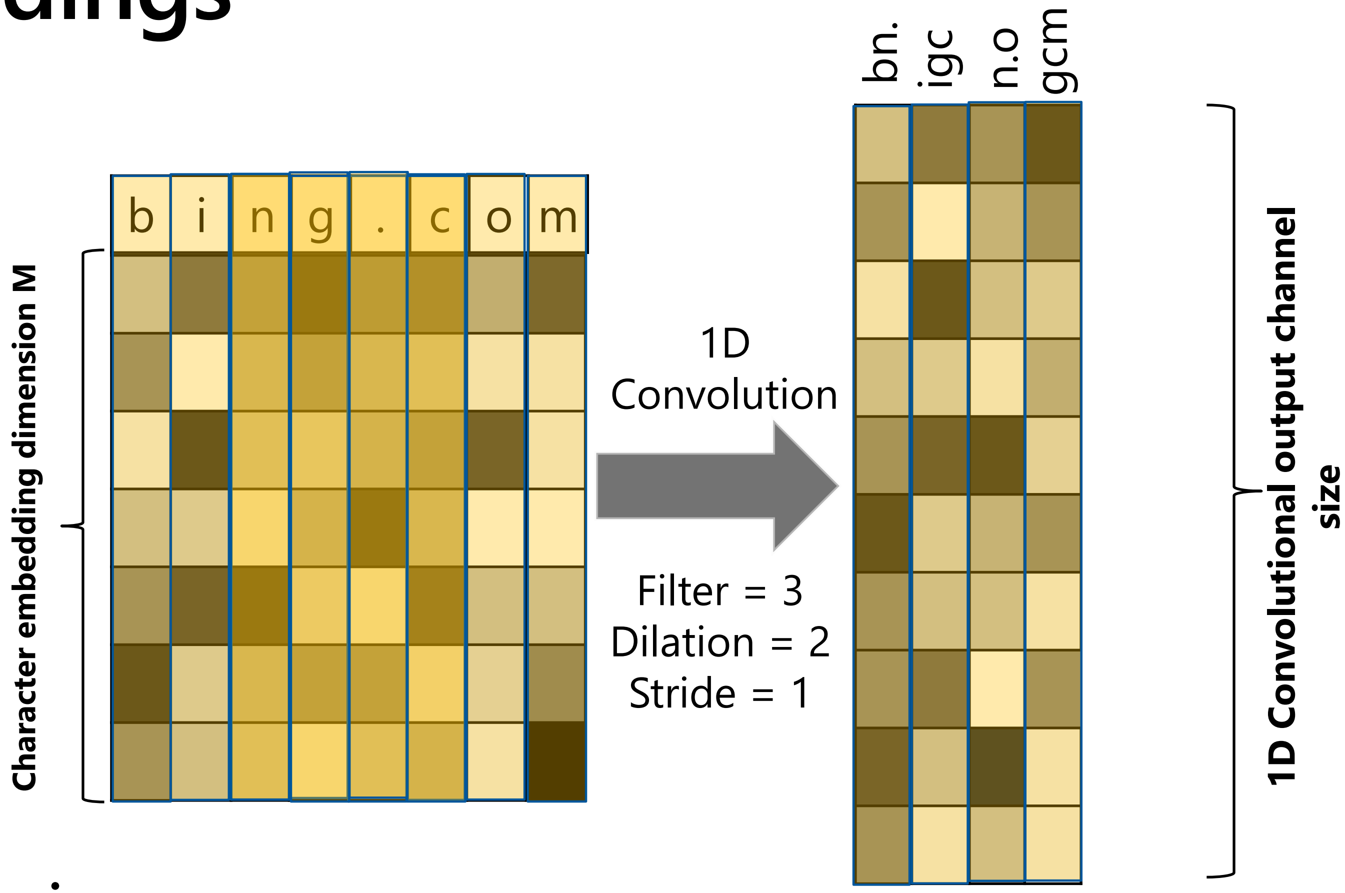
CNN on Character embeddings

- Assign a vector to represent each character in the alphabet of observed characters
- CNN with fixed filter size N is analogous to extracting N-grams



CNN on Character embeddings

- CNN with dilation resembles extracting skip-grams

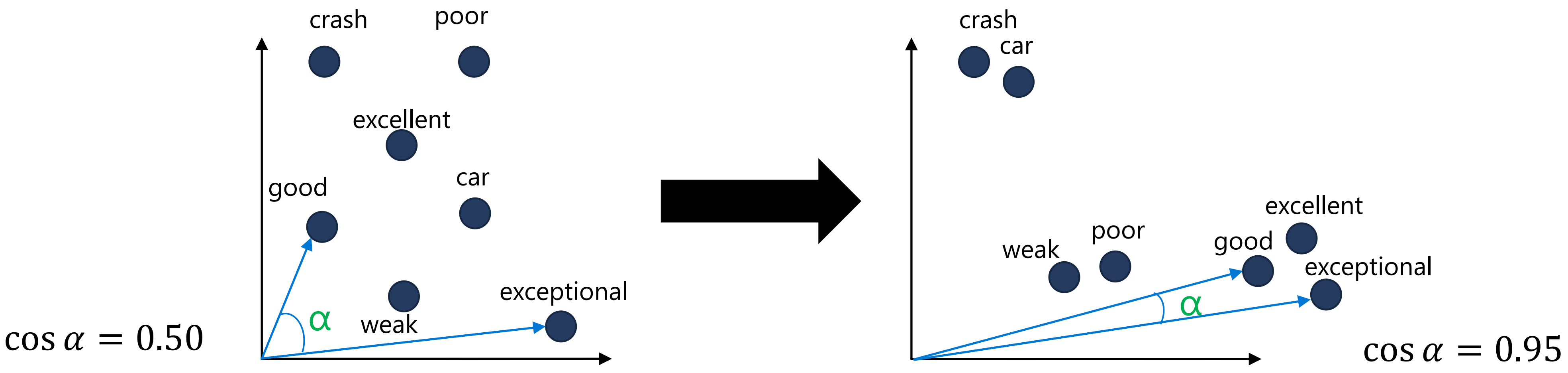


- Previous methods use “fixed” filter size
- That effectively limits model’s flexibility to “fixed” n-grams. E.g. only 3-grams

Convolution on Word Embedding

Word Embedding Concepts

- First proposed to capture semantic similarity in NLP
- Given a corpus of natural language text, randomly assign an N-dimensional vector to each word
- During training, move vectors closer if they appear in the same context
- Trained in unsupervised settings (Word2Vec, FastText, GloVe)



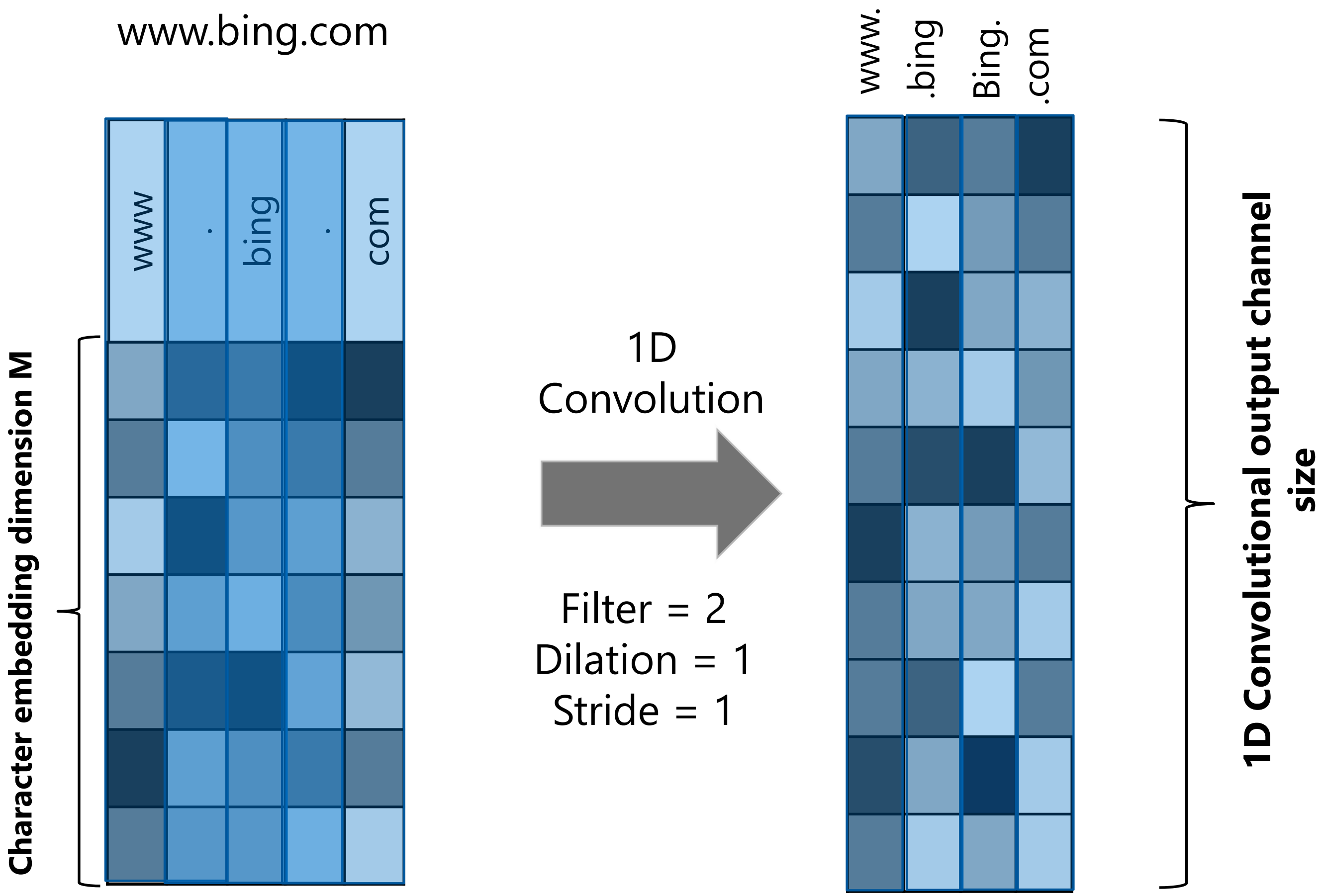
"More than a good movie; it's really exceptional"

FastText Word embedding on URL

- Tokenize URL based on special characters [/.;$:] and treat it as a sentence!
- Treat each special character as a word too
- Train FastText word embedding on the whole training data
- Use the trained embeddings in the downstream classification task

CNN on Word embeddings

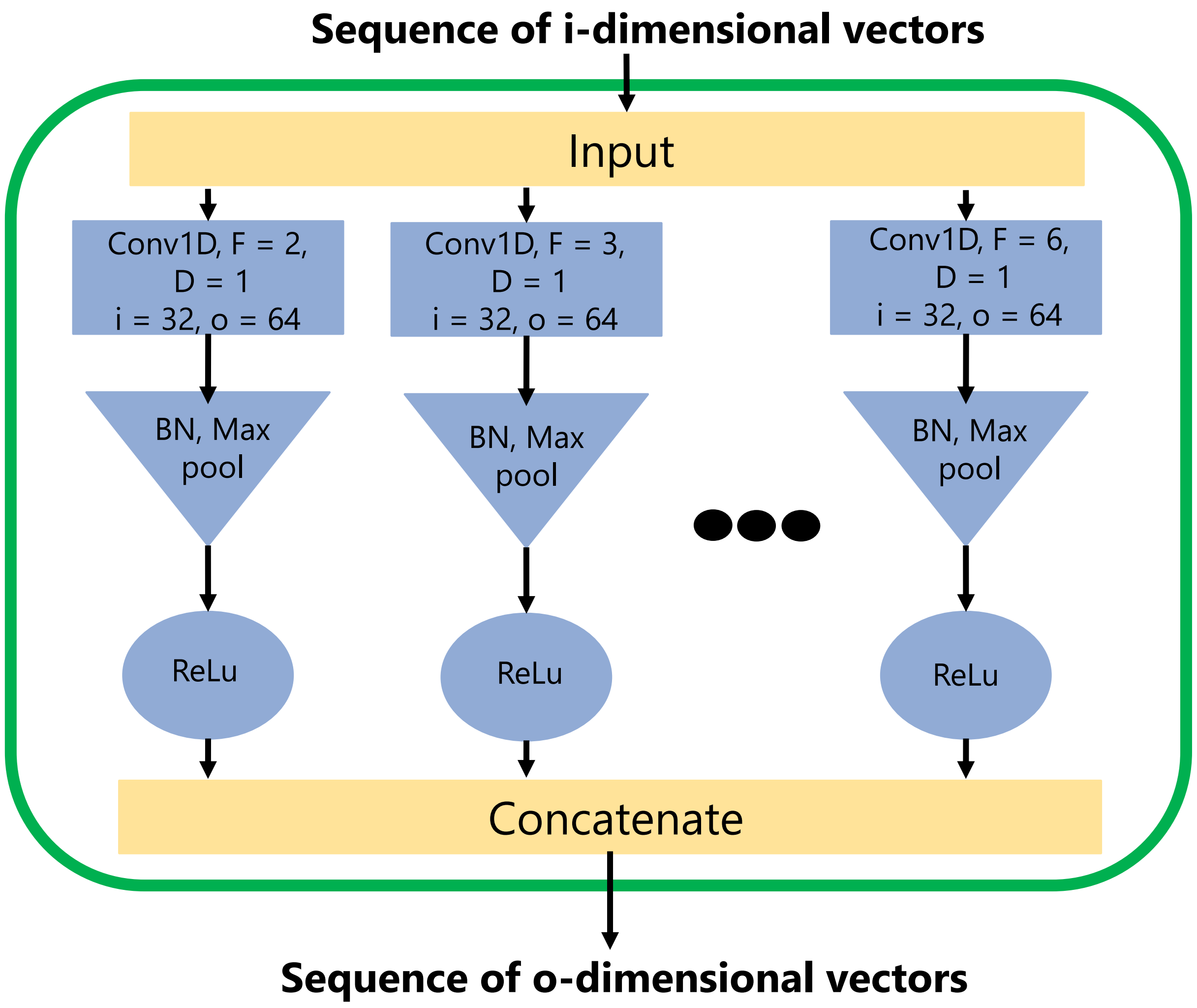
- Each word is represented by the FastText embedding vector
- CNN with fixed filter size N is analogous to extracting word N-grams
- Using fixed filter size limits the model flexibility
- Why not more?



Texception Block
&
Texception Model

Texception Block

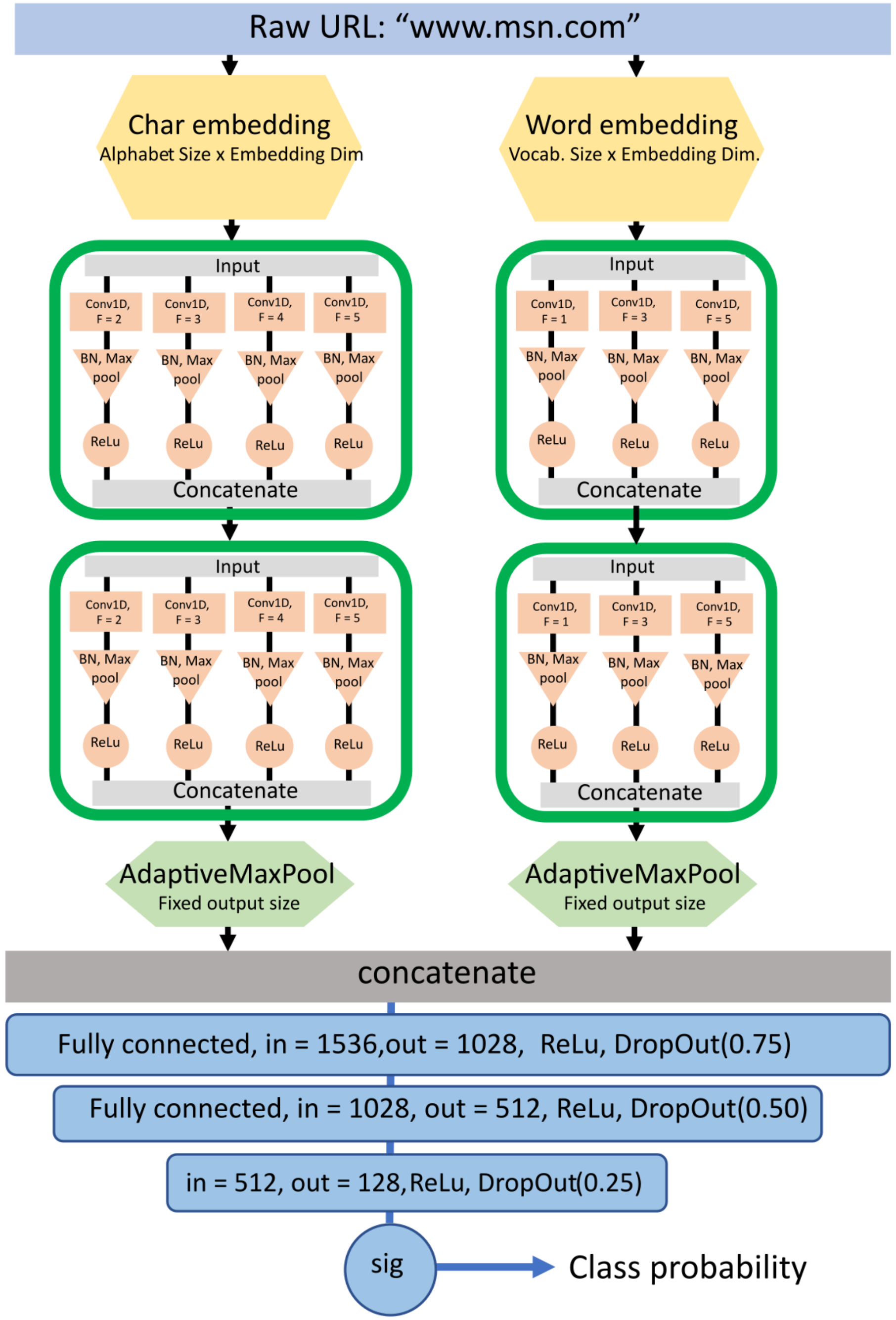
F = filter size
D = Dilation
i = input channel
o = output channel
w = Max pool window



- Multiple parallel Convolutional layers inside
- Each with Batch Normalization, Max Pooling, ReLu
- Features extracted from each path are concatenated and outputted

Texception Model

- Multiple branches to extract information from text
- Can grow wider or deeper at each branch
- Data scientist controls bottleneck of information flow
- Provides sufficient flexibility to learn from text



Hyper parameters

Parameter		value
Characters Branch	embedding dimension	32
	number of blocks	1
	block filters	[2,3,4,5]
	Adaptive MaxPool output	32,32
	maximum characters	1000
Words Branch	embedding dimension	32
	number of blocks	1
	block filters	[1,3,5]
	Adaptive MaxPool output	32,16
	maximum words	50
FastText Model	minimum words to include	50
	vocabulary size	120000
	window size	7
	n-grams	2-6
	embedding dimension	32
	epochs trained	30

Table 1. Hyperparameters used for experimentation.

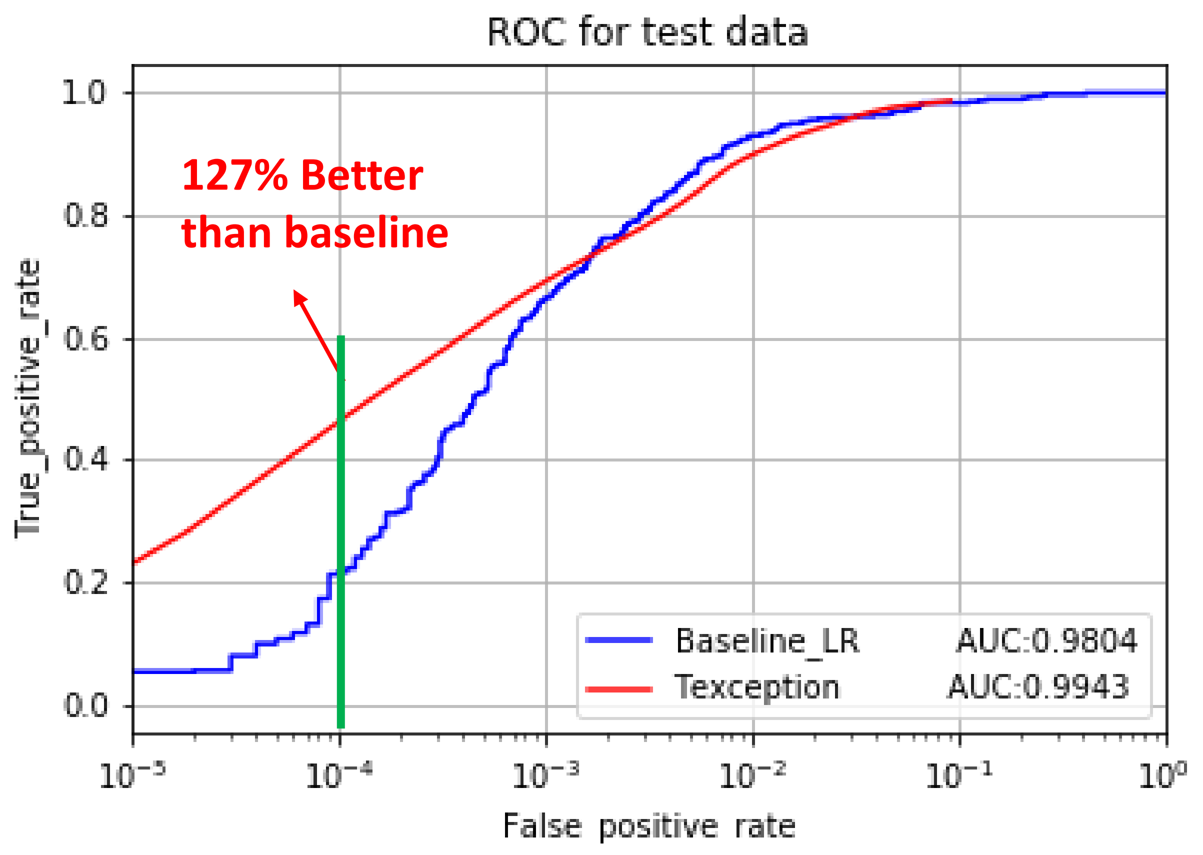
Experimentation

Train set:

- 2 weeks of traffic labeled by proprietary labeling systems
- Clean class down sampled to keep positive class ratio at 5%
- 1.7M samples in training set
- 20% used as validation set

Test set:

- No down sampling to keep the distribution like the real production data
- Positive class ratio at 0.01%
- 20M samples



Results

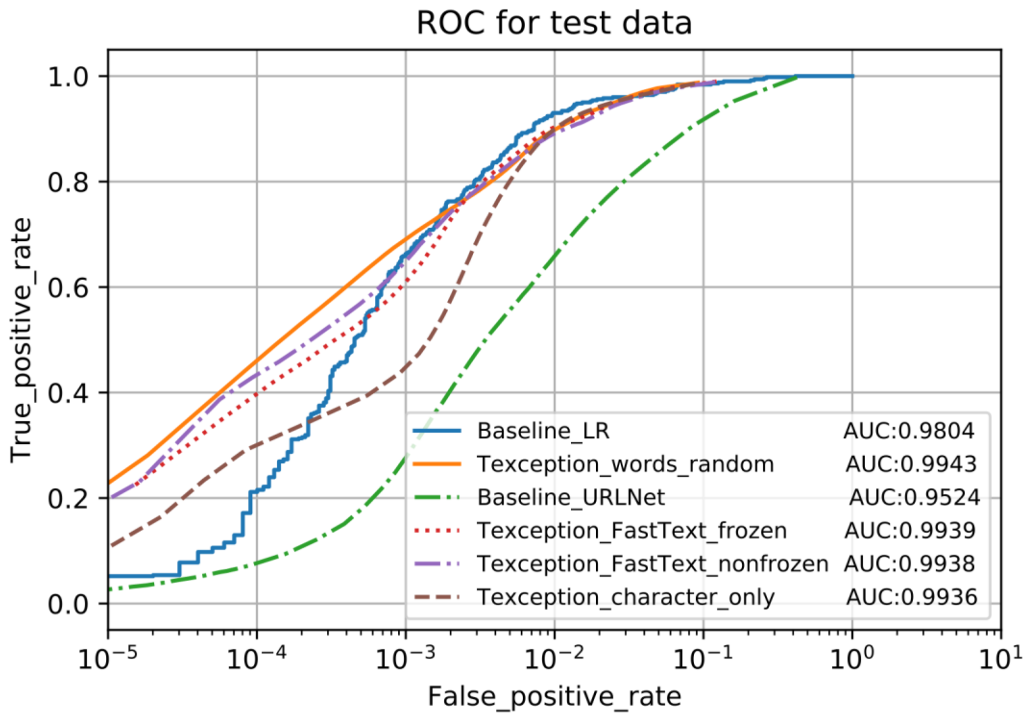
Superior performance at very high precision

- Are word embeddings useful?
- Is it better to use pre-trained embeddings or just start with random weights and let the model learn the embeddings itself?
- If using pretrained word weights, is it better to freeze the word weights or let the model adjust them during training?

Good to use word embeddings
Even better to randomly initialize them!

Results

What is the effect of word embedding?



Good to use word embeddings
Even better to randomly initialize them!

Results

Moving beyond URL

Texception is applicable where data is in the form of text string

- File name
- CTPH
- Command lines

Not limited to security data

- Texception is a generic model for text classification

Thank You!